

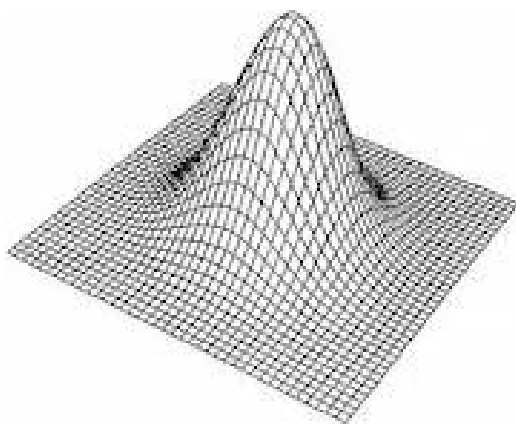


ESTADÍSTICA II

ESTADÍSTICA INFERENCIAL PARAMÉTRICA, NO PARAMÉTRICA Y MULTIVARIANTE

Notas de apoyo para el curso
Primera Edición, febrero de 2012

Dr. Oscar Valdemar De la Torre Torres



Academia de Matemáticas
Facultad de Contaduría y Ciencias Administrativas

Universidad Michoacana de San Nicolás de Hidalgo
(Cuna de héroes, crisol de pensadores...)



Resumen:

Las presentes notas de clase son para el curso de Estadística II que comprende los temas de muestreo, inferencia estadística paramétrica y no paramétrica y análisis multivariante, en concreto, análisis de regresión. Las mismas se redactan con la finalidad de dar al alumno de licenciatura un apoyo de estudio y síntesis de la temática tratada durante la clase.

Uso de estas notas del profesor

Por favor, dese el tiempo de leer esto. Le tomará, a lo mucho, diez minutos.

El objetivo de las presentes es hacer que su aprendizaje sea ameno y simple y que no sienta usted una presión psicológica tal que solo se enfoque en obtener una buena nota para su promedio o, en el peor de los casos, simplemente aprobar la materia para continuar con su licenciatura.

Como lo apreciará tanto en la clase introductoria al curso como en el plan de trabajo del mismo publicado en el sitio que previamente le indicaron, la finalidad del profesor y de las presentes es que usted enfoque sus energías a solamente aprender, estudiar, hacer el trabajo que le asignen y aprovechar el tiempo de clase. Por tanto, estos apuntes tienen un diseño simple pero que se busca sea pedagógico para usted. Adicional a lo ameno que se busca redactar, durante el contenido verá definiciones que se resaltarán y que será su obligación aprender y memorizar. Tenga usted la confianza de que no le será complicado recordarlos durante clase. Sin embargo, el conocer estos conceptos le garantizará tener una buena respuesta tanto en las pruebas de control como en las evaluaciones parciales ya que dichas definiciones y conceptos serán preguntados en las mismas. Por ejemplo:

“La Filosofía, como lo señalan los académicos de la Universidad de La Sapienza, se define como ‘El conjunto de concepciones sobre los principios y las causas del ser de las cosas, del universo y del hombre’, situación consistente con su origen etimológico ‘*Philos*’, que significa amigo y ‘*Sophia*’ que significa sabiduría...”

Después del párrafo, usted verá algo así:

Filosofía: El conjunto de concepciones sobre los principios y las causas del ser de las cosas, del universo y del hombre.

También podrán venir definiciones, comentarios o fórmulas durante el contenido que no pueden separarse como una definición independiente pero que, sin embargo, usted nunca debe olvidar y que se señalan con un fondo gris ya que son clave para su aprendizaje:

Para la definición de contenidos:



“... Como se puede apreciar, gracias a la iniciativa de George Washington y Lafayette, Luis XVI apoyó la independencia de Estados Unidos y mandó a la ruina económica a Francia. Esta ruina generó la revolución francesa que llevaría a Napoléon Bonaparte al poder. Es entonces que **el asenso de Napoléon como emperador y su invasión a España, fue el principal acontecimiento histórico europeo que motivó la independencia de México...**”

Este concepto difícilmente lo olvidará al estar resaltado y, a su vez, será la respuesta de una pregunta de prueba de control en clase o examen del tipo:

¿Cuál fue el principal acontecimiento histórico europeo, adicional a muchos otros suscitados en la Nueva España, que motivó la independencia de México?

Usted ya sabe la respuesta (NOTA: la pregunta en un examen puede venir de diversas formas y redacciones, con respuestas de opción múltiple, completar, etc. Este es un simple ejemplo de lectura de las presentes notas)

Para la definición de fórmulas:

“Por tanto, con la derivación previamente empleada, el área de un círculo se define como:

$$A = \pi \cdot r^2$$

Fórmula 1 área de un círculo:

$$A = \pi \cdot r^2$$

El Dr. De la Torre espera que estas notas sean de su provecho. No se deje impresionar si cree que es mucho material para estudiar. Ya verá usted que la carga de materia es amena y fácil de llevar.

Es importante señalar que las presentes son una parte fundamental de la materia pero en ningún momento se está afirmando que son la única fuente que debe usted estudiar y repasar. En muchas ocasiones se revisarán temas y se harán comentarios que pueden no venir en estas líneas y que, sin embargo, pueden preguntarse en las pruebas de control o examen. Por tanto, se le sugiere llevar sus propias notas a mano en clase, asistir a la misma y poner atención a todos los comentarios e indicaciones hechos por el profesor para evitar omisiones.

Dr. Oscar De la Torre.



Índice temático

| | |
|--|-----------|
| Resumen: | 2 |
| Uso de estas notas del profesor | 2 |
| 1 Introducción: Repaso de Estadística, conceptos y definiciones. | 7 |
| 1.1 ¿Para qué estudiamos esto?, ¿Cómo se come?, ¿Es un desperdicio de mi tiempo el estudiar Estadística si soy contador, administrador o informático? | 7 |
| 1.2 Repaso de conceptos y definiciones de Estadística I. | 9 |
| 1.2.1 La probabilidad ¿Qué es y cómo se cuantifica? | 11 |
| 1.3 Medidas de tendencia central y medidas de dispersión. | 14 |
| 1.3.1 La media, la mediana y la moda | 14 |
| 1.3.2 La varianza y la desviación estándar ¿qué significan? y ¿Por qué la calculamos la varianza elevando al cuadrado las diferencias respecto a la media? | 16 |
| 1.3.3 Reglas de dedo para calcular la media y la desviación estándar: | 20 |
| 1.4 Cálculo de probabilidades: los histogramas, las funciones y distribuciones de probabilidad. | 21 |
| 1.4.1 Mapa mental de lo hasta ahora visto | 21 |
| 1.4.2 Eventos aleatorios (variables aleatorias) discretos y continuos | 21 |
| 1.4.3 Cálculo de probabilidades en variables aleatorias discretas: El histograma. | 22 |
| 1.4.4 Distribuciones de probabilidad. | 26 |
| 1.4.5 Funciones de densidad de probabilidad | 28 |
| 1.4.5.1 Cálculo de probabilidades con función de densidad de probabilidad normal o gaussiana. | 32 |
| 1.4.6 La función de densidad de probabilidad normal estándar. | 34 |
| 1.4.6.1 Regla de dedo para comprender por qué utilizar una distribución normal estándar: | 37 |
| 1.4.7 El cálculo de la probabilidad utilizando la normal estándar y las tablas correspondientes. | 37 |
| 1.4.7.1 Diferentes formas de calcular una probabilidad. Los valores de probabilidad acumulada. | 39 |
| 2 Teoría del muestreo | 47 |
| 2.1 Tipos de muestreo | 47 |
| 2.2 Muestreo aleatorio simple | 48 |
| 2.3 Muestreo sistemático | 48 |
| 2.4 Muestreo estratificado. | 49 |
| 2.5 Muestreo de racimo. | 49 |
| 2.6 Diferencias operativas en cada uno de los tipos de muestreo y determinación del empleado en Estadística Inferencial. | 50 |
| 2.7 Diseño de un experimento: el proceso que se sigue para tomar decisiones. | 50 |



| | | |
|-------------|---|------------|
| 2.8 | Distribuciones de probabilidad muestrales | 52 |
| 2.8.1 | Las estadísticas necesarias para calcular la distribución normal muestral | 55 |
| 2.8.2 | Media muestral | 55 |
| 2.8.3 | Error estándar | 56 |
| 2.8.4 | Cálculo de probabilidades con muestras. | 59 |
| 2.9 | El teorema del límite central y una primera forma de determinar el tamaño adecuado de la muestra | 61 |
| 2.10 | El multiplicador de población finita | 64 |
| 3 | Estimaciones puntuales y de intervalo. La base de la inferencia estadística. | 66 |
| 3.1 | Consideraciones para calcular verdaderas estimaciones de intervalo | 67 |
| 3.1.1 | El verdadero cálculo del error muestral cuando se desconoce la desviación estándar de la población. | 69 |
| 3.1.2 | La estimación de intervalo. | 69 |
| 3.2 | ¿Qué pasa cuando nuestra muestra de datos no es grande? La distribución t-Student | 76 |
| 3.2.1 | Los parámetros para calcular la distribución t-Student y su empleo para el cálculo de estimaciones de intervalo. | 77 |
| 3.3 | Estimaciones de intervalo para comparar medias. | 80 |
| 3.3.1 | Estimaciones de intervalo para muestras apareadas grandes y pequeñas. | 80 |
| 3.3.1.1 | Estimación de intervalo para muestras apareadas grandes | 80 |
| 3.3.1.2 | Estimación de intervalo para muestras apareadas pequeñas | 85 |
| 3.3.2 | Estimaciones de intervalo para muestras independientes. | 86 |
| 3.4 | ¿Cómo determinar el intervalo de confianza? | 89 |
| 3.5 | ¿Cómo determinar el tamaño de muestra cuando se busca incrementar la precisión del intervalo de confianza? | 90 |
| 4 | Prueba de hipótesis: La técnica clásica | 94 |
| 4.1 | Comprobación de hipótesis de una sola muestra. | 95 |
| 4.1.1 | Ejemplos de los diferentes tipos de prueba de hipótesis con técnica clásica aplicados a una muestra simple. | 97 |
| 4.1.1.1 | Pruebas de hipótesis para demostrar igualdad de la media muestral con una media poblacional conocida o hipotética. | 97 |
| 4.1.1.2 | Pruebas de hipótesis para demostrar desigualdad de la media muestral con una media poblacional conocida o hipotética. | 107 |
| 4.1.1.3 | Ejemplos de pruebas de hipótesis de cola superior. | 115 |
| 4.1.1.4 | Ejemplos de pruebas de hipótesis de cola inferior. | 123 |
| 4.2 | ¿Cuándo se utiliza la escala original y cuándo la estandarizada? | 125 |
| 4.3 | ¿Qué se hace cuando se desconoce la desviación estándar poblacional? | 126 |
| 4.4 | Pruebas de hipótesis para comparar muestras. | 127 |
| 5 | Prueba de hipótesis: Las técnicas Ji-cuadrada y ANOVA | 131 |



| | | |
|------------|---|------------|
| 5.1 | La técnica Ji-Cuadrada | 131 |
| 5.1.1 | Prueba de hipótesis para demostrar independencia. | 131 |
| 5.1.2 | Distribución de probabilidad ji-cuadrada. | 135 |
| 5.1.3 | Algunas consideraciones a tomar con la prueba ji-cuadrada. | 141 |
| 5.1.4 | Prueba de hipótesis ji cuadrada para bondad de ajuste (determinar la función de probabilidad a emplear en un grupo de datos). | 141 |
| 5.1.5 | Prueba de hipótesis ji-cuadrada para hacer inferencias sobre la varianza de una sola población (o muestra). | 146 |
| 5.1.6 | Haciendo estimaciones de intervalos de varianzas. | 148 |
| 5.2 | Prueba ANOVA. | 149 |
| 5.2.1 | La función de probabilidad F. | 154 |
| 5.2.2 | La prueba F. | 154 |
| 5.2.3 | Prueba ANOVA para probar la igualdad en la varianza entre dos muestras. El caso de la cola superior. | 156 |
| 6 | Estadística multivariada: Regresión lineal simple y multivariada. | 160 |
| 6.1 | El coeficiente de correlación y su interacción con la covarianza. | 165 |
| 6.2 | El modelo regresión lineal simple para establecer relaciones estadísticas entre variables y hacer pronósticos básicos. | 167 |
| 6.2.1 | Determinación de los coeficientes del modelo de regresión. | 167 |



1 Introducción: Repaso de Estadística, conceptos y definiciones.

1.1 ¿Para qué estudiamos esto?, ¿Cómo se come?, ¿Es un desperdicio de mi tiempo el estudiar Estadística si soy contador, administrador o informático?

Antes de iniciar con la parte prominentemente técnica y bonita de la materia, es importante sensibilizar al alumno de la necesidad de la misma y cómo podrá esta cambiar su vida y la de los seres que lo rodean a través del conocimiento que asimile.

Un primer impulso que todos los alumnos tienen es creer que, por el hecho de estudiar para contadores, administradores o informáticos, la Estadística sirve solo para rellenar la currícula que se ofrece en la universidad y que nunca la utilizarán en su vida futura. Esto es realmente tanto falso como trágico ya que, como futuros profesionistas independientes, investigadores, dueños de empresas o cabezas de gobierno (por citar casos de dónde se ejercerá su maravillosa profesión y en donde se necesita de esta materia) deberán tomar decisiones.

En otros casos usted se deberá saber si lo que se está haciendo es o será realmente apropiado, por lo que es de necesidad conocer si lo que se planea realizar; como puede ser un negocio, la apertura de una empresa o línea de producto, la realización de una inversión, la implementación de un sistema informático o el diseño de una campaña política o social, le será necesario e incluso rentable tener información debidamente procesada para planear, decidir, ejecutar y controlar de manera exitosa.

A lo largo de sus estudios le han enseñado muchas cosas como puede ser el manejo de los registros contables y la información financiera de su empresa u organismo. También le han instruido cómo hacer algunos estudios de mercado, algo de microeconomía aplicada a la empresa y se la ha dicho cómo son el proceso administrativo y el de producción. Sin embargo todo esto no se manda solo y recae en algo muy importante (sin demeritar otras aplicaciones de la Estadística) que nunca debe olvidar:

Usted como alumno debe tomar una decisión y ¿cómo podrá hacerlo usted si no tiene información que le oriente o, preferentemente, le dé la certeza de que lo que hace está bien?

En este punto es donde la Estadística cobra importancia. Para dar una idea de esta relevancia, piense usted en la empresa Apple Inc. Cuando el señor Jobs retomó las riendas Macintosh (así se llamaba antes), esta era una empresa que vendía computadoras y tenía una tecnología más avanzada que las propias PC's que trabajaban con Windows. Sin embargo, era una empresa que vendía poco en relación a sus competidores y tenía pérdidas financieras, debido a que los directivos anteriores veían a su empresa como proveedora de equipos de cómputo avanzado para



arquitectos, ingenieros y diseñadores. Es decir, se veían como una empresa de nicho y apostaban a que la calidad prevalecería sobre el precio¹.

Sin embargo, Steve Jobs apostó a otro segmento de mercado que poco había sido explotado o de interés para ellos: El del ciudadano promedio (estudiante, ama de casa, joven, anciano, fotógrafo etc.). La facilidad de uso de la PC en comparación a una Macintosh era algo que necesitaban sortear. La primera propuesta que Jobs hizo fue cambiar el microprocesador propio de Mac por los producidos por la empresa Intel. Esto de tal forma que Apple Inc. redujera costos de producción y lograra que los nuevos usuarios de Mac pudieran utilizar Windows y sus aplicaciones que solo funcionaban en dicho sistema operativo, ya que detectaron que pocos programas o aplicaciones de la vida cotidiana corrían o se diseñaban para Mac.

Aquí la pregunta de interés será plantear ¿Cómo llegó Jobs a esa decisión y cómo convenció a su junta directiva para tomar ese paso de olvidar el microprocesador propio de Apple para cambiarlo por uno de Intel? Las labores de negociación que llevó a cabo no son de interés y dudamos mucho que a Apple Inc. desee publicar las minutas de su reunión, ya que es información clasificada. Lo que sí se puede suponer es qué información presentó a su junta. La respuesta es muy simple: Les presentó un estudio de mercado donde preponderan los resultados logrados con la Estadística.

De entrada, su director de Marketing o Mercadotecnia tuvo que realizar un muestreo de racimo (veremos en breve qué es ese concepto) en el que realizó encuestas y obtuvo cifras de diferentes usuarios de computadoras, comenzando por profesionales gráficos (arquitectos, músicos, diseñadores, ingenieros, etc.) y terminó con grupos como estudiantes y amas de casa en donde demostró que la nueva Mac con Windows tenía las mismas prestaciones de accesibilidad y facilidad de manejo que una PC pero con mucha mayor capacidad de cómputo. Esto aunado a las prestaciones de procesamiento, calidad y rapidez de inicio de la computadora (algo que nos molesta a los usuarios de PC). Lo que Jobs pudo hacer, con la ayuda de su staff de marketing, producción, diseño informático y finanzas, fue simplemente hacer un muestreo, hacer una inferencia (normalmente distribuida o no²) respecto a su muestra y demostrar una hipótesis: **“Dado que una Mac cuesta lo mismo que una PC de alto rendimiento, el usuario promedio prefiere una Mac a una PC porque su desempeño y su calidad es superior.”** Para comprobar como válido ese enunciado, Jobs y su gente tuvo que hacer un proceso de análisis estadístico que aprenderá usted a realizar en esta clase. Una vez que demostró eso, la junta directiva de Apple Inc. decidió invertir miles de millones de dólares en las nuevas MacBook pro y air, negocio que les representó una entrada de dinero tan grande al grado de que hoy en día Apple Inc. es una compañía más cara que Exxon, la gran compañía petrolera de Estados Unidos, sin tener todos los activos y patentes de esta última.

¹ No siempre la microeconomía clásica y pura que nos enseñan en la carrera es completa. Aquí entra la estadística para mejorarla.

² En breve analizaremos un poco de ello.



Si este ejemplo no le fue suficiente, piense ahora en un inversionista que maneja su portafolio de inversiones en la Bolsa Mexicana de Valores y en la Bolsa de Nueva York. Este individuo tiene el siguiente objetivo en su dinero: lograr cada semana un rendimiento positivo promedio de 0.5% con una variabilidad de $\pm 1\%$. Es decir que tiene como objetivo perder máximo 0.5% y que la tasa de rendimiento que logre se encuentre en el intervalo dado por $[-0.5\%, +1.5\%]$, siendo su objetivo siempre ganar en promedio 0.5% a la semana (con semanas buenas y malas). ¿Qué debe de hacer? Lo primero que debe realizar el individuo es obtener información histórica de los precios de las acciones en las que quiere invertir y luego debe calcular los rendimientos semanales que tuvo cada una. Después de ellos, debe generar una muestra en cada caso, hacer un portafolio a su gusto (eso no lo veremos en esta materia) y, ya que tiene el diseño del mismo, debe responder, con Estadística inferencial que veremos aquí, si efectivamente ese portafolio de muchas acciones y bonos le paga un rendimiento del 0.5% promedio a la semana, de tal forma que tenga fluctuaciones de $\pm 1\%$. Lo que haría ese inversionista es algo que usted va a aprender en Estadística II.

Es entonces, después de estos dos ejemplos, que usted puede replantearse si realmente le servirá o no esta materia en su vida.

Como puede apreciar, la Estadística es la materia donde un administrador, un contador moderno, un informático, un economista, un político culto y preparado o cualquier profesionista que dirija una empresa, un grupo de personas o agrupación social debe conocer para desenvolverse con fluidez, de lo contrario tomará decisiones poco informadas y cometer errores que cuesten mucho.

1.2 Repaso de conceptos y definiciones de Estadística I.

En el presente sub apartado se recordarán algunos conceptos de la materia de Estadística I que serán punto de partida.

De entrada hay que recordar que la Estadística es una rama de la Matemática consistente en **“El conjunto de técnicas de recolección, presentación y correcto análisis de información numérica relacionada con facilitar la toma decisiones frente a situaciones de riesgo (falta de certeza)”**. Aquí usted podrá identificar un elemento y circunstancia de la vida real: Los individuos, en lo cotidiano, estamos sujetos a tomar decisiones con falta de certeza. Y aquí es donde hacemos un primer paréntesis.

Estadística: El conjunto de técnicas de recolección, presentación y correcto análisis de información numérica relacionada con facilitar la toma decisiones frente a situaciones de riesgo (falta de certeza).

En la vida en general, según lo sugiere la teoría matemática de la decisión, se conciben cuatro escenarios en los que tomamos decisiones los individuos:



1. Escenario de certeza: En este escenario el individuo sabe con seguridad las consecuencias de la decisión que tome.
2. Escenario de riesgo: En este escenario el individuo carece de certeza alguna y puede cuantificar o determinar los diferentes resultados futuros de su decisión.
3. Escenario de incertidumbre: Aquí el individuo carece también de certeza pero sabe que la Estadística no le será de utilidad, por tanto, no puede cuantificar los diferentes resultados futuros de su decisión.
4. Escenario de conflicto: En el mismo se pueden o no conocer los resultados futuros. Sin embargo, estos no dependen de cuestiones estadísticas sino de los gustos e intenciones de otros individuos que no se pueden saber a ciencia cierta.

En relación al escenario de certeza se puede decir que, en nuestro universo y vida en general, este escenario es prácticamente teórico (por no decir que inexistente). Sin embargo, en algunos casos, podría decirse que casi existe la “certeza” de lo que suceda cuando decidamos. Por ejemplo, podemos saber que en Morelia es de noche a las 11:00 PM. Aunque esto puede ser diferente en ciudades como Oslo donde en el verano se tiene un sol de las 11:00 de la mañana cuando son las 11:00 PM, en el contexto de un moreliano, la afirmación anterior es casi cierta.

Escenario de certeza: Escenario en el que el individuo sabe con seguridad las consecuencias de la decisión que tome.

Escenario de riesgo: Escenario en el que el individuo carece de certeza alguna y puede cuantificar o determinar los diferentes resultados futuros de su decisión con la Estadística.

Escenario de incertidumbre: Escenario en el que el individuo sabe que la Estadística no le será de utilidad ya que no puede cuantificar los diferentes resultados futuros de su decisión.

Escenario de conflicto: Escenario en el que el individuo puede o no conocer los resultados futuros. Sin embargo, estos no dependen de cuestiones estadísticas; sino de los gustos e intenciones de otros individuos que no se pueden saber a ciencia cierta.

En lo que se refiere al **escenario de riesgo**, este será en el que usted se contextualizará ya que utilizará la Estadística para aproximar los posibles resultados que tendrá usted al decidir por un proyecto, empresa, inversión o actividad.

Un ejemplo del escenario de incertidumbre podría ser que usted intente determinar qué tan factible es que le caiga un meteorito a su nuevo restaurante. Eso, en algunos casos que no interesan a esta materia, podría determinarse. Baste simplificar un poco la vida y decir que ese potencial fenómeno no se puede modelar con la Estadística y por tanto es incierto saber qué eventos astronómicos podrían influir en la seguridad de nuestra empresa.



Por último, como ejemplo de un caso de escenario en conflicto, pueden tenerse las negociaciones en la cámara de diputados. Por ejemplo, piense en un tema difícil como es una reforma fiscal. Algunos partidos propondrán algo y los otros quizá no cedan en su postura de aceptar o no dicha propuesta, por lo que los posibles resultados se podrían conocer, podrían modelarse con la Estadística pero, a decir verdad, el resultado no depende del modelado de un **evento aleatorio** sino de la voluntad de las partes en la negociación.

La razón de hacer este breve paréntesis radica en resaltar que usted realizará, por lo general, decisiones en un entorno donde podría conocer, o al menos tener una idea genérica, de los posibles resultados que tendrá como consecuencia las decisiones que tome o conocer realmente si la información que tiene es apropiada.

Por ejemplo, si usted produce botellas de agua de 1 l, debe revisar que su línea de producción cumpla con el contenido para evitar un problema con PROFECO. Para ello usted obtiene una muestra de botellas y mide la cantidad que contiene cada una. Como el proceso de envasado tiene múltiples factores que pueden hacer que no se tenga exactamente 1 l en todas las botellas, usted está en un escenario de riesgo donde diferentes niveles de llenado pueden fluctuar en .7, .8, .99, .5 litros. Este es un escenario de riesgo en el que usted tiene múltiples resultados que potencialmente conoce.

Por tanto y para fines de la materia, todas las decisiones que deba usted efectuar en su negocio, las realizará en un contexto de riesgo.

1.2.1 La probabilidad ¿Qué es y cómo se cuantifica?

Ya que se estableció el tipo de escenario en donde usted tomará decisiones, es de necesidad recordar un elemento de importancia: La probabilidad. Esta se define como **“Una medida numérica que cuantifica numéricamente la posibilidad de que un resultado o evento se presente”**.

En un español más simple, la probabilidad es un número que indica a quien lo utiliza qué posibilidad se tiene de que algo suceda. El ejemplo más común se escucha en los noticieros: *“El día de hoy se tendrán nublados con una probabilidad de precipitación pluvial de 80%”*. Esto indica que el día de hoy es muy posible que llueva.

Probabilidad: “Una medida numérica que cuantifica numéricamente la posibilidad de que un resultado o evento se presente”.



En virtud de la definición de probabilidad dada, es de interés observar la definición de lo que se conoce como “evento”. Como se estableció previamente, la probabilidad es el número que determina qué tan posible es que se de ese evento. Por tanto el mismo se define como:

Evento: “El futuro acontecimiento que resultará de cualquier acción tomada en el presente”.

El evento no es más que el resultado futuro que se logra con la decisión que tomemos como administradores, contadores o informáticos. Por ejemplo el lanzar una moneda (jugar un “volado”) tiene dos eventos: cara o sol.

Sin embargo, dado que la moneda lanzada y, en específico, el resultado que se logre, es algo sujeto al azar, se tiene que este evento es un “**evento aleatorio**” y el hecho de lanzar una moneda se conoce como “**Experimento aleatorio**”

Todos los eventos que se estudien en esta materia se considerarán “eventos aleatorios” y las decisiones que deba tomar, desde un punto de vista Estadístico, se considerarán como un “**Experimento aleatorio**”. En este experimento aleatorio, todos los posibles eventos aleatorios que puedan existir en el mismo forman un conjunto llamado “**Espacio muestral**”. En el caso de la moneda que estudiamos previamente, el experimento aleatorio es jugar al “volado”, el espacio muestral es un conjunto de solo dos eventos aleatorios: tener cara o tener sol.

Evento aleatorio: “Son los resultados o acontecimientos cuyo valor, dada una decisión previa, están sujetos al azar”.

Experimento aleatorio: “Es una actividad sujeta a las leyes de la probabilidad en la que se puede obtener uno solo de los eventos aleatorios que conforman el espacio muestral”.

Espacio muestral: “Es el conjunto de posibles eventos aleatorios (resultados) que pueden tenerse en un experimento aleatorio”.

Algo que es importante saber ahora que se definió lo que es un evento aleatorio, un experimento aleatorio y la probabilidad, es que existen dos tipos de probabilidades:

Probabilidad subjetiva: Es una medida numérica que expresa un grado personal o teórico de que un evento suceda.

Por ejemplo usted puede creer que una señora embarazada tiene 50% de probabilidades de dar a luz una niña o 50% de lograr un niño.

Probabilidad objetiva: Es una medida numérica que cuantifica la posibilidad de que un evento aleatorio suceda en relación al total de eventos de un espacio muestral.



En un español más plano, la cantidad de veces que usted puede lograr un “tres” al lanzar un dado es uno. Es decir usted solo puede lograr un tres si lanza un dado ya que el dado solo tiene impreso dicho número una vez. Sin embargo el dado tiene seis números. Es decir, en nuestra terminología estadística, el experimento aleatorio de lanzar un dado tiene un espacio muestral consistente en seis eventos aleatorios:

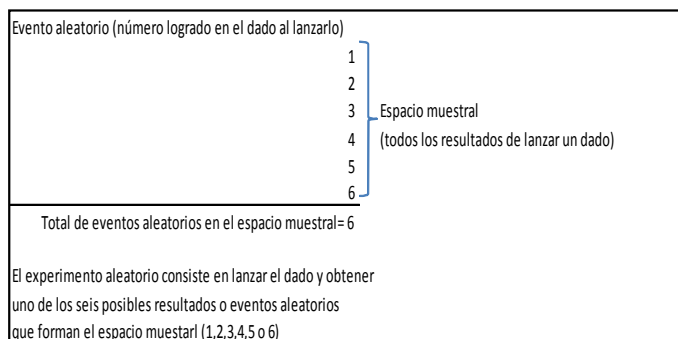


Figura 1 Ilustración del espacio muestral del evento aleatorio "lanzar un dado" donde se exponen los posibles "eventos aleatorios".

En base al experimento aleatorio (denotado con X) anterior, la probabilidad de tener un 3 en el experimento aleatorio llamado “lanzamiento de dado” es:

Fórmula 1:

$$probabilidad = p(x) = \frac{\#eventos\ aleatorios}{Tamaño\ de\ esp.\ muestral}$$

$$p(x) = \frac{\#eventos\ aleatorios}{Tamaño\ de\ esp.\ muestral} = \frac{1}{6} = 16.666\%$$

Esto significa que la medida numérica de la posibilidad de obtener un 3 en el lanzamiento del dado es de 16.6666% que se logra de dividir el número de eventos aleatorios de interés (#) entre el tamaño del espacio muestral que es de 6.

¿Cuál es la diferencia entonces entre la probabilidad objetiva y la subjetiva?

La diferencia radica en que la subjetiva se basa en cuestiones teóricas, como es la creencia de la probabilidad de que un recién nacido nazca mujer u hombre. La objetiva se basa en definir, a la luz de posibles resultados previamente definidos y con datos observados, la posibilidad de que un evento aleatorio se presente dado el número de veces que puede suceder en relación al número de veces que este y el número del resto de resultados puede acontecer (como en la fórmula 1).



Ahora, es importante señalar que en esta materia siempre se trabajará, salvo que se diga lo contrario, con probabilidades objetivas.

1.3 Medidas de tendencia central y medidas de dispersión.

1.3.1 La media, la mediana y la moda

Hasta ahora se ha dado una definición introductoria de la probabilidad de que suceda un evento aleatorio o resultado esperado y se han dado dos ejemplos sencillos. Ahora recuerde usted el evento aleatorio de la línea de producción de botellas de agua previamente mencionado. Usted no sabe ni sabrá con seguridad cuáles serán los posibles resultados del espacio muestral del nivel de llenado de sus botellas. Este puede ser tan grande como:

espacio muestral= {0 l., 0.00000001 l., 0.0000023 l., 0.55647726 l., 1 l.,...}

Es entonces que usted, para organizar la vasta información que tiene y tener un par de números en la cabeza que le resuman todos los datos, recurre a otro tipo de medidas que vio en Estadística I y que se conocen como medidas descriptivas. Las medidas descriptivas a las que se refiere esto son las medidas de tendencia central y las de dispersión y, para esto, se requiere una observación de diferentes elementos de una población o conjunto de características buscadas en un grupo de objetos que la presentan.

Por ejemplo, usted simplemente se dedica a medir el nivel llenado de todas las botellas de agua producidas a lo largo de la vida de su fábrica y el conjunto de datos que logre de todas sus botellas se llama **población**.

Población: Conjunto de todas las observaciones posibles sobre una característica de interés observada.

Aquí es importante señalar que la observación, como su nombre lo dice, es el resultado o valor de un evento aleatorio que se tiene una vez que se realiza el experimento aleatorio.

Es decir, para fines del ejemplo de las botellas, el experimento aleatorio es el nivel de llenado de una botella con múltiples valores o eventos aleatorios en el espacio muestral (infinita cantidad de los mismos)³, los cuales se observaron una vez que se realizó el evento de llenarla.

En base a toda la información que recabe usted con esta medición, registro y conteo podrá calcular probabilidades pero ¿Qué pasa si usted produjo en los, digamos, cinco años de vida de su fábrica, 1 millón de botellas?, ¿Cómo va a procesar toda esa información para calcular la

³ En breve se verá la diferencia entre una magnitud continua y una discreta.



probabilidad de que este embotellando entre 0.8l y 1l de agua en cada una?, es decir, ¿cómo sabe usted que se están cumpliendo con los estándares de calidad en el llenado de tal forma que no tenga problemas con sus clientes, y PROFECO?

Para responder esto, se emplean las medidas de tendencia central y de dispersión.

Las medidas de tendencia central son, como su nombre lo dice, aquellas que identifican el comportamiento más común en la característica buscada y las tres más empleadas son la media (o promedio), la mediana y la moda.

La media o promedio (μ): Es la medida de tendencia central que se obtiene de sumar los valores de todas las observaciones (x_i) de la población y dividir dicha suma entre el número de observaciones (n).

Fórmula 2:

$$\mu = \frac{\sum x_i}{n}$$

Por ejemplo piense en la población de 10 lanzamientos de dados:

$$X = \{1, 4, 2, 3, 4, 4, 5, 6\}$$

La media (denotada por μ) de la misma será:

$$\mu = \frac{1+4+2+3+4+4+5+6}{8} = 3.625$$

La mediana: Es el valor de la observación que, una vez ordenada la población de la menor observación a la mayor, que se encuentra exactamente a la mitad de la población.

La condición necesaria para que exista la mediana es que, al establecerse la misma, se cuente el mismo número de observaciones arriba y debajo de la misma. Por ejemplo piense en la siguiente población de números:

$$X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

Aquí claramente la mediana es el 5 ya que es el número que se encuentra a la mitad y el que tiene el mismo número de observaciones a la izquierda y a la derecha.

Moda: Es el valor de evento muestral que presenta el mayor número de observaciones en la población estudiada.



Siguiendo el ejemplo de los 10 lanzamientos de dados,

$$X = \{1, 4, 2, 3, 4, 4, 5, 6\}$$

se observa que la moda sería el 4 ya que es el número que más veces aparece en la población .

Las medidas de dispersión, como su nombre lo indican, determinan el nivel de separación (dispersión) que todas las observaciones de una población tienen respecto a una de sus medidas de tendencia central.

Las dos medidas de dispersión comúnmente empleadas son la varianza (o su raíz cuadrada la desviación estándar) y la desviación media absoluta.

1.3.2 La varianza y la desviación estándar ¿qué significan? y ¿Por qué la calculamos la varianza elevando al cuadrado las diferencias respecto a la media?

La varianza es, quizá, la medida de dispersión más empleada en la Estadística y en todo tipo de aplicaciones. La misma simplemente se dedica a medir el tamaño promedio de separación que las diferentes observaciones de la población tienen respecto a u media (μ).

Varianza: La separación promedio que tienen las observaciones de una población respecto a su media.

Si se piensa de nuevo en nuestro ejemplo del nivel de llenado de las botellas de agua, se puede apreciar que nuestros millones de botellas tienen diferentes separaciones respecto al llenado promedio. Para ilustrar esto suponga usted que la población de observaciones de botellas lleva a un nivel de llenado promedio de 0.89l ($\mu = 0.89l$) y que tiene solo una población de tres botellas producidas con 0.78 l, 0.92 l y 0.84 l. La separación respectiva respecto a la media en cada caso es de -0.11 l, 0.11 l y 0.02 l. La varianza simplemente determina la separación promedio. Es decir, el promedio de -0.11 l, 0.11 l y 0.02 l.

Pero **¡Alto!**, vea bien usted los valores. Si recordamos la fórmula del promedio tenemos que sumar los tres valores anteriores y dividir entre tres, lo que nos llevaría a una varianza muy pequeña y mal calculada:

$$\frac{-0.11 \text{ l} + 0.11 \text{ l} + 0.02 \text{ l}}{3} = \frac{0.02 \text{ l}}{3} = 0.006 \text{ l}$$



Esto le estaría diciendo erróneamente que la varianza o desviación promedio respecto a la media no existe. Cosa que claramente no es cierta. Para dar una mayor idea, recuerde ahora el ejemplo de los dados y la población de posibles resultados:

| Valor del dado | Promedio o media | Diferencia respecto a media |
|--|------------------|-----------------------------|
| 1 | 3.5 | -2.5 |
| 2 | | -1.5 |
| 3 | | -0.5 |
| 4 | | 0.5 |
| 5 | | 1.5 |
| 6 | | 2.5 |
| Suma de las diferencias | | 0 |
| Número de observaciones | | 6 |
| Promedio de diferencias respecto a media(varianza) | | =0/6=0 (¡?!) |

Tabla 1 Cálculo erróneo de la varianza en el evento aleatorio "lanzamiento de dado".

Cuando se calculan las diferencias de cada resultado (u observación) respecto a la media o promedio de resultados (que es de 3.5), se puede observar que realmente existen diferentes separaciones del valor de cada observación respecto a la media (última columna de la tabla 1). Sin embargo, como medida de dispersión, a usted no le interesa saber todas las diferencias sino su valor promedio. Es decir, el grado de separación medio.

Retomando la fórmula 2 puesta en página previas, usted tendrá que sumar las diferencias respecto a la media (columna de la derecha). Sin embargo, aquí se puede usted llevar una mala sorpresa al ver que la suma da cero y, al calcular la diferencia media, el cálculo nos dice que no existe separación alguna. ¿Será cierto eso? Si usted revisa la tabla estudiada y los valores de las diferencias podrá observar que no. Entonces ¿Qué se sugiere hacer? Para calcular la varianza, simplemente se eleva al cuadrado las diferencias respecto a la media expuestas en la columna derecha de la tabla anterior, se suman y se dividen entre el número de observaciones. Esto sería:

| Valor del dado | Promedio o media | Diferencia respecto a media | Diferencia elevada al cuadrado |
|--|------------------|-----------------------------|--------------------------------|
| 1 | 3.5 | -2.5 | 6.25 |
| 2 | | -1.5 | 2.25 |
| 3 | | -0.5 | 0.25 |
| 4 | | 0.5 | 0.25 |
| 5 | | 1.5 | 2.25 |
| 6 | | 2.5 | 6.25 |
| Suma de las diferencias | | 0 | 17.5 |
| Número de observaciones | | 6 | 6 |
| Promedio de diferencias respecto a media(varianza) = 0/6=0 (¡?!) | | | cálculo malo: |
| | | | cálculo bueno (varianza): |
| | | | 2.91666667 |

Tabla 2 Cálculo correcto de la varianza en el evento aleatorio "lanzamiento de dado".

Es entonces que, con el ejercicio anterior, usted puede ahora saber por qué debemos calcular la varianza (denotada por s^2) elevando al cuadrado las diferencias de las observaciones respecto a la media de la población:

Fórmula 3:



$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

La fórmula anterior, para el ejemplo de los dados expuesto en la columna sombreada de la tabla anterior sería:

$$\begin{aligned} \sigma^2 &= \frac{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2}{6} \\ &\quad + \frac{(4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2}{6} \\ &= \frac{6.25 + 2.25 + 0.25}{6} \\ &= \frac{+0.25 + 2.25 + 6.25}{6} = \frac{17.5}{6} = 2.9166... \end{aligned}$$

Ahora usted podrá pensar que este número está un poco raro ya que a usted no le hace sentido elevar al cuadrado las diferencias. Sin embargo, en la Matemática se pueden hacer cambios y trucos discrecionales sin que la realidad cambie y esto que se realiza (elevar al cuadrado las diferencias de las observaciones respecto a la media) se trata de un mero “ajustito” matemático para que salgan las cuentas.

Para ahorrarle la crisis existencial, una vez que se calcula la varianza, lo que muchas veces se hace (y no será la excepción aquí) es simplemente calcular la raíz cuadrada de la varianza para obtener la desviación estándar (denotada simplemente por σ). Esto, para el ejemplo de la varianza del dado, sería:

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.916666} = 1.7078...$$

Fórmula 4:

$$\sigma = \sqrt{\sigma^2}$$

Entonces ahora podrá hacerle sentido hacer la siguiente afirmación a sus clientes o a sí mismo. **“En el lanzamiento de un dado, el valor promedio de los seis posibles resultados (media) es de 3.5 el cual tiene variaciones potenciales de ± 1.70 . Es decir, en promedio esperaríamos tener un 3.5 con posibilidad de sacar al menos un 1.8 o un 5.2”**

Tranquilícese, bien es cierto que estos números son imposibles. El ejemplo del dado se presenta para ilustrarle a usted la forma de cálculo de media y desviación estándar. Para dejarle con algo más palpable, piense ahora en un corredor de bolsa. Este individuo quiere especular con el valor de una acción de telecomunicaciones y desea saber qué precio tendrá la misma el día de mañana en tres escenarios: uno factible, uno optimista y uno pesimista. Lo que este individuo realiza entonces es obtener el una muestra del precio histórico de esa acción del último mes (la forma de



determinar el tamaño de muestra se verá en breve) o últimos 30 días y calcular el promedio con la fórmula 1 y la desviación estándar con las fórmulas 3 y luego la 4.

Lo que, posteriormente hará, es tomar la media como escenario factible y a esa media sumarle la desviación estándar para obtener el escenario positivo o restársela para lograr el negativo. Suponga usted que el corredor obtiene las siguientes medidas de tendencia central y dispersión:

$$\mu = 23$$

$$\sigma = 2.5$$

Dado que el escenario factible es la media, El inversionista podría terminar con una inversión que valga \$23 el día de mañana. Para calcular el escenario optimista deberá hacer la siguiente suma:

$$\mu + \sigma = 23 + 2.5 = 25.5$$

Para el escenario pesimista deberá hacer lo siguiente:

$$\mu + \sigma = 23 - 2.5 = 20.5$$

Esto llevará al corredor a tener la siguiente tabla de escenarios y decidir si quiere comprar la acción:

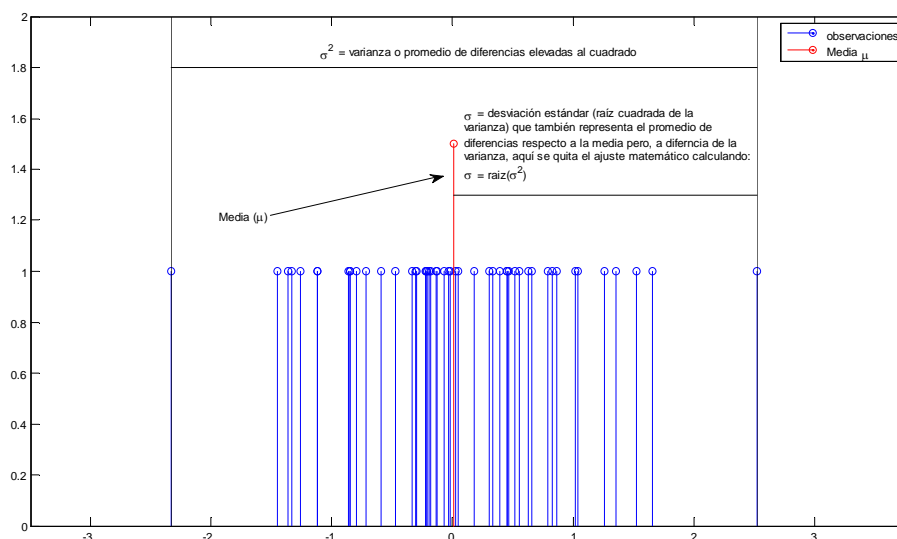
| Escenario | cálculo | Precio esperado en la acción |
|-----------|----------------------------------|------------------------------|
| Optimista | $\mu + \sigma = 23 + 2.5 = 25.5$ | 25.5 |
| Factible | $\mu = 23$ | 23 |
| Pesimista | $\mu + \sigma = 23 - 2.5 = 20.5$ | 20.5 |

Tabla 3 Escenarios esperados para del precio de una acción en base a un análisis con medidas de tendencia central y de dispersión.

Lo que el profesor

espera con este ejemplo es que se logre entender que la media (μ) es el resultado promedio que podría esperar en un evento aleatorio (como las botellas, el nivel de ventas de una Mac o el precio esperado en la acción por parte de un inversionista), dada la población de observaciones y que la varianza (σ^2) o si prefiere la desviación estándar (σ)⁴, que es más fácil de interpretar, miden simplemente el grado de separación promedio que todas las observaciones de la población tienen respecto a la media (μ). En términos gráficos esto sería:

⁴ Que es la raíz cuadrada de la varianza



Gráfica 1 Ilustración de la media, la varianza y la desviación estándar de una "población" compuesta de "observaciones" resultantes de la realización de un "evento aleatorio" en un "experimento aleatorio".

Ya que se repasó lo que es la media, la varianza y la desviación estándar, es de necesidad observar lo siguiente: En la mayoría de los casos se utiliza la media como medida tendencia central y la desviación estándar para hacer análisis estadístico y esto, salvo en los casos que se exprese lo contrario, será aplicable en la materia.

1.3.3 Reglas de dedo para calcular la media y la desviación estándar:



Ahora se le da la receta "de cocina" para calcular estas dos importantes medidas o estadísticas:

Media:

1. Tome todas las observaciones de su población (o muestra como se verá en breve)
2. Suma los valores numéricos de las observaciones.
3. Cuente el número de observaciones que tiene.
4. Divida la suma de valores numéricos de las observaciones entre el número de las mismas.

Desviación estándar:

Varianza:

1. Recuerde que debe calcularse la varianza para obtener este valor. Por tanto, debe calcularse primero la media.
2. A cada valor numérico de cada observación se le resta el valor de la media (vea columna "diferencias respecto a la media" en la tabla 2). Es decir, se calcula la diferencia entre cada valor numérico de cada observación respecto a la media.



3. Las diferencias calculadas anteriormente se elevan al cuadrado (vea columna “diferencia elevada al cuadrado” en la tabla 2).
4. Se suman las diferencias calculadas.
5. Se divide esta suma entre el número de observaciones.

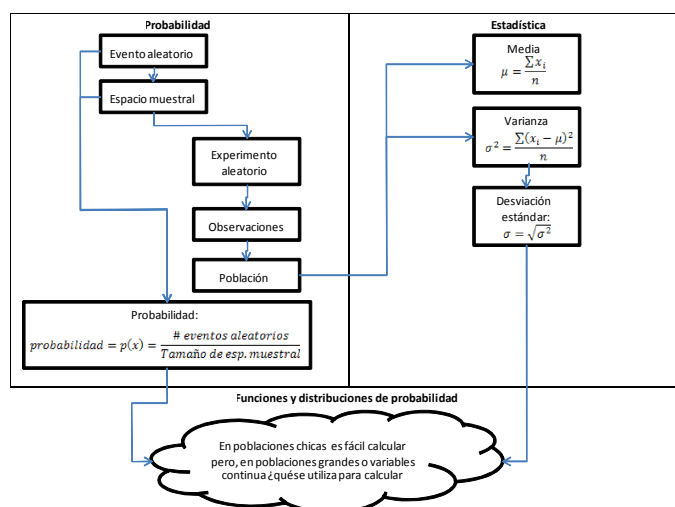
Ahora sí, la Desviación estándar:

6. En los pasos A a F se calculó la varianza. Si usted quiere utilizarla, está bien pero es más recomendable utilizar la desviación estándar que se calcula simplemente sacando la raíz cuadrada de la varianza lograda en el paso F.

1.4 Cálculo de probabilidades: los histogramas, las funciones y distribuciones de probabilidad.

1.4.1 Mapa mental de lo hasta ahora visto

Hasta ahora hemos visto qué es un evento aleatorio, cómo se calcula, de manera básica e intuitiva la probabilidad y, por otro lado, separado pero relacionado, hemos visitado las dos medidas o estadísticas que más utilizaremos: la media y la desviación estándar. En términos del tema que nos interesa esto es:



1.4.2 Eventos aleatorios (variables aleatorias) discretos y continuos

Cuando la población de observaciones que se tiene es la de un experimento aleatorio sencillo como es el lanzamiento de una o dos monedas, el lanzamiento de uno o dos dados, los resultados de un juego de cartas o las calificaciones de un grupo de clase, es muy simple aplicar la fórmula 1:



$$probabilidad = p(x) = \frac{\#eventos\ aleatorios}{Tamaño\ de\ esp.muestral}$$

Sin embargo, en poblaciones grandes o eventos continuos (ahorita conocerá el término), la forma de hacer esto es diferente. Antes de hablar de ello, es necesario saber qué es un evento aleatorio discreto o un evento discreto y qué es un evento continuo.

Evento aleatorio discreto: Es aquel cuyo conjunto de posibles resultados o acontecimientos tienen una cantidad que se puede contar aunque sea esta muy grande.

Evento aleatorio continuo: Es aquel cuyo conjunto de posibles resultados o acontecimientos tienen una cantidad que **no** se puede contar ya que esta es un número infinito.

Un ejemplo de evento aleatorio discreto puede ser el dado revisado, el número de mujeres profesionistas de la contabilidad en la ciudad de Morelia o el número de butacas rotas en la universidad. Estos tres ejemplos de poblaciones tienen eventos (ser mujer profesionista de la contabilidad, ser butaca rota) que se pueden contar. Es decir son **finitos**.

Un ejemplo de un evento aleatorio continuo serían los posibles valores que puede tomar la temperatura en Morelia en cierto día. Puede tener valores como 10.51°, 10.53243°, etc. Otro ejemplo serían los rendimientos diarios o por hora que puede tener el índice de la bolsa de valores. Dado que los posibles resultados en estos eventos pueden ser prácticamente infinitos, se dice que estos eventos son continuos.

Hasta ahora hemos hablado de “eventos aleatorios”. Sin embargo, en términos matemáticos, el nombre “evento” no dice mucho. Recuerde usted en sus clases de matemáticas I y II que en la Matemática (y por ende en la Estadística al ser una parte de la Matemática) utiliza ecuaciones (funciones), que tienen variables, para explicar o resolver problemas de nuestra vida cotidiana. Es entonces que aquí cambiaremos, para hablar con mayor propiedad matemática, el nombre de **evento aleatorio** por el de variable aleatoria.

Esto es, apréndase esta regla de dedo:

Evento aleatorio = Variable aleatoria (en Matemáticas)

1.4.3 Cálculo de probabilidades en variables aleatorias discretas: El histograma.

Recordemos la fórmula 1:

$$probabilidad = p(x) = \frac{\#eventos\ aleatorios}{Tamaño\ de\ esp.muestral}$$



Existen ocasiones en que las variables aleatorias arrojan observaciones que se repiten. Por ejemplo, piense en el número de coches de cinco posibles colores: blanco, negro, rojo, azul y verde. Si cuenta todos los casos posibles en el estacionamiento de la universidad podrá tener resultados como este:

| Color | Número de automóviles |
|--------|-----------------------|
| Blanco | 10 |
| Negro | 30 |
| Rojo | 40 |
| Azul | 5 |
| Verde | 6 |
| Total | 91 |

Tabla 4 Tabla de frecuencias de los coche en el estacionamiento de una ciudad

Ahora realice este ejercicio: Dada una población total de 91 automóviles, usted se pone afuera de su aula o salón de clase y hace un juego con sus compañer@s consistente en adivinar de qué color será el próximo coche que entre al estacionamiento. ¿Qué color elegiría? Muy simple, dadas sus capacidades como estadístico que viene a aprender, simplemente calcula la probabilidad de cada uno de los colores aplicando la fórmula 1 en los valores de la tabla cuatro. Por ejemplo, existen 10 coches blancos en la universidad y, dividiendo esta cantidad entre el total de coches que es 91, se llega a una probabilidad de suceso de 10.99%. Esto es:

$$\begin{aligned} \text{prob.cocheblanco} &= p(x_i) = \# \text{cocheblanco} / \text{totalcoches} \\ &= 10 / 91 = 10.99\% \end{aligned}$$

Para el juego de elegir un color de coche, se llega a la siguiente tabla de probabilidades:

| Color | Número de automóviles | Probabilidad de suceso |
|--------|-----------------------|------------------------|
| Blanco | 10 | 10.99% |
| Negro | 30 | 32.97% |
| Rojo | 40 | 43.96% |
| Azul | 5 | 5.49% |
| Verde | 6 | 6.59% |
| Total | 91 | |

Tabla 5 Probabilidad de que el siguiente coche que entre en el estacionamiento sea de determinado color de los dados en tabla 3.

Si desea ganarle a sus amig@s simplemente dirá que el siguiente coche será o rojo o negro (Elija por uno de los dos o ambos) ya que son los dos casos que mayor probabilidad de suceso o de ocurrencia tienen.



Este ejercicio es sencillo pero ahora recordemos el ejemplo de los niveles de llenado de sus botellas de agua ¿Se imagina la tabla que tendrá que calcular? Sería inmensa, al menos 1,000 celdas. Para simplificar la tarea, los estadísticos idearon una técnica de organización llamada **histograma de frecuencias**.

Para dar una idea y recordar lo que es un histograma de frecuencias, piense usted las cajas de verdura que pueden encontrarse en los mercados. En algunos comercios del mismo hay quienes separan las cajas de aguacate en función del peso de los mismos.

Para ilustrar un ejemplo, piense en un comerciante teórico que separa los aguacates de la siguiente manera: Pone los aguacates que pesan de 50 gramos o menos en una caja, los de 51 a 101 en otra, los de 102 a 152 gramos en otra y así sucesivamente hasta crear 10 grupos o **intervalos** de 50 gramos de **rango** que lleguen hasta los 500 g (vea tabla 6 para observar cómo quedaron los grupos o intervalos). Suponga que usted se encuentra con el comerciante comprando aguacates y este le dice que le venderá en 78 pesos tres aguacates si le permite a dicho comerciante elegir de manera aleatoria en las diez cajas. Ahora suponga que el comerciante tiene un total de 1,000 aguacates en inventario repartido de la manera que se presenta en el cuadro siguiente:

| | Intervalo (rango de 50 g) | Frecuencia |
|----------------|---------------------------|------------|
| Peso en gramos | 50 o menos | 20 |
| | 51-101 | 78 |
| | 102-152 | 112 |
| | 153-203 | 64 |
| | 204-254 | 184 |
| | 255-305 | 186 |
| | 306-356 | 142 |
| | 357-407 | 154 |
| | 408-458 | 46 |
| | 459-500 | 14 |
| | Total de aguacates | 1000 |

Tabla 6 Tabla de frecuencia de la clasificación de los aguacates que se encuentran en cada caja o “intervalo” con un “rango” de 50 gramos de diferencia.

Como se puede apreciar, la tabla anterior es muy parecida a la tabla 3 en donde se presenta la frecuencia del número de automóviles dado un color. Lo único que cambia aquí es que la clasificación no es en función de una cualidad como el color sino en función del peso que puede ser, como se dijo antes, una **variable aleatoria continua**. Por ejemplo, en el primer grupo o intervalo (0 a 50 gramos) puede haber un aguacate de 48.5 g, otro de 30.5, otro de 49.98 y así sucesivamente en los 20 miembros de la caja o intervalo.

Si usted desea saber qué **rango** de peso es más probable que resulte cuando el comerciante seleccione el primero de 10 aguacates, simplemente repite el cálculo de la tabla 4 y divide el



número de aguacates que se encuentran en cada caja o **intervalo** entre el total que tiene el comerciante en inventario (1,000). Por ejemplo, para el grupo o intervalo de aguacates que pesan de 153 a 203 gramos, la probabilidad de que el comerciante le de una fruta de este grupo se daría por:

$$p(153-203 \text{ g}) = \frac{\# \text{ aguacates (153-203 g)}}{\text{Total de aguacates}} = \frac{64}{1,000} = 6.4\%$$

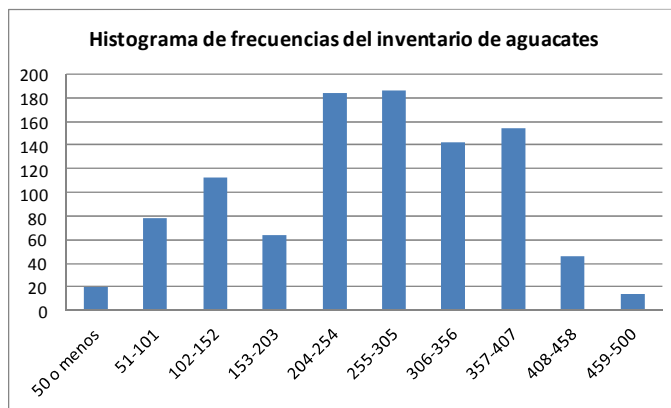
Para todos los grupos de aguacates se llega a la siguiente tabla de frecuencias y de probabilidades:

| | Intervalo (rango de 50 g) | Frecuencia | Probabilidad |
|----------------|---------------------------|------------|--------------|
| Peso en gramos | 50 o menos | 20 | 2.0000% |
| | 51-101 | 78 | 7.8000% |
| | 102-152 | 112 | 11.2000% |
| | 153-203 | 64 | 6.4000% |
| | 204-254 | 184 | 18.4000% |
| | 255-305 | 186 | 18.6000% |
| | 306-356 | 142 | 14.2000% |
| | 357-407 | 154 | 15.4000% |
| | 408-458 | 46 | 4.6000% |
| | 459-500 | 14 | 1.4000% |
| | Total de aguacates | 1000 | |

Tabla 7 Tabla de frecuencias y probabilidades dado el número de frecuencia de aguacates en cada caja o intervalo de peso.

Como puede apreciar, es más probable que el primer aguacate pese entre 204 y 254 gramos o 255 y 305 gramos ya que son las dos cajas o intervalos de peso con mayor número de aguacates y, por ende, mayores probabilidades de ser seleccionados.

La tabla 7 representa lo que se conoce como **una tabla de frecuencias** y la representación gráfica se da en la gráfica 2. Esta gráfica se conoce como **histograma de frecuencias** y se utiliza mucho en Estadística para analizar visualmente datos o en lo que más adelante conoceremos como **Estadística no paramétrica**. En esta es donde no se calculan parámetros como la media o la desviación estándar previamente revisados. Eso se deja para una discusión posterior.



Gráfica 2 histograma de frecuencias del inventario de aguacates dados los 10 intervalos con rango de 50 gramos.

Los histogramas de frecuencias no son más que la representación gráfica de la tabla de frecuencias y nos dan una idea de cuáles son los valores con mayores probabilidades de ser observados. Una definición más formal se puede dar por:

Histograma de frecuencias: Representación gráfica de una distribución de frecuencia de una variable aleatoria continua.

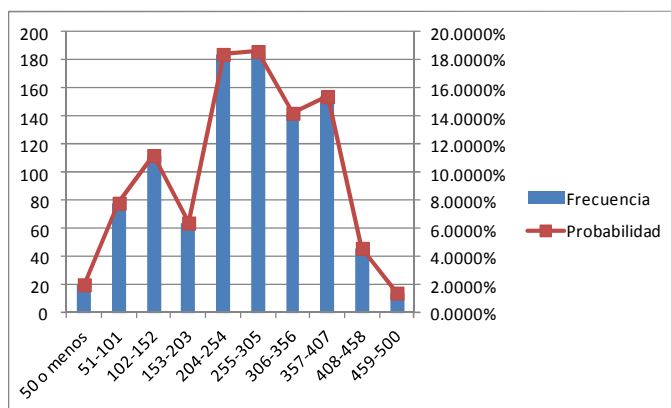
Para cerrar el tema del histograma, se ve que esta herramienta, que parte de la tabla de frecuencias, es de utilidad para organizar muchas observaciones o datos de una población y convertir un problema de variables aleatorias continuas en uno de variables discretas.

1.4.4 Distribuciones de probabilidad.

En la definición de histograma se acaba de identificar un término que será fundamental en la Estadística inferencial: La distribución de frecuencias. Esta no es más que la forma en que se acomodan las diferentes frecuencias de suceso de los eventos aleatorios dado un intervalo dado.

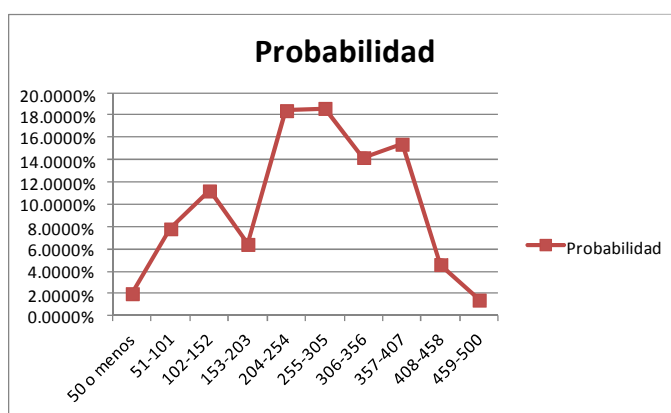
Nótese en la gráfica 2 cómo las diferentes frecuencias se acomodan describiendo un fenómeno de interés para usted como contador, administrador o informático: “El proveedor del comerciante solo produce aguacates de peso alto ya que las mayores frecuencias se encuentran entre los 204 y los 407 gramos. Ya la mayor parte de los aguacates **se distribuye** en estas cajas o intervalos.

¿Qué pasa ahora si en lugar de graficar el histograma se presenta una línea donde se muestra la distribución de probabilidades calculadas en la tabla 6? Observe usted:



Gráfica 3 Histograma de distribución de frecuencias y gráfica de distribución de probabilidades del inventario de aguacates estudiado.

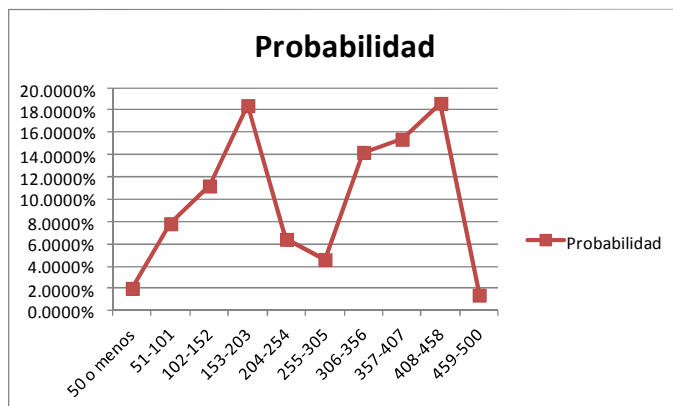
¿Y si quitamos el histograma por que buscamos calcular solo probabilidades? Vea ahora usted:



Gráfica 4 Distribución de probabilidad del inventario de aguacates estudiado.

Es entonces que llegamos a la distribución de probabilidad. Quizá esa línea toda quebrada nos dice cosas muy parciales y es fácil intuir que los pesos más probables de encontrar en los 10 aguacates que nos venda el comerciante sean entre 204 y 407 gramos. Pero ¿Qué pasa ahora si cambiamos un poco las frecuencias o el inventario es un poco diferente? Vea usted la gráfica 5.

¿Le dice a usted algo esta distribución de probabilidad? Ahora tenemos dos posibles valores de peso de aguacate: 153-203 gramos y 405-458. ¿Podría usted fijar un patrón estadístico con esto? La respuesta es que sí se podría pero la estadística que se utilice no será paramétrica. Utilizar la misma es muy sencillo pero eso se verá posteriormente.



Gráfica 5 Distribución de probabilidad del inventario de aguacates estudiado cambiando un poco los valores de frecuencia.

1.4.5 Funciones de densidad de probabilidad

Hasta ahora se ha hablado de una distribución de probabilidad obtenida totalmente de los datos de la población y vemos que se debe seguir la siguiente receta:



1. Obtener todos los datos u observaciones de la población.
2. Organizarlos de menor a mayor.
3. Definir una cantidad de grupos o intervalos que se acomode a su análisis (2,3,10,100, etc.)
4. La diferencia entre el valor máximo y el mínimo divídala entre el número de intervalos que desee calcular y con eso logra el rango:

Fórmula 5:

$$rango = \frac{(V_{\max pob} - V_{\min pob})}{n}$$

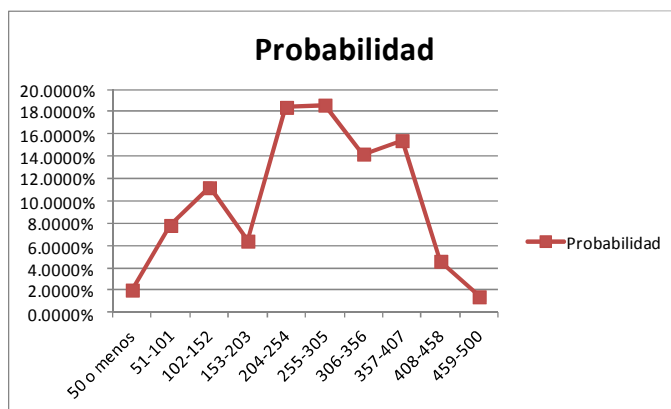
En donde $V_{\max pob}$ representa el valor máximo de la población, $V_{\min pob}$ el valor mínimo y n el número de intervalos o grupos que desea calcular.

5. Clasifique todas las observaciones en cada uno de los n intervalos que creó con la fórmula 5.
6. Cuente el número de observaciones en cada clasificación o intervalo.
7. Cuente el número total de observaciones.
8. Calcule la probabilidad de suceso que tiene cada intervalo al utilizar la fórmula 1 como sigue:

$$p(x) = \frac{\# \text{observaciones en intervalo}}{\text{total de observaciones en población}}$$



9. Ya que tiene estos valores, realice una gráfica como la 4 y con esto tendrá la distribución de probabilidad:



Esto parece fácil de hacer y lo es cuando se tiene una población tan pequeña y manejable como la distribución de pesos del inventario total de aguacates del comerciante estudiado. Sin embargo, ¿qué haría usted si fuera a determinar la distribución de probabilidad de la temperatura en Morelia el día de hoy?, ¿cómo haría usted para calcular la distribución de probabilidad del precio de una acción que cotiza en bolsa? Ahí los valores son muchos y, ahora se presenta otra pregunta ¿por qué elegir 10 intervalos, cajas o grupos? ¿puede elegir más, como digamos, 1 millón de grupos?

¿Se ha puesto a pensar que, si hace 10 grupos de temperaturas, tendrá probabilidades muy inexactas? Por ejemplo: llegaría a afirmaciones como “La probabilidad de que la temperatura de mañana sea entre 20 ° y 30° es de 30%” quizá usted busque algo como “La probabilidad de que la temperatura sea de entre 20.5° y 21.5° es de 2%”.

Hasta ahora, puede usted observar que el calcular histogramas de frecuencias y las respectivas tablas ayuda mucho para calcular probabilidades de manera intuitiva. Sin embargo, hay otra forma más rápida que requiere menos trabajo y de calcular solo nuestras dos medidas estadísticas de interés: media y desviación estándar. Esta forma consiste, en lugar de hacer tanto trabajo o “talacha” como es contar, clasificar y calcular probabilidades a mano, en solo calcular dos medidas (media y varianza) e insertar los valores de estas en una formulita.

Estas formulitas simples se llaman **funciones de densidad de probabilidad** y existen muchas en la Estadística. Sin embargo, en esta materia veremos solo cuatro de ellas por el uso práctico que se dará en sus labores como profesionalista:

1. La función de densidad de probabilidad “gaussiana” o “normal”.
2. La distribución t-Student.
3. La distribución Xi cuadrada.
4. La distribución F.



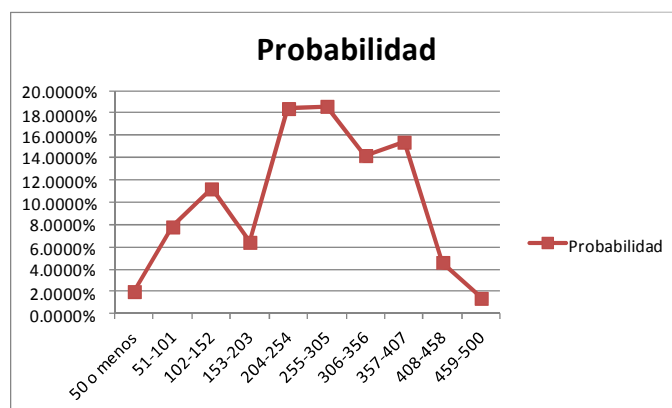
Función de densidad de probabilidad: función matemática que nos sirve para calcular probabilidades de manera más simple (con menos pasos) que con los histogramas. Son más exactas y sirven para cuando tenemos muchos datos o los posibles valores de las observaciones pueden ser infinitamente diferentes.

Función de densidad de probabilidad normal o gaussiana: Función de densidad de probabilidad que es la más utilizada y requiere de solo tres parámetros para su cálculo, el valor aleatorio (x_i) al que se le determinará la probabilidad, la media (μ) y la desviación estándar (σ).

Las últimas dos las estudiaremos de manera simple cuando se entre al tema de comprobación de hipótesis.

Sin embargo, es de necesidad observar que la más usada, por comodidad y en base al Teorema del Límite Central que veremos más adelante, es la normal y, cuando trabajamos con muestras en lugar de poblaciones, la t-Student.

Para ilustrar la función de densidad gaussiana o normal, vea usted de nuevo la gráfica 4 del inventario de aguacates, la cual se organiza con un histograma de frecuencias conformado de 10 grupos o intervalos dados en la tabla 7:

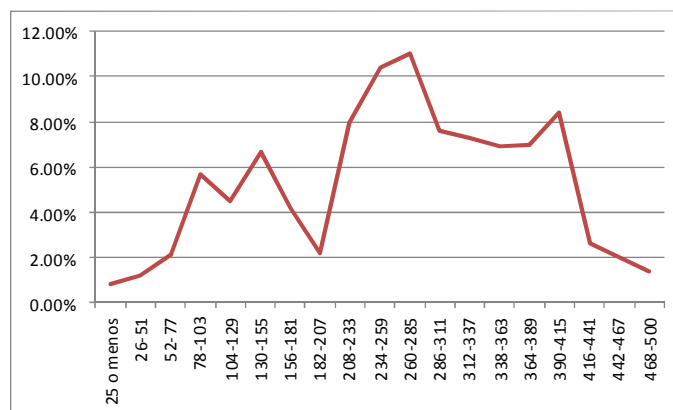


Si quisiéramos incrementar la precisión de nuestro cálculo de probabilidades, podríamos incrementar el número de intervalos de 10 a 20. Con esto, tendríamos la tabla de distribución de frecuencias y distribución de probabilidad dada en la tabla 8, así como la gráfica de distribución de probabilidad de la gráfica 6:

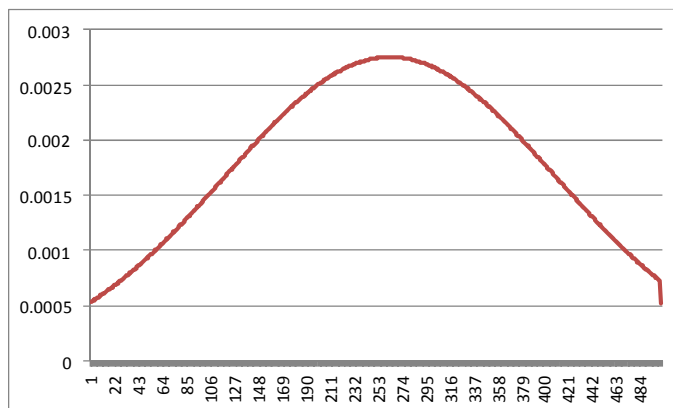


| | Frecuencia | Probabilidad histograma |
|------------|------------|-------------------------|
| 25 o menos | 8 | 0.80% |
| 26-51 | 12 | 1.20% |
| 52-77 | 21 | 2.10% |
| 78-103 | 57 | 5.70% |
| 104-129 | 45 | 4.50% |
| 130-155 | 67 | 6.70% |
| 156-181 | 42 | 4.20% |
| 182-207 | 22 | 2.20% |
| 208-233 | 80 | 8.00% |
| 234-259 | 104 | 10.40% |
| 260-285 | 110 | 11.00% |
| 286-311 | 76 | 7.60% |
| 312-337 | 73 | 7.30% |
| 338-363 | 69 | 6.90% |
| 364-389 | 70 | 7.00% |
| 390-415 | 84 | 8.40% |
| 416-441 | 26 | 2.60% |
| 442-467 | 20 | 2.00% |
| 468-500 | 14 | 1.40% |
| Total | 1000 | |

Tabla 8 Tabla de distribución de frecuencias y de distribución de probabilidades para el ejemplo de los aguacates cuando se incrementa el número de grupos de 10 a 20.



Gráfica 6 Gráfica de la distribución de probabilidad del ejemplo de los aguacates incrementando el número de grupos o intervalos de 10 a 20.



Gráfica 7 Gráfica de distribución de probabilidad para un número de grupos o intervalos de 500.

Ahora imagine usted ¿qué pasaría si tuviéramos un número de grupos o intervalo tan grande como 500? De entrada la tabla sería muy grande que no cabría aquí, la cantidad de trabajo sería mucha y podría generarnos muchos inconvenientes. Sin embargo, suponiendo que exista esa tabla y esa gráfica, la distribución de probabilidad sería algo como lo expuesto en la gráfica 7.

¿Qué sucedió? Primero se inició trabajando con una muestra grande y 10 grupos intervalos donde se iban a clasificar el número de observaciones (los diferentes aguacates en virtud de su peso), segundo, se quiso incrementar la precisión del cálculo de probabilidad ya que se tenían probabilidades para rangos muy abiertos de datos (0 a 50 gramos, 51 a 101 gramos, etc.). Lo que se buscó fue, para incrementar la precisión del cálculo de nuestras probabilidades, reducir el **rango** o tamaño del grupo para incrementar la precisión (y también se incrementó el número de cálculos).

Al incrementar el número de grupos la gráfica de distribución de probabilidad pasó de tener una forma quebrada o accidentada (gráfica 4) a una muy **suavizada** (gráfica 7), la cual nos dice que los pesos de aguacate observados en el inventario se centran alrededor de una media de 264 gramos y tiene una desviación estándar (separación promedio respecto a la media) de 144.9672.

Para calcular una probabilidad como la lograda en la gráfica 7, se necesitan agrupar las observaciones en muchos intervalos y aplicar muchas veces la fórmula 1:

$$p(x) = \frac{\text{\#observaciones en intervalo}}{\text{total de observaciones en población}}$$

Sin embargo, todo este esfuerzo se puede reducir a tres simples pasos:

1.4.5.1 Cálculo de probabilidades con función de densidad de probabilidad normal o gaussiana.



Regla de dedo para hacerlo:



1. De todos los datos que se tienen se calcula la media utilizando la fórmula 2:

$$\mu = \frac{\sum x_i}{n}$$

2. Con la media se calcula la desviación estándar observando que primero debe calcularse la varianza con la fórmula 3:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$$

Aplicar luego la raíz cuadrada a la misma para llegar a la desviación estándar, como lo indica la fórmula 4:

$$\sigma = \sqrt{\sigma^2}$$

3. Ya que se tienen estos dos simples cálculos que puede hacerlos Excel (como veremos en breve), se aplica la formulita de probabilidad conocida como **función de densidad de probabilidad gaussiana**:

Fórmula 5, función de densidad de probabilidad gaussiana:

$$\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

NOTA: ¡No se asuste! No tiene que aprenderse la fórmula. No se le preguntará en examen. Lo que el profesor desea hacer con la misma es hacerle ver una cosa simple: la forma de calcular probabilidades.

Por favor, vea detenidamente la función de densidad de probabilidad gaussiana dada en la fórmula 5. ¿Qué parámetros ya tiene calculados de los datos? Usted puede apreciar que todos los números como e o π ya están dados y no tiene que hacer nada con ellos⁵.

Sin embargo, x_i es el valor de cada resultado esperado. Por ejemplo, si usted espera la probabilidad de sacar un aguacate de 13.789 gramos, se tiene entonces $x_i = 13.789$. Las otras dos variables μ y σ ya las conoce y las calculó con la fórmula 1 de la media y la 4 de la desviación estándar.

⁵ e valdrá siempre 2.7182... y $\pi=3.1416$...



La idea que le quiere transmitir el profesor es simplemente decirle que, por más difícil que se vea la función de densidad de probabilidad de la fórmula 5, en realidad esta le ayuda a simplificar los cálculos ya que, de derivar tantos intervalos como quiera, contar cuánto hay en cada caja, grupo o intervalo y calcular a mano todas las probabilidades, se reduce a simplemente, con todos los datos que tiene a mano x_i , calcular μ y σ y aplicar, a través de la computadora, la fórmula 5.

1.4.6 La función de densidad de probabilidad normal estándar.

Para continuar con el tema de probabilidades y la introducción de esta materia, hay que hacer una última abstracción y hablar de un tipo de función de densidad que es la misma que la gaussiana o la normal; solo que se presenta en otro “idioma” o forma de interpretación.

Para exponer la idea de la distribución de probabilidad normal estándar, imagine usted ahora a dos comerciantes que venden aguacates. Uno vive aquí en Morelia y el otro es una prima de él que vive en Chicago. Los dos tienen el mismo proveedor de fruta y, un día hablando por teléfono, quisieron saber si la calidad de aguacates que les vendía el proveedor (que es el mismo para ambos) era la misma. Para definir si la calidad es la misma, utilizaron una función de densidad de probabilidad gaussiana o normal (por que los dos tienen un inventario de miles de aguacates y les da flojera hacer histogramas).

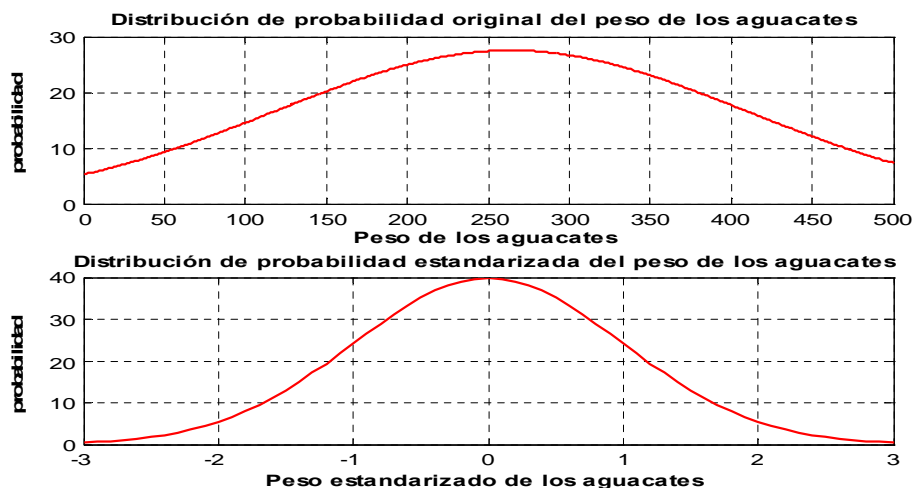
Sin embargo, se toparon con un problema muy grande: el comerciante mexicano registra el peso de sus aguacates en gramos y su prima que vive en Chicago lo tiene en onzas. Si calculan los dos su media y su desviación estándar, verán que los valores no son el mismo y no saben si se debe a un problema debido a la calidad de aguacates del proveedor o a uno de escala. ¿Cómo le hacen para poder comparar sus datos?

En Estadística hay una acción llamada “estandarizar” que consiste en hacer comparables variables aleatorias que, por naturaleza o escala de medida, son diferentes.

Por tanto, lo que se hace es ajustar los datos del inventario de aguacates en las diferentes escalas a valores que sean comparables al aplicar el siguiente ajuste o **estandarización**.

Fórmula 6 Cálculo del valor Z para estandarizar variables:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

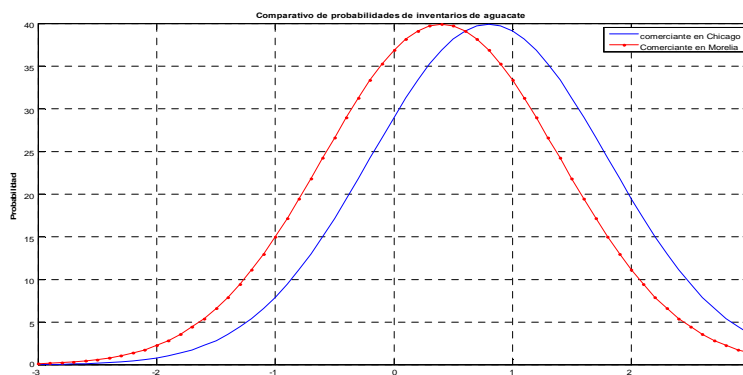


Gráfica 8 Lo que ocurre con la función de densidad de probabilidad normal cuando se estandariza.

Para apreciar el efecto de la estandarización, véase el cambio de función de densidad de probabilidad del comerciante de aguacates mexicano.

Como puede apreciar en la gráfica 8, la media se convierte de un valor de 264 gramos a un cero (en breve veremos qué sucede) y todos los valores cambian de escala de -3 a 3. Es decir, pierden medidas de unidad y se convierten en números reales comparables.

Por ejemplo, suponga ahora que tanto la comerciante de Chicago como el comerciante de Morelia, estandarizan su inventario de aguacates y comparan la función de densidad de probabilidad normal estándar. Entonces podrían llegar a una gráfica como la siguiente:



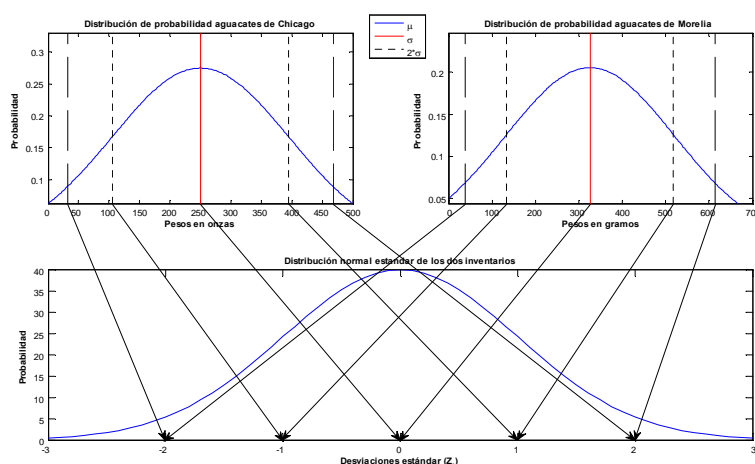
Gráfica 9 Comparativo de distribuciones de probabilidad con valores estandarizados de las observaciones.

Una característica peculiar de los datos es que estos, en lugar de estar unos en gramos y otros en libras, se miden en términos de desviaciones estándar o valores Z_i y es entonces que los inventarios de aguacates de los dos comerciantes pueden ser medidos con las mismas unidades para que los dos puedan contestar la pregunta del problema que se plantearon: Definir si la



calidad del aguacate, medida por su peso, es la misma. De entrada, se puede ver que las distribuciones de probabilidad no son la misma. Si fueran la misma, estarían una sobre la otra. Este tipo de circunstancias son lo que da entrada a lo que en dos temas más por revisar se llama **“comprobación de hipótesis”**. En este caso, se busca comprobar la hipótesis de que los dos inventarios de los comerciantes tienen la misma calidad (mismo peso).

Para terminar de platicar sobre la distribución normal estándar, simplemente se describe cómo se transformaron los datos o la distribución de probabilidad de los dos inventarios con medias y desviaciones estándar en unidades diferentes (gramos y onzas) a unidades más similares como son “desviaciones estándar” o valores Z_i .



Gráfica 10 El cambio de escala y valores que sufre la función de densidad de probabilidad del inventario de los dos comerciantes cuando se estandarizan.

Nótese cómo la media de los dos inventarios se hizo cero y las desviaciones estándar se hicieron comparables al adoptar magnitudes de números reales sin escala que aquí se presentan con valores en el intervalo dado por: $[-3, +3]$. Aunque se pueden tener distribuciones de probabilidad con intervalos mayores, digamos $[-6, +6]$ o mayores, este tipo de casos son poco comunes o propios de materias que requieren de una Estadística más amplia como la utilizada por economistas, ingenieros industriales, ingenieros financieros, actuarios y otro tipo de profesionistas que le ayudarán en su vida profesional. Dado que el objetivo de la presente materia es instruirle en problemas propios de la administración, dejando situaciones teóricamente más amplias a otros profesionistas; lo que debe preocuparle es aprender estos conceptos para entenderse bien con este tipo de profesionistas.

Para poder hacer que una variable aleatoria, como es los valores que el peso de un inventario de aguacates puede tener, sea comparable, esta debe expresarse en términos ya no de sus valores originales sino estandarizados. Esto lleva a la siguiente regla nemotécnica o de dedo:



1.4.6.1 Regla de dedo para comprender por qué utilizar una distribución normal estándar:



1. Cuando se desean comparar dos poblaciones cuyas unidades de medida no sean las mismas o, peor aún, cuando no se tienen desviaciones estándar comparables, se debe utilizar ya no una función de densidad de probabilidad normal común y corriente; sino una estandarizada.
2. Para poder utilizar una distribución normal estándar, es necesario ya no utilizar los valores originales de nuestro inventario sino más bien hacer una operación que se conoce como **“Estandarizar los valores de la variable”**.
3. La estandarización de valores se logra con la fórmula 6:

$$Z_i = \frac{x_i - \mu}{\sigma}$$

1.4.7 El cálculo de la probabilidad utilizando la normal estándar y las tablas correspondientes.

Ahora usted ha visto la principal función de densidad de probabilidad que se utiliza en la Estadística para ciencias administrativas: La distribución normal estándar. Ya que usted estandarice los valores de sus variables aleatorias, estará usted en capacidad de saber cómo se calculan probabilidades de eventos cuando se tienen solamente los datos de las observaciones de la población con que se trabaja al utilizar la distribución de probabilidad normal estándar.

Para ello se utilizarán dos posibles herramientas. Una de ellas es Excel (a la que se elaboró un tutorial especial) y otra las tablas de probabilidades que aparecen en muchos libros de texto y que usted podrá bajar fácilmente en la siguiente liga:

<http://www.droscardelatorre.com/classmat/UMSNH/FCCA/ESTADISTICAII/tablaavaloresz.htm>

Como se dijo antes en la regla de dedo recién expuesta, para poder trabajar con distribuciones de probabilidad normal estándar, será necesario convertir los valores de las variables aleatorias en valores estandarizados utilizando la fórmula 6.

Cuando tiene usted el valor estandarizado o Z_i , simplemente se remite a la tabla como la citada en la liga anterior y determina la probabilidad de suceso. Para dar un ejemplo de cómo hacer esto, piense usted de nuevo en el ejemplo del nivel de llenado de las botellas de agua que se citó previamente. Suponga ahora que el llenado medio de las últimas 2,000 botellas ha sido de $\mu=910$ ml con una desviación estándar de $\sigma=75.3$ ml. Conociendo estos dos simples datos, diga ahora usted ¿Cuál es la probabilidad de que la siguiente botella que se tome aleatoriamente de la línea de producción tenga un llenado de $x_i = 970$ ml?



Para responder esto rápidamente, usted deberá primero estandarizar el valor objetivo x_i .
Esto es:

$$Z_i = \frac{x_i - \mu}{\sigma} = \frac{970 - 910}{75.3} = 0.79681275$$

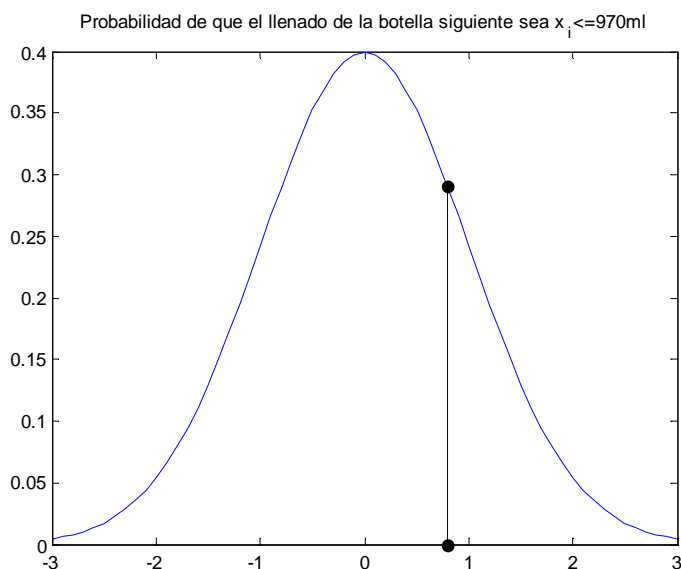
Para trabajar con una tabla de probabilidades como la dada, es necesario redondear a dos dígitos el valor Z_i . Esto sería $Z_i = 0.80$. Con este valor se llegaría a la siguiente probabilidad en la tabla dada⁶:

| Z | 0.00 | 0.01 | |
|------|-----------|-----------|-------|
| 0.00 | 0.0000000 | 0.0039894 | 0.007 |
| 0.10 | 0.0398278 | 0.0437953 | 0.047 |
| 0.20 | 0.0792597 | 0.0831662 | 0.087 |
| 0.30 | 0.1179114 | 0.1217195 | 0.125 |
| 0.40 | 0.1554217 | 0.1590970 | 0.162 |
| 0.50 | 0.1914625 | 0.1949743 | 0.198 |
| 0.60 | 0.2257469 | 0.2290691 | 0.232 |
| 0.70 | 0.2580363 | 0.2611479 | 0.264 |
| 0.80 | 0.2881446 | 0.2910299 | 0.293 |
| 0.90 | 0.3159399 | 0.3185887 | 0.321 |
| 1.00 | 0.3413447 | 0.3437524 | 0.346 |

Ilustración 1 Determinación de una probabilidad dado el valor Z del ejemplo y la tabla de probabilidades

Esto llevaría a observar que la probabilidad de tener un nivel de llenado de 970 ml es de 22.88%, la cual se puede representar como en la gráfica 11 en donde se relaciona ahora el punto del valor Z_i del valor que se desea buscar con la función de densidad de probabilidad normal estándar.

⁶ Si el valor Z_i hubiera sido 0.81, la probabilidad hubiese sido la de la columna del lado derecho (0.2910299 o 29.10%) y así sucesivamente.



Gráfica 11 La probabilidad de obtener un llenado de 970 ml dado $\mu=910$ ml y $\sigma=75.3$ ml que lleva a $Z=0.7968$.

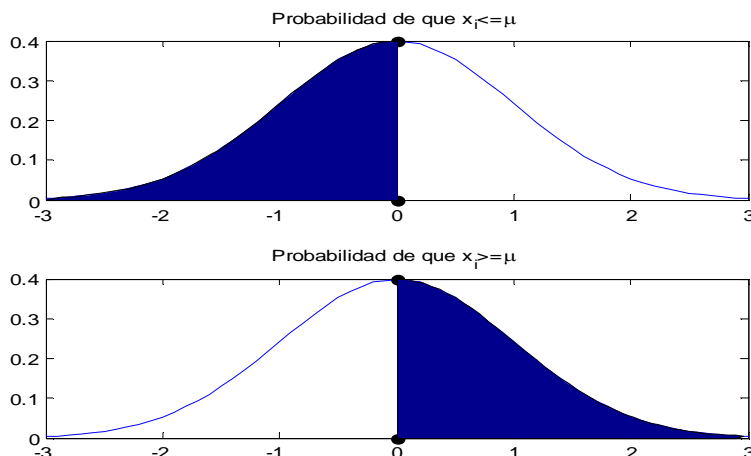
1.4.7.1 Diferentes formas de calcular una probabilidad. Los valores de probabilidad acumulada.

En base a lo revisado, se puede apreciar que el valor de la probabilidad es muy puntual si solo se desea saber cuánto vale la probabilidad de un valor determinado como puede ser un nivel de llenado específico de 970 ml. Sin embargo, en la vida cotidiana, las probabilidades se determinan en base a intervalos de datos.

Esto implica que lo que en realidad nos está dando la tabla es la probabilidad de que la siguiente botella de agua tenga un nivel de llenado de 910 ml (valor medio) a 970 ml. Esto es así ya que la fórmula de cálculo de la función de densidad de probabilidad así lo pide.

Si deseamos conocer la probabilidad de tener valores iguales a 910 ml (μ) (o sea entre 910 ml y 910 ml) usted podrá observar en la tabla que da una probabilidad de **cero** (probabilidad de un valor $Z_i=0$ es cero en la ilustración 1 que corresponde a la tabla de probabilidades). Por tanto las probabilidades de la desviación normal estándar trabajan con intervalos de valores y no con valores puntuales.

Por ejemplo, la probabilidad de tener niveles de llenado iguales o mayores que el nivel medio de $\mu=910$ ml es de 50%. Esto se ilustra en la parte inferior de la gráfica 12. En la superior se expone el caso contrario: El nivel de llenado es menor o igual a la media de 910 ml. Como se puede notar, el valor total del área de la superficie sombreada tiene una magnitud de 50%. Es decir, la probabilidad de tener ya sea valores mayores e iguales que la media o menores e iguales que la media.



Gráfica 12 Probabilidad de que x_i sea menor o igual a μ (parte superior) o de que sea mayor o igual a dicho valor (parte inferior).

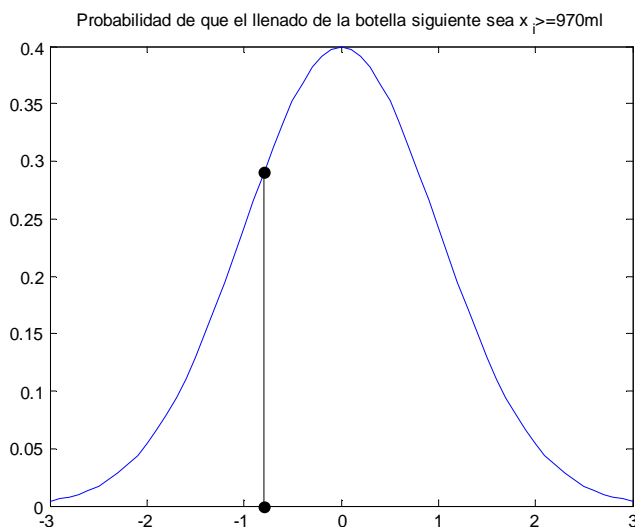
En base a lo previamente descrito, en la vida cotidiana se pueden tener los siguientes casos de cuantificación de probabilidades (sigamos con el ejemplo del nivel de llenado de botellas):

1. La probabilidad de que el valor del evento aleatorio sea menor o igual a b (por ejemplo que sea mayor o igual a 970 ml).
2. La probabilidad de que el valor del evento aleatorio sea mayor o igual a b por ejemplo que sea menor o igual a 970 ml).
3. La probabilidad de que el valor del evento se encuentre entre a y b (por ejemplo que el valor futuro se encuentre entre el valor medio de 910 ml y 970 ml).

Para poder responder estas preguntas, primero observe la tabla de probabilidades que tiene a mano y que bajó de la liga previamente mencionada páginas atrás. Usted podrá apreciar que solo le dan los valores Z_i positivos o a la derecha del cero. ¿Qué pasaría ahora si lo que usted desea calcular es la probabilidad de tener un nivel de llenado de 850 ml partiendo de la misma media de $\mu=910$ ml y $\sigma=75.3$ ml? Ahora usted tendría un valor Z_i de:

$$Z_i = \frac{x_i - \mu}{\sigma} = \frac{850 - 910}{75.3} = -0.79681275$$

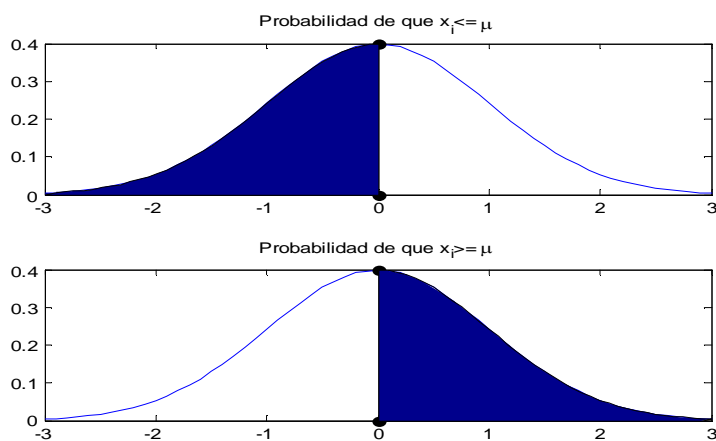
Este lleva a una probabilidad de 28.72% que se ilustra en la gráfica 12.



Gráfica 13 La probabilidad de obtener un llenado de 850 ml dado $\mu=910$ ml y $\sigma=75.3$ ml que lleva a $Z=0.7968$.

¿Para qué se le expuso este otro ejemplo que es el inverso del anterior? Piense usted que quiere resolver el primer caso o pregunta de las tres planteadas anteriormente: ¿Cuál es la probabilidad de que la botella que se tome aleatoriamente de la línea de producción tenga un llenado menor o igual a 970 ml?

Para responder esto, es de necesidad observar que la probabilidad de tener eventos menores o iguales a la media de la población o mayores o iguales a dicho valor siempre será, como se vio en la gráfica 12, de 50% en la distribución normal estándar:



Para resolver el problema ¿cuál es la probabilidad de que el nivel de llenado de la botella de agua sea menor o igual a 970 ml? Se calcularía la probabilidad como sigue:

1. Se sabe que el valor que se busca determinar como objetivo (x_i) es mayor a la media. Por lo tanto, se tiene la probabilidad de que una variable aleatoria adopte valores menores o

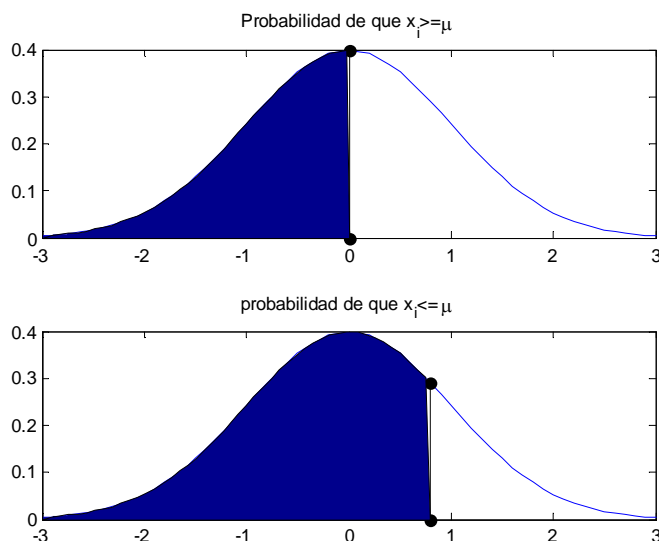


iguales la media que, de antemano por lo visto en la parte superior de la gráfica 12, se sabe es de 50%. O sea $p(X \leq \mu) = 50\%$.

2. Se sabe, por el ejercicio previo, que la probabilidad de que la botella tenga un nivel entre la media ($\mu=910$ ml) y 970 ml es de 22.88%. O sea $p(\mu \leq x_i) = 22.88\%$.
3. Por tanto, si se suman las probabilidades de tener un valor menor o igual a la media (μ) y el valor puntual de tener un llenado entre la media ($\mu=910$ ml) y 970 ml, se llega entonces a una probabilidad total de:

$$\begin{aligned} p(X \leq 970\text{ml}) &= p(X \leq \mu) + p(\mu \leq x_i) \\ &= 50\% + 22.88\% = 72.88\% \end{aligned}$$

Esto se ilustra en la gráfica 14.



Gráfica 14 probabilidad de que el nivel de llenado sea menor o igual a 970 ml.

Ahora téngase presente la segunda pregunta ¿cuál es la probabilidad de que el nivel de llenado sea mayor o igual a 850 ml? Para responder esto se siguieron los mismos pasos anteriores de la siguiente forma:

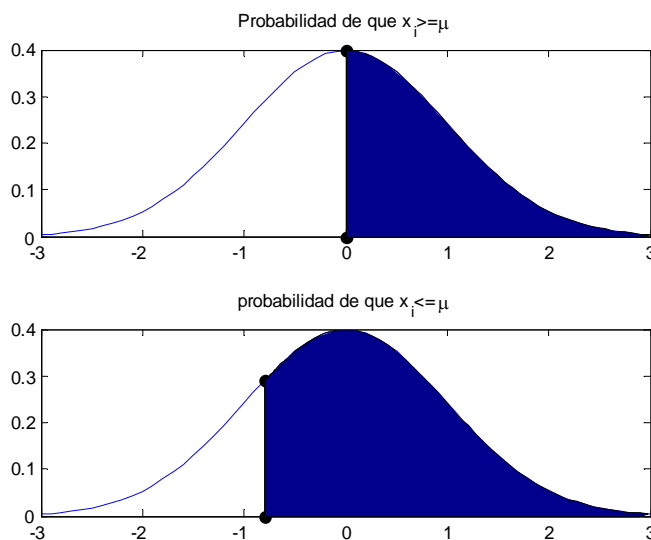
1. Se sabe que el valor que se busca determinar como objetivo (x_i) es menor a la media. Por lo tanto, se tiene la probabilidad de que una variable aleatoria adopte valores mayores o iguales la media que, de antemano, se conoce como de 50% por lo visto en la parte inferior de la gráfica 12. O sea $p(X \geq \mu) = 50\%$.
2. Se sabe, por el ejercicio previo, que la probabilidad de que la botella tenga un nivel puntual o específico de 850 ml es de 22.88%. O sea $p(x_i \leq \mu) = 22.88\%$.



3. Por tanto, si se suman las probabilidades de tener un valor mayor o igual a la media (μ) y el valor puntual de tener un llenado de 850 ml, se llega entonces a una probabilidad total de:

$$\begin{aligned} p(x_i \leq 860ml) &= p(X \geq \mu) + p(x_i \leq \mu) \\ &= 50\% + 22.88\% = 72.88\% \end{aligned}$$

El análisis visual se presenta en la gráfica 15. En la parte superior se expone la probabilidad acumulada hasta la media y en la inferior la probabilidad total.



Gráfica 15 probabilidad de que el nivel de llenado sea menor o igual a 970 ml.

Para responder el último tipo de pregunta: ¿cuál es la probabilidad de que el nivel de llenado de la botella esté entre 860 ml y 970 ml? Simplemente se obtienen las probabilidades de ambos casos partiendo de sus valores Z_i :

1. Se sabe, por el ejercicio previo, que la probabilidad de que la botella tenga un nivel puntual o específico de 970 ml es de 22.88%. O sea $p(\mu \leq x_i) = 22.88\%$.
2. Se sabe, por el ejercicio previo, que la probabilidad de que la botella tenga un nivel puntual o específico de 850 ml es de 22.88%. O sea $p(x_i \leq \mu) = 22.88\%$.
3. Por tanto, se suman las dos probabilidades de suceso: y se llega a:

$$\begin{aligned} p(860ml \leq x_i \leq 970ml) &= p(x_i \leq \mu) + p(\mu \leq x_i) \\ &= 28.72\% + 28.72\% = 57.44\% \end{aligned}$$



Como puede apreciar usted, no es tan difícil calcular la probabilidad de un evento utilizando la probabilidad normal estándar. Simplemente deberá usted pensar que su punto de referencia es la media y definir si el valor buscado se encuentra por encima o por debajo de esta.

Como un último tipo de problema de cálculo de probabilidades que podría presentarse en su vida cotidiana se tiene el siguiente ejemplo: Determine usted cuál es la probabilidad de que la siguiente botella que tome de muestra tenga un nivel de llenado entre 940 y 960 ml. Como puede apreciar, los dos valores buscados se encuentran arriba de la media. Por tanto, lo que debe calcular son los dos valores Z de cada caso y luego restar las probabilidades. Esto es, siguiendo la tabla de probabilidades dada:

1. Calcular la probabilidad de que el nivel de llenado se encuentre el nivel de la media ($\mu=901$ ml) y el de 940 ml:

$$p(\mu \leq 940ml) = p(z_i = 0.39840637) = 15.45\%$$

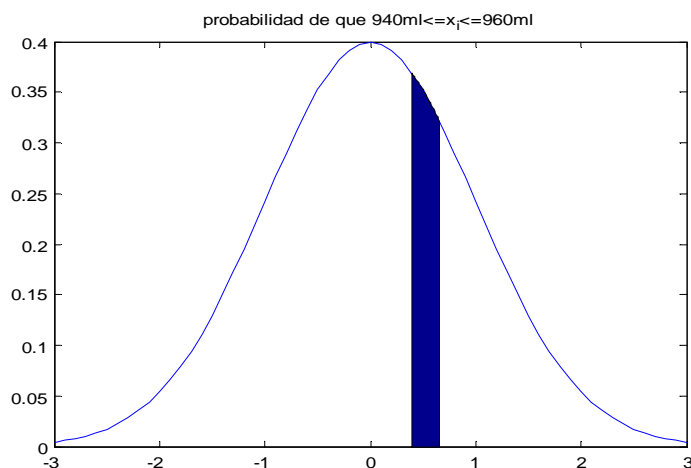
2. Calcular la probabilidad de que el nivel de llenado tenga una magnitud entre el nivel de la media ($\mu=901$ ml) y el de 960 ml:

$$p(\mu \leq 960ml) = p(z_i = 0.66401062) = 24.66\%$$

3. Restar a la probabilidad mayor, la probabilidad menor y, con esto, se tiene la probabilidad de que la botella tenga un nivel de llenado entre 940 ml y 960 ml:

$$\begin{aligned} p(940ml \leq x_i \leq 960ml) &= p(\mu \leq 960ml) - p(\mu \leq 940ml) \\ &= 24.66\% - 15.45\% = 9.18\% \end{aligned}$$

Esto se puede apreciar en la gráfica 16 en donde se marca el área sombreada con las probabilidades de suceso.



Gráfica 16 Probabilidad de que el nivel de llenado de la botella se encuentre entre 940 ml y 960 ml.



Múltiples casos como este se pueden presentar. Para ello se le sugiere siempre guiarse cuál es el valor menor del intervalo (por ejemplo 940 ml para el caso estudiado) y cuál es el mayor (por ejemplo 970 ml) y ver si estos tienen la media o promedio dentro de ellos.

Por tanto las siguientes recetas o reglas de dedo podrían servirle como guías generales para el cálculo de probabilidades con la distribución normal estándar:

Probabilidades cuando se tiene un intervalo con números finitos:

1. Defina como a el valor menor del intervalo. Por ejemplo si usted define la probabilidad que el nivel de llenado esté entre 490 ml y 960 ml, $a = 940ml$.
2. Defina como b el valor mayor del intervalo. Por ejemplo si usted define la probabilidad que el nivel de llenado esté entre 490 ml y 960 ml, $b = 960ml$.
3. Calcule el valor Z_i de cada intervalo. O sea el valor de 940 ml y 960 ml:

$$Z_i = \frac{x_i - \mu}{\sigma} = \frac{940 - 910}{75.3} = 0.39840637$$

$$Z_i = \frac{x_i - \mu}{\sigma} = \frac{960 - 910}{75.3} = 0.66401062$$

4. Determine las probabilidades con tablas:

$$p(Z_{940ml} = 0.39840637) = 15.48\%$$

$$p(Z_{960ml} = 0.66401062) = 24.66\%$$

5. Reste la probabilidad de b a la de a y llegará a la probabilidad buscada:

$$p(940ml \leq x_i \leq 960ml) = 24.66\% - 15.48\% = 9.18\%$$

Probabilidades cuando se tiene un intervalo con números infinitos:

Este caso, por más raro que suene, simplemente consiste en determinar cuál es la probabilidad de que un valor aleatorio sea mayor o igual o menor o igual a un número determinado. Por ejemplo, la probabilidad de tener cualquier nivel de llenado menor o igual que 960 ml o mayor e igual que 860 ml. Los pasos a seguir, como se vio en los ejemplos correspondientes, son:

1. Si el valor objetivo está arriba del promedio de los datos, se debe determinar la probabilidad de que los valores aleatorios sean mayores o iguales que la media. Como usted bien sabe por la gráfica 12, esta será siempre de 50%. Lo mismo sucede cuando



busca usted la probabilidad de que los valores sean menores o iguales que la media que también siempre es de 50%.

2. Calcule ahora los valores Z de que el valor aleatorio se encuentre entre la media y el número objetivo buscado. Por ejemplo:

$$p(\mu \leq 970ml) = 22.88\%$$

$$p(940ml \leq \mu) = 22.88\%$$

3. Ya que tiene la probabilidad del valor buscado respecto a la media y la probabilidad del resto de valores aleatorios infinitos respecto a μ , simplemente suma estas probabilidades. En el caso de los dos ejemplos vistos, esto sería:

$$\begin{aligned} p(X \leq 970ml) &= p(X \leq \mu) + p(\mu \leq x_i) \\ &= 50\% + 22.88\% = 72.88\% \end{aligned}$$

Ya para cerrar este tema de introducción a Estadística II, se le sugiere consultar el tutorial de cómo calcular las probabilidades utilizando Excel, el cual puede conseguirlo en la siguiente liga de internet:

www.drocardelatorre.com/classmat/UMSNH/FCCA/ESTADISTICAII/probnormestexcel.html



2 Teoría del muestreo

Ya que se revisaron los principales conceptos de Estadística I (Estadística descriptiva y probabilidad) que serán pieza fundamental de las revisiones de la Estadística inferencial, es necesario conocer otra parte fundamental que es la Teoría del muestreo.

De entrada, es de necesidad observar que en todo lo revisado hasta ahora se ha supuesto que el conjunto de datos con que se trabaja son muestras. Sin embargo, en la realidad es difícil trabajar con poblaciones enteras sino, más bien, con muestras. Para ilustrar esto, se recuerda la definición de población:

Población: Conjunto de todas las observaciones posibles sobre una característica de interés observada.

Lo que es de interés destacar en la definición es la frase “de todas las observaciones posibles”. En casos de la vida real, por ejemplo el precio que ha tenido, tiene y puede tener el precio de una acción; los valores puntuales de la temperatura que puede tener Morelia en toda la existencia de la ciudad, los pesos de los aguacates que un comerciante tuvo, tiene y tendrá, etc. son datos que difícilmente se conocerán y, si se logran, será muy caro obtenerlos.

Otro ejemplo clásico es conocer el número de habitantes por casa en México. Lo que hace el INEGI es simplemente evitarse la pena de tener que ir a tocar, de puerta en puerta, a todas las casas y preguntar el número de habitantes al tomar una muestra de determinadas casas, en determinadas ciudades y en determinadas zonas geográficas para tratar de estimar el número de habitantes en el país (esto es muestro por racimos como veremos en breve).

Lo que en la vida real se hace, como el INEGI, es trabajar no con la población de datos sino con una muestra de los mismos. Es decir, una parte de estos.

Muestra: subconjunto de una población de la cual se deriva.

2.1 Tipos de muestreo

En este tema simplemente se explorarán las características relacionadas a la forma de hacer muestras. En el siguiente tema: la inferencia, se observará que el cálculo de parámetros como la media y la desviación estándar cambian en una muestra respecto a una población.

Una parte de importancia a observar es que una muestra, según el tipo de estudio que se haga, se realiza de diferentes formas. Por ejemplo, la muestra de un grupo de aguacates en inventario o la que se obtiene con las botellas de agua extraídas de una línea de producción se forma de manera



diferente a la que emplea una empresa de mercadotecnia para probar la demanda de un producto. Esta diferencia radica en el uso que se dará a los datos. Por ejemplo, el tener que saber cuántas botellas de agua no satisfacen los estándares de calidad es una aplicación diferente a saber ¿cuál es la demanda de bebidas alcohólicas en el sector de clase media de una sociedad tanto en mujeres como en hombres?

En virtud de esto, se tienen cuatro tipos de muestreo o forma de hacer muestras comúnmente utilizados:

1. Muestreo aleatorio simple.
2. Muestreo sistemático.
3. Muestreo de racimo.
4. Muestreo estratificado.

2.2 Muestreo aleatorio simple

Como su nombre lo indica, consiste en seleccionar, de manera aleatoria, una serie de observaciones, objetos o datos de una población sin seguir algún tipo de agrupamiento específico. Un ejemplo simple, retomando el caso de los niveles de llenado de las botellas, sería ir una directamente de la línea de producción, luego dos y luego una y así sucesivamente hasta llegar a un número determinado de botellas u observaciones. Por ejemplo, 500.

Otra forma de hacer muestras aleatorias sería, por ejemplo usted desea hacer una muestra de afiliados al seguro social y elige de manera aleatoria a estos en función de los últimos tres números de folio de afiliación. Para esto, usted utiliza un generador de números aleatorios, como el que tiene Excel y elije primero al usuario cuyo número de seguro social termina en 472, luego corre otro número y elije al usuario con número 589 y así sucesivamente hasta que tenga un número determinado de usuarios.

Este tipo de muestreo es el más común pero tiene la limitante de que se elijen muestras aleatorias y algún tipo de característica (como puede ser color, procedencia, género, etc. En un grupo de personas) puede no ser tomada en cuenta.

2.3 Muestreo sistemático

Este tipo de muestreo consiste en elegir a un objeto en función de intervalos predeterminados. Por ejemplo, piense usted que tiene 2,000 cajas de aguacate foliadas todas y listas para empacarse a Estados Unidos. Ahora elige primero la caja número 20, luego la 40 y así sucesivamente hasta la 2,000. Esto le deja con una muestra de 100 cajas a las que le puede realizar el estudio estadístico que necesita.



El muestreo sistemático es muy útil. Sin embargo, tiene una limitante llamada **introducción de sesgo**. Para ilustrar la idea, se le da un ejemplo: Suponga que usted es dueño de una cadena de farmacias y desea muestrear el nivel de ventas de sus sucursales en Morelia haciendo el muestreo solo los días lunes. De entrada esto puede ser bueno y práctico. Sin embargo, puede tener la limitante de que el patrón de consumo de sus clientes es bajo los días lunes ya que es inicio de semana y desean gastar en otras cosas su dinero. Claramente, de hacer este tipo de muestreo, usted estaría estimando ventas menores y correría el riesgo de tomar decisiones mal informadas.

2.4 Muestreo estratificado.

En este tipo de muestreo, se divide la población de datos en grupos homogéneos (mujeres y hombres, intervalo de pesos, etc.) y se determina qué proporción representa cada estrato o grupo. Cuando se analizan las características y parámetros como media, desviación estándar, etc., se ponderan los mismos en función de su representación o proporción de peso respecto la población total y con esa ponderación se obtienen los parámetros y probabilidades totales de dicha población con este tipo de muestra.

Por ejemplo, piense usted que desea saber el número medio de personas que entran a sus farmacias en función de su edad. Por ejemplo, tendría usted una tabla como la siguiente:

| Estrato | Porcentaje del total | Frecuencia | Frecuencia ponderada (porcentaje x frecuencia) |
|--|----------------------|------------|---|
| menos de 18 años | 30% | 50 | 15 |
| 19-39 años | 40% | 230 | 92 |
| 40-69 años | 20% | 187 | 37.4 |
| 69 o más | 10% | 74 | 7.4 |
| Media ponderada de las frecuencias de la población | | | 151.8 |

Tabla 9 Muestreo estratificado y parámetros calculados con el mismo.

2.5 Muestreo de racimo.

Esta forma de muestrear se parece a la anterior, con la diferencia de que primero se hacen estratos y luego se seleccionan miembros, datos u observaciones de cada uno de los estratos de una manera aleatoria. Por ejemplo, usted desea saber cuántas televisiones existen en la ciudad de Morelia. Entonces, usted divide la ciudad en colonias y elige, de cada colonia y de manera aleatoria, una serie de casas, toca la puerta y pregunta el número de televisiones que hay en cada una. Con esto toma muestras aleatorias no de la totalidad de la población sino de cada uno de los grupos que usted formó.



2.6 Diferencias operativas en cada uno de los tipos de muestreo y determinación del empleado en Estadística Inferencial.

En el siguiente cuadro se destacan las principales diferencias operativas o de ejecución de cada uno de los tipos de muestreo estudiados.

| Cualidad | Tipo de muestreo | | | |
|------------------------------------|---|--|--------------------------------|----------------------------|
| | Aleatorio simple | Sistematizado | Racimo | Estratificado |
| Selección de datos para la muestra | Se toman datos directamente de la población | Se toman datos directamente de la población | Por grupos | Por grupos |
| Selección de datos | Aleatoria en el total de la población | Siguiendo una regla matemática (cada 20 números, los días lunes, etc.) | Aleatoria dentro de cada grupo | Total de datos en el grupo |

De los 4 tipos de muestreo vistos, el que se utilizará en Estadística inferencial es el aleatorio simple.

La afirmación anterior surge del hecho de que los otros tres tipos no son más que técnicas específicas de recolección de datos que derivan en una muestra a la que se calcularán probabilidades muestrales y con los cuales usted, como profesional tomador de decisiones, deberá realizar el análisis estadístico que requiera. En pocas palabras, los diferentes métodos de muestreo se aproximan al muestreo aleatorio.

2.7 Diseño de un experimento: el proceso que se sigue para tomar decisiones.

En Estadística aplicada a los negocios, es importante conducir de manera apropiada la toma de decisiones. Si usted a esta altura ya llevó una clase de Métodos de investigación o metodología de la investigación, recordará el método científico. Aunque éste último es más apropiado para la generación de conocimiento científico, la forma en cómo se llega a una conclusión y a la toma de decisiones en los negocios es muy similar.

Para poder decidir usted sobre algo, primero debe plantearse un objetivo del cual se elabora una hipótesis, luego se comprueba la misma y se decide en base a esta conclusión. Por ejemplo, en el caso de Steve Jobs, él tenía como objetivo determinar si su modelo de computadora tendría




mayor preferencia respecto a su competencia. Para esto, tuvo que plantear una hipótesis a demostrar: “La computadora de mi compañía es más preferida que la de mi competencia”. Para demostrar esta hipótesis, tuvo que utilizar la Estadística y hacer un experimento consistente en formar una muestra estratificada por tipo de usuario (arquitectos, ingenieros, matemáticos, programadores, diseñadores, amas de casa, estudiantes), calcular estadísticas (media y varianza muestrales) y luego aplicar una técnica de Estadística inferencial que veremos en breve llamada “comprobación de hipótesis”. Si con esta técnica demuestra como válida su hipótesis, lo que la compañía de Jobs hará es lanzar al mercado su nueva computadora.

En este subtema, dado este ejemplo, se visitan los pasos que deben seguirse en este proceso de análisis estadístico llamado “diseño de experimento”.

El diseño de experimento o pasos del análisis estadístico a seguir son los siguientes (Se utilizará el ejemplo de los comerciantes de aguacate):

1. **Definir el objetivo:** Los comerciantes definieron como objetivo determinar que la calidad de sus inventarios es la misma.
2. **Definir lo que se medirá:** Aquí los comerciantes definieron “calidad” como el peso de sus aguacates. En pocas palabras pusieron una hipótesis dada por: “Si nuestros inventarios de aguacates tienen el mismo peso, comparten la misma calidad”.
3. **Definir el tamaño de muestra:** Aquí los comerciantes decidieron no trabajar con la totalidad de su inventario porque son miles de aguacates pero acordaron tomar una muestra de 200 aguacates (cómo definir este número lo veremos en breve).
4. **Analizar los datos:** Aquí se emplean técnicas estadísticas, como es la comprobación de hipótesis, para concluir si el objetivo planteado se cumple o no. Por ejemplo, los comerciantes determinaron, con técnicas estadísticas, que sus inventarios son iguales.
5. **Conclusión y toma de decisiones:** En este punto, en base al diseño del experimento seguido hasta ahora, se concluye que los inventarios tienen la misma calidad y toman la decisión de no reclamar al proveedor.

A manera de síntesis de dedo, se presentan los pasos del diseño experimental o proceso de análisis estadístico: 

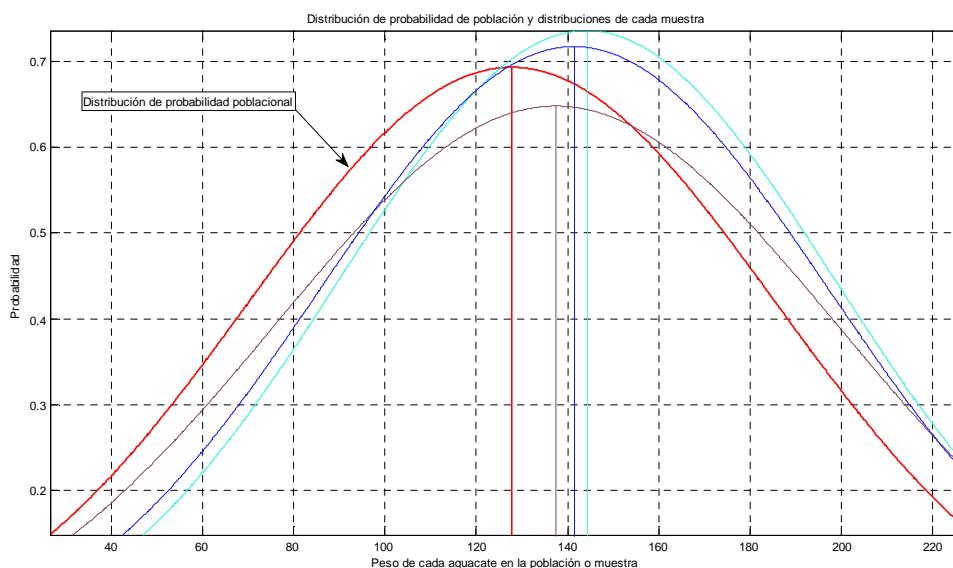
1. **Definir el objetivo.**
2. **Definir lo que se medirá** (Aquí se plantea una hipótesis estadística).
3. **Definir el tamaño de muestra.**
4. **Analizar los datos.**
5. **Conclusión y toma de decisiones.**



2.8 Distribuciones de probabilidad muestrales

Hasta ahora, se ha trabajado con el supuesto de que los datos que se han estudiado pertenecen a una población. Es decir, se ha supuesto que los datos con que se trabaja es la totalidad que se pueden tener. Sin embargo, al introducirnos en este nuevo tema de Teoría del muestreo, hemos visto que, en la mayoría de las ocasiones, es difícil obtener y manipular todos los datos de una población. Por ejemplo, a los comerciantes les era costoso y tedioso hacer un análisis estadístico de la totalidad de su inventario de aguacates con miles de piezas. Más bien, lo que hicieron es tomar unos cuantos (una muestra) para hacer inferencias sobre las propiedades del resto de la población.

¿Qué sucede cuando hacen esto? Para responder esto imagine que tiene una población total de 5,000 aguacates con diferentes niveles de peso en gramos. Esta población total tendrá una distribución de probabilidad determinada. Ahora, si se extrae una muestra de 30 aguacates, esta tendrá un promedio y una desviación estándar y, a su vez, una distribución de probabilidad con una forma y valores determinados. Si se repite dos veces más el ejercicio de muestreo, se verá que las medias, las desviaciones estándar y las distribuciones de probabilidad son diferentes por lo que, si se está trabajando con una muestra, es muy probable que esta tenga fluctuaciones en dicha media. Para ilustrar esto se tiene la siguiente gráfica:



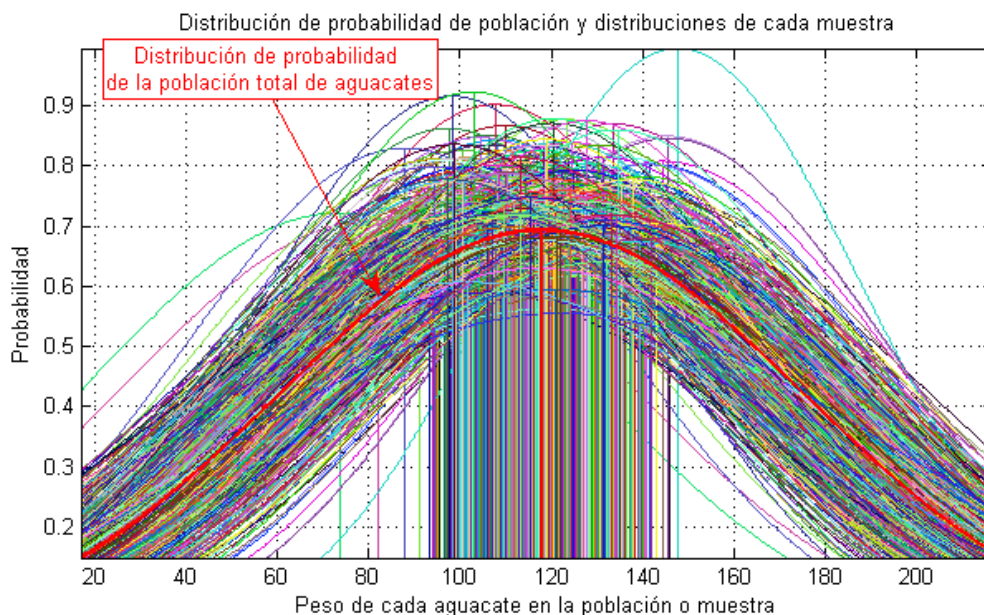
Gráfica 17 Comparativo de la media y distribución de probabilidad (dada también por la desviación estándar) de la población total de aguacates y de las tres muestras generadas de manera aleatoria.

En la misma, se puede apreciar la diferencia en la forma de las distribuciones de probabilidad y las medias de las tres muestras generadas con el método de muestreo aleatorio simple (se tomaron 30 aguacates de manera aleatoria de un total de 5,000). Es decir, no son estables. Por lo tanto, el



hacer un análisis estadístico con muestreo observando que tanto la media como la desviación estándar pueden ser diferentes de muestra en muestra, implica que se tiene incertidumbre o poca seguridad de tomar decisiones ya que la media y desviación estándar de la muestra no son la misma que la de la población.

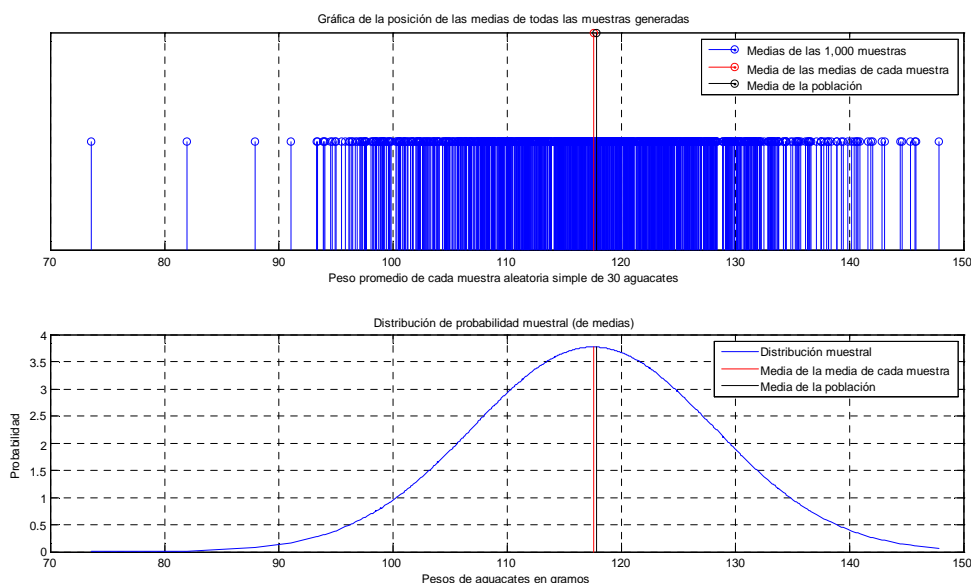
Para dar mayor idea, suponga usted que se generan 1,000 muestras diferentes de 30 aguacates cada una. El comportamiento de las medias y distribuciones de probabilidad en cada caso se presentan en la gráfica 18.



Gráfica 18 Comportamiento de la media, desviación estándar y distribución de probabilidad de cada una de las 1,000 muestras de 30 aguacates generadas con el método de muestreo aleatorio simple.

En la misma se aprecian todas las distribuciones de probabilidad de las 1,000 muestras generadas con el método de muestreo aleatorio simple. También se señala la distribución de probabilidad poblacional.

Aquí, de cuenta nueva, se puede observar cómo la media y las probabilidades de cada muestra cambian.



Gráfica 19 (Parte superior) todas las medias de las muestras generadas, el promedio de las medias y la media poblacional. (Parte inferior), la distribución de probabilidad muestral, el promedio de medias de cada muestra y la media poblacional.

Todas estas medias, dado que son muestrales, pueden también tener una distribución de probabilidad. Esta se conoce como **la distribución de probabilidad muestral** y la formación de la misma se presenta en la gráfica 19. Primero concentre su atención en la parte superior. Ahí se le presentan las medias de todas las muestras generadas. Luego se calcula un promedio (o media) de cada una de esas medias y se compara el valor de la media de la población de aguacates (todo el inventario).

Ahora, en la parte inferior, se genera, con el promedio de medias de cada muestra y la correspondiente desviación estándar⁷, la distribución de probabilidad muestral.

Dado que en la mayoría de las ocasiones, usted utilizará muestras y no poblaciones⁸, esta distribución de probabilidad muestral será la que se utilizará. Ahora, dos preguntas naturales que usted podría tener serían ¿Existe una diferencia entre la normal estándar y la muestral? y ¿Cómo se calcula esta función de probabilidad muestral? La respuesta a la primera pregunta es: No en términos de cálculo, salvo un pequeño ajuste que veremos en breve, no cambia en lo absoluto. Es más incluso usted podrá seguir utilizando las tablas de probabilidad normal estándar que empleó previamente. La segunda pregunta, relativa a la fórmula de cálculo se ve en el siguiente subtema.

⁷ Que ahora la llamaremos **error estándar** (en breve veremos el por qué del nombre),

⁸ Incluso en ocasiones, como es el caso de los precios de una acción o la temperatura de un lugar, usted no podrá conocer la población verdadera.



2.8.1 Las estadísticas necesarias para calcular la distribución normal muestral

Hasta ahora se ha revisado cómo se genera una función de probabilidad normal estándar y se ha hecho énfasis en observar que esta se revisó suponiendo que los datos con que se trabaja son **poblaciones**. Sin embargo, usted tendrá en su poder para trabajar, y salvo que el problema que usted resuelva sea diferente, muestras.

Se ha visto también que, para calcular probabilidades a través de una función de densidad, usted debe tener la media y la desviación estándar que son una medida de tendencia central y de dispersión respectivamente. Cuando usted trabaja con poblaciones, las medidas que son insumos necesarios para el cálculo de probabilidades se llaman **Parámetros**. Es decir, si los datos que usted tiene para analizar son la media y la desviación estándar. A estos dos se les denomina parámetros de su función de probabilidad.

Sin embargo, para una función de probabilidad cuando usted tiene muestras, los insumos son los mismos y se llaman ahora **estadísticas o medidas estadísticas**. Y estas estadísticas son la media y el error estándar (**recuerde que así le llamamos a la desviación estándar cuando tenemos datos de muestras**).

Para afirmar la idea, se tiene el siguiente cuadro de resumen que el profesor espera le sea de utilidad para memorizar:

| Tipo de medida | Parámetro (población) | Estadística (muestra) |
|-------------------|---------------------------------|--------------------------------------|
| Tendencia Central | Media poblacional μ | Media muestral σ |
| Dispersión | Desviación estándar σ | Error estándar $\sigma_{\bar{x}}$ |

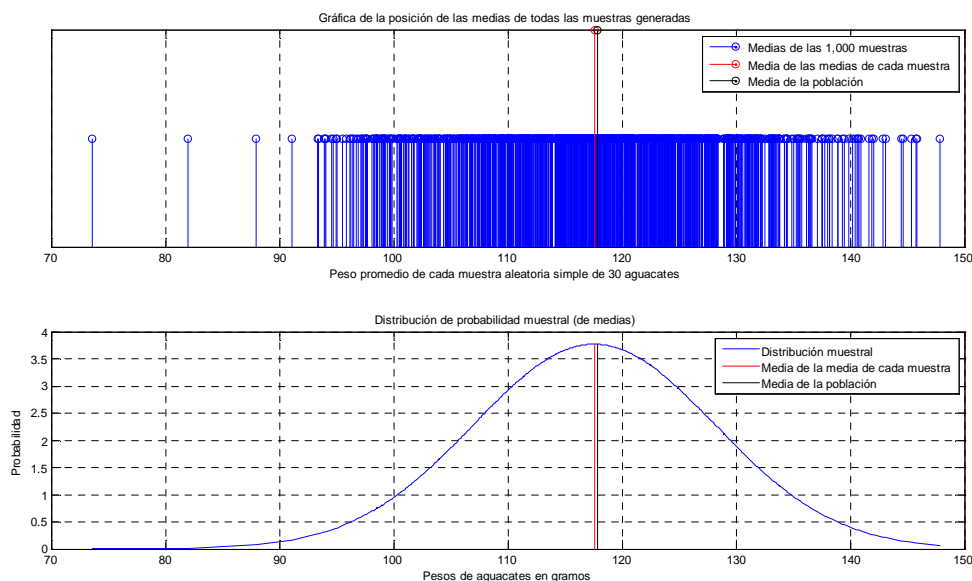
Gráfica 20 Parámetros (población) y estadísticas (muestra) empleados en el cálculo de probabilidades.

2.8.2 Media muestral

Como puede apreciar, la función de probabilidad normal estándar sigue utilizándose. Lo único que cambian son la forma de calcular la media y la desviación estándar. Para el caso de la media muestral, que ahora se denota como \bar{x} , simplemente se calcula igual para todos los datos repitiendo la fórmula 1:

**Fórmula 7 Cálculo de la media muestral:**

$$\bar{x} = \mu = \frac{\sum x_i}{n}$$

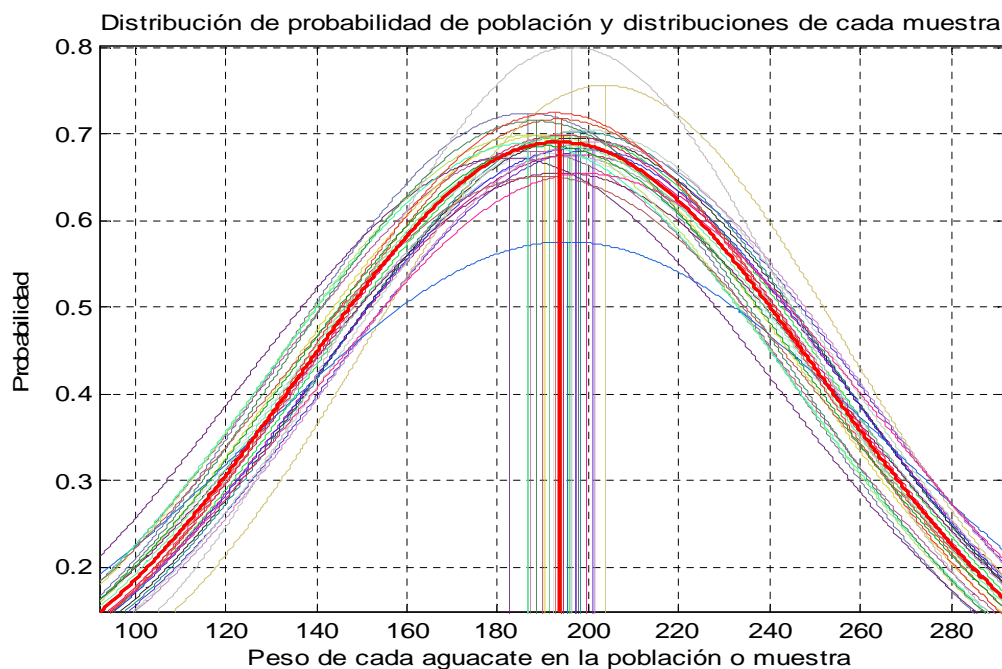


Para demostrar que la media poblacional puede aproximarse con la media de nuestra muestra observe usted de nuevo la parte superior e inferior de la gráfica 19 presentada de nuevo anteriormente. Note como, con 1,000 muestras, el promedio de cada una se acerca mucho (se aproxima) a la media poblacional. Por tanto, si se hubieran corrido 10,000 muestras en lugar de 1,000, la aproximación hubiera sido mayor al grado de que la media de nuestra muestra y la poblacional prácticamente serían la misma. Por tanto, con este simple ejercicio mental, puede usted comprobar que es válido emplear la media de la muestra calculándola con la media convencional.

Como una nota adicional, también es de importancia observar que esta aproximación se cumple incluso si se tienen múltiples muestras con diferente tamaño. Es decir, una muestra de 30 aguacates, otra de 200 y así sucesivamente.

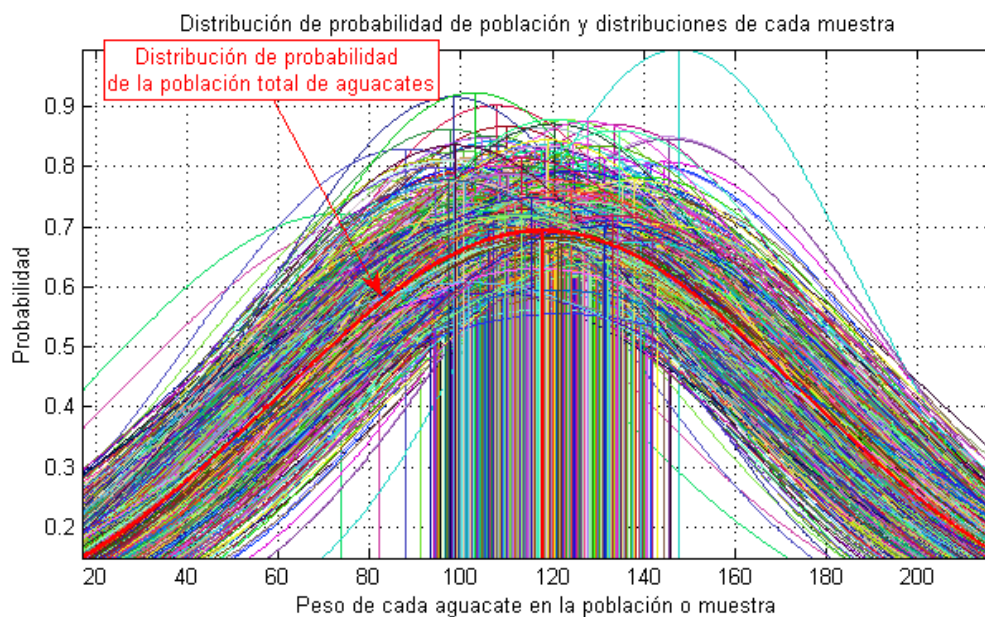
2.8.3 Error estándar

En el sub tema anterior se dijo que el tamaño de la muestra no influía en el cálculo de la media muestral y que sería la misma media para población que para muestra. Sin embargo, en el caso de la desviación estándar aplicable a una muestra, mejor conocida como **error estándar**, la cosa cambia. Para ilustrar la idea, observe primero la gráfica 21 en donde se generan 30 muestras diferentes con diferentes tamaños. Es decir, una muestra de 30 aguacates, otra de 55 y así sucesivamente.



Gráfica 21 Distribuciones de probabilidad de diferentes muestras con diferente tamaño.

Ahora veamos de nuevo la gráfica 18:



La diferencia entre la gráfica 21 y la 18 está en que la 21 se hicieron muestras aleatorias con diferente tamaño (en muchos casos más de 30) y en la 18 se hicieron muchas muestras con el mismo tamaño: 30 aguacates.



La idea que se busca resaltar de la gráfica 21 es que, conforme aumenta el tamaño de la muestra, el error estándar⁹ se aproxima mucho a la desviación estándar de la población. Por tanto, la forma en que se calcula la desviación estándar en una muestra, mejor conocida como **error estándar**, se da ahora por la siguiente fórmula sí y solo si se conoce la desviación estándar de toda la población:

Fórmula 8 Cálculo del error estándar:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n}}}{\sqrt{n}}$$

En un español más plano, lo que usted tiene que hacer para calcular el error estándar es simplemente calcular la desviación estándar (σ) de su muestra y luego dividirla entre la raíz cuadrada de su número de datos (\sqrt{n}).

Esta fórmula se utilizará, de momento, como válida ya que se presupone que se conoce la desviación estándar de la población. Sin embargo, no siempre es así y en el siguiente tema se verá cómo se hace para calcular la desviación estándar de una muestra.

Como su nombre lo dice, el error estándar es el error o separación que la media de su muestra tiene respecto a la verdadera media de la población y este se aproxima muy bien al simplemente dividir la desviación estándar de su muestra entre la raíz cuadrada del número de observaciones que integran su muestra.

Error estándar: separación que la media de su muestra tiene respecto a la verdadera media de la población y este se aproxima muy bien al simplemente dividir la desviación estándar de su muestra entre la raíz cuadrada del número de observaciones que integran su muestra.

Para cerrar el tema de la distribución de probabilidad muestral, es decir la distribución normal estándar aplicada a muestras, el profesor completa la tabla 20 y le presenta los cálculos necesarios para que usted pueda recordarlos y armar un mapa mental con ellos:

⁹ Que, como vimos, es el nombre que se le da a la desviación estándar cuando se trabaja con muestras.



| Tipo de medida | Parámetro (población) | Cálculo | Estadística (muestra) | Cálculo |
|-------------------|----------------------------------|--------------------------------------|--------------------------------|---|
| Tendencia Central | Media poblacional μ | $\mu = \frac{\sum x_i}{n}$ | Media muestral σ | $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$ |
| Dispersión | Desviación estándar \bar{x} | $\bar{x} = \mu = \frac{\sum x_i}{n}$ | Error estándar σ_x^- | $\sigma_x^- = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n}}}{\sqrt{n}}$ |

Tabla 10 Parámetros (población) y estadísticas (muestra) empleados en las funciones de probabilidad y su método de cálculo.

2.8.4 Cálculo de probabilidades con muestras.

Para calcular la probabilidad en una muestra se sigue utilizando la misma tabla de distribución normal estándar y se siguen los mismos métodos de cálculo previamente vistos. Lo único que cambia es la fórmula 6 a la que se le sustituye la desviación estándar por el error estándar. Esto es:

$$Z_i = \frac{x_i - \mu}{\sigma} \rightarrow Z_i = \frac{x_i - \mu}{\sigma_x^-}$$

Con lo anterior se establece entonces el cálculo de valores Z como sigue:

Fórmula 9 Cálculo del valor Z para estandarizar variables en muestras:

$$Z_i = \frac{x_i - \mu}{\sigma_x^-}$$

Dicho lo anterior usted puede seguir los pasos para calcular probabilidades que se le presentan en la tabla 11 ya sea cuando emplea datos de una población o de una muestra. Como aprecia, solo cambia el cálculo del error estándar.



| Población | Muestra |
|--|---|
| Calcular la media | Calcule la media de su muestra |
| Calcular la desviación estándar | Calcular la desviación estándar de su muestra |
| | Obtenga el error estándar dividiendo la desviación estándar entre la raíz cuadrada de "n" o número de observaciones |
| Estandarize la variable aleatoria a la que quiere calcularle la probabilidad | Estandarize la variable aleatoria a la que quiere calcularle la probabilidad |
| Busque en tablas el valor Z logrado anteriormente | Busque en tablas el valor Z logrado anteriormente |

Tabla 11 Pasos a seguir para el cálculo de probabilidades ya sea en poblaciones o en muestras.

Para ilustrar todo esto con un ejemplo, piense que usted es el comerciante de aguacates de Morelia y que toma una muestra de 30 piezas que le lleva a un peso promedio de 150 g. y una desviación estándar de los datos de 80 g. Usted quiere determinar la probabilidad de que, si toma otro aguacate más, este tenga un peso menor o igual a 180 grs.

Para responder esto, recuerde usted que la probabilidad de que el peso de los aguacates sea menor o igual a su peso promedio es de 50%:

$$p(x_i \leq \mu) = 50\%$$

Ahora que se tiene esto, solo debe determinarse la probabilidad de que el aguacate pese entre el valor del promedio (150 grs.) y 180 grs. Lo que debe usted hacer primero es estandarizar el valor de $x_i = 180 \text{ grs.}$ y para ello debe, con la desviación estándar que calculó, obtener el **error estándar**:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{30}} = \frac{80}{1.7320} = 46.1880$$

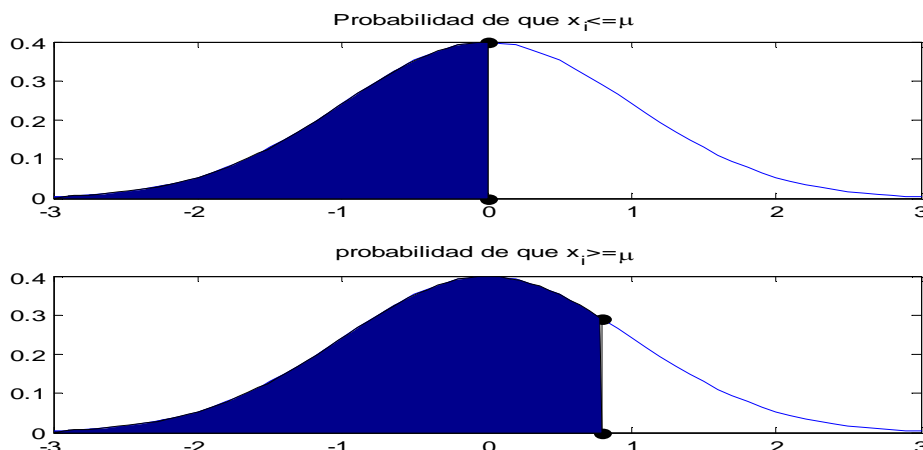
Ya que se tiene el error estándar se obtiene el valor Z_i :

$$Z_i = \frac{x_i - \mu}{\sigma_x} = \frac{180 - 150}{46.1880} = \frac{30}{46.1880} = 0.64519$$



Ya que se tiene este valor Z_i , se busca la probabilidad de que el aguacate pese entre el valor de la media (150 grs.) y 180 grs. Esto lleva a una probabilidad de

$$p(\mu \leq x_i \leq 180 \text{ grs}) = p(Z_i) = p(0.6451) = 23.89\%$$



$$\begin{aligned} p(\mu \leq x_i \leq 180 \text{ grs}) &= p(x_i \leq \mu) + p(\mu \leq x_i \leq 180 \text{ grs}) \\ &= 50\% + 23.89\% = 73.89\% \end{aligned}$$

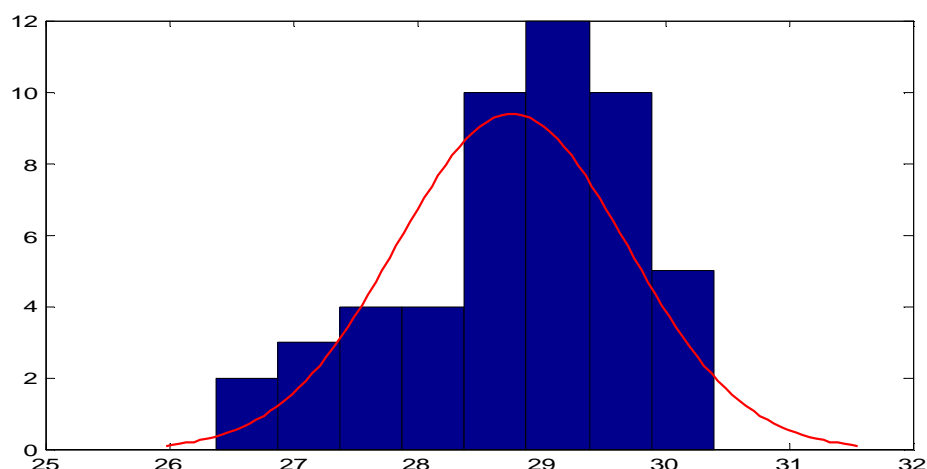
Entonces, ya que se tiene la probabilidad de que todos los pesos de aguacates posibles sean menores o iguales a la media y, posteriormente que el peso del aguacate seleccionado se encuentre entre la media de 150 grs. Y 180 grs. Se suman las dos probabilidades calculadas y se logra la probabilidad de que el siguiente aguacate seleccionado tenga un peso menor o igual a 180 g. La explicación gráfica se dio anteriormente.

Como se aprecia, el método de cálculo de probabilidades con muestras sigue siendo el mismo. Lo único que cambia es que se calcula el **error estándar** y en lugar de la desviación estándar.

2.9 El teorema del límite central y una primera forma de determinar el tamaño adecuado de la muestra

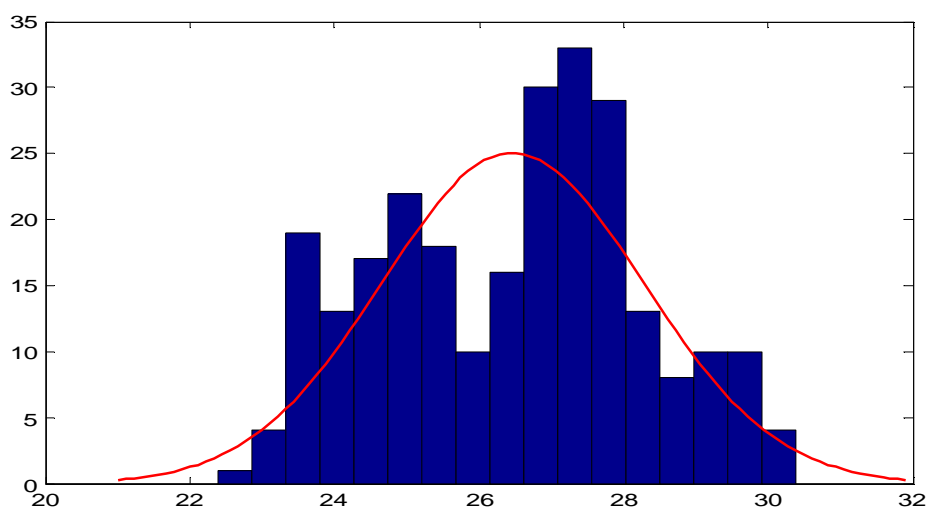
Hasta ahora se ha trabajado con el supuesto de que las variables aleatorias que se estudian están normalmente distribuidas. Sin embargo puede darse el caso de que esto no sea así. Cuando usted, con técnicas de las que se revisarán algunas en temas posteriores, detecta que los datos con que trabaja no están normalmente distribuidos, puede seguir manejando el supuesto de normalidad si incrementa el número de datos de su muestra.

Por ejemplo, Suponga usted que tiene una muestra de los 50 precios más recientes de una acción que cotiza en bolsa. Si usted graficara la distribución de probabilidad, tendría algo como esto:



Gráfica 22 Distribución de probabilidad de una muestra de 50 precios de una acción que cotiza en bolsa.

Claramente, por el histograma, se puede observar que, al compararlo con una distribución de probabilidad normal, su muestra no está normalmente distribuida. ¿Qué pasa ahora si usted incrementa la muestra de 50 a 250?



Gráfica 23 Distribución de probabilidad de una muestra de 250 precios de una acción que cotiza en bolsa.

Bien es cierto que el histograma sugiere que la distribución de probabilidad es todo menos “normal” o “gaussiana”. Sin embargo, el tipo de distribución de probabilidad se parece cada vez más a una normal conforme el número de datos se incrementa.

Entonces, el **Teorema del límite central** dice que debe incrementarse el número de datos para poder hacer que la muestra sea normal si esta no lo es. La siguiente pregunta que podría



plantearse usted sería ¿Qué tan grande debe ser mi muestra para darle validez al Teorema del Límite Central? La mayoría de los estadísticos sugiere que debe cumplirse una de las siguientes condiciones para considerar la muestra “grande”:

1. Que el número de observaciones de la muestra sea mayor de 30. Esto es: $n \geq 30$
2. O que el número de datos de su muestra tenga una proporción menor o igual al 5% de la población total.

Por ejemplo, si usted quiere determinar cuál es el tamaño apropiado de muestra para un análisis estadístico aplicado el precio de una acción o la temperatura de Morelia, usted deberá elegir 30 o más datos para considerar “grande” su muestra y darle validez al teorema del límite central. Casos como estos dos son situaciones en las que usted desconoce el verdadero tamaño de la población total.

En otras circunstancias usted podrá conocer o aproximar el tamaño total de su población. Por ejemplo la población de habitantes de México era de aproximadamente 110 millones de habitantes a finales del año 2011. Si usted fuese un director del INEGI y desea determinar el nivel de empleo de México, deberá encuestar, a lo mucho, a 5.5 millones de personas. Es decir, el 5% de 110 millones. Si usted entrevista a 7 millones, a pesar de ser muchos datos su muestra ya no se considerará grande. ¿Por qué la contradicción de pedir una muestra mayor de 30 o menor al 5% de la población conocida? Esa respuesta se la dejamos a los matemáticos. Sea suficiente para usted saber esto y aplicar la regla de dedo como se le presenta.

Teorema del límite central:

“Una muestra de datos que no tenga una distribución de probabilidad normal podrá suponerse que está normalmente distribuida si

- *su número de datos es mayor o igual a 30.*
- *si se conoce el tamaño total de la población (y estas es muy grande), su número de datos es menor o igual al 5% de la población total...”*

Por tanto, para regla general de usted y a reserva de ver otras más, si usted quiere determinar si su muestra está normalmente distribuida y no sabe cuántos datos debe tener como mínimo, recuerde el teorema del límite central revisado que le pide mínimo 30 datos. Por tanto, un tamaño adecuado de muestra, como regla general, es de 30 observaciones.

Habrà ocasiones en que usted no tenga capacidad material de obtener 30 datos sino menos. Para esos casos se utiliza una distribución t-Student. De momento, no la veremos para que usted sea capaz de asimilar estas ideas. Este tipo de distribución se revisará en el tema de comprobación de hipótesis.



2.10 El multiplicador de población finita

Para finalizar la revisión de la Teoría del muestreo que contempla este segundo tema del curso de Estadística II. Es de interés observar algo más del cálculo del error estándar. Líneas atrás se le observó que, cuando se trabaja con muestras, no se utiliza la desviación estándar sino el **error estándar** y que este se calcula simplemente al dividir la desviación estándar de su muestra entre la raíz cuadrada del número de observaciones en la misma:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n}}}{\sqrt{n}}$$

Este cálculo es el que casi siempre se utilizará. Esto es así porque muchos fenómenos que estudiamos en las Ciencias Administrativas tienen poblaciones cuyos tamaños desconocemos. Es decir, **son poblaciones infinitas**. Ejemplos de esto son el nivel de llenado de las botellas de agua que produce, la temperatura de Morelia, los precios de una acción, el inventario que tuvo, tiene y tendrá de aguacates, etc. Sin embargo, habrá casos en los que usted conozca muy bien el tamaño de su población y tenga que verse en la necesidad de hacer un muestreo dado lo costoso que le resulta sacar datos del total de su población. Un ejemplo puede ser un estudio de mercado como el que hizo Steve Jobs. Por ejemplo, él sabía cuántos arquitectos había en Estados Unidos. Por tanto, tuvo que hacer un ajuste adicional al error estándar para poder calcularlo bien y determinar la distribución de probabilidad:

Fórmula 10 el error estándar calculado con el multiplicador de la población finita:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{n-1}}$$

Esto es, al error estándar tuvo que multiplicarlo por el término conocido como multiplicador de población finita dado por:

$$\sqrt{\frac{N-n}{n-1}}$$

En donde N es el tamaño ya conocido de la población y n es el de la muestra que usted utiliza.

Cuando usted desee levantar una encuesta de ambiente laboral en su empresa, usted conoce de antemano el número de empleados en su nómina (N). Como usted posee o trabaja en una empresa con, digamos, 12 mil empleados, le resulta difícil procesar mucha información, por lo que



podría hacer un tipo de muestreo (digamos estratificado o aleatorio simple) y se acerca a 60 empleados diferentes a realizarle la encuesta.

Como puede ver, su muestra es igual al 5% del total de empleados por lo que es de tamaño grande. Al error estándar debería aplicarle, siempre que conozca el tamaño de la población, la siguiente multiplicación:

$$\sigma_x^- = \frac{\sigma}{\sqrt{60}} \times \sqrt{\frac{12,000 - 60}{60 - 1}}$$

NOTA: Si usted no conoce el tamaño total de la población o esta tiene un número infinito de datos (como la temperatura o el precio de una acción), no deberá aplicar el multiplicador de la población finita sino simplemente determinar el error estándar como sigue:

$$\sigma_x^- = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n}}}{\sqrt{n}}$$

Retomando el ejemplo de la encuesta a los empleados de su empresa, suponga que les hace una simple pregunta: Del 1 al 10, con 1 muy bajo y 10 muy alto, diga usted si se siente contento en esta empresa. Después de mandar por correo electrónico la pregunta a 60 personas, usted observa que la calificación media de felicidad en la empresa es de $\mu=7.8$ con una desviación estándar en la muestra de $\sigma=1.2$. Entonces, el error muestral de esta encuesta para la totalidad de su empresa sería de:

$$\sigma_x^- = \frac{1.2}{\sqrt{60}} \times \sqrt{\frac{12,000 - 60}{60 - 1}} = 2.20$$

Con este dato, deberá usted hacer los cálculos de valor Z de la fórmula 9 y calcular la probabilidad con tablas:

$$Z_i = \frac{x_i - \mu}{\sigma_x^-}$$

Ya que se conoció un poco sobre la Teoría del Muestreo y que se observó la forma de calcular las principales **estadísticas** para el cálculo de probabilidades muestrales, es de necesidad de pasar al tema de inferencia, de tal forma que podamos ya tener los fundamentos para comprobar hipótesis estadísticas.



3 Estimaciones puntuales y de intervalo. La base de la inferencia estadística.

En el tema anterior se introdujo el concepto de que la media muestral puede ser diferente respecto a la poblacional e incluso entre muestras. Eso llevó a observar como momentáneamente válida la aproximación de dicha media poblacional μ a través de la media de la muestra con que se trabaja \bar{x} . Sin embargo, sigue presente el efecto de la incertidumbre que se genera al ver fluctuar \bar{x} . Aquí es donde las estimaciones puntuales y de intervalo cobran vida como conceptos.

Durante su vida profesional, usted, para tomar decisiones, deberá siempre hacer estimaciones de qué sucederá en el futuro. Se vio al inicio de estas notas que usted siempre decidirá en un entorno de riesgo y uno de los riesgos a considerar es el hecho de que su muestra, si tiene datos de muestrales, es toda la información que tendrá para decidir.

Cuando usted estime qué sucederá en el futuro, puede hacer dos tipos de estimaciones:

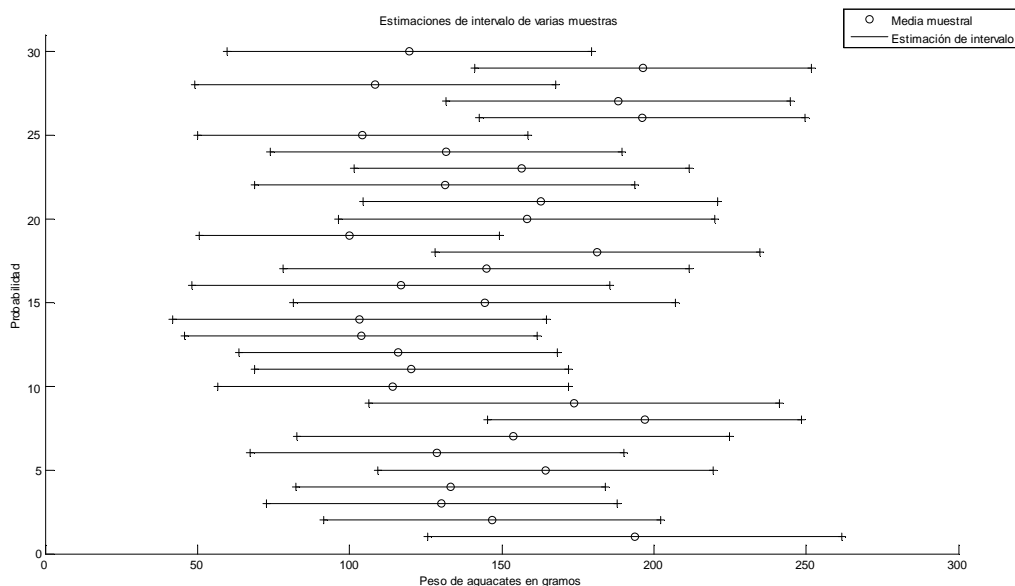
Estimaciones puntuales: Es un solo número que se utiliza para estimar un parámetro de la población: la media poblacional.

La estimación puntual consiste simplemente en decir qué sucederá exactamente según sus cálculos estadísticos. Por ejemplo, si usted dice que mañana a las 12:00 de la mañana la temperatura será de 20.15°, usted está haciendo una estimación puntual. Otro ejemplo será que usted diga que la siguiente botella de agua que tome de la línea de producción, tendrá un llenado de 997.83 l. o que el precio de una acción será de \$23.57 el día de mañana. La cualidad que tienen estas estimaciones es que son números exactos, números puntuales.

La forma más común de obtener este tipo de estimación puntual es simplemente utilizar la media muestral \bar{x} . Sin embargo, el utilizar \bar{x} está sujeto a cometer errores de muestra por lo que es más apropiado hacer afirmaciones o estimaciones del tipo: El día de mañana a las 12:00 la temperatura podría estar entre 20° y 21° o el nivel de llenado de la botella puede estar entre 996 l y 997l. Este tipo de afirmación se conoce como **estimación de intervalo**.

Estimación de intervalo: Es un rango de valores que se utiliza para estimar un parámetro de la población.

Para dar una idea más clara de una estimación puntual y de una de intervalo se retoma el ejemplo del inventario de aguacates del comerciante de Morelia, se generan 30 muestras aleatorias con 30 aguacates cada una y se llega a la siguiente gráfica:

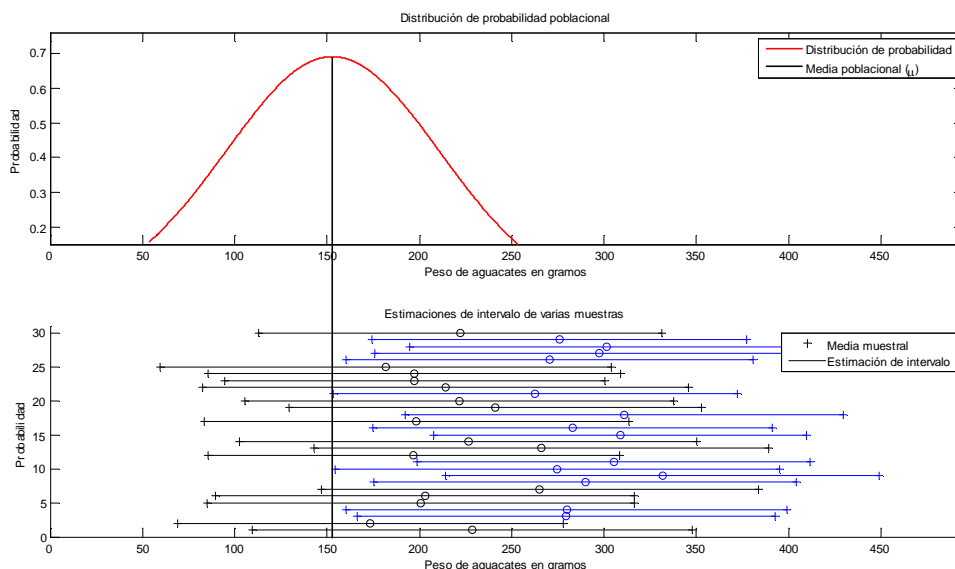


Gráfica 24 Estimaciones puntuales (media muestral representadas con círculos) y de intervalo de 30 muestras diferentes con 30 aguacates para el inventario total (población) del comerciante.

En la misma se presentan las estimaciones puntuales de 30 muestras diferentes que tienen 30 aguacates cada una. Dicha estimación puntual se logra con la media muestral \bar{x} y se marca con un círculo. La estimación de intervalo tiene dos valores. Uno superior (el extremo derecho de cada círculo o media muestral) que se logra de la siguiente forma: $\lim.\sup erior = \bar{x} + \sigma_{\bar{x}}$ y otro inferior que se logra con la siguiente resta: $\lim.\inf erior = \bar{x} - \sigma_{\bar{x}}$. Por ejemplo, en la primera muestra de la gráfica (la de hasta arriba) se tiene una estimación puntual o media muestral de $\bar{x} = 119.6$, un límite superior de $\lim.\sup erior = \bar{x} + \sigma_{\bar{x}} = 179.5$ y uno inferior de $\lim.\inf erior = \bar{x} - \sigma_{\bar{x}} = 59.67$.

3.1 Consideraciones para calcular verdaderas estimaciones de intervalo

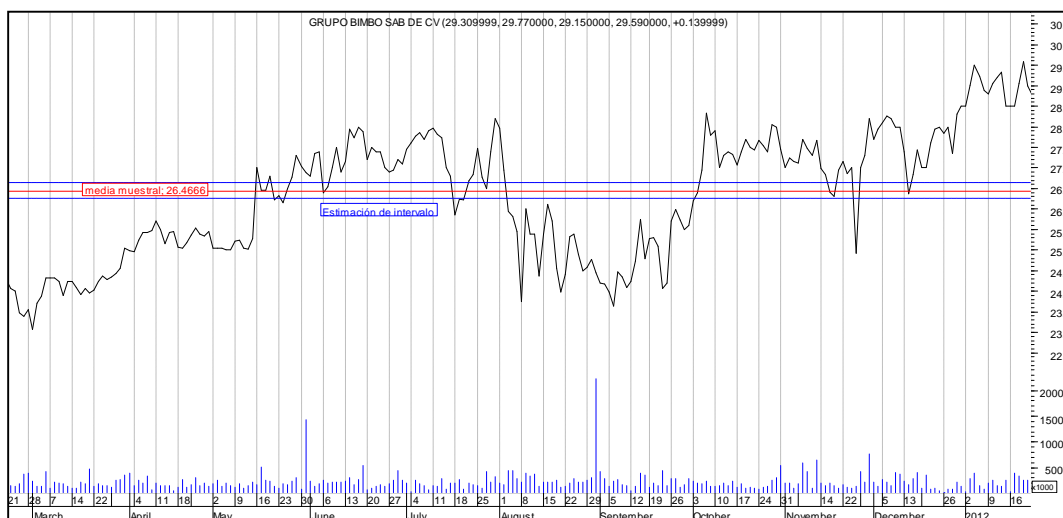
En la gráfica 24 se expusieron 30 muestras diferentes las cuales tienen diferentes medias muestrales \bar{x} y diferentes intervalos dados por $\lim.\sup erior = \bar{x} + \sigma_{\bar{x}}$ y $\lim.\inf erior = \bar{x} - \sigma_{\bar{x}}$. Si se recuerda que la media muestral \bar{x} puede fluctuar respecto a la poblacional μ , se aprecia en la siguiente gráfica en la que se exponen las 30 muestras aleatorias en comparación a la media poblacional.



Gráfica 25 Comparativo de las estimaciones de intervalo con la media poblacional y su distribución de probabilidad

En la misma se marcaron, en un color más claro (azul si está viendo en la computadora estas notas), aquellos intervalos cuyo rango de valores (límite inferior a límite superior) no contiene la media poblacional. Es decir, están sesgados respecto a dicha media.

En ejercicios anteriores, se ha trabajado con estimaciones de intervalo. Para darse una idea, suponga usted que la muestra que tiene de un año del precio de una acción se da por el siguiente comportamiento:



Gráfica 26 histórico de un año del precio de BIMBOA en la BMV. Fuente: Reuters metastock con Esignal.



Si usted quisiera hacer una estimación puntual del precio para el día siguiente en dicha acción podría calcular la media muestral¹⁰ que es de $\bar{x} = 26.4666$. Sin embargo, usted sabe que esta media no es la misma que la de la población ya que podría fluctuar de muestra en muestra. Por tanto se podría calcular el error estándar de la media muestral

3.1.1 El verdadero cálculo del error muestral cuando se desconoce la desviación estándar de la población.

Sin embargo, algo que se mencionó al inicio de este tema es que se está suponiendo que se conoce la desviación estándar de la población y en realidad lo que se está calculando la de una muestra. En el caso de muestras, lo que debe de hacerse es hacer un pequeño ajuste para calcular la desviación estándar muestral que ahora se denota como s :

Fórmula 11 Cálculo del error muestral cuando se desconoce la desviación estándar de la población:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n-1}}}{\sqrt{n}}$$

Nótese que simplemente, en lugar de dividir entre n , ahora se hace entre $n-1$. Entonces el cálculo del error muestral se hizo de esta forma:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.8164}{\sqrt{258}} = 0.1333$$

3.1.2 La estimación de intervalo.

Ya que tiene usted la estimación puntual (\bar{x}) del precio de la acción, que reconoce que este valor puede cambiar de muestra en muestra y que tiene el cálculo del error estándar de la muestra calculado con la fórmula 11, procederá usted a hacer una afirmación de este tipo: ***“El precio de la acción se estima que sea de \$26.4666 y, con un 95% de confianza, se espera que ese valor oscile entre \$26.6533 y \$28.2802.”*** Si usted observa la gráfica 26, quizá no le sea muy preciso el pronóstico en el sentido de que el precio esperado y su intervalo están muy abajo. Con el análisis de regresión podremos mejorar la precisión. Baste con suponer, de momento, que la media muestral es buen pronóstico del valor futuro.

Un ejemplo que puede ser más apropiado podría ser el del comerciante de aguacates. De una muestra de 30 aguacates, usted puede llegar a una media muestral de $\bar{x} = 153.0547g$ y un error

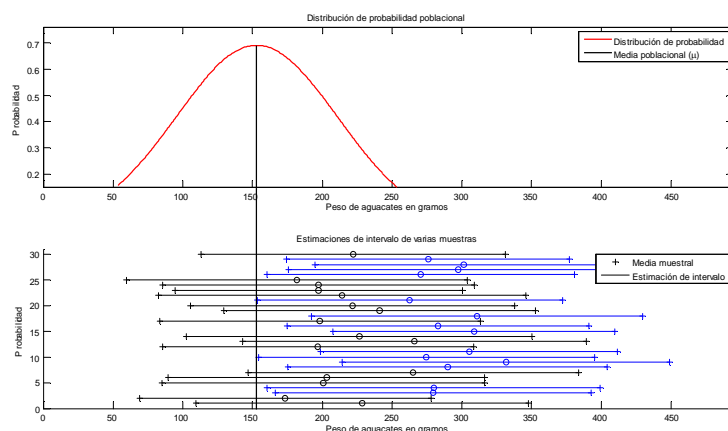
¹⁰ En el tema de regresión se le enseñará a hacer mejores pronósticos.



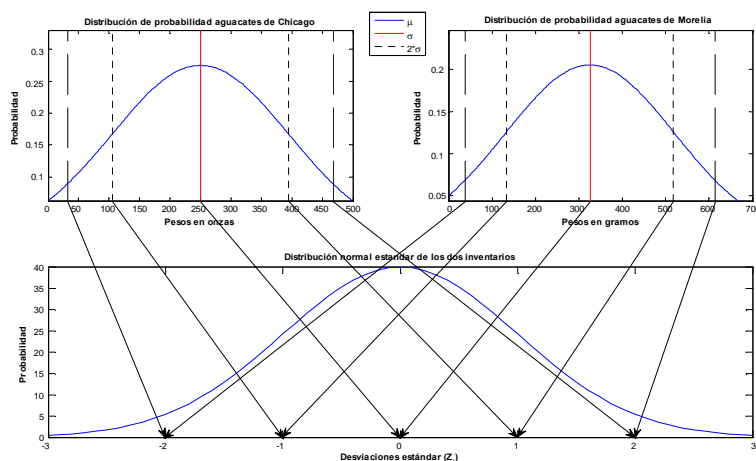
estándar calculado con la fórmula 11 de $\sigma_x = 57.6770g$. Con estos datos podría decir que su estimación del peso de un solo aguacate que tome de manera aleatoria podría ser de 153.0547 g. y que esta estimación podría, con un 95% de probabilidad, variar entre 58.1844 g y 247.9249 g.

¿Cómo se hicieron este tipo de estimaciones?, ¿Cómo fue que se llegó a ese 95% de probabilidad para definir un nivel de confianza (como el de 95% que se dio en la estimación) es necesario observar de nuevo la gráfica 25 y recordar la

Gráfica 25:



Gráfica 10:



Recuerde usted que, cuando se estandarizó la muestra con:

$$Z_i = \frac{x_i - \mu}{\sigma_x}$$



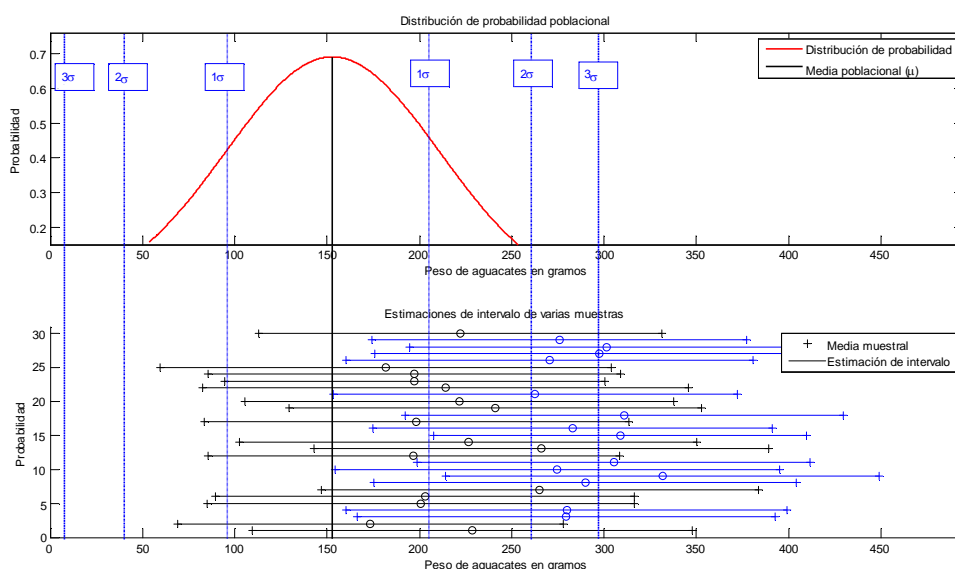
La desviación estándar se convirtió de tener su valor original (para el ejemplo de la muestra del aguacatero de 57.677g.) a uno de $Z=1$ (véase la gráfica 1). Si se grafica la desviación estándar de 57.677 y se suma a la media muestral directamente para obtener el intervalo superior de:

$$\text{int. superior} = 153.0547 + 57.677g = 210.7317g$$

Y uno inferior de

$$\text{int. inferior} = 153.0547 - 57.677g = 95.377g$$

Estos dos límites o intervalos equivales a **“una desviación estándar”** o **“un error estándar”** si se quiere ver como muestra y se ponen en la gráfica 25 para llegar a la 26:



En la misma se ponen 3 tipos de línea punteada. La primera de ella marca tanto el intervalo superior (a la derecha de la media) como el inferior (a la izquierda) calculados previamente con un valor de 210.7137 g y 95.377g respectivamente. Si se le calcula a ambos el valor Z se tendrán los siguientes valores:

$$Z_i = \frac{210.7137g - 153.0547}{57.677} = 1$$

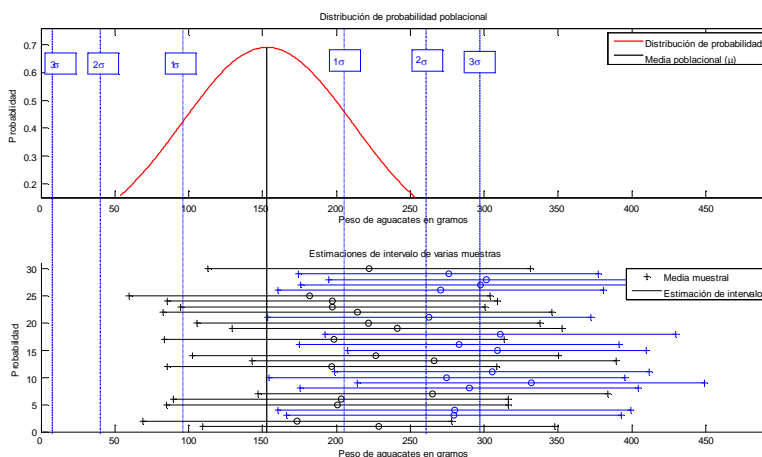
$$Z_i = \frac{95.377g - 153.0547}{57.677} = -1$$

Es decir, en términos de la desviación normal estándar. El error estándar de la muestra representa, reiterando para la probabilidad normal estándar, una sola desviación estándar. Es decir $Z=1$.



Para seguir ahora con la idea de la probabilidad que se le pone a la afirmación ***“El siguiente aguacate que se tome aleatoriamente podría pesar 153.0547 y esta estimación podría tener, con 95% de confianza, una fluctuación de 58.1844 g y 247.9249”***, obsérvese la gráfica 26. Note cómo no todos los intervalos de confianza de las 35 muestras generadas aleatoriamente tienen la media poblacional como parte de sus valores. Si se hicieran más muestras aleatorias, digamos unas 1,000 de 30 aguacates se tendría que el porcentaje de muestras que contendrían a μ sería de solo el 66%. ¿Cómo se obtuvo esta probabilidad? Recuerde usted el cálculo de probabilidades con valores Z. Nosotros ya tenemos dos valores Z correspondientes a $\lim.\text{superior} = \bar{x} + \sigma_{\bar{x}}$ y $\lim.\text{inferior} = \bar{x} - \sigma_{\bar{x}}$. Que son $Z_{\max} = 1$ y $Z_i = -1$ respectivamente. Por favor, ahora recordemos el valor en tabla de probabilidades cuánto vale cada uno.

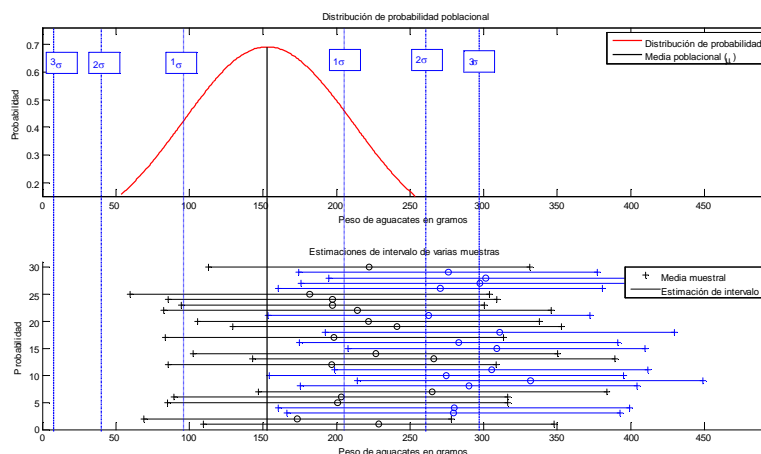
Esto lleva a valores de probabilidad de $p(Z_{\max}) = 34.13447\%$ y $P(Z_{\min}) = 34.1344\%$ respectivamente. Si se suman, se observa que se llega a probabilidad de 68.2689% que es la que representa el grado de confianza de nuestras estimaciones. Con esta probabilidad que, de momento interpretamos como nivel de confianza, se esperaría que todas las diferentes muestras contengan, dentro de los valores de sus intervalos de confianza, a μ o al menos $\mu \pm \sigma_{\bar{x}}$. Para darse una idea de si es cierto esto o no, observe de nuevo la gráfica 26:



Las estimaciones de intervalo que no tienen μ como parte de sus valores se señalan en color claro (azul si usted imprimió a color estas notas o las ve en la computadora). Aquellos intervalos que contienen a μ o al menos el valor de $\mu \pm \sigma_{\bar{x}}$ tienen valores que se encuentran dentro del área definida por las dos líneas marcadas con 1σ (una desviación estándar o error muestral alrededor de la media –poblacional o muestral-). La cantidad de muestras que contienen ya sea a μ o valores dados por $\mu \pm \sigma_{\bar{x}}$, es alrededor de 68.2689% del total de muestras generadas.



¿Qué pasa ahora si, en lugar de sumar o restar el error estándar ($\mu \pm \sigma$), sumamos y restamos este valor multiplicado por dos ($\mu \pm (2 \times \sigma)$)? Veamos qué sucedió en la gráfica 26:



Usted se volvió más tolerante, está permitiendo que algunas muestras que no contienen μ por lo menos tengan oportunidad de contener $\mu \pm (2 \times \sigma)$. Esto implica perder precisión en la estimación puntual pero incrementar la probabilidad de que fluctúe \bar{x} en más valores. Si usted calcula los valores Z de $\mu \pm (2 \times \sigma)$ llegará a ver que son $p(Z_{\max}) = 47.7249\%$ y $p(Z_{\min}) = 47.7249\%$ que, cuando se suman, llevan a una probabilidad de 95.4499%.

Ahora qué probabilidad se tiene cuando se es más tolerante digamos a $\mu \pm (3 \times \sigma)$. Si usted calcula la probabilidad de valores de $Z=3$ y $Z=-3$, llegará a valores de $p(Z_{\max}) = 49.8650\%$ y $p(Z_{\min}) = 49.8650\%$ que sumados llevan a una probabilidad total de 99.7302%. Es decir, está tomando como confiables las fluctuaciones del 99.7302% de todas las posibles muestras que se pueden generar de manera aleatoria.

Ahora, para generar ese “grado de confianza” cómo se determinó. En la vida real, salvo que usted esté en áreas de trabajo como es la administración de riesgos de un banco o casa de bolsa, en el departamento de calidad de una fábrica o trabajos afines, usted no piensa lo siguiente: “Voy a hacer estimaciones de intervalo con ± 2 desviaciones estándar. Usted como futuro empresari@, contador@, director@ de empresa u organismo social y/o gubernamental piensa más bien como **“Quiero saber cuál va a ser el valor promedio del inventario de aguacates que comercializo con un (a manera de ejemplo) 95% de confianza en mis estimaciones”**.

¿Qué pasos debe usted seguir para lograr esto? Simple, lo hace usted al revés con la probabilidad.

¿Cómo es eso? Siga esta simple receta de cocina o receta de dedo: 🍪

1. Saque su muestra. ¿Cómo? Utilice el método, de los previamente vistos (aleatorio simple, sistematizado, estratificado o de racimo), que mejor se acomode a sus objetivos.



2. Defina el tamaño de muestra de tal forma que esté distribuida normal o gaussianamente o al menos aproxime esta función de probabilidad. Si se trata de datos infinitos de (propios de variables aleatorias continuas) como son la temperatura o el precio de una acción, simplemente apele al teorema del límite central y defina una muestra mayor o igual a 30.
3. De los datos que tiene calcule la media muestral:

$$\bar{x} = \mu = \frac{\sum x_i}{n}$$

4. Con esta media muestral calcule el error estándar. Si, como el caso de la temperatura o el precio de una acción, no conoce el tamaño total de la población de datos, emplee la siguiente fórmula (la 11):

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n-1}}}{\sqrt{n}}$$

Si conoce el tamaño total de la población (como en el caso del censo de población y vivienda), el cual se denota con N, puede utilizar la siguiente fórmula (la 10):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{n-1}}$$

5. Determine el nivel o intervalo de confianza que desea darle a sus estimaciones. Para fines del ejemplo que llevamos, piense en un 95% (usted puede elegir el que quiera de 0% a 100%).
6. Ahora haga la operación inversa en las tablas de probabilidad normal estándar. Primero reste a 95% el 50% ya que usted busca un solo valor Z a la derecha de la media. Es decir arriba de $Z_i = 0$ en la tabla. Esto le llevará a 45%.
7. Ahora busque en la tabla ¿qué valor tiene una probabilidad de 45%?:



Para las instrucciones de uso, por favor consulte la liga:

<http://www.droscardelatorre.com/classmat/UMSNH/FCCA/ESTADISTICAII/tutorialtabla.html>

| Z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|------|-----------|-----------|-----------|-----------|-----------|-----------|
| 0.00 | 0.5000000 | 0.0039894 | 0.0079783 | 0.0119665 | 0.0159534 | 0.0199388 |
| 0.10 | 0.0398278 | 0.0437953 | 0.0477584 | 0.0517168 | 0.0556700 | 0.0596177 |
| 0.20 | 0.0792597 | 0.0831662 | 0.0870644 | 0.0909541 | 0.0948349 | 0.0987063 |
| 0.30 | 0.1179114 | 0.1217195 | 0.1255158 | 0.1293000 | 0.1330717 | 0.1368307 |
| 0.40 | 0.1554217 | 0.1590970 | 0.1627573 | 0.1664022 | 0.1700314 | 0.1736448 |
| 0.50 | 0.1914625 | 0.1949743 | 0.1984682 | 0.2019440 | 0.2054015 | 0.2088403 |
| 0.60 | 0.2257469 | 0.2290691 | 0.2323711 | 0.2356527 | 0.2389137 | 0.2421539 |
| 0.70 | 0.2580363 | 0.2611479 | 0.2642375 | 0.2673049 | 0.2703500 | 0.2733726 |
| 0.80 | 0.2881446 | 0.2910299 | 0.2938919 | 0.2967306 | 0.2995458 | 0.3023375 |
| 0.90 | 0.3159399 | 0.3185887 | 0.3212136 | 0.3238145 | 0.3263912 | 0.3289439 |
| 1.00 | 0.3413447 | 0.3437524 | 0.3461358 | 0.3484950 | 0.3508300 | 0.3531409 |
| 1.10 | 0.3643339 | 0.3665005 | 0.3686431 | 0.3707619 | 0.3728568 | 0.3749281 |
| 1.20 | 0.3849303 | 0.3868606 | 0.3887676 | 0.3906514 | 0.3925123 | 0.3943502 |
| 1.30 | 0.4031995 | 0.4049021 | 0.4065825 | 0.4082409 | 0.4098773 | 0.4114920 |
| 1.40 | 0.4192433 | 0.4207302 | 0.4221962 | 0.4236415 | 0.4250663 | 0.4264707 |
| 1.50 | 0.4331928 | 0.4344783 | 0.4357445 | 0.4369916 | 0.4382198 | 0.4394292 |
| 1.60 | 0.4452007 | 0.4463011 | 0.4473839 | 0.4484493 | 0.4494974 | 0.4505285 |
| 1.70 | 0.4554345 | 0.4563671 | 0.4572838 | 0.4581849 | 0.4590705 | 0.4599408 |
| 1.80 | 0.4640697 | 0.4648521 | 0.4656205 | 0.4663750 | 0.4671159 | 0.4678432 |

Como puede apreciar en la tabla que se le dio, la probabilidad que más aproxima el 45% es la dada por la fila 1.6 y la columna 0.04. Es decir, esta probabilidad tiene un valor Z de 1.64. por tanto, ya tiene ahora el valor Z del intervalo o nivel de confianza que desea usted poner: $Z_i = 1.64$

8. Ahora que tiene \bar{x} , σ_x^- y Z_i , simplemente calcula sus límites superior e inferior donde cree que fluctuará, con ese 95% de confianza, su estimación puntual dada por \bar{x} :

$$\text{Int.superior} = \bar{x} + (Z_i \times \sigma_x^-) = 153.0547g + (1.64 \times 57.677g) = 247.6449g$$

$$\text{int.inferior} = \bar{x} - (Z_i \times \sigma_x^-) = 153.0547g - (1.64 \times 57.677g) = 54.4644g$$

Fórmula 1 Cálculo de las estimaciones de intervalo para muestra grande

$$\text{Límite o intervalo superior} = \bar{x} + (Z_i \times \sigma_x^-)$$

$$\text{Estimación puntual} = \bar{x}$$

$$\text{Límite o intervalo inferior} = \bar{x} - (Z_i \times \sigma_x^-)$$

9. Ahora sí puede ya hacer la afirmación que busca: **“Para el siguiente mes se esperaría que el peso promedio del inventario de aguacates de mi empresa sea de 153.0547g y que este fluctúe, con un 95% de confianza, entre 54.4644g y 247.6449g”**.

¿Fácil no?



3.2 ¿Qué pasa cuando nuestra muestra de datos no es grande? La distribución t-Student

Hasta ahora se ha trabajado con el supuesto de que los datos (sean de población o de muestra) están normalmente distribuidos ya sea porque así nos conviene o porque hemos trabajado con muestras con más de 30 datos, situación que satisface el Teorema del Límite Central previamente revisado.

Sin embargo, no siempre se tiene la posibilidad de tener muestras de 30 datos sino más pequeñas. Un ejemplo muy claro puede estar en la contabilidad de una empresa. Suponga que usted desea hacer un análisis estadístico y calcular la distribución de probabilidad del ROI¹¹ y que solo tiene 12 trimestres de información. Claramente la distribución normal estándar no es de utilidad porque viola el Teorema del límite central. ¿Qué se hace entonces? ¿Qué función de probabilidad se puede utilizar?

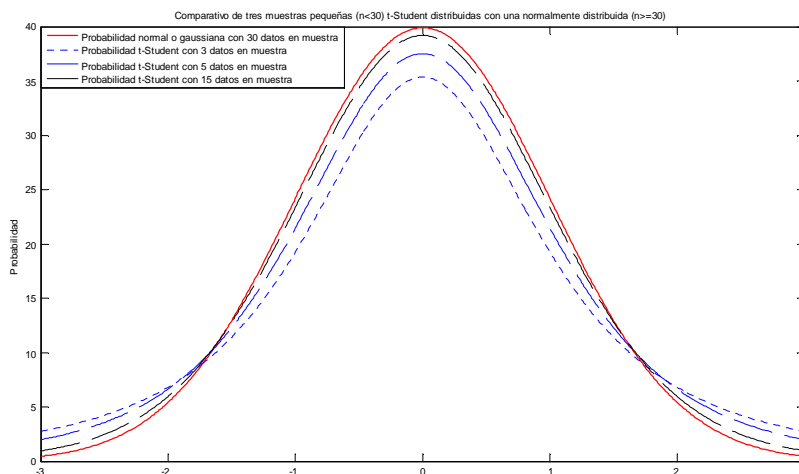
Muy simple: Hay un tipo de función de probabilidad, de los cuatro que revisaremos en el curso, que sirve para este fin. Esta se llama **distribución t-Student** o simplemente **distribución t**.

Esta distribución fue propuesta por W.S. Gosset quien era un trabajador de la cervecería Guinness en Dublín. El hombre era un aficionado a la Estadística y, como la cervecería prohibía a sus empleados hacer publicaciones científicas y académicas, Gosset utilizó el pseudónimo de "Student" para poder publicar su artículo y hacer su gran aportación a la Estadística.

Antes de hablar de los tres parámetros (uno más respecto a la normal) que se necesitan para calcular la distribución t-Student imagine usted que tiene tres muestras con la misma media:

1. Una con más de 30 datos a la que le podemos calcular la probabilidad normal.
2. Una con solo 5 datos que se le calcula una función de probabilidad t-Student.
3. Una con 20 datos que también se le calcula una función de probabilidad t-Student.

¹¹ Recuerde que el ROI es la rentabilidad del capital que se tiene invertido en la empresa:
 $ROI = \text{Utilidad neta} / \text{capital contable}.$



Gráfica 27 Comparativo de una muestra considerada “grande” y que está normalmente distribuida con 3 muestras consideradas “pequeñas” que están t-Student distribuidas.

Estas tres muestras se presentan en la grafica 27. Note usted cómo la muestra más pequeña (la de 3 observaciones) tiene una distribución t-Student cuya forma es muy parecida a la normal. Conforme se aumenta el número de observaciones en la muestra (de 3 a 5 y de 5 a 15) la forma de la distribución t-Student en cada caso se aproxima más a la normal o gaussiana. Esta situación es consistente con el Teorema del límite central y deja algo muy interesante para usted:

No siempre se tienen muestras con una cantidad de datos u observaciones mayor o igual a 30. Cuando esto sucede, los datos no están normalmente distribuidos pero se pueden hacer estimaciones de intervalo utilizando la distribución t-Student.

3.2.1 Los parámetros para calcular la distribución t-Student y su empleo para el cálculo de estimaciones de intervalo.

Se ha visto previamente que la distribución normal, a parte del valor de la variable aleatoria x_i , necesita solo dos simples **parámetros o estadísticas**¹² que son la media y la desviación estándar. Para el caso de la distribución t-Student se siguen utilizando estos dos más uno llamado Grados de libertad (denotado como GL o ν).

Este último (los grados de libertad) será el número de mayor importancia para calcular probabilidades.

¿Qué es esto de los grados de libertad?

¹² Recuerde que si sus datos son una muestra la media y la desviación estándar se llaman estadísticas y, si son de una población, se llaman parámetros.



Grados de libertad: Número de valores de una muestra que podemos especificar libremente, una vez que se sabe la media de la muestra.

Para dar una idea de los grados de libertad, suponga usted que tiene una muestra de solo dos datos que le lleva a un promedio o media muestral de 3.5:

$$\frac{a+b}{2} = 3.5$$

Si usted libremente elige el número 6, observará que el siguiente número necesario para llegar a un promedio de 3.5 es 6:

$$b = (3.5 \times 2) - a = (3.5 \times 2) - 1 = 6$$

De estos dos datos que conforman su muestra, uno de ellos lo especificó libremente y el otro es un valor forzado que debe cumplir con el promedio. Por tanto, la forma en que determinamos los grados de libertad ν aquí y en cualquier muestra de cualquier tamaño se daría por:

Función de densidad de probabilidad t-Student: Función de densidad de probabilidad que es la más utilizada y requiere de solo cuatro parámetros para su cálculo, el valor aleatorio (x_i) al que se le determinará la probabilidad, la media (μ), la desviación estándar (σ) y los grados de libertad. A diferencia de la normal o normal estándar, se emplea cuando nuestra muestra tiene menos de 30 datos (es muestra pequeña).

Fórmula 12 Determinación de los grados de libertad para la distribución t-Student:

$$\nu = n - 1$$

Es decir, si le restamos 1 (uno) al tamaño de nuestra muestra llegamos a los grados de libertad.

Cuando se tienen muestras pequeñas se puede realizar el cálculo de la estimación puntual y la de intervalo. Lo único que cambia en el cálculo de los intervalos es que no se utilizará un valor Z_i sino uno t. La forma de determinar los valores t es empleando las tablas de probabilidades y valores críticos que existen en la mayoría de los libros de Estadística y de las cuales puede bajar una, elaborada por el profesor, desde la siguiente liga:

www.drocardelatorre.com/classmat/UMSNH/FCCA/ESTADISTICAII/probtstudentexcel.html



Para calcular estimaciones puntuales y de intervalo suponga el siguiente caso del ejemplo del inventario de aguacates: El comerciante decide hacer una muestra más pequeña de solo 15 aguacates. Esto es así porque el resto del inventario (4,988 aguacates en total) ya los tiene empacados y listos para mandarlos a la central de abastos. Por tanto, le quedaron solo esos 12 aguacates. En base a estos quiere determinar ¿qué calidad, medida en peso, tendrá el siguiente inventario que le remita su proveedor? Para ello cuenta con los siguientes pesos en su muestra:

| Aguacate | Peso (g) | Aguacate | Peso (g) |
|----------|----------|----------|----------|
| 1 | 170.89 | 7 | 127 |
| 2 | 185.97 | 8 | 116.8 |
| 3 | 190.74 | 9 | 99.5 |
| 4 | 229.14 | 10 | 107.59 |
| 5 | 145.3 | 11 | 112.34 |
| 6 | 98 | 12 | 108.7654 |

| | |
|-------------------------|------------|
| Media muestral | 141.00295 |
| Desviación estándar (s) | 43.1767122 |
| Error estándar | 12.4640432 |

| | |
|------------------------|------------|
| Intervalo de confianza | 95% |
| Nivel de significancia | 5% |
| Grados de libertad | 11 |
| Valor t | 2.20098516 |

| Tipo de estimación | Fórmula | Estimación |
|--------------------|---|------------|
| Intervalo superior | $\bar{x} + (t \times \sigma_{\bar{x}})$ | 168.436124 |
| Estimación puntual | \bar{x} | 141.00295 |
| Intervalo inferior | $\bar{x} - (t \times \sigma_{\bar{x}})$ | 113.569776 |

De estos pesos, calcula la media muestral, la desviación estándar de su muestra con la fórmula 11 y el correspondiente error estándar con la fórmula 8.

Posteriormente, lo que se hace es determinar el valor t. Para ello, el comerciante decide dar un 95% de confianza a sus estimaciones. A diferencia del cálculo de intervalos con la probabilidad normal estándar en donde utilizamos una tabla de valores Z dada la probabilidad (tabla 2), se utilizará la tabla 4 correspondiente a los valores t dada la probabilidad. La diferencia aquí, respecto a la de valores Z (tabla 3), es que no se le da la probabilidad de suceso sino que tiene que buscar algo llamado **nivel de significancia** (denotado con una α) que no es más que el resultado de restar a 100% probabilidad el intervalo de confianza dado:

Fórmula 13 Cálculo del nivel de significancia dado un intervalo de confianza buscado en la estimación de intervalo:

$$\text{nivel de significancia} = \alpha = 100\% - (\% \text{ de intervalo de confianza})$$

En el caso del comerciante de aguacates que emplea un intervalo de confianza de 95%, se llega a un nivel de significancia de $\alpha=5\%$ o $\alpha=0.05$. Con los grados de libertad (11) y este nivel de significancia se busca el valor en tablas que es de 2.200985:



Ya que se tienen todos estos datos, el aguacatero puede hacer las siguientes estimaciones tanto puntuales como de intervalo:

Tablas de valores t relativos a su área de probabilidad.

| Grados de libertad/probabilidad crítica | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|-----------|-----------|-----------|-----------|-----------|
| 1 | 63.656741 | 31.820516 | 21.204949 | 15.894545 | 12.706205 |
| 2 | 9.924843 | 6.964557 | 5.642778 | 4.848732 | 4.302653 |
| 3 | 5.840909 | 4.540703 | 3.896046 | 3.481909 | 3.182446 |
| 4 | 4.604095 | 3.746947 | 3.297630 | 2.998528 | 2.776445 |
| 5 | 4.032143 | 3.364930 | 3.002875 | 2.756509 | 2.570582 |
| 6 | 3.707428 | 3.142668 | 2.828928 | 2.612242 | 2.446912 |
| 7 | 3.499483 | 2.997952 | 2.714573 | 2.516752 | 2.364624 |
| 8 | 3.355387 | 2.896459 | 2.633814 | 2.448985 | 2.306004 |
| 9 | 3.249836 | 2.821438 | 2.573804 | 2.398441 | 2.262157 |
| 10 | 3.169273 | 2.763769 | 2.527484 | 2.359315 | 2.228139 |
| 11 | 3.105807 | 2.718079 | 2.490664 | 2.328140 | 2.200985 |
| 12 | 3.054540 | 2.680998 | 2.460700 | 2.302722 | 2.178813 |
| 13 | 3.012276 | 2.650309 | 2.435845 | 2.281604 | 2.160369 |
| 14 | 2.976843 | 2.624494 | 2.414898 | 2.263781 | 2.144787 |
| 15 | 2.946713 | 2.602480 | 2.397005 | 2.248540 | 2.131450 |
| 16 | 2.920782 | 2.583487 | 2.381545 | 2.235358 | 2.119905 |
| 17 | 2.898231 | 2.566934 | 2.368055 | 2.223845 | 2.109816 |

Ilustración 2 Selección de valor t dada una significancia de 5% (95% de intervalo de confianza) y 11 grados de libertad.

Con ello puede hacer la siguiente afirmación: ***“El próximo inventario de aguacates tendrá un peso promedio de 141.0029 g y este valor podría fluctuar, con un 95% de confianza o de probabilidad, entre 168.4361 g y 113.5697 g.”***

Con este ejemplo, usted puede darse una idea de cómo hacer estimaciones puntuales y de intervalo cuando su muestra es pequeña (menor de 30 observaciones).

3.3 Estimaciones de intervalo para comparar medias.

3.3.1 Estimaciones de intervalo para muestras apareadas grandes y pequeñas.

3.3.1.1 Estimación de intervalo para muestras apareadas grandes

Recuerde usted que, por el Teorema del Límite central, se puede considerar una muestra como “grande” si tiene más de 30 observaciones e incluso se puede suponer que está normalmente distribuida si el tamaño de dicha muestra es menor al 5% del tamaño de la población total, si es que se sabe. Si usted ve detenidamente, se tienen 4 grupos de 20 individuos que dan un total de 80 diferencias o diferencias de calificaciones entre computadoras (D_i). Por tanto, se puede aceptar el supuesto de que es muestra grande y de que está normalmente distribuida.

Con lo hasta ahora revisado usted podrá hacer estimaciones aplicadas a una muestra. Sin embargo esta técnica de estimación o **inferencia** la puede también aplicar usted para comparar muestras. Recordemos al Sr. Steve Jobs con quien iniciamos estas notas del profesor ¿Qué hizo el Sr. Jobs para determinar que su computadora es mejor que la otra? Al principio de las notas, se mencionó



que probablemente el Sr. Jobs primero hizo un muestreo de racimo. Recordando la nota legal, se mencionó que esta es una mera suposición y resulta ser lo que muchos analistas de mercado o **mercadólogos** harían por su empresa para saber la superioridad de su producto respecto al de la competencia.

Recordando los comentarios iniciales del muestreo de racimo, se observó que este consiste en separar una población en diferentes grupos de interés y luego tomar una muestra de cada segmento, estrato o grupo de interés para tomar una muestra aleatoria de cada uno. ¿Qué pudo hacer el Sr. Jobs? De entrada separó su población objetivo (usuarios de computadoras) en cuatro grupos o estratos de interés:

1. Arquitectos ingenieros, matemáticos, físicos, investigadores y profesionistas que ocupen procesamiento de cálculo.
2. Diseñadores gráficos, artistas de medios, músicos y gente que ocupe procesamiento gráfico.
3. Amas de casa, estudiantes y gente mayor.
4. Contadores, abogados, economistas, financieros y otros profesionistas.

De cada uno de ellos seleccionó y entrevistó a 20 individuos, a los que les aplicó un cuestionario con una serie de preguntas sobre capacidad de procesamiento, facilidad de manejo, costo y calidad. Para esto, el Sr. Jobs les prestó a cada individuo tanto una Macbook como una PC de alta potencia durante un periodo de tiempo y luego les aplicó el cuestionario. Las calificaciones que arroja el mismo iban de 0 para una mala calificación de calidad, precio y procesamiento, es decir, una preferencia muy baja, hasta un 10 que es el nivel máximo de preferencia que pueden tener por la computadora (Mac o PC).

Los resultados de cada uno de los 20 individuos de los cuatro grupos se presenta en la página siguiente en la tabla 12. Para iniciar, en cada individuo, se calcula la diferencia (\bar{D}) entre la calificación dada a la PC o a la Macbook en cada individuo de cada estrato. Esta se calculó simplemente como:

Fórmula 14 Cálculo de la diferencia entre dos muestras:

$$D_i = x_{a,i} - x_{b,i} = \text{Calificación individuo } i \text{ a la Mac} - \text{Calificación individuo } i \text{ a la PC}$$

Como se verá, la forma de calcular estimaciones de intervalo puede aplicarse a muestras apareadas, cuyas variables tienen una influencia entre sí o a muestras que no tienen relación alguna. Es decir, son independientes.

En este primer ejemplo se supondrá que sí existe una relación de forma alguna en los resultados observados. ¿Cuál es la justificación? Simplemente que las calificaciones las está dando el mismo individuo. En breve veremos un caso donde se tienen dos calificaciones de dos individuos del



mismo grupo que constituye una diferencia entre muestras independientes y retomaremos, para ese caso, a la comerciante de aguacates en Chicago y el comerciante de Morelia. Regresemos a nuestro ejemplo.

Vea usted la tabla del muestreo de la página anterior. Lo primero que se hizo fue calcular la diferencia promedio muestral ¿Qué es esto? si aplica a los datos de cada computadora la diferencia (\bar{D}_i) con la fórmula 14 previamente vista, usted obtendrá 20 diferencias que puede tratar como si fueran simples observaciones (x_i).

Para calcular la media de diferencias y la desviación estándar muestral de estas aplicado a muestras apareadas, simplemente se siguen las siguientes fórmulas:

Fórmula 15 Cálculo de la media muestral de diferencias entre dos muestras apareadas o relacionadas:

$$\bar{D} = \frac{\sum D_i}{n}$$

Fórmula 16 Cálculo de la desviación estándar muestral de las diferencias entre dos muestras apareadas o relacionadas:

$$s_D = \sqrt{\frac{\sum D_i^2 - \bar{D}}{n-1}}$$

Fórmula 17 Cálculo del error estándar dada la desviación estándar muestral:

$$\sigma_x = \frac{s_D}{\sqrt{n}}$$

Como aprecia, simplemente se generan las dos medidas estadísticas de interés (media muestral y error estándar) y se procede a realizar la estimación de intervalo de D_i de la siguiente forma:

Fórmula 18 Cálculo de las estimaciones de intervalo para diferencias entre muestras grandes ($n \geq 30$):

$$\text{int. superior} = \bar{D} + (Z_i \times \sigma_{\bar{D}})$$

$$\text{Estimación puntual} = \bar{D}$$

$$\text{int. inferior} = \bar{D} - (Z_i \times \sigma_{\bar{D}})$$



¿Cuál sería el criterio para que el Sr. Jobs demostrara la preferencia que tiene su computadora respecto a la competencia? Simplemente que la diferencia de calificaciones entre la Mac y la PC fuera superior.

| Arquitectos ingenieros, matemáticos, físicos, investigadores y profesionistas que ocupen procesamiento de cálculo | | | |
|---|-----|------------|-------------|
| Observación | PC | Macbook | Macbook-PC |
| 1 | 8.3 | 4.38788967 | -3.91211033 |
| 2 | 0.5 | 4.92209773 | 4.42209773 |
| 3 | 4.3 | 9.67313362 | 5.37313362 |
| 4 | 1.4 | 2.30747633 | 0.90747633 |
| 5 | 5.1 | 2.48006728 | -2.61993272 |
| 6 | 5.6 | 10 | 4.4 |
| 7 | 3.4 | 10 | 6.6 |
| 8 | 6.7 | 5.23247557 | -1.46752443 |
| 9 | 1.1 | 8.47871744 | 7.37871744 |
| 10 | 2.8 | 6.07953997 | 3.27953997 |
| 11 | 2.8 | 8.22183159 | 5.42183159 |
| 12 | 3 | 10 | 7 |
| 13 | 2.4 | 10 | 7.6 |
| 14 | 3.7 | 10 | 6.3 |
| 15 | 2.9 | 10 | 7.1 |
| 16 | 4.7 | 10 | 5.3 |
| 17 | 8.5 | 5.08368353 | -3.41631647 |
| 18 | 8.2 | 10 | 1.8 |
| 19 | 7.2 | 9.12360095 | 1.92360095 |
| 20 | 9.6 | 4.91739203 | -4.68260797 |
| Diferencia media (\bar{D}) | | | 2.93539529 |
| Error estándar de la diferencia ($\sigma_{\bar{D}}$) | | | 0.92373069 |
| Proporción respecto al total de datos | | | 25% |

| Diseñadores gráficos, artistas de medios, músicos y gente que ocupe procesamiento gráfico. | | | |
|--|-----|------------|-------------|
| Observación | PC | Macbook | Macbook-PC |
| 1 | 7.1 | 12.5621048 | 5.46210477 |
| 2 | 0.4 | 0.32056992 | -0.07943008 |
| 3 | 4.6 | 7.11687863 | 2.51687863 |
| 4 | 6.1 | 7.66900193 | 1.56900193 |
| 5 | 6.6 | 11.6677116 | 5.06771158 |
| 6 | 5 | 4.98404484 | -0.01595516 |
| 7 | 2.8 | 3.36454431 | 0.56454431 |
| 8 | 9.9 | 13.0430403 | 3.14304035 |
| 9 | 6.9 | 12.9501221 | 6.05012214 |
| 10 | 3.6 | 6.11445288 | 2.51445288 |
| 11 | 6.6 | 11.6726286 | 5.07262863 |
| 12 | 1.6 | 1.97519172 | 0.37519172 |
| 13 | 5.9 | 4.20549523 | -1.69450477 |
| 14 | 0.4 | 0.0696755 | -0.3303245 |
| 15 | 2.7 | 4.31664637 | 1.61664637 |
| 16 | 6.3 | 10.556557 | 4.25655698 |
| 17 | 5.3 | 4.90381497 | -0.39618503 |
| 18 | 1.6 | 1.16549764 | -0.43450236 |
| 19 | 6.6 | 10.9010258 | 4.30102575 |
| 20 | 8.3 | 3.45679279 | -4.84320721 |
| Diferencia media (\bar{D}) | | | 1.73578985 |
| Error estándar de la diferencia ($\sigma_{\bar{D}}$) | | | 0.62632949 |
| Proporción respecto al total de datos | | | 25% |

| Amas de casa, estudiantes y gente mayor | | | |
|--|-----|------------|-------------|
| Observación | PC | Macbook | Macbook-PC |
| 1 | 5.5 | 7.04566704 | 1.54566704 |
| 2 | 3.5 | 3.78663203 | 0.28663203 |
| 3 | 5.2 | 8.50010688 | 3.30010688 |
| 4 | 1.2 | 1.61805527 | 0.41805527 |
| 5 | 9.9 | 9.79880549 | -0.10119451 |
| 6 | 7.1 | 4.30906344 | -2.79093656 |
| 7 | 9 | 16.1562832 | 7.15628317 |
| 8 | 4.6 | 4.60385059 | 0.00385059 |
| 9 | 5 | 6.80478736 | 1.80478736 |
| 10 | 3.8 | 4.91203839 | 1.11203839 |
| 11 | 7.9 | 1.19808511 | -6.70191489 |
| 12 | 4.9 | 7.85772766 | 2.95772766 |
| 13 | 1.7 | 3.25415281 | 1.55415281 |
| 14 | 2.8 | 0.88413087 | -1.91586913 |
| 15 | 0.7 | 0.20733566 | -0.49266434 |
| 16 | 8.1 | 14.7641442 | 6.66414421 |
| 17 | 3.1 | 4.88957935 | 1.78957935 |
| 18 | 4.4 | 1.21353582 | -3.18646418 |
| 19 | 6.7 | 5.20629118 | -1.49370882 |
| 20 | 5.7 | 9.44867827 | 3.74867827 |
| Diferencia media (\bar{D}) | | | 0.78294753 |
| Error estándar de la diferencia ($\sigma_{\bar{D}}$) | | | 0.72334319 |
| Proporción respecto al total de datos | | | 25% |

| Contadores, abogados, economistas, financieros y otros profesionistas | | | |
|---|-----|------------|-------------|
| Observación | PC | Macbook | Macbook-PC |
| 1 | 8.6 | 14.1055919 | 5.50559194 |
| 2 | 4.9 | 4.43132116 | -0.46867884 |
| 3 | 8.1 | 14.4405188 | 6.34051884 |
| 4 | 7 | 8.58863868 | 1.58863868 |
| 5 | 0.4 | 0.5222463 | 0.1222463 |
| 6 | 3.7 | 0.20552333 | -3.49447667 |
| 7 | 6.4 | 12.1265764 | 5.72657641 |
| 8 | 7.1 | 1.7105593 | -5.3894407 |
| 9 | 1 | 1.7270246 | 0.7270246 |
| 10 | 5.4 | 4.00333502 | -1.39666498 |
| 11 | 0.8 | 1.53742399 | 0.73742399 |
| 12 | 6.5 | 7.46199589 | 0.96199589 |
| 13 | 9.2 | 7.83612541 | -1.36387459 |
| 14 | 3 | 5.19618253 | 2.19618253 |
| 15 | 1.3 | 1.04122643 | -0.25877357 |
| 16 | 4.4 | 7.87412544 | 3.47412544 |
| 17 | 7.1 | 14.0571228 | 6.9571228 |
| 18 | 2.8 | 5.50171221 | 2.70171221 |
| 19 | 4.2 | 6.03230968 | 1.83230968 |
| 20 | 5.8 | 4.63536319 | -1.16463681 |
| Diferencia media (\bar{D}) | | | 1.26674616 |
| Error estándar de la diferencia ($\sigma_{\bar{D}}$) | | | 0.72259254 |
| Proporción respecto al total de datos | | | 25% |

Estadísticas de toda la muestra:

| | |
|--|------------|
| Diferencia media del total de la muestra (\bar{D}) | 1.68021971 |
| Error estándar de la diferencia del total de la muestra ($\sigma_{\bar{D}}$) | 0.74899898 |

Tabla 12 Resultados de un ejemplo de muestreo por racimo que pudo aplicar Apple para determinar la superioridad de su computadora respecto a la de la competencia.



Recuerde usted que se está trabajando con muestras generadas por racimo. Entonces debe usted calcular primero la media de las diferencias en cada estrato (\bar{D}) y el error estándar ($\sigma_{\bar{D}}$). Debajo de la tabla de las muestras de los 4 estratos se ponen dichos valores. Lo que procede, al igual que el muestreo estratificado, es a ponderar el peso que dichas estadísticas tienen en cada estrato. Como las muestras son iguales, se les da una ponderación de $\frac{1}{4}=25\%$. Por tanto, la media y el error estándar de la diferencia de toda la muestra tomada se calcula como:

$$\begin{aligned}\bar{D} &= (25\% \times \bar{D}_{Estrato1}) + (25\% \times \bar{D}_{Estrato2}) + (25\% \times \bar{D}_{Estrato3}) + (25\% \times \bar{D}_{Estrato4}) \\ \bar{D} &= (25\% \times 2.9353) + (25\% \times 1.7357) + (25\% \times 0.7829) + (25\% \times 1.2667) \\ \bar{D} &= 1.6802 \\ \sigma_{\bar{D}} &= (25\% \times \sigma_{\bar{D}, Estrato1}) + (25\% \times \sigma_{\bar{D}, Estrato2}) + (25\% \times \sigma_{\bar{D}, Estrato3}) + (25\% \times \sigma_{\bar{D}, Estrato4}) \\ \sigma_{\bar{D}} &= (25\% \times 0.9237) + (25\% \times 0.6263) + (25\% \times 0.7233) + (25\% \times 0.7225) \\ \sigma_{\bar{D}} &= 0.7489\end{aligned}$$

Suponga ahora que el Sr. Jobs quiere hacer ahora estimaciones de los datos y hacer una campaña publicitaria para afirmar que su computadora es más preferida que la de la competencia. Para poder hacerla, debe hacer las estimaciones tanto puntual como de intervalo y, para ello, decidió que se utilizara un nivel de confianza de 95%.

Dado este nivel de confianza, se observa, como en la ilustración 2, que el valor Z que corresponde a dicha probabilidad o confianza es de $z_i = 1.6449$. Por tanto, las estimaciones del Sr. Jobs quedarían como sigue:

$$\text{int. superior} = 1.6802 + (1.6449 \times 0.7489) = 2.9122$$

$$\text{Estimación puntual} = 1.6802$$

$$\text{int. inferior} = 1.6802 - (1.6449 \times 0.7489) = 0.4482$$

| Estimación de intervalo (muestra grande) | |
|--|--------|
| Intervalo de confianza | 95% |
| Valor Z | 1.6449 |
| Intervalo superior (muestra grande) | 2.9122 |
| Estimación puntual | 1.6802 |
| Intervalo inferior (muestra grande) | 0.4482 |

Tabla 13 Estimaciones de intervalo de la diferencia de calificación entre la Mac y la PC realizadas por el Sr. Jobs al aplicar las formas de hacer estimaciones puntuales y de intervalo con la fórmula 18.



Con estas estimaciones, el Sr. Jobs podría hacer la siguiente afirmación: ***“La calificación de preferencia una Mac respecto a una PC será superior en el mercado en 1.6802 puntos y esta puede fluctuar como máximo y mínimo, con un nivel de confianza de 95%, entre 2.9122 y 0.4482”***. Por tanto, partiendo del criterio de que la superioridad de la Mac se da por un valor positivo de la diferencia de calificaciones recibidas entre esta y la competencia, se puede observar que incluso en el escenario más pesimista, dado con el intervalo inferior, la diferencia de las preferencias del consumidor está a favor de la computadora de Apple.

Con este ejemplo del Sr. Jobs se puede apreciar que la estimación de intervalo de una sola muestra tiene la misma validez que la de la comparación entre dos muestras diferentes o de la diferencia de muestras.

Ahora, con lo que se ha trabajado en el ejemplo del Sr. Jobs es con una muestra que se considera grande. Por tanto, se aplica una estimación de intervalo con valores Z que ayudarán a determinar el grado de confianza suponiendo que los valores de dicha muestra se distribuyen normalmente. Sin embargo ¿qué hubiera pasado si la muestra con que se trabaja fuese pequeña?

3.3.1.2 Estimación de intervalo para muestras apareadas pequeñas

Como se vio previamente para trabajar con estimaciones de muestra pequeñas, las fórmulas de cálculo del límite superior e inferior de la estimación de intervalo siguen siendo los mismos. Lo único que cambiaba en la fórmula es el valor Z_i por el valor t_i en la fórmula 18:

Fórmula 19 Cálculo de las estimaciones de intervalo para diferencias entre muestras pequeñas ($n < 30$):

$$\text{int. superior} = \bar{D} + (t_i \times \sigma_{\bar{D}})$$

$$\text{Estimación puntual} = \bar{D}$$

$$\text{int. inferior} = \bar{D} - (t_i \times \sigma_{\bar{D}})$$

Para simplificar el ejemplo del Sr. Jobs, supongamos que los datos de la tabla 12, no son muestra grande sino pequeña y, para esto suponga que los valores de \bar{D} y $\sigma_{\bar{D}}$ de dicha tabla son de una muestra de solo 26 observaciones. Por tanto, si el número de observaciones es de 26, los grados de libertad son $\nu = 25$. Por tanto, al buscar en la tabla t para un nivel de significancia de 5% (el inverso de un nivel de confianza de 95%) con 25 grados de libertad, se llega a un valor $t_i = 2.0595$. Esto es:



| Grados de libertad/probabilidad crítica | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 |
|---|----------|----------|----------|----------|----------|
| 21 | 2.831360 | 2.517648 | 2.327792 | 2.189427 | 2.079614 |
| 22 | 2.818756 | 2.508325 | 2.320160 | 2.182893 | 2.073873 |
| 23 | 2.807336 | 2.499867 | 2.313231 | 2.176958 | 2.068658 |
| 24 | 2.796939 | 2.492159 | 2.306913 | 2.171545 | 2.063899 |
| 25 | 2.787436 | 2.485107 | 2.301130 | 2.166587 | 2.059539 |
| 26 | 2.778715 | 2.478630 | 2.295815 | 2.162029 | 2.055529 |
| 27 | 2.770683 | 2.472660 | 2.290914 | 2.157825 | 2.051830 |
| 28 | 2.763262 | 2.467140 | 2.286380 | 2.153935 | 2.048407 |
| 29 | 2.756386 | 2.462021 | 2.282175 | 2.150325 | 2.045230 |
| 30 | 2.749996 | 2.457262 | 2.278262 | 2.146966 | 2.042272 |
| Distribución normal | 2.326348 | 2.053749 | 1.880794 | 1.750686 | 1.644854 |

Por tanto, ahora la estimación de intervalo, con los valores de \bar{D} y $\sigma_{\bar{D}}$ que ya se tienen se puede hacer como sigue:

$$\text{int. superior} = \bar{D} + (t_i \times \sigma_{\bar{D}}) = 1.6802 + (2.0595 \times 0.7489) = 3.2228$$

$$\text{Estimación puntual} = 1.6802$$

$$\text{int. inferior} = 1.6802 - (2.0595 \times 0.7489) = 0.1376$$

| Estimación de intervalo (suponiendo que es muestra pequeña) | |
|---|--------|
| Intervalo de confianza | 5% |
| Valor t | 2.0595 |
| Intervalo superior (muestra grande) | 3.2228 |
| Estimación puntual | 1.6802 |
| Intervalo inferior (muestra grande) | 0.1376 |

Tabla 14 Estimaciones de intervalo de la diferencia de calificación entre la Mac y la PC realizadas por el Sr. Jobs al aplicar las formas de hacer estimaciones puntuales y de intervalo con la fórmula 19 (suponiendo que la muestra es pequeña. O sea de $n < 30$).

Hasta ahora se han hecho estimaciones de intervalo de diferencias entre muestras apareadas grandes y pequeñas que se suponen tienen algún grado de dependencia. Sin embargo, en algunas ocasiones, esto no siempre es así. Ahora veamos el caso de estimaciones de intervalo de diferencias cuando las muestras son independientes.

3.3.2 Estimaciones de intervalo para muestras independientes.

Para poder calcular las estimaciones de intervalo de diferencias de muestras independientes, lo único que debe cambiar es la forma de calcular tanto la media de las diferencias como el error estándar de las mismas. Esto es:

Fórmula 20 Cálculo de la media de la diferencia para muestras independientes:

$$\bar{D} = \bar{x}_a - \bar{x}_b$$

**Fórmula 21 Cálculo del error estándar de la diferencia de dos muestras independientes:**

$$\sigma_{\bar{D}} = \sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}} \text{ (Para muestras grandes } n \geq 30)$$
$$\sigma_{\bar{D}} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}, s^2 = \frac{\sum x_i}{n-1} \text{ (Para muestras pequeñas } n < 30)$$

Es decir, para calcular la media de las diferencias (\bar{D}) en el caso de muestras independientes no se tiene que seguir un primer paso de calcular una diferencia de observaciones y luego calcularle la media a dichas diferencias. Lo que se hace es calcular la media de las observaciones de cada muestra y luego restarlas como en la fórmula 20.

Para el caso del error estándar ($\sigma_{\bar{D}}$), simplemente se dividen las varianzas de muestras grandes o pequeñas, según sea el caso, entre la raíz del número de observaciones, se suma ese resultado en cada caso y luego se saca la raíz cuadrada como en la fórmula 21.

Con estos valores, los cálculos de estimación puntual, límite superior y límite inferior, siguen siendo los mismos que los de la fórmula 18 para muestras grandes ($n \geq 30$) como pequeñas ($n < 30$):

Para muestras grandes:

$$\text{int. superior} = \bar{D} + (Z_i \times \sigma_{\bar{D}})$$

$$\text{Estimación puntual} = \bar{D}$$

$$\text{int. inferior} = \bar{D} - (Z_i \times \sigma_{\bar{D}})$$

Para muestras pequeñas:

$$\text{int. superior} = \bar{D} + (t_i \times \sigma_{\bar{D}})$$

$$\text{Estimación puntual} = \bar{D}$$

$$\text{int. inferior} = \bar{D} - (t_i \times \sigma_{\bar{D}})$$



| El comerciante de Morelia | | | |
|---------------------------|----------|----------|----------|
| Aguacate | Peso (g) | Aguacate | Peso (g) |
| 1 | 170.89 | 7 | 127 |
| 2 | 185.97 | 8 | 116.8 |
| 3 | 190.74 | 9 | 99.5 |
| 4 | 229.14 | 10 | 107.59 |
| 5 | 145.3 | 11 | 112.34 |
| 6 | 98 | 12 | 108.7654 |

| | |
|-------------------------|-------------|
| Media muestral | 141.00295 |
| Desviación estándar (s) | 43.17671219 |
| Error estándar | 12.4640432 |

| La comerciante de Chicago | | | |
|---------------------------|-------------|----------|------------|
| Aguacate | Peso (g) | Aguacate | Peso (g) |
| 1 | 261.9059912 | 7 | 78.7437381 |
| 2 | 345.7353594 | 8 | 135.551006 |
| 3 | 221.8713694 | 9 | 59.5118592 |
| 4 | 9.731710016 | 10 | 121.142403 |
| 5 | 154.3676559 | 11 | 132.966079 |
| 6 | 122.0928027 | 12 | 83.2955785 |

| | |
|-------------------------|-------------|
| Media muestral | 143.9096294 |
| Desviación estándar (s) | 92.96894523 |
| Error estándar | 26.83782278 |

| | | | |
|---|------------|------------------------|------------|
| Diferencia media de las dos muestras (\bar{D}) | 2.90667936 | Intervalo de confianza | 95% |
| Error estándar de las dos muestras ($\sigma_{\bar{D}}$) | 15.8987434 | Grados de libertad | 11 |
| | | Valor Z | 1.64485363 |
| | | Nivel de significancia | 5% |
| | | Valor t | 2.20098516 |

| Estimaciones suponiendo muestra grande | | |
|--|---|-------------|
| Tipo de estimación | Fórmula | Estimación |
| Intervalo superior | $\bar{D} + (Z_i \times \sigma_{\bar{D}})$ | 29.057785 |
| Estimación puntual | \bar{D} | 2.90667936 |
| Intervalo inferior | $\bar{D} - (Z_i \times \sigma_{\bar{D}})$ | -23.2444263 |

| Estimaciones como la muestra pequeña que es | | |
|---|---|------------|
| Tipo de estimación | Fórmula | Estimación |
| Intervalo superior | $\bar{D} + (t_i \times \sigma_{\bar{D}})$ | 37.8995775 |
| Estimación puntual | \bar{D} | 2.90667936 |
| Intervalo inferior | $\bar{D} - (t_i \times \sigma_{\bar{D}})$ | -32.086219 |

Tabla 15 Ejemplo aplicado a los dos comerciantes de aguacate: La empresaria de Chicago y el de Morelia. Quieren demostrar, con un 95% de confianza, que el inventario de la empresaria de Chicago es mayor en calidad que el de Morelia.

Para ilustrar con un ejemplo, salgamos del ejemplo del Sr. Jobs (porque hemos dicho que sus muestras no son independientes) y ahora pasemos de nuevo con el ejemplo de los dos comerciantes de aguacate: La empresaria de Chicago y el de Morelia. Este ejemplo se presenta en la tabla 15 de la página anterior.

Suponga usted que ambos tomaron una muestra de 12 aguacates (son muestras pequeñas ya que $n < 30$), quieren hacer estimaciones de intervalo con 95% de confianza y la comerciante de Chicago quiere saber si su inventario tendrá una mayor calidad (mayor peso promedio) que la de su homónimo de Morelia.

Si observa en la tabla 15, las medias de diferencias (\bar{D}) y el error estándar de las mismas (calculada como s^2 , se divide entre $n-1$ y no entre n , con la fórmula 16 al ser una muestra pequeña $n < 30$). Estos valores están sombreados. Ahora, También se calculó la diferencia media y el error estándar de las dos muestras independientes como en las fórmulas 20 y 21. Estos valores también se sombrearon.



Con estos valores, los grados de libertad de las muestras (recuerde que se calculan como $\nu = n - 1 = 12 - 1$ (para este ejemplo) y la probabilidad o nivel de confianza de 95% (que lleva a una significancia de 5% para el valor t), se llegan a los valores Z_i y t_i y a las estimaciones tanto puntuales como de intervalo de las diferencias observadas en cada muestra.

Para el primer caso, en donde se supone que esta muestra es grande ($n \geq 30$) ya que las muestras son de $n=12$, se tiene, como estimación puntual, que el inventario que seguirá recibiendo la empresaria de Chicago será superior en calidad que el de Morelia al tener, en promedio 2.0966 g de más en los aguacates que recibe de su proveedor (que es el mismo que el de Morelia). A su vez, la empresaria observa que, con un 95% de confianza, el inventario de aguacates que reciba en el futuro podrá variar en calidad entre -23.2424 g (menor calidad que el de Morelia) y 29.0577 g.

Para el segundo caso, dado que en realidad la muestra con que se trabaja es de $n=12 < 30$, se observa que la calidad promedio o estimación puntual del inventario futuro de la empresaria de Chicago será mayor que el del moreliano. Sin embargo, con un 95% de confianza su estimación podría variar en un rango de -32.0862 g a 37.8995 g de diferencia en la calidad de sus frutas.

Con este ejemplo, se ilustra cómo se hacen estimaciones puntuales y de intervalo para diferencias de muestras cuyos valores son independientes, ya sea pequeñas o grandes.

Con esto se cierra el tema de la Teoría del Muestreo. Antes de proceder a un tema muy importante y sencillo (una vez que se domina esto) es necesario responder dos preguntas que quedaron sueltas:

1. Se ha dicho que las estimaciones de intervalo tienen un grado de confianza de X%. ¿cómo se determina el mismo? Ya que, si incrementamos el grado de confianza o lo reducimos, podemos manipular las cifras de nuestras estimaciones.
2. Se ha hablado de muestras grandes y muestras pequeñas y se sugiere que el tamaño apropiado es que $n \geq 30$ para que los datos se acepte el supuesto de que los datos distribuyan normalmente. ¿Hay otra forma de determinar el tamaño óptimo de la muestra?
3. Hasta ahora se está trabajando con muestras “bien portadas”. Es decir que no se da la presencia de lo que se conoce como **datos atípicos**. Estos ¿qué son? Datos cuyo valor se sale notablemente del resto. Por ejemplo, un aguacate de medio kilo en el ejemplo de los aguacateros.

3.4 ¿Cómo determinar el intervalo de confianza?

Existen muchas técnicas que nos ayudan a calibrar estadísticamente el nivel de confianza que se imprimirá a las estimaciones de intervalo. Sin embargo, estas salen de la óptica y grado de



exigencia del curso ya que en el mismo se le enseñará a dominar las principales técnicas estadísticas de utilidad para su vida profesional. Si usted desea profundizar en esto, puede cursar una maestría en administración o una en finanzas que logre ese grado de profundización o puede consultar fuentes más avanzadas en Econometría o análisis de datos multivariante.

Para usted, sea suficiente saber que un nivel de confianza de 90% o mayor es más que suficiente y que no debe de bajar de dicho valor para poder generar buenas estimaciones.

3.5 ¿Cómo determinar el tamaño de muestra cuando se busca incrementar la precisión del intervalo de confianza?

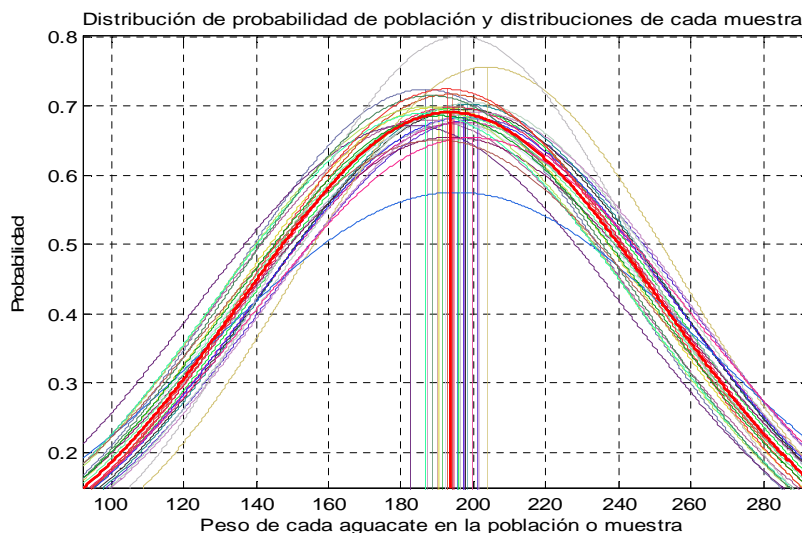
Ya para finalizar el tema de la Teoría del muestreo es necesario completar un poco más una pregunta que se planteó previamente ¿Qué tan grande debe ser la muestra para tener un estudio estadístico adecuado? Esta pregunta se respondió en una primera instancia con el Teorema del límite central que sugiere que la muestra sea mayor o igual a 30 observaciones para poder suponer que los datos se distribuyen normalmente. Sin embargo, el mismo se aplica cuando la población de datos del fenómeno en estudio es muy grande y, por ende, se desconoce el verdadero valor de la desviación estándar poblacional (σ).

Puede darse el caso de que usted sí conozca la desviación estándar de la población y es entonces cuando usted puede determinar el tamaño de la muestra que debe utilizar para poder ser más preciso en sus estimaciones puntuales de intervalo. Es decir lograr que su media muestral y error estándar se aproximen a los mismos parámetros calculados en la población.

Cuando se logra que la media muestral sea igual o muy próxima a la de la población, se dice que la muestra es insesgada. Cuando la media muestral es mayor que la poblacional, se dice que tiene sesgo positivo y, cuando sucede lo contrario, tienen sesgo negativo.

Cuando se logra que el error estándar sea igual o aproximado a la desviación estándar de la población, se dice que la muestra es eficiente.

Como se ha visto al inicio del tema de la Teoría del muestreo, se busca hacer estimaciones precisas y, para esto, se tiene el siguiente comportamiento. Recuerde usted la gráfica 21:



Recuerde también cómo, conforme el tamaño de la muestra se incrementaba (el número de observaciones era mayor), tanto la media muestral como el error estándar de la muestra se aproximan al de la población y la precisión de la estimación de intervalo podría incrementarse (ausencia de sesgo y eficiencia), al reducirse la fluctuación de la media muestral alrededor de la media poblacional y al tener más datos que hacían que $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \sigma$ al ser n cada vez más grande y, dada la división, aproximar el valor de $\sigma_{\bar{x}}$ al de σ .

Por tanto, si no quiere utilizar el Teorema del límite central, conoce ahora el valor de la desviación estándar poblacional y desea ahorrar costos para no hacer muestras demasiado grandes, podría preguntarse ¿qué tan grande debe ser la muestra para aumentar la precisión de la estimación de intervalo dado un intervalo de confianza si ya conozco la desviación estándar poblacional?

Para poder responder esto, primero piense ¿Qué tipo de casos pueden ser aquellos en donde si conozca la desviación estándar poblacional y en los que me obligue a hacer muestras?

Para ilustrar un caso similar, regresemos con los comerciantes de aguacates. Hablemos de nuevo de la empresaria de Chicago. Suponga que existen estadísticas dadas por APEAM (Asociación de Productores y Empacadores de Aguacate de Michoacán) en las que determinan, el peso promedio y la desviación estándar de los pesos de las frutas producidas durante muchos años. Dada la naturaleza de los datos, podríamos decir, para fines de nuestro ejemplo, que todos los datos procesados son de una población. Esta población tiene una desviación estándar de $\sigma = 250.15$ g.

¿Cómo determinará la comerciante de Chicago el tamaño de muestra óptimo? Si ella tiene una muestra de 12 aguacates, con una media muestral de 141.0029 g y un error estándar de 12.4640 g, éste tendría una estimación de intervalo de 168.4361 g y 113.5697 g. Ese intervalo tiene un



rango de 54.8663 g y, tal vez, usted lo considere demasiado grande. Quizá usted quiera saber: con el mismo 95% de confianza, ¿Qué tan grande debe hacerse la muestra para que la fluctuación de la estimación de intervalo sea de $\pm 100g$ respecto a la media muestral (\bar{x}) con un 95% de confianza?

Para llegar a la fórmula que nos ayudará a responder esto, haremos uso de conocimientos de álgebra muy elementales que usted ya domina porque es el álgebra que se le enseñó en la secundaria. De entrada, recordemos la fórmula para calcular el intervalo superior con muestras grandes (es decir cuando se acepta el supuesto de normalidad):

$$\bar{x} + (Z_i \times \sigma_{\bar{x}})$$

Dado que se busca que el intervalo fluctúe solo $\pm 10g$ respecto a la media muestral, se puede tener la siguiente definición:

$$\bar{x} + (Z_i \times \sigma_{\bar{x}}) = \bar{x} + 10g$$

O sea:

$$Z_i \times \sigma_{\bar{x}} = 10g$$

Ahora se recuerda la fórmula 8 para calcular el error estándar:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Esta se sustituye en la ecuación anterior y nos da:

$$Z_i \times \sigma_{\bar{x}} = Z_i \times \frac{\sigma}{\sqrt{n}} = 10g$$

Entonces ahora se despeja la n ¿cómo? Observe:

$$Z_i \times \frac{\sigma}{\sqrt{n}} = 10g$$

$$Z_i \times \sigma = 10g \times \sqrt{n}$$

$$\frac{Z_i \times \sigma}{10g} = \sqrt{n}$$

$$n = \left(\frac{Z_i \times \sigma}{10g} \right)^2$$



Si definimos el grado de fluctuación o **error tolerable** como $e = 10g$, se llega entonces a la siguiente fórmula para determinar el tamaño óptimo de la muestra cuando se conoce la desviación estándar de la población y se busca dar un intervalo de confianza de x% (ejemplo 95%).

Fórmula 22 Determinación del tamaño de muestra óptimo cuando se conoce la desviación estándar poblacional:

$$n = \left(\frac{Z_i \times \sigma}{e} \right)^2$$

Siguiendo el ejemplo de la comerciante de aguacates de Chicago, si ella busca una precisión de 10 g con un intervalo de confianza de 95%, se tiene el siguiente tamaño de muestra óptimo:

$$n = \left(\frac{1.6445 \times 100.15g}{10g} \right)^2 = 271.3666$$

Es decir, que si quiere, con un 95% de confianza hacer estimaciones cuya precisión de estimación de intervalo solo se aleje ± 100 g de la puntual., debe tener una muestra de 271 aguacates. Es decir, si quiere que su muestra sea **confiable**, esta debe cumplir que su media muestral sea cercana a la poblacional (no esté sesgada) y que el error estándar sea muy aproximado o igual a la desviación estándar de la población (sea eficiente). Si se permite una lejanía de ± 100 g de la media poblacional, se debe incrementar la precisión de la muestra al incrementar el número de datos. Y la cantidad apropiada de aguacates dentro de la muestra, deberá ser de 271.

Con esto, se da fin al tema de la Teoría del muestreo. El siguiente tema está muy relacionado y es consecuencia del presente. Como verá, una vez que domina la Teoría del muestreo, la comprobación de hipótesis es natural de estudiar y comprender.



4 Prueba de hipótesis: La técnica clásica

Hasta ahora se ha visto una de las aplicaciones de la Estadística inferencial que es la estimación de valores futuros dados los datos muestrales con que se cuenta. Ahora se revisará una de las técnicas más útiles y necesarias de la misma, la cual no será excepción en aplicaciones de su empresa y futuras materias de su carrera como pueden ser Producción, Administración de la calidad o Finanzas. Esta técnica de la que se habla es: la prueba de hipótesis.

En el subtema 2.7 se habló de un proceso de 5 pasos que se debe seguir en el proceso de toma de decisiones utilizando la Estadística y que es el mismo que deberá usted aplicar en su vida cotidiana:

1. **Definir el objetivo:** Definir el objetivo de la decisión que se va a hacer. Por ejemplo, determinar si la calidad del inventario es buena o no, si el número de piezas desperdiciadas es mayor a cierta cantidad, si el número de trimestres con pérdida es mayor a cierto objetivo o si el número de conexiones fallidas en un sistema de cómputo es mayor a determinada cantidad objetivo que se define en los estándares de calidad de la empresa de comunicaciones. Estos ejemplos se dan por citar algunos casos de lo que podría presentársele en su vida cotidiana.
2. **Definir lo que se medirá:** Aquí usted definirá cuál será la variable que delimitará la toma de sus decisiones. Por ejemplo “La cantidad de desperdicio” o el número de trimestres con pérdidas”.
3. **Definir el tamaño de muestra:** Esto es de vital importancia y se ha revisado en temas anteriores. Si usted no conoce el verdadero tamaño de la población ni sus parámetros como son la media y la desviación estándar, entonces apelará al teorema del límite central. Si se encuentra al caso contrario, usted empleará la fórmula 22 si y solo si se le proporciona algún valor que corresponda a la desviación estándar de dicha población.
4. **Analizar los datos:** Aquí se pueden utilizar varias técnicas de análisis. De entrada pueden ser las técnicas de estimación (puntual y de intervalo) y la comprobación de hipótesis.
5. **Conclusión y toma de decisiones:** Para fines del tema que interesa, una vez que se aplica la comprobación de hipótesis, se tiene una conclusión de la que se toma una decisión en la empresa.

Las pruebas de hipótesis, al igual que las estimaciones (las cuales son la base de la prueba de hipótesis) se pueden hacer para muestras simples o para comparar muestras (diferencias de muestras). En un primer subtema iniciaremos con la prueba de hipótesis para muestras simples o una sola muestra.



4.1 Comprobación de hipótesis de una sola muestra.

Para exponer el concepto de la prueba de hipótesis se deben recordar tanto la forma de hacer estimaciones como el cálculo de probabilidades empleando valores Z o valores t. Para iniciar con la exposición de la idea recordemos al ejemplo de los comerciantes de aguacate. En concreto, centremos la atención de la empresaria de Chicago. Suponga usted que ella busca definir que la calidad de su inventario (recordemos que este concepto está medido a través del peso de cada fruta) debe ser mayor a 3.8 onzas (Oz.) para decir que tiene buena calidad. Suponga que la empresaria toma una muestra de 30 aguacates de su inventario total de 5,000 y la experiencia de inventarios previos le dice que la desviación estándar en el peso de los aguacates es de 1.1 Oz. Es decir, aquí no se tiene medida la desviación estándar de una población pero se supone que ésta desviación estándar, que se logra con la experiencia de inventarios previos, es una aproximación adecuada¹³.

Para poder responder esta pregunta de si el inventario cubre los estándares de calidad de la empresa, la comerciante de Chicago tuvo que hacer los tres primeros pasos del proceso de toma de decisiones con la Estadística:

1. **Definir el objetivo:** Determinar si el estándar de calidad mínimo requerido se cumple en el inventario.
2. **Definir lo que se medirá:** Definir “calidad” como sinónimo de peso de la fruta: Más peso=más calidad.
3. **Definir el tamaño de muestra:** En base al Teorema del límite central previamente visto, la empresaria decide hacer una muestra de 30 piezas.

El cuarto paso correspondería al análisis de datos y es en ese punto donde se realiza la comprobación de hipótesis. En términos generales, lo que se busca hacer en una comprobación de hipótesis con la técnica clásica es determinar que, dada la muestra de datos que se tiene, la media de la misma es igual, más grande o más pequeña que la media de la población o media objetivo.

Definición de prueba de hipótesis: Método para evaluar creencias o afirmaciones sobre la realidad en base en la evidencia estadística, de tal forma que se determine la validez de dichas creencias o afirmaciones.

Para poder comprobar la hipótesis es necesario establecer que, en el caso de la técnica clásica, lo que se busca demostrar es que la media de la muestra empleada es parecida, inferior o superior a una media poblacional o a una aproximación de la misma dada a través de una media hipotética u objetivo.

¹³ En ocasiones no se tienen los datos exactos de tamaño de población, media poblacional o desviación estándar poblacional. Sin embargo, suponer los valores de estas medidas con los observados en experiencias previas puede ser válido y de mucha utilidad en la práctica cotidiana.



En este punto se hace mención de un primer concepto a resaltar consistente en la **media hipotética de la población** (μ_{H_0}), que no es más que definir “un nivel de media objetivo” si se desconoce el de la población.

El objetivo que se busca lograr con la prueba de hipótesis es determinar una de las siguientes posibilidades de hipótesis:

| D | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|---|---|----------------|--|---|--|
| 1 | Determinar si la media de la muestra que se tiene es igual a la de su población | Dos colas | $H_0: -IC < \bar{X} < IC$ | $H_0: -ZC < Z_{\bar{X}} < ZC$ | $H_0: -tC < t_{\bar{X}} < tC$ |
| 2 | Determinar si la media de la muestra que se tiene es diferente a la de su población | Dos colas | $H_0: \bar{X} < -IC \text{ o } IC < \bar{X}$ | $H_0: Z_{\bar{X}} < -ZC \text{ o } ZC < Z_{\bar{X}}$ | $H_0: t_{\bar{X}} < -tC \text{ o } tC < t_{\bar{X}}$ |
| 3 | Determinar si la media de la muestra es superior a la de su población | Cola inferior | $H_0: IC < \bar{X}$ | $H_0: ZC < Z_{\bar{X}}$ | $H_0: tC < t_{\bar{X}}$ |
| 4 | Determinar si la media de la muestra es inferior a la de su población | Cola superior | $H_0: \bar{X} < -IC$ | $H_0: Z_{\bar{X}} < -ZC$ | $H_0: t_{\bar{X}} < -tC$ |

Tabla 16 Reglas de decisión para las pruebas de hipótesis de una sola muestra.

Este tipo de hipótesis y el tipo de prueba (cola inferior, cola superior o dos colas) se verán a detalle en breve. Lo que desea resaltar es que, en la técnica estadística que interesa, se busca determinar si la media de una muestra es igual, superior o inferior a un estándar teórico.

También se revisará a mayor detalle el tipo de escala a utilizar (original o estandarizada) y se estudiarán, con ejemplos, las diferentes reglas de selección.

Como se ha visto, la comprobación de hipótesis es un procedimiento estadístico, comprendido en el cuarto paso del proceso de toma de decisiones, el cual, a su vez, realiza los siguientes pasos:

El proceso de comprobación de hipótesis:

- 1. Definir una hipótesis nula a demostrar:** Se plantea el enunciado (hipótesis) a demostrar y se plantean la hipótesis nula (H_0) y la alternativa (H_a).
- 2. Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior.**
- 3. Se define el grado de significancia:** Este es el contrario al intervalo de confianza. Es decir, si se tiene 5% de significancia, se tiene 95% de confianza. Aquí se calculan los valores Z o t con las expresiones de las formulas 24 y 25.
- 4. Se define si se trabaja con la escala original o con una estandarizada.**
- 5. Se define la regla de aceptación.**
- 6. Se comparan los valores críticos fijados con los estadísticos (valor Z o media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a).**



Fórmula 24: Cálculo del valor Z para una prueba de hipótesis con la técnica clásica:

$$Z = \frac{\bar{X} - \mu_{H_0}}{\sigma}$$

Fórmula 25: Cálculo del valor t para una prueba de hipótesis con la técnica clásica:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma}$$

Para poder ejemplificar lo expuesto y dar concreción a la forma de realizar una prueba de hipótesis con la técnica clásica, se hacen una serie de ejemplos para que usted asimile el significado de cada paso.

4.1.1 Ejemplos de los diferentes tipos de prueba de hipótesis con técnica clásica aplicados a una muestra simple.

4.1.1.1 Pruebas de hipótesis para demostrar igualdad de la media muestral con una media poblacional conocida o hipotética.

En un primer acercamiento se demostrará la igualdad que tiene la muestra respecto a la media objetivo o poblacional según sea el caso. Se tomará como caso de estudio el inventario de la comerciante de Chicago y se harán ligeros cambios a los estadísticos y parámetros para ilustrar mejor el empleo de la prueba de hipótesis en diferentes circunstancias.

4.1.1.1.1 Prueba de hipótesis para demostración de igualdad empleando muestras grandes.

De entrada, la comerciante de Chicago tiene el siguiente inventario, al cual se le establecen la media hipotética de $\mu_{H_0} = 3.80z$. y una desviación estándar poblacional, determinada con la experiencia previa de la empresaria, de $\sigma = 1.10z$.



| La comerciante de Chicago (muestra grande) | | | |
|--|------------|----------|-------------|
| Aguacate | Peso (g) | Aguacate | Peso (g) |
| 1 | 9.24156638 | 16 | 6.889437022 |
| 2 | 12.199554 | 17 | 2.805266224 |
| 3 | 7.82891212 | 18 | 1.263366478 |
| 4 | 0.34339132 | 19 | 0.321894217 |
| 5 | 5.44698856 | 20 | 0.689824563 |
| 6 | 4.30814406 | 21 | 5.14666579 |
| 7 | 2.77853698 | 22 | 0.934611953 |
| 8 | 4.78302773 | 23 | 6.680220548 |
| 9 | 2.09992446 | 24 | 3.119794535 |
| 10 | 4.27460843 | 25 | 3.4652034 |
| 11 | 4.69181648 | 26 | 3.796020078 |
| 12 | 2.93915238 | 27 | 5.120936133 |
| 13 | 5.27263691 | 28 | 8.525283542 |
| 14 | 8.44771181 | 29 | 11.79793331 |
| 15 | 2.32162703 | 30 | 4.563482046 |

| Parámetros poblacionales | |
|--------------------------|-----|
| μ_{H_0} | 3.8 |
| σ | 1.1 |

| | | |
|---------------------|--------------------|-------------|
| Media muestral | \bar{X} | 4.736584616 |
| Desviación estándar | | 3.12868851 |
| Error estándar | $\sigma_{\bar{x}}$ | 0.903174577 |

El objetivo de la empresaria para este primer caso es determinar que un embarque de 5,000 piezas de aguacate que le acaba de llegar se ajusta a su estándar de calidad de 3.8 onzas que es el valor que define a la media poblacional hipotética (μ_{H_0}). Para ello tomó 30 piezas de dicho embarque y siguió los siguientes pasos para demostrar que la calidad del mismo se ajusta al objetivo planteado. Esto lo hizo siguiendo los pasos que se presentan:

- Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) igual a 3.8 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} = 3.8$$

$$H_a : \bar{X} \neq 3.8$$

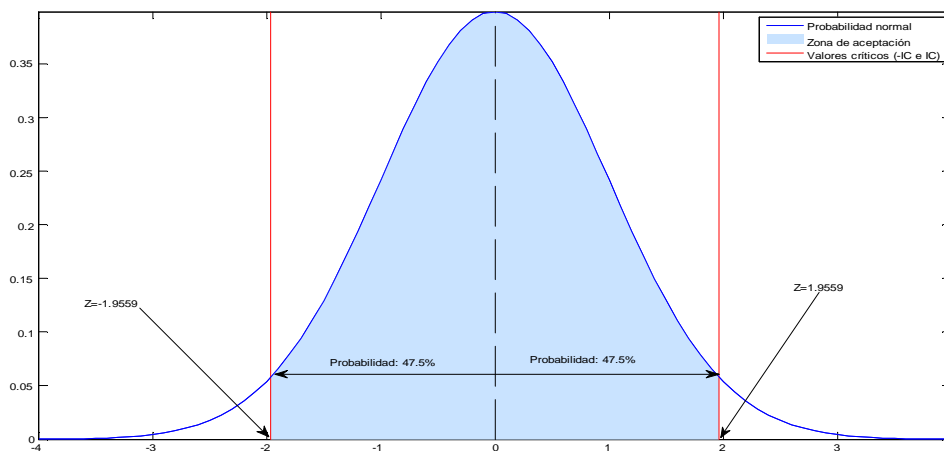
- Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 1, ya que se busca demostrar una igualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de la muestra que se tiene es igual a la de su población | Dos colas | $H_0 : -IC < \bar{X} < IC$ | $H_0 : -ZC < Z_{\bar{X}} < ZC$ | $H_0 : -tC < t_{\bar{X}} < tC$ |

- Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.



4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 2.5%. Al ser esta una prueba de dos colas, debe buscar un valor Z en tablas que corresponda a 47.5% de probabilidad (recuerde que es un 95% de confianza o 5% de significancia que se determina con 47.5% de probabilidad arriba de la media y 47.5% debajo de la misma al ser prueba de dos colas). Esto le lleva a un valor Z de 1.9599.



Gráfica 28 Determinación del valor Z (caso de muestra grande) para una prueba de hipótesis de dos colas.

5. **Se define si se trabaja con la escala original o con una estandarizada:** En este ejemplo, la empresaria decidió trabajar con la escala original por lo que utilizó el valor Z para definir los valores críticos ($-IC, IC$) correspondientes al intervalo de confianza con los que aceptará o rechazará la hipótesis. Esto la llevó a determinar los siguientes valores críticos:

$$-IC = \mu_{H_0} + (Z \cdot \sigma) = 3.8 - (1.9599 \cdot 1.1) = 1.6440$$

$$IC = \mu_{H_0} + (Z \cdot \sigma) = 3.8 + (1.9599 \cdot 1.1) = 5.9559$$

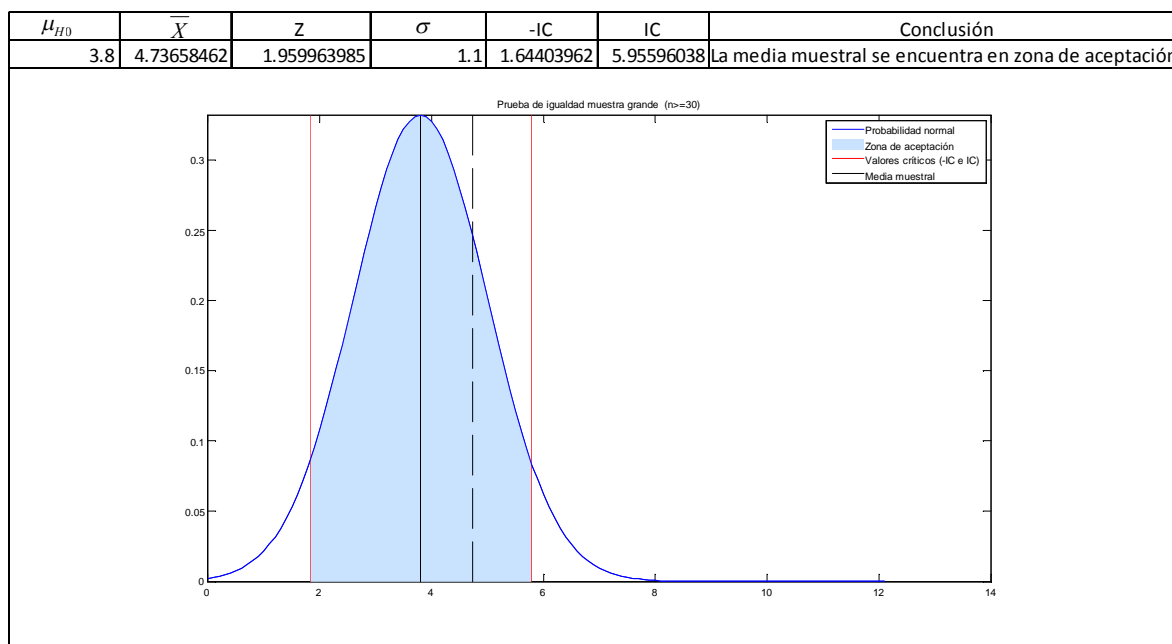
6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad (prueba de dos colas) se definió, en la gráfica anterior y como zona de aceptación, a todos los valores de \bar{X} que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $-IC < \bar{X} < IC$.

Aceptar H_a : Si $\bar{X} < -IC$ o $IC < \bar{X}$.



7. Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a): Con esto, se tienen los siguientes resultados:



Conclusión: En base a los datos que tiene la empresaria de Chicago, ella puede concluir que el embarque de 5,000 aguacates cumple con los estándares de calidad que tiene establecidos ya que la media muestral de una muestra aleatoria de 30 aguacates es estadísticamente igual al peso objetivo planteado de μ_{H0} .

4.1.1.1.2 Prueba de hipótesis para demostración de igualdad empleando una muestra grande y una escala estandarizada.

Ahora se realizará la prueba de hipótesis cambiando la escala original por una escala estandarizada. Es decir, se aplicará la fórmula del cálculo del valor Z dada en la fórmula 9 a la media muestral a contrastar, considerando que es muestra grande. Esto lleva al cálculo de los estadísticos de la forma en que se expresa en la fórmula 24 para muestra grande:

$$z = \frac{\bar{x} - \mu_{h0}}{\sigma_x^-}$$

Con esto se hace una prueba de hipótesis siguiendo los pasos establecidos:



1. **Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) igual a 3.8 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} = 3.8$$

$$H_a : \bar{X} \neq 3.8$$

2. **Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 1, ya que se busca demostrar una igualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de la muestra que se tiene es igual a la de su población | Dos colas | $H_0 : -IC < \bar{X} < IC$ | $H_0 : -ZC < Z_{\bar{X}} < ZC$ | $H_0 : -tC < t_{\bar{X}} < tC$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 2.5%. Al ser esta una prueba de dos colas, debe buscar un valor Z en tablas que corresponda a 47.5% de probabilidad. Esto le lleva a un valor Z de 1.9599.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala estandarizada, lo que se busca es determinar los valores críticos (IC) como los valores Z que corresponden a

$$-IC = \text{Valor Z de intervalo inferior 2.5\%} = -1.9599$$

$$IC = \text{Valor Z de intervalo superior 97.5\%} = 1.9599$$

Ya que se tienen los valores críticos de la prueba, se procede a calcular el estadístico de prueba, en este caso un valor Z dado por la fórmula 24:

$$Z = \frac{\bar{X} - \mu_{H_0}}{\sigma} = \frac{4.7365 - 3.8}{1.1} = 0.8514$$

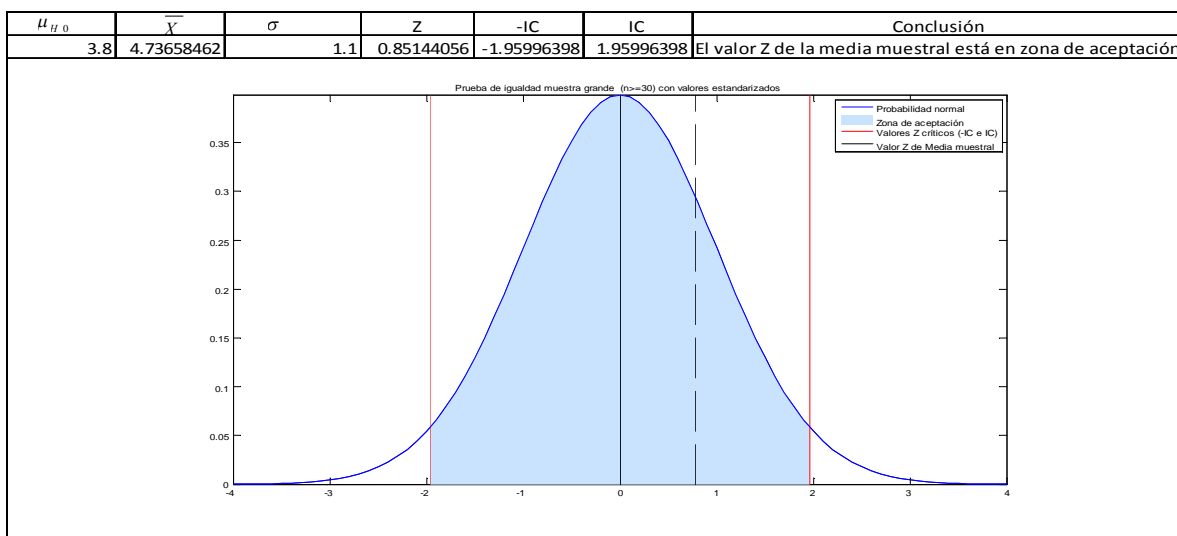
6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad (prueba de dos colas) se definió, en la gráfica anterior, como zona de aceptación a todos los valores de Z que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $-IC < Z < IC$.



Aceptar H_a : Si $Z < -IC$ o $IC < Z$.

7. Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a): Se tienen los siguientes resultados:



Conclusión: Por tanto, en base a los datos que tiene la empresaria de Chicago, ella llega a la misma conclusión de la prueba de hipótesis anterior, con la diferencia de que se utilizó una escala diferente.

4.1.1.1.3 Prueba de hipótesis para un caso de demostración de igualdad con una muestra pequeña con escala original.

Note usted cómo se empleó, para las pruebas anteriores, la escala original y el valor Z ya sea para realizar las estimaciones de intervalo o para definir el estadístico de prueba en una escala estandarizada. Sin embargo ¿Qué hubiera sucedido si la empresaria hubiera tomado sólo 15 aguacates en lugar de 30 y hubiese decidido emplear la escala original (onzas)? En este punto, la muestra sería pequeña y la media muestral sería ahora de $\bar{X} = 5.1318$:



| Chicago (muestra pequeña n=15) | |
|--------------------------------|------------|
| Aguacate | Peso (g) |
| 1 | 9.24156638 |
| 2 | 12.199554 |
| 3 | 7.82891212 |
| 4 | 0.34339132 |
| 5 | 5.44698856 |
| 6 | 4.30814406 |
| 7 | 2.77853698 |
| 8 | 4.78302773 |
| 9 | 2.09992446 |
| 10 | 4.27460843 |
| 11 | 4.69181648 |
| 12 | 2.93915238 |
| 13 | 5.27263691 |
| 14 | 8.44771181 |
| 15 | 2.32162703 |

| | | |
|---------------------|--------------------|-------------|
| Media muestral | \bar{X} | 5.131839908 |
| Desviación estándar | | 3.12868851 |
| Error estándar | $\sigma_{\bar{x}}$ | 0.903174577 |

El objetivo es demostrar desigualdad para una media hipotética de $\mu_{H_0} = 3.8 \text{ Oz}$. Para ello, se siguieron los pasos que se presentan a continuación:

- Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) diferente a 3.8 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} = 3.8$$

$$H_a : \bar{X} \neq 3.8$$

- Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 1, ya que se busca demostrar una igualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de la muestra que se tiene es igual a la de su población | Dos colas | $H_0 : -IC < \bar{X} < IC$ | $H_0 : -ZC < Z_{\bar{x}} < ZC$ | $H_0 : -tC < t_{\bar{x}} < tC$ |



3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra pequeña, se emplea la distribución t-Student o un valor t.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 15 piezas. Por tanto empresaria decide utilizar un valor t que corresponda a un nivel de significancia de 2.5%, al ser esta una prueba de dos colas. Con esto, debe buscar un valor en tablas que corresponda a 2.5% de probabilidad con 14 grados de libertad, lo que la lleva a un t de 2.1447.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala original, se establecen los siguientes valores críticos:

$$-IC = \mu_{H_0} + (t \cdot \sigma) = 3.8 - (2.1447 \cdot 1.1) = 1.4407$$

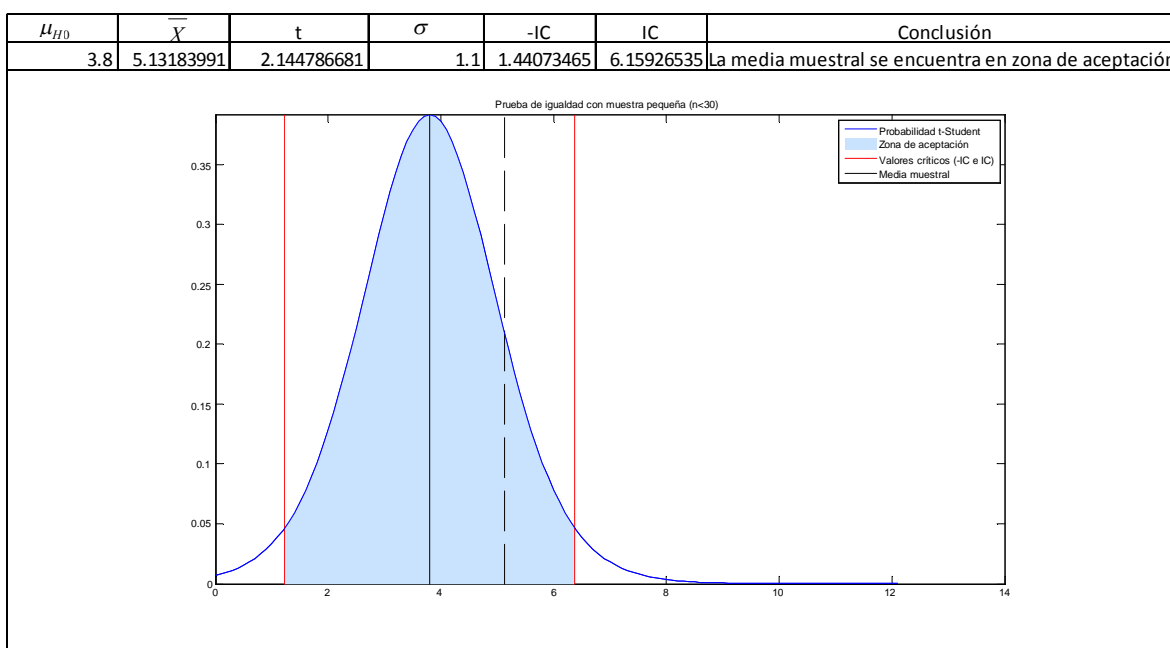
$$IC = \mu_{H_0} + (t \cdot \sigma) = 3.8 + (2.1447 \cdot 1.1) = 6.1592$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad (prueba de dos colas) se define como zona de aceptación a todos los valores de \bar{X} que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $\bar{X} < -IC$ o $IC < \bar{X}$.

Aceptar H_a : Si $-IC < \bar{X} < IC$.

7. **Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Se tienen los siguientes resultados:





Conclusión: Por tanto, en base a los datos que tiene la empresaria de Chicago, ella puede concluir que el embarque de 5,000 aguacates cumple con los estándares de calidad que ella tiene establecidos ya que la media de una muestra aleatoria de 15 aguacates (muestra pequeña) es estadísticamente igual a la media hipotética u objetivo dada por μ_{H_0} y esto le dice que, dada la información de esas pocas piezas, el resto del embarque puede ser aceptado y comprado.

4.1.1.1.4 Prueba de hipótesis para demostración de igualdad empleando una muestra pequeña y una escala estandarizada.

Ahora se procederá a realizar la prueba de igualdad con muestra pequeña como la anterior pero utilizando valores estandarizados. Para ello se tienen los siguientes pasos:

- 1. Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) igual a 3.8 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} = 3.8$$

$$H_a : \bar{X} \neq 3.8$$

- 2. Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 1, ya que se busca demostrar una igualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de la muestra que se tiene es igual a la de su población | Dos colas | $H_0 : -IC < \bar{X} < IC$ | $H_0 : -ZC < Z_{\bar{X}} < ZC$ | $H_0 : -tC < t_{\bar{X}} < tC$ |

- 3. Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra pequeña, se emplea la t-Student y, por ende, se emplea un valor t.
- 4. Se define el grado de significancia:** La muestra con que se trabaja es de 15 piezas. Por tanto empresaria decide utilizar un valor t que corresponda a un nivel de significancia de 2.5%. Al ser esta una prueba de dos colas, debe buscar un valor t en tablas que corresponda a 47.5% de probabilidad (recuerde que es un 47.5% de probabilidad arriba de la media y 47.5% debajo de la misma). Esto le lleva a un valor t de 2.1447.
- 5. Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala estandarizada, lo que se busca es determinar los valores críticos (IC) como los valores t que corresponden a:



$$-IC = \text{Valor } t \text{ de intervalo inferior } 2.5\% = -2.1447$$

$$IC = \text{Valor } t \text{ de intervalo superior } 97.5\% = 2.1447$$

Ya que se tienen los valores críticos de la prueba, se procede a calcular el estadístico de prueba, en este caso un valor t dado por la fórmula 25:

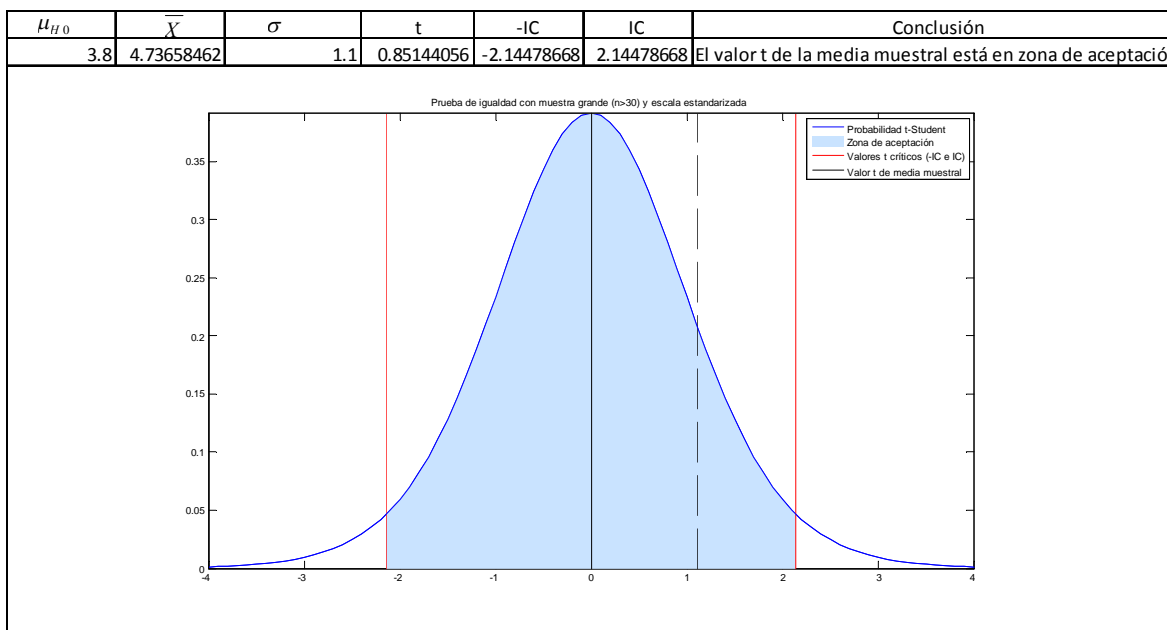
$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma} = \frac{4.7365 - 3.8}{1.1} = 0.8514$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad (prueba de dos colas) se definió, como zona de aceptación, a todos los valores de t que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $-IC < t < IC$.

Aceptar H_a : Si $t < -IC$ o $IC < t$.

7. **Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Se tienen los siguientes resultados:



Conclusión: Por tanto, en base a los datos que tiene la empresaria de Chicago, ella llega a la misma conclusión de la prueba de hipótesis anterior, con la diferencia de que se utilizó una escala diferente. Es decir, una estandarizada.



4.1.1.2 Pruebas de hipótesis para demostrar desigualdad de la media muestral con una media poblacional conocida o hipotética.

Ahora corresponde el caso de demostrar que existe una desigualdad entre la media muestral de los datos que se procesan y la media poblacional o la media objetivo (μ_{H_0}).

Para poder utilizar este ejemplo, suponga ahora que, por alguna circunstancia peculiar, la empresaria de Chicago desea que el inventario de aguacates sea diferente de $\mu_{H_0} = 8.9\text{Oz}$. Es decir, puede tener cualquier peso superior o inferior diferente a 8.9 Oz. Entonces, la comerciante buscará hacer una prueba de hipótesis en donde busca comprobar que dicha desigualdad existe. A continuación se le presentan los 4 casos de igualdad estudiados pero como desigualdades.

Los datos del problema para el caso de una prueba de hipótesis de desigualdad, como se ha visto, serán los mismos salvo el valor de la media hipotética $\mu_{H_0} = 8.9\text{Oz}$. A su vez, es de necesidad observar que, para el caso de desigualdad, las reglas de aceptación de la hipótesis nula (H_0) cambian por las siguientes marcadas con el ID 2 en la tabla 16:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 2 | Determinar si la media de la muestra que se tiene es diferente a la de su población | Dos colas | $H_0: \bar{X} < -IC \text{ o } IC < \bar{X}$ | $H_0: Z_{\bar{X}} < -ZC \text{ o } ZC < Z_{\bar{X}}$ | $H_0: t_{\bar{X}} < -tC \text{ o } tC < t_{\bar{X}}$ |

A su vez, es de necesidad recordar que para la muestra grande (con 30 piezas para el análisis), la media muestral es de $\bar{X} = 4.7365$ y la de la muestra pequeña (15 piezas) de $\bar{X} = 5.1318$. Para la desviación estándar poblacional se supone como válida la que la empresaria ha tomado en base a su experiencia previa con este proveedor: $\sigma = 1.1$.

Con estos datos a mano, podremos entonces realizar el análisis estadístico correspondiente para determinar si la calidad de dicha muestra es diferente o no al objetivo o hipótesis teórica marcados de $\mu_{H_0} = 8.9\text{Oz}$.

4.1.1.2.1 Prueba de hipótesis para demostración de desigualdad empleando muestras grandes y escala original.

Ahora se demostrará la desigualdad de la muestra tomada, respecto al objetivo planteado tomando la escala original de valores. Para ello, se siguen estos pasos:

- 1. Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) diferente a 8.9 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:



$$H_0: \bar{X} \neq 8.9$$

$$H_a: \bar{X} = 8.9$$

2. **Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 2, ya que se busca demostrar una desigualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 2 | Determinar si la media de la muestra que se tiene es diferente a la de su población | Dos colas | $H_0: \bar{X} < -IC \text{ o } IC < \bar{X}$ | $H_0: Z_{\bar{X}} < -ZC \text{ o } ZC < Z_{\bar{X}}$ | $H_0: t_{\bar{X}} < -tC \text{ o } tC < t_{\bar{X}}$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de dos colas, debe buscar un valor Z en tablas que corresponda a 47.5% de probabilidad (recuerde que es un 47.5% de probabilidad arriba de la media y 47.5% debajo de la misma). Esto le lleva a un valor Z de 1.9599.
5. **Se define si se trabaja con la escala original o con una estandarizada:** En este ejemplo, la empresaria decidió trabajar con la escala original por lo que utilizó el valor Z para definir los valores críticos $(-IC, IC)$ del intervalo de confianza con los que aceptará o rechazará la hipótesis. Esto la llevó a determinar los siguientes valores críticos:

$$-IC = \mu_{H_0} + (Z \cdot \sigma) = 8.9 - (1.9599 \cdot 1.1) = 6.7440$$

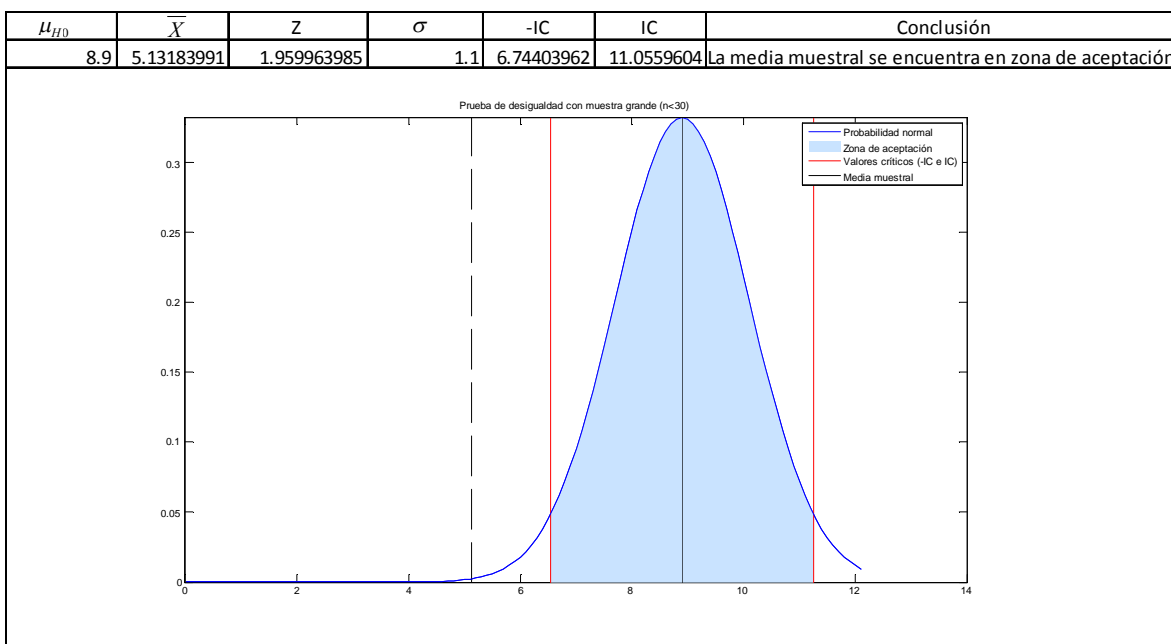
$$IC = \mu_{H_0} + (Z \cdot \sigma) = 8.9 + (1.9599 \cdot 1.1) = 11.0559$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de desigualdad (prueba de dos colas) se definió, como zona de aceptación, a todos los valores de \bar{X} que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $-IC < \bar{X} < IC$.

Aceptar H_a : Si $\bar{X} < -IC$ o $IC < \bar{X}$.

7. **Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Con esto, se tienen los siguientes resultados:



Conclusión: En base a los datos que tiene la empresaria de Chicago, ella puede concluir que el embarque de 5,000 aguacates cumple con los estándares de calidad que tiene establecidos ya que la media muestral de una muestra aleatoria de 30 aguacates es estadísticamente diferente al peso objetivo planteado de $\mu_{H0} = 8.90z$.

4.1.1.2.2 Prueba de hipótesis para demostración de desigualdad empleando una muestra grande y una escala estandarizada.

Ahora se realizará la prueba de hipótesis cambiando la escala original por una escala estandarizada. Es decir, se aplicará la fórmula del cálculo del valor Z dada en la fórmula 9 a la media muestral a contrastar, considerando si es muestra grande. Esto lleva al cálculo de los estadísticos de la forma en que se expresa en la fórmula 24 para muestra grande:

$$z = \frac{\bar{x} - \mu_{H0}}{\sigma_x^-}$$

Retomando el objetivo de la empresaria de Chicago quien desea demostrar que un determinado embarque de aguacates tiene un peso diferente a la media hipotética de $\mu_{H0} \neq 8.90z$. y una desviación estándar poblacional (determinada con la experiencia previa de la empresaria) de $\sigma = 1.10z$, la comprobación de hipótesis llevaría a un estadístico de prueba Z dado por (recuerde que la muestra es de 30 piezas y la media muestral de 4.7365):



$$z = \frac{4.7365 - 8.9}{1.1} = -3.7849$$

Con esto se hace una prueba de hipótesis siguiendo los pasos establecidos:

1. **Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) diferente a 8.9 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0: \bar{X} \neq 8.9$$

$$H_a: \bar{X} = 8.9$$

2. **Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 2, ya que se busca demostrar una igualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 2 | Determinar si la media de la muestra que se tiene es diferente a la de su población | Dos colas | $H_0: \bar{X} < -IC \text{ o } IC < \bar{X}$ | $H_0: Z_{\bar{X}} < -ZC \text{ o } ZC < Z_{\bar{X}}$ | $H_0: t_{\bar{X}} < -tC \text{ o } tC < t_{\bar{X}}$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 2.5%. Al ser esta una prueba de dos colas, debe buscar un valor Z en tablas que corresponda a 47.5% de probabilidad. Esto le lleva a un valor Z de 1.9599.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala estandarizada, lo que se busca es determinar los valores críticos (IC) como los valores Z que corresponden a

$$-IC = \text{Valor Z de intervalo inferior } 2.5\% = -1.9599$$

$$IC = \text{Valor Z de intervalo superior } 97.5\% = 1.9599$$

Ya que se tienen los valores críticos de la prueba, se procede a utilizar el estadístico de prueba calculado previamente. En este caso un valor Z dado por la fórmula 24:

$$Z = \frac{\bar{X} - \mu_{H0}}{\sigma} = \frac{4.7365 - 8.9}{1.1} = -3.7849$$

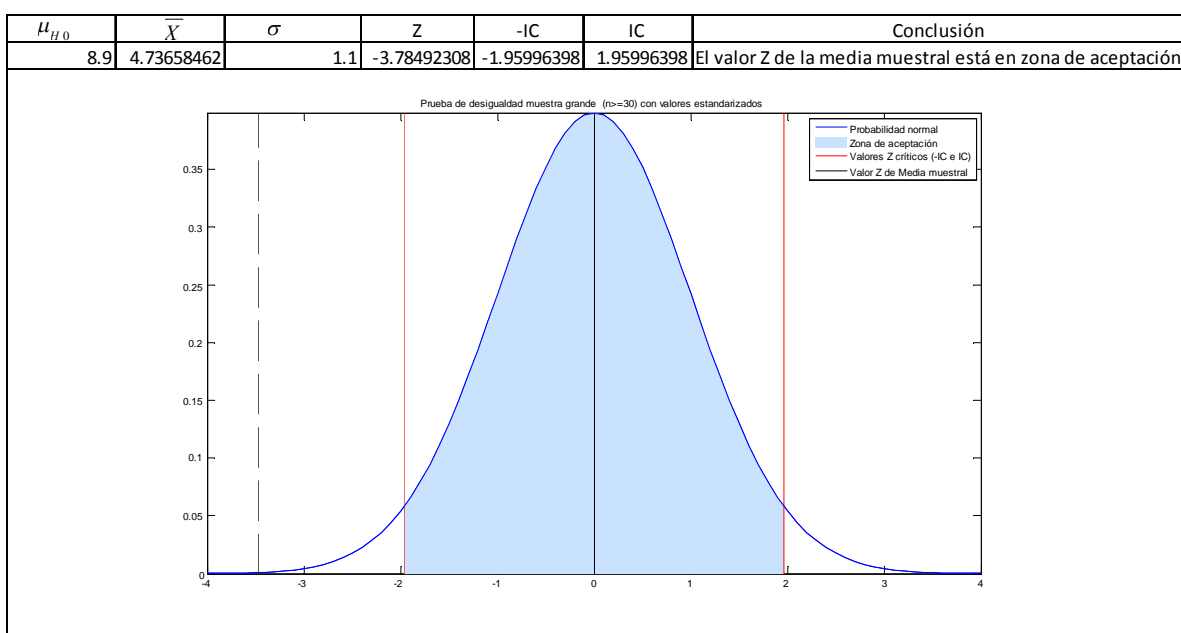


6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de desigualdad (prueba de dos colas) se definió, como zona de aceptación, a todos los valores de Z que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $-IC < Z < IC$.

Aceptar H_a : Si $Z < -IC$ o $IC < Z$.

7. **Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Se tienen los siguientes resultados:



Conclusión: Por tanto, en base a los datos que tiene la empresaria de Chicago, ella llega a la misma conclusión de la prueba de hipótesis anterior, con la diferencia de que se utilizó una escala diferente. Esto es, la calidad de la muestra es diferente al objetivo de 8.90z, por tanto debe aceptarse el embarque.

4.1.1.2.3 Prueba de hipótesis para un caso de demostración de desigualdad con una muestra pequeña con escala original.

Ahora se retoma el caso de una muestra pequeña con 15 piezas que tiene una media muestral de $\bar{X} = 5.1318$ y un objetivo de demostrar desigualdad para una media hipotética de $\mu_{H_0} = 8.90z$. Con estos datos iniciales, para este tipo de prueba de hipótesis, se siguieron los pasos que se presentan:



1. **Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) diferente a 8.9 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} \neq 8.9$$

$$H_a : \bar{X} = 8.9$$

2. **Se determina, dada la hipótesis, si es prueba de dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 2, ya que se busca demostrar una desigualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 2 | Determinar si la media de la muestra que se tiene es diferente a la de su población | Dos colas | $H_0 : \bar{X} < -IC \text{ o } IC < \bar{X}$ | $H_0 : Z_{\bar{X}} < -ZC \text{ o } ZC < Z_{\bar{X}}$ | $H_0 : t_{\bar{X}} < -tC \text{ o } tC < t_{\bar{X}}$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra pequeña, se emplea la distribución t-Student o un valor t.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 15 piezas. Por tanto, la empresaria decide utilizar un valor t que corresponda a un nivel de significancia de 2.5%, al ser esta una prueba de dos colas, debe buscar un valor en tablas que corresponda a 2.5% de probabilidad con 14 grados de libertad, lo que la lleva a un t de 2.1447.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala original, se establecen los siguientes valores críticos:

$$-IC = \mu_{H_0} + (t \cdot \sigma) = 8.9 - (2.1447 \cdot 1.1) = 6.5407$$

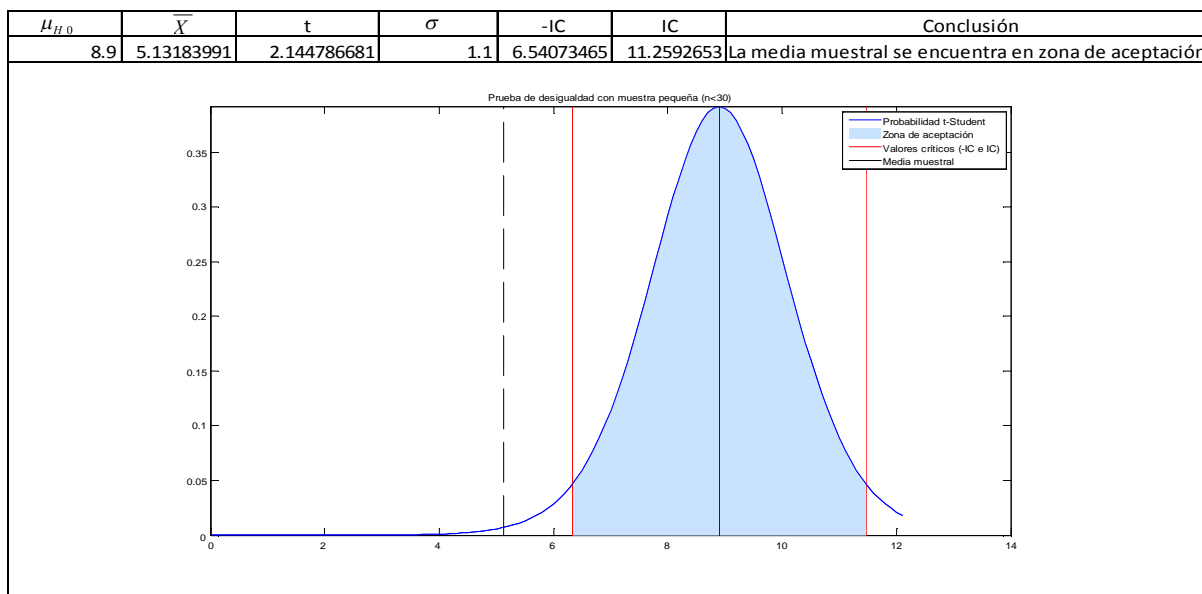
$$IC = \mu_{H_0} + (t \cdot \sigma) = 8.9 + (2.1447 \cdot 1.1) = 11.2592$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad (prueba de dos colas) se definió, como zona de aceptación, a todos los valores de t que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $\bar{X} < -IC$ o $IC < \bar{X}$.

Aceptar H_a : Si $-IC < \bar{X} < IC$.

7. **Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):**



Conclusión: Con lo anterior, la empresaria de Chicago puede aceptar el embarque enviado ya que la calidad del mismo es diferente al objetivo establecido de 8.9 Oz.

4.1.1.2.4 Prueba de hipótesis para demostración de desigualdad empleando una muestra pequeña y una escala estandarizada.

Ahora se procederá a realizar la prueba de desigualdad con muestra pequeña como la anterior pero utilizando valores estandarizados. Para ello se tienen los siguientes pasos:

- Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) diferente a 8.9 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0: \bar{X} \neq 8.9$$

$$H_a: \bar{X} = 8.9$$

- Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 1, ya que se busca demostrar una desigualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de la muestra que se tiene es igual a la de su población | Dos colas | $H_0: -IC < \bar{X} < IC$ | $H_0: -ZC < Z_{\bar{X}} < ZC$ | $H_0: -tC < t_{\bar{X}} < tC$ |



3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra pequeña, se emplea la t-Student y, por ende, se emplea un valor t.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 15 piezas. Por tanto empresaria decide utilizar un valor t que corresponda a un nivel de significancia de 2.5%. Al ser esta una prueba de dos colas, debe buscar un valor en tablas que corresponda a esto y a 14 grados de libertad, lo que le lleva a un t de 2.1447.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala estandarizada, lo que se busca es determinar los valores críticos (IC) como los valores t que corresponden a:

$$-IC = \text{Valor t de intervalo inferior 2.5\%} = -2.1447$$

$$IC = \text{Valor t de intervalo superior 97.5\%} = 2.1447$$

Ya que se tienen los valores críticos de la prueba, se procede a calcular el estadístico de prueba, en este caso un valor t dado por la fórmula 25:

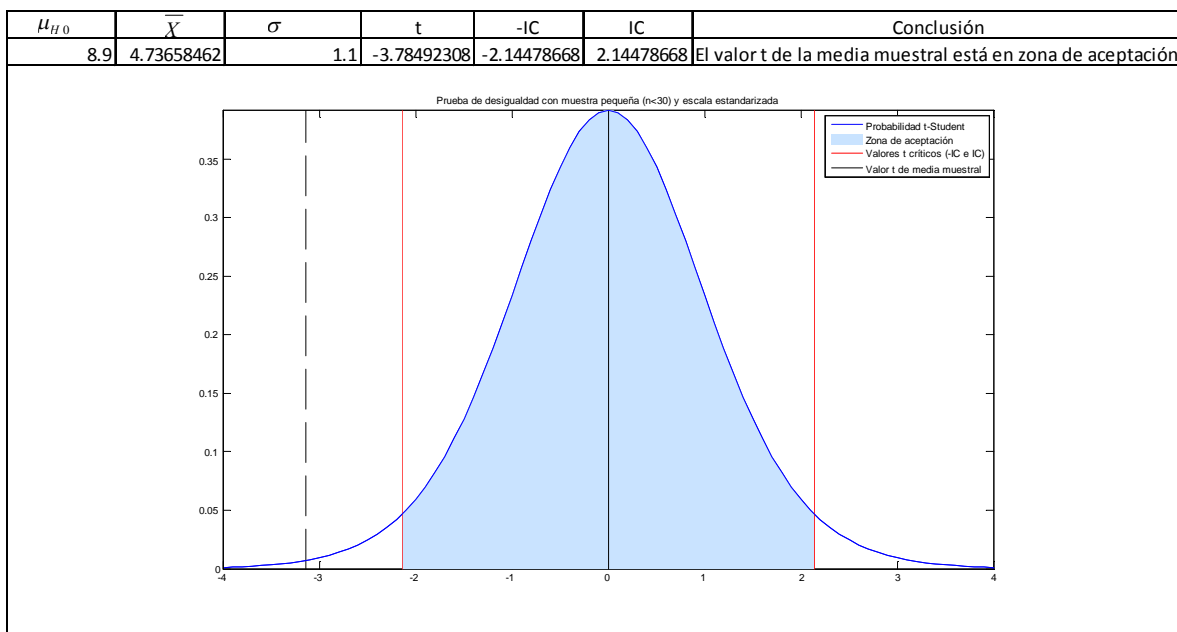
$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma} = \frac{5.1318 - 8.9}{1.1} = -3.4256$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad (prueba de dos colas) se definió, como zona de aceptación a todos los valores de t que se encuentren entre $-IC$ y IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $t < -IC$ o $IC < t$.

Aceptar H_a : Si $-IC < t < IC$.

7. **Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Se tienen los siguientes resultados:



Conclusión: Por tanto, en base a los datos que tiene la empresaria de Chicago, ella llega a la misma conclusión de la prueba de hipótesis anterior, con la diferencia de que se utilizó una escala diferente. Es decir, una estandarizada.

4.1.1.3 Ejemplos de pruebas de hipótesis de cola superior.

Ahora se harán los 4 casos previamente vistos consistentes en diferentes escalas (original o estandarizada) así como diferentes tamaños de muestra (grande y pequeña). Para ello, se seguirán manejando las mismas muestras, la misma desviación estándar poblacional de $\sigma = 1.1$. Lo único que cambiará es el objetivo de la empresaria. Ella buscará demostrar que la calidad mínima que debe tener el embarque para ser aceptado es de 2.5Oz. Es decir, hará una prueba de hipótesis de cola superior y demostrará que el peso del embarque tiene una calidad superior a 259Oz. Ahora se analizará cada caso específico:

4.1.1.3.1 Prueba de hipótesis de cola superior empleando muestra grande y escala original.

Los pasos que la empresaria siguió fueron:

1. **Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) mayor a 2.5 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:



$$H_0 : \bar{X} > 2.5$$

$$H_a : \bar{X} \leq 2.5$$

2. **Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de una cola establecida con la hipótesis señalada con ID 3, ya que se busca demostrar una prueba de cola superior:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 3 | Determinar si la media de la muestra es superior a la de su población | Cola superior | $H_0 : IC < \bar{X}$ | $H_0 : ZC < Z_{\bar{X}}$ | $H_0 : tC < t_{\bar{X}}$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de una cola (cola superior), debe buscar un valor Z en tablas que corresponda a 95% de probabilidad. Esto le lleva a un valor Z de 1.6448.
5. **Se define si se trabaja con la escala original o con una estandarizada:** En este ejemplo, la empresaria decidió trabajar con la escala original por lo que utilizó el valor Z para definir el valor crítico superior (IC) del intervalo de confianza con los que aceptará o rechazará la hipótesis. Esto la llevó a determinar el siguiente valor crítico:

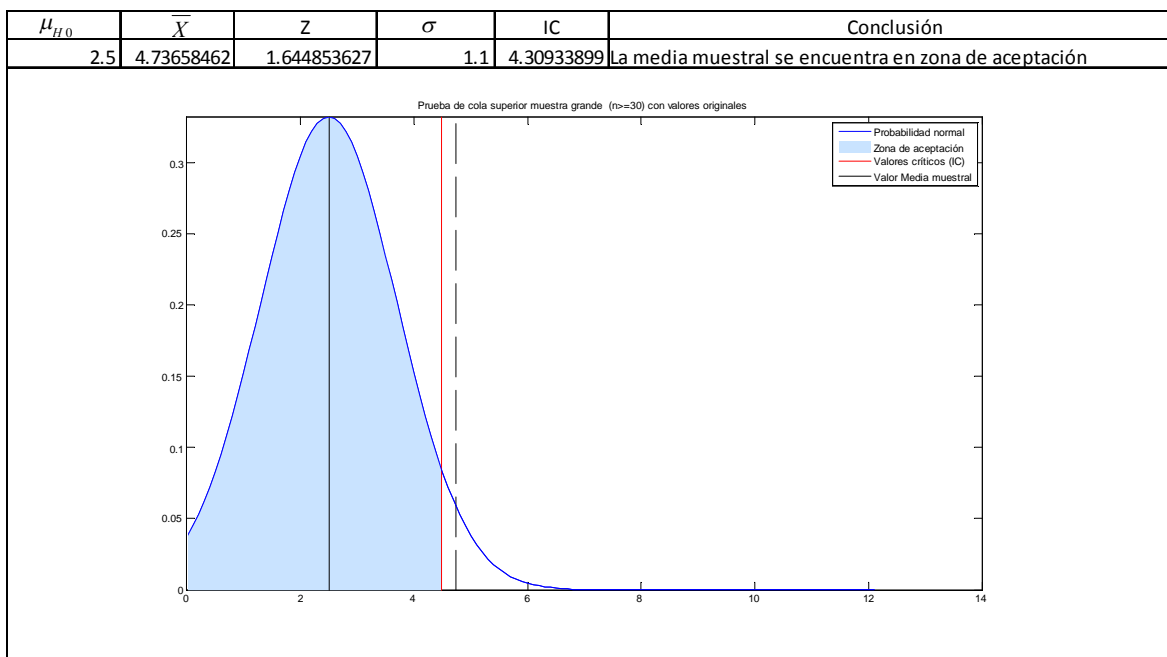
$$IC = \mu_{H_0} + (Z \cdot \sigma) = 2.5 + (1.6448 \cdot 1.1) = 4.3093$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de cola superior, se definió, como zona de aceptación, a todos los valores de \bar{X} que se encuentren arriba de IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $IC < \bar{X}$.

Aceptar H_a : Si $\bar{X} \leq IC$.

7. **Se comparan el valor crítico fijado con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Con esto, se tienen los siguientes resultados:



Conclusión: En base a los datos que tiene la empresaria de Chicago, ella puede concluir que el embarque de 5,000 aguacates cumple con los estándares de calidad que tiene establecidos ya que la media de una muestra aleatoria de 30 aguacates es estadísticamente superior al peso objetivo planteado de μ_{H_0} .

4.1.1.3.2 Prueba de hipótesis de cola superior empleando muestra grande y escala estandarizada.

Ahora se hará la misma prueba de hipótesis de cola superior con muestra grande empleando una escala estandarizada. Para ello se siguieron estos pasos:

- Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) mayor a 2.5 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} > 2.5$$

$$H_a : \bar{X} \leq 2.5$$

- Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 3, ya que se busca demostrar una prueba de cola superior:



| D | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|---|---|----------------|--|---|--|
| 3 | Determinar si la media de la muestra es superior a la de su población | Cola superior | $H_0 : IC < \bar{X}$ | $H_0 : ZC < Z_{\bar{X}}$ | $H_0 : tC < t_{\bar{X}}$ |

- Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
- Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de una cola (cola superior), debe buscar un valor Z en tablas que corresponda a 95% de probabilidad. Esto le lleva a un valor Z de 1.6448.
- Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala estandarizada, lo que se busca es determinar el valor crítico (IC) del intervalo superior como el valor Z que corresponde a:

$$IC = \text{Valor Z de intervalo superior 95\%} = 1.6448$$

Ya que se tienen los valores críticos de la prueba, se procede a tomar el estadístico de prueba calculado previamente. En este caso un valor Z dado por la fórmula 24:

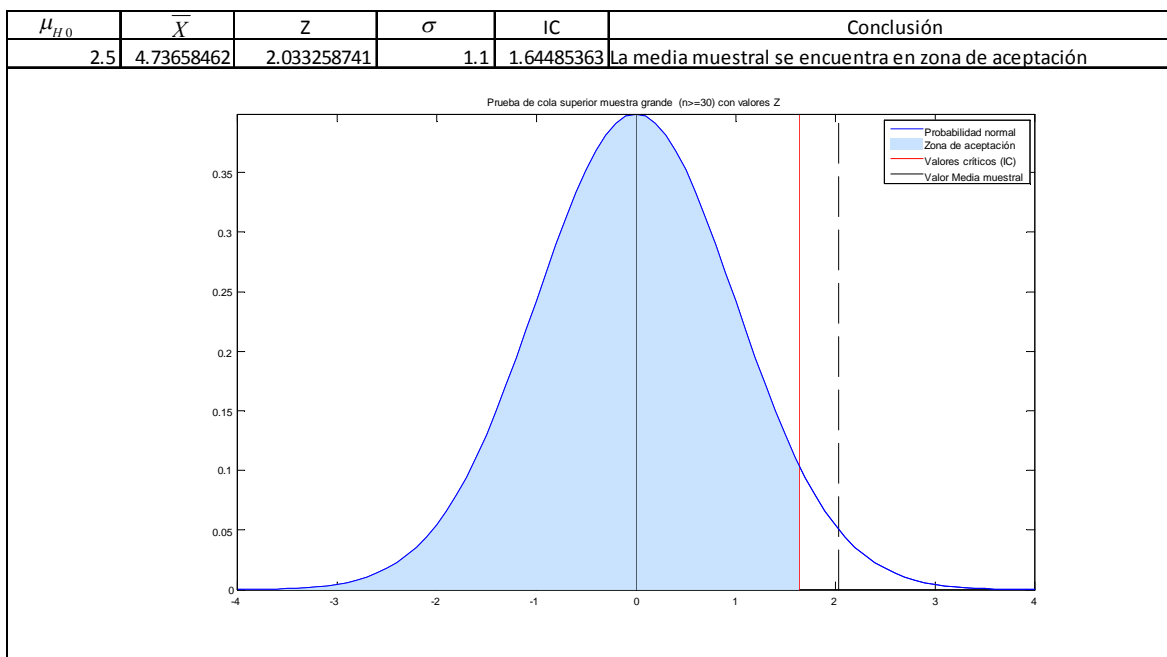
$$Z = \frac{\bar{X} - \mu_{H_0}}{\sigma} = \frac{4.7365 - 2.5}{1.1} = 2.0332$$

- Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de cola superior se definió, como zona de aceptación, a todos los valores de \bar{X} que se encuentren arriba de IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $IC < Z$.

Aceptar H_a : Si $Z \leq IC$.

- Se compara el valor crítico fijado con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Se tienen los siguientes resultados:



Conclusión: Por tanto, en base a los datos que tiene la empresaria de Chicago, ella llega a la misma conclusión de la prueba de hipótesis anterior, con la diferencia de que se utilizó una escala diferente.

4.1.1.3.3 Prueba de hipótesis de cola superior con muestra pequeña y escala original.

Ahora se tomará el caso de la muestra pequeña (15 aguacates) con su media muestral de 5.1318 Oz y la misma desviación estándar poblacional de 1.1 Oz. En este caso se trabajará con la escala original. Para ello, la empresaria siguió estos pasos:

- Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) mayor a 2.5 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} > 2.5$$

$$H_a : \bar{X} \leq 2.5$$

- Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de una cola establecida con la hipótesis señalada con ID 3, ya que se busca demostrar una prueba de cola superior:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 3 | Determinar si la media de la muestra es superior a la de su población | Cola superior | $H_0 : IC < \bar{X}$ | $H_0 : ZC < Z_{\bar{X}}$ | $H_0 : tC < t_{\bar{X}}$ |



3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra pequeña, se emplea la distribución t-Student y, por ende, se emplea un valor t.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 15 piezas. Por tanto, la empresaria decide utilizar un valor t que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de una cola (cola superior), debe buscar un valor t en tablas que corresponda a 5% de probabilidad y 14 grados de libertad. Esto le lleva a un valor t de 1.7613.
5. **Se define si se trabaja con la escala original o con una estandarizada:** En este ejemplo, la empresaria decidió trabajar con la escala original por lo que utilizó el valor Z para definir el valor crítico superior (IC) del intervalo de confianza con el que aceptará o rechazará la hipótesis. Esto la llevó a determinar el siguiente valor críticos:

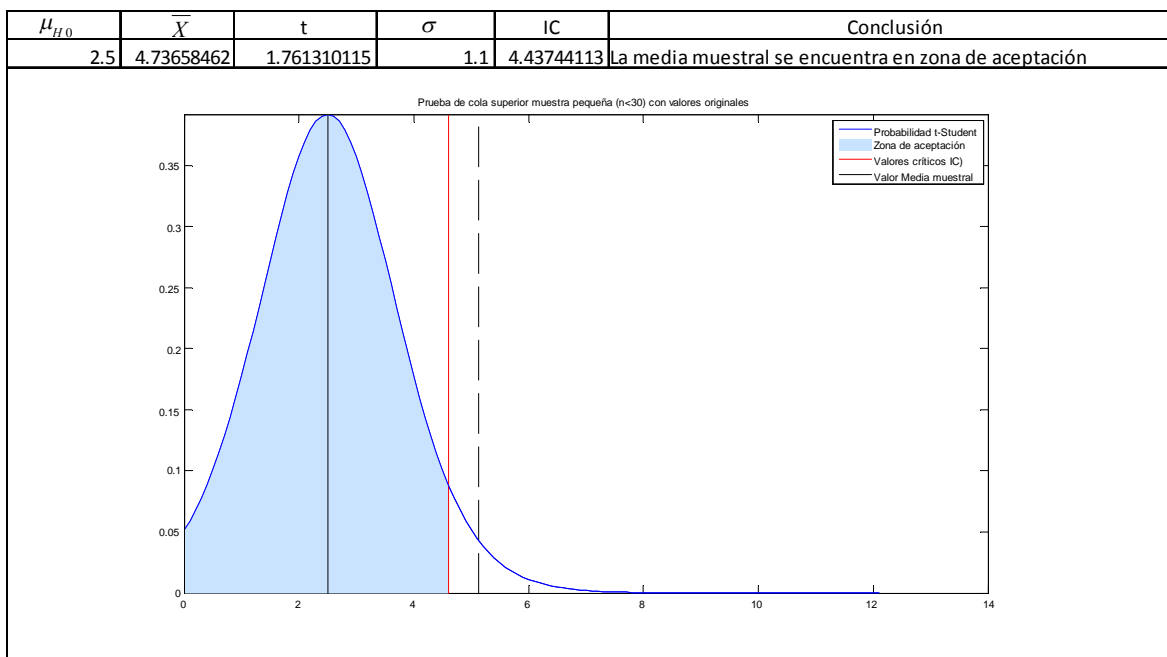
$$IC = \mu_{H_0} + (t \cdot \sigma) = 2.5 + (1.7613 \cdot 1.1) = 4.4374$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de cola superior se definió, como zona de aceptación, a todos los valores de \bar{X} que se encuentren arriba de IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $IC < \bar{X}$.

Aceptar H_a : Si $\bar{X} \leq IC$.

7. **Se compara el valor crítico fijado con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Con esto, se tienen los siguientes resultados:



Conclusión: Con lo anterior la empresaria de Chicago puede aceptar el embarque enviado ya que la calidad del mismo es superior al objetivo de calidad establecido de 2.5 Oz.

4.1.1.3.4 Prueba de hipótesis de cola superior con muestra pequeña y escala de estandarizada.

- Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) mayor a 2.5 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{X} > 2.5$$

$$H_a : \bar{X} \leq 2.5$$

- Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de una cola establecida con la hipótesis señalada con ID 3, ya que se busca demostrar una prueba de cola superior:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 3 | Determinar si la media de la muestra es superior a la de su población | Cola superior | $H_0 : IC < \bar{X}$ | $H_0 : ZC < Z_{\bar{X}}$ | $H_0 : tC < t_{\bar{X}}$ |

- Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra pequeña, se emplea la distribución t-Student y, por ende, se emplea un valor t.



4. **Se define el grado de significancia:** La muestra con que se trabaja es de 15 piezas. Por tanto, la empresaria decide utilizar un valor t que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de una cola (cola superior), debe buscar un valor t en tablas que corresponda a 5% de probabilidad y 14 grados de libertad. Esto le lleva a un valor t de 1.7613.
5. **Se define si se trabaja con la escala original o con una estandarizada:** En este ejemplo, la empresaria decidió trabajar con la escala original por lo que utilizó el valor Z para definir el valor crítico superior (IC) del intervalo de confianza con los que aceptará o rechazará la hipótesis. Esto la llevó a determinar al siguiente valor crítico:

$$IC = 1.7613$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de cola superior se definió, como zona de aceptación, a todos los valores de \bar{X} que se encuentren arriba de IC . Esto lleva a la siguiente regla de aceptación:

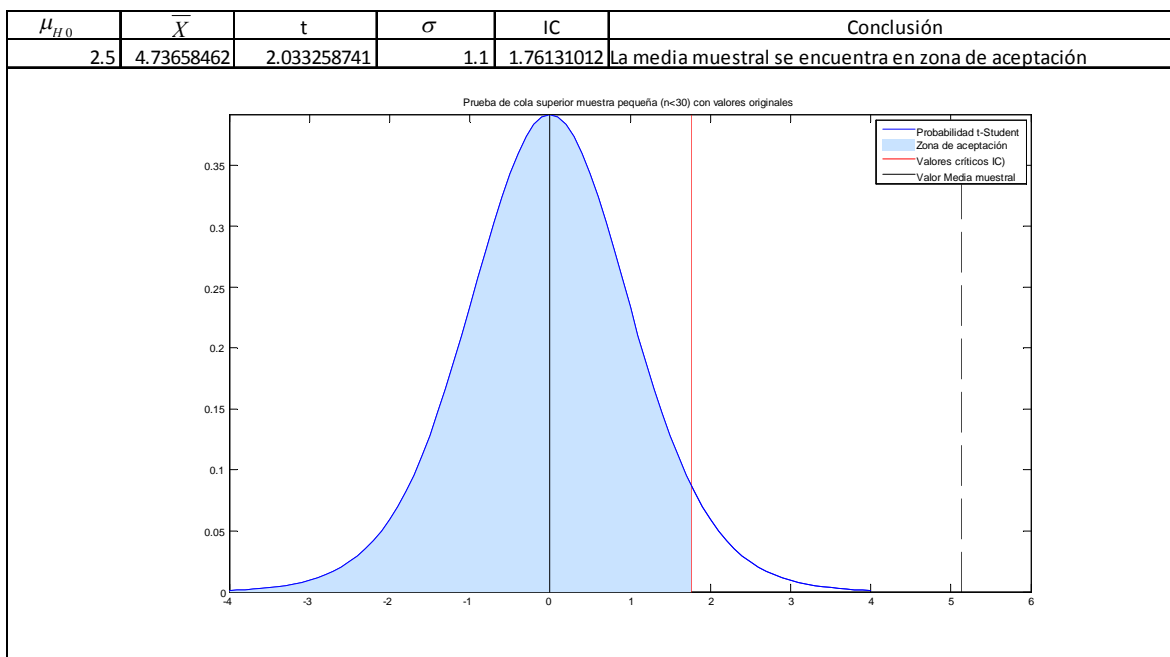
Aceptar H_0 : Si $IC < \bar{X}$.

Aceptar H_a : Si $\bar{X} \leq IC$.

Ya que se tienen los valores críticos de la prueba, se procede a tomar el estadístico de prueba calculado previamente. En este caso un valor Z dado por la fórmula 25:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma} = \frac{4.7365 - 2.5}{1.1} = 2.0332$$

7. **Se compara el valor crítico fijado con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):** Con esto, se tienen los siguientes resultados:



Conclusión: Se llega a la misma conclusión del caso anterior, con la diferencia de que se empleó una escala estandarizada para realizar la prueba.

4.1.1.4 Ejemplos de pruebas de hipótesis de cola inferior.

Para el caso de la prueba de hipótesis de cola inferior se tiene la misma lógica de análisis que las pruebas de cola inferior, con la diferencia de que las reglas de decisión para aceptar la hipótesis nula (H_0) se dan por el renglón o ID 3 de la tabla 16:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 4 | Determinar si la media de la muestra es inferior a la de su población | Cola inferior | $H_0: \bar{X} < -IC$ | $H_0: Z_{\bar{X}} < -ZC$ | $H_0: t_{\bar{X}} < -tC$ |

Por cuestión de espacio, no se hará una exposición extensa de los cuatro posibles casos dado el tamaño de muestra (grande o pequeña) y tipo de escala (original o estandarizada). Lo que se hará será simplemente citar un ejemplo consistente en una muestra grande con escala estandarizada. Usted podrá observar que los pasos a seguir en los otros tres casos son similares a las 4 pruebas previas de cola superior. Solo cambiará, como se ha mencionado, la regla de aceptación de H_0 .

Para exponer este ejemplo, se supondrá ahora que la empresaria no quiere aguacates con un peso menor a 7 Oz, por lo que deberá demostrar que el embarque que ha recibido se ajusta a dicho estándar de calidad. Para ello, siguió estos pasos:



1. **Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El embarque de aguacates recibido tiene una calidad (peso) menor a 7 Oz”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0: \bar{X} < 7$$

$$H_a: \bar{X} \geq 7$$

2. **Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 16, que se utiliza una prueba de hipótesis de una cola establecida con la hipótesis señalada con ID 3, ya que se busca demostrar una prueba de cola inferior:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 4 | Determinar si la media de la muestra es inferior a la de su población | Cola inferior | $H_0: \bar{X} < -IC$ | $H_0: Z_{\bar{X}} < -ZC$ | $H_0: t_{\bar{X}} < -tC$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de una cola (cola inferior), debe buscar un valor Z en tablas que corresponda a 95% de probabilidad. Esto le lleva a un valor Z negativo de -1.6448.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Dado que ahora se trabaja con escala estandarizada, lo que se busca es determinar el valor crítico (IC) del intervalo inferior como el valor Z que corresponde a:

$$-IC = \text{Valor Z de intervalo inferior 95\%} = -1.6448$$

6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de cola superior se definió, como zona de aceptación, a todos los valores de \bar{X} que se encuentren arriba de IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $Z < -IC$.

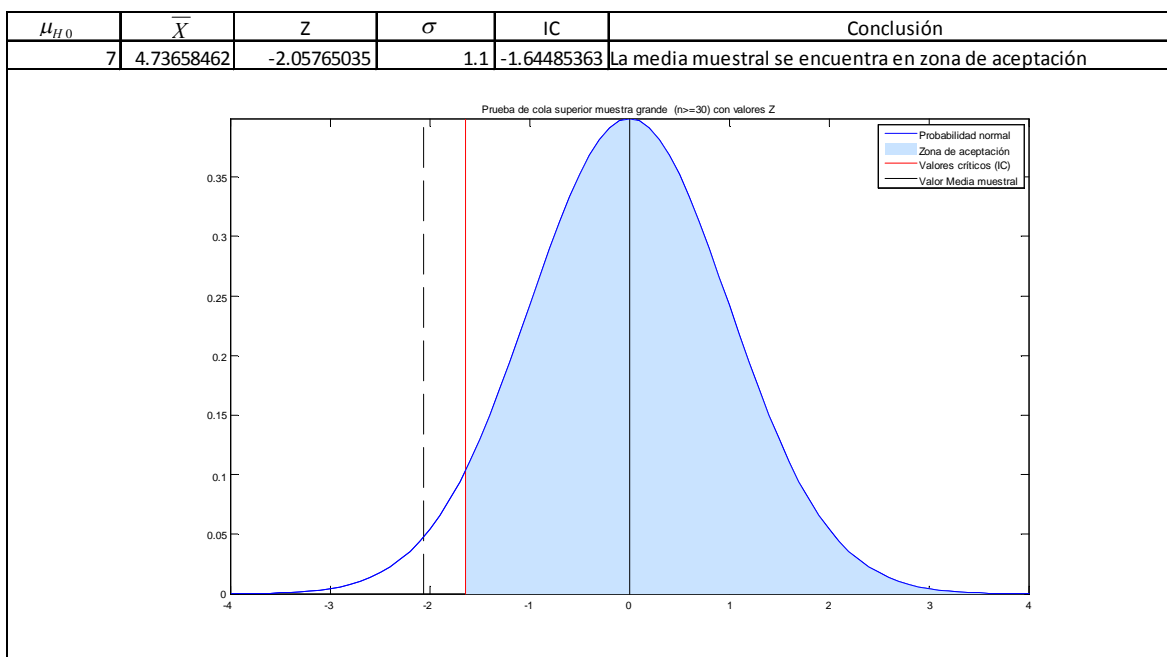
Aceptar H_a : Si $-IC \leq Z$.

Ya que se tienen los valores críticos de la prueba, se procede a tomar el estadístico de prueba calculado previamente. En este caso un valor Z dado por la fórmula 24:

$$Z = \frac{\bar{X} - \mu_{H0}}{\sigma} = \frac{4.7365 - 7}{1.1} = -2.0576$$



7. Se compara el valor crítico fijado con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):



Conclusión: Con los datos de la muestra de 30 piezas, se observa que el embarque tiene una calidad o peso inferior a 7 Oz, por lo que podrá aceptar el mismo.

4.2 ¿Cuándo se utiliza la escala original y cuándo la estandarizada?

Hasta el momento se ha observado que una de las variantes que puede tener la comprobación de hipótesis se refiere a la escala empleada. Esta puede ser la escala original o la escala estandarizada, lograda al aplicar las fórmulas 24 o 25 según sea el tamaño de la muestra. Sin embargo, poco se ha dicho sobre el criterio para utilizar una escala u otra. En realidad, no existen reglas generales para decidir. Más bien la selección se da en función del tipo de problema y las preferencias del analista.

Sin embargo, algo que puede ser de utilidad para elegir la escala estandarizada es el hecho de que ésta sirve para homologar escalas. Por ejemplo, la variable estandarizada, como veremos en breve, sirve más para comparar inventarios con variabilidades de peso diferentes. Tal es el caso de los empresarios aguacateros de Morelia y Chicago. Por tanto, la selección de la escala es netamente personal a inherente al analista.



4.3 ¿Qué se hace cuando se desconoce la desviación estándar poblacional?

Hasta ahora se ha trabajado bajo el supuesto de que se conoce la desviación estándar poblacional o se supone una partiendo de la experiencia propia. Por ejemplo, la empresaria de Chicago partió de lo que ha observado con su proveedor, en el sentido de fijar la desviación estándar del peso de aguacates que ha recibido a lo largo de la historia como de 1.1Oz. Sin embargo, no siempre se conoce este valor por lo que debe calcularse para poder determinar los estadísticos Z o t.

Cuando el tamaño de la muestra es grande ($n \geq 30$), se calcula el error estándar con la fórmula 8:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Cuando el tamaño de la muestra es pequeña ($n < 30$), se calcula el error estándar como en la fórmula 11:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (x_i - \mu)^2}{n-1}}}{\sqrt{n}}$$

Ya que se calculó el error estándar, éste se sustituye en el cálculo del estadístico Z o t en las fórmulas 24 o 25, según sea el tamaño de muestra:

$$Z = \frac{\bar{X} - \mu_{H0}}{\sigma_{\bar{X}}}$$

$$t = \frac{\bar{X} - \mu_{H0}}{\sigma_{\bar{X}}}$$

Este estadístico se utiliza para determinar los valores críticos de la prueba de hipótesis (IC) con las fórmulas de las estimaciones de intervalo si se trabaja con la escala original o se emplea directamente para definir el estadístico de prueba de la hipótesis. También sirve para determinar los estadísticos Z y t si se utiliza una prueba con escala estandarizada.



4.4 Pruebas de hipótesis para comparar muestras.

En este caso específico ya no se busca demostrar que los parámetros de una muestra se ajustan a los de una población o a un valor hipotético dado por μ_{H_0} . Más bien se busca comparar su media, buscando probar que estas sean iguales, mayor en la muestra A respecto a la B o viceversa. Las posibles pruebas de hipótesis y las reglas de aceptación de H_0 para los posibles casos se presentan en la siguiente tabla:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de una muestra es igual a la de otra | Dos colas | $H_0: -DC < \bar{D} < DC$ | $H_0: -ZC < Z_{\bar{D}} < ZC$ | $H_0: -tC < t_{\bar{D}} < tC$ |
| 2 | Determinar si la media de una muestra es diferente a la de otra | Dos colas | $H_0: \bar{D} < -DC \text{ o } DC < \bar{D}$ | $H_0: Z_{\bar{D}} < -ZC \text{ o } ZC < Z_{\bar{D}}$ | $H_0: t_{\bar{D}} < -tC \text{ o } tC < t_{\bar{D}}$ |
| 3 | Determinar si la media de una muestra es superior a la de otra | Cola inferior | $H_0: DC < \bar{D}$ | $H_0: ZC < Z_{\bar{D}}$ | $H_0: tC < t_{\bar{D}}$ |
| 4 | Determinar si la media de una muestra es inferior a la de otra | Cola superior | $H_0: \bar{D} < -DC$ | $H_0: Z_{\bar{D}} < -ZC$ | $H_0: t_{\bar{D}} < -tC$ |

Tabla 17 Reglas de aceptación de la hipótesis nula aplicada a la prueba de hipótesis de diferencias de muestras.

Para ilustrar el empleo de este tipo de prueba de hipótesis, tómese la idea original que tenían los dos empresarios aguacateros de comparar la calidad de sus inventarios, con la finalidad de saber si el proveedor, que es el mismo para ambos, les vende la misma calidad. Lo primero que hacen los empresarios es homologar sus escalas de medida, por lo que deciden trabajar con onzas. Posteriormente, deciden probar la calidad estableciendo como objetivo determinar que la calidad que ambos reciben es la misma.

Para poder establecer el ejercicio, recuerde usted el tema de estimaciones de intervalo de diferencias. En concreto, que las muestras pueden ser independientes o relacionadas. Para este caso, dado que ambos reciben aguacates del mismo proveedor y este puede determinar qué calidad mandar a cada cliente, se supondrá que la muestra está relacionada, acoplada o apareada. Es decir, las muestras no son independientes. Por tanto, deben calcularse diferencias entre los valores de cada observación de cada muestra. En este ejemplo se comparará, como referencia, el inventario de la empresaria de Chicago para el que se realizan los cálculos de diferencias entre observaciones de muestra como sugiere la fórmula 14:

$$D_i = x_{a,i} - x_{b,i} = \text{Peso de aguacate } i \text{ en inventario Chicago} - \text{Peso de aguacate } i \text{ en inventario Morelia}$$

Posterior a ello se calculan la media muestral de las diferencias y el error estándar siguiendo el criterio de muestra acoplada:

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n} = -0.6932$$



$$\sigma_{\bar{D}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{\sum (D_i - \bar{D})^2}{n}}}{\sqrt{n}} = 0.8928$$

Dado que se desconoce la desviación estándar poblacional y no se tiene un dato que se pueda conocer en base a la experiencia, el cual sea de utilidad para aproximar como desviación estándar poblacional, se emplea el error estándar anterior como tal.

| La comerciante de Chicago (muestra grande) | | | |
|--|------------|----------|------------|
| Aguacate | Peso (g) | Aguacate | Peso (g) |
| 1 | 9.24156638 | 16 | 6.88943702 |
| 2 | 12.199554 | 17 | 2.80526622 |
| 3 | 7.82891212 | 18 | 1.26336648 |
| 4 | 0.34339132 | 19 | 0.32189422 |
| 5 | 5.44698856 | 20 | 0.68982456 |
| 6 | 4.30814406 | 21 | 5.14666579 |
| 7 | 2.77853698 | 22 | 0.93461195 |
| 8 | 4.78302773 | 23 | 6.68022055 |
| 9 | 2.09992446 | 24 | 3.11979453 |
| 10 | 4.27460843 | 25 | 3.4652034 |
| 11 | 4.69181648 | 26 | 3.79602008 |
| 12 | 2.93915238 | 27 | 5.12093613 |
| 13 | 5.27263691 | 28 | 8.52528354 |
| 14 | 8.44771181 | 29 | 11.7979333 |
| 15 | 2.32162703 | 30 | 4.56348205 |

| El comerciante de Morelia (muestra grande) | | | |
|--|------------|----------|-------------|
| Aguacate | Peso (g) | Aguacate | Peso (g) |
| 1 | 6.02999294 | 16 | 4.630940879 |
| 2 | 6.56210303 | 17 | 11.49065308 |
| 3 | 6.73041637 | 18 | 12.11979191 |
| 4 | 8.08539167 | 19 | 14.93468555 |
| 5 | 5.12702893 | 20 | 8.008499815 |
| 6 | 3.45800988 | 21 | 3.382731702 |
| 7 | 4.48129852 | 22 | 6.325966735 |
| 8 | 4.1213832 | 23 | 1.246781817 |
| 9 | 3.5109386 | 24 | 1.4895245 |
| 10 | 3.79640085 | 25 | 0.879262348 |
| 11 | 3.96400847 | 26 | 2.183030375 |
| 12 | 3.83787579 | 27 | 3.842599967 |
| 13 | 6.90813581 | 28 | 10.89484725 |
| 14 | 4.18758912 | 29 | 7.111256588 |
| 15 | 1.70988128 | 30 | 1.844722109 |

| Tabla de diferencias entre muestras | | | |
|-------------------------------------|-------------|----------|-------------|
| Aguacate | Peso (g) | Aguacate | Peso (g) |
| 1 | 3.21157344 | 16 | 2.25849614 |
| 2 | 5.63745093 | 17 | -8.68538685 |
| 3 | 1.09849575 | 18 | -10.8564254 |
| 4 | -7.74200035 | 19 | -14.6127913 |
| 5 | 0.31995963 | 20 | -7.31867525 |
| 6 | 0.85013418 | 21 | 1.76393409 |
| 7 | -1.70276153 | 22 | -5.39135478 |
| 8 | 0.66164453 | 23 | 5.43343873 |
| 9 | -1.41101414 | 24 | 1.63027003 |
| 10 | 0.47820758 | 25 | 2.58594105 |
| 11 | 0.72780801 | 26 | 1.6129897 |
| 12 | -0.89872341 | 27 | 1.27833617 |
| 13 | -1.6354989 | 28 | -2.3695637 |
| 14 | 4.26012269 | 29 | 4.68667673 |
| 15 | 0.61174575 | 30 | 2.71875994 |

| Valores estadísticos para prueba de hipótesis(diferencias) | |
|--|--------------|
| Media de diferencias | -0.693273688 |
| Error estándar de diferencias | 0.892810705 |
| Estadístico Z de diferencias | -0.776506917 |
| Valor crítico superior para la prueba | 1.959963985 |
| Valor crítico inferior para la prueba | -1.959963985 |

Ya que se tienen estas medidas estadísticas, se procede a realizar la prueba de hipótesis.



1. **Definir una hipótesis nula a demostrar:** La hipótesis a demostrar sería: “El inventario de aguacates de ambos comerciantes es igual”. Esto se representa con la siguiente hipótesis nula a demostrar y su alternativa:

$$H_0 : \bar{D} = 0$$

$$H_a : \bar{D} \neq 0$$

2. **Se determina, dada la hipótesis, si es prueba dos colas, cola superior y cola inferior:** Aquí es importante observar, siguiendo las recomendaciones de la tabla 17, que se utiliza una prueba de hipótesis de dos colas establecida con la hipótesis señalada con ID 1, ya que se busca demostrar una prueba de igualdad:

| ID | Objetivo | Tipo de prueba | Regla de aceptación con escala original (muestra pequeña o grande) | Regla de aceptación con escala estandarizada (muestra grande) | Regla de aceptación con escala estandarizada (muestra pequeña) |
|----|---|----------------|--|---|--|
| 1 | Determinar si la media de una muestra es igual a la de otra | Dos colas | $H_0 : -DC < \bar{D} < DC$ | $H_0 : -ZC < Z_{\bar{D}} < ZC$ | $H_0 : -tC < t_{\bar{D}} < tC$ |

3. **Se determina la función de probabilidad a utilizar:** En este caso, al ser muestra grande, se emplea la gaussiana (normal estándar) y, por ende, se emplea un valor Z.
4. **Se define el grado de significancia:** La muestra con que se trabaja es de 30 piezas. Por tanto, la empresaria decide utilizar un valor Z que corresponda a un nivel de significancia de 5%. Al ser esta una prueba de dos colas, se debe buscar un valor Z en tablas que corresponda a 97.5% de probabilidad. Esto le lleva a un valor Z de 1.9599.
5. **Se define si se trabaja con la escala original o con una estandarizada:** Se determina trabajar con escala estandarizada. Con esto, lo que se busca es determinar el valor crítico (IC) del intervalo superior como el valor Z que corresponde a:

$$IC = \text{Valor Z de intervalo superior } 97.5\% = 1.9599$$

Ya que se tienen los valores críticos de la prueba, se procede a calcular el estadístico de prueba con valor Z dado por la fórmula 24:

$$Z = \frac{\bar{X} - \mu_{H_0}}{\sigma_{\bar{D}}} = \frac{-0.6932 - 0}{0.8928} = -0.7765$$

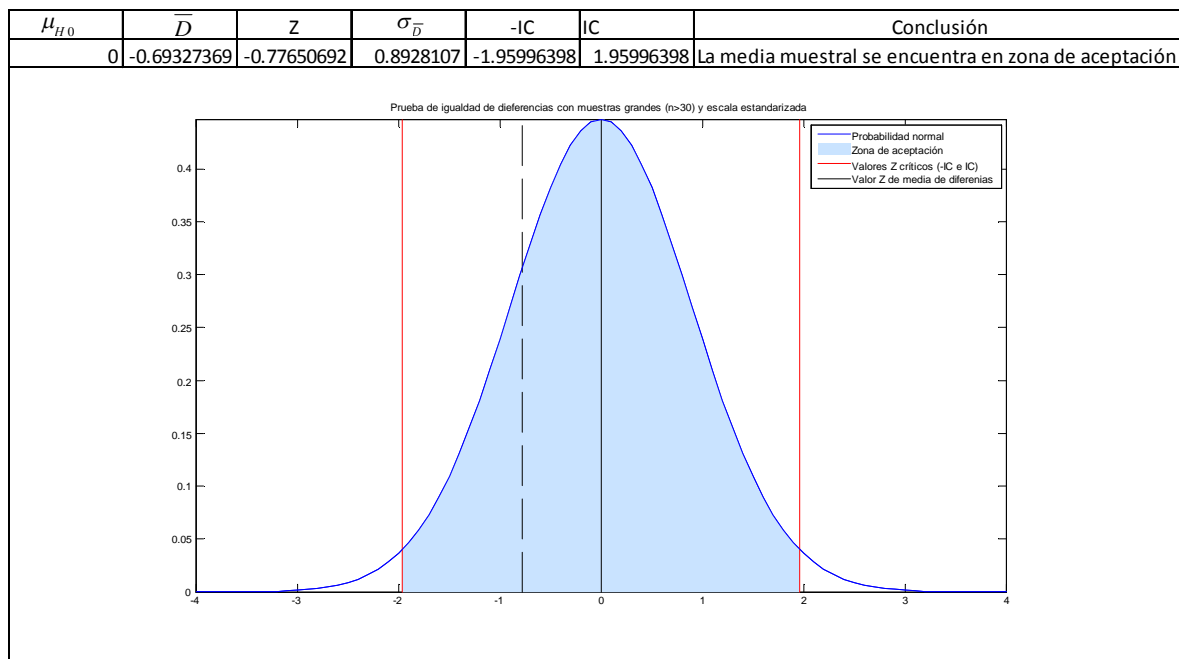
6. **Se define la regla de aceptación:** Dado que la prueba a realizar es una prueba de igualdad de dos colas, se definió, como zona de aceptación, a todos los valores de Z que se encuentren entre $-IC$ e IC . Esto lleva a la siguiente regla de aceptación:

Aceptar H_0 : Si $-IC < Z < IC$.

Aceptar H_a : Si $Z \leq -IC$ o $Z \geq IC$.



7. Se comparan los valores críticos fijados con el estadístico (media muestral) y se determina si se acepta la hipótesis nula (H_0) o se abre paso a la alternativa (H_a):



Conclusión: En base a las diferencias calculadas entre el inventario de la empresaria de Chicago y el de Morelia, se concluye que la calidad que reciben ambos de su proveedor es la misma. Por tanto, se acepta el hecho de que reciben el mismo tipo de aguacate.



5 Prueba de hipótesis: Las técnicas Ji- cuadrada y ANOVA

Hasta ahora se ha visto el caso en el que se realizan pruebas de hipótesis comprobando medias muestrales respecto a una media poblacional o una media hipotética. A su vez, se hicieron comparaciones de diferencias de medias muestrales entre dos poblaciones diferentes.

Dentro de los supuestos que se han manejado es determinar que existe dependencia o independencia en dichas poblaciones, por un lado y que están ya sea normalmente distribuidas (si se trata de una población o muestra grande) o t-Student distribuidas si se trata de una muestra pequeña.

En este tema específico utilizaremos un tipo de técnica de comprobación de hipótesis conocido como la técnica Ji-Cuadrada, la cual nos servirá para realizar dos cosas:

1. Determinar si dos o más variables o atributos de interés son independientes en base a los datos obtenidos en la muestra.
2. Determinar si el comportamiento de los datos con que se cuenta se explican o no con una distribución de probabilidad determinada como puede ser la normal, la t-Student u otro tipo de casos como son la F, la binomial, la Weibull, la Gumbel, la Poisson, la uniforme u otras.
3. Determinar si la varianza de una muestra es igual, inferior o superior a cierto valor hipotético, poblacional u objetivo.

Otro tipo de técnica de comprobación de hipótesis que se revisará será la prueba ANOVA (siglas en idioma inglés de Análisis de varianza –Analysis Of VAriance-) en la cual no se comparan medias directamente, como en la técnica clásica; sino que se contrastan las varianzas. Esta prueba permitirá, a su vez, ya no comparar solo dos muestras de manera conjunta. Más bien, ayudará a realizar lo siguiente:

1. Comparar 2 o más muestras o poblaciones al mismo tiempo con la finalidad de determinar si son o no iguales sus medias.
2. Comparar 2 o más muestras o poblaciones al mismo tiempo con la finalidad de determinar si son o no iguales o diferentes sus varianzas.

5.1 La técnica Ji-Cuadrada

5.1.1 Prueba de hipótesis para demostrar independencia.

Ahora se revisará el empleo de la técnica Ji-Cuadrada para determinar si los atributos de dos o más variables en una muestra o población son independientes o no. Por ejemplo, Steve Jobs pudo hacer una encuesta más amplia y detallada que la previamente vista en donde se asignaron calificaciones, y ahora preguntar a diferentes individuos de los cuatro segmentos o estratos



previamente estudiados¹⁴ si preferirían Mac dados algunos atributos de la misma como son rapidez de arranque, el tamaño de la computadora y la compatibilidad con Windows. Después de preguntar esto, Jobs pudo sospechar que los diferentes resultados de cada atributo se relacionan a las necesidades profesionales o personales de cada individuo en cada estrato realizado. Es entonces que pudo plantearse el cuestionamiento de si ¿Realmente los atributos estudiados y el estrato al que pertenece el individuo muestreado tienen relación cercana para determinar la preferencia de los individuos?

Este cuestionamiento se lo planteó después de aplicar las encuestas, recolectar los datos y organizar los mismos a través de una tabla de contingencia como la siguiente:

| | | Estrato profesional | | | | |
|------------------|-------------------------------------|---------------------|-----------|-----------|-----------|-----------|
| Atributo/estrato | | Estrato 1 | Estrato 2 | Estrato 3 | Estrato 4 | Total |
| Atributo | Arranque | 10 | 4 | 6 | 6 | 26 |
| | Tamaño de computadora | 2 | 11 | 4 | 8 | 25 |
| | Compatibilidad con Windows | 8 | 5 | 10 | 6 | 29 |
| | Total muestreado por Estrato | 20 | 20 | 20 | 20 | 80 |

Tabla 18 Tabla de contingencia con los tres atributos del muestreo de Jobs.

Tabla de contingencia: Tabla que contiene R renglones y C columnas. Cada renglón corresponde a un nivel de una variable; cada columna, a un nivel de otra variable. Los datos del cuerpo de la tabla son las frecuencias con que ocurre cada combinación de variables y los totales en cada extremo son la suma de esas frecuencias por renglón o por columna.

El enunciado de la hipótesis a plantear en este caso sería: ***“Las variables atributo de la computadora y estrato profesional están relacionadas entre sí y, por tanto pueden influir en la preferencia del usuario de la computadora.”***

Para fines de comprobación de hipótesis debe plantearse la hipótesis nula H_o . Para ello se define la frecuencia observada de cada combinación de atributos y estratos de la tabla 18 como sigue:

¹⁴ Estrato 1: Arquitectos ingenieros, matemáticos, físicos, investigadores y profesionistas que ocupen procesamiento de cálculo.

Estrato 2: Diseñadores gráficos, artistas de medios, músicos y gente que ocupe procesamiento gráfico.

Estrato 3: Amas de casa, estudiantes y gente mayor.

Estrato 4: Contadores, abogados, economistas, financieros y otros profesionistas.



| | | | Estrato profesional | | | | |
|------------------------------|----|----------------------------|---------------------|-----------|-----------|-----------|-------|
| Atributo/estrato | | | Estrato 1 | Estrato 2 | Estrato 3 | Estrato 4 | Total |
| Atributo | Pa | Arranque | pa1 | pa2 | pa3 | pa4 | 26 |
| | Pt | Tamaño de computadora | pt1 | pt2 | pt3 | pt4 | 25 |
| | Pc | Compatibilidad con Windows | pc1 | pc2 | pc3 | pc4 | 29 |
| Total muestreado por Estrato | | | 20 | 20 | 20 | 20 | 80 |

Tabla 19 Codificación como variables de las frecuencias observada según cada combinación de atributos y estratos en la tabla 18.

Con la codificación anterior se llega a la siguiente hipótesis nula y su alternativa:

Fórmula 23: Definición de la hipótesis nula de la prueba de hipótesis de dependencia del ejercicio de Steve jobs.

$$H_0 : p1a = pa2 = pa3 = pa4 = pt1 = pt2 = pt3 = pt4 = pc1 = pc2 = pc3 = pc4$$

$$H_a : p1a \neq pa2 \neq pa3 \neq pa4 \neq pt1 \neq pt2 \neq pt3 \neq pt4 \neq pc1 \neq pc2 \neq pc3 \neq pc4$$

Antes de iniciar, es de necesidad observar que esta prueba puede emplearse de manera inversa. Es decir, se puede emplear también para demostrar si hay independencia o no en los datos estudiados. Esto sería como sigue:

$$H_0 : pa1 \neq pa2 \neq pa3 \neq pa4 \neq pt1 \neq pt2 \neq pt3 \neq pt4 \neq pc1 \neq pc2 \neq pc3 \neq pc4$$

$$H_a : pa1 = pa2 = pa3 = pa4 = pt1 = pt2 = pt3 = pt4 = pc1 = pc2 = pc3 = pc4$$

El cambio formal u operativo en cada situación se da por la ubicación de la zona de aceptación. Para fines de este ejemplo, solo interesa demostrar que hay dependencia entre las dos variables y eso implicaría que las frecuencias observadas son estadísticamente (no numéricamente) iguales. Tal como se planteó en la fórmula

Como en toda prueba de hipótesis y al igual que el caso de la técnica clásica, se debe de trabajar con algún estadístico que preferentemente esté explicado por una función de densidad de probabilidad. En este caso se utiliza el estadístico ji-cuadrada dado por la siguiente función:

Fórmula 24: Estadístico ji-cuadrada para demostrar hipótesis.

$$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

En la misma surgen dos variables de vital importancia para esta técnica de comprobación de hipótesis. La primera de ellas es f_0 , que se refiere a la frecuencia observada en alguna celda de la tabla de contingencias. Por ejemplo, la frecuencia observada de las personas que prefieren la Mac



dado el arranque y que forman parte del estrato 1 (Arquitectos ingenieros, matemáticos, físicos, investigadores y profesionistas que ocupen procesamiento de cálculo) sería de:

$$f_{0,pa1} = 10$$

Lo mismo se determina para las frecuencias de las diferentes combinaciones de atributo y estrato. Las mismas se presentan en la tabla 18.

El segundo concepto de interés viene dado por f_e que es la frecuencia esperada para ese atributo. En este punto usted puede observar una pequeña similitud de esta técnica con la clásica. Se comparan valores observados contra valores esperados. La forma de determinar la frecuencia esperada se hace a través de las sumas de renglones (TR) y columnas (TC) y el número de observaciones (n).

Fórmula 25: Cálculo de la frecuencia esperada en cada combinación de variables o atributos.

$$f_e = \frac{TR \cdot TC}{n}$$

Siguiendo con el ejemplo de $pa1$, se tendría el siguiente cálculo de frecuencia relativa:

$$f_e = \frac{TR \cdot TC}{n} = \frac{20 \cdot 26}{80} = 6.5$$

Si se hace el cálculo de frecuencias esperadas para cada uno de los casos de interés se tiene la siguiente tabla de frecuencias esperadas:

| | | | Estrato profesional | | | | |
|----------|------------------------------|----------------------------|---------------------|-----------|-----------|-----------|-------|
| | | Atributo/estrato | Estrato 1 | Estrato 2 | Estrato 3 | Estrato 4 | Total |
| Atributo | Pa | Arranque | 6.5000 | 6.5000 | 6.5000 | 6.5000 | 26 |
| | Pt | Tamaño de computadora | 6.2500 | 6.2500 | 6.2500 | 6.2500 | 25 |
| | Pc | Compatibilidad con Windows | 7.2500 | 7.2500 | 7.2500 | 7.2500 | 29 |
| | Total muestreado por Estrato | | 20 | 20 | 20 | 20 | 80 |

Tabla 20 Tabla de frecuencias esperadas según cada combinación de atributos y estratos en la tabla 18.

Con esto se pueden sustituir ahora los valores de frecuencia observada de cada combinación de atributos de la tabla 18 y de frecuencia esperada de la tabla 20 para realizar el cálculo de la fórmula 24. Esto se detalla en la siguiente tabla de cálculos:



| Combinación de atributos | f_0 | f_e | $(f_0 - f_e)^2$ | $\frac{(f_0 - f_e)^2}{f_e}$ |
|--|-------|-------|-----------------|-----------------------------|
| pa1 | 10 | 6.5 | 12.25 | 1.88461538 |
| pa2 | 4 | 6.5 | 6.25 | 0.96153846 |
| pa3 | 6 | 6.5 | 0.25 | 0.03846154 |
| pa4 | 6 | 6.5 | 0.25 | 0.03846154 |
| pt1 | 2 | 6.25 | 18.0625 | 2.89 |
| pt2 | 11 | 6.25 | 22.5625 | 3.61 |
| pt3 | 4 | 6.25 | 5.0625 | 0.81 |
| pt4 | 8 | 6.25 | 3.0625 | 0.49 |
| pc1 | 8 | 7.25 | 0.5625 | 0.07758621 |
| pc2 | 5 | 7.25 | 5.0625 | 0.69827586 |
| pc3 | 10 | 7.25 | 7.5625 | 1.04310345 |
| pc4 | 6 | 7.25 | 1.5625 | 0.21551724 |
| $X^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$ | | | | 12.7575597 |

Tabla 21 Cálculo del estadístico ji-cuadrada para el ejemplo de la prueba de hipótesis de dependencia de Steve Jobs.

Ya que se tiene este estadístico ji-cuadrada, lo que corresponde realizar es el contraste del mismo con una zona de aceptación y de rechazo en el contexto de una distribución de probabilidad. Observe usted que no se trata de un contraste de técnica clásica con una distribución normal o una t-Student; sino un estadístico ji-cuadrada (o X^2). Por lo tanto, la distribución de probabilidad de interés es una que se denomina **distribución de probabilidad ji-cuadrada**. Ahora veremos de qué se trata la misma.

5.1.2 Distribución de probabilidad ji-cuadrada.

Recuerde usted que las distribuciones normal (gaussiana) y t-Student son simétricas e incluyen valores de probabilidad tanto a la izquierda como a la derecha del cero. Estas distribuciones de probabilidad son muy útiles para muchos fenómenos como los revisados. Sin embargo, cuando se trata con valores que solo son positivos, será de mucho interés tener una distribución de probabilidad que **nunca** tenga valores negativos en su distribución de probabilidad. Por ejemplo, el precio de una acción nunca tendrá valores negativos pero, en repetidas ocasiones de nuestros ejercicios, las estimaciones de intervalo llevaban a límites inferiores negativos que no era lógico que existan. Por ejemplo, podríamos tener la estimación de intervalo del precio de una acción, las calificaciones de un grupo de clase o el peso de los aguacates dadas por:



| | |
|--------------------|----|
| Límite superior | 6 |
| Estimación puntual | 2 |
| Límite inferior | -1 |

Observe cómo se tiene un límite inferior negativo en los valores del peso de aguacates. Claramente esto no es posible en términos lógicos. Sin embargo, usted hizo caso omiso en las estimaciones de intervalo y en la prueba de hipótesis de técnica clásica por que simplemente sustituyó el -1 por 0 y aproximó sus intervalos como sigue:

| | |
|--------------------|---|
| Límite superior | 6 |
| Estimación puntual | 2 |
| Límite inferior | 0 |

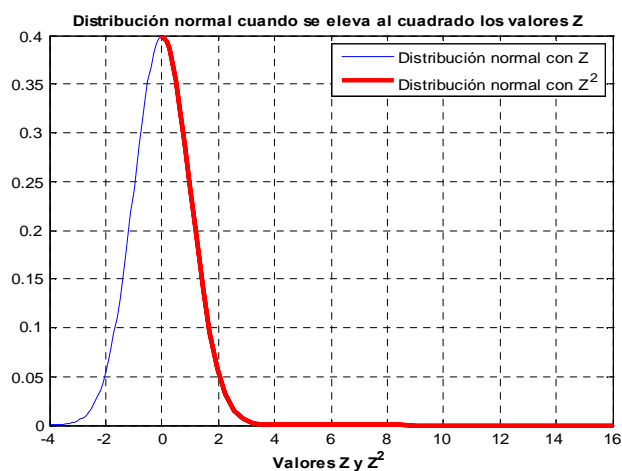
Ahora, hay casos en que el problema estudiado no nos permite darnos el lujo de tener valores negativos en los intervalos. Para estos casos se utilizan otro tipo de funciones de probabilidad como es el caso de la ji-cuadrada. De hecho esta es de las más socorridas para hacer estimaciones de intervalos y pruebas de hipótesis para actividades como son las siguientes:

- Determinar el número de personas que entran a un supermercado en una determinada hora (estará de acuerdo que no se pueden hacer estimaciones de intervalo que digan “entran -30 personas”).
- Aplicaciones de administración de riesgos crediticios como son, determinar el monto de pérdida potencial que un banco puede tener por hacer préstamos de tarjeta de crédito o definir ¿cuál es la probabilidad de que nuestra cartera de clientes acreditados nos haga perder X cantidad de dinero por incumplimiento de pago?
- Determinar si el número de conexiones a nuestra página por medio internet de internet es igual a un número objetivo.
- Determinar que la varianza de nuestros datos es igual, inferior o superior a un valor objetivo (no existen varianzas negativas de ahí la necesidad de determinar probabilidades únicamente de valores positivos).
- Las que, de momento, nos interesan más:
 - Determinar si dos variables, dada la frecuencia de la combinación de ambas en una tabla de contingencia, son dependientes o independientes.
 - Determinar si un conjunto de datos se distribuye o explica con una función de densidad de probabilidad determinada (prueba de bondad de ajuste).
 - Determinar si la varianza de una población, inferida a partir de una muestra de datos, se ajusta a un objetivo preestablecido.

¿De dónde surge la distribución ji-cuadrada? ¿cómo se determina de manera intuitiva? Recuerde usted que no nos permitimos tener números negativos en la distribución de probabilidad por



tanto ¿Qué pasa si, en una distribución normal estándar elevamos los valores Z al cuadrado? Vea usted qué sucede.



Gráfica 29 Generación intuitiva de la distribución χ^2 a partir de la normal.

Recuerde usted que los valores Z se obtienen de estandarizar los valores de x_i :

$$Z = \frac{x_i - \mu}{\sigma}$$

Por lo tanto la distribución de probabilidad de la línea gruesa se obtiene de elevar al cuadrado a x_i . Como un dato cultural que le será de mucho interés, la letra χ tiene una letra equivalente en el griego y esta letra se llama “ji” en ese alfabeto. Por tanto, si usted eleva al cuadrado x_i , tiene una x_i^2 (ji-cuadrada). De ahí viene el nombre de la distribución que nos interesa. Sin embargo, la derivación de esta distribución de probabilidad es mucho más sólida que esta mera explicación intuitiva. Su desarrollo y explicación a la luz de la matemática estadística sale de la óptica de las presentes notas del profesor. Sin embargo, solo para fines de ilustración, se presenta su fórmula de cálculo:

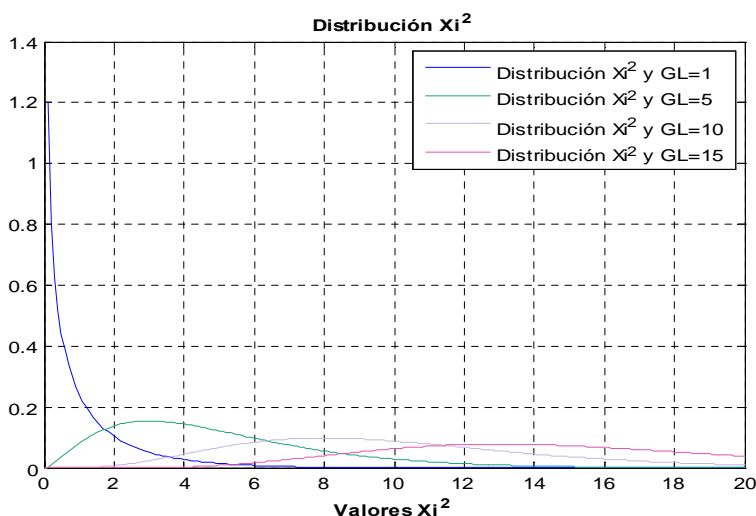
Fórmula 26: Función de densidad de probabilidad ji-cuadrada.

$$p(x_i^2) = \frac{x_i^{(v-2)} \cdot e^{-x_i/2}}{2^{v/2} \Gamma(v/2)}$$

Note usted cómo, a parte de los valores x_i , se emplean los grados de libertad (GL o v). El ¿por qué? lo veremos en breve de manera intuitiva. Baste ahora con observar que, adicional a los datos, se ocupan ciertos grados de libertad para determinar la función de probabilidad. Para



ilustrar el impacto de los mismos, véase la siguiente gráfica donde se emplean 1, 5 y 15 grados de libertad (GL o ν):



Gráfica 30 Cálculo de la distribución ji-cuadrada con diferentes grados de libertad.

Note usted cómo, conforme se incrementan los grados de libertad, la distribución ji-cuadrada se parece más a una normal o al menos es simétrica. La única diferencia de interés está en que se tiene una distribución de probabilidades en la que ninguno de los valores modelados es negativo, lo que asegura nuestro interés original expresado líneas atrás.

Ahora, aquí también se emplean grados de libertad como en la distribución t-Student. Sin embargo, la forma de calcularlos es diferente. En esta distribución se determinan como sigue:

Fórmula 27: Determinación de los grados de libertad en la función de densidad de probabilidad ji-cuadrada.

$$\nu = (\text{número de renglones}-1) \cdot (\text{número de columnas}-1)$$

La razón intuitiva de esto se puede identificar fácilmente a partir de la tabla de contingencia del ejemplo de Steve Jobs que nos interesa:

| | | Estrato profesional | | | | Total |
|------------------------------|----------------------------|---------------------|-----------|-----------|-----------|-------|
| Atributo/estrato | | Estrato 1 | Estrato 2 | Estrato 3 | Estrato 4 | |
| Pa | Arranque | pa1 | pa2 | pa3 | pa4 | 26 |
| Pt | Tamaño de computadora | pt1 | pt2 | pt3 | pt4 | 25 |
| Pc | Compatibilidad con Windows | pc1 | pc2 | pc3 | pc4 | 29 |
| Total muestreado por Estrato | | 20 | 20 | 20 | 20 | 80 |

Gráfica 31 Determinación de los grados de libertad con la tabla de contingencias del ejemplo de Steve Jobs.

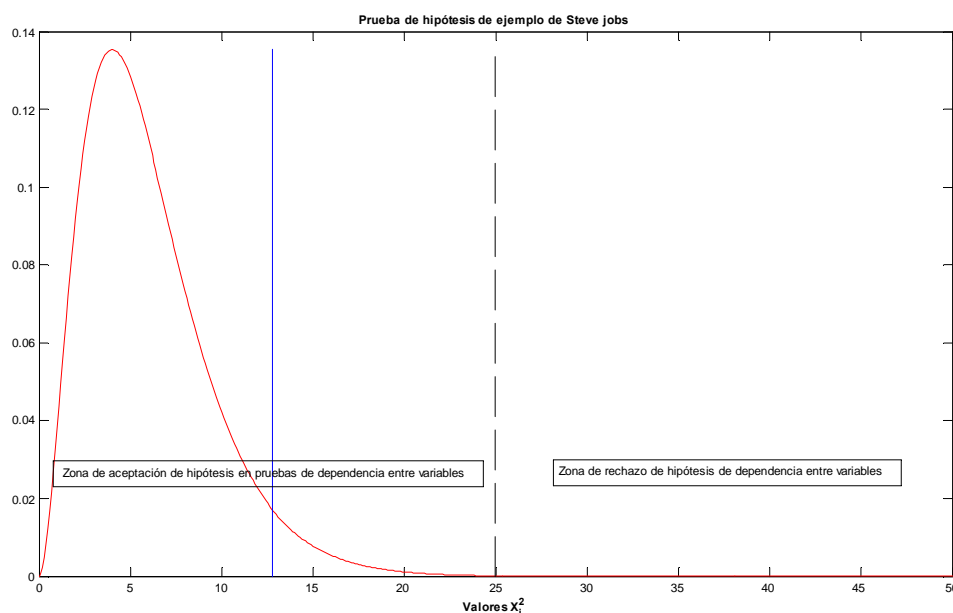


Para identificar el número de grados de libertad en la primera columna observe cómo pa1 y pt1 tienen valores de elección libre para usted. Sin embargo, el valor de pc1 ya no es libre porque su magnitud es fundamental para que la suma de pa1, pt1 y pc1 sea de 20. Lo propio sucede en las otras dos columnas. Ahora, si usted replica el ejercicio en los renglones y marca con amarillo las celdas que deben mantenerse sin cambio para que resulten los valores de la sumatoria por columnas o por renglones, observará que solo tiene usted seis valores de libre elección (pa1, pa2, pa3, pt1, pt2 y pt3). Si usted tiene una tabla de contingencias más complicada, puede utilizar la fórmula 26 y llegar al mismo resultado:

$$\nu = (\text{número de renglones}-1) \cdot (\text{número de columnas}-1)$$

$$\nu = (3-1) \cdot (4-1) = 2 \cdot 3 = 6$$

Por lo tanto, usted puede calcular valores críticos o χ^2_i (el equivalente a valores t o z) partiendo de un nivel de significancia deseado como puede ser, a manera de ejemplo, $\alpha=5\%$ y con los 6 grados de libertad del ejemplo de Jobs, usted llega a un valor $\chi^2_{i,\alpha,\nu} = 24.99$ puede elaborar la siguiente gráfica con un valor crítico de rechazo representado con una línea punteada:



Gráfica 32 Prueba de hipótesis de dependencia entre la variable atributo y estrato en el ejemplo de la preferencia por Mac hecho por Steve Jobs.

Note usted las siguientes situaciones de interés potencial:

1. La media global de frecuencia observada de las ocho combinaciones de variable (atributo y estrato, es decir, pa1, pa2, pa3, pa4, pt1, etc.) es de 6.667. O sea una frecuencia media (o



esperada) global de 6.66. Esta se aprecia claramente como el valor de mayor probabilidad en la distribución (note usted cómo los valores esperados de cada celda o combinación de variables es muy similar a este variable en la tabla 20).

2. Ahora, lo que se busca es determinar qué tan separada está la frecuencia de observaciones de preferencia en cada combinación de atributo y estrato respecto a un valor esperado determinado con la fórmula 25.

$$f_e = \frac{TR \cdot TC}{n}$$

Esta separación en cada caso se suma. Si la misma se encuentra un valor menor o igual a esa media global, se puede aceptar el hecho de que las variables están estrechamente relacionadas e influyen en la preferencia que una persona tiene por la Mac, dado el atributo y el estrato. Es decir, los profesionistas que requieren capacidad de cálculo (ingenieros, matemáticos, etc.) prefieren un arranque rápido y compatibilidad con Windows en relación a las amas de casa.

3. Ahora, si la diferencia es muy grande (es estadísticamente superior a la media de frecuencias por preferencia global en cada combinación de atributo y estrato), se deberá rechazar a la hipótesis de dependencia entre variables utilizando el estadístico Ji determinado con la fórmula 24:

$$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Que, para el ejemplo estudiado, es de:

$$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = 12.7575$$

En base al razonamiento anterior, la prueba que debe de realizarse es una prueba de cola superior y establecer la siguiente regla general de aceptación:

- Aceptar H_0 si $X^2 < x_{i,\alpha,v}^2$.
- Rechazarla en caso contrario.

En base a los cálculos hasta ahora desarrollados con la aplicación de la fórmula 24 al ejemplo y el valor crítico determinado con un nivel de significancia de 5% (se busca como 0.95 en la tabla de valores ji-cuadrada) y 6 grados de libertad que, para este caso es de $x_{i,5\%,6}^2 = 24.9958$, se tiene el siguiente resultado (que se aprecia claramente en la gráfica 32):



$$X^2 < x_{i,5\%,6}^2$$
$$12.7575 < 24.9958$$

En base al contraste de hipótesis realizado, se tienen elementos estadísticos suficientes para aceptar la hipótesis nula de que la variable estrato profesional y la variable atributo de la computadora están estrechamente relacionadas para determinar la preferencia del individuo por una Mac. Por lo tanto, Steve Jobs aplicó una encuesta adecuada y enfocó una adecuada estrategia de desarrollo, producción y marketing para vender una computadora que satisfaga las necesidades de diferentes personas y que tuviera una mayor demanda.

5.1.3 Algunas consideraciones a tomar con la prueba ji-cuadrada.

Existen dos situaciones que deben tenerse presente al momento de realizar pruebas con la técnica ji-cuadrada para los tres usos que interesan:

1. Nunca se deben trabajar con tablas de contingencia que tengan frecuencias menores a 5. Es decir, que el valor de una celda sea menor a 5. Si en algún momento se presentara este caso en dos o más celdas, podemos eliminar algunas categorías (renglón o columna) y combinar los valores de la (s) eliminada (s) con otra que esté en el mismo caso y así lograr frecuencias mayores o iguales a 5. Sin embargo, esto tiene la limitante de la pérdida de una o varias categorías y el examen de independencia entre variables quedaría muy parcial.
2. Si, por alguna circunstancia, el valor ji-cuadrada derivado con la fórmula 24 diera cero, debe sospecharse del resultado ya que se puede estar en presencia de un problema de una inapropiada recolección de datos.

5.1.4 Prueba de hipótesis ji cuadrada para bondad de ajuste (determinar la función de probabilidad a emplear en un grupo de datos).

Ahora se revisará uno de los usos más comunes que tiene la prueba con distribución ji-cuadrada: Determinar si el comportamiento de un grupo de datos se explica con alguna función de densidad determinada.

Como ha visto hasta ahora, el análisis de prueba con técnica ji-cuadrada se enfoca a trabajar con tablas de contingencias en donde se presentan las diferentes **frecuencias** observadas en las diferentes combinaciones de variables. Sin embargo, cuando se tiene una serie de datos, no siempre se puede hacer una tabla de frecuencias, salvo que sea el histograma, de los datos debido a que se deben fijar clases o intervalos discrecionales. Existen algunas otras distribuciones de probabilidad como la binomial, la Poisson u otras que se enfocan a eventos aleatorios discretos (Recuerde usted la definición correspondiente) y en estas se puede hacer un análisis de bondad de ajuste con técnica ji-cuadrada que no difiere mucho del anteriormente realizado. Si desea

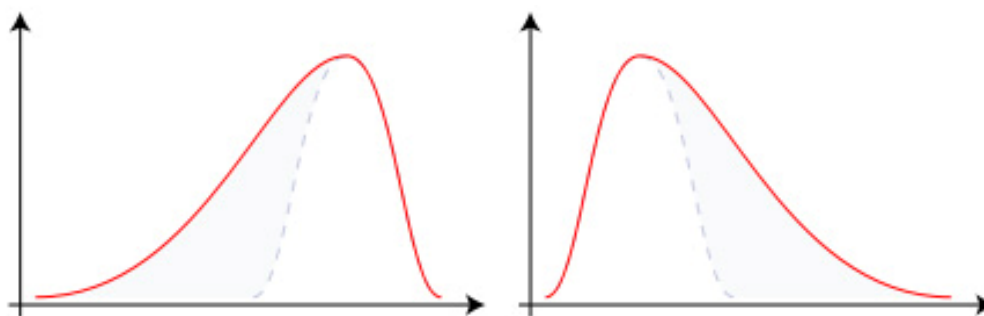


profundizar en el tema de prueba de bondad de ajuste de diferentes distribuciones de probabilidad puede usted consultar el libro de Levin y Rubin (2004, págs. 462-465) el cual resulta una excelente introducción al tema. El mismo no se desarrolla en las presentes notas debido a que la lógica de razonamiento podría salirse, de manera observable, de la lógica de presuponer que los datos con que se trabaja, están normal o t-Student distribuidos.

En estas notas nos limitaremos, con la finalidad de solo sensibilizarle al empleo de la distribución de probabilidad ji-cuadrada, a determinar si los datos con que se trabaja están o no normalmente distribuidos. Para ello, se utilizará el cálculo de un estadístico elaborado ampliamente utilizado en la Econometría y el análisis estadístico y el cuál fue elaborado por un Mexicano (Carlos Jarque¹⁵) y Anil Bera. El mismo se conoce como estadístico Jarque-Bera.

La lógica del estadístico Jarque-Bera parte de dos conceptos fundamentales inherentes a una función de probabilidad como es la normal:

- Sesgo: Se refiere a que la media y mediana no son iguales y, por lo tanto, la forma de la distribución de probabilidad normal no es igual de simétrica a la que debería esperarse:



Gráfica 33 Comparativo de sesgo negativo y positivo en una distribución de probabilidad normal respecto a su forma teórica correcta.

Note usted cómo la gráfica de la distribución de probabilidad no es tan simétrica como debería de ser (línea gris punteada) y tiene una cola más larga a la izquierda o a la derecha. Cuando la cola es más larga a la izquierda se dice que se tiene un **sesgo negativo** ya que es mayor la probabilidad de tener valores más negativos que positivos. En caso contrario, se dice que se tiene un **sesgo positivo** (Gráfica de la derecha) por que se tienen mayores probabilidades de tener valores más positivos que negativos.

¹⁵ Actuario y economista mexicano que ha sido representante del Banco Interamericano de Desarrollo en Europa y asesor del mismo, director de estudios económicos de Telmex, director del INEGI, director de Estadística en la ONU, Secretario del Plan Nacional de Desarrollo y Secretario de Desarrollo Social. Es uno de los científicos más citados de nuestro país.



Recuerde usted que una distribución de probabilidad normal debe ser **simétrica**, es decir, que se tengan la misma cantidad y probabilidades en los valores tanto positivos como negativos. Por lo tanto, si usted calcula las probabilidades de sus datos y la distribución de probabilidad tiene sesgo positivo o negativo, se tiene un claro indicio de que los datos pueden no ser normalmente distribuidos.

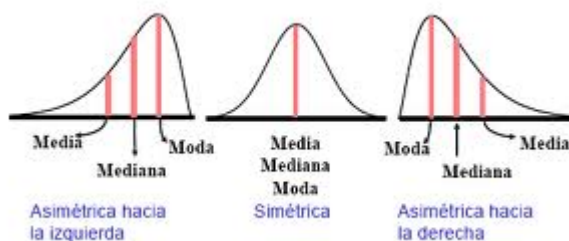
El sesgo de la distribución de probabilidad de los datos se calcula como sigue:

Fórmula 28: Determinación del sesgo en una función de probabilidad normal

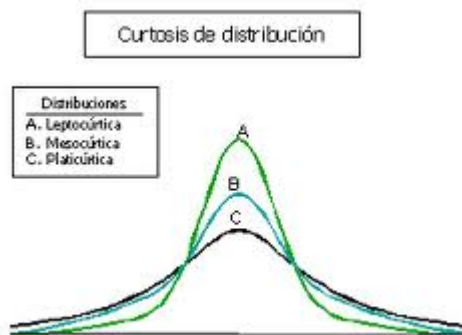
$$sesgo = \frac{\sum_{i=1}^n (x_i - \mu)^3 / n}{\sigma^2}$$

El valor apropiado del sesgo debe de ser de cero. Si el valor del sesgo, como previamente se vio, es positivo, se tiene un caso como el de la derecha en la gráfica 33. Si es negativo, se tiene el caso de la izquierda y los valores negativos son más probables de tenerse.

El término sesgo también puede conocerse como asimetría ya que hace alusión a esta situación observada en la distribución de probabilidad. El origen del sesgo o asimetría en una función de probabilidad normal se origina cuando la media, la mediana y la moda no son iguales. Esto se puede ilustrar a continuación:



- **Kurtosis (también se escribe en español como curtosis):** La kurtosis (o curtosis) se refiere a la propiedad del tamaño de las colas que tiene la distribución de probabilidad. Para ilustrar la idea se tiene la siguiente gráfica que compara diferentes niveles de kurtosis resultantes del tamaño de las colas de probabilidad:



Una función de probabilidad normal estándar debe tener la forma de la función de probabilidad b (mesokúrtica) ya que las probabilidades que se lograrían con los diferentes valores de x_i se apegan a la descripción de la función de probabilidad descrita en la fórmula 5. Cuando se tiene el caso de una probabilidad leptokúrtica (con colas más cortas que la distribución normal) se observa que las probabilidades de suceso se concentran en los valores cercanos a la media y le dan poca probabilidad (menor en relación a la distribución normal) a los valores más extremos a la derecha y a la izquierda.

Cuando se tiene el caso contrario (una distribución platikúrtica), los valores más extremos tienen mayor probabilidad de suceso que el caso de una normal.

Como se ha visto, una distribución normal debe ser mesokúrtica. Si no es así, la probabilidad que modela el comportamiento de los datos no es gaussiana. Por lo tanto debe calificarse el grado de kurtosis con la siguiente medida:

Fórmula 29: Determinación de la kurtosis:

$$kurtosis = \frac{\sum_{i=1}^n (x_i - \mu)^4 / n}{\sigma^2}$$

Si la kurtosis que se logra es de 3, la distribución que modela los datos es mesokúrtica y se puede suponer que es una normal. Si es menor a 3, la distribución es leptokúrtica (colas más cortas que la normal) y, en caso contrario, es platikúrtica (colas más largas de lo habitual).

Ya que se revisaron los conceptos de sesgo y kurtosis, se está en posibilidad de derivar el estadístico Jarque-Bera. Este se determina con la siguiente función:

Fórmula 30: Cálculo del estadístico Jarque-Bera:



$$JB = n \left[\frac{\text{sesgo}^2}{6} + \frac{(\text{kurtosis} - 3)^2}{24} \right]$$

Para mostrar la idea que se busca exponer respecto a este estadístico se retoma el ejemplo de la comerciante de Chicago para determinar si la función de probabilidad que explica el comportamiento de los datos de peso de su inventario es una normal. Como una pista inicial, se observa que no existen valores negativos (no existe, por ejemplo, un peso de -3.8 Oz.) por lo que una distribución como la ji-cuadrada podría ser más apropiada.

Sin embargo, se revisan los datos y se determinan los cálculos de sesgo, kurtosis y estadístico Jarque-Bera que lleva a un valor de 179.1553. Esto se aprecia en la siguiente ilustración:

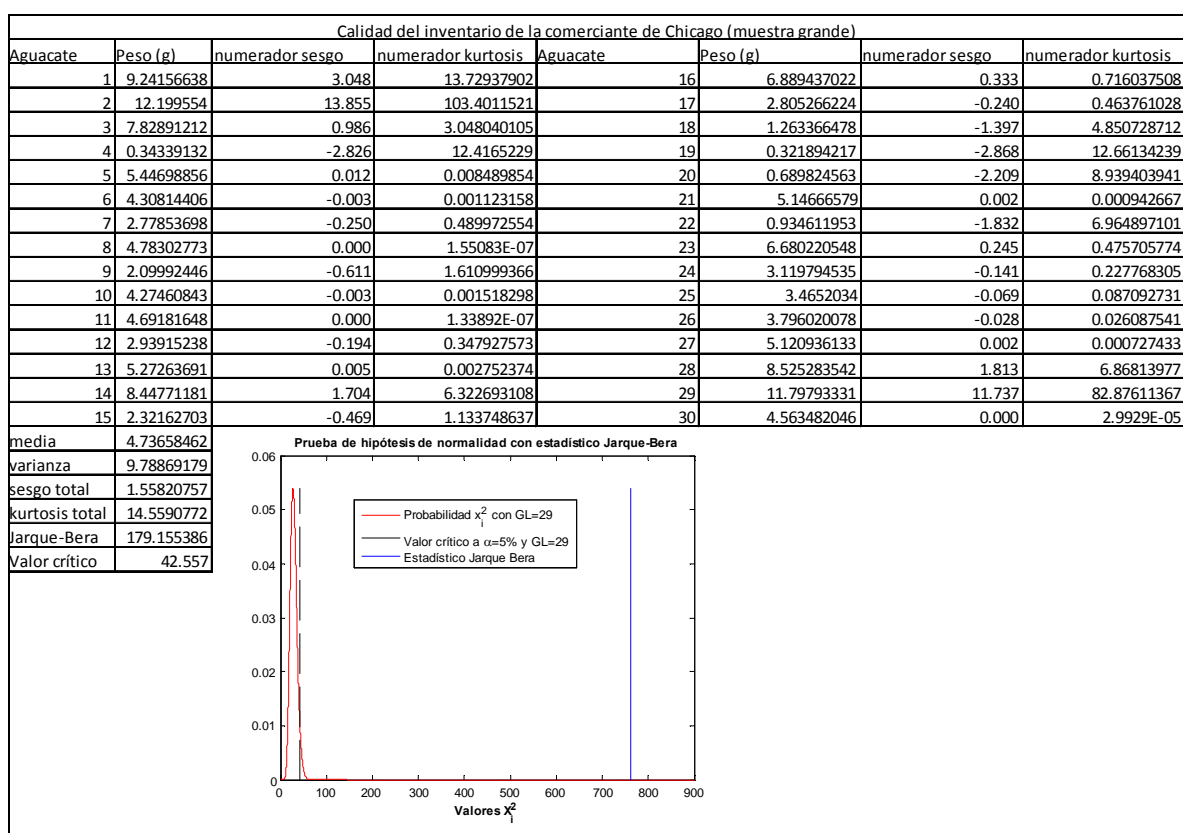


Ilustración 3 Resultado de aplicar la prueba de hipótesis de bondad de ajuste a la distribución normal con la técnica ji-cuadrada.

En la misma se aprecian las columnas tituladas “numerador sesgo” y “numerador kurtosis” que indican los términos $(x_i - \mu)^3 / n$ y $(x_i - \mu)^4 / n$ de las fórmulas 28 y 29 respectivamente. En la parte de abajo se calculan con estos términos la kurtosis y sesgo totales (de todos los datos de la muestra), así como el valor del estadístico Jarque-Bera con la fórmula 30).



Si se determina el valor crítico ji-cuadrado con 5% de significancia (95% en tabla ji-cuadrada) y 29 grados de libertad (en breve veremos cómo se calcularon los mismos) se tiene un valor crítico de 42.5570. La regla de aceptación o rechazo para la hipótesis de normalidad es igual que la de dependencia previamente revisada:

- “Se acepta la hipótesis de normalidad en los datos si el estadístico Jarque-Bera es menor al valor crítico dado el nivel de significancia y grados de libertad”.
- “Se rechaza la hipótesis de normalidad en los datos en caso contrario”.

Ya para concluir el tema de prueba de bondad de ajuste a una distribución normal con el estadístico Jarque-Bera se tiene que el cálculo de los grados de libertad es igual que en el caso de una distribución t-Student:

Fórmula 31: Determinación de los grados de libertad en la prueba de bondad de ajuste con la técnica ji-cuadrada

$$GL = \nu = n - 1$$

5.1.5 Prueba de hipótesis ji-cuadrada para hacer inferencias sobre la varianza de una sola población (o muestra).

En este tipo de prueba de hipótesis se determinará si la varianza que se calcula en una muestra de datos o población es igual, mayor o menor a algún nivel de varianza objetivo predeterminado ($\sigma_{H_0}^2$). La lógica de la prueba de hipótesis es muy similar al método de valores Z o t empleado en la técnica clásica. Es decir se debe determinar un estadístico ji-cuadrada que en breve se delimitará y extraer un valor ji-cuadrada tanto para el intervalo superior como el inferior (valores críticos), según el caso que aplique (si es prueba de una o dos colas).

Otra diferencia obvia y fundamental de la prueba ji-cuadrada es que no emplea una distribución normal o t-Student para fijar los mencionados valores críticos.

Para exponer la forma de realizar pruebas de hipótesis relativas a varianzas en una población se tiene de nuevo el ejemplo de la empresaria de Chicago. Anteriormente buscó demostrar que la calidad promedio de un embarque que le mandaron era igual a 3.8 Oz con un nivel de significancia (α) de 5%. Sin embargo se percató de que, a pesar de esto, la variabilidad del peso de cada aguacate dentro de la muestra respecto a su media muestral es muy alta. Por tanto, ahora buscó refinar su análisis y decir “Ok el inventario tiene una calidad promedio igual a la buscada. Sin embargo, tal vez, por ser un promedio, no se aprecie bien que hay piezas demasiado pesadas y otras demasiado ligeras por lo que la calidad buscada no es uniforme”. Esta falta de uniformidad es aproximada con la varianza de los pesos. Entonces, para no tener problemas con sus clientes (lo estadounidenses son muy exigentes con la calidad de su comida), la empresaria de Chicago busca



que ahora el embarque tenga una varianza de máximo 1.21 que se da por la desviación estándar que ella tiene por experiencia previa con su proveedor ($\sigma_{H_0}^2 = (1.1)^2$). Por tanto, lo que debe ella hacer es una prueba de hipótesis de cola inferior para demostrar que ahora la varianza es la que se ajusta a los objetivos establecidos. Por ejemplo, puede establecer la siguiente hipótesis nula y alternativa:

H_0 : "El embarque recibido tiene una varianza menor a 1.21 Oz."

H_a : "El embarque recibido tiene una varianza mayor a 1.21 Oz."

Para poder determinar esto, la empresaria calculó el siguiente estadístico ji-cuadrada:

Fórmula 32: Determinación del estadístico ji-cuadrada para pruebas de hipótesis de varianzas de una sola muestra.

$$X^2 = \frac{(n-1)s^2}{\sigma_{H_0}^2}$$

Recuerde que es la varianza de muestra pequeña empleada con la distribución t-Student:

$$s^2 = \frac{\sum (x_i - \mu)^2}{n-1}$$

Sustituyendo los valores del ejercicio, se llegó al siguiente estadístico:

$$X^2 = \frac{(30-1) \cdot 9.7886}{1.21} = 234.61$$

Ahora, se busca en tablas el estadístico equivalente a 5% de probabilidad (como es de cola inferior se busca tal como está en tablas. Es decir, 0.05) y 29 grados de libertad:

$$x_{i,5\%,29}^2 = 17.7084$$

Esto llevó a establecer las siguientes reglas de aceptación:

- Aceptar H_0 si $X^2 < x_{i,5\%,29}^2$.
- Rechazar H_0 en caso contrario

Al hacer la comparación se tiene el siguiente resultado:



$$X^2 > H_0$$

$$234.61 > 17.7084$$

Por lo tanto, se rechaza la hipótesis nula de que la varianza de los pesos contenidos en el embarque sea menor a la buscada y, por tanto, la empresaria deberá regresar el mismo a su proveedor argumentando que, a pesar de que el peso promedio se ajusta al objetivo buscado de 3.8 Oz., la variabilidad que tiene el mismo en su peso es contraproducente para sus políticas de calidad ya que es muy grande. Por tanto no puede aceptar este embarque y deben mandarle otro que cumpla tanto con el estándar de peso promedio como de varianza de peso esperada.

Los cálculos de la prueba de hipótesis revisada se exponen en la siguiente ilustración:

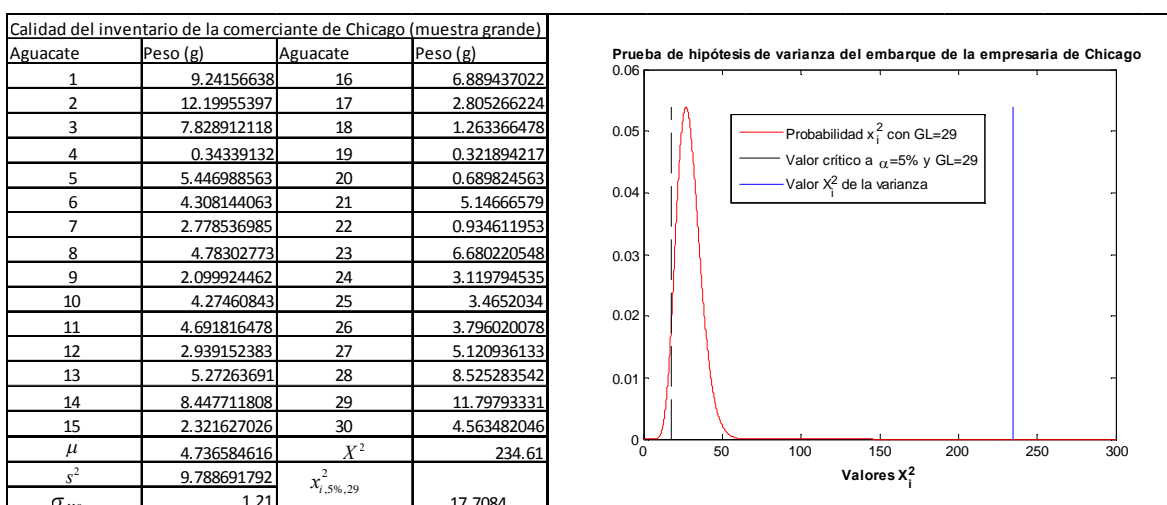


Ilustración 4 Prueba de hipótesis ji-cuadrada para la prueba de varianza realizada en el embarque de la empresaria de Chicago.

Usted puede replicar la prueba de hipótesis previamente estudiada si desea demostrar que la varianza de la muestra es igual, mayor o diferente de σ_{H0}^2 . Lo único que deberá hacer es obtener de las tablas ji-cuadrada los valores del valor crítico o límite ya sea superior e/o inferior, según corresponda el tipo de prueba (recuerde igualdad y desigualdad son pruebas de dos colas, inferioridad de cola inferior y viceversa con superioridad).

5.1.6 Haciendo estimaciones de intervalos de varianzas.

Así como se hicieron estimaciones de intervalos dado lo cambiante de la media muestral para determinar hasta donde podría fluctuar la misma (hacia arriba y/o hacia abajo), también se puede replicar el ejercicio en el caso de las varianzas. Lo único que se tiene que hacer es calcular los correspondientes intervalos de confianza a través de la siguiente expresión:



Fórmula 33: Determinación de los intervalos de confianza de varianzas empleando la distribución ji-cuadrada

$$\sigma_i^2 = \frac{(n-1)s^2}{X_s^2}, \quad \sigma_s^2 = \frac{(n-1)s^2}{X_i^2}$$

En la expresión anterior, X_i^2 y X_s^2 representan el valor crítico en tablas del nivel de probabilidad buscado y el número de colas. En este caso son dos colas y, si se busca calcular el estadístico para un intervalo de confianza de 95%, se debe de determinar (a diferencia del valor Z o t de las dos distribuciones anteriores) el valor ji-cuadrada de 2.5% y 97.5%.

Para ilustrar la idea con un ejemplo, se sigue trabajando con el caso de la empresaria de Chicago. Recuerde usted que su varianza muestral (s^2) es de 9.7886. Extrayendo valores de $X_i^2 = 16.0471$ y $X_s^2 = 45.7223$, se llega a los siguientes intervalos:

$$\sigma_i^2 = \frac{(30-1) \cdot 9.7886}{45.7223} = 6.2086, \quad \sigma_s^2 = \frac{(30-1) \cdot 9.7886}{16.0471} = 17.69$$

Por lo tanto, la empresaria de Chicago, si quisiera darse una idea de entre cuánto nivel de varianza podrían fluctuar diferentes muestras de este embarque, podría concluir que la variabilidad es muy alta a su objetivo de solo 1.10z y refrendar el rechazo del mismo (como en el tema anterior), debido a que la variabilidad del mismo no se ajusta al objetivo de 1.10z.

5.2 Prueba ANOVA.

Ahora toca revisar el segundo subtema de interés del presente: la prueba ANOVA (siglas de ANálisis Of VAriance). En temas previos se hicieron comparaciones de medias muestrales respecto a un objetivo hipotético o de diferencias entre dos muestras solamente. Lo propio se hizo para revisar la varianza de una sola muestra. Es decir, para determinar si esta era igual, superior o inferior a un objetivo determinado.

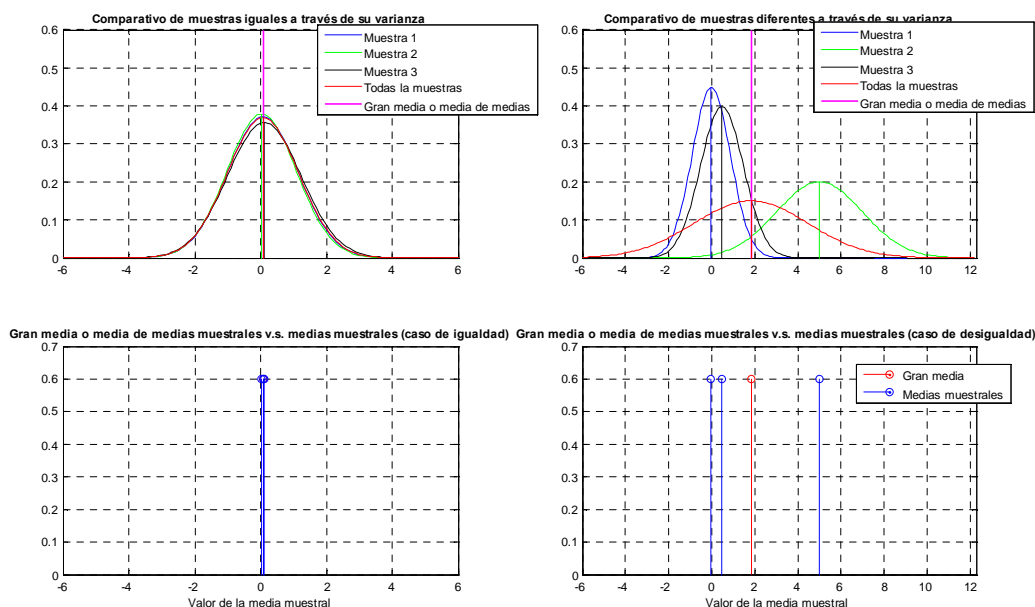
A pesar, de esto queda una pregunta en pie ¿Se pueden comparar medias o varianzas de dos o más poblaciones al mismo tiempo? Es decir, comparar la media de tres poblaciones o muestras al unísono o contrastar la magnitud de la varianza de dos o más de ellas.

La prueba ANOVA es una prueba muy poderosa y muy simple de interpretar. La misma nos servirá para comparar no solo dos medias a la vez sino más de dos medias de muestras que pueden ser independientes o estar acopladas, apareadas o relacionadas y que pueden venir de tamaños de muestras totalmente diferentes. Por ejemplo, piense usted que necesita comparar siete muestras diferentes correspondientes a siete inventarios de aguacate. Los métodos de la técnica clásica



serían bastante limitados si desea hacer pruebas de hipótesis (ya sea de muestras acopladas o independientes). Incluso el hacer esto sería laborioso ya que le implicaría hacer 21 pruebas de hipótesis diferentes. Es decir una prueba de hipótesis de la muestra 1 con la muestra 2 y así sucesivamente.

Para reducir todo este trabajo, la prueba ANOVA viene al rescate e incluso quizá sea la más adecuada en muchas aplicaciones de su futura vida profesional. Las nociones intuitivas de dicha prueba consisten en comparar dos muestras a través de su varianza y no de su media muestral. La idea se expone en la siguiente gráfica:



Gráfica 34 Comparación de tres muestras con media y varianza iguales y comparativo de muestras con estadísticos diferentes.

En la parte superior izquierda se puede apreciar el caso de tres muestras que son iguales. Esto al ser así sus varianzas y sus medias. Lo que interesa analizar son las varianzas ya que estas son consecuencia de las igualdades entre medias. Por tanto, si estas son estadísticamente iguales, se sobreentiende que las medias lo son.

Para poder dar validez a la afirmación anterior, es de necesidad observar que se presupondrá que las muestras están normalmente distribuidas, por lo que, si desea comprobar este supuesto, deberá aplicar la prueba de normalidad con la técnica ji-cuadrada previamente descrita.

Si se observa con detenimiento la gráfica 34, se puede notar que la desigualdad en la varianza de las tres muestras se puede atribuir a dos factores:



- La varianza existente entre las medias muestrales. Es decir, qué tan separadas están unas de otras respecto a un promedio de medias o gran media $\bar{\bar{X}}$.
- La varianza conjunta existente entre todos los datos de las muestras.

Para ilustrar mejor la idea, observe la gráfica superior izquierda. Usted verá que prácticamente no existe variabilidad o varianza entre las medias muestrales (son casi iguales) y, si la varianzas son iguales, la variabilidad total (graficada con una línea roja como “todas las muestras”) es prácticamente la misma que la existente entre las medias muestrales. De cumplirse esta posición, se llega a distribuciones de probabilidad con medias y varianzas prácticamente sobrepuestas unas sobre otras.

Ahora vea usted la gráfica superior derecha. En la misma son notorias las diferencias entre muestras ya que las medias muestrales se encuentran muy separadas. Esto lleva a una varianza entre medias muestrales observable (vea la gráfica inferior derecha en donde se exponen los valores de dichas medias y contraste su valor con el de la izquierda). Como consecuencia de esta situación, la varianza total de todos los datos de las tres muestras del lado derecho (función de probabilidad graficada con rojo) es mucho mayor que la de las tres muestras iguales expuestas a la izquierda. Es entonces que se llega a la noción intuitiva general de la prueba ANOVA:

Nociones intuitivas de la prueba ANOVA: “Dos o más muestras serán iguales si la varianza de sus medias muestrales es igual que la varianza total de los datos de las tres muestras en conjunto”

Para operacionalizar esta afirmación o noción intuitiva de una manera numérica, se tiene el estadístico F:

$$F = \frac{\text{Varianza entre medias muestrales}}{\text{Varianza total de los datos de las muestras en conjunto}}$$

¿Cómo se calculan estos dos tipos de varianza? Enfoquémonos en el numerador. Es decir, la varianza entre medias muestrales. Para ello revisemos las gráficas inferiores de la 34. Nótese cómo las medias muestrales tienen un grado de separación, el cual se determina calculando la gran media o media de medias y, con esta calculando la varianza de medias de la siguiente forma:

Fórmula 34: Cálculo de la varianza entre medias muestrales.

$$\sigma_b^2 = \frac{\sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}, \quad \bar{\bar{x}} = \frac{\sum_i \bar{x}_i}{k}$$



Es decir, la varianza entre medias se determina a partir de la gran media (\bar{x}) que no es más que una media de medias muestrales. De ahí se calcula la varianza entre medias siguiendo los siguientes pasos:

1. Se calculan las diferencias entre cada media muestral respecto a la gran media $\bar{x}_i - \bar{x}$.
2. Se elevan al cuadrado cada una de las diferencias $\left(\bar{x}_i - \bar{x}\right)^2$.
3. Se multiplica cada diferencia cuadrática calculada en el paso anterior por el tamaño de su respectiva muestra $n_i \left(\bar{x}_i - \bar{x}\right)^2$. Esto es, por ejemplo, si la muestra 1 tiene 35 observaciones, se multiplica la diferencia cuadrática de su media muestral respecto a la gran media por 35. Si la muestra 2 tiene 33 observaciones, se hace lo propio y así sucesivamente con cada muestra.
4. Se suman las diferencias cuadráticas y se dividen entre el número de muestras menos un

$$\text{grado de libertad} = \frac{\sum_i^k n_i \left(\bar{x}_i - \bar{x}\right)^2}{k-1}.$$

Esta varianza entre medias muestrales, en la forma en cómo se expresó en la fórmula 34, se conoce como la **“varianza dentro de columnas”**. Este término es muy común, más cuando se presenta lo que se conoce como “tabla ANOVA” que veremos en breve y que no es más que una forma de organizar el análisis de Varianza para su mayor comprensión.

Ahora, siguiendo con el cálculo del estadístico F, ¿Cómo se calcula el denominador o la varianza entre todos los datos de la población? Para llegar al concepto de varianza total de todas las muestras como se determina en la gráfica 34. Es de necesidad observar que la varianza total de todos los datos resulta de una media ponderada de todas las varianzas individuales de cada muestra. La varianza de cada muestra vimos cómo se calcula en la fórmula 11 en el tema de estimaciones de intervalos. Por tanto, se dará por asentado que se domina su cálculo. Es entonces que la forma en obtener la media ponderada de las varianzas, que se define como la varianza de todos los datos, se obtiene de la siguiente manera:

Fórmula 35: Cálculo de la varianza de todos los datos de la muestra.

$$\sigma_{\varpi}^2 = \sum_i^k \left(\frac{n_i - 1}{n_t - k} \right) s_i^2$$

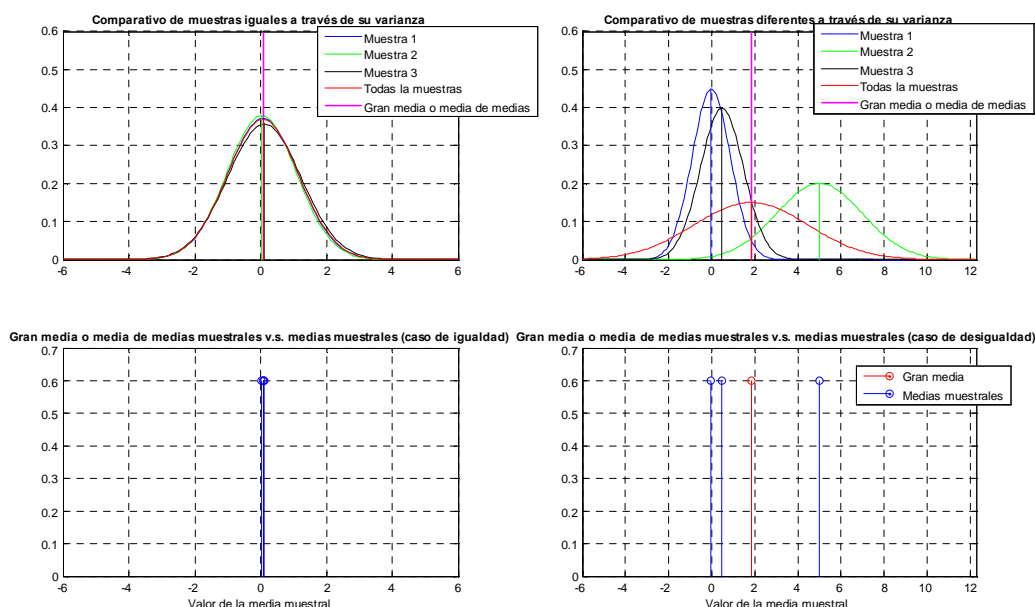
La expresión de varianza anterior se conoce también como **“varianza dentro de columnas”** y es también empleada en la tabla ANOVA. Con la varianza entre medias muestrales y la de todos los datos de las muestras se llega al estadístico F, de la forma en que se describió previamente:

**Fórmula 36: Cálculo del estadístico F para la prueba de igualdad entre muestras.**

$$F = \frac{\text{Varianza entre medias muestrales}}{\text{Varianza total de los datos de las muestras en conjunto}} = \frac{\text{Varianza entre columnas}}{\text{Varianza dentro de columnas}}$$
$$= \frac{\sum_i^k n_i (\bar{x}_i - \bar{x})^2}{k-1} = \frac{\sum_i^k \left(\frac{n_i-1}{n_i-k} \right) s_i^2}{k-1}$$

Las nociones generales intuitivas para emplear el estadístico F son:

1. Si el estadístico F tiene un valor cercano a 1, se llega a la observación de que las medias entre muestras son iguales debido a que la varianza entre muestras y la varianza entre todos los datos de las muestras son iguales, situación que lleva a la situación observada en las gráficas de la izquierda de la gráfica 34 que se presenta a continuación.
2. Si el estadístico F es muy grande, se observa que la diferencia o varianza entre medias muestrales es muy grande en proporción a la varianza de todos los datos de las muestras en conjunto. Por lo tanto, se llega a observar que las medias no son iguales y que nos encontramos en una situación como la expuesta en las gráficas de la derecha de la gráfica 34:



Ahora, una pregunta muy importante por resolver es: Si se tiene que el estadístico F es mayor a uno ¿Qué tanto es “grande”? Es decir ¿qué valor debe tener para decir que es grande? La



respuesta se logra a través de la prueba ANOVA empleando una función de probabilidad F de Fisher.

5.2.1 La función de probabilidad F.

¿Recuerda usted la distribución Ji-cuadrada? ¿Recuerda que la utilizábamos para calcular la probabilidad de eventos que solo pueden tener valores positivos como es el caso de los valores que puede adoptar la varianza de una variable aleatoria? Pues hagamos el siguiente silogismo o razonamiento:

1. Si la distribución ji-cuadrada se utiliza para determinar las probabilidades de una varianza.
2. Si por otro lado es estadístico F es la división entre dos varianzas (la varianza entre muestras y entre el total de datos).
3. Entonces la función de probabilidad F se da por la siguiente expresión:

Fórmula 37: Determinación de la probabilidad F a partir de la probabilidad ji-cuadrada.

$$F = \frac{P(\text{Varianza entre medias muestrales})}{P(\text{Varianza entre datos de las muestras en conjunto})} = \frac{X^2_{i,\alpha,k-1}}{X^2_{i,\alpha,n_T-1}} = F_{k-1,n_T-1}$$

La probabilidad anterior se logra de la tabla de probabilidades F como la presentada en la plataforma Moodle en el tema de la prueba ANOVA. Para determinar la misma se deben especificar los grados de libertad del numerados, que corresponden el número de muestras (k) empleadas menos 1 grado de libertad y el total de datos en las muestras en conjunto (n_T) menos un grado de libertad (para mayor referencia, consulte en la tabla el ejemplo que se da).

5.2.2 La prueba F.

Ahora que se observa que se tienen los múltiples datos necesarios como los grados de libertad del numerador y los del denominador se puede calcular el estadístico F y determinar un valor crítico F empleando las tablas correspondientes al emplear los grados de libertad del numerador y del denominador. Para ello será de necesidad utilizar las tablas de valores F como la que se presenta en la plataforma Moodle en el tema correspondiente a la prueba ANOVA.

Para ilustrar la prueba de hipótesis con la técnica ANOVA retomemos el ejemplo de los empresarios aguacateros de Morelia y Chicago. Lo que ellos quieren plantear es que el inventario que ellos reciben por parte de su proveedor (que es el mismo), tienen la misma calidad. Esto los lleva a plantear la siguiente hipótesis:



H_0 : La calidad de los dos inventarios es la misma. Esto es, $\bar{x}_{Morelia} = \bar{x}_{Chicago}$

H_a : La calidad de los dos inventarios es la misma. Esto es, $\bar{x}_{Morelia} \neq \bar{x}_{Chicago}$

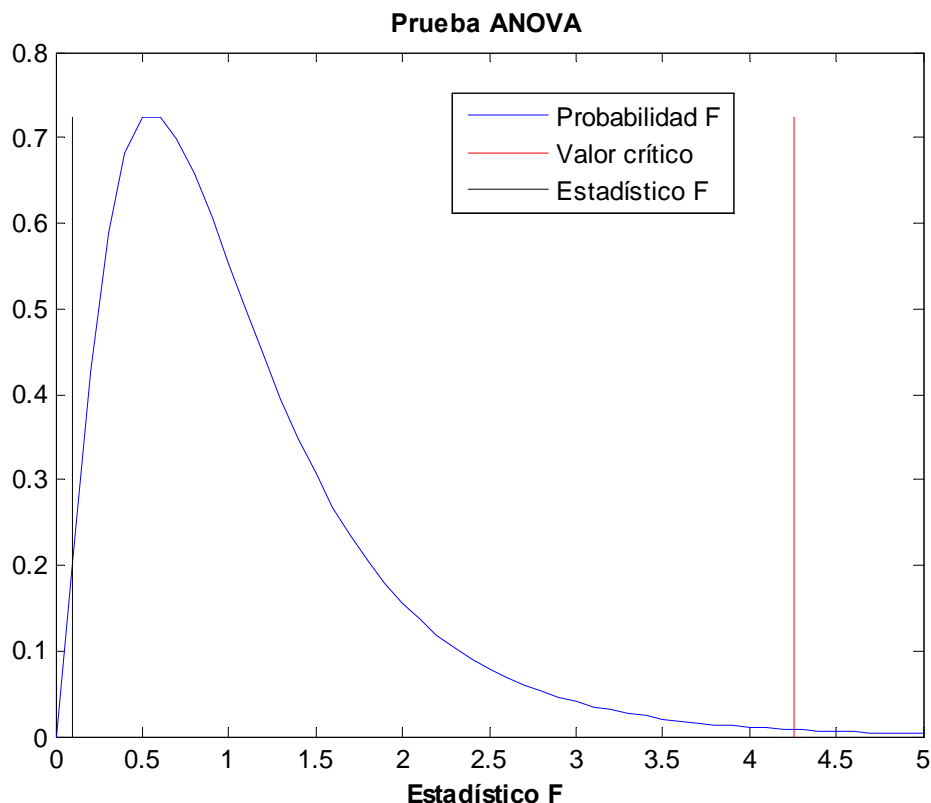
Para demostrar la hipótesis nula se hace el cálculo de la varianza entre medias muestrales y en la totalidad de los datos de las muestras que servirán para determinar el estadístico F:

| El comerciante de Morelia | | | La comerciante de Chicago | | |
|---------------------------|----------|---------------------|---------------------------|-------------|---------------------|
| Aguacate | Peso (g) | $(x_i - \bar{x})^2$ | Aguacate | Peso (g) | $(x_i - \bar{x})^2$ |
| 1 | 170.89 | 893.23576 | 1 | 261.9059912 | 13923.14141 |
| 2 | 185.97 | 2022.0356 | 2 | 345.7353594 | 41915.35948 |
| 3 | 190.74 | 2473.7741 | 3 | 221.8713694 | 6539.701259 |
| 4 | 229.14 | 7768.1396 | 4 | 9.731710016 | 17232.13845 |
| 5 | 145.3 | 18.464639 | 5 | 154.3676559 | 178.615363 |
| 6 | 98 | 1849.2537 | 6 | 122.0928027 | 357.5936694 |
| 7 | 127 | 196.08261 | 7 | 78.74373815 | 3876.209461 |
| 8 | 116.8 | 585.78279 | 8 | 135.5510059 | 29.72369489 |
| 9 | 99.5 | 1722.4949 | 9 | 59.51185924 | 6640.797873 |
| 10 | 107.59 | 1116.4252 | 10 | 121.1424029 | 394.4413313 |
| 11 | 112.34 | 821.5647 | 11 | 132.966079 | 64.59129562 |
| 12 | 108.7654 | 1039.2596 | 12 | 83.29557853 | 3330.140722 |
| Suma cuadrática | | 20506.513 | Suma cuadrática | | 94482.454 |
| Varianza muestral (n-1) | | 1864.2285 | Varianza muestral (n-1) | | 8589.314 |

| | | | |
|----------------|-----------|----------------|-------------|
| Media muestral | 141.00295 | Media muestral | 143.9096294 |
|----------------|-----------|----------------|-------------|

| | |
|----------------------------------|-------------|
| Media de medias | 142.4562897 |
| Varianza entre medias | 50.6927095 |
| Varianza total | 4999.520315 |
| Estadístico F | 0.01014 |
| Valor crítico F ($\alpha=5\%$) | 4.27934 |

Con los grados de libertad tanto del numerador como del denominador, que son $k - 1 = 2 - 1 = 1$ y $n_T - 1 = 24 - 1 = 23$ respectivamente, el cual da un valor crítico F de 4.27934 con un nivel de significancia (α) de 5%. Al comparar el estadístico F con el valor crítico se llega al siguiente contraste:



Esto implica que el valor F es menor al valor crítico. Para poder aceptar la hipótesis nula de que las medias son iguales dado que las varianzas entre la muestra de aguacates de la empresaria de Morelia y el de Chicago son iguales. Por tanto, se puede concluir que la calidad que ambos reciben, dada la prueba aplicada a las dos muestras, es la misma.

5.2.3 Prueba ANOVA para probar la igualdad en la varianza entre dos muestras. El caso de la cola superior.

Previo mente se estudió que la técnica clásica es de utilidad para comparar medias muestrales y que la ji-cuadrada lo es para contrastar varianzas respecto a una varianza objetivo. A su vez se acaba de revisar que la prueba ANOVA es muy poderosa para contrastar igualdad entre medias.

Adicional a la aplicación anterior, la prueba ANOVA puede ser utilizada para comprobar la igualdad estadística de la varianza entre dos muestras (solo dos y objetivo que no se logra con la técnica jicuada). Para ilustrar el método (que es muy sencillo y cambia poco respecto al anterior) se seguirá trabajando con el ejemplo de los dos empresarios aguacateros al querer demostrar la siguiente hipótesis:



$$H_0 : s_{Morelia}^2 = s_{Chicago}^2$$

$$H_a : s_{Morelia}^2 > s_{Chicago}^2$$

En el método a emplear, simplemente cambia la forma de determinar el estadístico por la siguiente forma funcional:

Fórmula 37: Cálculo del estadístico F para comprobar la igualdad en la varianza de dos muestras.

$$F = \frac{s_{Muestra1}^2}{s_{Muestra2}^2}$$

La lógica del estadístico de interés es la misma: “mientras más aproximado a 1 sea su valor, la igualdad de las varianzas será más evidente; una diferencia alta implica desigualdad”.

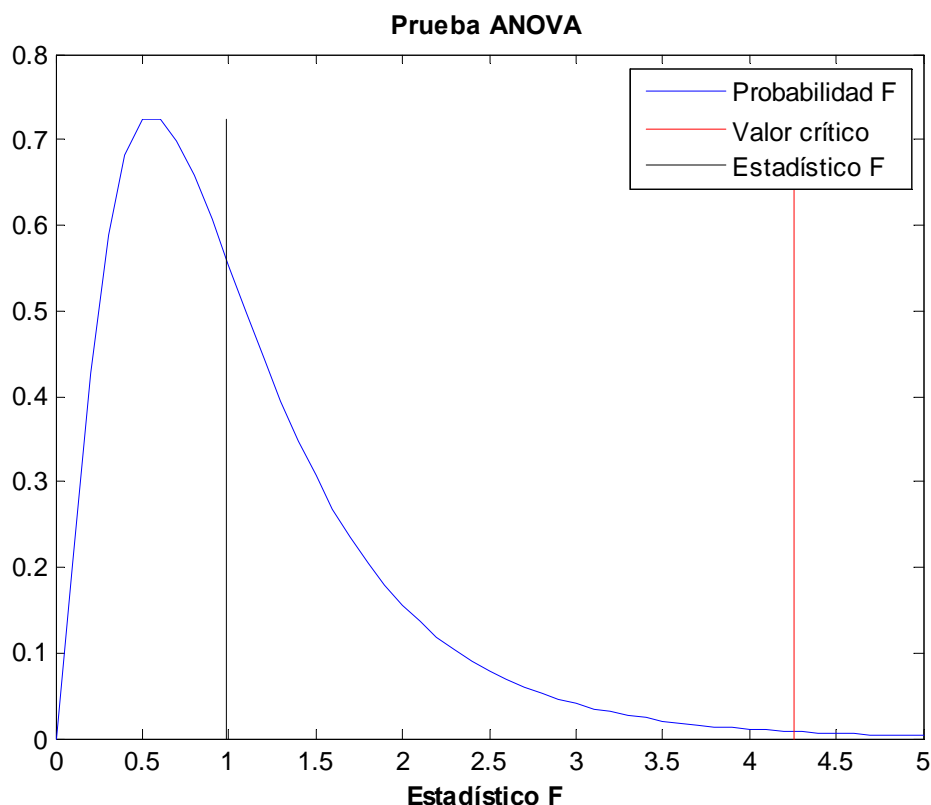
Para el ejemplo que interesa y reutilizando los datos que interesan, se tiene el siguiente cálculo:

$$F = \frac{s_{Morelia}^2}{s_{Chicago}^2} = \frac{141.00}{143.90} = 0.9798$$

Si se contrasta su magnitud con el valor crítico previamente determinado en el ejemplo de la igualdad de medias muestrales, se llega al siguiente resultado:

| | |
|-------------------------|------------|
| Varianza muestral (n-1) | 141.00295 |
| Varianza muestral (n-1) | 143.909629 |

| | |
|----------------------------------|------------|
| Media de medias | 142.45629 |
| Varianza entre medias | 50.6927095 |
| Varianza total | 4999.52031 |
| Estadístico F | 0.97980205 |
| Valor crítico F ($\alpha=5\%$) | 4.27934426 |



Con el resultado logrado se llega a la conclusión que ni la media muestral ni la varianza entre las dos muestras son diferentes, por lo que la calidad del producto recibido en ambos casos es la misma. En este caso particular se estableció la hipótesis de que o las varianzas son iguales; o la varianza de la muestra de la calidad del inventario de la empresaria de Morelia es mayor que el de Chicago.

Sin embargo ¿qué debe de hacerse cuando se desea una hipótesis alternativa como la siguiente:

$s^2_{Morelia} \neq s^2_{Chicago}$? Para poder hacer un contraste de esta naturaleza y tal como se vio previamente, se requiere de una prueba de dos colas, por lo que debe de determinarse la forma de determinar la cola inferior.

Dado que esta función de probabilidad, al igual que la Ji-cuadrada, es asimétrica y es resultado del cociente de dos funciones ji-cuadradas, no se tiene una forma directa de calcular el intervalo inferior, ya sea de la forma que se emplea en la distribución normal (con un valor F correspondiente a valores negativos ya que no existen F negativas) o la de la Ji-cuadrada en donde se determinaba $X^2_{f,0.5\%}$ y se utilizaba como se expresa en la fórmula 33. La única manera de calcular ese intervalo inferior es con la siguiente expresión:

Fórmula 38: Valor del intervalo inferior para prueba de dos colas con la prueba ANOVA.

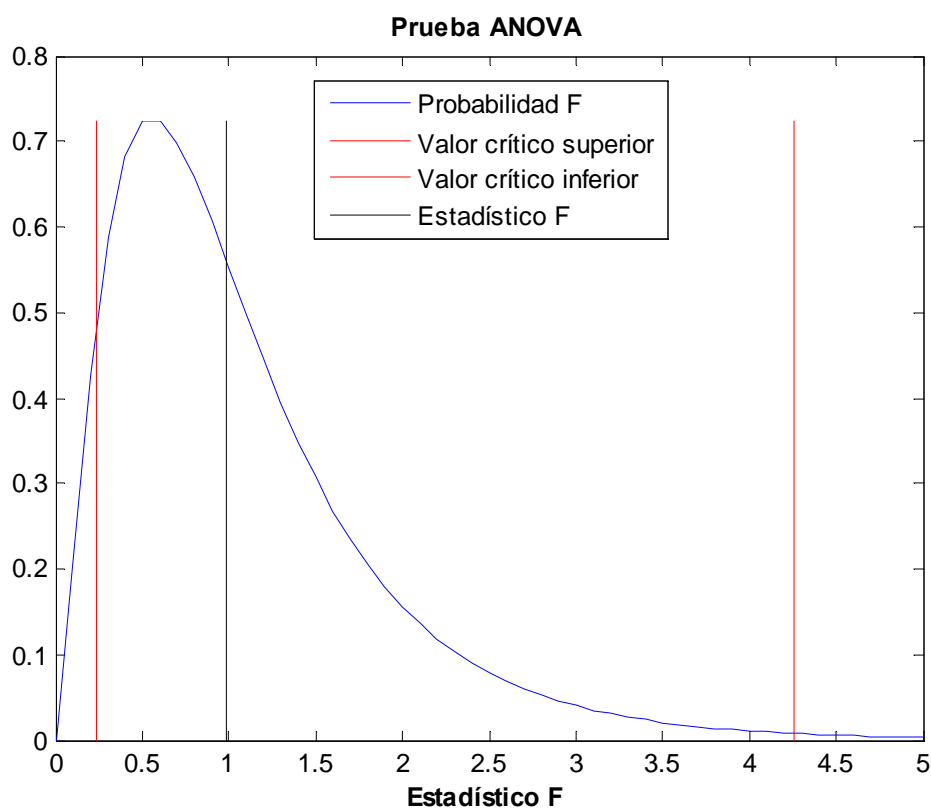


$$F(k-1, n_T-1, \alpha) = \frac{1}{F(k-1, n_T-1, \alpha)}$$

Para el ejemplo manejado hasta ahora se tienen los siguientes valores:

| | |
|-------------------------|------------|
| Varianza muestral (n-1) | 141.00295 |
| Varianza muestral (n-1) | 143.909629 |

| | |
|---|------------|
| Media de medias | 142.45629 |
| Varianza entre medias | 50.6927095 |
| Varianza total | 4999.52031 |
| Estadístico F | 0.97980205 |
| Valor crítico superior F ($\alpha=5\%$) | 4.27934426 |
| Valor crítico inferior F ($\alpha=5\%$) | 0.23368066 |



Dado que el estadístico F se encuentra dentro de los dos intervalos, se observa se acepta la hipótesis de igualdad entre las varianzas y se rechaza la de desigualdad.

Hasta este punto es que se revisan los métodos de comprobación de hipótesis más comunes. Estos son conocidos como **pruebas estadísticas paramétricas** ya que suponen la preexistencia tanto de

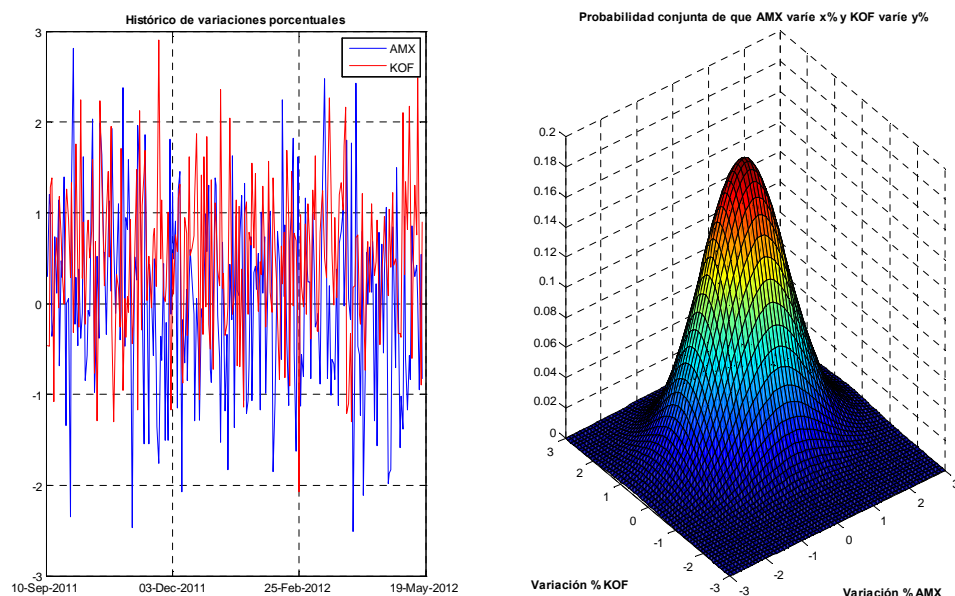


parámetros poblacionales como la suposición de normalidad o algún tipo de función de probabilidad en los datos aleatorios. Sin embargo, como se verá más adelante, existen otro tipo de pruebas a emplear cuando se desconoce la distribución de probabilidad, las cuales se denominan pruebas no paramétricas.

Baste con ahora seguir suponiendo que los datos se distribuyen de manera normal y que la Estadística con que se trabaja es paramétrica.

6 Estadística multivariada: Regresión lineal simple y multivariada.

Conceptos de estadística multivariada. Hasta el momento se ha trabajado con el contraste de muestras o poblaciones que tienen una distribución de probabilidad identidad y, en algunos casos, que tienen un comportamiento dependiente una de la otra. En la vida cotidiana, existen pocos fenómenos que no tienen algún tipo de relación estadística por lo que la probabilidad de suceso de dos o más eventos conjuntos debe de determinarse.

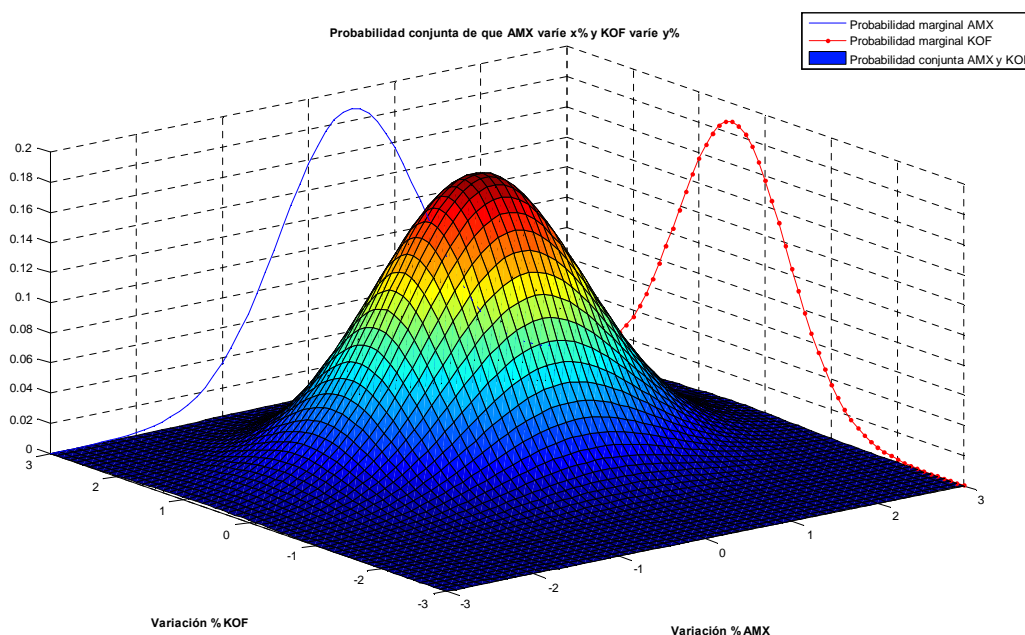


Gráfica 35 El comportamiento histórico de dos acciones y su probabilidad conjunta.

Un claro ejemplo en las ciencias económico-administrativas se puede encontrar en el comportamiento de los rendimientos pagados por una acción que cotiza en bolsa. Suponga usted que tiene un histórico de las variaciones porcentuales diarias del precio o rendimientos de dos acciones que cotizan en la bolsa han tenido a lo largo de un año. Supongamos que se trata de las empresas América Móvil (AMX) y Coca-Cola Femsa (KOF). Las variaciones porcentuales diarias del



precio de las dos acciones se presenta a la izquierda de la gráfica 35 y la probabilidad conjunta de que AMX vería un X% al mismo tiempo de que KOF varíe un Y% o viceversa se presenta a la derecha. Como se puede apreciar, se tiene una probabilidad gaussiana multivariada que, para los fines que nos interesan no se calculará en su totalidad y que, en realidad, tiene una interpretación bastante simple. Para exponer la idea, replantearemos la parte derecha de la gráfica 35 de la siguiente forma:



Gráfica 36 Probabilidades marginales y conjunta de la variación porcentual de las dos acciones de interés.

Note usted cómo se tienen dos probabilidades llamadas “marginales” Estas no son más que las distribuciones de probabilidad gaussianas de la variación porcentual del precio de AMX y de KOF respectivamente. Es decir las probabilidades de que AMX varíe un X% o de que KOF lo haga en un Y% independientemente del comportamiento de otras acciones. Sin embargo, la superficie en forma de campana o “montaña” colorida que se tiene al centro cuantifica la probabilidad de que AMX vería un X% dado o por la influencia de que KOF varía a su vez y al mismo tiempo un Y%.

Para poder calcular una probabilidad conjunta entre dos o más variables aleatorias es necesario suponer que ambas están relacionadas o acopladas. Es decir que no son independientes por lo que adicional a la media y la varianza (o desviación estándar) que se utilizan para calcular las funciones de probabilidad, debe entrar una nueva medida estadística de dispersión que cuantifique la variación conjunta e influenciada entre las dos variables aleatorias que nos interesan. Esta medida se conoce como la **covarianza**. Así como se definió a la varianza como el grado de separación promedio de las observaciones de una variable aleatoria respecto a su media, la covarianza se puede definir como.



Covarianza: “Grado de separación promedio que tiene una variable aleatoria respecto a su media dado el grado de separación promedio que otra variable aleatoria tiene también respecto a su media”.

La forma de calcular la covarianza es de la siguiente manera:

Fórmula 39: Cálculo de la covarianza.

$$\sigma_{x,y} = \frac{\sum_i^k (x_i - \bar{x})(y_i - \bar{y})}{n}$$

En donde la covarianza se denota como $\sigma_{x,y}$. Como se puede apreciar la covarianza se determina de una manera muy sencilla:

1. Se calcula la diferencia entre el valor de cada observación de x menos su media por la diferencia de cada observación de y menos su media.
2. Se suman las multiplicaciones de cada i observación de x e y .
3. La sumatoria se divide entre n para calcular el valor promedio.

A manera de reflexión respecto al método de cálculo, si observa detenidamente la fórmula 39, podrá observar que hay una estrecha relación entre la varianza de una variable aleatoria y la covarianza entre variables aleatorias. En específico usted podrá observar que si desea calcular la covarianza de la variable aleatoria 1 con la misma variable aleatoria 1, se llega a la covarianza:

$$\sigma_{x,x} = \frac{\sum_i^k (x_i - \bar{x})(x_i - \bar{x})}{n} = \frac{\sum_i^k (x_i - \bar{x})^2}{n} = \sigma_x^2$$

Ya que se tiene la covarianza entre la variable 1 y la 2, se calcula otra covarianza entre la variable 2 y la 1 y que se calcularon las correspondientes varianzas de la variable 1 y la variable 2, se está en capacidad de determinar el cálculo de una varianza total entre variables aleatorias destinando el mismo peso a cada variable aleatoria. Es decir, si se tienen dos variables, el peso (w) será de $w = 1/2 = 0.5$. Si son 3, será de $w = 1/3 = 0.33333...$ y así sucesivamente.

También, dados estos mismos pesos w que se dan a las variables aleatorias, se tiene el cálculo de una media multivariada global y de una varianza multivariada global:



Fórmula 40: Cálculo de la media global, la varianza global y la desviación estándar global entre variables aleatorias.

$$\text{Media global} = M = w_x \cdot \bar{x} + w_y \cdot \bar{y}$$

$$\text{Varianza global} = S^2 = w_x^2 \sigma_x^2 + w_x w_y \sigma_{x,y} + w_y w_x \sigma_{y,x} + w_y^2 \sigma_y^2$$

$$\text{Desviación estándar global} = S = \sqrt{S^2}$$

Quizá en este punto se esté preguntando ¿De qué me sirve tanto cálculo de probabilidad multivariada? Prioritariamente nos servirá para establecer que muchos fenómenos económicos y administrativos no son totalmente independientes y que, dada esa situación, tienen un comportamiento conjunto estadísticamente influenciado. De entrada se hará un ejercicio simple de cálculo de probabilidades conjuntas y, posteriormente se hará mención a otra forma alternativa de calcular la covarianza.

De todo lo revisado, la clave u objeto de interés es precisamente la covarianza ya que con ella podremos dar entrada al tema de regresión.

Retomando el tema que interesa, podemos observar que ya estamos en capacidad de calcular la probabilidad de que (si continuamos con el ejemplo de la fluctuación del precio de dos acciones) las dos acciones tengan una variación promedio conjunta de Z%. Para exponerlo, supongamos que tenemos los siguientes datos y cálculos de Media global y Desviación estándar global:

| Variación % promedio | |
|----------------------|--------|
| AMX | 6.21% |
| KOF | 49.53% |

| Ponderación | |
|-------------|--------|
| AMX | 50.00% |
| KOF | 50.00% |

| Medidas de dispersión | |
|----------------------------|---------|
| Varianza Δ% AMX | 104.15% |
| Varianza Δ% KOF | 68.53% |
| Varianza Δ% AMX con Δ% KOF | -3.08% |
| Varianza Δ% KOF con Δ% AMX | -3.08% |

| | |
|----------------------------|--------|
| Media global | 27.87% |
| Varianza global | 19.35% |
| Desviación estándar global | 43.99% |

¿Y qué son las ponderaciones? Esta es una pregunta que podría usted haberse hecho a esta altura. Incluso pudo preguntarse ¿por qué 50% a ambas y no 40% a AMX y 60% a KOF? En este punto es necesario observar que la llamada media global no es más que un **promedio ponderado** que no es más que una forma alternativa de calcular una media. La media tradicional como se calcula en la fórmula 2 se conoce como **media aritmética**. La media ponderada será siempre diferente en valor a la media aritmética excepto cuando se pone una ponderación proporcional a cada variable aleatoria. Esto fue justamente lo que sucedió. Se puso un valor de 0.5, ½ o 50% a cada variable y



con esto se hizo una media de medias. Sin embargo, usted puede cambiar a su gusto las ponderaciones y observar que tanto la media global como la varianza y desviación estándar globales hacen lo propio. Hagamos el caso de 40% a AMX y 60% en KOF:

| Variación % promedio | |
|----------------------|--------|
| AMX | 6.21% |
| KOF | 49.53% |

| Ponderación | |
|-------------|--------|
| AMX | 40.00% |
| KOF | 60.00% |

| Medidas de dispersión | |
|----------------------------|---------|
| Varianza Δ% AMX | 104.15% |
| Varianza Δ% KOF | 68.53% |
| Varianza Δ% AMX con Δ% KOF | -3.08% |
| Varianza Δ% KOF Δ% AMX | -3.08% |

| | |
|----------------------------|--------|
| Media global | 32.20% |
| Varianza global | 22.23% |
| Desviación estándar global | 47.15% |

Note cómo la media global y la desviación estándar global se incrementaron. Quizá una pregunta natural sea ¿Entonces de qué me sirve cambiar las ponderaciones si ya me queda claro que ponderaciones proporcionales (1/n) me da una media aritmética? En el caso de la Economía o las Ciencias administrativas, puede tener múltiples aplicaciones. Para el caso específico del ejemplo que nos interesa, usted puede elegir cuánto porcentaje del total de su patrimonio invertir en AMX y cuánto en KOF. Usted puede optar por 50% en cada una o hacer múltiples ponderaciones como al de 40% en AMX y 60% en KOF. En función de la ponderación que asigne será el rendimiento promedio que podría lograr (media global) en todo su portafolio de dos acciones y el grado de riesgo o variabilidad, dada por la desviación estándar, que resulte. Por ejemplo, usted al cambiar de ponderaciones proporcionales (1/n=50%) a las recientemente estudiadas, podrá observar que su rendimiento se incrementó pero también hizo lo propio el nivel de riesgo medido con la desviación estándar. Este dilema de cuánto invertir en cada acción que formará parte de nuestro portafolio es algo que los administradores de portafolios o de inversiones resuelven en su vida cotidiana. Este tema no se revisará y se dejará hasta aquí ya que sale de la óptica de la materia. Simplemente se menciona esto para dar una explicación a las ponderaciones que se asignan a cada variable.

Para ilustrar la idea del cálculo de probabilidades conjuntas, retomemos el ejemplo del portafolio con un nivel de inversión o ponderación del 50% en cada acción o variable aleatoria. Usted se podrá preguntar, dada esta ponderación o nivel de inversión en AMX y KOF ¿Cuál es la probabilidad de que mi portafolio tenga un rendimiento conjunto o total mayor o igual a 40%. Esto es muy sencillo de responder. Recordando el tema de cálculo probabilidades, lo que hacemos es estandarizar el valor de 40% como sigue:

$$Z = \frac{40\% - M}{S} = \frac{40\% - 27.87\%}{43.99\%} = 0.2757$$



Posteriormente, se busca el valor de la probabilidad de un valor Z en la tabla correspondiente y esto nos lleva a la siguiente probabilidad:

$$P(0.2757) = 10.864\%$$

Sin embargo, recuerde usted que esta probabilidad se da en el lado derecho de la campana gaussiana y como estamos buscando la probabilidad de que el rendimiento conjunto de las variables aleatorias o de todo el portafolios sea mayor o igual a 40%, se resta de 50% de probabilidad el 10.864% logrado y se tiene que:

$$P(40\%) = 39.136\%$$

Es decir, la probabilidad de que el portafolio o las dos variables aleatorias (variación porcentual del rendimiento de una acción) sea mayor o igual a 40% es de 39.13%.

6.1 El coeficiente de correlación y su interacción con la covarianza.

Si usted observa detenidamente la fórmula de cálculo de la covarianza puede preguntarse ¿Y qué hace que dos variables aleatorias se muevan en conjunto?:

$$\sigma_{x,y} = \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y})}{n}$$

El responder que causa esto es algo muy amplio de responder y que se investiga a la luz de las diferentes ciencias. Por ejemplo qué hace que el número de flores en un valle sea mayor conforme hay lluvia es algo que la biología, en especial la Botánica, nos puede responder. Qué hace que el precio de una acción se mueva en conjunto con el de otra es otra situación que se puede investigar con la Economía Financiera y así sucesivamente. Todas las causas de este fenómeno de interacción se responden con las respectivas ciencias de interés. Sin embargo responder ¿en qué magnitud y con qué grado de apego se mueven dos fenómenos modelados matemáticamente a través de dos variables aleatorias? Es algo que se responde gracias a la estadística. En qué magnitud se mueve una variable aleatoria de manera conjunta respecto al movimiento de otra se determina con la covarianza. Por ejemplo observe las dos covarianzas del ejercicio anterior. Esto nos dice que el precio de AMX varía en respecto a su media cuando el de KOF hace lo propio y la variación promedio que ambos tienen respecto a sus medias es de -3.08%. Sin embargo el -3.08% de covarianza no sale de la nada. Esto sale al grado de apego que existe entre las dos variables aleatorias.



Para ilustrar la idea usted piense en la luna y la tierra. La luna se mueve alrededor de la tierra y no del sol dado que la gravedad o grado de atracción entre ambas es mayor que el existente entre el sol y la luna. Por eso la luna no ha dejado a la tierra para irse con el sol. En las variables aleatorias sucede algo similar. Cuando estas son dependientes debe existir un grado de **correlación** o gravedad entre ellas para que puedan covariar. De lo contrario no existirá la covarianza y no existirá dependencia entre ellas.

Vea usted el coeficiente de correlación como el “pegamento” entre variables. Con este pegamento o nivel de gravedad usted puede calcular la covarianza como sigue:

Fórmula 41: Cálculo de la covarianza empleando el coeficiente de correlación de Pearson.

$$\sigma_{x,y} = \sigma_x \cdot \rho_{x,y} \cdot \sigma_y$$

En la expresión anterior, el coeficiente de correlación de Pearson o simplemente coeficiente de correlación $\rho_{x,y}$ (En donde $\rho_{x,y} = \rho_{y,x}$) es esa gravedad o pegamento del que se habla. Note usted cómo se tiene, de inicio, esa variabilidad individual de cada caso (σ_x y σ_y) y se relaciona con ese pegamento. Por tanto, la variable de interés es el coeficiente de correlación $\rho_{x,y}$. El coeficiente puede tener valores de -1 a 1. Es decir, un valor de $\rho_{x,y} = -1$ implicaría que con cada 1% que suba AMX, KOF baje 1% y viceversa. Cuando $\rho_{x,y} = 1$ el precio de AMX sube 1% cuando KOF hace lo propio y viceversa. Al incorporarse la variabilidad individual en cada caso (σ_x y σ_y) se llega a la covarianza que se presentó en la fórmula 39:

$$\sigma_{x,y} = \sigma_x \cdot \rho_{x,y} \cdot \sigma_y = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Algo que usted podría pensar sería: “bueno ya que tengo la covarianza se pueden hacer pronósticos”. Por ejemplo si se tiene la covarianza de la variación porcentual del PIB y de las ventas de una empresa, se podrían pronosticar las ventas de esta. Hasta cierto punto, el planteamiento suena correcto. Sin embargo, este no es del todo completo ya que se le podría preguntar ¿Cuánto variaría el nivel de ventas si el PIB sube en 7%? La covarianza le dice que la variación conjunta (covarianza) es, digamos 4.5% pero esto no implica que se establezca una elasticidad. Es decir que por cada 1% de incremento en PIB se incremente en 4.5% las ventas llevando a niveles de 31.5% (4.5% · 7%). La covarianza nos dice el sentido y en qué valor promedio varían las dos variables. Para poder calcular cuánto cambian las ventas de una empresa dado cada incremento porcentual del PIB se utiliza otra técnica mucho más precisa que emplea fuertemente las varianzas y covarianzas. Esta se conoce como la técnica de la regresión y nos será de mucha utilidad para establecer pronósticos.



6.2 El modelo regresión lineal simple para establecer relaciones estadísticas entre variables y hacer pronósticos básicos.

En este punto estamos entrando a una de las aplicaciones más importantes (sino la más importante) de la Estadística inferencial: la regresión. Con esta se logrará establecer la relación estadística entre variables de la forma:

$$y = \alpha + \beta \cdot x$$

Esto quiere decir que la variable y se determina por una ecuación matemática en donde a una constante α (que puede ser de cero o de otro valor positivo o negativo) se le suma un valor de x multiplicado por un número o constante β (que puede ser de cero o de otro valor positivo o negativo). Si usted observa detenidamente el modelo de regresión, podrá apreciar que será capaz de hacer estimaciones del tipo “por cada valor de x se tendrá un valor de y dado por $\alpha + \beta \cdot x$ ”. Es decir ya podrá decir cuánto valdrá y dado el de x . La clave aquí estará en calcular los coeficientes α y β . Esto es lo que determinaremos a continuación y a lo que le daremos una explicación gráfica.

6.2.1 Determinación de los coeficientes del modelo de regresión.

Para determinar los coeficientes de regresión, siendo esta el modelo resultante de la interacción entre las dos variables, se ocupa la covarianza que es la que cuantifica el grado de variación promedio conjunta entre variables, la varianza de la variable regresora (x)¹⁶, y la media muestral de la regresada (\bar{Y}):

¹⁶ La variable regresada es Y por lo que la regresora es X ya que esta “regresa” o determina el valor de Y .



7 Bibliografía

Levin, R., & Rubin, D. (2004). *Estadística para administración y economía* (7a. ed.). México, México: Pearson educación.