Inferencia Estadística

Dra. Graciela González Farías
Dr. Ulises Márquez

Maestría en Cómputo Estadístico.

CIMAT Monterrey.

Agradecimientos

En forma de agradecimiento, se enlistan personas que han contribuido de una u otra forma en la construcción de estas notas a través de los años:

- Víctor Muñiz
- Juan Antonio López
- Sigfrido Iglesias González
- Rodrigo Macías Paéz
- Edgar Jiménez
- Todos los estudiantes que han colaborado con sugerencias y comentarios sobre estas notas.

Estas notas son de uso exclusivo para enseñanza y no pretende la sustitución de los textos y artículos involucrados.

Temario

1. Variables aleatorias y distribuciones de probabilidad.
   a) Distribuciones de probabilidad de variables aleatorias discretas.
   b) Procesos de Poisson.
   c) Distribuciones de probabilidad de variables aleatorias continuas.
   d) Métodos gráficos para la identificación de distribuciones.
   e) Estimación de densidades.
   f) Distribuciones de probabilidad de vectores aleatorios.
   g) Esperanzas condicionales y regresión.
   h) Modelos jerárquicos, compuestos y mezclas de variables aleatorias.
   i) Transformaciones de variables aleatorias.
   j) Simulación de variables aleatorias.
   k) Convergencia de variables aleatorias y el Teorema del Límite Central.

Temario

2. Distribuciones muestrales y métodos de estimación.
   a) Propiedades de los estimadores.
   b) Estimadores no insesgados
   c) Distribuciones muestrales.
   d) Principio de máxima verosimilitud.
   e) Estimación puntual.
   f) Bootstrap y jackknife.

Temario

3. Pruebas de Hipótesis e intervalos de confianza.
   a) Definición de conceptos.
   b) Potencia de la prueba.
   c) Pruebas para dos poblaciones normales independientes.
   d) Pruebas para medias en muestras pareadas.
   e) Pruebas básicas de varianzas.
   f) Pruebas para proporciones.
   g) Conceptos de estimación bayesiana.
   h) Temas optativos de modelos para presentaciones finales, por ejemplo:
      1. Pruebas no-paramétricas clásicas.
      2. Pruebas de permutaciones.
      3. Estimación no paramétrica (suavizadores y splines).
      4. Modelos gráficos probabilístas.
      5. Entre muchos otros.

Evaluación y acreditación

- Dos exámenes parciales, 18 de septiembre y 5 de noviembre: 15 %, cada uno.
- Evaluación de las tareas (de 2 tipos) y actividades en clase y asistencia: 40 %.
- Un examen final, consistente en una exposición donde se entrega un reporte y se hace una presentación de 1/2 hora. La presentación debe incluir antecedentes, metodología, un ejemplo práctico y compartir el código. Deberán entregar a los instructores y a sus compañeros el resumen. Adicionalmente, deberán dejar un ejercicio sobre el tema a sus compañeros que calificarán en forma honesta: 30 %.

Las tareas tienen una frecuencia quincenal e incluyen TODOS los ejercicios dejados en las notas y requerirán en general el uso de recursos computacionales.

Textos

- **Larry Wasserman (2004) . All of Statistics, A concise course in Statistical Inference. Springer.**
- F.M. Dekking, C. Kraaikamp, H.P. Lopuhaa L.E. Meester (2005). A Modern Introduction to Probability and Statistics, Understanding Why and How. Springer text in Statistics.
- John A. Rice (1995). Mathematical Statistics and Data Analysis, Second Edition. Duxbury Press.
- Casella & Berger. (2002). Statistical Inference, Second Edition . Duxbury Press.
- Richard J. Larsen and Morris L. Marx (2011). An Introduction to Mathematical Statistics and its Applications. Fifth Edition. Prentice Hall.

Estimación por Intervalos

Estimación por Intervalos

Si se nos encomendara la tarea de estimar algún parámetro poblacional, podríamos intentarlo de dos formas: a) "Me late que...", b) "Interactuaré con..."

Utilizaremos la segunda forma, como hasta ahora lo hemos hecho!!! Nuestra elección equivale a navegar en el mundo de la inferencia estadística. En este mundo, interactuaremos con la población a través de muestras aleatorias. Claro que será una interacción planeada y sistemática basada en todos los conceptos estadísticos que hemos logrado dominar.

Anteriormente, hemos tratado de estimar parámetros poblacionales a través de la estimación puntual, esto es, identificar un estimador (función de la m.a.), que cumpla con ciertas condiciones (su calidad es importante). En principio, lo que se ha hecho es resumir la información de la m. a. de una manera adecuada y posteriormente evaluar la calidad del estimador.

Estimación por Intervalos

Es como una evolución que no ha terminado:

- Un primer paso en la obtención de infomación pudo haber sido tomar una m.a. de tamaño 1 y con ello intentar estimar el parámetro poblacional; una elección algo pobre.
- Un segundo paso, más razonable, es tomar una m.a. de tamaño mayor a 1, dos o más. Mejor elección. Ahora tenemos más información algo "esparcida".
- En el tercer paso, resumimos la información en un estadístico (estimador), no cualquier estadístico, e intentamos estimar el parámetro poblacional mediante el valor del estimador para una cierta m.a.

Estimación por Intervalos

Claro está, si tomamos otra m.a., muy posiblemente encontraremos otro valor para el estadístico. Si recordamos este carácter aleatorio de los estimadores, podrás aceptar que podemos utilizar más información (para estimar el valor del parámetro). Digamos,

- Su dispersión (mediante su varianza),
- Su distribución de probabilidad (comportamiento),

los cuales serían los pasos 4 y 5 en la evolución recién mencionada.

Inclusión de la Variabilidad

Inclusión de la Variabilidad

En la estimación de un parámetro para una población dada, además de sólo asignarle un valor único (obtenido del estimador), es factible incluir de alguna forma la información correspondiente a la variabilidad del estimador.

Por ejemplo, si queremos estimar μ, sabemos que podemos usar el valor de $\hat{\theta} = \overline{X}$, pero también sabemos que $V(\overline{X}) = \frac{\sigma^2}{n}$. ¿Cómo incluir esta información?

Una primera forma de incluir esta información es:

$$\hat{\theta} \pm \sqrt{V(\hat{\theta})}.$$

Inclusión de la Variabilidad

Así,

- con signo menos tenemos el extremo izquierdo de un intervalo localizado a una distancia de una desviación estándar del valor del estimador,
- con el signo más tenemos el extremo derecho de un intervalo localizado a una distancia de una desviación estándar del valor del estimador.

Ejemplo: Si en un problema obtenemos $\overline{X} = 3$ y $V(\overline{X}) = \frac{\sigma^2}{n} = 4$, entonces se podría reportar el intervalo

$$3 \pm \sqrt{4}.$$

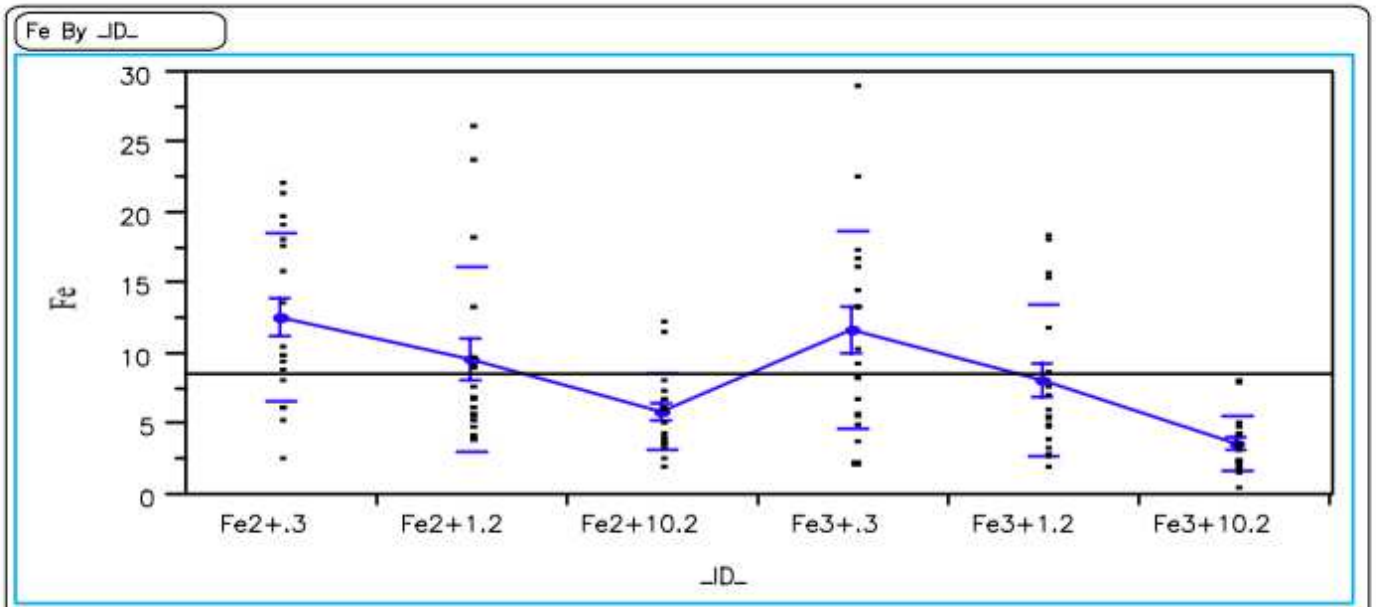
Sin embargo, si $V(\overline{X}) = 0.4$, entonces el intervalo

$$3 \pm \sqrt{0.4}$$

nos da más información que el primer intervalo (recuerda la desigualdad de Chebyshev, por ejemplo); esto es, la variabilidad juega un papel crucial.

Inclusión de la Variabilidad

En la práctica es común ver gráficas como:



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
Fe2+.3	18	12.6394	6.08209	1.4336
Fe2+1.2	18	9.6322	6.69122	1.5771
Fe2+10.2	18	5.9367	2.80678	0.6616
Fe3+.3	18	11.7500	7.02815	1.6566
Fe3+1.2	18	8.2039	5.44739	1.2840
Fe3+10.2	18	3.6989	2.03087	0.4787

Inclusión de la Variabilidad

De estas gráficas podemos construir algunas conclusiones preliminares. Por ejemplo, parece ser que los valores medios de fierro varían de acuerdo a las concentraciones y al tipo de Fe empledo. Así también algunos casos presentan menor variabilidad que otros.

Sin embargo, esto no nos permite más que describir la situación. La pregunta obligada es: ¿Son estas diferencias reales?

Recordemos que existe una distribución asociada con estas mediciones. En este caso estamos hablando de comparar el comportamiento de 6 poblaciones. Más adelante discutiremos algunos aspectos experimentales que nos permitirán establecer otras "propiedades" de nuestras poblaciones, en particular, el concepto de homogeneidad e independencia. Por ahora, veamos una forma "mejor" de trabajar con la variabilidad: el concepto de **Intervalo de Confianza**.

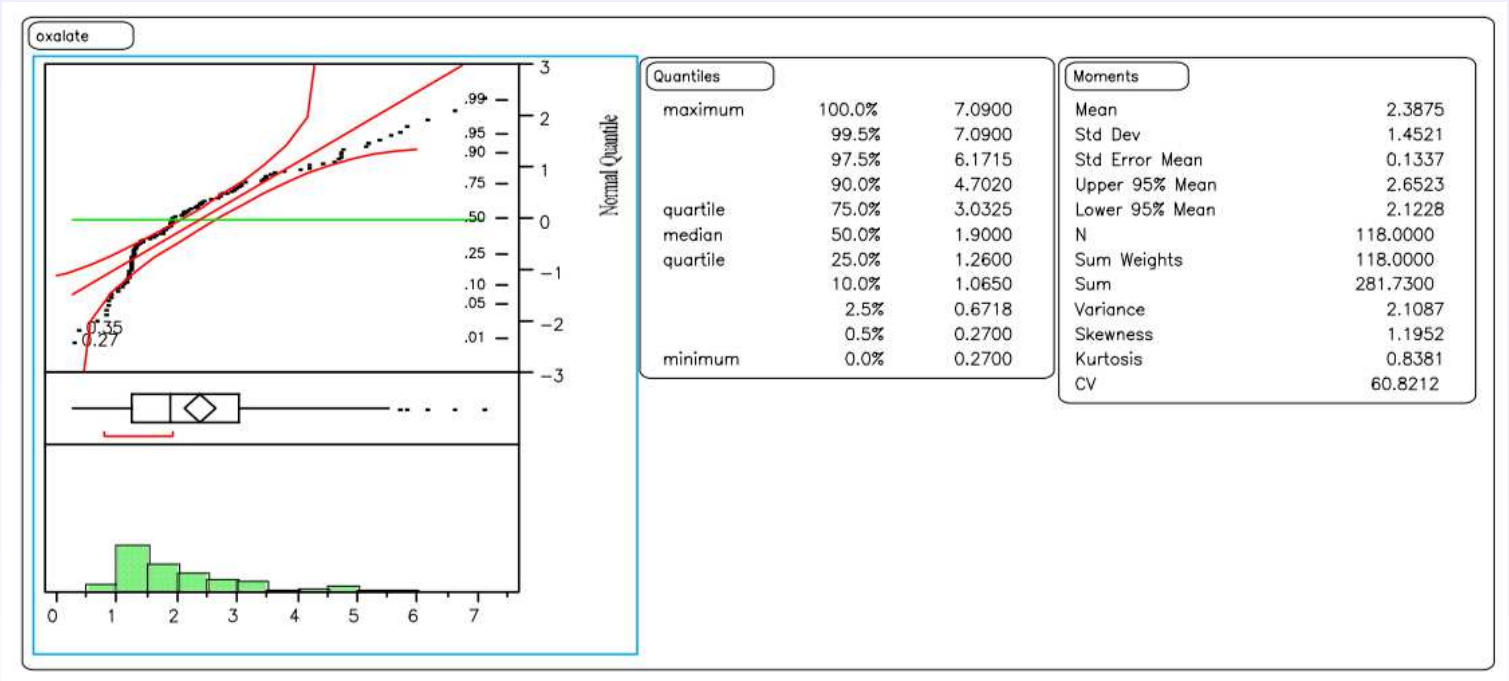
Inclusión de la Variabilidad

Veamos un ejemplo.

Ejemplo: En una rama de la industria alimentaria, se realizan en forma rutinaria mediciones del contenido de calcio en comida para animales (mascotas). El método estándar utiliza precipitación de oxalato de calcio seguida de tritanio; es una técnica que consume tiempo. Los resultados de 118 muestras (Heckman 1960), se dan en el archivo calcio.txt (columna 1). Ahora, con lo que hemos discutido hasta aquí, ¿qué podemos decir a partir del comportamiento de estos valores muestrales?

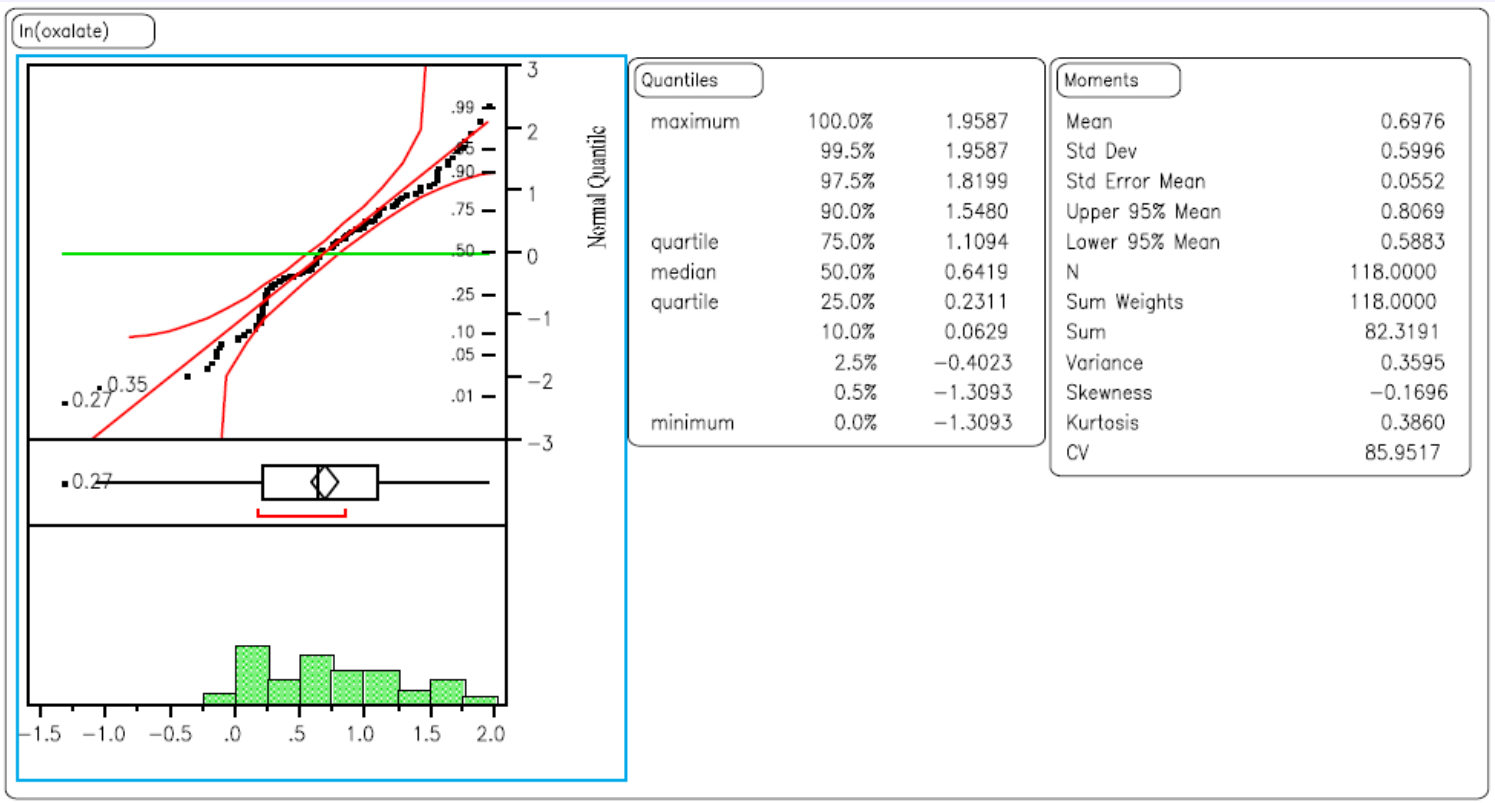
Sabemos cómo ajustar gráficamente un modelo probabilístico, y conocemos, al menos en algunos casos, lo que esperamos de nuestros valores muestrales a través de caracterizar las distribuciones muestrales. Veamos que pasa exactamemte para este ejemplo:

Inclusión de la Variabilidad



Inclusión de la Variabilidad

De las graficas anteriores y de los valores muestrales de simetría, curtosis, el desplazamiento entre la media y la mediana y los posibles valores extremos, es claro que nuestros datos no parecen provenir de una población con distribución Normal. Sin embargo, considerando una transformación de ellos, digamos sus logaritmos, observamos los siguientes cambios:



Inclusión de la Variabilidad

De estos resultados incluso nos atreveríamos a decir que la muestra posiblemente fue obtenida de una población (valores del logaritmo del contenido de calcio) normal, con la posible presencia de dos valores atípicos.

Por otra parte, como la muestra es en particular "grande", podríamos obtener el comportamiento aproximado de cantidades tales como la media muestral y la varianza muestral, usando el TLC. Para fines de ilustración pensaremos que la distribución de la población es normal (y por ende, la distribución de los valores de calcio, sería lognormal).

Inclusión de la Variabilidad

Bajo las condiciones anteriores, ¿qué nos dicen los valores de $\overline{x} = 0.6976$ y de $s^2 = 0.3595$? Estos son los estimadores puntuales de μ y σ^2 en la población. ¿Qué más podemos decir sobre μ y σ^2?

Nos gustaría no sólo describir la variabilidad de nuestros estimadores puntuales, sino aprovechar el hecho de que sabemos que la distribución de los valores de $\hat{X}$ es $Normal(\mu, \frac{\sigma}{\sqrt{n}})$ y que la distribución de S^2 es $\frac{n-1}{\sigma^2}\chi^2$.

Intervalos de Confianza

Intervalos de Confianza

Inclusión de la Distribución de Probabilidad.

Cuando incluimos la distribución muestral, además de la variabilidad, se dice que construimos un Intervalo de Confianza (IC). Existen muchas formas de construcción de intervalos de confianza, pero aquí sólo discutiremos el método del Pivote.

Definición (Pivote)

Un pivote es una función de los elementos de la muestra y del parámetro que se desea estimar, de tal forma que la función de probabilidad o de densidad de dicha función no dependa del parámetro a ser estimado.

Intervalos de Confianza

Así, el método del pivote consiste en:

- Encontrar una función de las observaciones y del parámetro que se desea estimar,

de tal forma que

- su distribución no contenga al valor del parámetro.

Más detalladamente:

Sea $X_1, X_2, \ldots, X_n$ una m.a. de una población. Nuestro objetivo es construir un intervalo de confianza para un parámetro θ.

Intervalos de Confianza

1.- Supongamos que que se puede encontrar una v.a. que depende de

- $X_1, X_2, \ldots, X_n$ y de θ,
- la distribución de esta v.a. no depende de ningún parámetro desconocido.

Sea $h(X_1, X_2, \ldots, X_n, \theta)$ (Posible Pivote).

Ejemplo: Población Normal, con σ^2 conocida. Se desea un IC para el parámetro μ. Si la población es normal, sabemos que

$$\overline{X} \sim N(\mu, \sigma^2) \quad \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

entonces,

$$h(X_1, X_2, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

(h depende de la muestra, a través de $\overline{X}$ y también de $\theta = \mu$)

$$h(X_1, X_2, \ldots, X_n; \mu) \sim N(0,1)$$

(la distribución de h no depende de parámetros desconocidos).

Notas: Aquí, $h(X_1, X_2, \ldots, X_n) = \overline{X}$.

Podríamos definir otro pivote, por ejemplo

$$\frac{X_1 - \mu}{\sigma} \sim N(0,1).$$

Intervalos de Confianza

2.- Para cualquier $\alpha \in (0,1)$ se pueden determinar constantes a y b (percentiles) tales que

$$P\{a < h(X_1, X_2, \ldots, X_n, \theta) < b\} = 1-\alpha,$$

y

(a, b) no dependen de θ.

Ejemplo: Población Normal, con σ^2 conocida (continuación).

$$P\{a < h(X_1, X_2, \ldots, X_n) < b\} = 1 - \alpha$$

$$\Rightarrow P\left(-z_{\alpha/2} < \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\sigma/2}\right) = 1 - \alpha.$$

Intervalos de Confianza

3.- Si las desigualdades anteriores se pueden manipular algebraicamente para aislar θ y obtener:

$$P(l(X_1, X_2, \ldots, X_n) < \theta < u(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$

entonces,

$$l(X_1, X_2, \ldots, X_n)$$

$$u(X_1, X_2, \ldots, X_n)$$

constituyen los límites inferior y superior del Intervalo de Confianza del 100(1- α) % para θ.

Nota: Observemos que como $l(X_1, X_2, \ldots, X_n)$ y $u(X_1, X_2, \ldots, X_n)$ dependen de la muestra $X_1, X_2, \ldots, X_n$, son variables aleatorias y por lo tanto, obtener una probabilidad sobre ellas tiene sentido. Pero una vez que evaluamos en una realización de una muestra específica, estos valores se convierten en constantes, esto es, para una muestra dada obtendremos uno de los múltiples IC que pueden formarse para μ, y allí es en donde el concepto de "nivel de confianza" entra en acción.

Ejemplo: Población Normal, con σ^2 conocida (continuación).
Despejando para μ:

$$P\left(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

multiplicado por -1 queda

$$P\left(z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} > \mu - \overline{X} > -z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

o equivalentemente

$$P\left(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu - \overline{X} < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

finalmente, sumando$\overline{X}$ en cada término

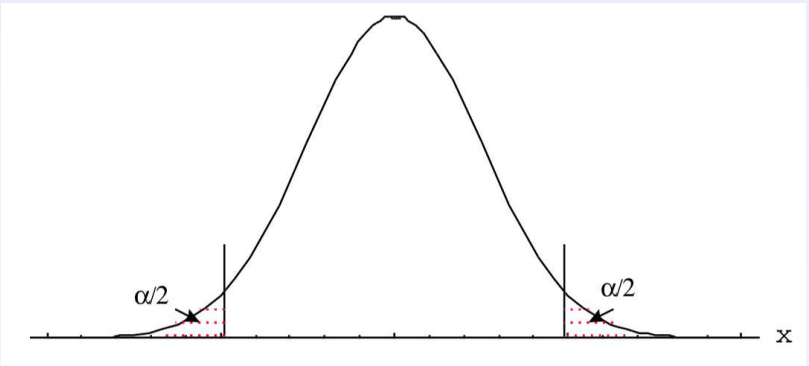
$$P\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

En este caso:

$$l(X_1, X_2, \ldots, X_n) = \overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$
$$u(X_1, X_2, \ldots, X_n) = \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

donde $z_{\alpha/2}$es un valor de la v.a. $Z \sim N(0, 1)$ la cual tiene a la derecha y bajo la curva normal un área de $\alpha/2$.

Intervalos de Confianza



¿Qué área existe a la izquierda de $-z_{\alpha/2}$?

$\therefore$ El intervalo de confianza para μ del 100(1-α)%, en poblaciones normales, con σ^2 conocida es:

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Gráficamente

Donde

$$P(Z > b) = \alpha/2 \quad P(Z < a) = \alpha/2$$

Intervalos de Confianza

¿Cómo se interpreta un Intervalo de Confianza?

- En el muestreo repetido, el IC contendrá con una confiabilidad del $(1-\alpha)$ al valor del parámetro.
- Dada una m.a. particular sólo podemos decir que el IC contiene o no al parámetro. Por ejemplo, si $l(X_1, X_2, \ldots, X_n) = 3$ y $u(X_1, X_2, \ldots, X_n) = 5$, entonces (3,5) contiene al parámetro con probabilidad uno o cero: está o no está.
- Un intervalo de confianza **NO** es un intervalo en el cual θ cae con probabilidad $(1-\alpha)$.
- Un intervalo de confianza no es único, existen muchos intervalos del mismo nivel de confianza para el mismo parámetro y con los mismos datos.

Intervalos de Confianza

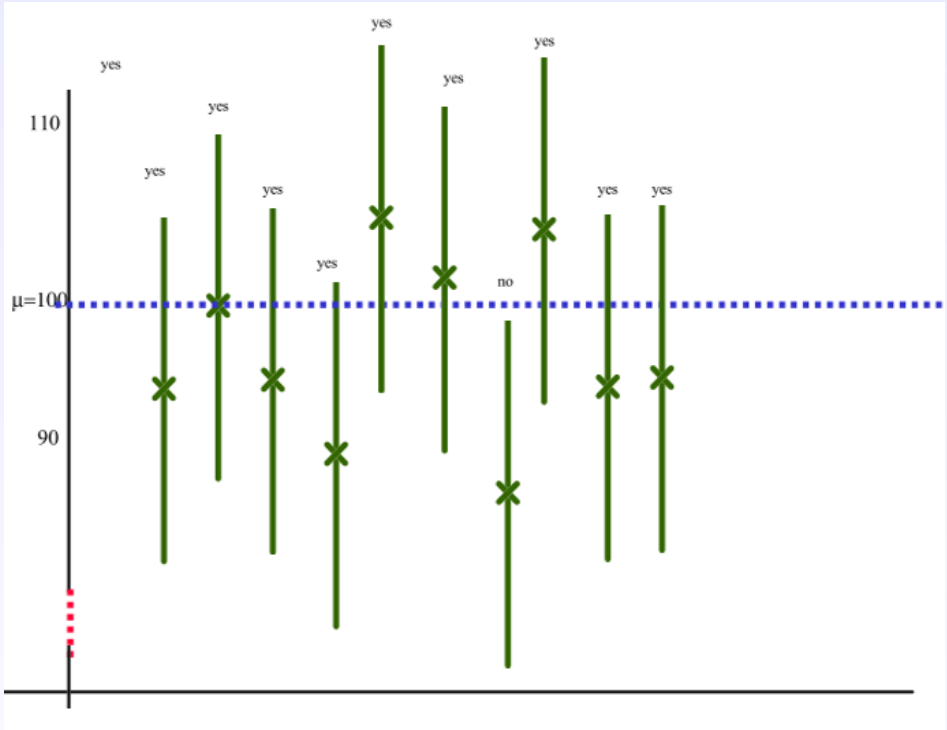
Debemos recalcar que cuando obtengas tu intervalo, éste tendrá una interpretación de **frecuencia**: si el experimento se repitiera un número grande de veces, una gran proporcion de los intervalos construidos de la manera propuesta contendría el valor real del parámetro; a esa proporción se le llama **nivel de confianza** $=(1-\alpha)$ (α es pequeña).

Ejemplo: Se tomaron 10 muestras de tamaño 7 de una población $N(\mu = 100, \sigma^2 = 100)$. Los IC del 95 % son

$$\overline{X_i} \pm 1.96 \cdot \frac{10}{\sqrt{7}} \equiv \overline{X_i} \pm 7.41 \quad i = 1, 2, \ldots, 10.$$

Así la longitud de todos los intervalos es de 14.82, pero el centro cambia con el valor de $\overline{X_i}$.

Intervalos de Confianza



Observa que la probabilidad de que el primer intervalo contenga a μ es 1, mientras que es cero para el séptimo intervalo. La confiabilidad debe entenderse como el porcentaje de veces que los IC's construidos con este método cubren el valor real de μ (manteniendo fijos α, n y σ).

Error en la Estimación

Error en la Estimación

Presentar un intervalo del cual aseguramos con una confianza determinada que contendrá al parámetro, nos da pie para poder hablar del error en nuestra estimación.

La expresión

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

podría interpretarse como la estimación puntual de μ (mediante $\overline{X}$), con un margen de error (incertidumbre, por decirlo como en física cuando se hace una medición) de $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ a la derecha e izquierda de la estimación de μ.

Error en la Estimación

Además, si llamamos

$$\text{Error de estimación} \equiv E \equiv z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

podemos despejar n

$$n = \left[\frac{z_{\alpha/2}\sigma}{E}\right]^2$$

lo cual nos da una forma de obtener el tamaño de muestra que genera el IC con confiabilidad de $(1-\alpha)$ dado un error fijado por el investigador.

Error en la Estimación

En la última expresión, n crece si:

- $z_{\alpha/2}$ crece, esto es, si aumenta la confiabilidad ($z_{\alpha/2}$ crecerá al aumentar $1-\alpha$).
- E disminuye, esto es, si queremos tener un error de estimación muy pequeño.
- La variabilidad, σ^2, de la población es grande.

Ejemplo: Con $\sigma^2 = 1$ y $E = 1$ tenemos para dos niveles de confianza

$1-\alpha = 0.9$	$1-\alpha = 0.95$
n=2.72$\rightarrow$ 3	n=3.84$\rightarrow$ 4

y con $\sigma^2 = 1$ y $1-\alpha = 0.95$ tenemos para dos errores de estimación

$E = \frac{1}{2}$	$E = 0.1$
n=15.36$\rightarrow$ 16	n=384.16$\rightarrow$ 385

en donde se redondea al siguiente entero.

Error en la Estimación

Ejercicio:

1. Encuentra en el anterior ejemplo los valores $z_{\alpha/2}$ utilizados.
2. Si se quisiera tener un IC del 100 % de confiabilidad, ¿qué pasa con $z_{\alpha/2}$? ¿Tendría este intervalo algún interés práctico?

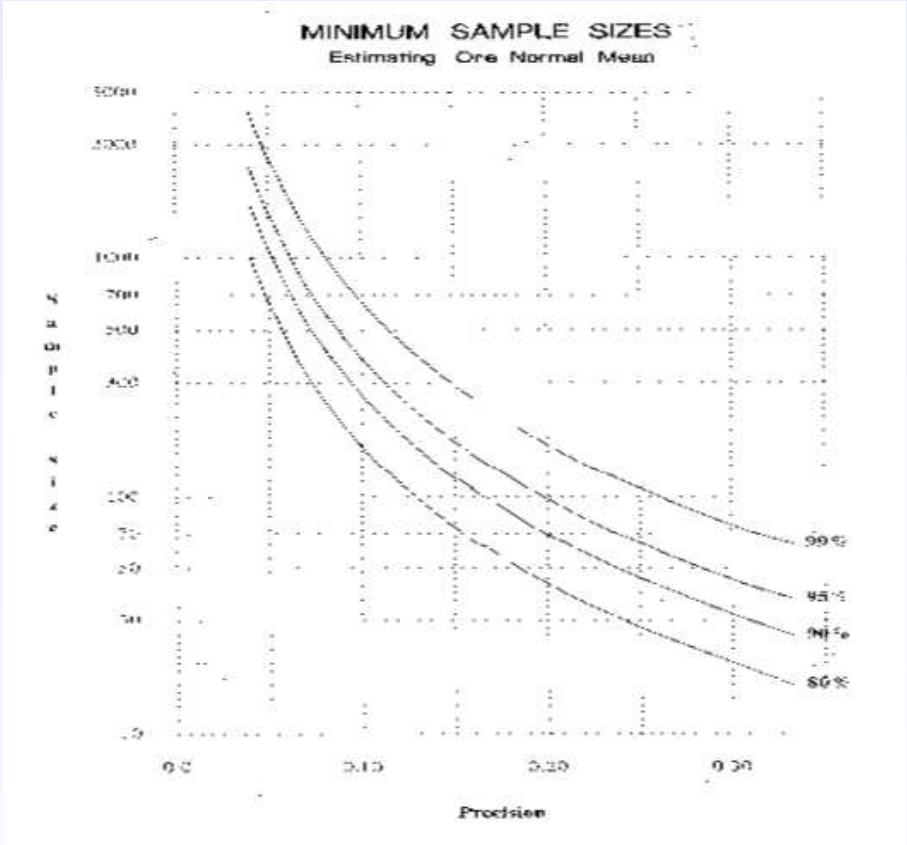
En la práctica, generalmente no se conoce el valor de σ^2. Sin embargo, aún se puede hallar un tamaño de muestra para una confianza y un error especificados. Se acostumbra trabajar el error en términos de desviaciones estándar, dado que así

$$n = \left[\frac{z_{\alpha/2}\sigma}{\delta\sigma}\right]^2 = \left[\frac{z_{\alpha/2}}{\delta}\right]^2 \quad \text{no depende de } \sigma$$

Aquí $E = \delta\sigma$, a δ se le conoce como la precisión. La siguiente gráfica ilustra lo anterior. [1]

[1] Brush, Gary G. (1988). How to Choose the Proper Sample Size. American Society for Quality Control. Vol. 12

Error en la Estimación



Observa que el pivote que hemos utilizado tiene la forma.

$$\frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}}.$$

Error en la Estimación

Esta estandarización aparece usualmente en la práctica (reemplazando las cantidades pertinentes según la teoría, se establece el comportamiento probabilístico de este cociente). Sin embargo, no es la única o la más conveniente en algunos casos.

Ejemplo: Un modelo teórico sugiere que el tiempo de ruptura de un fluido aislante entre electródos de un voltaje particular, sigue una distribución exponencial con parámetro β. Una m.a. de tiempos (en minutos) quedó determinada por:

41.53, 18.73, 2.99, 30.34, 12.33, 117.52, 73.02, 223.63, 4.00, 26.78.

Construir un IC del 95 % para el parámetro β.

Error en la Estimación

Aquí, $X = \textit{Tiempo de ruptura} \sim \textit{Exp}(\beta)$ y los datos representan la realización de una m.a. de tamaño 10. En este caso no será conveniente utilizar

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X} - \beta}{\frac{\beta}{\sqrt{n}}}, \qquad (V(X) = \beta^2)$$

ya que, del capítulo de transformaciones sabemos que, $Y = \sum_{i=1}^{10} X_i$ depende de la m.a. (como debe de ser), pero $Y \sim \textit{Gamma}(\alpha = 10, \beta)$ ($\neq$ normal).

Además, $W = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{10} X_i = \overline{X} \sim \textit{Gamma}(10, \frac{\beta}{n})$. Por lo tanto, al estandarizar de la anterior manera no obtendremos una distribución libre del parámetro β (¿Y TLC?). Veremos que se puede estandarizar de otra manera.

Error en la Estimación

Construcción del pivote.

1. Aprovecharemos la propiedad utilizada anteriormente, $Y \sim Gamma(\alpha, \beta)$, $\frac{Y}{n} \sim Gamma(\alpha, \frac{\beta}{n})$ para eliminar el parámetro β de la distribución de W.

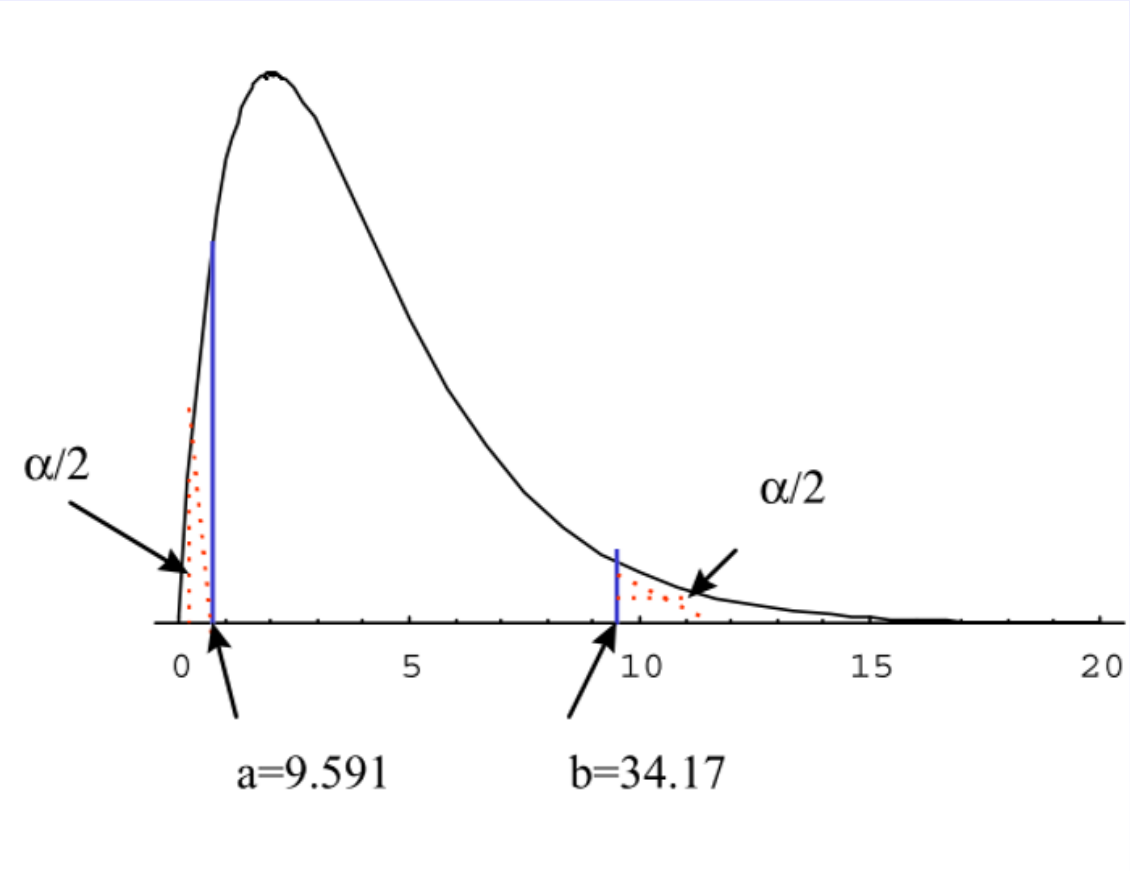
Si escogemos

$$W = \frac{2}{\beta} Y = \frac{2}{\beta} \sum_{i=1}^{10} X_i \sim Gamma(10, \frac{2}{\beta}\beta) = Gamma(10, 2) = \chi^2_{20},$$

esto es, una Ji-cuadrada con 20 grados de libertad ($\chi^2_n \equiv Gamma(\frac{n}{2}, 2)$). Entonces, W puede ser nuestro pivote, ya que depende del parámetro β, de la m. a. (a través de $\sum_{i=1}^{10} X_i$) pero su distribución no depende de β.

Error en la Estimación

2. Para encontrar a y b (percentiles), hacemos uso de las tablas de χ^2_{20} con $1 - \alpha = 0.95$ y de la siguiente gráfica.



Error en la Estimación

Así, tenemos que

$$P\{9.591 < W < 34.17\} = P\{9.591 < \frac{2}{\beta}\sum_{i=1}^{10} X_i < 34.17\} = 0.95$$

1. Ahora, aislamos el parámetro realizando las operaciones necesarias:

$$P\left\{\frac{2\sum_{i=1}^{10} X_i}{34.17} < \beta < \frac{2\sum_{i=1}^{10} X_i}{9.591}\right\} = 0.95$$

de donde, sustituyendo los datos, obtenemos que el IC del 95 % para β es:

$$\left(\frac{2(550.87)}{34.17}, \frac{2(550.87)}{9.591}\right) = (32.24, 114.87).$$

Podemos observar que el intervalo es muy amplio, lo cual refleja la variabilidad intrínseca de los tiempos de ruptura y el hecho de contar con una muestra pequeña.

Error en la Estimación

¿Qué pasó con el TLC? Dada una población cualquiera, sabemos que

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0,1) \quad n \to \infty,$$

por el TLC. De aquí es factible, si la muestra es grande, construir un IC aproximado para la media de cualquier distribución; este sería, con una confianza de $1 - \alpha$

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Intervalos de Confianza para el parámetro μ

Intervalos de Confianza para el parámetro μ

Hemos mostrado :

Caso 1: Población Normal, parámetro μ. σ^2 conocida.
Los pasos que muestra el uso del pivote nos llevaron a concluir que un IC para μ está dado por:

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Caso 2: Población Arbitraria, parámetro μ. σ^2 conocida.
De la discusión final en el ejemplo de los tiempos de ruptura, también debería quedar claro que si la población no es normal pero el tamaño de la muestra es suficientemente grande, podemos aplicar el TLC para determinar la distribución aproximada del pivote y así, siguiendo exactamente los mismos pasos mostrados en el Caso 1: construir el IC aproximado para μ

$$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Intervalos de Confianza para el parámetro μ

Veremos ahora otros casos de interés.

Caso 3: Población Normal, parámetro μ. σ^2 desconocida.
En este caso,

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1),$$

pero no conocemos σ y por lo tanto no podemos evaluar ni el límite superior $l(X_1, \ldots, X_n)$, ni el inferior $u(X_1, \ldots, X_n)$. No obstante, podríamos estimar σ^2 con el valor de S^2 y utilizar $h(X_1, \ldots, X_n; \theta) = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$.

Debería ser claro que este pivote no necesariamente sigue una distribución normal, ya que no aparece σ sino S – ver comentario y derivación de la distribución t en el capítulo cinco.

Intervalos de Confianza para el parámetro μ

Si usamos

$$\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1), \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{indepedendientes}$$

y creamos el cociente de la normal con la raíz cuadrada de la χ^2 dividida por sus grados de libertad, sabemos que esto será una v.a. t-Student con los grados de libertad de la Ji-cuadrada:

$$\frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{S}{\sigma}} = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

Intervalos de Confianza para el parámetro μ

Entonces, nuestro pivote sí puede ser

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}},$$

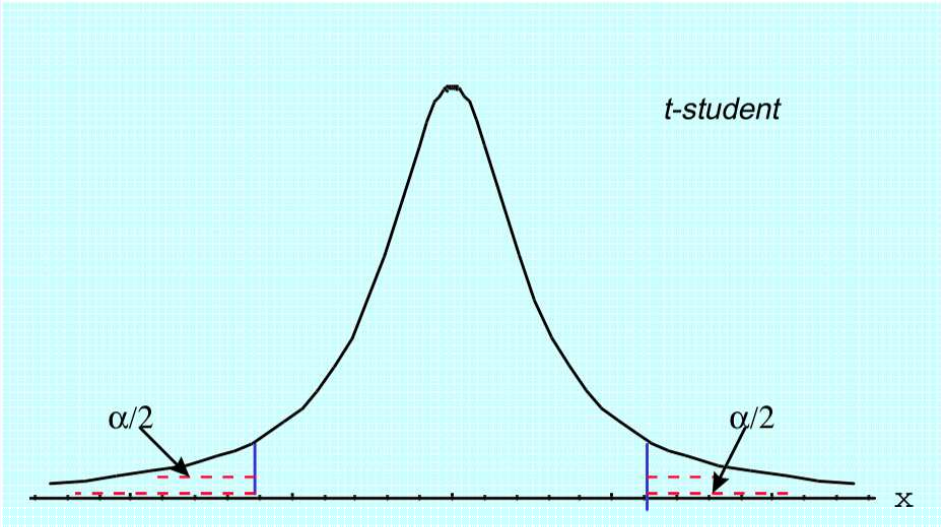
ya que depende de $X_1, X_2, \ldots, X_n$ a través de $\overline{X}$ y su distribución no depende de μ, ni de ningún parámetro desconocido (paso 1 del método). Ubicándonos en el paso 2, tenemos

$$P\left\{a < \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} < b\right\} = 1 - \alpha$$

de donde un IC del 100 % para μ es (del paso 3):

$$\overline{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

Intervalos de Confianza para el parámetro μ



Intervalos de Confianza para el parámetro μ

Caso 4. Población Arbitraria. Parámetro μ. σ^2 desconocida.
Si la muestra es grande también podemos aplicar el TLC de la siguiente manera:

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \sim N(0,1).$$

cuando n grande. La distribución anterior es de forma aproximada. El resultado se basa en el hecho de que S^2 se aproxima con alta probabilidad, al valor de σ^2 cuando es n es grande.

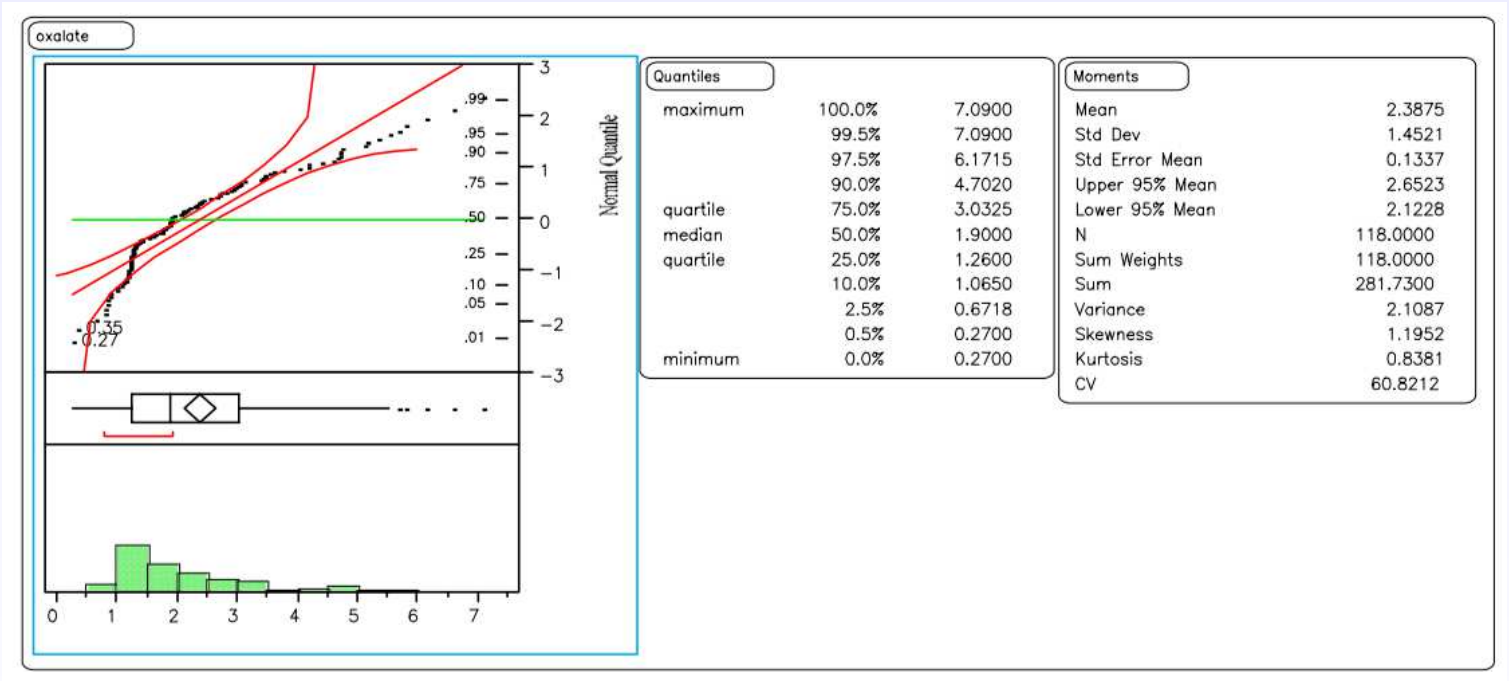
Inclusión de la Variabilidad

Veamos un ejemplo (**repetido**).

Ejemplo: En una rama de la industria alimentaria, se realizan en forma rutinaria mediciones del contenido de calcio en comida para animales (mascotas). El método estándar utiliza precipitación de oxalato de calcio seguida de tritanio; es una técnica que consume tiempo. Los resultados de 118 muestras (Heckman 1960), se dan en el archivo calcio.txt (columna 1). Ahora, con lo que hemos discutido hasta aquí, ¿qué podemos decir a partir del comportamiento de estos valores muestrales?

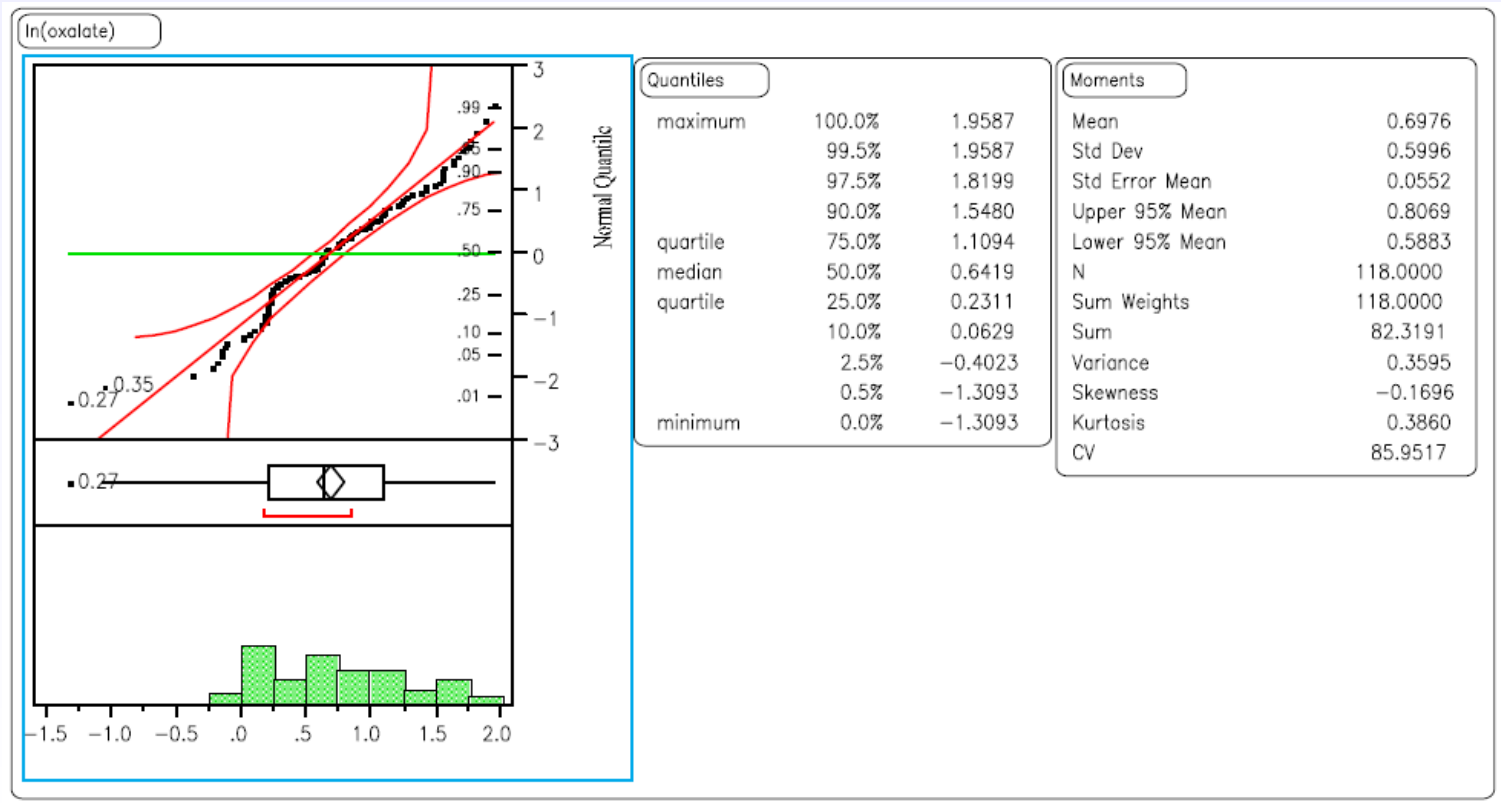
Sabemos cómo ajustar gráficamente un modelo probabilístico, y conocemos, al menos en algunos casos, lo que esperamos de nuestros valores muestrales a través de caracterizar las distribuciones muestrales. Veamos que pasa exactamemte para este ejemplo:

Inclusión de la Variabilidad



Inclusión de la Variabilidad

De las graficas anteriores y de los valores muestrales de simetría, curtosis, el desplazamiento entre la media y la mediana y los posibles valores extremos, es claro que nuestros datos no parecen provenir de una población con distribución Normal. Sin embargo, considerando una transformación de ellos, digamos sus logaritmos, observamos los siguientes cambios:



Inclusión de la Variabilidad

De estos resultados incluso nos atreveríamos a decir que la muestra posiblemente fue obtenida de una población (valores del logaritmo del contenido de calcio) normal, con la posible presencia de dos valores atípicos.

Por otra parte, como la muestra es en particular "grande", podríamos obtener el comportamiento aproximado de cantidades tales como la media muestral y la varianza muestral, usando el TLC. Para fines de ilustración pensaremos que la distribución de la población es normal (y por ende, la distribución de los valores de calcio, sería lognormal).

Inclusión de la Variabilidad

Bajo las condiciones anteriores, ¿qué nos dicen los valores de $\overline{x} = 0.6976$ y de $s^2 = 0.3595$? Estos son los estimadores puntuales de μ y σ^2 en la población. ¿Qué más podemos decir sobre μ y σ^2?

Nos gustaría no sólo describir la variabilidad de nuestros estimadores puntuales, sino aprovechar el hecho de que sabemos que la distribución de los valores de $\hat{X}$ es $Normal(\mu, \frac{\sigma}{\sqrt{n}})$ y que la distribución de S^2 es $\frac{n-1}{\sigma^2}\chi^2$.

Intervalos de Confianza para el parámetro μ

Si pensamos en el ejemplo de los contenidos de calcio, podemos construir el IC para el valor medio de calcio de varias formas:

i) Si consideramos el IC aproximado dado que el tamaño de muestra es suficientemente grande:

$$\overline{x} = 2.3875$$

$$z_{\alpha/2} = 1.96 \quad \alpha = 0.05$$

$$\frac{s}{\sqrt{n}} = \frac{1.4521}{\sqrt{118}} = 0.1337$$

Entonces el IC aproximado del 95 % para μ es : (2.16 , 2.61)

Intervalos de Confianza para el parámetro μ

2) Si consideramos los logaritmos y calculamos directamente el pivote que sigue una t-student:

$$\overline{x}_{\ln x} = 0.6976$$
$$t_{117} = 1.96 \quad \alpha = 0.05$$
$$\frac{s}{\sqrt{n}} = \frac{0.5996}{\sqrt{118}} = 0.0552$$

Entonces el IC del 95 % para $\mu_{\ln x}$es : (0.589, 0.806). E=0.225
Una forma de encontrar el IC para μ, la media de los valores de oxalato de calio es tomando la transformación inversa:

$$\left(e^{0.589}, e^{0.806}\right) = (1.802, 2.239)\ E = 0.2185$$

Intervalos de Confianza para el parámetro μ

Ejercicio: En este último intervalo deberíamos notar que la relación entre la media de la variable original y la de su transformación logarítmica está afectada por una constante que aquí no aparece. Recuerda la relación entre medias y varianzas que hemos establecido al final del capítulo 3, para la Normal y Lognormal. Trata de incorporar esta información para entender mejor la diferencia numérica entre estos intervalos.

Caso 5: En el caso que se utilice el estimador de máxima verosimilitud en una población dada, es posible, construir IC's aproximados:

$$\frac{\hat{\theta}_{MV} - \theta}{\sqrt{V(\hat{\theta}_{MV})}} \dot{\sim} N(0,1).$$

Intervalos de Confianza para el parámetro μ

En la expresión anterior, $V(\hat{\theta}_{MV})$ se puede calcular de manera aproximada como

$$V(\hat{\theta}_{MV}) = \left[- \left. \frac{\partial^2 l(\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}_{MV}} \right]^{-1}.$$

Por ejemplo, en el caso de una población $Exp(\beta)$, sabemos que $\hat{\beta}_{MV} = \overline{X}$, entonces

$$\left. \frac{\partial^2 l(\beta)}{\partial^2 \beta} \right|_{\beta=\hat{\beta}_{MV}} = \frac{n}{\overline{x}^2} - \frac{2n\overline{x}}{\overline{x}^3} = -\frac{n}{\overline{x}^2}.$$

Así,

$$V(\hat{\beta}_{MV}) = \left\{ - \left(-\frac{n}{\overline{x}^2} \right) \right\}^{-1} = \frac{\overline{x}^2}{n},$$

o sencillamente,

$$V(\hat{\beta}_{MV}) = V(\overline{X}) = \frac{\hat{\beta}^2_{MV}}{n}.$$

Intervalos de Confianza para el parámetro μ

Así, usando la normalidad asintótica del estimador de MV, tenemos

$$\frac{\hat{\beta}_{MV} - \beta}{\sqrt{V(\hat{\beta}_{MV})}} = \frac{\overline{X} - \beta}{\frac{\overline{X}}{\sqrt{n}}} \dot{\sim} N(0,1) \quad \text{cuando n es grande.}$$

En general, los IC's para un parámetro θ dado por la la normalidad asintótica del estimador de MV son de la forma

$$\hat{\theta}_{MV} \pm z_{\alpha/2}\sqrt{V(\hat{\theta}_{MV})};$$

así, en el ejemplo anterior se puede construir un intervalo de confianza como

$$\overline{x} \pm z_{\alpha/2}\sqrt{\frac{\overline{x^2}}{n}} = \overline{x} \pm z_{\alpha/2}\frac{\overline{x}}{\sqrt{n}}.$$

Estos intervalos son muy usados en la práctica estadística, dado que utilizan estimadores que son eficientes (al menos para muestras grandes); esto es, estimadores que tienen la mínima varianza y por lo mismo garantizan el menor error de estimación posible.

Más sobre Máxima Verosimilitud

Recordemos y observemos lo siguiente sobre la función de verosimilitud:

- La función de verosimilitud se define como

$$L_n(\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

  Observa que la verosimilitud es una función de los parámetros.

- La log-verosimilitud se define como

$$l_n(\theta) = \log L_n(\theta).$$

- $L_n(\theta) : \Theta \to [0, \infty)$ y no es una función de densidad, i.e. integrar con respecto a θ no integra 1.
- $\hat{\theta}_n = \hat{\theta}_{MLE}$ es el valor que maximiza a $L_n(\theta)$.
- Si $\hat{\theta}_n$ maximiza a $L_n(\theta)$, también maximiza a $l_n(\theta)$.

Más sobre Máxima Verosimilitud

Definición

La función "score" se define como

$$S(x;\theta) = \frac{\partial \log f(x;\theta)}{\partial \theta}$$

y la información de Fisher se define como

$$\mathcal{I}_n(\theta) = V_\theta\left(\sum_{i=1}^{n} S(x_i;\theta)\right) = \sum_{i=1}^{n} V_\theta\left(S(x_i;\theta)\right).$$

Para $n = 1$ se escribe simplemente como $\mathcal{I}(\theta)$. Se puede demostrar que $E_\theta(S(x;\theta)) = 0$. De aquí se sigue que $V_\theta(S(x;\theta)) = E_\theta(S^2(x;\theta))$.

Más sobre Máxima Verosimilitud

En efecto,

$$E_\theta[S(x;\theta)] = \int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)dx = \int \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial \theta} f(x;\theta)dx$$
$$= \int \frac{\partial f(x;\theta)}{\partial \theta} dx \overset{\text{cond. de reg.}}{=} \frac{\partial}{\partial \theta} \int f(x;\theta)dx = \frac{\partial}{\partial \theta}(1) = 0.$$

Además,

$$\mathcal{I}_n(\theta) = n\mathcal{I}(\theta),$$

$$\mathcal{I}(\theta) = -E_\theta\left(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\right) = -\int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)dx.$$

Todo esto bajo **condiciones de regularidad**[2].

[2] Ver el documento Regularity_Conditions.pdf

Más sobre Máxima Verosimilitud

Para justificar nuestra aseveración anterior, demostraremos

$$E(S^2(x;\theta)) = E\left[\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right].$$

Notemos lo siguiente

$$\frac{\partial}{\partial\theta}\log f(x,\theta) = \frac{1}{f(x,\theta)}\frac{\partial}{\partial\theta}f(x,\theta).$$

Así,

$$\begin{aligned}
E\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right] &= \int \frac{f(x;\theta)\frac{\partial^2}{\partial\theta^2}f(x;\theta) - [\frac{\partial}{\partial\theta}f(x;\theta)]^2}{f^2(x,\theta)} f(x,\theta)dx \\
&= \int \frac{\partial^2 f(x;\theta)}{\partial\theta^2}dx - \int \frac{1}{f^2(x;\theta)}[\frac{\partial}{\partial\theta}f(x;\theta)]^2 f(x;\theta)dx \\
&= \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx - \int [\frac{\partial}{\partial\theta}\log f(x;\theta)]^2 f(x;\theta)dx \\
&= -E[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2].
\end{aligned}$$

Más sobre Máxima Verosimilitud

Supongamos que θ_0 el verdadero valor de θ y que $\hat{\theta}_n = \hat{\theta}_{MV}$.

Teorema (Distribución asintótica del MLE)

$$\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,1)$$

A continuación haremos un bosquejo de la prueba de este resultado.

Tenemos que

$$0 = \frac{\partial l(\hat{\theta}_n)}{\partial \theta} \doteq \frac{\partial l(\theta_0)}{\partial \theta} + \frac{\partial^2 l(\theta_0)}{\partial \theta^2}(\hat{\theta}_n - \theta_0),$$

lo cual implica que

$$\hat{\theta}_n - \theta_0 = -\left[\frac{\partial^2 l(\theta_0)}{\partial \theta^2}\right]^{-1}\left[\frac{\partial l(\theta_0)}{\partial \theta}\right].$$

Más sobre Máxima Verosimilitud

De donde

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left[-\frac{1}{n}\frac{\partial^2 l(\theta_0)}{\partial \theta^2}\right]^{-1}\left[-\frac{1}{\sqrt{n}}\frac{\partial l(\theta_0)}{\partial \theta}\right].$$

Notemos que

$$\frac{\partial l(\theta_0)}{\partial \theta} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(x_i;\theta)}{\partial \theta_0}, \qquad \text{(por definición de } l_n(\theta))$$

y que

$$E\left[\frac{\partial \log f(x_i;\theta_0)}{\partial \theta}\right] = E[S(x_i;\theta_0)] = 0,$$

$$\text{Var}\left[\frac{\partial \log f(x_i;\theta)}{\partial \theta}\right] = E\left[\left(\frac{\partial \log f(x_i;\theta_0)}{\partial \theta_0}\right)^2\right] = \mathcal{I}(\theta_0).$$

Más sobre Máxima Verosimilitud

Recordatorio:

- TLC: $X_1, \ldots, X_n$ i.i.d. $(0, \sigma^2)$

$$\frac{1}{\sqrt{n}} \sum X_i \xrightarrow{d} N(0, \sigma^2),$$

- Ley Débil de los Grandes Números: $X_1, \ldots, X_n$ independientes (μ, σ^2), entonces

$$\frac{1}{n} \sum_{1}^{n} X_i \xrightarrow{P} \mu.$$

- Slutzky:

$$\begin{cases} X_n \xrightarrow{d} X \\ \\ Y_n \xrightarrow{P} a \end{cases} \Rightarrow X_n Y_n \xrightarrow{d} aX$$

Más sobre Máxima Verosimilitud

Consideremos $W_i = \frac{\partial}{\partial\theta} \log f(x_i;\theta_0)$. Como las X_i's son i.i.d., entonces las W_i's son i.i.d.; además, por lo hecho arriba, $E(W_i) = 0$ y $\text{Var}(W_i) = \mathcal{I}(\theta_0)$. Por lo tanto, el TLC

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \xrightarrow{d} N(0, \mathcal{I}(\theta_0)).$$

Además, tenemos que

$$-\frac{1}{n}\frac{\partial^2 l_n(\theta_0)}{\partial\theta^2} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(x_i;\theta_0), \qquad \text{(por definición de } l_n(\theta))$$

y de acuerdo a la identidad que relaciona la información de Fisher con la segunda derivada de la log-verosimilitud

$$E\left[\frac{\partial^2}{\partial\theta^2} \log f(x;\theta_0)\right] = -\mathcal{I}(\theta_0).$$

Más sobre Máxima Verosimilitud

Así, por la Ley de Grandes Números

$$-\frac{1}{n}\frac{\partial^2 l_n(\theta_0)}{\partial\theta^2} \xrightarrow{P} \mathcal{I}(\theta_0). \qquad \text{(la varianza)}$$

Por el Teorema de Slutsky

$$\sqrt{n}(\hat{\theta}-\theta_0) \doteq \underbrace{\left[-\frac{1}{n}\frac{\partial^2 \log(\theta_0)}{\partial\theta^2}\right]^{-1}}_{\xrightarrow{P}\mathcal{I}(\theta_0)} \underbrace{\left[\frac{1}{\sqrt{n}}\frac{\partial l_n(\theta_0)}{\partial\theta}\right]}_{\xrightarrow{d}N(0,\mathcal{I}(\theta_0))},$$

lo que implica que

$$\sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{d} N(0,\mathcal{I}^{-1}(\theta_0)).$$

Más sobre Máxima Verosimilitud

Para propósitos prácticos

$$\hat{\theta} \dot{\sim} N(\theta_0, [n\mathcal{I}(\theta_0)]^{-1})$$

o

$$\sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \dot{\sim} N(0, 1).$$

Además, si $\mathcal{I}(\theta)$ es una función continua, de la consistencia de los estimadores de máxima verosimilitud se sigue que

$$\mathcal{I}(\hat{\theta}_n) \xrightarrow{P} \mathcal{I}(\theta_0).$$

Así, el Teorema de Slutzky implica que

$$\sqrt{n\mathcal{I}(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) = \frac{\sqrt{\mathcal{I}(\hat{\theta}_n)}}{\sqrt{\mathcal{I}(\theta_0)}} \sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, 1).$$

Más sobre Máxima Verosimilitud

Lo anterior explica porque aproximamos $V(\hat{\theta}_{MV})$ por $\mathcal{I}(\hat{\theta}_{MV})^{-1}$.

Las ideas anteriores se pueden extender al caso multiparametral.

Sea $\theta = (\theta, \ldots, \theta_k)$ y $\hat{\theta} = (\hat{\theta}, \ldots, \hat{\theta}_k)$ el estimador de MLE. Sea $l_n = \sum_{i=1}^{n} \log f(x_i; \theta)$, $H_{jj} = \frac{\partial^2 l_n}{\partial \theta_j^2}$, $H_{jk} = \frac{\partial^2 l_n}{\partial \theta_j \partial \theta_k}$. La matriz de Información de Fisher se define como

$$\mathcal{I}_n(\theta) = - \begin{bmatrix} E_\theta(H_{11}) & E_\theta(H_{12}) & \cdots & E_\theta(H_{1k}) \\ \vdots & \vdots & \ddots & \vdots \\ E_\theta(H_{k1}) & E_\theta(H_{k2}) & \cdots & E_\theta(H_{kk}) \end{bmatrix}.$$

y definamos $J_n = \mathcal{I}_n^{-1}(\theta)$ el inverso de $\mathcal{I}_n$.

Más sobre Máxima Verosimilitud

Teorema

Bajo apropiadas condiciones de regularidad

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \mathcal{I}_1(\theta)^{-1}).$$

Si θ_j es la componente j-ésima de $\hat{\theta}$, entonces

$$\frac{(\hat{\theta}_j - \theta_j)}{\hat{se}_j} \xrightarrow{d} N(0,1)$$

donde $\hat{se}_j^2 = J_n(j,j)$ es el j-ésimo elemento de la diagonal. La covarianza aproximada de $\hat{\theta}_j$ con $\hat{\theta}_k$ es $Cov(\hat{\theta}_j, \hat{\theta}_k) \approx J_n(j,k)$.

Más sobre Máxima Verosimilitud

Optimalidad.

Supongamos que $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. El MLE $\hat{\theta}_n = \overline{X}_n$. Otro estimador razonable de θ sería la mediana, digamos $\tilde{\theta}$.

El MLE cumple con

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

y se puede mostrar que la mediana también converge a θ pero

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2 \frac{\pi}{2}).$$

Más sobre Máxima Verosimilitud

En general, dado dos estimadores (uno de MLE) para el mismo parámetro, definimos su eficiencia relativa (asymptotic relative efficiency) como

$$\text{ARE}(\tilde{\theta}_n, \hat{\theta}_n) = \frac{V(\hat{\theta}_n)}{V(\tilde{\theta}_n)},$$

siempre y cuando el modelo sea **correcto**.

Así, para el caso anterior

$$\text{ARE}(\tilde{\theta}_n, \hat{\theta}_n) = \frac{2}{\pi} \approx 0.64 \leq 1,$$

lo cual indica que la mediana solo usa una fracción de los datos!

Interpretación: Si $\tilde{\theta}_n$ se usa con un tamaño de muestra n, el número de observaciones necesarias para que $\hat{\theta}_n$ se desepeñe equivalentemente es $\text{ARE}(\tilde{\theta}_n, \hat{\theta}_n) \times n$.

Más sobre Máxima Verosimilitud

Teorema

Si $\hat{\theta}_n$ es el estimador de máxima verosimilitud y $\tilde{\theta}_n$ es cualquier otro estimador con

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(0, t^2),$$

entonces

$$ARE(\tilde{\theta}, \hat{\theta}) \leq 1.$$

Decimos que el estimador de máxima verosimilitud es eficiente o asintóticamente óptimo.

Más sobre Máxima Verosimilitud

Método Delta.

Supongamos que

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

y que g es una función diferenciable tal que $g'(\mu) \neq 0$. Entonces

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{d} N(0,1).$$

En otras palabras $Y_n \dot{\sim} N(\mu, \sigma^2/n) \Rightarrow g(Y_n) \dot{\sim} N(g(\mu), g'(\mu)^2\sigma^2/n)$.

Dem. (bajo el supuesto de que g es continuamente diferenciable):
Del Teorema de Taylor se sigue que

$$g(Y_n) = g(\mu) + g'(\mu^*)(Y_n - \mu),$$

donde $\mu^* \in (Y_n, \mu)$. Ya que $Y_n \xrightarrow{P} \mu \Rightarrow \mu^* \xrightarrow{P} \mu \Rightarrow g'(\mu^*) \xrightarrow{P} g'(\mu)$.

Más sobre Máxima Verosimilitud

Así,

$$g(Y_n) - g(\mu) = g'(\mu^*)(Y_n - \mu)$$

de donde

$$\sqrt{n}(g(Y_n) - g(\mu)) = g'(\mu^*)\sqrt{n}(Y_n - \mu).$$

Por lo tanto,

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{d} N(0, g'^2(\mu)\sigma^2).$$

Ejemplo: $X_1, \ldots, X_n$ i.i.d. $(\mu, \sigma^2 > 0)$.

Entonces, el TLC implica que $\sqrt{n}(\overline{X}_n - \mu)/\sigma \dot{\sim} N(0,1)$. Sea $W_n = e^{\overline{X}_n}$, es decir que $W_n = g(\overline{X}_n)$ con $g(s) = e^s$. Notemos que $g'(s) = e^s$. El método delta implica que $W_n \dot{\sim} N(e^\mu, e^{2\mu}\sigma^2/n)$.

Más sobre Máxima Verosimilitud

De la primera igualdad en la diapositiva anterior, podemos notar que

$$E(g(Y_n) - g(\mu))^2 = g'(\mu^*)E(Y_n - \mu)^2,$$

luego

$$\text{Var}(g(Y_n)) = g'(\mu^*)\text{Var}(Y_n).$$

Así obtenemos un método para aproximar la varianza bajo transformaciones suaves.

Más sobre Máxima Verosimilitud

Teorema (Método Delta Multivariado)

Supongamos que $Y_n = (Y_{n1}, \ldots, Y_{nk})$ *una sucesión de vectores aleatorios tales que*

$$\sqrt{n}(Y_n - \mu) \xrightarrow{d} N(0, \Sigma).$$

Sea $g : \mathbb{R}^k \to \mathbb{R}$ *de clase* C^1 *con gradiente*

$$\nabla g(y) = \begin{bmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{bmatrix}.$$

Denotemos por $\nabla_\mu = \nabla g(\mu)$ *y asumamos que los elementos de* ∇_μ *son distintos de 0. Entonces*

$$\sqrt{n}(g(Y_n) - g(\mu)) \xrightarrow{d} N(0, \nabla_\mu^T \Sigma \nabla_\mu).$$

Más sobre Máxima Verosimilitud

Ejemplo: Sean $(X_{11}, X_{21})', (X_{12}, X_{22})', \ldots, (X_{1n}, X_{2n})'$ i.i.d. con $\mu = (\mu_1, \mu_2)'$ y varianza Σ. Sean

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \qquad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}.$$

Definamos $Y_n = \overline{X}_1\overline{X}_2 = g(\overline{X}_1, \overline{X}_2)$, i.e. $g(s_1, s_2) = s_1 s_2$. Por el TLC sabemos que

$$\sqrt{n}(\overline{X}_1 - \mu_1, \overline{X}_2 - \mu_2)' \xrightarrow{d} N(0, \Sigma).$$

Como

$$\nabla g(s)' = (\frac{\partial g}{\partial s_1}, \frac{\partial g}{\partial s_2})' = (s_2, s_1)',$$

entonces

$$\nabla'_\mu \Sigma \nabla_\mu = [\mu_2, \mu_1] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} \mu_2 \\ \mu_1 \end{bmatrix} = \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22} \equiv a.$$

Más sobre Máxima Verosimilitud

Por lo tanto

$$\sqrt{n}(\overline{X}_1\overline{X}_2 - \mu_1\mu_2) \xrightarrow{d} N(0, a).$$

Más sobre Máxima Verosimilitud

Invarianza.

Sea $\tau = g(\theta)$ donde g es una función suave. El estimador de Máx. Veros. de τ es $\hat{\tau} = g(\hat{\theta})$. (Una propiedad super deseable).

Ejemplo: El estimador de MLE de σ^2 es $\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$ en el caso normal, lo que implica que $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ es el estimador de ML de la desviación estándar de σ.

La pregunta es, ¿qué distribución tiene este nuevo estimador $\hat{\tau}$?

El método delta implica que

$$\frac{\hat{\tau} - \tau}{\hat{se}(\hat{\tau})} \xrightarrow{d} N(0,1)$$

donde $\hat{\tau}_n = g(\hat{\theta}_n)$ y $\hat{se}(\tau_n) = |g'(\theta)|se(\hat{\theta}_n)$.

Más sobre Máxima Verosimilitud

Así, un intervalo de confianza del $(1-\alpha)\,\%$ está dado por

$$C_n = (\hat{\tau}_n - z_{\alpha/2}\hat{se}(\hat{\tau}_n), \hat{\tau}_n + z_{\alpha/2}\hat{se}(\hat{\tau}_n)),$$

es decir que

$$P_\theta(\tau \in C_n) \to 1-\alpha \qquad \text{cuando} \quad n \to \infty.$$

Nota: La varianza asintótica de $\hat{\theta}_n$

$$[n\mathcal{I}(\theta)]^{-1} = \left[-nE_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right]^{-1}\right] = \left[-E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right]^{-1}\right]$$
$$= -\left[E\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right]^{-1}\right].$$

Más sobre Máxima Verosimilitud

Ejemplo: Sea $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Sea $\tau = g(\mu, \sigma) = \sigma/\mu$. Entonces,

$$\mathcal{I}_n(\mu, \sigma^2) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}, \qquad (\textit{Demostrarlo!})$$

lo que implica que $J_n = \mathcal{I}_n^{-1}(\mu, \sigma^2) = \frac{1}{n}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2/2 \end{bmatrix}$. Luego, como

$$\nabla g = (-\sigma/\mu^2, 1/\mu)',$$

por lo tanto

$$\hat{se}(\hat{\tau}) = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{\hat{\mu}^4} + \frac{\hat{\sigma}^2}{2\hat{\mu}^2}}.$$

Error Relativo del Estimador

Hemo visto que la estandarización del tipo

$$\frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}},$$

es muy común.

De hecho, podemos notar que en la construcción del pivote para parámetros de localización como μ, la selección está guiada por el concepto de **error relativo del estimador** con respecto al parámetro:

$$\frac{\text{Error cometido}}{\text{Error Promedio}} \quad \acute{o} \quad \frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}} \quad \acute{o} \quad \frac{\text{Error}}{\text{Dispersión del error}}.$$

Intervalos de confianza para σ^2

INTERVALOS DE CONFIANZA PARA EL PARÁMETRO σ^2.

En el caso de parámetros de variabilidad, tales como σ^2, se sugiere medir

$$\frac{\hat{\theta}}{\overline{\theta}},$$

pues es posible, bajo condiciones adecuadas, generar la distribución del estimador (y/o la del pivote).

Intervalos de confianza para σ^2

Caso 1: Población Normal. Intervalo de Confianza para la Varianza.

Dada una m.a. $X_1, X_2, \ldots, X_n$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

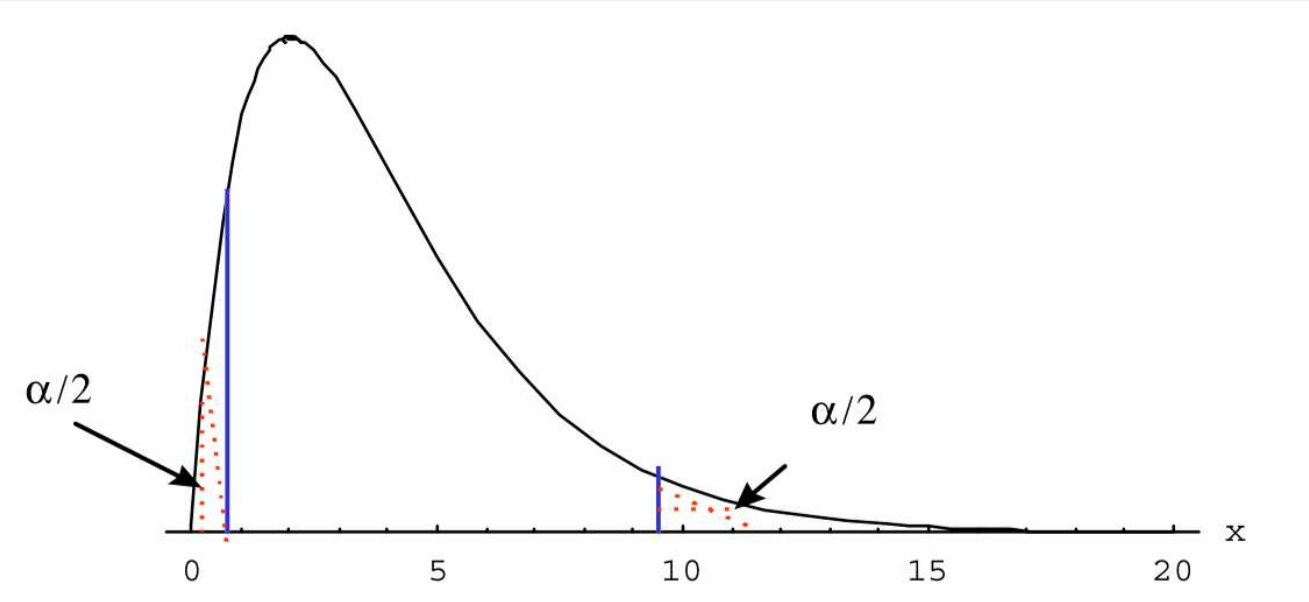
De aquí vemos que

$$\frac{\hat{\theta}}{\theta} = \frac{S^2}{\sigma^2} = (\frac{1}{n-1})\chi^2_{n-1}$$

es un múltiplo de una χ^2, por lo que $\frac{(n-1)S^2}{\sigma^2}$ es un pivote para la estimacion de σ^2.

Intervalos de confianza para σ^2

Consideremos $P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$, que gráficamente es



donde se encontró que $a = \chi^2_{1-\alpha/2}$ y $b = \chi^2_{\alpha/2}$.

Intervalos de confianza para σ^2

Tomando recíprocos en la desigualdad anterior se obtiene

$$P\left(\frac{1}{\chi^2_{\alpha/2}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha,$$

multiplicando ahora por $(n-1)S^2$ queda

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha.$$

Así, el IC del $100(1-\alpha)\,\%$ de confianza para σ^2 es

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right].$$

Intervalos de confianza para σ^2

Una forma simple de construir un IC para la desviación estándar, en una población normal, es sencillamente tomar la raíz cuadrada de los valores del intervalo anterior.

Así, el IC del $100(1-\alpha)\,\%$ de confianza para σ es

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}\right].$$

Existen otras formas de construir intervalos para σ, pero ésta es la más sencilla de todas y sus resultados son aceptables.

Intervalos de confianza para σ^2

Caso 2: Población Arbitraria. Intervalo de Confianza Aproximado para la varianza.

Cuando la población no sigue una distribución normal, pero el tamaño es grande, una forma de construir un IC aproximado para σ^2 es:

1. $S^2 \to \sigma^2$ es un estimador consistente.
2. Si $E[(X-\mu)^4]=\alpha_4\sigma^4$ existe (bajo poblaciones normales $E[(X-\mu)^4] = 3\sigma^4$), se propuso un estimador para α_4 mediante

$$d = \frac{n(n+1)\sum_{i=1}^{n}(X_i-\overline{X})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)};$$

d estima la curtosis.

Intervalos de confianza para σ^2

3. Sea

$$c = (1 + \frac{1}{2}d)^{-\frac{1}{2}},$$

entonces $\frac{c(n-1)S^2}{\sigma^2} \sim \chi^2_{c(n-1)}$ y procediendo como en el anterior ejemplo, un IC de $1-\alpha$ aproximado para σ^2 es

$$\left[\frac{c(n-1)S^2}{\chi^2_{c,\alpha/2}}, \frac{c(n-1)S^2}{\chi^2_{c,1-\alpha/2}}\right],$$

donde $\chi^2_{c,\alpha/2}, \chi^2_{c,1-\alpha/2}$ indican que han sido calculadas a partir de $\chi^2_{c(n-1)}$.

Existen otras alternativas como el procedimiento de Jackknife, Bootstrap, procedimientos no-paramétricos, etc.

Estimación por intervalos para dos poblaciones

Comparación de Dos Poblaciones

Comparación de Dos Poblaciones. Muestras Independientes.

Antes de continuar, cabe mencionar el significado de independencia entre dos poblaciones.

Por ejemplo, si quisiéramos comparar las resistencias de dos materiales de construcción, digamos dos tipos de vigas, podríamos tomar n_1 elementos de la viga 1, y n_2 elementos de la viga 2, aplicándoles esfuerzos a cada elemento hasta romperlos.

Si nuestra máquina aplicadora de esfuerzos se calienta de manera indeseable, podemos estar obteniendo mediciones erróneas; si cada viga se fabricó con alguna variación en sus especificaciones, podemos estar registrando esta variación indeseable en la comparación.

Comparación de Dos Poblaciones

Una forma de evitar estos casos indeseables sería tomar un sólo especímen de cada tipo, dividirlo en secciones y tomar varias mediciones, si es posible; claro, en forma aleatoria.

Generalmente son las condiciones experimentales las que nos dicen si asumir independencia.

Es conveniente mencionar que cuando se habla de experimentos independientes, de base se asume que las unidades del material sobre el que trabajamos son homogéneas en principio y que las acciones que le aplicamos (tratamientos) son lo único en lo que pueden diferir, por lo que es factible compararlos sin más, esto es, todas las otras causas posibles de variabilidad han sido aisladas o controladas de alguna manera.

Comparación de Dos Poblaciones

Cuando esto no es posible, entonces hablamos de poblaciones no independientes y lo que regularmente hacemos es partir nuestros elementos bajo estudio en grupos que sean muy parecidos dentro de sí, y lo más diferente posible entre sí. Este es el principio de lo que llamamos "bloques" en el área de Diseño de Experimentos.

La aplicación de cada "tratamiento = experimento aleatorio" es lo que genera nuestras poblaciones bajo estudio. En este curso nos concentraremos solo en la comparación de dos poblaciones.

Comparación de Dos Poblaciones

Caso 1: Se quiere determinar si dos formas de fierro son retenidas de manera distinta por el organismo. La razón de ello es que la forma de fierro (entre Fe^{2+} y Fe^{3+}), será la que se recomiende para su uso como complemento dietético.

Algunos elementos del Depto. de Industrias Alimentarias fueron llamados para hacer el estudio. Ellos decidieron plantear el siguiente experimento: 36 ratones con características similares fueron divididos en 2 grupos. De aquí, en forma aleatoria se seleccionó un grupo para el Fe^{3+} y el otro para el Fe^{2+}.[3] Se usó una sola concentración 10.2 milimolares.

[3] En el experimento original se usaron 3 concentraciones distintas para cada tipo de Fierro, lo cual nos da un total de 6 poblaciones a comparar en formas bastantes más interesantes. La metodología para ello es desarrollada en cursos posteriores.

Comparación de Dos Poblaciones

Así, tenemos definidas nuestras dos poblaciones bajo estudio. Se suministró el fierro en forma oral, un tiempo después se tomó un conteo (el fierro es medido mediante algunas técnicas de conteos radioactivos) y se calculó la cantidad de fierro retenida tomando la diferencia entre los conteos inicial y final.

Los datos aparecen en el archivo **iron.txt** (trabajarlo, los datos están en el site del curso).

Comparación de Dos Poblaciones

Caso 2: Se realizó un experimento para comparar dos métodos de medición del contenido de calcio en comida para animales. El método estándar usa precipitación de oxalato de calcio seguida de titration y es bastante tardada. Otro método que utiliza una flama fotométrica es bastante más rápido. Se realizó la medición del porcentaje de calcio en 118 muestras típicas empleando ambos métodos en cada muestra.

Los datos se encuentran en el archivo **calcio.txt** (trabajar los datos, los datos están en el site).

En el primer caso, todos los ratones se consideran material "homogéneo" y posteriormente la aplicación de los diferentes tratamientos forma nuestras poblaciones bajo estudio. Entonces decimos que estamos trabajando con **poblaciones independientes**.

Comparación de Dos Poblaciones

En el segundo caso, para hacer los métodos comparables, éstos son aplicados sobre la misma unidad, esto es, cada unidad es trabajada dos veces (en ocasiones se parten las unidades experimentales básicas y cada una de las subpartes es trabajada con cada método), entonces tendremos un total de 118 resultados, cada uno de ellos correspondiendo a las dos lecturas realizadas por cada método.

Notemos que en este segundo caso no nos interesa el valor medio del contenido de calcio en cada muestra, sino que las lecturas coincidan, por esa razón nuestros datos serán las parejas de valores y lo que estudiaremos son las diferencias de esos valores. Aquí, estamos trabajando con **poblaciones que no son independientes**.

Comparación de Dos Poblaciones

Otros casos típicos de poblaciones con dependencias son los que se clasifican como "antes y después". Por ejemplo, la presión arterial de pacientes antes y después de aplicar algún medicamento. De esa forma, la variabilidad entre cada individuo es separada al tomar por ejemplo las diferencias de estas lecturas.

Algunos ejemplos de aplicación en poblaciones que pueden ser consideradas como independientes son

- Aplicación de dos tratamientos a un especímen.
- Análisis de dos bandas de producción.
- Dos métodos de trabajo, etc.

Comparación de Dos Poblaciones

Siempre deben verificarse las condiciones experimentales para poder concluir sobre la independencia de éstas.

Más aún, las cosas comienzan en el otro extremo: dadas las mediciones que uno tiene la factibilidad de realizar, con el objeto de responder a las preguntas esenciales del problema, se lleva a cabo un plan de diseño que permita controlar las fuentes de variabilidad, si es que las hay. En otras palabras, somos nosotros quienes decidimos si estaremos trabajando con poblaciones independientes o no.

Comparación de Dos Poblaciones

¿En qué sentido comparamos las poblaciones?

Las poblaciones se pueden comparar respecto a su nivel medio o respecto a su variabilidad.

Una forma de hacer esta comparación es construyendo un IC para

$$\mu_1 - \mu_2 \qquad \acute{o} \qquad \frac{\sigma_1^2}{\sigma_2^2}.$$

Lo que acontinuación haremos es desarrollar algunos de los esquemas más sencillos para dichas comparaciones.

Comparación de Dos Poblaciones

Población 1: Distribución $N(\mu_1, \sigma_1^2)$ se toma una m.a. $X_1, X_2, \ldots, X_{n_1}$.

$\Uparrow$

INDEPENDIENTES

$\Downarrow$

Población 2: Distribución $N(\mu_2, \sigma_2^2)$ se toma una m.a. $Y_1, Y_2, \ldots, Y_{n_2}$.

(Notemos que cuando las poblaciones son independientes los tamaños de muestra no necesitan ser iguales)

Comparación de Dos Poblaciones

Primero con respecto a su nivel medio: $\mu_1 = \mu_2$.

Nota Importante: El fin primario de la técnica de IC es el de darnos un conjunto de valores plausible para nuestro parámetro de interés, pero muchas veces lo usamos de la siguiente forma:
Si estamos más que nada interesados en ver si $\mu_1 - \mu_2 = 0$, i.e. si las poblaciones tienen la misma respuesta media, si el IC contiene el valor 0 entonces diremos, con una confiabilidad del 100(1- α) %, que las poblaciones tienen el mismo comportamiento medio.

El mismo razonamiento se puede aplicar para otros valores de $\mu_1 - \mu_2$.

Comparación de Dos Poblaciones

En cada caso,

$$E(\overline{X}-\overline{Y}) = \mu_1-\mu_2, \quad V(\overline{X}-\overline{Y}) = \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}, \quad \overline{X}-\overline{Y} \sim N(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}).$$

a) Si σ_1^2 y σ_2^2 son conocidas:

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1),$$

el cual es pivote para $\mu_1 - \mu_2$.

Comparación de Dos Poblaciones

Entonces,

$$P\left(a < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < b\right) = 1 - \alpha,$$

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha;$$

y efectuando operaciones, el intervalo del $100(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$ con σ_1^2 y σ_2^2 conocidas es

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Comparación de Dos Poblaciones

b) Si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ conocida, se puede proceder como en el inciso a), ya que

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1);$$

así, el intervalo del $100(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$ es

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Comparación de Dos Poblaciones

c) Si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconocida, podríamos intentar usar

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}.$$

El problema es que σ^2 no es conocida; sera necesario estimarla.

Sabemos que

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1-1}, \qquad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}.$$

Tenemos dos fuentes de información para estimar a σ^2 y nos gustaría aprovecharlas a ambas, sin perder las propiedades que cada una de ellas tienen, a decir, el insesgamiento y su distribución.

Comparación de Dos Poblaciones

Definamos

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}.$$

Es fácil mostrar que $E(S_p^2) = \sigma^2$ y que $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$.

Más aún, se puede mostrar que esta forma de ponderar a los valores de varianzas muestrales, con respecto al tamaño de muestra, produce un estimador insesgado de mínima varianza.

Recordemos que la precisión de los estimadores es algo de lo que debemos preocuparnos, pues de otra forma, aunque tengamos un procedimiento válido, sus resultados son tan "opacos", que o no nos dicen nada o incluso nos confunden más.

Comparación de Dos Poblaciones

Ejercicio: Muestra las 3 propiedades mencionadas arriba para S_p^2. (Se entrega como tarea el 6 de noviembre.)

Ahora, $\overline{X}$, $\overline{Y}$, S_1^2, S_2^2 son independientes entre sí, de donde $\overline{X} - \overline{Y}$ y S_p^2 también lo son. Entonces

$$\frac{\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\frac{\frac{(n_1-1)S_1^2}{\sigma^2}+\frac{(n_2-1)S_2^2}{\sigma^2}}{(n_1+n_2-2)}}} \sim t_{n_1+n_2-2};$$

o sea que

$$\frac{\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{1}{\sigma}\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}}} \sim t_{n_1+n_2-2}.$$

Comparación de Dos Poblaciones

Sustituyendo S_p^2,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2};$$

así,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

es un pivote para la estimacion de $\mu_1 - \mu_2$.

(Notemos que el resultado final fue sólo sustituir σ por s_p, pero todos los argumentos anteriores nos permiten conocer la distribución de este nuevo pivote, lo cual es crítico para poder formar el IC y más adelante para realizar pruebas de hipótesis).

Comparación de Dos Poblaciones

Entonces, el intervalo de confianza se obtendría como:

$$P\left(-t_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Por lo tanto, el intervalo del $100(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$ con la misma varianza es

$$\overline{x} - \overline{y} \pm t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Comparación de Dos Poblaciones

d) Si $\sigma_1^2 \neq \sigma_2^2$ desconocidas, ya no se puede tomar un S_p^2 pues necesitamos a S_1^2 y S_2^2 para estimar σ_1^2 y σ_2^2 y además

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

no sigue una distribución t (este es el problema de Behrens-Fisher).

Comparación de Dos Poblaciones

Posibles Soluciones:

1) Si n_1 y n_2 son grandes se puede aplicar el TLC, de donde

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1).$$

Este resultado es válido aún en el caso en que las poblaciones no sigan una distribución normal. Entonces el IC del $100(1-\alpha)\,\%$ de confianza aproximado para $\mu_1 - \mu_2$ es

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Comparación de Dos Poblaciones

2) Si la población es aproximadamente normal, una solución fue dada por Welch (y Satterthwaite 1946):

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu,$$

donde

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2}\right]}.$$

Se reducen los grados de libertad de la t.

Comparación de Dos Poblaciones

Pensemos qué implicación tiene esta reducción en los grados de libertad. Recordemos que las gráficas de t con pocos grados de libertad tienen colas más pesadas, esto es, hay mayor variabilidad presente; además, si vemos la construcción de los IC, lo que nos interesa son probabilidades asignadas en las colas de la distribución. Compara los valores de los percentiles para α=0.05 , para diferentes t-Student por ejemplo, con 1, 5, 10 y 20 grados de libertad.

Notarás que esos valores van decreciendo y por lo mismo los IC mostrarán mayor precisión entre más grados de libertad tengamos para nuestra distribución.

Si quieres ser una persona instruida lee, entiende y se capaz de reproducir las notas
EL_NUMERO_EFECTIVO_DE_GRADOS_DE_LIBERTAD.pdf.

Intervalos de Confianza para Proporciones

Intervalos de Confianza para Proporciones.

Si ustedes buscan en Google el tema de intervalos de confianza para p, van a encontrar 119 millones de resultados. Consúltenlos o sigan estás notas ultra-básicas.

1) Si partimos de una población Bernoulli y queremos un IC para p, la proporción de éxitos, podemos utilizar el TLC

$$X_1, X_2, \ldots, X_n \sim Bernoulli(p) \Rightarrow E(X_i) = p,\ V(X_i) = pq.$$

Así,

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1), \qquad \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \sim N(0,1).$$

Entonces, un IC aproximado del $1 - \alpha$ para p es

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Intervalos de Confianza para Proporciones

2) En el caso de dos muestras (grandes) independientes, el IC para $p_1 - p_2$ es

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}} \sim N(0,1).$$

Entonces, el IC del $1 - \alpha$ es

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}.$$

Cocientes de varianzas

Ahora comparemos las poblaciones respecto a su variabilidad.

Cocientes de varianzas.

Se quiere encontrar un intervalo de confianza para $\frac{\sigma_1^2}{\sigma_2^2}$.

En forma análoga al caso de la comparación de medias, diremos que las poblaciones tienen la misma variabilidad si el cociente de varianzas toma el valor de 1. Así, si el IC contiene al uno, diremos que con una confiabilidad del $100(1-\alpha)\,\%$, las poblaciones se comportan en forma similar en cuanto a su variabilidad.

Cocientes de varianzas

Para construir un pivote requerimos dos condiciones fundamentales:

a) **Considerar dos poblaciones normales**

$$N(\mu_1, \sigma_1^2) \quad N(\mu_2, \sigma_2^2).$$

b) **Independientes**
basándose en muestras aleatorias de tamanos n_1 y n_2.

Se sabe que $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$ y éste constituye un pivote natural para $\frac{\sigma_1^2}{\sigma_2^2}$.

$$\Longrightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$$

$$\therefore \quad P\left(F_{1-\alpha/2,n_1-1,n_2-1} < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < F_{\alpha/2,n_1-1,n_2-1}\right) = 1-\alpha.$$

Cocientes de varianzas

Tomando recíprocos

$$P\left(\frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{S_2^2}{S_1^2}\cdot\frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{1-\alpha/2,n_1-1,n_2-1}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{1-\alpha/2,n_1-1,n_2-1}}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{S_1^2}{S_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\cdot F_{\alpha/2,n_2-1,n_1-1}\right) = 1-\alpha.$$

$\therefore$ El intervalo del $100(1-\alpha)\,\%$ para el cociente de varianzas es

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2,n_2-1,n_1-1}.$$

Comparación de medias para dos poblaciones no-independientes

Nota: Se puede mostrar que $F_{\alpha/2,n_2-1,n_1-1} = \frac{1}{F_{1-\alpha/2,n_1-1,n_2-1}}$, dependiendo del software que se use.

Comparación de Medias para dos poblaciones No Independientes (Análisis de datos pareados).

Supongamos que nuestra muestra consiste de n parejas seleccionadas independientemente

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

Esto significa que sólo tenemos n unidades experimentales y tomamos 2 mediciones sobre ellas, como lo describimos anteriormente, antes y después, método A, método B, etc., (existen algunas excepciones).

Comparación de medias para dos poblaciones no-independientes

Para eliminar el efecto del individuo o unidad experimental, tomamos las diferencias entre estos valores, definimos una nueva variable

$$D_i = X_i - Y_i \quad i = 1, 2, \ldots, n.$$

Es fácil ver que

$$E(D_i) = E(X_i - Y_i) = E(X_i) - E(Y_i) = \mu_1 - \mu_2 = \mu_D$$

Entonces, trabajamos con estas $D's$ como si fuesen nuestros valores muestrales directamente.

Comparación de medias para dos poblaciones no-independientes

1) Supongamos que las diferencias siguen una distribución Normal (típicamente no conocemos la varianza de las diferencias)[4] entonces aplicamos directamente los resultados vistos sobre IC para la media de una población normal con varianza desconocida

$$\text{Pivote: } T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1},$$

donde S_D es la desviación estándar de las diferencias entonces el IC del $100(1-\alpha)\,\%$ para μ_D está dado por

$$\overline{d} \pm t_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}.$$

[4] Notemos que $Var(D_i) = Var(X_i) + Var(Y_i) - 2Cov(X_i, Y_i)$ dado que los elementos dentro de cada pareja no son independientes, aunque las parejas si lo sean

Comparación de medias para dos poblaciones no-independientes

2) Si la muestra es grande, directamente aplicamos el TLC para las diferencias y construimos al igual que antes un IC aproximado

$$\overline{d} \pm z_{\alpha/2,n-1}\frac{s_D}{\sqrt{n}}.$$

Pruebas de hipótesis

Pruebas de Hipótesis

En los capítulos anteriores hemos discutido dos formas de inferencia, a saber:

- la estimación puntual; y
- la estimación por intervalos.

En ambos casos nos hallamos frente a los datos y nos interesa obtener de ellos información pertinente acerca de algún parámetro θ.

Consideremos ahora la siguiente situación: Nos hallamos frente a un proceso en el que una máquina acaba de ser ajustada para producir cilindros de metal con un diámetro de 80 milímetros. Conscientes de que existe una variación inherente a las magnitudes del diámetro, quisieramos saber si en promedio los diámetros cumplen con la medida especificada. Si μ representa ese promedio, queremos probar con mediciones reales hechas de los diámetros si en efecto $\mu = 80$ o por el contrario, $\mu \neq 80$.[5]

[5] Adaptado de Anthony J. Hayter. (1996). Probability and Statistics for Engineers and Scientists, International Thompson Publishing.

Pruebas de Hipótesis

¿Cuál es la estructura del problema ahora?

Por ciertas razones creemos que el valor del parámetro es un cierto valor específico, digamos θ_0 ; deseamos entonces verificar ese valor, es decir, determinar si es consistente con nuestros datos, o bien, estos algoritmos nos hacen pensar que es erróneo.

Buscaremos una regla construida a través de nuestras observaciones que nos guíe para decidir sobre una hipótesis acerca de algún componente del modelo probabilístico (típicamente sobre parámetros).

Existen muchas otras situaciones en donde nos interesa probar hipótesis de otra índole, a continuación mencionamos algunos ejemplos.

Pruebas de Hipótesis

Se cuenta con los registros de CO (en gr/Km) de cientos de automóviles en las estaciones de verificación[6] Anahuac, La Pastora y San Pedro, en el área Metropolitana de la Cd. de Monterrey, y nos gustaría, a través de esta información, evaluar la proporción de autos que pasan de la norma en esta ciudad.

El primer problema que se debe enfrentar aquí (además de verificar la calidad de los datos) es el de proponer un modelo distribucional, para lo cual se contemplarían algunas propuestas obtenidas a partir de los histogramas, seguidas de pruebas de falta de ajuste.

Una vez determinada apropiadamente la distribución y sus parámetros, estimar este valor (la proporción de autos fuera de la norma) no es muy complejo, pues representa a un percentil de dicha distribución.

[6] Para sorpresa de muchos, estas estaciones ya no están en funcionamiento!!!

Pruebas de Hipótesis

En este capítulo se discutirán pruebas que se desarrollan dentro de un modelo poblacional específico en consideración, por lo que su validez está ligada al modelo propuesto; además, nos restringiremos a pruebas sobre los valores de los parámetros.

No existen criterios de prueba que sean correctos para todas las situaciones prácticas; siempre deberemos considerar una serie de supuestos y tratar de verificar, en la medida de lo posible, su veracidad antes de aplicar cualquiera de las formulaciones que se discutirán en esta sección.

Pruebas de Hipótesis

¿Qué modelo probabilístico propones inicialmente para el ejemplo de los cilindros de metal?

Observa que la prueba que se construya para la hipótesis acerca de μ, se basará en la forma de la distribución que rige las observaciones, de la cual μ es su media.

Observa también que en nuestra hipótesis sobre μ se intenta recoger algún aspecto relevante del proceso (valor nominal de 80 mm.), el cual puede ser, y debe ser, representado adecuadamente dentro de un modelo de distribución (en este caso es μ , pero no siempre los valores nominales coinciden con los valores de los parámetros de los procesos o fenómenos estudiados).

Pruebas de Hipótesis

En circunstancias muy frecuentes, los valores de los parámetros poblacionales se encuentran asociados con los diferentes estados de los procesos o con el estatus de validez de una teoría, de forma tal que para ciertos valores de los parámetros ejercemos distintas acciones; por ejemplo, colocar el proceso a un nivel deseable, o bien mantenerlo en el nivel actual, argüir en favor o en contra de una teoría, etc.

En estas circunstancias estamos dispuestos a realizar acciones tomando como referencia los valores actuales de los parámetros.

Discutiremos a detalle los ingredientes y el proceso en sí de cómo probamos hipótesis del tipo que nos atañen aquí y estableceremos estos procedimientos en forma explícita para algunos casos específicos.

Pruebas de Hipótesis

Estos ingredientes son:

- **Planteamiento de Hipótesis**,
- Estadístico de Prueba,
- Región de Rechazo y Nivel de significancia,
- p-valor,
- Potencia de la Prueba,
- Conclusiones.

Pruebas de Hipótesis

Planteamiento de Hipótesis.

En el ejemplo anterior hemos visto que se tienen dos afirmaciones acerca del parámetro las cuales son puestas en competencia para ver cuál de ellas es más favorecida por la evidencia que representan las observaciones.

Definición

Una hipótesis es una afirmación acerca de un parámetro poblacional.

- *Llamamos Hipótesis Nula, H_0, a la afirmación de los valores del parámetro que nos interesa probar.*
- *Denotamos con H_A a la Hipótesis Alternativa, que representa el contrario de la nula.*
- *Ambas hipótesis son complementarias*.

Pruebas de Hipótesis

En la práctica, frecuentemente el investigador realiza una prueba de hipótesis con la intención deliberada de rechazar H_0 sobre la base de una evidencia suficientemente "fuerte" de que es falsa; en otras palabras, supone que es verdadera en la medida que la muestra no le indique lo contrario. Su interés primario es H_A, la hipótesis que se espera esté sustentada por los datos (en algunas ocasiones la Hipótesis Alternativa es llamada Hipótesis de Investigación).

La razón de esto es que tomar la decisión de si se rechaza o acepta una hipótesis conlleva un costo: El investigador que realiza una prueba de hipótesis se verá ante la situación de mantener o modificar su teoría; el farmacólogo que prueba la efectividad de un medicamento desea estar seguro de que el medicamento es realmente efectivo para recomendar su producción y uso; en la industria, alterar o modificar algún componente del proceso implica tiempos, recursos, etc.

Pruebas de Hipótesis

En estos casos, la situación que prevalece hasta el momento de hacer la prueba de hipótesis es la situación que se quiere modificar en un momento dado, siempre y cuando haya suficientes razones para hacerlo. La Hipótesis Nula se identifica con esta situación, mientras que la Hipótesis Alternativa representa la situación de cambio.

Una prueba de hipótesis recogerá la información de la muestra, nos dirá el grado de discrepancia entre ella y H_0, y nos dirá si éste es "grande" o no. De esta forma aceptará H_A en lugar de H_0 sólo si tiene un fuerte respaldo de la muestra.

En este contexto, las conclusiones que seamos capaces de obtener de la muestra no nos indican de manera definitiva si las hipótesis son verdaderas o falsas, simplemente si unas son más plausibles que otras.

Pruebas de Hipótesis

En general no deberíamos tomar la información y en base a ella plantear las preguntas a investigar; recuerda que todo esto se deriva de tener un problema real del cuál se concretizan preguntas en términos del modelo propuesto para su solución. Cuando se actúa en base a la información y se aplican las mismas metodologías de análisis, sesgas los resultados, a menos que sólo se usen con carácter informativo para indicar las líneas de investigación que se habrán de desarrollar en nuevos "experimentos".

Para realizar una prueba de hipótesis es necesario especificar con toda claridad y cuidado desde el inicio ambas hipótesis. En caso de que la muestra que tomamos de la población nos de evidencia de que la Hipótesis Nula es poco plausible, la rechazaremos y no rechazaremos la Hipótesis Alternativa.

Pruebas de Hipótesis

Según la situación y nuestro conocimiento acerca del problema podemos plantear distintas modalidades de hipótesis:

Hipótesis simples, para un valor simple del parámetro: $H_0 : \theta = \theta_0$, o equivalentemente, $H_0 : \theta - \theta_0 = 0$; y en consecuencia la hipótesis alternativa podría ser simple, $H_A : \theta = \theta_1$, o compuesta, $H_A : \theta \neq \theta_0$.

Alternativa simple: Cuando tenemos tal conocimiento del proceso estudiado que podemos probar un valor específico contra otro perfectamente delimitado.

Alternativa compuesta: Cuando no tenemos información en qué sentido nuestra Hipótesis Nula pueda ser falsa. Es decir, rechazaremos H_0, ya sea que la muestra nos indique que proviene de una población con parámetro $\theta > \theta_0$, o bien, con parámetro $\theta < \theta_0$; en cualquier caso rechazaremos H_0.

Pruebas de Hipótesis

Hipótesis compuestas. En esta discusión nos restringiremos a valores del parámetro dentro de un intervalo. Las posibles Hipótesis Nulas pueden ser

$$a)\ H_0 : \theta \leq \theta_0,$$

o

$$b)\ H_0 : \theta_0 \leq \theta,$$

las cuales conducen a Hipótesis Alternativas de la forma:

$$H_A : \theta > \theta_0, \text{ en el caso a),}$$

y

$$H_A : \theta < \theta_0, \text{ en el caso b).}$$

Si queremos probar que el parámetro θ se haya acotado superior o inferiormente. Desde el punto de vista de la evidencia aportada por la muestra, significa que rechazaremos H_0 para valores de θ mayores a θ_0, pero no para valores menores a θ_0, en el caso a); el caso b) se interpreta similarmente.

Pruebas de Hipótesis

¿Qué tipo de prueba hacer en cada situación?

La respuesta a esta pregunta está íntimamente ligada a nuestra información previa acerca del fenómeno estudiado. Al tiempo en que planteemos las hipótesis, debemos ser capaces de introducir en su formulación lo que ya conozcamos del proceso y en particular del parámetro.

A continuación estudiaremos algunos ejemplos.

Pruebas de Hipótesis

Ejemplos.

Ejemplo 1. Una tela es inapropiada para teñirla si su absorción de agua es menor al 55 %[7]. Supongamos que tenemos una muestra de telas de algodón y se desea saber si este tejido es apropiado para teñirlo. ¿Qué Hipótesis Nula deberá plantear el comprador del producto? Las hipótesis pueden ser las siguientes:

$$H^* : \theta \leq 55\,\% \text{ vs } H^+ : \theta > 55\,\%$$

Observa que el comprador desea probar para protegerse que realmente $\theta > 55\,\%$, ya que si esto es así puede recomendar las telas de algodón para teñirse. Por lo tanto,

$$H^* = H_0 : \theta \leq 55\,\% \text{ vs } H^+ = H_A : \theta > 55\,\%.$$

[7] Adaptado de Anthony J. Hayter. (1996). Probability and Statistics for Engineers and Scientists, International Thompson Publishing.

Pruebas de Hipótesis

Ejemplo 2. Un fabricante de automóviles asegura que sus vehículos logran un promedio de por lo menos 35 millas por galón de combustible en carretera[8]. Un grupo representante de consumidores desea poner a prueba lo que el fabricante dice, por lo que toma una selección aleatoria de los automóviles que el fabricante produce y los conduce en condiciones de carretera, tratando de mantener estas pruebas en condiciones homogéneas y midiendo la eficiencia del combustible.

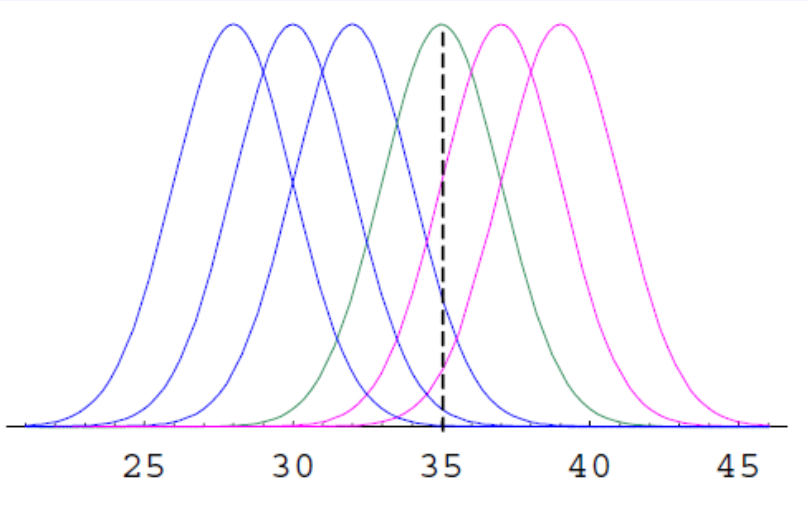
Evidentemente, el grupo representante de consumidores tiene la intención de demostrar que lo dicho por el fabricante no es correcto:

$$H_0 : \mu \geq 35 \text{ vs } H_A : \mu < 35$$

[8] Adaptado de Anthony J. Hayter. Ibid.

Pruebas de Hipótesis

Tanto en este problema como en el anterior, estamos considerando las distintas poblaciones correspondientes a las medias que se especifican en las hipótesis, lo cual no significa sean varias poblaciones las que se están muestreando; por el contrario, la población de la cual proviene la muestra es solamente una, pero al realizar la prueba se toman en cuenta los conjuntos de poblaciones para decidir a cuál es más plausible que pertenezca la muestra.



Pruebas de Hipótesis

Es conveniente notar que las hipótesis de mayor interés típicamente son compuestas, esto es, bajo H_0 se presenta la posibilidad de que, en el caso de considerar el valor medio, μ sea μ_0 o cualquier otro número mayor que o menor que μ_0, dependiendo de la situación bajo estudio.

Sin embargo, al poner en marcha el mecanismo de prueba, sólo necesitamos comparar las poblaciones bajo H_A con la población de la Hipótesis Nula cuyo parámetro tiene valor en la igualdad (el valor en la frontera del intervalo).

Si, como en este ejemplo, tenemos el juego de hipótesis:

$$H_0 : \mu \geq 35 \text{ vs } H_A : \mu < 35$$

y la información muestral no nos indica que la media sea menor a la de la población hipotética con $\mu = 35$ (mostrada en verde en la figura), con mucho menos razón esa información podría "derrocar" poblaciones con medias mayores a 35.

Pruebas de Hipótesis

Así las cosas, en términos de trabajo, siempre se considera la igualdad en la Hipótesis Nula, pero debe tenerse en cuenta el intervalo que realmente está siendo considerado para plantear de manera adecuada la Hipótesis Alternativa.

Ejemplo 3. Supongamos que se utilizan dos máquinas para producir laminillas de vidrio, A y B. Se pueden formular las siguientes hipótesis respecto de las medias de los grosores de las laminillas producidas por ambos procesos, de acuerdo a la información disponible:

a) Los procesos siguen la misma tendencia, $H_0 : \mu_A = \mu_B$, donde las μ's representan esa tendencia.

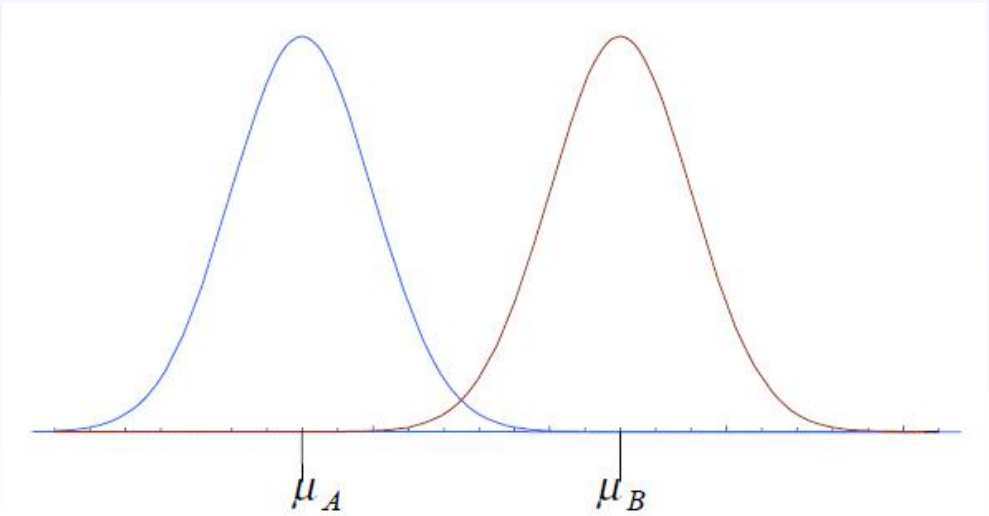
b) Las tendencias de los procesos son distintas entre sí por alguna cantidad específica: $H_0 : \mu_A - \mu_B = \mu_{A0} - \mu_{B0}$.

Pruebas de Hipótesis

c) Las tendencias de los procesos son distintas como máximo por una cantidad específica: $H_0 : \mu_A - \mu_B \leq \mu_{A0} - \mu_{B0}$ (la contraparte $H_0 : \mu_A - \mu_B \geq \mu_{A0} - \mu_{B0}$ también es formulable).

¿Qué hipótesis alternativas corresponden en cada caso?

En este problema sí tenemos más de una población: una con media μ_A y otra con μ_B:



Pruebas de Hipótesis

Cuando discutamos los diferentes tipos de errores y la forma de medirlos, en cada una de nuestras decisiones, tendremos mejor claridad sobre la forma más adecuada para la selección de nuestro juego de hipótesis.

Pruebas de Hipótesis

Estos ingredientes son:

- Planteamiento de Hipótesis,
- **Estadístico de Prueba**,
- Región de Rechazo y Nivel de significancia,
- *p*-valor,
- Potencia de la Prueba,
- Conclusiones.

Pruebas de Hipótesis

Estadístico de Prueba.

Es tiempo de plantearnos: ¿Cómo incluiremos la información de nuestra m.a. en la prueba?

Considerando la idea de lo que hace una Prueba de Hipótesis, podemos adelantar una respuesta diciendo:

"Es posible separar los valores muestrales para los que el valor θ_0 es plausible de aquéllos para los que no lo es". Es decir, en nuestro espacio muestral existen valores que nos hacen no rechazar H_0 y otros (el complemento) que nos hacen rechazarla.

Pruebas de Hipótesis

Al igual que en los dos capítulos anteriores, en los que de alguna forma resumimos la información de la muestra en un estadístico con características deseables, dar respuesta a este problema resulta difícil considerando simultáneamente los valores muestrales, por lo que utilizamos un estadístico de prueba W.

Por ejemplo, si nos interesa probar que μ es 4, de lo que hemos aprendido en estimación, lo natural sería:

a) obtener una muestra,

b) estimar μ a través de $\overline{x}$,

c) comparar qué "tan diferente" se ve el valor muestral del valor hipotetizado, por ejemplo midiendo la discrepancia de la siguiente manera: $(\overline{x} - 4)$.

Pruebas de Hipótesis

Medir la discrepancia entre el valor de la hipótesis y el valor muestral es una de las labores de los Estadísticos de Prueba. Así, si esa distancia es "grande" rechazamos H_0, esto es, no estamos dispuestos a creer que la población de donde obtuvimos la muestra realmente tiene un valor medio de 4.

Este proceso requiere de una forma universalmente aceptada, que nos permita establecer qué es grande y qué es pequeño con el menor subjetivismo posible. Una solución al problema se logra calculando la probabilidad de que, si mi muestra proviniera de una población con parámetro θ_0, encuentre una desviación tan grande como la que he calculado.

Esto es, se cambia el concepto de medir a través de distancias o distancias relativas por el de la probabilidad de haber observado dicha distancia.

Pruebas de Hipótesis

Aunque en principio esto puede parecer difícil o confuso, en realidad es muy simple, si contamos con el respaldo de: información clara sobre nuestro problema, la intuición natural de como medir, la construcción de modelos estadísticos y estimación de parámetros.

Esto tiene su precio (como todo en la vida), pues para poder asignar probabilidades a esas discrepancias, requerimos del conocimiento del modelo probabilístico que siguen esas distancias.

Pruebas de Hipótesis

Definición

La Prueba de Hipótesis se establece mediante :

- *Un estadístico $W(X_1, \ldots, X_n)$, que llamamos Estadístico de Prueba, el cual es una función de la muestra que recoge de ella la información acerca del parámetro, al mismo tiempo que*
- *su distanciamiento respecto del valor paramétrico propuesto en la Hipótesis Nula, y*
- *su distribución es conocida bajo la Hipótesis Nula y no depende de ningún otro parámetro desconocido.*

Pruebas de Hipótesis

Resumiendo las ideas anteriores:

Una regla que nos diga cuándo rechazar o cuándo no rechazar H0 basándose en la información muestral será el procedimiento de prueba de hipótesis. Más precisamente:

Definición

Un procedimiento de Prueba de Hipótesis es una regla que establece:

a) *Qué valores muestrales conllevan a la decisión de no rechazar H_0*

b) *Para qué valores de la muestra se toma la decisión de rechazar H_0 y no rechazar H_A.*

Pruebas de Hipótesis

La importancia de un procedimiento de esta naturaleza tiene que ver con el hecho de cómo obtener conclusiones de la muestra con un nivel de error controlable; lo que buscamos es decir cosas acerca de un parámetro que no conocemos sin equivocarnos "tanto". El procedimiento nos dirá el grado de error en el que incurro si decido aceptar la hipótesis como correcta o como incorrecta.

Pruebas de Hipótesis

Estos ingredientes son:

- Planteamiento de Hipótesis,
- Estadístico de Prueba,
- **Región de Rechazo y Nivel de significancia**,
- p-valor,
- Potencia de la Prueba,
- Conclusiones.

Pruebas de Hipótesis

Región de Rechazo y Nivel de Significancia.

Definición

El subconjunto del espacio muestral en el cual H_0 es rechazada se llama región de rechazo o región crítica. Su complemento se llama región de aceptación o de no rechazo.

Estas regiones se definen en términos del estadístico W y no en términos de la muestra directamente.

Denotaremos por conveniencia $\mathfrak{R}$ a esa región de rechazo.

En un problema particular, el cálculo del estadístico de prueba nos indica si rechazamos o aceptamos H_0; es decir, rechazamos si $w \in \mathfrak{R}$.

Pruebas de Hipótesis

Más adelante veremos cómo elegir la región de rechazo $\mathfrak{R}$ de tal forma que nuestra prueba tenga buenas propiedades. Por lo pronto, parece claro que si la muestra arroja un valor de un estimador de θ que esté muy alejado de θ_0, el valor hipótetico del parámetro, es poco plausible que la muestra provenga de una población regida por θ_0.

Veamos un ejemplo para tratar de poner en claro todo lo discutido hasta aquí.

Pruebas de Hipótesis

Ejemplo: Retomemos el ejemplo del diámetro de los cilindros de metal. Nuestras hipótesis serían:

$$H_0 : \mu = 80 \text{ vs } H_1 : \mu \neq 80.$$

Si en una muestra de tamaño 10, digamos, nuestro valor muestral de $\overline{X}$ es de 80.3 mm., **¿qué decisión tomaríamos?**

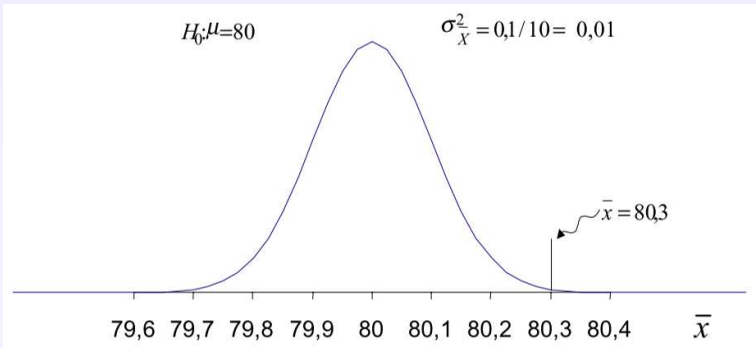
Observa que $\overline{X}$ es una variable aleatoria, y que por lo tanto posee variabilidad, misma que debería ser considerada al tiempo de nuestra decisión.

Si $X_1 \ldots, X_n$ forman una m. a. de una Normal(μ,σ^2), σ^2 conocida (μ no lo es), nuestra decisión se verá influenciada por los valores de σ^2 . Primero, bajo la hipótesis nula estaríamos diciendo que nuestra población es Normal(80,σ^2).

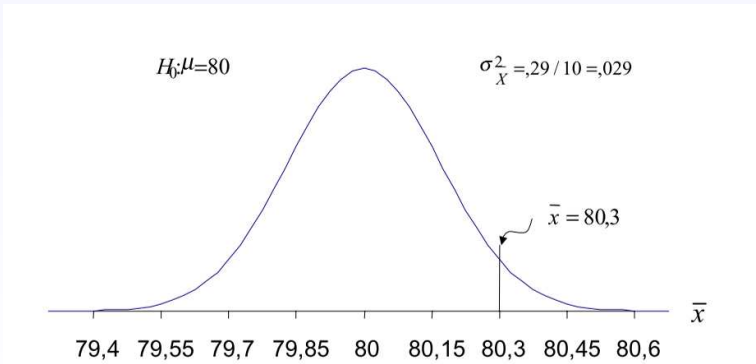
Pruebas de Hipótesis

Ejemplo:

Distribución Muestral de $\overline{X}$.



Si $\sigma^2 = 0.1$, nos veríamos inclinados a rechazar $H_0 : \mu = 80$: recuerda que la varianza de $\overline{X}$ es menor que la de cada variable, de hecho $\sigma_{\overline{X}} = \sigma/\sqrt{n} = \sqrt{0.1/10} = 0.1$, y además $\overline{X}$ hereda la forma de la distribución de la población normal).



Pero si $\sigma^2 = 0.29$, quizá decidiríamos no rechazar H_0, ¿no crees?

Pruebas de Hipótesis

¿Qué relación puedes establecer con los Intervalos de Confianza?

Observa que el IC del 95 % para μ cuando $\sigma^2 = 0.1$ es

$$\overline{x} \pm 1.96(.1) = 80.3 \pm 0.196 = [80.104, 80.496],$$

mientras que con $\sigma^2 = 0.29$ es

$$[79.966, 80.633];$$

el primer intervalo no contiene al parámetro, pero el segundo sí.

Pruebas de Hipótesis

En términos de hipótesis, **¿qué tan plausible parece ser en el primer caso, que la muestra provenga de una población con parámetro $\mu = 80$?**

Si $\overline{x} = 80.3$ mm., está muy alejada de esto, para ello basta considerar su varianza y su distribución: el desvío $\overline{x} - \mu = 0.3$ parece grande, ¿no? Notarás que la probabilidad de que aparezca el valor de 80.3 ú otro mayor es muy baja.

Sin embargo, en el **segundo caso** un distanciamiento de por lo menos 0.3 entre la media poblacional y la muestral es todavía esperable. Esto es, bajo una $Normal(80, 0.029)$ valores de 80.3 o mayores no se ven tan imposibles.

Entonces en el **segundo caso** el valor de $\mu = 80$, se pensaría como correcto, o por lo menos, como que no tenemos suficiente evidencia para pensar que el valor de μ propuesto en H_0, no sea el correcto.

Pruebas de Hipótesis

El desvío $\overline{x} - \mu$ ha resultado un buen candidato para ser un estadístico de prueba W si tratamos de probar una hipótesis acerca de la media; veremos cómo mejorarlo.

Ejercicio: Calcula las probabilidades mencionadas arriba.

Pruebas de Hipótesis

Cuando hacemos una hipótesis acerca de un parámetro, existen dos posibilidades: acertar o equivocarse (¿en qué?). Claro, la hipótesis, o es verdadera o es falsa, puede alguien argüir; pero nosotros no lo **sabemos**!!!

Por supuesto, si conocieramos el valor del parámetro, podríamos decidir cuál hipótesis es verdadera y cuál no. En esta situación es claro qué hipótesis aceptar y cuál rechazar sin cometer ningún error; nuestra decisión no tendrá asociada ningún grado de error.

Cuando estamos ante los datos, no tenemos de ningún modo la certidumbre de la situación anterior. De cara a esta incertidumbre sobre el parámetro y teniendo frente a nosotros la decisión de aceptar o rechazar la hipótesis, creamos el espacio para cometer **errores** en la toma de decisiones.

Pruebas de Hipótesis

Puedes leer la frase anterior en el sentido de que la decisión de aceptar la hipótesis como correcta está sujeta a error y que un procedimiento de prueba nos debe permitir el no equivocarnos en alto grado al tomar nuestra desición, o al menos, prevenirnos de ello ante esta situación.

No se interpreta como que el resultado final de una prueba de hipótesis es declarar la verdad o la falsedad de las Hipótesis Nula y Alternativa.

Pruebas de Hipótesis

Al equivocarnos, cometemos un error, que en el contexto de las pruebas de hipótesis estadísticas puede ser el Error Tipo I o el Error Tipo II:

	Valor de la Hipótesis	
Decisión	H_0 **cierta**	H_0 **falsa**
No rechazar H_0	OK!	Error Tipo II
Rechazar H_0	Error Tipo I	OK!

Tomando en cuenta las consideraciones anteriores, debemos ahora determinar claramente una región de rechazo con la cual decidir sobre H_0.

Como has podido observar, nuestra actitud respecto de la Hipótesis Nula ha sido más bien **conservadora**; rechazaremos sobre la base de una evidencia suficiente de que sea incorrecta. Esta actitud hace necesario protegerse del Error Tipo I.

Pruebas de Hipótesis

En otros términos, deseamos que la probabilidad de rechazar la Hipótesis Nula cuando en realidad es cierta sea baja:

$$\text{P(Rechazar } H_0 \mid H_0 \text{ es cierta)}=\alpha\text{, donde } \alpha \text{ es pequeña y } 0 \leq \alpha \leq 1$$

Queremos minimizar el número de veces que se tome la decisión de hacer un cambio, cuando en realidad no se necesita. Este error está asociado con lo que se conoce como una **falsa alarma**.

Definición

Una prueba cuya probabilidad de cometer el Error Tipo I es α se dice que es una prueba de nivel α (Nivel de significancia).

Pruebas de Hipótesis

Los valores de α elegidos con mayor frecuencia son: 0.1, 0.05, 0.01.

La decisión de cuál valor de α elegir en una prueba particular que realicemos tiene que ver forzosamente con la situación que estemos tratando.

Es la circunstancia en la que se presenta el problema con el que estamos tratando la que nos motiva a probar una hipótesis y es de acuerdo con ella que debemos fijar los criterios para evaluar la información recopilada para tal efecto. **Básicamente depende de qué tanto te cueste equivocarte en esta dirección**.

Pruebas de Hipótesis

En resumen, **¿cómo relacionamos estas consideraciones con la manera en que la muestra será utilizada en la construcción de la prueba?**

Como se ha establecido, normalmente usaremos un **estadístico de prueba** W.

Éste debe reunir ciertas condiciones para que sea útil en la regla de decisión; además de las ya mencionadas en su definición es conveniente remarcar que, como se trasluce en las ideas anteriores, muchas veces el estadístico W es una modificación del estimador de θ, con el objetivo medir el alejamiento de los valores muestrales respecto de H_0.

Pruebas de Hipótesis

La cuestión clave aquí es que la distribución del estadístico W dado el valor del parámetro en la Hipótesis Nula sea conocida, a fin de poder determinar las probabilidades de cometer el Error Tipo I para los distintos valores de W.

Los valores del estadístico W que nos conllevarán a rechazar o aceptar H_0, es decir, la región $\mathfrak{R}$, serán elegidos de tal forma que si H_0 es cierta, la probabilidad de que ocurran sea α.

O sea, fijaremos los valores de W que sean **poco probables bajo la Hipótesis Nula** y rechazaremos si es que éstos se presentan en una muestra particular que tomemos; al hacerlo así, disminuimos la probabilidad de equivocarnos en lo que respecta a rechazar H_0 cuando en realidad la población bajo estudio tiene el comportamiento descrito por ella.

Pruebas de Hipótesis

Discutiremos la implementación de las ideas anteriores en diferentes pruebas acompañándolas de un ejemplo.

Pruebas para Una Muestra.

Ejemplo: Para obtener información sobre la resistencia a la corrosión de un cierto tipo de conductor de acero, 35 especímenes son enterrados en suelo por dos años[9]. La penetración máxima para cada ejemplar es medida en milipulgadas (mils).

La muestra arroja un valor de $\overline{x} = 52.7$. Los conductores fueron fabricados bajo la especificación de que el promedio de penetración sea a lo mucho de 50 milipulgadas con $\sigma = 4.8$.

[9]Tomado de Jay L. Devore. (1991). Probability and Statistics for Engineering and the Sciences, 3era Edición, Duxbury Press.

Pruebas de Hipótesis

Un estadístico de prueba en este caso muy bien puede ser $\overline{X} - \mu_0$; contiene al estimador de μ y además ofrece una medida de discrepancia.

La cuestión es qué tan grande puede ser la discrepancia bajo H_0; si la Hipótesis Nula es cierta, ¿qué valores de $\overline{X} - \mu_0$ pueden considerarse como poco probables?

Supondremos que la muestra proviene de una población Normal(μ,σ^2). Como en este caso lo que nos interesa es saber si los conductores cumplen con la especificación, tenemos:

$$H_0 : \mu \leq \mu_0 = 50 \text{ vs } H_1 : \mu > \mu_0 = 50.$$

Nos interesa fijar un punto k a partir del cual, si la Hipótesis Nula es cierta, $P(\overline{X} - \mu_0 \geq k) = \alpha$, para que en caso de rechazar, dado que H_0 sea en realidad cierta, nuestra probabilidad de Error Tipo I sea baja.

Pruebas de Hipótesis

Es muy importante tomar en cuenta lo que afirma la Hipótesis Alternativa, ya que en función de ella es que la prueba tomará la dirección de los desvíos para rechazar.

Para este problema nos interesan los desvíos positivos de la forma $\overline{X} - \mu_0 \geq k$, ya que buscamos valores más grandes que μ_0 para rechazar.

Ahora,

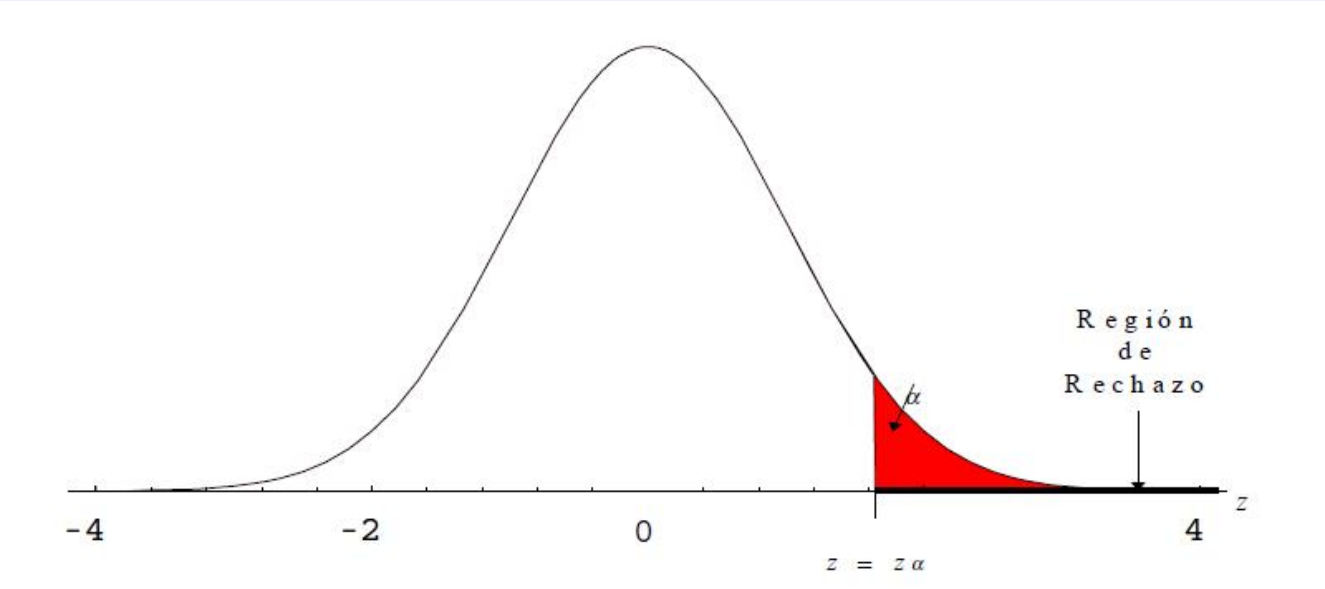
$$P(\overline{X} - \mu_0 \geq k) = P\left(\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{k}{\frac{\sigma}{\sqrt{n}}}\right) = P(Z \geq k^*)$$

donde $Z \sim N(0, 1)$.

Podemos escoger $k^* = z_\alpha$, de tal forma que $P(Z \geq z_\alpha) = \alpha = 0.05$; es decir, fijamos nuestro nivel de significancia en 0.05.

Pruebas de Hipótesis

Toma nota que también hemos fijado nuestro estadístico de prueba definitivo (Z) y la región de rechazo $\mathfrak{R}$, es decir, los valores de Z mayores a z_α.



De esta consideración anterior podemos obtener nuestra regla de prueba de hipótesis para este caso:

$$\text{Rechazar } H_0 \text{ si } z \geq z_\alpha \text{ donde } z_\alpha = 1.645.$$

Pruebas de Hipótesis

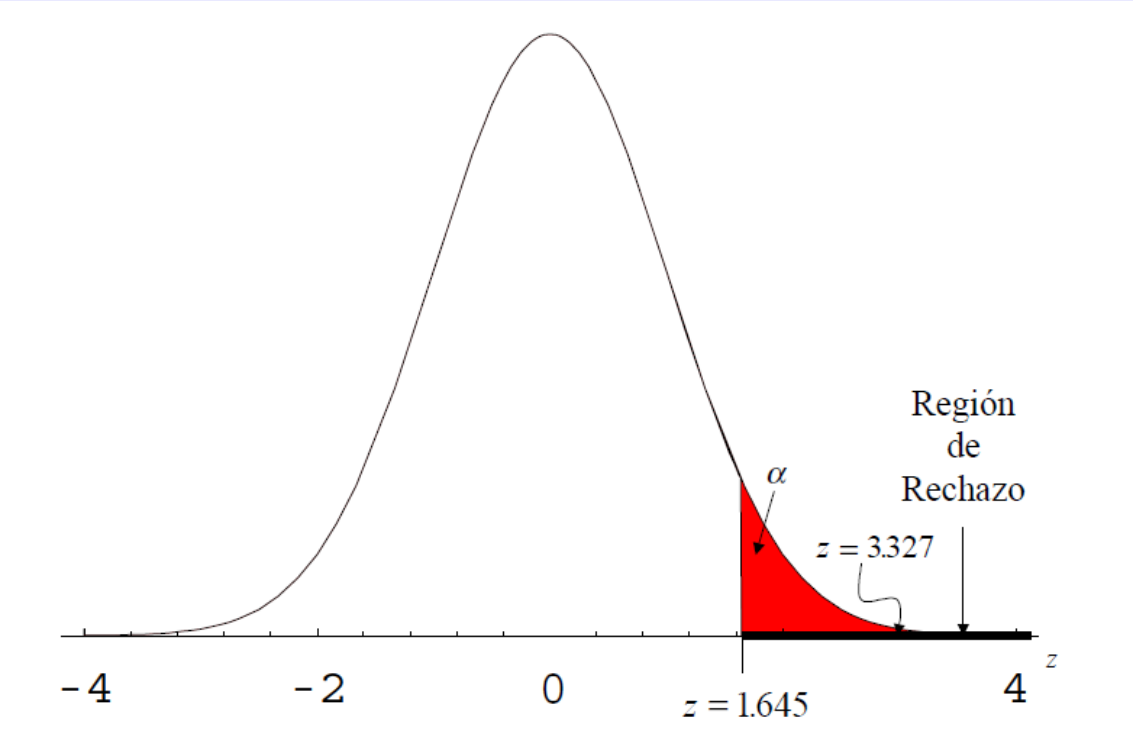
Tenemos ahora una procedimiento de prueba de hipótesis con un nivel de significancia de 0.05 para la media de una población Normal(μ,σ^2), σ^2 conocida. Llamamos nivel de significancia al hecho de declarar como significativos los desvíos $\overline{X} - \mu_o$ el 5 % de las veces si H_0 es cierta.

En este ejemplo

$$\frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{52.7 - 50}{\frac{4.8}{\sqrt{35}}} = 3.327 \geq 1.645 = z_{0.05},$$

por lo que rechazamos!! Nuestro valor calculado del estadístico de prueba cae en la región de rechazo, como lo muestra la siguiente figura.

Pruebas de Hipótesis



De todo esto podemos concluir que, tomando como base la información de la muestra, se rechaza H_0 ($\alpha = 0.05$), esto es, la muestra indica que el promedio de penetración de la corrosión en los especímenes de donde se tomó la muestra es mayor al límite superior especificado previamente.

Pruebas de Hipótesis

Nota: En pruebas para Hipótesis Nulas de la forma

$$H_0 : \theta \leq \theta_0$$

ó

$$H_0 : \theta \geq \theta_0,$$

en principio se debería fijar una región de rechazo para cada uno de los valores menores o iguales (mayores o iguales) a θ_0; sin embargo, con sólo hacerlo para el límite del intervalo es suficiente, ya que si se rechaza en base a esta región lo estaremos haciendo para todos los demás valores de θ_0 en el intervalo.

Pruebas de Hipótesis

En la práctica, muy frecuentemente no se conoce σ^2, por lo que es necesario estimarla. De esta forma, el estadístico de prueba para las hipótesis anteriores se contruye sustituyendo σ^2 por s^2:

$$\frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t \text{ de Student con } n-1 \text{ grados de libertad.}$$

Recordemos que esta distribución es factible dado que la muestra proviene de una población Normal. Así, es posible fijar la regla de rechazo como

$$\text{Rechazar } H_0 \text{ si } \frac{\sqrt{n}(\overline{x} - \mu_0)}{s} > t_{\alpha, n-1},$$

para un α específico.

Pruebas de Hipótesis

Al rechazar debemos estar conscientes que lo hacemos para un nivel de significancia α específico. Dejaremos la discusión pendiente para el segundo de nuestros errores y veremos primero una forma alternativa de tomar la decisión, sin usar α.

Esta metodología es la que está implementada en la mayoría de los softwares estadísticos y suele dar una información diferente a la discutida con el error tipo I.

Claro está que tiene sus puntos débiles, los cuales mencionaremos en su momento.

Pruebas de Hipótesis

Estos ingredientes son:

- Planteamiento de Hipótesis,
- Estadístico de Prueba,
- Región de Rechazo y Nivel de significancia,
- *p*-**valor**,
- Potencia de la Prueba,
- Conclusiones.

Pruebas de Hipótesis

El p-valor.

Imagina lo que alguien pensaría si le decimos que rechazamos la Hipótesis Nula en una prueba unidireccional con un nivel de significancia de 0.05 (como en el ejemplo anterior).

Seguramente ese alguien se preguntará qué tan grande fue el estadístico de prueba calculado o qué tan lejos estaba del punto crítico z_α.

Una manera bastante útil de presentar la información de los resultados de una prueba de hipótesis es:

- Señalar el valor calculado del estadísitico y,
- especificar su p-valor.

Pero, **¿qué es este p-valor?**

Pruebas de Hipótesis

Informalmente podríamos decir que el p-valor mide el grado de evidencia en contra de la hipótesis nula que refleja la muestra obtenida.

¿Cómo lo mide?

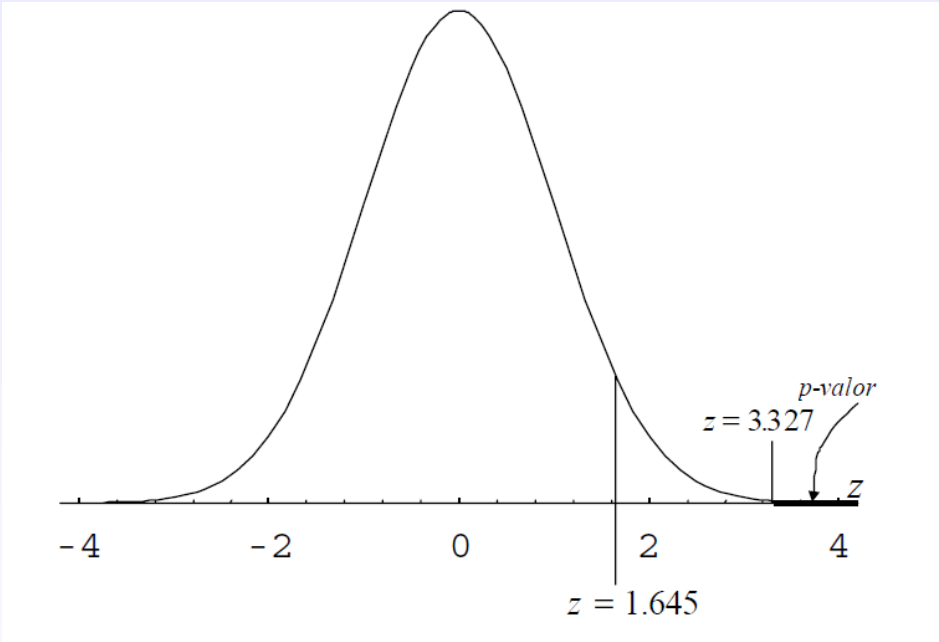
Eso lo discutiremos primero a través de un ejemplo y posteriormente lo definiremos en términos operativos.

Ejemplo: En el ejemplo que acabamos de revisar sobre el promedio de penetración de corrosión, podemos preguntarnos cuál es la probabilidad que ocurran valores mayores a $z = 3.327$, el valor del estadístico que hemos observado en la muestra.

El p-valor es precisamente esa probabilidad que se obtiene hasta después de que calculamos el valor del estadístico en base a los datos que dispongamos.

Pruebas de Hipótesis

Gráficamente, el *p*-valor está dado como lo muestra la siguiente Figura.



Observa cómo esa probabilidad está indicada en la gráfica por la pequeña sombra sobre los valores a la derecha de $z = 3.327$. De hecho,

$$P(Z \geq z_{observado}) = P(Z \geq 3.327) \approx 0.0005,$$

la cual es una probabilidad muy baja en este caso.

Pruebas de Hipótesis

Recuerda que el valor de Z está siendo calculado suponiendo que $\mu = 50$, es decir, el valor del parámetro bajo la Hipótesis Nula, por lo que el p-valor es una probabilidad que se obtiene sobre el supuesto de que el valor del parámetro en la Nula es el correcto.

De esta forma, es posible entender el p-valor como la probabilidad de que encontremos discrepancias respecto de nuestro valor nulo tan grandes o mayores como las observadas en nuestra muestra, si la Hipótesis Nula es cierta.

Un p-valor pequeño es un indicador de que nuestra Hipótesis Nula es incorrecta, en cuyo caso debemos rechazarla.

Pruebas de Hipótesis

Así, el p-valor es como si dijeramos:

"Si mi Hipótesis Nula fuese cierta, la probabilidad de que existan desviaciones tan grandes como las que observó es muy baja y no esperaría que se presentaran al tomar una muestra; sin embargo, la muestra que tengo enfrente me presenta tales desviaciones, por lo tanto, la Hipótesis Nula debe ser falsa, o 'La Hipótesis Nula no es plausible'."

Esta es la forma en que el p-valor mide la evidencia en contra de la Hipótesis Nula: de acuerdo a la improbabilidad de las discrepancias observadas.

Pruebas de Hipótesis

Operativamente, es la probabilidad de que el estadístico W tome valores a partir del valor observado w (calculado a partir de la muestra) en la dirección de la Hipótesis Alternativa:

p-valor=P(desviaciones tan grandes o mayores como las observadas en la muestra $\mid H_0$)

Por lo tanto, dado que el p-valor en nuestro ejemplo es muy bajo rechazamos H_0, concluyendo que el promedio de corrosión en los especímenes es mayor a 50.

¿Qué hemos hecho? ¡¡Hemos tomado una decisión de una manera alternativa a la metodología basada en el nivel de significancia!!

De hecho, ésta es la manera natural en que se plantea el rechazo o no de la Hipótesis Nula. Independientemente de fijar una probabilidad α arbitraria, nosotros podemos decidir qué tan grande o pequeña es la discrepancia a partir de los p-valores. No es necesario contar con un valor de α para tomar decisiones.

Pruebas de Hipótesis

Definamos ahora un poco más formalmente este *p*-valor y así podemos discutir mejor la relación entre el *p*-valor y el nivel de significancia.

Definición

El p-valor es el nivel de significancia más pequeño con el cual es posible rechazar H_0 dada la muestra. Indica la probabilidad más pequeña con la que declararíamos como significativa la discrepancia de la muestra respecto del valor del parámetro en H_0 .

A pesar de que en la definición se expresa el *p*-valor en términos de niveles de significancia, existen diferencias importantes entre el α y el *p*-valor.

El nivel de significancia α de nuestra prueba es **fijado de antemano y de una vez por todas** al hacer nuestra prueba, mientras que el *p*-valor es dependiente de los datos, en el sentido de que es una probabilidad que se calcula **posteriormente a lo que observamos**.

Pruebas de Hipótesis

Al discutir la probabilidad de Error Tipo II y la potencia de la prueba será evidente que el α de la prueba tiene la ventaja de permitir desarrollar ideas al respecto, cosa que no es posible con el *p*-valor

En la gráfica anterior es posible visualizar que si $z_{observado} = z_\alpha$, entonces el *p*-valor es igual al nivel de significancia que hayamos fijado de antemano para una prueba en particular.

Pruebas de Hipótesis

Sin embargo, la decisión sobre qué p-valor debe ser considerado como suficientemente pequeño para declarar significativo al valor que observemos del estadístico de prueba es independiente de que fijemos de antemano un α. No necesitamos comparar nuestro p-valor con un nivel de significancia que hubieramos fijado inicialmente.

Una vez más, si consideramos la situación en donde estemos realizando la prueba, podremos ponderar qué costo tendría declarar como significativos los desvíos a un cierto nivel.

Aún y con todo lo que implica el comentario anterior, en general, como regla, no se toma la decisión de rechazar la Hipótesis Nula en base a p-valores que sean más grandes que 0.20.

Discutamos ahora otros ejemplos de pruebas de hipótesis, incluyendo lo ahora sabemos del p-valor.

Pruebas de Hipótesis

Ejemplos:

Ejemplo 1. La suposición fuerte en el ejemplo del promedio de penetración de corrosión es que tenemos una m.a. de tamaño n de una $N(\mu, \sigma^2)$.

La pregunta es: ¿Qué pasa si la población de donde estoy muestreando no es normal?

En esta situación podemos apelar al TLC, ya que tenemos un tamaño de muestra suficientemente grande ($n = 35$). El estadístico de prueba es entonces

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \dot{\sim} N(0,1), \text{ bajo } H_0.$$

Pruebas de Hipótesis

Lo que cambia respecto de cuando asumimos normalidad es que ahora trabajamos con la distribución aproximada del estadístico Z.

Todo lo demás queda **igual**: el cálculo del estadístico, la región de rechazo, el p-valor, la decisión que tomemos, etc. En este sentido, **si queremos concluir sobre el promedio de corrosión lo haremos de la misma forma que cuando asumimos normalidad**.

Toma en cuenta que para que esta aproximación entre en escena ha sido necesario una tamaño de **muestra grande**; hay que cuidarse en otras situaciones de que haya las condiciones necesarias para aplicar el TLC. Existen otras soluciones alternativas conocidas como Pruebas No-parámetricas.

Pruebas de Hipótesis

Ejemplo 2. Una máquina llenadora de envases de bebida está bajo control si su promedio de llenado es de 12.2 onzas por envase, con una desviación estándar de .05oz como máximo[10].

Un supervisor desea saber, como parte de un chequeo periódico, si realmente la máquina se encuentra bajo control, para lo cual toma una muestra aleatoria sobre 26 envases distribuidos normalmente. La muestra arroja $\overline{x} = 12.102$ y $s^2 = 0.00335$.

La desviación estándar máxima permitida para declarar el proceso de llenado bajo control es de .05oz, por lo tanto $\sigma^2 = 0.0025$. Así,

$$H_0 : \sigma^2 \leq 0.0025 \text{ vs } H_A : \sigma^2 > 0.0025.$$

[10] Adaptado de Edward J. Dudewics y Satya N. Mishra. (1988). Modern Mathematical Statistics, John Wiley & Sons.

Pruebas de Hipótesis

Sabemos que S^2 es un buen estimador de σ^2, por lo que escogerlo como parte de nuestro estadístico de prueba parece razonable.

De esta forma, para valores de S^2 muy alejados de σ^2, en este caso para valores grandes, rechazaríamos H_0.

Notemos que tanto μ como σ^2 son desconocidos, no nos interesa hacer inferencias sobre μ pero necesitamos calcular $\overline{X}$ para lograr la estimación de σ^2, dado que $\hat{\sigma}^2 = S^2$ (razón por la cual pierde un grado de libertad).

Por otra parte, conocemos la distribución de una transformación de S^2:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Tenemos ahora los ingredientes básicos para construir la prueba.

Pruebas de Hipótesis

Queremos:

$$P(S^2 > k) = P\left(\frac{(n-1)S^2}{\sigma_0^2} > \frac{(n-1)k}{\sigma_0^2}\right) = P(\chi^2_{n-1} > k^*) = \alpha,$$

donde, de nuevo, k^* puede ser escogido de tal forma que esa probabilidad se cumpla; es decir, tomando el percentil correspondiente de una distribución Ji-Cuadrada con $n-1$ grados de libertad.

Para una prueba de nivel α, rechazamos si

$$u = \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}.$$

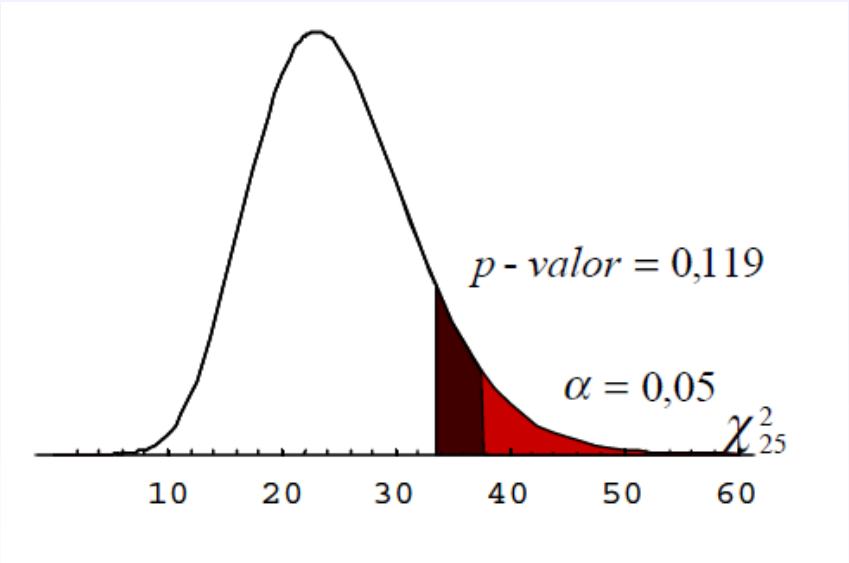
Fijemos en esta ocasión $\alpha = 0.05$. En R se puede comprobar que $\chi_{\alpha=0.05,25} = 37.65$.

Pruebas de Hipótesis

Calculemos el estadístico de prueba:

$$u = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(25)(0.00335)}{0.0025}.$$

El p-valor asociado a esta prueba es 0.119



lo cual indica que valores tan grandes o más como los de la varianza estimada en esta muestra son todavía factibles por fluctuaciones debidas al azar, aún y cuando $\sigma^2 \leq 0.0025$.

Pruebas de Hipótesis

Por lo tanto, no rechazamos $H_0 : \sigma^2 \leq 0.0025$ y podemos decirle al supervisor que, en base a la muestra que ha tomado, la máquina está bajo control y no hay necesidad de reajustarla.

Lectura adicional: sobre el p-valor: **The ASA Statement on p Values Context Process and Purpose.pdf**.

Pruebas de Hipótesis

Estos ingredientes son:

- Planteamiento de Hipótesis,
- Estadístico de Prueba,
- Región de Rechazo y Nivel de significancia,
- p-valor,
- **Potencia de la Prueba**,
- Conclusiones.

Potencia de la prueba

Potencia de la Prueba.

Probar una Hipótesis Nula H_0 con un procedimiento que se equivoca al rechazarla con probabilidad α pequeña es muy bueno pero, **¿qué hay de la probabilidad del Error Tipo II?**

Recordemos que la probabilidad del Error Tipo II es,

$$\text{P(No rechazar } H_0 | H_0 \text{ es falsa)}$$

es decir, nos referimos a la probabilidad de equivocarnos al no rechazar la Hipótesis Nula cuando la Hipótesis Alternativa es la correcta.

Llamaremos β a esta probabilidad.

Potencia de la prueba

La probabilidad complementaria $(1-\beta)$ está relacionada con la potencia de la prueba.

Definición

A la probabilidad de rechazar H_0 cuando H_A es cierta la llamaremos la Potencia de la Prueba

$$1-\beta = P(\textit{Rechazar}\, H_0 | H_A).$$

Al controlar la probabilidad del Error Tipo I en nuestra prueba no necesariamente lo hacemos con la del Error Tipo II.

Potencia de la prueba

Al construir una prueba que rechaza H_0 tomando en cuenta las discrepancias grandes entre ésta y la muestra, al mismo tiempo podemos estar pasando por alto discrepancias quizá no tan grandes, pero que son debidas a que realmente la muestra proviene de una población contemplada en la Alternativa y la prueba es "ciega" en reconocerlas como tales.

Una prueba ideal sería aquella que, para cualquier valor del parámetro en la Nula y en la Alternativa, sus probabilidades de Error Tipo I y Tipo II sean bajas. Esto se verá más claro a través del siguiente ejemplo.

Ejemplo: Observemos de nuevo qué pasa con el ejemplo de la corrosión de los conductores. Nuestras hipótesis son

$$H_0 : \mu \leq \mu_0 = 50 \quad vs \quad H_A : \mu > \mu_0 = 50.$$

Potencia de la prueba

Hemos fijado el cuantil $z_\alpha = 1.645$ como punto crítico de nuestro estadístico de prueba

$$Z = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \text{ donde } Z \sim N(0,1).$$

Debemos entonces ser capaces de volver a nuestras observaciones originales y ver a partir de qué valores de $\overline{X}$ estamos rechazando.

Sabemos que

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \quad \Longrightarrow \quad \overline{X} = \frac{\sigma}{\sqrt{n}} Z + \mu, \text{ donde } \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Potencia de la prueba

Recuerda que la distribución sobre la que trabajamos nuestras decisiones es la del estadístico de prueba, o equivalentemente la de $\overline{X}$.

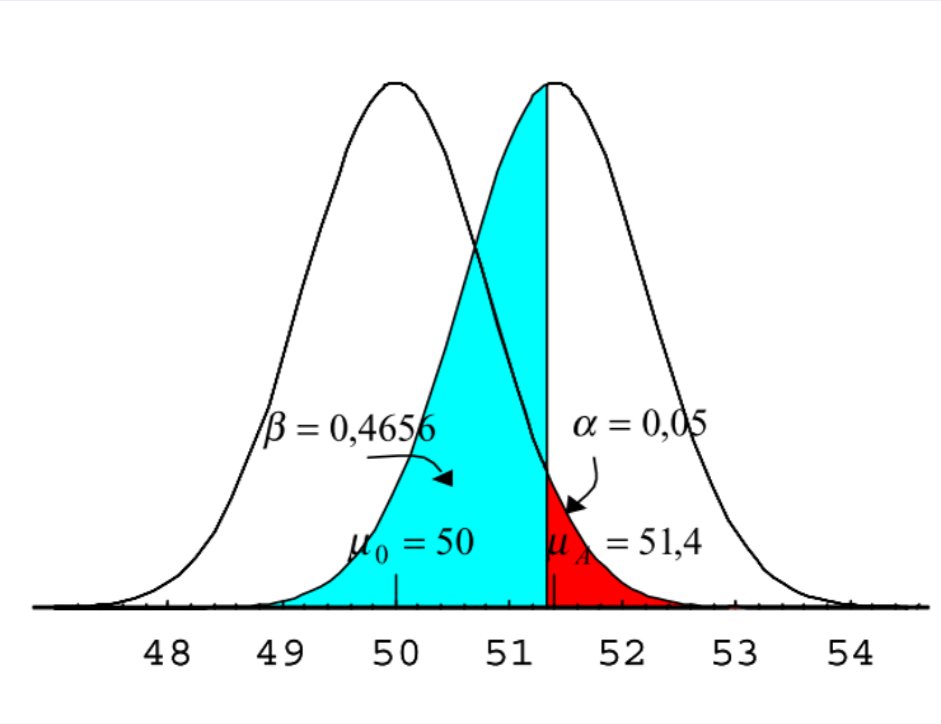
Por lo que

$$\overline{x}_\alpha = \frac{\sigma}{\sqrt{n}} z_\alpha + \mu_0 = \left(\frac{4.8}{\sqrt{35}} \right) (1.645) + 50 = 51.33$$

en este caso. De esta forma, hemos obtenido el cuantil en términos de la media muestral a partir del cual, para valores mayores a él, rechazaremos H_0; es decir, rechazaremos H_0 si $\overline{x} > \overline{x}_\alpha = 51.33$.

En la siguiente figura veremos lo que nos "cuesta" el ser tan precavidos.

Potencia de la prueba



La gráfica muestra las regiones de no-rechazo y de rechazo, junto con sus probabilidades; es decir, la región anterior a 51.33 y su probabilidad de ocurrencia bajo un valor $H_A(\mu_A = 51.4)$, así como la región a la izquierda y su probabilidad de ocurrencia bajo $H_0(\alpha)$.

Potencia de la prueba

Lo anterior significa que si la muestra en realidad proviene de una población con parámetro representado por $\mu_A = 51.4$, al no rechazar H_0 basados en $\overline{X} < 51.33$ nos estaremos equivocando con probabilidad $\beta = 0.4656$:

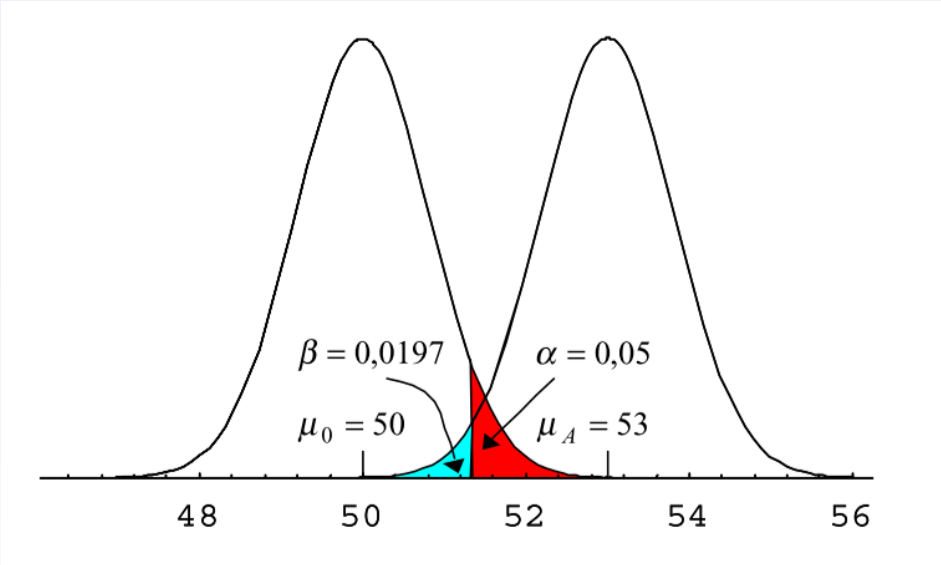
$$\begin{aligned}\beta &= P(\text{de aceptar } H_0 | H_A) = P(\text{de no rechazar } H_0 \text{ cuando es falsa}) \\ &= P(\text{Error tipo II cuando } \mu = 51.4) \\ &= P\left(\overline{X} \leq 51.33 \textit{ cuando } \overline{X} \sim N\left(51.4, \frac{23.04}{35}\right)\right) = 0.4656\end{aligned}$$

Tomando otro valor de μ en la Alternativa, por ejemplo, $\mu_A = 53$, tenemos

$$\beta = P\left(\overline{X} < \overline{x}_\alpha \textit{ cuando } \overline{X} \sim N\left(51.4, \frac{23.04}{35}\right)\right) = 0.0197.$$

Potencia de la prueba

La siguiente figura ilustra lo anterior.



,

Nuestra β pasa de 0.4656 a 0.0197; es decir, la probabilidad de que se presenten valores de $\overline{X}$ en la región de no-rechazo disminuye al considerar valores de μ_A alejados de μ_0.

Potencia de la prueba

El valor de β depende entonces del nivel de confusión entre las poblaciones definidas por H_0 y por H_A. Varias cosas más hay que observar en las gráficas.

• En los dos casos el α es la misma, ya que ésta se fija para H_0. Lo que cambia es la consideración de distintos valores particulares μ correspondientes a la Hipótesis Alternativa.

Aún y cuando $\mu_A = 51.4$, puede ocurrir en una muestra particular que $\overline{x} < 51.33$, y la regla nos dirá que rechacemos, cometiendo entonces el Error Tipo II. La probabilidad de que este error ocurra es 0.4656.

Al considerar otro valor de μ en la Alternativa, $\mu_A = 53$, un valor muestral de $\overline{X} < 51.33$ es mucho menos probable de que ocurra, y por lo tanto, también es menos probable que nos equivoquemos (Error Tipo II).

Potencia de la prueba

• Para cada valor de μ en la Hipótesis Alternativa tenemos un valor de β que le corresponde, dado el nivel de significancia α.

• Si α disminuye, β se incrementa, siempre que se considere el mismo valor de μ y de los demás parámetros involucrados en la forma distribucional.

Esto último implica que, en la implementación de una prueba en un problema particular, debe existir un equilibrio entre α y β.

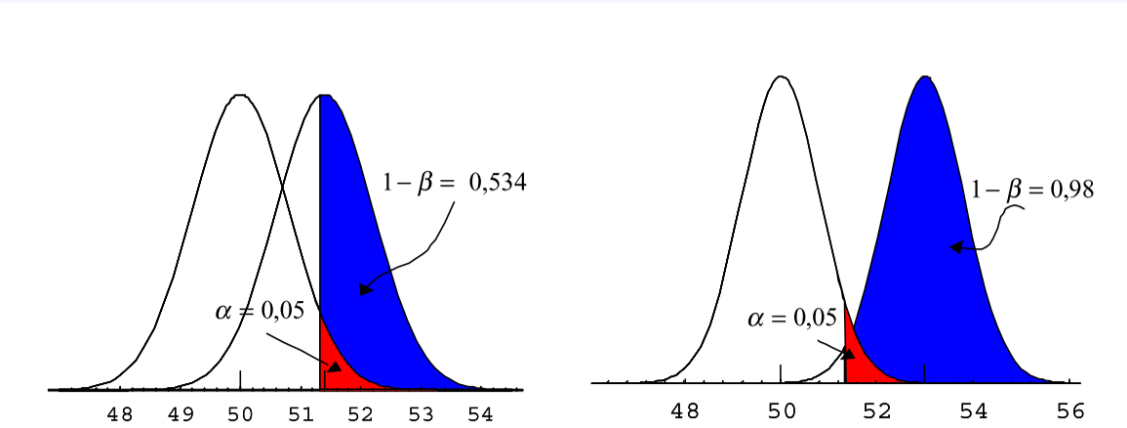
En la práctica se toma el valor más grande de α que pueda ser tolerado, lo cual significa tener muy en claro la situación en la que se está aplicando la prueba.

Potencia de la prueba

Los Errores Tipo I y II deben ser interpretados a la luz del problema que dio origen al uso de la prueba, ya que sólo así es posible evaluar las consecuencias de uno y otro tipo de error.

En algunas ocasiones es más grave tolerar el Error Tipo II que el Tipo I, por lo que es factible incrementar nuestra α.

La probabilidad complementaria a β, $1-\beta$, es la probabilidad de rechazar H_0 cuando H_A es cierta, que en las gráficas sería el área bajo la curva de la Alternativa a la derecha del cuantil 51.33.



Potencia de la prueba

Así, la Potencia de la prueba es 0.534 cuando $\mu_A = 51.4$ y 0.9803 cuando $\mu_A = 53$.

Evidentemente el cálculo de estas probabilidades implica el conocimiento de la distribución del estadístico de prueba bajo H_A. Para ello, como cuando tratabamos de establecer la región de rechazo, es conveniente estandarizar

$$\beta = P(\overline{X} < \overline{x}_\alpha | H_A) = P(\overline{X} < \overline{x}_\alpha \text{ cuando } \overline{X} \sim N(\mu_A, \frac{\sigma^2}{n}))$$

$$= P\left(\frac{\overline{X} - \mu_A}{\frac{\sigma}{\sqrt{n}}} < \frac{\overline{x}_\alpha - \mu_A}{\frac{\sigma}{\sqrt{n}}}\right) \quad \text{donde } \bar{x}_\alpha = \frac{\sigma}{\sqrt{n}} z_\alpha + \mu_0$$

(esto es, se debe estandarizar ahora con respecto a los parámetros propuestos bajo la hipótesis alternativa)

$$= P\left(Z < \frac{\frac{\sigma z_\alpha}{\sqrt{n}} + \mu_0 - \mu_A}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z < z_\alpha + \frac{\mu_0 - \mu_A}{\frac{\sigma}{\sqrt{n}}}\right) \qquad (Z \sim N(0,1))$$

$$= P\left(Z < z_\alpha - \frac{\mu_A - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z < z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) = \Phi\left(z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$$

donde Φ es la función de distribución acumulada de una $N(0,1)$.

Potencia de la prueba

El complemento $1 - \Phi\left(z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$ nos da la potencia de la prueba.

Cuando el interés es primordialmente el cálculo de la potencia, éste se facilita expresándola directamente en términos de la acumulada

$$\begin{aligned}\beta &= P\left(Z < z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) \\ &= P\left(Z > -z_\alpha + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) \quad \text{(ya que la distribución es simétrica)} \\ &= 1 - \Phi\left(-z_\alpha + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)\end{aligned}$$

por lo que es claro que la potencia de la prueba queda expresada por

$$\Phi\left(-z_\alpha + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) = 1 - \beta.$$

Recalcando: Las Hipótesis iniciales son de la forma:

$$H_0 : \mu \leq \mu_0 \quad vs \quad H_A : \mu > \mu_0.$$

Potencia de la prueba

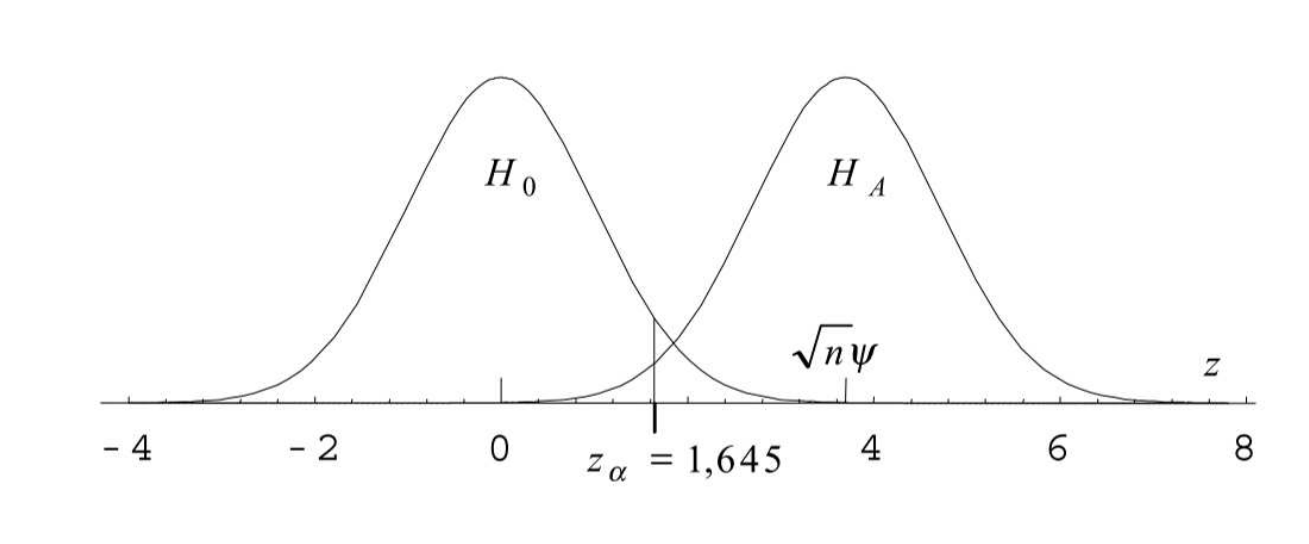
Recalcando: Las Hipótesis iniciales son de la forma:

$$H_0 : \mu \leq \mu_0 \quad vs \quad H_A : \mu > \mu_0.$$

Si llamamos ψ a $\frac{\mu_A - \mu_0}{\sigma}$, observamos que la potencia es el área a la derecha de z_α debajo de la curva centrada en $\frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma} = \sqrt{n}\psi$, esto es, $\sqrt{n}\psi$ representa la posición de μ_A en la escala estandarizada, z.

Entonces, ψ es un desvío estandarizado; representa el alejamiento estandarizado entre el valor de μ en la Hipótesis Nula y un valor particular de μ en la Alternativa. Un alejamiento mayor entre estos dos valores ocasiona un valor más grande de ψ, que se traduce en un mayor corrimiento hacia la derecha y en una mayor área, es decir, en una mayor potencia.

Potencia de la prueba



Considerar ψ en lugar de las desviaciones originales $\mu_A - \mu_0$ cobra su importancia en las aplicaciones prácticas, ya que normalmente se desconoce el valor de σ y se trabaja considerando directamente el valor del desvío estandarizado; es decir, se trabaja con las desviaciones pero asigándoles valores a sus estandarizaciones, en vez de tomar primero el desvío y luego estandarizarlo.

Potencia de la prueba

Ahora, desglosemos un poco la potencia:

- Contiene a z_α, por lo cual depende del nivel de significancia de la prueba;
- depende del desvío $\mu_A - \mu_0$;
- de σ^2; y
- de n.

En primer lugar, cada vez que consideremos la potencia de una prueba no podemos desligarla de los valores de cada una de las cantidades anteriores, por lo que hablaremos de la potencia de una prueba dados los valores de α, de $\mu_A - \mu_0$, etc.

Además, si $\psi = \frac{\mu_A - \mu_0}{\sigma}$ se hace grande, la potencia aumenta. De esta manera, si los valores de $\mu_A - \mu_0$ aumentan y se mantiene la varianza constante, la potencia se incrementa.

Potencia de la prueba

Esto significa que la prueba tenderá a rechazar con mayor probabilidad si la muestra en realidad proviene de una población con media μ_A muy alejada de μ_0.

Si las poblaciones hipotetizadas "casi" no se confunden, la probabilidad de tomar una decisión errónea se minimiza.

También es posible desplazar hacia la derecha ψ si la varianza σ^2/n se hace pequeña, lo cual es posible incrementando n, el tamaño de muestra.

De nuevo, dado que en la práctica típicamente no conocemos la varianza, es posible asociar un tamaño de muestra a la potencia de la prueba cuando tenemos en mente una desviación $\mu_A - \mu_0$, asignando valores a su estandarización ψ.

Veremos después esto último.

Potencia de la prueba

Para el caso en que σ^2 sea desconocida, se sustituye por S^2 y la potencia queda:

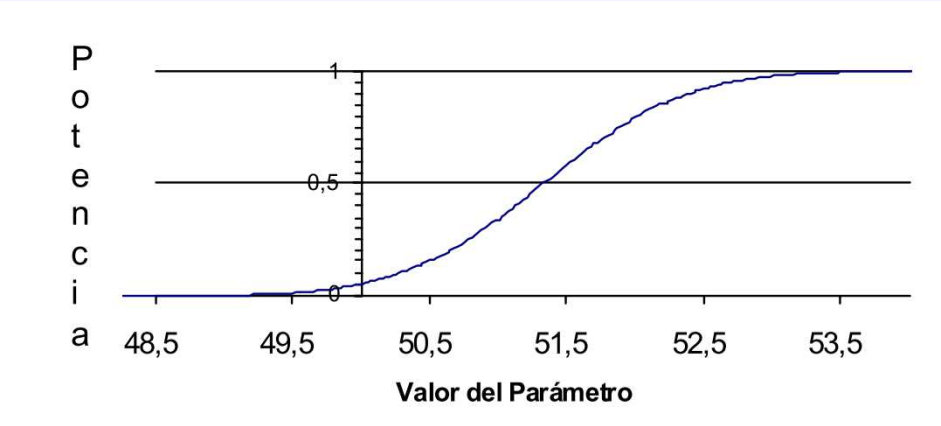
$$P\left(T < -t_{\alpha,n-1} + \frac{\sqrt{n}(\mu_A - \mu_0)}{s}\right).$$

Si te fijas, la expresión anterior depende de s, entonces no puedes calcular la potencia hasta que fijes n y saques la muestra y calcules n, por eso es importante recurrir a la t no central pues dependerá solo del parámetro de no centralidad, que es una función de la media!!!

Potencia de la prueba

Visto esto anterior, ¿cómo interpretar la potencia de una prueba?

Intuitivamente, la potencia de la prueba nos dice qué tan bueno es el procedimiento para detectar que nuestra muestra provenga de una población distinta de la que especifica la Hipótesis Nula, o sea, que nos indique rechazar la Hipótesis Nula cuando sea falsa (no es plausible).



En esta gráfica se muestra la potencia de la prueba para los distintos valores de μ tanto en la Hipótesis Alternativa como en la Nula.

Potencia de la prueba

Como puedes ver, para valores de μ anteriores a μ_0 inclusive, la potencia es menor o igual a 0.05, el α de la prueba.

Así, la potencia se puede plantear en un contexto más general como la probabilidad de rechazar la Hipótesis Nula para cualquier valor de μ.

Esto es consistente con nuestra intención inicial de rechazar H_0 con probabilidad α si en realidad el parámetro es μ_0; cuando fijamos α pequeño, queremos que la potencia sea baja para los valores admisibles del parámetro dentro de la Hipótesis Nula.

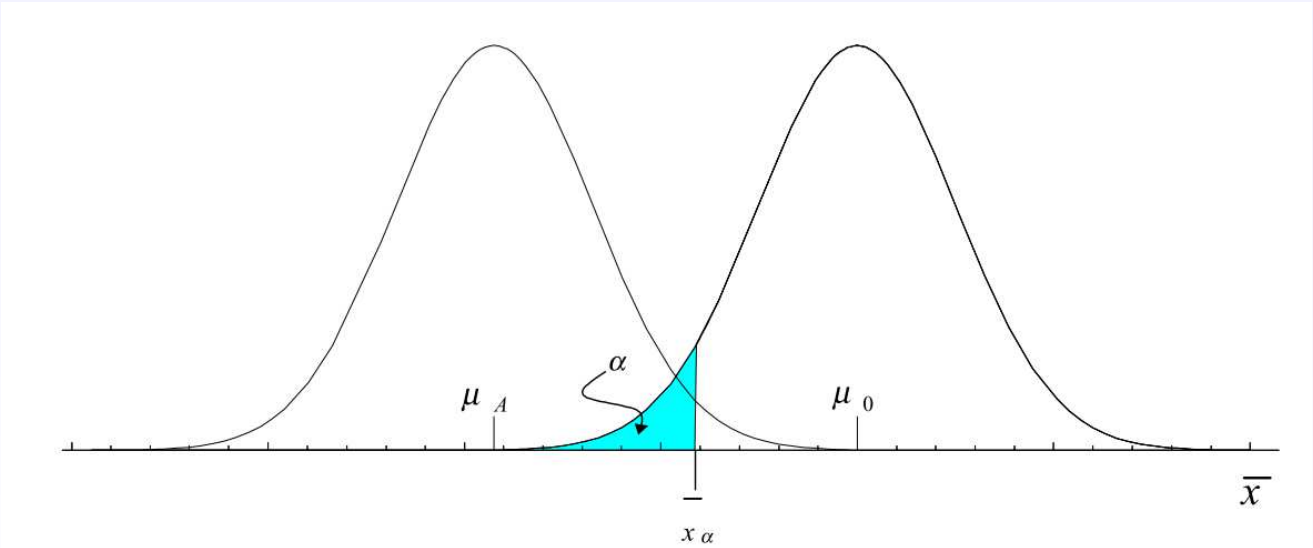
Potencia de la prueba

Consideremos ahora la potencia para el juego de hipótesis opuesto,

$$H_0 : \mu \geq \mu_0 \quad vs \quad H_A : \mu < \mu_0$$

para una población $N(\mu, \sigma^2)$ y σ^2conocida.

Lo que buscamos esta vez son valores del desvío $\overline{X} - \mu_0 \leq k$ tales que la probabilidad de que ocurran bajo la Hipótesis Nula $H_0 : \mu \geq \mu_0$ sea α. Es decir:



Potencia de la prueba

Analizando la gráfica, es claro que al estandarizar el desvío el percentil debajo del cual se encuentra una probabilidad α debe ser negativo.

De esta forma, procediendo de manera similar al anterior juego de hipótesis:

$$P(\overline{X} - \mu_0 < k) = P\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} < \frac{\sqrt{n}k}{\sigma}\right) = P(Z < -z_\alpha)$$

donde $Z \sim N(0,1)$ y $\frac{\sqrt{n}k}{\sigma} = -z_\alpha$ De esta forma,

$$\text{rechazamos } H_0 \text{ si } \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = z < -z_\alpha$$

Por lo tanto, si queremos considerar la regla de decisión y por consiguiente la región de rechazo en términos de los valores muestrales de $\overline{X}$, debemos fijar el percentil crítico así:

$$\overline{x}_\alpha = \frac{\sigma(-z_\alpha)}{\sqrt{n}} + \mu_0.$$

Potencia de la prueba

Ahora tendremos que considerar β de esta forma:

$$\beta = P(\overline{X} > \overline{x}_\alpha | H_A) = P(\overline{X} > \overline{x}_\alpha \text{ cuando } \overline{X} \sim N(\mu_A, \sigma^2))$$

$$= P\left(\frac{\overline{X} - \mu_A}{\sigma/\sqrt{n}} > \frac{\overline{x}_\alpha - \mu_A}{\sigma/\sqrt{n}}\right) = P\left(Z > \frac{\frac{\sigma(-z_\alpha)}{\sqrt{n}} + \mu_0 - \mu_A}{\sigma/\sqrt{n}}\right)$$

$$= P\left(Z > -z_\alpha + \frac{(\mu_0 - \mu_A)}{\sigma/\sqrt{n}}\right) = P\left(Z > -z_\alpha - \frac{(\mu_A - \mu_0)}{\sigma/\sqrt{n}}\right)$$

$$= P\left(Z > -z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$$

$$\Rightarrow \text{Potencia de la prueba} = P\left(Z \leq -z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$$

$$= \Phi\left(-z_\alpha - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right).$$

Potencia de la prueba

Procediendo de manera similar, si desconocemos σ^2 la potencia queda

$$P\left(T < -t_{\alpha,n-1} - \frac{\sqrt{n}(\mu_A - \mu_0)}{s}\right),$$

donde $T \sim t$ de Student con n-1 grados de libertad.

Nota: Sin embargo, no es equivalente al caso de la normal. Para calcularla se necesita S y para ello la muestra, por ende no se puede observar antes de sacar los datos, a diferencia del otro caso.

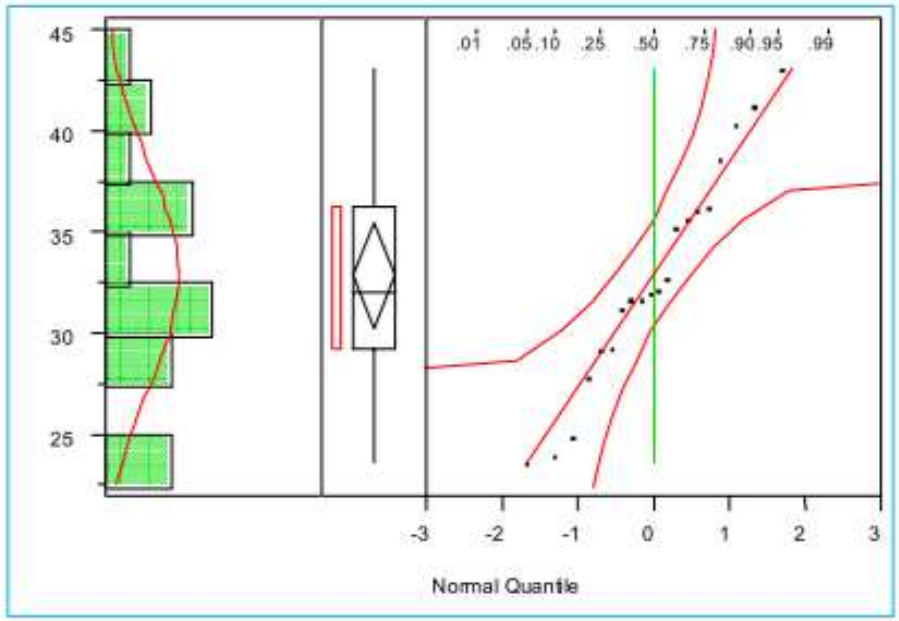
Potencia de la prueba

Ejemplos.

Ejemplos 1. Retomemos el ejemplo que se ha discutido sobre el rendimiento de combustible y el deseo del grupo de consumidores de verificar lo que el fabricante dice. La muestra que se tomó fue de 20 automóviles.

Realicemos la prueba para las hipótesis. Primero, verifiquemos normalidad.

Potencia de la prueba



Moments	
Mean	32,90450
Std Dev	5,55449
Std Error Mean	1,24202
Upper 95% Mean	35,50406
Lower 95% Mean	30,30494
N	20,00000
Sum Weights	20,00000

Test for Normality
Shapiro-Wilk W Test

W	Prob<W
0,967258	0,6914

Potencia de la prueba

¿Qué opinas al analizar el Q-Q Plot? Parecen normales, ¿no?

Notarás que aparece al final del informe la prueba para normalidad de Shapiro-Wilk. Se trata de una prueba de bondad de ajuste muy usada para probar la Hipótesis Nula de que la muestra viene de una población normal contra la Hipótesis Alternativa de que proviene de alguna otra población no normal. Su consideración equivale a verificar normalidad mediante un método no gráfico, sino propiamente estadístico.

El estadístico de prueba para este método es W. Tiene la característica de no tomar valores mayores a uno. Además, cuando la muestra proviene de una población normal, el estadístico W tiene la propiedad de acercarse a uno.

Potencia de la prueba

De esta forma, valores pequeños del estadístico indican falta de normalidad, mientras que cercanos a uno, la confirman. Un punto importante de esta prueba es que no requiere que los valores de los parámetros μ, σ^2 sean conocidos.

La interpretación de los resultados es del mismo tenor que para las pruebas que aquí hemos presentado:

- Sabiendo que las observaciones cumplen con ciertos supuestos distribucionales (en las pruebas hasta aquí consideradas la normalidad ha sido ese supuesto) y conociendo la distribución del estadístico de prueba, es posible fijar regiones de rechazo y calcular p-valores, dadas las hipótesis que estemos considerando.

Potencia de la prueba

- Nota además que para llegar a la conclusión última de una prueba es necesario establecer H_0 y ponderar la evidencia que la muestra presenta en contra de ella, ya sea a través del valor observado del estadístico de prueba (más precisamente, a través de su ubicación en la región de aceptación o de rechazo), o equivalentemente, a través del p-valor.

La prueba de Shapiro y Wilk para verificar Normalidad sólo tiene como supuesto base que se cuenta con una muestra aleatoria. Aunque no presentamos la forma explícita de W ni su densidad, sí presentamos lo que nos interesa para tomar una decisión: el valor observado del estadístico y el p-valor (Prob$<$W).

En general, si no conocemos los supuestos distribucionales, aún y cuando contemos con los estadísticos de prueba y sus p-valores, tomar decisiones en base únicamente a esta información puede llevarnos a errores graves.

Potencia de la prueba

Esto es, si haces análisis con software estadístico como el R, siempre te arrojará esas cantidades pero no por ello estarás siempre en libertad de hacer interpretaciones de las mismas, pues falta la parte clave, qué hipótesis se están probando y bajo que condiciones experimentales se tomó la muestra.

Ahora, en base a lo anterior veamos qué pasa con nuestros datos. Recordemos qué se obtuvo:

Test for Normality
Shapiro-Wilk W Test

W	Prob<W
0,967258	0,6914

¿Qué concluyes? Dado que el valor del estadístico está cercano a uno y el p-valor es grande, Aceptamos normalidad!!!

Potencia de la prueba

Refresquemos las hipótesis del problema:

$$H_0 : \mu \geq 35 \; y \; H_A < 35$$

y hagamos la prueba al nivel de 10 %.

- Como no conocemos σ^2, procederemos a estimarla con $s^2 = 30.8523$.
- Rechazaremos H_0 en favor de H_A si $\frac{\sqrt{n}(\overline{X}-\mu_0)}{S^2} < -t_{0.10,19}$.
- Calculando: $\frac{\sqrt{n}(\overline{X}-\mu_0)}{s^2} = -1.687 < -t_{0.10,19} = -1.328$, por lo cual rechazamos H_0 al nivel del 10 % .

Si realizaramos una prueba de nivel $\alpha = 0.05$, $-t_{0.05,19} = -1.729$, y . . . no rechazaríamos. Esto debe ponernos a pensar seriamente qué estamos haciendo cuando fijamos un valor específico de α.

Potencia de la prueba

La potencia de la prueba para los dos valores de α y el valor $\mu_A = 32$ en la Alternativa es:

$$P\left(T < -t_{0.10,19} + \frac{\sqrt{20}(35-32)}{5.554}\right) = P(T < 1.087) = 0.8546$$
$$P\left(T < -t_{0.05,19} + \frac{\sqrt{20}(35-32)}{5.554}\right) = P(T < 0.6864) = 0.7496$$

La razón por la cual calculamos la potencia para ese valor de μ_A es que los datos provienen en realidad de una población $Normal(32, 25)$ (fueron simulados).

En la práctica nosotros no sabemos cuál será el valor de la media bajo la alternativa por lo que se procedería a graficar las potencias para los dos valores de α.

Ejercicio: Construir esas gráficas para este caso particular.

Potencia de la prueba

La prueba al nivel del 10 % sí logró detectar la diferencia, pero la de nivel del 5 % no. Compara las potencias de ambas pruebas.

En el segundo caso nos tocó estar en el 25 % de los casos en que la prueba no detecta diferencias de 3, aunque éstas realmente existan. Recuerda que en este caso, puesto que los datos fueron simulados, sí conocemos el valor de μ, lo que nos posibilita afirmar que la diferencia existe.

Sin embargo, aquí debemos considerar qué consecuencias puede tener incurrir en el Error Tipo I y qué tan poderosa queremos que sea nuestra prueba.

Potencia de la prueba

Nota que el grupo de consumidores, al fijar un $\alpha = 0.10$, parece estar dispuesto a rechazar H_0 más frecuentemente en favor de detectar cambios en la media (los cubiertos por H_A). De esta forma, otorga pocas concesiones al productor de automóviles y cualquier indicación en la muestra de un promedio menor al especificado es tomado muy en serio.

Potencia de la prueba

Ejemplo 2: Veamos ahora la potencia de la prueba para la varianza revisando el ejemplo de la máquina de llenado de bebidas.

El par de hipótesis era:

$$H_0 : \sigma^2 \leq 0.0025 \quad vs \quad H_A : \sigma^2 > 0.0025$$

Por lo que nuestra regla para rechazar quedaba:

$$\text{Rechazar } H_0 \text{ si } u = \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha,\nu=n-1}$$

La potencia de la prueba es

$$\begin{aligned} P\left(U = \frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha,\nu=n-1} \text{ cuando } \sigma^2 = \sigma_A^2\right) &= P((n-1)S^2 > \sigma_0^2 \chi^2_{\alpha,\nu=n-1} \text{ cuando } \sigma^2 = \sigma_A^2) \\ &= P\left(\frac{(n-1)s^2}{\sigma_A^2} > \left(\frac{\sigma_0^2}{\sigma_A^2}\right)\chi^2_{\alpha,\nu=n-1}\right) \\ &= P\left(U^* > \left(\frac{\sigma_0^2}{\sigma_A^2}\right)\chi^2_{\alpha,\nu=n-1}\right) \end{aligned}$$

Potencia de la prueba

Ilustraremos el comportamiento de la potencia de esta prueba para distintos valores del cociente $\left(\frac{\sigma_0^2}{\sigma_A^2}\right)$:

	σ_0^2/σ_A^2					
	1	0.66	0.5	0.33	0.25	0.2
Potencia	0.05	0.456828	0.805375	0.982798	0.997956	0.999707

¿Qué te parece? Con este tamaño de muestra y un α de 0.05, la probabilidad de que rechacemos H_0 cuando la varianza en realidad es la mitad más grande que lo que dice la Nula es 0.456828; cuando es el doble, 0.805375; si la triplica, 0.982798, y así.

Podemos preguntarnos qué tan poderosa es la prueba, lo que dependerá de la magnitud del cociente considerado en cada ocasión.

Potencia de la prueba

Ejercicio: Realiza una prueba de nivel $\alpha = 0.01$ para las siguientes hipótesis: $H_0 : \sigma^2 = 8$ *vs* $H_A : \sigma^2 < 8$ a partir del archivo datvar.txt.

a) Verifica inicialmente normalidad y concluye sobre la plausibilidad de la misma en este conjunto de datos.

b) Utiliza el estadístico discutido en este capítulo y el que se presenta además en el capítulo anterior cuando la población no es normal.

c) Calcula la potencia de la prueba para valores del cociente 1.5, 2, 2.5, 3.

Pruebas de Dos Colas

Hasta ahora sólo hemos considerado pruebas para Hipótesis Nulas compuestas y Alternativas también compuestas (parámetros en un intervalo). Otro par de hipótesis muy común es el que toma un valor simple para el parámetro en la Nula y todos los demás valores que se admiten para el parámetro en la Alternativa:

$$H_0 : \theta = \theta_0 \qquad vs \qquad H_A : \theta \neq \theta_0.$$

Una prueba para estas hipótesis rechazará tanto con valores pequeños del estadístico, como con valores grandes, por lo cual, es necesario considerar las probabilidades de estos dos eventos conjuntamente para obtener el α deseado para la prueba.

Pruebas de Dos Colas

Así, en una prueba de hipótesis para la media, se desea que:

$$P\left(\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} > k^*\right) + P\left(\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} < k^{**}\right)$$
$$= P(Z > k^*) + P(Z < k^{**})$$

Tomamos $k^* = z_{\alpha/2}$ y $k^{**} = -z_{\alpha/2}$, lo que nos asegura que la suma de estas probabilidades sea α, ya que la distribución de Z es simétrica. De esta forma, la regla para rechazar H_0 queda:

Rechazar H_0 si $z < -z_{a/2}$ ó $z > z_{a/2}$, o bien, equivalentemente, si $|z| > z_{\alpha/2}$.

Pruebas de Dos Colas

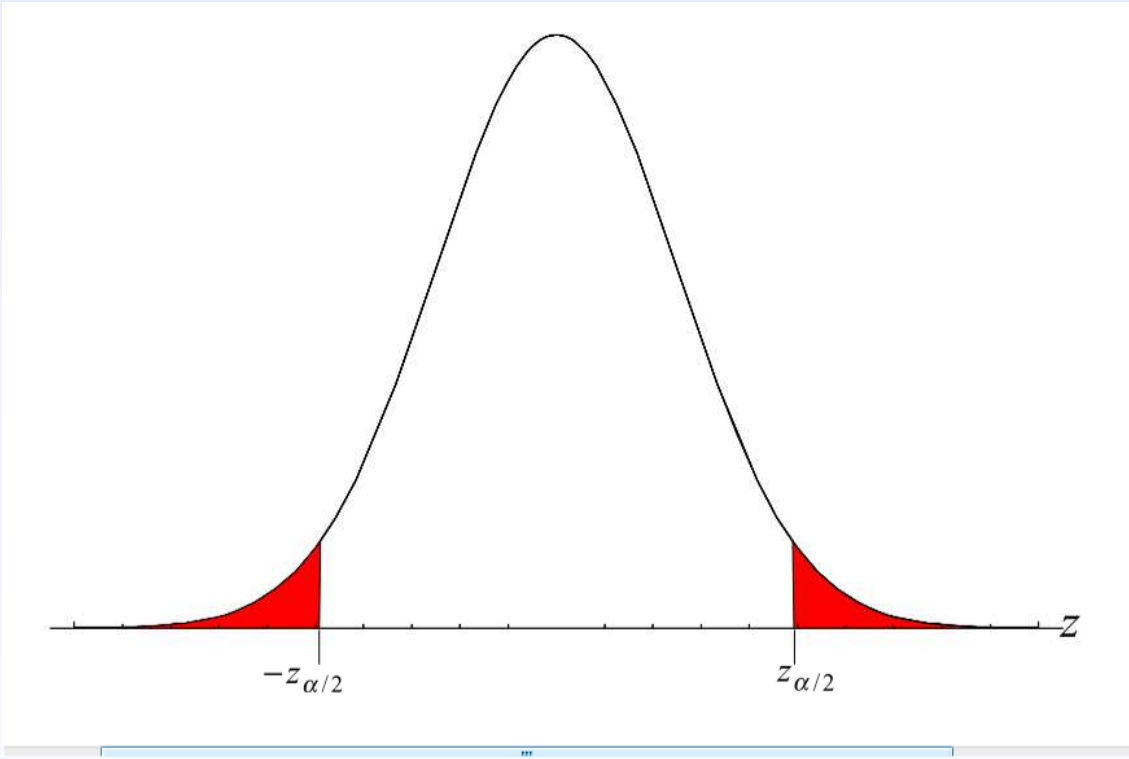


Figura: En esta situación, $z_{\alpha/2}$ es el percentil de la distribución $N(0,1)$ que deja una probabilidad de $\alpha/2$ para valores mayores que él.

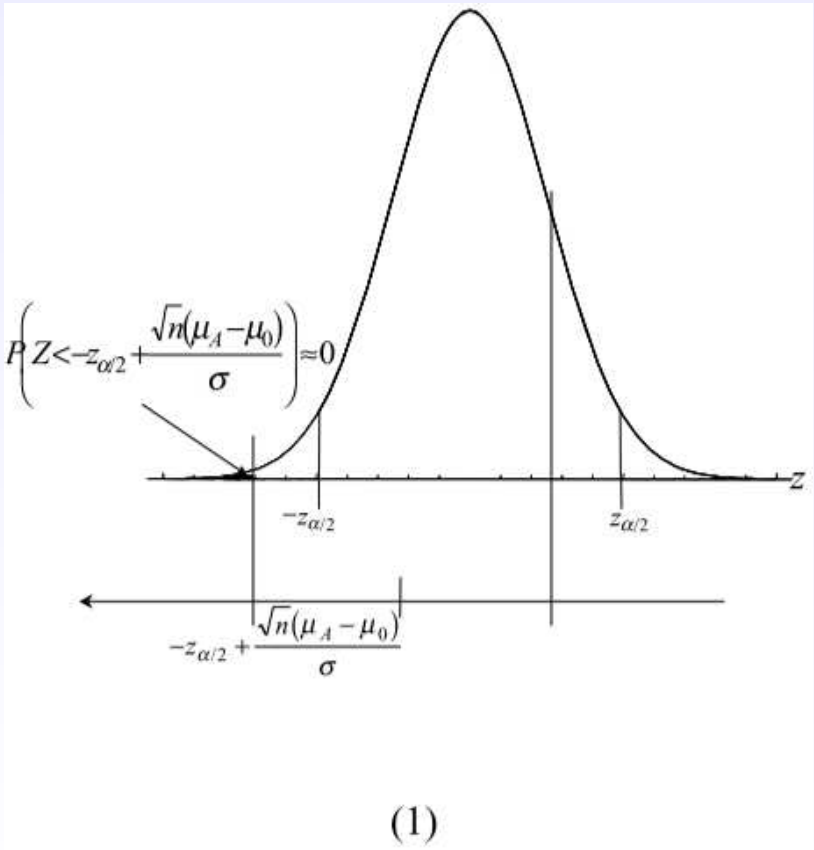
Pruebas de Dos Colas

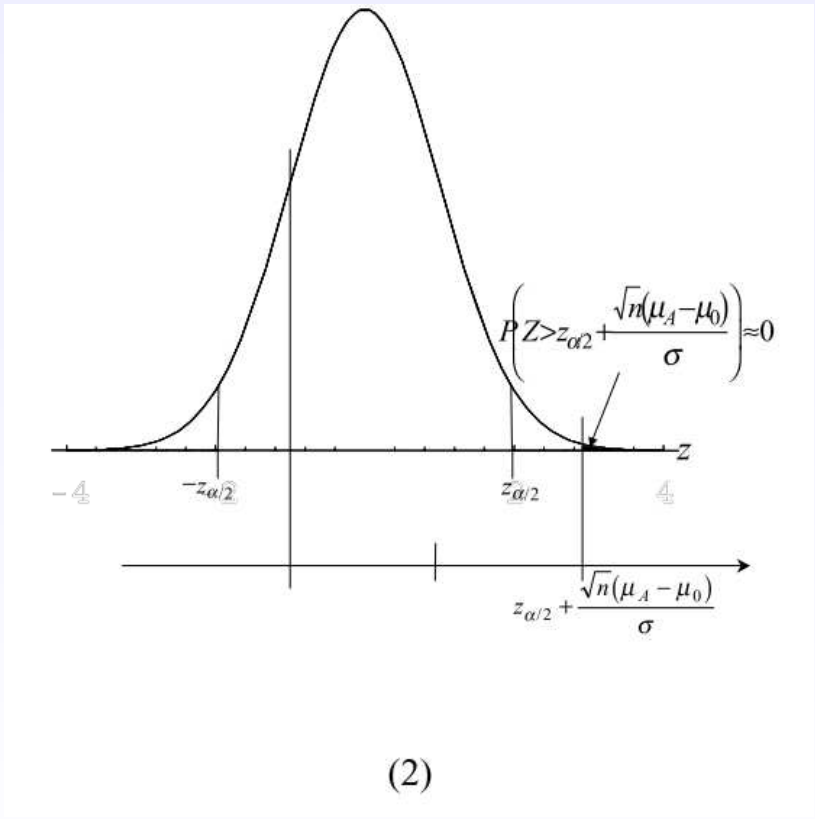
Entonces para la prueba de hipótesis de la media, adaptando las fórmulas para la potencia definidas anteriormente para las pruebas de una cola, la potencia de la prueba se define como:

$$
\begin{aligned}
1-\beta &= P\left(Z < -z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) + P\left(Z < -z_{\alpha/2} - \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) \\
&= P\left(Z < -z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) \\
&\quad + P\left(Z > z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) \quad \text{(por simetría de la distribución)}
\end{aligned}
$$

Pruebas de Dos Colas

Las siguientes gráficas ilustran el comportamiento de la potencia de la prueba de acuerdo a la magnitud y dirección del desvío $\frac{\sqrt{n}(\mu_A-\mu_0)}{\sigma}$:





Pruebas de Dos Colas

En las figuras anteriores podemos observar que:

1. Si $\mu_A < \mu_0$, entonces $\mu_A - \mu_0 < 0$ de tal forma que $P\left(Z < -z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$ se aproxima a cero conforme $\mu_A - \mu_0$ se hace grande (negativamente). Figura (1).
2. Observa que si $\mu_A > \mu_0$, entonces $\mu_A - \mu_0 > 0$, por lo que $P\left(Z > z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$ se acerca a cero conforme $\mu_A - \mu_0$ se hace muy grande (positivamente). Figura (2).

Por lo que la probabilidad de Error Tipo II queda:

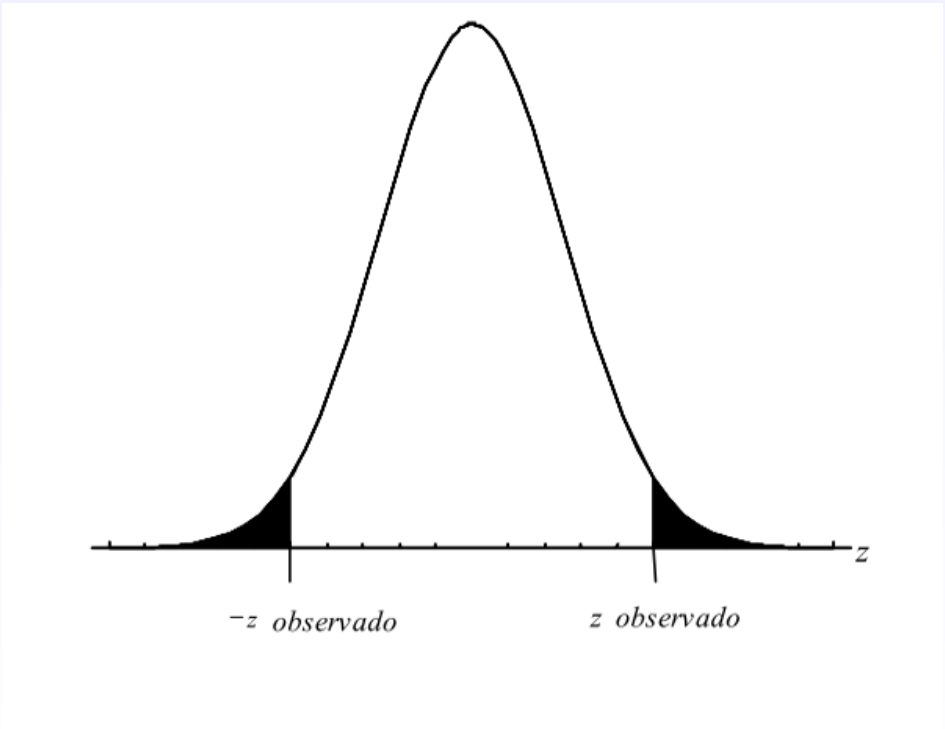
$$\beta = 1 - P\left(Z < -z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right) - P\left(Z > z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right)$$

$$= P\left(-z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma} < Z < z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}\right).$$

Pruebas de Dos Colas

El p-valor de la prueba queda expresado por:

$$\begin{aligned} p\text{-valor} &= P(Z < -z\ observado) + P(Z > z\ observado) \\ &= 2P(Z > |z\ observado|) \end{aligned}$$



Pruebas de Dos Colas

Ejemplo. El departamento de almacenamiento de una fábrica de harina debe inspeccionar cada lote de costales de harina que recibe. El administrador desea saber cada vez si el lote que recibe cumple con ciertas especificaciones en cuanto al porcentaje de absorción de humedad, a fin de hacer los ajustes necesarios en la humedad del lugar de almacenamiento y así preservar las propiedades del cargamento.

La harina debe tener un porcentaje promedio óptimo de absorción del 45 %, ya que con un incremento del 1 % el riesgo de fermentación es muy fuerte; por otra parte, si el porcentaje baja, sus propiedades se pierden debido a la deshidratación.

La práctica usual es la de tomar una muestra aleatoria sobre 30 costales, medir el porcentaje y realizar una prueba. Debido a esta práctica se sabe que las observaciones se distribuyen normalmente y que $\alpha = 5\,\%$.

Pruebas de Dos Colas

Los valores obtenidos fueron:

36.78	40.85	41.33	41.42	41.58	42.09
42.66	42.98	43.65	44.05	44.06	44.16
44.90	44.99	45.08	45.18	45.20	45.38
45.92	46.20	46.22	46.23	47.38	47.48
48.17	48.20	48.44	49.06	49.21	52.19

Las hipótesis son

$$H_0 : \mu = 45 \qquad vs \qquad H_A : \mu \neq 45$$

por lo que se rechazará si $\left|\frac{\sqrt{30}(\overline{X}-45)}{3}\right| > 1.96$, en una prueba de nivel del 5 %.

Pruebas de Dos Colas

Calculando el estadístico en base a los valores de la muestra:

$$\overline{x} = 45.035 \Rightarrow \left| \frac{\sqrt{30}(45.035 - 45)}{3} \right| = |0.0603| = 0.0603 < 1.96$$

por lo que no se rechaza H_0.

Si tomaramos la decisión en base al p-valor tendríamos:

$$\begin{aligned} p\text{-valor} = 2P(Z > |z\ observado|) &= 2P(Z > |0.0603|) \\ &\approx 2(0.4761) = 0.952 \end{aligned}$$

y no rechazaríamos.

Conclusión: El lote de costales de harina de donde proviene la muestra **no presenta** diferencias en el procentaje promedio de absorción de humedad respecto de la especificación de 45 %.

Pruebas de Dos Colas

Cotidianamente, los procesos en los que estamos interesados tienen una varianza de la cual quisieramos que fuera pequeña, ya que así se puede decir que el proceso está bajo control, realizar predicciones más precisas, etc. Es por esta razón que muchas pruebas de hipótesis acerca de la varianza son aplicadas para verificar algún incremento indeseable en la varianza.

Pensemos ahora en la necesidad de verificar si dado un evento la varianza se ha incrementado o disminuido. Interesa verificar también si ha disminuido, porque así podremos ganar precisión y obtener mejores inferencias, predicciones, etc.

Pruebas de Dos Colas

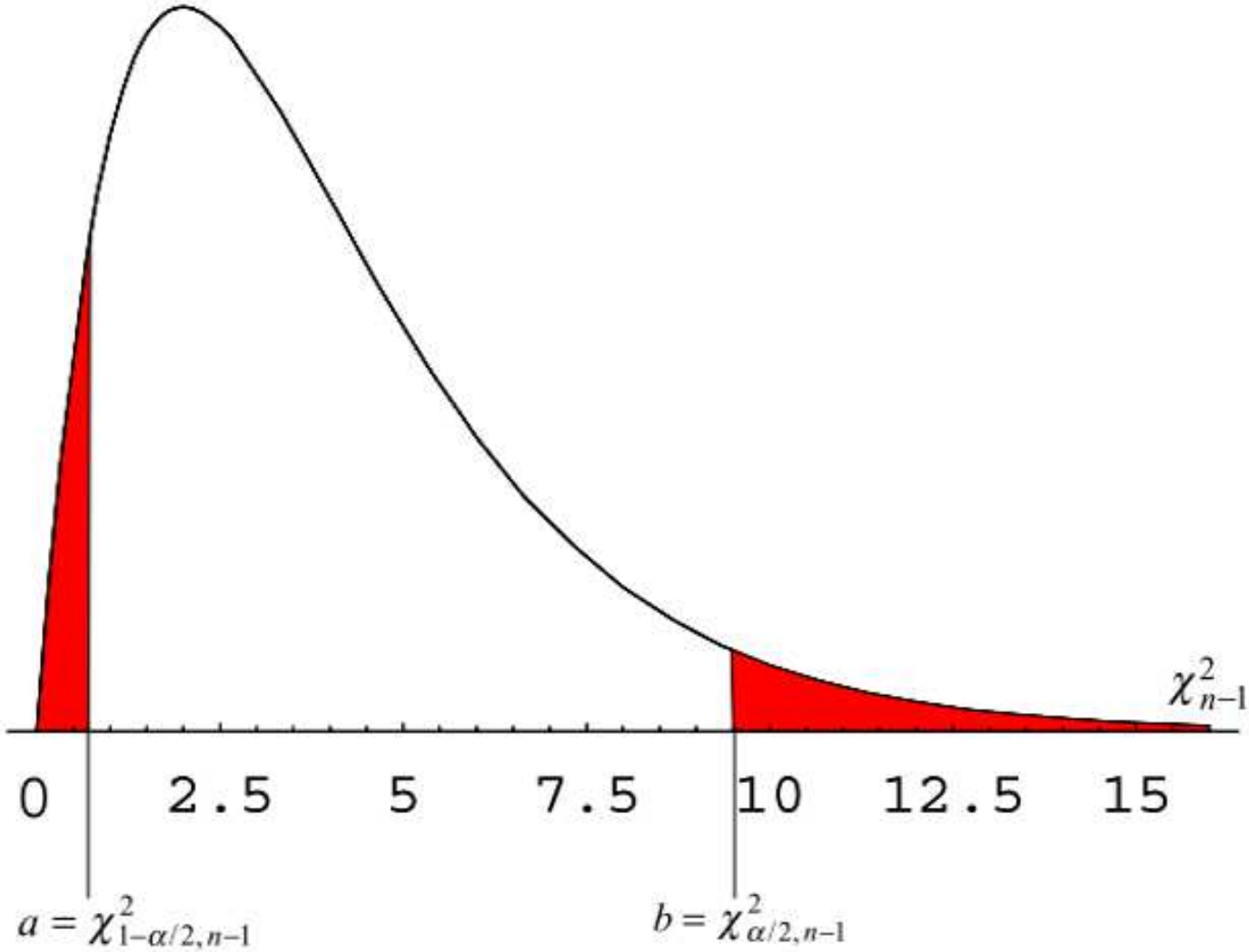
En estas circunstancias, el estadístico de prueba, bajo normalidad, sigue siendo $U = \frac{(n-1)S^2}{\sigma_0^2}$, pero ahora la región de rechazo que se busca es tal que

$$P(U < a) + P(U > b) = \alpha,$$

para una prueba de nivel α, ya que las hipótesis son

$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_A : \sigma^2 \neq \sigma_0^2.$$

Pruebas de Dos Colas



Pruebas de Dos Colas

Se acostumbra particionar la región de maneral que dejamos probabilidades de tamaño $\alpha/2$ de cada lado, aún cuando la distribución del estadístico de prueba no sea simétrica.

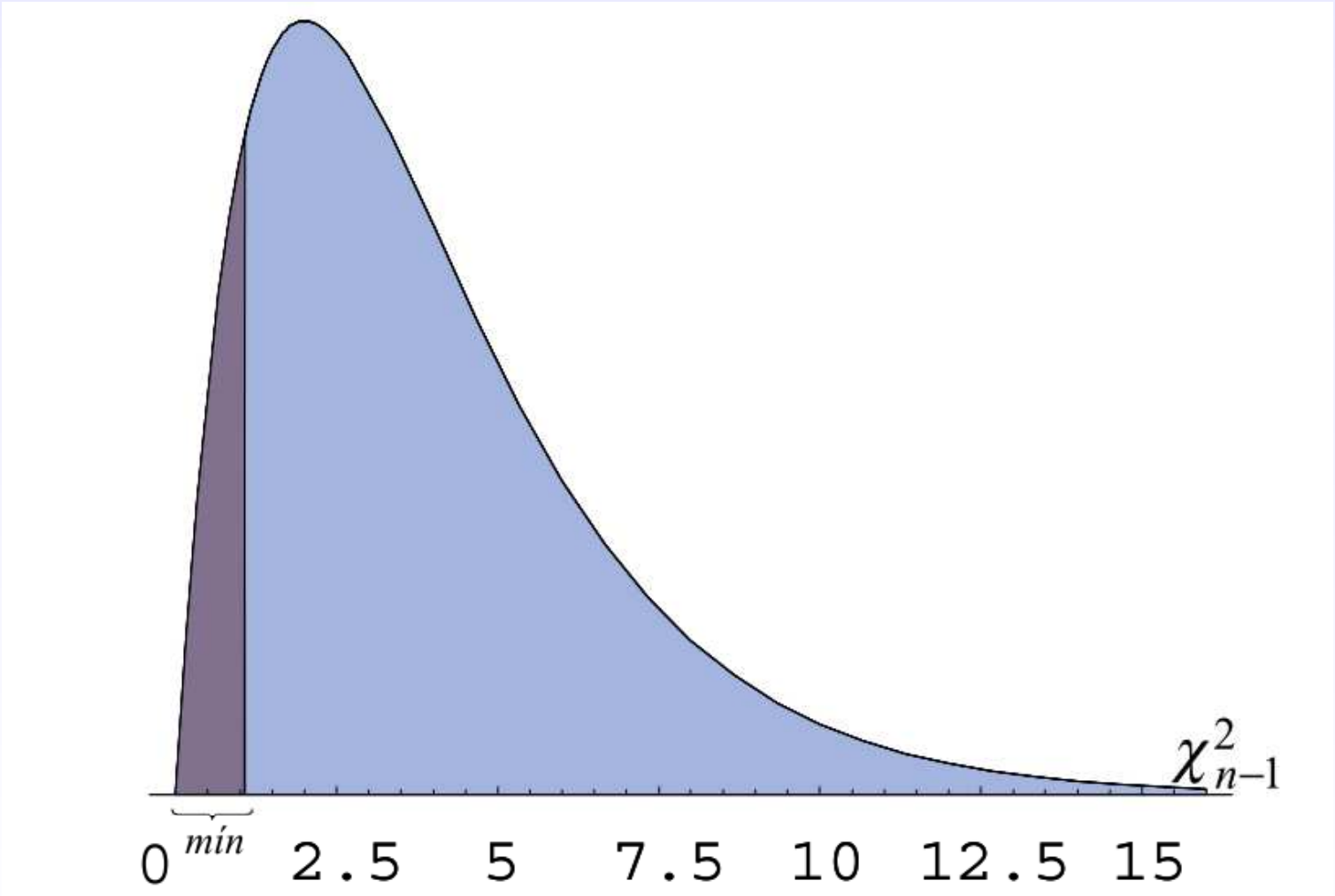
Esto es, se escogen los cuantiles correspondientes de una Ji-cuadrada con $n-1$ grados de libertad $a = \chi^2_{1-\alpha/2,n-1}$ y $b = \chi^2_{\alpha/2,n-1}$, y se rechaza si $u < \chi^2_{1-\alpha/2,n-1}$ ó $u > \chi^2_{\alpha/2,n-1}$.

El *p*-valor de una prueba de esta naturaleza puede ser calculado de la siguiente forma, aunque no es universalmente aceptada y esto sucede con todas las distribuciones de referencia no simétricas,

$$= 2\,\text{mín}\left\{P\left(\chi^2_{n-1} < U\ observado\right), P\left(\chi^2_{n-1} > U\ observado\right)\right\}$$

es decir, tomando dos veces la probabilidad más pequeña de las dos. La siguiente figura muestra esto gráficamente.

Pruebas de Dos Colas



Pruebas de Dos Colas

Ejercicio: La desviación estándar del diámetro interno de tuercas producidas por un proceso en una fábrica es de 0.0004cm. Un día, como parte de acciones correctivas producto de un diagnóstico en los componentes del proceso, son reemplazadas ciertas partes mecánicas y se espera que la variación del proceso cambie, pero no se sabe en qué dirección.

Para verificar algún cambio se toma una muestra aleatoria:

0,501081	0,499666	0,499195	0,500069	0,499803
0,500921	0,499513	0,499984	0,500451	0,500472
0,499908	0,500196	0,499234	0,500193	0,501199
0,499327	0,500036	0,500392	0,499124	0,501195
0,500131	0,499604	0,500323	0,500531	0,500599
0,499044	0,499379	0,499695	0,499600	0,499300
0,500165	0,499721	0,500026	0,498187	0,500645
0,500345	0,499962	0,500556	0,499923	0,499287
0,500179	0,499772	0,499380	0,499102	0,500477
0,49982	0,499415	0,500007	0,500609	0,500209

Pruebas de Dos Colas

Realiza un Q-Q Plot y verifica normalidad. Plantea las hipótesis que correspondan, haz la prueba y concluye: ¿Se ha alterado la varianza del proceso? (Los datos se encuentran en el archivo diam.txt).

Pruebas de Hipótesis

Estos ingredientes son:

- Planteamiento de Hipótesis,
- Estadístico de Prueba,
- Región de Rechazo y Nivel de significancia,
- p-valor,
- **Potencia de la Prueba**,
- Conclusiones.

Potencia y Tamaño de Muestra

Potencia y tamaño de muestra.

Ya comentamos que la potencia de la prueba puede ser mejorada incrementando el tamaño de la muestra en el caso de una prueba para la media μ, dado un valor específico de $\mu_A - \mu_0$. En el caso de las pruebas para varianzas el incremento del tamaño de la muestra también aumenta la potencia para un valor dado del cociente σ_0^2/σ_A^2.

¿Cómo verificarían vía simulación que el tamaño de la muestra aumenta la potencia en los casos anteriores?

Potencia y Tamaño de Muestra

Considerando las hipótesis:

$$H_0 : \mu \leq \mu_0 \quad vs \quad H_A : \mu > \mu_0,$$

deseamos que

$$1 - \beta = P\left(Z < \underbrace{-z_\alpha + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}}_{z^*} \right) = P(Z < z^*)$$

donde z^* es un percentil de una $N(0,1)$ que debe ser escogido para obtener la probabilidad deseada $1 - \beta$. Llamaremos z_β a ese percentil de la cola superior de una $N(0,1)$, debajo del cual hay una probabilidad $1 - \beta$, la potencia deseada para la prueba.

Potencia y Tamaño de Muestra

Así,

$$n = \frac{(z_\beta + z_\alpha)^2 \sigma^2}{(\mu_A - \mu_0)^2} = \frac{(z_\beta + z_\alpha)^2}{\Psi^2}.$$

Para las hipótesis:

$$H_0 : \mu \geq \mu_0 \quad vs \quad H_A : \mu < \mu_0$$

la fórmula queda igual.

Potencia y Tamaño de Muestra

En una prueba de dos colas debemos considerar las dos probabilidades complementarias:

$$P\left(Z < \underbrace{-z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}}_{-z_\beta^*}\right) \text{ y } P\left(Z > \underbrace{z_{\alpha/2} + \frac{\sqrt{n}(\mu_A - \mu_0)}{\sigma}}_{z_\beta^{**}}\right).$$

Así, debemos escoger $-z_\beta^*$ y z_β^{**} de forma que ambas probabilidades sumen $1-\beta$. Nota que los valores absolutos de $-z_\beta^*$ y z_β^{**} no son iguales.

Pensando la potencia de la prueba de acuerdo a la magnitud de su desvío, para la potencia que se desea en una aplicación específica una de las dos probabilidades se hace casi cero. Por esto, una aproximación al tamaño de muestra para una prueba de dos colas está dado por:

$$n = \frac{(z_\beta + z_{\alpha/2})^2}{\Psi^2}.$$

Potencia y Tamaño de Muestra

Ejercicio: Ya habíamos comentado que cuando se considera la potencia de una prueba de dos colas una de las dos probabilidades que es necesario calcular se vuelve despreciable conforme $\mu_A - \mu_0$ se hace grande o se hace muy pequeña, dependiendo de cuál de las dos direcciones de la Alternativa se estén considerando.

1. Calcula la potencia de una prueba bajo normalidad para las hipótesis $H_0 : \mu = \mu_0$ y $H_A : \mu \neq \mu_A$ en valores de $\sqrt{n}\Psi = -3, -2.5, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 2.5, 3$. Hazlo para valores de $\alpha = 0.05, 0.01$. A partir de la tabla que generes, haz la gráfica correspondiente, (en el eje x: $\sqrt{n}\Psi$, en el eje y: los valores de la potencia, para cada nivel α considerado).
2. Verifica a partir de que valores de $\sqrt{n}\psi$ una de las dos probabilidades es posible no tomarla en cuenta sin perder precisión. ¿Qué valores de las potencias calculadas desearías usar en una aplicación?

Potencia y Tamaño de Muestra

3. Nota la relación entre el tamaño de muestra para una prueba de medias de dos colas con respecto a la potencia que se desearía en una aplicación, cuál usarías?
4. ¿Qué criterios debes usar para seleccionar el tamaño de la muestra en aplicaciones concretas?

Potencia y Tamaño de Muestra

Quien realiza la prueba y desea asociarle una potencia particular debe proveer el valor de σ^2 y el valor del desvío $\mu_A - \mu_0$.

Una manera de lidiar con σ^2 desconocida es trabajar con el desvío estandarizado ψ.

Observemos que el tamaño de la muestra se hace grande para valores del desvío muy pequeños. Recordemos que existen circunstancias en que aumentar el tamaño de la muestra puede resultar caro. No tiene sentido incrementar el tamaño de muestra para la detección de valores de $\mu_A - \mu_0$ muy pequeños que en realidad no interesen; pero si desviaciones pequeñas son importantísimas, es necesario examinar el tamaño de muestra que se requiera para detectarlas. En este sentido, de nuevo, la consideración de la situación será lo que dicte cuál es el desvío mínimo que se quiere detectar con la prueba.

Potencia y Tamaño de Muestra

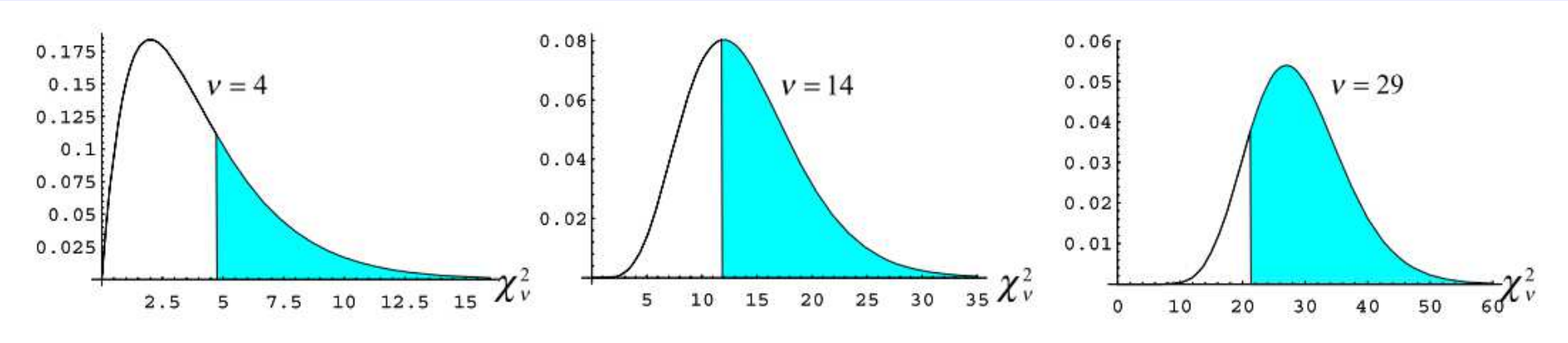
En lo que se refiere a los tamaños de muestra en relación a la potencia de la prueba para varianzas, no existen fórmulas en términos cerrados, por lo que se requiere consultar softwares adecuados.

La potencia de las pruebas para varianzas realizadas sobre la distribución Ji-Cuadrada se comporta algo distinto del caso de la distribución normal.

Mientras que en el caso normal es claro que conforme n crece, el desvío Ψ es desplazado, lo cual trae como consecuencia un incremento de la potencia, en la distribución Ji-Cuadrada no es tan evidente; con cada tamaño de muestra la forma de la distribución cambia, a diferencia de la normal, en la que lo único que cambiaba era su localidad (la desplazaba).

Ilustraremos este incremento en la potencia para tamaños de muestra 5, 15 y 30, cuando el cociente $\sigma_0^2/\sigma_A^2 = 0.05$ y el nivel de significancia de la prueba es de 0.05.

Potencia y Tamaño de Muestra



En cada gráfica, el punto a partir del cual el área se haya sombreada es $\frac{\sigma_0^2}{\sigma_A^2}\chi^2_{\alpha,\nu} = (0.05)\chi^2_{\alpha,\nu}$.

Las potencias para cada tamaño de muestra 5, 15 y 30 son 0.314, 0.618, 0.848, respectivamente. Así, si queremos ser capaces de detectar un incremento del doble en nuestra varianza con una prueba de este tipo y una muestra de tamaño 30, tendremos 0.848 de probabilidad de lograrlo si es que la muestra en realidad proviene de una población Normal con $\sigma_A^2 = 2\sigma_0^2$.

Potencia y Tamaño de Muestra

Ejemplo. Retomemos el ejemplo del porcentaje de absorción de humedad. Podemos preguntarnos si realmente ese tamaño de muestra es el indicado para detectar desviaciones del 1 % en las pruebas que se realizan.

Nuestras hipótesis son:

$$H_0 : \mu = 45 \quad vs \quad H_A : \mu \neq 45.$$

La potencia de la prueba queda expresada por:

$$1-\beta = P\left(Z < -z_{\alpha/2} + \frac{\sqrt{30}(\mu_A - 45)}{3}\right) + P\left(Z < z_{\alpha/2} + \frac{\sqrt{30}(\mu_A - 45)}{3}\right)$$

por lo que

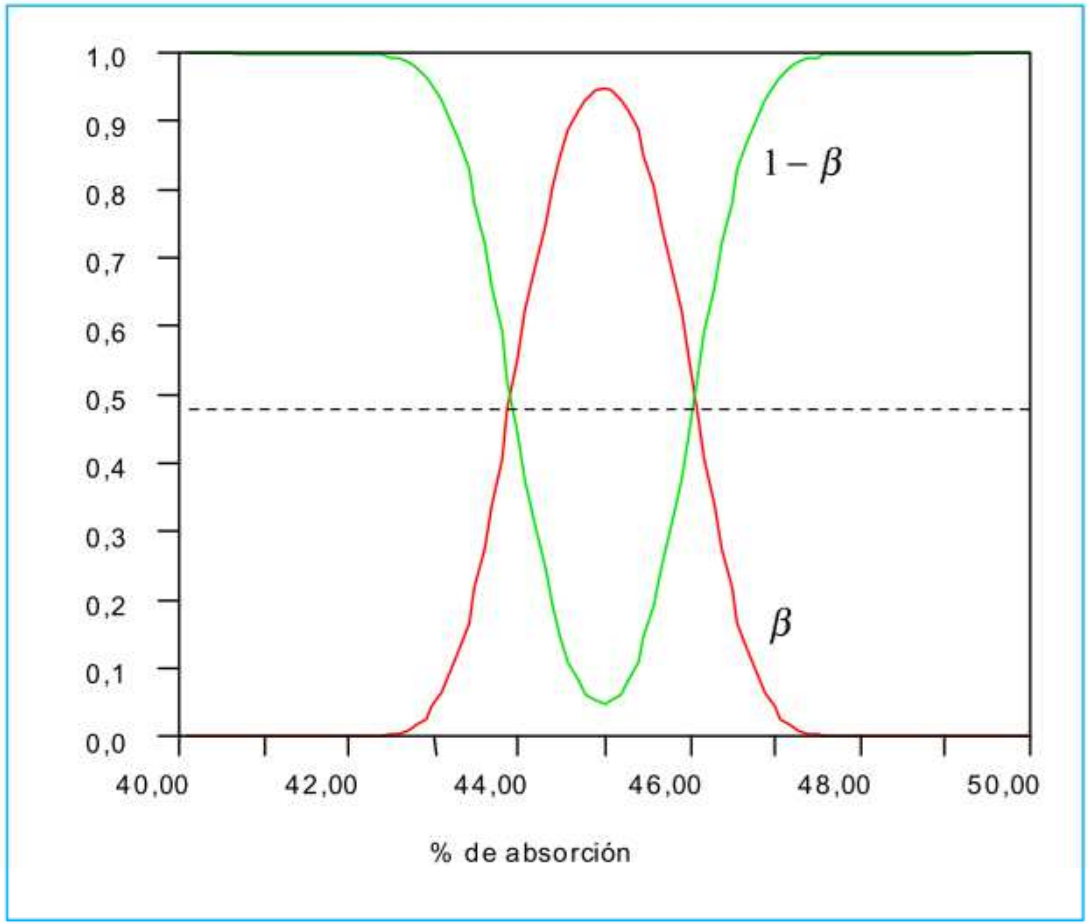
$$\beta = P\left(-z_{\alpha/2} + \frac{\sqrt{30}(\mu_A - 45)}{3} < Z < z_{\alpha/2} + \frac{\sqrt{30}(45 - \mu_A)}{3}\right).$$

Potencia y Tamaño de Muestra

Si fijamos nuestro nivel de significancia de la prueba en 5 %, $z_{\alpha/2} = 1.96$. De esta forma, la probabilidad de Error Tipo II y la potencia de la prueba con un nivel de significancia del 5 %, una desviación estándar de 3 y un tamaño de muestra de 30, si $\mu = 46$ *ó* 44, es de 0.553 y de 0.447, respectivamente.

En la siguiente gráfica se muestran las probabilidades anteriores para distintos valores de μ_A.

Potencia y Tamaño de Muestra



¿Qué podemos decir de esta práctica que había venido llevándose a cabo en la fábrica?

Potencia y Tamaño de Muestra

Dado ese tamaño de muestra, observa que se tienen casi las mismas probabilidades, tanto de detectar un incremento o un decremento del 1 % , como de no detectarlo con esta prueba. Es como si lanzaramos una moneda al aire!!! Evidentemente, si la desviación estándar fuera menor, el tamaño de muestra se acercaría más al indicado para detectar una desviación de esa magnitud.

Moraleja: analiza cada problemática por separado y decide sobre el tamaño de muestra de acuerdo a los valores especificos de los parámetros y nivel de significancia con los que tu prueba se realiza. La recomendación final para el administrador es que sería muy conveniente incrementar el tamaño de la muestra para detectar esa desviación.

Potencia y Tamaño de Muestra

Toma en cuenta, de esta breve discusión, que el tamaño de muestra responde a un fin global, el objetivo del estudio. En éste se hayan incluidos y articulados todos los demás aspectos: el α, la potencia de la prueba, las desviaciones que se desean detectar, las hipótesis, etc. La consideración conjunta de todos estos puntos entra en juego a la hora de decidir sobre el tamaño de muestra, a tal grado que no es posible desligar n de este fin englobador.

Busca software para consultar tablas de tamaños muestrales para diversas situaciones, entenderlos y ser capaz de ir más allá.

Cociente de verosimilitudes

Sea $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ un vector de parámetros y sea $\boldsymbol{X}$ un vector aleatorio con pdf $f_{\boldsymbol{\theta}}$. Consideremos la prueba de comparar la hipótesis nula $H_0 : \boldsymbol{X} \sim f_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ contra la alternativa $H_1 : \boldsymbol{X} \sim f_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$.

Una prueba de hipótesis donde se contrasta H_0 contra H_1 y se rechaza H_0 ssi $\lambda(\boldsymbol{X}) < c$, donde c es una constante y

$$\lambda(\boldsymbol{X}) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} f_{\boldsymbol{\theta}}(x_1, x_2 \ldots, x_n)}{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(x_1, x_2 \ldots, x_n)},$$

es llamada prueba de cociente de verosimilitudes generalizado.

El numerador del cociente de verosimilitudes λ es la mejor explicación de $\boldsymbol{X}$ (en el sentido de MV) que la hipótesis nula puede proveer, mientras que el denominador es la mejor explicación posible de $\boldsymbol{X}$. La hipótesis H_0 es rechazada si hay una mucho mejor explicación de $\boldsymbol{X}$ que la mejor proveída por H_0.

Cociente de verosimilitudes

Es claro que $0 \leq \lambda \leq 1$. La constante c se determina según el nivel de significancia α deseado mediante

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(\lambda(\boldsymbol{X}) < c) = \alpha.$$

Si la distribución de $\lambda(\boldsymbol{X})$ es continua, se puede obtener un c para cualquier valor de α. Cuando $\lambda(\boldsymbol{X})$ es una v.a. que no es continua, puede no ser posible encontrar un c que satisfaga exactamente la ecuación anterior.

Cociente de verosimilitudes

Ejemplo:

Consideremos el problema de comparar $\mu = \mu_0$ contra $\mu \neq \mu_0$ en una muestra Normal(μ, σ^2), donde μ y σ son desconocidos. Definamos $\boldsymbol{\Theta}_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}$ y $\boldsymbol{\Theta} = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. Si $\boldsymbol{\theta} = (\mu, \sigma^2)$, tenemos que

$$\begin{aligned} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} f_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \sup_{\sigma^2 > 0} \left\{ \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[\frac{-\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2} \right] \right\} \\ &= f_{\hat{\sigma}_0^2}(\boldsymbol{x}) \end{aligned}$$

donde $\hat{\sigma}_0^2$ es el EMV $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2$. Así,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{((2\pi)^{n/2}[\sum_{i=1}^n (x_i - \mu_0)^2]^{n/2}} e^{-n/2}.$$

Cociente de verosimilitudes

El EMV de $\boldsymbol{\theta} = (\mu, \sigma^2)$ cuando μ y σ^2 son desconocidos es $(\sum_{i=1}^n \frac{x_i}{n}, \sum_{i=1}^n \frac{(x_i - \bar{x})^2)}{n})$. Se sigue que

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sup_{\mu, \sigma^2} \left\{ \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left[\frac{-\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2} \right] \right\}$$

$$= \frac{1}{((2\pi)^{n/2} [\sum_{i=1}^n (x_i - \bar{x})^2]^{n/2}} e^{-n/2}.$$

Así,

$$\lambda(\boldsymbol{x}) = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right]^{n/2} = \left\{ \frac{1}{1 + \left[\frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \right\}^{n/2}.$$

Cociente de verosimilitudes

El test de cociente de verosimilitudes generalizado rechaza H_0 si

$$\lambda(\boldsymbol{X}) < c,$$

y ya que $\lambda(\boldsymbol{x})$ es una función decreciente de $\left[\frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^n n(x_i-\bar{x})^2}\right]$, rechazamos H_0 si

$$\left|\frac{\bar{x}-\mu_0}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}}\right| > c'$$

es decir, si

$$\left|\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}\right| > c''$$

donde $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i-\bar{x})^2$.

Cociente de verosimilitudes

El estadístico

$$t(\boldsymbol{X}) = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$$

tiene una distribución t_{n-1} (i.e. con $n-1$ grados de libertad). Bajo $H_0 : \mu = \mu_0$, tiene una distribución t_{n-1} central, pero bajo $H_1 : \mu \neq \mu_0$, $t(\boldsymbol{X})$ tiene una distribución t no central con $n-1$ grados de libertad y parámetro de no centralidad $\delta = (\mu - \mu_0)/\sigma$. Elegimos $c'' = t_{n-1,\alpha/2}$ de acuerdo con la distribución de $t(\boldsymbol{X})$ bajo H_0.

Ejercicio: Sea $X \sim$Binomial(n, p). Deseamos contrastar las hipótesis $H_0 : p \leq p_0$ contra $H_1 : p > p_0$. Calcule el estadístico $\lambda(X)$ de la prueba de cociente de verosimilitudes. Justifique $\lambda(X) < c$ si $x > c'$, para constantes c, c'. ¿Cómo podemos elegir c para un nivel de significancia α? ¿Existe c' para cualquier α?

Cociente de verosimilitudes

En el caso del ejercicio anterior, si tal c' no existe, elegimos un entero c' tal que

$$P_{p_0}(X > c') \leq \alpha \qquad \text{y} \qquad P_{p_0}(X > c' - 1) > \alpha.$$

En los casos anteriores, tuvimos que deducir la distribución de $\lambda(\boldsymbol{x})$ a partir de las propiedades de cada distribución. Sin emabargo, se cuenta con un resultado asintótico que facilita aplicar la prueba del cociente de verosimilitudes.

Teorema (Wilks)

Bajo algunas condiciones de regularidad en la densidad $f_{\boldsymbol{\theta}}$, bajo H_0 la v.a. $-2\log\lambda(\boldsymbol{X})$ se distribuye asintóticamente como una χ^2 con grados de libertad igual a la diferencia entre el número de parámetros independientes de $\boldsymbol{\Theta}$ y el número en $\boldsymbol{\Theta}_0$.

Cociente de verosimilitudes

Ejemplo: Anteriormente, analizamos la prueba de cociente de verosmilitudes para contrastar $\mu = \mu_0$ contra $\mu \neq \mu_0$ en una muestra Normal(μ, σ^2). En dicho caso encontramos que

$$\lambda(\boldsymbol{x}) = \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2}\right]^{n/2} = \left\{\frac{1}{1 + \left[\frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}\right\}^{n/2},$$

de donde se sigue que

$$-2 \log \lambda(\boldsymbol{x}) = n \log \left[1 + \left[\frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right] \sim \chi^2_1,$$

cuando $n \to \infty$.

No presentamos la demostración del Teorema de Wilks, pero en el material suplementario de la clase se puede encontrar el artículo original donde esto fue probado **TeoremaDeWilks.pdf**.

Cociente de verosimilitudes

Además, se incluyen otras dos referencias que estudian la distribución asintótica del cociente de verosimilitudes:**Asymptotic_Distribution_LR_Statistic.pdf** y **Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions.pdf**.

Ejercicio (Honors): Para el caso presentado en el ejemplo anterior, sin usar el Teorema de Wilks, justifique que la distribución asintótica de $-2\log\lambda(\boldsymbol{X})$ es χ^2_1.

Inferencia Bayesiana

Inferencia Bayesiana

La inferencia bayesiana se basa en los siguientes supuestos:

1. Elegimos una función de densidad $f(\theta)$ llamada **distribución a priori** que expresa nuestro grado de certidumbre o conocimiento sobre un parámetro θ antes de conocer los datos.
2. Elegimos un modelo estadístico $f(x|\theta)$ que refleje nuestro conocimiento sobre x dado θ. Note que ahora escribimos $f(x|\theta)$ en lugar de $f(x;\theta)$.
3. Después de observar los datos $X_1, \ldots, X_n$ actualizamos nuestro conocimiento y formamos la distribución **posteriori** $f(\theta|X_1, \ldots, X_n)$.

Inferencia Bayesiana

Supongamos que θ es discreto y que hay una sola sbservación X. Sea Θ el parámetro, así

$$\begin{aligned} P(\Theta = \theta | X = x) &= \frac{P(X = x, \Theta = \theta)}{P(X = x)} \\ &= \frac{P(X = x | \Theta = \theta) P(\Theta = \theta)}{\sum_\theta P(X = x | \Theta = \theta) P(\Theta = \theta)}, \end{aligned}$$

en virtud del *Teorema de Bayes*.

Inferencia Bayesiana

Para el caso continuo, usamos la función de densidad

$$f(\theta|X) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{1}$$

Si tenemos n observaciones $x_1, \ldots, x_n$, iid, reemplazamos $f(x|\theta)$ con $f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$.

Inferencia Bayesiana

Denotemos $X^n = (X_1, \ldots, X_n)$ y $x^n = (x_1, \ldots, x_n)$. Entonces

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta} \propto \mathcal{L}_n(\theta)f(\theta) \qquad (2)$$

En el lado derecho de 2 el término $\mathcal{L}_n(\theta)f(\theta)d\theta$ es constante y lo llamamos **constante de normalización**, la cual podemos descartar. En resumen podemos decir

posteriori es proporcional a las probabilidades anteriores

Dado que $\int f(\theta|x^n)d\theta = 1$, siempre podemos recuperar la constante, pero la omitiremos hasta que sea necesaria.

Inferencia Bayesiana

¿Qué hacemos con la posteriori?

Primero, podemos obtener una estimación puntual resumiendo el centro de la posteriori. Normalmente, usamos la media o la moda de la posteriori. La media de la posteriori es

$$\bar{\theta}_n = \int \theta f(\theta|x^n)d\theta = \frac{\int \theta \mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta} \tag{3}$$

También podemos obtener un intervalo bayesiano estimado. Definamos a y b por

$$\int_{-\infty}^{a} f(\theta|x^n)d\theta = \int_{b}^{\infty} f(\theta|x^n)d\theta = \alpha/2.$$

Inferencia Bayesiana

Sea $C = (a, b)$, entonces

$$P(\theta \in C|x^n) = \int_a^b f(\theta|x^n)d\theta = 1 - \alpha.$$

de modo que C es un $1 - \alpha$ intervalo posteriori.

Ejemplo 1

Sea $X_1, \ldots, X_n \sim$Bernoulli(p). Suponga que tomamos la distribución uniforme $f(p) = 1$ como a priori. Por el Teorema de Bayes la posteriori tiene la forma

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s(1-p)^{n-s} = p^{s+1-1}(1-p)^{n-s+1-1},$$

donde $s = \sum_i x_i$ es el número de aciertos.

Inferencia Bayesiana

Recordemos que una variable aleatoria tiene distribución Beta con parámetros α y β si su densidad es

$$f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

Vemos que la posteriori para p es una distribución Beta con parámetros $s+1$ y $n-s+1$. Esto es

$$f(p|x^n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} p^{(s+1)-1}(1-p)^{(n-s+1)-1}.$$

Finalmente, escribimos esta distribución como

$$p|x^n \sim Beta(s+1, n-s+1).$$

Inferencia Bayesiana

Note que hemos obtenido la constante de normalización sin necesidad de calcular la integral $\int \mathcal{L}_n(p) f(p) dp$. La media de una $Beta(\alpha, \beta)$ es $\alpha/(\alpha+\beta)$, de modo que el estimador de Bayes es

$$\bar{p} = \frac{s+1}{n+2}.$$

Es útil reescribir el estimador como

$$\bar{p} = \lambda_n \hat{p} + (1-\lambda_n)\tilde{p},$$

donde

$$\begin{aligned} \hat{p} &= s/n, \text{ es el estimador de máxima verosimilitud,} \\ \tilde{p} &= 1/2, \text{ es la media a priori y} \\ \lambda_n &= n/(n+2) \approx 1. \end{aligned}$$

Inferencia Bayesiana

Un intervalo posteriori del 95 % se puede obtener numéricamente, hallando a y b tal que

$$\int_a^b f(p|x^n)dp = 0.95.$$

Suppongamos ahora que en lugar de una uniforme para la distribución a priori, usamos la priori $p \sim Beta(\alpha, \beta)$. si repetimos los cálculos anteriores, vemos que

$$p|x^n \sim Beta(\alpha + s, \beta + n - s).$$

El caso anterior es justamente el caso especial $\alpha = \beta = 1$.

Inferencia Bayesiana

La media posteriori es

$$\bar{p} = \frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{n}{\alpha + \beta + n}\right)\hat{p} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)p_0,$$

donde $p_0 = \alpha/(\alpha + \beta)$ es la media a priori.

Inferencia Bayesiana

Ejemplo 2

Sea $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. Por simplicidad, asumamos σ conocido. Suponga que tomamos como a priori $\theta \sim N(a, b^2)$. La posteriori para θ es

$$\theta | X^n \sim N(a, b^2) \tag{4}$$

donde $\bar{\theta} = w\bar{X} + (1+w)a$, con

$$w = \frac{\frac{1}{se^2}}{\frac{1}{se^2} + \frac{1}{b^2}} \quad \text{y} \quad \frac{1}{\tau} = \frac{1}{se^2} + \frac{1}{b^2}$$

y $se = \sigma/\sqrt{n}$ es el error estándar del E.M.V de $\bar{X}$.

Inferencia Bayesiana

Veamos que en efecto, la posteriori es la mostrada

$$
\begin{aligned}
p(\theta|X) &\propto f(X|\theta)f(\theta) \\
&= \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(X_i-\theta)^2\right\}\right)\left(\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{b^2}} \exp\left\{-\frac{1}{2b^2}(\theta-a)^2\right\}\right) \\
&= \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2\right\} \frac{1}{(2\pi)^{1/2}} \frac{1}{(b^2)^{1/2}} \exp\left\{-\frac{1}{2b^2}(\theta-a)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2 - \frac{1}{2b^2}(\theta-a)^2\right\}
\end{aligned}
$$

Inferencia Bayesiana

Ahora

$$
\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2+\frac{1}{b^2}(\theta-a)^2 &= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}X_i^2-2n\theta\bar{X}+n\theta^2\right]+\frac{1}{b^2}\left[\theta^2-2\theta a+a^2\right] \\
&= -2\left[\frac{n\bar{X}}{\sigma^2}+\frac{a}{b^2}\right]\theta+\left[\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{1/2}\right]\theta^2+\square \\
&= \left(\sqrt{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right)^2\left[\theta^2-2\frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)}\left(\frac{n\bar{X}}{\sigma^2}+\frac{a}{b^2}\right)\theta\right]+\square \\
&= \left\{\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)^{-1}\right\}^{-1}\left\{\theta-\frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)}\left(\frac{n\bar{X}}{\sigma^2}+\frac{a}{b^2}\right)\right\}^2+\triangle \\
\Rightarrow \quad \theta|X &\sim N\left(\frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{b^2}\right)}\left(\frac{n\bar{X}}{\sigma^2}+\frac{a}{b^2}\right),\frac{1}{\frac{n}{\sigma^2}+\frac{1}{b^2}}\right).
\end{aligned}
$$

Inferencia Bayesiana

Note que $w \to 1$ y $\tau/se \to 1$ cuando $n \to \infty$. Así, para n grande, la posteriori es aproximadamente $N(\hat{\theta}, se^2)$. Lo mismo es cierto si n es fijo pero $b \to \infty$, lo cual corresponde a hacer que la priori se vuelva muy plana. Ahora hallemos $C = (c, d)$ tal que $P(\theta \in C|X^n) = 0.95$. Podemos escoger c tal que $P(\theta < c|X^n) = 0.025$ y d tal que $P(\theta > d|X^n) = 0.025$, de modo que queremos hallar c tal que

$$\begin{aligned} P(\theta < x|X^n) &= P\left(\frac{\theta - \bar{\theta}}{\tau} < \frac{c - \bar{\theta}}{\tau}|X^n\right) \\ &= P\left(Z < \frac{c - \bar{\theta}}{\tau}\right) = 0.025. \end{aligned}$$

Inferencia Bayesiana

Pero de probabilidades sabemos que $P(Z < -1.96) = 0.025$, así

$$\frac{c - \bar{\theta}}{\tau} = -1.96,$$

lo que implica que $c = \bar{\theta} - 1.96\tau$. Por analogía, $d = \bar{\theta} + 1.96\tau$. Así un intervalo bayesiano del 95 % es $\theta \pm 1.96\tau$. Dado que $\bar{\theta} \approx \hat{\theta}$ y $\tau \approx se$, entonces el intervalo bayesiano del 95 % es aproximado por $\hat{\theta} \pm 1.96se$, el cual es el intervalo de confianza frecuentista.

■

Funciones de parámetros

Funciones de parámetros

¿Cómo hacemos inferencia sobre una función $\tau = g(\theta)$?
De probabilidad sabemos, dada la densidad f_x de X obtener la densidad para $Y = g(X)$.
De manera similar, tenemos que la función de distribución acumulada posteriori para τ es

$$H(\tau|x^n) = P(g(\theta) \leq \tau) = \int_A f(\theta|x^n)d\theta,$$

donde $A = \{\theta : g(\theta) \leq \tau\}$. La densidad posteriori es

$$h(\tau|x^n) = H'(\tau|x^n).$$

Funciones de parámetros

Ejemplo 3

Sea $X_1, \ldots, X_n \sim$Bernoulli(p) y $f(p) = 1$, de modo que $p|X^n \sim Beta(s+1, n-s+1)$ con $s = \sum_{i=1}^{n} X_i$
Sea $\psi = \log(p/(1-p))$. Entonces

$$\begin{aligned} H(\psi|x^n) &= P(\Psi \leq \psi|x^n) = P\left(\log\left(\frac{p}{1-p}\right) \leq \psi|x^n\right) \\ &= P\left(p \leq \frac{e^\psi}{1+e^\psi}|x^n\right) = \int_0^{e^\psi/(1+e^\psi)} f(p|x^n)dp \\ &= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \int_0^{e^\psi/(1+e^\psi)} p^s(1-p)^{n-2}dp \end{aligned}$$

Funciones de parámetros

y

$$
\begin{aligned}
h(\psi|x^n) &= H'(\psi|x^n) \\
&= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{\partial(\frac{e^\psi}{1+e^\psi})}{\partial\psi}\right) \\
&= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s} \left(\frac{1}{1+e^\psi}\right)^2 \\
&= \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \left(\frac{e^\psi}{1+e^\psi}\right)^s \left(\frac{1}{1+e^\psi}\right)^{n-s+2}
\end{aligned}
$$

para $\psi \in \mathbb{R}$.

■

Simulación

Simulación

La posteriori a menudo se puede aproximar por simulación, supongamos que tenemos $\theta_1, \ldots, \theta_n \sim p(\theta|x^n)$. Entonces el histograma de $\theta_1, \ldots, \theta_B$ aproxima la densidad posteriori $p(\theta|x^n)$. Una aproximación de la media posteriori es

$$\bar{\theta}_n = \mathbb{E}(\theta|x^n) = \frac{1}{B}\sum_{j=1}^{B}\theta_j.$$

El intervalo posteriori $1-\alpha$ se puede aproximar por $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$, donde $\theta_{\alpha/2}$ es el $\alpha/2$ cuantil muestral de $\theta_1, \ldots, \theta_B$.
Una vez que tenemos una muestra $\theta_1, \ldots, \theta_B$ de $f(\theta|x^n)$, hacemos $\tau_i = g(\theta_i)$. Entonces $\tau_1, \ldots, \tau_B$ es la muestra de $f(\tau|x^n)$. Esto evita la necesidad de cualquier cálculo analítico.

Simulación

Ejemplo 4

Consideremos de nuevo el Ejemplo 3. Podemos aproximar la posteriori para ψ sin hacer ningún cálculo. Los pasos para ello son:

1. Simulamos $P_1, \ldots, P_B \sim Beta(s+1, n-s+1)$
2. Hacemos $\psi_i = \log(P_i/(1-P_i))$ para $i = 1, \ldots, B$. Se tiene que $\psi_1, \ldots, \psi_B$ son iid obtenidas de $h(\psi|x^n)$.
3. Realizamos un histograma de $\psi_1, \ldots, \psi_B$, este nos provee un estimado de $h(\psi|x^n)$.

Simulación

Ejemplo 5. dos Binomiales

Sean $X \sim Binomial(n_1, p_1)$ y $Y \sim Binomial(n_2, p_2)$. Queremos estimar $\delta = p_2 - p_1$.
El EMV es $\hat{\delta} = \hat{p}_2 - \hat{p}_1 = (Y/n_2) - (X/n_1)$.
El error estándar es

$$\hat{se} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Un intervalo del 95 % es $\hat{\delta} \pm 2\hat{se}$.

Simulación

Ahora consideremos el análisis bayesiano.
Suponga que usamos la priori $f(p_1, p_2) = f(p_1)f(p_2) = 1$, esto es, una priori plana sobre (p_1, p_2). La posteriori es

$$f(p_1, p_2|X, Y) \propto p_1^X (1 - p_1)^{n_1 - X} p_2^Y (1 - p_2)^{n_2 - Y}.$$

La media posteriori de δ es

$$\bar{\delta} = \int_0^1 \int_0^1 \delta(p_1, p_2) f(p_1, p_2|X, Y) = \int_0^1 \int_0^1 (p_2 - p_1) f(p_1, p_2|X, Y)$$

Si queremos la densidad posteriori de δ, debemos obtener primero la función de distribución acumulada

$$F(c|X, Y) = P(\delta \leq 0|X, Y) = \int_A f(p_1, p_2|X, Y),$$

Simulación

donde $A = \{(p_1, p_2) : p_2 - p_1 \leq c\}$, luego diferenciamos.
Para evitar todas estas integrales usamos simulación. Note que

$$f(p_1, p_2|X, Y) = f(p_1|X)f(p_2|Y),$$

lo que implica que p_1 y p_2 son independientes bajo la distribución posteriori.
También tenemos que

$$\begin{aligned} p_1|X &\sim Beta(X + 1, n_1 - X + 1) \\ p_2|Y &\sim Beta(Y + 1, n_2 - Y + 1) \end{aligned}$$

Simulación

Por consiguiente, podemos simular $(p_1^{(1)}, p_2^{(1)}), \ldots, (p_1^{(N)}, p_2^{(N)})$ de la posteriori, por medio de

$$\begin{aligned} p_1^{(i)} &\sim Beta(X+1, n_1 - X + 1) \\ p_2^{(i)} &\sim Beta(Y+1, n_2 - Y + 1) \end{aligned}$$

para $i = 1, \ldots, N$. Luego hacemos $\delta^{(i)} = p_2^{(i)} - p_1^{(i)}$, entonces

$$\bar{\delta} = \frac{1}{N} \sum_i \delta^{(i)}$$

Simulación

Supongamos $n_1 = n_2 = 10, X = 8, Y = 6$, el código en R es:

```
x<-8;y<-6
n1<-10;n2<-10
B<-10000
p1<-rbeta(B,x+1,n1-x+1)
p2<-rbeta(B,y+1,n2-y+1)
delta<-p2-p1
print(mean(delta))
left<-quantile(delta,0.025)
right<-quantile(delta,0.975)
print(c(left,right))
hist(delta, xlab = "delta=p2-p1", ylab = "f(delta|data)")
```

Propiedades de los procedimientos de Bayes para muestras grandes

Propiedades de los procedimientos de Bayes para muestras grandes

En los ejemplos de Bernoulli (Ejemplo 1) y Normal (Ejemplo 2) vimos que la media posteriori estaba cercana al estimador de máxima verosimilitud. Esto es cierto en general.

Teorema

Bajo condiciones apropiadas de regularidad, tenemos que la posteriori es aproximadamente $N(\hat{\theta}, \hat{se}^2)$, *donde* $\hat{\theta}_n$ *es el E.M.V y* $se = 1/\sqrt{nI(\hat{\theta}_n)}$. *Por consiguiente,* $\bar{\theta}_n \approx \hat{\theta}_n$. *También, si* $C = (\hat{\theta}_n - z_{\alpha/2}\hat{se}, \hat{\theta}_n + z_{\alpha/2}\hat{se})$ *es el intervalo de confianza asintótico frecuentista* $1-\alpha$, *entonces* C_n *es también un intervalo bayesiano posteriori* $1-\alpha$ *aproximado*

$$P(\theta \in C | X^n) \to 1 - \alpha.$$

Propiedades de los procedimientos de Bayes para muestras grandes

Demostración

Se puede demostrar que el efecto de la priori disminuye cuando n crece de modo que $f(\theta|x^n) \propto \mathcal{L}_n(\theta)f(\theta) \approx \mathcal{L}_n(\theta)$. Por consiguiente

$$\log f(\theta|x^n) \approx \ell(\theta)$$

Ahora,

$$\begin{aligned} \ell(\theta) &= \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + [(\theta - \hat{\theta})^2/2]\ell''(\hat{\theta}) \\ &= \ell(\hat{\theta}) + [(\theta - \hat{\theta})^2/2]\ell''(\hat{\theta}) \end{aligned}$$

dado que $\ell'(\hat{\theta}) = 0$. Tomando exponencial, obtenemos

Propiedades de los procedimientos de Bayes para muestras grandes

$$f(\theta|x^n) \propto \exp\left[-\frac{1}{2}\frac{(\theta-\hat{\theta})^2}{\sigma_n^2}\right],$$

donde $\sigma_n^2 = --1/\ell''(\hat{\theta}_n)$. Así, la posteriori de θ es aproximadamente normal con media $\hat{\theta}$ y varianza σ_n^2.
Sea $\ell_i = \log f(X_i|\theta)$, entonces

$$\begin{aligned}
\sigma_n^{-2} &= -\ell''(\hat{\theta}_n) = \sum_i -\ell_i''(\hat{\theta}_n) \\
&= n(\frac{1}{n})\sum_i -\ell_i''(\hat{\theta}_n) \approx n\mathbb{E}_\theta[-\ell_i''(\hat{\theta}_n)] \\
&= nI(\hat{\theta})
\end{aligned}$$

y por lo tanto $\sigma_n \approx se(\hat{\theta})$ ■

Propiedades de los procedimientos de Bayes para muestras grandes

Existe también un método delta bayesiano. Sea $\tau = g(\theta)$. Entonces

$$\tau | X^n \approx N(\hat{\tau}, \hat{se}^2),$$

donde $\hat{\tau} = g(\hat{\theta})$ y $\hat{se} = se|g'(\hat{\theta})|$

- El método delta nos proporciona una aproximación de los valores de τ y $Var\tau$, que se basa en una expansión en serie de Taylor de la función $g(\theta)$.
- El método delta da resultados *exactos* en el caso gaussiano/lineal.
- Si X no tiene una distribución gaussiana, es posible que $\mathbb{E}(X)$ y $Var(X)$ po sí mismos no proporcionen una descripción adecuada de $f_X(\theta)$

Prioris plana, prioris impropias y prioris no informativa

Prioris plana, prioris impropias y prioris no informativa

- Una pregunta importante en inferencia bayesiana es ¿De donde se obtiene $f(\theta)$ a priori?
- Una escuela de pensamiento, llamada *subjetivismo*, dice que la priori debe rechazar nuestra opinión subjetiva sobre θ antes de que se recojan los datos.
- Esto puede ser posible en algunos casos, pero impráctico en problemas complicados, especialmente si hay muchos parámetros.
- Una alternativa es tratar de definir alguna *priori no informativa*
- Un candidato obvio para una priori no informativa es usar una priori *plana* $f(\theta) \propto constante$.
- En el Ejemplo 1, tomamos $f(p) = 1$, lo que nos llevó a $p|X^n \sim Beta(s+1, n-s+1)$.

Prioris plana, prioris impropias y prioris no informativa

1) Prioris impropias

- Consideremos el ejemplo $N(\theta, 1)$.
- Supongamos que adoptamos una priori plana $f(\theta) \propto c$, donde $c > 0$ es constante.
- Note que $\int f(\theta) d\theta = \infty$, de modo que no es una densidad de probabilidad real en el sentido usual.
- Podemos llamar a tal priori una **priori impropia**
- Aún así, podemos usar el *Teorema de Bayes* y calcular la densidad posteriori $f(\theta) \propto \mathcal{L}_n(\theta) f(\theta) \propto \mathcal{L}_n(\theta)$.

Prioris plana, prioris impropias y prioris no informativa

- En el Ejemplo 2, (variable normal), esta dio $\theta|X^n \sim N(\bar{X}, \sigma^2/n)$
- Los estimadores puntual y por intervalo coinciden con sus contrapartes frecuentista.
- En general, las *priori impropia* no son un problema, siempre y cuando la posteriori resultante sea una distribución de probabilidad bien definida.

Prioris plana, prioris impropias y prioris no informativa

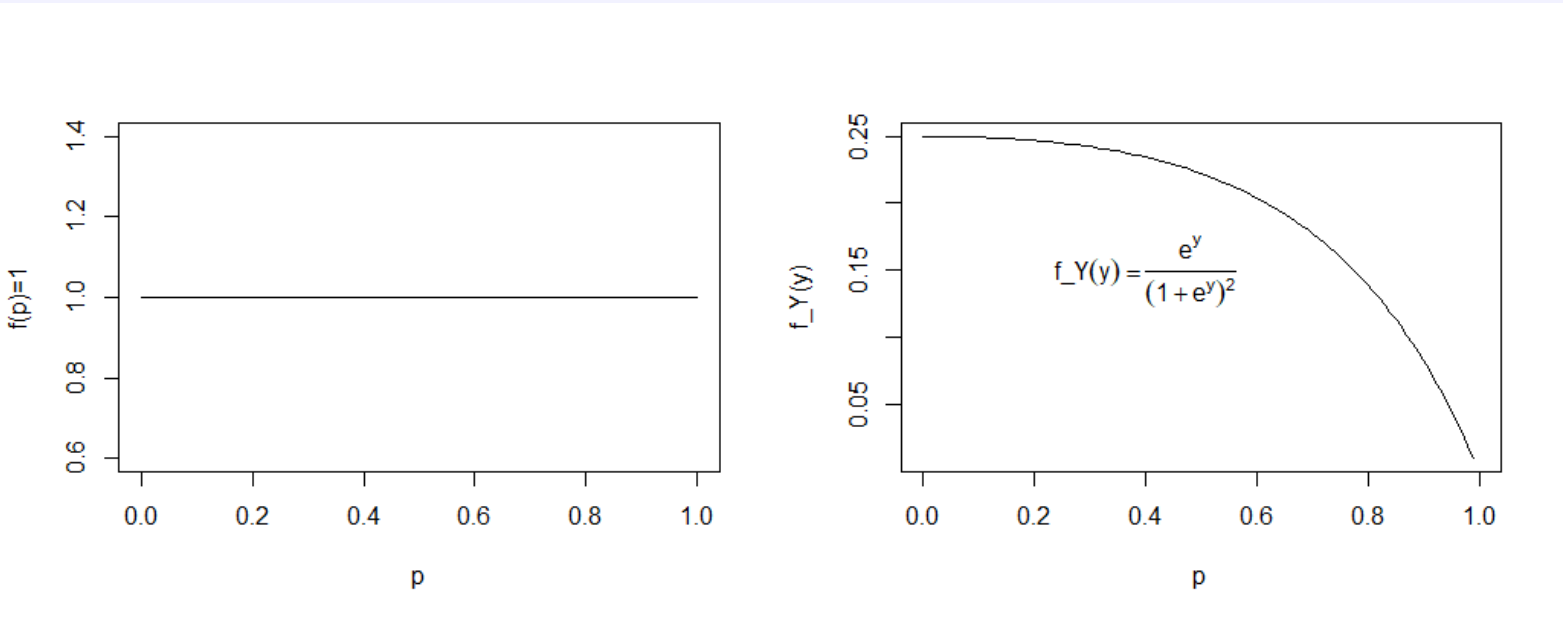
2) Priori planas no son invariantes

- Veamos de nuevo el ejemplo de la Bernoulli (Ejemplo 1) y consideremos una priori plana $f(p) = 1$.
- Recordemos que la priori plana presumiblemente representa nuestra falta de información sobre p antes del experimento.
- Sea $\psi = \log(p/(1-p))$
- Esto es una transformación sobre p y podemos calcular la distribución resultante para ψ.
- La misma es

$$f_\Psi(\psi) = \frac{e^\psi}{(1+e^\psi)^2}.$$

Prioris plana, prioris impropias y prioris no informativa

- ¿Podríamos argumentar que si no tenemos conocimiento sobre p, entonces tampoco tendríamos información sobre ψ y no deberíamos usar una priori plana para ψ?
- Esto contradice la priori $f_{\Psi}(\psi)$ para ψ que proviene de la priori plana para p.



Prioris plana, prioris impropias y prioris no informativa

- La noción de que una priori plana no está bien definida es porque una priori plana sobre un parámetro no implica una priori plana sobre una transformación del parámetro.
- Las priori planas no son **invariantes bajo transformaciones**.

Prioris plana, prioris impropias y prioris no informativa

3) Priori de Jeffreys

- Jeffreys ideó una *regla* para crear prioris
- La regla es: tome $f(\theta) \propto I(\theta)^{1/2}$, donde $I(\theta)$ es la función de información de Fisher.
- Esta regla resulta ser invariante bajo transformaciones.

Prioris plana, prioris impropias y prioris no informativa

Ejemplo 6

- Consideremos una $Bernoulli(p)$.
- Recordemos que

$$I(p) = \frac{1}{p(1-p)}$$

- La regla de Jeffreys dice que usemos la priori

$$f(p) \propto \sqrt{I(p)} = p^{-1/2}(1-p)^{-1/2}.$$

- Esta es una densidad $Beta(1/2, 1/2)$
- Esta densidad es muy cercana a una uniforme

Prioris plana, prioris impropias y prioris no informativa

En problemas de multiparámetros, la priori de Jeffreys se define como

$$f(\theta) \propto \sqrt{detI(\theta)},$$

donde $det(A)$ denota el determinante de la matriz A.

Problemas de multiparámetros

Problemas de multiparámetros

- En principio, los problemas de multiparámetros los podemos manejar de la misma manera
- Suponga que $\theta = (\theta_1, \ldots, \theta_p)$.
- La densidad posteriori es

$$p(\theta|X^n) \propto \mathcal{L}n(\theta)f(\theta).$$

- La pregunta ahora es: ¿cómo hacemos inferencia sobre un parámetro?
- La clave está en hallar la densidad marginal posteriori para el parámetro de interés.

Problemas de multiparámetros

- Suponga que queremos hacer inferencia sobre θ_1
- La marginal posteriori para θ_1 es

$$f(\theta_1|X^n) = \int \cdots \int f(\theta_1, \ldots, \theta_p|x^n) d\theta_2 \cdots d\theta_p.$$

- En la práctica, puede que no sea factible resolver esta integral.
- La simulación puede ser de gran ayuda.

Problemas de multiparámetros

- Tomamos una muestra aleatoria de la posteriori

$$\theta^1, \ldots, \theta^B \sim f(\theta|x^n),$$

donde el superíndice indica las diferentes muestras.

- Cada θ^j es un vector $(\theta_1^j, \ldots, \theta_p^j)$.
- Ahora, recopilamos la primera componente de cada muestra

$$\theta_1^1, \ldots, \theta_1^B.$$

- Estas son las muestras de $f(\theta_1|x^n)$ y así evitamos hacer integración.

Problemas de multiparámetros

Ejemplo 7. Comparación de dos binomiales

Suponga que tenemos n_1 pacientes de control y n_2 pacientes con tratamiento, y que X_1 pacientes de control sobreviven mientras que X_2 pacientes con tratamiento sobreviven.
Deseamos estimar $\tau = g(p_1, p_2) = p_2 - p_1$. Entonces

$$\begin{aligned} X_1 &\sim Binomial(n_1, p_1) \\ X_2 &\sim Binomial(n_2, p_2) \end{aligned}$$

Problemas de multiparámetros

Suponga que tomamos $f(p_1, p_2) = 1$. La posteriori es

$$f(p_1, p_2|x_1, x_2) \propto p_1^{x_1}(1-p_1)^{n_1-x_1} p_2^{x_2}(1-p_2)^{n_2-x_2}.$$

Note que (p_1, p_2) yacen en un rectángulo (de hecho en el cuadrado $[0,1] \times [0,1]$) y que

$$f(p_1, p_2|x_1, x_2) = f(p_1|x_1) f(p_2|x_2),$$

donde

$$\begin{aligned} f(p_1|x_1) &\propto p_1^{x_1}(1-p_1)^{n_1-x_1} \\ f(p_2|x_2) &\propto p_2^{x_2}(1-p_2)^{n_2-x_2} \end{aligned}$$

Problemas de multiparámetros

lo cual implica que p_1 y p_2 son independientes bajo la posteriori. Además

$$\begin{aligned} p_1|x_1 &\sim Beta(x_1+1, n_1-x_1+1) \\ p_2|x_2 &\sim Beta(x_2+1, n_2-x_2+1) \end{aligned}$$

Si simulamos

$$\begin{aligned} P_1^1, \ldots, P_1^B &\sim Beta(x_1+1, n_1-x_1+1) \\ P_2^1, \ldots, P_2^B &\sim Beta(x_2+1, n_2-x_2+1) \end{aligned}$$

entonces

$$\tau_b = P_{2,b} - P_{1,b}, \quad b = 1, \ldots, B,$$

es una muestra de $f(\tau|x_1, x_2)$ ■

Fortalezas y debilidades de la inferencia bayesiana

Fortalezas y debilidades de la inferencia bayesiana

- La inferencia bayesiana es atrayente cuando la información a priori está disponible dado que el teorema de Bayes es una forma natural de combinar información a priori con los datos.
- En modelos parámetricos, con muestras grandes, los métodos bayesianos y frecuentistas dan aproximadamente la misma inferencia.
- En general, no necesitan concordar.

Fortalezas y debilidades de la inferencia bayesiana

Ejemplo 8

Sea $X \sim N(\theta, 1)$ y suponga que usamos la priori $\theta \sim N(0, \tau^2)$. De (4) la posteriori es

$$\theta|x \sim N\left(\frac{x}{1+\frac{1}{\tau^2}}, \frac{1}{1+\frac{1}{\tau^2}}\right) = N(cx, c),$$

donde $c = \tau^2/(\tau^2+1)$. Un intervalo posteriori de $1-\alpha$ por ciento es $C = (a, b)$ donde

$$a = cx - \sqrt{c}z_{\alpha/2} \quad \text{y} \quad b = cx + \sqrt{c}z_{\alpha/2}.$$

Luego, $P(\theta \in C|X) = 1-\alpha$.

Fortalezas y debilidades de la inferencia bayesiana

Nos preguntamos ahora, desde la perspectiva frecuentista, ¿cuál es la cobertura de C, esto es, cuán frecuente este intervalo contendrá el valor real? La respuesta es

$$\begin{aligned}
P(a<\theta<b) &= P_\theta(cX - z_{\alpha/2}\sqrt{c} < \theta < cX + z_{\alpha/2}\sqrt{c}) \\
&= P_\theta\left(\frac{\theta - z_{\alpha/2}\sqrt{c} - c\theta}{c} < X - \theta < \frac{\theta + z_{\alpha/2}\sqrt{c} - c\theta}{c}\right) \\
&= P_\theta\left(\frac{\theta(1-c) - z_{\alpha/2}\sqrt{c}}{c} < Z < \frac{\theta(1-c) + z_{\alpha/2}\sqrt{c}}{c}\right) \\
&= \Phi\left(\frac{\theta(1-c) + z_{\alpha/2}\sqrt{c}}{c}\right) - \Phi\left(\frac{\theta(1-c) - z_{\alpha/2}\sqrt{c}}{c}\right),
\end{aligned}$$

donde $Z \sim N(0,1)$.

Fortalezas y debilidades de la inferencia bayesiana

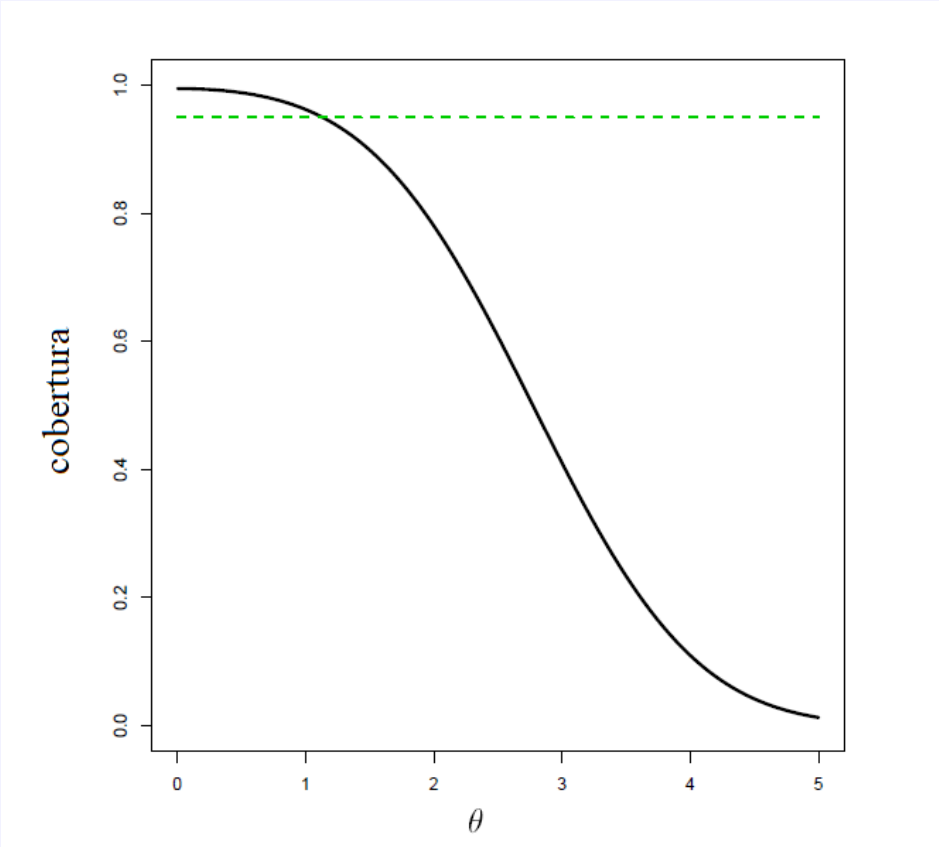


Figura: Cobertura frecuentista del 95 % del intervalo posterior bayesiano en función del valor real. La línea de puntos marca el nivel del 95 por ciento.

Fortalezas y debilidades de la inferencia bayesiana

A menos que el valor real de θ esté cerca de 0, la cobertura es muy pequeña. Así, tras un uso repetido el intervalo bayesiano 95 % puede contener el valor real con frecuencia cercana a 0. En contraste, un intervalo de confianza tiene cobertura del 95 % sin importar el valor real de θ. ■

Fortalezas y debilidades de la inferencia bayesiana

¿Qué concluimos de esto?

- Es importante entender que los métodos bayesianos y frecuentistas responden a diferentes preguntas.
- Si quiere combinar creencia a priori con datos use *inferencia bayesiana*.
- Si quiere construir procedimientos que garanticen rendimiento a largo plazo, tal como un intervalo de confianza, use *métodos frecuentistas*.
- Vale la pena destacar que es posible desarrollar métodos bayesianos no parámetricos similar a las estimaciones *plug-in* o *bootstrap*, pero hay que tener cuidado, ya que las propiedades frecuentistas de los métodos bayesianos no parámetricos pueden a veces ser pobres.

Jackknife y bootstrap

Estimación del error estadístico, donde error:

- sesgo de un estimador;
- error estándar de un estimador;
- tasa de error de una regla de predicción basada en datos.

En las técnicas que estudiaremos se empleará **poder computacional bruto en lugar de análisis teórico**.

Ejemplo: Supongamos que $X_1, \ldots, X_n$ son i.i.d. con distribución F y que observamos $X_1 = x_1, \ldots, X_n = x_n$. Calculamos $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ como un estimador de $E(X)$.

Jackknife y bootstrap

Los datos proveen, no solo el estimador $\overline{x}$, sino también un estimador de la precisión que tiene $\overline{x}$, a saber

$$\hat{\sigma}(\overline{x}) = \left(\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{1/2}.$$

Problema: Esta fórmula para la precisión de $\overline{x}$ no se extiende en forma obvia para otros estimadores, e.g. la mediana.

El bootstrap y jackknife pueden hacer esta generalización.

Definamos

$$\overline{x}_{(i)} = \frac{1}{n-1} \sum_{j \neq i} x_j = \frac{1}{n-1} \left(\sum_{j=1}^{n} x_j - x_i \right) = \frac{1}{n-1} \left(n\overline{x} - x_i \right).$$

Jackknife y bootstrap

Sea $x_{(\cdot)} = n^{-1}\sum_{i=1}^{n}\overline{x}_{(i)}$. El estimador de Jackknife de la desviación estándar de $\overline{x}$ es

$$\hat{\sigma}_J = \left(\frac{n-1}{n}\sum_{i=1}^{n}(\overline{x}_{(i)} - \overline{x}_{(\cdot)})^2\right)^{1/2}.$$

La vetaja del estimador anterior es que se generaliza a cualquier estimador $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. Lo único que hay que hacer es sustituir $\overline{x}_{(i)}$ por $\hat{\theta}_{(i)} = \hat{\theta}(X_1, \ldots X_{n-1}, X_{n+1}, \ldots, X_n)$ y $x_{(\cdot)}$ por $\hat{\theta}_{(\cdot)} = n^{-1}\sum_{i=1}^{n}\hat{\theta}_{(i)}$y ya está.

El bootstrap generaliza al estimador de la desviación estándar de $\overline{x}$ de una forma aparentemente diferente.

Jackknife y bootstrap

Sea $\hat{F}$ la función de distribución empírica de la cual pone una masa de $1/n$ en cada x_i. Sea $X_1^*, \ldots, X_n^*$ una muestra i.i.d de $\hat{F}$, esto es X_i^* es seleccionada (simulada) independientemente con reemplazo de $\{X_1, \ldots, X_n\}$.

Entonces,

$$\overline{X}^* = \sum_{i=1}^{n} X_i^*$$

tiene varianza

$$\text{Var.}\overline{X} = \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \overline{x}).$$

(Hacer cálculo en el pizarrón).

Jackknife y bootstrap

El estimador bootstrap del error estándar de un estimador $\hat{\theta}(X_1, \ldots, X_n)$ es

$$\hat{\sigma}_B = (\text{Var}.\hat{\theta}(X_1^*, \ldots, X_n^*))^{1/2}$$

Más adelante vemos a lo que le llamamos propiamente bootstrap.

Jackknife

En estadística, Jackknife es una técnica de remuestreo que es especialmente útil para estimar varianzas y sesgos.

El estimador Jackknife de un parámetro se encuentra al sistemáticamente dejar fuera cada una de las observaciones de un conjunto de datos y calcular el estimado y al final encontrar el promedio de estos cálculos.

Dada una muestra de tamaño n, el estimador Jackknife se encuentra al sumar los estimados de cada sub-muestra de tamaño $(n-1)$.

Jackknife es anterior a otras técnicas de remuestreo como bootstrap.

La técnica Jackknife fue desarrollada por Quenouille (1949) y expandida por John Tukey (1958), quién propuso el nombre Jackknife.

Jackknife es una aproximación lineal al bootstrap.

Jackknife

Recapítulando lo que habíamos dicho antes del Jackknife, supongamos que $X_1, \ldots, X_n$ son i.i.d. con distribución F y que observamos $X_1 = x_1, \ldots, X_n = x_n$. Calculamos $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$ como un estimador de $E(X)$.

Tenemos una medida de la precisión del estimador $\overline{x}$ dada por

$$\hat{\sigma}(\overline{x}) = \left(\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{1/2}.$$

Problema: Esta fórmula para la precisión de $\overline{x}$ no se extiende en forma obvia para otros estimadores, e.g. la mediana.

Jackknife

Mediante el método de Jackknife habíamos dicho que podemos hallar un estimador de la desviación estándar de $\overline{x}$ mediante

$$\hat{\sigma}_J = \left(\frac{n-1}{n} \sum_{i=1}^{n} (\overline{x}_{(i)} - \overline{x}_{(\cdot)})^2 \right)^{1/2}.$$

El estimador se generaliza a cualquier estimador $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ al reemplazar $\overline{x}_{(i)}$ por $\hat{\theta}_{(i)} = \hat{\theta}(X_1, \ldots X_{n-1}, X_{n+1}, \ldots, X_n)$ y $\overline{x}_{(\cdot)}$ por $\hat{\theta}_{(\cdot)} = n^{-1} \sum_{i=1}^{n} \hat{\theta}_{(i)}$.

A continuación veremos esto con más cuidado.

Jackknife

Sea $T_n = T(X_1, \ldots, X_n)$ un estimador de alguna cantidad θ y denotemos su sesgo por bias$(T_n) = E(T_n) - \theta$.

Llamaremos a la submuestra

$$X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$$

la i-**ésima muestra Jackknife** de $(X_1, \ldots, X_n)$.

Denotemos por $T_{(-i)}$ al estadístico T_n con la i-ésima observación removida, es decir

$$T_{(-i)} = T(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

Llamaremos a $T_{(-i)}$ la i-**ésima replica de Jackknife**. En otras palabras, $T_{(-i)}$ es el resultado de evaluar la i-ésima muestra Jackknife en el estadístico.

Jackknife

Para ejemplificar los conceptos anteriores, consideremos los siguientes datos de Manly (2007)

3.56,	0.69,	0.10,	1.84,	3.93,	1.25,	0.18,	1.13,	0.27,
0.50,	0.67,	0.01,	0.61,	0.82,	1.70,	0.39,	0.11,	1.20,
1.21,	0.72.							

Se desea estimar la desviación estándar de la población, i.e. $\theta = \sigma$.

- Los datos tienen 20 valores. Manly (2007) estima la raíz cuadrada del MLE de la varianza poniendo n en lugar de $n-1$ en el denominador, es decir

$$\hat{\theta} = \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = 1.03,$$

y no con la desviación estándar muestral la cual tiene $(n-1) = 19$ grados de libertad.

Jackknife

- Los renglones en la tabla, marcados del 1 al 20, tienen un valor removido dejando los 19 valores restantes.
- La i-ésima replica de Jackknife es

$$\hat{\sigma}_i = \sqrt{\frac{1}{19}\sum_{i \neq j}(x_i - \overline{x})^2},$$

la cual es calculada de los 19 valores (sin x_i) en la i-ésima muestra de Jackknife. Los valores están dados en la primera columna de SD en la tabla.

Más adelante, usaremos Jackknife para estimar σ de los datos anteriores.

Jackknife

El estimador Jackknife del sesgo se define como

$$b_{Jack} = (n-1)(\overline{T}_n - T_n)$$

donde $\hat{T}_{(\cdot)} = \overline{T}_n = n^{-1} \sum_i T_{(-i)}$. El estimador corregido del sesgo es $T_{Jack} = T_n - b_{Jack}$.

Jackknife

¿Por qué se define b_{Jack} de esta manera? Para muchos estadísticos se puede mostrar que

$$\text{bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right)$$

para algunos a y b.

Por ejemplo, sea $\sigma^2 = V(X_i)$ y $\hat{\sigma}^2 = n^{-1}\sum_i (X_i - \overline{X})^2$. Entonces,

$$\text{bias}(\hat{\sigma}^2) = -\frac{\sigma^2}{n},$$

es decir que $\hat{\sigma}^2$ tiene la forma anterior con $a = -\sigma^2$ y $b = 0$.

Jackknife

Cuando el sesgo tiene la forma descrita anteriormente, se cumple que

$$\text{bias}(T_{(-i)}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

Se sigue que el sesgo de $\overline{T}_n$ también tiene dicha forma. Por tanto,

$$\begin{aligned} E(b_{Jack}) &= (n-1)(\overline{T}_n - T_n) \\ &= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right) \\ &= \text{bias}(T_n) + O\left(\frac{1}{n^2}\right) \end{aligned}$$

lo cual muestra que el b_{Jack} estima el sesgo hasta un orden de $O(n^{-2})$.

Jackknife

Por un cálculo similar,

$$\text{bias}(T_{Jack}) = \frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right)$$

así que el sesgo de T_{Jack} es un orden de magnitud menor que el de T_n. T_{Jack} también puede escribirse como

$$T_{Jack} = \frac{1}{n}\sum_{i=1}^{n} \tilde{T}_i$$

donde

$$\tilde{T}_i = nT_n - (n-1)T_{(-i)}$$

son los llamados pseudo-valores.

Jackknife

El estimador Jackknife de $V(T_n)$ es

$$v_{Jack} = \frac{\tilde{s}^2}{n}$$

donde

$$\tilde{s}^2 = \frac{\sum_{i=1}^{n}(\tilde{T}_i - \frac{1}{n}\sum_{j=1}^{n}\tilde{T}_j)^2}{n-1}$$

es la varianza muestral de los pesudo-valores. Bajo condiciones adecuadas en T, se puede demostrar que v_{Jack} estima consitentemente $V(T_n)$.

Por ejemplo, si T es una función suave de la media muestral entonces la consistencia se cumple.

Jackknife

Teorema

Sea $\mu = E(X_1)$ *y* $\sigma^2 = V(X_1) < \infty$ *y supongamos que* $T_n = g(\overline{X}_n)$ *donde* g *tiene una derivada continua no-nula en* μ. *Entonces*

$$\frac{(T_n - g(\mu))}{\sigma_n} \xrightarrow{d} N(0,1)$$

donde $\sigma_n^2 = n^{-1}(g'(\mu))^2\sigma^2$. *El Jackknife es consistente significando*

$$\frac{v_{Jack}}{\sigma_n^2} \xrightarrow{c.s.} 1.$$

Jackknife

Consideremos nuevamente los datos

3.56, 0.69, 0.10, 1.84, 3.93, 1.25, 0.18, 1.13, 0.27,
0.50, 0.67, 0.01, 0.61, 0.82, 1.70, 0.39, 0.11, 1.20,
1.21, 0.72.

El objetivo era calcular la desviación estandar σ de la población. Así que en este ejemplo,

- El estimador sin corregir $\hat{\theta} = \hat{\sigma} = 1.03285 \approx 1.03$.
- Las replicas de Jackknife $(\sigma_{(-i)})$ de las 20 submuestras aparecen en la columna SD.
- Los 20 pseudo-valores $\tilde{\sigma}_i$ aparecen en la columna PVi.
- El promedio de los pseudo-valores $\sigma_{Jack} = 1.09616 \approx 1.096$ es el estimado de Jackknife corregido por el sesgo.
- Por lo tanto, $b_{Jack} = \widehat{\text{bias}}(\hat{\sigma}) \approx (19)(1.02952 - 1.03285) = -.06327$ (ligero error de aproximación).
- O puedes estimar el sesgo usando $\widehat{\text{bias}}(\hat{\sigma}) = \hat{\theta} - \hat{\theta}_{Jack} \approx 1.03285 - 1.09616 \approx -.06331$ (más exacto).

Jackknife

Ejercicio: Supongamos que queremos aplicar Jackknife en otras 3 situaciones:

1. Supongamos que queremos estimar $\theta = \mu$. ¿Qué pasaría si usamos la media muestral ($\hat{\theta} = \overline{x}$)?
2. Supongamos que queremos estimar $\theta = \sigma$. ¿Qué pasaría la desviación estándar muestral ($\hat{\theta} = s$)?
3. Supongamos que queremos estimar $\theta = \sigma^2$. ¿Qué pasaría si usamos la varianza muestral estándar ($\hat{\theta} = s^2$)?

Jackknife

Hay algunas cosas importantes que observar en estos resultados,

1. Sabemos que $\overline{x}$ y s^2 son estimadores insesgados de μ y σ. Aplicar el método de Jackknife a un estimador insesgado no tiene efecto. Es decir, $\hat{\theta}_{Jack} = \hat{\theta}$ y el sesgo estimado será 0.
2. Ambos estimados de σ corregidos por el sesgo siguen siendo sesgados, pero el sesgo se reduce comparado a los estimadores originales. Después de la corrección del sesgo, los dos valores de $\hat{\theta}_{Jack}$ para estimar σ son muy cercanos.
3. Los errores estándar pueden usarse para calcular intervalos de confianza

$$\hat{\theta}_{Jack} \pm 1.960 \cdot \text{se}(\hat{\theta}_{Jack}).$$

Observación:

$$\hat{\theta}_{Jack} \pm t^* \cdot \text{se}(\hat{\theta}_{Jack})$$

donde t^* es el valor crítico $1 - \alpha/2$ de una distribución t con $n - 1$ grados de libertad.

Limitaciones del jackknife

- El método de Jackknife puede fallar si el estadístico $\hat{\theta}$ no es suave. La suavidad implica que cambios relativamente pequeñosa los datos provocarán sólo un pequeño cambio en el estadístico.
- La mediana muestra es un ejemplo de un estadístico que no es suave.
- Por ejemplo, volviendo a los datos de Manly (2007), los valores ordenados son

0.01	0.10	0.11	0.18	0.27	0.39	0.50	0.61
0.67	0.69	0.72	0.82	1.13	1.20	1.21	1.25
1.70	1.84	3.56	3.93				

Notemos que hay dos estimadores Jackknife:

- 1/2 de los estimadores de Jackknife son iguales a 0.72 (al borrar alguno de los 10 primeros valores).
- 1/2 de los estimadores de Jackknife son iguales a 0.69 (al borrar alguno de los 10 últimos valores).

Limitaciones del jackknife

- Por lo tanto, el método de Jackknife no es bueno para estimar percentiles (como la mediana), o cuándo se usa un estimador no suave.
- Esto no será el caso cuando usemos el método de estimación bootstrap.

Jackknife

Teorema

Si $T(F) = F^{-1}(p)$ es el p-ésimo cuantil, entonces el estimador Jackknife de la varianza es inconsistente. Para la mediana ($p = 1/2$) tenemos que

$$v_{Jack}/\sigma_n^2 \xrightarrow{d} (\chi_2^2/2)^2$$

donde σ_n^2 es la varianza asintótica de la mediana muestral.

Ejemplo: Sea $T_n = \overline{X}_n$. Es fácil ver que $\tilde{T}_i = X_i$. Por tanto $T_{Jack} = T_n$, $b = 0$ y $v_{Jack} = S_n^2/n$ donde S_n^2 es la varianza muestral.

Bootstrap

Hay una conexión entre Jackknife y la función de influencia.

Recordemos que la función de influencia es

$$L_F(x) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}.$$

Supongamos que aproximamos $L_F(X_i)$ poniendo $F = \hat{F}_n$ y $\epsilon = -1/(n-1)$. Esto da lugar a la aproximación

Bootstrap

Considera una m.a. $X_1, \ldots, X_n$ de alguna población y un estadístico $\hat{\theta}$ calculado a partir de la muestra que es el estimador de algún parámetro θ de la población. Por ejemplo, θ puede ser la media de la población y $\hat{\theta}$ la media muestral.

Para poder entender que tan bien $\hat{\theta}$ estima θ, es necesario conocer la varianza de la v.a. $\hat{\theta}$. Sin embargo, para poder calcularla, uno necesita saber algo de la población.

Existe un método general para estimar la varianza de $\hat{\theta}$ a partir de la muestra. Este método es conocido como **bootstrap**.

La idea de bootstrap es tratar a la muestra como una nueva población.

Bootstrap

Pensemos a nuestra m.a. $X_1, \ldots, X_n$ de v.a.i.i.d. como una **población finita** y consideremos el experimento de **tomar una m.a. con reemplazo y ordenada de tamaño n de esta población finita**. Denotaremos esta nueva muestra por $X_1^*, \ldots, X_n^*$ y la llamaremos muestra bootstrap.

Así, $X_1^*, \ldots, X_n^*$ son v.a. independientes, cada una de las cuales puede tomar el valor X_i con probabilidad $1/n$. Es decir,

$$P(X_i^* = X_j) = \frac{1}{n}.$$

A partir de la muestra bootstrap podemos calcular la v.a. correspondiente $\hat{\theta}^*$.

Ejemplo: Si θ es la media de la población y $\hat{\theta}$ la media muestral de la m.a. $X_1, \ldots, X_n$, entonces $\hat{\theta}^*$ es la media muestral de $X_1^*, \ldots, X_n^*$.

Bootstrap

La idea es usar la varianza de $\hat{\theta}^*$ (con $X_1, \ldots, X_n$ fija) como un estimador de la varianza de $\hat{\theta}$. Esta varianza es llamada el **estimador bootstrap ideal**.

¿Cómo podemos calcular el estimador bootstrap ideal?

En principio, es simple. Existe un número finito n^n de muestras bootstrap, cada una con probabilidad $1/n^n$. Podríamos hacer lo siguiente:

1. Para cada una de tales muestras bootstrap, calcula el valor de $\hat{\theta}^*$.
2. Calcula la media de estos n^n números.
3. Entonces calcula la media del cuadrado de las desviaciones de su media. Este número es el estimador bootstrap ideal.

El problema con esto es, por supuesto, que n^n es un número enorme. Enumerar todas las muestras bootstrap de esta forma es impráctico, incluso para n's pequeños.

Bootstrap

El bootstrap es un método para estimar la varianza y la distribución de un estadístico $T_n = g(X_1, \ldots, X_n)$.

También se utiliza al bootstrap para construir intervalos de confianza.

Denotemos por $V_F(T_n)$ a la varianza de T_n. Añadimos el subíndice F para enfatizar que la varianza es una función de la función de ditribución F.

Si conocemos F podríamos, al menos en principio, calcular la varianza. Por ejemplo, si $T_n = n^{-1} \sum_i X_i$, entonces

$$V_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - (\int x dF(x))^2}{n},$$

la cuál claramente está en función de F.

Bootstrap

Denotemos por $\hat{F}_n$ a la distribución empírica de la m.a. $X_1, \ldots, X_n$.

Con bootstrap estimamos $V_F(T_n)$ a tráves de $V_{\hat{F}_n}(T_n)$. En otras palabras, utilizamos un estimador plug-in de la varianza.

Como $V_{\hat{F}_n}(T_n)$ puede ser muy difícil de calcular, lo aproximamos con un estimado simulado denotado por v_{boot}. Específicamente, seguimos los siguientes pasos.

Bootstrap

Algoritmo (Estimación de la varianza por bootstrap)

1. *Obten una muestra* $X_1^*, \ldots, X_n^* \sim \hat{F}_n$.
2. *Calcula* $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. *Repite los pasos 1 y 2, B veces para obtener* $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. *Pongamos*

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2$$

Por la Ley de Grandes Números,

$$v_{boot} \xrightarrow{c.s.} V_{\hat{F}_n}(T_n)$$

cuando $B \to \infty$.

El error estándar estimado de T_n es $\hat{se}_{boot} = \sqrt{v_{boot}}$.

Bootstrap

El siguiente diagrama ilustra la idea del bootstrap:

$$\begin{array}{lccccc} \text{Mundo real:} & F & \Rightarrow & X_1, \ldots, X_n & \Rightarrow & T_n = g(X_1, \ldots, X_n) \\ & n \Uparrow & & & & B \Uparrow \\ \text{Mundo bootstrap:} & \hat{F}_n & \Rightarrow & X_1^*, \ldots, X_n^* & \Rightarrow & T_n^* = g(X_1^*, \ldots, X_n^*) \end{array}$$

$$V_{F_n}(T_n) \overset{O(1/\sqrt{n})}{\approx} V_{\hat{F}_n}(T_n) \overset{O(1/\sqrt{B})}{\approx} v_{boot}.$$

¿Cómo simulamos de $\hat{F}_n$?

Bootstrap

Ya que $\hat{F}_n$ da probabilidad $1/n$ a cada observación de los datos originales, muestrar aleatoriamente n observaciones de $\hat{F}_n$ es lo mismo que obtener una muestra con reemplazamiento de tamaño n a partir de los datos originales.

Por lo tanto, el paso 1 del algoritmo anterior puede reemplazarse por

1. Obten una muestra $X_1^*, \ldots, X_n^*$ con reemplazamiento de $X_1, \ldots, X_n$.

Bootstrap puede usarse para aproximar la función de distribución acumulada de un estadístico T_n.

Sea $G_n(t) = P(T_n \leq t)$ la función de distribución de T_n. La aproximación bootstrap a G_n está dada por

$$\tilde{G}_n^*(t) = \frac{1}{B} \sum_{b=1}^{B} I(T_{n,b}^* \leq t).$$

Bootstrap

Hasta ahora, hemos estimado F de forma no-paramétrica. Existe también un bootstrap paramétrico. Si F_θ depende de un parámetro θ y $\hat{\theta}$ es un estimado de θ, entonces podemos simplemente muestrear de $F_{\hat{\theta}}$ en lugar de F_n.

Esto es igual de exacto, pero mucho más simple que el método delta.

Bootstrap

Existen diversas maneras de contruir intervalos de confianza bootstrap. Varían en la facilidad de calcular y en la exactitud.

Entre estas formas encontrarmos:

- Intervalo normal.
- Intervalo pivotal.
- Intervalo pivotal studentizado.
- Intervalo basado en percentiles.

Bootstrap

Intervalo normal. El intervalo bootstrap más sencillo es el intervalo normal

$$T_n \pm z_{\alpha/2} \hat{\text{se}}_{boot}$$

donde $\hat{\text{se}}_{boot}$ es el estimador bootstrap del error estándar.

Este intervalo no es muy excto a menos de que la distribución de T_n sea cercana a la normal.

Bootstrap

Intervalo pivotal. Sea $\theta = T(F)$ y $\hat{\theta}_n = T(\hat{F}_n)$ y definamos el pivote $R_n = \hat{\theta}_n - \hat{\theta}$.

Denotemos por $H(r)$ a la función de distribución del pivote R_n,

$$H(r) = P_F(R_n \leq r).$$

Sea $C_n^* = (a, b)$ donde

$$a = \hat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2}), \qquad b = \hat{\theta}_n - H^{-1}(\frac{\alpha}{2});$$

así

$$\begin{aligned} P(a \leq \theta \leq b) &= P(\hat{\theta}_n - b \leq R_n \leq \hat{\theta}_n - a) = H(\hat{\theta}_n - a) - H(\hat{\theta}_n - b) \\ &= H(H^{-1}(1 - \frac{\alpha}{2})) - H(H^{-1}(\frac{\alpha}{2})) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \end{aligned}$$

Bootstrap

Por tanto C_n^* es un intervalo de confianza exacto para θ, de confiabilidad $1-\alpha$. Desafortunadamente, a y b dependen de la distribución desconocida H.

Sin embargo, podemos tomar el estimado bootstrap de H

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(R_{n,b}^* \leq r),$$

donde $R_{n,b}^* = \hat{\theta}_{n,b}^* - \hat{\theta}_n$.

Denotemos por r_β^* al cuantil muestral β de $(R_{n,1}^*, \ldots, R_{n,B}^*)$ y por θ_β^* al cuantil muestral β de $(\theta_{n,1}^*, \ldots, \theta_{n,B}^*)$. Notemos que $r_\beta^* = \theta_\beta^* - \hat{\theta}_n$.

Bootstrap

Se sigue que un intervalo de confianza aproximado de $1-\alpha$ para θ es $C_n = (\hat{a}, \hat{b})$ donde

$$\hat{a} = \hat{\theta}_n - H^{-1}(1-\frac{\alpha}{2}) = \hat{\theta}_n - r^*_{1-\alpha/2} = 2\hat{\theta}_n - \theta^*_{1-\alpha/2},$$

$$\hat{b} = \hat{\theta}_n - H^{-1}(\frac{\alpha}{2}) = \hat{\theta}_n - r^*_{\alpha/2} = 2\hat{\theta}_n - \theta^*_{\alpha/2}.$$

En resumen, el intervalo de confianza bootstrap pivotal es

$$C_n = (2\hat{\theta}_n - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta}_n - \hat{\theta}^*_{\alpha/2}).$$

Típicamente este es un intervalo de confianza puntual y asintótico.

Bootstrap

Intervalo pivotal studentizado. Existe una versión diferente del intervalo pivotal que tiene algunas ventajas. Sean

$$Z_n = \frac{T_n - \theta}{\hat{\text{se}}_{boot}}$$

y

$$Z^*_{n,b} = \frac{T^*_{n,b} - T_n}{\hat{\text{se}}^*_b}$$

donde $\hat{\text{se}}^*_b$ es un estimador del error estándar de $T^*_{n,b}$ y no de T_n.

Ahora aplicamos el mismo razonamiento que en el intervalo pivotal. Los cuantiles muestrales del bootstrap $Z^*_{n,1}, \ldots, Z^*_{n,B}$ deberían de aproximar los verdadero cuantiles de la distribución de Z_n. Denotemos por z^*_α al cuantil muestral α de $Z^*_{n,1}, \ldots, Z^*_{n,B}$, entonces $P(Z_n \leq z^*_\alpha) \approx \alpha$.

Bootstrap

Sea $C_n = (T_n - z^*_{1-\alpha/2}\hat{\text{se}}_{boot}, T_n - z^*_{\alpha/2}\hat{\text{se}}_{boot})$, entonces

$$\begin{aligned} P(\theta \in C_n) &= P(T_n - z^*_{1-\alpha/2}\hat{\text{se}}_{boot} \leq \theta \leq T_n - z^*_{\alpha/2}\hat{\text{se}}_{boot}) \\ &= P(z^*_{\alpha/2} \leq \frac{T_n - \theta}{\hat{\text{se}}_{boot}} \leq z^*_{1-\alpha/2}) \\ &= P(z^*_{\alpha/2} \leq Z_n \leq z^*_{1-\alpha/2}) \\ &\approx \alpha. \end{aligned}$$

Este intervalo tiene mayor exactitud que todos los intervalos discutidos hasta ahora pero tiene un detalle: es necesario calcular $\hat{\text{se}}^*_b$ para cada muestra bootstrap. Esto puede requerir hacer un segundo bootstrap dentro de cada bootstrap.

Bootstrap

En resumen, el intervalo bootstrap pivotal studentizado está dado por

$$(T_n - z^*_{1-\alpha/2}\hat{\text{se}}_{boot}, T_n - z^*_{\alpha/2}\hat{\text{se}}_{boot}),$$

donde z^*_β es cuantil β de $Z^*_{n,1}, \ldots, Z^*_{n,B}$ y

$$Z^*_{n,b} = \frac{T^*_{n,b} - T_n}{\hat{\text{se}}^*_b}.$$

Intervalos basados en percentiles. El intervalo bootstrap de percentiles esta definido por

$$(T^*_{\alpha/2}, T^*_{1-\alpha/2}).$$

Bootstrap

Let us now compare the accuracy of the different confidence interval methods. Consideremos un intervalo de confianza 'one sided' de $1-\alpha$ dado por $[\hat{\theta}_\alpha, \infty)$. Nos gustaría que $P(\theta \leq \hat{\theta}) = \alpha$, pero usualmente esto solo se cumple de forma aproximada.

Si $P(\theta \leq \hat{\theta}) = \alpha + O(n^{-1/2})$ entonces decimos que el intervalo es exacto de primer orden. Si $P(\theta \leq \hat{\theta}) = \alpha + O(n^{-1})$ entonces decimos que el intervalo es exacto de segundo orden.

Método	Exactitud
Intervalo normal	exacto de primer orden
Intervalo pivotal	exacto de primer orden
Intervalo basado en percentiles	exacto de primer orden
Intervalo pivotal studentizado	exacto de segundo orden

Bootstrap

Bajo ciertas condiciones de regularidad se puede demostrar la validez del bootstrap. Aquí no las estudiaremos pero se debe de tener en cuenta que el bootstrap no debe de aplicarse ciegamente.

También se puede demostrar que el estimador bootstrap de la varianza es consistente bajo algunas condiciones de T. En general, las condiciones de consistencia para el bootstrap son más débiles que las necesarias para el jackknife. Por ejemplo, el estimador bootstrap para la varianza de la mediana es consitente, pero el estimador jackknife de la varianza de la mediana no lo es.

Anexo: Potencia de la prueba y Estadística Bayesiana.

Poder de la prueba

Poder de una prueba

- Consideremos la hipótesis simple H_0, con estadístico de prueba D. El p-valor asociado al valor observado $d = D(x)$ es

$$p\text{-valor} = P(D(x) \geq d \mid H_0)$$

- Si existe un valor d_x tal que $P(D \geq d_x|H_0) = \alpha$ entonces, se dice que α es un nivel se significancia alcanzable.
- Sea $P = \{p \mid p = P(D \geq d_x|H_0),$ para alguna $d_x = D(x)\}$. Esto es, P es el conjunto de p-valores, o niveles e significancia alcanzables.
- Dos estadísticos de prueba son comparables, si tienen el mismo conjunto de niveles de significancia alcanzables.
- Una región crítica de tamaño α para una prueba, es el conjunto, C_α, de resultados, x, para los cuales el nivel de significancia es menor o igual a α.

$$x \in C_\alpha \quad \Leftrightarrow \quad D(x) \geq d_x$$

Poder de una prueba

- Dado un tamaño α de una prueba, el **poder** de la prueba ante una alternativa H_1 es la probabilidad de haber tomado la decisión correcta al rechazar H_0 al nivel de significancia α.

$$K_\alpha = P(p\text{-valor} \leq \alpha | H_1) = P(X \in C_\alpha | H_1)$$

- Sean D y D' dos estadísticos de prueba comparables, con poderes K_α y K'_α respectivamente. Se dice que D es más poderoso que D' si $K_\alpha \geq K'_\alpha$ para todos los niveles alcanzables α.
- Se dice que D es el estadístico de prueba más poderoso para contrastar H_0 contra H_1, si D es más poderoso que cualquier otro estadístico comparable D'.

Ejemplo 12.10.2. Comparación de medias 1

- Consideremos un ejemplo básico. Supongamos $X \sim N(\mu, 1)$ y queremos decidir entre

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu = \mu_1$$

- Una adaptación del estadístico cociente de verosimilitudes es

$$D(x) = \frac{f_1(x)}{f_0(x)} = \exp\left((\mu_1 - \mu_0)x + \frac{1}{2}(\mu_0^2 - \mu_1^2)\right)$$

- Supongamos $\mu_1 > \mu_0$, entonces rechazamos H_0 si $D(x)$ es grande, mayor que cierta cantidad C.
- $D(x)$ es monótona creciente, entonces $D(x) \geq C$ si, y sólo si, $x \geq k$.

Ejemplo 12.10.2. Comparación de medias 2

- La región de rechazo o región crítica de tamaño α, es

$$C_\alpha = \{x | x \geq k\}$$

donde k es tal que $P(X \geq k | H_0) = \alpha$.

- Puede verse que

$$k = \mu_0 + z_\alpha$$

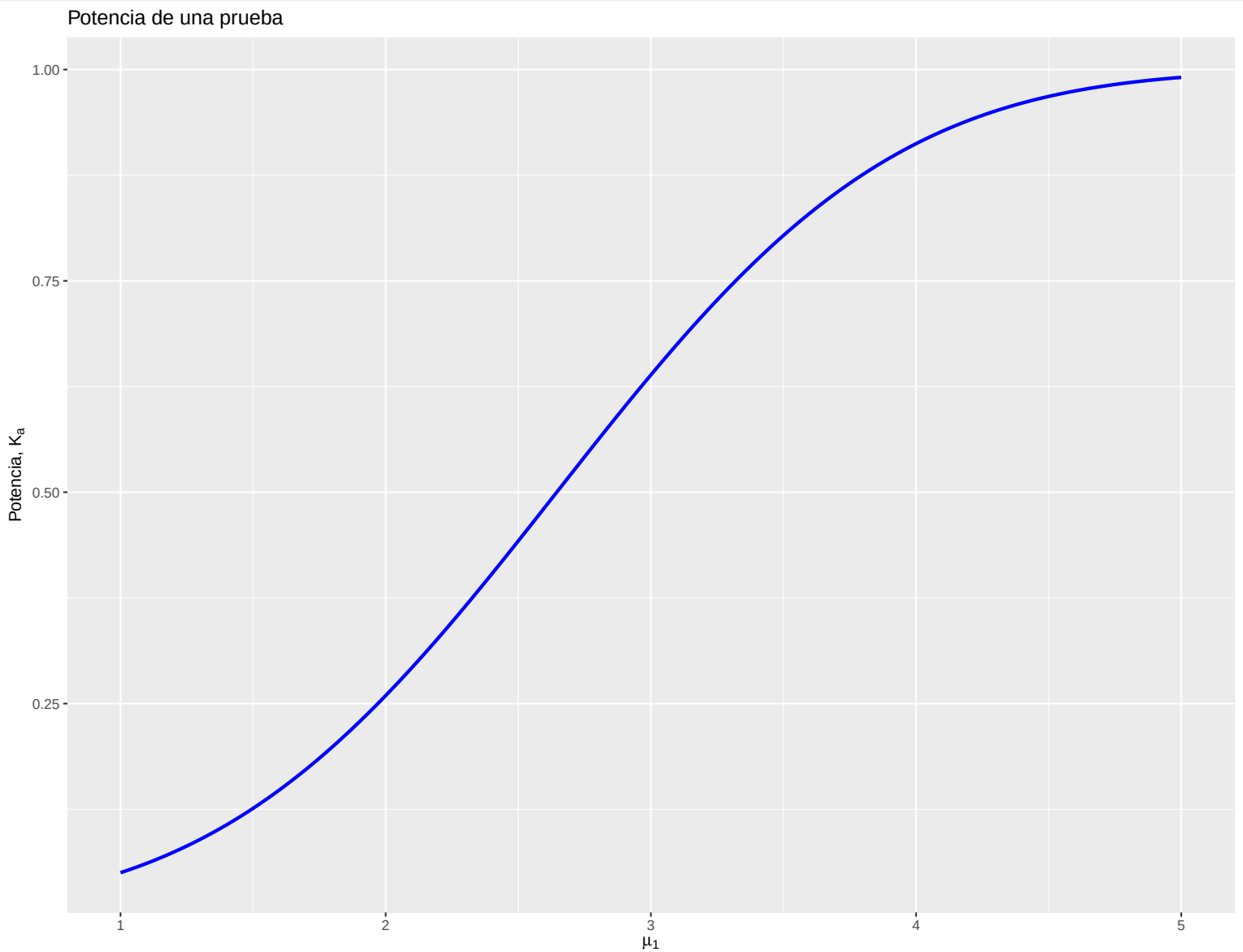
donde z_α es el cuantil en la cola derecha de la normal estándar.

- El poder correspondiente es

$$K_\alpha = P(X \geq k | H_1) = 1 - \Phi(k - \mu_1) = 1 - \Phi(z_\alpha - (\mu_1 - \mu_0))$$

mostramos una visualización en la siguiente gráfica.

Potencia de una prueba



Potencia

```
library(ggplot2)
M    <- 201
alfa <- .05
zalf <- qnorm( 1 - alfa )
mu0  <- 1
mu1  <- seq(1,5,length.out = M)
Ka   <- 1 - pnorm( zalf - (mu1-mu0) )
plot(mu1,Ka,type="l")

df   <- data.frame(mu1=mu1,Ka=Ka)

ggplot() +
  geom_line(aes(x=mu1,y=Ka), data=df, size=1.05, col="blue") +
  xlab(expression(mu[1])) +
  ylab( expression(paste("Potencia, ",K[a] )) ) +
  ggtitle("Potencia de una prueba")
```

Neyman-Pearson 1

- **Teorema**: Sean H_0 y H_1 hipótesis simples, y sean f_0 y f_1 los modelos de probabilidad respectivos. Entonces

$$D(x) = \frac{f_1(x)}{f_0(x)}$$

  da la prueba más poderosa para probar H_0 versus H_1.
- Sea α un nivel alcanzable para D y sea d_x el valor de D tal que

$$P(D \geq d_x | H_0) = \alpha$$

- La región crítica de tamaño α es

$$C_\alpha = \{x | D(x) \geq d_x\}$$

- Sea D' otro estadístico de prueba, comparable a D y sea C'_α la región crítica de tamaño α de D'

Neyman-Pearson 2

- Particionemos el espacio muestral S como

$$S = (C_\alpha \cap C'_\alpha) \cup (C_\alpha \cap \bar{C}'_\alpha) \cup (\bar{C}_\alpha \cap C'_\alpha) \cup (\bar{C}_\alpha \cap \bar{C}'_\alpha)$$

- Definamos las probabilidades de cada uno de los cuatro subconjuntos, bajo H_0 y bajo H_1

Bajo H_0

	C'_α	$\bar{C}'_\alpha$	
C_α	p_{11}	p_{12}	α
$\bar{C}_\alpha$	p_{21}	p_{22}	$1-\alpha$
	α	$1-\alpha$	

Bajo H_1

	C'_α	$\bar{C}'_\alpha$	
C_α	q_{11}	q_{12}	K_α
$\bar{C}_\alpha$	q_{21}	q_{22}	$1-K_\alpha$
	K'_α	$1-K'_\alpha$	

Neyman-Pearson 3

- Recuerde que el poder de D es la probabilidad de decisión correcta de rechazar H_0 dado que H_1 es cierta

$$K_\alpha = P(C_\alpha | H_1)$$

- Como C_α y C'_α son ambas de tamaño α

$$p_{11} + p_{12} = \alpha = p_{11} + p_{21} \quad \Rightarrow \quad p_{12} = p_{21}$$

- La diferencia de potencias entre D y D' es

$$K_\alpha - K'_\alpha = (q_{11} + q_{12}) - (q_{11} + q_{21}) = q_{12} - q_{21}$$

Neyman-Pearson 4

- Ahora
$$q_{12} = P(C_\alpha \cap \bar{C}'_\alpha | H_1) = \int I(C_\alpha \cap \bar{C}'_\alpha) f_1(x) dx$$
- por otro lado, si $x \in C_\alpha$, $D(x) \geq d_x$ y $f_1(x) \geq d_x f_0(x)$
- entonces
$$q_{12} \geq d_x \int I(C_\alpha \cap \bar{C}'_\alpha) f_0(x) dx = d_x P(C_\alpha \cap \bar{C}'_\alpha | H_0) = d_x p_{12}$$
- Similarmemte
$$q_{21} < d_x p_{21}$$
- Entonces
$$K_\alpha - K'_\alpha = q_{12} - q_{21} > d_x p_{12} - d_x p_{21} = 0$$

pues $p_{12} = p_{21}$, luego $K_\alpha > K'_\alpha$. Esto es válido para todo estadístico D' comparable a D y para todos los niveles de significancia alcanzables $\square$

Ejemplo 12.10.2 (extendido) 1

- Supongamos que $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, con σ^2 conocida, y queremos decidir entre

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu = \mu_1$$

- El resultado de Neyman-Pearson nos dice que consideremos, denotando $x = (x_1, \ldots, x_n)$

$$\begin{aligned} D(x) = \frac{f_1(x)}{f_0(x)} &= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_1)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n [\,(x_i - \mu_1)^2 - (x_i - \mu_0)^2\,]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n [\,(-2x_i(\mu_1 - \mu_0) + \mu_1^2 - \mu_0^2\,]\right\} \\ &= \exp\left\{\frac{n}{\sigma^2}(\mu_1 - \mu_0)\bar{x} - \frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2)\right\} \end{aligned}$$

Ejemplo 12.10.2 (extendido) 2

- Supongamos $\mu_1 > \mu_0$, entonces rechazamos H_0 si $D(x)$ es grande, mayor que cierta cantidad C

$$D(x) = \exp\left\{\frac{n}{\sigma^2}(\mu_1 - \mu_0)\bar{x} - \frac{n}{2\sigma^2}(\mu_1^2 - \mu_0^2)\right\} \geq C$$

- Con $\mu_1 > \mu_0$, $D(x)$ es monótona creciente en $\bar{x}$, entonces $D(x) \geq C$ si, y sólo si, $\bar{x} \geq k$.
- La región de rechazo o región crítica de tamaño α, es

$$C_\alpha = \{x | \bar{x} \geq k\}$$

donde k es tal que $P(\bar{X} \geq k | H_0) = \alpha$.

- Recuerde que

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

Ejemplo 12.10.2 (extendido) 3

- El valor críctico k se elige tal que

$$P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \geq \frac{k-\mu_0}{\sigma/\sqrt{n}} \,\middle|\, H_0\right) = P\left(Z \geq \frac{k-\mu_0}{\sigma/\sqrt{n}}\right) = \alpha$$

- de la última igualdad se obtiene

$$\frac{k-\mu_0}{\sigma/\sqrt{n}} = z_\alpha$$

$$k = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

donde z_α es el cuantil en la cola derecha de la normal estándar.

- La regla de decisión, de tamaño α, es: Rechazar $H_0 : \mu = \mu_0$ si

$$\bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

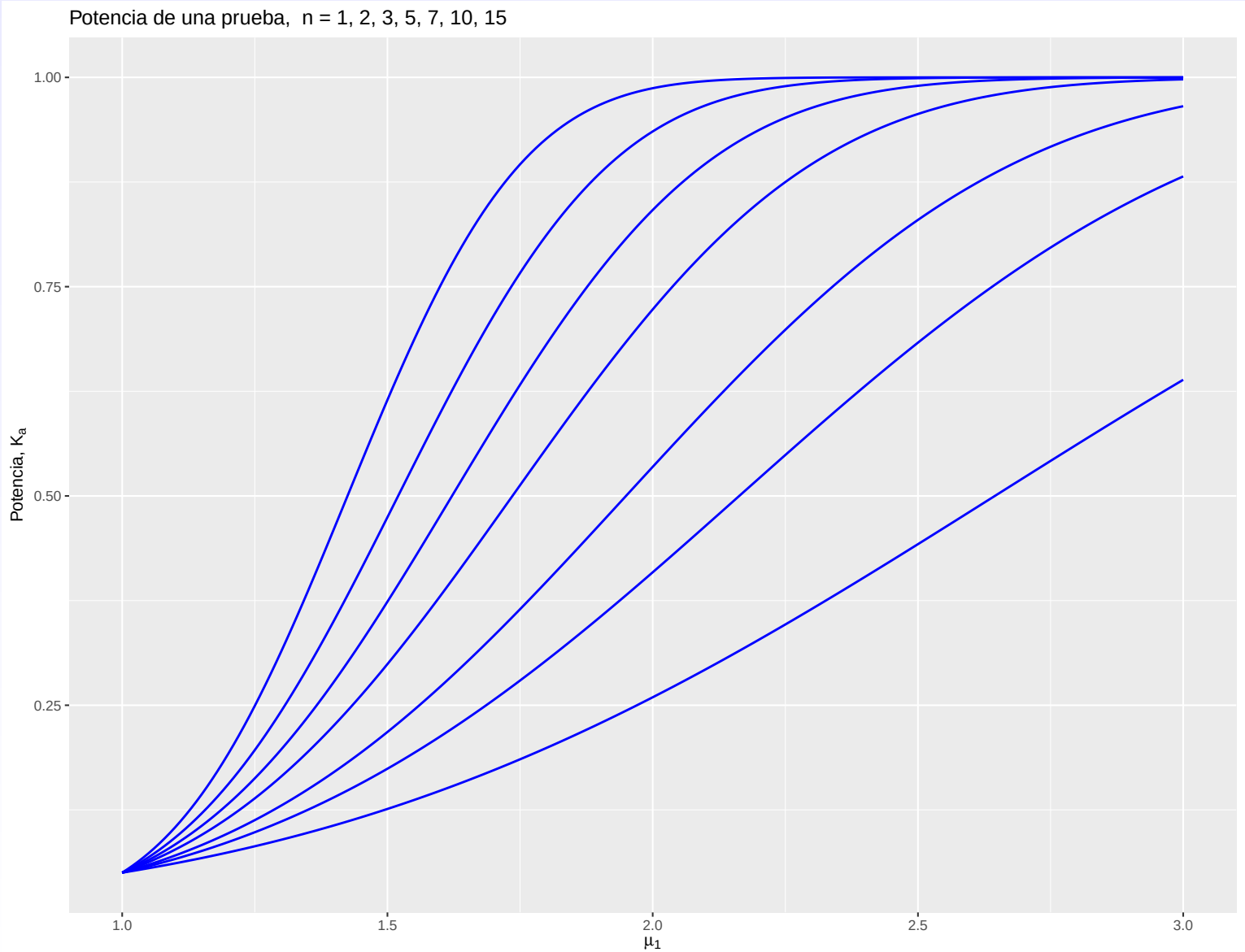
Ejemplo 12.10.2 (extendido) 4

- El poder correspondiente es

$$\begin{aligned} K_\alpha &= P\left(\bar{X} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \,\middle|\, H_1\right) \\ &= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \geq z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \,\middle|\, H_1\right) \\ &= P\left(Z \geq z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

- En la siguiente gráfica mostramos el caso $\mu_0 = 1$ y diferentes valores de μ_1, así como el efecto del tamaño de muestra.

Potencia de una prueba



Potencia

```
library(ggplot2)
library(ggpubr)
library(tidyverse)

alfa <- .05
zalf <- qnorm( 1 - alfa )
mu0  <- 1
sig  <- 1
nes  <- c(1,2,3,5,7,10,15)
N    <- length(nes)
power_curves <- tibble(Group = 1:N, nes)
plot_data <- pmap_df(power_curves,
          function(Group, nes) {
            tibble(Group = Group,
                   x = seq(1,3,length.out = 201),
                   y = 1 - pnorm( zalf - (x-mu0)/(sig/sqrt(nes)) ))})
graf <- ggplot(data = plot_data) +
          xlab( expression(mu[1]) ) +
          ylab( expression(paste("Potencia, ",K[a] )) ) +
          geom_line(aes(group = Group, x = x, y = y), colour="blue",
            size=.7, data=plot_data) +
          ggtitle("Potencia de una prueba,  n = 1, 2, 3, 5, 7, 10, 15")
plot(graf)
```

Ejemplo 12.10.2 (extendido) 5

- ¿Qué pasa si usamos otro estadístico de prueba?
- Observo $X_1, \ldots, X_n$ y me fijo en el **mínimo** de ellos, $X_{(1)}$.
- Suponga que decidimos rechazar $H_0 : \mu = \mu_0$, en favor de $H_1 : \mu = \mu_1$, donde $\mu_1 > \mu_0$, si $X_{(1)} > C$ para cierta constante C.
- ¿Qué valor de C hace que esta prueba sea de tamaño α?

$$
\begin{aligned}
P(\, X_{(1)} \geq C \mid H_0 \,) &= \alpha \\
P(\, X_{(1)} \leq C \mid H_0 \,) &= 1 - \alpha \\
1 - (\, 1 - P(\, X \leq C \mid H_0 \,) \,)^n &= 1 - \alpha \\
(\, 1 - P(\, Z \leq (C - \mu_0)/\sigma \,) \,)^n &= \alpha \\
P(\, Z \leq (C - \mu_0)/\sigma \,) &= 1 - \alpha^{1/n}
\end{aligned}
$$

- Entonces el punto crítico es $C = \mu_0 + z_{\alpha^{1/n}}\sigma$

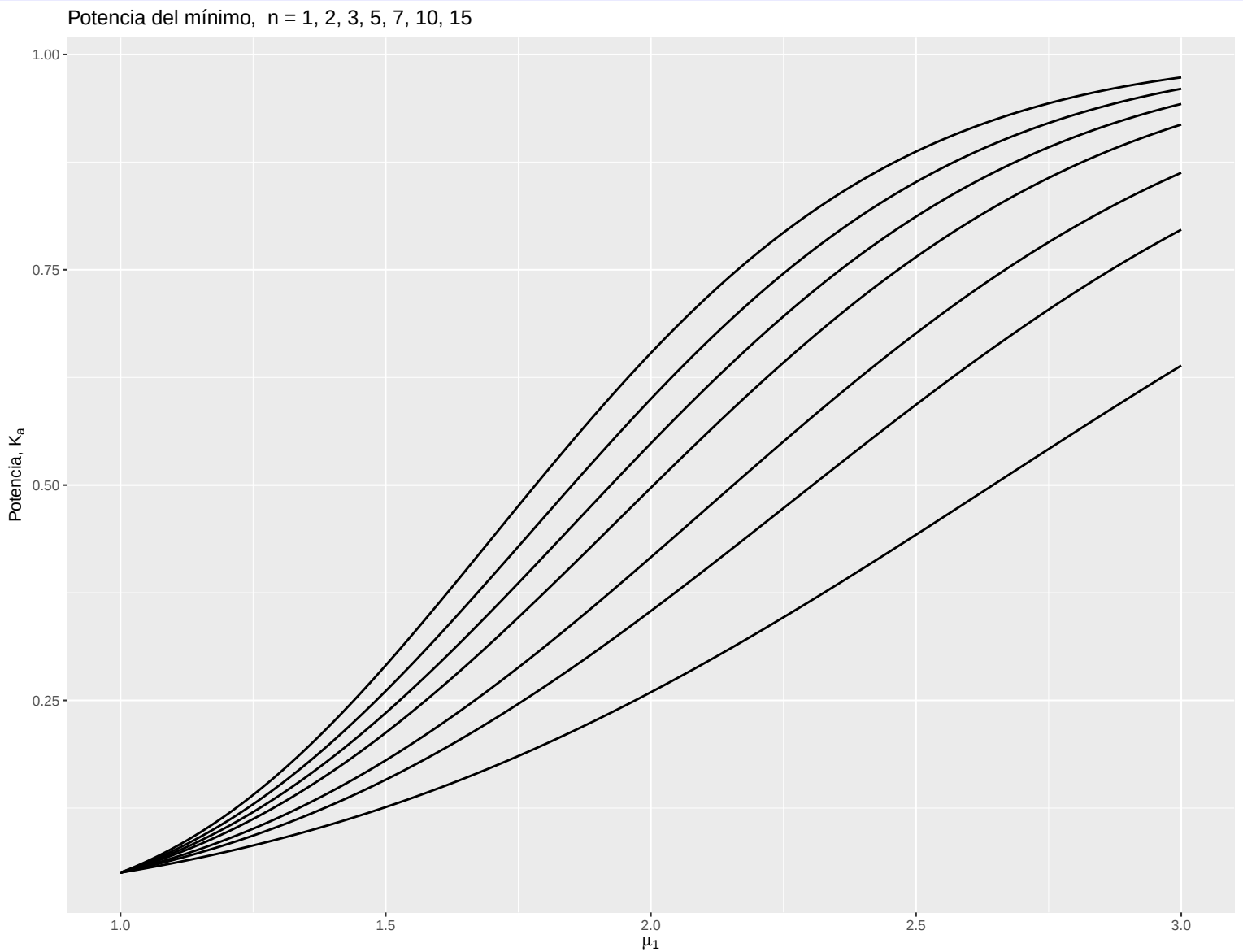
Ejemplo 12.10.2 (extendido) 6

- Ahora, la potencia, usando el mínimo:

$$\begin{aligned} K_\alpha &= P(\, X_{(1)} \geq C \mid H_1 \,) \\ &= 1 - P(\, X_{(1)} \leq C \mid H_1 \,) \\ &= 1 - [\, 1 - (\, 1 - P(\, X \leq C \mid H_1 \,) \,)^n \,] \\ &= (\, 1 - P(\, Z \leq (C - \mu_1)/\sigma \,) \,)^n \\ &= (\, 1 - P(\, Z \leq z_{\alpha^{1/n}} - (\mu_1 - \mu_0)/\sigma \,) \,)^n \\ K_\alpha &= (\, 1 - \Phi(\, z_{\alpha^{1/n}} - (\mu_1 - \mu_0)/\sigma \,) \,)^n \end{aligned}$$

- En la siguiente gráfica, mostarmos la curva de potencia. Observamos que estas curvas de potencia estan por abajo que las correspondientes a las de la prueba óptima, como lo asegura el teorema.

Potencia de la prueba con el mínimo

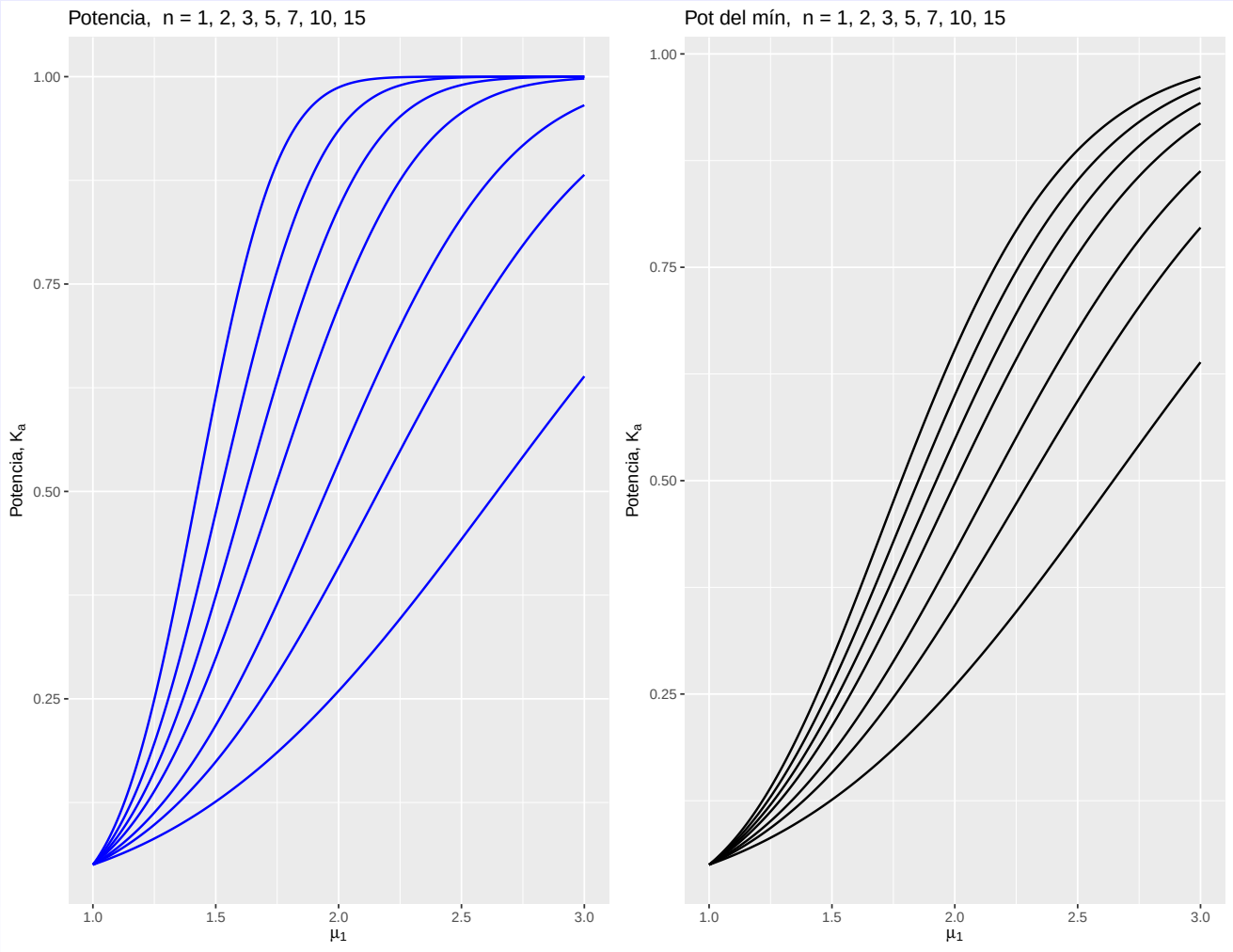


Potencia de la prueba con el mínimo

```
library(ggplot2)
library(ggpubr)
library(tidyverse)

alfa <- .05
zalf <- qnorm( 1 - alfa )
mu0  <- 1
sig  <- 1
nes  <- c(1,2,3,5,7,10,15)
N    <- length(nes)
power_curves2 <- tibble(Group = 1:N, nes)
plot_data2 <- pmap_df(power_curves2,
          function(Group, nes) {
            tibble(Group = Group,
             x = seq(1,3,length.out = 201),
             y =(1-pnorm(qnorm(1-alfa^(1/nes))-(x-mu0)/sig))^nes)})
graf2 <- ggplot(data = plot_data2) +
          xlab( expression(mu[1]) ) +
          ylab( expression(paste("Potencia, ",K[a] )) ) +
          geom_line(aes(group = Group, x = x, y = y),colour="black",
                    size=.7, data=plot_data2) +
          ggtitle( "Potencia del mínimo,  n = 1, 2, 3, 5, 7, 10, 15")
plot(graf2)
```

Comparación de potencias




```
gg      <- ggarrange(graf,graf2, ncol = 2, nrow = 1)
plot(gg)
```

Ejemplo 12.10.2 (extendido) 7

- Los cálculos anteriores usaron la distribución del mínimo. Aquí la recordamos.
- Sean $X_1, \ldots, X_n$ i.i.d. $F(x)$. Sea $X_{(1)} = \min\{X_1, \ldots, X_n\}$ y denotemos por $F_m(x)$ a la función de distribución del mínimo.
- Entonces

$$\begin{aligned} F_m(x) &= P(X_{(1)} \leq x) = 1 - P(X_{(1)} \geq x) \\ &= 1 - P(X_1 \geq x, \ldots, X_n \geq x) \\ &= 1 - \prod_{i=1}^{n} P(X_i \geq x) \\ &= 1 - \prod_{i=1}^{n} [1 - P(X_i \leq x)] \\ &= 1 - [\, 1 - P(X \leq x) \,]^n = 1 - [\, 1 - F(x) \,]^n \end{aligned}$$

Inferencia Bayesiana

Inferencia Bayesiana

- Dada una muestra aleatoria $x_1, \ldots, x_n$ de una población $f(x;\theta)$, los métodos de inferencia clásicos se basan en propiedades de la verosimilitud

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- El parámetro θ es una cantidad desconocida de interés. Toda la información previa o conocimiento que se tenga del fenómeno en cuestión, es reflejada en la verosimilitud.
- En inferencia Bayesiana, aunado a lo anterior, se conceptualiza que el desconocimiento o incertidumbre que se tenga sobre θ debe ser cuantificado mediante una distribución de probabilidad.
- El conocimiento, o desconocimiento, se representa mediante la distribución previa $\pi(\theta)$.

Inferencia Bayesiana

- Dada una muestra aleatoria $x = (x_1, \ldots, x_n)$ de una población $f(x;\theta)$, la inferencia Bayesiana se basa en las propiedades de la distribución posterior

$$\pi(\theta|x) = \frac{h(\theta, x)}{m(x)} = \frac{\pi(\theta)L(x|\theta)}{m(x)}$$

- donde

$$\begin{aligned} \pi(\theta|x) &= \text{ distribución posterior} \\ \pi(\theta) &= \text{ distribución previa o apriori} \\ L(x|\theta) &= \text{ verosimilitud} \\ m(x) &= \text{ distribución marginal de } x \end{aligned}$$

Inferencia Bayesiana

- Una forma de hacer estimación puntual es usar el MAP (Maximum aposteriori); esto es,
$$\widehat{\theta} = \arg\max_{\theta} \pi(\theta|x)$$
- Una forma de construir intervalos de confianza es mediante la región HPD (high posterior density) tal que
$$HPD = \{\theta \mid \pi(\theta|x) \geq k\}$$
donde k es tal que $P(HPD) = 1 - \alpha$
- Una forma de probar hipótesis es: Si quiero contrastar H_0 versus H_1, entonces la decisión es elegir la hipótesis con la probabilidad posterior más alta.

Ejemplo 16.2.2 (1)

- Inferencia para el caso binomial. Supongamos que tenemos una observación x de una distribución Binomial.
- Si lo anterior es lo único que sabemos, la inferencia se basa en la verosimilitud

$$L(\theta; x) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$$

- Si nuestro conocimiento previo sobre θ lo podemos reflejar en una previa, $\pi(\theta)$ entonces usamos la posterior $\pi(\theta|x)$.
- Supongamos una previa del tipo Beta

$$\pi(\theta) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \theta^{p-1}(1-\theta)^{q-1}$$

Ejemplo 16.2.2 (2)

- ¿Por qué una Beta?
  - Por conveniencia (conjugación)
  - Por que es sensata: La Beta es una distribución flexible que puede representar muchos conocimientos previos sobre θ
- Los métodos Bayesianos empíricos usan datos previos para definir distribuciones previas razonables
- La posterior es

$$\begin{aligned}\pi(\theta|x) &\propto \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\theta^{p-1}(1-\theta)^{q-1}\binom{n}{x}\theta^{x}(1-\theta)^{n-x} \\ &\propto \theta^{x+p-1}(1-\theta)^{n-x+q-1}\end{aligned}$$

esto es, la posterior es Beta con parámetros $x+p$ y $n-x+q$

Ejemplo 16.2.2 (3)

- Supongamos $x = 7$ y $n = 10$. El máximo verosímil es $\widehat{\theta} = \frac{7}{10}$. ¿Quién es el MAP si usamos una previa $Beta(p = 3, q = 2)$?
- Puede verse que la moda de una $Beta(\alpha, \beta)$ es $(\alpha - 1)/(\alpha + \beta - 2)$ (para $\alpha > 1$ y $\beta > 1$). Entonces

$$MAP = \frac{x + p - 1}{x + p + n - x + q - 2} = \frac{7 + 3 - 1}{7 + 3 + 10 - 7 + 2 - 2}$$
$$= \frac{10}{13} = .77$$

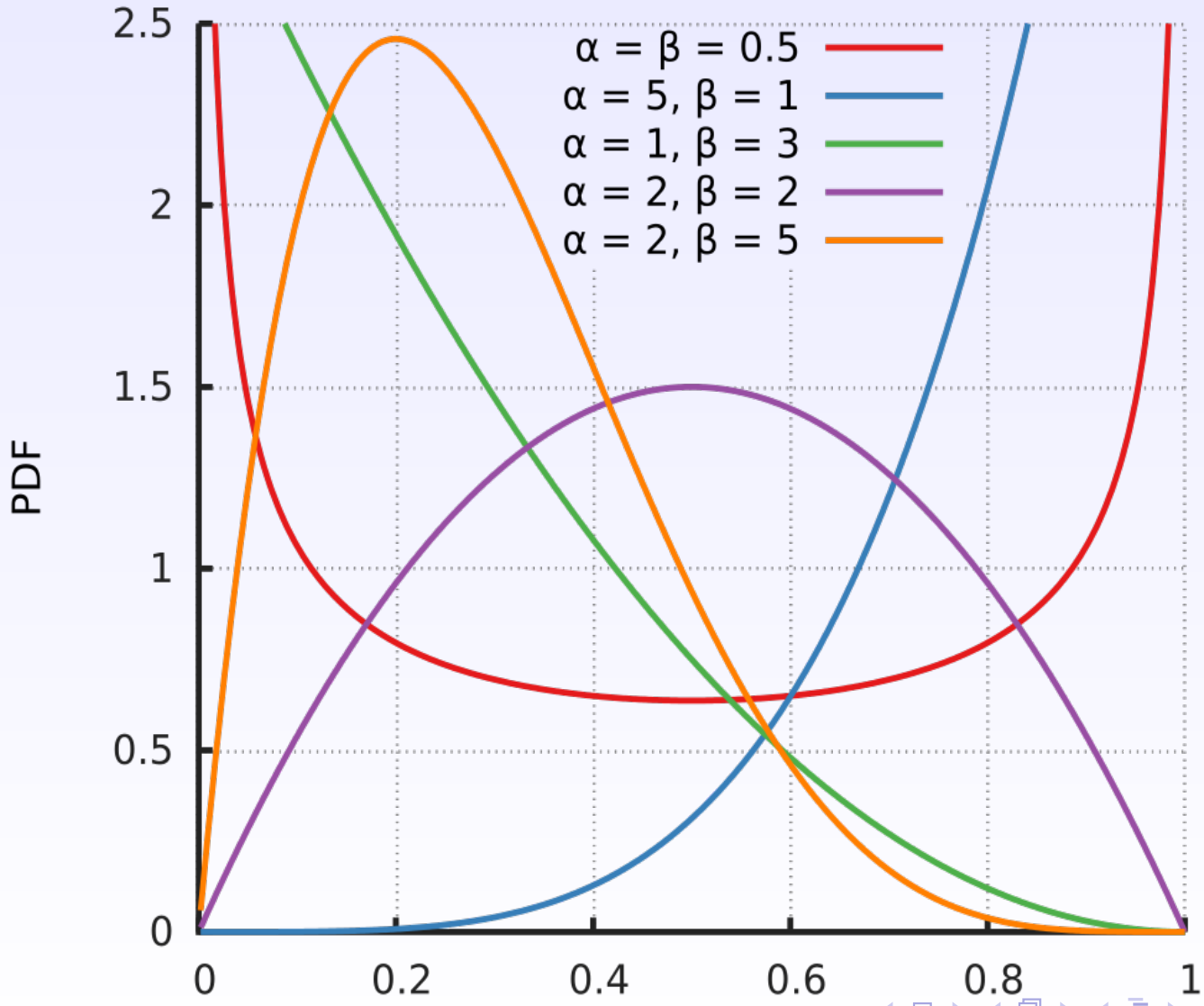
- Es fácil ver que

$$MAP = k_1 \frac{x}{n} + k_2 \frac{p - 1}{p + q - 2}$$

donde k_1 y k_2 están en $(0, 1)$ y $k_1 + k_2 = 1$, y $k_1 \to 1$ para n grande.

- El MAP es una combinación convexa entre el máximo verosímil y la moda apriori.

Algunas densidades Beta



Ejemplo, modelo normal (1)

- Supongamos una muestra aleatoria de una población normal

$$x_1, \ldots, x_n \text{ i.i.d. } N(\mu, \sigma^2)$$

con σ^2 conocida.

- Supongamos una previa normal informativa

$$\pi(\mu) = N(\mu; m, \tau^2)$$

- La posterior es

$$\pi(\mu|x) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \times \exp\left\{-\frac{1}{2\tau^2}(\mu-m)^2\right\}$$

$$\propto \exp\left\{-\frac{A}{2}\left(\mu^2 - 2\frac{1}{A}\left[\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2}m\right]\mu\right)\right\}$$

donde

$$A = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$$

Ejemplo, modelo normal (2)

- Entonces, la posterior es

$$\pi(\mu|x) \propto \exp\left\{-\frac{A}{2}\left(\mu - \frac{1}{A}\left[\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2}m\right]\right)^2\right\}$$

- Esto es

$$\pi(\mu|x) \sim N(\mu; M, A^{-1})$$

donde

$$M = \frac{1}{A}\left[\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2}m\right]$$

$$A = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$$

- Notamos que aquí tambien el MAP es una combinación convexa de dos estimadores, por un lado, el máximo verosímil y, por otro, la mejor estimación previa, m.
- Notamos que los pesos respectivos son inversamente proporcionales a

Distribución predictiva

- Consideremos la predicción de un dato nuevo, z, proveniente de la misma distribución $f(x|\theta)$, de la que fue obtenida la muestra $x_1, \ldots, x_n$
- Para hacer la predicción, podríamos usar $f(x|\widehat{\theta})$; sin embargo, esto no incorporaría la variabilidad inducida por la estimación de θ.
- La alternativa Bayesiana es usar la llamada "distribución predictiva".

$$\begin{aligned} p(z|x) &= \int p(z,\theta|x)d\theta \\ &= \int p(z|\theta,x)\pi(\theta|x)d\theta \\ &= \int p(z|\theta)\pi(\theta|x)d\theta \end{aligned}$$

asumiendo que el nuevo dato, condicional a θ, es independiente de los anteriores.

Distribución predictiva

- En el caso del ejemplo con una población normal, consideremos la predicción de $z = x_{n+1}$.
- Puede verse que (algunos detalles más adelante)

$$p(z|x) = N(M, \sigma^2 + A^{-1})$$

- La alternativa de usar

$$p(z) = N(\bar{x}, \sigma^2)$$

podría ser demasiado optimista, ¿Por qué?

Predictiva normal (1)

- Si $x_1, \ldots, x_n$ son i.i.d. $N(\mu, \sigma^2)$ con σ^2 conocida, vimos que la posterior es

$$\pi(\mu|x) = N(M, A^{-1})$$

$$\text{donde} \quad M = \frac{1}{A}\left[\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau^2 m}\right] \quad \text{y} \quad A = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$$

- Lo anterior se obtuvo para una previa informativa

$$\pi(\mu) = N(m, \tau^2)$$

- Ahora, la predictiva para z es

$$p(z|x) = \int p(z|\mu)\pi(\mu|x)d\mu$$

Predictiva normal (2)

- El kernel de la predictiva se obtiene de integrar

$$p(z|x) \propto \int \exp\left\{-\frac{1}{2\sigma^2}(z-\mu)^2\right\} \exp\left\{-\frac{A}{2}(\mu - M)^2\right\} d\mu$$

- La exponencial a integrar es (quitando términos que no dependan de μ ni de z)

$$\exp\left(-\frac{1}{2\sigma^2}z^2\right) \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}+A\right)\left[\mu^2 - 2\mu\left(\frac{1}{\sigma^2}+A\right)^{-1}\left(\frac{z}{\sigma^2}+AM\right)\right]\right\}$$

- Completando el cuadrado en μ se llega a que el integrando es una densidad normal, integrando a 1. Los términos que salen de la integral, puede verse que forman una densidad normal. Esto es,

$$p(z|x) \equiv N(M, \sigma^2 + A^{-1})$$

Predictiva Poisson (1)

- Supongamos muestra de conteos independientes con distribución Poisson(λ), $x_1, \ldots, x_n$.
- Supongamos también una apriori Gama para λ

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \equiv \text{Gama}(\alpha, \beta)$$

- La posterior es

$$\begin{aligned} \pi(\lambda|x) &\propto \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \times \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} \\ &\propto \lambda^{n\bar{x}+\alpha-1} e^{-(n+\beta)\lambda} \equiv \text{Gama}(n\bar{x}+\alpha, n+\beta) \end{aligned}$$

- El MAP es

$$MAP = \frac{n\bar{x}+\alpha-1}{n+\beta} = \left(\frac{n}{n+\beta}\right)\bar{x} + \left(\frac{\beta}{n+\beta}\right)\frac{\alpha-1}{\beta}$$

Predictiva Poisson (2)

- Ahora, la predictiva para un nuevo conteo z

$$
\begin{aligned}
p(z|x) &= \int f(z|\lambda)\pi(\lambda|x)d\lambda \\
&\propto \int \frac{1}{z!}\lambda^z e^{-\lambda}\ \lambda^{n\bar{x}+\alpha-1}e^{-(n+\beta)\lambda}d\lambda \\
&\propto \frac{1}{z!}\int \lambda^{z+n\bar{x}+\alpha-1}e^{-(n+\beta+1)\lambda}d\lambda \\
&\propto \frac{1}{z!}\ \frac{\Gamma(z+n\bar{x}+\alpha)}{(n+\beta+1)^{z+n\bar{x}+\alpha}} \\
p(z|x) &= \frac{\Gamma(z+n\bar{x}+\alpha)}{\Gamma(n\bar{x}+\alpha)\Gamma(z+1)}\ \left(\frac{n+\beta}{n+\beta+1}\right)^{n\bar{x}+\alpha}\left(\frac{1}{n+\beta+1}\right)^{z}
\end{aligned}
$$

- Recordar Binomial Negativa

Predictiva Poisson (3) (Recordar BN)

- La Binomial Negativa es $X =$ número de éxitos hasta que sucede el fracaso r

$$P(X = x) = \binom{x + r - 1}{x} p^x (1-p)^{r-1}(1-p)$$
$$= \frac{\Gamma(x+r)}{\Gamma(r)\Gamma(x+1)}(1-p)^r p^x$$

- Con media y varianza

$$\frac{pr}{1-p}, \qquad \frac{pr}{(1-p)^2}$$

- Las correspondencias con la predictiva son

$$x = z, \quad r = n\bar{x} + \alpha, \quad p = \frac{1}{n + \beta + 1}$$

Predictiva Poisson (4)

- Entonces, la predictiva para un nuevo conteo z es una Binomial Negativa con media y varianza dados por

$$\frac{n\bar{x}+\alpha}{n+\beta}, \qquad \frac{n\bar{x}+\alpha}{(n+\beta)^2}\cdot(n+\beta+1)$$

  igual media que la posterior, pero mayor varianza.
- Entonces, en vez de usar como modelo para la nueva z una distribución Poisson(λ_{MAP}) con

$$\text{media} = \lambda_{MAP}, \qquad \text{varianza} = \lambda_{MAP}$$

  usaríamos una Binomial Negativa con

$$\text{media} = \lambda_{MAP}, \qquad \text{varianza} = \lambda_{MAP}\cdot\frac{n+\beta+1}{n+\beta}$$

Ejemplo 16.3.2 Regresión lineal (1)

- Consideremos datos independientes $(x_1, y_1), \ldots, (x_n, y_n)$ donde postulamos que

$$y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \ldots, n$$

(consideramos el caso σ^2 conocida).

- La verosimilitud es

$$L(\alpha, \beta) \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right\}$$

- La maximización de L es equivalente a minimizar la suma de cuadrados residual

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} SCE(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Ejemplo 16.3.2 Regresión lineal (2)

- La minimización de $SCE(\alpha, \beta)$ lleva a un sistema de ecuaciones lineales con dos incógnitas

$$\alpha + \bar{x}\beta = \bar{y}$$

$$n\bar{x}\alpha + \left(\sum_{i=1}^{n} x_i^2\right)\beta = \sum_{i=1}^{n} x_i y_i$$

- con solución

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$

donde $S_{xx} = \sum(x_i - \bar{x})^2$ y $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$

- (ver sección 13.5 del Kalbfleisch)

Ejemplo 16.3.2 Regresión lineal (3)

- Un pequeño paréntesis: Veamos el ejemplo 13.5.1 donde se analiza un conjunto de datos usando regresión lineal
- Los datos son de 12 mujeres, para las cuales se registra su edad y su presión sistólica

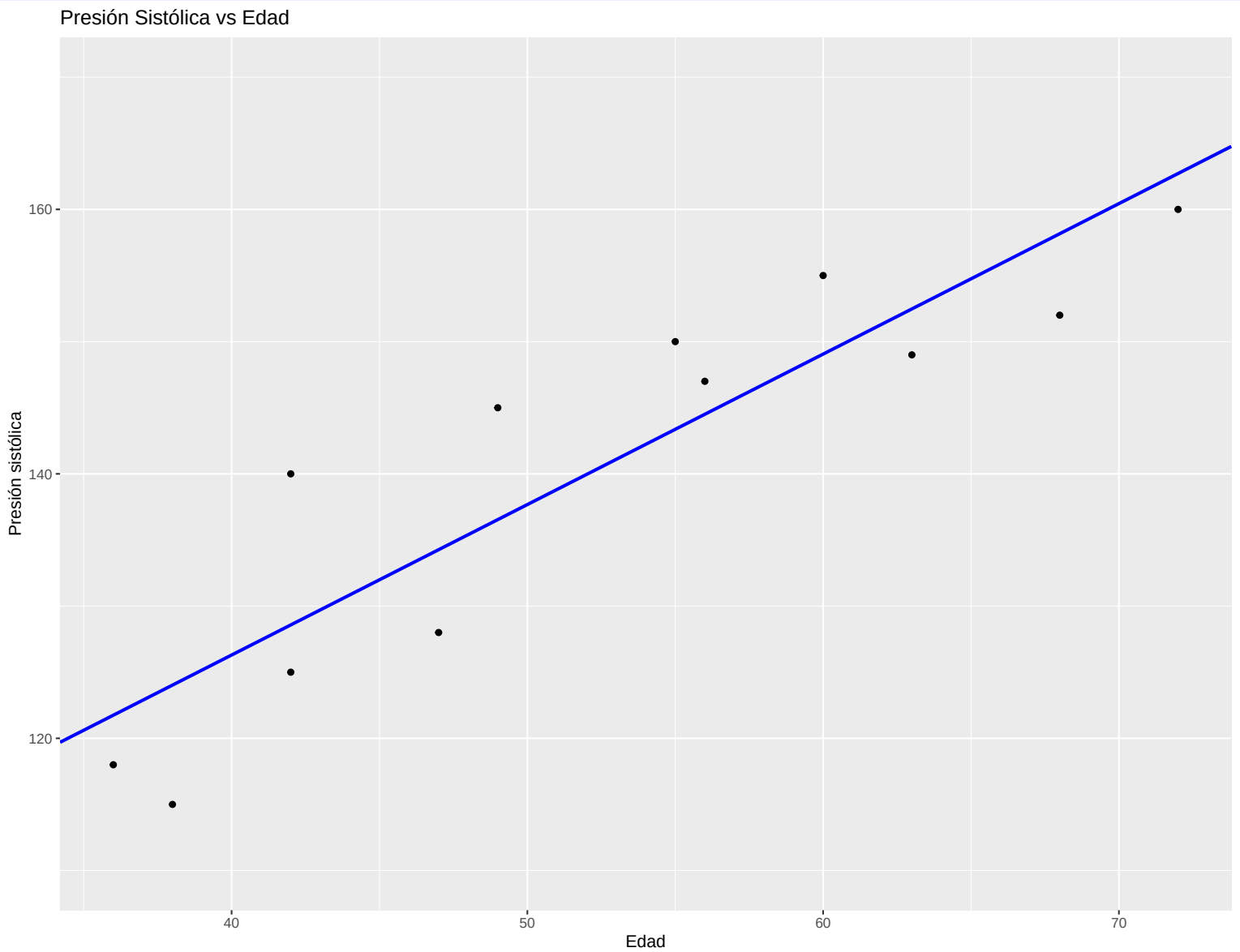
x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

- Los estimadores del intercepto y pendiente son

$$\widehat{\alpha} = 80.778 \qquad \widehat{\beta} = 1.138$$

cada año de edad incrementa, en promedio, 1.138 unidades de presión sistólica.

Ejemplo 16.3.2 Regresión lineal (4)



Ejemplo 16.3.2 Regresión lineal (5)

```
library(ggplot2)
x   <- c(56, 42, 72, 36, 63, 47, 55, 49, 38, 42, 68, 60)
y   <- c(147,125,160,118,149,128,150,145,115,140,152,155)
df  <- data.frame(x=x,y=y)
out <- lm(y~x)
coe <- out$coefficients
# (Intercept)           x
#   80.777730    1.138005
ggplot(data=df,aes(x=x, y=y)) +
  ylim(c(110,170)) +
  xlab("Edad") + ylab("Presión sistólica") +
  ggtitle("Presión Sistólica vs Edad") +
  geom_point(aes(x=x, y=y)) +
  geom_abline(slope = coe[2], intercept = coe[1],
              col="blue", size=1)
```

Ejemplo 16.3.2 Regresión lineal (6)

- Veamos ahora inferencia Bayesiana para regresión lineal
- En forma matricial, el modelo de regresión estándar se escribe como

$$y = X\beta + e, \qquad e \sim N(0, \sigma^2 I)$$

- El estimador de máxima verosimilitud, en forma matricial, puede verse que es

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

- Supongamos una previa normal (conjugada)

$$\beta \sim \pi(\beta) \equiv N(\beta_0, D), \qquad D = \begin{bmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{bmatrix}$$

Ejemplo 16.3.2 Regresión lineal (7)

- La posterior es
$$\pi(\beta|y) \propto \pi(\beta)L(\beta)$$
- Es un ejercicio estándar ver (aunque no lo veremos) que la posterior es Normal
$$\pi(\beta|y) \equiv N(M, A^{-1})$$
$$A = \left(D^{-1} + \frac{1}{\sigma^2}X^T X\right)$$
$$M = A^{-1}\left(D^{-1}\beta_0 + \frac{1}{\sigma^2}X^T y\right)$$
- M es el máximo aposteriori (MAP): $M = (\widehat{\alpha}, \widehat{\beta})^T$
- La estimación Bayesiana para la presión media de una mujer de edad x_0 es
$$X_0^T M = \widehat{\alpha} + \widehat{\beta}x_0$$

Ejemplo 16.3.2 Regresión lineal (8)

- La predicción Bayesiana para la presión de una mujer de edad x_0 se obtiene de la distribución predictiva.
- Si tenemos interés de predicción en

$$y_0 = X_0^T \beta + e_0 \sim N(X_0^T \beta, \sigma^2)$$

- entonces, toda la información disponible se encuentra en la posterior

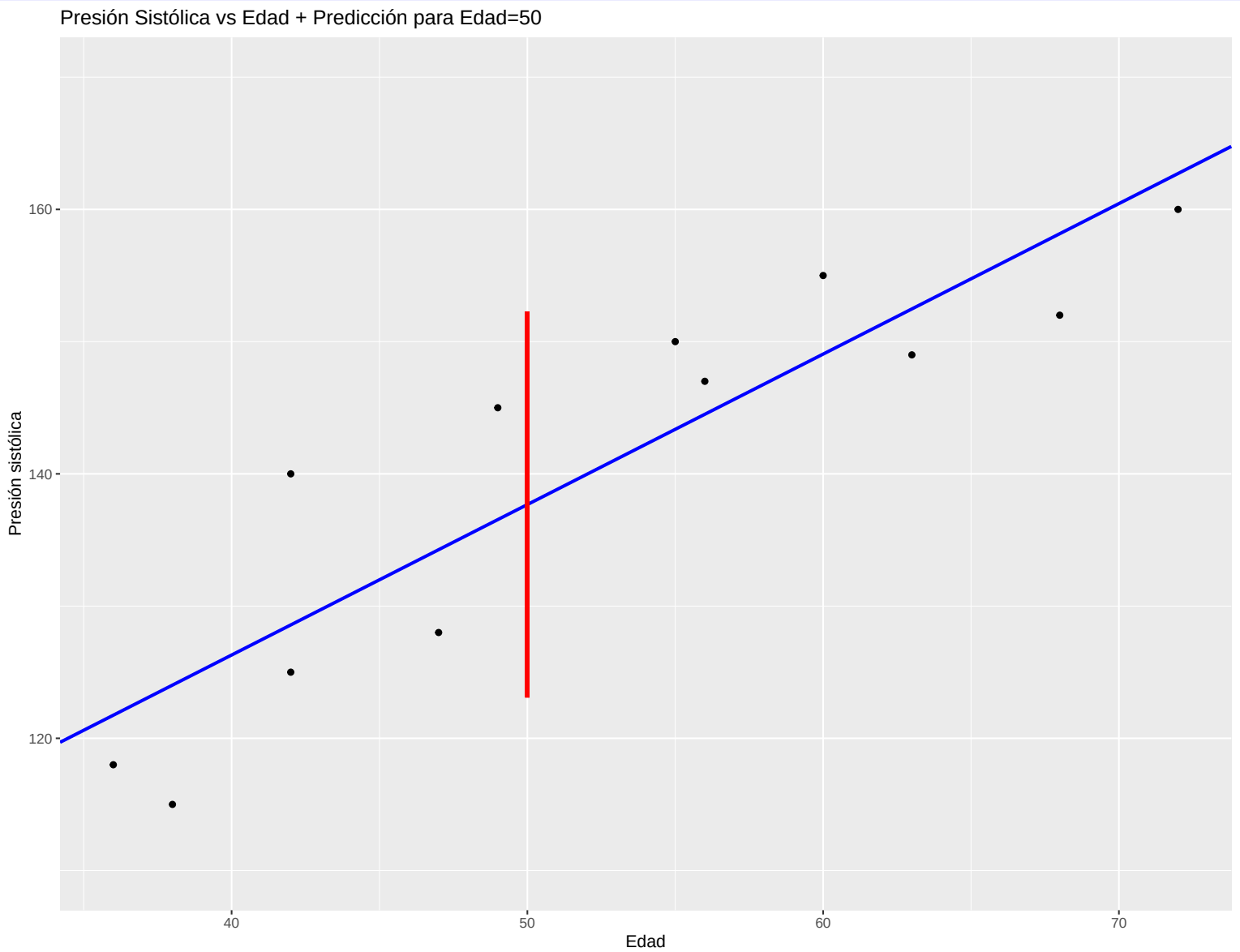
$$X_0^T \beta \sim N(X_0^T M, X_0^T A^{-1} X_0)$$

- Por lo tanto, como vimos antes, la distribución predictiva es

$$Y_0 \sim N(X_0^T M,\ \sigma^2 X_0^T A^{-1} X_0)$$

- La siguiente gráfica muestra los cálculos para la predicción de presión para una mujer de 50 años.

Ejemplo 16.3.2 Regresión lineal (9)



Ejemplo 16.3.2 Regresión lineal (10)

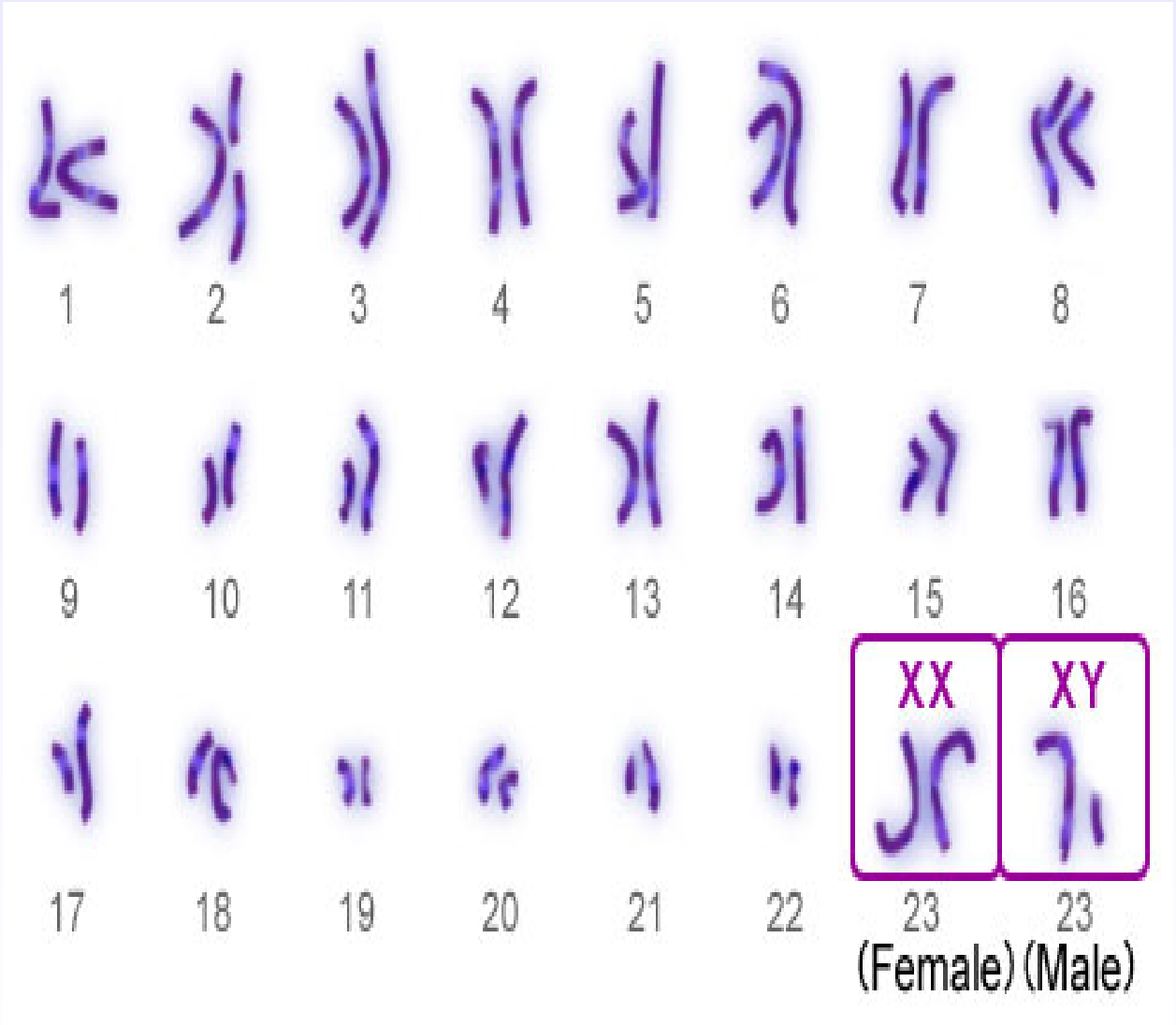
```
# Predicción de presión para edad=50
n   <- length(y)
x0  <- c(1,50)
X   <- cbind(rep(1,n),x)
D   <- diag(c(100,10))
b0  <- c(coe[1],0)
s2  <- sig^2
A   <- solve(D)+t(X)%*%X/s2
M   <- solve(A) %*% (solve(D)%*%b0 + t(X)%*%y/s2)
px0 <- t(M) %*% x0
Vx0 <- s2 + sum( x0 * solve(A,x0) )

ggplot(data=df,aes(x=x, y=y)) +
  ylim(c(110,170)) +
  xlab("Edad") + ylab("Presión sistólica") +
  ggtitle("Presión Sistólica vs Edad + Predicción para Edad=50") +
  geom_point(aes(x=x, y=y)) +
  geom_abline(slope = coe[2], intercept = coe[1],col="blue",size=1) +
  geom_segment(aes(x=50,y=px0-2*sqrt(Vx0),
                   xend=50,yend=px0+2*sqrt(Vx0) ),col="red",size=1.3)
```

Ejemplo 16.2.1 Hemofilia

- En este ejemplo se muestra el uso de la previa para incorporar explícitamente información previa.
- Pero primero, un vistazo rápido al fenómeno de la Hemofilia.
- La hemofilia es una enfermedad transmitida de padres a hijos y esta ligada a un gen localizado en el cromosoma X. Los hombres tienen XY y las mujeres XX. Los decendientes reciben un cromosoma del padre y uno de la madre.
- Un hijo recibe el cromosoma Y del padre y una X de la madre. Una hija recibe el X del padre y una de las X's de la madre.

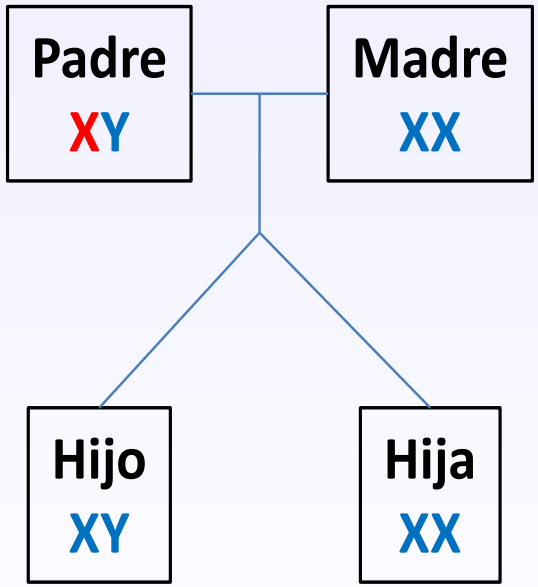
Cromosomas humanos



Algunos cálculos (simplificados) sobre Hemofilia (1)

- Supongamos padre (P) con hemofilia y madre (M) no portadora.
- ¿Cuál es la probabilidad de bebé con hemofilia, o portador del gen?

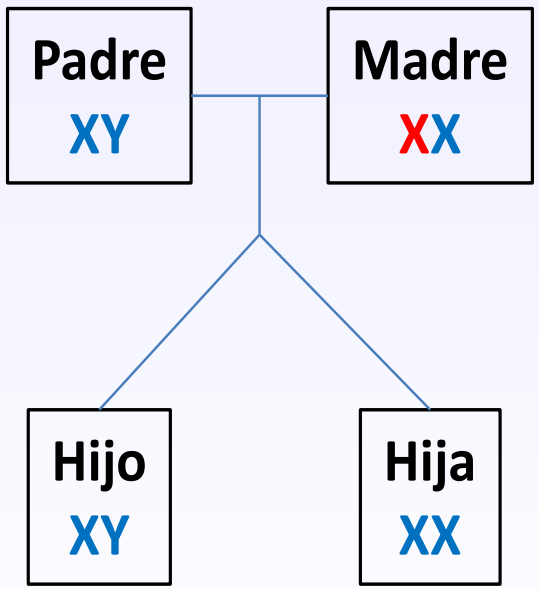
$$P(\text{ hemof} \mid \text{niño }) = 0$$
$$P(\text{ hemof} \mid \text{niña }) = 0$$
$$P(\text{ porta} \mid \text{niña }) = 1$$



Algunos cálculos (simplificados) sobre Hemofilia (2)

- Supongamos P sin hemofilia y M portadora.
- ¿Cuál es la probabilidad de bebé con hemofilia, o portador del gen?

$$P(\text{ hemof } | \text{ niño }) = \frac{1}{2}$$
$$P(\text{ hemof } | \text{ niña }) = 0$$
$$P(\text{ porta } | \text{ niña }) = \frac{1}{2}$$



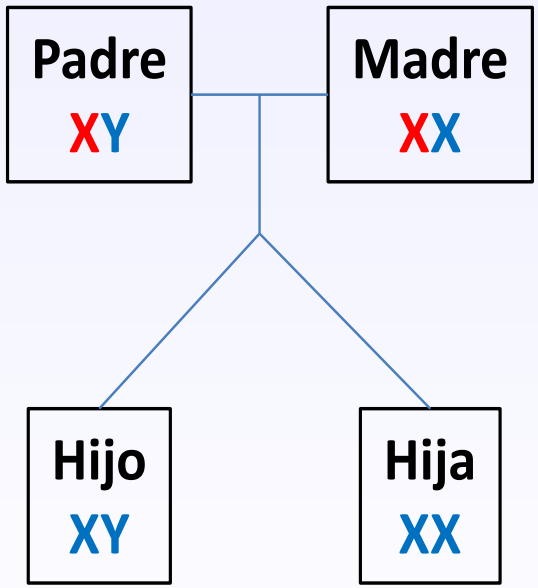
Algunos cálculos (simplificados) sobre Hemofilia (3)

- Supongamos P con hemofilia y M portadora.
- ¿Cuál es la probabilidad de bebé con hemofilia, o portador del gen?

$$P(\text{ hemof} \mid \text{niño }) = \frac{1}{2}$$

$$P(\text{ hemof} \mid \text{niña }) = \frac{1}{2}$$

$$P(\text{ porta} \mid \text{niña }) = 1$$



Ejemplo 16.2.1

- Supongamos el caso de una mujer con n hijos varones, de los cuales x tienen hemofilia. La probabilidad de este evento es

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$$

  donde θ es la probabilidad de que un hijo dado tenga hemofilia.
- Dado que se observan x casos en n eventos. ¿Cómo estimar θ? (si, es un escenario algo forzado, pero bueno, sigamos $\cdots$)
- Si no tenemos ninguna otra información, entonces la verosimilitud anterior es nuestra única herramienta.
- Pero, supongamos que si sabemos algo: Sus padres son normales, pero tiene un hermano con hemofilia. ¿Cómo estimar θ?

Ejemplo 16.2.1

- Si sus padres son normales, pero su hermano tiene hemofilia, entonces la madre de la mujer tiene que ser portadora del gen. Por lo tanto, ella tiene una probabilidad apriori de $1/2$ de ser portadora.
- Entonces, antes de saber el estado de sus hijos, la probabilidad de hijo con hemofilia era $\theta = 1/2$ (si es que ella era portadora) o la probabilidad era $\theta = 0$ (si es que no era portadora) (no estamos considerando mutaciones).
- Con esta información, una distribución apriori es

$$\pi(\theta) = \begin{cases} \frac{1}{2} & \text{para } \theta = 0 \\ \\ \frac{1}{2} & \text{para } \theta = \frac{1}{2} \end{cases}$$

- ¿Cuál es la distribución posterior?

Ejemplo 16.2.1

- La distribución posterior es

$$\pi(\theta|x) \propto \pi(\theta)L(x|\theta)$$

- Si $x > 0$, entonces

$$\pi(\theta|x) = \begin{cases} \pi(\theta = 0|x) = 0 \\ \\ \pi(\theta = \frac{1}{2}|x) \propto \left(\frac{1}{2}\right)^n \end{cases}$$

esto es

$$\pi(\theta|x) = \begin{cases} 0 & \text{para } \theta = 0 \\ \\ 1 & \text{para } \theta = \frac{1}{2} \end{cases}$$

Ejemplo 16.2.1

- Si $x = 0$, entonces

$$\pi(\theta|x) = \begin{cases} \pi(\theta = 0|x) = \frac{1}{2} \\ \\ \pi(\theta = \frac{1}{2}|x) \propto \left(\frac{1}{2}\right)^{n+1} \end{cases}$$

esto es

$$\pi(\theta|x) = \begin{cases} \frac{2^n}{2^n+1} & \text{para } \theta = 0 \\ \\ \frac{1}{2^n+1} & \text{para } \theta = \frac{1}{2} \end{cases}$$

- Si la mujer tiene al menos un hijo hemofílico, entonces $x > 0$ y por lo tanto $\theta = 1/2$.
- Si la mujer no tiene hijos hemofílicos, entonces $x = 0$ y la probabilidad de que ella se portadora decrece con n.

Ejemplo Poisson (1)

- Considere conteos Poisson $y_1, \ldots, y_n$. El modelo estándar es

$$y_i \sim \text{Poison}(\lambda), \quad i = 1, \ldots, n$$

- Con frecuencia, los conteos son hechos en áreas de diferentes tamaños, o en poblaciones de diferente tamaño, o en tiempos de diferente duración. Para tomar esto en consideración, los datos los representamos como

$$(y_1, x_1), \cdots, (y_n, x_n)$$

donde las x_i's (llamadas "exposiciones") son los tamaños de las diferentes unidades de donde vienen los conteos y_i's.

- El modelo es entonces

$$y_i \sim \text{Poison}(\theta x_i), \quad i = 1, \ldots, n$$

donde θ es la tasa de ocurrencias por unidad.

Ejemplo Poisson (2)

- Ejemplo tomado de (pág 45)
  - Gelman *et al.* (2014) Bayesian Data Analysis. Chapman & Hall
- Supongamos una ciudad de 200,000 habitantes, en donde, en un año, murieron 3 personas por asma. Un modelo probabilístico puede ser

$$y \sim \text{Poisson}(\theta x)$$

  donde θ es el número de muertes por asma (por cada 100,000 habitantes) y $x = 2.0$ para la ciudad en cuestión.
- Podemos usar información epidemiológica conocida para tener una idea de la tasa de muerte y poder construir una distribución apriori para θ.

Ejemplo Poisson (3) Elicitación de previa

- Los datos de mortalidad por asma a nivel mundial, indican que es raro tener tasas por arriba de 1.5 en países occidentales, típicamente estas tasa andan alrededor de 0.6 muertes por cada 100,000 habitantes.
- Por conveniencia (conjugación) elgimos un miembro de la familia Gama, que tenga las características anteriores

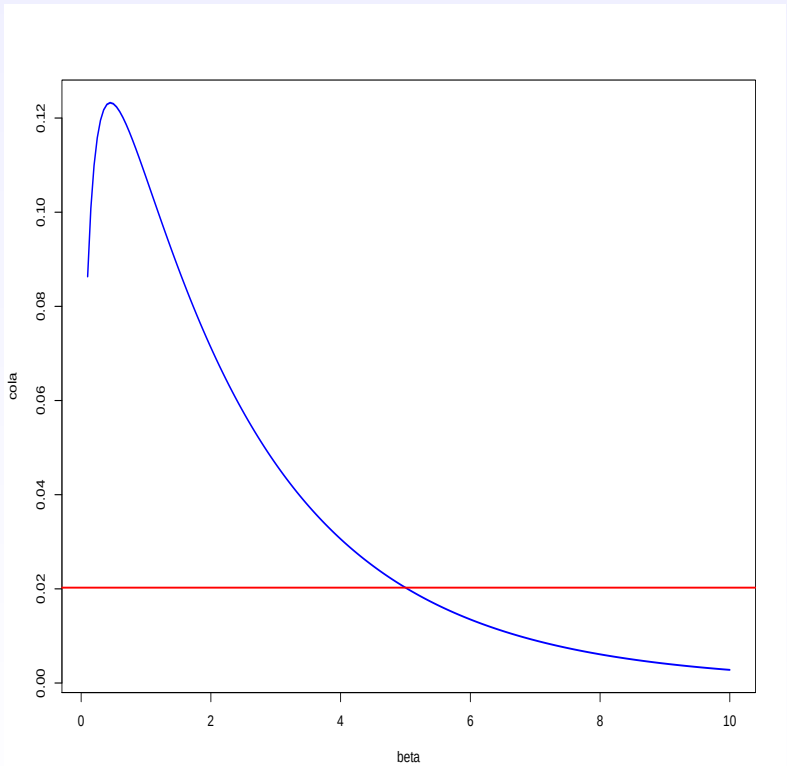
$$\frac{\alpha}{\beta} \approx 0.6 \qquad P(W \geq 1.5) \leq 0.05$$

- De acuerdo al pequeño programa que se encuentra en la siguiente lámina (básicamente, a prueba y error), un par de valores razonables para la apriori Gama son

$$\alpha = 3 \qquad \beta = 5$$

Ejemplo Poisson (4) Elicitación de previa

```
beta   <- seq(0.1,10,length.out = 201)
cola   <- 1 - pgamma(1.5, shape=0.6*beta, rate=beta)
plot(beta,cola,type="l", lwd=2,col="blue")
beta0  <- 5
cola0  <- 1 - pgamma(1.5, shape=0.6*beta0, rate=beta0)  # 0.02025672
abline(h=cola0,lwd=2,col="red")
```



Ejemplo Poisson (5) Distribución posterior

- La posterior para θ es

$$\pi(\theta|y) \propto \pi(\theta)\ L(\theta)$$

- Para el caso de apriori Gama y verosimilitud Poisson, tenemos (vimos estos cálculos antes)

$$\pi(\theta|y) \propto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\ (x\theta)^y e^{-x\theta} \propto \theta^{y+\alpha-1}e^{-(x+\beta)\theta}$$

- La posterior es Gamma

$$\pi(\theta|y) \equiv \text{Gama}(y+\alpha,\ x+\beta)$$

- con los datos del ejemplo

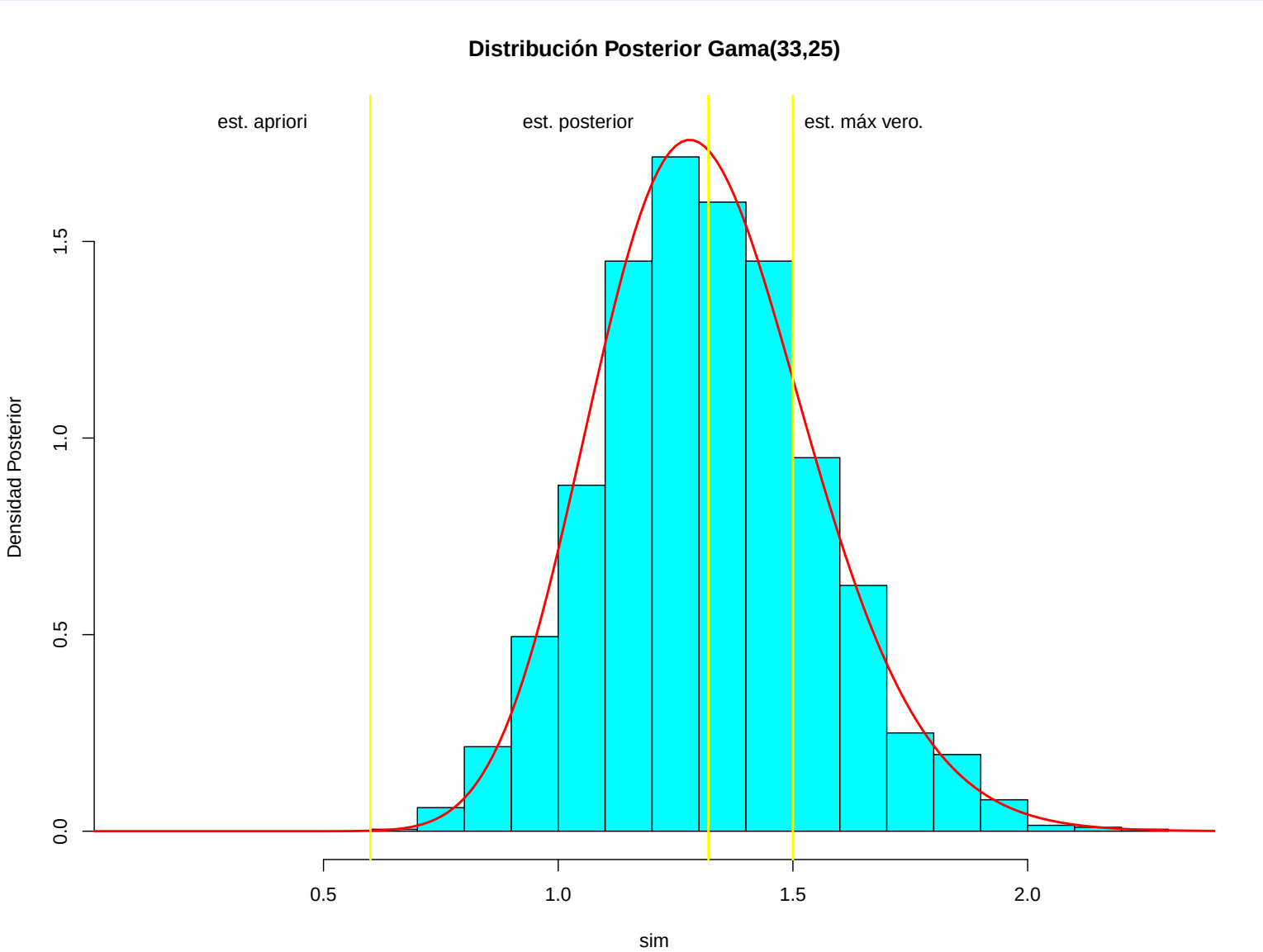
$$\pi(\theta|y) \equiv \text{Gama}(3+3,\ 2+5) = \text{Gama}(6,7)$$

Ejemplo Poisson (6) Simulación de posterior

En este caso, no es necesario simular de la posterior, pues todas sus propiedades se conocen en forma analítica. En otros casos, la simulación de la posterior es la única opción viable y los métodos MCMC son extremadamente útiles.

```
M    <- 2000
set.seed(8656)
sim  <- rgamma(M,shape=6,rate=7)
hist(sim,breaks=20,col="cyan",freq=FALSE,ylim=c(0,1.8),xlim=c(.1,2.31),
     main="Distribución Posterior  Gama(6,7)",ylab="Densidad Posterior")
tet  <- seq(0,2.5,length.out = 201)
post <- dgamma(tet,shape=6,rate=7)
lines(tet,post,lwd=2,col="red")
abline(v=c(3/5,6/7,3/2),lwd=2,col="yellow")
text(0.2,1.3,"est. apriori",pos=4)
text(0.86,1.3,"est. posterior",pos=4)
text(1.5,1.3,"est. máx vero.",pos=4)
```

Ejemplo Poisson (7) Gráfica de posterior



Ejemplo Poisson (8)

- Hemos comentado sobre la relación entre las diferentes estimaciones

$$\text{Est. Posterior} = w_1 \text{ Est. apriori} + w_2 \text{ Est. Máx. Versímil}$$

donde $w_1 + w_2 = 1$ con $w_i \geq 0$. En la gráfica visualizamos como el MAP esta bastante cargado hacia la estimación apriori ("shrinkage" grande).

- Supongamos ahora, que tenemos 10 años de datos de mortalidad para esa misma ciudad y supongamos que $\sum y_i = 30$ para esos $n = 10$ años (esto es, se observó la misma tasa de mortalidad).
- Entonces la verosimilitud es

$$L(\theta) \propto \theta^{\sum y_i} e^{-nx\theta}$$

- y la posterior es, con menos shrinkage

$$\pi(\theta|y) \equiv \text{Gama}\left(\sum y_i + \alpha,\ nx + \beta\right) = \text{Gama}(33, 25)$$

Ejemplo Poisson (9) Simulación de posterior

```
M    <- 2000
set.seed(8656)
sim  <- rgamma(M,shape=33,rate=25)
hist(sim,breaks=20,col="cyan",freq=FALSE,ylim=c(0,1.8),xlim=c(.1,2.31),
   main="Distribución Posterior G(33,25)",ylab="Densidad Posterior")
tet  <- seq(0,2.5,length.out = 201)
post <- dgamma(tet,shape=33,rate=25)
lines(tet,post,lwd=2,col="red")
abline(v=c(3/5,33/25,3/2),lwd=2,col="yellow")
text(0.25,1.8,"est. apriori",pos=4)
text(0.9,1.8,"est. posterior",pos=4)
text(1.5,1.8,"est. máx vero.",pos=4)
```

Ejemplo Poisson (10) Gráfica de posterior