EYP1113 - Probabilidad y Estadística

Capítulo 6: Inferencia Estadística

Ana María Araneda Ricardo Aravena Ricardo Olea

Departamento de Estadística
Pontificia Universidad Católica de Chile

Primer Semestre 2012

Probability Concepts in Engineering

Alfredo H-S. Ang† and Wilson H. Tang‡

† University of Illinois at Urbana-Champaing and University of California, Irvine

‡ Hong Kong University of Science & Technology

Contenido I

Contenido II

- Intervalo de Confianza para la media
- Intervalo de Confianza para una proporción

6. Teoría de la medida

Estimación

Definiciones y Propiedades

En los capítulos previos hemos visto de manera introductoria como, dada una distribución ($p_X(x)$, $f_X(x)$ o $F_X(x)$) de una variable aleatoria X y el valor de sus parámetros, obtener probabilidades.

El calculo de probabilidades depende del valor de los parámetros.

Por tanto, nos interesa disponer de métodos que permitan seleccionar adecuadamente valores de estos para las distribuciones de importancia práctica.

Para realizar lo anteriormente expuesto, requerimos información "del mundo real".

Por ejemplo, datos referente a la pluviosidad en cierta área, intensidad y frecuencia de los movimientos telúricos, conteos, velocidades y flujo de vehículos en cierta intersección o vía, etc.

Estimación

Definiciones y Propiedades

Con base a estos datos, los parámetros pueden ser estimados estadísticamente, y con información sobre el fenómeno inferir la distribución de probabilidad.

Mundo Real

$\downarrow$

Colección de Datos

$\downarrow$

Estimación de Parámetros
(Escoger Distribución)

$\downarrow$

Calculo de Probabilidades

$\downarrow$

Información para Decidir−Hacer−Diseñar

Estimación

Definiciones y Propiedades

La estimación clásica de parámetros consiste en dos tipos:

- **Puntal**: simplemente indica un valor único, basado en los datos, para representar el parámetro de interés.
- **Intervalar**: entrega un conjunto de valores (intervalo) donde el parámetro puede estar con cierto nivel de confianza.

Estimación

Definiciones y Propiedades

Propiedades deseables para un estimador puntual

- **Insesgamiento**: valor esperado del estimador sea igual al parámetro de interés.
- **Consistencia**: implica que si $n \to \infty$, el estimador converge al parámetro (propiedad asintótica)
- **Eficiencia**: se refiere a que la varianza del estimador. Dado un conjunto de datos, θ_1 es más eficiente que θ_2 para estimar θ si tiene menor varianza.
- **Suficiencia**: un estimador se dice suficiente si utiliza toda la información contenida en la muestra para estimar el parámetro.

Estimación Puntual

Métodos de Estimación

Método de los Momentos

En términos generales, el método propone igualar los momentos teóricos no centrales de una variable aleatoria X denotados por μ_k, con los momentos empíricos, basados en los datos, m_k, y despejar los parámetros de interés. Es decir,

$$\mu_k = E(X^k) \quad \text{y} \quad m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$\Rightarrow \mu_k = m_k, \quad k = 1, 2, \ldots$$

Estimación Puntual

Métodos de Estimación

Ejercicio Sean $X_1, \ldots, X_n$ una muestra aleatoria cuya distribución de probabilidad está dada por:

- Bernoulli(p)
- Poisson(λ)
- Exponencial(ν)
- Gamma(k, ν)
- Normal(μ, σ)

Obtenga los estimadores de momentos de los parámetros.

Estimación Puntual

Métodos de Estimación

Método de Máxima Verosimilitud

Otro método de estimación puntual es el denominado método de máxima verosimilitud (MV).

En contraste con el método de los momentos, el método de máxima verosimilitud deriva directamente en estimador puntual del parámetro de interés.

Sea X variable aleatoria con función de probabilidad $f_X(x,\theta)$, donde θ es el parámetro de interés.

Dada una muestra (es decir, valores observados) $x_1,\ldots,x_n$, nos preguntamos cuál es el valor más probable de θ que produzca estos valores.

Estimación Puntual

Métodos de Estimación

Método de Máxima Verosimilitud

En otras palabras, para los diferentes valores de θ, cuál es el valor que maximiza la verosimilitud de los valores observados $x_1,\ldots,x_n$.

La función de verosimilitud de una muestra aleatoria $x_1,\ldots,x_n$, es decir, independiente e idénticamente distribuida es:

$$
\begin{aligned}
L(x_1, x_2, \ldots, x_n, \theta) &= f_{X_1,\ldots,X_n}(x_1, \ldots, x_n, \theta) \\
&= f_{X_1}(x_1, \theta) \times \cdots \times f_{X_n}(x_n, \theta), \quad \text{por independencia} \\
&= \prod_{i=1}^{n} f_X(x_i, \theta), \quad \text{por idéntica distribución}
\end{aligned}
$$

Se define el estimador de máxima verosimilitud (EMV) como el valor de θ que maximiza la función de verosimilitud L.

Estimación Puntual

Métodos de Estimación

Método de Máxima Verosimilitud

Es decir, es la solución de

$$\frac{\partial}{\partial\,\theta}\,L(x_1,\,x_2,\ldots,\,x_n,\theta) = 0 \Rightarrow \hat{\theta} = \theta$$

Maximizar L es equivalente a maximizar $\ln L$, es decir,

$$\frac{\partial}{\partial\,\theta}\,\ln L(x_1,\,x_2,\ldots,\,x_n,\theta) = 0$$

Estimación Puntual

Métodos de Estimación

Método de Máxima Verosimilitud

Si la función de distribución (discreta o continua) depende de más de un parámetro, digamos $\theta_1, \ldots, \theta_m$, los EMV respectivos son las soluciones de las m ecuaciones:

$$\frac{\partial}{\partial \theta_j} \ln L(x_1, x_2, \ldots, x_n, \theta_1, \ldots, \theta_m) = 0 \quad j = 1, 2, \ldots, m$$

Los EMV son estimadores que poseen las propiedades deseables descritas anteriormente.

En particular, para n grande, son "los mejores" estimadores (en el sentido de varianza mínima)

Estimación Puntual

Métodos de Estimación

Ejercicio

Sean $X_1, \ldots, X_n$ una muestra aleatoria cuya distribución de probabilidad está dada por:

- Bernoulli(p)
- Poisson(λ)
- Exponencial(ν)
- Gamma(k, ν)
- Normal(μ, σ)

Obtenga los estimadores máximo verosímiles de los parámetros.

Estimación Puntual

Métodos de Estimación

Propiedades de los Estimadores Máximo Verosímiles

- Asintóticamente Insesgados: $E(\hat{\theta}_n) \to \theta$, cuando $n \to \infty$.
- Varianza alcanza la cota de Cramer-Rao:

$$\text{Var}(\hat{\theta}_n) = \frac{1}{I_n(\theta)},$$

con $I_n(\theta) = -E\left[\frac{\partial^2}{\partial\,\theta^2} \ln\, L(\theta)\right]$.

- Distribución Asintótica: Normal.
- Invarianza: Si $\hat{\theta}_n$ es el estimador máximo verosímil de θ, entonces $g(\hat{\theta}_n)$ es el estimador máximo verosímil de $g(\theta)$ cuya distribución asintótica es

$$g(\hat{\theta}) \;\dot{\sim}\; \text{Normal}\left(g(\theta),\, \sqrt{\frac{[g'(\theta)]^2}{I_n(\theta)}}\right)$$

Distribuciones Muestrales

Distribución de la Media con Varianza Conocida

Sea $X_1, \ldots, X_n$ una sucesión de variables aleatorias independientes con función de probabilidad $p_X(x)$ o de densidad $f_X(x)$.

Si $E(X) = \mu$ y $\text{Var}(X) = \sigma^2$, entonces el valor esperado y varianza de $\overline{X}_n$ son

$$\mu_{\overline{X}_n} = E(\overline{X}_n) = \mu$$

$$\sigma^2_{\overline{X}_n} = \text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$$

Distribuciones Muestrales

Distribución de la Media con Varianza Conocida

Si la distribución subyacente es Normal, entonces

$$\overline{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

En el caso que la distribución NO sea Normal, por el Teorema del Límite Central para n grande se cumple que

$$\overline{X}_n \overset{\text{aprox}}{\sim} \text{Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \overset{\text{aprox}}{\sim} \text{Normal}(0, 1)$$

Distribuciones Muestrales

Distribución de la Media con Varianza Desconocida

Sea $X_1,\ldots,X_n$ una sucesión de variables aleatorias independientes con función de probabilidad $p_X(x)$ o de densidad $f_X(x)$ tal que $E(X) = \mu$ y $\text{Var}(X) = \sigma^2$.

Generalmente la varianza poblacional es desconocida.

Para tal caso, si reemplazamos σ por su estimador muestral S^2 se tiene que

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim \text{t-student}(n-1)$$

con

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Distribuciones Muestrales

Distribución de la Media con Varianza Desconocida

Un variable aleatoria T tiene distribución t-student si su función de densidad está dada por:

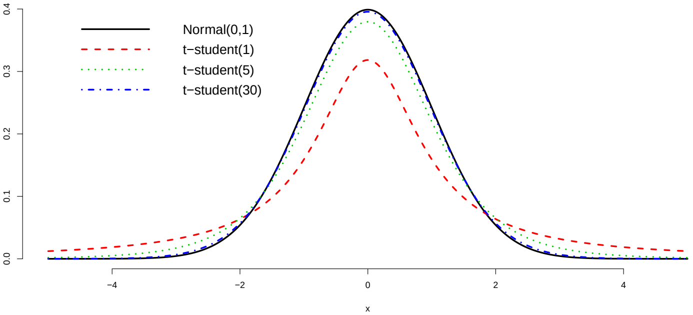
$$f_T(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\,\nu}\,\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty$$

W. S. Gosset (1908)

Distribuciones Muestrales

Distribución de la Media con Varianza Desconocida

Normal Estándar vs. t-Student



Distribuciones Muestrales

Distribución de la Varianza Muestral

La varianza muestral está definida como

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)$$

La cual cumple con la siguiente igualdad

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i-\mu-\overline{X}_n+\overline{X}_n)^2\\ &= \sum_{i=1}^{n}(X_i-\overline{X}_n)^2+n(\overline{X}_n-\mu)^2\\ &= (n-1)\,S^2 + +n(\overline{X}_n-\mu)^2\end{aligned}$$

Ejercicio: Muestre que S^2 es un estimador insesgado para σ^2.

Distribuciones Muestrales

Distribución de la Varianza Muestral

Ejercicio

Sean Z y U variables aleatorias independientes con distribución Normal(0,1) y Chi-Cuadrado(ν) respectivamente.

Muestre que:

$$\frac{Z}{\sqrt{U/\nu}} \sim \text{t-student}(\nu)$$

Justifique con este resultado que

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim \text{t-student}(n-1)$$

Prueba de Hipótesis

Introducción

Una prueba de hipótesis es un método estadístico inferencial para la toma de decisiones sobre una población en base a la información proporcionada por los datos de una muestra.

La inferencia puede hacerse con respecto a uno o más parámetros de la población o también para un modelo de distribución.

Prueba de Hipótesis

Introducción

Prueba de Hipótesis Una hipótesis es una afirmación con respecto a uno o más parámetros de una población.

Usualmente son dos las hipótesis que se contrastan son:

- Hipótesis nula, H_0.
- Hipótesis alternativa, H_a.

Ejemplo: Cuando interesa inferir sobre el valor de un parámetro μ de la población las hipótesis a contrastar son generalmente:

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{vs} \quad \mathrm{H}_a : \mu \neq \mu_0$$

Prueba de Hipótesis

Procedimiento para una Prueba de Hipótesis

Los pasos necesarios en las pruebas de hipótesis son:

1. Defina la hipótesis nula y alternativa.

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{vs} \quad \mathrm{H}_a : \mu \neq \mu_0$$

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{vs} \quad \mathrm{H}_a : \mu > \mu_0$$

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{vs} \quad \mathrm{H}_a : \mu < \mu_0$$

2. Identificar la prueba estadística adecuada y su distribución.
3. Basado en una muestra de datos observados estime el estadístico de prueba.
4. Especifique el nivel de significancia.

Prueba de Hipótesis

Procedimiento para una Prueba de Hipótesis

Dado que el estadístico de prueba es una variable aleatoria, la probabilidad de una decisión errónea puede ser controlada. Los errores que se pueden cometer son:

- **Error Tipo I**: Se rechaza H_0 dado que era correcta.
- **Error Tipo II**: No se rechaza H_0 dado que no era correcta.

La probabilidad de **Error Tipo I** se denota como α, la cual corresponde al nivel de significancia de la prueba de hipótesis.

Prueba de Hipótesis

Procedimiento para una Prueba de Hipótesis

Prueba de Hipótesis

Ejemplos

Sea $X_1,\ldots,X_n$ una muestra aleatoria proveniente de una distribución Normal$(\mu,\, \sigma)$, determine una prueba de hipótesis para

- μ, con σ^2 conocido.
- μ, con σ^2 desconocido.
- σ^2.

Prueba de Hipótesis

Valor-p y Potencia

- Valor-p
- Potencia.

Prueba de Hipótesis

Ejercicios

Ejemplo 6.5

Suponga que se especifica que un tipo de barra debe tener una resistencia media igual o superior 38 psi. para ser utilizada en la construcción.

De un lote recién llegado, se extrae una muestra de 25 barras las que ofrece en promedio una resistencia de 37.5 psi.

Al ingeniero a cargo le preocupa que las barras no cumplan las especificaciones, por esta razón decide realizar una prueba de hipótesis. ¿Envía de vuelta el camión? Utilice un $\alpha = 5\%$. Suponga que las

resistencias tienen distribución normal con desviación estándar conocida igual a 3.0 psi.

Ejemplo 6.5

Considere ahora que la desviación estándar es desconocida y que su estimación muestral es 3.0 psi.

Prueba de Hipótesis

Ejercicios

Ejemplo 6.7

Suponga que la muestra se incremento a 41 barras y que la media y desviación estándar muestral son 37.6 psi. y 3.75 psi. respectivamente. Para un nivel a $= 2.5\%$ pruebe las siguientes hipótesis:

$$\mathrm{H}_0 : \sigma^2 = \sigma_0^2 \qquad \text{vs.} \quad \mathrm{H}_a : \sigma^2 > \sigma_0^2$$

con $\sigma_0^2 = 9.0$.

Intervalos de Confianza

Intervalo de Confianza para la media

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población cuya distribución es Normal$(\mu,\, \sigma)$.

Ya vimos que un estimador insesgado y consistente para μ esta dado por

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \text{Normal}\left(\mu,\, \frac{\sigma}{\sqrt{n}}\right)$$

Intervalos de Confianza

Intervalo de Confianza para la media

Intervalo de Confianza para μ con σ conocido

Tenemos que

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0,1)$$

Luego, se puede mostrar que

$$< \mu >_{1-\alpha} \in \ \overline{X}_n \pm k_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Intervalos de Confianza

Intervalo de Confianza para la media

Intervalo de Confianza para μ con σ desconocido

Tenemos que

$$T_n = \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim \text{t-student}(n-1)$$

Luego, se puede mostrar que

$$< \mu >_{1-\alpha} \in \ \overline{X}_n \pm t_{1-\alpha/2}(n-1) \cdot \frac{S}{\sqrt{n}}$$

Intervalos de Confianza

Intervalo de Confianza para la media

Determinación del Tamaño de Muestra

Como se aprecia en la construcción de los Intervalos de Confianza, el tamaño de muestra es fundamental.

Al observar el Intervalo de Confianza para μ, se aprecia que el semi-ancho esta dado por:

$$k_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} = w$$

Lo anterior se conoce como **Error de Estimación**.

Intervalos de Confianza

Intervalo de Confianza para la media

Determinación del Tamaño de Muestra

Por lo tanto, para una precisión w dada, es posible determinar el tamaño de muestra necesaria, con σ y α fijos, dado por

$$n = \left(\frac{k_{1-\alpha/2}\, \sigma}{w} \right)^2$$

Nota: Alternativamente también se puede determinar un tamaño muestral a partir controlando por los errores tipo I y II de una prueba de hipótesis.

Intervalos de Confianza

Intervalo de Confianza para una proporción

Consideremos ahora una muestra aleatoria $X_1,\ldots,X_n$ proveniente de una población cuya distribución es Bernoulli(p).

Ya vimos que un estimador insesgado y consistente para p esta dado por:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\text{aprox.}}{\sim} \text{Normal}\left(p,\, \sqrt{\frac{p\,(1-p)}{n}}\right)$$

Intervalos de Confianza

Intervalo de Confianza para una proporción

Intervalo de Confianza para p (Tamaños muestrales grandes)

Tenemos que

$$T_n = \frac{\overline{X}_n - p}{S/\sqrt{n}} \overset{\text{aprox.}}{\sim} \text{Normal}(0,1),$$

con $S^2 = \overline{X}_n(1 - \overline{X}_n)$.

Luego, se puede mostrar que

$$< p >_{1-\alpha} \in \overline{X}_n \pm k_{1-\alpha/2} \cdot \sqrt{\frac{\overline{X}_n\,(1 - \overline{X}_n)}{n}}$$

Intervalos de Confianza

Intervalo de Confianza para la media

Consideremos nuevamente una muestra aleatoria $X_1,\ldots,X_n$ proveniente de una población cuya distribución es Normal$(\mu,\,\sigma)$.

Recordemos que un estimador insesgado y consistente para σ^2 esta dado por:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \Rightarrow \frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Intervalos de Confianza

Intervalo de Confianza para la media

Intervalo de Confianza para σ^2

Tenemos que

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

Luego, se puede mostrar que

$$< \sigma^2 >_{1-\alpha} \in \left[\frac{(n-1)\,S^2}{c_{1-\alpha/2}(n-1)};\ \frac{(n-1)\,S^2}{c_{\alpha/2}(n-1)}\right]$$

Teoría de la medida

En variados problemas, el interés es determinar una cantidad (por ejemplo, d una distancia entre dos puntos, desconocida), para lo cual se toman varias mediciones $d_1, d_2, \ldots, d_n$.

Con base a lo visto, d_1, d_2,..., d_n se pueden considerar valores de variables aleatorias independientes D_1,..., D_n.

Por lo tanto el interés esta en:

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

con valor esperado δ.

Teoría de la medida

Por tanto, la media muestral es

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

y la varianza muestral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$$

Luego, un intervalo de confianza al $(1-\alpha)100\%$ para δ esta dado por

$$< \delta >_{1-\alpha} \in \overline{d} \pm t_{1-\alpha/2}(n-1)\cdot\frac{S}{\sqrt{n}}$$