EYP 1027 Modelos Probabilísticos

Clase 19

Profesor: Reinaldo B. Arellano-Valle

Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2020

Contenido I

Muestras aleatorias y distribuciones muestrales

Conceptos básicos

Como ya sabemos, una distribución de probabilidad, también llamada distribución poblacional (o simplemente población), describe el compartamiento probabilístico de una determinada variable aleatoria.

También sabemos que toda distribución de probabilidad o población puede ser descrita o representada por su fda (F) o bien por su fmp (f) en el caso discreto o fdp (f) en el caso continuo.

Normalmente no es posible conocer toda la población, por lo sólo disponemos de una muestra representativa de dicha población.

A partir de la muestra, podemos inferir sobre caraterísticas poblacionales, llamadas parámetros, que nos interesa conocer.

Para ello, se requiere saber primero la distribución de probabilidad (exacta o asintótica) de las versiones muestrales de tales caracteristicas, las cuales son denominadas estadísticos.

Muestras aleatorias y distribuciones muestrales

Definición 1.1

Se dice que $X_1, \ldots, X_n$ es una muestra aleatoria de tamaño n (ma(n)) de una población con fmp o fdp f si $X_1, \ldots, X_n$ son variables aleatorias mutuamente independientes e identicamente distribuidas (iid) f

Como es obvio, si $X_1, \ldots, X_n$ es una ma(n) de f, entonces la fmp o fdp conjunta de la ma(n) es,

$$f\left(x_1, \ldots, x_n\right) = \prod_{i=1}^{n} f\left(x_i\right).$$

Cuando la ma(n) proviene de una familia paramétrica $f(x;\theta)$, entonces la fmp o fdp conjunta de $X_1, \ldots, X_n$ es,

$$f\left(x_1, \ldots, x_n;\theta\right) = \prod_{i=1}^{n} f\left(x_i;\theta\right),$$

donde el parámetro θ (posiblemente vectorial) etiqueta la fmp o fdp.

Ejemplo 1.1

Sea $X_1, \ldots, X_n$ una ma(n) de una población $\exp(\lambda)$, donde $\lambda > 0$ es el parámetro asociado. Por ejemplo, $X_1, \ldots, X_n$ podrían corresponder a los tiempos de vida (medidos en años y fracciones) de n artefactos electrónicos idénticos, los que se hacen funcionar hasta que fallen. La fdp conjunta de la ma(n) es,

$$\begin{aligned} f\left(x_1, \ldots, x_n; \lambda\right) &= \prod_{i=1}^{n} f\left(x_i; \lambda\right) \\ &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}, \quad x_i > 0, \quad i = 1, \ldots, n. \end{aligned}$$

Con esta fdp se pueden estudiar diferentes propidades de la muestra; por ejemplo, para calcular la probabilidad de que todos los artefactos duren más de 2 años,

$$P(X_1 > 2, \ldots, X_n > 2) = \prod_{1=1}^{n} P(X_i > 2) = \exp\{-2n\lambda\}.$$

A partir de la fda conjunta de la ma(n) también se puede demostrar que la variable aleatoria $\sum_{i=1}^{n} X_i$ tiene distribucion $Gama(n, \lambda)$, y por tanto que

$$2\lambda \sum_{i=1}^{n} X_i \sim \chi^2_{2n}.$$

Este resultado juega un rol importante cuando se desea inferir sobre el verdadero valor del parámetro λ, es decir, aquel valor efectivamente generó la ma(n) observada.

Muestras aleatorias y distribuciones muestrales

Definición 1.2

Sea $T(x_1, \ldots, x_n)$ una función de valor real o bien vectorial, y sea $X_1, \ldots, X_n$ una ma(n) de una población. Entonces la variable o vector aleatoria(o) $T = T(X_1, \ldots, X_n)$ se denomina estadístico. La distribución de probabilidad de un estadístico T se llama distribución muestral de T.

La definición de un estadístico es muy amplia; cualquier función de la ma(n) que no dependa de ningún parámetro desconocido es un estadístico. Por ejemplo, un estadístico puede ser la propia ma(n), el mínimo o el máximo de de la muestra, el promedio de la muestra o alguna medida de la variabilidad de las observaciones de la muestra.

Habitualmente, los parámetros poblacionales de mayor interés son la media μ y la varianza σ^2 de la población.

Por ende, al disponer de una ma(n) de dicha población, es natural preocuparse de la media $\bar{X}$ y la varianza S^2 de la muestra.

Definición 1.3

Dada una muestra aleatoria $X_1, \ldots, X_n$ de una población f, dos estadísticos de inteés inmediato son:

(a) La media muestral, definida como,

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

(b) La varianza muestral, definida como,

$$S^2 = \frac{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

También se define la desviación estándar muestral como, $S = \sqrt{S^2}$.

Muestras aleatorias y distribuciones muestrales

Teorema 1.1

Para cualquier secuencia de números reales $x_1, \ldots, x_n$, sea $\bar{x} = \sum_{i=1}^{n} x_i/n$. Entonces,

i) $\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$ para todo $\mu \in \mathbb{R}$;

ii) $\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$, es decir, para $\hat{\mu} = \bar{x}$ la distancia $\sum_{i=1}^{n} (x_i - \mu)^2$ es mínima;

iii) $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})\, x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Tarea: Pruebe que $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Demostración 1.1

La demostración de ii) es directa de la parte i). Las demostraciones de i) y iii) quedan de ejercio.

Lema 1.1

Sea $X_1, \ldots, X_n$ una ma(n) de f, y sea g una fución real valorada. Entonces,

i) $g(X_1), \ldots, g(X_n)$ son variables aleatorias iid;

ii) $\mathrm{E}\left\{\sum_{i=1}^{n} g\left(X_i\right)\right\} = n\mathrm{E}\left\{g\left(X_1\right)\right\}$ provisto que la esperanza exista;

iii) $\mathrm{Var}\left\{\sum_{i=1}^{n} g\left(X_i\right)\right\} = n\mathrm{Var}\left\{g\left(X_1\right)\right\}$ provisto que la varianza exista.

Demostración 1.2

Resultados ya probados en un contexto más general.

Teorema 1.2

Sea $X_1, \ldots, X_n$ una ma(n) desde una población con media μ y varianza σ^2 finitas. Entonces,

i) $\mathrm{E}(\bar{X}) = \mu$;
ii) $\mathrm{Var}(\bar{X}) = \sigma^2/n$;
iii) $\mathrm{E}(S^2) = \sigma^2$.

Conclusión: La media muestral $\bar{X}$ y la varianza muestral S^2 de una ma(n) son estimadores insesgados, respectivamente, de la media poblacional μ la varianza poblacional σ^2.

Demostración 1.3

i) y ii) ya fuerón probados, por lo que sólo probaremos iii). En efecto, note primero que

$$
\begin{aligned}
\mathrm{E}\{(n-1)S^2\} &= \mathrm{E}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\
&= \sum_{i=1}^{n} \mathrm{E}\left(X_i^2\right) - n\mathrm{E}\left(\bar{X}^2\right) \\
&= n\mathrm{E}X_1^2 - n\mathrm{E}\bar{X}^2 \\
&= n\left(\sigma^2+\mu^2\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right) = (n-1)\sigma^2,
\end{aligned}
$$

de donde se obtiene el resultado.

Teorema 1.3

Sea $X_1, \ldots, X_n$ una ma(n) desde una población con fgm $M(t)$. Entonces la fgm de la media muestral es,

$$M_{\bar{X}}(t) = \{M(t/n)\}^n$$

Demostración 1.4

Tenemos que,

$$M_{\bar{X}}(t) = \mathrm{E}\left(e^{t\bar{X}}\right) = \mathrm{E}\left(\prod_{i=1}^{n} e^{tX_i/n}\right) = \prod_{i=1}^{n} M_{X_i}(t/n) = \{M_X(t/n)\}^n$$

El resultado anterior proporciona una forma muy simple para obtener la distribución muestral de $\bar{X}$ cuando $M_{\bar{X}}(t)$ es una fgm conocida, como se ilustra en el siguiente ejemplo.

Ejemplo 1.2

Distribución de la media Sea $X_1, \ldots, X_n$ una ma(n) desde una población $N(\mu, \sigma^2)$. Entonces la fgm de la media muestral $\bar{X}$ es,

$$
\begin{aligned}
M_{\bar{X}}(t) &= \left\{ \exp\left(\mu \frac{t}{n} + \frac{\sigma^2 (t/n)^2}{2} \right) \right\}^n \\
&= \exp\left\{ n \left(\mu \frac{t}{n} + \frac{\sigma^2 (t/n)^2}{2} \right) \right\} = \exp\left\{ \mu t + \frac{\left(\sigma^2/n\right) t^2}{2} \right\}.
\end{aligned}
$$

Así, si la ma(n) proviene de una población $N(\mu, \sigma^2)$, entonces la distribución muestral de $\bar{X}$ es $N(\mu, \sigma^2/n)$.

Ejemplo 1.3

Se tiene una máquina de llenado para vaciar 500gr de cereal en una caja de cartón. Suponga que la cantidad de cereal que se coloca en cada caja es una variable aleatoria normalmente distribuida con media 500gr y desviación estándar igual a 20gr.

Para verificar que el peso promedio de cada caja se mantiene en 500gr se toma una muestra aleatoria de 25 de éstas cajas en forma periódica y se pesa el contenido de cada una de ellas.

El gerente de la planta ha decidido detener el proceso y encontrar la falla cada vez que el valor promedio de la muestra sea mayor de 510gr o menor de 490gr. Obtenga la probabilidad de detener el proceso.

Sea $X_1, X_2, \ldots, X_{25}$ una ma(25), las cuales representan la cantidad de cereal contenido en las cajas de una muestra aleatoria dada.

Por hipótesis $X_i \sim N(500, 20^2),\ i = 1, 2, \ldots, 25$. Luego $\bar{X} \sim N(500, 20^2/25)$.

La probabilidad deseada es igual a uno menos la probabilidad de que $\bar{X}$ se encuentre entre 490 y 510gr.

$$\begin{aligned} P(\text{ Detención del proceso}) &= 1 - P(490 < \bar{X} < 510) \\ &= 1 - P\left(\frac{490-500}{4} < Z < \frac{510-500}{4}\right) \\ &= 1 - P(-2.5 < Z < 2.5) \\ &= 0.0124. \end{aligned}$$

Teorema 1.4

Sea $X_1, \ldots, X_n$ una ma(n) desde una distribución $N(\mu, \sigma^2)$, y sea $\bar{X} = \sum_{i=1}^n X_i/n$ y $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$. Entonces,

i) $\bar{X}$ y S^2 son variables aleatorias independientes;

ii) $\bar{X}$ tiene una distribución $N(\mu, \sigma^2/n)$;

iii) $(n-1)S^2/\sigma^2$ distribuye chicuadrado con $n-1$ grados de libertad.

Demostración 1.5

Para probar i) note primero que $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$, de modo que $X_1 - \bar{X} = -\sum_{i=2}^{n}(X_i - \bar{X})$. Luego,

$$
\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \\
&= \frac{1}{n-1}\left\{\left(X_1 - \bar{X}\right)^2 + \sum_{i=2}^{n}\left(X_i - \bar{X}\right)^2\right\} \\
&= \frac{1}{n-1}\left\{\left(\sum_{i=2}^{n}\left(X_i - \bar{X}\right)\right)^2 + \sum_{i=2}^{n}\left(X_i - \bar{X}\right)^2\right\}.
\end{aligned}
$$

Es decir, S^2 es una función de $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$ solamente. Entonce, basta mostrar que estas variables son son independientes de $\bar{X}$.

Suponga (sin pérdida de generalidad) que $\mu = 0$ y $\sigma = 1$. Entonces, la fdp conjunta de la muestra $X_1, \ldots, X_n$, está dada por,

$$f(x_1, \ldots, x_n) = (2\pi)^{-n/2} e^{-\sum_{i=1}^n x_i^2/2}, \quad -\infty < x_i < \infty.$$

Considere la transformación uno-a-uno dada por,

$$Y_1 = \bar{X},\ Y_2 = X_2 - \bar{X}, ..., Y_n = X_n - \bar{X}.$$

Esta es una transformación lineal en $(X_1, \ldots, X_n)$ con un jacobiano igual a $1/n$.

Luego,

$$\begin{aligned} f\left(y_1,\ldots,y_n\right) &= n(2\pi)^{-n/2} e^{-\left(y_1-\sum_{i=2}^n y_i\right)^2/2} \\ &\quad \times e^{-\sum_{i=2}^n (y_i+y_1)^2/2}, \quad -\infty < y_i < \infty, \quad i=1,\ldots,n, \\ &= (n/(2\pi))^{1/2}\, e^{\left(-ny_1^2\right)/2} \\ &\quad \times n^{1/2}(2\pi)^{-(n-1)/2} e^{-\left\{\sum_{i=2}^n y_i^2+\left(\sum_{i=2}^n y_i\right)^2\right\}/2}, \\ &\qquad\qquad -\infty < y_i < \infty, \quad i=1,\ldots,n. \end{aligned}$$

Dado que la fdp conjunta de $Y_1,\ldots,Y_n$ se puede escribir como el producto de las marginales, se deduce que $Y_1=\bar{X}$ es independiente de $Y_2=X_2-\bar{X},\ldots,Y_n=X_n-\bar{X}$ y, por lo tanto, que $\bar{X}$ es independiente de S^2.

Para probar iii), recuerde que

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 = \sum_{i=1}^{n}(X_i - \mu)^2;$$

es decir,

$$(n-1)S^2 + n(\bar{X} - \mu)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 .$$

Al dividir ambos miembros de la expresión anterior por la varianza poblacional σ^2, se tiene

$$\frac{(n-1)S^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2}$$

o

$$\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 = \sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^2}. \quad (*)$$

Muestras aleatorias y distribuciones muestrales

Pero $(X_i - \mu)/\sigma \overset{\text{iid}}{\sim} N(0, 1)$, de modo que $(X_i - \mu)^2/\sigma^2 \overset{\text{iid}}{\sim} \chi^2_1$ y, por tanto, $\sum_{i=1}^{n}(X_i - \mu)^2/\sigma^2 \sim \chi^2_n$.

Similar, como $\bar{X} \sim N(\mu, \sigma^2/n)$, se tiene que $\{(\bar{X} - \mu)/\sigma/\sqrt{n}\}^2 \sim \chi^2_1$.

Pero como fue probado, $(n-1)S^2/\sigma^2$ y $[(\bar{X} - \mu)/\sigma/\sqrt{n}]^2$ son variables aleatorias independientes, entonces, de $(*)$ se concluye que $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

La distribución $t-$Student Sean X e Y variables aleatorias independientes tales que $X \sim N(0,1)$ e $Y \sim \chi^2_{(k)}$. Considere la transformación

$$g(x,y) = \left(x/\sqrt{y/k}, y\right)$$

La transformación inversa está dada por,

$$h(x,y) = (x\sqrt{y/k}, y),$$

mientras que el Jacobiano es,

$$J(x,y) = \sqrt{y/k}.$$

Por lo tanto, la fdp conjunta de $Z = X/\sqrt{Y/k}$ y $W = Y$ esta dada por,

$$f_{Z,W}(z,w) = \sqrt{w/k} f_{X,Y}\left(z\sqrt{w/k}, w\right), \quad x \in \mathbb{R}, \quad w > 0,$$

donde,
$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{(1/2)^{k/2}}{\Gamma(k/2)} y^{k/2-1} e^{-y/2}, \quad x \in \mathbb{R}, \quad y > 0.$

Integrando $f_{Z,W}(z,w)$ con respecto a w, se obtiene la fdp marginal de Z, es decir,

$$f_Z(z) = \frac{\Gamma\left((k+1)/2\right)}{\sqrt{k\pi}\,\Gamma\left(k/2\right)} \left(1 + \frac{z^2}{k}\right)^{-(k+1)/2}, \quad z \in \mathbb{R},$$

la cual corresponde a la fdp de una distribución $t-$Student con k grados de libertad.

Notación: $Z \sim t_k$.

La distribución F Sean X e Y variables aleatorias independientes tales que $X \sim \chi^2_m$ e $Y \sim \chi^2_n$. Considere la transformación,

$$g(x, y) = \left(\frac{x/m}{y/n}, y \right)$$

La función inversa de g está dada por,

$$h(x, y) = \left(\frac{m}{n} xy, y \right),$$

y tiene un Jacobiano igual a $J(x, y) = \dfrac{m}{n} y$. Luego, la fdp conjunta de $Z = \frac{nX}{mY}$ y $W = Y$ está dada por,

$$f_{Z,W}(z,w) =$$
$$\frac{(1/2)^{m/2}\,(1/2)^{n/2}}{\Gamma\left(m/2\right)\Gamma\left(n/2\right)} \left(\frac{m}{n} zw\right)^{\frac{m}{2}-1} w^{\frac{n}{2}-1} \mathrm{e}^{\left\{-\frac{1}{2}w\left(\frac{m}{n}z+1\right)\right\}},$$
$$z > 0, w > 0$$

Integrando con respecto a w encontramos la fdp de Z, dada por la siguiente expresión,

$$f_Z(z) = \frac{\Gamma\left((m+n)/2\right)}{\Gamma\left(m/2\right)\Gamma\left(n/2\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1} \left(1 + \frac{m}{n} z\right)^{-(m+n)/2}, \quad z > 0,$$

la cual corresponde a la fdp de una distribución F con m gl en el numerador y n gl en el denominador.

Notación: $Z \sim F_{m,n}$.

Teorema 1.5

Sea $X_1, \ldots, X_n$ una ma(n) desde una distribución $N(\mu, \sigma^2)$. Entonces la variable aleatoria $(\bar{X} - \mu)/(S/\sqrt{n})$ tiene una distribución $t-$Student con $n-1$ grados de libertad.

Demostración 1.6

Basta notar que $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)$ y que es independiente de $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Luego la variable aleatoria $T = (\bar{X} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$. Además,

$$
\begin{aligned}
\mathrm{E}(T) &= 0, \qquad \text{si } n-1 > 1 \\
\mathrm{Var}(T) &= \frac{n-1}{n-2}, \qquad \text{si } n-2 > 2
\end{aligned}
$$

Teorema 1.6

Sea $X_1, \ldots, X_n$ una ma(n) de una población $N(\mu_X, \sigma_X^2)$, y sea $Y_1, \ldots, Y_m$ una ma(m) una población independiente $N(\mu_Y, \sigma_Y^2)$. La variable aleatoria $F = \left(S_X^2/\sigma_X^2\right) / \left(S_Y^2/\sigma_Y^2\right)$ tiene una distribución F con $n-1$ y $m-1$ grados de libertad.

Demostración 1.7

Notemos que $(n-1)S_X^2/\sigma_X^2 \sim \chi_{n-1}^2$ y que es independiente de $(m-1)S_Y^2/\sigma_Y^2 \sim \chi_{m-1}^2$.

Luego la variable aleatoria $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$. Además,

$$\mathrm{E}(F) = \frac{m-1}{m-3}, \quad \text{si } m-3 > 1$$

Muestras aleatorias y distribuciones muestrales

Teorema 1.7

i) Si $X \sim F_{p,q}$, entonces $1/X \sim F_{q,p}$; esto es, el recíproco de una variable aleatoria F también tiene una distribución F.

ii) Si $X \sim t_q$, entonces $X^2 \sim F_{1,q}$

iii) Si $X \sim F_{p,q}$, entonces $(p/q)X/(1+(p/q)X) \sim \text{Beta}(p/2, q/2)$.

Problema: Qué hacemos si no tenemos normalidad?

References

Blanco, L., Arunachalam, V. y Dharmaraja, S. (2012). *Introduction to Probability and Stochastic Processes.* John Wiley, New Jersey.

Casella, G. y Berger, R.L. (2002). *Statistical Inference.* Second Edition. Duxbury, California.

DeGroot, M. y Schervish, M.J. (2012). *Probability and statistics.* Fourth Edition. Addison-Wesley, New York.