Distribución en el muestreo

Contenido

$X \sim F$

F Binomial

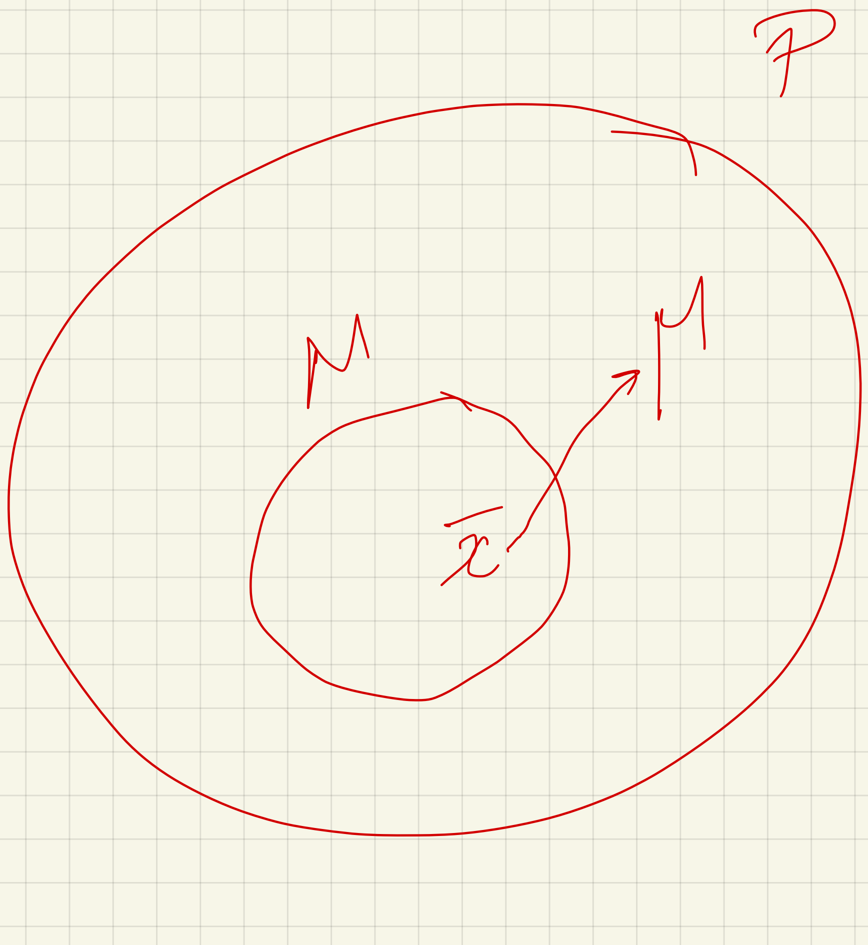
F Poisson

F Exponencial

F Normal

$$P(X \le x) = F(x)$$

$$\boxed{x_1, \; x_2, \dots, x_n} \longrightarrow F = ?$$



Introducción

Trabajaremos con funciones de variables aleatorias $Y_1, \cdots, Y_n$, observadas en una muestra aleatoria seleccionada de una población de interés. Estas son idénticamente distribuidas e independientes.

Supongamos que deseamos estimar una media poblacional μ. Si observamos una muestra aleatoria de n observaciones $y_1, \cdots, y_n$ parecerá razonable estimar μ con la media muestral

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$\overline{Y}$ es una variable aleatoria (*v.a*) definida como

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

es decir, $\overline{Y}$ está en función de las observaciones de la muestra de tamaño n.

$\overline{Y}$ es una función de (sólo) las variables aleatorias $Y_1, \cdots, Y_n$ y del tamaño de la muestra n, por lo tanto la v.a $\overline{Y}$ es un ejemplo de estadístico.

Definición

Un **estadístico** es una función de las variables aleatorias observables en una muestra y constantes conocidas.

Ejemplos de estadísticos

Variables aleatorias.

$$\bar{Y}$$

$$S^2$$

$$Y_{(n)} = \max(Y_1, Y_2, \dots, Y_n)$$

$$Y_{(1)} = \min(Y_1, Y_2, \dots, Y_n)$$

$$R = Y_{(n)} - Y_{(1)}$$

Mediana

$$\hat{P}$$

Como todos los estadísticos son funciones de v.a’s, entonces ellos mismos son v.a’s, que tienen distribuciones de probabilidad y que llamaremos **sus distribuciones muestrales**.

Desde un punto de vista más practico, la distribución muestral de un estadístico proporciona un modelo teórico para el histograma de frecuencias relativas de los posibles valores del estadístico que observamos por medio de muestreo repetido.

Experimento

Lanzar un dado tres veces.

$$\bar{Y} = \frac{1}{3}\sum_{i=1}^{3} Y_i$$

$$= \frac{1}{3}(Y_1 + Y_2 + Y_3)$$

$Y_1 \quad Y_2 \quad Y_3$

$\downarrow \quad \downarrow \quad \downarrow$

$6 \quad 6 \quad 6 \qquad |S| = 6^3 = 216$

X : Suma de las caras de los tres dados.

$x \in \{3, 4, \ldots, 18\}$

Y_1, Y_2, Y_3 son igualmente distribuidos e independientes con distribución uniforme discreta

Distribución uniforme discreta

$Y_i = y$	1	2	3	4	5	6
$P(Y_i = y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$Y_i \sim U(1, 2, 3, 4, 5, 6)$

con $i = 1, 2, 3.$

$$E(Y_i) = \sum_{i=1}^{6} y_i \, P(Y_i = y_i)$$

$$= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right)$$

$$= \frac{1}{6}\left[1 + 2 + 3 + 4 + 5 + 6\right] = \frac{21}{6} = 3.5$$

$$V(Y_i) = E(Y_i^2) - \left[E(Y_i)\right]^2$$

$$E(Y_i^2) = \frac{1}{6}\left[1+4+9+16+25+36\right]$$

$$= \frac{1}{6}(91) = 15.1\overline{6}$$

$$Var(Y_i) = 15.1\overline{6} - (3.5)^2$$

$$= 15.1\overline{6} - 12.25$$

$$= 2.91\overline{6}$$

Entonces vamos a caracterizar el estadístico $\overline{Y}$

1. ¿Cuál es la distribución de $\overline{Y}$?
2. ¿$E(\overline{Y})$?

3 ¿ $Var(\overline{Y})$?

R/. 2

$$E(\overline{Y}) = E\left(\frac{1}{3}\sum_{i=1}^{3} Y_i\right) = \frac{1}{3}E\left(\sum_{i=1}^{3} Y_i\right)$$

$$= \frac{1}{3}\sum_{i=1}^{3} E(Y_i) = \frac{1}{3}\sum_{i=1}^{3} 3.5$$

$$= \frac{\cancel{3}(3.5)}{\cancel{3}} = 3.5.$$

R/. 3

$$\operatorname{Var}\left(\frac{1}{3}\sum_{i=1}^{3} Y_i\right)$$

$$= \left(\frac{1}{3}\right)^2 \operatorname{Var}\left(\sum_{i=1}^{3} Y_i\right)$$

$$\overset{iid}{=} \frac{1}{9}\sum_{i=1}^{3} \operatorname{Var}(Y_i)$$

$$= \frac{1}{9}\sum_{i=1}^{3} 2.91\overline{6}$$

$$= \frac{\cancel{3}(2.91\overline{6})}{\cancel{9}} = \frac{2.91\overline{6}}{3} = 0.9722$$

$$\sigma_{\bar{Y}} = \sqrt{\frac{Var(Y_i)}{3}} = \frac{\sigma_{Y_i}}{\sqrt{3}} = 0.9860.$$

1 ¿Cómo podemos deducir la distribución de la v.a $\bar{Y}$?

Los posibles valores de la v.a $X = Y_1 + Y_2 + Y_3$ son (Julian) $3, 4, 5, \ldots, 18$ y $\bar{Y} = \frac{X}{3}$

X	$\bar{Y} = \frac{X}{3}$	$P(\bar{Y} = \bar{y})$
3	1	1/216
4	4/3	3/216
5	5/3	6/216
6	2	
7	7/3	
8	8/3	
9	3	
10	10/3	
11	11/3	
12	4	
13	13/3	
14	14/3	
15	5	
16	16/3	
17	17/3	
18	6	

$$P(\bar{Y} = 1) = P(X = 3) = P(Y_1 = 1, Y_2 = 1, Y_3 = 1)$$
$$= \frac{1}{216}$$

$$P\left(\bar{Y} = \frac{4}{3}\right) = P(X = 4) =$$

$$P\left(\overline{Y}=\frac{4}{3}\right) = P(X=4)$$

$$= P(Y_1=1, Y_2=1, Y_3=2)$$

$$+ P(Y_1=1, Y_2=2, Y_3=1)$$

$$+ P(Y_1=2, Y_2=1, Y_3=1)$$

$$= \frac{1}{216} + \frac{1}{216} + \frac{1}{216} = \frac{3}{216}$$

$$P\left(\overline{Y}=\frac{5}{3}\right) = P(x=5)$$

$$= P(Y_1=2, Y_2=2, Y_3=1$$

$$+ P(Y_1=1, Y_2=2, Y_3=2)$$

$$+ P(Y_1=2, Y_2=1, Y_3=2)$$

$$+ P(Y_1=3, Y_2=1, Y_3=1)$$

$$+ P(Y_1=1, Y_2=3, Y_3=1)$$

$$+ P(Y_1=1, Y_2=1, Y_3=3)$$

$$= \frac{1}{216} + \frac{1}{216} + \frac{1}{216} + \frac{1}{216} + \frac{1}{216} + \frac{1}{216}$$

$$= \frac{6}{216}$$

las probabilidades $P\left(\overline{Y}=\frac{i}{3}\right)$, $i=6,7,\dots,18$ se obtienen de manera análoga $\blacksquare$

Teorema 7.1 (Según estadística matemática con aplicaciones)

Sea $Y_1, Y_2, \cdots, Y_n$ una muestra aleatoria de tamaño n de una distribución normal con media μ y varianza σ^2. Entonces

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

está distribuida normalmente con media $\mu_{\overline{Y}} = \mu$ y varianza $\sigma^2_{\overline{Y}} = \sigma^2/n$.

Ejemplo 1:

Una máquina embotelladora puede ser regulada para que descargue un promedio de μ onzas por botella. Se ha observado que la cantidad de líquido dosificado por la maquina está distribuida normalmente con $\sigma = 1{,}0$ onza. Una muestra de $n = 9$ botellas se selecciona aleatoriamente de la producción de la máquina en un día determinado (todas embotelladas con el mismo ajuste de la máquina) y las onzas de contenido líquido se mide para cada una. Determine la probabilidad de que la media muestral se encuentre a no más de 0.3 onza de la verdadera μ para el ajuste seleccionado de la máquina.

$Y_1, Y_2, \ldots, Y_9$ denota el contenido en onzas de las botellas que se van a observar, entonces sabemos que

$$Y_i \sim N(\mu, \sigma^2 = 1), \quad i = 1, \ldots, 9.$$

Por el teorema 7.1, $\overline{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i$ se distribuye también normal pero con parámetros.

$$\mu_{\overline{Y}} = \mu_{Y_i} = \mu \quad y \quad \sigma^2_{\overline{Y}} = \frac{\sigma^2_{Y_i}}{9} = \frac{1}{9}$$

$\sigma_{\overline{Y}} = \sqrt{\frac{1}{9}} = \frac{1}{3}$ entonces nos interesa

$$P(|\bar{Y}-\mu| \leq 0.3) = P(-0.3 \leq \bar{Y}-\mu \leq 0.3)$$

$$= P\left(\frac{-0.3}{\sigma_{\bar{Y}}} \leq \frac{\bar{Y}-\mu}{\sigma_{\bar{Y}}} \leq \frac{0.3}{\sigma_{\bar{Y}}}\right)$$

$$= P\left(\frac{-0.3}{\frac{1}{3}} \leq Z \leq \frac{0.3}{\frac{1}{3}}\right)$$

$$= P((-0.3)(3) \leq Z \leq 3(0.3))$$

$$= P(-0.9 \leq Z \leq 0.9)$$

$$= pnorm(0.9) - pnorm(-0.9)$$

$$= 0.6318$$

Ejemplo 2

Bajo el mismo contexto del ejemplo , determinar ¿Cuántas observaciones deben estar incluídas en la muestra si deseamos que $\overline{Y}$ se encuentre a no más de 0.3 onzas de μ con probabilidad 0.95?

$$P(|\bar{Y}-\mu| < 0.3) = 0.95$$

$$Y_i \sim N(\mu, 1) \qquad \bar{Y} \sim N(\mu_{\bar{Y}}, \sigma_{\bar{Y}})$$

$$\mu_{\bar{Y}} = \mu \qquad \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}} = \frac{1}{\sqrt{n}}$$

1. $P(-0.3 < \bar{Y}-\mu < 0.3) = 0.95$

2. $P\left(-\frac{0.3}{\sigma_{\bar{Y}}} < \frac{\bar{Y}-\mu}{\sigma_{\bar{Y}}} < \frac{0.3}{\sigma_{\bar{Y}}}\right) = 0.95$

3 $P\left(-\frac{0.3}{\frac{1}{\sqrt{n}}} < Z < \frac{0.3}{\frac{1}{\sqrt{n}}}\right) = 0.95$

4. $P(-0.3\sqrt{n} < Z < 0.3\sqrt{n}) = 0.95$

5. $P(-0.3\sqrt{n} < Z < 0.3\sqrt{n}) = 0.95$

6. $F(0.3\sqrt{n}) - F(-0.3\sqrt{n}) = 0.95$

7. $1 - 2F(-0.3\sqrt{n}) = 0.95$

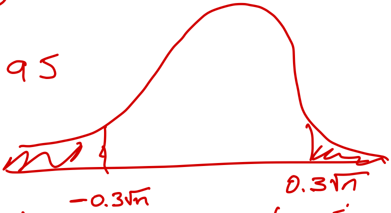
8. $1 - 0.95 = 2F(-0.3\sqrt{n})$

$0.05 = 2F(-0.3\sqrt{n})$

$\frac{0.05}{2} = F(-0.3\sqrt{n})$

$0.025 = F(-0.3\sqrt{n})$

$qnorm(0.025) = -0.3\sqrt{n}$



$\sqrt{n} = \frac{qnorm(0.025)}{-0.3}$

$\sqrt{n} = \frac{-1.9599}{-0.3}$

$\sqrt{n} = 6.533$

$n = 42.68 \approx 43$

Teorema del límite central:

Sean $Y_1, Y_2, \cdots, Y_n$ variables aleatorias independientes y distribuidas idénticamente con $E(Y_i) = \mu$ y $V(Y_i) = \sigma^2 < \infty$. Definamos

$$U_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \text{ donde } \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Entonces la función de distribución de U_n converge hacia la función de distribución normal estándar cuando $n \to \infty$. Esto es,

$$\lim_{n\to\infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ para toda } u.$$

$$U_n = \frac{\bar{Y}_n - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{\frac{\sum_{i=1}^{n} Y_i - n\mu}{n}}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{n\sigma}$$

$$= \frac{\sqrt{n}}{n} \frac{\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{\sigma} = \frac{n^{1/2}}{n} \frac{\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{\sigma}$$

$$= \frac{1}{\sqrt{n}} \frac{\left(\sum_{i=1}^{n} Y_i - n\mu\right)}{\sigma} = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}}$$

Nota $\sum_{i=1}^{n} Y_i \sim N(n\mu, \sigma\sqrt{n})$

$$\overline{Y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$Y_1, Y_2, \ldots, Y_n$ iid , $E(Y_i) = \mu$ y

$Var(Y_i) = \sigma^2$

Ejemplo

Las calificaciones de exámenes para todos los estudiantes de último año de preparatoria, en cierto estado, tienen media de 60 y varianza de 64. una muestra aleatoria de $n = 100$ estudiantes de una escuela preparatoria grande tuvo una calificación media de 58. ¿Hay evidencia para sugerir que en nivel de conocimiento de esta escuela sea inferior? (Calcule la probabilidad de que la media muestral sea 58 cuando $n = 100$.)

X_i Calificación promedio de la escuela

$\overline{X}$ Calificación promedio del estado.

$E(X_i) = 60$ $n = 100$

$\overline{X} \sim N\left(60, \frac{8}{\sqrt{100}}\right)$ Promedio observado es 58

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ donde X_i es lo observado en el i-ésimo colegio del estado.

$X_i \sim F$? $E(X_i) = 60$ $Var(X_i) = 64$

$$P(\overline{X} \leq 58) = P\left(\frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} \leq \frac{58 - \mu_{\overline{x}}}{\sigma_{\overline{x}}}\right)$$

$$P\left(Z \le \frac{\frac{58-60}{8}}{10}\right)$$

$$= P\left(Z \le \frac{-2 \cdot 10}{8}\right)$$

$$= P(Z \le -2.5)$$

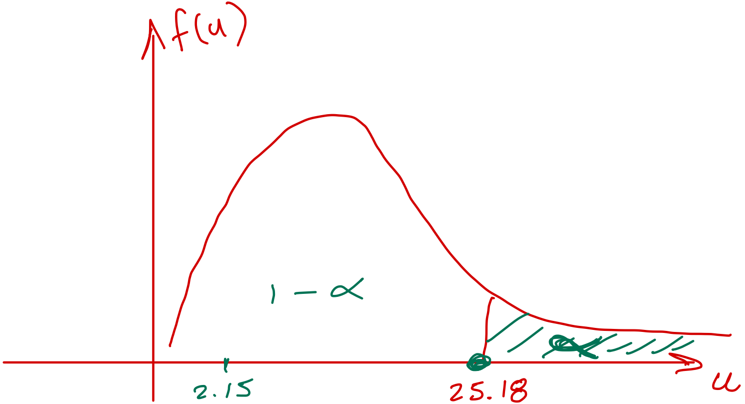
$$= \text{pnorm}(-2.5) = 0.0062.$$

Teorema 7.2

Si $Y_1,\ Y_2, \cdots,\ Y_n$ está definida como en el Teorema 7.1. Entonces $Z_i = (Y_i - \mu)/\sigma$ son variables aleatorias normales estándar e independientes, $i = 1, 2, \cdots, n$, y

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

tienen una distribución χ^2 con n grados de libertad (gl).



$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

implica que $$P(\chi^2 \leq \chi^2_\alpha) = 1 - \alpha$$

$$P(\chi^2 > \chi^2_\alpha) = \alpha \textbf{ implica que } P(\chi^2 \leq \chi^2_\alpha) = 1 - \alpha$$

qchiq(0.995,10)

= 25.18

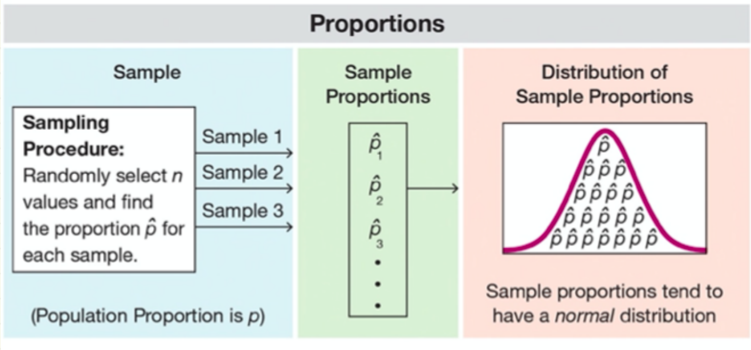
qchiq(0.995,10,
lower.tail=T)

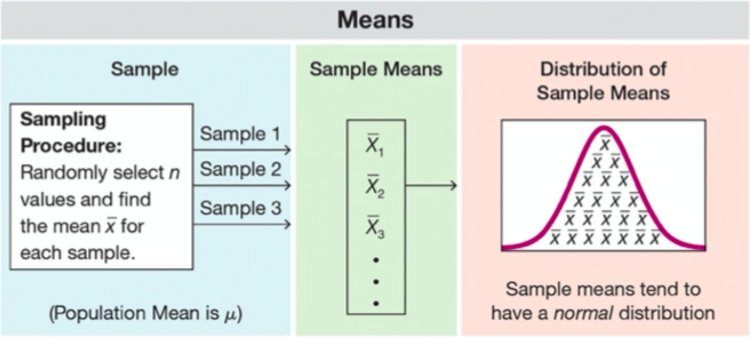
= 25.18

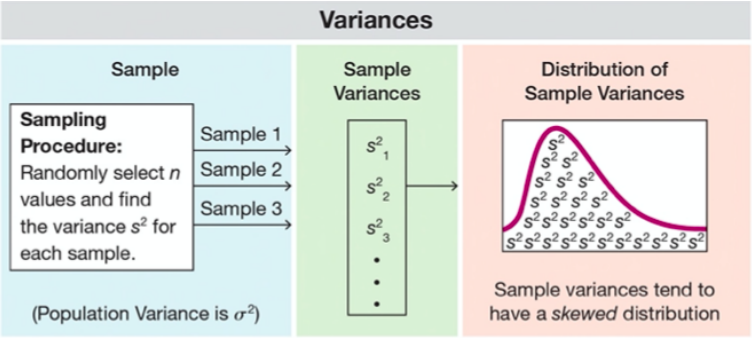
qchisq(0.995, 10,
lower.tail = F)

= 2.15

α	$1-\alpha$	$\chi^2_\alpha = p_{1-\alpha} = \phi_{1-\alpha}$
0.005	0.995	$p_{0.995} = 25.18818$
0.01	0.99	$p_{0.99} = 23.20928$
0.025	0.975	$p_{0.975} = 20.48318$
0.05	0.95	$p_{0.95} = 18.30704$
0.1	0.9	$p_{0.9} = 15.98718$
0.9	0.1	$p_{0.1} = 4.86518$
0.95	0.05	$p_{0.05} = 3.9403$
0.975	0.025	$p_{0.025} = 3.24697$
0.99	0.01	$p_{0.01} = 2.55821$
0.995	0.005	$p_{0.005} = 2.15585$


Proportions
Sample
Sampling Procedure:
Randomly select n values and find the proportion p̂ for each sample.
Sample 1
Sample 2
Sample 3
(Population Proportion is p)
Sample Proportions
p̂1
p̂2
p̂3
Distribution of Sample Proportions
Sample proportions tend to have a normal distribution


Means
Sample
Sampling Procedure:
Randomly select n values and find the mean x̄ for each sample.
Sample 1
Sample 2
Sample 3
(Population Mean is μ)
Sample Means
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
Distribution of Sample Means
Sample means tend to have a normal distribution


Variances
Sample
Sampling Procedure:
Randomly select n values and find the variance s² for each sample.
Sample 1
Sample 2
Sample 3
(Population Variance is σ²)
Sample Variances
Distribution of Sample Variances
Sample variances tend to have a skewed distribution

$$
\begin{array}{cccc}
Y_1 & Y_2 & \cdots & Y_{10000} \\
1 = y_1^{(1)} & 0 & & 1 \\
0 = y_1^{(2)} & 0 & & 0 \\
1 = y_1^{(3)} & 0 & & 1 \\
0 & 1 & & 0 \\
\vdots & \vdots & & \vdots \\
0 & 0 & & 0 \\
\hline
\hat{p}_1 & \hat{p}_2 & & \hat{p}_{1000}
\end{array}
$$

100

$$
\hat{p}_1 = \frac{\sum_{i=1}^{100} y_1^{(i)}}{100} \qquad \hat{p}_2 = \frac{\sum_{i=1}^{100} y_2^{(i)}}{100} \qquad \cdots \qquad \hat{p}_{10000}
$$

$$X_1 \quad X_2 \quad \cdots \quad X_{10000} \sim Bin(n, p)$$

$$20 \quad 80 \quad \quad 35$$

$$\hat{p}_1 = \frac{20}{100} \quad \hat{p}_2 = \frac{80}{100} \quad \cdots \quad \hat{p}_{10000} = \frac{35}{100} \sim Bin(1, p)$$

$$\hat{p} = \frac{X}{n} \sim Bin(1, p)$$

$$\hat{p} \sim \underline{N(p, \sqrt{pq})}$$

$$P_1 \quad P_2 \quad P_3 \quad \cdots \quad P_{10000}$$

$$\begin{array}{cccc} p_1^{(1)} & p_2^{(1)} & p_3^{(1)} & p_{10000}^{(1)} \\ p_1^{(2)} & p_2^{(2} & p_3^{(2)} & p_{10000}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ p_1^{(30)} & p_2^{(30)} & p_3^{(30)} & p_{10000}^{(30} \end{array}$$

$$\hat{p}_1 = \frac{1}{30}\sum_{i=1}^{30} p_1^{(i)} \qquad \hat{p}_2 = \frac{1}{30}\sum_{i=1}^{30} p_2^{(i)} \quad \cdots \quad \hat{p}_{10000} = \frac{1}{30}\sum_{i=1}^{30} p_{10000}^{(i)}$$

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{30}}\right)$$

Distribución normal estándar de las medias muestrales

Siempre que la distribución de ls medias muestrales en el muestreo es un distribución normal, podemos calcular una **variable aleatoria normal estandarizada**, Z, que tiene una media de 0 y una varianza de 1:

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

Resultados de la distribución de las medias muestrales en el muestreo

Sea $\overline{X}$ la media muestral de una muestra aleatoria de n observaciones de una población tiene una media μ_X y una varianza σ^2. En ese caso,

1. La distribución de $\overline{X}$ en el muestreo tiene la media

$$E(\overline{X}) = \mu_X \tag{2}$$

Resultados de la distribución de las medias muestrales en el muestreo

2. La distribución de $\overline{X}$ en el muestreo tiene la desviación típica (error típico)

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{3}$$

   Se llama error típico.

3. Si el tamaño de la muestra, n, no es pequeño en comparación con la población, N, el error típico de $\overline{X}$ es

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{4}$$

Resultados de la distribución de las medias muestrales en el muestreo

4. Si la distribución de la población de la que procede la muestra es normal y, por lo tanto, la distribución de las medias muestrales en el muestreo es normal, la variable aleatoria

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \tag{5}$$

sigue una distribución normal estándar de media 0 y varianza 1.

Formulación del teorema de límite central

Sea $X_1, X_2, \cdots, X_n$ un conjunto de n variables aleatorias independientes que tienen distribuciones idénticas con una media μ y una varianza σ^2. Sea X la suma y $\overline{X}$ la media de estas variables aleatorias. A medida que aumenta n, el **teorema de límite central** establece que la distribución de

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{X - n\mu_X}{\sqrt{n\sigma^2}} \tag{6}$$

tiende a la distribución normal estándar.

Proporción muestral

Sea X el número de éxitos es una muestra binomial de n observaciones cuyo parámetro es P. El parámetro es la proporción de miembros de la población que tienen una característica de interés. La **proporción muestral** es

$$\hat{P} = \frac{X}{n} \tag{7}$$

X es la suma de un conjunto de n variables aleatorias de Bernoulli independientes, cada una de las cuales tienen una probabilidad de éxito P.

Por lo tanto, $\hat{P}$ es la media de un conjunto de variables aleatorias y se aplican los resultados que hemos obtenido en los apartados anteriores para las medias muestrales.
Además, puede utilizarse el teorema del límite central para sostener que la distribución de probabilidades de $\hat{P}$ puede considerarse una variable aleatoria que sigue una distribución normal.

Distribución de la proporción muestral en el muestreo

Sea $\hat{P}$ la proporción muestral de éxitos en una muestra aleatoria extraída de una población en la que la proporción de éxitos es P. En ese caso,

1. La distribución de $\hat{P}$ en el muestreo tiene una media P.

$$E(\hat{P}) = P. \tag{8}$$

2. La distribución de p en el muestreo tiene una desvianción típica

$$\sigma_{\hat{P}} = \sqrt{\frac{P(1-P)}{n}} \tag{9}$$

3. Si el tamaño de la muestra es grande, la variable aleatoria

$$Z = \frac{\hat{P} - P}{\sigma_{\hat{P}}} \tag{10}$$

está distribuida aproximadamente como una normal estándar. Esta aproximación es buena si

$$nP(1 - P) > 9$$

Ejemplo

Un gerente de un grupo de hospitales cree que el 30 % de todos los pacientes generan facturas que se cobran con 2 meses de retraso.
Se toma una muestra aleatoria de 200 pacientes.

a. Calcula $\sigma_{\hat{P}}$.

$$\sigma_{\hat{P}} = \sqrt{\frac{0.3(0.7)}{200}}$$

$$= 0.0324037$$

b. $P(\hat{P} < 0{,}25)$ $= P\left(\frac{\hat{p} - p}{\sigma_{\hat{p}}} < \frac{0.25 - p}{\sigma_{\hat{p}}} \right)$

$$= P\left(Z < \frac{0.25 - 0.3}{0.0324} \right)$$

$$= P\left(Z < \frac{-0.05}{0.0324} \right)$$

$$= P(Z < -1.543033)$$

$$= F(-1.543033)$$

$$= 0.0614132$$

c. $P(\hat{P} > 0{,}33) = P\left(\frac{\hat{p} - p}{\sigma_{\hat{p}}} > \frac{0.33 - 0.3}{0.0324}\right)$

$= P(Z > 0.9258201)$

$= 1 - F(0.9258201)$

$= 0.1772697.$

Varianza muestral

Sea $x_1, x_2, \cdots, x_n$ una muestra aleatoria de observaciones de una población. La cantidad

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

se llama **varianza muestral** y su raíz cuadrada, s, se llama desvianción típica muestral. Dada una muestra aleatoria específica, podríamos calcular la varianza muestral y ésta sería diferente para cada muestra aleatoria, debido a las diferencias entre las observaciones muestrales.

Distribución ji-cuadrado de varianzas muestrales y poblacionales

Dada una muestra aleatoria de n observaciones procedenbtes de una población que sigue una distribución normal cuya varianza poblacional es σ^2 y cuya varianza muestral resultante es s^2, puede demostrarse que

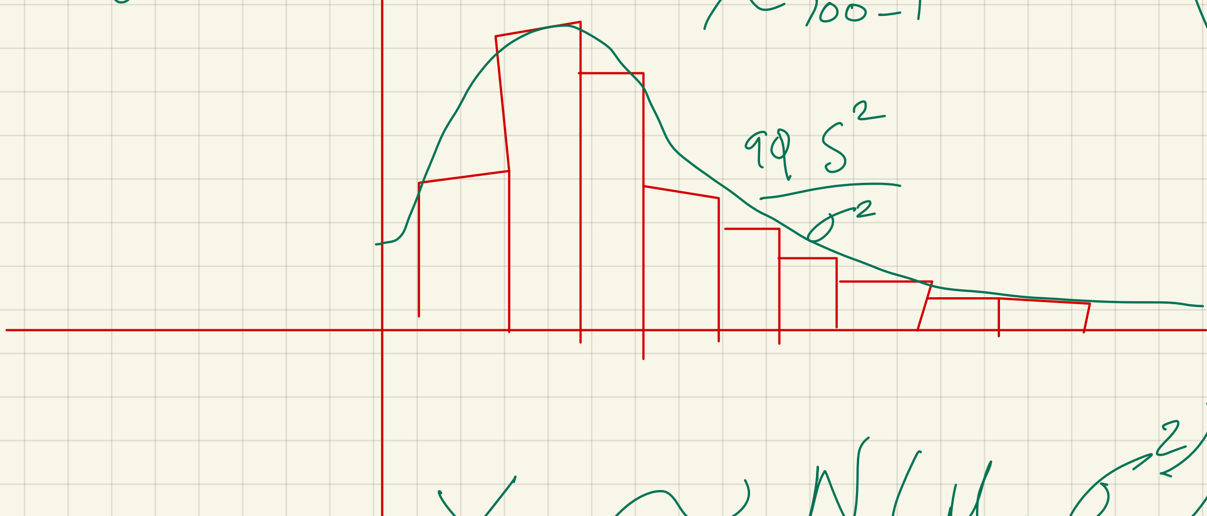
$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sigma^2}$$

tiene una distribución conocida con el nombre de **distribución** χ^2 (**ji-cuadrado**) con $n-1$ grados de libertad.

$$
\begin{array}{lcccc}
 & X_1 & X_2 & \cdots & X_{1500} \\
\text{100 obs.}\left\{\begin{array}{l} \\ \\ \\ \\ \\ \end{array}\right. &
\begin{array}{c} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \\ \vdots \\ x_{100}^{(1)} \end{array} &
\begin{array}{c} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \\ \vdots \\ x_{100}^{(2)} \end{array} &
\begin{array}{c} \cdots \\ \cdots \\ \cdots \\ \ddots \\ \cdots \end{array} &
\begin{array}{c} x_1^{(1500)} \\ x_2^{(1500)} \\ x_3^{(1500)} \\ \vdots \\ x_{100}^{(1500)} \end{array} \\
\hline
 & S_1^2 & S_2^2 & & S_{1500}^2
\end{array}
$$

$$
99\frac{S_1^2}{\sigma^2} \qquad \frac{99 S_2^2}{\sigma^2} \qquad \frac{99 S_{1500}^2}{\sigma^2}
$$

$$\chi^2_{100-1}$$



$$X_i \sim N(\mu, \sigma^2)$$

$$i = 1, \ldots, 1500$$

$$\text{mean}\left(\frac{99 S^2}{\sigma^2}\right) = \, ?$$

Distribución de las varianzas muestrales en el muestreo

Sea s^2 la varianza muestral de una muestra aleatoria de n observaciones procendentes de una población que tiene una varianza σ^2. En ese caso,

1. La distribución de s^2 en el muestreo tien una media σ^2:

$$E(s^2) = \sigma^2 \tag{11}$$

2. La varianza de la distribución de s^2 es el muestreo depende de la distribución de la población subyancente. Si esa distribución es normal, entonces

$$Var(s^2) = \frac{2\sigma^4}{n-1} \tag{12}$$

3. Si la distribución de la población es normal, entonces $\frac{(n-1)s^2}{\sigma^2}$ se distribuye como una $\chi^2_{(n-1)}$.

Si $X \sim \chi^2_{(n-1)}$ entonces $E(X) = n-1$ 1

$Var(X) = 2(n-1)$ 2

$$X = \frac{(n-1)S^2}{\sigma^2} \quad 3$$

Reemplazando 3 en 1

$$E(X) = E\left(\frac{(n-1)S^2}{\sigma^2}\right) = \frac{n-1}{\sigma^2} E(S^2)$$

$$\Rightarrow \frac{n-1}{\sigma^2} E(S^2) = n-1$$

$$\boxed{E(S^2) = \sigma^2} \checkmark$$

Reemplazando 3 en 2 se tiene que

$$Var(X) = Var\left(\frac{(n-1)}{\sigma^2} S^2\right)$$

$$= \frac{(n-1)^2}{\sigma^4} Var(S^2)$$

$$\overset{4}{=} 2(n-1)$$

Despejando $Var(S^2)$ en 4 se tiene que

$$Var(S^2) = \frac{2\cancel{(n-1)}}{(n-1)^{\cancel{2}}} \cdot \sigma^4$$

$$= \frac{2\sigma^4}{n-1}$$

Ejemplo

El tiempo que tarda una unidad de procesamiento en procesar un determinado tipo de trabajo, se distribuye normalmente con una media de 20 segundos y una desviación estándar de 3 segundos. Si se observa una muestra de 15 de estos trabajos ¿Cuál es la probabilidad de que la varianza de la muestra exceda los 12 segundos?

$\mu = 20$ seg $\quad \sigma = 3$ seg $\quad n = 15$

grados de libertad (gl) = 14

$$P(S^2 > 12) = P\left(\frac{S^2(n-1)}{\sigma^2} > \frac{12(n-1)}{\sigma^2}\right)$$

$$= P\left(\chi^2_{14} > \frac{12(14)}{9}\right)$$

$$= P\left(\chi^2_{14} > 18.67\right)$$

$$= 1 - P\left(\chi^2_{14} < 18.67\right)$$

$$= 1 - F(18.67)$$

$$= 1 - \text{pchisq}(18.67, 14)$$

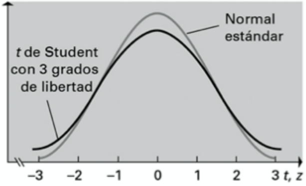
$$= 1 - 0.8221 = 0.1779 \approx 0.18$$

Distribución t de Student

Dada una muestra aleatoria de n observaciones, de media $\overline{x}$ y desviación estándar típica s, extraída de una población que sigue una distribución normal de media μ, la variable aliatoria t sigue la **distribución t de Student** con $(n-1)$ grados de libertad y viene dada por

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



gura 8.7. Funciones de densidad de la distribución normal estándar y la distribución t de Student con 3 grados de libertad.

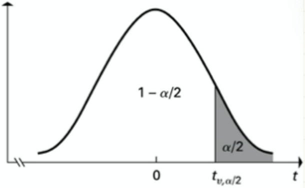


Figura 8.8. $P(t_v > t_{v,\alpha/2}) = \alpha/2$, donde t_v es una variable aleatoria t de Student con v grados de libertad.

Ejemplo

Los precios de la gasolina experimentaron una vertiginosa subida en los primeros años de este siglo. Supongamos que se ha realizado recientemente un estudio con camioneros que tenían más o menos el mismo número de años de experiencia para comprobar el comportamiento de 24 camiones de un determinado modelo en la misma autopista.

Determine los extremos del intervalo, que con un 90 % de confianza contiene a la verdadera media poblacional del consumo de combustible, suponiendo que el consumo, en millas por galón, de estos 24 camiones es

15.5	21	18.5	19.3	19.7	16.9
20.2	14.5	16.5	19.2	18.7	18.2
18	17.5	18.5	20.5	18.6	19.1
19.8	18	19.8	18.2	20.3	21.8

Solución Se calcula $\bar{x}$ y S

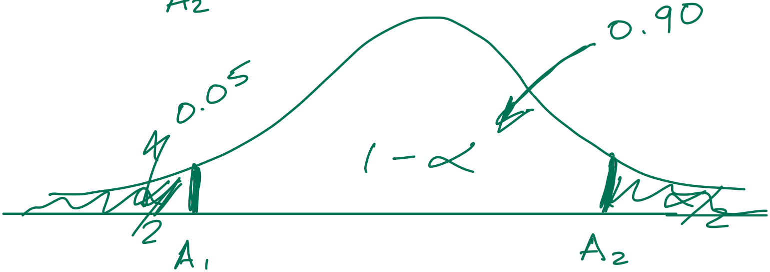
$\bar{x} = 18.68$ $\quad S = 1.69526$, $\quad n = 24$

$$P(a_1 < \mu < a_2) = 0.9$$

$$P\left(\frac{a_1 - \bar{x}}{\frac{S}{\sqrt{n}}} < \frac{\mu - \bar{x}}{\frac{S}{\sqrt{n}}} < \frac{a_2 - \bar{x}}{\frac{S}{\sqrt{n}}} \right)$$

$$P\left(\frac{a_1 - 18.68}{\frac{1.69526}{\sqrt{24}}} < t < \frac{a_2 - 18.68}{\frac{1.69526}{\sqrt{24}}} \right)$$

$$F\Bigg(\underbrace{\frac{a_2 - 18.68}{\frac{1.69526}{\sqrt{24}}}}_{A_2}\Bigg) - F\Bigg(\underbrace{\frac{a_1 - 18.68}{\frac{1.6926}{\sqrt{24}}}}_{A_1}\Bigg) = 0.90.$$



$$1 - \alpha + \frac{\alpha}{2} = 1 - \frac{\alpha}{2}$$

$$F\left(\frac{a_2 - 18.68}{\frac{1.69526}{\sqrt{24}}}\right) = 0.95$$

$$\Longrightarrow \quad qt(0.95, 23) = \frac{a_2 - 18.68}{\frac{1.69526}{\sqrt{24}}}$$

$$[qt(0.95, 23)] \frac{1.69526}{\sqrt{24}} + 18.68 = a_2$$

$$19.273 = a_2$$

$$\boxed{a_2 = \bar{x} + qt\left(1 - \frac{\alpha}{2}, n-1\right) \cdot \frac{s}{\sqrt{n}}}$$

$$P\left(t < \frac{a_1 - 18.68}{0.3460} \right) = 0.05$$

$$F\left(\frac{a_1 - 18.68}{0.3460} \right) = 0.05$$

$$\frac{a_1 - 18.68}{0.3460} = qt(0.05, 23)$$

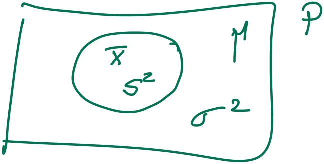
$$a_1 = 0.3460\, qt(0.05, 23) + 18.68$$

$$a_1 = 18.087$$

$$\boxed{a_1 = \overline{X} - q\left(1 - \alpha/2, n-1\right) \cdot \frac{S}{\sqrt{n}}}$$

Problema

La resistencia a la tensión de un tipo de alambre está distribuida normalmente con media desconocida μ y varianza desconocida σ^2. Seis trozos de alambre se seleccionaron aleatoriamente de un rollo largo y se mide la resistencia para cada uno. La media poblacional μ y varianza σ^2 pueden ser estimadas por $\overline{Y}$ y S^2, respectivamente



a. Encuentre la probabilidad de que $\overline{Y}$ esté a no más de $2S/\sqrt{n}$ de la verdadera media poblacional μ.

$$P\left(|\overline{Y}-\mu| < \frac{2S}{\sqrt{n}}\right)$$

$$= P\left(-\frac{2S}{\sqrt{n}} < \overline{Y}-\mu < \frac{2S}{\sqrt{n}}\right)$$

$$= P\left(\frac{-\frac{2\cancel{S}}{\cancel{\sqrt{n}}}}{\frac{\cancel{S}}{\cancel{\sqrt{n}}}} < \boxed{\frac{\overline{Y}-\mu}{\frac{S}{\sqrt{n}}}} < \frac{\frac{2\cancel{S}}{\cancel{\sqrt{n}}}}{\frac{\cancel{S}}{\cancel{\sqrt{n}}}}\right) = P(-2 < t < 2)$$

$$= F(2) - F(-2)$$

$$= pt(2,5) - pt(-2,5)$$

$$= 0.898 \approx 0.9$$

b. Encuentre la probabilidad de que $\overline{Y}$ esté a no más de $2\sigma/\sqrt{n}$ de la verdadera media poblacional μ y compare con el resultado anterior.

$$P\left(|\overline{Y} - \mu| < \frac{2\sigma}{\sqrt{n}} \right) =$$

$$P\left(-\frac{2\sigma}{\sqrt{n}} < \overline{Y} - \mu < \frac{2\sigma}{\sqrt{n}} \right) =$$

$$P\left(\frac{-\frac{2\sigma}{\cancel{\sqrt{n}}}}{\frac{\sigma}{\cancel{\sqrt{n}}}} < \frac{\overline{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{\frac{2\sigma}{\cancel{\sqrt{n}}}}{\frac{\sigma}{\cancel{\sqrt{n}}}} \right) = P(-2 < Z < 2)$$

$$= F(2) - F(-2)$$

$$= \text{pnorm}(2) - \text{pnorm}(-2)$$

$$= 0.9544$$

Hay tres cuestiones que se deben de tener en presentes para evitar que haya confusión respecto a las distribuciones muestrales fundamentales:

1. **NO** se puede usar el teorema central de límite (TCL) a menos que se conozca σ. Para usar el TCL cuando no se conoce σ, se debe de remplazar con s, la desviación estándar de la muestra.

2. El estadístico t **NO** es un resultado del TCL y $x_1, \cdots, x_n$ deben de provenir de una distribución normal $(x;\ \mu, \sigma)$ para que

$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t - \text{student}_{(n-1)}$$

por supuesto, s es tan sólo una estimación de σ.

3. Aunque el concepto de grados de libertad es nuevo en este punto, debería ser muy intuitivo, ya que es razonable que la naturaleza de la distribución de s y también de t deban de depender de la cantidad de información en la muestra $x_1, \cdots, x_n$.

2. El estadístico t **NO** es un resultado del TCL y $x_1, \cdots, x_n$ deben de provenir de una distribución normal $(x;\ \mu,\ \sigma)$ para que

$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t - \text{student}_{(n-1)}$$

por supuesto, s es tan sólo una estimación de σ.

3. Aunque el concepto de grados de libertad es nuevo en este punto, debería ser muy intuitivo, ya que es razonable que la naturaleza de la distribución de s y también de t deban de depender de la cantidad de información en la muestra $x_1, \cdots, x_n$.