INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN540- Métodos Estadísticos para Economía y gestión

Profesores: Marcelo Henríquez, Felipe Avilés
Profesor Auxiliar: José Miguel Carrasco

Guía probabilidades y estadística:

La presente guía contiene la definición y ejemplos de todos los elementos requeridos para introducirse al curso. Todo esto tiene, por lo mismo, un carácter recordatorio y no para el aprendizaje de dichos contenidos. En caso de requerir mayor información, se sugiere la búsqueda de textos especializados en probabilidades y estadística

1- Conteo:

(Principio aditivo del conteo) Sean A y B dos sucesos que no pueden ocurrir simultáneamente. Si A ocurre de a maneras distintas y B ocurre de b maneras distintas, el número de maneras en el cual puede ocurrir A o B es A +B

Ejemplo: Se tienen 6 banderas de señalización, dos rojas, dos verdes y dos azules. ¿Cuántas señales distintas pueden hacerse con una o dos banderas a la vez?

Solución:

Si denotamos las banderas rojas, verdes y azules por R, V y A, respectivamente, vemos que con una bandera a la vez se pueden hacer 3 señales distintas:

R
V
A

Con dos banderas a la vez se puede hacer las siguientes señales (sacando, por ejemplo, una primera y después la otra):

R R
R V
R A
V R
V V
V A
A R
A V
A A

Entonces, si se utilizan dos banderas, se pueden hacer 9 señales distintas. Luego, con una o dos banderas se podrán realizar 3+9= 12 señales diferentes. Observa que, como se establece en la definición, se trata de dos sucesos A y B descritos como:

A: Se hacen señales con una sola bandera

B: Se hacen señales con dos banderas.

Y que ambos no pueden ocurrir simultáneamente, ya que si se decide hacer señales con una bandera se descarta la segunda alternativa y viceversa.
(Principio multiplicativo de conteo): Si un suceso puede ocurrir en a maneras e, independientemente, un segundo suceso puede ocurrir en b maneras, entonces el número de maneras en que ambos, A y B, pueden ocurrir ab.
A este principio también se le denomina principio fundamental de conteo.

1.1- Permutación

Son arreglos o selecciones ordenadas de k objetos o símbolos, tomados de un conjunto que tiene n objetos o símbolos. El orden de aparición de los objetos es importante (Es un arreglo).

Ejemplo: abc, acb, bca, bac. cab, cba, son permutaciones de las letras a, b, c.
Su fórmula matemática es:

$$P(n,k) = \frac{n!}{(n-k)!}$$ Con k <n

1.2- Combinatoria

El número de subconjuntos de r elementos que se pueden seleccionar de un conjunto de n elementos se llaman combinaciones de n elementos tomados de r en r y se notan por Cn,r, calculadas por:

$$C_{n,r} = \binom{n}{r} = \frac{n!}{(n-r)!r!}$$

Ejemplo:
Una placa de circuito impreso tiene 8 localizaciones diferentes donde se puede colocar una componente. Si se van a colocar 5 componentes idénticas sobre la placa, ¿cuántos diseños diferentes se pueden construir?
Cada diseño es un subconjunto de las 8 localizaciones que van a contener componentes. De la ecuación anterior se deduce que el número de diseños posibles es:

$$C_{8,5} = \binom{8}{5} = \frac{8!}{(8-5)!5!} = 56$$

2- Variables Aleatorias

Una **variable aleatoria** (v.a.) *X* es una función real definida en el espacio muestral asociado a un experimento aleatorio, Ω.[1]
Ejemplo:

Supongamos que se lanzan dos monedas al aire. El espacio muestral, esto es, el conjunto de resultados elementales posibles asociado al experimento, es

Ω = {cc, cx, xc, xx},

donde (c representa "sale cara" y x, "sale cruz").

Podemos asignar entonces a cada suceso elemental del experimento el número de caras obtenidas. De este modo se definiría la variable aleatoria X como la función

$$X : \Omega \to \mathbb{R}$$

dada por

$$cc \rightarrow 2$$
$$cx, xc \rightarrow 1$$
$$xx \rightarrow 0$$

El recorrido o rango de esta función, R_X, es el conjunto
$R_X = \{0, 1, 2\}$
2.1 Tipos de variables aleatorias:

- Variable aleatoria discreta: una v.a. es discreta si su recorrido es un conjunto discreto.

- Variable aleatoria continua: una v.a. es continua si su recorrido no es un conjunto numerable. Intuitivamente esto significa que el conjunto de posibles valores de la variable abarca todo un intervalo de números reales. Por ejemplo, la variable que asigna la estatura a una persona extraída de una determinada población es una variable continua ya que, teóricamente, todo valor entre, pongamos por caso, 0 y 2,50 m, es posible

- Variable aleatoria independiente: Supongamos que "X" y "Y" son variables aleatorias discretas. Si los eventos X = x / Y = y son variables aleatorias independientes. En tal caso: P(X = x, Y = y) = P(X = x) P (Y = y). De manera equivalente: f(x,y) = f1(x)f2(y). :Inversamente, si para todo "x" e "y" la funcion de probabilidad conjunta f(x,y) puede expresarse sólo como el producto de una función de "x" sola y como una función dc y solo (las cuales son entonces funciones de probabilidad marginal de "X" e "Y"), "X" e "Y" son independientes. Sin embargo, si f(x,y),no puede exprarsarse de tal manera, entonces "X" e "Y" son independientes.

3- Densidad/Funciones de probabilidad

3.1- La **distribución de probabilidad** de una v.a. X, también llamada función de distribución de X es la función FX(x), que asigna a cada evento definido sobre X una probabilidad dada por la siguiente expresión:

$$F_X(x) = P(X \leq x)$$

y de manera que se cumplan las siguientes tres condiciones:

1. $\lim_{x \to -\infty} F(x) = 0$ y $\lim_{x \to \infty} F(x) = 1$
2. Es continua por la derecha.
3. Es monótona no decreciente.

La distribución de probabilidad de una v.a. describe teóricamente la forma en que varían los resultados de un experimento aleatorio.

3.2- La **función de densidad de probabilidad** (FDP) o, simplemente, función de densidad, representada comúnmente como f(x), se utiliza con el propósito de conocer cómo se distribuyen las probabilidades de un suceso o evento, en relación al resultado del suceso.

La FDP es la derivada (ordinaria o en el sentido de las distribuciones) de la función de distribución de probabilidad F(x), o de manera inversa, la función de distribución es la integral de la función de densidad:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

La función de densidad de una v.a. determina la concentración de probabilidad alrededor de los valores de una variable aleatoria continua.

4- Esperanza

Es el promedio o valor medio de una variable X y está dada por:

$E(X) = \sum_{x} xp(x)$ si x es discreta

$E(X) = \int xf(x)dx$ si x es continua

Ejemplo:

Sea X una Variable Aleatoria Discreta con su función:

x	1	3	5
P(X=x)	0.25	0.50	0.25

Hallar la esperanza matemática de Y= (x -3). Utilizando la definición de esperanza matemática tenemos:

X= {1,3,5}

$E(X) = \sum_{x} ((x-3)^2 P(X=x) = (1-3)^2(0.25) + (3-3)^2(0.50) + (5-3)^2(0,25) = 2$

5- Varianza y covarianza

La varianza es la esperanza del cuadrado de la desviación de dicha variable respecto a su media. Se trata de una medida de la dispersión de dicha variable aleatoria.

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] \quad \mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2$$

$$\mathrm{Var}(X) = \int (x-\mu)^2 p(x)\,dx \quad \mu = \int x\,p(x)\,dx$$

La covarianza es una medida de dispersión conjunta de dos variables estadísticas.

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}\left((X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))\right) = \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\mathrm{E}(X_j)$$

Ejemplo: Sean X e Y variables aleatorias discretas con distribución de probabilidad F_{xy} dada por

X	0	1	1	2	2
Y	0	1	2	1	2
F_{xy}	1/8	¼	1/8	1/8	3/8

Encuentre Esperanza de X e Y, Varianza y Covarianza

$$\mathbf{E}[X] = 0\, f_{xy}(0,0) + 1\, f_{xy}(1,1) + 1\, f_{xy}(1,2) + 2\, f_{xy}(2,1) + 2\, f_{xy}(2,2) = \frac{11}{8}$$

$$\mathbf{E}[Y] = 0\, f_{xy}(0,0) + 1\, f_{xy}(1,1) + 2\, f_{xy}(1,2) + 1\, f_{xy}(2,1) + 2\, f_{xy}(2,2) = \frac{11}{8}$$

$$!\mathbf{E}[X] = \sum\sum x\, f(x,y) = \sum x\, f_x(x)!$$

$$\mathbf{E}\left[X^2\right] = \sum\sum x^2\, f_{xy}(x,y) = \frac{19}{8} \Rightarrow \mathbf{V}[X] = \frac{19}{8} - \left(\frac{11}{8}\right)^2 = \frac{31}{64}$$

$$\mathbf{E}\left[Y^2\right] = \sum\sum y^2\, f_{xy}(x,y) = \frac{19}{8} \Rightarrow \mathbf{V}[Y] = \frac{31}{64}$$

$$\mathbf{E}[XY] = \sum\sum x\, y\, f_{xy}(x,y) = \frac{18}{8} = \frac{9}{4}$$

6- Teorema de Tchebyshev

Si una variable aleatoria tiene una varianza o desviación estándar pequeña, esperaríamos que la mayoría de los valores se agrupan alrededor de la media. Por lo tanto, la probabilidad de que una variable aleatoria tome un valor dentro de cierto intervalo alrededor de la media es mayor que para una variable aleatoria similar con una desviación estándar mayor si pensamos en la probabilidad en términos de una área, esperaríamos una distribución continua con un valor grande de σ que indique una variabilidad mayor y, por lo tanto, esperaríamos que el área este extendida. Sin embargo, una desviación estándar pequeña debería tener la mayor parte de su área cercana a μ.

7- Ley de Grandes Números

Establece que la frecuencia relativa de los resultados de un cierto experimento aleatorio, tienden a estabilizarse en cierto número, que es precísamente la probabilidad , cuando el experimento se realiza muchas veces. Ese valor es la media μ del sistema

8- Teorema Central del Límite

El teorema del límite central o teorema central del límite indica que, en condiciones muy generales, la distribución de la suma de variables aleatorias tiende a una distribución normal (también llamada distribución gaussiana o curva de Gauss o campana de Gauss) cuando la cantidad de variables es muy grande.

9- Distribución t-student

es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño.

$$T = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}},$$

donde

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

10- Distribución chi-cuadrado

La distribución χ^2 tiene muchas aplicaciones en inferencia estadística, por ejemplo en la denominada prueba χ^2 utilizada como prueba de independencia y como prueba de bondad de ajuste y en la estimación de varianzas. También está

involucrada en el problema de estimar la media de una población normalmente distribuida y en el problema de estimar la pendiente de una recta de regresión lineal, a través de su papel en la distribución t de Student, y participa en todos los problemas de análisis de varianza, por su papel en la distribución F de Snedecor, que es la distribución del cociente de dos variables aleatorias independientes con distribución χ^2.
Su función de densidad es:

$$f(x;k) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{(k/2)-1} e^{-x/2} & \text{para } x \geq 0 \\ 0 & \text{para } x < 0 \end{cases}$$

11- Distribución F-Fisher

Una variable aleatoria de distribución F se construye como el siguiente cociente:

$$F = \frac{U_1/d_1}{U_2/d_2}$$

donde

U1 y U2 siguen una distribución ji-cuadrada con d1 y d2 grados de libertad respectivamente, y U1 y U2 son estadísticamente independientes.

La distribución F aparece frecuentemente como la distribución nula de una prueba estadística, especialmente en el análisis de varianza.

12- Inferencia: Definición de test de hipótesis

El test de hipótesis juzga si una propiedad que se supone cumple una población estadística es compatible con lo observado en una muestra de dicha población

Ejemplo
Las puntuaciones en un test que mide la variable creatividad siguen, en la población general de adolescentes, una distribución Normal de media 11,5. En un centro escolar que ha implantado un programa de estimulación de la creatividad una muestra de 30 alumnos ha proporcionado las siguientes puntuaciones:
11, 9, 12, 17, 8, 11, 9, 4, 5, 9, 14, 9, 17, 24, 19, 10, 17, 17, 8, 23, 8, 6, 14, 16, 6, 7, 15, 20, 14, 15.
A un nivel de confianza del 95% ¿Puede afirmarse que el programa es efectivo
Respuesta:
1º Ho m = 11,5 //Hipótesis nula
2º H1 m > 11,5
3º El estadístico de contraste en este caso es:

$$t = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n-1}}}$$

4º La media muestral es 12,47 y la desviación típica de la muestra es 5,22, sustituyendo en el estadístico estos valores se obtiene:

$$t = \frac{12,47 - 11,5}{\frac{5,22}{\sqrt{29}}} = 1,00$$

5º Como el contraste es unilateral, buscamos en las tablas de la t de Student, con 29 grados de libertad, el valor que deja por debajo de sí una probabilidad de 0,95, que resulta ser 1,699

6º El valor del estadístico es menor que el valor crítico, por consiguiente se acepta la hipótesis nula.

7º La interpretación sería que no hay evidencia de que el programa sea efectivo.

13- Errores de tipo I y tipo II

El error de tipo I también llamado error de tipo alfa es el error que se comete cuando el investigador rechaza la hipótesis nula (Ho) siendo ésta verdadera en la población.

el error de tipo II, también llamado error de tipo beta, se comete cuando el investigador no rechaza la hipótesis nula siendo ésta falsa en la población.

	H_0 es cierta	H_1 es cierta
Se escogió H_0	No hay error	Error de tipo II
Se escogió H_1	Error de tipo I	No hay error

14- Test de medias y test de diferencia de medias

(Test de medias) Con la notación que habitualmente se utiliza en el contraste de hipótesis tendremos que **m** es la media de la población, **s** la desviación típica de la población, s la desviación típica de la muestra, n es el tamaño de muestra, X la media de la muestra, y Z o t es el estadístico.

Con relación al contraste de medias, suelen emplearse dos tipos de pruebas, los tests unilaterales o los tests bilaterales, que tienen, respectivamente, las siguientes estructuras.

	Test unilateral	Test bilateral
Hipótesis nula	$H_o : \mu = \mu_o$	$H_o : \mu = \mu_o$
Hipótesis alternativa	$H_a : \mu \neq \mu_o$	$H_a : \mu \neq \mu_o$
Estadístico, con distribución N(0,1)	$Z = \frac{X - \mu_o}{\sigma/\sqrt{n}}$	$Z = \frac{X - \mu_o}{\sigma/\sqrt{n}}$
Nivel de significación(generalmente)	$\alpha = 0.05$	$\alpha = 0.05$
Región crítica	$Z_o > 1.645$	$-1,96 \leq Z_{o/2} \leq 1,96$
Criterio aceptación Ho	$Z < Z_o$	$-1,96 \leq Z \leq 1,96$

Ejemplo . Un laboratorio farmacéutico afirma que el antiinflamatorio fabricado por ellos elimina la inflamación en 14 minutos en los casos corrientes.

Con el objeto de comprobar estadísticamente esta afirmación, eligimos al azar 18 pacientes con inflamaciones varias y tomamos como variable de respuesta el tiempo

transcurrido entre la administración del antiinflamatorio y el momento en que desaparece la inflamación. Además, nos dicen que la variable tiempo transcurrido entre la administración del antiinflamatorio y el momento en que desaparece la inflamación sigue una distribución normal de media 14 y desviación 7. El tiempo medio de respuesta de la muestra fue de 19 minutos.
Se pide comprobar la afirmación del laboratorio a un nivel de significación de 0.05.

Solución.

Primero consideremos los datos que tenemos.

X = 19, **m** = 14, **s** = 7, n = 18
Planteemos ahora las hipótesis de este test. Queremos contrastar la hipótesis nula a partir de la afirmación de la empresa que dice que la inflamación desaparece en 14 minutos; así pues, tendremos:

Hipótesis nula $\rightarrow H_o$: **m** = 14

La hipótesis alternativa será el caso desfavorable, en esta ocasión para la empresa, y puede escribirse:

Hipótesis alternativa $\rightarrow$ Ha : **m**> 14

Procederemos aceptando de entrada la hipótesis nula (m = 14), calculando el estadístico y observando si se sitúa en la región crítica. Si así sucediera, rechazaríamos la creencia inicial de aceptación de la hipótesis nula.

Sustituyendo los parámetros de la población y de la muestra en el estadístico tenemos:

$$Z = \frac{X - \mu_o}{\sigma/\sqrt{n}} = \frac{19 - 14}{7/\sqrt{18}} = 3{,}03$$

Con lo que podemos observar que el estadístico se sitúa en la región crítica y ,por lo tanto no sigue el criterio de aceptación de la hipótesis nula.

De ese modo, rechazaríamos la hipótesis H_o de que **m** = 14 y concluimos que a un nivel 0.05 el tiempo medio de eliminar la inflamación por este antiinflamatorio es superior a 14 minutos.

(Test de diferencias de medias) Sean X_1 y X_2 dos medias muestrales de dos poblaciones. Los tamaños de cada una de estas muestras son n_1 y n_2 respectivamente. Queremos observar si la diferencia entre las medias es significativa o no, es decir, comprobar si podemos aceptar que $\mathbf{m}_1 = \mathbf{m}_2$.

Tenemos:

Hipótesis nula	$H_o : \mu_1 - \mu_2$
Hipótesis alternativa	$H_a : \mu_1 - \mu_2 \neq 0$
Estadístico, con distribución N(0,1)	$Z = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/\sqrt{n_1} + \sigma_2^2/\sqrt{n_2}}}$
Nivel de significación(generalmente)	$\alpha = 0.05$
Región crítica	$-1{,}96 \leq Z_{\alpha/2} \leq 1{,}96$
Criterio aceptación H_o	$-1{,}96 \leq Z \leq 1{,}96$

Si las desviaciones de las poblaciones son desconocidas y sólo conocemos las desviaciones muestrales, tendremos que considerar la distribución t de Student en vez de la normal.

Ejemplo . Se conocen los datos de dos muestras de dos poblaciones, que son los siguientes:

$X_1 = 74$	$X_2 = 78$
$S_1^2 = 225$	$S_2^2 = 169$
$N_1 = 42$	$N_2 = 56$

Se pide contrastar estadísticamente si hay diferencia entre las dos poblaciones, a un nivel de significación del 0.05.

Las dos poblaciones siguen una distribución Normal N($\mathbf{m}_1$, $\mathbf{s}_1$) y N($\mathbf{m}_2$, $\mathbf{s}_2$)

Solución.

Sabemos que las distribuciones de las dos poblaciones son Normales, pero desconocemos el valor de su desviación, sólo conocemos el valor de la desviación típica de las muestras. Por ahora, planteemos las hipótesis:

- Hipótesis nula → $H_o : \mathbf{m}_1 - m_2 = 0$, es decir, $m_1 = m_2$
- Hipótesis alternativa → $H_a : m_1 - m_2 \neq 0$, es decir, $m_1 \neq m_2$

Aunque el estadístico que correspondería a este test es el asociado a una distribución T-Student, por ser las desviaciones de las poblaciones desconocidas, como el tamaño de las muestras es elevado y sabemos que una distribución T-Student con muchos grados de libertad se aproximaba mucho a una Normal, utilizaremos el siguiente estadístico:

Estadístico → $Z = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\left[\left(\sigma_1^2/\sqrt{n_1}\right) + \left(\sigma_2^2/\sqrt{n_2}\right)\right]}}$ con distribución N(0,1)

Con los datos de la población y de la muestra, calculamos el estadístico, aceptando, por ahora, la hipótesis nula ($m_1 = m_2$), y observemos en que región se sitúa el estadístico.

$$Z = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{\sqrt{\left[\left(\sigma_1^2 / \sqrt{n_1}\right) + \left(\sigma_2^2 / \sqrt{n_2}\right)\right]}} = \frac{(74 - 78) - 0}{\sqrt{\frac{225}{42} + \frac{169}{56}}} = -1,38$$

Como podemos ver, el estadístico se sitúa en la región de aceptación de la hipótesis nula, con lo que aceptaríamos la H_o ($m_1 = m_2$), y podríamos concluir que, a un nivel de significación de 0.05, las dos poblaciones se pueden considerar iguales estadísticamente.