Análisis de Datos para la Toma de Decisiones

Clase 3

Christian Araya

Instituto de Estadística
PUCV

Marzo 2020

Paréntesis: Elementos de Probabilidad

Introducción

- Sea E un experimento aleatorio, Ω el espacio muestral asociado. Se define una variable aleatoria (X) como una función real, es decir: $X : \Omega \to \mathbb{R}$, tal que $\omega \to X(\omega) = x$.
- Conceptualmente, lo que está haciendo la variable aleatoria es identificar un aspecto numérico de interés en un experimento aleatorio.
- Ejemplos: número de personas que ingresa a una estación de metro entre las 7:30 y 8:15 am; número de caras obtenidas al lanzar una moneda 300 veces; cantidad de alumnos del curso que están de cumpleaños el día 19 de abril.

Conceptos importantes

- Llamaremos R_X al recorrido de la función X (variable aleatoria), es decir, representa el subconjunto de $\mathbb{R}$ donde la variable toma sus valores.
- Si R_X es numerable, se dice que la variable aleatoria es discreta; si R_X es no numerable, se trata de una variable aleatoria continua.
- Ejemplo: supongamos que el experimento E es lanzar una moneda perfecta dos veces. Ya sabemos que el espacio muestral corresponde a: $\Omega = \{$(C,C)(C,S)(S,C)(S,S)$\}$
- Si definimos X : número de sellos en los dos lanzamientos, entonces la v.a. toma los siguientes valores:
- $X(C,C) = 0$
- $X(C,S) = 1$
- $X(S,C) = 1$
- $X(S,S) = 2$

Conceptos importantes

- Si definimos X : número de sellos en los dos lanzamientos, entonces la v.a. toma los siguientes valores:
- $X(C, C) = 0$
- $X(C, S) = 1$
- $X(S, C) = 1$
- $X(S, S) = 2$
- Notar que la función no es inyectiva.
- En este caso, el espacio numérico (recorrido) de X es: $R_X = \{0,1,2\}$

Estadísticos

Introducción

- En adelante, trabajaremos con una colección de variables aleatorias que llamaremos muestra: $Y_1, \dots, Y_n$
- Esta muestra provendrá de una población de interés. Las variables aleatorias serán independientes e idénticamente distribuidas (tendrán la misma distribución).
- Algunas funciones de las variables aleatorias son utilizadas para estimar parámetros de la población, los que desconocemos generalmente.
- Por ejemplo, supongamos que queremos estimar la media de la población, μ. Si tenemos n valores para las variables aleatorias: $y_1, \dots, y_n$, suena razonable que μ sea estimado a partir de: $\bar{Y} = \sum_{i=1}^{n} \frac{1}{n} y_i$

Estadístico

- $\bar{Y} = \sum_{i=1}^{n} \frac{1}{n} y_i$ es una función de la muestra (depende sólo de los valores observados de las variables aleatorias y de la constante n).
- Se define estadístico como cualquier función de la muestra (las variables aleatorias observables) y algunas constantes conocidas.
- Otros ejemplos son: S y S^2; $\max(Y_1, \ldots, Y_n) = Y_{(n)}$; $\min(Y_1, \ldots, Y_n) = Y_{(1)}$; el rango $Y_{(n)} - Y_{(1)}$.
- Los estadísticos se usan para hacer inferencia sobre parámetros que se desconocen en la población y, como son una función de variables aleatorias, también son variables aleatorias.
- Los estadísticos, a partir de lo anterior, tienen distribución de probabilidades (sus distribuciones asociadas al muestreo).

Distribuciones relacionadas con la dist. Normal

- Sea $Y_1, \ldots, Y_n$ una muestra aleatoria de tamaño n de una distribución Normal con una media μ y varianza σ^2.
- Se define la media muestral como: $\bar{Y} = \dfrac{1}{n} \sum_{i=1}^{n} Y_i$
- Sabemos que $\bar{Y}$ se distribuye Normal también, con $\mu_{\bar{Y}} = \mu$ y $\sigma^2_{\bar{Y}} = \sigma^2/n$
- De aquí se deriva que $Z = \dfrac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \sqrt{n} \cdot (\dfrac{\bar{Y} - \mu}{\sigma})$ sigue una distribución Normal Estándar.

Ejemplo 1)

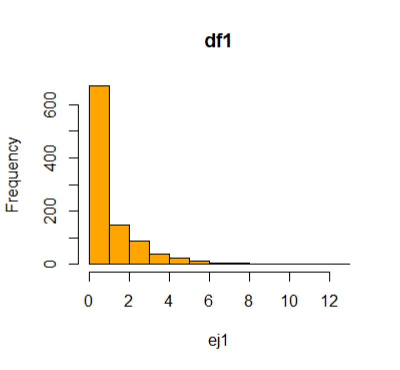
- Una máquina embotelladora puede ser regulada de modo que proporciona en promedio μ ml por botella de jugo que llena. Se ha observado que la cantidad de líquido que entrega la máquina sigue una distribución Normal, con $\sigma = 1$ ml. Una muestra de $n = 9$ botellas se selecciona al azar de una partida que la máquina ha embotellado en un día y se mide el contenido de cada una. Encuentre la probabilidad de que el promedio de la muestra tenga una diferencia de a lo más 0,3 con respecto a la media de la población μ para el proceso de embotellado en estudio.

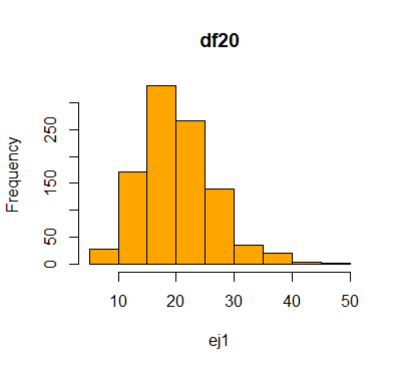
Ejemplo 2)

- Una máquina embotelladora puede ser regulada de modo que proporciona en promedio μ ml por botella de jugo que llena. Se ha observado que la cantidad de líquido que entrega la máquina sigue una distribución Normal, con $\sigma = 1$ ml. Una muestra de $n = 9$ botellas se selecciona al azar de una partida que la máquina ha embotellado en un día y se mide el contenido de cada una. Encuentre la probabilidad de que el promedio de la muestra tenga una diferencia de a lo más 0,3 con respecto a la media de la población μ para el proceso de embotellado en estudio.
- Cuál debe ser el tamaño muestral que se requiere considerar si se busca que $\bar{Y}$ tenga una diferencia de 0,3 con respecto a μ, con probabilidad 0,95.

χ^2

- Sea $Y_1, \dots, Y_n$ una muestra aleatoria de tamaño n de una distribución Normal con una media μ y varianza σ^2.
- Sabemos que $Z_i = (Y_i - \mu)/\sigma$, con i entre 1 y n, son independientes (Y_i lo son) y siguen una distribución Normal Estándar.
- Luego: $\sum_{i=1}^{n} Z_i^2$ tiene una distribución χ^2 con n grados de libertad, siendo n el parámetro para esta distribución.
- A diferencia de la distribución Normal, en esta distribución se observa un nivel de asimetría vinculado al valor de n.


df1
Frequency
0
200
400
600
0
2
4
6
8
10
12
ej1


df20
Frequency
0
50
150
250
10
20
30
40
50
ej1

Uso de Distribución χ^2

- Sea $Y_1, \ldots, Y_n$ una muestra aleatoria de tamaño n de una distribución Normal con una media μ y varianza σ^2.
- Luego: $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ tiene una distribución χ^2 con n-1 grados de libertad.
- Además, $\bar{Y}$ y S^2 son variables aleatorias independientes.

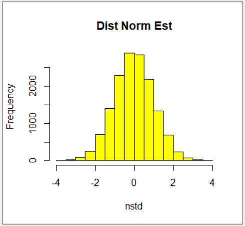
Extensiones

- Sea Z una variable aleatoria Normal Estándar y sea W una variable aleatoria χ^2 con ν grados de libertad. Luego, si Z y W son independientes, se tiene que $T = \dfrac{Z}{\sqrt{\dfrac{W}{\nu}}}$ sigue una distribución t con ν grados de libertad.

Extensiones

- A partir de lo anterior, si $Y_1, \ldots, Y_n$ es una muestra aleatoria de tamaño n de una distribución Normal con una media μ y varianza σ^2, se sabe que $Z = \dfrac{\sqrt{n}(\bar{Y} - \mu)}{\sigma}$ sigue una distribución Normal Estándar.
- También sabemos que $W = \dfrac{(n-1)S^2}{\sigma^2}$ tiene distribución χ^2 con $\nu = n - 1$ grados de libertad.
- Además, Z y W son independientes porque $\bar{Y}$ y S^2 son independientes.
- Esto forma parte del Lema de Fisher.
- Entonces, $T = \dfrac{Z}{\sqrt{\dfrac{W}{\nu}}} = \dfrac{\sqrt{n}(\bar{Y} - \mu)}{S}$ sigue una distribución t con $\nu = n - 1$ grados de libertad.

- La distribución t es simétrica con respecto al 0, al igual que la distribución Normal Estándar, pero se dice que tiene colas más pesadas.
- A continuación se muestra un set de datos ($n = 15000$) generados a partir de una distribución Normal Estándar y de una distribución t con $\nu = 10$ grados de libertad.



Distribución F

- Sean W_1 y W_2 variables aleatorias independientes, distribuidas según χ^2 con ν_1 y ν_2 grados de libertad respectivamente. Luego: $F = \dfrac{W_1/\nu_1}{W_2/\nu_2}$ sigue una distribución F con (ν_1,ν_2) grados de libertad. (También se habla de grados de libertad de numerador y denominador).

Distribución F

- A partir de lo anterior, consideremos **dos** muestras provenientes de **dos** poblaciones independientes. Se sabe que $W_1 = \dfrac{(n_1 - 1)S_1^2}{\sigma_1^2}$ y $W_2 = \dfrac{(n_2 - 1)S_2^2}{\sigma_2^2}$ tienen distribuciones χ^2 independientes con $\nu_1 = n_1 - 1$ y $\nu_2 = n_2 - 1$ grados de libertad, respectivamente (n_1 y n_2 son los tamaños de cada muestra).
- Entonces, $F = \dfrac{W_1/\nu_1}{W_2/\nu_2} = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ sigue una distribución F con $(\nu_1 - 1, \nu_2 - 1)$ grados de libertad.
- Este estadístico será empleado para decidir pruebas de hipótesis relacionadas con varianzas.

Distribución Normal $N(0, 1)$

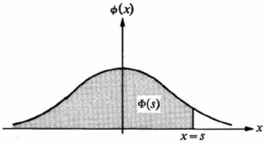


FIGURE 9.3

Distribución χ^2

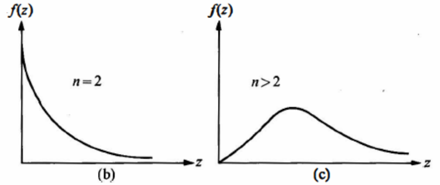
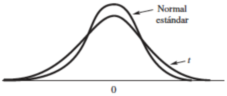


FIGURE 9.7

Fuente: Introductory Probability and Statistical Applications. Meyer, Paul L.

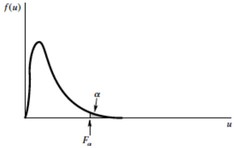
Distribución Normal $N(0, 1)$ vs t-Student

FIGURA 7.3 Comparación de las funciones de densidad normal estándar y t



Distribución F

FIGURA 7.4 Una típica función de densidad de probabilidad F



Fuente: Estadística Matemática con Aplicaciones. Wackerly, Mendenhall, Scheaffer.