Análisis de Datos para la Toma de Decisiones

Clase 4

Christian Araya

Instituto de Estadística
PUCV

Marzo 2020

Recordatorio clase pasada

Estadísticos, muestras

- En adelante, trabajaremos con una colección de variables aleatorias que llamaremos muestra: $Y_1, \ldots, Y_n$
- 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: $y_1, \ldots, y_n$, 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; máx$(Y_1, \ldots, Y_n) = Y_{(n)}$; mín$(Y_1, \ldots, Y_n) = Y_{(1)}$; 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: $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$
- 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 $Z = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_{\bar{Y}}} = \sqrt{n} \cdot (\frac{\bar{Y} - \mu}{\sigma})$ sigue una distribución Normal Estándar.

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.
- 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.

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: $\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ 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 $T = \dfrac{Z}{\sqrt{\dfrac{W}{\nu}}}$

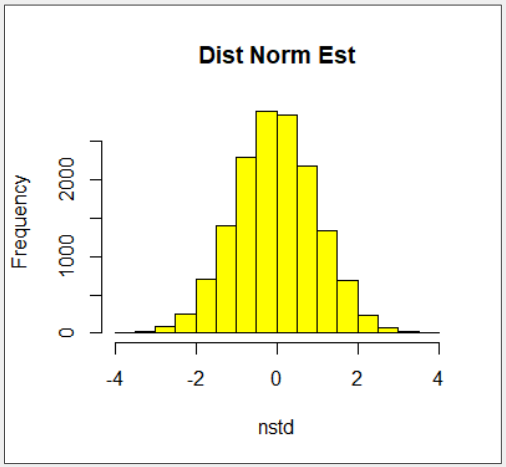
  sigue una distribución t con ν grados de libertad.

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.

Distribuciones en muestras

- 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 $W_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2}$ y $W_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2}$ 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, $F = \frac{W_1/\nu_1}{W_2/\nu_2} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ 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 $N(0, 1)$

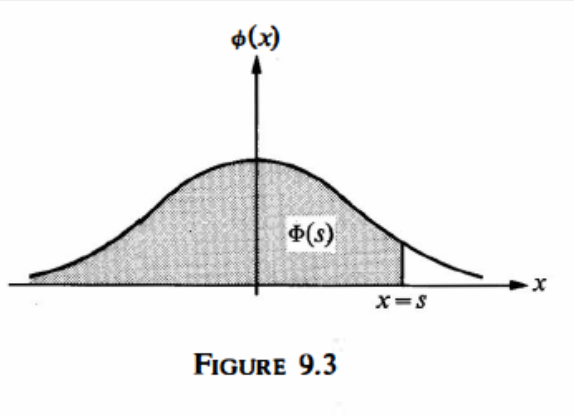


FIGURE 9.3

Distribución χ^2

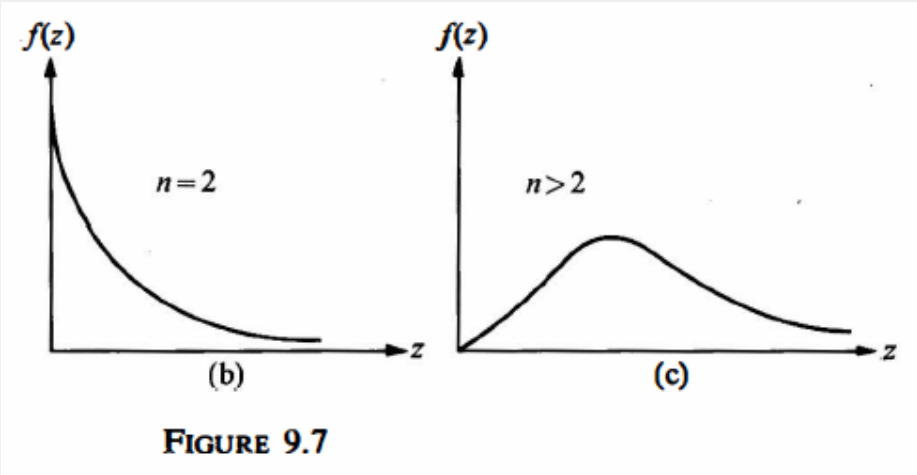
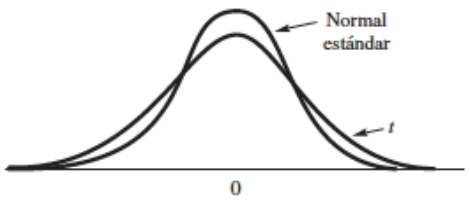


FIGURE 9.7

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

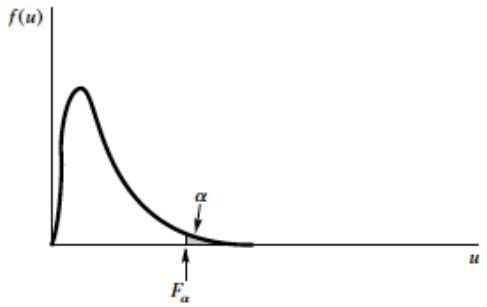
Distribución Normal $N(0, 1)$ vs t-Student

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.

Pruebas de Hipótesis

Hipótesis

- Es una afirmación o suposición. No es un hecho establecido. Esperamos poder aportar evidencia estadística para tomar alguna decisión.

Hipótesis de Investigación

- El investigador, en el contexto de un estudio científico, define una hipótesis en base a su experiencia o conocimiento de un fenómeno.
- Por ejemplo, podría postular que dos tratamientos, con dos medicamentos diferentes, generan efectos distintos (uno mejor que el otro) en pacientes con una determinada enfermedad.

Hipótesis de Investigación

- Para dar respaldo a la hipótesis del investigador, se efectúa una **prueba de hipótesis** o **test de hipótesis**.
- De esta manera, con una metodología estructurada, se puede aportar evidencia para tomar alguna decisión.

Conceptos

- Existen dos hipótesis: una **hipótesis nula** y una **hipótesis alternativa**.
- La primera es H_0 y la segunda, H_1.
- La hipótesis nula plantea una aseveración y, todo lo que sea opuesto a ella, está en la hipótesis alternativa.
- Por ejemplo: si H_0 plantea que la media de la población μ es igual a 75, la hipótesis alternativa H_1 indica que μ es distinto de 75.
- Otro ejemplo: si H_0 plantea que la media de la población μ es menor o igual a 75, la hipótesis alternativa H_1 indica que μ es mayor a 75.

Convenciones

- ¿Cómo definimos las hipótesis? (H_0 y H_1)
- Para construirlas, usaremos la siguiente guía/convención: cada vez que queramos probar si algún parámetro (o alguna diferencia de parámetros) es **igual** a un valor, ésta será H_0.
- En H_1 se adjuntará lo opuesto: que es **diferente**.
- Este tipo de prueba de hipótesis se denomina test a dos colas.
- Para proceder con los cálculos de la prueba, asumiremos que H_0 es verdadera y calcularemos indicadores que nos entregarán evidencia.

Convenciones

- Cada vez que queramos probar si algún parámetro (o alguna diferencia de parámetros) es **mayor o menor a un valor**, ésta será H_1.
- En H_0 se adjuntará lo opuesto, que dependerá de lo que estemos analizando.
- Este tipo de prueba de hipótesis se denomina test a una cola.
- Para proceder con los cálculos de la prueba, asumiremos que H_0 es verdadera, pero trabajando como si fuera una igualdad.
- Veremos que al final, trataremos de recopilar evidencia, a través de cálculos (algún estadístico, por ejemplo), para decidir si debemos rechazar o no la hipótesis nula, en favor de la alternativa.

Errores tipo I y tipo II

- El único problema de la metodología, es que no está libre de errores.
- Podríamos cometer errores de dos tipos:

		Estado Real de la Naturaleza	
		H_0 es Verdadera	H_0 es Falsa
Decisión	No se Rechaza H_0	Decisión Correcta	Error Tipo II
	Se Rechaza H_0	Error Tipo I	Decisión Correcta

Conceptos Importantes

		Estado Real de la Naturaleza	
		H_0 es Verdadera	H_0 es Falsa
Decisión	No se Rechaza H_0	Decisión Correcta	Error Tipo II
	Se Rechaza H_0	Error Tipo I	Decisión Correcta

α y β

- **Tipo I:** se denomina α o nivel de significancia.
- **Tipo II:** se denomina β.
- En general, se trabaja minimizando/controlando α al definir el nivel de significancia de la prueba (en general, al 5 %).

Conceptos Importantes

		Estado Real de la Naturaleza	
		H_0 es Verdadera	H_0 es Falsa
Decisión	No se Rechaza H_0	Decisión Correcta	Error Tipo II
	Se Rechaza H_0	Error Tipo I	Decisión Correcta

Tipo I vs Tipo II

- El experimentador controla el error Tipo I al hacer su prueba.
- Históricamente se postuló que el error Tipo I era el más grave, por ende se enfocó la metodología sobre este indicador.
- Ejemplo: mujer embarazada/no embarazada; paciente con diagnóstico de VIH+/VIH-.

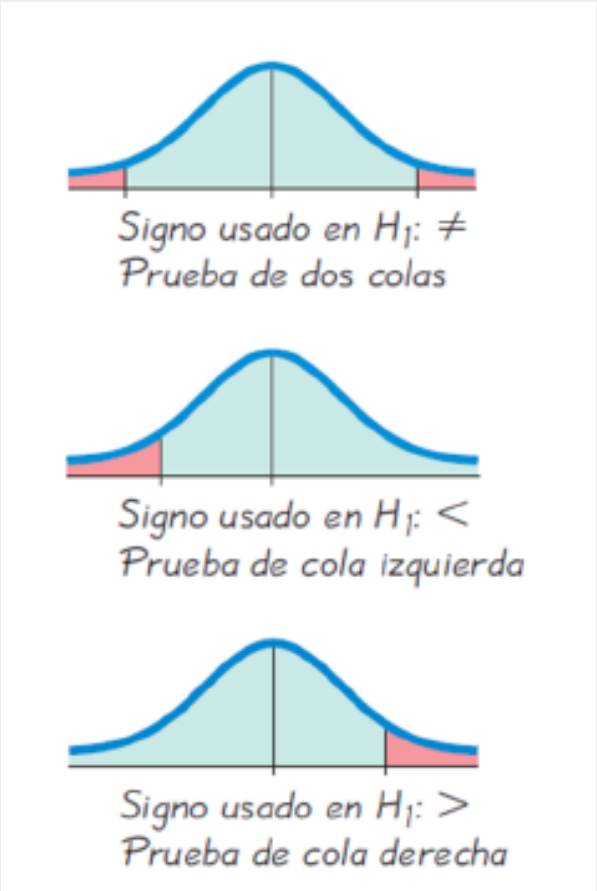
Pruebas de Hipótesis para la Media y Varianza (1 población)

Media

- Sea una muestra aleatoria $X_1, ..., X_n$ proveniente desde una población Normal, con media μ y varianza σ^2.
- Planteamos las hipótesis que correspondan de acuerdo al problema.
- Identificamos el estadístico:
- Si σ^2 es conocido, se calcula $Z_{obs} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} \sim N(0, 1)$
- Si σ^2 es desconocido y $n < 30$, se calcula $T_{obs} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S} \sim t(n-1)$
- Si σ^2 es desconocido y $n \geq 30$, se calcula $Z_{obs} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S} \sim N(0, 1)$

Media

- Se rechaza H_0 si el estadístico de prueba cae dentro de la región crítica de rechazo.



Ejemplo 1

- Se tienen 23 mediciones de tiempos de despacho (días) de un medicamento para la bodega de farmacia que opera en una clínica privada. Usted sabe de antemano que dichos tiempos provienen de una población normal, con media y varianza desconocidas.
- Como meta para la continuidad del contrato, se le pide plantear un test de hipótesis que le permita estudiar si el tiempo promedio de despacho es inferior a 31 (días).
- Trabaje al nivel de significancia de 5 %.

Ejemplo 2

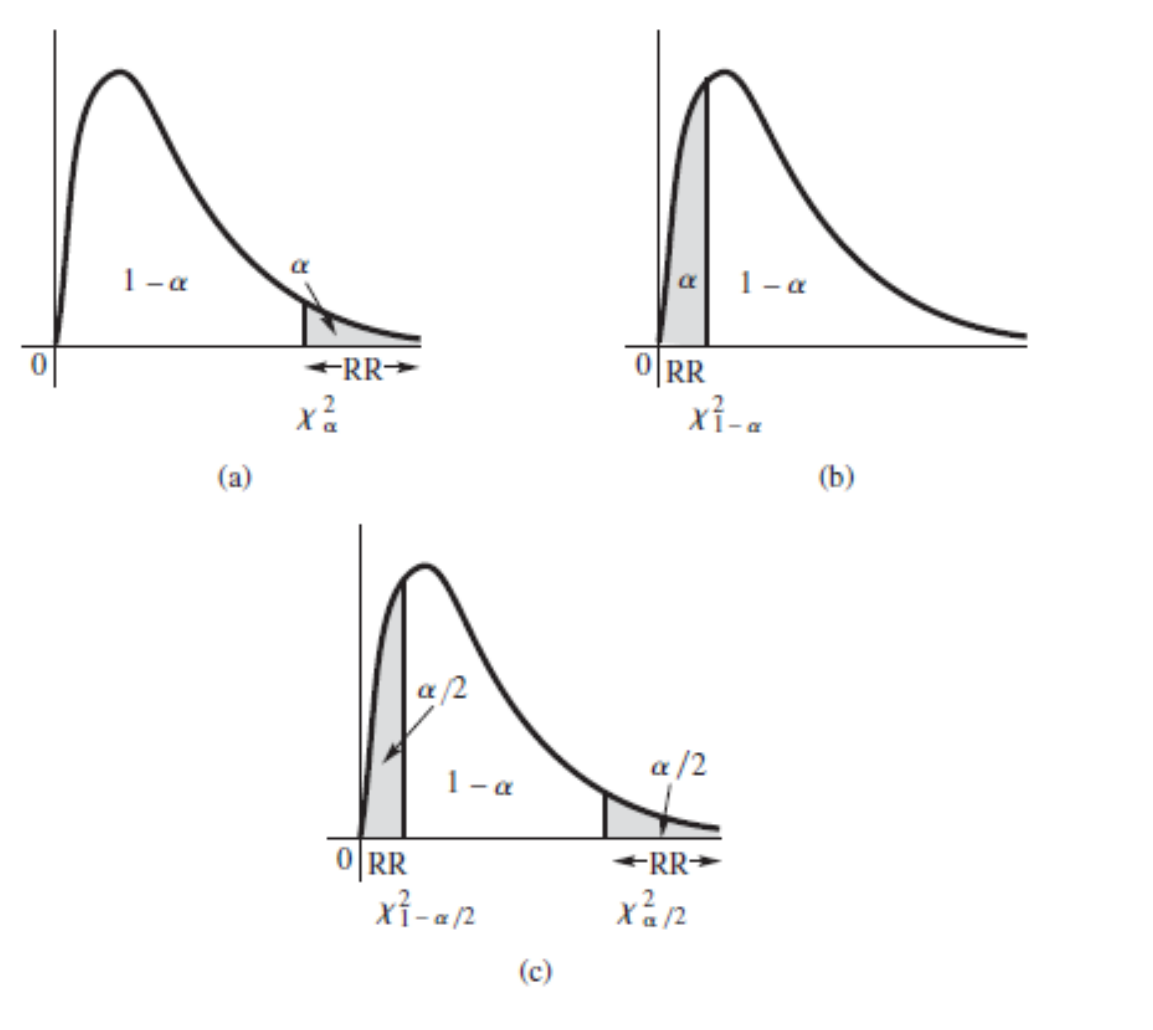
- Las mediciones corresponden a los Kg de arroz con que una máquina rellena los paquetes de una famosa marca. La muestra se ha tomado con el propósito de realizar un control sobre las promesas efectuadas por el proveedor del equipo.
- La empresa productora de arroz está interesada en analizar si el promedio de Kg de arroz, parámetro de funcionamiento de la máquina, se puede validar en 0,995 Kg según la información contenida en la muestra.
- Considere que la desviación estándar informada por el fabricante es de 0,24 Kg y que se sabe que los Kg de arroz siguen una distribución Normal.
- Trabaje al 5 % de significancia. ¿Qué puede concluir?
- ¿Cambiaría su veredicto si trabaja al 7 % de significancia?

Varianza

- Sea una muestra aleatoria $X_1, ..., X_n$ proveniente desde una población Normal, con media μ y varianza σ^2, ambas desconocidas.
- Exploraremos la prueba de hipótesis que permite verificar si $H_0 : \sigma^2 = \sigma_0^2$ para un valor específico, en un test a dos colas.
- El estadístico a construir es $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$, el que distribuye según χ^2 con $n-1$ grados de libertad.
- Recordemos que para un test a una cola, el estadístico es el mismo, solamente cambia el planteamiento de las hipótesis y la definición de las regiones de rechazo.

Varianza

- El texto de Wackerly y Mendenhall (*Estadística Matemática con Aplicaciones*) introduce los siguientes gráficos para guiar en la identificación de la zona de rechazo.



Ejemplo 3

- Considere la misma muestra aleatoria de los Kg de arroz del ejemplo anterior. Ahora, asuma que provienen de una población Normal, de media y varianza desconocidas.
- El SERNAC ha decidido impartir una multa a la empresa arrocera si la variabilidad del contenido de sus paquetes, medida a través de la varianza, sobrepasa los 0,12 (Kg^2).
- Efectúe un test de hipótesis al nivel de significancia de 5 % para estudiar si está cumpliendo o no la normativa.