

# ANÁLISIS DE DATOS PARA LA TOMA DE DECISIONES

CLASE 5

CHRISTIAN ARAYA

INSTITUTO DE ESTADÍSTICA  
PUCV

2020



## Definición

- Escoger de antemano el nivel  $\alpha$  para decidir la prueba de hipótesis tiene origen en lo conversado previamente: la creencia histórica de que el error tipo I debe ser controlado puesto que es el peor tipo de error en que una prueba de hipótesis puede incurrir.
- En las aplicaciones de Estadística en diversas áreas, el criterio del valor-p es la forma más común para decidir un test de hipótesis.
- Sobre el valor-p: **mientras menor sea el valor-p de la prueba, mayor es la evidencia de que debería rechazarse la hipótesis nula.**
- Comparamos el valor-p con el nivel de significancia que exigimos al test (típicamente  $\alpha = 0,05$ ).

# CÓMO EJECUTAR UNA PRUEBA DE HIPÓTESIS

- Si se quiere adoptar el criterio de fijar  $\alpha$  de antemano:

## Forma 1

1. Definir  $H_0$  y  $H_1$ .
2. Elegir un nivel  $\alpha$  fijo.
3. Seleccionar un estadístico de prueba y encontrar la región crítica.
4. Rechazar  $H_0$  si el estadístico de prueba está en la región crítica.
5. Concluir o interpretar en el contexto del problema.

# CÓMO EJECUTAR UNA PRUEBA DE HIPÓTESIS

- Si se quiere adoptar el criterio de calcular el valor-p:

## Forma 2

1. Definir  $H_0$  y  $H_1$ .
2. Seleccionar un estadístico de prueba.
3. Calcular el valor-p, poniendo atención si la prueba es de 1 o 2 colas.
4. Concluir en base al valor-p, comparándolo con el nivel  $\alpha$  deseado para la prueba. Si es inferior, se rechaza  $H_0$ .

**NORMALIDAD**

# **PRUEBAS DE HIPÓTESIS PARA VERIFICAR NORMALIDAD**

# SHAPIRO-WILK VERSUS KOLMOGOROV-SMIRNOV

- Hasta el momento, hemos asumido Normalidad en los datos.
- Para comprobar la hipótesis de Normalidad para una muestra, se puede iniciar el Análisis Exploratorio graficando un histograma o un boxplot.
- Se pueden efectuar gráficos más complejos como un QQ-Plot (típicamente empleado para estudiar la Normalidad de los residuos en una Regresión Lineal).
- Se pueden calcular el coeficiente de Asimetría y de Curtosis, pero ninguna de estas herramientas tiene la fuerza de un test de hipótesis propiamente tal.

# SHAPIRO-WILK VERSUS KOLMOGOROV-SMIRNOV

## KS

- La prueba de KS se utiliza para diagnosticar si un set de datos proviene de una población Normal.
- Su uso se recomienda para muestras grandes:  $n \geq 50$
- En la Hipótesis Nula, se establece que los datos provienen de una distribución Normal.
- KS utiliza un estadístico de prueba que representa la máxima diferencia, en valor absoluto, entre la distribución muestral y la distribución (acumulada) teórica.



## Shapiro

- La prueba de Shapiro tiene el mismo fin.
- Su uso se recomienda para muestras pequeñas:  $n < 50$
- Shapiro utiliza un estadístico de prueba que se construye ordenando los datos de menor a mayor y se calculan las diferencias desde los extremos hacia el centro. Luego se aplican coeficientes definidos por los autores y se realiza una suma de las diferencias.
- Se dice que Shapiro puede ser utilizado en muestras grandes, pero su mejor desempeño es evidente en muestras pequeñas. Por otro lado, Kolmogorov Smirnov es más robusto frente a distorsiones en las colas de la distribución que determinan los datos.

# **PRUEBA DE INDEPENDENCIA (DATOS CATEGÓRICOS)**

## Definición

- La prueba de bondad de ajuste de Chi Cuadrado, tiene una versión para estudiar la hipótesis de independencia entre dos variables categóricas.
- **Por ejemplo:** se desea investigar si en una organización, existe alguna relación entre el género y nivel de ingreso de sus trabajadores, siendo esta variable expresada en términos de niveles (bajo, medio, alto).
- Se ha recogido una muestra aleatoria de 1000 personas, cuya información fue resumida en la siguiente tabla:

	Nivel de Ingresos		
Género	Bajo	Medio	Alto
Masculino	182	213	203
Femenino	154	138	110

# $\chi^2$ PARA DATOS CATEGÓRICOS

## Ejemplo

- La información presentada de este modo es una tabla de doble entrada, tabla cruzada o de contingencia.

	Nivel de Ingresos		
Género	Bajo	Medio	Alto
Masculino	182	213	203
Femenino	154	138	110

## Estadístico:

- $$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

# $\chi^2$ PARA DATOS CATEGÓRICOS

	Nivel de Ingresos		
Género	Bajo	Medio	Alto
Masculino	182	213	203
Femenino	154	138	110

## Regla:

- $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$
- La regla indica: se rechaza la hipótesis nula para valores grandes del estadístico, esto es, superiores al valor crítico:  $\chi^2_{\alpha}$  con  $(\text{filas}-1) * (\text{columnas}-1)$  grados de libertad.