Estadística No Paramétrica

Clase 1

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Contenidos del curso

Estadística No Parámetrica:

- Introducción
- Generalidades sobre técnicas no paramétricas
- Pruebas para una muestra
- Pruebas para dos muestras independientes
- Pruebas para dos muestras relacionadas
- Pruebas para varias muestras independientes
- Pruebas para varias muestras relacionadas
- Jackknife, Bootstrap paramétrico y no paramétrico
- Regresión no paramétrica

Bibliografía

Referencias bibliográficas:

Obligatoria

- **Conover, W.J (1999)**. *Practical Nonparametric Statistics (3rd Ed.)*
- **Siegel S.** *Diseño experimental no paramétrico.*
- **Wasserman, L (2006)**. *All of Nonparametric Statistics*

Complementaria

- **Hollander, M and Wolfe D.A (1972)**. *Nonparametric Statistical Methods.*
- **Daniel W. W (1978)**. *Applied Nonparametric Statistics*

Tipos de evaluaciones

Tipo de evaluación	Ponderación (% del total)
Pruebas	60%
Presentaciones grupales e informes	20%
Tareas	10%
Co-evaluación	10%

Software

- R (principal del curso)
  - The R project: `www.r-project.org`
  - Disponible en Windows, MacOSX, Linux
- `Python` (uso opcional)

¿Por que usar R?

- R es un software de uso libre.
- No necesita una licencia.
- Cualquiera puede usar o modificar los códigos disponibles ('source').
- Sigue presentando un amplio desarrollo y crecimiento (a diferencia del software SPSS que ha ido disminuyendo su popularidad)
- Es uno de los softwares más utilizados por los Data Scientist para el análisis de datos y creación de modelos predictivos.

¿Para que sirve R?

R contiene una variedad de 'librerías' base para ser diferentes tipos de análisis estadísticos y más de 12000 librerías adicionales que hans sido desarrolladas. Estas librerías nos permiten trabajar con:

- Distribuciones de probabilidad.
- Test estadísticos.
- Modelado lineal, no lineal, semiparamétrico, no parametríco, etc.
- Análisis multivariado.
- Series de tiempo.
- Estadística espacial.
- Mapas.
- Machine learning, Deep learning.
- ....

Adicionales con R

Además de permitir realizar análisis estadísticos, `R` se ha convertido en un ambiente de desarrollo con extensiones tales como:

- Desarrollo de API's.
- Interfaz con `Shiny`.
- Interfaz con `LaTeX` mediante `Rmarkdown`.
- Interfaz con *c++* a través de `Rcpp`.
- Interfaz con álgebra lineal a través de `RcppArmadillo`.
- Interfaz con análisis númerico a través de `RcppNumerical`.
- Creación de páginas web con `blogdown`
- ....

Estadística paramétrica

Caracteristicas principales

- Los parámetros son desconocidos y fijos en el tiempo. Estos determinan la caractecteristicas de una población.
- La estimación y la inferencia estan basados en supuestos distribucionales en la función de distribucion.

Introducción

Estadística no paramétrica

Caracteristicas principales

- No se asume una forma conocida para la función de distribución.
- Se requieren pocos supuestos en el proceso de estimación y test para la población de estudio.
- Igualmente se realizan procesos de estimación y test de hipótesis para los parámetros de la población.

Introducción

Ventajas

- Se requieren pocos supuestos sobre los datos obtenidos sobre la población en estudio.
- Permite la estimación exacta en los test comparativos de los p-values y/o estimación exacta de los intervalos de confianza sin asumir supuesto de distribución Normal.
- Sin problemas de estimación en muestras pequeñas.
- Generalmente son sencillos de aplicar y sencillos de comprender.
- Relativamente insensible a valores atípicos.
- Puede ser apicable cuando la teoría de la distribución Normal no puede ser utilizada.
- Los avances computacionales permiten estimaciones eficientes en los test no paramétricos.

Función de distribución

Comencemos defininiendo una variable aleatoria Y la cual esta determinada por su función de distribución acumulada de la forma:

$$F(y) = P(Y \leq y)$$

Lo anterior es puede ser aplicado para una variable aleatoria discreta o una variable aleatoria continua.
La distribución de Y esta deteminada únicamente por:

- Función de densidad de probabilidad $(pdf) \Rightarrow f(y)$ si Y es v.a continua.
- Función de masa de probabilidad $(pmf) \Rightarrow f(y) = P(Y = y)$ si Y es una v.a discreta.

Función de distribución acumulada

CDF

$F(y) = P(Y \leq y)$ para una v.a continua.

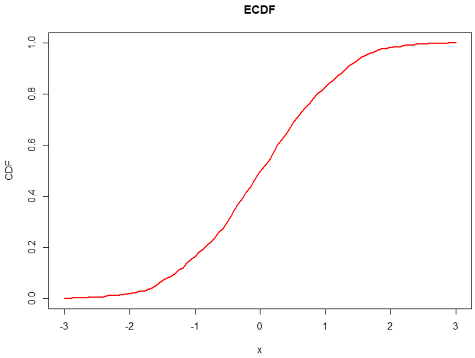
```
n = 1000
y = rnorm(n, mean = 0, sd = 1)
mean(y)
var(y)
```

```
x = seq(-3, 3, length = 100)
ecdf.fun = ecdf(z)  #Crea una CDF
class(ecdf.fun)#  La función CDF con el argumento 'class'
```

Función de distribución acumulada

CDF

$F(y) = P(Y \leq y)$ para una v.a continua.



Función de distribución acumulada

CDF

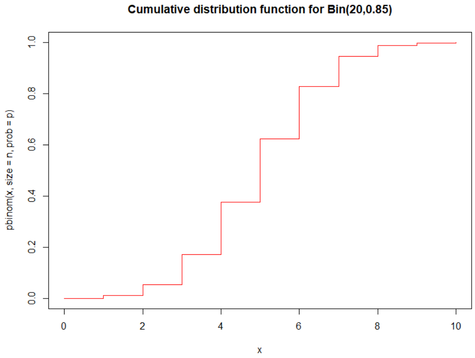
$F(y) = P(Y \leq y)$ para una v.a discreta es la misma que para una v.a continua pero a través de una función en 'intervalos'.

$$\mathtt{n} = 10$$
$$\mathtt{p} = 0.5$$
$$\mathtt{dbinom}(1, \mathtt{size} = \mathtt{n}, \mathtt{prob} = \mathtt{p})$$
$$\mathtt{x} < -0 : \mathtt{n}$$

Función de distribución acumulada

CDF

$F(y) = P(Y \leq y)$ para una v.a discreta es la misma que para una v.a continua pero a través de una función en 'intervalos'.



Función de distribución de probabilidad

PDF

Considerando una v.a aleatoria continua Y, la función de densidad de probabilidad (pdf, siglas en inglés) denotada como $f(y)$, determina la región más probable.

$P(a < Y \leq b) = F(b) - F(a) = \int_a^b f(y)dy$

Métodos paramétricos

Si nos enfocamos en los tradicionales métodos paramétricos, la función de distribución está gobernada por parámetros, por ejemplo:

- Distribución Normal: $\mathscr{N}(\mu, \sigma^2)$
- Distribución de Poisson: λ
- Distribución Gamma: $Ga(a, b)$

Métodos paramétricos

Comparación de medias de dos grupos

Si asumimos que tenemos dos muestras aleatorias desde dos grupos, llamemosles $y_1, ..., y_n$ y $z_1, ..., z_m$, para observaciones independientes una de otra. Para determinar si las medias son distintas podemos realizar un clásico test paramétrico bajo los siguientes supuestos:

- $Y_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- $Z_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- $\sigma_1 = \sigma_2 = \sigma_{\text{total}}$

Métodos paramétricos

Test de hipótesis

- Generalmente tenemos una hipótesis nula (por ejemplo, asumir que las medias de los grupos son iguales).
- Si la hipótesis nula se cumple entonces el 'estadístico - t' tiene cierta distribución de probabilidad.
- Observamos el valor actual de 't'.
- Determinamos que tan distinto es este valor comparado con la distribución nula del 'estadístico - t'.

Métodos paramétricos

Test de hipótesis: ejemplo

Comparación de dos medias en dos grupos distintos

- Calculamos el 'estadístico - t' para las dos muestras aleatorias observadas (independientes):
  $t = \frac{\bar{y}-\bar{z}}{\hat{\sigma}\sqrt{\frac{1}{m}+\frac{1}{n}}}$
  donde $\bar{y}$ e $\bar{z}$ son las medias de cada muestra y $\hat{\sigma}$ es una estimación de la desviación estándar.
- t es el 'estadístico - t' para estas dos muestras y T la correspondiente variable aleatoria.
- Hipótesis nula $\Rightarrow \mu_1 = \mu_2$ para las muestras y y z respectivamente.
- Si la Hipótesis nula es verdadera entonces T tiene una distribución t_{n+m-2}.

Métodos paramétricos

p - value

- Para verificar la Hipótesis nula, calculamos la probabilidad que T podría tomar valores en los extremos de los valores observados.
- Esta probabilidad es conocida como $p - value$.
- La distribución de probabilidad utilizada es la distribución que tomaría T si la Hipótesis nula fuera cierta.
- Por ejemplo: Si la distribución de T es simétrica en torno a 0 y, además observamos que $T = t$, entones el $p - value$ (en ambos extremos): $p = P(|T| \geq |t|)$

Métodos paramétricos

Test de Hipótesis

Para llevar a cabo un test de Hipótesis necesitamos comparar el $p - value$ con un valor dado. Este valor dado le llamamos nivel de significancia

- El nivel de significancia generalmente se denota por α con un valor de $\alpha = 0.05$.
- Si $p - value < \alpha$ entonces decimos que hay evidencia para rechazar la Hipótesis nula. Esto por que el valor observado t era poco probable si la Hipótesis nula fuera cierta.

Métodos paramétricos

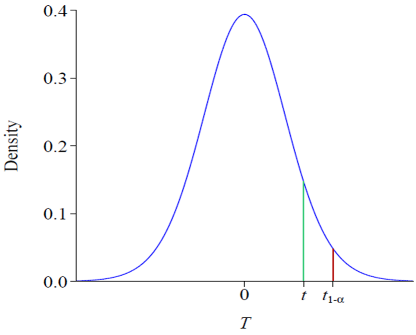
Test de Hipótesis: ejemplo

- Si $p - value < \alpha$ entonces decimos que hay evidencia para rechazar la Hipótesis nula. Esto por que el valor observado t era poco probable si la Hipótesis nula fuera cierta.

Métodos paramétricos

Test de Hipótesis: ejemplo

- $p - value$ es el área al lado derecho de t
- α es el área al lado derecho de $t_{1-\alpha}$.
- No podemos rechazar la Hipótesis nula si $t < t_{1-\alpha}$.



Métodos paramétricos

Observaciones del 'estadístico - t' (t)

- Todos los supuestos se satisfacen.
- Puede llevar a errores ante muestras pequeñas.
- El téorema del límite central puede ayudar para muestras grandes.
- Si cada Y_i y Z_i no son se asumen como normalmente distribuidos, las medias muestrales son aproximadamente normales para muestras grandes.

este es el final.... por ahora....