

ÍNDICE DE LA PRESENTACIÓN

- 1.- Introducción
- 2.- Índices de centralización
Tipos, Propiedades, Selección
- 3.- Índices de dispersión
Tipos, Propiedades, Selección
- 4.- Estimación de la distribución de mediciones

INTRODUCCIÓN AL ANÁLISIS DE MEDICIONES

Los experimentos de evaluación del funcionamiento de los computadores basados en mediciones generan muchos datos

Es necesario resumir los datos medidos
Este proceso se denomina ANÁLISIS DE DATOS
No confundir con el ANÁLISIS DEL FUNCIONAMIENTO

Después de tomar una muestra de una variable $\{x_1 \dots x_n\}$
Hay que caracterizar la muestra mediante 2 valores:

$\left\{ \begin{array}{l} \text{Su valor medio} \Rightarrow \text{Índices de centralización} \\ \text{Su dispersión} \Rightarrow \text{Índices de dispersión} \end{array} \right.$

Es muy importante ...

- 1) Saber cuándo/cómo usar índices de centralización/dispersión
- 2) Reconocer si se han usado correctamente los índices

ÍNDICES DE CENTRALIZACIÓN

Un índice de centralización es un número que representa de un modo la tendencia central que siguen las observaciones de una muestra

Media

Suma de las observaciones / número de Obs

$$x = \frac{1}{n} \sum_{i=1}^n x_i$$

Mediana muestral

Valor que cae en el centro de las observaciones tras ordenarlas
Si nº Obs es par tomar la media de las 2 Obs centrales

Moda muestral

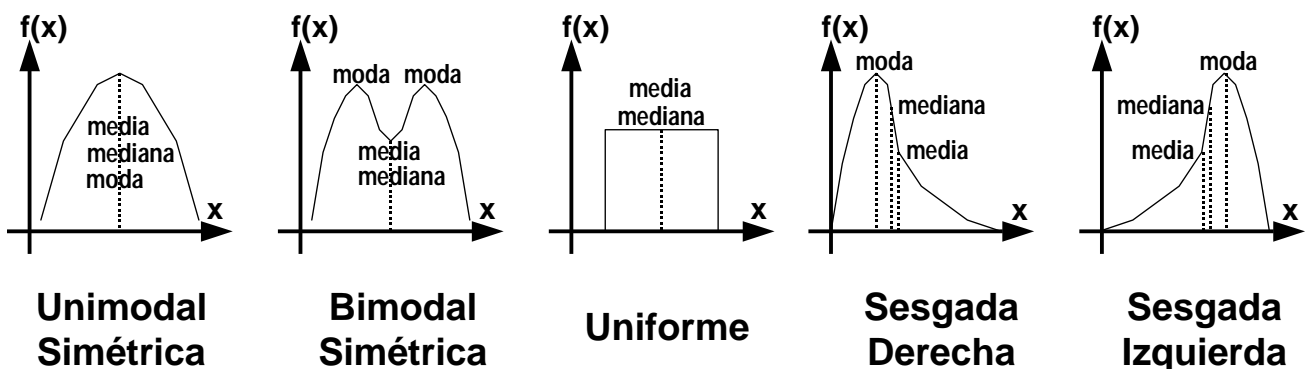
Es el centro de la clase con mayor frecuencia en un histograma
Es el valor (categoría) que aparece con mayor frecuencia

PROPIEDADES DE ÍNDICES DE CENTRALIZACIÓN

Existencia y Unicidad

La Media y la Mediana EXISTEN siempre y son ÚNICAS
Moda puede NO EXISTIR y puede haber VARIAS modas

Relaciones y diferencias



PROPIEDADES DE ÍNDICES DE CENTRALIZACIÓN

Efecto de observaciones anómalas (outliers) sobre los índices

MEDIA

Usa todas las Obs de la muestra ponderándolas por igual
Le afectan MUCHO las Obs anómalas (más con pocas Obs)

MEDIANA y MODA

Ignoran mucha información sobre las observaciones
Les afecta POCO la presencia de Obs anómalas

Tiempos de ejecución { 10, 20, 15, 18, 16 }

media = 15.8

mediana de { 10, 15, 16, 18, 20 } = 16

Al añadir la 6 observación (anómala) { 200 }

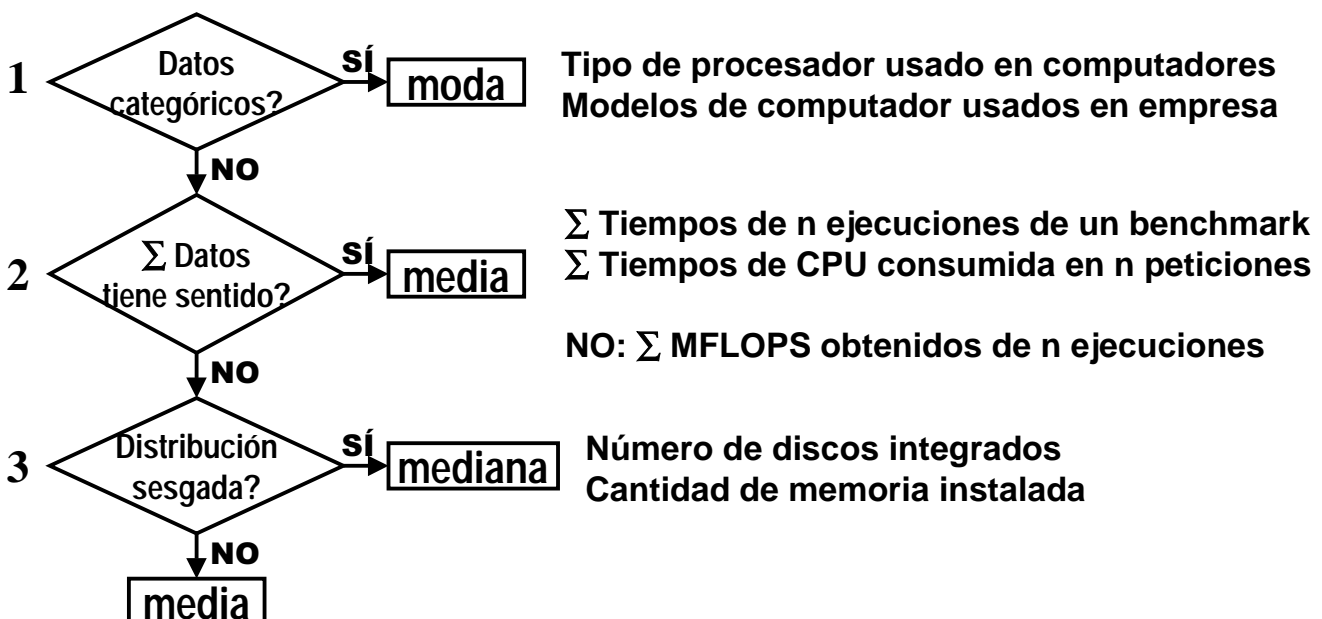
media = 46.5 que supera a casi todas las observaciones

mediana de { 10, 15, 16, 18, 20, 200 } = (16+18) / 2 = 17



SELECCIÓN DE UN ÍNDICE DE CENTRALIZACIÓN

El mejor índice = f(tipo + características) de los datos



ÍNDICES DE DISPERSIÓN

Un índice de dispersión es un número que representa de un modo la variabilidad de las observaciones de una muestra

Rango

$$\text{Rango} = \text{máx}(x_i) - \text{mín}(x_i)$$

Número de discos integrados
Cantidad de memoria instalada

Varianza

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

Desviación estándar

$$s = \sqrt{s^2}$$

Coef de Variación

$$COV = \frac{s}{\bar{x}}$$

ÍNDICES DE DISPERSIÓN

Percentiles

El percentil α es el valor de la variable observada x_α tal que el $\alpha\%$ de las observaciones son $\leq x_\alpha$

$$\Pr(x \leq x_\alpha) = \alpha\%$$

Percentiles 5 y 95 (ó 10 y 90) \Leftrightarrow Mínimo y Máximo

Ventaja: Calculables para todas las variables (p.j. las inacotadas)

Cuartiles

Q1, Q2, Q3 = Percentiles 25, 50 y 75

Semirango Intercuartílico SRI = $(Q3 - Q1) / 2 = (x_{0.75} - x_{0.25}) / 2$

Indica el rango que concentra el 50% de las observaciones

Desviación absoluta media

$$DAM = \frac{1}{n-1} \sum_{i=1}^n |x_i - \bar{x}|$$

PROPIEDADES DE ÍNDICES DE DISPERSIÓN

Efecto de observaciones anómalas (outliers) sobre los índices

| Índice | Efecto o impacto |
|----------|---|
| Rango | Muchísimo: las anomalías definen el propio rango |
| Varianza | Mucho: al elevar al cuadrado las desviaciones |
| DAM | Moderado: al usar el valor absoluto de las desviaciones |
| SRI | Poco impacto: al afectar poco una Obs a los cuartiles |

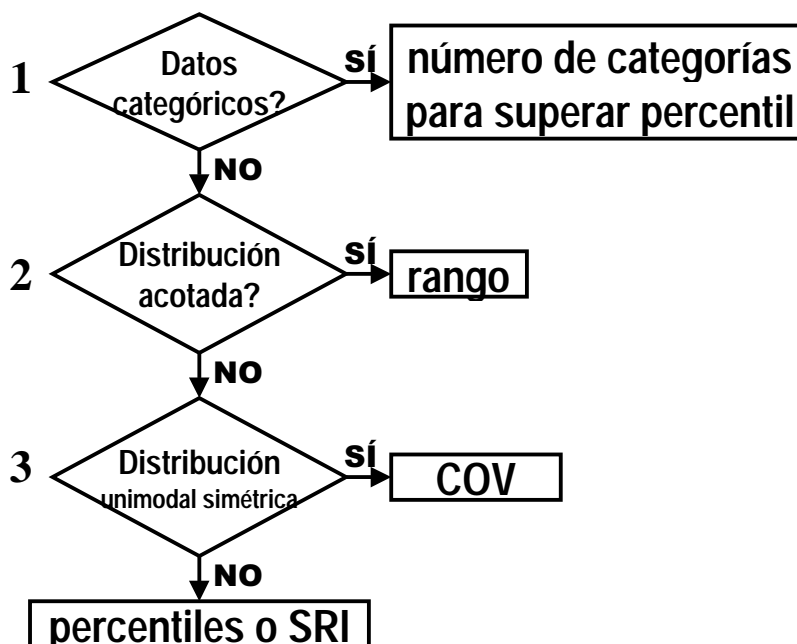
Si la distribución es muy sesgada \Rightarrow Hay muchas Obs con dispersión alta

Entonces: (Mediana + SRI) es mejor que (Media + Desviación estándar)

En general, si se usa la mediana como IC se usa el SRI como ID

SELECCIÓN DE UN ÍNDICE DE DISPERSIÓN

El mejor índice = $f(\text{tipo} + \text{características})$ de los datos



ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

1. Introducción

Problema: Los modelos necesitan datos de entrada.

Ej.: cadencia de llegadas, tamaño de las peticiones, etc.

Solución: Medir datos de entrada y utilizarlos para especificar una distribución. Aproximaciones (de peor a mejor):

- Utilizar directamente los **valores medidos**. Problema: limitados a los datos disponibles, que habitualmente son escasos
- Utilizar los datos para definir una **distribución empírica**. Problemas:
 - a) Sólo hay valores entre el máx. y el mín. medidos
 - b) Posibles irregularidades (cuando hay pocos datos)
- Utilizar una técnica estándar de inferencia para ajustar los valores a una **distribución teórica**.
 - Ventaja: forma compacta de representar los datos.
 - Problema: a veces no se encuentra una distribución teórica que ajuste

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

2. Distribuciones de probabilidad útiles

Consultar tablas para conocer uso y parámetros de distribuciones útiles en simulación. Ej.: Exponencial para tiempo entre llegadas

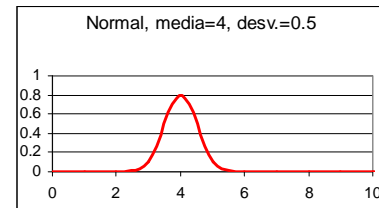
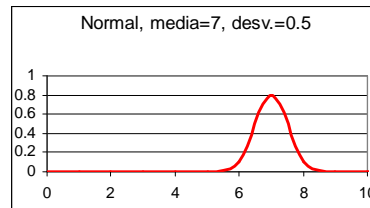
Parámetros de las distribuciones continuas:

- De **localización**: indican la posición en el eje x de la distribución. Habitualmente se da la media o el extremo inferior
- De **escala**: determinan la escala de la distribución. Un cambio en este parámetro comprime o expande
- De **forma**: determinan la forma de la distribución

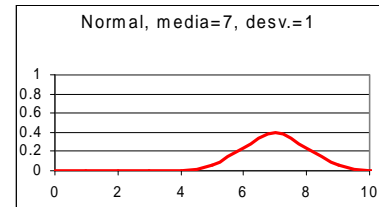
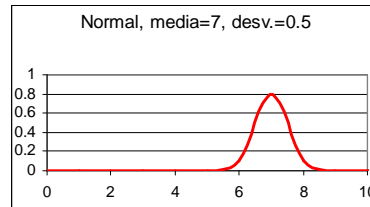
ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

2. Distribuciones de probabilidad útiles

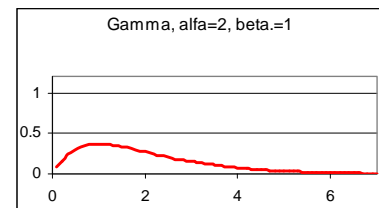
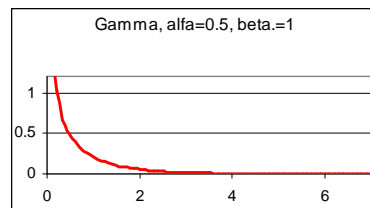
Variación de la localización



Variación de la escala



Variación de la forma



ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

2. Distribuciones de probabilidad útiles

Distribuciones empíricas

Se usan cuando no se puede ajustar ninguna distribución teórica

Se ordenan los puntos medidos y se unen por tramos lineales

Sea $X_{(i)}$ el i -ésimo valor más pequeño de las observaciones. La distribución empírica F viene dada por:

$$F(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ \frac{i-1}{n-1} + \frac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & \text{si } X_{(i)} \leq x < X_{(i+1)} \quad \forall i = 1, 2, \dots, n-1 \\ 1 & \text{si } x \geq X_{(n)} \end{cases}$$

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

3. Técnicas para establecer la independencia de las muestras

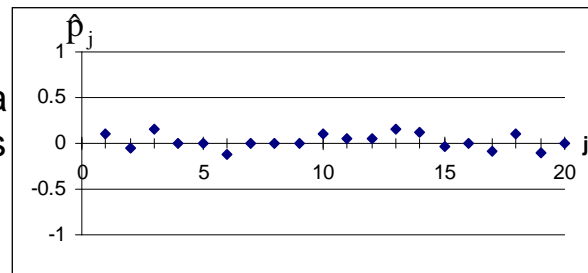
Necesidad de estas técnicas: los métodos de estimación de distribuciones necesitan en muchos casos que los datos sean independientes

Técnica 1: Gráfico de correlación

$$\hat{\rho}_j = \text{correlación de la muestra} = \frac{\hat{C}_j}{S^2(n)} \quad \text{donde} \quad \hat{C}_j = \frac{\sum_{i=1}^{n-j} (X_i - \bar{X}(n))(X_{i+j} - \bar{X}(n))}{n-j}$$

Cuanto más se aleje $\hat{\rho}_j$ de cero, más correladas estarán las muestras

Ejemplo: Gráfico de correlación para datos exponenciales independientes



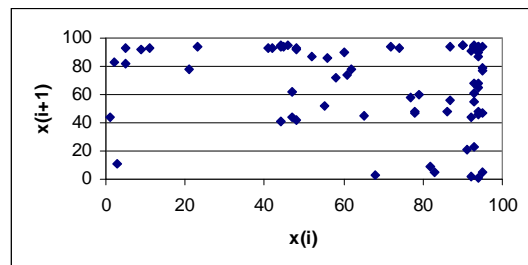
ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

3. Técnicas para establecer la independencia de las muestras

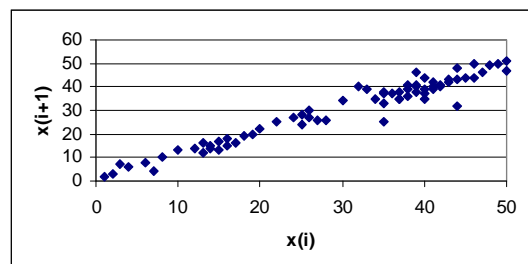
Técnica 2: Diagrama de dispersión

Representar los pares (x_i, x_{i+1}) para $i=1, 2, 3, \dots, n-1$

Si son independientes,
entonces distribución aleatoria



Si son dependientes,
entonces siguen una línea



ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

Tres pasos:

- 1) Seleccionar una distribución o familia de distribuciones
- 2) Obtener parámetros para la distribución
- 3) Contrastar que se ajusta razonablemente a los datos de partida

Paso 1: Selección de una familia de distribuciones

Usar conocimiento del problema y técnicas heurísticas

Técnica 1: Resumen estadístico

| Estadístico | Tipo de distr. | Uso |
|---|----------------|--|
| Mínimo, máximo | C, D | Aproximación del rango |
| Media (μ), mediana ($x_{0.5}$) | C, D | Si coinciden, entonces dist. simétrica |
| Varianza (σ^2) | C, D | Medida de la variabilidad |
| Coefficiente de variación ($CV=\sigma/\mu$) | C | CV tiende a 1 sugiere exponencial CV > 1 sugiere Weibull o Gamma |
| Razón de Lexis ($\tau=\sigma^2/\mu$) | D | $\tau = 1$ sugiere Poisson $\tau < 1$ sugiere Binomial $\tau > 1$ sugiere Binomial negativa |
| Coefficiente de asimetría (v) | C, D | $v = 0$ implica distr. simétrica $v > 0$ implica distr. sesgada a la dcha. $v < 0$ implica distr. sesgada a la izda. |

C=Continua
D=Discreta

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

Técnica 2: Histogramas y gráficos de líneas

Comparar histogramas y gráficos de líneas con los de las dist. teóricas

Problema de los histogramas: seleccionar ancho de celda. Directrices:

- Todos los intervalos deben ser iguales
- Han de eliminarse los puntos muy dispersos
- No utilizar tamaño ni muy grande ni muy pequeño

Regla empírica: N^0 de intervalos $\equiv k = 1 + \log_2 n$ con $n \equiv$ número de datos

Utilidad dudosa de esta regla. Solución: Probar con varios intervalos

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

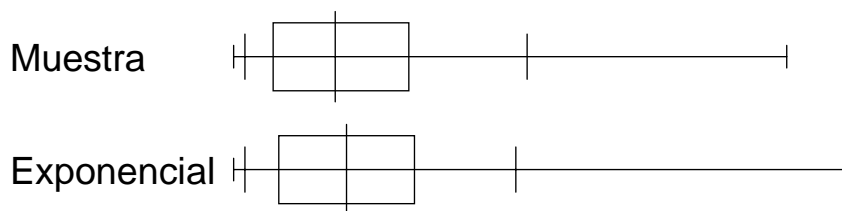
Técnica 3: Resumen de cuantiles y representación por cajas

Objetivo: determinar si la distr. es simétrica o desviada a dcha. o izda.

Cuantil q de $F(x) \equiv X_q$ tal que $F(X_q)=q$

| Cuantil | Profundidad | Valores | Punto medio |
|-----------|-------------------------------|-----------------------------|-----------------------------|
| Mediana | $i=(n+1)/2$ | $x_{0.5}$ | |
| Cuartiles | $j=(\lfloor i \rfloor + 1)/2$ | $x_{0.25} \quad x_{0.75}$ | $(x_{0.25} + x_{0.75})/2$ |
| Octiles | $k=(\lfloor j \rfloor + 1)/2$ | $x_{0.125} \quad x_{0.875}$ | $(x_{0.125} + x_{0.875})/2$ |
| Extremos | 1 | $x_{(1)} \quad x_{(n)}$ | $(x_{(1)} + x_{(n)})/2$ |

Si los puntos medios son similares, entonces la distr. es simétrica



ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

Paso 2: Estimación de parámetros

Se utiliza el estimador de máxima verosimilitud (MLE). Para algunas distr. se calcula de forma sencilla:

- Obtener función de probabilidad : $L(\theta)$ (θ representa el estimador)
- Obtener la función de probabilidad logarítmica: $l(\theta)=\ln[L(\theta)]$
- Derivar con respecto al parámetro: $dl/d\theta$
- Igualar a cero y resolver la ecuación: $(dl/d\theta)=0$

Cuando no es sencillo, mirarlo en las tablas

Se utiliza el MLE porque cumple ciertas propiedades estadísticas convenientes

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

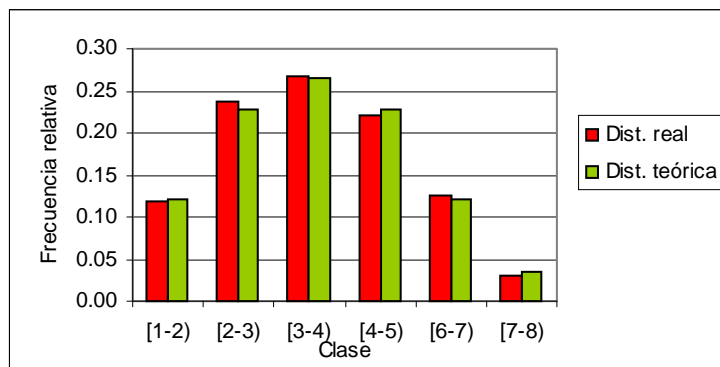
4. Estimación de una función de distribución

Paso 3: Determinar la representatividad del ajuste

Se utilizan métodos heurísticos y técnicas de análisis de hipótesis

Técnica 1: Comparación de frecuencias

Representar histograma de la distr. real y de la teórica para el mismo número de intervalos y tamaño de intervalos



$$\text{dist. real} = \frac{\text{n}^\circ \text{ de datos en el intervalo}}{\text{n}^\circ \text{ de datos total}}$$

$$\text{dist. teórica en el intervalo } j = \int_{b_{j-1}}^{b_j} \hat{f}(x) dx$$

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

Técnica 2: Gráficos de probabilidades

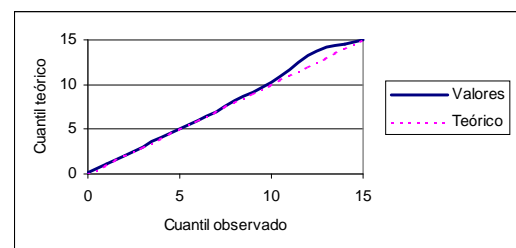
2.1 Gráfico cuantil-cuantil (Q-Q)

Siendo $\hat{F}(x)$ la dist. teórica, $\tilde{F}(x)$ la dist. real y $q_i = (i-0.5)/n$ para $i=1, 2, \dots, n$

Se dibuja:

eje-x: $x_{q_i}^S = \tilde{F}^{-1}(q_i)$

eje-y: $x_{q_i}^M = \hat{F}^{-1}(q_i)$

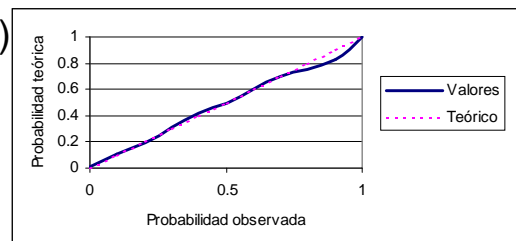


2.2 Gráfico Probabilidad-Probabilidad (P-P)

Para diversos x_i se dibuja:

eje-x: $\tilde{F}(x_{(i)})$

eje-y: $\hat{F}(x_{(i)})$



ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

Técnica 3: Comparación de gráficos de cajas

Los gráficos de cajas tienen que ser similares. Para la distr. teórica, usar como extremos inicial y final $1/(2n)$ y $1-(1/(2n))$ respectivamente

Técnica 4: Comparación mediante test de hipótesis χ^2

H_0 : Las X_i son variables aleatorias IID con función de distribución \hat{F}

Un fallo al rechazar no debe interpretarse como " H_0 es cierta" porque para n pequeño el test no es muy sensible a las diferencias

Rechazar la hipótesis tampoco debe interpretarse como " H_0 es falsa" porque para n grandes rechaza casi siempre

ESTIMACIÓN DE LA DISTRIBUCIÓN DE MEDICIONES

4. Estimación de una función de distribución

Pasos:

- Se dividen los valores en k intervalos adyacentes: $[a_0, a_1), \dots, [a_{k-1}, a_k)$ donde a_0 puede ser $-\infty$ y a_k puede ser $+\infty$
- Se calcula $N_j = n^o$ de X_i que pertenecen a cada intervalo $[a_{j-1}, a_j)$
- Se calcula la proporción esperada p_j de las X_i que caerían en el intervalo j a partir de la distribución objetivo

- Distribuciones continuas: $p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x) dx$

- Distribuciones discretas: $p_j = \sum_{a_{j-1} \leq x_i < a_j} \hat{p}(x_i)$

- Se calcula el estadístico del test: $\chi^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}$

Cuanto menor sea χ^2 mayor la semejanza de la distr. real a la teórica

Normalmente se intenta que $p_j = 1/k \forall k, k \geq 3$ y $np_j \geq 5$

Se rechaza si $\chi^2 > \chi^2_{k-1, 1-\alpha}$.