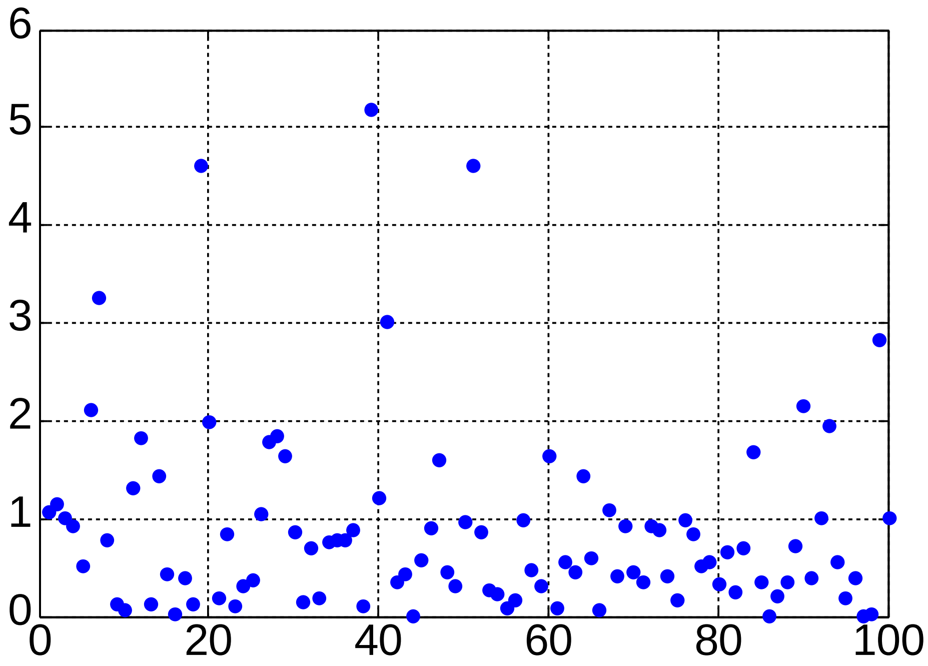
# Presentando los Datos

**Modelado y Análisis de Redes de Telecomunicaciones**

FACULTAD DE INGENIERIA
UNIVERSIDAD DE LA REPUBLICA
URUGUAY

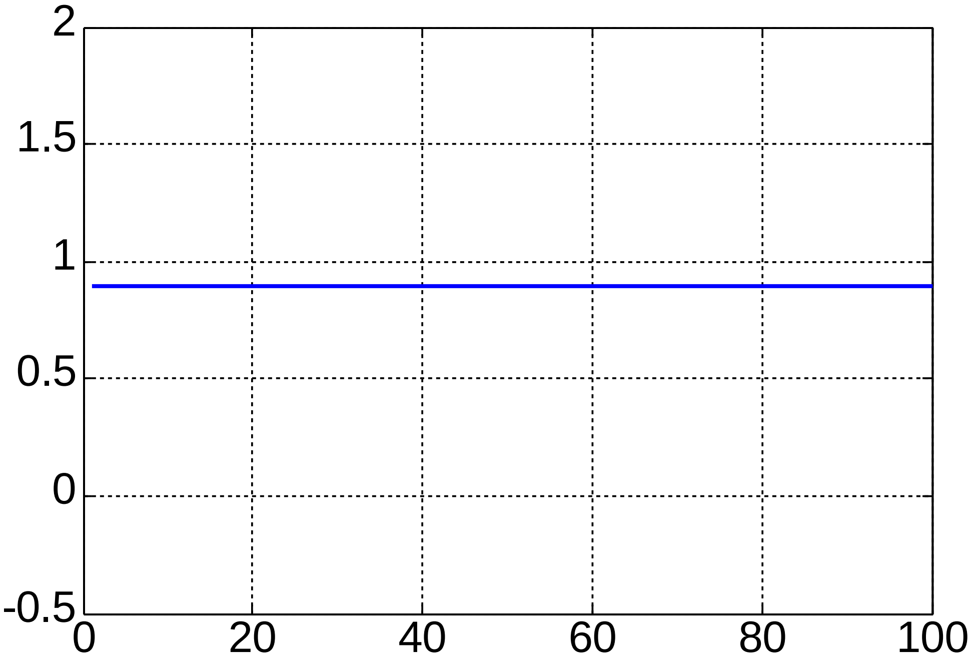
# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿todas?



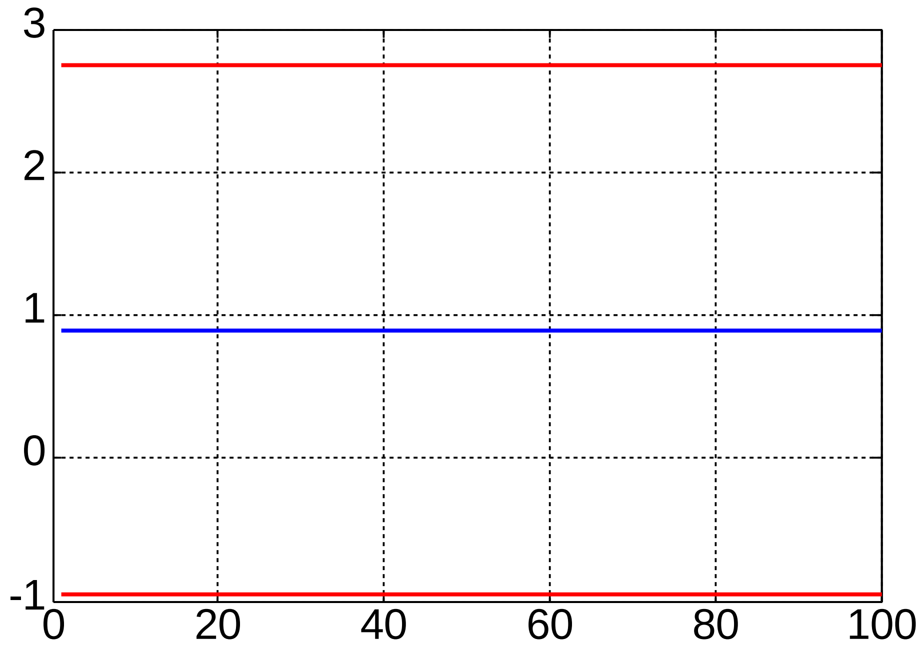
# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿el promedio?



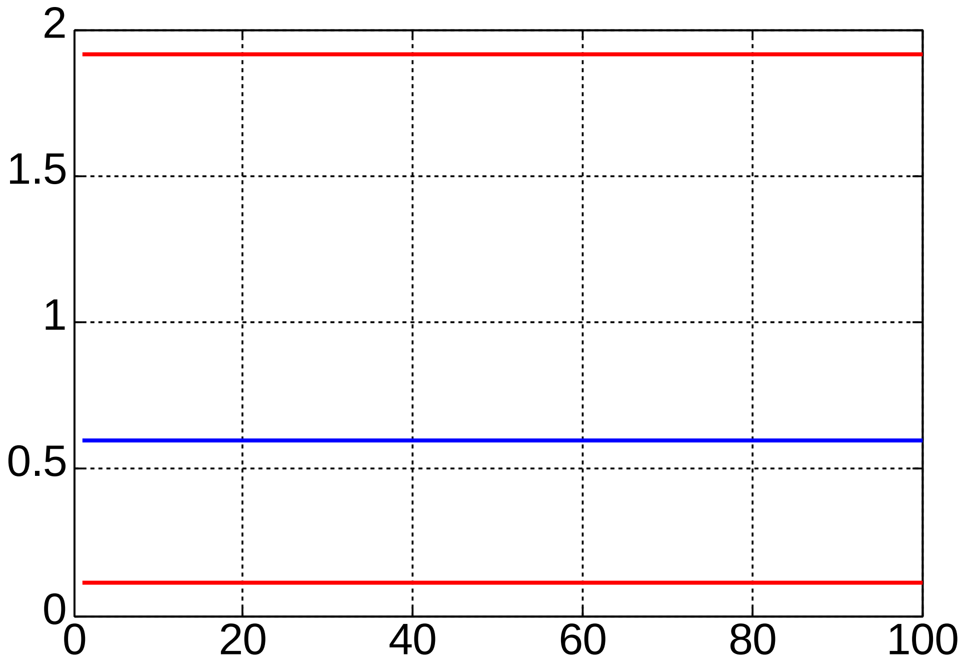
# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿asumo que los datos son normales y le muestro la media más/menos dos veces la varianza?



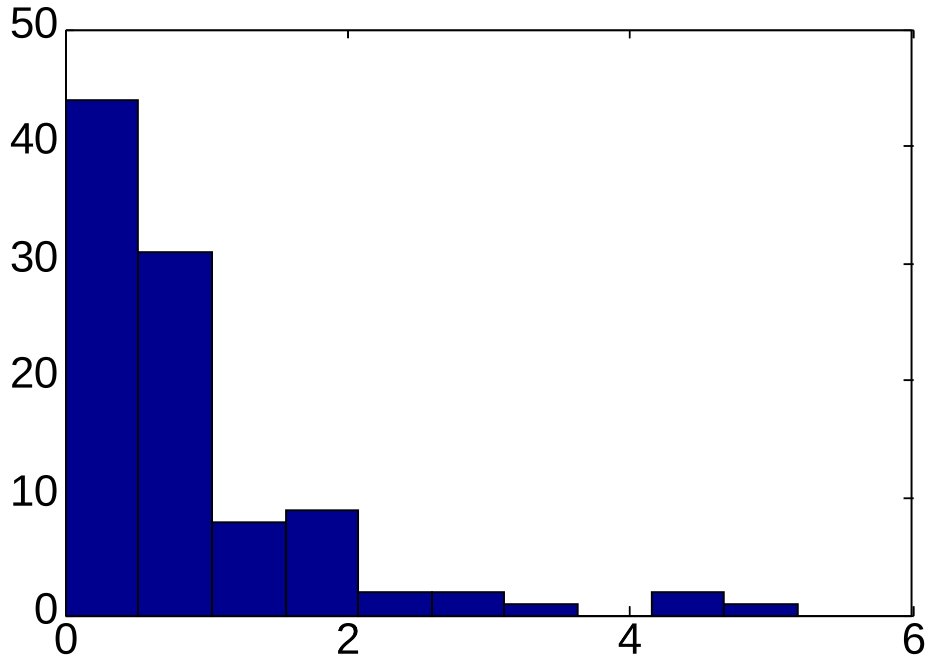
# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿muestro la mediana y un par de cuantiles?



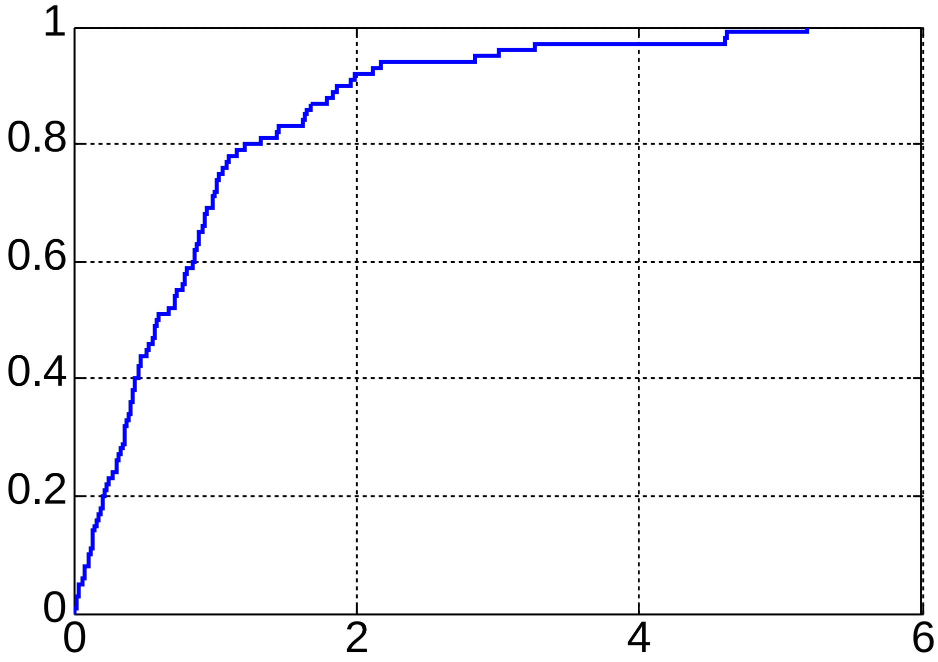
# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿muestro un histograma?



# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿muestro un distribución empírica?



# Presentando los Datos

- Estoy haciendo el benchmarking de un procesador
  - Ejecuto cien veces la misma serie de operaciones y mido el tiempo que demora en ejecutarlas
  - ¿qué le muestro a mi jefe?
    - ¿Qué garantías tengo acerca de la correctitud o sentido de lo que mostré?
    - I.e. ¿El jefe me pregunta “¿y porqué 100 valores?” y qué le digo?

# Agenda

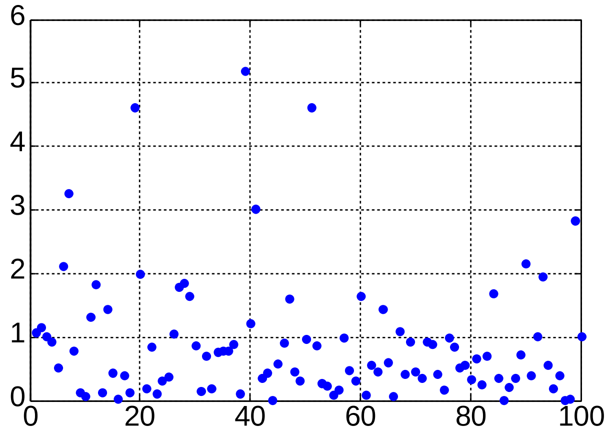
- Histogramas y distribuciones empíricas
- (Media,Varianza) y (Mediana,Cuantiles)
- Intervalos de Confianza
- Intervalos de Predicción
- Comentario Finales

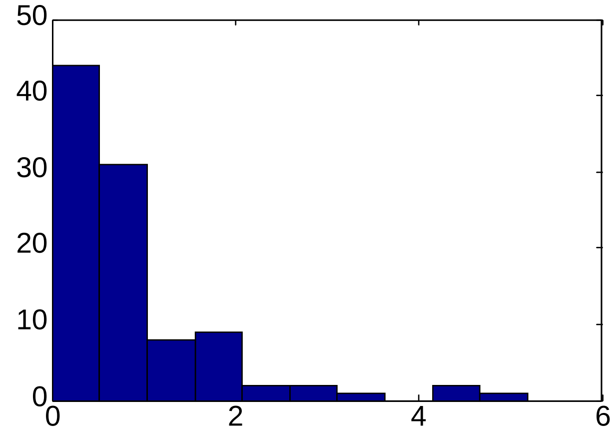
# Histograma

- Sean $x_1,\ldots,x_n$ $n$ realizaciones de un experimento
- Sean $m=\min\{x_1,\ldots,x_n\}$ y $M=\max\{x_1,\ldots,x_n\}$
  - Tomo $k$ sub-intervalos (*bins*) $B_i$ ($i=1,\ldots,k$) de ancho $w=(M-m)/k$ como $B_i = (\ m+(i-1)w\ ;\ m+iw\ ]$
  - La función histograma $h(x)$ se define para $x \in B_i$ como:

$$h(x) = \sum_{i=1}^{n} 1_{\{x_i \in B_i\}}$$

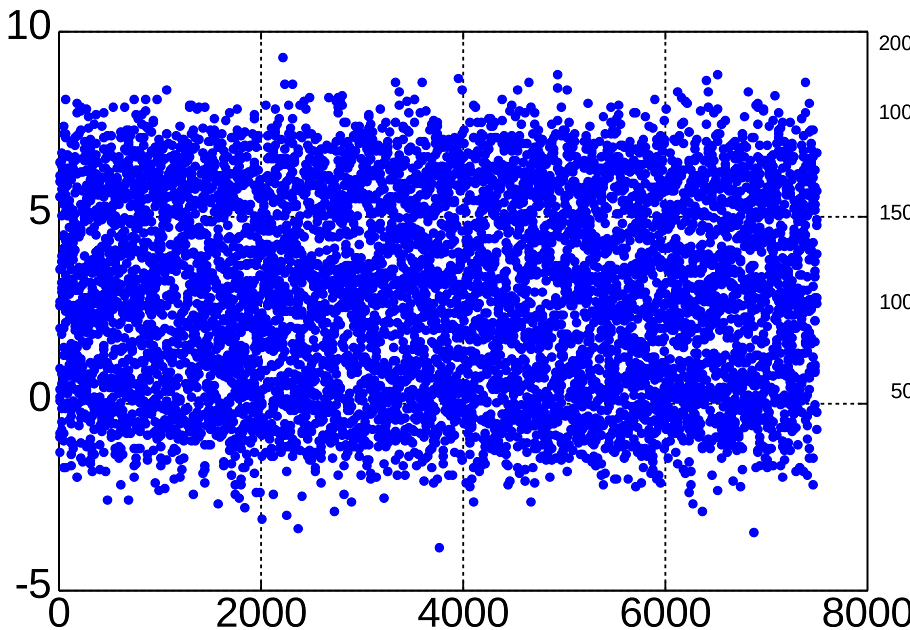
i.e. la cantidad de datos que cayeron en el *i*-ésimo bin

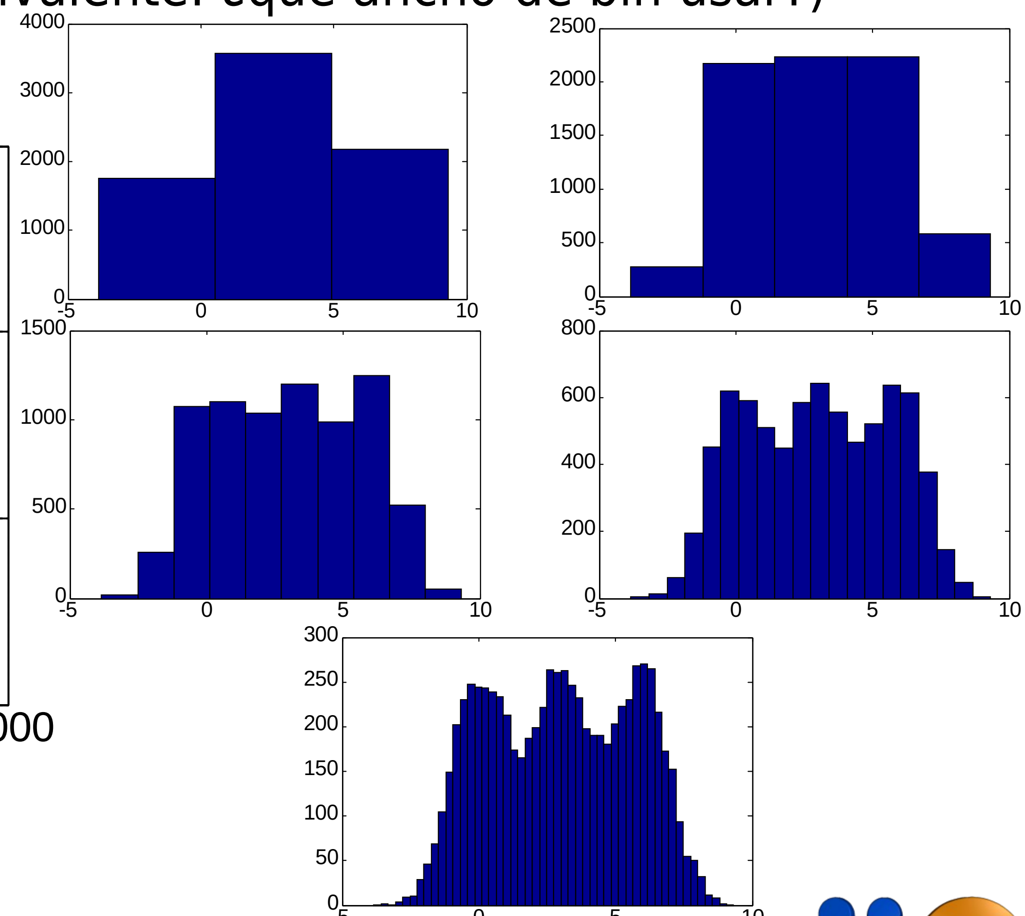


# Histograma

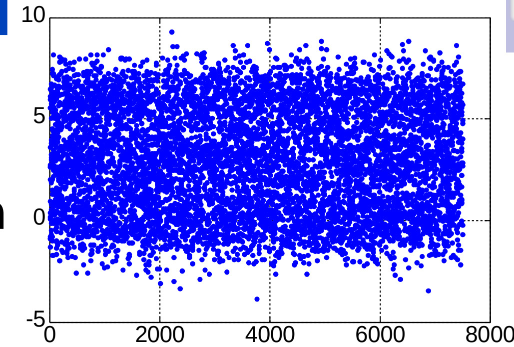
- Dos problemas:
  - No es suave
  - ¿cuántos bins usar? (o equivalente: ¿qué ancho de bin usar?)

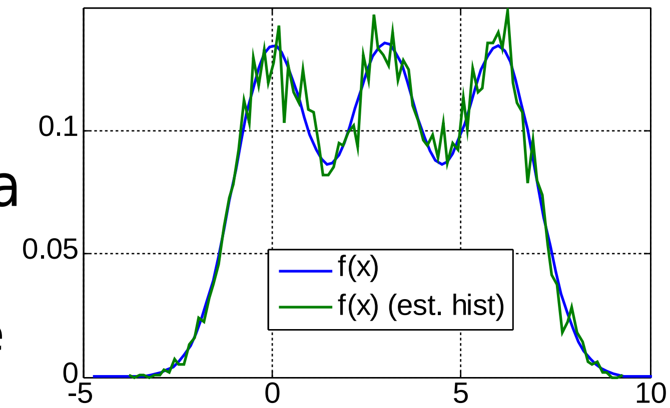


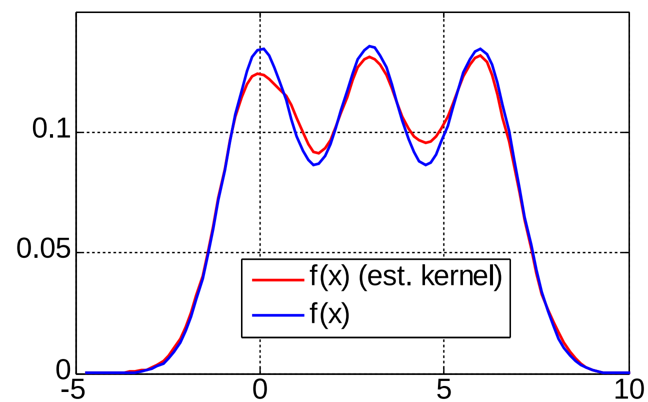
# Histograma

- ¿Qué es el histograma?
  - Sean $X_1,\ldots,X_n$ V.A. iid con densidad $f(x)$ y $x_1,\ldots,x_n$ una realización
  - Si x está en $B_i$ entonces $h(x)/n \rightarrow P(X_i \in B_i) \approx f(x)w$
  - Por lo tanto $h(x)$ no es más que una aproximación tosca de la densidad (a excepción del re-escalado)
- Existen otras técnicas más sofisticadas para estimar la densidad de $X_i$
  - Ejemplo: estimación por kernels (donde $K(x)$ puede ser por ejemplo la campana de Gauss o un rectángulo de ancho $h$)



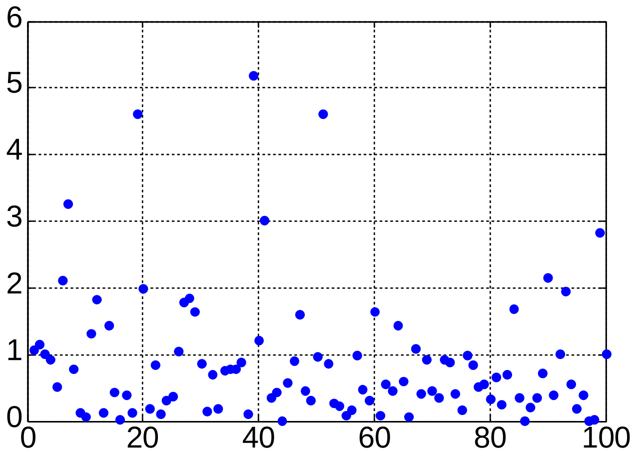

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

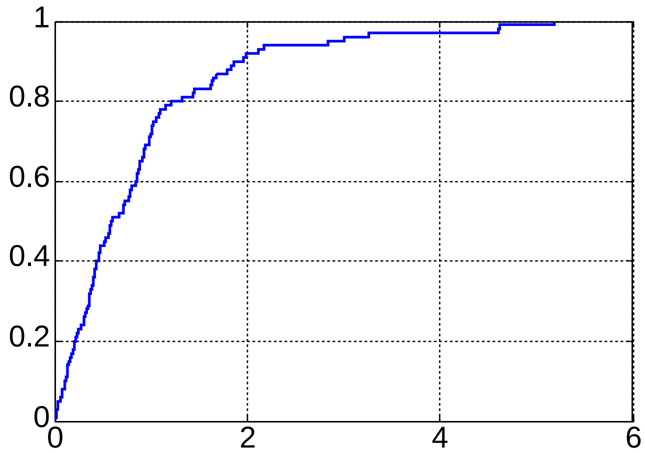


# Distribución Empírica

- La distribución empírica $F_n(x)$ de los valores $x_1,\ldots,x_n$ se define como:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$$




- Sean $X_1,\ldots,X_n$ V.A. iid con distribución $F(x)$ y $x_1,\ldots,x_n$ una realización
  - Entonces $F_n(x)$ es una estimación de la función de distribución de $X_i$

# Distribución Empírica

- Sean $X_1,\ldots,X_n$ V.A. iid con distribución $F(x)$ y $x_1,\ldots,x_n$ una realización
  - Entonces $F_n(x)$ es una estimación de la función de distribución de $X_i$
- Esto se justifica por el teorema de Glivenko-Cantelli

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \underset{n}{\to} 0 \quad \text{c.s.}$$

  - Prueba parcial (¿Porqué no es completa?):

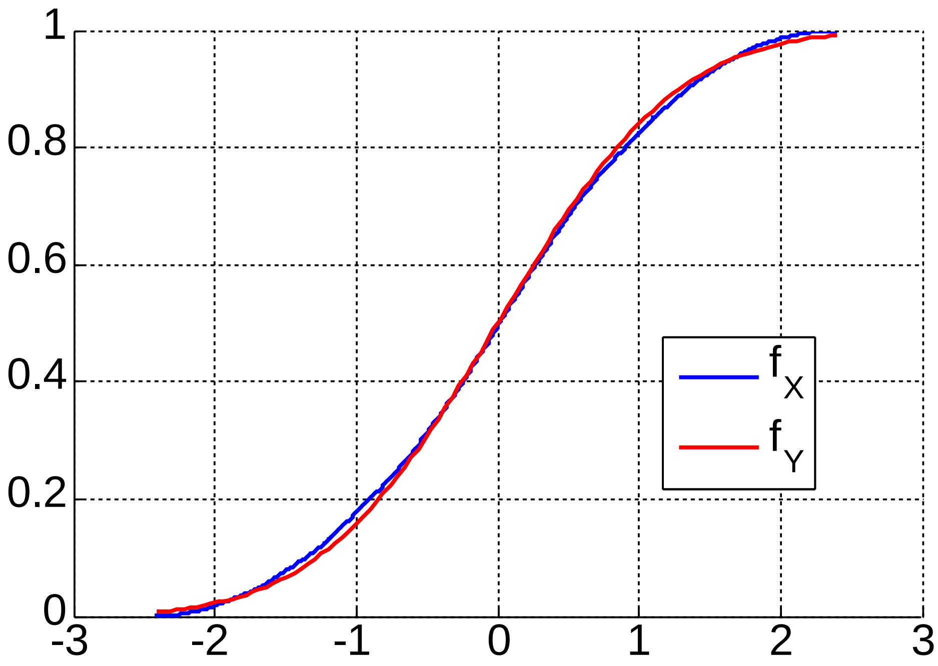
$$Y_i = 1_{\{X_i \le t\}} \Rightarrow E\{Y_i\} = P(X_i \le t) = F(t)$$

$$\underset{\text{LFGN }(Y_i \text{ iid})}{\Rightarrow} F_n(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i \underset{n}{\to} E\{Y_1\} = F(t) \quad \text{c.s.}$$

- Por lo general es preferible utilizar la ECDF que una estimación de densidades pues la primera no tiene parámetros

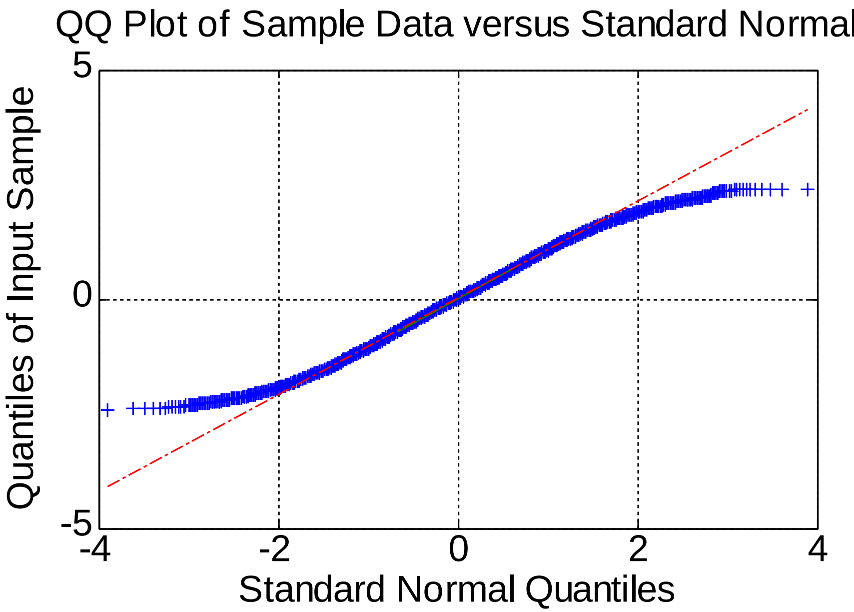
# Comparación de Distribuciones

- Hasta ahora veníamos comparando distribuciones graficando sus distribuciones
  - ¿qué tiene de malo?
  - E.g. $X=(X_1+X_2)/2$ con $X_i$~U[-2.45,2.45] e $Y \sim N_{0,1}$



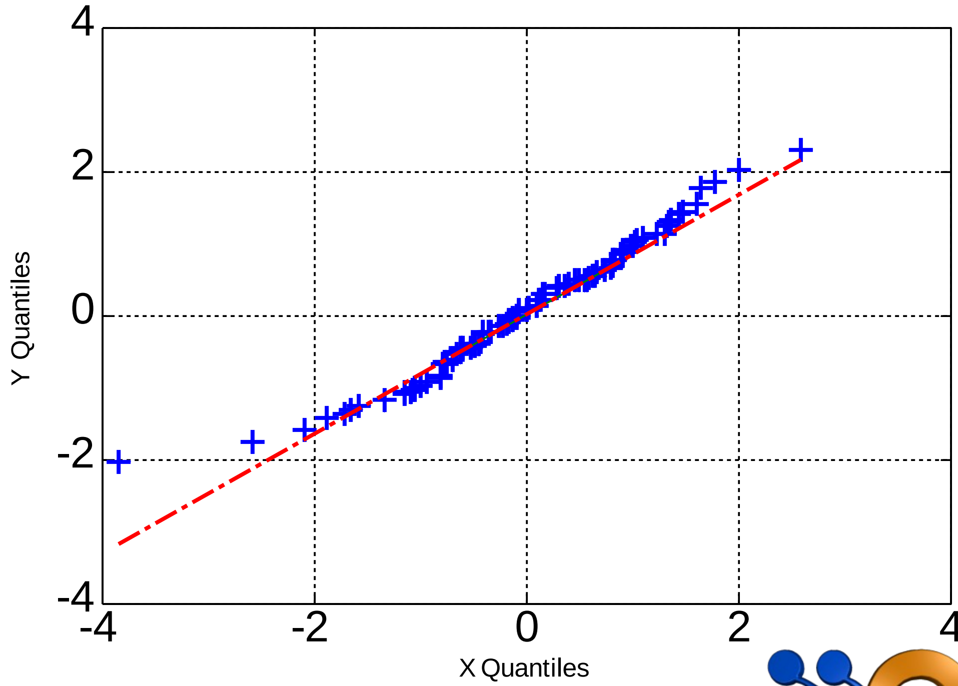
# Comparación de Distribuciones

- Para una comparación visual de las distribuciones es mejor el **QQ plot**
- La idea es graficar para ciertos valores de *q* el valor del cuantil de orden *q* de las dos distribuciones $(\alpha_q^X, \alpha_q^Y)$
- Si las distribuciones son iguales (salvo un re-escalado y/o una traslación) entonces los puntos debería seguir una recta



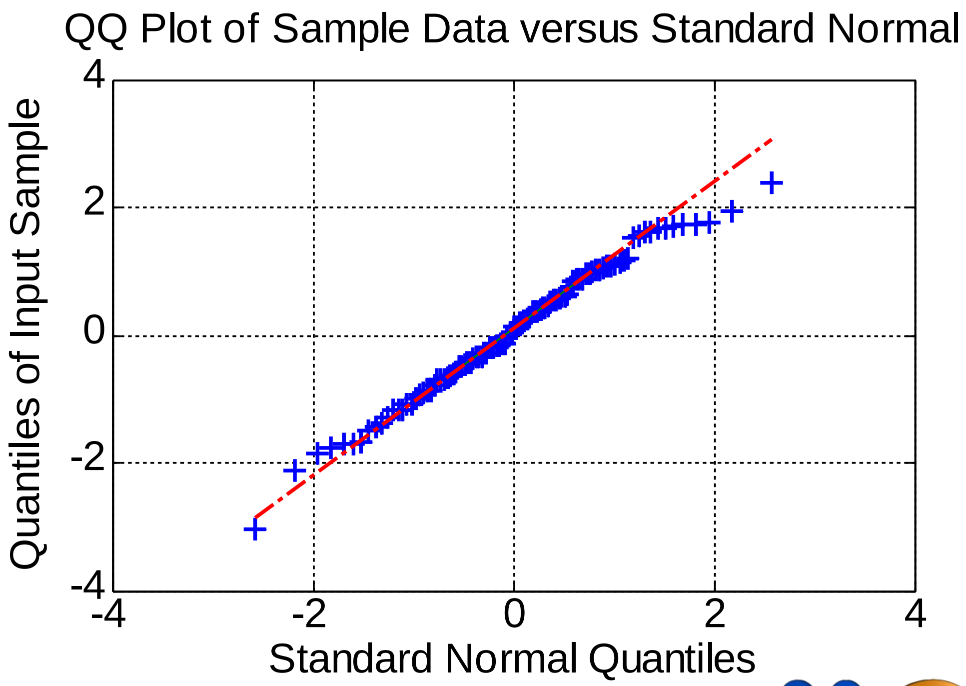
# Comparación de Distribuciones

- En realidad, su uso más corriente es para:
  - Comparar dos muestras a ver si provienen de la misma distribución
    - Sean $x_{(1)},\ldots,x_{(n)}$ los valores de $x_1,\ldots,x_n$ ordenados de menor a mayor
    - Para comparar dos muestras se grafica $(x_{(i)}, y_{(i)})$ para $i=1,\ldots,n$
    - La idea es que $x_{(i)}, y_{(i)}$ son buenas estimaciones del mismo cuantil
    - Ejemplo: dos muestras normales (atención que en este caso hay que ser más “tolerante” con las desviaciones de la recta)



# Comparación de Distribuciones

- En realidad, su uso más corriente es para:
  - Comparar una muestra para ver si proviene de una distribución dada
  - En este caso se grafica $(x_{(i)}, E\{X_{(i)}\})$
  - ¿$E\{X_{(i)}\}$? Una buena aproximación si $F$ (la distribución de $X$) es creciente es $F^{-1}(i/(n+1))$
  - Ejemplo: una muestra normal



# Agenda

- Histogramas y distribuciones empíricas
- (Media,Varianza) y (Mediana,Cuantiles)
- Intervalos de Confianza
- Intervalos de Predicción
- Comentario Finales

# Media

- En vez de mostrar todos los datos o de mostrar toda la distribución, muchas veces es necesario algo más concreto
  - Un valor representativo y su variabilidad alrededor de él
  - Dos candidatos
    - Media y varianza
    - Mediana y cuantiles

# Media y Varianza

- Sean $X_1,\ldots,X_n$ V.A.
  - La media empírica se define:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - La varianza empírica se define (ver en el práctico el porqué de estas dos definiciones y porqué $s_n^2$ es mejor):

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \quad \text{o} \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- IMPORTANTE: notar que estos indicadores son V.A. y por lo tanto son aleatorios

# Media y Varianza

- Ejemplo: la media empírica
  - En el caso que $X_1,\ldots,X_n$ tengan esperanza y sean para todas la misma:

$$E\left\{\overline{X}_n\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = \frac{1}{n}\sum_{i=1}^{n} E\{X_i\} = E\{X_1\}$$

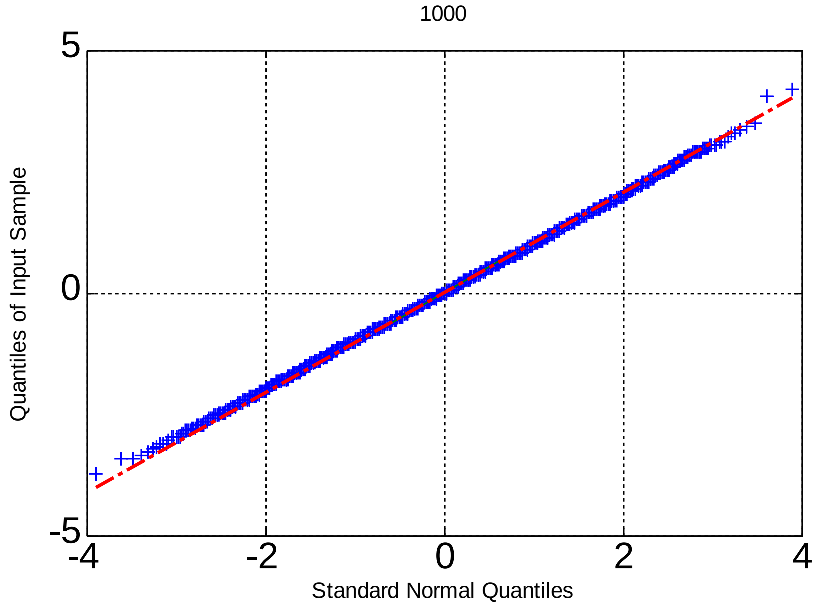
  - Además, si $n$ es suficientemente grande, $X_1, \ldots, X_n$ son iid y tienen varianza $\sigma^2$ finita:

$$\frac{\overline{X}_n - E\{X_1\}}{\sigma/\sqrt{n}} \sim N_{0,1} \Rightarrow \overline{X}_n \sim N_{E\{X_1\},\sigma^2/n}$$

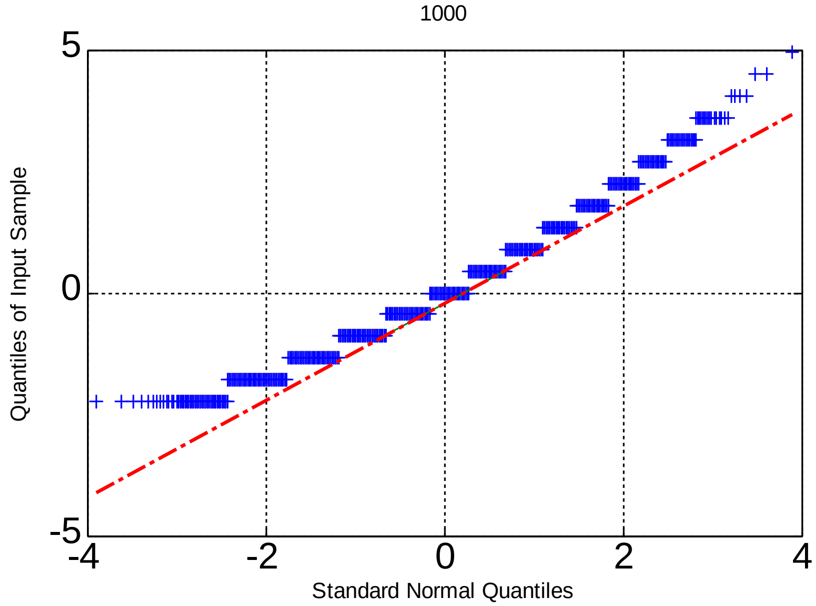
# Media y Varianza

- Al respecto del *n* grande y la aplicación del TCL
  - $X_i \sim \text{Ber}(p)$ ($P(X_i=1)=p$; $P(X_i=0)=1-p$)
  - 1000 muestras para hacer el promedio ($n=1000$)

$p=0.5$




$p=0.005$



# Mediana y Cuantiles

- Sean $X_1,\ldots,X_n$ V.A. y $X_{(1)},\ldots,X_{(n)}$ las mismas V.A. ordenadas de menor a mayor
- El cuantil empírico de órden $q$ se define como

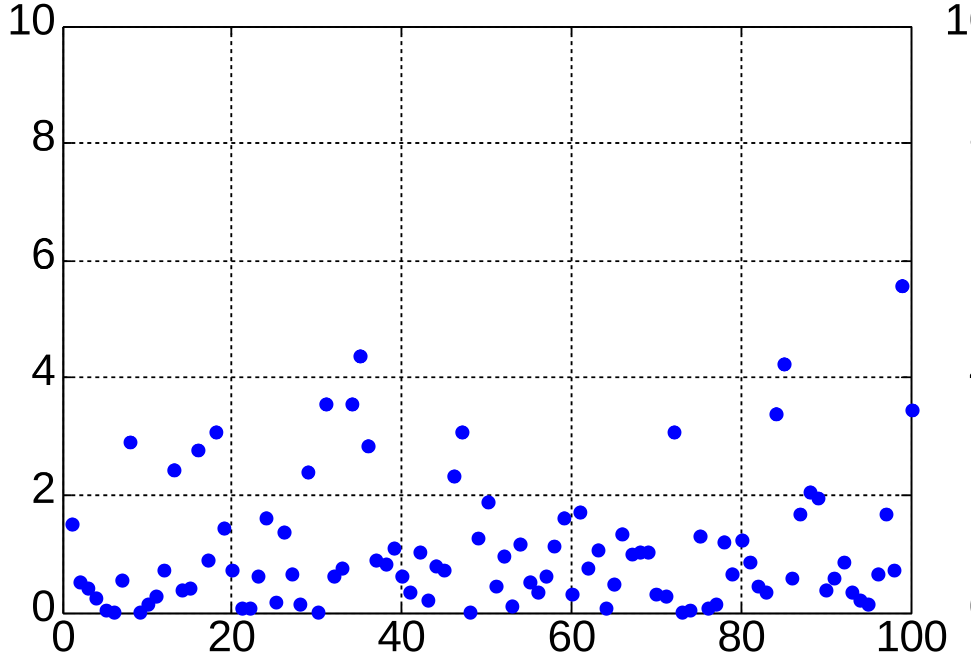
$$\alpha_{n,q} = \frac{X_{(k')} + X_{(k'')}}{2}$$

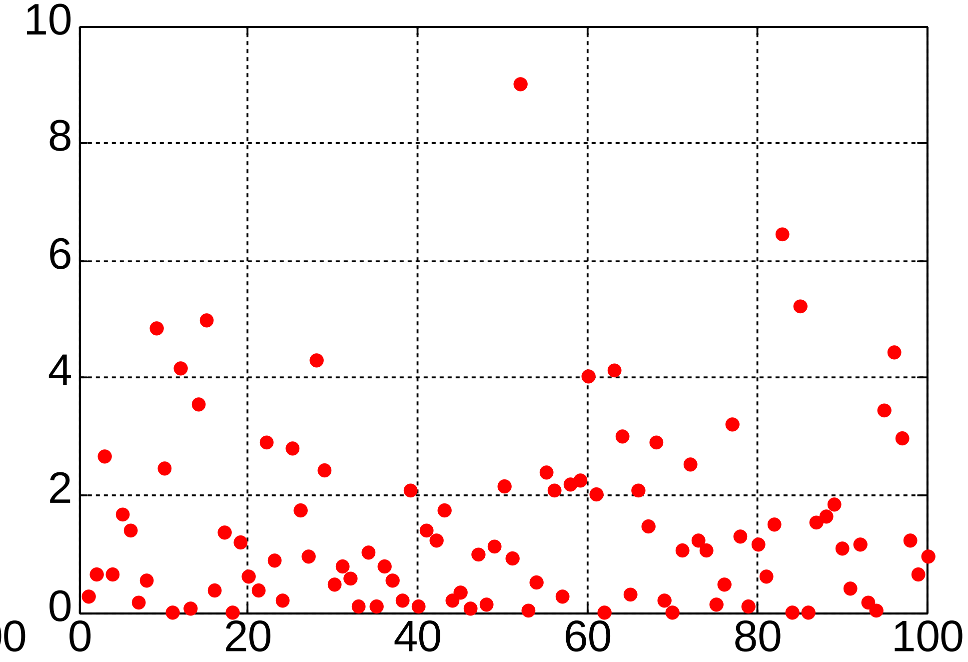
$$\text{donde} \quad k' = \lfloor qn + (1-q) \rfloor \text{ y } k'' = \lceil qn + (1-q) \rceil$$

- Ejemplo: la mediana empírica
  - Si $n$ es par => $\alpha_{n,0.5}=(X_{n/2}+X_{(n+1)/2})/2$
  - Si $n$ es impar => $\alpha_{n,0.5}=X_{(n+1)/2}$

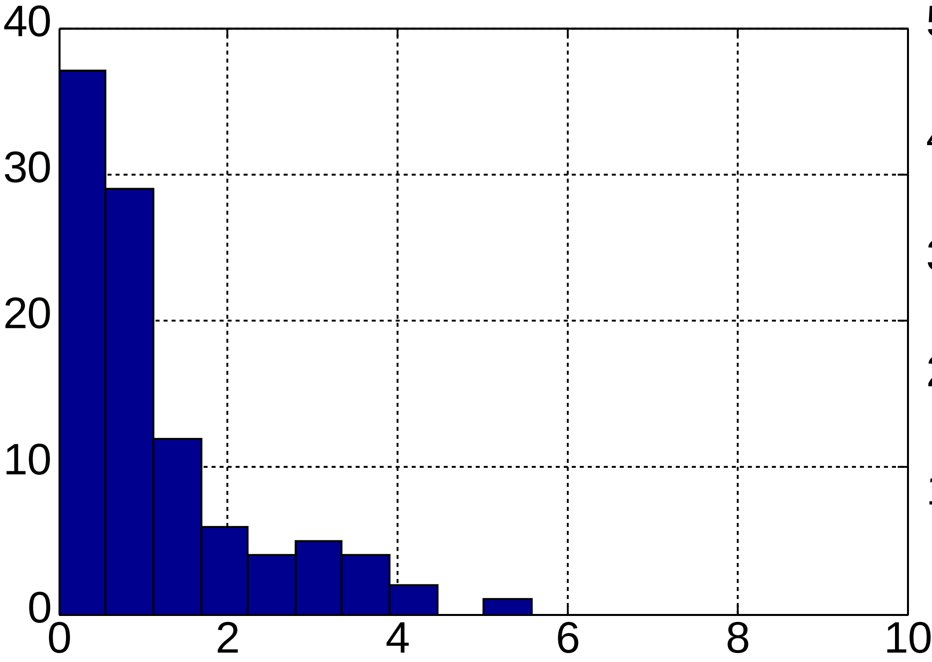
# Ejemplos gráficos

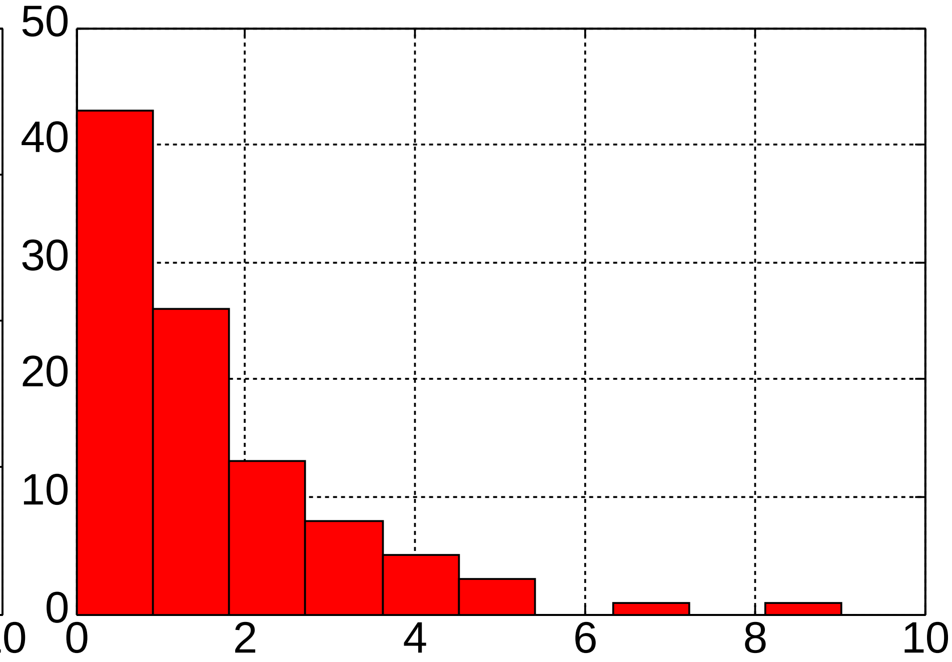
- El mismo benchmarking sobre dos procesadores



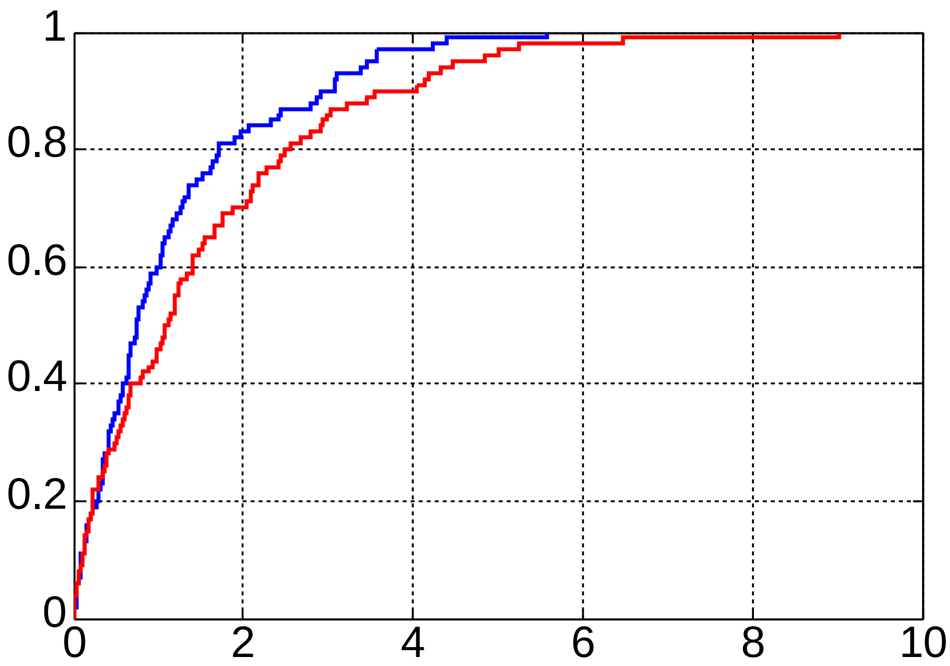



# Histogramas

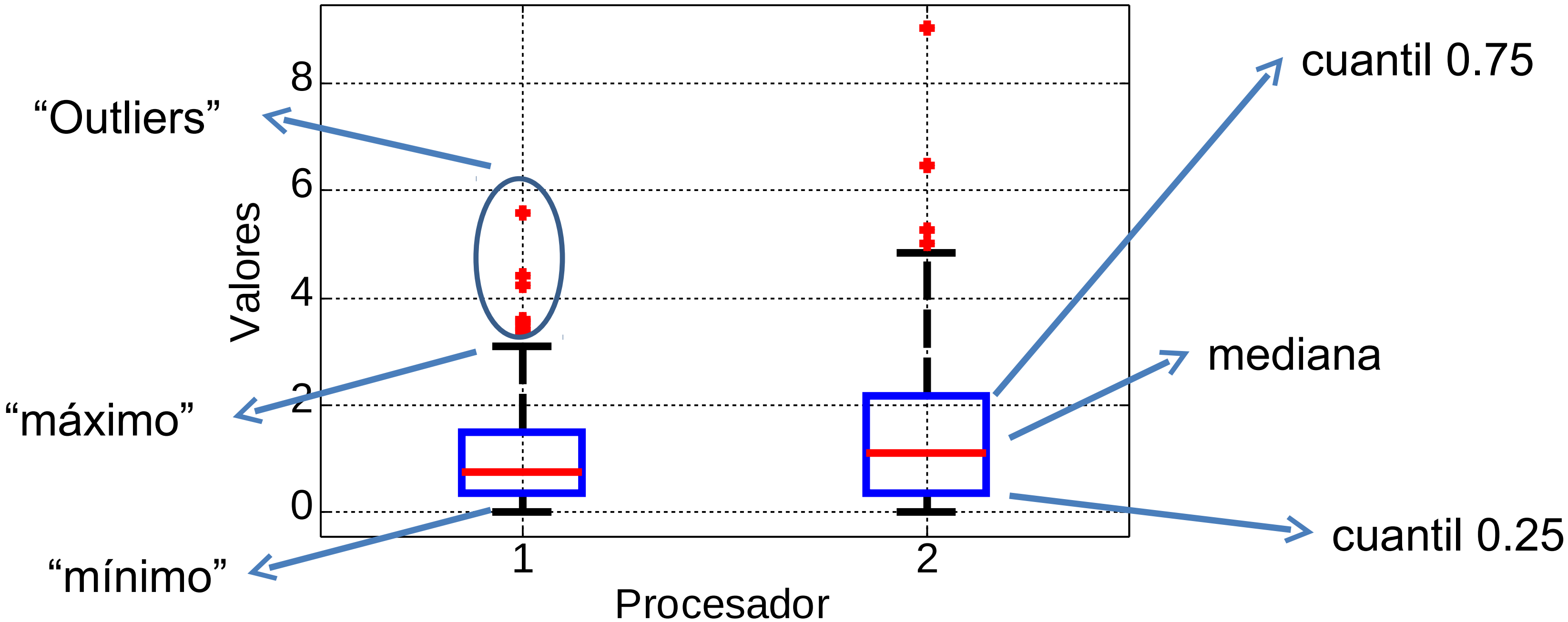


# ECDFs



# Boxplots



"Outliers"
"máximo"
"mínimo"
cuantil 0.75
mediana
cuantil 0.25
8
6
4
2
0
Valores
1
2
Procesador

# Agenda

- Histogramas y distribuciones empíricas
- (Media,Varianza) y (Mediana,Cuantiles)
- Intervalos de Confianza
- Intervalos de Predicción
- Comentario Finales

# Intervalos de Confianza

- Imaginemos que tenemos un montón de datos y nos decidimos a mostrar a modo de resumen la media
- Asumamos que estos *n* números fueron generados por *n* V.A. iid $X_1,\ldots,X_n$
  - En realidad idealmente quisiéramos mostrar $E\{X_1\}$
  - Sabemos que la mejor aproximación es el promedio $\Sigma X_i/n$
  - Pero también sabemos que el promedio es otra V.A.
  - Dos preguntas:
    - ¿Qué tan cerca está el promedio de la esperanza?
    - ¿A partir de $X_1,\ldots,X_n$ podemos calcular un intervalo en el cual la esperanza se encuentre con alta probabilidad?
      - ATENCIÓN: A partir de $X_1,\ldots,X_n$ solamente (nada sabemos por ejemplo de $F()$)

# Intervalos de Confianza

- Más en general
  - Sea *m* un parámetro de la distribución *F*() cualquiera (la esperanza, la varianza, etc.)
    - *F*() se asume fijo pero desconocido (por lo tanto también *m*)
  - Sean $X_1,\ldots,X_n$ *n* V.A. iid con distribución *F*()
  - Un intervalo de confianza de nivel $\gamma$ para el parámetro *m* es un intervalo de la forma $(u(X_1,\ldots,X_n),v(X_1,\ldots,X_n))$ tal que:

$$P\left(u(X_1,\ldots,X_n) < m < v(X_1,\ldots,X_n)\right) \geq \gamma$$

- REPITO: m es fijo, lo aleatorio es el intervalo
  - Imaginen que sé calcular un intervalo de confianza a nivel 0.95 para la esperanza
  - Genero 1000 realizaciones de las *n* V.A. $X_1,\ldots,X_n$ (con $X_1$~U[0,1]) y a cada una le calculo el intervalo
  - ¿Cuántos de esos intervalos contendrán a 0.5?

# Intervalos de Confianza

- Un primer caso "sencillo": Intervalo de confianza para la esperanza
  - Sean $X_1,\dots,X_n$ $n$ V.A. iid con distribución $F()$ tal que existen la esperanza y la desviación estándar $\sigma$
  - Sabemos que si $n$ es suficientemente grande (ver práctico):

$$\left.\begin{aligned} & s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \underset{n}{\rightarrow} \sigma^2 \quad \text{c.s.} \\ & \frac{\overline{X}_n - E\{X_1\}}{\sigma/\sqrt{n}} \sim N_{0,1} \end{aligned}\right\} \Rightarrow \frac{\overline{X}_n - E\{X_1\}}{s_n/\sqrt{n}} \sim N_{0,1}$$

- Entonces UN POSIBLE intervalo de confianza de nivel $\gamma$ para la media es:

$$\left(\overline{X}_n - z_{\frac{\gamma+1}{2}}\frac{s_n}{\sqrt{n}}, \overline{X}_n + z_{\frac{\gamma+1}{2}}\frac{s_n}{\sqrt{n}}\right) \quad \text{con } F_{N_{0,1}}\left(z_{\frac{\gamma+1}{2}}\right) = \frac{\gamma+1}{2}$$

# Intervalos de Confianza

- Un segundo caso no tan “sencillo”: Intervalo de confianza para la varianza
  - Sean $X_1,\ldots,X_n$ *n* V.A. iid con distribución $N_{0,1}$
  - Se puede probar que:

$$\frac{\overline{X}_n - E\{X_1\}}{s_n/\sqrt{n}} \sim t_{n-1} \qquad (n-1)\frac{s_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

  - $\chi^2_n$ se llama distribución chi-cuadrado con *n* grados de libertad
    - Es la distribución de la V.A. definida como la suma de *n* V.A. iid normales elevadas al cuadrado
  - $t_n$ se llama distribución de t-student
    - Es la distribución de la V.A. definida como *X*/sqrt(*Y*/*n*) con *X* normal, *Y* chi-cuadrado con *n* grados de libertad y ambas independientes
- Entonces UN POSIBLE I. de C. de nivel $\gamma$ para la varianza es:

$$\left(s_n\sqrt{\frac{n-1}{\xi}}, s_n\sqrt{\frac{n-1}{\zeta}}\right) \quad \text{con } F_{\chi^2_{n-1}}(\xi) = \frac{1+\gamma}{2} \text{ y } F_{\chi^2_{n-1}}(\zeta) = \frac{1-\gamma}{2}$$

# Intervalos de Confianza

- ¿Cómo saber si la distribución $F()$ tiene esperanza y/o varianza si no la conozco a priori?
- Siempre que sea posible, es mejor utilizar cuantiles
  - ¿Cómo calcular intervalos de confianza para los cuantiles?
- Sean $X_1,\ldots,X_n$ $n$ V.A. iid con distribución $F()$ tal que $F()$ tiene densidad y $X_{(1)},\ldots,X_{(n)}$ los estadísticos de orden
  - Sea $\alpha_q$ el cuantil de orden $q \Rightarrow F(\alpha_q)=q$
  - Un intervalo de confianza de nivel $\gamma$ para $\alpha_q$ es:

$$(X_{(j)},X_{(k)})$$

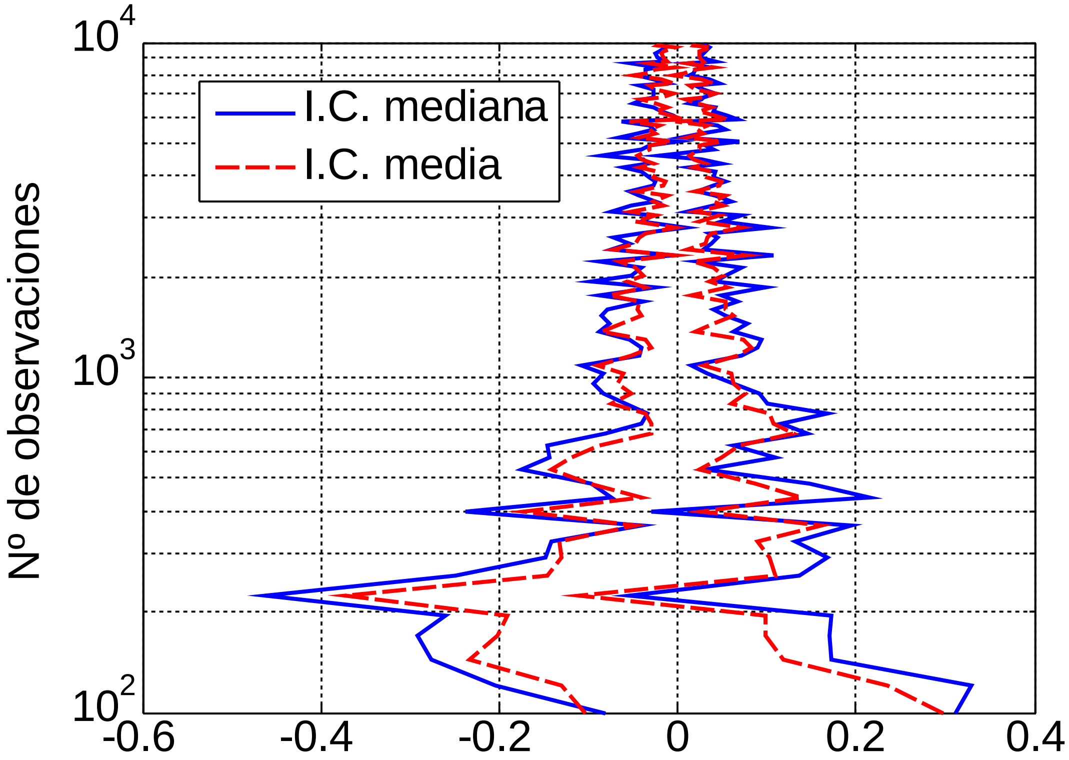
Donde j y k satisfacen

$$F_{\mathrm{Bin}_{n,q}}(k-1) - F_{\mathrm{Bin}_{n,q}}(j-1) \geq \gamma$$

  - $X \sim \mathrm{Bin}_{n,q}$ se denomina binomial de parámetros $n$ y $q$, definida como $X=\Sigma Y_i$ ($i$=1, …,$n$) con $P(Y_i=1)=q$ y $P(Y_i=0)=1-q$ ($Y_i$ independientes)
- Ejemplo: I. de C. de nivel 0.95 para la mediana con n=10 es $(X_{(2)},X_{(9)})$

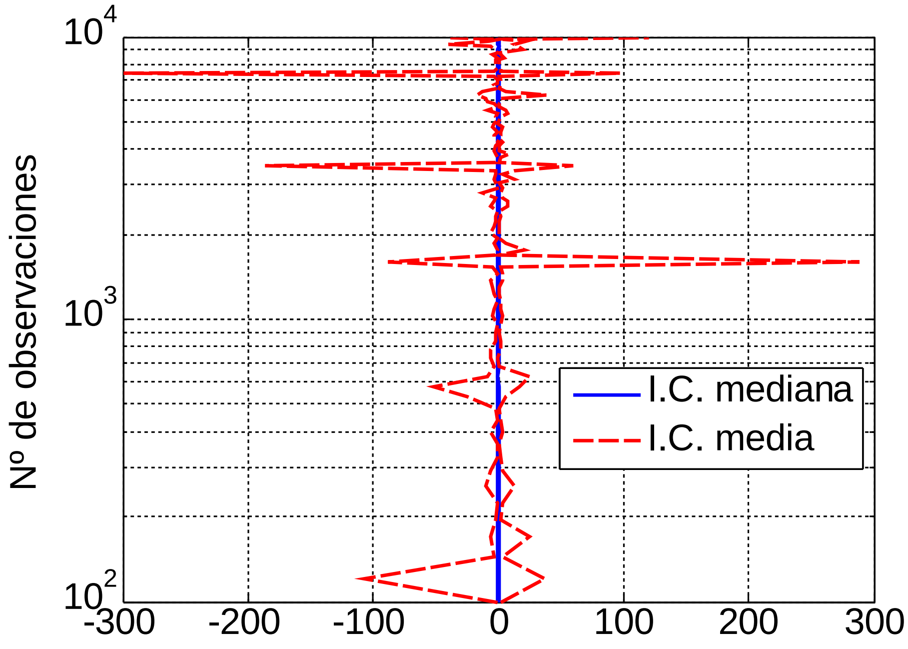
# Intervalos de Confianza

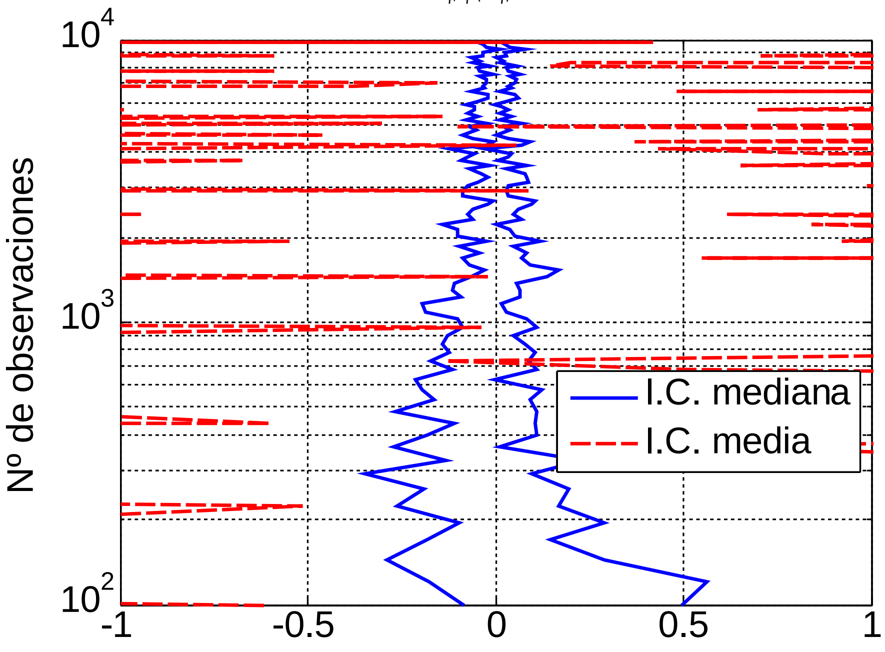
- Ejemplo: $X_1,\ldots,X_n$ $n$ V.A. iid con distribución $N_{0,1}$
  - Dos intervalos de confianza a nivel 0.95
    - Para la media
    - Para la mediana (cuando $nq$ y $n(1-q)$ son grandes se puede aproximar la distribución binomial por una normal $N_{nq,np(1-q)}$ )



# Intervalos de Confianza

- Ejemplo: $X_1,\ldots,X_n$ *n* V.A. iid con distribución $\text{Cauchy}_{0,1}$
  - Dos intervalos de confianza a nivel 0.95
    - Para la media (creyéndome equivocadamente que los datos son normales)
    - Para la mediana (cuando $nq$ y $n(1-q)$ son grandes se puede aproximar la distribución binomial por una normal $N_{nq,np(1-q)}$ )



# Intervalos de Confianza

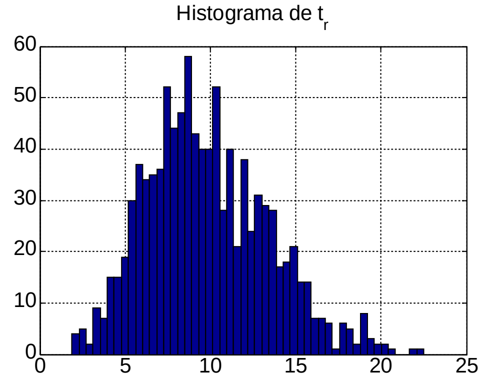
- Sean $X_1,\ldots,X_n$ $n$ V.A. iid y $x_1,\ldots,x_n$ una realización
- Asumamos que queremos calcular un I. de C. de un estadístico $t(X_1,\ldots,X_n)$ pero cuya distribución no sabemos calcular
  - Posible idea: estimar la distribución de $t(X_1,\ldots,X_n)$ a través de su empírica
    - Problema: Tenemos solo una muestra de $t(X_1,\ldots,X_n)$ ($t(x_1,\ldots,x_n)$)
  - Posible idea 2: si tuviéramos R muestras de las V.A. $X_1^r,\ldots,X_n^r$ ($r$=1,…,$R$) tales que se distribuyen como $X_1,\ldots,X_n$ podríamos generar R muestras de $t(X_1,\ldots,X_n)$ evaluando $t(X_1^r,\ldots,X_n^r)$
    - Problema:¿cómo generar las R muestras $X_1^r,\ldots,X_n^r$?
    - Posible respuesta: resampling!
  - Resampling:
    - Una muestra de $X_1^r,\ldots,X_n^r$ se puede generar sorteando **con repetición** n valores de $x_1,\ldots,x_n$
- Lo que acabamos de describir es básicamente el método de **bootstrap**

# Intervalos de Confianza

- Ejemplo: tengo diez muestras de la duración en meses de un cierto equipo
  - x = [4 3 0 14 10 2 39 0 9 16]
  - Asumo que responden al modelo que venimos trabajando hasta ahora (i.e. son la realización de $n$ V.A. iid $X_1,\ldots,X_n$)
  - Quisiera un intervalo de confianza para la duración promedio a nivel 0.95
    - Los datos no son normales y además no puedo aplicar la aproximación normal porque no tengo suficientes muestras
    - No me interesa la mediana, quiero la media
  - Genero $R$=999 muestras con resampling (e.g. $x_r$=[0 4 0 4 9 39 0 0 9 10]), lo que me genera $R$=999 muestras del promedio (e.g. $t_r$=7.5)
  - Una estimación de un I. de C. a nivel $\gamma$ es:

$$(t_{((1-\gamma)R/2)}\ ,\ t_{(\gamma R/2)})$$

i.e. un intervalo conteniendo $\gamma$R de las muestras de $t_r$ (en este caso (3.8,18))



# Intervalos de Confianza

- Algoritmo de bootstrap para intervalos de confianza de un estadístico $t(X_1,\ldots,X_n)$ con $X_1,\ldots,X_n$ $n$ V.A. iid y $x_1,\ldots,x_n$ una realización
  - Parámetros: $r_0$ (e.g. $r_0$=25) y $\gamma$

1: $R = \lceil 2r_0/(1-\gamma) \rceil - 1$
2: **for** $r = 1, \ldots, R$ **do**
3: Sean $(x_1^r, \ldots, x_n^r)$ el resultado de sortear $n$ valores **con reposición** de $(x_1, \ldots, x_n)$
4: Sea $t^r = t(x_1^r, \ldots, x_n^r)$
5: **end for**
6: $(t^{(1)}, \ldots, t^{(R)}) = \mathrm{sort}(t^1, \ldots, t^R)$
7: El intervalo de confianza es $\left[t^{(r_0)}, t^{(R+1-r_0)}\right]$

# Agenda

- Histogramas y distribuciones empíricas
- (Media,Varianza) y (Mediana,Cuantiles)
- Intervalos de Confianza
- Intervalos de Predicción
- Comentario Finales

# Intervalos de Predicción

- Quizá no se hayan dado cuenta, pero lo que estimamos con bootstrap no fue un intervalo de confianza
  - Calculamos un intervalo para el cual:

$$P\left(u(X_1,\ldots,X_n) < t(X_1,\ldots,X_n) < v(X_1,\ldots,X_n)\right) \geq \gamma$$

- Un intervalo de confianza hubiera sido por ejemplo un intervalo tal que:

$$P\left(u(X_1,\ldots,X_n) < E\left\{t(X_1,\ldots,X_n)\right\} < v(X_1,\ldots,X_n)\right) \geq \gamma$$

- Lo primero se denomina **intervalo de predicción**
  - Es una medida de la **variabilidad** de $t(X_1,\ldots,X_n)$
  - En cambio, un intervalo de confianza es una medida de la **precisión** en mi estimación de $E\{t(X_1,\ldots,X_n)\}$

# Intervalos de Predicción

- **Definición:**
  - Sean $X_1,\ldots,X_n,X_{n+1}$ $n+1$ V.A. (no necesariamente iid)
  - Un intervalo de predicción de nivel g para $X_{n+1}$ es un intervalo de la forma $(u(X_1,\ldots,X_n),v(X_1,\ldots,X_n))$ tal que:

$$P\left(u(X_1,\ldots,X_n) < X_{n+1} < v(X_1,\ldots,X_n)\right) \geq \gamma$$

- **Teorema (caso iid):**
  - Sean $X_1,\ldots,X_n,X_{n+1}$ $n+1$ V.A. iid con distribución $F()$ tal que $F()$ tiene densidad
  - Sean $X_{(1)},\ldots,X_{(n)}$ los estadísticos de orden de $X_1,\ldots,X_n$
  - Entonces para $1 \leq j \leq k \leq n$:

$$P\left(X_{(j)} \leq X_{n+1} \leq X_{(k)}\right) = \frac{k-j}{n+1}$$

  - Por lo tanto para $\alpha \geq 2/(n+1)$ el intervalo $[X_{((n+1)\alpha/2)},X_{((n+1)(1-\alpha/2))}]$ es un intervalo de predicción de nivel al menos $\gamma=1-\alpha$

# Intervalos de predicción

- Hay resultados más poderosos que el anterior
  - Por ejemplo si puedo asumir que los datos son iid y además normales
- De todas formas es importante quedarse con el significado del intervalo de predicción
  - Si realizo muchas realizaciones de $X_{n+1}$, ésta estará al menos $\gamma$ veces dentro del intervalo de predicción
- Caso iid
  - Las $X_i$ me ayudan a predecir $X_{n+1}$
  - ¿Contradicción?

# Agenda

- Histogramas y distribuciones empíricas
- (Media,Varianza) y (Mediana,Cuantiles)
- Intervalos de Confianza
- Intervalos de Predicción
- Comentario Finales

# Comentarios Finales

- Todas las técnicas mencionadas necesitan "iid-ismo"
- ¿Cómo hago que los datos sean iid?
  - Si controlo el experimento (como en una simulación) está en mí lograr que los datos sean iid
    - Por ejemplo, si es una simulación tengo que generar los resultados mediante tiradas independientes en el generador aleatorio
    - Si por ejemplo estoy midiendo cuánto demora un servidor ante un conjunto de pedidos, tengo que elegir en cada experimento el pedido al azar con reposición
  - Si NO controlo el experimento (e.g. son medidas de campo) difícilmente los datos sean iid
    - Por lo general aleatorizando las medidas se corrige el problema
    - Ejemplo: medidas del tiempo de atención de pedidos http en un servidor web
      - Probablemente los tiempos de atención estén muy correlacionados en el tiempo
      - Puedo generar un conjunto nuevo tomando cada medida con probabilidad $p$ (si sigue sin ser independientes genero uno nuevo con una $p$ menor que la anterior)

# Comentarios Finales

- ¿Cómo verifico que mis datos hayan sido generadas por V.A. iid?
  - Lag-Plots
    - Una gráfica de $(x_i, x_{i-k})$ para varios valores de $k$
    - Si son independientes no se deberían ver tendencias en las curvas
  - Tests
    - Las próximas clases
- ¿Qué pasa si calculo un intervalo de confianza pensando que fueron generadas por una V.A. iid y no es verdad?
  - Si los datos no son independientes y el calculo asume que sí, el intervalo de confianza resultante va a ser más pequeño de lo que un intervalo correcto sería

# Comentarios Finales

- Hipótesis para el intervalos de confianza de la media
  - ¿Cómo saber si mi *n* es suficientemente grande?
  - ¿Cómo saber si *F*() tiene esperanza y varianza finitas?
  - Lo que se requiere en realidad es que

$$\frac{\overline{X}_n - E\{X_1\}}{\sigma/\sqrt{n}} \sim N_{0,1}$$

  - Tengo solo una muestra del promedio
    - ¿Cómo obtengo su distribución?
    - Bootstrap!
    - Verifico que tenga distribución normal (sin importar la media o la varianza)
- Ejemplo de la duración de los productos
  ($x$ = [4 3 0 14 10 2 39 0 9 16])