Unidad 3: “Introducción a la probabilidad y estadística aplicada”

Cátedra de Métricas del Software

Universidad Tecnológica Nacional
Facultad Regional Córdoba

- Diego Rubio
- Álvaro Ruiz de Mendarozqueta
- Natalia Andriano
- Juan Pablo Bruno

Objetivos Específicos

- Comprender los **conceptos básicos** de probabilidad y estadística aplicada para el **análisis de métricas** asociadas a la administración del **proceso de desarrollo y mejora continua de software**

Preaviso!

- El contenido de esta unidad fue simplificado para cubrir los objetivos y conocimientos necesarios para el correcto entendimiento del resto del curso.

Agenda

- Recolección y Sistematización de Datos
- Tipos y escalas de datos
- Calidad de datos
- Estadística descriptiva
  - Posición
  - Dispersión
  - Forma
- Muestreo aleatorio, estratificado y sistemático
- Probabilidades
  - Distribuciones de probabilidades
  - Teorema central del límite

Recolección y Sistematización de Datos

- Sumamente importante para obtener datos adecuados
  - A veces con el mismo esfuerzo se puede obtener mucha más información si se planifica correctamente
- Técnicas como Checksheets, Stem&Leaf usualmente en la industria
  - Recolección automática vía software es más típico en la industria del software
  - …aunque a veces son muchos Excels…

Tipos de datos

- De acuerdo a excel:
- De acuerdo al lenguaje de programación
  - C, C++, Java, VB, C#, etc…
- Importante para nosotros:
  - Discretos
  - Continuos

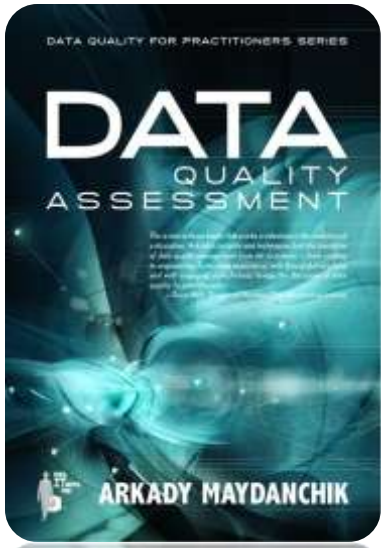
Nos permitirá decidir técnicas a usar, tamaños de muestra, etc

Escala

- Un conjunto de valores con propiedades definidas [ISO 14598-1].
- Tipos de Escala
  - Nominal Ej: Juan, Pedro, Mariano
  - Ordinal (orden pero no diferencia). Ej: Grande, Mediano, Chico
  - Intervalo (sin cero absoluto) Ej: Temperatura en F
  - Radio Ej: Peso
  - Absoluta Ej: Número de personas en el proyecto
- La escala determina el tipo de análisis que se puede hacer sobre el conjunto de valores.

Calidad de datos

- Validar antes de analizar los datos.
- Los análisis pueden variar desde muy sencillos hasta incluso MSA (análisis de sistema de medición)
- Algunos ejemplos básicos:
  - Tipo correcto (e.g. continuo, discreto)
  - Formato correcto (e.g. escala)
  - Dentro de rango especificados
  - Completos
  - Aritméticamente correctos
  - Válidos (nivel básico)



Ejemplo de libros en la temática
http://www.dataqualitygroup.com

Estadística descriptiva

- Posición
  Dispersión
  Forma

Estadística descriptiva

- Descripción sumarizada de una serie de datos
- Los **datos discretos (atributos)** puede ser usualmente sumarizados por cuentas, proporciones o gráficos temporales de estas.
- Los **datos continuos (variables)** pueden ser sumarizados por:
  - Posición
  - Dispersión
  - Forma

- Nota: Las estadísticas descriptivas son números basados en muestras de la población. Son estimaciones puntuales de las características de la distribución de la población subyacente.

Medidas de posición

- **Dada la siguiente serie:**

Observación (x)	Valor (y)
1	3
2	5
3	7
4	158
5	7

- **¿Cuál es el “centro” de estos datos?**

Medidas de posición

- DEPENDE!

Medidas de posición

- Existen varias maneras de medir la posición de una distribución.
- En este curso veremos 2 (dos):
  - Media
    - Promedio de los datos

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

  - Mediana
    - Percentil 50°
    - Mitad de los datos por encima y mitad por debajo

- Nota: otra medida de posición usualmente conocida es la Moda, ¿se acuerdan?

Medidas de posición

- **Dada la siguiente serie:**

Observación (x)	Valor (y)
1	3
2	5
3	7
4	158
5	7

- **¿Cuál es el "centro" de estos datos?**
  - ☑ Media = 36.2
  - ☑ Mediana = 7
  - ☑ Moda = 7

Nota: la media es mucho más sensible a puntos extremos

Medidas de posición

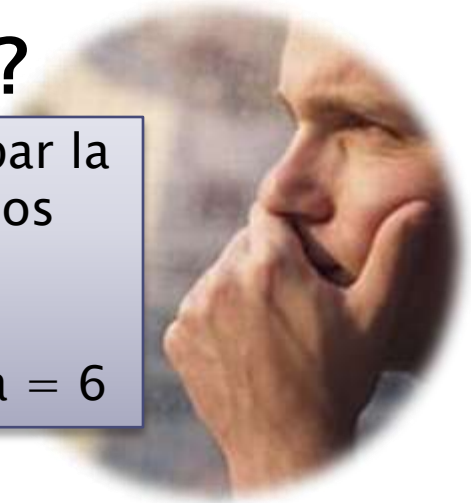
- **Dada esta nueva serie:**

Observación (x)	Valor (y)
1	1
2	5
3	7
4	10
5	8
6	5

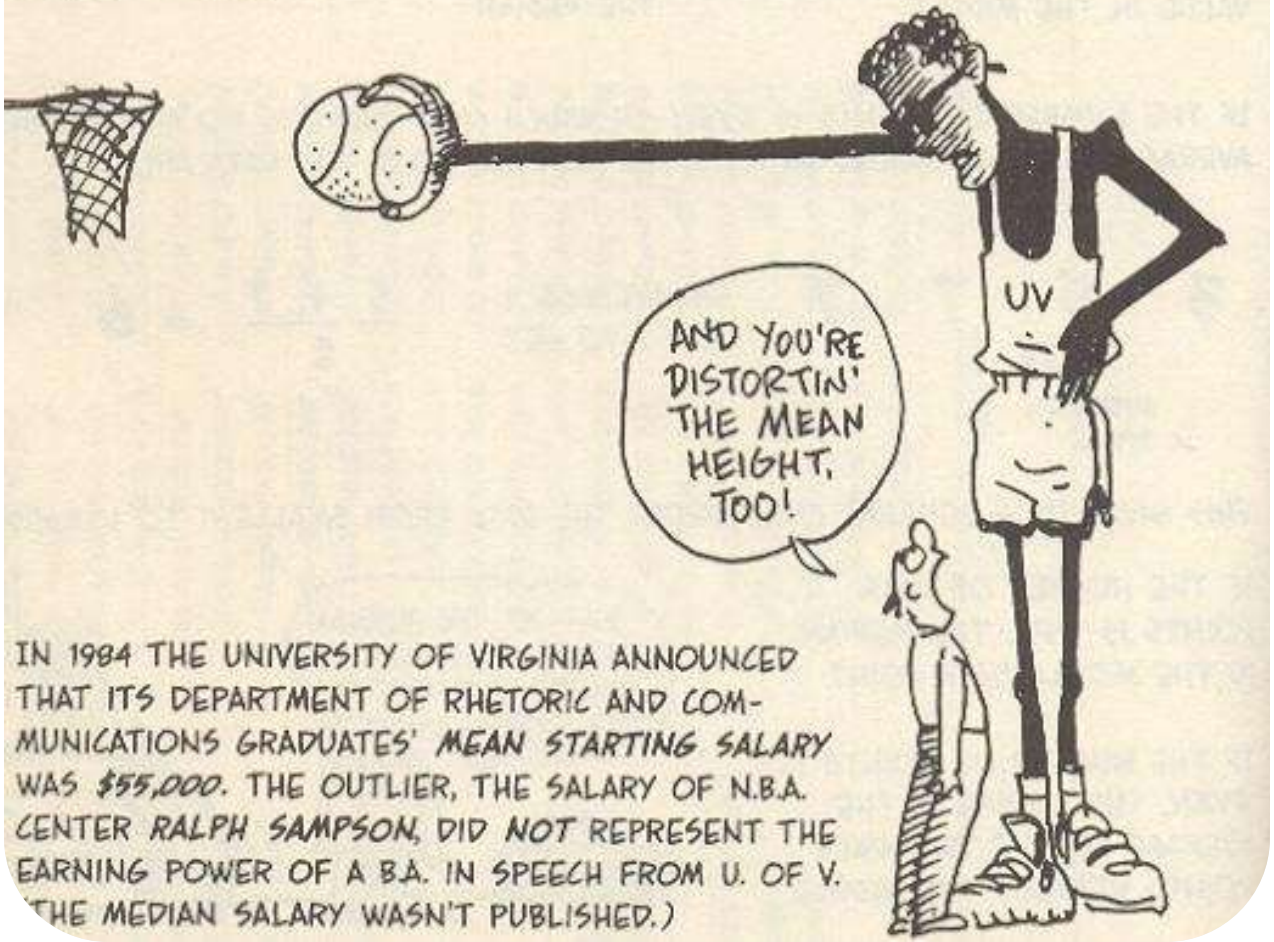
- **¿Cuál es el "centro" de estos datos?**
  - ☑ Media = 6
  - ☑ Mediana = ?

Cuando el número de datos es par la mediana es igual a la media de los dos números centrales

1 – 5 – 5 – 7 – 8 – 10 → Mediana = 6



HY MORE THAN ONE MEASURE OF THE CENTER? EACH HAS ADVANTAGES. FOR EXAMPLE, THE *MEDIAN* IS NOT SENSITIVE TO *OUTLIERS*, OR EXTREME VALUES NOT TYPICAL OF THE REST OF THE DATA. SUPPOSE IN OUR SMALL TV-WATCHING GROUP, ONE PERSON WATCHES *200* HOURS PER WEEK. THEN OUR DATA ARE 3, 5, 7, 7, 200. THE MEDIAN, 7, IS UNCHANGED, BUT THE MEAN IS NOW $\bar{x} = 45.8$!



IN 1984 THE UNIVERSITY OF VIRGINIA ANNOUNCED THAT ITS DEPARTMENT OF RHETORIC AND COMMUNICATIONS GRADUATES' *MEAN STARTING SALARY* WAS *$55,000*. THE OUTLIER, THE SALARY OF N.B.A. CENTER *RALPH SAMPSON*, DID *NOT* REPRESENT THE EARNING POWER OF A B.A. IN SPEECH FROM U. OF V. (THE MEDIAN SALARY WASN'T PUBLISHED.)

Reference: The cartoon Guide to statistics, Larry Gonick, Woollcott Smith,
ISBN: 0-06-273102-5

Medidas de dispersión

- Al igual que para las medidas de posición existen varias
- En este curso veremos:
  - Rango
    - Mide la distancia entre los puntos extremos
    - **R = Max{data} – Min{data}**
  - Varianza
    - Promedio del cuadrado de la distancia de los puntos a la media

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

  - Desviación estándar
    - Raíz cuadrada de la varianza

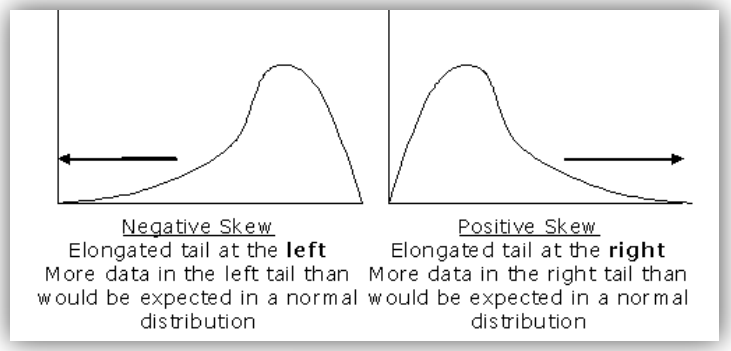
$$s = \sqrt{s^2}$$

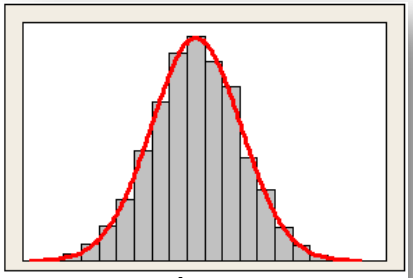
Medidas de dispersión

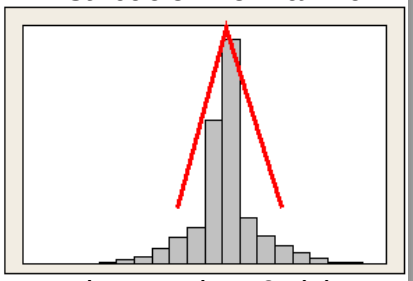
- Algunas consideraciones:
  - La desviación estándar se mide en la misma unidad que la media
  - El rango es mucho más sensible a puntos extremos.
  - En general para tamaños moderados de n, digamos $n > 10$, la desviación estándar es una mejor medida de la dispersión.
    - Hace mejor uso de toda la información disponible!
  - Sin embargo, para muestras pequeñas como las utilizadas frecuentemente en las cartas de control ($n = 4, 5, 6$) el rango suele ser una buena medida.

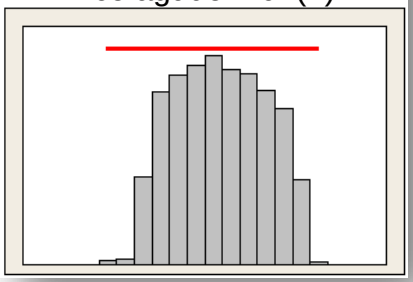
Forma

- Es la distribución aproximadamente simétrica?
  - Skewness $$skewness = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{(N-1)s^3}$$
- Es la distribución plana o con un marcado pico?
  - Kurtosis $$kurtosis = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4}{(N-1)s^4} - 3$$
- ¿Existe alguna distribución subyacente a la que la podamos asociar?
  - ¿Podemos rechazar la normalidad?
- En conclusión: ¿Qué nos dice la forma de los datos?




Distribución normal = 0


Pico agudo > 0 (+)

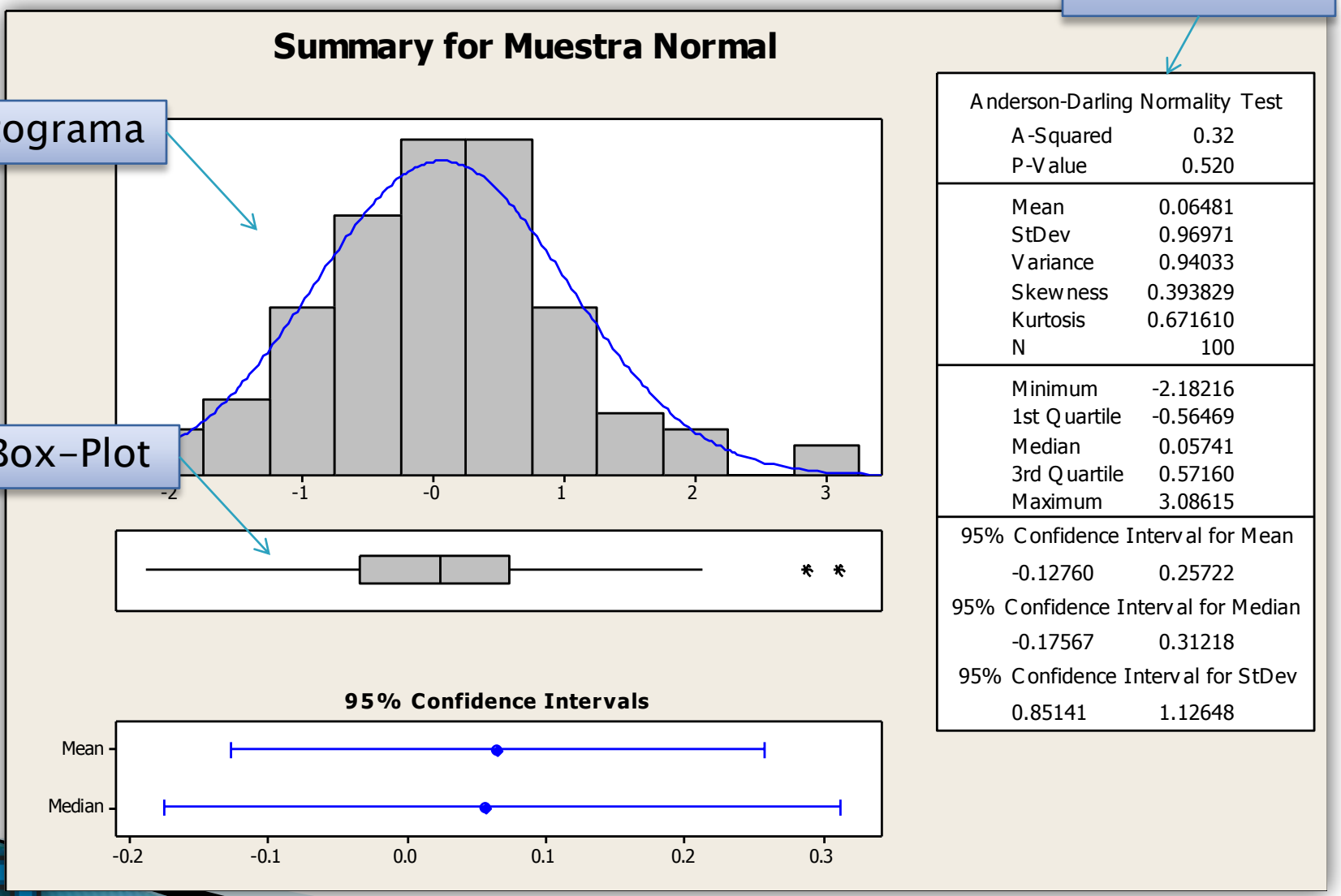

Distribución plana < 0 (-)

Sumarizando..

Estadística básica

Histograma

Box-Plot



Anderson-Darling Normality Test	
A-Squared	0.32
P-Value	0.520
Mean	0.06481
StDev	0.96971
Variance	0.94033
Skewness	0.393829
Kurtosis	0.671610
N	100
Minimum	-2.18216
1st Quartile	-0.56469
Median	0.05741
3rd Quartile	0.57160
Maximum	3.08615
95% Confidence Interval for Mean	
-0.12760	0.25722
95% Confidence Interval for Median	
-0.17567	0.31218
95% Confidence Interval for StDev	
0.85141	1.12648

Muestreo

- *"…People hate to waste time doing unnecessary work, and one thing statistics can do is tell us exactly* ***how lazy we can afford to be****…"*

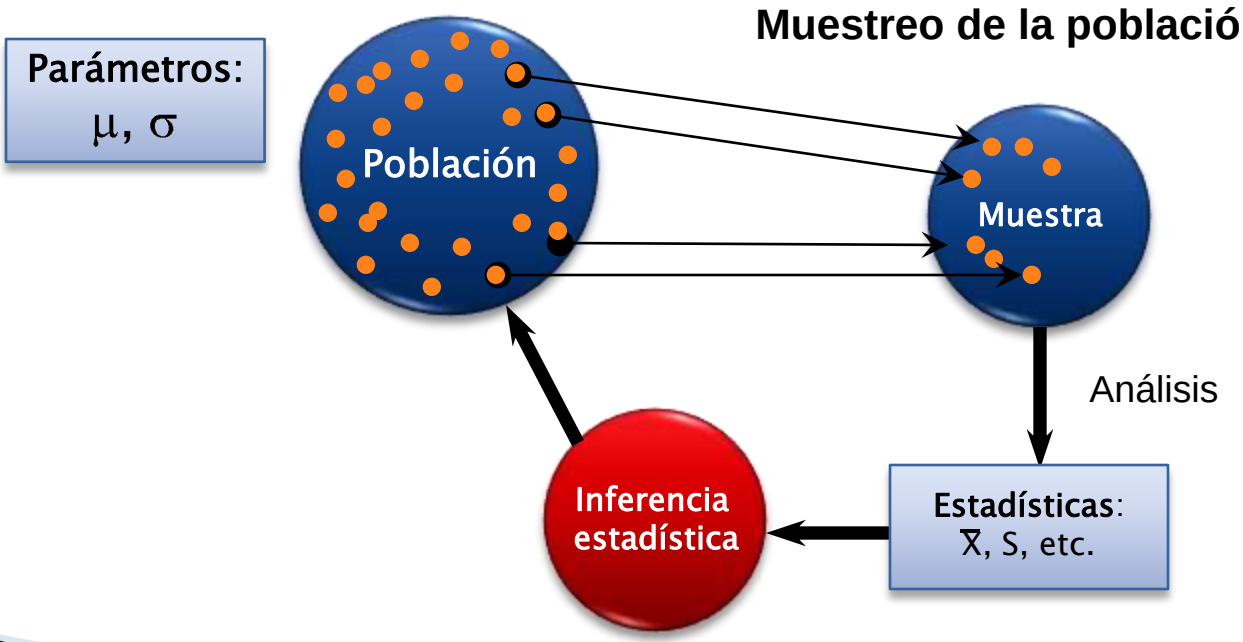
The cartoon Guide to statistics

Muestreo

- Problema:
  - Tamaño de las colecciones en el “mundo real”
  - Costo de recolección de datos
  - Posibilidad de recolección
- Solución:
  - Tomar una parte del todo e inferir sobre él.
  - Tomar decisiones con un costo y riesgo razonable

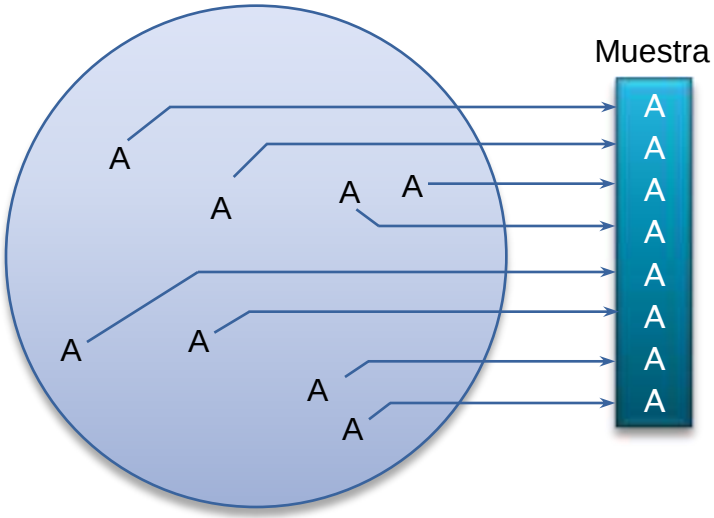
¿Qué es el muestreo?

- Sacar conclusiones de la población basado en la muestra → **Inferencia estadística**
- Decisiones razonables de negocio →**Nivel de confianza**



Tipos básicos de Muestreo

Muestreo aleatorio simple



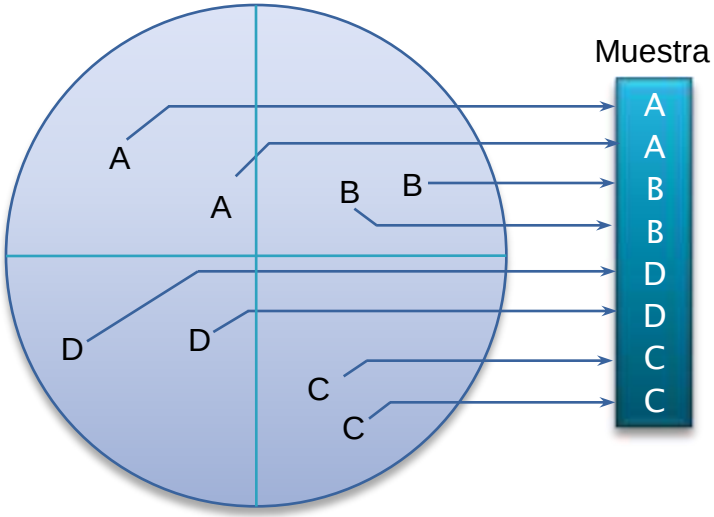
Muestreo aleatorio dentro de la población

Propiedades muy importantes:

- **Independencia:** la selección de una unidad no tiene influencia en la selección de otras.
- **Sin bias**: cada unidad tiene la misma chance de ser elegido.
- El procedimiento asegura que todas las posibles muestras de *n* objetos tomados de la población, tienen la misma probabilidad se ser elegidos

Tipos básicos de Muestreo

Muestreo estratificado simple



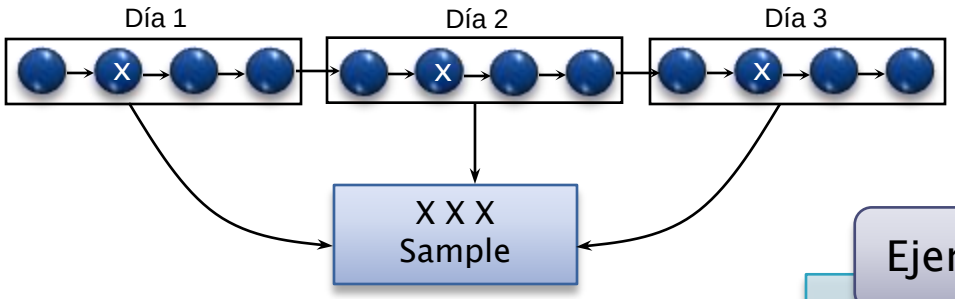
Muestreo aleatorio dentro de cada categoría (e.g. lugar, turno, producto, etc.)

Procedimiento:

- Dividir la población en unidades homogéneas (estratos) y realizar un muestreo aleatorio simple para cada grupo.

Tipos básicos de Muestreo

Muestreo sistemático



Ejemplos:

- Muestreo en un paso determinado del proceso en cada día (hora, semana, mes)
- Empezar con una unidad aleatoria y luego seleccionar cada *K* unidades.

Algunas reglas generales…

- Tamaño de muestra (n) es función de:
  - Riesgo de tomar la decisión equivocada
  - Variabilidad de la población
  - Diferencia a ser detectado (y/o precisión)
  - En general:
    - A menor riesgo (↓), mayor tamaño de muestra(↑).
    - A mayor variación en la población(↑), mayor tamaño de muestra(↑).
    - A menor diferencia a detectar(↓), menor tamaño de muestra(↓).
- Algunos factores a tener en cuenta:
  - Costo.
  - Facilidad.
  - Representatividad.
  - Variación y estabilidad de la población.
- Tanto tomar de más como de menos se traducen en pérdidas:
  - Por lo general se comienza tomando de más y luego se reduce

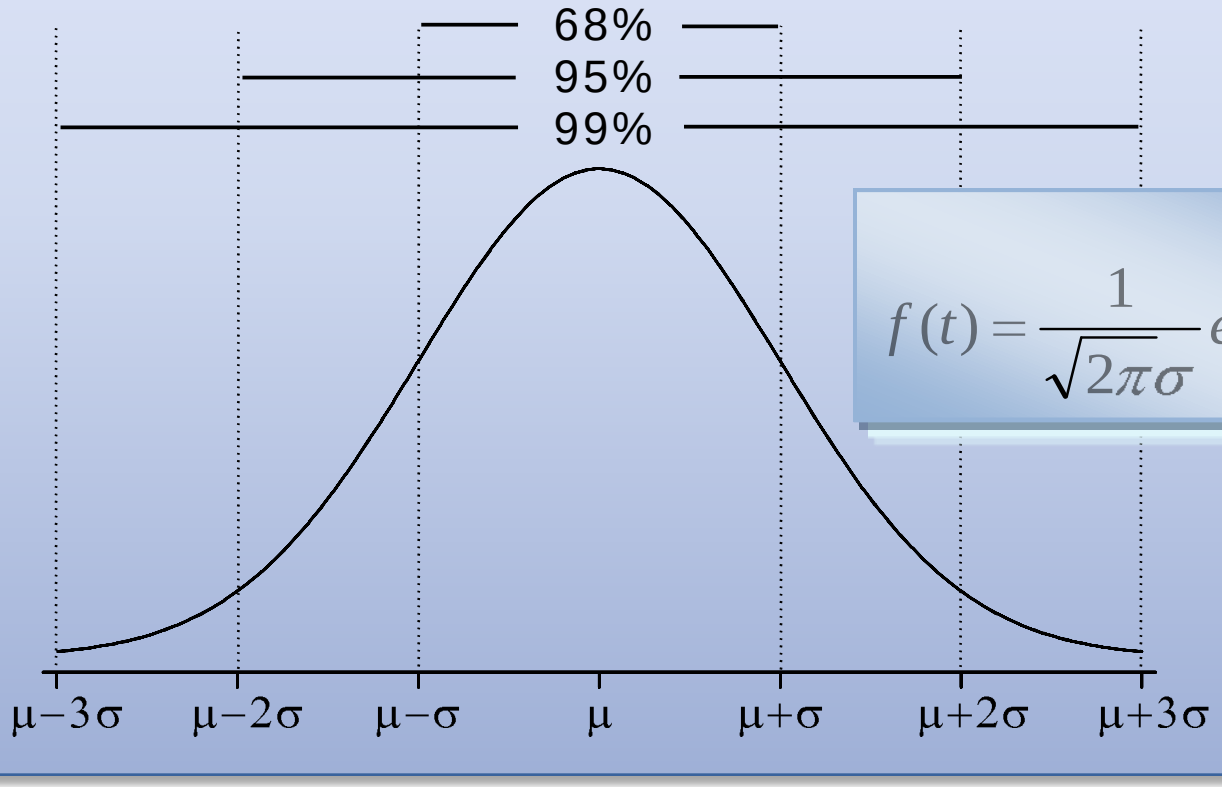
Aleatoriedad!!

"…Sin diseños aleatorios, es imposible hacer análisis estadísticos que requieren independencia, sin importar como se modifiquen los datos.
La belleza de la aleatoriedad es que "garantiza estadísticamente" la exactitud del muestreo…"

Distribuciones de probabilidades

Distribución Normal

Probabilidades útiles de la distribución normal (aprox.)



$$f(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{(t-\mu)^2}{2\sigma^2}\right)}$$

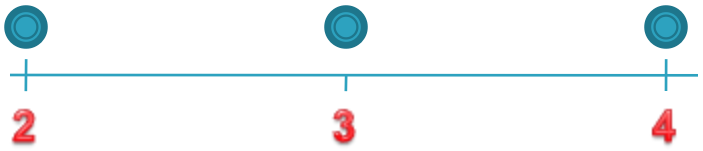
Distribución Normal

- Muchos de los análisis que se realizan habitualmente asumen cierta normalidad en los datos…
- Pero la distribución normal es conocida por ser la menos normal de las distribuciones…al menos en software… ☺
- ¿Que hacer?
  - Transformar (e.g. Box Cox)
  - Usar otra distribución (e.g. exponencial, Weibull, etc)
  - Aprovechar el teorema central del límite

Teorema central del límite

- Sin importar la forma de la población, la distribución de las medias muestreales es aproximadamente normal si el tamaño de la muestra es lo suficientemente grande.

- Ejercicio en clase:
  - Población:



Conceptos Claves (1 /2)

- Recolección y Sistematización de Datos
- Tipos y escalas de datos
  - Continuo vs discreto
  - NOIR → nominal, ordinal, intervalo, radio (absoluto)
- Calidad de datos
  - Tipo, escala, Dentro de rango, Completos, Aritméticamente correctos, Válidos (nivel básico)
- Estadística descriptiva
  - Posición
    - Media, Mediana (moda)
  - Dispersión
    - Rango, desviación estándar (varianza)
  - Forma

Conceptos Claves (2/2)

- Muestreo:
  - Tipos:
    - Aleatorio simple
    - Estratificado
    - Sistemático
  - Propiedades importantes:
    - Independencia
    - Falta de Bias
  - ALEATORIEDAD
- Probabilidades
  - Distribuciones de probabilidades
    - Normal (1 σ → 68%, 2 σ →95%, 3 σ → 99%)
  - Teorema central del límite
    - "…la distribución de las medias muestreales es aproximadamente normal si el tamaño de la muestra es lo suficientemente grande…"

Lecturas Obligatorias

Autor	Título	Editor	Referencia

Lecturas Recomendadas

Autor	Título	Editor	Referencia
Larry Gonick, Woollcott Smith	The cartoon Guide to statistics	HarperReour se. 1993.	0-06-273102-5

Bibliografía

Autor	Título	Editor	Referencia
Larry Gonick, Woollcott Smith	The cartoon Guide to statistics	HarperReourse. 1993.	0-06-273102-5

Versión

Versión	Fecha	Descripción	Autor
1.0.0_Draft_A	Feb-2008	Primera versión adaptada de Material previo.	Diego Rubio
1.0.0	Apr-2008	A línea base	Diego Rubio