CEU
Universidad
San Pablo

Practical Statistical Questions

Session 1

Carlos Óscar Sánchez Sorzano, Ph.D.

Madrid, July 7th 2008

Course outline

1. I would like to know the intuitive definition and use of …: The basics
   1. Descriptive vs inferential statistics
   2. Statistic vs parameter. What is a sampling distribution?
   3. Types of variables
   4. Parametric vs non-parametric statistics
   5. What to measure? Central tendency, differences, variability, skewness and kurtosis, association
   6. Use and abuse of the normal distribution
   7. Is my data really independent?

1.1 Descriptive vs Inferential Statistics

Statistics (="state arithmetic")

Descriptive: describe data

- How rich are our citizens on average? → Central Tendency
- Are there many differences between rich and poor? → Variability
- Are more intelligent people richer? → Association
- How many people earn this money? → Probability distribution
- Tools: tables (all kinds of summaries), graphs (all kind of plots), distributions (joint, conditional, marginal, …), statistics (mean, variance, correlation coefficient, histogram, …)

Inferential: derive conclusions and make predictions

- Is my country so rich as my neighbors? → Inference
- To measure richness, do I have to consider EVERYONE? → Sampling
- If I don't consider everyone, how reliable is my estimate? → Confidence
- Is our economy in recession? → Prediction
- What will be the impact of an expensive oil? → Modelling
- Tools: Hypothesis testing, Confidence intervals, Parameter estimation, Experiment design, Sampling, Time models, Statistical models (ANOVA, Generalized Linear Models, …)

1.1 Descriptive vs Inferential Statistics

Of 350 randomly selected people in the town of Luserna, Italy, 280 people had the last name Nicolussi.

Which of the following sentences is descriptive and which is inferential:

1. 80% of THESE people of Luserna has Nicolussi as last name.
2. 80% of THE people of ITALY has Nicolussi as last name.

On the last 3 Sundays, Henry D. Carsalesman sold 2, 1, and 0 new cars respectively.

Which of the following sentences is descriptive and which is inferential:

1. Henry averaged 1 new car sold of the last 3 sundays.
2. Henry never sells more than 2 cars on a Sunday

What is the problem with the following sentence:

3. Henry sold no car last Sunday because he fell asleep inside one of the cars.

Source: http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Data_Descr_Infer.htm

1.1 Descriptive vs Inferential Statistics

The last four semesters an instructor taught Intermediate Algebra, the following numbers of people passed the class: 17, 19, 4, 20

Which of the following conclusions can be obtained from purely descriptive measures and which can be obtained by inferential methods?

a) The last four semesters the instructor taught Intermediate Algebra, an average of 15 people passed the classs
b) The next time the instructor teaches Intermediate Algebra, we can expect approximately 15 people to pass the class.
c) This instructor will never pass more than 20 people in an Intermediate Algebra class.
d) The last four semesters the instructor taught Intermediate Algebra, no more than 20 people passed the class.
e) Only 5 people passed one semester because the instructor was in a bad mood the entire semester.
f) The instructor passed 20 people the last time he taught the class to keep the administration off of his back for poor results.
g) The instructor passes so few people in his Intermediate Algebra classes because he doesn't like teaching that class.

Source: http://infinity.cos.edu/faculty/woodbury/Stats/Tutorial/Data_Descr_Infer.htm

1.2 Statistic vs. Parameter. What is a sampling distribution?

Statistic: characteristic of a sample

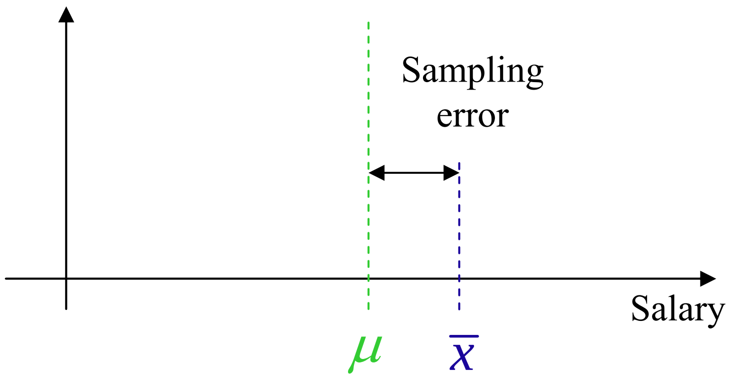
What is the average salary of 2000 people randomly sampled in Spain?

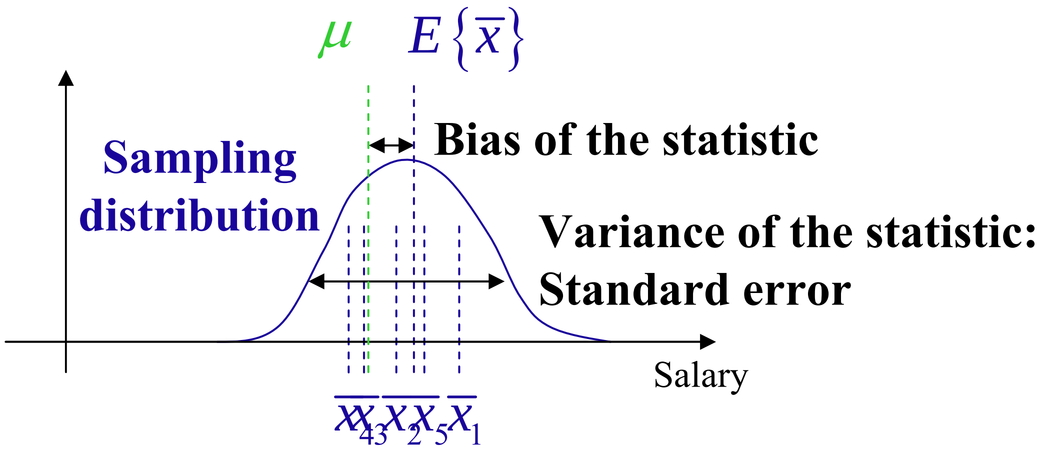
$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Parameter: characteristic of a population

What is the average salary of all Spaniards?

μ




μ $E\left\{\overline{x}\right\}$

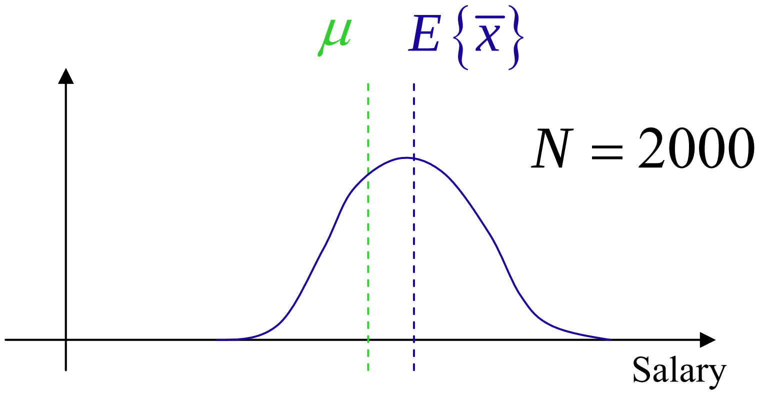
Sampling distribution

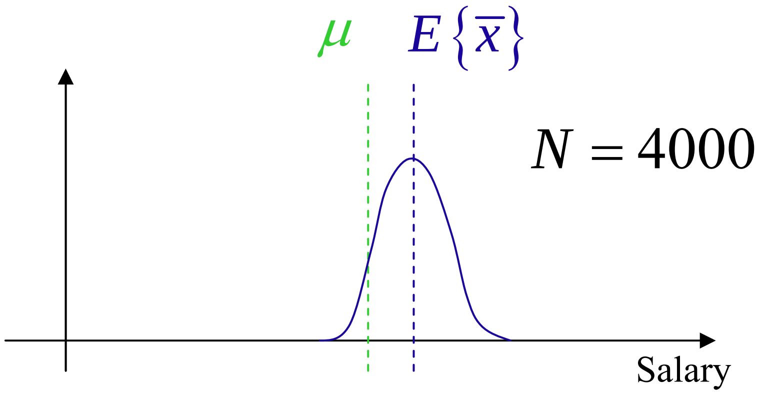
Bias of the statistic

Variance of the statistic: Standard error

Salary
$\overline{x}_4\overline{x}_3\overline{x}_2\overline{x}_5\overline{x}_1$

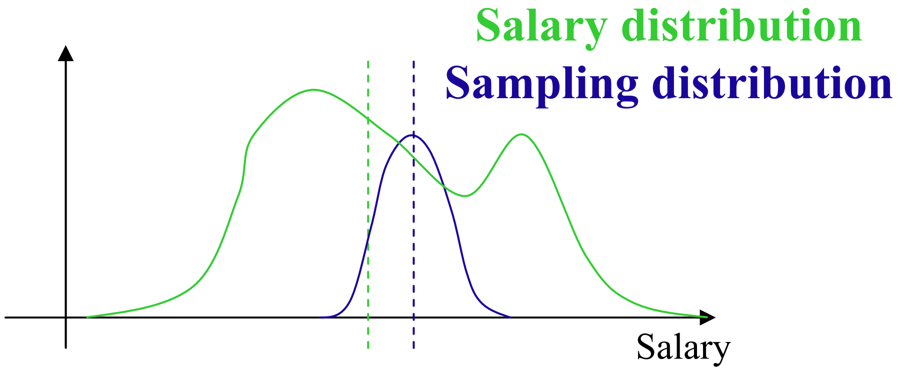
1.2 Statistic vs. Parameter. What is a sampling distribution?





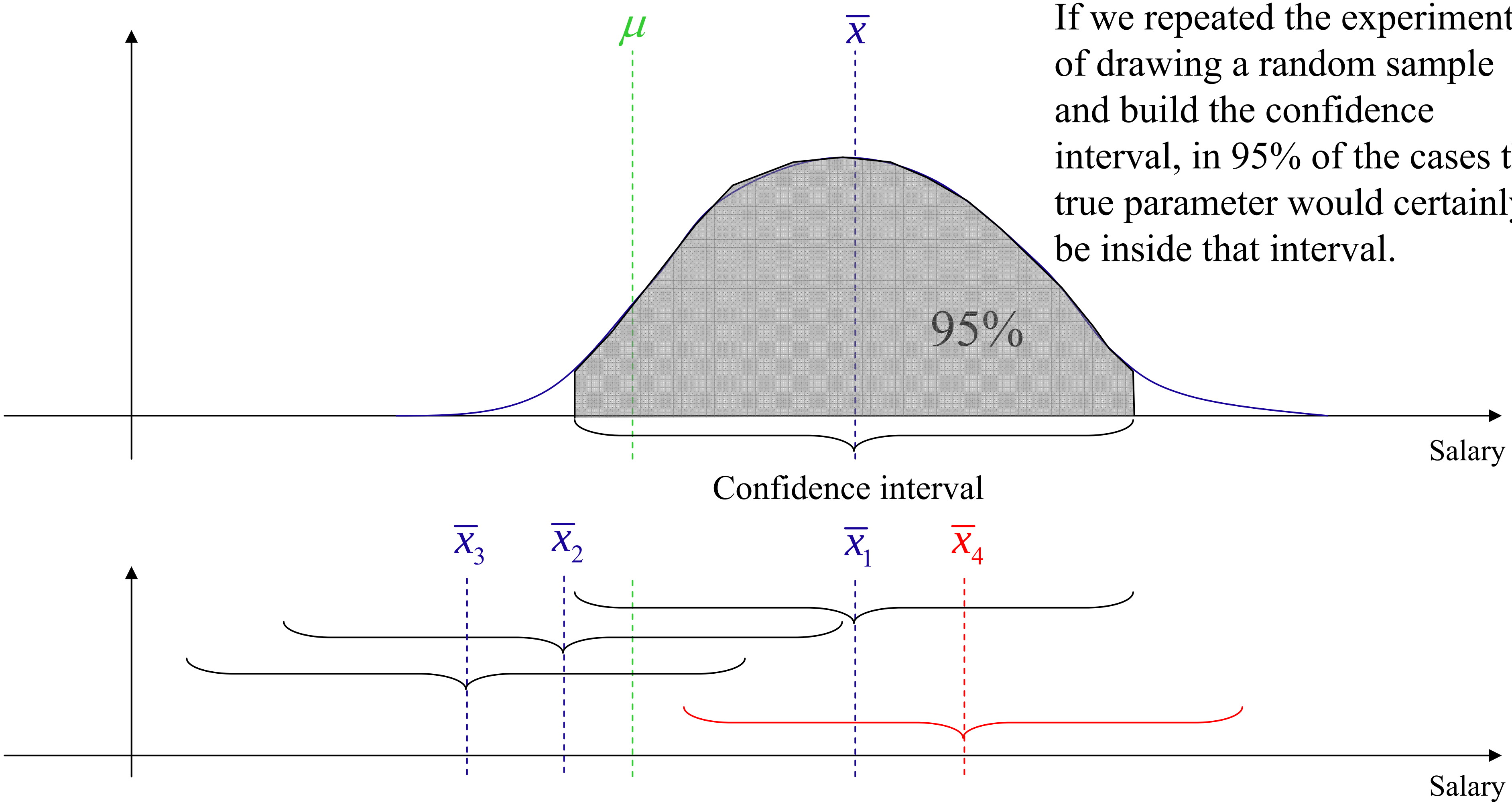
Unbiased $\mu - E\{\overline{x}\} = 0$

Asymptotically unbiased $\lim\limits_{N \to \infty} \mu - E\{\overline{x}\} = 0$



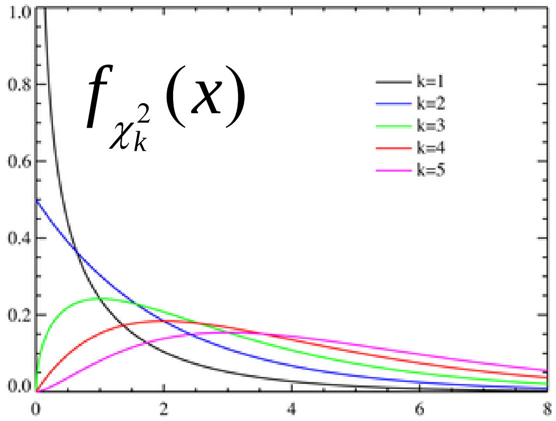
Sampling distribution: distribution of the statistic if all possible samples of size N were drawn from a given population

1.2 Statistic vs. Parameter. What is a sampling distribution?



If we repeated the experiment of drawing a random sample and build the confidence interval, in 95% of the cases the true parameter would certainly be inside that interval.

1.2 Statistic vs. Parameter. What is a sampling distribution?



Sometimes the distribution of the statistic is known

$$X_i \sim N(\mu,\sigma^2) \Rightarrow \frac{1}{N}\sum_{i=1}^{N} X_i \sim N\left(\mu,\frac{\sigma^2}{N}\right)$$

$$\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi_N^2$$

Sometimes the distribution of the statistic is NOT known, but still the mean is well behaved

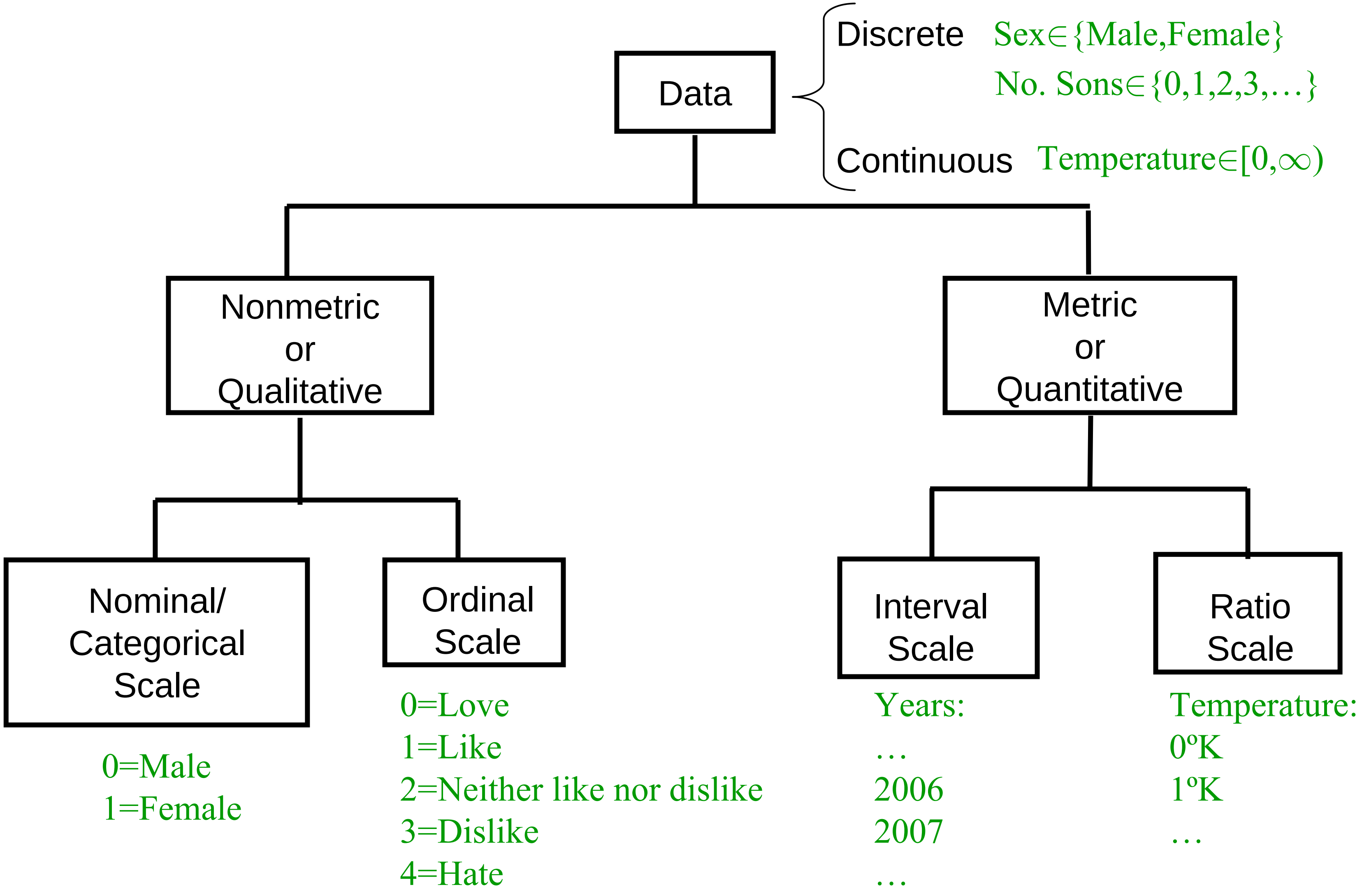
$$\begin{aligned} E\{X_i\} &= \mu \\ Var\{X_i\} &= \sigma^2 \end{aligned} \quad \Rightarrow \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_i \sim N\left(\mu,\frac{\sigma^2}{N}\right)$$

Central limit theorem!!

But:

- The sample must be truly random
- Averages based on samples whose size is more than 30 are reasonably Gaussian

1.3 Introduction: Types of variables



Data

Discrete Sex∈{Male,Female}

No. Sons∈{0,1,2,3,…}

Continuous Temperature∈[0,∞)

Nonmetric or Qualitative

Metric or Quantitative

Nominal/ Categorical Scale

0=Male
1=Female

Ordinal Scale

0=Love
1=Like
2=Neither like nor dislike
3=Dislike
4=Hate

Interval Scale

Years:
…
2006
2007
…

Ratio Scale

Temperature:
0ºK
1ºK
…

1.3 Introduction: Types of variables

Coding of categorical variables

Hair Colour
{Brown, Blond, Black, Red} $\xrightarrow{\text{No order}}$ $\left(x_{Brown}, x_{Blond}, x_{Black}, x_{Red}\right) \in \{0,1\}^4$

Peter: Black	Peter: $\{0,0,1,0\}$
Molly: Blond	Molly: $\{0,1,0,0\}$
Charles: Brown	Charles: $\{1,0,0,0\}$

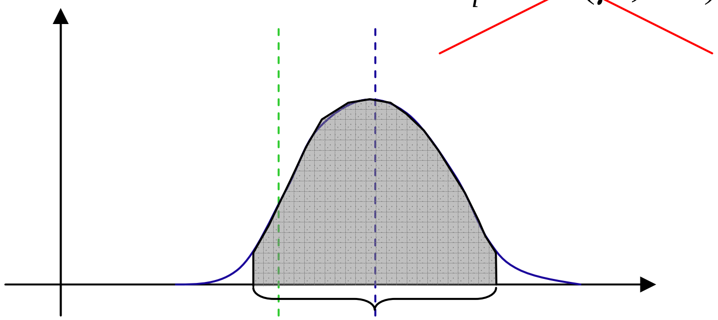
Company size
{Small, Medium, Big} $\xrightarrow{\text{Implicit order}}$ $x_{size} \in \{0,1,2\}$

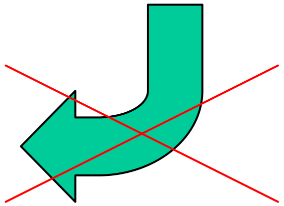
Company A: Big	Company A: 2
Company B: Small	Company B: 0
Company C: Medium	Company C: 1

1.4 Parametric vs. Non-parametric Statistics

Parameter estimation

$$\xcancel{X_i \sim N(\mu,\sigma^2)} \Rightarrow \frac{1}{N}\sum_{i=1}^{N} X_i \sim \xcancel{N\left(\mu,\frac{\sigma^2}{N}\right)}$$





Solution: Resampling (bootstrap, jacknife, …)

Hypothesis testing

Cannot use statistical tests based on any assumption about the distribution of the underlying variable (t-test, F-tests, χ^2-tests, …)

Solution:

- discretize the data and use a test for categorical/ordinal data (non-parametric tests)
- use randomized tests

1.5 What to measure? Central tendency

During the last 6 months the rentability of your account has been:
5%, 5%, 5%, -5%, -5%, -5%. Which is the average rentability of your account?

Arithmetic mean

$$x^*_{AM} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

(-) Very sensitive to large outliers, not too meaningful for certain distributions

(+) Unique, unbiased estimate of the population mean,
better suited for symmetric distributions

$$x^*_{AM} = \frac{1}{6}(5+5+5-5-5-5) = 0\%$$

Property $E\left\{x^*_{AM}\right\} = \mu$

$$x^*_{AM} = \frac{1}{6}(1.05+1.05+1.05+0.95+0.95+0.95) = 1 = 0\%$$

Geometric mean

$$x^*_{GM} = \sqrt[N]{\prod_{i=1}^{N} x_i} \Rightarrow \log x^*_{GM} = \frac{1}{N}\sum_{i=1}^{N} \log x_i$$

(-) Very sensitive to outliers

(+) Unique, used for the mean of ratios and percent changes,
less sensitive to asymmetric distributions

$$x^*_{GM} = \sqrt[6]{1.05\cdot 1.05\cdot 1.05\cdot 0.95\cdot 0.95\cdot 0.95} = 0.9987 = -0.13\%$$

Which is right? 1 ? $\rightarrow 1050 \rightarrow 1102.5 \rightarrow 1157.6 \rightarrow 1099.7 \rightarrow 1044.8 \rightarrow 992.5$

1.5 What to measure? Central tendency

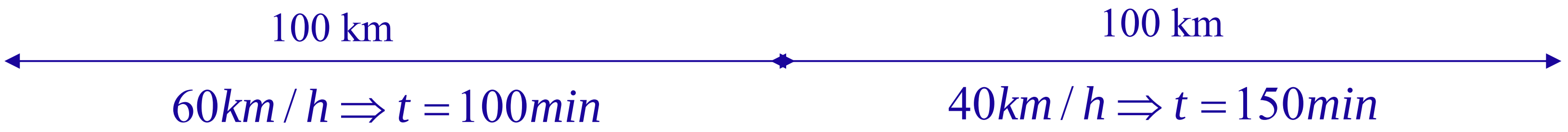
Harmonic mean

(-) Very sensitive to small outliers

(+) Usually used for the average of rates,
less sensitive to large outliers

$$x^*_{HM} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}} \Rightarrow \frac{1}{x^*_{HM}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}$$

A car travels 200km. The first 100 km at a speed of 60km/h, and the second 100 km at a speed of 40 km/h.



$$60km/h \Rightarrow t = 100min \qquad 40km/h \Rightarrow t = 150min$$

$$x^*_{HM} = \frac{1}{\frac{1}{2}\left(\frac{1}{60}+\frac{1}{40}\right)} = 48km/h$$

$$x^*_{AM} = \frac{1}{2}\left(60+40\right) = 50km/h$$

Which is the right average speed?

1.5 What to measure? Central tendency

Property: For positive numbers $\boxed{x^*_{HM} \leq x^*_{GM} \leq x^*_{AM}}$

More affected by extreme large values
Less affected by extreme small values

Less affected by extreme values
More affected by extreme small values

Generalization: Generalized mean $x^* = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^p \right)^{\frac{1}{p}}$

Minimum	$p = -\infty$
Harmonic mean	$p = -1$
Geometric mean	$p = 0$
Arithmetic mean	$p = 1$
Quadratic mean	$p = 2$
Maximum	$p = \infty$

1.5 What to measure? Robust central tendency

During the last 6 months the rentability of your account has been:
5%, 3%, 7%, -15%, 6%, 30%. Which is the average rentability of your account?

x_1	x_2	x_3	x_4	x_5	x_6
$x_{(3)}$	$x_{(2)}$	$x_{(5)}$	$x_{(1)}$	$x_{(4)}$	$x_{(6)}$

<u>Trimmed mean, truncated mean, Windsor mean:</u>

Remove p% of the extreme values on each side

$$x^* = \frac{1}{4}\left(x_{(2)} + x_{(3)} + x_{(4)} + x_{(5)}\right) = \frac{1}{4}\left(3+5+6+7\right) = 5.25\%$$

<u>Median</u>

Which is the central sorted value? (50% of the distribution is below that value) It is not unique

Any value between $x_{(3)} = 5\%$ and $x_{(4)} = 6\%$

<u>Winsorized mean:</u>

Substitute p% of the extreme values on each side

$$x^* = \frac{1}{6}\left(x_{(2)} + x_{(2)} + x_{(3)} + x_{(4)} + x_{(5)} + x_{(5)}\right) = \frac{1}{6}\left(3+3+5+6+7+7\right) = 5.1\overset{\frown}{6}\%$$

1.5 What to measure? Robust central tendency

M-estimators

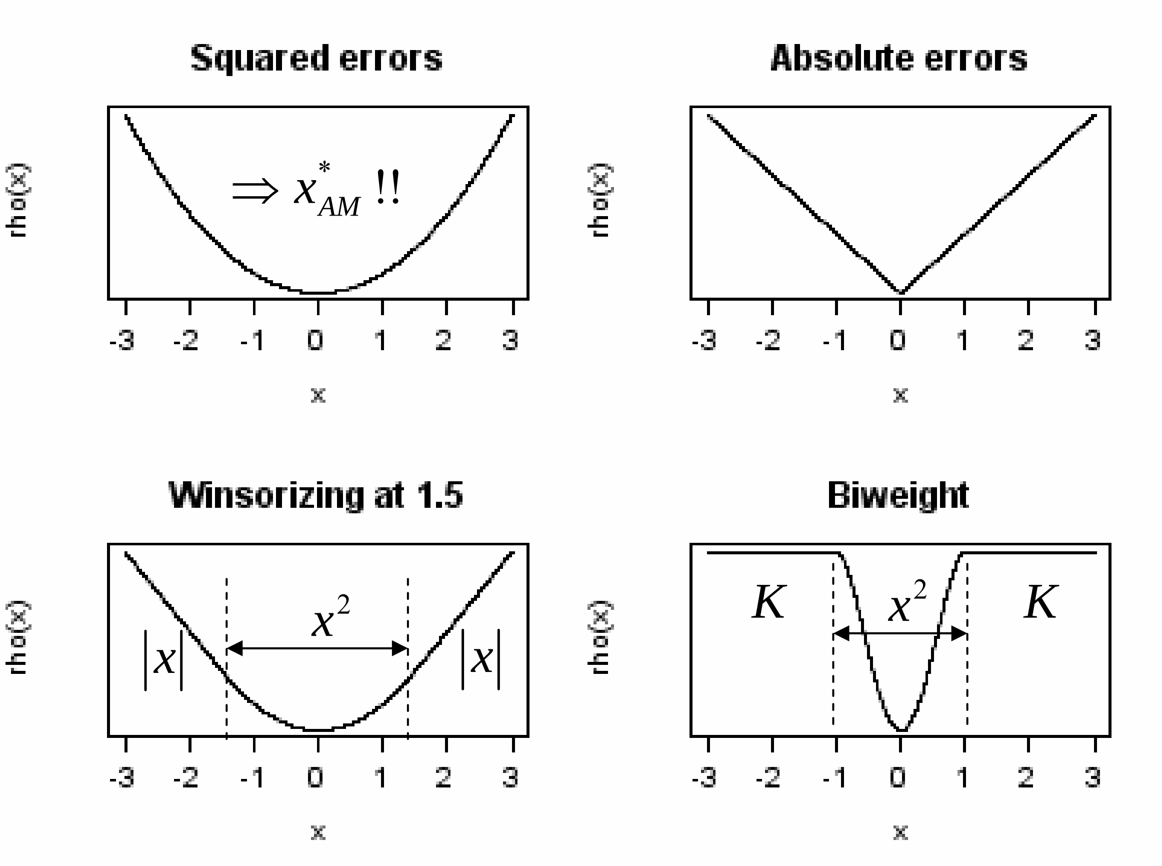
Give different weight to different values

$$x^* = \arg\min_x \frac{1}{N}\sum_{i=1}^{N} \rho(x_i - x)$$

R and L-estimators

Now in disuse

The distribution of robust statistics is usually unknown and has to be estimated experimentally (e.g., bootstrap resampling)



1.5 What to measure? Central tendency

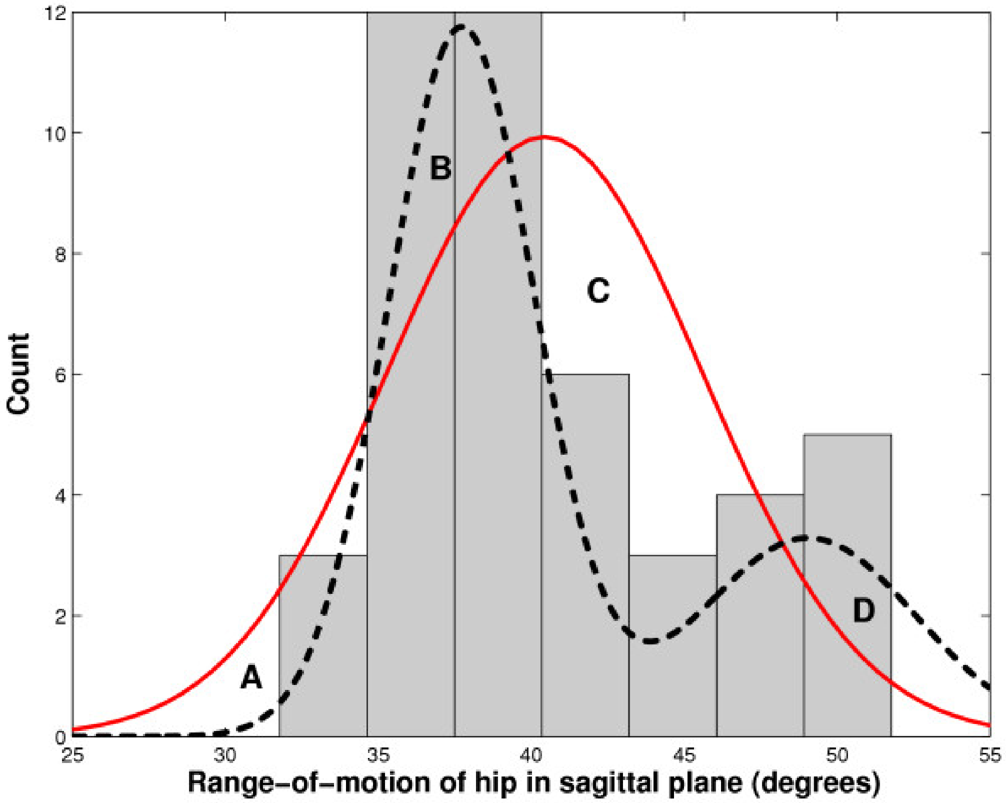
Mode:

Most frequently occurring

(-) Not unique (multimodal)

(+) representative of the most "typical" result

$$x^* = \arg\max f_X(x)$$



If a variable is multimodal, most central measures fail!

1.5 What to measure? Central tendency

- What is the geometric mean of {-2,-2,-2,-2}? Why is it so wrong?

- The arithmetic mean of {2,5,15,20,30} is 14.4, the geometric mean is 9.8, the harmonic mean is 5.9, the median is 15. Which is the right central value?

1.5 What to measure? Differences

An engineer tries to determine if a certain modification makes his motor to waste less power. He makes measurements of the power consumed with and without modifications (the motors tested are different in each set). The nominal consumption of the motors is 750W, but they have from factory an unknown standard deviation around 20W. He obtains the following data:

Unmodified motor (Watts): 741, 716, 753, 756, 727 $\bar{x} = 738.6$
Modified motor (Watts): 764, 764, 739, 747, 743 $\bar{y} = 751.4$

Not robust measure of unpaired differences $d^* = \bar{y} - \bar{x}$

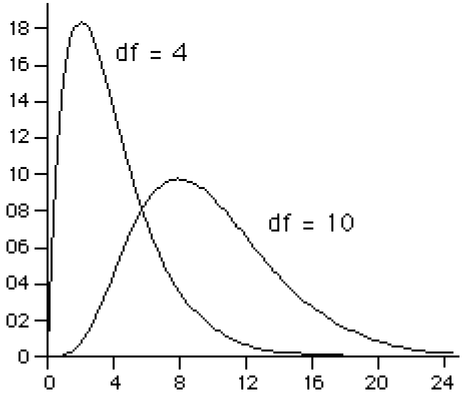
Robust measure of unpaired differences $d^* = median\left\{y_i - x_j\right\}$

If the measures are paired (for instance, the motors are first measured, the modified and remeasured), then we should first compute the difference.

Difference: 23, 48, -14, -9, 16 $d^* = \bar{d}$

1.5 What to measure? Variability

During the last 6 months the rentability of an investment product has been:
-5%, 10%, 20%, -15%, 0%, 30% (geometric mean=5.59%)
The rentability of another one has been: 4%, 4%, 4%, 4%, 4%, 4%
Which investment is preferrable for a month?



Variance

(-) In squared units
(+) Very useful in analytical expressions

$$\sigma^2 = E\left\{(X-\mu)^2\right\}$$

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 \qquad E\left\{s_N^2\right\} \underset{\uparrow}{=} \frac{N-1}{N}\sigma^2$$

Subestimation of the variance

$$\boxed{X_i \sim N(\mu,\sigma^2) \Rightarrow (N-1)\frac{s^2}{\sigma^2} \sim \chi^2_{N-1}}$$

$$s^2\left\{0.95,1.10,1.20,0.85,1.00,1.30\right\} = 0.0232$$

Rentability=5.59±2.32%

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2 \qquad E\left\{s^2\right\} = \sigma^2$$

1.5 What to measure? Variability

Standard deviation

(+) In natural units,

provides intuitive information about variability

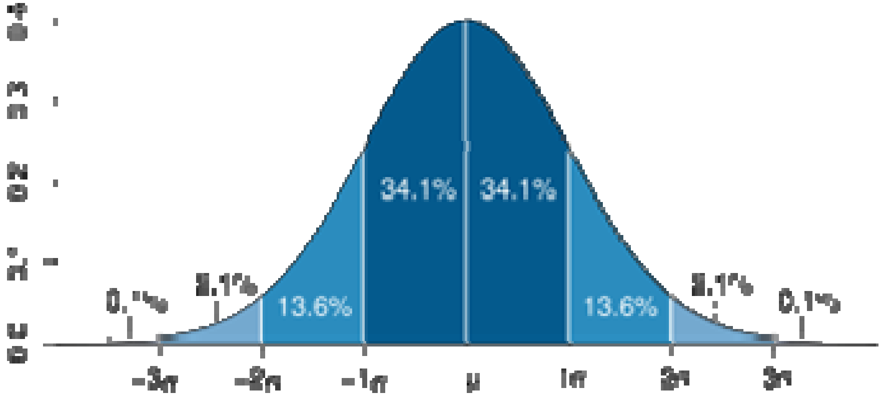
Natural estimator of measurement precision

Natural estimator of range excursions

$$s=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})^2}$$

Rentability=5.59±√0.0232=5.59±15.23%

$$\boxed{X_i \sim N(\mu,\sigma^2) \Rightarrow \sqrt{N-1}\,\frac{s}{\sigma} \sim \chi_{N-1}}$$



Tchebychev’s Inequality

$$\Pr\left\{\mu-K\sigma \le X \le \mu+K\sigma\right\}=1-\frac{1}{K^2}$$

At least 50% of the values are within √2 standard deviations from the mean.
At least 75% of the values are within 2 standard deviations from the mean.
At least 89% of the values are within 3 standard deviations from the mean.
At least 94% of the values are within 4 standard deviations from the mean.
At least 96% of the values are within 5 standard deviations from the mean.
At least 97% of the values are within 6 standard deviations from the mean.
At least 98% of the values are within 7 standard deviations from the mean.

For any distribution!!!

1.5 What to measure? Variability

Percentiles

(-) Difficult to handle in equations
(+) Intuitive definition and meaning
(+) Robust measure of variability

$$\Pr\left\{X \leq x^{*}\right\} = q$$

Someone has an IQ score of 115. Is he clever, very clever, or not clever at all?

Deciles

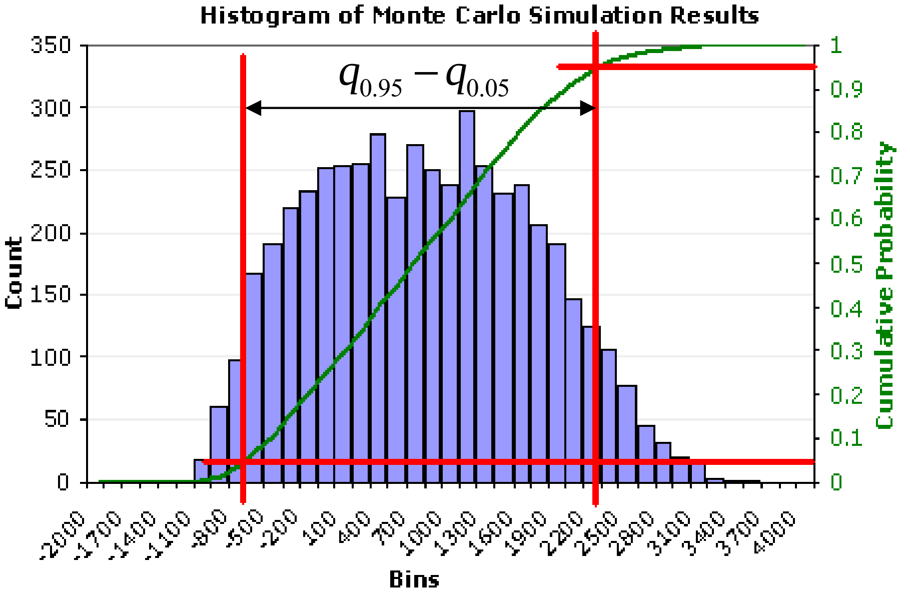
$q_{0.10}, q_{0.20}, q_{0.30}, q_{0.40}, q_{0.50}$
$q_{0.60}, q_{0.70}, q_{0.80}, q_{0.90}$

$$q_{0.90} - q_{0.10}$$

Quartiles

$q_{0.25}, q_{0.50}, q_{0.75}$

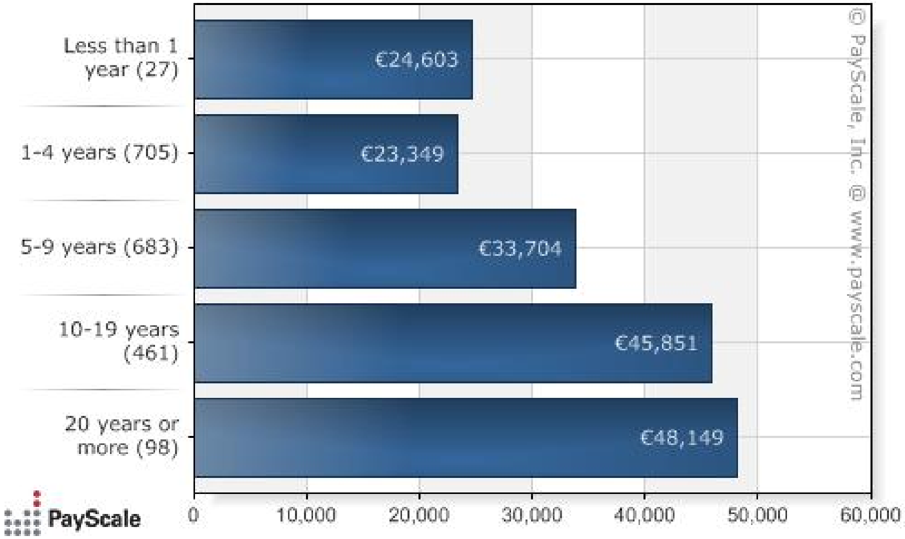
$$q_{0.75} - q_{0.25}$$



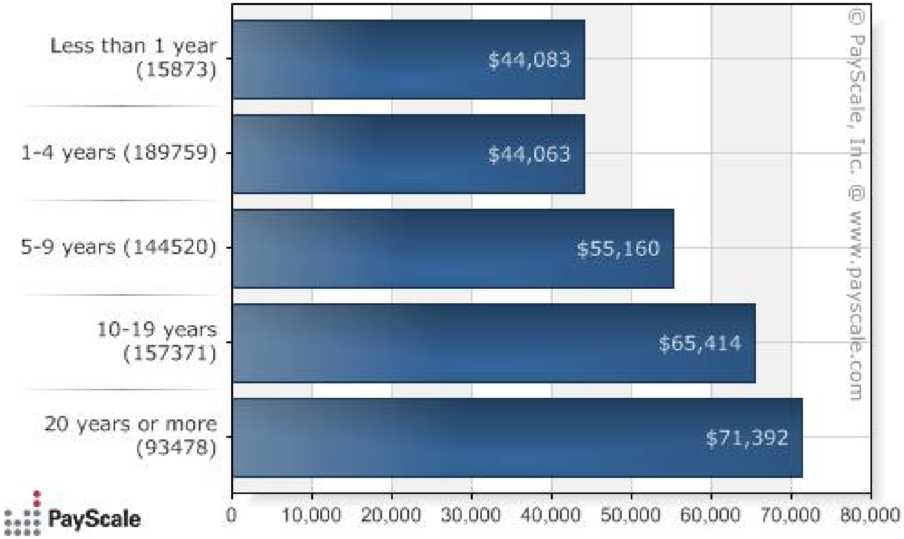
1.5 What to measure? Variability

Coefficient of variation

Median salary in Spain by years of experience



Median salary in US by years of experience



In which country you can have more progress along your career?

1.5 What to measure? Skewness

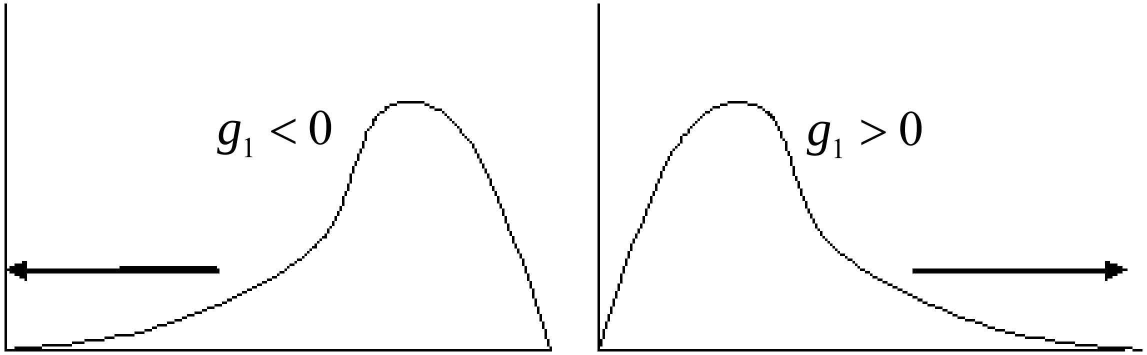
Skewness: Measure of the assymetry of a distribution

$$\gamma_1 = \frac{E\left\{\left(X-\mu\right)^3\right\}}{\sigma^3}$$

Unbiased estimator

$$g_1 = \frac{m_3}{s^3}$$

$$m_3 = \frac{N\sum_{i=1}^{N}(x_i-\overline{x})^3}{(N-1)(N-2)}$$



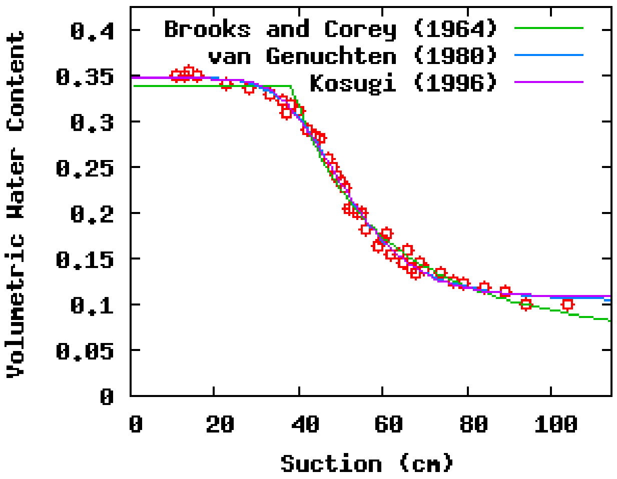
Negative Skew
Elongated tail at the **left**
More data in the left tail than would be expected in a normal distribution

$$\mu < Med < Mode$$

Positive Skew
Elongated tail at the **right**
More data in the right tail than would be expected in a normal distribution

$$Mode > Med > \mu$$

The residuals of a fitting should not be skew! Otherwise, it would mean that positive errors are more likely than negative or viceversa. This is the rationale behind some goodness-of-fit tests.



1.5 What to measure? Correlation/Association

Is there any relationship between education, free-time and salary?

Person	Education (0-10)	Education	Free-time (hours/week)	Salary $	Salary
A	10	High	10	70K	High
B	8	High	15	75K	High
C	5	Medium	27	40K	Medium
D	3	Low	30	20K	Low

Pearson's correlation coefficient

$$\rho = \frac{E\left\{\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right\}}{\sigma_X \sigma_Y} \in [-1,1]$$

$$r = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{s_X s_Y}$$

$Salary \uparrow \Rightarrow FreeTime \downarrow$ ↓

$Education \uparrow \Rightarrow Salary \uparrow$ ↓

Correlation	Negative	Positive
Small	−0.3 to −0.1	0.1 to 0.3
Medium	−0.5 to −0.3	0.3 to 0.5
Large	−1.0 to −0.5	0.5 to 1.0

1.5 What to measure? Correlation/Association

Correlation between two ordinal variables? Kendall’s tau

Is there any relationship between education and salary?

Person	Education	Salary $
A	10	70K
B	8	75K
C	5	40K
D	3	20K

Person	Education	Salary $
A	1st	2nd
B	2nd	1st
C	3rd	3rd
D	4th	4th

P=Concordant pairs

Person A

Education: (A>B) (A>C) (A>D)
Salary: (A>C) (A>D)

2

Person B

Education: (B>C) (B>D)
Salary: (B>A) (B>C) (B>D)

2

Person C

Education: (C>D)
Salary: (C>D)

1

Person D

Education:
Salary:

0

$$\tau = \frac{P}{\frac{N(N-1)}{2}}$$

$$\tau = \frac{2+2+1+0}{\frac{4(4-1)}{2}} = \frac{5}{6} = 0.83$$

1.5 What to measure? Correlation/Association

Correlation between two ordinal variables? Spearman’s rho

Is there any relationship between education and salary?

Person	Education	Salary $
A	10	70K
B	8	75K
C	5	40K
D	3	20K

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

Person	Education	Salary $	di
A	1st	2nd	-1
B	2nd	1st	1
C	3rd	3rd	0
D	4th	4th	0

$$\rho = 1 - \frac{6((-1)^2 + 1^2 + 0^2 + 0^2)}{4(4^2-1)} = 0.81$$

1.5 What to measure? Correlation/Association

Other correlation flavours:

- Correlation coefficient: How much of Y can I explain given X?
- Multiple correlation coefficient: How much of Y can I explain given X_1 and X_2?
- Partial correlation coefficient: How much of Y can I explain given X_1 once I remove the variability of Y due to X_2?
- Part correlation coefficient: How much of Y can I explain given X_1 once I remove the variability of X_1 due to X_2?

CEU Universidad San Pablo

1.6 Use and abuse of the normal distribution

Univariate $$X \sim N(\mu,\sigma^2) \Rightarrow f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

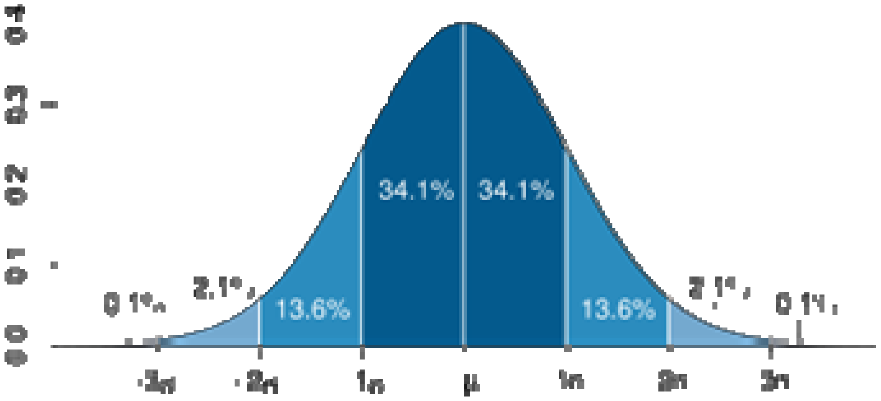
Multivariate $$\mathbf{X} \sim N(\boldsymbol{\mu},\Sigma) \Rightarrow f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}\text{-}\boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}\text{-}\boldsymbol{\mu})}$$

Covariance matrix

Use: Normalization

$$X \sim N(\mu,\sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$$

Z-score



Compute the z-score of the IQ $(\mu = 100, \sigma = 15)$ of:
Napoleon Bonaparte (emperor): 145
Gary Kasparov (chess): 190

$$z_{Napoleon} = \frac{145-100}{15} = 3$$

$$z_{Kasparov} = \frac{190-100}{15} = 6$$

1.6 Use and abuse of the normal distribution

Use: Computation of probabilties IF the underlying variable is normally distributed

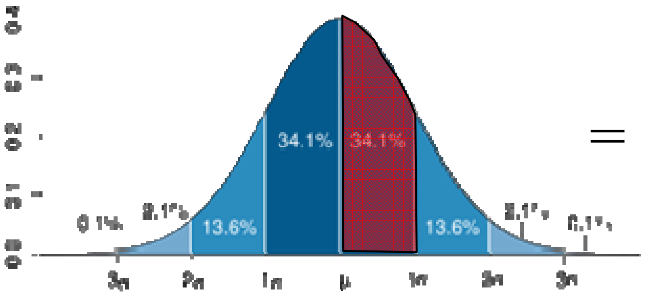
$$X \sim N(\mu,\sigma^2)$$

What is the probability of having an IQ between 100 and 115?

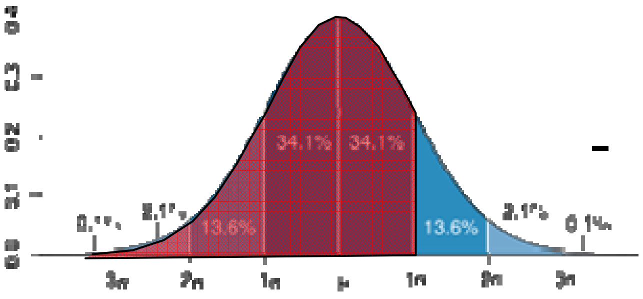
$$\Pr\{100 \le IQ \le 115\} = \int_{100}^{115} \frac{1}{\sqrt{2\pi 15^2}} e^{-\frac{1}{2}\left(\frac{x-100}{15}\right)^2} dx = \int_{0}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx - \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.341$$

Normalization

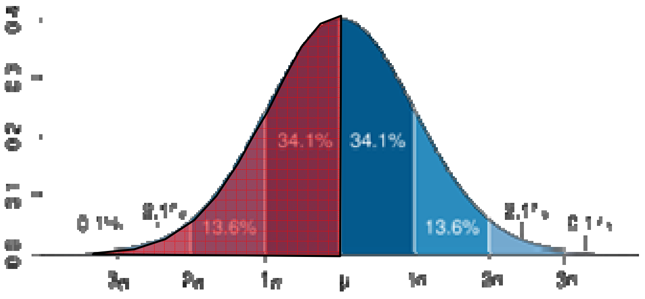
Use of tabulated values



=



-



1.6 Use and abuse of the normal distribution

Use: Computation of probabilties IF the underlying variable is normally distributed

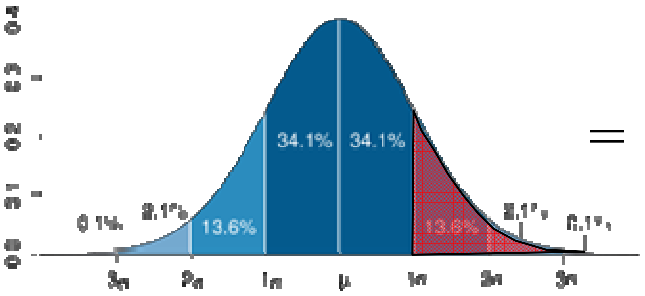
$$X \sim N(\mu, \sigma^2)$$

What is the probability of having an IQ larger than 115?

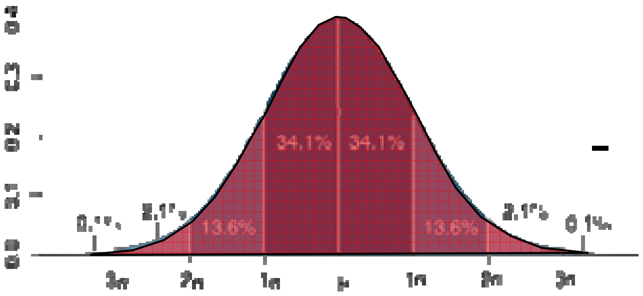
$$\Pr\{100 \leq IQ \leq 115\} = \int_{115}^{\infty} \frac{1}{\sqrt{2\pi 15^2}} e^{-\frac{1}{2}\left(\frac{x-100}{15}\right)^2} dx = \int_{1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1 - \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0.159$$

Normalization

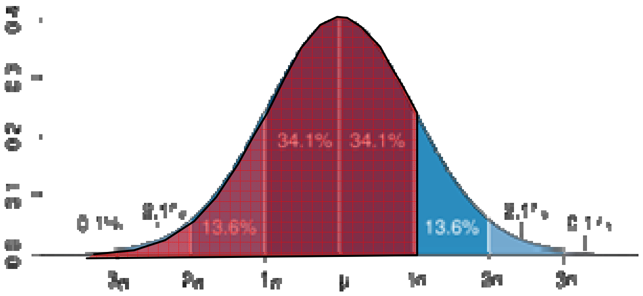
Use of tabulated values



=



-



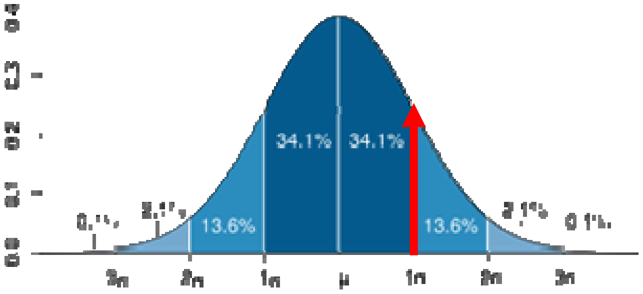
1.6 Use and abuse of the normal distribution

Abuse: Computation of probabilties of a single point

What is the probability of having an IQ exactly equal to 115?

$$\Pr\left\{IQ=115\right\}=0$$

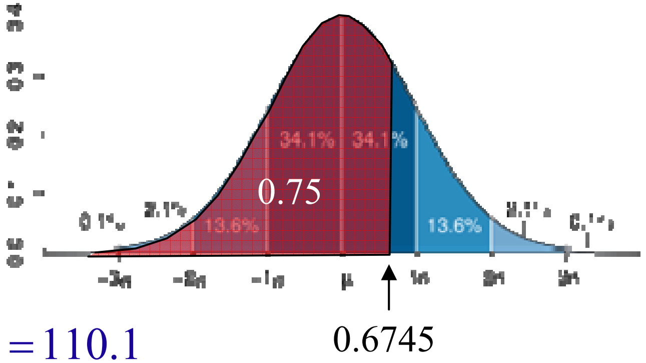
$$Likelihood\left\{IQ=115\right\}=Likelihood\left\{z_{IQ}=1\right\}=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}}$$



Use: Computation of percentiles

Which is the IQ percentile of 75%?

$$\int_{-\infty}^{q_{0.75}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx=75\Rightarrow q_{0.75}=0.6745$$



$$IQ_{0.75}=\mu_{IQ}+q_{0.75}\sigma_{IQ}=100+0.6745\cdot 15=110.1$$

1.6 Use and abuse of the normal distribution

Abuse: Assumption of normality

Many natural fenomena are normally distributed (thanks to the central limit theorem):

- Error in measurements
- Light intensity
- Counting problems when the count number is very high (persons in the metro at peak hour)
- Length of hair
- The logarithm of weight, height, skin surface, … of a person

But many others are not

- The number of people entering a train station in a given minute is not normal, but the number of people entering all the train stations in the world at a given minute is normal.

Many distributions of mathematical operations are normal

$$X_i \sim N \xrightarrow{\hspace{2cm}} aX_1 + bX_2; a + bX_1 \sim N$$

But many others are not

$$X_i \sim N \xrightarrow{\hspace{2cm}} \frac{X_1}{X_2} \sim Cauchy; e^X \sim LogNormal; \frac{\sum X_i^2}{\sum X_j^2} \sim F - Snedecor$$

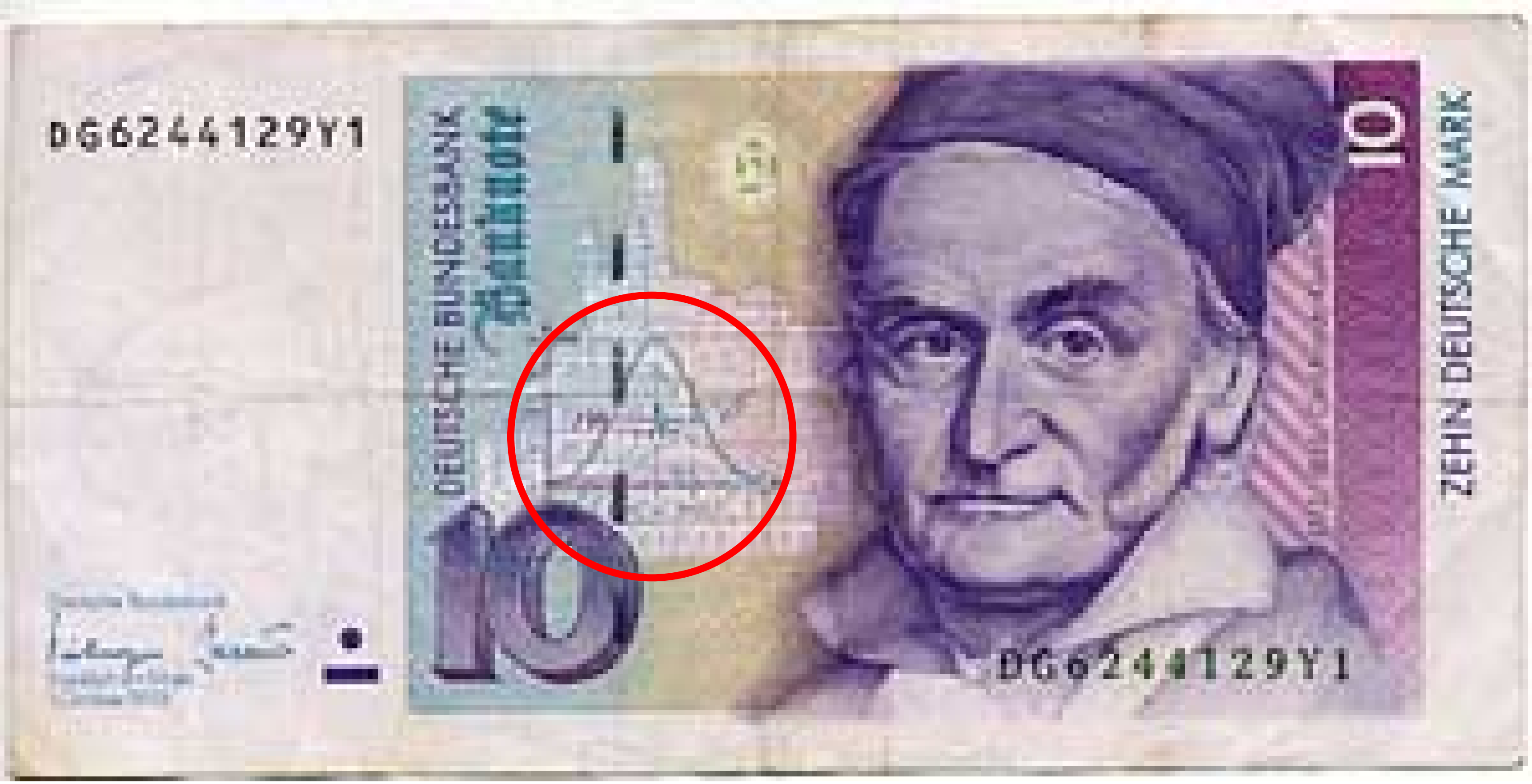
$$\sum X_i^2 \sim \chi^2; \sqrt{\sum X_i^2} \sim \chi; \sqrt{X_1^2 + X_2^2} \sim Rayleigh$$

Some distributions can be safely approximated by the normal distribution

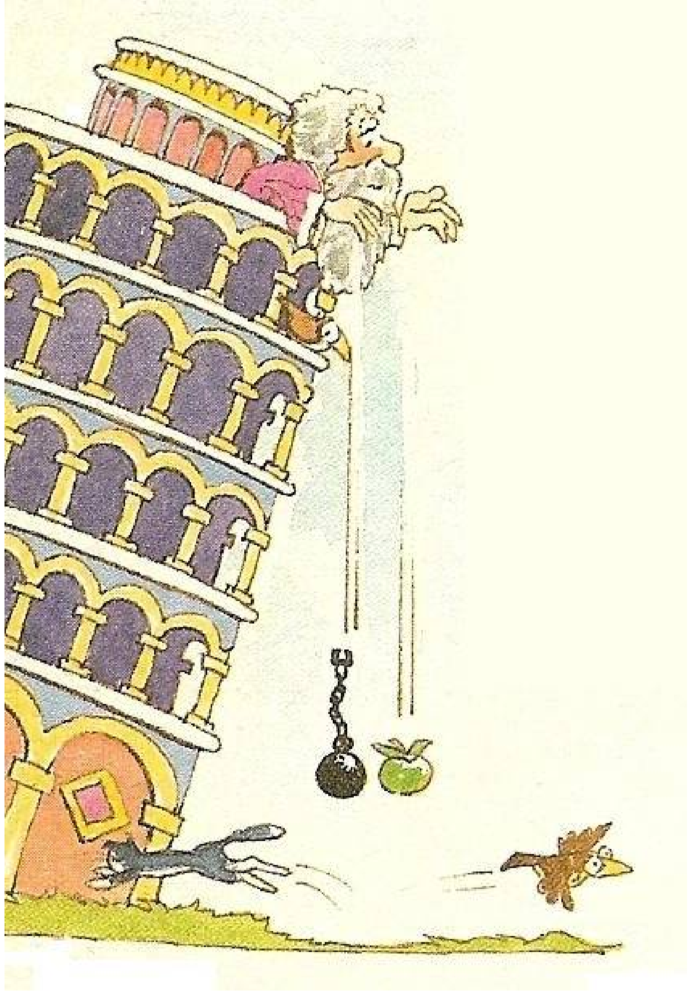
- Binomial $np > 10 \; and \; np(1-p) > 10$, Poisson $\lambda > 1000$

1.6 Use and abuse of the normal distribution

Abuse: banknotes



1.6 Use and abuse of the normal distribution



$$\begin{tikzcd}
t(\text{sec}) \sim N(t_0,\sigma^2) \arrow[dr] \arrow[ddr] & \\
& t(\text{msec}) \sim N \\
& h = \frac{1}{2} g t^2 \sim N
\end{tikzcd}$$

1.7 Is my data really independent?

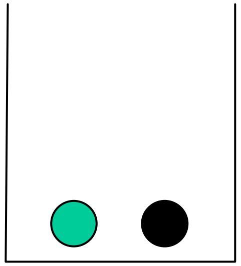
Independence is different from mutual exclusion

In general,

$$\begin{tikzcd}
p(A\cap B)=p(A)\,p(B\mid A) \arrow[d, "p(B\mid A)=0"] \arrow[dd, bend left, "\text{Knowing A does not give any information about the next event}"] \\
p(A\cap B)=0 \\
p(A\cap B)=p(A)\,p(B)
\end{tikzcd}$$

Mutual exclusion is when two results are impossible to happen at the same time.

Independence is when the probability of an event does not depend on the results that we have had previously.



Example: Sampling with and without replacement

What is the probability of taking a black ball as second draw, if the first draw is green?

1.7 Is my data really independent?

Sampling without replacement

In general samples are not independent except if the population is so large that it does not matter.

Sampling with replacement

Samples may be independent. However, they may not be independent (see Example 1)
Examples: tossing a coin, rolling a dice

Random sample: all samples of the same size have equal probability of being selected

Example 1: Study about child removal after abuse, 30% of the members were related to each other because when a child is removed from a family, normally, the rest of his/her siblings are also removed. Answers for all the siblings are correlated.

Example 2: Study about watching violent scenes at the University. If someone encourages his roommate to take part in this study about violence, and the roommate accepts, he is already biased in his answers even if he is acting as control watching non-violent scenes.

Consequence: The sampling distributions are not what they are expected to be, and all the confidence intervals and hypothesis testing may be seriously compromised.

1.7 Is my data really independent?

- A newspaper makes a survey to see how many of its readers like playing videogames. The survey is announced in the paper version of the newspaper but it has to be filled on the web. After processing they publish that 66% of the newspaper readers like videogames. Is there anything wrong with this conclusion?

1.7 Is my data really independent?

“A blond woman with a ponytail snatched a purse from another woman. The thief fled in a yellow car driven by a black man with a beard and moustache”.

A woman matching this description was found. The prosecution assigned the following probabilities: blond hair (1/3), ponytail (1/10), yellow car (1/10), black man with beard (1/10), moustache (1/4), interracial couple in car (1/1000). The multiplication of all these probabilities was 1/12M and the California Supreme Court convicted the woman in 1964.

Is there anything wrong with the reasoning?