Data Analysis

Patrice Koehl
Department of Biological Sciences
National University of Singapore

http://www.cs.ucdavis.edu/~koehl/Teaching/BL5229
koehl@cs.ucdavis.edu

Data analysis

- Statistics of a sample
  - Central tendency
  - Variation
  - Normal distribution
- Inference
  - From sample to population
  - P-value

Data analysis

- Statistics of a sample
  - Central tendency
  - Variation
  - Normal distribution
- Inference
  - From sample to population
  - P-value

Measures of Central Tendency

- Mean … the average score

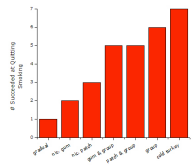
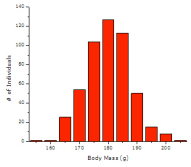
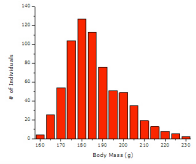
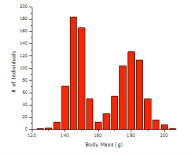
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Median … the value that lies in the middle after ranking all the scores

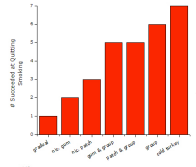
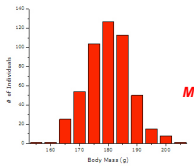
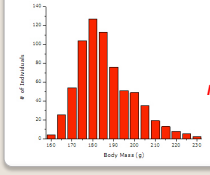
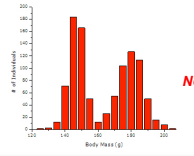
$$X_M = \begin{cases} X_{n/2+1} & n \text{ odd} \\ \dfrac{X_{n/2} + X_{n/2+1}}{2} & n \text{ even} \end{cases}$$

- Mode … the most frequently occurring score

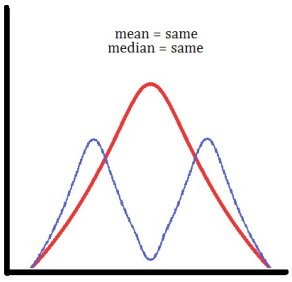
Which Measure should you use?

Which Measure should you use?


Mode

Mean

Median

None

Measures of Central Tendency
Attention: danger!



Variation or Spread of Distributions

- Range

$$Range = X_{Max} - X_{Min}$$

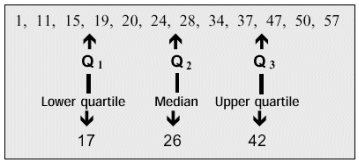
- Variance and Standard Deviation

$$Var(X) = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

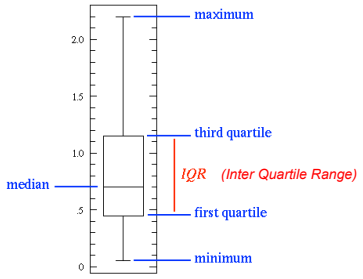
$$Std(X) = \sigma = \sqrt{Var(X)}$$

Variation or Spread of Distributions

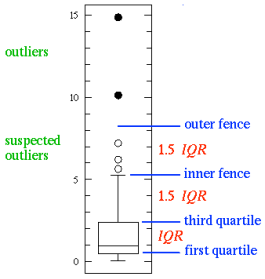
- Quartiles



Visualization of Variation: the Box Plot



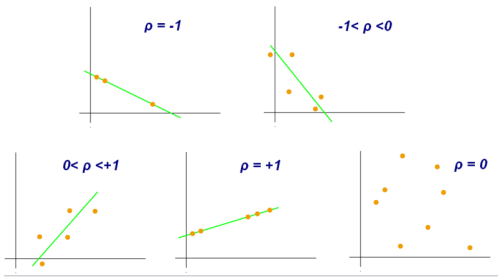
Detecting outliers



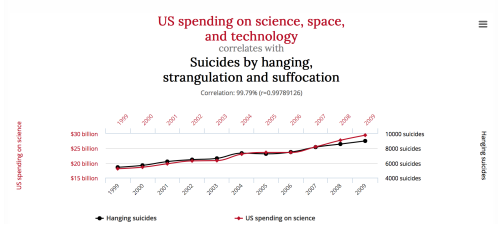
Correlation between two distributions

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

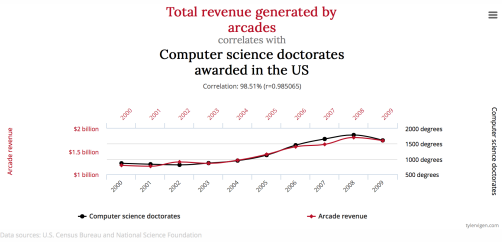
Correlation between two distributions



Correlation does not mean causality



Correlation does not mean causality

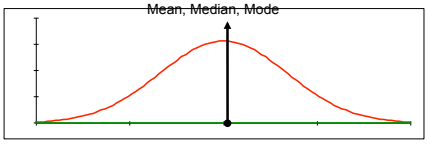


Data analysis

- Statistics of a sample
  - Central tendency
  - Variation
  - Normal distribution
- Inference
  - From sample to population
  - P-value

The Normal Distribution Curve

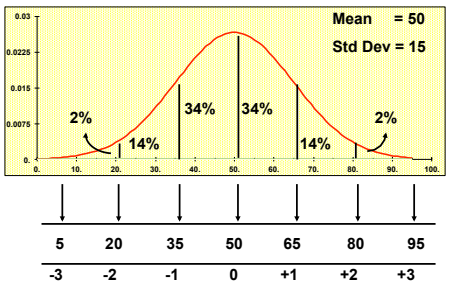
In everyday life many variables such as height, weight, shoe size and exam marks all tend to be normally distributed, that is, they all tend to look like:



It is bell-shaped and symmetrical about the mean

The mean, median and mode are equal

Interpreting a normal distribution



Statistical Inference

The process of making guesses about the truth from a sample

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X}_n)^2}{n-1}$$

Truth (not observable)

Population parameters

$$\mu = \frac{\sum_{i=1}^{N} x}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Sample (observation)

Make guesses about the whole population

The Central Limit Theorem

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

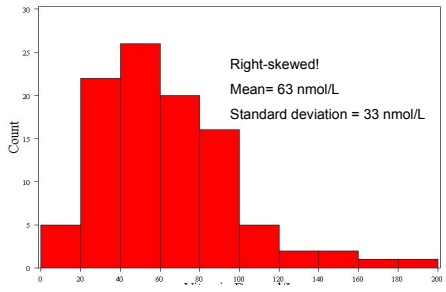
1. have mean: $\mu_{\bar{x}} = \mu$

2. have standard deviation: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

   (standard error)

3. be approximately normally distributed regardless of the shape of the parent population (normality improves with larger *n*)

Vitamin D



Right-skewed!

Mean= 63 nmol/L

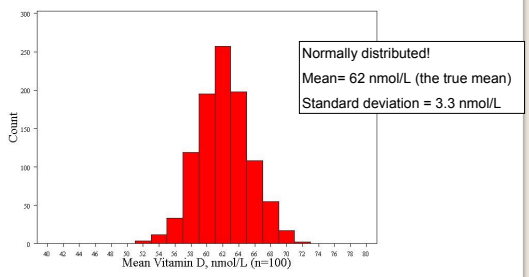
Standard deviation = 33 nmol/L

Lee DM, Tajar A, Ulubaev A, et al. Association between 25-hydroxyvitamin D levels and cognitive performance in middle-aged and older European men. J Neurol Neurosurg Psychiatry. 2009 Jul;80(7):722-9.

Distribution of the sample mean, computer simulation…

- Specify the underlying distribution of vitamin D in all European men aged 40 to 79.
  - Right-skewed
  - Standard deviation = 33 nmol/L
  - True mean = 62 nmol/L
- Select a random sample of 100 virtual men from the population.
- Calculate the mean vitamin D for the sample.
- Repeat steps (2) and (3) a large number of times (say 1000 times).
- Explore the distribution of the 1000 means.

Distribution of sample mean: vitamin D



Normally distributed!

Mean= 62 nmol/L (the true mean)

Standard deviation = 3.3 nmol/L

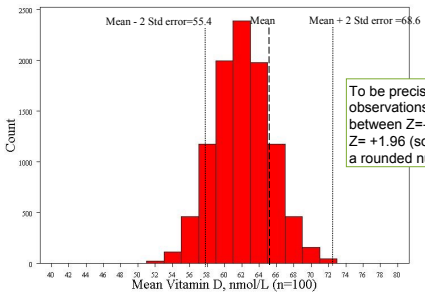
Confidence interval

Given a sample and its statistics (mean and standard deviation), is it possible to get an estimate of the true mean?

The confidence interval is set to capture the true effect "most of the time".

For example, a 95% confidence interval should include the true effect about 95% of the time.

Recall: 68-95-99.7 rule for normal distributions! These is a 95% chance that the sample mean will fall within two standard errors of the true mean= 62 +/- 2*3.3 = 55.4 nmol/L to 68.6 nmol/L



To be precise, 95% of observations fall between Z=-1.96 and Z= +1.96 (so the "2" is a rounded number)…

Confidence interval

The value of the statistic in the sample (mean)

point estimate ± ***(measure of how confident we want to be)*** · ***(standard error)***

Standard error of the statistics

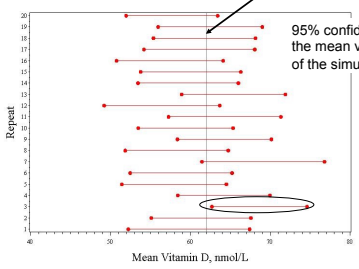
From a Z table or a T table, depending on the sampling distribution of the statistic.

Confidence Level	*Z value*
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.8%	3.08
99.9%	3.27

Confidence interval: simulation for Vitamin D study

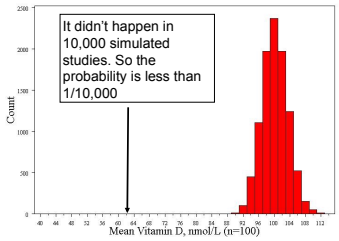
Vertical line indicates the true mean (62)

95% confidence intervals for the mean vitamin D for each of the simulated studies.



Hypothesis Testing: P-value

What's the probability of seeing a sample mean of 63 nmol/L if the true mean is 100 nmol/L?



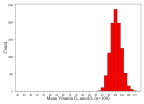
P-value is the **probability** that we would have seen our data just by chance if the null hypothesis (null value) is true.

Small p-values mean the null value is unlikely given our data.

Hypothesis Testing

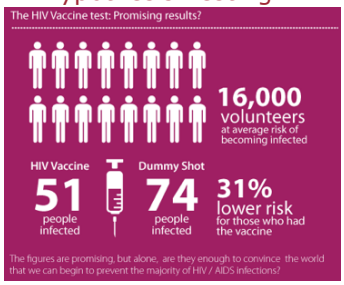
Steps:

1. Define your hypotheses (null, alternative)
   Mean = 100
2. Specify your null distribution



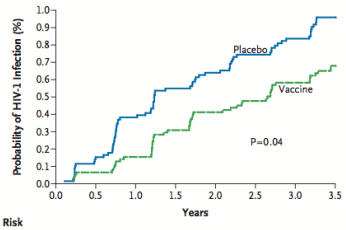
3. Do an experiment
   X = 63
4. Calculate the p-value of what you observed
   p < 0.001
5. Reject or fail to reject (~accept) the null hypothesis
   reject

Hypothesis Testing



(http://www.ngpharma.com/news/possible-HIV-vaccine/
http://news.bbc.co.uk/go/pr/fr/-/2/hi/health/8272113.stm
Rerks-Ngarm et al, New Eng. J. of Medicine, 361, 2209 (2009))

Hypothesis Testing



VE=31%

No. at Risk					
Placebo	8198	7775	7643	7441	7325
Vaccine	8197	7797	7665	7471	7347
Cumulative No. of Infections					
Placebo		30	50	65	74
Vaccine		12	32	45	51

(Rerks-Ngarm et al, New Eng. J. of Medicine, 361, 2209 (2009))

Hypothesis Testing

Null hypothesis: VE = 0 %

P-value = 0.04. This means:

P(Data/Null) = 0.04

However, **this does not mean P(Null/Data) = 0.04**!

A Bayesian Approach: prior

User new evidence to update beliefs

Likelihood function

Prior probability

$$P(Model / Data) = \frac{P(Data / Model)P(Model)}{P(Data)}$$

Posterior probability

Model evidence (Independent of Model)

Numbers can be misleading....

Example: suppose a drug test is 99% sensitive and 99% specific.

(Namely, *P(+|User) = 0.99* and *P(+|Non user) = 0.01*)

Suppose that 0.5% of people are users of the drug.

If a random individual tests positive, what is the probability she is a user?

A Bayesian Approach

Bayes 's theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$P(User \mid +) = \frac{P(+ \mid User)P(User)}{P(+)} = \frac{P(+ \mid User)P(User)}{P(+ \mid User)P(User) + P(+ / NonUser)P(NonUser)}$$

$$P(User \mid +) = 33.2\%$$

Beware of lurking variables!

A real example from a medical study* comparing the success rates of two treatments of kidney stones:

	Treatment A	Treatment B
Patients	78% (273/350)	**83% (289/350)**

**Charig et al, Br Med J, 292, 879 (1986)*

Beware of lurking variables!

A real example from a medical study* comparing the success rates of two treatments of kidney stones:

	Treatment A	Treatment B
Small Stones	**93% (81/87)**	87% (234/270)
Large Stones	**73% (192/263)**	69% (55/80)
Patients	78% (273/350)	**83% (289/350)**

What is happening here?

**Charig et al, Br Med J, 292, 879 (1986)*