Analyze the Data

Human-Computer Interaction

Prof. Dr. habil. Ansgar Scherp

(Some slides based on Prof. Klaus Troitzsch)

C A U
Christian-Albrechts-Universität zu Kiel

ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Course on HCI: A Bird‘s Eye View

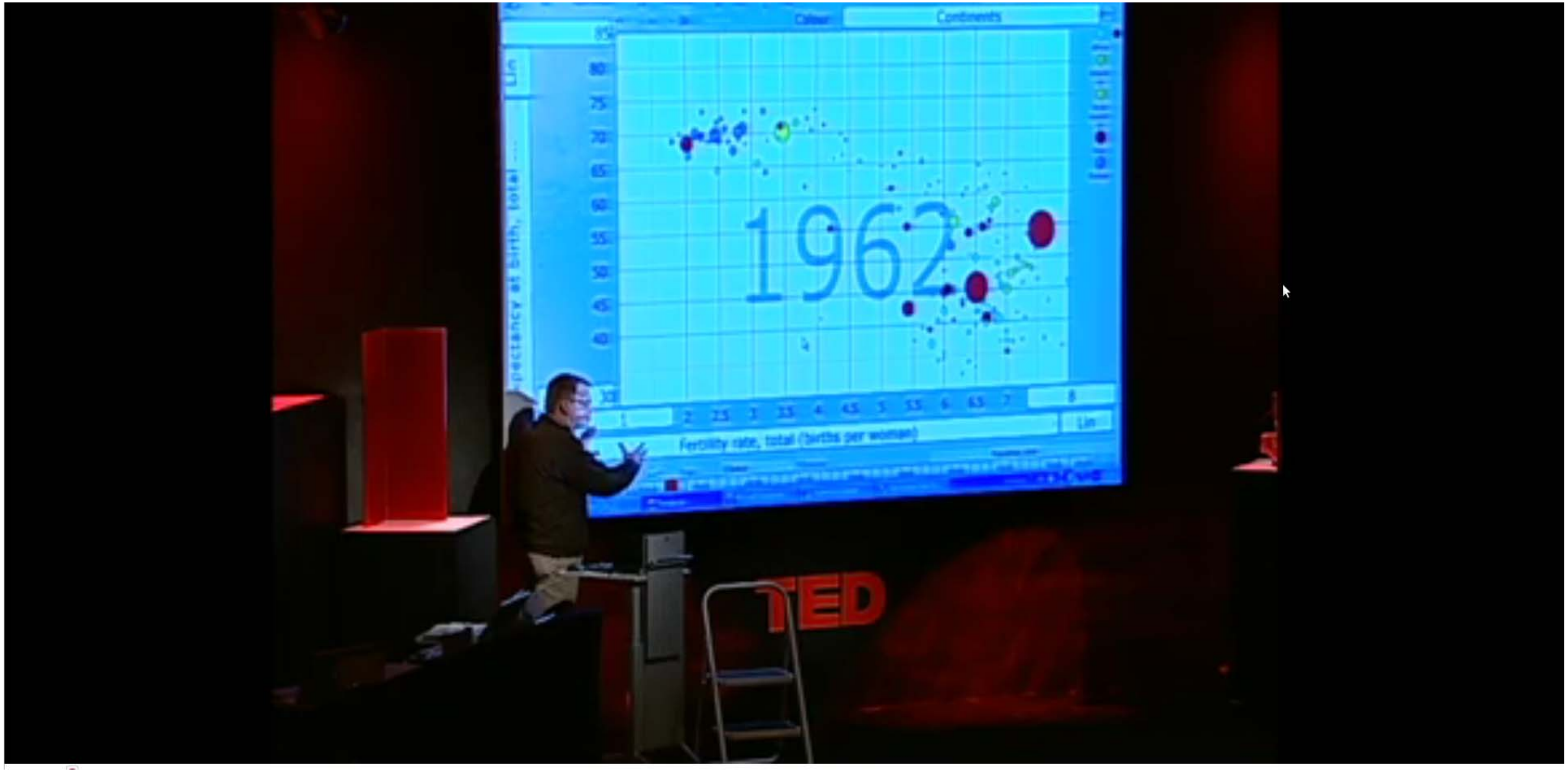
- History and Fundamentals
- Usability Testing vs. Experimental Research
- Identify a research hypothesis
- Specify the design of the study
- Run a pilot study
- Recruit participants
- Run the data collection sessions
- **Analyze the data**
- Report the results

Experimental HCI research (user studies)

Learning Goals

- Understand measures of central tendency and spread from descriptive statistics.
- Reflect on why inferential statistics is needed.
- What is a Type I and Type II error?
- How to estimate the number of participants needed in your experiment?
- Qualify in choosing and applying appropriate statistical methods for data analysis.
- Reflect on the difference between parametric and non-parametric significance tests.
- What is the effect size and confidence interval of your experiment outcome?

Best Stats You‘ve Ever Seen



- Hans Rosling, http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

Analyzing the Data

- Often equally or even more labor-intensive than collecting the data
- Not uncommon that researchers say
  - „Our data collection took 4 weeks, we need now several months to analyze the data.“

- Which statistical method to use?
- How to interpret significance test results?

Data Preparation

1. Step: Cleaning Up Data

- Data may have errors
- Check for reasonable entry
  - Mistyped age: “223” → probably: “23”
  - Compare “age” with “years of computing experience”
- Logging correct?
  - Documented performance rating unreasonably low or high
  - Multiple data about same participant grouped correctly (time, result, questionnaire data, …)

1. Step: Fixing Problems

- Contact participant and ask for correct age
- Remove problematic data item when participant cannot be contacted, e.g., online study
- Treat missing data item in statistical analysis, e.g., replace by mean value, use method that can deal with missing values, etc.

2. Step: Coding the Input Data

- Coding needs to be consistent
- Particularly challenging when coding is conducted by more than one coder

ID	Age	Gender	Highest degree	Experience in software A
P1	34	male	College	Yes
P2	28	female	Graduate	No
P3	21	female	High school	No

2. Step: Coding the Input Data

ID	Age	Gender	Highest degree	Experience in software A
P1	34	male	College	Yes
P2	28	female	Graduate	No
P3	21	female	High school	No

Coding

ID	Age	Gender	Highest degree	Experience in software A
P1	34	1	2	1
P2	28	0	3	0
P3	21	0	1	0

Levels of Measurement Revisited

- Dichotomous: two discrete values
- Nominal: multiple discrete but unranked values
- Ordinal: ranked values, unequal differences
- Interval: ranked values, equal differences
- Ratio: interval plus true zero

ID	Age	Gender	Highest degree	Experience in software A
P1	34	1	2	1
P2	28	0	3	0
P3	21	0	1	0

3. Step: Organizing Data

- Accommodate to requirements of statistical software

- Familiarize yourself with the analysis tools
  – IBM SPSS (available through the university)
  – PSPP (open source alternative to SPSS)
  – R and RStudio (open source)
  – Gnumeric (open source)
  – Compute analysis manually using OpenOffice, LibreOffice, Excel, …

4. Apply Statistical Analysis

- Statistics
  - Science of collecting, organizing, and analyzing data

I. Descriptive Statistics
  - Organizing and summarizing data

II. Inferential Statistics
  - Using data to form conclusions

Population and Sampling

- Sample (“small” n)
  - Subset of the population
  - Our representative group of participants
- Population (“large” N)
  - Entire group you are interested in studying

Only in rare cases the values of an attribute for an entire population are known, e.g., unemployment statistics.
But what about the height and music taste of 18-year old boys?

Descriptive Statistics

I. Descriptive Statistics

a) Measures of central tendency
   - Mode
   - Median
   - Mean

b) Measures of spread / dispersion
   - Range
   - Mean deviation
   - Variance
   - Standard deviation
   - Standard error

a) Measures of Central Tendency

- “central tendency” = where the bulk of data is
- Find single score that best represents the sample

- Mode
  - Value with the highest frequency
  - Applicable to all levels of measurement
- Example
  - {12,15,24,24,<u>25</u>,<u>25</u>,<u>25</u>,40,45} → Mode: 25

a) Measures of Central Tendency

- Disadvantages of mode
  - Dataset can have two (bimodal) or several modes (multimodal)
  - This makes it messy to summarize data
  - Mode can change dramatically
- Example: Feature in application is rated
  - 4 x “poor” and 5 x “good” → Mode: “good”
  - Add 1 “poor” sample → Mode: “poor”, “good”
  - Another “poor” sample → Mode: “poor”

a) Measures of Central Tendency

- Median
  - Middle score in a data set of ranked values
  - When even number of values, average middle scores

$$\tilde{x} = \begin{cases} x_i, i = \dfrac{n+1}{2} & \text{if } n \text{ odd} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \text{if } n \text{ even} \end{cases}, x_i \le x_{i+1}, i = 1 \ldots n$$

a) Measures of Central Tendency

- Examples
  - {12,14,19,<u>22</u>,27,35,40} → $\tilde{x} = 22$
  - {7,12,<u>17</u>,<u>23</u>,37,42} → $\tilde{x} = (17 + 23)/2 = 20$

$$\tilde{x} = \begin{cases} x_i, i = \dfrac{n+1}{2} & \text{if } n \text{ odd} \\ \dfrac{x_{n/2} + x_{(n/2)+1}}{2} & \text{if } n \text{ even} \end{cases}, x_i \leq x_{i+1}, i = 1 \ldots n$$

a) Measures of Central Tendency

- Mean (arithmetic average, short: AVG, M)
  - Sum of all scores, divided by # of scores
  - Defined for interval and ratio variables

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

$\bar{X}$ in the context of modeling a random variable

- Example
  - {12,14,19,22,27,35,40} → $\bar{x} = 24.14$

Some Characteristics

- Median (and mode) is relatively unaffected by extreme scores (*outliers*) compared to mean
- Median is also relatively unaffected by a lot of scores at one end of scale (*skewness*)
- Median is mathematically not very useful (ignores most scores in the data set)
- Variability of mode and median is high
- Mean is a value not necessarily been observed in the sample
- Mean is best representing the data set !

b) Measures of Spread

- Range
  - Difference of minimum and maximum score
  - Applicable on interval and ratio
  - Trivial measure of spread

$$R = x_{\max} - x_{\min}$$

- Example
  - {12,14,19,22,27,35,40} → R=40-12=28

b) Measures of Spread

- Mean deviation (= average deviation, short: AD)
  - Mean of all absolute differences with respect to arithmetic mean
  - Only seldom used

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

- Defined on interval and ratio variables

b) Measures of Spread

- Variance (short: VAR)
  - Mean of all squared distances with respect to arithmetic mean

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

- Standard deviation (short: SD)
  - Square root of variance
  - Most common measure of deviation
  - Sum of squared error (SSE)

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

b) Measures of Spread

- Standard Error (short: SE)
  - Standard deviation of a sampling distribution of a statistic
  - Most commonly the mean

- Standard error of the mean

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = s \cdot \sqrt{\frac{1}{n}}$$

Excursion: Standard Error

- Mean over a set of n scores: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Consider the sampling variable X_i (a random variable)
  - Its distribution describes the likelihood that a specific value is observed at the i-th sampling
  - x_i is the value of X_i
- Then we can define the ***estimator***

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

- for independent random variables $X_1, \dots, X_n$ with identical distribution and variance σ^2

Excursion: Standard Error (cont'd)

- The standard error of $\bar{X}$ is then defined as

$$\sigma(\bar{X}) = \sqrt{VAR(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

Standard deviation of the population

- With

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

$$= \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

See definition (previous slide)

Degrees of Freedom (df)

- So far, we computed VAR / SD of the sample n
- To make inferences to the entire population
  - Use n-1 instead of n
  - Otherwise we would fix the VAR / SD and do not leave room for the sample mean

- Mathematically: when mean of a sample is known to be 4 and we have 10 scores, than one can freely choose 9 scores

Introduction to: Inferential Statistics

II. Inferential Statistics

- Significance tests
- Hypothesis testing
- Type I and II error

- Parametric tests
- Non-parametric tests

Significance Tests: Why?

- Naïve approach
  - Compare two mean values of error rate for sample n_1 and n_2 to detect difference
- Example
  - Mike's height is 186 cm, Mary's is 174 cm.
    → So Mike is taller than Mary.
  - Average height of 3 males is 165 cm and average height of 3 females is 180 cm.
    → So females are taller than males.

Significance Tests: Why? (cont.)

- First statement is OK
- What is wrong with the second statement?
  - Common sense
  - Selected sample is not representative
  - Falsification: find 3 other males and females, where average height is other way round
  - Sizes of comparison groups too small
  - Inappropriate sampling?

Significance Tests: Why? (cont.)

- Groups can be directly compared, when
  - All values of all group members are known
  - There is no uncertainty involved

- When N is large, we can only sample $n \ll N$
- Significance test needed to determine if results of sampling population n can be generalized to the entire population N

!

Hypothesis Testing

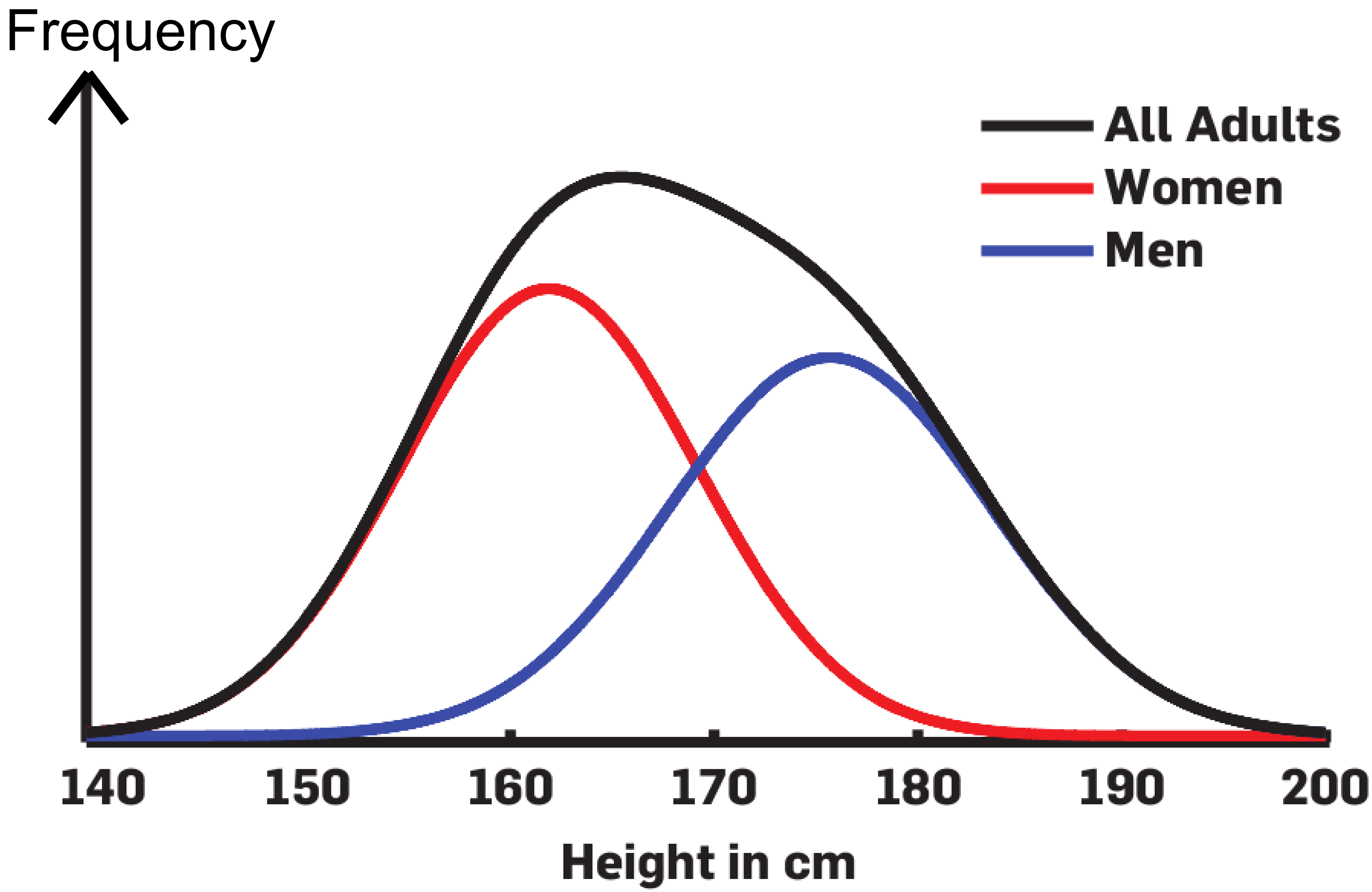
- H_0 is contrasted with H_1 to determine likelihood that H_0 is true
- In our example
  - H_0: There is no difference in the average height of male and female.

- We chose a significance level of $\alpha = 5\%$
- If test shows significance, we can conclude that in 95% of the time, the test result correctly applies to the entire population

Normal Distribution*(s)*

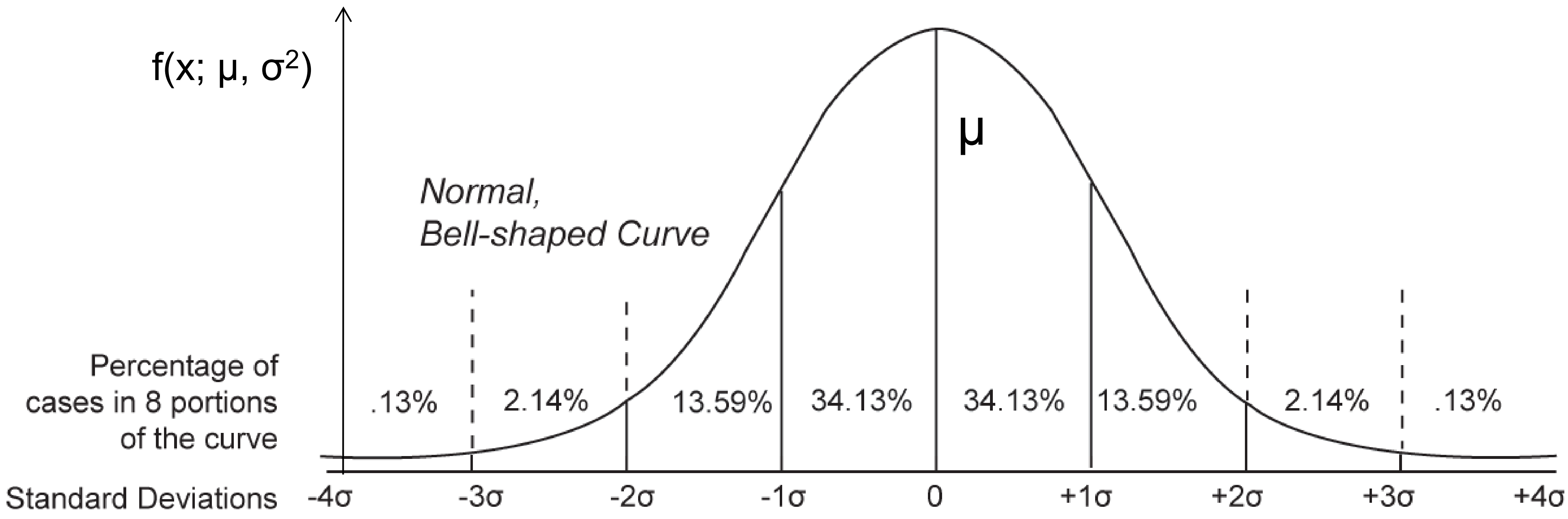
- Observation
  - Majority of measures x_i fall in the area around the mean of the population
- For example, distribution of
  - Heights of a population
  - Student grades
  - Etc.

Normal Distribution: Examples



Normal Distribution
(Gauss-distribution)

μ = mean of population
σ = standard deviation of population
$f(x; \mu, \sigma^2)$ = distribution function

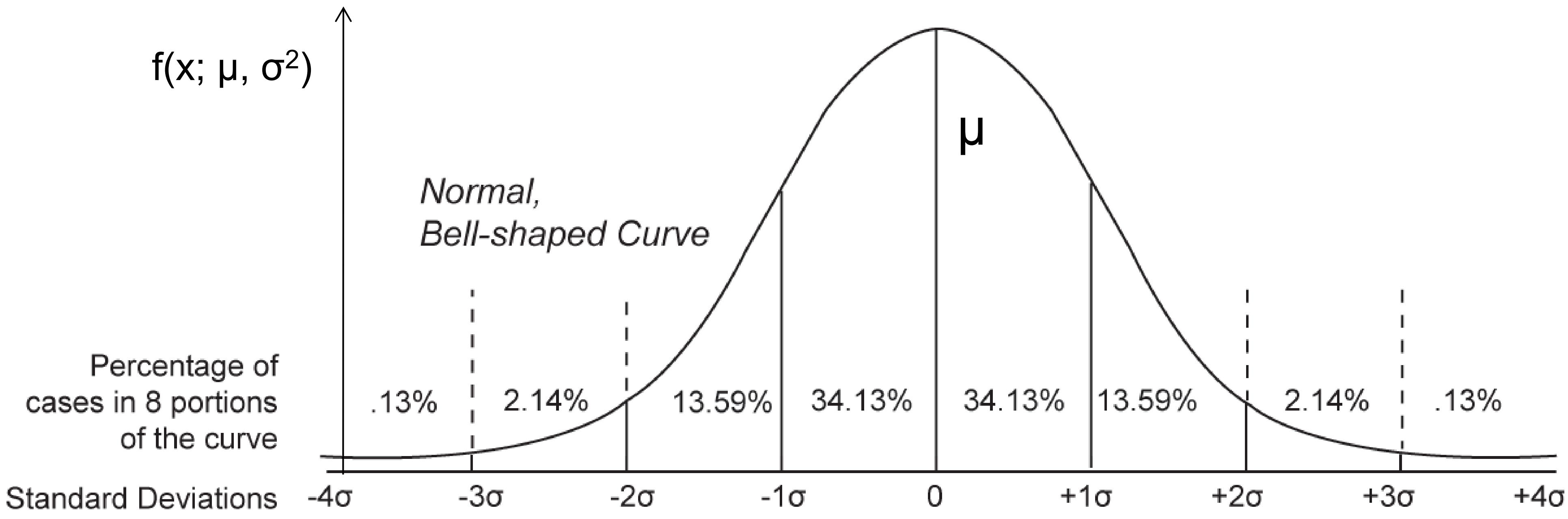


$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Diagram source: Wikipedia

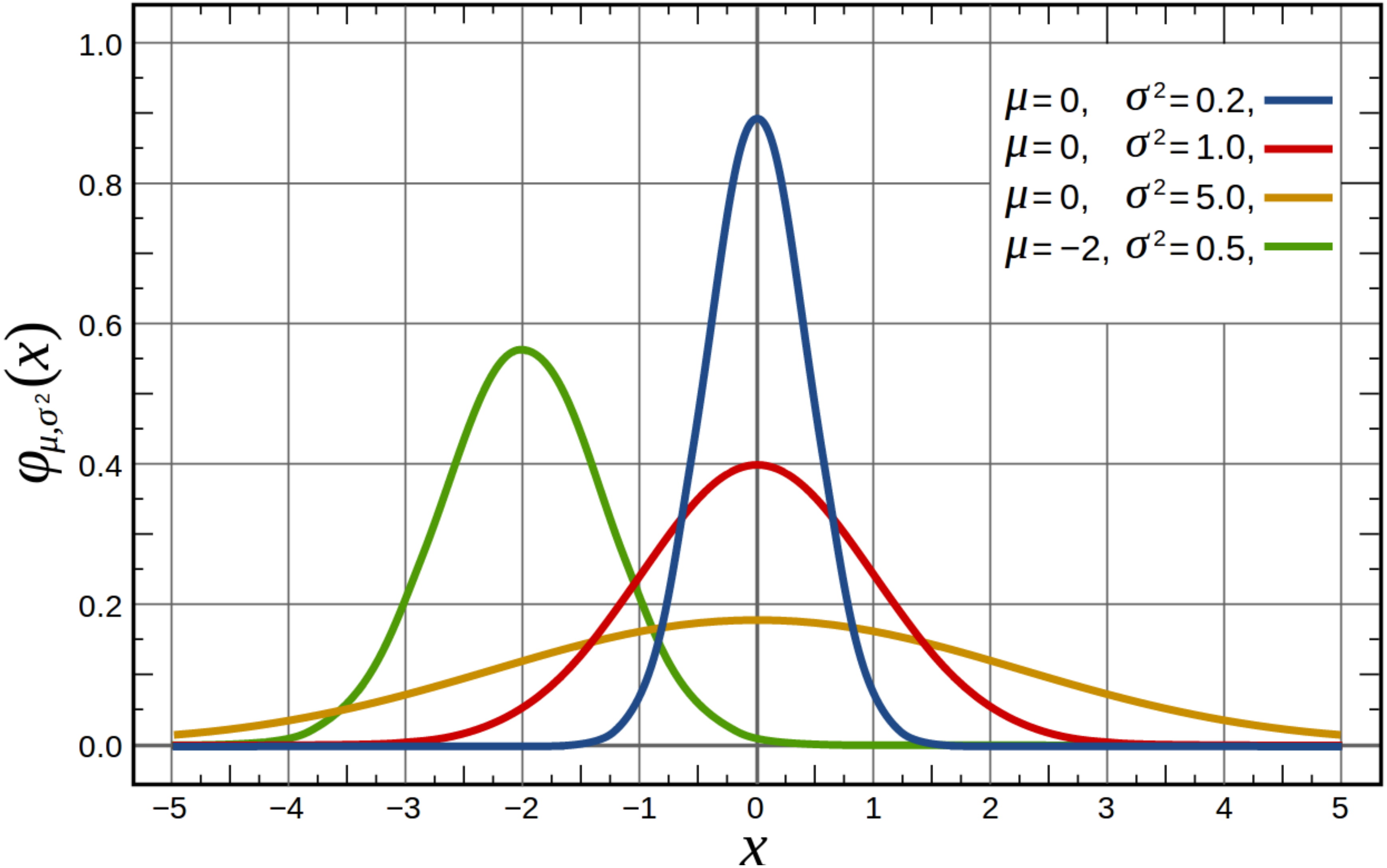
Normal Distribution

μ = mean of population
σ = standard deviation of population
$f(x; \mu, \sigma^2)$ = distribution function



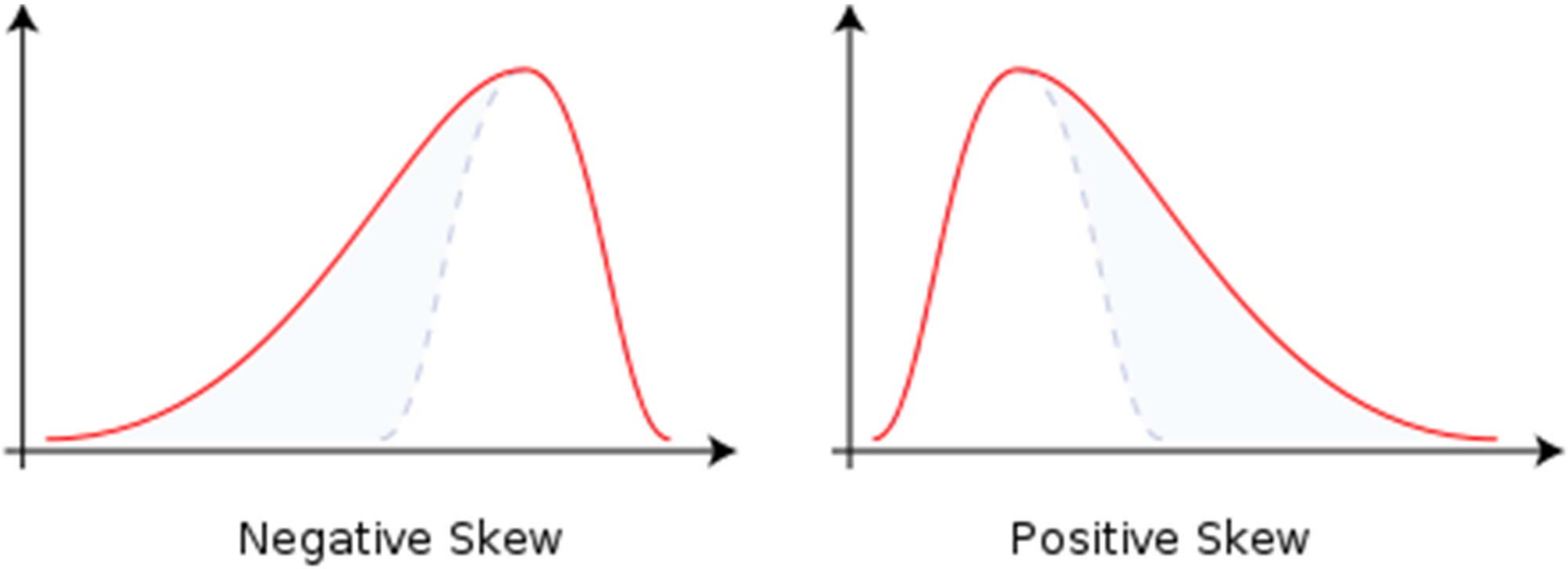
- Empirical rule
  - ~68% of the values lie within $[\mu-1\sigma, \mu+1\sigma]$
  - ~95% of the values lie within $[\mu-2\sigma, \mu+2\sigma]$
  - ~99.7% of the values lie within $[\mu-3\sigma, \mu+3\sigma]$

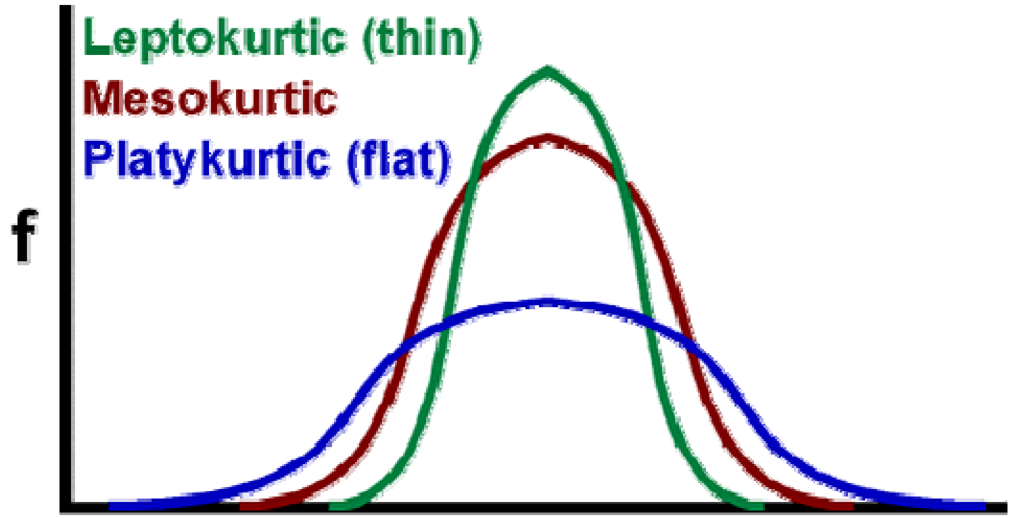
Normal Distribution: Examples



- http://upload.wikimedia.org/wikipedia/commons/7/74/Normal_Distribution_PDF.svg

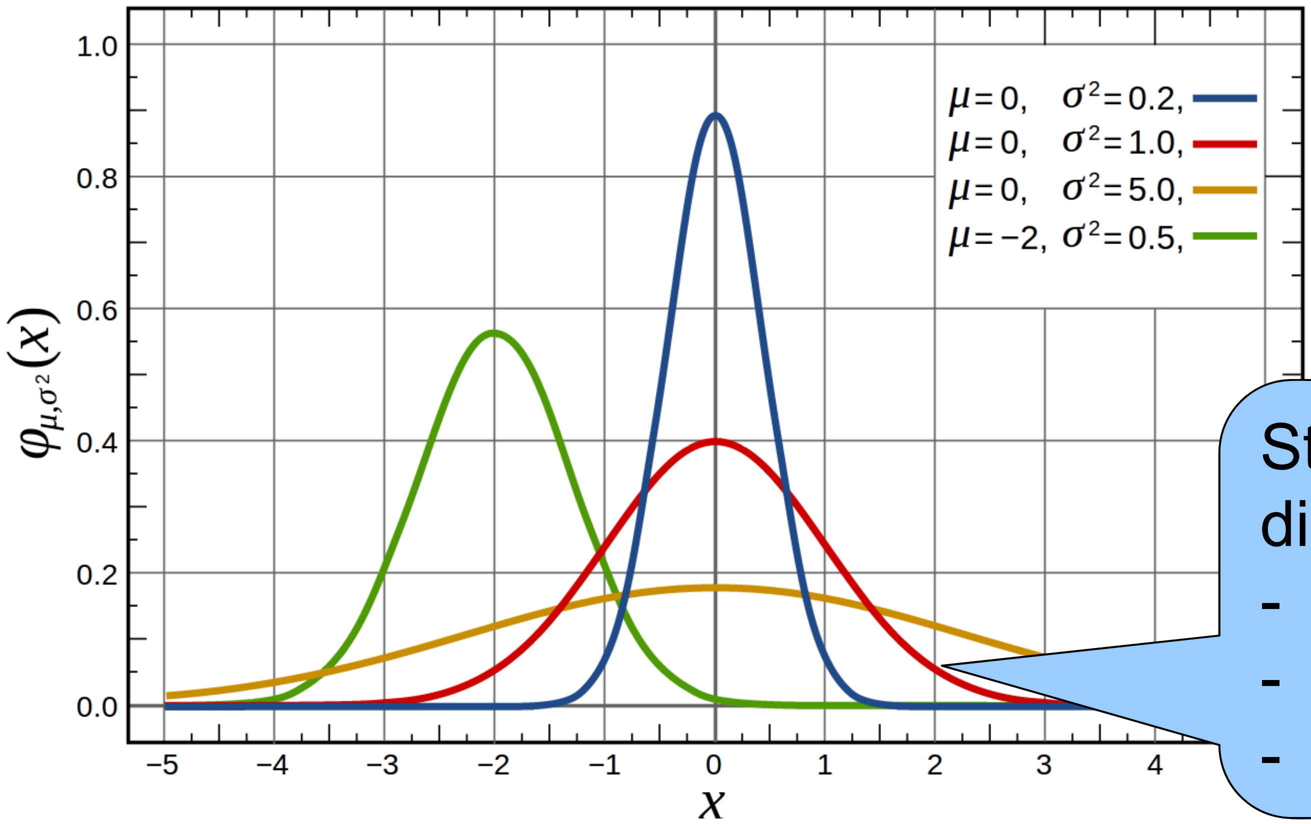
Skewness and Kurtosis





Standard Normal Distribution

- Special case of the normal distribution
- Distribution when a normal random variable has a mean of zero and a standard deviation of one



$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Standard normal distribution aka
- standard score
- z-score
- normal score

Standard Normal Distribution

- Any normal variable X can be transformed to a Z score
- Calculate standard score z from a "raw" score x with

$$z = \frac{x - \mu}{\sigma}$$

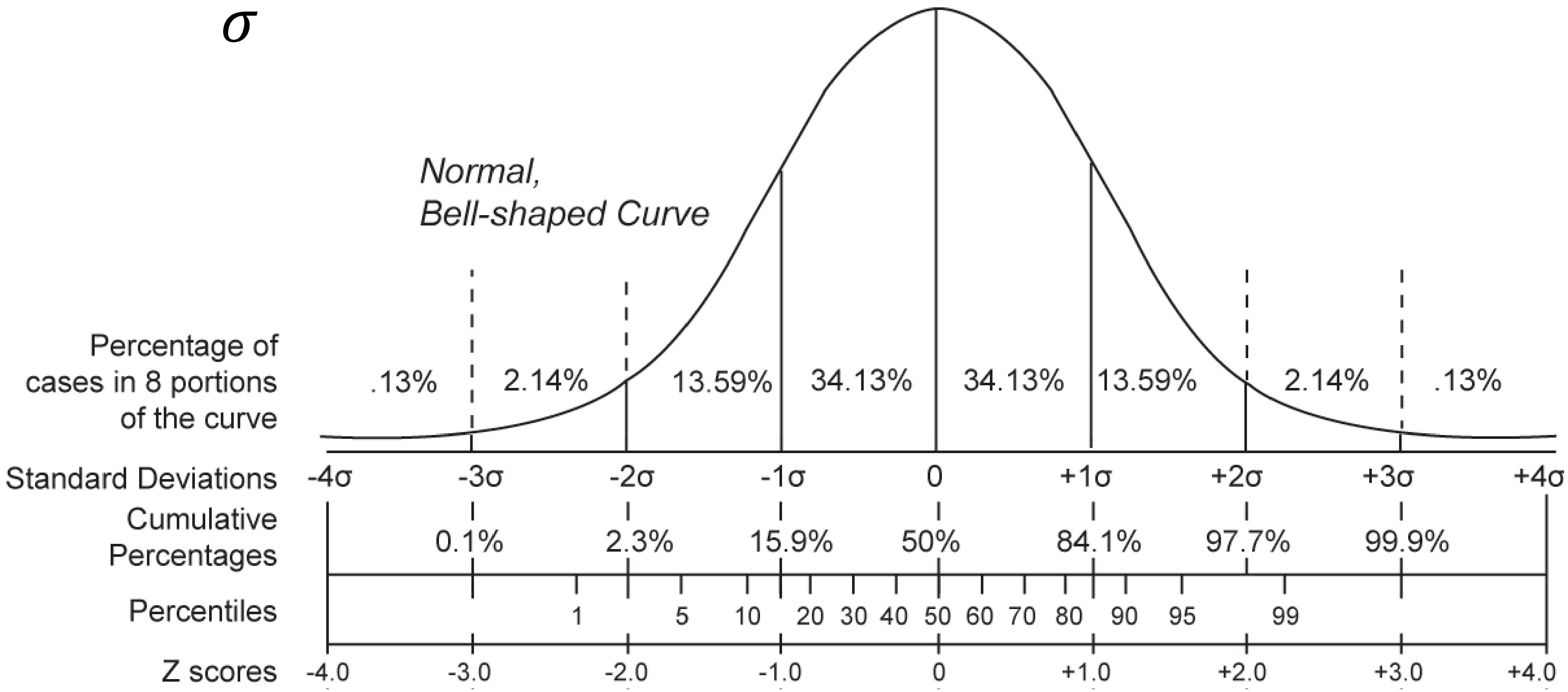
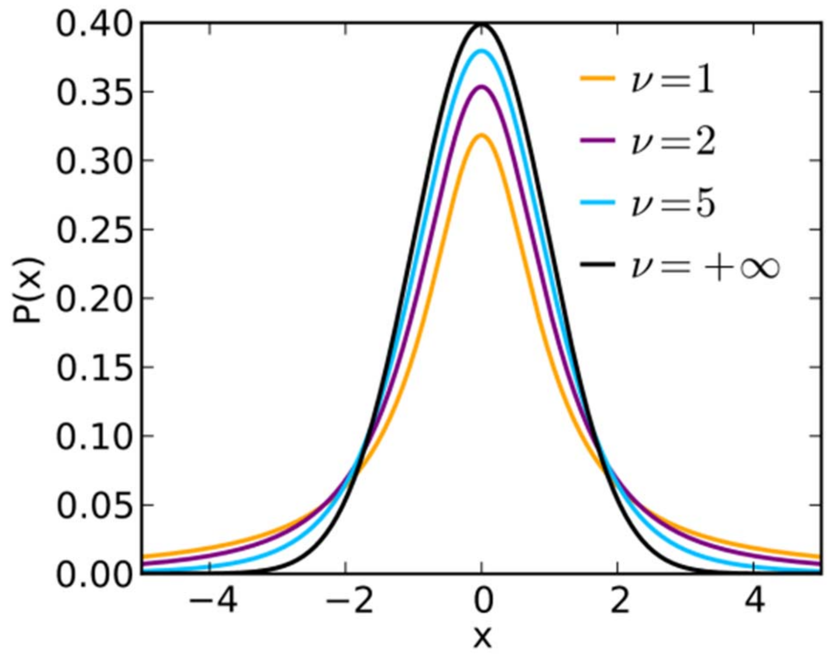


Diagram source: Wikipedia

Student's t-distribution

- ***Estimator*** for the mean of normally distributed data
  - Does not follow a normal distribution if sample size is small
  - And: true VAR / SD of population is unknown and has to be estimated from the sample



- Rather: estimator follows the so-called t-distribution

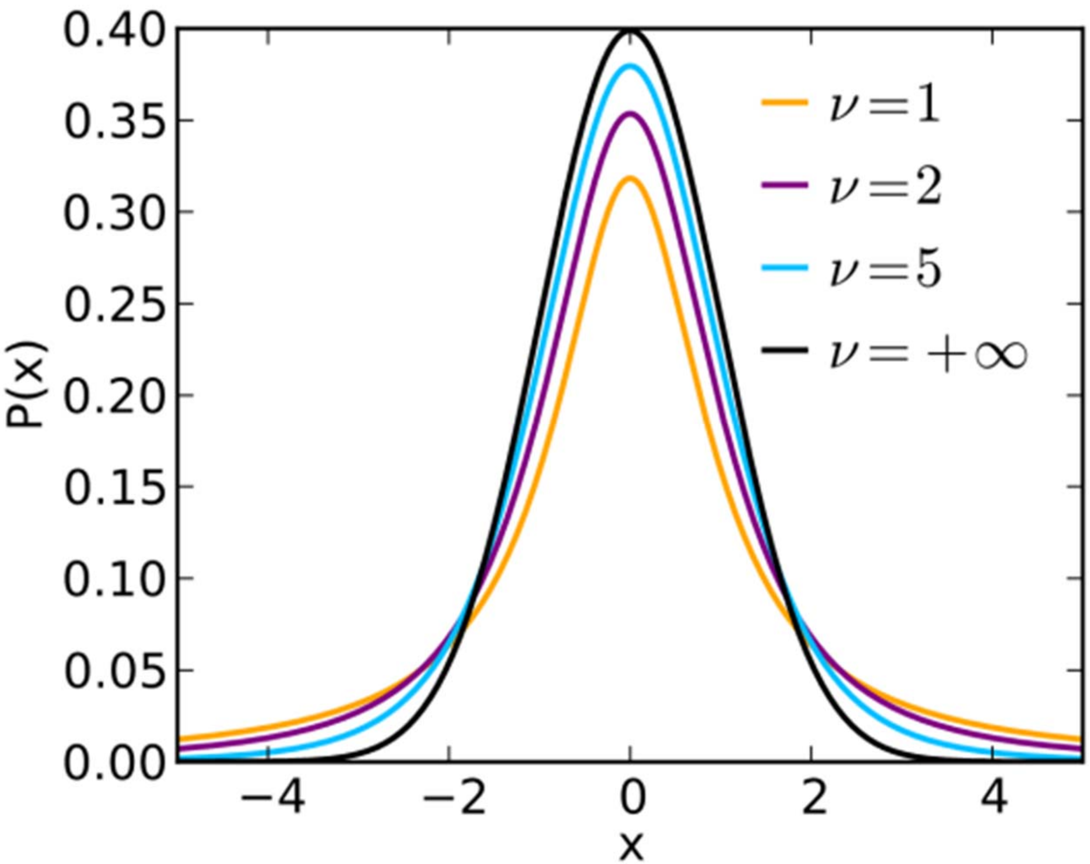
- Normal distribution describes the entire population
- t-distribution describes the distribution of samples drawn from the population

!

Student's t-distribution (cont'd)

- t-distribution: describes the distribution of samples drawn from the population
- Thus, t-distribution is different for each sample size **!**

- $\nu = \mathrm{n} - 1$, i.e., the degree of freedom of the sample size (ν the letter after μ ;-)
- Larger n → slimmer the distributions
- $n \to \infty$: t-distribution converges to the normal distribution



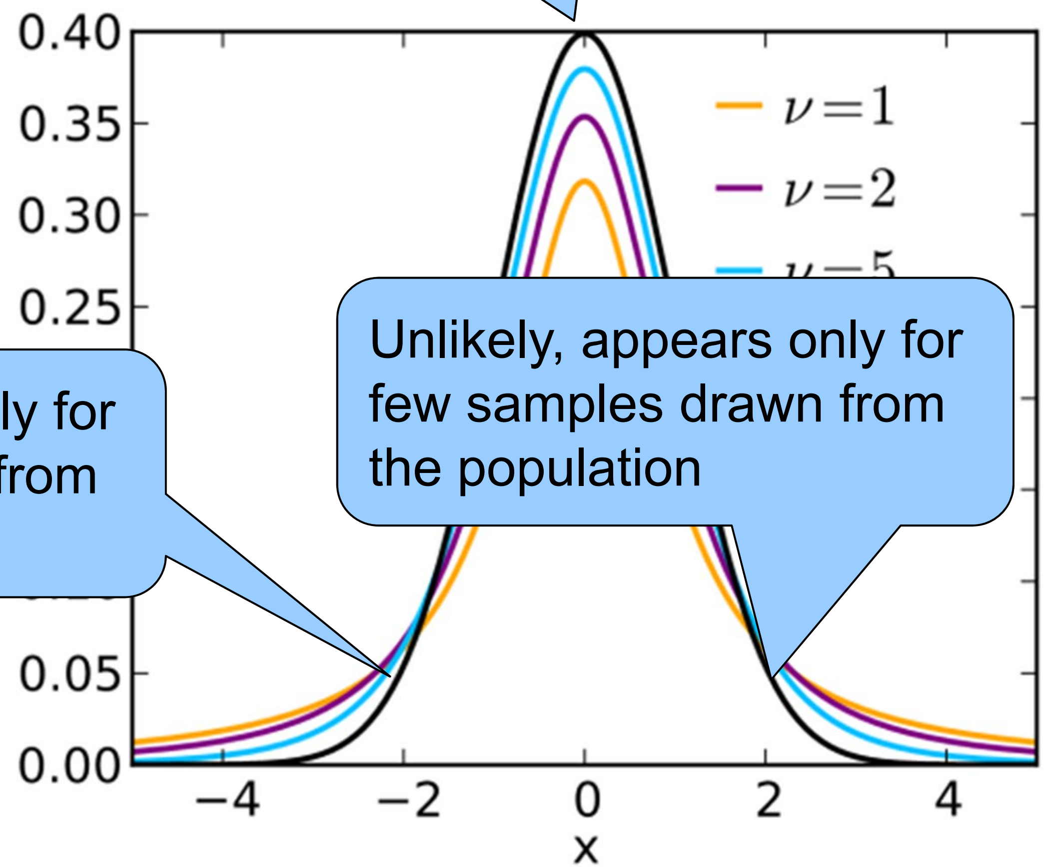
Student's t-distribution (cont'd)

- t-distribution: describes the distri
  drawn from the population
- Thus, t-distribution is different for ea ample size

Most likely, will appear for most samples drawn from the population

!

- $\nu = \mathrm{n} - 1$, i.e., the degree of freedom of the sample size (ν the letter after μ :-)
- Larger n
  the distr
- $n \rightarrow \infty$: t-distribution converges to the normal distribution



Unlikely, appears only for few samples drawn from the population

Unlikely, appears only for few samples drawn from the population

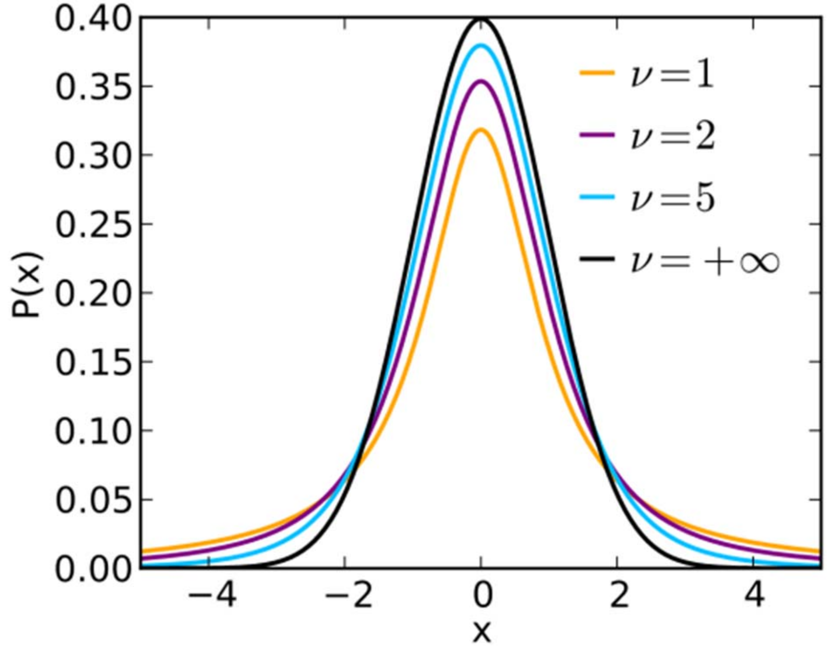
Formally: Student's t-distribution

$$f(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Also: shifted via parameter μ

with Gamma function

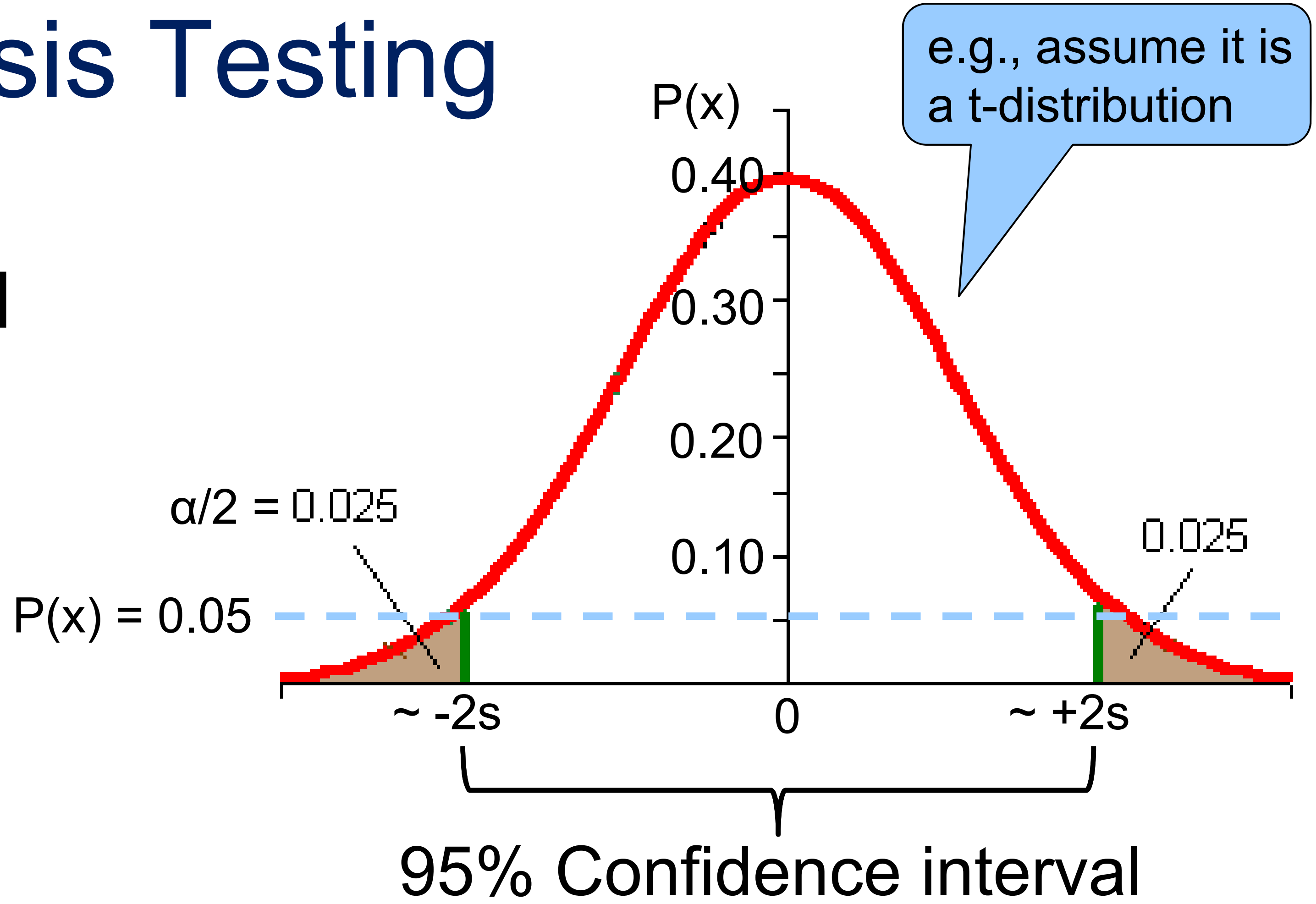
$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\,\mathrm{d}t$$



- t-distribution: used to compute the difference from the mean of the sample to the true mean of the population
- Hypotheses tests using the t-distribution: t-tests

Hypothesis Testing

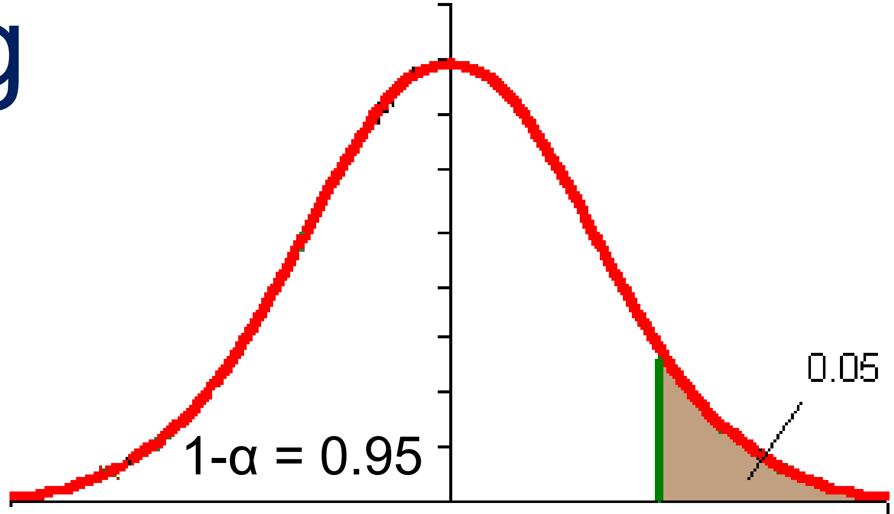
- α = .05
- Two-tailed



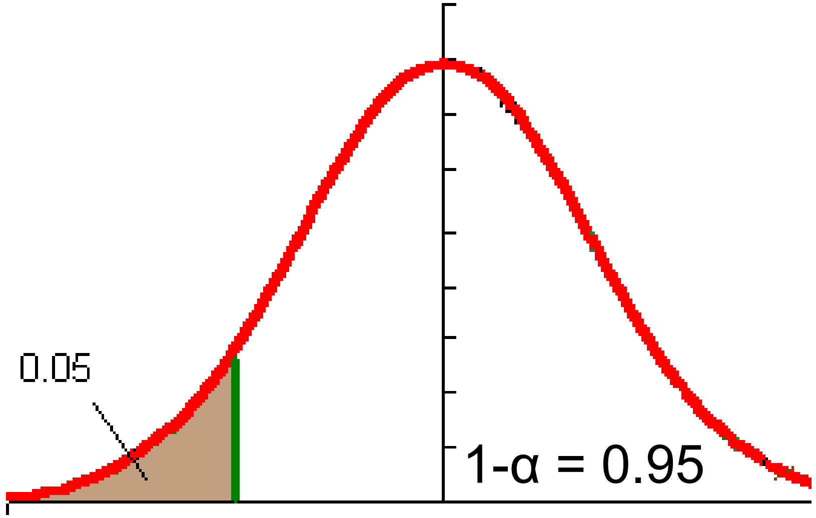
- Test statistics falls in $\sim[-2s, +2s]$ → we keep H_0
- Otherwise: Likely that H_0 is not true; accept H_1

Hypothesis Testing

- α = .05
- One-tailed (right)



- One-tailed (left)



- Testing if means is higher or lower

Type I and Type II Errors

- Significance tests are subject to the risk of error

- Type I error (“false positive”, α error)
  - Rejecting H_0 although it is true and should not be rejected
- Type II error (“false negative”, β error)
  - Not rejecting H_0 when it is false and should be rejected (i.e., H_1 is true)

- Commonly accepted: α=5%, β=20% **!**

Type I and Type II Errors

- Type I errors are generally believed to be worse than Type II errors **!**

- Example: judicial case

		Jury decision	
		Not guilty	Guilty
Reality	Not guilty	✓	Type I error
	Guilty	Type II error	✓

- Example: pregnancy test

Type I and Type II Errors

- Example
  - Comparing user interface designs for ATMs

		Study conclusion	
		No difference	Touchscreen ATM is easier to use
Reality	No difference	√	Type I error
	Touchscreen ATM is easier to use	Type II error	√

Type I and Type II Errors

- Example
  - Comparing user interface designs for ATMs

		Study conclusion	
		H_0 No difference	H_1 Touchscreen ATM is easier to use
Reality	H_0 No difference	√	Type I error
	H_1 Touchscreen ATM is easier to use	Type II error	√

Correct inference: True negative

False positive

False negative

Correct inference: True positive

Type I and Type II Errors

- Type I errors are mistakes of "gullibility".
  - May result in a condition worse than the current state

- Type II errors are mistakes of "blindness"
  - Can cost the opportunity to improve the current state

Controlling Risks of Errors

- Probability of making a Type I error: α (also called significance level)
- Probability of making a Type II error: β

Aka: statistical power, sensitivity

- Power of a test is $1-\beta$
  - Probability of successfully rejecting H_0 when it is false and should be rejected
  - Or: probability that the test can detect a treatment effect, if it is present

!

Controlling Risks of Errors (cont.)

- Reducing Type II error
  - Generally suggested to have a relatively large sample size
  - So that difference can be observed, even when the effect size is relatively small
- Rationale: increased sample size boosts the statistical power of a test (and it is fairly easy to add more participants to your experiment …)

Controlling Risks of Errors

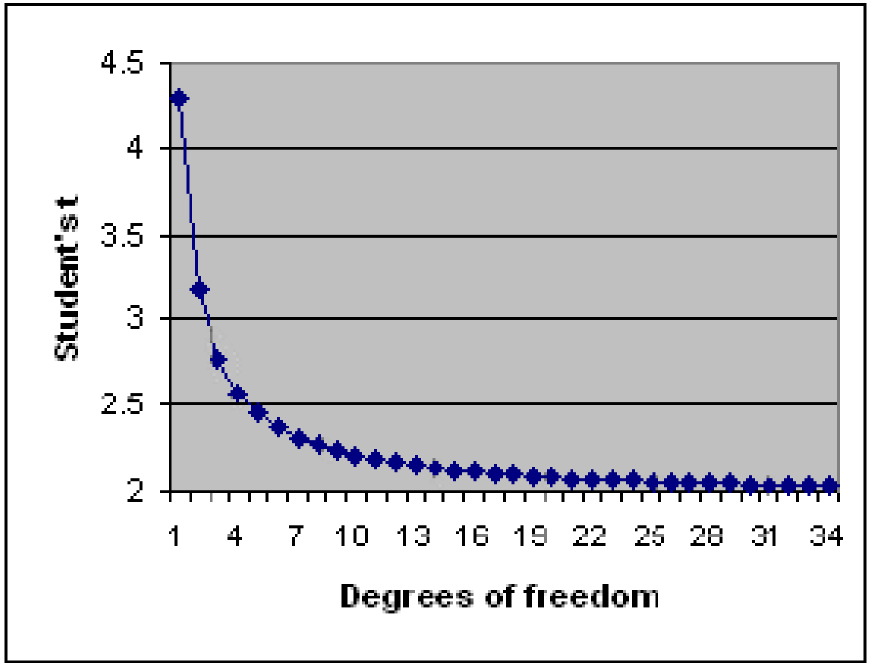
- α and β are interrelated
  - Under the same conditions, decreasing α reduces the chance of making Type I errors
  - But increases chance of making Type II errors

- A very low p value ($p < .05$) is widely adopted to control the occurrence of Type I errors **!**

Estimating the Number of Participants

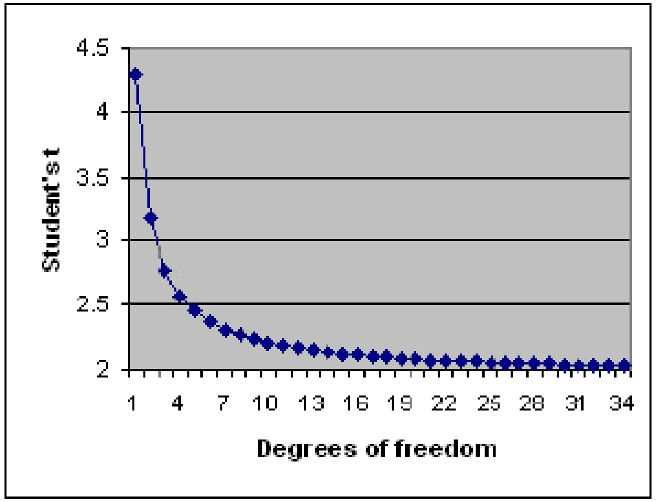
Mead‘s Resource Equation (Revisited)

- Law of diminishing returns
  - Adding one more animal to a small experiment gives good returns in terms of increased power
  - However, doing the same to a large experiment will hardly increase the power at all
- Plot shows the 5% critical value of the t-test plotted against the df
- Mead suggests: design experiments to give a good estimate of error, but should not be so big that they waste resources



Mead's Resource Equation (1988)

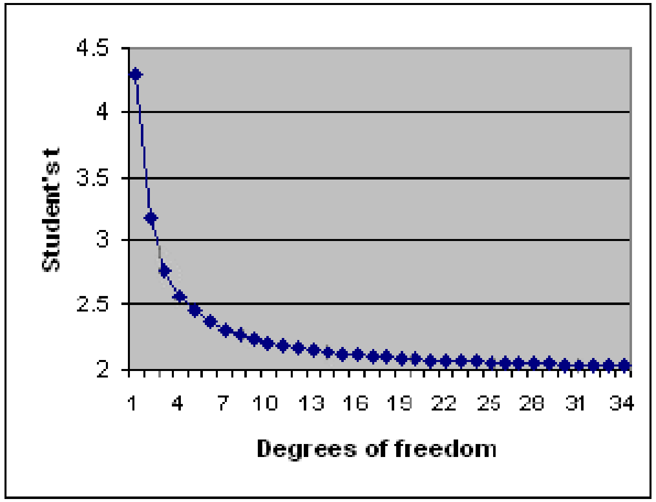
- Easy to use equation
  - $E = N - B - T$
- Error degrees of freedom E should be between 10-20



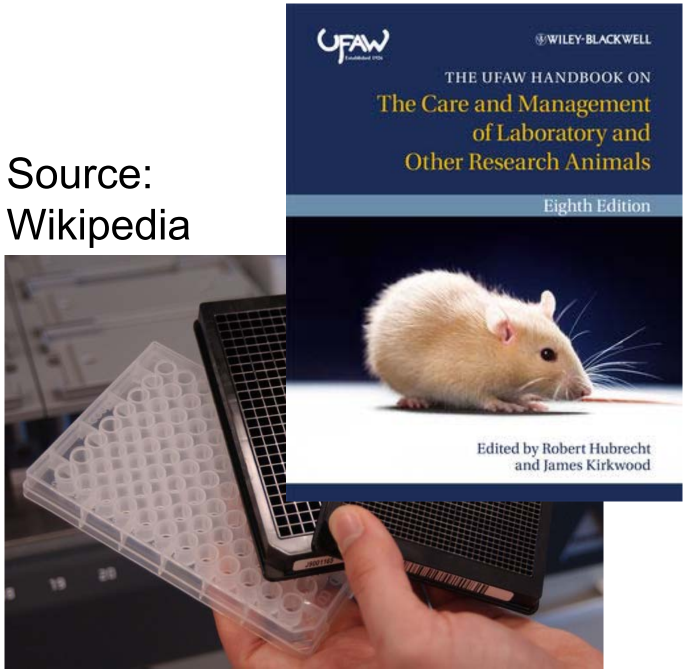
- Treatment component, T, corresponding to the questions being asked (number of conditions)
- Blocking component B, representing environmental effects allowed for in the design (e.g., special diet)
- Error component E, being used to estimate the variance, s^2 which is used for calculating the standard errors for treatment comparisons

Mead's Resource Equation (1988)

- Easy to use equation
  - $E = N - B - T$
- Error degrees of freedom E should be between 10-20



Source: Wikipedia



- But: limits should not be applied too rigidly
- Good arguments can be to make E being 25-30 or more to ensure equal group sizes
- Can even go higher if experimental units are "cheap"

Power Analysis

- Determine how many participants one needs for a specific "power" of the significance test
- Should be preferred over Mead's approach
- But more difficult when, e.g.,
  - More than two treatments
  - Preconditions need to be fulfilled, e.g., t-test
- Power of a test is: $\Pi = 1 - \beta$
  - Probability of successfully rejecting H_0 when it is false and should be rejected
  - "1 minus probability of making a Type II error"

Power Analysis (Continued)

- Power of a test is: $\Pi = 1 - \beta$
- Power between .80 and .90 is commonly accepted

!

Inacceptable in medical tests, e.g., cancer tests!!! Why?

- Power of .80 is known as "4-1 rule" (i.e., when $\beta = .20$)
  - 20% probability of Type II error
  - 80% probability of successfully rejecting H_0 when it is false and should be rejected

Power Analysis (Continued)

- Power of (two-group) t test depends on
  a) Difference between means (effect size)
  b) Residual variance (variance in the scores)
  c) Sample size

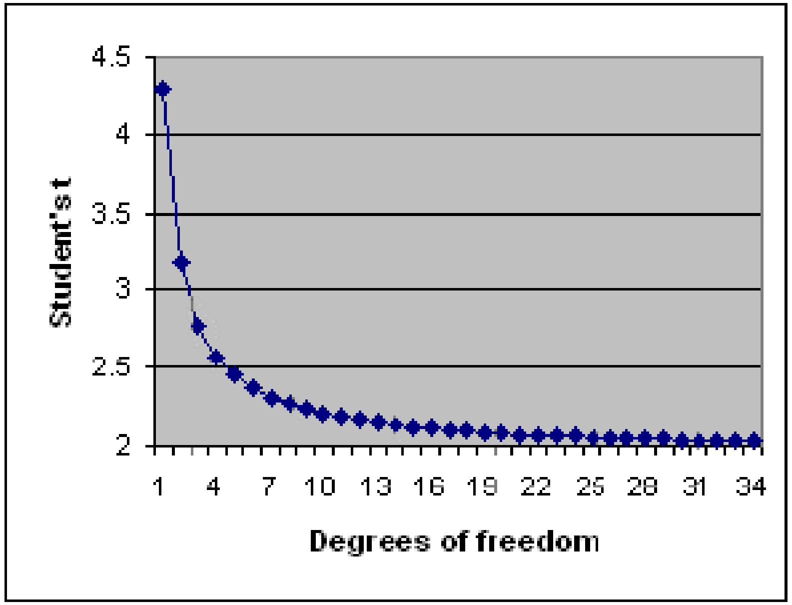
Your opinion on the suggestions for a) and b)?

- For enhancing the power you may choose ...
  a) more “extreme” treatments
  b) participants who are “relatively similar” (e.g., in animal research: use same strain)
  c) add more participants to increase the df

Power Analysis (Continued)

- Regarding: c) add more participants

- Increase of df of the t test requires less difference between means to become significant



- Reduces standard error (SE) because of the factor $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

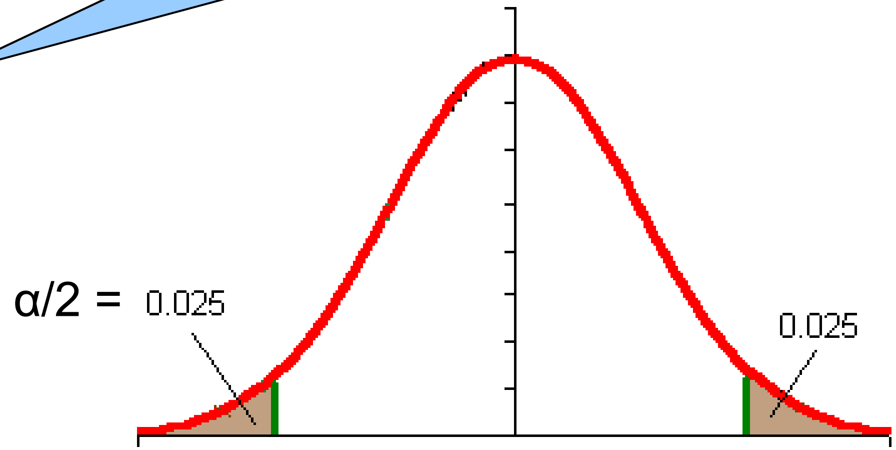
- If sample size is fixed: SE of mean difference is reduced by having n_1 equal n_2 because of pooled variance to estimate σ

We will see t-test in detail …

Determine Power Π

Magnitude of difference between the means of two groups

- *Input*: given values for
  - Cohen's d (effect size)
  - Sample size n
  - Significance level α
- *Output*: statistical power Π



- Where Cohen's d is estimated from the sample:
  - Small effect: $d \geq .2$
  - Medium effect: $d \geq .5$
  - Large effect: $d \geq .8$

$$d = t \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Stay tuned …

Determine Power Π: Power Tables

- Two-group t test with $\alpha = .05$ and $n = n_1 = n_2$
- n is sample size in each group, so total study is $2n$

n	$d = .2$	$d = .5$	$d = .8$
10	.07	.18	.39
20	.09	.33	.69
40	.14	.60	.94
80	.24	.88	.99
100	.29	.94	.99
200	.51	.99	.99

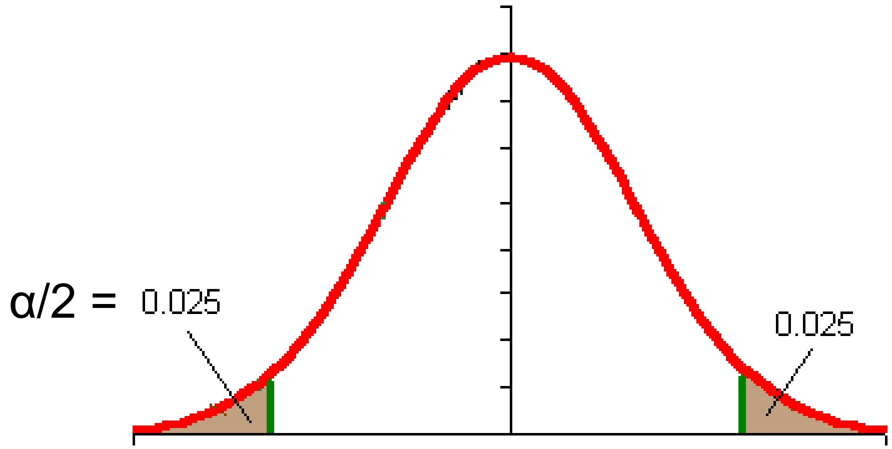
e.g., for a retrospective computation of power Π

- Example: if $d = .5$ and there are $n = 20$ participants in each group, than chances of rejecting H_0 is $\Pi = .33$

Source: David A Kenny: Statistics for the social and behavioral sciences, 1987

Determine Power Π

- *Input*: given values for
  - Cohen's d (effect size)
  - Significance level α
  - Desired power Π
- *Output*: sample size n



- Where Cohen's d is estimated from the sample:
  - Small effect: $d \geq .2$
  - Medium effect: $d \geq .5$
  - Large effect: $d \geq .8$

$$d = t \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Stay tuned …

Determine Power Π: Power Tables

- Two-group t test with $\alpha = .05$ and $n = n_1 = n_2$
- n is sample size in each group, so total study is $2n$

Π	$d = .2$	$d = .5$	$d = .8$
.25	84	14	6
.50	193	32	13
.60	246	40	16
.70	310	50	20
.80	393	64	26
.90	526	85	34
.95	651	105	42
.99	920	148	58

Apply for estimating the sample size you need

!

Example: if $d = .5$ and desired power $\Pi = .80$, one would need $n = 64$ participants *per group*

Do you make any interesting observations?

Source: David A Kenny: Statistics for the social and behavioral sciences, 1987

Determine Power Π: Power Tables

- Two-group t test with $\alpha = .05$ and $n = n_1 = n_2$
- n is sample size in each group, so total study is $2n$

Π	$d = .2$	$d = .5$	$d = .8$
.25	84	14	6
.50	193	32	13
.60	246	40	16
.70	310	50	20
.80	393	64	26
.90	526	85	34
.95	651	105	42
.99	920	148	58

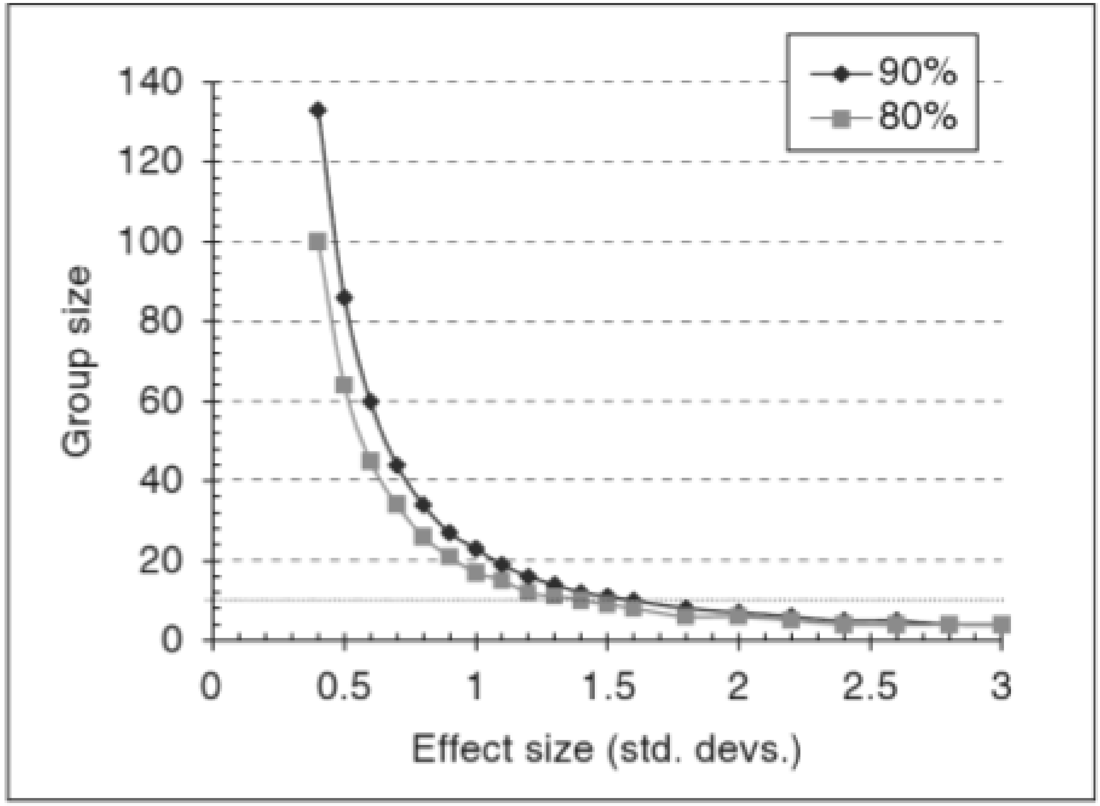
Observation 1:
Larger power requires more participants

Observation 2:
The larger the effect, you need to be able to detect it with *fewer* participants **!**

Source: David A Kenny: Statistics for the social and behavioral sciences, 1987

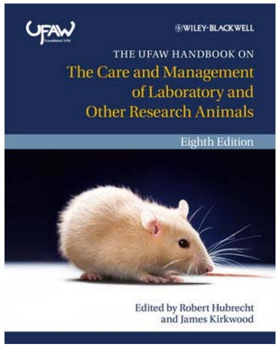
Observation 2 Continued

- Two-group t test with $\alpha = .05$ and power $\Pi = .80/.90$
- Estimated sample size per group can be read from the graph as function of the effect size in standard deviation



- Note: it is useful to express the effect size in SD since it puts everything on the same units

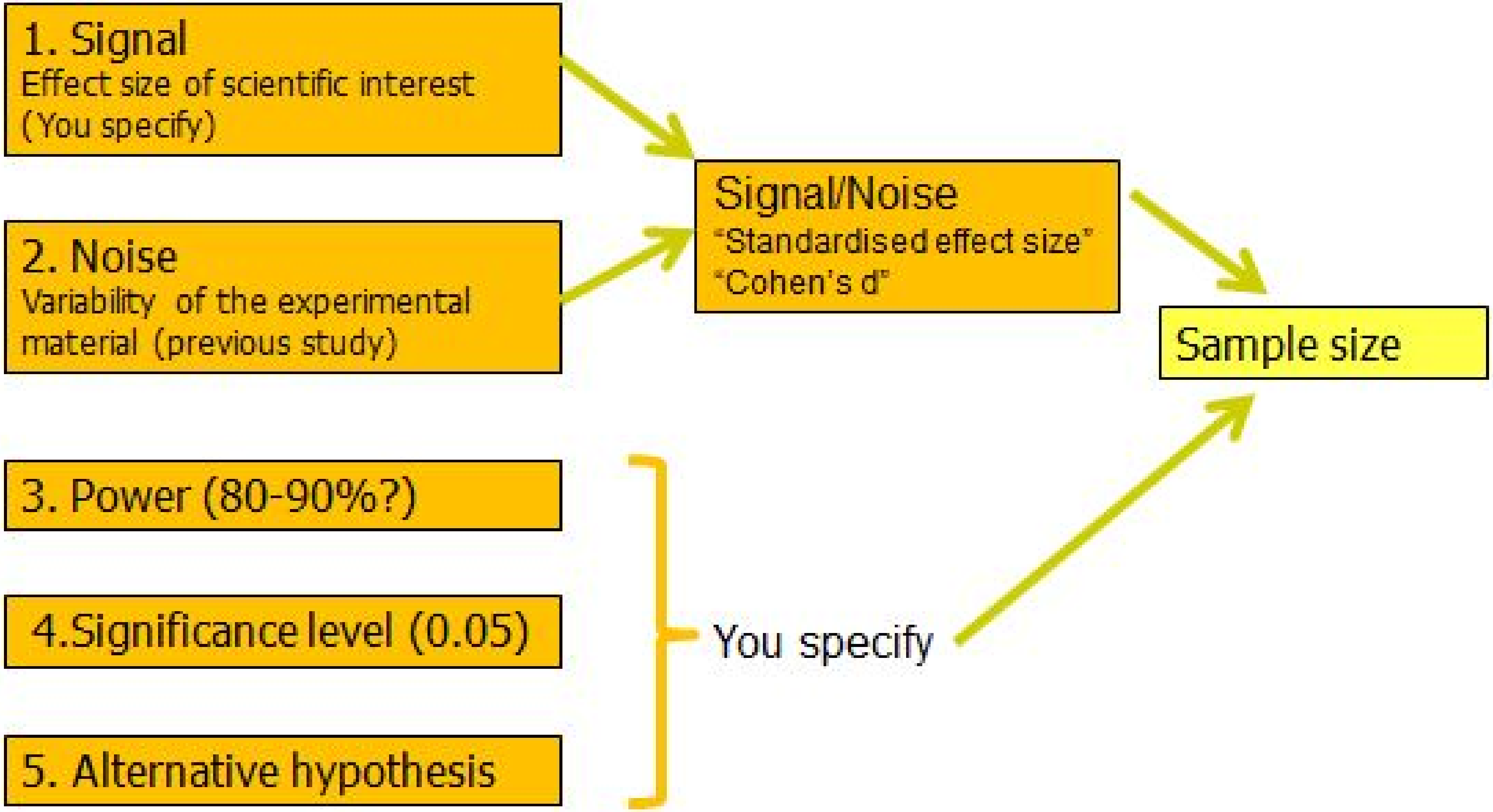
Source:



Summary: Mead vs. Power Analysis

- Preference to use power analysis over Mead‘s resource equation whenever possible
- But: Power analysis not so easy when there are more than two groups because it is more difficult (though not impossible) to specify the effect size of interest
- But: Power analysis requires knowledge about the estimate of the standard deviation
- Mead‘s resource equation method is useful when there is no such previous estimate
- Always better than „relying on tradition“

Power Analysis: Visual

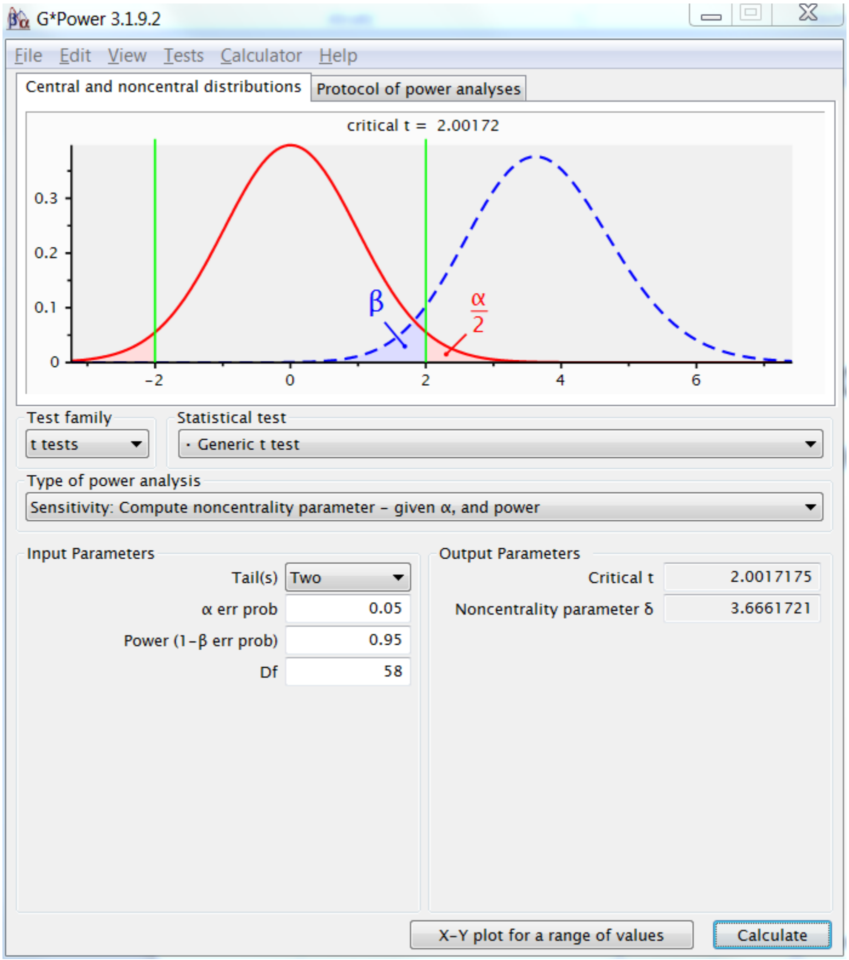


Source: http://www.3rs-reduction.co.uk/html/6__power_and_sample_size.html

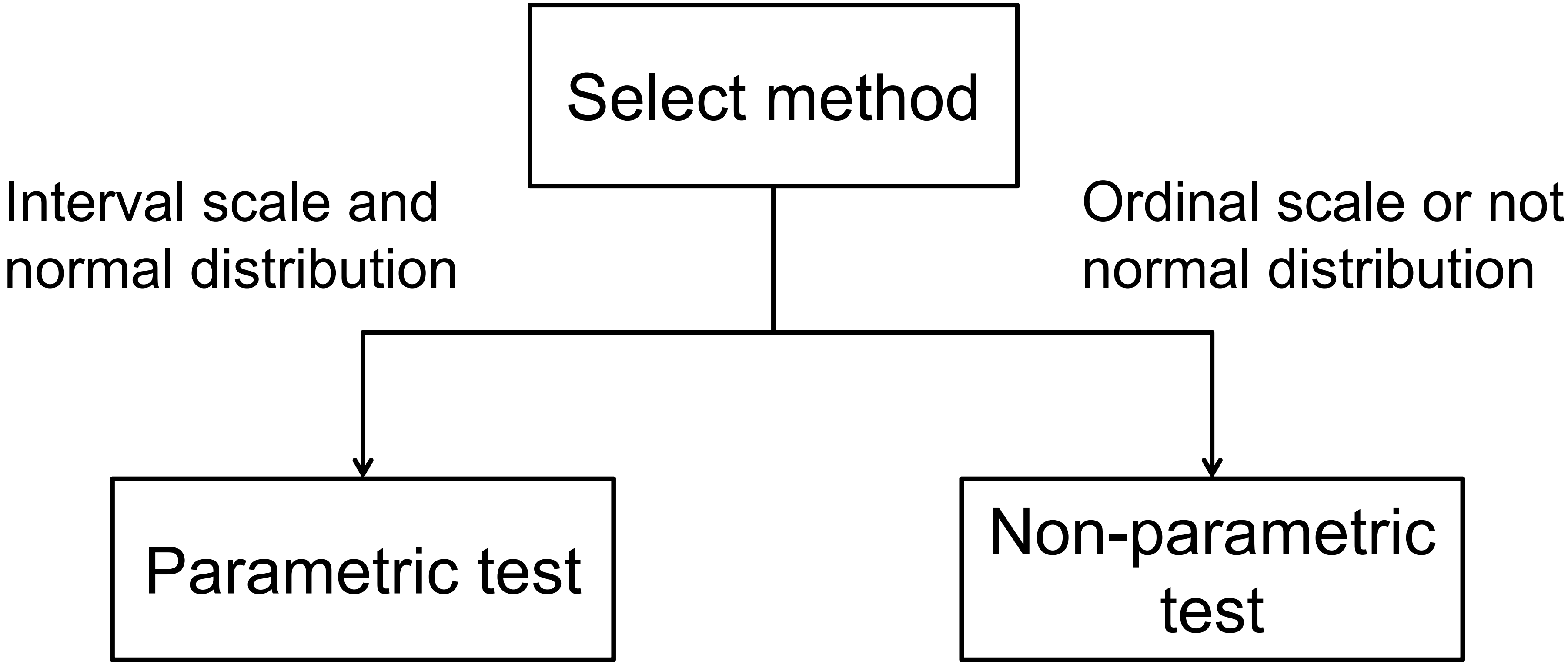
Power Analysis: Tool Support

- Example: G*Power, http://www.gpower.hhu.de/
  - Supports different statistical tests
  - Available for Windows and Mac

- R statistical package
- …



Selecting a Significance Test (cont.)



→Number of conditions?
→Experiment design?

→Number of conditions?
→Experiment design?

Continued on Part 2 …