INFOMSCIP 2019-2020
lectures 16&17
Oct 28&31, 2019

Sources & recommended references:

- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B. and Wesslén, A., 2012. Experimentation in software engineering. Springer Science & Business Media, chapter 10 on "Analysis and Interpretation"

  Although focused on software engineering research, the book is a very good source for research in computer science (esp. experimental research) in general. You can access it for free via the UU library website.
- Alan Dix: "Statistics for HCI", CHI 2017 course, http://alandix.com/statistics/
- J. Edward Swan II: "The Replication Crisis in Empirical Science: Implications for Human Subject Research in Virtual Environments", Tutorial presented at *IEEE Virtual Reality 2018*, Reutlingen, Germany, March 18, 2018, http://web.cse.msstate.edu/~swan/teaching/tutorials/Swan-VR2018-Tutorial-Replication-Crisis.pdf
- Course notes to "Empirical Research Methods for Computer Scientists, University of Toronto, 2016, http://www.cs.toronto.edu/~sme/CSC2130/
- The INFOARM course in the MBI program (slides by Pawel Wozniak)
- Various other sources; some referenced, others not (since lots of this is common knowledge from statistics). If you need background information on particular aspects (e.g., what kind of significance test to use and how), Wikipedia is actually quite helpful for this topic.

Disclaimer: These slides may contain copyrighted material
that is used for pure educational purposes based on the Fair Use policy.
http://en.wikipedia.org/wiki/Fair_use

Scientific perspectives on GMT (INFOMSCIP)

Statistics

- Descriptive statistics
  Basics (mean, variance, …), graphical visualization
- Inferential statistics
  Hypothesis testing, significance tests.
- Problems, issues, challenges
  The replication crisis, p-hacking, biases

The real world

Simulate, model, describe, …

Chaotic, varying contexts (including random ones), huge numbers, …

E.g.: **population** (people visiting or living in Tokyo)

Interpret, explain, predict, …

The scientific world

Models the real world in a controllable, manipulatable, observable way that is best suited for research goal

E.g.: **sample** of people observed, tracked and analyzed in a particular time frame at a particular place under particular conditions

Scientific research (HCI focus)

Exploration → **Validation** → **Explanation**

Finding questions

- Ethnography, literature
- Observations
- End-user interviews
- (Big) data analysis
- Ideas, creative thinking
- …

Answering them

- Experiments
- Large scale survey
- Quantitative data
- …

Finding why and how

- Qualitative data
- Theoretical models
- Mechanisms
- ...

What is the best model, measure, etc. to simulate the real world?

How can we interpret our observations and data and draw conclusions (and are those reliable and generalizable)?

Statistics (i.e., the mathematics of data and how to interpret it) can help with this.

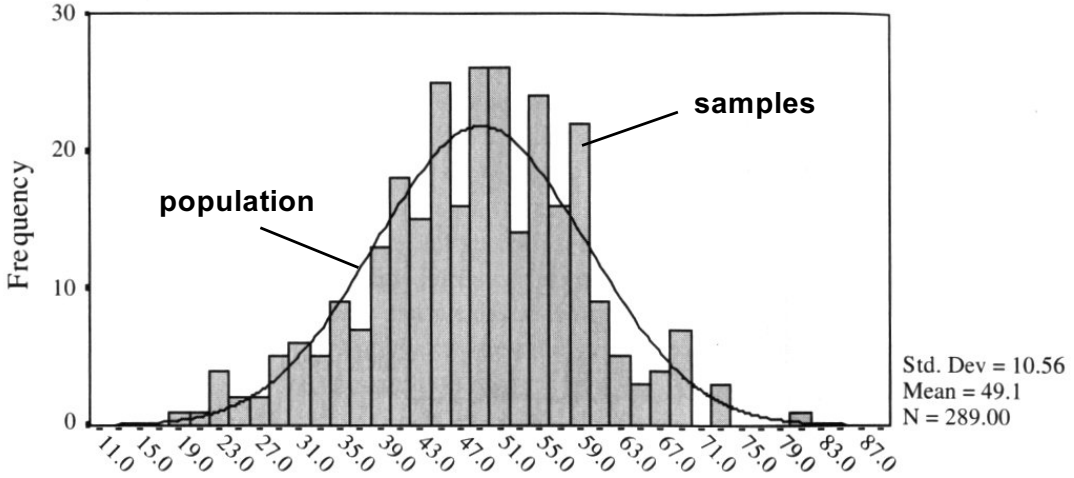
Important questions:

What does the data say?

Descriptive statistics

Describe and graphically present interesting and relevant aspects of the data set.

E.g., average/mean, variance, median, distribution, …



What conclusions can we draw from it?

Predictive / inferential / inductive statistics

Deduce properties of probability distributions to predict future data samples based on those seen in the past.

E.g., hypothesis and significance testing, …

Statistics

- **Descriptive statistics**
  Basics (mean, variance, …), graphical visualization
- Inferential statistics
  Hypothesis testing, significance tests.
- Problems, issues, challenges
  The replication crisis, p-hacking, biases

Data comes in various ways (continuous/discrete, …) and scales.
Common scale types for empirical research are:

Nominal
Data indicating categories. No notion of ordering.
Examples: technique (A/B), gender (male/female), handedness (left/right/both)

Ordinal
Data that can be ordered. Yet, differences between any two values may not be equal (relative distances cannot be compared).

Examples: grades, software complexity

Interval
Data that can be ordered and differences between two values are the same (relative distances are meaningful), but there is no absolute zero. Note: this allows us to have meaningful negative values.

Examples: temperatures in C or F (0 C and 0 F exist, but are artificially defined)

Ratio
Data that can be ordered, differences between two values are the same (relative distances are meaningful), there is an absolute zero. Note: this means that meaningful negative values do not exist (in statistics)

Examples: length, weight, height, time, speed, error rate

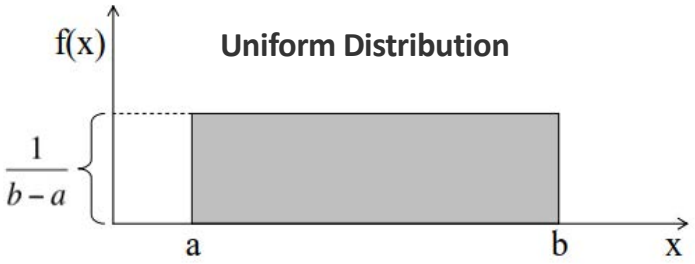
Data distributions

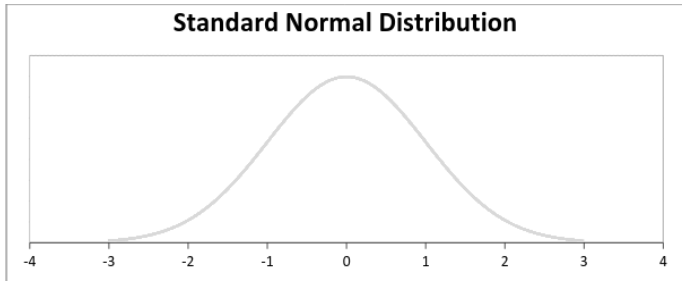
Knowing how data is distributed is the first step towards understanding how a variable is likely to predict another variable.

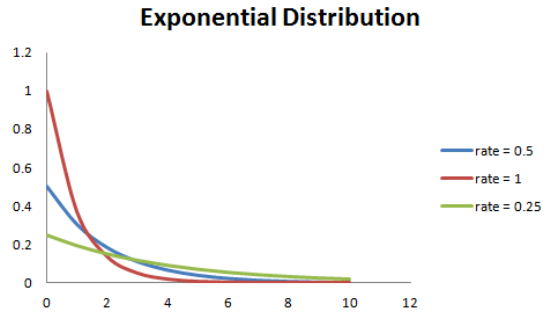
This ***distribution*** can serve as a model of the data.

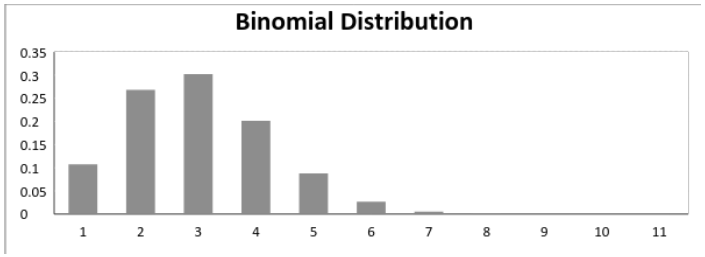
We can accept or reject this model after collecting ***empirical evidence*** (via experimental data).

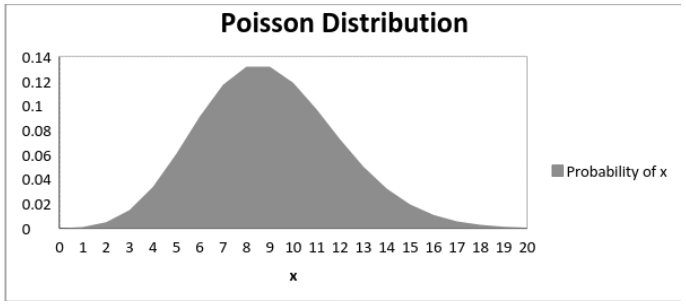
Models allow us to organize our data in a meaningful way.







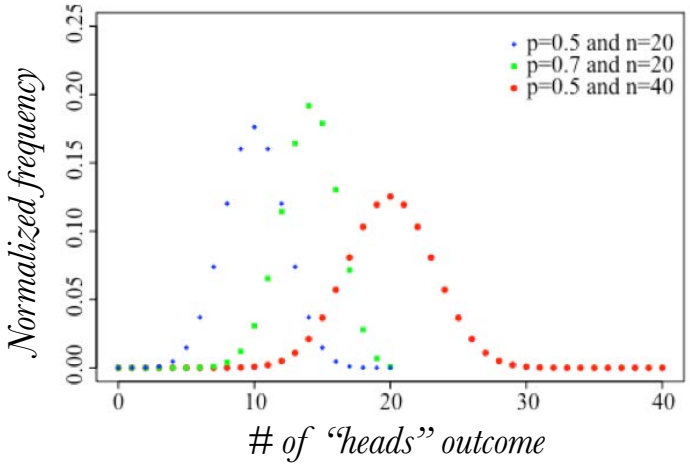




Source: 6 Common Probability Distributions every data science professional should know

https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/

Binomial Distribution

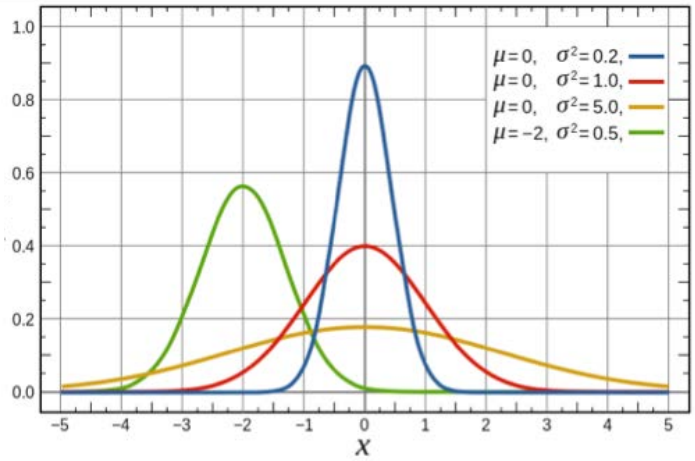


$$P(x) = \frac{N!}{x!(N-x)!}\pi^x(1-\pi)^{N-x}$$

N = number of flips
x = number of desired outcome
Π = probability of the desired outcome (0.5=chance)

The discrete probability distribution of the number of successes (x) in a sequence of *N* independent yes/no experiments, each of which yields success with probability *p*.

Normal Distribution



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

μ = true mean of the distribution
σ^2 = variance of the observations
x = value of the observation(s)

Measures of central tendency

Indicate a 'middle' of a data set. May be interpreted as an estimation of the expectation of the stochastic variable from which the data points in the data set are sampled.

Average / mean

Data points $x_1, x_2, \ldots, x_n$

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

Median

Middle value of a data set; following that the number of samples that are higher than the median is the same as the number of samples that are lower than the median. (for even data sets numbers, usually the average of the two middle values is taken)

Percentile x_p

Denotes the percentile where **p% of samples lie below this value**. (Note: median = x_{50})

Mode

The **most commonly occurring sample**. Calculated by counting the number of samples for each unique value and selecting the value with the highest count. (If an odd number of samples have the same occurrence count, the mode may be selected as the middle value of the most common samples)

Geometric mean

Calculated as the **n:th root of the product of all samples** as follows:

$$\sqrt[n]{\prod_{i=1}^{n} x_i}$$

Be careful with such data

Simpson's Paradox

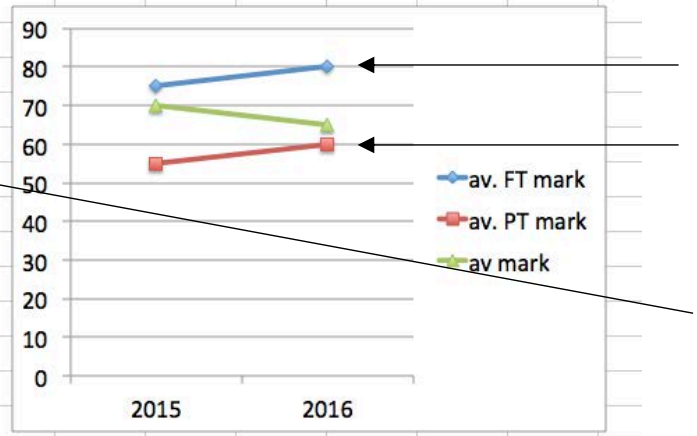
You run a course with some full-time (FT) and some part-time (PT) students
According to your calculations:

- Average FT student marks went up compared to last year
- Average PT student marks also went up compared to last year

Yet, the university complains your marks are going down. Who is right?

You might both be correct!

	2015	2016
av. FT mark	75	80
av. PT mark	55	60
nos FT	30	30
nos PT	10	90
av mark	70	65



Both FT and PT marks increased, so you are right!

But the number of PTs also increased strongly (yet their increase in grades is not that strong). Thus, the average of both groups actually decreases!

Simpson's paradox, or the Yule–**Simpson** effect, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined. It is sometimes given the descriptive title reversal **paradox** or amalgamation **paradox**.

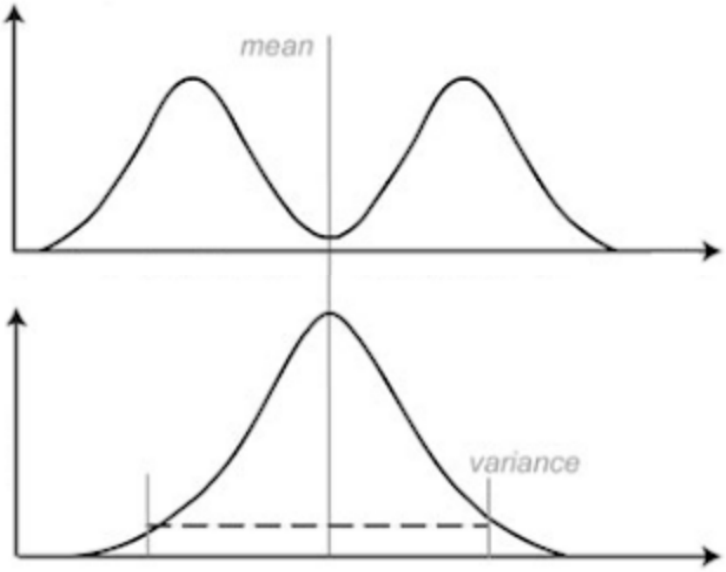
Important: When reporting averages, always state precisely over what the average is taken! (People often forget that in diagrams and on slides, for example)

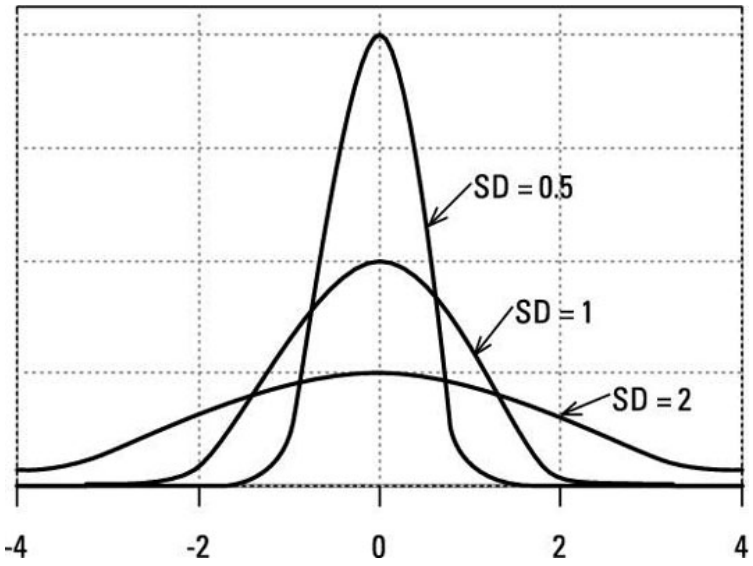
Be careful with such data

Also, don't just look at the average when interpreting your data.
E.g. here, the overall system A has a lower error rate than system B, but system B is better for experts.

		error rate	
	number	system A	system B
novice	30	3.7%	7.4%
expert	10	9.6%	2.7%
all	40	5.2%	6.2%

Remember: this is only a measure of *central* tendency





Measures of dispersion

Convey information of the dispersion of the data set by measuring the level of variation from the central tendency to see how spread or concentrated the data is.

Variance

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

Range

$$x_{max} - x_{min}$$

Variation interval

$$(x_{min},\ x_{max})$$

Standard deviation

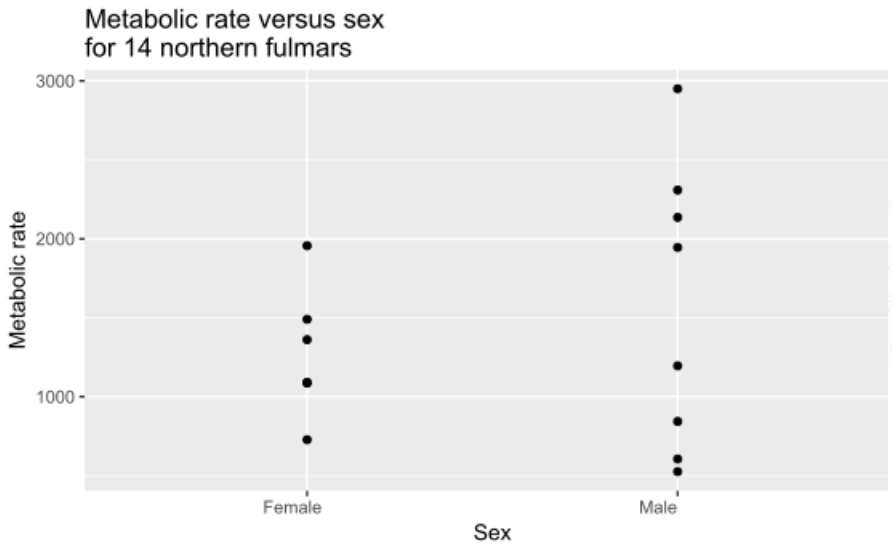
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

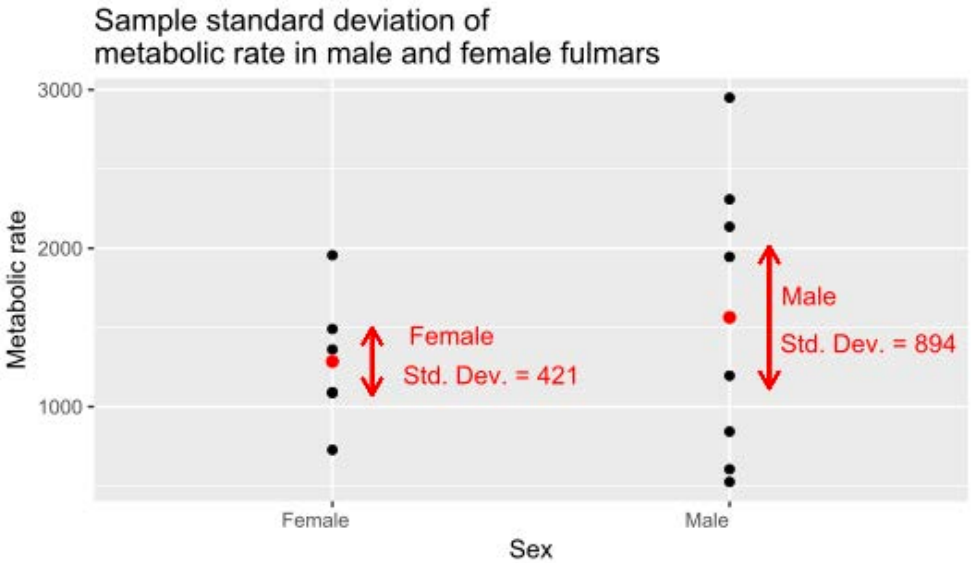
Frequency, relative frequency

Counts the frequency of each data value.
The relative frequency is calculated by dividing the frequency of each value by the total number of samples.

Coefficient of variation

$$100\frac{s}{x}$$





Measures of dependency

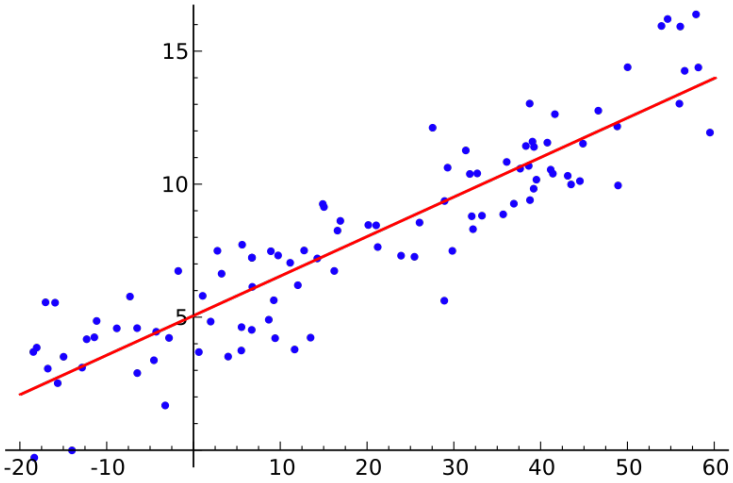
When the data set consists of related samples in pairs $(x_i; y_i)$ from two stochastic variables, X and Y, it is often interesting to examine the dependency between these variables.

Linear regression
… (and many others)

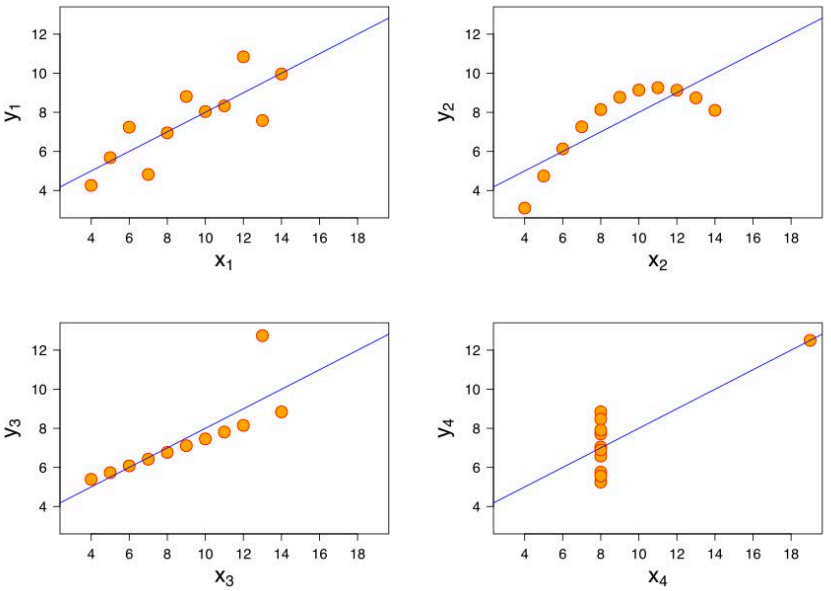
Multivariate analysis, e.g.:

- Multiple regression
- Principle component analysis (PCA)
- Cluster analysis
- Discriminant analysis

If X and Y are related through a linear function $y = a + bx$, we can estimate this function by applying **linear regression**. Regression means fitting the data points to a curve (in the linear case: a line).



Be careful! All these have approximately the same linear regression line (as well as nearly identical means, standard deviations, and correlations) but are graphically very different. This illustrates the pitfalls of relying solely on a fitted model to understand the relationship between variables.



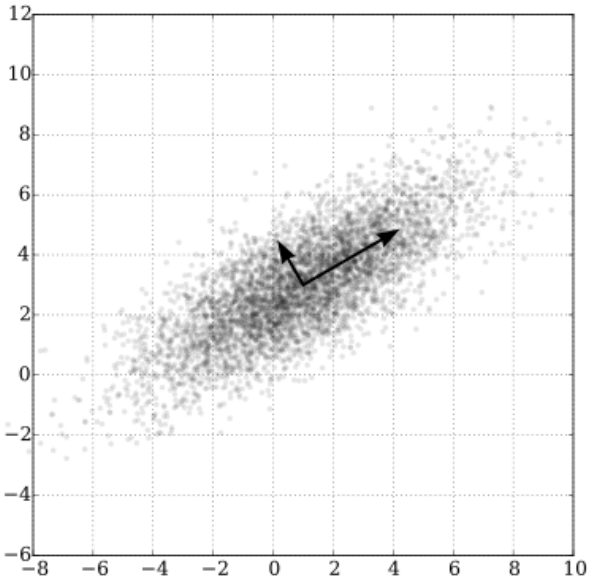
https://en.wikipedia.org/wiki/Linear_regression

Measures of dependency

When the data set consists of related samples in pairs (x_i, y_i) from two stochastic variables, X and Y, it is often interesting to examine the dependency between these variables.

Linear regression
… (and many others)

Multivariate analysis, e.g.:
- Multiple regression
- ***Principle component analysis (PCA)***
- Cluster analysis
- Discriminant analysis



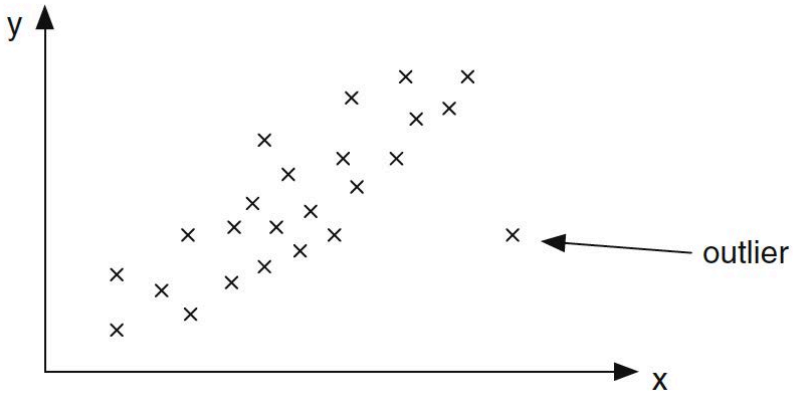
PCA of a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in roughly the (0.866, 0.5) direction and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.

https://en.wikipedia.org/wiki/Principal_component_analysis

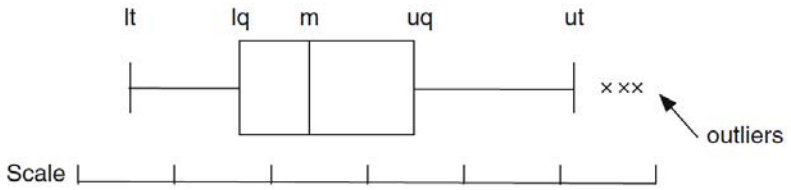
Graphical visualization

When describing a data set, quantitative measures of central tendency, dispersion, and dependency, can be combined with graphical visualization techniques.
Graphs are very illustrative and give a good overview of the data set.

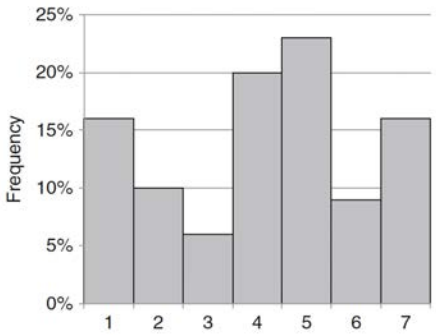
Scatter plot



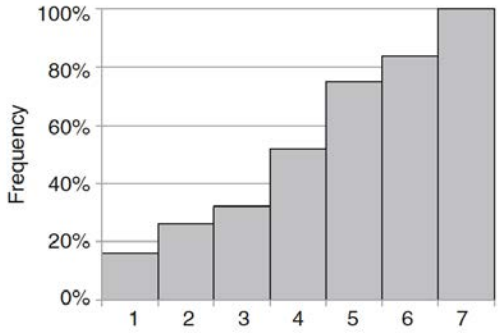
Box plot



Histogram

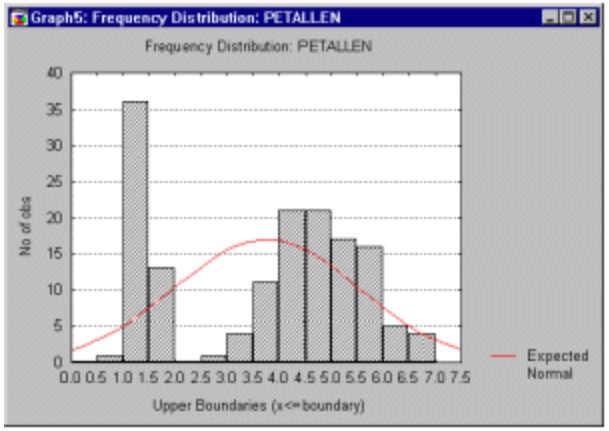


Cumulative histogram



Checking your data is normal

- Draw a Histogram
- Compute the mean and standard deviation
- Superimpose the expected normal curve over the histogram



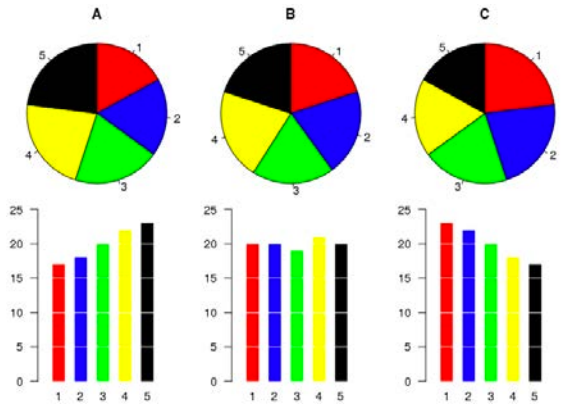
http://www.statsoft.com/textbook/stathome.html

Generally, **be careful with visuals**, e.g., with the scale: smaller scaled can highlight smaller differences better, but might also exaggerate minor irrelevant variations.

Visualizations are **good to identify trends and oddities**,
that can (and most often should) then be studied more objectively.

Also:

Three sets of percentages, plotted as both pie charts and bar charts. Comparing the data on bar charts is generally easier.



Hint: for bar charts, always think about sorting (rule of thumb: generally any order is better than random order, e.g., subjects)

Statisticians generally regard pie charts as a poor method of displaying information, and they are uncommon in scientific literature. One reason is that it is more difficult for comparisons to be made between the size of items in a chart when area is used instead of length and when different items are shown as different shapes.

(*From Wikipedia, "Pie charts"*)

Always chose the most appropriate visualization type based on data and aspect that you want to study or illustrate.

Scientific perspectives on GMT (INFOMSCIP)

Statistics

- Descriptive statistics
  Basics (mean, variance, …), graphical visualization
- **Inferential statistics**
  Hypothesis testing, significance tests.
- Problems, issues, challenges
  The replication crisis, p-hacking, biases

The real world

Chaotic, varying contexts (including random ones), huge numbers, …

E.g.: **population**
(people visiting or living in Tokyo)

Get data (e.g., from population)

Draw conclusions, make **predictions** (e.g., about population)

The scientific world

Models the real world in a controllable, manipulatable, observable way that is best suited for research goal

E.g.: **sample** of people observed, tracked and analyzed in a particular time frame at a particular place under particular conditions

Inferential statistics

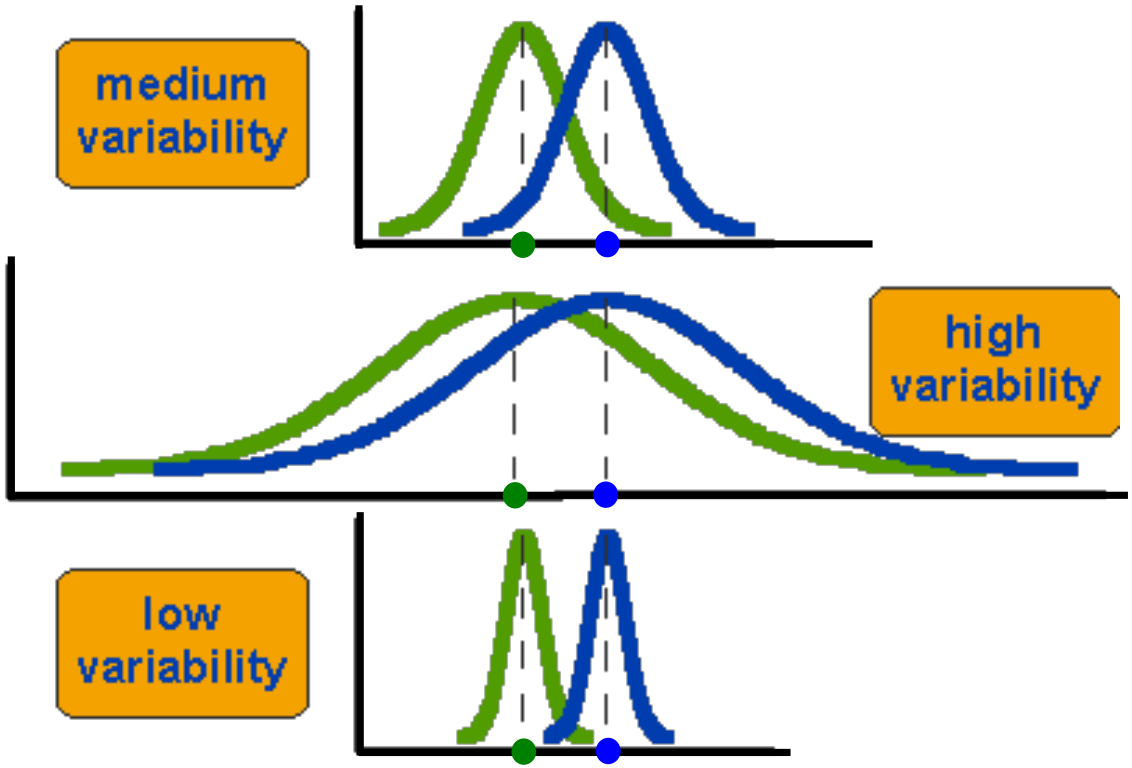
Drawing conclusions from your data and making predictions

Descriptive statistics only describe your data. **Inferential statistics** aim at interpreting them and verifying their applicability to similar scenarios.

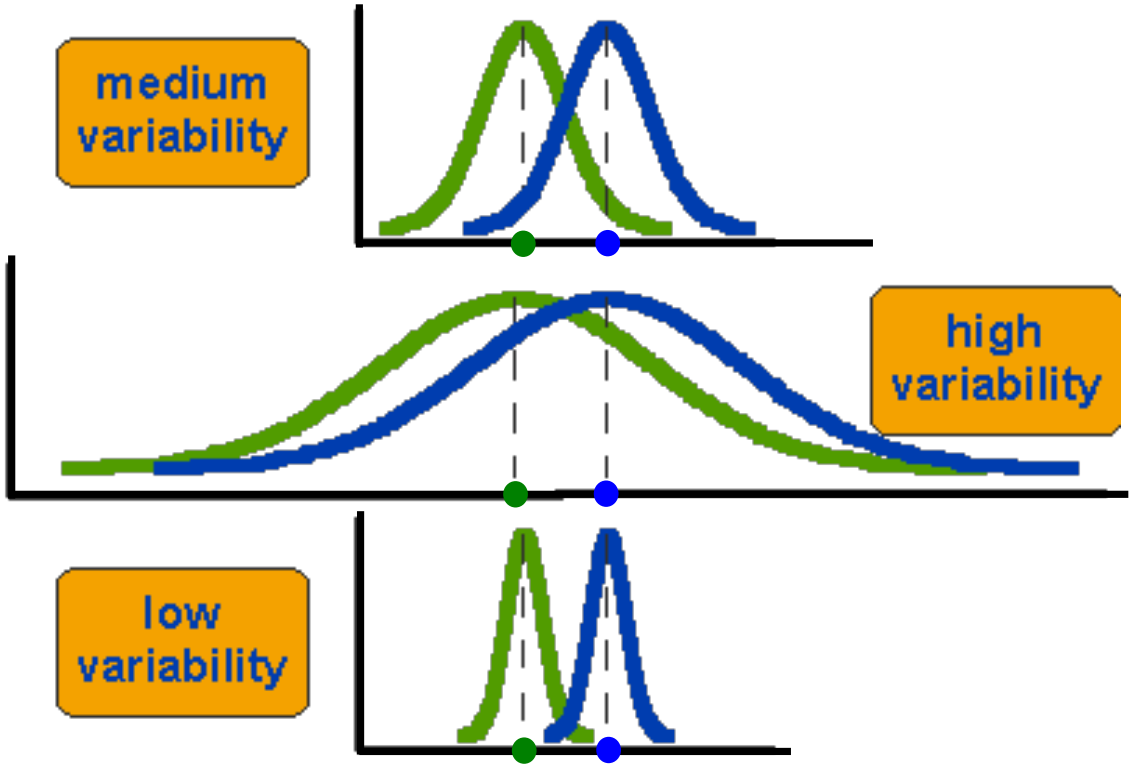
How do we know that our results are representative and not just by chance?

Example:
3 different experiments

- Almost identical averages
- But different variability



It appears that only the data from the last case represents a general result that does not just illustrate the analyzed sample data, but should apply to other samples from the same population as well.



Statistical hypothesis testing (aka Null Hypothesis Significance Testing, NHST) is a statistical method to verify such an intuitive assumption. It verifies whether the factor we are talking about has the effect on our observation.

The following slides are from the tutorial "**The Replication Crisis in Empirical Science: Implications for Human Subject Research in Virtual Environments**" by J. Edward Swan II, presented at *IEEE Virtual Reality 2018*, Germany, March 18, 2018 (***with slight modifications and additions for this course***)

http://web.cse.msstate.edu/~swan/teaching/tutorials/Swan-VR2018-Tutorial-Replication-Crisis.pdf

References

For the "Hypothesis testing" slides:

[Cohen 1994] J Cohen, "The Earth is Round ($p < .05$)", *American Psychologist*, 49(12), pages 997–1003.

[Cohen 1988] J Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[Howell 2002] DC Howell, *Statistical Methods for Psychology*, 5th edition, Duxbury, Pacific Grove, CA, 2002.

[Living Swan et al 2003] MA Livingston, JE Swan II, JL Gabbard, TH Höllerer, D Hix, SJ Julier, Y Baillot, D Brown, "Resolving Multiple Occluded Layers in Augmented Reality", The 2nd *International Symposium on Mixed and Augmented Reality* (ISMAR), 56–65, 2003.

[Swan et al 2003] JE Swan II, JL Gabbard, D Hix, RS Schulman, KP Kim, "A Comparative Study of User Performance in a Map-Based Virtual Environment", Technical Papers, *IEEE Virtual Reality*, 259–266, 2003.

For the "Reproducibility project / replication crisis" slides:

[Economist 2013] "Unreliable Research: Trouble at the Lab", *The Economist*, 18 Oct 2013.

[Freedman 2010] Freedman, D. H., "Lies, Damned Lies, and Medical Science: Dr. John Ioannidis Exposes the Bad Science of Colleagues", *The Atlantic*, Nov 2010.

[Groby 2016] Gobry, P.-E., "Big Science is Broken", *The Week*, 18 April 2016. [Hen Thom 2017] Henderson, D., Thomson, K., "What Makes Science True?", *NOVA Video Short*, 1 Jan 2017. http://www.pbs.org/wgbh/nova/body/reproduce-science.html

[Ioannidis 2005] Ioannidis, J. P. A., "Why Most Published Research Findings Are False", *PLOS Medicine*, 2(8), e124., 2005. http://doi.org/10.1371/journal.pmed.0020124

[OSC 2015] Open Science Collaboration, "Estimating the Reproducibility of Psychological Science", *Science*, 349(6251), 2015, DOI: 10.1126/science.aac4716

[OSC 2012] Open Science Collaboration, "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science", *Perspectives on Psychological Science*, 7(6), 657–660, 2012. http://doi.org/10.1177/1745691612462588

[Prinz et al. 2011] Prinz, F., Schlange, T., & Asadullah, K., "Believe it or not: How much can we rely on published data on potential drug targets?", *Nature Reviews Drug Discovery*, 10(9), 712–712, 2011. http://doi.org/10.1038/nrd3439-c1

[Rehman 2013] Rehman, J., "Cancer research in crisis: Are the drugs we count on based on bad science?", *Salon*, 1 Sep 2013.

[Young 2016] Young, E. (2016, March 4). "Psychology's Replication Crisis Can't Be Wished Away", *The Atlantic*, 4 Mar 2016.

[Young 2015] Young, E., "How Reliable Are Psychology Studies?: Brian Nosek's Reproducibility Project Finds Many Psychology Studies Unreliable", *The Atlantic*, 25 Aug 2015.

Hypothesis testing – Basic principle

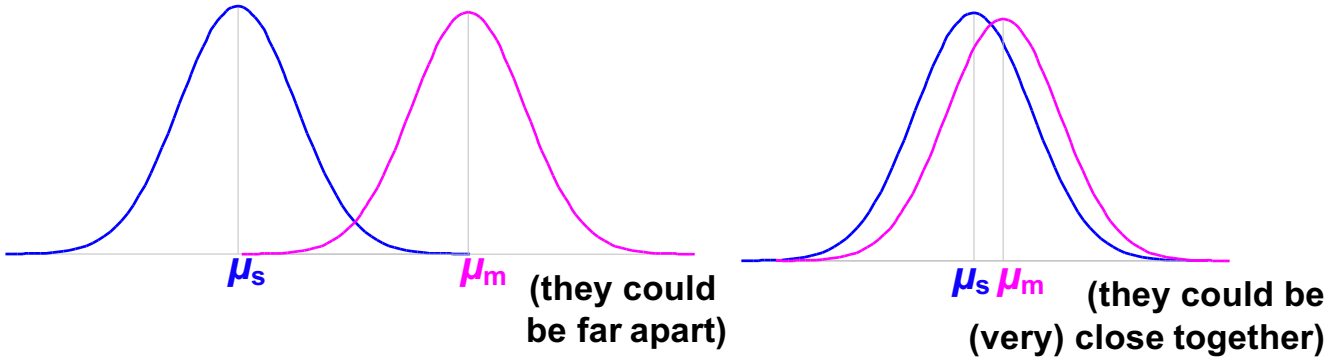
Assume an **experiment** where we ***navigate in VR*** and want to measure the impact of stereo (3D) versus mono vision on navigation time.

Assume we have two populations for *time to navigate* based on measures from two experiments, one with stereo, one with mono vision:

μ_s: *stereo time* and μ_m: *mono time*

Perhaps there are two populations:

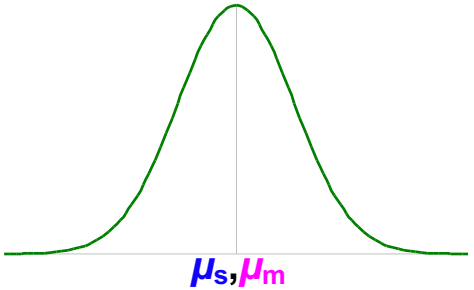
$\mu_s - \mu_m = d$



This would mean *vision* has an impact on navigation time.

Perhaps there is one population:

$\mu_s - \mu_m = 0$



This would mean it hasn't.

Hypothesis testing: terminology

Null hypothesis H_0

Assumes that there are no real underlying trends or patterns in the experiment setting.

Generally, we want to show that there *is* a difference, thus reject this hypothesis

Alternative hypothesis H_1

The hypothesis in favor of which the null hypothesis is rejected.

Thus, when we reject H_0, we can assume that H_1 is true.
(Note: sometimes H_a *is used instead of* H_0.)

Statistical tests exist to verify, if a null hypothesis can be rejected.

Which test to chose depends on the type of data and its distribution.

But they all follow the same **basic principle**.
(Thus, if you pick the right one, you can use it as a black box.)

Hypothesis testing: basic procedure

1. Develop testable **hypothesis** H_1: $\mu_s - \mu_m = d$
   (E.g., subjects faster under stereo viewing)

2. Develop **null hypothesis** H_0: $\mu_s - \mu_m = 0$
   Logical opposite of testable hypothesis

3. Construct sampling distribution assuming H_0 is true.

4. Run an experiment and collect samples; yielding **sampling statistic *X***.
   (E.g., measure subjects under stereo and mono conditions)

5. Referring to sampling distribution,
   calculate **conditional probability of seeing *X* given H_0: *p*(*X* | H_0).**
   - If probability is low ($p \leq 0.05$), we are *unlikely* to see X when H_0 is true.
     We *reject* H_0 and *embrace* H_1.
   - If probability is not low ($p > 0.05$), we are *likely* to see X when H_0 is true.
     We do *not reject* H_0.

Example 1: VE Navigation with Stereo Viewing [Swan et al. 2003]

1. Hypothesis H_1: $\mu_s - \mu_m = d$
   Subjects faster under stereo viewing.

2. Null hypothesis H_0: $\mu_s - \mu_m = 0$
   Subjects same speed whether stereo or mono viewing.

3. Constructed sampling distribution assuming H_0 is true.

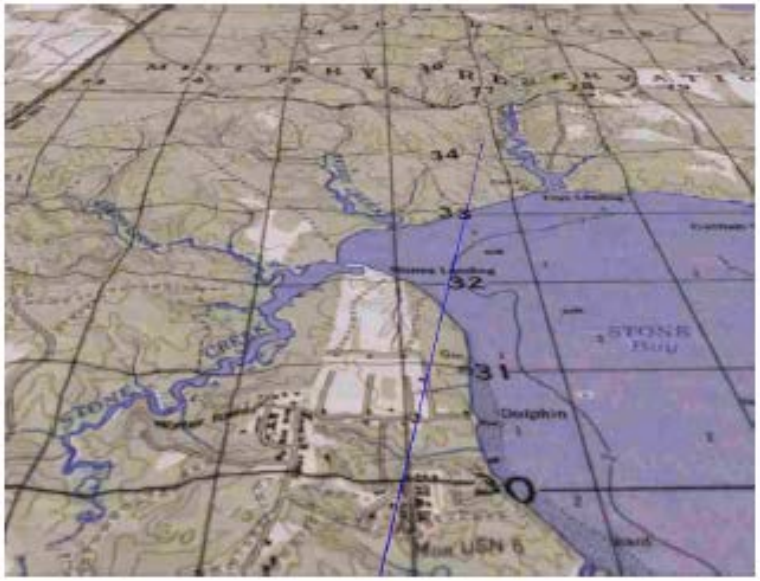


Figure 1: Typical user view of the *Dragon* map during the experiment.

4. Ran an experiment and collected samples:
   32 participants, collected 128 samples
   $X_s = 36.431$ sec; $X_m = 34.449$ sec; $X_s - X_m = 1.983$ sec

5. Calculated conditional probability of seeing 1.983 sec given H_0: $p(1.983 \text{ sec} \mid H_0) = 0.445$.

   $p = 0.445$ not low, we are likely to see 1.983 sec when H_0 is true.
   We do not reject H_0.

   This experiment did *not* tell us that subjects were faster under stereo viewing.

Example 2: Effect of Intensity on AR Occluded Layer Perception [Living Swan et al. 2003]

1. Hypothesis H_1: $\mu_c - \mu_d = d$
   Tested constant and decreasing intensity.
   Subjects faster under decreasing intensity.
2. Null hypothesis H_0: $\mu_c - \mu_d = 0$
   Subjects same speed whether constant or decreasing intensity.
3. Constructed sampling distribution assuming H_0 is true.
4. Ran an experiment & collected samples:
   8 participants, collected 1728 samples
   $X_c = 2592.4$ msec; $X_d = 2339.9$ msec; $X_c - X_d = 252.5$ msec
5. Calculated conditional probability of seeing 252.5 msec given H_0: $p(\ 252.5 \text{ msec} \mid H_0\ ) = 0.008$.
   $p = 0.008$ is low ($p \le 0.01$); we are unlikely to see 252.5 msec when H_0 is true.
   We reject H_0 and embrace H_1.
   This experiment suggests that subjects are faster under decreasing intensity.

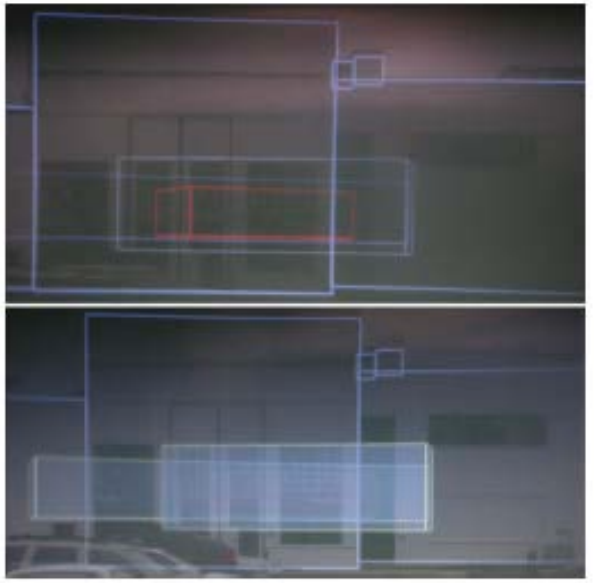


Figure 1. Before-and-after pictures of one of our visualization techniques.

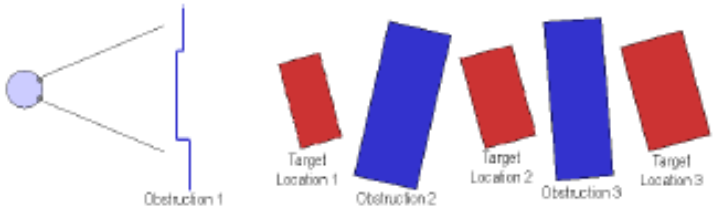


Figure 3. The experimental design (not to scale)

Some Considerations...

- **The conditional probability $p(X \mid H_0)$**
  - Much of statistics involves how to calculate this probability; source of most of statistic's complexity
  - Logic of hypothesis testing the same regardless of how $p(X \mid H_0)$ is calculated
  - If you can calculate $p(X \mid H_0)$, you can test a hypothesis

- **The null hypothesis H_0**
  - H_0 usually in form $f(\mu_1, \mu_2, \ldots) = 0$

    Gives hypothesis testing a double-negative logic: assume H_0 as the opposite of H_1, then reject H_0

    Philosophy is that we can never prove $f = 0$, because 0 is point value in domain of real numbers
  - H_1 usually in form $f(\mu_1, \mu_2, \ldots) \neq 0$

    We don't know what value it will take, but main interest is that it is not 0

When We Reject H_0

Calculate $\alpha = p(X \mid H_0)$, when do we reject H_0?

- In science generally, $\alpha = 0.05$
- But, just a social convention

What can we say when we reject H_0 at $\alpha = 0.008$?

- "If H_0 is true, there is only an 0.008 probability of getting our results, and this is unlikely."
  Correct!
- "There is only a 0.008 probability that our result is in error."
  Wrong, this statement refers to $p(H_0)$, but that's not what we calculated.
- "There is only a 0.008 probability that H_0 could have been true in this experiment."
  Wrong, this statement refers to $p(H_0 \mid X)$, but that's not what we calculated.

[Cohen 1994]

Important:

- If the null hypothesis is rejected, it can be stated that the hypothesis is false with a given significance (α).
- If it is not rejected, nothing can be said about the outcome.

When We Don't Reject H_0

What can we say when we don't reject H_0 at $\alpha = 0.445$?

- "We have proved that H_0 is true."
  "Our experiment indicates that H_0 is true."
  Both wrong! Hypothesis testing cannot prove H_0: $f(\mu_1, \mu_2, \ldots) = 0$.

Statisticians do not agree on what failing to reject H_0 means.

- Conservative viewpoint (Fisher):
  We must suspend judgment, and cannot say anything about the truth of H_0.
- Alternative viewpoint (Neyman & Pearson):
  We can accept H_0 if we have sufficient experimental power,
  and therefore a low probability of type II error.

[Howell 2002, p 99]

Probabilistic Reasoning

If hypothesis testing was **absolute**:

- If H_0 is true, then X cannot occur…
  however, X has occurred … therefore H_0 is false.

- e.g.: If a person is a Martian, then they are not a member of Congress (true)…
  this person is a member of Congress … therefore they are not a Martian.
  (correct result)

- e.g.: If a person is an American, then they are not a member of Congress (false)…
  this person is a member of Congress … therefore they are not an American.
  (incorrect result, but correct logical reasoning)

p	q	$p \rightarrow q$	$\neg q \rightarrow \neg p$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

$$\frac{p \rightarrow q \quad \neg q}{\rightarrow \neg p}$$

modus tollens

[Cohen 1994]

Probabilistic Reasoning

However, hypothesis testing is **probabilistic**:

- If H_0 is true, then X is highly unlikely…
  however, X has occurred… therefore H_0 is highly unlikely.

- e.g.: If a person is an American,
  then they are probably not a member of Congress (true, right?)…
  This person is a member of Congress…
  therefore they are probably not an American.
  (incorrect result, but correct hypothesis testing reasoning)

p	q	$p \rightarrow q$	$\neg q \rightarrow \neg p$
T	T	T	T
T	F	F	F
F	T	T	T
F	F	T	T

$p \rightarrow q$
$\neg q$
$\rightarrow \neg p$

modus tollens

[Cohen 1994]

Testing for statistical significance

In **statistical hypothesis testing**, a result has statistical significance when it is very unlikely to have occurred given the null hypothesis. More precisely, a study's defined **significance level, α**, is the probability of the study rejecting the null hypothesis, given that it were true; and the ***p*-value** of a result, ***p***, is the probability of obtaining a result at least as extreme, given that the null hypothesis were true. The result is **statistically significant,** by the standards of the study, when $p < \alpha$. The significance level for a study is chosen before data collection, and typically set to 5% or much lower, depending on the field of study.

Wikipedia ("statistical significance")

For the statistical tests:

- ***p*-value** (we calculate this) - probability that a relationship observed in the sample happened by chance
- **Alpha level** (selected a priori) - a threshold for *p* at which we will accept that a relationship did not happen by chance (typically 0.1 or 0.05)
  - If $p < \alpha$, we say the result was significant
  - This allows us to fix the probability of a type I error in advance

Student's t test

For testing whether two samples really are different

- Given: **two experimental treatments, one dependent variable**
- Assumes that:
  - the variables are normally distributed in each treatment
  - the variances for the treatments are similar
  - the sample sizes for the treatments do not differ hugely
- Basis: difference between the means of samples from two normal distributions is itself normally distributed.
- The t-test checks whether the treatments are significantly different

Procedure

- H_0: "There is no difference in the population means from which the samples are drawn"
- Choose a significance level (e.g. 0.05)
- Calculate *t* as $t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(SE_A)^2 + (SE_B)^2}}$ where $SE = \frac{SD}{\sqrt{N}}$
- Look up the value for *t*, with degrees of freedom df = $(n_A + n_B) - 2$
- If calculated value of *t* is greater than the lookup value, reject H_0

Analysis of Variance (ANOVA)

Generalization of t-test for >2 treatments

- Given: **n experimental treatments, one dependent variable**
- Assumes that …:
  - the variables are normally distributed in each treatment
  - the variances for the treatments are similar
  - the sample sizes for the treatments do not differ hugely

  (It's okay to deviate slightly from these assumptions for larger samples sizes)
- Works by analyzing how much of the total variance is due to differences within groups, and how much is due to differences across groups.

Procedure:

- H_0: "There is no difference in the population means across all treatments"
- Compute the *F*-statistic:
  - *F* = (found variation of the group averages) / (expected variation of the group averages)

  (Note: don't do this by hand!)
- If H_0 is true, we would expect F=1
- Note: ANOVA tells you whether there is a significant difference, but does not tell you which treatment(s) are different (use post-hoc tests for this).

General comments

- There are various different **tests for statistical significance**.
- Which one to chose depends on your experiment. In particular:
  - What data you have (e.g., discrete / continuous, …)
  - How the data is distributed (e.g., normal distribution, …)
  - How many independent variables and levels you have
  - …

- They all follow the same **procedure** though, i.e.,
  calculating the probability to observe your sample X
  under the assumption that the null hypothesis H_0 is true, i.e., $p(X \mid H_0)$.

- **Statistics tools** (SPSS, R, Excel, …) are commonly used to do these tests.
- The difficult part is choosing the right test, data, alpha level, etc.

- It is also important to **report test results** accurately
- What values to report depends on the test, e.g.
  - t-tests require to report a *t*-value and the *p*-value
  - ANOVA requires to report an *F*-value and the *p*-value

So many tests. So many options. Which one must I choose???

So many tests. So many options. Which one must I choose???

Independent variable	Dependent variable	
Parametric		
Two valued	Normal	Student's t-test on difference of means
Discrete	Normal	ANOVA (Analysis Of VAriance)
Continuous	Normal	Linear (or non-linear) regression factor analysis
Non-parametric		
Two valued	Continuous	Wilcoxon (or Mann-Whitney) rank-sum test
Discrete	Continuous	Rank-sum version of ANOVA
Continuous	Continuous	Spearman's rank correlation
Contingency tests		
Two valued	Discrete	No special test, see next entry
Discrete	Discrete	Contingency table and chi-square test
Continuous	Discrete	(Rare) Group independent variable and then as above

So many tests. So many options. Which one must I choose???

The terms like paired, ordinal, factors, levels, etc. can be quite confusing at first, but once you understand what they mean, it becomes a pretty straightforward approach of classifying your experimental design and then picking the right test.

Helpful resources for picking the right tests:

http://yatani.jp/teaching/doku.php?id=hcistats:start
See esp. the table under "What statistical test should I use?"

Another very helpful site:
https://www.graphpad.com/support/faqid/1790/

- The most difficult task is picking the right test.
- For the actual test, usually tools like SPSS or R are used (simple test can be done with Excel, too).

- You need a good understanding of the basics to:
  a) pick the right test,
  b) get the parameters right when using the tools, and
  c) report the results correctly.

Every statistical model is relying on **specific assumption** regarding, e.g., distribution, independence and scales. For example:

Normality
If a test assumes that the data is normally distributed, a Chi-2 test can be made to assess to which degree the assumption is fulfilled.

Independence
If the test assumes that the data is a sample from several independent stochastic variables, it is necessary to check that there is no correlation between the sample sets. This may be checked with scatter plots and by calculating correlation coefficients.

Residuals
In many statistical models, there is a term that represents the residuals (statistical error). It is often assumed that the residuals are normally distributed. A common way to check this property is to plot the residuals in a scatter plot and see that there is no specific trends in the data (the distribution looks random).

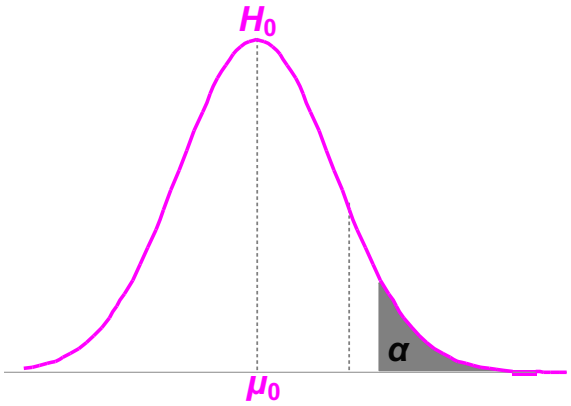
Power, effect size, p-value

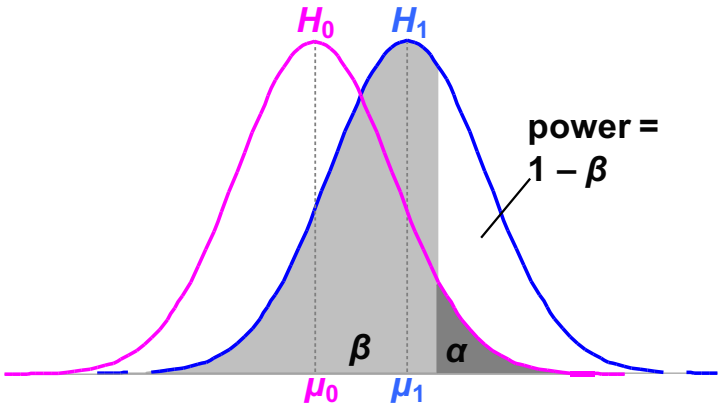
Interpreting α, β, and Power

		Decision	
		Reject H_0	Don't reject H_0
True state of the world	H_0 false	A (correct) result! $p = 1 - \beta$ = power	Type II error $p = \beta$
	H_0 true	Type I error $p = \alpha$	Argue H_0? $p = 1 - \alpha$

If H_0 is true:

α is probability we make a type I error (false positive): we think we have a result, but we are wrong



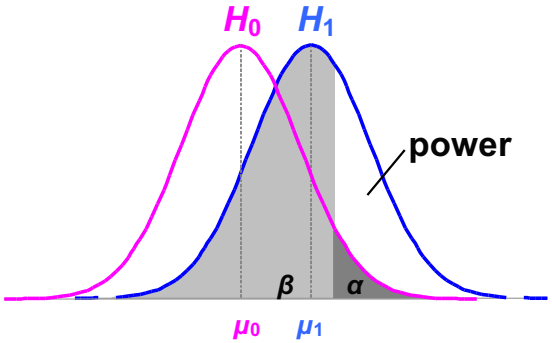


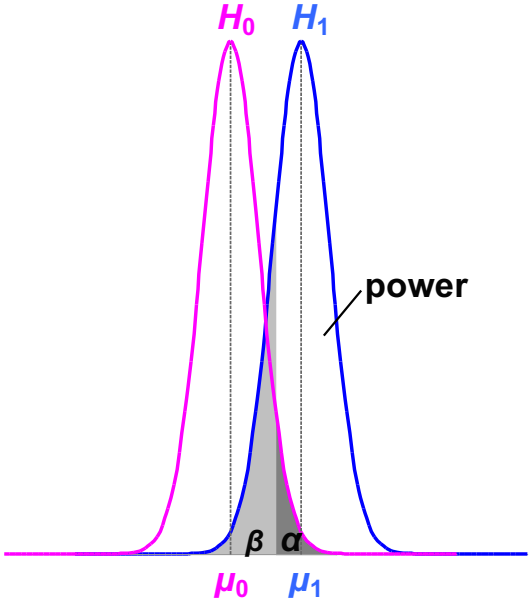
If H_1 is true:

β is probability we make a type II error (false negative): a result was there, but we missed it

Power is a more common term than β

In general, we want our results to have high *power*.
General **strategies to increase power** (i.e., get stronger results) are:





Increase number of ***samples***
(e.g., more subjects)

- The standard approach …
  but square root often means
  very large increases

Reduce ***noise*** in your data

- Better control conditions
  (physics approach)
- Measure other factors and fit
  (e.g. age, experience)

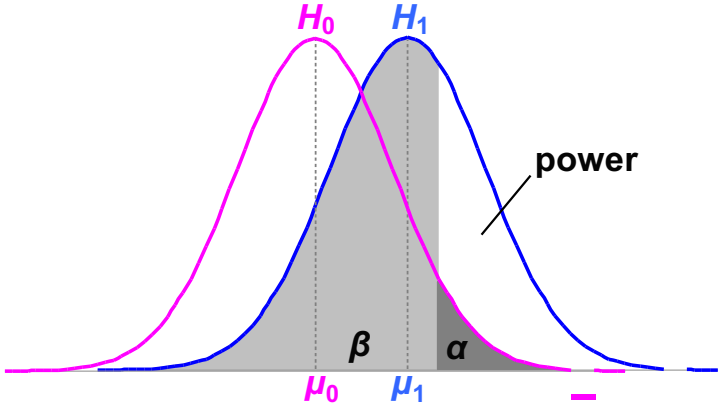
Increase ***effect size***

- Manipulate sensitivity
  (e.g. photo back of crowd!)

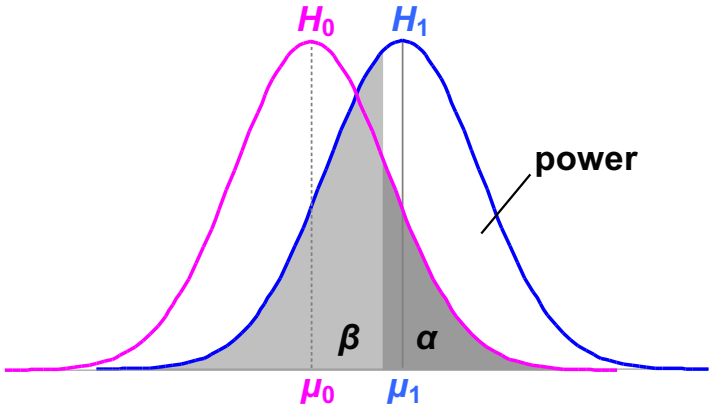
Increasing Power by Increasing α

Illustrates α / power tradeoff

- Increasing α:
  - Increases power
  - Decreases type II error
  - Increases type I error



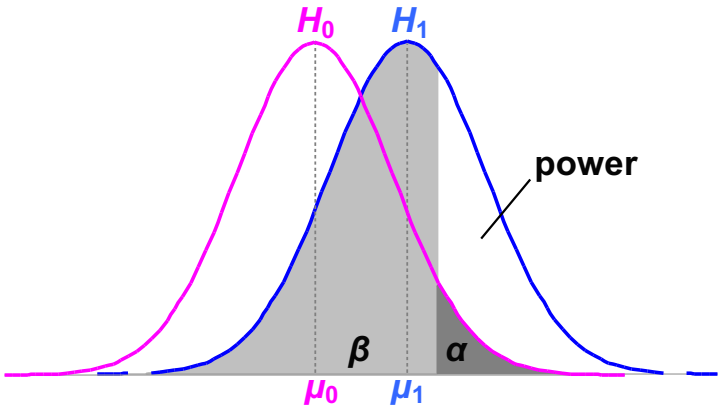
- Decreasing α:
  - Decreases power
  - Increases type II error
  - Decreases type I error



$\Rightarrow$ We are technically "trading" the decrease of the likelihood of one type of error for the increase of the one for another type of error.

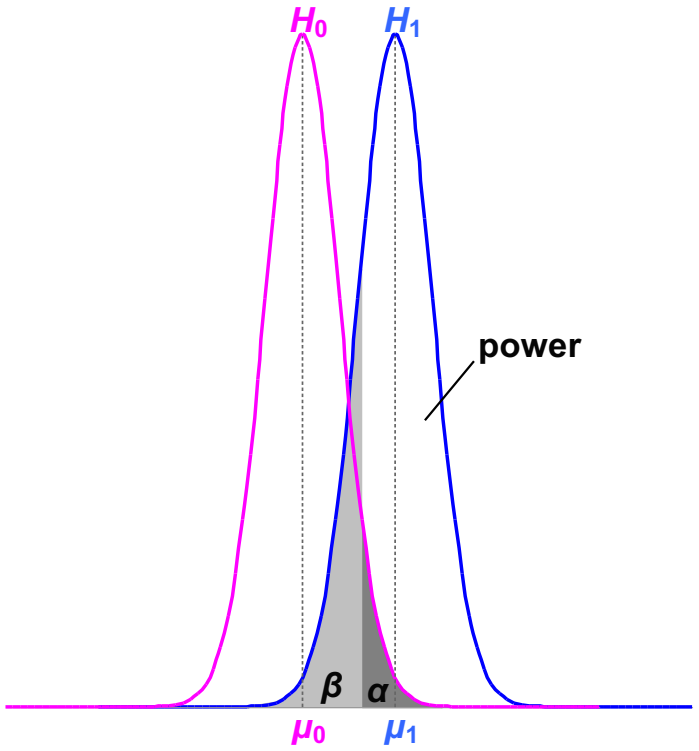
$\Rightarrow$ Be careful with statements such as: "results are close to significance level"

Increasing Power by Collecting More Data



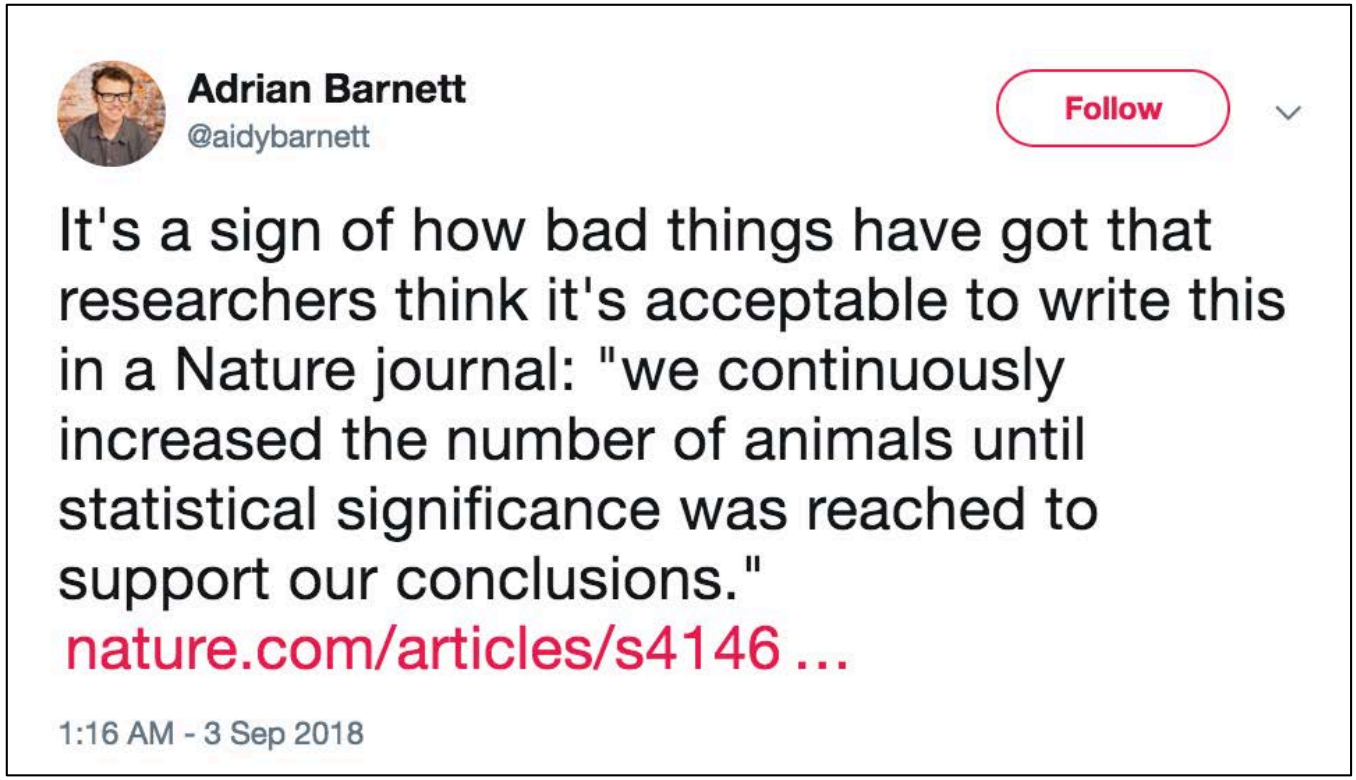
- Increasing sample size (*N*):
  - Decreases variance
  - Increases power
  - Decreases type II error
  - *α* and type I error stay the same

There are techniques that give the value of *N* required for a certain power level.

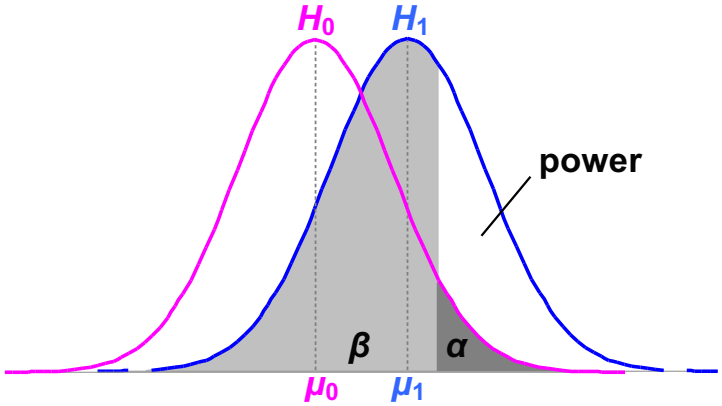


Here, effect size remains the same, but variance drops by half.

But be careful with this …

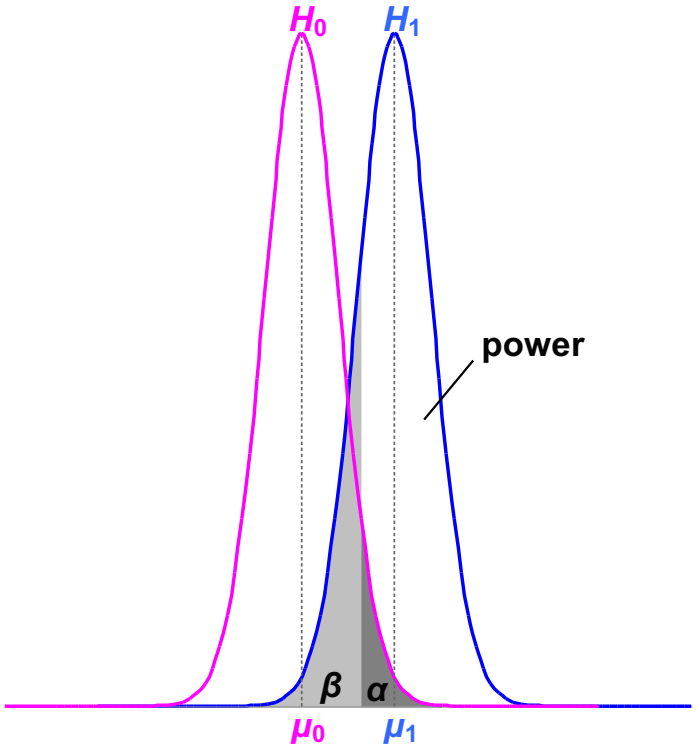


Increasing Power by Decreasing Noise



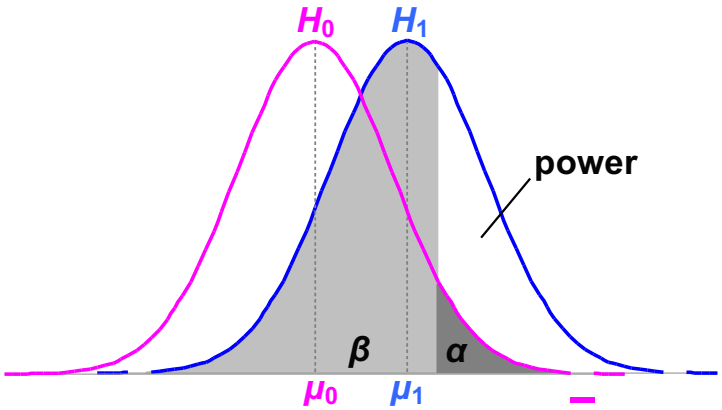
- Decreasing experimental noise:
  - Decreases variance
  - Increases power
  - Decreases type II error
  - *a* and type I error stay the same

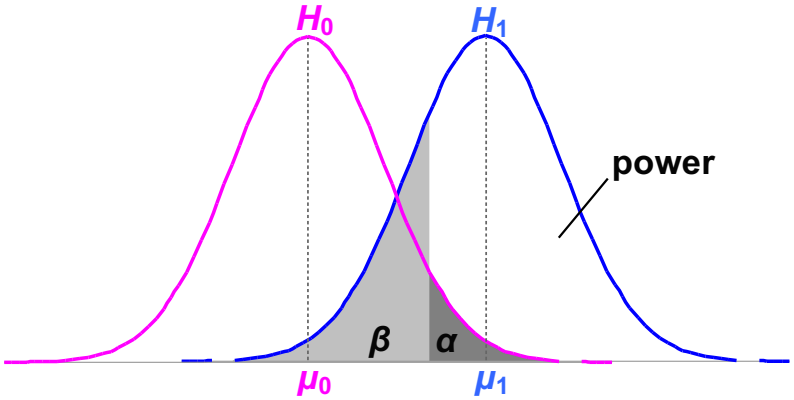
More careful experimental results give lower noise.



Again, effect size remains the same, but variance drops by half.

Increasing Power by Measuring a Bigger Effect





- If the effect size is large:
  - Power increases
  - Type II error decreases
  - α and type I error stay the same

Unsurprisingly, large effects are easier to detect than small effects

Using Power

- Need α, effect size, and sample size for power:
  power = $f(\alpha, |\mu_0 - \mu_1|, N)$

- Problem for VR / AR:
  - Effect size $|\mu_0 - \mu_1|$ hard to know in our field
    - Population parameters estimated from prior studies
    - But our field is so new, not many prior studies
  - Can find effect sizes in more mature fields

- Post-hoc power analysis:

  $$\text{effect size} = |X_0 - X_1|$$

  - Then, calculate power for experiment
  - But this makes statisticians grumble
    (e.g. [Howell 2002] [Cohen 1988])
  - Same information as *p* value

Other Uses for Power

1. Number samples needed for certain power level:

$$N = f(\text{ power}, \alpha, \ |\mu_0 - \mu_1| \text{ or } |X_0 - X_1| \ )$$

   – Number extra samples needed for more powerful result
   – Gives "rational basis" for deciding N
   – Cohen [1988] recommends $\alpha = 0.05$, power = 0.80

2. Effect size that will be detectable:

$$|\mu_0 - \mu_1| = f(N, \text{ power}, \alpha)$$

3. Significance level needed:

$$\alpha = f(\ |\mu_0 - \mu_1| \text{ or } |X_0 - X_1|, N, \text{ power })$$

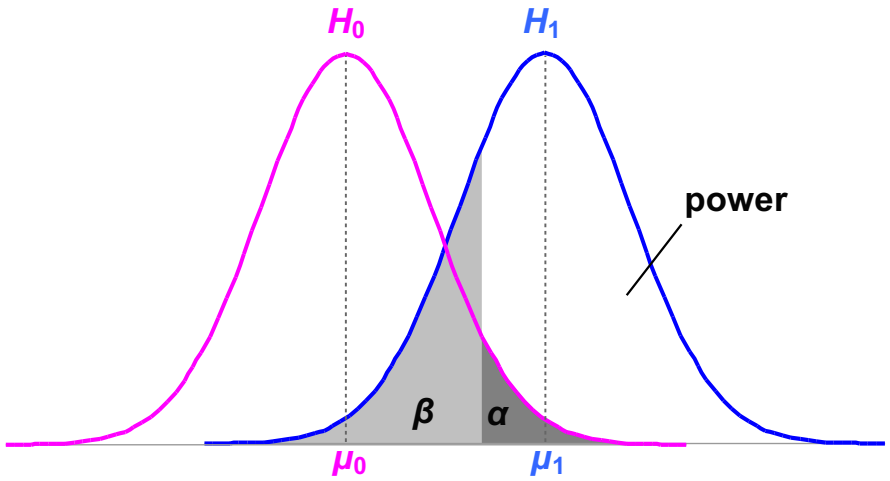
(1) is the most common power usage

[Cohen 1988]

Arguing the Null Hypothesis

Cannot directly argue H_0: $\mu_s - \mu_m = 0$.
But we can argue that $|\mu_0 - \mu_1| < d$.

- Thus, we have bound our effect size by d.
- If d is *small*, effectively argued null hypothesis.
- Cohen recommends $\alpha = 0.05$, power = 0.20



[Cohen 1988, p 16]

Scientific perspectives on GMT (INFOMSCIP)

Statistics

- Descriptive statistics
  Basics (mean, variance, …), graphical visualization
- Inferential statistics
  Hypothesis testing, significance tests.
- **Problems, issues, challenges**
  The replication crisis, p-hacking, biases

⇒ ***To be continued next time …***