Quantitative Evaluation

CS160: User Interfaces

John Canny

Assignments Reminder

- Contextual Inquiry due today
- Low-Fi Prototype out today, due in two weeks
- PPA2 due next Monday

Topics

- Studies with subjects - Ethics
- Designing controlled experiments with subjects
- Basic Stats

Quantitative Studies

Quantitative

Make measurements on interfaces to e.g. determine which is more effective under some measure

Determine whether **apparent** differences are **significant** (i.e. probably reproducible) vs. random.

Approach

Figure out what is important to measure and how to measure it:

- Time, errors, number of keystrokes, mouse gestures,…

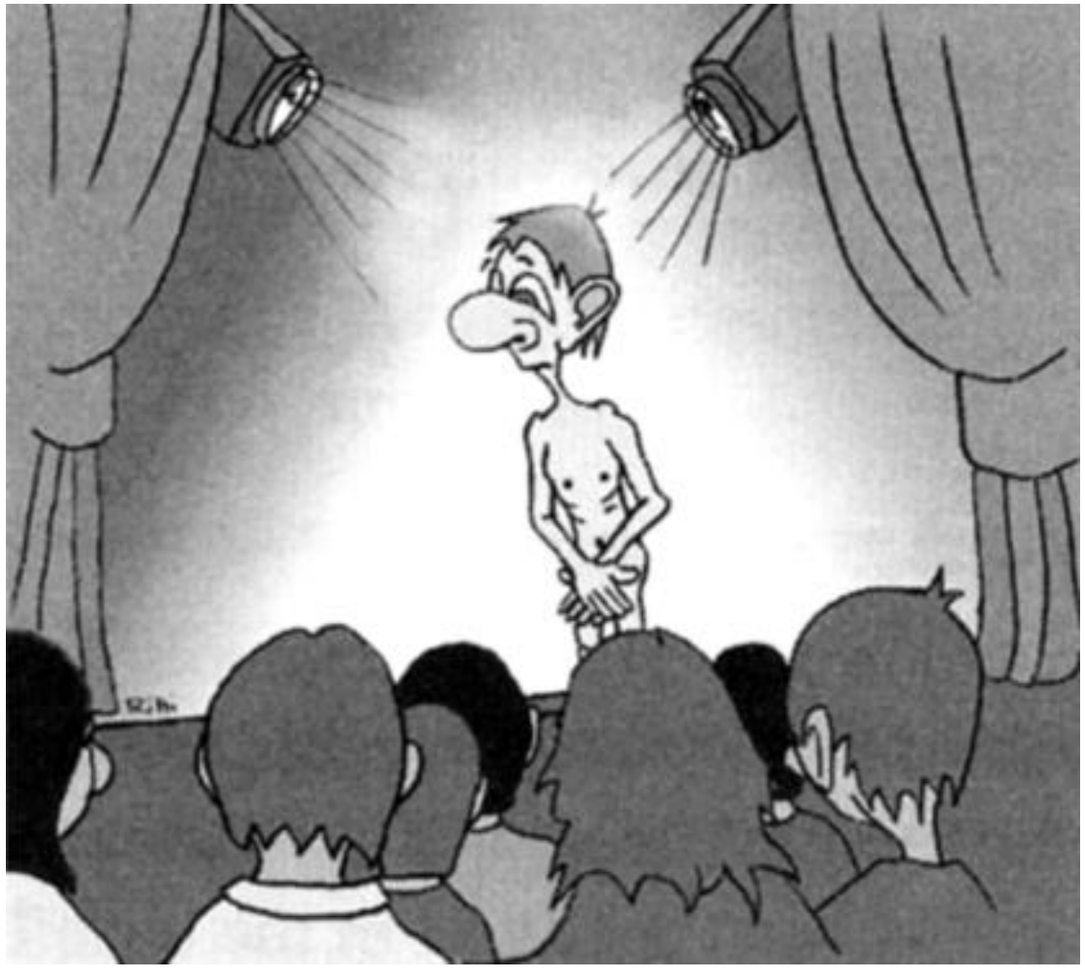
How to **control** for other effects which might influence the results.

Studies with real users - Managing Study Participants

The Participants' Standpoint

Testing is a distressing experience

- Pressure to perform
- Feeling of inadequacy
- Looking like a fool in front of your peers, your boss, …



(from "Paper Prototyping" by Snyder)

Ethics: Stanford Prison Experiment

1971 Experiment by Phil Zimbardo at Stanford

- 24 Participants – half prisoners, half guards ($15 a day)
- Basement of Stanford Psychology bldg turned into mock prison
- Guards given batons, military style uniform, mirror glasses,…
- Prisoners wore smocks (no underwear), thong sandals, pantyhose caps

Experiment quickly got out of hand

- Prisoners suffered and accepted sadistic treatment
- Prison became unsanitary/inhospitable
- Prisoner riot put down with use of fire extinguishers
- Guards volunteered to work extra hours

Zimbardo terminated experiment early

- Grad student Christina Maslach objected to experiment
- Important to check protocol with ethics review boards

[from Wikipedia]

Ethics

Was it useful?

- "…that's the most valuable kind of information that you can have and that certainly a society needs it" (Zimbardo)

Was it ethical?

- Could we have gathered this knowledge by other means?

http://www.prisonexp.org/slide-42.htm

The Three Belmont Principles

- **Respect for Persons**
  - Have a meaningful consent process: give information, and let prospective subjects freely chose to participate
- **Beneficience**
  - Minimize the risk of harm to subjects, maximize benefits
- **Justice**
  - Use fair procedures to select subjects (balance burdens & benefits)

To ensure adherence to principles, most schools require Institutional Review Board (IRB) approval of research involving human subjects.

Treating Subjects With Respect

Follow human subject protocols

- Individual test results will be kept confidential
- Users can stop the test at any time
- Users are aware (and understand) the monitoring technique
- Their performance will have not implication on their life
- Records will be made anonymous if possible
  - Video face blurring (Youtube), Audio distortion

Use standard informed consent form

- Especially for quantitative tests
- Be aware of legal requirements

Privacy and Confidentiality

- **Privacy**: having control over the extent, timing, and circumstances of sharing oneself with others.

- **Confidentiality**: the treatment of information that an individual has disclosed with the expectation that it will not be divulged

- Examples where privacy could be violated or confidentiality may be breached in HCI studies?

Beneficience: Example

- MERL DiamondTouch:
  - User capacitively coupled to table through seating pad.
  - No danger for normal users, but possibly increased risk for participants with pacemakers.
  - Inform subjects in consent!

http://www.merl.com/projects/images/DiamondTouch.jpg

Justice

- Subjects in the study should accurately reflect the target population
- Target population may include:
  - Men and Women
  - People will disabilities
  - Left-hand people
  - Elders
  - Non-native speakers
  - Children
  - Color-blind people
- E.g. if you don't include subjects from all the target populations, you wont be able to discover features of the design that are difficult or impossible for them to use.
- Also avoid excessive burden on an easily-available population (e.g. fellow students) to balance the load.

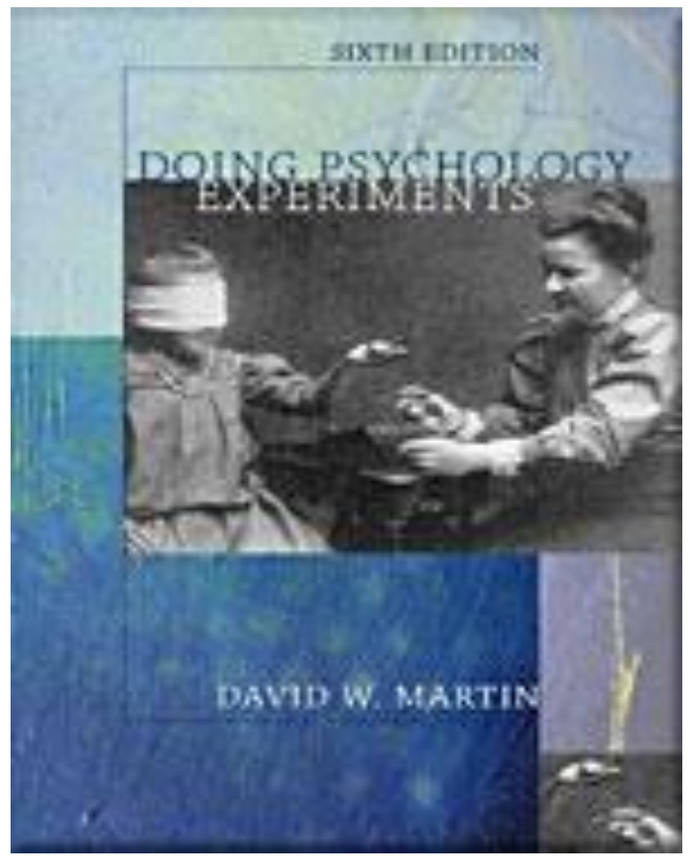
Conducting the Experiment

- Before the experiment
  - Have them read and sign the consent form
  - Explain the goal of the experiment in a way accessible to users
  - Be careful about **demand effects**
    (Participants biased towards experimenter's hypothesis)
  - Answer questions
- During the experiment
  - Stay neutral
  - Never indicate displeasure with users performance
- After the experiment
  - Debrief users (Inform users about the goal of the experiment)
  - Answer any questions they have

If you want to learn more…

- Online human subjects certification courses:
  - E.g., http://phrp.nihtraining.com/users/login.php
- The Belmont Report: Ethical Principles and Guidelines for the protection of human subjects of research
  - 1979 Government report that describes the basic ethical principles that should underly the conduct of research involving human subjects
  - http://archive.org/details/belmontreporteth00unit

Designing Controlled Experiments



Doing Psychology Experiments
David W. Martin

Steps in Designing an Experiment

1. State a lucid, testable hypothesis
2. Identify variables (independent, dependent, control, random)
3. Design the experimental protocol
4. Choose user population
5. Apply for human subjects protocol review
6. Run pilot studies
7. Run the experiment
8. Perform statistical analysis
9. Draw conclusions

Experiment Design

- **Independent variable:**
  Something you control, the condition you want to vary to see what affect it has: e.g. age of users. Is often a *discrete* variable: e.g. several interface designs.
- **Dependent variable:**
  Something you measure, like completion time, number of errors, user survey results.
- **Hypothesis:**
  What you believe will be true about the influence of independent variables on dependent variables: one design will be faster than others etc.

Experiment Design

- **Control variables**
  - Attributes that will be fixed throughout experiment

Control variables help deal with confounds – attributes that can effect DVs but are not modeled, e.g. subject's fluency with video games. Instead of letting this vary, you fix it in the subject selection process.

- **Random variables**
  - Attributes that you do not deliberately vary (IV) or fix (CV).
  - Usually intended to model the population realistically.

- Note, you can often improve the analysis by including RV labels – e.g. male/female, age (in decades), education level…

Common Dependent Variables in HCI

- Performance metrics:
  - Task success (binary or multi-level)
  - Task completion time
  - Errors (slips, mistakes) per task
  - Efficiency (cognitive & physical effort)
  - Learnability
- Satisfaction metrics:
  - Self-report on ease of use, frustration, etc.

Satisfaction Metric: Likert Scales

- Respondents rate their level of agreement to a statement

"Overall, I am satisfied with the ease of completing the tasks in this scenario"

1: Strongly Disagree
2: Disagree
3: Neither agree nor disagree
4: Agree
5: Strongly agree

Choosing Subjects

- Pick balanced sample reflecting intended user population
  - Novices, experts
  - Age group
  - Sex
  - ….

- Example
  - 12 non-colorblind right-handed adults (male & female)

- Population group can also be an IV or a controlled variable

Example: Multiview

Example: Multiview

- **Independent variable:**
  Form of meeting between groups (face-to-face, normal video-conference, quasi-3D conference).
- **Dependent variable:**
  profit from investment (a trust measure)
- **Hypothesis:**
  Directional (quasi-3D) video will improve trust relative to normal video-conferencing.
- **Secondary Hypotheses:**
  Face-to-face > normal video-conferencing
  Face-to-face >? directional video

Example: Multiview

- **Control variables:** group size, task, duration
- **Random variables:** age, gender, education

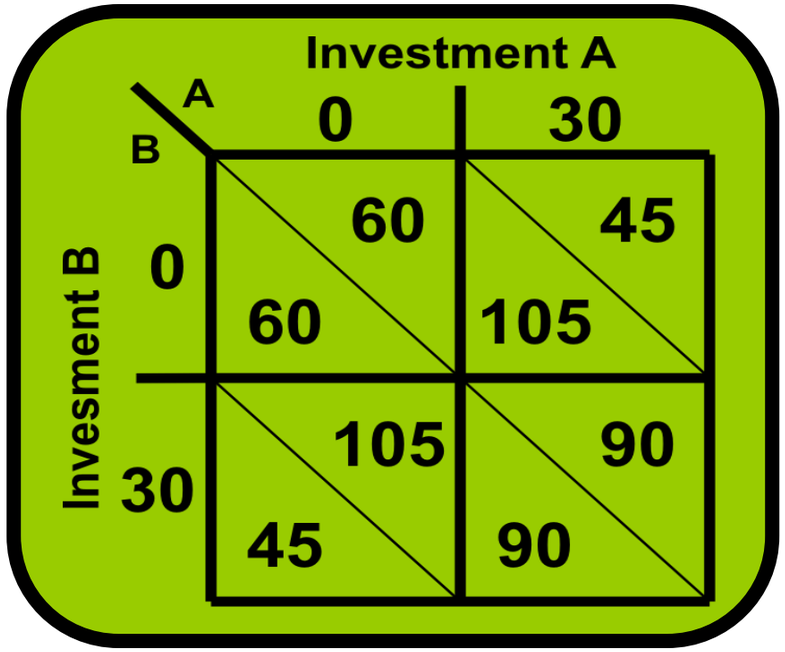
We fix control variables to reduce unnecessary noise in the results – to make the "signal" stronger and easier to verify.

We allow random variables to vary so the result represents reality: what real groups will look like.

Task design

In this game, we design the task as a "prisoner's dilemma" task – teams gain by cooperating but can get short-term gain by defecting. Users have to trust each for max gain.

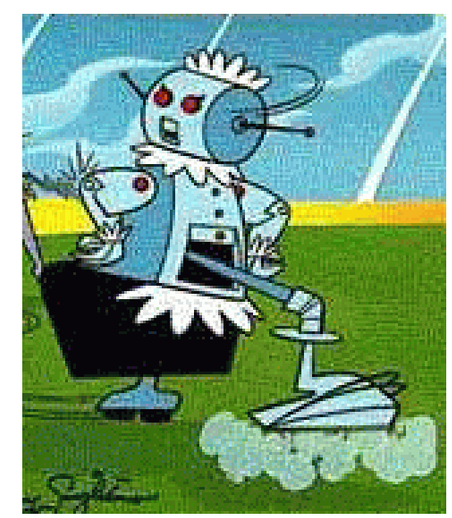
We see how much effect this has in a one-hour session.



Between Subjects Design

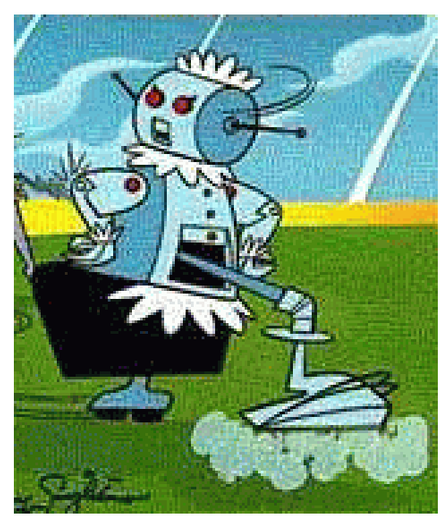
Wilma and Betty use one interface

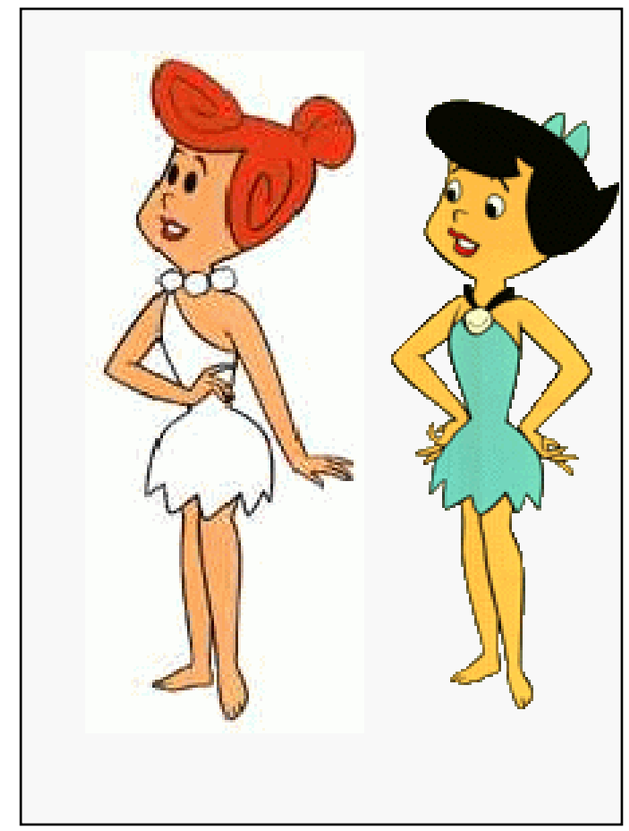
Dino and Fred use the other



Within Subjects Design

Everyone uses both interfaces





Between vs. Within Subjects

Between subjects

- +/- Participants cannot compare conditions
- + Can collect more data for a given condition
- - Need more participants

Within subjects

- + Compare one person across conditions to isolate effects of individual diffs
- + Requires fewer participants, possibly less overall time
- - Fatigue effects
- - Bias due to ordering/learning effects

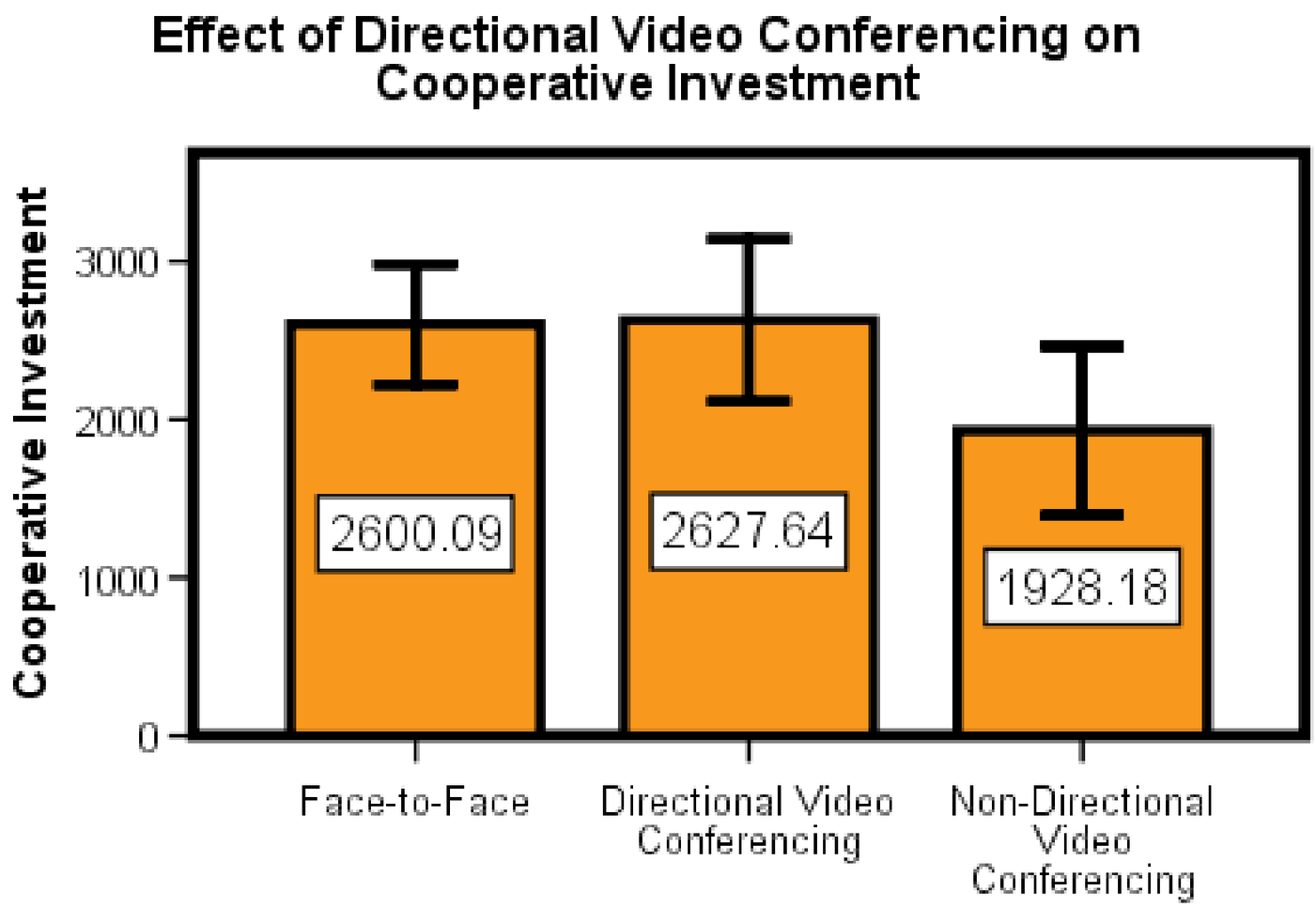
Between vs. Within Subjects

Often the choice is forced:

- If the task is time-consuming, each subject will only be able to complete one condition. Between-subjects is necessary (true for Multiview).
- If the task is short, filling a one-hour session will consume several conditions. Within-subjects is natural.

Example: Multiview

Result:



But could this be due to chance? Stay tuned…

Statistics without Tears

Statistics means never having to say you're certain
– Phil Stark

You can “prove” certain statements with near-certainty under strong assumptions. But you can also “prove” non-facts when those assumptions are violated.

We’ll concentrate on doing the easy cases well.

Hypothesis

Most experiments in HCI and social science make use of "inferential statistics."

Such methods don't directly provide support for a hypothesis. Instead they provide evidence *against* a **null hypothesis**.

Null Hypothesis: Something that must be false if the hypothesis is true, e.g. no difference between control and treatment groups.

Null Hypothesis

e.g. for the hypothesis interfaceA faster than interfaceB, the null hypothesis would be that *the times are the same*.

Note: refuting the null hypothesis typically **does not prove the hypothesis**.

Anything else, however unlikely, that could cause the measurement difference could be the real explanation. Standard tests don't consider any of these situations.

Variable types

Categorical Variables: {hair color}, {conservative, liberal}, - {set of buttons the user might click} – no natural ordering.

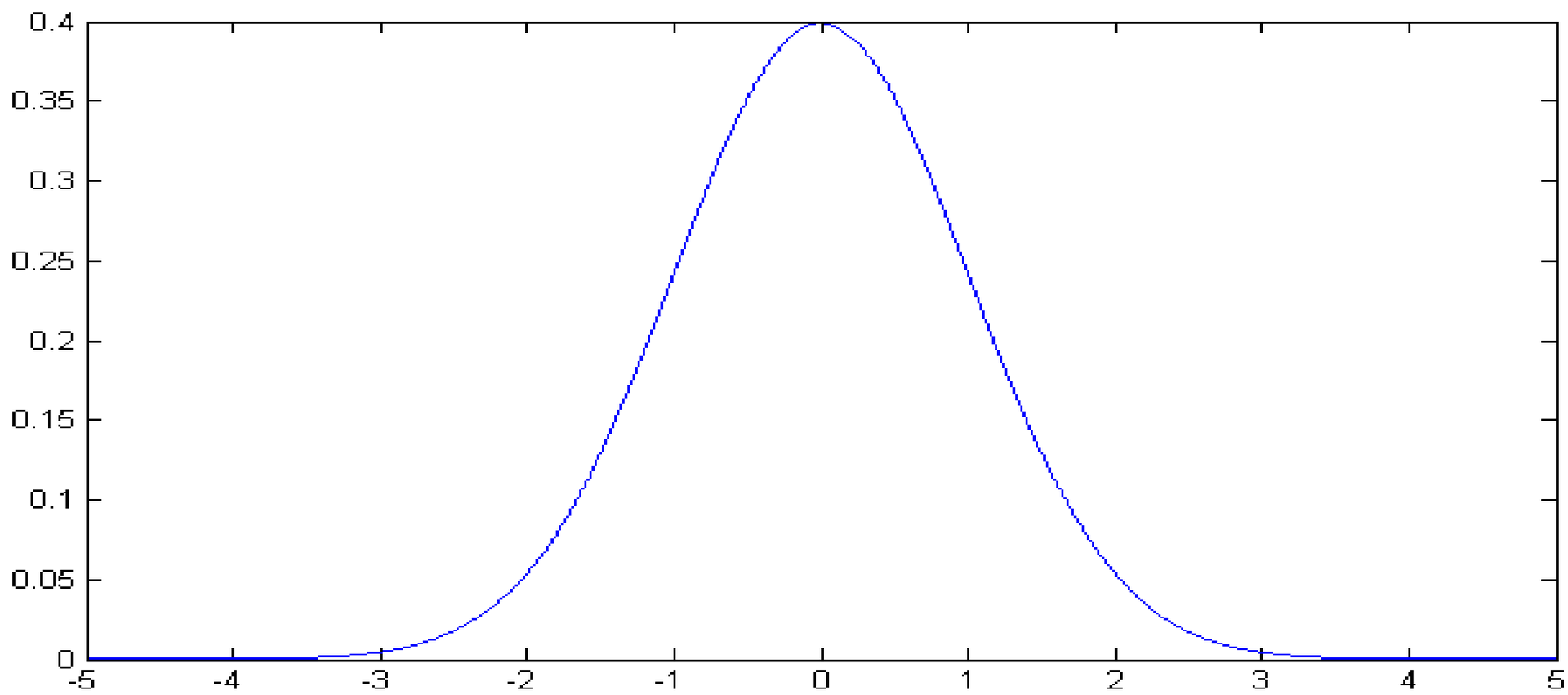
Ordinal Variables: Have a natural order, e.g. {XS, S, M, L, XL}, but for which the sizes/differences are not constant.

Interval Variables: Ordered variables where the intervals between groups are equal. E.g. income $20k-30k, $30k-40k, $40k-$50k

Distributions

For ordered variables, what really matters is the distribution of the variable, i.e. the probability it lies in a range of values.

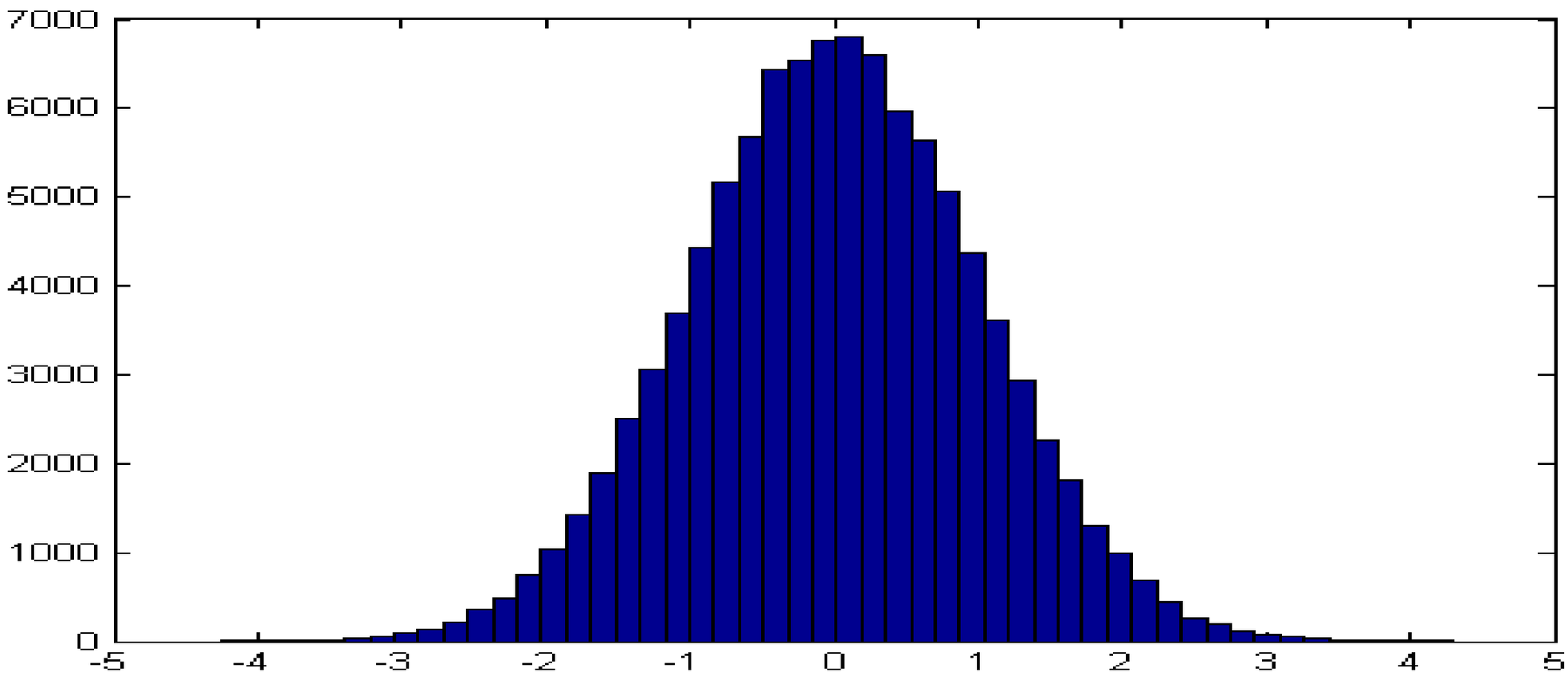
The figure below is a probability density function (pdf) which is the limit of the probability the variable lies in an interval, divided by its size.



Distributions

For ordered variables, what really matters is the distribution of the variable, i.e. the probability it lies in a range of values.

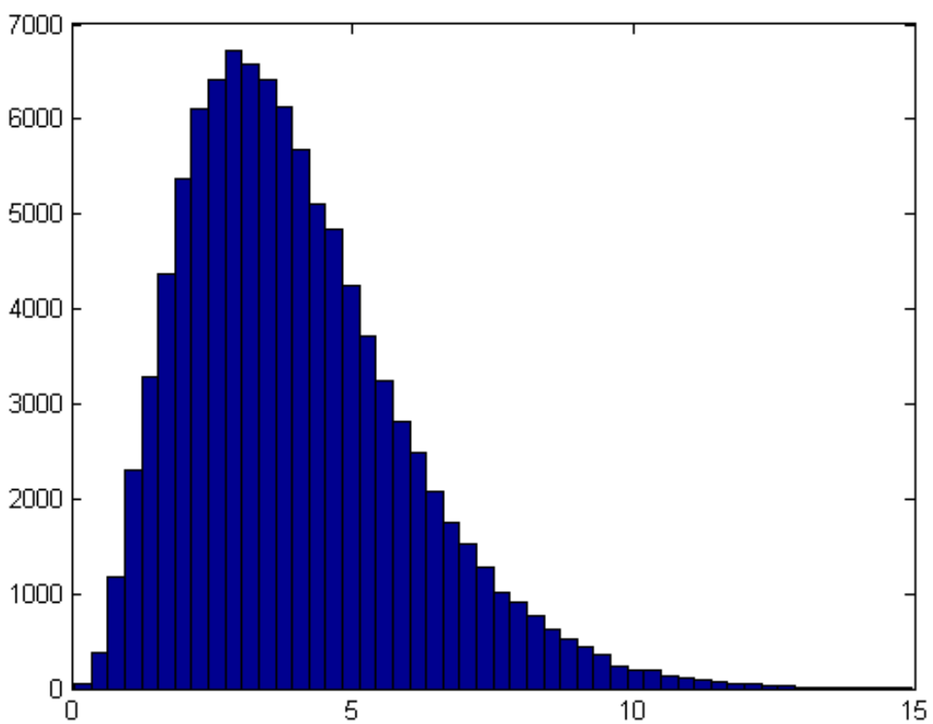
We can approximate that with a **Histogram**, which counts how many samples have values in a given range.

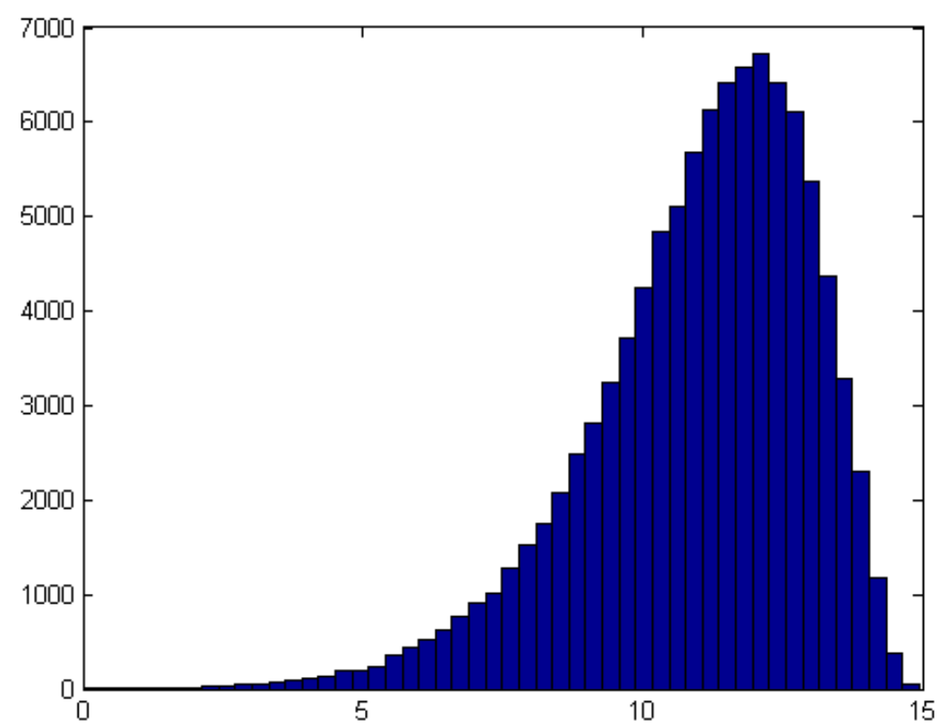


Mean and Median

Recall that the **mean** of a set of values is the numerical average. The **median** is the element in the middle of the sorted list of elements.

What is the relationship between mean and median for these examples?



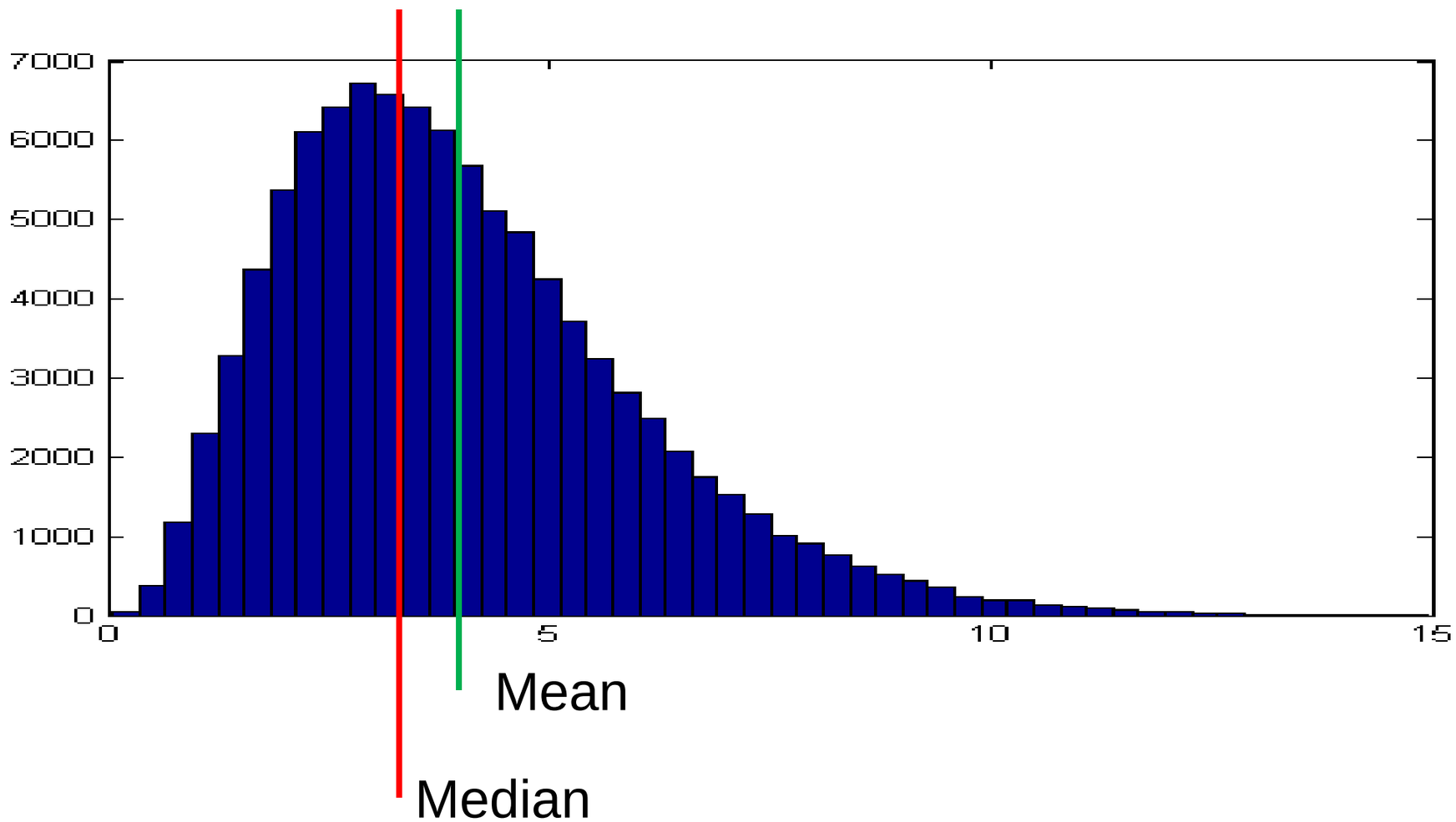


These distributions are *skewed*

Mean and Median

The median keeps equal numbers of elements (equal curve areas) on either side. It is not influenced by magnitude.

The mean is sensitive to values, the larger the values, the larger the mean. So it will move toward the "tail" of the distribution.

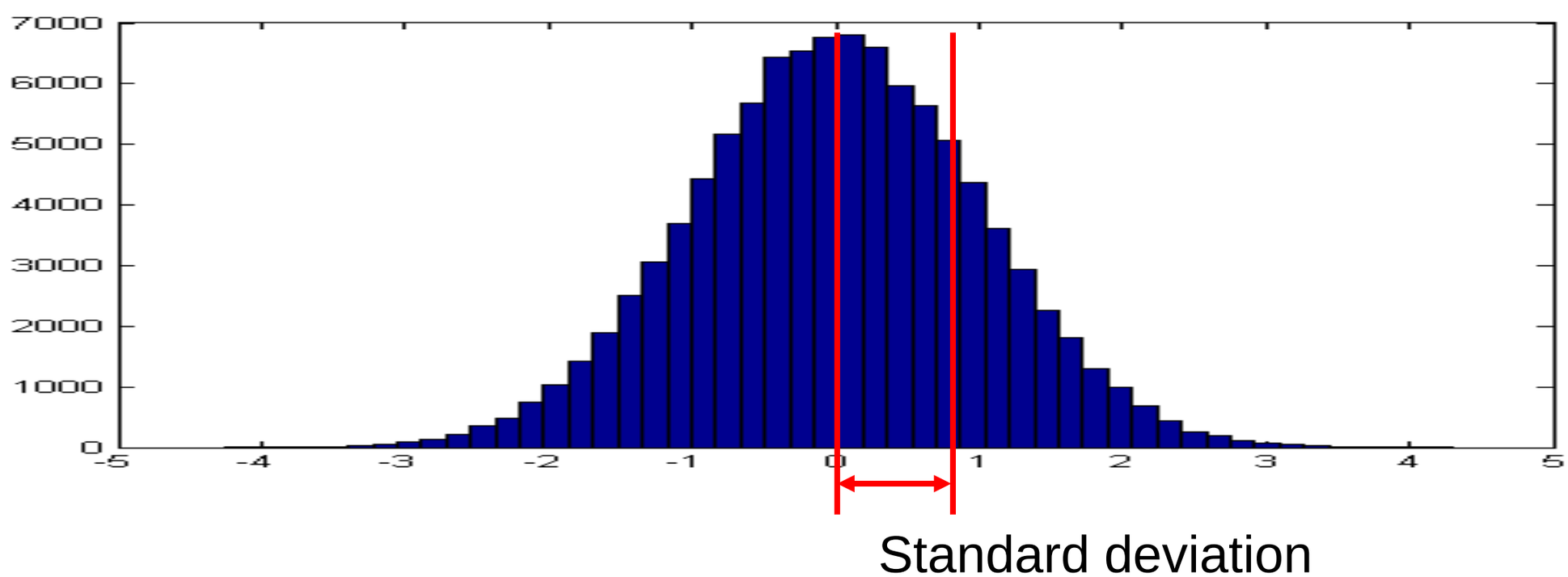


Variance and Standard Deviation

Is a measure of the width of a distribution. Specifically, it is the average squared deviation of samples from their mean:

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

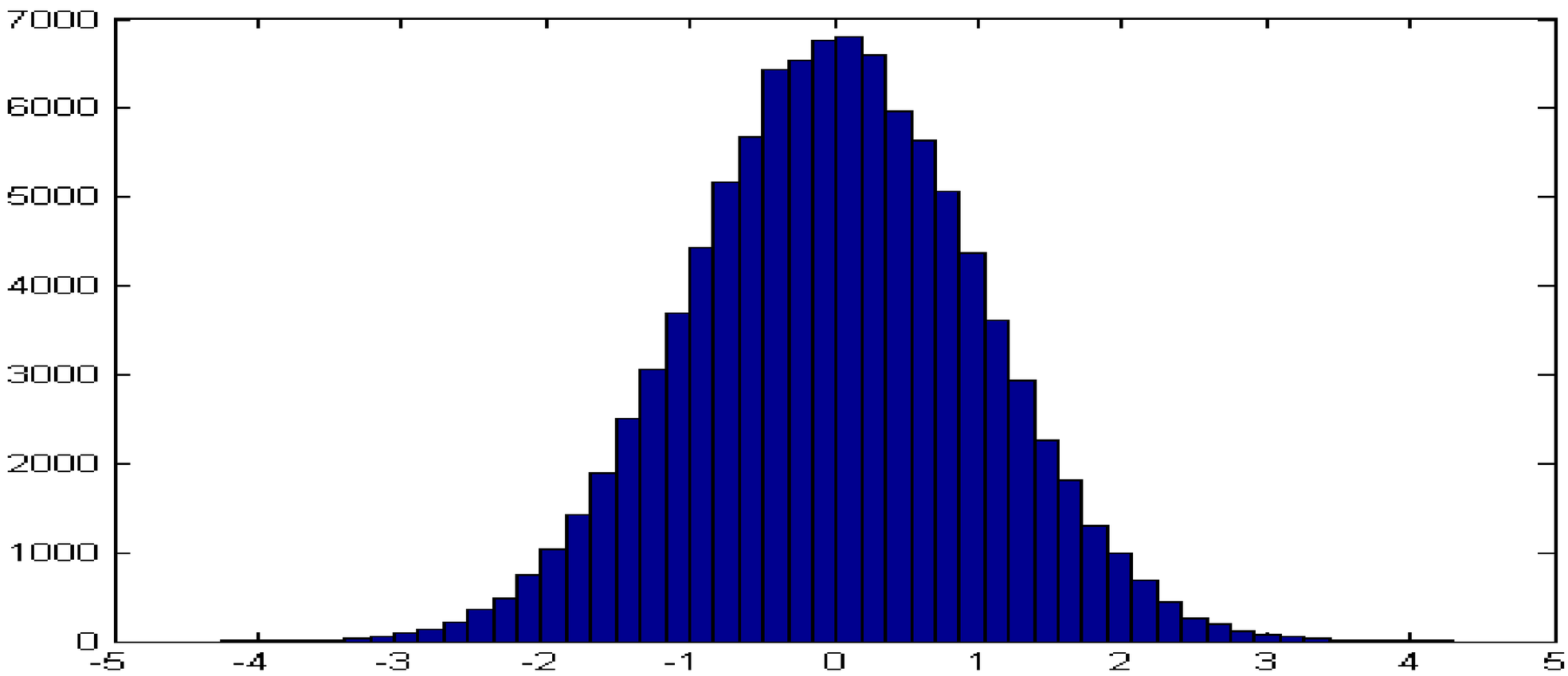
The related quantity called **standard deviation** is the square root of variance and can be used to measure the width of the distribution:



Normal Assumption

For many datasets of continuous or even discrete data, we assume that the data are **normally** distributed.

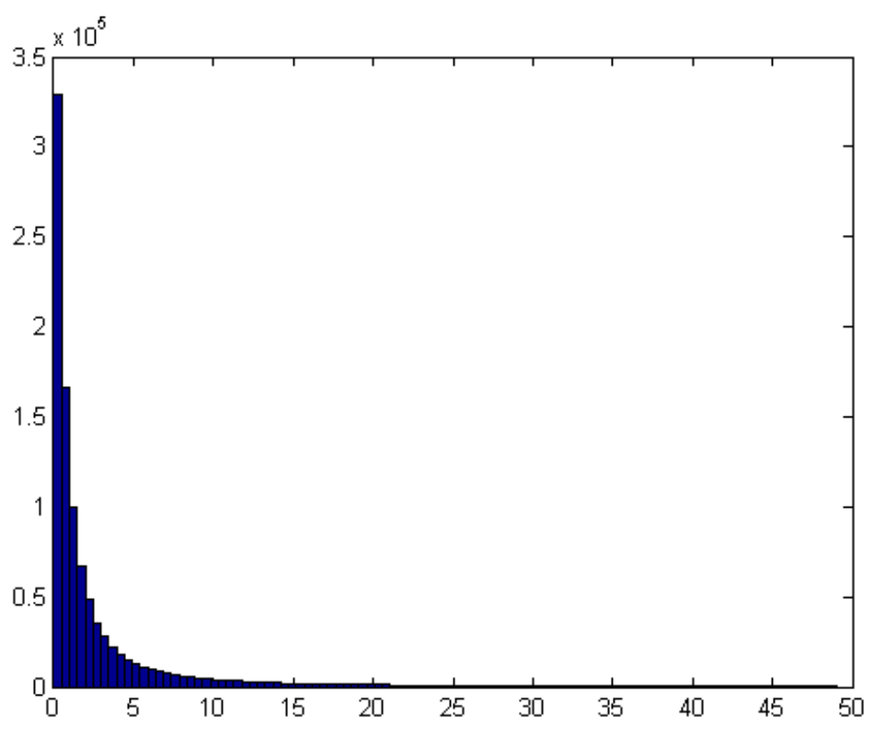
Then we can use only the means and variances of the data, since a normal distribution is completely described by mean and variance.



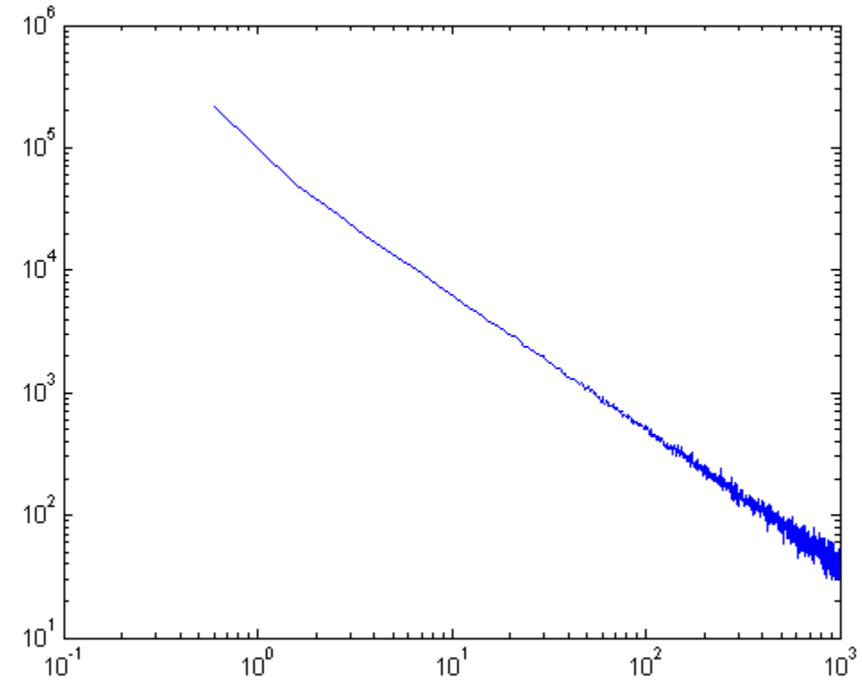
Long-tailed distributions

Quite a few measurements in HCI and social science exhibit power-law distributions, where p(x) is a negative power of the rank of x, e.g.

- Number of times users visit a web site
- Number of times user types a given word
- Size of friend networks



Sorted (rank) histogram



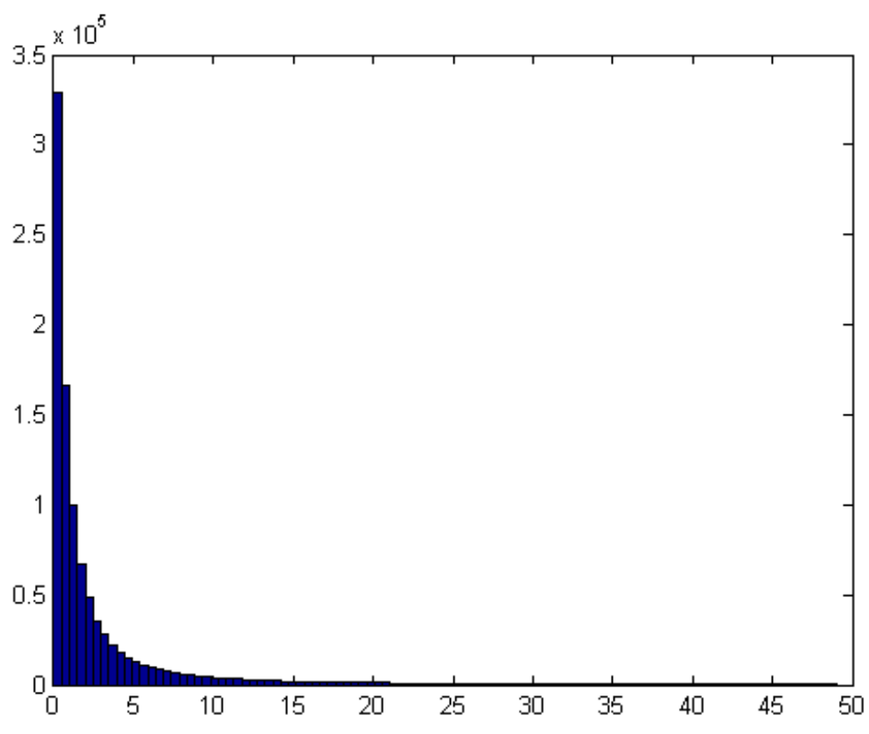
Log-log histogram

Long-tailed distributions

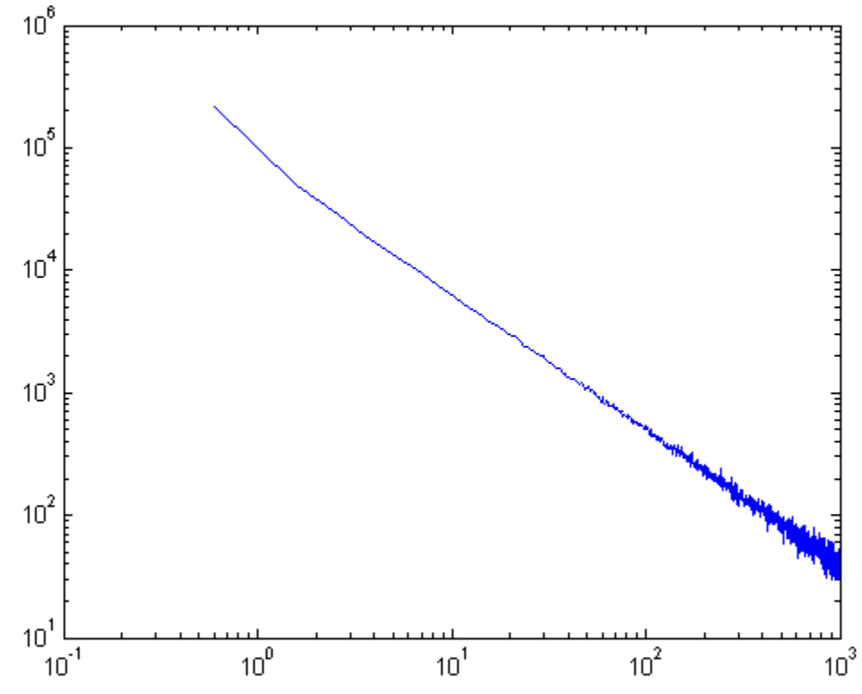
One or both of the mean and variance **may be infinite** for these distributions (Power law or Pareto distributions).

Even if computable, mean and variance will be too unreliable to use.

You will need to reparametrize, or use a non-parametric test, to deal with these types of variables.



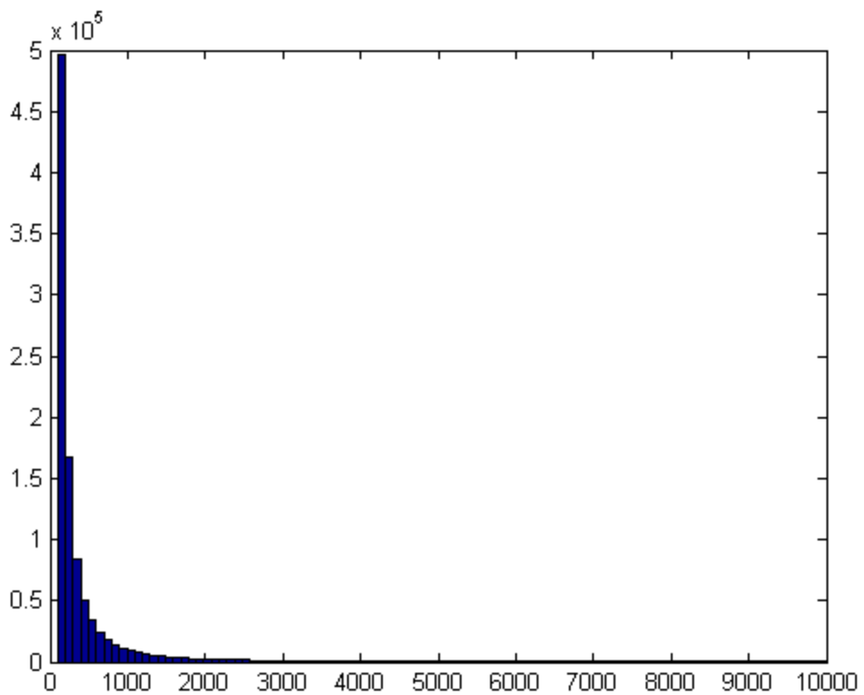
Histogram



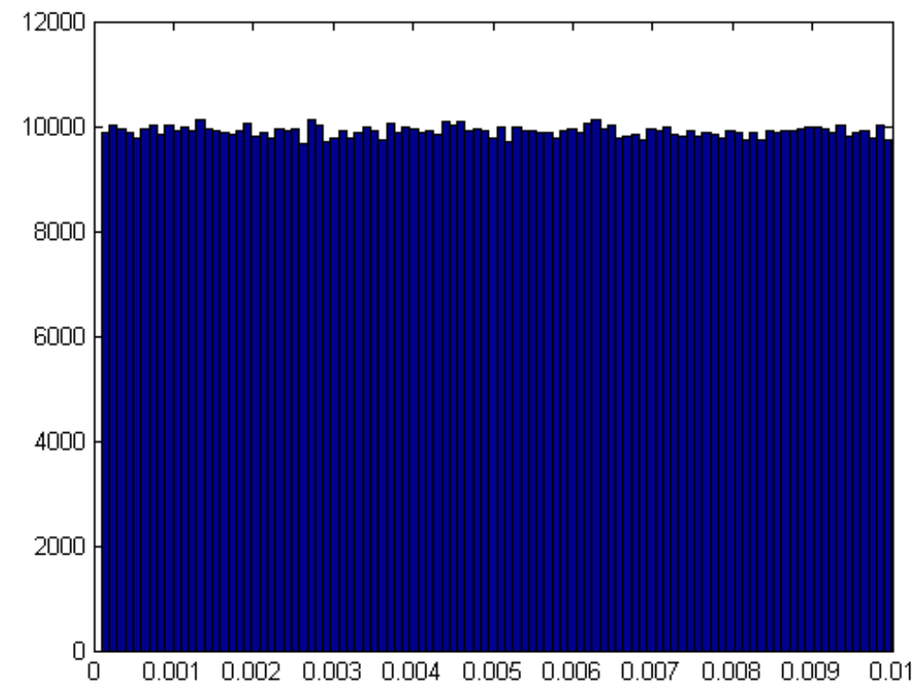
Log-log histogram

Long-tailed distributions

e.g. reparametrize with x0 = 1/x



Histogram of x



Histogram of x0

Other Assumptions

Independence: measured variables are assumed to be sums of **independent,** normal, random variables that represent the effects of the independent variables. E.g Multiview:

$$\textbf{Investment} = \mathbf{S_C} + \mathbf{S_U}$$

Where $\mathbf{S_C}$ is the effect of the condition on investment, and $\mathbf{S_U}$ is the random variation for that user.

For Multiview, n is the number of **groups**, not users.

A simple within-subjects test

Suppose we have just one independent variable with two levels (two discrete values), and one dependent variable.

Suppose the design is within-subjects, then we can subtract for each user the scores for condition_B from condition_A, i.e. $s = s_B - s_A$.

Per-subject variation will be eliminated this way. This is a **"paired-sample"** test.

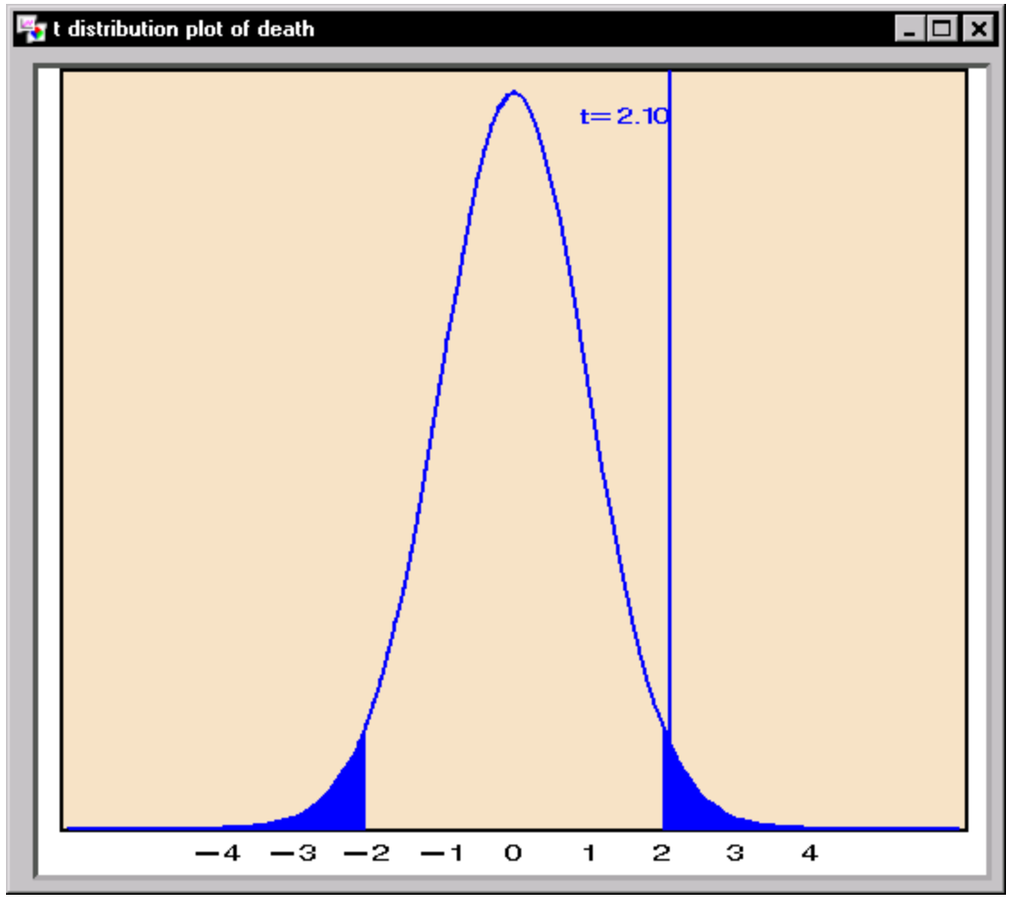
We obtain a list of differences, one per user. Under the null hypothesis (no difference between conditions), this list should have mean zero.

Paired or One-sample t-test

There is a black-box test to determine the probability that a set of values came from a normal distribution with mean zero.

It is the single-sample (or paired-sample) t-test.

In fact it needs only the mean and variance of the sample.



Paired or One-sample t-test

The t-statistic is defined as:

$$t = \sqrt{n}\frac{\bar{X}}{s}$$

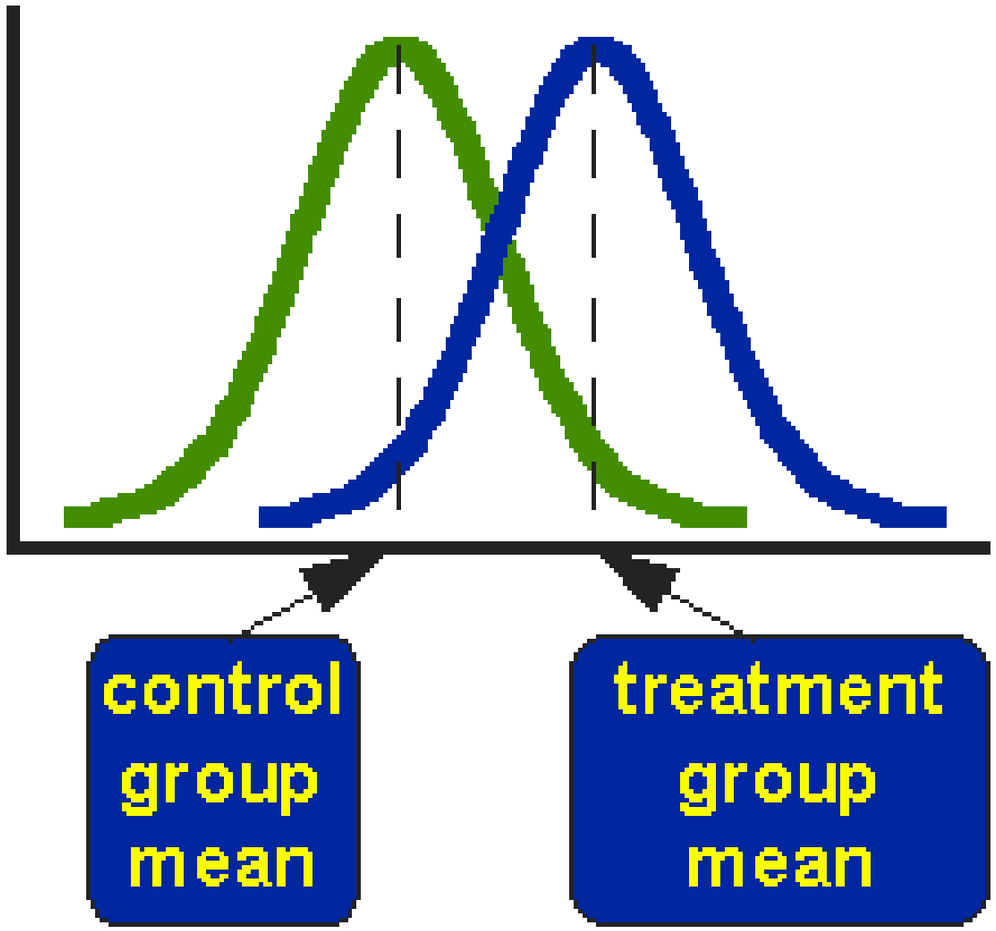
Where $\bar{X}$ is the sample mean, s is the sample standard deviation, and n is the number of samples.

The distribution of this statistic depends on the number of *degrees of freedom*, which is n-1.

Two-sample t-test

If we have **between-subjects** data, we can still use the method just described, but it wont be very effective. If we take differences between two different people, their random variation will tend to mask systematic effects.

A **two-sample t-test** tests just what we want: whether two distinct samples come from the same distribution.



t-statistic

The t-statistic was invented by William Sealy Gosset, A Chemist working for Guinness Breweries, in 1908.

Gosset had to publish under the pseudonym “Student”, hence “Student’s t-test”

t-statistic

We gave the t-statistic earlier for a single sample.

The ***two-sample*** statistic is: $t = \dfrac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$

Where $S_{X_1X_2} = \sqrt{\dfrac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$

And $S_{X_1X_2}$ is the pooled standard deviation for the two samples.

You compute the t-statistic for your experiment and then find the p-value from a table (or Matlab or SPSS).

Sensitivity and Experiment Size

The t-statistic value generally increases with the number of independent measurements (number of subjects) n.

Large values of t refute the null hypothesis, so it is **easier for your test to succeed the more measurements you make**.

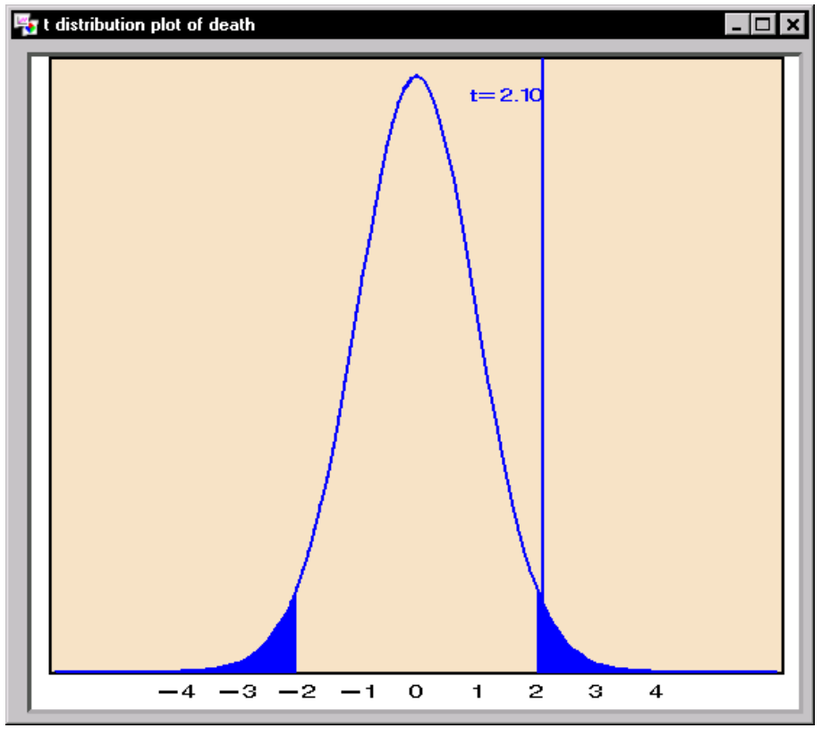
Statistic(s): the core of Statistics

Virtually every statistical test uses a *statistic*.

A statistic is a real-valued function of the observations.

Since most observations are unique, their probability is close to zero.

A statistic measures deviation from the norm, and allows us to measure the probability of values ***at least as large*** as the observed value.



Discrete Data

For discrete ordinal data (usually count data), other methods are preferred. They include:

Fisher's exact test: for 2x2 *contingency (count) tables* such as those for "method A produced more errors than method B".

CHI-squared statistic. Its distribution allows approximate significance testing on count data.

Permutation tests (see later).

Significance – a line in the sand

Hypothesis testing is a **probabilistic** process.
It will never tell you "X is true" or "X is false."

So researchers have come to declare that certain probabilities represent "statistically significant" effects.

Significance: is an a-priori determined probability σ, such as 0.05 or 0.01, such that when Pr(Observation | Null Hypothesis) < σ, the result can be declared to be "statistically significant".

P-values

Both t-tests produce probabilities Pr(Observation | Null Hypothesis) that we can check against the significance threshold to see if we can call our results "statistically significant."

This Pr(Observation | Null hypothesis) is called a **p-value**.

Testing errors

"Statistically significant" outcomes will happen by chance, even when the null hypothesis is true, at a rate given by the p-value.

i.e. for p = 0.05, in 1/20 experiments in which the null hypothesis is true, a positive test will result, and the null hypothesis will be rejected.

This is called a **type-I error**. These are serious. You concluded something was true that may not be true.

If an experiment fails to reject the null hypothesis when its false, there is a **type-II error**. These are inconvenient, but less destructive, you haven't "proved" a falsehood.

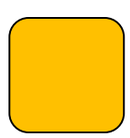
Publication Bias

Many outlets (journals, conferences) prefer to publish significant results rather than tests that were not.

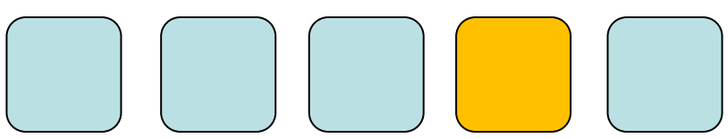
Authors themselves tend not to submit non-significant results.

What's wrong with this?

E.g. suppose for every published result significant at 0.05, there were 4 other experiments that were not?



probability of success by chance = 1/20 = 0.05



probability of success by chance = 1/4 =0.25

Avoid Many Comparisons

Each time you try something, you have another chance of a false positive.

e.g. if you have 6 conditions, there are (6 2) = 15 pairs of conditions to test, and one will very likely be significant by chance.*

So concentrate on your main (most important) hypothesis and test that first.

* The probability of this is less than 15*0.05 because the tests are not independent

If you don’t succeed at first…

Here are some typical p-values for a borderline-significant t-test (median p-value is 0.05) on different randomly selected groups of 20 subjects:

0.0019	0.2891	0.0429	0.0095	0.0078	0.0427
0.0433	0.5866	0.0593	0.0100	0.0015	0.0487

Some of these values would be considered “extremely significant”, while others not at all.

If you believe the result, try the experiment again!

Consider more subjects, but do cost-benefit analysis.

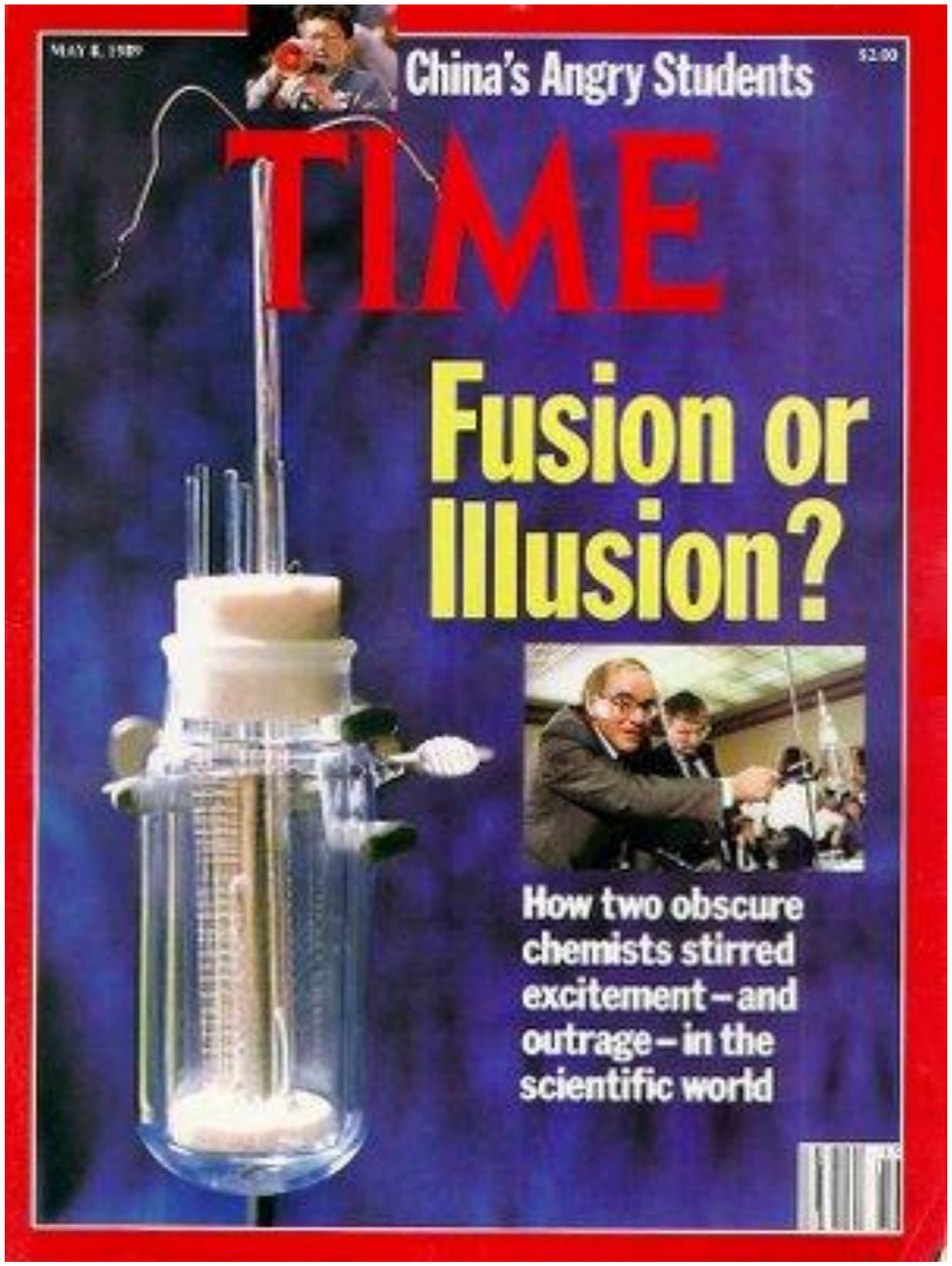
Biggest reporting mistake

If a test does not produce a significant effect, e.g. $p = 0.1$, it does not mean the original hypothesis doesn't hold, just that the **experiment failed to demonstrate a strong enough result**.

Avoid saying "there was no significant difference between A and B."

Very often there will be, and you will find it if you do the experiment again. Remember $p = 0.05$ is just an arbitrary convention, and p-values from real experiments vary all over the map.

Testing errors



Other problems

The New York Times

Research

Search All NYTimes.com Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

RESEARCH | FITNESS & NUTRITION | MONEY & POLICY | VIEWS | HEALTH GUIDE

Search Health 3,000+ Topics

Go

For First Time, AIDS Vaccine Shows Some Success

By DONALD G. McNEIL Jr.
Published: September 24, 2009

Scientists said Thursday that a new AIDS vaccine, the first ever declared to protect a significant minority of humans against the disease, would be studied to answer two fundamental questions: why it worked in some people but not in others, and why those infected despite vaccination got no benefit at all.

The vaccine — known as RV 144, a combination of two genetically engineered vaccines, neither of which

SIGN IN TO RECOMMEND
TWITTER
COMMENTS (33)
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE
SHARE

Well
Tara Parker-Pope on Health

Tips for Navigating Medicare
October 16, 2009

Show Off Your Vegetables With Pasta
October 16, 2009

High-Deductible Health Plans: Better for You or Your Employer?
October 16, 2009

The Roving Runner: Prospect Park
October 16, 2009

Alternative Medicine Cabinet: Thyme for Toenails
October 15, 2009

TicketWatch - Theater Offers by E-Mail

Sign up for ticket offers from Broadway shows

RV144 Aids vaccine

A huge controlled study (16,000 volunteers, $110M) of an AIDS vaccine in Thailand found that the vaccine had a significant effect on subjects ($p = 0.045$).

By itself, this sounds like a major success.

But the difficulty lies in the context: this is one of many AIDS vaccine trials (> 30) that are underway or completed. The probability of a 0.05-significant result in one of these studies assuming all vaccines are ineffective, is 80%.

Multiple Comparisons

If there are many tests, the significance level should be lowered to make sure that results are not just due to chance.

Bonferroni discounting reduces the significance threshold exactly by the number of experiments.

e.g. if you have 10 experiments, you should use a significance threshold of 0.05/10 = 0.005 for each one.

This guarantees that the probability of a type-I error in the collection of experiments is less than the significance threshold.

Complex Experiment Designs

What if there are more than two values for the independent variable, or more than one independent variable?

The simplest approach is to conduct many paired t-tests and apply Bonferroni correction. However, this approaches weakens the power (sensitivity) of the test.

If there are k levels of a random variable, that means $\binom{k}{2}$ pairs to test, and significance thresholds have to be lowered by that amount.

Complex Designs (Between Subjects)

What do we do if we have more than two levels of the independent variable, or more than one variable?

In a **between-subjects** design, the answer is straightforward. We can still represent each subject's score as a sum of components due to the independent variables, plus individual variation. The analysis method is called:

ANOVA: Analysis of Variance. Allows tests of statistical effects of any one variable, or group of variables.

ANOVA

Single factor analysis of variance (ANOVA)

- Compare means for 3 or more levels of a single independent variable (2 reduces to a t-test).

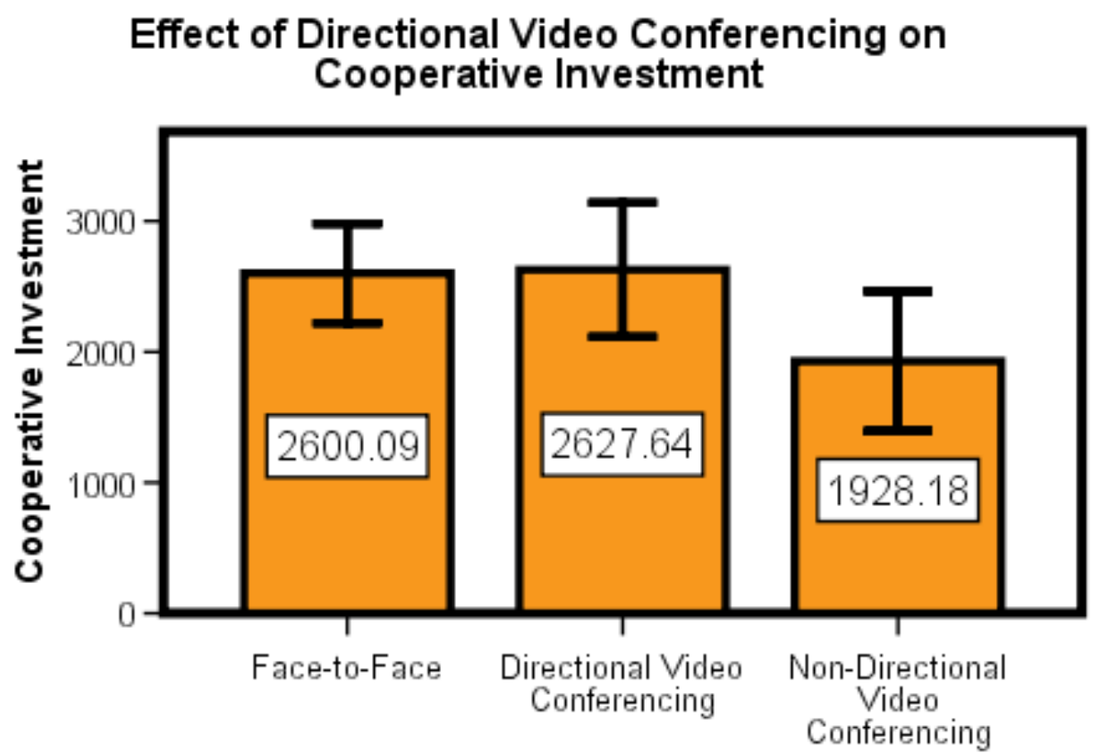
Multi-Way Analysis of variance (n-Way ANOVA)

- Compare more than one independent variable
- Can find interactions between independent variables

ANOVA tests whether means differ, but does not tell us ***which*** means differ – for this we must perform pairwise t-tests

ANOVA example

Multiview: Single between-subjects variable (Factor), the kind of interaction: Face-to-Face, Directional Video, Non-directional video:



The null hypothesis is that all means are the same. An ANOVA test determines they ***are not***, but does not tell us how.

Complex Designs (Within Subjects)

MANOVA, or Multivariate analysis. Which treats all the measurements made on each subject as distinct variables.

It then discovers the correlations between variables and uses a statistic that takes those into account.

MANOVA is more complex to understand, but is a safer black box than RM-ANOVA.

Core Concepts

- Variables – independent, dependent, control, random
- Data distributions: skew, mean, median, variance
- Hypothesis – Initial and then a null hypothesis
- Test statistic to measure “how unusual” the data are
- Significance – probability of type I errors
- P-values – probabilities derived from the statistic

Process

- Make a clear hypothesis before you start.
- Look at sample data before you decide how to test.
- Pick a design that is as simple as possible.
- Make sure you collect all the data you need.
- Commit to the experiment, publish everything.
- If it doesn’t work, consider redoing the experiment.