A Very Brief Intro to Statistics: t-tests

Slides by Ruth Rosenholtz

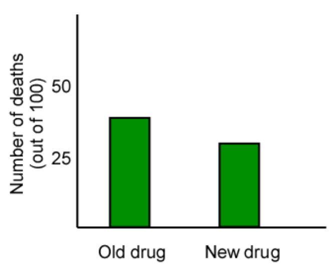
t Test at a glance

$$t = \frac{\text{Difference between groups (means)}}{\text{Normal variability within group(s)}}$$

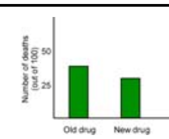
- If *t* is large, the difference between groups is **much bigger** than the normal variability within groups.
  - Therefore, two groups are significantly different from each other
- If *t* is small, the difference between groups is **much smaller** than the normal variability within groups.
  - Therefore, two groups are **not** significantly different from each other

Does a new drug cure cancer better than the old drug?

- The data:



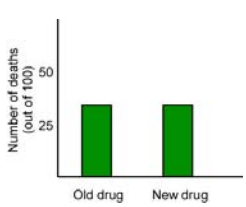
Does a new drug cure cancer better than the old drug?



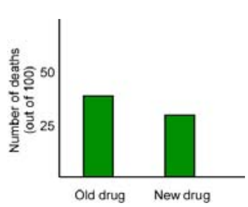
- There's an empirical difference between the old drug and the new drug.
- But is it due to a systematic factor (e.g. the new drug works better) or due to chance?
- If we gave the new drug to 100 more people, would we expect to continue to see improvement over the old drug? Do we expect this effect to *generalize*?

Alt: Is the difference between data & theory due to systematic factors + chance, or to chance alone?

- "Theory" = no difference between the drugs:



- Data:



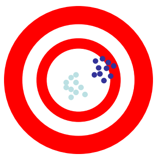
Also called the "null hypothesis"

Chance vs. systematic factors

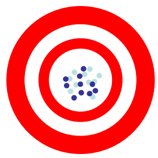
- A *systematic* factor is an influence that contributes a predictable advantage to a subgroup of our observations.
  - E.G. a longevity gain to elderly people who remain active.
  - E.G. a health benefit to people who take a new drug.
- A *chance* factor is an influence that contributes haphazardly (randomly) to each observation, and is unpredictable.
  - E.G. measurement error

Systematic + chance vs. chance alone: Is archer A better than archer B?

- Likely systematic + chance variation:



- Likely due to chance alone:



Observed effects can be due to:

A. Chance effects alone (all chance variation).
  - Often occurs. Often boring because it suggests the effects we're seeing are just random.
  - *Null hypothesis*

B. Systematic effects plus chance.
  - Often occurs. Interesting because there's at least some systematic factor.
  - *Alternative hypothesis*

C. Systematic effects alone (no chance variation).
  - We're interested in systematic effects, but this almost never happens!

An important part of statistics is determining whether we've got case A or B.

We have a natural tendency to over-estimate the influence of systematic factors

- The lottery is entirely a game of chance (no skill), yet subjects often act as if they have some control over the outcome. (Langer, 1975).
- We tend to feel that a person who is grumpy the first time we meet them is fundamentally a grumpy person. (The "fundamental attribution error," Ross, 1977.)

The purpose of statistics

- As researchers, we need a principled way of analyzing data, to protect us from inventing elaborate explanations for effects in data that could have occurred predominantly due to chance.

Example

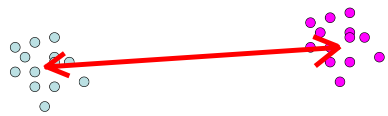
- You have subjects memorize lists of words, and record how many they can remember.
- Does the number they can remember depend upon word length?

	long words	short words
	4	4
	8	5
	9	6
	6	4
	6	5
	9	6
mean	7	5

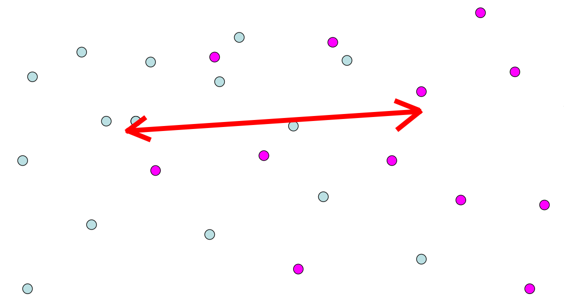
Today we'll test whether the difference in means is "significant," using a "t-test"

- "Significant" = a difference in means this big is unlikely to have occurred by chance
  - Thus there's likely to be a systematic, generalizable effect.
- Let's get some intuitions: what might determine whether or not we think a difference in means is "significant"?

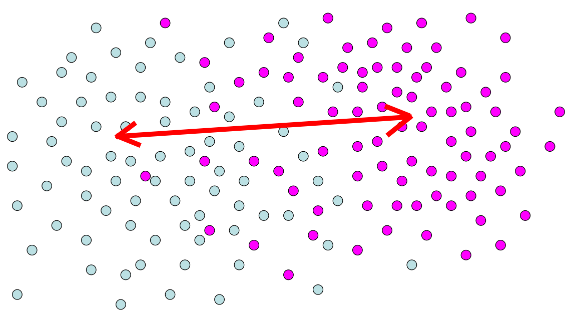
Is the difference in mean of these 2 groups systematic, or just due to chance?



What about this difference in mean?



What about this difference in mean?



Normal distribution of data for two conditions

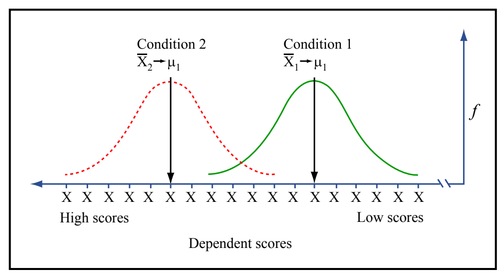


Figure by MIT OpenCourseWare.

Intuitions: Significant difference in means

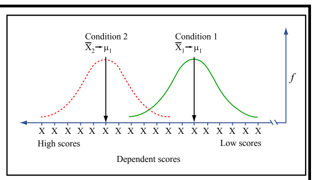


Figure by MIT OpenCourseWare.

- Occurs when the difference in means is large compared to the spread (e.g. variance s^2 or standard deviation s) of the data.
  - $t_{stat} \approx (m_1 - m_2) / s$
- Depends upon the number of samples.
  - With more samples, we're willing to say a difference is significant even if the variance is a bit larger compared to the difference in means.
    - $t_{stat} = (m_1 - m_2)$ / standard error (SE)
    - Standard error = something like s/sqrt(n)
    - Degree of freedom (df) describe the number of samples (n-1)

t-tests

- In general, we'll compute from our data some t_{stat}, of the form:

  $t_{stat} = (m_1 - m_2)/SE$

- t_{stat} is a measure of how reliable of a difference we're seeing between the two conditions.

- If this number is "big enough" we'll say that there is a *significant difference* between the two conditions.

The *t* Test

$$t = \frac{\text{Difference between groups (means)}}{\text{Normal variability within group (or standard error SE)}}$$

- If *t* is large, the difference between groups is **much bigger** than the normal variability within groups.
  - Therefore, two groups are significantly different from each other
- If *t* is small, the difference between groups is **much smaller** than the normal variability within groups.
  - Therefore, two groups are **not** significantly different from each other

t-tests

- Would like to set a threshold, t_{crit}, such that $\mathbf{t_{stat} > t_{crit}}$ means the difference we see between the conditions is unlikely to have occurred by chance (and thus there's likely to be a real systematic difference between the two conditions).

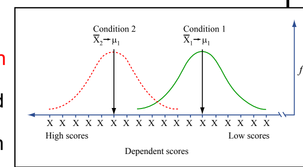


Figure by MIT OpenCourseWare.

- How big is t_{stat} likely to be if there's actually *no difference between the two conditions?*

Read t-crit estimation in a table

In the word experiment

(cf. t-testdemo excel file):

df=5, level of significance is 0.05

Figure removed due to copyright restriction.

OK, so here's the general plan:

- Compute t_{stat} and df from your data (cf. T-TestDemo.xls)
- Decide upon a level of *confidence (significance)*. 99% and 95% are typical. => *significance level,* α = 0.01 or 0.05
- From this, and a t-table, find t_{crit}
- Compare t_{stat} to this threshold.
  - If $|t_{stat}|>|t_{crit}|$, "the difference is significant", there's likely an actual difference between the two conditions.
  - If not, the difference is "not significant."

3 kinds of t-tests

- Case 1: The two samples are *related*, i.e. not independent (e.g. the same subject did the 2 conditions of your experiment)
- Case 2: The samples are independent (e.g. different subjects), and the variances of the populations are *equal*.
- Case 3: The samples are independent, and the variances of the populations are *not equal*.

All tests are of the same form. We just need to know, for each case, how to compute SE (and thus t_{stat}), and what is df.

Case 1: When do you have related or paired samples?

- When you test each subject on both conditions.
  - E.G. You ask 100 subjects two geography questions: one about France, and the other about Great Britain. You then want to compare scores on the France question to scores on the Great Britain question.
  - These two samples (answer, France, & answer, GB) are not independent – someone getting the France question right may be good at geography, and thus more likely to get the GB question right.

Case 1: When do you have related or paired samples?

- When you have "matched samples".
  - E.G. You want to compare weight-loss diets A and B.
  - How well the two diets work may well depend upon factors such as:
    - How overweight is the dieter to begin with?
    - How much exercise do they get per week?
  - Match each participant in group A as nearly as possible to a participant in group B who is similarly overweight, and gets a similar amount of exercise per week.

Excel demo: Related samples t-test

- Let x_i and y_i be a pair in the experimental design
  - The scores of a matched pair of participants, or
  - The scores of a particular participant, on the two conditions of the experiment
- Let $D_i = (x_i – y_i)$
- Compute SE = stdev(D_i)/sqrt(n)
- $t_{stat} = (m_1 – m_2)/SE$,
- df = n-1 = # of pairs - 1

Case 2: Independent samples, equal variances

- Independent samples may occur, for instance, when the subjects in condition A are different from the subjects in condition B (e.g. most drug testing).
- Either the sample variances look very similar, or there are theoretical reasons to believe the variances are roughly the same in the two conditions.

Excel demo
Case 2: Independent samples, equal variances

- $t_{stat} = (m_1 – m_2)/SE$
- $SE = sqrt(s_{pool}^2\ (1/n_1 + 1/n_2))$
- $s_{pool}^2 = [(n_1 – 1)s_1^2 + (n_2 – 1)s_2^2]/(n_1 + n_2 – 2)$
- This is like an average of estimates s_1^2 and s_2^2, weighted by their degrees of freedom, $(n_1 – 1)$ and $(n_2 – 1)$, i.e. essentially by the number of samples used to compute s_1^2 and s_2^2.
- $df = n_1 + n_2 – 2$

Case 3: Independent samples, variances not equal

- The samples variances may be very different, or one may have theoretical reasons to suspect that the variances are not the same in the two conditions.
  - E.G. the response of healthy people to a drug may be more uniform than the response of sick people.
  - E.G. one high school may have students with a bigger range in the education of the students' parents, and one might thus expect a bigger range of test scores.

Excel demo: Case 3: Independent samples, variances not equal

- $t_{stat} = (m_1 – m_2)/SE$
- $SE = sqrt(s_1^2/n_1 + s_2^2/n_2)$
- For equal variances: d.f. = $n_1 + n_2 – 2$
- Unequal variances:

$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

How many subjects per level (condition) should you run?

- How many subjects to use depend on how much variability you expect in your data
- The more subjects you have, the less the means of the data will deviate from their true value
- The usual way of representing this error of measurement is called the **standard error of the mean (s.e.m)**
- Increasing the number of subjects does not decrease the error of measurement in a linear way.
- **Nb participants ? ~ 10 / condition, from 12-20 participants, results should be stable**

Figure removed due to copyright restriction.

How many subjects should you test?

- Doubling the number of subjects (from 10 to 20) reduces the s.e.m by only 30 % (theoretical case)

Figures removed due to copyright restriction.