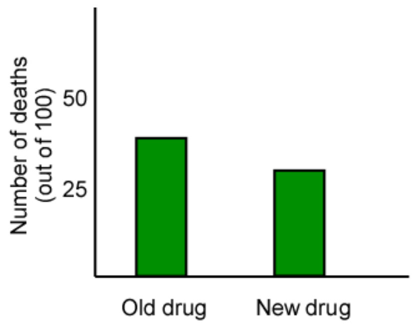
A Very Brief Intro to Statistics: t-tests

Ruth Rosenholtz

Instructor, 9.07, Spring 2006

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

- The data:



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

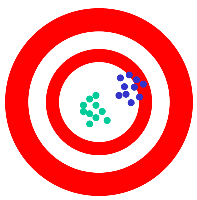
- There's an empirical difference between the old drug and the new drug.
- But is it due to a systematic factor (e.g. the new drug works better) or due to chance?
- If we gave the new drug to 100 more people, would we expect to continue to see improvement over the old drug? Do we expect this effect to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

Chance vs. systematic factors

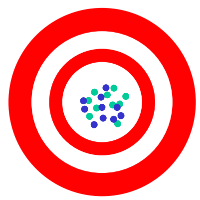
- A *systematic* factor is an influence that contributes a predictable advantage to a subgroup of our observations.
  – E.G. a longevity gain to elderly people who remain active.
  – E.G. a health benefit to people who take a new drug.
- A *chance* factor is an influence that contributes haphazardly (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:



No chance variation

Image removed due to copyright reasons.

On a scale from 1 to 10, rate your experience at MIT so far.
Engineering majors:
7, 7, 7, 7, 7, 7, 7, 7, 7, 7, …
BCS majors:
8, 8, 8, 8, 8, 8, 8, 8, 8, 8, …

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

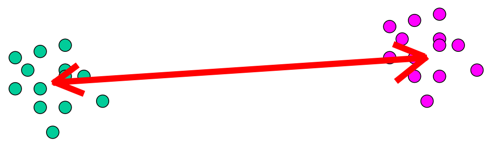
- You have subjects memorize lists of words, and record how many they can remember.
- Does the number they can remember depend upon word length?

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

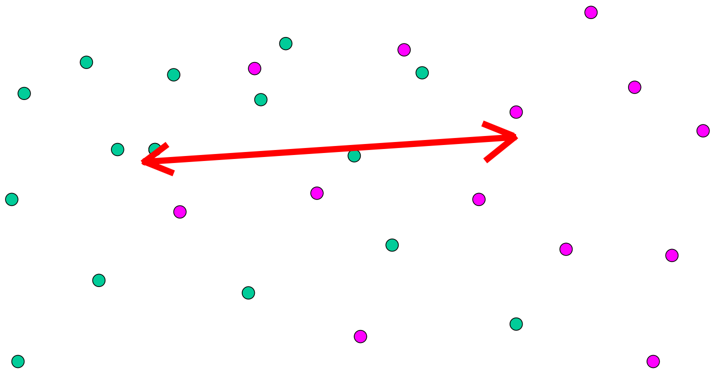
Today we'll test whether the difference in means is "significant," using a "t-test"

- "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"?

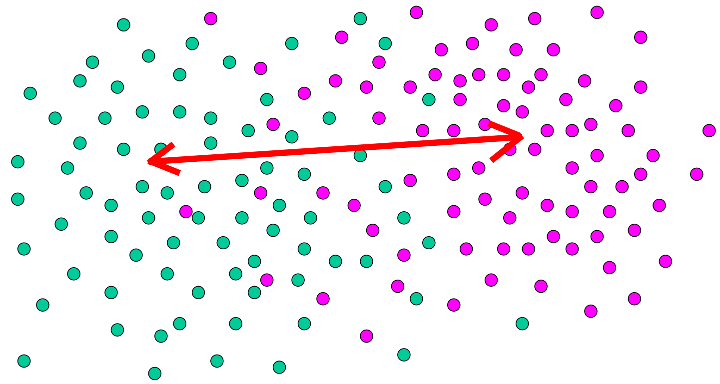
Is the difference in mean of these 2 groups systematic, or just due to chance?



What about this difference in mean?



What about this difference in mean?



Intuitions: Significant difference in means

- 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)

t-tests

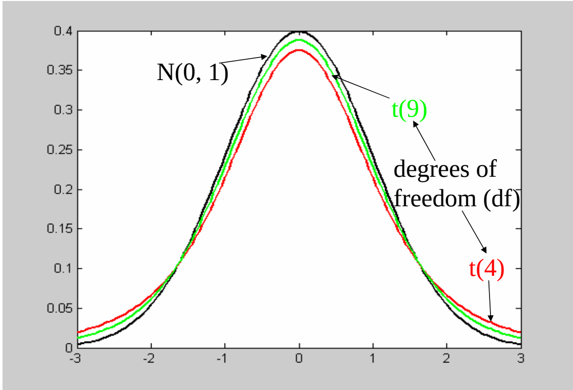
- In general, we'll compute from our data some t_{stat}, of the form:

  $t_{stat} = (m_1 - m_2)/SE$
- 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.
- How do we decide if it is "big enough"?

t-tests

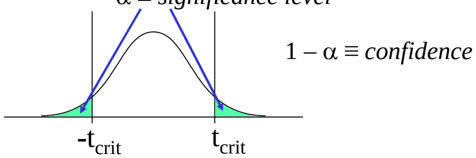
- Would like to set a threshold, t_{crit}, such that $t_{stat}>t_{crit}$ means the difference we see between the conditions is unlikely to have occurred by chance (and thus there's likely to be a real systematic difference between the two conditions).
- Well, how big is t_{stat} likely to be if there's actually *no difference between the two conditions?*
  - (Any difference we see in the data is due to chance.)

Theoretical distribution of $t_{stat} = (m_1 – m_2)/SE$, assuming the two conditions differ only by chance



Confidence and t_{crit}

- You could see in your data a big value for t_{stat}, even if there's no real difference between the conditions. But it's unlikely.
- How unlikely? $P(|t_{stat}|>|t_{crit}|) = \alpha$.



OK, so here's the general plan:

- Compute t_{stat} and df from your data (exact details to follow)
- Decide upon a level of *confidence*. 99% and 95% are typical. => *significance level*, $\alpha = 0.01$ 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.
- Case 2: The samples are independent, 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 have a "repeated measures" experimental design, i.e. 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.

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 (repeated measures)
- 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

Excel demo

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.

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.
- $df = n_1 + n_2 – 2$

Excel demo

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.

Case 3: Independent samples, variances not equal

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

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

Excel demo

Summary of two-sample tests for a significant difference in mean

When to do this test	Standard error	Degrees of freedom
Small sample, $\sigma_1^2=\sigma_2^2$	$\sqrt{s_{pool}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	$n_1 + n_2 - 2$
Small sample, $\sigma_1^2\neq\sigma_2^2$	$\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$
Related samples	$s_D/\sqrt{n}$	$n - 1$