

Two-sample hypothesis testing, II

9.07

3/16/2004

Small sample tests for the difference between two independent means

- For two-sample tests of the difference in mean, things get a little confusing, here, because there are several cases.
- Case 1: The sample size is small, and the standard deviations of the populations are *equal*.
- Case 2: The sample size is small, and the standard deviations of the populations are *not equal*.

Inhomogeneity of variance

- Last time, we talked about Case 1, which assumed that the variances for sample 1 and sample 2 were equal.
- Sometimes, either theoretically, or from the data, it may be clear that this is not a good assumption.
- Note: the equal-variance t-test is actually pretty robust to reasonable differences in the variances, if the sample sizes, n_1 and n_2 are (nearly) equal.
 - *When in doubt about the variances of your two samples, use samples of (nearly) the same size.*

Case 2: Variances not equal

- Sometimes, however, it either isn't possible to have an equal number in each sample, or the variances are very different.
- In which case, we move on to Case 2, the t-test for difference in means when the variances are not equal.

Case 2: Variances not equal

- Basically, one can deal with unequal variances by making a correction in the value for degrees of freedom.
- Equal variances: d.f. = $n_1 + n_2 - 2$
- Unequal variances:

$$\text{d.f.} = \frac{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)^2}{\frac{(\sigma_1^2 / n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2 / n_2)^2}{n_2 - 1}}$$

Note on corrected degrees of freedom

- There are several equations out there for correcting the number of degrees of freedom.
- This equation is a bit on the conservative side – it will lead to an overestimate of the p-value.
- An even easier conservative correction:
$$\text{d.f.} = \min(n_1 - 1, n_2 - 1)$$
- You will NOT be required to memorize any of these equations for an exam.
- Use the one on the previous slide for your homework.

Case 2: Variances not equal

- Once we make this correction, we proceed as with a usual t-test, using the equation for SE from last time, for unequal variances.
- $SE(\text{difference in means}) = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
- $t_{\text{obt}} = (\text{observed} - \text{expected})/SE$
- Compare with t_{crit} from a t-table, for d.f. degrees of freedom, from the previous slide.

Example

- The math test scores of 16 students from one high school showed a mean of 107, with a standard deviation of 10.
- 11 students from another high school had a mean score of 98, and a standard deviation of 15.
- Is there a significant difference between the scores for the two groups, at the $\alpha=0.05$ level?

Set up the null and alternative hypotheses, and find t_{crit}

- $H_0: \mu_1 - \mu_2 = 0$
- $H_a: \mu_1 - \mu_2 \neq 0$
- $s_1^2 = 10^2; s_2^2 = 15^2; n_1 = 16; n_2 = 11$

$$\text{d.f.} = \frac{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)^2}{\frac{(\sigma_1^2 / n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2 / n_2)^2}{n_2 - 1}}$$

- $\text{d.f.} = (100/16 + 225/11)^2 /$
 $[(100/16)^2/15 + (225/11)^2/10]$
 ≈ 16

Determining t_{crit}

- d.f. = 16, $\alpha=0.05$
- Two-tailed test ($H_a: \mu_1 - \mu_2 \neq 0$)
- $t_{\text{crit}} = 2.12$

Compute t_{obt} , and compare to t_{crit}

- $m_1 = 107$; $m_2 = 98$, $s_1 = 10$; $s_2 = 15$; $n_1 = 16$;
 $n_2 = 11$
- $t_{\text{obt}} = (\text{observed} - \text{expected})/\text{SE}$
- Observed difference in means = $107 - 98 = 9$
- $\text{SE} = \text{sqrt}(\sigma_1^2/n_1 + \sigma_2^2/n_2)$
 $= \text{sqrt}(10^2/16 + 15^2/11) = 5.17$
- $t_{\text{obt}} = 9/5.17 = 1.74$
- $t_{\text{obt}} < t_{\text{crit}}$, so the difference between the two schools is not significant.

Two-sample hypothesis testing for a difference in mean

- So, we've covered the following cases:
 - Large sample size, independent samples
 - Small sample size, independent samples, equal variance
 - Small sample size, independent samples, unequal variance
- There's essentially one more case to go: related samples.

But first, consider the following example

- An owner of a large taxi fleet wants to compare the gas mileage with gas A and gas B.
- She randomly picks 100 of her cabs, and randomly divides them into two groups, A and B, each with 50 cabs.
- Group A gets gas A, group B gets gas B

Gas mileage example

- After a day of driving, she gets the following results:

	mean mpg	std. dev.
Group A	25	5.00
Group B	26	4.00

- $m_1 - m_2 = -1$. Is this significant? The samples are large enough that we can use a z-test.

Gas mileage example

- $H_0: \mu_1 - \mu_2 = 0; H_a: \mu_1 - \mu_2 \neq 0$
- $z_{\text{obt}} = -1/\text{SE}$
- $\text{SE} = \text{sqrt}(s_1^2/n_1 + s_2^2/n_2)$
 $= \text{sqrt}(25/50 + 16/50) = 0.905$
- $z_{\text{obt}} = -1.1 \rightarrow p=0.27$
- No significant difference.

What could the fleet owner have done better?

- Though gas B seems to be slightly better than gas A, there was no way this difference was going to be significant, because the variance in each sample was too high.
- I.E. gas mileages varied widely from one cab to the next. Why?
 - Cars vary greatly in their gas mileage, and drivers vary greatly in how they drive.

What could the fleet owner have done better?

- This kind of variability (variability in cars and drivers) is basically irrelevant to what the fleet owner wants to find out. Is there some way she could do the experiment so that she can essentially separate out this irrelevant variability from the variability she's interested in (due to the change in gasoline)?
- Yes: use a matched-sample or repeated-measures experimental design.

Matched samples: Weight-loss example

- Suppose you want to compare weight-loss diet A with diet 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?
- You would like to make sure that the subjects in group A (trying diet A) are approximately the same according to these factors as the subjects in group B.

Matched-samples experimental design

- One way to try, as much as possible, to match relevant characteristics of the people in group A to the characteristics of the people in group B:
 - 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.
- This is called a “matched samples” design.
- If group A does better than group B, this cannot be due to differences in amt. overweight, or amt. of exercise, because these are the same in the two groups.

- What you're essentially trying to do, here, is to remove a source of variability in your data, i.e. from some participants being more likely to respond to a diet than others.
- Lowering the variability will lower the standard error, and make the test more powerful.

Matched-samples design

- Another example: You want to compare a standard tennis racket to a new tennis racket. In particular, does one of them lead to more good serves?
- Test racket A with a number of tennis players in group A, racket B with group B.
- Each member of group A is matched with a member of group B, according to their serving ability.
- Any difference in serves between group A and group B cannot be due to difference in ability, because the same abilities are present in both samples.

Matched-samples design

- Study effects of economic well-being on marital happiness in men, vs. on marital happiness in women.
- A natural comparison is to compare the happiness of each husband in the study with that of his wife.
- Another common “natural” comparison is to look at twins.

Repeated-measures experimental design

- Match each participant with him/her/itself.
- Each participant is tested under all conditions.
- Test each person with tennis racket A and tennis racket B. Do they serve better with A than B?
- Test each car with gasoline A and gasoline B – does one of them lead to better mileage?
- Does each student study better while listening to classical music, or while listening to rock?

Related samples hypothesis testing

- In related samples designs, the matched samples, or repeated measures, mean that there is a *dependence* within the resulting pairs.
- Recall:
 - A, B independent \Leftrightarrow the occurrence of A has no effect on the probability of occurrence of B.

Related samples hypothesis testing

- Dependence of pairs in related samples designs:
 - If subject A_i does well on diet A, it's more likely subject B_i does well on diet B, because they are matched for factors that influence dieting success.
 - If player A_i does well with racket A, this is likely in part due to player A_i having good tennis skill. Since player B_i is matched for skill with A_i , it's likely that if A_i is doing well, then B_i is also doing well.

Related samples hypothesis testing

- Dependence of pairs in related samples designs:
 - If a husband reports high marital happiness, this increases the chances that we'll find his wife reports high marital happiness.
 - If a subject does well on condition A of a reaction time experiment, this increase the chances that he will do well on condition B of the experiment, since his performance is likely due in part to the fact that he's good at reaction time experiments.

Why does dependence of the two samples mean we have to do a different test?

- For independent samples,

$$\text{var}(m_1 - m_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

- For dependent samples,

$$\text{var}(m_1 - m_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2 - 2 \text{cov}(m_1, m_2)$$

Why does dependence of the two samples mean we have to do a different test?

- “COV”?
 - Covariance(x, y) = $E(xy) - E(x)E(y)$
 - If x & y are independent, = 0.
 - We’ll talk more about this later, when we discuss regression and correlation.
- For a typical repeated-measures or matched-samples experiment, $cov(m_1, m_2) > 0$.
 - If a tennis player does well at serving with racket A, they will tend to also do well with racket B.
 - Negative covariance would happen if doing well with racket A tended to go with doing poorly with racket B.

Why does dependence of the two samples mean we have to do a different test?

- So, if $\text{cov}(m_1, m_2) > 0$,
$$\text{var}(m_1 - m_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2 - 2 \text{cov}(m_1, m_2)$$
$$< \sigma_1^2/n_1 + \sigma_2^2/n_2$$
- This is just the reduction in variance you were aiming for, in using a matched-samples or repeated-measures design.

Why does dependence of the two samples mean we have to do a different test?

- If $\text{cov}(m_1, m_2) > 0$,
$$\text{var}(m_1 - m_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2 - 2 \text{cov}(m_1, m_2)$$
$$< \sigma_1^2/n_1 + \sigma_2^2/n_2$$
- The standard t-test will tend to overestimate the variance, and thus the SE, if the samples are dependent.
- The standard t-test will be too conservative, in this situation, i.e. it will overestimate p.

So, how do we do a related samples t-test?

- Let x_i be the i th observation in sample 1.
Let y_i be the i th observation in sample 2.
- x_i and y_i are 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)
- Create a new random variable, D , where
$$D_i = (x_i - y_i)$$
- Do a standard *one-sample* z- or t-test on this new random variable.

Back to the taxi cab example

- Instead of having two different groups of cars try the two types of gas, a better design is to have the same cars (and drivers) try the two kinds of gasoline on two different days.
 - Typically, randomize which cars try gas A on day 1, and which try it on day 2, in case order matters.
- The fleet owner does this experiment with only 10 cabs (before it was 100!), and gets the following results:

Results of experiment

Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
...
Mean	25.20	25.80	-0.60
Std. dev.	4.27	4.10	0.61

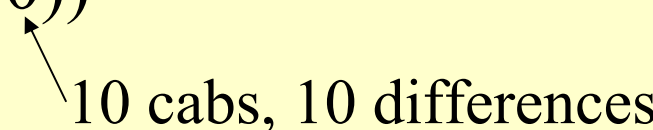
Means and std. dev's of gas A and B about the same -- they have the same source of variability as before.

Results of experiment

Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
...
Mean	25.20	25.80	-0.60
Std. dev.	4.27	4.10	0.61

But the std. dev. of the difference is very small.
Comparing the two gasolines within a single car
eliminates variability between taxis.

t-test for related samples

- $H_0: \mu_D = 0; H_a: \mu_D \neq 0$
- $t_{\text{obt}} = (m_D - 0)/SE(D)$
 $= -0.60/(0.61/\text{sqrt}(10))$
 $= -3.11$

- To look up p in a t-table, use d.f.=N-1=9
- $\rightarrow p = 0.0125$
Gas B gives significantly better mileage than gas A.

Note

- Because we switched to a repeated measures design, and thus reduced irrelevant variability in the data, we were able to find a significant difference with a sample size of only 10.
- Previously, we had found no significant difference with 50 cars trying each gasoline.

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

When to do this test	Standard error	Degrees of freedom
Large sample size	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	not applicable
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, $\sigma_1^2 \neq \sigma_2^2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$\frac{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)^2}{\frac{(\sigma_1^2 / n_1)^2}{n_1 - 1} + \frac{(\sigma_2^2 / n_2)^2}{n_2 - 1}}$
Related samples	s_D / \sqrt{n}	$n - 1$

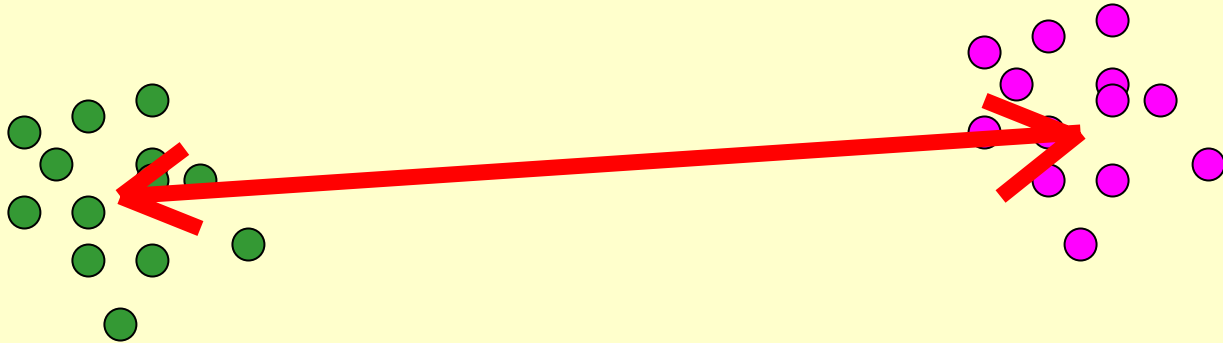
Confidence intervals

- Of course, in all of these cases you could also look at confidence intervals
- $(\mu_1 - \mu_2)$ falls within $(m_1 - m_2) \pm t_{\text{crit}} \cdot \text{SE}$, with a confidence corresponding to t_{crit} .

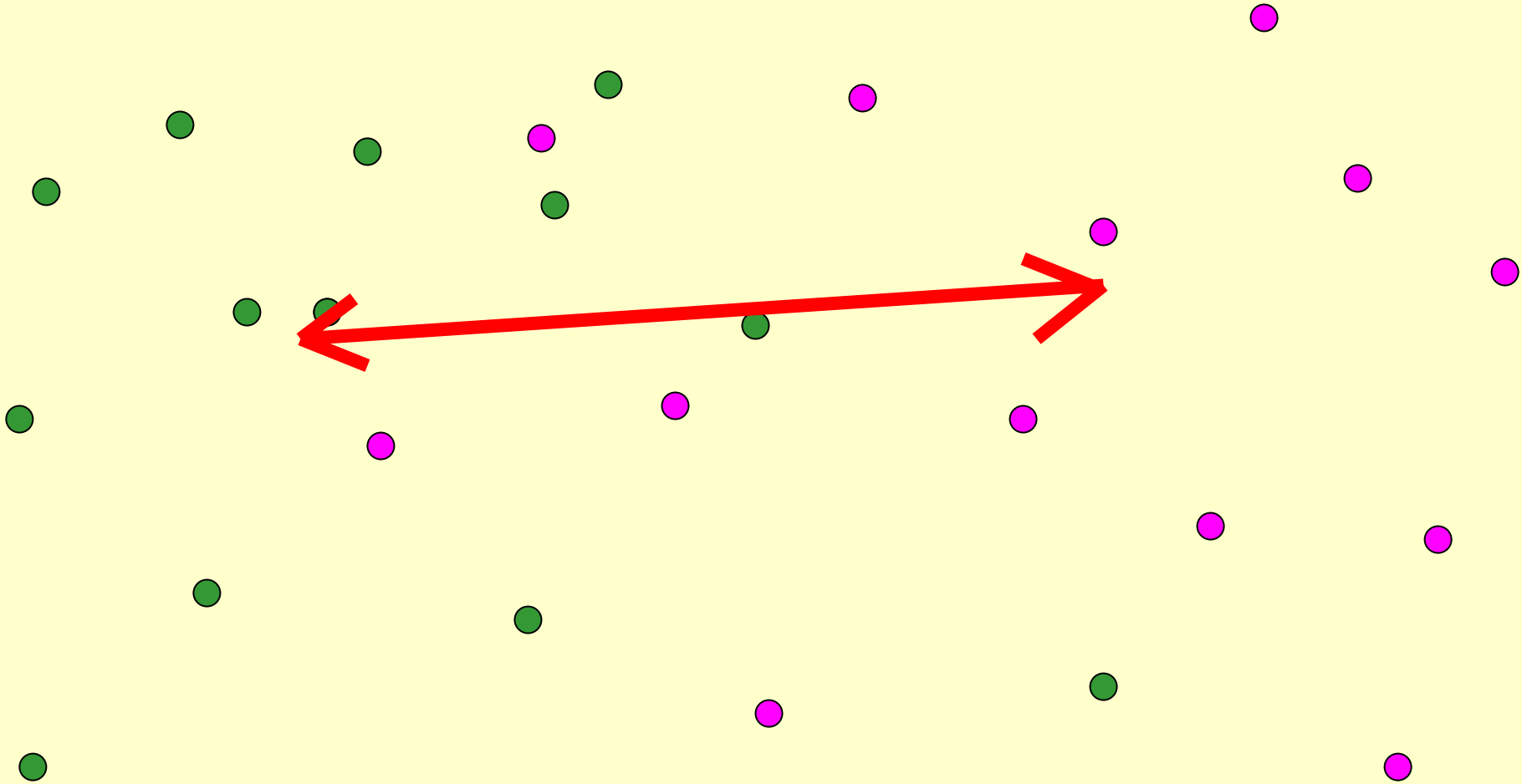
Taking a step back

- What happens to the significance of a t-test, if the difference between the two group means gets larger? (What happens to p ?)
- What happens to the significance of a t-test, if the standard deviation of the groups gets larger?
- *This is inherent to the problem of distinguishing whether there is a systematic effect or if the results are just due to chance. It is not just due to our choosing a t-test.*

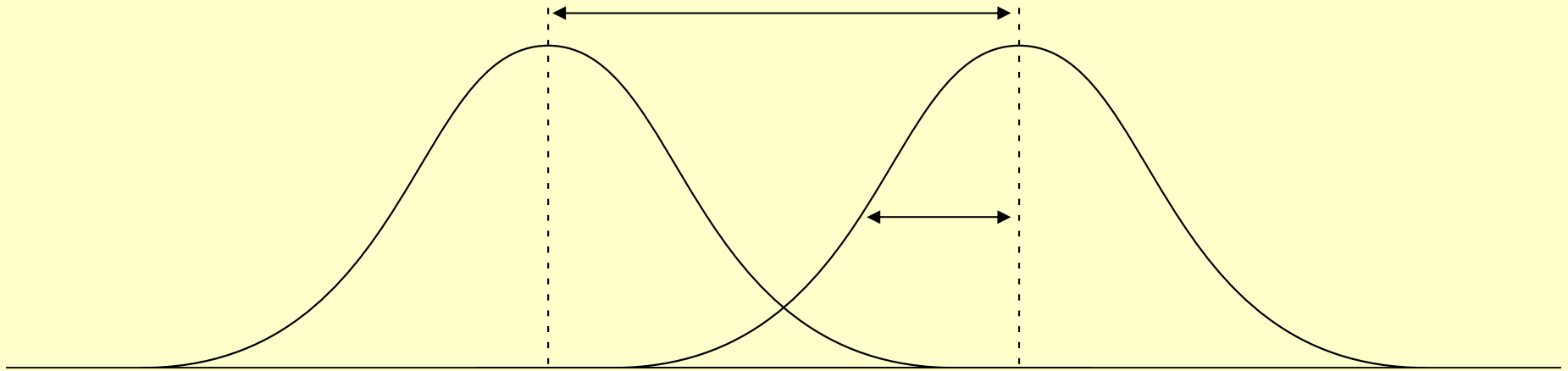
Is the difference in mean of these 2 groups systematic, or just due to chance?



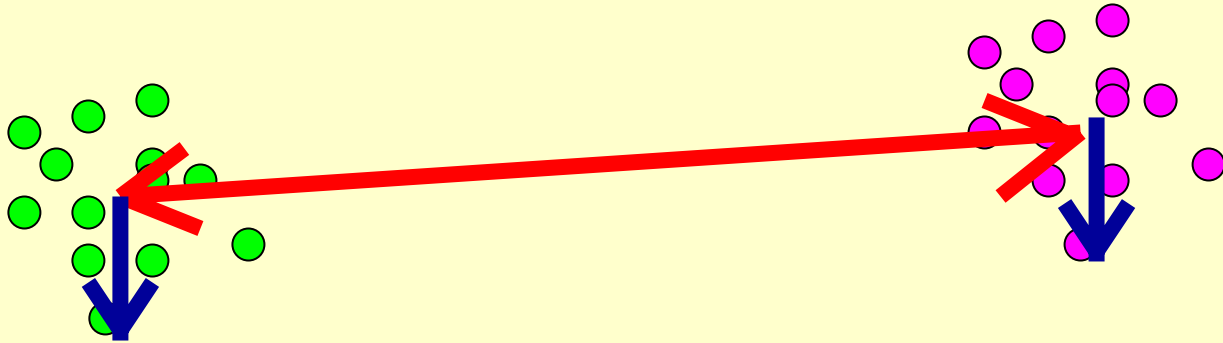
What about this difference in mean?



- A difference in the means of two groups will be significant if it is sufficiently large, compared to the variability within each group.

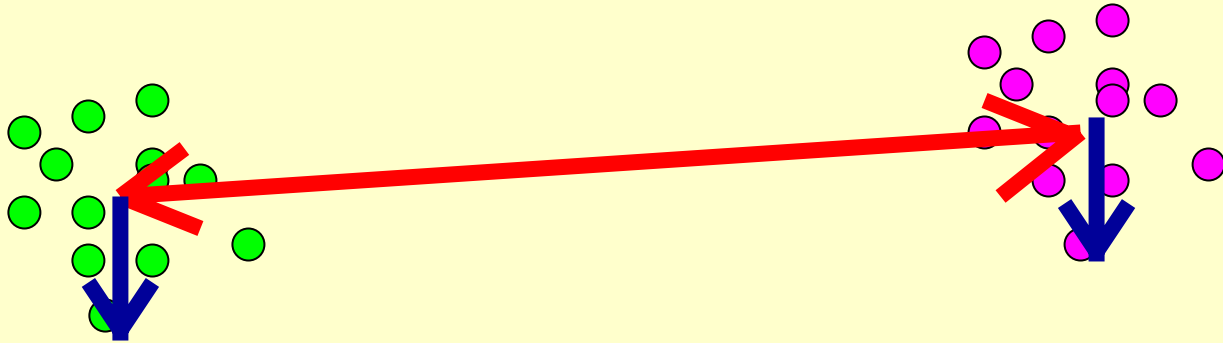


Put another way:



The essential idea of statistics is to compare some measure of the **variability between means of groups** with some measure of the **variability within those groups**.

Yet another way:



Essentially, we are comparing the *signal* (the difference between groups that we are interested in) to the *noise* (the irrelevant variation within each group).

Variability and effect size

- If the variability between groups (the “signal”) is large when compared to the variability within groups (the “noise”) then the *size of the effect* is large.
- A measure of effect size is useful both in:
 - Helping to understand the *importance* of the result
 - Meta-analysis – comparing results across studies

Measures of effect size

- There are a number of possible measures of effect size that you can use.
- Can p (from your statistical test) be a measure of effect size?
 - No, it's not a good measure, since p depends upon the sample size.
 - If we want to compare across experiments, we don't want experiments with larger N to appear to have larger effects. This doesn't make any sense.

Measures of effect size

- Difference in mean for the two groups, $m_1 - m_2$
 - Independent of sample size
 - Nice that it's in the same units as the variable you're measuring (the *dependent variable*, e.g. RT, percent correct, gas mileage)
 - However, could be large and yet we wouldn't reject the null hypothesis
 - We may not know a big effect when we see one
 - Is a difference in RT of 100 ms large? 50 ms?

Measures of effect size

- Standardized effect size

$d = (m_1 - m_2)/\sigma$, where σ^2 is a measure of variance within each of the groups, e.g. s_i^2 or s_{pool}^2

- Cohen's rule of thumb for this measure of effect size:

d=0.20 low

d=0.50 medium

d=0.80 high

Measures of effect size

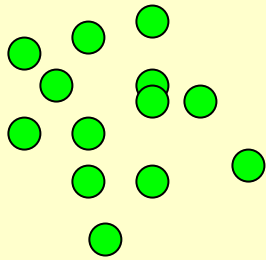
- What proportion of the total variance in the data is accounted for by the effect?
 - This concept will come up again when we talk about correlation analysis and ANOVA, so it's worth discussing in more detail.

Recall dependent vs. independent variables

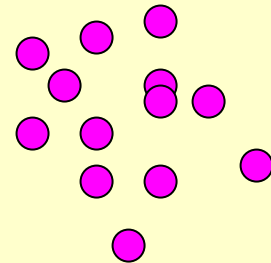
- An experimenter tests two groups under two conditions, looking for evidence of a statistical relation.
 - E.G. test group A on tennis racket A, group B on tennis racket B. Does the racket affect proportion of good serves?
- The manipulation under control of the experimenter (racket type) is the *independent variable*. “**X**”
- The resulting performance, not under the experimenter’s control, is the *dependent variable* (proportion of good serves). “**Y**”

Predicting the value of Y

- If X has a significant effect, then the situation might look like this:



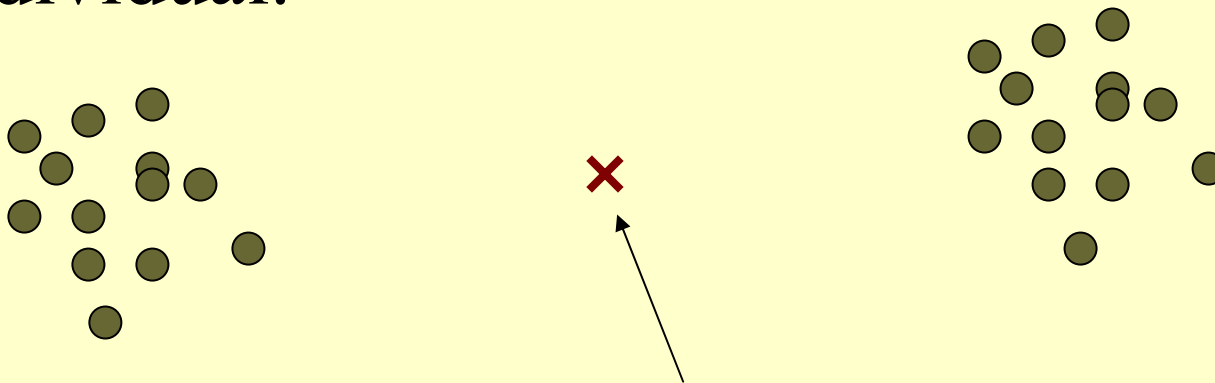
$X = A$



$X = B$

Predicting the value of Y

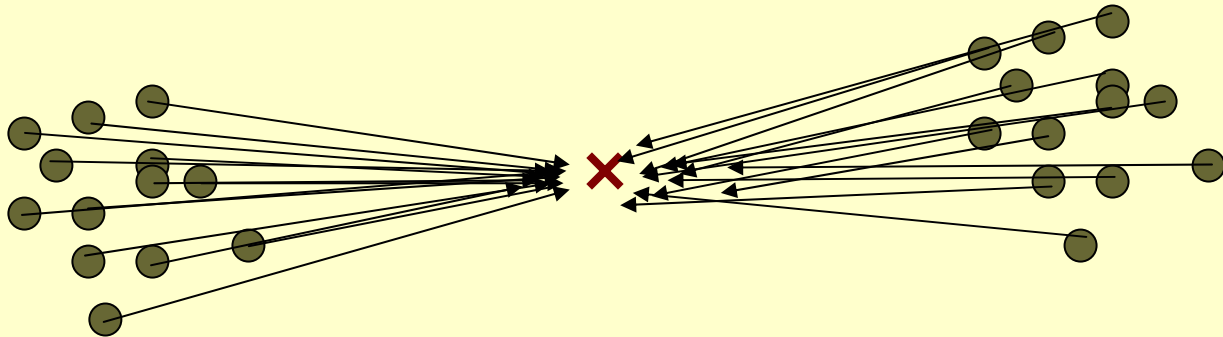
- Suppose I pick a random individual from one of these two groups (you don't know which), and ask you to estimate Y for that individual.



It would be hard to guess! The best you could probably hope for is to guess the mean of all the Y values (at least your error would be 0 on average)

How far off would your guess be?

- The variance about the mean Y score, σ_Y^2 , gives a measure of your uncertainty about the Y scores.

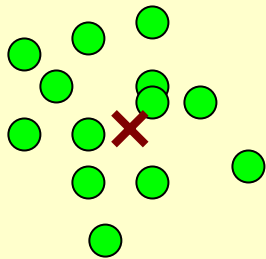


Predicting the value of Y when you know X

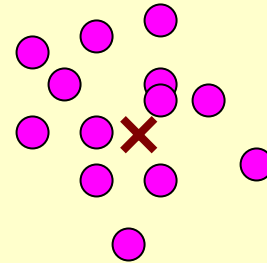
- Now suppose that I told you the value of X (which racket a player used), and again asked you to predict Y (the proportion of good serves).
- This would be somewhat easier.

Predicting the value of Y when you know X

- Your best guess is again a mean, but this time it's the mean Y for the given value of X .



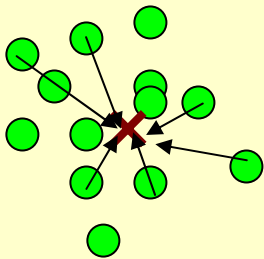
$X = A$



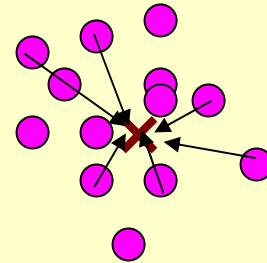
$X = B$

How far off would your guess be, now?

- The variance about the mean score *for that value of X* , $\sigma_{Y|X}^2$ gives a measure of your uncertainty.



$X = A$



$X = B$

The strength of the relationship between X and Y

- is reflected in the extent to which knowing X reduces your uncertainty about Y.
- Reduction in uncertainty = $\sigma_Y^2 - \sigma_{Y|X}^2$
- Relative reduction in uncertainty:
$$\omega^2 = (\sigma_Y^2 - \sigma_{Y|X}^2) / \sigma_Y^2$$
- This is the *proportion of variance in Y accounted for by X*.

$$\begin{aligned} & (\text{total variation} - \text{variation left over}) / (\text{total variation}) \\ & = (\text{variation accounted for}) / (\text{total variation}) \end{aligned}$$

Back to effect size

- This proportion of the variance accounted for by the effect of the independent variable can also serve as a measure of effect size.
- Version that's easy to calculate once you've done your statistical test (see extra reading on the MIT Server):

$$r_{pb}^2 = \frac{t_{obt}^2}{t_{obt}^2 + d.f.}$$

t_{obt} value from t-test

degrees of freedom from t-test

Summary of one- and two-sample hypothesis testing for a difference in mean

- We've talked about z- and t-tests, for both independent and related samples
- We talked about power of a test
 - And how to find it, at least for a z-test. It's more complicated for other kinds of tests.
- We talked about estimating the sample size(s) you need to show that an effect of a given size is significant.
 - One can do a similar thing with finding the sample size such that your test has a certain power.
- We talked about measures of effect size

Steps for hypothesis testing (in general), revisited

Steps from before:

1. Formulate hypotheses H_0 and H_a
2. Select desired significance level, α
3. Collect and summarize the data, using the test statistic that will assess the evidence against the null hypothesis
4. If the null hypothesis were true, what is the probability of observing a test statistic at least as extreme as the one we observed? (Get the p-value)
5. Check the significance by comparing p with α .
6. Report the results

Steps for hypothesis testing: the long version

Steps:

1. Formulate hypotheses H_0 and H_a
2. Select desired significance level, α
3. Specify the effect size of interest (“an effect of 1% would be important”, or “large”)
4. Specify the desired level of power for an effect of this size
5. Determine the proper size of the sample(s)
6. Collect and summarize the data, using the test statistic that will assess the evidence against the null hypothesis

Steps for hypothesis testing: the long version

7. If the null hypothesis were true, what is the probability of observing a test statistic at least as extreme as the one we observed? (Get the p-value)
8. Check the significance by comparing p with α .
9. Report the results
 - A. Means, std. deviations, type of test, degrees of freedom, value of test statistic, p-value
 - B. Post-hoc measure of the size of effect
 - C. Possibly post-hoc measure of the power of the test