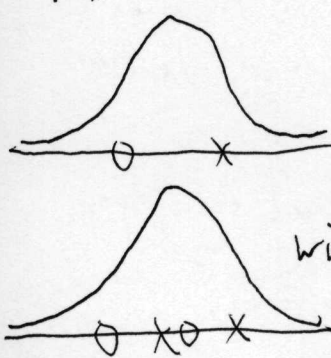


Hypothesis Testing and Statistical Significance

Wk 11

Computer scientists sometimes structure their experiments like other scientists: they have a hypothesis, they run an experiment to test it, and the results are statistically significant. A human-computer interaction researcher may want to argue that people are engaged longer when they use one kind of interface over another. An AI researcher may want to argue that a system performs well on random samples of data compared to a baseline algorithm. While some CS researchers may work in fields where it's acceptable to compare two numbers without any statistical reasoning, other communities might look at such results and say, sure, but is this just a fluke, or is it statistically significant?

Suppose that random samples are being drawn from the same normal distribution. If we just draw two, it's clear that one value will be bigger than the other almost all the time. If we draw two groups, it'll still be the case that the means will almost certainly be slightly different. Are the populations of X's and O's really different here?



No, any differences we see are pure chance — we could have had them switch which was bigger the next time out.

This is an example of a null hypothesis, a hypothesis that is sort of boring and the opposite of what we want to prove. If we want to prove the X's and O's have different means, we need to consider the null hypothesis that they don't.

Statistical significance is achieved by showing the null hypothesis is very unlikely, or rather, by showing that the results are very unlikely given the null hypothesis. We compute the likelihood that our null hypothesis would have generated the data, and if it's

Very low, we conclude our hypothesis is correct instead. Wkll 57

The likelihood of the data under the null hypothesis is sometimes called the "p-value," and the threshold for declaring significance, the α value. 5%, or $p < 0.05$ is a commonly accepted threshold; if the chance of seeing this data or more extreme data is less than 5% under the null hypothesis, we reject the null hypothesis and accept the alternative. Statistical significance is achieved.

So, for example, suppose we're trying to show group X ~~is~~ is taller than group Y on average. Our null hypothesis is ~~that there is no difference between the groups~~ that there is no difference between the groups. (Not that Y is ~~is~~ taller, which itself would be a claim in need of statistical proof.) We run our experiment and calculate that there is only a 3% chance of seeing such a big difference in heights if they were drawn from the same distribution. We conclude that they really are different, and if X is taller on average, we declare the result statistically significant.

Now, there are many misconceptions around this idea of statistical significance. Many researchers like to report better p-value thresholds if they happen to have really low probabilities of the null hypothesis - for example, $p < 0.001$. But this shouldn't be confused with effect size, even though they might call this "highly significant." It means that they're very confident the null hypothesis isn't true. But the effect could still be small - they may have achieved confidence with lots of samples instead of a big effect.

Another potential misreading is that if $p > 5\%$ the null hypothesis is proven. But, this could just mean you didn't run enough trials or get enough samples. The null hypothesis isn't proven - it's just not disproven.

Another thing that might strike you as odd is that the "significance" is hardly about the discovered effect at all - it's just showing that there is some effect. Researchers will often spend a big part of their conclusions explaining why they think something happened. The statistics just confirm that it happened - the explanation has to be analyzed on its own merits.

Lastly, note that if your null hypothesis is badly constructed, then rejecting it doesn't necessarily prove what you want. It's pretty common, for example, for researchers to run experiments that use 5-point scales. If the wrong statistical assumptions are made for the null hypothesis, like the mean being normally distributed despite the hard cutoffs for possible values, this could make the data unlikely under the null hypothesis in a way that has nothing to do with the hypothesis being tested.

Type I and II Errors

This is just a little terminology: a Type I error means rejecting the null hypothesis when you shouldn't, and a Type II error means failing to reject a null hypothesis when you should. α , the significance threshold, is setting a bound on Type I error probability; if $\alpha = 0.05$, the chance of a Type I error is < 0.05 . The chance of a Type II error is sometimes denoted with β , and the "power" of a statistical test is $1 - \beta$. (The main factors affecting power are sample size, effect size, and the variance.)

One - and Two - Tailed Tests

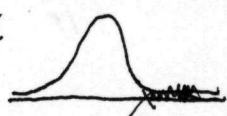
When doing hypothesis testing, we want to know the chance that the null hypothesis would produce a result at least as extreme as the one we got. But, there are two possible interpretations of this.

Interpretation 1: The absolute value of the difference from the mean is at least as extreme as this value. ("Two-tailed")

Lwk 11 58



Interpretation 2: The difference is at least as extreme in this direction as the measured point. ("One-tailed")



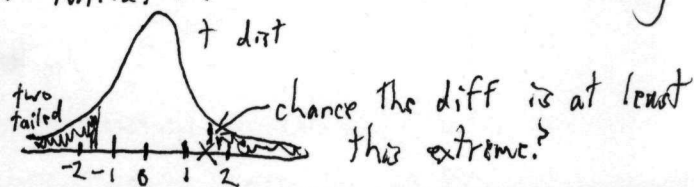
There's an interesting effect of this choice, which is that the two-tailed interpretation makes the data "twice as expected" under the null hypothesis. This accordingly makes the null hypothesis more difficult to disprove, which is rarely what a researcher wants. One could make the argument for some experiments that we never would have entertained a hypothesis going the other direction, and therefore "at least as extreme" should only refer to the direction of your actual hypothesis.

But, in practice, this is usually sketchy, since most scientists would happily reverse their hypothesis and publish a different one if the data supported it. Use two-tailed statistical tests, which don't assume the direction of effect as a foregone conclusion - or else be prepared to encounter skepticism.

T-tests for differences in populations

A very common use of hypothesis testing is to run an experimental group versus a control group, measure both, and determine whether one measurement is significantly larger or smaller than the other. The null hypothesis is "no difference."

The null hypothesis plays out on a t -distribution, since we don't inherently know σ^2 . Generally we also won't make any assumptions about the direction of effect, so it's two-tailed. And what we're measuring is a difference between means, scaled to be in terms of the standard error.

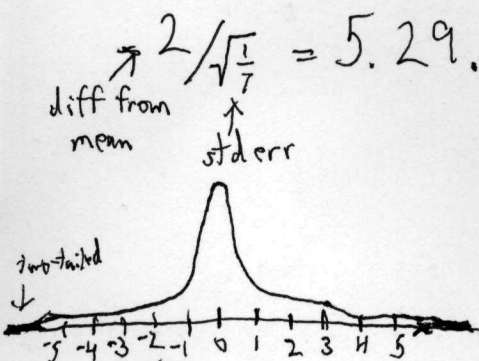


The difference in means $\bar{X}_1 - \bar{X}_2$ we can calculate straightforwardly, finding the mean of each group. The ~~standard error~~ for this difference is $\sqrt{\frac{\sigma^2}{N} + \frac{\sigma^2}{N}} = \sqrt{\frac{2\sigma^2}{N}}$, assuming the two experimental groups have the same N .

We can calculate the "mean squared error" our estimate of σ^2 , as the average of our two unbiased estimates $\frac{\sum (X - \bar{X})^2}{n-1}$. Plugging all that in gives us our standard error; then we can use a t distribution with $2(N-1)$ degrees of freedom to find the likelihood of a difference at least as extreme.

For example, when we talked about confidence intervals, we had an example of 21 people in each group, one with mean 10, the other with mean 12 and MSE 1.5. We calculated the standard error to be $\sqrt{\frac{3}{21}} = \sqrt{\frac{1}{7}}$ and the range of values that were 95% likely to be $2 \pm \sqrt{\frac{1}{7}} \cdot 2.021 = [1.24, 2.764]$.

That's almost what we need to do for the significance test, only we're comparing our standardized value - our " t " value - to a distribution centered around 0. Our value, 2, has a t -value of $2 / \sqrt{\frac{1}{7}} = 5.29$. If we plug this value into a function or look it up in a table of t -values with 40 degrees of freedom, we find that it's well beyond the value of 3.551 we'd need for even $p < 0.001$. So we can reject the null hypothesis and report a nice p -value.



In short, the steps for testing the hypothesis that two means are different:

- 1) Compute the means and unbiased variances for the two groups.
- 2) Compute the overall ~~standard error~~ of the difference of means as $\sqrt{\frac{2\sigma^2}{N}}$ where σ^2 is estimated with the average of the unbiased variance estimates.
- 3) Compute the ~~standard error~~ ^{t -value from} the difference of the means: $\frac{\text{diff}}{\text{standard error}}$
- 4) Using $2(N-1)$ as the degrees of freedom, look up in a t -table or function what t -value is necessary for two-tailed significance ($p < 0.05$). If the value is much larger, consider reporting a better threshold (like $p < 0.0001$).

Extra t-test problems.

Group 1: $\mu_1 = 20$ $\sigma_{unbiased}^2 = 4$

Group 2: $\mu_1 = 21$

56 supp

$\sigma_{unbiased}^2 = 2$

$N = 10$ for both

Average $\sigma_{unbiased}^2 = 3$ $Stderr = \sqrt{\frac{2.3}{10}}$

Diff = 1 $t\text{-value} \approx \frac{1}{\sqrt{\frac{6}{10}}} = 1.29$

Crit value
 $df = 2(9) = 18: 2.10$

Results not significant yet.

95% confidence interval: $1 \pm 2.10 \cdot \sqrt{\frac{6}{10}} = \boxed{1 \pm 7.75}$

As above, but increase sample size to 40 both σ^2 to 0.5

$Stderr = \sqrt{\frac{2 \cdot 0.5}{40}} = \sqrt{\frac{1}{40}}$

$t\text{-value} = \frac{1}{\sqrt{\frac{1}{40}}} = 6.32$

Now we're definitely significant.

Confidence interval $1 \pm \sqrt{\frac{1}{40}} \cdot 2.10 = 1 \pm 0.332$

-
- Why? Confidence intervals & hypothesis testing ($p < 0.05$)
 - Finish confidence intervals for 1 value, conf for 2 values (56)
 - Meaning of p-value, null hypothesis
 - T-test: 59

T-tests for difference from a ~~value~~ value

wk 11 59

It's more common to compare populations, but conceivably you could want to decide whether a value is different from a target value, like 5. This works exactly the same, except one of the numbers in the "difference" has no variance and no additional degrees of freedom. Thus the standard error used is just $\frac{\sigma}{\sqrt{N}}$ instead of

$$\sqrt{\frac{2\sigma^2}{N}} = \sqrt{\frac{\sigma^2}{N} + \frac{\sigma^2}{N}}$$

(Again, N is the number of individuals in one group.)

Chi-square tests

T-tests are good for finding significant effects for continuous values; finding that one mean is bigger than another, for example. Chi-square tests are good for finding significance in boolean or discrete values. For a Boolean example, we might wonder whether Mac users are more likely than Windows users to install our product. Each sample is just an answer to two Boolean questions: Mac? And, Software Installed? Our datapoints aren't continuous, and treating them as 0 or 1 won't result in a normal distribution, so we need a different approach to decide whether a claim like "Mac users install our software less" has statistical significance.