Lecture 6: Hypothesis Testing

Thomas Chadefaux

PO7001: Quantitative Methods I

Tests of Significance

Significance tests

- Confidence interval: estimate a population parameter, such as the mean.
- Significance test: assess the truth of a statement about the population parameters
- E.g.: Young people are more left-leaning that older people. I.e., $x_{young} < x_{old}$

Significance Tests: Hypotheses

We typically state a *Null Hypothesis*, which is the statement being tested. Usually the null hypothesis is that there is no difference between two groups. - The null hypothesis is typically denoted as H_0 - e.g.: H_0 : there is no difference in the mean political leaning of young and old voters - Usually you also state an alternative hypothesis, H_1 - E.g., H_1: the means are not the same

The logic of significance tests

- Suppose we observe a difference between the two means of $+3$.
- Is this difference large or small?
- One way to answer the question is to compute the probability of obtaining a difference this large (or larger) than 3 assuming that, in fact, there is no difference in the true means.
- Suppose we find that this probability is 17%. Then we might conclude that this is not very rare and hence that the difference might simply be due to chance. If on the other hand we find that that probability is 0.0001, then we might reject the hypothesis that there is no difference
- Our goal here is therefore to figure out how to obtain these probabilities.

Obtaining a z-score

First we want to normalize our data to be able to compare it to a distribution we know

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{SD of the estimate}}$$

Test statistics: How to do it?

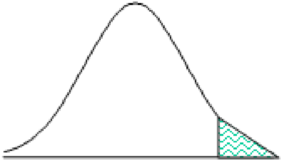
- First, standardize your variable:

$$z = \frac{estimate - hypothesizedvalue}{\text{standard deviation of the estimate}}$$

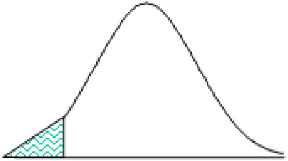
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- E.g.:
- H_0 : The average salary of TCD and UCD graduates is the same as the average salary of DCU students.
- We observe that the average difference in salary between TCD/UCD and DCU is 4000
- The standard deviation of that difference is 3000
- Question: is this difference significantly different from 0?
- so $z = \frac{4000-0}{3000} = 4/3$

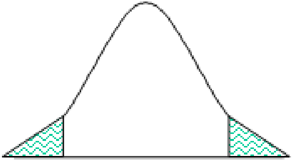
One-tailed vs two-tailed tests



Positive one-tailed test



Negative one-tailed test



Two-tailed test

The p-value

Suppose for example that we get $z = 4/3$ as above. We want to know the probability that we observe a value of z as extreme or more extreme than $4/3$. Because we are using a two-sided alternative, we want to find:

$$P(Z \leq -4/3 \text{ OR } Z \geq 4/3,$$

where $z \sim N(0, 1)$

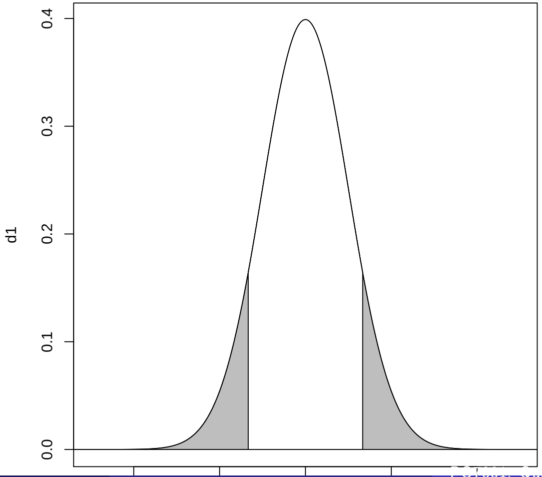
p-value

The p-value is the probability, assuming H_0 is true, that the test statistic would take a value as extreme as what is actually observed. The smaller the P-value, the stronger the evidence against H_0. (from Moore et al., p.413)

The key to calculating the P-value is the sampling distribution of the test statistic. In most cases we will only use the normal or t distribution (more on that later)

But how do I find $P(Z \leq -1.2ORZ \geq 1.2$?

Note that p is really the area under the sampling distribution curve (here the normal distribution) above and below $4/3$ (or whatever value you got for z).



But how do I find $P(Z \leq -4/3 ORZ \geq 4/3?$

- Graphically, it looks like $P(Z \leq -4/3 \text{ OR } Z \geq 4/3$ is fairly large, i.e., it seems probable that we would get $|Z| \geq -4/3$ by chance.
- Let's ask R to calculate the area below -4/3:

```r
pnorm(-4/3)
```

```
## [1] 0.09121122
```

This tells us that 9.1% of the area under the curve is below $z = -4/3$. Since the curve is symmetric, we know that 9.1% is also above 4/3. So in total, 18.2% of the data is within the grey area. In other words, if the null hypothesis were true, then an observation as extreme (or more) as ours (4/3) would occur in 18.2% of cases. Clearly this is not very rare, and so we *fail to reject the null hypothesis*. - Note: we never "accept" the null hypothesis

Another example

- Suppose now that collect a sample of IQ at Irish Universities and want to know whether these IQs are, on average, different from those of the average population. We have no prior expectation about the direction (i.e., smarter/dumber) and so use a two-sided test.
- We observe a sample of size 16 with mean 110 and standard deviation 15. Is this significantly different from the population mean of 100? And what is the probability that we would observe such a sample mean?

Another example (cont'd)

- First, calculate SE:

$$SE = s/\sqrt{n} = 15/4 = 3.75$$

$$z = \frac{110 - 100}{SE} = \frac{10}{3.75} = 2.66$$

  This is telling us that our sample mean is 2.66 standard deviations larger than the population mean.

- Is this a lot? $P(Z < -2.66 \text{ OR } Z > 2.66)$

```r
pnorm(-2.66)
```

```
## [1] 0.003907033
```

There is only a 0.3% chance that we would observe this sample by chance.

But wait, we don't know σ

- In the real world, you do not usually observe σ, which is a population parameter. All you have is your sample, from which you need to draw inferences about the general population.
- So we need to *estime* σ. A simple way of estimating σ is to simply use our sample's standard deviation. I.e., we will use s to estimate σ.
- Previously we used σ to calculate the standard deviation of the sampling distribution, in order to obtain a z-score ($z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$). Now we will just use s and get a statistic called t:

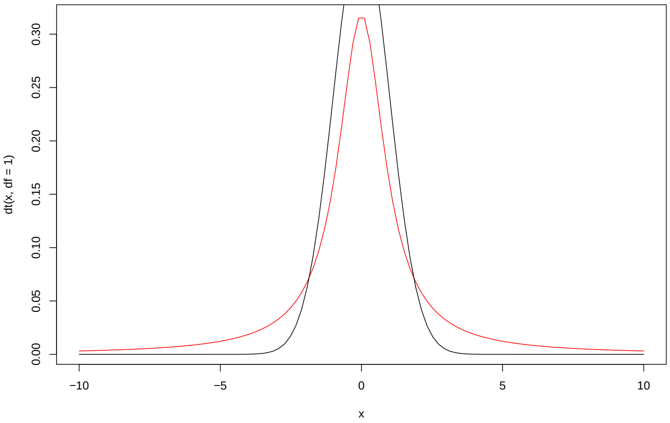
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

t has the t distribution with $n - 1$ degrees of freedom.

But wait, we don't know σ (cont'd)

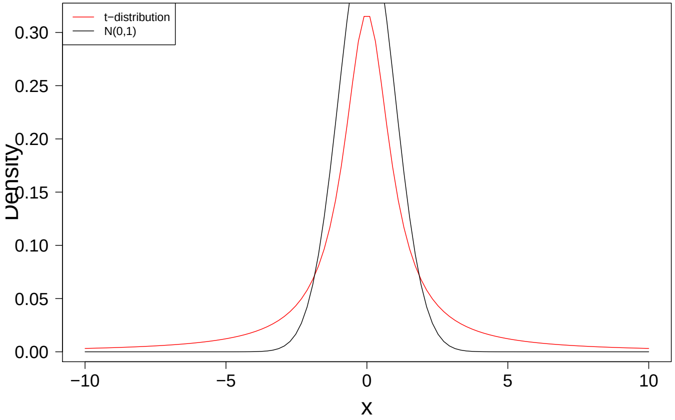
- As n becomes larger, the t distribution becomes more and more like the normal distribution. But for low levels of n, they can differ significantly
- The t distributions have more probability in the tails and less in the center. This greater spread is due to the extra variability caused by substituting the random variable s for the fixed parameter σ

the t distribution

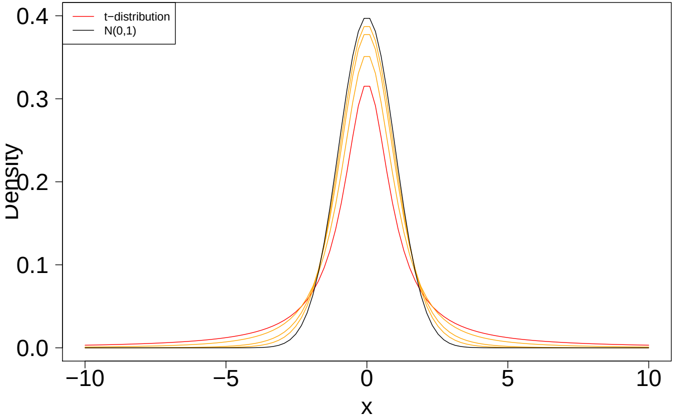


the *t* distribution

t distribution with 1 df:



the t distribution as *n* increases:



the one-sampled t-test

- To test the hypothesis $H_0 : \mu = \mu_0$ based on a random sample of size n, compute the one-sampled t statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- for a random variable T with a $t(n-1)$ distribution, the p value for a test of H_0 against $H_a : \mu \neq \mu_0$ is $2P(t \geq |t|)$

Significance test: an example

Suppose that we want to compare the mean political orientation in the US and the UK. More specifically, we want to test:

$$H_0 : \mu = 0.1$$

$$H_a : \mu \neq 0.1$$

Suppose that $n = 100$, $\bar{x} = 0.5$ and $s = 2$. Then the t statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.5 - 0.1}{2/10} = 2$$

This means that the sample mean $\bar{x}$ is 2 standard deviations away from the null hypothesized value $\mu = 0.1$. Because there are 99 degrees of freedom, this t statistic has the $t(99)$ distribution. So we now find $2P(T\ geq2)$ (question: why 2?):

```
1-pt(q = 2, df = 99)
```

```
## [1] 0.02411985
```

Example (cont'd)

Here I tested the null hypothesis again the alternative *two-sided* alternative that $\mu \neq 0.1$. Why two-sided? Because I had no prior beliefs about whether the average in the US would be higher/lower than in the UK. If I had suspected that the U.S. average would be smaller, I could have used a one-sided test. - However, you *should not* look at the data first then decide to do a one-sided test in the direction indicated by the data. If in doubt, use a two-sided test! (In fact, I'd almost always recommend a two-sided test)

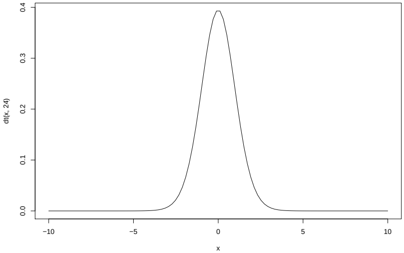
Another example

- Suppose we observe a sample of size 25 with mean $\bar{x} = 2$, standard deviation 4.
- Can we conclude that $\bar{x}$ is significantly different from 0?
- First we calculate the t statistic:

$$t = \frac{2 - 0}{4/\sqrt{25}} = \frac{2}{4/5} = 2.5$$

Another example (cont'd)

Let's look at 2.5 on the graph of a t-distribution with 24 degrees of freedoms, we find that that is a pretty high number:



Another example (cont'd)

To formalize this, we calculate the p-value, i.e., the probability that, if the null hypothesis were true, we would obtain more than 2.5 or less than -2.5 by chance. I.e., what is the area under the curve above 2.5 and below -2.5?

```
2*pt(-2.5, 24)
```

```
## [1] 0.01965418
```

That is a low probability indeed, meaning that there is only a 2% chance that we would observe such a value by chance. We say that the result is significant at the 5% level (actually, we could even say at the 2% level, but we typically use cutoffs such as 5%, 1%, 0.1%)

Comparing two means

- Unlike the matched pairs designs studied earlier, there is no matching of the units in the two samples, and the two samples may be of different sizes.
- Natural estimator of the difference $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2$, and so we need to know its sampling distribution.
- The variance of the difference $\bar{x}_1 - \bar{x}_2$ is the sum of their variances:

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- So the two-sample t-test is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\bar{\mu}_1 - \bar{\mu}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Inference for Proportions

Inference for a single Proportion

- Count data rather than measurements
- When you have at least 15 observations:
- The sample proportion is

$$\hat{p} = \frac{X}{n},$$

where X is the number of successes.

- The standard error of $\hat{p}$ is

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- So the confidence interval is

$$\hat{p} \pm m$$

Inference for a single Proportion: example

- In a sample of TCD students, researchers found that 4,000 were supportive of the CETA agreement, whereas 8,000 were opposed to it.
- The proportion of supporters is therefore

$$\hat{p} = \frac{4000}{12000} = 1/3$$

- To find the 95% CI, first compute the SE:

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = sqrt\frac{(1/3)(1-1/3)}{12,000} = 0.00430$$

  Approximately 95% of the time, $\hat{p}$ will be within 1.96 standard errors of the true p. So the CI is:

$$\hat{p} \pm z \times SE_{\hat{p}} = 1/3 \pm 1.96 \times 0.00430 = (0.3249, 0.3418)$$

- So we estimate with 95% confidence that between 32.5% and 34.2% of students support the CETA agreement.

Significance test for a single proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Confidence interval for a difference in proportions

- The estimate of the difference is

$$D = \hat{p}_1 - \hat{p}_2$$

- The standard error of D is

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- So the CI is:

$$D \pm z \times SE_D$$

The “normal distribution” assumption

- Most of our tests rely on the assumption that the underlying data is normally distributed.
- But most data is not normally distributed and either varies a bit or dramatically from it.
- In most cases, this is not a problem. Why? Because our tests rely on the normality of the *sampling distribution*, NOT of the underlying data.
- Thanks to the Central Limit Theorem, we know that most sampling distributions will be normally distributed provided that:
- The distribution is not overly skewed / there are no large outliers
- We have enough observations

What if you data is not normal AND small sample?

- use another distribution.
- Try to normalize the data. (Esp. for skewed data): take the log
- Use a nonparametric procedure. These don't assume that the distribution of the population has any specific form.

Chi-squared test of independence

Chi-squared test of independence: Introduction

- A measure of dependence between two categorical variables
- The chi-squared test (also written as χ^2-test) is defined as:

  $$\chi^2 = \sum_i \frac{(n_{observed} - n_{expected})^2}{n_{expected}}$$

  The degrees of freedom are calculated as $df = (r - 1)(c - 1)$, where r is the number of rows in the table and c the number of columns.
- If the expected counts and the observed counts are very different, a large value of χ^2 will result. Large values of χ^2 provide evidence against the null hypothesis

χ^2-test: an example

Do dictatorship experience war more often on their territories than democracies?

	Democracy	Dictatorship
No War	40	74
War	3	11

We want to calculate

$$\chi^2 = \sum_i \frac{(n_{observed} - n_{expected})^2}{n_{expected}}$$

But how do we calculate the expected values?

χ^2: calculating expected values for a cross-table

Basic intuition: if the two variables were independent, their relative proportions should be similar to the *marginal* distributions. E.g., the proportion of democracies at war should be similar to the proportion of countries at war. I.e., the probability of a democracy at war if democracy and war were independent would be:

$$\hat{p}_{democracy\&war} = \hat{p}_{democracy} \times \hat{p}_{war}$$

- So first, what is $\hat{p}_{democracy}$? $\hat{p}_{democracy} = 43/128$ and $\hat{p}_{war} = 14/128$.
- So if the two variables were independent, we would expect the bottom-left cell to be: $128 \times \hat{p}_{democracy\&war} = 128 \times \frac{43}{128} \times \frac{14}{128} = 4.7$

χ^2: calculating expected values for a cross-table (cont'd)

- We do the same for every cell and obtain the table of *expected* values:

	Democracy	Dictatorship
No War	38.3	75.7
War	4.7	9.3

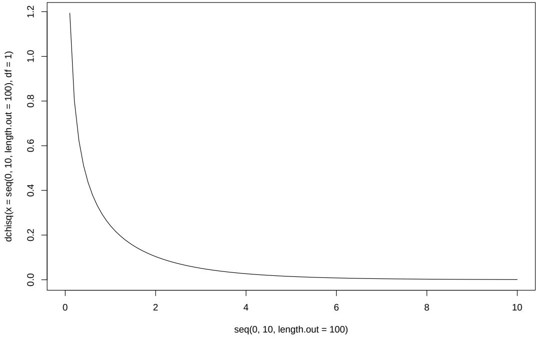
Calculating χ^2

Now we calculate χ^2 as:

$$\chi^2 = \frac{(40-38.3)^2}{38.3} + \frac{(74-75.7)^2}{75.7} + \frac{(3-4.7)^2}{4.7} + \frac{(11-9.3)^2}{9.3} = 1.042$$

- How many df do we have? $(2-1) \times (2-1) = 1$.

So what does a $\chi^2(1)$ look like?



```r
1-pchisq(1.042, df = 1)
```

```
## [1] 0.3073568
```

So we fail to reject the Null hypothesis that the two variables are independent.

χ^2: Another example

Observed:

	Republican	Democrat	Independent
Male	200	150	50
Female	250	300	50

χ^2: Another example (cont'd)

Expected:

	Republican	Democrat	Independent
Male	.4*.45*1000 =180	.4*.45*1000=180	.4*.1*1000=40
Female	.6*.45*1000=270	.6*.45*1000=270	.6*.1*1000=60

so

$$\chi^2 = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} +$$
$$\frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60} = 16.2$$

How many df? 2

```
1-pchisq(16.2,2)
```

```
## [1] 0.0003035391
```