Lecture 6: Hypothesis Testing

Thomas Chadefaux

PO7001

Tests of Significance

Confidence intervals and 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., $x_{young} < x_{old}$

Confidence intervals and Significance tests

- Confidence interval: use if you want to estimate a population parameter
- Hypothesis test: You have a hypothesized value and want to determine the likelihood that a population with that parameter would produce a sample as different as your sample

Confidence intervals and Significance tests

Examples

- What is the average political orientation of young voters?
  - Estimate mean and report confidence interval
- Does the average political orientation of young voters differ from the one of older voters
  - Use a hypothesis test

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

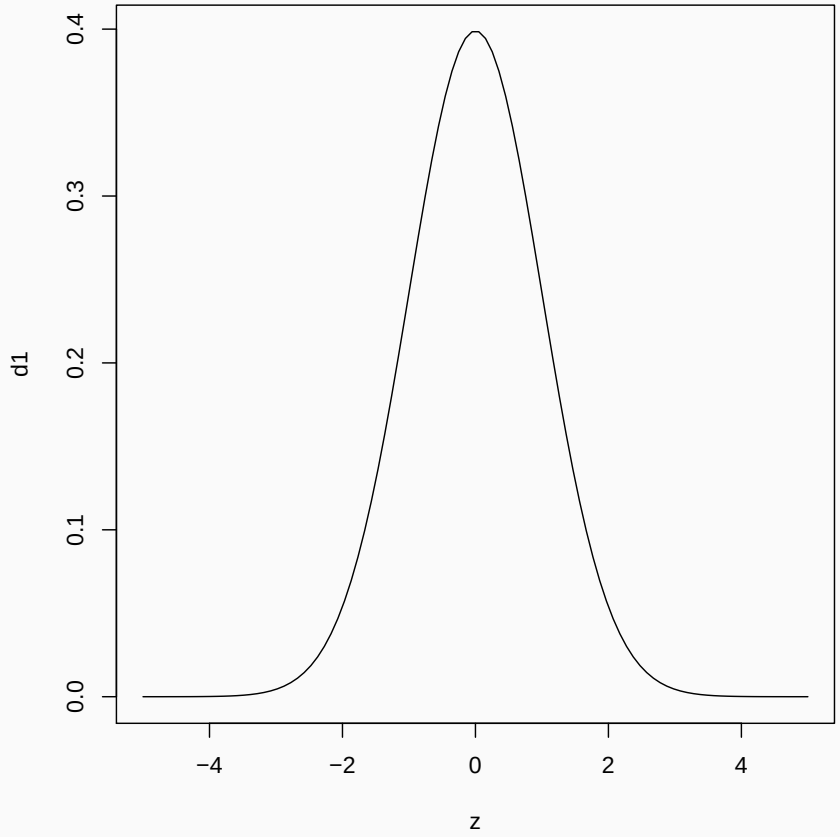
Significance Tests: Hypotheses

Intuitively, we are asking the following question: Suppose the true mean political leaning of young voters was 0.

But of course, even if the true mean is 0, sampling error is such that, when I take a sample, I will almost always get a mean different from 0. Sometimes -0.1, sometimes 1.2, etc.

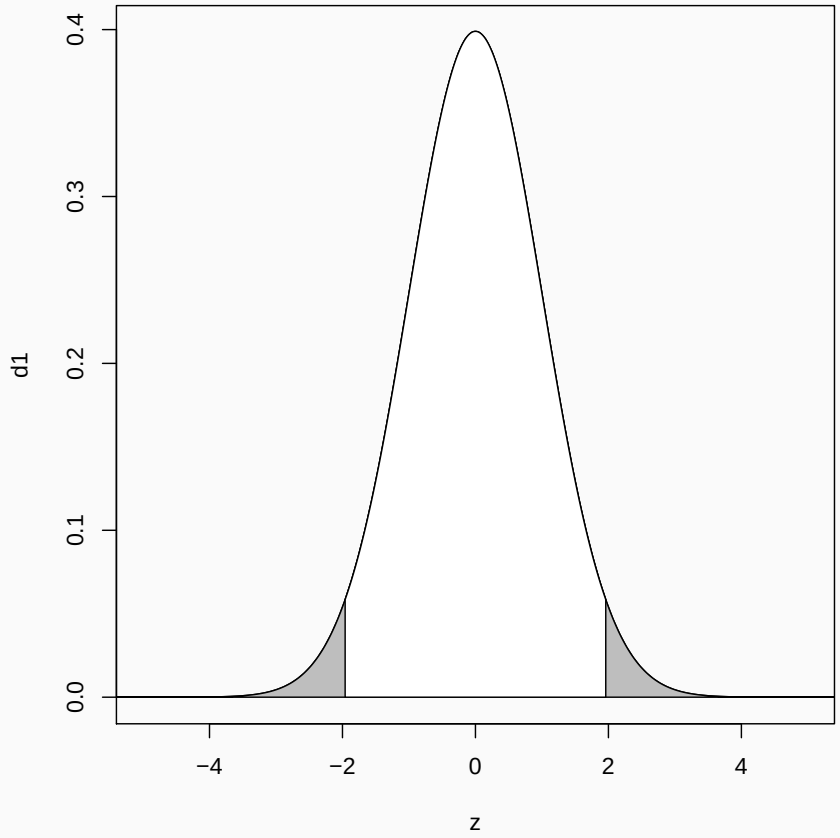
I.e., I should observe sample means that follow a distribution such as:

Significance Tests: Hypotheses



Significance Tests: Hypotheses

If the null hypothesis were true, it is unlikely that I would observe a sample mean in the grey regions:



The logic of significance tests

So when I get my sample mean, I want to know whether it falls within that grey region. If it does, then I'll conclude it is unlikely that this sample was drawn from that distribution, and hence conclude that I can "reject" the Null hypothesis.

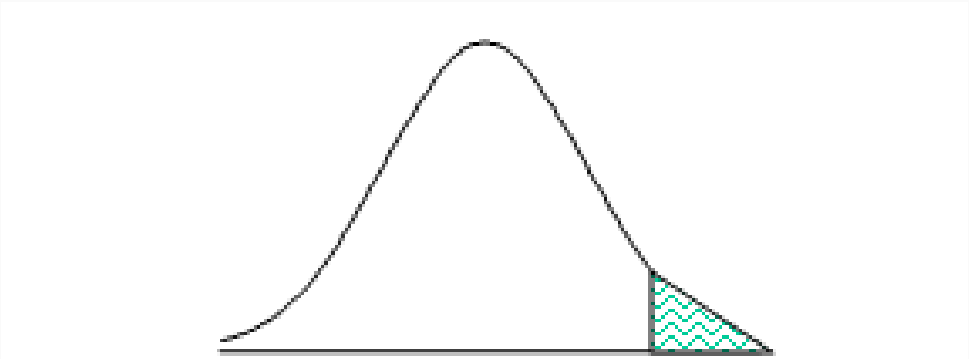
If, however, my sample mean is NOT in that grey region, then it is very plausible that I would get a sample mean this size by chance, and hence will NOT reject the null hypothesis.

The logic of significance tests

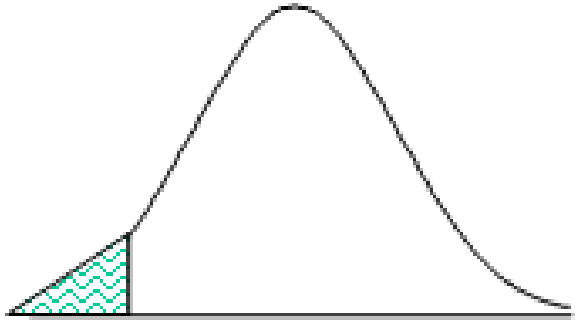
Same thing, slightly differently:

- 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.

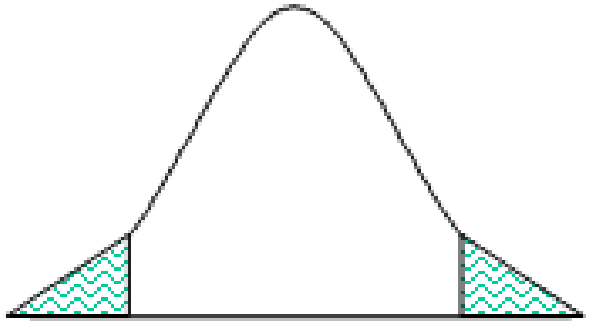
(aside):One-tailed vs two-tailed tests



Positive one-tailed test



Negative one-tailed test



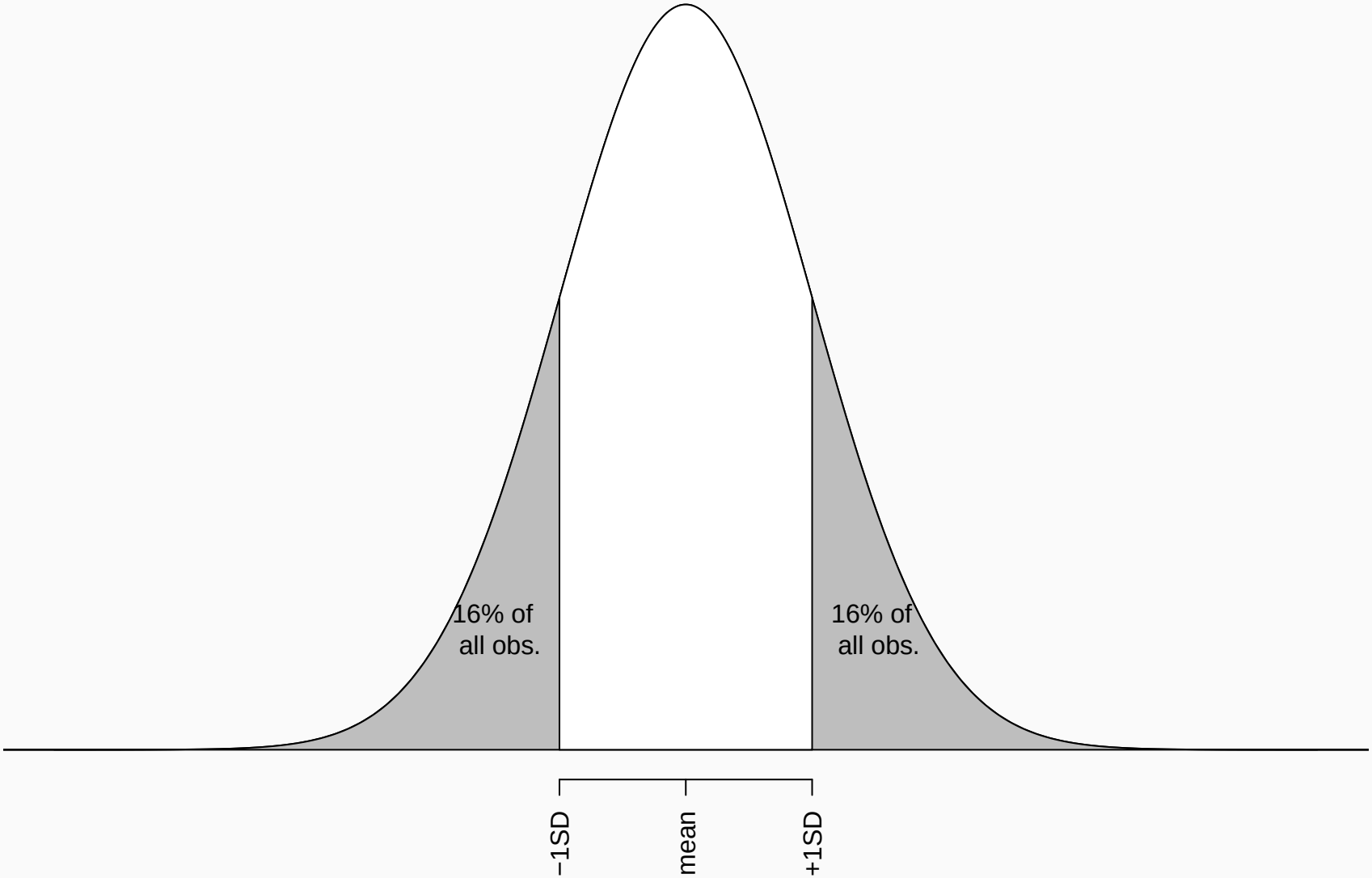
Two-tailed test

Proportions under the normal distribution

Proportions under the normal distribution

```
## Loading required package: spam

## Loading required package: dotCall64

## Loading required package: grid

## Spam version 2.3-0 (2019-09-13) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduc
## and overview of this package.
## Help for individual functions is also obtained by adding
## suffix '.spam' to the function name, e.g. 'help( chol.sp

##
## Attaching package: 'spam'

## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

Proportions under the normal distribution

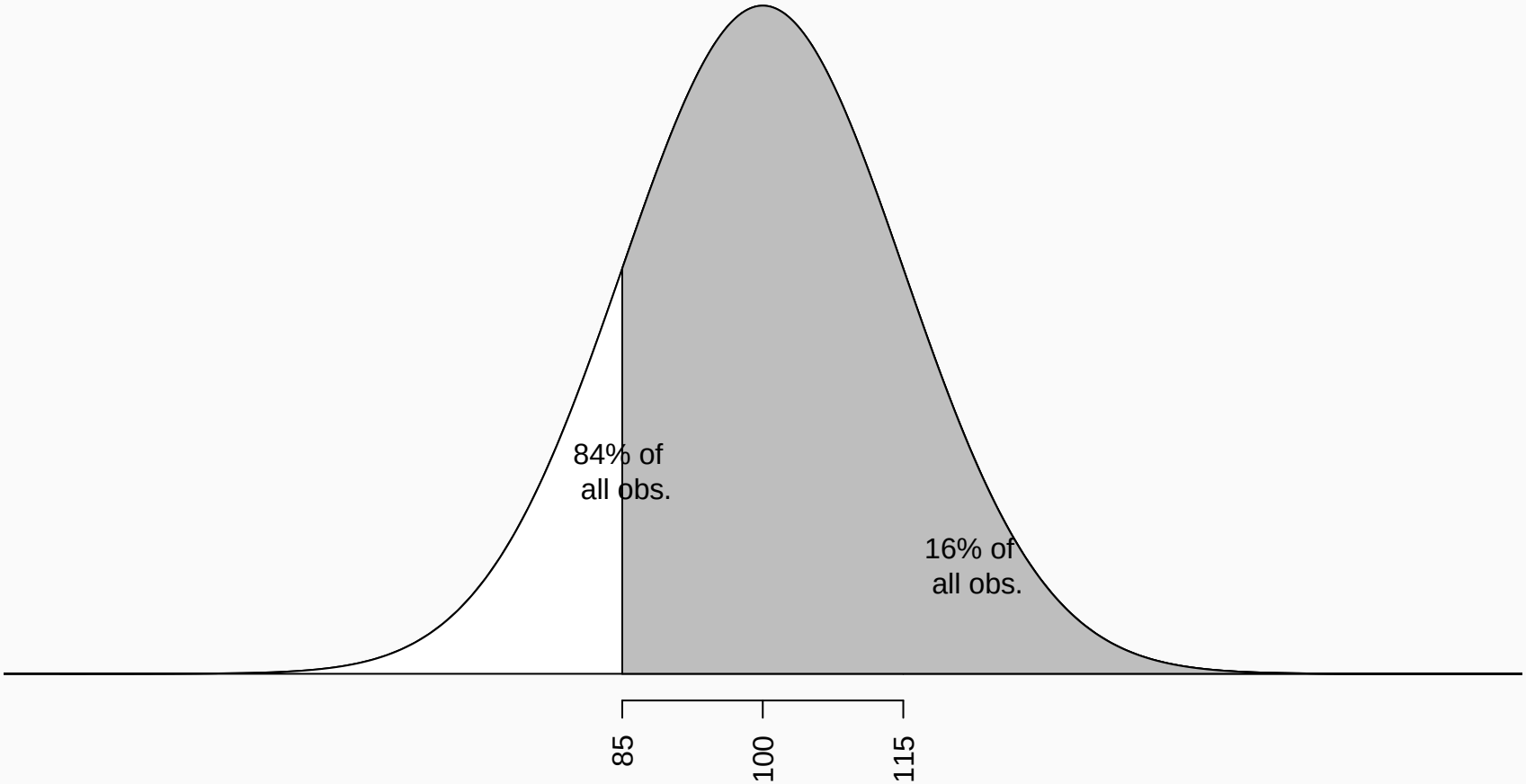


An example

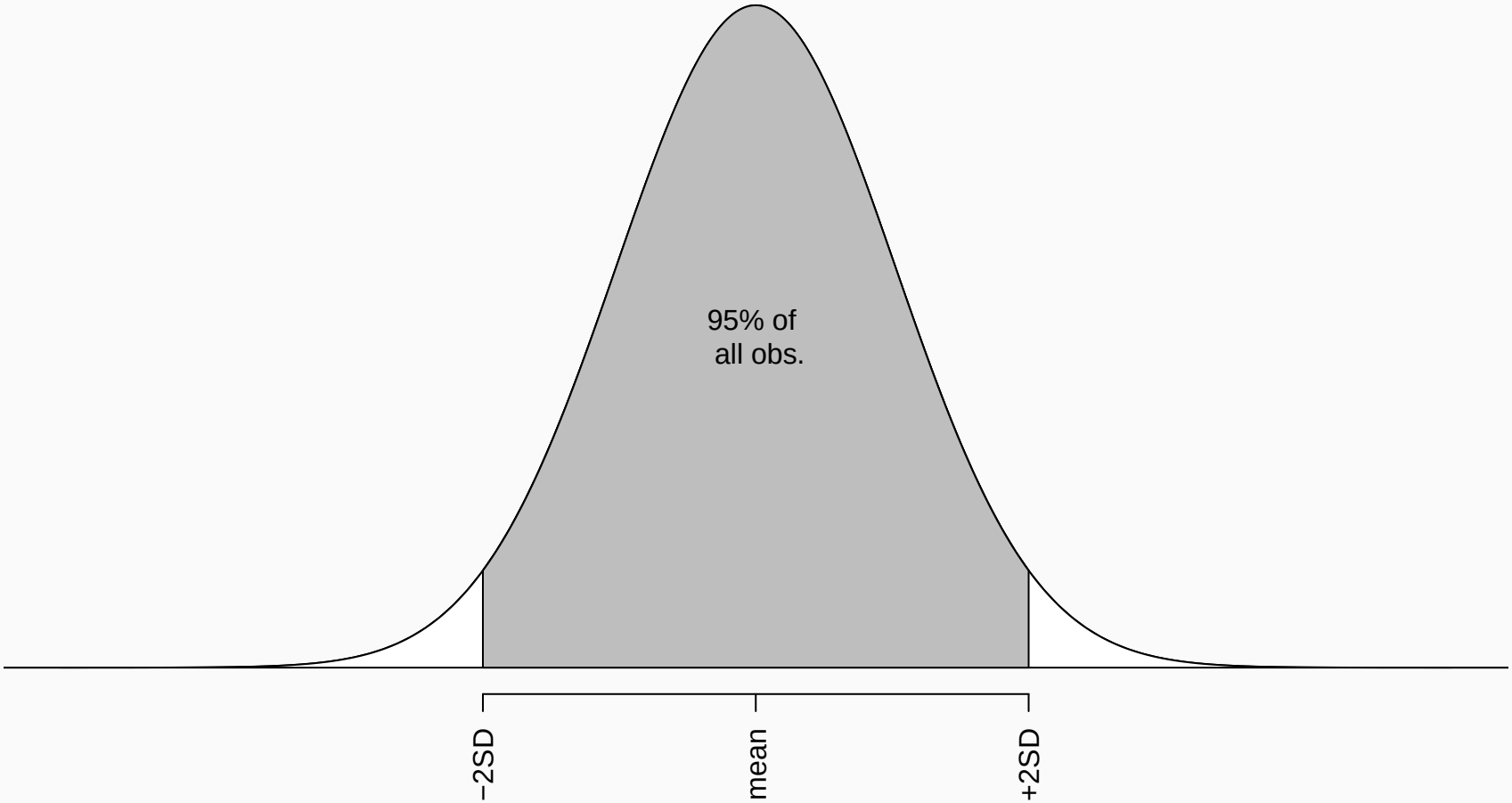
Mean IQ= 100, SD = 15. What proportion of the population has $IQ > 85$?

An example

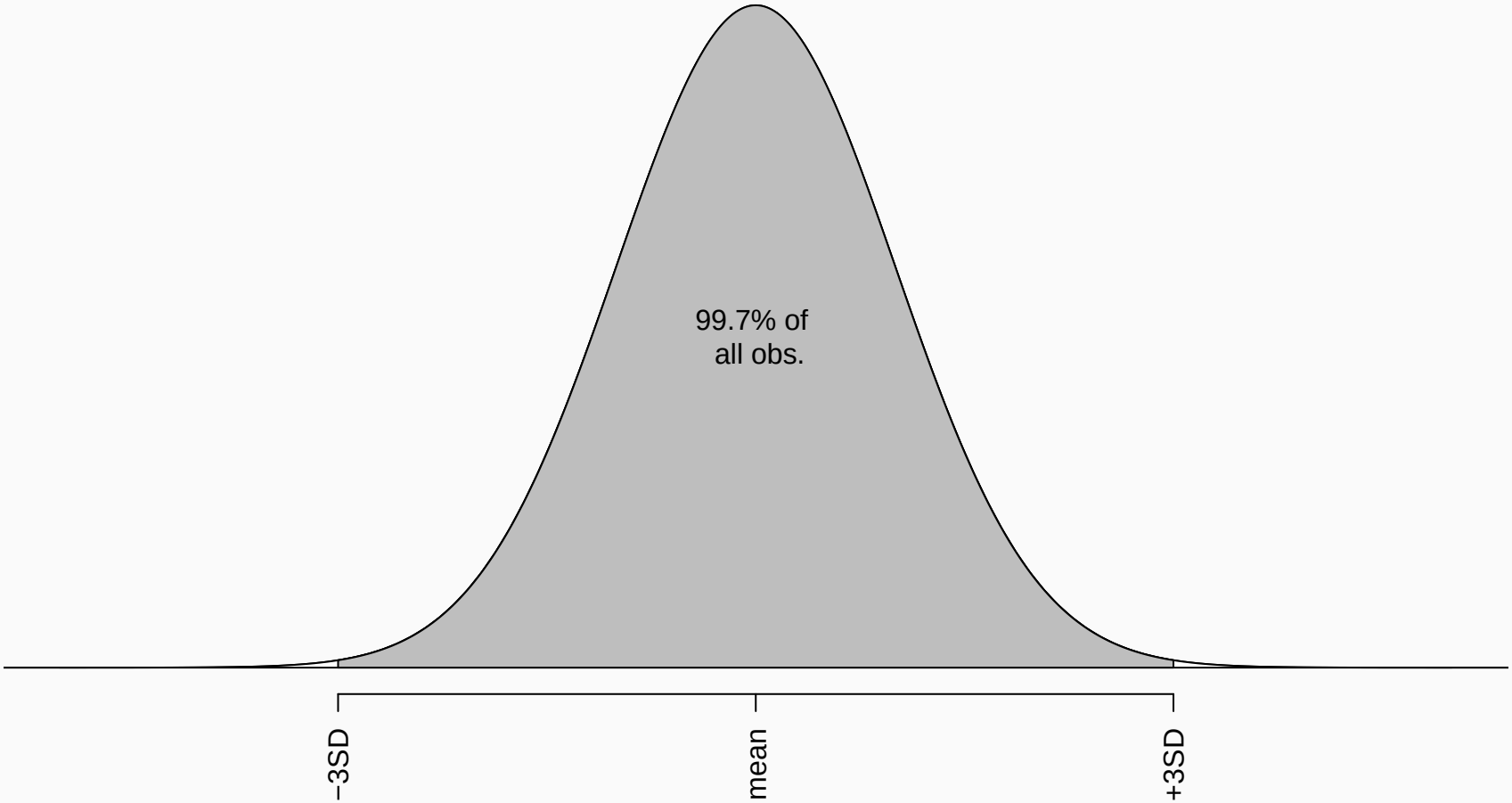
Mean IQ= 100, SD = 15. What proportion of the population has $IQ > 85$?



95% of the area under the curve



99% of the area under the curve



Comparing values

Knowing the mean (μ) and standard deviation (σ) of a distribution gives us information about its internal proportions. E.g., what percentage of people have $IQ < 90$, etc.

It also allows us to **compare values** that come from different distributions

Comparing values. An example

- Niamh gets a 67 on her Econ exam
- Eoin gets a 72 on his Law exam

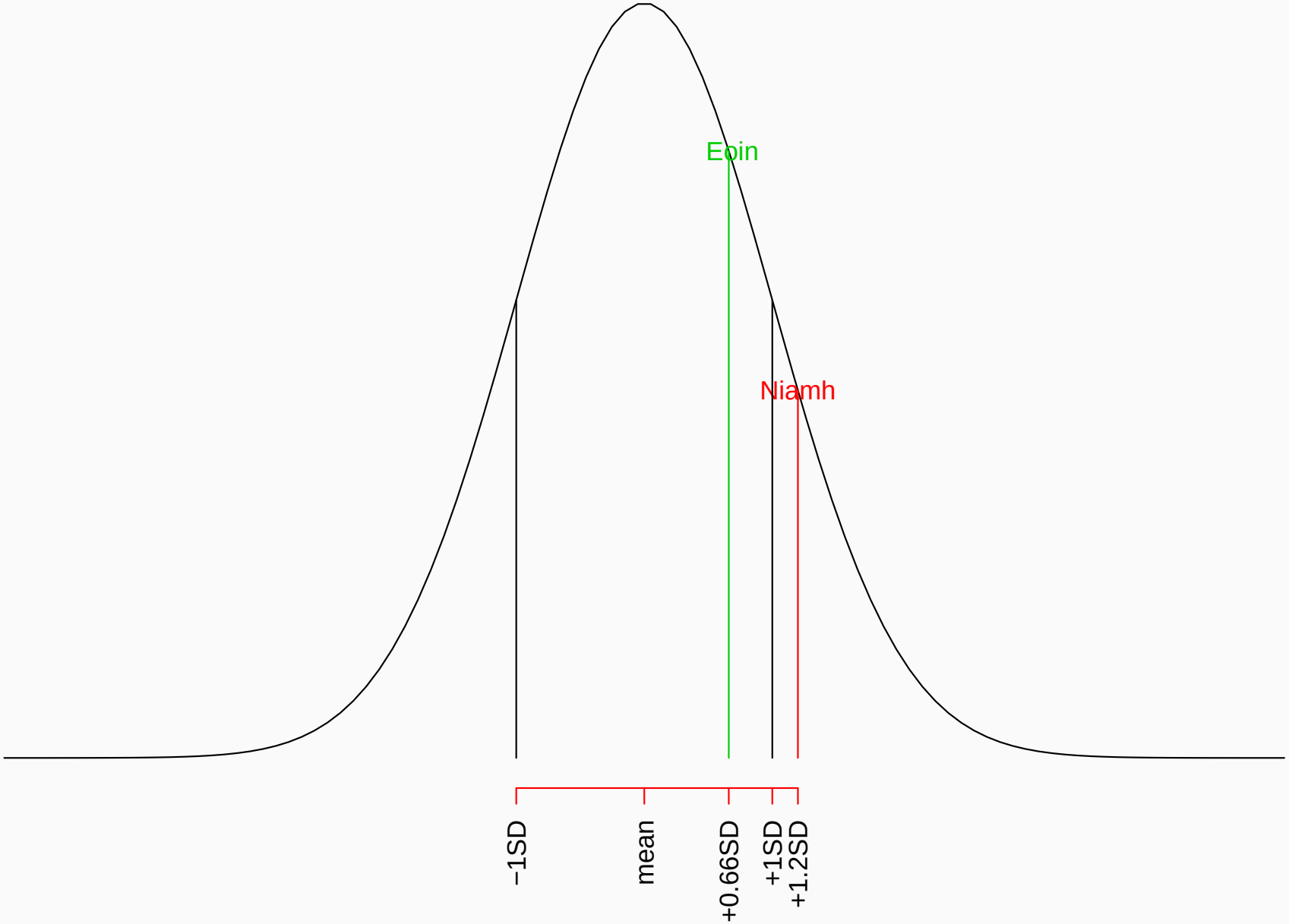
Assume that the distribution of marks in both modules is normal. Which student has the "better" score? Technically it is Eoin, but maybe Eoin received the worst mark of his law class, and Niamh the best mark of her Econ class. So we need info about other students' marks

Comparing values. An example

Suppose that we know that: 1. the mean mark in Econ is 55, sd = 10 2. the mean mark in Law is 60, sd=18.

Now can we compare them better?

Comparing values. An example



Comparing values. An example

Note that we measured scores in standard deviations. We have translated the observations of both distributions into numbers of standard deviations above or below the means of their distributions.

These "standard deviations" measures are often call z-values.

z-scores

z-scores: intro

A z-score is a unit-free, standardized score that, regardless of the original units of measurement, indicates how many standard deviations a score is above or below the mean of its distribution.

$$z = \frac{X - \mu}{\sigma}$$

X is the original score, μ is the mean, and σ is the standard deviation.

The z score tells you how many standard deviations above/below the mean it is.

z-scores: an example

Going back to Niamh and John, we can now easily calculate their z-scores:

- Niamh: $\frac{67-55}{10} = 1.2$
- Eoin: $\frac{72-60}{18} = 2/3$

z-scores: the standard normal distribution

The standard normal distribution is a special case of the normal distribution, for which $\mu = 0$ and $\sigma = 1$

If the original distribution approximates a normal curve, then the shift to standard or z-scores will produce a distribution that is standard normal, i.e. has mean 0 and SD 1.

z-scores: the standard normal distribution

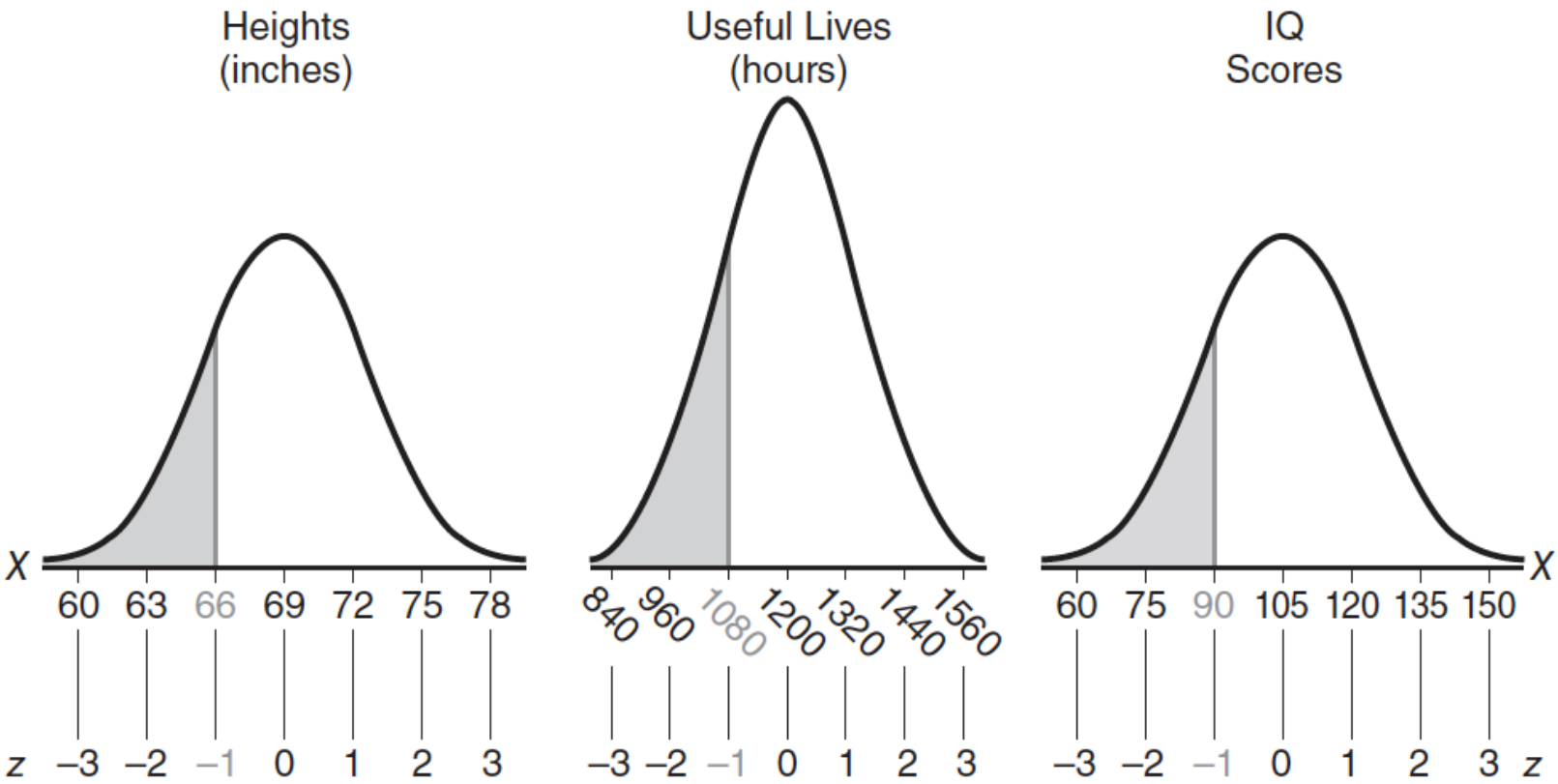


FIGURE 5.4

Converting three normal curves to the standard normal curve.

Source: Witte, p. 87:

Significance tests: How to do it?

Obtaining a z-score

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

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

As discussed, these z-scores tell us how many SD away from the mean each observation is.

Test statistics: How to do it?

- First, standardize your variable:

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

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

- 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 €400 (n = 100)
  - The standard deviation of that difference is €9000
  - Question: is this difference significantly different from 0?

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad (1)$$

$$= \frac{400 - 0}{9000/30} = 4/3 \quad (2)$$

The p-value

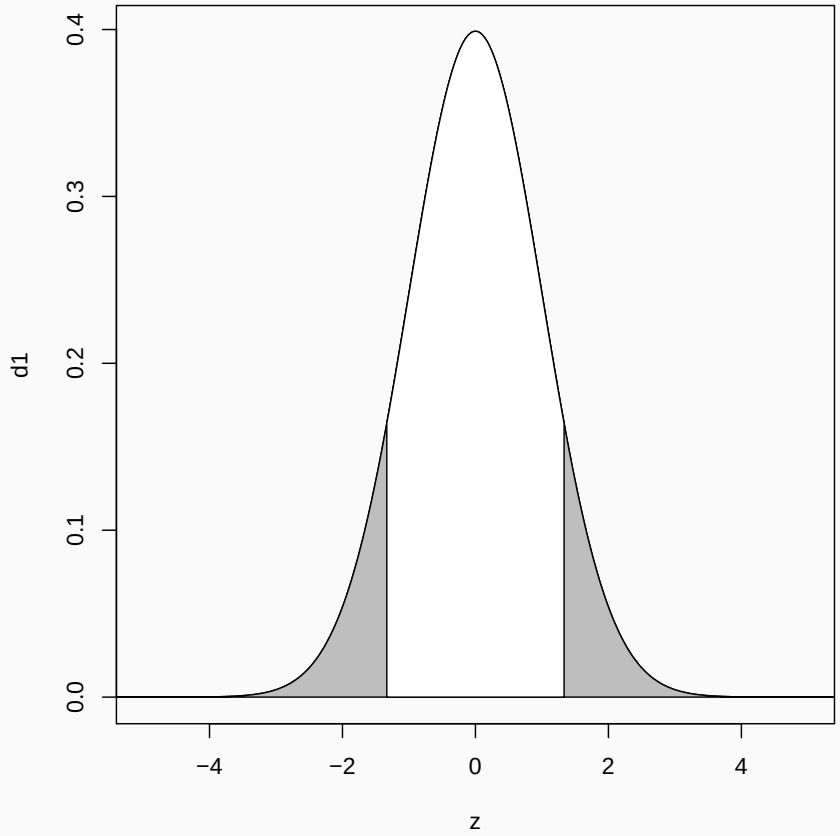
So we got $z = 5$ as above. We want to know the probability that we would observe a value this extreme (i.e., either greater than 4/3 or smaller than -4/3) if the null hypothesis were true.

I.e., because we are using a two-sided alternative, we want to find:

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

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

The p-value



- 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.

The p-value

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 $P(Z \leq -4/3 OR Z \geq 4/3$?

- 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).
- So let's ask R to calculate the area below -4/3:

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

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

The p-value

```
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

Jumping ahead

Note that -4/3 is (very loosely, more on why it's loosely later – don't forget!) the value of the t-statistic later, and 0.18 is the p-value associated with that statistic

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? $P(Z < -2.66 \text{ OR } Z > 2.66)$

```
2*pnorm(-2.66)
```

```
## [1] 0.007814065
```

There is only a 0.78% 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 *estimate* 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.

(if you must know: how the t statistic arises)

Intuition: we don't know σ so we cannot use it in our formula. So Gosset used another variable, V below, with a know distribution, that cancels out the *sigma*. We are left with a statistic without the population variance, which is one we want.

(if you must know: how the t statistic arises)

(From Wikipedia:)

Let $X_1, \ldots, X_n$ be independent realizations of the normally-distributed, random variable X, which has an expected value μ and variance σ^2. Let

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

be the sample mean, and

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

be an unbiased estimate of the variance from the sample.

(if you must know: how the t statistic arises)

It can be shown that the random variable

$$V = (n-1)\frac{S_n^2}{\sigma^2}$$

has a chi-squared distribution with $\nu = n-1\nu = n-1$ degrees of freedom. Then we can show that

$$Z = \left(\overline{X}_n - \mu\right)\frac{\sqrt{n}}{\sigma}$$

is normally distributed with mean 0 and variance 1, since the sample mean $\overline{X}_n$ is normally distributed with mean μ and variance σ^2/n. Moreover, it is possible to show that these two random variables (the normally distributed one Z and the chi-squared-distributed one V) are independent.

(if you must know: how the t statistic arises)

Consequently the pivotal quantity

$$T \equiv \frac{Z}{\sqrt{V/\nu}} = \left(\overline{X}_n - \mu\right) \frac{\sqrt{n}}{S_n},$$

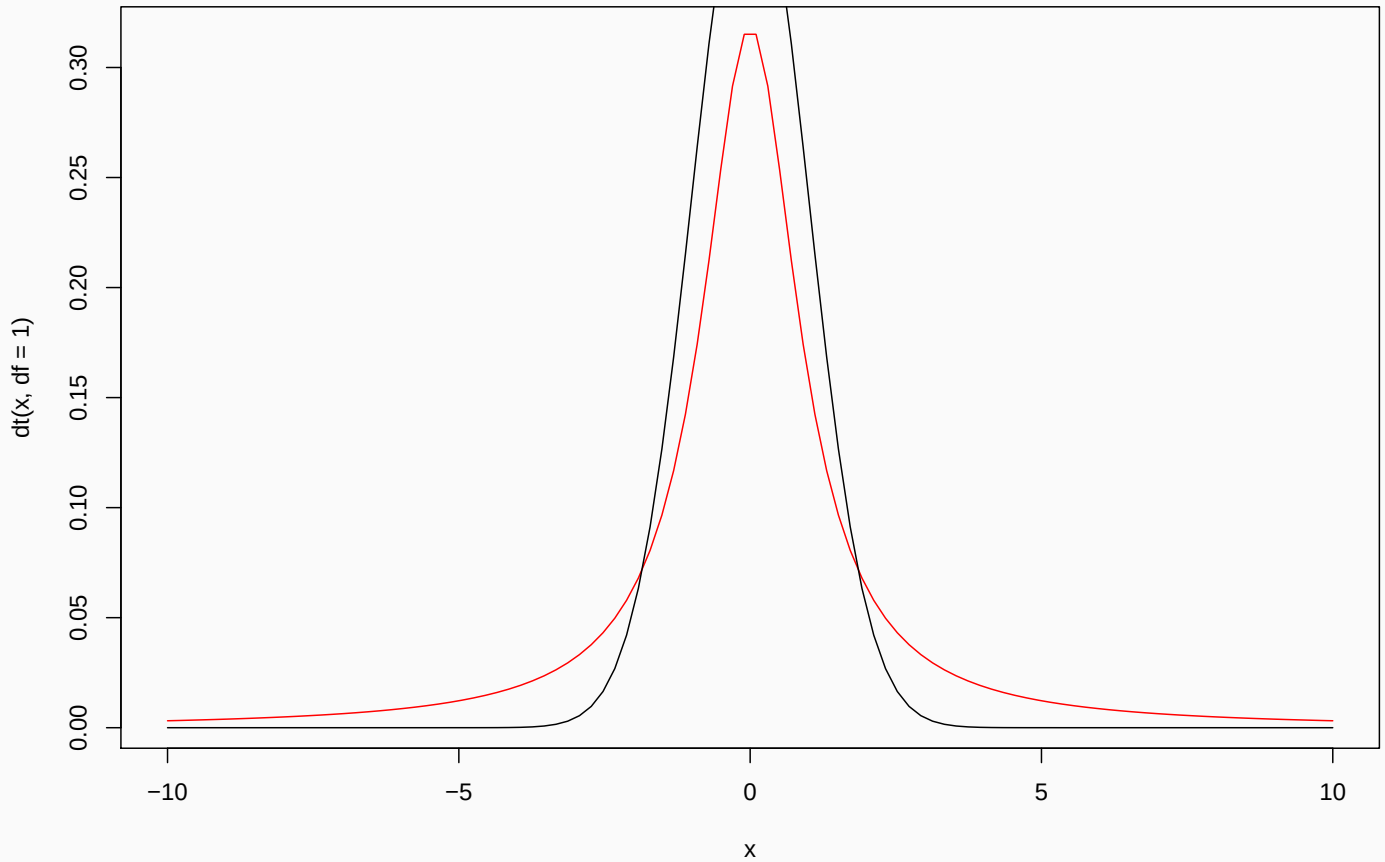
which differs from Z in that the exact standard deviation σ is replaced by the random variable S_n, has a Student's t-distribution.

Notice that the unknown population variance σ^2 does not appear in T, since it was in both the numerator and the denominator, so it canceled.

But wait, we don't know σ (cont'd)

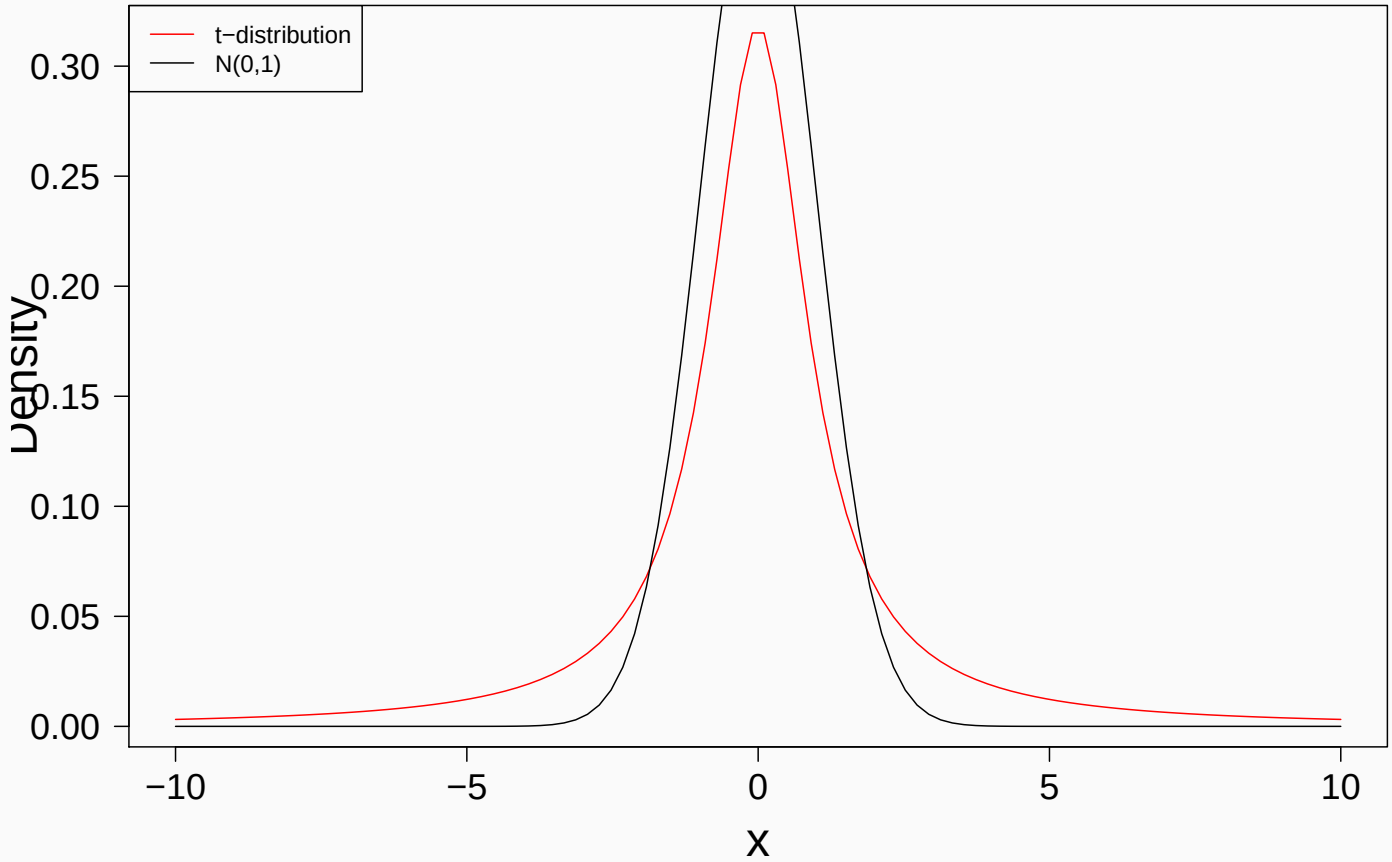
- As n becomes larger, the t distribution becomes more and more like the normal distribution. But for low levels of n, they can differ significantly
- The t distributions have more probability in the tails and less in the center. This greater spread is due to the extra variability caused by substituting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

The t distribution

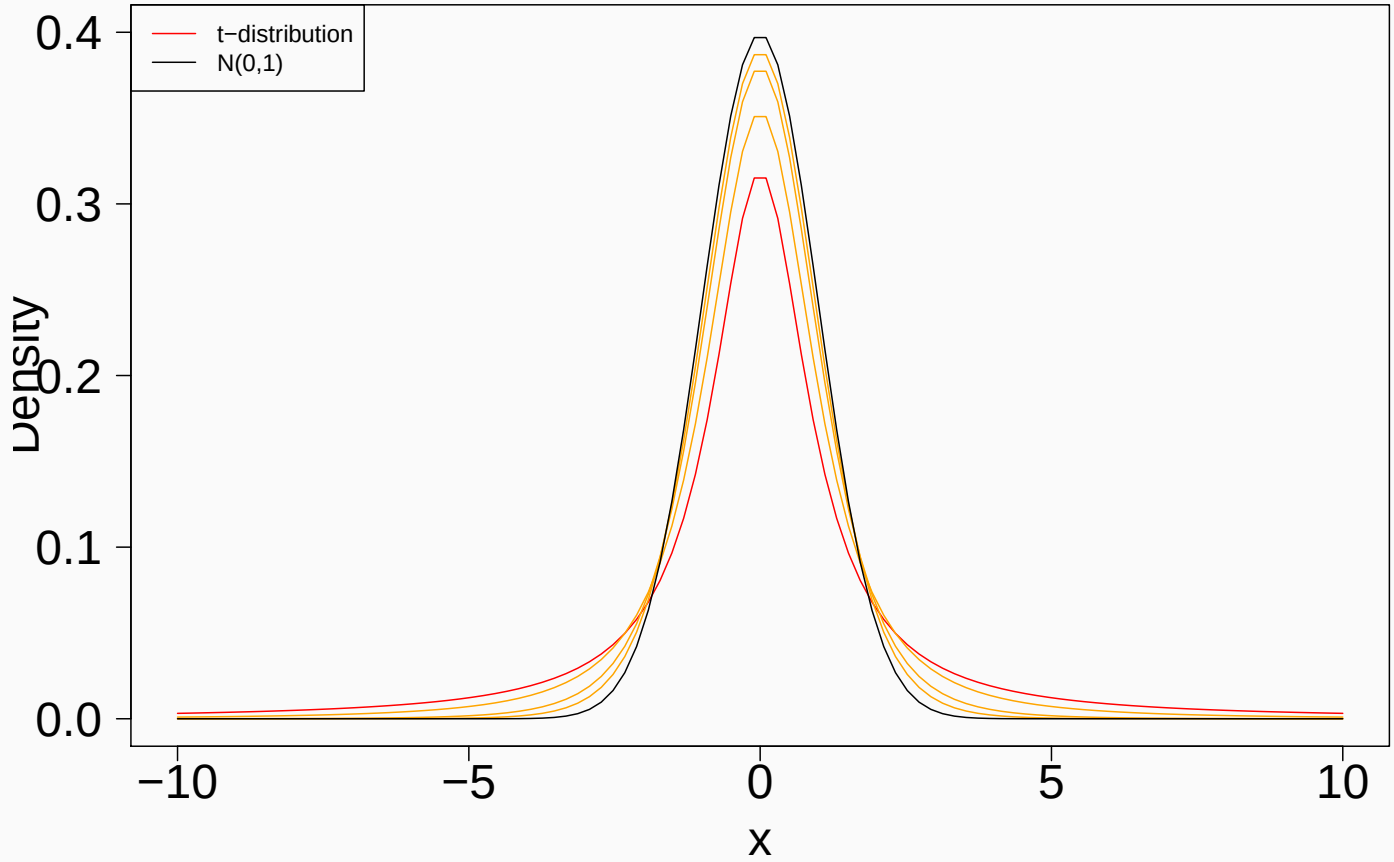


the t distribution

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 (100 - 1), this t statistic has the $t(99)$ distribution.

Example (cont'd)

So we now find $2P(T \geq 2)$ (question: why 2?):

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

```
## [1] 0.04823969
```

(question: why 1-pt?)

Compare with "normal" assumption:

```
2*(pnorm(-2))
```

```
## [1] 0.04550026
```

This result in incompatible with a mean of 0.1 at the $\alpha = 0.05$ level.

Note on 1 vs 2-sided

- 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)

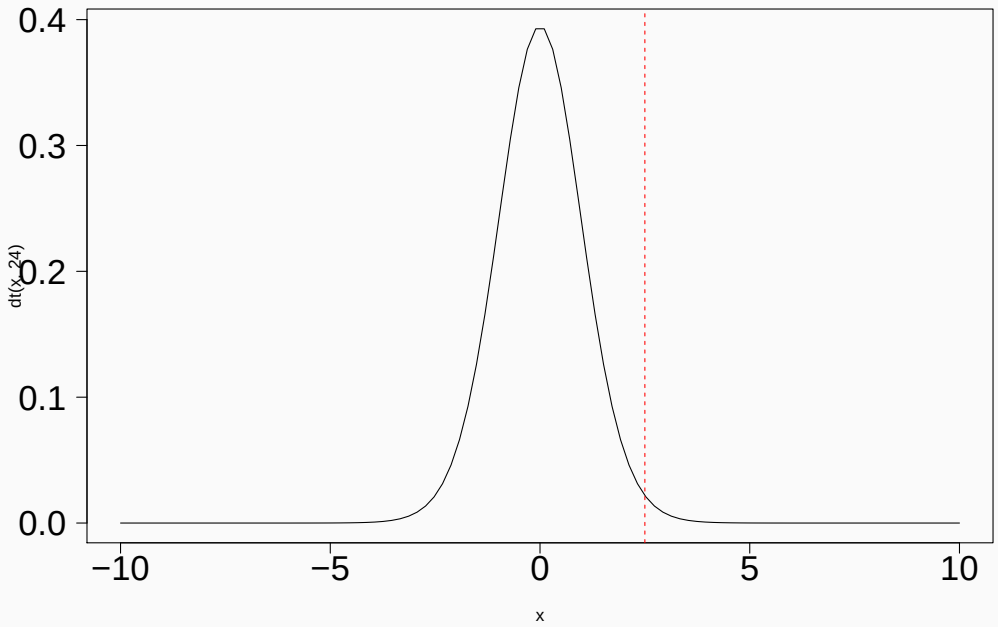
Another example

- 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 large 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%)

t or z?

When you don't know the standard deviation of the population (and you usually don't), use a t-test. In short, almost always use a t-test.

In practice, it only matters if $n < 30$. For $n \geq 30$, there is practically no difference.

Comparing two samples

Two samples, matched

Same units sampled twice. E.g., before and after measure of the same units.

We want to know if there is a difference before/after, and so simply calculate a t-test on the mean *difference*. It's EXACTLY the same as before, except that now we use the mean difference ($\overline{x}_D$), rather than the mean:

$$t = \frac{\overline{x_D} - \mu_D}{s_D/\sqrt{n}}$$

Comparing two means, independent samples

- Unlike the matched pairs designs above, there is no matching of the units in the two samples, and the two samples may be of different sizes.
- The natural estimator of the difference is $\mu_1 - \mu_2$ 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:

$$\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

- So the two-sample t-test is:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\bar{\mu_1} - \bar{\mu_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

two-sample t-test: an example

```
x1 <- c(1,2,3,4)
x2 <- c(3,5,6,7)
```

1. state the hypothesis:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

two-sample t-test: an example

Step 2: choose a significance level α. Let's say $\alpha = 0.05$

two-sample t-test: an example

Step 3: gather what we need, i.e., the mean and the standard deviation of each sample:

```
mean(x1)
```

```
## [1] 2.5
```

```
mean(x2)
```

```
## [1] 5.25
```

```
var(x1)
```

```
## [1] 1.666667
```

```
var(x2)
```

```
## [1] 2.916667
```

two-sample t-test: an example

Now apply the formula:

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\bar{\mu_1} - \bar{\mu_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

$$\approx \frac{(2.5 - 5.25) - 0}{\sqrt{1.66/4 + 2.99/4}} \tag{4}$$

$$\approx -2.57 \tag{5}$$

two-sample t-test: an example

Let's check with R:

```
t.test(x1,x2)
```

```
##
##  Welch Two Sample t-test
##
## data:  x1 and x2
## t = -2.569, df = 5.5846, p-value = 0.04522
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -5.41722062 -0.08277938
## sample estimates:
## mean of x mean of y
##      2.50      5.25
```

(NB: degrees of freedom for a two-sample t-test)

In case you are wondering how R calculated the degrees of freedom here. Because we are not assuming equal variance, the distribution of the test statistic is approximated as an ordinary Student's t-distribution with the degrees of freedom calculated using

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

This is known as the Welch–Satterthwaite equation.

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

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

(NB: How did we derive this?)

First note that we are dealing with a sample in which there are 1s and 0s. Suppose the number of 1s is m, and the number of 0s is $n - m$. I.e. the mean is $\overline{x}(= \hat{p}) = m/n$

(NB: How did we derive this?)

Then the standard deviation will be:

$$\sum_i (x - \overline{x}) \tag{6}$$

$$= m(1 - \overline{x})^2 + (n - m)(0 - \overline{x})^2 \tag{7}$$

$$= m(1 - m/n)^2 + (n - m)(0 - m/n)^2 \tag{8}$$

$$= m(1 + m^2/n^2 - 2m/n) + (n - m)(m^2/n^2) \tag{9}$$

$$= m + -m^2/n \tag{10}$$

$$= m(1 - m/n) \tag{11}$$

$$= n\hat{p}(1 - \hat{p}) \tag{12}$$

(NB: How did we derive this?)

Divide all this by n to get the variance:

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

(NB: How did we derive this?)

Then the SE, just like for the mean, will be:

$$SE = \frac{s_{\hat{p}}}{\sqrt{n}} \tag{13}$$

$$= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{14}$$

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.

Inference for a single Proportion: example

- Check with R:

```
prop.test(4000, 12000)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  4000 out of 12000, null probability 0.5
## X-squared = 1332.7, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3249121 0.3418617
## sample estimates:
##         p
## 0.3333333
```

Significance test for a single proportion

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

The “normal distribution” assumption (reminder)

The "normal distribution" assumption (reminder)

- 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. More on this in the next lecture!

An aside: taking the log

A logarithm is the power to which a number must be raised in order to get some other number

For example, the base ten logarithm of 100 is 2, because ten raised to the power of two is 100:

$$log_{10}(100) = 2.$$

similarly

$$log_{10}(10000) = 4$$

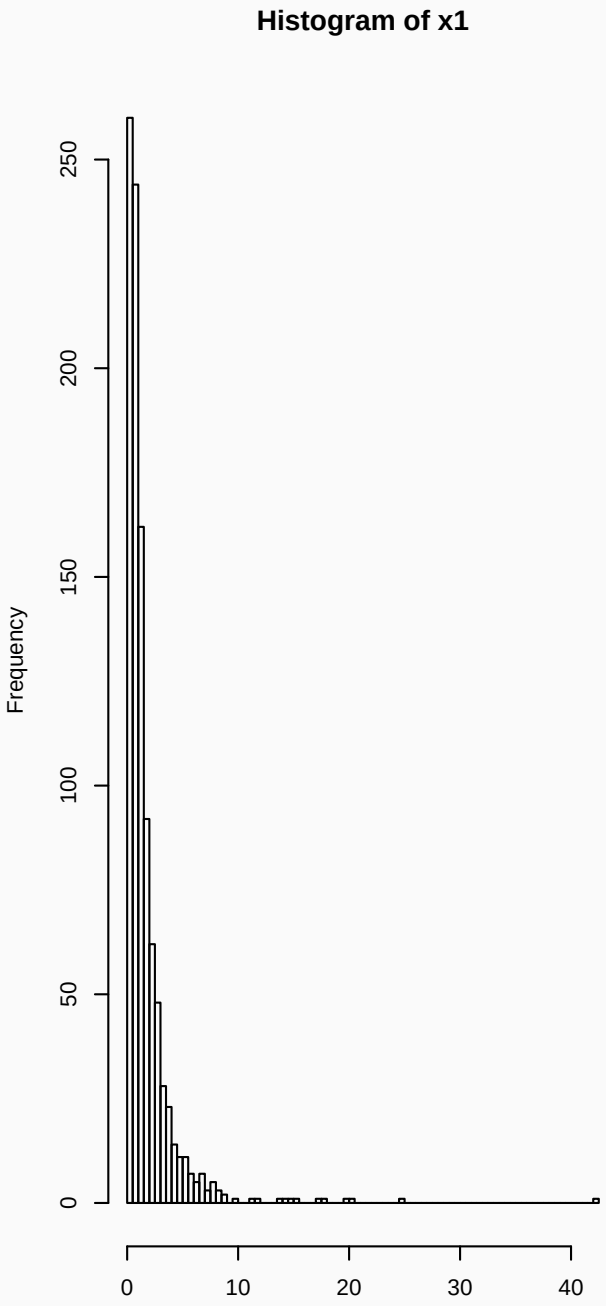
and

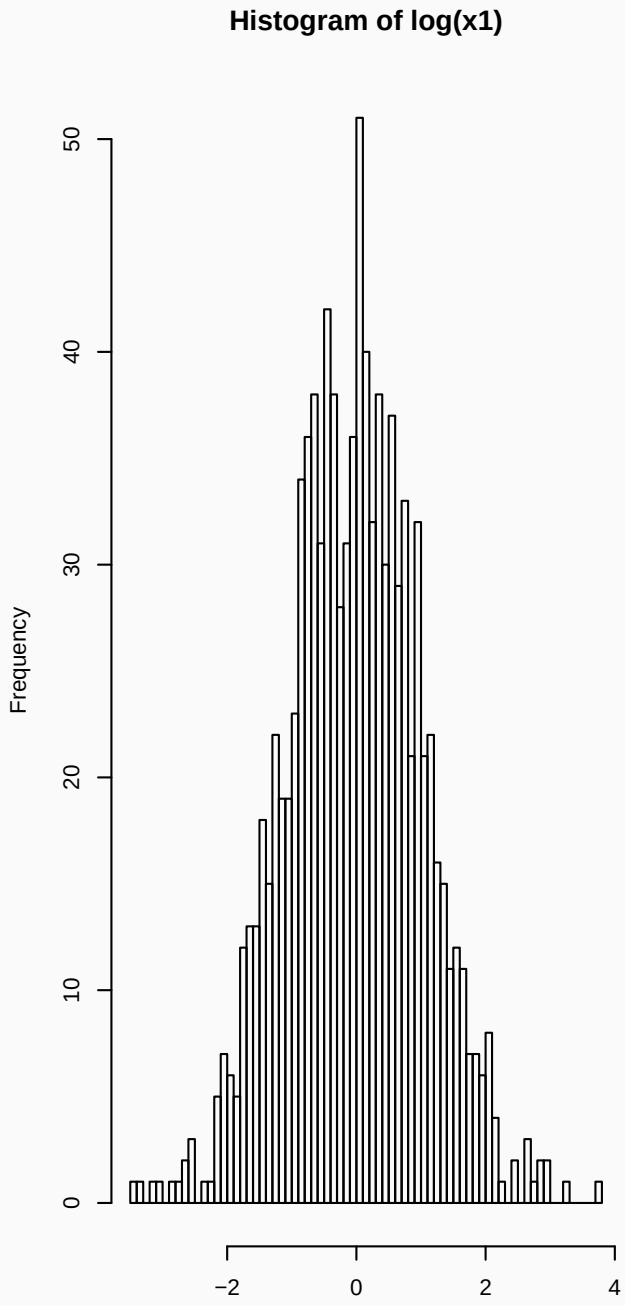
$$log_e(e) = 1$$

$$log_e(2e) = 2$$

etc.

What happens when you log a skewed variable?





One way to think about hypothesis tests:

$$\text{test} = \frac{\text{Signal}}{\text{Noise}}$$

Sigmal = the magnitude of the estimate. E.g. difference in mean

Noise = the standard deviation of the estimate

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

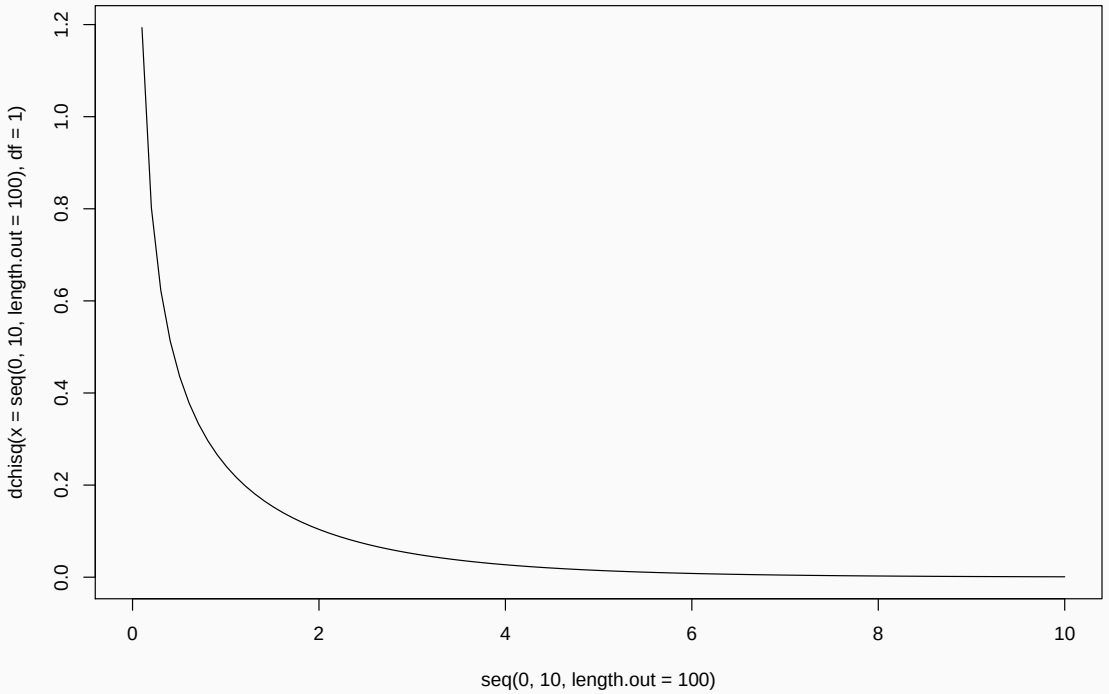
Calculating χ^2

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?




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

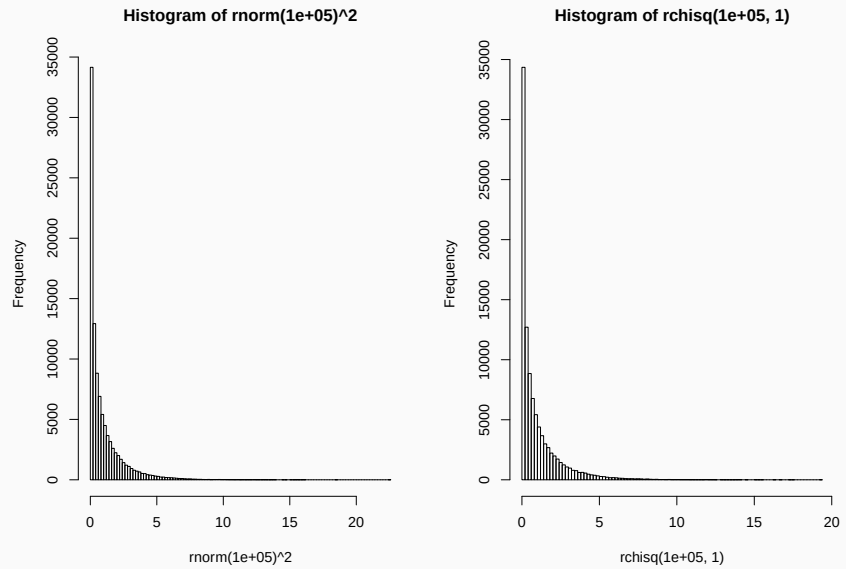
```
## [1] 0.3073568
```

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

Aside note on χ^2 distributions

- Chi square random variables are sums of squared independent standard normal random variables
- So the $\chi^2(1)$ is really a squared standard normal dist.
- Don't take my word for it:

```
par(mfrow=c(1,2))
hist(rnorm(100000)^2, breaks=100)
hist(rchisq(100000, 1), breaks=100)
```



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
```