Hypothesis testing

Vera Demberg

Saarland University

WS 2021/22

Table of Contents

1. Hypothesis testing
2. Type I and type II errors
3. Power
4. One- vs. two-tailed hypotheses
5. Confidence Intervals
6. Parametric assumptions

Table of Contents

1 Hypothesis testing

2 Type I and type II errors

3 Power

4 One- vs. two-tailed hypotheses

5 Confidence Intervals

6 Parametric assumptions

Rude Motorist example

Study by Doob and Gross (1968)

Study Goal: *to investigate the influence of perceived social status.*

When an old, beat-up (low-status) car failed to start when the traffic light turned green, 84% of the time the driver of the second car in line honked the horn. However, when the stopped car was an expensive, high-status car, the following driver only honked 50% of cases.

Two possible explanations:

- difference between 84% and 50% is attributable to sampling error, therefore, we cannot conclude that perceived social status influences horn-honking behaviour.
- difference between 84% and 50% is large and reliable. We conclude that people are less likely to honk at drivers of high-status cars.

Samples representing parking study

Derivation by sampling experiment:

Idea: we check whether a difference of 6.88 seconds can be expected to occur by chance. For this, we simulate the experiment 10 000 times.
If we find that a difference of 6.88 seconds is likely to occur by chance (= we find that this often happens in simulation), we conclude that it doesn't make a difference whether anybody is waiting.
If we find that a difference of 6.88 seconds is unlikely to occur by chance, we conclude that having somebody waiting does affect the time people take.

A little simulation in R...

Let's try to do a little simulation in R:

L3-examples.R

Samples representing parking study

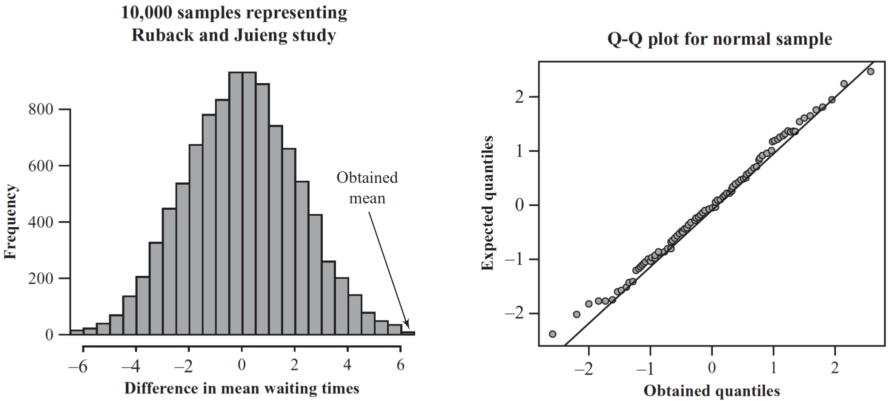


Figure 4.1 Distribution of difference between means, each based on 25 observations

Hypothesis testing

In **hypothesis testing**, our goal is to calculate how large the chances are that an observation came about *only due to sampling error*.

This is also called the **null hypothesis**: it usually holds that the effect we are interesting in showing does NOT hold (= any difference just due to sampling error).

(Remember that with samples that are smaller than the whole population, we can always expect to see some sampling error. That is, even if there is no effect of a manipulation whatsowever, we still expect means to be different.)

Why not prove the alternative hypothesis (instead of disproving the null hypothesis)?

What we actually want is to prove the alternative hypothesis.

Why not prove the alternative hypothesis (instead of disproving the null hypothesis)?

What we actually want is to prove the alternative hypothesis.

However, just by making observations, it's impossible to prove that something is true.
But we can prove that something is false.

Why not prove the alternative hypothesis (instead of disproving the null hypothesis)?

What we actually want is to prove the alternative hypothesis.

However, just by making observations, it's impossible to prove that something is true.
But we can prove that something is false.

Example: White hedgehog

Observing 3000 brown hedgehogs does not prove that white hedgehogs don't exist.
However, observing a single white hedgehog proves that not all hedgehogs are brown.

The limits of hypothesis testing

- the mathematical edifice that justifies hypothesis testing is built under the assumption that the null hypothesis is true
- so, all we can do within this statistical framework is
  - either decide that we do not have sufficient evidence to reject the null hypothesis (i.e., we therefore accept the null)
  - or decide that we have sufficient evidence to reject the null
- **however**, the latter does **NOT** mean that we have proved the alternative hypothesis!
- an alternative approach which avoids this: → Bayesian statistics

What if we can't reject the null hypothesis?

Different stances:

- proved null hypothesis to be true?
- accept the null hypothesis?
- retain the null hypothesis?
- withhold judgment and collect more data?

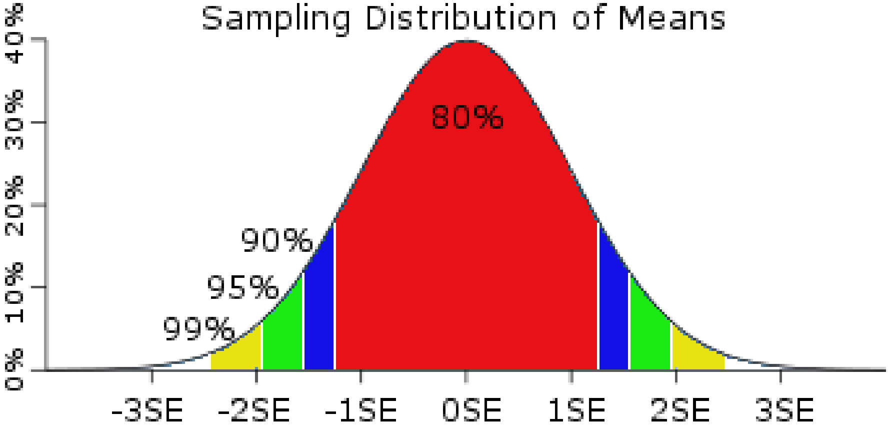
Decision making

Say we found that a sample mean (e.g. 84% honks) is very unlikely (p=0.002) to come from the same population as another sample mean (e.g. 50% honks).

Then we need to decide whether a probability of 0.2% chance that they were coming from the same distribution is low enough for us to conclude that they were coming from different distributions.

The value which we set for rejecting a null hypothesis (**the rejection level**) is essentially arbitrary. 5% is a commonly used value.

Rejection Level



Test statistics

- Examples include: the *t*-test, the chi-square-test (χ^2-test), the F-test
- they are all random variables for which we know the distribution (i.e., we know their density functions, so we know how likely different values are to occur)

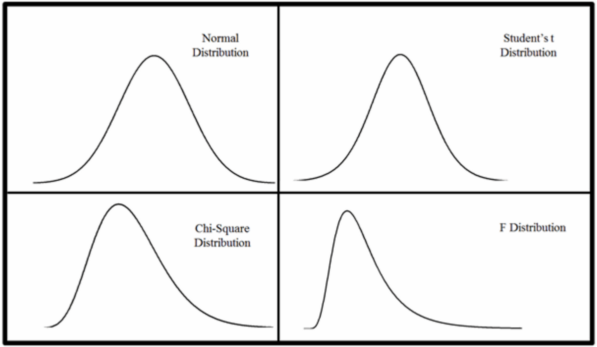
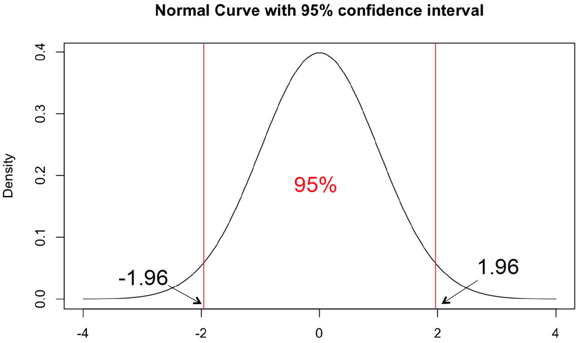


Table of Contents

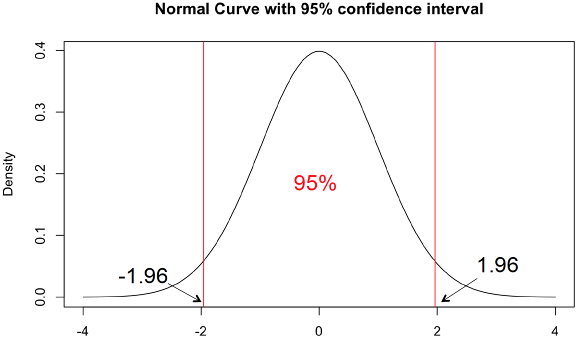
Type I error

- let us assume that our test statistics has a normal distribution
- and that we have standardized it, so that it has μ equal to 0 and σ equal to 1
- then, we know that 95% of the distribution lies between -1.96 and $+1.96$



Type I error

- alternatively, we can say that values of the standardised test statistics that are lower than -1,96 or higher than +1.96 are unlikely to have occurred by chance, as they represent only 5% of the distribution
- what does this mean?



Type I error... cntd

if the null hypothesis were true, and we were to draw another sample (repeat our experiment), and compute another value of the test statistics

- we would have a 95% chance that this value would still lie between -1.96 and +1.96, but only a 5% chance that it would lie outwith these values
- thus, values of the test statistics that lie outside this interval are "good news"
    - they are unlikey to have occured by pure chance
    - therefore, we conclude it's due to whatever factor we are interested in

Type I error...cntd 2

BUT:

- it is important to remember that those values **could** have occured by pure chance
- precisely, there is less than a 1 in 20 chance (i.e., a probability of less than .05), that any such value would have occured by pure chance
- a significance level of 0.05 means that in 1 out of 20 samples or experiments, we would conclude we found an interesting difference when in fact there isn't any

Type I error – Definition

- this is called a “false positive” or a Type I error
- and the probability associated with it is normally referred to as α
- we conventionally agree that a Type I error rate of $\alpha = .05$ is acceptable
- differences associated with values of the test statistics that fall in the 5% of the distribution are therefore deemed '**significant**'

But wouldn’t it be better to put this value lower?

Type II error

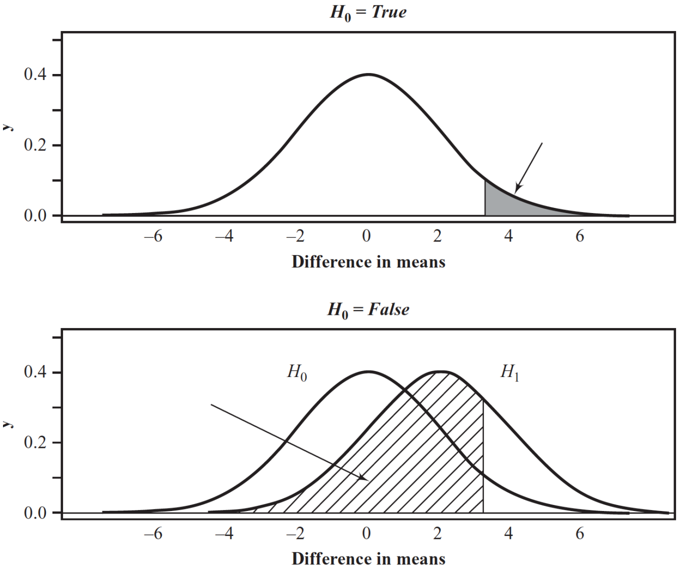
It can also happen that we fail to detect a true difference.

Type II error

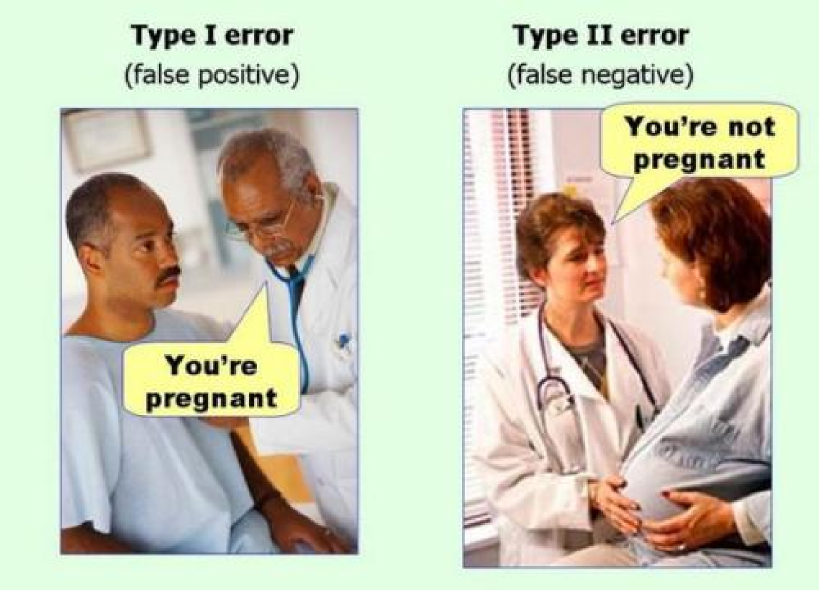
A false negative: we do not reject the null hypotheses when we should have rejected it

- the probability associated with a Type II error is indicated with β
- it tends to be higher the lower α gets, but the relation between the two is not a simple one

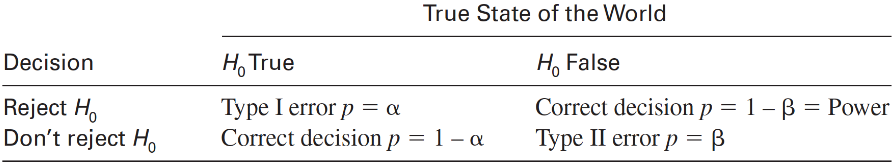
Type I and Type II error (signif level 0.01)



Type I and Type II error



Type I and Type II error



	True State of the World	
Decision	H_0 True	H_0 False
Reject H_0	Type I error $p = \alpha$	Correct decision $p = 1 - \beta =$ Power
Don’t reject H_0	Correct decision $p = 1 - \alpha$	Type II error $p = \beta$

Take notes for yourself

Let's go back to our IQ example:
Average IQ is 100, with standard deviation of 15 points.
Imagine, there is a novel brain training technique that you developed. A participant that you recruited used that technique and when you test him, you find he has an IQ of 145.

- Draw the distribution of IQs in the general population.
- What is H0 in this study?
- What is H1 in this study?
- What would be a type I error in this study?
- What would be a type II error in this study?
- What is the z score that corresponds to the observed IQ of 145? What does this number tell you? How can you get from this number to a *p* value?

Table of Contents

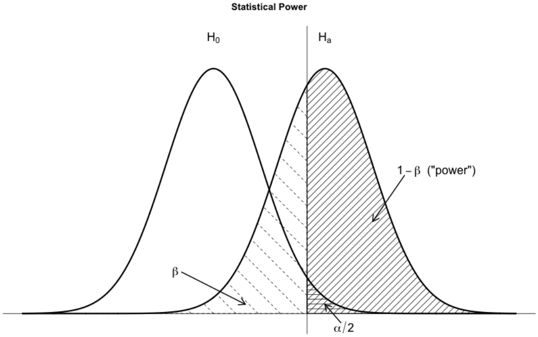
1. Hypothesis testing
2. Type I and type II errors
3. Power
4. One- vs. two-tailed hypotheses
5. Confidence Intervals
6. Parametric assumptions

Power

Power

The probability of detecting an effect if there is one, or simply 1 - β

- depends on sample size
- depends on the effect size
- depends on the significance level (α)



Significance - what does it mean?

It is important to remind yourself that:

- a **significant** result is not necessarily an **important** result
- a **significant** effect is not necessarily a **strong** effect
- a **significant** result could have occured by chance
- a **non-significant** result could have been a real effect
- science is not about finding significant results, but **finding better ways of describing our data**

Effect size

We just said that a significant result is not necessarily an important result or a strong effect.

We can quantify this more meaningfully by reporting **effect size**.

For example, we can specify how many more times somebody is likely to honk, or we how many milliseconds less overlap there is between speakers who know each other well.

More formally, we can quantify the effect size in terms of the standard deviation of the distribution, for example, we can say that the effect is “half the size of the standard deviation”

Table of Contents

1. Hypothesis testing
2. Type I and type II errors
3. Power
4. One- vs. two-tailed hypotheses
5. Confidence Intervals
6. Parametric assumptions

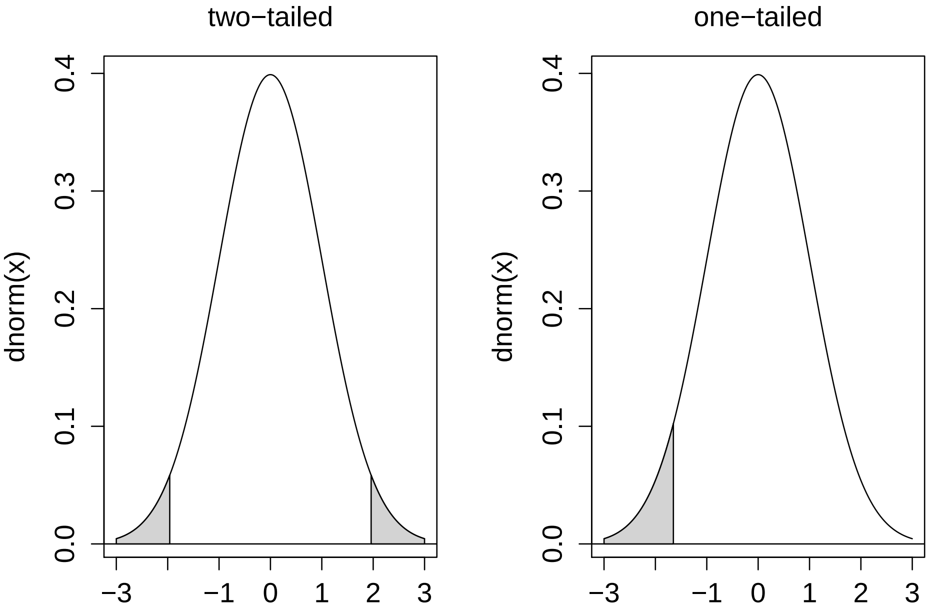
One- vs. two-tailed hypotheses

Spot the difference:

- people talk **more slowly** when they know somebody is talking at the same time as them H_1
- strangers overlap **at a different rate** than close friends H_1

One- vs. two-tailed hypotheses (and tests)

- people talk **more slowly** when they know somebody is talking at the same time as them H_1
  - one-tailed hypothesis: predicts a difference and the direction of the difference (more slowly, not faster)
- strangers overlap **at a different rate** than close friends H_1
  - two-tailed hypothesis: predicts a difference but not the direction of the difference (the rate could be higher or lower)


two−tailed
one−tailed
dnorm(x)
dnorm(x)
0.0
0.1
0.2
0.3
0.4
−3
−1
0
1
2
3

Note of caution

- as a general rule, you should phrase your hypotheses as two-tailed, and incorporate this assumption into your statistical tests
- the reason for this is two-fold:
    - first, it is rarely the case in computer science or linguistics that we can make specific predictions about the direction of an effect before running an experiment or checking a corpus; there are exceptions to this, of course (e.g., more frequent words are processed more quickly than less frequent words)
    - second, two-tailed tests are more conservative

One- and two-tailed hypotheses

- remember that with two-tailed hypotheses, we make no prediction with regard to the direction of the difference
- this means, that we do not predict the sign ($-$ or $+$) of the test statistics
- therefore, we distribute our 5% equally on both sides, so that we actually carve out 2.5% on each side
- however, if we have a one-tailed hypothesis, we “put all our eggs in one basket”

Two caveats against one-tailed tests

1. if we are wrong about the directionality of the effect, we might miss an otherwise significant difference
2. when we are not wrong about the directionality, we are relaxing our concerns about Type I error; this is an issue when you decide to run a one-tailed test after you have seen the data!

Table of Contents

1. Hypothesis testing
2. Type I and type II errors
3. Power
4. One- vs. two-tailed hypotheses
5. Confidence Intervals
6. Parametric assumptions

Confidence Interval

For an estimated parameter (the mean for our sample), the confidence interval tells us where the real population mean would lie with the level of certainty given by the confidence level.

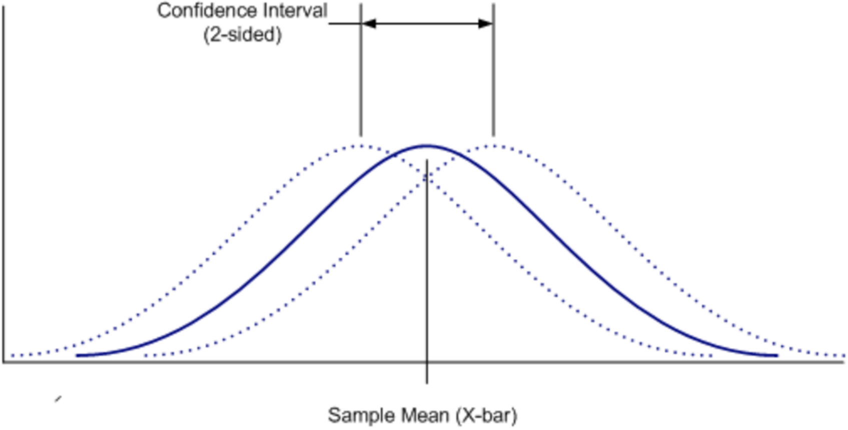


Illustration of confidence interval

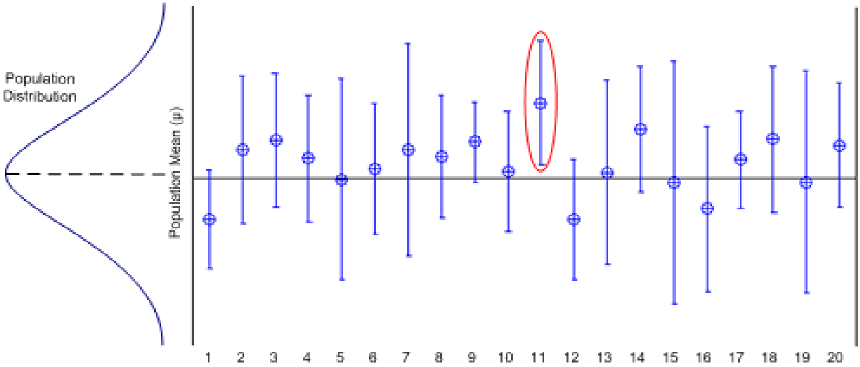


Table of Contents

1 Hypothesis testing

2 Type I and type II errors

3 Power

4 One- vs. two-tailed hypotheses

5 Confidence Intervals

6 Parametric assumptions

Parametric assumptions

Some of the very important test statistics (t-test and ANOVA) we will consider in this course can only be appropriately applied when ALL of the following conditions hold:

1. the D.V.* is a continuous random variable whose values have been **independently sampled** from the same population
2. the D.V. **follows a specific distribution** (normal distribution, binomial distribution, t distribution, poisson distribution...)
3. the D.V. has **similar variances across values of the I.V.**† (homogeneity of variance)

*D.V. = dependent variable or response variable

†I.V.= independent variable or predictor variable

Parametric assumptions

Some of the very important test statistics (t-test and ANOVA) we will consider in this course can only be appropriately applied when ALL of the following conditions hold:

1. the D.V.* is a continuous random variable whose values have been **independently sampled** from the same population
2. the D.V. **follows a specific distribution** (normal distribution, binomial distribution, t distribution, poisson distribution...)
3. the D.V. has **similar variances across values of the I.V.**† (homogeneity of variance)

Tests which have these assumptions are called *parametric*.

*D.V. = dependent variable or response variable

†I.V.= independent variable or predictor variable

Parametric assumptions

Some of the very important test statistics (t-test and ANOVA) we will consider in this course can only be appropriately applied when ALL of the following conditions hold:

1. the D.V.* is a continuous random variable whose values have been **independently sampled** from the same population
2. the D.V. **follows a specific distribution** (normal distribution, binomial distribution, t distribution, poisson distribution...)
3. the D.V. has **similar variances across values of the I.V.**† (homogeneity of variance)

Tests which have these assumptions are called *parametric*.

"non-parametric" is any test that does not rely on any assumption about the distribution of the D.V., and we'll look at a couple of those, too.

*D.V. = dependent variable or response variable

†I.V.= independent variable or predictor variable

What if my data are not normally distributed?

It is often still possible to use a parametric test, after conducting a transformation of the D.V. and / or after removing outliers.

What if my data are not normally distributed?

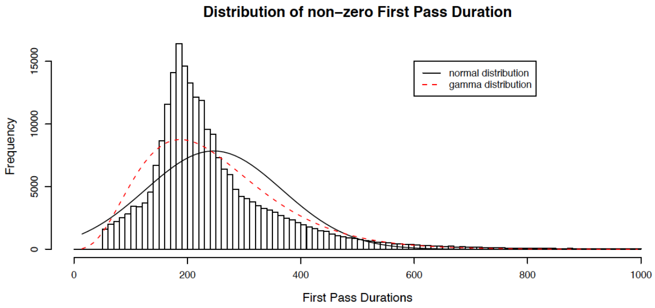
It is often still possible to use a parametric test, after conducting a transformation of the D.V. and / or after removing outliers.

Transformations:

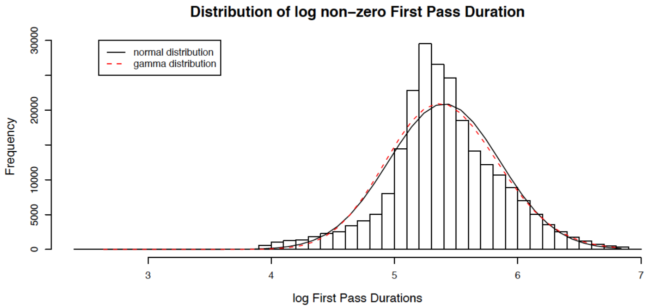
- transforming the D.V. (y) amounts to applying a suitable mathematical function *f*, so that:

$$f(y) \sim N(\mu, \sigma)$$

- most often the problem is that the distribution is not symmetric, but has a long tail to the right
  - people are rarely unusually fast; they are mostly unusually slow!
  - most words are very infrequent, with a few words being very frequent
- in this instance, commonly used functions include the logarithmic function, log(y), or the inverse function $\frac{1}{y}$



Distribution of non–zero First Pass Duration



Distribution of log non–zero First Pass Duration

Summary

What you should know from this lecture:

- logics behind hypothesis testing
- null hypothesis vs. alternative hypothesis
- limits of hypothesis testing
- type I error and type II error
- power of an experiment, and what it depends on
- when to choose one vs. two tails in hypothesis testing, what this means for querying z values in the table or in R
- confidence intervals – what they tell you about the population mean
- parametric assumptions: assume that DV is independently sampled, follows a specific distribution, homogeneity of variance

Reading: chapter 3 in Howell, chapter 11 in Navarro

Exercises

Imagine I told you that Germany played football against France yesterday and won with 20 to 13. You would probably decide that I'm confused or talking about something other than football. Actually, when you do that, you are rejecting a null hypothesis.

- What is the null hypothesis?
- Outline the hypothesis testing procedure that you just applied.

More exercises (this is about the logic, not the actual numbers)

The new batch of graduate students at a large state university has a mean GRE verbal score of 650 with a standard deviation of 50. (The scores are reasonably normally distributed.) One student, whose mother happens to be on the board of the trustees, was admitted with a GRE of 490.

- draw a rough sketch of the distribution of GRE scores.
- should the local newspaper editor write a scathing editorial about favoritism?
- Why might the GRE scores for admitted students be normally distributed (or not)?
- What would be a Type I error here?
- What would be a Type II error here?
- Why might I want to adopt a one-tailed test here, and which tail should I choose?
- What would happen if I chose the wrong tail?

More exercises (this is about the logic, not the actual numbers)

The new batch of graduate students at a large state university has a mean GRE verbal score of 650 with a standard deviation of 50. (The scores are reasonably normally distributed.) One student, whose mother happens to be on the board of the trustees, was admitted with a GRE of 490.

- draw a rough sketch of the distribution of GRE scores.
- should the local newspaper editor write a scathing editorial about favoritism? $z = (490 - 650)/50 = -3.2$; probability$= 0.0007$
- Why might the GRE scores for admitted students be normally distributed (or not)?
- What would be a Type I error here?
- What would be a Type II error here?
- Why might I want to adopt a one-tailed test here, and which tail should I choose?
- What would happen if I chose the wrong tail?

Additional exercise

You are working on a new algorithm for aligning two streams of data. For a standard alignment problem, the existing algorithms take 4.5 seconds, with a standard deviation of .5 seconds. Your own algorithm takes 3.8 seconds.

- draw a rough sketch of the distribution of runtimes.
- indicate the rejection interval for a two-tailed test.
- what would the rejection interval for a one-tailed test look like, and which tail should you choose?
- is your algorithm significantly faster than the others on this problem?
- What would be a Type I error here?
- What would be a Type II error here?