STAT0032: Introduction to Statistical Data Science

Dr. Francois-Xavier Briol

Department of Statistical Science
University College London

Week 2: Hypothesis Testing

Challenging Assumptions and Conclusions

- We may study data which appear to have interesting properties
  - **Are such properties real**?
- Data does not speak for itself.
  - We need to make assumptions
  - **Are those assumptions true?**

A First Encounter with Hypothesis Testing

A simple example for illustration

- Suppose we were to offer a training course on Statistical Data Science and wish to determine if there is an imbalance in the sex of individuals interested in this topic.
- With a large sample and large discrepancy in proportions this might be easy to conclude:
  - In such cases there might be no need for formal statistical inference.
- However, what if we observe 15 out of 40 participants are female?
  - How strong is this as evidence against balance?
  - We will assume for simplicity that the proportion of women is not greater than 0.5.

A hypothesis testing approach

- What is the hypothesis we would like to test?
  - Is the probability of the event of any given student being female 0.5?
- The technical term for this is the **null hypothesis.**
- The general approach first assumes that the null hypothesis is true:
  - Under this assumption, what is the probability of observing the available data?
  - Similar to legal trials: we assume will the null hypothesis holds unless proved with high certainty that it does not.

Test statistic

- A test statistic is a summary of the data
  - This concept has been introduced previously.
- A test statistic is a summary which can falsify the null hypothesis, if it is indeed false.
- There are different test statistics which could be chosen
  - Careful choice could make the calculations involved easier.
- Our Example: The number of female students in the class provides a summary for the proportion of female students interested in Statistical Data Science.

Complementary assumptions

- On top of the null hypothesis, we often need to make further assumptions to characterise the test statistic.
- For our considered example, we will assume that each student decides to enrol on the course independently.
  - Thus, the sex of each student is independent of each other.
- We might encode the sex of students numerically and assume these values are realisations from a random variable:
  - 0 for males, 1 for females
  - We then have independent Bernoulli random variables (“coin flips”)

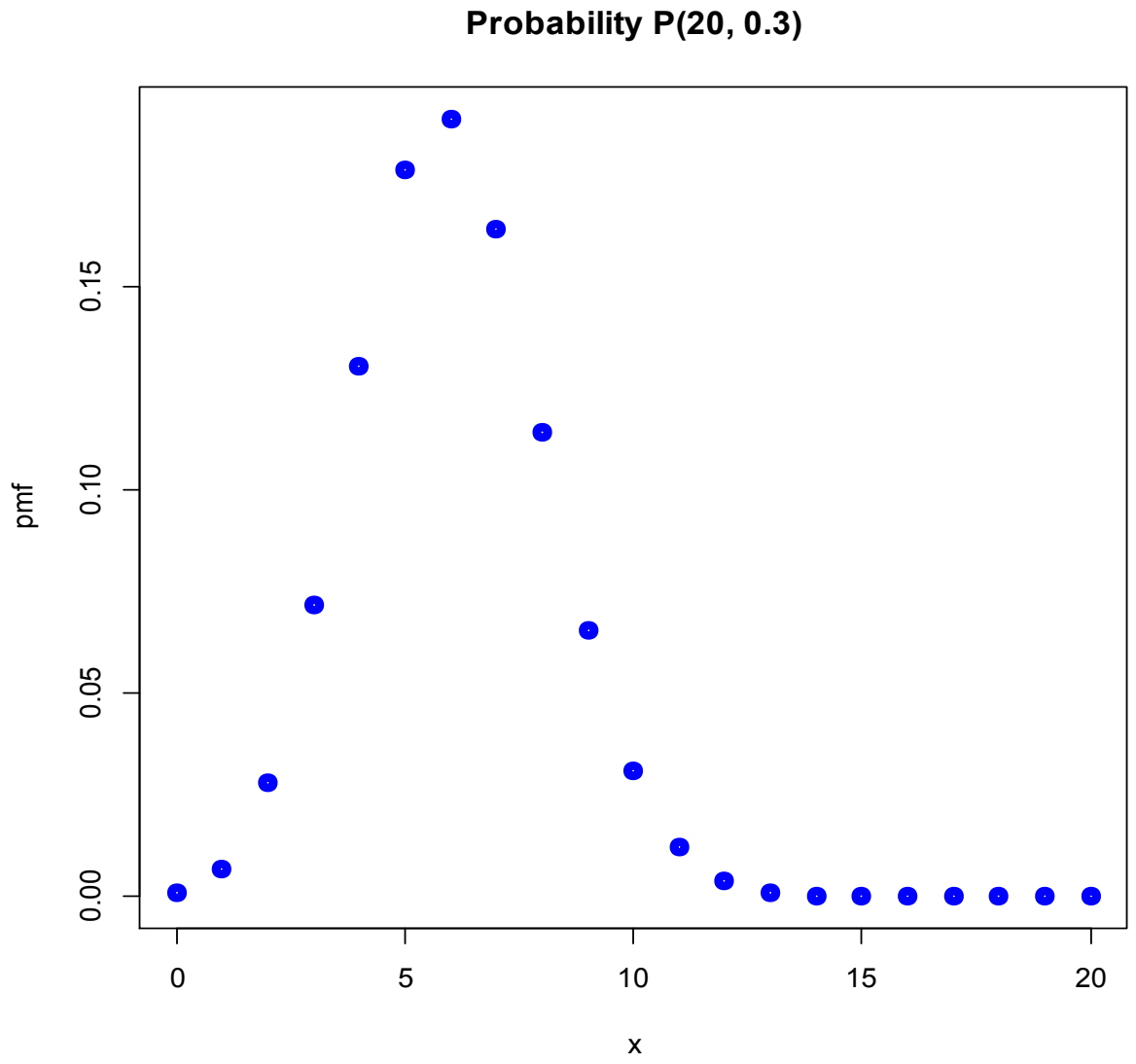
The Binomial distribution

- If we have *n* independent Bernoulli trials, each with probability θ, we have a binomial distribution.

$$X \sim Bin(n, \theta)$$

$$P(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

- PMF refers to the probability mass function.
- R code example.



Returning to our test statistic

- In our example we observe Bernoulli variables (0's and 1's) in an **independent, identically distributed (i.i.d)** way

$$Y_1, \ldots, Y_{40} \sim Bernoulli(0.5)$$

- That is, we can show that the sum of these variables is binomially distributed

$$X \equiv \sum_{i=1}^{40} Y_i \sim Bin(40, 0.5)$$

- More generally, the sum is binomially distributed with parameters *n* and θ

Relevancy

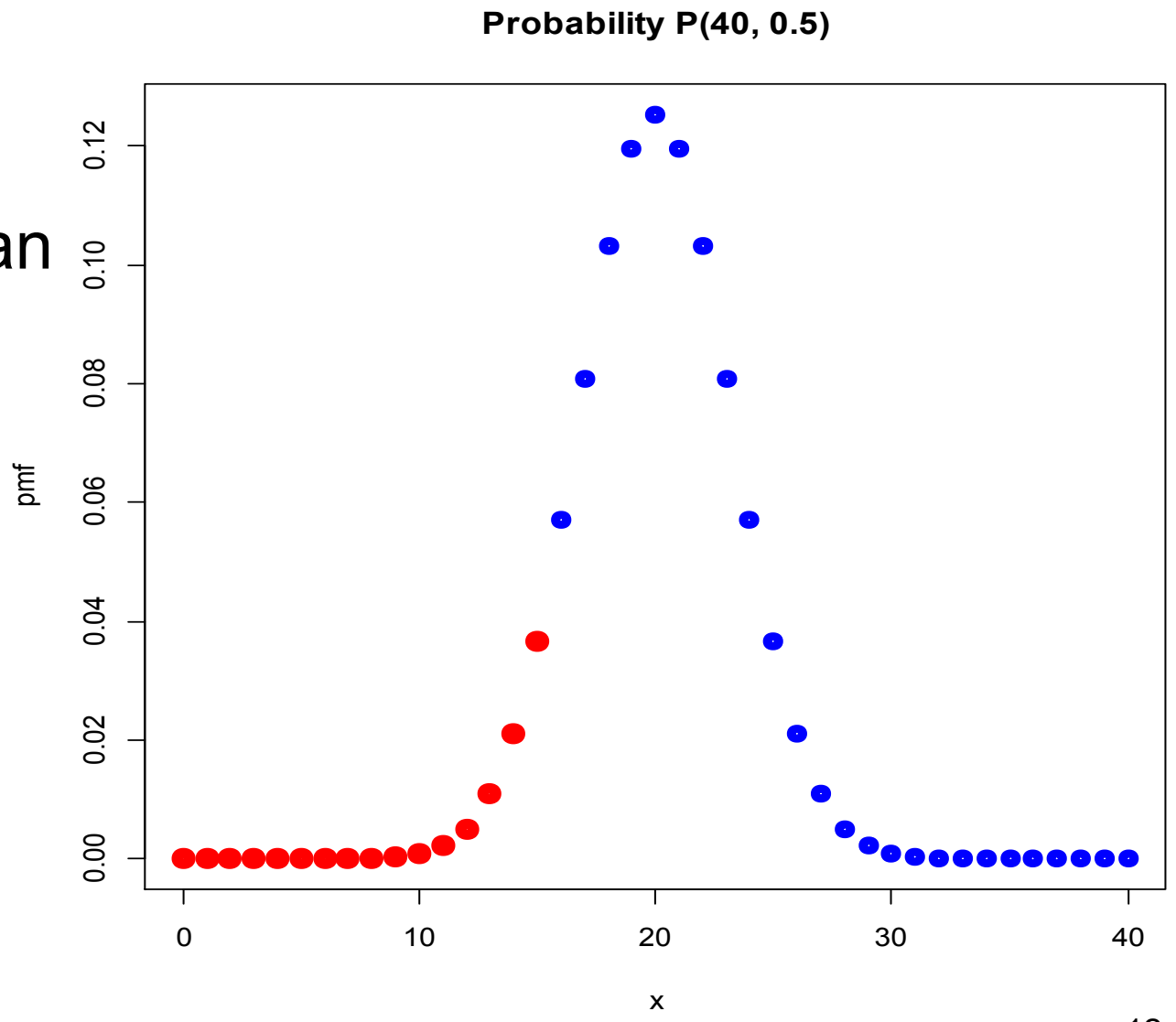
- We can characterise whether the value of x observed in our example (15), is likely under the null hypothesis H_0: $\theta = 0.5$
- We will in fact characterise how probable values of X of size 15 or smaller are
  - We assumed that θ is not greater than 0.5
  - Values of X less then or equal to 15 are therefore all of those which are as or more extreme than our observation

The p-value

- The probability of obtaining results as or more extreme than that observed, assuming H_0 is true, is the **p-value**

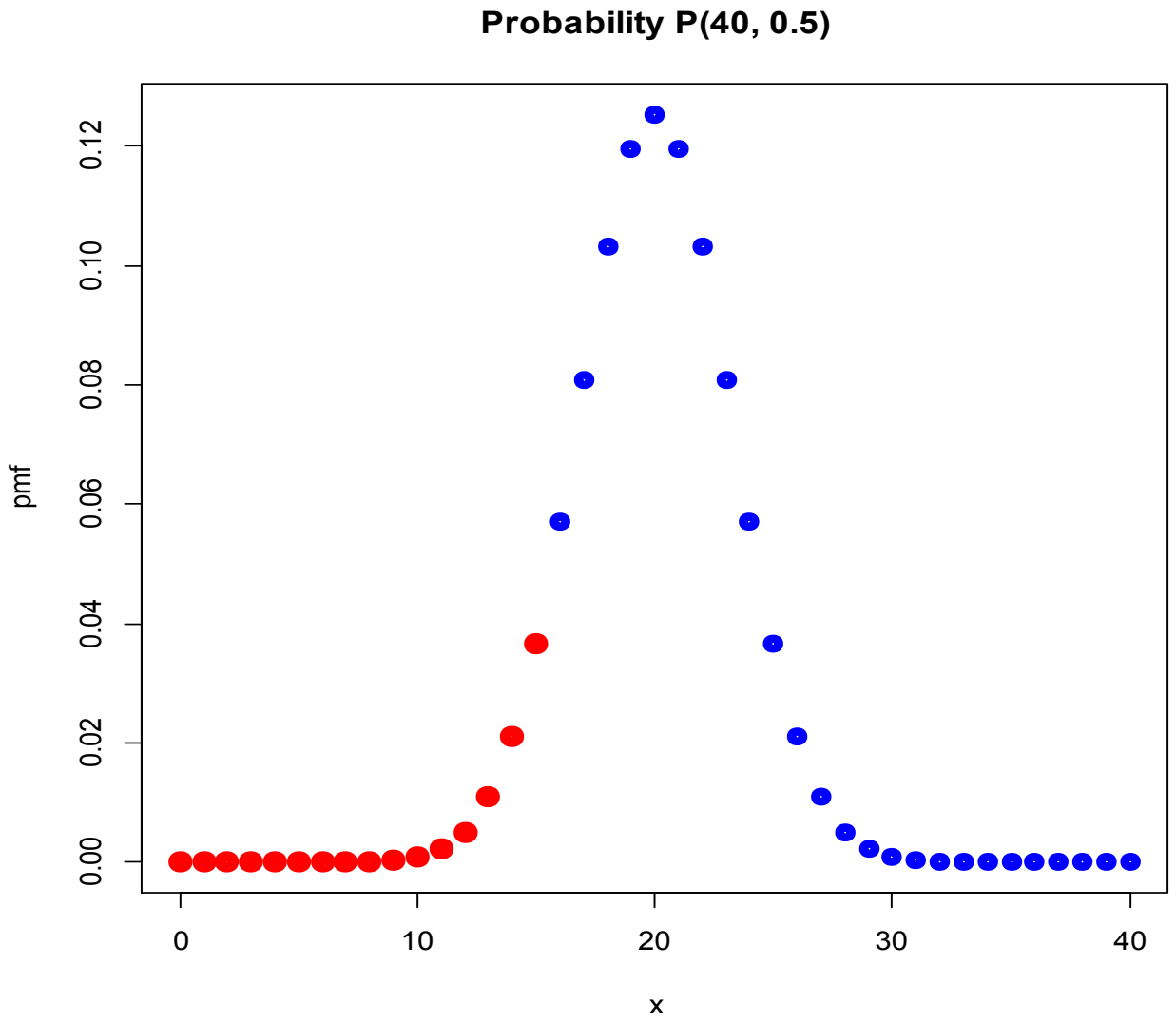
$$p \equiv P(X \leq 15; H_0)$$

$$p = \sum_{x=0}^{15} \binom{40}{x} 0.5^x (1 - 0.5)^{(40-x)}$$



The p-value

- Performing the required sum (by hand or by calculator or in R) gives us a p-value of approximately 0.07.
  - What is our conclusion in light of this?
- Decision thresholds are typically used to determine judge p-values.



Interpreting the p-value

- The p-value is the probability of observing a test statistic, *X*, as or more extreme than the value *x* seen in the data, under the assumption that the null hypothesis, H_0, is true
- The p-value is most certainly **not the probability of H_0 being true**
- We may refer back to some fundamentals of probability to confirm this difference

P-value is not the probability H_0 is true

- The rules of conditional probability state that

$$P(A, B) = P(A \mid B)P(B)$$

- As a result, we may present the probability of H_0 being true given test statistic *T* as follows

$$P(H_0 \mid T = t) = \frac{P(T = t \mid H_0)P(H_0)}{P(T = t)}$$

- This expression requires us to define the probability of H_0 being true, which is not always easy
  - A much deeper discussion of this approach is provided in the course STAT0031: Applied Bayesian Methods.

A logical analogy

- In logic implications may be reversed to provide what is known as the contrapositive

$$A \Rightarrow B$$

"If A is a statistician, then A is a data scientist"

$$\neg B \Rightarrow \neg A$$

"If A is not a data scientist, then A is not a statistician"

- The unwritten logic of hypothesis testing is that:
  - H_0 should imply with high probability the data values which we observe.
  - If instead we observe sufficiently extreme values under H_0 we may consider H_0 to have been disproved by an informal contrapositive argument.
  - Everything is of course complicated by the fact we are working with statements "with high probability".

Rejecting H_0

- One aspect of the contrapositive analogy is useful

$$\neg B \Rightarrow \neg A$$

- If we observe unusual/extreme data there is an indication that something within our assumptions is awry:
  - This could be the assumption of the parameter θ.
  - It might also relate to other implicit assumptions.

Critical Regions as Alternatives to p-values

- Rather than determining a p-value, we may determine a critical region for the test statistic
  - The set of all test statistic values which would cause us to reject H_0
  - $P(X \leq 15) = 0.077$
  - $P(X \leq 14) = 0.040$
  - $\Rightarrow CR = \{0,1,2, \ldots, 14\}$

- We may therefore simply compare our observed value to the critical region to judge whether to reject H_0

A memorable warning example

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

A Hypothesis Testing Procedure

Hypothesis Testing Procedure

- 1) Specify a null and alternative hypothesis
- 2) Specify the level of the test
- 3) Specify a suitable test statistic
- 4) Determine the distribution of the test statistic under H_0
- 5) Determine what it means to be “more extreme” by considering H_0 and H_1
- 6) Determine the corresponding p-value or critical region
- 7) Reject H_0 if
  - the p-value is less than the level of the test
  - or if the test statistic is inside the critical region

Testing Procedure for our Example

1) Specify a null and alternative hypothesis

$$H_0 : \theta = 0.5 \quad \text{The proportion of males and females is identical}$$
$$H_1 : \theta < 0.5 \quad \text{There is a smaller proportion of females than males}$$

2) Specify the level of the test $\text{Level} = 0.05$

3) Specify a suitable test statistic $X = \text{The number of females}$

4) Determine the distribution of the test statistic under H_0

$$X \sim \text{Binomial}(40, 0.5)$$

Testing procedure

- 5) Determine what it means to be “more extreme” by considering H_0 and H_1
  - H1: $\theta < 0.5$, so smaller values of X are more extreme

- 6) Determine the corresponding p-value $$p = P(X \leq 15) = 0.077$$

- 7) Reject H_0 if the p-value is less than the level of the test
  - $p > 0.05$, so we fail to reject H_0 in this instance
  - Conclude that the we do not have enough evidence to show that the proportion of females and males differs.

Selecting a Test Statistic and Level

Designing a Testing Procedure

- Given a problem we want to answer, there are many tests we could come up with. How do pick a good one?
  - We could have considered a different hypothesis…
  - We could have picked a different test statistic…
  - We could have fixed another confidence level…
  - We could have tried to collect more data...

Statistical Power

- Statistical hypothesis testing involves a trade-off between two drawbacks
  - False positives: rejecting H_0 when it is true
  - False negatives: not rejecting H_0 when it is false
  - Note that not rejecting H_0 is not the same as accepting H_0
- The power of a hypothesis test is the probability of avoiding a false negative
  - In other words, the probability of correctly rejecting the null hypothesis when it is false.
  - Choosing a test with maximal power may be one criterion to answer a question.

Setting a significance level

- We may define a rule under which we reject H_0
  - We reject H_0 if the probability of obtaining an outcome as or more extreme than that observed is < or = to 0.05 under the assumption that H_0 is true.
- This provides two things to consider
  - The p-value itself is a random variable giving the smallest level at which the test would reject the null hypothesis (see the next slide).
  - How does the probability of the p-value being less than 0.05 vary depending upon the manner in which H_0 is false?

P-values as random variables

Data is a realisation from some random variable. → Test Statistic is a realisation from some random variable → P-value is a realisation from some random variable

- We may consider the p-value to be a black-box function of the data

$$p_v(x) = \sum_{i=0}^{x} \binom{40}{i} 0.5^i (1-0.5)^{40-i} = F(x)$$

This is the CDF of a Bin(40,0.5)

- This black function may be calculated for a fixed data set providing the summary statistic *x* (15, in our example)

- The expression $p_v(X)$ with an upper case X indicates that the p-value is random since the data generating process is also random

P-value distribution

- If we assume that X is continuously distributed for simplicity then we may determine the CDF of the p-value

$$P(F(X) \leq z) = P(F^{-1}(F(X)) \leq F^{-1}(z)) = P(X \leq F^{-1}(z)) = z$$

p-value as function of data (itself a r.v.)

- We have already seen a distribution with CDF, P($Z \leq z$) = z for z in [0, 1]: Uniform[0,1]

- That is, under H_0 p-values are uniformly distributed on [0, 1]
  - Does this make intuitive sense?
  - What are the implications of this result?

Error control

- If H_0 is true and we reject H_0 only when the test statistic is below the 0.05 quantile, then the probability of erroneously rejecting H_0 when it is true is 0.05:

$$P(X \leq F^{-1}(0.05); H_0) = 0.05$$

- We say that below the 0.05 quantile is the critical region of this test and that the Type I error rate is 0.05 (I'll get back to type I errors…).

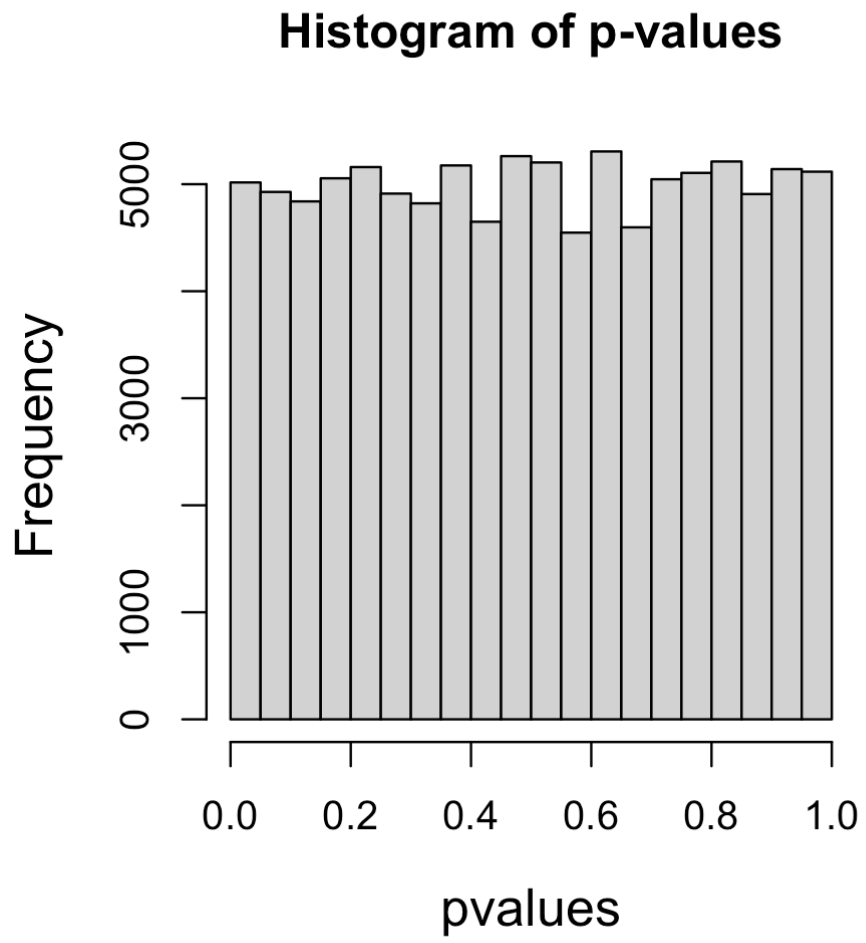
Frequentist interpretation and practical motivation

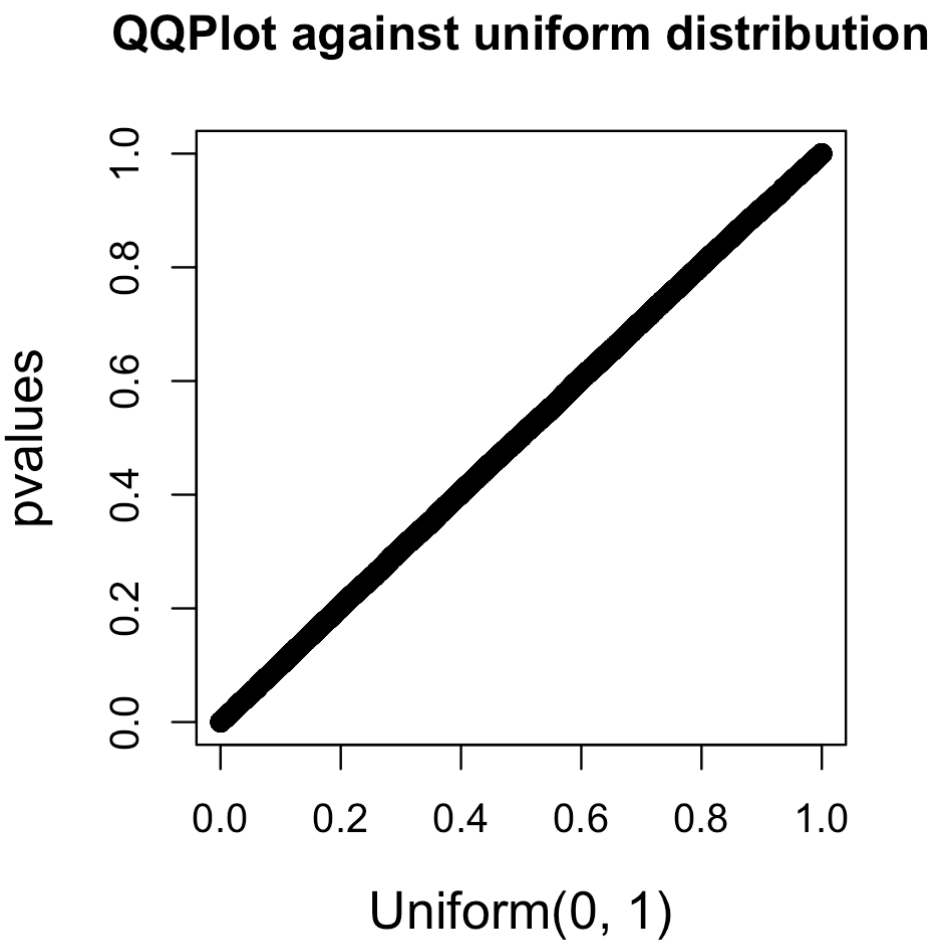
- Statisticians are not expected to collect data of the same phenomenon over and over again
  - Error calibration is about using the procedure over a long range of problems.
- The provided arguments are an idealisation
  - There will often be approximations (eg. the distribution of X is often not known exactly) and mistakes.
  - However, the aim is to be “less wrong”, if we do the appropriate thing.

Distribution of the p-value under H_0

- Earlier we showed the distribution of the p-value, given that the H_0 is true is simply uniform on the interval [0,1].
- This can be confirmed empirically using R
  - We sample a large number of random binomial variables (I will use 100000 from a Binomial(40000,0.5)).
  - For each we determine the p-value, the probability of observing a result as or more extreme (in this case less) than that observed.
  - We plot a histogram of the resulting p-values as an estimate of the p-value density.
  - We also compare the results to the Uniform[0,1] distribution using a Q-Q plot.

Distribution of the p-value under H_0 (in R)





Level

- In the example presented earlier the threshold probability of 0.05 was the **level** of the test
  - In general, the choice of level is problem-dependent.
  - 0.05 is a common example in scientific literature, but its motivation is not always justified.
- The choice of a particular level may be guided by the need to trade off **Type I** and **Type II** errors.

Type I and Type II errors

- Type I errors occur when we reject the null hypothesis, H_0, when it is true (I called this “false positives” earlier).
- On the other hand, Type II errors occur when we fail to reject H_0 when it is false (I called this “false negatives” earlier).
- The **probability of avoiding a Type II error is the power** of the test
  - The probability that we reject H_0 given that it is false.
  - Unlike the level of the test, which we specify, the power generally depends upon what the true hypothesis is.

Power of a Test

- The power of a test varies with sample size
  - The distribution of the test statistic changes with sample size
- The power of a test also varies with the level of the test
  - Changes in the level of the test change the rejection region
- When we describe a trade-off, we mean level vs. power at a fixed sample size
  - Increasing sample size will increase power without changing the level

Power investigation: Alternative Values

- We may investigate how the power of the test varies for a range of alternative values of θ, the true state of nature (i.e. the actual proportion of women in the class)

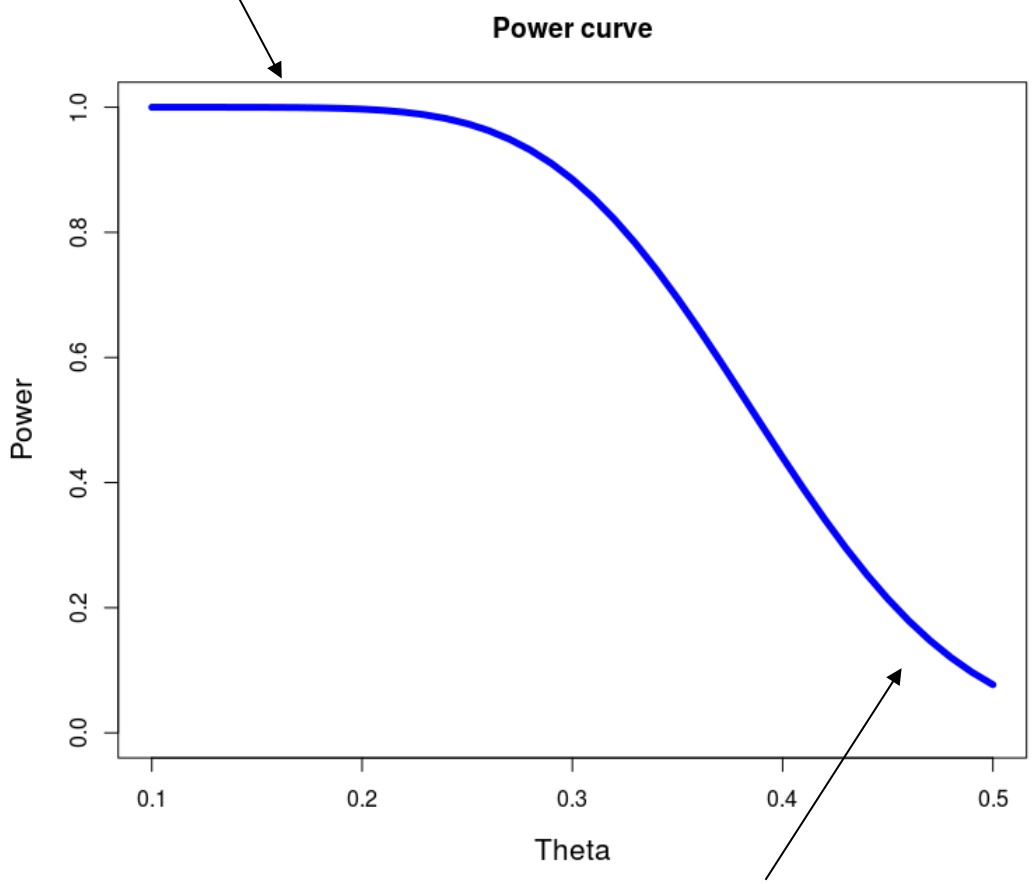
$$\text{Power} = P(X \in CR|\theta)$$

$$P(X \in CR|\theta = 0.2) = 0.992$$
$$P(X \in CR|\theta = 0.3) = 0.807$$
$$P(X \in CR|\theta = 0.45) = 0.133$$

Power investigation

- Plotting the power for each value of θ between 0 and 0.5 allows us to see how it changes with the true value of the parameter
  - As θ deviates further from 0.5, our test is more effective at detecting this difference

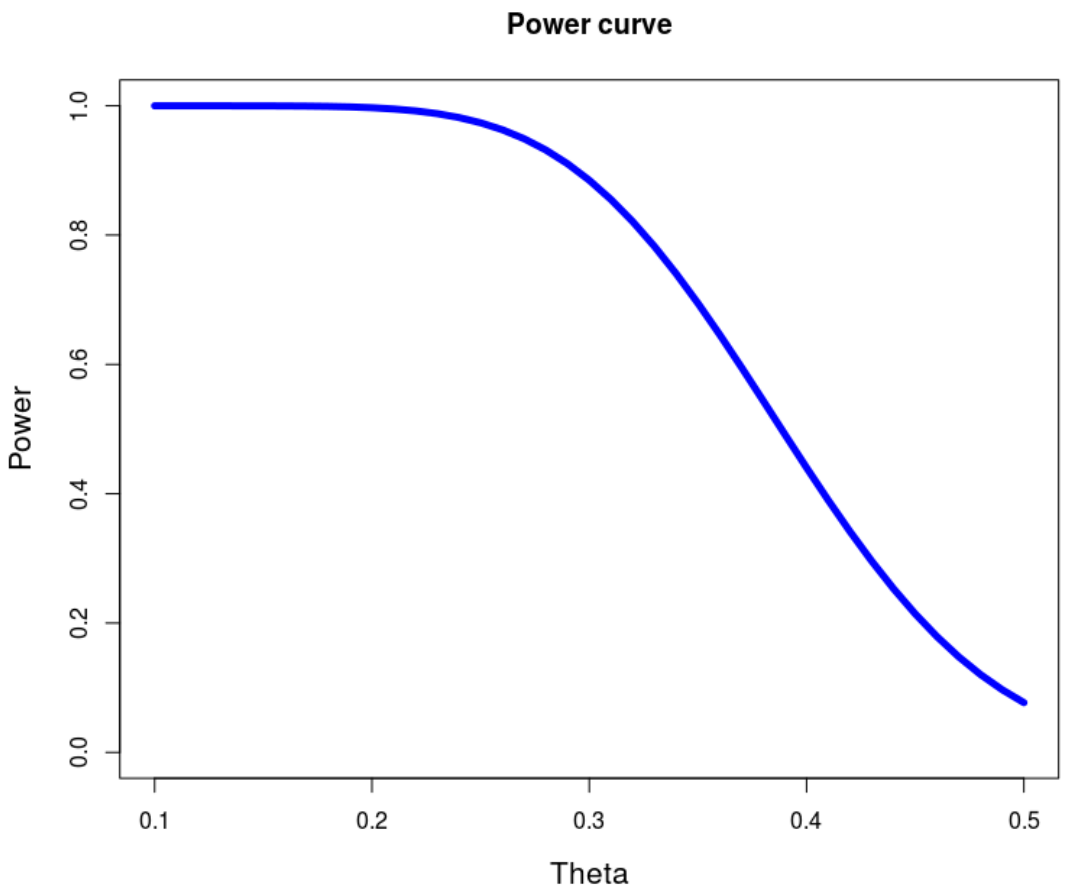
We usually reject the null when it is false

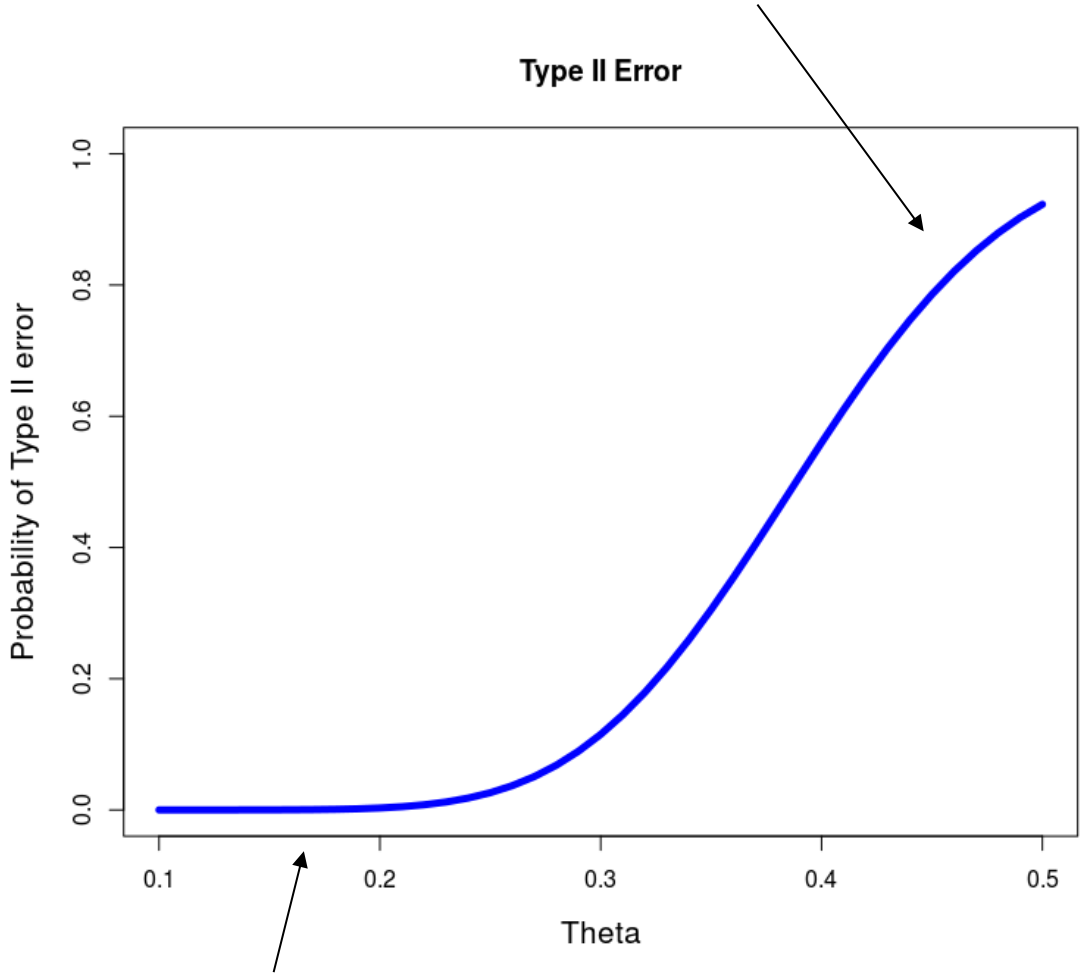


We often fail to reject the null when it is false

Power investigation

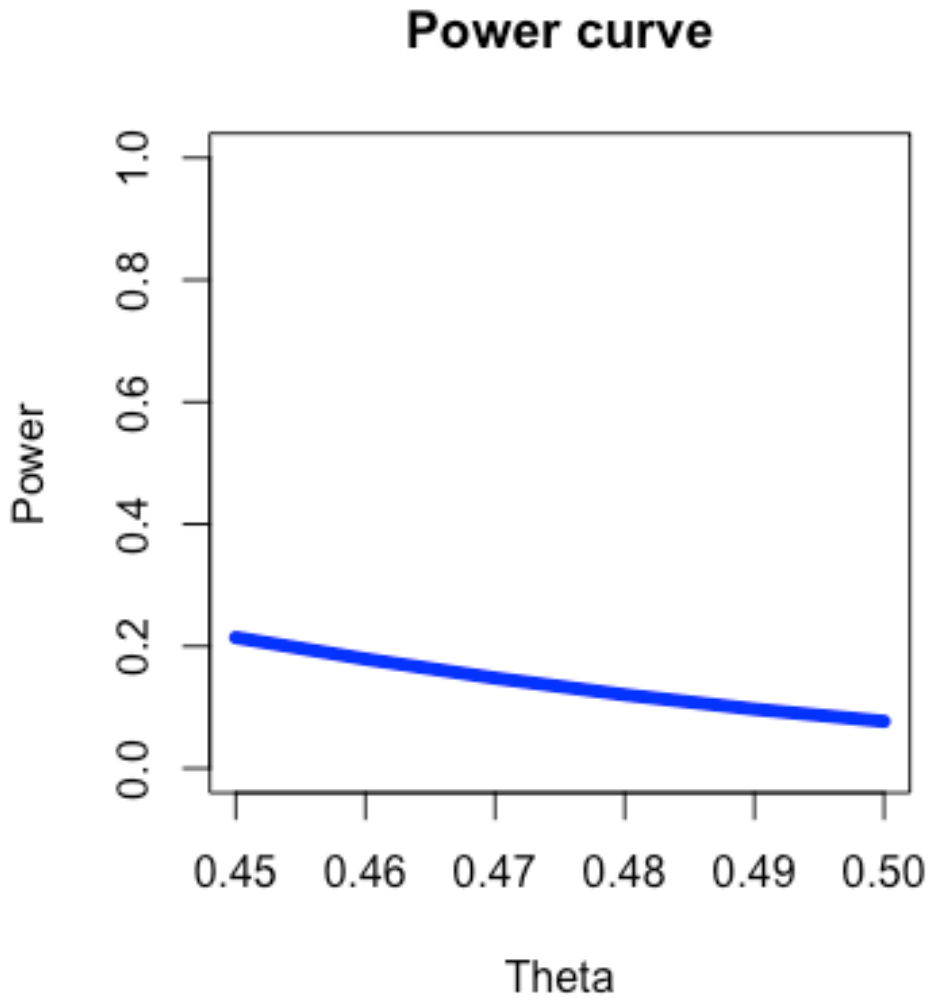
We often fail to reject the null when it is false





We usually reject the null when it is false

Power investigation (Zooming in on High Theta Values)



More powerful tests

- The power equals 1-rate of type II error.

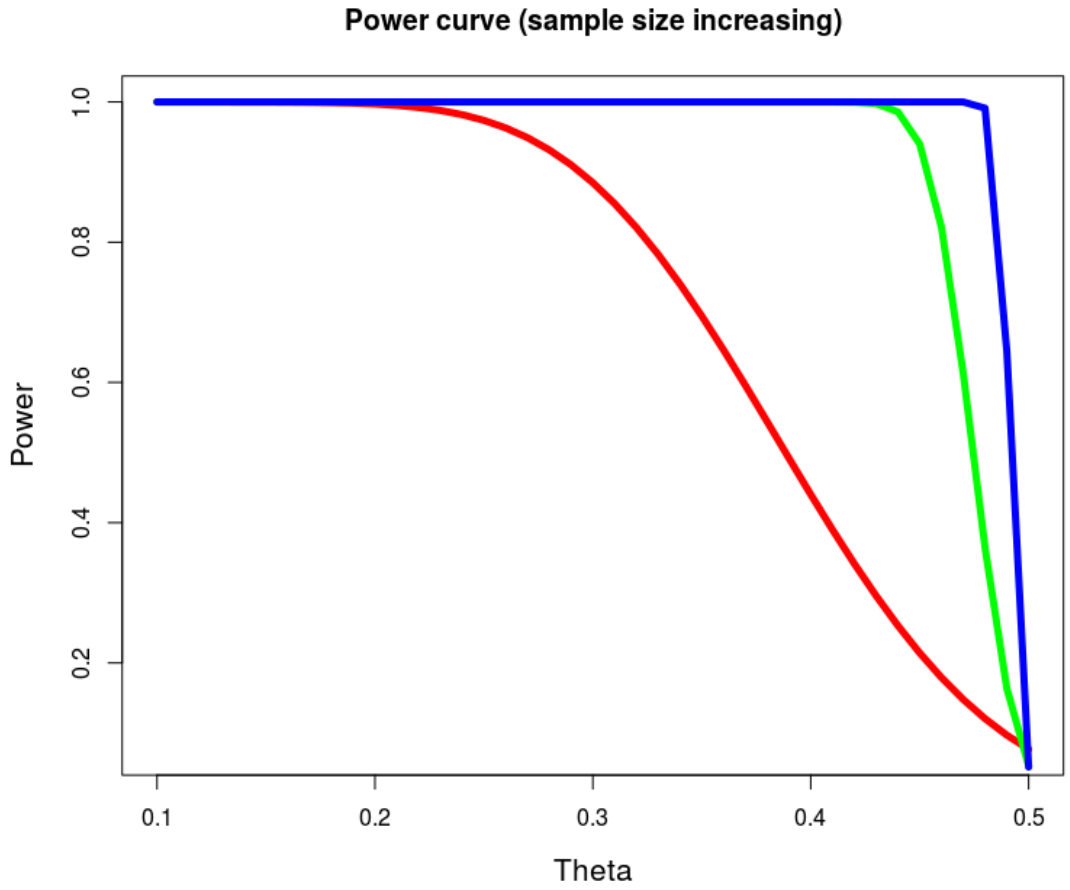
- Given that we cannot change the underlying value of θ, how might we increase the power of our test?
  - Collect more data (increase the sample size).
  - Allow for a higher Type I error (increase the level of the test).
  - Use a better test statistic.
  - Make stronger assumptions.

- These possibilities might be used before collecting data, during the study design phase

Get more data

- In the same way that we could plot the power of our test for varying underlying θ, we may produce similar plots for varying sample size

- N = 40 (red)
- N = 1 000 (green)
- N = 10 000 (blue)



Allow for a higher Type I error

- If you increase the level of the test, you increase the size of the critical region and make it easier to distinguish H_0 from the alternatives

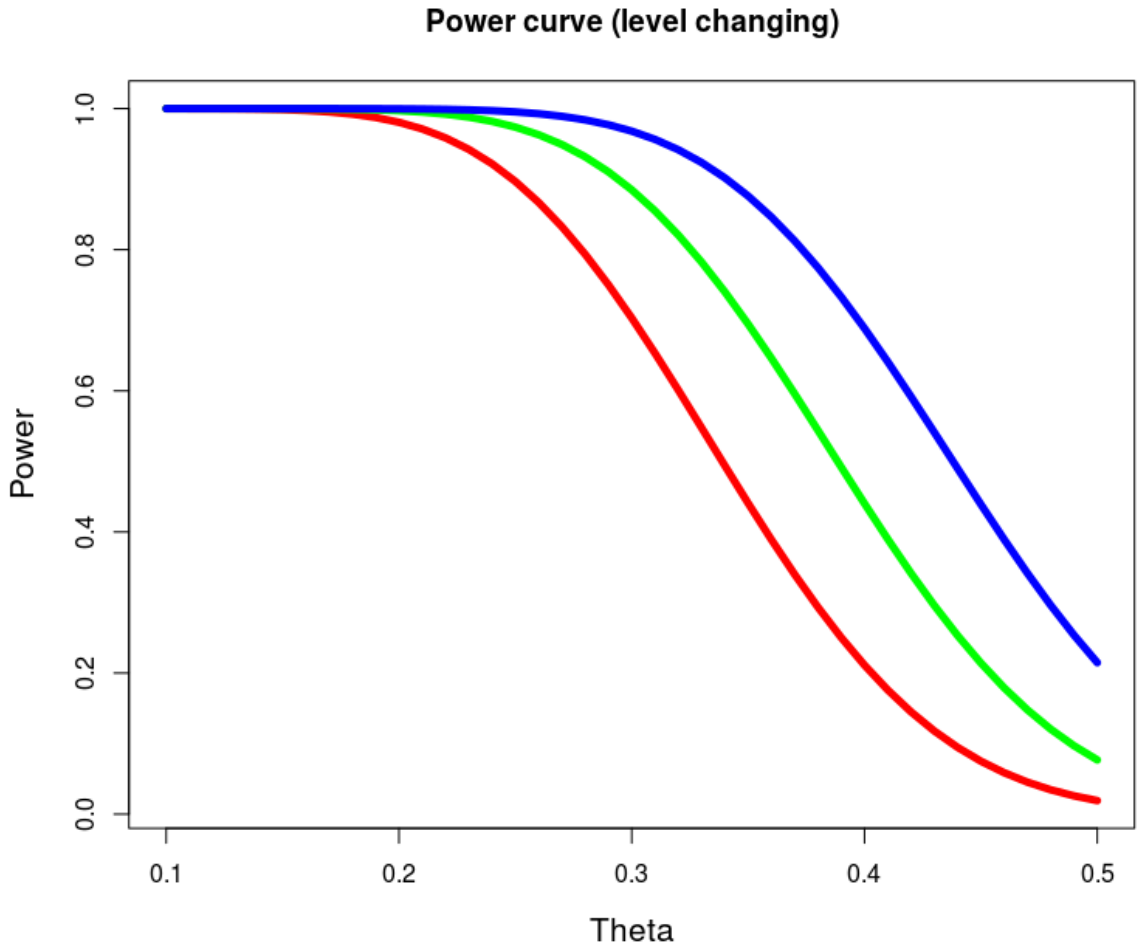
- If you decrease the level of the test, you reduce the size of the critical region and make it more difficult to distinguish H_0 from the alternatives

- Level 0.01, critical region is [0,12]
- Level 0.05, critical region is [0,14]
- Level 0.20, critical region is [0,16]

Allow for a higher Type I error

- We can plot the power functions again.

- Level 0.01 (red)
- Level 0.05 (green)
- Level 0.20 (blue)



Uniformly Most Powerful Test

- It is straightforward to consider valid, but less effective test statistics
  - Instead of using the number of females out of the 40 total students, we could consider the number of females in the first 20 students to enter the room.
  - The test carried out using this statistic has worse power than the test carried out using *X*.
- Finding a superior test statistic may not be easy
  - There are such things as *uniformly most powerful tests*, but we won't discuss these in any detail.
  - The example tests to come are known to have good power.

The Important Part of the Testing Procedure

- 1) Specify a null and alternative hypothesis (next section)

$$H_0 : \theta = 0.5 \quad \text{The proportion of males and females is identical}$$

$$H_1 : \theta < 0.5 \quad \text{There is a smaller proportion of females than males}$$

- 2) Specify the level of the test.
  - Bearing in mind the need to balance probabilities of Type I and Type II errors
  - Reducing the level reduces the probability of a Type I error
  - Increasing the level reduces the probability of a Type II error

$$\text{Level} = 0.05$$

- 3) Specify a suitable test statistic.
  - Bearing in mind the impact on the power of the test.

$$X = \text{The number of females}$$

Selecting a Hypothesis

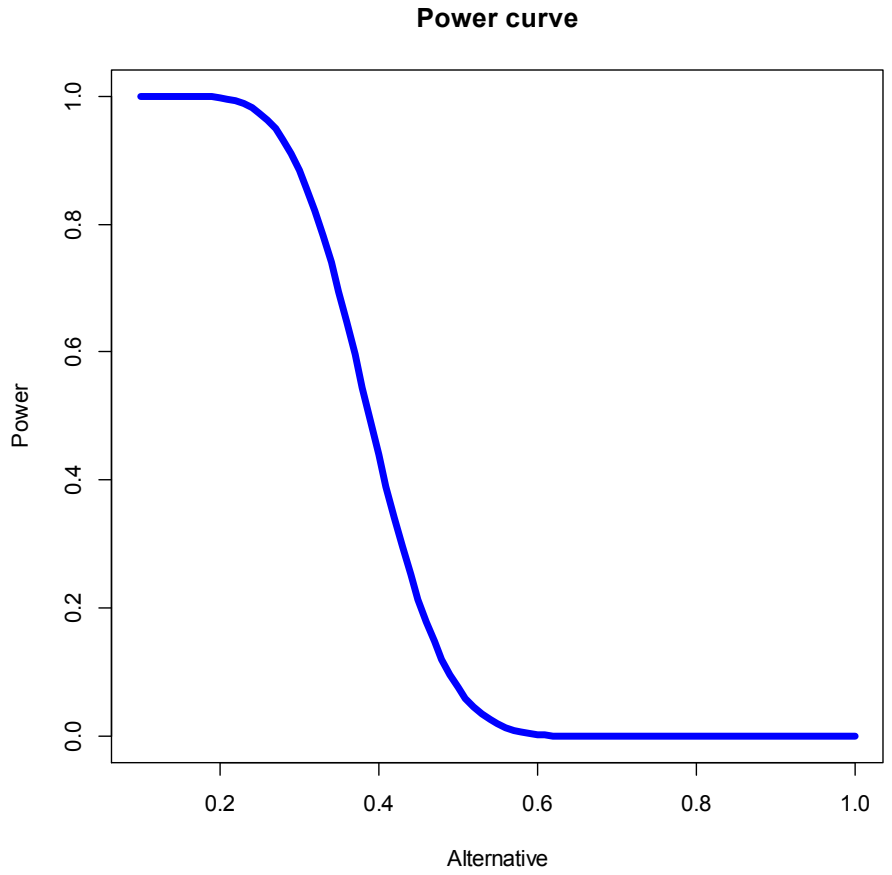
Making stronger assumptions

- If we make stronger assumptions with the consequence of reducing our set of alternatives we may increase the power of our test.
- Consider a different alternative hypothesis:

$$
\begin{aligned}
H_0 : &\quad \theta = 0.5 \\
H_1 : &\quad \theta \neq 0.5
\end{aligned}
$$

Making stronger assumptions

- If we do not update our definition of more extreme (equivalent to keeping the identical critical region) we have terrible power for some alternative values of θ.



One Sided vs Two Sided Tests

- We may update our definition of more extreme to encompass large positive and negative deviations from the null hypothesis specification $\theta = 0.5$.
- Our critical region is then the union of the two sets, for c_z the z^{th} quantile of the distribution under H_0

$$\{X \leq c_{0.025}\} \cup \{X \geq c_{0.975}\}$$
$$\{X \leq 13\} \cup \{X \geq 27\}$$

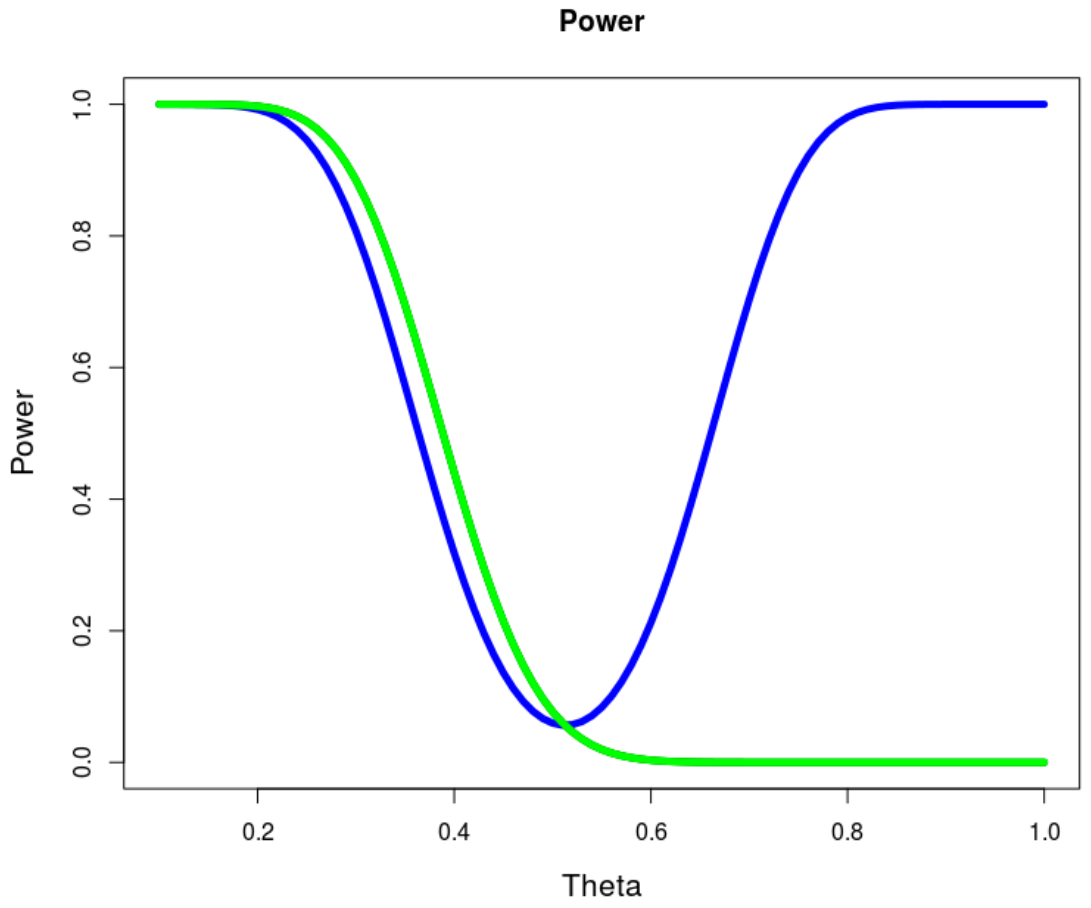
- This is referred to as a **two-tailed test**, in comparison to the **one-tailed test** we considered previously.

Making Stronger Assumptions

- We may again plot the power curve for the two specifications of alternative hypothesis

- One-tailed test (green)
- Two-tailed test (blue)

- The one-tailed test is more powerful for $\theta < 0.05$



Composite hypotheses

- In principle, a null hypothesis can postulate more than one value for the target state of nature
  - For example

$$H_0 : \quad \theta \geq 0.5$$
$$H_1 : \quad \theta < 0.5$$

- Such a hypothesis is known as a **composite hypothesis**
  - In the case where H_0 specifies a single value it is referred to as a simple hypothesis.

Composite hypotheses

- We will not go into any further detail on composite hypotheses, but it is worth knowing that they exist.
- In our context, it suffices to say that we can look at the “hardest” value to falsify ($\theta = 0.5$) and proceed with this as a simple hypothesis.
- For other states of nature in H_0 (for example $\theta = 0.6$) our Type I error rate will be of smaller size than the designed level (e.g. 0.05)
  - However, in the worst case scenario ($\theta = 0.5$) we know we are still controlling the Type I error rate at 0.05

An important note

- When performing a hypothesis test a large p-value may be caused two different things

1. H_0 may be true
2. H_0 may be false, but the power of our test is too low to detect this

Strategy

- For a given level, pick the test which maximises power regardless of the true hypothesis
  - This is easier said than done
  - Only in some cases are there uniformly most powerful tests (tests which are at least as good as any other test for any value of the true hypothesis)
- In the following slides we will introduce some common tests and their applications, without detailed mathematical discussions

Historical Note: Neyman-Pearson

- The framework of controlling Type I error and minimising Type II error was introduced by Neyman and Pearson in the early 20th century
  - As a result, this has since become known as the Neyman-Pearson framework
- Both Neyman and Pearson have links to UCL, having worked for the University at points during their careers

Some Useful Tests

Warning

- The following slides may sound like an intense laundry list of techniques
  - The most important part for now is understanding the general logic behind the testing procedures.
  - With practice the specific applications, pros and cons of each of the tests will become clear.
- Further details are provided in courses STAT0029 (Statistical Design of Investigations and STAT0030 (Statistical Computing).
- For each test, try to think of every step of the hypothesis testing procedure. You may also find more details in the Wasserman book, which I would highly recommend going through.

Hypothesis Testing Procedure

- 1) Specify a null and alternative hypothesis
- 2) Specify the level of the test
- 3) Specify a suitable test statistic
- 4) Determine the distribution of the test statistic under H_0
- 5) Determine what it means to be “more extreme” by considering H_0 and H_1
- 6) Determine the corresponding p-value or critical region
- 7) Reject H_0 if
  - the p-value is less than the level of the test
  - or if the test statistic is outside the critical region

The t-test

Original motivation: yields of barley

- As a result, we will illustrate the t-test with a problem of quality control for the Guinness stout by William Gosset (a.k.a. Student).
- Say you are measuring barley concentration in small beer samples. Your sample is assumed to follow some unknown iid Gaussian distribution:

$$X_i \sim N(\mu, \sigma^2), i = 1,2, \dots, n$$

The t-test

- We would like to know if we are correctly manufacturing the Guinness with a target mean barley concentration of $\boldsymbol{\mu_0}$.
- We may formulate this hypothesis test as:

$$H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$$

- Notice that in most cases the alternative hypothesis is simply the negation of the null hypothesis.

T-test: Distribution of the test statistic

- Gosset proposed the following test statistic:

$$T = \frac{\sqrt{n}\,(\bar{X}_n - \mu_0)}{S_n} \sim \mathcal{T}(n-1)$$

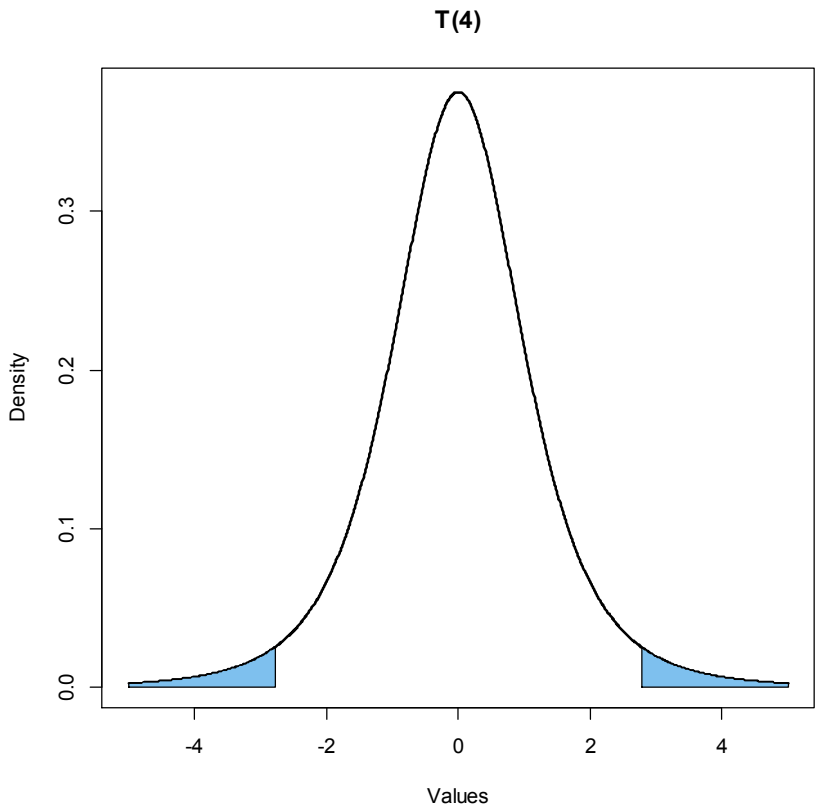
$\bar{X}_n$: sample mean

S_n: sample standard deviation

- Notice that this is called a student-t distribution with n-1 degrees of freedom.
- The formula may be ugly, but for large n it is essentially a Normal(0,1) distribution.

T-test: Extreme Values

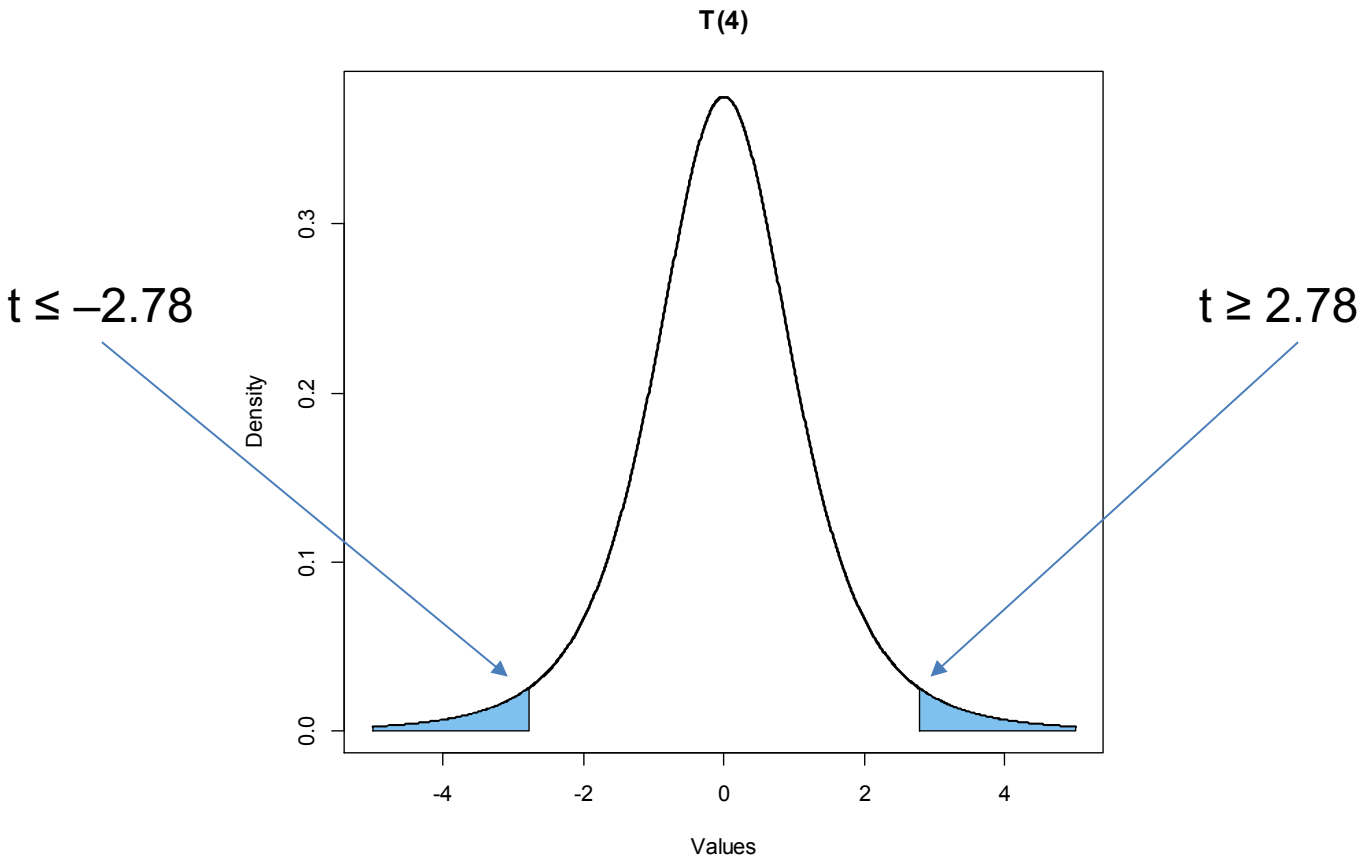
- We may now consider the probability of "extreme values"
- The true mean may differ from μ_0 by being smaller or larger. We therefore consider a two-tailed test in this example.
- The blue shaded region in the figure is the critical region.



T-test: Extreme Values

- We reject H_0 with level α only if the test statistic is smaller than the respective $\alpha/2$ quantile, or larger than the $1 - \alpha/2$ quantile.

- For instance, if $n = 5$, $\alpha = 0.05$ we obtain the image to the right.



The Wald Test

- Recall our previous discussion of the Central limit Theorem?
  - For short: averages of large samples approach Gaussian rand. variables.

- The t-test is motivated by small sample sizes, normally distributed.

- The **Wald test** uses the same test statistic as the t-test, but the interpretation is different, and as a result the distribution of the test statistic also differs.
  - Samples can be from “any” distribution.
  - Sample sizes must be assumed to be “big enough” that the CLT applies.
  - Hence the distribution of the test statistic is Normal(0,1).
  - Remember that the t-distribution approaches N(0,1) as n grows.

- Read Chapter 10 from the Wasserman book and do Q3 from Exercise Sheet 2.

Testing more complicated hypotheses

- Testing equalities of means is one thing.
  - The t-test is a useful tool in such cases.
- We can however think of testing more general assumptions.
  - How might we test other **constraints** on distributions?

Testing Independence

- Consider an example in which we wish to test whether two discrete variables are independent.

- In probabilistic terms, independency is equivalent to the constraint on the joint distribution that $P(X, Y) = P(X)P(Y)$.
  - An implication of this result and the rules of conditional probability is that $P(Y|X) = P(Y)$.
  - That is, the outcome of X has no impact on the outcome of Y.

Contingency Tables

- Tests of independence are often applied to data summarised by what is known as a **contingency table**.
  - D_j = depression, sibling j
  - A_j = dependence on alcohol, sibling j

		$D_1 = 0$		$D_1 = 1$	
		$D_2 = 0$	$D_2 = 1$	$D_2 = 0$	$D_2 = 1$
$A_1 = 0$	$A_2 = 0$	288	80	92	51
	$A_2 = 1$	15	9	7	10
$A_1 = 1$	$A_2 = 0$	8	4	8	9
	$A_2 = 1$	3	2	4	7

- https://arxiv.org/pdf/0707.3794.pdf

- We may wish to determine if there is an association between depression and alcohol dependency across different subjects.

Chi-square test

- The **chi-square test** compares “expected” versus “observed” outcomes.
- In a nutshell: in this case, for every combination of the two variables we may compare the frequency of co-occurrence against the product of the marginal frequencies.
  - We may derive a test statistic using a particular way of aggregating these numbers.
- The resulting statistic, Pearson’s statistic, has a so-called chi-square distribution.
  - In general, this test can be used to check whether a particular probability mass function explains some observed **multinomial data**.
- See Wasserman book and Exercise Sheet 2 Q4 & Q7 for more details.

Paired tests

- As a final example hypothesis test, consider comparing measurements which are tied to a single unit.
  - A person, a beer vat etc.
- Such measurements are typical when we apply a treatment to an individual and wish to contrast the results pre- and post-treatment.
  - More information on such tests might be seen in the course STAT0029: Statistical Design of Investigations.

Paired tests: Signed Rank Test

- A **signed rank test** (also known as the Wilcoxon test) compares the differences of two measurements, $X^{(i)}$ and $Y^{(i)}$, from a single individual i.

- The idea is to build data as if it came from the null hypothesis.
  - For that, build differences $X^{(i)} - Y^{(i)}$ for all possible pairs and consider their sign.
  - Under the null, it is possible to find the distribution of a test statistic based upon these rearrangements.

- The null hypothesis in question is whether $P\left(X^{(i)} > Y^{(i)}\right) = P\left(Y^{(i)} > X^{(i)}\right)$.
  - How might we approach such a question if we assumed that the observations are normally distributed with the same variance?

Ongoing Debates about Statistical Significance

Before we conclude on hypothesis testing

- Let's remind ourselves of why we are doing this:
  - As statisticians, we want to be **explicit** about our assumptions and think about whether these are reasonable.
  - We might want to test whether this is the case.
- Hypothesis tests can be useful, but are not a silver bullet to all statistical problems and they can be misused.
  - Lies, damned lies and statistics.

Objections to hypothesis testing

1. The null hypothesis is effectively always false, especially depending up the amount of precision used.
    - Is Gosset's mean concentration $\boldsymbol{\mu_0}$, or is it actually $\boldsymbol{\mu}_0+10^{-20}$?

2. Dichotomisation of decisions can lead to inconsistencies.
    - We can accept some null hypothesis H_0^i, and reject another hypothesis H_0^j, even if H_0^i implies H_0^j.

- All of this can be confusing and leads people to misuse tests.

Why do hypothesis testing?

- A common industrial practice is A/B testing.
  - Do two treatments, say offering a product with two different variations (price, colour, user interface, etc.).
  - Perform a test to determine whether the distribution of outcomes (sales, customer satisfaction, etc.) change in some way (mean, variance, maximum value, etc.).
  - Specify the null hypothesis as the “no change” hypothesis.
- Kohavi et al. (2009) "Controlled experiments on the web: survey and practical guide“, http://dl.acm.org/citation.cfm?id=1485091.

Why do hypothesis testing?

- Granted, you may be (should be?) interested in the size of the effect.
  - For example, the difference in sales means.
- However, do you have a large enough sample in order to distinguish it from zero?
- The machinery of hypothesis testing helps to tell you whether you are asking a ridiculous question to begin with.
  - That is, estimating effect size when the sample size you have can't really distinguish it from zero.

Why do hypothesis testing?

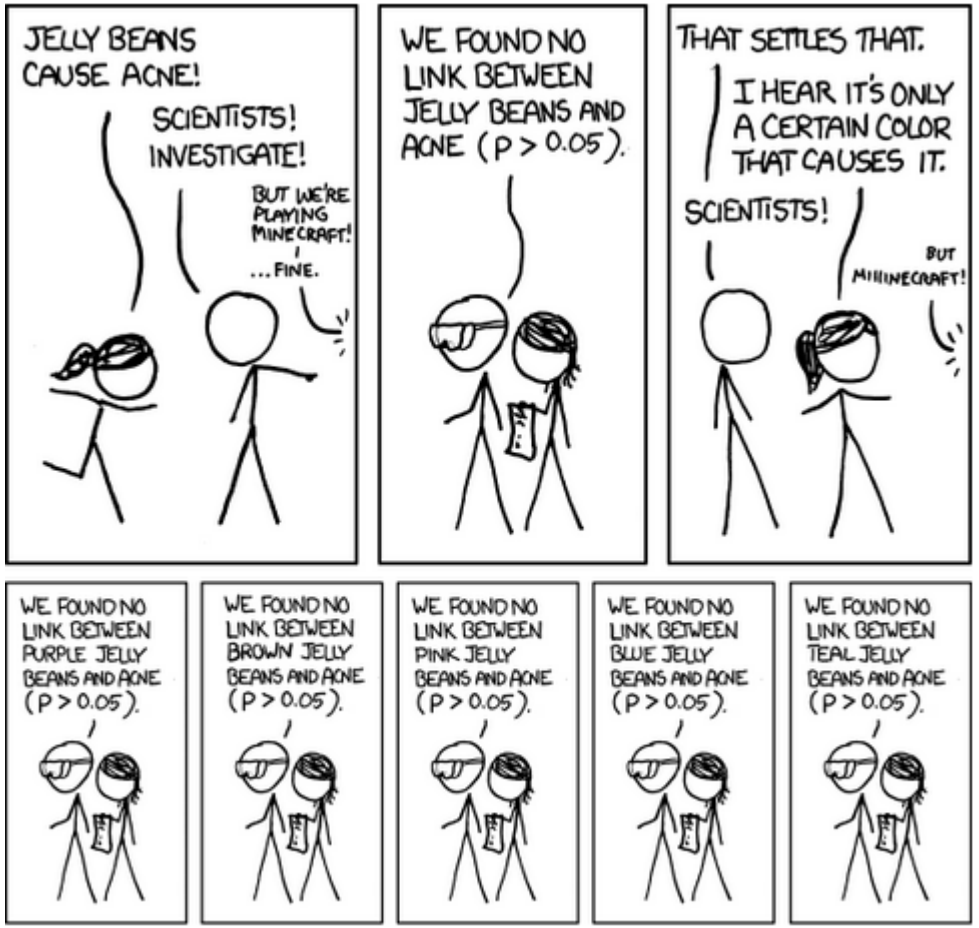
- Physical/social/psychological measurements have practical limits.
- A null hypothesis can be highly precise, and yet not falsifiable with the given technology.
- Moreover, background knowledge might tell us that the precision of the null is good enough.
  - An example is discovery of the Higgs Boson.
  - van Dyk (2014). "The Role of Statistics in the Discovery of a Higgs Boson".
  - http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-062713-085841

Why avoid hypothesis testing?

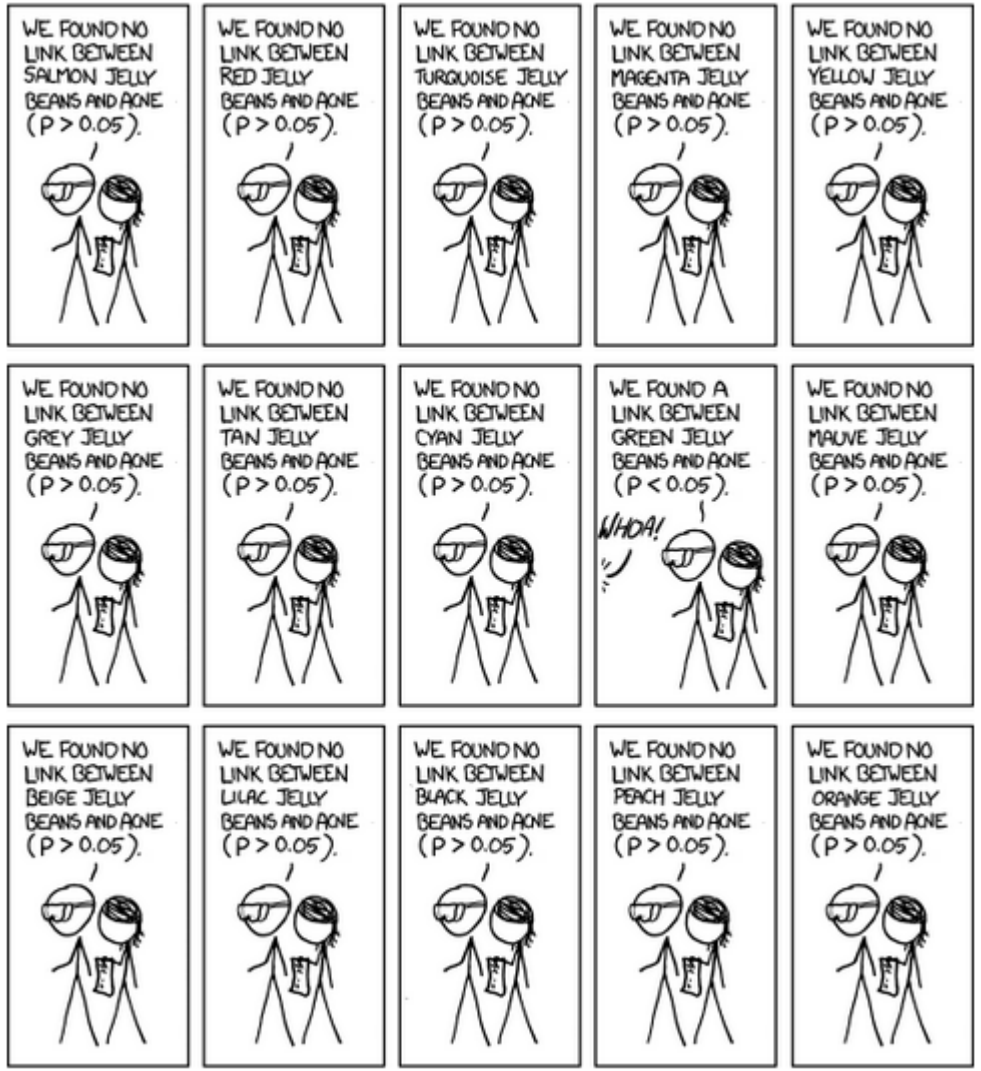
- Essentially, if you think that you can get away without doing anything else.

- Hypothesis testing is (or should be) hardly convincing by itself.
  - Effect size matters.
  - Validation of assumptions/sample size may be just the starting point of a solid analysis.
  - **Statistical significance is not practical significance.**

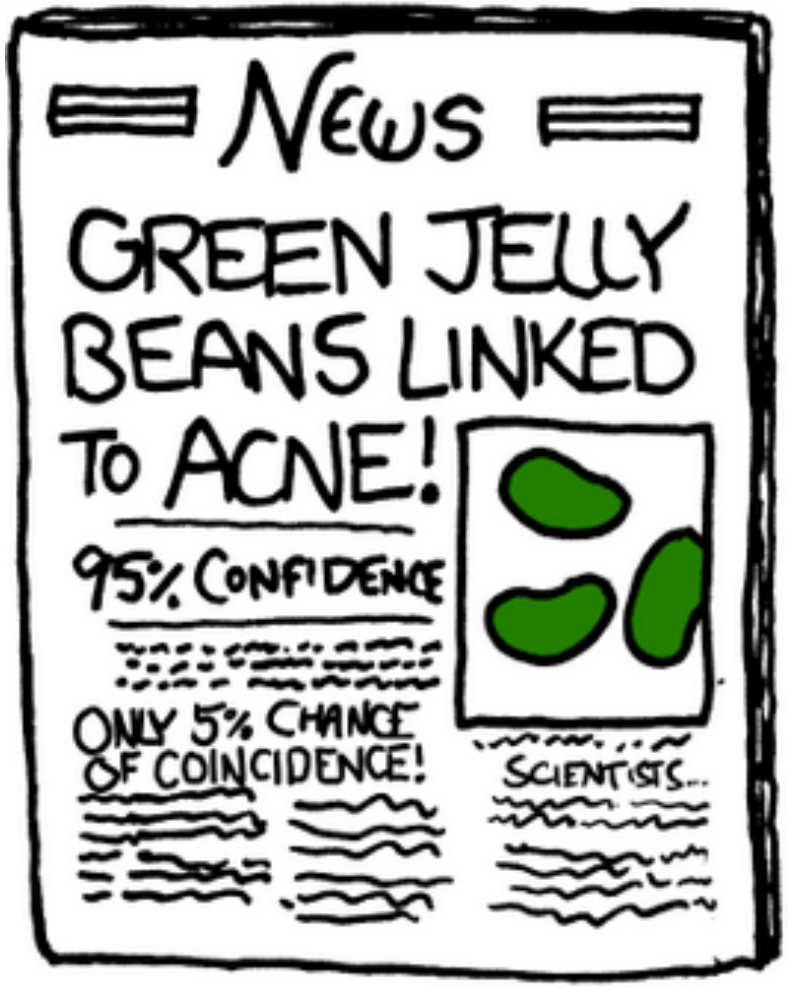
The sociology of hypothesis testing and p-hacking



https://xkcd.com/882/



The sociology of hypothesis testing and p-hacking



The sociology of hypothesis testing and p-hacking

- There are often perverse incentives for “p-hacking”, that is, selectively reporting p-values.
  - Multiple tests on a single data set are not independent.
  - The p-value of a single hypothesis test is uniformly distributed, allowing us to specify that the Type I error (false positive) probability is simply the size of the test, α.
  - **The minimum of a set of p-values is not uniformly distributed** and as a result it is much more difficult to determine the error rates.

- Hopefully, you will always be responsible.

The sociology of hypothesis testing and p-hacking

- There are two different types of bad incentive.

1. "The null is bad."
   - If I am proposing a new treatment and the null is a zero difference with respect to the old treatment.
   - Down with the p-value!

2. "The null is good."
   - If I'm proposing a model and the null is "the model generated the data".
   - Up with the p-value.

Multiple testing

- There is a considerable literature on multiple testing and so we will not go into any great detail during these lectures.

- One of the simplest correction techniques for multiple testing is the **Bonferroni correction**.
  - This makes use of the result that $P(A \cup B) \leq P(A) + P(B)$.

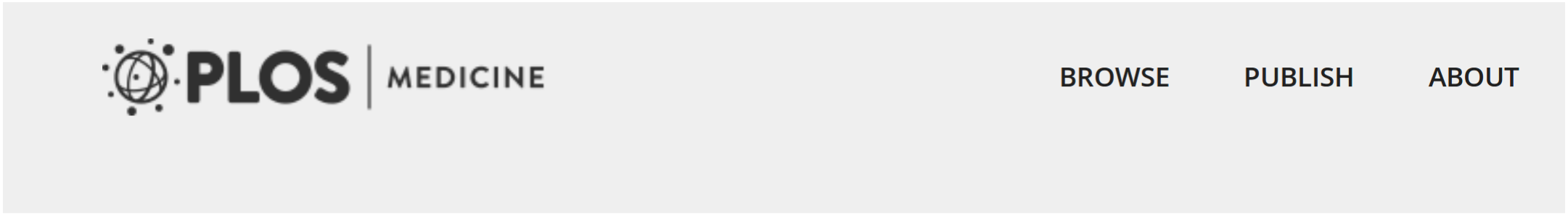
- If you have k hypotheses to test, think of A's and B's as the Type I error events.

Multiple testing

- Without knowing the (perhaps very complicated) joint distribution of these events, it suffices to control the level of the joint test by changing the level α of each test to α/k.

- This correction will control the combined level, but the resulting probability of a Type I error may be much smaller than α.
  - Recall that reducing the Type I error rate often leads to an inflated Type II error rate and as a consequence poor power is likely to follow.

The replication crisis



OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Getting rid of statistical significance?

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

A well-known issue...

MENU

nature
International journal of science

Letter | Published: 21 September 1935

Statistical Tests

K. A. FISHER

Nature **136**, 474 (1935) | Download Citation
1110 Accesses | **36** Citations | **9** Altmetric | Metrics »

Abstract

IN a letter to NATURE of August 24, Prof. Karl Pearson states: " From my point of view, the tests are used to ascertain whether a reasonable *graduation* curve has been achieved, not to assert whether one or another hypothesis is true or false."

"For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning."

Take home message

- The framework of hypothesis testing allows us to put assumptions to the test by deriving their consequences to the observed data.
- Despite its shortcomings, hypothesis tests are a common type of diagnostics.
  - As long as we do not (or only in rare cases) take them as the ultimate goal of an analysis they can be valuable.