STATG006: Introduction to Statistical Data Science

Thomas Honnor

t.honnor@ucl.ac.uk

Department of Statistical Science

University College London

Chapter 2

Statistical assessment: Hypothesis testing and confidence intervals

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?**

- Our assumptions imply conclusions
  - **How would our conclusions change if we had observed different data?**
  - What can we say about the long-run behaviour of our conclusions?

Outline of this Chapter

- Hypothesis tests and p-values
  - The good, the bad and the ugly

- Confidence intervals

- Computational aspects
  - Central Limit Theorem
  - The bootstrap

Hypothesis Testing

A preliminary example

A simple example for illustration

- Suppose we were to offer a training course and wish to determine if there is a gender imbalance

- 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 gender balance?
  - We will assume for simplicity that the proportion of females 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?
  - Why consider the specific value 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?

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
  - Intuitively, the number of female students provides a summary for our example

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 students sex numerically as a random variable
  - 0 for males, 1 for females
  - We then have independent Bernouilli random variables (“coin flips”)

The Binomial distribution

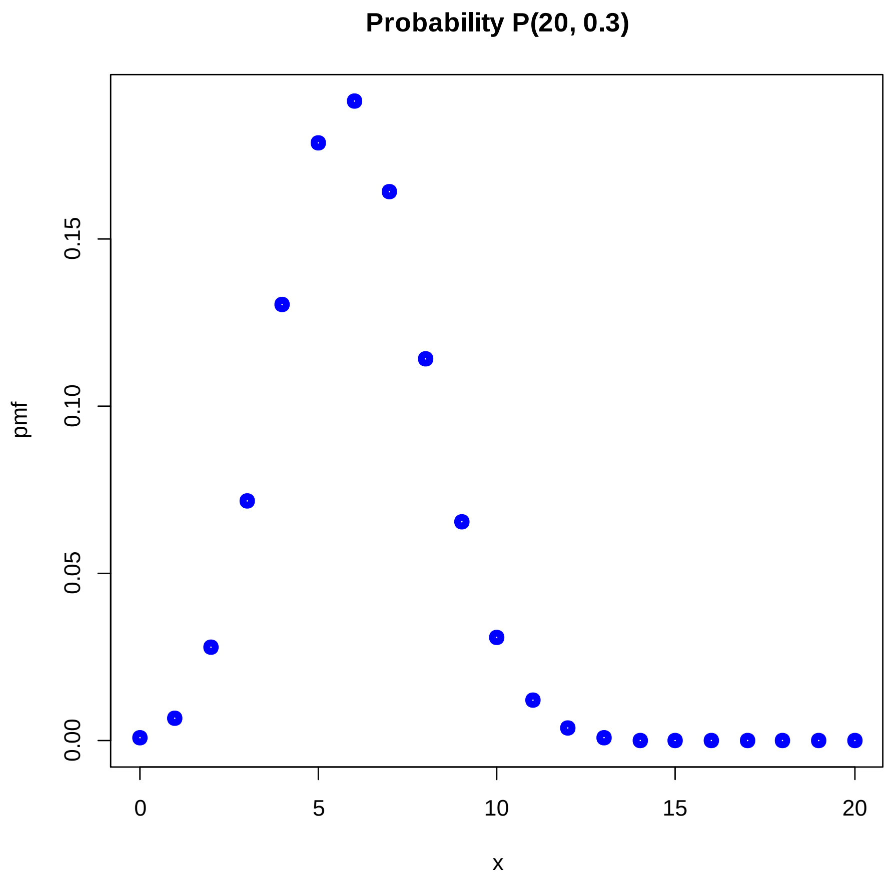
- If we have *n* independent Bernouilli trials, each with probability *θ*, we have a binomial distribution.

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

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

- Pmf refers to the probability mass function

- R code example



Returning to our test statistic

- In our example we observe Bernouilli 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: $\theta = 0.5$

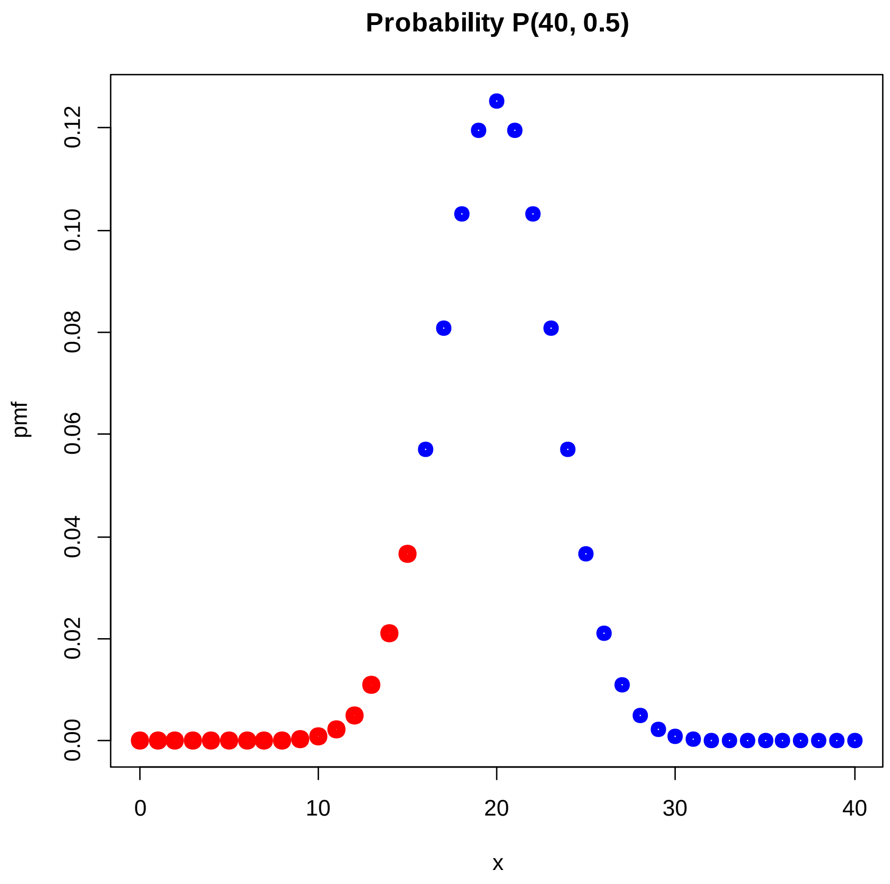
- We will in fact characterise how probable values of X of size 15 or smaller are
  - We assumed that θ is not greater than 0.5
  - Values of X less then or equal to 15 are therefore all of those which are as or more extreme than our observation

The p-value

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

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

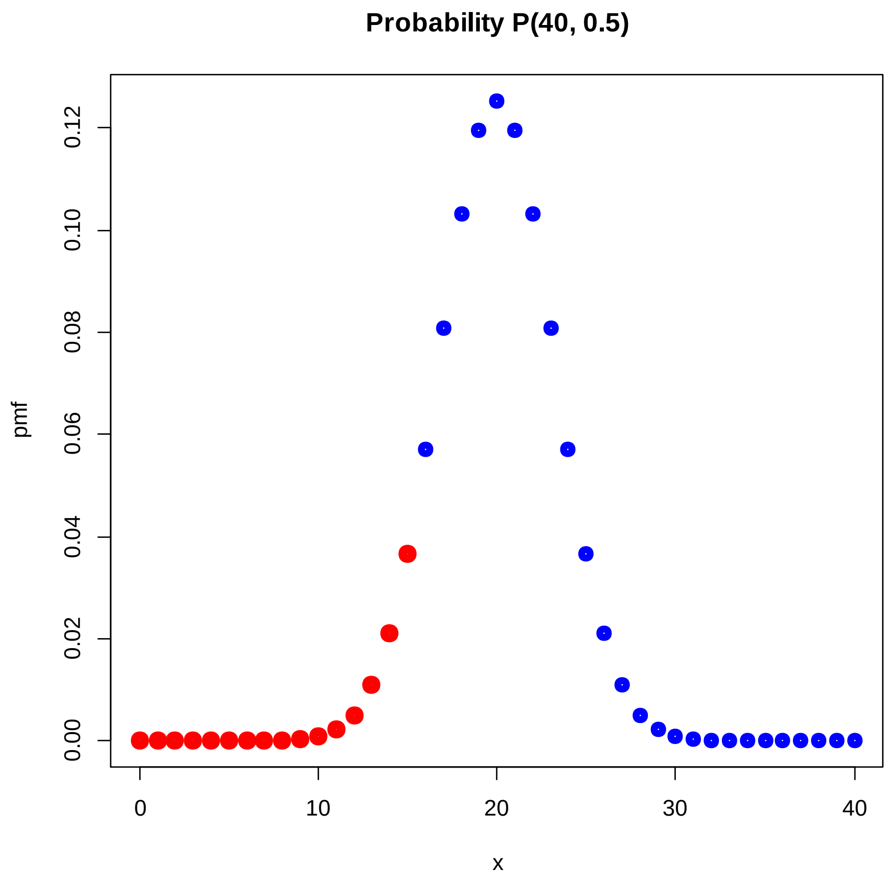
$$p = \sum_{x=0}^{15} \binom{40}{x} 0.5^x (1 - 0.5)^{(40-x)}$$



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 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 STATG004: Bayesian Analysis

A logical analogy

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

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

- 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

A logical analogy

- When used in practise with a threshold of 0.05 this is an informal method of reasoning and can be easily criticised
- In future we will see another interpretation based upon long-run trade-offs between ‘false positives’ and ‘false negatives’
- Ultimately, we will also present a pragmatic guide on when and why to use null hypothesis testing
  - This is not a tool without controversies

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

- Can we point out what else might falsify H_0 in our example?

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

Hypothesis Testing

Probabilities, power and types of error

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
  - However, calculation of the power requires specification of an alternative hypothesis

Returning to the example

- 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 the observed is less than or equal 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
  - 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

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

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

- 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

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

- 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 is the critical region of this test and that the Type I error rate is 0.05

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

- In the previous slides (lecture) 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
  - 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

Level

- In the example presented in the previous slides (lecture) 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 II errors

- We stated previously that a Type I error occurs when we reject the null hypothesis, H_0, when it is true

- On the other hand a Type II error occurs when we fail to reject H_0 when it is false

- 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

Type II error

- 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

Testing procedure

- Specify a null and alternative hypothesis

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

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

- 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

$$\text{Level} = 0.05$$

- Specify a suitable test statistic

$$X = \text{The number of females} = 15$$

Testing procedure

- Determine the distribution of the test statistic under H_0

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

- 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

- Determine the corresponding p-value

$$p = P(X \leq 15)$$
$$= 0.077$$

- Reject H0 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 proportion of females and males is identical

Alternative procedure

- 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

$$P(X \leq 15) = 0.77$$

Should be 0.077

$$P(X \leq 14) = 0.40$$

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

Power investigation

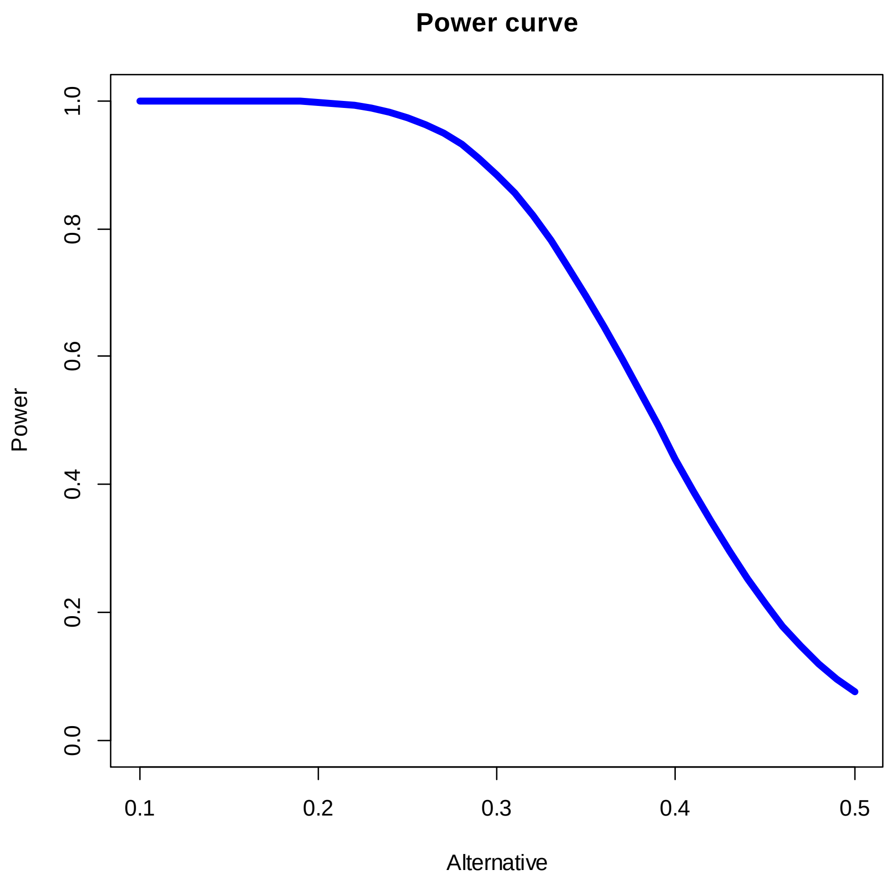
- We may investigate how the power of the test varies for a range of alternative values of θ, the true state of nature

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

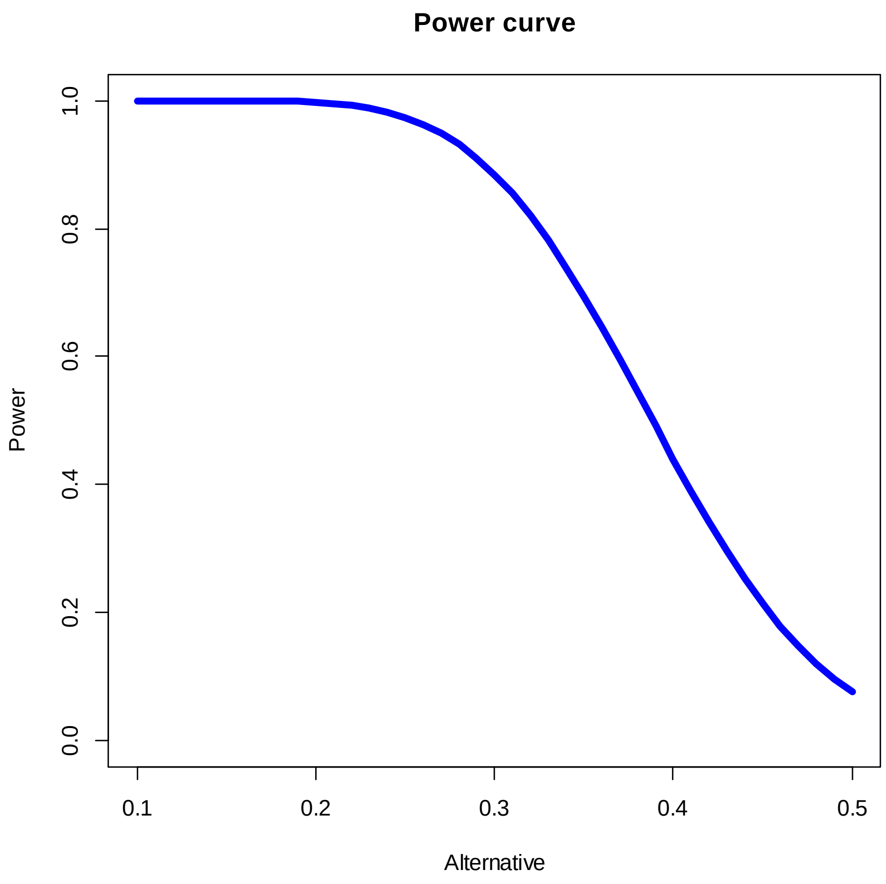
$$P(X \in CR|\theta = 0.2) = 0.992$$
$$P(X \in CR|\theta = 0.3) = 0.807$$
$$P(X \in CR|\theta = 0.45) = 0.133$$

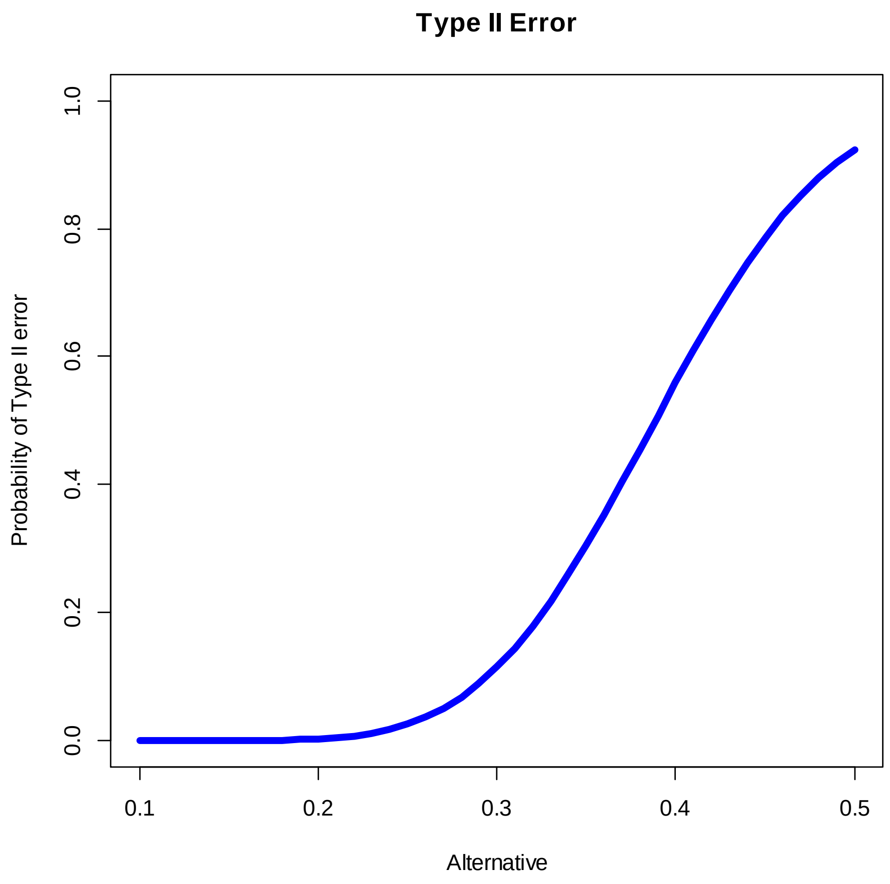
Power investigation

- 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



Power investigation





More powerful tests

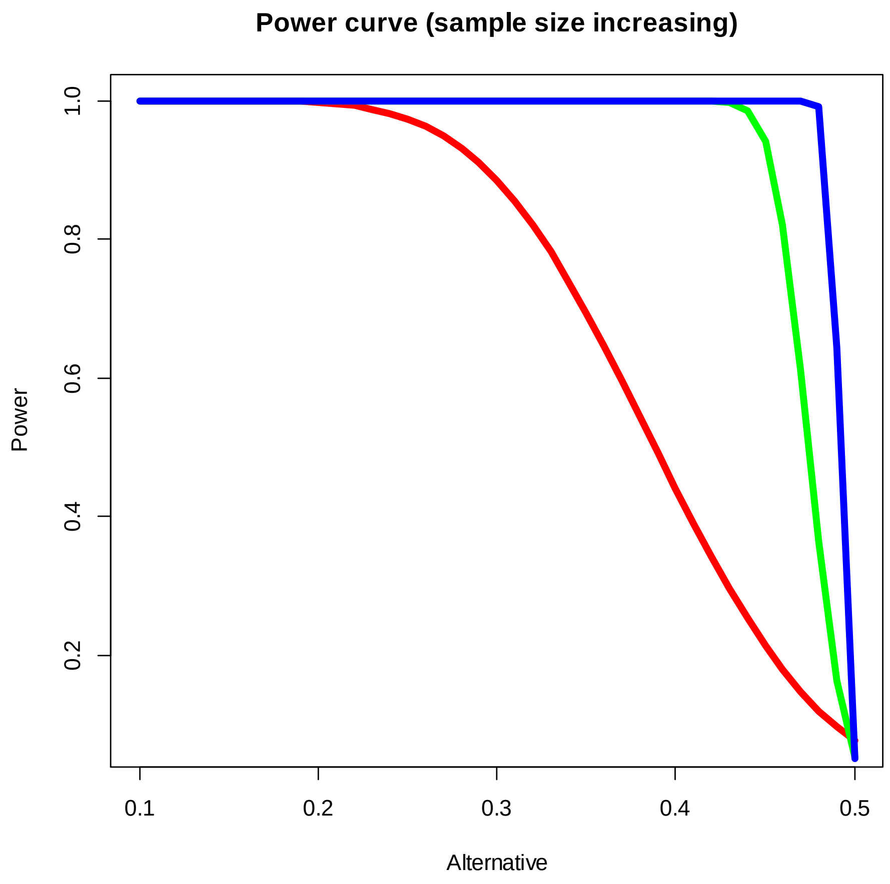
- Given that we cannot change the underlying value of θ, how might we increase the power of our test?
  - Collect more data (increase the sample size)
  - Allow for a higher Type I error (increase the level of the test)
  - Use a better test statistic
  - Make stronger assumptions

- 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

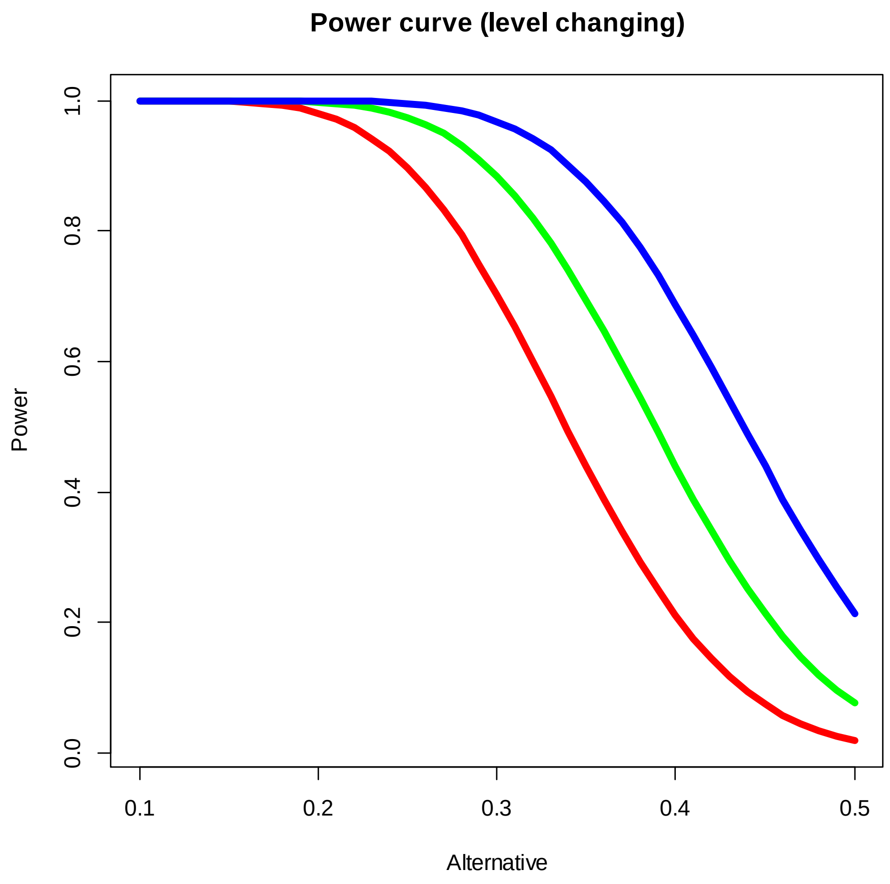
- If you decrease the level of the test, you reduce the size of the critical region and make it more difficult to distinguish H_0 from the alternatives

- 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)



Choice of test statistic

- 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

Making stronger assumptions

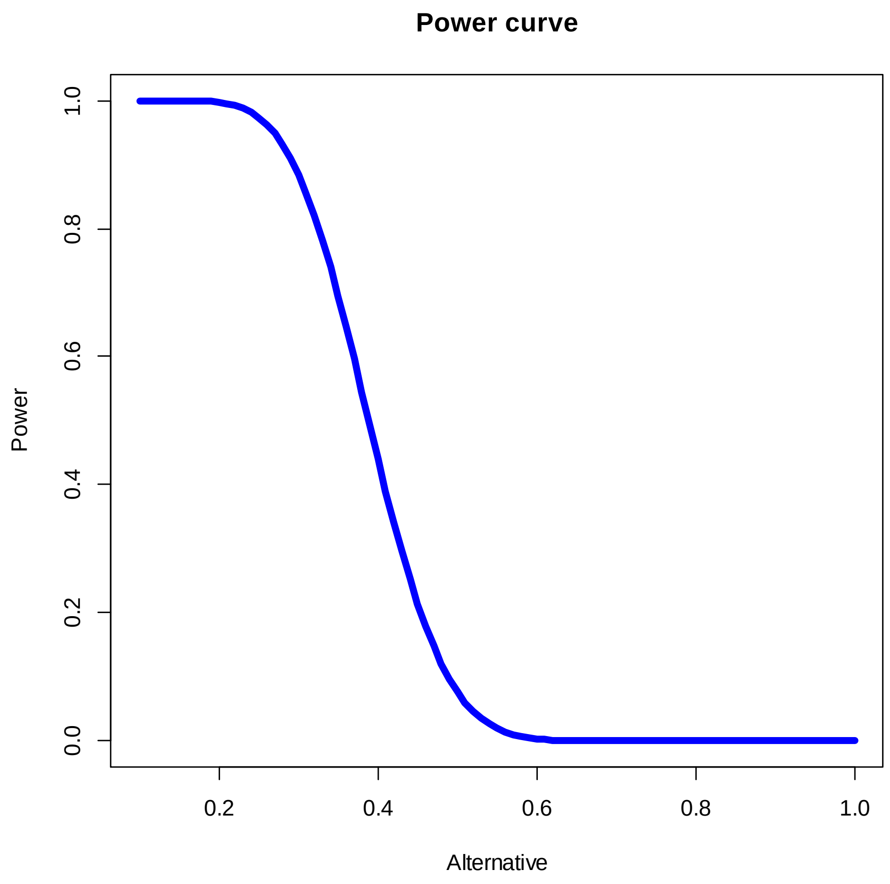
- If we make stronger assumptions with the consequence of reducing our set of alternatives we may increase the power of our test

- 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



Making stronger assumptions

- 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

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

$$\{X \leq 13\} \cup \{X \geq 27\}$$

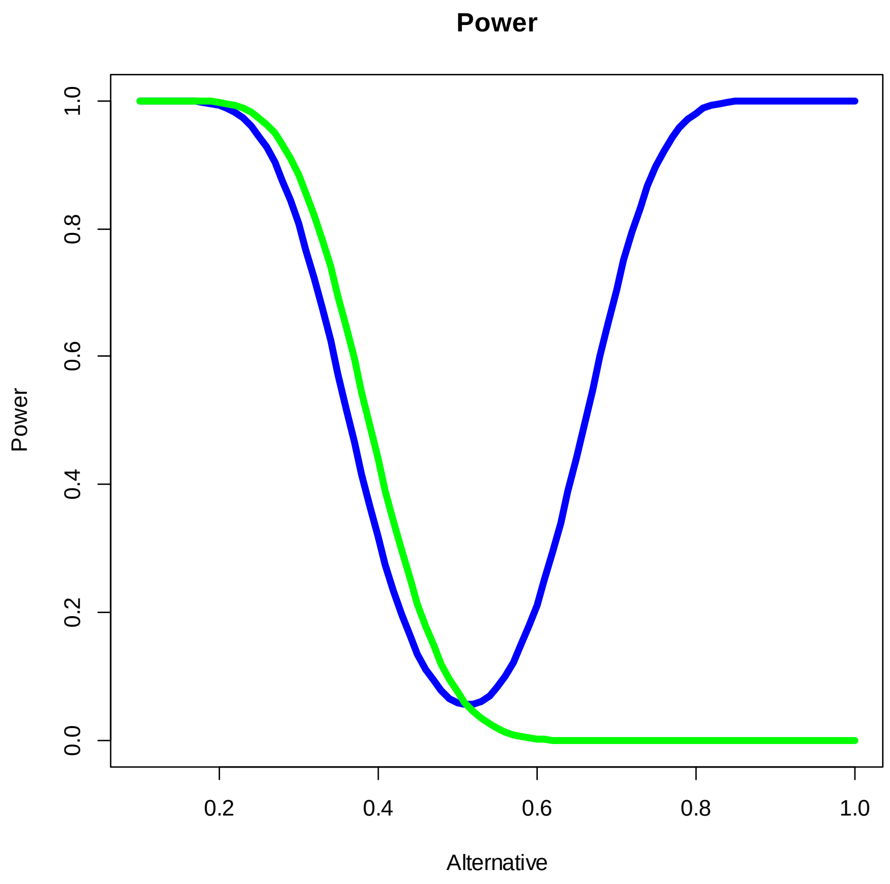
- This is referred to as a **two-tailed test**, in comparison to the **one-tailed test** we considered previously

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

$$\begin{aligned} H_0 &: \quad \theta \geq 0.5 \\ H_1 &: \quad \theta < 0.5 \end{aligned}$$

- 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

- H_0 may be true

- 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

- 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 both have links to UCL, having worked for the University at points during their careers

Hypothesis Testing

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 practise the specific applications, pros and cons of each of the tests will become clear

- Further details are provided in courses STATG002 and STATG003 and in the later chapters of the STAT1005 notes