FIT2086 Lecture 5
Hypothesis Testing and Model Selection

Daniel F. Schmidt

Faculty of Information Technology, Monash University

September 20, 2017

Outline

Revision from last week (1)

- Central limit theorem; if $Y_1, \ldots, Y_n$ are RVs with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ then

$$\sum_{i=1}^{n} Y_i \xrightarrow{d} N\left(n\mu, n\sigma^2\right)$$

- Implies distribution of the sample mean $\bar{Y}$ for $Y_1, \ldots, Y_n$ with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$ satisfies

$$\bar{Y} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Revision from last week (2)

- $100(1-\alpha)\%$ confidence intervals: cover the true population parameter for 95% of possible samples we could draw from our population
- $100(1-\alpha)\%$ confidence interval for mean μ of normal population with unknown variance σ^2:

$$\left(\hat{\mu} - t_{\alpha/2,n-1}\sqrt{\sigma^2/n},\ \hat{\mu} + t_{\alpha/2,n-1}\sqrt{\sigma^2/n}\right)$$

where

$$\sqrt{\sigma^2/n} = \sqrt{\mathbb{V}\left[\hat{\mu}\right]}$$

is the standard error, and $t_{\alpha/2,n-1}$ is the $100(1-\alpha/2)$ percentile of a t distribution with degrees-of-freedom $n-1$

Outline

Modelling data (1)

- Over the last two weeks we have looked at parameter estimation
- In week 3 we examined point estimation using maximum likelihood
  - Selecting our "best guess" at a single value of the parameter
- Last week we examined interval estimation using confidence intervals
  - Give a range of plausible values for the unknown population parameter

Modelling data (2)

- This week we are looking at the evidence in the data about certain hypotheses
- In statistical parlance, a hypothesis is usually expressed in terms of parametric distributions
- We might be asking:
  - Are the parameters of a model equal to some specific value?
  - Does one model fit the data better than another?
- Most common statistical hypothesis testing problems can be expressed using one of these two questions

Hypothesis Testing (1)

- Let us begin with the first question
- We ask whether there is evidence against a null hypothesis
- More formally, we say we are testing

$$\begin{aligned} H_0 &: \text{Null hypothesis} \\ &\text{vs} \\ H_A &: \text{Alternative hypothesis} \end{aligned}$$

  on the basis of our observed data $\mathbf{y}$
- What does this mean?

Hypothesis Testing (2)

- We are taking the null hypothesis as our default position
- Then asking how much evidence the data carries against the null hypothesis?
- Imagine we model the population using a normal distribution; then, we might set up the hypothesis:

$$H_0 \quad : \quad \mu = \mu_0$$

vs

$$H_A \quad : \quad \mu \neq \mu_0$$

- We are asking: "is there sufficient evidence in the data to dismiss the hypothesis that μ is equal to some fixed value μ_0?"

Hypothesis Testing (3)

- For example, imagine we found from a very large study that the average height of European people is 1.7m
- We measure the heights of a sample of Chinese people
- We might ask – are Chinese people on average the same height as Europeans?
- We can then set up the hypothesis:

$$H_0 \quad : \quad \mu = 1.7m$$

vs

$$H_A \quad : \quad \mu \neq 1.7m$$

- Obviously the sample mean will never be exactly 1.7 even if that is the population average height of Chinese people.
- So how do we scientifically try and answer this question on the basis of the data?

Hypothesis Testing (4)

- We use the Neyman-Pearson framework
- In this approach, we are interested in the evidence against the null hypothesis.
- To do this, we ask: "How likely would it be to see our data sample $\mathbf{y}$ by chance if the *null hypothesis were true*?"
- So key ideas
  - We assume null hypothesis is true;
  - we calculate the probability of observing our sample by chance if it were true.
- The smaller this probability, the stronger the evidence against our null being true

Testing μ with known variance (1)

- Let us first look at the following problem
- Assume our population is normally distributed with known variance σ^2, unknown mean
- Given a sample $y_1, \ldots, y_n$ from our population, our test is:

$$H_0 \quad : \quad \mu = \mu_0$$

vs

$$H_A \quad : \quad \mu \neq \mu_0$$

- As previously mentioned, the ML estimate $\hat{\mu} \neq \mu_0$ just due to random chance, even if the population mean μ is equal to μ_0
- So instead ask: how unlikely is the estimate $\hat{\mu}$ we have observed if the population mean was $\mu = \mu_0$?

Testing μ with known variance (2)

- Under our assumptions, if null was true then

$$Y_1, \ldots, Y_n \sim N(\mu_0, \sigma^2)$$

- Our maximum likelihood estimate of the population mean is the sample mean

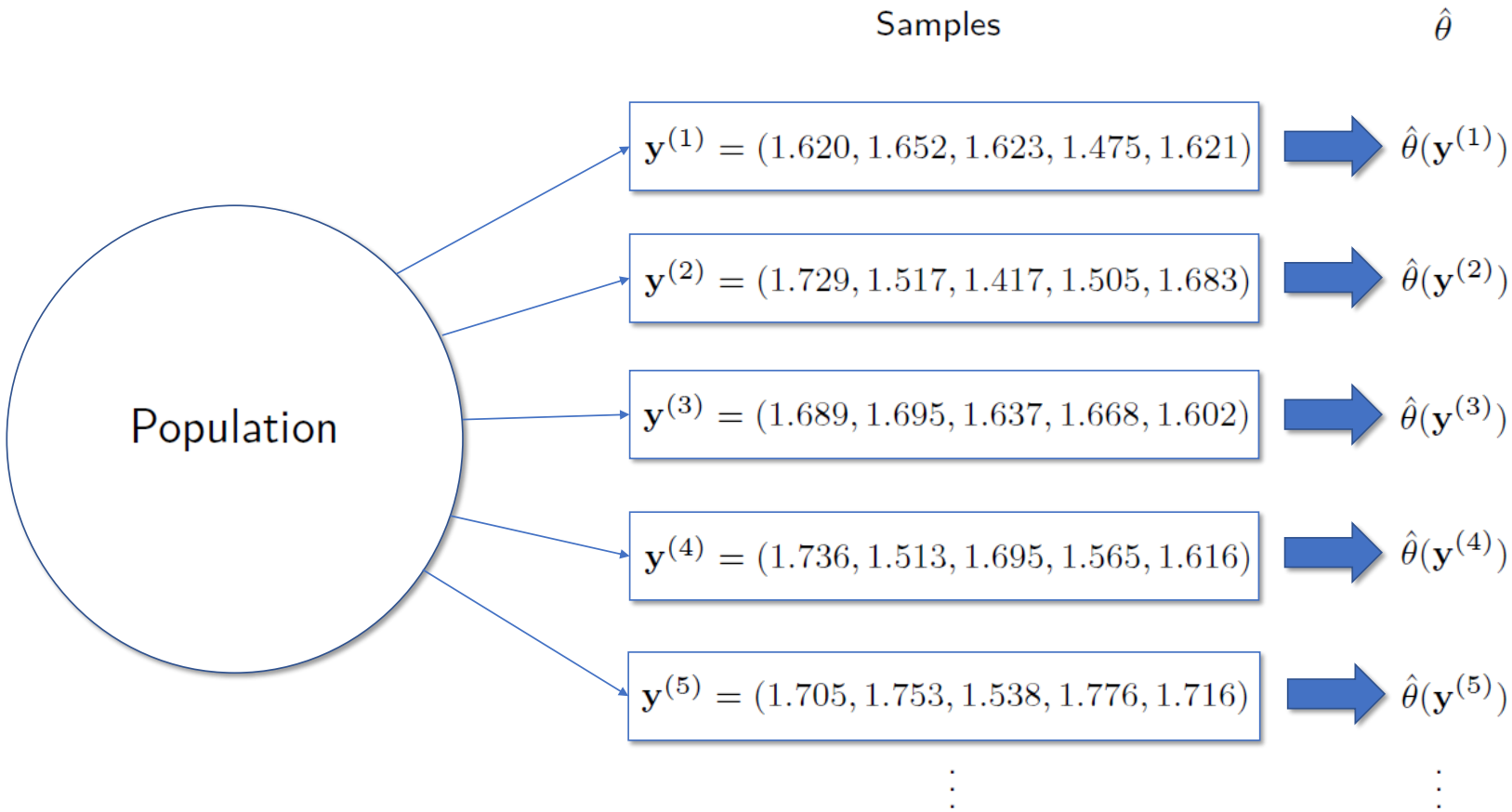
$$\hat{\mu} \equiv \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- Under this assumed population model, we can recall the sampling distribution of the mean is

$$\hat{\mu} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

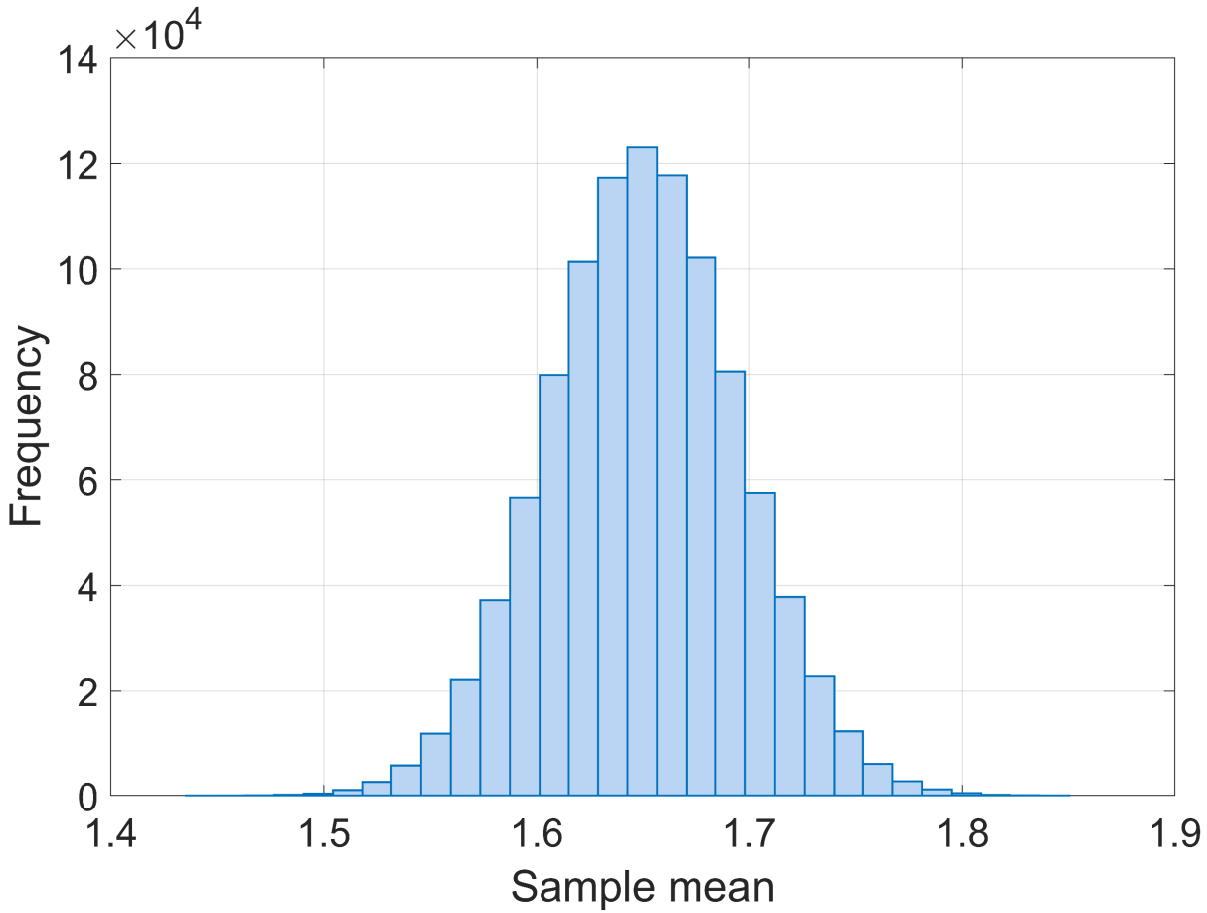
- This is the distribution of the sample mean $\hat{\mu}$ if we repeatedly took samples of size n from our population

Sampling Distributions



An (infinite) number of different random samples can be drawn from a population. Each sample would lead to a potentially different estimate $\hat{\theta}$ of the population parameter θ. The distribution of these estimates is called the sampling distribution of $\hat{\theta}$.

Sampling Distribution of the Mean



Histogram of sample means of $1,000,000$ different data samples, each of size $n = 5$, generated from a $N(\mu = 1.65, \sigma = 0.1)$ distribution.

Testing μ with known variance (3)

- Imagine we have observed a sample $\mathbf{y} = (y_1, \ldots, y_n)$
- The difference between $\hat{\mu}$ and μ_0 is a measure of how much the sample differs from the mean in our null hypothesis
- $\hat{\mu}$ will never equal μ_0, even if the population mean is μ_0, just because of randomness in our sampling
- However, the bigger the difference, the more the sample is at odds with our null hypothesis assumptions
- How to determine how likely it would be to see a difference of this size (or greater) just by chance?

Testing μ with known variance (3)

- Imagine we have observed a sample $\mathbf{y} = (y_1, \ldots, y_n)$
- The difference between $\hat{\mu}$ and μ_0 is a measure of how much the sample differs from the mean in our null hypothesis
- $\hat{\mu}$ will never equal μ_0, even if the population mean is μ_0, just because of randomness in our sampling
- However, the bigger the difference, the more the sample is at odds with our null hypothesis assumptions
- How to determine how likely it would be to see a difference of this size (or greater) just by chance?

Testing μ with known variance (4)

- If the null *is true*, then sampling distribution of $\hat{\mu}$ is

$$\hat{\mu} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

- Calculate the z-score for our estimate $\hat{\mu}$ under the assumption the null hypothesis is true

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}$$

  which represents a standardised difference between the null μ_0 and our sample estimate $\hat{\mu}$
  - It tells us how many standard errors, $\sigma/\sqrt{n}$, the estimate $\hat{\mu}$ is away from the null $\mu = \mu_0$
- If the null is true the z-score satisfies

$$z_{\hat{\mu}} \sim N(0, 1)$$

Testing μ with known variance (5)

- The probability of seeing a standardised difference from μ_0 of $z_{\hat{\mu}}$ or greater, in either direction is

$$\begin{aligned} p &= 1 - \mathbb{P}(-|z_{\hat{\mu}}| < Z < |z_{\hat{\mu}}|) \\ &= \mathbb{P}(Z < -|z_{\hat{\mu}}|) + \mathbb{P}(Z > |z_{\hat{\mu}}|) \end{aligned}$$

  where $Z \sim N(0, 1)$.

  - We ignore the sign as a big difference in either direction (positive or negative) is strong evidence against the null

- By symmetry of the normal, we can write the above as

$$p = 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|)$$

- We call p a p-value. We can calculate it in R using

$$\texttt{pval} = 2 * \texttt{pnorm}(-\texttt{abs}(\texttt{z}))$$

Testing μ with known variance (5)

- The probability of seeing a standardised difference from μ_0 of $z_{\hat{\mu}}$ or greater, in either direction is

$$\begin{aligned} p &= 1 - \mathbb{P}(-|z_{\hat{\mu}}| < Z < |z_{\hat{\mu}}|) \\ &= \mathbb{P}(Z < -|z_{\hat{\mu}}|) + \mathbb{P}(Z > |z_{\hat{\mu}}|) \end{aligned}$$

  where $Z \sim N(0, 1)$.

  - We ignore the sign as a big difference in either direction (positive or negative) is strong evidence against the null

- By symmetry of the normal, we can write the above as

$$p = 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|)$$

- We call p a p-value. We can calculate it in R using

$$\texttt{pval} = 2 * \texttt{pnorm}(-\texttt{abs}(\texttt{z}))$$

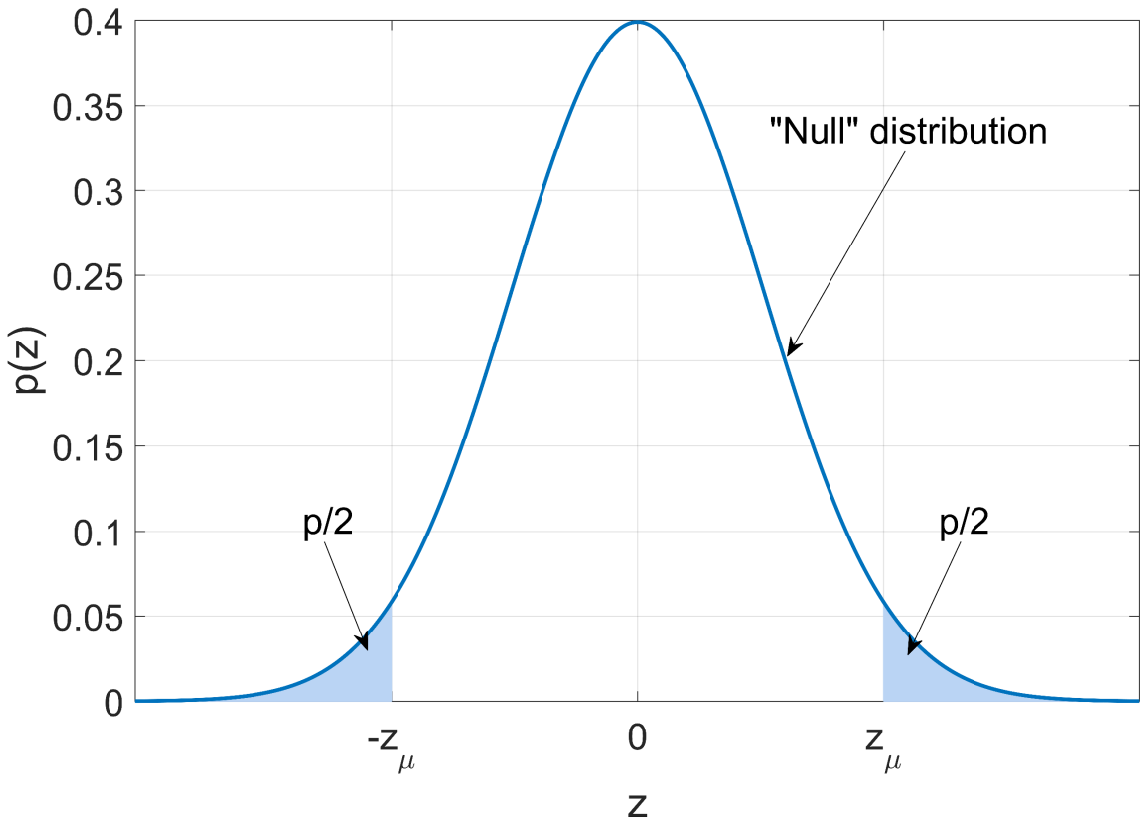
p-values (1)

- So in this case, the p-value is the probability of observing a sample for which the difference between μ_0 and the sample mean $\hat{\mu}$ is greater than $|\mu_0 - \hat{\mu}|$ in either direction, if the null was true.
  - The smaller the p-value, the more improbable such a sample would be
  - A smaller p-value is therefore stronger evidence against the null being true
- We can informally grade the p-value: for
  - $p > 0.05$ we have weak/no evidence against the null;
  - $0.01 < p < 0.05$ we have moderate evidence against the null;
  - $p < 0.01$ we have strong evidence against the null.
- We refer to the quantity that we use to compute our p-value (in this case, a z-score) as a test statistic.

p-values (1)

- So in this case, the p-value is the probability of observing a sample for which the difference between μ_0 and the sample mean $\hat{\mu}$ is greater than $|\mu_0 - \hat{\mu}|$ in either direction, if the null was true.
  - The smaller the p-value, the more improbable such a sample would be
  - A smaller p-value is therefore stronger evidence against the null being true
- We can informally grade the p-value: for
  - $p > 0.05$ we have weak/no evidence against the null;
  - $0.01 < p < 0.05$ we have moderate evidence against the null;
  - $p < 0.01$ we have strong evidence against the null.
- We refer to the quantity that we use to compute our p-value (in this case, a z-score) as a test statistic.

p-values (2)



Null distribution and an observed z-score, $z_{\hat{\mu}}$. The probability in the shaded areas is the probability that $Z \sim N(0, 1)$ would be greater than or less than $|z_{\hat{\mu}}|$ (the p-value). This is the probability of that a sample from the population would result in a standardised difference of $|z_{\hat{\mu}}|$ or greater, *if the null distribution was true*.

Example: Testing if $\mu = \mu_0$ (1)

- For US women aged between 20 to 34 years of age, the population body mass index (BMI) has
  - an approximate mean of $26.8 kg/m^2$; and
  - an approximate standard deviation of $4.5 kg/m^2$.

  (Source: Center for Disease Control)
- We have BMI measured on a sample of women aged 20-34 from the Pima ethnic group, without diabetes:

$$\mathbf{y} = (46.8, 27.8, 32.5, 39.5, 32.8, 31.0, 26.2, 20.8)$$

- Using this data, can we say whether women aged 20-34 in this Pima cohort have the same average BMI as the general US population?

Example: Testing if $\mu = \mu_0$ (2)

- We want to test:
  - $H_0 : \mu = 26.8$ vs $H_A : \mu \neq 26.8$,

  μ is the population mean BMI of Pima women aged 20-34.
- The estimated mean $\hat{\mu}$ from our sample is

$$\hat{\mu} = 32.175$$

- From this we can calculate the z-score as

$$z_{\hat{\mu}} = \frac{32.175 - 26.8}{(4.5/\sqrt{8})} = 3.3784$$

- This yields a p-value of

$$\begin{aligned} 1 - \mathbb{P}(-z_{\hat{\mu}} < Z < z_{\hat{\mu}}) &= 2 * \texttt{pnorm}(-\texttt{abs}(3.3784)) \\ &= 7.29 \times 10^{-4} \end{aligned}$$

Example: Testing if $\mu = \mu_0$ (3)

- How to interpret?
- A p-value of 7.29×10^{-4} can be interpreted as follows:

  If the null was true, i.e., Pima ethnic women aged 20-34 have the same BMI as the average US woman aged 20-34, then the chance of observing a sample with as an extreme, or more extreme, difference from the null as the one that we saw would be less than $1/1371$.

- So quite unlikely to happen just by vagaries of sampling $\Rightarrow$ strong evidence against the null.

One Sided Tests (1)

- Assume our population is normally distributed with known variance σ^2, unknown mean
- Given a sample $y_1, \ldots, y_n$ we want to test

$$H_0 \quad : \quad \mu \leq \mu_0$$

vs

$$H_A \quad : \quad \mu > \mu_0$$

- This is called a one-sided test
- Has a similar solution to the previous example, which is a two-sided test

One Sided Tests (2)

- For this problem, our test statistic is once again the z-score

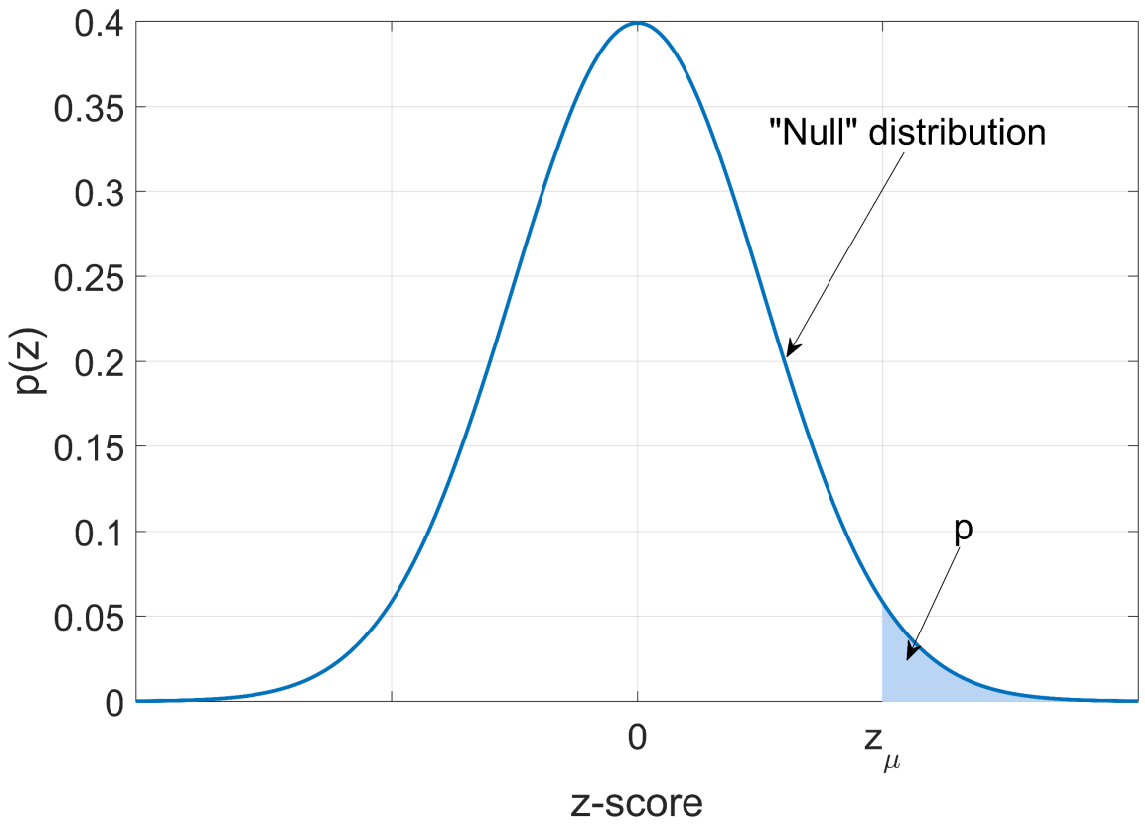
$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})}$$

  where $\hat{\mu}$ is our ML estimate of the mean (equivalent to the sample mean)
- However, this time we treat standardised differences $z_{\hat{\mu}}$ that are large and positive as evidence against the null
- So the p-value is the probability of seeing a z-score *at least* as large as $z_{\hat{\mu}}$, i.e.,

$$p = \mathbb{P}(Z > z_{\hat{\mu}}) = 1 - \mathbb{P}(Z < z_{\hat{\mu}})$$

  where $Z \sim N(0, 1)$ (note we do not take absolute of $z_{\hat{\mu}}$)

One Sided Tests (3)



Null distribution and an observed z-score, $z_{\hat{\mu}}$. The probability in the shaded areas is the probability that $Z \sim N(0,1)$ would be greater than $z_{\hat{\mu}}$ (the p-value for the one-sided test $H_0 : \mu = \mu_0$ vs $H_A : \mu \geq \mu_0$). This is the probability of that a sample from the population would result in a standardised difference of $z_{\hat{\mu}}$ or greater, *if the null distribution was true.*

One Sided Tests (4)

- We can also test

$$H_0 \quad : \quad \mu \geq \mu_0$$

vs

$$H_A \quad : \quad \mu < \mu_0$$

- This time we are treat standardised differences $z_{\hat{\mu}}$ that are large and negative as evidence against the null
- So the p-value is the probability of seeing a z-score *as small as, or smaller than* $z_{\hat{\mu}}$, i.e.,

$$p = \mathbb{P}(Z < z_{\hat{\mu}})$$

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

Example: One Sided Test

- Using our BMI measured on a sample of women aged 20-34 from the Pima ethnic group, without diabetes we can test

$$H_0 : \mu \geq 26.8 \text{ vs } H_A : \mu < 26.8$$

where the population standard deviation $\sigma = 4.5$.

- Recall our z-score was

$$z_{\hat{\mu}} = 3.3784$$

- So our p-value is

$$\begin{aligned} \mathbb{P}(Z < z_{\hat{\mu}}) &= \texttt{pnorm}(3.3784) \\ &\approx 0.9996 \end{aligned}$$

$\Rightarrow$ no evidence against the null

Testing μ with known variance – Key Slide

- Assume population follows normal distribution with unknown mean and known variance σ^2; testing inequality of μ
  1. First calculate the ML estimate of the mean/sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

  2. Then calculate the z-score

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})}$$

  3. Then calculate the p-value:

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases} .$$

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

Understanding hypothesis testing

- A misconception is that a large p-value proves the null is true
- The p-value represents evidence against the null
  $\Rightarrow$ little evidence against the null does not prove it is true
- So for example, if we have:
  - Large estimated differences from null;
  - Small sample size;
  - p-values in the "gray" $0.05 - 0.2$ region

  are inconclusive; it is hard to determine if only reason we did not have stronger evidence was simply because of sample size
- Smaller sample sizes $=$ larger standard errors $=$ smaller standardised differences $z_{\hat{\mu}}$

Testing μ with unknown variance (1)

- Let us now relax the assumption and inequality of the mean

$$H_0 \quad : \quad \mu = \mu_0$$

vs

$$H_A \quad : \quad \mu \neq \mu_0$$

  under the assumption that the population is normal with unknown μ and σ^2

- We estimate the variance using the unbiased estimator

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

- We then use the t-test

Testing μ with unknown variance (2) – Key Slide

- Then our test statistic is a t-score

$$t_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\hat{\sigma}/\sqrt{n})}$$

where the unknown population σ is replaced with our estimate

- If the null was true, then

$$t_{\hat{\mu}} \sim \mathrm{T}(n-1)$$

where $\mathrm{T}(d)$ denotes a standard t-distribution with d degrees-of-freedom

- The p-value is then

$$p = \begin{cases} 2\,\mathbb{P}(T < -|t_{\hat{\mu}}|)) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(T < t_{\hat{\mu}}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(T < t_{\hat{\mu}}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases}.$$

where $T \sim \mathrm{T}(n-1)$.

Testing difference of means, known variances (1)

- Often we are interested in the difference between two samples
- Imagine we have a cohort of people in a medical trial
  - At the start of the trial, all participants' weights are measured and recorded (Sample **x**, population mean μ_x)
  - The participants are then administered a drug targeting weight loss
  - At the end of the trial, everyone's weight is remeasured and recorded (Sample **y**, population mean μ_y)
- To see if the drug had any effect, we can try to estimate the population mean difference in weights pre- and post-trial

$$\mu_x - \mu_y$$

- If no difference at population level, $\mu_x = \mu_y \Rightarrow \mu_x - \mu_y = 0$

Testing difference of means, known variances (2)

- Assume both samples come from normal populations with unknown means μ_x and μ_y and known variances σ_x^2 and σ_y^2
- Formally, we are testing

$$H_0 \quad : \quad \mu_x = \mu_y$$

vs

$$H_A \quad : \quad \mu_x \neq \mu_y$$

- If the populations from which the two samples came have the same mean, their difference will have a mean of zero at the population level

Testing difference of means, known variances (3)

- Estimate the sample means of the two samples:

$$\hat{\mu}_x = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i, \quad \hat{\mu}_y = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$$

  where n_x and n_y are the sizes of the two samples
- Then, under the null distribution the difference follows

$$\hat{\mu}_x - \hat{\mu}_y \sim N\left(0, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

- Our test statistic is the z-score for the difference in means

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Testing difference of means, known variances (4)

- The p-value is then

$$p = 2\,\mathbb{P}\left(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|\right)$$

which tells us the probability of observing a (standardised) difference between the sample means of $|z_{(\hat{\mu}_x - \hat{\mu}_y)}|$ or greater in either direction, if the null was true

- For testing $H_0 : \mu_x \geq \mu_y$ vs $H_A : \mu_x < \mu_y$ we can compute

$$p = \mathbb{P}\left(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}\right)$$

which can also be used to test $\mu_x > \mu_y$ by noting this is the same as $\mu_y < \mu_x$.

Testing difference of means, unknown variances (1)

- If we want to relax the assumption that σ_x^2, σ_y^2 are known the problem becomes trickier
- Assume that $\sigma_x^2 = \sigma_y^2 = \sigma^2$, i.e., unknown but equal
  $\Rightarrow$ Then we can still use a t-test
- Estimate the population variances for each sample

$$\hat{\sigma}_x^2 = \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (x_i - \hat{\mu}_x)^2, \quad \hat{\sigma}_y^2 = \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (y_i - \hat{\mu}_y)^2$$

- The next step is to form a pooled estimate of σ^2:

$$\hat{\sigma}_p^2 = \frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}$$

Testing difference of means, unknown variances (2)

- Our test statistic is then a t-score of the form

$$t_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\hat{\sigma}_p^2(1/n + 1/m)}} \tag{1}$$

  which follows a $\mathrm{T}(n_x + n_y - 2)$ distribution.
- Our p-value is then

$$p = 2\,\mathbb{P}\left(T < -|t_{(\hat{\mu}_x - \hat{\mu}_y)}|\right)$$

  where $T \sim \mathrm{T}(n_x + n_y - 2)$.
- If `tdiff` is a variable containing our t-score (1) then

$$\texttt{p} = 2 * \texttt{pt}(-\texttt{abs}(\texttt{zdiff}))$$

  will give us our p-value.

Testing difference of means, unknown variances (3)

- If we relax assumption that $\sigma_x^2 = \sigma_y^2$ things get hard
- An approximate p-value can be computed by substituting estimates $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ into the formulae for known variance
- This give us the test statistic

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

  which is approximately $N(0, 1)$ for large samples.
- We can then find approximate p-values using:

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(Z < z_{(\hat{\mu}_x - \hat{\mu}_y)}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases} .$$

- More exact but complicated procedures exist; `t.test()` in R implements some of these

Testing Bernoulli populations

- We can also apply hypothesis testing to binary data
- This is an important application as we are often testing if rates of events occuring have been changed, or if they meet certain requirements
- For example, we can imagine a production line making electronic components. They guarantee that the failure rate of components is less than some amount θ_0
- After obtaining a sample and observing a failure rate in that sample, a customer could test to see if the advertised failure rate was achieved

Testing a Bernoulli population (1)

- Assume our population is Bernoulli distributed with success probability θ
- Given a sample, we want to test

$$H_0 \quad : \quad \theta = \theta_0$$

vs

$$H_A \quad : \quad \theta \neq \theta_0$$

- Derive an approximate test based on the central limit theorem
- Recall our estimate of the population success probability is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{m}{n}$$

where m is the number of successes in our data $\mathbf{y}$

Testing a Bernoulli population (2)

- If the null hypothesis was true, then by the CLT

$$\hat{\theta} - \theta_0 \xrightarrow{d} N\left(0, \frac{\theta_0(1-\theta_0)}{n}\right)$$

- Our test statistic is then the approximate z-score

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$$

- We can then calculate two or one-sided approximate p-values

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{\hat{\theta}}|) & \text{if } H_0 : \theta = \theta_0 \text{ vs } H_A : \theta \neq \theta_0 \\ 1 - \mathbb{P}(Z < z_{\hat{\theta}}) & \text{if } H_0 : \theta \leq \theta_0 \text{ vs } H_A : \theta > \theta_0 \\ \mathbb{P}(Z < z_{\hat{\theta}}) & \text{if } H_0 : \theta \geq \theta_0 \text{ vs } H_A : \theta < \theta_0 \end{cases} .$$

where $Z \sim N(0,1)$.

Testing two Bernoulli populations (1)

- Now consider testing equality of two Bernoulli populations
- Given two samples $\mathbf{x}$ and $\mathbf{y}$ of binary data, test

$$H_0 \quad : \quad \theta_x = \theta_y$$

vs

$$H_A \quad : \quad \theta_x \neq \theta_y$$

where θ_x, θ_y are the population success probabilities

- Under the null hypothesis, $\theta_x = \theta_y = \theta$
- We use a pooled estimate of θ

$$\hat{\theta}_p = \frac{m_x + m_y}{n_x + n_y}$$

where m_x, m_y are the number of successes in the two samples, and n_x, n_y is the total number of trials

Testing two Bernoulli populations (2)

- In this case our test statistic is

$$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(1/n_x + 1/n_y)}}$$

  which approximately follows an $N(0,1)$ if the null is true

- We can then get approximate p-values using

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\theta}_x - \hat{\theta}_y)}|) & \text{if } H_0 : \theta = \theta_0 \text{ vs } H_A : \theta \neq \theta_0 \\ 1 - \mathbb{P}(Z < z_{(\hat{\theta}_x - \hat{\theta}_y)}) & \text{if } H_0 : \theta \leq \theta_0 \text{ vs } H_A : \theta > \theta_0 \\ \mathbb{P}(Z < z_{(\hat{\theta}_x - \hat{\theta}_y)}) & \text{if } H_0 : \theta \geq \theta_0 \text{ vs } H_A : \theta < \theta_0 \end{cases}.$$

Testing Bernoulli populations

- There exist more exact methods for computing p-values when testing Bernoulli populations
- They make use of properties of the Binomial distribution
- In R:
  - `binom.test()` can be used to test a single Bernoulli sample
  - `prop.test()` can be used to test difference in Bernoulli samples
- See Ross (Chapter 8) for more details on these.

Example: Testing a Bernoulli Population

- Imagine we run a survey asking $n = 60$ people whether they prefer holidaying in France or Spain
  - $m = 37$ people preferred France, so $\hat{\theta} = 37/60 \approx 0.6166$
  - Is there a real preference for France ($\theta \neq \frac{1}{2}$) or is this just random chance ($\theta = \frac{1}{2}$)?
- The approximate z-score is

$$z_{\hat{\theta}} = \frac{(37/60) - 1/2}{\sqrt{(1/2)(1-1/2)/60}} \approx 1.807$$

giving an approximate p-value of

$$2\,\mathbb{P}(Z < -1.807) = 2 * \texttt{pnorm}(-1.807) \approx 0.0707$$

- Exact p-value: `binom.test(x=37,n=60,p=0.5)`$= 0.0924$

Outline

Decision making (1)

- So far we have computed p-values as evidence against the null
- What if we are asked to make a decision regarding our hypothesis?
- We could decide that if the evidence was sufficiently strong, we could reject the null hypothesis.
- For example, we could say that if we see a sample that has probability α or less of arising by chance
  if the null distribution was true, then the evidence is strong enough to reject the null

Decision making (2)

- Formally, we reject the null hypothesis at a significance level of α if we reject the null when $p < \alpha$
- Sometimes people say the result is "statistically significant"
- A common convention to take $\alpha = 0.05$; why?
- Remember, we cannot prove the null; only accumulate evidence against it
- Is this procedure any good? What properties does it have?

Type I and II Errors (1)

- Rejecting the null when $p < \alpha$ implies we reject the null if the sample we observe resulted in a test statistic that has probability $\leq \alpha$ of occurring by chance, if the null was true
- If we reject the null when it is true, we erroneously reject it
- Erroneously rejecting the null is called a "false discovery", a "false positive" or a Type I error
- We make a false discovery $100\alpha\%$ of the time
  $\Rightarrow$ we control the Type I error rate at α

Type I and II Errors (2)

- So if we make α very small we will have very small probability of making a false discovery
- Why not set $\alpha = 0$ then?
- Consider the case when the null is not true; i.e., the alternative is true
- Erroneously accepting the null when the alternative is true is called a "false negative", or a Type II error
- The smaller the threshold of rejection α, the stronger the evidence is needed to reject the null
  $\Rightarrow$ increases the Type II error rate, which we call β

Type I and II Errors (2)

- So if we make α very small we will have very small probability of making a false discovery
- Why not set $\alpha = 0$ then?
- Consider the case when the null is not true; i.e., the alternative is true
- Erroneously accepting the null when the alternative is true is called a "false negative", or a Type II error
- The smaller the threshold of rejection α, the stronger the evidence is needed to reject the null
  $\Rightarrow$ increases the Type II error rate, which we call β

Type I and II Errors (3)

- In statistics it is more common to talk about the power
- This is the probability that a test will correctly reject the null
  $\Rightarrow$ i.e., if the alternative is true and we reject the null
- The power is $1 - \beta$. It clearly depends on α
  $\Rightarrow$ the smaller the α, the smaller the power $1 - \beta$
- It also depends on the underlying population parameters

Type I and II Errors (4)

		Null hypothesis (H0) is	
		Valid (True)	**Invalid (False)**
Judgment:	**Reject**	Type I error *(False positive)*	Correct *(True positive)*
	Do not reject (accept)	Correct *(True negative)*	Type II error *(False negative)*

Power (1)

- We demonstrate how power depends on the population
- Consider testing the means of two normal populatons μ_x and μ_y from two samples with known variances σ_x^2 and σ_y^2
- The test statistic we use is the z-score

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

- The difference at the population level

$$\mu_x - \mu_y$$

is sometimes called the effect size.

Power (2)

- Under our assumption

$$\hat{\mu}_x \sim N\left(\mu_x, \frac{\sigma^2}{n_x}\right), \quad \hat{\mu}_y \sim N\left(\mu_y, \frac{\sigma^2}{n_y}\right)$$

so that the z-score for the difference in sample means follows

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} \sim N\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}, 1\right)$$

under repeated sampling from our population.

Power (3)

- The quantity

$$\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$$

  can be viewed as the *standardised* effect size
- It is the difference in means relative to the variability in the estimates
- The absolute value of this increases with ...
  - increasing population effect size $|\mu_x - \mu_y|$
  - increasing sample sizes n_x, n_y
  - decreasing population variances σ_x^2, σ_y^2
- Bigger standardised effect sizes mean we would see larger (absolute) z-scores on average if we repeatedly sampled from our populutions and computed $z_{\hat{\mu}_x - \hat{\mu}_y}$

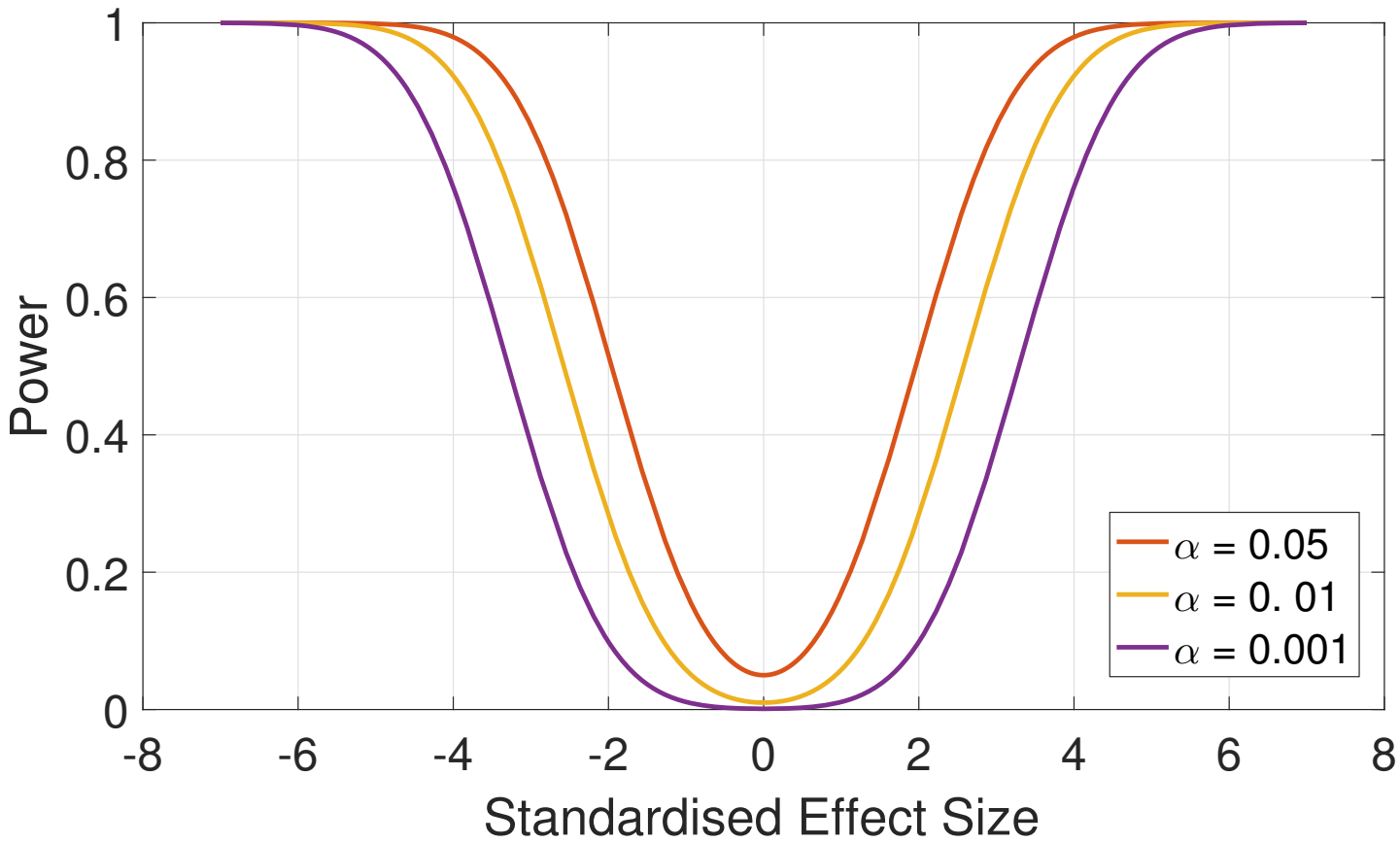
Power (4)

- Recall the p-value is given by

$$p = 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_x - \hat{\mu}_y)}|)$$

  where $Z \sim N(0, 1)$.
- Large values of $|z_{(\hat{\mu}_x - \hat{\mu}_y)}|$ unlikely under the null ($\mu_x = \mu_y$)
  $\Rightarrow$ they lead to small p-values
- The bigger the standardised effect size, the smaller the average p-value obtained under repeated sampling
  $\Rightarrow$ greater chance to correctly reject the null (true positive!)
- So the power increases with increasing effect size

Power (5)



Plot of power (probability of correctly rejecting the null hypothesis) as a function of standardised effect size. Note that when the standardised effect size is zero, the null is true and the plot shows the probability of a false discovery (which is equal to α)

Power and Significance – Key Slide

- These ideas behind power apply to all tests, but the definition of the "effect size" may vary
- In summary, assume we test a hypothesis by rejecting the null if the p-value satisfies $p < \alpha$; then
  - The probability of making a false discovery is α
  - The power $1 - \beta$ (probability of making a true discovery) is:
    - smaller for smaller significance level α;
    - greater for larger population effect size;
    - greater for larger sample sizes.

Reading/Terms to Revise

- Reading for this week: Chapter 8 of Ross.
- Terms you should know:
  - Hypothesis test;
  - p-value;
  - One sided and two sided test;
  - Tests of means of normal populations
  - Tests for Bernoulli populations
  - Significance level, Type I error;
  - Power, Type II error;
- Next week we will cover linear regression.