Unit Schedule: Modules

Module	Week	Content	Ross
1.	1	introduction to modelling	1,2
2.	2	probability refresher	3
	3	random vars & expected values	4
	4	special distributions	5
3.	5	**statistical inference**	6&7
	6	**confidence intervals**	7
	7	**hypothesis testing**	8
4.	8	dependence & linear regression	9
	9	classification, clustering & mixtures	
5.	10	random numbers & simulation	15(bits)
	11	basic machine learning	
6.	12	modelling, validation and review	

Revision at *https://flux.qa/NYHZTZ*

FIT5197 Modelling for Data Analysis

Module 3

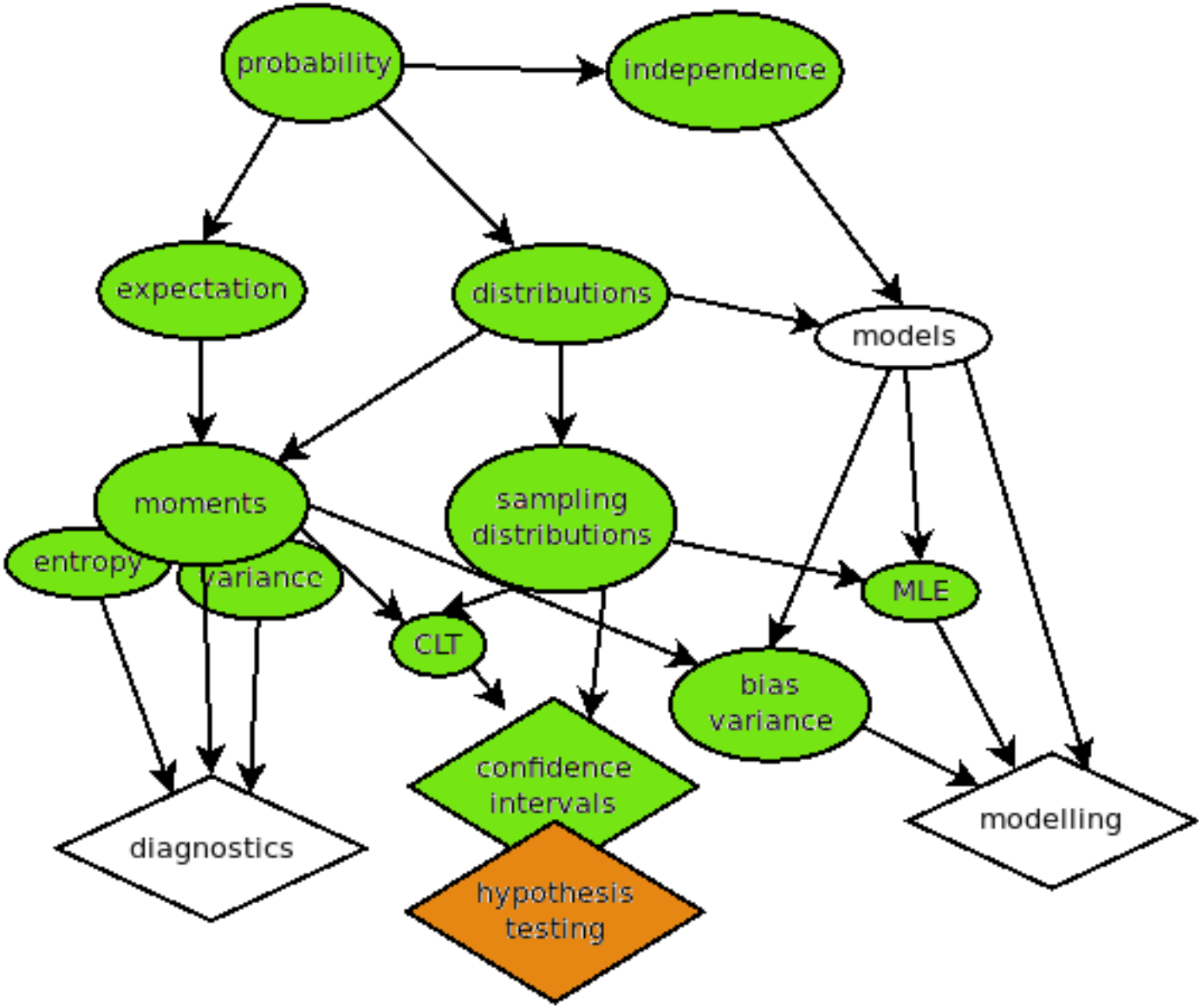
Hypothesis Testing

2019 Lecture 7

Monash University

Outline

Concept Map


probability
independence
expectation
distributions
models
moments
sampling distributions
entropy
variance
MLE
CLT
bias variance
confidence intervals
modelling
diagnostics
hypothesis testing

What is a Confidence Interval?

Have sample of size n with data $X_1, ..., X_n$ and wish to estimate parameter θ using $\hat{\theta} = f(X_1, ..., X_n)$.

What is a confidence interval (CI) with confidence $100(1-\alpha)\%$?

The parameter values θ inside the confidence interval have the property that the parameter estimate $f(X_1, ..., X_n)$ could reasonably have been generated by θ. In fact $f(X_1, ..., X_n)$ lies in a most probable interval, which has probability $(1-\alpha)$, for the parameter estimate $\hat{\theta}$ given θ.

This tells you about the probability of sampling $\hat{\theta}$ given (or assuming the truth of) θ.

What a CI is not?

(optional)

- A confidence interval says **nothing** about the probability of θ given the data.
- A fully probabilistic question is: What is the probability of θ given the data $X_1, ..., X_n$?
  - That is, what is $p(\theta|X_1, ..., X_n)$?
- A fully specified model θ gives us the **sample likelihood** $p(X_1, ..., X_n|\theta)$.
- But what is the **prior** $p(\theta)$?

Example: let the model be Gaussian with unknown mean μ and known standard deviation of 1, so what is our prior on μ, $p(\mu)$?

- hypthesis testing is supposed to be objective, so generally people don't want to specify priors

- a Bayesian/probabilistic version of a confidence interval is the *credible interval*

What is a Confidence Interval?

(optional)

Have sample of size n with points $X_1, ..., X_n$ and wish to estimate parameter θ using $\hat{\theta} = f(X_1, ..., X_n)$.

To construct a confidence interval by first principles:

1. construct the distribution for samples $X_1, ..., X_n$
2. infer the distribution for estimates $\hat{\theta}$ which will be in terms of θ, called the samping distribution
3. for a given value of θ get a most probable interval with total probability $(1 - \alpha)$ for $\hat{\theta}$ given θ, say $(\theta_L(\theta), \theta_U(\theta))$
4. turn this around, infer the confidence interval on θ, say $\left(\hat{\theta}_L(X_1, ..., X_n), \hat{\theta}_U(X_1, ..., X_n)\right)$ such that

$$\begin{aligned} & \theta_L(\theta) < f(X_1, ..., X_n) < \theta_U(\theta) \\ \equiv \quad & \hat{\theta}_L(X_1, ..., X_n) < \theta < \hat{\theta}_U(X_1, ..., X_n)) \end{aligned}$$

Two-sided CIs Gaussian

Assume dataset of count n with mean $\bar{X}$ and sample variance S^2.

Assumptions	Parameter	Interval
Gaussian, σ^2 known	μ	$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Gaussian, σ^2 unknown	μ	$\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$
Gaussian, μ unknown	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$

NB. you need to look up Z, t and χ^2 tables

Two-sided CIs 2 Gaussians

Assume dataset of count n with mean $\bar{X}$ and sample variance S^2. Also, second dataset of count m with mean $\bar{Y}$ and sample variance T^2.

Assumptions	Parameter	Interval
Gaussian, σ_1^2, σ_2^2 known	$\mu_1 - \mu_2$	$\bar{X} - \bar{Y} \pm Z_{\alpha/2}\sqrt{\sigma_1^2/n + \sigma_2^2/m}$
Gaussian, $\sigma_1^2 = \sigma_2^2$ unknown but equal	$\mu_1 - \mu_2$	$\bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\sqrt{\frac{1}{n} + \frac{1}{m}}S_P$ for $S_P^2 = \frac{(n-1)S^2+(m-1)T^2}{n+m-2}$
Gaussian, $\sigma_1^2 \neq \sigma_2^2$ unknown, using CLT	$\mu_1 - \mu_2$	use 1st case for $\sigma_1^2 = S^2$, $\sigma_2^2 = T^2$, assuming n, m are large

NB. you need to look up Z, t and χ^2 tables

Two-sided CIs Using CLT

For Poisson, assume dataset of count n with mean $\hat{X}$. For Bernoulli, assume dataset of count n with mean $\hat{p}$. Also a 2nd dataset of count m with mean $\hat{q}$.

Assumptions	Parameter	Interval
Poisson, λ unknown, using CLT	λ	$\hat{X} \pm Z_{\alpha/2}\sqrt{\hat{X}/n}$
Bernoulli, θ unknown, using CLT	θ	$\hat{p} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$
Bernoulli, θ_1, θ_2 unknown, using CLT	$\theta_1 - \theta_2$	$\hat{p} - \hat{q} \pm Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + \hat{q}(1-\hat{q})/m}$

NB. you need to look up for Z, also hope $\hat{p}, \hat{q} \gg 0$ and $\ll 1$ so normal approximation to binomial works

One-sided CIs

1. pick the left or right bound of the corresponding two-sided CI, as required
2. apply using 2α instead of α

Example: for Gaussian, σ^2 unknown, estimating μ with one-sided on the right, use

$$\bar{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}$$

Outline

Using Hypothesis Testing

Scientific Questions: have a well-defined Boolean question

(I) does drug A cause side-effect B?

(II) does Algorithm A produce better predictions than Algorithm B?

Take Measurements: take suitable measurements

(I) give some patients Drug A, give others a placebo, and measure side-effect B

(II) run both Algorithm A and Algorithm B on test data and record accuracy of predictions

Hypothesis Testing: (I) test equality of proportions in two Bernoulli populations

(II) test for difference between two Gaussian means

generally, we need to know available recipes and choose an appropriate one for our comparison task

Modelling data

- 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 Context

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

$$H_0 \quad : \quad \text{Null hypothesis}$$
$$\text{vs}$$
$$H_A \quad : \quad \text{Alternative hypothesis}$$

  on the basis of our observed data **y**
- What does this mean?

See *what is a null hypothesis?*

Null Hypothesis

- 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?"

Why the Null Hypothesis?

- The null hypothesis must be a fully specified generative model
- So it provides a known model from which the sample likelihood can be evaluated.
- Examples:
  - model family unknown mean μ and known variance σ^2, and $H_0 \ : \ \mu = \mu_0$
  - model family unknown mean μ and unknown variance σ^2, and $H_0 \ : \ \mu = \mu_0, \sigma^2 = \sigma_0^2$
- Counter-examples:
  - model family unknown mean μ and unknown variance σ^2, and $H_0 \ : \ \mu = \mu_0$
    - cannot generate data because do not know σ^2

Hypothesis Testing Example

- 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?

Neyman-Pearson

- 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 **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

- 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:

$$\begin{aligned} H_0 &: \quad \mu = \mu_0 \\ &\text{vs} \\ H_A &: \quad \mu \neq \mu_0 \end{aligned}$$

- 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

- 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

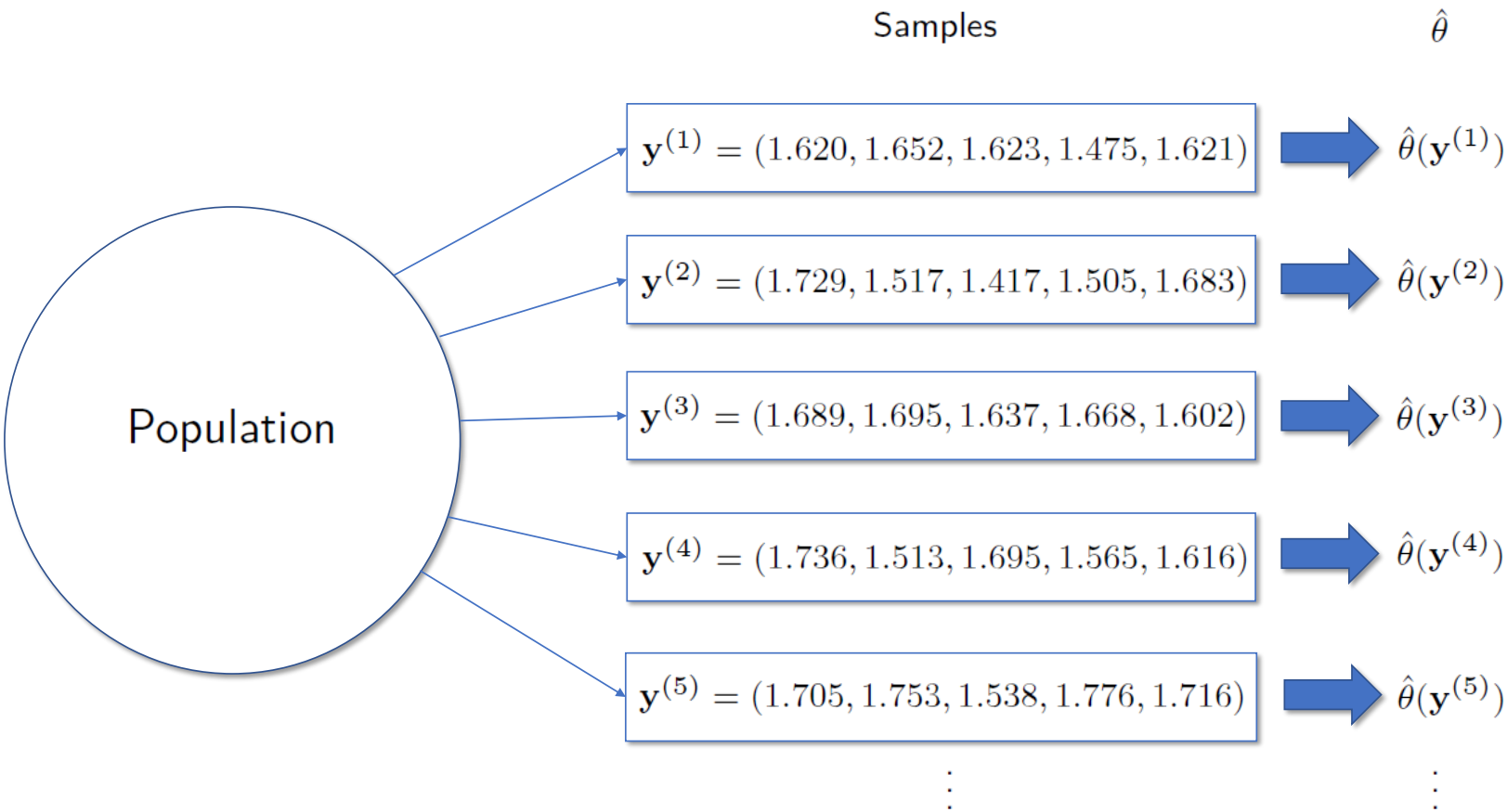


Figure: An (infinite) number of different random samples can be drawn from a population. Each sample would lead to a potentially different estimate $\hat{\theta}$ of the population parameter θ. The distribution of these estimates is called the sampling distribution of $\hat{\theta}$.

Sampling Distribution: Mean

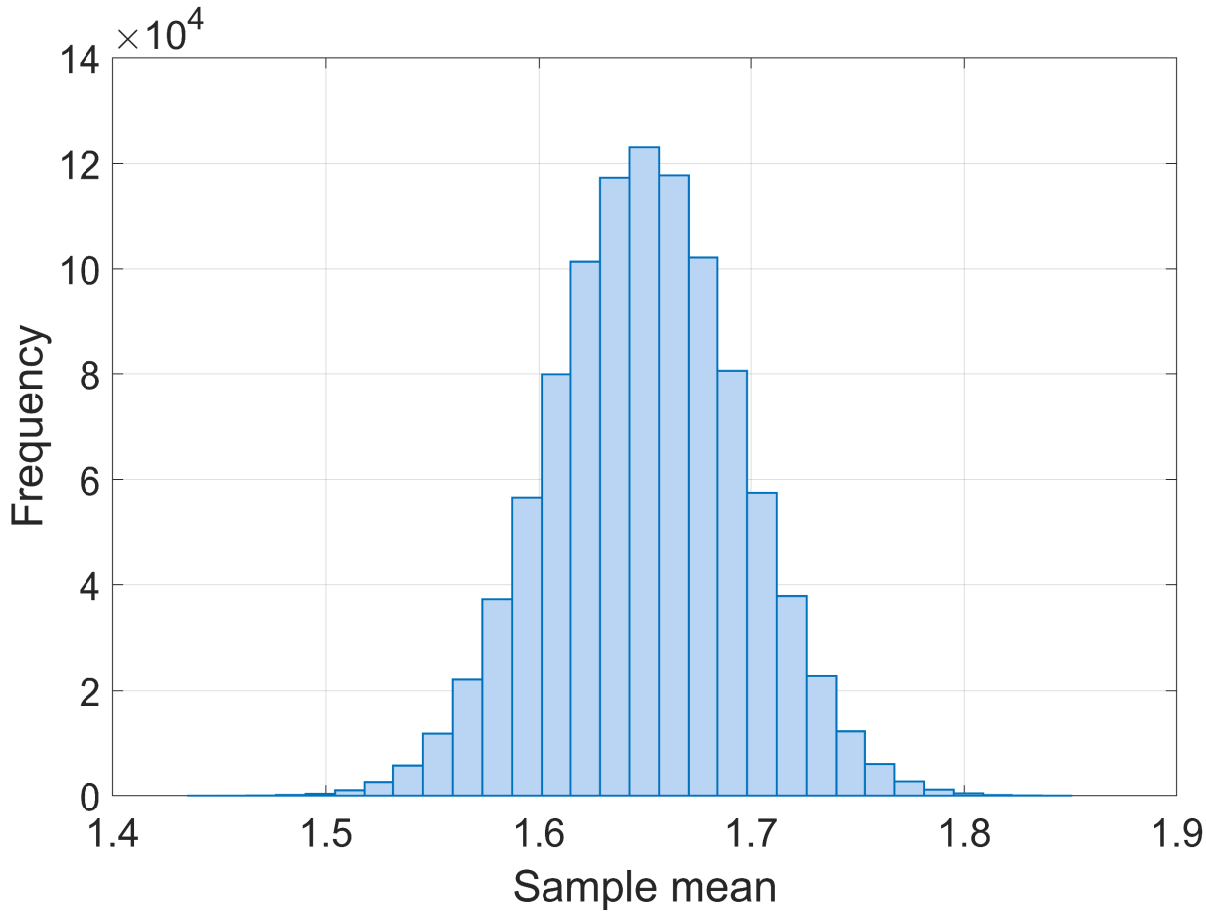


Figure: Histogram of sample means of $1,000,000$ different data samples, each of size $n = 5$, generated from a $N(\mu = 1.65, \sigma = 0.1)$ distribution.

Testing μ with known variance

- 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

- 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

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

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

  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

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

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

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

p-values: Plot

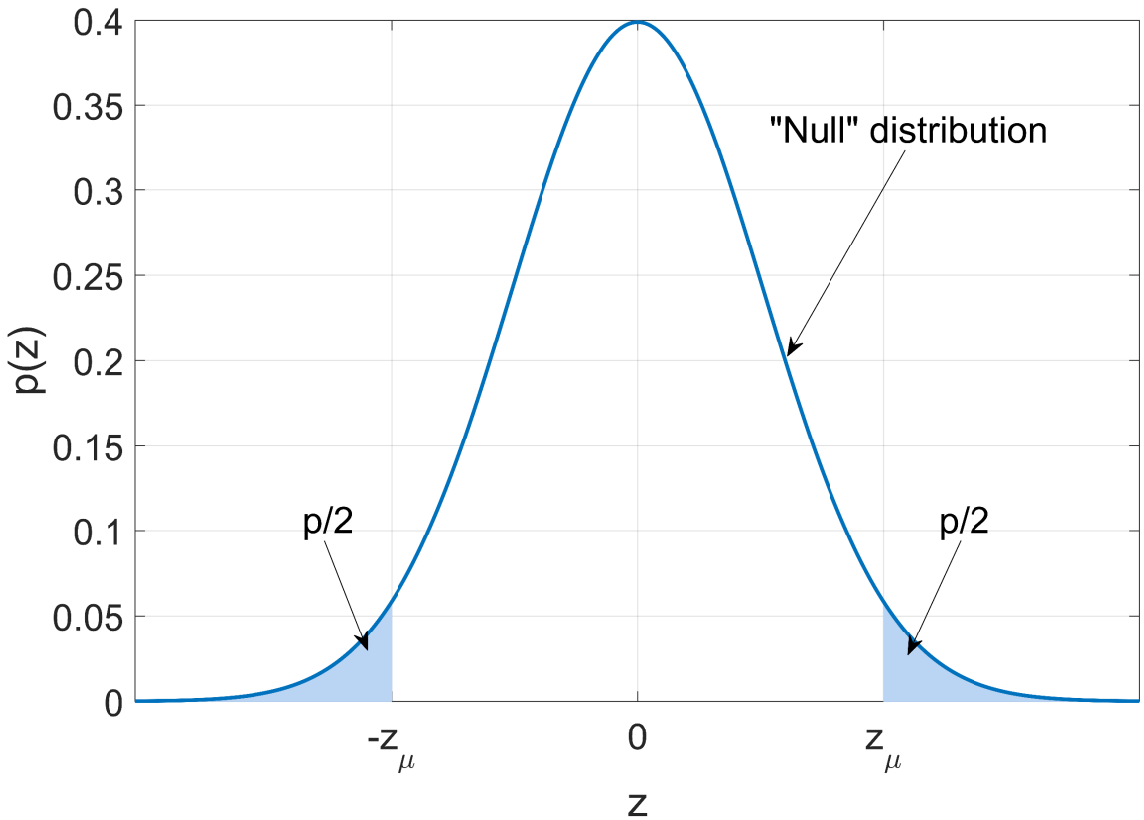


Figure: Null distribution and an observed z-score, $z_{\hat{\mu}}$. The probability in the shaded areas is the probability that $Z \sim N(0, 1)$ would be greater than or less than $|z_{\hat{\mu}}|$ (the p-value). This is the probability of that a sample from the population would result in a standardised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

p-values

- 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.

Example: Testing if $\mu = \mu_0$

- 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 $\mu = \mu_0$

- 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

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

Example: Interpretation

- 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 $\Longrightarrow$ strong evidence against the null.

Outline

One Sided Tests

- 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

- 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.,

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

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

One Sided Tests: Plot

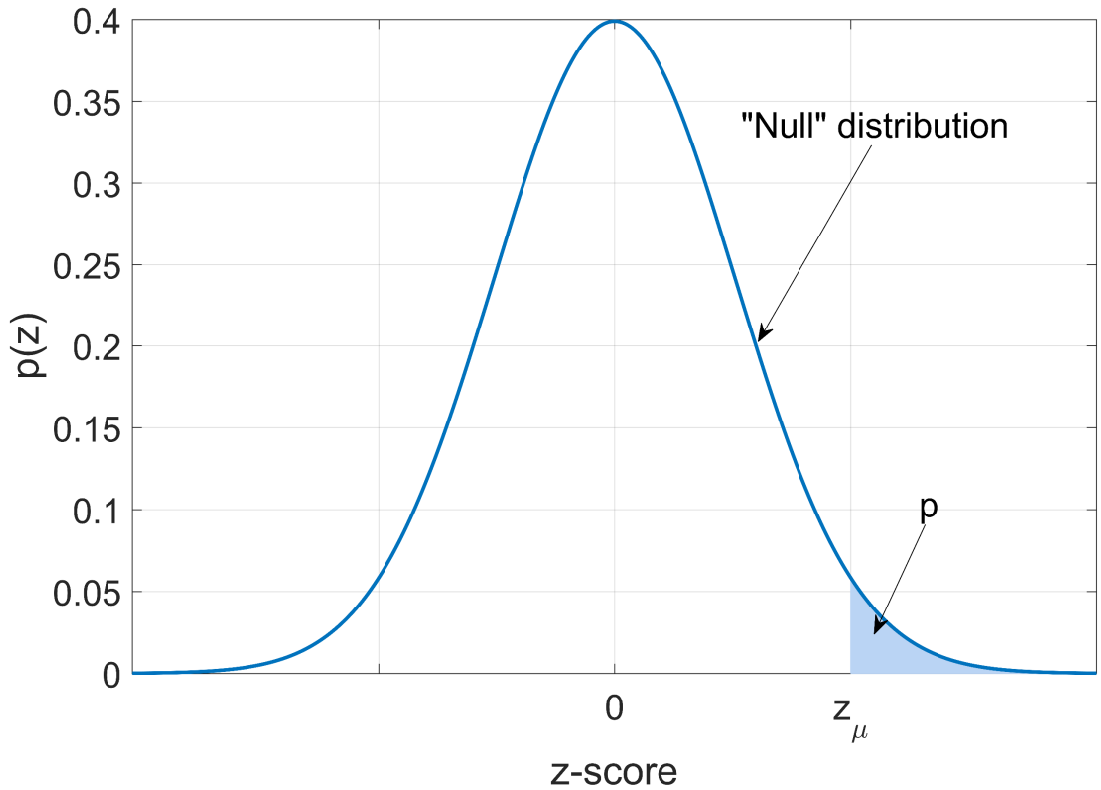


Figure: Null distribution and an observed z-score, $z_{\hat{\mu}}$. The probability in the shaded areas is the probability that $Z \sim N(0, 1)$ would be greater than $z_{\hat{\mu}}$ (the p-value for the one-sided test $H_0 : \mu < \mu_0$ vs $H_A : \mu \geq \mu_0$). This is the probability of that a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

Another One Sided Test

- 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.,

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

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

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

$\Longrightarrow$ no evidence against the null

Testing μ with known variance

- 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:

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

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

Understanding it all

- A misconception is that a large *p*-value proves the null is true
- The *p*-value represents evidence against the null $\Longrightarrow$ 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 σ^2

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

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ &\text{vs} \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

  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 σ^2

- 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

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

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

Testing difference of means

- 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

- 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

$$\begin{aligned} H_0 &: \mu_x = \mu_y \\ &\text{vs} \\ H_A &: \mu_x \neq \mu_y \end{aligned}$$

- 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

- 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

- The p-value is then

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

  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

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

  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 (2)

- 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**
  $\Longrightarrow$ 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 (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

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

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

$$\mathtt{p} = 2 * \mathtt{pt}(-\mathtt{abs}(\mathtt{tdiff}), \mathtt{n_x} + \mathtt{n_y} - 2)$$

  will give us our p-value.

Testing difference of means (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:

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

- 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

- 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

- 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

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

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

Testing two Bernoulli popul.s

- Now consider testing equality of two Bernoulli populations
- Given two samples **x** and **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 popul.s

- 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

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

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: Bernoulli

- 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

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

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

Outline

Decision making

- 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, cont.

- 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

- 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
  $\Longrightarrow$ we **control** the Type I error rate at α

Type I versus Type II Errors

- 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
  $\Longrightarrow$ 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
  $\Longrightarrow$ i.e., if the alternative is true and we reject the null
- The power is $1 - \beta$. It clearly depends on α
  $\Longrightarrow$ the smaller the α, the smaller the power $1 - \beta$
- It also depends on the underlying population parameters, or "truth," so difficult to evaluate

Type I and II Errors: Table

		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)*

Outline

P-Hacking

Methods for defeating hypothesis testing:

Weak significance: make α low, so 0.05, and you're guaranteed to defeat it 1/20 times

Repeated testing: repeat the test $1/\alpha$ times and one of the tests is likely to appear significant

Repeated hypothesis: try testing $1/\alpha$ different hypotheses, and one of the tests is likely to appear significant.

Due to the rewards for getting positive results in testing, this happens more often than we would like!

See *The problem with p values*