FIT2086 Lecture 5
Hypothesis Testing and Model Selection

Daniel F. Schmidt

Faculty of Information Technology, Monash University

August 26, 2019

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 $n-1$

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 often 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

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

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:

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

- 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)

- Recall Georg Ohm's famous law relating current (I), voltage (V) and resistance (R):

$$V = IR$$

- Imagine we want to perform an experiment to test this law
  - Given a certain resistance and current, measure voltage
  - Compare against the voltage predicted by Ohm's law
- We set $R = 100\Omega$ and $I = 0.01\mathrm{A}$
  - Ohm's law predicts $V = 100 \cdot 0.01 = 1\mathrm{V}$
- We go to lab, select 100Ω resistor and set current to 0.01 amps
  - Using voltmeter we obtain a reading of 0.956 volts
- Due to measurement error and tolerances (up to 20% variation for cheap resistors) we would not expect to see exactly 1 volt
  - Is the difference between measured and predicted big enough to be incompatible with Ohm's law?

Hypothesis Testing (4)

- We use the Neyman-Pearson framework to answer questions like this
- We are interested in the evidence against the null hypothesis contained in our data
- 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:

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

- 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 (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

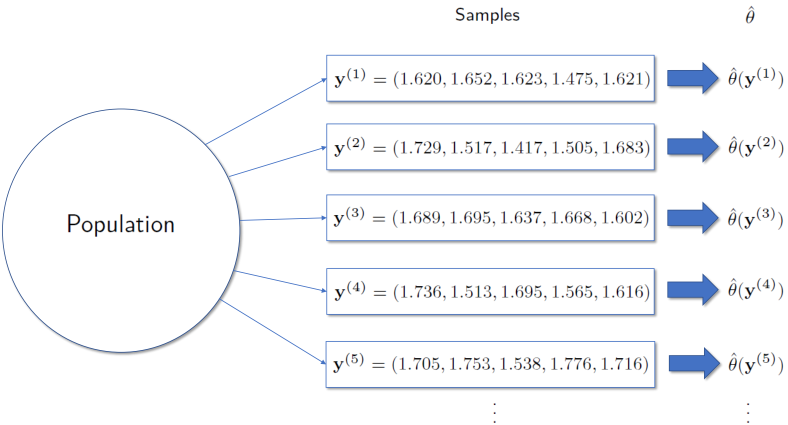
$$\hat{\mu} \equiv \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

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

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

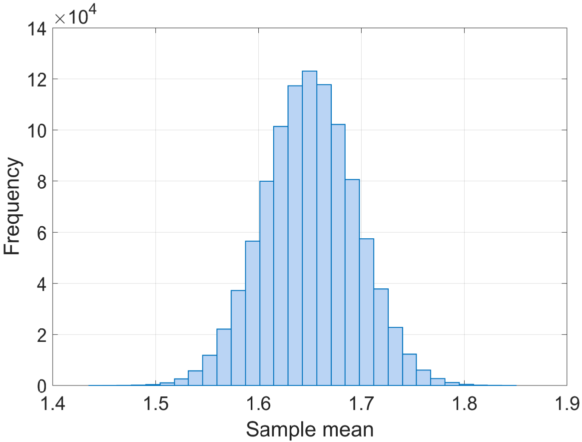
- This is the distribution of the sample mean $\hat{\mu}$ if we repeatedly took samples of size n from our population

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 (or is inconsistent) with our null hypothesis assumptions
- How to determine how likely it would be to see a difference/degree of inconsistency 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

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

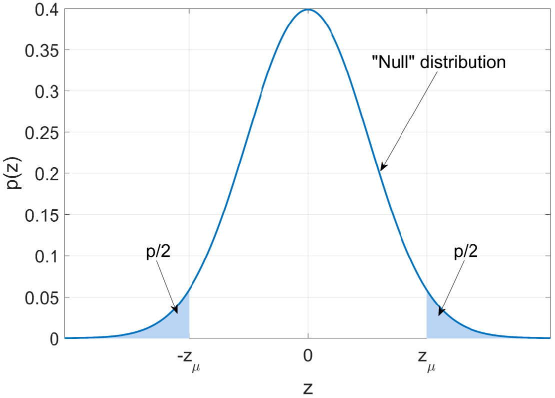
- We call p a p-value. We can calculate it in R using

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

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*.

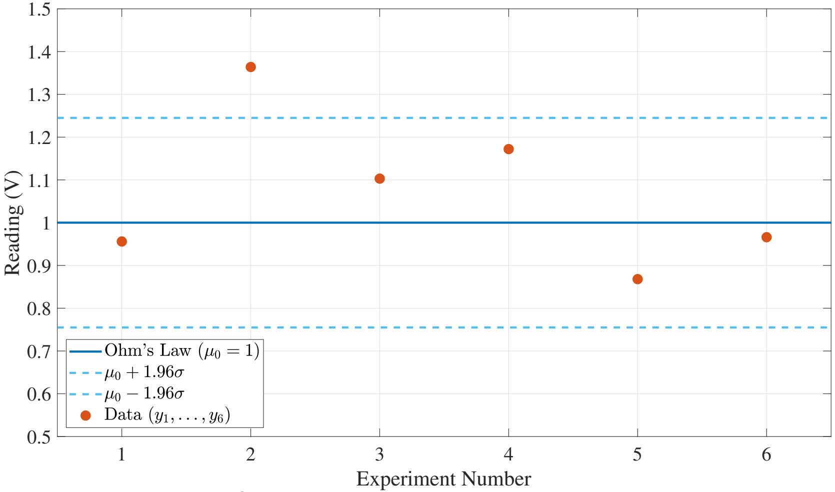
Example: Testing if $\mu = \mu_0$ (1)

- Let us return to our Ohm's law example
- After factoring component tolerances and measurement accuracy specifications we estimate the error to be around $+/-0.25\text{V}$
  - This is approximately $\sigma = 0.125$ (remember the $+/-1.96\sigma$ rule)
- We decide to repeat the experiment $n = 6$ times
  - Nominal resistance of 100Ω, nominal current of 0.01A
  - Ohm's law predicts $V = IR = 100 \cdot 0.01 = 1\text{V}$
- We obtain the readings

$$\mathbf{y} = (0.956, 1.364, 1.103, 1.172, 0.868, 0.966)\text{V}$$

- By using this data, can we test whether Ohm's law seems to hold?

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



Voltage as predicted by Ohm's law and experimental data. Notice that the data appears quite consistent with the null distribution ($\mu_0 = 1$, $\sigma = 0.125$)

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

- We want to test:
  - $H_0 : \mu = 1$ vs $H_A : \mu \neq 1$,

  μ is the population mean voltage of our readings
- The estimated mean $\hat{\mu}$ from our sample is

$$\hat{\mu} = 1.0715$$

- From this we can calculate the z-score as

$$z_{\hat{\mu}} = \frac{1.0715 - 1}{(0.125/\sqrt{6})} = 1.4011$$

- This yields a p-value of

$$\begin{aligned} 1 - \mathbb{P}(-z_{\hat{\mu}} < Z < z_{\hat{\mu}}) &= 2 * \texttt{pnorm}(-\texttt{abs}(1.4011)) \\ &= 0.1612 \end{aligned}$$

Example: Testing if $\mu = \mu_0$ (4)

- How to interpret?
- A p-value of 0.1612 can be interpreted as follows:
  If the null was true, i.e., the voltage for this experiment is predicted by Ohm's law, 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 around $1/6.2$.
- That is, we expect, if Ohm's law is true, about 1 in 6 experiments of $n = 6$ readings would yield a sample mean $\hat{\mu}$ that differed from $\mu_0 = 1$ by 0.0715 or greater, in either direction, just by chance
- 1 in 6 events are pretty common
  $\Rightarrow$ very weak evidence against the null.
- It appears our experiment data is not incompatible with Ohm's law
  - For this particular experimental setup

Example: Testing if $\mu = \mu_0$ (5)

- What if I propose my own law?
- "Schmidt's law" predicts voltage to be

$$V = 1.5IR$$

  so that in our experimental setup we would predict $V = 1.5\text{V}$
- From this we can calculate the z-score as

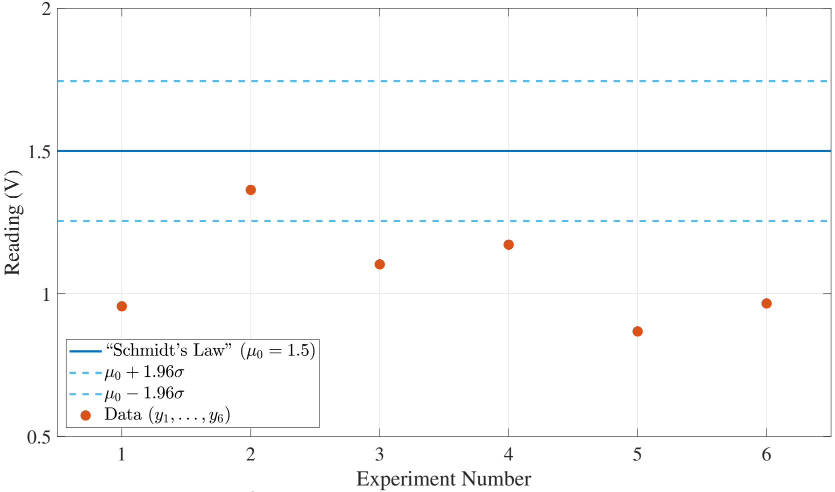
$$z_{\hat{\mu}} = \frac{1.0715 - 1.5}{(0.125/\sqrt{6})} \approx -8.36$$

- This yields a p-value of

$$\begin{aligned} 1 - \mathbb{P}(-z_{\hat{\mu}} < Z < z_{\hat{\mu}}) &= 2 * \texttt{pnorm}(-\texttt{abs}(8.36)) \\ &\approx 4.5 \times 10^{-17} \end{aligned}$$

- Very strong evidence against "Schmidt's law"!

Example: Testing if $\mu = \mu_0$ (6)



Voltage as predicted by "Schmidt's law" and experimental data. Notice that the data appears to be quite inconsistent with the null distribution ($\mu_0 = 1.5$, $\sigma = 0.125$)

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

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ &\text{vs} \\ H_A &: \mu > \mu_0 \end{aligned}$$

- 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

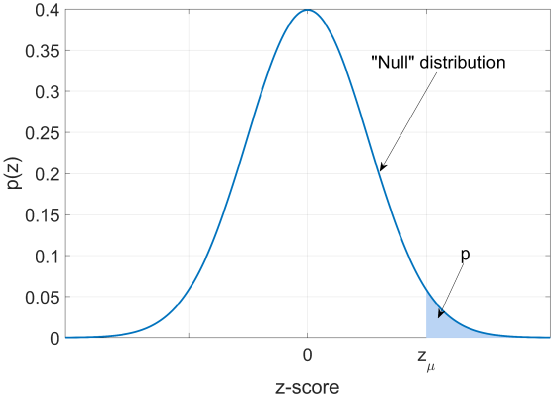
$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})}$$

  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

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

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

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 assumptions and test inequality of the mean

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

  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 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

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

- 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, 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

$$\texttt{p} = 2 * \texttt{pt}(-\texttt{abs}(\texttt{zdiff}), \texttt{nx} + \texttt{ny} - 2)$$

  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:

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

- 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

$$\begin{aligned} H_0 &: \quad \theta = \theta_0 \\ &\text{vs} \\ H_A &: \quad \theta \neq \theta_0 \end{aligned}$$

- 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

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

  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

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

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)` $= 0.0924$

Statistical Significance

- 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.
- We would set a threshold α
  - If the p-value is less than α, we could say it is "statistically significant"
  - A typical value might is $\alpha = 0.05$
- If we did this, then we would accept that $100\alpha\%$ of the time, we would erroneously reject the null
  - p-value is the probability of seeing as extreme, or more extreme inconsistency with the null just by chance even if the null was true
- This approach is becoming more and more deprecated
  - Instead, encouraged to report and consider evidence

Why we can't "prove the null" (1)

- A large p-value does not provide evidence for the null
  - We can only use small p-values as evidence against the null
- To see why, recall our Ohm's law experiment
- Imagine there was a competing law

$$V = \frac{R^2 \sqrt{I}}{1000}$$

- Different from Ohm's law (and clearly wrong ...)
  - For $R = 100$ and $I = 0.01$ both give same prediction of $V = 1\mathrm{V}$
- Therefore, if we only did experiment for $R = 100$ and $I = 0.01$ we would get large p-value as before and not reject this new law

Why we can't "prove the null" (2)

- But if we tried another experiment, say $R = 200$ and $I = 0.01$:
  - Ohm's law predicts $\mu_0 = 2\mathrm{V}$
  - This new law predicts $\mu_0 = 4\mathrm{V}$
- If we collected data we would find:
  - A large p-value for Ohm's law
  - A small p-value for this new law; so we could reject it
- But still can't accept Ohm's law as there could conceivably be a setup for which it fails to work ("is falsified")
- Consider Newtonian phsyics vs Relativity
  - For wide range of setups, Newtonian laws consistent with nature
  - For extreme setups, Netwonian laws inconsistent

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
- Next week we will cover linear regression.