Hypothesis Testing

A Demonstration of the One-sample t-test

Dr James Baglin (james.baglin@rmit.edu.au)

Hypothesis Testing Logic

- You might not realise, but we have already been introduced to many of the concepts behind hypothesis testing: sampling distributions, standard error and confidence intervals.
- Hypothesis testing formalises these concepts into an inferential decision making process.

Dr James Baglin (james.baglin@rmit.edu.au)

Hypothesis Testing Logic

Example: Comment on how usual you consider the Galos sample mean to be assuming IQ scores have a population mean of 100. Is there evidence that the population mean for IQ scores is not 100? Justify your conclusion by referring to the output from the sampling distribution."

- Hypothesis testing provides a set of rules to help us decide what should be considered "unusual".
- It's all about the Null hypothesis!

Hypothesis Testing Logic

The logic is simple...

- We begin by stating an assumption about the world, we will call this the **Null hypothesis, H_0**.
- **H_0** is bleak and uninteresting. It's the status quo. Nothing is happening.
    - The population mean IQ score is 100.
- Next, we state an opposing viewpoint that contradicts H_0, we will call this the **Alternate hypothesis, H_A**.
- **H_A** is what we set out to establish in the population, but we need to gather some evidence to support it.
    - The population mean IQ scores are not 100.
- The burden of proof is always on **H_A**. To support **H_A** we must rule out **H_0** beyond a reasonable doubt. In other words, reject **H_0**.

$H_0: \mu = 100$

$H_A: \mu \neq 100$

Dr James Baglin (james.baglin@rmit.edu.au)

Possible alternative Hypothesis

Under null Hypothesis we always assume a given value for the population parameter (mean) μ or (proportion) P:

H0 : μ = given value μ_0 or H0 : P= P0

Alternative Hypothesis HA formulate the research (experimental) question which are;

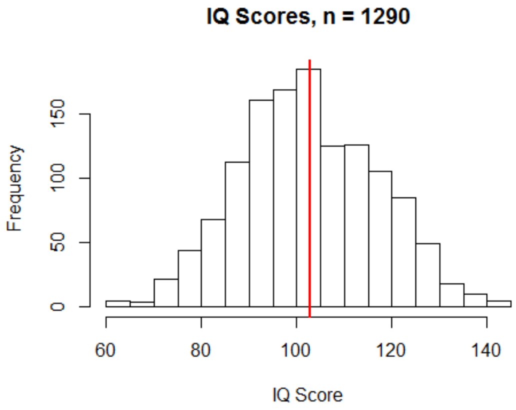
HA: $\mu > \mu_0$ or Ha: $\mu < \mu_0$ one sided test

or

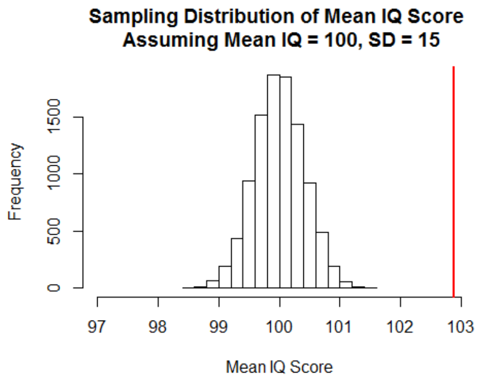
Ha : : $\mu \neq \mu_0$ two sided test

Hypothesis Testing Logic

- We will need some convincing evidence to reject $\mathbf{H_0}$.
- We gather a sample data from the population and using our knowledge of sampling distributions and confidence intervals, calculate how probable these results would be assuming H_0 is true.
- We define "unusual" as there being a less than 5% chance for a result to occur, or a result even more extreme, assuming $\mathbf{H_0}$ **is true**. We will call this 5% "line in the sand" the **significance (α) level** of the test.



We observe $\bar{x} = 102.89$



If we assume $\mu = 100$ and $\sigma = 15$...

Dr James Baglin (james.baglin@rmit.edu.au)

Approach to make the decision

- The general approach to hypothesis testing is as follows;
  - Assume that the null hypothesis is true.
  - Look for evidence to suggest that the null hypothesis is false based on a sample from the population;

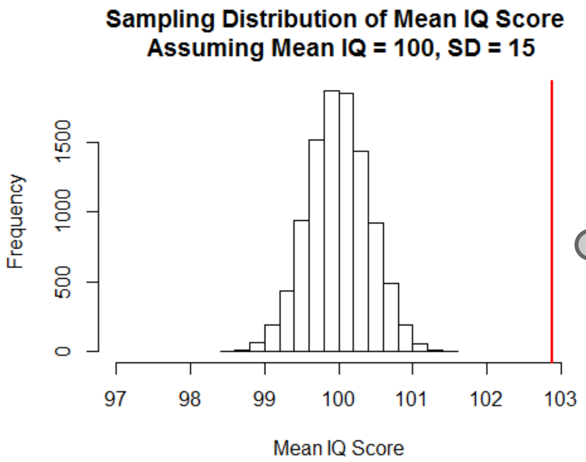
a. using the critical value of the test and see if is falling in the rejection region or

b. Is its *p*–value is less than the significant level α or

c. Does the (1-α)% Confidence Interval based on the sample mean contains the value of the population parameter stated under H0.

Hypothesis Testing Logic

- Next, we make a decision to **reject** or to **fail to reject H_0**, based on the probability of the data under H_0 and the criteria set by **α**.
- We **reject H_0** when the probability of the data under **H_0** < **α** or our **100(1 - α)%** misses **H_0**. These decisions rules are our burden of proof.
- Otherwise, we must **fail to reject H_0.**



Knowing what we do about the nature of sampling distributions of the mean, the probability of a sample mean of $\bar{x} = 102.89$, assuming $\mu = 100$, is really low...so low in fact, that we should reject H_0!

Calculating p-value for different alternative hypothesis

Under assumption that the sample is selected from a normal population (or at least mound-shaped distribution) the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

has Student's *t* distribution with n-1 degrees of freedom when n (sample size) is **less than 25.**

Calculating p-value for different alternative hypothesis

- Example: Prior to 1990 it was thought that the average oral human body temperature of a healthy adult was 37°Celsius (C). Investigators at that time were interested to know if this mean was correct. They gathered a sample of 130 adults and measured their oral body temperature. The dataset, Body_temp.csv, can be downloaded from the data repository. The descriptive statistics produced using R are reported below.

```
> favstats(Body_temp, data = Body_temp)
min    Q1   median Q3   max  mean   sd      n   missing
35.7 36.6 36.8   37.1 38.2 36.81 0.4074 130    0
```

Calculating p-value for different alternative hypothesis

H0: $\mu = 37°C$

HA: $\mu < 37°C$

- P-value = p = $Pr(\bar{x} < 36.81 | \mu = 37)$
- If $p < \alpha$, reject H0
- If $p \geq \alpha$, fail to reject H0
- Pay attention on how the sign for probability in p-value ($Pr(\bar{x} < 36.81 | \mu = 37)$)
- follows the sign in alternative hypothesis (HA: $\mu < 37°C$)
- To calculate the one-sided p-value, we need to convert the mean into t- test statistic and calculate P-value = p = $Pr(t < -5.38 | t = 0)$ where t is defined by:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Dr James Baglin (james.baglin@rmit.edu.au)

Calculating p-value for different alternative hypothesis

```
> pt(t,df=n-1)
[1] 1.680393e-07
```

Since p-value is very small we reject H0 and claim that the test is very significant and mean temperatures is less than 37.

- Things get a little strange if we use a two-tailed test (HA: $\mu \neq 37°C$). Because we need to take into account the mean also falling 5.38 SE above the mean, the p-value for a two-tailed test becomes:

P-value = $p = Pr(t < 5.83|t = 0) + Pr(t > 5.83|t = 0)$

As the t-distribution is symmetric, the two probabilities to be added are exactly the same. Therefore, a short cut to a two-tailed p-value can be calculated as:

$p = Pr(t < 5.83|t = 0)*2$,

```
> pt(t,df=n-1)*2
[1] 3.360785e-07
```

Calculating p-value for different alternative hypothesis

p = Pr(t < 5.83|t = 0)*2,

```
> pt(t,df=n-1)*2

[1] 3.360785e-07
```

The p value method

When to Reject H_0	When to Fail to reject H_0

• If the p value is less than the significance level (α)
• Example:
  ○ $p = 0.01$
  ○ $\alpha = 0.05$
Here, 0.01 is less than 0.05
• Rejecting H_0 suggests that there is a significant effect.
	• If the p value is greater than the significance level (α)
	• Example:
	  ○ $p = 0.09$
	  ○ $\alpha = 0.05$
	Here 0.09 is greater than 0.05
	• Failing to reject H_0 suggests that there is no significant effect.

Dr James Baglin (james.baglin@rmit.edu.au)

Confidence Interval Approach

If we use a confidence interval to test H0 for a one-sample *t*-test, we will automatically use a two-tailed hypothesis test.

- That's because most confidence intervals divide the significance level by 2 in their calculations. One-sided confidence intervals can be calculated, but are not typically supported by most statistical software.
- So let's test H0 using a confidence interval. First, we calculate the 95% *CI* for the sample mean x̄ = 36.81. Recall, when the population standard deviation is unknown, the 95% CI is calculated as:

$$\bar{x} \pm t_{n-1,1-(\alpha/2)} \frac{s}{\sqrt{n}}$$

Confidence Interval Approach

> confint(t.test(~ Body_temp, data = Body_temp))

mean of x lower   upper   level

36.80769 36.73699 36.87839   0.95000

95% *CI* for the sample mean, $\bar{x}$ = 36.81 [36.74, 36.87]

If the 95% *CI* does not capture the value of μ under H0, reject H0

If the 95% *CI* captures the value of μ under H0, fail to reject H0

We recall H0: μ = 37. Is μ = 37 captured by the 95% CI [36.74, 36.87]?

No. Therefore, our decision should be to reject H0.

Dr James Baglin (james.baglin@rmit.edu.au)

Using R to carry out hypothesis testing

The one-sample *t*-test in R:

```
> t.test(~ Body_temp, data=Body_temp ,mu = 37, alternative="less")


One Sample t-test


data:  data$Body_temp
t = -5.3818, df = 129, p-value = 1.68e-07
alternative hypothesis: true mean is less than 37
95 percent confidence interval:
     -Inf 36.86689
sample estimates:
mean of x
 36.80769
```


Dr James Baglin (james.baglin@rmit.edu.au)

Rejection Region Approach for one and two sided alternative

Two sided alternative:

```
> t<-qt(0.025,130-1,lower.tail=FALSE) #save t*
> mu <- 37 #Assign mu
> s <- sd(Body_temp$Bo dy_temp) #Assign sd
> n <-length(Body_temp$Body_temp) #Assign n
> se <-s/sqrt(n) #Calculate se
> mu + (t*se) #Determine lower critical mean
[1] 37.07
> mu - (t* se) #Determine upper critical mean
[1] 36.93
```

The lower t critical mean is equal to 36.93 and the upper critical region starts at 37.07.

Rejection Region Approach for one and two sided alternative

For two sided test is

$$\mu - t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}, \quad \mu + t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}$$

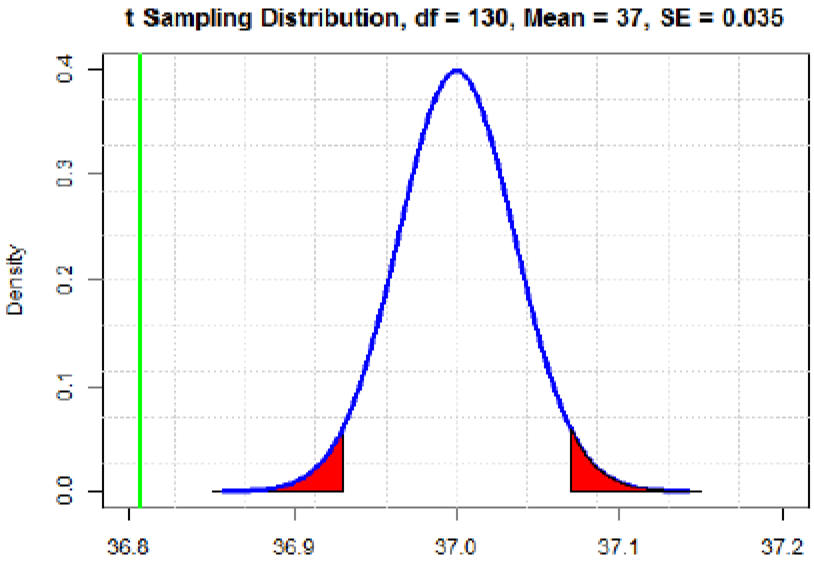
For one sided (H0: μ > given value or H0: μ < given value)

$$\mu - t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}, \qquad \mu + t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}$$

Dr James Baglin (james.baglin@rmit.edu.au)

Rejection Region Approach for one and two sided alternative

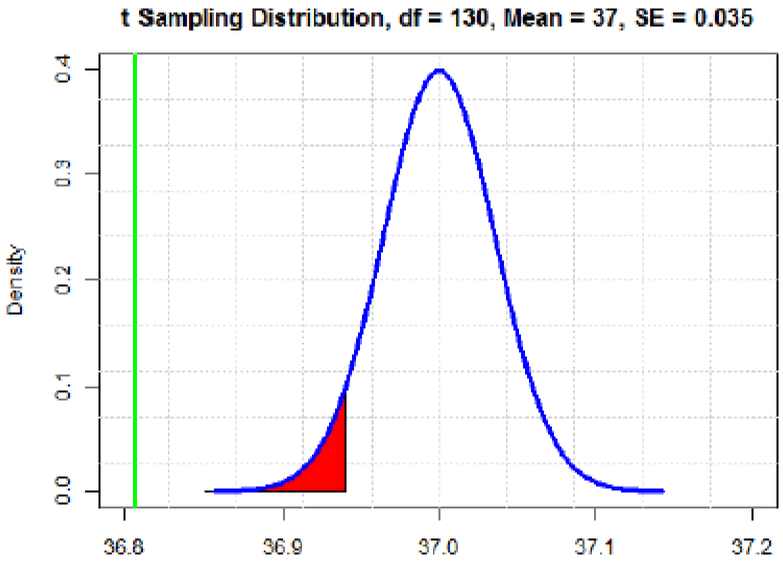
Two sided alternative



Rejection Region Approach for one and two sided alternative

One sided HA: μ < 37

```
t<-qt(0.05,130-1,lower.tail = TRUE) #save t-crit
> mu <- 37 #Assign mu
> s <- sd(Body_temp$Body_temp) #Assign sd
> n <-length(Body_temp$Body_temp) #Assign r
> se <-s/sqrt(n) #Calculate se
> mu + (t*se) #Determine critical mean
[1] 36.94079988
```

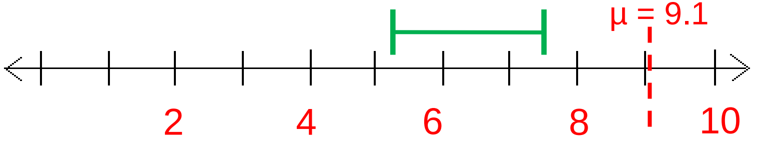
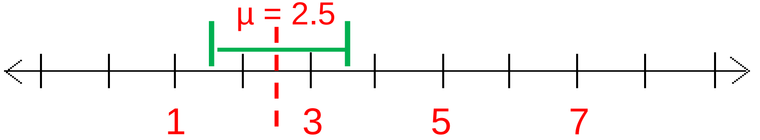
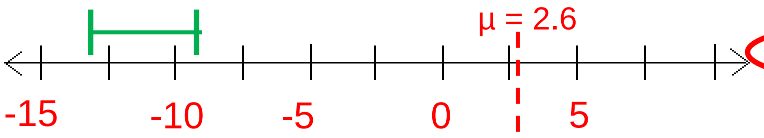
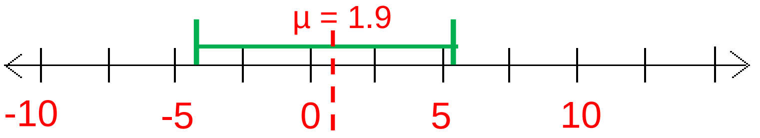


Question 4 Complete the table below regarding hypothesis testing

Method	Decision: Reject H_0 (Significant)	Decision: Fail to reject H_0 (Non Significant)
Critical value	If the test statistic lays within the critical region	If the test statistic does not lay within the critical region
Confidence interval	If the *CI* does not capture μ	If the *CI* does capture μ
p-value	$p < \alpha$	$p > \alpha$

Method	Decision (Significant) Reject H_0	Decision (Non Significant) Fail to reject H_0
Confidence interval	If the *CI* <u>does not</u> capture μ	If the *CI* <u>does</u> capture μ

<u>Question 7</u> Sketch 95% confidence intervals for the following scenarios. Also, indicate μ on your graphs and state the correct decision to make regarding H_0.

95% CI = (5.5, 7.5) μ = 9.1		Reject H_0 / Fail to reject H_0
95% CI = (1.7, 3.4) μ = 2.5		Reject H_0 / Fail to reject H_0
95% CI = (-13, -9) μ = 2.6		Reject H_0 / Fail to reject H_0
95% CI = (-4, 5.5) μ = 1.9		Reject H_0 / Fail to reject H_0

	Significant	Non Significant
	Decision	
Method	Reject H_0	Fail to reject H_0
p-value	$p < \alpha$	$p > \alpha$

Question 5 Determine the correct decision regarding H_0 for the following scenarios:

a. A *p*-value of 0.962 was discovered. The researchers were testing a $\alpha = 0.05$.
Reject H_0/Fail to reject H_0 because: $p > \alpha$ Significant/Not Significant.

b. A *p*-value of 0.001 was discovered. The researchers were testing a $\alpha = 0.01$.
Reject H_0/Fail to reject H_0 because: $p < \alpha$ Significant/Not Significant.

c. A *p*-value of 0.049 was discovered. The researchers were testing a $\alpha = 0.05$.
Reject H_0/Fail to reject H_0 because: $p < \alpha$ Significant/Not Significant.

d. A *p*-value of 0.049 was discovered. The researchers were testing a $\alpha = 0.01$.
Reject H_0/Fail to reject H_0 because: $p > \alpha$ Significant/Not Significant.

Rejection Region Approach for one and two sided alternative

One sided HA: μ > 37

t<-qt(0.05,130-1,lower.tail = TRUE) #save t-crit

> mu <- 37 #Assign mu

> s <- sd(Body_temp$Body_temp) #Assign sd

> n <-length(Body_temp$Body_temp) #Assign n

> se <-s/sqrt(n) #Calculate se

- mu - (t*se) #Determine critical mean
- [1] 37.0592
- So the value of μ =37 under H0 is not greater than 37.0592 therefore, we do not reject H0 based on rejection region.

Hypothesis Testing Logic

```
> library(mosaic)

> favstats(~IQ, data = IQ)
 min Q1 median  Q3 max     mean       sd    n missing
  60 93    102 113 144 102.8876 14.44732 1290       0

> t.test(~IQ, data = IQ, mu = 100)

	One Sample t-test

data:  data$IQ
t = 7.1787, df = 1289, p-value = 1.185e-12
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 102.0985 103.6767
sample estimates:
mean of x
 102.8876
```

How low? The *p*-value tells us that the probability of obtaining $\bar{x} = 102.89$, or a sample mean more extreme, assuming $\mu = 100$ is $< .001$.

The 95% CI of mean IQ does not capture $\mu = 100$. Not even close! Things are not looking good for H_0!

Dr James Baglin (james.baglin@rmit.edu.au)

Hypothesis Testing Logic

- ***p*-value**:
  - The probability of observing a sample mean IQ, or one more extreme, assuming $\mu = 100$. If the *p*-value is really small, the sample mean is really unlikely. This suggests our assumption that $\mu = 100$ (H_0) might be wrong.
- **95% CI**:
  - We know that 95% of confidence intervals for the mean, calculated under the same conditions using independent samples from the population, will capture the population mean.
  - If our 95% *CI* of the mean does not capture $\mu = 100$, there are two possibilities:
    - 1) It's one of those pesky 5% of *CI* that do not capture $\mu = 100$
    - 2) Our 95% CI comes from a population mean not equal to 100
  - 1) above will happen less than 5% of the time. Therefore, 2) above is more likely.

Hypothesis Testing Logic

- When we **Reject H_0**, we conclude the results of the test are **statistically significant.** This means we found evidence to support H_A.
- If we **fail to reject H_0,** this means there was insufficient evidence to reject H_0 and the results were **not statistically significant.**
- Don't worry, we will get plenty of practice applying this logic!

Decision: Reject H_0: $\mu = 100$ as *p*-value < .001 and 95% CI [102.1, 103.7] did not capture H_0

Conclusion:

The estimated population mean IQ based on the sample data was $\bar{x} = 102.89$, 95% CI [102.1, 103.7]. A one-sample t-test found the mean IQ score to be significantly different to the previously assumed population mean of 100, $t(df = 1289) = 7.18$, $p < .001$.

Module 7 Learning Objectives

Let's practice the following:

- Explain the process and logic of Null Hypothesis Significance Testing (NHST).
- State and test the assumptions behind the different *t*-tests.
- Determine when a one-sample *t*-test should be applied.
- Use technology to compute and interpret a one-sample *t*-test.

https://sites.google.com/a/rmit.edu.au/intro-to-stats/home/module-7

Dr James Baglin (james.baglin@rmit.edu.au)

Class Activity - Simple Reaction Time

- The average human reaction time (RT) for a simple RT task is said to be 268 milliseconds (ms)
- We will test this claim by measuring the reaction times of the class and treating them as a random sample.
    1. Measure your average RT across five tries using the following online test - http://www.humanbenchmark.com/tests/reactiontime
    2. Upload your results to the Google form (no trolling!) - http://goo.gl/forms/nlA08pvkc7 (login required)
    3. When instructed, download results from the Data Repository - `Reaction Time.csv`
- Import the data into RStudio and Perform a one-sample *t*-test to determine if there is statistical evidence that the mean human reaction time is different to 268 ms.

Dr James Baglin (james.baglin@rmit.edu.au)

Class Activity - Reaction Times

- **Hypotheses for the one-sample *t*-test**:

$$H_0: \mu = 268 \text{ ms}$$

$$H_A: \mu \neq 268 \text{ ms}$$

- **Assumptions**:
  - Known population mean, μ, unknown population standard deviation, σ.
  - Population data are normally distributed or large sample used ($n > 30$)
- **Decision Rules**:
  - Reject H_0:
    - If p-value < 0.05 (α significance level)
    - If 95% *CI* of the mean IQ does not capture H_0: $\mu = 268$
  - Otherwise, fail to reject H_0.
- **Conclusion**:
  - Test will be statistically significant if we reject H_0
  - Otherwise, the test is not statistically significant.

Class Activity - Reaction Times - R Code

- Use the following R code to perform the one-sample *t*-test for the class reaction time data.

```
favstats(~Average.RT, data = Reaction.Time)

hist(Reaction.Time$Average.RT,col="grey")
abline(v = 268, col = "red",lwd = 2) #Pop mean
abline(v=mean(Reaction.Time$Average.RT),
       col = "blue", lwd = 2) #Sample Mean

favstats(~Average.RT, data = subset(Reaction.Time,
                                    subset = Average.RT < 600))

t.test(~Average.RT, data = subset(Reaction.Time,
                                  subset = Average.RT < 600) , mu = 268)
```

Someone had some issues with the RT test. Remove the outlier.

Class Activity - Reactio Times - Interpretation

One-sample t-test results:

- Mean = 352, SD = 98
- t = 3.20
- 95% *CI* (310, 458)
- p-value = 0.002
- Decision: Reject H_0

What do we conclude?

- What conclusion can we draw from the results of the one-sample t-test?
- What are the limitations of this investigation?
- How could the investigation be improved?