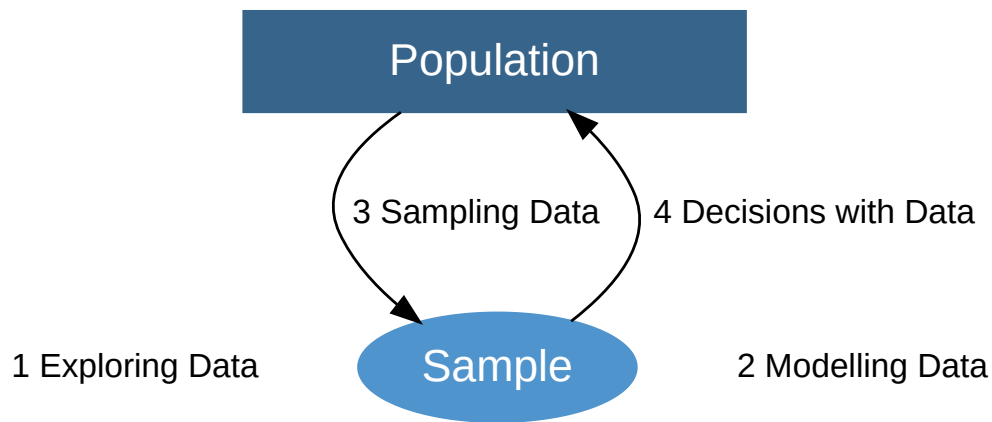


1 Sample Z and T Tests

Decisions with Data | Tests for a mean

© University of Sydney DATA1001/1901

Unit Overview





Module4 Decisions with Data

Test for a Proportion

How can we make evidence based decisions? Is an observed result due to chance or something else? How can we test whether a population has a certain proportion?

Tests for a Mean

How can we test whether a population has a certain mean? Or whether 2 populations have the same mean?

Tests for a Relationship [DATA1001/MATH1115]

How can we test whether 2 variables are linearly related? How can we test whether a categorical variable is in certain proportions?

Topic 33 Z and T Tests

Data Story | Caffeine and Endurance

The Z-Test

The T-Test

The Paired T-Test

Summary

Data Story



Caffeine and Endurance

I've heard that caffeine can give that extra boost during a workout to improve my performance, but I know that in high amounts it can also lead to testing positive for a banned substance.

Is caffeine something I should incorporate into my nutrition plan? Are there certain forms that are better than others?



Caffeine and Endurance

- **Caffeine** is a central nervous system stimulant.
- Collegiate and Professional Sports Dietitians Association (**CPSDA**) provides guidelines for consumption of caffeine before training or events.  **Guidelines**
 - “Consume 2-6 milligrams of caffeine per kilogram of body weight (one to three cups of brewed coffee for a 150-pound individual) one hour before cardiovascular endurance training or up to 20 minutes of high intensity training. Performance-enhancing effects may last up to four hours.”
- Why does it work? The body has a limited supply of glycogen. Caffeine encourages the depletion of fat for energy, rather than glycogen.  **Mechanism**

HOW MUCH CAFFEINE ARE YOU CONSUMING?

Caffeine-Containing Food Product	Amount of Caffeine (mg)
8 ounces of home-brewed drip coffee	80-100
8 ounces of instant coffee	65-100
2 ounces of espresso (latte, cappuccino, Americano)	100
8 ounces of decaffeinated coffee	5
8 ounces of brewed tea	50
12 ounces of caffeine-containing soft drinks	35-55
8 ounces of energy drink	80
Energy bar with caffeine	50 or 100
1.5 ounces of dark chocolate	30
2 caplets of Excedrin	130
1 caffeine tablet	200

NOTE: Exact amounts may vary between product brand and types.
Approximately 10 grams, or 80-100
8-ounce cups of coffee, is considered
the lethal dose of caffeine.



Caffeine: Is it legal in sports?

- Caffeine is considered a performance enhancer.
 - The National Collegiate Athletics Association ([NCAA](#)) currently rules that the maximum amount allowed in urine is 15 microgram/mL.
 - Pre-2003, the maximum amount allowed at the Olympics was 12 microgram/mL.
- The issue is that different people metabolise caffeine at different rates.
 - The Australia Heptathlete [Alex Watson](#) was sent home from the 1988 Summer Olympics for caffeine levels.
 - He drank 10+ cups coffee during competition from the event drink stand!
- In 2017 the World Anti-Doping Agency ([WADA](#)) added [Caffeine](#) to their monitoring programme to establish use among athletes.
- Will it be banned in the future?



 Alex Watson



Statistical Thinking

With the person next to you:

- How would you do an experiment to establish caffeine improved performance?
- How would you do an experiment to identify what dose of caffeine elevates urinary caffeine concentration above legal thresholds?

Research Data

Consider data from the following [research article](#):

W.J. Pasma, M.A. van Baak, A.E. Jeukendrup, A. de Haan (1995). "The Effect of Different Dosages of Caffeine on Endurance Performance Time", International Journal of Sports Medicine, Vol. 16, pp. 225-230.

Details:

- Sample: 9 elite cyclists.
- Experiment: Double blind, random order administration of caffeine capsules.
 - Dosage: 0, 5, 9, 13 mg caffeine per kg body weight.
 - Test performed once per week over 4 weeks.
 - Ride to exhaustion on stationary bike.
- Measurements: time to exhaustion, blood and urine samples.

A quick look at the data

Focus on the time (in minutes) to exhaustion for 0 and 13 mg caffeine per kg body weight.

```
caf0 = c(36.05, 52.47, 56.55, 45.2, 35.25, 66.38, 40.57, 57.15, 28.34)  
caf13 = c(37.55, 59.3, 79.12, 58.33, 70.54, 69.47, 46.48, 66.35, 36.2)  
mean(caf0)
```

```
## [1] 46.44
```

```
sd(caf0)
```

```
## [1] 12.48826
```

```
mean(caf13)
```

```
## [1] 58.14889
```

```
sd(caf13)
```

```
## [1] 15.13416
```



Does caffeine seem to affect exhaustion? If so, how?

The Z Test

Research Question 1

Is the mean time to exhaustion with no caffeine (“baseline”) equal to 45 minutes.

3 main steps for any hypothesis test

1. Set up research question

H: Hypothesis H_0 vs H_1

2. Weigh up evidence

A: Assumptions

T: Test Statistic

P: P-value

3. Explain conclusion

C: Conclusion

H: Hypotheses

- The **null hypothesis** H_0 assumes that any difference between the observed mean (OV) and the expected mean (EV) is due to chance alone. For the baseline, we assume that the mean of the box is 45.

$$H_0 : \text{mean exhaustion time with no caffeine} = 45 \text{ (mins)}$$

- The **alternative hypothesis** H_1 assumes that any difference between the OV and EV is NOT due to chance alone. Here, we choose a 2 sided alternative.

$$H_1 : \text{mean exhaustion time with no caffeine} \neq 45$$

A: Assumptions

- We assume the sample of cyclists is random - they are all independent (eg none related to each other).
- We assume the population is Normal. We want the distribution of the Sample Sum or Mean to be Normal, and as the sample size of 9 is small, the CLT would not be valid.
- We assume we know the population SD. For example, suppose a large scale study on the exhaustion time without caffeine was conducted previously, and the SD was 12 mins.

T: Test statistic for 1 Sample Z Test

The test statistic is used to measure the “gap” between the data (OV) and the EV based on the null hypothesis.

$$\text{test statistic} = \frac{\text{observed value (OV)} - \text{expected value (EV)}}{\text{standard error (SE)}}$$

$$\text{test statistic} = \frac{\text{observed mean} - \text{expected mean}}{\text{SD} / \sqrt{n}}$$

where SD is the known population SD.

```
teststat = (mean(caf0) - 45)/(12/sqrt(length(caf0)))  
teststat
```

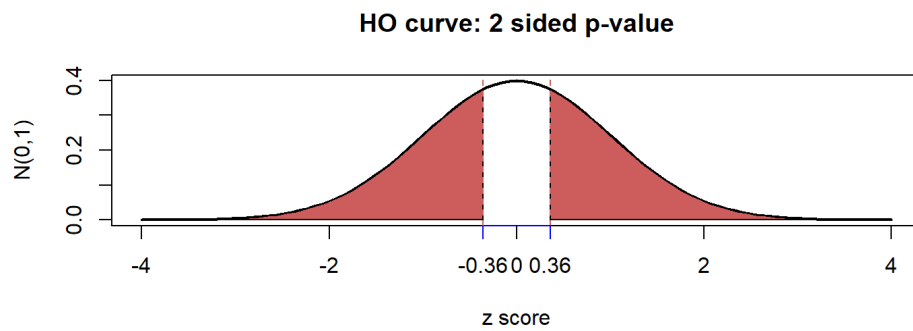
```
## [1] 0.36
```

Note: The test statistic is a measure of the difference between the observed and expected values on a standardised scale to facilitate comparison. A large value provides evidence against H_0 . How “large” is “large”? We define it in terms of the p-value.

P-Value

- The p-value is the chance of observing the test statistic (or more extreme) if H_0 is true.
- If the observed value is much bigger than 45 (EV), the the tail area (p-value) will become small. The significance level α represents the minimum tail area before we Reject H_0 .
- We look at 1 or 2 tails of the distribution, depending on what H_1 is.
 - If H_1 is \neq a certain mean, then we calculate the p-value in 2 tails.
 - If H_1 is $>$ or $<$ a certain mean, then we calculate the p-value in 1 tail.
- If we have a symmetric distribution (like the Normal), then the 2 sided p-value is just double the 1 sided p-value.

H_1 : 2 sided: Mean exhaustion time with no caffeine is not 45.



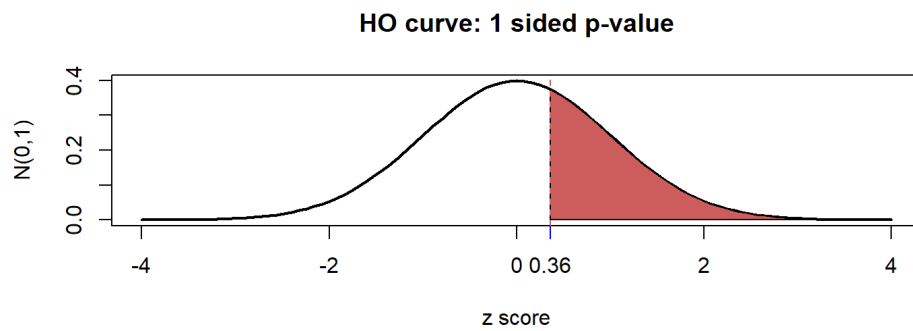
```
# Note lower.tail=F gives the upper tail area  
2 * pnorm(teststat, lower.tail = F)
```

```
## [1] 0.7188471
```

Conclusion [2 sided H_1]

This means: There is about a 72 in 100 chance of getting a test statistic like this (or even more extreme), if the null hypothesis is true. Therefore the data appears consistent with the null hypothesis.

H_1 : 1 sided: Mean exhaustion time with no caffeine is > 45 .



```
# Note lower.tail=F gives the upper tail area  
pnorm(teststat, lower.tail = F)
```

```
## [1] 0.3594236
```

Conclusion [1 sided H_1]

This means: There is about a 36 in 100 chance of getting a test statistic like this (or even more extreme), if the null hypothesis is true.

Size of P-value

- How small does the p-value need to be before we reject H_0 ?
- There is no simple answer to this question, it depends on context.
- By **convention** we choose a significance level: α , often 0.05.
- Commonly we say:

P-value	Conclusion
p-value < 5%	significant or statistically significant
p-value < 1%	highly significant

- We always report the value of the p-value, so other investigators can weigh it up in context.

Summary: Z Test [1 sided]



Z Test

H: H_0 : population mean = c vs H_1 : population mean < c (or >)

A: Sample is random. Population is Normal, or sample size is large enough for the CLT. Population SD is known.

T: Test statistic =
$$\frac{\text{observed mean (OV)} - \text{expected mean (EV)}}{SD/\sqrt{n}}$$

P: Use Normal curve to find tail area for observed test statistic.

C: Retain or Reject H_0 .

Summary: Z Test [2 sided]



Z Test

H: H_0 : population mean = c vs H_1 : population mean \neq c

A: Sample is random. Population is Normal, or sample size is large enough for the CLT. Population SD is known.

T: Test statistic =
$$\frac{\text{observed mean (OV)} - \text{expected mean (EV)}}{SD/\sqrt{n}}$$

P: Use Normal curve to find 2 tail areas for observed test statistic.

C: Retain or Reject H_0 .

Comparing the Z Test to the Proportion Test

How does the use of the Z Test here differ from the Proportion Test for the peanut allergy treatment?

- Both tests use a Z Test (based on the Normal). However, here the Z-test is used for continuous data (time), whereas the peanut data was binary (success or failure of the probiotic treatment).
- The Z-test can only apply to binary data when there is a large sample size, allowing the use of the CLT. If the sample proportion is not Normal, we may have to bootstrap the distribution based on the sample proportion.
- Here, we calculated the SE for the sample **Mean**, whereas with the peanut allergy example we calculated the SE for the sample **Sum** (the number of successful treatments).

- Here, we assume we know the population SD (eg from a previous study). For the peanut allergy data, the SD of the population was worked out as $\sqrt{p(1-p)}$, using the true proportion p in H_0 .

The T-Test

When we don't know the population SD ...

To use the Z-test, we need to know the **population SD**, and this is often not known!

Solution 1: Estimate the population SD from the sample SD, and use the Z Test.

- This estimation will add extra variability to the test statistic, as the sample SD varies from sample to sample.
- For large samples, the difference between the population SD and sample SD should be small, and so this may be appropriate. However, for small samples, the difference will be more noticeable, and hence, we should use the T-Test.

Solution 2: Use the T Test

The T-Test

- [W.S. Gosset](#) (1876-1936) invented a similar test to the Z Test, which uses the **sample SD** and the T distribution.
- The T distribution varies in shape according to the sample size. The smaller the sample size is, the more variable the sample SD is, and hence the distribution of the test statistic will be “wider”. The degree of “wide-ness” (also called **degree of freedom**) depends on the sample size and here it is $n - 1$. We write such a distribution as t_{n-1} .
- See the difference [here](#)

Summary: T Test [1 sided]



T Test

H: H_0 : population mean = c vs H_1 : population mean < c (or >)

A: Sample is random. Population is Normal, or sample size is large enough for the CLT.

T: Test statistic =
$$\frac{\text{observed mean} - \text{population mean}}{\text{sample SD} / \sqrt{n}}$$

P: Use t_{n-1} curve to find tail area for observed test statistic.

C: Retain or Reject H_0 .

Summary: T Test [2 sided]



T Test

H: H_0 : population mean = c vs H_1 : population mean $\neq c$

A: Sample is random. Population is Normal, or sample size is large enough for the CLT.

T: Test statistic =
$$\frac{\text{observed mean} - \text{population mean}}{\text{sample SD} / \sqrt{n}}$$

P: Use t_{n-1} curve to find the 2 tail areas for the observed test statistic.

C: Retain or Reject H_0 .

Comparing the Z and T Tests

Test	Test Statistic	P-value curve
Z	Test statistic = $\frac{\text{observed mean} - \text{population mean}}{\text{population SD}/\sqrt{n}}$	Normal
T	Test statistic = $\frac{\text{observed mean} - \text{population mean}}{\text{sample SD}/\sqrt{n}}$	t_{n-1}

The sample SD is found using `sd` in R.

Applying the T-Test to the Baseline data

H

H_0 : mean exhaustion time with no caffeine = 45 (mins)

H_1 : mean exhaustion time with no caffeine \neq 45

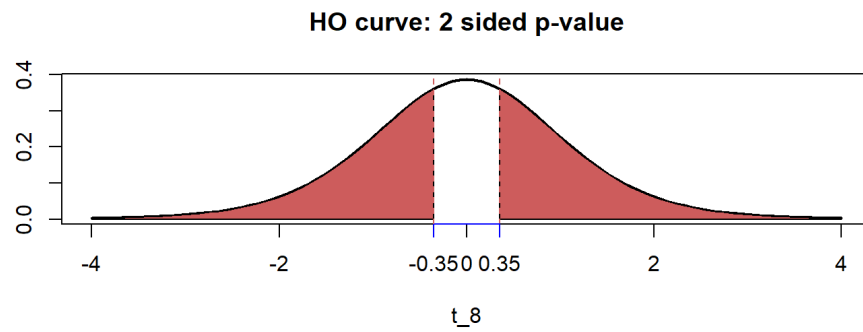
A

- We assume the sample of cyclists is random - they are all independent (eg none related to each other).
- We assume the population is Normal. We want the distribution of the Sample Sum or Mean to be Normal, and as the sample size of 9 is small, the CLT would not be valid.

T

```
teststat1 = (mean(caf0) - 45)/(sd(caf0)/sqrt(length(caf0)))  
teststat1
```

```
## [1] 0.3459249
```

P

```
pt(teststat1,8,lower.tail=F) # 1 tailed  
2*pt(teststat1,8,lower.tail=F) # 2 tailed
```

C

- Notice that the P-value here is slightly larger here than the Z Test, due to the extra wide-ness for a small sample size of $n = 9$ making the tail area larger.
- However, the conclusion is the same as for the Z Test - we retain H_0 .

A speedy way to do this in R

The `t.test` command calculates the test statistic and p-value for us!

```
t.test(caf0, mu = 45)
```

```
##
## One Sample t-test
##
## data:  caf0
## t = 0.34592, df = 8, p-value = 0.7383
## alternative hypothesis: true mean is not equal to 45
## 95 percent confidence interval:
##  36.84067 56.03933
## sample estimates:
## mean of x
##      46.44
```

Note:

- The 95% CI is equivalent to a 2 sided Test with $\alpha = 0.05$.
- As 45 falls inside the CI hence we retain H_0 .

The Paired T Test

Research Question 2

Is the mean time to exhaustion with no caffeine (“baseline”) the same as 13mg caffeine? ie Is there any difference between having caffeine or not?

Difference between 0 and 13 caffeine levels

- Since each cyclist had each dosage level, we now focus on the difference in time (in minutes) to exhaustion for 0 and 13 mg caffeine per kg body weight.
- This 1 sample of “differences” is now our focus.

```
caf0 = c(36.05, 52.47, 56.55, 45.2, 35.25, 66.38, 40.57, 57.15, 28.34)
caf13 = c(37.55, 59.3, 79.12, 58.33, 70.54, 69.47, 46.48, 66.35, 36.2)
cafdiff = caf13 - caf0 # This 1 sample is our focus: the 'differences'.
mean(cafdiff)
```

```
## [1] 11.70889
```

```
sd(cafdiff)
```

```
## [1] 10.79987
```

Using t.test

Again, the `t.test` command easily calculates the test statistic and p-value for us!

```
t.test(cafdiff, mu = 0)
```

```
##  
## One Sample t-test  
##  
## data: cafdiff  
## t = 3.2525, df = 8, p-value = 0.01166  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
##  3.407372 20.010405  
## sample estimates:  
## mean of x  
## 11.70889
```

Summary of T Test

H

H_0 : mean exhaustion time for the differences = 0 (mins)

H_1 : mean exhaustion time for the differences $\neq 0$

A

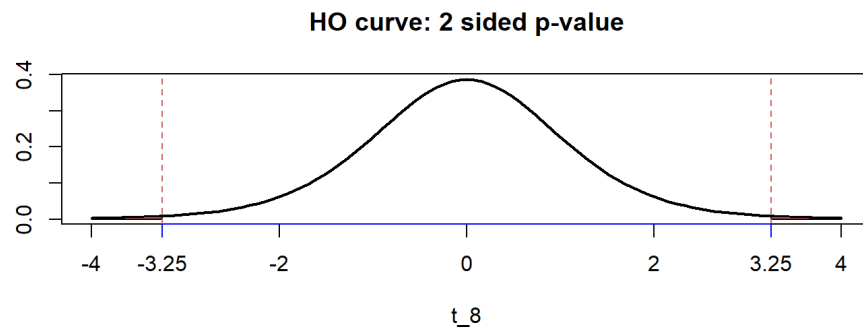
- We assume the sample of cyclists is random - they are all independent (eg none related to each other).
- We assume the population of differences is Normal.

T

```
teststat2 = (mean(cafdiff) - 0)/(sd(cafdiff)/sqrt(length(cafdiff)))  
teststat2
```

```
## [1] 3.252508
```

P



```
2 * pt(teststat2, 8, lower.tail = F) # 2 tailed
```

```
## [1] 0.01165724
```

C

- As the p-value is so small, we reject H_0 and conclude that “there is a difference” - ie caffeine consumption affects endurance.

Note on SD vs SE

- The **SD** tells us how far each individual cyclist varies from the mean in **this sample** of 9 riders.
- The **SE** tells us how far the **sample means** vary from the true population mean, for all elite cyclists.
- Since the sample size of 9 is small, the distribution of sample means may not be Normal under the CLT. If we use the Normal to construct a CI, we need to assume that the distribution of the box (population for each measurement) is Normal.

Extension: How to calculate the CI by hand

```
# T value for 95% CI (ie upper tail is 5%/2 = 0.025)
tval = qt(0.025, 8, lower.tail = F)
tval
```

```
## [1] 2.306004
```

```
mean(cafdiff)
```

```
## [1] 11.70889
```

```
sd(cafdiff)
```

```
## [1] 10.79987
```

```
c(mean(cafdiff) - tval * sd(cafdiff)/sqrt(9), mean(cafdiff) + tval * sd(cafdiff)/sqrt(9))
```

```
## [1] 3.407372 20.010405
```

Summary

- The 1 sample Z and T Tests are used to test a hypothesis about a certain mean.
- If the population SD is unknown, we use the T-Test, especially in the case of small samples.