STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 17: Difference Between Two Means

Outline

- Previous lectures:
  - Explored statistical models for normally distributed data
  - Data are modelled as normal with mean μ and variance σ^2
  - Found confidence interval for μ
  - Hypothesis test for μ
- Today: begin to look at relationships between variables
  - Relationship between a continuous variable and a categorical variable
  - Continuous variable: can take any value
    - e.g. height, weight, time to run 100 m
    - It could be limited a range (e.g. height must be positive)
  - Categorical variable: represents categories or groups
    - e.g. sex, country of birth, blood type, etc.

Motivation

- What is the effect of sensory deprivation?[1]
  - Study designed to explore this question, where all participants were prisoners
- Twenty participants were selected
  - 82 inmates initially volunteered
    - Removed: medically unfit, low IQ, history of behaviour or psychiatric problems in prison
- The 20 participants were randomly allocated into two groups
  - Solitary confinement
  - Control (ordinary prison life)
- EEG[2] frequencies were obtained on day 7
  - Is there a difference in arousal levels? (as measured by EEG frequency)

[1] From Journal of Abnormal Psychology, 1972, **79**, 54–59

[2] EEG (Electroencephalogram) measures the frequency of brain waves

Data: EEG frequencies

- Import the data

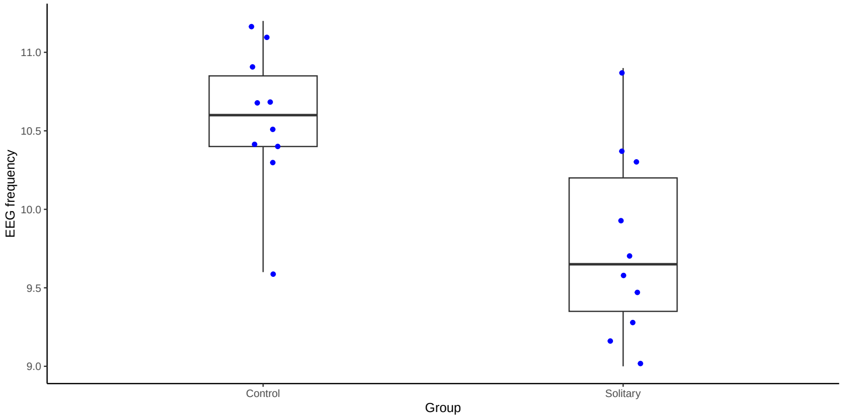
```r
EEG = read.csv('EEG.csv')
```

- Have a look at the data:

```r
head(EEG)
```

```
##     Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
```

Visualise the data

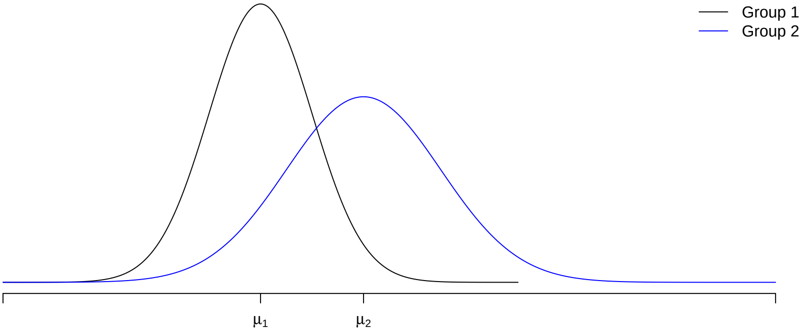


https://mathstatfiles.otago.ac.nz/STAT115/GEEplot.r

Problem

- We have looked at models:
  - Data are normally distributed with mean μ and variance σ^2
  - Focus has been on the estimation of a (single) mean μ
- We need to extend our model to allow for two groups of data
  - Group 1 (experimental): normally distributed with mean μ_1 and variance σ_1^2
  - Group 2 (control): normally distributed with mean μ_2 and variance σ_2^2
- Interest is in the difference in means between the two groups
  - $\mu_1 - \mu_2$ (or $\mu_2 - \mu_1$)
- Difference in the mean arousal level between the deprived and the controls

Model (graphical representation)



Other examples

- There are other applications we could have used to motivate:
  - Cuckoos are avian brood parasites: they lay their eggs in the nest of other birds
    - Compare the length of cuckoo eggs in wren and robin nests
  - Explore differences in chemical composition of wine or olives
    - Different cultivars (wine)
    - Different regions (olives)
  - Comparing athletic performance
    - Comparing resistance training and traditional training for athletes in some sport
  - Survival time for breast cancer patients
    - Comparing candidate drug and placebo
  - Gene expression in a section of the brain
    - Comparing diseased, with healthy controls
  - You will see a variety of examples in Assignments

How to find a confidence interval

- Much of what we have learned previously 'carries over'
- Use statistics (from sample) to estimate parameters (from population)
  - Parameter: $\mu_1 - \mu_2$
  - Statistic: $\bar{y}_1 - \bar{y}_2$
- Standard error for $\bar{y}_1 - \bar{y}_2$
  - Tells us about the variation in $\bar{y}_1 - \bar{y}_2$ in repeated samples
  - Estimated standard error: $s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- The confidence interval is given as

$$\underbrace{\bar{y}_1 - \bar{y}_2}_{\text{statistic}} \pm \underbrace{t_{\nu,1-\alpha/2}}_{\text{multiplier}} \underbrace{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}_{\text{standard error}}$$

Standard error

- The standard error is different from before, but similar
  - Follows from variance rules (Lecture 9)
  - Observations in the two groups are independent

$$
\begin{aligned}
Var(\bar{y}_1 - \bar{y}_2) &= Var(\bar{y}_1) + Var(\bar{y}_2) \\
&= \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}
\end{aligned}
$$

Multiplier

- The multiplier is again given by the t-distribution
  - The use of the t-distribution relies on an approximation
    - Approximation is accurate provided we have more than a handful of observations ($n_1 > 5, n_2 > 5$)
- The degrees of freedom, ν, we use is given by a complicated formula
  - You have no need to know or learn this

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

- If software isn't available, simpler approximations for ν are sometimes used
  - e.g. using smaller of $n_1 - 1$ and $n_2 - 1$
  - Conservative

Calculating the confidence interval

- We could calculate the confidence interval by hand:
  - Find the sample mean in each group: $\bar{y}_1, \bar{y}_2$
  - Find the sample variance in each group: s_1^2, s_2^2
  - Find the standard error
  - Calculate the degrees of freedom
  - Find the t-multiplier
  - Construct the confidence interval
- Tedious task
  - Important to know how the interval is constructed
    - You may be asked to do various aspects of it for assignment/test/exam
  - Easier to use R to calculate the interval

In R

- We use the same function as before: `t.test`
  - This requires us to have the data for each group separately
  - Currently our data are in a single data frame

```
head(EEG)

##     Group Freq
## 1 Control 10.7
## 2 Control 10.7
## 3 Control 10.4
## 4 Control 10.9
## 5 Control 10.5
## 6 Control 10.3
```

- The variable `Group` distinguishes which group the observation is from
  - Either `Control` or `Solitary`

In R

- There are several ways in R we could separate into two groups
  - We will use `subset`
    - Subsets the data based on a specified criteria
  - Only cover 'basic' data handling in STAT115
    - See STAT 260

```
control = subset(EEG, Group == "Control")
solitary = subset(EEG, Group == "Solitary")
```

- We use two equal signs (==) to *check* equality
  - `Group == "Solitary"` is checking which observations are `Solitary`

In R

- Check each of these objects

```
control
```

```
##       Group Freq
## 1   Control 10.7
## 2   Control 10.7
## 3   Control 10.4
## 4   Control 10.9
## 5   Control 10.5
## 6   Control 10.3
## 7   Control  9.6
## 8   Control 11.1
## 9   Control 11.2
## 10  Control 10.4
```

```
solitary
```

```
##        Group Freq
## 11 Solitary  9.6
## 12 Solitary 10.4
## 13 Solitary  9.7
## 14 Solitary 10.3
## 15 Solitary  9.2
## 16 Solitary  9.3
## 17 Solitary  9.9
## 18 Solitary  9.5
## 19 Solitary  9.0
## 20 Solitary 10.9
```

In R

- Each of the groups is a separate argument in `t.test`

```
out = t.test(control$Freq, solitary$Freq)
out

##
##  Welch Two Sample t-test
##
## data:  control$Freq and solitary$Freq
## t = 3.4, df = 17, p-value = 0.004
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2969 1.3031
## sample estimates:
## mean of x mean of y
##     10.58      9.78
```

R output

- R calculates the degrees of freedom for us: $\nu = 16.875$
- R gives us the means

```
out$estimate # gives the samples means of the two groups

## mean of x mean of y
##     10.58      9.78

out$estimate[1] - out$estimate[2] # find the diff in sample means

## mean of x
##       0.8
```

- When interpreting, we must be careful to not confuse the order
  - Mean of x corresponds to the first argument: controls
  - Mean of y corresponds to the second argument: solitary
  - Confidence interval is for $\mu_x - \mu_y$, or $\mu_{\text{control}} - \mu_{\text{solitary}}$

Confidence interval

- The confidence interval is

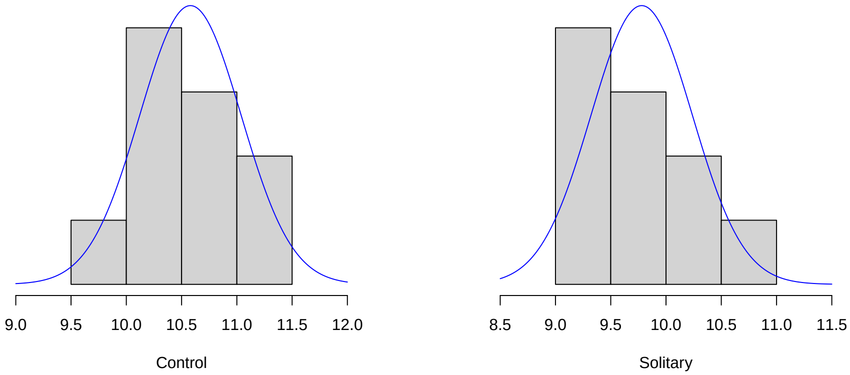
```
out$conf.int

## [1] 0.2969 1.3031
## attr(,"conf.level")
## [1] 0.95
```

- We are 95% confident that the mean EEG frequency for the control group is between (0.2969, 1.3031) higher than those in solitary confinement
- The confidence interval has the same properties as before
  - In the long run, we would expect 95% of the confidence intervals we calculate to include the true difference $\mu_1 - \mu_2$
    - If we were to repeatedly sample from the population and repeat this analysis

Checking assumptions

- We are assuming a normal model for each group
- Check fitted model



Checking assumptions

- Do the data show departures fromn normality?
- Enough to make us cautious
  - Small sample size: normality assumption very important
    - It is hardest to assess normality assumptions, when it matters the most
- Want to be cautious in our conclusions

Hypothesis test

- This study was set up to look into a specific hypothesis
  - Confirmatory
- Theory was that sensory deprivation changes EEG frequency
- Null hypothesis: status quo / assumption of no difference
  - The two groups have the same mean: $\mu_1 = \mu_2$
  - $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$
- The alternative hypothesis
  - The two groups differ: $\mu_1 \neq \mu_2$
  - $\mathrm{H}_A : \mu_1 - \mu_2 \neq 0$

Hypothesis test

- The same function (`t.test`) is used to calculate a hypothesis test

```
out = t.test(control$Freq, solitary$Freq)
out
##
##  Welch Two Sample t-test
##
## data:  control$Freq and solitary$Freq
## t = 3.4, df = 17, p-value = 0.004
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2969 1.3031
## sample estimates:
## mean of x mean of y
##     10.58      9.78
```

Interpretation

- The p-value is 0.0038
  - Evidence of incompatibility between data and null hypothesis
  - Data provide support for the alternative hypothesis
    - Difference in EEG frequency between the control and solitary groups
- Given the small sample and cautiousness in checking assumptions
  - We have provided evidence in support of EEG differing
  - Larger studies desirable to provide further confirmation

Confidence intervals vs hypothesis testing

- In this example we look at both confidence intervals and hypothesis test
- The p-value does not tell us how strong an effect is
  - We could have p-value of 0.05 with $\bar{y}_1 - \bar{y}_2 = 10$
    - Small sample size
  - We could have p-value of 0.001 with $\bar{y}_1 - \bar{y}_2 = 0.002$
    - Large sample size
- Confidence interval gives an interval estimate of effect

Independent groups

- We have assumed the two groups are independent
  - Important assumption
- What does that mean?
  - The outcome from one group does not affect the outcome from the other group
- This will not always be the case:
  - Students take a test before undertaking a course
  - Same students undertake the same test after the course
    - Same participants in each ‘group’
    - It is likely that someone who scored well in first test will also score well in the second test
- Look into this more next lecture

Summary

- First look at relationship between variables
  - How EEG frequency varies by sensory deprivation
- Relationship between a continuous variable and a categorical variable
  - EEG frequency (continuous); sensory deprivation yes/no (categorical)