Week01 - Summary

Review of Basic Statistical Concepts

Population

Examples:

- Blood pressure readings of all people in Australia
- The number of languages spoken from all currently enrolled students in University of Sydney

Sample

Examples:

- Blood pressure readings of 1000 randomly selected people in Australia
- The number of languages spoken from 500 randomly selected students currently enrolled in University of Sydney

Parameters vs. Statistic

- A parameter is a fixed number (usually unknown). A statistic is a variable whose value varies from sample to sample

Visualisation Packages

Simple Example data frames for Plots

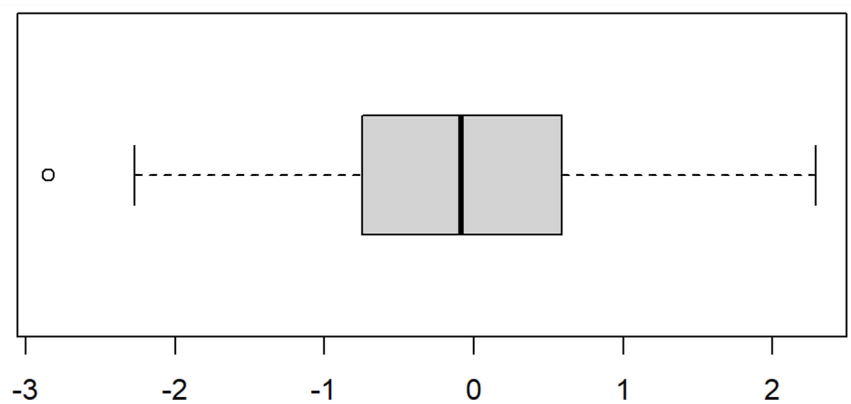
```
example.dat <- data.frame(x = rnorm(100), y = runif(100),
                          cat = sample(LETTERS[1:2], prob = c(1, 3), size = 100, replace = TRUE))

head(example.dat)
```

```
           x         y cat
1  0.37573068 0.8676167   A
2 -1.70387817 0.6760221   B
3 -1.64878643 0.7621811   A
4  0.09658172 0.2585820   B
5  0.74011371 0.2326891   A
6 -0.86970148 0.3605919   B
```

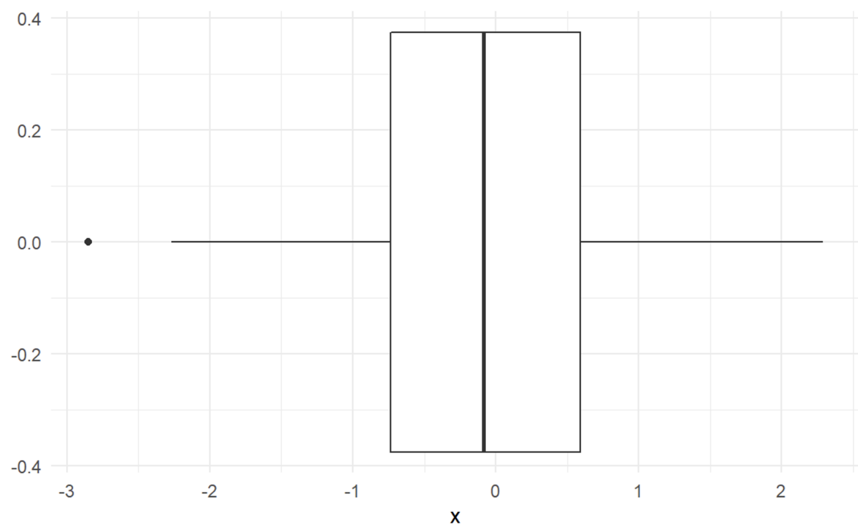
Single Numeric Variable: Boxplot in base `R`

```
boxplot(example.dat$x, horizontal = TRUE)
# by default the boxplot is vertical
```



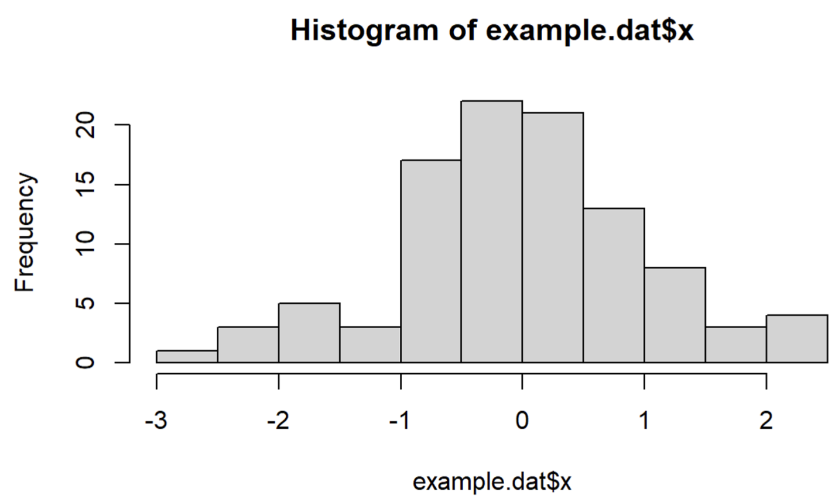
Single Numeric Variable: Boxplot in `ggplot2`

```
library(ggplot2)
ggplot(example.dat, aes(x = x)) + geom_boxplot() + theme_minimal()
# theme_minimal() gives a basic background
# mapping is done using aes()
```



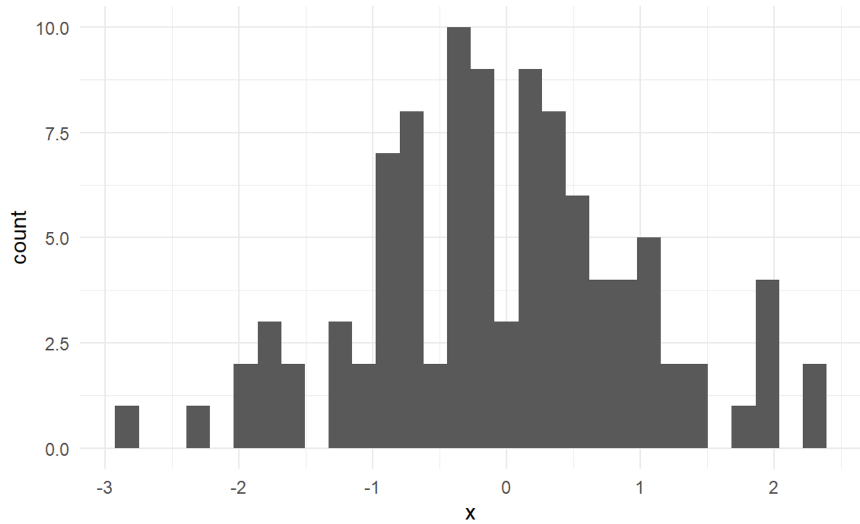
Single Numeric Variable: Histogram in base `R`

```
hist(example.dat$x)
```



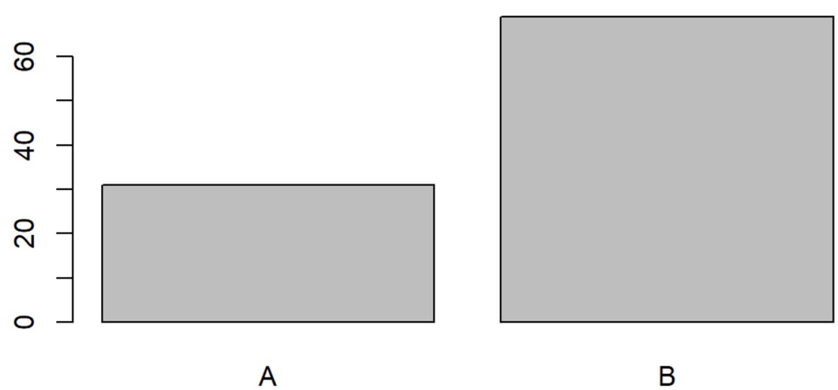
Single Numeric Variable: Histogram in `ggplot2`

```
ggplot(example.dat, aes(x = x)) + geom_histogram() + theme_minimal()
# you can specify the number of bins, which controls the bandwidth of the histogram
```



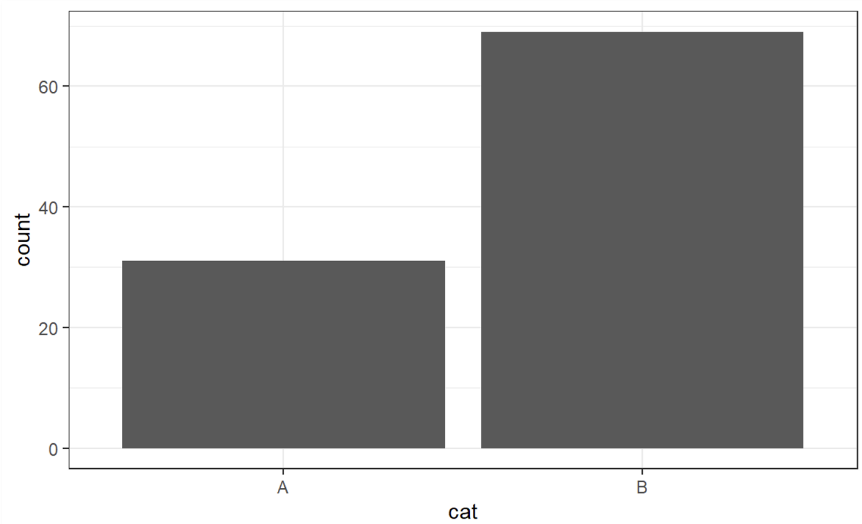
Single Categorical Variable: Bar Plot in base `R`

```
barplot(table(example.dat$cat))
# table() creates the number of counts for each variable
```



Single Categorical Variable: Bar Plot `ggplot2`

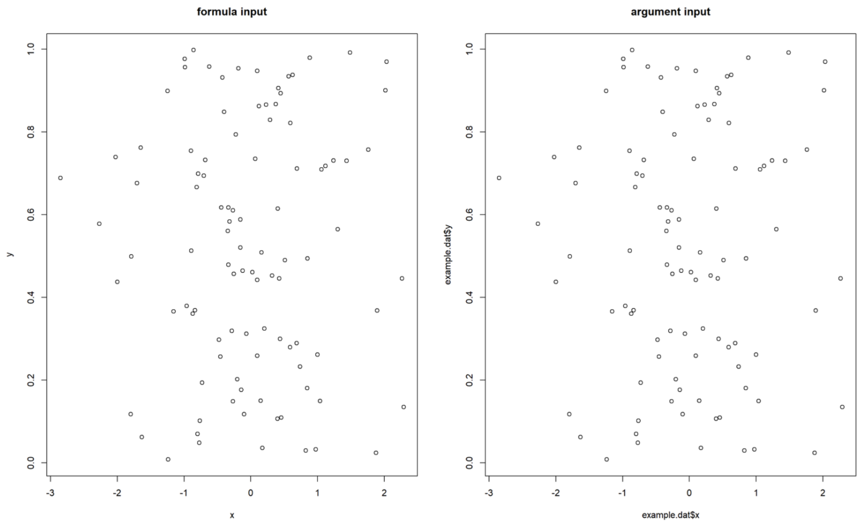
```
ggplot(example.dat, aes(x = cat)) + geom_bar() + theme_bw() # Change the theme
```



Two Numeric Variables: Scatterplot in base `R`

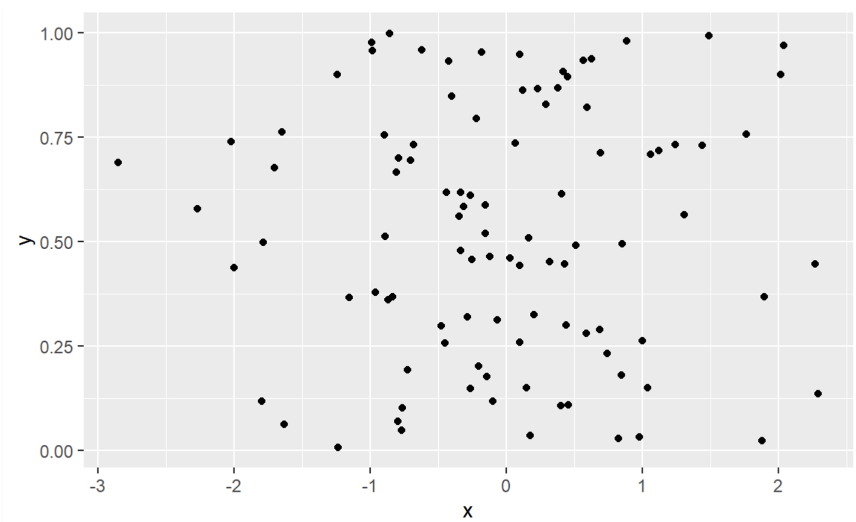
```
# These two plot commands are near equivalent
# this method you need to specify your dataset
plot(y ~ x, data = example.dat, main = "formula input")

plot(example.dat$x, example.dat$y, main = "argument input")
```



Two Numeric Variables: Scatterplot in `ggplot2`

```
ggplot(example.dat, aes(x = x, y = y)) + geom_point() # default theme here
```



Coding with R

Base `R` and the tidyverse

- The `tidyverse` has a (somewhat) standardised syntax (pipes `|>` or `%>%` are key except for `+` in `ggplot2`)
- Produces more human readable code however not as stable as base, breaking changes occurs as `tidyverse` develops

- Core base `R`
  - Good for production level code
  - Stable
  - Function syntax inconsistent

Homogeneous vs. Non-homogeneous Data Types in R

Homogenous	Non-homogeneous

Vector - Sequence of data elements of the same basic data type	List - More general structure containing other objects (including possibly other lists)
Matrix - Collection of data elements in a 2-dimensional array with rows and columns	Data frame - Used for storing data, each column can be a different basic type - All columns must have the same length

Vectors

```
new.vector <- c(1, 2, 3)
class(new.vector)

length(new.vector)

new.vector[1:2]

new.vector <- c(1, 2, "hello")
class(new.vector)
```

Matrix

```
A <- matrix(c(2, 4, 3, 1, 7, 8), nrow = 3)
# Unless specified otherwise, it will fill the matrix by column.
```

List

```
vector.a <- c(1, 2, 3)
vector.b <- c("hello", "world", "!!")
new.list <- list(c(vector.a, vector.b))
new.list

new.list <- list(vector.a, vector.b)
new.list
```

Data Frames

```
head(warpbreaks) # just an example
class(warpbreaks)
head(warpbreaks$wool)
str(warpbreaks)
names(warpbreaks)

# we can check if a variable is a list using:
# is.hist(warpbreaks) - in this case
# should be true, since a dataframe is a special kind of list
```

Line of best fit

$Y = \beta_0 + \beta_1 X + \varepsilon$

- X is the predictor (feature or independent variable)
- Y is the response (target or dependent variable)
- β_0 is the intercept of the regression line
- β_1 is the slope of the regression line
- ε is the unexplained variation or random error

Residual sum of squares

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

- Can show by simple calculus the following:
  - Regression (slope) coefficient:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{cov(x,y)}{var(x)}$$

- Intercept: $b_0 = \bar{y} - b_1\bar{x}$

- This leads to the estimated regression line:

$$\hat{y} = b_0 + b_1 x$$

- Least squares regression line since it minimises the residual sum of squares

Simple linear regression

Fitting a linear model

```
lm.fit <- lm(Price ~ BuildingArea, data = st.kilda.data)
summary(lm.fit)
```

```
Call:
lm(formula = Price ~ BuildingArea, data = st.kilda.data)

Residuals:
    Min      1Q  Median      3Q     Max
-817415 -201614  -85181   19895 3403199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -129484.0    91775.9  -1.411    0.161
BuildingArea   11209.5      799.8  14.015   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 490300 on 99 degrees of freedom
Multiple R-squared:  0.6649,    Adjusted R-squared:  0.6615
F-statistic: 196.4 on 1 and 99 DF,  p-value: < 2.2e-16
```

$\hat{y} = -129484 + 11209 \times \text{BuildingArea}$

$\widehat{\text{Price}} = -129484 + 11209 \times \text{Area}$

Standard error of population mean

- Consider single population estimation problem

- Wish to estimate some mean, μ, of some random variable Y.
- If Y_i is sampled then $\hat{\mu} = \overline{Y}$ estimates μ with
- $Var(\hat{\mu}) = \left(SE(\hat{\mu})\right)^2 = \frac{\sigma^2}{n}$
- σ^2 is the variance of Y_i
- n is the sample size.

$SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$.

Standard error of regression coefficient estimates

- Same concept applies to the regression estimates

$$SE(\widehat{\beta_0}) = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \qquad SE(\widehat{\beta_1}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

where $\sigma^2 = Var(\varepsilon)$

- As $n \to \infty$, $SE(\hat{\beta}_0) \to 0$ and $SE(\hat{\beta}_1) \to 0$

- Interestingly, if the x_i are more spread out, the standard errors will be smaller
  - more leverage to estimate the parameters

Using standard errors to compute confidence intervals

```
summary(lm.fit) # Truncated output with coefficient table
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -129484.0    91775.9  -1.411    0.161
BuildingArea     11209.5      799.8  14.015   <2e-16 ***
---

Residual standard error: 490300 on 99 degrees of freedom
...
```

n-2

We can use the standard error to estimate the 95% confidence interval as:

- $\left(\hat{\beta}_1 - t_{n-2,0.975} SE(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,0.975} SE(\hat{\beta}_1)\right) = b_1 \pm t_{n-2,0.975} SE(b_1) = b_1 \pm t_{99,0.975} SE(b_1)$

- In our housing example, the 95% confidence interval for the coefficient of `BuildingArea` is [9622.6968, 12796.3032]

Confidence intervals of regression coefficients

- More directly on R code:

```
confint(lm.fit)
```

```
                     2.5 %     97.5 %
(Intercept)   -311587.233 52619.18
BuildingArea     9622.491 12796.50
```

- Is exact, no precision lost to rounding error and easy to change

```
confint(lm.fit, level = 0.99)
```

```
                  0.5 %     99.5 %
(Intercept)  -370524.63 111556.57
BuildingArea    9108.86  13310.13
```

Is `BuildingArea` a good predictor of price?

- Refer to the code for summary(lm.fit)
  - Linear regression assumes $Y = \beta_0 + \beta_1 X + \varepsilon$
  - If `BuildingArea` is not linearly related to `Price`, then $\beta_1 = 0$
  - Can conduct a test of significance $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

$$t = \frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)} \overset{H_0}{=} \frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}$$

- The p-values for each significance test in the last column
- Recall, p-value gives the probability of observing your test statistic (and other scenarios support H_1), assuming H_0 is true
- Small p-value here gives very little evidence to support the claim that there is no relationship between `Price` and `BuildingArea`

Estimating the price of a 100 sum house in St Kilda

```
new.100 <- data.frame(BuildingArea = 100)
predict(lm.fit, new.100, interval = "confidence")
```

```
       fit      lwr     upr
1 991465.5 894562.7 1088368
```

```
predict(lm.fit, new.100, interval = "prediction")
```

```
       fit      lwr     upr
1 991465.5 13820.26 1969111
```

Extending Simple Linear Regression

Goodness of fit statistic

- Goodness of fit is measured by the coefficient of determination or R^2

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

- It measures the proportion of variation in the response Y, explained by the linear regression on X
  - A value of 0 indicates none of the variance in Y can be explained linearly by X
  - A value of 1 indicates all of the variance in Y can be explained linearly by X

Multiple regression with `lm`

```
multi.lm <- lm(Price/1000 ~ Type + BuildingArea, data = st.kilda.data)
summary(multi.lm)
```

```
Call:
lm(formula = Price/1000 ~ Type + BuildingArea, data = st.kilda.data)

Residuals:
   Min     1Q Median     3Q    Max
-700.3 -173.1  -65.9   18.6 3389.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   342.865    186.764   1.836  0.06945 .
Typet        -613.953    286.272  -2.145  0.03448 *
Typeu        -408.417    139.915  -2.919  0.00436 **
BuildingArea    9.533      1.014   9.398 2.68e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 469.5 on 97 degrees of freedom
Multiple R-squared:  0.6989,	Adjusted R-squared:  0.6896
F-statistic: 75.06 on 3 and 97 DF,  p-value: < 2.2e-16
```

$$\widehat{Price} = \begin{cases} 342 + 9.5\,Area, & \text{if house} \\ 342 - 613 + 9.5\,Area, & \text{if town house} \\ 342 - 408 + 9.5\,Area, & \text{if unit} \end{cases}$$

Interpreting Regression Models

- Simple case

  $$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

  - β_1: the average change in Y for each unit increase in X_1
  - β_0: the average of Y when $X_1 = 0$

- Multiple regression case

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

  - β_p: The average change in Y for each single unit increase in X_p, holding all the other predictors fixed