Introduction to Basic R: Statistical Analysis

Introduction to Bioinformatics

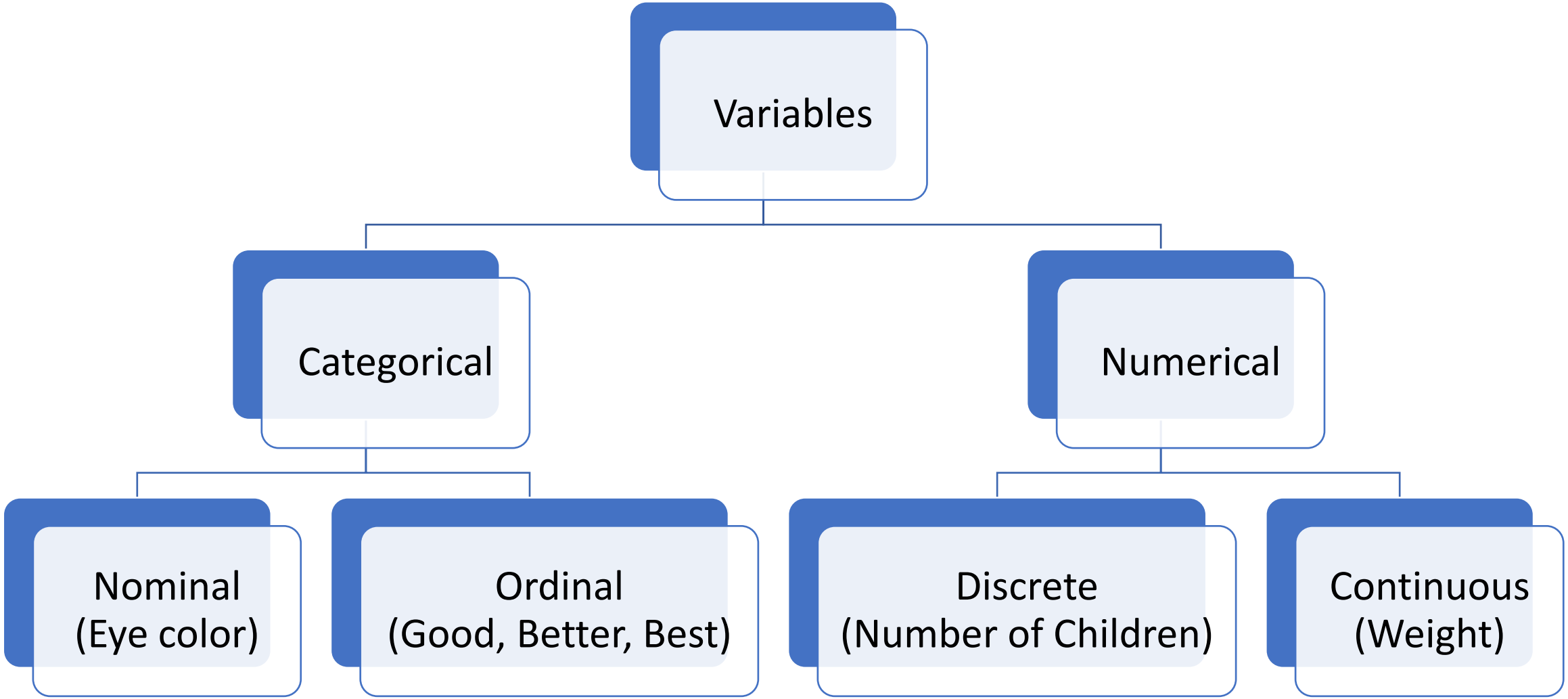
Presented by Tran Ba Thien

March 12 2023

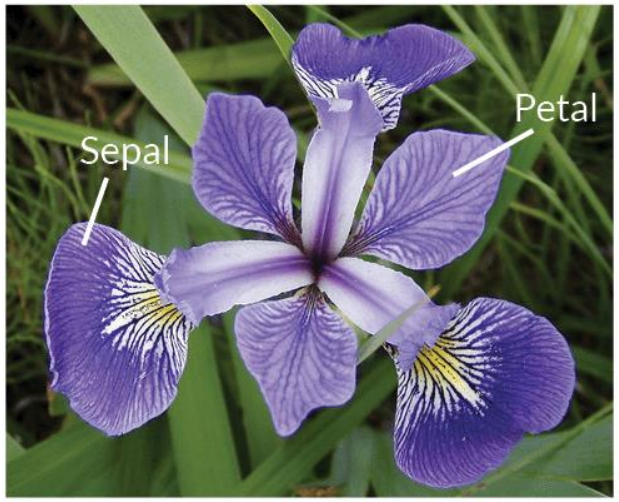
Content

- Data in the nutshell
- Descriptive statistics
  - Variance
  - Covariance
  - Correlation
- Data distribution
- Inferential statistics
  - Parametric test: t-test and ANOVA
  - Non-Parametric test: Wilcoxon and Kruskal-Wallis
- Linear Regression

Types of Variables



Iris data in the nutshell

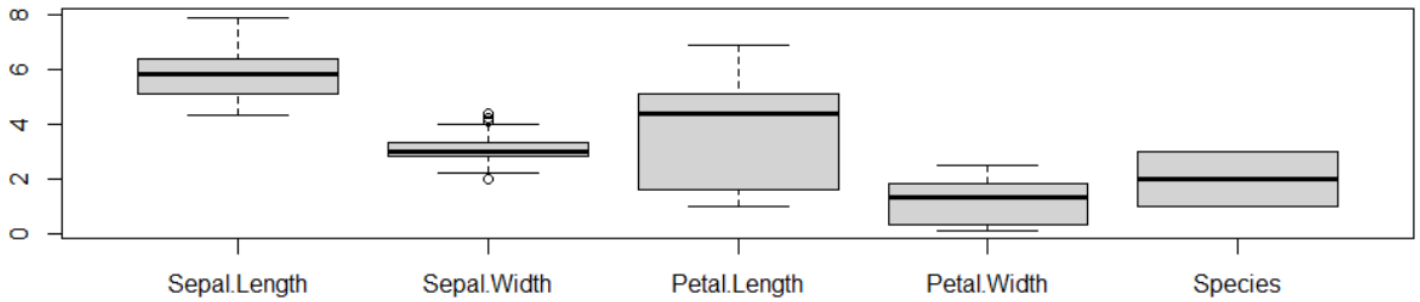


Iris Versicolor

```
> dt2 <- iris
> str(dt2)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> summary(dt2)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
> boxplot(dt2)
```

Iris Setosa

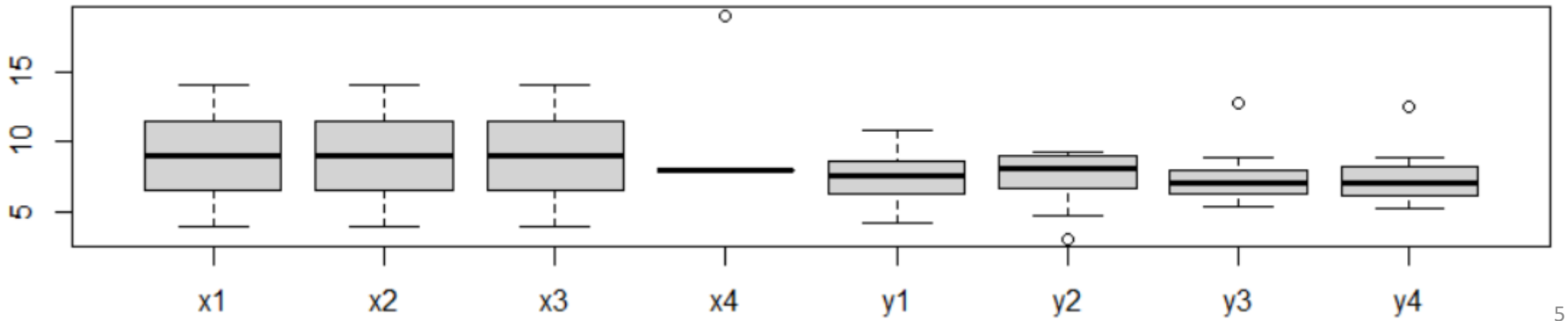
Iris Virginica



Data Science Example - Iris dataset (inpe.br)

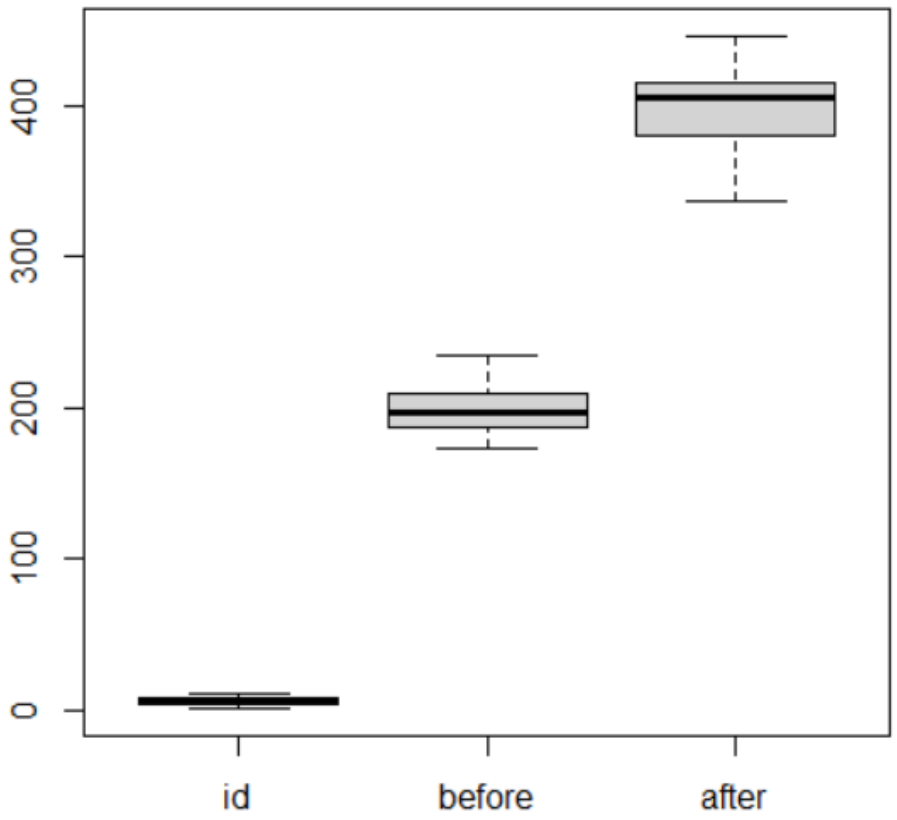
Anscombe data in the nutshell

```
> dt3 <- anscombe
> str(dt3)
'data.frame':   11 obs. of  8 variables:
 $ x1: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x2: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x3: num  10 8 13 9 11 14 6 4 12 7 ...
 $ x4: num  8 8 8 8 8 8 8 19 8 8 ...
 $ y1: num  8.04 6.95 7.58 8.81 8.33 ...
 $ y2: num  9.14 8.14 8.74 8.77 9.26 8.1 6.13 3.1 9.13 7.26 ...
 $ y3: num  7.46 6.77 12.74 7.11 7.81 ...
 $ y4: num  6.58 5.76 7.71 8.84 8.47 7.04 5.25 12.5 5.56 7.91 ...
> summary(dt3)
       x1              x2              x3              x4              y1              y2              y3              y4
 Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8   Min.   : 4.260   Min.   :3.100   Min.   : 5.39   Min.   : 5.250
 1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8   1st Qu.: 6.315   1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170
 Median : 9.0   Median : 9.0   Median : 9.0   Median : 8   Median : 7.580   Median :8.140   Median : 7.11   Median : 7.040
 Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9   Mean   : 7.501   Mean   :7.501   Mean   : 7.50   Mean   : 7.501
 3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8   3rd Qu.: 8.570   3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190
 Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19   Max.   :10.840   Max.   :9.260   Max.   :12.74   Max.   :12.500
> boxplot(dt3)
```

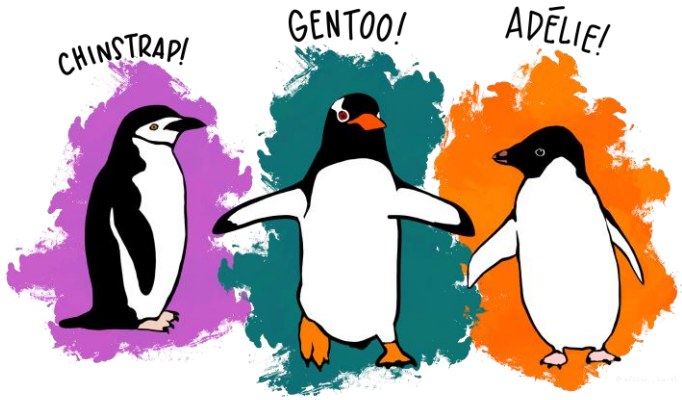


mice2 data in the nutshell

```
> library(datarium)
> data("mice2", package = "datarium")
> head(mice2, 3)
  id before after
1  1  187.2 429.5
2  2  194.2 404.4
3  3  231.7 405.6
> str(mice2)
'data.frame':   10 obs. of  3 variables:
 $ id    : int  1 2 3 4 5 6 7 8 9 10
 $ before: num  187 194 232 200 202 ...
 $ after : num  430 404 406 397 378 ...
> summary(mice2)
       id            before          after      
 Min.   : 1.00   Min.   :172.4   Min.   :337.0  
 1st Qu.: 3.25   1st Qu.:187.8   1st Qu.:384.5  
 Median : 5.50   Median :197.3   Median :405.0  
 Mean   : 5.50   Mean   :200.6   Mean   :400.0  
 3rd Qu.: 7.75   3rd Qu.:206.9   3rd Qu.:412.8  
 Max.   :10.00   Max.   :235.0   Max.   :445.8  
> boxplot(mice2)
```

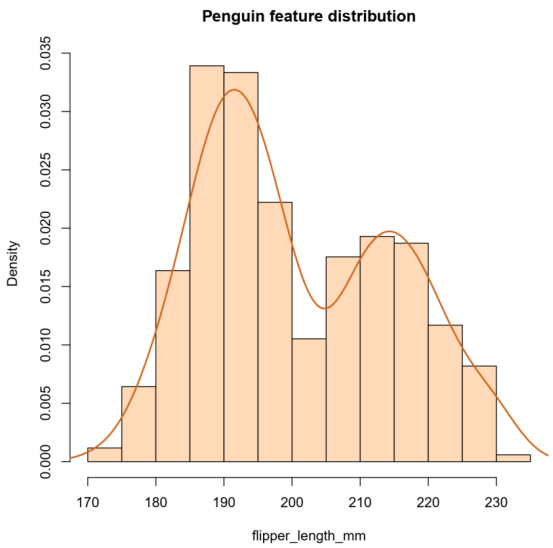


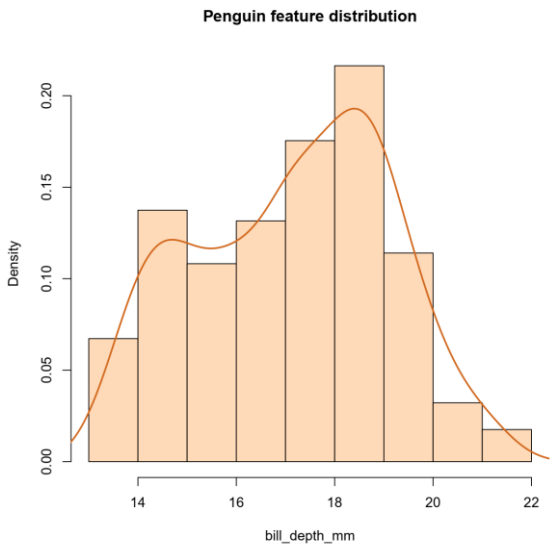
Palmerpenguins data in the nutshell

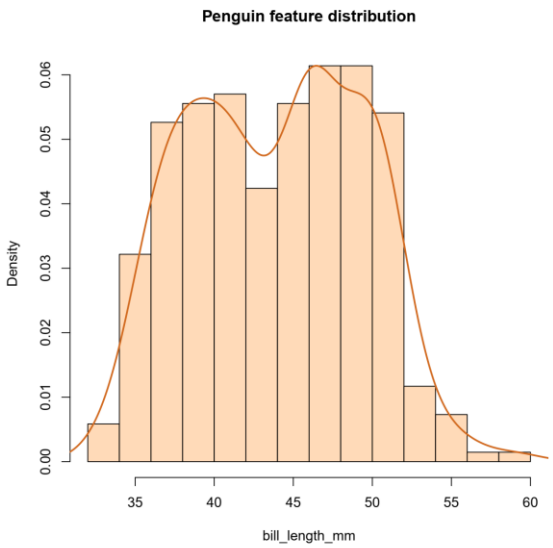


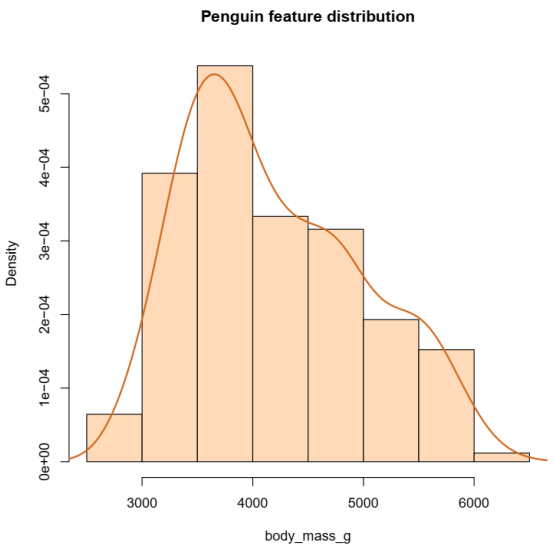

```
> library(palmerpenguins)
> head(penguins)
# A tibble: 6 × 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex     year
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int> <fct>  <int>
1 Adelie  Torgersen           39.1          18.7               181        3750 male    2007
2 Adelie  Torgersen           39.5          17.4               186        3800 female  2007
3 Adelie  Torgersen           40.3          18                 195        3250 female  2007
4 Adelie  Torgersen           NA            NA                  NA          NA NA      2007
5 Adelie  Torgersen           36.7          19.3               193        3450 female  2007
6 Adelie  Torgersen           39.3          20.6               190        3650 male    2007
> str(penguins)
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
> dt <- as.data.frame(penguins)
> dt$year <- as.factor(dt$year)
> summary(dt)
      species          island    bill_length_mm  bill_depth_mm   flipper_length_mm  body_mass_g       sex         year
 Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10   Min.   :172.0     Min.   :2700   female:165   2007:110
 Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0     1st Qu.:3550   male  :168   2008:114
 Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30   Median :197.0     Median :4050   NA's  : 11   2009:120
                                 Mean   :43.92   Mean   :17.15   Mean   :200.9     Mean   :4202
                                 3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0     3rd Qu.:4750
                                 Max.   :59.60   Max.   :21.50   Max.   :231.0     Max.   :6300
                                 NA's   :2       NA's   :2       NA's   :2         NA's   :2
```

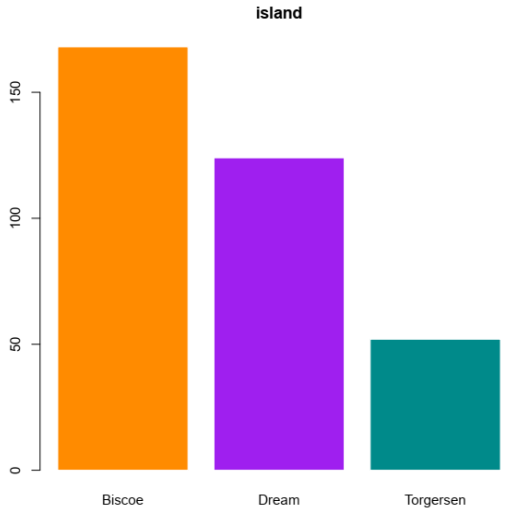
Palmerpenguins data in the nutshell

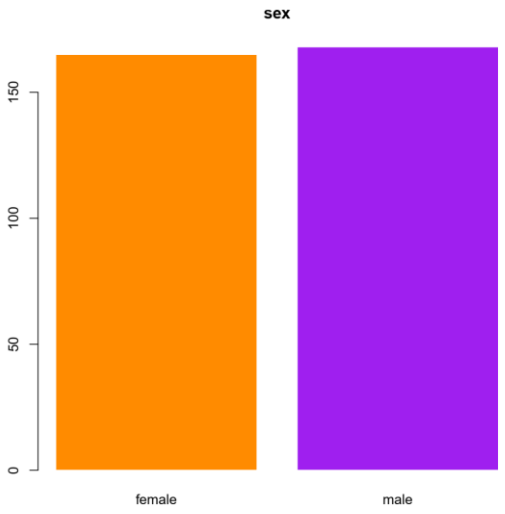


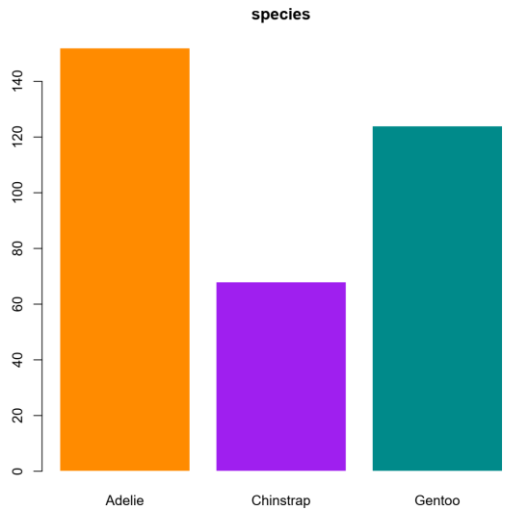


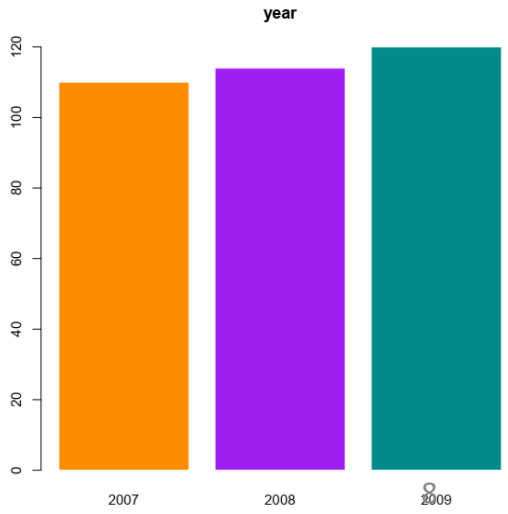






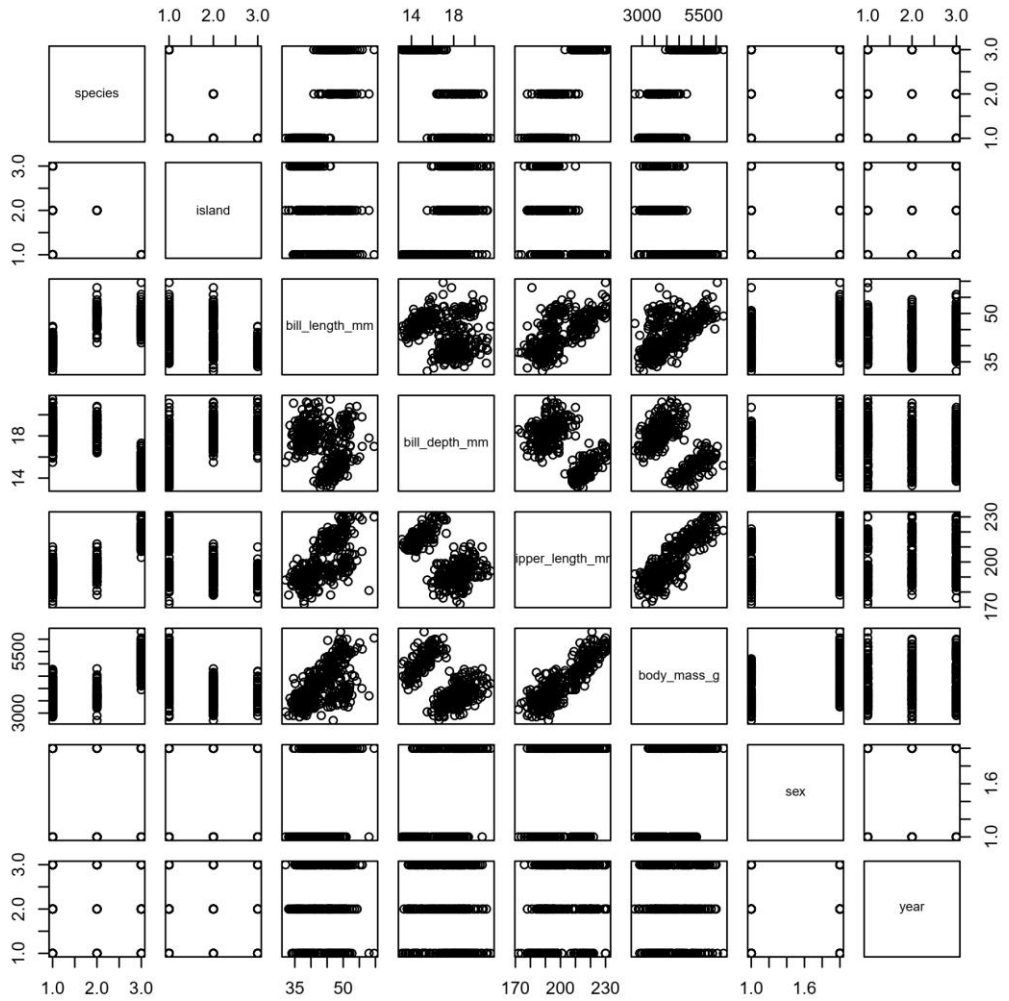






Palmerpenguins data in the nutshell

```
> pdf(file=file.path(paste0(plot_dir, "Penguin_pairs.pdf")))
> pairs(dt)
> dev.off()
null device
          1
```

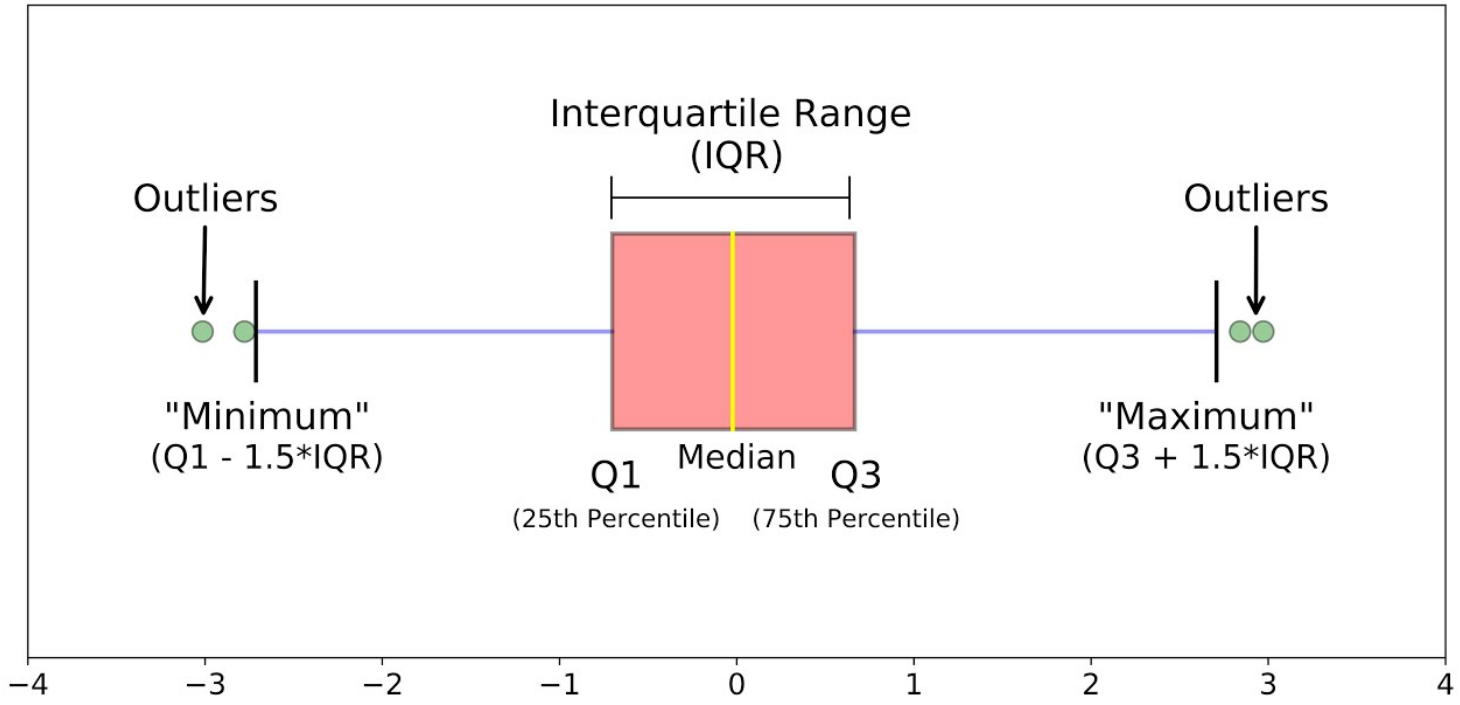



```
> for (i in names(dt[, 3:6])) {
+   print(i)
+   pdf(file=file.path(paste0(plot_dir, "Penguin_", i, ".pdf")))
+   # Histogram
+   hist(dt[,i],
+        col="peachpuff",
+        border="black",
+        prob = TRUE,
+        xlab = i,
+        main = "Penguin feature distribution")
+   # Density
+   lines(density(dt[,i], na.rm=TRUE),
+         lwd = 2,
+         col = "chocolate")
+   dev.off()
+ }
[1] "bill_length_mm"
[1] "bill_depth_mm"
[1] "flipper_length_mm"
[1] "body_mass_g"
```

```
> for (u in names(dt[, c(1:2,7:8)])) {
+   print(u)
+   pdf(file=file.path(paste0(plot_dir, "Penguin_", u, ".pdf")))
+   # Barplot
+   barplot(table(dt[, u]),
+           border="white",
+           col = c("#56B4E9", "#E69F00", "lightblue"),
+           main = u)
+   dev.off()
+ }
[1] "species"
[1] "island"
[1] "sex"
[1] "year"
```

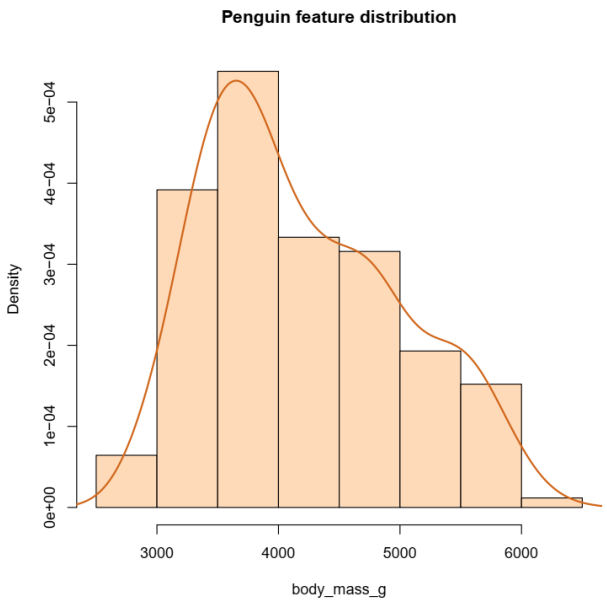
Basic descriptive statistics

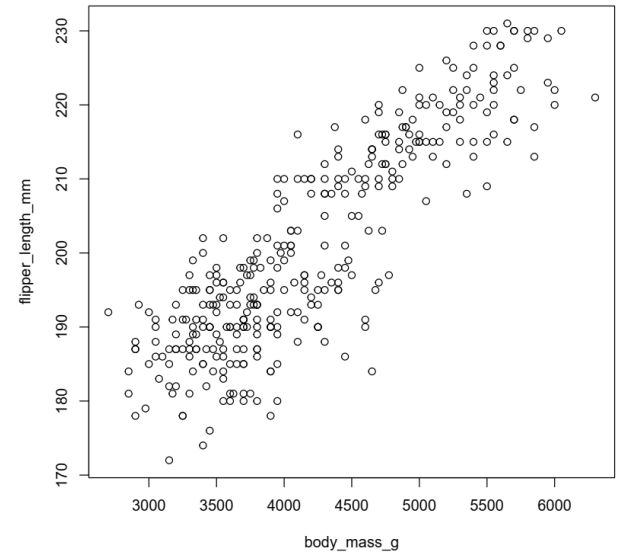
Min	min()
Q1	quantile(…, p=0.25)
Mean	mean()
Median (Q2)	median() / quantile(…, p=0.50)
Q3	quantile(…, p=0.75)
Max	max()
Standard deviation	sd()

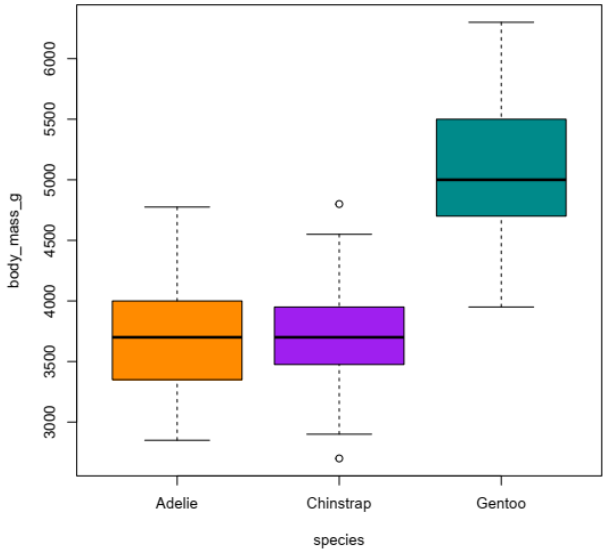


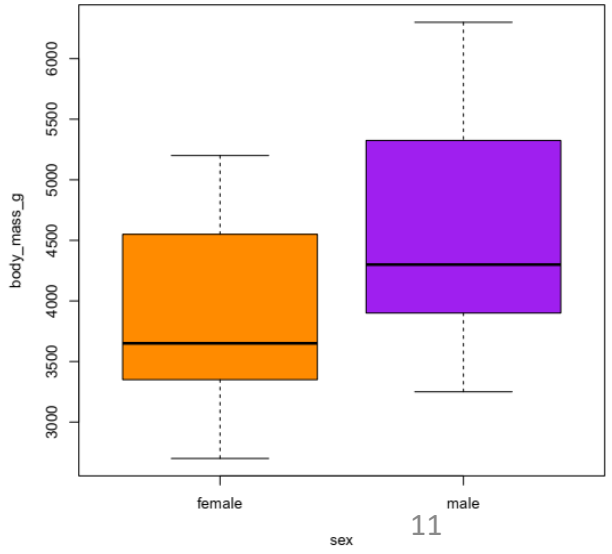
Penguins question!!!

- How well distributed is the data?
- What is the relationship between the variables?

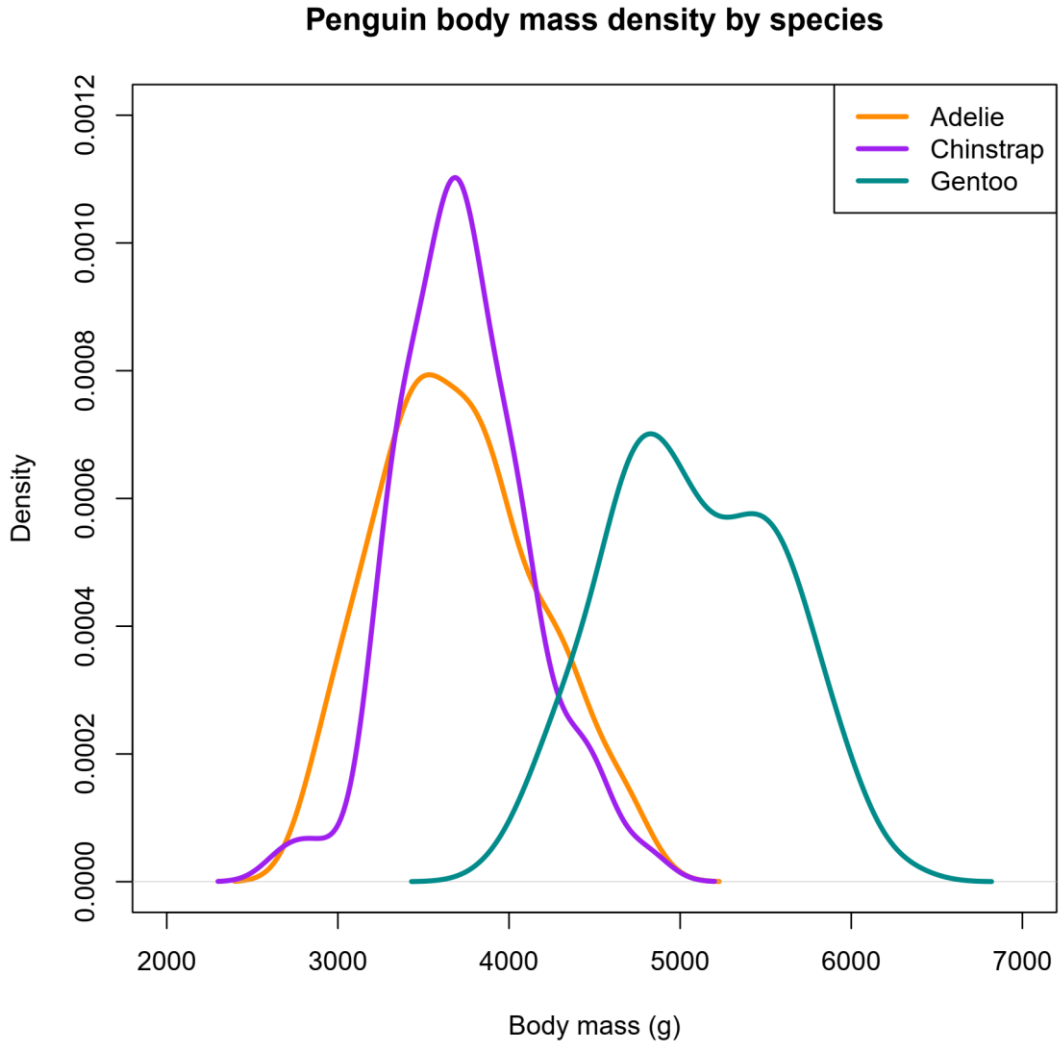








Variance (σ^2)



- Measure the dispersion of a set of data values
- Measure how far each number in a data set is from the mean
- Formula:

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

S^2 = sample variance

x_i = value of one observation

$\bar{x}$ = mean of all observations

n = number of observations

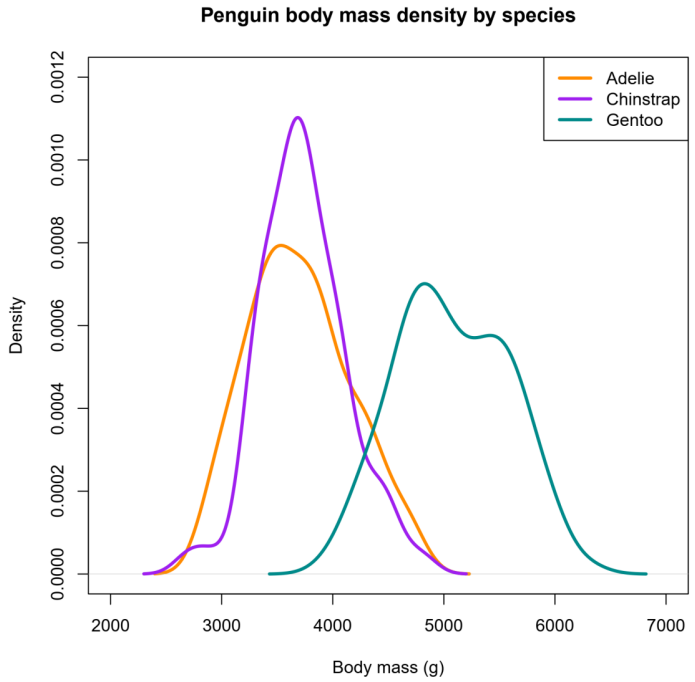
What Is Variance in Statistics? Definition, Formula, and Example (investopedia.com)

Variance (σ^2)

Syntax:

var(x, y = NULL, na.rm = FALSE, use)

```
> var_mass <- var(dt$body_mass_g)
> var_mass
[1] NA
> table(is.na(dt$body_mass_g))

FALSE  TRUE
  342     2
> var_mass <- var(dt$body_mass_g, na.rm=T)
> var_mass
[1] 643131.1
> cat("Variance of penguins body mass: ", var_mass, "\n")
Variance of penguins body mass:  643131.1
> sd(dt$body_mass_g, na.rm=T)
[1] 801.9545
> sqrt(var_mass)
[1] 801.9545

> var_mass_flipper <- var(dt$body_mass_g,
+                         dt$flipper_length_mm,
+                         na.rm=T)
> var_mass_flipper
[1] 9824.416
```




```
> tapply(dt$body_mass_g, dt$species, function(x) {
+   variances <- var(x, na.rm=T)
+   return(variances)
+   })
   Adelie Chinstrap    Gentoo
 210282.9  147713.5  254133.2
```

Covariance

- Measure the degree of correlation between two random variables
- Shows the overall volatility between two variables.
- Can take any value from -∞ to +∞
- Formula:

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$cov_{x,y}$ = sample covariance

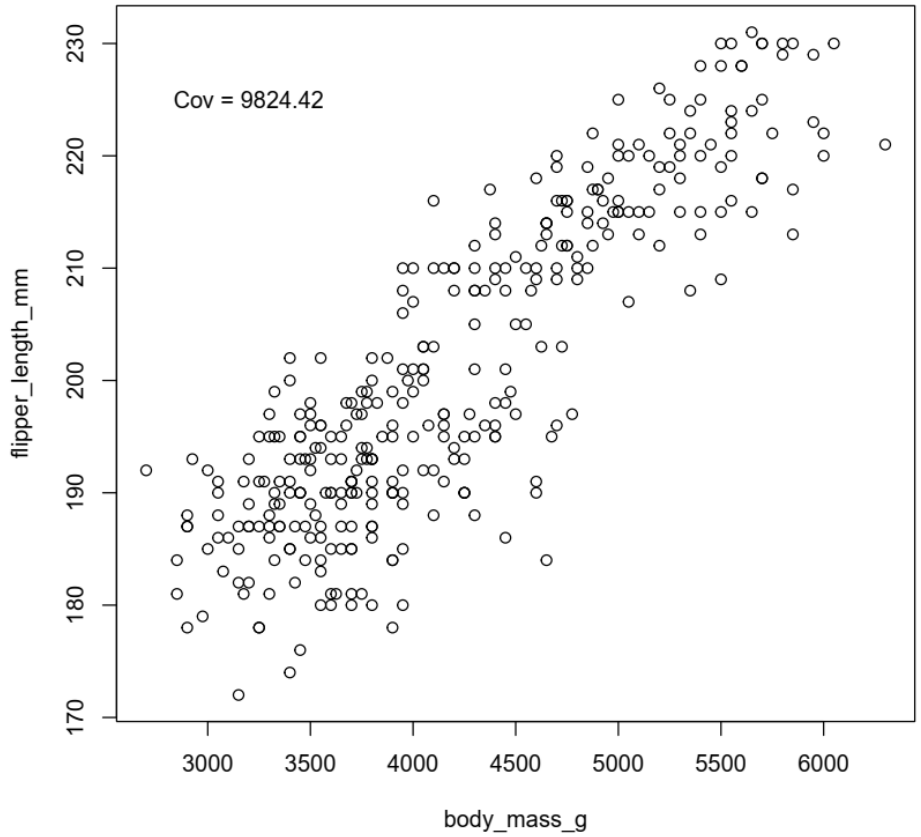
x_i = value of one x observation

y_i = value of one y observation

$\bar{x}$ = mean of all x observations

$\bar{y}$ = mean of all y observations

n = number of observations



Covariance: Formula, Definition, Types, and Examples (investopedia.com)

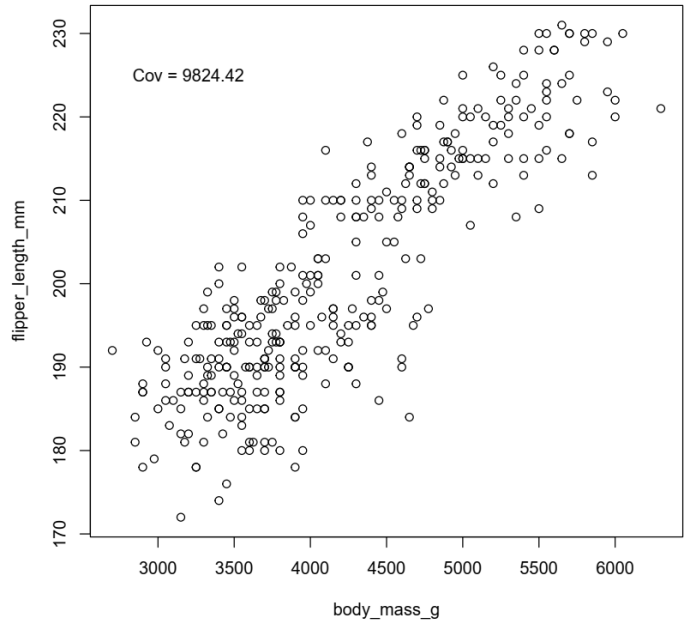
Covariance

Syntax:

cov(x, y = NULL, use = "everything",

method = c("pearson", "kendall", "spearman"))

```
> cov_mass_flipper <- cov(dt$body_mass_g,
+                         dt$flipper_length_mm,
+                         na.rm=T)
Error in cov(dt$body_mass_g, dt$flipper_length_mm, na.rm = T) :
  unused argument (na.rm = T)
> table(is.na(dt$body_mass_g), is.na(dt$flipper_length_mm))

       FALSE TRUE
  FALSE   342    0
  TRUE      0    2
> cov_mass_flipper <- cov(dt$body_mass_g,
+                         dt$flipper_length_mm,
+                         use="complete.obs")
> cov_mass_flipper
[1] 9824.416
> var_mass_flipper
[1] 9824.416
> cat("Covariance of penguins body mass and flipper length: ", cov_mass_flipper, "\n")
Covariance of penguins body mass and flipper length:  9824.416
```



Correlation

- Describe the relationship between two variables
- Show how much the values of one variable are associated with the values of another variable
- Can take value from -1 to +1
- Formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

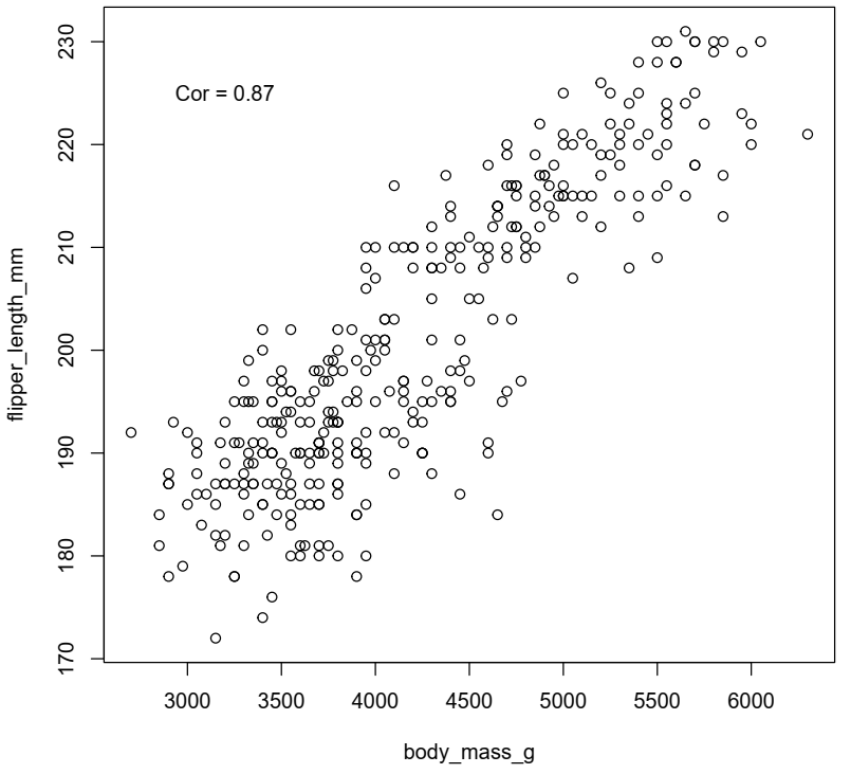
r = sample correlation

x_i = value of one x observation

y_i = value of one y observation

$\bar{x}$ = mean of all x observations

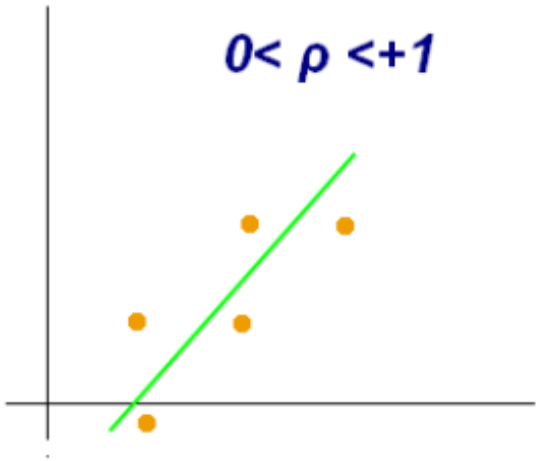
$\bar{y}$ = mean of all y observations

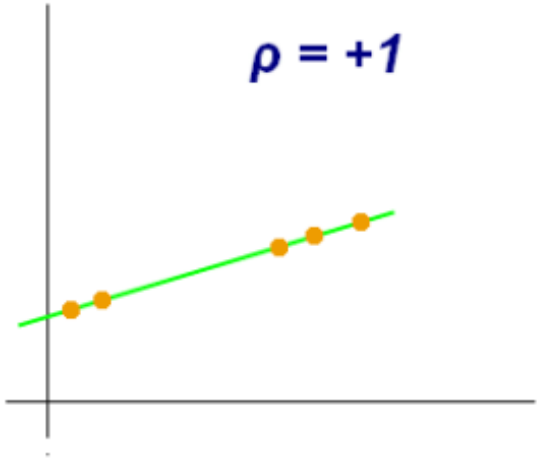


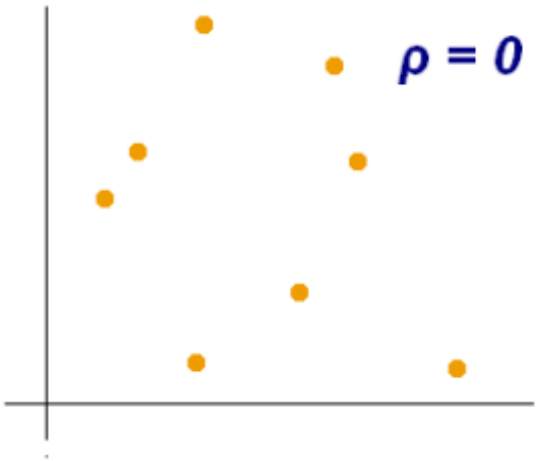
Correlation: What It Means in Finance and the Formula for Calculating It (investopedia.com)

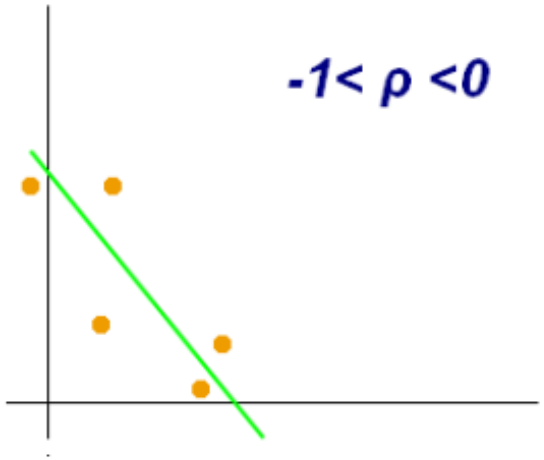
Covariance vs Correlation | Top 5 Differences (with Infographics) (wallstreetmojo.com)

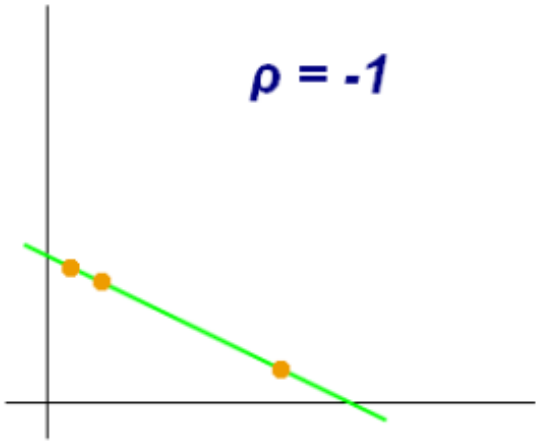
Correlation









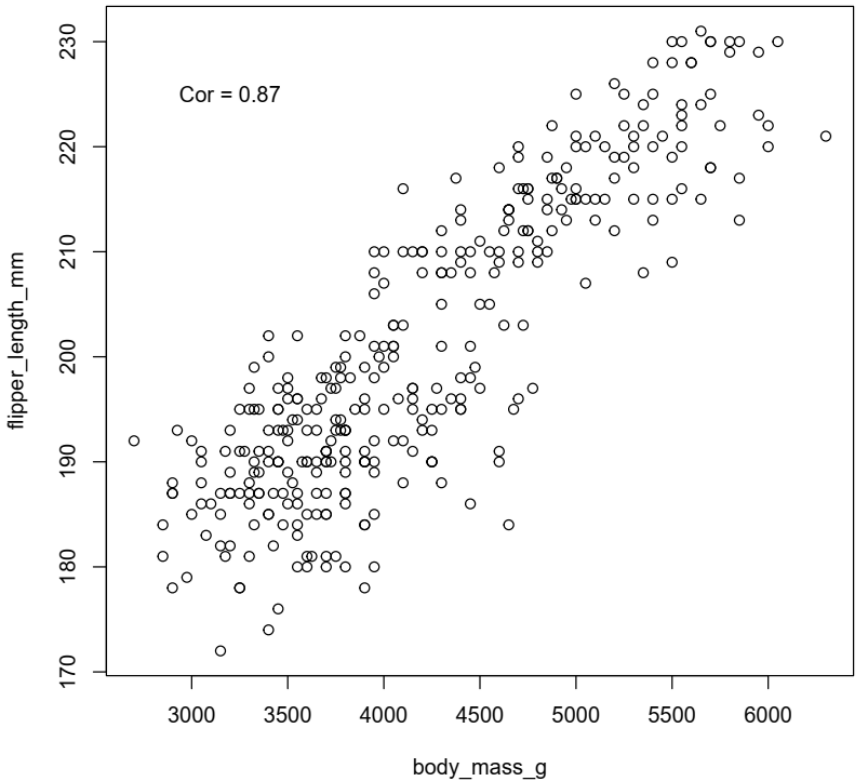


Correlation

Syntax:

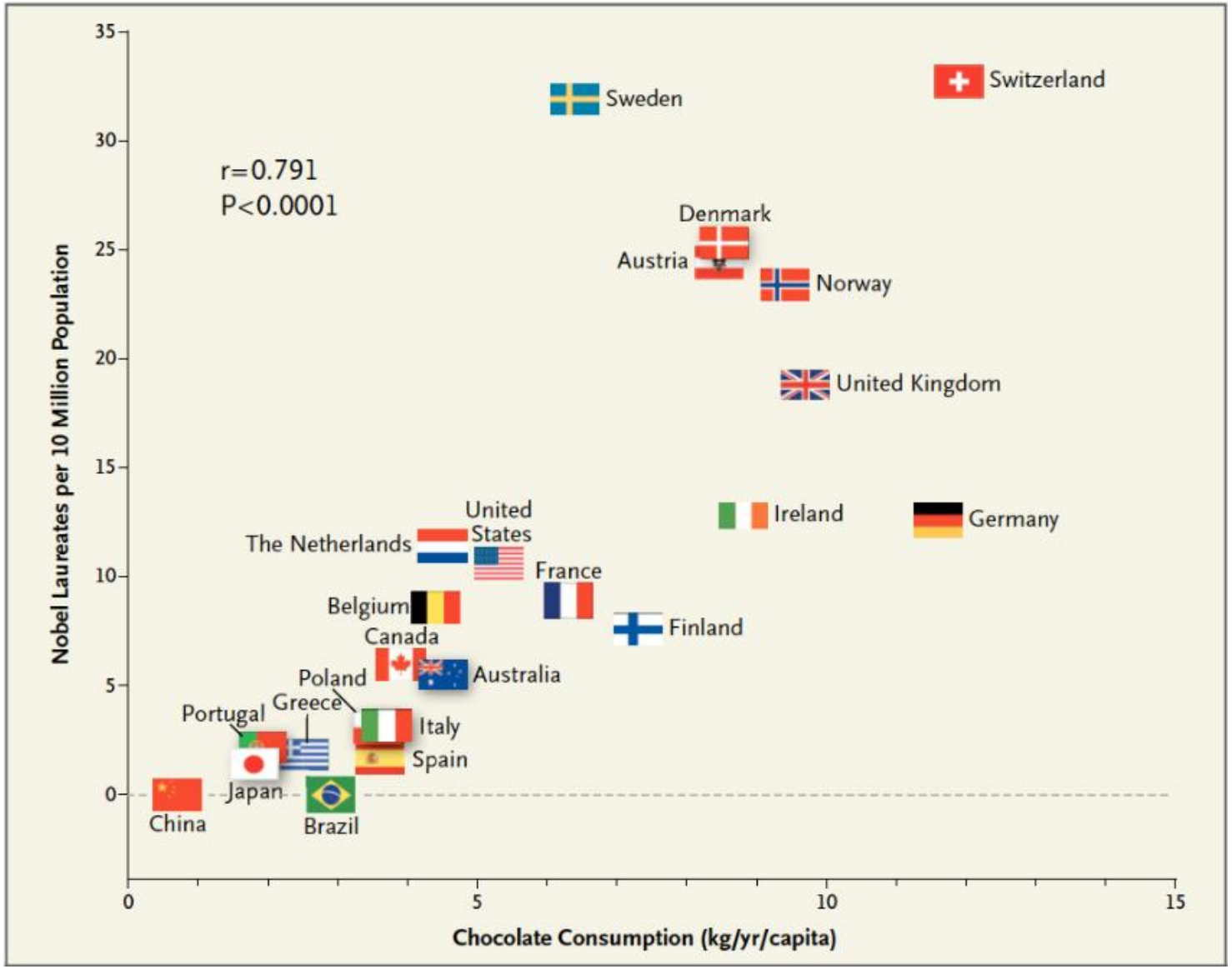
cor(x, y = NULL, use = "everything",

method = c("pearson", "kendall", "spearman"))




```
> cor_mass_flipper <- cor(dt$body_mass_g,
+                         dt$flipper_length_mm,
+                         use="complete.obs")
> cor_mass_flipper
[1] 0.8712018
> cov_mass_flipper/(sd(dt$body_mass_g, na.rm=T)*sd(dt$flipper_length_mm, na.rm=T))
[1] 0.8712018
> cat("Correlation of penguins body mass and flipper length: ", cor_mass_flipper, "\n")
Correlation of penguins body mass and flipper length:  0.8712018
```

Correlation indicates association, not causation

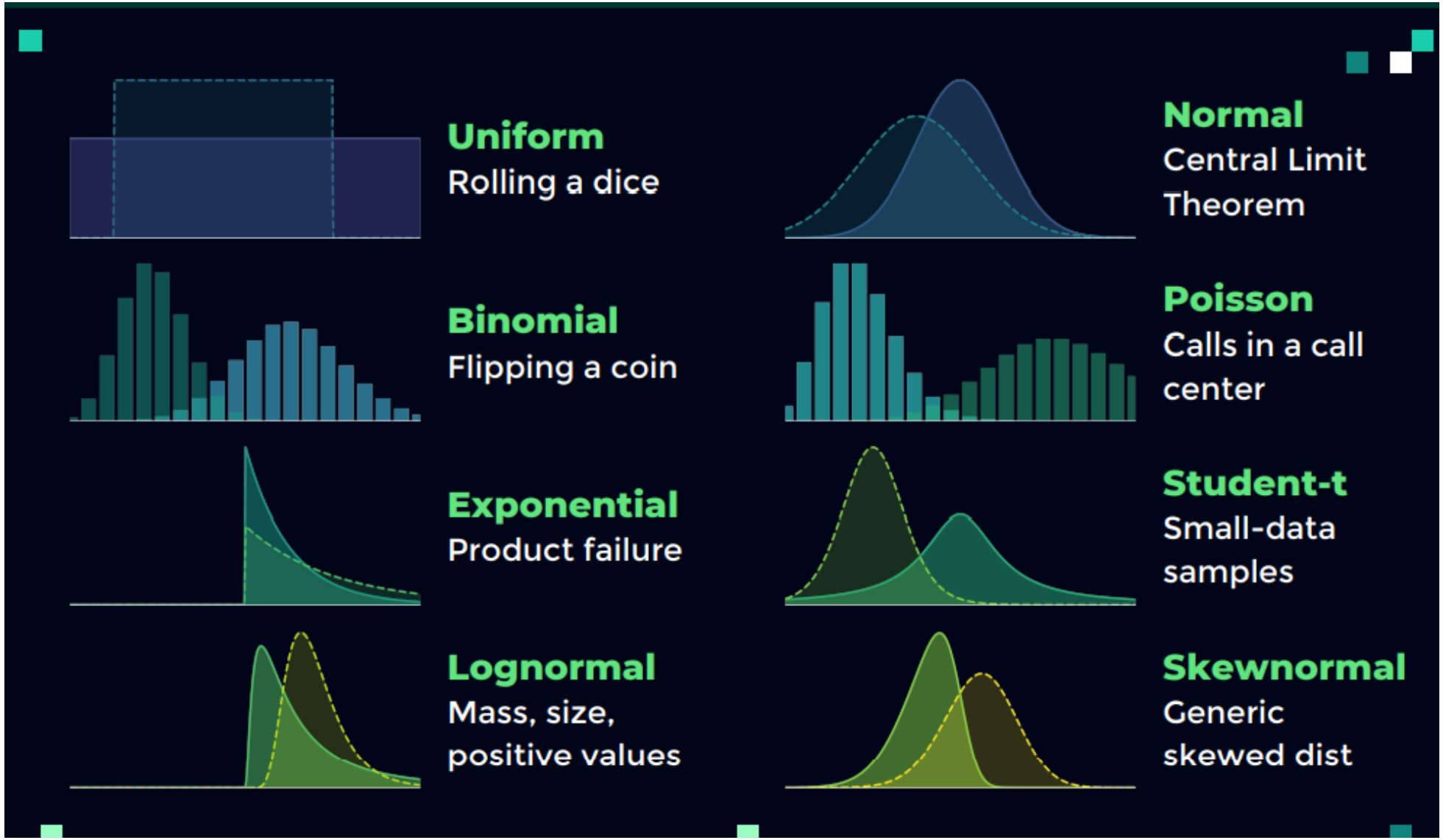


http://www.nejm.org/doi/full/10.1056/NEJMon1211064

Classwork 1

1. If the law requires women to marry only men 2 years older than themselves, what is the correlation of the ages between all pairs of couples (husbands and wives)?
2. Use *Iris* data in R and do as requested:
   a. Visualize the *Sepal.Length*, *Sepal.Width* and *Species*.
   b. Calculate variance, covariance, correlation between *Sepal.Length* and *Sepal.Width*
   c. Calculate variance, covariance, correlation between *Sepal.Length* and *Species*
3. Use *anscombe* data in R and do as requested:
   a. Calculate the variance, covariance, and correlation between the *x* variable and the corresponding *y* variable. Give a comment.
   b. Use scatter plot to visualize the *x* variable and the corresponding *y* variable. Give a comment.

Data distribution



All distribution in statistics | Data Science and Machine Learning | Kaggle

When will we using parametric tests?

- The data are normally distributed or can be transformed to be normally distributed (Logarithmic, square root, inverse , box-cox transformation and quantile normalization)
- The variance of the population is the same for all groups being compared
- The data are independent and randomly sampled
- The dependent variable is continuous or at least ordinal

How to check if data is normally distributed?

➢Visual Method

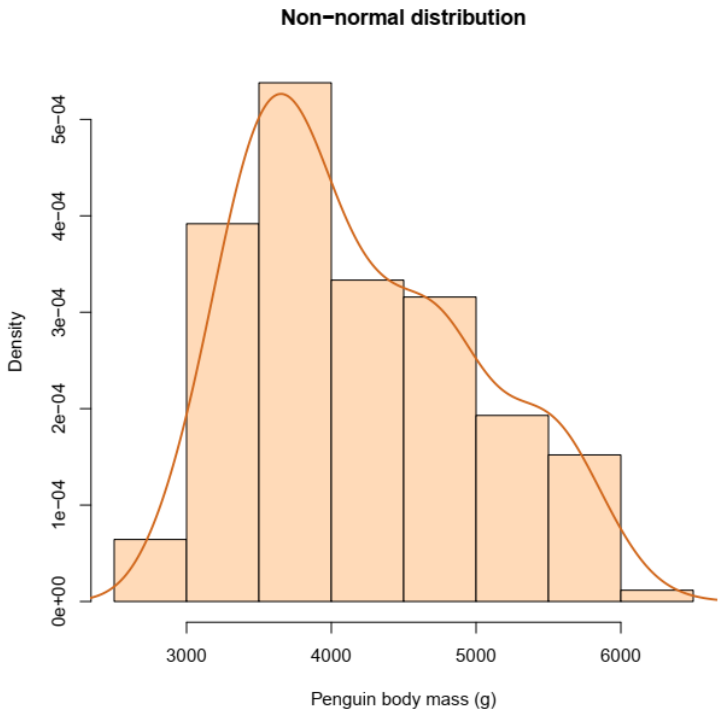
- Histogram
- Q-Q plot

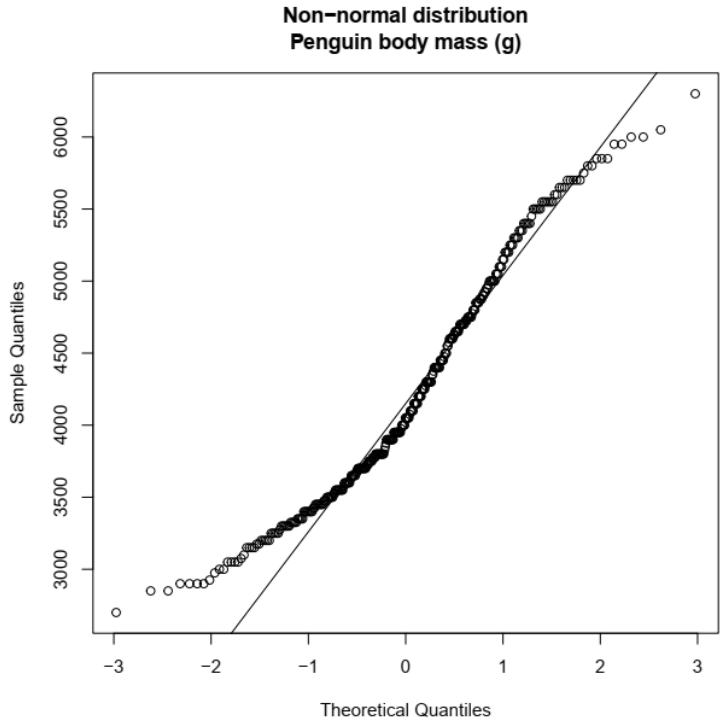
➢Statistical Test

- Shapiro-Wilk Test
- Kolmogorov-Smirnov Test

How to Test for Normality in R : Statistics in R : Data Sharkie
How to Test for Normality in R (4 Methods) – Statology

Check normality of penguins body mass






```
> shapiro.test(dt$body_mass_g)

        Shapiro-Wilk normality test

data:  dt$body_mass_g
W = 0.95921, p-value = 3.679e-08
```

```
> ks.test(dt$body_mass_g, "pnorm",
+         mean=mean(dt$body_mass_g, na.rm=T),
+         sd=sd(dt$body_mass_g, na.rm=T))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dt$body_mass_g
D = 0.10408, p-value = 0.00121
alternative hypothesis: two-sided
```

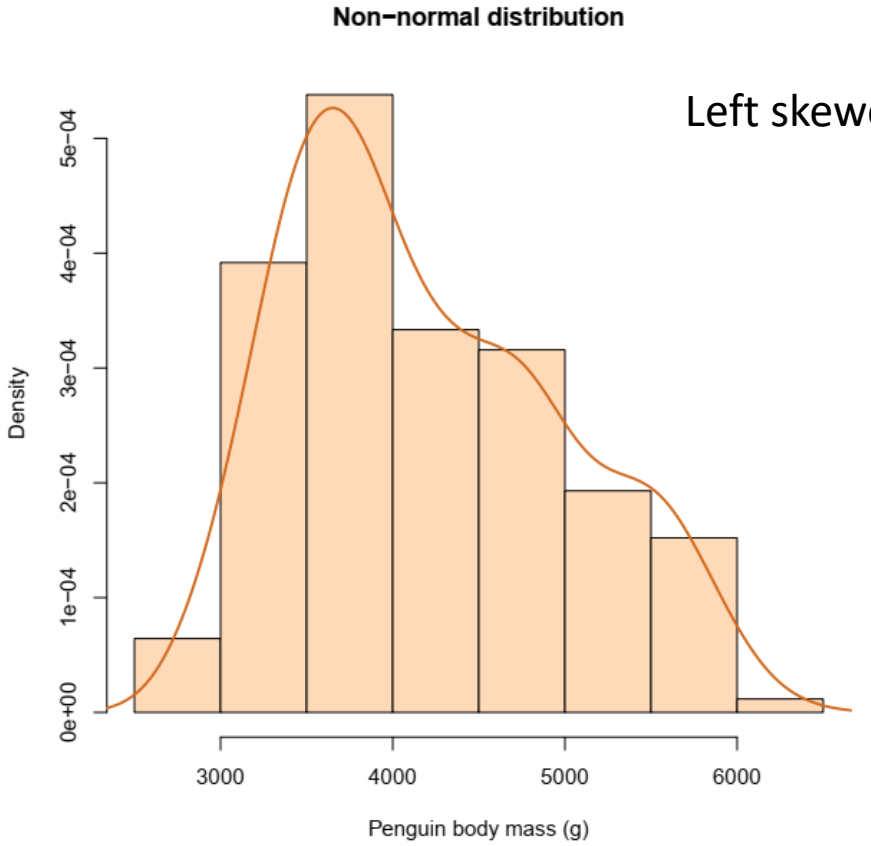
Simulate a normal distributed body mass

```
> dt$sim_mass <- NA
> mean_mass <- mean(dt$body_mass_g, na.rm=T)
> sd_mass <- sd(dt$body_mass_g, na.rm=T)
> p_values_sw <- rep(0, 9)
> p_values_ks <- rep(0, 9)
> iteration <- 0
> while (any(p_values_sw < 0.8) | any(p_values_ks < 0.8)) {
+   iteration <- iteration + 1
+   # cat(paste0("Number of iterations: ", iteration, "\n"))
+   dt$sim_mass <- rnorm(nrow(dt),
+                        mean = mean_mass,
+                        sd = sd_mass)
+
+   p_values_sw <- sapply(c("male", "female"), function(x) {
+     sapply(c("Adelie", "Chinstrap", "Gentoo"), function(y) {
+       shapiro.test(dt[!is.na(dt$sim_mass) & (dt$sex == x & dt$species == y), "sim_mass"])$p.value
+     })
+   })
+
+   p_values_ks <- sapply(c("male", "female"), function(x) {
+     sapply(c("Adelie", "Chinstrap", "Gentoo"), function(y) {
+       ks.test(dt[!is.na(dt$sim_mass) & (dt$sex == x & dt$species == y), "sim_mass"], "pnorm",
+               mean = mean(dt[dt$sex == x & dt$species == y, "sim_mass"], na.rm = TRUE),
+               sd = sd(dt[dt$sex == x & dt$species == y, "sim_mass"], na.rm = TRUE))$p.value
+     })
+   })
+ }
> cat(paste0("Reach condition at iterations: ", iteration, "\n"))
Reach condition at iterations: 21743
```

[A Guide to dnorm, pnorm, rnorm, and qnorm in R - GeeksforGeeks]

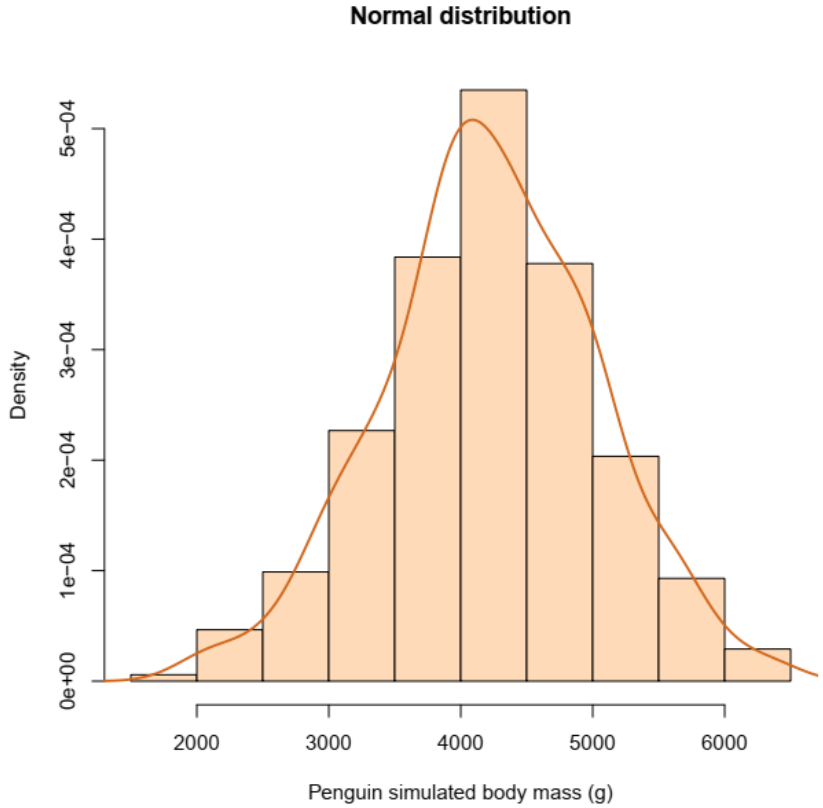
Histogram normality check

```
> hist(dt$body_mass_g,
+      col="peachpuff",
+      border="black",
+      prob = TRUE,
+      xlab = "Penguin body mass (g)",
+      main = "Non-normal distribution")
> lines(density(dt$body_mass_g, na.rm=TRUE),
+       lwd = 2,
+       col = "chocolate")
```



Left skewed

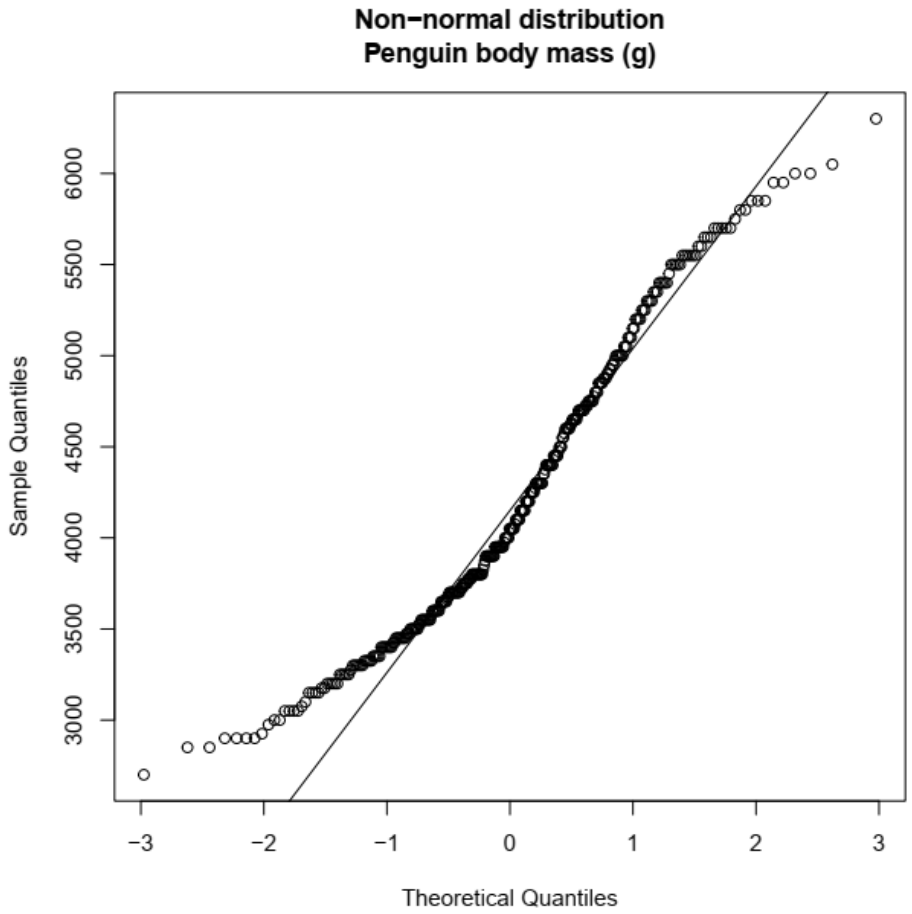
```
> hist(dt$sim_mass,
+      col="peachpuff",
+      border="black",
+      prob = TRUE,
+      xlab = "Penguin simulated body mass (g)",
+      main = "Normal distribution")
> lines(density(dt$sim_mass, na.rm=TRUE),
+       lwd = 2,
+       col = "chocolate")
```



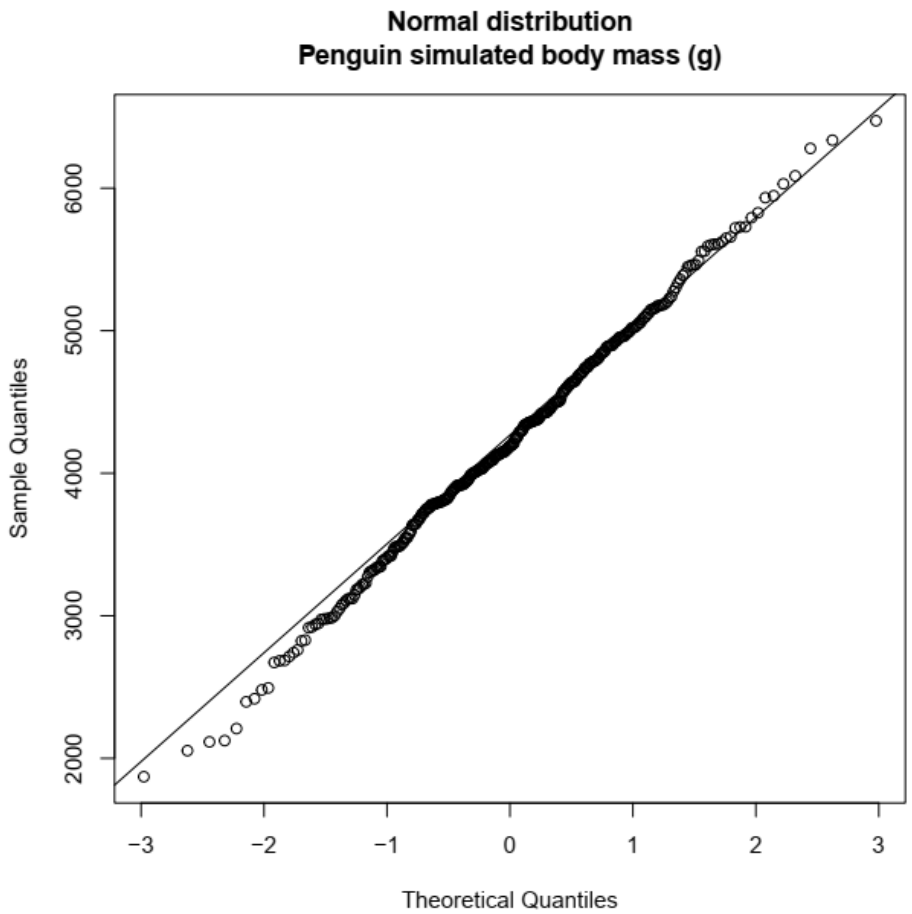
Symmetric

Q-Q plot normality check

```
> qqnorm(dt$body_mass_g, main='Non-normal distribution\nPenguin body mass (g)')
> qqline(dt$body_mass_g)
```




```
> qqnorm(dt$sim_mass, main='Normal distribution\nPenguin simulated body mass (g)')
> qqline(dt$sim_mass)
```



Statistical test normality check

Non-normal distribution

Normal distribution

Shapiro-Wilk Test

```
> shapiro.test(dt$body_mass_g)

        Shapiro-Wilk normality test

data:  dt$body_mass_g
W = 0.95921, p-value = 3.679e-08
```

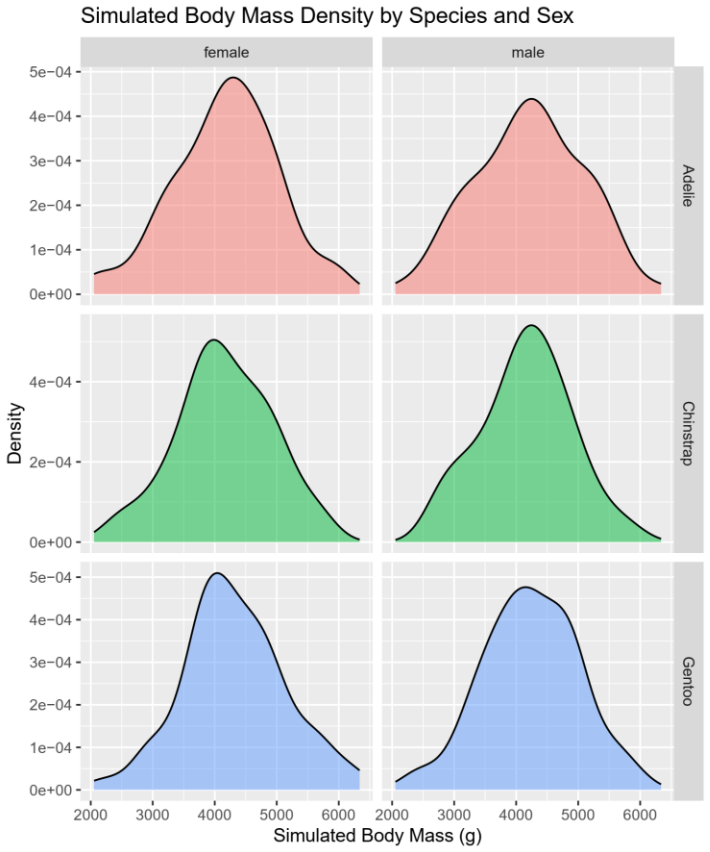
```
> shapiro.test(dt$sim_mass)

        Shapiro-Wilk normality test

data:  dt$sim_mass
W = 0.99749, p-value = 0.883
```

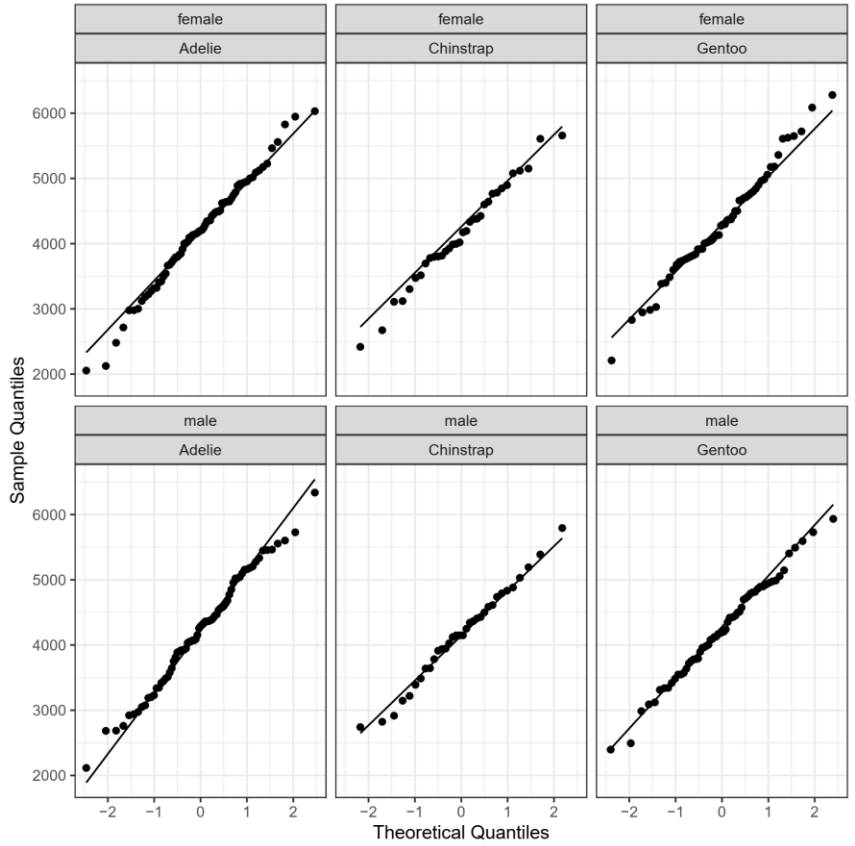
Kolmogorov-Smirnov Test

```
> ks.test(dt$body_mass_g, "pnorm",
+         mean=mean(dt$body_mass_g, na.rm=T),
+         sd=sd(dt$body_mass_g, na.rm=T))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dt$body_mass_g
D = 0.10408, p-value = 0.00121
alternative hypothesis: two-sided
```

```
> ks.test(dt$sim_mass, "pnorm",
+         mean=mean(dt$sim_mass),
+         sd=sd(dt$sim_mass))

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  dt$sim_mass
D = 0.036344, p-value = 0.7538
alternative hypothesis: two-sided
```

Check normality for each combination

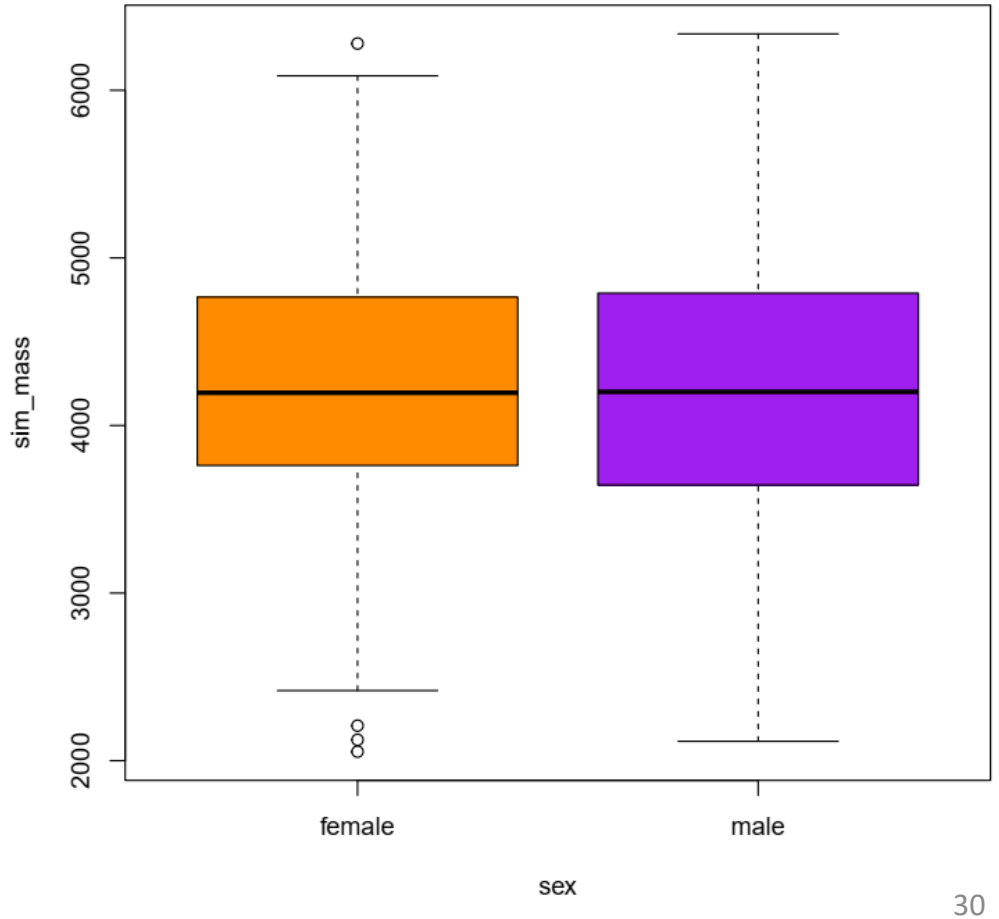





```
> sapply(c("male", "female"), function(x) {
+   sapply(c("Adelie", "Chinstrap", "Gentoo"), function
(y) {
+     shapiro.test(dt[!is.na(dt) & (dt$sex == x & dt$spe
cies == y), "sim_mass"])$p.value
+   })
+ })
              male    female
Adelie    0.8832532 0.8311959
Chinstrap 0.9607401 0.9540578
Gentoo    0.9646967 0.8879197
> sapply(c("male", "female"), function(x) {
+   sapply(c("Adelie", "Chinstrap", "Gentoo"), function
(y) {
+     ks.test(dt[!is.na(dt$sim_mass) & (dt$sex == x & dt
$species == y), "sim_mass"], "pnorm",
+             mean = mean(dt[dt$sex == x & dt$species ==
y, "sim_mass"], na.rm = TRUE),
+             sd = sd(dt[dt$sex == x & dt$species == y,
"sim_mass"], na.rm = TRUE))$p.value
+   })
+ })
              male    female
Adelie    0.9599492 0.9188871
Chinstrap 0.9833130 0.9809331
Gentoo    0.9827294 0.9219875
```

Penguins question!!!

- Is there any statistical different simulated mass between penguins gender?

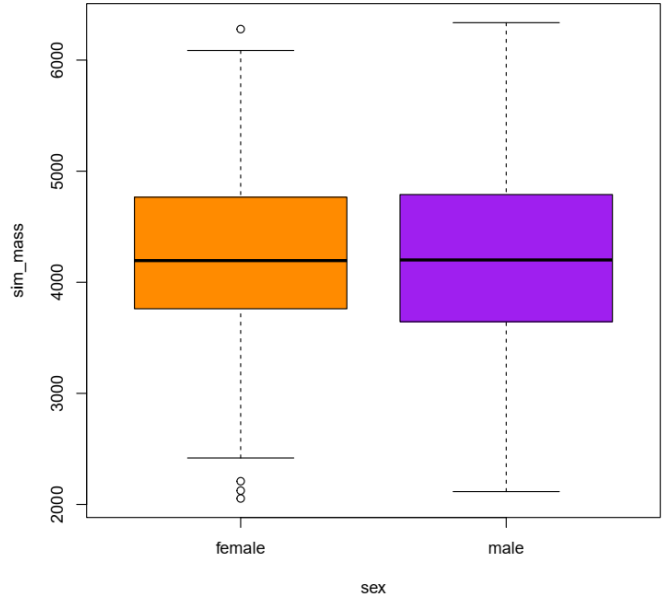


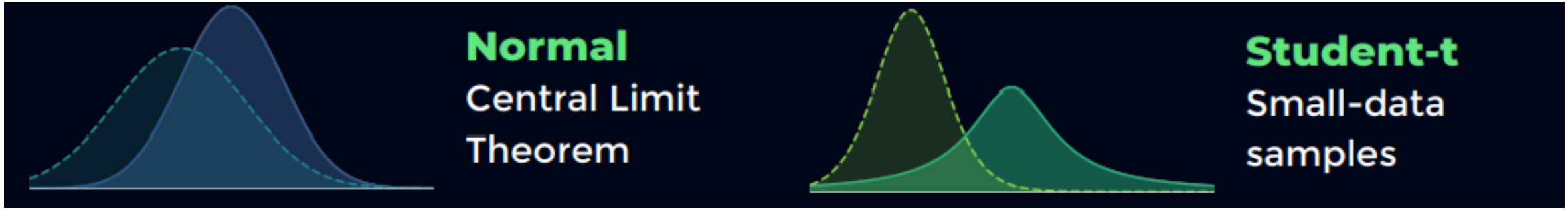
Parametric test Vs. Non-parametric test

	Parametric test	Non-parametric test
2 Numerical variable (Correlation)	Pearson method	Spearman method
1 Numerical variable 2 Categorical variable (Between groups)	Independent t-test	Wilcoxon Rank Sum test (Mann-Whitney U test)
1 Numerical variable 2 Categorical variable (Within groups)	Paired t-test	Wilcoxon Signed Rank test
1 Numerical variable >2 Categorical variable	One-way ANOVA	Kruskal-Wallis test
2 Categorical variable	-	Chi-squared test

Student t-test

- Type of parametric statistical test
- Determine whether two groups of data are significantly different from each other
- Based on the t-distribution (probability distribution similar to the normal distribution)





Student t-test

Syntax:

```
t.test(    x, y = NULL,
           alternative = c("two.sided", "less", "greater"),
           mu = 0,
           paired = FALSE,
           var.equal = FALSE,
           conf.level = 0.95, ...)
```

```
> head(dt)
  species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g    sex year sim_mass
1  Adelie Torgersen           39.1          18.7               181        3750   male 2007 4050.258
2  Adelie Torgersen           39.5          17.4               186        3800 female 2007 3496.960
3  Adelie Torgersen           40.3          18.0               195        3250 female 2007 3826.199
4  Adelie Torgersen             NA            NA                NA          NA   <NA> 2007 3815.542
5  Adelie Torgersen           36.7          19.3               193        3450 female 2007 4355.975
6  Adelie Torgersen           39.3          20.6               190        3650   male 2007 5204.582
> penguin_male <- dt[dt$sex == "male", "sim_mass"]
> summary(penguin_male)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2116    3645    4201    4212    4788    6337      11
> penguin_female <- dt[dt$sex == "female", "sim_mass"]
> summary(penguin_female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2054    3761    4195    4221    4766    6279      11
```

```
> ttest1 <- t.test(penguin_male, penguin_female)
> ttest1

        Welch Two Sample t-test

data:  penguin_male and penguin_female
t = -0.10503, df = 330.43, p-value = 0.9164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -183.0088  164.4577
sample estimates:
mean of x mean of y
 4211.822  4221.098

> ttest1$p.value
[1] 0.9164185
> ttest1$statistic
         t
-0.1050265
> ttest1$estimate
mean of x mean of y
 4211.822  4221.098
```

Student t-test

Syntax:

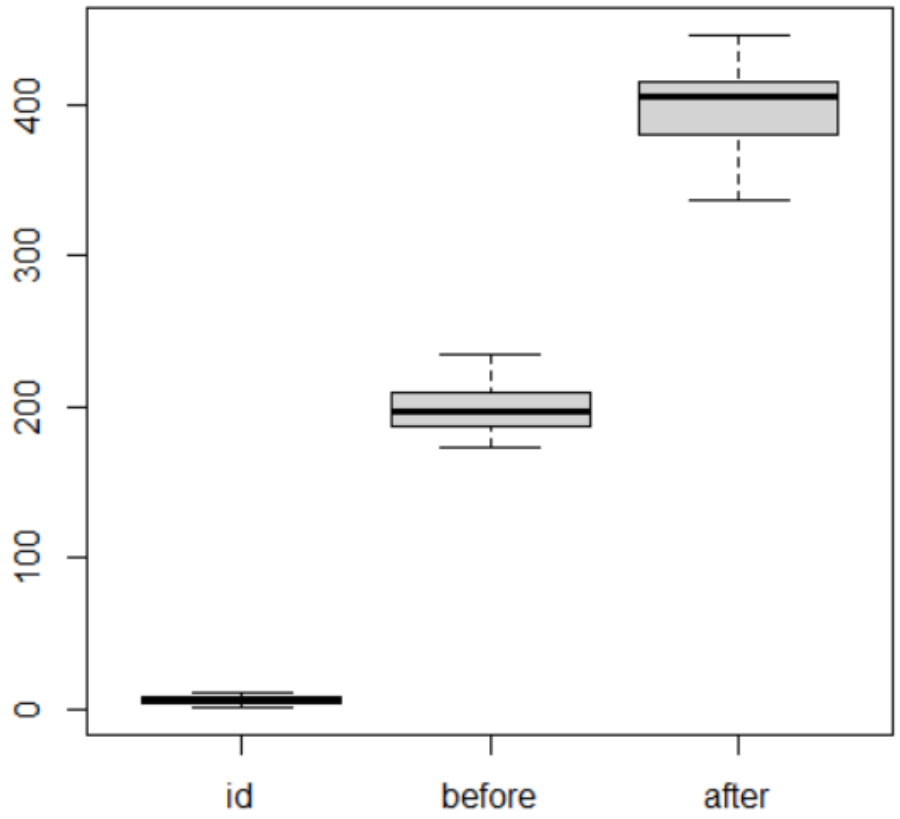
t.test(formula, data)

```
> ttest2 <- t.test(sim_mass ~ sex, data=dt)
> ttest2

        Welch Two Sample t-test

data:  sim_mass by sex
t = 0.10503, df = 330.43, p-value = 0.9164
alternative hypothesis: true difference in means between group female and group male is not equal to 0
95 percent confidence interval:
 -164.4577  183.0088
sample estimates:
mean in group female   mean in group male
            4221.098             4211.822

> table(summary(ttest1) == summary(ttest2))

TRUE
  30
```

```
> ttest2$p.value
[1] 0.9164185
> ttest2$statistic
        t
0.1050265
> ttest2$estimate
mean in group female   mean in group male
            4221.098             4211.822
```

mice2 data in the nutshell

```
> library(datarium)
> data("mice2", package = "datarium")
> head(mice2, 3)
  id before after
1  1  187.2 429.5
2  2  194.2 404.4
3  3  231.7 405.6
> str(mice2)
'data.frame':   10 obs. of  3 variables:
 $ id    : int  1 2 3 4 5 6 7 8 9 10
 $ before: num  187 194 232 200 202 ...
 $ after : num  430 404 406 397 378 ...
> summary(mice2)
       id            before          after      
 Min.   : 1.00   Min.   :172.4   Min.   :337.0  
 1st Qu.: 3.25   1st Qu.:187.8   1st Qu.:384.5  
 Median : 5.50   Median :197.3   Median :405.0  
 Mean   : 5.50   Mean   :200.6   Mean   :400.0  
 3rd Qu.: 7.75   3rd Qu.:206.9   3rd Qu.:412.8  
 Max.   :10.00   Max.   :235.0   Max.   :445.8  
> boxplot(mice2)
```



Student t-test

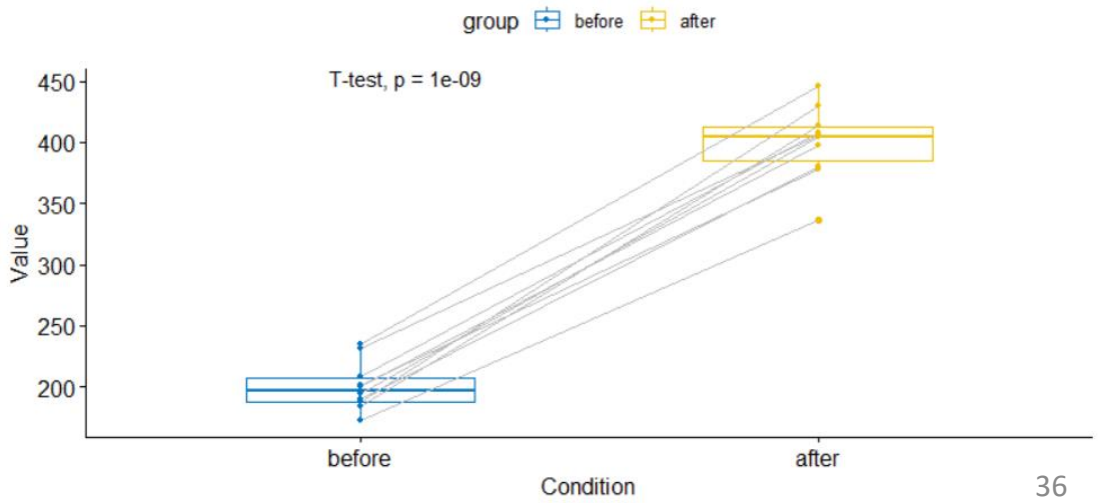
Un-paired t-test

```
> res_unpaired <- t.test(mice2$before, mice2$after, paired = FALSE)
> res_unpaired

        Welch Two Sample t-test

data:  mice2$before and mice2$after
t = -17.453, df = 15.667, p-value = 1.099e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -223.7515 -175.2085
sample estimates:
mean of x mean of y
   200.56    400.04
```

```
> library(reshape2)
> mice.melt <- melt(mice2, id = 'id')
> names(mice.melt)[2:3] <- c("group", "weight")
> library(ggpubr)
> ggpaired(mice.melt, x = "group", y = "weight",
+          color = "group", line.color = "gray", line.size = 0.4,
+          palette = "jco")+
+   stat_compare_means(paired = TRUE, method = "t.test")
```

Paired t-test

```
> res_paired <- t.test(mice2$before, mice2$after, paired = TRUE)
> res_paired

        Paired t-test

data:  mice2$before and mice2$after
t = -25.546, df = 9, p-value = 1.039e-09
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -217.1442 -181.8158
sample estimates:
mean difference
        -199.48
```

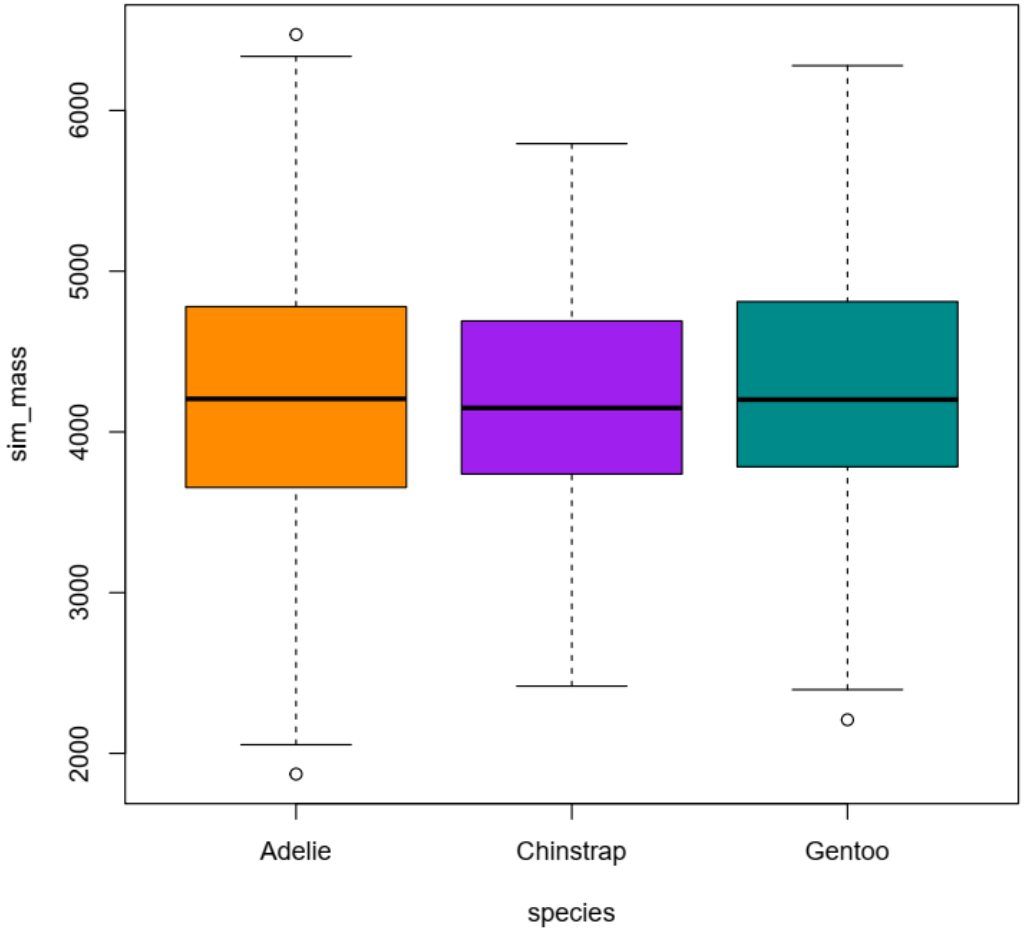


Penguins question!!!

- Is there any statistical difference simulated mass among penguins species group?

ANOVA test

- Type of parametric statistical test
- Compare the means of three or more groups of data
- Compare the variance between the group means (a) to the variance of each group's (b)
- If the a > b → significant difference in the means of the groups.



ANOVA test

Syntax:

```
aov(formula,
    data = NULL,
    projections = FALSE,
    qr = TRUE,
    contrasts = NULL, ...)
```

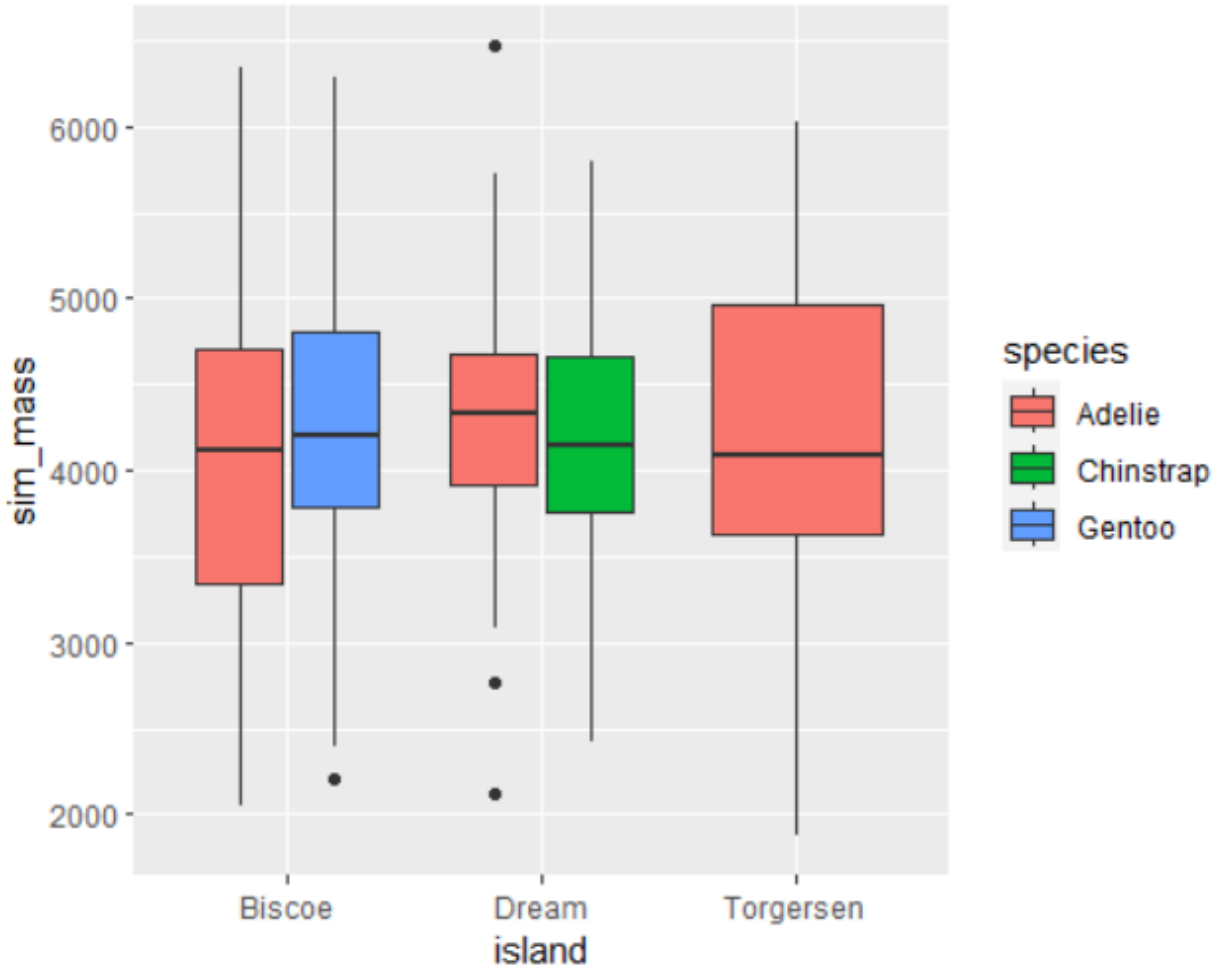
```
> anova1way <- aov(sim_mass ~ species, data=dt)
> sum_anova1way <- summary(anova1way)
> sum_anova1way
             Df    Sum Sq Mean Sq F value Pr(>F)
species       2    838247  419124   0.627  0.535
Residuals   341 227990943  668595
> sum_anova1way[[1]]$`Pr(>F)`
[1] 0.5348749        NA
> sum_anova1way[[1]]$`F value`
[1] 0.626872       NA
```

Penguins question!!!

- Is there any statistical different simulated mass in different penguins species and island?

Two way ANOVA test

```
> anova2way <- aov(sim_mass ~ species + island, data=dt)
> sum_anova2way <- summary(anova2way)
> sum_anova2way
             Df    Sum Sq Mean Sq F value Pr(>F)
species       2    838247  419124   0.629  0.534
island        2   2031013 1015506   1.524  0.219
Residuals   339 225959930  666548
> sum_anova2way[[1]]$`Pr(>F)`
[1] 0.5338538 0.2194297        NA
> sum_anova2way[[1]]$`F value`
[1] 0.6287968 1.5235298        NA
> TukeyHSD(anova2way)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = sim_mass ~ species + island, data = dt)

$species
                     diff       lwr      upr     p adj
Chinstrap-Adelie -38.59059 -318.9838 241.8026 0.9437918
Gentoo-Adelie     86.49959 -146.0707 319.0699 0.6561091
Gentoo-Chinstrap 125.09018 -164.9230 415.1033 0.5676622

$island
                     diff       lwr      upr     p adj
Dream-Biscoe      97.93389 -129.6064 325.4742 0.5690336
Torgersen-Biscoe -28.58766 -333.5789 276.4036 0.9735191
Torgersen-Dream  -126.52154 -444.0453 191.0022 0.6166197
```

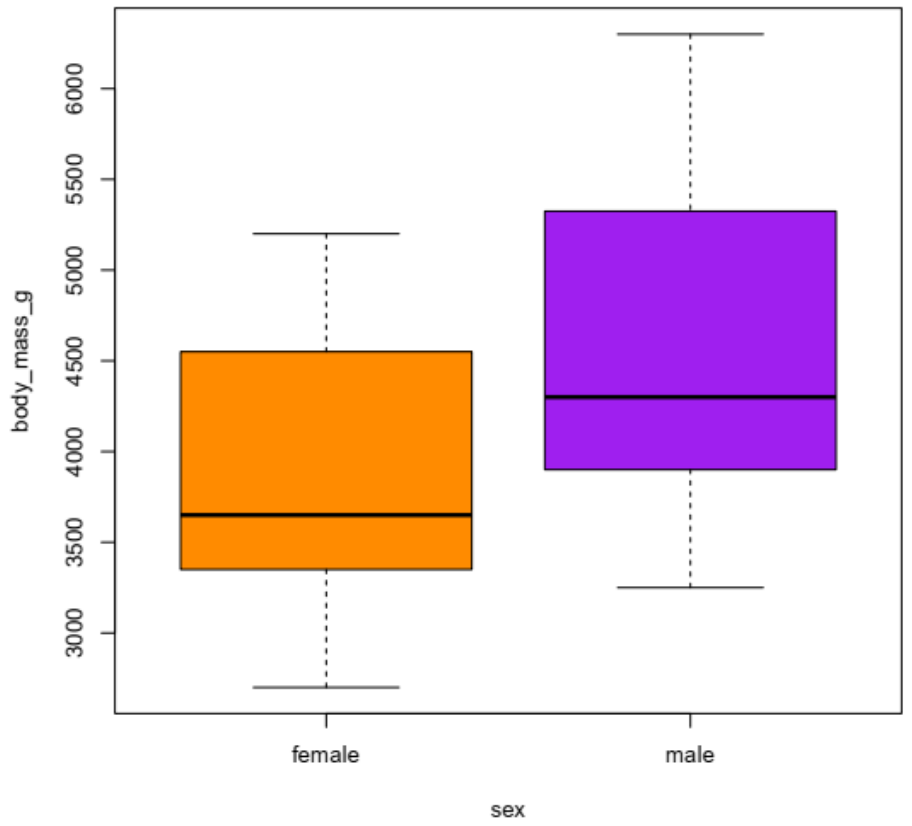


Penguins question!!!

- Forget the pseudo for now! Is there any statistical different real mass between gender group?

Wilcoxon test

- Type of non-parametric statistical test
- Compare two groups of data.
- Alternative to the t-test



Wilcoxon test

Syntax:

```
wilcox.test(formula, data)

wilcox.test(x, y = NULL,
            alternative = c("two.sided",
            "less", "greater"),
            mu = 0,
            paired = FALSE,
            exact = NULL,
            correct = TRUE,
            conf.int = FALSE,
            conf.level = 0.95,
            tol.root = 1e-4,
           digits.rank = Inf, ...)
```

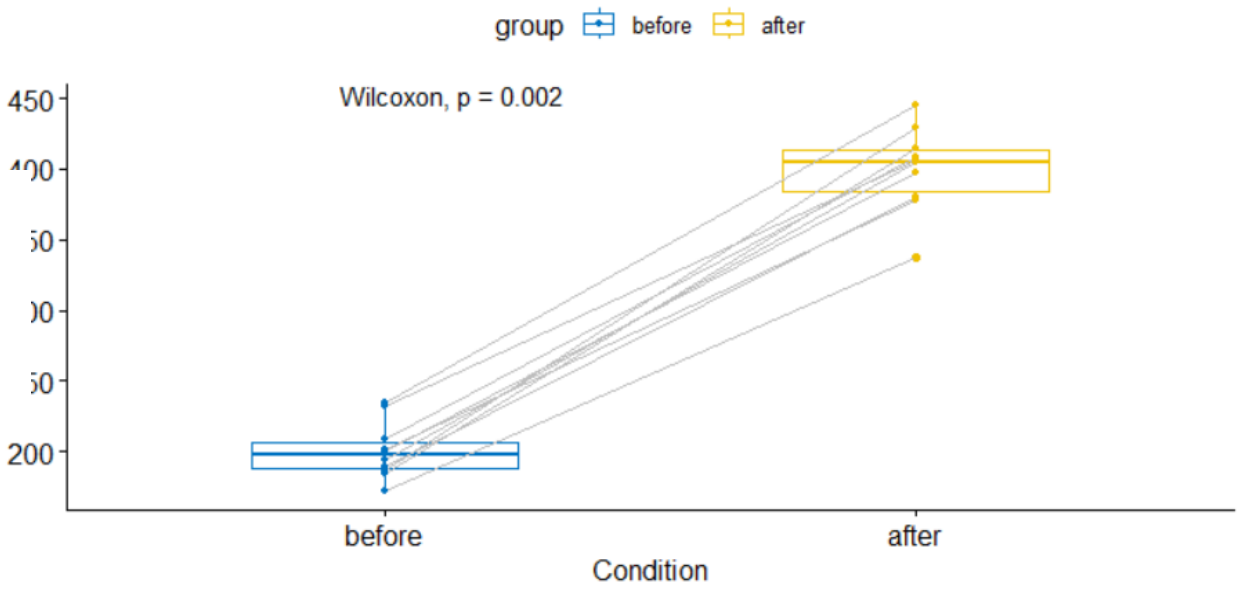
```
> penguin_male <- dt[dt$sex == "male", "body_mass_g"]
> summary(penguin_male)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   3250    3900    4300    4546    5312    6300      11
> penguin_female <- dt[dt$sex == "female", "body_mass_g"]
> summary(penguin_female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   2700    3350    3650    3862    4550    5200      11
> wilcoxon1 <- wilcox.test(body_mass_g ~ sex, data=dt)
> wilcoxon1

        Wilcoxon rank sum test with continuity correction

data:  body_mass_g by sex
W = 6874.5, p-value = 1.813e-15
alternative hypothesis: true location shift is not equal to 0

> wilcoxon2 <- wilcox.test(penguin_female, penguin_male)
> wilcoxon2

        Wilcoxon rank sum test with continuity correction

data:  penguin_female and penguin_male
W = 6874.5, p-value = 1.813e-15
alternative hypothesis: true location shift is not equal to 0

> wilcoxon2$p.value
[1] 1.813334e-15
> wilcoxon2$statistic
     W
6874.5
```

Wilcoxon rank sum test (Mann-Whitney U test) vs. Wilcoxon signed rank test

```
> wilcoxon3 <- wilcox.test(mice2$before, mice2$after, paired = FALSE)
> wilcoxon3

        Wilcoxon rank sum exact test

data:  mice2$before and mice2$after
W = 0, p-value = 1.083e-05
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcoxon4 <- wilcox.test(mice2$before, mice2$after, paired = TRUE)
> wilcoxon4

        Wilcoxon signed rank exact test

data:  mice2$before and mice2$after
V = 0, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

```
> library(reshape2)
> mice.melt <- melt(mice2, id = 'id')
> names(mice.melt)[2:3] <- c("group", "weight")
> library(ggpubr)
> ggpaired(mice.melt, x = "group", y = "weight",
+          color = "group", line.color = "gray", line.size = 0.4,
+          palette = "jco")+
+   stat_compare_means(paired = TRUE, method = "wilcox.test")
```



Penguins question!!!

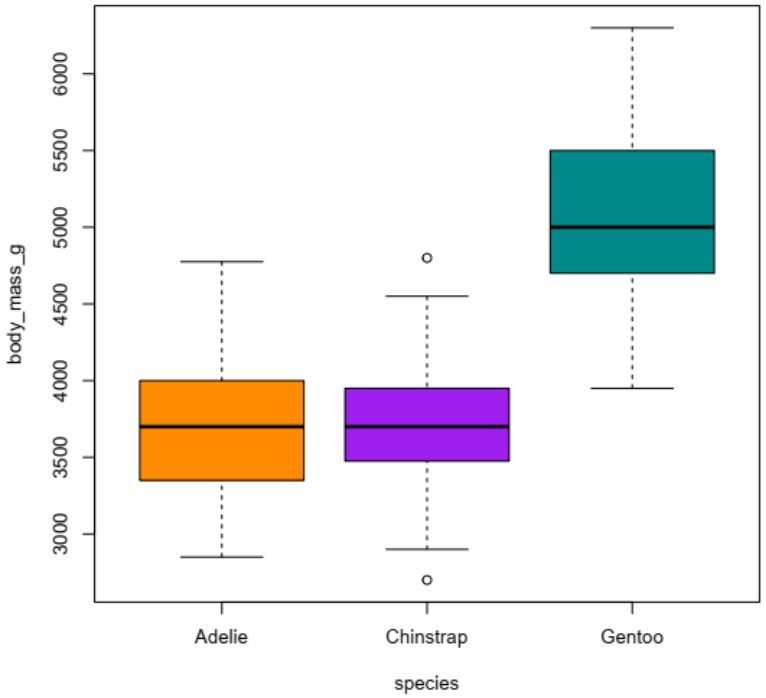
- Is there any statistical difference real mass among the penguins species group?

Kruskal-Wallis test

- Type of non-parametric statistical test
- Compare three or more groups of data.
- Alternative to the ANOVA

Kruskal-Wallis test

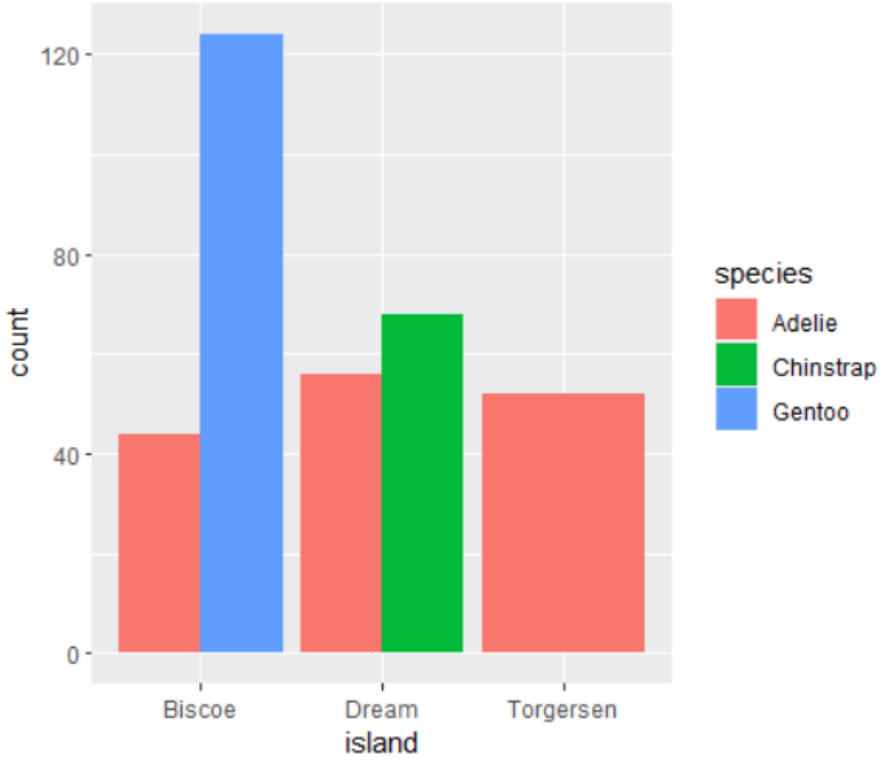
kruskal.test(formula,
data,
subset,
na.action, ...)

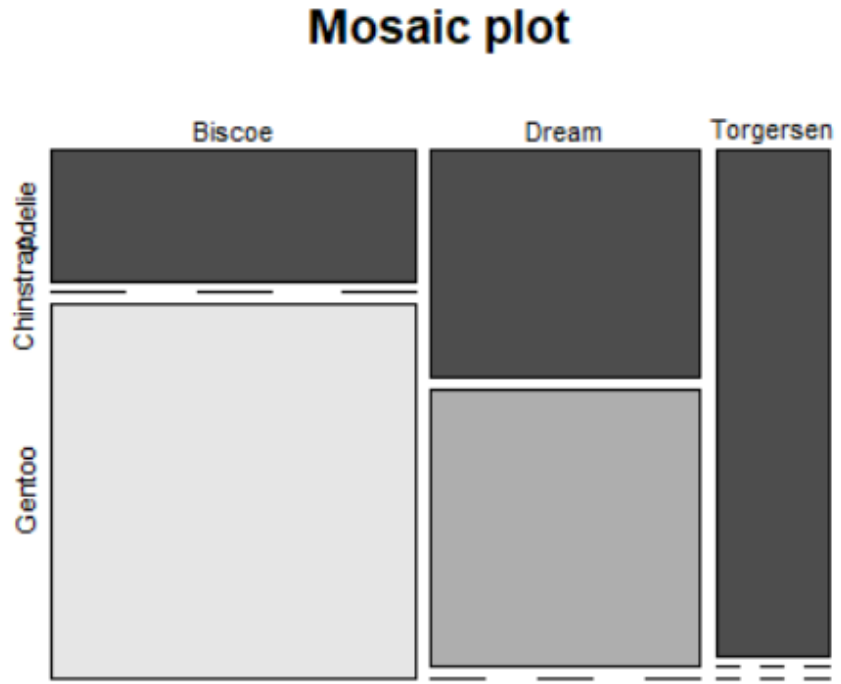



```
> krus_wal <- kruskal.test(body_mass_g ~ species, data=dt)
> krus_wal

        Kruskal-Wallis rank sum test

data:  body_mass_g by species
Kruskal-Wallis chi-squared = 217.6, df = 2, p-value < 2.2e-16

> krus_wal$p.value
[1] 5.609512e-48
> krus_wal$statistic
Kruskal-Wallis chi-squared
                  217.5992
> pairwise.wilcox.test(dt$body_mass_g, dt$species,
+                      p.adjust.method = "BH")

        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  dt$body_mass_g and dt$species

          Adelie Chinstrap
Chinstrap 0.49   -
Gentoo    <2e-16 <2e-16

P value adjustment method: BH
```

Penguins question!!!

- Is there any statistical difference between the penguins island and species group?





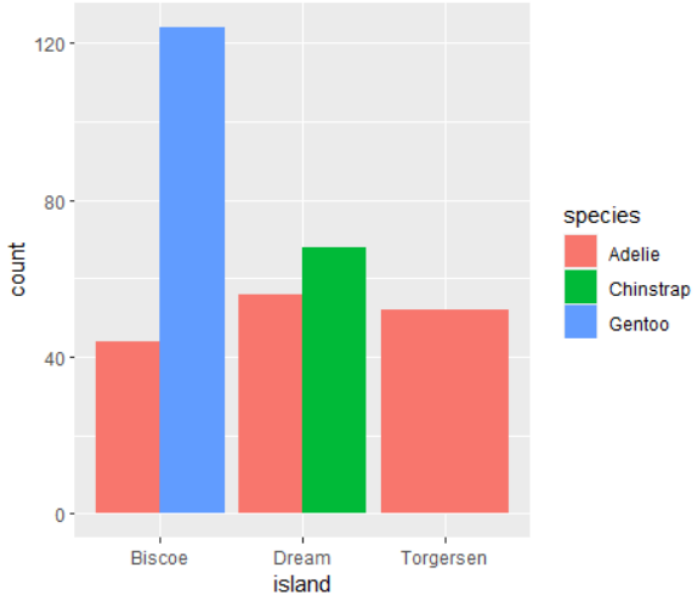

```
> ggplot(dt, aes(island, ..count.., fill = species)) +
+   geom_bar(position = "dodge")
```

```
> mosaicplot(table(dt$island, dt$species),
+            main = "Mosaic plot", color = TRUE)
```

Chi-square test

- Determine if there is a significant association between two categorical variables

```
chisq.test(x, y = NULL,
           correct = TRUE,
           p = rep(1/length(x),
           length(x)),
           rescale.p = FALSE,
           simulate.p.value = FALSE,
           B = 2000)
```




```
> tb3 <- table(dt$island, dt$species)
> tb3

           Adelie Chinstrap Gentoo
  Biscoe       44         0    124
  Dream        56        68      0
  Torgersen    52         0      0
> chisq.test(tb3)

        Pearson's Chi-squared test

data:  tb3
X-squared = 299.55, df = 4, p-value < 2.2e-16

> summary(tb3)
Number of cases in table: 344
Number of factors: 2
Test for independence of all factors:
        Chisq = 299.55, df = 4, p-value = 1.355e-63
```

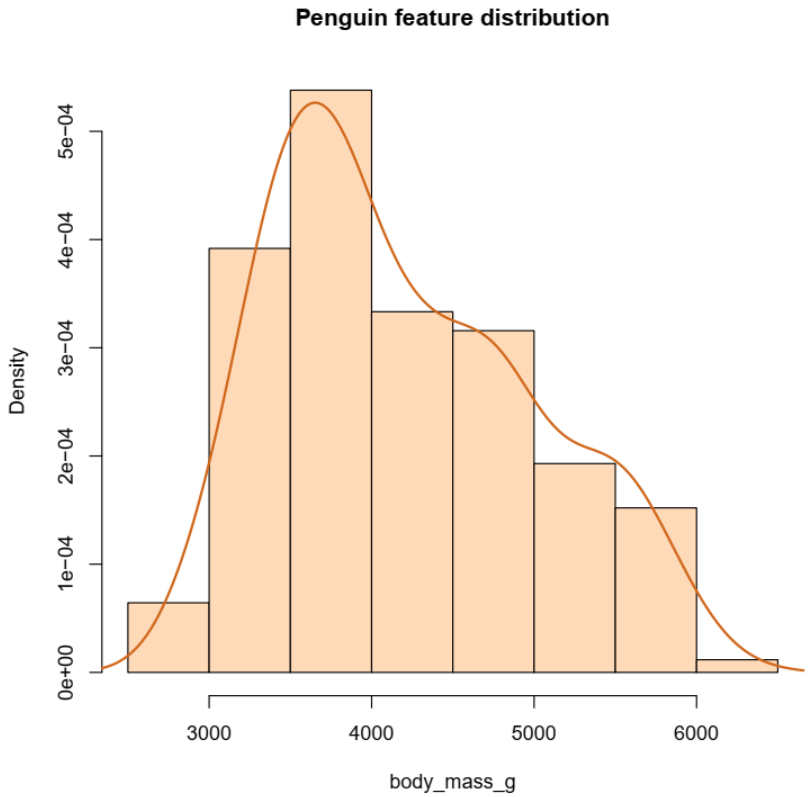
Classwork 2

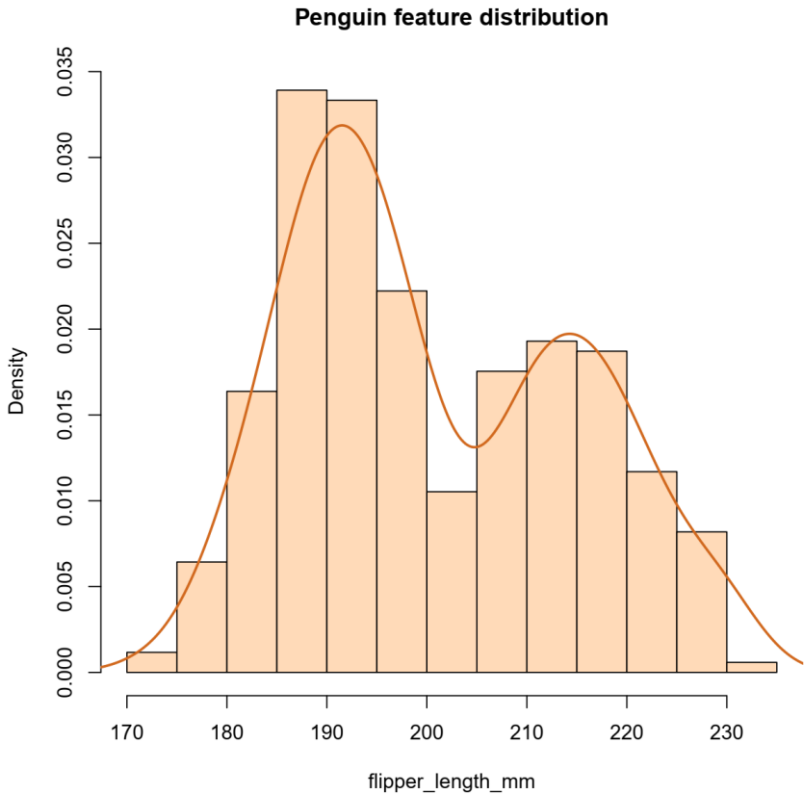
Use *Iris* data in R and do as requested:

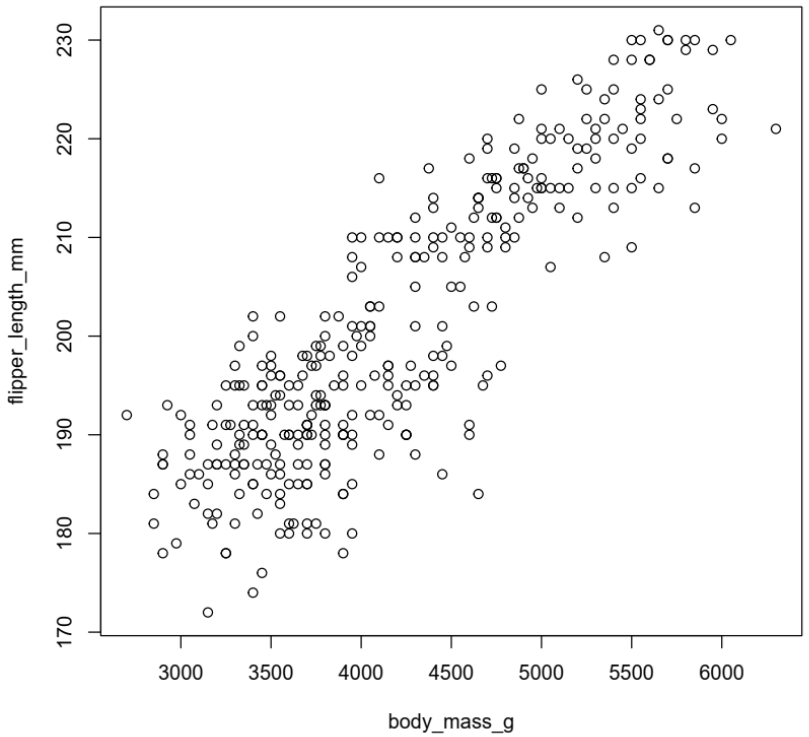
1. Check if the data is normally distributed.
2. Is there statistical difference *Sepal.Length* among *species*?
3. Is there statistical difference *Petal.Length* between *virginica* and the other species?

Penguins question!!!

- What is the relationship between the real body mass and flipper length?







Linear regression

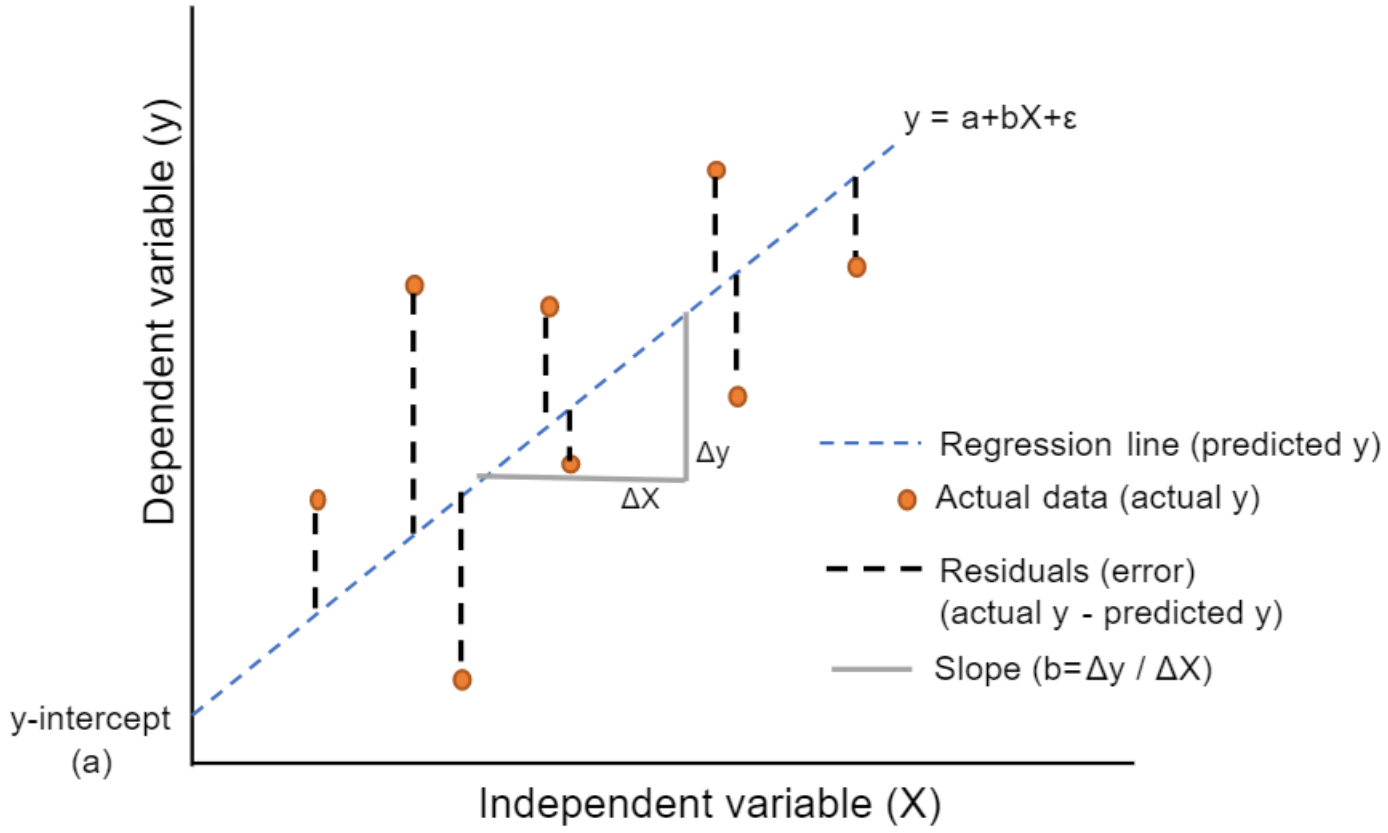
- A supervised machine learning technique
- Establish a relationship between a dependent variable and one or more independent variables.
- Determine if there is a linear relationship between two or more variables and how strong that relationship is.
- Simple linear regression equation:

$$y = b_0 + b_1 x$$

- Multiple linear regression equation:

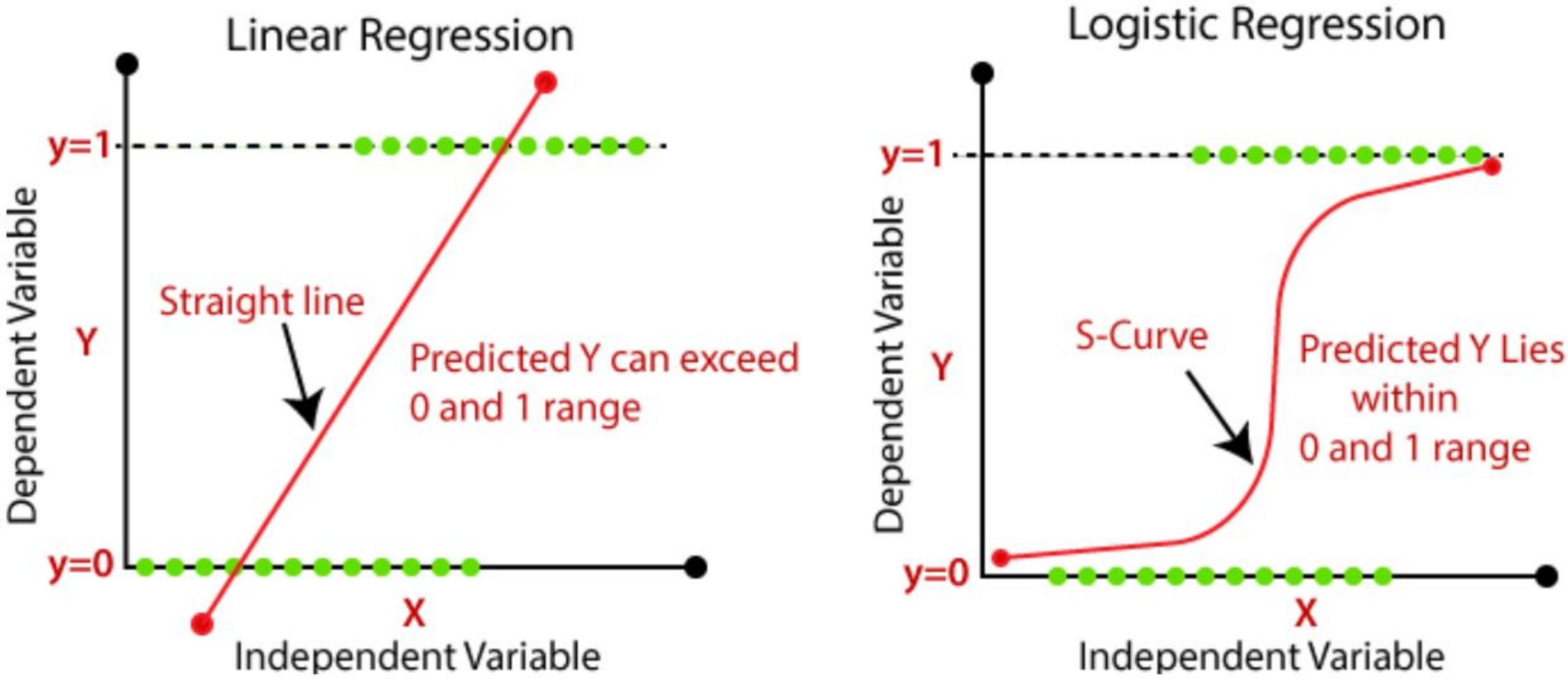
$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + + b_n x_n + \varepsilon$$

Linear regression



Find the smallest residual square by Least square method

Linear regression vs. Logistic regession



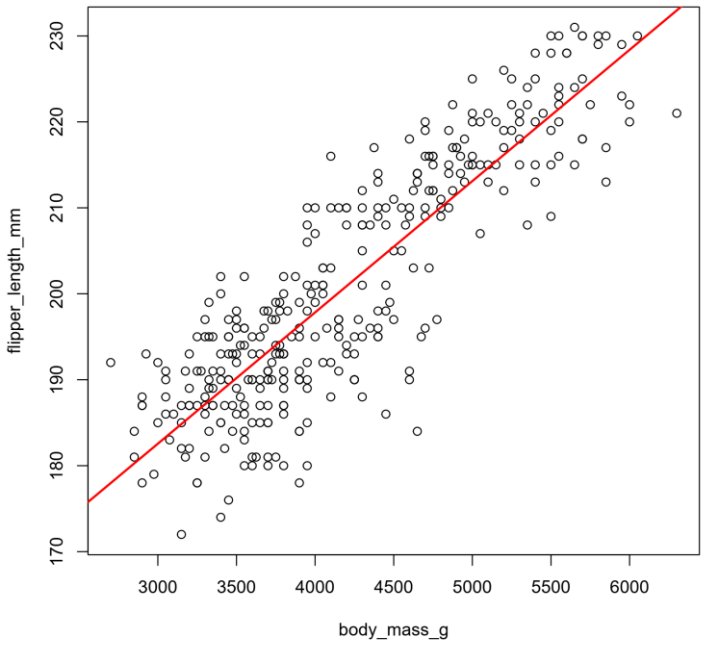
Linear regression

Syntax:

```
lm(formula,
    data,
    subset,
    weights,
    na.action,
    method = "qr",
    model = TRUE,
    x = FALSE,
    y = FALSE,
    qr = TRUE,
    singular.ok = TRUE,
    contrasts = NULL,
    offset, ...)
```

```
> model1 <- lm(flipper_length_mm ~ body_mass_g, dt)
> model1

Call:
lm(formula = flipper_length_mm ~ body_mass_g, data = dt)

Coefficients:
(Intercept)  body_mass_g
  136.72956      0.01528

> summary(model1)

Call:
lm(formula = flipper_length_mm ~ body_mass_g, data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-23.7626  -4.9138   0.9891   5.1166  16.6392

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.367e+02  1.997e+00   68.47   <2e-16 ***
body_mass_g 1.528e-02  4.668e-04   32.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.913 on 340 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.759,     Adjusted R-squared:  0.7583
F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

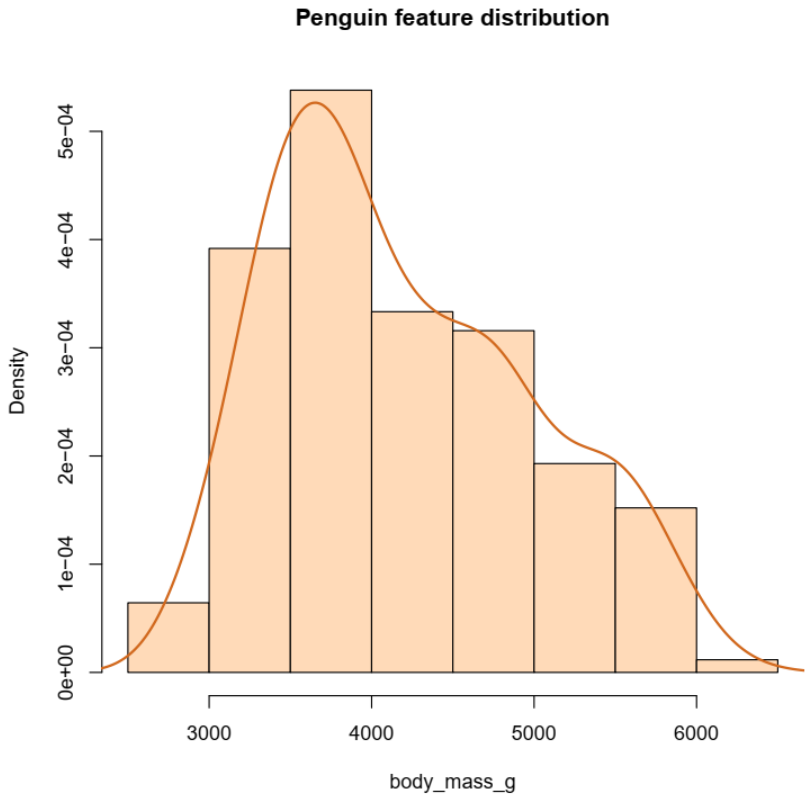
```
> plot(flipper_length_mm ~ body_mass_g, dt)
> abline(model1, lwd=2, col="red")
```

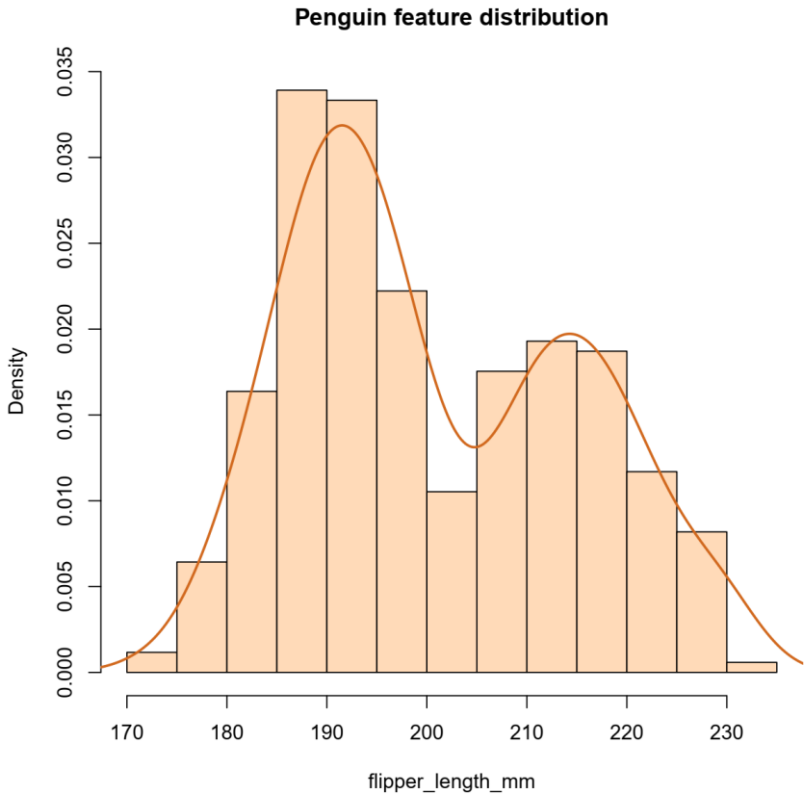


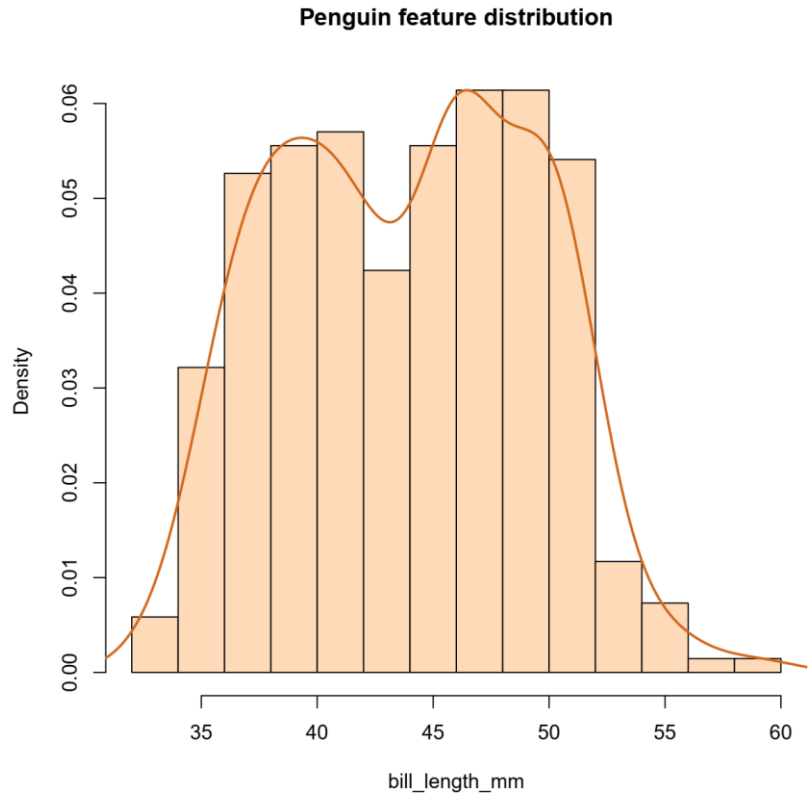
$$y = 1.367 \times 10^{2} + 1.528 \times 10^{-2} x$$

Penguins question!!!

- What is the relationship between the flipper length with body mass and bill length?



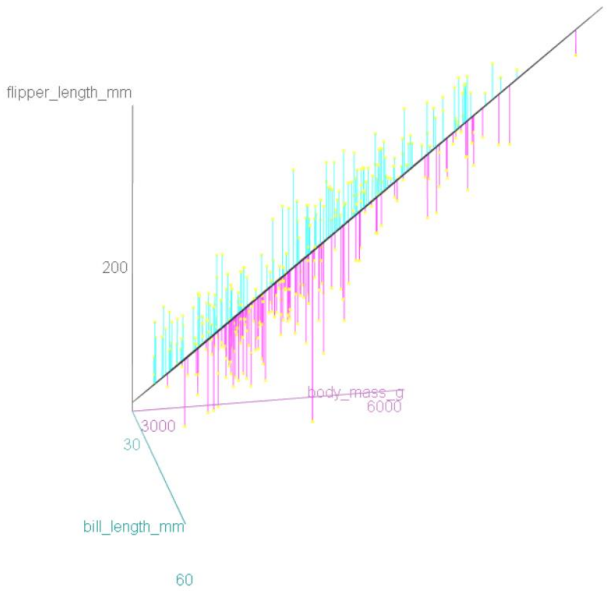


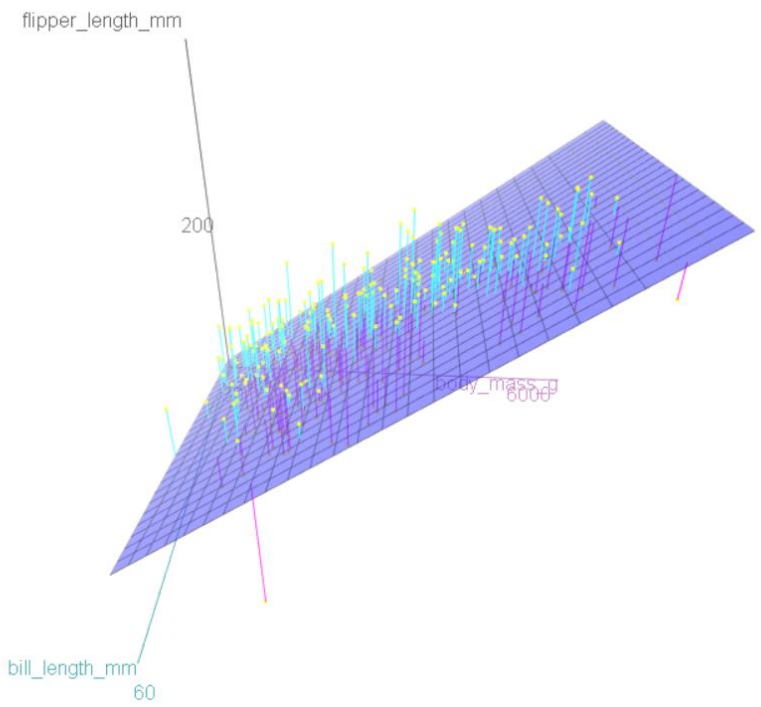


Multiple linear regression

```
> model2 <- lm(flipper_length_mm ~ body_mass_g + bill_length_mm, dt)
> model2

Call:
lm(formula = flipper_length_mm ~ body_mass_g + bill_length_mm,
    data = dt)

Coefficients:
   (Intercept)     body_mass_g  bill_length_mm
     121.95604         0.01305         0.54922

> summary(model2)

Call:
lm(formula = flipper_length_mm ~ body_mass_g + bill_length_mm,
    data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-21.0989  -4.5520   0.3379   4.8942  16.0953

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.220e+02  2.855e+00  42.715  < 2e-16 ***
body_mass_g    1.305e-02  5.452e-04  23.939  < 2e-16 ***
bill_length_mm 5.492e-01  8.008e-02   6.859 3.31e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.488 on 339 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7884,    Adjusted R-squared:  0.7871
F-statistic: 631.4 on 2 and 339 DF,  p-value: < 2.2e-16
```

$$y = 1.22 \times 10^{2} + 1.305 \times 10^{-2} x_1 + 5.492 \times 10^{-1} x_2$$






```
> scatter3d(flipper_length_mm ~ body_mass_g + bill_length_mm, dt)
```

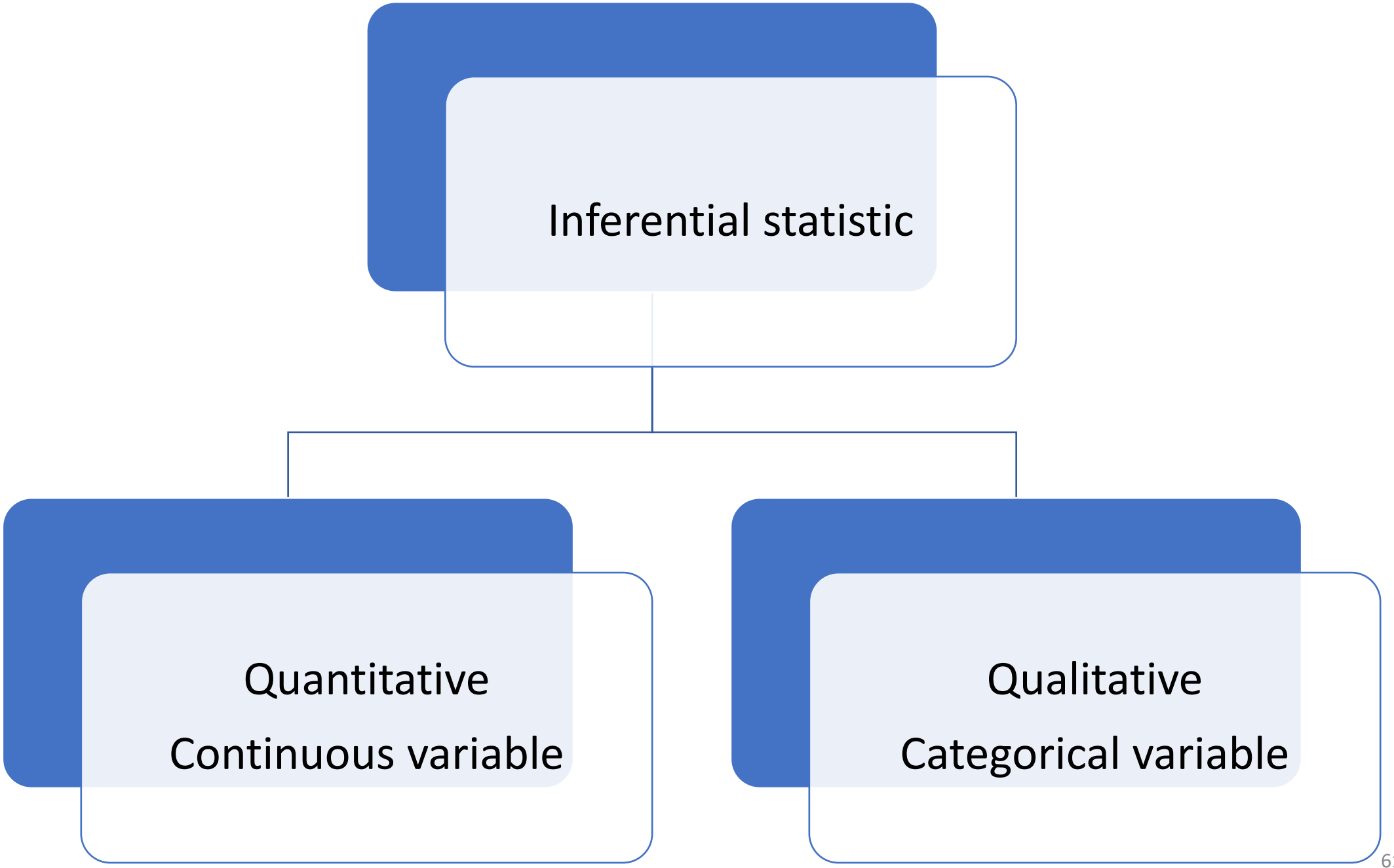
Classwork 3

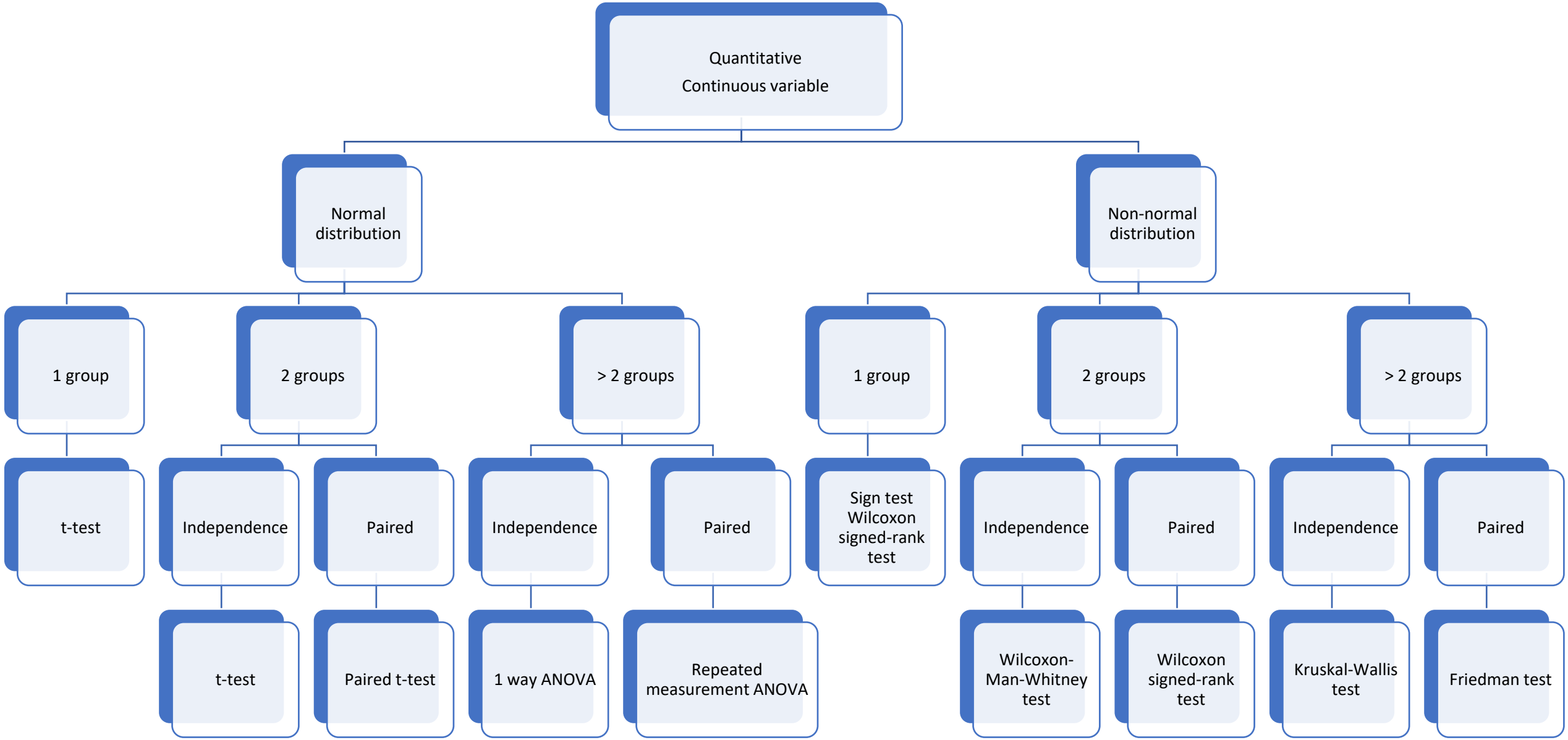
Use *Iris* data in R and do as requested:

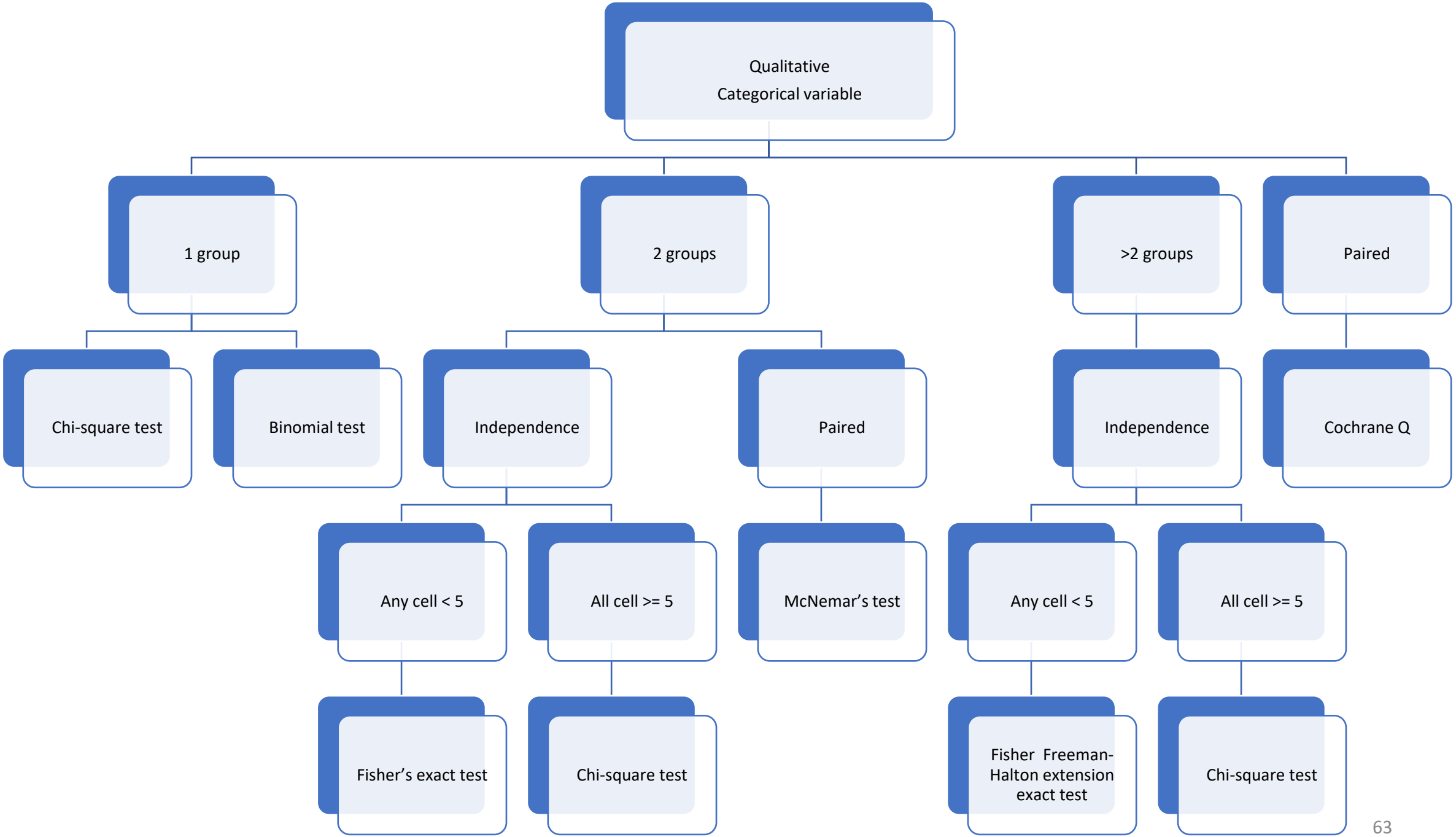
1. Perform linear regression to show the relationship of *Sepal.Length* and *Petal.Length*. Give a comment.
2. Perform linear regression to show the relationship of *Species* and 4 other variables. Give a comment.
3. Let plot these models.

Summary

Type		Name	Function in R
Descriptive statistics		Variance	var()
		Covariance	cov()
		Correlation	cor()
Inferential statistics	Parametric tests	T-test	t.test()
		ANOVA	aov()
	Non-parametric tests	Wilcoxon test	wilcox.test()
		Kruskal-Wallis test	kruskal.test()
		Chi-squared test	chisq.test()
	Normality tests	Shapiro-Wilk Test	shapiro.test()
		Kolmogorov-Smirnov Test	ks.test()
Supervised machine learning		Linear Regression	lm()


Inferential statistic
Quantitative
Continuous variable
Qualitative
Categorical variable


Quantitative
Continuous variable
Normal distribution
Non-normal distribution
1 group
2 groups
> 2 groups
1 group
2 groups
> 2 groups
t-test
Independence
Paired
Independence
Paired
Sign test
Wilcoxon signed-rank test
Independence
Paired
Independence
Paired
t-test
Paired t-test
1 way ANOVA
Repeated measurement ANOVA
Wilcoxon-Man-Whitney test
Wilcoxon signed-rank test
Kruskal-Wallis test
Friedman test


Qualitative
Categorical variable
1 group
2 groups
>2 groups
Paired
Chi-square test
Binomial test
Independence
Paired
Independence
Cochrane Q
Any cell < 5
All cell >= 5
McNemar's test
Any cell < 5
All cell >= 5
Fisher's exact test
Chi-square test
Fisher Freeman-Halton extension exact test
Chi-square test

Classwork 1

1. If the law requires women to marry only men 2 years older than themselves, what is the correlation of the ages between all pairs of couples (husbands and wives)?
2. Use *Iris* data in R and do as requested:
   a. Visualize the *Sepal.Length*, *Sepal.Width* and *Species*.
   b. Calculate variance, covariance, correlation between *Sepal.Length* and *Sepal.Width*
   c. Calculate variance, covariance, correlation between *Sepal.Length* and *Species*
3. Use *anscombe* data in R and do as requested:
   a. Calculate the variance, covariance, and correlation between the *x* variable and the corresponding *y* variable. Give a comment.
   b. Use scatter plot to visualize the *x* variable and the corresponding *y* variable. Give a comment.

Classwork 2

Use *Iris* data in R and do as requested:

1. Check if the data is normally distributed.
2. Is there statistical difference *Sepal.Length* among *species*?
3. Is there statistical difference *Petal.Length* between *virginica* and the other species?

Classwork 3

Use *Iris* data in R and do as requested:

1. Perform linear regression to show the relationship of *Sepal.Length* and *Petal.Length*. Give a comment.
2. Perform linear regression to show the relationship of *Species* and 4 other variables. Give a comment.
3. Let plot these models.

Homework

Download the data from: https://drive.google.com/file/d/1tOdeLpEzhEcsDPU6Vz_dIQsV0UZy0bz0/view?usp=share_link

1. Perform descriptive statistics of cg17348029 and cg21549285. Give a comment.
2. Is there any different of cg17348029 value among statuses groups?
3. Is there any different of cg21549285 value between two gender groups?
4. Are gender difference associated with the difference in the disease statuses?
5. Is the age by the disease statuses in this study normally distributed?
6. Indicate the relationship among the disease statuses, age, value of cg17348029 and cg21549285.
7. Let visualize these results.

Homework

```
> dim(BetaMatrix)
[1] 201 301
> head(BetaMatrix)[1:10]
     probeID sample_011 sample_024 sample_026 sample_030 sample_033 sample_039 sample_055 sample_057 sample_064
1 cg27142757       0.57       0.64       0.59       0.53       0.54       0.54       0.46       0.49       0.50
2 cg22652934       0.31       0.34       0.27       0.27       0.24       0.31       0.17       0.25       0.20
3 cg18448767       0.67       0.68       0.66       0.60       0.65       0.68       0.59       0.61       0.61
4 cg08282375       0.34       0.46       0.47       0.37       0.36       0.43       0.35       0.40       0.30
5 cg05167074       0.52       0.53       0.54       0.49       0.46       0.54       0.42       0.46       0.50
6 cg21549285       0.73       0.73       0.64       0.76       0.63       0.72       0.85       0.56       0.67
```

```
> dim(SampleSheet)
[1] 100   5
> head(SampleSheet)
    sample_ID      study status sex age
1  sample_513 Konigsberg   Mild   M  62
2 sample_1036   Barturen Severe   F  58
3  sample_425 Konigsberg   Mild   F  30
4  sample_944   Barturen Severe   F  93
5  sample_365 Konigsberg Severe   F  50
6  sample_739   Barturen Severe   F  83
> table(SampleSheet$status)

Healthy    Mild  Severe
     26      41      33
```

Thanks for listening!
Have a nice week!