STAT9006: Two-Sample Statistical Inference with *R*

Dr Seán Lacey

Outline

Hypothesis testing

- Null and alternative hypothesis
- Types of error
- Level of significance

Inferential statistics

- One-sample tests
- Difference between two samples

Power Analysis

Exercises

- One-sample t-test
- Independent-samples t-test
- Paired-samples t-test

Null and alternative hypothesis

- A **hypothesis is a claim** or statement about a property of a population. A hypothesis test (or test of significance) is a standard procedure for testing a claim about a property of a population.
- The **null hypothesis** is a statement about the value of a population parameter. The main objective is to test the null hypothesis directly. The result will either be to reject H_0 or to fail to reject H_0.
- The **alternative hypothesis** (denoted by H_1 or H_A) is the statement that the parameter has a value that somehow differs from the null hypothesis.

Types of error

There are two ways in which the results of a hypothesis test may be wrong :

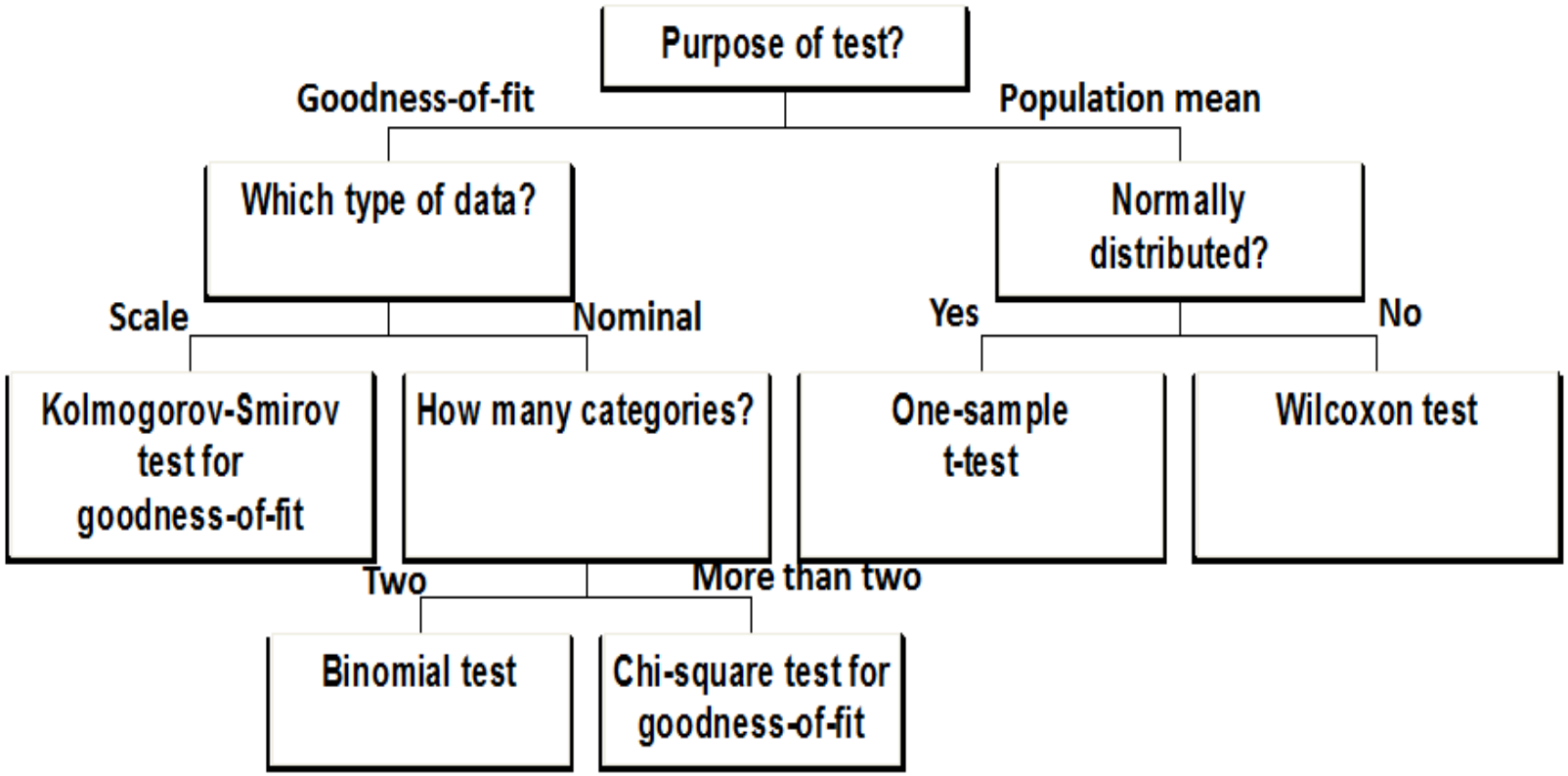
- **Type *I* error** - the mistake of **rejecting the null hypothesis when it is true**. The symbol α is used to represent the probability of a type *I* error. For example, the accused are found guilty when in fact they are innocent.
- **Type *II* error** - mistake of **failing to reject the null hypothesis when it is false**. The symbol β is used to represent the probability of a type *II* error. For example, the accused are found innocent when in fact they are guilty.

	H_0 is correct	H_1 is correct
Reject H_0	Type *I* error	Correct decision
Fail to reject H_0	Correct decision	Type *II* error

Level of significance

- The mistake of rejecting the null hypothesis when it is true.
- α is used to represent the probability of a type *I* error i.e., level of significance.
- If $p \leq \alpha$, then reject H_0 - i.e., **a statistically significant** result.
- If $p > \alpha$, then fail to reject H_0 - i.e., **a statistically insignificant** result.

Selecting a one-sample test



One-sample tests

Goodness-of-fit

1. Scaled data:
   - Compares a sample distribution with a hypothetical distribution i.e., ascertain if the sample is drawn from a normal population;
   - H_0: **No difference** exists between the sample and hypothetical distribution;
   - H_1: **A difference** exists between the sample and hypothetical distribution.
2. Nominal data:
   - Confirms the existence of preferences among a range of choices, or the fairness of a coin or a die.
   - H_0: **No preference** exists ... ;
   - H_1: **A preference** exists ... .

Example: Normality

- Consider IQs of 50 people. Have these scores been drawn from a normal population? (Import the IQ dataset)
- If the sample is normally distributed (symmetric/bell-shaped) then the Shapiro-Wilk test will yield an insignificant result (i.e., $p > 0.05$).
- Before doing a test of normality, you should look at mean, median, histogram and skewness of the measurement.

Example: Normality

```
# apply the three checks, followed with test
# 1: histogram is bell-shaped
library(ggplot2)
range(DF$IQ) #determining the spread of the data
ggplot(DF,aes(x=DF$IQ))+geom_histogram(breaks=seq(70,140,10))+theme_bw()+
labs(x = "IQ score", y = "Frequency")+scale_y_continuous(breaks=seq(0,14,2))+
scale_x_continuous(breaks=seq(70,140,10))

# 2: mean roughly equal median and
# 3: -1<skewness<1
# 4: Do the test
library(psych) # required for skew()
(Norm<-DF %>% summarise("Sample size"=n(),Mean = mean(IQ), Median = median(IQ), Skewness=skew(IQ),
                        "Normally distributed"=ifelse(shapiro.test(IQ)$p.value>0.05,"Yes","No")))
t(Norm)
```

One-sample tests

One-sample t-test

- Tests the null hypothesis that a sample has been drawn from a population with a mean of specified value.
- H_0: No difference exists between the sample and population.
- H_1: A difference exists between the sample population.

Example: From the IQ dataset, determine whether or not the mean IQ score is:

1. Different from 100;
2. Less than 100;
3. Greater than 100.

Example: One-sample difference

```
> #one sample t-test
> (result<-t.test(DF$IQ,mu=100)) #2-tailed test

        One Sample t-test

data:  DF$IQ
t = 1.0471, df = 49, p-value = 0.3002
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  98.02017 106.28783
sample estimates:
mean of x
  102.154

> t.test(DF$IQ,mu=100,alternative = "less") #left-tailed test

        One Sample t-test

data:  DF$IQ
t = 1.0471, df = 49, p-value = 0.8499
alternative hypothesis: true mean is less than 100
95 percent confidence interval:
     -Inf 105.6028
sample estimates:
mean of x
  102.154
```

Example: One-sample difference

```
> t.test(DF$IQ,mu=100,alternative = "greater") #right-tailed test

        One Sample t-test

data:  DF$IQ
t = 1.0471, df = 49, p-value = 0.1501
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 98.70522      Inf
sample estimates:
mean of x
  102.154

>
> #if the single measurement is not normally distributed
> wilcox.test(DF$IQ,mu=100,conf.int = 0.95)

        Wilcoxon signed rank test with continuity correction

data:  DF$IQ
V = 714.5, p-value = 0.3127
alternative hypothesis: true location is not equal to 100
95 percent confidence interval:
  97.65002 106.50002
sample estimates:
(pseudo)median
         102.4
```

Statistical significance

When a statistic is significant:

- It simply means that you are very sure that the statistic is **reliable**;
- It does not allude to the importance of the finding;
- It does not mean the finding has any decision-making utility.

After finding a significant relationship/difference, it is important to evaluate the strength of the relationship/difference.

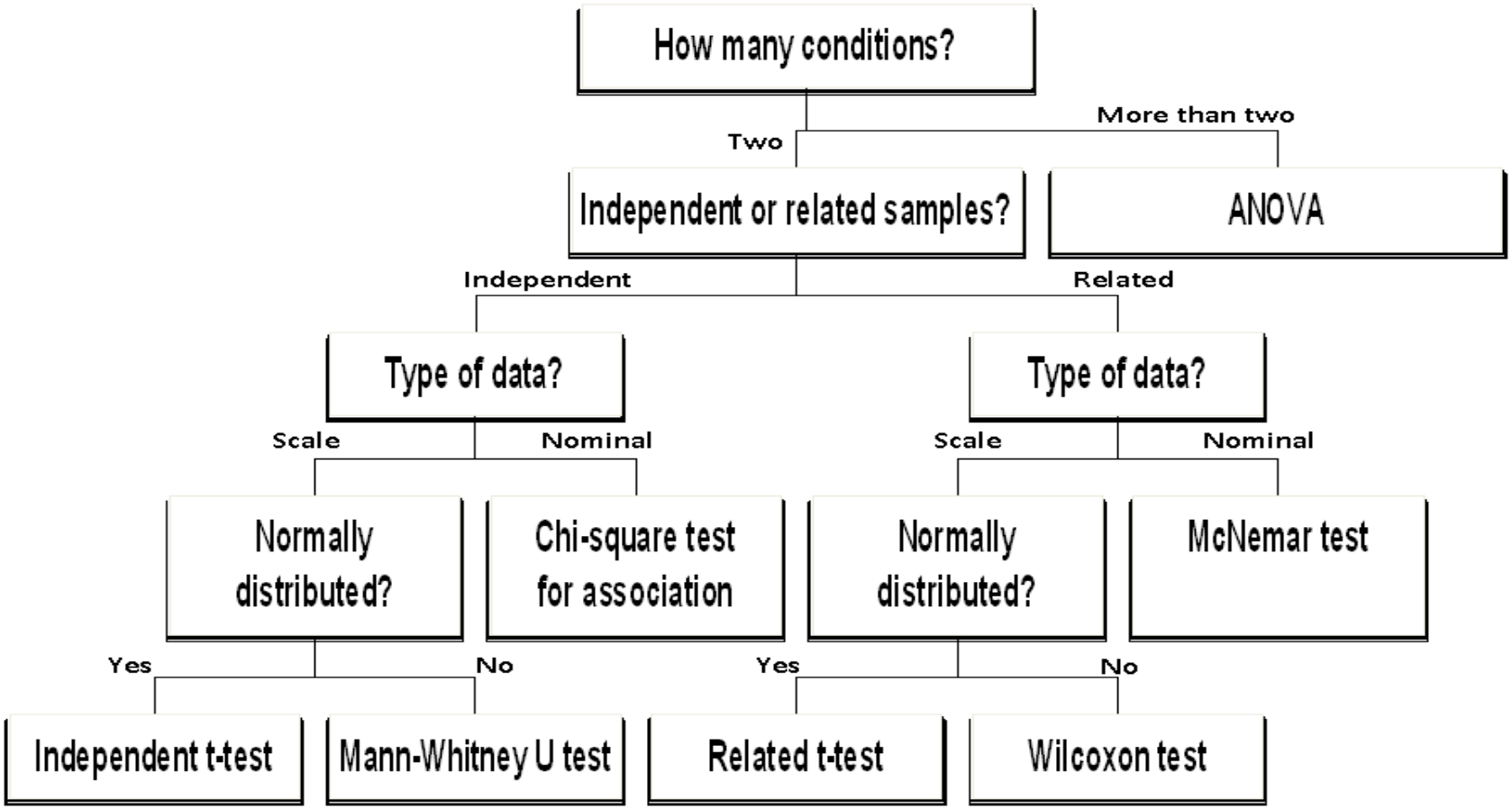
Effect size

The strength of the difference observed is called the effect size (standardised mean difference). The following table displays the classification of Cohen's effect size:

Size of effect	Effect size
Small	$0.2 \leq d < 0.5$
Medium	$0.5 \leq d < 0.8$
Large	$d \geq 0.8$

```
> # effect size ... standardised mean difference
> library(lsr)
> cohensD(DF$IQ,mu=100)
[1] 0.1480855
```

Selecting a test for differences between **means**



Type of samples

- H_0: **No difference** exists between the two population means.
- H_1: **A difference** exists between the two population means.

Independent samples

- The sample values selected from one population are not related or somehow paired with the sample values selected from the other population.
- Both samples are simple random samples.

Related samples

- If the values in one sample are related to the values in the other sample, the samples are dependent. Such samples are often referred to as matched pairs or paired samples.
- The samples are simple random samples.

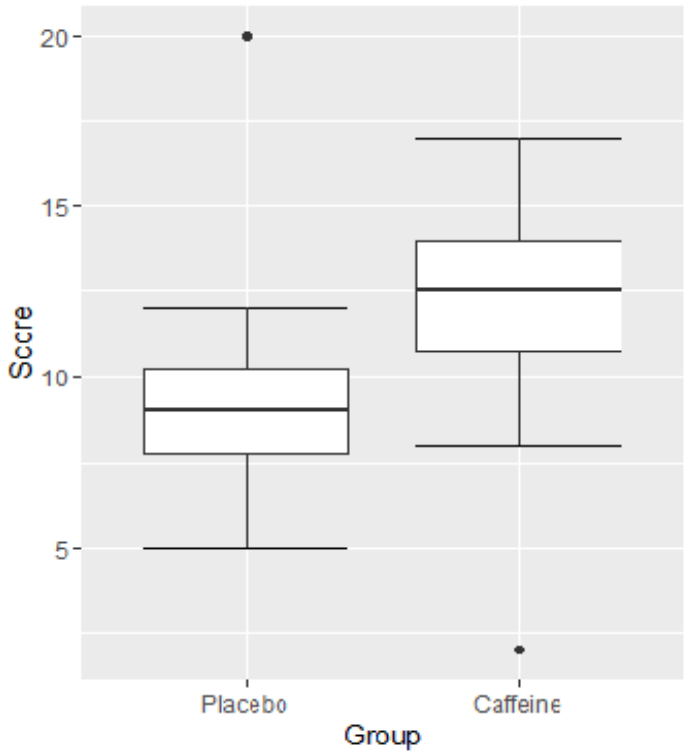
Example: Independent-samples

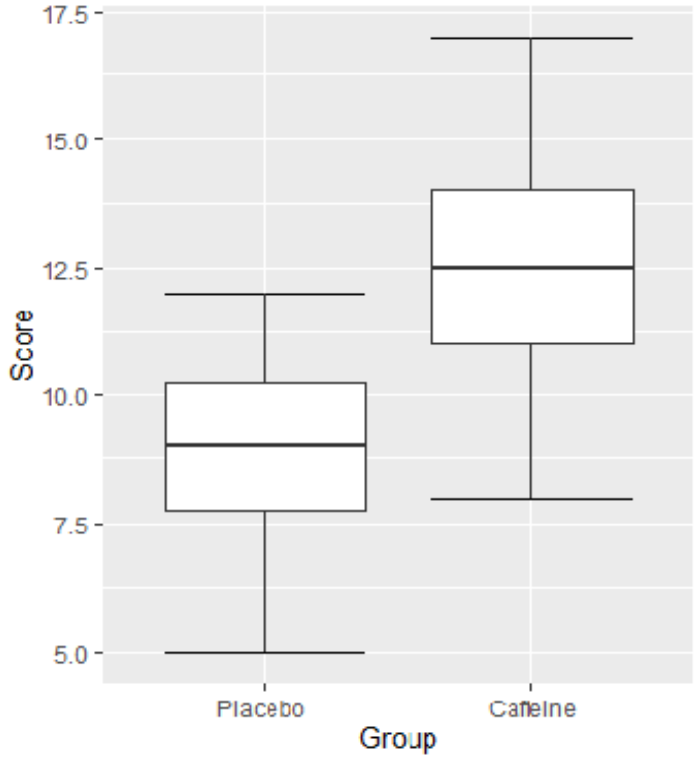
Consider an experiment in which half the participants were tested on shooting accuracy after ingesting a dose of caffeine; the remaining participants took a placebo and took the same test. (Import the Independent *t*-test dataset)

1. Check the data properties and for anomalies such as extreme values or skewed distributions:

```
# properties
is.factor(Indep$Group) # returns FALSE
Indep$Group<-factor(Indep$Group,levels=c("Placebo","Caffeine"))
is.numeric(Indep$Score) # returns TRUE
```

Example: Independent-samples






```
# explore the data
(bp1<-ggplot(Indep,aes(x=Group, y=Score))+stat_boxplot(geom = "errorbar")+geom_boxplot()+
    labs(x = "Group", y = "Score"))

# outliers are present ... trhough investigation outliers are found to be typos
min(Indep$Score[Indep$Group=="Caffeine"]) # locating exact value of the outlier
Indep$Score[Indep$Score == 2] <- 12 # let us say this is a typo and should have been 12
max(Indep$Score[Indep$Group=="Placebo"]) # locating exact value of the outlier
Indep$Score[Indep$Score == 20] <- 12 # let us say this is a typo and should have been 12
#rerun the boxplot
bp1 # edit in your own time
```

Assumptions

2 Check assumptions for an independent *t*-test:

```
> # normality
> Norm2<-Indep %>% group_by(Group) %>% summarise("Sample size"=n(),Mean = mean(Score),
+                                                Median = median(Score), Skewness=skew(Score),
+                         "Normally distributed"=ifelse(shapiro.test(Score)$p.value>0.05,"Yes","No"))
> t(Norm2)
                    [,1]          [,2]
Group               "Placebo"     "Caffeine"
Sample size         "20"          "20"
Mean                " 8.85"       "12.40"
Median              " 9.0"        "12.5"
Skewness            "-0.12839058" "-0.09918414"
Normally distributed "Yes"        "Yes"
>
> # test for homogeneity of variances
> library(car)
> leveneTest(Score ~ Group,center=mean, data=Indep)
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value Pr(>F)
group  1  0.1172  0.734
      38
```

Example: Independent-samples

3 Depending on which assumptions are met (or violated), perform one of the following tests:

```
> # independent t-test (if measurements are normally distributed)
> # t.test(Indep$score ~ Indep$Group, var.equal=TRUE) # gives same output
> t.test(Score ~ Group, var.equal=TRUE, data=Indep)

        Two Sample t-test

data:  Score by Group
t = -5.1674, df = 38, p-value = 7.862e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.940766 -2.159234
sample estimates:
 mean in group Placebo mean in group Caffeine
                  8.85                  12.40
```

Example: Independent-samples

```
> # Mann-Whitney U test (if measurements are not normally distributed)
> wilcox.test(Indep$Score ~ Indep$Group,exact=F)

        Wilcoxon rank sum test with continuity correction

data:  Indep$Score by Indep$Group
W = 49.5, p-value = 4.486e-05
alternative hypothesis: true location shift is not equal to 0

>
> # independent t-test (if measurements are normally distributed), but equal variances are not assumed
> t.test(Indep$Score ~ Indep$Group, var.equal=FALSE)

        Welch Two Sample t-test

data:  Indep$Score by Indep$Group
t = -5.1674, df = 37.424, p-value = 8.166e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.941469 -2.158531
sample estimates:
 mean in group Placebo mean in group Caffeine
                  8.85                  12.40
```

Example: Independent-samples

4 Effect size (standardised mean difference):

```
> # effect size for shooting accuracy ...standardised mean difference
> library(effsize)
> cohen.d(Indep$Score ~ Indep$Group)

Cohen's d

d estimate: -1.634066 (large)
95 percent confidence interval:
     lower      upper
-2.3733911 -0.8947399
```

Example: Paired-samples

Consider the experiment of reported and measured heights (in inches) of male statistics students. (Import the Paired t-test dataset)

Check the data properties and for anomalies such as extreme values or skewed distributions:

```
# need to change data from long to wide
library(tidyr)
Paired<-gather(Dep,Type,Height,2:3)
View(Paired)

# properties
is.factor(Paired$Type)
Paired$Type<-factor(Paired$Type,levels=c("Reported","Measured"))
is.numeric(Paired$Height)
```

Assumptions

```
# explore
ggplot(Paired,aes(x=Type, y=Height))+stat_boxplot(geom = "errorbar")+geom_boxplot()+
  labs(x = "", y = "Height (inches)")+
coord_cartesian(ylim=c(65, 85)) + scale_y_continuous(breaks=seq(65,85,5))

# normality
Norm4<-Paired %>% group_by(Type) %>% summarise("Sample size"=n(),Mean = mean(Height),
                Median = median(Height), Skewness=skew(Height),
                "Normally distributed"=ifelse(shapiro.test(Height)$p.value>0.05,"Yes","No"))
t(Norm4)
```

Example: Paired-samples

Depending on which assumptions are met (or violated), perform one of the following tests:

```
> # Proceed leaving outlier in
> t.test(Paired$Height ~ Paired$Type, paired=TRUE)

        Paired t-test

data:  Paired$Height by Paired$Type
t = 1.9751, df = 11, p-value = 0.07389
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1462991  2.7046324
sample estimates:
mean of the differences
               1.279167

> wilcox.test(Paired$Height ~ Paired$Type, paired=TRUE,conf.int=T,exact=F) # outputs CI as well

        Wilcoxon signed rank test with continuity correction

data:  Paired$Height by Paired$Type
V = 72, p-value = 0.01061
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.1500574 1.6999527
sample estimates:
(pseudo)median
     0.8280165
```

Example: Paired-samples

Effect size

```
> cohen.d(Paired$Height ~ Paired$Type, paired=TRUE)

Cohen's d

d estimate: 0.2567765 (small)
95 percent confidence interval:
      lower       upper
-0.01725042  0.53080336
```

Example: Paired-samples (outlier removed)

```
# Remove the third subject
library(dplyr)
Paired2<-filter(Paired,Case != "C") # create new dataframe with subject C removed

# copy and paste old script and change Paired to Paired2
ggplot(Paired2,aes(x=Type, y=Height))+stat_boxplot(geom = "errorbar")+
  geom_boxplot()+labs(x = "", y = "Height (inches)")+
  coord_cartesian(ylim=c(65, 75)) + scale_y_continuous(breaks=seq(65,75,1)) #+

# normality
Norm5<-Paired2 %>% group_by(Type) %>% summarise("Sample size"=n(),Mean = mean(Height),
                  Median = median(Height), Skewness=skew(Height),
                  "Normally distributed"=ifelse(shapiro.test(Height)$p.value>0.05,"Yes","No"))
t(Norm5)

# test and effect size
t.test(Paired2$Height ~ Paired2$Type, paired=TRUE)
wilcox.test(Paired2$Height ~ Paired2$Type, paired=TRUE,conf.int=T,exact=F) # outputs CI as well
cohen.d(Paired2$Height ~ Paired2$Type, paired=TRUE)
```

Parameters

- Deciding on an appropriate sample size must involve consideration of a variety of input parameters in advance of the study - e.g., level of significance (α), effect size, statistical power (1-β).
- The power of a test is the capacity of a test to detect statistical significance if it exists.

Required sample size calculation

```
> # test and effect size on triglyceride data
> t.test(Long$Response ~ Long$Time, paired=TRUE)

        Paired t-test

data:  Long$Response by Long$Time
t = 1.2, df = 15, p-value = 0.2487
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.91541  39.04041
sample estimates:
mean of the differences
                14.0625

> (effect<-cohen.d(Long$Response ~ Long$Time, paired=T))

Cohen's d

d estimate: 0.4811598 (small)
95 percent confidence interval:
     lower      upper
-0.3838207  1.3461403
>
> # required sample size for statistically significant effect
> power.t.test(delta=effect$estimate,sig.level = 0.05,power=0.8)

     Two-sample t test power calculation

              n = 68.77827
          delta = 0.4811598
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

One-sample *t*-test

- A manufacturer of high-performance automobiles produces disc brakes that must measure 322 millimeters in diameter. Quality control randomly draws 16 discs made by each of eight production machines and measures their diameters.
- This example uses the file brakes.sav. Use the Brakes dataset to determine whether or not the mean diameters of the brakes in each sample significantly differ from 322 millimeters.

Independent-samples t-test

- An analyst at a department store wants to evaluate a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months, and half received a standard seasonal ad.
- Open the creditpromo dataset and compare the spending of the two groups.

Paired-samples *t*-test

- A physician is evaluating a new diet for her patients with a family history of heart disease. To test the effectiveness of this diet, 16 patients are placed on the diet for 6 months. Their weights and triglyceride levels are measured before and after the study, and the physician wants to know if either set of measurements has changed.
- This example uses the dietstudy dataset, to determine whether there is a statistically significant difference between the pre- and post-diet weights and triglyceride levels of these patients.