ECON20003 QM2

Tutorial 3 Semester 1, 2022

Chin Quek

Department of Economics

March 2022

Objectives

- Numerical Descriptive Measures
- Introduction to Statistical Inference
  - Describing a single population (Confidence Interval and Hypothesis testing)
- Nonparametric Tests for a Population Central Location
  - One-sample Sign test
  - One-sample Wilcoxon signed rank test

Reminder: Always create an R Project and R script when starting on a new exercise.

Install the `pastecs` package if you have not already done so, then in the *R Script* file enter the line:

```
library(pastecs)        # stat.desc()
library(DescTools)      # SignTest()
library(exactRankTests) # wilcox.exact()
```

Numerical Descriptive Measures

Here's a preview of the dataset `hs0`, containing 195 observations and 11 variables.

```
##   gender  id race ses schtyp  prgtype read write math science socst
## 1      0  70    4   1      1  general   57    52   41      47    57
## 2      1 121    4   2      1   vocati   68    59   53      63    61
## 3      0  86    4   3      1  general   44    33   54      58    31
## 4      0 141    4   3      1   vocati   63    44   47      53    56
## 5      0 172    4   2      1 academic   47    52   57      53    61
## 6      0 113    4   2      1 academic   44    52   51      63    61
```

Using R, how do we compute the:

- Mean of `read`?
- Quantile values of `read`, `write`, `math`, `science` and `socst`?

Recall that package `pastecs` has an easy to use function called `stat.desc` to display a table of descriptive statistics for a list of variables.

- How would we obtain descriptive statistics using `stat.desc()` and rounded off to 2 decimal places?

```r
attach(hs0)

mean(read)
```

```
## [1] 51.98974
```

```r
sapply(hs0[7:11], quantile)
```

```
##      read write math science socst
## 0%     28  31.0   33      26    26
## 25%    44  45.5   45      44    46
## 50%    50  54.0   52      53    52
## 75%    60  60.0   58      58    61
## 100%   76  67.0   75      74    71
```

```r
summary(hs0[7:11])
```

```
##       read           write            math          science     
##  Min.   :28.00   Min.   :31.00   Min.   :33.00   Min.   :26.00  
##  1st Qu.:44.00   1st Qu.:45.50   1st Qu.:45.00   1st Qu.:44.00  
##  Median :50.00   Median :54.00   Median :52.00   Median :53.00  
##  Mean   :51.99   Mean   :52.74   Mean   :52.43   Mean   :51.66  
##  3rd Qu.:60.00   3rd Qu.:60.00   3rd Qu.:58.00   3rd Qu.:58.00  
##  Max.   :76.00   Max.   :67.00   Max.   :75.00   Max.   :74.00  
##      socst      
##  Min.   :26.00  
##  1st Qu.:46.00  
##  Median :52.00  
##  Mean   :52.36  
##  3rd Qu.:61.00  
##  Max.   :71.00
```

```r
round(stat.desc(hs0[-6], 2))
```

```
##              gender   id race ses schtyp read write math science socst
## median            1  100    4   2      1   50    54   52      53    52
## mean              1  100    3   2      1   52    53   52      52    52
## SE.mean           0    4    0   0      0    1     1    1       1     1
## CI.mean.0.95      0    8    0   0      0    1     1    1       1     2
## var               0 3285    1   1      0  103    91   85      97   114
## std.dev           0   57    1   1      0   10    10    9      10    11
## coef.var          1    1    0   0      0    0     0    0       0     0
```

Here is a list of some commonly used descriptive statistics for a single variable in R

- `mean()`: arithmetic mean
- `sd()`: *sample* standard deviation
- `var()`: *sample* variance
- `median()`: median

Pages 5 and 6 are important for interpreting the various descriptive statistics!!!

```
One Sample t-test

data:  Expenses

t = 0.024844, df = 59, p-value = 0.9803

alternative hypothesis: true mean is not equal
to 170

95 percent confidence interval:

 101.5273 240.1944

sample estimates:

mean of x

 170.8608
```

A Student t distribution or t distribution is a .....(a)..... distribution used in .....(b)..... when the .....(c)..... variance is not known. The t-statistics for .....(d)..... is a standardized value of the .....(e)..... when .....(f)..... is unknown and replaced by .....(g).....

a. continuous / discrete
b. descriptive statistics / statistical inference
c. population / sample
d. μ / η
e. sample mean / sample variance
f. σ / s
g. σ / s

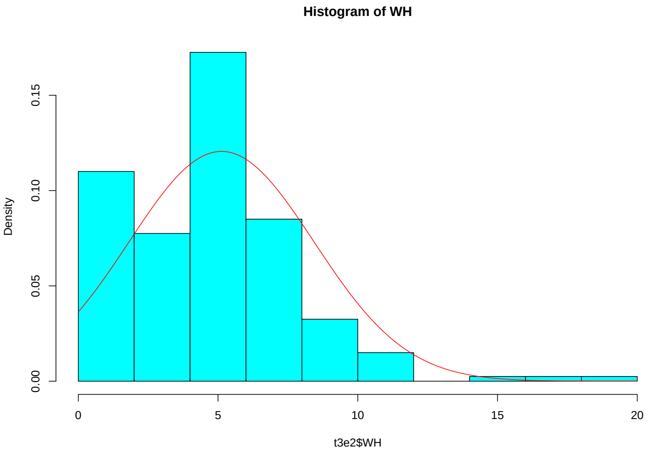
Checking for normality

What are the different ways to check for normality?

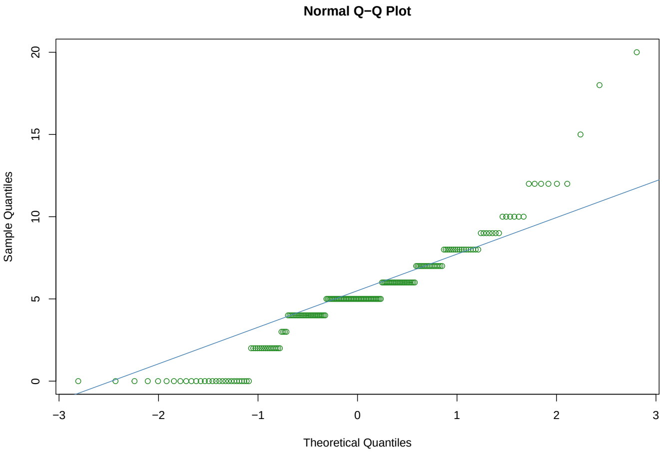
- Is it sufficient to perform just one check?

Using the data from Exercise 2, the following graphs are plotted.

```r
hist(t3e2$WH, freq =FALSE, col = "cyan", main = "Histogram of WH")
lines(seq(0, 20, by = 0.1),
      dnorm(seq(0, 20, by = 0.1),
            mean(t3e2$WH), sd(t3e2$WH)),
      col= "red")
```



```r
qqnorm(t3e2$WH, main = "Normal Q-Q Plot",
xlab = "Theoretical Quantiles", ylab = "Sample Quantiles" ,
        col = "forestgreen")
qqline(t3e2$WH, col = "steelblue")
```



We can also check normality numerically with descriptive measures. Interpret the results below.

```
round(stat.desc(t3e2$WH),3)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
##      200.000       28.000        0.000        0.000       20.000       20.000
##          sum       median         mean      SE.mean CI.mean.0.95          var
##     1025.000        5.000        5.125        0.234        0.461       10.954
##      std.dev     coef.var
##        3.310        0.646
```

```
round(stat.desc(t3e2$WH, basic = FALSE, desc = FALSE,
                norm = TRUE), 3)
```

```
##    skewness    skew.2SE    kurtosis    kurt.2SE normtest.W normtest.p
##       0.778       2.262       2.278       3.329      0.925      0.000
```

Kurtosis in the R output refers to Excess Kurtosis

Checking normality using Shapiro-Wilk hypothesis test

```
shapiro.test(t3e2$WH)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  t3e2$WH
## W = 0.92484, p-value = 1.338e-08
```

- How do we interpret this R output?

Key summary:

- If the sample data is more or less normally distributed, the points on the QQ plot are on or close to this reference line, while **substantial departures** from it are indicative of a lack of normality.

Key summary:

- If the sample data is more or less normally distributed, the points on the QQ plot are on or close to this reference line, while **substantial departures** from it are indicative of a lack of normality.

  Evaluations of graphs for normality are INSUFFICIENT. We must supplement these visual checks with numerical descriptive statistics checks!

Key summary:

- If the sample data is more or less normally distributed, the points on the QQ plot are on or close to this reference line, while **substantial departures** from it are indicative of a lack of normality.

  Evaluations of graphs for normality are INSUFFICIENT. We must supplement these visual checks with numerical descriptive statistics checks!

- If the mean and median are relatively close together this would indicate that the data is symmetrically distributed.
- Using S and K: Ratio values of above 1 would indicate that at the 5% significance level that S is different from zero and K is different from 3, and hence the population is unlikely normal.
- For Shapiro-Wilk test, the null hypothesis is that the population is normally distributed.

Are normality checks very important?

YES!

Please take note!

Why do we still go through the normality checks? Can't we just mention that we have LARGE sample sizes?

Are normality checks very important?

YES!

Please take note!

Why do we still go through the normality checks? Can't we just mention that we have LARGE sample sizes?

While the sample size is large enough to rely on CLT, then we can assume normality for the distribution of say $\bar{X}$. However, this is for the distribution of $\bar{X}$ and does not apply for the X.

- As an example, for one sample, if X is normal, $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is a t random variable.

So it is still important to check if the sampled population is normal when the population standard deviation is **unknown**.

Exercise 2

A growing concern for educators in Australia is the number of teenagers who have part-time jobs while they attend high school. It is generally believed that the amount of time teenagers spend working is deducted from the amount of time devoted to schoolwork. To investigate this problem, a school guidance counsellor took a random sample of 200 15-year-old high school students and asked how many hours per week each worked at a part-time job (WH). The results are stored in the *t3e2* file. Estimate with 95% confidence the mean amount of time all 15-year-old high school students devote per week to part-time jobs.

What are the statistics provided in the background?

The sample ($n = 200$) is drawn form a population of 15 year old students in Australia. The variable is the number of part-time hours worked per week.

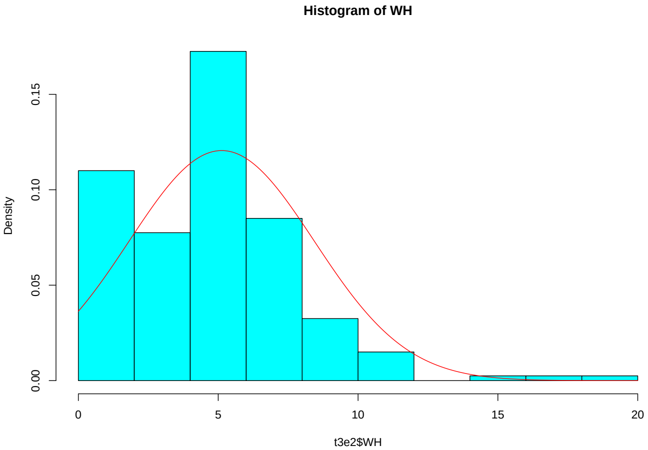
What is/are the necessary condition(s) that need to be met to estimate the 95% confidence interval?

Hints:

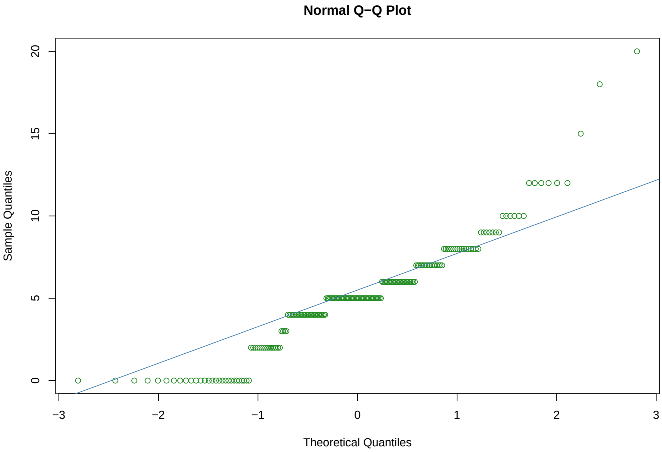
- Is the population standard deviation (σ) given in the question?
- Is the data normally distributed?

Checking normality visually

Overlaying the new histogram with a normal curve that is characterised by the same mean and standard deviation than the sample data



Checking normality visually



Checking normality numerically with descriptive measures

Comparison of sample means and medians

Using sample means and medians, what can we conclude about the distribution of the data?

```
##  nbr.val nbr.null   nbr.na      min      max    range      sum
##      200       28        0        0       20       20     1025

##       median         mean      SE.mean CI.mean.0.95          var
##        5.000        5.125        0.234        0.461       10.954

##  std.dev coef.var
##    3.310    0.646
```

Checking normality numerically with descriptive measures

Comparison of sample means and medians

Using sample means and medians, what can we conclude about the distribution of the data?

```
##   nbr.val nbr.null    nbr.na       min       max     range       sum
##       200       28         0         0        20        20      1025

##        median          mean       SE.mean CI.mean.0.95           var
##         5.000         5.125         0.234        0.461        10.954

##   std.dev coef.var
##     3.310    0.646
```

- Sample mean of *WH* is 5.125 and its sample median is 5.000.
- They are not equal to each other, but very similar. Their ratio is 5.125 / 5 = 1.025, so the sample mean is only 2.5% larger than the sample median.
- Hence, the distribution of *WH* might be symmetrical.

Checking normality numerically with descriptive measures

Skewness and kurtosis

```
##    skewness    skew.2SE    kurtosis    kurt.2SE normtest.W normtest.p
##       0.778       2.262       2.278       3.329      0.925      0.000
```

Checking normality numerically with descriptive measures

Skewness and kurtosis

```
##    skewness   skew.2SE   kurtosis   kurt.2SE normtest.W normtest.p
##       0.778      2.262      2.278      3.329      0.925      0.000
```

- Skewness is 0.778: positive suggesting that this sample of *WH* is skewed to the right
- Excess kurtosis is 2.278: positive, and hence kurtosis $= 2.278 + 3 = 5.278$ is larger than 3, suggesting that this sample of WH is leptokurtic, i.e. has thinner tails than the normal distribution (more "peakedness").
- The point estimates of *SK* and *K* imply that this sample of *WH* is certainly *not* normally distributed.

Checking normality numerically with descriptive measures

Skewness and kurtosis

```
##    skewness    skew.2SE    kurtosis    kurt.2SE normtest.W normtest.p
##       0.778       2.262       2.278       3.329      0.925      0.000
```

- Skewness is 0.778: positive suggesting that this sample of *WH* is skewed to the right
- Excess kurtosis is 2.278: positive, and hence kurtosis $= 2.278 + 3 = 5.278$ is larger than 3, suggesting that this sample of WH is leptokurtic, i.e. has thinner tails than the normal distribution (more "peakedness").
- The point estimates of *SK* and *K* imply that this sample of *WH* is certainly *not* normally distributed.

Note

Note, however, that this does NOT allow us to claim that the population of *WH* is not normal either.
To do so, we need to consider their standard errors as well.

Standard errors of SK and K

$$s_{\widehat{SK}} = \sqrt{\frac{6}{n}} = \sqrt{\frac{6}{200}} = 0.173$$

$$s_{\widehat{K}} = \sqrt{\frac{24}{n}} = \sqrt{\frac{24}{200}} = 0.346$$

$$\frac{\widehat{SK}}{2s_{\widehat{SK}}} = \frac{0.778}{2 \times 0.173} = 2.249 \text{ and}$$

$$\frac{\hat{K} - 3}{2s_{\widehat{K}}} = \frac{2.778}{2 \times 0.346} = 3.292$$

How do we use the earlier R printout to approximate 2.249 and 3.292? What do these values indicate about normality?

The calculated ratio values are a bit larger than the ones reported by R because we used approximate standard errors while R used the exact ones.

Checking normality numerically using SW test

```
##
##  Shapiro-Wilk normality test
##
## data:  t3e2$WH
## W = 0.92484, p-value = 1.338e-08
```

What can we conclude from the results of the SW test?

Checking normality numerically using SW test

```
##
##  Shapiro-Wilk normality test
##
## data:  t3e2$WH
## W = 0.92484, p-value = 1.338e-08
```

What can we conclude from the results of the SW test?

The p-value of this test is 1.338e-08 $=$ 0.00000001338. It is practically zero, so we can reject the null hypothesis of normality at any reasonable significance level.

So, all things considered, there is no ground to assume that the population of *WH* is normally distributed.

Note, however, that the histogram does not exhibit *extreme* non-normality, so at the given *large* sample size (n = 200), we can still use the CI estimator $\bar{x} \pm t_{(\alpha/2,n-1)} s_{\bar{x}}$.

We already have $\bar{x}$, the standard error $s_{\bar{x}}$ and to get the critical t value, $t_{(\alpha/2,n-1)}$, we can use

```
qt(0.975,199)
```

```
## [1] 1.971957
```

Putting all this together gives

$$\bar{x} \pm t_{(\alpha/2,n-1)} s_{\bar{x}} = 5.125 \pm (1.972)(0.234) = 5.125 \pm 0.461 = [4.464, 5.586]$$

we conclude with 95% confidence that the mean number of hours all 15-year-old high school students devote per week to part-time jobs is within the [4.664; 5.586] interval.

Exercise 3

In quality control applications of hypothesis testing, the null and alternative hypotheses are frequently specified as

H_0 :The production process is performing satisfactorily

H_A :The process is performing in an unsatisfactory manner

An injection moulder produces plastic golf tees. The process is designed to produce tees with a mean weight of 0.250 ounce.

To investigate whether the injection moulder is operating satisfactorily, 40 tees were randomly sampled from the last hour's production.

Their weights (in ounces) are saved in the *t3e3* file.

> *Do the data provide sufficient evidence at the 1% significance level to conclude that the process is not operating satisfactorily?*

The variable of interest is the Weight of a tee, and the null and alternative hypotheses are:

$H_0 : \mu = 0.25$

$H_A : \mu \neq 0.25$

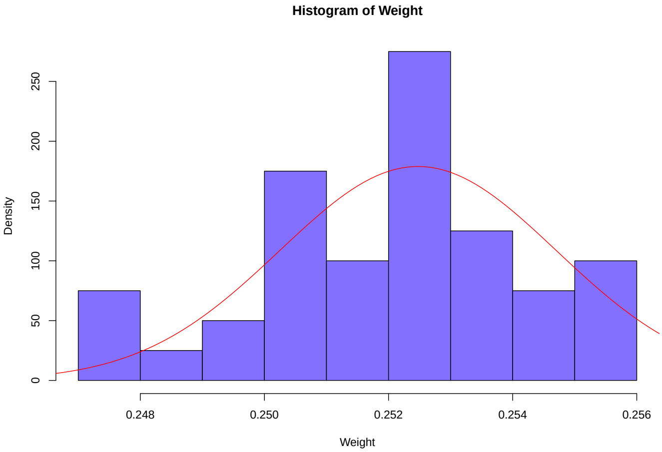
Both the shape and the standard deviation of the population of tee weights are unknown.

However, the sample size is reasonably large ($n = 40$), so the sample mean is approximately normally distributed. Consequently, granted that the population is not extremely non-normal, the test statistic is

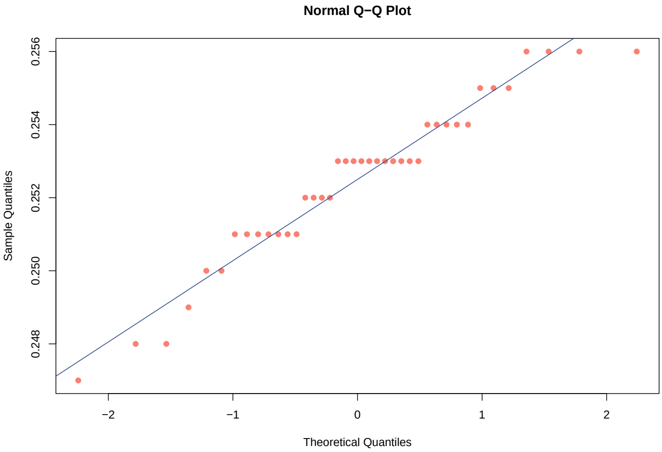
$$t = \frac{\bar{x} - \mu_{x,0}}{s_{\bar{x}}} \sim t_{n-1}$$

Based on the following outputs, do we have evidence of normality?

Histogram evidence



Normal QQ plot evidence



Descriptive statistics and the SW test

```
##         median          mean       SE.mean CI.mean.0.95           var       std.dev
##         0.2530        0.2525        0.0004       0.0007        0.0000        0.0022
##       coef.var      skewness      skew.2SE     kurtosis      kurt.2SE    normtest.W
##         0.0088       -0.4411       -0.5900      -0.2490       -0.1699        0.9501
##     normtest.p
##         0.0763
```

Calculation of observed test statistic:

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} - \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.2525 - 0.2500}{0.002/\sqrt{40}} = 7.187$$

How can we obtain the critical values?

The critical values are:

$$\pm t_{\alpha/2,n-1} = \pm t_{0.005,39} \approx t_{0.005,40} = \pm 2.704$$

or use

```r
qt(0.995,39)
```

```
## [1] 2.707913
```

What is the decision and conclusion?

Hence, we can reject H_0 at the 1% significance level if the absolute value of the observed test statistic is larger than 2.708.

What about a one-tailed test?

What is the decision and conclusion?

Hence, we can reject H_0 at the 1% significance level if the absolute value of the observed test statistic is larger than 2.708.

What about a one-tailed test?

Using R, run the following:

```
t.test(t3e3$Weight, mu = 0.25, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  t3e3$Weight
## t = 7.0188, df = 39, p-value = 2.019e-08
## alternative hypothesis: true mean is not equal to 0.25
## 99 percent confidence interval:
##  0.2515201 0.2534299
## sample estimates:
## mean of x
##  0.252475
```

The one-sample **sign** test assumes that:

i. The data is a random sample of independent observations.
ii. The variable of interest is qualitative or quantitative.
iii. The measurement scale is at least ordinal.

The one-sample **Wilcoxon signed ranks** test serves the same purpose than the sign test, but some of the assumptions required by this test are stronger.

i. The data is a random sample of independent observations.
ii. The variable of interest is quantitative and continuous.
iii. The measurement scale is interval or ratio.
iv. The distribution of the sampled population is symmetric.

Exercise 4

A courier service in Brisbane advertises that its average delivery time is less than six hours for local deliveries. A random sample of the amount of time this courier takes to deliver packages to an address across town produced the delivery times (DT, rounded to the nearest hour) saved in the *t3e4* Excel file.

Does this sample provide sufficient evidence to support the courier's advertisement, at the 5% level of significance?

> *Let's have a think about an appropriate statistical technique to answer this question.*

Exercise 4

A courier service in Brisbane advertises that its average delivery time is less than six hours for local deliveries. A random sample of the amount of time this courier takes to deliver packages to an address across town produced the delivery times (DT, rounded to the nearest hour) saved in the *t3e4* Excel file.

Does this sample provide sufficient evidence to support the courier's advertisement, at the 5% level of significance?

> *Let's have a think about an appropriate statistical technique to answer this question.*

The sample size is only 10, far too small to check normality with reasonable certainty.

- Better to rely on a nonparametric procedure, the sign test or/and the Wilcoxon signed ranks test.

  Also recall how the Wilcoxon signed ranks test requires that the sampled population is symmetric. This cannot be verified here.

$H_0 : \eta = 6$

$H_A : \eta < 6$

Sign test

- $n < 10$: use binomial table to obtain p-value
- $n \geq 10$: binomial distribution can be approximated using a normal distribution, $B(n, 0.5) \sim N(0.5n, 0.5\sqrt{n})$

Wilcoxon signed rank test

- $6 \leq n \leq 30$: use wilcoxon signed rank critical values table
- $n > 30$: sampling distribution of T can be approximated with a normal distribution

$$T \sim N(\mu_T, \sigma_T) \qquad \mu_T = \frac{n(n+1)}{4} \qquad \sigma_T^2 = \frac{n(n+1)(2n+1)}{24}$$

Sign Test using R

- $n^* = 8$, $S_- = 5$, $S_+ = 3$

To reject H_0, the number of negative deviations needs to be sufficiently large.

Using S_+ as the test statistic, p-value $= \Pr(S \leq 3) = 0.3633$

```
pbinom(3,8,0.5)
```

```
## [1] 0.3632813
```

Using S_- as the test statistic, p-value $= \Pr(S \geq 5) = 1 - \Pr(S \leq 4) = 0.3633$

```
1-pbinom(4,8,0.5)
```

```
## [1] 0.3632813
```

```
pbinom(4,8,0.5,lower.tail=F)
```

```
## [1] 0.3632813
```

```r
binom.test(3, 8, alternative = "less")
```

```
##
##  Exact binomial test
##
## data:  3 and 8
## number of successes = 3, number of trials = 8, p-value = 0.3633
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.7107592
## sample estimates:
## probability of success
##                  0.375
```

```r
binom.test(5, 8, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  5 and 8
## number of successes = 5, number of trials = 8, p-value = 0.3633
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.2892408 1.0000000
## sample estimates:
## probability of success
##                  0.625
```

$H_0 : \eta = 6$

$H_A : \eta < 6$

```
attach(t3e4)
SignTest(DT, mu = 6, alternative = "less")
```

```
##
##  One-sample Sign-Test
##
## data:  DT
## S = 3, number of differences = 8, p-value = 0.3633
## alternative hypothesis: true median is less than 6
## 98.9 percent confidence interval:
##  -Inf    7
## sample estimates:
## median of the differences
##                       5.5
```

Wilcoxon signed rank test

Manually, bit more tedious, having to rank the non-zero *absolute* deviations.

- Then take the sum of the ranks assigned to negative and positive deviations.
- For a left-tail test, reject H_0 if $T = T_+ \leq T_L$. Otherwise, reject H_0 if $T = T_- \geq T_U$.

Using R:

```
wilcox.exact(t3e4$DT, mu = 6,
             alternative = "two.sided")
```

```
##
##  Exact Wilcoxon signed rank test
##
## data:  t3e4$DT
## V = 13.5, p-value = 0.6016
## alternative hypothesis: true mu is not equal to 6
```

DT	Median under H_0	Deviation	Abs deviation	Rank
7	6	1	1	
3	6	-3	3	
4	6	-2	2	
6	6	0	0	
10	6	4	4	
5	6	-1	1	
6	6	0	0	
4	6	-2	2	
3	6	-3	3	
8	6	2	2	

Exercise 5

In Exercise 3, we tried to find out whether the injection moulder of plastic golf tees was operating satisfactorily in the sense that the mean weight of tees was not significantly different from the specified 0.250 ounce.

For the sake of illustration, let's try a non-parametric test.

a. Use the one sample sign test to determine whether the median weight of tees differs from 0.250 ounce.

Set up the hypotheses:

$H_0 : \eta = 0.25$

$H_A : \eta \neq 0.25$

Sign Test using R

- $n^* = 38$, $S_- = 4$, $S_+ = 34$

For a two-tail test, p-value $= 2 \times \min\left(p_R = \Pr(S \geq S_+),\ p_L = \Pr(S \leq S_+)\right)$

```
# Pr(S>=34)
1-pbinom(33,38,0.5)
```

```
## [1] 3.019268e-07
```

```
# Pr(S<=34)
pbinom(34,38,0.5)
```

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

```
# p-value
2*(1-pbinom(33,38,0.5))
```

```
## [1] 6.038536e-07
```

```r
binom.test(34, 38, conf.level = 0.99)
```

```
##
##  Exact binomial test
##
## data:  34 and 38
## number of successes = 34, number of trials = 38, p-value = 6.039e-07
## alternative hypothesis: true probability of success is not equal to 0.5
## 99 percent confidence interval:
##  0.7042125 0.9817423
## sample estimates:
## probability of success
##              0.8947368
```

```r
SignTest(t3e3$Weight, mu = 0.25, conf.level = 0.99)
```

```
##
##  One-sample Sign-Test
##
## data:  t3e3$Weight
## S = 34, number of differences = 38, p-value = 6.039e-07
## alternative hypothesis: true median is not equal to 0.25
## 99.4 percent confidence interval:
##  0.251 0.254
## sample estimates:
## median of the differences
##                     0.253
```

The p-value is practically zero (6.039e-07), so we can reject the null hypothesis at the 1% significance level.

```r
SignTest(t3e3$Weight, mu = 0.25, conf.level = 0.99)
```

```
##
##  One-sample Sign-Test
##
## data:  t3e3$Weight
## S = 34, number of differences = 38, p-value = 6.039e-07
## alternative hypothesis: true median is not equal to 0.25
## 99.4 percent confidence interval:
##  0.251 0.254
## sample estimates:
## median of the differences
##                      0.253
```

The p-value is practically zero (6.039e-07), so we can reject the null hypothesis at the 1% significance level.

Using normal approximation to the binomial distribution

$$S \sim N\left(0.5n^*, 0.5\sqrt{n^*}\right)$$

Wilcoxon signed ranks test using R

```r
wilcox.exact(t3e3$Weight, mu = 0.25)
```

```
##
##  Exact Wilcoxon signed rank test
##
## data:  t3e3$Weight
## V = 693, p-value = 1.483e-07
## alternative hypothesis: true mu is not equal to 0.25
```

The p-value is practically zero, so H_0 can be rejected at any reasonable significance level.