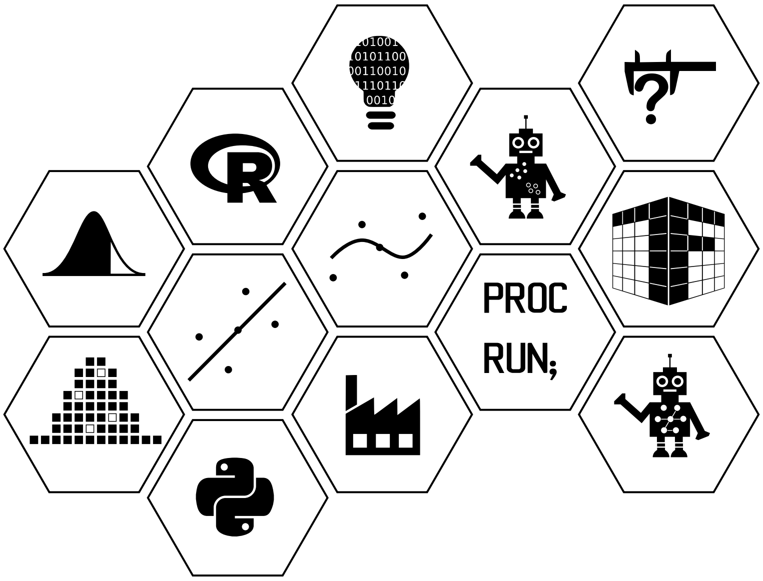
Learning from Data/Data Science Foundations

Week 4: Classical tests



DATA ANALYTICS
GLASGOW

Classical statistical tests

In the initial weeks of the course we have introduced the idea of using summary statistics from a sample of data to enable us to infer conclusions about population parameters for a variable of interest. Summary statistics have variability and, hence, uncertainty associated with them and we need to account for this appropriately in order to make reliable conclusions for a population of interest.

Sometimes, it is plausible to assume that we actually know the distributional form of the variable of interest (in the population). In this section, we are going to start by assuming that a sample of continuous data comes from a normally-distributed population. In other words, we will adopt a normal probability model for the population. This means that, implicitly, we are claiming to know everything about the population apart from the values of the model's two unknown parameters: the population mean (μ) and variance (σ^2). This is known as *parametric inference*.

The more we believe we know about the population to begin with, the stronger the inferences we should be able to draw on the basis of the data. On the other hand, the inferences we draw could well be invalid if the assumptions we choose to make are invalid.

Parametric statistical inference naturally focuses on the unknown parameters of the model. In particular, we might wish to:

- find a point estimate (*best guess*) of a parameter's value;

use our distributional assumption to

- find an interval estimate (*range of plausible values*) for a parameter;
- test hypotheses about a parameter.

This section will cover statistical tests that are appropriate for continuous and binary data that have arisen from one and two populations.

Specifically, we'll investigate inference for the population mean μ and the population proportion θ. We will use the sample statistics of the sample mean and sample proportion as estimates of these population parameters. We refer to them as point estimates and denote them as follows: $\hat{\mu}$ and $\hat{\theta}$.

Week 4 learning material aims

The material in week 4 covers:

- calculating confidence intervals and constructing hypothesis tests for one/two sample problems;

- computing the intervals and tests in `R`;
- interpreting output from confidence intervals and hypothesis tests.

T-tests and confidence intervals

In the case of continuous data that we can assume have arisen from a normal distribution, we can use simple summary statistics for the mean and standard deviation of our summary data along with the properties of the normal distribution introduced in **Probability and Stochastic Models** or **Probability and Sampling Fundamentals/Sampling Fundamentals** to perform inference for our populations of interest. Specifically, we might be interested in performing hypothesis tests regarding the population mean and computing confidence intervals for the population mean.

One-sample t-interval and t-test

Suppose we are interested in constructing a confidence interval for the population mean μ.

If we assume that $X_1, X_2, \ldots, X_n$ are independent random variables, that are all normally distributed $N(\mu, \sigma^2)$, then it can also be shown (see **Probability and Stochastic Models** or **Probability and Sampling Fundamentals/Sampling Fundamentals**) that $\bar{X}$ is normally distributed. So,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

The only random quantity in the formula for Z is $\bar{X}$. This means that any probability statement we make about Z is equivalent to a probability statement about $\bar{X}$. This is an example of a **pivotal function**.

Definition 1

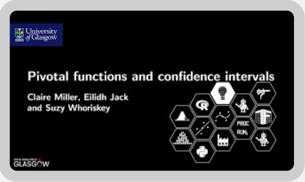
A **pivotal function** is a function of the data, $\mathbf{X}$, and a parameter of interest, θ, which, when regarded as a random variable calculated at θ_T (the true value of θ), has a probability distribution whose form does not depend on any unknown parameter. We usually denote a pivotal function by $PIV(\theta_T, \mathbf{X})$. Note: here we are using θ to represent any population parameter.

This video explains the idea of a pivotal function and describes how to use a pivotal function to create a 95% confidence interval:

Video

Pivotal functions and interval estimation

Duration 13:57



Pivotal functions and confidence intervals

Claire Miller, Eilidh Jack and Suzy Whoriskey

We found a pivotal function for μ_T (the true value of μ) above (assuming σ is known):

$$PIV(\mu_T, \mathbf{X}) = \frac{\bar{X} - \mu_T}{\sigma_T/\sqrt{n}} = Z.$$

This pivotal function has an $N(0,1)$ distribution whatever the value of μ_T.

Now assume that $X_1, X_2, \ldots, X_n$ are independent random variables, each with a $N(\mu_T, \sigma_T^2)$ distribution, and that we wish to estimate the population mean μ, but that both μ_T and σ_T (the population standard deviation) are unknown. This is now a two-parameter model.

The pivotal function for μ_T found previously is not a pivotal function in this case since it depends on the unknown parameter σ. A pivotal function for μ_T in this case is found by replacing σ_T with its estimator, s (the sample standard deviation).

It can be shown that:

$$t = \frac{\bar{X} - \mu_T}{s/\sqrt{n}} \sim t(n-1)$$

where $t(n-1)$ is the Student's t distribution with $n-1$ degrees of freedom.

Supplement 1

For more about the t-distribution see chapter 2 of the e-book: Foundations of Applied Statistical Methods (linked to on the Moodle page).

As shown in the first video for this week, all t distributions are symmetric and unimodal with expected value 0, just like the $N(0,1)$ distribution. However, t distributions are less peaked and more spread out than the $N(0,1)$ distribution.

The statistic t is identical to Z, except that the known value σ_T is replaced by an estimate s. This further source of uncertainty causes the distribution to be more spread out. As $n \rightarrow \infty$, then the $t(n-1)$ distribution tends towards the $N(0,1)$ distribution.

Having identified a suitable pivotal function (if one exists), then we can construct a confidence interval for our parameter of interest. There is no general method for deriving a pivotal function.

It is possible to produce $100c\%$ confidence intervals for any value of c in the range $0 < c < 1$.

Let $t_{1-\frac{(1-c)}{2}}(n-1)$ denote the value of a t random variable such that:

$$P(t \leq t_{1-\frac{(1-c)}{2}}(n-1)) = 1 - \frac{(1-c)}{2}.$$

For example, let $t_{0.975}(n-1)$ denote the value of a t random variable such that:
$P(t \leq t_{0.975}(n-1)) = 0.975, c = 0.95.$

Since the $t(n-1)$ distribution is symmetric around 0, there is probability 0.95 (or 95%) that:

$$-t_{0.975}(n-1) \leq t = \frac{\bar{X} - \mu_T}{s/\sqrt{n}} \leq t_{0.975}(n-1)$$

i.e.

$$\mu_T \leq \bar{X} + t_{0.975}(n-1)s/\sqrt{n}$$

and

$$\mu_T \geq \bar{X} - t_{0.975}(n-1)s/\sqrt{n}$$

simultaneously.

This means that **a 95% confidence interval for the population mean**, $\boldsymbol{\mu}$ (when σ is unknown), is:

$$\left(\bar{x} - t_{0.975}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

i.e.

$$\left(\bar{x} \pm t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

where $\bar{x}$ is the observed value of $\bar{X}$.

Intervals with 95% coverage are obtained using $t_{0.975}()$ and intervals with 99% coverage are obtained using $t_{0.995}()$. The greater the coverage required, i.e. the larger the value of c that is used, the wider the confidence interval becomes.

In general, we say that the form of a confidence interval is as follows:

$$\text{estimate} \pm \text{t-value} \times \text{estimated standard error}$$

Supplement 2

A visualisation of the t-distribution and normal distribution relationship as the degrees of freedom for the t-distribution change is provided at the link here: Normal and t distribution visualisation

Example 1

IQ data cont.....

As a simple example consider the IQ data introduced in week 3. We will produce a 95% confidence interval for the population mean IQ, $n = 39$.

IQ <- c(94, 89, 83, 99, 94, 90, 94, 87, 89, 92, 92, 94, 91, 86, 90, 91, 82, 86, 88, 97, 96, 95, 87, 103, 91, 87, 91, 89, 98, 102, 105, 95, 88, 92, 90, 101, 98, 96, 83)

The summary statistics are:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3595}{39} = 92.18$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 30.68.$$

The 97.5th percentile of the t-distribution for $n-1$ degrees of freedom can be found by referring to a statistical table containing the quantiles of a t-distribution. The first video for this week illustrates how to find the appropriate value from the statistical table for the t-distribution by locating the value for the degrees of freedom on the rows of the table and the quantile of interest in the columns. A copy of this table is available on Moodle for this week's learning material.

The appropriate value can also be found in `R` at the 97.5th percentile or 0.975 quantile for df=$n-1$, which for the IQ example is 38 using:

```
qt(0.975, df=38)
```

```
R Console

[1] 2.024394
```

A 95% confidence interval for the population mean IQ μ_T (when σ_T is unknown) is:

$$\left(\bar{x} \pm t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

$$\left(92.18 - t_{0.975}(39-1)\frac{\sqrt{30.68}}{\sqrt{39}}, 92.18 + t_{0.975}(39-1)\frac{\sqrt{30.68}}{\sqrt{39}}\right)$$

$$t_{0.975}(38) = 2.024$$

$$\left(92.18 - 2.024\frac{\sqrt{30.68}}{\sqrt{39}}, 92.16 + 2.024\frac{\sqrt{30.68}}{\sqrt{39}}\right)$$

$$(92.18 - 2.024 \times (0.8869), 92.18 + 2.024 \times (0.8869))$$

$$(90.385, 93.975)\,.$$

Conclusion

It can therefore be concluded that the population mean IQ is highly likely to lie in the range 90.39 to 93.98, with a point estimate for $\hat{\mu}$ of 92.18.

Note: we would usually round the results here to a sensible number of significant figures or decimal places. Usually either 2 decimal places or 3 significant figures is a useful rule.

Hypothesis Testing

Reminder: In hypothesis testing we ask if the data provide sufficient evidence to reject the null hypothesis. If not we do not reject the null hypothesis.

Framework

- Specify H$_0$ and H$_1$;
- Define a test statistic (TS);
- Compute the observed value of the TS from sample data.

A p-value allows us to assess if we have evidence to reject H$_0$ or if we cannot reject H$_0$ based on the evidence. Typically, **Reject** H$_0$ if the p-value is < 0.05 (for a two-sided test), else **do not reject**.

The same pivotal function that was used above may also be used to define a test statistic and hence carry out hypothesis tests about the parameter μ.

Example 2

IQ data cont.....

Suppose that we wish to investigate the following hypotheses about the population mean μ_T for the IQ data:

$\text{H}_0 : \mu_T = 91$

$\text{H}_1 : \mu_T \neq 91$

Under H$_0$:

$$t = \frac{\bar{X} - \mu_T}{s/\sqrt{n}} \sim t(n-1).$$

This means that t is a suitable test statistic. We reject H$_0$ in favour of H$_1$ when $|t|$ is too *large* to be consistent with H$_0$.

Adopting a significance level of $\alpha = 0.05$, we reject H$_0$ in favour of H$_1$ if:

$$|t| > t_{0.975}(n-1)$$

This very famous statistical procedure is known as a **one-sample t-test**.

The conclusions drawn from the one-sample t-test (with a two-sided alternative hypothesis) are guaranteed to agree with those from the confidence interval previously discussed, in the following sense. The null hypothesis H$_0$: $\mu_T = 91$ is rejected at a significance level of $\alpha = 0.05$ if and only if a $100(1-\alpha)\%$ confidence interval for μ_T does not include the value 91.

The function `t.test()` can be used in `R` to perform this test and produce both a *p*-value for the hypothesis test and a confidence interval.

The `R` command and results are displayed below.

```
IQ <- c(94,  89,  83, 99, 94, 90, 94,  87,  89, 92, 92,
94,  91,  86, 90, 91, 82, 86,  88,  97, 96, 95,
87, 103,  91, 87, 91, 89, 98, 102, 105, 95, 88,
92,  90, 101, 98, 96, 83)

t.test(IQ, mu=91)
```

```
R Console
One Sample t-test


data:  IQ
t = 1.3299, df = 38, p-value = 0.1915
alternative hypothesis: true mean is not equal to 91
95 percent confidence interval:
90.38404 93.97493
sample estimates:
mean of x
92.17949
```

Conclusion

Since the *p*-value is > 0.05 at 0.192, we do not reject H$_0$ and conclude that there is insufficient evidence that the population mean IQ is different from 91.

(Note: the interpretation of the confidence interval (90.4, 94.0) has been discussed previously. Since the confidence interval contains the value 91, it provides the same conclusion as the p-value that there is insufficient evidence that the population mean is different from 91. Also, note that *t=1.3299* referred to in the output is the observed value of the test statistic.)

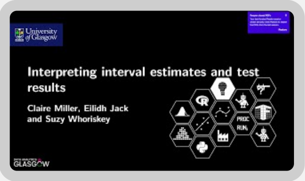
If we now consider the general population mean IQ of 100, since 100 is not contained in this 95% confidence interval then 100 does not appear to be a plausible value for the population mean based on the evidence from this sample of data. Why might this be? Possible reasons are: the sample size may not be large enough, the individuals in the study may not be representative of the whole population or the general population mean may be different from 100.

This video [1] provides an explanation of how to interpret the results of a 95% confidence interval and the results from `R` for a hypothesis test. Examples are provided for one and two sample intervals and tests:

Video

Interpreting confidence intervals and hypothesis test results

Duration 12:16



Interpreting interval estimates and test results

Claire Miller, Eilidh Jack and Suzy Whoriskey

Assumptions

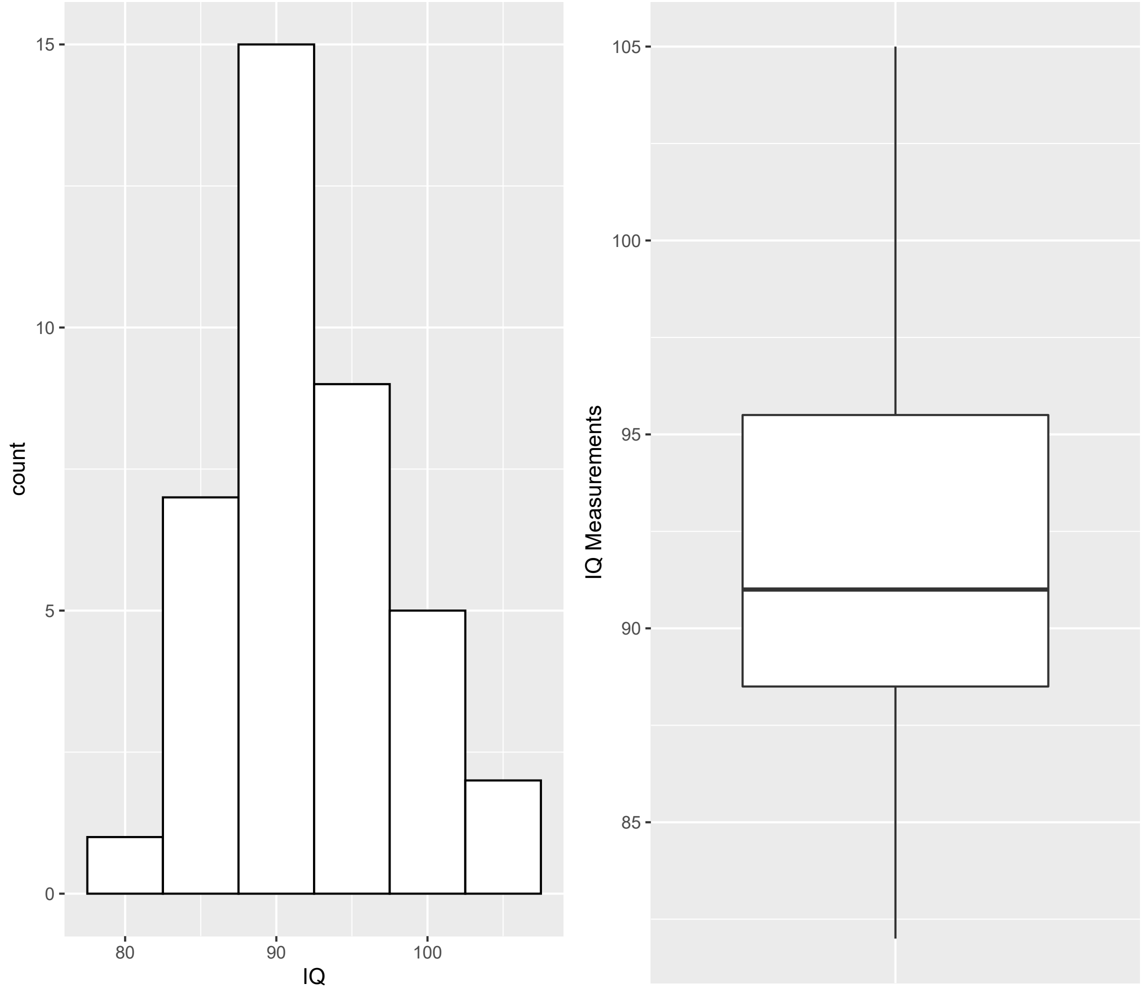
For the one-sample t-interval and test we have assumed that:

- our data $x_1 \ldots, x_n$ have arisen from a normal distribution;
- our data points $x_1, \ldots, x_n$ are independent of one another.

We should check to see if we think these assumptions are reasonable.

A histogram and boxplot for the IQ data are displayed below:

```
IQ <- data.frame(iq=c(94,  89,  83, 99, 94, 90, 94,  87,  89, 92, 92,
94,  91,  86, 90, 91, 82, 86,  88,  97, 96, 95,
87, 103,  91, 87, 91, 89, 98, 102, 105, 95, 88,
92,  90, 101, 98, 96, 83))

hist<-ggplot(IQ, aes(x=iq))+
  geom_histogram(binwidth=5, fill="white", color="black")+
  xlab("IQ")

boxplot<-ggplot(IQ, aes(x="", y=iq))+
  geom_boxplot()+
  ylab("IQ Measurements")+
  theme(axis.title.x=element_blank(),axis.text.x = element_blank(),
                                axis.ticks.x = element_blank())
grid.arrange(hist, boxplot, ncol=2)
```



Since the histogram is reasonably bell-shaped the normality assumption appears reasonable. The boxplot also supports that the distribution is reasonably symmetric but is not enough on its own to assess normality.

Additionally, the IQ measurements are for a random sample of 39 individuals and there is no reason to suggest that the IQ measurement of one individual should be related to another individual.

Task 1

A random sample of 20 houses in a postcode area in Scotland had their value advised by a surveyor. The data are in £1000's and are stored in the vector below:

price <-c(120, 110, 108, 100, 150, 106, 100, 100, 114, 130, 122, 100, 120, 130, 115, 112, 126, 110, 120, 128)

Tasks:

1. Produce a histogram of the data and comment on the assumptions for a one-sample t-test.
2. Produce a confidence interval for the population mean house price by hand.
3. Use a one-sample t-test in `R` to investigate the hypothesis that the population mean house price is £118,000.

Note: The statistical table for the quantiles of the t-distribution can be used to select the correct critical value from the table. The pivotal functions sheet can be used to help formulate the confidence interval. Both of these are available on Moodle.

Notes:

Paired t-test

Recall, that when we have paired data one thing to consider is whether or not it is natural to reduce it to one sample of data by taking differences. When we have more than one measurement from the same *object* or *individual* these *repeated measurements* are not independent, and hence the second assumption of independent observations for our tests is not appropriate. If it is sensible to take differences for each pair of measurements then the problem is reduced to one sample of differences arising from a population of differences. In this case, a one-sample t-interval and t-test can be computed using the sample of differences, which are assumed to be independent of one another.

One-sided tests

One-sided tests have significance levels of α = 0.025 (or α = 0.005). These significance levels are exactly half those of the corresponding two-sided tests (e.g. for α = 0.05 or 0.01), since a test with a one-sided alternative hypothesis has a p-value that is exactly half the p-value of the corresponding test with a two-sided alternative hypothesis.

Non-parametric tests

If the normality assumption does not seem reasonable, then you might want to consider a nonparametric test. Nonparametric tests make fewer assumptions and, for example, the **Wilcoxon Signed Ranks** test could be used in preference to a procedure based on the t distribution, if the assumption of normality is dubious. The Wilcoxon signed ranks test performs inference on the population median. In practice, the t-test performs well even when the assumption of normality is dubious and hence if you apply both a t-test and the Wilcoxon signed ranks test you will usually find that the conclusions are very similar.

Supplement 3

Supplementary material with outline information on basic nonparametric tests is available here: Nonparametric

Two-sample t-interval and t-test

Now let's extend this to the situation where we have data that have arisen from two populations and we are interested in comparing the population means.

Denote the sample values for sample 1 by $X_1, X_2, \ldots, X_m$ and for sample 2 by $Y_1, Y_2, \ldots, Y_n$, with sample means $\bar{X}$ and $\bar{Y}$ and sample standard deviations s_1 and s_2. Let μ_1 and μ_2 denote the corresponding population means and let σ_1 and σ_2 denote the population standard deviations.

We make the following assumptions:

- All recorded values are independent observations from their respective populations.
- The distribution of the variable of interest is the same in both populations, except, possibly, for a difference in the population means. In particular, this requires that the population standard deviations are equal, i.e. $\sigma_1 = \sigma_2 (= \sigma_T$, say)
- The distribution of the variable of interest is normal in both the populations.

With the above assumptions, a pivotal function for the difference in population means, $\mu_1 - \mu_2$, is derived as follows.

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_T^2}{m}\right)$$

independently of

$$\bar{Y} \sim N\left(\mu_2, \frac{\sigma_T^2}{n}\right).$$

It can then be shown using probability theory that:

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \left(\frac{1}{m} + \frac{1}{n}\right)\sigma_T^2\right)$$

i.e.

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\sigma_T^2}} \sim N(0, 1).$$

Since we typically do not know the true value σ_T^2, we obtain the pivotal function:

$$PIV(\mu_1-\mu_2; X, Y) = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\left(\frac{1}{m}+\frac{1}{n}\right)\hat{\sigma}^2}} \sim t(m+n-2).$$

The second assumption suggests that we could obtain a better estimate of the common standard deviation, $\hat{\sigma}$, by pooling the information from the two samples. The resulting, pooled estimate of $\hat{\sigma}$ is usually denoted s_p, where:

$$s_p^2 = \frac{(m-1)s_1^2+(n-1)s_2^2}{m+n-2} = \frac{\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{j=1}^{n}(Y_j-\bar{Y})^2}{m+n-2}.$$

Supplement 4

See Chapter 3 of the e-book: Foundations of Applied Statistical Methods (linked to on the Moodle page) for more information on the pivotal function for a two-sample t-interval.

It follows from the results above that a **symmetric 95% confidence interval for a difference in population means, $\mu_1-\mu_2$,** is given by:

$$(\bar{X}-\bar{Y}) \pm t_{0.975}(m+n-2)\sqrt{s_p^2\left(\frac{1}{m}+\frac{1}{n}\right)}$$

If this interval contains the value 0, then it is plausible that $\mu_1-\mu_2=0$, i.e. it is plausible that $\mu_1=\mu_2$. We may then conclude that there is insufficient evidence that the two population means are different. If the confidence interval does not contain the value 0, then it is not plausible that $\mu_1=\mu_2$; and hence there is evidence that the two population means are different.

Note: In $t(m+n-2)$ the degrees of freedom $m+n-2$ is determined by the total sample size $m+n$ and then subtracting the number of parameters we are estimating, which in this case is 2 for $\hat{\mu_1}$ and $\hat{\mu_2}$.

Example 3

Preferred Room Temperatures

In a controlled environment laboratory, 10 men and 10 women were tested individually to determine the room temperature (^{0}F) they found to be most comfortable. The following results were obtained:

```
Women (X)          75   77   78   79   77   73   78   79   78   80

Men (Y)            74   72   77   76   76   73   75   73   74   75
```

We will now calculate a 95% confidence interval for the difference in the two population means i.e. is there a difference in the mean preferred temperature for males and females?

Let μ_1, μ_2 denote the population mean preferred temperatures for women and men respectively.

A 95% confidence interval for the difference in population means $\mu_1 - \mu_2$:

$$\bar{x} = \sum_{i=1}^{10} \frac{x_i}{10} = 77.4, \quad \bar{y} = \sum_{i=1}^{10} \frac{y_i}{10} = 74.5$$

$$s_p^2 = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2 + \sum_{j=1}^{10}(y_j - \bar{y})^2}{10 + 10 - 2} = 3.3833$$

^Not sure what that number is
Follow Lee (SEE NOTES)

$$(\bar{x} - \bar{y}) \pm t_{0.975}(m + n - 2)\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}$$

$$(77.4 - 74.5) \pm t_{0.975}(18)\sqrt{3.3833\left(\frac{1}{10} + \frac{1}{10}\right)}$$

meanX - meanY/
sqrt((sd(X)/sqrt(n)))^2 + (sd(Y)/sqrt(n)))^2)

^THIS GIVES DIFFERENCE IN STANDARD ERROR
in this case, .8226

$$2.9 \pm 2.101 \times 0.8226$$

$$2.9 \pm 1.728$$

$$(1.17, 4.63).$$

A two-sample t-test can be performed in `R` using the command `t.test()`. The results from `R` are displayed below. In this case the hypotheses are:

$\text{H}_0 : \mu_1 = \mu_2$

$\text{H}_1 : \mu_1 \neq \mu_2$

Use the following command to perform the two-sample t-test. This assumes that the population variances for each population are equal.

```
Temps <- c(74, 72, 77, 76, 76, 73, 75, 73, 74, 75, 75, 77, 78, 79, 77,
73, 78, 79, 78, 80)
group <- c(rep("M",10), rep("F",10))

t.test(Temps~group, var.equal=TRUE)
```

```
R Console
Two Sample t-test


data:  Temps by group
t = 3.5254, df = 18, p-value = 0.002416
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
1.171787 4.628213
sample estimates:
mean in group F mean in group M
77.4            74.5
```

Conclusion

Since the confidence interval (1.17, 4.63) does not contain zero and the p-value is < 0.05 at 0.002, there is a statistically significant difference between the mean preferred room temperatures of males and females. The mean difference is highly likely to lie between 1.17 and 4.63 ^oF with females highly likely to prefer higher temperatures.

Nonparametric test

If the assumption of normality is dubious then the **Mann-Whitney test** is a nonparametric test to compare the population medians of two populations. See the supplementary material at Nonparametric for more information.

Task 2

These data are the starting salaries of 50 Scottish adults who are administrative assistants; 25 of whom work in the public sector and 25 work in the private sector. The Salary vector in `R` contains all of the data in £1,000's with group identifiers provided in the vector Sector for public and for private.

```
Salary <- c(17.7, 17.2, 20.2, 34.0, 36.4, 11.3, 24.0, 17.6, 26.0, 25.7, 17.2, 14.1, 22.0, 17.2, 20.9,
16.8, 19.3, 15.8, 27.0, 20.4, 25.5, 30.1, 28.3, 29.5, 31.6, 18.9, 10.5, 17.5, 13.1, 13.0, 18.2, 22.0,
13.0, 25.0, 12.2, 10.3, 15.5, 24.4, 11.8, 15.0, 25.6, 11.8, 22.8, 19.4, 12.3, 22.7, 27.3, 16.0, 11.0,
12.6)

Sector<- c(rep("private",25), rep("public",25))
```

Tasks:

1. Produce a histogram for each of the two groups separately.
2. Use a two-sample t-test to investigate the hypothesis that the population mean salaries are different for the public and private sector.

Supplement 5

When it is of interest to investigate differences between 3 or more population means, single factor analysis of variance can be used. See Chapter 4 of the e-book: Foundations of Applied Statistical Methods (linked to on the Moodle page) for more information on one-way ANOVA. However, note that this might be more accessible after getting to grips with the ideas of regression modelling and tests in the **Predictive Modelling** course.

One and two proportion tests

Some of the questions considered in the examples in the initial weeks involved population proportions and quite clearly the best estimates of these are the corresponding sample proportions.

However, one has to realise that the sample proportions are just the best guess at the population equivalent and if a different sample of the same size was taken it is virtually certain that the sample proportion would not be the same as the first sample proportion and so on for subsequent possible samples (of the same size).

It is also essential to note that the larger the sample size the less variability one would see in such sample proportions.

Assumptions

Suppose $X \sim Bi(n, \theta)$. Then, X is the number of successes in n independent trials, where each trial has success probability θ. The sample proportion $\hat{\theta}$ is clearly the best estimator of the population proportion θ.

As you've seen in the course **Probability and Stochastic Models** or **Probability and Sampling Fundamentals/Sampling Fundamentals**, we can use a normal approximation to the binomial distribution and hence,

$$X \sim N(n\theta, n\theta(1-\theta)).$$

This result requires a discrete distribution to be approximated by a continuous one. This introduces some inconsistencies, since the probability that a normal random variable equals any particular value is 0. Using the so-called continuity correction improves the approximation (see the **Probability and Stochastic Models** or **Probability and Sampling Fundamentals/Sampling Fundamentals** course for revision here).

The central limit theorem tells us that, approximately, for sufficiently large values of n (sometimes defined to be $n >= 20$, $n\theta >= 5$ and $n(\theta(1-\theta)) >= 5$), the test statistic of interest is:

$$\frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} \sim N(0,1).$$

Therefore, taking $\hat{\theta} = X/n$, it follows that an approximate pivotal function for θ is

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{\theta}(1-\hat{\theta})/n}} \sim N(0,1)$$

and similarly, in the two sample case:

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\hat{\theta}_1(1-\hat{\theta}_1)/n_1 + \hat{\theta}_2(1-\hat{\theta}_2)/n_2}} \sim N(0,1).$$

The data for the following examples are contained in the Rdata object for week 4, which can be found at: RData, under the data object **blooddata**.

Example 4

Is the taste bitter?

Phenylthiocarbamide (PTC) is a chemical that tastes bitter to some people ("tasters") and bland to everyone else ("non-tasters"), a response which is controlled by two forms of the TAS2R38 gene. A sample of individuals from across the world had their tasting status determined.

Initially, we are interested in answering the following question:

- What percentage of living humans are tasters?

The dataset contains the tasting status in column "Taster" ("Yes" = taster; "No" = non-taster) and the geographical information in column "Hemisphere" ("N" = northern; "S" = southern).

The number of tasters (X) in the sample = 156, with a sample size of $n = 213$, and so an estimate for the population proportion of tasters $\hat{\theta} = 156/213 = 0.732$.

We will construct a 95% confidence interval for the population proportion of tasters:

$$\hat{\theta} \pm N\left(0, 1; 1 - \frac{(1-0.95)}{2}\right)\sqrt{\frac{(\hat{\theta}(1-\hat{\theta}))}{n}}$$

$$\hat{\theta} \pm 1.96\sqrt{\frac{(\hat{\theta}(1-\hat{\theta}))}{n}}$$

$$0.732 \pm 1.96\sqrt{\frac{(0.732(1-0.732))}{213}}$$

$$0.732 \pm 1.96 \times 0.03035$$

$$(0.6725, 0.7915).$$

The 97.5th percentile of the $\mathrm{N}(0, 1)$ distribution is 1.96.

Conclusion

It is highly likely that between 67% and 79% of the population are tasters of PTC, with a point estimate of 73%.

Hypothesis test in `R`

Initially, summarising the data:

```
table(blooddata$Taster)
```

R Console

```
No Yes
57 156
```

```
prop.table(table(blooddata$Taster))
```

R Console

```
No        Yes
0.2676056 0.7323944
```

In the sample of data, 73.2% can taste PTC.

The command `prop.test` in `R` can be used to produce this confidence interval and to perform hypothesis tests for one proportion.

```
number <- table(blooddata$Taster)
number
```

R Console

```
No Yes
57 156
```

```
prop.test(number)
```

R Console

```
1-sample proportions test with continuity correction

data:  number, null probability 0.5
X-squared = 45.089, df = 1, p-value = 1.883e-11
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2105102 0.3332564
sample estimates:
p
0.2676056
```

The results from the proportion test in `R` have taken the first entry in the object `number` and hence refer to a 95% confidence interval for those that are not tasters. A 95% confidence interval for those that are tasters can be found by (1-lower limit for non-tasters, 1-upper limit for non-tasters) of the interval.

Example 5

Is the taste bitter? cont.....

In addition to the earlier question in example 4, i.e. `what percentage of living humans are tasters?', we are also interested in extending this to consider:

Does the percentage differ between the northern and the southern hemisphere?

The dataset contains the tasting status in column "Taster" ("Yes" = taster; "No" = non-taster) and the geographical information in column "Hemisphere" ("N" = northern; "S" = southern).

A two proportion confidence interval and hypothesis test can also be performed in `R` using the `prop.test` command.

```
number <- table(blooddata$Hemisphere, blooddata$Taster)
prop.table(table(blooddata$Hemisphere, blooddata$Taster), margin=1)
```

R Console

```
No        Yes
N 0.2873563 0.7126437
S 0.1794872 0.8205128
```

```
prop.test(number)
```

R Console

```
2-sample test for equality of proportions with continuity correction

data:  number
X-squared = 1.3811, df = 1, p-value = 0.2399
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.04576387  0.26150215
sample estimates:
prop 1    prop 2
0.2873563 0.1794872
```

This provides a 95% confidence interval for the difference between the proportion of the people that are not tasters for the northern and southern hemispheres. Since the p-value is > 0.05 and the confidence interval includes 0 then there is insufficient evidence of a difference in the proportion of non-tasters between the populations of the two hemispheres.

Task 3

In the previous example for 'Is the taste bitter?' the two proportion interval was given in terms of a difference between the northern and southern hemispheres for non-tasters. State the 95% confidence interval for the difference in the population proportion of **tasters** between the northern and southern hemispheres.

Supplement 6

In the situation where we have more than 2 categories that are of interest then we have what we refer to as categorical data and generally assume a multinomial distribution (where the binomial distribution and two categories is a special case). We'll revisit categorical data later in the course once we've introduced the ideas of likelihood. However, at this point it's useful to mention two other classical statistical tests that are popular:

- a test of association and,
- a test to compare multinomial populations

Test of association

This test is useful if we are interested in investigating if two categorical variables have a relationship i.e. are associated. For example, are people with blond hair more likely to have blue eyes than people with black hair?

Test to compare multinomial populations

This test is useful if we are interested in investigating if the distribution of frequencies for different levels of a categorical variable are different between two populations. For example, are the percentages of people that will vote for the Labour, Conservative and Liberal Democrats parties at the next election different in Scotland compared to England?

For more information on tests for categorial data see Chapter 13 of the e-book: Statistical inference: a short course (linked to on the Moodle page) for more information.

Learning outcomes for week 4

By the end of week 4, you should be able to:

- use pivotal functions to obtain confidence intervals within normal models;
- compute one and two-sample t-intervals and interpret the results;
- interpret `R` output from one and two-sample t-tests;
- state the assumptions for one and two-sample t-tests and one and two-sample proportion tests;
- interpret `R` output for one and two proportion tests.

Review exercises, selected video solutions and written answers to all tasks/review exercises are provided overleaf.

Review exercises

Task 4

The same individuals that had their Phenylthiocarbamide (PTC) tasting status recorded, from examples 4 & 5 and task 3, also had their blood group determined in the experiment. This was classified as Blood Group O (`Yes') or not, for males (M) and females (F). There were two questions of interest;

(a) What is the percentage of living humans with Blood Group O?

(b) Does the percentage differ between males and females?

Below you'll find the `R` output from the investigation to answer these questions of interest. Comment and interpret.

```
# Raw counts by Blood Group
table(blooddata$BloodgroupO)
```

R Console

```
No Yes
136  77
```

```
# Transform the raw counts by Blood Group into proportions
prop.table(table(blooddata$BloodgroupO))
```

R Console

```
No        Yes
0.6384977 0.3615023
```

```
rawcounts <- table(blooddata$BloodgroupO)
prop.test(77,213)
```

R Console

```
1-sample proportions test with continuity correction

data:  77 out of 213, null probability 0.5
X-squared = 15.793, df = 1, p-value = 7.065e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.2977464 0.4303422
sample estimates:
p
0.3615023
```

```
# Raw counts for Blood Group by sex
table(blooddata$Sex,blooddata$BloodgroupO)
```

R Console

```
No Yes
F 63  33
M 73  44
```

```
# Calculate proportions by columns
prop.table(table(blooddata$BloodgroupO,blooddata$Sex),
margin=2)
```

R Console

```
F         M
No  0.6562500 0.6239316
Yes 0.3437500 0.3760684
```

```
prop.test(c(33,44), c(96,117))
```

R Console

```
2-sample test for equality of proportions with continuity
correction

data:  c(33, 44) out of c(96, 117)
X-squared = 0.11914, df = 1, p-value = 0.73
```

```
alternative hypothesis: two.sided
95 percent confidence interval:
-0.1711479  0.1065112
sample estimates:
prop 1    prop 2
0.3437500 0.3760684
```

Task 5

As part of a study of health and growth in school children in Glasgow, 43 ten year old boys had their triceps skinfold thickness measured. All measurements were made by the same measurer using standard calipers, with summaries of:

$$n = 43, \quad \sum x_i = 346.2 \quad \sum x_i^2 = 3066.94.$$

A large scale study in the South East of England has established that the mean triceps skinfold thickness of ten year old boys in that area is 9.3mm. Is there any evidence that the mean triceps skinfold thickness in Glasgow is different from that in South East England? (Hint: produce a 95% confidence interval for the population mean.)

Task 6

For this task there is a dataset containing times taken for 70 individuals to read words of colours. This is a paired dataset as each individual has two measurements: one reading the words in black text, and the other reading the words in different coloured text (different to the colour of the word). It is of interest to investigate whether or not there is a time difference in reading these words in coloured or black text. Read in the dataset and investigate, forming some informal conclusions from your plots and summaries and then carrying out a formal test. You can find the dataset at: RData in **Week 4 - review exercises(RData file)** in the object **colours**: you have subject number (Subj), time reading colour text (DiffCol) and time reading black text (Black). (Hint: consider a one-sample t-test here.)

Task 7

A new scoring system for the severity of illness for patients entering an Intensive Care Unit (ICU) had been adopted by researchers in Glasgow. This sickness score gets progressively higher the worse the condition of the patient. As part of a recent study of the usefulness of such a score in predicting whether patients entering the ICU will recover successfully (R) or not (NR), a random sample of 40 patients had their sickness score measured on admission to the ICU and their final status (i.e. recover or not) recorded. The resulting data were as follows:

```
#Score for patients who recovered successfully
ICU$R
```

R Console

```
[1] 14 10 13 10 17 18 12 19 17 10 10 15  4  9 13 14 12 12  9 18
```

```
#Score for patients who did not recover successfully
ICU$NR
```

R Console

```
[1] 14 26 22 20 20 23 20 23 18  8 13  9 26 25 30 26 22 25 24 20
```

Where the scores for patients who recovered successfully are denoted as x_1 and patients who did not recover successfully as x_2. Some summary statistics are:

Group	i	n_i	$\bar{x}_i$	s_i^2
R	1	20	12.8	14.48
NR	2	20	20.7	33.91

Assume both underlying distributions can be adequately represented by normal distributions. Is there any evidence of a significant difference between the population mean (μ_i) sickness score for the two groups? (Hint: produce a 95% confidence interval for $\mu_1 - \mu_2$. You could also carry out a t-test in `R`, data are in the object **ICU** in the `R` data file, RData in **Week 4 - review exercises(RData file)**).

Task 8

As part of a physiology practical session, the mean haemoglobin (gdl^{-1}) levels of 14 male and 11 female medical students were determined.

```
## Male Summary
summary(Physiology$Haemo[Physiology$Sex=="M"])
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.80   14.03   14.70   14.56   14.90   16.10
```

```
## Female Summary
summary(Physiology$Haemo[Physiology$Sex=="F"])
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
11.20   12.65   14.40   13.59   14.75   15.20
```

(a) Display the data and comment appropriately (Data are in the `R` object **Physiology** in the `R` Data file RData in **Week 4 - review exercises(RData file)**).

(b) A two-sample t-test comparing the population mean (μ) haemoglobin for males (m) and females (f) yielded the result $0.05 < p < 0.10$. Provide clear interpretation of this result. (There is no need for any further calculations.)

(c) A 95% confidence interval for $\mu_m - \mu_f$ is (-0.06, 2.00) gdl^{-1}. Provide clear interpretation of this result and compare with part (b).

Answer 1

Price data

```
Price <-data.frame(price=c(120, 110, 108, 100, 150, 106, 100,
100, 114, 130, 122, 100, 120,
          130, 115, 112, 126, 110, 120, 128))
library(ggplot2)
library(gridExtra)

hist<-ggplot(Price, aes(x=price))+
  geom_histogram(binwidth=5, fill="white", color="black")+
  xlab("1,000s GBP")

boxplot<-ggplot(Price, aes(x="", y=price))+
  geom_boxplot()+
  ylab("1,000s GBP")+
  theme(axis.title.x=element_blank(),axis.text.x =
element_blank(),
        axis.ticks.x = element_blank())

grid.arrange(hist, boxplot, ncol=2)
```

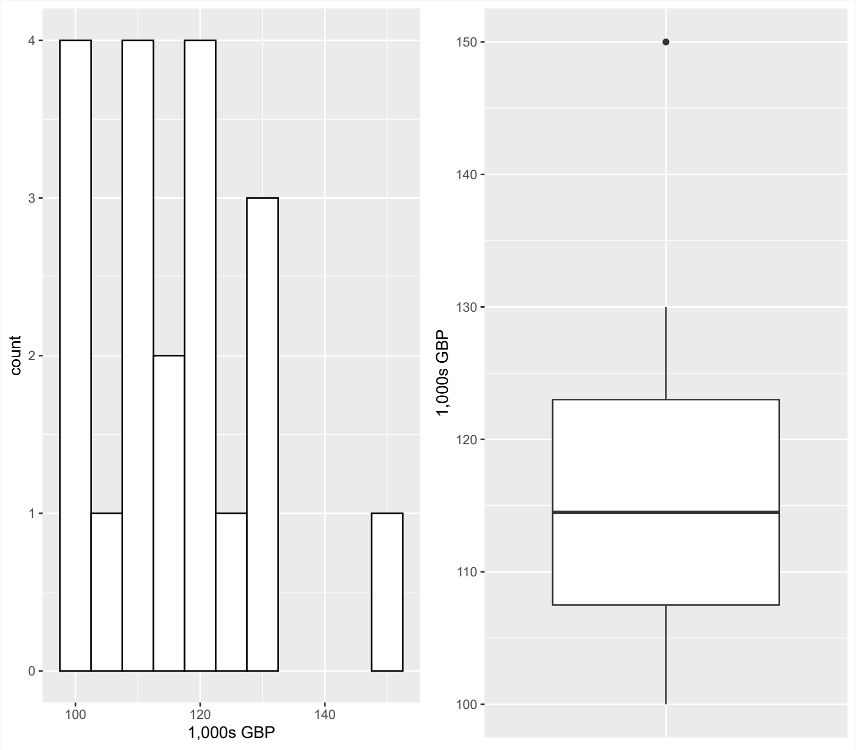


Figure 1

The boxplot looks reasonably symmetric with one outlier. However, the histogram looks to be skewed to the right. The histogram does not appear to follow a bell-shaped curve and hence the assumption of a normal distribution might be a little dubious here.

A 95% confidence interval for the population mean house price μ_T (when σ_T is unknown) is:

$$\left(\bar{x} \pm t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

$$\left(116.05 - t_{0.975}(20-1)\frac{\sqrt{166.16}}{\sqrt{20}}, 116.05 + t_{0.975}(19-1)\frac{\sqrt{166.16}}{\sqrt{20}}\right)$$

$$t_{0.975}(19) = 2.093$$

$$\left(116.05 - 2.093\frac{\sqrt{166.16}}{\sqrt{20}}, 116.05 + 2.093\frac{\sqrt{166.16}}{\sqrt{20}}\right)$$

$$(116.05 - 2.093 \times (2.882), 116.05 + 2.093 \times (2.882))$$

$$(110.0, 122.1)\,.$$

Conclusion

It can therefore be concluded that the population mean house price is highly likely to lie in the range £110,000 to £122,100, with a point estimate for $\hat{\mu}$ of £116,050. However, we should be cautious around this interpretation since the assumption of normality appeared a little dubious. A larger sample size or nonparametric test may be required to check the conclusion of this result.

```
t.test(Price$price, mu=118)
```

R Console

```
One Sample t-test


data:  Price$price
t = -0.67654, df = 19, p-value = 0.5069
alternative hypothesis: true mean is not equal to 118
95 percent confidence interval:
110.0172 122.0828
```

```
sample estimates:
mean of x
116.05
```

Since the p-value is greater than 0.05, we would not reject the null hypothesis and conclude that there is insufficient evidence that the population mean house price is not equal to £118,000. This is also suggested by the confidence interval since the value 118 is contained within the interval.

Answer 2

Salary data

```
Salary <- c(17.7, 17.2, 20.2, 34.0, 36.4, 11.3, 24.0, 17.6,
26.0, 25.7, 17.2, 14.1, 22.0, 17.2,
            20.9, 16.8, 19.3, 15.8, 27.0, 20.4, 25.5, 30.1,
28.3, 29.5, 31.6, 18.9, 10.5, 17.5,
            13.1, 13.0, 18.2, 22.0, 13.0, 25.0, 12.2, 10.3,
15.5, 24.4, 11.8, 15.0, 25.6, 11.8,
            22.8, 19.4, 12.3, 22.7, 27.3, 16.0, 11.0, 12.6)

Sector<- c(rep("private",25), rep("public",25))

Salary.Sec<-data.frame(Salary = Salary, Sector=Sector)

# New facet label names for supp variable
sec.labs <- c(private="Private sector", public="Public sector")

ggplot(Salary.Sec, aes(x=Salary))+
  geom_histogram(binwidth=3, fill="white", color="black")+
  facet_grid(~Sector,labeller = labeller(Sector = sec.labs))+
  xlab("Salary")
```

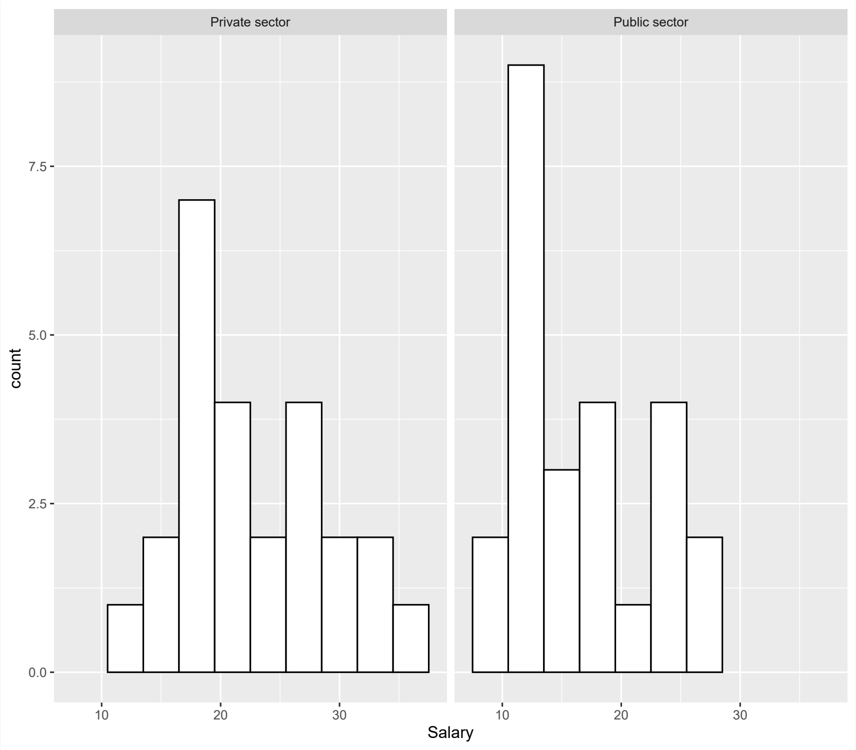


Figure 2

```
ggplot(Salary.Sec, aes(x=Sector, y=Salary))+
  geom_boxplot()+
  ylab("Job group")+
  xlab("Salary 1,000s GBP")
```

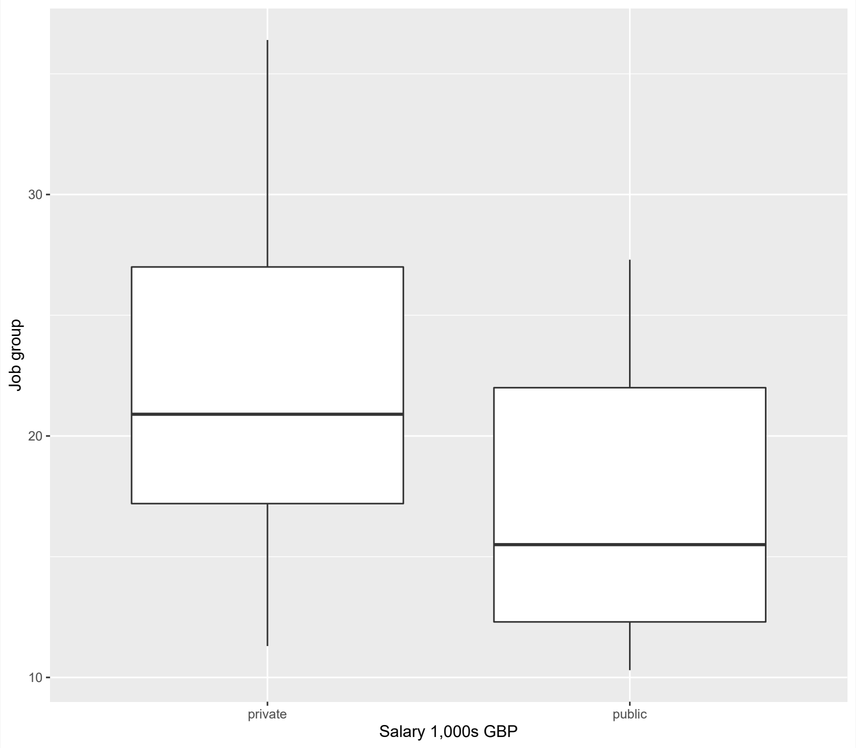


Figure 3

```
t.test(Salary~Sector, var.equal=TRUE)
```

R Console

```
Two Sample t-test

data:  Salary by Sector
t = 3.3933, df = 48, p-value = 0.001392
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
2.34536 9.16664
sample estimates:
mean in group private  mean in group public
22.632                16.876
```

The boxplot suggests that the two samples have a similar shape and spread. However, there is some evidence from the histogram that while the private sector distribution appears to follow a bell-shaped curve the assumption of normality is dubious for the public sector.

From the two-sample t-test results, the p-value is much less than 0.05 at 0.001 and hence we would reject the null hypothesis (which in this case is that there is no difference between the two population means) and conclude that there is a statistically significant difference between the two population means.

The 95% confidence interval provides us with further information. The confidence interval does not include zero, which again indicates a statistically significant difference. It is also entirely positive which tells us that the values for the private sector are higher than the values for the public sector. Therefore, there is evidence of a statistically significant difference between the population mean salaries for those in the private and public sector. The salaries in the private sector are highly likely to be higher by between £2,345 and £9,167.

A 95% confidence interval for the difference in population means $\mu_1 - \mu_2$:

$$\bar{x} = \sum_{i=1}^{25} \frac{x_i}{25} = 22.632, \quad \bar{y} = \sum_{i=1}^{25} \frac{y_i}{25} = 16.876$$

$$s_p^2 = \frac{\sum_{i=1}^{25}(x_i - \bar{x})^2 + \sum_{j=1}^{25}(y_j - \bar{y})^2}{25 + 25 - 2} = 35.968$$

$$(\bar{x} - \bar{y}) \pm t_{0.975}(m + n - 2)\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}$$

$$(22.632 - 16.876) \pm t_{0.975}(48)\sqrt{35.968\left(\frac{1}{25} + \frac{1}{25}\right)}$$

$$5.756 \pm 2.0106 \times 1.696$$

$$5.756 \pm 3.40998$$

$$(2.346, 9.166).$$

Answer 3

In order to find the 95% confidence interval for the difference in the population proportion of **tasters** between the northern and southern hemispheres, we would simply reverse the confidence interval produced for the difference for non-tasters. Therefore, the 95% confidence interval for the difference in the proportion of tasters between the northern and southern hemisphere is

$$(-0.262, 0.046)$$

i.e. there is insufficient evidence of a difference between the population proportions and hence between the proportion of tasters in the population of the northern and southern hemisphere.

Answer 4

Blood group data:

(a) Of the 213 subjects, 77 have Blood Group O, which is 36.2%. So we expect about $\frac{1}{3}$ of individuals to have Blood Group O, but need to test this formally using a proportions test.

The `R` output given in the question is a 1-sample proportions test, where here we are testing the hypotheses; $H_0 : p = 0.5 \text{ v.s. } H_1 \neq 0.5$, where p is the population proportion of people with Blood Group O. In the `R` output, the p-value is very small, 7.065e-05 (< 0.001), and is therefore much less than $\alpha = 0.05$, our significance level.

In this case, we reject the null hypothesis and conclude that there is evidence that $p \neq 0.5$. In other words, the proportion of living humans with Blood Group O is highly likely not to equal 50%.

For a more informative conclusion we look to the 95% Confidence Interval in the ouput. This interval is (0.29, 0.43). So, worldwide percentage of humans living with Blood Group O is highly likely to fall between 30% and 43%.

(b) We gain little knowledge from looking at the raw counts table in the `R` output for Blood Group by sex, the proportions table is more informative. From the table, 34% of the female sample have Blood Group O and 38% of the male sample have Blood Group O, a difference of 4% with slightly more in the male sample. However, to see whether this is meaningful, we have to test this formally using a two-sample t-test for the equality of proportions.

In the `R` output, this test shows the results from testing the hypotheses;

$H_0 : p_f = p_m \text{ v.s. } H_1 : p_f \neq p_m$, where p_f and p_m are the population proportions of females and males with Blood Group O, respectively.

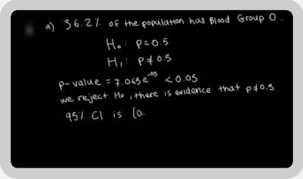
The p-value is 0.73 which is $> \alpha = 0.05$ and so we cannot reject the null hypothesis, that the population proportion of females with Blood Group O is equal to the population proportion of males with Blood Group O.

An approximate 95% confidence interval for the difference in humans with Blood Group O between the sexes is (-0.17, 0.11). This interval includes zero, and so we can conclude that there is not enough evidence to suggest that there is a difference between genders.

Video

Video model answers

Duration 4:40



a) 36.2% of the population has Blood Group O.

H_0: $p=0.5$

H_1: $p \neq 0.5$

p-value $= 7.065e^{-09} < 0.05$

we reject H_0, there is evidence that $p \neq 0.5$

95% CI is (0.

Answer 5

Mean tricep skinfold thickness:

To calculate a 95% Confidence Interval for the population mean tricep skinfold thickness in Glasgow, we use the following formula;

$$(\bar{x} - t_{0.975}(n-1)\frac{s}{\sqrt{n}} \ , \ \bar{x} + t_{0.975}(n-1)\frac{s}{\sqrt{n}})$$

Using the measurements given in the question, the interval is calculated as shown.

$$n = 43 \ , \ \sum_i x_i = 346.2 \ , \ \sum_i x_i^2 = 3066.94$$

We can calculate $\bar{x} = \frac{\sum_i x_i}{n} = \frac{346.2}{43} = 8.05 \text{ rounded to 2.d.p}$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\left(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}\right) \\ &= \frac{1}{43-1}\left(3066.94 - \frac{(346.2)^2}{43}\right) \\ &= \frac{1}{42}(3066.94 - 2787.31) \\ &= \frac{1}{42} \times 279.627 \\ &= 6.657796235 \\ &= 6.66(2.\text{d.p}) \end{aligned}$$

Next, we need to find s^2. This is the formula from Task 2 in your Week 2 Learning Material.

Therefore, $s = \sqrt{s^2} = \sqrt{6.66} = 2.58$.

So, our 95% Confidence Interval for the population mean tricep skinfold thickness in Glasgow is,

$$\left(\bar{x} \pm t_{0.975}(n-1)\frac{s}{\sqrt{n}}\right)$$

i.e.

$$\left(8.05 - t_{0.975}(43-1)\frac{2.58}{\sqrt{43}}\ ,\ 8.05 + t_{0.975}(43-1)\frac{2.58}{\sqrt{43}}\right)$$

i.e.

$$(8.05 - (2.0181 \times 0.393)\ ,\ 8.050 + (2.0181 \times 0.393))$$

i.e.

$$(8.05 - 0.794\ ,\ 8.05 + 0.794)$$

i.e.

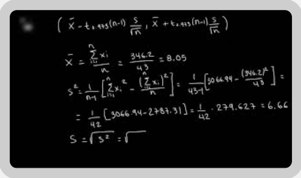
$$(7.256\ ,\ 8.844)$$

The 95% Confidence Interval is (7.256, 8.844) and so we interpret this as the population mean tricep skinfold thickness in Glasgow is highly likely to lie between 7.26mm and 8.84mm. This interval does not contain 9.3mm and so we can conclude that there is evidence that the Glasgow population mean is statistically different than that in South East England. The Glasgow population mean appears to be lower than in South East England

Video

Video model answers

Duration 5:43



Answer 6

We start this analysis by looking at the summary statistics of the data, including the mean, median and spread. It is also useful to visualise the data, a simple scatterplot is sufficient.

```
summary(colours$Black)
```

R Console

```
Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
29.85   38.20   40.97   41.00   44.45   52.21
```

```
summary(colours$DiffCol)
```

R Console

```
Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
29.13   38.76   44.33   43.30   46.96   56.55
```

```
sd(colours$Black)
```

R Console

```
[1] 4.840233
```

```
sd(colours$DiffCol)
```

R Console

```
[1] 6.14971
```

```
ggplot(colours, aes(x=DiffCol, y=Black))+
  geom_point()+
  geom_abline(intercept = 0, slope = 1)+
  xlab("Time reading in colour")+
  ylab("Time reading in black")
```

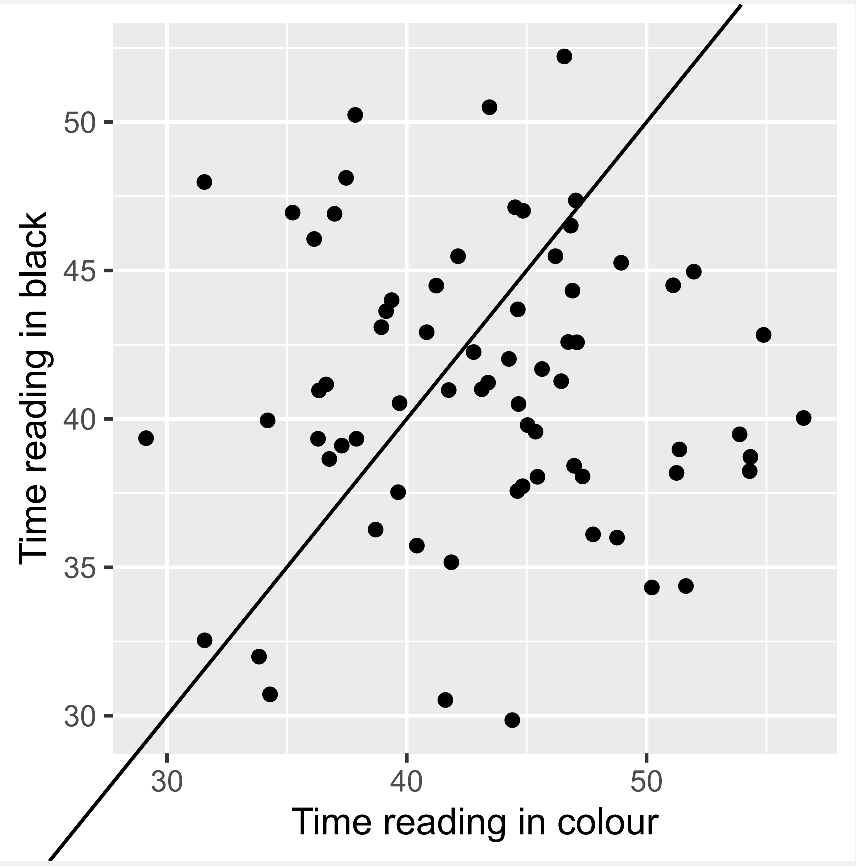


Figure 4

From the plot of Different Colours plotted against Black text, there doesn't appear to be a huge difference between the two measurements. Most of the points are equally scattered at either side of the line of equality. This is further emphasised when we look at the means and standard deviations of both groups. The mean time for Black text is 41.0 seconds and 43.3 seconds for Coloured text. The standard deviations for each group are 4.84 for Black text and 6.15 for Coloured text. Again, this indicates there is only a small difference between the two groups since the spreads over lap.

The line of equality has been superimposed on the plot and a very slight difference in the number of points on either side of this line can be seen. It could be that there is a longer time taken for reading in coloured text, however there is not sufficient evidence from our informal analysis to support this. More formal analysis must be carried out on this paired data problem.

Firstly, we create a new variable time (time for black words minus time for coloured words) within the dataset - the difference in time taken to read. To test our hypotheses, a one-sample t-test will be carried out. We are testing the following hypotheses,
$H_0 : \mu_d = 0 \text{ v.s. } H_1 : \mu_d \neq 0.$
The results of the one-sample t-test are shown below, where μ_d represents the population mean difference in time reading black text minus time reading coloured text.

R Console

```
One Sample t-test


data:  time
t = -2.4639, df = 69, p-value = 0.01624
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-4.1622755 -0.4377245
sample estimates:
mean of x
-2.3
```

The results of the t-test give us a p-value of 0.01624. This value is $< \alpha = 0.05$ and so we reject the null hypothesis, and conclude that the population mean difference is not equal to zero. For further information, we look to our 95% Confidence Interval. The interval does not contain zero, reiterating the results from the p-value, and so there is evidence that the population mean difference is not zero. The confidence interval is (-4.2, -0.4), which is wholly negative. We interpret this as; it is highly likely that the mean difference between reading the black text and the coloured text lies between -4.2 and -0.4 seconds. Since this interval is wholly negative, it indicates that reading in coloured text increases the time taken to read by between 0.4 and 4.2 seconds.

Answer 7

For this question, we are interested in whether or not there is a difference between the sickness score for each group in ICU, recovered successfully and not recovered successfully.

Let's denote group $\mathrm{i} = 1$ to be the group that recovered successfully (R) and $\mathrm{i} = 2$ the group that did not recover (NR). From the summary statistics given in the question, the mean sickness score for each group appears to be different, with the group that recovered having a lower sickness score of 12.8 compared to the group that did not recover with 20.7.

A 95% Confidence Interval for the difference in group means is given by,

$$(\bar{x} - \bar{y}) \pm t_{0.975}(m + n - 2)\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}$$

where m is the sample size of Group 1 (i.e. R) and n is the sample size of Group 2 (i.e. NR). The pooled estimate of $\hat{\sigma}$, denoted s_p is,

$$\begin{aligned} s_p^2 &= \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \\ &= \frac{((20-1) \times 14.48) + ((20-1) \times 33.91}{20+20-2} \\ &= \frac{275.12 + 644.29}{38} \end{aligned}$$

Using this result to calculate our 95% confidence interval gives;

$$(12.8 - 20.7) \pm t_{0.975}(20 + 20 - 2)\sqrt{24.195\left(\frac{1}{20} + \frac{1}{20}\right)}$$

$$(-7.9) \pm t_{0.975}(38) \times 1.555$$

$$(-7.9) \pm 2.0244 \times 1.555$$

$$(-7.9) \pm 3.147942$$

$$\text{i.e.} \, (-11.05\,, \, -4.80)$$

Interpreting our 95% Confidence Interval, it is highly likely that the difference in mean sickness score between the recovered successfully and the not recovered groups lies between -11.0 and -4.8. This interval does not contain zero indicating that there is in fact evidence of a statistically significant difference between the population mean sickness scores of each group. This interval is wholly negative, suggesting that the mean sickness score for the untreated (and not recovered) group is higher than that of the recovered group, which, from the question, indicates a worse condition of the patient.

Our hypotheses are; $H_0 : \mu_1 = \mu_2$ v.s. $H_1 : \mu_1 \neq \mu_2$.

The following `R` output shows the result of a two-sample t-test, assuming equal variances.

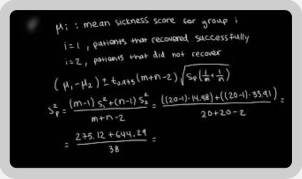
```
Two Sample t-test


data:  ICU$R and ICU$NR
t = -5.0789, df = 38, p-value = 1.037e-05
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
-11.048876  -4.751124
sample estimates:
mean of x mean of y
12.8      20.7
```

Here, the p-value is $< \alpha = 0.05$ and so we reject H_0 and conclude that there is a difference in population mean sickness score between the two groups. This agrees with our result from the 95% Confidence Interval.

Video

Video model answers

Duration 9:38



μ_i : mean sickness score for group i

i=1 , patients that recovered successfully

i=2 , patients that did not recover

$$(\mu_1-\mu_2) \pm t_{0.975}(m+n-2)\sqrt{S_p\left(\frac{1}{m}+\frac{1}{n}\right)}$$

$$S_p^2 = \frac{(m-1)S_1^2+(n-1)S_2^2}{m+n-2} = \frac{((20-1)\cdot 14.48)+((20-1)\cdot 33.91)}{20+20-2} =$$

$$= \frac{275.12+644.29}{38} =$$

Answer 8

Physiology task:

(a)

```
##Male Summary
summary(Physiology$Haemo[Physiology$Sex=="M"])
```

R Console

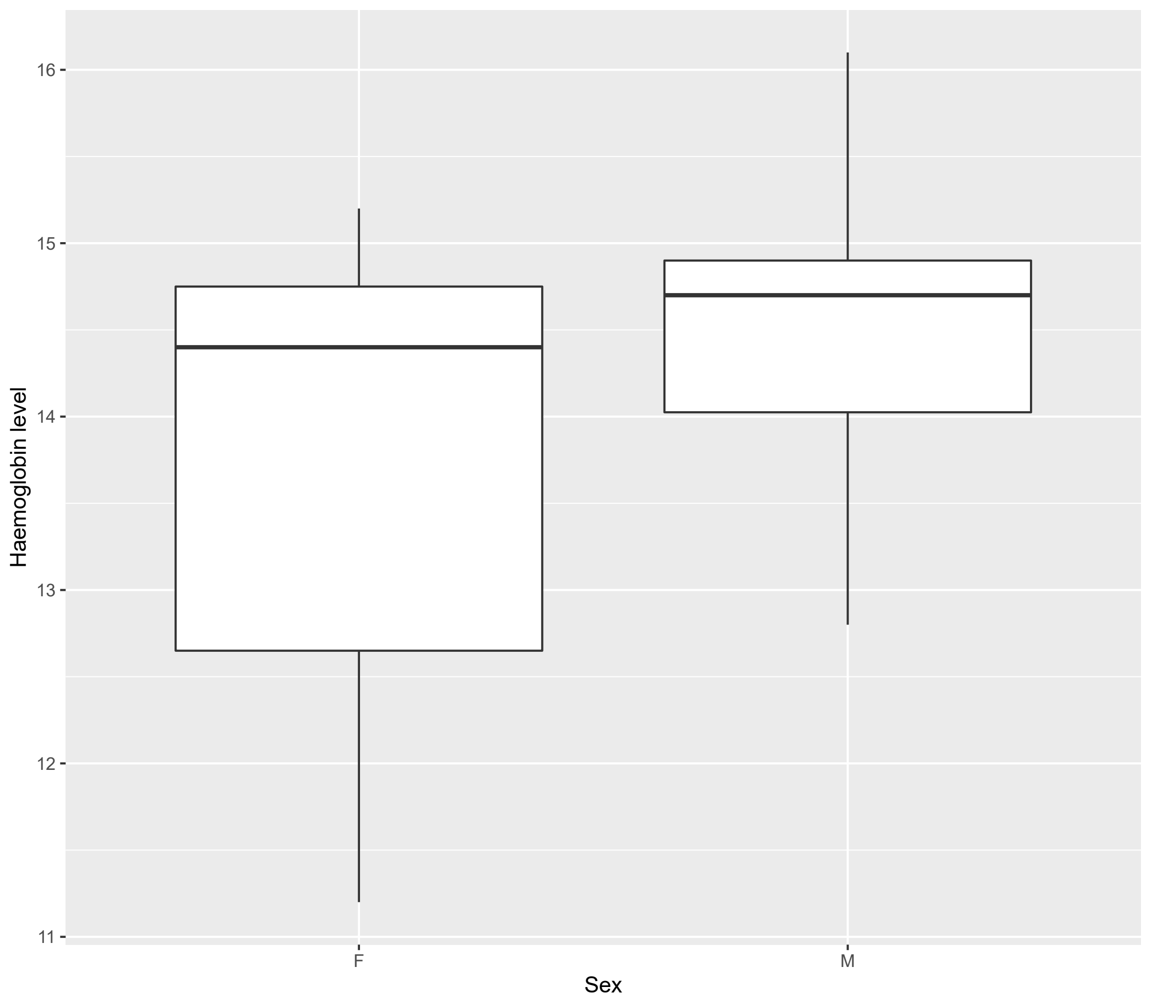
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
12.80   14.03   14.70   14.56   14.90   16.10
```

```
##Female Summary
summary(Physiology$Haemo[Physiology$Sex=="F"])
```

R Console

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
11.20   12.65   14.40   13.59   14.75   15.20
```

```
ggplot(Physiology, aes(x=Sex, y=Haemo))+
  geom_boxplot()+
  ylab("Haemoglobin level")+
  xlab("Sex")
```



The median haemoglobin levels from males (14.7) and females (14.4) are similar with males having a slightly higher haemoglobin level. The boxplots appear roughly symmetric (although this could be dubious), with females looking questionable with a higher median. Based on the sample data, it's possible that males will have slightly higher population mean haemoglobin levels. We could include a histogram of the data to check normality.

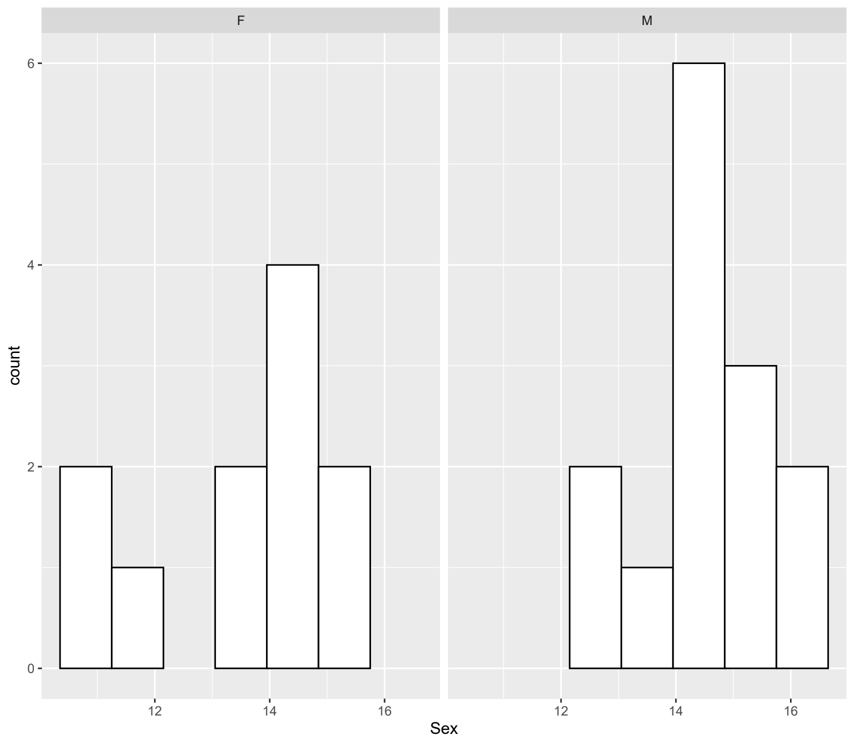


Figure 5

We can see that normality seems dubious from these histograms, as there is a lack of a bell-shaped curve.

(b) The fact that our p-value lies between 0.05 and 0.10, indicates that we should not reject the null hypothesis (H$_0$: μ_m = μ_f) and conclude that there is insufficient evidence of a difference between male and female haemoglobin levels. However, this p-value is bordering on a statistical difference (lying within this range so close to our significance level of 0.05) and so it may be worthwhile investigating further with a larger sample size.

(c) This 95% confidence interval contains zero and so there is insufficient evidence of a difference between population average haemoglobin levels of males and females. The interval just contains zero, so the difference is bordering on statistical significance (this agrees with part (b)). However, from the confidence interval, the range of plausible values is fairly wide (given the range of the data) and so this lack of precision of our estimates of (μ_m - μ_f) is clear and a larger sample is required for a more precise estimation.

Footnotes

1. Many thanks to Suzy Whoriskey for all her contributions to the development of the course material ↩