Chapter 13

Associations

13.1 Introduction

By the phrase "describing relationships" in statistics, we mean that we are interested in how two (or more) variables relate to each other. Two variables are **positively related** when an increase in one variable tends to accompany an increase in the other. They are **negatively related** when an increase in one variable tends to accompany a decrease in the other.

Most observational studies and experiments record observations on multiple variables (not just one). It therefore makes sense to think about how variables might be related. In what follows, we explore how to calculate different measures of correlation and how to make the appropriate choice for a variety of scenarios.

Statistical Inference: Does a relationship among a sample indicate a relationship in the population?

13.2 Describing Categorical Relationships

13.2.1 Fisher's Exact Test

Fisher's exact test is named after Ronald Fisher, often referred to as the father of modern-day statistics, who invented it to assess a Muriel Bristol's claim that she could tell whether a cup of tea was made by adding tea to milk or adding milk to tea simply from tasting it.

Example 13.1. Fisher designed an experiment to test her assertion in which eight cups of tea were made, four in each way. The eight cups of tea would then be presented to Muriel in random order. Muriel knows that half of the cups were made by adding tea to milk and half were made by adding milk to tea and so her guess will include four of each type.

In practice, we will simply report the number of observations in each cell of the table but here we've included how we refer to each cell which will be important for defining key values. The data is reported below in a 2×2 contingency table – 2×2 denotes 2 rows and 2 columns (not including the totals).

	Tea-First	Milk First	Total
Guessed Tea-First	O_{11} =3	O_{12} =1	R_1 =4
Guessed Milk First	O_{21} =1	O_{22} =3	R_2 =4
Total	C_1 =4	C_2 =4	n =8

We can create an object in to store this table in R as follows.

```
> O11<-3
> O12<-1
> O21<-1
> O22<-3
> tea.tab<-matrix(data=c(O11,O12,O21,O22),
+                 nrow = 2,
+                 ncol = 2,
+                 byrow = TRUE)
> colnames(tea.tab)<-c("Tea.First","Milk.First")
> rownames(tea.tab)<-c("Guessed.Tea.First","Guessed.Milk.First")
> tea.tab
Tea.First Milk.First
Guessed.Tea.First          3          1
Guessed.Milk.First         1          3
```

Definition 13.2. The **phi coefficient of correlation** is a measure of the the relationship between two binary variables. Below, we calculate this correlation between Muriel's guesses and the method with which the tea was made.

$$\phi = \frac{O_{11}O_{22} - O_{12}O_{21}}{\sqrt{R_1R_2C_1C_2}}$$

We can interpret values of ϕ as follows:

- -1.0 to -0.7 strong negative association
- -0.7 to -0.5 moderate negative association
- -0.5 to -0.3 weak negative association
- -0.3 to +0.3 little or no association
- +0.3 to +0.5 weak positive association
- +0.5 to +0.7 moderate positive association
- +0.7 to +1.0 strong positive association

For Fisher's experiment,

$$\begin{aligned}\phi &= \frac{O_{11}O_{22} - O_{12}O21}{\sqrt{R_1R_2C_1C_2}} \\ &= \frac{(3)(3) - (1)(1)}{\sqrt{(4)(4)(4)(4)}} \\ &= \frac{9-1}{\sqrt{256}} \\ &= \frac{8}{16} = 0.50.\end{aligned}$$

Since $\phi > 0$, Muriel's guesses are positively correlated the method with which the tea was made; this correlation is weak to moderate. We can calculate the ϕ correlation in R as follows.

```
> library(rcompanion)
> phi(tea.tab)
phi
0.5
```

Definition 13.3. Fisher's Exact Test is a methodology for assessing the relationship between two categorical variables, each having two observable values. Fisher's Exact Test application to tables larger than 2×2 isn't straightforward is computationally difficult. To this end, we'll explore the formulas for Fisher's Exact Test for a 2×2 table and save Fisher's test for larger tables for R.

Fisher's exact test assesses the relationship between two variables using the **odds ratio** which represents the odds of the occurrence of one observed value of one variable given the observed value of a second variable; e.g., the odds of survival given the sex of the passenger is female.

The **Mantel-Haenszel odds ratio** is calculated as follows for a 2×2 table.

$$\text{OR}_{\text{MH}} = \frac{O_{11}/O_{21}}{O_{12}/O_{22}}$$

The **conditional maximum likelihood estimation (CMLE) odds ratio** requires optimization to calculate, but has been shown to provide less biased estimates.

$$\begin{aligned} a &= \max\{0, R_1 - C_2\} \\ b &= \min\{R_1, C_1\} \\ L(\Psi) &= \frac{\binom{C_1}{O_{11}}\binom{C_2}{O_{12}}\Psi^{O_{11}}}{\sum_{u=a_k}^{b_k}\binom{C_1}{O_{11}}\binom{C_2}{O_{12}}\Psi^{u}} \\ \text{OR}_{\text{CMLE}} &= \underset{\Psi}{\text{argmax}}\ L(\Psi). \end{aligned}$$

These values are interpreted as follows

$\text{OR} = 1 \implies$ observation of row 1 does not affect odds of an observation of column 1
$\text{OR} < 1 \implies$ observation of row 1 is associated with lower odds of an observation of column 1
$\text{OR} > 1 \implies$ observation of row 1 is associated with higher odds of an observation of column 1.

The Mantel-Haenszel odds ratio for our researchers is

$$\begin{aligned} \text{OR} &= \frac{O_{11}/O_{21}}{O_{12}/O_{22}} \\ &= \frac{3/1}{1/3} \\ &= 9. \end{aligned}$$

This indicates that Muriel guessing tea-first is associated with higher odds that the cup was made by adding tea-first. More specifically, if Muriel guesses tea-first the cup of tea is nine times more

likely to be made tea-first compared to milk-first. This provides evidence among the sample that Muriel has the ability to guess.

The CMLE odds ratio is set up as

$$\begin{aligned} a &= \max\{0, 4-4\} = 0 \\ b &= \min\{4, 4\} = 4 \\ L(\Psi) &= \frac{\binom{4}{3}\binom{4}{1}\Psi^3}{\sum_{u=0}^{4}\binom{4}{3}\binom{4}{1}\Psi^u} \\ \text{OR}_{\text{CMLE}} &= \underset{\Psi}{\text{argmax}}\ L(\Psi). \end{aligned}$$

This complicated maximization can be done in `R` as follows.

```
> OR.mle<-function(OR){
+   a=max(0,4-4)
+   b=min(4,4)
+   seq<-seq(a,b)
+   denom=0
+   for(u in seq){
+     denom=denom+choose(4,u)*choose(4,4-u)*OR^u
+   }
+   choose(4,3)*choose(4,1)*OR^3/denom
+ }
> optimize(f=OR.mle,interval=c(0, 100),maximum = TRUE)
$`maximum`
[1] 6.408326

$objective
[1] 0.5629951
```

Thus, it follows that

$$\text{OR}_{\text{CMLE}} = 6.408326.$$

The CMLE odds ratio, also greater than one, indicates that when Muriel guesses that the tea was made by adding tea-first there are higher odds that the cup is truly made with tea-first. More specifically, if Muriel guesses tea-first the cup of tea is 6.4083 times more likely to be made tea-first compared to milk-first.

Research Question: The ϕ correlation and odds ratios calculated provide evidence among the sample data that Muriel has some ability to guess between tea-first and milk-first. However, is this enough evidence to suggest that the Muriel can discern between the two methods of making tea more generally?

Hypotheses

The **null hypothesis**, H_0, for a Fisher's exact test is that two categorical variables are, in fact, independent among the population; e.g.,

$$H_0 : \text{OR} = 1 \implies \text{Muriel's guesses are independent of method}$$

The possible **alternative hypotheses**, H_a for a Fisher's exact test is that the two categorical variables are dependent among the population; e.g.,

$H_a : \text{OR} \neq 1 \implies$ Muriel's guess of tea-first is associated with the tea-first method
$H_a : \text{OR} < 1 \implies$ Muriel's guess of tea-first is associated with lower odds of the tea-first method
$H_a : \text{OR} > 1 \implies$ Muriel's guess of tea-first is associated with higher odds of the tea-first method

We would like to test for ability, OR> 1; e.g.,

$H_0 : \text{OR} = 1 \implies$ the guess of Muriel is independent of method
$H_a : \text{OR} > 1 \implies$ a guess of tea-first is associated with higher odds that it was tea-first.

Remark: An odds ratio less than one indicates evidence that Muriel guesses the opposite of the method used.

Assumptions

The assumptions of Fisher's exact test are as follows.

- The two variables are categorical.
- Observations are independent.
  - One observation doesn't affect later observations.
  - Each sample can be represented in one and only one cell of the table.
- There are fixed column and row totals.
  - This is a rare condition.

These assumptions hold for our researchers.

- The two variables are categorical.
  - guess – tea-first or milk first
  - method – tea-first or milk first
- Observations are independent.
  - One observation doesn't affect later observations.
  - Each sample can be represented in one and only one cell of the table.
- There are fixed column and row totals
  - The lady tasting tea knew that there were 4 cups made with tea-first and 4 cups made with milk first. This ensures that the row and column totals will be 4; e.g., they are fixed.

Test Statistic

For Fisher's exact test on a 2×2 table, we take the test statistic to be the number of observations O_{11}. However, the contingency table is necessary as we need many values to calculate a p-value. The table, in conjunction with the calculated odds ratio, give us a lot of interpretable information.

	Tea-First	Milk First	Total
Guessed Tea-First	O_{11} =3	O_{12} =1	R_1 =4
Guessed Milk First	O_{21} =1	O_{22} =3	R_2 =4
Total	C_1 =4	C_2 =4	T =8

P-Value

As in all of our hypothesis testing procedures, we require a sampling distribution to calculate the p-value of this hypothesis test. Under the null hypothesis, the test statistic of Fisher's exact test, O_{11}, has a hypergeometric distribution.

Result: The p value of Fisher's exact test is then calculated with the hypergeometric distribution using the following sampling distribution,

$$O_{11} \sim \text{hypergeometric}(m = C_1, n = C_2, k = R_1).$$

For our researchers,

$$O_{11} \sim \text{hypergeometric}(m = 4, n = 4, k = 4).$$

Graphs of the PDF and CDF is seen in Figure 13.2.1 and are created with the following `R` code.

```
> ggdat<-data.frame(x=(-1:5),
+                   f1=dhyper(x=(-1:5),m=4,n=4,k=4),
+                   F1=phyper(q=(-1:5),m=4,n=4,k=4))
> g1<-ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,5)+
+   ylim(0,0.55)+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Hypergeometric PMF",subtitle="m=4, n=4, k=4")
> ggdat.openpoints<-data.frame(x=ggdat$x,
+                              y=phyper(ggdat$x-1,m=4,n=4,k=4))
> ggdat.closedpoints<-data.frame(x=ggdat$x,
+                                y=phyper(ggdat$x,m=4,n=4,k=4))
> g1.CDF<-ggplot(data=ggdat, aes(x = x, y = F1)) +
+   geom_step()+
+   geom_point(data = ggdat.openpoints, aes(x = x, y = y),shape=1) +
+   geom_point(data = ggdat.closedpoints, aes(x = x, y = y)) +
+   geom_hline(yintercept=0)+
+   theme_bw()+
```

```
+   xlab("X")+
+   ylab(bquote(F[x](x)))+
+   ggtitle("Hypergeometric CDF",subtitle="m=4, n=4, k=4")
> grid.arrange(g1,g1.CDF,ncol=2)
```

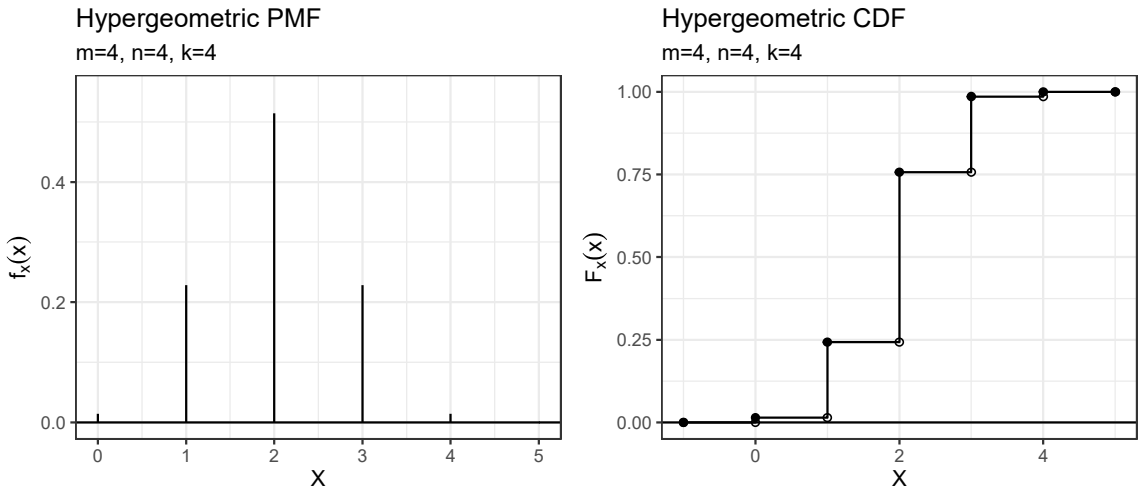

Figure 13.2.1: Hypergeometric PDF(left) and CDF(right) for $m = 4$, $n = 4$, $k = 4$; note that the possible values of X are $0, 1, 2, 3, 4$.

The p-value associated with this hypothesis test is calculated as follows for the observation $O_{11} = 3$.

$$P(O_{11} \geq 3) = P(O_{11} = 3) + P(O_{11} > 3)$$

We can ask for this probability in R as follows.

```
> dhyper(x=3,m=4,n=4,k=4)+phyper(q=3,m=4,n=4,k=4,lower.tail=FALSE)
[1] 0.2428571
```

Decision Making

We fail to reject the null hypothesis when the observed data is not considered unusual under H_0. A prespecified significance level of $\alpha = 0.05$ dictates that we reject the null hypothesis when the probability of the observed data or more extreme under H_0 is less than α.

Figure 13.2.2 shows that the observed data does not fall in the rejection region for this test and is created with the following code in R.

```
> alpha<-0.05
> ggdat.highlight<-data.frame(x=2:4,
+                             y=dhyper(x=(2:4),m=4,n=4,k=4))
> ggplot(data=ggdat,aes(x=x))+
+   geom_linerange(aes(ymax=f1), ymin=0)+
+   geom_ribbon(data=subset(ggdat,x>=qhyper(p=1-alpha,m=4,n=4,k=4)),aes(ymax=f1),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
```

```
+   geom_linerange(data=ggdat.highlight,aes(ymax=y),ymin=0,color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlim(0,5)+
+   ylim(0,0.55)+
+   xlab("X")+
+   ylab(bquote(f[x](x)))+
+   ggtitle("Fisher's Test of Independence",
+           subtitle=bquote(H[0]*":"~OR==1~"versus"~H[a]*":"~OR>1))+
+   annotate("text", x=4, y=0.1,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text", x=2.5,y=0.40, label="P-value=0.2429",size=3.5)
```

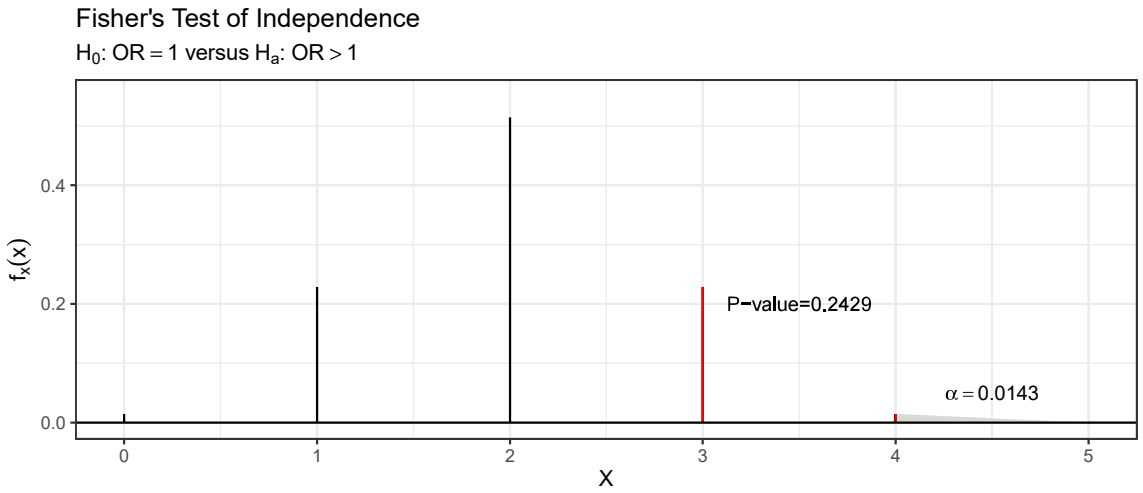

Figure 13.2.2: P-value as calculated for Example 13.1 with the hypergeometric distribution where $m = 4$, $n = 4$, $k = 4$.

Remark: We can see in the graph that even though we asked for a $\alpha = 0.05$ significance level, we can only perform the test at the $\alpha = 0.0143$ significance level due the the discrete sampling distribution. This is what makes the approximate results for inference about $\hat{p}$ better than the exact approaches.

Q1: Is the test statistic O_{11} in the rejection region?
A1: This question asks if the observed O_{11} is in the top $\alpha \times 100$ percentile of O_{11} under the null hypothesis. We can ask R for this percentile of the hypergeometric distribution.

```
> qhyper(p=0.95,m=4,n=4,k=4)
[1] 3
```

This requires some investigation due to the discrete nature of the sampling distribution. Should the rejection region be $X \geq 3$ or $X > 3$.

```
> dhyper(x=3,m=4,n=4,k=4)+phyper(q=3,m=4,n=4,k=4,lower.tail=FALSE)
[1] 0.2428571
```

```
> phyper(q=3,m=4,n=4,k=4,lower.tail=FALSE)
[1] 0.01428571
```

To observe at least 95% confidence, we would reject the null hypothesis for any observation of $\mathcal{X}_1^2$ greater than 3.

For our researchers: Since $O_{11} = 3$, it is not in the rejection region and so we fail to reject the null hypothesis.

Q2: Does the p-value indicate that the observation lies in the rejection region?
A2: If the p-value is less than alpha then the observation must lie in the rejection region. Thus, we would reject the null hypothesis for any observation that has a p-value less than 0.05.
For our researchers: The p-value for this test, p-value= 0.2428571, is greater than $\alpha = 0.05$ which indicates that we should fail to reject the null hypothesis.

Conclusion: We say "there is not sufficient evidence to suggest that Muriel can discern between the two methods of making tea (OR = 6.408309, p-value = 0.2429)."

We can run ask `R` to conduct this test directly as follows.

```
> fisher.test(x=tea.tab,alternative="greater")

Fisher's Exact Test for Count Data

data:  tea.tab
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
0.3135693       Inf
sample estimates:
odds ratio
6.408309
```

Remark: The `fisher.test` function also provides a 95% confidence interval for the odds ratio.

```
> fisher.test(x=tea.tab)

Fisher's Exact Test for Count Data

data:  tea.tab
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.2117329 621.9337505
sample estimates:
odds ratio
6.408309
```

Here, we can interpret this interval by saying "we are 95% confident that the true population odds ratio is between 0.2117329 and 621.9337505, the interval is wide because of the small sample size.

Note that because "1" is on the interval it is possible that they are independent, which matches the interpretation of Fisher's Test.

Summary

In this section, we succinctly summarize the five steps for Fisher's Test. While it might be simple to check this table and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Step One	H_0: OR$= 1$ H_a: OR< 1	H_0: OR$= 1$ H_a: OR> 1	H_0: OR$= 1$ H_a: OR$\neq 1$
Step Two	Check Assumptions		
Step Three	O_{11}		
Step Four	$P(O_{11} \leq o_{11})$	$P(O_{11} \geq o_{11})$	$1 - P(O_{11} \leq [o_{11}, (n+m) - o_{11}])$
Step Five	Reject H_0 if p-value$< \alpha$ or if O_{11} is in the rejection region.		

For our researchers,

Step One	H_0: OR$= 1 \Longrightarrow$ Muriel's guesses are independent of method H_a: OR$> 1 \Longrightarrow$ Muriel's guess of tea-first is associated with higher odds of the tea-first method
Step Two	Variables are categorical reasonably independent fixed column and row totals sample size assumption is met $20 < 891$
Step Three	$o_{11} = 3$
Step Four	$P(O_{11} \geq o_{11}) = 0.2428571$
Step Five	p-value $> 0.05 \longrightarrow$ fail to reject H_0

13.2.2 Chi-Square Independence Test

The Chi-square independence test is a nonparametric alternative to the Fisher test that straightforwardly works for tables of various dimensions.

Example 13.4. ? conducted a poll with a sample of 1993 registered voters from May 10-14, 2018 . The poll asked about the voters age, race/ethnicity, gender, educational attainment, and region and a variety of political questions including the following.

> *As you may know, the United States and other countries made a deal in 2015 to lift economic sanctions against Iran in exchange for Iran agreeing not to manufacture nuclear weapons. This agreement is sometimes referred to as the Iran nuclear deal. Knowing this, do you support or oppose this agreement?*

Data on 1,694 respondents who answered the two questions of interest are reported below in a 3×5 contingency table – 3×5 denotes 3 rows and 5 columns.

	Strongly support	Somewhat support	Somewhat oppose	Strongly oppose	No opinion	Total
Some college or less	O_{11} =291	O_{12} =375	O_{13} =131	O_{14} =216	O_{15} =19	R_1 =1032
Bachelor degree	O_{21} =111	O_{22} =152	O_{23} =58	O_{24} =82	O_{25} =14	R_2 =417
Post-graduate	O_{31} =65	O_{32} =86	O_{33} =29	O_{34} =51	O_{35} =14	R_3 =245
Total	C_1 =467	C_2 =613	C_3 =218	C_4 =349	C_5 =47	n =1694

The data is loaded in R as follows.

```
> R1<-c(291,375,131,216,19)
> R2<-c(111,152,58, 82, 14)
> R3<-c(65, 86, 28, 51, 14)
> Iran.tab<-matrix(data=c(R1,R2,R3),
+                  nrow = 3,
+                  ncol = 5,
+                  byrow = TRUE)
> colnames(Iran.tab)<-c("Strongly support","Somewhat support","Somewhat oppose","Strongly oppose","No opinion")
> rownames(Iran.tab)<-c("Some college or less","Bachelor degree", "Post-graduate")
> Iran.tab
                    Strongly support Somewhat support Somewhat oppose Strongly oppose No opinion
Some college or less              291              375              131              216         19
Bachelor degree                   111              152               58               82         14
Post-graduate                      65               86               28               51         14
> (row.sums<-rowSums(Iran.tab))
Some college or less      Bachelor degree        Post-graduate
1032                  417                  244
>(col.sums<-colSums(Iran.tab))
Strongly support Somewhat support  Somewhat oppose  Strongly oppose       No opinion
467              613               217              349               47
>(obs<-sum(Iran.tab))
[1] 1693
```

Research Question: Is there enough evidence to suggest a relationship between education and support for the Iran nuclear deal?

Definition 13.5. The **chi-square independence** hypothesis test is used to assess whether or not two categorical variables are independent using their observation frequencies. Here, we want to test whether this observed dependence between sex and survival status is simply due to random chance.

Hypotheses

The **null hypothesis**, H_0, for a chi-square independence hypothesis test is that two categorical variables are, in fact, independent; e.g.,

$$H_0 : \text{the two categorical variables of interest are independent.}$$

The **alternative hypothesis**, H_a for a chi-square independence hypothesis test is the opposite of the null hypothesis – that the two categorical variables are dependent; e.g.,

$$H_a : \text{the two categorical variables of interest are dependent.}$$

For our researchers,

H_0 : the education of registered voters is independent of support for the Iran nuclear deal

H_a : the education of registered voters is dependent of support for the Iran nuclear deal.

Equivalently, we could state these hypotheses as,

H_0 : the education of registered voters is not related to support for the Iran nuclear deal
H_a : the education of registered voters is related to support for the Iran nuclear deal.

Assumptions

When completing a chi-squared independence hypothesis test we have to check our assumptions for using such methodology. To do so, we need to introduce a new key value – expected counts.

Definition 13.6. The **expected count** is the expected number of observations (out of n) that we would expect to see in a cell if the observations were truly independent; e.g., the null hypothesis is true. This can be calculated for the cell in row i and column j as follows.

$$E_{ij} = \frac{\text{Total of Row } i \times \text{Total of Column } j}{n}.$$

Note: The expected counts are the counts we'd expect in each cell of the table should the events truly be independent.

The assumptions for the chi-squared independence hypothesis test include the following.

- The two variables are categorical.
  - Each category has two or more mutually exclusive observable values.
- Observations are independent.
  - One observation doesn't affect later observations.
  - Each individual can be represented in one and only one cell of the table.
- The sample size is at least the number of cells in the table multiplied by 5.
- 80% of the expected counts are greater than 5.
- None of the expected counts are less than one.

Remark: The sample size and the expected count assumptions are highly related. These assumptions are fully explained using the Titanic example below.

To test these hypotheses using the chi-squared independence hypothesis test we check the assumptions below.

- The two variables are categorical
  - education – some college or less, bachelor degree, or post-graduate
  - support status – strongly support, somewhat support, somewhat oppose, strongly oppose, no opinion.
- Observations are independent
  - each individual is only reported once

- it can be reasonably assumed that observations are independent – a registered voter's outcome doesn't affect another registered voter.

- The sample size is at least the number of cells in the table multiplied by 5;

$$\text{number of cells} \times 5 = 15 \times 5 = 75 < 1694.$$

- Expected count assumptions are met
  - 100% of the expected counts are greater than 5

$$\begin{aligned}
E_{11} &= \frac{1032 \times 467}{1694} = 284.5006 & E_{21} &= \frac{417 \times 467}{1694} = 118.0952 & E_{31} &= \frac{245 \times 467}{1694} = 67.54132 \\
E_{12} &= \frac{1032 \times 613}{1694} = 373.4451 & E_{22} &= \frac{417 \times 613}{1694} = 155.0158 & E_{32} &= \frac{245 \times 613}{1694} = 88.65702 \\
E_{13} &= \frac{1032 \times 218}{1694} = 132.8076 & E_{23} &= \frac{417 \times 218}{1694} = 55.12796 & E_{33} &= \frac{245 \times 218}{1694} = 31.52893 \\
E_{14} &= \frac{1032 \times 349}{1694} = 212.6139 & E_{24} &= \frac{417 \times 349}{1694} = 88.25531 & E_{34} &= \frac{245 \times 349}{1694} = 50.47521 \\
E_{15} &= \frac{1032 \times 47}{1694} = 28.63282 & E_{25} &= \frac{417 \times 47}{1694} = 11.88539 & E_{35} &= \frac{245 \times 47}{1694} = 6.797521
\end{aligned}$$

  - none of the expected counts are less than one.

The assumptions are met and so we can move forward with the test.

Test Statistic

The test statistic for a chi-squared independence hypothesis test is

$$\mathcal{X}^{2*} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Remark: By construction, if the number of observations, O_{ij}, is equal to the expected number of observations under independence, E_{ij}, for every cell in the table then the test statistic is zero. If they are different, the test statistic grows.

Definition 13.7. A **Yates' continuity correction** is often used when conducting a chi-squared independence hypothesis test on a 2×2 table. This correction is not dissimilar to the correction used when conducting a two-sample hypothesis test about two population proportions.

Here, the chi-squared test statistic is corrected as follows.

$$\mathcal{X}^{2*} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(|O_{ij} - E_{ij}| - \frac{1}{2}\right)^2}{E_{ij}}$$

We can ask `R` for the corrected version, which it does by default as the preferred method in the 2×2 case by passing in the argument `correct=TRUE`.

For our researchers, the test statistic can be calculated using the following table in which the expected counts have been appended to each cell in parentheses.

	Strongly support	Somewhat support	Somewhat oppose	Strongly oppose	No opinion	Total
Some college or less	291 (284.5)	375 (373.4)	131 (132.8)	216 (212.6)	19 (28.6)	1032
Bachelor degree	111 (118.1)	152 (155.0)	58 (55.1)	82 (88.3)	14 (11.9)	417
Post-graduate	65 (67.5)	86 (88.7)	29 (31.5)	51 (50.5)	14 (6.8)	245
Total	467	613	218	349	47	1694

We calculate these in R as follows.

```
> (E.R1<-row.sums[1]*col.sums/obs)
Strongly support Somewhat support  Somewhat oppose  Strongly oppose       No opinion
284.66864        373.66568         132.27643        212.73952         28.64973
> (E.R2<-row.sums[2]*col.sums/obs)
Strongly support Somewhat support  Somewhat oppose  Strongly oppose       No opinion
115.02599        150.98701          53.44891         85.96161         11.57649
> (E.R3<-row.sums[3]*col.sums/obs)
Strongly support Somewhat support  Somewhat oppose  Strongly oppose       No opinion
67.305375        88.347312         31.274660        50.298878         6.773774
> (Iran.Etab<-matrix(data=c(E.R1,E.R2,E.R3),
+                    nrow = 3,
+                    ncol = 5,
+                    byrow = TRUE))
[,1]      [,2]      [,3]      [,4]      [,5]
[1,] 284.66864 373.66568 132.27643 212.73952 28.649734
[2,] 115.02599 150.98701  53.44891  85.96161 11.576491
[3,]  67.30538  88.34731  31.27466  50.29888  6.773774
```

The test statistic is calculated as follows.

$$\begin{aligned}\mathcal{X}^{2*} =& \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\\ =& \frac{(291-284.5006)^2}{284.5006}+\frac{(375-373.4451)^2}{373.4451}+\frac{(131-132.8076)^2}{132.8076}+\frac{(216-212.6139)^2}{212.6139}+\frac{(19-28.63282)^2}{28.63282}\\ &+\frac{(111-118.0952)^2}{118.0952}+\frac{(152-155.0158)^2}{155.0158}+\frac{(58-55.12796)^2}{55.12796}+\frac{(82-88.25531)^2}{88.25531}+\frac{(14-11.88539)^2}{11.88539}\\ &+\frac{(65-67.54132)^2}{67.54132}+\frac{(86-88.65702)^2}{88.65702}+\frac{(29-31.52893)^2}{31.52893}+\frac{(51-50.47521)^2}{50.47521}+\frac{(14-6.797521)^2}{6.797521}\\ =& 12.94349\end{aligned}$$

We calculate this in R as follows.

```
> (test.stat<-sum((Iran.tab-Iran.Etab)^2/Iran.Etab))
[1] 12.88611
```

P-value

We require the sampling distribution of $\mathcal{X}^{2*}$ to calculate the p-value of this hypothesis test. The sampling distribution of $\mathcal{X}^{2*}$ is estimated well by the **chi-squared** ($\mathcal{X}^2_v$) distribution.

Definition 13.8. The degrees of freedom (v) for a chi-squared independence hypothesis test is calculated as

$$v = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

Result: The p value of the chi-squared independence hypothesis test is then calculated as the right tail probability; e.g.,

$$p\text{-value} = P(\mathcal{X}^2_v > \mathcal{X}^{2*}) = 1 - P(\mathcal{X}^2_v \leq \mathcal{X}^2)$$

For our researchers,

$$v = (3-1) \times (5-1) = 8$$

and the p-value is calculated as

$$p\text{-value} = P(\mathcal{X}^2_8 > 12.88611) = 1 - P(\mathcal{X}^2_8 \leq 12.88611).$$

This value can be calculated in R as follows. We calculate these in R as follows.

```
> 1-pchisq(q=test.stat,df=8)
[1] 0.1158289
```

We fail to reject the null hypothesis since the p-value is less than 0.05. We conclude that there is not sufficient evidence to suggest that there is a relationship between educational attainment and support of the Iran nuclear deal among registered voters ($\mathcal{X}^2_8 = 12.88611$, p-value= 0.1158289).

Decision Making

We reject the null hypothesis when the observed data is unusual under H_0. A prespecified significance level of $\alpha = 0.05$ dictates that we reject the null hypothesis when the probability of the observed data or more extreme under H_0 is less than α.

Figure 13.2.3 shows that the observed data does not fall in the rejection region for this test.

```
> alpha<-0.05 #significance level
> v<-(nrow(Iran.tab)-1)*(ncol(Iran.tab)-1)
> ggdat<-data.frame(chisq=seq(from=0,to=30,by=0.01),
+                   f=dchisq(seq(from=0,to=30,by=0.01),df=v))
> #plot a point at the 0.95 quantile
> ggdat.highlight<-data.frame(x=test.stat,
+                             y=0)
> ggplot(data=ggdat,aes(x=chisq,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,chisq>=qchisq(1-alpha,df=v)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA,alpha=0.5)+
+   geom_ribbon(data=subset(ggdat,chisq>=test.stat),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab(bquote(chi^2))+
```

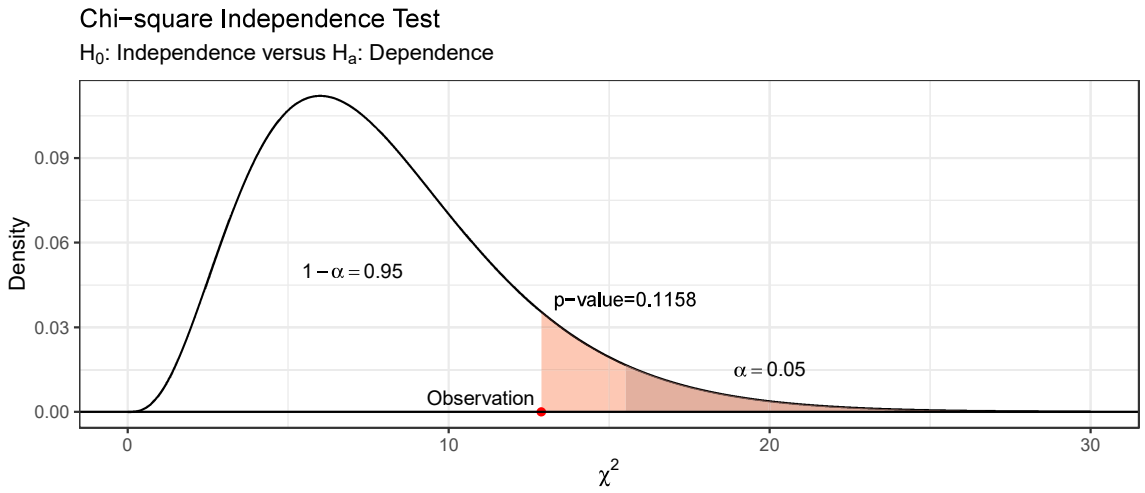

Figure 13.2.3: Chi-squared ($v = 8$) PDF with rejection region shaded in grey, the observation highlighted with a red point, and the p-value shaded in red.

```
+   ylab("Density")+
+   ggtitle("Chi-square Independence Test",
+           subtitle=bquote(H[0]*": Independence versus "*H[a]*": Dependence"))+
+   annotate("text",x=20,y=0.015,label=deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text",x=7,y=0.05,label=deparse(bquote(1-alpha==0.95)),parse=TRUE,size=3.5)+
+   annotate("text",x=11,y=0.005,label="Observation",size=3.5)
+   annotate("text",x=15.5,y=0.04,label="p-value=0.1158",size=3.5)
```

While this is clear from the graph, we are curious about how to decide numerically whether or not the observed data is in the rejection region, noting we did this as we asked R to create the graph.

Q1: Is the test statistic $\mathcal{X}^{2*}$ in the rejection region?
A1: This question asks if the observed $\mathcal{X}^{2*} = 12.88611$ is in the top $\alpha \times 100$ percentile of $\mathcal{X}^2_8$ under the null hypothesis. We can ask R for this percentile of $\mathcal{X}^2_8$ using the sampling distribution.

```
> qchisq(p=0.95,df=8)
[1] 15.50731
```

Thus, we would reject the null hypothesis for any observation of $\mathcal{X}^2_8$ greater than 15.50731.

For our researchers: Since $\mathcal{X}^{2*} = 12.88611$, it is not in the rejection region and so we fail to reject the null hypothesis.

Q2: Does the p-value indicate that the observation lies in the rejection region?
A2: If the p-value is less than alpha then the observation must lie in the rejection region. Thus, we would reject the null hypothesis for any observation that has a p-value less than 0.05.
For our researchers: The p-value for this test, p-value= 0.1158289, is greater than $\alpha = 0.05$ which indicates that we should fail to reject the null hypothesis.

Conclusion: We say "there is not sufficient evidence to suggest that there is a relationship between educational attainment and support of the Iran nuclear deal among registered voters ($\mathcal{X}^2_1 = 12.8861$, p-value= 0.1158)."

Remark: The answers to these questions will always match. In practice, quantitative researchers report the test statistic and p-value when drawing their conclusions.

This test can be conducted directly in R as follows.

```
> chisq.test(Iran.tab)

Pearson's Chi-squared test

data:  Iran.tab
X-squared = 12.886, df = 8, p-value = 0.1158
```

Remark: The reason for this non-parametric test is that for large tables Fisher's test is not applicable. It's instructive to consider using Fisher's exact test for this example. With this example we have too much data to efficiently calculate Fisher's exact test; not to mention, the fixed column and row total assumption is problematic here.

```
> fisher.test(tab,alternative="two.sided")
Error in fisher.test(tab,alternative="two.sided") : FEXACT error 6.
LDKEY is too small for this problem.
Try increasing the size of the workspace.
```

This is due to the computational cost associated with the large observation counts. To avoid this complication, R makes available a methodology that approximates the p-value based on simulations; this can be done as follows.

```
> fisher.test(tab,simulate.p.value=TRUE,B=1000)

Fisher's Exact Test for Count Data with simulated
p-value (based on 1000 replicates)

data:  Iran.tab
p-value = 0.1598
alternative hypothesis: two.sided
```

We note that the p-value is slightly larger as Fisher's exact test is more conservative than the chi-squared independence hypothesis test.

We also note that the chi-squared test requires a large sample size. If we don't have enough data to satisfy the assumptions, R will warn us that we might not be using to correct approach. Below, we see the warning when running the `chisq.test()` function on Muriel's guesses about tea preparation.

```
> chisq.test(tea.tab)

Pearson's Chi-squared test with Yates' continuity correction

data:  tea.tab
X-squared = 0.5, df = 1, p-value = 0.4795
```

```
Warning message:
In chisq.test(tea.tab) : Chi-squared approximation may be incorrect
```

Summary

In this section, we succinctly summarize the five steps for the chi-squared independence test. While it might be simple to check this table and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Step One	H_0: Variables are independent H_a: Variables are dependent
Step Two	Check Assumptions
Step Three	$\mathcal{X}^{2*} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
Step Four	$P(\mathcal{X}^2 > \mathcal{X}^{2*})$
Step Five	Reject H_0 if p-value$< \alpha$ or if $\mathcal{X}^{2*}$ is in the rejection region

For our researchers,

Step One	H_0: Sex and survival are independent H_a: Sex and survival are dependent
Step Two	Variables are categorical reasonably independent sample size assumption is met $20 < 891$ 100% of expected counts are greater than 5 non of the expected counts are less than 1
Step Three	$\mathcal{X}^{2*} = 263.0503$
Step Four	$P(\mathcal{X}^2 > \mathcal{X}^{2*}) < 0.0001$
Step Five	p-value $< 0.05 \longrightarrow$ reject H_0

Definition 13.9. Cramér's V correlation is equivalent to the phi coefficient of correlation for a 2×2 contingency table. For an $r \times c$ we can calculate the correlation as

$$v = \sqrt{\frac{\mathcal{X}^{2*}/n}{\min(r-1, c-1)}}$$

where

- $\mathcal{X}^{2*}$ is the test statistic from an uncorrected chi-square independence test
- r is the number of rows in the table
- c is the number columns.

This estimator has been shown to be biased, and tends to overestimate the the true association. A bias-correct version of this statistic (Bergsma, 2013) is calculated as follows.

$$v_c = \sqrt{\frac{\max\left(0, \frac{\mathcal{X}^{2*}}{n} - \frac{(r-1)(c-1)}{n-1}\right)}{\min\left(\frac{(r-1)^2}{n-1} - 1, \frac{(c-1)^2}{n-1} - 1\right)}}$$

We can interpret the values of Cramer's V the same way we did ϕ when the table is 2×2, but Cramer's V decreases as the dimension of the table increases. Cohen (2013), provides insight into how we can account for the size of the table, with respect to $k = \min(r, c)$, in our interpretations which is summarized in Table 13.9.

	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
strong association	>0.70	>0.495	>0.404	>0.350	>0.313
moderate association	>0.50	>0.354	>0.289	>0.250	>0.224
weak association	>0.30	>0.212	>0.173	>0.150	>0.134
little or no association	>0.00	>0.000	>0.000	>0.000	>0.000

Remark: Because the ϕ correlation coefficient is a special case of Cramer's V, we will simply use Cramer's V going forward. This will be advantageous, because the calculation of the bias-corrected version is simple to do in R.

For Fisher's experiment involving Muriel's guesses on tea preparation we ask for the uncorrected chi-squared statistic

```
> chisq.test(tea.tab,correct=FALSE)

Pearson's Chi-squared test

data:  tea.tab
X-squared = 2, df = 1, p-value = 0.1573

Warning message:
In chisq.test(tea.tab, correct = FALSE) :
Chi-squared approximation may be incorrect
```

and calculate Cramer's V as follows.

$$\begin{aligned} v &= \sqrt{\frac{\mathcal{X}^{2*}/n}{\min(r-1, c-1)}} \\ &= \sqrt{\frac{2/8}{\min(2-1, 2-1)}} \\ &= \sqrt{\frac{2}{8}} \\ &= \sqrt{0.25} \\ &= 0.50 \end{aligned}$$

With respect to the values in Table 13.9, $v = 0.50$ indicates weak to moderate association between Muriel's guesses and the method with which the tea was made. We can ask for the uncorrected, and bias-corrected Cramer's V values directly in R as follows.

```
> cramerV(tea.tab)
Cramer V
0.5
> cramerV(tea.tab,bias.correct = TRUE)
```

```
Cramer V
0.3536
```

We can also ask for a confidence interval, via bootstrapping, using the following R code.

```
> library(RVAideMemoire)
> cramer.test(tea.tab,conf.level = 0.95)

Cramér's association coefficient

data:  tea.tab
X-squared = 0.5, df = 1, p-value = 0.4795
alternative hypothesis: true association is not equal to 0
95 percent confidence interval:
0.00 0.75
sample estimates:
V
0.25
```

For the Iran deal poll involving questions about the relationship between educational attainment and support for the Iran deal, we can calculate Cramer's V as follows.

$$\begin{aligned} v &= \sqrt{\frac{\mathcal{X}^{2*}/n}{\min(r-1, c-1)}} \\ &= \sqrt{\frac{12.886/1694}{\min(3-1, 5-1)}} \\ &= \sqrt{\frac{12.886}{1694(2)}} \\ &= \sqrt{0.003803424} \\ &= 0.0616719 \end{aligned}$$

With respect to the values in Table 13.9, $v = 0.0617$ indicates there is little to no relationship between educational attainment and support of the Iran nuclear deal among registered voters. We can ask for the uncorrected, and bias-corrected Cramer's V values directly in R as follows.

```
> cramerV(Iran.tab)
Cramer V
0.06169
> cramerV(Iran.tab,bias.correct = TRUE)
Cramer V
0.03799
```

We can also ask for a confidence interval, via bootstrapping, using the following R code.

```
> cramer.test(Iran.tab,conf.level=0.95)

Cramér's association coefficient
```

```
data:  Iran.tab
X-squared = 12.886, df = 8, p-value = 0.1158
alternative hypothesis: true association is not equal to 0
95 percent confidence interval:
0.04578157 0.11375049
sample estimates:
V
0.06169037

Warning message:
In cramer.test(Iran.tab, conf.level = 0.95) :
at least 1 level contains less than 5% of total number of individuals
```

13.3 Describing Continuous Relationships: Scatterplots and Correlation

13.3.1 Introduction

Example 13.10. In Japan, where space is limited, there's an increased need for space saving in processing sludge, a mixture of liquid and solid waste. The sludge thickening–dewatering process decreases the volume of waste by removing liquid from the sludge. Thickening removes liquid, but still allows the sludge to be pumped as a liquid and dewatering sludge leads to solid sludge that can be transported by truck.

Watanabe and Tanaka (1999) developed a new compression machine for processing sewage sludge. This new system conditions sludge with an inorganic coagulant, peletizes it, and removes water with belt press filter. The benefit of their new methodology yields a dewatered sludge with increased concentration of solids in a fraction of time – 10-20 minutes compared to the conventional 12 hour procedure.

The engineers show that their new system produces compressed pellets with small moisture content (y) and high filtration rates (x) which is where their time savings come from.

$$x = \text{belt press filtration rate (measured in kg/m/hr)}$$
$$y = \text{moisture of compressed pellets (measured as a \%)}.$$

To describe the resulting moisture of compressed pellets (y) as a function of the belt press filtration rate, the engineers took a random sample of $n = 20$ individual sewage specimens.

Specimen	x	y	Specimen	x	y
1	125.3	77.9	11	159.5	79.9
2	98.2	76.8	12	145.8	79.0
3	201.4	81.5	13	75.1	76.7
4	147.3	79.8	14	151.4	78.2
5	145.9	78.2	15	144.2	79.5
6	124.7	78.3	16	125.0	78.1
7	112.2	77.5	17	198.8	81.5
8	120.2	77.0	18	132.5	77.0
9	161.2	80.1	19	159.6	79.0
10	178.9	80.2	20	110.7	78.6

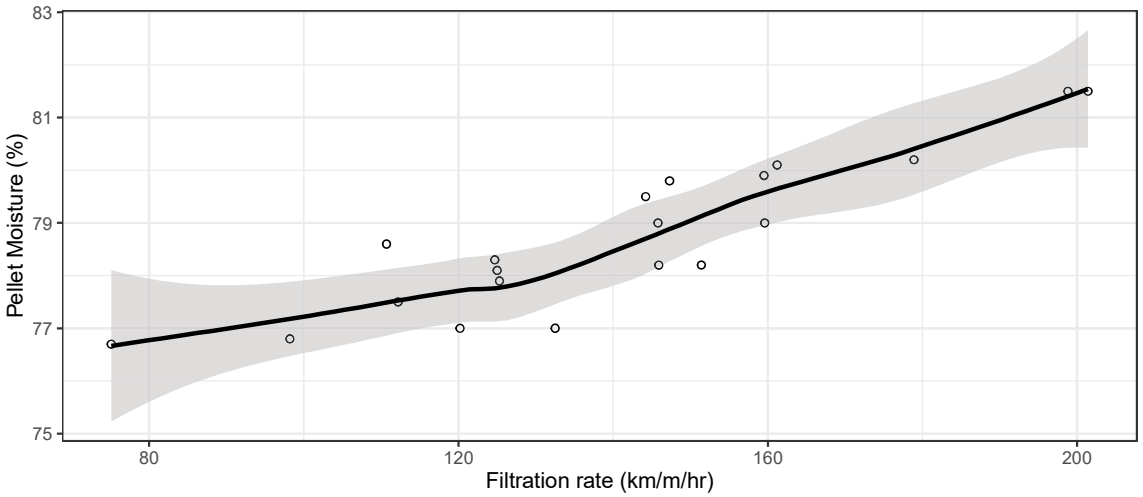

Figure 13.3.4: Scatterplot of filtration rate (x, measured in kg/m/hr) and pellet moisture (y, measured as a %) for a sample of $n = 20$ sewage specimens.

Definition 13.11. A **scatterplot** is a graphical display that shows the relationship between two quantitative variables measured on the same individuals.

- The values of one variable appear on the horizontal axis; the values of the other variable appear on the vertical axis.
- Scatterplots give a visual impression of how the two variables behave together.
- Figure 13.3.4 shows the scatterplot between filtration rate (x) and pellet moisture (y) in Example 13.10. This graph shows a **positive linear relationship** between the two variables.

This graph is obtained in `R` as follows.

```
> filt.rate<-c(125.3,98.2,201.4,147.3,145.9,124.7,112.2,120.2,161.2,178.9,
+              159.5,145.8,75.1,151.4,144.2,125,198.8,132.5,159.6,110.7)
> moisture<-c(77.9,76.8,81.5,79.8,78.2,78.3,77.5,77,80.1,80.2,
+             79.9,79,76.7,78.2,79.5,78.1,81.5,77,79,78.6)
> ggdat<-data.frame(filt.rate=filt.rate,moisture=moisture)
> ggplot(ggdat, aes(x=filt.rate, y=moisture)) +
+   geom_point(shape=1)+
+   geom_smooth(alpha=0.25,color="black",method="loess")+
+   theme_bw()+
+   xlab("Filtration rate (km/m/hr)")+
+   ylab("Pellet Moisture (%)")
```

Example 13.12. Elementary school performance in California is based on a standardized exam called the Stanford 9 standardized test. The data, made available by Croissant (2016), include information from tests administered 1998-1999 to 5th grade students for all 420 K-6 and K-8 districts in California. Suppose we want to examine the relationship between

x = percentage of students qualifying for reduced-price lunch

y = average math score.

A scatterplot of these observations is shown in Figure 13.3.5. This graph shows a **negative linear relationship** between the two variables. This graph is obtained in R as follows.

```
> install.packages("Ecdat")
> library("Ecdat")
> data(Caschool)
> ggdat<-data.frame(mealpct=Caschool$mealpct,mathscr=Caschool$mathscr)
> ggplot(ggdat, aes(x=mealpct, y=mathscr)) +
+   geom_point(shape=1)+
+   geom_smooth(alpha=0.25,color="black",method="loess")+
+   theme_bw()+
+   xlab("Students qualifying for reduced-price lunch (%)")+
+   ylab("Average math score")
```

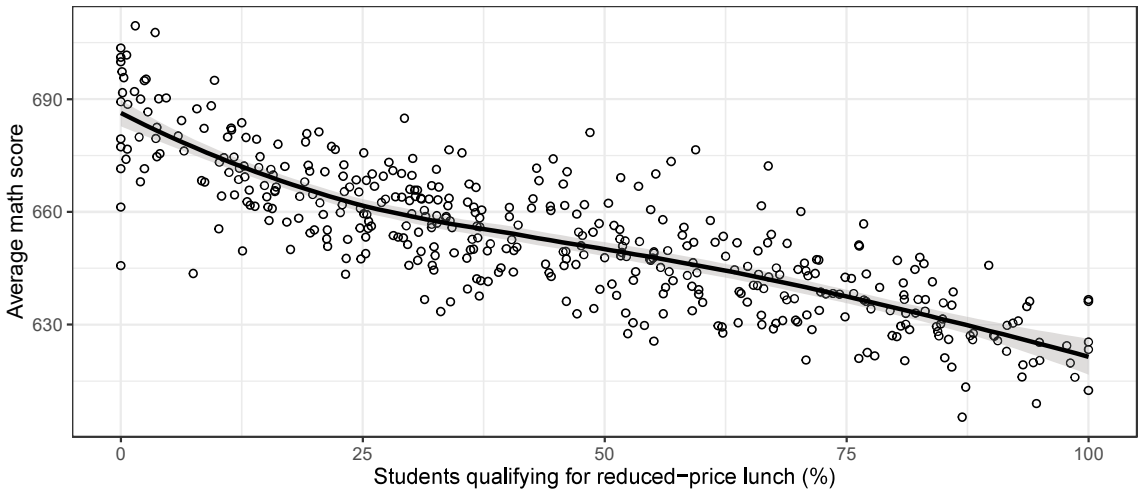

Figure 13.3.5: Scatterplot of the percentage of students who qualify for reduced-price lunch (x) and average math score on a standardized exam (y) for $n = 420$ schools in California.

13.3.2 Interpreting scatterplots

Remark: Constructing scatterplots is easy. Interpreting them is more important. The first thing we should remember is **statistical inference**.

- Scatterplots are used to show the relationship between observations for two quantitative variables.
- If the observations we have are from a representative **sample**, then the scatterplot presents an impression of the underlying relationship for the **population**.
- Therefore, by interpreting characteristics we see in the scatterplots, we are interpreting what may be "going on" in the larger population of individuals.

Interpretation: We will focus on the following characteristics when we examine and describe scatterplots:

1. Overall pattern:
    - Form: Are there straight-line (linear) patterns or curved patterns? Do observations tend to fall into clusters?
    - Direction: Are the variables positively related or negatively related?
    - Strength: Is the relationship strong, moderate, or mild? Perhaps there is no relationship (e.g., a random scatter of points)?
2. Deviations from the overall pattern (e.g., outliers, etc.).

Definitions: Two quantitative variables are **positively related** when an increase in one variable tends to accompany an increase in the other. They are **negatively related** when an increase in one variable tends to accompany a decrease in the other.

- Example 13.10: Filtration rate (x) and pellet moisture (y) are positively related.
- Example 13.12: The percentage of students who qualify for reduced-price lunch (x) and average math score (y) are negatively related.

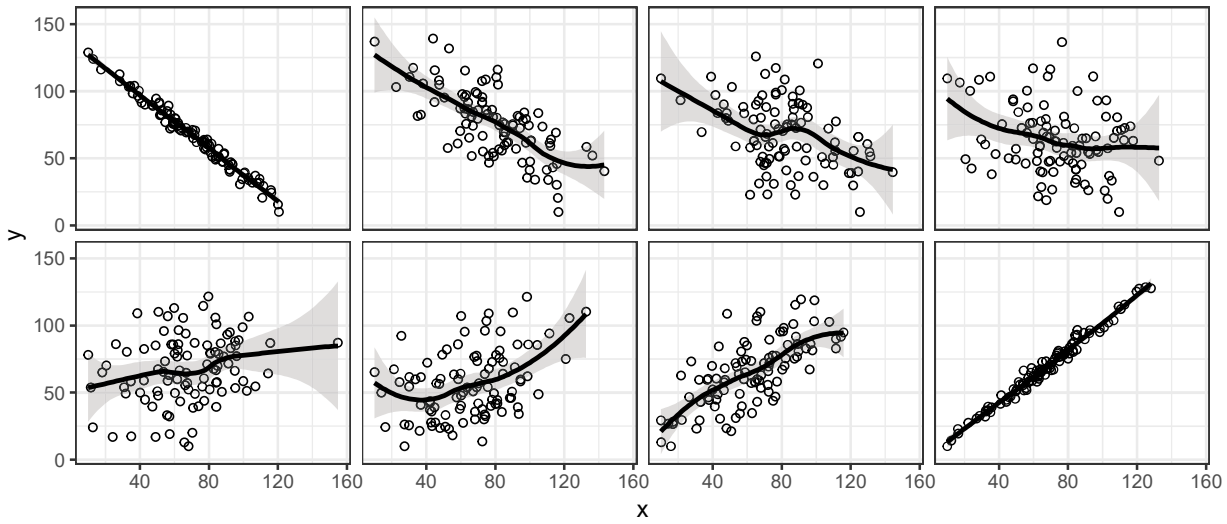

Figure 13.3.6: Scatterplot examples. **Top row:** very strong, moderate, weak, very weak negative linear relationship. **Bottom row:** very weak, weak, moderate, very strong positive linear relationship.

13.3.3 Correlation

Goal: We would like to study the relationship between two quantitative variables, x and y. Scatterplots give us a visual display of the relationship. We now wish to summarize this relationship **numerically**.

Recall our discussion about variance in Chapter 3. We said that variance measures the expected squared distance from the mean; e.g., for data $y_1, y_2, \ldots y_n$ we calculate the sample variance as

$$s_y = \frac{\sum_{i=1}^{n} (y - \bar{y})^2}{(n-1)}.$$

In essence, what we're going to measure with correlation is how the variability of one random variable, Y, is explained by its relationship with another random variable X.

Definition 13.13. The **Pearson correlation** (r) is a numerical summary that describes the strength and direction of the straight-line (linear) relationship between two quantitative variables. With a sample of n individuals, the correlation is computed by the following formula:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right),$$

where $\overline{x}$ and $\overline{y}$ are the sample means and s_x and s_y are the sample standard deviations. Note that the terms

$$\frac{x-\overline{x}}{s_x} \quad \text{and} \quad \frac{y-\overline{y}}{s_y}$$

are the **sample standardized values** of x and y, respectively.

Important: We will use `R` to compute the correlation with real data. It will be our job to understand what its value means.

Let's discuss the important facts about the correlation r.

Definition 13.14. The **correlation** is a numerical summary that describes the strength and direction of the relationship between two variables.

- **Fact 1:** If two variables have a positive relationship then the correlation is positive. If two variables have a negative relationship then the correlation is positive.
- **Fact 2:** The correlation is always between -1 and 1.
- **Fact 3:** The correlation r is **unitless**.
- **Fact 4:** The correlation between two variables X and Y is the same as the correlation between Y and X.
- **Fact 5:** The correlation only measures the strength and the direction of the relationship between two variables.
- **Fact 6:** The value of the correlation can be highly affected by **outliers** in some cases.

A general guideline for interpreting correlations is

- -1.0 to -0.7 strong negative association
- -0.7 to -0.5 moderate negative association
- -0.5 to -0.3 weak negative association
- -0.3 to +0.3 little or no association
- +0.3 to +0.5 weak positive association
- +0.5 to +0.7 moderate positive association
- +0.7 to +1.0 strong positive association

Fact 1: If x and y have a positive linear relationship, then $r > 0$. If x and y have a negative linear relationship, then $r < 0$. Note, $r = 0$ means that x and y have no linear relationship.

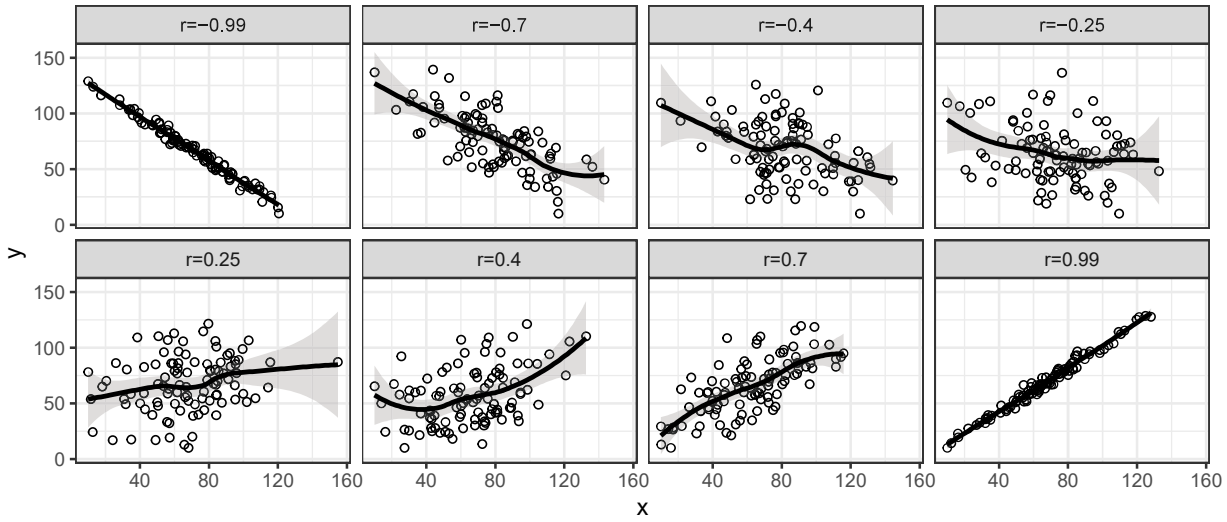

Figure 13.3.7: Scatterplot examples. **Top row:** $r < 0$. **Bottom row:** $r > 0$.

Fact 2: The correlation r is always between -1 and 1; i.e.,

$$-1 \leq r \leq 1.$$

What happens at the endpoints?

- If $r = 1$, then all of the data fall on a straight line with **positive** slope.
- If $r = -1$, then all of the data fall on a straight line with **negative** slope.
- In either case, the relationship between the two variables x and y is **perfectly linear**.

Remark: Perfect relationships are a rarity with real life data; e.g., see the scatterplots in Examples 13.10 and 13.12. There are almost always other sources of variation that make a relationship not perfect (e.g., lurking variables, etc.).

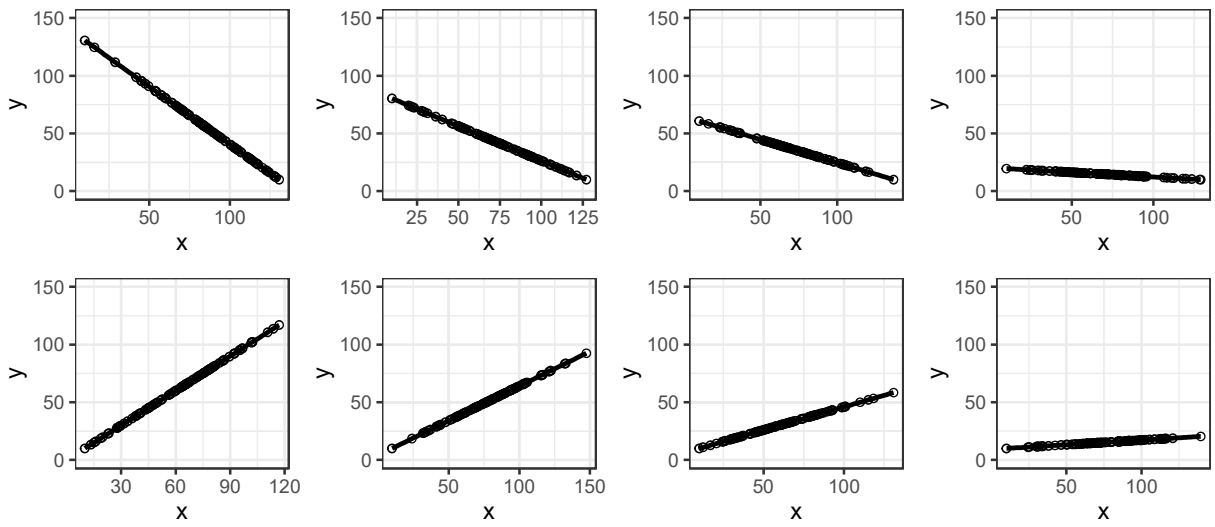

Figure 13.3.8: Scatterplot examples. It is important to note that slope only affects the sign of r, not the value. For the examples above, **Top row:** $r = -1$. **Bottom row:** $r = 1$. This is because the points have a perfect linear relationship.

Illustration: Let's use `R` to calculate the correlation r for the data in Examples 13.10 and 13.12.

Continuing Example 13.10:

```
> cor(filt.rate,moisture,method = "pearson")
[1] 0.8943937
```

The correlation between filtration rate (x) and pellet moisture (y) is $r \approx 0.89$. This represents a strong, positive linear relationship.

Continuing Example 13.12:

```
> cor(Caschool$mealpct,Caschool$mathscr,"pearson")
[1] -0.8230145
```

The correlation between the percentage of students who qualify for reduced-price lunch (x) and the average math score (y) is $r \approx -0.82$. This represents a strong, negative linear relationship.

Fact 3: The correlation r is **unitless**; i.e., there are no units attached to it (e.g., dollars, cm, etc.). This also means that you could change the units of your data (e.g., inches to cm, percentages to proportions, etc.), and this would not change the value of r.

Continuing Example 13.10: If we take the moisture of compressed pellets measured as a proportion instead of a percentage, we see the same correlation.

```
> moisture_prop<-moisture/100
> cor(filt.rate,moisture_prop,method = "pearson")
[1] 0.8943937
```

Continuing Example 13.12: If we take the math score as a percentage of total points instead of the raw score, we see the same correlation.

```
> score_perc<-Caschool$mathscr/750
cor(Caschool$mealpct,score_perc,method = "pearson")
[1] -0.8230145
```

Fact 4: When calculating the correlation r, it makes no difference what you call x and what you call y; the correlation will be the same. In other words, the correlation r **ignores** the distinction between which variable is the explanatory variable x and which one is the response variable y.

Continuing Example 13.10:

```
> cor(moisture,filt.rate,method = "pearson")
[1] 0.8943937
```

We get the same answer, $r \approx 0.89$. This represents a strong, positive linear relationship.

Continuing Example 13.12:

```
> cor(Caschool$mathscr,Caschool$mealpct,"pearson")
[1] -0.8230145
```

We get the same answer, $r \approx -0.82$. This represents a strong, negative linear relationship.

Fact 5: The correlation r only measures the strength and the direction of **straight-line (linear) relationships**.

- The correlation does not describe a curved relationship, no matter how strong that relationship is.

Example 13.15. Croissant (2016) provide data from US Census Tables and `www.measuringworth.com` about

$$x = \text{Real GDP}$$
$$y = \text{Persons per Family}$$

A scatterplot of the data is shown in Figure 13.3.9.

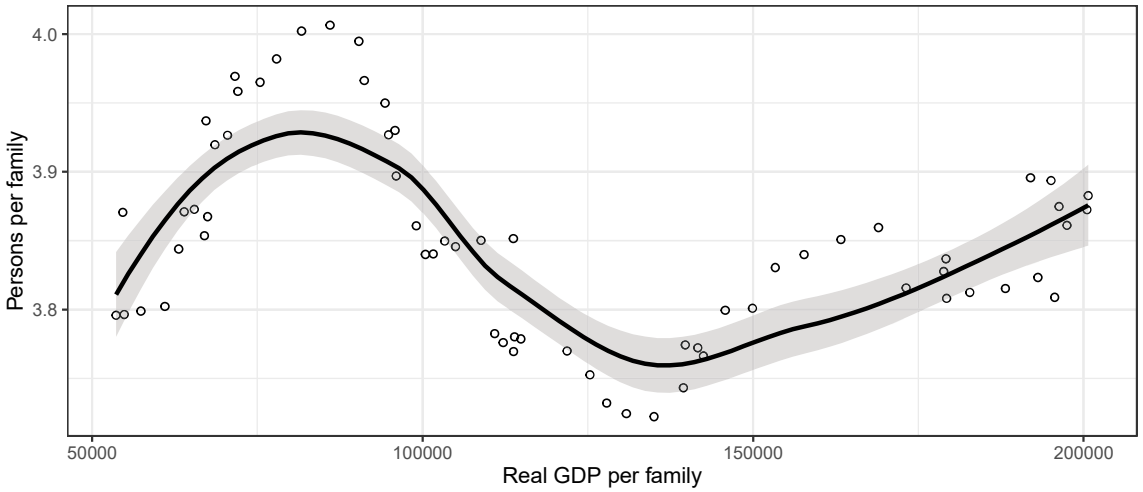

Figure 13.3.9: Scatterplot of Real GDP (x) and persons per family (y) for a sample of 66 years. A polynomial curve has been added to emphasize the relationship.

Discussion: Clearly, there is a relationship between Real GDP (x) and persons per family (y). However, the relationship is not a straight-line relationship. It is better described as a **curved** (i.e., polynomial) relationship. Because the relationship in Figure 13.3.9 is not a linear one, the correlation r here is useless. It does not describe curved relationships.

```
> library("Ecdat")
> data(incomeInequality)
> ggdat<-data.frame(realGDPperFamily=incomeInequality$realGDPperFamily,
+                   personsPerFamily=incomeInequality$personsPerFamily)
> ggplot(ggdat, aes(x=realGDPperFamily, y=personsPerFamily)) +
+   geom_point(shape=1)+
+   geom_smooth(alpha=0.25,color="black",method="loess")+
+   theme_bw()+
+   xlab("Real GDP per family")+
+   ylab("Persons per family")
> cor(incomeInequality$realGDPperFamily,incomeInequality$personsPerFamily,
+       method="pearson")
[1] -0.3080172
```

The correlation $r \approx -0.31$, which represents a weak linear relationship. However, the (curved) relationship between x and y here is pretty strong.

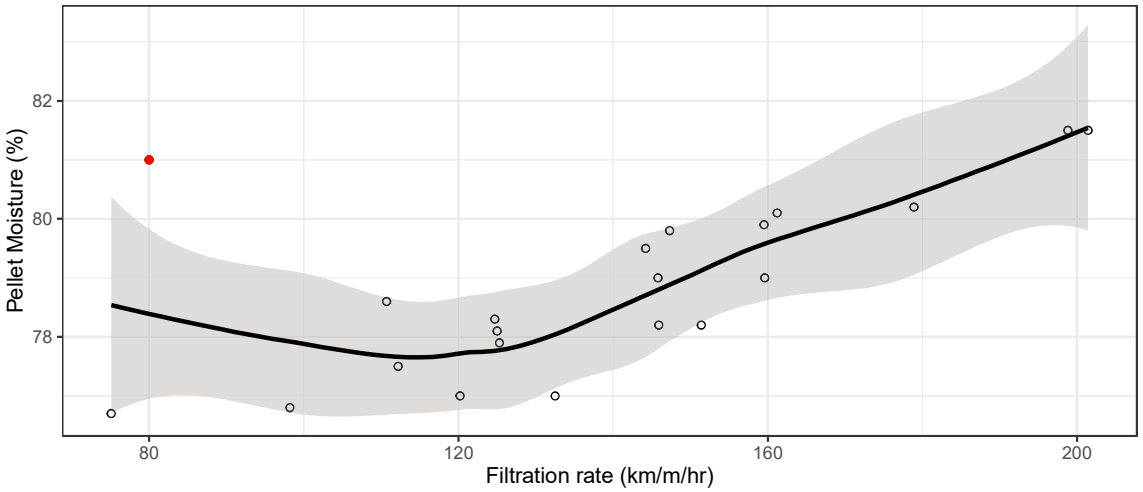

Figure 13.3.10: Scatterplot of filtration rate (x, measured in kg/m/hr) and pellet moisture (y, measured as a %) for a sample of $n = 20$ sewage specimens from Example 13.10 with outlying observation added (in red).

Fact 6: The value of the correlation r can be highly affected by **outliers**. Just one or two outliers might completely change one's impression of the strength of the relationship.

In Example 13.10, suppose we added just one observation in Figure 13.3.10. Note what this does to the value of r:

```
> cor(filt.rate,moisture,method = "pearson")
[1] 0.8943937
> filt.rate2<-c(125.3,98.2,201.4,147.3,145.9,124.7,112.2,120.2,161.2,178.9,
+               159.5,145.8,75.1,151.4,144.2,125,198.8,132.5,159.6,110.7,
+               80)#Outlier Added
> moisture2<-c(77.9,76.8,81.5,79.8,78.2,78.3,77.5,77,80.1,80.2,
+              79.9,79,76.7,78.2,79.5,78.1,81.5,77,79,78.6,
+              81)#Outlier Added
)
>  cor(filt.rate2,moisture2,method = "pearson")
[1] 0.6442389
```

The value of the correlation has changed from $r \approx 0.89$ to $r \approx 0.64$ after adding only one observation!

Important: Always plot your data first!

13.4 Nonparametric Alternatives

What if the relationship isn't linear, or the data is discrete?

Definition 13.16. Spearman's rho (r_s) is a rank-based correlation which determines the strength and direction of a monotone relationship between x and y. The calculation is done on the ranks of the data and not the data itself which lends itself to use for both continuous and discrete data.

With a sample of size n observations with unique x and y values, so there are no tied ranks, this correlation is computed as

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where d_i is the difference of the ranks of each observations' x and y values.
Remark: This measure of correlation is robust to outliers, unlike the Pearson correlation.

Later, a tie-corrected version was created

$$r_s^* = \frac{\left(\frac{n(n^2-1)}{6} - \sum_{i=1}^{n} d_i^2 - T_x - T_y\right)}{\sqrt{\left(\frac{n(n^2-1)}{6} - 2T_x\right)\left(\frac{n(n^2-1)}{6} - 2T_y\right)}},$$

where T_x is the number of tied ranks for x observations and T_y is the number of tied ranks for y observations. **Remark:** These are equivalent to the Pearson correlation on the ranks of X and Y.

Definition 13.17. Kendall's Tau-b (τ_b) is also a rank-based correlation which determines the strength and direction of a monotone relationship between x and y. Kendall's Tau-b, however, is appropriately used when there are tied ranks when the sample space of x and the sample space for y are the same size. Kendall's Tau-b is calculated by

$$\tau_b = \frac{n_c - n_d}{\sqrt{\left(\frac{n(n-1)}{2} - \sum_i \frac{T_{x_i}(T_{x_i}-1)}{2}\right)\left(\frac{n(n-1)}{2} - \sum_i \frac{T_{y_i}(T_{y_i}-1)}{2}\right)}},$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, T_{x_i} is the number of observations of the i$^{\text{th}}$ tied x observation, and T_{y_i} is the number of observations of the i$^{\text{th}}$ tied y observation.

Concordant and discordant pairs are decided by ordering x in ascending order and considering the ranks of the corresponding y values. For each observation, we count concordant or discordant pairs as the number of subsequent observations that agree; e.g., for an observation i we can define observation $j > i$ as **concordant** if

$$x_i < x_j \text{ and } y_i < y_j \quad \text{or} \quad x_i > x_j \text{ and } y_i > y_j,$$

and **discordant** if

$$x_i < x_j \text{ and } y_i > y_j \quad \text{or} \quad x_i > x_j \text{ and } y_i < y_j.$$

When the sample space of x and the sample space for y are finite and not the same size, we can use Kendall's Tau-c (τ_c) to assess a monotone relationship between x and y; e.g., X might take integer values from 1 to 5 while Y takes integer values from 1 to 100. We can calculate τ_c using

$$\tau_c = \frac{2(n_c - n_d)}{n^2 \frac{\min(r,c)-1}{\min(r,c)}},$$

where r is the number of unique observations of x and c is the number of unique observations of y.

Remark: This measure of correlation, like Spearman's rank correlation, is robust to outliers, but provides better approximations with small sample sizes. Kendall's Tau-b is usually smaller than r_s.

These alternative approaches extend Pearson's correlation beyond linear relationships. Instead of requiring a linearity, the Spearman's rank and the Kendall Tau correlation measure **monotone** relationships. If as x increases, y is non-decreasing this is a **monotone increasing** relationship. If as x increases, y is non-increasing this is a **monotone decreasing** relationship. This distinction is made in Figure 13.4.11 where we explore linear and monotone relationships between x and y.

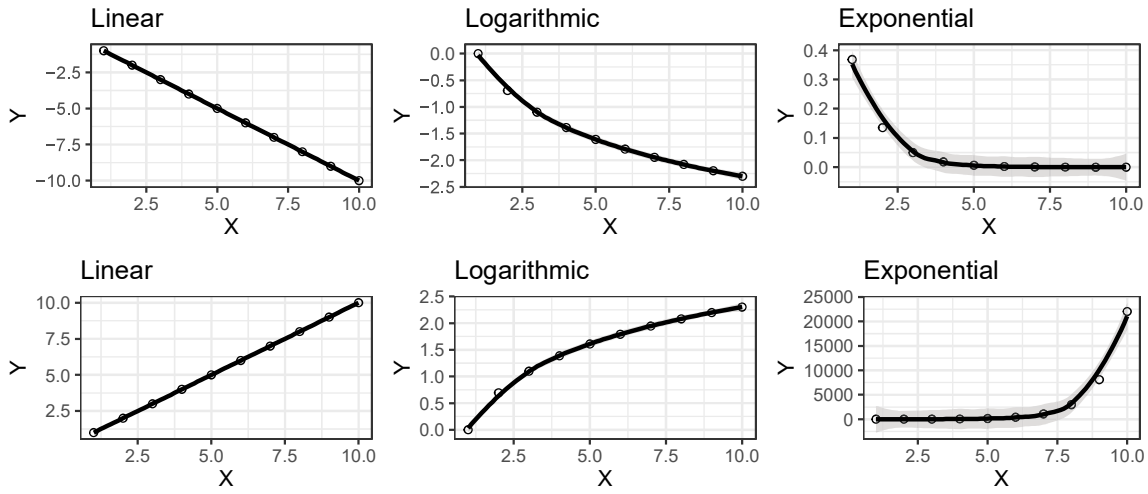

Figure 13.4.11: **Top left:** $r = -1$, $r_s = -1$, and $\tau_b = \tau_c = -1$; **top center:** $r = -0.9517$, $r_s = -1$, and $\tau_b = \tau_c = -1$; **top right:** $r = -0.7169$, $r_s = -1$, and $\tau_b = \tau_c = -1$. **Bottom left:** $r = 1$, $r_s = 1$, and $\tau_b = \tau_c = 1$; **bottom center:** $r = 0.9517$, $r_s = 1$, and $\tau_b = \tau_c = 1$; **bottom right:** $r = 0.7169$, $r_s = 1$, and $\tau_b = \tau_c = 1$.

Remark: Note that all the plots above have "perfect" relationships, but the Pearson correlation only captures the linear relationship whereas the Spearman's rank and Kendall's tau-b and tau-c correlations capture linear and the monotone relationships.

Example 13.18. Edmonson et al. (1979) explored the treatment of patients with advanced ovarian carcinoma using either cyclophosphamide alone (1 g/m2) or cyclophosphamide (500 mg/m2) plus adriamycin (40 mg/m2) by IV injection every 3 weeks. The combination regimen produced superior results to that of cyclophosphamide alone as depicted in the survival plot displayed in Figure 13.4.12. The data for this experiment can be accessed in in `R` as follows.

```
> library(survival)
> data(ovarian)
```

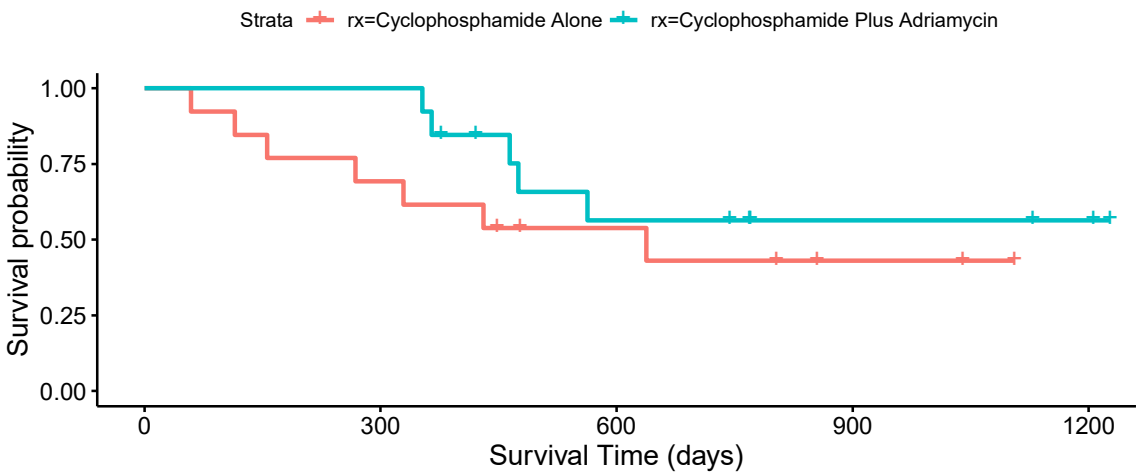

Figure 13.4.12: The survival curves for cyclophosphamide alone or cyclophosphamide plus adriamycin; '+' denotes experiment dropouts.

A possible concern about the analysis is how comparable the treatment groups are – is the superiority of the two-drug treatment due to the treatment regimen or is it due to the characteristics of the sample receiving the treatment? We explore this question by plotting the age of participants in each group as seen in Figure 13.4.13, and created with the following R code.

```
> ovarian.nodropouts<-ovarian[which(ovarian$fustat==0),]
> ggdat<-data.frame(rx=ovarian.nodropouts$rx,age=ovarian.nodropouts$age)
> ggplot(data=ggdat,aes(x=rx,y=age))+
+   geom_violin(fill="lightblue")+
+   geom_boxplot(width=0.25)+
+   theme_bw()+
+   xlab("Treatment")+
+   ylab("Age (years)")+
+   ggtitle("Age of Participants by Treatment Group",
+           subtitle="For non-dropouts")
```

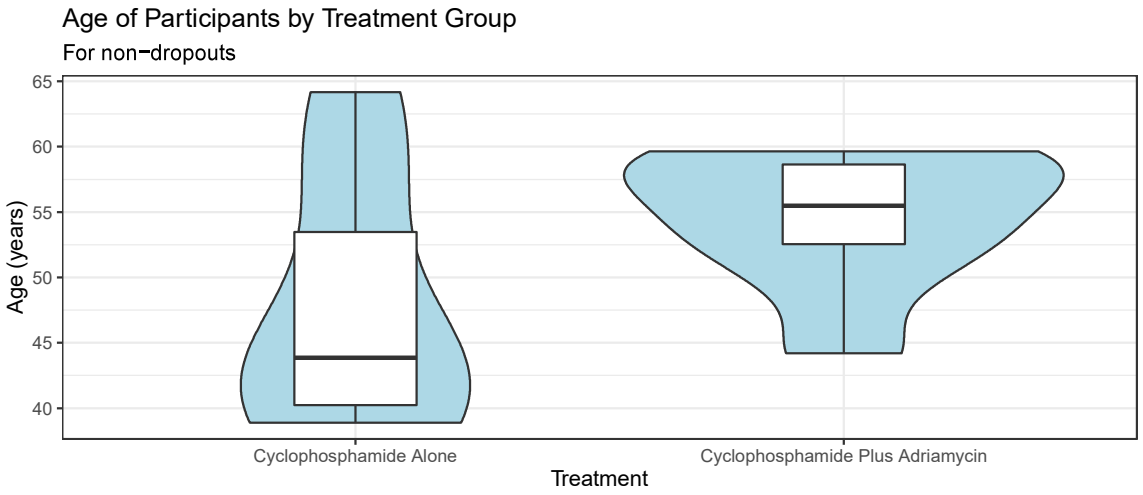

Figure 13.4.13: The age distributions for the cyclophosphamide alone or cyclophosphamide plus adriamycin treatment groups.

After consulting a graph about the ages of non-dropouts in each treatment group, we might be concerned about the comparisons we're making. The cyclophosphamide plus adriamycin treatment group appears to have older participants.

Research Question: Is the survival time dependent on age of the patient? If so, does this indicate that the superiority of the combination regimen under or over-stated?

A scatterplot of the $n = 14$ patients who died during the experiment can be seen in Figure 13.4.14 and created with the following R code.

```
> ggdat<-data.frame(age=ovarian.nodropouts$age,futime=ovarian.nodropouts$futime)
> ggplot(data=ggdat,aes(x=age, y=futime))+
+   geom_point(shape=1)+
+   geom_smooth(alpha=0.25,color="black",method="loess")+
+   theme_bw()+
+   xlab("Age (years)")+
+   ylab("Survival Time (days)")
```

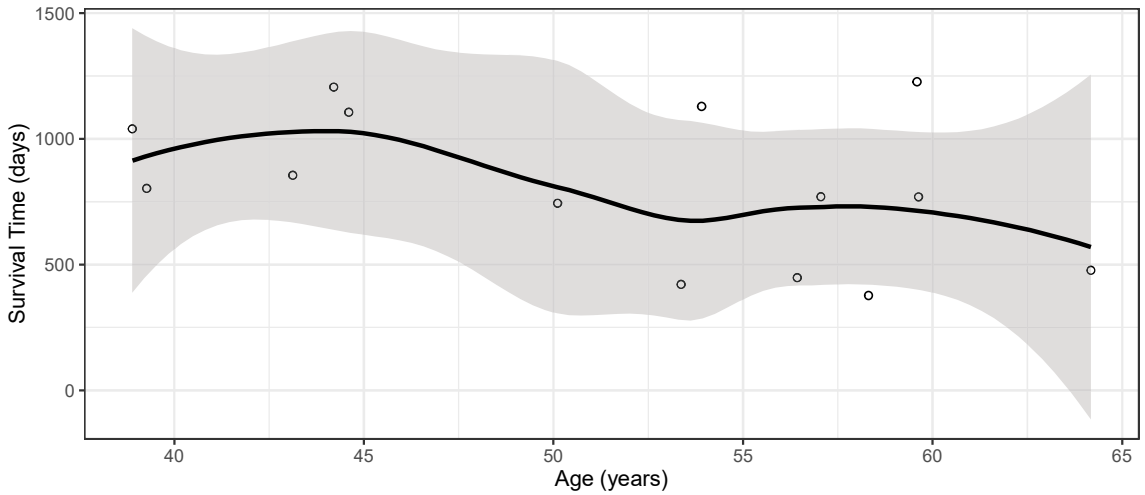

Figure 13.4.14: A scatterplot of observations of age and survival time for $n = 14$ non-dropouts.

To calculate the Spearman's rank correlation, we need to consider the ranks of the x and y observations; the key values for this calculation are in the table below.

Obs.	x	y	rank(x)	rank(y)	d_i	d_i^2
1	53.3644	421	7	2	5	25
2	56.4301	448	9	3	6	36
3	64.1753	477	14	4	10	100
4	50.1096	744	6	5	1	1
5	59.6301	769	13	6	7	49
6	57.0521	770	10	7	3	9
7	39.2712	803	2	8	-6	36
8	43.1233	855	3	9	-6	36
9	38.8932	1040	1	10	-9	81
10	44.6000	1106	5	11	-6	36
11	53.9068	1129	8	12	-4	16
12	44.2055	1206	4	13	-9	81
13	59.5890	1227	12	14	-2	4
14	58.3096	377	11	1	10	100
Total	722.6602	11372				610

Since there are no ties among ranks, the Spearman's rank correlation is calculated as

$$\begin{aligned} r_s &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \\ &= 1 - \frac{6(610)}{14(14^2-1)} \\ &= -0.3406593 \end{aligned}$$

This indicates a weak negative correlation between age and survival time. Of course, this is labor intensive and we will use `R` to calculate this value. The following code creates the table used to calculate the Spearman's rank by hand

```
> x<-ovarian.nodropouts$age
> y<-ovarian.nodropouts$futime
> rx<-rank(x=x)
> ry<-rank(x=y)
> di<-ry-rx
> di_squared<-di^2
> (rs_table<-cbind(x,y,rx,ry,di,di_squared))
           x    y rx ry  di di_squared
[1,] 53.3644  421  7  2  -5         25
[2,] 56.4301  448  9  3  -6         36
[3,] 64.1753  477 14  4 -10        100
[4,] 50.1096  744  6  5  -1          1
[5,] 59.6301  769 13  6  -7         49
[6,] 57.0521  770 10  7  -3          9
[7,] 39.2712  803  2  8   6         36
[8,] 43.1233  855  3  9   6         36
[9,] 38.8932 1040  1 10   9         81
[10,] 44.6000 1106  5 11   6         36
[11,] 53.9068 1129  8 12   4         16
[12,] 44.2055 1206  4 13   9         81
[13,] 59.5890 1227 12 14   2          4
[14,] 58.3096  377 11  1 -10        100
> sum(di_squared)
[1] 610
```

and the Spearman's rank can be asked for directly

```
> cor(x,y,method="spearman")
[1] -0.3406593
```

noting this value matches our by-hand calculation.

To calculate Kendall's tau-b, we need to consider the ranks of the x and y observations again, but in a slightly different way; the key values for this calculation are in the table below.

Obs.	x	y	rank(x)	rank(y)	Concordant	Discordant
9	38.8932	1040	1	10	4	9
7	39.2712	803	2	8	5	7
8	43.1233	855	3	9	4	7
12	44.2055	1206	4	13	1	9
10	44.6000	1106	5	11	2	7
4	50.1096	744	6	5	4	4
1	53.3644	421	7	2	6	1
11	53.9068	1129	8	12	1	5
2	56.4301	448	9	3	4	1
6	57.0521	770	10	7	1	3
14	58.3096	377	11	1	3	0
13	59.5890	1227	12	14	0	2
5	59.6301	769	13	6	0	1
3	64.1753	477	14	4	0	0
Total	722.6602	11372			35	56

The following code creates the table used to calculate the Kendall's Tau-b by hand

```
> x<-ovarian.nodropouts$age
> y<-ovarian.nodropouts$futime
> rx<-rank(x=x)
> ry<-rank(x=y)
> kendall.tab<-cbind(x,y,rx,ry)
> kendall.tab<-kendall.tab[order(x),]
> n<-nrow(kendall.tab)
> concordant<-c() #a place to save concordant pair counts
> discordant<-c() #a place to save discordant pair counts
> for(i in 1:n){
+   curr.x<-kendall.tab[i,3] #take the ith row's x rank
+   curr.y<-kendall.tab[i,4] #take the ith row's y rank
+   concordant_pairs<-c(which(kendall.tab[i:n,3]>curr.x&kendall.tab[i:n,4]>curr.y),
+                       which(kendall.tab[i:n,3]<curr.x&kendall.tab[i:n,4]<curr.y))
+   discordant_pairs<-c(which(kendall.tab[i:n,3]>curr.x&kendall.tab[i:n,4]<curr.y),
+                       which(kendall.tab[i:n,3]<curr.x&kendall.tab[i:n,4]>curr.y))
+   concordant<-c(concordant,length(concordant_pairs))
+   discordant<-c(discordant,length(discordant_pairs))
+ }
> (kendall.tab<-cbind(kendall.tab,concordant,discordant))
           x    y rx ry concordant discordant
 [1,] 38.8932 1040  1 10          4          9
 [2,] 39.2712  803  2  8          5          7
 [3,] 43.1233  855  3  9          4          7
 [4,] 44.2055 1206  4 13          1          9
 [5,] 44.6000 1106  5 11          2          7
 [6,] 50.1096  744  6  5          4          4
 [7,] 53.3644  421  7  2          6          1
 [8,] 53.9068 1129  8 12          1          5
 [9,] 56.4301  448  9  3          4          1
```

```
[10,] 57.0521  770 10  7          1          3
[11,] 58.3096  377 11  1          3          0
[12,] 59.5890 1227 12 14          0          2
[13,] 59.6301  769 13  6          0          1
[14,] 64.1753  477 14  4          0          0
> sum(concordant) #total concordant
[1] 35
> sum(discordant) #total discordant
[1] 56
> tab<-table(x,y)
> Tx<-rowSums(tab)
> Ty<-colSums(tab)
> (sum(concordant)-sum(discordant))/sqrt((((n*(n-1)/2)-sum(Tx*(Tx-1)/2))*((n*(n-1)/2)-sum(Ty*(Ty
[1] -0.2307692
```

We can then calculate Kendall's tau-b and tau-c as follows.

$$\begin{aligned}\tau_b &= \frac{n_c - n_d}{\sqrt{\left(\frac{n(n-1)}{2} - T_x\right)\left(\frac{n(n-1)}{2} - T_y\right)}} \\ &= \frac{35 - 56}{\sqrt{\left(\frac{14(14-1)}{2} - 0\right)\left(\frac{14(14-1)}{2} - 0\right)}} \\ &= \frac{-21}{\sqrt{8281}} \\ &= -0.2307692\end{aligned}$$

$$\begin{aligned}\tau_c &= \frac{2(n_c - n_d)}{n^2 \frac{\min(r,c)-1}{\min(r,c)}} \\ &= \frac{2(35 - 56)}{14^2 \frac{\min(14,14)-1}{\min(14,14)}} \\ &= \frac{2(35 - 56)}{14^2\left(\frac{13}{14}\right)} \\ &= -0.2307692\end{aligned}$$

This, like the Spearman's rank, indicates a weak negative correlation between age and survival time. We will use `R` to calculate this value.

Kendall's Tau-b and Tau-c can be asked for directly in `R` as follows. Note that the `StuartTauC()` function is rather slow for larger sample sizes.

```
> cor(x,y,method="kendall")
[1] -0.2307692
> library(DescTools)
> StuartTauC(x,y)
[1] -0.2307692
```

Remark: This value matches our by-hand calculations.

Note: When we decide to calculate Kendall's tau we have two choices, tau-b (τ_b) and tau-c (τ_c). These values were the same in this example because there were no ties. We will generally default to using Kendall's tau-b (τ_b), which is the value returned by the `cor()` function in `R`.

Research Question: Is the survival time dependent on age of the patient? If so, does this indicate that the superiority of the combination regimen under or over-stated?

Q: Can we conclusively state that the effect is under-stating the true superiority of the combined regimen?
A: No, it's possible that there is some confounding and that the combined regimen works well for older patients. It may be that the prognosis is significantly better for older patients with this treatment, despite the negative association between age and survival time.

Example 13.19. Data from the Australian Health Survey from 1977-1978, made available by **?**, gives information about 5,190 subjects. The association between X, the number of doctor visits in the past two weeks, and Y, the number of illnesses in the past two weeks, is of interest.

	Number of Doctor Visits (Y)									
Number of Illnesses (X)	0	1	2	3	4	5	6	7	8	9
0	1449	91	12	1	1	0	0	0	0	0
1	1317	247	40	9	10	7	3	3	2	0
2	688	188	50	8	4	1	2	3	2	0
3	376	128	26	6	3	1	1	1	0	0
4	174	70	20	1	5	0	3	1	0	0
5	137	58	26	5	1	0	3	4	1	1

Table 13.4.1: Cross tabulation of X and Y from the Australian Health Survey

Calculating Spearman's rho and Kendall's tau-b here would be infeasible as we would have to find concordant and discordant pairs for 5,190 observations. This would take a very long time; instead, we lean on `R`'s computational capabilities.

```
> library("AER")
> data("DoctorVisits")
> x<-DoctorVisits$illness
> y<-DoctorVisits$visits
> (tab<-table(x,y))
  y
x    0    1    2    3    4    5    6    7    8    9
0 1449   91   12    1    1    0    0    0    0    0
1 1317  247   40    9   10    7    3    3    2    0
2  688  188   50    8    4    1    2    3    2    0
3  376  128   26    6    3    1    1    1    0    0
4  174   70   20    1    5    0    3    1    0    0
5  137   58   26    5    1    0    3    4    1    1
> cor(x,y,method="spearman")
[1] 0.2626792
> cor(x,y,method="kendall")
[1] 0.2332083
> StuartTauC(x,y)
[1] 0.1423136
```

These values, $r_s = 0.2626792$ and $\tau_c = 0.1423136$ indicate a little to no positive correlation between the number of doctor visits and the number of illnesses.

Remark: We choose to report τ_c here due to the different scale of x and y, six versus ten unique observations. However, we do note that τ_b does not report an unreasonable value here.

Fact 1: If x and y have a monotone increasing relationship, then $r_s > 0$ and $\tau_b > 0$. If x and y have a monotone decreasing relationship, then $r_s < 0$ and $\tau_b < 0$. Note, $r_s = 1$ and $\tau_b = 1$ means that x and y have no monotone relationship.

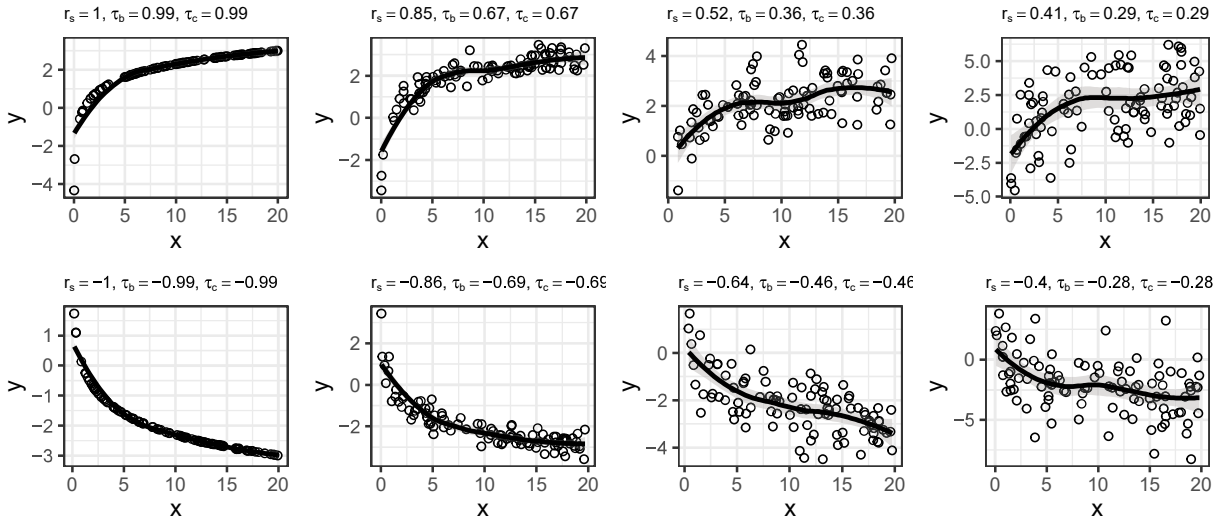

Figure 13.4.15: Scatterplot examples, $n = 100$. **Top row:** $r_s < 0$, $\tau_b < 0$, and $\tau_c < 0$. **Bottom row:** $r_s > 0$ and , $\tau_b > 0$, and $\tau_c > 0$.

Remark Kendall's Tau is usually smaller than Spearman's rho, however they're usually similar enough to lead to the same inference. Note that Kendall's Tau is generally preferred, particularly for small sample sizes.

Fact 2: Spearman's rho (r_s) and Kendall's tau (τ_b, τ_c) are always between -1 and 1; i.e.,

$$-1 \leq r_s \leq 1 \qquad -1 \leq \tau_b \leq 1 \qquad -1 \leq \tau_c \leq 1.$$

What happens at the endpoints?

- If $r_s = 1$, $\tau_b = 1$, $\tau_c = -1$, then all of the data follow an exact **monotone increasing** pattern
- If $r_s = -1$, $\tau_b = -1$, $\tau_c = -1$, then all of the data follow an exact **monotone increasing** pattern
- In either case, the relationship between x and y is perfectly **monotone**.

Fact 3: Similar to the Pearson correlation coefficient (r), Spearman's rho (r_s), Kendall's tau (τ_b, τ_c) are unitless and so they are retain their value even if we change the units of our data.

Continuing Example 13.18 If we take the survival time measured in weeks instead of days, we see the same Spearman's rho and Kendall's tau-b values.

```
> survival_in_weeks<-ovarian.nodropouts$futime/7
> cor(ovarian.nodropouts$age,ovarian.nodropouts$futime,method="spearman")
[1] -0.3406593
> cor(ovarian.nodropouts$age,survival_in_weeks,method="spearman")
[1] -0.3406593
> cor(ovarian.nodropouts$age,ovarian.nodropouts$futime,method="kendall")
[1] -0.2307692
> cor(ovarian.nodropouts$age,survival_in_weeks,method="kendall")
[1] -0.2307692
> StuartTauC(ovarian.nodropouts$age,ovarian.nodropouts$futime)
[1] -0.2307692
> StuartTauC(ovarian.nodropouts$age,survival_in_weeks)
[1] -0.2307692
```

Fact 4: When calculating Spearman's rho (r_s) or Kendall's tau (τ_b, τ_c), it makes no difference what you call x and what you call y. In other words, r_s and τ_b ignore the distinction between which variable is the explanatory variable x and which one is the response variable y.

Continuing Example 13.18

```
> cor(ovarian.nodropouts$age,ovarian.nodropouts$futime,method="spearman")
[1] -0.3406593
> cor(ovarian.nodropouts$futime,ovarian.nodropouts$age,method="spearman")
[1] -0.3406593
> cor(ovarian.nodropouts$age,ovarian.nodropouts$futime,method="kendall")
[1] -0.2307692
> cor(ovarian.nodropouts$futime,ovarian.nodropouts$age,method="kendall")
[1] -0.2307692
> StuartTauC(ovarian.nodropouts$age,ovarian.nodropouts$futime)
[1] -0.2307692
> StuartTauC(ovarian.nodropouts$futime,ovarian.nodropouts$age)
[1] -0.2307692
```

We get the same solution for both Spearman's rho (r_s) or Kendall's tau (τ_b, τ_c) and retain the interpretation of a very weak monotone decreasing relationship between age and survival time.

Continuing Example 13.19

```
> n<-nrow(DoctorVisits)
> x<-DoctorVisits$illness
> y<-DoctorVisits$visits
> cor(x,y,method="spearman")
[1] 0.2626792
> cor(y,x,method="spearman")
[1] 0.2626792
> cor(x,y,method="kendall")
[1] 0.2332083
> cor(y,x,method="kendall")
[1] 0.2332083
> StuartTauC(x,y)
```

```
[1] 0.1423136
> StuartTauC(y,x)
[1] 0.1423136
```

We get the same solution for both Spearman's rho (r_s) or Kendall's tau (τ_b, τ_c) and retain the interpretation of a little to no monotone increasing relationship between the number of illnesses and number of doctor visits.

Fact 5: Spearman's rho (r_s) or Kendall's tau (τ_b, τ_c) measure the strength and direction of **monotone** relationships.

Continuing Example 13.15 Recall the relationship of Real GDP per family and Persons per family depicted in Figure 13.4.16. The relationship is not a linear one nor a monotone one as the relationship changes directions several times. Like Pearson's correlation, Spearman's rho and Kendall's tau-b are useless here. They do not describe non-monotone relationships. Recall Pearson's correlation was calculated to be $r = -0.3080172$.

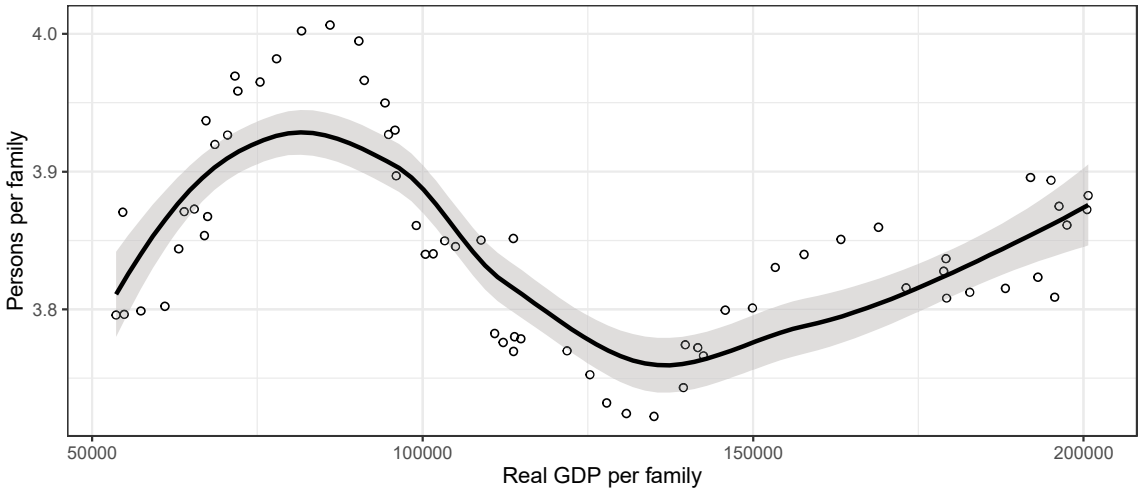

Figure 13.4.16: Scatterplot of Real GDP (x) and persons per family (y) for a sample of 66 years. A polynomial curve has been added to emphasize the relationship.

```
> library("Ecdat")
> data(Caschool)
> cor(Caschool$mealpct,Caschool$mathscr,method="spearman")
[1] -0.8254765
> cor(Caschool$mealpct,Caschool$mathscr,method="kendall")
[1] -0.638797
> StuartTauC(Caschool$mealpct,Caschool$mathscr)
[1] -0.6386605
```

Fact 6: The value of Spearmans's rho and Kendall's Tau are not as highly affected by **outliers** compared to Pearson's correlation. These approaches are robust to outliers due to using the ranks of the observations (Spearman) and whether or not the observations are concordant, instead of the raw values.

13.5 Inference about Pearson's Correlation

13.5.1 Hypothesis Test for Pearson's Correlation

As a motivating example, recall Example 13.10, where we discussed the relationship between the belt filtration rate of a new compression machine for processing sewage sludge (x) and the moisture content of compressed pellets (y).

Research Question: Is there enough evidence to suggest that there is a positive correlation between belt filtration rate and moisture content in compressed pellets from the evidence the sample provides?

Remark: Note that the question does not ask whether or not a higher belt filtration rate *causes* the moisture content to decrease. Correlation tools only yield results about the two items tendencies to happen together, not causation.

Hypotheses The **null hypothesis**, H_0, for a Pearson's correlation hypothesis test is that two continuous variables are, in fact, independent; e.g.,

$$H_0 : \rho = 0 \longrightarrow \text{the two continuous variables are independent}$$

The **alternative hypotheses**, H_a, for a Pearson's correlation hypothesis test is that two continuous variables are dependent. There are three ways dependence might be of interest; e.g.,

$$H_a : \rho < 0 \longrightarrow \text{the two continuous variables negatively correlated}$$
$$H_a : \rho > 0 \longrightarrow \text{the two continuous variables are positively correlated}$$
$$H_a : \rho > 0 \longrightarrow \text{the two continuous variables are dependent.}$$

For our researchers,

$$H_0 : \rho = 0 \longrightarrow \text{the belt filtration rate and pellet moisture are independent}$$
$$H_a : \rho > 0 \longrightarrow \text{higher belt filtration rates are associated with more pellet moisture}$$

Assumptions The assumptions for Pearson's correlation hypothesis test are

1. the two variables are continuous
2. there is a linear relationship – this can be evaluated via a scatterplot
3. there are no outliers
4. the two variables should have an approximate bivariate Gaussian distribution
   - we will check this using the mvnorm.etest() function from the "energy" package in R (Rizzo and Szekely, 2018)
   - this is important for the approximate sampling distribution to hold
5. pairs of observations are independent
6. sample is generalizable

Remark: Normality is not required to calculate Pearson correlation, but this assumption is necessary for the methodology described below.

For our researchers,

1. the two variables are continuous
2. there is a linear relationship as seen in Figure 13.3.4
3. there are no visible outliers in Figure 13.3.4
4. the two variables should be approximately normally distributed

   H_0 : data follows a bivariate normal versus H_a : data does not follow a bivariate normal

   ```
   > mvnorm.etest(x=cbind(filt.rate,moisture),R=1000)

   Energy test of multivariate normality: estimated parameters

   data:  x, sample size 20, dimension 2, replicates 1000
   E-statistic = 0.48953, p-value = 0.877
   ```

   Here, there is not enough evidence to suggest that the data is not normally distributed ($E = 0.48953, p = 0.877$), thus this assumption is met.
5. pairs of observations are reasonably independent
6. a random sample is generalizable to the population.

Remark: Here we are really putting faith in our researchers sampling design. We hope that the sampling design was random so that it would provide an independent and representative sample. Noting that these assumptions are ubiquitous hypothesis testing assumptions emphasizes that sampling design is paramount to any statistical analysis. This reminds us of a quote from Ronald Fisher.

> *"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."*

Test Statistic

Definition 13.20. The test statistic for Pearson's correlation hypothesis test is

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

For our researchers,

$$\begin{aligned} t^* &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ &= \frac{0.8943937\sqrt{20-2}}{\sqrt{1-0.8943937^2}} \\ &= 8.483693 \end{aligned}$$

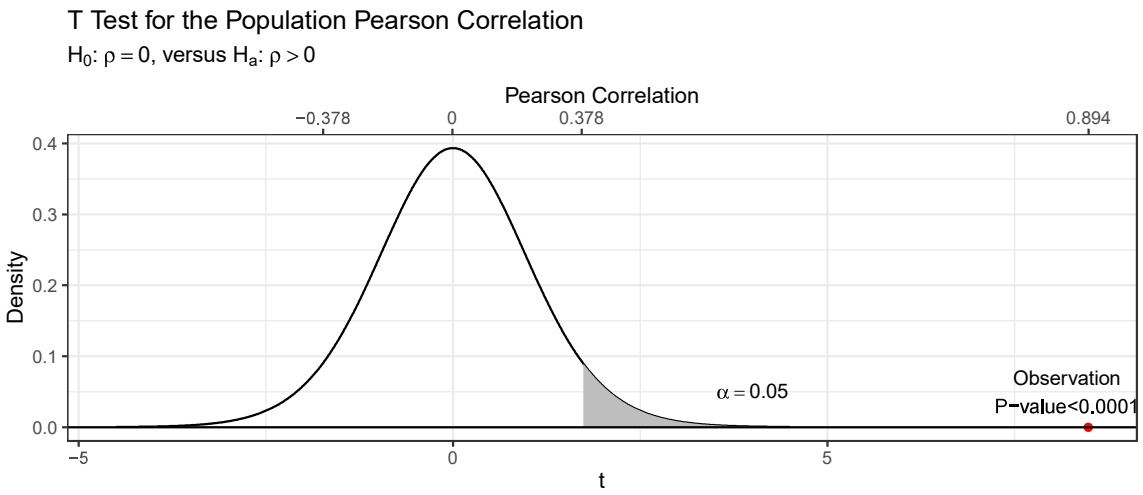

Figure 13.5.17: P-value as calculated for the hypothesis test about the Pearson correlation for Example 13.10.

P-value

The sampling distribution of t^* is given by the Student T distribution with $n-2$ degrees of freedom. We calculate the p-value using this sampling distribution to calculate probabilities about making observations more extreme than what we've observed.

For our researchers,

$$\begin{aligned} p\text{-value} &= P(\text{observed statistic or more extreme}|H_0 \text{ is true}) \\ &= P(r > 0.8943937|\rho = 0) \\ &= P(T_{20-2} > 8.483693) \\ &= 1 - P(T_{18} \leq 8.483693) \\ &= 0.0000000525854 \times 10^{-8} \end{aligned}$$

This value can be calculated in R as follows.

```
> n<-length(filt.rate)
> (r<-cor(filt.rate,moisture,method="pearson"))
[1] 0.8943937
> (t.obs<-r*sqrt((n-2)/(1-r^2)))
[1] 8.483693
> pt(test.stat,df=n-2,lower.tail=FALSE)
[1] 5.258601e-08
```

Decision Making Figure 13.5.17 shows that the observed data falls far in the rejection region for this test. There is significant evidence that the population Pearson correlation is greater than zero ($t = 8.48$, $p < 0.0001$).

Summary

In this section, we succinctly summarize the five steps for the hypothesis test about a population Pearson correlation. While it might be simple to check these tables and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Step One	$H_a : \rho < 0$	$H_a : \rho > 0$	$H_a : \rho \neq 0$
Step Two	Check Assumptions		
Step Three	$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$		
Step Four	$P(T_{n-2} < t^*)$	$P(T_{n-2} > t^*)$	$2P(T_{n-2} < -\lvert t^* \rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if t^* is in the rejection region		

For our researchers,

Step One	$H_0 : \rho = 0$ $H_a : \rho > 0$
Step Two	The variables are continuous. A scatterplot shows a linear relationship. There are no visible outliers. the two varaibles are approximately normally distributed. Sample is generalizable. Observations are independent.
Step Three	$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 8.483693$
Step Four	p-value $= P(T_{18} < t^*) < 0.0001$
Step Five	p-value$< 0.05 \longrightarrow$ Reject H_0.

This t test for the population Pearson correlation can be calculated in `R` as follows.

```
> cor.test(filt.rate,moisture_prop,method = "pearson",alternative = "greater")

Pearson's product-moment correlation

data:  filt.rate and moisture_prop
t = 8.4837, df = 18, p-value = 5.259e-08
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.7796716 1.0000000
sample estimates:
cor
0.8943937
```

In the case where the bivariate Gaussian assumption is not reasonable, we can conduct a **permutation test**. Like bootstrapping, a permutation test allows us to "peak" at the attributes of a sampling distribution by resampling the observed data. Specifically, we shuffle the order (or permute) observations without replacement, unlike bootstrapping.

Intuition: If the data are not correlated, then if we shuffle the observations the correlation won't change by very much and if the data are correlated, then if we shuffle the observations we should see a large change in the correlations.

This test can be calculated in R as follows.

```
> pearson.perm.test <- function(x,y,R=1000,alternative="two.sided",supplyCorrs=FALSE){
+   n = length(x) #length(x)=length(y)
+   if(n!=length(y)){
+     stop("\n 'x' and 'y' must have the same length")
+   }
+   r = cor(x,y,method="pearson") #sample correlation
+   #a place to store correlations on resampled
+   corrs<-rep(NA,R)
+   for(i in 1:R){
+     #randomly generate observations to sample
+     curr.samp<-sample(x=1:n,size=n,replace=FALSE)
+     #save correlation for this sample
+     corrs[i]<-cor(x,y[curr.samp],method="pearson")
+   }
+   #calculate p-values as the proportion of correlations
+   of resampled 'more extreme' than observed
+   if(alternative=="less"){
+     p.value = sum(corrs <= r) / R
+   } else if(alternative=="greater"){
+     p.value = sum(corrs >= r) / R
+   } else{
+     p.value = sum(abs(corrs) >= abs(r)) / R
+   }
+   if(supplyCorrs==TRUE){
+     list(r=r,p.value=p.value,corrs=corrs)
+   }else{
+     list(r=r,p.value=p.value)
+   }
+ }
> perm.test<-pearson.perm.test(filt.rate,moisture,supplyCorrs=TRUE)
> perm.test[1:2]
$r
[1] 0.8943937
$p.value
[1] 0
```

Figure 13.5.18, created with the R code below, shows that the observation is much larger than any of the observed Pearson correlations when we shuffle the data. Thus, there is significant evidence that the population Pearson correlation is greater than zero.

```
> ggdat<-data.frame(corrs=perm.test$corrs)
> ggplot(data=ggdat,aes(x=corrs))+
+   geom_histogram(aes(y=..density..),
+                  fill="lightblue",
```

```
+                   color="black",
+                   bins=20)+
+   geom_hline(yintercept=0)+
+   geom_vline(xintercept=perm.test$r,color="red")+
+   theme_bw()+
+   xlab("Pearson Correlations")+
+   ylab("Density")+
+   ggtitle("Permutation Test For Pearson's Correlation",
+           subtitle=bquote(H[0]*":"~rho==0*", versus "*H[a]*":"~rho>0))
```

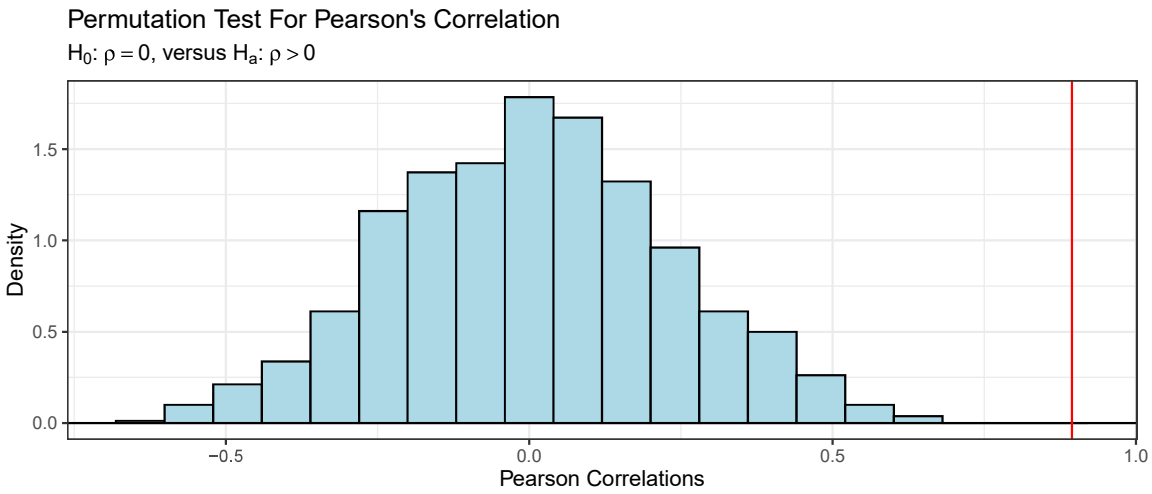

Figure 13.5.18: P-value as calculated for the permutation hypothesis test about the Pearson correlation for Example 13.10.

13.5.2 Confidence Interval for Pearson's Correlation

Using the assumption of bivariate normality, we transform r to z_r; e.g.,

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

The transformed variable z_r is approximately Gaussian with standard error $\sqrt{1/(n-3)}$. To find the $(1-\alpha)\times 100\%$ confidence interval for the population Pearson correlation we calculate upper and lower bounds

$$z_l = z_r - z_{1-\alpha/2}\sqrt{\left(\frac{1}{n-3}\right)}$$
$$z_u = z_r + z_{1-\alpha/2}\sqrt{\left(\frac{1}{n-3}\right)},$$

and transform back to r,

$$\begin{aligned} r_l &= \frac{e^{2z_l} - 1}{e^{2z_l} + 1} \\ u_u &= \frac{e^{2z_u} - 1}{e^{2z_u} + 1}. \end{aligned}$$

For our researchers, we can calculate a 95% confidence interval as follows. First, we transform r to z_r,

$$\begin{aligned} z_r &= \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \\ &= \frac{1}{2} \ln \left(\frac{1+0.8943937}{1-(0.8943937)} \right) \\ &= 1.443468 \end{aligned}$$

The upper and lower bounds are calculated as

$$\begin{aligned} z_l &= z_r - z_{1-\alpha/2} \sqrt{\left(\frac{1}{n-3} \right)} \\ &= 1.443468 - z_{0.975} \sqrt{\left(\frac{1}{20-3} \right)} \\ &= 1.443468 - 1.959964 \sqrt{\left(\frac{1}{17} \right)} \\ &= 0.968107 \\ z_u &= z_r + z_{1-\alpha/2} \sqrt{\left(\frac{1}{n-3} \right)}, \\ &= 1.443468 + z_{0.975} \sqrt{\left(\frac{1}{20-3} \right)} \\ &= 1.443468 + 1.959964 \sqrt{\left(\frac{1}{17} \right)} \\ &= 1.918829, \end{aligned}$$

which we can transform back to r

$$\begin{aligned} r_l &= \frac{e^{2z_l} - 1}{e^{2z_l} + 1} \\ &= \frac{e^{2(0.968107)} - 1}{e^{2(0.968107)} + 1} \\ &= 0.7478712 \\ r_u &= \frac{e^{2z_u} - 1}{e^{2z_u} + 1} \\ &= \frac{e^{2(1.918829)} - 1}{e^{2(1.918829)} + 1} \\ &= 0.9578207. \end{aligned}$$

This indicates that we are 95% confident the true population Pearson correlation, ρ is between 0.7478712 and 0.9578207, the linear association is strong and positive.

The confidence interval for the population Pearson correlation can be calculated in R as follows.

```
> cor.test(filt.rate,moisture,method="pearson")

Pearson's product-moment correlation

data:  filt.rate and moisture
t = 8.4837, df = 18, p-value = 1.052e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7478712 0.9578207
sample estimates:
cor
0.8943937
```

In the case where the bivariate Gaussian assumption is not reasonable, we can construct a bootstrap confidence interval for the population Pearson correlation. This can be done with the following R code.

```
> library("boot")
> boot.pearson<-function(data,indices){
+   d<-data[indices,]
+   return(cor(d$filt.rate,d$moisture,method="pearson"))
+ }
> boot.dat<-data.frame(filt.rate=filt.rate,moisture=moisture)
> boot<-boot(boot.dat,R=1000,statistic=boot.pearson)
> boot.ci(boot.out=boot,conf=0.95,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.7489,  0.9612 )
Calculations and Intervals on Original Scale
```

This indicates that we are 95% confident the true population Pearson correlation, ρ is between 0.7489 and 0.9612, the linear association is strong and positive.

13.6 Inference about Spearman's Rank Correlation

13.6.1 Hypothesis Test for Spearman's Rank Correlation

As a motivating example, recall Example 13.10, where we discussed the relationship between the belt filtration rate of a new compression machine for processing sewage sludge (x) and the moisture content of compressed pellets (y).

Research Question: Is there enough evidence to suggest that there is a positive correlation between belt filtration rate and moisture content in compressed pellets from the evidence the sample provides?

Remark: Note that the question does not ask whether or not a higher belt filtration rate *causes* the moisture content to decrease. Correlation tools only yield results about the two items tendencies to happen together, not causation.

Hypotheses The **null hypothesis**, H_0, for a Spearmans's correlation hypothesis test is that two quantitative variables are, in fact, independent; e.g.,

$$H_0 : \rho_s = 0 \longrightarrow \text{the two quantitative variables are independent}$$

The **alternative hypotheses**, H_a, for a Spearman's correlation hypothesis test is that two quantitative variables are dependent. There are three ways dependence might be of interest; e.g.,

$$H_a : \rho_s < 0 \longrightarrow \text{the two quantitative variables negatively correlated}$$
$$H_a : \rho_s > 0 \longrightarrow \text{the two quantitative variables are positively correlated}$$
$$H_a : \rho_s > 0 \longrightarrow \text{the two quantitative variables are dependent.}$$

For our researchers,

$$H_0 : \rho_s = 0 \longrightarrow \text{the belt filtration rate and pellet moisture are independent}$$
$$H_a : \rho_s > 0 \longrightarrow \text{higher belt filtration rates are associated with more pellet moisture}$$

Assumptions The assumptions for Spearmans's correlation hypothesis test are

1. the two variables are quantitative
2. there is a monotone relationship – this can be evaluated via a scatterplot
3. the two variables should have an approximate bivariate Gaussian distribution
    - we will check this using the mvnorm.etest() function from the "energy" package in R (Rizzo and Szekely, 2018)
    - this is important for the approximate sampling distribution to hold
4. pairs of observations are independent
5. sample is generalizable

Remark: Normality is not required to calculate Spearman's correlation, but this assumption is necessary for the methodology described below.

For our researchers,

1. the two variables are quantitative
2. there is a linear relationship as seen in Figure 13.3.4
3. there are no visible outliers in Figure 13.3.4
4. the two variables should be approximately normally distributed

   H_0 : data follows a bivariate normal versus H_a : data does not follow a bivariate normal

   ```
   > mvnorm.etest(x=cbind(filt.rate,moisture),R=1000)

   Energy test of multivariate normality: estimated parameters

   data:  x, sample size 20, dimension 2, replicates 1000
   E-statistic = 0.48953, p-value = 0.877
   ```

   Here, there is not enough evidence to suggest that the data is not normally distributed ($E = 0.48953, p = 0.877$), thus this assumption is met.
5. pairs of observations are reasonably independent
6. a random sample is generalizable to the population.

Test Statistic

Definition 13.21. The test statistic for Spearman's correlation hypothesis test is

$$t^* = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}},$$

or, with a continuity correction,

$$t^* = \frac{\tilde{r_s}\sqrt{n-2}}{\sqrt{1-\tilde{r_s}^2}},$$

where

$$\tilde{r_s} = 1 - \frac{\frac{(n^3-n)(1-r_s)}{6}}{\frac{n(n^2-1)}{6}+1}.$$

For our researchers, we calculate the adjusted correlation as

$$\tilde{r_s} = 1 - \frac{\frac{(n^3-n)(1-r_s)}{6}}{\frac{n(n^2-1)}{6}+1} = 1 - \frac{\frac{(20^3-20)(1-0.8524106)}{6}}{\frac{20(20^2-1)}{6}+1} = 1 - \frac{\frac{(20^3-20)(1-0.8524106)}{6}}{\frac{20(20^2-1)}{6}+1} = 1 - 0.1474785 = 0.8525215$$

and the corrected test statistic as

$$\begin{aligned} t^* &= \frac{\tilde{r_s}\sqrt{n-2}}{\sqrt{1-\tilde{r_s}^2}} \\ &= \frac{0.8525215\sqrt{20-2}}{\sqrt{1-0.8525215^2}} \\ &= 6.919833. \end{aligned}$$

P-value
The sampling distribution of t^* is given by the Student T distribution with $n-2$ degrees of freedom. We calculate the p-value using this sampling distribution to calculate probabilities about making observations more extreme than what we've observed.

For our researchers,

$$\begin{aligned} p\text{-value} &= P(\text{observed statistic or more extreme}|H_0 \text{ is true}) \\ &= P(r_s > 0.8527343|\rho_s = 0) \\ &= P(T_{20-2} > 6.919833) \\ &= 1 - P(T_{18} \leq 6.919833) \\ &= 9.059894 \times 10^{-07} \end{aligned}$$

These values can be calculated in R as follows.

```
> (r<-cor(filt.rate,moisture,method="spearman"))
[1] 0.8524106
> r.adjusted<-1-((n^3-n)*(1-r)/6)/((n*(n^2-1))/6+1)
> t.obs<-(temp/sqrt((1-temp^2)/(n-2)))
> pt(t.obs,df=n-2,lower.tail=FALSE)
[1] 9.059894e-07
```

Decision Making Figure 13.6.19 shows that the observed data falls far in the rejection region for this test.

```
> alpha<-0.05
> ggdat<-data.frame(t=seq(-4.5,4.5,0.01),
+                   f=dt(x=seq(-4.5,4.5,0.01),df=n-2))
> ggdat.highlight<-data.frame(x=t.obs,y=0)
> axis.labels<-round(c(0,qt(0.95,df=n-2)/sqrt(qt(0.95,df=n-2)^2 + (n-2)),r),3)
> ggplot(data=ggdat,aes(x=t,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,t>=qt(1-alpha,df=n-2)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,t>=abs(t.obs)),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("t")+
+   ylab("Density")+
+   ggtitle("T Test for the Population Spearman's rank",
+           subtitle=bquote(H[0]*":"~rho[s]==0*", versus "*H[a]*":"~rho[s]>0))+
+   annotate("text",x=4,y=0.05,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text",x=6.5,y=0.05,label="Observation\n P-value<0.0001",size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+        breaks=c(0,qt(1-alpha,n-1),t.obs),
+        labels=axis.labels,name="Spearman's rank"))
```

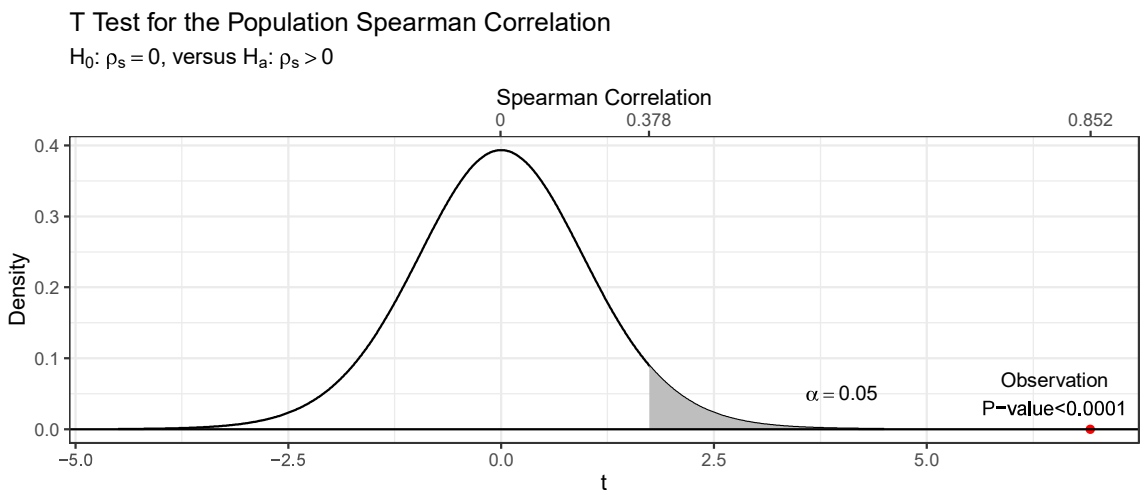

Figure 13.6.19: P-value as calculated for the hypothesis test about the Spearman's rank for Example 13.10.

There is significant evidence that the population Spearman's rank is greater than zero ($t = 6.919833$, $p < 0.0001$).

Summary In this section, we succinctly summarize the five steps for the hypothesis test about a population Spearman's rank. While it might be simple to check these tables and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Step One	$H_a : \rho_s < 0$	$H_a : \rho_s > 0$	$H_a : \rho_s \neq 0$
Step Two	Check Assumptions		
Step Three	$t^* = \frac{\tilde{r_s}\sqrt{n-2}}{\sqrt{1-\tilde{r_s}^2}}$		
Step Four	$P(T_{n-2} < t^*)$	$P(T_{n-2} > t^*)$	$2P(T_{n-2} < -\lvert t^*\rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if t^* is in the rejection region		

For our researchers,

Step One	$H_0 : \rho = 0$ $H_a : \rho > 0$
Step Two	The variables are quantitativre. A scatterplot shows a monotone relationship. There are no visible outliers. the two varaibles are approximately normally distributed. Sample is generalizable. Observations are independent.
Step Three	$t^* = \frac{\tilde{r_s}\sqrt{n-2}}{\sqrt{1-\tilde{r_s}^2}} = 6.919833$
Step Four	p-value $= P(T_{18} < t^*) < 0.0001$
Step Five	p-value$< 0.05 \longrightarrow$ Reject H_0.

This t test for the population Spearman's rank can be calculated in R as follows.

```
> cor.test(filt.rate,moisture,method="spearman",alternative="greater",continuity=TRUE)

Spearman's rank correlation rho

data:  filt.rate and moisture
S = 196.29, p-value = 9.06e-07
alternative hypothesis: true rho is greater than 0
sample estimates:
rho
0.8524106

Warning message:
In cor.test.default(filt.rate, moisture, method = "spearman", alternative = "greater") :
Cannot compute exact p-value with ties
```

In the case where the bivariate Gaussian assumption is not reasonable, we can conduct a **permutation test**. Like bootstrapping, a permutation test allows us to "peak" at the attributes of a sampling distribution by resampling the observed data. Specifically, we shuffle the order (or permute) observations without replacement, unlike bootstrapping.

Intuition: If the data are not correlated, then if we shuffle the observations the correlation won't change by very much and if the data are correlated, then if we shuffle the observations we should see a large change in the correlations.

This test can be calculated in R as follows.

```
> spearman.perm.test <- function(x,y,R=1000,alternative="two.sided",supplyCorrs=FALSE){
+   n = length(x) #length(x)=length(y)
+   if(n!=length(y)){
+     stop("\n 'x' and 'y' must have the same length")
+   }
+   r = cor(x,y,method="spearman") #sample correlation
+   #a place to store correlations on resampled
+   corrs<-rep(NA,R)
+   for(i in 1:R){
+     #randomly generate observations to sample
+     curr.samp<-sample(x=1:n,size=n,replace=FALSE)
+     #save correlation for this sample
+     corrs[i]<-cor(x,y[curr.samp],method="spearman")
+   }
+   #calculate p-values as the proportion of correlations
+   of resampled 'more extreme' than observed
+   if(alternative=="less"){
+     p.value = sum(corrs <= r) / R
+   } else if(alternative=="greater"){
+     p.value = sum(corrs >= r) / R
+   } else{
+     p.value = sum(abs(corrs) >= abs(r)) / R
```

```
+   }
+   if(supplyCorrs==TRUE){
+     list(r=r,p.value=p.value,corrs=corrs)
+   }else{
+     list(r=r,p.value=p.value)
+   }
+ }
> perm.test<-pearson.perm.test(filt.rate,moisture,supplyCorrs=TRUE)
> perm.test[1:2]
$r
[1] 0.8524106
$p.value
[1] 0
```

Figure 13.6.20, created with the R code below, shows that the observation is much larger than any of the observed Spearman's rank correlations when we shuffle the data. Thus, there is significant evidence that the population Spearman's rank correlation is greater than zero.

```
> ggdat<-data.frame(corrs=perm.test$corrs)
> ggplot(data=ggdat,aes(x=corrs))+
+   geom_histogram(aes(y=..density..),
+                  fill="lightblue",
+                  color="black",
+                  bins=20)+
+   geom_hline(yintercept=0)+
+   geom_vline(xintercept=perm.test$r,color="red")+
+   theme_bw()+
+   xlab("Spearman's ranks")+
+   ylab("Density")+
+   ggtitle("Permutation Test For Spearman's Correlation",
+           subtitle=bquote(H[0]*":"~rho[s]==0*", versus "*H[a]*":"~rho[s]>0))
```

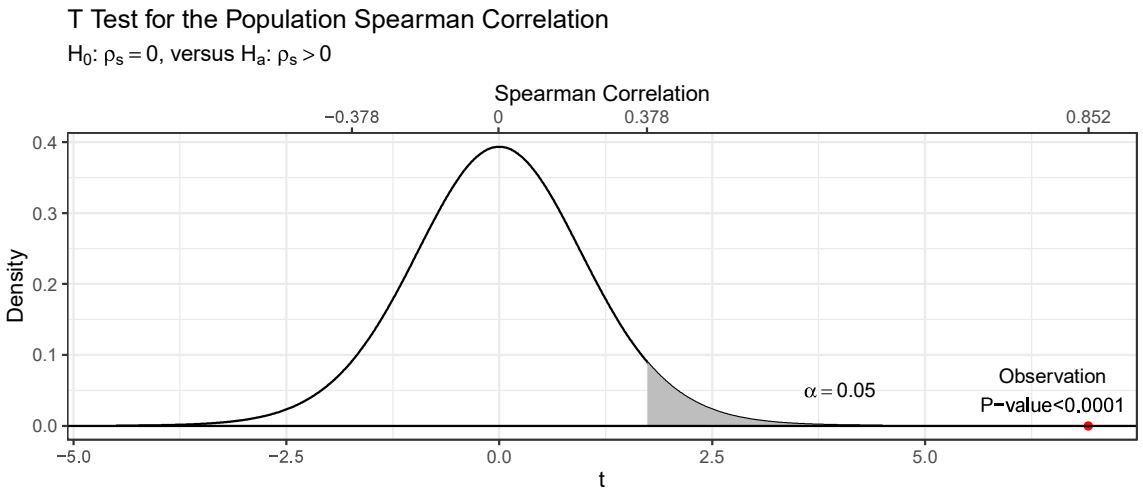

Figure 13.6.20: P-value as calculated for the permutation hypothesis test about the Spearman's rank correlation for Example 13.10.

13.6.2 Confidence Interval for Spearman's Rank Correlation

Using the assumption of bivariate normality, we transform r_s to z_r; e.g.,

$$z_r = \frac{1}{2}\ln\left(\frac{1+r_s}{1-r_s}\right).$$

The transformed variable z_r is approximately Gaussian with standard error $\sqrt{\frac{1+\frac{r_s^2}{2}}{n-3}}$. To find the $(1-\alpha)\times 100\%$ confidence interval for the population Spearman's rank correlation we calculate upper and lower bounds

$$z_l = z_r - z_{1-\alpha/2}\sqrt{\frac{1+\frac{r^2}{2}}{n-3}}$$
$$z_u = z_r + z_{1-\alpha/2}\sqrt{\frac{1+\frac{r^2}{2}}{n-3}},$$

and transform back to r,

$$r_l = \frac{e^{2z_l}-1}{e^{2z_l}+1}$$
$$u_u = \frac{e^{2z_u}-1}{e^{2z_u}+1}.$$

For our researchers, we can calculate a 95% confidence interval as follows. First, we transform r_s to z_r,

$$\begin{aligned} z_r &= \frac{1}{2}\ln\left(\frac{1+r_s}{1-r_s}\right) \\ &= \frac{1}{2}\ln\left(\frac{1+0.8524106}{1-(0.8524106)}\right) \\ &= 1.264905 \end{aligned}$$

The upper and lower bounds are calculated as

$$\begin{aligned}
z_l &= z_r - z_{1-\alpha/2}\sqrt{\frac{1+\frac{r_s^2}{2}}{n-3}} \\
&= 1.264905 - z_{0.975}\sqrt{\frac{1+\frac{0.8524106^2}{2}}{20-3}} \\
&= 1.264905 - 1.959964\sqrt{\frac{1+\frac{0.7266038}{2}}{17}} \\
&= 0.8572496 \\
z_u &= z_r + z_{1-\alpha/2}\sqrt{\frac{1+\frac{r_s^2}{2}}{n-3}}, \\
&= 1.264905 + z_{0.975}\sqrt{\frac{1+\frac{0.8524106^2}{2}}{20-3}} \\
&= 1.264905 + 1.959964\sqrt{\frac{1+\frac{0.7266038}{2}}{17}} \\
&= 1.819939,
\end{aligned}$$

which we can transform back to r

$$\begin{aligned}
r_l &= \frac{e^{2z_l}-1}{e^{2z_l}+1} \\
&= \frac{e^{2(0.8572496)}-1}{e^{2(0.8572496)}+1} \\
&= 0.6948379 \\
r_u &= \frac{e^{2z_u}-1}{e^{2z_u}+1} \\
&= \frac{e^{2(1.819939)}-1}{e^{2(1.819939)}+1} \\
&= 0.9488323.
\end{aligned}$$

This indicates that we are 95% confident the true population Spearman's rank, ρ is between 0.6948379 and 0.9488323, the monotone association is strong and positive.

The confidence interval for the population Pearson correlation can be calculated in `R` as follows.

```
> cor.test(filt.rate,moisture,method="spearman")

Spearman's rank correlation rho

data:  filt.rate and moisture
S = 196.29, p-value = 1.824e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.8524106
```

```
Warning message:
In cor.test.default(filt.rate, moisture, method = "spearman") :
Cannot compute exact p-value with ties
```

In the case where the bivariate Gaussian assumption is not reasonable, we can construct a bootstrap confidence interval for the population Pearson correlation. This can be done with the following R code.

```
> library("boot")
> boot.spearman<-function(data,indices){
+   d<-data[indices,]
+   return(cor(d$filt.rate,d$moisture,method="spearman"))
+ }
> boot.dat<-data.frame(filt.rate=filt.rate,moisture=moisture)
> boot<-boot(boot.dat,R=1000,statistic=boot.spearman)
> boot.ci(boot.out=boot,conf=0.95,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot, conf = 0.95, type = "perc")

Intervals :
Level     Percentile
95%   ( 0.5708,  0.9678 )
Calculations and Intervals on Original Scale
```

This indicates that we are 95% confident the true population Spearman's rank correlation, ρ_s is between 0.5708 and 0.9678, the monotone association is strong and positive.

13.7 Inference about Kendall's Tau-b Correlation

13.7.1 Hypothesis Test for Kendall's Tau-b Correlation

As a motivating example, recall Example 13.10, where we discussed the relationship between the belt filtration rate of a new compression machine for processing sewage sludge (x) and the moisture content of compressed pellets (y).

Research Question: Is there enough evidence to suggest that there is a positive correlation between belt filtration rate and moisture content in compressed pellets from the evidence the sample provides?

Remark: Note that the question does not ask whether or not a higher belt filtration rate *causes* the moisture content to decrease. Correlation tools only yield results about the two items tendencies to happen together, not causation.

Hypotheses The **null hypothesis**, H_0, for a Kendall's tau-b correlation hypothesis test is that two quantitative variables are, in fact, independent; e.g.,

$$H_0 : \rho_\tau = 0 \longrightarrow \text{the two quantitative variables are independent}$$

The **alternative hypotheses**, H_a, for a Kendall's tau-b correlation hypothesis test is that two quantitative variables are dependent. There are three ways dependence might be of interest; e.g.,

$$H_a : \rho_\tau < 0 \longrightarrow \text{the two quantitative variables negatively correlated}$$
$$H_a : \rho_\tau > 0 \longrightarrow \text{the two quantitative variables are positively correlated}$$
$$H_a : \rho_\tau > 0 \longrightarrow \text{the two quantitative variables are dependent.}$$

For our researchers,

$$H_0 : \rho_\tau = 0 \longrightarrow \text{the belt filtration rate and pellet moisture are independent}$$
$$H_a : \rho_\tau > 0 \longrightarrow \text{higher belt filtration rates are associated with more pellet moisture}$$

Assumptions The assumptions for Kendall's tau-b correlation hypothesis test are

1. the two variables are quantitative
2. there is a monotone relationship – this can be evaluated via a scatterplot
3. the two variables should have an approximate bivariate Gaussian distribution
   - we will check this using the mvnorm.etest() function from the "energy" package in R (Rizzo and Szekely, 2018)
   - this approach is robust to the bivariate normality assumption
4. pairs of observations are independent
5. sample is generalizable

For our researchers,

1. the two variables are quantitative
2. there is a linear relationship as seen in Figure 13.3.4
3. there are no visible outliers in Figure 13.3.4
4. the two variables should be approximately normally distributed

   H_0 : data follows a bivariate normal versus H_a : data does not follow a bivariate normal

   ```
   > mvnorm.etest(x=cbind(filt.rate,moisture),R=1000)

   Energy test of multivariate normality: estimated parameters

   data:  x, sample size 20, dimension 2, replicates 1000
   E-statistic = 0.48953, p-value = 0.877
   ```

   Here, there is not enough evidence to suggest that the data is not normally distributed ($E = 0.48953, p = 0.877$), thus this assumption is met.

5. pairs of observations are reasonably independent

6. a random sample is generalizable to the population.

Test Statistic

Definition 13.22. The test statistic for Kendall's tau-b correlation hypothesis test is

$$z = \frac{n_c - n_d}{\sigma},$$

or, with a continuity correction

$$z_c = \frac{n_c - n_d + c}{\sigma},$$

where

$$c = \begin{cases} -1 & n_c - n_d > 1 \\ 1 & n_c - n_d < 1 \end{cases}$$

$$==$$

$$\sigma = \frac{(n^2-n)(2n+5) - \sum_i T_{x_i}(T_{x_i}-1)(2T_{x_i}+5) - \sum_i T_{y_i}(T_{y_i}-1)(2T_{y_i}+5)}{18} + \frac{\left(\sum_i T_{x_i}(T_{x_i}-1)(T_{x_i}-2)\right)\left(\sum_i T_{y_i}(T_{y_i}-1)(T_{y_i}-2)\right)}{9(n^2-n)(n-2)} + \frac{\left(\sum_i T_{x_i}(T_{x_i}-1)\right)\left(\sum_i T_{y_i}(T_{y_i}-1)\right)}{2(n^2-n)}$$

For our researchers, we calculate the corrected statistic as

$$\begin{aligned} z_c^* &= \frac{n_c - n_d + c}{\sigma} \\ &= \frac{159 - 27 + (-1)}{30.75711} \\ &= \frac{131}{30.75711} \\ &= 4.259177 \end{aligned}$$

using the values calculated in R as follows.

```
> x<-filt.rate
> y<-moisture
> rx<-rank(x=x)
> ry<-rank(x=y)
> kendall.tab<-cbind(x,y,rx,ry)
> kendall.tab<-kendall.tab[order(x),]
> n<-nrow(kendall.tab)
> concordant<-c() #a place to save concordant pair counts
> discordant<-c() #a place to save discordant pair counts
> for(i in 1:n){ #for each row
+   curr.x<-kendall.tab[i,3] #take the ith row's x rank
+   curr.y<-kendall.tab[i,4] #take the ith row's y rank
+   concordant_pairs<-c(which(kendall.tab[i:n,3]>curr.x&kendall.tab[i:n,4]>curr.y),
+                       which(kendall.tab[i:n,3]<curr.x&kendall.tab[i:n,4]<curr.y))
+   discordant_pairs<-c(which(kendall.tab[i:n,3]>curr.x&kendall.tab[i:n,4]<curr.y),
```

```
+                         which(kendall.tab[i:n,3]<curr.x&kendall.tab[i:n,4]>curr.y))
+   concordant<-c(concordant,length(concordant_pairs))
+   discordant<-c(discordant,length(discordant_pairs))
+ }
> kendall.tab<-cbind(kendall.tab,concordant,discordant)
> sum(concordant) #total concordant
[1] 159
> sum(discordant) #total discordant
[1] 27
> tab<-table(x,y)
> Tx<-rowSums(tab)
> Ty<-colSums(tab)
> (sum(concordant)-sum(discordant))/
+    sqrt((((n*(n-1)/2)-sum(Tx*(Tx-1)/2))*((n*(n-1)/2)-sum(Ty*(Ty-1)/2))))
[1] 0.7021674
> sig2<-(1/18)*((n^2-n)*(2*n+5)-sum(Tx*(Tx-1)*(2*Tx+5))-sum(Ty*(Ty-1)*(2*Ty+5)))+
+      (sum(Tx*(Tx-1)*(Tx-2))*sum(Ty*(Ty-1)*(Ty-2)))/(9*(n^2-n)*(n-2))+
+      (sum(Tx*(Tx-1))*sum(Ty*(Ty-1)))/(2*(n^2-n))
> sqrt(sig2)
[1] 30.75711
```

P-value
The sampling distribution of z^* is given by the standard Gaussian distribution. We calculate the p-value using this sampling distribution to calculate probabilities about making observations more extreme than what we've observed.

For our researchers,

$$\begin{aligned} p\text{-value} &= P(\text{observed statistic or more extreme}|H_0 \text{ is true}) \\ &= P(r > 0.7021674|\rho_\tau = 0) \\ &= P(Z > 4.259177) \\ &= 1 - P(Z \leq 4.259177) \\ &= 1.025903 \times 10^{-5} \end{aligned}$$

This value can be calculated in R as follows.

```
> S<-sum(concordant)-sum(discordant)
> (test.stat<- (S)/sqrt(sig))
[1] 4.29169
> (test.stat.corrected<- (S+(-1*sign(S)))/sqrt(sig))
[1] 4.259177
> pnorm(test.stat.corrected,lower.tail=FALSE)
[1] 1.025903e-05
```

Decision Making Figure 13.7.21 shows that the observed data falls far in the rejection region for this test. We used the following function to solve the test statistic equation for τ_b to find the corresponding value, noting that

$$n_c - n_d = \tau_b \sqrt{(n^2 - n)\sum_i T_{x_i}(T_{x_i} - 1) - (n^2 - n)\sum_i T_{y_i}(T_{y_i} - 1)}.$$

```
> solve_tau<-function(kendalls.tab,z,tau,correct=TRUE){
+   kendalls.tab<-data.frame(kendalls.tab)
+   tab<-table(kendalls.tab$x,kendalls.tab$y)
+   n<-nrow(tab)
+   Tx<-rowSums(tab)
+   Ty<-colSums(tab)
+   sig2<-(1/18)*((n^2-n)*(2*n+5)-sum(Tx*(Tx-1)*(2*Tx+5))-sum(Ty*(Ty-1)*(2*Ty+5)))+
+     (sum(Tx*(Tx-1)*(Tx-2))*sum(Ty*(Ty-1)*(Ty-2)))/(9*(n^2-n)*(n-2))+
+     (sum(Tx*(Tx-1))*sum(Ty*(Ty-1)))/(2*(n^2-n))
+   S<-tau*sqrt(((n^2-n)-sum(Tx*(Tx-1)))*((n^2-n)-sum(Ty*(Ty-1))))
+   if(correct==TRUE){
+     (S+(-1*sign(S)))/sqrt(sig) - z
+   }else{
+     (S)/sqrt(sig) - z
+   }
+ }
```

Solving the above function for τ, given z, yields the corresponding Kendall's tau-b correlation. We create Figure 13.7.21 as follows.

```
> ggdat<-data.frame(z=seq(-4.5,4.5,0.01),
+                   f=dnorm(x=seq(-4.5,4.5,0.01)))
> ggdat.highlight<-data.frame(x=test.stat,y=0)
> alpha<-0.05
> critical<-uniroot(f=solve_tau,interval=c(0,1),
+                kendalls.tab=kendall.tab,z=qnorm(1-alpha),correct=TRUE)$root
> axis.labels<-c(round(c(0,critical,r),4))
> ggplot(data=ggdat,aes(x=z,y=f))+
+   geom_line()+
+   geom_ribbon(data=subset(ggdat,z>=qnorm(1-alpha)),aes(ymax=f),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,z>=test.stat),aes(ymax=f),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("z")+
+   ylab("Density")+
+   ggtitle("Z Test for the Populaiton Kendall's Tau-b",
+           subtitle=bquote(H[0]*":"~rho[tau]==0*", versus "*H[a]*":"~rho[tau]>0))+
+   annotate("text",x=2.5,y=0.07,
+            label= deparse(bquote(alpha==0.05)),parse=TRUE,size=3.5)+
+   annotate("text",x=4.25,y=0.05,label="Observation\n p-value<0.0001",size=3.5)+
+   scale_x_continuous(sec.axis = sec_axis(~.,
+         breaks=c(0,qnorm(1-alpha),test.stat),
+         labels = axis.labels,name="Kendall's Tau-b Correlation"))
```

There is significant evidence that the population Kendall's tau-b correlation is greater than zero ($z = 4.259177$, $p < 0.0001$).

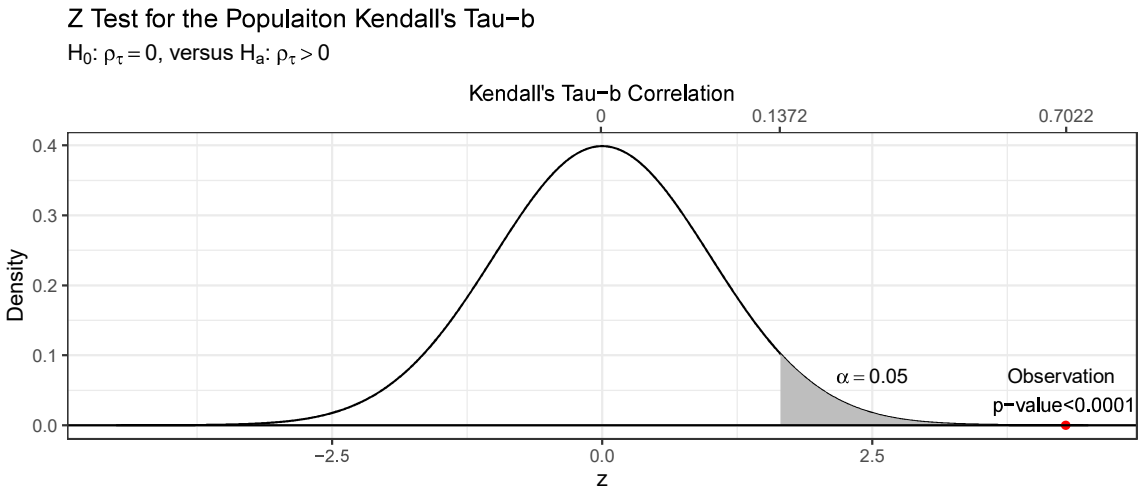

Figure 13.7.21: P-value as calculated for the hypothesis test about the Kendall's tau-b correlation for Example 13.10.

Summary In this section, we succinctly summarize the five steps for the hypothesis test about a population Kendall's tau-b correlation. While it might be simple to check these tables and simply churn through the formulas, it is important to understand the story and interpretations explained in the sections above.

Step One	$H_a : \rho_\tau < 0$	$H_a : \rho_\tau > 0$	$H_a : \rho_\tau \neq 0$
Step Two	Check Assumptions		
Step Three	$z^* = \frac{n_c - n_d + c}{\sigma}$		
Step Four	$P(Z < z^*)$	$P(Z > z^*)$	$2P(Z < -\lvert z^* \rvert)$
Step Five	Reject H_0 if p-value$< \alpha$ or if z^* is in the rejection region		

For our researchers,

Step One	$H_0 : \rho = 0$ $H_a : \rho > 0$
Step Two	The variables are quantitativre. A scatterplot shows a monotone relationship. There are no visible outliers. the two varaibles are approximately normally distributed. Sample is generalizable. Observations are independent.
Step Three	$z^* = \frac{n_c - n_d + c}{\sigma} = 4.259177$
Step Four	p-value $= P(Z < z^*) < 0.0001$
Step Five	p-value$< 0.05 \longrightarrow$ Reject H_0.

This t test for the population Kendall's tau-b correlation can be calculated in `R` as follows.

```
> cor.test(filt.rate,moisture,method="kendall",alternative="greater",continuity=TRUE)

Kendall's rank correlation tau

data:  filt.rate and moisture
z = 4.2592, p-value = 1.026e-05
alternative hypothesis: true tau is greater than 0
sample estimates:
tau
0.7021674

Warning message:
In cor.test.default(filt.rate, moisture, method = "kendall", alternative = "greater",  :
Cannot compute exact p-value with ties
```

In the case where the bivariate Gaussian assumption is grossly violated, we can conduct a **permutation test**. Like bootstrapping, a permutation test allows us to "peak" at the attributes of a sampling distribution by resampling the observed data. Specifically, we shuffle the order (or permute) observations without replacement, unlike bootstrapping.

Intuition: If the data are not correlated, then if we shuffle the observations the correlation won't change by very much and if the data are correlated, then if we shuffle the observations we should see a large change in the correlations.

This test can be calculated in R as follows.

```
> kendall.perm.test <- function(x,y,R=1000,alternative="two.sided",supplyCorrs=FALSE){
+   n = length(x) #length(x)=length(y)
+   if(n!=length(y)){
+     stop("\n 'x' and 'y' must have the same length")
+   }
+   r = cor(x,y,method="kendall") #sample correlation
+   #a place to store correlations on resampled
+   corrs<-rep(NA,R)
+   for(i in 1:R){
+     #randomly generate observations to sample
+     curr.samp<-sample(x=1:n,size=n,replace=FALSE)
+     #save correlation for this sample
+     corrs[i]<-cor(x,y[curr.samp],method="kendall")
+   }
+   #calculate p-values as the proportion of correlations
+   of resampled 'more extreme' than observed
+   if(alternative=="less"){
+     p.value = sum(corrs <= r) / R
+   } else if(alternative=="greater"){
+     p.value = sum(corrs >= r) / R
+   } else{
+     p.value = sum(abs(corrs) >= abs(r)) / R
+   }
+   if(supplyCorrs==TRUE){
+     list(r=r,p.value=p.value,corrs=corrs)
```

```
+   }else{
+     list(r=r,p.value=p.value)
+   }
+ }
> perm.test<-kendall.perm.test(filt.rate,moisture,supplyCorrs=TRUE)
> perm.test[1:2]
$r
[1] 0.7021674

$p.value
[1] 0
```

Figure 13.7.22, created with the R code below, shows that the observation is much larger than any of the observed Kendall's tau-b correlations when we shuffle the data. Thus, there is significant evidence that the population Kendall's tau-b correlation is greater than zero.

```
> ggdat<-data.frame(corrs=perm.test$corrs)
> ggplot(data=ggdat,aes(x=corrs))+
+   geom_histogram(aes(y=..density..),
+                  fill="lightblue",
+                  color="black",
+                  bins=20)+
+   geom_hline(yintercept=0)+
+   geom_vline(xintercept=perm.test$r,color="red")+
+   theme_bw()+
+   xlab("Kendall's Tau-b Correlations")+
+   ylab("Density")+
+   ggtitle("Permutation Test For Kendall's Tau-b Correlation",
+           subtitle=bquote(H[0]*":"~rho[tau]==0*", versus "*H[a]*":"~rho[tau]>0))
```

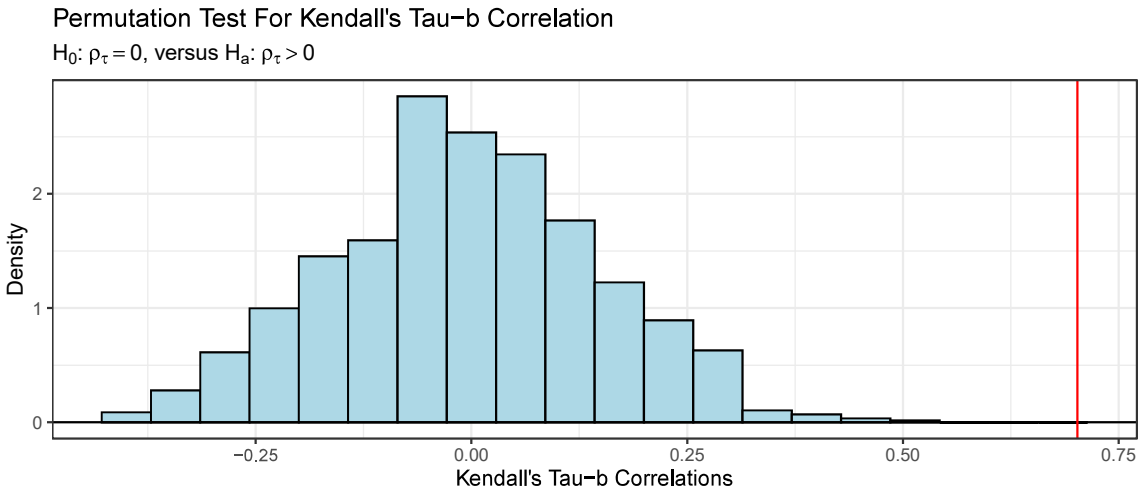

Figure 13.7.22: P-value as calculated for the permutation hypothesis test about the Kendall correlation for Example 13.10.

13.7.2 Confidence Interval for Kendall's Tau-b Correlation

Using the assumption of bivariate normality, we transform τ_b to z_r; e.g.,

$$z_r = \frac{1}{2}\ln\left(\frac{1+\tau_b}{1-\tau_b}\right).$$

The transformed variable z_r is approximately Gaussian with standard error $\sqrt{\frac{0.437}{n-4}}$. To find the $(1-\alpha)\times 100\%$ confidence interval for the population Kendall's tau-b correlation we calculate upper and lower bounds

$$\begin{aligned} z_l &= z_r - z_{1-\alpha/2}\sqrt{\frac{0.437}{n-4}} \\ z_u &= z_r + z_{1-\alpha/2}\sqrt{\frac{0.437}{n-4}}, \end{aligned}$$

and transform back to r,

$$\begin{aligned} r_l &= \frac{e^{2z_l}-1}{e^{2z_l}+1} \\ u_u &= \frac{e^{2z_u}-1}{e^{2z_u}+1}. \end{aligned}$$

For our researchers, we can calculate a 95% confidence interval as follows. First, we transform τ_b to z_r,

$$\begin{aligned} z_r &= \frac{1}{2}\ln\left(\frac{1+\tau_b}{1-\tau_b}\right) \\ &= \frac{1}{2}\ln\left(\frac{1+0.7021674}{1-(0.7021674)}\right) \\ &= 0.871563 \end{aligned}$$

The upper and lower bounds are calculated as

$$\begin{aligned} z_l &= z_r - z_{1-\alpha/2}\sqrt{\frac{0.437}{20-4}} \\ &= 0.871563 - z_{0.975}\sqrt{\frac{0.437}{16}} \\ &= 0.871563 - 1.959964\sqrt{0.0273125} \\ &= 0.6336589 \\ z_u &= z_r + z_{1-\alpha/2}\sqrt{\frac{0.437}{20-4}} \\ &= 0.871563 + z_{0.975}\sqrt{\frac{0.437}{16}} \\ &= 0.871563 + 1.959964\sqrt{0.0273125} \\ &= 1.195476, \end{aligned}$$

which we can transform back to r

$$\begin{aligned} r_l &= \frac{e^{2z_l}-1}{e^{2z_l}+1} \\ &= \frac{e^{2(0.6336589)}-1}{e^{2(0.6336589)}+1} \\ &= 0.5605665 \\ r_u &= \frac{e^{2z_u}-1}{e^{2z_u}+1} \\ &= \frac{e^{2(1.195476)}-1}{e^{2(1.195476)}+1} \\ &= 0.8322696. \end{aligned}$$

This indicates that we are 95% confident the true population Kendall's tau-b correlation, ρ_τ is between 0.5605665 and 0.8322696, the monotone association is strong and positive.

The confidence interval for the population Pearson correlation can be calculated in R as follows.

```
> cor.test(filt.rate,moisture,method="kendall",continuity=TRUE)

Kendall's rank correlation tau

data:  filt.rate and moisture
z = 4.2592, p-value = 2.052e-05
alternative hypothesis: true tau is not equal to 0
sample estimates:
tau
0.7021674

Warning message:
In cor.test.default(filt.rate, moisture, method = "kendall", continuity = TRUE) :
Cannot compute exact p-value with ties
```

In the case where the bivariate Gaussian assumption is not reasonable, we can construct a bootstrap confidence interval for the population Pearson correlation. This can be done with the following R code.

```
> library("boot")
> boot.kendall<-function(data,indices){
+   d<-data[indices,]
+   return(cor(d$filt.rate,d$moisture,method="kendall"))
+ }
> boot.dat<-data.frame(filt.rate=filt.rate,moisture=moisture)
> boot<-boot(boot.dat,R=1000,statistic=boot.kendall)
> boot.ci(boot.out=boot,conf=0.95,type="perc")
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot, conf = 0.95, type = "perc")
```

```
Intervals :
Level     Percentile
95%   ( 0.4193,  0.8983 )
Calculations and Intervals on Original Scale
```

This indicates that we are 95% confident the true population Kendall's tau-b correlation, ρ is between 0.4193 and 0.8983, the monotone association is strong and positive.

13.8 An Interesting Note About Bias

The Spearman's rank and Kendall's tau-b statistics are biased, while Pearson's correlation is unbiased. These unbiased statistics can be corrected so that we have three unbiased estimators for population correlations.

$$\begin{aligned}\widehat{\rho} &= r \\ \widehat{\rho}_s &= 2\sin\left(\frac{\pi}{6}r_s\right) \\ \widehat{\rho}_\tau &= \sin\left(\frac{\pi}{2}\tau_b\right)\end{aligned}$$

Xu et al. (2013) introduced a weighted average of these unbiased estimators

$$\widehat{\rho}_m = 2\sin\left(\frac{r_s}{-}\frac{\pi(\tau_b - r_s)}{2(n-2)}\right).$$

Simulation studies conducted by Xu et al. (2013) show the following.

- For an unbiased estimator $\widehat{\rho}_m$ is most appropriate.
- To minimize MSE then we choose
  - τ_b when the relationship is strong
  - r_s when the relationship is weak
- In larger sample sizes, τ_b outperforms r_s