

Solutions for Computational Probability and Statistics

Ken Horton

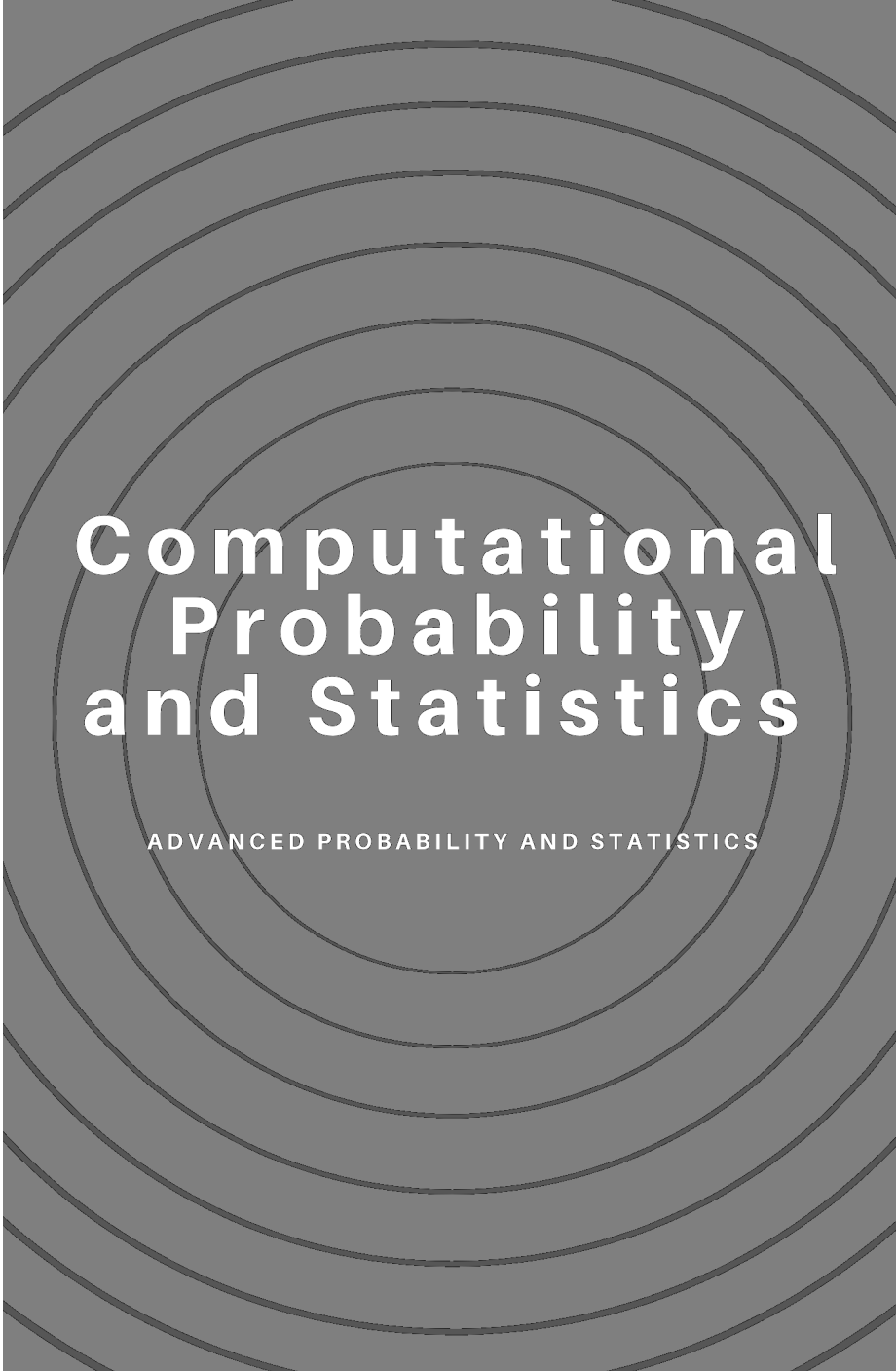
Kris Pruitt

Bradley Warner

2021-03-11

Contents

Preface

The book cover features a series of concentric circles in a dark gray color, centered on a lighter gray background. The circles are of varying radii, creating a tunnel-like or ripple effect. The title is printed in white, bold, sans-serif font, centered within the circles.

Computational Probability and Statistics

ADVANCED PROBABILITY AND STATISTICS

Contained in this volume are the solutions to homework problems in the Computational Probability and Statistics book.

0.1 Book Structure and How to Use It

This solution manual is setup to match the structure of the accompanying book.

The learning outcomes for this course are to use computational and mathematical statistical/probabilistic concepts for:

- a. Developing probabilistic models
- b. Developing statistical models for inference and description
- c. Advancing practical and theoretical analytic experience and skills

0.2 Packages

These notes make use of the following packages in R **knitr** (?), **rmarkdown** (?), **mosaic** (?), **mosaicCalc** (?), **tidyverse** (?), **ISLR** (?), **vcd** (?), **ggplot2** (?), **MASS** (?), **openintro** (?), **broom** (?), **infer** (?), **ISLR** (?), **kableExtra** (?), **DT** (?).



This book is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

0.3 File Creation Information

- File creation date: 2021-03-11
- Windows version: Windows 10 x64 (build 18362)
- R version 3.6.3 (2020-02-29)

Part I

**Descriptive Statistical
Modeling**

Chapter 1

Case Study

1.1 Objectives

- 1) Use R for basic analysis and visualization.
- 2) Compile a report using `knitr`.

1.2 Homework

Load `tidyverse`, `mosaic`, and `knitr` packages.

```
library(tidyverse)
library(mosaic)
library(knitr)
```

1.2.1 Problem 1

1. **Stent study continued.** Complete a similar analysis for the stent data but this time for the one year data. In particular
 - a. Read the data into your working directory.

```
stent_study <- read_csv('data/stent_study.csv')
```

- b. Complete similar steps as in the class notes.
 - i. Use `inspect` on the data.
 - ii. Create a table of `outcome365` and `group`. Comment on the results.
 - iii. Create a barchart of the data.

```
inspect(stent_study)
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

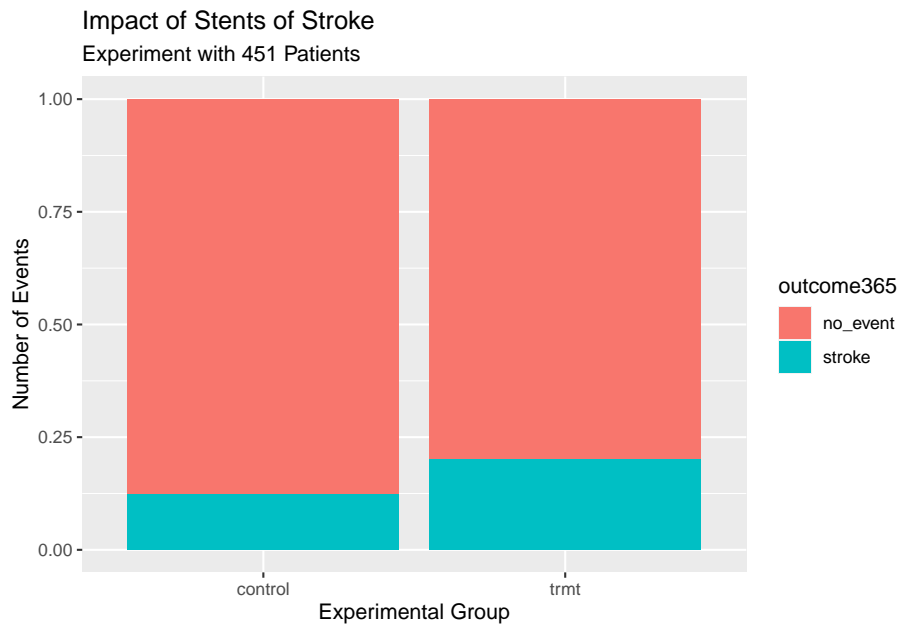
```
##
## categorical variables:
##      name      class levels  n missing
## 1   group character      2 451        0
## 2 outcome30 character      2 451        0
## 3 outcome365 character      2 451        0
##                                     distribution
## 1 control (50.3%), trmt (49.7%)
## 2 no_event (89.8%), stroke (10.2%)
## 3 no_event (83.8%), stroke (16.2%)
```

```
tally(outcome365~group,data=stent_study,format="proportion",margins = TRUE)
```

```
##           group
## outcome365 control    trmt
##   no_event 0.8766520 0.7991071
##   stroke   0.1233480 0.2008929
##   Total    1.0000000 1.0000000
```

Patients in the treatment group had a higher proportion of strokes than those in the control group after one year. The treatment does not appear to help the rate of strokes and in fact may hurt it.

```
stent_study %>%
  gf_props(~group,fill=~outcome365,position='fill') %>%
  gf_labs(title="Impact of Stents of Stroke",
  subtitle='Experiment with 451 Patients',
  x="Experimental Group",
  y="Number of Events")
```



1.2.2 Problem 2

2. **Migraine and acupuncture.** A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 patients in the control group received placebo acupuncture (needle insertion at nonacupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free.¹

The data is in the file `migraine_study.csv` in the folder `data`.

Complete the following work:

- Read the data an object called `migraine_study`.

¹G. Allais et al. “Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints”. In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

```
migraine_study <- read_csv("data/migraine_study.csv")
```

```
head(migraine_study)
```

```
## # A tibble: 6 x 2
##   group    pain_free
##   <chr>    <chr>
## 1 treatment yes
## 2 treatment yes
## 3 treatment yes
## 4 treatment yes
## 5 treatment yes
## 6 treatment yes
```

b. Create a table of the data.

```
tally(pain_free~group, data=migraine_study, format="proportion", margin=TRUE)
```

```
##           group
## pain_free control treatment
##    no    0.95652174 0.76744186
##    yes    0.04347826 0.23255814
##    Total 1.00000000 1.00000000
```

c. Report the percent of patients in the treatment group who were pain free 24 hours after receiving acupuncture.

There are 23.2% of the treatment group pain free.

d. Repeat for the control group.

There are only 4.3% of the control group pain free.

e. At first glance, does acupuncture appear to be an effective treatment for migraines? Explain your reasoning.

Yes, a substantial increase in the percentage of patients pain free after acupuncture versus those with no acupuncture, so it appears to be effective.

f. Do the data provide convincing evidence that there is a real pain reduction for those patients in the treatment group? Or do you think that the observed difference might just be due to chance?

Either of these is acceptable: i. We could get slightly different group estimates even if there is no real difference. Though the difference is big, I'm skeptical the results show a real difference and think this might be due to chance.

ii. The difference in these rates looks pretty big, and so I suspect acupuncture is having a positive impact on pain.

3. Compile, `knit`, this report into a pdf.

Complete on your computer or server.

Chapter 2

Data Basics

2.1 Objectives

- 1) Define and use properly in context all new terminology to include but not limited to case, observational unit, variables, data frame, associated variables, independent, and discrete and continuous variables.
- 2) Identify and define the different types of variables.
- 3) From reading a study, explain the research question.
- 4) Create a scatterplot in R and determine the association of two numerical variables from the plot.

2.2 Homework

Identify study components

Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question in the studies described below.

2.2.1 Problem 1

1. Researchers collected data to examine the relationship between pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter

(PM₁₀) in $\mu\text{g}/\text{m}^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM₁₀ and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.¹

- i. The cases are 143,196 eligible study subjects who were born in Southern California between 1989 and 1993.
- ii. The variables are measurements of carbon monoxide (CO), nitrogen dioxide, ozone, and particulate matter less than $10\mu\text{m}$ (PM10) collected at air-quality-monitoring stations as well as length of gestation. All of these variables are continuous numerical variables.
- iii. The research question was **Is there an association between air pollution exposure and preterm births?**

2.2.2 Problem 2

2. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.²
 - i. The cases are 600 adult patients aged 18-69 years diagnosed and currently treated for asthma.
 - ii. The variables were whether or not the patient practiced the Buteyko method (categorical) and measures of quality of life, activity, asthma symptoms and medication reduction of the patients (categorical, ordinal). It may also be reasonable to treat the ratings on a scale of 1 to 10 as discrete numerical variables.
 - iii. The research question was **Do asthmatic patients who practice the Buteyko method experience improvement in their condition?**

¹B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502–511.

²J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

Chapter 3

Overview of Data Collection Principles

3.1 Objectives

- 1) Define and use properly in context all new terminology.
- 2) From a description of a research project, at a minimum be able to describe the population of interest, the generalizability of the study, the response and predictor variables, differentiate whether it is observational or experimental, and determine the type of sample.

3.2 Homework

3.2.1 Problem 1

1. **Generalizability and causality.** Identify the population of interest and the sample in the studies described below, these are the same studies from the previous lesson. Also comment on whether or not the results of the study can be generalized to the population and if the findings of the study can be used to establish causal relationships.
 - a. Researchers collected data to examine the relationship between pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter

(PM₁₀) in $\mu\text{g}/\text{m}^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM₁₀ and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.¹

The population of interest is all births. The sample consists of the 143,196 births between 1989 and 1993 in Southern California. If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

- b. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.²

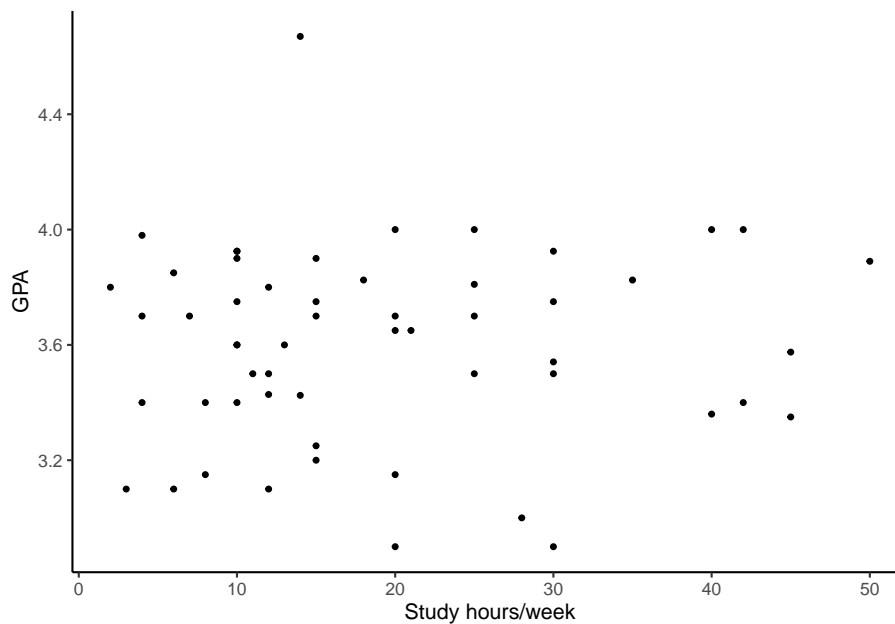
The population is all 18-69 year olds diagnosed and currently treated for asthma. The sample is the 600 adult patients aged 18-69 years diagnosed and currently treated for asthma. Since the sample is not random (voluntary) the results cannot be generalized to the population at large. However, since the study is an experiment, the findings can be used to establish causal relationships.

¹B. Ritz et al. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". In: *Epidemiology* 11.5 (2000), pp. 502-511.

²J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: *Thorax* 58 (2003).

3.2.2 Problem 2

2. **GPA and study time.** A survey was conducted on 55 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about their GPA and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.



- What is the explanatory variable and what is the response variable?
- Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
- Is this an experiment or an observational study?
- Can we conclude that studying longer hours leads to higher GPAs?

Solutions

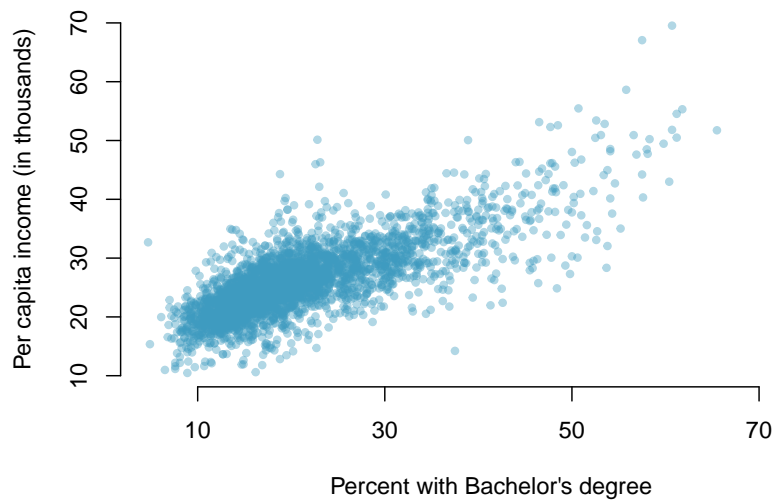
- The explanatory variable is the number of study hours per week, and the response variable is GPA.
- There is a somewhat weak positive relationship between the two variables, though the data become more sparse as the number of study hours increases. One respondent reported a GPA above 4.0, which is clearly a data error. Also, there are a few respondents who reported unusually high study hours (60 and 70 hours/week). It should also be noted that

the variability in GPA is much higher for students who study less than those who study more, also might be due to the fact that there aren't many respondents who reported studying higher hours.

- c. This is an observational study.
- d. Since this is an observational study, we cannot conclude that there is a causal relationship between the two variables even though there appears to be an association.

3.2.3 Problem 3

3. **Income and education** The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.



- What are the explanatory and response variables?
- Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
- Can we conclude that having a bachelor's degree increases one's income?

Solutions

- The explanatory variable is percent of population with a bachelor's degree and the response variable is per capita income (in thousands).
- There is a strong positive linear relationship between the two variables. As the percentage of population with a bachelor's degree increases the per capita income increases as well. There are very few counties where more than 60% of the population have a bachelor's degree and very few counties that have a more than \$50,000 in per capita income.
- This is an observational study so we cannot make a causal statement based on the results. However, we can say that having a higher percentage of population with bachelor's degree is associated with a higher per capita income.

Chapter 4

Studies

4.1 Objectives

- 1) Define and use properly in context all new terminology.
- 2) Given a study description, be able to identify and explain the study using correct terms.
- 3) Given a scenario, describe flaws in reasoning and propose study and sampling designs.

4.2 Homework

4.2.1 Problem 1

1. **Propose a sampling strategy.** A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.
 - a. What type of study is this? Observational study.
 - b. Suggest a sampling strategy for carrying out this study. Stratified sample, sample randomly within each section.

4.2.2 Problem 2

2. **Flawed reasoning.** Identify the flaw in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

- a. Students at an elementary school are given a questionnaire that they are required to return after their parents have completed it. One of the questions asked is, *Do you find that your work schedule makes it difficult for you to spend time with your kids after school?* Of the parents who replied, 85% said *no*. Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

Solution

Non-responders may have a different response to this question. The parents who returned the surveys are probably those who do not have difficulty spending time with their kids after school. Parents who work might not have returned the surveys since they probably have a busier schedule.

- b. A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

Solution

It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might be in a lower socio-economic status than the respondents.

4.2.3 Problem 3

3. **Sampling strategies.** A Math 377 student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Four research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.
- a. He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
 - b. He gives out the survey only to his friends, and makes sure each one of them fills out the survey.
 - c. He posts a link to an online survey on his Facebook wall and asks his friends to fill out the survey.
 - d. He stands outside the QRC and asks every third person that walks out the door to fill out the survey.

Solution

- a. Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population.
- b. Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey.
- c. Convenience sample. This will have a similar issues to handing out surveys to friends.
- d. Convenience sample. Same.

4.2.4 Problem 4

4. **Vitamin supplements.** In order to assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four medication groups, and the placebo group had the shortest duration of symptoms.

- a. Was this an experiment or an observational study? Why?
- b. What are the explanatory and response variables in this study?
- c. Were the patients blinded to their treatment?
- d. Was this study double-blind?
- e. Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

Solution

- a. Experiment, since the researchers randomly assigned different treatments to the participants.
- b. Response variable: Duration of the cold.
Explanatory variable: Treatment, with 4 levels; placebo, 1g, 3g, 3g with additives.
- c. The patients were blinded as they did not know which treatment they received.
- d. The study was double-blind with respect to the researchers evaluating the patients, but the nurses who briefly interacted with patients during the distribution of the medication were not blinded. (It was partially double-blind.)
- e. Since the patients were randomly assigned to the treatment groups and they are blinded we would expect about an equal number of patients in each group to not adhere to the treatment. While this means that final results of the study will be based on fewer number of participants, non-adherence does not introduce a confounding variable to the study.

4.2.5 Problem 5

5. **Exercise and mental health.** A researcher is interested in the effects of exercise on mental health and she proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41-55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.
- What type of study is this?
 - What are the treatment and control groups in this study?
 - Does this study make use of blocking? If so, what is the blocking variable?
 - Does this study make use of blinding?
 - Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.
 - Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

Solution

- This is an experiment since we assigned subjects to the exercise program.
- The treatment is exercise twice a week and control is no exercise.
- Yes, the blocking variable is age.
- No, the study is not blinded since the patients will know whether or not they are exercising.
- Since this is an experiment, we can make a causal statement. Since the sample is random, the causal statement can be generalized to the population at large. However, we should be cautious about making a causal statement because of a possible placebo effect.
- It would be very difficult, if not impossible, to successfully conduct this study since randomly sampled people cannot be required to participate in a clinical trial.

Chapter 5

Numerical Data

5.1 Objectives

- 1) Define and use properly in context all new terminology.
- 2) Generate in **R** summary statistics for a numeric variable including breaking down by cases.
- 3) Generate in **R** appropriate graphical summaries of numerical variables.
- 4) Be able to interpret and explain output both graphically and numerically.

5.2 Homework

5.2.1 Problem 1

1. Mammals exploratory

Data were collected on 39 species of mammals distributed over 13 orders. The data is in the **openintro** package as **mammals**

- a. Using **help**, report the units for the variable **BrainWt**.

```
?mammals
```

- b. Using **inspect** how many variables are numeric?

```
inspect(mammals)
```

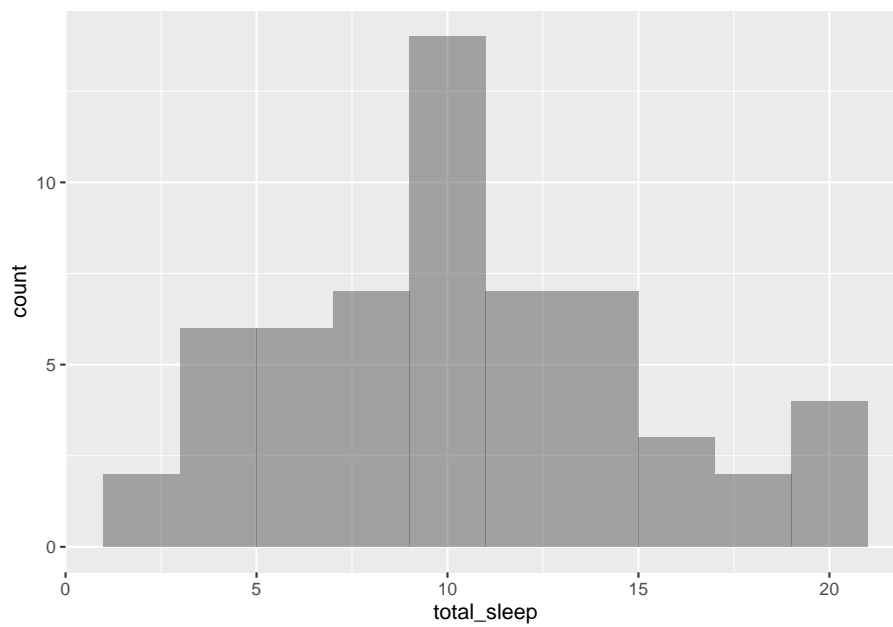
```
##
## categorical variables:
##   name  class levels  n missing
## 1 species factor    62 62      0
##                                     distribution
## 1 Africanelephant (1.6%) ...
##
## quantitative variables:
##   name  class  min    Q1  median    Q3    max    mean
## ...1   body_wt numeric 0.005 0.600 3.3425 48.2025 6654.0 198.789984
## ...2   brain_wt numeric 0.140 4.250 17.2500 166.0000 5712.0 283.134194
## ...3 non_dreaming numeric 2.100 6.250 8.3500 11.0000 17.9 8.672917
## ...4   dreaming numeric 0.000 0.900 1.8000 2.5500 6.6 1.972000
## ...5 total_sleep numeric 2.600 8.050 10.4500 13.2000 19.9 10.532759
## ...6   life_span numeric 2.000 6.625 15.1000 27.7500 100.0 19.877586
## ...7   gestation numeric 12.000 35.750 79.0000 207.5000 645.0 142.353448
## ...8   predation integer 1.000 2.000 3.0000 4.0000 5.0 2.870968
## ...9   exposure integer 1.000 1.000 2.0000 4.0000 5.0 2.419355
## ...10  danger integer 1.000 1.000 2.0000 4.0000 5.0 2.612903
##           sd  n missing
## ...1  899.158011 62      0
## ...2  930.278942 62      0
## ...3   3.666452 48     14
## ...4   1.442651 50     12
## ...5   4.606760 58      4
## ...6  18.206255 58      4
## ...7 146.805039 58      4
## ...8   1.476414 62      0
## ...9   1.604792 62      0
## ...10  1.441252 62      0
```

c. What type of variable is `danger`?

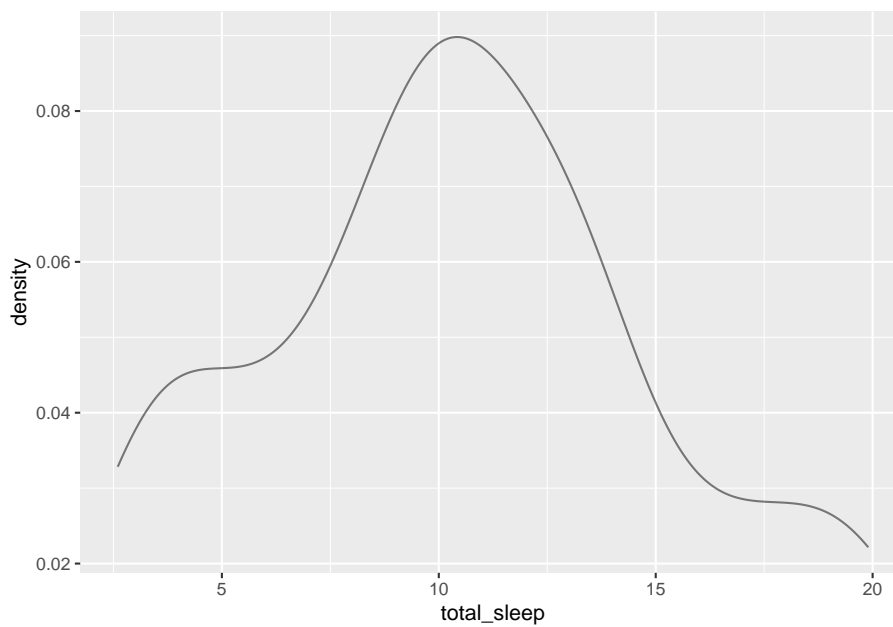
Categorical

d. Create a histogram of `total_sleep` and describe the distribution.

```
gf_histogram(~total_sleep, data=mammals, binwidth = 2)
```

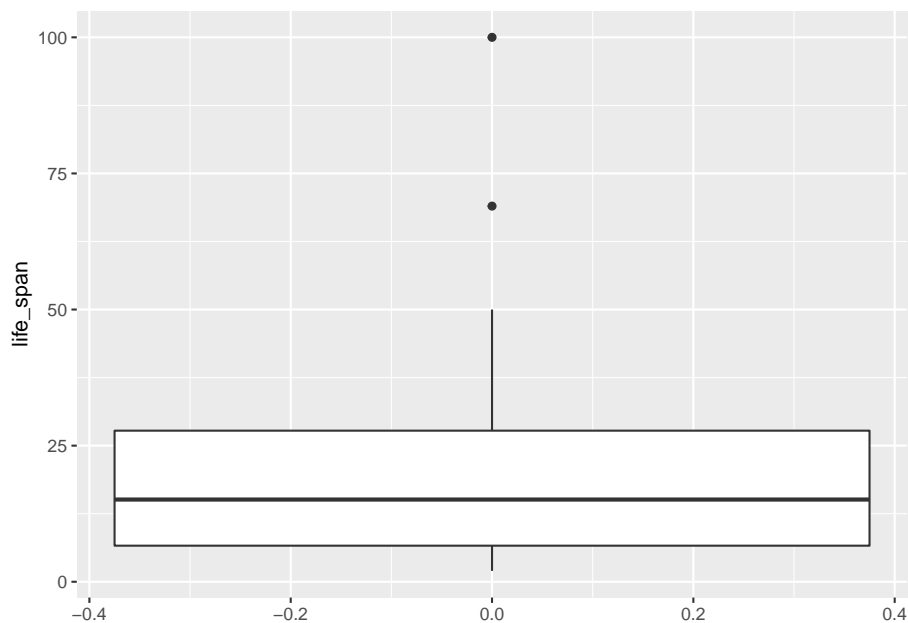
```
gf_dens(~total_sleep,data=mammals)
```



The distribution is unimodal and skewed to the right. It appears it is centered around the value of 11.

e. Create a boxplot of `life_span` and describe the distribution.

```
gf_boxplot(~life_span,data=mammals)
```



f. Report the mean and median life span of a mammal.

```
mean(~life_span,data=mammals,na.rm=TRUE)
```

```
## [1] 19.87759
```

```
median(~life_span,data=mammals,na.rm=TRUE)
```

```
## [1] 15.1
```

g. Calculate the summary statistics for `LifeSpan` broken down by `Danger`.

```
favstats(life_span~danger,data=mammals)
```

```
##   danger  min    Q1 median    Q3   max   mean    sd  n missing
## 1      1  3.0  7.700 17.60 32.500 100.0 24.20556 23.53829 18      1
## 2      2  2.3  4.500 10.40 13.000  50.0 12.92308 13.15948 13      1
## 3      3  2.0  4.175  5.35  7.875  38.6  9.43750 11.99559  8      2
## 4      4  2.6  9.775 22.10 27.000  69.0 23.11000 18.75482 10      0
## 5      5 17.0 20.000 23.60 30.000  46.0 26.95556 10.18910  9      0
```

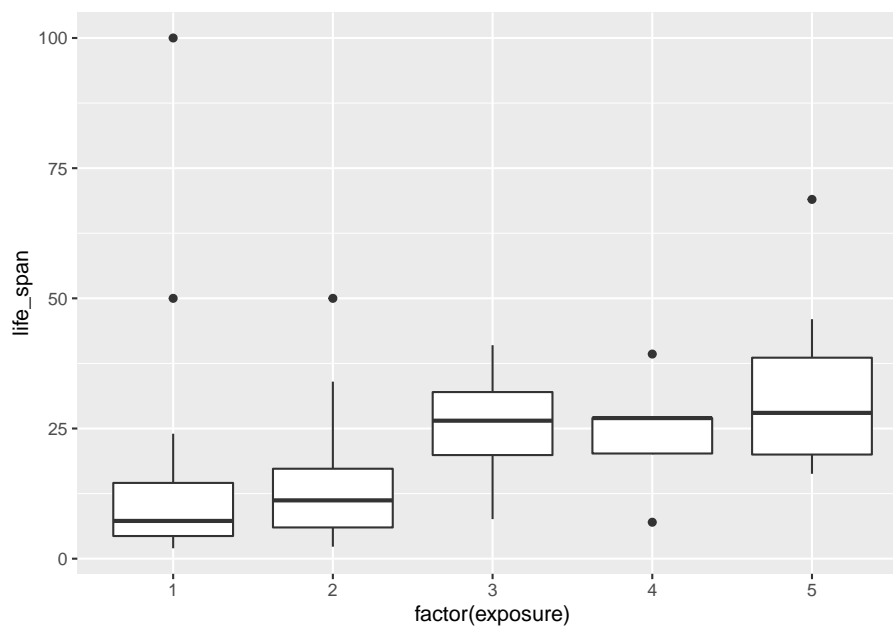
5.2.2 Problem 2

2. Mammals life spans

Continue using the `mammals` data set.

- a. Create side-by-side boxplots for `life_span` broken down by `exposure`.
Note: you will have to change `exposure` to a `factor()`. Report on any findings.

```
mammals %>%  
  gf_boxplot(life_span~factor(exposure))
```



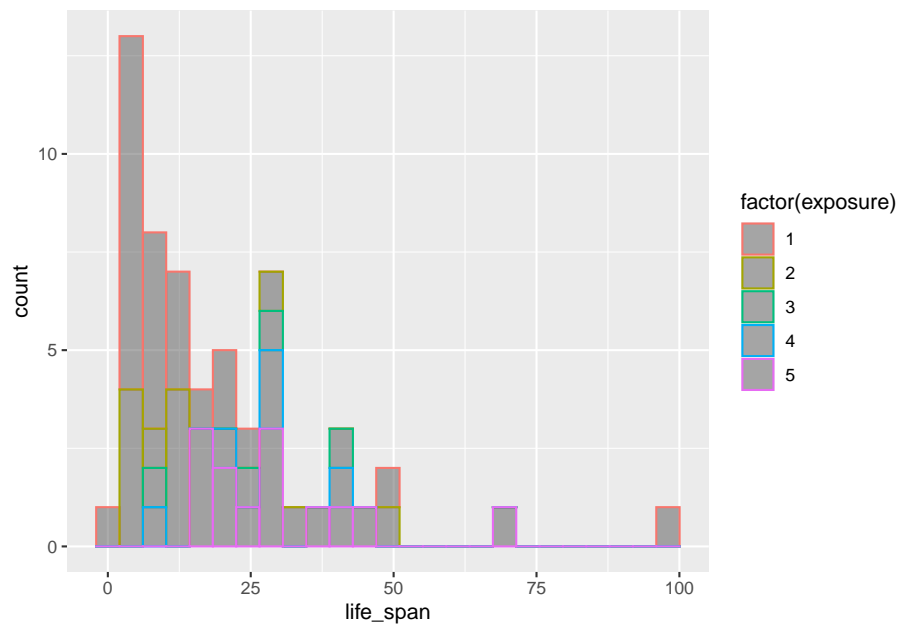
- b. What happened to the median and third quartile in exposure group 4?

```
favstats(life_span~factor(exposure), data=mammals)
```

##	factor(exposure)	min	Q1	median	Q3	max	mean	sd	n	missing
## 1	1	2.0	4.35	7.25	14.550	100.0	14.55000	20.98594	24	3
## 2	2	2.3	6.00	11.20	17.275	50.0	15.39167	14.55819	12	1
## 3	3	7.6	19.90	26.50	32.000	41.0	25.40000	13.84582	4	0
## 4	4	7.0	20.20	27.00	27.000	39.3	24.10000	11.78431	5	0
## 5	5	16.3	20.00	28.00	38.600	69.0	30.53077	14.98084	13	0

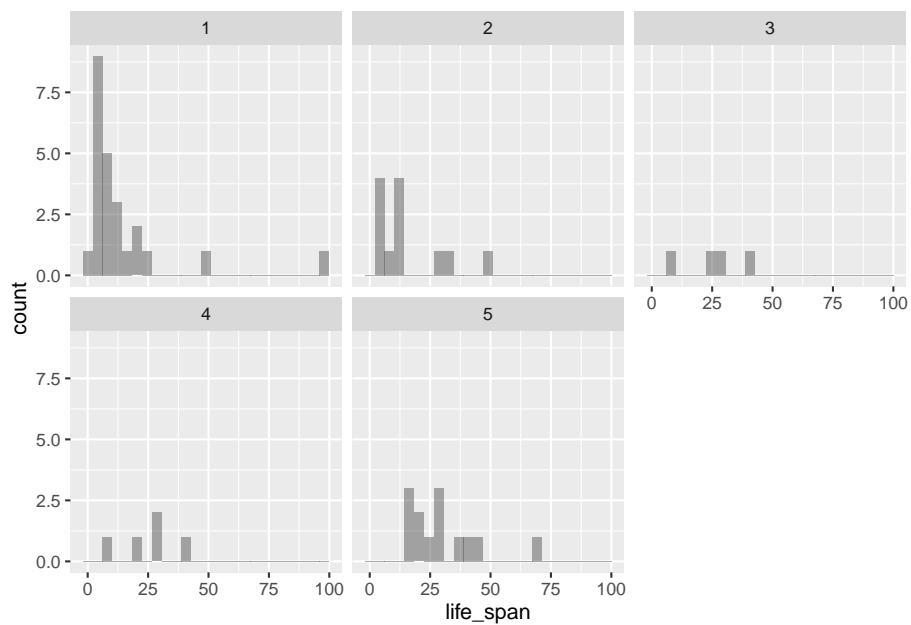
- c. Create faceted histograms. What are the shortcomings of this plot?

```
gf_histogram(~life_span,color=~factor(exposure),data=mammals)
```



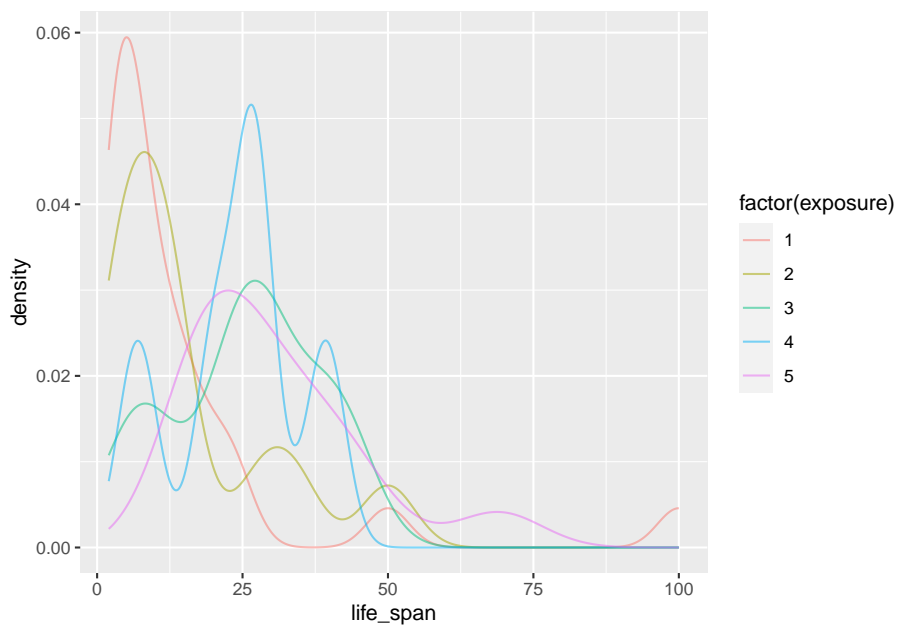
This is awful.

```
gf_histogram(~life_span|factor(exposure),data=mammals)
```

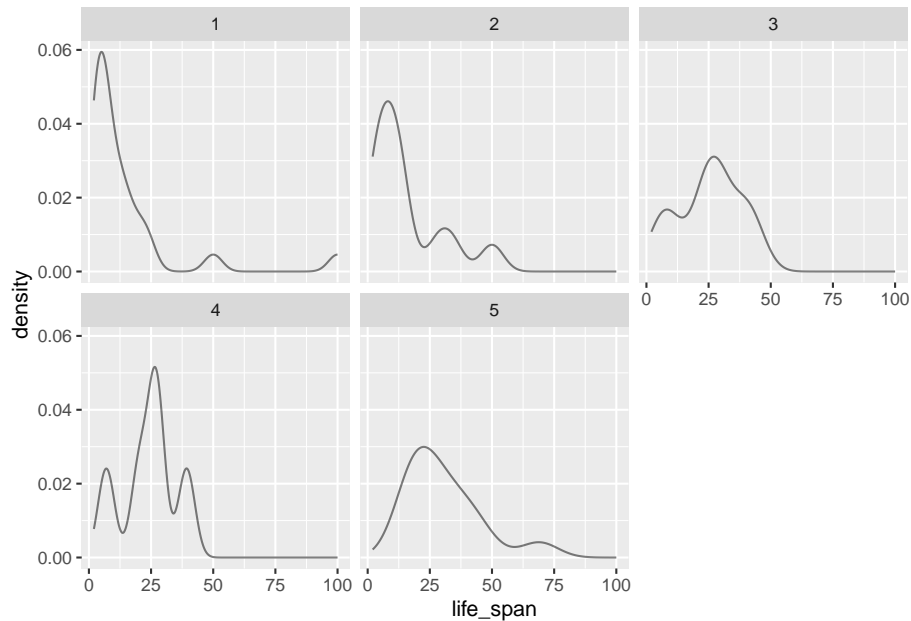


Not enough data for each histogram; some of the histograms provide little to no information. Let's do density plots.

```
gf_dens(~life_span,color=~factor(exposure),data=mammals)
```



```
gf_dens(~life_span|factor(exposure),data=mammals)
```



Which do you think is the best graph?

- d. Create a new variable `exposed` that is a factor with level `Low` if exposure is 1 or 2 and `High` otherwise.

```
mammals <- mammals %>%
  mutate(exposed=factor(ifelse((exposure==1)|(exposure==2),"Low","High")))
```

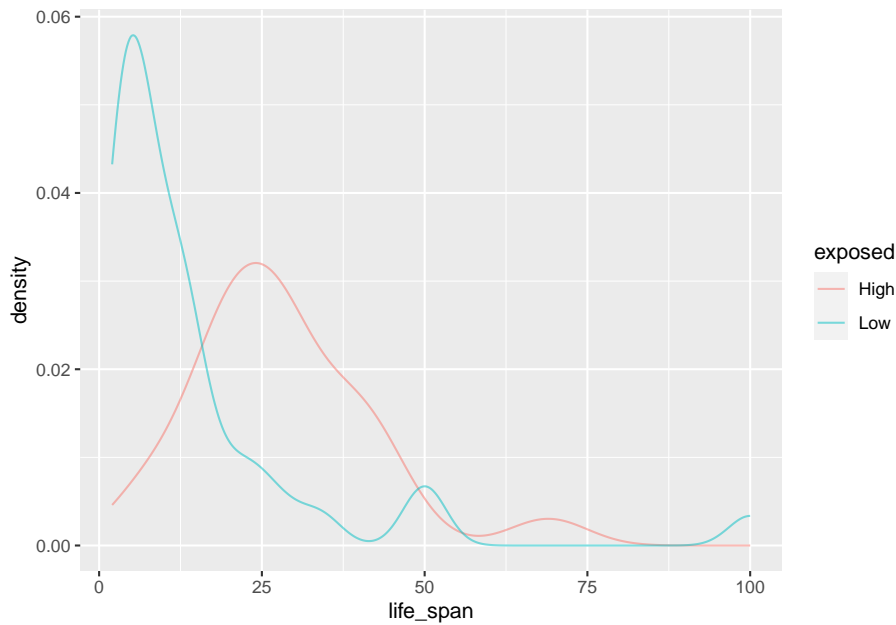
```
inspect(mammals)
```

```
##
## categorical variables:
##   name  class levels  n missing
## 1 species factor    62 62      0
## 2 exposed factor     2 62      0
##
##                                     distribution
## 1 Africanelephant (1.6%) ...
## 2 Low (64.5%), High (35.5%)
##
## quantitative variables:
##   name  class  min    Q1  median    Q3   max    mean
```

```
## ...1      body_wt numeric  0.005  0.600  3.3425  48.2025 6654.0 198.789984
## ...2      brain_wt numeric  0.140  4.250 17.2500 166.0000 5712.0 283.134194
## ...3 non_dreaming numeric  2.100  6.250  8.3500  11.0000  17.9   8.672917
## ...4      dreaming numeric  0.000  0.900  1.8000  2.5500   6.6   1.972000
## ...5    total_sleep numeric  2.600  8.050 10.4500  13.2000  19.9  10.532759
## ...6      life_span numeric  2.000  6.625 15.1000  27.7500 100.0  19.877586
## ...7      gestation numeric 12.000 35.750 79.0000 207.5000 645.0 142.353448
## ...8      predation integer  1.000  2.000  3.0000  4.0000   5.0   2.870968
## ...9      exposure integer  1.000  1.000  2.0000  4.0000   5.0   2.419355
## ...10     danger integer  1.000  1.000  2.0000  4.0000   5.0   2.612903
##          sd n missing
## ...1  899.158011 62      0
## ...2  930.278942 62      0
## ...3   3.666452 48     14
## ...4   1.442651 50     12
## ...5   4.606760 58      4
## ...6  18.206255 58      4
## ...7 146.805039 58      4
## ...8   1.476414 62      0
## ...9   1.604792 62      0
## ...10  1.441252 62      0
```

e. Repeat part c with the new variable.

```
gf_dens(~life_span,color=~exposed,data=mammals)
```

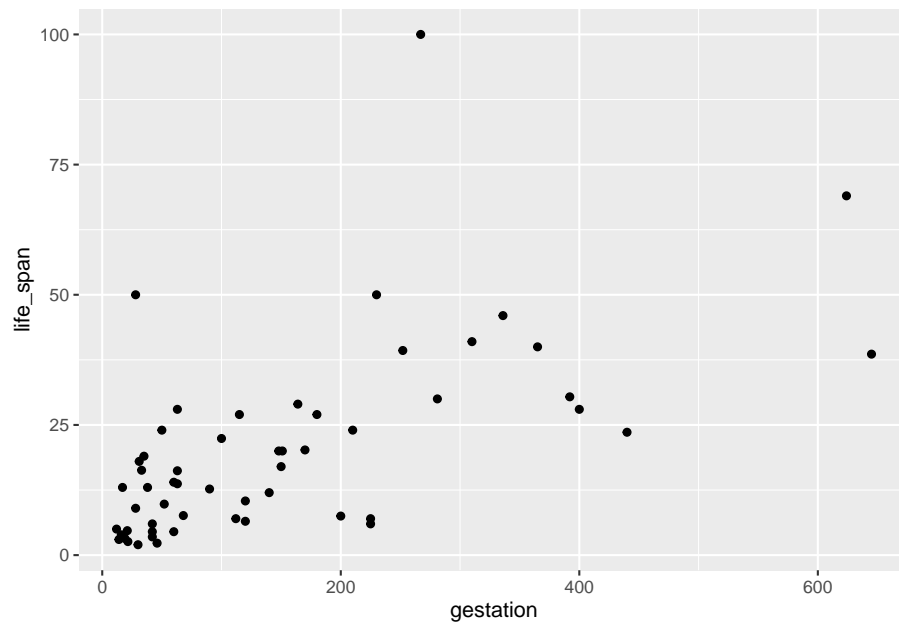


5.2.3 Problem 3

3. Mammals life spans continued

- a. Create a scatterplot of life span versus length of gestation.

```
mammals %>%  
  gf_point(life_span~gestation)
```



- b. What type of an association is apparent between life span and length of gestation?

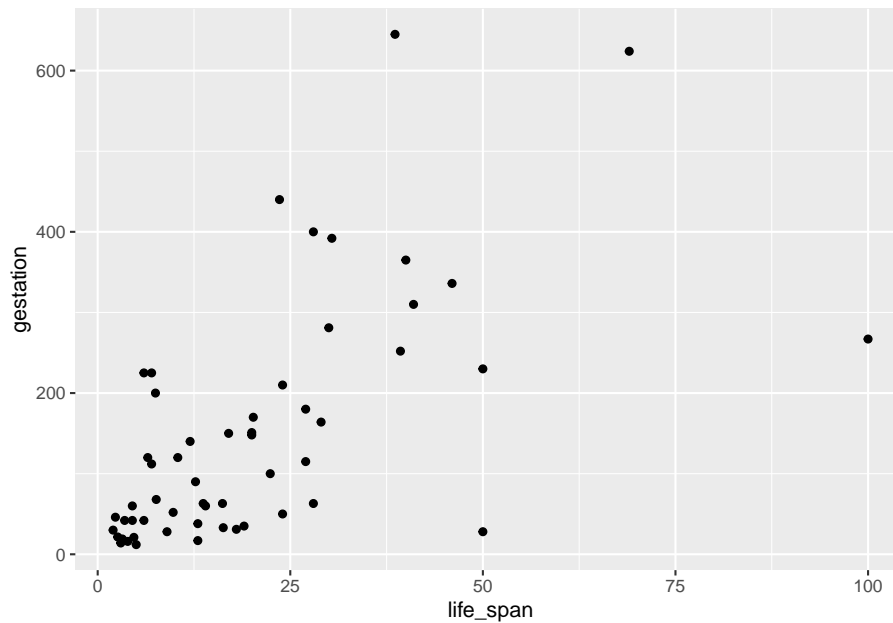
It is a weak positive association.

- c. What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

The same as this is observational data there is no reason to believe is a causal relationship just by looking at the data. Switching the axis will preserve the association.

- d. Create the new scatterplot suggested in c.


```
mammals %>%  
gf_point(gestation~life_span)
```



e. Are life span and length of gestation independent? Explain your reasoning.

No there is an association and it appears to be linear. If the plot looked like a “shotgun” blast, we would consider the variables to be independent. However, remember there may be confounding variables that could impact the association between these variables.

Chapter 6

Categorical Data

6.1 Objectives

- 1) Define and use properly in context all new terminology.
- 2) Generate in R tables for categorical variable(s).
- 3) Generate in R appropriate graphical summaries of categorical and numerical variables.
- 4) Be able to interpret and explain output both graphically and numerically.

6.2 Homework

Make sure your plots have a title and the axes are labeled.

6.2.1 Problem 1

1. Views on immigration

910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country.

The data is in the `openintro` package in the `immigration` data object.

- a. How many levels of *political* are there?

```
levels(immigration$political)
```

```
## [1] "conservative" "liberal"      "moderate"
```

```
inspect(immigration)
```

```
##
## categorical variables:
##      name  class levels   n missing
## 1 response factor      4 910        0
## 2 political factor      3 910        0
##                                     distribution
## 1 Leave the country (38.5%) ...
## 2 conservative (40.9%), moderate (39.9%) ...
```

There are three levels for `political` and they are conservative, liberal, and moderate.

- b. Create a table using `tally`.

```
round(tally(~response+political,data=immigration,format="percent",margins = TRUE),2)
```

```
##
##      response      political
##      conservative liberal moderate Total
## Apply for citizenship      6.26  11.10  13.19 30.55
## Guest worker              13.30   3.08  12.42 28.79
## Leave the country          19.67   4.95  13.85 38.46
## Not sure                   1.65   0.11   0.44  2.20
## Total                     40.88  19.23  39.89 100.00
```

- c. What percent of these Tampa, FL voters identify themselves as conservatives?

From the table, 40.88% of voters identified themselves as conservatives.

- d. What percent of these Tampa, FL voters are in favor of the citizenship option?

Again, from the table 30.55% of the voters favor the citizenship option.

- e. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

From the table, 6.26% of the voters are conservative and favor the citizenship option.

- f. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates and liberal share this view?

We need a different table for this question.

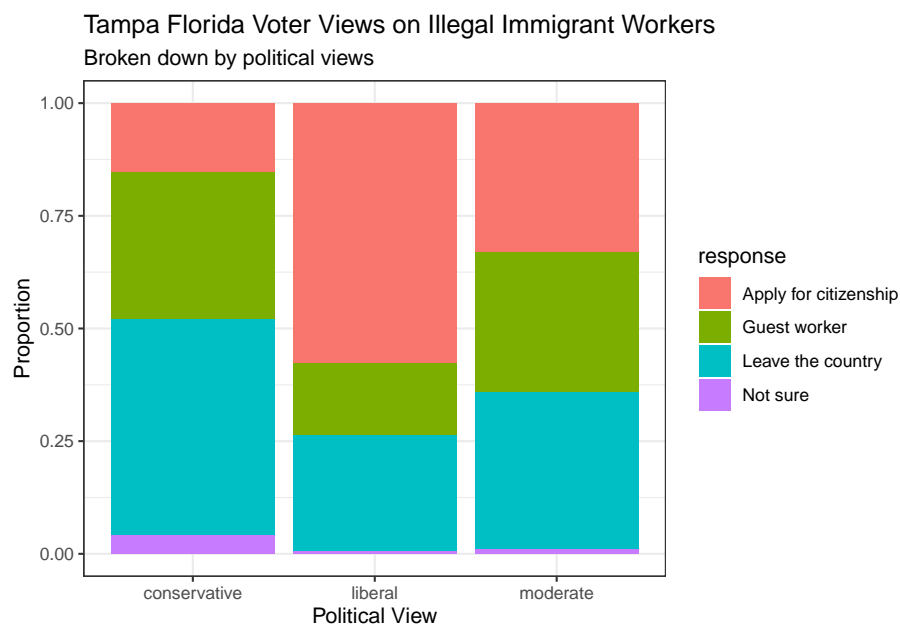
```
round(tally(response~political,data=immigration,format="percent",margins = TRUE),2)
```

	political		
response	conservative	liberal	moderate
Apply for citizenship	15.32	57.71	33.06
Guest worker	32.53	16.00	31.13
Leave the country	48.12	25.71	34.71
Not sure	4.03	0.57	1.10
Total	100.00	100.00	100.00

Of the conservative voters, 15.32% are in favor of the citizenship option. The numbers are 57.71% for liberals and 33.06% for moderates.

- g. Create a stacked bar chart.

```
immigration %>%
  gf_props(~political,fill=~response,position="fill") %>%
  gf_labs(title="Tampa Florida Voter Views on Illegal Immigrant Workers",
          subtitle="Broken down by political views",x="Political View",y="Proportion") %>%
  gf_theme(theme_bw())
```



- h. Using your plot, do political ideology and views on immigration appear to be independent? Explain your reasoning.

The percentages of Tampa, FL conservatives, moderates, and liberals who are in favor of illegal immigrants working in the US staying and applying for citizenship are quite different from one another. Therefore, the two variables appear to be dependent.

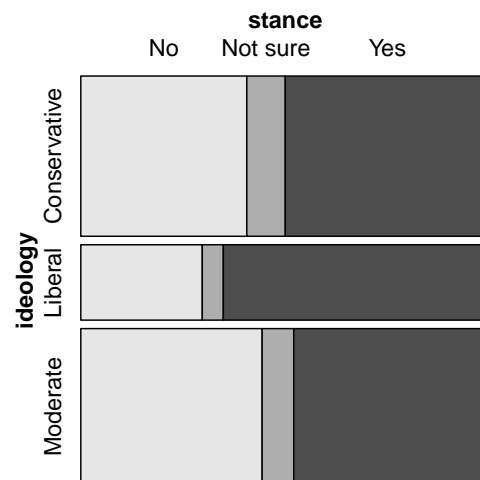
6.2.2 Problem 2

2. **Views on the DREAM Act** The same survey from Exercise 1 also asked respondents if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children.

The data is in the `openintro` package in the `dream` data object.

- a. Create a **mosaic** plot.

```
mosaic(stance~ideology,data=dream,sub="Voter views on illegal worker status")
```



Voter views on illegal worker status

- b. Based on the mosaic plot, are views on the DREAM Act and political ideology independent?

The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates the variables are dependent.

6.2.3 Problem 3

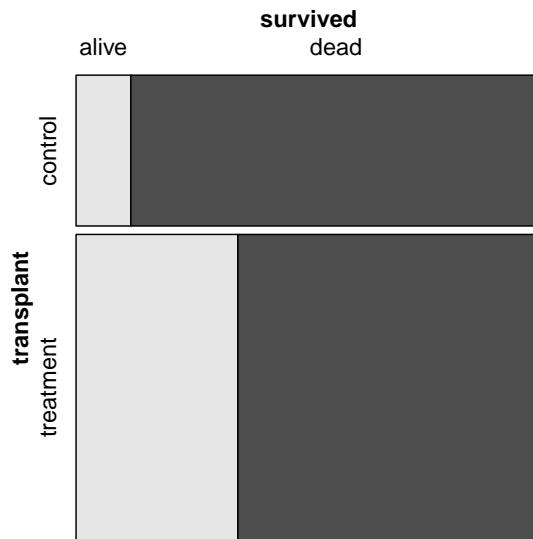
3. Heart transplants

The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable *transplant* indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called *survived* was used to indicate whether or not the patient was alive at the end of the study.

The data is in the `openintro` package and is called `heart_transplant`.

- a. Create a **mosaic** plot.

```
mosaic(survived~transplant,data=heart_transplant)
```

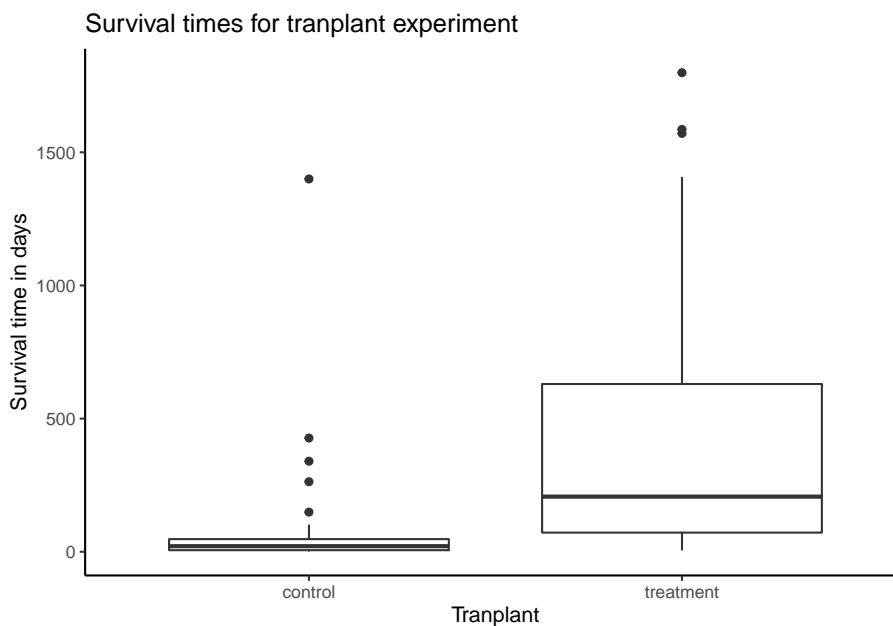


- b. Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.

Proportion of patients who are alive at the end of the study is higher in the treatment group than in the control group. These data suggest that survival is not independent of whether or not the patient got a transplant.

- c. Using *survtime* create side-by-side boxplots for the control and treatment groups.

```
heart_transplant %>%
  gf_boxplot(survtime~transplant) %>%
  gf_labs(title="Survival times for tranplant experiment",
          sub="Treatment group had the transplant",x="Tranplant",y="Survival time in days") %>%
  gf_theme(theme_classic())
```



- d. What do the box plots suggest about the efficacy (effectiveness) of transplants?

The shape of the distribution of survival times in both groups is right skewed with one very clear outlier for the control group and other possible outliers in both groups on the high end. The median survival time for the control group is much lower than the median survival time for the treatment group; patients who got a transplant typically lived longer. Tying this together with the much lower variability in the control group, evident by a much smaller IQR than the treatment group (about 50 days versus 500 days), and we can see that

patients who did not get a heart transplant tended to consistently die quite early relative to those who did have a transplant. Overall, very few patients without transplants made it beyond a year while nearly half of the transplant patients survived at least one year. It should also be noted that while the first and third quartiles of the treatment group is higher than those for the control group, the IQR for the treatment group is much bigger, indicating that there is more variability in survival times in the treatment group.

Part II

Probability Modeling

Chapter 7

Case Study

7.1 Objectives

- 1) Use R to simulate a probabilistic model.
- 2) Use basic counting methods.

7.2 Homework

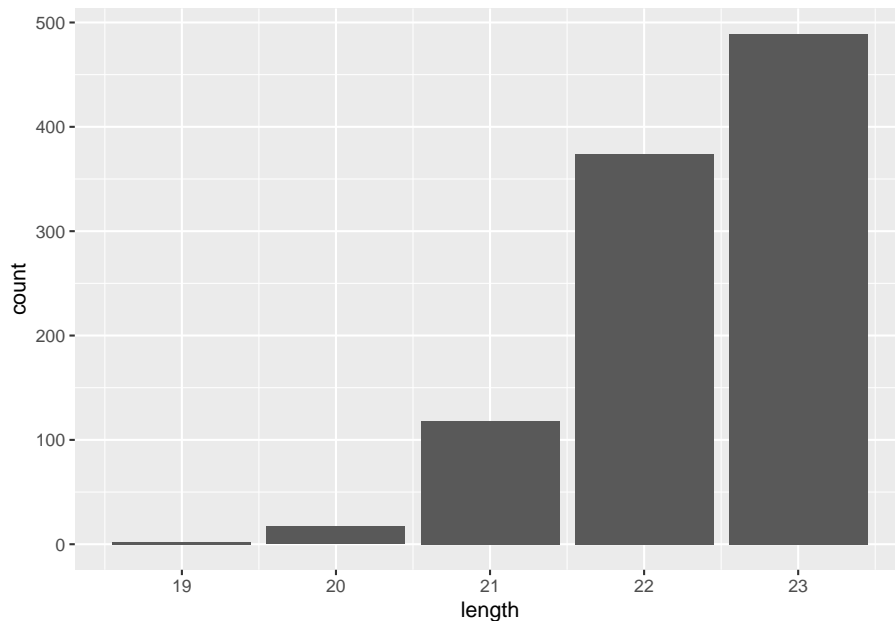
7.2.1 Problem 1

1. **Exactly 2 people with the same birthday - Simulation.** Complete a similar analysis for case where exactly 2 people in a room of 23 people have the same birthday. In this exercise you will use a computational simulation.
 - a. Create a new R Markdown file and create a report. Yes, we know you could use this file but we want you to practice generating your own report.
 - b. Simulate having 23 people in the class with each day of the year equally likely. Find the cases where exactly 2 people have the same birthday, you will have to alter the code from the Notes more than changing 18 to 23.
 - c. Plot the frequency of occurrences as a bar chart.
 - d. Estimate the probability of exactly two people having the same birthday.

```
(do(10000)*length(unique(sample(days,size=23,replace = TRUE)))) %>%
  mutate(match=if_else(length==22,1,0)) %>%
  summarise(prob=mean(match))
```

```
##      prob
## 1 0.362
```

```
(do(1000)*length(unique(sample(days,size=23,replace = TRUE)))) %>%
  gf_bar(~length)
```



7.2.2 Problem 2

2. **Exactly 2 people with the same birthday - Mathematical.** Repeat problem 1 but do it mathematically. As a big hint, you will need to use the `choose()` function. The idea is that with 23 people we need to choose 2 of them to match. We thus need to multiply, the multiplication rule again, by `choose(23,2)`. If you are having trouble, work with a total of 3 people in the room first.
 - a. Find a formula to determine the exact probability of exactly 2 people in a room of 23 having the same birthday.

- b. Generalize your solution to any number n people in the room and create a function.
- c. Vectorize the function.
- d. Plot the probability of exactly 2 people having the same birthday versus number of people in the room.
- e. Comment on the shape of the curve and explain it.
- f. knit and compile your report.

For two people we have

```
choose(23,2)*prod(365:344)/365^23
```

```
## [1] 0.3634222
```

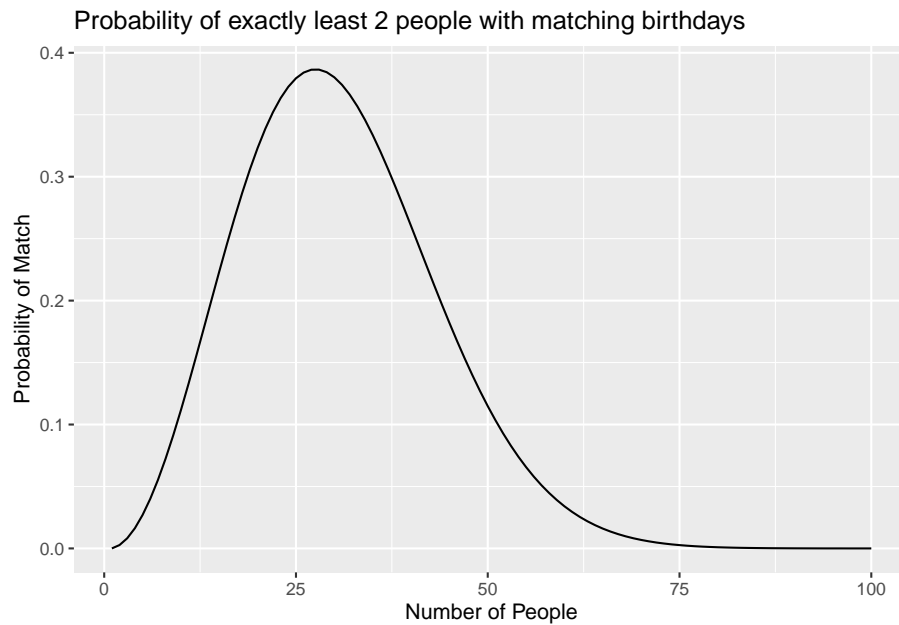
```
exactly_two <- function(n){  
  choose(n,2)*prod(365:(365-(n-2)))/365^n  
}
```

```
exactly_two(23)
```

```
## [1] 0.3634222
```

```
exactly_two <- Vectorize(exactly_two)
```

```
gf_line(exactly_two(1:100)~ seq(1,100),  
        xlab="Number of People",  
        ylab="Probability of Match",  
        title="Probability of exactly least 2 people with matching birthdays")
```



By the way, exactly three matches in simulation is hard. We have to table the data

```
set.seed(10)
temp <- table(sample(days,size=23,replace = TRUE))
temp
```

```
##
##  13  24  50  72  92 110 137 143 154 155 211 231 263 271 285 330 338 342 344 351
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   2   1   1   1
## 365
##   1
```

```
(sum(temp==2) == 2)+0
```

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

```
(do(10000)*((sum(table(sample(days,size=23,replace = TRUE)) == 3)==1)+0)) %>%
  summarise(prob=mean(result))
```

```
##      prob
## 1 0.0117
```


Two sets that have same but different birthday

```
(do(10000)*((sum(table(sample(days,size=23,replace = TRUE)) == 2)==2)+0)) %>%
  summarise(prob=mean(result))
```

```
##      prob
## 1 0.1139
```

```
(do(10000)*length(unique(sample(days,size=23,replace = TRUE)))) %>%
  mutate(match=if_else(length==21,1,0)) %>%
  summarise(prob=mean(match))
```

```
##      prob
## 1 0.1187
```

Mathematically exactly 3 is easy. Simulation seems to be off a little or the math formula is off.

```
choose(23,3)*prod(365:345)/365^23
```

```
## [1] 0.007395218
```


Chapter 8

Probability Rules

8.1 Objectives

- 1) Define and use properly in context all new terminology related to probability to include but not limited to: outcome, event, sample space, probability.
- 2) Apply basic probability and counting rules to find probabilities.
- 3) Describe the basic axioms of probability.
- 4) Use R to calculate and simulate probabilities of events.

8.2 Homework

8.2.1 Problem 1

1. Let A , B and C be events such that $P(A) = 0.5$, $P(B) = 0.3$, and $P(C) = 0.4$. Also, we know that $P(A \cap B) = 0.2$, $P(B \cap C) = 0.12$, $P(A \cap C) = 0.1$, and $P(A \cap B \cap C) = 0.05$. Find the following:

- a. $P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.5 + 0.3 - 0.2 = 0.6$$

- b. $P(A \cup B \cup C)$

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \\ &= 0.5 + 0.3 + 0.4 - 0.2 - 0.12 - 0.1 + 0.05 = 0.83 \end{aligned}$$

c. $P(B' \cap C')$

$$\begin{aligned} P(B' \cap C') &= P((B \cup C)') = 1 - P(B \cup C) = 1 - [P(B) + P(C) - P(B \cap C)] \\ &= 1 - (0.3 + 0.4 - 0.12) = 0.42 \end{aligned}$$

d. $P(A \cup (B \cap C))$

$$P(A \cup (B \cap C)) = P(A) + P(B \cap C) - P(A \cap B \cap C) = 0.5 + 0.12 - 0.05 = 0.57$$

e. $P((A \cup B \cup C) \cap (A \cap B \cap C)')$

$$P((A \cup B \cup C) \cap (A \cap B \cap C)') = P(A \cup B \cup C) - P(A \cap B \cap C) = 0.83 - 0.05 = 0.78$$

8.2.2 Problem 2

2. Consider the example of the family in the reading. What is the probability that the family has at least one boy?

$$P(\text{at least one boy}) = 1 - P(\text{no boys}) = 1 - P(\text{GGG}) = 1 - \frac{1}{8} = 0.875$$

8.2.3 Problem 3

3. The Birthday Problem Revisited.

- a. Suppose there are $n = 20$ students in a classroom. My birthday, the instructor, is April 3rd. What is the probability that at least one student shares my birthday? Assume only 365 days in a year and assume that all birthdays are equally likely.

$$P(\text{at least one other person shares my bday}) = 1 - P(\text{no one else has my bday}) =$$

$$1 - \left(\frac{364}{365}\right)^{20} = 0.0534$$

- b. In R, find the probability that at least one other person shares my birthday for each value of n from 1 to 80. Plot these probabilities with n on the x -axis and probability on the y -axis. At what value of n would the probability be at least 50%?

Generalizing,

$$P(\text{at least one other person shares my bday}) = 1 - P(\text{no one else has my bday}) = 1 - \left(\frac{364}{365}\right)^n$$

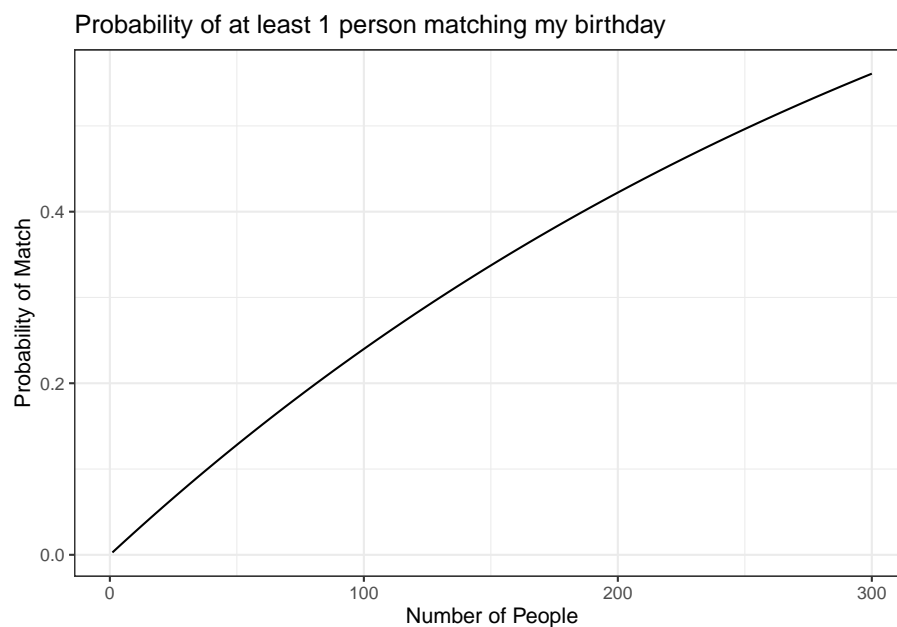
```
n<-1:300
mybday<-function(x) 1-(364/365)^x
mybday <- Vectorize(mybday)
```

Check our function.

```
mybday(20)
```

```
## [1] 0.05339153
```

```
gf_line(mybday(n) ~ n,
        xlab="Number of People",
        ylab="Probability of Match",
        title="Probability of at least 1 person matching my birthday") %>%
  gf_theme(theme_bw)
```



```
prob <- mybday(n)
which(prob>= .5)
```

```
## [1] 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271
## [20] 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290
## [39] 291 292 293 294 295 296 297 298 299 300
```

So 253 people.

8.2.4 Problem 4

4. Thinking of the cards again. Answer the following questions:

a. Define two events that are mutually exclusive.

The first card drawn is red.
The first card drawn is black.

b. Define two events that are independent.

The first card drawn is black.
The first card drawn is a face card.

c. Define an event and its complement.

The first card drawn is less than 5.
The first card drawn is equal to or more than 5.

8.2.5 Problem 5

5. Consider the license plate example from the reading.

a. What is the probability that a license plate contains **exactly** one “B”?

```
#fourth spot
num4<-10*10*10*1*25*25

#fifth spot
num5<-10*10*10*25*1*25

#sixth spot
num6<-10*10*10*25*25*1

denom<-10*10*10*26*26*26

(num4+num5+num6)/denom
```

```
## [1] 0.1066796
```

- b. What is the probability that a license plate contains **at least one** “B”?

$$1 - P(\text{no B's})$$

```
num0<-10*10*10*25*25*25
1-num0/denom
```

```
## [1] 0.1110036
```

8.2.6 Problem 6

6. Consider the party example in the reading.
- a. Suppose 8 people showed up to the party dressed as zombies. What is the probability that all three awards are won by people dressed as zombies?

$$\frac{8 \cdot 7 \cdot 6}{25 \cdot 24 \cdot 23}$$

```
(8*7*6)/(25*24*23)
```

```
## [1] 0.02434783
```

- b. What is the probability that zombies win “most creative” and “funniest” but not “scariest”?

$$\frac{8 \cdot 17 \cdot 7}{25 \cdot 24 \cdot 23}$$

```
(8*17*7)/(25*24*23)
```

```
## [1] 0.06898551
```

8.2.7 Problem 7

7. Consider the cards example from the reading.
- a. How many ways can we obtain a “two pairs” (2 of one number, 2 of another, and the final different)?

We have to pick the rank of the two pairs.

$$\binom{13}{2}$$

Notice here the order does matter because a pair of Kings and 4s is the same as a pair of 4s and Kings. This is different from the full house example. Make sure you understand this point.

Now we have to pick two of the four cards for each rank

$$\binom{4}{2} \binom{4}{2}$$

And finally we need the last card to come from the 44 remaining cards so that we don't get a full house.

$$\binom{44}{1}$$

Putting it all together:

$$\binom{13}{2} \binom{4}{2} \binom{4}{2} \binom{44}{1}$$

```
choose(13,2)*choose(4,2)*choose(4,2)*choose(44,1)
```

```
## [1] 123552
```

- b. What is the probability of drawing a “four of a kind” (four cards of the same value)?

$$P(4 \text{ of a kind}) = \frac{\binom{13}{1} \binom{4}{4} \binom{48}{1}}{\binom{52}{5}}$$

```
(13*1*48)/choose(52,5)
```

```
## [1] 0.000240096
```

8.2.8 Problem 8

8. Advanced Question: Consider rolling 5 dice. What is the **probability** of a pour resulting in a full house?

First pick the value for the three of a kind, there are 6. Then pick the value from the remaining 5 for the two of a kind. This is actually a permutation. There are 30 distinct “flavors” of full house (three 1’s & two 2’s, three 1’s & two 3’s, etc.). In the reading we did this as

$$\binom{6}{1} \times \binom{5}{1}$$

We now have the 5 dice. We have to select three to have the same value and the order doesn’t matter since they are the same value. Thus we multiple by $\binom{5}{3}$. Divide this by the total distinct ways the dice could have landed (assuming order matters).

$$P(\text{full house}) = \frac{30 \times \frac{5!}{3!2!}}{6^5}$$

$$P(\text{full house}) = \frac{\binom{6}{1} \times \binom{5}{1} \times \binom{5}{3}}{6^5}$$

```
30*10/(6^5)
```

```
## [1] 0.03858025
```

Simulating is tough so let’s write some code that may help.

```
set.seed(23)
temp<-table(sample(1:6,size=5,replace=TRUE))
temp
```

```
##
## 1 3 4 5
## 1 2 1 1
```

```
sum(temp==2) & sum(temp==3)
```

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

```
temp<-c(1,1,1,2,2)
temp<-table(temp)
temp
```

```
## temp
## 1 2
## 3 2
```

```
sum(temp==2) & sum(temp==3)
```

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

Let's write a function.

```
full_house <-function(x){  
  temp<-table(x)  
  sum(temp==2) & sum(temp==3)  
}
```

```
temp<-c(1,1,1,2,2)  
full_house(temp)
```

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

```
set.seed(751)  
results<-do(10000)*full_house(sample(1:6,size=5,replace=TRUE))  
mean(~full_house,data=results)
```

```
## [1] 0.039
```