Classical and Bayesian Statistics

Problems 1

Problem 1.1

(∗∗) a) The following table lists the civilian deaths in WW1 and WW2. We consider the Allied death in WW2. What is *very* problematic about that part of the table?

Hint: Consider the total number of death.

CIVILIANS

(a) World War I - Not known

(b) World War II

Allied

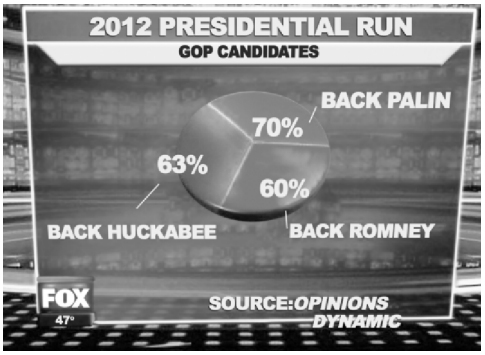
United Kingdom	60,595
Belgium	90,000
China	An enormous number
Denmark	Unknown
France	152,000
Netherlands	242,000
Norway	3,638
U.S.S.R.	6,000,000
	6,548,233

Enemy

Germany	800,000
Austria	125,000
Italy	180,000
Japan	600,000
Poland	5,000,000
Yugoslavia	Large number
	6,705,000

(∗∗) b) In the following figure, a presidential election prediction in the USA is shown in a so-called *pie chart*.

What is problematic about this specific graphic?



(∗∗) c) A little bit of critical thinking*.

I) A fictitious lady called Linda is described as follows: Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Which alternative is more probable?

i) Linda is a bank teller.

ii) Linda is a bank teller and is active in the feminist movement.

Explain your answer.

II) An individual has been described by a neighbour as follows:

"Steve is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail".

Which alternative is more probable?

i) Steve is a librarian.

ii) Steve is a farmer.

Explain your answer.

III) A bat and ball cost \$1.10. The bat costs one dollar more than the ball. How much does the ball cost?

i) \$0.20 ii) \$0.10 iii) \$0.05

IV) What causes more deaths?

i) Heart diseases.

ii) All accidents combined.

(∗ ∗ ∗) d) A true story from WW II†.

So here's the question. You don't want your planes to get shot down by enemy fighters, so you armor them. But armor makes the plane heavier, and heavier planes are less maneuverable and use more fuel. Armoring the planes too much is a problem; armoring the planes too little is a problem. Somewhere inbetween there's an optimum.

The military came to the SRG (Statistical research group; a highly influential US think tank in WWII) with some data they thought might be useful to find this optimum. When American planes came back from engagements over Europe, they were covered in bullet holes. But the damage wasn't uniformly distributed

*From: *Thinking, fast and slow*; Daniel Kahneman
†From: *How not to be wrong*; Jordan Ellenberg

across the aircraft. There were more bullet holes in the fuselage, not so many in the engines (see Table 1).

Section of the plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of the plane	1.8

Table 1: Bullets in plane

The officers saw an opportunity for efficiency; you can get the same protection with less armor if you concentrate the armor on the places with the greatest need, where the planes are getting hit the most, i.e., the fuselage and asked the SRG But exactly how much more armor belonged on those parts of the plane.

The officers didn't get the answer they expected. They should strengthen the engine they were told. Can you explain why?

Problem 1.2

In the Wikipedia article

https://en.wikipedia.org/wiki/Heights_of_presidents_and_presidential_candidates_of_the_United_States

the body height of US presidents and their challengers in the presidential elections are listed (see Table 2). It was mentioned that the taller candidate typically wins the elections.

In this exercise, we examine the data of the presidential elections since they are broadcast on television.

We create two vectors for the corresponding heights

```
winner <- c(193, 183, 191, 185, 185, 182, 182, 188, 188, 188, 185, 185, 177,
            182, 182, 193, 183, 179, 179, 175)
opponent <- c(163, 191, 165, 187, 175, 193, 185, 187, 188, 173, 180, 177, 183,
              185, 180, 180, 182, 178, 178, 173)
```

a) Determine the length of the two vectors. This way you can check if there are the same number of entries in both vectors.

b) Determine the entries 6. to 10. entry of the vector winner. Use the square brackets notation winner[...] for this purpose.

c) Determine the 3rd, 5th and 10th to 12th entry.

Year	Winner	Height	Opponent	Height
2024	Donald Trump	191 cm	Kamala Harris	163 cm
2020	Joe Biden	183 cm	Donald Trump	191 cm
2016	Donald Trump	191 cm	Hillary Clinton	165 cm
2012	Barack Obama	185 cm	Mitt Romney	187 cm
2008	Barack Obama	185 cm	John McCain	175 cm
2004	George W. Bush	182 cm	John Kerry	193 cm
2000	George W. Bush	182 cm	Al Gore	185 cm
1996	Bill Clinton	188 cm	Bob Dole	187 cm
1992	Bill Clinton	188 cm	George H. W. Bush	188 cm
1988	George H. W. Bush	188 cm	Michael Dukakis	173 cm
1984	Ronald Reagan	185 cm	Walter Mondale	180 cm
1980	Ronald Reagan	185 cm	Jimmy Carter	177 cm
1976	Jimmy Carter	177 cm	Gerald Ford	183 cm
1972	Richard Nixon	182 cm	George McGovern	185 cm
1968	Richard Nixon	182 cm	Hubert Humphrey	180 cm
1964	Lyndon B. Johnson	193 cm	Barry Goldwater	180 cm
1960	John F. Kennedy	183 cm	Richard Nixon	182 cm
1956	Dwight D. Eisenhower	179 cm	Adlai Stevenson	178 cm
1952	Dwight D. Eisenhower	179 cm	Adlai Stevenson	178 cm
1948	Harry S. Truman	175 cm	Thomas Dewey	173 cm

Table 2: Heights of presidents and their challengers since 1948

d) The Washington Post found out that the entries for Bill Clinton (8th and 9th entry) are too small. He has a height of 189 cm. Change the entries in the vector winner accordingly and output the new vector again.

e) The claim is that the winners are taller than their challengers. Check this by comparing the means of the vectors.

f) Determine the average height *differences*.

g) Determine the variance s^2 and the standard deviation s of the vector winner.

h) Determine these values without using the implemented functions (for practicing the handling of R). The variance is defined by

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n - 1}$$

Problem 1.3

In a class, the following grades were achieved in a statistics test:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

Alter three grades in the dataset such that the median remains the same, but the mean decreases as significantly as possible.

Use the sort(...) command to do this.

Classical and Bayesian Statistic

Sample solution for Problems 1

Solution 1.1

a) The main problem is that the total number of death is accurate to the dead, but only of the *known* numbers. The dead of China with "An enormous number" and Denmark with "Unknown" were not included. Accordingly, the total number cannot be correct, especially not accurate to the dead.

Another problem is the indication of the numbers. Norway had 3638 death, an exact figure, while the Soviet Union had 6 000 000 death, which must be an estimate.

Giving the total number *accurate* to the individual dead makes no sense. This is a pretence of accuracy that is not existing.

b) The percentages do not add up to 100 %, but almost 200 %. The reason for this is that the participants in the poll could indicate more than one candidate. However, this creates the problem that the candidates are very difficult to compare with each other.

c) I) Reading the description of Linda one would expect that the second answer is correct, because it "fits" the description better.

However, this argument does not hold. The description tells us nothing about Linda's job, so we have to rely on the whole population and there are more bank tellers than banktellers, which are *also* active in the feminist movement.

II) Also Steve sounds like your typical local librarian, we know nothing more about Steve. And as there are way more farmers than librarians, we have to conclude that Steve is more likely a farmer.

III) The usual initial answer is \$0.10 which is wrong. As we skim the statement, the word "more" is overlooked, so \$0.05 is correct.

IV) This is a problem of perception. We read or hear quite a lot about accidents because they are more or less spectacular. Heart diseases, however, are publicly almost never heard of. And there are more death by heart diseases than accidents*.

d) The SRG's insight was simply to ask: where are the *missing* holes? The ones that would have been all over the engine casing, if the damage had been spread *equally* all over the plane (the numbers give are per square foot and should be the same all over the plane)?

*See for example https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

The missing bullet holes were on the missing planes. The reason planes were coming back with fewer hits to the engine is that planes that got hit in the engine weren't coming back. Whereas the large number of planes returning to base with a thoroughly Swiss-cheesed fuselage is pretty strong evidence that hits to the fuselage can (and therefore should) be tolerated.

If you go the recovery room at the hospital, you'll see a lot more people with bullet holes in their legs than people with bullet holes in their chests. But that's not because people don't get shot in the chest; it's because the people who get shot in the chest don't recover.

In cognitive science, this phenomenon is called the *survival bias* and is everywhere in our thinking. Recently, the Covid vaccination opponents have said that the vaccination does not work because they know someone vaccinated who has contracted Covid-19. No word about the very high proportion of the population where the vaccin *did* work.

Solution 1.2

a) We create two vectors for the corresponding body heights

```
winner <- c(193, 183, 191, 185, 185, 182, 182, 188, 188, 188, 185, 185, 177,
            182, 182, 193, 183, 179, 179, 175)
opponent <- c(163, 191, 165, 187, 175, 193, 185, 187, 188, 173, 180, 177, 183,
              185, 180, 180, 182, 178, 178, 173)
```

a) Length of the two vectors

```
length(winner)
[1] 20

length(opponent)
[1] 20
```

b) 6th to 10th entries of the vector winner

```
winner[6:10]
[1] 182 182 188 188 188
```

c) 3rd, 5th and 10th to 12th entries.

```
winner[c(3, 5, 10:12)]
[1] 191 185 188 185 185
```

d) Change 8th and 9th entries:

```
winner[8] <- 189
winner[9] <- 189

# or winner[c(8,9)] <- 189

winner
[1] 193 183 191 185 185 182 182 189 189 188 185 185 177 182 182 193 183 179
    179 175
```

e) Compare the mean values of the vectors

```
mean(winner)
[1] 184.35

mean(opponent)
[1] 180.15
```

f) Average of height differences:

```
diff <- winner - opponent

mean(diff)
[1] 4.2
```

The winner of the election is thus on average about 4.2 cm taller.

g) Variance s^2 and the standard deviation s of the vector winner.

```
var(winner)
[1] 25.08158

sd(winner)
[1] 5.008151
```

h) Without using the implemented functions.

```
winner_var <- sum((winner - mean(winner))^2) / (length(winner)-1)
winner_var
[1] 25.08158

winner_sd <- sqrt(winner_var)
winner_sd
[1] 5.008151
```

Solution 1.3

The original dataset have the following values:

```
grades_1 <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6,4, 3.7, 5, 5.2, 4.5,
              3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1)
```

For the median and mean we obtain:

```
median(grades_1)
[1] 4.65

mean(grades_1)
[1] 4.5125
```

First we reorder the data values according to their size:

```
grades_2 <- sort(grades_1)
grades_2
[1] 2.3 2.4 2.8 3.3 3.6 3.7 3.9 4.0 4.2 4.2 4.5 4.5 4.8 4.9 5.0 5.0 5.1 5.2 5.5 5.6 5.9
    5.9 6.0 6.0
```

Since the number of grades is even, the median is formed from the mean of $x_{(12)}$ and $x_{(13)}$. Thus, if we change grades smaller than $x_{(12)}$, the median will not change. Accordingly, we change the grade values of $x_{(9)}, x_{(10)}, x_{(11)}$ in the most extreme way, namely to 1. This leaves the median unchanged, but decreases the average value as much as possible.

```
grades_2[c(9,10,11)] <- 1
grades_2
[1] 2.3 2.4 2.8 3.3 3.6 3.7 3.9 4.0 1.0 1.0 1.0 4.5 4.8 4.9 5.0 5.0 5.1 5.2 5.5 5.6
    5.9 5.9 6.0 6.0

sort(grades_2)
[1] 1.0 1.0 1.0 2.3 2.4 2.8 3.3 3.6 3.7 3.9 4.0 4.5 4.8 4.9 5.0 5.0 5.1 5.2 5.5 5.6
    5.9 5.9 6.0 6.0

median(grades_2)
[1] 4.65

mean(grades_2)
[1] 4.1
```