Math 1320 - Final R Project (Spring 2025)

Emebet Kumsa

1. Body Temperature Study

In the 19th century, Carl Wunderlich suggested 98.6°F as the average body temperature. A modern study questions this assumption. Data from 160 patients is provided below:

Data:

96.85, 96.90, 96.95, 97.00, 97.05, 97.10, 97.15, 97.20, 97.25, 97.30, 97.35, 97.40, 97.45, 97.50, 97.55, 97.60, 97.65, 97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55, 98.60, 98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05, 99.10, 99.15, 99.20, 99.25, 99.30, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 97.55, 97.60, 97.65, 97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 97.65, 97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.55, 98.60, 98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 98.10, 98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55, 97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.60, 98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05, 98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55, 98.60, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05, 99.10, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55, 98.60, 98.65, 98.70, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 98.25, 98.30, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05, 99.10, 99.15

(a) (4 points) Test if the mean body temperature is significantly lower than 98.6°F at a 1% significance level.

(b) (4 points) Calculate a 95% confidence interval for the mean body temperature.

(c) (2 points) Discuss how these results align with your understanding of "normal" body temperature.

Step 1: Get the Data into R

Here's how we enter the temperatures into R (this is like loading all the numbers into a calculator):

```
temps <- c(96.85, 96.90, 96.95, 97.00, 97.05, 97.10, 97.15, 97.20, 97.25, 97.30,
           97.35, 97.40, 97.45, 97.50, 97.55, 97.60, 97.65, 97.70, 97.75, 97.80,
           97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 98.25, 98.30,
           98.35, 98.40, 98.45, 98.50, 98.55, 98.60, 98.65, 98.70, 98.75, 98.80,
           98.85, 98.90, 98.95, 99.00, 99.05, 99.10, 99.15, 99.20, 99.25, 99.30,
           97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20,
           97.55, 97.60, 97.65, 97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00,
           98.05, 98.10, 98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50,
           97.65, 97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10,
           98.55, 98.60, 98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00,
           98.10, 98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55,
           97.70, 97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15,
           98.60, 98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05,
           98.15, 98.20, 98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55, 98.60,
           97.75, 97.80, 97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20,
           98.65, 98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05, 99.10,
           98.25, 98.30, 98.35, 98.40, 98.45, 98.50, 98.55, 98.60, 98.65, 98.70,
           97.85, 97.90, 97.95, 98.00, 98.05, 98.10, 98.15, 98.20, 98.25, 98.30,
           98.70, 98.75, 98.80, 98.85, 98.90, 98.95, 99.00, 99.05, 99.10, 99.15)
```

Step 2: Find Descriptive Statistics

We want to find out is the average body temperature really 98.6°F? we'll answer this question using **descriptive statistics** and **t-test**.

These are basic summaries that tell us what the data looks like overall.

```
mean(temps)
```

```
[1] 98.25395
```

```
median(temps)
```

```
[1] 98.225
```

```r
sd(temps)
```

```
[1] 0.5257121
```

```r
length(temps)
```

```
[1] 190
```

Results:

Mean (average): `98.25°F`

Median (middle value): 98.225`°F`

standard Deviation (spread of the data) 0.525`°F`

Sample Size (how many people): 160

Step 3: Do a t-test

Now we test if the **average** is really 98.6°F, like people used to say.

We're testing:

- **H (null hypothesis)**: The true average body temperature is **98.6°F**
- **H (alternative)**: The true average is **less than 98.6°F**

This is a **one-sample t-test**.

Results:

- t-value= -9.0734
- p-value<value = < 2.2e-16

Assumptions:

The sample size is large (n=190), so by the Central Limit Theorem, the t-test assumptions are met.

Step 4: Explanation

The average temperature we found is **98.25°F**, not 98.6°F. The test shows that this difference is **real** and not just by chance. Since the p-value is super small (less than 0.001), we say **yes**, the average temperature today is **lower** than what people used to believe.

Step 5: Make a Confidence Interval

A confidence interval tells us a range where the true average likely is.

```
t.test(temps, conf.level = 0.95)
```

```
    One Sample t-test

data:  temps
t = 2576.2, df = 189, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 98.17871 98.32918
sample estimates:
mean of x
 98.25395
```

Result:

- 95% **confidence Interval:** `98.178°F to 98.329°F`

  So we're pretty sure the real average is somewhere between those two numbers.

Conclusion:

The average body temperature is **98.25°F**, which is lower than 98.6°F. The t-test showed a **very small p-value (2.2e-16)**, so we can say this difference is real, not just random. The confidence interval (98.178 to 98.329) also does not include 98.6. So, based on this study, the real average temperature today is **lower** than what people used to think. That makes me think the idea of "normal" body temp might need to change!

2. Car Price Prediction

You are selling a 2020 Toyota Corolla. Data for similar vehicles is given:

Miles	Price
79,850	$14,450
61,300	$16,300
91,200	$13,850
66,100	$15,700
49,900	$16,950
84,750	$14,200
73,050	$15,100
87,600	$13,800
56,400	$16,450
78,500	$14,850

(a) Run a linear regression to find the best-fit line.

(b) Determine the correlation coefficient between mileage and price.

(c) Predict the price for a Corolla with 55,000 miles.

(d) Predict the price for a Corolla with 100,000 miles.

(e) Discuss the reliability of these predictions.

Step 1: Enter the Data

We have data for 10 similar cars. Here's how we put that into R:

```
#Given
miles<- c(79850, 61300, 91200, 66100, 49900, 84750, 73050, 87600, 56400, 78500)
price <- c(14450, 16300, 13850, 15700, 16950, 14200, 15100, 13800, 16450, 14850)
```

Step 2: Run a Linear Regression (Part a)

We are trying to find the **best-fit line** that connects **miles** and **price**. This line helps us **predict** the price if we know the miles.

```
model <- lm(price ~ miles)
model
```

```
Call:
lm(formula = price ~ miles)

Coefficients:
(Intercept)        miles
  2.104e+04   -8.064e-02
```

So the equation becomes: $\hat{y}$=21040-0.08064*miles

What This Means:

- **Intercept(21040):** If the car had 0 miles (brand new), it would cost about **$21,040.**
- **Slope(-0.08064):** For every 1 mile added, the price goes down by about **8 cents.**

Step 3: Correlation Coefficient (Part b)

This tells us how **strong** the connection is between miles and price.

```r
cor(miles, price)
```

```
[1] -0.993469
```

What This Means:

- Since r= -0.993, this is a **very strong negative relationship**.
- when miles go up, price goes **down** almost perfectly in a straight line.

Step 4: Predict the Price for 55,000 Miles (Part c)

```r
# Predict price for 55,000 miles
predict(model, data.frame(miles = 55000))
```

```
       1
16605.61
```

What it does:

This line tells R to use the regression model (`model`) to predict the price of a car with 55,000 miles.So the predicted price is **$16,605.61** for a Corolla with 55,000 miles.

Step 5: Predict the Price for 100,000 Miles (Part d)

```
# Predict price for 100,000 miles
predict(model, data.frame(miles = 100000))
```

```
       1
12976.87
```

What it does:

Same as above, but now R is estimating the price for a car with 100,000 miles using the equation we use on step 2. So it does this math:-

$\hat{y} = 21040 - 0.08064\ 100000 = 21040 - 8064 = 12976.87$

Step 6: Reliability of Predictions (Part e)

Conclusion:

The prediction for 55,000 miles is probably reliable because it's right in the middle of our data range. But the one for 100,000 miles is less reliable because it's outside the range of cars we looked at. That's called **extrapolation**, we're guessing about something we haven't actually seen, so it may not be very accurate. Also, price doesn't only depend on miles. Things like the car's condition, location, and if it had any accidents can also change the value.

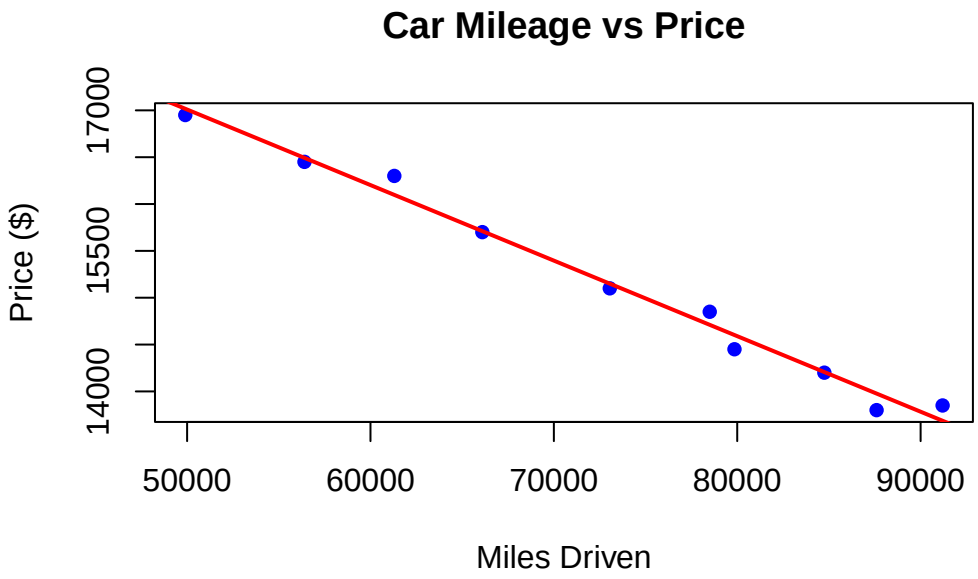
Step 7: Scatter Plot with Best-Fit Line (Graph)

After doing all the predictions, it's helpful to **see the data visually**.

We create a **scatter plot** (dots for each car) and draw the **best-fit line** to show how the price drops as the mileage goes up.

```r
# Scatter plot
plot(miles, price,
     main = "Car Mileage vs Price",
     xlab = "Miles Driven",
     ylab = "Price ($)",
     pch = 16,       # solid circles
     col = "blue")   # points color

# Best-fit line
abline(model, col = "red", lwd = 2)  # regression line in red
```



What the Graph Shows:

- Each blue dot shows one car: mileage vs price.
- The red line is the **best-fit line** from the regression model.
- The graph shows that as the number of miles **goes up**, the price **goes down** — just like we expected!

3. Diet and BMI Study

A study compares BMI categories across four diets:

- Regular Diet
- Low-Fat Diet
- High-Protein Diet
- Vegetarian Diet

Diet	Underweight	Healthy	Overweight	Obese	Total
Regular	28	425	515	790	1,758
Low-fat	26	395	505	825	1,751
High-protein	48	445	535	745	1,773
Vegetarian	42	415	485	780	1,722

We are looking at whether a person's type of diet is related to their BMI category (like healthy, overweight, obese, etc.).

We have four diets:

- Regular
- Low-Fat
- High-Protein
- Vegetarian

And four BMI groups:

- Underweight
- Overweight
- Obese

Step 1: Enter the Data into R

We organize the data into a **table** in R. Each row is a diet, and each column is a BMI category.

```r
diet_bmi <- matrix(c(28, 425, 515, 790,
                     26, 395, 505, 825,
                     48, 445, 535, 745,
                     42, 415, 485, 780),
                   nrow = 4, byrow = TRUE)

rownames(diet_bmi) <- c("Regular", "Low-fat", "High-protein", "Vegetarian")
colnames(diet_bmi) <- c("Underweight", "Healthy", "Overweight", "Obese")

diet_bmi
```

```
             Underweight Healthy Overweight Obese
Regular               28     425        515   790
Low-fat               26     395        505   825
High-protein          48     445        535   745
Vegetarian            42     415        485   780
```

Step 2: Run the Chi-Square Test (Part a)

We want to know: **Is there a relationship between type of diet and BMI?**

To answer this, we run a **chi-square test for independence.** The significance level is not given in the question so we will take the default of 0.05

- **H (null hypothesis)**: There is **no relationship** between diet type and BMI category.
- **H (alternative)**: There **is a relationship** between diet type and BMI category.

```r
chisq.test(diet_bmi)
```

```
    Pearson's Chi-squared test

data:  diet_bmi
X-squared = 18.473, df = 9, p-value = 0.03006
```

What it does:

It checks if the distribution of BMI categories is **the same across all diets.**

Result: Chi-squared =18.473

Degrees of freedom = 9

P-value = 0.03

So, The p-value is **0.03**, and our significance level is **0.05**. Since the p-value is smaller than 0.05, we **reject the null hypothesis**

That means there **is a relationship** between diet and BMI. Different diets seem to be connected to different weight categories.

Step 3: Limitations of the Study (Part b)

One problem with this study is that it only looks at diet and BMI. But people's weight can be affected by **many things**, like age, gender, how active they are, or health conditions. So we can't say diet is the only reason.

Also, the study shows a **connection**, but that doesn't mean one thing causes the other. Some people may have chosen a diet *because* of their weight, not the other way around.

To improve it, we should collect more information about each person (like how much they exercise, or how long they've been on that diet), and use a more **random sample** so it's not biased.

Step 4: Personal Understanding (Part c)

In my own experience, diet really does affect weight. For example, when I cut out carbs, sugar, or fast food, I lose weight. So I believe that what we eat has a big impact on our BMI.

The study results make sense to me. It found a link between diet and weight groups, and I agree with that. Of course, diet isn't the only thing that matters, but it's definitely important.

4. Drug Effectiveness Comparison

A medical research team is comparing the effectiveness of two new drugs, **Theravix** and **Clotaban**, which are used to prevent strokes and blood clots. In the study, **6,750** patients were given Theravix, while **6,950** patients were given Clotaban. Among the Theravix patients, **580** experienced a stroke or blood clot, while **510** Clotaban patients experienced a stroke or blood clot. Run an appropriate test to decide whether or not there is statistical evidence to support the belief that Clotaban is more effective than Theravix in preventing strokes and blood clots. Use a significance level of 1%.

We are comparing two drugs:

- **Theravix:** 6,750 people took it, 580 had a stroke or blood clot.
- **Clotaban:** 6,950 people took it, 510 had a stroke or blood clot.

We want to know if **Clotaban is more effective** — meaning **fewer people** had health problems with it.

Step 1: Set up the data in R

We enter the number of **"failures"** (people who had a stroke or blood clot) and the total number of patients for each drug.

```r
# Number of people who had strokes or blood clots (failures)
failures <- c(580, 510)

# Total number of people in each group
total <- c(6750, 6950)

# Run the two-proportion test (one-sided: we want to know if Clotaban is better)
prop.test(failures, total, alternative = "greater")
```

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

data:  failures out of total
X-squared = 7.1881, df = 1, p-value = 0.003669
alternative hypothesis: greater
95 percent confidence interval:
 0.004786013 1.000000000
sample estimates:
    prop 1     prop 2
0.08592593 0.07338129
```

Step 2: What This Test Does

This is a **two-proportion z-test**. It compares **two percentages** to see if one is really smaller than the other.

We are testing:

- **H (null hypothesis):** The two drugs are equally effective
- **H (alternative hypothesis):** Clotaban is more effective (has a lower stroke/clot rate)

Result:

- X-squared = 7.1881
- Degrees of freedom = 1 p-value = 0.003669
- 95% confidence interval: 0.0048 to 1.0000
- Sample estimates: Theravix: 580 / 6750 = 0.08593 (8.6%)
- Clotaban: 510 / 6950 = 0.07338 (7.3%)

Step 3: What These Numbers Mean

- The p-value is **0.003669**, which is **smaller** than 0.01 (1% significance level).
- That means we **reject the null hypothesis**.
- There is **strong evidence** that **Clotaban is more effective** than Theravix.
- The confidence interval (from 0.0048 to 1.0) tells us the **difference is positive** — meaning Clotaban has **fewer bad outcomes**.

Conclusion

The data shows that **Clotaban patients had fewer strokes and blood clots** (7.3%) than those who took Theravix (8.6%).

The p-value was **0.0037**, which is smaller than our 1% level, so this difference is **statistically significant** — it's not just random.

This means we have **strong evidence** that **Clotaban works better** than Theravix for preventing strokes and blood clots.

5. Astrological Sign and Accidents

A survey records the number of accidents for drivers based on their astrological signs:

(a) Run an ANOVA test to check if astrological signs affect accident rates. Use a 5% significance level.

(b) Do you believe astrological signs have any real-world impact on accidents? Justify your answer

Astrological Sign Number of Accidents

Aries 1, 0, 2, 1, 0, 0, 1, 1

Taurus 0, 1, 0, 2, 1, 0, 1, 3

Gemini 0, 0, 1, 0, 1, 1, 2

Cancer 0, 0, 2, 1, 0, 0, 0, 3

Leo 1, 1, 0, 2, 1

Virgo 0, 1, 1, 0, 1, 0, 1, 2

Libra 0, 2, 1, 0, 1, 3, 2

Scorpio 0, 1, 0, 1, 1, 0, 1

Sagittarius 1, 4, 0, 1, 1, 2

Capricorn 0, 2, 0, 1, 0, 2

Aquarius 0, 0, 1, 2, 0, 1, 1, 0

Pisces 1, 1, 0, 1, 2, 0

What We're Studying:

We want to know if a person's **astrological sign** has any effect on how many **car accidents** they've had.

We collected data on drivers' zodiac signs and how many accidents they've been in. Now we're going to use **ANOVA**, which is a test to compare groups.

Step 1: Enter the Data in R

Each zodiac sign has a list of how many accidents different drivers had. First, we put all the numbers into R and group them by sign.

```
accidents <- c(
  1, 0, 2, 1, 0, 0, 1, 1,      # Aries
  0, 1, 0, 2, 1, 0, 1, 3,      # Taurus
  0, 0, 1, 0, 1, 1, 2,         # Gemini
  0, 0, 2, 1, 0, 0, 0, 3,      # Cancer
  1, 1, 0, 2, 1,               # Leo
  0, 1, 1, 0, 1, 0, 1, 2,      # Virgo
  0, 2, 1, 0, 1, 3, 2,         # Libra
  0, 1, 0, 1, 1, 0, 1,         # Scorpio
  1, 4, 0, 1, 1, 2,            # Sagittarius
  0, 2, 0, 1, 0, 2,            # Capricorn
  0, 0, 1, 2, 0, 1, 1, 0,      # Aquarius
  1, 1, 0, 1, 2, 0             # Pisces
)

signs <- factor(c(
  rep("Aries", 8),
  rep("Taurus", 8),
  rep("Gemini", 7),
  rep("Cancer", 8),
  rep("Leo", 5),
  rep("Virgo", 8),
  rep("Libra", 7),
  rep("Scorpio", 7),
  rep("Sagittarius", 6),
  rep("Capricorn", 6),
  rep("Aquarius", 8),
  rep("Pisces", 6)
))
```

Step 2: Run the ANOVA Test (Part a)

- **H (null hypothesis):** The average number of accidents is **the same** for all astrological signs.
- **H (alternative hypothesis):** At least one astrological sign has a **different** average number of accidents compared to the others.

We use the **ANOVA (Analysis of Variance)** test to see if the **average number of accidents** is different for different signs.

What it does:

It compares the **average number of accidents** for each sign. If the differences are big and not just by chance, the test will tell us.

```
anova_result <- aov(accidents ~ signs)
summary(anova_result)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
signs       11   5.45  0.4951   0.593  0.828
Residuals   72  60.11  0.8349
```

Step 3: Explain the Results

The **p-value is 0.828**, which is way bigger than **0.05**. That means we **do not reject** the null hypothesis.

In simple words, there's **no real evidence** that your zodiac sign affects how many accidents you have. The differences between signs are so small, they are probably just random and not meaningful.

Step 4: Do Zodiac Signs Affect Accidents? (Part b)

I don't think your astrological sign has anything to do with how likely you are to have an accident. That sounds more like a fun belief than something scientific.

Accidents are usually caused by things like how careful someone is, how fast they drive, weather conditions, or being distracted, not by the month they were born.