Week 3 Lab Statistical Computing

I'm going to try something different this week for the lab. Instead of giving you an R notebook, you'll have to create one yourself. At the end try knitting it to HTML and see does it display correctly.
Important: You are not expected to complete all questions in this lab. Choose 5-6 exercises that interest you across different test types. Focus on understanding the concepts rather than completing every single problem.

1 Some Distributions (Quick Warm-up)

Do this section quickly - it's just to refresh your memory about distributions before we use them to check test assumptions. Don't spend more than 10 minutes here.

In lectures, we looked at various discrete probability functions. These can be drawn using the following code (for a Poisson example)

```
barplot(dpois(0:15, 5), names=0:15)
```

and for a continuous density (eg. Normal) we can use

```
curve(dnorm(x, mean=1, sd=3), from=-8, to=10)
```

In R 'help(distributions)' will give a list of the distributions available in (base) R .

For each of the following, make an appropriate plot. These plots will help you visualize the distributions we'll assume when checking test assumptions later.

1. Poisson with mean 12
2. Negative Binomial with size 3 and prob 0.5
3. Normal with mean 10 and standard deviation 5
4. Gamma with shape 2 and scale 0.5
5. Chi-square with 4 degrees of freedom

2 Hypothesis Testing

For every test make sure you always do the following

- State the Null and Alternative Hypothesis
- State your level of significance
- Check assumptions (normality, equal variances where applicable)
- Choose appropriate test based on assumption checks
- Interpret the results correctly
- Write what the results means in English

3 Smaller Tests

For a basic start, we'll do a couple of smaller tests.

3.1 Study Time Before and After Intervention

A study skills workshop claims to improve student study habits. Ten students tracked their weekly study hours before and after attending the workshop. The data is:

Before Workshop	12	15	10	18	14	11	16	13	17	12
After Workshop	15	18	13	20	16	14	18	15	19	14

Use an appropriate hypothesis test to see if the workshop had a significant effect on study hours. Make sure to check the assumption of normality of the differences.

```
before <- c(12, 15, 10, 18, 14, 11, 16, 13, 17, 12)
after <- c(15, 18, 13, 20, 16, 14, 18, 15, 19, 14)
```

3.2 Mobile-phone usage

Suppose a study of mobile-phone usage for a user gives the following lengths for the calls

$$12.8, 3.5, 2.9, 9.4, 8.7, .7, .2, 2.8, 1.9, 2.8, 3.1, 15.8$$

What is an appropriate test for centre? First, look at a stem and leaf plot

```
x <- c(12.8, 3.5, 2.9, 9.4, 8.7, .7, .2, 2.8, 1.9, 2.8, 3.1, 15.8)
stem(x)
```

The distribution looks skewed with a possibly heavy tail - not normally distributed

Try the Shapiro-Wilk Normality test, you should see it is not normally distributed data.

A t-test is ruled out. Instead, a test for the median is done.

Suppose $H_0 : M = 5$, and the alternative is $H_1 : M > 5$.
To test this with R we can use the wilcox.test

Definition 1 *Wilcoxon Rank Sum and Signed Rank Tests Description Performs one- and two-sample Wilcoxon tests on vectors of data; the latter is also known as 'Mann-Whitney' test.*

3.3 Compare Schools

A researcher wants to compare the exam scores of students from two different schools. The scores from School A (n = 15) and School B (n = 15) are given below:

School A	School B
78	82
85	88
90	84
76	79
82	85
79	83
88	87
91	90
85	86
80	81
83	85
87	89
92	91
84	87
81	83

Perform an appropriate hypothesis test to check if there is a statistically significant difference with $\alpha = 0.05$ and interpret the results. Make sure you check the variances and normality. Boxplots would be my advice.

```
school_A <- c(78, 85, 90, 76, 82, 79, 88, 91, 85, 80, 83, 87, 92, 84, 81)
school_B <- c(82, 88, 84, 79, 85, 83, 87, 90, 86, 81, 85, 89, 91, 87, 83)
```

3.4 Placebo vs Drug

A medical researcher is testing whether a new drug affects blood pressure differently for two groups. The recorded blood pressure reductions (mmHg) are as follows:

Drug Group	Placebo Group
8.1	4.5
7.6	10.2
8.9	3.8
9.2	12.1
7.8	4.0
8.3	11.3
9.1	3.9
8.7	10.8
9.0	5.2
8.5	9.7
8.6	
9.3	

Perform an appropriate hypothesis test to check if there is a statistically significant difference with $\alpha = 0.05$ and interpret the results. Make sure you check the variances and normality using boxplots and Shapiro-Wilk tests.
Note: If one group appears less clearly normal than the other, consider: (1) the t-test is reasonably robust to moderate departures from normality, especially with similar sample sizes, or (2) you could use a non-parametric alternative like the Mann-Whitney U test (Wilcoxon rank-sum test).

```
drug_group <- c(8.1, 7.6, 8.9, 9.2, 7.8, 8.3, 9.1, 8.7, 9.0, 8.5, 8.6, 9.3)
placebo_group <- c(4.5, 10.2, 3.8, 12.1, 4.0, 11.3, 3.9, 10.8, 5.2, 9.7)
```

3.5 Penalty Shoot Outs

Originally I was going to get you to download the dataset `https://www.kaggle.com/datasets/jandimovski/world-cup-penalty-shootouts-2022` and see if you could determine whether there is enough statistical evidence to say if there is a difference between going first or second in a penalty shootout. This ended up being too messy in how the data had to be read in so I'm just going to say

Of the 35 penalty shootouts in the World Cup from 1982-2022

- 17 were won by the team that went first
- 18 were won by the team that went second

Is there enough statistical evidence to say if there is an advantage to either way? With 0.05 level of significance. There are two tests here that are applicable - binom.test and prop.test. You could even do this using chi-squared test too, but pick one of the others.

3.6 Two surveys

A survey is taken two times over the course of two weeks. The pollsters wish to see if there is a difference in the results as there has been a new advertising campaign run. Here is the data

	Week 1	Week 2
Favorable	45	56
Unfavorable	35	47

The standard hypothesis test is $H_0 : \pi_1 = \pi_2$ against the alternative (two-sided) $H_1 : \pi_1 \neq \pi_2$. Like the previous one, this is a proportion test (prop.test) except it has two proportions so we need vectors, c() for first row and then c(firstrow+secondrow)

```
# Two surveys
c(45, 56), c(45+35, 56+47)
```

4 Chi-Squared Tests

For chi-squared tests, always check that the test requirements are met:

- All expected frequencies should be at least 1
- At least 80% of expected frequencies should be 5 or greater
- If these conditions are not met, consider combining categories or using an alternative test (like Fisher's exact test for 2x2 tables)

```
result <- chisq.test(data)
result$expected
```

Will allow you to see the expected table and check these frequencies.

4.1 Dice rolls

If we toss a die 150 times and find that we have the following distribution of rolls is the die fair?

face	1	2	3	4	5	6
Number of rolls	22	21	22	27	22	36

```
observed_dice <- c(22, 21, 22, 27, 22, 36)
```

4.2 Letters

The letter distribution of the 5 most popular letters in the English language is known to be approximately

letter	E	T	N	R	O
freq.	29	21	17	17	16

That is when either E,T,N,R,O appear, on average 29 times out of 100 it is an E and not the other 4. This information is useful in cryptography to break some basic secret codes. Suppose a text is analysed and the number of E,T,N,R and Os are counted. The following distribution is found

letter	E	T	N	R	O
freq.	100	110	80	55	14

Is there sufficient evidence (at the 0.05 significance level) to say whether the text was not from the English language?

```
expected_props <- c(29, 21, 17, 17, 16)/100
observed_letters <- c(100, 110, 80, 55, 14)
```

4.3 Star Trek

You wish to investigate whether the colour of a uniform worn by a Star Trek officer is independent from the number of deaths on the show. The follow data was gathered :

Colour	**No. Alive**	**No. Dead**	**Total**
Blue	135	10	145
Gold	19	10	29
Red	215	46	261
Totals	369	66	435

Carry out, in detail, an appropriate test to see if there is evidence at $\alpha = 0.05$ of a relationship between colour shirt an officer wears and how many of them live or die.

```
# Star Trek - create matrix
star_trek <- matrix(c(135, 10, 19, 10, 215, 46), nrow=3, byrow=TRUE)
rownames(star_trek) <- c("Blue", "Gold", "Red")
colnames(star_trek) <- c("Alive", "Dead")
```

4.4 Belt Example

		Injury Level			
		None	minimal	minor	major
Seat Belt	Yes	12,813	647	359	42
	No	65,963	4,000	2,642	3,037

```
yesbelt <- c(12813, 647, 359, 42)
nobelt <- c(65963, 4000, 2642, 3037)
```

5 Tests Using Datasets

These datasets are either available on Moodle or you need to download them from the linked site

5.1 Difference between car origins

`https://archive.ics.uci.edu/dataset/9/auto+mpg` is the dataset we will use.
One of the columns (origin) says where the car is from

1. USA
2. Europe
3. Japan

You will need to read this data in to a dataframe (check the data format using `read.table()` with `header=TRUE` and be aware of missing values (`na.strings`) marked as "?"). This may help after that

```
auto_data$origin <- factor(auto_data$origin, levels = c(1, 2, 3),
                           labels = c("USA", "Europe", "Japan"))
```

if you would prefer the rows to have USA/Europe/Japan instead of numbers and can get one column of one country with

```
usa_mpg <- auto_data$mpg[auto_data$origin == "USA"]
```

Perform multiple t-tests, to see if there are differences between USA vs Japan, USA vs Europe and Europe vs Japan cars in different categories. You choose the columns and permutations and then do a write up about your results.

5.2 Differences in Marks

Take the dataset from Moodle called `SampleDataSet2014.csv`. This describes multiple students in a university.

In particular it has placement test scores (out of 100 points) for four subject areas: English, Reading, Math, and Writing.

Perform t-tests to see if there is a statistically significant difference between English and Maths scores.

There were 409 cases with non-missing English scores, and 422 cases with non-missing Math scores, but only 398 cases with non-missing observations for both variables. You will only want to take into consideration the ones with both Maths and English scores

When you've done that, do similar with say English and Writing and whatever other combinations you want.

5.3 Smokers

This goes back to the `SampleDataSet2014.csv` again. Suppose we want to test for an association between smoking behaviour (nonsmoker, current smoker, or past smoker) and gender (male or female) (we'll use $\alpha = 0.05$).

You will need to read in the data and make a cross-tab so that you have a contingency table. Try it out