# Lab 5- Conditional Probability

This lab contains some numbered questions. You are required to submit:

1. One pdf document (Lab 5.pdf) that contains:
   - your written answers to the question and any charts produced
   - the commands you used to answer the question, when asked

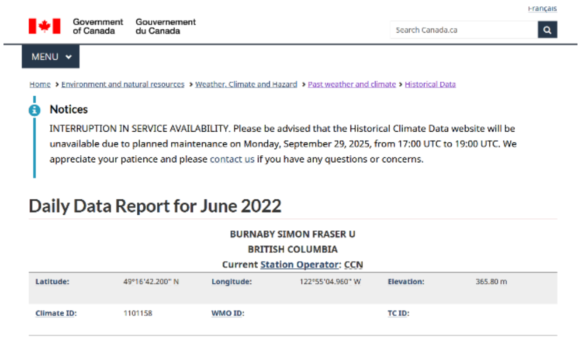
2. One R script (Lab 5.R) that contains all the code used to complete the lab. The script must run without errors.

Submit your files to the Lab 5 folder by 11:59pm two school days from now.

Packages required: **dplyr, mosaic, readxl**

## Introduction

Weather data collected daily at SFU from 1965-2022 is available on the Government of Canada website.


Français
Government of Canada
Gouvernement du Canada
Search Canada.ca
MENU
Home › Environment and natural resources › Weather, Climate and Hazard › Past weather and climate › Historical Data

Notices

INTERRUPTION IN SERVICE AVAILABILITY. Please be advised that the Historical Climate Data website will be unavailable due to planned maintenance on Monday, September 29, 2025, from 17:00 UTC to 19:00 UTC. We appreciate your patience and please contact us if you have any questions or concerns.

Daily Data Report for June 2022

BURNABY SIMON FRASER U<br>
BRITISH COLUMBIA<br>
Current Station Operator: CCN

| | | | | | |
|---|---|---|---|---|---|
| Latitude: | 49°16'42.200" N | Longitude: | 122°55'04.960" W | Elevation: | 365.80 m |
| Climate ID: | 1101158 | WMO ID: | | TC ID: | |

In this lab, you are going to examine this data set to answer questions such as:

- Is precipitation (rain/snow) more likely on Mondays?
- In which season is precipitation most likely?
- Do either high or low temperatures affect the probability of precipitation?
- Does the season influence the dependence of precipitation on temperature?

This will involve *filtering* the data, *segmenting* the data, and calculating *conditional probabilities* from data.

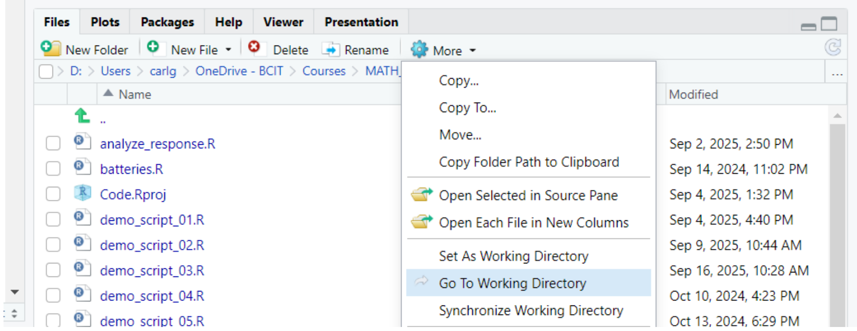
**This lab combines skills from Labs 1 to 4. If you forget something from those labs, you should review them.**

## Lab Procedure

For the purposes of dealing with files, RStudio designates a certain file system location as the “working directory”. It is a good idea to set the working directory appropriately. For instance, you could use:

```
> setwd("C:/Users/cgladish/OneDrive - BCIT/Courses/MATH_3042/Code")
```

When you import data files, file names are interpreted relative to the working directory. Verify that you have set it properly by clicking “Go To Working Directory”:


Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New File
Delete
Rename
More
D: > Users > carlg > OneDrive - BCIT > Courses > MATH
Name
Modified
Copy...
Copy To...
Move...
Copy Folder Path to Clipboard
Open Selected in Source Pane
Open Each File in New Columns
Set As Working Directory
Go To Working Directory
Synchronize Working Directory

| Name | Modified |
|---|---|
| .. | |
| analyze_response.R | Sep 2, 2025, 2:50 PM |
| batteries.R | Sep 14, 2024, 11:02 PM |
| Code.Rproj | Sep 4, 2025, 1:32 PM |
| demo_script_01.R | Sep 4, 2025, 4:40 PM |
| demo_script_02.R | Sep 9, 2025, 10:44 AM |
| demo_script_03.R | Sep 16, 2025, 10:28 AM |
| demo_script_04.R | Oct 10, 2024, 4:23 PM |
| demo_script_05.R | Oct 13, 2024, 6:29 PM |

Download the file “SFU.Weather.with.Precip.csv” (found on Learning Hub) and save it somewhere. For instance, suppose you save it to the following location (relative to the working directory)

```
> data.file <- "../Data/SFU.Weather.with.Precip.csv"
```

You can then import it into R using:

```
> SFU.Weather <- read.csv(data.file, colClasses = c("Date", "numeric", "chara
cter", "numeric", "numeric", "numeric", "numeric"))
```

The argument **colClasses** tells R the data type for each column in the csv file. The resulting data frame contains:

- **Date/Year/Month/Day**
- **Temp.deg.C** – daily temperature at SFU in degrees Celsius
- **Total.Precip.mm** – total rain and/or snow (melted) in mm

First, we will need to establish the order of the months, so run the command:

```
> months <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep",
"Oct", "Nov", "Dec")
```

This character vector will be used later when plotting bar charts.

We will need both **Temp.deg.C** and **Total.Precip.mm** not to be NA. Before any further calculations, apply a filter to reject rows where either variable is NA.

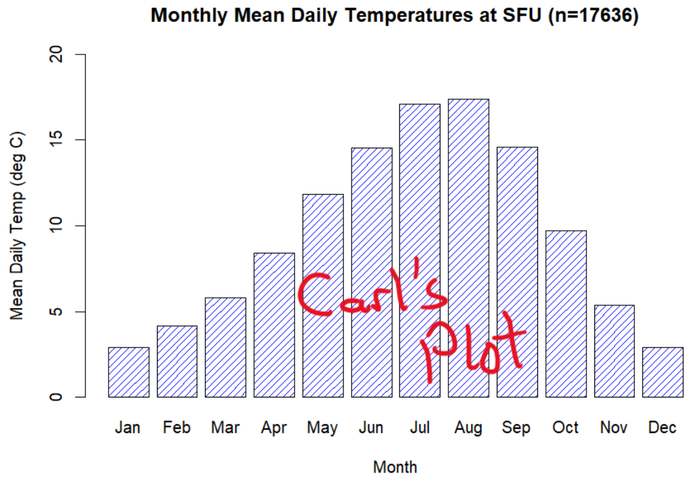
```
> SFU.Weather <- filter( SFU.Weather, !is.na(Temp.deg.C), !is.na(Total.Preci
p.mm))
```

1. In Lab 3 you learned how to calculate sample statistics for a numerical variable $X$ grouped by a categorical variable. Write code to generate mean and standard deviations of **Temp.deg.C** grouped by **Month**. (Save your code in your R script.)

```
> library(mosaic)
> Monthly.Mean.Temp <- mean(<something>)
> Monthly.SD.Temp    <-   sd(<something>)
```

Use the vector `Monthly.Mean.Temp` to produce a bar plot showing the mean daily temperature each month at SFU. Copy/paste the chart to your written answers. Which two months are warmest and which two are coldest?

```
> barplot(Monthly.Mean.Temp[months],<other arguments>)
```

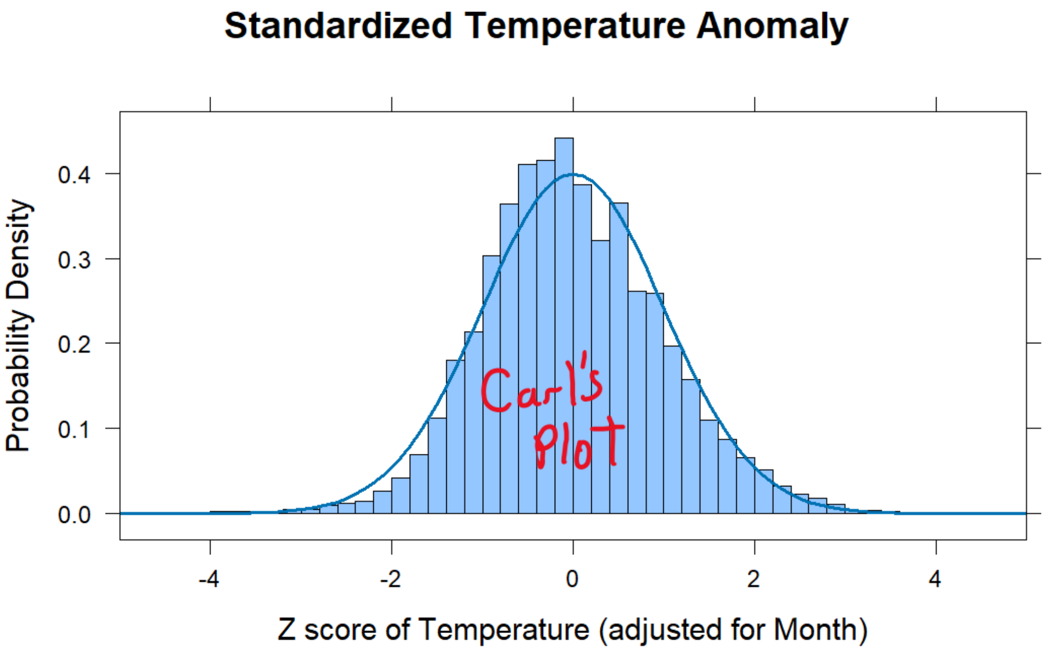


Next, we will use the monthly mean and standard deviation values to calculate a $Z$-score for temperature for each day. For each day, we will use the mean value for the month in which that day occurs since the mean changes greatly from month to month.

```
SFU.Weather$Z.Temp.Anomaly <-
  with(SFU.Weather,
       (Temp.deg.C - Monthly.Mean.Temp[Month]) /
          Monthly.SD.Temp[Month])
```

Note: using the R function **with()** allows you to access variables in **SFU.Weather** without the $ notation.

2. Plot a mosaic-style histogram of **Z.Temp.Anomaly** and overlay a normal distribution on top (using arguments **fit** and **type**) to verify that the $Z$-score we calculated is normally distributed. Copy/paste your chart into your written answers.



Next we will *segment* the data frame according to temperature anomaly $Z$-scores.

- *Cold* days are days where $Z < -0.4307$
- *Mid* days are days where $-0.4307 \leq Z < 0.4307$
- *Warm* days are days where $Z \geq 0.4307$

```
SFU.Weather$Temp.Seg <- ifelse( SFU.Weather$Z.Temp.Anomaly < -0.4307, "Cold",
                        ifelse( SFU.Weather$Z.Temp.Anomaly <= 0.4307, "Mid",
                                "Warm"))
```

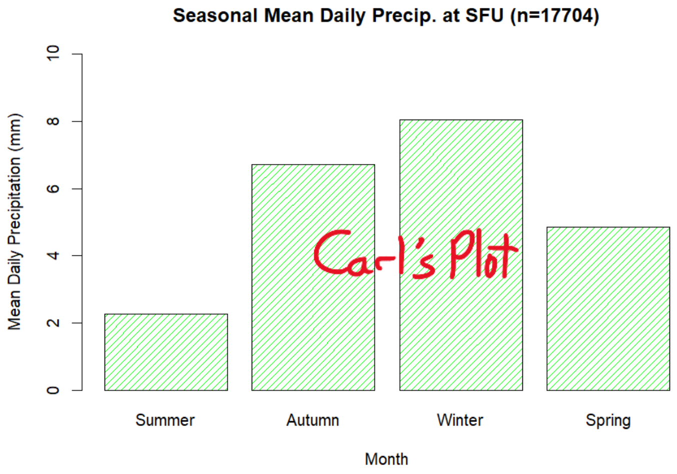
Check the help file for **ifelse()** if you don't see how it works from this example. (Note: for a standard normal distribution, the probability of $Z < -0.4307$ is 33.33%.)

3. Suppose you select a random day at SFU. Use the data (not assumptions) to calculate each of the probabilities:

```
• P.Cold <- <something> #Probability of a "Cold" day
• P.Mid  <- <something> #Probability of a "Mid" day
• P.Warm <- <something> #Probability of a "Warm" day
```

4. Write code that segments the data frame according to seasons.
   - *Summer* is Jun, Jul, Aug
   - *Autumn* is Sep, Oct, Nov
   - *Winter* is Dec, Jan, Feb
   - *Spring* is Mar, Apr, May

   (i.e, Create a variable called **Season** that is attached to the **SFU.Weather** data frame and contains either “Summer”, “Autumn”, “Winter”, or “Spring” for each day.)
   Use the variable **Season** to produce a seasonally averaged plot of Precipitation, as shown below.



We see that winter is the wettest season. Now we will investigate *which* days in winter are the wettest. The following code shows that 48.76% of Mondays have some precipitation (> 0 mm).

```
> SFU.Weather$Precipitation <- (SFU.Weather$Total.Precip.mm > 0)
> SFU.Weather$Day.of.Week <- weekdays( SFU.Weather$Date)
> dow.tab <- table(SFU.Weather[c("Day.of.Week","Precipitation")])
> dow.tab["Monday","TRUE"] / rowSums(dow.tab)["Monday"]
   Monday
0.4876297
```

5. Find the probability that a randomly selected day at SFU has some precipitation. Is the event of precipitation independent of the event that a day is Monday? Explain.

6. It seems plausible that colder days are wetter days (it *feels* like that, at least). Generate a table like **dow.tab** above that shows how many days had precipitation or not for each level of **Temp.Seg** (“Cold”, “Mid”, “Warm”). Use the table to find each of the conditional probabilities:

- $P(\text{Precip} > 0 \mid \text{Cold})$
- $P(\text{Precip} > 0 \mid \text{Mid})$
- $P(\text{Precip} > 0 \mid \text{Warm})$

Is the hypothesis that Cold days (for that month) are more likely to have precipitation supported by our data?

7. Use the probabilities you calculated in questions 3 and 6 along with Bayes' Theorem to calculate the conditional probability

$$P(\text{ Warm} \mid \text{Precip} > \ 0)$$

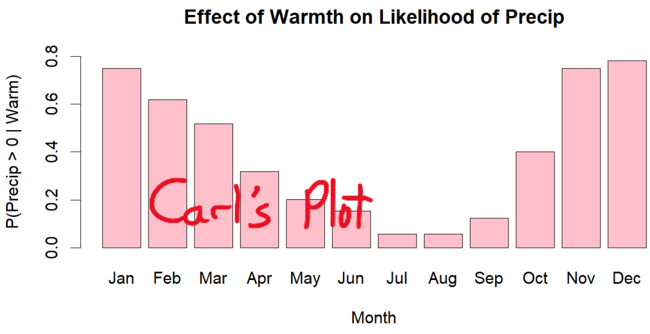
Record a suitable version of Bayes' Theorem and the details of your numerical calculation in your written solution.

8. Now calculate each of the following probabilities from the data. You will first need to make an appropriate frequency table using **table()**.

- $P(\text{ Precip} > 0 \mid \text{Warm and Summer})$
- $P(\text{ Precip} > 0 \mid \text{Warm and Autumn})$
- $P(\text{ Precip} > 0 \mid \text{Warm and Winter})$
- $P(\text{ Precip} > 0 \mid \text{Warm and Spring})$

Summarize your findings in two or three clear sentences. For instance, you could start with "In general, precipitation is less likely given warm temperatures. However, taking into account the effect of seasons, we find that ...."

9. Now we use **Month** as a conditioning variable. Which month has the highest probability of precipitation given that it is a Warm day? (In other words, in which month is the effect of warmth on precipitation the greatest?) Write R code that can generate the answer without manual calculation. Also, write code to generate a bar plot that shows this conditional probability for each month.


Effect of Warmth on Likelihood of Precip
P(Precip > 0 | Warm)
0.0
0.2
0.4
0.6
0.8
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Carl's Plot