QQ plots

Put your name here

Put the date here

Introduction

When we have numerical data, we are often interested in knowing if such data is normally distributed. One can always look at a histogram, of course. In this assignment we will learn about QQ plots, another method for assessing the normality of data. When sample sizes are small (rendering histograms less meaningful), a QQ plot will often be easier to read.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document and work back and forth between this R Markdown file and the PDF output as you work through this module.

When you are finished with the assignment, knit to PDF one last time, proofread the PDF file **carefully**, export the PDF file to your computer, and then submit your assignment.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

Be sure to remove the line `## Add code here to [do some task]...` when you have added your own code.

Sometimes you will be asked to type up your thoughts. That will appear in the document as follows:

Please write up your answer here.

Again, please be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. If you need to use some R code as well, you can use inline R code inside the block between `\begin{answer}` and `\end{answer}`, or if you need an R code chunk, please go outside the `answer` block and start a new code chunk.

Load Packages

We load the standard `mosaic` package.

```
library(mosaic)
```

QQ plots

All of the work we do with normal models assumes that a normal model is appropriate. When we want to summarize data using a normal model, this means that the data distribution should be reasonably unimodal, symmetric, and with no serious outliers.

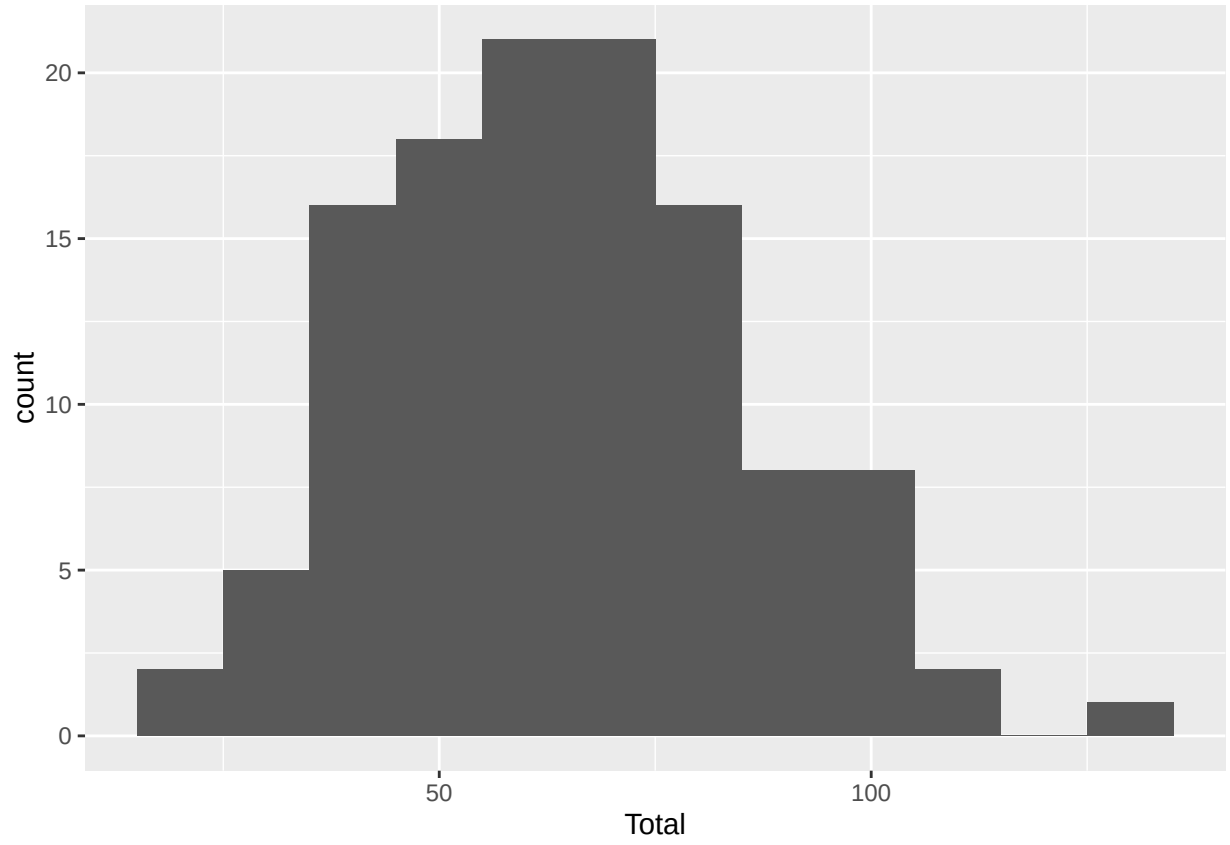
We can, of course, use a histogram to check this. But a histogram can be highly sensitive to the choice of bins. Furthermore, for small sample sizes, histograms look "chunky", making it hard to test this assumption.

An easier way to check normality is to use a *quantile-quantile plot*, typically called a *QQ plot* or sometimes a *normal probability plot*. We won't get into the technicalities of how this plot works. Suffice it to say that if data is normally distributed, the points of a QQ plot should lie along a diagonal line.

Here is an example. The total snowfall in Grand Rapids, Michigan has been recorded every year since 1893. This data is included with the `mosaic` package in the data frame `SnowGR`. A histogram (with reasonable binning) shows that the data is nearly normal.

```
ggplot(SnowGR, aes(x = Total)) +
    geom_histogram(binwidth = 10)
```

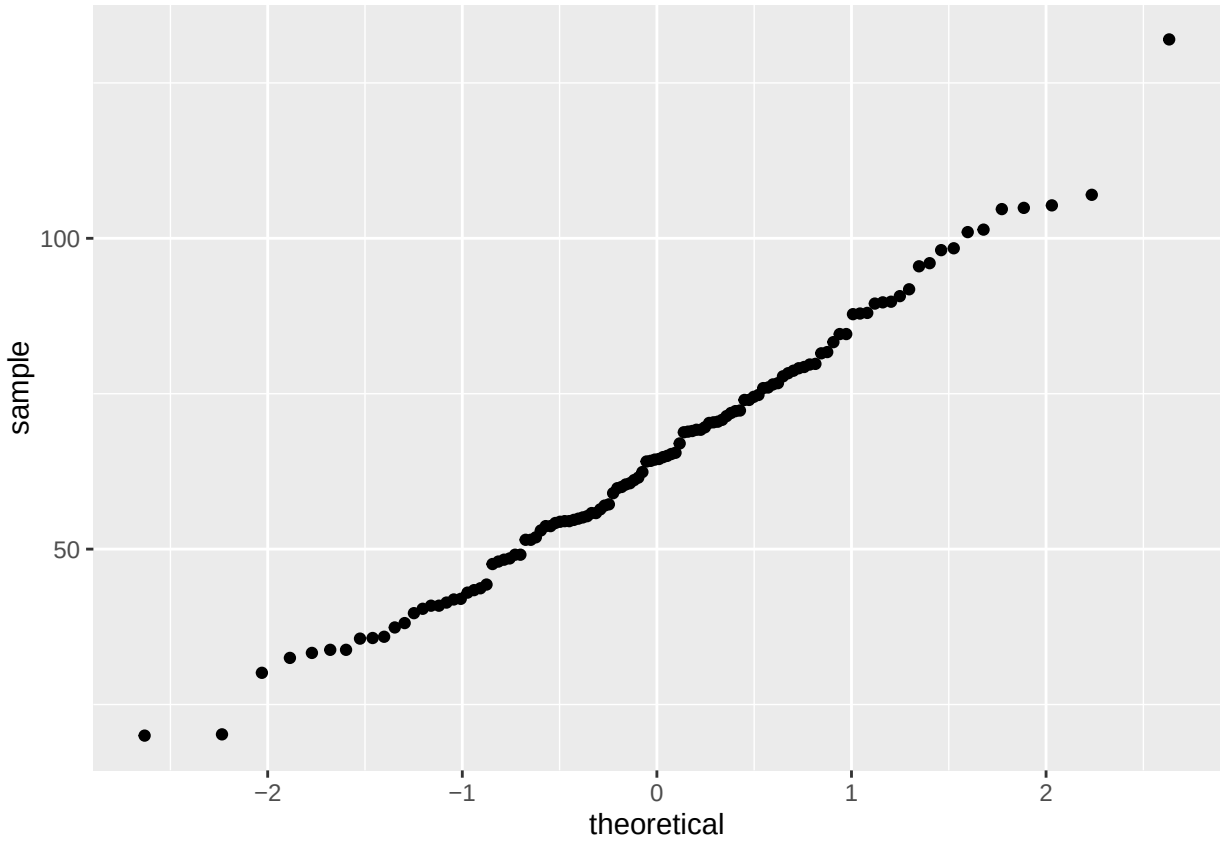
```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



Here is the QQ plot for the same data. Notice that the aesthetics are a little different. Instead of `x`, we have to use `sample`.

```
ggplot(SnowGR, aes(sample = Total)) +
    geom_qq()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_qq).
```

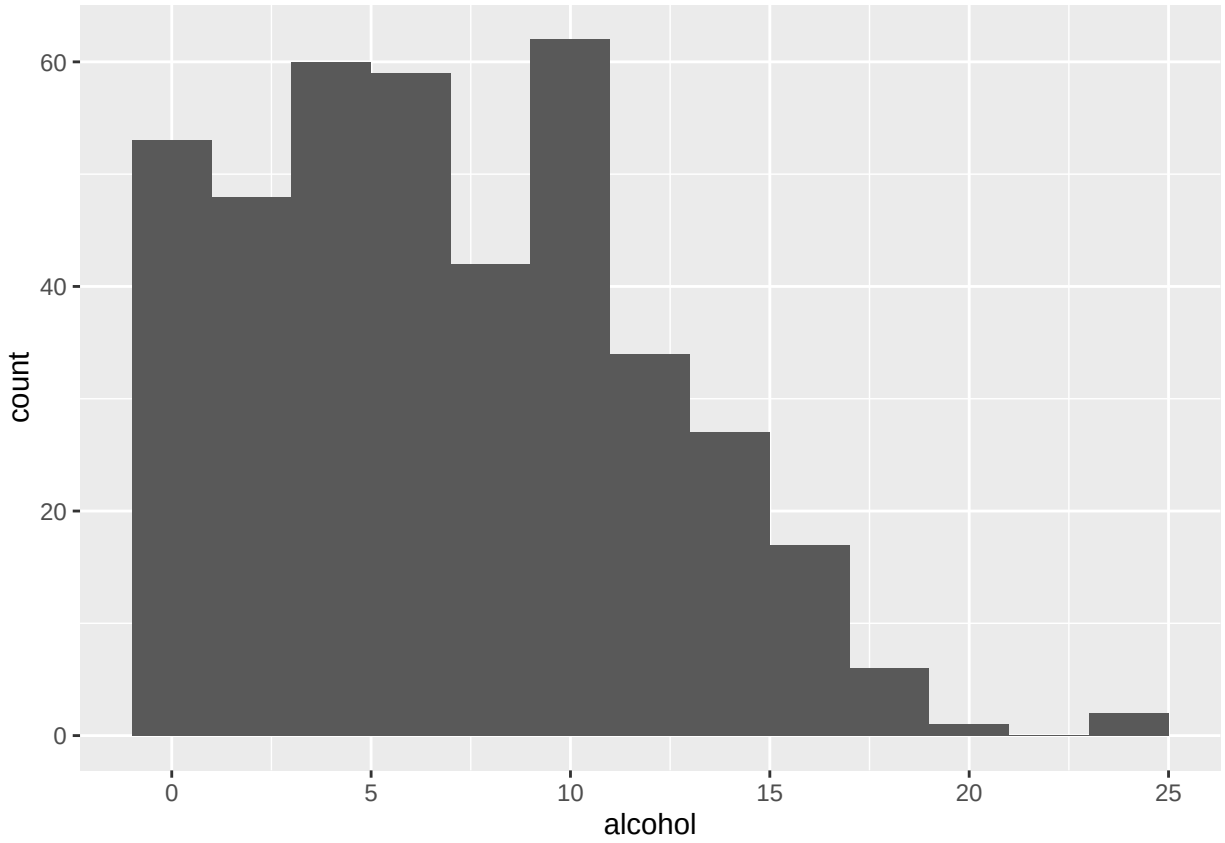


(The warning is because there is one missing value in the data.)

Other than a few points here and there, the bulk of the data is lined up nicely. There's a minor outlier, and that can be seen in both the histogram and the QQ plot.

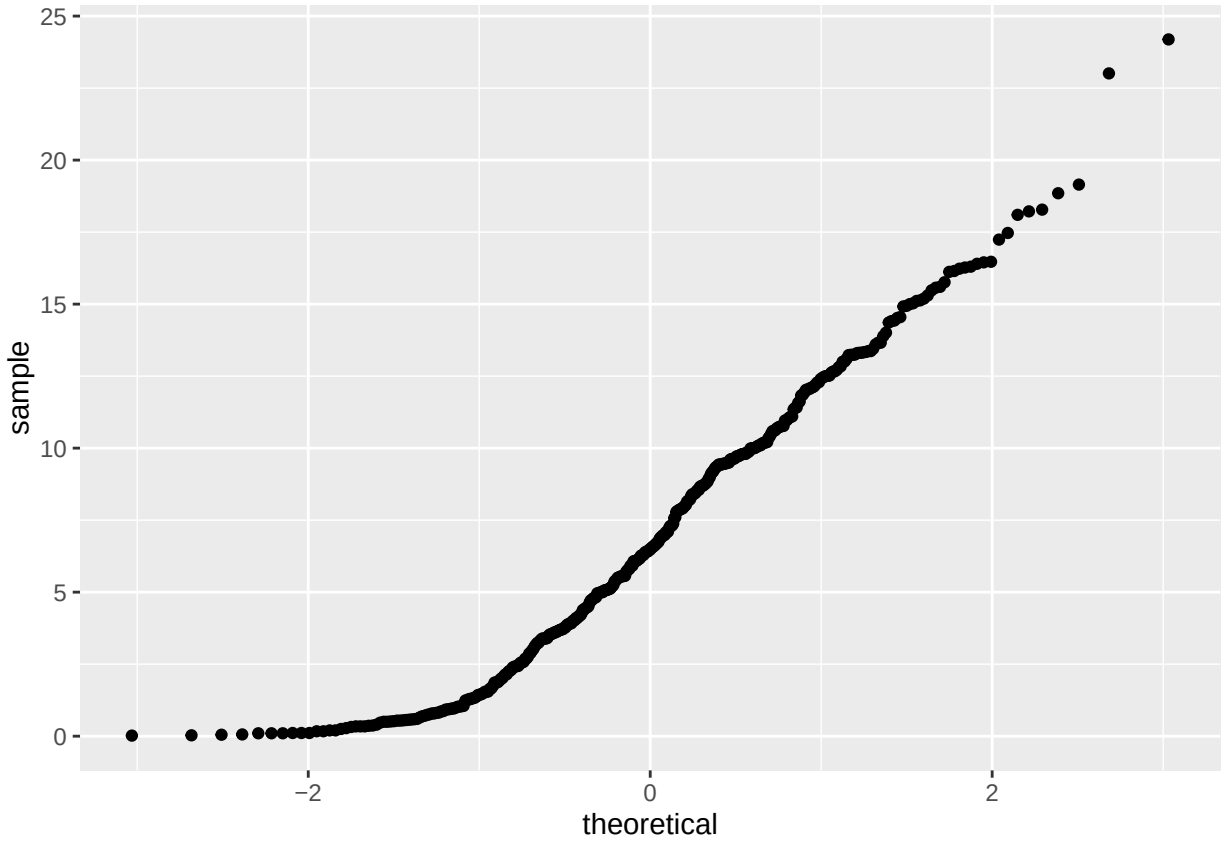
Contrast that with skewed data. For example, the `Alcohol` dataset contains per capita consumption (in liters) of alcohol for various countries over several years. The alcohol consumption variable is highly skewed, as one can see in the histogram.

```
ggplot(Alcohol, aes(x = alcohol)) +
    geom_histogram(binwidth = 2)
```



It is also apparent in the QQ plot that the data is not normally distributed.

```
ggplot(Alcohol, aes(sample = alcohol)) +
    geom_qq()
```



The path of dots is sharply curved, indicating a lack of normality.

Your turn

Find a data set with a numerical variable that is nearly normal in its distribution. (It can be something we've already seen in a past assignment, or if you're really ambitious, you're welcome to find a new data set.) Plot both a histogram and a QQ plot to demonstrate that the data is nearly normal. No need for a written response. Just plot the graphs.

```
## Add code here to plot a histogram and a QQ plot.
```

Now find a data set with a numerical variable that is skewed in its distribution. Plot both a histogram and a QQ plot to demonstrate that the data is not nearly normal. Again, no need for a written response. Just plot the graphs.

```
## Add code here to plot a histogram and a QQ plot.
```

Conclusion

A QQ plot is a welcome addition to our data visualization arsenal. When the points of a QQ plot fall along a relative straight diagonal line, that means our data is nearly normal. Deviations from normality can easily be seen in a QQ plot, sometimes more easily than in a histogram (especially for small sample sizes).