Statistics for Engineering, 4C3/6C3, 2012

Assignment 2

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 23 January 2012**

Assignment objectives:

- Use a table of normal distributions to calculate probabilities
- Summarizing data my means and standard deviations, and their robust equivalent
- Ability to downloaded data and analyze it

Question 1 [2]

Estimate the following:

1. Without using tables or a computer: the cumulative area under the normal distribution between 15 and 35, with mean of 25 and standard deviation of 5.
2. The same as part 1, but using a table of normal distributions from the course notes (or another statistics textbook).
3. Between which lower and upper bounds will we find 60% probability of an event occurring, using the standardized (z) normal distribution? Calculate your answer using a printed table, ensuring that the two bounds are symmetrical about zero.
4. Convert these dimensionless z-bounds to real-world bounds for a process with mean of 100 kg and a standard deviation of 25 kg.
5. Verify your previous two answers using R, or other computer software.

Solution

1. The distance between the mean and the given bounds is equal to 10, which is the same as 2σ: so the cumulative area is roughly 95%.
2. The bounds must be transformed into z-values:

$$z_{\mathrm{L}} = \frac{15 - 25}{5} = -2$$
$$z_{\mathrm{U}} = \frac{25 - 25}{5} = +2$$

   From tables, the cumulative area up to -2 is 0.02275, and the cumulative area up to $+2$ is 0.9773. So the area from $z = -2$ up to $z = +2$ is $0.9773 - 0.02275 = 0.955$, or 95.5%.

   So the cumulative area between 15 and 35 is 95.5%.

3. There are infinitely many lower and upper bounds that contain 60% of the area under the standard normal distribution (mean of zero, variance of 1). However, for symmetrical bounds we would require 30% area to left of zero and 30% area above zero, since the mean of the distribution, $z = 0$ has 50% area below and 50% area above it.

   Using the tables, we look for the z-value that gives a cumulative area of 20% from $-\infty$. The tables in the notes only list the area at 10% and 30%. A quick visual interpolation shows $z \approx -0.85$ (though any value close to that will do). Similarly, a z-value that contains a cumulative area of 80% from $-\infty$ is $z \approx 0.85$.

So the lower bound is -0.85 and upper bound is 0.85.

4. Back-calculating the z-value gives:
   - $x = z_{\text{L}}s + m = (-0.85)(25) + 100 = 78.8$ kg for the lower bound
   - $x = z_{\text{U}}s + m = (+0.85)(25) + 100 = 121$ kg for the upper bound
5. Using R for part 3:

```
> qnorm(0.2)
[1] -0.8416212

> qnorm(0.8)
[1] 0.8416212
```

and for part 4:

```
> qnorm(0.2, 100, 25)
[1] 78.95947

> qnorm(0.8, 100, 25)
[1] 121.0405
```

Question 2 [3]

A chicken facility produces bags filled with breaded chicken strips. The advertised weight for each package is 750 grams. Each bag contains between 8 and 15 strips, given that each chicken strip is between 40 an 80 grams and from a uniform distribution. The company sets their target fill weight at 790 grams to avoid breaking regulations that require an accurate package labelling.

1. If we take a large sample of bagged chicken strips and weigh each bag, from which distribution will we expect these (bag) weights to come from?
2. Clearly explain why.
3. If the standard deviation of this large sample of bag weights is 12 grams, out of 10,000 customers, how many will purchase bags below the advertised 750g weight?

Solution

1. The total bag weight is expected to come from the normal distribution.
2. From the central limit theorem, if we take independent samples (weight of each chick strip is independent) and if the samples come from a distribution of finite variance (which is true for the uniform distribution), then the average of the samples, $\overline{x}$, will tend to be from a normal distribution.

   The average of the samples can be defined as $\overline{x} = \dfrac{1}{n} \cdot \sum_{i}^{n} x_i$. Since $\overline{x}$ is normally distributed from the CLT, then so will a scalar multiple of that value, $n\overline{x} = \sum_{i}^{n} x_i$. Note the right hand side is simply the total bag weight, which explains why the bag weight is normally distributed.

   Note this question, like all real-world problems, contains extra data that is not required to solve the problem.
3. We are essentially told the population standard deviation is 12g, and we expect the population mean to be 790g, the target weight. Currently the company overfills each bag by 40g (about 5% overage). But there are a few

customers that receive bag weights below 750g:

$$z = \frac{750 - 790}{12} = \frac{-40}{12} = -3.33$$

This corresponds to `pnorm(-40/12)*10000 = 4.29`, or about 4 customers in 10,000. Alternatively use `pnorm(750, mean=790, sd=12)*10000` to get the same result.

Question 3 [3]

1. Compute the mean, median, standard deviation and MAD for salt content of various potato chips in this report (page 22) as described in the the article from the Globe and Mail on 24 September 2009.
2. Plot a box plot of the data and report the interquartile range (IQR). Comment on the 3 measures of spread you have calculated: standard deviation, MAD, and interquartile range.
3. Comment on the effectiveness of the visualization plots used in the PDF report.

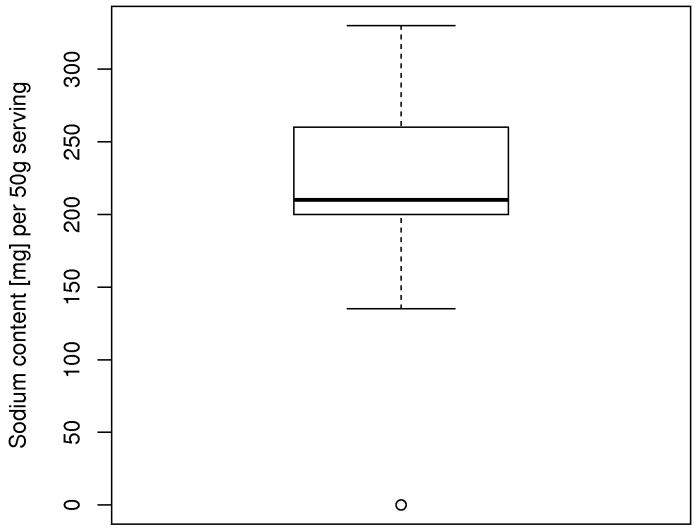
Solution

1. The raw sodium content data are: $[330, 290, 270, 260, 240, 240, 210, 210, 200, 200, 160, 135, 0]$. So in R we could write:

```
sodium <- c(330, 290, 270, 260, 240, 240, 210, 210, 200, 200, 160, 135, 0)
mean(sodium)
median(sodium)
sd(sodium)
mad(sodium)
```

   and obtain mean of 211 mg, a median of 210 mg, standard deviation of 82.2 mg and a MAD of 74.1 mg.

2. The box plot is:



   The positive skew is apparent, though that is primarily due to the "0" value; try removing it and redrawing the box plot.

   The interquartile range can be calculated `summary(sodium)`:

```
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0   200.0   210.0   211.2   260.0   330.0
```

so the IQR is $260 - 200 = 60$ mg.

The three measures of spread reported here are 60 mg, 74.1 mg and 82.2 mg: they are all roughly the same. The MAD and interquartile range are more resistant to outliers than the standard deviation. Since the value of "0" is easily considered an outlier in the context of this data set, it is not surprising to see the standard deviation value being higher than the other 2 measures.

3. *Solution from Nahomi Mahaffy (2012 class)*: The visualization is not very effective. The bars do not clearly demonstrate differences in the sodium amount, as different sizes are difficult to distinguish. The bars could be removed and this information could be displayed in a table that would demonstrate the same message, more clearly, and without extra ink. It is helpful that the chips varieties are organized based on their sodium content (from highest to lowest). The percentages given alongside the data are confusing; it seems to be illustrating percentage relative to 200mg of sodium per 50g of serving, but it does not indicate why this value was chosen (and the value is different for each graph in the document). The percentage values should be explained in the figure heading or description.

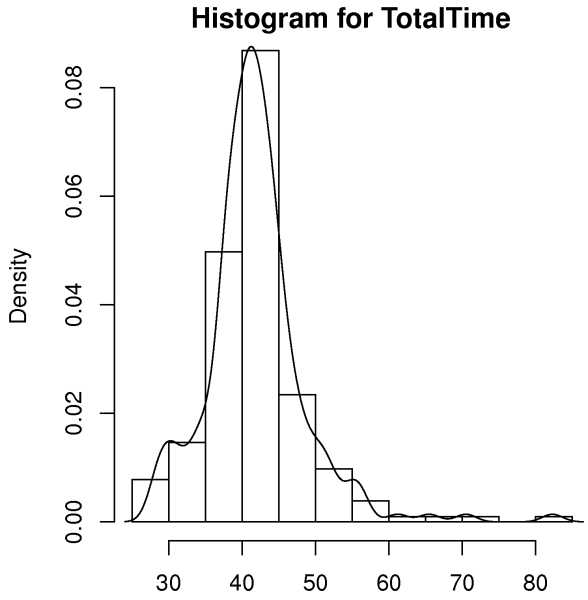
Question 4 [4]

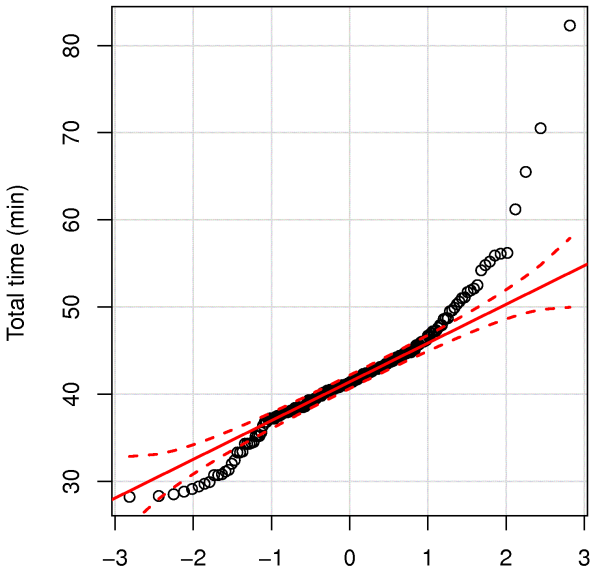
Data characterizing 200 commuting trips of your instructor was visualized in the previous assignment.

1. Plot a histogram of the `TotalTime` variable (the total time for the commute) to confirm the variable is not normally distributed.
2. How would you characterize the distribution of the `TotalTime` variable? Give reasons *why* the variable is not normally distributed.
3. Confirm the variable is not normally distributed by using a suitable, visual statistical test.
4. The 407 highway speeds are almost always much faster than the 403. Does the `MaxSpeed` variable (the maximum speed recorded during the entire trip, usually while travelling the 407) follow a normal distribution. Plot both a histogram and a q-q plot to check.

Solution

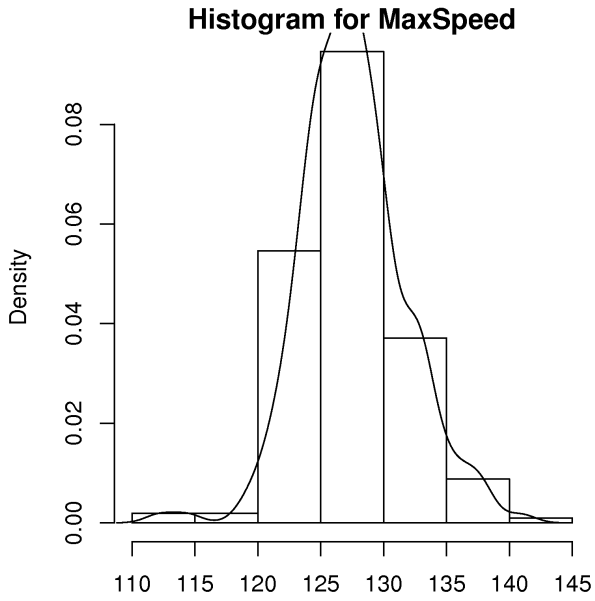
1. The histogram for total time doesn't appear to be normally distributed.
2. Instead, the total time has a strong positive skew, most trips are around 40 to 45 minutes, and quite a few that take much longer. This is expected: most of the times the highways are relatively free-flowing, so most trips are around the average duration. However, when there are accidents or bad weather the trips take much, much longer. There are no possible conditions that can produce trips significantly shorter than the average. So a positive skew is expected.
3. A q-q plot is shown, together with the histogram, to prove the variable is not normally distributed.

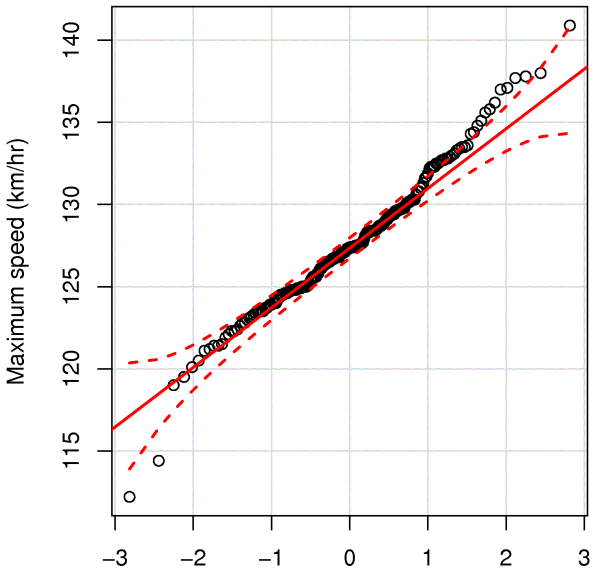




4. The `MaxSpeed` is roughly normally distributed. The histogram shows a more balanced, symmetrical distribution and this is confirmed by the q-q plot, whose 95% confidence lines contain most the data.

   The maximum speed makes sense to be normally distributed, since some days the traffic tends to move slower overall and other days faster. Notice the sharp cut at 120 km/hr: this is because a portion of the journey is always taken on the 407 everyday, and the 407 tends to have average speeds of 120 km/hr. I.e. the maximum trip speed is always recorded somewhere on the 407.





The R code for this question:

```
travel <- read.csv('http://datasets.connectmv.com/file/travel-times.csv')
summary(travel)

# Confirm it is not normally distributed
bitmap('travel-times-totaltime.png', pointsize=14, res=300,
       type="png256", width=10, height=5)
layout(matrix(c(1,2), 1, 2)) # layout plot in a 1x2 matrix
par(mar=c(2, 4, 1, 0.2))    # (bottom, left, top, right) spacing around plot
hist(travel$TotalTime, freq=FALSE, main="Histogram for TotalTime")
```

```
lines(density(travel$TotalTime))
library(car)
qqPlot(travel$TotalTime, ylab="Total time (min)")
dev.off()

bitmap('travel-times-maxspeed.png', pointsize=14, res=300,
       type="png256", width=10, height=5)
layout(matrix(c(1,2), 1, 2)) # layout plot in a 1x2 matrix
par(mar=c(2, 4, 1, 0.2))    # (bottom, left, top, right) spacing around plot
hist(travel$MaxSpeed, freq=FALSE, main="Histogram for MaxSpeed")
lines(density(travel$MaxSpeed))
qqPlot(travel$MaxSpeed, ylab="Maximum speed (km/hr)")
dev.off()
```

Question 5 [3]

In this question we investigate the stock prices for the Canadian National Railway Company (ticker `CNR` on the Toronto Stock Exchange).

- Visit http://finance.yahoo.com/
- Type in `CNR.TO` in the symbol (ticker) box
- Click **Historical Prices** in the left column
- Change the date range from 01 March 2011 to 01 January 2012
- Click **Get Prices** to get the "Daily" prices of the stock
- Scroll to the bottom of the page and click "Download to spreadsheet" to download a CSV file

Once you have loaded the CSV file into R, answer the following questions regarding the `Adj.Close` column (the price at which stock closes at end of the trading day, after adjusting for stock splits and dividends paid)

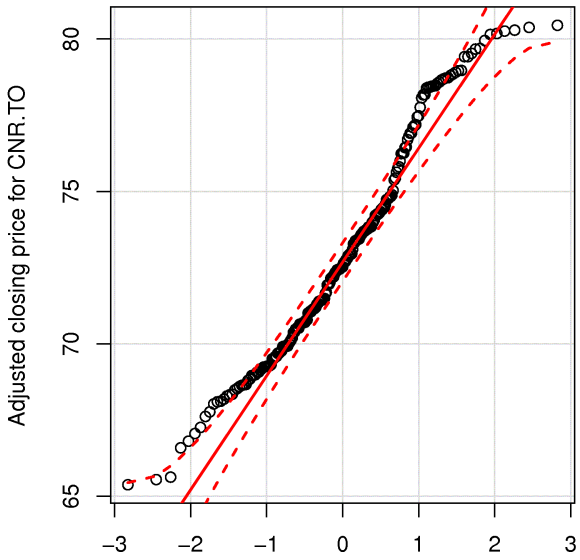
1. Are these closing prices from a normal distribution? Test your answer with a q-q plot.
2. Estimate the distribution's location and spread, assuming the data are from a normal distribution. 600-level students must use the `fitdistr` function in R from the MASS package.
3. Are these data points independent?
4. What is the probability of observing a stock value above $ 77.00 ?

Note: the purpose of this exercise is more for you to become comfortable with web-based data retrieval, which is common in most companies.

Solution

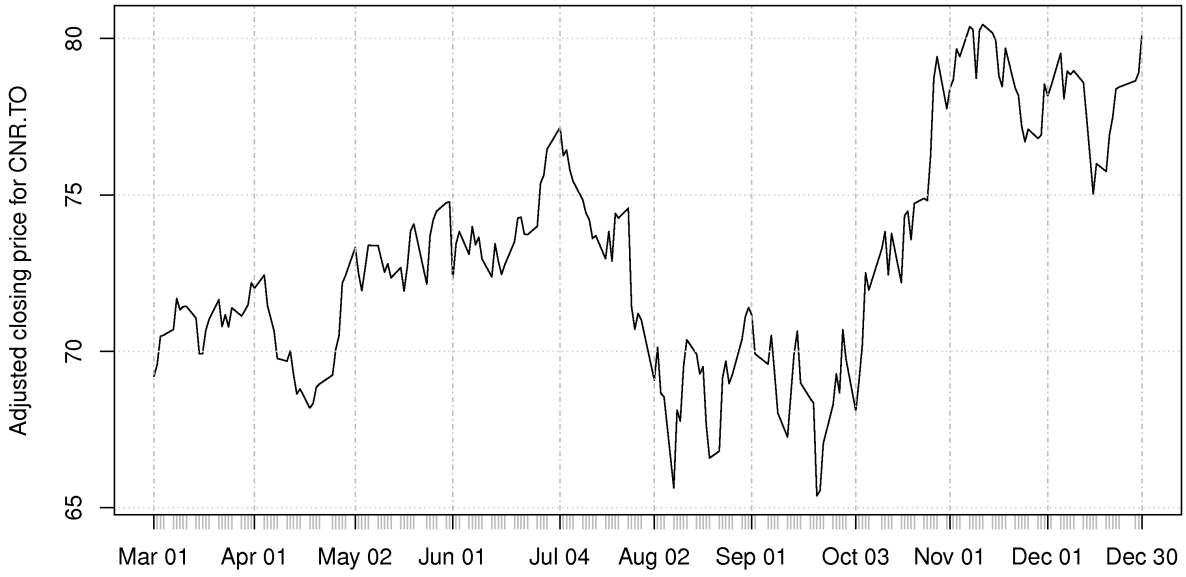
After downloading the 211 closing stock prices:

1. Yes: most of the data points fall within the q-q plot limits.



2. Location is given either by the mean, \$73.00, or median, \$72.53. Spread can be given by either the standard deviation, \$3.60, or the MAD, \$3.65. The `fitdistr` function from the MASS package reports a mean of \$73.00 (confidence interval of ± 0.25) and a standard deviation of \$3.60 (and CI of ± 0.18).

3. The data points from day-to-day are not expected to be independent. We can see this visually:



   There is a clear relationship in time between sequential points; prices the day before have a strong influence on prices in the following days.

4. The probability is given by finding the fraction of the distribution above \$77.00, and is 13.3% when using the location and spread values calculated previously. Note that the this calculation **does not** require the data to be independent.

The R code for this question:

```
# Save stock prices to CSV file:
CNR.TO <- read.csv('stock-prices.csv')
summary(CNR.TO)
library(car)
```

```
bitmap('stock-prices-qqplot.png', pointsize=14, res=300,
       type="png256", width=5, height=5)
par(mar=c(2, 4, 1, 0.2))   # (bottom, left, top, right) spacing around plot
qqPlot(CNR.TO$Adj.Close, ylab="Adjusted closing price for CNR.TO")
dev.off()

# Location and spread
c(mean(CNR.TO$Adj.Close), median(CNR.TO$Adj.Close))
c(sd(CNR.TO$Adj.Close), mad(CNR.TO$Adj.Close))
library(MASS)
fitdistr(CNR.TO$Adj.Close, 'normal')

# Independent? Can we see it visually? Defintely!
bitmap('stock-prices-timeseries.png', pointsize=14, res=300,
       type="png256", width=10, height=5)
par(mar=c(2, 4, 1, 0.2))   # (bottom, left, top, right) spacing around plot
plot(CNR.TO$Adj.Close, type="l", ylab="Adjusted closing price for CNR.TO")
dev.off()

# Use the xts library for better plots; search the software tutorial for "xts" to see how.
library(xts)
date.order <- as.Date(CNR.TO$Date, format="%Y-%m-%d")
CNR.TO$Date
date.order
Adj.Close <- xts(CNR.TO$Adj.Close, order.by=date.order)
bitmap('stock-prices-timeseries-xts.png', res=300, pointsize=14, width=10, height=5)
par(mar=c(2, 4, 1, 0.2))
plot(Adj.Close, ylab="Adjusted closing price for CNR.TO", main="")
dev.off()


# Use the autocorrelation function (acf) to check lack of independence:
# (we will introduce this function later on)
acf(CNR.TO$Adj.Close, lag=40)
acf(diff(CNR.TO$Adj.Close), lag=5)

# Probability:
1-pnorm(77, mean=mean(CNR.TO$Adj.Close), sd=sd(CNR.TO$Adj.Close))
```

END