Statistics for Engineers, 4C3/6C3

Assignment 2

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 19 January 2011**

Assignment objectives

- A review of basic probability, histograms and sample statistics.
- Collect data from multiple sources, consolidate it, and analyze it.
- Deal with issues that are prevalent in real data sets.
- Improve your skills with R (if you are using R for the course).

Notes:

- I would normally expect you to spend between 3 and 5 hours outside of class on assignments. This assignment should take about that long. Answer with bullet points, not in full paragraphs.
- **Numbers in bold** next to the question are the grading points. Read more about the assignment grading system.
- 600-level students must complete all the question; 400-level students may attempt the 600 level question for extra credit. Also 600-level students must read the paper by PJ Rousseeuw, "Tutorial to Robust Statistics".

Question 1 [1]

Recall from class that $\mu = \mathcal{E}(x) = \frac{1}{N}\sum x$ and $\mathcal{V}\{x\} = \mathcal{E}\left\{(x-\mu)^2\right\} = \sigma^2 = \frac{1}{N}\sum(x-\mu)^2$.

1. What is the expected value thrown of a fair 12-sided dice?
2. What is the expected variance of a fair 12-sided dice?
3. Simulate 10,000 throws in R, MATLAB, or Python from this dice and see if your answers match those above. Record the average value from the 10,000 throws.
4. Repeat the simulation for the average value of the dice a total of 10 times. Calculate and report the mean and standard deviation of these 10 simulations and *comment* on the results.

Solution

The objective of this question is to recall basic probability rules.

1. Let X represent a discrete random variable for the event of throwing a fair die. Let x_i for $i = 1, \ldots, 12$ represent the numerical or realized values of the outcome of the random event given by X. Now we can define the expected value of X as,

$$\mathcal{E}(X) = \sum_{i=1}^{12} x_i P(x_i)$$

where the probability of obtaining a value of $1, \ldots, 12$ is $P(x_i) = 1/N = 1/12 \; \forall \; i = 1, \ldots, 12$. So, we have,

$$\mathcal{E}(X) = \frac{1}{N}\sum_{i=1}^{12} x_i = \frac{1}{12}\left(1 + 2 + \cdots + 12\right) = \mathbf{6.5}$$

2. Continuing the notation from the above question we can derive the expected variance as,

$$\begin{aligned}\mathcal{V}(X) &= \mathcal{E}\left\{[X - \mathcal{E}(X)]^2\right\} \\ &= \mathcal{E}(X^2) - [\mathcal{E}(X)]^2\end{aligned}$$

where $\mathcal{E}(X^2) = \sum_i x_i^2 P(x_i)$. So we can now calculate $\mathcal{V}(X)$ as,

$$\begin{aligned}\mathcal{V}(X) &= \sum_{i=1}^{12} x_i^2 P(x_i) - \left[\sum_{i=1}^{12} x_i P(x_i)\right]^2 \\ &= \frac{1}{12}(1^2 + 2^2 + \cdots + 12^{12}) - [6.5]^2 \approx \mathbf{11.9167}\end{aligned}$$

3. Simulating 10,000 throws corresponds to 10,000 independent and mutually exclusive random events, each with an outcome in the set $\mathcal{S} = 1, 2, \ldots, 12$. The sample mean and variance from my sample was:

$$\begin{aligned}\overline{x} &= 6.4925 \\ s^2 &= 11.77915\end{aligned}$$

R code

```
x.data <- as.integer(runif(10000, 1, 13))

# Verify that it is roughly uniformly distributed
# across 12 bins
hist(x.data, breaks=seq(0,12))

x.mean <- mean(x.data)
x.var <- var(x.data)
c(x.mean, x.var)
```

MATLAB code

```
x.data = randi([1,12], 10000,1 );
x.mean = mean(x.data);
x.var = var(x.data,1);
```

4. Repeating the above simulation 10 times (i.e., 10 independent experiments) produces 10 different estimates of μ and σ^2. Note, everyone's answer should be slightly different, and different each time you run the simulation.

R code

```
N <- 10
n <- 10000
x.mean <- numeric(N)
x.var <- numeric(N)
for (i in 1:N) {
  x.data <- as.integer(runif(n, 1, 13))
  x.mean[i] <- mean(x.data)
  x.var[i] <- var(x.data)
}

x.mean
# [1] 6.5527 6.4148 6.4759 6.4967 6.4465
# [6] 6.5062 6.5171 6.4671 6.5715 6.5485

x.var
```

```
# [1] 11.86561 11.84353 12.00102 11.89658 11.82552
# [6] 11.83147 11.95224 11.88555 11.81589 11.73869

# You should run the code several times and verify whether
# the following values are around their expected, theoretical
# levels.  Some runs should be above, and other runs below
# the theoretical values.
# This is the same as increasing "N" in the first line.

# Is it around 6.5?
mean(x.mean)

# Is it around 11.9167?
mean(x.var)

# Is it around \sigma^2 / n = 11.9167/10000 = 0.00119167 ?
var(x.mean)
```

MATLAB code

```
x.data=zeros(10000,10);
for i=1:10
  x.data(:,i) = randi([1,12],10000,1);
end
x.mean = mean(x.data,1);
x.var = var(x.data,1,1);

mean(x.mean)
var(x.mean)
mean(x.var)
```

Note that each $\overline{x} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$, where $n = 10000$. We know what σ^2 is in this case: it is our theoretical value of **11.92**, calculated earlier, and for $n = 10000$ samples, our $\overline{x} \sim \mathcal{N}\left(6.5, 0.00119167\right)$.

Calculating the average of those 10 means, let's call that $\overline{\overline{x}}$, shows values around 6.5, the theoretical mean.

Calculate the variance of those 10 means shows numbers that are around 0.00119167, as expected.

Question 2 [1.5]

In the class last week I mentioned an example of independence. I said that if I take the grade for each question in an exam for a student, calculate the grade per question, then the average of those N grades will be normally distributed, even if the grades in individual question are not. For example: if there are 10 questions, and your grades for each question was 100%, 95%, 26%, 78%, ... 87%, then your average will be as if it came from a normal distribution.

1. This example was faulty: what was wrong with my reasoning?
2. 600-level students: However, when I look at the average grades from any exam, without fail they are always normally distributed. What's going on here?

Solution

1. Unfortunately, I chose my example in class too hastily, without thinking about the details. The grades for every student are not independent, because that student (as long as they are not receiving external help), will likely do well in all questions, or poorly in all questions, or only well in the section(s) they have studied. So each student's grade for the individual questions will be related.

2. **600-level** students: The central limit theorem tells us that samples from *any distribution with finite variance* (each question in the exam has a different distribution, but has finite variance), that the average of those values (the average grade of each student) will be normally distributed, as long as we took our samples independently (which we did not have for the grades example).

   So we are only breaking the independence assumption of the central limit theorem. That means we should take a look at why we've assumed independence between two sampled values.

   To do this, first let's look at the case when we do have independence, and for simplicity, let's assume every question in the exam also had a normal distribution with the same mean, μ and the same variance, σ^2 (really restrictive, but you will see why in a minute). We know that this case leads to:

$$\overline{x}_j \sim \mathcal{N}(\mu, \sigma^2/N)$$

   which says the average grade for student j, call it $\overline{x}_j$, comes from a normal distribution with that mean μ, and with standard deviation of σ^2/N, where N is the total number of questions. This is the usual formula we have seen in class; but where did this formula come from? Recall that:

$$\overline{x}_j = \frac{1}{N}x_{j,1} + \frac{1}{N}x_{j,2} + \ldots + \frac{1}{N}x_{j,N}$$

   where each student, j, obtained a grade for question, $1, 2, \ldots, n, \ldots N$. Let's call that grade $x_{j,n}$, and recall that we have assumed $x_{j,n} \sim \mathcal{N}(\mu, \sigma^2)$. The mean and standard deviation of $\overline{x}_j$, *crucially assuming independence* between each $x_{j,n}$ value, can then be found from:

$$\begin{aligned}
\mathcal{E}(\overline{x}_j) &= \mathcal{E}\left(\frac{1}{N}x_{j,1} + \frac{1}{N}x_{j,2} + \ldots + \frac{1}{N}x_{j,N}\right) \\
&= \frac{1}{N}\mathcal{E}(x_{j,1}) + \frac{1}{N}\mathcal{E}(x_{j,2}) + \ldots + \frac{1}{N}\mathcal{E}(x_{j,N}) \\
&= \frac{1}{N}\mu + \frac{1}{N}\mu + \ldots + \frac{1}{N}\mu \\
&= \mu \qquad \text{(this is expected)} \\
\mathcal{V}(\overline{x}_j) &= \mathcal{V}\left(\frac{1}{N}x_{j,1} + \frac{1}{N}x_{j,2} + \ldots + \frac{1}{N}x_{j,N}\right) \\
&= \frac{1}{N^2}\mathcal{V}(x_{j,1}) + \frac{1}{N^2}\mathcal{V}(x_{j,2}) + \ldots + \frac{1}{N^2}\mathcal{V}(x_{j,N}) \qquad \text{(this is why we require independence)} \\
&= \frac{N}{N^2}\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

   This also explains where the σ^2/N term, used in the t-distribution, comes from. The above derivation relies on two properties you should be familiar with (see a good stats textbook, e.g. Box, Hunter and Hunter):

$$\begin{aligned}
\mathcal{V}(x+y) &= \mathcal{V}(x) + \mathcal{V}(y) + 2\text{Cov}(x, y) \\
\mathcal{V}(x+y) &= \mathcal{V}(x) + \mathcal{V}(y) + 2\mathcal{E}\big[(x - \mathcal{E}[x])(y - \mathcal{E}[y])\big] \\
\mathcal{V}(ax) &= a^2\mathcal{V}(x)
\end{aligned}$$

   and independence implies that $\text{Cov}(x, y) = 0$.

   So relaxing our assumption of independent $x_{j,n}$ values shows that we cannot combine the variances in an easy way, but we do see that the correct variance will be a larger number (if the student grades within an exam are positively correlated - the usual case), or a smaller number (if the grades are negatively correlated within each student's exam).

   Also, relaxing the assumption that each question has the same variance, we just replace σ^2 with σ_n^2 in the formula for $\mathcal{V}(\overline{x})$. Relaxing the assumption of equal means, μ, for each question requires we use μ_j instead of μ. Note that σ_n^2 and μ_n can come from *any distribution*, not just the normal distribution.

But, the central limit theorem tells us average grade for a student, $\overline{x}_j$, will be as if it came from a normal distribution. However, because we do not have independence, and we don't know the individual σ_n^2 and μ_n values for each question, we cannot *estimate the parameters* of that normal distribution.

So, to conclude: it is correct that the average grades from the exam for every student will be as if they came from a normal distribution, only we can't calculate (estimate) that distribution's parameters. I always find my course grades to be normally distributed when examining the qq-plot.

Question 3 [1]

Write a few *bullet point* notes on the purpose of feedback control, and its effect on variability of process quality.

Solution

- Purpose is to keep the process close to a desired set point (or mean).
- Sometimes used to maintain the process variability within a desired tolerance limit (or standard deviation).
- Lowers the variability of the process outputs (i.e., narrow the distribution) by actually introducing *greater* variability into the process, to counteract external variation in the the process inputs. For example, variation from the raw materials, or ambient conditions, such as seasonal temperature are process inputs.
- Feedback control allows us to move the process operation closer to targets, without less likelihood of deviation outside these limits. (In the next section on process monitoring we will learn how to track and quantify this).

Question 4 [1.5]

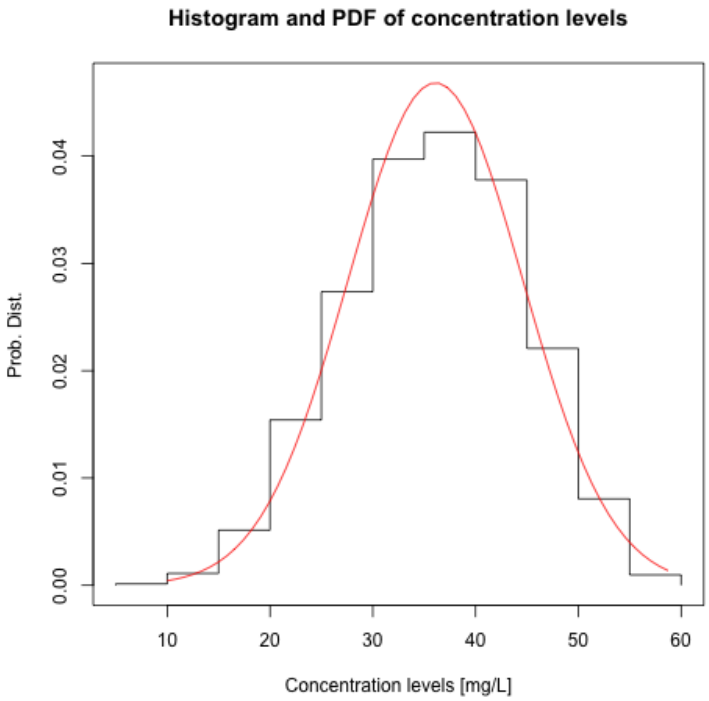
The ammonia concentration in your wastewater treatment plant is measured every 6 hours. The data for one year are available from the dataset website.

1. Use a visualization plot to hypothesize from which distribution the data might come. Which distribution do you think is most likely?
2. Confirm your answer using a suitable plot.
3. Estimate location and spread statistics assuming the data are from a normal distribution.
4. What if I told you that these measured values are not independent. How does it affect your answer?
5. What is the probability of having an ammonia concentration greater than 40 mg/L when:
    - you may use only the data (do not use *any* estimated statistics)
    - you use the estimated statistics for the distribution?

    Note: Answer this entire question using computer software to calculate values from the normal distribution. But also make sure you can answer the last part of the question by hand, if given the mean and variance, and using the table of normal distributions. Print out the table and bring it with you to the midterm and final exam. The computer answer should agree with your hand-calculated value.

Solution

1. To visualize/hypothesize which distribution the data might come from, use a histogram, a plot of the estimated frequency density, or simply a comparison of a histogram and normal PDF. We show a combined histogram and normal PDF as follows,



2. An appropriate distribution appears to be the normal distribution, however the right hand side tail (upper tail), of the plot shown below, is slightly heavier, outside the given limits, than would be found on the normal distribution. This bias may have a small effect on our results - by estimating a standard deviation that is larger than would have been from a true normal distribution.

3. Assuming the data are normal, we can calculate the distribution's parameters as $\bar{x} = \hat{\mu} = 36.1$ and $s = \hat{\sigma} = 8.52$.

4. The fact that the *data* are not independent is not an issue. To calculate estimates of the parameter's distribution (the mean and standard deviation) we do not need to assume independence. One way to see this: if I randomly reorder the data, I will still get the same value for the mean and standard deviation. The assumption of independence is required for the central limit theorem, but we have not used that theorem here.

5. The probability of having an ammonia concentration greater than 40 mg/L:

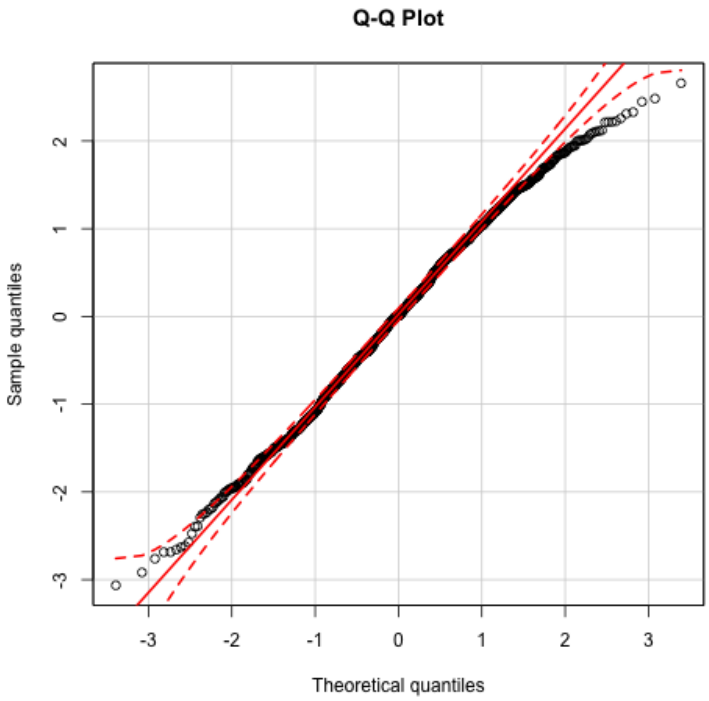
   - when using only the data: 34.4% (see code below)
   - when using the estimated parameters of the distribution: 32.3% (see code below)

   We should use the t-distribution to answer the last part, but at this stage we had not yet looked at the t-distribution. However, the large number of observations (1440) means the t-distribution is no different than the normal distribution. But note that the t-distribution requires the assumption that the data are normally distributed, and independent. We are better off *using the raw data* to estimate the probability in this case, without making these restrictive assumptions.

R code

```
# read in raw data
data <- read.csv('http://datasets.connectmv.com/file/ammonia.csv')
summary(data$Ammonia) # just to check we've got the right data

# place raw data into concen.x
concen.x <- data$Ammonia
# standardize the data
concen.mean <- mean(concen.x)   # 36.095
```




```
concen.sd <- sd(concen.x)       # 8.52
concen.z <- (concen.x-concen.mean)/concen.sd

# plot histogram
png(file='../images/Q4-histogram.png')
hist(concen.x,xlab="Concentration levels [mg/L]",
    main="Histogram of concentration levels")
dev.off()

# 600-level: could plot histogram with normal pdf to compare
# the distribution of the data
png(file='../images/Q4-histogram-npdf.png')
h <- hist(concen.x,breaks=11)
x_hist <- c(min(h$breaks),h$breaks)
y_hist <- c(0,h$density,0)
x_fit <- seq(min(concen.x),max(concen.x),length=60)
y_nfit <- dnorm(x_fit,mean=concen.mean,sd=concen.sd)
plot(x_hist,y_hist,type="s",ylim=c(0,max(y_hist,y_nfit)),
    xlab="Concentration levels [mg/L]",ylab="Prob. Dist.",
    main="Histogram and PDF of concentration levels")
lines(x_fit,y_nfit,type="l",col="red")
dev.off()

# 600-level: could estimate the frequency density of the data
png(file='../images/Q4-density.png')
plot(density(concen.x),xlab="Concentration Levels [mg/L]",
    ylab="Density",main="Density estimate of concentration levels")
dev.off()

# 400 and 600-level: verify normality using a qq-plot
library(car)
png(file='../images/Q4-qqplot.png')
```

```
qqPlot(concen.z,distribution="norm", ylab="Sample quantiles",
    xlab="Theoretical quantiles",main="Q-Q Plot")
dev.off()

# The distribution appears normal, apart from the right-hand-side tail

# 600-level: we can estimate distribution parameters using fitdistr()
# from the MASS library (uses an optimization-based maximum liklihood
# estimator) to find: mu = 36.0949931, sd = 8.5159694 -- very close
# to those estimated above.
library(MASS)
fitdistr(concen.x, "normal")

# determine, using raw data only, the probability of x > 40
level <- 40
p_est1 <- sum(concen.x>level)/length(concen.x)

# determine, using data statistics, the probability of x > 40
z <- (level-concen.mean)/concen.sd
p_est2 <- 1-pnorm(z)

# or more directly (without standardization) using
p_est3 <- 1-pnorm(level,mean=concen.mean,sd=concen.sd)

# We should have used the t-distribution, if we assume normality (OK)
# and independence (not checked for). The t-distribution is more
# appropriate, because we've used an estimate of sigma. However, with
# such a large number of degrees of freedom, our estimates are
# essentially the same.
p_est4 <- 1-pt(z, df=(length(concen.x)-1))
```

MATLAB code

```
% import data
data = textscan(fopen('ammonia.csv','r'),'%f','HeaderLines',1);

% place raw data into concen.x
concen.x = data{1};
% standardize the data
concen.mean = mean(concen.x);
concen.std = std(concen.x,0,1);
concen.z = (concen.x-concen.mean)/concen.std;

% plot histogram
figure;
hist(concen.x);
xlabel('Concentration Levels [mg/L]');
ylabel('Frequency');
title('Histogram of Concentration Levels');

% Alternatively, plot histogram with normal pdf to
% compare distribution of data
figure;
histfit(concen.x,11,'normal');
xlabel('Concentration Levels [mg/L]');
ylabel('Frequency');

% to verify normality use a normality or qq-plot
figure;
probplot('normal',concen.x)
```

```
figure;
qqplot(concen.z);

% determine, using raw data only, the probability
level = 40;
p_est1 = sum(concen.x > level)/length(concen.x);

% estimate parameters of the distribution
z = (level-concen.mean)/concen.std;
p_est2 = 1-normcdf(z,0,1);

% or more directly (without standardization) using
p_est3 = 1-normcdf(level,concen.mean,concen.std);

% We should have used the t-distribution, if we assume normality (OK)
% and independence (not checked for). The t-distribution is more
% appropriate, because we've used an estimate of sigma. However, with
% such a large number of degrees of freedom, our estimates are
% essentially the same.
p_est4 = 1-tcdf(z,length(concen.x)-1);
```

Question 5 [1]

One of the questions we posed at the start of the course was: given the yields from a batch bioreactor system for the last 3 years (300 data points; we run a new batch about every 3 to 4 days).

1. What sort of distribution do the yield data have?
2. A recorded yield value was was less than 60%, what are the chances of that occurring? Express your answer as: *there's a 1 in X chance* of it occurring.
3. Which assumptions do you have to make for the second part of this question?

Solution

1. Assume the 300 data points represent an entire population. Plot a `qqPlot(...)` using the `car` package (you could also try using a normal probability plot, e.g. the `probplot()` function from the `e1071` package):

   Since the data seem to agree with the above plot, we can conclude they follow a normal distribution.

2. We need to find the probability that the yield, Y, is less than or equal to 60, stated as $P(Y \leq 60)$. If we assume $Y \sim \mathcal{N}(\mu, \sigma^2)$ then we first need to find the z-value bound corresponding to 60, and then find the probability of finding values below, or equal to that bound.

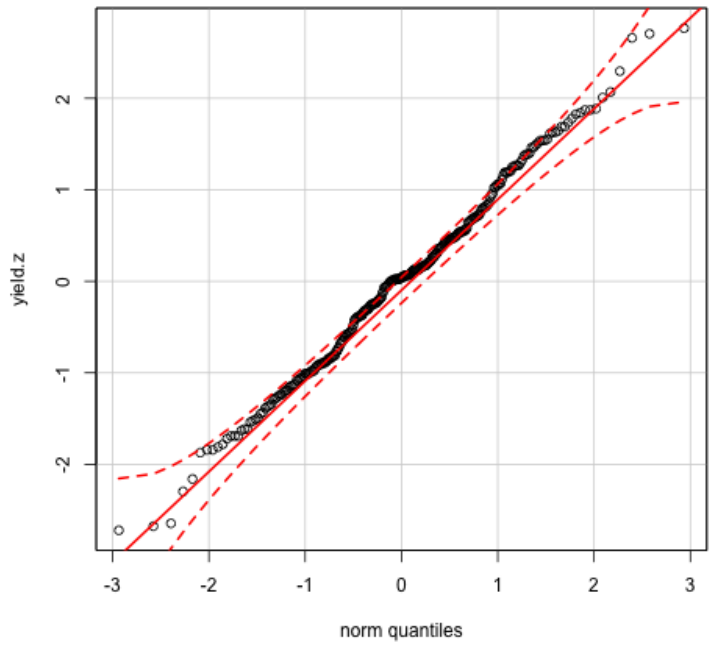
   $$z_{\text{bound}} = \frac{y - \mu}{\sigma} = \frac{60 - 80.353}{6.597} = -3.085$$

   In this data set of 300 numbers there are zero entries below this limit. But using the distribution's fit, we can calculate the probability as `pnorm(-3.085)`, which is ≈ 0.001. This is equivalent to saying that there is a *1 in 1000 chance* of achieving a yield less than 60%.

3. We only had to assume the data are normally distributed - we did not need the data to be independent - in order to use the estimated parameters from the distribution to calculate the probability.

R code

```
# import data
data <- read.csv('http://datasets.connectmv.com/file/batch-yields.csv')
```




```
# determine statistics
summary(data)
yield.x <- data$Yield
yield.mean <- mean(yield.x)
yield.var <- var(yield.x)*299/300
yield.sd <- sd(yield.x)*sqrt(299/300)
yield.z <- (yield.x-yield.mean)/yield.sd # standardize data

# Check distribution of data via a probability plot
library(e1071)
probplot(yield.x,xlab="Batch Yield")

# Rather use a qqplot with limits
library(car)
png(file='../images/Q5-qqplot.png')
qqPlot(yield.z)
dev.off()

# Could also use the standard qqplot in R as,
# qqnorm(yield.x, datax = TRUE);
# qqline(yield.x, datax = TRUE)

# assuming normal distribution determine probability of x < 60
z <- (60-yield.mean)/yield.sd
p <- pnorm(z,0,1)
```

MATLAB code

```
% import data
data = textscan(fopen('batch-yields.csv','r'),'%f','HeaderLines',1);
yield.x = data{1};

% determine statistics
```

```
yield.mean = mean(yield.x);
yield.var = var(yield.x,1,1);
yield.std = std(yield.x,1,1);
yield.z = (yield.x-yield.mean)/yield.std;

% Check distribution of data via a probability plot
figure;
probplot('normal',yield.x)
xlabel('Batch Yield')

% Or a qqplot
figure;
qqplot(yield.z);

% assuming normal distribution determine probability
z = (60-yield.mean)/yield.std;
p = normcdf(z,0,1);
```

Question 6 [1]

Use the section on Historical data from Environment Canada's website and use the `Customized Search` option to obtain data for the `HAMILTON A` station from 1990 to 2000. Use the settings as `Year=1990`, and `Data interval=Monthly` and request the data for 1990, then click `Next year` to go to 1991 and so on.

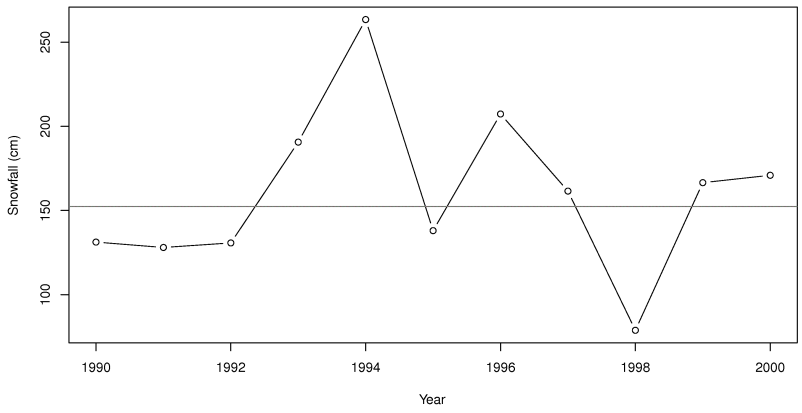
- For each year from 1990 to 2000, record the total snowfall and the average of the `Mean temp` column over the 12 months (the sums and averages are reported at the bottom of the table).
- Plot these 2 variables against time.
- Perform a test to verify whether the 11 total snowfall values are possibly from a normal distribution.
- Now retrieve the long-term averages for these data from a different section of their website (use the same location, `HAMILTON A`, and check that the data range is 1961 to 1990). Superimpose the long-term average as a horizontal line on your previous plot.

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

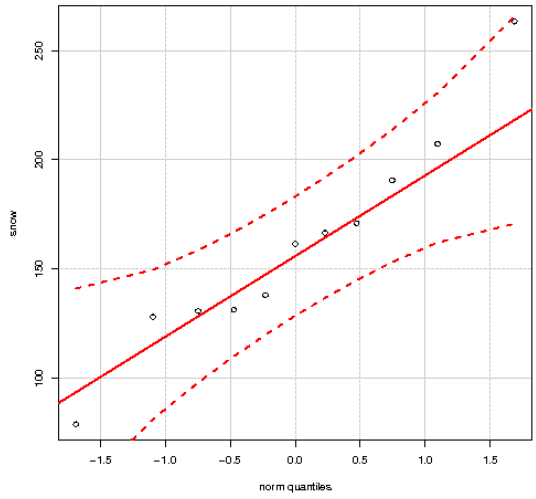
- The snow data are $[131.2, 128.0, 130.7, 190.6, 263.4, 138.0, 207.3, 161.5, 78.8, 166.5, 170.9]$.

  The mean temperature data are $[8.6, 8.6, 6.9, 7.1, 7.1, 7.7, 6.9, 7.3, 9.8, 8.8, 7.6]$.
- A plot against time for these two, with the long-term average (1961 to 1990 values) superimposed



where the 1961 to 1990 average snowfall was 152.4 cm per year, and the average temperature was 7.6 Celsius.

- The qq-plot for the snowfall data shows that these 11 values could quite possibly come from a normal distribution.



END