STAT 305 D Homework 3

Due February 7, 2012 at 12:40 PM in class

1. Here is dataset giving the number of printer jams per day of a receipt printer in a supermarket for 17 days.

$$43, 100, 500, 23, 89, 89, 89, 89, 72, 72, 72, 21, 32, 41, 39, 47, 56$$

   - Identify the sample. The 17 days during which the number of printer jams was recorded.
   - Identify the population. All the days (past, present, and future) during which we *could* record or *could have* recorded the number of printer jams.
   - Calculate the sample mean. $\overline{x} = 86.7059$.
   - Calculate the median. Sort the values and pick the middle value. Doing that gives a median of 72.
   - Calculate the mode. Mode = 89 because 89 appears most frequently.
   - Calculate the sample variance. $s^2 = \frac{1}{17}[(43 - 86.7059)^2 + (100 - 86.7059)^2 + \cdots] = 1.1989 \times 10^4$.
   - Calcluate the sample standard deviation. $s = \sqrt{s^2} = 109.4936$
   - Calculate the range. Range = 500 - 21 = 479.

2. Chapter 3 section 2 part of exercise 1 (page 92):

The following are data (from *Introduction to Contemporary Statistical Methods* by L. H. Koopmans) on the impact strength of sheets of insulating material cut in two different ways. (The values are in ft lb.)

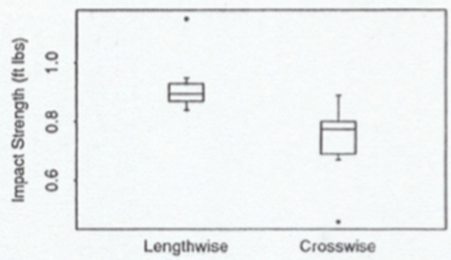
Lengthwise Cuts	Crosswise Cuts
1.15	.89
.84	.69
.88	.46
.91	.85
.86	.73
.88	.67
.92	.78
.87	.77
.93	.80
.95	.79

For the lengthwise sample, $Q(0.25) = 0.870, Q(0.5) = 0.895$, and $Q(0.75) = 0.930$. For the crosswise sample, $Q(0.25) = 0.690, Q(0.5) = 0.775$, and $Q(0.75) = 0.800$.

a. Find the IQR. IQR = Q(0.75) - Q(0.25) = 0.930 - 0.870 = 0.06

b. Draw (to scale) carefully labeled side-by-side boxplots for comparing the two cutting methods. Discuss what these show about the two methods.

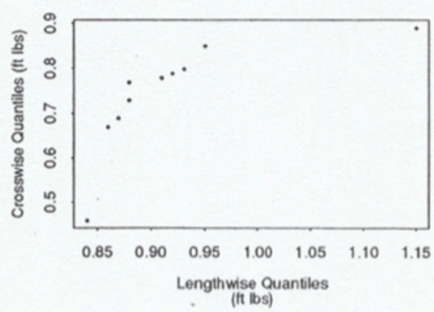
(b)



On the whole, the impact strengths are larger and more consistent for lengthwise cuts. Each method produced an unusual impact strength value (outlier).

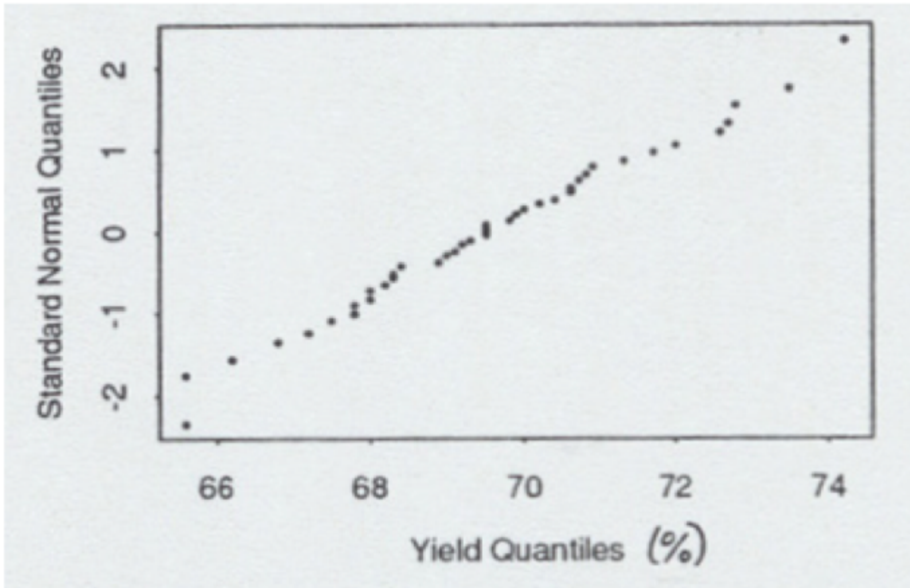
c. Make and discuss the appearance of a Q-Q (quantile-quantile) plot for comparing the shapes of these two data sets.

(c)



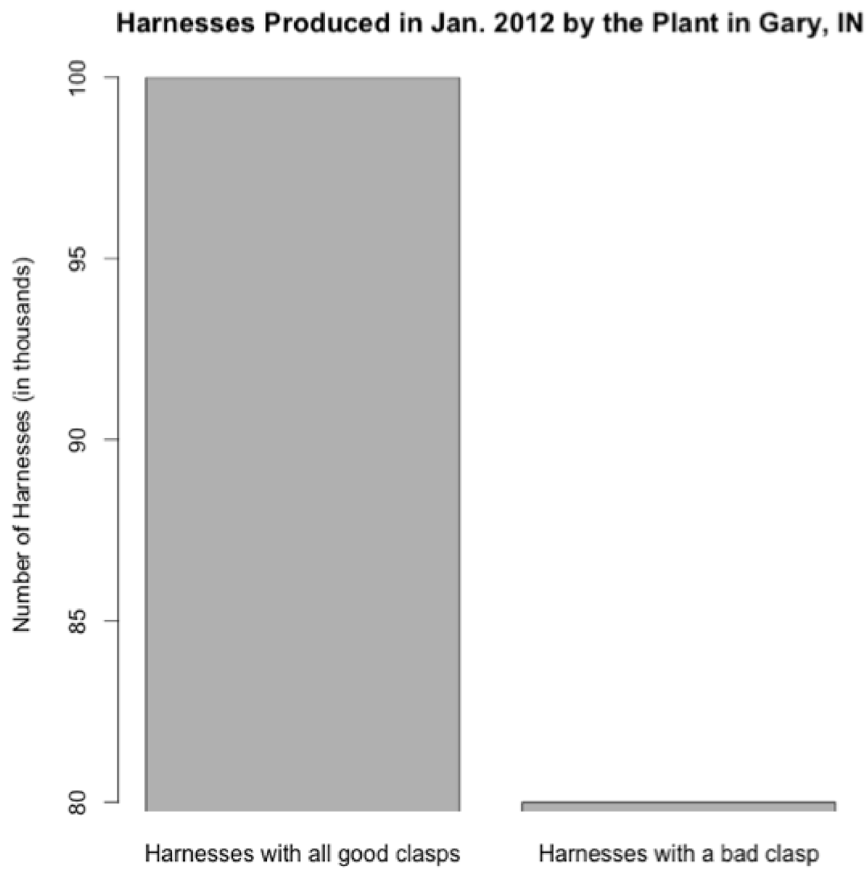
The non-linearity of the Q-Q plot indicates that the overall shapes of these two data sets are not the same. The lengthwise cuts had an unusually large data point ("long right tail"), whereas the crosswise cuts had an unusually small data point ("long left tail"). Without these two outliers, the data sets would have similar shapes, since the rest of the Q-Q plot is fairly linear.

3. Below is a normal quantile plot of the yield data from problem 1 (Chapter 3 Section 1 Exercise 1, page 77). What does this plot tell you about the distributional shape of the yield data? Justify your answer based on the shape of the points in the plot.



Since the data in the above normal quantile plot follow a straight line, the yield values are symmetric and bell-shaped (normally distributed). In practice, you dont usually get data better than this.

4. You work for one of the main manufactures of fire rescue harnesses in the United States. In the news one day, you read about a firefighter who was using one of your companys harnesses in a rescue and suffered an accident because a clasp on his harness came undone. You contact the firefighter, obtain the harness that failed, and determine that the bad clasp was originally faulty when it came out of the plant in Gary, IN. Concerned that other harnesses may have faulty clasps, you consult your supervisor. He informs you that he is already investigating the problem, and he shows you the following bar chart:



1. Is the above plot an honest representation of the data? Why or why not? If there is there anything misleading about the plot, what is it?

   The above plot is dishonest and misleading. The vertical axis starts from 79.75 thousand harnesses, hiding most of the 80,000 faulty harnesses corresponding to the right-hand bar.

2. Should the company arrange a recall of the harnesses produced in Jan. 2012 by the Gary plant, or is a recall unnecessary?

   With 80,000 faulty harnesses out of 180,000 harnesses produced in Jan. 2012, a recall is necessary.

5. Vardeman and Jobe chapter 4 section 1 problem 3 parts a-c (page 140). Data can be found at `http://www.will-landau.com/stat305/data/csv/polypolyols.csv`. You're welcome to use a spreadsheet program to do this problem, but please show your calculations. If you use Excel formulas, for example, please write them down. The article Polyglycol Modified Poly (Ethylene

Ether Carbonate) Polyols by Molecular Weight Advancement by R. Harris (Journal of Applied Poly- mer Science, 1990) contains some data on the effect of reaction temperature on the molecular weight of resulting poly polyols. The data for eight experimental runs at temperatures 165°C and above are as follows. Here, x is pot temperature (°) and y is average molecular weight.

x	y
165.00	808.00
176.00	940.00
188.00	1183.00
205.00	1545.00
220.00	2012.00
235.00	2362.00
250.00	2742.00
260.00	2935.00

a. What fraction of the observed raw variation in y is accounted for by a linear equation in x?

b. Fit a linear relationship $y \approx b_0 + b_1 x$ to these data via least squares. About what change in average molecular weight seems to accompany a $1°C$ increase in pot temperature (at least over the experimental range of temperatures)?

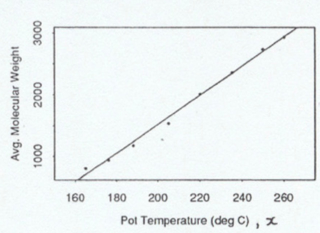
c. Compute and plot residuals from the linear relationship fit in (b). Discuss what they suggest about the appropriateness of that fitted equation. (Plot residuals versus x and residuals versus $\widehat{y}$.)

(a) $R^2 = .994$.

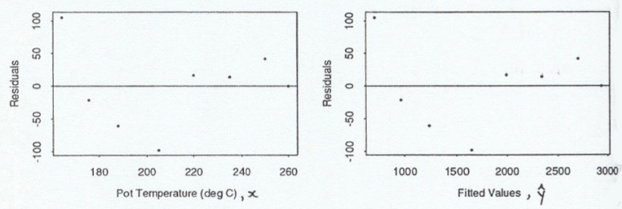
(b) The least squares equation is

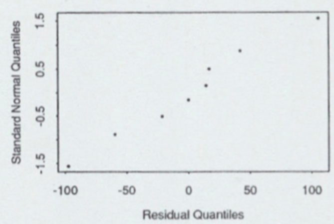
$$\hat{y} = -3174.6 + 23.5x.$$

β_1 represents the "true" average change in molecular weight that accompanies a 1°C increase in pot temperature (assuming that a straight-line model is correct). $b_1 = 23.5$ is a data-based approximation of this value.



(c) The residuals are: 105.36, −21.13, −60.11, −97.58, 16.95, 14.48, 42.00, and .02.





It is difficult to evaluate the appropriateness of the fitted equation based on so little data. The plots of residuals versus x and residuals versus $\hat{y}$ do not contain any obvious patterns, and thus provide no evidence that the equation is inappropriate. The normal plot of residuals is fairly linear, providing no evidence that the residuals are not bell-shaped.

6. Vardeman and Jobe chapter 4 section 1 problem 4 parts a-c(page 140). Data can be found at http://www.will-landau.com/stat305/data/csv/tools.csv. Please do this problem by hand.

Upon changing measurement scales, nonlinear relationships between two variables can sometimes be made linear. The article "The Effect of Experimental Error on the Determination of the Optimum Metal-Cutting Conditions" by Ermer and Wu (*The Journal of Engineering for Industry*, 1967) contains a data set gathered in a study of tool life in a turning operation. The data here are part of that data set.

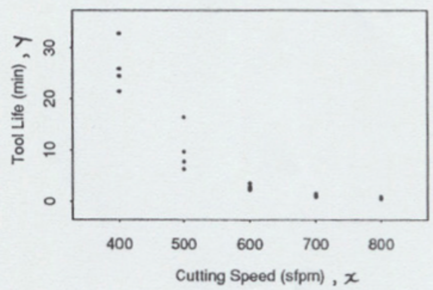
Cutting Speed, x (sfpm)	Tool Life, y (min)
800	1.00, 0.90, 0.74, 0.66
700	1.00, 1.20, 1.50, 1.60
600	2.35, 2.65, 3.00, 3.60
500	6.40, 7.80, 9.80, 16.50
400	21.50, 24.50, 26.00, 33.00

(a) Plot y versus x and calculate R^2 for fitting a linear function of x to y. Does the relationship $y \approx \beta_0 + \beta_1 x$ look like a reasonable explanation of tool life in terms of cutting speed?

(b) Take natural logs of both x and y and repeat part (a) with these log cutting speeds and log tool lives.

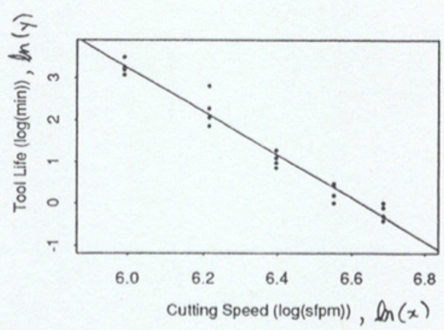
(c) Using the logged variables as in (b), fit a linear relationship between the two variables using least squares. Based on this fitted equation, what tool life would you predict for a cutting speed of 550? What approximate relationship between x and y is implied by a linear approximate relationship between $\ln(x)$ and $\ln(y)$? (Give an equation for this relationship.) By the way, Taylor's equation for tool life is $yx^{\alpha} = C$.

(a)



The scatterplot is not linear, so the given straight-line relationship does not seem appropriate. The least squares line is $\hat{y} = 44.075 - .059650x$. The corresponding R^2 is .723.

(b)



This scatterplot is much more linear, and a straight-line relationship seems appropriate for the transformed variables. The least squares line is $\widehat{\ln y} = 34.344 - 5.1857 \ln x$. The corresponding R^2 is .965.

(c) The least squares line is given in part (b). For $x = 550$,

$$\widehat{\ln y} = 34.344 - 5.1857 \ln(550) = 1.6229 \ln(\text{minutes}),$$

so $\hat{y} = e^{1.6229} = 5.07$ minutes. The implied relationship between x and y is

$$\begin{aligned} y &\approx e^{\beta_0 + \beta_1 \ln x} \\ y &\approx e^{\beta_0} e^{\ln x^{\beta_1}} \\ y &\approx e^{\beta_0} x^{\beta_1}. \end{aligned}$$

With slight rearrangement, this is the same as Taylor's equation for tool life.

7. Weekly feedback. You get full credit as long as you write something.

   1. Is there any aspect of the subject matter that you currently struggle

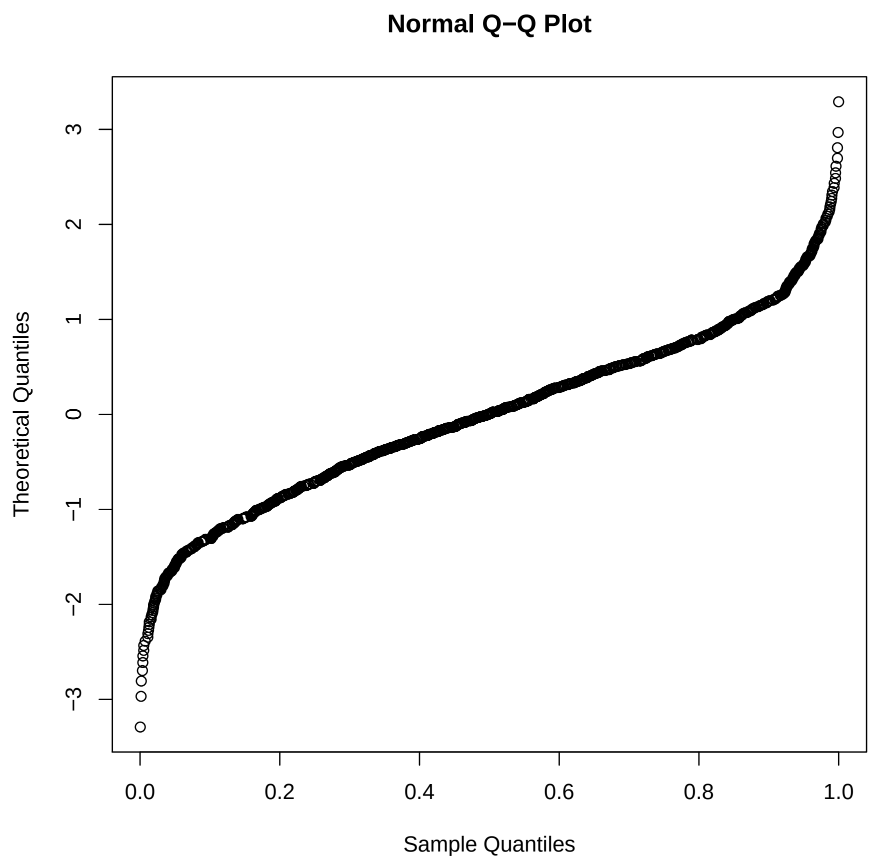
with? If so, what specifically do you find difficult or confusing? The more detailed you are, the better I can help you.

You got full credit as long as you wrote something.

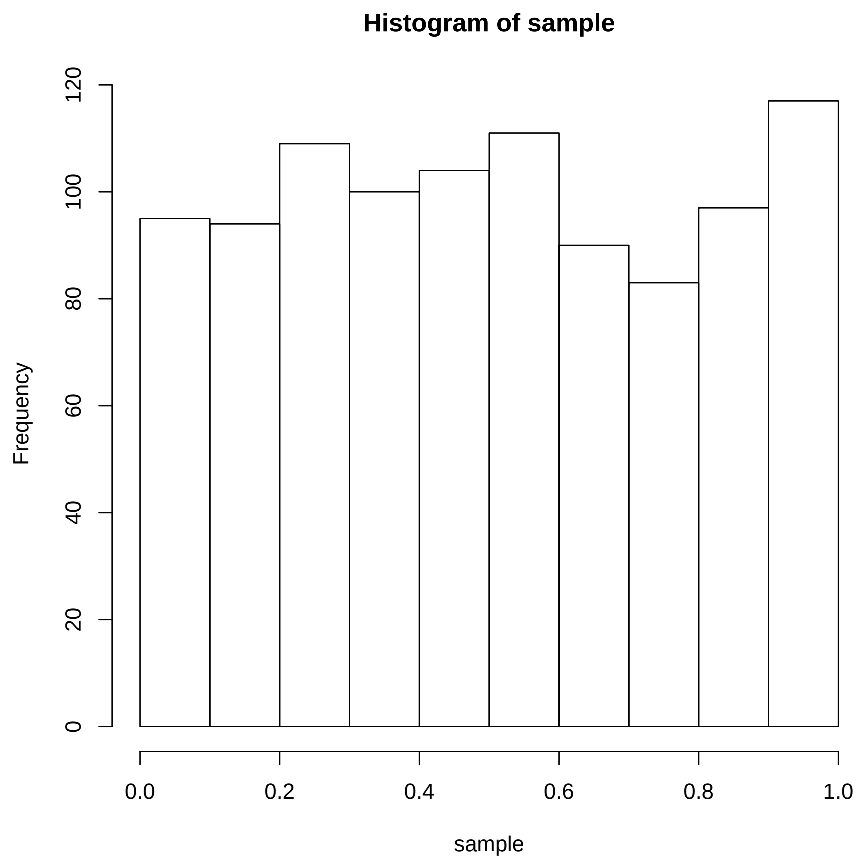
2. Do you have any questions or concerns about the material, class logistics, or anything else? If so, fire away.

You got full credit as long as you wrote something.

8. EXTRA CREDIT. Using the normal quantile plot below, draw the approximate distributional shape of the original data.



Here is the histogram of the original data. The distributional shape is approximately uniform.



However, there are shapes besides uniform that are also correct, such as...

You could have also put something like: