NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

IOMS Department

Statistics and Data Analysis

Professor William Greene
Phone: 212.998.0876
Office: KMC 7-90
Home page: http://people.stern.nyu.edu/wgreene
Email: wgreene@stern.nyu.edu
Course web page: http://people.stern.nyu.edu/wgreene/Statistics/Outline.htm

Assignment 2

Notes:
The data sets for this homework (and for the other problem sets for this course) are all stored on the home page for this course. You can find links to all of them on the course outline, at the bottom with the links to the problem sets themselves.

Part I. Binomial Probability

1. Let r be a binomial random variable. Compute $P_r(r)$ for each of the following situations:

a. n= 10, θ = .2, r= 3
b. n= 4, θ = .4, r= 2
c. n= 16, θ = .7, r= 12

SOLUTION: For part (a), the requested probability is

$$\binom{10}{3}0.2^8 \times 0.8^7 = 120 \times 0.008 \times .2097152 = .201326592 \approx 0.2013$$

For part (b), we need

$$\binom{4}{2}0.4^2 \times 0.6^2 = 6 \times 0.16 \times .36 = 0.3456$$

Part (c) requests

$$\binom{16}{12}0.7^{12} \times 0.3^4 = \binom{16}{4}0.7^{12} \times 0.3^4$$

Minitab will make this easy:

Probability Density Function

```
Binomial with n = 16 and p = 0.700000
x P( X = x)
12.00 0.2040
```

Thus, the requested probability is 0.2040.

2. A chain of motels has adopted a policy of giving a 3% discount to customers who pay in cash rather than by credit cards. Its experience is that 30% of all customers take the discount. Let Y = number of discount takers among the next 20 customers.

a. Do you think the binomial assumptions are reasonable in this situation?

b. Assuming that the binomial probabilities apply, find the probability that exactly 5 of the next 20 customers take the discount.

c. Find P(5 or fewer customers take the discount).

d. What is the most probable number of discount takers in the next 20customers?

SOLUTION: The binomial assumption is quite reasonable. Parts (b) through (d) suggest that it would be convenient to see the whole distribution. Thus, we'll use Minitab. Given below are the probabilities and then the cumulative probabilities. The output was edited by taking out the highly unlikely outcomes 16 through 20.

```
Binomial with n = 20 and p = 0.300000
x P( X = x) P( X <= x)
 0.00 0.0008 0.0008
 1.00 0.0068 0.0076
 2.00 0.0278 0.0355
 3.00 0.0716 0.1071
 4.00 0.1304 0.2375
 5.00 0.1789 0.4164
 6.00 0.1916 0.6080
 7.00 0.1643 0.7723
 8.00 0.1144 0.8867
 9.00 0.0654 0.9520
10.00 0.0308 0.9829
11.00 0.0120 0.9949
12.00 0.0039 0.9987
13.00 0.0010 0.9997
14.00 0.0002 1.0000
15.00 0.0000 1.0000
```

For part (b), we see that the probability of exactly 5 is 0.1789.
For part (c), the probability of 5 or fewer is 0.4164.

The probability listing (called the density by Minitab) has a peak probability of 0.1916 at $y = 6$. This is the most probable number of discount takers.

3. The admissions office of a small, selective liberal-arts college will only offer admission to applicants who have a certain mix of accomplishments, including a combined SAT score of 1,400 or more. Based on past records, the head of admissions feels that the probability is 0.66 that an admitted applicant will come to the college. If 500 applicants are admitted, what is the probability that 340 or more will come? Note that "340 or more" means the set of values {340, 341, 342, 343, …, 499, 500}.

SOLUTION: You can get Minitab to find the cumulative binomial probability for $n = 500$, $p = 0.66$, and for 339 or fewer "successes."

Cumulative Distribution Function

```
Binomial with n = 500 and p = 0.66
x P( X <= x )
339 0.814839
```

The probability that 339 or fewer people will come is 0.814839. Thus the probability is 1 – 0.814839 = 0.185161, about 18.52%, that 340 or more will come. You can also use the normal approximation. For this, μ would be 500(.66) = 330 and σ would be sqr(100(.66)(.34)) = 10.59. We are looking for the probability that $x \geq 340$. Making the correction, we will be looking for the normal probability that $x \geq 339.5$. Standardizing this, we get Prob($z \geq (339.5 – 330)/10.59$) = Prob($z \geq 9.5/10.59$) = 1 – Prob($z \leq .897$). Using the normal table, the probability for .89 is .8133 and for .90 it is .8159. Going 7 tenths of the distance (linearly interpolating), this would be .8151. 1 - .8151 = .1849. That is a very good estimate of the correct value of .185161.

4. Suppose that a full-repair warranty is offered with each new Power-Up foodprocessor. If the probability that any individual food processor will be returned forneeded warranty repairs within one year is 0.11, and if a certain store sells 83 of these,find the probabilities that

a. at most 10 food processors will be returned for warranty repairs;
b. at least 10 food processors will be returned for warranty repairs;
c. exactly 10 food processors will be returned for warranty repairs;
d. not more than 15 food processors will be returned for warranty repairs.

SOLUTION: Minitab solves these routinely. We ask for the binomial distribution with $n = 83$ and $p = 0.11$. For part (a), we ask for the cumulative probability up through $x = 10$. This gives us

```
Binomial with n = 83 and p = 0.110000
x P( X <= x )
10.00 0.6969
```

Thus, $0.6969 \approx 0.70$ is the probability that at most 10 will be returned. For (b), we ask for the cumulative probability up through $x = 9$. This gives us

```
Binomial with n = 83 and p = 0.110000
x P( X <= x )
9.00 0.5694
```

This corresponds to $P[X \leq 9]$. Thus, our solution to (b) is $P[X \geq 10] = 1 - 0.5694 = 0.4306 \approx 0.43$. For (c), we ask for the (simple) probability for $x = 10$. This gives us

```
Binomial with n = 83 and p = 0.110000
x P( X = x )
10.00 0.1275
```

Thus the probability of having *exactly* ten food processors returned is $0.1275 \approx 0.13$. For (d) we recognize that "not more than 15" means exactly the same as "at most 15."
This is similar to (a). We find then

```
Binomial with n = 83 and p = 0.110000
x P( X <= x )
15.00 0.9820
```

Then the probability of having not more than 15 returns is $0.9820 \approx 0.98$. Note that you can also use the normal approximation as we did in the previous problem.

Part II. Poisson Probability

5. The rate of home sales at a small real estate agency is 1.3 per day. We'll assume that a Poisson phenomenon can represent these home sales.

a. Find the probability that no homes will be sold on Monday.
b. Find the probability that one home will be sold on Monday.
c. Find the probability that two homes will be sold on Monday.
d. Find the probability that more than two homes will be sold on Monday.

SOLUTION:

(a) The probability is $e^{-1.3}\frac{1.3^0}{0!} = e^{-1.3} \approx 0.272532$

(b) The probability is $e^{-1.3}\frac{1.3^1}{1!} = e^{-1.3}1.3 \approx 0.354291$

(c) The probability is $e^{-1.3}\frac{1.3^2}{2!} = e^{-1.3}\frac{1.69}{2} \approx 0.230289$

6. For each of the following situations, indicate whether the model should be binomial or Poisson or something else.

a. The number of major forest fires to strike Colorado in calendar year 2006.
b. The number of trading days in the month of October that the stock of General Electric will go up in value.
c. The number of plays at craps, out of 50 attempted, that are winners.
d. The number of prize coupons, out of 800 inserted into cereal boxes, that are returned to collect the prizes.
e. The number of visitors to your web site on 25 FEB 2007.
f. The number of dead squirrels found on one mile of highway 93, on May15. (Such schemes are actually used to estimate animal populations.)
g. The number of expense account claims with inadequate documentation, in a sample of 10 selected from a master file of 280.
h. The number of mattresses, out of 140 sold during the month of May, returned by the customers.
i. The number of diamonds in a single hand at hearts. (In the game of hearts, a single hand consists of 13 cards dealt from the deck of 52.)
j. The number of customers, out of 418 who made a purchase at Windham Supermarket, who purchased milk.

SOLUTION:

(a) is Poisson.
(b) is binomial. There are n trading days, and we let p be the probability of a stock price improvement. This is not a perfect model, as there may be day-to-day dependence and p may change during the month. Nonetheless, the binomial is the best simple model.
(c) is binomial.
(d) is binomial. Each of the 800 coupons represents a separate yes-or-no trial.
(e) is Poisson.
(f) is Poisson. It's possible to think of this as binomial if you imagine that there are n squirrels in the area and that each runs a success-or-failure dash across the highway. This would probably be too simplistic. Instead the population biologists will think of this as Poisson, where the rate λ is proportional to the squirrel population.
(g) is hypergeometric. This has $N = 280$ and $n = 10$. I.e., something else
(h) is binomial.
(i) is hypergeometric. This has $N = 52$, $n = 13$, and $D = 13$. I.e., something else.
(j) is binomial.

Part III. Normal Distribution

7. It is maintained that, in a quiet equity market with no news, the daily number of shares trades of EquiNimbus Corporation will be approximately normally distributed with mean 280,000 and with standard deviation 32,000. Find the probability that the number of shares traded tomorrow will be at most 325,000.

SOLUTION: Let X be the (random) number of shares traded tomorrow. The assumptions indicate that X will be normally distributed with mean $\mu = 280{,}000$ and with standard deviation $\sigma = 32{,}000$. Then

$$\text{P}[\, X \le 325{,}000 \,] = P\left[\frac{X - 280{,}000}{32{,}000} \le \frac{325{,}000 - 280{,}000}{32{,}000} \right] \approx \text{P}[\, Z \le 1.41 \,]$$

$$= 0.50 + \text{P}[\, 0 \le Z \le 1.41 \,] = 0.50 + 0.4207 = 0.9207$$

The ≈ in this demonstration refers to (*a*) approximating (X-280,000)/32,000 as standard normal. If we were told that X were exactly (rather than approximately) normal, then this distribution would be normal. (*b*) rounding the calculation (325,000 - 280,000)/32,000 − to 1.41.

8. The quantity produced daily at the Milesite cement factory is approximately normally distributed with mean 0.82 and standard deviation 0.14. Production is independent from one day to the next. The units are in *millions* of pounds. Find the probability that the total production for the next 20 days will between16 and 17 million pounds. HINT: The total will be between 16 and 17 if and only if the average is between 16/20 = 0.80 and 17/20= 0.85.

SOLUTION: If $\bar{X}$ denotes the mean of the 20 days, then $\bar{X}$ has a normal distribution with mean 0.82 and with standard deviation $0.14/\sqrt{20} \approx 0.0313$. Then

$$\text{P}[\, 0.80 < X < 0.85 \,] = P\left[\frac{0.80 - 0.82}{0.0313} < \frac{\bar{X} - 0.82}{0.0313} < \frac{0.85 - 0.82}{0.0313} \right]$$

$$\approx \text{P}[\, -0.64 < Z < 0.96 \,] = \text{P}[\, 0 \le Z < 0.64 \,] + \text{P}[\, 0 \le Z < 0.96 \,]$$

$$= 0.2389 + 0.3315 = 0.5704$$

You can also do this directly in terms of the total $T = n\bar{X}$. Just note that the expected value of T is $\text{E}(T) = n\mu = 20 \times 0.82 = 16.40$ and that the standard deviation of T is

$\text{SD}(T) = \sigma\sqrt{n} = 0.14 \times \sqrt{20}\ \ 20 \approx 0.6261$. Then

$$\text{P}[\, 16 < T < 17 \,] = P\left[\frac{16 - 16.40}{0.6261} < \frac{T - 16.40}{0.6261} < \frac{17 - 16.40}{0.6261} \right] \approx \text{P}[\, -0.64 < Z < 0.96 \,]$$

and this gets us to the same solution.

Part IV. Law of Large Numbers.

9. In Notes (slides) 10, we looked at the idea that in estimating a mean of a population, a larger sample is better than a small one. "Better" is quantified in the idea of the "standard error of the mean," which is computed as $\sigma/\sqrt{n}$ for a sample of n observations. A useful question to consider is "how much better?" Suppose I have drawn a sample of 10,000 observations on the number of minutes that arriving flights are late at airports around the world. I find that the sample mean is 24.75 and the sample standard deviation is 9.32. What is the estimate of the standard error of the mean? Now, the question. How much better would you say a sample of 100,000 observations would be?

10. We have found (and will continue to find) many uses for the empirical rule: 95% of almost any distribution will lie within two standard deviations of the mean. One of the ways we use this result is to form a plausible range of values around an estimate of the mean of a population that we can feel accounts for the uncertainty (sampling variability) of that estimator. For the results in problem 1 above, what would you report as your plausible range of values for the true average number of minutes late for flights assuming that the sample used is 10,000 flights?

$\bar{x} = 24.75$, s = 9.32.

Sample size = n = 10,000: $\frac{s}{\sqrt{n}} = \frac{9.32}{\sqrt{10,000}} = .0932$

Confidence interval = $24.75 \pm 1.96(.0932)$ = (24.56 to 24.93)

Sample size = n = 100,000: $\frac{s}{\sqrt{n}} = \frac{9.32}{\sqrt{100,000}} = .0295$

Confidence interval = $24.75 \pm 1.96(.0295)$ = (24.69 to 24.81)

Part V. Statistical Quality Control

11. In the DataStor case, the management gets into some trouble because of an error in statistical methodology made by one of the quality control engineers. What is the error? Describe the problem in a short paragraph. (Note, the authors of the case answer this question specifically in their description of the statistical investigation.)

The quality control engineer recommended using the mean of 8 tests, rather than the individual tests. The bounds on the quality control chart were drawn at + and – 2 standard deviations, but the standard deviation of a mean would be s/sqr(8). So, the bounds were much too far apart, and it looked like the production process was under control when it was not.

Part VII. Extra Problems

EXTRA. At the Sweet Easter Company, internet orders for the very expensive Super Chocolate Bunny ($90 each) come in at a rate of 0.9 per day, and it is believed that this phenomenon can be described as a Poisson random variable with $\lambda = 0.9$. Find

a. the probability that there will be three or more orders on any day.
b. the probability that, in a five-day work week, there will be at least one day on which there are three or more orders.

SOLUTION:
a. Let X be the random number of orders on any day. The probability that $X \geq 3$ is equal to $1 - P[X \leq 2]$. We have, using Minitab, $P[X \leq 2] = 0.93715$. This means that the probability of three or more orders is $1 - 0.9371 = 0.0629 \approx 6\%$.
b. Assuming days are independent, you have 5 independent "trials" in which "success" is having 3 or more orders. The probability of a success is .0629. The question then asks for the probability of having at least one success in 5 trials. That will be 1 minus the probability of having zero successes in 5 trials. That is, the probability of the event in part b is the probability of 5 consecutive days with 2 or fewer orders. This will be $1 - .93715^5 = 0.2773 \approx 27\%$.

EXTRA.. Suppose that Z represents a standard normal random variable. Don't forget that "standard" here says $\mu = 0$ and $\sigma = 1$.

a. Find the probability $P[Z > 1.42]$.
b. Find the probability $P[-0.22 < Z < -0.13]$.
c. Find the probability $P[| Z | \leq 0.90]$.
d. Find the value h for which $P[| Z | > h] = 0.08$.

SOLUTION:
(a) $P[Z > 1.42] = 0.50 - P[0 \leq Z \leq 1.42] = 0.50 - 0.4222 = 0.0778$.
(b) $P[-0.22 < Z < -0.13] = P[0.13 < Z < 0.22] = P[0 \leq Z < 0.22] - P[0 \leq Z < 0.13]$
$= 0.0871 - 0.0517 = 0.0354$.
(c) $P[| Z | \leq 0.90] = 2 \times P[0 \leq Z \leq 0.90] = 2 \times 0.3159 = 0.6318$.
(d) The request $P[| Z | > h] = 0.08$ must b e converted into a form usable in the printed table. Since 0.08 of the probability is shared equally between the positive and negative tails of the distribution, we must have $P[Z > h] = 0.04$. Equivalently we can say $P[0 \leq Z \leq h] = 0.46$. We will now look for the value 0.4600 in the body of the table. We find $P[0 \leq Z \leq 1.75] = 0.4599$; this is as close as we're going to get. Let's use $h = 1.75$. You could interpolate to get a more precise solution, but this effort is generally not worth the trouble.

EXTRA.Suppose that X is a normal random variable with mean 4,500 and with standard deviation 1,000. Find the probability

a. $P[X < 5{,}000]$
b. $P[X > 3{,}500]$
c. $P[4{,}000 \leq X \leq 5{,}000]$
d. $P[| X - 4{,}000 | > 800]$

SOLUTION:

(a) P[$X < 5{,}000$] =

$$P\left[\frac{X-4{,}500}{1{,}000} < \frac{5{,}000-4{,}500}{1{,}000}\right] = P[Z < 0.5] = .5 + P[0 \le Z \le .5] = .5 + .1915 = .6915$$

(b) P[$X > 3{,}500$] =

$$P\left[\frac{X-4{,}500}{1{,}000} > \frac{3{,}500-4{,}500}{1{,}000}\right] = P[Z > -1.0] = .5 + P[0 \le Z \le 1.0] = .5 + .3413 = .8413$$

(c) P[$4{,}000 \le X \le 5{,}000$] = $P\left[\frac{4{,}000-4{,}500}{1{,}000} \le \frac{X-4{,}500}{1{,}000} \le \frac{5{,}000-4{,}500}{1{,}000}\right]$

= P[$-0.5 \le Z \le 0.5$] = $2 \times$ P[$0 \le Z \le 0.5$] = $2 \times 0.1915 = 0.3830$

(d) P[$|X - 4{,}000| > 800$] = P[$(X - 4{,}000) > 800$] + P[$(X - 4{,}000) < -800$]

= P[$X > 4{,}800$] + P[$X < 3{,}200$]

These two probabilities can be developed separately.

P[$X > 4{,}800$] = $P\left[\frac{X-4{,}500}{1{,}000} > \frac{4{,}800-4{,}500}{1{,}000}\right] = P[Z > 0.30]$

= 0.50 - P[$0 \le Z \le 0.30$] = 0.50 - 0.1179 = 0.3821

P[$X < 3{,}200$] = $P\left[\frac{X-4{,}500}{1{,}000} < \frac{3{,}200-4{,}500}{1{,}000}\right] = P[Z < -1.30]$

= P[$Z > 1.3$] = 0.50 - P[$0 \le Z \le 1.3$]

= 0.50 - 0.4032 = 0.0968

Finally, P[$X > 4{,}800$] + P[$X < 3{,}200$] = 0.3821 + 0.0968 = 0.4789.

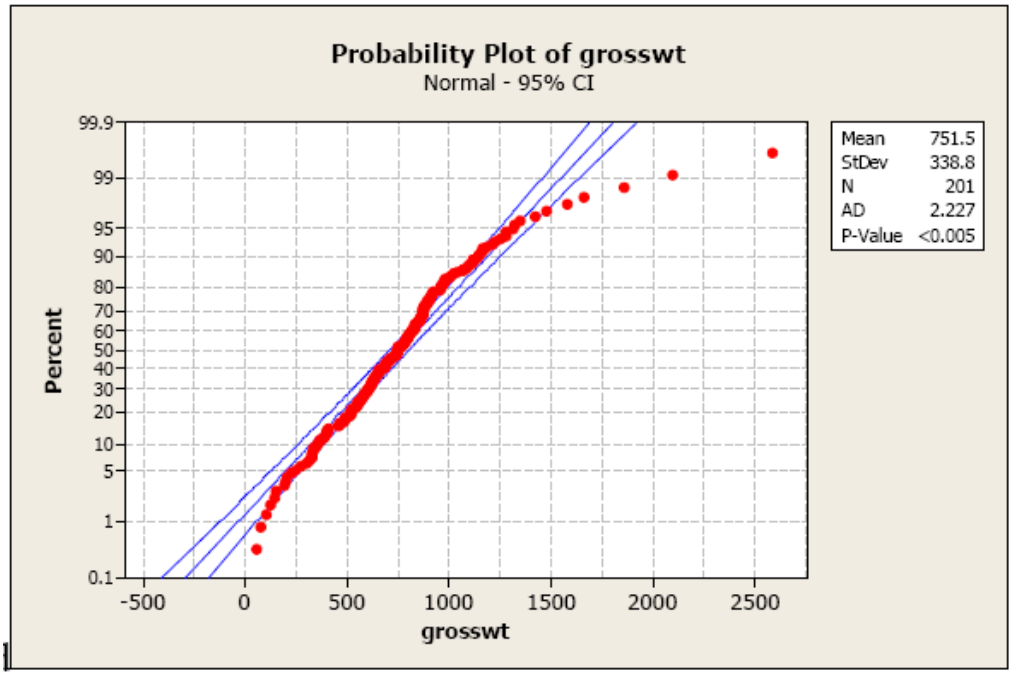
EXTRA.Stan's Deli is situated inside a large industrial park. The weekday gross sales at Stan's average $1,240, with a standard deviation of $180. Find the probability that the average over the next 40 weekdays will exceed $1,200. Please note the assumptions that are used in making the calculation.

SOLUTION: The only assumptions needed are that the sales amounts are independent of each other and come from the same distribution. We have no need to assume that the distribution is normal, as the Central Limit theorem will assure us that X is normal. Let X_1, X_2, …, X_{40} be the random amounts for these 40 weekdays. We must assume that these random variables are statistically independent, each with the mean $1,240 and each with the standard deviation $180. Let $\bar{X}$ be the average of these 40 values. The mean of the distribution of $\bar{X}$ is $1,240 and the standard deviation of this distribution is $180/\sqrt{40} \approx$ $28.46. We then proceed as follows:

P[$\bar{X} >$ \$1,200] = $P\left[\frac{\bar{X}-\$1{,}240}{\$28.46} > \frac{\$1{,}200-\$1{,}240}{\$28.46}\right] \approx$ P[$Z > -1.41$]

= P[$Z < 1.41$] = 0.50 + P[$0 \le Z < 1.41$] = 0.50 + 0.4207 = 0.9207

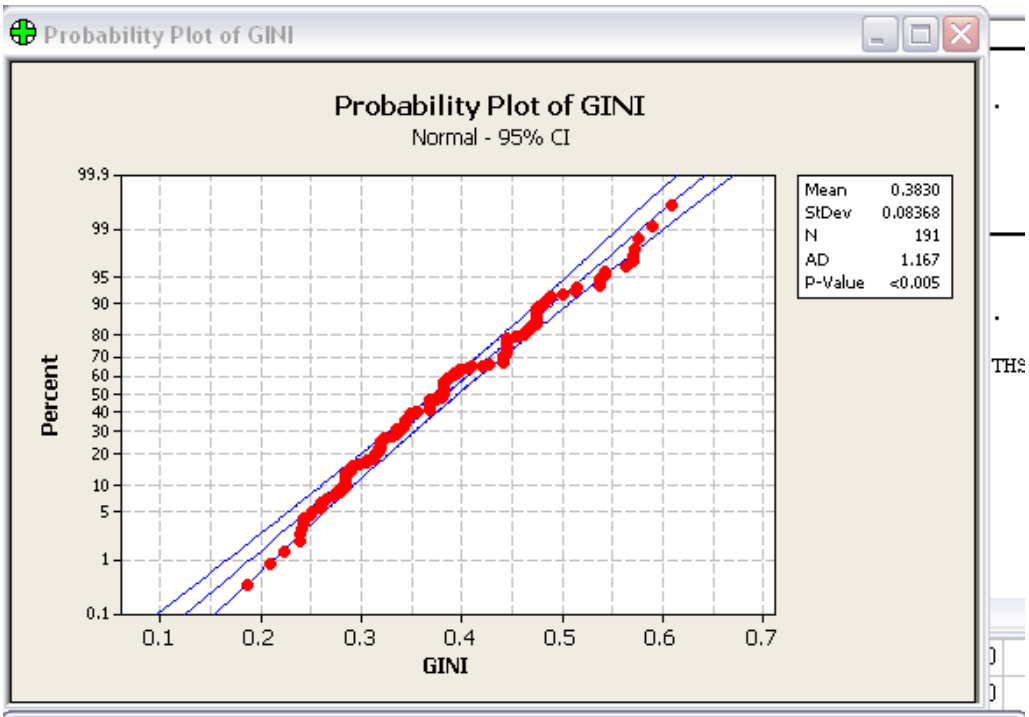
20. We often wish to determine whether our data can be considered normally distributed. There's an approximate graphical method, based on the normal probability plot. This plot is available in Minitab through **Graph** ⇒ **Probability Plot** ⇒ **Single**. Use all the default options and just ask for the single data column you wish to investigate. The display will look like this:

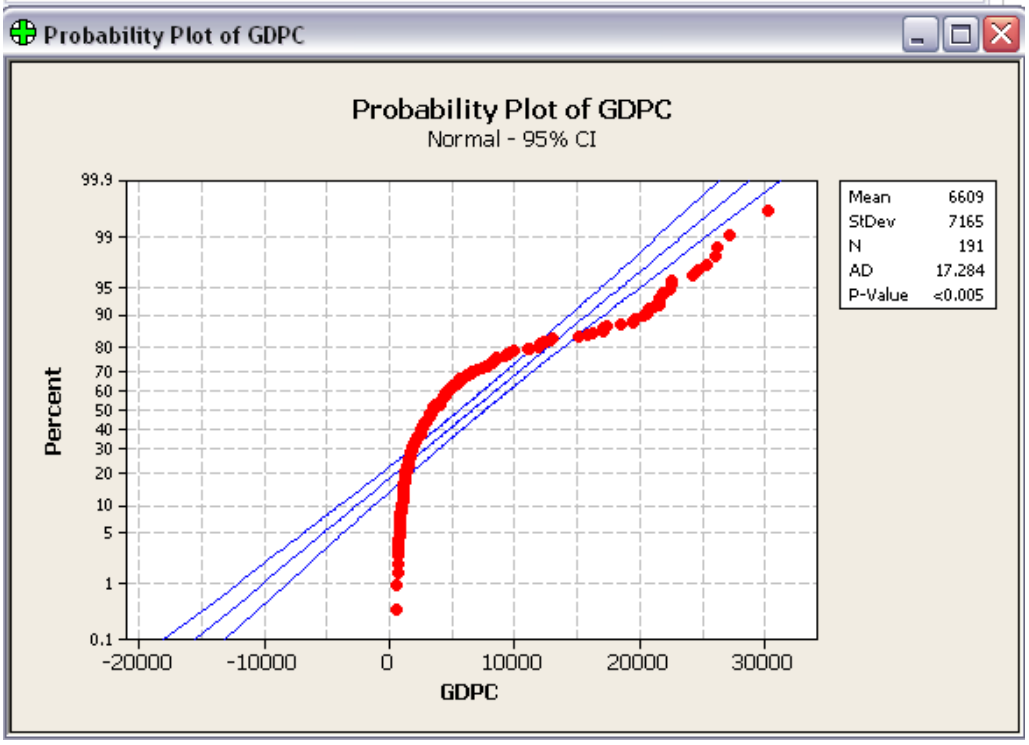


This picture shows the variable named *grosswt* The horizontal axis gives the data scale, so we see that grosswt has a lowest value around 50 to a highest value around 2,600. The vertical axis is "percent of data less than." For example, about 20% of the data values are less than 500, and about 85% of the data values are less than 1,000. The plot is designed so that data from a normal distribution will fall close to the oblique straight line. Because of sampling noise, this will not happen perfectly. We generally agree that the data are acceptably normal when all (or nearly all) the points lie between the curved bands. For the plot above, the data are *not* normally distributed. The lower tail sticks out of the bands, the upper tail sticks out of the bands, and the section of data values are 1,000 wanders outside the bands. The values are positively skewed. The second-largest value, about 2,100 is 2,100 751.5 338.8 − ≈ 4 standard deviations above average, and the largest is even more extreme; this will not happen on a set of normally-distributed data with $n = 201$.

EXTRA.To see a stark comparison of a variable that is approximately normally distributed to one that definitely is not, use the **WHO-HealthStudy.mpj** data file that we used in Homework 1. Produce a normal probability plot for the variagle GINI (Gini coefficient of income inequality). Then do the same for the variable GDPC (per capita gross domestic product).

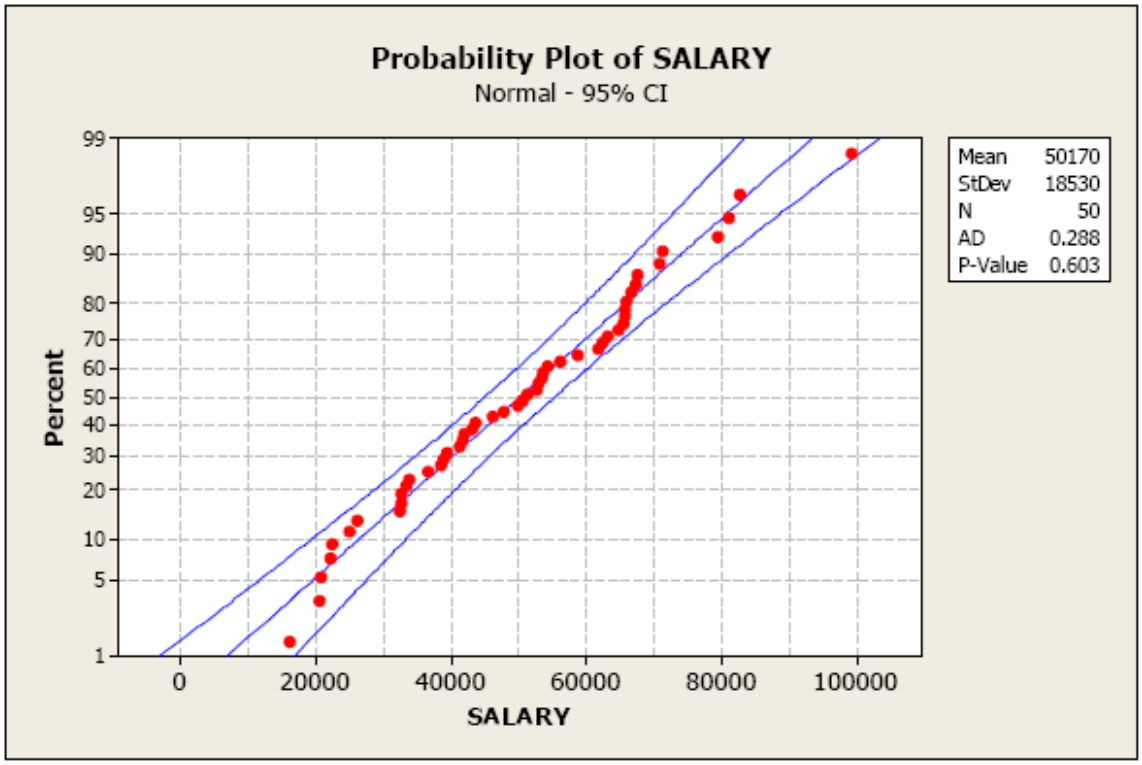
SOLUTION. The GINI coefficient appears to fit well within the bounds predicted by the normal distribution. The GDPC data definitely do not. We saw the skewness of this variable in assignment 1, which is once again reflected here.


Probability Plot of GINI
Probability Plot of GINI
Normal - 95% CI
Mean 0.3830
StDev 0.08368
N 191
AD 1.167
P-Value <0.005
Percent
GINI
99.9
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7


Probability Plot of GDPC
Probability Plot of GDPC
Normal - 95% CI
Mean 6609
StDev 7165
N 191
AD 17.284
P-Value <0.005
Percent
GDPC
99.9
99
95
90
80
70
60
50
40
30
20
10
5
1
0.1
-20000
-10000
0
10000
20000
30000

EXTRA.Examine the salary data in file **salary.mpj** (on the course outline). Would you consider the variable SALARY to have a normal distribution?

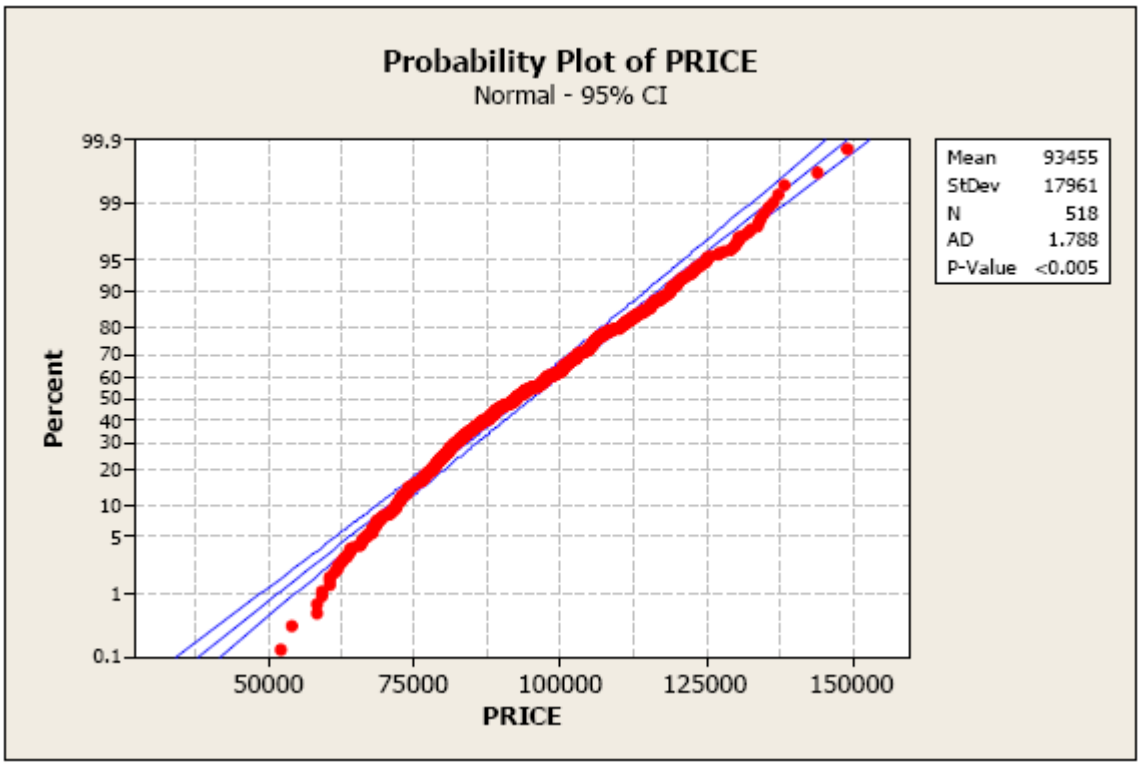
SOLUTION: The normal probability plot looks like this:



The points stay inside the curved bands, so we'll have to say that the values seem to be sampled (as far as we can tell) from a normal population.

EXTRA.. Consider the data in file **Easton.mpj**. Would you consider the variable *Price* to come from a normal population? This judgment is somewhat harder than the previous.

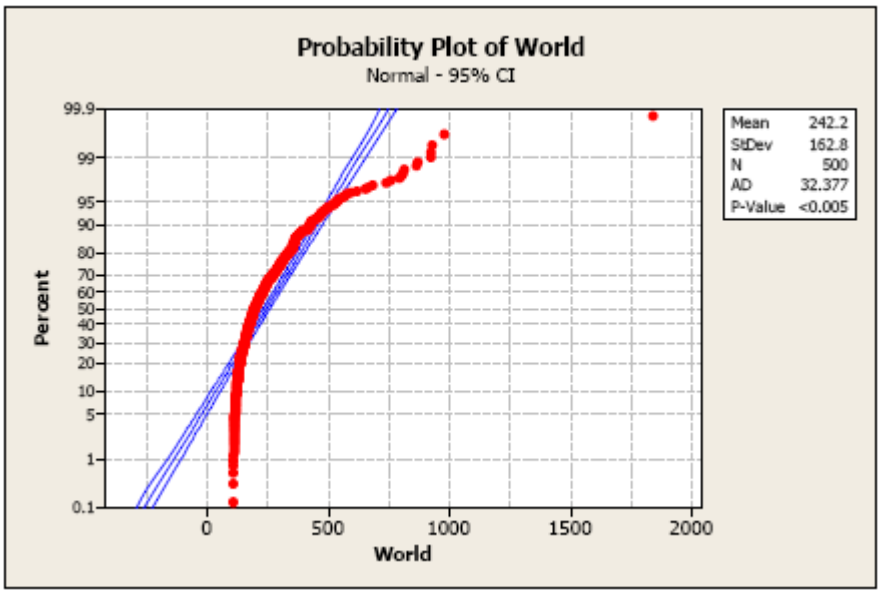
SOLUTION: The probability plot is the following:



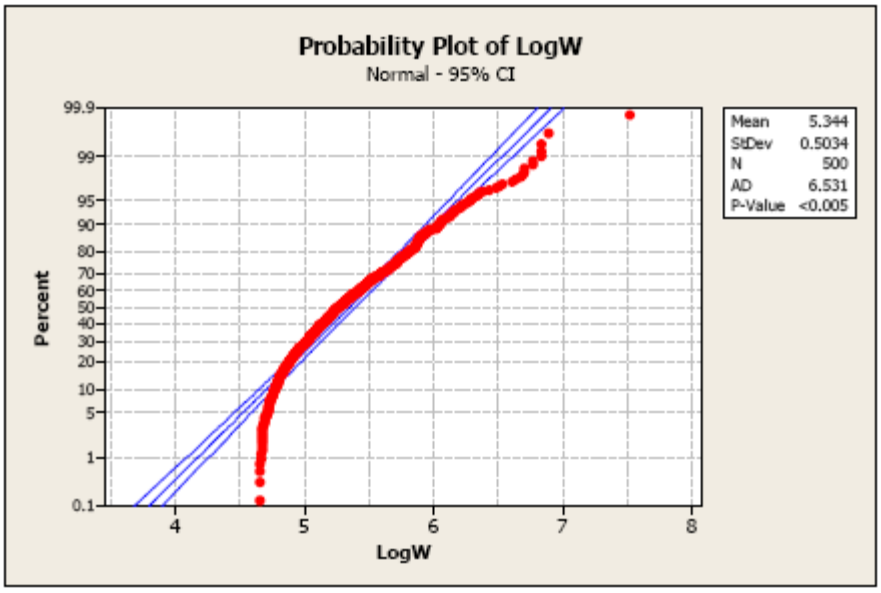
At the low end, there are about eight or ten points that stay outside the bands. The sample size however is very large at $n = 518$. Most statisticians would be willing to treat these data as normal.

EXTRA. Examine the file **Movies9OCT2003.mpj**. Consider the data column World (for world-wide gross movie revenues). Would those values be considered normally distributed?

SOLUTION: The picture is this:



This is outrageously non-normal. Sometimes this type of non-normality is cured by taking logarithms. Here that doesn't work, as you can see from the following:



Assignment 2

EXTRA.. The figure below (that we discussed in class) is a probability tree that (it is suggested) is to be used to help a potential litgant decide whether to settle or to litigate.a case. For now, let's focus on the center of the tree and ignore the decision branch (Litigate,Do Nothing) and the issue of Causation. Use the following symbols: L = Jury Finds Liability, ~L = No Finding of Liability, F = Find Document, ~F = Do Not Find document. Reading off the figure, can see that P(L|F) = .6, P(~L|F) = .4, P(L|~F)=.3, P(~L|~F) = .7. We also see that P(F) = .4 and P(~F)=.6. So, the picture and the figures show the probabilities that the jury will (or won't) find liability given that a document was found (or not). Use Bayes Theorem to reverse this calculation. What is the probability that a document will be found given that the jury finds liability. That is, compute P(F|L) from these results.

SOLUTION

We have directly from the figure
P(L|F) = .6,
P(~L|F) = .4,
P(L|~F)=.3,
P(~L|~F) = .7.

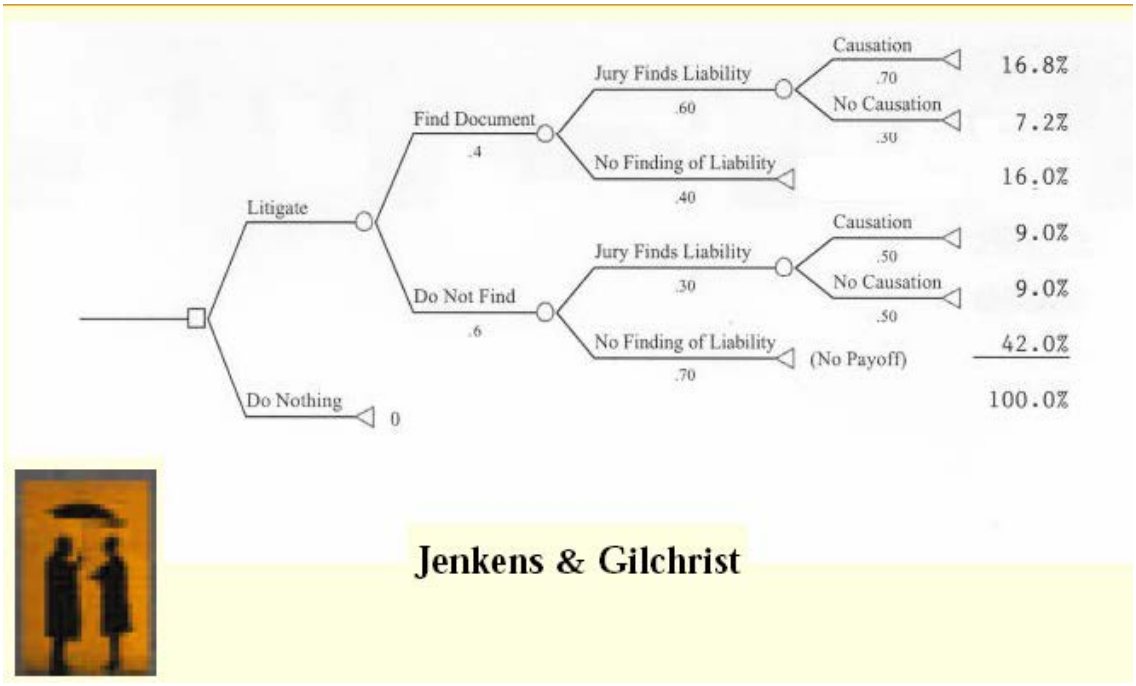
We also see that P(F) = .4 and P(~F)=.6. We seek P(F|L).

Using the definition,

P(F|L) = P(F and L)/P(L).

Using Bayes Theorem,

P(F|L) = P(L|F)P(F) / [P(L|F)P(F) + P(L|~F)P(~F)] = [.6(.4)]/[.6(.4)+.3(.6)] = 4/7 = .572.



Assignment 2

EXTRA.. (This is Exercise HOG 3.14, page 97.) A survey of workers in two plants of a manufacturing firm includes the question, "How effective is management in responding to legitimate grievances of workers?" In plant 1, 48 of 192 workers respond "poor"; in plant 2, 80 of 248 workers respond "poor." An employee of the manufacturing firm is to be selected randomly. Let *A* be the event "worker comes from plant 1" and let *B* be the event "response is poor."

a. Find P(*A*), P(*B*), and P(*A* and *B*).

b. Are the events *A* and *B* independent?

c. Find P(*B* | *A*) and P(*B* | not *A*). Are they equal?

You might find it helpful to create this display:

	Poor	Not Poor	TOTAL
Plant 1	48	144	192
Plant 2	80	168	248
TOTAL	128	312	440

SOLUTION:

(a) Observe that there are 192 + 248 = 440 workers in all. Also, there are 48 + 80 = 128 "poor" responses in all. Then P(*A*) = 192/440 ≈ 0.4364, P(*B*) = 128/440 ≈ 0.2909, and P(*A* ∩ *B*) = 48/440 ≈ 0.1091.

(b) Note that P(*A*) × P(*B*) = 0.4364 × 0.2909 ≈ 0.1269, and this is *not* equal to P(*A* ∩ *B*) = 0.1091. The values 0.1091 and 0.1269 are not very close.

(c) P(*B* | *A*) = 48/192 = 0.25 and P(*B* | *A*′) = 80/248 ≈ 0.3226. These are unequal, and it supports the non-independence noted in part (b).

EXTRA.A survey of female visitors, ages 15 to 18, to an upscale suburban shopping mall asked these questions:

Do you have your own cell phone (not shared with others)?

Do you have a portable CD music player (Walkman or iPod or similar device)?

Do you have a body piercing somewhere besides your ears?

The responses were the following:

YES to	Count
cell phone, portable CD, piercing	18
cell phone, portable CD	37
cell phone, piercing	19
portable CD, piercing	9
cell phone	54
portable CD	15
piercing	2
(none)	14
TOTAL	168

Assuming that the sampled group is representative of some population, estimate

a. P(piercing)
b. P(cell phone)
c. P(cell phone | piercing)
d. P(piercing | cell phone)
e. Indicate which is bigger P(piercing | cell phone) or P(piercing | portable CD).

SOLUTION:

(a) The number with piercings is 18 + 19 + 9 + 2 = 48, so the estimated probability is 48/168 ≈ 0.2857.

(b) The number with cell phones is 18 + 37 + 19 + 54 = 128, and the estimated probability is 128/168 ≈ 0.7619.

(c) Among the 48 with piercings, there were 18 + 19 = 37 with cell phones, so this conditional probability is 37/48 ≈ 0.7708.

(d) Among the 128 will cell phones, there were 18 + 19 = 37 with piercings, so this conditional probability is 37/128 ≈ 0.2891.

(e) There are 18 + 37 + 9 + 15 = 79 with portable CD devices, and 18 + 9 = 27 of these also had piercings. Thus P(piercing | portable CD) is estimated by 27/79 ≈ 0.3418. It seems that P(piercing | portable CD) > P(piercing | cell phone), at least for this data set.

A note on precision:

There were 168 teenagers in this data set. The number 168 has three significant figures, so it seems odd to report results to four significant figures, as given in these solutions. This extra precision, seemingly pointless, guarantees that we can reconstruct the original counts from the reported proportions. You can see the difficulty imposed by reporting the numbers to whole percents (two significant figures). The ratio 45/168 ≈ 0.2678 would have been rounded to 27%.

The ratio 46/168 ≈ 0.2738 would also have been rounded to 27%. A report that indicated “27% of the 168 subjects said that . . .” would not allow you to reconstruct the original count.