Answers

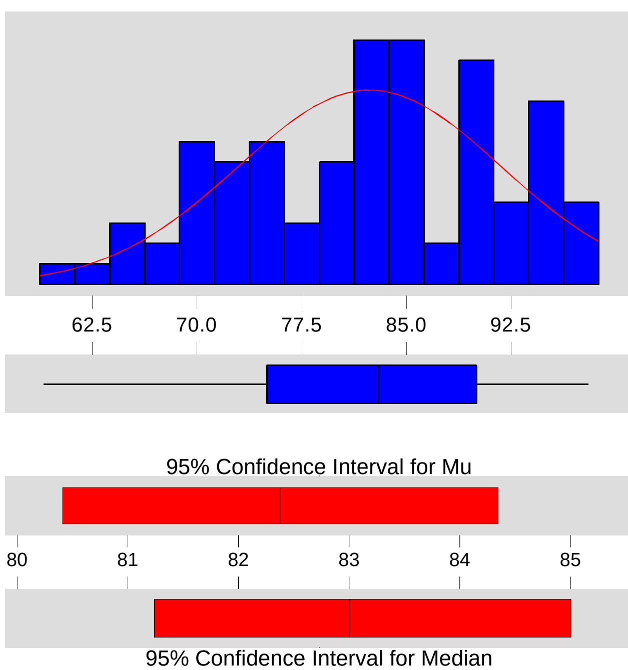
1. Let S be the S&P return for a randomly chosen day and $\overline{S}$ be the mean return for a random sample of 20 days.
   (a) $E(\overline{S}) = E(S) = .00032$
   (b) $SD(\overline{S}) = SD(S)/\sqrt{n} = .00859/\sqrt{20} = .00192$
   (c)

$$P(\overline{S} > .005) = P\left(Z > \frac{.005 - .00032}{.00192}\right) = P(Z > 2.438) = .00074$$

Without doing any calculations we know that it is more likely that an individual day's return will be greater than .005 than that the mean return will be, since the standard deviation of $\overline{S}$ is smaller than that of S.

2.(a) The graphical output for descriptive statistics gives all of the information we need:

Descriptive Statistics



Variable: Freshman

Anderson-Darling Normality Test	
A-Squared:	0.794
P-Value:	0.038
Mean	82.3778
StDev	9.3983
Variance	88.3276
Skewness	-3.2E-01
Kurtosis	-7.8E-01
N	90
Minimum	59.0000
1st Quartile	75.0000
Median	83.0000
3rd Quartile	90.0000
Maximum	98.0000
95% Confidence Interval for Mu	
80.4093	84.3462
95% Confidence Interval for Sigma	
8.1973	11.0148
95% Confidence Interval for Median	
81.2373	85.0000

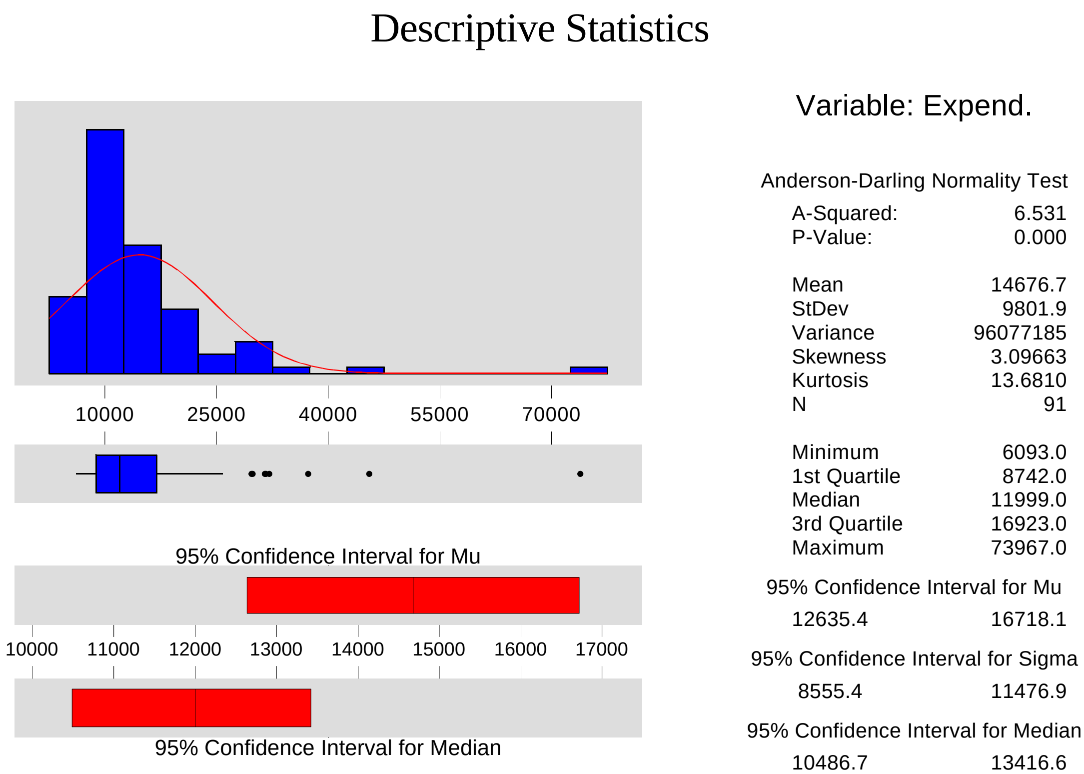
So, the 95% confidence interval for the true average freshman retention rate is (80.409, 84.346). The interval requires that the variable be reasonably normally distributed, and that the sample be a random sample from a larger population. The latter requirement is met, and the former is not too bad.

(b) The prediction interval is then

$$\overline{X} \pm t^{89}_{.025} s \sqrt{1 + \frac{1}{90}} = 82.378 \pm (1.987)(9.398)\sqrt{1.0111} = 82.378 \pm 18.777 = (63.601, 101.155).$$

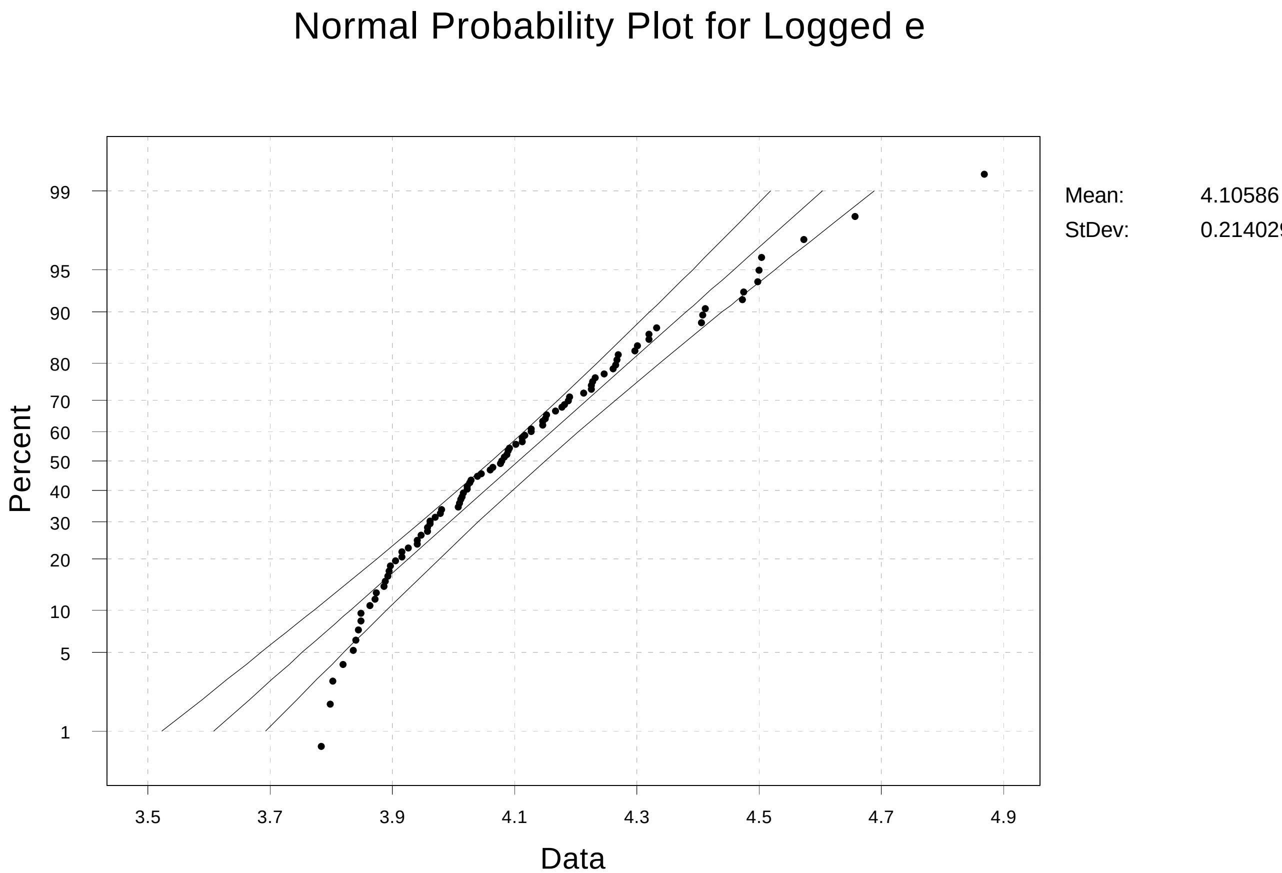
Note that the upper limit of the interval goes past the highest possible value of 100, which reflects the slight left tail in the distribution. The actual value for NYU, by the way, was 86.

(c) Here is a graphical representation of the expenditure variable:



The confidence interval for average (over all research universities) mean (over all students within the university) expenditure per student is thus (12635, 16718). This interval is not very useful, however, since the expenditure variable is very long right–tailed.

(d) Given the long right tail in the distribution, a prediction interval must be constructed in the correct scale. The way to do it is to log the variable, construct the prediction interval, and then antilog. Here is a normal plot for the logged expenditure variable, which looks more Gaussian (it's still not great, being right–tailed, but there's no stronger transformation than the log to try):



The prediction interval in the logged scale has the form

$$\overline{X} \pm t^{90}_{.025} s \sqrt{1+\frac{1}{91}} = 4.106 \pm (1.987)(.214)\sqrt{1.011} = 4.106 \pm .4276 = (3.678, 4.534).$$

Antilogging this interval gives $(10^{3.678}, 10^{4.534})$, or $(4764, 34198)$. The actual value for NYU was 24135. Note that this interval is much more reasonable than the one done in the original scale, which is

$$14677 \pm (1.987)(9802)(1.0055) = 14677 \pm 19584 = (-4907, 34261).$$

(e) In the sample 30 of the 90 institutions have freshman retention rate less than 80%, so a confidence interval for the probability that a research university has retention rate less than 80% is

$$\overline{p} \pm z_{.025}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = \frac{1}{3} \pm (1.96)\sqrt{\frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}{90}} = .333 \pm .0974.$$

The continuity–corrected interval adds $1/180 = .0056$ to the $\pm$ portion of the interval, giving $.333 \pm .103$.

3.(a) From the marginal distribution determined earlier, the confidence interval has the form

$$\frac{203}{362} \pm (1.645)\sqrt{\frac{\left(\frac{203}{362}\right)\left(1-\frac{203}{362}\right)}{362}},$$

or $.561 \pm .043 = (.518, .604)$. The continuity–corrected interval is $.561 \pm (.043 + .001) = (.517, .605)$.

(b) Using the joint distribution of employment opportunities and salary size, and counting the number of respondents who rated salary as more important than employment opportunities gives the confidence interval

$$\frac{58}{362} \pm (1.645)\sqrt{\frac{\left(\frac{58}{362}\right)\left(1 - \frac{58}{362}\right)}{362}},$$

or $.16 \pm .032 = (.128, .192)$. The continuity–corrected interval is $.16 \pm (.032 + .001) = (.127, .193)$.

4.(a) Here is MINITAB output:

```
Confidence Intervals

Variable      N       Mean     StDev  SE Mean        95.0 % CI
Weekly p     88      0.880     8.186    0.873  (  -0.855,    2.614)
```

The confidence interval for true average price change includes zero, so the observed value is not significantly different from zero. The price change variable is reasonably Gaussian here.

(b) A prediction interval for the price change of a randomly chosen stock is $.88 \pm (1.988)(8.186)\sqrt{1.011} = .88 \pm 16.363$. The wide interval reflects the large variability of stock prices.

(c) Here are values separated by market:

```
Descriptive Statistics

Variable  Market           N      Mean   Median  Tr Mean    StDev  SE Mean
Weekly p  OTC             51      0.64     0.00     0.09     9.43     1.32
          NYSE            24     -0.03    -0.85    -0.58     5.96     1.22
          ASE             13      3.49     1.30     2.79     6.18     1.71

Variable  Market         Min       Max       Q1       Q3
Weekly p  OTC         -19.40     34.20    -5.00     4.40
          NYSE         -7.60     19.70    -4.55     1.38
          ASE          -4.10     18.80    -1.05     6.30
```

The confidence intervals for each market are as follows:

$$\text{OTC} : .64 \pm (2.009)(9.43)/\sqrt{51} = .64 \pm 2.65$$
$$\text{NYSE} : -.03 \pm (2.069)(5.96)/\sqrt{24} = -.03 \pm 2.52$$
$$\text{ASE} : 3.49 \pm (2.179)(6.18)/\sqrt{13} = 3.49 \pm 3.73$$

All three intervals contain zero, so the average returns in each were not significantly different from zero. Note that the ASE interval was close, but didn't quite make it because of the small sample size. A sample of size 15 with the same mean and standard deviation would have given an interval that did not contain zero. The distributions of price change within each group are reasonably Gaussian here.

(d) Here are prediction intervals:

$$\text{OTC} : .64 \pm (2.009)(9.43)(1.01) = .64 \pm 19.134$$
$$\text{NYSE} : -.03 \pm (2.069)(5.96)(1.021) = -.03 \pm 12.59$$
$$\text{ASE} : 3.49 \pm (2.179)(6.18)(1.038) = 3.49 \pm 13.978$$

Note that the intervals for NYSE and ASE are narrower than that for all of the stocks together, which reflects the gains in treating the markets separately (despite the reduction in effective sample size).

5.(a) The mean of $\overline{X}$ is just the population mean μ, or .00039.

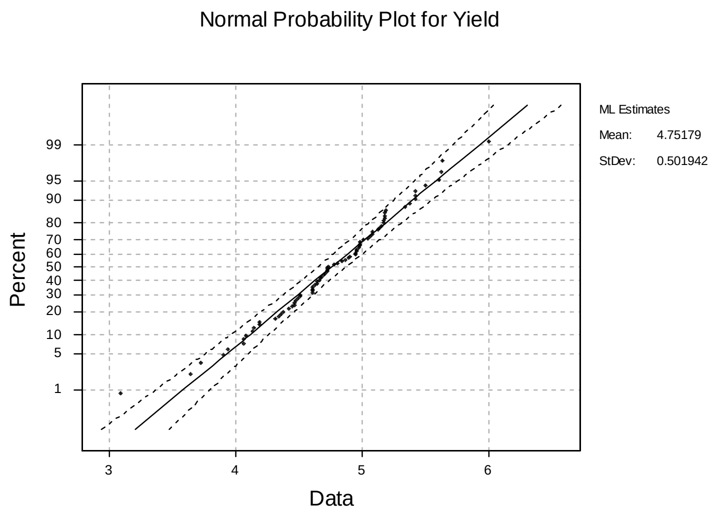
(b) The standard deviation (or standard error) of $\overline{X}$ is $\sigma/\sqrt{n}$, or $.00721/\sqrt{30} = .00132$.

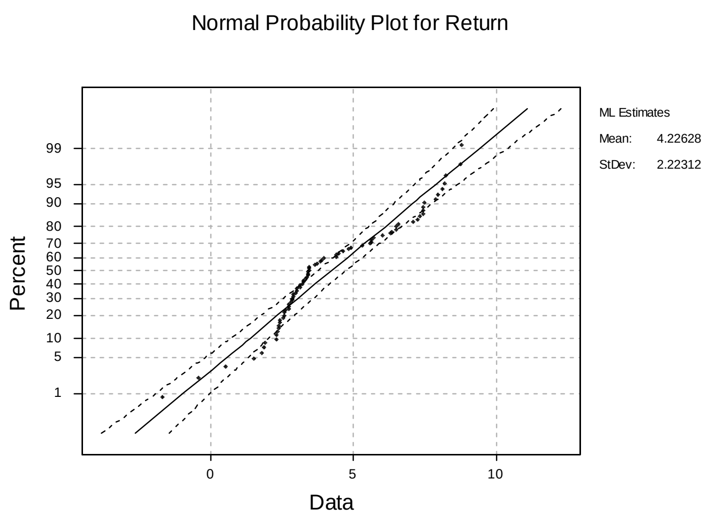
(c)

$$P(\overline{X} > .01) = P\left(Z > \frac{.01 - .00039}{.00132}\right)$$
$$= P(Z > 7.28)$$
$$\approx 0$$

It is more likely that the sample mean return will be less than .01 than an individual return, since the mean is much more tightly centered around the population mean. In fact, the probability of individual return being less than .01 is about .91.

6. Before we do anything, we should recognize that we're assuming that our data are a random sample from a roughly Gaussian distribution. Normal plots for return and yield show that the while the yield variable is reasonably Gaussian (with one possibly unusual low value), returns are not, being bimodal. In fact, as the structure of the data in the file suggests, there are two well–defined subgroups here: no load funds, and load funds.






```
Stem-and-leaf of Return    N  = 78
Leaf Unit = 0.10
```

```
   1    -1 7
   2    -0 4
   3     0 5
   7     1 5789
  27     2 33333445556777888889
 (20)    3 00112223344444467889
  31     4 445689
  25     5 36667
  20     6 023456
  14     7 023444489
   5     8 11278
```

(a) We want a confidence interval, which has the form $\overline{X} \pm t_{\alpha/2}^{(n-1)} s/\sqrt{n}$. Since $n = 78$, $t_{\alpha/2}^{(n-1)} = 1.99$. Here is the answer from Minitab:

```
Variable     N      Mean    StDev  SE Mean       95.0 % CI
Return      78     4.226    2.238    0.253  (   3.722,   4.731)
```

The nonnormality of returns here is less worrying, since the sample size is somewhat large, so this interval might be okay. On the other hand, the bimodal nature of the distribution leaves open the question of whether estimating a single mean is reasonable anyway.

(b) Now we want a prediction interval, which has the form $\overline{X} \pm t_{\alpha/2}^{(n-1)} s\sqrt{1 + 1/n}$. From the output above, we thus obtain an interval of $4.226 \pm (1.99)(2.238)\sqrt{1.013} = 4.226 \pm 4.482 = (-.256, 8.708)$. The negative return is ridiculous, of course, but that just reflects the nonnormality of returns.

(c) We want a confidence interval; here's the output:

```
Variable     N      Mean    StDev  SE Mean       95.0 % CI
Yield       78    4.7518   0.5052   0.0572  (  4.6379,  4.8657)
```

Given the reasonable normality of the sample, this is probably a reasonable representation of the average yield. Note that yields are larger than returns on average.

(d) The prediction interval for a specific yield is

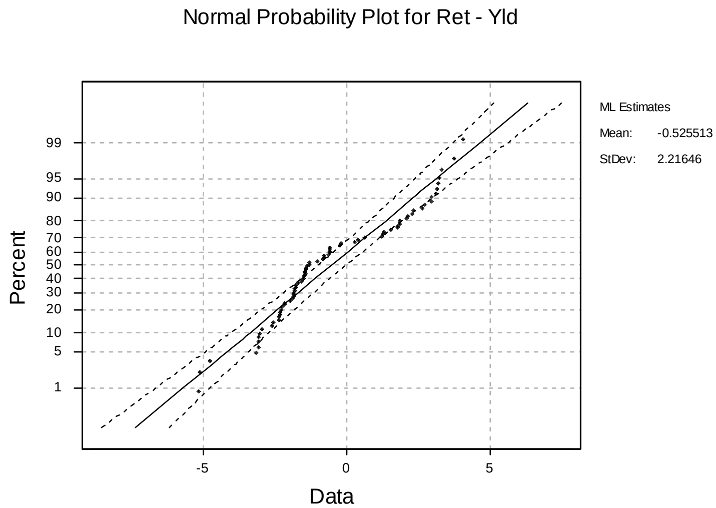
$$4.752 \pm (1.99)(0.505)\sqrt{1.013} = 4.752 \pm 1.011 = (3.741, 5.763).$$

Yields are much less variable than returns, which is interesting, given the very different kinds of costs of different funds. This suggests that funds believe that customers just look at the yield (because they don't recognize that it is the return that actually determines how much money you make).

(e) You can't just compare the two confidence intervals above to answer this question, since the key source of variability is in the *difference* between yield and return. Since the difference between mean yield and mean return is just the mean of the difference between yield and return, the trick is to construct a confidence interval using the variable `Return` − `Yield`. Here's the resultant confidence interval:

```
Variable     N      Mean    StDev  SE Mean       95.0 % CI
Ret - Yl    78    -0.526    2.231    0.253  (  -1.028,  -0.023)
```

Here's a normal plot of this variable, which shows some of the same bimodality as in the Return distribution:



The confidence interval for the mean difference between return and yield does not include zero, which implies that yields are significantly greater than returns. The moral is to look at the return, not the yield!

7. We're constructing 90% confidence intervals for a binomial, so the key critical value is $z_{.05} = 1.645$. I'll give below the exact interval as constructed by Minitab, and the "+2" interval constructed by hand.

(a) We have 29 out of 30 favorable reports from people who received money, so $\tilde{p} = (29+2)/(30+4) = .912$. The confidence intervals are as follows:

```
Sample       X      N  Sample p        90.0 % CI
1           29     30  0.966667  (0.851404, 0.998292)
```

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n} = .912 \pm .085 = (.827, .997)$$

(b) We have 10 out of 17 neutral reports from people who received money, so $\tilde{p} = (10+2)/(17+4) = .571$. The confidence intervals are as follows:

```
Sample       X      N  Sample p        90.0 % CI
1           10     17  0.588235  (0.364009, 0.788092)
```

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n} = .571 \pm .197 = (.374, .768)$$

(c) We have 8 out of 23 critical reports from people who received money, so $\tilde{p} = (8+2)/(23+4) = .37$. The confidence intervals are as follows:

```
Sample       X      N  Sample p        90.0 % CI
1            8     23  0.347826  (0.186344, 0.540456)
```

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n} = .37 \pm .166 = (.204, .536)$$

Note that in all three cases the "+2" intervals are pretty similar to the exact ones.

(d) We haven't talked about any way to formally assess this, but based on the confidence intervals, it certainly seems that it's much more likely to find authors who've received money among the favorable reports than among the critical reports, with neutral reports somewhere in between. In fact, formal investigation of this pattern confirms that there are real differences in these proportions between the three types of reports, so there is a relationship between monetary support and how favorable the report is.

8.(a) The confidence interval output shows that 0 is inside the interval, so the average is not significantly different from zero (the distribution of price changes is reasonably Gaussian, so this interval should be okay, as should the one in (b)).

```
Variable      N       Mean     StDev   SE Mean         95.0 % CI
Weekly p     30       1.30      7.49      1.37   (    -1.50,     4.10)
```

(b) The prediction interval is

$$1.30 \pm (2.05)(7.49)\sqrt{1.033} = 1.30 \pm 15.61 = (-14.31, 16.91).$$

Note the large variability in price changes.

(c) Here is the output we need for parts (c) and (d):

Variable	Exchange	N	Mean	Median	TrMean	StDev
Weekly p	AMEX	6	-5.95	-4.86	-5.95	6.31
	NASDAQ	10	5.99	8.53	6.05	8.86
	NYSE	14	1.05	0.42	0.60	3.79

Variable	Exchange	SE Mean	Minimum	Maximum	Q1	Q3
Weekly p	AMEX	2.57	-14.29	3.39	-12.35	-1.69
	NASDAQ	2.80	-6.96	18.48	-2.37	12.90
	NYSE	1.01	-3.32	10.92	-1.71	3.51

Thus, confidence intervals for average price change for each exchange are as follows:

$$\text{AMEX} : -5.95 \pm (2.57)(6.31/\sqrt{6}) = -5.95 \pm 6.60$$
$$\text{NASDAQ} : 5.99 \pm (2.26)(8.86/\sqrt{10}) = 5.99 \pm 6.33$$
$$\text{NYSE} : 1.05 \pm (2.16)(3.79/\sqrt{14}) = 1.05 \pm 2.18$$

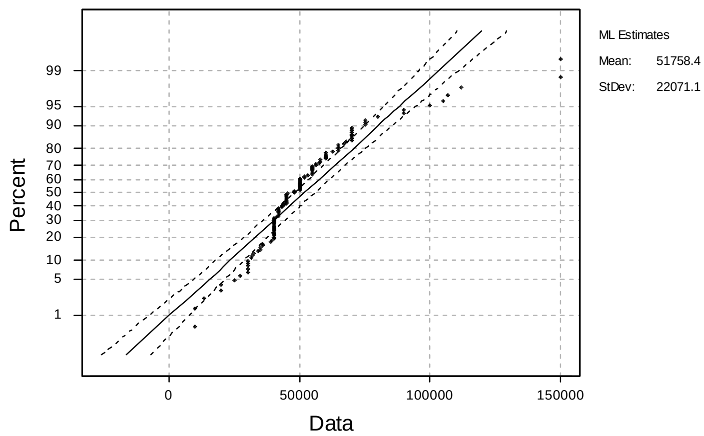
Thus, while the AMEX interval is almost completely below 0 and the NASDAQ interval is almost completely above it, none of the intervals actually exclude zero.

(d) Here are the three prediction intervals:

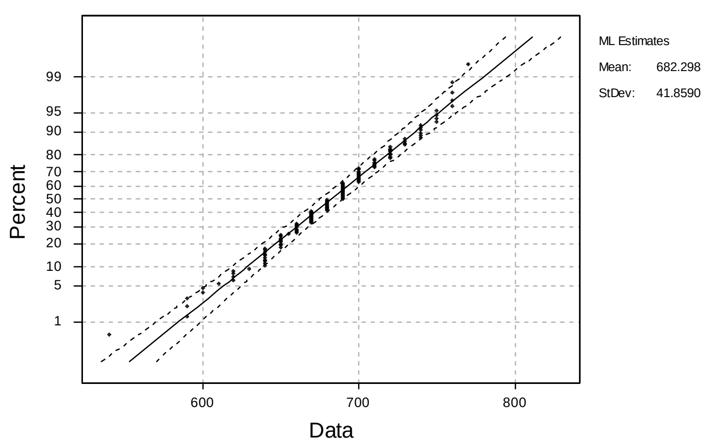
$$\text{AMEX} : -5.95 \pm (2.57)(6.31)\sqrt{1 + 1/6} = -5.95 \pm 17.52$$
$$\text{NASDAQ} : 5.99 \pm (2.26)(8.86)\sqrt{1 + 1/10} = 5.99 \pm 21.00$$
$$\text{NYSE} : 1.05 \pm (2.16)(3.79)\sqrt{1 + 1/14} = 1.05 \pm 8.47$$

9. Before we do anything, we should recognize that we're assuming that our data are a random sample from a roughly Gaussian distribution. Normal plots for previous salary and GMAT show that the while the GMAT variable is reasonably Gaussian (with one possibly unusual low value), past salaries are not, being right–tailed and (maybe) bimodal. The latter possibility can be seen from the hint of two separate lines in the normal plot.

Normal Probability Plot for Previous sal



Normal Probability Plot for GMAT



(a) We want a confidence interval, which has the form $\overline{X} \pm t_{\alpha/2}^{(n-1)} s/\sqrt{n}$. Since $n = 119$, $t_{\alpha/2}^{(n-1)} = 1.98$. Here is the answer from `Minitab`:

```
Variable     N      Mean    StDev  SE Mean       95.0 % CI
Previous   119     51758    22164     2032  (   47735,   55782)
```

The nonnormality of salaries here is not very worrying, since the sample size is reasonably large, so this interval is probably okay.

(b) Now we want a prediction interval, which has the form $\overline{X} \pm t_{\alpha/2}^{(n-1)} s\sqrt{1+1/n}$. From the output above, we thus obtain an interval of $51758 \pm (1.98)(22164)\sqrt{1.008} = 51758 \pm 44060 = (7698, 95818)$. The values seem a bit small (although the actual minimum value in the dataset is \$10,000, only three people had previous salary less than \$20,000; the maximum is \$150,000, and three people had previous salary over \$110,000). That might just reflect the nonnormality of the salaries. We might consider logging previous salary and constructing a prediction interval in the logged space, since this reduces the right tail. Here are the numbers we need:

```
Variable            N         N*      Mean     Median     TrMean     StDev
Logged p          119          6    4.6780     4.6812     4.6833    0.1822
```

Thus, a prediction interval for logged previous salary is

$$4.678 \pm (1.98)(.1822)\sqrt{1.008} = 4.678 \pm .3622 = (4.3158, 5.0402);$$

antilogging gives a prediction interval for previous salary of (20692, 109698), which seems more in line with the observed data.

(c) We want a confidence interval; here's the output:

```
Variable      N       Mean    StDev  SE Mean        95.0 % CI
GMAT        124     682.30    42.03     3.77  (  674.83,  689.77)
```

Given the reasonable normality of the sample, this is probably a good representation of average GMAT.

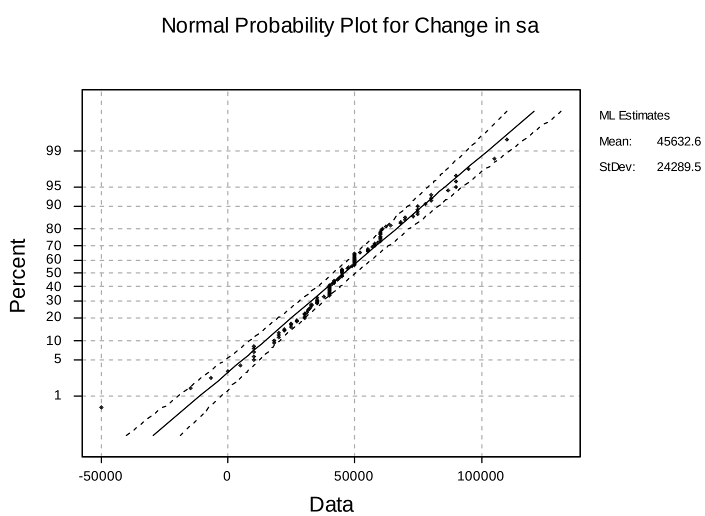
(d) The prediction interval for a specific GMAT is

$$682.3 \pm (1.98)(42.03)\sqrt{1.008} = 682.3 \pm 83.6 = (598.7, 765.9).$$

(e) You can't just compare the confidence intervals for the two salary variables to answer this question, since the key source of variability is in the *difference* between previous and anticipated salary. Since the difference between mean previous salary and mean anticipated salary is just the mean of the difference between previous salary and anticipated salary, the trick is to construct a confidence interval using the variable `Anticipated salary − Previous salary`. Here's the resultant confidence interval:

```
Variable      N       Mean    StDev  SE Mean        95.0 % CI
Change i    115      45633    24396     2275  (   41126,   50139)
```

Here's a normal plot of this variable, which looks reasonable. Note that three people expected to be earning after graduation less than what they were earning before they entered Stern, including one person who expects to be making \$50,000 less(!):



The confidence interval for the mean change in salary includes \$50,000, which implies that \$50,000 is not unreasonable as a guess for the average expected increase in salary for members of the class of 2000.

10. We're constructing 99% confidence intervals for a binomial, so the key critical value is $z_{.005} = 2.576$. I'll give the exact interval as constructed by `Minitab`, and the "+2" interval constructed by hand. We

have 735 out of 1007 adults who believe in the conspiracy, so $\tilde{p} = (735 + 2)/(1007 + 4) = .729$. The confidence intervals are as follows:

```
Sample      X      N  Sample p        95.0 % CI
1         735   1007  0.729891  (0.701324, 0.757106)
```

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1-\tilde{p})/n} = .729 \pm .036 = (.693, .765)$$

As you would expect with such a large sample, the two intervals are very similar to each other.

11. The only trick here is to recognize that you can get `Minitab` to convert values to zeroes or ones based on the conditions given (the easiest way to do this is using the `Calculator`). After that, the program can be used to construct exact confidence intervals:

```
    Variable          X      N  Sample p        95.0 % CI
(a) Cost over 40     48    161  0.298137  (0.228701, 0.375172)
(b) Food over 20     46    161  0.285714  (0.217379, 0.362122)
(c) All over 20       9    161  0.055901  (0.025877, 0.103462)
```

Presumably those 9 restaurants in part (c) are worth checking out!

12. The key point to recognize here is that it is differences between mid cap or small cap returns and S&P returns that matter in parts (a) and (c), respectively, and the difference between mid cap and small cap return that matters in part (d). Then, if 0 is in the confidence interval, the average returns are not different from each other, while if it is not, they are.
   (a) The confidence interval output shows that 0 is inside the interval, so the average is not significantly different from zero. That is, mid caps do slightly worse than the S&P Composite during corrections, but not significantly so.

```
Variable     N      Mean    StDev  SE Mean        95.0 % CI
Correcti    14     -0.70     5.98     1.60  (   -4.16,     2.76)
```

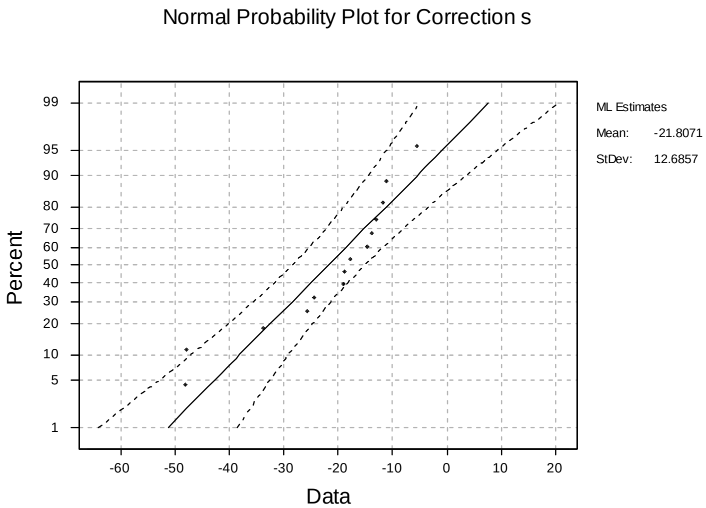
   There is, however, one outlier: from October 1976 through February 1978 the S&P dropped 11.4%, while mid caps actually went up 4.9%. If this observation is omitted from the dataset, here is the result:

```
Variable     N      Mean    StDev  SE Mean        95.0 % CI
Correcti    13    -2.008    3.587    0.995  (  -4.176,    0.160)
```

   Removing that unusual correction period makes the difference between mid caps and the S&P Composite larger, and almost statistically significant. With such a small sample size, this is worth noting. Indeed, small caps do even worse during market corrections; over the last 45 years, anyway, when market corrections come, the littler guys get hit worse than the big guys do.
   (b) The prediction interval is

$$-21.81 \pm (2.16)(13.16)\sqrt{1.071} = -21.81 \pm 28.43 = (-50.24, 6.62).$$

   The normal plot of this variable doesn't look too bad here.



(c) A normal plot of the difference between small cap return and S&P return looks fine, so the confidence interval is as follows:

Variable	N	Mean	StDev	SE Mean	95.0 % CI
Post-cor	14	10.39	13.16	3.52	(2.80, 17.99)

Thus, small cap stocks significantly outperform the S&P Composite during the 12 months after a market correction, on average. Note, of course, that a prediction interval for any one market correction would be very wide, with the small cap return ranging from roughly 18 percentage points below the S&P Composite to more than 45 percentage points above it, so this isn't a foolproof strategy! The other trick is to be able to recognize when a market correction is over while you're still in it!

(d) A normal plot of the difference between mid cap and small cap return looks reasonable. Here's the confidence interval:

Variable	N	Mean	StDev	SE Mean	95.0 % CI
Post-cor	14	-2.80	6.59	1.76	(-6.61, 1.01)

Mid caps underperform small caps in the post–correction period, although based on these data, not significantly so.