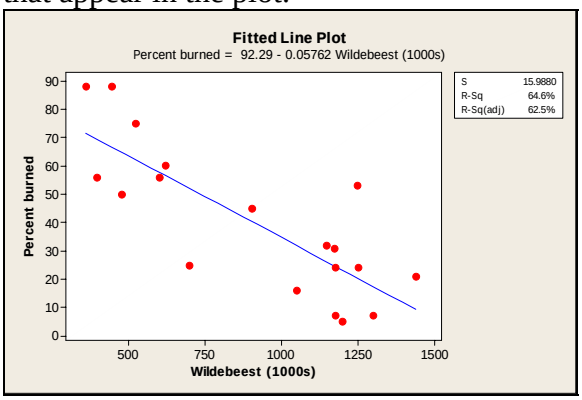
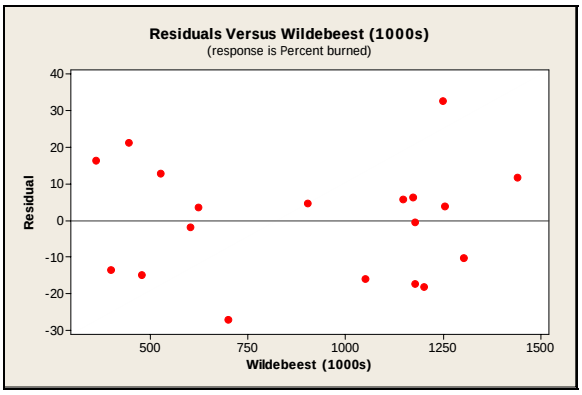
Part I Review Exercises

I.1 *Who?* The individuals are 19 years. *What?* The variables measured are wildebeest abundance (in thousands of animals) and the percent of grass area burned in the same year. *Why?* There is a claim that more wildebeest reduce the percent of grasslands burned. *When, where, how, and by whom?* We are not told when these data were collected. However, we know the data are from long-term records from the Serengeti National Park in Tanzania. *Graph***:** The scatterplot below (on the left) shows a moderately strong, negative, fairly linear relationship between the percent of grass area burned and wildebeest abundance. There are no unusual points that appear in the plot.





Numerical summaries: For these data, $\bar{x} = 904.8, s_x = 364.0, \bar{y} = 40.16, s_y = 26.10$, and $r = -0.803$. *Model*: The line on the plot is the least-squares regression line of percent of grass area burned on wildebeest abundance. The regression equation is $\hat{y} = 92.29 - 0.05762x$. A residual plot is shown above (on the right). *Interpretation*: The scatterplot shows a negative association. That is, areas with less grass burned tend to have a higher wildebeest abundance. The overall pattern is moderately linear ($r = -0.803$). The slope of the regression line suggests that for every increase of 1000 wildebeest, the percent of grassy area burned decreases by about 5.8. According to the *y*-intercept, an area with no wildebeest would have 92.29 percent of grass area burned. It does not make sense to interpret the *y*-intercept due to extrapolation. The residual plot shows a fairly "random" scatter of points around the "residual = 0" line. There is one large positive residual at 1249 thousand wildebeest. Since $r^2 = 0.646$, 64.6% of the variation in percent of grass area burned is explained by the least-squares regression of percent of grass area burned on wildebeest abundance. That leaves 35.4% of the variation in percent of grass area burned unexplained by the linear relationship.

I.2 (a) The marginal distribution of reasons for all students is

Save time	21.2%
Easy	21.2%
Low price	27.7%
Live far from stores	8.2%
No pressure to buy	7.1%
Other reason	14.7%

Note: The percentages total 100.1%, due to rounding error. (b) The conditional distributions of American and East Asian students are

	American	East Asian
Save time	25.2%	14.5%
Easy	24.3%	15.9%
Low price	14.8%	49.3%
Live far from stores	9.6%	5.8%
No pressure to buy	8.7%	4.3%
Other reason	17.4%	10.1%

Note: The percentages for East Asian students total 99.9%, due to rounding error. (c) A higher percentage of American students than East Asian students buy from catalogs because it saves them time (25.2% versus 14.5%) and it is easy (24.3% versus 15.9%). A higher percentage of East Asian students than American students buy from catalogs because of the low price (49.3% versus 14.8%).

I.3 (a) Since we know the weights of seeds of a variety of winged bean are approximately Normal, we can use the Normal model to find the percent of seeds that weigh more than 500 mg. First, we standardize 500 mg:

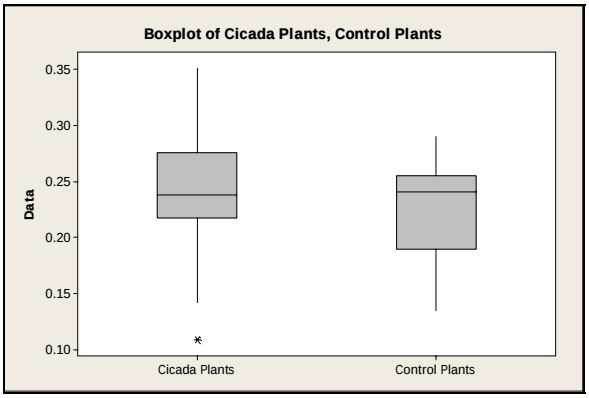
$$z = \frac{x - \mu}{\sigma} = \frac{500 - 525}{110} = \frac{-25}{110} = -0.23$$

Using Table A, we find the proportion of the standard Normal curve that lies to the left of $z = -0.23$ to be 0.4090, which means that $1 - 0.4090 = 0.5910$ lies to the right of $z = -0.23$. Thus, 59.1% of seeds weigh more than 500 mg. (b) We need to find the z-score with 10% (or 0.10) to its left. The value $z = -1.28$ has proportion 0.1003 to its left, which is the closest proportion to 0.10. Now, we need to find the value of x for the seed weights that gives us $z = -1.28$:

$$\begin{aligned} -1.28 &= \frac{x - 525}{110} \\ -1.28(110) &= x - 525 \\ 525 - 1.28(110) &= x \\ 384.2 &= x \end{aligned}$$

If we discard the lightest 10% of these seeds, the smallest weight among the remaining seeds is 384.2 mg.

I.4 *Who?* The individuals are American bellflower plants. *What?* The explanatory variable is whether cicadas were placed under the plant (categorical) and the response variable is seed mass in milligrams (quantitative). *Why?* The researcher wants to investigate whether cicadas serve as fertilizer and increase plant growth. *When, where, how, and by whom?* We are not told when these data were collected. However, we know the data come from 39 cicada plants and 33 control plants on the forest floor in the eastern United States. *Graphs***:** We can compare the cicada plants and the control plants with a side-by-side boxplot and a back-to-back stemplot. In the stemplot, the stems are listed in the middle and the leaves are placed on the left for cicada plants and on the right for control plants.



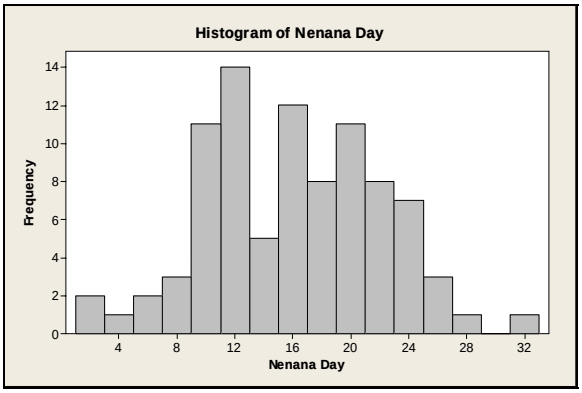

```
Stem-and-leaf of Cicada Plants and Control Plants
                Leaf Unit = 0.010

Cicada     |    | Control
     0     | 1  | 3
     4     | 1  | 445
     7     | 1  | 77
    99     | 1  | 89999
   111100  | 2  | 0111
3333332222 | 2  | 2
     5544  | 2  | 4444445555
  7777666  | 2  | 66666
     999   | 2  | 89
     110   | 3  |
           | 3  |
       5   | 3  |
```

Numerical summaries: Here are summary statistics for the two distributions:

Variable	Mean	s	Min	Q1	M	Q3	Max	IQR
Cicada Plants	0.24264	0.04759	0.1090	0.2170	0.2380	0.2760	0.3510	0.0590
Control Plants	0.22209	0.04307	0.1350	0.1900	0.2410	0.2550	0.2900	0.0650

Interpretation: The distribution of seed mass (in mg) is a bit right-skewed for the cicada plants. One cicada plant had an unusually low seed mass (0.109 mg). For the control plants, the distribution of seed mass (in mg) is somewhat left-skewed. While the median seed mass is about the same for both the cicada plants and the control plants, the seed mass for the cicada plants is higher than the seed mass for the control plants at the first and third quartiles (and at the maximum). The mean seed mass is higher for the cicada plants. The standard deviation is larger for the cicada plants, while the *IQR* is larger for the control plants. Because of the outlier in the seed mass for the cicada plants and the skewness of both distributions, we should use the resistant medians and *IQR*s in our numerical comparisons. The median and *IQR* are both smaller for the cicada plants than for the control plants. However, the first and third quartiles and the maximum are greater for the cicada plants than for the control plants. We might want to do more research to see if we come up with more conclusive data.

I.5 A histogram of the date of ice breakup (number of days since April 20) on the Tanana River shows the data well.



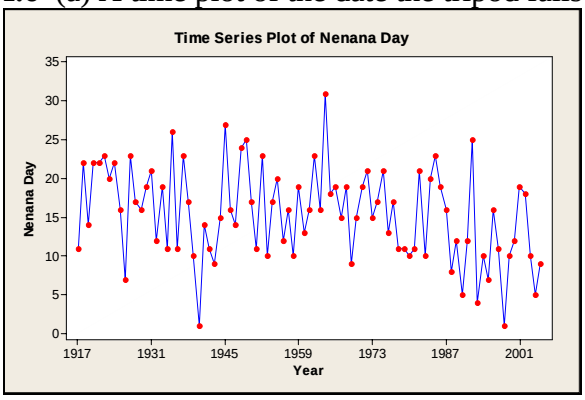
Because the distribution is slightly right-skewed, it is appropriate to use the five-number summary (and *IQR*) to describe the data. Alternatively, since the distribution is roughly

symmetric with no outliers, it is appropriate to use the mean and standard deviation to describe center and spread.

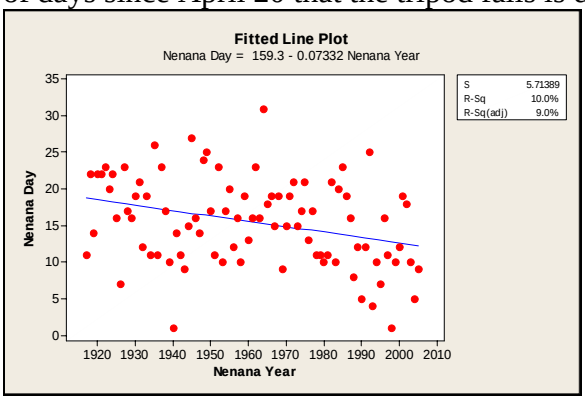
Variable	Mean	*s*	Minimum	Q1	Median	Q3	Maximum	IQR
Nenana Day	15.483	5.989	1.000	11.000	16.000	20.000	31.000	9.000

The median date for ice breakup occurs 16 days after April 20, which is May 6.

I.6 (a) A time plot of the date the tripod falls against the year is shown below.

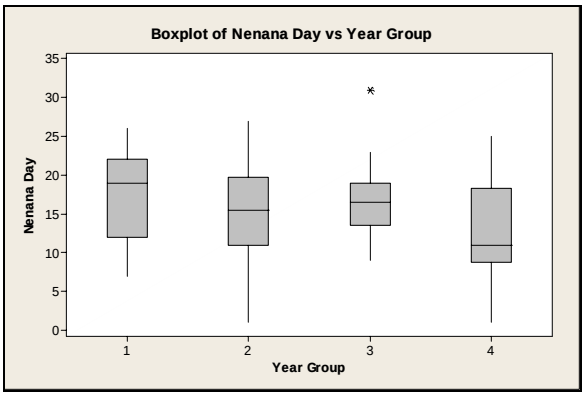


(b) A regression line added to a plot of the days against year shows, on average, that the number of days since April 20 that the tripod falls is decreasing as the years go by.



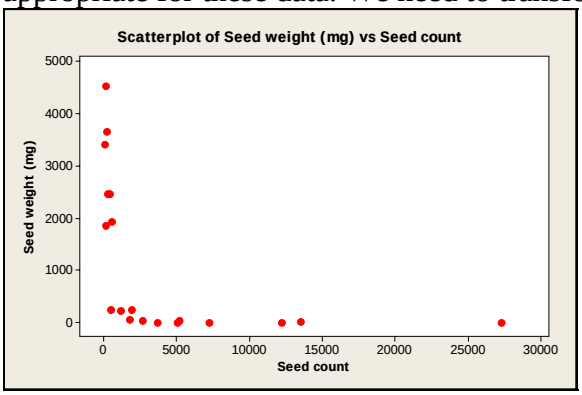
(c) According to R-Sq in the fitted line plot above, 10.0% of the variation in ice breakup time is accounted for by the time trend.

I.7 Grouping the data into year groups (1 = 1917 to 1939, 2 = 1940 to 1959, 3 = 1960 to 1979, 4 = 1980 to 2005), we can see that the median time to tripod drop is generally decreasing over time. The median is approximately equal for the time periods 1940 to 1959 and 1960 to 1979. However, the median looks noticeably higher for the time period 1917 to 1939 and noticeably lower for the time period 1980 to 2005.

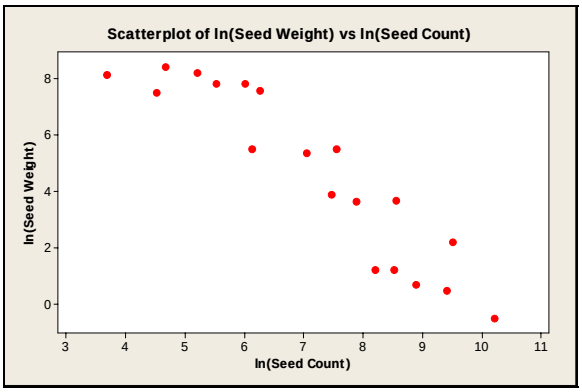


I.8 This is an observational study, so we cannot prove that online instruction is more effective than classroom teaching. There are other factors that we must consider. These arise when we ask the question "What might be different about students who choose online instruction over classroom instruction?" Some factors to consider are: age of the students (e.g., older students may work full time and find it easier to take an online course, but these students might be more serious about doing well in the course), aptitude of the students (e.g., those who are proficient with computers and choose online instruction might also be better students).

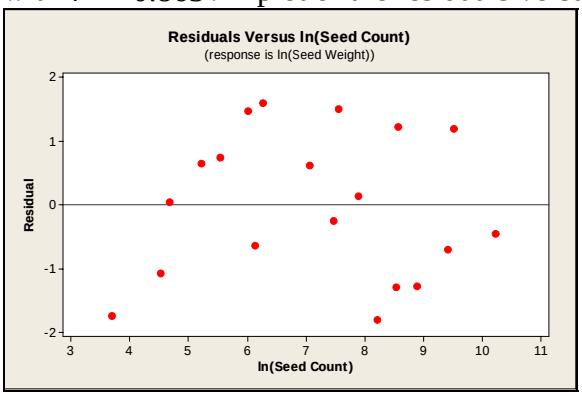
I.9 *Who?* The individuals are several common tree species. *What?* The variables are seed count and seed weight (mg). *Why?* We wonder if trees with heavy seeds tend to produce fewer seeds than trees with light seeds. *When, where, how, and by whom?* These data come from many studies compiled in Greene and Johnson's "Estimating the mean annual seed production of trees," which was published in *Ecology*, volume 75 (1994). *Graphs***:** We first examine a scatterplot of seed weight versus seed count. The plot shows that a linear relationship is not appropriate for these data. We need to transform the data.



Taking the natural log of both seed count and seed weight gives us a relationship that looks more linear.



Numerical summaries: The correlation between ln(Seed Weight) and ln(Seed Count) is -0.929. *Model*: The least-squares regression equation is ln(Seed Weight) = $15.5 - 1.52$ ln(Seed Count), with $r^2 = 0.863$. A plot of the residuals versus ln(Seed Count) is shown below.



There appears to be fairly random scatter in the residual plot, so the regression we have performed seems appropriate. We now perform an inverse transformation on the linear regression equation:

$$\ln(\text{Seed Weight}) = 15.5 - 1.52 \ln(\text{Seed Count})$$

$$e^{\ln(\text{Seed Weight})} = e^{15.5 - 1.52 \ln(\text{Seed Count})}$$

$$(\text{Seed Weight}) = e^{15.5} \times e^{-1.52 \ln(\text{Seed Count})}$$

$$(\text{Seed Weight}) = e^{15.5} \times (\text{Seed Count})^{-1.52}$$

This is the power model for the original data. *Interpretation*: The relationship between seed count and seed weight is not linear. However, we have found a power model that works well to describe this relationship. The relationship we found tells us that 86.3% of the variability in ln(Seed Weight) is accounted for by the least-squares regression on ln(Seed Count).

I.10 (a) Smaller cars tend to get better gas mileage than larger cars. More than 50% of large cars get less gas mileage than the midsize car with the worst gas mileage. All large cars get less gas mileage than 75% of the subcompact and compact cars. Subcompact cars get the best gas mileage, on average, but they also have the most variability. Compact cars get slightly worse gas mileage than subcompact cars, but there is still a lot of variability for the compact cars. Overall, as the size of the car increases, the gas mileage noticeably decreases. (b) For each additional penny in the cost of gas, the sale of high MPG cars increases by 0.101690%, on average. A more practical way to look at this relationship is to say that for each additional 10 cents spent on gas, the sale of high MPG cars increases about 1.02%, on average. The *y*-intercept says that if gas

cost nothing, the high MPG cars sales would be about 9.6% of the car sales market. This does not make any sense, since we need to extrapolate outside of the range of the data to make this statement. (c) The predicted sales of high MPG cars for that month is

$$\text{High mpg Car\%} = 9.63594 + 0.101690(150) = 24.89$$

That is, we predict high MPG cars to represent about 24.89% of sales that month. The actual sales of high MPG cars were about 25.8%. The residual is $25.8\% - 24.89\% = 0.91\%$. (d) 45% of the variation in the sale of high MPG cars (%) is accounted for by the least-squares relationship with gas price in the current month.