ELEMENTARY STATISTICS: A WORKBOOK

by

VIRGINIA LEE WEBER

A writing project submitted in partial fulfillment

of

MASTER OF SCIENCE

in

Statistics

Approved:

_____

Chairperson, Graduate Committee

_____

Committee Member

_____

Committee Member

Montana State University
Bozeman, Montana

May, 1998

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Being concurrently a student and an instructor of statistics for the past two and a half years at Montana State University, I have identified what has been for me, a stumbling block in both endeavors.

Sustaining interest and understanding the relevance of statistics were but two of my shortcomings as a student. Motivating interest and demonstrating the relevance of statistics are but two of my shortcomings as an instructor.

Valid examples are not easily acquired and the multifariousness of applications limited. But statistics is an integral part of many occupations yielding actual applications of statistics which should be made available to illustrate statistical concepts and techniques in the classroom. The absence of a compilation of real life applications with real data in every field requiring knowledge of statistics severely restricts the breadth of both instruction and learning in an introductory statistics classroom.

This paper represents a general approach to stimulate the interest in and demonstrate the relevance of statistics to students and teachers alike.

CHAPTER II

THE PROBLEM

The problem is twofold.

First, the absence of a real life applications resource featuring real data from various fields severely restricts the breadth of instruction in an introductory statistics classroom. Since statistics is used in many pursuits there are genuine applications of statistics that are procurable and may be used to illustrate statistical concepts and techniques in the classroom.

Recognizing the importance of such illustration, instructors often motivate new concepts with examples. They demonstrate mechanics and reinforce statistical notions using more examples. Where are these examples coming from? Are they statistically valid in their application? Are their interpretations cogent in the context of the field of application?

It is difficult to formulate (or make up) a usance of a statistical concept. Hence models intended for the classroom use fabricated data and are largely applicable to only a couple of fields with which the instructor is abreast. Oftentimes, every statistical topic is motivated via representatives from the same general area, say sports, biology, or industry.

But generally, student enrollment in an introductory statistics class consist of twenty or more different majors. Students who are sports fans, biologist, or interested in industry are motivated to learn statistics because it increases their knowledge about something they already know and like. The instructor holds their interest and the students make valuable contributions to the classroom through their attendance and participation.

However, many more students in the same classroom are not interested in sports, biology, or industry. Their attendance is poor and their participation deficient. They struggle with the statistical concepts and lack the motivation to sustain them through that struggle. Yet, instructors avoid examples from the fields of interest to these students. The instructor's unfamiliarity of a field, like finance or economics, to which he or she is attempting to apply a statistical concept may be noticed and challenged by students of finance or economics. This creates a potential threat to the instructor in the form of the loss of credibility.

The instructor's goal is to teach the fundamentals of statistics to all his or her students. To hold the interest of as many students as possible, instructors need access to actual statistical applications from other fields which are, by nature, based on real data.

The second aspect of the problem is that the lack of real life examples seriously restricts the scope of learning by introductory statistics students. In order to ingrain statistical concepts, students require practice as well as relevant and motivative illustrations in the classroom.

Presently, students practice through homework assignments. Those assignments may or may not relate to their field of study. They are unable to carry the statistical concept over to their own area of interest. The homework does not inspire learning and they loose interest.

So where is inspiration? The student's goal is to learn statistics in the context of the field requiring it. They will need to apply statistics to this field in the marketplace. They are interested in this field and want to know more about it. The study of actual applications of statistics by those occupations which utilize statistics would reinforce statistical concepts and techniques to the student.

Further, by nature we tend to practice more at something we enjoy. Students would work more problem sets and learn more statistics if given actual scenarios with actual data, analysis, and interpretation from the experts in their own field.

CHAPTER III

THE SOLUTION

The compilation of real life applications with real data, analysis, and interpretations in every field requiring knowledge of statistics into a statistics workbook (hereafter referred to as the workbook) would not only benefit teacher and student, but also education and the marketplace.

By reviewing applications from other fields, instructors are more diverse in their presentation of statistical concepts. A wider range of students is reached. Student involvement increases. Learning is mutual as discussions about the mechanics of statistics become more in-depth discussions about the applications and interpretations of statistics. Class preparation time is reduced for the instructor as well making more time available for research or other responsibilities.

Given a resource containing actual scenarios from various fields with actual data, analysis, and interpretations from an expert, students would find problem sets to work that demonstrated the statistical concepts in their field of interest. This would prepare the student for both classroom discussions about statistics and the application of the subject to his or her future occupation.

Education benefits in that the statistics classroom would be more productive with a higher success rate, the needs of the students, their major departments, and the marketplace having been met. And lastly, the marketplace profits as more highly skilled workers join the work force.

CHAPTER IV

IMPLEMENTATION

Writing a statistics workbook of this nature is a multifaceted task. Statistics is an integral part of many occupations and there are true applications of statistics that would be useful to illustrate statistical concepts and techniques in the classroom.

Compiling a book of actual statistical applications requires contribution by both the statistician and the field expert. Together, the statistician and the field expert can transcribed actual scenarios and present them in such a fashion as to lead students into deeper thought about both statistics and the field of application.

The statistician must insure that an application from a certain field is appropriate in teaching a given concept or technique. He or she must ensure that the application is statistically valid. The contributor proficient in the field ensures appropriate application and interpretation in context of the field of interest. For example, the statistician ensures that the center of a distribution of customer savings accounts is found properly while the banker must ensure that the interpretation of that center is appropriate.

Such a workbook would need to be organized by field of interest as well as by topic of concern. It may be more appropriate to have separate workbooks for business, current events, natural sciences, and the soft sciences. Each statistical topic could be presented through problem sets based on true scenarios.

A statistics workbook of this nature provides a powerful tool. Teachers instruct from it, students are motivated through it, and the workplace has a forum to express to educators and future employees alike what the workplace requisitions.

CHAPTER V

PROBLEM SETS

The following dataset contains the number (in millions) of hours worked in iron and steel mills between 1980 and 1992.

758 474 419 356 354 363 360 350 304 293

a) Describe the distribution of hours worked in iron and steel mills using the 5-number summary.

b) What does the 5 # summary reveal about the variability of man-hours worked in iron and steel mills between the years of 1980 and 1992?

c) Describe the distribution of hours worked in iron and steel mills using a modified boxplot. Label both axis's and identify the 5 # summary, the IQR, all outliers, and the direction of skewness.

The following dataset contains the number (in millions) of hours worked in iron and steel mills between 1980 and 1992.

758 474 419 356 354 363 360 350 304 293

d) Describe the spread in the distribution of hours worked in iron and steel mills for the years 1980 - 1992 using a histogram with an interval width of 50 (millions) hours starting at 27.5.

e) Describe the spread of the distribution of hours worked in iron and steel mills using standard deviation.

f) Compare the differences in the measurements of the spread of the distribution of hours worked in iron and steel mills using standard deviation. In this case, should we use IQR or the standard deviation? Explain.

The annual cost of tuition and fees in Nebraska[1] four year colleges in 1997 are given below.

$3,700	$1,900	$8,300	$11,800	$12,200	$11,000
$11,000	$7,500	$11,000	$12,300	$4,300	$8,300
$10,700	$1,900	$9,600	$2,100	$2,300	$2,700
$2,600	$1,900	$6,700			

i. What is the **variable of interest**?

ii. What is the **unit of measure**?

A. ***Frequency tables, relative frequency tables, and histograms***

1. Construct a **frequency** and **relative frequency table** of the annual cost of tuition and fees in Nebraska four year colleges in 1997. Construct these numerical summaries with a class width of $2500.

2. What is the difference between the two **numerical summaries**? Which table represents a count and which represents a proportion?

3. Construct a **frequency histogram** that corresponds to your frequency table.

---

1 Peterson's (1997). Guide to Four-Year Colleges 1998, "Nebraska" pgs 713-724.

4. Approximate the **median** of the **distribution**.

5. Is the distribution **skewed left**, **skewed right**, or **approximately symmetric**?

6. Does the **mean** lie to the left or the right of the median or is it approximately equal to the median?

7. Which method of reporting the **center** and **spread** is most appropriate for this dataset? Explain.

8. Construct a **relative frequency histogram** that corresponds to your relative frequency table.

9. The 50th **percentile** of tuition cost in Nebraska four-year colleges is approximately $_____

10. Describe the **shape** of the distribution.

11. The average tuition in Nebraska four year colleges is _____ the 50th percentile of tuition cost in that state.

12. Which method of reporting the **center** and **spread** is most appropriate for this dataset? Explain.

13. What is the difference between the two histograms?

14. Why didn't the center, spread and shape of the distribution change with the different histograms?

B. Five Number Summary and Boxplots

1. Find the **five number summary** of the distribution Report the summary in terms of 1997 tuition cost for Nebraska four-year colleges.

2. Does your report in B1 confirm you **estimates** of the median in questions A4 and A9? Explain.

3. Describe the center, spread and shape of the distribution in terms of 1997 tuition cost for Nebraska four-year colleges.

4. Does your report in B1 confirm the shape of the distribution as chosen in questions A5 and A10? Explain.

5. Construct a **boxplot** of the tuition cost.

6. What numerical summary is used to construct a boxplot?

7. What graphical technique reveals **statistics** which are **resistant** measures of center and spread?

8. What is the difference between a boxplot and a modified boxplot?

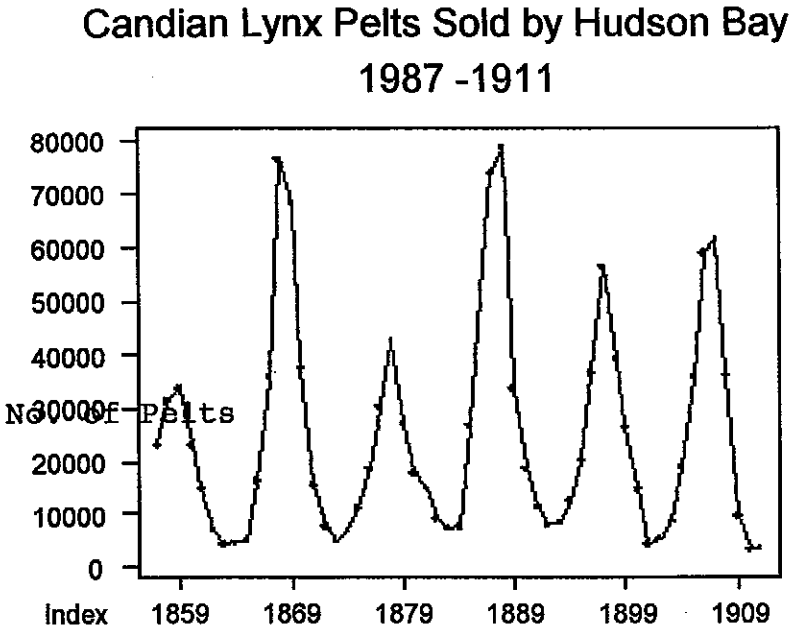
9. The whiskers in a modified box plot extend to **1.5IQR** from **M** when

   A. both the minimum and the maximum data points are less than 1.5IQR from the median.
   B. none of the data points exceed 1.5IQR.
   C. at least one data point is beyond 1.5IQR from the median.
   D. a data point exist that is 1.5IQR from the median.
   E. Both C and D are correct.

10. When should stemplots, five number summaries and boxplots be used?

11. Frequency tables and histograms should be used when there are _____ present.

The plot below represents the numbers of Canadian Lynx[1] pelts sold by the Hudson's Bay Company for the years 1857 - 1911.



1. The plot above is called a

   A. bar graph.
   B. boxplot.
   C. time plot.
   D. frequency histogram.
   E. frequency table.

2. Identify three cyclic patterns in the plot above.

---

[1] Andrew, D. F. and Herzberg, A. M. (1985). Data: A Collection of Problems from Many Fields for the Student and the Research Worker, "Canadian Lynx Trappings" pgs 13-15

3. What would you expect to happen in the next decade? Justify your answer.

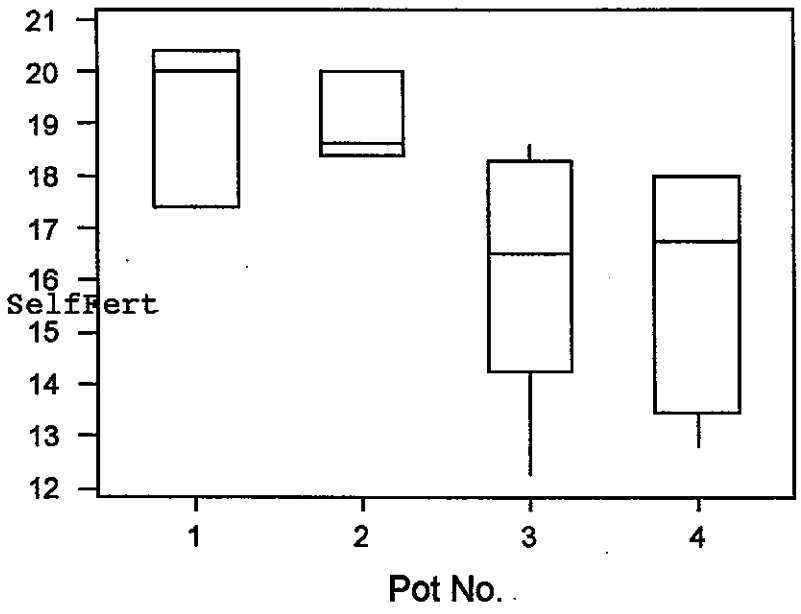
4. Describe the overall trend in the data.

5. T/F: It is possible to identify seasonal variations from this plot.

6. Identify any irregular fluctuations. Explain.

7. Do you think that the endangered species act and other such legislation had an impact the cycles and trend of Canadian Lynx sales after 1911? Describe the potential impact.

Darwin[1] was interested in comparing the effects of cross and self fertilization of plants. One of his studies involved the self fertilization of a plant called *Zea Mays*. The graph below summarizes the distributions of heights (in inches) for the self-fertilized *Zea Mays* that Darwin grew in 4 different pots.



1. T/F: The distribution of heights of the self-fertilized *Zea Mays* grown in pots 1 and 2 have no variability because their plots do not have whiskers.

2. T/F: The tallest plant grown in pot 4 is the same height as the 75th percentile of the distribution of heights in pot 4.

3. T/F: The variable of interest is the pot number.

4. The graph above is a

   A. boxplot.
   B. side-by-side boxplot.
   C. frequency histogram.
   D. bar graph.
   E. scatterplot.

---

[1] Andrew, D. F. and Herzberg, A. M. (1985). Data: A Collection of Problems from Many Fields for the Student and the Research Worker, "Darwin's Data on Growth Rate of Plants" pgs 9 - 12.

5. Rank the pots based on their median height (from tallest to shortest). Justify your answer.

   A. 1-2-3-4
   B. 2-1-3-4
   C. 2-1-4-3
   D. 4-3-2-1
   E. 1-2-4-3

6. Which pot produced plants that had the least difference in heights and which pot produced plants with the greatest difference in heights? Justify your answer

   A. 1 least, 2 greatest
   B. 2 least, 3 greatest
   C. 3 least, 4 greatest
   D. 2 least, 4 greatest
   E. can not be determined.

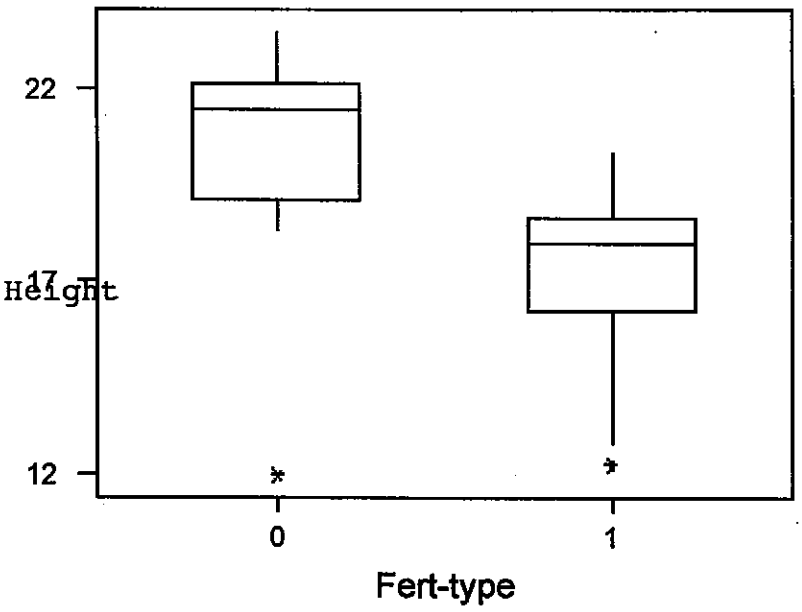
7. Which pot contained the tallest *Zea May*? Explain.

   A. 1
   B. 2
   C. 3
   D. 4
   E. Can not be determined.

8. Which of the following statements is **true**:

   A. pots 1 and 4 are skewed left, pot 2 is skewed right and pot 3 is symmetric.
   B. pots 3 and 4 are skewed left and pots 1 and 2 are skewed right.
   C. pots 1, 3, and 4 are skewed left and pot 2 is skewed right.
   D. pot 2 is skewed left and the rest are skewed right.
   E. pot 2 is skewed left, pot 3 is symmetric and pots 1 and 4 are skewed right.

The plot below represents the distributions of heights in the cross fertilized *Zea Mays* (Fert-type 0) and the self-fertilized *Zea Mays* (Fert-Type 1). Answer the following questions:



9. T/F: All of the self-fertilized plants are shorter than the median height of the cross-pollinated plants.

10. T/F: The cross-pollinated plants are more variable in height then the self pollinated plants.

11. T/F: Both distributions have one outlier.

12. T/F: The star and the lower whisker on the cross-pollinated data (Fert-type 0) indicates that none of the cross-pollinated plants had a height between 12 inches and approximately 18 inches.

13. T/F: Both distributions are skewed right.

14. T/F: The average height of a cross-pollinated *Zea May* is less than the second quartile height of the same distribution.

The proportions of female students in 30 Alabama four year colleges as reported by Peterson's Guide to Four-Year Colleges (1997)[1] are summarized below.

Min =	10%
Q1 =	53%
Med =	57%
Q3 =	61%
Max =	85%

1. Based on this information, we know that

   A. less than a quarter of the colleges surveyed had more men enrolled than women.
   B. 3/4 of the colleges surveyed have student bodies consisting of more than 50% women.
   C. the average proportion of women in these Alabama colleges in 1997 is less than 57%.
   D. A, B, and C are all true.
   E. none of the above statements are true.

2. T/F: Half of the colleges surveyed had a female student enrollment in excess of 53% but less than 61%.

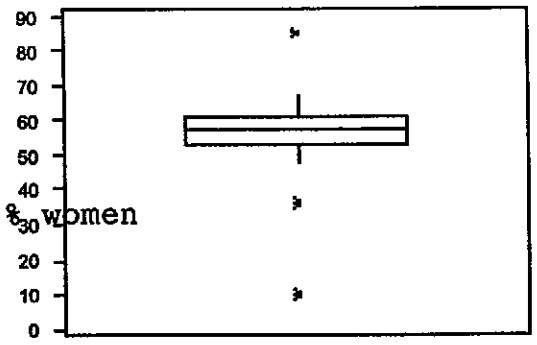
3. T/F: At least one of the 30 Alabama colleges has a male student enrollment of 90%.

4. T/F: The women students out number the men approximately 6 to 1 in at least one of the 30 colleges.

---

[1] Peterson's (1997). Guide to Four-Year Colleges 1998, "Alabama" pgs 244-259.

Below is a graphical representation of the five number summary given on the previous page. Use this plot to answer the following questions.



5. The variable of interest is:

   A. the percentage of women enrolled at Montana State University.
   B. the percentage of women enrolled in college.
   C. the percentage of women enrolled at an Alabama college.
   D. the percentage of women enrolled at a four-year Alabama college.
   E. the percentage of women enrolled at a four-year Alabama college in 1997.

6. T/F: 25% of the schools surveyed have student bodies composed of more than 61% women.

7. T/F: At least two of the thirty schools have more male students than female.

8. T/F: On average, more men attended the 30 Alabama four-year colleges in 1997 than men.

CHAPTER VI

CONCLUSION

A statistics workbook of this nature must be a collaborative endeavor. Teachers, students, and the marketplace must all contribute. First a workbook that meets the needs of teachers and students today. And then the subdivision of traditional introductory statistics service courses into field oriented subsections to meet the needs of teachers and students tomorrow.

The problem sets in this paper do not represent the type of collaborative effort that I feel is needed. They merely portray a general model to accomplish that which a teacher wishes to realize: students successfully incorporating statistical concepts and techniques into their lives and futures.

I consider a future implementation of this work a near reality.

BIBLIOGRAPHY

1. Andrew, D. F. and Herzberg, A. M. Data: A Collection of Problems from Many Fields for the Student and the Research Worker, "Canadian Lynx Trappings" pgs 13 -15

2. Andrew, D. F. and Herzberg, A. M. Data: A Collection of Problems from Many Fields for the Student and the Research Worker, "Darwin's Data on Growth Rate on Plants" pgs 9 -12

3. Peterson's (1997), Guide to Four-Year Colleges 1998, "Alabama" pgs 244 - 259

4. Peterson's (1997), Guide to Four-Year Colleges 1998, "Nebraska" pgs 713 - 724

APPENDIX I

PROBLEM SET SOLUTIONS

Business - Manufacturing

Spread - Labor

The following dataset contains the number (in millions) of hours worked in iron and steel mills between 1980 and 1992.

758 474 419 356 354 363 360 350 304 293

a) Describe the distribution of hours worked in iron and steel mills using the 5-number summary.

Min = 293 million man hours
Q1 = 350 million man hours
M = 358 million man hours
Q3 = 419 million man hours
Max = 758 million man hours

Five years had fewer than 358 million man hours of production. One year saw a total number of man hours worked of 758 million.

b) What does the 5 # summary reveal about the variability of man-hours worked in iron and steel mills between the years of 1980 and 1992?

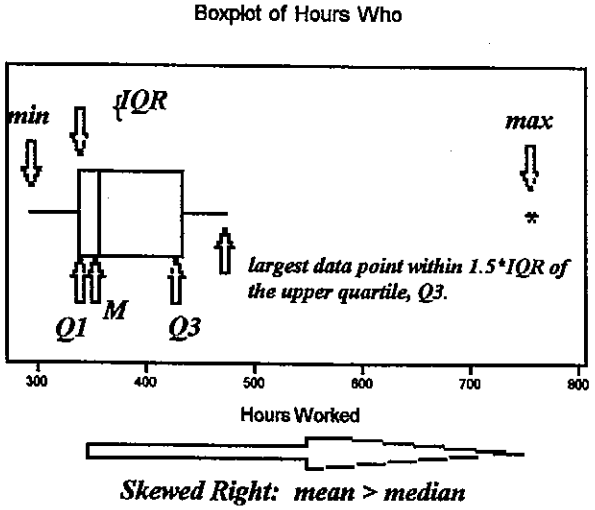
The overall range of man-hours worked during this period is Max - Min = 758 - 293 = 465 million man hours. The interquartile range is Q3 - Q1 = 419 - 350 = 69 million man hours. This suggest that outliers may exist which would effect the overall range but not the interquartile range.

We may also note that M - min is only 65 million man hours while Max - M is 400 million. This indicates that the distribution of hours is skewed to the right. this will pull the mean to the right of the median. Thus the average man hours worked exceeds the median man hours worked in the steel mills.

Further analysis reveals that the max data point lies beyond 1.5*IQR of the upper quartile. This indicates that the maximum data point is an outlier. That is to say that during one of the years studied, there was an unusually high number of man hours worked.

Either this is actually the case or this large number was recorded in error. To determine this, we might want to study production levels for the same time period. We would look for a single year in which production was unusually high.

c) Describe the distribution of hours worked in iron and steel mills using a modified boxplot. Label both axis's and identify the 5 # summary, the IQR, all outliers, and the direction of skewness.

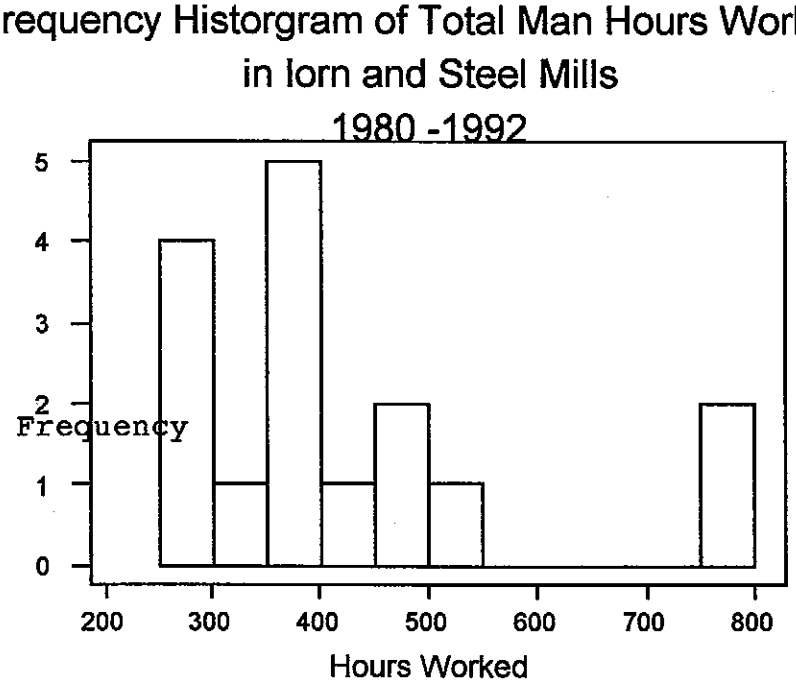


Skewed Right: mean > median

The following dataset contains the number (in millions) of hours worked in iron and steel mills between 1980 and 1992.

758 474 419 356 354 363 360 350 304 293

d) Describe the spread in the distribution of hours worked in iron and steel mills for the years 1980 - 1992 using a histogram with an interval width of 50 (millions) hours starting at 27.5.



The histogram shows that the overall range is approximately 800 - 250 = 550 million man hours. An obvious outlier is also revealed. We recognize that this outlier will cause the standard deviation to be inflated or greater than it should be with respect to the rest of the data.

e) Describe the spread of the distribution of hours worked in iron and steel mills using standard deviation.

$$s^2 = (1/(n-1))(\Sigma x^2 - n^* x^2)$$
$$= (1/9)(1788987 - 10*162489.61)$$
$$= (1/9)(164090.9)$$
$$= 18232.32 \text{ (million man hours)}^2$$

$$s = (18232.32)^{1/2} = 135.03 \text{ million man hours.}$$

f) Compare the differences in the measurements of the spread of the distribution of hours worked in iron and steel mills using standard deviation. In this case, should we use IQR or the standard deviation? Explain.

The IQR is a resistant measure of spread and indicates that the middle 50% of the years surveyed had a spread of 69 million man hours. The standard deviation is a non resistant measure of spread which reports the average distance of a randomly selected data point from the mean of the distribution. The reported standard deviation for this case is almost double the IQR because of the significant effect of the outlier on the standard deviation. It is more appropriate to report the spread of this distribution in terms of IQR because of the outlier present.

The annual cost of tuition and fees in Nebraska[1] four year colleges in 1997 are given below.

$3,700	$1,900	$8,300	$11,800	$12,200	$11,000
$11,000	$7,500	$11,000	$12,300	$4,300	$8,300
$10,700	$1,900	$9,600	$2,100	$2,300	$2,700
$2,600	$1,9000	$6,700			

i. What is the **variable of interest**?

The annual cost of tuition and fees in Nebraska our year colleges in 1997.

ii. What is the **unit of measure**?

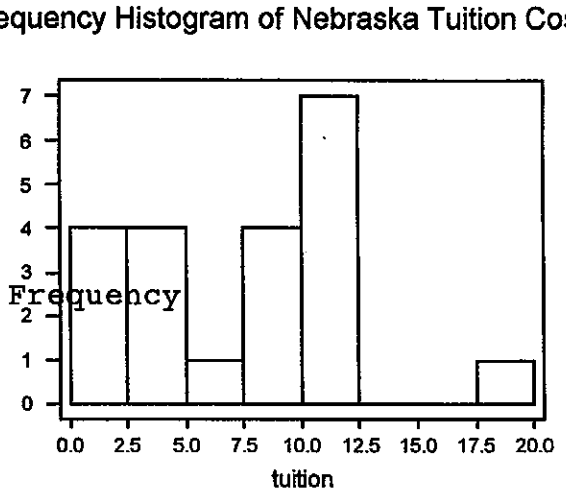
Dollars ($)

A. ***Frequency tables, relative frequency tables, and histograms***

1. Construct a **frequency** and **relative frequency table** of the annual cost of tuition and fees in Nebraska our year colleges in 1997. Construct these numerical summaries with a class width of $2500. What is the difference between the two **numerical summaries**? Which table represents a count and which represents a proportion?

Class	Freq.	Rel. Freq.
$0 < tuition < $2,500	4	.19
$2,500 < tuition < $5,000	4	.19
$5,000 < tuition < $7,500	1	.05
$7,500 < tuition < $10,000	4	.19
$10,000 < tuition < $12,500	7	.33
$12,500 < tuition < $15,000	0	0
$15,000 < tuition < $17,500	0	0
$17,5000 < tuition < $20,000	1	.05
	count	*proportion*

3. Construct a **frequency histogram** that corresponds to your frequency table.



[1] Peterson's (1997). Guide to Four-Year Colleges 1998, "Nebraska" pgs 713-724.

4. Approximate the **median** of the **distribution**.

Nine of twenty one schools have tuition lower than $7,500 so M > $7,500.

5. Is the distribution **skewed left**, **skewed right**, or **approximately symmetric**?

The distribution is skewed left since M > $7,500. Min - M > $7,500 - 0 = $7,500 while M - Max < 12.5 - 7.5 = 5.0. This implies that some schools had unusually low tuition..

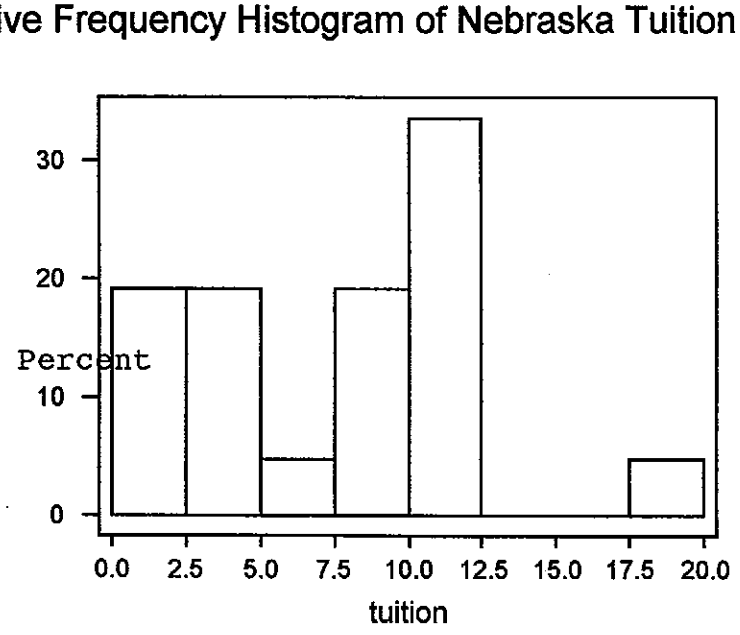
6. Does the **mean** lie to the left or the right of the median or is it approximately equal to the median?

The mean is pulled toward the lower tuition which is left of the median. Thus the distribution is skewed left.

7. Which method of reporting the **center** and **spread** is most appropriate for this dataset? Explain.

Since the distribution of tuition is skewed we want to use resistant techniques like the five number summary and boxplot. Skew due to unusually low tuition cost at some schools.

8. Construct a **relative frequency histogram** that corresponds to your relative frequency table.



9. The 50th **percentile** of tuition cost in Nebraska four-year colleges is approximately ***$7,500***

10. Describe the **shape** of the distribution.

The data have a bimodal distribution that is skewed left since the distance from min to approximate center (M) is greater than the distance from the median to the max (M - max).

11. The average tuition in Nebraska four year colleges is ***less than*** the 50th percentile of tuition cost in that state.

12. Which method of reporting the **center** and **spread** is most appropriate for this dataset? Explain.

Since the distribution of tuition is skewed we want to use resistant techniques like the five number summary and boxplot. Skew due to unusually low tuition cost at some schools.

13. What is the difference between the two histograms?

The frequency histogram represents the number of colleges which fall into each class of tuition cost. The relative frequency histogram reports the proportion of colleges which fall into each class of tuition

14. Why didn't the center, spread and shape of the distribution change with the different histograms?

The frequency histogram represents the number of colleges which fall into each class of tuition cost relative to the sample size n. The relative frequency histogram reports the proportion of colleges which fall into each class of tuition relative to the whole (relative to one).

B. Five Number Summary and Boxplots

1. Find the **five number summary** of the distribution Report the summary in terms of 1997 tuition cost for Nebraska four-year colleges.

Min	*Q1*	*M*	*Q3*	*Max*
1900	*$2,650*	*$8,300*	$11,000	*$19,000*

The lowest tuition in Nebraska four-year colleges in 1997 was $1,900 while the most expensive school had an annual tuition of $19,000. The middle range of tuition cost was between $2,650 and $11,000. Half of the schools had tuition under $8,300.

2. Does your report in B1 confirm you **estimates** of the median in questions A4 and A9? Explain.

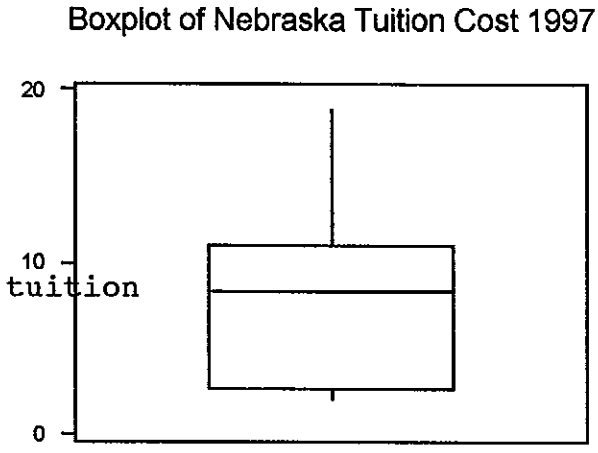
Yes, the median is $8,300 which is larger than $7,500 which was our lower bound on M.

3. Describe the center, spread and shape of the distribution in terms of 1997 tuition cost for Nebraska four-year colleges.

The distribution of tuition is skewed left with the average tuition cost being less than the median tuition cost of $8,300. The interquartile range of tuition is $8,350.

4. Does your report in B1 confirm the shape of the distribution as chosen in questions A5 and A10? Explain.

5. Construct a **boxplot** of the tuition cost.



6. What numerical summary is used to construct a boxplot?

The five number summary is a numerical technique of summarizing a distribution by its quartiles. The boxplot is a graphical representation of this numerical summary.

7. What graphical technique reveals **statistics** which are **resistant** measures of center and spread?

The five number summary is a numerical summary that provides statistics which remain relatively constant irrespective of outliers. And since the boxplot is a graphical representation of the five number summary it also provides statistics which are resistant.

8. What is the difference between a boxplot and a modified boxplot?

The modified boxplot shows outliers and its whiskers extend only to the last data point within 1.5IQR of the median. The boxplot draws whiskers to the min and the max from their respective quartiles.

9. The whiskers in a modified box plot extend to **1.5IQR** from **M** when

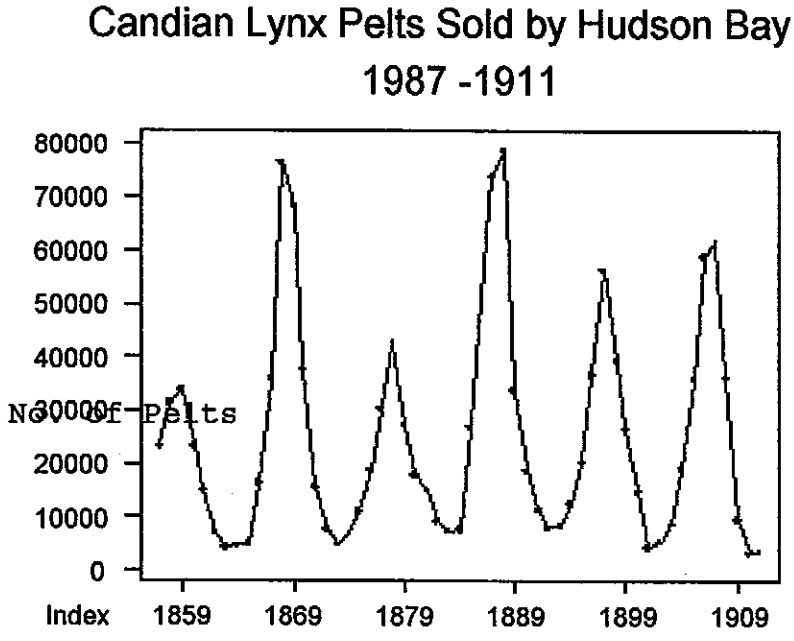
A. both the minimum and the maximum data points are less than 1.5IQR from the median.
B. none of the data points exceed 1.5IQR.
C. at least one data point is beyond 1.5IQR from the median.
D. a data point exist that is 1.5IQR from the median.
E. Both C and D are correct.

10. When should stemplots, five number summaries and boxplots be used?

The stemplot should be used for small data sets. The five number summary and the boxplots should be used when outliers exist.

11. Frequency tables and histograms should be used when the data set is ***large*** .

The plot below represents the numbers of Canadian Lynx[1] pelts sold by the Hudson's Bay Company for the years 1857 - 1911.



1. The plot above is called a

   A. bar graph.
   B. boxplot.
   C. time plot.
   D. frequency histogram.
   E. frequency table.

2. Identify three cyclic patterns in the plot above.

   1. ***Ten year cycle with highs near the end of each decade and lows in mid decade. Sates of pelts (and number of pelts available) decline for five years from the last decades high. Midway through the decade a minimum is reached and number of pelts sold climbs steadily for the next five years until they peak out at decades end. Then the cycle repeats.***

   2. ***Twenty year cycle of two consecutive decades with the second 10 year cyclic pattern being very exaggerated it peak. Thus, every twenty years the number of sales skyrocket after a fast five year trek up. It then fell back to its regular 10 year low. Then the next decade resumes its decade cycle.***

   3 ***Twenty year cycle where the second peak is higher than the first for each two consecutive decades.***

---

[1] Andrew, D. F. and Herzberg, A. M. (1985). Data: A Collection of Problems from Many Fields for the Student and the Research Worker, "Canadian Lynx Trappings" pgs 13-15

3. What would you expect to happen in the next decade? Justify your answer.

We would expect that sales of pelts would increase for the next few years. I would expect the next decade to have a lower max then the decade previous.

4. Describe the overall trend in the data.

The overall trend is generally increasing because each peak is taller than the one previous (ignoring the twenty year peaks).

5. T/F: It is possible to identify seasonal variations from this plot.

No. Seasonal variations are patterns within each year. The data must have within year information. This plot gives annual totals and therefore we look for cycles, trends, and irregular fluctuations but not seasonal variations.

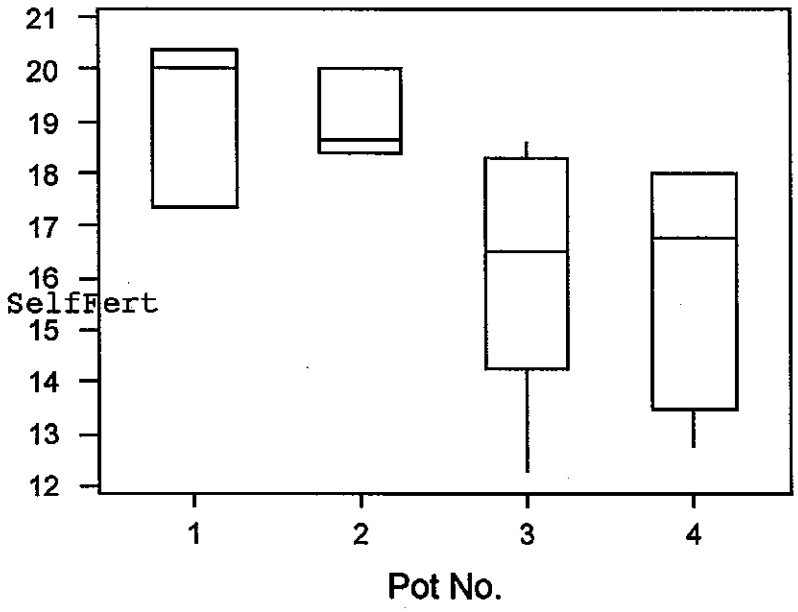
6. Identify any irregular fluctuations. Explain.

The last peak may be considered an irregular fluctuation in the twenty year cycle of exaggerated maximum in sales in the twentieth year of the cycle.

7. Do you think that the endangered species act and other such legislation had an impact the cycles and trend of Canadian Lynx sales after 1911? Describe the potential impact.

Such legislation must impact the distribution of sales. As trapping and hunting becomes more restrictive, the umber of pelts available decreases. The effect of such legislation is to reduce the and twenty year peaks. It may also drive the entire distribution with lower minimum sales.

Darwin[1] was interested in comparing the effects of cross and self fertilization of plants. One of his studies involved the self fertilization of a plant called *Zea Mays*. The graph above summarizes the distributions of heights (in inches) for the self-fertilized *Zea Mays* that Darwin grew in 4 different pots.



1. T/F: The distribution of heights of the self-fertilized *Zea Mays* grown in pots 1 and 2 have no variability because their plots do not have whiskers.

   False: Variability in a boxplot is measured by the box not the whiskers.

2. T/F: The tallest plant grown in pot 4 is the same height as the 75th percentile of the distribution of heights in pot 4.

   True: The boxplot implies that the tallest twenty five percent of the plants in pot 4 are all the same height.

3. T/F: The variable of interest is the pot number.

   False: The variable of interest is the height of the Zea May plants grown in the pots.

4. The graph above is a

   A. boxplot.
   B. side-by-side boxplot.
   C. frequency histogram.
   D. bar graph.
   E. scatterplot.

---

[1] Andrew, D. F. and Herzberg, A. M. (1985). Data: A Collection of Problems from Many Fields for the Student and the Research Worker, "Darwin's Data on Growth Rate of Plants" pgs 9 - 12.

5. Rank the pots based on their median height (from tallest to shortest).

   A. 1-2-3-4
   B. 2-1-3-4
   C. 2-1-4-3
   D. 4-3-2-1
   E. 1-2-4-3

   The median height of the plants in each pot is represented by the line which divides the box. Therefore, to rank the pots by median height simply compare the position of the line dividing each box. Since box 1 has the highest median line, plants in pot one have a larger median height, etc.

6. Which pot produced plants that had the least difference in heights and which pot produced plants with the greatest difference in heights?

   A. 1 least, 2 greatest
   B. 2 least, 3 greatest
   C. 3 least, 4 greatest
   D. 2 least, 4 greatest
   E. can not be determined.

   This question is interested in the overall range of heights in each pot. Therefore we must compare each distribution from the ends of the whiskers on the boxplot. Pot 2 hasn't any whiskers and also has the smallest box. Thus the plants in pot 2 are the most uniform in height. Pot 3 contains the plants with the most variability because the entire boxplot (including the whiskers) is larger than the other three boxplots.

7. Which pot contained the tallest *Zea May*?

   A. 1
   B. 2
   C. 3
   D. 4
   E. Can not be determined.

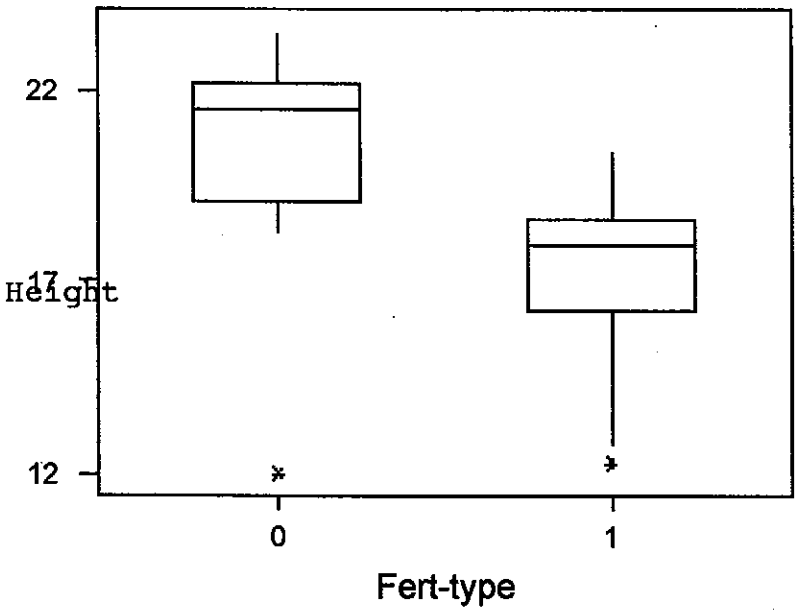
   Here we are comparing the highest point of each plot. 25% of the plants in pot one are taller than all the plants in each of the other three pots. Therefore pot 1 contains the tallest Zea May.

8. Which of the following statements is **true**:

   A. pots 1 and 4 are skewed left, pot 2 is skewed right and pot 3 is symmetric.
   B. pots 3 and 4 are skewed left and pots 1 and 2 are skewed right.
   C. pots 1, 3, and 4 are skewed left and pot 2 is skewed right.
   D. pot 2 is skewed left and the rest are skewed right.
   E. pot 2 is skewed left, pot 3 is symmetric and pots 1 and 4 are skewed right.

   All the pots are skewed left except pot 2 because the distance between the shortest plant and the median height in these plots all exceed the difference between the median and tallest plant in each pot. This is reversed for pot 2 which is skewed right.

The plot above represents the distributions of heights in the cross fertilized *Zea Mays* (Fert-type 0) and the self-fertilized *Zea Mays* (Fert-Type 1). Answer the following questions:



9. T/F: All of the self-fertilized plants are shorter than the median height of the cross-pollinated plants.

   True: Comparing the highest point of the top whisker for fert-type 1 to the bar dividing the box of fert-type 0 shows that the tallest self-fertilized plant is shorter than half of the crossed fertilized plants.

10. T/F: The cross-pollinated plants are more variable in height then the self pollinated plants.

    True: The middle 50% of the cross fertilized plants are between 19 and 22 inches tall, a difference of 4 inches. The middle 50% of the self fertilized plants are between 16 and 19 inches tall, a difference of three inches. Therefore the cross pollinated Zea Mays are more variable in their height.

11. T/F: Both distributions have one outlier.

    True: The distribution of both crossed and self fertilized plants contained an unusually small plant, shorter than the median height - 1.5 IQR.

12. T/F: The star and the lower whisker on the cross-pollinated data (Fert-type 0) indicates that none of the cross-pollinated plants had a height between 12 inches and approximately 18 inches.

    True: The star represents a single unusually small plant is 12 inches tall. The lowest point of the first quartile whisker indicates that the second shortest plant is 18 inches tall. Thus, no crossed fertilized plants were between 12 and 18 inches tall.

13. T/F: Both distributions are skewed right.

False: ***The shorter half of the plants for both fertilization types are more variable in height then the taller 50% of the plants This indicates that the average height of both distributions will be pulled toward the shorter plants. Therefore, both the crossed fertilized and the self fertilized plants are skewed left.***

14. T/F: The average height of a cross-pollinated *Zea May* is less than the second quartile height of the same distribution.

True: ***The average height is the mean height which we know from the previous problem has been pulled toward the shorter plants. The second quartile height is the median height which is unaffected by the presence of shorter plants. That is, the mean height is less than the median height.***

The proportions of female students in 30 Alabama four year colleges as reported by Peterson's Guide to Four-Year Colleges (1997)[1] are summarized below.

Min =	10%
Q1 =	53%
Med =	57%
Q3 =	61%
Max =	85%

1. Based on this information, we know that

   A. less than a quarter of the colleges surveyed had more men enrolled than women.
   B. 3/4 of the colleges surveyed have student bodies consisting of more than 50% women.
   C. the average proportion of women in these Alabama colleges in 1997 is less than 57%.
   D. A, B, and C are all true.
   E. none of the above statements are true.

   A and B true because Q1 = 53% female enrollment. Hence, less than 1/4 of the schools could have more men than women and at least 3/4 of the four year colleges in Alabama had more than 50% women enrolled. C is also true because M - Min = 47 % and Max - M = 28%. This means that the lower 50% of the data is more variable than the upper 50%. The mean will be pulled to the left of M = 57% as a result. This implies that the average female enrollment is less than 57%.

2. T/F: Half of the colleges surveyed had a female student enrollment in excess of 53% but less than 61%.

   True: The interquartile range of the data is Q3 - Q1. That is, the middle 50% of the data lies between Q1 = 53% and Q3 = 61%.

3. T/F: At least one of the 30 Alabama colleges has a male student enrollment of 90%.

   True: The school with the lowest female enrollment had only 10% women. Therefore, the remaining 90% of the enrollment must consist of men.

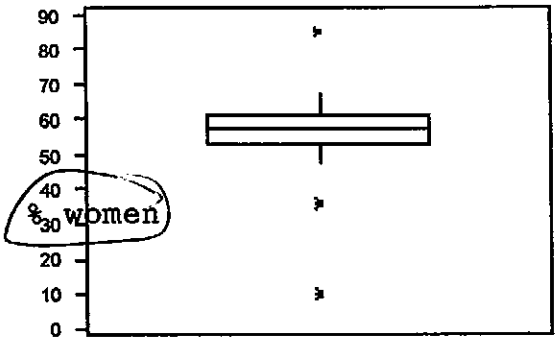
4. T/F: The women students out number the men approximately 6 to 1 in at least one of the 30 colleges.

   True: The school with the maximum female enrollment of 85% has only 15% male enrollment. Since 15 is approximately 1/6th of 85, women out number men about 6 to 1 at this school.

---

[1] Peterson's (1997). Guide to Four-Year Colleges 1998, "Alabama" pgs 244-259.

Below is a graphical representation of the five number summary given on the previous page. Use this plot to answer the following questions.



5. The variable of interest is:

   A. the percentage of women enrolled at Montana State University.
   B. the percentage of women enrolled in college.
   C. the percentage of women enrolled at an Alabama college.
   D. the percentage of women enrolled at a four-year Alabama college.
   E. the percentage of women enrolled at a four-year Alabama college in 1997.

6. **T/F**: 25% of the schools surveyed have student bodies composed of more than 61% women.

   True: ***The third quartile is Q3*** = 61%. ***Therefore, 25% of the schools surveyed have student bodies composed of more than 61% women.***

7. **T/F**: At least two of the thirty schools have more male students than female.

   True: The two stars at the bottom of the plot represent schools with 10% and approximately 35% women enrollment. This implies a male enrollment of 90% and 65% at those same schools. Therefore two schools have a male majority.

8. T/**F**: On average, more men attended the 30 Alabama four-year colleges in 1997 than men.

   False: ***The distribution of female enrollment in Alabama four-year colleges in 1997 is skewed left. The mean is less than 57% but more than Q1 = 53%. Therefore more women attended these schools then men.***

APPENDIX II

DATA SETS

U.S. Steel Data

Year	Exports	Imports	Employee	Hours Wo	Revenue	Assets
1980	4.1	15.1	399	758	37.7	30.8
1981	2.9	19.9	391	753	43.8	31.7
1982	1.8	16.7	289	526	28.7	27.9
1983	1.2	17.1	243	475	25.0	25.5
1984	1.0	26.2	236	474	30.3	26.2
1985	1.0	24.3	208	419	28.4	24.0
1986	1.0	20.7	175	356	25.0	21.0
1987	1.1	20.4	163	354	27.1	21.9
1988	2.1	20.9	169	363	32.7	24.2
1989	4.6	17.3	169	360	31.8	24.6
1990	4.3	17.2	164	350	30.9	28.3
1991	6.3	15.8	146	304	27.1	27.4
1992	4.3	17.1	140	293	26.9	28.2
1993	4.0	19.5	127	274	29.5	30.6
1994	3.8	30.1	126	273	33.5	32.0
1995	7.1	24.4	123	269		

1997 Tuition Cost for Nebraska Colleges

tuition
3.7
11.0
10.7
2.6
1.9
7.5
1.9
19.0
8.3
11.0
9.6
6.7
11.8
12.3
2.1
12.2
4.3
2.3
11.0
8.3
2.7

Canadian Lynx Pelts Sold by Hudson Bay Co. 1857 - 1911

No. of P	In Thous
23362	23.36
31642	31.60
33757	33.76
23226	23.23
15178	15.18
7272	7.27
4448	4.45
4926	4.93
5437	5.44
16498	16.50
35971	35.97
76556	76.56
68392	68.39
37447	37.45
15686	15.69
7942	7.94
5123	5.12
7106	7.11
11250	11.25
18774	18.77
30508	30.51
42834	42.83
27345	27.35
17834	17.83
15386	15.39
9443	9.44
7599	7.60
8061	8.06
27187	27.19
51511	51.51
74050	74.05
78773	78.77
33899	33.90
18886	18.89
11520	11.52
8352	8.35
8660	8.66
12902	12.90
20331	20.33
36853	36.85
56407	56.41
39437	39.44
26761	26.76
15185	15.19
4473	4.47
5781	5.78
9117	9.12
19267	19.27
36116	36.12
58850	58.85
61478	61.48
36300	36.30
9704	9.70
3410	3.41
3774	3.77

Crossed Fertilized vs Self Fertilized Zea May Data

Pot No.	Crossed	SelfFert	Fert-typ	Height
1	23.500	17.375	0	23.500
1	12.000	20.375	0	12.000
1	21.000	20.000	0	21.000
2	22.000	20.000	0	22.000
2	19.125	18.375	0	19.125
2	21.500	18.625	0	21.500
3	22.125	18.625	0	22.125
3	20.375	12.250	0	20.375
3	18.250	16.500	0	18.250
3	21.625	18.000	0	21.625
3	23.250	16.250	0	23.250
4	21.000	18.000	0	21.000
4	22.125	12.750	0	22.125
4	23.000	15.500	0	23.000
4	12.000	18.000	0	12.000
			1	17.375
			1	20.375
			1	20.000
			1	20.000
			1	18.375
			1	18.625
			1	18.625
			1	12.250
			1	16.500
			1	18.000
			1	16.250
			1	18.000
			1	12.750
			1	15.500
			1	18.000