

Exam #2 Review – Chapter 13 Test in Book pp. 785 – 787
Plus additional problems on review.

1. I would guess that $\frac{1}{2}$ of all U.S. workers were working in farm occupations in 1890. Note: We use a bar graph versus a histogram when using categorical or qualitative data. The years 1820, 1840, can be thought of as categories not continuous variables.
2. No, the graph shows percents, not raw numbers. If the population grew, then the *number* of workers could increase even if the percent stayed the same. All we can say from the data is that the percent is 4 times what it was before.
3. About 3 million.
4. In 1940, there were about 6.2 million farms with an average of 174 acres each. That's about 1,079 million acres. In 2000, there were just about 2 million farms at 434 acres per farm on average. That's 868 million total acres. So more land was being farmed in 1940. The percent increase is $(1079 - 868)/868 * 100$, or about 24 or 25%.
5. See answer in back of book
6. See back of book. You might also describe shape, estimate center and spread. For some options for doing that, see below.

The histogram in the back of the book (A-50) looks like it is roughly symmetric.

As measures of center the mean and the median look to be about in the interval 16 – 20. So the mean and median would be about 18 client contacts.

As measure of spread we know the range is 30. We could calculate the standard deviation directly from the given data, but one way to estimate it is to use the empirical rule -- 68% of data falls one standard deviation from the mean and 68% of 22 is about 15. So what spread accounts for 15 of the 22 observations \rightarrow 11 – 15 gives 6 observations, 16-20 gives 7, and 21-25 gives 4, so the spread of 11-25 accounts for $6 + 7 + 4 = 17$ so a little over 68%, but that would be a good measure. The standard deviation would be half the length of the interval from 11 – 25 or about 7 client days. **(this is close to the true standard deviation, which is about 6.1)**

7. If the limits were 7 – 9 the width would be 3 client days. To cover the range of $30 - 7 = 23$ client days, we would need, $23/3 = 7 \frac{2}{3}$, so we'd need 8 classes to cover the full range.

13 – 17. See back of book.

18. We might also consider: are there any outliers, and if so construct a modified boxplot. What is shape of distribution – how does the stem-and-leaf plot show this? How does the boxplot show this?

$$\text{IQR: } 43 - 29.5 = 13.5$$

$$1.5 \times \text{IQR} = 20.25$$

$$\text{Outliers are below } 29.5 - 20.25 = 9.25$$

$$\text{Outliers are above } 43 + 20.25 = 63.25 \rightarrow \text{There are no outliers.}$$

The distribution is skewed to the right based on the stem plot. If you look at the boxplot in the back of the book you will see Q_1 and the median are closer together.

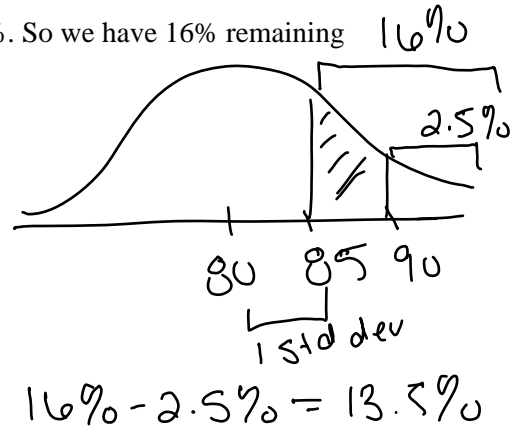
19. Use empirical rule (or Table 10 if you prefer)

70 – 90 lie 2 standard deviations around mean → empirical rule states that 95% of observations will lie within two std. dev. of mean

20. This is the area *NOT* between $z = 3$ and $z = -3$. The empirical rule says 99.7% of the data are in that range, so 0.3% of scores are outside that range.

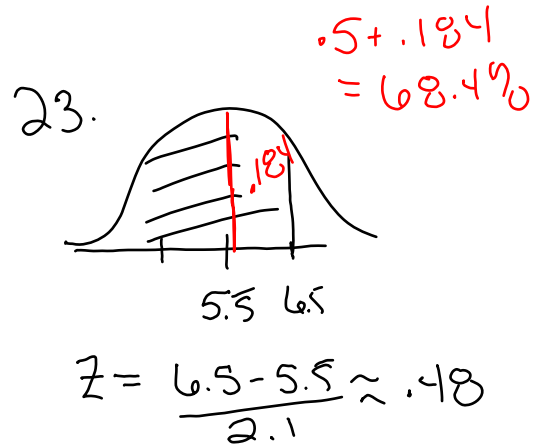
21. Less than 75 is less than $z = -1$. Between the mean and $z = -1$ are 34%. So we have 16% remaining below 75.

22. Between 85 and 90 is the area between $z = 1$ and $z = 2$. Between 0 and 2, we have 47.5%. Between 0 and 1 we have 34%. This leaves 13.5% between 1 and 2, so that's our answer.



(The diagram shows another way to see it, by using the tails.)

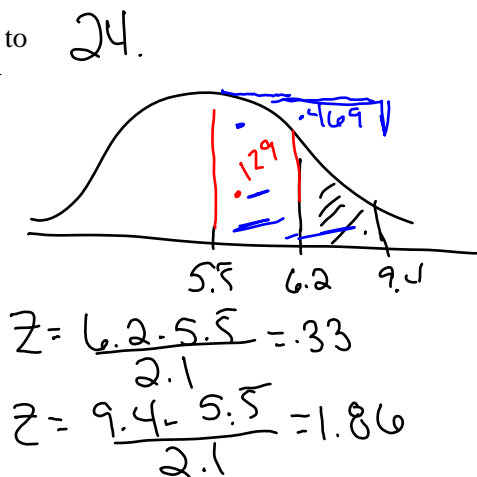
23. $z = (6.5 - 5.5) / 2.1 = .48$. Look up $z = .48$ and find $A = .184 = 18.4\%$. So a total of 68.4% (that's 50% + 18.4%) of the trees are less than 6.5 meters tall.



80% confidence interval: $z = 1.28$.

Interval: $(5.5 \pm 1.28 * 2.1) = 2.8$ to 8.2 meters tall

24. That's the area between $z = .33$ and $z = 1.86$. We'll have to subtract the areas after we look them up in Table 10: $0.469 - 0.129 = 0.371$, or 37.1% of trees.



25. Winning Averages

We need to use weighted averages since there were not always the same number of teams in each league.

$$\text{East: } \frac{76.4 \times 5 + 80.0 \times 5}{10} = 78.2 \quad (\text{Note: weighted average was not really needed here})$$

since each league had 5 teams in its East division. Our answer will be the same as if we did a simple average of 76.4 and 80.0.)

$$\text{Central: } \frac{78.0 \times 5 + 78.3 \times 6}{11} = 78.16 \quad (\text{Note: it would be 78.15 if you did a simple average:})$$

close, but not the same)

$$\text{West: } \frac{91.5 \times 4 + 78.3 \times 5}{9} = 84.17 \quad (\text{Note: it would be 84.9 if you did a simple average:})$$

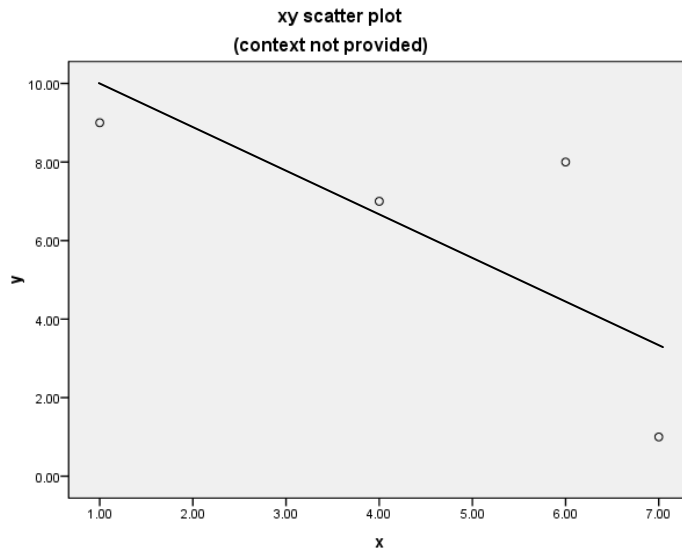
close, not the same).

So the West had the highest winning averages.

27. Need to use weighted averages again since there are different numbers of teams used to create each average.

$$\bar{w} = \frac{5(76.5) + 5(78.0) + 4(91.5) + 5(80.0) + 6(78.3) + 5(78.3)}{30} = 79.99$$

28. Here's the scatter diagram, with a 'best fit line' drawn by eye.



29. The equation of my 'by eye' best fit line (above) can be found by using the endpoints of my line. These appear to be approximately (1, 10) and (7, 3.5). So $m = -5.5/6$, or about -0.92. If $y = -0.92x + b$, then we can use one of these points (say, (1,10)), to solve for b. Work is below.

$$10 = -0.92(1) + b \rightarrow b = 10.92.$$

Therefore, the equation is $y = -0.92x + 10.92$. (Your answer will be different if you sketched a slightly different 'best fit' line.)

The *least squares regression line* computation is based on the following numbers:

	Sum
x	18.00
y	25.00
xy	92.00
x^2	102.00
y^2	195.00

Using the formulas on pg. 779, we calculate: $a = \frac{4(92) - (18)(25)}{4(102) - (18)^2} = \frac{-82}{84} \approx -0.98$ and

$$b = \frac{25 - (-0.98)(18)}{4} = \frac{42.64}{4} \approx 10.66, \text{ so the regression line should be } y = -0.98x + 10.66.$$

Comparing our 'by eye' answer with the least squares regression, we see the two lines have similar slopes and similar y-intercept. Looks pretty good!!

The slope of the LSR line (or best fit line) is about -0.98, so for every unit increase in the x-coordinate, there is a corresponding decrease in the y-coordinate of about -0.98 units. (Note that because the context is not provided here, we cannot be more specific.)

The r^2 value, which can be obtained by a calculator or computer, is about 0.5, which describes a moderately to strong linear trend. Because the slope of the line is downward to the right, the r -value itself will be negative: specifically, about -0.7, which is the *negative* square root of r^2 .

VALUES ARE APPROXIMATE!

30. When $x = 3$, my 'by eye' line equation gives $y = -0.92 \cdot 3 + 10.92 = 7.9$.

When $x = 3$, the least squares regression equation gives $y = -0.98 \cdot 3 + 10.66 = 7.72$.

31. Using the formula on pg. 780, we calculate the following for the *least squares regression*

$$\text{line: } r = \frac{4(92) - (18)(25)}{\sqrt{4(102) - 18^2} \cdot \sqrt{4(195) - 25^2}} = \frac{-82}{\sqrt{84} \cdot \sqrt{155}} \approx \frac{-82}{114.11} \approx -0.72. \text{ This represents a moderate to}$$

strong negative linear relationship between x and y .

Additional Problems

- Measures of center: mean (average), median (middle number), mode (one or two most frequent; otherwise, no mode). Measures of spread: range (max minus min), standard deviation (a sort of average distance from the mean for all data points), IQR (Q3 minus Q1), variance (square of standard deviation). Measures of position: z-score, percentile, decile, quartile.

- a. Stemplot.

<u>Atlanta</u>	<u>Philadelphia</u>
86 2 5	
8644222221 3	000022346668899
74400 4	0000
532200 5	0348
30 6	1
0 7	

For the Atlanta distribution:

Min = 26
 Q1 = 32
 Med = 40
 Q3 = 52
 Max = 70

The data appears to be skewed slightly to the right, with an IQR of 20 and a median of 40 stories.

For Philadelphia distribution:

Min = 25
 Q1 = 32
 Med = 38
 Q3 = 40
 Max = 61

The data appears to be skewed to the right, with an IQR of 8 and a median of 38.

- Atlanta appears to have taller buildings since Q1, Med, Q3 and Max are all higher for Atlanta than for Philadelphia.

c. IQR = 8, and $1.5 \cdot 8 = 12$. If we add that to Q3, we get 52. So yes, 61 is an outlier (for that matter, so are 53, 54, and 58.)

d. Measures of center: The mean would change, but the median would not.

Measures of spread: The IQR would not change, but the range and standard deviation (and variance) would be affected.

e. 80% CI has $z = 1.28$, while the 95% CI has $z = 2$. Confidence interval is then calculated, in either case, as $3.5 \pm z \cdot 2.1$. In the 80% case, we have: (0.8, 6.2), whereas in the 95% case we have: (-0.7, 7.7). So at the 80% confidence level, we have sufficient evidence to conclude that the difference is statistically significant and so Atlanta has significantly taller buildings. However, the 95% confidence interval includes 0, so we cannot say there is a significant difference between the two cities' buildings at the higher 95% level of confidence.

3. 50% of the data are in any two adjacent quartiles. So reviewing our options, we find that only option b (20 to 27) which ranges from the min to the median will contain 50% of the data.

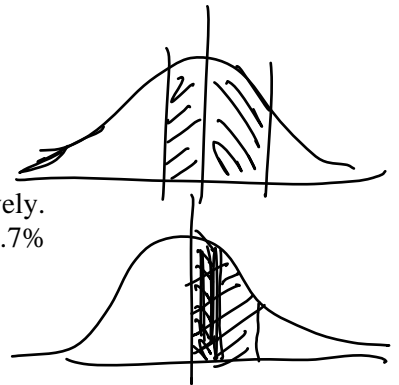
4. a) 10 hamburgers is $z = (10 - 14)/5 = -0.8$. Using table 10, we find $A = 0.288$.

20 hamburgers is $z = (20 - 14)/5 = 1.2$. Using table 10, we find $A = 0.385$.

Adding these results gives 0.673, so about 67% of the time he eats between 10 and 20 hamburgers per month.

b) Here the z-scores are +0.8 and 1.2, and again $A = 0.288$ and $A = 0.385$, respectively.

We must subtract the areas to find the area in between. This gives 0.097, so about 9.7% of the time he eats between 18 and 20 hamburgers per month.



c) At least 11; that's $z > -0.6$. Using Table 10, we have $.226 + .500 = .726$, or 72.6%.

Between 4 and 11; that's z between -2 and -0.6. We subtract $47.5\% - 22.6\% = 24.9\%$.

d) To be at the 68th percentile: There would be 50% below the mean and 18% above the mean. So look up $A = .18$ in Table 10 → We find that the z-score is about 0.47. So we start at the mean and add .47 standard deviations to it: we have $x = 14 + 0.47(5) = 16.35$. So he must eat an average of 16.35 hamburgers per month.

5. The datasets have means of about 5.6 and 7.7, respectively. The average difference is thus about 2.1 units. The mean absolute deviations are about 1.2 and 1.4, respectively, or about 1.3 on average. So the 2.1 unit difference in means is about 1.6 times the mean absolute deviation. Because the gap is larger than the mean absolute deviation, one could argue (albeit somewhat informally) that the difference between the means is a statistically significant one.

6. The mean is $(2 + 7 + 8 + 9 + 9) / 5 = 7$.

The sum of the squared deviations are $(-5)^2 + 0^2 + 1^2 + 2^2 + 2^2 = 25 + 1 + 4 + 4 = 34$.

So $s = \sqrt{34/4} = 2.92$ (rounded).