Math 207
Introduction to Statistics

Exam II Review Exercises

1. **Standard Normal Curve**: Sketch and compute the specified area under the normal curve. Write down the R command that you use.

a) $z < 0.5$

b) $z > 1.5$

c) $-1.5 < z < 0.5$

d) What z is greater than 25% of all z scores?

2. **Blood Pressure**: Suppose that for a particular population systolic blood pressures average out to 120 with an sd of 6 and have a normal distribution.

a) What percent of the scores were below 123?

b) What percent were above 129?

c) What percent were between 111 and 123?

d) What score was greater than 80% of all the test scores?

3. **Linear Regression**: In a certain class, midterm scores average out to 70 with an SD of 15, as do scores on the final. The correlation between midterm scores and final scores is about 0.50.

a) Write down the equation for the regression line for predicting the final exam score.

b) Sketch the regression line.

c) Estimate the average final exam scores for students whose midterm scores were 85, 40 and 70.

d) Estimate the midterm score for a student who scored 90 on the final exam.

4. **Concept of Correlation**:

a) For a representative sample of people age 25 and older, would the correlation between education level (number of years completed in school) and income be positive or negative? Explain.

b) The correlation between education level and blood pressure turns out to be negative? How do you account for this association?

c) Suppose that in a particular population, fathers were always 3" taller than their sons. What can you say about the correlation?

d) Suppose that in a particular population, sisters were always 2" shorter than their brothers. What can you say about the correlation?

5. **Regression and the Normal Curve**: A survey of trees in a state park found that: follows.

average girth $\approx$ 13 inches,	sd $\approx$ 3 inches	
average height $\approx$ 76 feet,	SD $\approx$ 6 feet	$r \approx 0.5$

a) Predict the height of a tree that has an eight inch girth.

b) Among all trees with an eight inch girth, about 95% had a height in what range?

6. **NHANESIII Data Set**:

a) In the exam data set, what information does the variable `PEPPACE` give? What kind of variable is it?

b) In the lab data set, what information does the variable `PBPSI` give? What kind of variable is it? What are its units?

c) In the exam data set, what information does the variable `BMPARML` give? Find the average of all the non-blank values (as you did for other variables in Report 2).

7. **Computing the Correlation and RMS Error**. Here is some data:

x	y	z_x	z_y	$z_x\, z_y$
0	1			
2	5			
4	9			

a) Complete the table then calculate the correlation.

b) Someone uses the line $y = 3\,x - 1$ to approximate the data. Compute the RMS error for this line.

c) What is the RMS error for the regression line?

Exam II Review Exercises

1. **Standard Normal Curve**: Sketch and compute the specified area under the normal curve. Write down the R command that you use.

a) $z < 0.5$. Using R, `pnorm(0.5)=0.691 = 69.1%`.

b) $z > 1.5$. Using R, `1 - pnorm(1.5)=0.0668 = 6.68%`.

c) $-1.5 < z < 0.5$. Using R, `pnorm(0.5) - pnorm(-1.5)=0.625 = 62.5%`.

d) What z is greater than 25% of all z scores? Using R, `qnorm(0.25)=-0.674`.

2. **Blood Pressure**: Suppose that for a particular population systolic blood pressures average out to 120 with an sd of 6 and have a normal distribution.

a) What percent of the scores were below 123?

If $x = 123$, then $z = (123 - 120)/6 = 3/6 = 0.5$. As in #1a, the area is 69.1%.

b) What percent were above 129?

If $x = 129$, then $z = (129 - 120)/6 = 9/6 = 1.5$. As in #1b, the area is 6.68%.

c) What percent were between 111 and 123?

The z scores are 0.5 and $(111 - 120)/6 = -9/6 = -1.5$. As in #1c, the area is 62.5%.

d) What blood pressure was greater than 25% of all the blood pressures?

As in #1d, the z-score is -0.674. Solve $-0.674 = (x - 120)/6$ to get $x = 116.0$.

3. **Linear Regression**: In a certain class, midterm scores average out to 70 with an SD of 15, as do scores on the final. The correlation between midterm scores and final scores is about 0.50.

a) Write down the equation for the regression line for predicting the final exam score.

$y - 70 = 0.5\,\frac{15}{15}\,(x - 70)$ which simplifies to $y - 70 = 0.5\,(x - 70)$.

b) Sketch the regression line. The line goes through the point $(x,\, y) = (70,\, 70)$ and has slope $= 0.5$.

c) Estimate the average final exam scores for students whose midterm scores were 85, 40 and 70.

If $x = 85$, then $y - 70 = 0.5\,(85 - 70) = 0.5 \times 15 = 7.5$. So, $y = 77.5$.

If $x = 40$, then $y - 70 = 0.5\,(40 - 70) = 0.5 \times (-30) = -15$. So, $y = 55$.

If $x = 70$, then $y - 70 = 0.5\,(70 - 70) = 0$. So, $y = 70$.

d) Estimate the midterm score for a student who scored 90 on the final exam.

We need a new regression line! Let $y =$ final exam score and $x =$ midterm score as in the answers above. Our new equation is

$$x - 70 = 0.5\,(y - 70)$$

If $y = 90$, then $x - 70 = 0.5\,(90 - 70) = 0.5 \times 20 = 10$. So, $x = 80$.

Warning: If we used the equation from a), substituted $y = 90$ then solved for x, we would get $x = 110$ which doesn't make sense.

4. **Concept of Correlation**:

a) For a representative sample of people age 25 and older, would the correlation between education level (number of years completed in school) and income be positive or negative? Explain.

It would be positive. On average, completing more years in school (high school diploma, bachelors degree, ...) results in a better-paying job.

b) The correlation between education level and blood pressure turns out to be negative? How do you account for this association?

Since higher education leads to a better salary and job, perhaps those with higher education also have better health care. Maybe they have studied nutrition.

c) Suppose that in a particular population, fathers were always 3" taller than their sons. What can you say about the correlation?

The correlation is equal to 1. All the data falls on the line $y = x + 3$ where x is the son's height, y is the father's height and both are measured in inches.

d) Suppose that in a particular population, sisters were always 2" shorter than their brothers. What can you say about the correlation?

The correlation is equal to 1. All the data falls on the line $y = x - 2$ where x is brother's height and y is sisters height.

5. **Regression and the Normal Curve**: A survey of trees in a state park found that: follows.

average girth $\approx$ 13 inches,	sd $\approx$ 3 inches	
average height $\approx$ 76 feet,	sd $\approx$ 6 feet	$r \approx 0.5$

a) Predict the height of a tree that has an eight inch girth.

The equation for the regression line is height $- 76 = 0.5\frac{6}{3}$ (girth $- 13$) which simplifies to height $- 76 =$ girth $- 13$ or simply height $=$ girth $+ 63$. So, if girth $= 8$ inches, then height $= 71$ feet.

b) Among all trees with an eight inch girth, about 95% had a height in what range?

The RMS for regression is $\text{sd}_y \sqrt{1 - r^2} = 6\sqrt{1 - (0.5)^2} = 6\sqrt{0.75} \approx 5.2$. So, 95% of the trees had heights in the range $71 \pm 2 \times 5.2$ feet $= 71 \pm 10.4$ feet.

6. **NHANESIII Data Set**: Use the variable classes from our notes (unordered qualitative, ordered qualitative, discrete, continuous) not the ones from the text.

a) In the exam data set, what information does the variable `PEPPACE` give? What kind of variable is it?

It indicates whether or not the patient had a pacemaker (page 204). It is an unordered, qualitative variable (yes or no) although it is recorded as a number in the data set.

b) In the lab data set, what information does the variable `PBPSI` give? What kind of variable is it? What are its units?

It gives information about lead levels in the blood. It is a continuous variable. The units are micro-moles per liter.

c) In the exam data set, what information does the variable `BMPARML` give? Find the average of all the non-blank values (as you did for other variables in Report 2).

It is upper arm length measured in cm.

```
exam =read.csv("/Users/ralphwojtowicz/Desktop/math207/data/exam.csv", header=TRUE)
e1 = exam[exam$BMPARML != 8888, ]
mean(e1$BMPARML, na.rm=TRUE)
```

The result is 30.3 cm. See Report 1 for a similar example. Be sure to use the location of exam.csv on your own computer. See page 200 of the exam data set for the error code.

7. **Computing the Correlation and RMS Error**. Here is some data:

x	y	z_x	z_y	$z_x\, z_y$
0	1	-1	-1	1
2	5	0	0	0
4	9	1	1	1

a) Complete the table then calculate the correlation.

$\text{mean}_x = (0+2+4)/3 = 6/3 = 2$. $\text{sd}_x = \sqrt{(4+0+4)/2} = sqrt8/2 = \sqrt{4} = 2$. If $x = 0$, then $z = (0-2)/2 = -1$. The other values are in the table.

$\text{mean}_y = (1+5+9)/3 = 15/3 = 5$. $\text{sd}_y = \sqrt{(16+0+16)/2} = \sqrt{32/2} = \sqrt{16} = 4$. If $y = 1$, then $z = (1-5)/4 = -1$. The other values are in the table.

$r = \sqrt{(1+0+1)/2} = \sqrt{1} = 1$.

b) Someone uses the line $y = 3\,x - 1$ to approximate the data. Compute the RMS error for this line.

x	y	predicted y	error
0	1	-1	-2
2	5	5	0
4	9	11	2

The RMS error is $\sqrt{(4+0+4)/3} = \sqrt{8/3} \approx 1.6$

c) What is the RMS error for the regression line?

It is zero since the points fall on a line. We can use the formula: $\text{sd}_y\sqrt{1-r^2} = \text{sd}_y\sqrt{1-1} = 0$.