XP83 Statistics Final

Fall, 2012

Name: _____

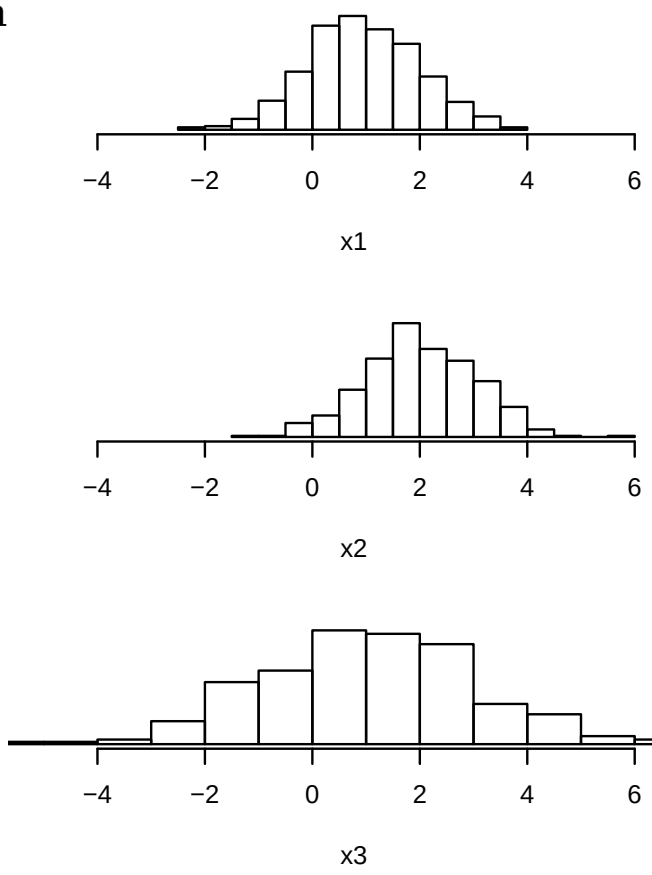
I pledge my honor that I have not violated the Honor Code:

_____

Note:

- You have three hours.
- You may use a pen and a calculator.
- A formula sheet has been provided.
- There are 12 questions.
- Each part of each question is worth 2 points.

1 Question



The three histograms above depict observations on the variables x_1, x_2, and x_3.

Each has average either 1 or 2 and standard deviation either 1 or 2.

(1.1) The mean of x_1 is _____ . The standard deviation of x_1 is _____ .

(1.2) The mean of x_2 is _____ . The standard deviation of x_2 is _____ .

(1.3) The mean of x_3 is _____ . The standard deviation of x_3 is _____ .

(1.4) What is the variance of x_2 ?

(1.5) Give an interval which should contain roughly 95% of the x_1 values.

2 Question

In the countries (conret.xls.txt) data, the usa variables tells us what returns were for a series on months on a portfolio made up of American assets.

The average usa return is .01346.
The standard deviation of usa returns is .0333.

usa is the returns you would have gotten if you put all your money into the american portfolio.

Suppose (unrealistically) that there was an investment available that would give you a return of .002 for sure each month. This is a "riskless" asset.

Suppose instead of putting all your money in usa, you put 60% in the riskless asset and 40% in usa.

Your returns for each month would have been

$$rp = .6\,(.002) + .4\,\text{usa}$$

where usa means the usa return for a given month.

2.1

If you had done this, what would be the mean of your returns?

2.2

If you had done this, what would be the standard deviation of your returns?

3 Question

Suppose the distribution of the random variable X is given by the following table

```
x       p(x)
--      ---
1        .25
2        .5
3        ??
```

3.1

What is $P(X = 3)$?

3.2

What is $P(X < 3)$?

3.3

What is $E(X)$?

3.4

What is $Var(X)$?

3.5

What is σ_X.

4 Question

The table below gives the joint distribution of X and Y.

		X	
		0	1
Y	0	.36	.24
	1	.24	.16

4.1

What is $P(X = 0, Y = 1)$?

4.2

What is $P(X = 0)$?

4.3

What is $P(X = 0 \,|\, Y = 1)$?

4.4

Are X and Y independent?

4.5

Is X a Bernoulli random variable?

4.6

What is $E(X)$?

4.7

What is $Var(X)$?

4.8

Are X and Y iid?

4.9

What is the covariance between X and Y?

5 Question

Suppose

$$R_1 \sim N(.2, .01), \;\; R_2 \sim N(.1, .01).$$

The correlation between R_1 and R_2 is .7.

Let

$$P = .4\, R_1 + .6\, R_2.$$

5.1

What is the covariance between R_1 and R_2?

5.2

What is $E(P)$?

5.3

What is $Var(P)$?

6 Question

Let R denote the uncertain return on an asset next period.
Our uncertainty is represented by

$$R \sim N(.1, .01).$$

6.1

What is $E(R)$?

6.2

What is $Var(R)$?

6.3

What is σ_R ?

6.4

What is $P(R > 0)$?

Working with returns makes us work small numbers. In some cases, a change of a half of one percent is a big deal. If mortgage rates go from 4.5% to 4.0% that matters and that is a change from .045 to .04. Often people work in terms of *basis points*. 100 basis points = 1%.

So, if we want to represent a return as a "percent" we would multiply by 100 (.1 is 10%) and if we want to represent a return in basis points we multiply by 10,000 (.01 is 1% is 100 basis points).
In basis points,

$$B = 10000\,R$$

6.5

What is $E(B)$?

6.6

What is σ_B?

Suppose R_1 and R_2 are iid $N(.1, .01)$.
Let $B_1 = 10000R_1$ and $B_2 = 10000R_2$
(the B's are the R's expressed in basis points).

Suppose we are interested in the difference in the returns expressed as basis points.

Let

$$D = B_1 - B_2.$$

6.7

What is $E(D)$?

6.8

What is $Var(D)$?

6.9

Give an interval such that there is a 95% chance D will end up being in it.

7 Question

Suppose you are in charge of making a part are you are confident that the part making process is "under control" in that whether or not a part is defective is iid Bernoulli with a 1 representing a defect and a 0 representing a good one.

So, your model is $Y_i \sim \text{Bernoulli}(p)$ iid.

However, you still need to learn about p where p is the probability that a part is defective.

7.1

Suppose you collect data on 1,000 parts and find that 95 of them are defective.
What is your estimate of p?

7.2

Give a 95% confidence interval for p.

7.3

Suppose your boss claims that $p = .15$.

Using this data, test the null hypothesis $H_0 : p = .15$.

7.4

Do you think you have evidence to refute your boss's claim?

8 Question

A company wishes to asses the effectiveness of a one day training program.

To make as assessment 100 members of the sales force were randomly selected and then 50 were randomly selected to take the training while the remaining 50 did not.

For each of the salespersons, number of units sold was collected for the week prior to the training and for the week after the training.

So, for each of 100 salespersons we have:
wk1: sales prior to training
wk2: sales after training
T: 1 if they got the training, 0 if they did not.
To analysize the data, the following multiple regression model was fit:

$$wk2 = \alpha + \beta_1\, wk1 + \beta_2\, T + \epsilon.$$

Here is the regression ouput:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.47174    7.02630   2.629  0.00996 **
wk1          0.81866    0.06943  11.791  < 2e-16 ***
T           -0.84577    1.47809  -0.572  0.56851
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.311 on 97 degrees of freedom
Multiple R-squared: 0.5983,Adjusted R-squared:  0.59
F-statistic: 72.22 on 2 and 97 DF,  p-value: < 2.2e-16
```

So, for example, the estimate of β_2 is -.84577 and the associated standard error, t-statistic, and p-value (for $H_0 : \beta_2 = 0$) are 1.47809, -.572, and .56851 respectively.

8.1

Given the sales in wk1 is 95, give the 95% plug-in predictive interval for sales in wk2 for someone who had the training.

8.2

Give the 95% confidence interval for β_1.

8.3

Test the null hypothesis $\beta_1 = 0$.

8.4

Give the 95% confidence interval for β_2.

8.5

Test the null hypothesis $\beta_2 = 0$.

8.6

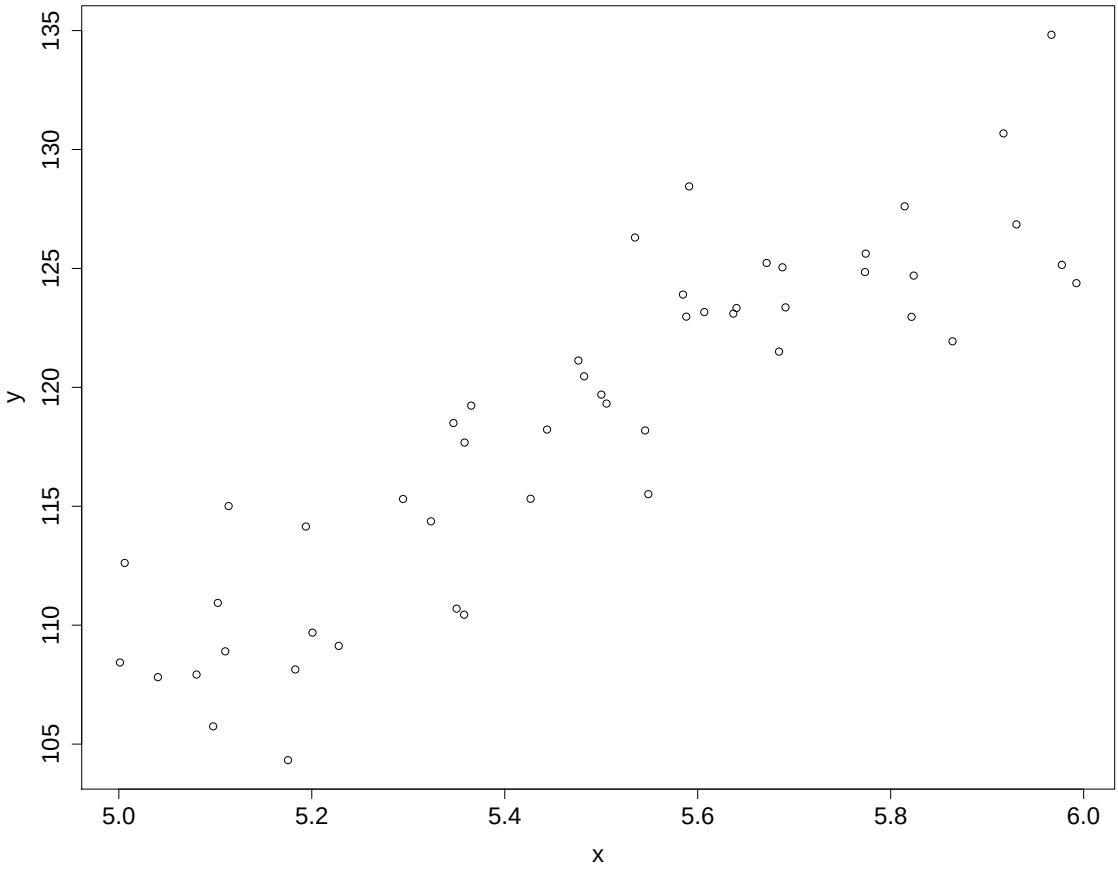
A manager looks at the ouput and comments "wow, I know how to interpret the coefficient of a dummy", this says the effect of the training is actually negative, that's interesting!

Is this a reasonable reaction to the ouput?

8.7

Test the null hypothesis $\beta_1 = 1$.

9 Question



The following questions refer to the x and y graphed above.
Circle your choice.

9.1

The correlation between x and y is
(a) -.78 (b) .96 (c) .23 (d) .89

9.2

The least squares estimate of the intercept is
(a) 25 (b) 105 (c) -5 (d) 56

9.3

The least squares estimate of the slope is
(a) 22.5 (b) 5.8 (c) -12.4 (d) 56.0

9.4

The least squares estimate of σ is
(a) .15 (b) 3.4 (c) 11.3 (d) .01

9.5

R^2 is
(a) .56 (b) .89 (c) .98 (d) .79

9.6

The p-value for testing whether the intercept is equal to 0 is
(a) .001 (b) .59 (c) -.34 (d) .02

9.7

The largest residual is
(a) 7.5 (b) .038 (c) 523.9 (d) 26.7

9.8

The regression plug-in prediction for y given $x = 5.6$ is
(a) 110 (b) -5 (c) 22.53 (d) 121.1

10 Question

(1) If the probability of a defect is p, and parts are iid, the the probability of getting 20 good parts in a row is $(1-p)^{20}$.

T F

(2) The cdf of the standard normal distribution evaluated at -1 is pretty close to .32.

T F

(3) The pdf of the uniform distribution on (-1,1) evaluated at .5 is .5.

T F

(4) To include a categorical independent variable having k possible levels in a multiple regression, we use k dummy variables.

T F

(5) If the p-value is very small, the confidence interval must be small as well.

T F

(6) In the regression model we assume ϵ is independent of Y.

T F

(7) If we toss a fair coin 100 times, we would not be very surprised to get .58 for the fraction of heads.

T F

(8) If $x_1 < x_2$ then $F(x_1) \leq F(x_2)$ where F is a cdf.

T F

(9) If $x_1 < x_2$ then $f(x_1) \leq f(x_2)$ where f is a pdf.

T F

(10) If $X \sim N(4, 4)$, then $Y = -4X + 10 \sim N(-6, 64)$.

T F

(11) If Y_i are iid $N(\mu, \sigma^2)$ then the probability the next 10 Y are greater than μ is $.5^{10}$

T F

(12) $cor(x, y) = cor(ax + b, y)$ where $a > 0$ and b are constants.

T F

(13) The correlation and covariance always have the same sign.

T F

(14) In multiple regression when you add an explanatory variable (an x) to the regression, R^2 cannot go down.

T F

(15) The residuals for a least squares line sum to zero.

T F

(16) In a SLR, R^2 is equal to the square of the sample correlation between the observed Y and X values.

T F

(17) The R^2 for a regression of Y onto X is the same as the R^2 for the regression of X onto Y.

T F

(18) In SLR, $Corr(Y, \hat{Y})$ is always 1.

T F

(19) In SLR, $Corr(X, \hat{Y})$ is always 1.

T F

(20) In SLR, if the $R^2 > .8$, we are guaranteed to make an accurate, precise prediction.

T F

11 Question

An investigator would like to survey a set of people in order to learn what fraction of them use illegal drugs. The survey involves the standard approach of randomly selecting a subset of individuals to survey from the complete list of population members.

The investigator is concerned that potential respondents will be reluctant to answer a question about drug usage truthfully.

The investigator will have each respondent flip a coin.
If it comes up tails, the respondent will answer 1 (yes) or 0 (no) to the question "Is the first digit of your social security number even".
If the coin comes up heads, the respondent will answer 1 (yes) or 0 (no) to the question "do you use illegal drugs".

Thus, each respondent will answer 1 or 0 (yes or no) but the investigator does not know which of the two questions the respondent is actually replying to. The hope is that since the investigator does not know which question was asked, the respondent will give the correct answer.

Let Q be the random variable which is 1 if the coin comes up heads (the drugs question is asked) and 0 if the coin comes up tails (the digit question is asked).

Let R be the random variable representing the answer (1 for yes, 0 for no).

Thus, $P(Q = 1) = P(Q = 0) = .5$.

The investigator believes that $P(R = 1 \mid Q = 0) = .5$.

The investigator would like to know $P(R = 1 \mid Q = 1)$.

Since we will use $P(R = 1 \mid Q = 1)$ a lot, to simplify notation let's also call it p_1:
$p_1 = P(R = 1 \mid Q = 1)$.

First, let's look at things from the point of the respondent.
He might wonder if the investigator can guess what question was asked. For example, if drug use is very low in the population, then a no answer might suggest it was the drug question that was asked.

To investigate this, a respondent supposes that prior to collecting the data, the investigator might believe $p_1 = .1$

11.1

Suppose $p_1 = .1$, what is $P(Q = 1, R = 1)$.

11.2

Suppose $p_1 = .1$. what is $P(R = 1)$?

11.3

Suppose $p_1 = .1$, what is $P(Q = 1 \mid R = 1)$?

11.4

Suppose $p_1 = .1$, what is $P(Q = 1 \mid R = 0)$?

11.5

How do the above probabilities make the respondent feel?
Can the investigator guess the question in a way that matters to the respondent?

Now let's look at things from the point of view of the investigator.
He is trying to esitmate $p_1 = P(R = 1 \mid Q = 1)$.
The survey allows him to estimate $P(R = 1)$.
To simplify notation, let's call this p, $p = P(R = 1)$.

11.6

Note that given p_1, we can figure out p.

Write p as a linear function of p_1.

Note that you can check your function by plugging in $p_1 = .1$ and making sure you get the same answer as you got above!

11.7

Now suppose the survey is done and 450 out of 1,000 respondents answer yes.
What is your estimate of p?

11.8

Now suppose the survey is done and 450 out of 1,000 respondents answer yes.
What is your estimate of p_1?

11.9

Is your estimator for p_1 unbiased?

11.10

Now suppose the survey is done and 450 out of 1,000 respondents answer yes.
Give a 95% confidence interval for p.

11.11

Now suppose the survey is done and 450 out of 1,000 respondents answer yes.
Give a 95% confidence interval for p_1.

11.12

Now suppose the investigator cheated and observed the question.

It actually worked out the each question was asked 500 times and 250 of the digit questions were answered yes and 200 of the drug questions were answered yes.

Give this additional information, give a 95% confidence interval for p_1.

11.13

Which p_1 confidence interval is smaller, the one where did not know about which question was asked or the one where you did?

Why does this make sense?

12 Question

Suppose

$$Y = 20 + 5X_1 + 3X_2 + \epsilon, \;\; \epsilon \sim N(0, 25).$$

and $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, X_1, X_2, ϵ all independent.

12.1

What is $E(Y)$?

12.2

What is $Var(Y)$?

12.3

Given $X_1 = 1$, what is $E(Y)$?

12.4

Given $X_1 = 1$, what is $Var(Y)$?

12.5

Given $X_1 = 1$, $X_2 = -1$, what is $E(Y)$?

12.6

Given $X_1 = 1$, $X_2 = -1$, what is $Var(Y)$?