Statistics 2017 - Midterm

NAME: ____

You have 2 hours.
There are 7 questions, each part of each question is worth 2 points.
You can use only a pen, a calculator, and a hand written cheat-sheet (both sides).

THE CONTENTS OF THIS EXAM ARE CONFIDENTIAL.
DO NOT DISCUSS THEM WITH ANYONE.

I pledge my honor that I have not violated the Honor Code
during this examination.

SIGNATURE: ____

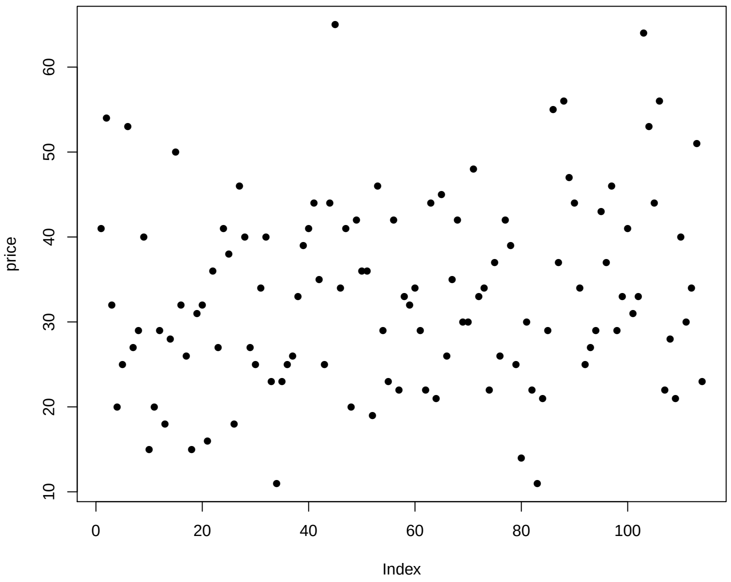
1 Question

Below are data collected from a sample of 114 restaurants.

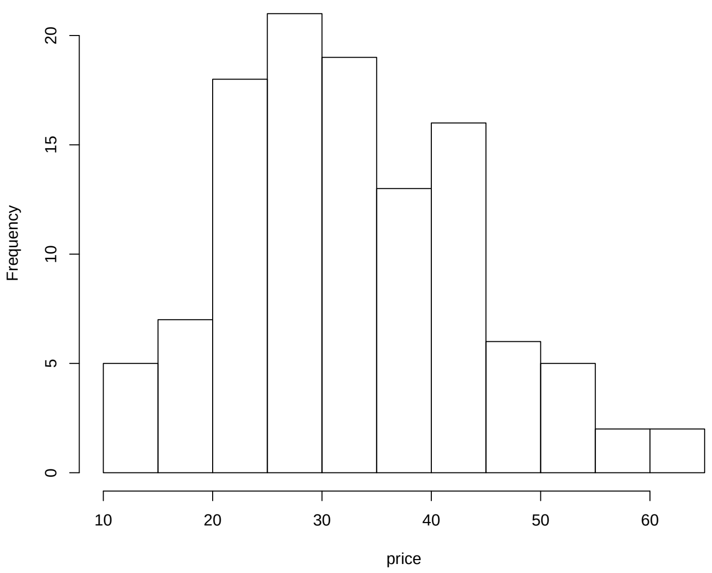
For each restaurant we collect the typical price of a meal in dollars.

Below we have the sequence plot (just the numbers plotted in the order in the data file) and the histogram of `y=price`.

(a) sequence plot of prices



(b) histogram of prices



1.1

The average `price` is:

(a) 11 (b) 65 (c) 33 (d) -22

1.2

The standard deviation of `price` is:

(a) 11 (b) 121 (c) -4 (d) 5

1.3

The variance of `price` is:

(a) 11.78 (b) 123.5 (c) -4.78 (d) 78.4

1.4

The third Quartile (or 75% quantile (or percentile)) of `price` is:

(a) 33 (b) 11 (c) 65 (d) 41

1.5

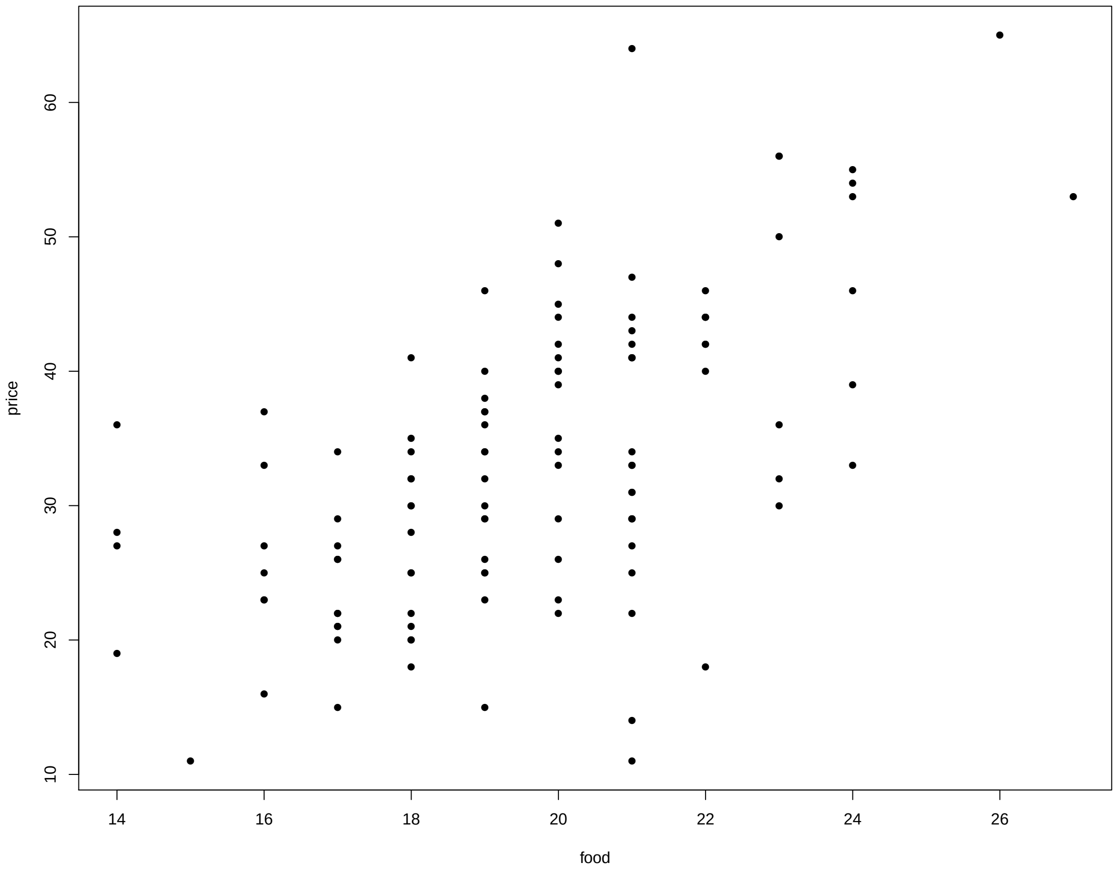
Give an interval which should contain roughly 95% of the `price` values.

2 Question

For each restaurant we also have a rating for the quality of the food.
We record these values in the variable `food`.

How is `price` related to `food`?

Below is the scatter-plot of `price` vs. `food`.



2.1

The correlation between `food` and `price` is:

(a) .95 (b) .12 (c) .60 (d) -.95

2.2

The slope from the linear regression of `price` on `food` is:

(a) -14.5 (b) .6 (c) 1.67 (d) 2.6

2.3

The intercept from the linear regression of `price` on `food` is:

(a) 20 (b) -18 (c) 10 (d) 0

2.4

The standard deviation of `food` is:

(a) 2.5 (b) -17.4 (c) 6.25 (d) 27.8

2.5

The covariance between `food` and `price` is:

(a) 47.8 (b) -12.4 (c) 2.5 (d) 16.9

2.6

Suppose you know the food rating for a restaurant is 18.

What would be your guess (prediction) for the price?

3 Question

Suppose you are uncertain about the time it will take to complete a project.

Let T be the random variable representing the time in days.

t	$P(T = t)$
1	.2
2	.5
3	.3

3.1

What is $P(T > 1)$, the probability of the project taking more than 1 day?

3.2

What is $E(T)$, the expected value of the number of days T?

3.3

What is $Var(T)$, the variance?

3.4

What is σ_T, the standard deviation of T?

Suppose you have a fixed cost of 10 (thousand dollars) and a per day cost of 5.

If we let the random variable denoting the cost of the project be denoted by C, then we have:

$$C = 10 + 5T$$

3.5

What is $E(C)$, the expected value of C?

3.6

What is $Var(C)$, the variance of C?

3.7

What is σ_C, the standard deviation of C?

4 Question

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

		X	
		0	1
Y	0	.12	.20
	1	.12	.56

4.1

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

4.2

What is $P(Y = 1)$?

4.3

What is $P(Y = 1 \mid X = 1)$?

4.4

Are X and Y independent?

4.5

What is the distribution of X?

4.6

What is $E(X)$?

4.7

What is $Var(X)$?

4.8

Are X and Y IID?

4.9

Given you know $X = 0$ what is the distribution of Y?

4.10

Given you know $X = 0$ what is the variance of Y?

4.11

Which is bigger, the variance of Y given you know $X = 0$ or the variance of Y given you know $X = 1$.

4.12

When do you “know more about Y”, when you learn $X = 0$ or when you learn $X = 1$.

5 Question

Suppose $R \sim N(.06, .18^2)$ represents our beliefs about about the return on an asset over the next period.

5.1

What is $P(.06 < R)$?

5.2

What is $P(-.3 < R < .42)$?

5.3

What is $P(-.12 < R < .24)$?

5.4

What is $P(-.3 < R)$?

5.5

What is $P(0 < R)$, the probability of a positive return?

(a) .95 (b) .975 (c) .05 (d) .63

5.6

What is the normal CDF (cumulative distribution function)
evaluated at 0 given $\mu = .06$, $\sigma = .18$?

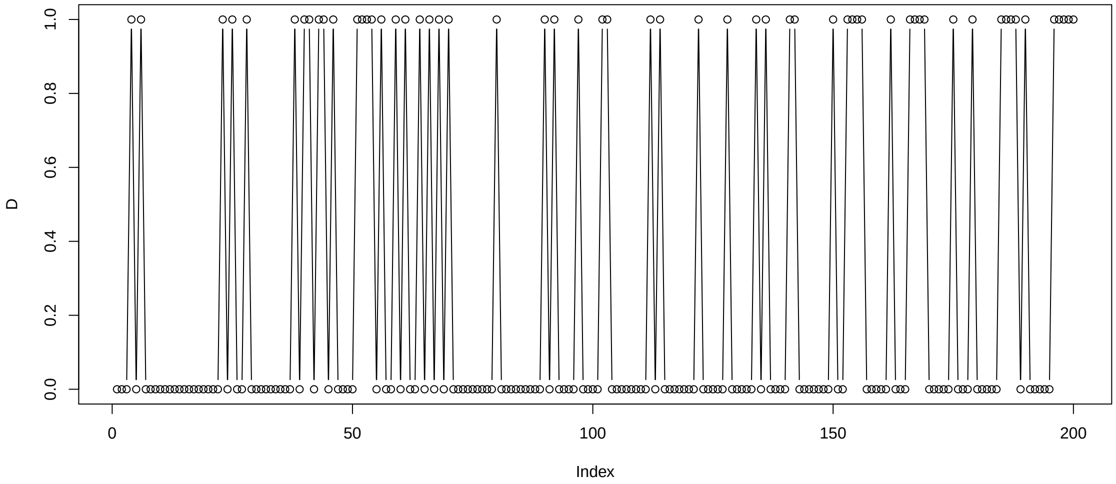
6 Problem

You are in charge of a manufacturing process which produces a single part.

Each time you make a part you inspect it and classify it as either defective: $D = 1$ or not: $D = 0$.

Below are the results from 200 parts, where we have plotted the D values in sequence.

58 of the 200 parts are defective giving an observed defect rate of $58/200 = 0.29$.



6.1

Is it reasonable to model the occurrence of defects in the manufacturing process as IID Bernoulli?

Discuss.

For the rest of this problem *assume* the $D_i \sim$ Bernoulli(.3).
Let D_1, D_2, and D_3 be next three parts so that $D_i \sim$ Bernoulli(.3), $IID, i = 1, 2, 3$.

6.2

What is the probability that the next 3 parts made are all defective $(D_i = 1, i = 1, 2, 3)$?

6.3

What is the probability that the next 3 parts made are all not defective $(D_i = 0, i = 1, 2, 3)$?

Let $T_3 = D_1 + D_2 + D_3$.

T_3 is number of defects out of the next three.

6.4

What is $P(T = 3)$?

6.5

What is $P(T = 1)$?

7 Problem

7.1

Suppose you are considering sending a promotion to one of your customers.

It costs you 2 dollars to send out the promotion.
If they respond you get 100 dollars. If they don't respond you (obviously) get nothing.

Based on past data you think the probability they will respond is .03, or

$$P \sim \text{Bernoulli}(.03),$$

where $P = 1$ means they respond (purchase) and $P = 0$ means they do not.

Should you send out the promotion?

Well, we don't think about just one customer, we think about all our customers.

If we use the same Bernoulli model for each customer, we will either send out the promotion to all of them, or to none of them.

This does not seem right.

We would like to *target* our customers.
This could mean of lot of things, but in this case let's think about which customers are more likely to respond since it makes sense to send the promotion to them.

Based on past purchasing behavior we divide our customers into three groups depending on the amount they have spent with us in the past.

Let $a = 1$, $a = 2$, and $a = 3$ denote low, medium, and high amounts.

Thinking of A (spending level) and P (purchase on future promotion) as random variables we estimate the conditional distribution of A given P from past data:

```
    p(a|P=0) p(a|P=1)
a=1     0.87     0.53
a=2     0.05     0.14
a=3     0.08     0.33
```

So, for example, *given* the customer purchases, the probability they are in amount level 3 is .33.

Let's still suppose $P(P = 1) = .03$.

7.2

What is $P(A = 1, P = 1)$?

7.3

What is $P(A = 1, P = 0)$?

7.4

What is $P(A = 1)$?

7.5

What is $P(P = 1 \,|\, A = 1)$?

7.6

Given a customer is in spending amount level 1 ($A = 1$), should you send them the promotion?

7.7

Given a customer is in spending amount level 3 ($A = 3$), should you send them the promotion?

CHEAT SHEET:

Given numbers $y_1, y_2, \ldots, y_n$ the mean is

$$\bar{y} \ = \ \frac{y_1 + y_2 + \ldots + y_n}{n} \ = \ \frac{1}{n}\sum_{i=1}^{n} y_i$$

Recall: $\sum_{i=1}^{n} y_i$ means for each i from 1 to n add in y_i.

The sample variance is the average squared distance to the mean:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

The sample standard deviation is the square root of the variance.

$$s_y = \sqrt{s_y^2}$$

The sample covariance between x and y is

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The sample correlation between x and y is

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y}$$

For fitted regression line: $y = a + b\,x$:

$$\hat{y}_i = a + b\,x_i.$$

and

$$e_i = y_i - \hat{y}_i$$

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\,\bar{x}.$$

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \ = \frac{s_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x}$$

Discrete Random Variable:
Given $E(X) = \sum x_i \, P(X = x_i)$.
$Var(X) = \sum (x_i - E(X))^2 \, P(X = x_i)$.
$sd(X) = \sqrt{Var(X)}$.

Joint distributions:
$p(x, y) = p(x) \, p(y|x)$,
X and Y are independent if $p(y|x) = p(y)$ for all x, or $p(x, y) = p(x)p(y)$.
Bayes: $p(x|y) = \frac{p(y|x)p(x)}{\sum p(y|x_i)p(x_i)}$.

Normal:
If $X \sim N(\mu, \sigma^2)$ then
$P(\mu - 2\sigma < X < \mu + 2\sigma) = .95$
$P(\mu - \sigma < X < \mu + \sigma) = .68$