2020 Summer STAT 22000 Practice Midterm Exam Solutions

1. [Computations] [18 points]

(a) (6pts) Find the mean, median and standard deviation (SD) for the data below. Show your work to receive full credit.

$$4, \quad 1, \quad 3, \quad 7, \quad 4, \quad -1$$

Mean: ____________ **Median:** ____________ **SD:** ____________

Answer: The sorted list is: $-1, \quad 1, \quad 3, \quad 4, \quad 4, \quad 7.$

$$\text{(2pts) Median} = (3+4)/2 = 3.5, \text{ the average of the two middle numbers in the sorted list}$$

$$\text{(2pts) Mean} = \frac{-1+1+3+4+4+7}{6} = \frac{18}{6} = 3$$

$$\text{(2pts) SD} = \sqrt{\frac{(-1-3)^2+(1-3)^2+(3-3)^2+(4-3)^2+(4-3)^2+(7-3)^2}{5}}$$

$$= \sqrt{\frac{16+4+0+1+16}{5}} = \sqrt{\frac{38}{5}} \approx 2.76.$$

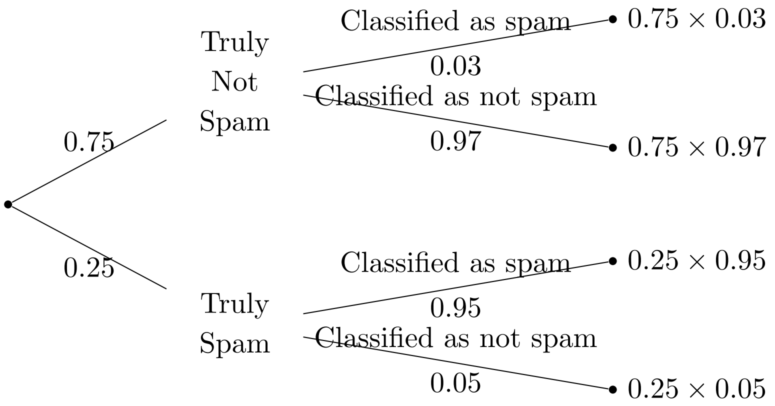
[Grading: Take 1pt off if dividing by $n = 6$ rather than by $n-1 = 6-1 = 5$, which gives SD $= \sqrt{38/6} \approx 2.52$.]

(b) (4pts) Each email message Mary receives is either "spam" or "not spam". Her email software has a built in spam filter that classifies each incoming email message as either "spam" or "not spam". This spam filter is not perfect, however.

- It misclassifies 5% of the truly "spam" emails as "not spam", and
- it misclassifies 3% of the truly "not spam" emails as "spam".
- Suppose 25% of all incoming email messages truly are "spam".

What percentage of the messages marked as "spam" by the software are truly "not spam"? In other words, given that an incoming email message is classified as "spam", what is the probability that it is truly "not spam"?

Answer:



$$P(\text{Truly not spam}|\text{Classified as spam}) = \frac{0.75 \times 0.03}{0.75 \times 0.03 + 0.25 \times 0.95} = \frac{0.0225}{0.0225 + 0.2375} \approx \boxed{0.0865}.$$

(c) (4pts) According to the *National Health and Nutrition Examination Survey*, published by the *National Center for Health Statistics*, the serum (noncellular portion of blood) total cholesterol level of U.S. females 20 years old or older is normally distributed with a mean of 206 mg/dL (milligrams per deciliter) and a standard deviation of 44.7 mg/dL. Determine the percentage of U.S. females 20 years old or older who have a serum total cholesterol level between 150 mg/dL and 250 mg/dL.

Answer: **0.7309.** Let X be the serum total cholesterol level of U.S. females age 20 or over. From the problem, we know $X \sim N(\mu = 206, \sigma = 44.7)$. So

$$\begin{aligned}
P(150 < X < 250) &= P\left(\frac{150-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{250-\mu}{\sigma}\right) \\
&= P\left(\frac{150-206}{44.7} < \underbrace{\frac{X-206}{44.7}}_{=Z} < \frac{250-206}{44.7}\right) \\
&\approx P(-1.25 < Z < 0.98) \\
&= P(Z < 0.98) - P(Z < -1.25) \\
&= 0.8365 - 0.1056 = 0.7309.
\end{aligned}$$

Or one may calculate using the R command:

```
> pnorm(250, m=206, s=44.7)-pnorm(150, m=206, s=44.7)
[1] 0.7323859
```

(d) (4pts) Continue the previous problem. What is the 25th percentile of serum total cholesterol level of U.S. females 20 years old or older?

Answer: The z scores for 25% percentile is about $z = -0.675$ (any value between -0.67 and -0.68 is acceptable). So the corresponding serum total cholesterol level is

$$\mu + \sigma z = 206 + (44.7)(-0.675) \approx 175.8$$

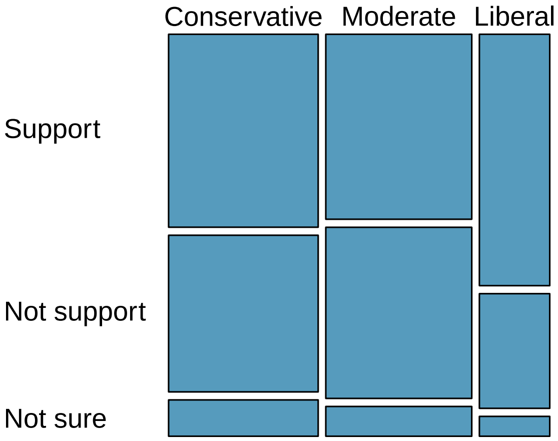
(Depending on the value of z, values between 175.6 and 176.1 are acceptable.)

Or one may calculate using the R command:

```
> qnorm(0.25, m=206, s=44.7)
[1] 175.8503
```

2. [True or False] [11 points]

(a) (3pts) A random sample of registered voters from Tampa, FL were asked if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. The survey also collected information on the political ideology of the respondents. Based on the mosaic plot shown below, determine whether each of the following statement is TRUE or FALSE.



(i) [✓ TRUE] or FALSE: There are more conservative people than liberal people in the sample

(ii) TRUE or [✓ FALSE]: Views on the DREAM Act appear to be independent of political ideology

(iii) TRUE or [✓ FALSE]: Less than half of liberal people in the sample supported the DREAM act

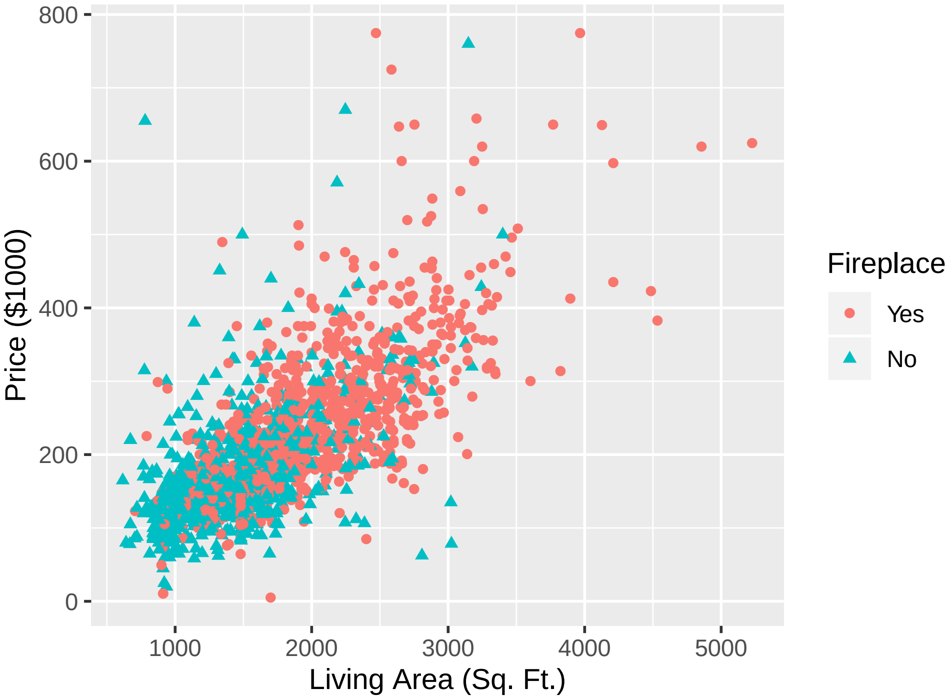
(b) (4pts) A random sample of 1,728 homes was taken from public records of the Saratoga County in New York state in 2007. The scatterplot shows the prices and the sizes of the living area of the 1728 homes. Based on the mosaic plot shown below, determine whether each of the following statement is TRUE or FALSE.

(i) [✓ TRUE] or FALSE: The prices of homes appear to increase linearly with the sizes of the living area.

(ii) [✓ TRUE] or FALSE: The larger the homes, the higher the variability of the prices.

(iii) [✓ TRUE] or FALSE: Larger homes are more likely to have a fireplace.

(iv) TRUE or [✓ FALSE]: The histogram of the sizes of those homes is left-skewed.



(c) (4pts) In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss. Answer the following questions.

For each of the following statements below, determine whether it is TRUE or FALSE.

(i) TRUE or [✓ FALSE]: This is an observational study.

(ii) [✓ TRUE] or FALSE: This study blocks on gender.

(iii) [✓ TRUE] or FALSE: Blinding is used in this study because subjects in the control group received a placebo, so that they were blinded to the treatment they received.

(iv) TRUE or [✓ FALSE]: We can safely generalize the conclusion (that chia seeds have no significant effect on appetite or weight loss) from the participants to the population at large because randomization is used when assigning participants to groups.

Answer: (i) This is an experiment, because the investigator can determine whether a subject go to the treatment group or the control group. (iv) Whether we can make a causal conclusion depends on whether the subjects are randomly sampled from the population. In this study, the subjects were volunteers, not randomly sampled from the population, the conclusion cannot be generalized to the population at large. It is randomization (that subjects are randomly assigned to the treatment and control groups) enables us to make a causal conclusion.

3. [Multiple Choices] [14 points]

Circle the correct answer for each question. There is **only one correct answer** in each question.
Answer part (a-c) based on the table below, summarizes whether respondents smoked or not and whether they had ever divorced or not for persons in the 1991-1993 General Social Surveys who had ever been married.

	Ever Divorced?		
Smoke?	Yes	No	Total
Yes	238	247	485
No	374	810	1184
Total	612	1057	1669

(a) (2pts) What percentage of the overall sample did not smoke and had not ever been divorced?

(i) [✓ 0.485]
(iv) 0.709

Answer: **(iv)** $810/1669 \approx 0.485$

(b) (2pts) Which proportions should we examine if we want to compare the proportion of the people were smokers among those who had ever divorced and those who had not?

(i) 238/485 vs 374/1184
(ii) [✓ 238/612 vs 247/1057]
(iii) 238/1669 vs 247/1669
(iv) 485/1669 vs 612/1669
(v) 238/485 vs 247/485

(c) (2pts) Is smoking habits independent of whether persons ever married had ever divorced or not.

(i) Yes, because Pr(smokers | ever divorced) $\approx$ Pr(smokers)

(ii) Yes, because Pr(smokers | ever divorced) $\approx$ Pr(ever divorced)

(iii) [✓ No, because Pr(smokers | ever divorced) $\not\approx$ Pr(smokers)]

(iv) No, Pr(smokers | ever divorced) $\not\approx$ Pr(ever divorced).

(d) (2pts) Which of the following information can be determined from the histogram of a variable but not from its boxplot? Circle the correct answer.

(i) Whether the distribution of a variable is symmetric or skewed

(ii) Whether there is a outlier

(iii) [✓ The number of modes in the distribution of a variable]

(iv) Where the median is located (roughly)

(e) (2pts) When a statistic, like the median, is said to be resistant to outliers, this means that

(i) it is impossible for the data to have any outliers.

(ii) the statistic is greatly influenced by the value of the outliers.

(iii) [✓ the statistic is not greatly influenced by the value of the outliers.]

(iv) the statistic itself is an outlier

(v) the statistic itself cannot be an outlier

(f) (2pts) A die that is equally likely to land on any of its 6 faces. Consider the following 2 games.

Game A: Rolling the die 60 times, and you win \$1 if it shows an ace (one spot) 5 to 15 times.

Game B: Rolling the die 600 times, and you win \$1 if it shows an ace (one spot) 95 to 105 times.

Which game, A or B, is better (meaning you are more likely to win)? Circle the correct answer.

(i) [✓ A is better]

(ii) B is better

(iii) A and B are equally good.

Not covered in the midterm

(g) (2pts) Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

(i) [✓ $n = 25$, $p = 0.45$]

(ii) $n = 100$, $p = 0.95$

(iii) $n = 150$, $p = 0.05$

(iv) $n = 500$, $p = 0.01$

Not covered in the midterm

Answer: A rule of thumb for a binomial distribution to be approximately normal, np and $n(1-p)$ must be both greater than 10. For (i), both $np = 25 \times 0.45 = 11.25 > 10$; $n(1-p) = 25 \times 0.55 = 13.75 > 10$, so normal approximation is appropriate.

For (ii), $n(1-p) = 100 \times 0.05 = 5 < 10$, for (iii), $np = 150 \times 0.05 = 7.5 < 10$, for (iv), $np = 500 \times 0.01 = 5 < 10$. all the sample sizes p are too small.

4. [Job Satisfaction] [9 points]

The table one the right shows the first 8 observations from a sample of 200 individuals, who reported their age, race, income, and job satisfaction score on a scale from 0 to 100.

Age	Race	Income	Score
21	W	less than \$10,000	29
33	B	\$20,000-30,000	32
41	B	more than \$100,000	84
26	A	\$30,000-40,000	78
22	O	\$10,000-20,000	87
19	A	\$40,000-50,000	42
34	W	\$50,000-75,000	21
26	W	less than \$10,000	91
⋮	⋮	⋮	⋮

(a) (2pts) Which of the following best describes the `Income` variable?

(i) categorical, nominal [✓ (ii) categorical, ordinal]
(iii) quantitative, continuous (iv) quantitative, discrete

(c) (2pts) Which type of plot would be most useful for visualizing the relationship between `Race` and job satisfaction `Score`? (Circle one)

(i) [✓ side by side boxplot] (ii) scatter plot (iii) dot plot
(iv) single boxplot (v) histogram (vi) mosaic plot

(d) (2pts) Below are some summary statistics from the `score` variable. For each of the statements below, determine whether they are TRUE or FALSE. **No explanation is required.**

```
min Q1 median   Q3 max   mean        sd   n missing
 30 57   69.5   77  99 65.075 16.09361 200       0
```

(i) TRUE or [✓ FALSE]: the minimum value of 30 would be identified as an outlier in a box plot
Reason: Q1 -1.5 IQR $= 57 - 1.5(77 - 57) = 27 < 30$.

(ii) [✓ TRUE] or FALSE: there were about the same number of survey respondents who reported job satisfaction scores less than 57 as survey respondents who reported job satisfaction scores greater than 77

(e) (3pts) (This continues part (c)). Is the distribution of `score` right-skewed or left-skewed or (approximately) symmetric. Please support your answer with two pieces of evidence.

Answer: [✓ Left-skewed]. The reasons can be any two of the following

- Mean $= 65.075 <$ Median $= 69.5$
- Q3 $-$ Median $= 77 - 69.5 = 7.5 <$ Median $-$ Q1 $= 69.5 - 57 = 13.5$, indicating that right tail is shorter than the left tail, so it is left-skewed
- max $-$ Q3 $= 99 - 77 = 22 <$ Q1 $-$ min $= 57 - 30 = 27$, indicating that right tail is shorter than the left tail, so it is left-skewed
- max $-$ Median $= 99 - 69.5 = 29.5 <$ Median $-$ min $= 69.5 - 30 = 39.5$, indicating that right tail is shorter than the left tail, so it is left-skewed

5. [Allowance] [20 points]

A boy has his allowance determined every Sunday by drawing one paper bill from an urn containing 5 bills in total: three \$1 bills, one \$5 bill, and one \$10 bill.

$$\boxed{\$1}\ \boxed{\$1}\ \boxed{\$1}\ \boxed{\$5}\ \boxed{\$10}.$$

His parents refill the urn every week, so the content of the urn stays the same every week.

(a) (4pts) Let A be the event that the student draws the ten-dollar bill this Sunday, and B be the event that the student draws the ten-dollar bill the next Sunday.
(i) Are the events A and B disjoint? Explain briefly.
(ii) Are the events A and B independent? Explain briefly.

Answer:

(i) A and B are not disjoint because his parents refill the urn every week, so it is possible that he gets \$10 in both weeks.

(ii) A and B are independent for his parents refill the urn every week, and hence the outcomes of draws are independent from week to week.

(b) (2pts) What is the probability that the boy draws a one-dollar bill for 5 consecutive weeks?

Answer: $(3/5)^5 = 0.07776 \approx \boxed{0.078}$.

(c) (3pts) What is the probability that the boy gets a ten-dollar bill at least once in 5 consecutive weeks?

Answer: $\mathrm{P}(\text{at least once}) = 1 - \mathrm{P}(\text{never gets a \$10}) = 1 - (4/5)^5 \approx \boxed{0.67}$.

(d) (3pts) What is the expected amount of allowance the boy gets in a week?

Answer: The probability distribution of his allowance X in a week is

X	1	5	10
$P(X)$	3/5	1/5	1/5

So

$$E(X) = \$1 \times \frac{3}{5} + \$5 \times \frac{1}{5} + \$10 \times \frac{1}{5} = \frac{\$18}{5} = \boxed{\$3.6.}$$

(e) (3pts) What is the standard deviation of the amount of allowance the boy gets in a week?

Answer: The variance is

$$\text{Var}(X) = (1-3.6)^2 \times \frac{3}{5} + (5-3.6)^2 \times \frac{1}{5} + (10-3.6)^2 \times \frac{1}{5} = 12.64.$$

So the standard deviation is $\sqrt{12.64} \approx \boxed{\$3.555}$.

[*Grading*]: Give 0pt if the variance is computed as the sample variance of 1,1,1,5,10,

$$\frac{(1-3.6)^2 + (1-3.6)^2 + (1-3.6)^2 + (5-3.6)^2 + (10-3.6)^2}{5-1} = 15.8.$$

(f) (5pts) What is the expected value of the amount of allowance that the boy can get in a year? And with what standard deviation? Assume a year has 52 weeks.

Answer: Let X_i be the amount the boy can get in the ith week. The amount he can get in 52 weeks (a year) is $X_1 + X_2 + \cdots + X_{52}$. Observe that the X_i's are independent because the urn stays the same from week to week, and the distribution of X_i's are the same as X in the previous part. So, we know $E(X_i) = 3.6$, and $V(X_i) = 12.64$ for all i. The expected amount of his allowance in a year is hence:

$$E(X_1 + X_2 + \cdots + X_{52}) = E(X_1) + E(X_2) + \cdots + E(X_{52}) = 3.6 \times 52 = \boxed{\$187.2}.$$

and the variance is

$$V(X_1 + X_2 + \cdots + X_{52}) = V(X_1) + V(X_2) + \cdots + V(X_{52}) = 12.64 \times 52 = 657.28$$

So the SD of his allowance in a year is $\sqrt{657.28} \approx \boxed{\$25.64}$.

6. [Underage Drinking] [10 points]

Data collected by the Substance Abuse and Mental Health Services Administration (SAMSHA) suggests that 69.7% of 18-20 year olds consumed alcoholic beverages in 2008.

(a) (3pts) Suppose a random sample of ten 18-20 year olds is taken. Calculate the probability that exactly 7 out of 10 randomly sampled 18-20 year olds consumed an alcoholic drink in 2008.

Answer: The number X who have drunk in 2008 in the sample has the binomial distribution $Bin(n = 10, p = 0.697)$. The chance that $X = 7$ is

$$\begin{aligned} P(X = 7) &= \frac{10!}{7!\,3!} \cdot (0.697)^7 \cdot (1 - 0.697)^3 \\ &= \frac{10 \times 9 \times 8}{3 \times 2 \times 1} \cdot (0.697)^7 \cdot (1 - 0.697)^3 \\ &= 120 \cdot (0.697)^7 \cdot (1 - 0.697)^3 \approx \boxed{0.267} \end{aligned}$$

For part (b-c) below, we now consider a random sample of size 1000 from the population of 18-20 year olds.

(b) (3pts) How many people would you expect to have consumed alcoholic beverages in a sample a random sample of size 1000 from the population of 18-20 year olds? And with what standard deviation?

Answer: Now $X \sim Bin(n = 1000, p = 0.697)$.

So the expected value is $np = 1000 \times 0.697 = \boxed{697}$. ....(1pt)

The SD is $\sqrt{np(1-p)} = \sqrt{1000(0.697)(1 - 0.697)} \approx \boxed{14.53}$. .... (2pts)

Not covered in the midterm

(c) (4pts) What is the probability that 746 or more people in this sample have consumed alcoholic beverages?

Hint: Use normal approximation, but you don't have to use continuity correction.

Answer: By normal approximation of binomial, we know

$$Bin(n = 1000, p = 0.697) \approx N(np, \sqrt{np(1-p)}) = N(697, 14.53)$$

So

$$P(X \geq 746) = P(Z \geq \frac{746 - 697}{14.53}) \approx P(Z \geq 3.372) = 1 - 0.9996 \approx 0.0004.$$

We can see that, the probability to obtain a sample of size 1000 to observe 49 or more people above the expected number (697) to have consumed alcoholic beverages is very small.

Not covered in the midterm

7. [Explanation] [18 points]

(a) The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.

(i) (3pts) Why did they study men and women and the different age groups separately? Explain briefly.

Answer: Age and gender are both confounding variables for health status. Good observational studies control for known confounding variables.

(ii) (3pts) The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly.

Answer: This is the wrong conclusion to draw. Ex-smokers are a self-selected group, and many people give up smoking because they are sick. So recent ex-smokers include a lot of sick people.

(b) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Forty-three percent (43%) of the students returned the questionnaire, and of those who returned, 85% answered "no" to the question. Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

(i) (3pts) Do you think the percentage of parents having no difficulty spending time with their kids after school was a lot higher than, a lot lower than, or somewhat around 85%? Explain.

Answer: The true percentage is probably a lot $\boxed{\text{lower than 85\%}}$. Non-responders may have a different response to this question. The parents who returned the surveys are probably those who do not have difficulty spending time with their kids after school. Parents who work might not have returned the surveys since they probably have a busier schedule.

(ii) (3pts) Suggest what the school officials should have done differently to better estimate the true percentage.

Answer: The school officials should try to $\boxed{\text{increase the response rate}}$, like ask students to remind their parents about the questionnaire or providing reward to for those who completed the questionnaire early and so on.

(c) (3pts) A hypothetical university has two departments, A and B. There are 2000 male applicants, of whom half apply to each department. There are 1100 female applicants: 100 apply to department A and 1000 to department B. Department A admits 50% of the men who apply and 60% of the women. Department B admits 25% of the men who apply and 30% of the women. What percentage of male applicants that were admitted? What percentage of female applicants that were admitted? Which percentage was higher?

	Male		Female	
Department	number of applicants	percent admitted	number of applicants	percent admitted
A	1000	50%	100	60%
B	1000	25%	1000	30%
Total	2000	?%	1100	? %

Answer:

	Male			Female		
Dept	number of applicants	percent admitted	number admitted	number of applicants	percent admitted	number admitted
A	1000	50%	$1000 \times 0.50 = 500$	100	60%	$100 \times 0.6 = 60$
B	1000	25%	$1000 \times 0.25 = 250$	1000	30%	$1000 \times 0.3 = 300$
Total	2000	$\frac{750}{2000} = 37.5\%$	$500 + 250 = 750$	1100	$\frac{360}{1100} \approx 32.7\%$	$60 + 300 = 360$

In the computation above, we see in total 750 out of the 2000 male applicants and 360 of the 1100 female applicants were admitted. So the percentage of women admitted was 37.5%, lower than the percentage of men admitted 32.7%.

(d) (3pts) Continue the previous part. Explain why the female applicants were more likely to be admitted than male applicants within each department but overall female applicants were less likely to be admitted than male applicants when applicants in the two departments are combined.

Answer: This is because the variable “Department” confounds the relationship between “Gender of Applicant” and “Admitted or Not”. We can see that “Department” is associated with “Admitted or Not” as Department A had higher admission rates (50% and 60%) than Department B (25% and 30%). “Department” is also associated with “Gender of Applicant” as females mostly applied for Department B and males were equally likely to applied for either department. As most female applicants applied for Department B which had lower admission rates, this lowered the overall admission rate for females.

This is an example of the Simpson’s paradox.

[Grading: Simply saying this is the Simpson’s paradox is not enough since Simpson’s paradox doesn’t always occur. One must point out the association between “Department” and “Gender of Applicant” and “Admitted or Not” and how the relation between “Gender of Applicant” and “Admitted or Not” get flipped when “Department” is ignored.]