Problem 1. (15 points). Your latest business idea is a discount knock-off of home sharing services, which you plan to call GroundC&C (for "cook and clean" since you plan to save money by providing no hospitality services to guests.)

Furthermore, rather than using fancy machine learning algorithms to match guests with suitable properties, you plan to hire Aggies to manage the reservation process at minimum wage. The first question is whether Aggies are efficient and productive enough to manage this task profitably.

You conduct an experiment, seeing how many bookings can be made per hour, and obtain the following data:

$$15 \quad 20 \quad 25 \quad 35 \quad 40 \quad 45$$

Assuming that this number follows a normal distribution, answer the following questions.

(a) Provide an interval containing the number of bookings that will be made in the next hour, with 95% confidence.

(b) With 90% confidence, give an interval on the mean number of bookings per hour. For the purposes of this experiment, you only want a lower bound on the interval: your only concern is whether the number of bookings exceeds a certain minimum level, and don't care if it's higher.

Solution: With this data we have $\bar{x} = 30$, $s = 11.8$, and $n = 6$.

(a) This is a *prediction interval* with $\alpha = 0.05$, so the interval takes the form $\bar{x} \pm t_{n-1,\alpha/2}\sqrt{1 + 1/n}s$, or $(-2.86, 62.86)$. *(Oops, I didn't check the numbers. Turns out the lower bound is negative.*

(b) This is a *one-sided confidence interval* with $\alpha = 0.1$, so the lower bound on the interval is $\bar{x} - t_{n-1,\alpha} = 22.9$.

Problem 2. (15 points). Based on the results of your experiment, you proceed with trying to obtain capital to start your business. Unfortunately all the banks you speak to are extremely skeptical about your plan, so as a last resort you plan to approach your sketchy brother-in-law, who you are sure will loan you the money. While you are fairly sure he is not connected with organized crime, you still want to be very sure you can pay him back before taking out the loan.

Therefore, you conduct an experiment, running computer simulations to forecast what your monthy profit will be. You will only take out the loan from your brother-in-law if you are very convinced ($\alpha = 0.01$) that your monthly profit will be more than \$10,000. After running 20 simulations to get a sample, you find a mean profit of \$11,000 with a standard deviation of \$500.

(a) What are your null and alternative hypotheses?

(b) State what a Type I and Type II error are in regular English, without using any statistical jargon.

(c) What is the P value for this test?

(d) Do you approach your brother-in-law for a loan? Explain your answer.

Solution:

(a) $H_0 : \mu = 10000$, $H_a : \mu > 10000$

(b) Type I error: taking out the loan when you cannot repay it. Type II error: not taking out the loan when you can repay it.

(c) The t score is $(\bar{x} - \mu_0)/(s/\sqrt{n}) = 8.94$. The corresponding P value is essentially zero. *(Another oops, I did not check the numbers and did not mean for the t value to be so large.)*

(d) Yes, you take out the loan; based on the P value you can reject the null hypothesis for any level of significance.

Problem 3. (20 points). Your business is up and running — congratulations! After the first few months, you notice that many of your customers are causing large amounts of damage to the properties they are staying in. In particular, the amount of damage X seems to be uniformly distributed between \$0 and \$1000. On the other hand, the amount they pay to stay in these properties Y is uniformly distributed between \$500 and \$1500. Assume these values are independent.

(a) What are the mean and standard deviation of the amount of damage?

(b) Write the joint PDF for X and Y.

(c) What is the probability a customer causes more damage than the amount they pay?

(d) What is the correlation coefficient between X and Y?

Solution:

(a) Using properties of the uniform distribtion, $EX = 500$ and $\sigma_X = 289$.

(b) $f_{X,Y}(x,y) = 1/1000^2$ for $(x,y) \in [0,1000] \times [500,1500]$ and zero otherwise.

(c) $\int_{x=500}^{1000} \int_{y=500}^{x} f_{X,Y}(x,y)\ dy\ dx = 1/8$. An easy way to compute this is to compute areas.

(d) X and Y are independent, so $\rho_{X,Y} = 0$.

Problem 4. (15 points). Over time, you also notice that there seems to be a relationship between the rating given by guests (1–5 stars), and the number of flights of stairs they have to climb to access their room. This table shows the fraction of *all* stays that fall into different combinations of these values.

		Flights of stairs B			
		2	4	6	8
Star rating A	1	0.10	0.04	0.06	0.05
	2	0.10	0.08	0.07	0
	3	0.05	0.10	0.10	0
	4	0.04	0.03	0.03	0
	5	0.01	0.05	0.04	0.05

(a) Write the marginal PMFs for A and B.

(b) Are A and B independent? (No credit without explanation.)

(c) If a guest has to walk up 8 flights of stairs, what is the probability they give a 5-star rating?

(d) What is the probability that *exactly one* of the next three guests (in any of your properties) will leave a 5-star rating?

Solution:

(a) $p_A(a) = 1/4, 1/4, 1/4, 1/10$, and $3/20$ for $a = 1, 2, 3, 4, 5$ respectively. $p_B(b) = 3/10, 3/10, 3/10$, and $1/10$ for $b = 2, 4, 6, 8$ respectively.

(b) No, because $p_{A,B}(2, 8) = 0$ while $p_A(2)$ and $p_B(8)$ are both nonzero (or any similar calculation).

(c) $P(A = 5 \mid B = 8) = 0.05/0.10 = 1/2$.

(d) Let X be how many of the next three guests leave a 5-star rating. X is binomial with $n = 3$ and $p = p_A(5) = 3/20$, so $P(X = 1) = 2601/8000 \approx 0.325$.

Problem 5. (15 points). In addition to having guests rate the properties, you conduct a second survey to see how easy to use your app is, with a rating of 1 to 3. (1 = useless, 2 = annoying but eventually worked, 3 = good). In reality, 25% of customers find your app useless, 50% find it annoying but workable, and 25% find it good, and the true average app rating is 2.0.

However, your obsequious assistant is administering the survey, and is secretly inflating the scores you see in an attempt to curry favor with you. Specifically, any score less than 3 is increased by 1 before you see it. (So a customer that reports a 1 will be seen in your records as a 2; but a customer reporting a score of 3 will be correctly seen as a 3.).

Being impatient, you decide to use the first score you see as an estimator for the average customer satisfaction with your app (and not take any other information into account).

(a) What is the PMF of the first score that you see?

(b) What is the bias of this estimator?

(c) What is its mean squared error?

Solution:

(a) Let X be the first score you see; $P(X = 2) = 1/4$ and $P(X = 3) = 3/4$.

(b) $E[X - 2] = 0.75$ using LOTUS.

(c) $E[(X - 2)^2] = 0.75$ also using LOTUS.

Problem 6. (20 points). After five months of operation, you collect the following data for the number of customers and your profit in each month.

Number of customers	75	77	81	95	104
Profit ($)	8400	8800	9300	11400	10000

(a) What is the best-fit line for estimating profit from number of customers?

(b) What is the R^2 value of this line?

(c) Assuming you can have 100 customers in a given month, what is the probability that your profit will be high enough to make the payment on your shady brother-in-law's loan that month? (Remember you need a profit of $10,000 to make the loan payment in a month.)

Solution: For this problem $\sum x = 432$, $\sum y = 47900$, $\sum xy = 4.1839 \times 10^6$, $\sum x^2 = 37956$, and $\sum y^2 = 4.6445 \times 10^8$, so $S_{xx} = 631.2$, $S_{yy} = 5.568 \times 10^6$, and $S_{xy} = 45340$. Further $\bar{x} = 86.4$, $\bar{y} = 9580$, $SSE = 2.311 \times 10^6$ and $SST = 5.568 \times 10^6$.

(a) $\beta_1 = S_{xy}/S_{xx} = 71.8$ and $\beta_0 = \bar{y} - \beta_1 \bar{x} = 3374$, so the line is $y = 3374 + 71.8x$.

(b) $R^2 = 1 - SSE/SST = 0.585$.

(c) $\sigma = \sqrt{SSE/(n-2)} = 878$, so $P(Y > 10000) = P(\epsilon > -557) = P(Z > -0.634) = 0.74$.