Midterm Exam #1 Name: EID:

STA 371G, Statistics and Modeling, Spring 2015

- Please answer all problems in the space provided on the exam. The full score is 100.
- Read each question carefully and clearly present your answers.
- You must show all your work and give a complete explanation. No credit will be given for only the answer without an explanation/equation.
- The exam is closed-book. You are allowed one page of notes. You may use a calculator.

$Z \sim N(0,1)$

x	P(Z<x)	x	P(Z<x)	x	P(Z<x)	x	P(Z<x)
-3	0.0013	-1.5	0.0668	0	0.5	1.5	0.9332
-2.95	0.0016	-1.45	0.0735	0.05	0.5199	1.55	0.9394
-2.9	0.0019	-1.4	0.0808	0.1	0.5398	1.6	0.9452
-2.85	0.0022	-1.35	0.0885	0.15	0.5596	1.65	0.9505
-2.8	0.0026	-1.3	0.0968	0.2	0.5793	1.7	0.9554
-2.75	0.003	-1.25	0.1056	0.25	0.5987	1.75	0.9599
-2.7	0.0035	-1.2	0.1151	0.3	0.6179	1.8	0.9641
-2.65	0.004	-1.15	0.1251	0.35	0.6368	1.85	0.9678
-2.6	0.0047	-1.1	0.1357	0.4	0.6554	1.9	0.9713
-2.55	0.0054	-1.05	0.1469	0.45	0.6736	1.95	0.9744
-2.5	0.0062	-1	0.1587	0.5	0.6915	2	0.9772
-2.45	0.0071	-0.95	0.1711	0.55	0.7088	2.05	0.9798
-2.4	0.0082	-0.9	0.1841	0.6	0.7257	2.1	0.9821
-2.35	0.0094	-0.85	0.1977	0.65	0.7422	2.15	0.9842
-2.3	0.0107	-0.8	0.2119	0.7	0.758	2.2	0.9861
-2.25	0.0122	-0.75	0.2266	0.75	0.7734	2.25	0.9878
-2.2	0.0139	-0.7	0.242	0.8	0.7881	2.3	0.9893
-2.15	0.0158	-0.65	0.2578	0.85	0.8023	2.35	0.9906
-2.1	0.0179	-0.6	0.2743	0.9	0.8159	2.4	0.9918
-2.05	0.0202	-0.55	0.2912	0.95	0.8289	2.45	0.9929
-2	0.0228	-0.5	0.3085	1	0.8413	2.5	0.9938
-1.95	0.0256	-0.45	0.3264	1.05	0.8531	2.55	0.9946
-1.9	0.0287	-0.4	0.3446	1.1	0.8643	2.6	0.9953
-1.85	0.0322	-0.35	0.3632	1.15	0.8749	2.65	0.996
-1.8	0.0359	-0.3	0.3821	1.2	0.8849	2.7	0.9965
-1.75	0.0401	-0.25	0.4013	1.25	0.8944	2.75	0.997
-1.7	0.0446	-0.2	0.4207	1.3	0.9032	2.8	0.9974
-1.65	0.0495	-0.15	0.4404	1.35	0.9115	2.85	0.9978
-1.6	0.0548	-0.1	0.4602	1.4	0.9192	2.9	0.9981
-1.55	0.0606	-0.05	0.4801	1.45	0.9265	2.95	0.9984
-1.5	0.0668	0	0.5	1.5	0.9332	3	0.9987

Problem 1 (10 points)

A construction company has to complete a project no later than four months from now or there will be significant cost overruns. The manager of the construction company believes that there are four possible values for the random variable X, the number of months from now it will take to complete the project: 1, 2, 3, and 4. The manager currently thinks that the probabilities for these four possibilities are in the ratio of 2 to 5 to 2 to 1. That is to say, $X = 1$ is two times more likely than $X = 4$.

(a) (**2 points**) Find the probability distribution of X.

The random variable X has four possible random outcomes: 1, 2, 3 and 4. Since the summation of the probabilities of all possible outcomes must be equal to one, we have

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1.$$

Since $P(X = 1) = 2P(X = 4)$, $P(X = 2) = 5P(X = 4)$ and $P(X = 3) = 2P(X = 4)$, we have

$$2P(X = 4) + 5P(X = 4) + 2P(X = 4) + P(X = 4) = 10P(X = 4) = 1.$$

Therefore

$$\begin{aligned} P(X = 1) &= 0.2 \\ P(X = 2) &= 0.5 \\ P(X = 3) &= 0.2 \\ P(X = 4) &= 0.1 \end{aligned}$$

(b) (**2 points**) What is the expected completion time of this project from now?

The expected completion time of the project is also the mean of the random variable, which can be calculated as

$$\begin{aligned} E(X) &= 1P(X = 1) + 2P(X = 2) + 3P(X = 3) + 4P(X = 4) \\ &= 1 \times 0.2 + 2 \times 0.5 + 3 \times 0.2 + 4 \times 0.1 \\ &= 2.2 \end{aligned}$$

(c) (**2 points**) How much variability exists around the expected completion time? (Hint: calculate the variance/standard deviation)

To measure the variability around $E(X)$, we use the variance $Var(X)$ or standard deviation $sd(X)$. The variance is

$$\begin{aligned} Var(X) =& (1 - E(X))^2 P(X = 1) + (2 - E(X))^2 P(X = 2) \\ &+ (3 - E(X))^2 P(X = 3) + (4 - E(X))^2 P(X = 4) \\ =& (1 - 2.2)^2 \times 0.2 + (2 - 2.2)^2 \times 0.5 \\ &+ (3 - 2.2)^2 \times 0.2 + (4 - 2.2)^2 \times 0.1 \\ =& 0.76 \end{aligned}$$

The standard deviation is

$$sd(X) = \sqrt{Var(X)} = \sqrt{0.76} = 0.87$$

Note that we can also calculate the variance as

$$\begin{aligned} Var(X) =& E(X^2) - (E(X))^2 \\ =& (1)^2 P(X=1) + (2)^2 P(X=2) \\ & + (3)^2 P(X=3) + (4)^2 P(X=4) - 2.85^2 \\ =& 1^2 \times 0.2 + 2^2 \times 0.5 \\ & + 3^2 \times 0.2 + 4^2 \times 0.1 - 2.2^2 \\ =& 0.76 \end{aligned}$$

(d) (**2 points**) Suppose that the manager will get a bonus of $Y = 10 - 2X$ thousand dollars. Find the mean and variance of the manager's bonus.

$$P(Y=8) = 0.2$$
$$P(Y=6) = 0.5$$
$$P(Y=4) = 0.2$$
$$P(Y=2) = 0.1$$

$E(Y) = 10 - 2E(X) = 10 - 2 * 2.2 = 5.6$

$Var(Y) = 4Var(X) = 4 * 0.76 = 3.04$

(e) (**2 points**) What's the probability for the manager to get a bonus that is more than \$5600?
$P(Y > 5600) = P(Y=6) + P(Y=8) = 0.7$

Problem 2 (10 points)

Suppose that 34% of the UT 2014 BBA graduates work in Houston, TX, 10% of the UT 2014 BBA graduates are in the energy industry, and 80% of the UT 2014 BBA graduates who are in the energy industry work in Houston, TX.

(a) (2 points) If we randomly choose a UT 2014 BBA graduate, what is the probability that this person is in the energy industry and works in Houston, TX.

Let $E = 1$ if the student is in the energy industry and $E = 0$ otherwise, and $H = 1$ if the student works in Houston and $H = 0$ otherwise. Since $P(E=1) = 0.10$, $P(H=1) = 0.34$ and $P(H=1|E=1) = 0.80$, we have

$$P(E=1, H=1) = P(E=1)P(H=1|E=1) = 0.10 \times 0.80 = 8\%$$

(b) (2 points) If we randomly choose a UT 2014 BBA graduate, what is the probability that this person is in the energy industry and does *not* work in Houston, TX?

$$P(E=1,H=0) = P(E=1) - P(E=1,H=1) = 0.10 - 0.08 = 2\%$$

(c) (2 points) If we randomly choose a UT 2014 BBA graduate, what is the probability that this person works neither in the energy industry nor in Houston, TX?

$$P(E=0,H=0) = P(H=0) - P(E=1,H=0) = (1-0.34) - 0.02 = 64\%$$

(d) (2 points) If we randomly choose a person from the UT 2014 BBA graduates who work in Houston, TX, what is the probability that this person is in the energy industry?

$$P(E=1|H=1) = \frac{P(E=1,H=1)}{P(H=1)} = \frac{0.08}{0.34} = 23.5\%$$

(e) (2 points) If we randomly choose a person from the UT 2014 BBA graduates who do *not* work in Houston, what is the probability that this person is in the energy industry?

$$P(E=1|H=0) = \frac{P(E=1,H=0)}{P(H=0)} = \frac{0.02}{1-0.34} = 3.03\%$$

Comment: one easy way to do this kind of problem is using a table shown below.

	Energy (E=1)	Not Energy (M=0)	
Houston (H=1)	0.08	0.26	0.34
Not Houston (H=0)	0.02	0.64	0.66
	0.10	0.90	1

Problem 3 (5 points)

Suppose $X \sim \mathcal{N}(10,4)$, i.e., X is normal distributed with mean 10 and variance 4. Compute:

(a) (1 points) $P(X=10)$

$P(X=10) = 0$

(b) (2 points) $P(X>14)$

$$P(X>14) = P(Z > \frac{14-10}{\sqrt{4}}) = P(Z>2) \approx 0.025 \quad \text{(or 2.5\%)}$$

Rule of thumb: In the normal distribution, about 95% of the probability is between -2 and 2 standard deviations.

(c) (2 points) $P(8 < X < 10)$

$$P(8 \leq X \leq 10) = P\left(\frac{8-10}{\sqrt{4}} \leq Z \leq \frac{10-10}{\sqrt{2}}\right) = P(-1 < Z < 0) = P(-1 < Z < 1)/2 = 0.34$$

Problem 4 (5 points)

A company can purchase raw material from either Supplier A or Supplier B and is concerned about the amounts of impurity the material contains. A review of the records for each supplier indicates that the percentage purity levels in consignments of the raw material follow normal distributions with the means and standard deviations given in the table below. The company want to ensure the purity level in a consignment to exceed 93 percent and want to purchase from the supplier more likely to meet that specification. Which supplier should the company choose?

	Mean	Standard Deviation
Supplier A	94.8	0.8
Supplier B	95.0	1.0

Let X_A represent the percentage of purity level in a randomly chosen consignment of raw material from Supplier A. Therefore, $X_A \sim N\left(94.8, 0.8^2\right)$. Similarly, X_B represents the percentage of purity level in a randomly chosen consignment of raw material from Supplier B, and, $X_B \sim \left(95, 1^2\right)$.

We need to compute $P(X_A > 93)$ and $P(X_B > 93)$.

$$P(X_A > 93) = P\left(Z > \frac{93-94.8}{0.8}\right) = P(Z > -2.25)$$

and

$$P(X_B > 93) = P\left(Z > \frac{93-95}{1}\right) = P(Z > -2)$$

Since $P(Z > -2.25) > P(Z > -2)$, we conclude that Supplier A is better and should be chosen.

Problem 5 (20 points)

The figure below shows the 2014 McCombs BBA Salary Survey, which is based on 600 voluntary reports.

BBA FULL-TIME PROFILES **View by** BBA Salary Survey 2014

BBA Salary Survey 2014	Average	Median	Standard Deviation
Full-Time Overall Salaries	$58,769	$60,000	$10,780

Assume the annual salary for a 2014 BBA graduate follows a normal distribution, whose mean is equal to the "Average" reported in the Survey and whose standard deviation is equal to the "Standard Deviation" reported in the Survey. Answer Questions (a)-(e).

(a) (4 points) Find the 95% Confidence Interval for the annual salary of a 2014 BBA graduate.

Since $X \sim \mathcal{N}(, 10780^2)$, we have the 95% Confidence Interval as

$$58769 \pm 2 \times 10780$$

$$(37209, 80329)$$

(b) (4 points) Find the probability for a 2014 BBA graduate to have an annual salary that is between \$58,769 and \$69,549.

$$P(58,769 < X < 69,549) = P(X < 69,549) - P(X \leq 58,769) = P(Z < 1) - P(Z \leq 0)$$
$$= 34\%$$

(c) (4 points) If a 2014 BBA graduate decided to reject a job offer immediately if he/she received an offer with an annual salary that was among the bottom 16%, what would be the least amount of annual salary for him/her to not reject the offer immediately?

Suppose the least amount of salary is x, then

$$P(X < x) = 0.16$$
$$P(\frac{X - 58769}{10780} < \frac{x - 58769}{10780}) = 0.16$$
$$P(Z < \frac{x - 58769}{10780}) = 0.16$$
$$\frac{x - 58769}{10780} = -1$$
$$x = 58769 - 10780 = 47989$$

(d) (4 points) Describe the distribution of the average annual salary of 25 randomly selected 2014 BBA graduates.

$\bar{X} \sim \mathcal{N}(58769, (10780/\sqrt{25})^2)$, or $\bar{X} \sim \mathcal{N}(58769, (2156)^2)$, or $\bar{X} \sim \mathcal{N}(58769, 4648336)$

The average annual salary of \$58,769 reported in the survey provides an estimate of the TRUE average annual salary, which can only be obtained if all 2014 BBA graduates reported their salaries.

(e) (4 points) Does the average annual salary reported in this survey provide an accurate estimation of the true average annual salary of all 2014 BBA graduates? If Yes, provide your explanations and find the 95% Confidence Interval of the true average salary of all 2014 BBA graduates. If No, provide your explanations and give suggestions on how to improve the estimation accuracy.

Yes. $58769 \pm 2 \times \frac{10780}{\sqrt{600}}$, or 58769 ± 880, or $(57889, 59657)$

Comment: if your answers differed from mine but you provided your arguments for these yes/no questions, you were given partial or even full credits.

Problem 6 (10 points)

For a "Yes/No" question, suppose that the proportion of people in the population that would answer the question with "Yes" is p. If we survey n people randomly selected from a large population with this "Yes/No" question, where the population size is considerably larger than n, then the number of "Yes" from a random sample of n people can be considered as a binomial random variable

$$X \sim \text{Binomial}(n, p).$$

If n is not too small and p is not too close to 0 or 1, then $X \sim \text{Binomial}(n, p)$ can be further approximated with a normal random variable as

$$X \sim \mathcal{N}(np, np(1-p)),$$

where np is mean and $np(1-p)$ is the variance.

Suppose in a recent survey, 90 out of 108 randomly selected McCombs BBA students supported introducing an elective course on business analytics. Answer Question (a).

(a) (5 points) Based on this survey, find the 95% Confidence Interval for the true proportion of McCombs BBA students who support introducing an elective course on business analytics.

Since $\hat{p} = 90/108 = 0.8333$, we have the 95% Confidence Interval for the true percentage as

$$0.8333 \pm 2 \times \sqrt{\frac{0.8333 \times (1 - 0.8333)}{108}} = 0.8333 \pm 0.0717$$

$$(0.762, 0.905)$$

For Fall 2014, there were 4,515 BBA students in total and 3,931 of them were Texas residents. Answer Question (b).

(b) (5 points) For Fall 2014, if you randomly select 80 BBA students, what would be the 95% Confidence Interval of the number of Texas residents among these randomly selected 80 BBA students?

Since $p = 3931/4515 = 0.8707$, so approximately

$$X \sim \mathcal{N}(80 * 0.8707, 80 * 0.8707 * (1 - 0.8707)) = \mathcal{N}(70, 3^2)$$

Thus the 95% Confidence Interval is

$$70 \pm 2 \times 3$$

or $[64, 76]$

Problem 7 (20 points)

The table below shows the fuel economy of four randomly selected cars. The weight of a car is measured in thousands of kilograms and the fuel economy is measured in KPL (kilometers per liter). Let X denote Weight and Y denote Fuel Economy (Note that four data points are usually far from enough in practice. We choose four points only for illustration purpose).

Weight (X)	2	1.5	1.8	1.1
Fuel Economy (Y)	7	8	5	10

(a) (2 points) Calculate the sample means of X and Y.

$\bar{x} = 1.6$, $\bar{y} = 7.5$

(b) (2 points) Calculate the sample standard deviations of X and Y.

$s_x = 0.392$, $s_y = 2.082$

(c) (2 points) Calculate the sample covariance between X and Y.

$Cov(X, Y) = -0.667$

(d) (2 points) Calculate the sample correlation between X and Y.

$r_{xy} = Corr(X, Y) = \frac{-0.667}{0.392 \times 2.082} = -0.817$

(e) (2 points) Suppose we use simple linear regression to describe how the Fuel Economy changes as a linear function of the Weight. Calculate the least squares estimates of the intercept and slope. $b_1 = r_{xy}\frac{s_y}{s_x} = -4.34$, $b_0 = \bar{y} - b_1\bar{x} = 14.44$

(f) (2 points) What are the units of the intercept and slope.

Intercept: KPL

Slope: KPL/1000 kilograms

(g) (2 points) Suppose we change the units of Weight from 1000 kilograms to 1000 pounds (1000 kilograms = 2204.62 pounds), what would be the new intercept and slope?

Intercept: 14.44 KPL

Slope:

-4.34 KPL/1000 kilograms $= -4.34$ KPL/2204.62 pounds $= (-4.34 \times \frac{1000}{2204.62})$ KPL/1000 pounds $= -1.97$ KPL/1000 pounds

(h) (2 points) Suppose we not only change the units of Weight from 1000 kilograms to 1000 pounds, but also change the units of Fuel Economy from KPL (kilometers per litter) to MPG (miles per gallon), where 1 KPL = 2.352 MPG, what would be the new intercept and slope?

Intercept: 14.44×2.352 MPG $= 33.96$ MPG

Slope:

-1.97×2.352 MPG/1000 pounds $= -4.63$ MPG/1000 pounds

(i) (2 points) Calculate the coefficient of determination R^2 and explain its meaning.

$R^2 = (-0.817)^2 = 0.667$: the proportion of variation in Y explained by X

(j) (2 points) Based on this analysis, briefly describe your understanding of the relationship between the Weight and Fuel Economy of a car.

The KPL tends to decrease by -4.34 as the weight increases by 1000 kilograms. But one should notice that there are only four points and hence ...

Problem 8 (20 points)

The federal Class III milk price, although not the same as, is closely related to the California mailbox price that a milk farmer in California receives for his milk. Based on the monthly milk price data from May 2004 to May 2007, one can run a simple linear regression model to regress the federal Class III milk price on the California mailbox price. The milk price is measured with \$/cwt, where cwt is a unit of measurement that is roughly 100 pound of milk.

The simple linear regression results are presented in the table below:

	A	B	C	D	E	F	G	H
1	**Linear Regression**							
2								
3	**Regression Statistics**							
4	*R*							
5	*R Square*	B5=?						
6	*Adjusted R Square*							
7	*Standard Error*	0.54						
8	*Total number of observations*	37						
9	**Class III = b0 + b1 * Mailbox**							
10								
11	**ANOVA**							
12		*d.f.*	*SS*	*MS*	*F*	*p-level*		
13	*Regression*	1.	138.11		465.63179	0.E+0		
14	*Residual*	35.	C14=?					
15	*Total*	36.	148.49					
16								
17		*Coefficients*	*Standard Error*	*LCL*	*UCL*	*t Stat*	*p-level*	*H0 (5%) rejected?*
18	**Intercept**	-0.93	0.69346	-2.33371	0.48189	-1.3352	0.19043	*No*
19	**Mailbox**	1.14	0.05268	1.02983	1.24372	21.5785	0.E+0	*Yes*
20	*T (5%)*	2.03011						
21	*LCL - Lower value of a reliable interval (LCL)*							
22	*UCL - Upper value of a reliable interval (UCL)*							

Based on the results presented in the table, answers Questions (a)-(d).

(a) (2 points) Suppose the estimated simple linear regression line is expressed as

$$\text{Class III Price} = b_0 + b_1 \times \text{Mailbox Price},$$

what's the values of b_0 and b_1?

$b_0 = -0.93$

$b_1 = 1.14$

(b) (2 points) What's the value in cell C14?

148.49-138.11=10.39

(c) (2 points) What's the value in cell B5?

$138.11/148.49 = 0.930$

(d) (2 points) We choose b_0 and b_1 to minimize which value in the table?

$C14$ or $B7$

Consider the regression model

$$\text{ClassIII}_t = \beta_0 + \beta_1 \text{Mailbox}_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2),$$

where ClassIII_t represents the milk price in month t for the federal Class III milk and Mailbox_t represents the California mailbox price in month t. Supposing it is true that $\beta_0 = b_0$, $\beta_1 = b_1$ and $\sigma = 0.54$, answers Questions (e)-(g).

(e) (4 points) Suppose the California Mailbox Price is \$12/cwt, what's the 95% Prediction Interval for the price of the federal Class III milk?

$$(-0.93 + 1.14 \times 12) \pm 2 \times 0.54$$

$$12.75 \pm 1.08$$

(f) (4 points) Suppose the California Mailbox Price is \$12/cwt, what's the probability that the federal Class III milk will be lower than \$10/cwt ?

Since given $X = 10$, we have $Y \sim \mathcal{N}(12.75, 0.54^2)$ and hence

$$P(Y < 10) = P(Z < \frac{10 - 12.75}{0.54}) = P(Z < -5.1) \approx 0$$

(g) (4 points) In order to hedge the risk of low milk price in California, in this February, a California milk farmer purchased a put option on the federal Class III milk with a strike price of \$14/cwt. The payoff from the put option is zero, if the strike price is lower than or equal to the Class III milk price, and is equal to the strike price of the put option MINUS the Class III milk price if the strike price is higher than the Class III milk price.

Suppose the California Mailbox Price is \$12/cwt in this August, and the total cost of purchasing and trading the August put option is \$0.79/cwt, what's the probability that this farmer will make a net revenue (mailbox price PLUS payoff from the put option MINUS cost of purchasing and trading the put option) of more than \$13/cwt for his milk in this August?

Since given $X = 12$, we have $Y \sim \mathcal{N}(12.75, 0.54^2)$ and hence

$$P(12 + (14 - Y) - 0.79 > 13) = P(Y < 12.21) = P(Z < -1) = 16\%$$

(Note that if $Y > 14$ when $X = 12$, there is no chance to have a net revenue of \$13/cwt or more.)