STA 113 Name:________________________

Probability and Statistics in Engineering: Final Exam

2:00-5:00 pm Saturday, May 4, 1996

This is a **closed-book** examination, so please do not refer to your notes, the text, or to any other books **except** that you may use **one page** of your own notes, if you wish; if you do so, you must write your name on that page and turn it in along with your examination. A few pages of formulas, tables, and blank work-sheets are appended to the exam. Please do not use your own scratch paper; you may use your own calculator but may not share one.

Please ask Yang or me to explain anything you don't understand in the questions (wording, definitions, *etc.*). It is to your advantage to write your solutions as clearly as possible, and to show enough of your work to make clear how you got your answer. Good luck! Please sign the Duke Honor Code:

I have neither given nor received aid on this examination.

____

signature

1.	/25
2.	/25
3.	/25
4.	/25
5.	/25
6.	/25
	/150

STA 113

Probability and Statistics in Engineering: Final Exam

1. The probability distribution of two random variables X and Y is given by the following table:

		Y	
		5	10
X	10	.10	.40
	20	.30	.20

a. Give the following probabilities:

$\mathsf{P}[X = 10] =$ ______ $\mathsf{P}[Y = 5] =$ ______ $\mathsf{P}[X = 5] =$ ______

$\mathsf{P}[X = 10 \mid Y = 5] =$ ___ $\mathsf{P}[Y = 5 \mid X = 10] =$ ___ $\mathsf{P}[X = 10 \cap Y = 5] =$ ___

b. Are X and Y independent? Justify your answer.

c. Give the possible values z and probability distribution function $p(z)$ for the sum $Z = (X + 2Y)$:

d. Find the following expectations:

$\mathsf{E}[X] =$ ______ $\mathsf{E}[2Y] =$ ______ $\mathsf{E}[X+2Y] =$ ______

$\mathsf{E}[X^2] =$ ______ $\mathsf{E}[(2Y)^2] =$ ______ $\mathsf{E}[(X+2Y)^2] =$ ______

e. Give the following variances:

Var $[X] =$ ______ **Var** $[2Y] =$ ______ **Var** $[X+2Y] =$ ______

Probability and Statistics in Engineering: Final Exam

2. Sixty-four matched pairs of mice are placed in exposure chambers for two hours; within each pair, one is exposed to a combustion byproduct, sulphur dioxide (SO_2), while the other is exposed only to ambient air and the discomforts of the exposure chamber. A technician assigns a Respiratory Impairment (RI) score X_i to each mouse and records the RI score for the SO_2 mouse, the control mouse, and their difference. Here is a summary of the RI scores:

Group	Count	$\sum X_i$	$\bar{X}$	$\sum X_i^2$	S^2
SO_2:	64	704.00	11.00	11776.00	64.00
Control:	64	640.00	10.00	10432.00	64.00
Differences:	64	64.00	1.00	1072.00	16.00

a. What hypothesis do you think is under test here? What is the alternative?
H_0: H_1:

b. What statistical test would you recommend for this hypothesis? Justify your answer. If any degree-of-freedom parameters are necessary, please give them:
Procedure: Degrees of Freedom:______________

c. Perform the test you suggest. Be sure to specify a significance level (your choice) and say whether or not you would reject H_0; give a P-value if possible.

d. Once the mice are assigned to matched pairs, how should it be decided which mouse gets the SO_2 and which gets ambient air? Is it necessary to use a chance mechanism (coin flip, *etc.*) or is it OK for the technician to just pick either mouse for the treatment, then assign the other one to the control? Why?

Probability and Statistics in Engineering: Final Exam

3a. We would like to see whether Lake A, a pristine North Carolina mountain lake, differs in Dissolved Oxygen Deficit (DOD) from Lake B, a similar lake located downstream from a mountain resort. Repeated measurements X_i^A are available from Lake A, X_j^B from Lake B. Which statistical procedure should we use?

○ Paired t test	○ One-way ANOVA F test	○ Simple linear regression
○ Unpaired t test	○ Two-way ANOVA F test	○ Multiple linear regression

Why?

3b. We would like to see whether the compound D, a potential drug, interferes with the binding of the endogenous enzyme E to red blood cells. In repeated independent measurements red blood cells are exposed to independently varied levels of D and of radiolabled E; following exposure the amount of radiolabled E bound to red blood cells is measured and recorded. Which statistical procedure should we use?

○ Paired t test	○ One-way ANOVA F test	○ Simple linear regression
○ Unpaired t test	○ Two-way ANOVA F test	○ Multiple linear regression

Why?

3c. We would like to predict the operating temperature of an electrical component as a function of the applied voltage, switching frequency, and heat-sink size. Which statistical procedure should we use?

○ Paired t test	○ One-way ANOVA F test	○ Simple linear regression
○ Unpaired t test	○ Two-way ANOVA F test	○ Multiple linear regression

Why?

4. In simple linear regression we usually have in mind a model of the form

$$Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$$

for unknown parameters β_0, β_1. Statistical software programs like *SAS* and *Minitab* provide estimates $\hat\beta_0, \hat\beta_1$ for the unknown parameters to "best" fit the data, *i.e.*, to make the *predictions* $\hat Y(X) = \hat\beta_0 + X\hat\beta_1$ "close" to the real observations Y_i, or (equivalently) to minimize the *residuals*

$$e_i = [Y_i - \hat Y_i] = Y_i - \hat\beta_0 - X_i\hat\beta_1$$

a. There are lots of ways we could imagine for measuring how far a line is from the data points, or for measuring how "small" are the residuals. How are the estimates $\hat\beta_0, \hat\beta_1$ determined in linear regression? That is, exactly what measure of "goodness of fit" do they optimize?

b. Even though the variance $\sigma^2 = \mathsf{E}[(Y_i - \beta_0 - X_i\beta_1)^2]$ is the same for every X_i, the prediction intervals $\hat y \pm t_{\alpha/2}s_{(y-\hat y)}$ are not equally long. For which X can we best predict Y, i.e., where is $s^2_{(y-\hat y)}$ *smallest*? (Don't calculate it!)

c. The equation for a (deterministic) straight line is $y = \beta_0 + x\beta_1$; the line passes through $(x, y) = (0, 1)$ and $(x, y) = (2, 3)$. Find β_0 and β_1.

d. True or false? Circle one; no explanations needed.

T F The regression line always passes through $(\bar X, \bar Y)$
T F The sample average residual $\overline{e_i}$ is always zero
T F The regression line always passes through $(0, 0)$
T F The sample average error $\overline{\epsilon_i}$ is always zero
T F If $\mathsf{Cov}(X, Y) > 0$ then the regression slope $\beta_1 > 0$.

Probability and Statistics in Engineering: Final Exam

5. A small fraction p of the population experiences an undesirable side-effect from aspirin. In a sample of $n = 40$ subjects, let X be the number who show the side-effect.

a. What is the probability distribution of X? Why?

b. If $X = 4$ showed the side-effect, find the maximum likelihood estimator of p. Show your work; it's not the answer $\hat{p} = 4/40 = .10$ that's important, it's the *method*.

c. Find a 90% confidence interval for p (still with $X = 4$). You may assume that the sample is large enough to justify a normal approximation.

Probability and Statistics in Engineering: Final Exam

6. Every class day Kelly rides the campus bus from East to West; she keeps track of the number of minutes she waits for the bus each day. If she waits 4 minutes each trip, on average, with a standard deviation of 2 minutes,

a. What is the expected total number of minutes she will wait in the 64 class days of next Fall semester?

b. What is the variance of the total number of minutes she will wait in the 64 class days of next Fall?

c. What is the (approximate) probability that she will spend less than four hours (total) waiting for the bus next Fall? Show your work.

d. What did you have to assume before you could answer parts b. or c. above?

Formulas from Midterms

$$(n)_k = \frac{n!}{(n-k)!} = \overbrace{(n)(n-1)\cdots(n-k+1)}^{k\text{ terms}}$$

$$\text{“}n\text{ choose }k\text{”: } \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{(n)(n-1)\cdots(n-k+1)}{(k)(k-1)\cdots(1)}$$

Distributions

Binomial	$\Pr[X=x] = \binom{n}{x} p^x q^{(n-x)},\ x = 0,1,\ldots,n$	$\mu = np,\ \sigma^2 = npq$
Poisson	$\Pr[X=k] = \lambda^k e^{-\lambda}/k!,\ k = 0,1,\ldots$	$\mu = \lambda,\ \sigma^2 = \lambda$
Geometric	$\Pr[Y=y] = pq^{y-1},\ y = 1,2,\ldots$	$\mu = 1/p,\ \sigma^2 = q/p^2$
Neg. Binomial	$\Pr[Y=y] = \binom{y-1}{r-1} p^r q^{y-r},\ y = r, r{+}1,\ldots$	$\mu = r/p,\ \sigma^2 = rq/p^2$
Hypergeometric	$\Pr[Y=y] = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}},\ y = 0,1,\ldots,n$	$\mu = \frac{nr}{N},\ \sigma^2 =$ a mess
Exponential	$\Pr[X \le x] = 1 - e^{-x/\lambda},\ x \in (0,\infty)$	$\mu = \lambda,\ \sigma^2 = \lambda^2$
	$f(x) = \frac{1}{\lambda} e^{-x/\lambda},\ x > 0;\ 0,\ x \le 0$	
Uniform	$\Pr[X \le x] = (x-a)/(b-a),\ x \in (a,b)$	$\mu = \frac{a+b}{2},\ \sigma^2 = \frac{(b-a)^2}{12}$
	$f(x) = \frac{1}{b-a},\ a<x<b;\ 0,\ x \notin (a,b)$	

$\sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} = (a+b)^n$ $\qquad$ $\int_0^\infty t^n e^{-t}\,dt = n!$

$\sum_{k=0}^{\infty} ar^k = a/(1-r)$ $\qquad$ $\sum_{k=0}^{n} ar^k = (a - ar^{n+1})/(1-r)$

$\Pr[X \le x] = F(x) = \int_{-\infty}^{x} f(t)\,dt$ $\qquad$ $f(x) = F'(x)$

$\mathsf{E}[g(X)] = \int g(x) f(x)\,dx$ $\qquad$ $\mathsf{E}[g(X)] = \sum_x g(x) P(x)$

$\mu = \mathsf{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx$ $\qquad$ $\sigma^2 = \mathsf{E}[(X-\mu)^2] = \mathsf{E}[X^2] - \mu^2$

A **and** $B : A \cap B$ $\qquad$ **not** $A : A^c$ $\qquad$ A **or** $B : A \cup B$

$\mathsf{P}[A|B] = \mathsf{P}[A \cap B]/\mathsf{P}[B]$ $\qquad$ $\mathsf{P}[B|A] = \mathsf{P}[A \cap B]/\mathsf{P}[A]$

$\mathsf{P}[A \cap B] = \mathsf{P}[A|B]\,\mathsf{P}[B]$ $\qquad$ $\mathsf{P}[A \cup B] = \mathsf{P}[A] + \mathsf{P}[B] - \mathsf{P}[A \cap B]$

Independent $\Leftrightarrow \mathsf{P}[A \cap B] = \mathsf{P}[A]\,\mathsf{P}[B]$; $\qquad$ Exclusive $\Leftrightarrow \mathsf{P}[A \cap B] = 0$

Bayes: $\mathsf{P}[A_i|B] = \frac{\mathsf{P}[B|A_i]\mathsf{P}[A_i]}{\mathsf{P}[B|A_1]\mathsf{P}[A_1] + \ldots + \mathsf{P}[B|A_k]\mathsf{P}[A_k]}$ $\quad$ $A_i \cap A_j = \emptyset, \quad \mathsf{P}[A_1] + \ldots + \mathsf{P}[A_k] = 1$

	Continuous	Discrete
Mean	$\mu = \mathsf{E}[X] = \int_{-\infty}^{\infty} x\, f(x) dx$	$= \sum x p(x)$
Variance	$\sigma^2 = \mathsf{V}[X] = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x) dx$	$= \sum (x-\mu)^2 p(x)$
	$= \mathsf{E}[X^2] - \mathsf{E}[X]^2$	$= \mathsf{E}[X^2] - \mathsf{E}[X]^2$
Expectation	$\mathsf{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f(x,y) dx\, dy$	$= \sum_x \sum_y g(x,y)\, p(x,y)$
Covariance	$\mathsf{Cov}[X,Y] = \iint_{-\infty,-\infty}^{\infty,\infty} (x-\mu_x)(y-\mu_y)\, f(x,y) dx\, dy$	$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)\, p(x,y)$
	$= \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]$	$= \mathsf{E}[XY] - \mathsf{E}[X]\mathsf{E}[Y]$
Correlation	$\rho[X,Y] = \mathsf{Cov}[X,Y]/\sigma_X\sigma_Y$	
Conditional	$f(x\|y) = f(x,y)/f_2(y)$	$= p(x,y)/p_2(y)$
Marginals	$f_1(x) = \int_{-\infty}^{\infty} f(x,y) dy$	$= \sum_y p(x,y)$
	$f_2(y) = \int_{-\infty}^{\infty} f(x,y) dx$	$= \sum_x p(x,y)$
Independent	$\Leftrightarrow f(x,y) = f_1(x)\, f_2(y)$	$\Leftrightarrow p(x,y) = p_1(x)\, p_2(y)$
	$\Leftrightarrow f(x\|y) = f_1(x)$	$\Leftrightarrow p(x\|y) = p_1(x)$
	$\Leftrightarrow f(y\|x) = f_2(y)$	$\Leftrightarrow p(y\|x) = p_2(y)$

Conf. Intervals	$\mu = \bar{x} \pm z_{\alpha/2} S/\sqrt{n}$	Large Sample
	$= \bar{x} \pm t_{\alpha/2} S/\sqrt{n}$	Small Sample (normal)
	$p = \hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\,\hat{q}\,/n}$	Population Proportion
Diff. of Means	$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	Large Samples
	$= (\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	Small Samples (normal, same σ^2)
	$= \overline{x_1 - x_2} \pm z_{\alpha/2} s_\Delta/\sqrt{n}$	Matched Pairs, large n
	$= \overline{x_1 - x_2} \pm t_{\alpha/2} s_\Delta/\sqrt{n}$	Matched Pairs, small n, normal
Sample Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	
Sample Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	
Pooled Estimate	$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$	