Mathematics 375 – Probability and Statistics 1
Solutions for Final Examination – December 17, 2005

Directions: Do problems I - X in the blue exam booklet. There are 200 possible regular points and 20 Extra Credit points. Notes:

1) If an answer can be expressed in terms of factorials, binomial coefficients, the gamma function, etc., it is *preferable* to leave it in that form.
2) Even if you don't see how to derive a complete solution to a problem, don't leave it blank. Partial credit will be given for relevant definitions, ideas that could lead to solutions, and so forth.

I. 20 students in a statistics class were asked to report the number of pets in their families, giving the following data:

$$\begin{array}{cccccccccc} 3, & 3, & 2, & 4, & 1, & 0, & 1, & 2, & 3, & 5, \\ 1, & 2, & 2, & 1, & 1, & 0, & 2, & 4, & 3, & 1 \end{array}$$

A) (10) Construct a relative frequency histogram for this data using one "bin" for each integer value.

Solution: There are two 0's, six 1's, five 2's, four 3's, two 4's, and one 5 in the data. So the relative frequency histogram will show a rectangle of height $2/20 = 1/10$ for the value 0, a rectangles of height $6/20 = 3/10$ for the value 1, $5/20 = 1/4$ for the value 2, $4/20 = 1/5$ for the value 3, $2/20 = 1/10$ for the value 4, and $1/20$ for the value 5.

B) (10) How many data points are within one sample standard deviation of the sample mean?

Solution: The sample mean is $\overline{x} = (\sum_{i=1}^{20} x_i)/20 = 2.05$. The sample standard deviation is

$$s = \sqrt{\frac{1}{19}\sum_{i=1}^{20}(x_i - \overline{x})^2} \doteq 1.36$$

So the interval $[\overline{x} - s, \overline{x} + s]$ is approximately $[.69, 3.41]$. 15 out of the 20 data points are in this range.

II. Suppose that you have a key ring with N keys, exactly one of which is your house key. You get home after dark and can't see the keys or the ring. So you try the keys in your front door lock one by one until you find the right one.

A) (5) Assuming that you are not careful and you mix up the keys before each try, what is the probability that you first find the correct key on the yth try ($y = 1, 2, \ldots$)?

Solution: Since you mix the keys back up each time, you have a probability of finding the correct key of $1/N$ on each try, and the tries are independent. The number of the try on which you first find the key is a *geometric* random variable with $p = 1/N$ and $q = (N-1)/N$. The probability that you find the key first on the yth trial is $((N-1)/N)^{y-1}(1/N)$.

B) (5) Let Y_A be the discrete random variable giving the number of the trial on which you find the correct key by the method from part A. What is the expected value of Y_A?

Solution: From the formula for geometric random variables, $E(Y_A) = 1/p = N$.

C) (5) Same question as part A, but now you put aside each key as you try it so it does not get mixed back in with the ones you have not tried.

Solution: Now, you have a probability of $1/N$ of finding the correct key on the first trial and a probability of $(N-1)/N$ of not finding the correct one. If you don't find the correct one, then the first choice is removed from the "pool" and you have a $1/(N-1)$ chance of finding the correct key on the second try, and an $(N-2)/(N-1)$ probability of not finding the correct one. Similarly for the later trials. The probability of finding the correct key on the yth trial is

$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{N-3}{N-2} \cdots \frac{N-y+1}{N-y+2} \cdot \frac{1}{N-y+1} = \frac{1}{N}$$

(the same for all trials $y = 1, \ldots, N$), and zero for $y > N$.

D) (5) Let Y_C be the discrete random variable giving the number of the trial on which you first find the correct key by the method from part C. What is the expected value of Y_C? Which method is superior?

Solution:

$$E(Y_C) = 1 \cdot \frac{1}{N} + 2 \cdot \frac{1}{N} + \cdots + N \cdot \frac{1}{N} = \frac{N(N+1)}{2} \cdot \frac{1}{N} = \frac{N+1}{2}$$

As should be intuitively clear, this method is superior to the one from part A. You'll find the correct key faster on average this way!

III. A tea distributor randomly places a miniature porcelain figurine of one of 10 different endangered animals in each of its boxes of tea.

A) (15) What is the probability of getting a complete set of all 10 animals if you purchase exactly 10 boxes of the tea? Explain how you are deriving your answer.

Solution: Think of buying the boxes one by one and placing the animals you get in an ordered row. There are 10^{10} different (ordered) collections of 10 animals (including possible duplicates). Of those collections, $10 \cdot 9 \cdot 8 \cdots 2 \cdot 1 = 10!$ consist of 10 distinct animals. So the probability of getting a complete set when you buy just 10 boxes is

$$\frac{10!}{10^{10}}$$

(This is $\doteq .000363$ – not very likely!)

B) *Extra Credit* (10) What is the expected number of boxes of tea you need to purchase to obtain a complete set of all 10 animals? (Hint: You get one in the first box you buy. What is the expected number of boxes until you get a second, different animal? After you have two, what is the expected number of boxes to get the third? etc.)

Solution: After you get the first animal, since the animals are random, you have a $9/10$ chance of getting something different from the first with each box from that point on. The number of the box on which you first get something different from the first is *geometric* with $p = 9/10$, so the expected number of boxes until you get something different is $1/p = 10/9$. Once you do get a different second animal, you're in the same situation, but now there is a probability of $p = 8/10$ that you get something different from the first two distinct ones. You expect a further $10/8$ tries to get a third distinct animal. Keep going the same way. The expected number of boxes for a complete set is:

$$1 + 10/9 + 10/8 + \cdots + 10/2 + 10/1 = 10\left(\frac{1}{10} + \frac{1}{9} + \frac{1}{8} + \cdots + \frac{1}{2} + \frac{1}{1}\right) \doteq 29.29$$

(smaller than you might expect!)

IV.

A) (5) State Bayes' Rule.

Solution: Bayes' Rule states that if a sample space S is partitioned into subsets $S_1, \ldots, S_k$ and $E \subset S$ is an event, then

$$P(S_j|E) = \frac{P(E|S_j)P(S_j)}{\sum_{i=1}^{k} P(E|S_i)P(S_i)}$$

B) (15) According to the American Lung Association, 7% of the population has lung disease. Of those people having lung disease, 90% are smokers. Of those not having lung disease, 74.7% are nonsmokers. What is the probability that a smoker has lung disease?

Solution: In the sample space S consisting of the whole population, we have the partition $S = L \cup \overline{L}$, where L are the people with lung disease. We also have $A \subset S$, the set of smokers. The given information is $P(L) = .07$ and $P(A|L) = .9$, $P(\overline{A}|\overline{L}) = .747$.

We want $P(L|A)$. Because of the "switch" we see that Bayes' Rule will apply to this situation. Applying the complement rule,

$$P(L|A) = \frac{P(A|L)P(L)}{P(A|L)P(L) + P(A|\overline{L})P(\overline{L})} = \frac{(.9)(.07)}{(.9)(.07) + (.253)(.93)} \doteq .211$$

V. The number of customers arriving at a furniture store in a one-hour period has a Poisson distribution with mean 5.

A) (10) What is the probability that more than 7 customers arrive in one hour?

Solution: From the Poisson table for $\lambda = 5$, $P(Y > 7) = 1 - P(Y \le 7) = 1 - .867 = .133$.

B) (10) What is the probability that more than 7 customers arrive in exactly 4 out of the 8 hours that the store is open in one day? (Assume the numbers in each hour are independent.)

Solution: Since the hours are independent, the number Y of hours where more than 7 customers arrive is a binomial random variable with $n = 8$, $p = .133$, $q = .867$:

$$P(Y = 4) = \binom{8}{4}(.133)^4(.867)^4 \doteq .0124$$

VI. The pdf for a continuous random variable Y has the form

$$f(y) = \begin{cases} k(9 - y^2) & \text{if } -3 \le y \le 3 \\ 0 & \text{otherwise} \end{cases}$$

A) (10) Determine the value of k.

Solution: The value of k is determined by the equation

$$1 = k \int_{-3}^{3} 9 - y^2 \; dy = k \left(9y - y^3/3\right|_{3}^{3} = 36k$$

So $k = 1/36$.

B) (15) What are the expected value and variance of Y?

$$E(Y) = \int_{-3}^{3} y \cdot (9 - y^2)/36 \; dy = 0$$

(Note no computation needed here – integral of an odd function over a symmetric interval!) Then

$$V(Y) = \int_{-3}^{3} y^2 \cdot (9 - y^2)/36 \; dy - 0^2 = \frac{9}{5}$$

VII. Professional golfers' tee shot distances are normally distributed with $\mu = 282.2$ yards and $\sigma = 8.12$ yards.

A) (10) What is the probability that a tee shot goes more than 300 yards?

Solution: Standardize and use the standard normal table:

$$P(Y > 300) = P\left(\frac{Y - 282.2}{8.12} > \frac{300 - 282.2}{8.12}\right) = P(Z > 2.19) = .0143.$$

B) (10) What is the probability that a tee shot goes between 270 and 290 yards?

Solution: Same idea:

$$\begin{aligned} P(270 < Y < 290) &\doteq P(-1.50 < Z < .96) \\ &= 1 - P(Z < -1.50) - P(Z > .96) = 1 - .0668 - .1685 = .7647 \end{aligned}$$

VIII. (15) Two phone calls are made independently in a fixed 1-hour period. Each call occurs at a random time uniformly distributed on the interval $[0, 1]$. Find the joint density function for the times Y_1, Y_2 of the two calls and use it to find the probability that the two calls are made within 5 minutes of each other (i.e. $|Y_1 - Y_2| \le 5/60$).

Solution: Because the times are independent, the joint density is the product of the uniform densities for Y_1 and Y_2:

$$f(y_1, y_2) = \begin{cases} 1 & \text{if } 0 \le y_1, y_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Note this is nonzero exactly on the unit square in the (y_1, y_2) plane. The region R corresponding to the condition $|y_1 - y_2| \le 1/12$ is a diagonal strip down the middle of the square bounded by the lines $y_2 = y_1 + 1/12$ and $y_2 = y_1 - 12$. Then

$$P(|Y_1 - Y_2| \le 1/12) = \int\int_R f(y_1, y_2)\, dA$$

is the area of R. We can find this by subtracting the areas of the two triangles outside the strip:

$$\text{area}(R) = 1 - 2 \cdot \frac{1}{2} \cdot \frac{11}{12} \cdot \frac{11}{12} = \frac{23}{144}$$

IX. Two random variables Y_1, Y_2 have joint density

$$f(y_1, y_2) = \begin{cases} 2e^{-y_1 - y_2} & \text{if } 0 \le y_1 \le y_2 < \infty \\ 0 & \text{otherwise} \end{cases}$$

A) (10) Find $P(Y_2 \le 2 + Y_1)$. (Draw a picture!)

Solution: The region where the joint density is nonzero is the region in the first quadrant above the line $y_2 = y_1$. So to compute $P(Y_2 \leq 2 + Y_1)$, note the line $y_2 = y_1 + 2$ is parallel to the boundary line and

$$\begin{aligned} P(Y_2 \leq 2 + Y_1) &= \int_0^\infty \int_{y_1}^{y_1+2} 2e^{-y_1-y_2}\, dy_2 dy_1 \\ &= \int_0^\infty -2e^{-2y_1-2} + 2e^{-2y_1}\, dy_1 \\ &= 1 - e^{-2} \end{aligned}$$

B) (10) Find the marginal density $f_1(y_1)$. What is the distribution of Y_1?

Solution:

$$f_1(y_1) = \int_{y_1}^\infty 2e^{-y_1-y_2}\, dy_2 = 2e^{-2y_1}$$

(if $y_1 > 0$; 0 otherwise). This is the density for an exponential random variable with $\beta = 1/2$.

C) (10) Let $U = 3Y_1 + 2Y_2$. Find the mean μ_U.

Solution: By linearity, $E(U) = 3E(Y_1) + 2E(Y_2)$. Using part B, $E(Y_1) = \frac{1}{2}$. Now setting up the double integral with dy_2 outside:

$$\begin{aligned} E(Y_2) &= \int_0^\infty \int_0^{y_2} y_2 \cdot 2e^{-y_1-y_2}\, dy_1 dy_2 \\ &= \int_0^\infty -2y_2e^{-2y_2} + 2y_2e^{-y_2}\, dy_2 \\ &= -\frac{2}{4}\Gamma(2) + 2\Gamma(2) = \frac{3}{2} \end{aligned}$$

So $E(U) = 3 \cdot \frac{1}{2} + 2 \cdot 3/2 = \frac{9}{2}$.

D) *Extra Credit* (10) Find the variance σ_U^2 for U from part C.

Answer: $V(U) = \frac{41}{4}$. The easiest way to do this is *not to compute the covariance matrix.* Instead, since you already have the mean from part C, just compute $E(U^2) - (9/2)^2$.

X.

A) (10) Let Y have a gamma distribution with $\alpha = 3.2$, and $\beta = 1$. Derive a general formula for $E(Y^k)$, where $k \geq 1$ is an integer. (See note 1 in the directions!)

Solution:

$$E(Y^k) = \int_0^\infty y^k \cdot \frac{y^{2.2}e^{-y}}{\Gamma(3.2)}\ dy = \frac{\Gamma(3.2+k)}{\Gamma(3.2)}$$

B) (10) Let $Y_1, \ldots, Y_n$ be independent, all with gamma distributions with the same α, and β. Using moment-generating functions, determine the distribution of the mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

Solution: By independence,

$$m_{\overline{Y}}(t) = \prod_{i=1}^n m_{Y_i}(t/n) = \left(\frac{1}{(1-(\beta/n)t)^\alpha}\right)^n = \frac{1}{(1-(\beta/n)t)^{n\alpha}}$$

By the Uniqueness Theorem, $\overline{Y}$ also has a gamma distribution with parameters $n\alpha, \beta/n$.