Statistics 403 Exam 1

October 27, 2009

Instructions:

- You may not use notes, books, formula cards, etc.
- You may use a calculator, but only for arithmetic.
- If you are asked for a probability that cannot be calculated by hand, you can express your answer either in terms of normal probabilities (e.g. $P(Z < \ldots)$, $P(Z > \ldots)$), or in terms of R commands (e.g. `pnorm(...)`).
- If you need a quantile, define it as a symbol (e.g. $Q_{0.3}$) and give your answer in terms of the symbol.
- Every problem is worth 15 points.
- Partial credit will be given. Express you answers clearly and show work where appropriate.

1. Order the following three distributions in terms of their variances (i.e. state which has the smallest variance, which has the middle variance value, and which has the greatest variance).

a.

x	-1	0	1
$P(X = x)$	0.2	0.6	0.2

b.

x	-2	0	2
$P(X = x)$	0.3	0.4	0.3

c.

x	-1	0	1
$P(X = x)$	0.3	0.4	0.3

Solution: You can answer this question numerically – the variance of a, b, and c, respectively, are 0.4, 2.4, and 0.6, so b > c > a.

It's good to think about how these problems can be answered without actually computing the variance (in case you see a problem where it is too tedious to do the calculation).

All three distributions have mean zero, so the variances are determined by how far the squared values are likely to be from zero.

b and c have the same probabilities, but the two points in b are further from the zero than the points in c. Therefore the variance of b is greater than the variance of c.

a and c have the same sample space, but in a there is a higher probability of observing a zero, so the variance of c is greater than the variance of a.

2. Suppose two judges are rating music performances. The first judge rates with a mean of 8 and a standard deviation of 2. The second judge rates with a mean of 9 and a standard deviation of 3. The correlation between the two judges' scores is r. The score received by each performer is the sum of the two judges' scores.

   (a) What is the expected value of the score?

   Solution: If we let X and Y be the two scores, the expected values add, so it is $E(X + Y) = EX + EY = 8 + 9 = 17$.

   (b) If the correlation is $r = 0.5$, what is the variance of the score?

   Solution: The variances satisfy

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X, Y).$$

   The covariance is $0.5 \times 2 \times 3 = 3$, so the variance is $4 + 9 + 6 = 19$.

   (c) What values of r give the least and greatest variance for the score?

   Solution: The variance for a general value of r is $13 + 12r$. This is maximized when $r = 1$ and minimized when $r = -1$. We also will accept an answer stating that it is minimized when $r = 0$, in case you are thinking that correlations must be non-negative (this is not true in general, but I mentioned several times in class that correlation in these settings tend to be positive, so we will accept this answer).

3. (a) We collect independent data from a population that has expected value 3 and variance 4. Our goal is to construct a 95% confidence interval for EX that is 1/2 unit wide. What sample size is required?

   Solution: The width of the 95% CI is $4 \cdot \mathrm{SE} \approx 4 \cdot 2/\sqrt{n}$. If we set $8/\sqrt{n} = 1/2$ and solve for n we get $n = 256$.

   (b) Now suppose we collect the data sequentially $X_1, X_2, \ldots$, such that the correlation between any two consecutive values in the sequence is 0.5. If the expected value and variance are still 3 and 4, what sample size is required to get a confidence interval for EX that is 1/2 unit wide?

   Solution: The width of the 95% CI is still $4 \cdot \mathrm{SE}$, but the standard error is now different due to the correlation. For a sample of size n we have a covariance matrix with a total value of $4n$ on the diagonal. There are also $2(n - 1)$ nonzero values in the covariance matrix resulting from the covariances between adjacent pairs. Each of these covariances is $0.5 \cdot 2 \cdot 2 = 2$. Thus the total of the covariance matrix is $4n + 4(n-1) = 8n - 4$, so the variance of $\bar{X}$ is $(8n-4)/n^2$. Thus we need to solve

$$4\sqrt{8n-4}/n = 1/2,$$

or

$$n^2 - 512n + 256 = 0.$$

From the quadratic equation we have

$$\frac{512 \pm \sqrt{512^2 - 4 \cdot 256}}{2} = \frac{512 \pm 511}{2}.$$

Since $n = 1/2$ doesn't solve the original equation $4\sqrt{8n-4}/n = 1/2$, we must take the solution $(512 + 511)/2 \approx 512$.

4. Suppose two researchers have each developed a new treatment for a disease. Let A and B denote the the levels of symptoms in patients treated by the two researchers, respectively, and let Y denote the level of symptoms in an untreated patient.

   Suppose further that the two treatments are equally effective on average, so $EA = EB$. However, the standard deviations of the two treatments are not equal: specifically, $\mathrm{SD}(A) = 2 \times \mathrm{SD}(B)$; also, $\mathrm{SD}(B) = \mathrm{SD}(Y)$.

   The first researcher plans to treat 20 subjects, and to collect data for 10 controls. The second researcher plans to treat 10 subjects, and to collect data for 10 controls.

   State which researcher has greater power for detecting a beneficial treatment effect using a two sample comparison of means, or state that the powers are equal. Briefly explain your reasoning.

   Solution: Since $EA = EB$ the unstandardized effect sizes for the two researchers are the same. Thus the power will be maximized when the variance for estimating the mean difference is minimized. The mean difference can be estimated as $\bar{A} - \bar{Y}$ for the first researcher's study and $\bar{B} - \bar{Y}$ for the second researcher's study. Letting σ_A, σ_B, and σ_Y denote the three standard deviations, the variances of the mean differences are

$$\sigma_A^2/20 + \sigma_Y^2/10 = 4\sigma_Y^2/20 + \sigma_Y^2/10 = 3\sigma_Y^2/10.$$

   and

$$\sigma_B^2/10 + \sigma_Y^2/10 = \sigma_Y^2/10 + \sigma_Y^2/10 = \sigma_Y^2/5.$$

Thus the second researcher has the lower variance and hence the greater power.

5. (a) Is the correlation coefficient of the following data positive, negative or zero? Briefly explain your reasoning.

X	Y
25	20
5	30
10	30
0	40

Solution: The mean of the first column is 10 and the mean of the second column is 30. If we center the columns we get

$X - \bar{X}$	$Y - \bar{Y}$
15	-10
-5	0
0	0
-10	10

The signs are always opposite (or else one of the values is zero). Thus when X is greater than its mean, Y is less than or equal to its mean, and vice versa. This implies that $(X - \bar{X})(Y - \bar{Y}) \leq 0$ so the covariance is negative.

If you choose to treat this as a population rather than as a data set, the same argument can be used except you replace $\bar{X}$ with EX and $\bar{Y}$ with EY.

(b) Is the correlation coefficient of the following data greater than, less than, or equal to the correlation coefficient of the data shown in part a?

X	Y
75	20
15	30
30	30
0	40

Solution: The first column of this data set is 3 times the first column of the data set in part a, and the second column is the same as in part a. Let X^* and Y^* refer to the data in part b, to distinguish it from the data in part a. Then

$$\begin{aligned} \text{cor}(X^*, Y^*) &= \frac{\text{cov}(X^*, Y^*)}{\text{SD}(X^*)\text{SD}(Y^*)} \\ &= \frac{\text{cov}(3X, Y)}{\text{SD}(3X)\text{SD}(Y)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{3\text{cov}(X,Y)}{3\text{SD}(X),\text{SD}(Y)} \\
&= \frac{\text{cov}(X,Y)}{\text{SD}(X)\text{SD}(Y)} \\
&= \text{cor}(X,Y).
\end{aligned}
$$

Thus the correlation coefficients are the same for parts a and b.

6. Answer the questions below about the following cumulative distribution function.

x	0	1	2	3
$P(X \le x)$	0.2	0.3	0.5	1.0

(a) What is the probability distribution of X?

Solution:

x	0	1	2	3
$P(X = x)$	0.2	0.1	0.2	0.5

(b) What is $E(1/(X+1))$?

Solution:

$$0.2 \cdot \frac{1}{0+1} + 0.1 \cdot \frac{1}{1+1} + 0.2 \cdot \frac{1}{2+1} + 0.5 \cdot \frac{1}{3+1} \approx 0.44.$$

(c) What is $1/((EX)+1)$?

Solution:

$$EX = 0.2 \cdot 0 + 0.1 \cdot 1 + 0.2 \cdot 2 + 0.5 \cdot 3 = 2$$

so $1/(EX+1) = 1/3$.