Solutions Midterm BUEC 333: Version v

June 17, 2013, 14:30-16:20

Grading guidelines

- Give partial points if the answer is wrong, but part of the steps is correct. Or if there is some reasonable explanation. No need to make it more complicated than 0.5 points (round up: if you want to give somebody 0.32 credit for some explanation, just make it 0.5).
- Some answers are version-dependent. I give answers that depend on v, the version number.
- Total number of points = 20. This does not include the bonus question.
- Bonus question: this is multiple choice. 1 point for the right answer. Getting it right means selecting the right answer (A), even if it comes without an explanation. Give partial credit for the wrong choice, but a sensible explanation.

1 Probability

Consider the joint probability distribution represented in Table 1. For questions (1)-(6), give a numerical answer, and show how you obtained the answer. No explanation: no points.

		X_1	
		0	v
X_2	1	0.05	$0.v$
	2	0.05	$0.7 - 0.v$
	3	0.1	0.1

Table 1: Joint probability distribution.

1. Use the table to find $P(X_1 = v, X_2 = 1)$.
   1 point. For each version v, the answer is $0.v$.

2. Compute $P(X_1 = 0)$.
   1 point. 0.2.

3. Compute $P(X_2 = 2 | X_1 = v)$.
   1 point. Regardless of the version, note that $P(X_1 = v) = 0.8$. Then, $P(X_2 = 2 | X_1 = v) = \frac{P(X_2=2, X_1=v)}{P(X_1=v)} = \frac{0.7-0.v}{0.8} = \frac{7-v}{8}$. The answer depends on the version.

4. Compute $\mathrm{E}(X_2)$.
   1 point. First, get the marginal distribution of X_2. $P(X_2 = 1) = 0.v + 0.05$, $P(X_2 = 2) =$

$0.75 - 0.v$, *and* $P(X_2 = 3) = 0.2$. *Then, the correct answer is*

$$\begin{aligned} E(X_2) &= 1 * P(X_2 = 1) + 2 * P(X_2 = 2) + 3 * P(X_2 = 3) \\ &= 0.v + 0.05 + 1.5 - 0.v - 0.v + 3 * 0.2 \\ &= 2.15 - 0.v \end{aligned}$$

5. Compute $\text{var}(X_1)$.
   2 points, one for getting $E(X_1)$ *and one for then getting the variance.* $E(X_1) = 0 * P(X_1 = 0) + v * 0.8 = 0.8 * v$. *For the variance,*

$$\begin{aligned} var(X_1) &= P(X_1 = 0) * (0 - 0.8 * v)^2 + P(X_1 = v)(v - 0.8v)^2 \\ &= 0.2 * 0.64 * v^2 + 0.8 * 0.04 * v^2 \\ &= 0.16 * v^2. \end{aligned}$$

6. Compute $\text{E}(X_2 | X_1 = 0)$.
   1 point. Need to derive the conditional distribution first, which is $\{1/4, 1/4, 1/2\}$. *Then the answer follows from* $1 * 1/4 + 2 * 1/4 + 3 * 1/2 = 2.25$

7. Are X_1 and X_2 independent? Explain your answer.
   1 point. There are several ways to show this, all of which should get full points. Accept other answers only if they really make sense, but it is hard for me to see correct, alternative answers. Do not give points for an answer without an explanation. The three ways of showing this: (i) They show that the covariance or correlation is not zero. This implies dependence. (ii) They construct a counterexample based on the definition of independence, $P(X_1 = a, X_2 = b) = P(X_1 = a) P(X_2 = b)$. *(iii) They show that one of the conditional expectations or variance depends on the value of the RV that you are conditioning on. For example,* $E(X_2 | X_1 = 0) \neq E(X_2 | X_1 = v)$.

2 Statistics

Now, consider a random variable X with $\text{E}(X) = \mu_X$ and $\text{var}(X) = \sigma_X^2$. To estimate the parameter μ_X, you have available the random sample

$$\{X_1, \cdots, X_{10}\} = \{v, 4, 5, 6, 3, 2, 7, 8, 3, v\}.$$

1. Compute the sample mean $\bar{X}$.
   1 point. $\bar{x} = \frac{2v+38}{10} = 3.8 + \frac{v}{5} = 4 + (v - 1) * 0.2$. *For* $v = 1$, *it is* 4, *for* $v = 2$, *it is* 4.2, *etc.*

2. Is the sample mean a random variable? Is it unbiased? Is it consistent?
   1 point. Yes, yes, yes.

3. Compute the sample variance s_X^2.
   1 point. Formula is complicated due to the presence of v, *but you can lookup the answers in the following R output.*

```
> for(v in seq(1,6)) {
+   X <- c(v,4,5,6,3,2,7,8,3,v)
+    print(c(v,mean(X),var(X)))
```

```
+ }

[1] 1 4 6
[1] 2 4.200000 4.844444
[1] 3 4.400000 4.044444
[1] 4 4.6 3.6
[1] 5 4.800000 3.511111
[1] 6 5.000000 3.777778
```

4. In general, is it true that $s_X^2 = \sigma_X^2$? If yes: why? If no: what **can** you say about the relationship between s_X^2 and σ_X^2?
   2 points. One for knowing that it is not true. One for the additional remark that $E\left(s_X^2\right) = \sigma_X^2$ or the equivalent statement in words ("the sample variance is unbiased for the population variance").

5. Assume that the number of observations in the random sample above, $\{X_1, \cdots, X_{10}\}$, is "very, very large". Construct a 95% confidence interval for μ_X.
   1 point. They should know the critical value by hard (the one for 95%). Read the answer from the R output below. For each row, the first item is v, the second one is the lower bound of the 95%-CI, and the third one is the upper bound.

```
> for(v in seq(1,6)) {
+   X <- c(v,4,5,6,3,2,7,8,3,v)
+   n <- length(X)
+   xbar <- mean(X)
+   s <- sqrt(var(X))
+   z.alpha <- 1.96
+   lb <- xbar - z.alpha*s/sqrt(n)
+   ub <- xbar + z.alpha*s/sqrt(n)
+   print(c(v,lb,ub),digits=3)
+ }

[1] 1.00 2.48 5.52
[1] 2.00 2.84 5.56
[1] 3.00 3.15 5.65
[1] 4.00 3.42 5.78
[1] 5.00 3.64 5.96
[1] 6.0 3.8 6.2
```

6. Why do we have to make this assumption ("very, very large")?
   1 point. So that we can use the central limit theorem + a little more explanation.

3 Regression analysis

1. In

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

what is the interpretation of β_1? Be precise.
1 point for the correct answer, which would be something about the effect of a change in X_i of one unit, on the change in Y_i. MAKE SURE that they mention the "everything else equal" stuff.

2. What is the difference between a residual and an error term? Draw a picture to illustrate your answer.
2 points. 1 for the explanation, and 1 for the picture. The picture should have a population regression function (PRF) and a sample regression function (SRF). The error term is the distance to the PRF, the residual is the distance to the SRF. The error term is a population concept; the residual is a sample concept.

3. What is the smallest possible value the R^2 can take? What is the largest value? What does it mean when the $R^2 = 0.v$?
2 points. 1 for smallest=0 and largest=1. The other point for the right answer to the third question (see book).

Bonus question (multiple choice)

From Kahnemann, D. and Tversky, A. (1974), Judgement under Uncertainty: Heuristics and Biases, Science 185 (4157), p. 1124-1131.

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?
>
> **A.** the smaller hospital
>
> **B.** the bigger hospital
>
> **C.** about the same (within 5 percent of each other)

The correct answer is A. For each baby, the probability of it being a boy/girl is 0.5. The RV that denotes whether it is a girl ($Y = 1$) or a boy ($Y = 0$) is therefore Bernoulli with mean 0.5 and variance 0.25. Let X_1 be the proportion of girls in the small hospital, and X_2 the proportion of girls in the larger hospital. It is easy to see that $\frac{X_1 = \sum_{j=1}^{15} Y_i}{15}$ where Y_j denotes the jth baby that is born. $E(X_1) = \frac{15*0.5}{15} = 0.5$ and $var(X_1) = \frac{0.25}{15}$ whereas for the larger hospital, we have $E(X_2) = 0.5$ and $var(X_2) = \frac{0.25}{45}$. So we have two random variable, X_1 and X_2, which have the same mean, but a different variance. The RV with the largest variance has less mass in the center, and so it has more mass away from the center. We are interested in $P(X_k > 0.6)$, which is a question about mass away from the center.