W271 Lab 1

Brandon Shurick

May 2016

1 Question 1

Question ignored, as per instructions.

2 Question 2

Suppose for events A and B, $Pr(A) = p \leq 1/2$, $Pr(B) = q$, where $\frac{1}{4} < q < \frac{1}{2}$. These are the only information we have about the events.

1. What are the maximum and minimum possible values for $Pr(A \cup B)$?

   The maximum possible value for a union of two probabilities would be the sum of the largest possible probabilities, less the smallest possible combined probability of both events, in this case, where $p = 1/2$ and $q \approx 1/2$, and $p \cap q = 0$. So, $P(A) + P(B) - P(A \cap B) \approx 1$.

   The minimum possible value for a union of two probabilities would be the sum of the smallest possible probabilities, less the largest possible combined probability of both events, in this case, where $p = 0$ and $q \approx 1/4$, and $p \cap q = 0$. So, $P(A) + P(B) - P(A \cap B) \approx 1/4$.

2. What are the maximum and minimum possible values for $Pr(A|B)$?

   Given Bayes's formula, $P(A|B) = \frac{P(A)*P(B|A)}{P(B)}$, the maximum values for $P(A)$, $P(B)$, and $P(B|A)$ would be $P(A) = 1/2$, $P(B) \approx 1/2$, and $P(B|A) \approx 1$. Given the formula, $P(A|B) = \frac{1/2*1}{1/2} \approx 1$.

   The minimum values for $P(A)$, $P(B)$, and $P(B|A)$ would be $P(A) = 0$, $P(B) \approx 1/4$, and $P(B|A) \approx 1/4$. Given the formula, $P(A|B) = \frac{0*1/4}{1/4} = 0$.

3 Question 3

For a randomly selected county in the United States, let X represent the proportion of adults over age 65 who are employed, or the elderly employment rate.

Then, X is restricted to a value between zero and one. Suppose that the cumulative distribution function for X is given by $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. Find the probability that the elderly employment rate is at least .6 (60%).

The function $F(x) = 3x^2 - 2x^3 = -x^2(2x-3)$ has a definite integral $\int_{0.6}^{1} 3x^2 - 2x^3 = 0.5 - 0.1512 = 0.3488$.

4 Question 4

Just prior to jury selection for O. J. Simpson's murder trial in 1995, a poll found that about 20% of the adult population believed Simpson was innocent (after much of the physical evidence in the case had been revealed to the public). Ignore the fact that this 20% is an estimate based on a sub-sample from the population; for illustration, take it as the true percentage of people who thought Simpson was innocent prior to jury selection. Assume that the 12 jurors were selected randomly and independently from the population (although this turned out not to be true).

1. Find the probability that the jury had at least one member who believed in Simpson's innocence prior to jury selection. [Hint: Define the Binomial(12,.20) random variable X to be the number of jurors believing in Simpson's innocence.]

   Using the density function for $Binomial(12, 0.2)$, the probability of having 0 jurors believe in Simpson's innocence prior to jury selection is 6.9%, so the probability of 1 or more jurors is 100%-6.9% or 93.1%.

2. Find the probability that the jury had at least two members who believed in Simpson's innocence.[Hint:$P(X \geq 2) = 1 - P(X \leq 1)$, and $P(X \leq 1) = P(X = 0) + P(X = 1)$.]

   Using the density function for $Binomial(12, 0.2)$, the probability of having 0 jurors believe in Simpson's innocence is 6.9%, and the probability of having 1 jurors believe in Simpson's innocence is 20.6%, and so combined $P(X \leq 1) = 6.9\% + 20.6\% = 27.5\%$. Thus, $P(X \geq 2) = 1 - 27.5\% = 72.5\%$.

5 Question 5

(Requires calculus) Let X denote the prison sentence, in years, for people convicted of auto theft in a particular state in the United States. Suppose that the pdf of X is given by $f(x) = (1/9)x^2$, $0 < x < 3$. Use integration to find the expected prison sentence.

$\int_0^3 (1/9)x^2 * x = (1/36)(3)^4 - (1/36)(0)^4 = 2.25$

6 Question 6

If a basketball player is a 74% free throw shooter, then, on average, how many free throws will he or she make in a game with eight free throw attempts?

The expected value of $Binomial(8, 0.74) = 8*0.74 = 5.92$ or ≈ 6 free throws.

7 Question 7

You are hired by the governor to study whether a tax on liquor has decreased average liquor consumption in your state. You are able to obtain, for a sample of individuals selected at random, the difference in liquor consumption (in ounces) for the years before and after the tax. For person i who is sampled randomly from the population, Y_i denotes the change in liquor consumption. Treat these as a random sample from a $Normal(\mu, \sigma)$ distribution.

1. The null hypothesis is that there was no change in average liquor consumption. State this formally in terms of μ.

   The null hypothesis would be stated as $H_0 : \mu = 0$, in which case there was no change from before to after the tax was introduced.

2. The alternative is that there was a decline in liquor consumption; state the alternative in terms of μ.

   The alternative hypothesis would be stated as $H_1 : \mu < 0$, in which case the mean change in liquor consumption decreased in the sample population from before to after the tax was introduced.

3. Now, suppose your sample size is $n = 900$ and you obtain the estimates $\bar{y} = -32.8$ and $s = 466.4$. Calculate the t statistic for testing H_0 against H_1; obtain the p-value for the test. (Because of the large sample size, just use the standard normal distribution tabulated in Table G.1.) Do you reject H_0 at the 5% level? At the 1% level?

   The t statistic in this case is $t = \sqrt{n} * (\bar{y} - \mu_0)/s = \sqrt{900} * (-32.8 - 0)/466.4 = -2.11$, with $p - value = 0.0143$. H_0 would be rejected at a 5% level, but not at a 1% level.

4. Would you say that the estimated fall in consumption is large in magnitude? Comment on the practical versus statistical significance of this estimate.

   I would say that a mean drop in consumption of 32.8 ounces is not a large magnitude. This would translate to a mean drop in consumption of one-bottle-and-a-half per year. Although it is not clear what the average amount of liquor a consumer typically purchases, it seems as though this is not a practically significant change in purchasing behavior, although it is a statistically significant change.

5. What has been implicitly assumed in your analysis about other determinants of liquor consumption over the two-year period in order to infer causality from the tax change to liquor consumption?

   It must be assumed that no determinant of liquor consumption over the two-year period had any effect on liquor consumption, other than the tax change.

8 Question 8

The New York Times (2/5/90) reported three-point shooting performance for the top 10 three-point shooters in the NBA. The following table summarizes these data:

Player	FGA-FGM
Mark Price	429-188
Trent Tucker	833-345
Dale Ellis	1,149-472
Craig Hodges	1,016-396
Danny Ainge	1,051-406
Byron Scott	676-260
Reggie Miller	416-159
Larry Bird	1,206-455
Jon Sundvold	440-166
Brian Taylor	417-157

Note: FGA = field goals attempted and FGM = field goals made.

For a given player, the outcome of a particular shot can be modeled as a Bernoulli (zero-one) variable: if Y_i is the outcome of shot i,then $Y_i = 1$ if the shot is made, and $Y_i = 0$ if the shot is missed. Let θ denote the probability of making any particular three-point shot attempt. The natural estimator of θ is $\bar{Y} = FGM/FGA$.

1. Estimate θ for Mark Price.

   $\theta = FGM/FGA = 188/429 \approx 0.4382$

2. Find the standard deviation of the estimator $\bar{Y}$ in terms of θ and the number of shot attempts, n. The standard deviation of $\theta = 0.0206$; the standard deviation of $FGA = 326.71$

3. The asymptotic distribution of $(\bar{Y} - \theta)/se(\bar{Y})$ is standard normal, where $se(\bar{Y}) = \sqrt{\bar{Y}(1 - \bar{Y})/n}$. Use this fact to test $H_0 : \theta = .5$ against $H_1 : \theta < .5$ for Mark Price. Use a 1% significance level.

   In this case, $(\bar{Y} - \theta)/se(\bar{Y}) = -7.96$, which equates to a p-value of $9e - 16$ and is well below the 1% significance threshold.

9 Question 9

Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. A human rights group suspects foul play and hires you to test the validity of the dictator's claim. You have a budget that allows you to randomly sample 200 voters from the country.

1. Let X be the number of yes votes obtained from a random sample of 200 out of the entire voting population. What is the expected value of X if, in fact, 65% of all voters supported the dictator?

   The expected value of X would be 65% of 200, or 130.

2. What is the standard deviation of X, again assuming that the true fraction voting yes in the plebiscite is .65?

   The standard deviation of 130 yes's (yes = 1) and 70 no's (no = 0) is 0.478167.

3. Now, you collect your sample of 200, and you find that 115 people actually voted yes. Use the CLT to approximate the probability that you would find 115 or fewer yes votes from a random sample of 200 if, in fact, 65% of the entire population voted yes.

   The Z-score would be $Z = (\bar{y} - \mu_0)/(S_n/\sqrt{n}$,

   or $(0.575 - 0.65)/(0.4781665/\sqrt{200}) = -2.218182$ which equates to a p-value of 0.0138, meaning that there is approximately a 1.4% probability of observing 115 or fewer votes if, in fact, the true percentage of yes votes for the population was 65%.

4. How would you explain the relevance of the number in part (iii) to someone who does not have training in statistics?

   The p-value represents how unusual our data is, and is used in statistical tests to give reasonable evidence that we should reject the null hypothesis.

10 Question 10

Before a strike prematurely ended the 1994 major league baseball season, Tony Gwynn of the San Diego Padres had 165 hits in 419 at bats, for a .394 batting average. There was discussion about whether Gwynn was a potential .400 hitter that year. This issue can be couched in terms of Gwynn's probability of getting a hit on a particular at bat, call it θ. Let Y_i be the $Bernoulli(\theta)$ indicator equal to unity if Gwynn gets a hit during his ith at bat, and zero otherwise. Then, $Y_1, Y_2, ..., Y_n$ is a random sample from a $Bernoulli(\theta)$ distribution, where θ is the probability of success, and n = 419.

Our best point estimate of θ is Gwynn's batting average, which is just the proportion of successes: $\bar{y} = .394$. Using the fact that $se(\bar{y}) = \sqrt{\bar{y}(1-\bar{y})/n}$, construct an approximate 95% confidence interval for u, using the standard normal distribution. Would you say there is strong evidence against Gwynn's being a potential .400 hitter? Explain.

A normalized value for Gwinn's batting average would be equal to $(\bar{y} - \mu_0)/se(\bar{y})$, or $(0.394 - 0.4)/(\sqrt{0.394 * (1 - 0.394)/419}) = -0.2513$. This z-score would equate to a p-value of 0.40, which is not significant, which means that there is not enough evidence to reject the null hypothesis that Gwinn's true batting average is 0.400.