STA 2101/442 Assignment Two[1]

Please bring your R printouts to the quiz on Friday Sept. 26th. The non-computer parts are just practice for the quiz, and are not to be handed in.

1. A medical researcher conducts a study using twenty-seven litters of cancer-prone mice. Two members are randomly selected from each litter, and all mice are subjected to daily doses of cigarette smoke. For each pair of mice, one is randomly assigned to Drug A and one to Drug B. Time (in weeks) until the first clinical sign of cancer is recorded. State a reasonable model for these data. Remember, a statistical model is a set of assertions that partly specify the probability distribution of the observable data. For simplicity, you may assume that the study continues until all the mice get cancer, and that log time until cancer has a normal distribution.

2. Suppose that volunteer patients undergoing elective surgery at a large hospital are randomly assigned to one of three different pain killing drugs, and one week after surgery they rate the amount of pain they have experienced on a scale from zero (no pain) to 100 (extreme pain). State a reasonable model for these data. For simplicity, you may assume normality.

3. In a risky type of brain surgery, seventy-five percent of patients survive for at least 24 hours after the surgery. But at a hospital that usually achieves this success rate, 15 out of the last 30 patients have died. Could this be due to chance?

   (a) State a reasonable model for these data. For simplicity, assume that the last 30 patients to undergo the surgery form a simple independent random sample.

   (b) What is the parameter space?

   (c) Without any derivation, estimate the parameter in your model. Your answer is a number.

   (d) Give an approximate 95% confidence interval for the (recent) probability of survival. Your answer is a set of two numbers.

   (e) What is the null hypothesis corresponding to the *main question*, in symbols?

   (f) What is the critical value (or values) of the test statistic at $\alpha = 0.05$ for a 2-sided test? The answer is a number or a pair of numbers.

   (g) Calculate a reasonable test statistic. Your answer is a number. Show some work.

      i. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

      ii. Using R, calculate the p-value. Make sure it's on the printout you bring to the quiz.

      iii. Do the data provide convincing evidence against the null hypothesis?

      iv. In plain, non-statistical language, what do you conclude? Your answer is a statement about surviving this surgery.

4. A polling firm plans to ask a random sample of registered voters in Quebec whether Quebec should separate from Canada and become an independent nation: Yes or No. They would like to be able to say that their results are expected to be accurate within three percentage points, nineteen times out of twenty.

[1] Copyright information is at the end of the last page.

(a) Suppose the population percent favouring independence is 25%. What sample size is required to achieve the desired margin of error?

(b) Suppose the population percent favouring independence is 40%. What sample size is required to achieve the desired margin of error?

(c) What sample size would be required if you were unwilling to make any assumptions about the true percentage favouring independence?

5. For years, brand awareness for Big Red chewing gum has been stuck at about 6%, meaning that about 6% of consumers who chew gum say they remember hearing about Big Red gum. The gum company is planning an advertising campaign to increase brand awareness, in the hope that increased brand awareness will lead to increased sales.

   The advertising agency has a problem. With the budget they have been given to purchase media (air time and so on), they are confident they can move brand awareness a little – perhaps to 8%. In the old days, they could tell the client they had increased awareness by 33% and start to celebrate, but now the client has fallen under the influence of a U of T graduate who insists that a null hypothesis be rejected at the $\alpha = 0.05$ level with a non-directional test before they admit that anything actually worked. A market research analyst from the adversing agency took a market research analyst from the gum company out to lunch, and they agreed on the test statistic

$$Z = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}.$$

   Now, the advertising agency has to decide how many people they need to survey when they measure brand awareness, in order to have a good chance of rejecting the null hypothesis. It's important, because if the client thinks the advertising didn't work, they might get a new advertising agency. On the other hand, they also don't want to survey more people than necessary, because that's expensive.

   Suppose they want to be 90% sure of rejecting H_0 if they manage to increase brand awareness to 8%. What sample size do they need? I will start you out. You want the smallest (integer) sample size so that $Pr\{|Z| > 1.96\} \geq 0.90$. Here are some points to consider.

   - The null hypothesis, of course, is $\theta = \theta_0$. What is θ_0? The answer is a specific number in this problems.
   - Power is being calculated under the assumption a true parameter value of $\theta = 0.08$.
   - When I calculate the probability indicated above (power), I get an expression in n, and my answer emerges in terms of Φ, the cumulative distribution function of a standard normal. That is, $\Phi(0) = \frac{1}{2}$, and so on. Φ is exactly R's `pnorm` function.

   Again, what sample size is required? I suggest you use R to calculate power for different values of n, until you find the smallest n that makes the power at least 0.90. Please do the paper and pencil calculations, and then obtain the answer using R and bring your printout to the quiz.

6. Recall the chimney vent damper example, in which X_i was energy consumption with vent damper active and Y_i was was energy consumption with vent damper inactive, independently for $i = 1, \ldots, n$. All the parts of this question are related to that example.

   (a) Suppose $E(X_i) = \mu_1$, $E(Y_i) = \mu_2$, $Var(X_i) = \sigma_1^2$, $Var(Y_i) = \sigma_2^2$, and $Cov(X_i, Y_i) = \sigma_{12}$. The *correlation* between two random variables X and Y is defined by $Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)\,SD(Y)}$, so that $Corr(X_i, Y_i) = \frac{Cov(X,Y)}{\sigma_1\sigma_2}$. What is $Corr(\overline{X}, \overline{Y})$? Show your work.

   (b) What is $E(\overline{X} - \overline{Y})$?

   (c) What is $Var(\overline{X} - \overline{Y})$? Show your work.

   (d) For 90 independent standard normal random variables, what is the probability of getting at least one larger than 5 in absolute value? Calculate the value with R.

7. Here is a more specific model for the vent damper data (it was given in lecture, except that this time there is no assumption of normality).

   Independently for $i = 1, \ldots, n$, we observe (X_i, Y_i) where

$$\begin{aligned} X_i &= \delta + Z_i + \epsilon_{i1} \\ Y_i &= \phantom{\delta +{}} Z_i + \epsilon_{i2} \end{aligned}$$

   with

   - δ the effect of having the vent damper active (a constant.)
   - $E(Z_i) = \mu_z$, $Var(Z_i) = \sigma_z^2$
   - $E(\epsilon_{i1}) = E(\epsilon_{i2}) = 0$,
   - $Var(\epsilon_{i1}) = \sigma_1^2$, $Var(\epsilon_{i2}) = \sigma_2^2$
   - Z_i and ϵ_{ij} all independent.

   Here are two questions:

   (a) Calculate $Cov(X_i, Y_i)$. Why does this tell you that the covariance cannot be negative?

   (b) Under the assumption (null hypothesis) that the vent damper had absolutely no effect, we would have $\delta = 0$, and also the distributions of all the ϵ_{ij} would be identical. Under these circumstances,

      i. Should the distributions of $\overline{X}$ and $\overline{Y}$ be the same?

      ii. Should $\overline{X}$ and $\overline{Y}$ be independent? Why or why not?

      iii. Assuming that all the distributions are continuous, try to show that $Pr\{\overline{X} > \overline{Y}\} = \frac{1}{2}$. Did it work? If it worked, we can justify a sign test.
