STA 2101/442 Assignment Two[1]

Please bring your R printouts to the quiz on Friday Sept. 27th. The non-computer parts are just practice for the quiz, and are not to be handed in.

For some of the convergence questions, the following may be helpful. A *degenerate* random variable is one that equals a particular constant with probability one. That is, $P(X = a) = 1$. Convergent sequences of constants may be viewed as sequences of degenerate random variables that converge almost surely, and hence in probability. This allows you to use the Slutsky lemmas when some of the sequences are sequences of constants.

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.

   (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.

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

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.

   (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.

   (a) 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? Please obtain the answer using R and bring your printout to the quiz.

   (b) Please check your calculation with a simulation. Specifically, use the required sample size you found, and estimate the probability of rejecting H_0. Use a Monte Carlo sample size of 10,000. Bring your printout to the quiz.

6. This is about how to simulate from a continuous univariate distribution. Let the random variable X have a continuous distribution with density $f_X(x)$ and cumulative distribution function $F_X(x)$. Suppose the cumulative distribution function is strictly increasing over the set of x values where $0 < F_X(x) < 1$, so that $F_X(x)$ has an inverse. Let U have a uniform distribution over the interval $(0, 1)$. Show that the random variable $Y = F_X^{-1}(U)$ has the same distribution as X. Hint: You will need an expression for $F_U(u) = Pr\{U \leq u\}$, where $0 \leq u \leq 1$.

7. Let X be an arbitrary random variable. Show $\frac{X}{n} \xrightarrow{p} 0$. Please use Slutsky lemmas rather than definitions, though the proof using the definition of convergence in probability is fairly short.

8. Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} T$. Show $T_n \xrightarrow{p} \theta$. Please use Slutsky lemmas rather than definitions.

9. Let $X_1, \ldots, X_n$ be a random sample from a Binomial distribution with parameters 3 and θ. That is,
$$P(X_i = x_i) = \binom{3}{x_i} \theta^{x_i} (1-\theta)^{3-x_i},$$
for $x_i = 0, 1, 2, 3$. Find the maximum likelihood estimator of θ, and show that it is strongly consistent.

10. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density
$$f(x; \tau) = \frac{\tau^{1/2}}{\sqrt{2\pi}} e^{-\frac{\tau x^2}{2}},$$
where the parameter $\tau > 0$. Let
$$\widehat{\tau} = \frac{n}{\sum_{i=1}^n X_i^2}.$$
Is $\widehat{\tau}$ a consistent estimator of τ? Answer Yes or No and prove your answer. Hint: You can just write down $E(X^2)$ by inspection. This is a very familiar distribution.

11. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Prove that the sample variance $S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$ is a strongly consistent estimator of σ^2.

12. Independently for $i = 1, \ldots, n$, let
$$Y_i = \beta X_i + \epsilon_i,$$
where $E(X_i) = E(\epsilon_i) = 0$, $Var(X_i) = \sigma_X^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i. Let
$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$
Is $\widehat{\beta}$ a consistent estimator of β? Answer Yes or No and prove your answer.

13. In this problem, you'll use (without proof) the *variance rule*, which says that if θ is a real constant and $T_1, T_2, \ldots$ is a sequence of random variables with

$$\lim_{n\to\infty} E(T_n) = \theta \text{ and } \lim_{n\to\infty} Var(T_n) = 0,$$

then $T_n \xrightarrow{P} \theta$.

In Problem 12, the independent variables are random. Here they are fixed constants, which is more standard (though a little strange if you think about it). Accordingly, let

$$Y_i = \beta x_i + \epsilon_i$$

for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are a random sample from a distribution with expected value zero and variance σ^2, and β and σ^2 are unknown constants.

(a) What is $E(Y_i)$?

(b) What is $Var(Y_i)$?

(c) Find the Least Squares estimate of β by minimizing $Q = \sum_{i=1}^{n}(Y_i - \beta x_i)^2$ over all values of β. Let $\widehat{\beta}_n$ denote the point at which Q is minimal.

(d) Is $\widehat{\beta}_n$ unbiased? Answer Yes or No and show your work.

(e) Give a sufficient condition for $\widehat{\beta}_n$ to be consistent. Show your work. Remember, in this model the x_i are fixed constants, not random variables.

(f) Let $\widehat{\beta}_{2,n} = \frac{\overline{Y}_n}{\overline{x}_n}$. Is $\widehat{\beta}_{2,n}$ unbiased? Consistent? Answer Yes or No to each question and show your work.

(g) Prove that $\widehat{\beta}_n$ is a more accurate estimator than $\widehat{\beta}_{2,n}$ in the sense that it has smaller variance. Hint: The sample variance of the independent variable values cannot be negative.

14. Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = \beta = \theta > 0$. That is, the density is

$$f(x;\theta) = \frac{1}{\theta^\theta \Gamma(\theta)} e^{-x/\theta} x^{\theta-1},$$

for $x > 0$. Let $\widehat{\theta} = \overline{X}_n$. Is $\widehat{\theta}$ a consistent estimator of θ? Answer Yes or No and prove your answer.
