STA 2101/442 Assignment 2[1]

These questions are practice for the midterm and final exam, and are not to be handed in.

1. 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?

2. 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, junk email, pop-up ads 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. My answer is 1653.

3. Suppose that several studies have tested the same hypothesis or similar hypotheses. For example, six studies could all be testing the effect of a treatment for arthritis. Two studies rejected the null hypothesis, one was almost significant, and the other three were clearly non-significant. What should be concluded? Ideally, one would pool the data from the six studies, but in practice the raw data are not available. All we have are the published test statistics. How do we combine the information we have and come to an overall conclusion? That is the main task of *meta-analysis*. In this question you will develop some simple, standard tools for meta-analysis.

   (a) Let the test statistic T be continuous, with pdf $f(t)$ and cdf $F(t)$ under the null hypothesis. The null hypothesis is rejected if $T > c$. Show that if H_0 is true, the distribution of the p-value is $U(0,1)$. Derive the density. Start with the cumulative distribution function of the p-value: $Pr\{P \leq x\} = \ldots$.

   (b) Suppose H_0 is false. Would you expect the distribution of the p-value to still be uniform? Pick one of the alternatives below. You are not asked to derive anything for now.

      i. The distribution should still be uniform.
      ii. We would expect more small p-values.
      iii. We would expect more large p-values.

   (c) Let $P_i \sim U(0,1)$. Show that $Y_i = -2\ln(P_i)$ has a χ^2 distribution. What are the degrees of freedom?

   (d) Let $P_1, \ldots P_n$ be a random sample of p-values with the null hypotheses all true, and let $Y = \sum_{i=1}^{n} -2\ln(P_i)$. What is the distribution of Y? Only derive it (using moment-generating functions) if you don't know the answer.

   (e) Let $P_i \sim U(0,1)$, and denote the cumulative distribution function of the standard normal by $\Phi(x)$.

i. What is the distribution of $Y_i = \Phi^{-1}(1 - P_i)$? Show your work.

ii. If H_0 is false and P_i is not uniform, would you expect Y_i to be bigger, or smaller? Why?

(f) Let $P_1, \ldots P_n$ be a random sample of p-values.

i. Propose a test statistic based on your answer to Question 3(e)i.

ii. What is the null hypothesis of your test?

iii. What is the distribution of your test statistic under the null hypothesis? Only derive it (using moment-generating functions) if you don't know the answer.

iv. Would you reject the null hypotesis when your test statistic has big values, or when it has small values? Which one?

(g) Suppose we observe the following random sample of p-values: `0.016 0.188 0.638 0.148 0.917 0.124 0.695`.

i. For the test statistic of Question 3d,

A. What is the critical value at $\alpha = 0.05$? The answer is a number.

B. What is the value of the test statistic? The answer is a number.

C. Do you reject the null hypothesis? Yes or No.

D. What if anything do you conclude?

ii. For the test statistic of Question 3(e)i,

A. What is the critical value at $\alpha = 0.05$? The answer is a number.

B. What is the value of the test statistic? The answer is a number.

C. Do you reject the null hypothesis? Yes or No.

D. What if anything do you conclude?

4. Suppose $\sqrt{n}(T_n - \theta) \xrightarrow{d} T$. Show $T_n \xrightarrow{p} \theta$. Please use Slutsky lemmas rather than definitions. Hint: Think of the sequence of constants $\frac{1}{\sqrt{n}}$ as a sequence of degenerate random variables (variance zero) that converge almost surely and hence in probability to zero. Now you can use a Slutsky lemma.

5. 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.

6. 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.

7. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ. Show that $T_n = \frac{1}{n+400}\sum_{i=1}^{n} X_i$ is a strongly consistent estimator of μ.

8. 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.

9. 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

$$\widehat{\beta}_n = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}.$$

Is $\widehat{\beta}_n$ a consistent estimator of β? Answer Yes or No and prove your answer.

10. 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 9, 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) Use the same estimator as in Problem 9. Is $\widehat{\beta}_n$ unbiased? Answer Yes or No and show your work.

(d) Suppose that the sequence of constants $\sum_{i=1}^{n} x_i^2 \to \infty$ as $n \to \infty$. Does this guarantee $\widehat{\beta}_n$ will be consistent? Answer Yes or No. Show your work.

(e) 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. Do you need a condition on the x_i values ?

(f) 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 explanatory variable values cannot be negative.

11. Let X be a random variable with expected value μ and variance σ^2. Show $\frac{X}{n} \xrightarrow{p} 0$.

12. 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.

13. Here is an integral you cannot do in closed form, and numerical integration is challenging. For example, R's `integrate` function fails.

$$\int_0^{1/2} e^{\cos(1/x)}\, dx$$

Using R, approximate the integral with Monte Carlo integration, and give a 99% confidence interval for your answer. You need to produce 3 numbers: the estimate, a lower confidence limit and an upper confidence limit.