

## STA 2101/442 Assignment 2<sup>1</sup>

Questions 4, 5 and 6g are to be done with R; please bring your printouts to the quiz and be prepared to hand them in if requested. The other questions are practice for the quiz on Friday Oct. 7, 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.
  - (a) 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.
  - (b) What is the parameter space for your model?
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).
  - (a) State a reasonable model for these data. For simplicity, you may assume normality.
  - (b) What is the parameter space?
3. 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?
4. 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 advertising agency took a market research analyst from the gum company out to lunch, and they agreed on the test statistic

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

---

<sup>1</sup>This assignment was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a [Creative Commons Attribution - ShareAlike 3.0 Unported License](https://creativecommons.org/licenses/by-sa/3.0/). Use any part of it as you like and share the result freely. The L<sup>A</sup>T<sub>E</sub>X source code is available from the course website: <http://www.utstat.toronto.edu/~brunner/oldclass/appliedf16>

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  $\theta_0$ ? 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  $\Phi$ , the cumulative distribution function of a standard normal. That is,  $\Phi(0) = \frac{1}{2}$ , and so on.  $\Phi$  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.

5. Let  $X_1, \dots, X_{n_1}$  be a random sample from a possibly non-normal distribution with expected value  $\mu_1$  and variance  $\sigma_1^2$ . Independently of  $X_1, \dots, X_{n_1}$ , let  $Y_1, \dots, Y_{n_2}$  be a random sample from a possibly non-normal distribution with expected value  $\mu_2$  and variance  $\sigma_2^2$ . The interest is in testing  $H_0 : \mu_1 = \mu_2$ . This could be a test for difference between a treatment and control group. The usual assumptions of normality and equal variance are unbelievable, so an independent  $t$ -test is hard to justify. On the other hand, collecting a lot of data is not too expensive; it's easy to get  $n_1$  and  $n_2$  both above 25, so that the Central Limit Theorem applies. The question is, how big do the sample sizes need to be in order to obtain good power?

The significance level will be  $\alpha = 0.05$ , and the test statistic will be

$$Z^2 = \left( \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \right)^2 \approx \left( \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)^2,$$

where it is assumed that the sample sizes are large enough so that sample variances are close approximations of population variances. The null hypothesis will be rejected when  $Z^2$  is greater than some critical value.

- (a) What is the approximate distribution of  $Z$  under the null hypothesis? Just write down the answer.
- (b) What is the approximate distribution of  $Z^2$  under the null hypothesis? Just write down the answer.
- (c) What is the critical value? The answer is a number.
- (d) When  $\mu_1 \neq \mu_2$ , the test statistic  $Z^2$  has an approximate non-central chi-squared distribution. What is the non-centrality parameter? Show your work.
- (e) Suppose that  $n_1 = n_2 = n$  while  $\sigma_1^2 = 144$  and  $\sigma_2^2 = 225$ . If the true difference between  $\mu_1$  and  $\mu_2$  is 5,
  - i. What is the power of the test for  $n = 50$ ?
  - ii. What is the power of the test for  $n = 500$ ?
  - iii. What is the minimum  $n$  required for a power of 0.90?

6. 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:  $Pr\{P \leq x\} = \dots$
- (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.
- The distribution should still be uniform.
  - We would expect more small  $p$ -values.
  - 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  $\chi^2$  distribution. What are the degrees of freedom?
- (d) Let  $P_1, \dots, P_n$  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  $\Phi(x)$ .
- What is the distribution of  $Y_i = \Phi^{-1}(1 - P_i)$ ? Show your work.
  - If  $H_0$  is false and  $P_i$  is not uniform, would you expect  $Y_i$  to be bigger, or smaller? Why?
- (f) Let  $P_1, \dots, P_n$  be a random sample of  $p$ -values.
- Propose a test statistic based on your answer to Question 6(e)i.
  - What is the null hypothesis of your test?
  - 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.
  - Would you reject the null hypothesis 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.
- For the test statistic of Question 6d,
    - What is the critical value at  $\alpha = 0.05$ ? The answer is a number.
    - What is the value of the test statistic? The answer is a number.
    - Do you reject the null hypothesis? Yes or No.
    - What if anything do you conclude?
  - For the test statistic of Question 6(e)i,
    - What is the critical value at  $\alpha = 0.05$ ? The answer is a number.
    - What is the value of the test statistic? The answer is a number.
    - Do you reject the null hypothesis? Yes or No.
    - What if anything do you conclude?