STAC51: Categorical Data Analysis

Assignment 1

Deadline to hand in: **Feb. 1 (Monday) 11:59 pm, 2021**

Total: 100 points

Note: Submit the R markdown file along with your compiled pdf file and/or hand-written part. You should provide comments quoting necessary values from your outputs so that it will be easy for TAs to grade your work.

Whenever you are using an R for generating random numbers, set seed to **your student number**. This can be done by simply adding the command ***set.seed(your student number)*** before generating the random number.

Q. 1 (10 pts) Let $(Y_1, Y_2, \ldots, Y_k) \sim Multinomial(n, \pi_1, \pi_2, \ldots, \pi_k)$, then

(a) Calculate the moment generating function (this will be a multivariate MGF).

(b) Show that $E(Y_j) = n\pi_j$

(c) Show that $Var(Y_j) = n\pi_j(1 - \pi_j)$

(d) Show that $Cov(Y_i, Y_j) = -n\pi_i\pi_j$

(e) Show that for $C = 2$, the $Cor(Y_1, Y_2) = -1$. Explain why?

Note: you can use the Multivariate MGF. Recall if $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ be a multivariate random variable, then the MGF is defined as:

$$M_{\mathbf{Y}}(t_1, t_2, \ldots, t_k) = E(exp(t_1Y_1 + \ldots + t_kY_k))$$

Now use the partial derivative on t_j or t_i, t_j to achieve the results.

Q. 2 (24 pts) At any given time, soft-drink dispensers may harbor bacteria such as Chryseobacterium meningosepticum that can cause illness. To estimate the proportion of contaminated soft-drink dispensers in a community in Toronto, researchers randomly sampled 30 dispensers and found 5 to be contaminated with Chryseobacterium meningosepticum. Let π be the proportion of contaminated dispensers in the population.

(a) We want to test whether 10% of the dispensers in the population were contaminated. Compute the test statistics for a score test, a Wald test, and a likelihood ratio test for the hypotheses

$$H_0 : \pi = 0.1 \quad \text{v.s.} \quad H_a : \pi \neq 0.1$$

and report the 3 P-values.

(b) Find the 90% Wald confidence interval for π.

(c) Find the 90% score confidence interval for π.

(d) Find the 90% Agresti-Coull confidence interval for π.

(e) Verify your score test and score confidence interval using R.

(f) Verify Wald, Agresti-Coull confidence interval using R.

Q. 3 (10 pts) Continue the previous problem with y = 5 and n = 30. In this problem, we will find the likelihood-ratio test based confidence interval for π:

(a) Find ℓ_0, the maximized likelihood under $H_0 : \pi = \pi_0$

(b) Find ℓ_1, the maximized likelihood over all possible π values.

(c) Find the likelihood-ratio test statistic for testing, $H_0 : \pi = \pi_0$.

(d) How big the likelihood-ratio test statistic must be at least to be significant at 0.1 level? Use R command, qchisq().

(e) Compute the 90% likelihood-ratio test-based confidence interval.

Q. 4 (20 pts) Observed (or true) coverage and the targeted coverage probabilities of confidence intervals are not necessary equal. In this question, we will calculate the observed (or true) coverage probability of Wald confidence intervals using two methods: Mote Carlo simulation and direct calculation.

(a) (Monte Carlo simulation) Generate $N = 100,000$ observations on Y where $Y \sim Bin(n, \pi)$, where $n = 25$ and $\pi = 0.06$. From each observation generated, calculate a Wald 95% confidence interval for the population proportion (π). (Note: This means you are calculating 100,000 confidence intervals). Calculate the proportion of these Wald intervals that contain 0.06 (the value of π). Comment on your results.

(b) (Direct calculation) In order to calculate the coverage probability for a known value of π, calculate a confidence interval for every possible value of y ($y = 0, 1, \ldots, n$) and check whether true value of the parameter is in the confidence interval calculated. Below are the steps:

i. Find all possible intervals that one could have with $y = 0, \ldots, n$

ii. Form $I(y) = 1$ if the interval for a y contains π and 0 otherwise

iii. Calculate the true confidence level as

$$\sum_{y=0}^{n} I(y) \binom{n}{y} (\pi)^y (1 - \pi)^{n-y}$$

Q. 5 (20 pts) In this question, we will also calculate and plot the true coverage probabilities of Wald confidence intervals for proportions (i.e. Binomial parameter) based on a sample of given size (n), but this time we calculate the coverage probabilities for many values of π making a plot of coverage probability versus π.

(a) (5 pts) For a Binomial sample of size n = 25, use the method in part (b) of the previous question (i.e. direct calculation) to calculate the coverage probability of a 95 % confidence interval for $\pi = 0.01, 0.02, \ldots, 0.99$ and plot them against π. Draw a horizontal line through the target probability 0.95. Comment on what you learned from your plot.

(b) (5 pts) Repeat part (a) above with n = 500 and plot both the curves on the same plot. Compare and comment on your findings.

(c) (10 pts) Repeat part (a) for Wald, Wilson (Score), Agresti-Coull and Clopper-Preason confidence intervals and plot the coverage probabilities versus π for all four confidence intervals on one graph (i.e all four curves on the same system of axes). Do not use **built-in R functions from "binom" package** for these confidence intervals.

Use four different colours for easy comparison. Compare and comment on your results. (Note that in this part, we are using the same values as in part (a) above, i.e n = 25, 95% confidence interval and $\pi = 0.01, 0.02, \ldots, 0.99$)

Q. 6 (16 pts) The power of the test is the probability of rejecting the null hypothesis, given the null hypothesis is false. (Given that a specific alternative hypothesis is true). Consider the score test for the binomial parameter π.

(a) Show that the power of the score test of the null hypothesis, $H_0 : \pi = \pi_0$ against the alternative $H_1 : \pi > \pi_0$ is given by:

$$Power(\pi) = p\left(Z > \frac{\pi_0 - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} + Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}}\right)$$

(b) Calculate the power of the test of $H_0 : \pi = 0.5_0$ against the alternative $H_1 : \pi > \pi_0$ based on a sample of size 100 and $\alpha = 0.05$, when the true value of $\pi = 0.55$ (i.e. Calculate Power(0.55)). Comment on your result.

(c) Plot the power curve (use R) for $0.4 < \pi < 0.7$. (with $\pi_0 = 0.5$, $n = 100$, $\alpha = 0.01$). Comment on your results.

(d) Repeat part (c) above but with n = 200. Repeat again with n = 300. Plot all three curves (i.e. for n = 100, 200, 300) on the same frame for easy comparison. Use three different colours for the three curve. Comment on your results.