Topic 6 – Population Proportions and Logistic Regression

Topic 6a (Part I) - Basics/Review

Until now, the response variable has been numeric. We will now consider inference upon categorical response variables. Specifically, consider a binary response variable where the response is classified as either a "success" or a "failure". First, we discuss inferential procedures on one/two proportion(s), which means reviewing earlier STAT courses.

NOTE: All methods are based on large-sample approximations such that *Z*-tools are permissible. This requires use of the standard normal table (or *Z*-table). Still, given the similarity of *Z*-tools to *T*-tools, the same procedure for parameter estimation and inference shall be followed.

One-Sample Procedure

1. Parameter: π = population proportion of "successes"
2. Estimator: $\hat{\pi}$ = sample proportion of "successes" in sample of size n
3. Sampling Distribution of Estimator:

   For large n ($n\hat{\pi} \geq 15$, $n(1-\hat{\pi}) \geq 15$), it can be shown that:

$$\hat{\pi} \dot{\sim} N\left(\mu_{\hat{\pi}} = \pi, \sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}} \right)$$

4. i) Test statistic for testing H_0: $\pi = \pi_0$ (assuming random sample and large n)

$$Z_0 = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

   → p-values obtained from Z-table

4. ii) An *approximate* $(1-\alpha)100\%$ CI for π is (same assumptions as testing)

$$\hat{\pi} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

   → Substituting $\hat{\pi}$ for π allowed because of large n, but only allows for *approximate* CI.

Two-Sample Procedure

1. Parameter: $\pi_1 - \pi_2$ = difference in prop'n of "successes" from two pop'ns, Y_1 and Y_2.
2. Estimator: $\hat{\pi}_1 - \hat{\pi}_2$ = difference in proportion of "successes" from independent and random samples of size n_1 and n_2, from Y_1 and Y_2, respectively.
3. Sampling Distribution of Estimator:

   For large n_1 and n_2 ($n\hat{\pi}_1 \geq 15$, $n(1-\hat{\pi}_1) \geq 15$, $n\hat{\pi}_2 \geq 15$, $n(1-\hat{\pi}_2) \geq 15$), then:

$$\hat{\pi}_1 - \hat{\pi}_2 \dot{\sim} N\left(\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2, \sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n} + \frac{\pi_2(1-\pi_2)}{n}} \right)$$

4. i) Test statistic for testing H_0: $\pi_1 - \pi_2 = 0$

The purpose is to test equality such that either sample should estimate π well. Ergo, a weighted average of $\hat{\pi}_1$ and $\hat{\pi}_2$ that gives more weight to the larger sample is used. The *combined estimate of the common population proportion* is

$$\hat{\pi}_c = \frac{"total_successes"}{n_1 + n_2} = \frac{n_1\hat{\pi}_1 + n_2\hat{\pi}_2}{n_1 + n_2}$$

Then, assuming independent & random samples of size n_1 and n_2, we have

$$Z_0 = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}_c(1-\hat{\pi}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

→ *p*-values obtained from *Z*-table

4. ii) An *approximate* $(1 - \alpha)100\%$ CI for $\pi_1 - \pi_2$ does not make the same assumption as the hypothesis test does, so we substitute similar to the one-sample procedure here.

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2} \times \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

→ Substituting $\hat{\pi}_1$ and $\hat{\pi}_2$ for π_1 and π_2 allowed because of large *n*, but only allows for *approximate* CI.

Topic 6a (Part II) - Odds and Odds Ratio

Another way to analyse data from binary response variables is to compare the odds of success vs. the odds of failure.

The Odds

Def'n: If π represents the proportion of individuals satisfying some criteria (success), then the odds (ω) of a person satisfying this criteria is

$$\omega = \frac{proportion_of_successes}{proportion_of_failure} = \frac{\pi}{1-\pi}$$

Note: Odds are in favour of the event, gambling odds are usually *against* the event.

Example:

a) Suppose 80% of individuals will suffer from a cold this winter. Thus, $\pi = 0.8$. Then, the odds of suffering from a cold are

$$\omega = \frac{\pi}{1-\pi} = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = 4$$

Or, there is a 4 to 1 chance that any individual will suffer from a cold.

b) Suppose you're told that the odds of catching the flu are 2 (or 2 to 1). What is the probability of catching the flu? The 2 to 1 odds mean that for every 2 individuals that will catch the flu, 1 will not. Thus, the probability (π) of catching the flu is 2/3 = 0.67 or 67%. Or,

$$\pi = \frac{\omega}{1+\omega} = \frac{2}{1+2} = \frac{2}{3} \approx 0.667$$

Properties of Odds

1. Boundaries: $0 \le \pi \le 1$, whereas $\omega > 0$.
2. If $\pi = ½$, then $\omega = 1$ (a.k.a. odds are "fifty-fifty").
3. Odds are not defined for proportions that are exactly 0 or 1.
4. If the odds of a 'yes' outcome are ω, the odds of a 'no' outcome are $1/\omega$.
5. If the odds of a 'yes' outcome are ω, the probability of 'yes' is $\pi = \frac{\omega}{1+\omega}$.

The Odds Ratio

Def'n: For two populations with proportions π_1 and π_2, the odds ratio is defined as

$$\phi = \frac{\omega_1}{\omega_2} = \left(\frac{\pi_1}{1-\pi_1}\right) \Big/ \left(\frac{\pi_2}{1-\pi_2}\right) = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

The odds ratio is a more desirable measure than a difference in population proportions because of three reasons:

1. In practice, the odds ratio tends to remain more nearly constant over levels of confounding variables.
2. The odds ratio is the *only* parameter that can be used to compare two groups of binary responses from a retrospective study.
3. The comparison of odds extends nicely to regression analysis.

Sampling Distribution of the Sample Log Odds Ratio

Inferential procedures for the odds ratio are based on the natural log of the odds ratio as it is a bit more normal than ϕ itself. It can be shown that for large samples (n_1 from population 1 & n_2 from population 2) that $\ln(\hat{\phi}) = \log(\hat{\phi})$ is approx. normal such that

1. Mean($\log(\hat{\phi})$) = $\log(\phi)$
2. S.D.($\log(\hat{\phi})$) = $\sqrt{\frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)}}$

Inferential Procedures

4. i) Test statistic for testing H_0: $\log(\phi) = 0$ → H_0: $\phi = 1$

$$Z_0 = \frac{\log(\hat{\phi}) - 0}{S.E.(\log(\hat{\phi}))} \qquad \text{where S.E.}(\log(\hat{\phi})) = \sqrt{\frac{1}{n_1\hat{\pi}_c(1-\hat{\pi}_c)} + \frac{1}{n_2\hat{\pi}_c(1-\hat{\pi}_c)}}$$

→ *p*-values obtained from *Z*-table

4. ii) An *approximate* $(1 - \alpha)100\%$ CI for $\log(\phi)$ is

$$\log(\hat{\phi}) \pm z_{\alpha/2} \times \sqrt{\frac{1}{n_1\hat{\pi}_1(1-\hat{\pi}_1)} + \frac{1}{n_2\hat{\pi}_2(1-\hat{\pi}_2)}}$$

→ Substituting $\hat{\pi}_1$ and $\hat{\pi}_2$ for π_1 and π_2 allowed because of large *n*, but only allows for *approximate* CI.

→ If a CI for $\log(\phi)$ is (L, U), then a CI for ϕ is (e^L, e^U).

Topic 6b - Binary Logistic Regression

Def'n: A generalized linear model (GLM) is a probability model in which the mean of a response variable is related to explanatory variables through a regression equation. As before, $\mu = \mu(Y \mid X_1, \ldots, X_k)$ indicates the mean response and the regression structure is linear in unknown parameters, or

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Consider now a specified function of μ, a *link function*, which is equal to the structure

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

For a normally distributed response variable (as in Topic 5), the natural choice is the *identity link*, $g(\mu) = \mu$. Binary responses are NOT normally distributed, BUT the log-odds ratio is. Consider the *logit* function where

$$g(\pi) = \text{logit}(\pi) = \log[\pi/(1 - \pi)]$$

where, to represent the population mean, π is used instead of μ because the mean is a proportion or probability. Thus, for logistic regression, we have

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

PLEASE NOTE: Logistic regression is not "linear" regression because the equation for $\mu(Y \mid X_1, \ldots, X_k)$ is not linear in terms of β. Yet, it is still a "GLM" because of the linearity the link function has with the regression coefficients.

Interpretation of Model/Coefficients

Def'n: The inverse of the logit function is called the logistic function. If $\text{logit}(\pi) = \eta$,

$$\pi = \frac{e^{\eta}}{1 + e^{\eta}}$$

This function allows for calculating probabilities, for a given set of values for the explanatory variable(s). Extending to odds, it can be shown (can YOU?) that

$$\omega = e^{\eta}$$

and, for the ratio of the odds at $X_1 = A$ relative to the odds at $X_1 = B$, for fixed values of the other X's,

$$\frac{\omega_A}{\omega_B} = e^{\beta_1 (A - B)}$$

For an individual regression coefficient under the logistic regression model, a hypothesis test for H_0: $\beta_j = 0$ is

$$Z_0 = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)}$$

Since this test is based on *approximate* normality of maximum likelihood estimates, it is referred to as *Wald's test*. (*T*-tools only apply to normally-distributed response variables.) NOTE: SPSS output will not calculate Z_0 or find the corresponding *p*-value range. The "Wald" column (and given *p*-value) refers to a test statistic that follows a χ^2_1 distribution that CANNOT do one-tailed tests.

The confidence interval for an individual regression coefficient is

$$\hat{\beta}_j \pm z_{\alpha/2} \times SE(\hat{\beta}_j)$$

Note that interpretation to odds requires back-transformation.