

Statistics 403 Problem Set 9

Due in lab on Friday, November 13th

1. Suppose we observe data in an observational study with a treatment variable $T = 0, 1$ and a confounding variable $S = 0, 1$, such that the outcome satisfies

$$E(Y|T, S) = \alpha_T + \beta_S,$$

where $\beta_0 = 0$. Suppose also that $P(S = 1|T = 1) = p_1$ and $P(S = 1|T = 0) = p_0$. We collect data on outcomes (Y) and the treatment variable (T), but we are unable to obtain data on the confounder S . Suppose that \bar{Y}_T is the average outcome of our n_T treated subjects, and \bar{Y}_U is the average outcome of our n_U untreated subjects. Let $D = \bar{Y}_T - \bar{Y}_U$. Suppose we use $D \pm 2\text{SE}(D)$ as a 95% confidence interval for $\alpha_1 - \alpha_0$. What is the coverage probability of this interval?

Solution: The coverage probability is

$$P(D - 2\text{SE}(D) \leq \alpha_1 - \alpha_0 \leq D + 2\text{SE}(D))$$

The expected value of the estimator is given in the notes:

$$ED = \alpha_1 - \alpha_0 + \beta_0(p_1 - p_0),$$

and the standard error is

$$\sigma\sqrt{1/n_T + 1/n_U}.$$

So we can calculate the coverage probability as:

$$\begin{aligned} &P(D - 2\text{SE}(D) \leq \alpha_1 - \alpha_0 \leq D + 2\text{SE}(D)) \\ &= P(-2 \leq (D - (\alpha_1 - \alpha_0))/\text{SE}(D) \leq 2) \\ &= P(-2 \leq (D + \beta_1(p_1 - p_0) - (\alpha_1 - \alpha_0 + \beta_1(p_1 - p_0)))/\text{SE}(D) \leq 2) \\ &= P(-2 - \beta_1(p_1 - p_0)/\text{SE}(D) \leq (D - ED)/\text{SE}(D) \leq 2 - \beta_1(p_1 - p_0)/\text{SE}(D)) \\ &= P(-2 - \beta_1(p_1 - p_0)/\text{SE}(D) \leq Z \leq 2 - \beta_1(p_1 - p_0)/\text{SE}(D)). \end{aligned}$$

If we are given β_1 , $p_1 - p_0$, σ , n_T , and n_U , we can calculate the coverage probability of the interval. The only situations in which we will get the nominal 95% coverage probability are if $\beta_1 = 0$ (no effect of the confounder on the outcome) or $p_1 = p_0$ (perfect balance).

2. Suppose we have a cluster sample of $q = 20$ clusters of equal size $n_i = 10$, for which the intraclass correlation coefficient is r . We plan to use the data to carry out a level $\alpha = 0.05$ hypothesis test that $EX = 0$ (this is a one-sample test). If we carry out the test as if the data were an iid sample, what will be the actual level of the test (i.e. the probability of rejecting the null hypothesis when the null hypothesis is true)?

Solution: The actual standard error is $\sigma^2/n + \tau^2/q = \sigma^2/200 + \tau^2/20$, but if we treat the data as an iid sample, we would have a standard deviation of $(\sigma^2 + \tau^2)/n = (\sigma^2 + \tau^2)/200$. The level of the test with test statistic T is $P(T > 2)$, calculated under the null hypothesis. Since we are using the wrong standard deviation, our test statistic is

$$T = \sqrt{200}\bar{X}/\sqrt{\sigma^2 + \tau^2}.$$

Thus the level is

$$\begin{aligned} P(T > 2) &= P(\sqrt{200}\bar{X}/\sqrt{\sigma^2 + \tau^2} > 2) \\ &= P(\bar{X} > 2\sqrt{\sigma^2 + \tau^2}/\sqrt{200}) \\ &= P\left(\bar{X}/\sqrt{\sigma^2/200 + \tau^2/20} > 2\sqrt{\frac{(\sigma^2 + \tau^2)/200}{\sigma^2/200 + \tau^2/20}}\right) \\ &= P\left(Z > 2\sqrt{\frac{\sigma^2 + \tau^2}{\sigma^2 + 10\tau^2}}\right) \\ &= P\left(Z > 2\sqrt{1/\frac{\sigma^2 + 10\tau^2}{\sigma^2 + \tau^2}}\right) \\ &= P\left(Z > 2\sqrt{1/(1 + 9r)}\right). \end{aligned}$$

A similar calculation gives us $P(T < -2) = P\left(Z < -2\sqrt{1/(1 + 9r)}\right)$. If $r = 0$, we get the correct size $P(Z > 2) + P(Z < -2) \approx 0.05$. Otherwise the level will be greater than 0.05. This means that we are overstating the evidence against the null hypothesis.

3. Suppose we are studying a population in which 70% of the individuals are female, and 30% are male. We collect a sample consisting of a fraction w_f of females, and w_m of males, and we form the averages \bar{X}_f and \bar{X}_m of the female and male data. Then we form the statistic $(\bar{X}_f + \bar{X}_m)/2$. Let E denote the expected value of this statistic.

(a) What is E ?

Solution: Let $\mu_f = EX_f$ and $\mu_m = EX_m$, where X_f and X_m are individual data points from the female and male subpopulations, respectively. The expected value of the statistic is

$$\begin{aligned}
E &= E\left(\frac{\bar{X}_f + \bar{X}_m}{2}\right) \\
&= \frac{E\bar{X}_f + E\bar{X}_m}{2} \\
&= \frac{EX_f + EX_m}{2} \\
&= \frac{\mu_f + \mu_m}{2}.
\end{aligned}$$

- (b) Suppose that the expected values of the female and male data, are, respectively, μ_f and μ_m . What is the expected value μ of a data value randomly sampled from the entire population?

Solution: Let X be the value for a randomly selected person, then apply the double expectation theorem: $EX = EE(X|\text{gender})$. Specifically, $EX = \mu_f P(\text{female}) + \mu_m P(\text{male}) = 0.7 \cdot \mu_f + 0.3 \cdot \mu_m$.

- (c) What is the relative difference $(E - \mu)/\mu$?

Solution: Using the results from (a) and (b), we get

$$\begin{aligned}
(E - \mu)/\mu &= \frac{(\mu_f + \mu_m)/2 - (0.7\mu_f + 0.3\mu_m)}{0.7\mu_f + 0.3\mu_m} \\
&= \frac{-0.2\mu_f + 0.2\mu_m}{0.7\mu_f + 0.3\mu_m} \\
&= 2 \frac{\mu_m - \mu_f}{7\mu_f + 3\mu_m}.
\end{aligned}$$

4. We are performing a comparative analysis of the expected value of an outcome Y based on a binary treatment variable $T = 0, 1$. We plan to use stratification to minimize the effect of a quantitative confounding variable X . We plan to use three strata, such that the proportion of treated subjects in stratum k is Q_T^k , the expected values of treated and untreated subjects in stratum k are $\mu_T^{(k)}$ and $\mu_U^{(k)}$, respectively, and the standard deviations of treated and untreated subjects in stratum k are $\sigma_T^{(k)}$ and $\sigma_U^{(k)}$, respectively.

	Q_T^k	$\mu_T^{(k)}$	$\mu_U^{(k)}$	$\sigma_T^{(k)}$	$\sigma_U^{(k)}$
1	0.1	0.7	0.3	0.4	0.2
2	0.5	1.0	0.5	0.3	0.4
3	0.8	1.3	0.9	0.4	0.5

- (a) What are the expected values of D_1 , D_2 , and D_3 ?

Solution:

$$ED_1 = 0.7 - 0.3 = 0.4.$$

$$ED_2 = 1 - 0.5 = 0.5.$$

$$ED_3 = 1.3 - 0.9 = 0.4.$$

(b) What is the variance of $(D_1 + D_2 + D_3)/3$?

Solution: Assuming sample size m per stratum, the variances of the D_i are

$$\text{var}(D_1) = 0.4^2/(0.1m) + 0.2^2/(0.9m) = 1.64/m.$$

$$\text{var}(D_2) = 0.3^2/(0.5m) + 0.4^2/(0.5m) = 0.5/m.$$

$$\text{var}(D_3) = 0.4^2/(0.8m) + 0.5^2/(0.2m) = 1.45/m.$$

Thus the variance of $(D_1 + D_2 + D_3)/3$ is $(1.64 + 0.5 + 1.45)/(9m) = 0.4/m$.

(c) What is the variance of $\bar{D} = w_1D_1 + w_2D_2 + w_3D_3$, using weights w_j as defined on slide 29?

Solutions: The weights are

$$w_1 = \frac{1/1.64}{1/1.64 + 1/0.5 + 1/1.45} = 0.18$$

$$w_2 = \frac{1/0.5}{1/1.64 + 1/0.5 + 1/1.45} = 0.61$$

$$w_3 = \frac{1/1.45}{1/1.64 + 1/0.5 + 1/1.45} = 0.21$$

Note that the middle stratum has the best balance, and gets the highest weight. The variance of the optimally weighted sum of the D_i is

$$w_1^2\text{var}(D_1) + w_2^2\text{var}(D_2) + w_3^2\text{var}(D_3) = 0.18^2 \cdot 1.64/m + 0.61^2 \cdot 0.5/m + 0.21^2 \cdot 1.45/m \approx 0.30/m.$$

Note that the variance using the optimal weights is lower than the variance using equal weights.