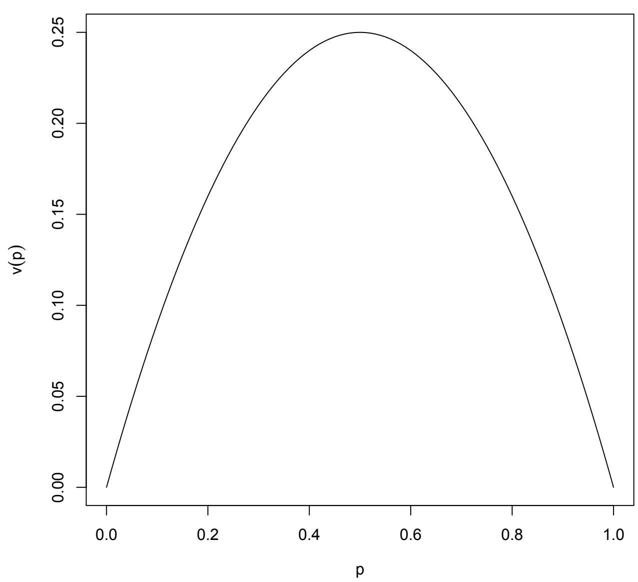
MATH10282 Introduction to Statistics
Semester 2, 2019/2020
Example Sheet 10 - Solutions

1. (i) The width of the $100(1-\alpha)\%$ CI is given by the difference of the end-points, $2z_{1-\alpha/2}\widehat{\text{s.e.}}(\hat{p})$. For fixed α, this is proportional to $\widehat{\text{s.e.}}(\hat{p})$.

(ii) The graph of $v(p)$, $p \in (0,1)$, is shown below:



(iii) Completing the square, $v(p) = p(1-p) = p - p^2 = -\left(p - \frac{1}{2}\right)^2 + \frac{1}{4}$. Thus $v(p) \leq \frac{1}{4}$ with equality when $p = \frac{1}{2}$. This can also be shown by differentiating, and solving $\frac{dv}{dp} = 0$. Hence,

$$\text{s.e.}(\hat{p}) \leq \sqrt{\frac{1}{4n}}\,.$$

(iv) If $n = 1000$, the width of the 95% CI is at most

$$2z_{0.975}\sqrt{\frac{1}{4n}} = 0.062\,.$$

(N.B. $z_{0.975} = 1.96$.)

2. We have

$$n_A = 150\,, \quad \bar{x}_A = 1386\,, \quad s_A = 114\,,$$
$$n_B = 200\,, \quad \bar{x}_B = 1218\,, \quad s_B = 98\,.$$

We are not told that the data are normally distributed, and so we use the confidence intervals for the difference of two non-normal means, where the variances are unknown (Ch 7, Part II, slide 18).

(i) Assuming $\sigma_A^2 \neq \sigma_B^2$, note that $z_{0.975} = 1.96$. The approximate 95% CI for $\mu_A - \mu_B$ has end points

$$\begin{aligned}&(1386 - 1218) \pm 1.96 \times \sqrt{\frac{114^2}{150} + \frac{98^2}{200}}\\ &= 168 \pm 1.96\sqrt{134.66}\,.\end{aligned}$$

Hence the approximate 95% confidence interval for $\mu_A - \mu_B$ is (145.26, 190.74).

(ii) Assuming that $\sigma_A^2 = \sigma_B^2 = \sigma^2$, we have that

$$\hat{\sigma}^2 = \frac{149 \times 114^2 + 199 \times 98^2}{150 + 200 - 2} = 11056.32\,.$$

Hence the 95% CI for $\mu_A - \mu_B$ has end points

$$\begin{aligned}&(1386 - 1218) \pm 1.96\sqrt{11056.32 \times \left(\frac{1}{150} + \frac{1}{200}\right)}\\ &= 168 \pm 1.96 \times \sqrt{128.9904}\,.\end{aligned}$$

Hence the approximate 95% confidence interval for $\mu_A - \mu_B$ is (145.74, 190.26). This is slightly narrower than before. In both cases, the confidence interval excludes 0, and so we conclude that it is not plausible that $\mu_A = \mu_B$.

3. (i) We have that

$$\begin{aligned}\mathrm{E}(\hat{\mu}) &= \mathrm{E}\left(\frac{n\bar{X}_n + m\bar{Y}_m}{n+m}\right) = \frac{1}{n+m}\,\mathrm{E}(nX_n + m\bar{Y}_m)\\ &= \frac{1}{n+m}\left[n\,\mathrm{E}(X_n) + m\,\mathrm{E}(\bar{Y}_m)\right] = \frac{n\mu + m\mu}{n+m} = \mu\,.\end{aligned}$$

Hence $\hat{\mu}$ is unbiased for μ. Similarly,

$$\begin{aligned}\mathrm{E}(\hat{\sigma}^2) &= \frac{1}{n+m-2}\,\mathrm{E}\left[(n-1)S_X^2 + (m-1)S_Y^2\right]\\ &= \frac{1}{n+m-2}\left\{(n-1)\,\mathrm{E}[S_X^2] + (m-1)\,\mathrm{E}[S_Y^2]\right\}\\ &= \frac{1}{n+m-2}\left\{(n-1)\sigma^2 + (m-1)\sigma^2\right\}\\ &\quad \text{since } S_X^2 \text{ and } S_Y^2 \text{ are both unbiased estimators of } \sigma^2\\ &= \frac{(n+m-2)\sigma^2}{n+m-2} = \sigma^2\,.\end{aligned}$$

(ii) For the variance, note that

$$\begin{aligned}\operatorname{Var}(\hat{\mu}) &= \frac{1}{(n+m)^2}\operatorname{Var}(n\bar{X}_n + m\bar{Y}_m)\\ &= \frac{1}{(n+m)^2}\left[\operatorname{Var}(n\bar{X}_n) + \operatorname{Var}(m\bar{Y}_m)\right]\end{aligned}$$

since by independence of the two samples, $\bar{X}_n$ and $\bar{Y}_m$ are independent.

$$\begin{aligned}&= \frac{n^2\operatorname{Var}(\bar{X}_n) + m^2\operatorname{Var}(\bar{Y}_n)}{(n+m)^2}\\ &= \frac{n^2(\sigma^2/n) + m^2(\sigma^2/m)}{(n+m)^2} = \frac{(n+m)\sigma^2}{(n+m)^2} = \frac{\sigma^2}{n+m}\,.\end{aligned}$$

Thus, if both $n \to \infty$ and $m \to \infty$, then $\operatorname{Var}(\hat{\mu}) \to 0$.

4. We have

Brand A: $n_1 = 12$, $\bar{x}_1 = 21.8$, $s_1 = 8.7$

Brand B: $n_2 = 12$, $\bar{x}_2 = 18.9$, $s_2 = 7.5$

The common variance is estimated by

$$\hat{\sigma}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{11\times 8.7^2 + 11\times 7.5^2}{22} = 65.97\,,$$

The 95% CI for $\mu_A - \mu_B$ has end points

$$(\bar{x}_A - \bar{x}_B) \pm t_{0.975}\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (21.8 - 18.9) \pm 2.074\sqrt{\frac{2\times 65.97}{12}}\,,$$

where $t_{0.975} = 2.074$ is the 0.975 point of a t distribution with $n_1 + n_2 - 2 = 22$ degrees of freedom. Hence the 95% CI for $\mu_A - \mu_B$ is $(-3.98, 9.78)$. The interval contains zero, thus given the data it is plausible that $\mu_A = \mu_B$.

5. Let p_1 denote the (population) proportion of registered voters who turned out to vote in California and p_2 be the (population) proportion of registered voters who turned out to vote in Colorado. We have that

$$\begin{aligned}n_1 &= 288, \quad r_1 = 141, \quad \hat{p}_1 = 141/288 = 0.4896\\ n_2 &= 216, \quad r_2 = 125, \quad \hat{p}_2 = 125/216 = 0.5787\end{aligned}$$

The end points of the 95% CI for $p_1 - p_2$ are

$$\begin{aligned}&\hat{p}_1 - \hat{p}_2 \pm z_{0.975}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\\ &= (0.4896 - 0.5787) \pm 1.96\sqrt{\frac{0.4896\times 0.5104}{288} + \frac{0.5787\times 0.4213}{216}}\end{aligned}$$

Hence the 95% CI for $p_1 - p_2$ is $(-0.177, -0.0015)$. This interval does not contain zero, and so given the data it is *not* plausible that $p_1 = p_2$. Thus, we conclude that $p_1 \neq p_2$.