ACM 157 MIDTERM

(1)

(a)

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2$$

$$\mathbb{E}(X) = \alpha\pi + (1-\pi)\beta$$

$$\Rightarrow \mathbb{E}(X)^2 = (\alpha\pi + \beta - \beta\pi)^2$$

$$\mathbb{E}(X^2) = \sum_x x^2 p(x) = \alpha^2\pi + \beta^2(1-\pi)$$

$$= \alpha^2\pi + \beta^2 - \beta^2\pi$$

$$\Rightarrow \mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

$$\Rightarrow = \alpha^2\pi + \beta^2 - \beta^2\pi - \left[(\alpha\pi + \beta - \beta\pi)^2\right]$$

$$= (\pi - \pi^2)\alpha^2 + 2(\pi - 2)\pi\alpha\beta - (\pi - 2)\pi\beta^2$$

$$= \pi(\alpha^2 - 2\alpha\beta + \beta^2) - \pi^2(\alpha^2 - 2\alpha\beta + \beta^2)$$

$$= (\alpha - \beta)^2(\pi - \pi^2)$$

variance of population $\rightarrow$ $\boxed{\mathbb{V} = (\alpha - \beta)^2(\pi - \pi^2)}$

(b) Variance maximized at $\frac{d}{d\pi}\mathbb{V} = 0$

$$\Rightarrow \frac{d}{d\pi}\mathbb{V} = \frac{d}{d\pi}(\alpha-\beta)^2(\pi - \pi^2) = (\alpha-\beta)^2(1 - 2\pi) = 0 \Rightarrow \pi = .5$$

$$\frac{d^2}{d\pi^2}\mathbb{V} = (\alpha-\beta)^2(-2) < 0 \Rightarrow \frac{d}{d\pi}\mathbb{V} = 0$$, finds maximum by 2nd derivative test

Validated with 2nd derivative test $\frac{d}{d\pi}\mathbb{V} = 0$, finds the value of π yielding maximum variance which is $\pi = .5$

(2)

(a) To achieve same accuracy standard errors must be equal.

$$se[\bar{X}_n] = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n-1}{N-1}}$$

$$se_g = \frac{\sigma_g}{\sqrt{n_g}}\sqrt{1-\frac{n_g-1}{N_g-1}}, \quad se_u = \frac{\sigma_u}{\sqrt{n_u}}\sqrt{1-\frac{n_u-1}{N_u-1}}$$

$se_g = se_u$, since $\sigma^2_g = \sigma^2_u \Rightarrow \sigma_g = \sigma_u$

$$\Rightarrow se_g = se_u \Rightarrow \frac{1}{\sqrt{n_g}}\sqrt{1-\frac{n_g-1}{N_g-1}} = \frac{1}{\sqrt{n_u}}\sqrt{1-\frac{n_u-1}{N_u-1}}$$

$$\Rightarrow \frac{1}{10}\sqrt{1-\frac{99}{1298}} = \frac{1}{\sqrt{n_u}}\cdot\sqrt{1-\frac{n_u-1}{937}}$$

$$\frac{\sqrt{\frac{109}{118}}}{10} = \frac{1}{\sqrt{n_u}}\cdot\sqrt{\frac{938-n_u}{937}}$$

$$\Rightarrow \frac{109}{118}n_u = 100\left(\frac{938-n_u}{937}\right) \Rightarrow n_u \approx 97.1483$$

Theoretically $n_u = 97.1483$ (at least 98 students), sample size should be used in the undergraduate survey to achieve the same level of accuracy as in the graduate student survey.

(2)(b) If $n_g = N_g$

$$se_g = \frac{\sqrt{\sigma^2_g}}{\sqrt{n_g}} \cdot \sqrt{1 - \frac{n_g - 1}{N_g - 2}} = \frac{\sqrt{\sigma^2_g}}{\sqrt{n_g}} \cdot 0 = \underline{0}$$

$$se_u \Rightarrow \text{needs to} = 0 = \frac{\sqrt{\sigma^2_u}}{\sqrt{n_u}} \cdot \sqrt{1 - \frac{n_u - 1}{N_u - 2}}$$

$$\frac{\sqrt{\sigma^2_u}}{\sqrt{n_u}} \neq 0 \Rightarrow \sqrt{1 - \frac{n_u - 1}{N_u - 2}} = 0 \Longrightarrow \underline{n_u = N_u}$$

= 938

We notice that $n_u = N_u$ is required for accuracy to be at least grad level, this intuitively makes sense as $n_g = N_g$ means the whole graduate sampled so no SE. Thus to have 0 SE for undergrads we must sample the entire N_u. Since SE goes to 0, the assumption $\sigma_g^2 = \sigma_u^2$ does not matter in this specific case but would matter if the finite population correction doesn't go to 0.

(3) We can show this by proving the probability density of $Q = X_{(n)}/\theta$ does not depend on θ, notice

$P(Q < x) = P(X_{(n)}/\theta < x) =$ [illegible]

s.t. $x \in [0,1]$ since $Q = X_{(n)}/\theta \in [0,1]$

we can then equate this to:

$P(Q < x) = P(X_{(n)}/\theta < x) = \mathbb{P}(X_{(n)} < \theta x)$.

where $\theta x \in [0, \theta]$. From here, notice

$$\mathbb{P}(X_{(n)} < y) = \mathbb{P}(\forall i \in [1,n] : X_i < y) = \left(\frac{y}{\theta}\right)^n$$ since

all x_i are sampled i.i.d from $U[0,\theta]$, thus we can get:

$$\mathbb{P}(X_{(n)} < \theta x) = \left(\frac{\theta x}{\theta}\right)^n = x^n$$

thus we get $P(Q < x) = x^n$ for $x \in [0,1]$ for $x > 1$ $P(Q < x) = 1$. Thus we see the density function and consequently the pdf are only dependent on x not θ and thus Q is a pivot as it does not depend on θ.

(3b) Given from (a) $P(Q \leq x) = x^n$ we can write $P(y \leq Q \leq x) = x^n - y^n$ for $x, y \in [0,1]$, to construct a confidence interval on Q we only need:

$P(y \leq Q \leq x) = x^n - y^n = 1-\alpha$, notice setting $x = 1$ yields:

$\Rightarrow P(y \leq Q \leq 1) = 1 - y^n$, to find the $1-\alpha$ confidence interval we can then just set $1 - y^n = 1-\alpha \Rightarrow y = \alpha^{1/n}$

$\Rightarrow P(\alpha^{1/n} \leq Q \leq 1) = 1-\alpha$, substituting $Q = X_n/\theta$ we get.

$$P(\alpha^{1/n} \leq Q \leq 1) = P\left(\alpha^{1/n} \leq \frac{X_n}{\theta} \leq 1\right) \overset{=1-\alpha}{\Rightarrow} P\left(1 \leq \frac{\theta}{X_n} \leq \alpha^{-1/n}\right) = 1-\alpha$$

rearranging terms, we conclude: $\boxed{P\left(X_n \leq \theta \leq X_n \cdot \alpha^{-1/n}\right) = 1-\alpha}$ for $\alpha \in [0,1]$

Notice this interval was constructed on the pivot Q, with distribution independent of θ, so we can build a confidence interval on θ using pivot distribution of Q not dependent on θ.

(4)(a)

An unbiased estimate of $p_1 - p_2$ would be to subtract the unbiased estimates $\hat{p}_1 - \hat{p}_2$ to make these unbiased:

$E(\hat{p}_1) - p_1 = 0$, consider $\hat{p}_1 = \frac{1}{n} \sum_{i=2}^{n} \mathbb{I}(x_i)$

where x_i is a 1 or 0 based on if the patient recovered, then: $E(\hat{p}_1) = \frac{1}{n} \sum_{i=2}^{n} E(\mathbb{I}(x_i))$

$= \frac{1}{n} \cdot n(p) = p$ so $E(\hat{p}_1) - p_1 = p_1 - p_1 = 0$

thus $\frac{\#\text{ of recovered}}{\text{total group population}}$ is a good unbiased estimate of $\hat{p}_1$ and $\hat{p}_2$ and give specific estimates:

$\hat{p}_1 = \frac{400}{450}$; $\hat{p}_2 = \frac{375}{450}$

which gives $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = \frac{400 - 375}{450} = \frac{25}{450}$

$\Rightarrow \boxed{\hat{\theta} = \frac{1}{18}}$

(4b)

$$se[\hat{\theta}] = \sqrt{V[\hat{\theta}]} = \sqrt{V[\hat{P}_1 - \hat{P}_2]}$$, since $\hat{P}_1, \hat{P}_2$ are generated from different groups/treatments: they can be considered independent.

$$\sqrt{V[\hat{P}_1 - \hat{P}_2]} = \sqrt{V[\hat{P}_1] + V[\hat{P}_2]} \Rightarrow V[\hat{P}_i] = se[\hat{P}_i]^2$$

$$= \sqrt{se[\hat{P}_1]^2 + se[\hat{P}_2]^2}$$

$$se[P_2] = \frac{S_2}{\sqrt{n}}, \quad s_1^2 = \frac{1}{n-1}\sum_{i=2}^{\hat{}} (x_i - \bar{x}_n)^2$$

$$= \frac{1}{449}\left(400 \times \left(\frac{50}{450}\right)^2 + 50 \times \left(\frac{-400}{450}\right)^2\right)$$

$$se[P_2] = \frac{S_2}{\sqrt{n}}, \quad s_2^2 = \frac{1}{n-1}\sum_{i=2}^{\hat{}} (x_i - \bar{x}_n)^2$$

$$= \frac{1}{449}\left(375 \times \left(\frac{75}{450}\right)^2 + 75\left(\frac{-375}{450}\right)^2\right)$$

$$se[P_1] = \frac{\sqrt{s_1^2}}{\sqrt{450}} \approx .0148, \quad se[P_2] = \frac{\sqrt{s_2^2}}{\sqrt{450}} \approx .01739$$

$$\sqrt{V[\hat{P}_1 - \hat{P}_2]} = \sqrt{se[\hat{P}_1]^2 + se[\hat{P}_2]^2} \approx .02285$$

$$se[\hat{\theta}] \approx .02285$$

(5)

(a) $\mathbb{E}[\bar{X}_n^* \mid X_1, \dots X_n]$, $\bar{X}_n^* = \frac{1}{n}\sum_{i=1}^{n} X_i^*$

$= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^* \mid X_1, \dots X_n\right]$, ~~by ...~~

by Linearity of expectation: $\Rightarrow = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i^* \mid X_1, \dots X_n)$

since $X_1 \dots X_n \sim F$ sampled i.i.d and X_i^* sampled from $X_1 \dots X_n$ by SRS: $\mathbb{E}(X_i^* \mid X_1 \dots X_n) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\Rightarrow = \frac{1}{n}(n) \cdot \mathbb{E}(X_i^* \mid X_1 \dots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ from $X_1 \dots X_n$

so we conclude $\mathbb{E}[\bar{X}_n^* \mid X_1, \dots X_n] = \frac{1}{n}\sum_{i=1}^{n} X_i$

(b) By law of total expectation: $\mathbb{E}[\bar{X}_n^*] =$ ... from (a)

$= \mathbb{E}[\mathbb{E}[\bar{X}_n^* \mid X_1 \dots X_n]] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$

by linearity of expectation this equates to: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i)$

since X_i are sampled i.i.d from $\sim F$ with mean μ, $\mathbb{E}(X_i) = \mu \quad \forall X_i \in [1, n]$, thus

$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \left(\frac{1}{n}\right)(n)(\mu) = \mu$

thus we conclude $\mathbb{E}[\bar{X}_n^*] = \mu$

(6)

(a) From the definition of the pmf $f(x;\theta)$

we can solve for $\hat{\theta}_{MOM}$ using the first moment and the population mean defined by the PMF:

$$\mu = 1(\theta) + 2(1-\theta) = 2-\theta$$

$$\hat{\mu} = \bar{X}_n = \frac{1}{n}\sum X_i = \frac{1}{5}(7) = \frac{7}{5}$$

} produces linear system with 1 eq. 1 unknown

$$\mu \hat{=} \hat{\mu} = \frac{7}{5} = 2-\theta \Rightarrow \boxed{\hat{\theta}_{MOM} = 3/5}$$

(b) $L(\theta) = \prod_{i=1}^{5} f(x_i;\theta)$, given $X_1 \cdots X_5$ are an i.i.d sample $\sim f(x;\theta)$:

$\Rightarrow L(\theta) = \theta^3(1-\theta)^2$

$\Rightarrow \log(L(\theta)) = 3\log(\theta) + 2\log(1-\theta)$, to find MLE find max of $\log(L(\theta))$ by finding $\frac{d}{d\theta}\log(L(\theta)) = 0$

$$\Rightarrow \frac{d}{d\theta}\log(L(\theta)) = \frac{3}{\theta} - \frac{2}{1-\theta} = 0 \Rightarrow$$

$$\Rightarrow 3-3\theta = 2\theta \Rightarrow \theta = \frac{3}{5}$$

2nd derivative test $\Rightarrow \frac{d^2}{d\theta}\log(L(\theta)) = -\frac{3}{\theta^2} - \frac{2}{(1-\theta)^2} < 0$

$\Rightarrow$ second derivative test satisfied $\frac{d}{d\theta}\log(L(\theta)) = 0$ finds maximum

Thus we conclude $\hat{\theta}_{MLE} = 3/5$