PSETs Landing Page*

Anish Krishna Lakkapragada

This is the documentation for using my PSET PDFs responsibly. I post these LaTEX'd PSETs (1) as an education resource for friends at other universities, fellow Yalies, and all those interested and (2) for quick reference. These PSETs are not to be used irresponsibly; only look at the solution after giving each problem an honest attempt. **If YOU USE THESE PSETS TO CHEAT, YOU ARE NOT ONLY STUPID BUT YOU ARE CHEATING YOURSELF OUT OF THE ABILITY TO FALL IN LOVE WITH MATH.** Furthermore, I am not smarter than you and my solutions did not always get a perfect score.

The general format for accessing the (one-indexed) `N`th assigned PSET PDF of a Yale course with course number `CODE` is:

`https://anish.lakkapragada.com/notes/TYPE-CODE/psets/N.pdf`

where `TYPE` is `stats` or `math`. Similarily, to access my solution for this PSET you can go to:

`https://anish.lakkapragada.com/notes/TYPE-CODE/sols/N.pdf`

These PSETs and associated solution PDFs are synchronized daily at 4:20AM with my computer files through a Cronjob Shell Script. If you want to contribute any corrections, please email `anish.lakkapragada@yale.edu`.

*Note that PDF here is referring to Portable Document Format, not to be confused with the veritable Probability Density Function.

STATS 242 HW 1

January 22, 2025

Number of late days: 0; Collaborators: None.

1.

a) For $i = 1 \dots \frac{n}{2}$, let the Bernoulli variables $F_i \sim Bern(\theta + \delta)$ and $M_i \sim Bern(\theta - \delta)$ represent whether the ith female or male (respectively) voter supports Harris. Therefore, $\hat{\theta}$ is given by:

$$\hat{\theta} = \frac{\Sigma_{i=1}^{\frac{n}{2}} F_i + M_i}{n}$$

and so we have that $\mathbb{E}[\hat{\theta}]$ is given by:

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\frac{\Sigma_{i=1}^{\frac{n}{2}} F_i + M_i}{n}] = \frac{1}{n}\Sigma_{i=1}^{\frac{n}{2}}\mathbb{E}[F_i + M_i] = \frac{1}{n}\Sigma_{i=1}^{\frac{n}{2}}\mathbb{E}[F_i] + \mathbb{E}[M_i] = \frac{1}{n}[\frac{n}{2}(\theta + \delta + \theta - \delta)]$$
$$= \frac{1}{n}[\frac{n}{2}(2\theta)] = \frac{n\theta}{n} = \theta$$

Because $\mathbb{E}[\hat{\theta}] = \theta$, the bias of $\hat{\theta}$ is zero.

b)

$$Var(\hat{\theta}) = \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\theta]^2$$

Let random variables $\hat{\theta}_F = \frac{1}{0.5n}\Sigma_{i=1}^{\frac{n}{2}} F_i$ and $\hat{\theta}_M = \frac{1}{0.5n}\Sigma_{i=1}^{\frac{n}{2}} M_i$ define the sampling parameters from the female and male independent simple random samples, respectively. Thus, we have that $\hat{\theta} = \frac{(\hat{\theta}_F + \hat{\theta}_M)}{2}$. Note $\hat{\theta}_F$ and $\hat{\theta}_M$ are independent as their respective simple random samples are independent. Thus, we can compute $Var(\hat{\theta}) = Var(\frac{\hat{\theta}_F + \hat{\theta}_M}{2}) = \frac{1}{4}[Var(\hat{\theta}_F) + Var(\hat{\theta}_M)]$. Because both sampling parameters $\hat{\theta_F}$ and $\hat{\theta}_M$ are drawn from simple random samples, their respective variances can be given by $\frac{(\theta+\delta)(1-\theta-\delta)}{0.5n}(1 - \frac{0.5n-1}{0.5N-1})$ and $\frac{(\theta-\delta)(1-\theta+\delta)}{0.5n}(1 - \frac{0.5n-1}{0.5N-1})$. Thus, putting it all together:

$$Var(\hat{\theta}) = \frac{1}{4}[Var(\hat{\theta}_F) + Var(\hat{\theta}_M)] = \frac{1}{4}\frac{(\theta+\delta)(1-\theta-\delta)+(\theta-\delta)(1-\theta+\delta)}{0.5n}(1-\frac{0.5n-1}{0.5N-1}) =$$
$$\frac{(\theta-\theta^2-\delta^2)}{n}\frac{0.5N-0.5n}{0.5N-1} = \frac{(\theta-\theta^2-\delta^2)}{n}\frac{N-n}{N-2}$$

c) We first simplify the given quantity in this question to:

$$Var[\hat{\theta}] = \frac{\theta-\theta^2}{n}(1-\frac{n-1}{N-1}) = \frac{\theta-\theta^2}{n}(\frac{N-n}{N-1})$$

Given N is significantly larger than n, both $\frac{N-n}{n-1}$ and $\frac{N-n}{N-2}$ are approximately equal to one. This means the variance computed in part (b) is approximately $\frac{\delta^2}{n}$ smaller than the quantity given in this question.

2.

a) For $i = 1 \dots n$, let X_i be a Bernoulli variable that models whether the ith person in the survey support Harris. This probability is equivalent to the expected percentage of the sample that supports Harris, or the expected number of people in the survey who support Harris divided by the total number of people in the survey.

The number of people who support Harris in the survey can be given by θN, the number of people who support Harris in the population, times $\frac{p}{N}$ (probability each person who support Harris in the population was chosen.)

The number of people who support other candidates in the survey can be given by $(1-\theta)N$, the number of people who support other candidates in the population, times $\frac{q}{N}$ (probability each person who supports another candidate in the population was chosen.)

Thus, $\mathbb{E}[X_i]$ is given by:

$$\mathbb{E}[X_i] = P(X_i = 1) = \frac{\theta N \cdot \frac{p}{N}}{\theta N \cdot \frac{p}{N} + (1-\theta)N \cdot \frac{q}{N}} = \frac{p\theta}{p\theta + (1-\theta)q}$$

Given $p\theta + (1-\theta)q = 1$, we have that:

$$\mathbb{E}[X_i] = p\theta$$

Given this, we can compute the bias of $\hat{\theta}$ as:

$$\mathbb{E}[\hat{\theta}] - \theta = \mathbb{E}[\frac{1}{n}\Sigma_{i=1}^{n} X_i] - \theta = \frac{1}{n} \cdot n \cdot \mathbb{E}[X_1] - \theta = p\theta - \theta$$

For $\hat{\theta}$ to be unbiased, $p\theta - \theta = 0$, or p (and q) has to equal to one.

b) Given $Var(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = p\theta(1-p\theta)$, we can use the Central Limit Theorem to approximate $\hat{\theta} \sim \mathcal{N}(\mathbb{E}[X_i], \frac{Var(X_i)}{n})$ or $\hat{\theta} \sim \mathcal{N}(p\theta, \frac{p\theta - p^2\theta^2}{n})$. Plugging in $\theta = 0.5, p = 1.02, q = 0.98$, we have:

i. $P(\hat{\theta} > 0.5) = 0.5793$ for $n = 100$

ii. $P(\hat{\theta} > 0.5) = 0.7365$ for $n = 1000$

iii. $P(\hat{\theta} > 0.5) = 0.97727$ for $n = 10000$

As n increases, $P(\hat{\theta} > 0.5)$ increases.

3.

22 We first compute the joint distribution $f_{X,Y}(x,y)$. Because (X,Y) is uniformly distributed over the region R where $x^2 + y^2 \leq 1$, we have $\iint_R f_{X,Y}(x,y) = 1$. Note that R forms a circle with radius one, and so its area is pi. Because $f_{X,Y}(x,y)$ is constant over region R as (X,Y) is uniformly distributed, $\iint_R f_{X,Y}(x,y) = Area_R \cdot f_{X,Y} = 1$, or that for $(x,y) \in R$, $f_{X,Y}(x,y) = \frac{1}{\pi}$. Putting it all together, we have:

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & \text{for } x^2 + y^2 \leq 1 \\ 0 & \text{else} \end{cases}$$

We now are ready to compute $Cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. We compute $\mathbb{E}[XY], \mathbb{E}[X]$, and $\mathbb{E}[Y]$ below:

(1) $\mathbb{E}[XY]$

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dxdy = \int_{-1}^{1}\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{xy}{\pi}dxdy = \int_{-1}^{1} \frac{x^2y}{2\pi}\Big|_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} dxdy$$
$$= \int_{-1}^{1} 0dxdy = 0$$

(2) $\mathbb{E}[X]$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{X,Y}(x,y)dxdy = \int_{-1}^{1}\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{x}{\pi}dxdy = \int_{-1}^{1} \frac{x^2}{2\pi}\Big|_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} dxdy = \int_{-1}^{1} 0dx = 0$$

(3) $\mathbb{E}[Y]$

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{X,Y}(x,y)dxdy = \int_{-1}^{1}\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} \frac{y}{\pi}dxdy = \int_{-1}^{1} \frac{xy}{\pi}\Big|_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} dxdy =$$
$$\int_{-1}^{1} 2\frac{y\sqrt{1-y^2}}{\pi}dy = \frac{2}{\pi}\int_{-1}^{1} y\sqrt{1-y^2}dy$$

Because $y\sqrt{1-y^2}$ is an odd function, it's integral from -1 to 1 will be 0. Thus:

$$\mathbb{E}[Y] = \frac{2}{\pi}\int_{-1}^{1} y\sqrt{1-y^2}dy = 0$$

Thus, $Cov[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot 0 = 0$. X and Y are not independent because their joint distribution (X, Y) is not defined over a rectangular region, but instead a circle. As a simple demonstration of this idea, let us assume that X and Y are independent. If so, then:

$$P(X > 0.5|Y > 0.9) = P(X > 0.5)$$

Note that $P(X > 0.5)$ is obviously nonzero. However, $P(X > 0.5|Y > 0.9) = 0$ as if $Y > 0.9$, then in order for $x^2 + y^2 \leq 1$, x must be in the range $[-\sqrt{1 - 0.9^2}, \sqrt{1 - 0.9^2}]$ and so x cannot be greater than 0.5. Thus, we have shown $P(X > 0.5|Y > 0.9) = 0 \neq P(X > 0.5)$ and so X and Y are demonstrated to not be independent.

4.

Note that because X is a normal distribution, it is symmetric and so $P(X > 0) = 0.5$.

$$P(X + Y > 0|X > 0) = \frac{P(X + Y > 0 \cap X > 0)}{P(X > 0)} = 2P(X + Y > 0 \cap X > 0)$$

The event $X + Y > 0 \cap X > 0$ can be given by $(X > 0 \cap Y > 0) \cup (X > -Y \cap Y < 0)$. Thus, we have that:

$$P(X + Y > 0|X > 0) = 2P((X > 0 \cap Y > 0) \cup (X > -Y \cap Y < 0))$$
$$= 2[P(X > 0 \cap Y > 0) + P(X > -Y \cap Y < 0)] = 2[P(X > 0)P(Y > 0) + P(X > -Y \cap Y < 0)]$$
$$= 2[0.5(0.5) + P(X > -Y \cap Y < 0)]$$

$P(X > -Y \cap Y < 0)$ takes up half of the space in the third quadrant, and because the joint PDF of (X, Y) is rotationally symmetric about the origin, we know that $P(X > -Y \cap Y < 0) = 0.125$. Thus:

$$P(X + Y > 0|X > 0) = 2[0.25 + 0.125] = 2[0.375] = 0.75$$