Gov 2001: Problem Set 2

Spring 2025

February 17, 2025

1 Variance

1.1 Alternative expression

Prove that we can write variance as below:

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

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[(\mathbb{E}[X])^2].\end{aligned}$$

Since $\mathbb{E}[X]$ is a constant, its expectation remains the same:

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

Substituting this back:

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.\end{aligned}$$

Thus, we have proven the variance formula:

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

1.2 Counter example

Show an example where $V[X+Y] = V[X] + V[Y]$ does not hold, and explain why.

when they are dependent. such as $Y = 2X$, then we have $V[X+Y] = V(3X) = 9V(X)$, but $V[X] + V[Y] = V(X) + 4V(X) = 5V(X)$

2 Ranking Continuous R.V.s

Let $X_1, \dots, X_n$ be i.i.d. from a continuous distribution. For any permutation $a_1, a_2, \dots, a_n$ of $1, 2, \dots, n$, calculate the following probability:

$$\mathbb{P}(X_{a_1} < X_{a_2} < \cdots < X_{a_n})$$

$\frac{1}{n!}$, all orderings are equally likely. pmf for each individual point is 0.

2.1 Discrete R.V.

Would your result hold for discrete R.V.s? Why?

No. Pmf of drvs are not evenly distributed. For example, we need to remove the cases where any two or more rvs are equal, before dividing by $n!$

3 Binomial Distribution

A common approximation for a Binomial random variable $X \sim \text{Bin}(n, p)$ is a Poisson random variable $Y \sim \text{Pois}(\lambda)$ where $\lambda = n \cdot p$. One of the reasons for this approximation is that the expectations of Binomial and Poisson random variables match (both np, in our case). We will explore this a bit further.

(a) Holding n fixed, would Y better approximate X if $p = \frac{1}{10}$ or if $p = \frac{1}{2}$? Compare the variances of X and Y by checking the difference and ratio of the two variances.

- **Step 1: Compute Variances**
  - For a Binomial random variable $X \sim \text{Bin}(n, p)$, the variance is:

    $$\mathbb{V}[X] = np(1-p).$$

  - For a Poisson random variable $Y \sim \text{Pois}(\lambda)$ with $\lambda = np$, the variance is:

    $$\mathbb{V}[Y] = \lambda = np.$$

- **Step 2: Compare Variances**
  - **Variance Difference:**

    $$\begin{aligned}\mathbb{V}[X] - \mathbb{V}[Y] &= np(1-p) - np \\ &= np - np^2 - np \\ &= -np^2.\end{aligned}$$

    Since $np^2 \geq 0$, the variance of the Binomial distribution is always less than or equal to that of the Poisson approximation.

- **Variance Ratio:**

$$\begin{aligned}\frac{\mathbb{V}[X]}{\mathbb{V}[Y]} &= \frac{np(1-p)}{np} \\ &= 1-p.\end{aligned}$$

- **Step 3: Compare for $p = \frac{1}{10}$ and $p = \frac{1}{2}$**

  - **For $p = \frac{1}{10}$:**

$$\begin{aligned}\mathbb{V}[X] - \mathbb{V}[Y] &= -n\left(\frac{1}{10}\right)^2 = -\frac{n}{100}, \\ \frac{\mathbb{V}[X]}{\mathbb{V}[Y]} &= 1 - \frac{1}{10} = \frac{9}{10} = 0.9.\end{aligned}$$

  - **For $p = \frac{1}{2}$:**

$$\begin{aligned}\mathbb{V}[X] - \mathbb{V}[Y] &= -n\left(\frac{1}{2}\right)^2 = -\frac{n}{4}, \\ \frac{\mathbb{V}[X]}{\mathbb{V}[Y]} &= 1 - \frac{1}{2} = \frac{1}{2} = 0.5.\end{aligned}$$

- **Conclusion**

  - The variance difference is smaller (closer to zero) for $p = \frac{1}{10}$, meaning the Poisson variance better approximates the Binomial variance in this case.
  - The variance ratio is closer to 1 when $p = \frac{1}{10}$, suggesting that the relative difference in variance is smaller, making the approximation more accurate.

  Thus, the Poisson approximation $Y \sim \text{Pois}(\lambda)$ **better approximates $X \sim \text{Bin}(n, p)$ when $p = \frac{1}{10}$ compared to when $p = \frac{1}{2}$** because the variance of the Poisson distribution is closer to that of the Binomial distribution when p is small.

(b) Holding p fixed, would Y better approximate X if n = 1,000 or if n = 100,000? Compare the variances of X and Y by checking the difference and ratio of the two variances.

- **Step 1: Compare for $n = 1,000$ and $n = 100,000$**

  - **For $n = 1,000$:**

$$\begin{aligned}\mathbb{V}[X] - \mathbb{V}[Y] &= -1000p^2, \\ \frac{\mathbb{V}[X]}{\mathbb{V}[Y]} &= 1-p.\end{aligned}$$

- **For $n = 100,000$:**

$$\mathbb{V}[X] - \mathbb{V}[Y] = -100,000p^2,$$
$$\frac{\mathbb{V}[X]}{\mathbb{V}[Y]} = 1 - p.$$

- **Conclusion**
  - The absolute variance difference, $-np^2$, increases as n increases, meaning that the Poisson variance deviates more from the Binomial variance for larger n.
  - However, the ratio $\frac{\mathbb{V}[X]}{\mathbb{V}[Y]} = 1 - p$ remains unchanged, showing that the relative difference does not depend on n.
  - The Poisson approximation is generally better when n is large, because as $n \to \infty$ and $p \to 0$ in such a way that np remains constant, the Binomial distribution converges to the Poisson distribution.

  Thus, the Poisson approximation $Y \sim \text{Pois}(\lambda)$ **better approximates $X \sim \text{Bin}(n, p)$ when $n = 100,000$ compared to when $n = 1,000$**, especially when p is small, because the Binomial distribution approaches the Poisson distribution in this limit.

4 Expected Value

Suppose you're interested in studying the distribution of political ideology in the US, a random variable that we'll call X. Individuals are placed on a continuous one-dimensional ideology scale that varies from -1 to 1, where lower score are more liberal. Instead of having to do sampling to estimate the distribution, the Data Generating God comes to you in a dream and tells you the unnormalized distribution of this random variable, which is as follows:

$$f(x) = \begin{cases} c(\frac{1}{2}x + 1) & -1 \leq x \leq 1, \\ 0 & \text{else} \end{cases}$$

where c is a normalizing constant[1].

(a) Find the value of c that would make $f(x)$ a valid probability distribution function, $f_X(x)$.

- A valid probability density function must satisfy the condition:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

  Since $f(x)$ is zero outside $[-1, 1]$, we compute:

$$\int_{-1}^{1} c\left(\frac{1}{2}x + 1\right) dx = 1.$$

[1] Note that this distribution is completely made up and is not based on actual data.

- Evaluating the integral:

$$c \int_{-1}^{1} \left(\frac{1}{2}x + 1 \right) dx = 1.$$

- Compute the integral:

$$\int \left(\frac{1}{2}x + 1 \right) dx = \frac{1}{2}\frac{x^2}{2} + x = \frac{x^2}{4} + x.$$

- Evaluating from -1 to 1:

$$\begin{aligned} \left(\frac{1^2}{4} + 1 \right) - \left(\frac{(-1)^2}{4} + (-1) \right) &= \left(\frac{1}{4} + 1 \right) - \left(\frac{1}{4} - 1 \right) \\ &= \left(\frac{5}{4} \right) - \left(-\frac{3}{4} \right) \\ &= \frac{5}{4} + \frac{3}{4} = \frac{8}{4} = 2. \end{aligned}$$

- Solving for c:

$$\begin{aligned} c \cdot 2 &= 1 \\ c &= \frac{1}{2}. \end{aligned}$$

Thus, the normalized PDF is:

$$f_X(x) = \begin{cases} \frac{1}{2}\left(\frac{1}{2}x + 1\right), & -1 \le x \le 1, \\ 0, & \text{else.} \end{cases}$$

(b) Calculate $E[X]$.

- The expectation is given by:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

  Since $f_X(x)$ is zero outside $[-1, 1]$, we compute:

$$\mathbb{E}[X] = \int_{-1}^{1} x \cdot \frac{1}{2} \left(\frac{1}{2}x + 1 \right) dx.$$

- Expanding:

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{2} \int_{-1}^{1} x \left(\frac{1}{2}x + 1 \right) dx \\ &= \frac{1}{2} \int_{-1}^{1} \left(\frac{1}{2}x^2 + x \right) dx. \end{aligned}$$

- Compute the individual integrals:

$$\int x^2 dx = \frac{x^3}{3}, \quad \int x dx = \frac{x^2}{2}.$$

- Evaluating from -1 to 1:

$$\left[\frac{x^3}{3}\right]_{-1}^{1} = \left(\frac{1^3}{3} - \frac{(-1)^3}{3}\right) = \left(\frac{1}{3} - \left(-\frac{1}{3}\right)\right) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3},$$
$$\left[\frac{x^2}{2}\right]_{-1}^{1} = \left(\frac{1^2}{2} - \frac{(-1)^2}{2}\right) = \left(\frac{1}{2} - \frac{1}{2}\right) = 0.$$

- Substituting back:

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{2}\left(\frac{1}{2} \cdot \frac{2}{3} + 0\right) \\ &= \frac{1}{2} \times \frac{1}{3} \\ &= \frac{1}{6}.\end{aligned}$$

Thus, the expected value of X is:

$$\mathbb{E}[X] = \frac{1}{6}.$$

5 Count Data

Suppose $X \sim \text{Pois}(\lambda)$, where λ is fixed but unknown.

An estimator is a function of the data and the **bias** of an estimator, $f(X)$, is defined as $E[f(X)] - \theta$, where θ is the **estimand** (an unknown quantity we would like to estimate from the observable data).

For instance our estimand could be λ, and we know by the properties of a Poisson random variable that the bias of the estimator, $f(X) = X$, is $E[X] - \lambda = \lambda - \lambda = 0$. We call an estimator with 0 bias an **unbiased** estimator.

For this question, suppose that our estimand is λ^3 rather than λ.

(a) Show that X^3 is **not** an unbiased estimator of λ^3 and specify the bias as a function of λ.

Hints:

1. You may use the following result: if $X \sim \text{Pois}(\lambda)$, then $E[X \cdot g(X)] = \lambda E[g(X+1)]$ for any function $g(\cdot)$.

2. You may use the result for $E[X^2]$ derived in lecture and section, (i.e., no need to derive it again).

3. Remember following operation using alternative representation of variance:

$$\mathbb{E}[X^3] = \lambda\mathbb{E}[(X+1)^2] = \lambda[\mathrm{Var}(X+1) + (\mathbb{E}[X+1])^2] = \lambda(\lambda+\lambda^2+2\lambda+1) = \lambda^3+3\lambda^2+\lambda$$

We are given that $X \sim \mathrm{Pois}(\lambda)$, meaning the expectation of X is:

$$\mathbb{E}[X] = \lambda.$$

The bias of an estimator $f(X)$ for estimating λ^3 is defined as:

$$\mathrm{Bias}(f(X)) = \mathbb{E}[f(X)] - \lambda^3.$$

For this problem, we choose $f(X) = X^3$, so we need to compute $\mathbb{E}[X^3]$.

- **Step 1: Compute $\mathbb{E}[X^3]$**

  We can use the hint, let $g(X) = X^2$ and given that $\mathbb{E}[X] = \mathrm{Var}(X) = \lambda$:

  $$\mathbb{E}[X^3] = \lambda\mathbb{E}[(X+1)^2] = \lambda[\mathrm{Var}(X+1) + (\mathbb{E}[X+1])^2] = \lambda(\lambda+\lambda^2+2\lambda+1) = \lambda^3+3\lambda^2+\lambda$$

  Or the standard method: The moments of a Poisson-distributed random variable are known from properties of factorial moments. Specifically, the third raw moment of a Poisson random variable is:

  $$\mathbb{E}[X^3] = \lambda^3 + 3\lambda^2 + \lambda.$$

  This follows from the general formula for moments of a Poisson variable:

  $$\mathbb{E}[X^r] = \sum_{k=0}^{\infty} k^r \frac{e^{-\lambda}\lambda^k}{k!}.$$

  Using standard derivations or lookup tables, we obtain:

  $$\mathbb{E}[X^3] = \lambda^3 + 3\lambda^2 + \lambda.$$

- **Step 2: Compute the Bias**

  The bias is given by:

  $$\begin{aligned}\mathrm{Bias}(X^3) &= \mathbb{E}[X^3] - \lambda^3\\ &= (\lambda^3 + 3\lambda^2 + \lambda) - \lambda^3\\ &= 3\lambda^2 + \lambda.\end{aligned}$$

- **Conclusion**

  Since $\text{Bias}(X^3) = 3\lambda^2 + \lambda \neq 0$, the estimator X^3 is **not** an unbiased estimator of λ^3. Instead, it overestimates λ^3 by an amount equal to $3\lambda^2 + \lambda$.

(b) Suppose $\lambda = 5$. Use $150,000$ simulations to validate your result to part (a). That is, calculate the bias of you estimator from both the simulations and the analytical results. Let's use CGIS Zipcode for seed.

```
   set.seed(02138)

# Given parameters
lambda_val <- 5
num_simulations <- 150000

# Simulate X from Poisson(lambda)
X_samples <- rpois(num_simulations, lambda_val)

# Compute the estimator f(X) = X^3
X_cubed_samples <- X_samples^3

# Compute the simulated expectation E[X^3]
E_X3_simulated <- mean(X_cubed_samples)

# Compute analytical expectation E[X^3] = lambda^3 + 3*lambda^2 + lambda
E_X3_analytical <- lambda_val^3 + 3 * lambda_val^2 + lambda_val

# Compute bias: Bias(X^3) = E[X^3] - lambda^3
bias_simulated <- E_X3_simulated - lambda_val^3
bias_analytical <- E_X3_analytical - lambda_val^3
```

Analytical: 80 , simulated: 79.17433 .
Note, this will be different if you use python / other software that is not R.