

# 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`. Similarly, 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 2

January 29, 2025

---

**Number of late days: 0; Collaborators: None.**

1.

For  $1 \leq i \leq n$ , let us define r.v.  $B_i \sim \text{Bern}(p)$ , where for  $i \neq j$ ,  $B_i$  and  $B_j$  are independent. Then, we have that  $X = \sum_{i=1}^n B_i$ . Because all  $B_i$ s are independent, we have that:

$$M_X(t) = M_{\sum_{i=1}^n B_i}(t) = \prod_{i=1}^n M_{B_i}(t)$$

The MGF of r.v.  $B_i$  can be computed as:  $M_{B_i}(t) = \mathbb{E}[e^{tB_i}] = e^{t(1)}P(B_i = 1) + e^{t(0)}P(B_i = 0) = pe^t + (1 - p) = 1 + p(e^t - 1)$ . Thus, we have  $M_X(t)$  as:

$$M_X(t) = \prod_{i=1}^n M_{B_i}(t) = \prod_{i=1}^n [1 + p(e^t - 1)] = [1 + p(e^t - 1)]^n$$

2.

We first start by computing the distributions of  $X_1$  and  $X_2$ . Note that because  $Z_1$  and  $Z_2$  are independent normal distributions, their sum forms a normal distribution as well.

## 1. Distribution of $X_1$

Since  $c_1 Z_1 \sim \mathcal{N}(0, c_1^2)$  and  $d_1 Z_1 \sim \mathcal{N}(0, d_1^2)$ ,  $c_1 Z_1 + d_1 Z_2 \sim \mathcal{N}(0, c_1^2 + d_1^2)$ . Finally,  $e_1$  is just a constant and so it doesn't affect the variance so  $X_1 = c_1 Z_1 + d_1 Z_2 + e_1 \sim \mathcal{N}(e_1, c_1^2 + d_1^2)$ . Therefore we can set  $e_1 = \mu_1$ , the mean of  $X_1$ . Furthermore, the variance  $\sigma_1^2$  of  $X_1$  is equal to  $c_1^2 + d_1^2$ .

## 2. Distribution of $X_2$

We use identical reasoning as with before.  $c_2 Z_1 \sim \mathcal{N}(0, c_2^2)$  and  $d_2 Z_2 \sim \mathcal{N}(0, d_2^2)$ , so we have  $c_2 Z_2 + d_2 Z_2 \sim \mathcal{N}(0, c_2^2 + d_2^2)$ . Thus,  $X_2 = c_2 Z_2 + d_2 Z_2 + e_2 \sim \mathcal{N}(e_2, c_2^2 + d_2^2)$ . Therefore we can set  $e_2 = \mu_2$ , the mean of  $X_2$ . Furthermore, the variance  $\sigma_2^2$  of  $X_2$  is equal to  $c_2^2 + d_2^2$ .

We now have  $c_1, c_2, d_1, d_2$  remaining to assign. We compute the correlation  $\rho$  between  $X_1$  and  $X_2$ , starting by computing the covariance between  $X_1$  and  $X_2$ :

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \text{Cov}(c_1 Z_1 + d_1 Z_2 + e_1, c_2 Z_1 + d_2 Z_2 + e_2) \\ &= c_1 \text{Cov}(Z_1, c_2 Z_1 + d_2 Z_2 + e_2) + d_1 \text{Cov}(Z_2, c_2 Z_1 + d_2 Z_2 + e_2) + \text{Cov}(e_1, c_2 Z_1 + d_2 Z_2 + e_2)\end{aligned}$$

Note that because  $e_1$  is a fixed constant,  $\text{Cov}(e_1, c_2 Z_1 + d_2 Z_2 + e_2) = 0$ . Also note that because  $Z_1$  and  $Z_2$  are independent,  $\text{Cov}(Z_1, Z_2) = 0$ .

$$\begin{aligned}\text{Cov}(X_1, X_2) &= c_1 [c_2 \text{Cov}(Z_1, Z_1) + d_2 \text{Cov}(Z_1, Z_2) + \text{Cov}(Z_1, e_2)] + d_1 \text{Cov}(Z_2, c_2 Z_1 + d_2 Z_2 + e_2) \\ &= c_1 [c_2 + 0 + 0] + d_1 [c_2 \text{Cov}(Z_2, Z_1) + d_2 \text{Cov}(Z_2, Z_2) + \text{Cov}(Z_2, e_2)] = c_1 c_2 + d_1 [0 + d_2 + 0] \\ &= c_1 c_2 + d_1 d_2\end{aligned}$$

and now we compute the correlation  $\rho$  between  $X_1$  and  $X_2$  as:

$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}} = \frac{c_1 c_2 + d_1 d_2}{\sigma_1 \sigma_2}$$

This gives us three equations:

$$\begin{aligned}\rho \sigma_1 \sigma_2 &= c_1 c_2 + d_1 d_2 \\ c_1^2 + d_1^2 &= \sigma_1^2 \\ c_2^2 + d_2^2 &= \sigma_2^2\end{aligned}$$

to solve for four variables. As such, we arbitrarily choose to set  $c_1 = 0$ , giving us  $d_1 = \sigma_1$  and  $\rho \sigma_1 \sigma_2 = d_1 d_2 = \sigma_1 d_2 \implies d_2 = \rho \sigma_2$ . Finally, we can solve for  $c_2^2 = \sigma_2^2 - d_2^2 = \sigma_2^2 - \rho^2 \sigma_2^2 = \sigma_2^2 (1 - \rho^2) \implies c_2 = \sigma_2 \sqrt{1 - \rho^2}$ .

As a summary of my answer,

$$\begin{aligned}e_1 &= \mu_1 \\ e_2 &= \mu_2 \\ c_1 &= 0 \\ c_2 &= \sigma_2 \sqrt{1 - \rho^2} \\ d_1 &= \sigma_1 \\ d_2 &= \rho \sigma_2\end{aligned}$$

3.

- (a) We first show  $\mathbb{E}[\hat{I}_n(f)] = I(f)$ . Note that  $\frac{f(X_i)}{g(X_i)}$  is a random variable and so its expectation is given to us by  $\mathbb{E}[\frac{f(X_i)}{g(X_i)}] = \int_{-\infty}^{\infty} \frac{f(u)}{g(u)} g(u) du = \int_a^b \frac{f(u)}{g(u)} g(u) du = \int_a^b f(u) du = I(f)$ . Thus:

$$\mathbb{E}[\hat{I}_n(f)] = \mathbb{E}[\frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)}] = \frac{1}{n} \cdot n \cdot \mathbb{E}[\frac{f(X_i)}{g(X_i)}] = I(f)$$

Notice that  $\hat{I}_n(f)$  is essentially an average of  $n$  random variables, each of with an expectation of  $I(f)$ . Thus,  $\hat{I}_n(f) \rightarrow I(f)$  in probability as  $n \rightarrow \infty$  due to the (Weak) Law of Large Numbers.

- b) We first compute  $\text{Var}[\hat{I}_n(f)]$ . We begin by computing the variance of random variable  $\frac{f(X_i)}{g(X_i)}$ :

$$\text{Var}[\frac{f(X_i)}{g(X_i)}] = \mathbb{E}[(\frac{f(X_i)}{g(X_i)})^2] - \mathbb{E}[\frac{f(X_i)}{g(X_i)}]^2 = \int_a^b \frac{f^2(u)}{g^2(u)} g(u) du - I^2(f) = \int_a^b \frac{f^2(u)}{g(u)} du - I^2(f)$$

Let us call define this quantity to be  $\sigma^2 = \int_a^b \frac{f^2(u)}{g(u)} du - I^2(f) \in \mathbb{R}$ . and so we have:

$$\text{Var}[\hat{I}_n(f)] = \text{Var}[\frac{1}{n} \sum_{i=1}^n \frac{f(X_i)}{g(X_i)}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[\frac{f(X_i)}{g(X_i)}] = \frac{1}{n} \sigma^2$$

Let us define  $c_n = \frac{\sqrt{n}}{\sigma} \in \mathbb{R}$ . Thus, by the Central Limit Theorem, we have that:

$$c_n(\hat{I}_n(f) - I(f)) \rightarrow \mathcal{N}(0, 1) \text{ as } n \rightarrow \infty$$

- c) In this problem,  $a = 0$  and  $b = 1$ ,  $f(x) = \cos(2\pi x)$ , and  $g(x) = 1$  if  $x \in [0, 1]$  and 0 otherwise. We first compute  $I(f)$ :

$$I(f) = \int_0^1 \cos(2\pi x) dx = \sin(2\pi x) \Big|_0^1 = \sin(2\pi) - \sin(0) = 0 - 0 = 0$$

and then  $\sigma^2$ :

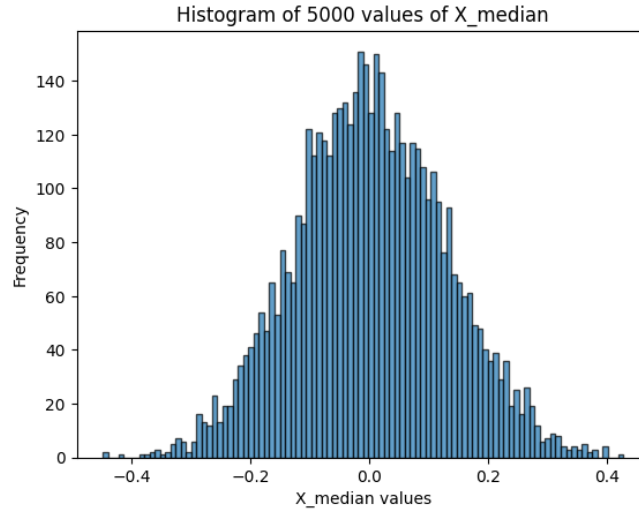
$$\begin{aligned} \sigma^2 &= \int_0^1 \frac{\cos^2(2\pi x)}{1} dx - I^2(f) = \int_0^1 [\frac{1}{2} + \frac{\cos(4\pi x)}{2}] dx = [\frac{x}{2} + \frac{\sin(4\pi x)}{8}] \Big|_0^1 \\ &= \frac{1}{2} + \frac{1}{8}(\sin(4\pi) - \sin(0)) = \frac{1}{2} \end{aligned}$$

and so  $c_n$  is given by:

$$c_n = \frac{\sqrt{n}}{\sigma} = \sqrt{\frac{n}{\sigma^2}} = \sqrt{2n}$$

4.

From my simulation, the mean and standard deviation of  $X_{median}$  are given by 0.00178 and 0.1265 respectively. Below is the histogram of the 5000 values of  $X_{median}$  from my simulation:



Based on the above histogram, we can see that the sampling distribution of  $X_{median}$  follows a normal distribution. We now derive the standard deviation of the sample mean  $\bar{X}$ . Note that 99 observations is enough for us to apply the Central Limit Theorem with confidence. Thus<sup>1</sup>,  $\bar{X} = \frac{X_1 + \dots + X_{99}}{99} \sim \mathcal{N}(\mathbb{E}[X_1], \frac{\text{Var}(X_1)}{99})$  or  $\bar{X} \sim \mathcal{N}(0, \frac{1}{99})$ . Thus the analytically-computed standard deviation of sample mean  $\bar{X}$  is  $\sqrt{\frac{1}{99}}$  or 0.1005, which is less than my calculated simulated standard deviation for  $X_{median}$  (0.1265). According to my simulation,  $X_{median}$  is more variable than  $\bar{X}$ .

```
1 # %%
2
3 # Run all imports first and then write helper functions.
4
5 import numpy as np
6 import math
7 import matplotlib.pyplot as plt
8
9 def get_N_obs_iid_standard_normal(N):
10     return np.random.normal(0, 1, N)
11
12 def compute_median(arr):
13     # assume arr is numpy
14     return np.median(arr)
```

<sup>1</sup>Note that  $X_1 \dots X_{99}$  are identical distributions and so they all have the same mean and variances.

```

15 # %%
16 N_SIMULATIONS = 5000
17 X_medians = []
18 for _ in range(N_SIMULATIONS):
19     samples = get_N_obs_iid_standard_normal(99)
20     X_median_curr = compute_median(samples)
21     X_medians.append(X_median_curr)
22
23 X_medians = np.array(X_medians)
24 X_median_mean = np.mean(X_medians)
25 X_median_std = np.std(X_medians)
26
27 print(f"X_median mean: {X_median_mean} and std: {X_median_std}")
28 # %%
29
30 """
31 Plot a histogram.
32 """
33
34 plt.hist(X_medians, bins=100, edgecolor='black', alpha=0.7)
35
36 # Add labels and title
37 plt.xlabel('X_median values')
38 plt.ylabel('Frequency')
39 plt.title('Histogram of 5000 values of X_median')
40
41 # %%

```