# 6 - Sample Average & Variance, Hypothesis Tests

| | |
|---|---|
| Topic | Hypothesis Testing, Sample Average & Variance |
| Comments | |
| Created | @January 31, 2022 4:09 PM |
| Date | @January 31, 2022 |
| Last Edited | @January 31, 2022 5:18 PM |

## Sample Average

**Sample Mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i$ are i.i.d. variables

Sample average is a random variable

Where $n$ is size of sample, $N$ is size of population

**Population Mean** $\mu_X = \frac{1}{N}\sum_{i=1}^{N} X_i$

The one and only population mean, population mean is not a random variable

Property 1: Sample Mean is “correct” since $\mathbb{E}[\bar{X}] = \mu_X$ “Sample mean is an unbiased estimator of the population mean”

Property 2: Sample Mean is a “good” estimator since the variance of Sample Mean (which helps us quantify uncertainty in the sample mean) goes down as $n$ increases (formally: $\frac{\sigma_X^2}{n} \to 0$ as $n \to \infty$)

$$\begin{aligned}
\mathrm{Var}(\bar{X}) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) \\
&= \frac{1}{n^2}\sum \mathrm{Var}(X_i) \text{ Due to Independence, we can extract } \frac{1}{n} \\
&= \frac{1}{n^2}\sum \sigma_X^2 \text{ By substitution of } \mathrm{Var}(X_i) \\
&= \frac{1}{n^2}(n\sigma_X^2) \text{ Simplifying the summation} \\
&= \frac{\sigma_X^2}{n} \text{ Cancelling } n \cdot \frac{1}{n}
\end{aligned}$$

Properties used:

$\mathrm{Var}(a + bX + cY) = \mathrm{Var}(bX) + \mathrm{Var}(cY) + 2\mathrm{Cov}(bX, cY) = b^2\mathrm{Var}(X) + c^2\mathrm{Var}(Y) + \cancel{2bc\mathrm{Cov}(X, Y)}$

If $X, Y$ are independent, then their covariance is 0. Otherwise, keep the covariance term

$\mathrm{Var}(X_i) = \sigma_X^2 \ \forall i$ Where $\forall$ is the symbol "for all"

## Sample Variance

Sample Variance: $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Population Variance: $\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{n}(X_i - \mu_X)^2$

We use sample variance to estimate population variance (This is the kind of proof that students should feel comfortable with)

$$
\begin{aligned}
\mathbb{E}[S_X^2] &= \mathbb{E}[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2] \text{ Applying Df. of } S_X \\
&= \frac{1}{n-1}\mathbb{E}[\sum X_i^2 - n\bar{X}^2] \text{ Linearity of expectation, and applying Property 1} \\
&= \frac{1}{n-1}(\sum \mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}^2]) \text{ Applying expectation} \\
&= \frac{1}{n-1}(n\mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X}^2]) \text{ Expanding summation} \\
&= \frac{1}{n-1}(n(\sigma_X^2 + \mu_X^2) - n(\frac{\sigma_X^2}{n} + \mu_X^2)) \text{ Using Property 3} \\
&= \frac{1}{n-1}(n\sigma_X^2 - \sigma_X^2) \\
&= \sigma_X^2
\end{aligned}
$$

Hence using $n-1$ ensures $S_X^2$ is an unbiased estimator

Properties:

1. $\sum(X_i - \bar{X})^2 = (\sum X_i^2) - n\bar{X}^2$
2. $\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$
3. $\mathbb{E}[X_i^2] = \sigma_X^2 + \mu_X^2$ By rearranging terms from Property 2

Note: $\sum \mathbb{E}[X] = n\mathbb{E}[X]$ since $\mathbb{E}[X]$ is just a number, summing up $n$ copies of it is equivalent to $n\mathbb{E}[X]$ (Many students are confused about this)

## Hypothesis Testing

Motivation: We want to test claims like “The average height of Cal students is 170cm” or “The height of the typical Cal student is not significantly different from national average”

**Null Hypothesis**: $H_0 : \mu_X = \mu_{X,0}$

$\mu_X$ is the true population parameter, what we are trying to estimate

$\mu_{X,0}$ is the hypothesized value, our best guess

**Alternative Hypothesis**: $H_a : \mu_X \neq \mu_{X,0}$

Assertion that the true population parameter is not equal to the hypothesized value

Example: If we think the average height is 170cm, then $H_0 : \mu_X = 170$ and $H_a : \mu_X \neq 170$

Once we get enough data, we run some analysis which supports either:

(i) Reject the null hypothesis

(ii) Fail to reject the null hypothesis (We never “prove” null hypothesis)

| Truth \ Decision | "Accept" $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ true | ✔️ | Type I error |
| $H_0$ false | Type II error | ✔️ |

Idea: We are always dealing with uncertainty. We hope that we are right (and we will be right most of the times if we do hypothesis testing properly), but Type I and Type II errors do occur from time to time

Size of the Test: $\mathbb{P}(\text{Type I error}) = \mathbb{P}(\text{Reject } H_0 | H_0 \text{ true})$

Also known as the “significance level” of the test, $= \alpha$

Power of the Test: $\mathbb{P}(\text{Reject } H_0 | H_0 \text{ false})$

Instructive Exercise: Try to use the table above to illustrate size and power