Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture # 12

Sampling Distributions and Estimation – Part I

Weighing a Random Sample

Bag Contains 100 Candies

Estimate total weight of candies by weighing a random sample of size 5 and multiplying the result by 20.

Your Chance to Win

The bag of candies and a digital scale will make their way around the room during the lecture. Each student gets a chance to draw 5 candies and weigh them.

Student with closest estimate wins the bag of candy!

Weighing a Random Sample

Procedure

When the bag and scale reach your team, do the following:

1. Fold the top of the bag over and shake to randomize.
2. Randomly draw 5 candies without replacement.
3. Weigh your sample and record the result in grams.
4. Calculate your estimate: 20 times the weight of your sample.
5. Replace your sample and shake again to re-randomize.
6. Pass bag and scale to next student.

What is this all about?

(Sample) Statistic

Any function of the data *alone*, e.g. sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Statistics

1. Estimation – Use data to construct educated guess (Estimate) about true value of population parameter.
2. Inference – Quantify uncertainty about estimate using:
   - Confidence Intervals
   - Hypothesis Testing

How does Probability connect to Statistics?

Random Sampling!

Random Sampling

(Simple) Random Sample

Select a sample of n objects from a population in such a way that:

1. Each member of the population has the same probability of being selected
2. The fact that one individual is selected does not affect the chance that any other individual is selected
3. Each sample of size n is equally likely to be selected

Random Sampling

In other words:

$$X_1, X_2, \ldots, X_n \sim \text{iid } f(x)$$

is a Random Sample

Statistics

Sample is drawn randomly, so sample statistics are *also random*.

Use what we know about probability theory to analyze the *distribution* of a statistic under random sampling.

Sample with or Without Replacement?

Strictly speaking, random samples should be drawn *with replacement*, otherwise there is *dependence*. In practice, this doesn't matter much so long as the population is large relative to the sample.

Candy Example In Progress: 100 is large relative to 5.

Estimator versus Estimate

Estimator

An estimator is a function $T(X_1, \ldots, X_n)$ of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

Estimator versus Estimate

Estimator

An estimator is a function $T(X_1, \ldots, X_n)$ of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

Sampling Distribution

The probability distribution of an Estimator is called a *sampling distribution*.

Estimator versus Estimate

Estimator

An estimator is a function $T(X_1, \ldots, X_n)$ of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

Sampling Distribution

The probability distribution of an Estimator is called a *sampling distribution*.

Estimate

An estimate is a function $T(x_1, \ldots, x_n)$ of the *observed data*, i.e. the *realizations* of the random variables we use to represent random sampling. An estimate is a *constant* since the observed data are *constants*

$\boxed{\text{Population: } f(x)}$ *Probability Distribution*

$$\boxed{\text{Population: } f(x)}$$

Probability Distribution

Random Sample of Size n

Random Variables

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$$

$\boxed{\text{Population: } f(x)}$ *Probability Distribution*

Random Sample of Size n

Random Variables $\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$

$\boxed{x_1^{(1)}, \ldots, x_n^{(1)}}$

Population: $f(x)$

Probability Distribution

Random Sample of Size n

Random Variables

$X_1, X_2, \ldots, X_n \sim \text{iid } f(x)$

$x_1^{(1)}, \ldots, x_n^{(1)}$

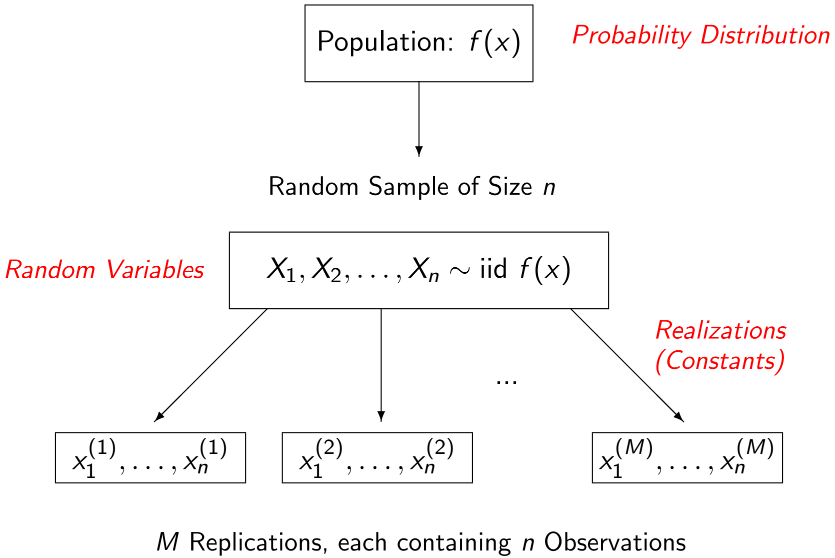
$x_1^{(2)}, \ldots, x_n^{(2)}$

Population: $f(x)$ *Probability Distribution*

Random Sample of Size n

Random Variables $X_1, X_2, \ldots, X_n \sim \text{iid } f(x)$

$x_1^{(1)}, \ldots, x_n^{(1)}$ $x_1^{(2)}, \ldots, x_n^{(2)}$ $\cdots$


Population: $f(x)$
Probability Distribution
Random Sample of Size n
Random Variables
$X_1, X_2, \ldots, X_n \sim \text{iid } f(x)$
Realizations
(Constants)
...
$x_1^{(1)}, \ldots, x_n^{(1)}$
$x_1^{(2)}, \ldots, x_n^{(2)}$
$x_1^{(M)}, \ldots, x_n^{(M)}$
M Replications, each containing n Observations

Random Sample of Size n

Random Variables $\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$

Random Sample of Size n

Random Variables

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \longrightarrow \boxed{T(X_1, \ldots, X_n)}$$

Estimator

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Random Sample of Size n

Estimator

Random Variables

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \longrightarrow \boxed{T(X_1, \ldots, X_n)}$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\downarrow$$

Observations (Data)

Constants

$$\boxed{x_1, \ldots, x_n}$$

Random Sample of Size n

Estimator

Random Variables

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \longrightarrow \boxed{T(X_1, \ldots, X_n)}$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Observations (Data)

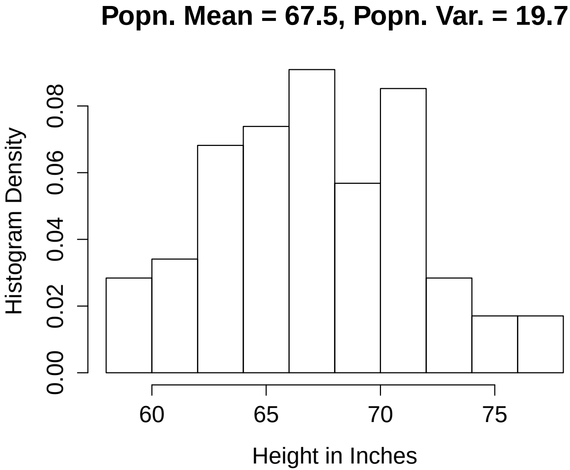
Estimate

Constants

$$\boxed{x_1, \ldots, x_n} \longrightarrow \boxed{T(x_1, \ldots, x_n)}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Population: All Students in the Class



Random Sample of Size n

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$$

Random Sample of Size n

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$$

$$\boxed{x_1^{(1)}, \ldots, x_n^{(1)}}$$

Random Sample of Size n

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$$

$$\downarrow$$

$$\boxed{x_1^{(1)}, \ldots, x_n^{(1)}}$$

$$\downarrow$$

$$\bar{x}_1$$

Random Sample of Size n

$$\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \\
\bar{x}_1 &
\end{tikzcd}$$

Random Sample of Size n

$$
\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] \\
\bar{x}_1 & \bar{x}_2
\end{tikzcd}
$$

Random Sample of Size n

$$\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] & \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] & \cdots \\
\bar{x}_1 & \bar{x}_2 & \cdots
\end{tikzcd}$$

Random Sample of Size n

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$$

$$\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \quad \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \quad \cdots \quad \boxed{x_1^{(M)}, \ldots, x_n^{(M)}}$$

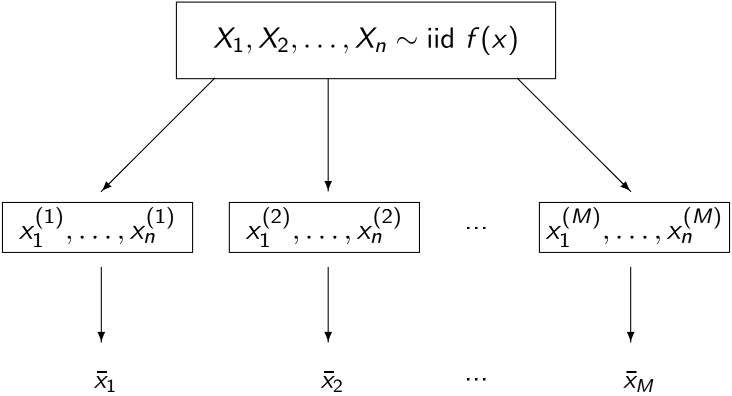
$$\bar{x}_1 \quad \bar{x}_2 \quad \cdots$$

Random Sample of Size n

$$\boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)}$$

$$\begin{array}{cccc}
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} & \cdots & \boxed{x_1^{(M)}, \ldots, x_n^{(M)}} \\
\downarrow & \downarrow & & \downarrow \\
\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_M
\end{array}$$

Random Sample of Size n



M Replications yield M *different estimates*

Random Sample of Size n

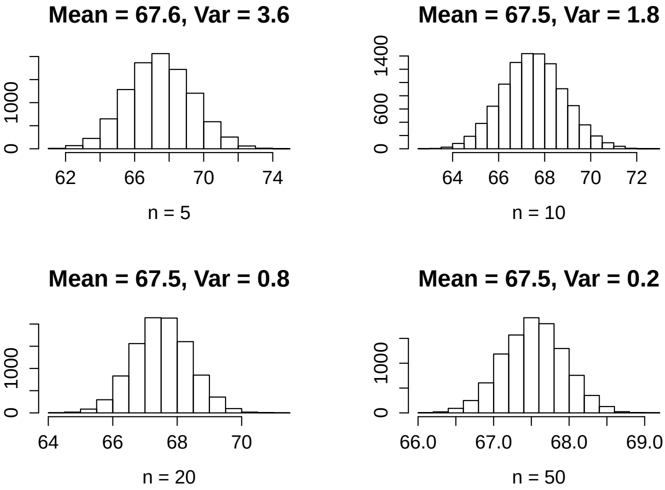
$$
\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \arrow[drr] & & \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] & \cdots & \boxed{x_1^{(M)}, \ldots, x_n^{(M)}} \arrow[d] \\
\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_M
\end{tikzcd}
$$

M Replications yield M *different estimates*

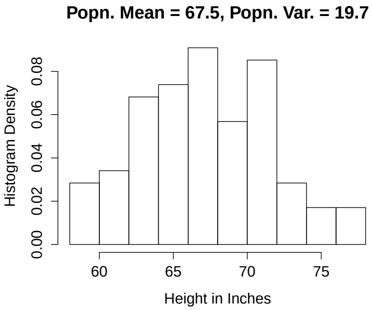
Sampling Distribution: Infinite Replications

Histograms of sampling distribution of sample mean $\bar{X}_n$

Random Sampling With Replacement, 10000 Reps. Each



Population Distribution vs. Sampling Distribution of $\bar{X}_n$



	Sampling Dist. of $\bar{X}_n$	
n	Mean	Variance
5	67.6	3.6
10	67.5	1.8
20	67.5	0.8
50	67.5	0.2

Two Things to Notice:

1. Sampling dist. "correct on average"
2. Sampling variability decreases with n

$X_1, \ldots, X_9 \sim$ iid with $\mu = 5$, $\sigma^2 = 36$.

Calculate:

$$E(\bar{X}) = E\left[\frac{1}{9}(X_1 + X_2 + \ldots + X_9)\right]$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] =$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \dots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu =$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} =$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu$$

Hence, sample mean is "correct on average." The formal term for this is *unbiased*.

$X_1, \ldots, X_9 \sim$ iid with $\mu = 5$, $\sigma^2 = 36$.

Calculate:

$$Var(\bar{X}) = Var\left[\frac{1}{9}(X_1 + X_2 + \ldots + X_9)\right]$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$Var[\bar{X}_n] \quad = \quad Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&=
\end{aligned}
$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 =
\end{aligned}
$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} =
\end{aligned}
$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned} Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned} Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Hence the variance of the sample mean *decreases linearly with sample size.*

$X_1, \ldots, X_9 \sim$ iid with $\mu = 5$, $\sigma^2 = 36$.

Calculate:

$$SD(\bar{X}) = SD\left[\frac{1}{9}(X_1 + X_2 + \ldots + X_9)\right]$$

Standard Error

Std. Dev. of estimator's sampling dist. is called standard error.

Standard Error of the Sample Mean

$$SE(\bar{X}_n) = \sqrt{Var\left(\bar{X}_n\right)} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$$

Why $(n-1)$ for sample variance?

Why $(n-1)$ for sample variance?

We will show that having $n-1$ in the denominator ensures:

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \sigma^2$$

under random sampling.

Why $(n-1)$ for sample variance?

Step # 1 – Tedious but straightforward algebra gives:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-n(\bar{X}-\mu)^2$$

You are not responsible for proving Step #1 on an exam.

$$
\begin{aligned}
\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 &= \sum_{i=1}^{n}\left(X_i-\mu+\mu-\bar{X}\right)^2=\sum_{i=1}^{n}\left[(X_i-\mu)-(\bar{X}-\mu)\right]^2 \\
&= \sum_{i=1}^{n}\left[(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}(X_i-\mu)^2-\sum_{i=1}^{n}2(X_i-\mu)(\bar{X}-\mu)+\sum_{i=1}^{n}(\bar{X}-\mu)^2 \\
&= \left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+n(\bar{X}-\mu)^2 \\
&= \left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-2(\bar{X}-\mu)\left(\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\mu\right)+n(\bar{X}-\mu)^2 \\
&= \left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-2(\bar{X}-\mu)(n\bar{X}-n\mu)+n(\bar{X}-\mu)^2 \\
&= \left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-2n(\bar{X}-\mu)^2+n(\bar{X}-\mu)^2 \\
&= \left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-n(\bar{X}-\mu)^2
\end{aligned}
$$

Why $(n-1)$ for sample variance?

Step # 2 – Take Expectations of Step # 1:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= E\left[\left\{\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right\}-n(\bar{X}-\mu)^2\right] \\
&=
\end{aligned}
$$

Why $(n-1)$ for sample variance?

Step # 2 – Take Expectations of Step # 1:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= E\left[\left\{\sum_{i=1}^{n}(X_i-\mu)^2\right\}-n(\bar{X}-\mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-E\left[n(\bar{X}-\mu)^2\right] \\
&=
\end{aligned}
$$

Why $(n-1)$ for sample variance?

Step # 2 – Take Expectations of Step # 1:

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= E\left[\left\{\sum_{i=1}^{n}(X_i-\mu)^2\right\}-n(\bar{X}-\mu)^2\right] \\
&= E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-E\left[n(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]-n\,E\left[(\bar{X}-\mu)^2\right]
\end{aligned}
$$

Where we have used the linearity of expectation.

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - n\, E\left[(\bar{X}-\mu)^2\right] \\
&=
\end{aligned}
$$

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[\left(X_i-\mu\right)^2\right] - n\,E\left[(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,E\left[(\bar{X}-E[\bar{X}])^2\right] \\
&=
\end{aligned}
$$

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - n\, E\left[(\bar{X} - \mu)^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\, E\left[(\bar{X} - E[\bar{X}])^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\, Var(\bar{X}) = n\sigma^2 - \sigma^2 \\
&=
\end{aligned}$$

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - n\,E\left[(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,E\left[(\bar{X}-E[\bar{X}])^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,Var(\bar{X}) = n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}
$$

Since we showed earlier today that $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \sigma^2/n$ under this random sampling assumption.

Why $(n-1)$ for sample variance?

Finally – Divide Step # 3 by $(n-1)$:

$$E[S^2] \;=\; E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=\frac{(n-1)\sigma^2}{n-1}=\sigma^2$$

Hence, having $(n-1)$ in the denominator ensures that the sample variance is "correct on average," that is *unbiased*.