Probability Theory

Applications for Data Science
Module 5: Expectation, Variance, Covariance, and Correlation

Anne Dougherty

March 21, 2021

TABLE OF CONTENTS

Central Limit Theorem

This module is concerned with the Central Limit Thm.

In this video, we will discuss a random sample and the Law of Large Numbers.

We'll conclude with the CLT

At the end of this module, students should be able to

- Understand the definition of a random sample.
- Understand the Law of Large Numbers.
- Understand and use the Central Limit Theorem (CLT).
- Explain the implications of the CLT to the calculation and estimation of the mean.

Recall

For a random variable X, we need either the probability mass function $p(k) = P(X = k)$ or density function $f(x)$ to compute *a probability or to find*

- $\mu_X = E(X) = \sum_k kP(X = k)$ or $\mu_X = \int_{-\infty}^{\infty} xf(x)\, dx$
- $\sigma_X^2 = V(X) = E[(X - \mu_X)^2] = \sum_k (k - \mu_X)^2 P(X = k)$ or $\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\, dx$

Question: What if we don't know how a random variable is distributed? What if we don't know the mean or the variance?

Statistical Inference: In future courses, we will be focusing on making "statistical inferences" about the true mean and true variance of a population by using sample datasets. Before we do, we need to finish laying the groundwork.

Definition: $X_1, X_2, \ldots, X_n$ are a **random sample** of size n if

- $X_1, X_2, \ldots, X_n$ are independent ← we saw what it means for 2 r.v. to be indep.
- each random variable has the same distribution

We say that these X_i's are *iid*, independent and identically distributed.

This is an extension of what we did in the previous module where we worked with 2 r.v. Now, we're working with n of them. The same ideas still hold.

mean + variance are 2 quantities that help to summarize our data.

We use **estimators** to summarize our iid sample. For example, suppose we want to understand the distribution of adult female heights in a certain area. We plan to select n women at random and measure their height. Suppose the height of the i^{th} woman is denoted by X_i. $X_1, X_2, \ldots, X_n$ are iid with mean μ.

An **estimator** of μ is denoted $\bar{X}$ and $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

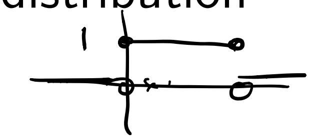
$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)$$

we saw last time that $E(aX+bY) = aE(X)+bE(Y)$

$$= \frac{1}{n}\underbrace{\sum_{k=1}^{n} E(X_k)} = \frac{1}{n}\sum_{k=1}^{n}\mu = \mu$$

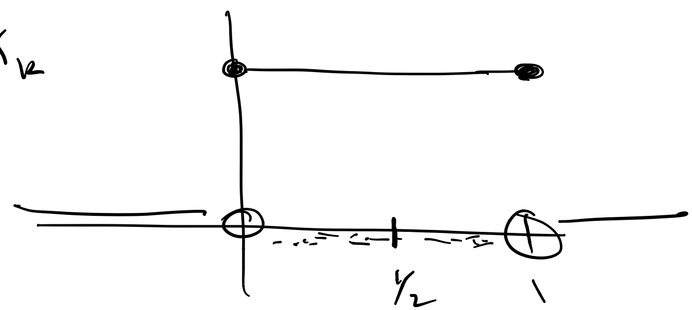
The Law of Large Numbers is fairly technical. However, it says that under most conditions, if $X_1, X_2, \ldots, X_n$ is a random sample with $E(X_k) = \mu$, then $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$, converges to μ in the limit as n goes to infinity. Previous slide, $E(\bar{X}) = \mu$.
The LLN says more.,

Example: Let $X_1, X_2, \ldots, X_n$ each have a uniform distribution on $[0, 1]$. For each X_i, $f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{if } x < 0 \text{ or } x > 1 \end{cases}$



$E(X_i) = 1/2$

$\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$



The LLN (& CLT) tells us that what we do in statistics is justified. If we want to understand more of the population, we have to look at more information. One r.v. is not enough.

What about the variance? Given a random sample $X_1, X_2, \ldots, X_n$ with $V(X_i) = \sigma^2$,

~~$V(\bar{X}) =$~~ Recall: previous video $V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab\,Cov(X, Y)$

If X & Y are indep, $Cov(X, Y) = 0$.

So: $V(aX + bY) = a^2 V(X) + b^2 V(Y)$

Now, apply this to $V(\bar{X})$

$$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

Note: as n increases, $\underbrace{\text{variance}}_{\text{spread of the distribution}}$ becomes smaller.

As n increases, $\bar{X}$ gets closer to μ; & variance gets smaller.

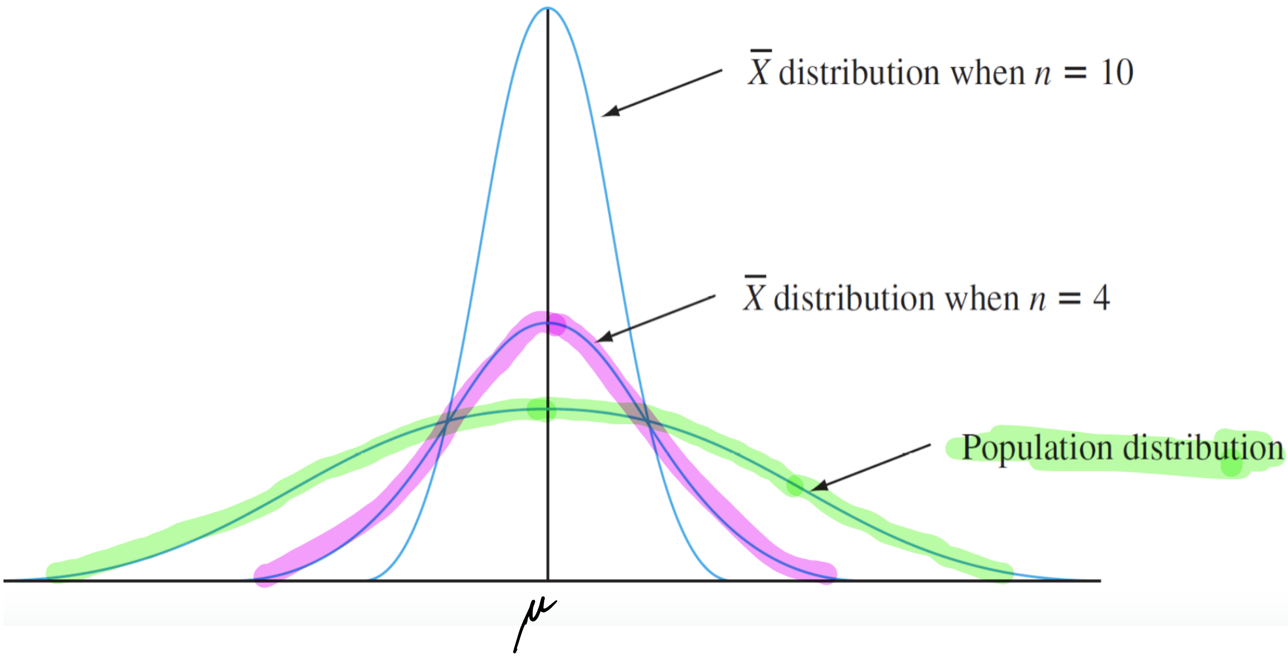
We use estimators to summarize our iid sample. Any estimator, including the sample mean, $\bar{X}$, is a random variable (since it is based on a random sample).

This means that $\bar{X}$ has a distribution of it's own, which is referred to as the **sampling distribution of the sample mean**. This sampling distribution depends on:

- the sample size n
- the population distribution of the X_i
- the method of sampling (ideally, want method to produce indep X_i's

Great, but what is the **distribution** of the sample mean?

Proposition: If $X_1, X_2, \ldots, X_n$ be iid with $X_i \sim N(\mu, \sigma^2)$. Then, $\bar{X} \sim N(\mu, \sigma^2/n)$



We know everything there is to know about the distribution of the sample mean when the population distribution is normal. What if the population distribution is not normal?

- When the population distribution is non-normal, averaging produces a distribution that is more bell-shaped than the one being sampled.
- A reasonable conjecture is that if n is large, a suitable normal curve will approximate the actual distribution of the sample mean.
- The formal statement of this result is one of the most important theorems in probability and statistics: **Central Limit Theorem**

Central Limit Theorem Let $X_1, X_2, \ldots, X_n$ be a random sample with $E(X_i) = \mu$ and $V(X_i) = \sigma_2$. If n is sufficiently large, $\bar{X}$ has approximately a normal distribution with mean $\mu_{\bar{X}}$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$.

We write $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$

The larger the value of n, the better the approximation. Typical rule of thumb: $n \geq 30$.