Probability Theory

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

Anne Dougherty

March 23, 2021

TABLE OF CONTENTS

Central Limit Theorem

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.

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

Proposition: If $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim N(\mu_i, \sigma_i^2)$ then $\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$.

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu_i$$

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

$X_1, X_2, \ldots, X_n$ are indep

Extend to $\sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right) \rightarrow$ assumes $X_1, \ldots X_n$ indep.

Suppose you have 3 errands to do in three different stores. Let T_i be the time to make the i^{th} purchase for $i = 1, 2, 3$. Let T_4 be the total walking time between stores. Suppose $T_1 \sim N(15, 16)$, $T_2 \sim N(5, 1)$, $T_3 \sim N(8, 4)$, and $T_4 \sim N(12, 9)$. Assume T_1, T_2, T_3, T_4 are independent . If you leave at 10 in the morning and you want tell a colleague, "I'll be back by time t", what should t be so that you will return by that time with probability .99?

Let $T_0 = T_1 + T_2 + T_3 + T_4$ = total time

$E(T_0) = 15 + 5 + 8 + 12 = 40$

$V(T_0) = 16 + 1 + 4 + 9 = 30$

$T_0 \sim N(40, 30)$

Want to find t so that $P(T_0 \leq t) = .99$

$$T_0 \sim N(40, 30)$$

$$P(T_0 \le t) = .99$$

$$P\left(\frac{T_0 - 40}{\sqrt{30}} \le \frac{t-40}{\sqrt{30}} \right) = .99$$

$$\Rightarrow \quad P\left(Z \le \frac{t-40}{\sqrt{30}} \right) = \Phi\left(\frac{t-40}{\sqrt{30}} \right) = .99$$

Look up: $\Phi(2.33) = .99$ $\quad \Rightarrow \quad \dfrac{t-40}{\sqrt{30}} = 2.33$

$$\Rightarrow \quad t = 52.76 \text{ min.}$$

Return by 10:52.76 with prob. .99

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 $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$.

You want to verify that 25-kg bags of fertilizer are being filled to the appropriate amount. You select a random sample of 50 bags of fertilizer and weigh them. Let X_i be the weight of the i^{th} bag for $i = 1, 2, \ldots 50$. You expect $E(X_i) = 25$ and $V(X_i) = .5$. Let $\bar{X} = (1/50)\sum_{i=1}^{50} X_i$.

Find $P(24.75 \leq \bar{X} \leq 25.25)$

From CLT

$\bar{X} \sim N(25, \frac{.5}{50})$

$\bar{X} \sim N(25, .01)$

$$P\left(\frac{24.75-25}{\sqrt{.01}} \leq \frac{\bar{X}-25}{\sqrt{.01}} \leq \frac{25.25-25}{\sqrt{.01}}\right)$$

$$= P(-2.5 \leq Z \leq 2.5)$$

$$= \Phi(2.5) - \Phi(-2.5) \approx .9876$$

Suppose $E(X_i) = 24.5$, that is, the bags are underfilled, and $V(X) = .5$. Now, find $P(24.75 \leq \bar{X} \leq 25.25)$.

<u>Now:</u> $\bar{X} \sim N(24.5, \frac{.5}{50}) = N(24.5, .01)$

$$P\left(\frac{24.75 - 24.5}{\sqrt{.01}} \leq \frac{X - 24.5}{\sqrt{.01}} \leq \frac{25.25 - 24.5}{\sqrt{.01}} \right)$$

$$= P(2.5 \leq Z \leq 7.5)$$

$$= \Phi(7.5) - \Phi(2.5)$$

$$\cong 1.000 - .9938 = .0062$$

In a statistics class of 36 students, past experience indicates that 53% of the students will score at or above 80%. For a randomly selected exam, find the probability at at least 20 students will score above 80%.

Let $X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ student scores at or above } 80\% \\ 0 & \text{if below } 80\% \end{cases}$

Let $X = \sum_{i=1}^{n} X_i = \#$ of success, $X \sim Bin(np, np(1-p))$

$X \sim Bin(19.08, 8.97)$

$$P(X \geq 20) = P(X \leq 19.5)$$

$$= P\left(\frac{X - 19.08}{\sqrt{8.97}} \leq \frac{19.5 - 19.08}{\sqrt{8.97}}\right)$$

$$= P(Z \leq .14) = \Phi(.14) \simeq .5557$$

Example: Normal approximation to the binomial. If $X \sim Bin(n, p)$ then X counts the number of successes in n independent Bernoulli trials, each with probability of success p. We know:

$$E(X) = np \; V(X) = np(1-p)$$

So, by CLT, $\dfrac{X - np}{\sqrt{np(1-p)}} \approx N(\mu = np, \sigma^2 = np(1-p)$.

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment. can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when n is sufficiently large. The problem is that the accuracy of the approximation for a particular n depends on the shape of the original underlying distribution being sampled.