1 Mean, Variance, and Standard Deviation

Mean The mean is the average value of a variable. The population mean $\mu = E(Y)$ describes the entire population. The sample mean $\bar{y}$ is calculated from n observed values and is used to estimate the population mean:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The population mean is a fixed, usually unknown value. The sample mean can vary across samples.

Variance Variance measures the spread of a variable around its mean. The population variance and sample variance are

$$\sigma^2 = E[(Y-\mu)^2], \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

The sample variance s^2 is used to estimate the population variance σ^2.

Standard Deviation Standard deviation is the square root of variance and has the same units as the original variable:

$$\sigma = \sqrt{\sigma^2}, \qquad s = \sqrt{s^2}.$$

Exercise A random sample of three observations from a population is

$$1, \quad 2, \quad 3.$$

1. Calculate the sample mean.
2. Calculate the sample variance.
3. Calculate the sample standard deviation.
4. What is the population mean? Can you determine it from this sample?

2 CDF and Quantile Function

Cumulative Distribution Function (CDF) The CDF gives the probability that a random variable Y is less than or equal to a given value y:

$$F(y) = P(Y \leq y).$$

In words, $F(y)$ asks: "What proportion of values are at or below y?" For example, if Y is hourly wage and $F(20) = 0.60$, then 60% of workers earn \$20 per hour or less.

The probability of an interval can be calculated as

$$P(a < Y \leq b) = F(b) - F(a).$$

Quantile Function The quantile function gives the value corresponding to a cumulative probability p. For a continuous, strictly increasing CDF,

$$Q(p) = F^{-1}(p).$$

In words, $Q(p)$ asks: "What value marks the bottom $100p$%?" For example, $Q(0.60) = 20$ means that \$20 per hour is the 60th percentile of wages.

The CDF maps a value to a probability; the quantile function maps a probability to a value.

The median is $Q(0.50)$. The central 90% interval is

$$[Q(0.05),\ Q(0.95)].$$

For a continuous distribution, this interval leaves 5% of the distribution below the lower endpoint and 5% above the upper endpoint.

Exercise 2 Suppose hourly wage Y has a continuous, strictly increasing CDF. The following values are given:

y	10	15	20	30	50
$F(y)$	0.05	0.25	0.50	0.75	0.95

1. What is the probability that a worker earns \$15 per hour or less?
2. Calculate $P(15 < Y \leq 30)$.
3. Find $Q(0.50)$ and $Q(0.95)$. Interpret each value.
4. Find the central 90% interval of hourly wages. What proportion of workers earn more than its upper endpoint?

3 Normal Distribution

Normal Distribution A normal distribution is bell-shaped and symmetric around its mean. We write

$$Y \sim N(\mu, \sigma^2),$$

where μ is the mean and σ^2 is the variance. The mean determines the center, and the standard deviation σ determines the spread.

Standard Normal Distribution The standard normal distribution has mean 0 and variance 1. If $Y \sim N(\mu, \sigma^2)$, then

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1).$$

The standardized value, or **z-score**, tells us how many standard deviations a value is above or below the mean. For example, $z = 2$ means two standard deviations above the mean.

Let Φ denote the standard normal CDF. Then

$$P(Y \leq y) = \Phi\left(\frac{y - \mu}{\sigma}\right).$$

Normal Quantiles and Central Intervals To find a quantile of Y, multiply the standard normal quantile by σ and add μ:

$$Q_Y(p) = \mu + \sigma\Phi^{-1}(p).$$

Useful standard normal quantiles are

$$\Phi^{-1}(0.95) \approx 1.645, \qquad \Phi^{-1}(0.975) \approx 1.96.$$

Therefore, the central intervals are

$$\text{Central } 90\%: \quad [\mu - 1.645\sigma,\ \mu + 1.645\sigma],$$

$$\text{Central } 95\%: \quad [\mu - 1.96\sigma,\ \mu + 1.96\sigma].$$

Each interval leaves equal probability in the two tails. For example, the central 95% interval leaves 2.5% below its lower endpoint and 2.5% above its upper endpoint.

Exercise 3 Suppose test scores follow a normal distribution:

$$Y \sim N(100, 10^2).$$

Use the following values:

$$\Phi(2) \approx 0.9772, \qquad \Phi^{-1}(0.95) \approx 1.645, \qquad \Phi^{-1}(0.975) \approx 1.96.$$

1. What are the mean, variance, and standard deviation of Y?
2. Calculate the z-score for a score of 120. Interpret your answer.
3. Calculate the probability that a student scores above 120.
4. Find the 95th percentile of test scores.
5. Find the central 95% interval of test scores.

6. Is the 95th percentile the upper endpoint of the central 95% interval? Explain.

Answers

Exercise 1

A random sample of three observations from a population is

$$1, \quad 2, \quad 3.$$

1. Calculate the sample mean.

$$\bar{y} = \frac{1+2+3}{3} = 2.$$

2. Calculate the sample variance.

$$s^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3-1} = 1.$$

3. Calculate the sample standard deviation.

$$s = \sqrt{s^2} = \sqrt{1} = 1.$$

4. What is the population mean? Can you determine it from this sample?

   The population mean μ is unknown and cannot be determined from this sample alone. The sample mean $\bar{y} = 2$ is an estimate of μ, not necessarily its true value.

Exercise 2

Suppose hourly wage Y has a continuous, strictly increasing CDF. The following values are given:

y	10	15	20	30	50
$F(y)$	0.05	0.25	0.50	0.75	0.95

1. What is the probability that a worker earns \$15 per hour or less?

$$P(Y \leq 15) = F(15) = 0.25.$$

   Thus, 25% of workers earn \$15 per hour or less.

2. Calculate $P(15 < Y \leq 30)$.

$$P(15 < Y \leq 30) = F(30) - F(15) = 0.75 - 0.25 = 0.50.$$

3. Find $Q(0.50)$ and $Q(0.95)$. Interpret each value.

$$Q(0.50) = 20, \qquad Q(0.95) = 50.$$

   The median hourly wage is \$20: 50% of workers earn \$20 or less. The 95th percentile is \$50: 95% of workers earn \$50 or less.

4. Find the central 90% interval of hourly wages. What proportion of workers earn more than its upper endpoint?

$$[Q(0.05),\, Q(0.95)] = [10,\, 50].$$

The central 90% interval is \$10 to \$50 per hour. The proportion above the upper endpoint is

$$P(Y > 50) = 1 - F(50) = 1 - 0.95 = 0.05.$$

Exercise 3

Suppose test scores follow a normal distribution:

$$Y \sim N(100, 10^2).$$

Use the following values:

$$\Phi(2) \approx 0.9772, \qquad \Phi^{-1}(0.95) \approx 1.645, \qquad \Phi^{-1}(0.975) \approx 1.96.$$

1. What are the mean, variance, and standard deviation of Y?

$$\mu = 100, \qquad \sigma^2 = 100, \qquad \sigma = 10.$$

2. Calculate the z-score for a score of 120. Interpret your answer.

$$z = \frac{120 - 100}{10} = 2.$$

A score of 120 is two standard deviations above the mean.

3. Calculate the probability that a student scores above 120.

$$P(Y > 120) = P(Z > 2) = 1 - \Phi(2) \approx 1 - 0.9772 = 0.0228.$$

Approximately 2.28% of students score above 120.

4. Find the 95th percentile of test scores.

$$Q_Y(0.95) = 100 + 10(1.645) = 116.45.$$

Approximately 95% of students score at or below 116.45.

5. Find the central 95% interval of test scores.

$$[100 - 1.96(10),\, 100 + 1.96(10)] = [80.4,\, 119.6].$$

6. Is the 95th percentile the upper endpoint of the central 95% interval? Explain.

No. The 95th percentile is 116.45, with 5% of scores above it. The central 95% interval leaves 2.5% in each tail, so its upper endpoint is the 97.5th percentile, 119.6.