STA 103 Lecture 7: Standardization, Percentiles

Instructor: Wookyeong Song

Department of Statistics, University of California, Davis

Aug 19th, 2025

Properties of Expectation and Variance

Recall, suppose we have RVs X_1 and X_2, then we have the following formulas:

- $E(X_1 + X_2) = E(X_1) + E(X_2)$.
- If X_1 and X_2 are independent then,

  $$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

  Otherwise,

  $$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2).$$

Properties of Expectation and Variance

Suppose we have RVs $X_1, \ldots, X_n$, then

- $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$.

- If $X_1, X_2, \ldots, X_n$ are pairwise independent, i.e., for any pairs (X_i, X_j), X_i and X_j are independent, then

  $$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots \mathrm{Var}(X_n).$$

  Otherwise, we have

  $$\begin{aligned}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = {} & \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots \mathrm{Var}(X_n) \\ & + \sum_{i \neq j} 2\mathrm{Cov}(X_i, X_j).\end{aligned}$$

Expectation and Variance of Sample Mean

Consequence of the formula, if $X_1, \ldots, X_n$ are pairwise independent with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, for all $i = 1, \ldots, n$,

- The expectation of sample mean $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ is given by

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n)$$
$$= \frac{1}{n}\left\{E(X_1) + E(X_2) + \cdots + E(X_n)\right\} = \frac{1}{n} \times n \times \mu = \mu.$$

- The variance of the sample mean is

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n)$$
$$= \frac{1}{n^2}\left\{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)\right\}$$
$$= \frac{1}{n^2} \times n \times \sigma^2 = \frac{\sigma^2}{n}.$$

Expectation and Variance of Sample Mean

- **Example**: Your friend just flipped a coin 100 times and got 80 heads. Is this unusual? Let $X_i = 1$ if ith flip is heads, and $X_i = 0$ if ith flip is tails.

- Your friend observed

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{100} = 0.8$$

- **Question 1**: Find $E(\bar{X})$ and $\mathrm{sd}(\bar{X})$.

Expectation and Variance of Sample Mean

- **Answer**: Since X_i follows **i.i.d.** Bernoulli distribution, $X_i \sim$ Bernoulli(p), with parameter $p = 1/2$. Then $E(X_1) = p = \frac{1}{2}$ and $\mathrm{Var}(X_1) = p(1-p) = \frac{1}{4}$.

- Then we have

$$\begin{aligned} E(\bar{X}) &= E(X_1) = 0.5, \\ \mathrm{Var}(\bar{X}) &= \frac{\mathrm{Var}(X_1)}{100} = \frac{1}{400}, \\ \mathrm{sd}(\bar{X}) &= \sqrt{\frac{1}{400}} = \frac{1}{\sqrt{400}} = \frac{1}{20} = 0.05. \end{aligned}$$

- **Question 2**: Is your friend's observation unusual?

- **Answer**: Then we would expect $\bar{X} \approx E(\bar{X}) \pm \mathrm{sd}(\bar{X}) = 0.5 \pm 0.05$. But your friend observed $\bar{X} = 0.8$ way out here so we can say very unusual.

Identifying unusual outcomes of RV using Z-score

- Recall, the Z-score of a RV X is another RV Z,

$$Z = \frac{X - E(X)}{\mathrm{sd}(X)}.$$

- The Z-score shows a common scale, so it could be used for identifying unusual outcomes of a random variables as well.

- You can say that the Z-scores are a bit like percentiles in that they allow you to assess how rare a sample of X is. Later we will see how to translate a Z-score into a percentile for Normal RVs.

- Here is very informal guideline when underlying distribution is unknown, especially not assumed to be Normal distribution:
  - Z-score within $\pm$ 2 is typical
  - Z-score 3 or -3 is rare but possible
  - Z-score greater than $|4|$ is extremely rare

Z-score

- In the previous example your friend observed $\bar{X} = 0.8$ when $E(\bar{X}) = 0.5, \mathrm{sd}(\bar{X}) = 0.05$.

- Then the Z-score of $\bar{X}$ above example is given by

$$Z = \frac{\bar{X} - E(\bar{X})}{\mathrm{sd}(\bar{X})} = \frac{0.8 - 0.5}{0.05} = 6.$$

- This is so rare there something must be wrong.

- The best way to think of a Z-score is that it quantifies how many $\mathrm{sd}(X)$'s can fit between what we would expect X to be (i.e. $E(X)$) and what we actually observed X to be.

- The $Z = 6$ implies $\bar{X} - E(\bar{X}) = 6 \times \mathrm{sd}(\bar{X})$, meaning that observed $\bar{X}$ is 6 times the typical fluctuation $\mathrm{sd}(\bar{X})$ of what we expect.

Special Case of Standardization: Normal Distribution

- There is exact guideline when RV X is assumed to be Normal distribution.

- Recall important Fact about Normal distribution:
  - If X is Gaussian (i.e., $N(a, b)$) then the Z-score for X is standard Gaussian

$$Z = \frac{X - a}{\sqrt{b}} \sim N(0, 1).$$

- Then, Z-score of Normal RV X is one-to-one with percentiles.

Quantile

- Quantile of the standard Normal Distribution $Z \sim N(0, 1)$: The p quantile (or $(100 \times p)$th percentile) of a standard normal distribution is the value z_p such that:

$$P(Z \leq z_p) = p, \quad 0 < p < 1.$$

- **Interpretation**: The z_p is the value below which a proportion p of the data falls.

- **Example**: "Your income is in the top 1%." Meaning that the percentile of your income is greater than 99%. We call it your income is 99th percentile (or 0.99 quantile).

Finding $100p$th percentile using the Z-table

- Step 1) Identify the cumulative probability $p \in (0, 1)$, or $100p$th percentile.
- Step 2) Use the Z-table (which gives $P(Z \leq z_p)$) to find the critical value z_p that corresponds as closely as possible to the desired probability p.
- Step 3) If the exact critical value is not in the table (combination of row name and column name), use the closet value.

Finding $100p$th percentile using the Z-table

- **Example 1**: Suppose $Y \sim N(0, 1)$. Here is a list of some possible values for Y along with their associated percentiles.

Raw value Y	Percentile
-1.89	?
-0.39	?
0.11	?

- Since Y is already standard normal distribution, we do **not** need to generate Z-score. Raw value Y is already Z-score.

- Finding what percentile of raw value Y is to calculate p such tat $P(Y \leq -1.89) = p$. Similarly, we need to find p such that $P(Y \leq -0.39) = p$ and $P(Y \leq 0.11) = p$. Then, $100 \times p$ percentiles are the answers.

Finding $100p$th percentile using the Z-table

- **Example 2**: Suppose $W \sim N(-27, 9)$. Here is a list of some possible values for W along with their associated percentiles.

Raw value W	Z-score	Percentile
-32.67	?	?
-28.17	?	?
-26.67	?	?

- Since W is **not** standard normal distribution, we **do** need to generate Z-score.

- **Takeaway**: There is a one-to-one correspondence among 1) raw value, 2) Z-score and 3) Percentile.

Finding Z-score using percentiles

- **Example 3**: Suppose $X \sim N(5, 4)$. The $X = x$ is observed to be at the 94.5th percentile. What was the Z-score for that observation? What was the actual value x of the observed X?

- This situation can be displayed by:

Raw value X	Z-score	Percentile
?	?	94.5

- The RV $X = x$ is observed to be 94.5th percentile:

$$P(X \leq x) = 0.945.$$

- Since there is one-to-one correspondence between Percentile and Z-score, we can obtain Z-score in the table $z = \frac{x - E(X)}{\mathrm{sd}(X)} = 1.598$.

- Then, the actual value x is given by

$$x = E(X) + 1.598 \times \mathrm{sd}(X) = 5 + 2 \times 1.598 = 8.196.$$

Finding Z-score using percentiles

- **Example 3**: Suppose $X \sim N(5, 4)$. The $X = x$ is observed to be at the 94.5th percentile. What was the Z-score for that observation? What was the actual value x of the observed X?
- To sum up, this situation can be displayed by:

Raw value X	Z-score	Percentile
8.196	1.598	94.5

Identifying Unusual Outcomes of Gaussian RV

- **Example 4**: Suppose we want the critical value z^* such that the central 95% of the standard normal distribution lies between $-z^*$ and $+z^*$.

- Then, we can formalize it $P(-z^* < Z < z^*) = 0.95$, where $Z \sim N(0, 1)$.

- To find z^*, we use the symmetry of the standard normal distribution, then $P(Z < -z^*) = 0.975$.

- From Z-table we have $z_{0.975} = 1.96$.

- **Interpretation**: Under standard Gaussian distribution, Z-score within ± 1.96 guaranteed 95% confidence interval. Observing Z-score greater than 1.96 or less than -1.96 is rare, occurring only about 5% of the time."

Identifying Unusual Outcomes of Gaussian RV

- **Example 5**: Similarly, we can find the critical value z^* such that the central 99% of the standard normal distribution lies between $-z^*$ and $+z^*$.

- Then, we can formalize it $P(-z^* < Z < z^*) = 0.99$, where $Z \sim N(0, 1)$.

- To find z^*, we use the symmetry of the standard normal distribution, then $P(Z < -z^*) = 0.995$.

- From Z-table we have $z_{0.995} = 2.58$.

- **Interpretation**: Under standard Gaussian distribution, Z-score within ± 2.58 guaranteed 99% confidence interval. Observing Z-score greater than 2.58 or less than -2.58 is extremely rare, occurring only about 1% of the time."

Case Study

- **Demand Forecasting for a Retail Chain**: A national retail chain tracks weekly sales of a popular seasonal product, portable air conditioners, across its regional stores. The company aims to estimate the appropriate inventory level to meet expected demand without overstocking.

- Let X be the weekly demand for the air conditioners. Historical data show that mean weekly demand is 500 units and standard deviation 60 units. Also, X is known to follow normal distribution.

- **Question 1**: Identify the distribution of X with parameters.

- **Answer**: X is known to be normal distribution with mean 500 and standard deviation 60.

Case Study

- **Demand Forecasting for a Retail Chain**: A national retail chain tracks weekly sales of a popular seasonal product, portable air conditioners, across its regional stores. The company aims to estimate the appropriate inventory level to meet expected demand without overstocking.

- Let X be the weekly demand for the air conditioners. Historical data show that mean weekly demand is 500 units and standard deviation 60 units. Also, X is known to follow normal distribution.

- **Question 2**: Find the probability that weekly demand exceeds 600 units. (Hint: $P(X \geq 600)$ and standardization).

Case Study

- **Demand Forecasting for a Retail Chain**: A national retail chain tracks weekly sales of a popular seasonal product, portable air conditioners, across its regional stores. The company aims to estimate the appropriate inventory level to meet expected demand without overstocking.

- Let X be the weekly demand for the air conditioners. Historical data show that mean weekly demand is 500 units and standard deviation 60 units. Also, X is known to follow normal distribution.

- **Question 3**: Determine the inventory level that meet demand in 90% of weeks. (Hint: $x = E(X) + z_{0.90} \times \mathrm{sd}(X)$).