Ch4 Exploring the data

Jih-Chang Yu

August 26, 2023

Agenda

Sample Mean, Sample Median, Mode

Range, Sample Variance, Sample Std. Deviation

Percentiles, Quartiles

Skewness, Kurtosis

Sample Mean, Sample Median, Mode

Sample Mean

The mean of a dataset $X_1, X_2, \ldots, X_n$ is calculated as:

$$\text{Mean} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

where n is the total number of data points.

Sample Median

The median, denoted by M, of a dataset is obtained by arranging the data in ascending order and identifying:

$$\Pr(X < M) < 0.5 \quad \text{and} \quad \Pr(X \leq M) \geq 0.5$$

In the data, we can evaluate the median (refers as the sample median)

- If n (number of data points) is odd: the value at the middle position.
- If n is even: the average of the two middle values.

where n is the total number of data points.

Mode

The mode of a dataset is the value that appears most frequently.

- A dataset can have one or more modes.
- If all values have the same frequency, the dataset has no mode.
- If multiple values have the highest frequency, it's a multimodal dataset.

- **Sample Mean**: Sum of data divided by the number of data points.
- **Sample Median**: Middle value when data is sorted.
- **Mode**: Most frequent value in the dataset.

Properties:

- Mean is sensitive to outliers.
- Median is robust to outliers.
- Mode may not exist or be unique.

Range, Sample Variance, Sample Std. Deviation

Range: Mathematical Definition

The range of a dataset $X_1, X_2, \ldots, X_n$ is the difference between the maximum and minimum values:

$$\text{Range} = \max(X_1, X_2, \ldots, X_n) - \min(X_1, X_2, \ldots, X_n)$$

Sample Variance: Mathematical Definition

The sample variance of a dataset $X_1, X_2, \ldots, X_n$ with mean $\bar{X}$ is calculated as:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

where n is the total number of data points.
The sample standard deviation S of a dataset is the square root of the sample variance S^2, representing the spread of data points:

$$S = \sqrt{S^2}$$

It provides a measure of the dispersion or variability within the dataset.

- **Range**: Difference between max and min values.
- **Sample Variance**: Avg. of squared differences from the mean.
- **Sample Std. Deviation**: Square root of variance.

Interpretation:

- Larger variance/standard deviation indicates greater data dispersion.

Percentiles, Quartiles

Percentiles, Quartiles

- **Percentiles**: Values dividing data into percentiles.
- **Quartiles**: Values dividing data into four parts.

Use Case:

- Median is the 50th percentile.
- Quartiles help identify data spread and skewness.

Skewness, Kurtosis

Measurement of Skewness

Skewness is a measure of the asymmetry of the probability distribution of a dataset. It indicates the direction and extent of the departure from symmetry around the mean.

The skewness S_k is calculated using the following formula:

$$S_k = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{n \cdot S^3}.$$

where S_k is an estimator of $E(X-\mu)^3/\sigma^3$;
$E(X)=\mu,\ Var(X)=\sigma^2$.

Properties of Skewness

- Positive skewness ($S_k > 0$): Right-skewed distribution (longer right tail).

- Negative skewness ($S_k < 0$): Left-skewed distribution (longer left tail).

Kurtosis Measurement

- Kurtosis measures the degree of ”tailedness” of the probability distribution of a dataset. It quantifies how much the distribution’s tails deviate from the tails of a normal distribution.
- In statistics, kurtosis measures the peakedness of the probability distribution of a real-valued random variable. Higher kurtosis implies that increased variance is caused by infrequent extreme deviations from the mean, whether they are larger or smaller.

The kurtosis K is calculated using the following formula:

$$K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4} - 3$$

where K is an estimator of $E(X - \mu)^4/\sigma^4 - 3$.

Properties of kurtosis

- High kurtosis ($K > 0$): Heavy tails, more extreme values.

- Low kurtosis ($K < 0$): Light tails, fewer extreme values.

- **Skewness**: Measure of data distribution's asymmetry.
- **Kurtosis**: Measure of distribution's tail heaviness.

Implications:

- Positive skewness: Right-skewed distribution.
- Negative skewness: Left-skewed distribution.
- High kurtosis: Heavy tails, peaked distribution.
- Low kurtosis: Light tails, flat distribution.

Likelihood Function

Definition
The likelihood function is defined as the probability of observing the data given a particular set of parameter values in a statistical model.

- Let θ represent the unknown parameters in the model.
- Let X denote the observed data.
- The likelihood function, denoted as $L(\theta \mid X)$, is given by:

$$L(\theta \mid X) = P(X \mid \theta)$$

Maximum Likelihood Estimation (MLE)

- MLE is a method for estimating the parameters of a statistical model.
- It is based on finding the parameter values that maximize the likelihood function.

Steps of MLE

1. Formulate the likelihood function $L(\theta \mid X)$ based on the statistical model and observed data X.
2. Take the natural logarithm of the likelihood function to obtain the log-likelihood function $\ln L(\theta \mid X)$.

Steps of MLE (continued)

3. Maximize the log-likelihood function with respect to the parameters θ to find the MLE estimates.

- MLE estimates are often considered as the "best-fit" values of the parameters given the observed data.
- MLE has desirable asymptotic properties such as consistency and asymptotic normality.

MLE of Normal Distribution

Central Limit Theorem

Central Limit Theorem
Let $X_1, X_2, \cdots, X_n$ be a sequence of independent and identically distributed random variables with mean μ and standard deviation σ. As n approaches infinity, the distribution of the standardized sum (or average)

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

converges to a standard normal distribution, i.e., Z follows a standard normal distribution with mean 0 and standard deviation 1.

Confidence Intervals

- Introduction to confidence intervals
- Basic concepts
- Calculation of confidence intervals
- Interpretation of confidence intervals

Introduction

In statistics, a confidence interval is a range of values calculated from a sample of data that is likely to contain the true population parameter with a certain level of confidence. It provides a measure of the uncertainty associated with estimating an unknown parameter. Confidence intervals are widely used in inferential statistics and play a crucial role in hypothesis testing and estimation.

Basic Concepts

- Point estimate: Single value that estimates an unknown parameter
- Standard error: Measures the variability of a point estimate
- Confidence level: Probability that the true parameter lies within the confidence interval

Interpretation

Confidence intervals provide a range of plausible values for the population parameter. They represent the uncertainty associated with the estimation process and allow researchers to make informed decisions. It's important to note that confidence intervals are not probability distributions for the parameter, but rather ranges of values that are consistent with the observed data and chosen confidence level.

Confidence Interval Calculation with CLT

1. Collect a random sample from the population of interest.
2. Calculate the sample mean ($\bar{x}$) and the sample standard deviation (s).
3. Determine the desired confidence level (e.g., 95%, 99%). Find the critical value (z) corresponding to the chosen confidence level.
4. Calculate the margin of error (E) using the formula:

   $$E = z \times \frac{s}{\sqrt{n}}$$

   where n is the sample size.
5. Construct the confidence interval by adding and subtracting the margin of error from the sample mean:

$$\text{Lower bound} = \bar{x} - E \quad \text{and} \quad \text{Upper bound} = \bar{x} + E$$

Confidence Intervals for Binomial Distribution

- For a binomial distribution, confidence intervals estimate the success probability p.
- Steps to calculate confidence intervals for a binomial distribution:

1. Calculate sample proportion: $\hat{p} = \frac{X}{n}$, where X is the number of successes and n is the sample size.
2. Calculate standard error: $SE = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$.
3. Choose a confidence level (e.g., 95
4. Find the critical value $z_{1-\alpha/2}$ corresponding to the chosen confidence level (e.g., $z_{1-\alpha/2} \approx 1.96$ for 95

$$\text{Lower limit} = \hat{p} - z \cdot SE$$

$$\text{Upper limit} = \hat{p} + z \cdot SE$$

Confidence Intervals for Poisson Distribution

- For a Poisson distribution, confidence intervals estimate the average event rate λ.
- Steps to calculate confidence intervals for a Poisson distribution:

1. Calculate the sample average $\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$, where x_i are the observed event counts and n is the sample size.
2. Calculate standard error: $SE = \sqrt{\frac{\hat{\lambda}}{n}}$.
3. Choose a confidence level (e.g., 95
4. Find the critical value $z_{1-\alpha/2}$ corresponding to the chosen confidence level (e.g., $z_{1-\alpha/2} \approx 1.96$ for 95
5. Calculate confidence interval:

$$\text{Lower limit} = \hat{\lambda} - z_{1-\alpha/2} \cdot SE$$

$$\text{Upper limit} = \hat{\lambda} + z_{1-\alpha/2} \cdot SE$$

Confidence Intervals for Exponential Distribution

- For an exponential distribution, confidence intervals estimate the mean μ.
- Steps to calculate confidence intervals for an exponential distribution:

1. Calculate the sample mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$, where x_i are the observed values and n is the sample size.
2. Calculate standard error: $SE = \frac{\hat{\mu}}{\sqrt{n}}$.
3. Choose a confidence level (e.g., 95
4. Find the critical value $z_{1-\alpha/2}$ corresponding to the chosen confidence level (e.g., $z_{1-\alpha/2} \approx 1.96$ for 95
5. Calculate confidence interval:

$$\text{Lower limit} = \hat{\mu} - z_{1-\alpha/2} \cdot SE$$

$$\text{Upper limit} = \hat{\mu} + z_{1-\alpha/2} \cdot SE$$

Confidence Intervals for Normal Distribution

- For a normal distribution, confidence intervals estimate the population mean μ.
- Steps to calculate confidence intervals for a normal distribution:

1. Calculate the sample mean $\bar{x}$ and the sample standard deviation s from the data ; Calculate standard error: $SE = \frac{s}{\sqrt{n}}$, where n is the sample size.
2. Find the critical values $z_{\alpha/2}$ and $z_{1-\alpha/2}$ corresponding to the chosen confidence level, where α is the significance level (e.g., $\alpha = 0.05$ for 95
3. Calculate confidence interval:

$$\text{Lower limit} = \bar{x} - z_{1-\alpha/2} \cdot SE$$

$$\text{Upper limit} = \bar{x} + z_{1-\alpha/2} \cdot SE$$

Take-home message