Statistics Notes

Devesh

December 24, 2024

1 Measurements of Central Tendency

1.1 Mean

The mean, or average, is the sum of all values divided by the number of values. It is a measure of the central tendency of a set of numbers.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1.2 Median

The median is the middle value in a list of numbers. To find the median, the numbers must be arranged in numerical order. If there is an even number of observations, the median is the average of the two middle numbers.

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

1.3 Mode

The mode is the value that appears most frequently in a data set. A set of numbers may have one mode, more than one mode, or no mode at all.

$$\text{Mode} = \text{most frequent value in the data set}$$

1.4 Variance

Variance measures how far a set of numbers are spread out from their average value. It is the average of the squared differences from the mean.

$$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

1.5 Standard Deviation

The standard deviation is the square root of the variance. It provides a measure of the average distance from the mean.

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

1.6 Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. If the skewness is negative, the data are skewed to the left, meaning the left tail is longer or fatter than the right tail. If the skewness is positive, the data are skewed to the right, meaning the right tail is longer or fatter than the left tail. A skewness of zero indicates that the data are perfectly symmetrical.

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \mu)^3}{n\sigma^3}$$

1.7 Standard Error

The standard error is the standard deviation of the sampling distribution of a statistic, most commonly of the mean. It provides an estimate of the variability of the sample mean.

$$\text{Standard Error} = \text{SE} = \frac{\sigma}{\sqrt{n}}$$

1.8 Gaussian Distribution

The Gaussian distribution, also known as the normal distribution, is a continuous probability distribution characterized by a bell-shaped curve. It is defined by two parameters: the mean (μ) and the standard deviation (σ). The probability density function (PDF) of a Gaussian distribution is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where: - x is the variable - μ is the mean - σ is the standard deviation
The Gaussian distribution is symmetric about the mean, and its shape is determined by the standard deviation. A larger standard deviation results in a wider and flatter curve, while a smaller standard deviation results in a narrower and taller curve.

2 Central Limit Theorem

The Central Limit Theorem (CLT) states that the distribution of the sample mean of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables. This theorem is fundamental in statistics because it allows for the use of normal distribution approximations in various statistical methods.
Formally, if $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with mean μ and variance σ^2, then the sample mean $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$ for sufficiently large n. Mathematically, this can be expressed as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

The CLT is important because it justifies the use of the normal distribution in many practical applications, even when the underlying data do not follow a normal distribution.

3 Bayes' Theorem

Bayes' Theorem is a fundamental theorem in probability theory that describes how to update the probability of a hypothesis based on new evidence.
Bayes' Theorem is stated as:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where: - $P(A \mid B)$ is the posterior probability, the probability of hypothesis A given the evidence B. - $P(B \mid A)$ is the likelihood, the probability of evidence B given that hypothesis A is true. - $P(A)$ is the prior probability, the initial probability of hypothesis A before seeing the evidence. - $P(B)$ is the marginal likelihood, the total probability of the evidence under all possible hypotheses.

3.1 Prior Probability

The prior probability, $P(A)$, represents our initial belief about the probability of the hypothesis before observing any evidence. It is based on previous knowledge or assumptions.

3.2 Likelihood

The likelihood, $P(B \mid A)$, represents the probability of observing the evidence given that the hypothesis is true. It quantifies how well the hypothesis explains the evidence.

3.3 Posterior Probability

The posterior probability, $P(A \mid B)$, represents the updated probability of the hypothesis after observing the evidence. It combines the prior probability and the likelihood to provide a new probability based on the evidence.

3.4 Marginal Likelihood

The marginal likelihood, $P(B)$, represents the total probability of observing the evidence under all possible hypotheses. It is calculated as:

$$P(B) = \sum_i P(B \mid A_i) P(A_i)$$

where A_i represents all possible hypotheses.

3.5 Example: Medical Diagnosis

Suppose a patient is tested for a rare disease. The disease has a prevalence (prior probability) of 0.1% ($P(D) = 0.001$). The test has a sensitivity (true positive rate) of 99% ($P(T \mid D) = 0.99$) and a specificity (true negative rate) of 99% ($P(\neg T \mid \neg D) = 0.99$).

We want to find the probability that the patient has the disease given a positive test result ($P(D \mid T)$).

First, calculate the probability of a positive test result ($P(T)$):

$$P(T) = P(T \mid D)P(D) + P(T \mid \neg D)P(\neg D)$$

$$P(T) = (0.99 \times 0.001) + (0.01 \times 0.999) = 0.00099 + 0.00999 = 0.01098$$

Next, apply Bayes' Theorem:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T)}$$

$$P(D \mid T) = \frac{0.99 \times 0.001}{0.01098} \approx 0.0902$$

So, the probability that the patient has the disease given a positive test result is approximately 9.02

4 Z-Score

The Z-score measures how many standard deviations an element is from the mean, allowing for comparison between different data sets.

The Z-score formula is:

$$Z = \frac{X - \mu}{\sigma}$$

where: - X is the value - μ is the mean - σ is the standard deviation

A Z-score of 0 indicates the element is at the mean. Positive or negative Z-scores indicate the element is above or below the mean, respectively.

5 P-Values

A p-value is the probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true. It is used in hypothesis testing to determine the significance of the results. The p-value is calculated using the cumulative distribution function (CDF) of the test statistic under the null hypothesis. A smaller p-value indicates stronger evidence against the null hypothesis.

5.1 Example: Probability of Sampling at Least 2.5 Standard Deviations from the Mean

To find the probability of sampling at least 2.5 standard deviations from the mean in a standard normal distribution, we can use the Z-score and the properties of the normal distribution.
The Z-score formula is:

$$Z = \frac{X - \mu}{\sigma}$$

For a standard normal distribution, $\mu = 0$ and $\sigma = 1$. Therefore, the Z-score for 2.5 standard deviations from the mean is:

$$Z = 2.5$$

The probability of sampling at least 2.5 standard deviations from the mean is the sum of the probabilities in the two tails of the distribution:

$$P(|Z| \geq 2.5) = P(Z \leq -2.5) + P(Z \geq 2.5)$$

Using the standard normal distribution table or a calculator, we find:

$$P(Z \geq 2.5) \approx 0.0062$$

Since the normal distribution is symmetric:

$$P(Z \leq -2.5) = P(Z \geq 2.5) \approx 0.0062$$

Therefore, the total probability is:

$$P(|Z| \geq 2.5) = 2 \times 0.0062 = 0.0124$$

So, the probability of sampling at least 2.5 standard deviations from the mean is approximately 0.0124, or 1.24

6 Null Hypothesis

The null hypothesis, denoted as H_0, is a statement that there is no effect or no difference, and it serves as the default or starting assumption in hypothesis testing. It is the hypothesis that researchers aim to test against.
In hypothesis testing, the null hypothesis is typically tested against an alternative hypothesis, denoted as H_a or H_1, which represents a new effect or difference that the researcher wants to prove.
The steps to test a null hypothesis are as follows:

- Formulate the null hypothesis (H_0) and the alternative hypothesis (H_a).
- Choose a significance level (α), commonly set at 0.05.
- Collect data and calculate a test statistic.
- Determine the p-value, which is the probability of observing the test statistic or something more extreme under the null hypothesis.
- Compare the p-value to the significance level:

- If $p \leq \alpha$, reject the null hypothesis (H_0).
- If $p > \alpha$, fail to reject the null hypothesis (H_0).

Rejecting the null hypothesis suggests that there is sufficient evidence to support the alternative hypothesis. Failing to reject the null hypothesis suggests that there is not enough evidence to support the alternative hypothesis.

6.1 Example 1

Suppose we want to test whether a new drug is effective in lowering blood pressure. The null and alternative hypotheses might be:

$$H_0 : \text{The new drug has no effect on blood pressure.}$$

$$H_a : \text{The new drug lowers blood pressure.}$$

Let's say we conduct an experiment with 30 participants and measure their blood pressure before and after taking the drug. The mean decrease in blood pressure is found to be 5 mmHg with a standard deviation of 8 mmHg.

We perform a one-sample t-test to determine if the mean decrease is significantly different from zero. The test statistic is calculated as:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{5 - 0}{8/\sqrt{30}} \approx 3.42$$

Using a t-distribution table with 29 degrees of freedom, we find the p-value corresponding to $t = 3.42$. The p-value is approximately 0.002.

Since the p-value (0.002) is less than the chosen significance level (0.05), we reject the null hypothesis and conclude that the new drug is effective in lowering blood pressure.

6.2 Example 2

With a fair coin, the probability of throwing six heads or six tails in a six-coin-flip experiment is 0.03125 ($\rho = -0.015625$ for either of six heads or six tails). If a friend of yours hands you a coin, the null hypothesis (the baseline assumed by the fair-toss distribution) would be that the coin is fair. If you test this coin by flipping it six times and it comes up heads on all six or tails on all six, this observation would suggest that you should reject the null hypothesis (theres a good chance that the coin is not fair) because chance alone would facilitate such an observation less than 5% of the time, i.e., .

7 Binomial Distribution

The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. It is defined by two parameters: n (the number of trials) and p (the probability of success in each trial).

The probability mass function (PMF) of the binomial distribution is given by:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where: - X is the random variable representing the number of successes - k is the number of successes - $\binom{n}{k}$ is the binomial coefficient, calculated as $\frac{n!}{k!(n-k)!}$ - p is the probability of success - $1 - p$ is the probability of failure

7.1 Example: Coin Toss

Suppose we have a fair coin (i.e., $p = 0.5$) and we flip it 10 times. We want to find the probability of getting exactly 6 heads.
Here, $n = 10$ and $p = 0.5$. The number of successes $k = 6$.
Using the binomial formula:

$$P(X = 6) = \binom{10}{6} (0.5)^6 (0.5)^4$$

First, calculate the binomial coefficient:

$$\binom{10}{6} = \frac{10!}{6!4!} = 210$$

Then, calculate the probability:

$$P(X = 6) = 210 \times (0.5)^{10} = 210 \times \frac{1}{1024} \approx 0.205$$

8 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method used to estimate the parameters of a statistical model. The goal of MLE is to find the parameter values that maximize the likelihood function, which measures how well the model explains the observed data.
Given a set of independent and identically distributed data points $X = \{x_1, x_2, \ldots, x_n\}$ and a probability density function $f(x; \theta)$ parameterized by θ, the likelihood function $L(\theta)$ is defined as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

The log-likelihood function, which is often easier to work with, is given by:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

The MLE for θ is the value that maximizes the log-likelihood function:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\theta)$$

8.1 Example: Estimating the Mean of a Normal Distribution

Suppose we have a sample $X = \{x_1, x_2, \ldots, x_n\}$ from a normal distribution with unknown mean μ and known variance σ^2. The probability density function is:

$$f(x; \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

The log-likelihood function is:

$$\ell(\mu) = \sum_{i=1}^{n} \log f(x_i; \mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

To find the MLE for μ, we take the derivative of $\ell(\mu)$ with respect to μ and set it to zero:

$$\frac{\partial \ell(\mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

Solving for μ, we get:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Thus, the MLE for the mean μ of a normal distribution is the sample mean.
So, the probability of getting exactly 6 heads in 10 flips of a fair coin is approximately 0.205, or 20.5%.

9 T-Test

A t-test is a statistical test used to compare the means of two groups. It helps determine if the differences between the groups are statistically significant.

9.1 One-Sample T-Test

A one-sample t-test compares the mean of a single sample to a known value (usually the population mean).
The test statistic is calculated as:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

where: - $\bar{X}$ is the sample mean - μ_0 is the known value (population mean) - s is the sample standard deviation - n is the sample size

9.2 Two-Sample T-Test

A two-sample t-test compares the means of two independent samples to determine if they come from populations with equal means.
The test statistic is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where: - $\bar{X}_1$ and $\bar{X}_2$ are the sample means - s_1 and s_2 are the sample standard deviations - n_1 and n_2 are the sample sizes

9.3 Paired T-Test

A paired t-test compares the means of two related groups. It is used when the samples are dependent, such as measurements taken before and after a treatment on the same subjects.
The test statistic is calculated as:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}}$$

where: - $\bar{D}$ is the mean of the differences between paired observations - s_D is the standard deviation of the differences - n is the number of pairs

9.4 Example: One-Sample T-Test

Suppose we want to test if the average height of a sample of 20 students is different from the known average height of 170 cm. The sample mean height is 172 cm with a standard deviation of 5 cm.
The null hypothesis is:

$$H_0 : \mu = 170$$

The test statistic is:

$$t = \frac{172 - 170}{5/\sqrt{20}} = \frac{2}{1.118} \approx 1.79$$

Using a t-distribution table with 19 degrees of freedom, we find the p-value corresponding to $t = 1.79$. If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that the average height is significantly different from 170 cm.

10 Z-Test

A Z-test is a statistical test used to determine whether there is a significant difference between the means of two groups, or between a sample mean and a known population mean, when the population variance is known and the sample size is large (typically $n > 30$).
The test statistic for a one-sample Z-test is calculated as:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

where: - $\bar{X}$ is the sample mean - μ_0 is the population mean - σ is the population standard deviation - n is the sample size
For a two-sample Z-test, the test statistic is calculated as:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where: - $\bar{X}_1$ and $\bar{X}_2$ are the sample means - σ_1 and σ_2 are the population standard deviations - n_1 and n_2 are the sample sizes

10.1 Example: One-Sample Z-Test

Suppose we want to test if the average weight of a sample of 50 people is different from the known average weight of 70 kg. The sample mean weight is 72 kg with a population standard deviation of 10 kg.
The null hypothesis is:

$$H_0 : \mu = 70$$

The test statistic is:

$$Z = \frac{72 - 70}{10/\sqrt{50}} = \frac{2}{1.414} \approx 1.41$$

Using the standard normal distribution table, we find the p-value corresponding to $Z = 1.41$. If the p-value is less than the chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that the average weight is significantly different from 70 kg.

10.2 Difference Between T-Test and Z-Test

The T-test and Z-test are both statistical tests used to compare means, but they are used under different conditions and have different assumptions.

- **T-Test:**
  - Used when the sample size is small (typically $n < 30$).
  - Used when the population variance is unknown.
  - The test statistic follows a t-distribution.
  - More appropriate for small sample sizes because it accounts for the additional uncertainty in the estimate of the population standard deviation.
- **Z-Test:**
  - Used when the sample size is large (typically $n > 30$).
  - Used when the population variance is known.
  - The test statistic follows a standard normal distribution (Z-distribution).
  - More appropriate for large sample sizes because the sample mean will be approximately normally distributed due to the Central Limit Theorem.

In summary, the main differences between the T-test and Z-test lie in the sample size and whether the population variance is known. The T-test is used for smaller samples with unknown population variance, while the Z-test is used for larger samples with known population variance.

11 Type I and Type II Errors

In hypothesis testing, two types of errors can occur:

11.1 Type I Error

A Type I error occurs when the null hypothesis (H_0) is rejected when it is actually true. This is also known as a "false positive" or "alpha error." The probability of making a Type I error is denoted by α, which is the significance level of the test.

$$\text{Type I Error} = P(\text{Reject } H_0 \mid H_0 \text{ is true}) = \alpha$$

11.2 Type II Error

A Type II error occurs when the null hypothesis (H_0) is not rejected when it is actually false. This is also known as a "false negative" or "beta error." The probability of making a Type II error is denoted by β.

$$\text{Type II Error} = P(\text{Fail to reject } H_0 \mid H_0 \text{ is false}) = \beta$$

11.3 Power of a Test

The power of a test is the probability of correctly rejecting the null hypothesis when it is false. It is calculated as $1 - \beta$.

$$\text{Power} = 1 - \beta$$

The power of a test increases with larger sample sizes, larger effect sizes, and higher significance levels.

11.4 Example

Suppose we are testing a new drug and set the significance level (α) at 0.05. If the null hypothesis is that the drug has no effect, a Type I error would occur if we conclude that the drug is effective when it is not. A Type II error would occur if we conclude that the drug is not effective when it actually is.

By understanding and controlling the probabilities of Type I and Type II errors, researchers can design more reliable and valid experiments.

12 Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, given the average number of times the event occurs over that interval. It is defined by a single parameter λ (the average rate of occurrence).

The probability mass function (PMF) of the Poisson distribution is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where: - X is the random variable representing the number of events - k is the number of events - λ is the average rate of occurrence - e is the base of the natural logarithm

12.1 Example: Number of Emails

Suppose the average number of emails received per hour is 5. We want to find the probability of receiving exactly 3 emails in the next hour.

Here, $\lambda = 5$ and $k = 3$.

Using the Poisson formula:

$$P(X=3) = \frac{5^3 e^{-5}}{3!} = \frac{125 \times 0.0067}{6} \approx 0.1404$$

So, the probability of receiving exactly 3 emails in the next hour is approximately 0.1404, or 14.04

13 Confidence Intervals

A confidence interval is a range of values, derived from sample statistics, that is likely to contain the value of an unknown population parameter. The interval has an associated confidence level that quantifies the level of confidence that the parameter lies within the interval.

13.1 Confidence Interval for the Mean

For a population with a normal distribution and known standard deviation, the confidence interval for the mean μ is given by:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where: - $\bar{X}$ is the sample mean - $Z_{\alpha/2}$ is the Z-score corresponding to the desired confidence level - σ is the population standard deviation - n is the sample size
If the population standard deviation is unknown, the t-distribution is used instead of the normal distribution:

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

where: - $t_{\alpha/2,n-1}$ is the t-score with $n-1$ degrees of freedom - s is the sample standard deviation

13.2 Example: Confidence Interval for the Mean

Suppose we have a sample of 25 students with a mean test score of 80 and a standard deviation of 10. We want to calculate the 95% confidence interval for the population mean.
Using the t-distribution (since the population standard deviation is unknown):

$$\bar{X} = 80, \quad s = 10, \quad n = 25, \quad t_{\alpha/2,24} \approx 2.064$$

The confidence interval is:

$$80 \pm 2.064 \frac{10}{\sqrt{25}} = 80 \pm 4.128$$

So, the 95% confidence interval for the population mean is $[75.872, 84.128]$.

13.3 Confidence Interval for Proportions

For a population proportion p, the confidence interval is given by:

$$\hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where: - $\hat{p}$ is the sample proportion - $Z_{\alpha/2}$ is the Z-score corresponding to the desired confidence level - n is the sample size

13.4 Example: Confidence Interval for Proportions

Suppose we have a sample of 200 voters, and 120 of them support a particular candidate. We want to calculate the 95% confidence interval for the population proportion.

$$\hat{p} = \frac{120}{200} = 0.6, \quad Z_{\alpha/2} \approx 1.96$$

The confidence interval is:

$$0.6 \pm 1.96\sqrt{\frac{0.6 \times 0.4}{200}} = 0.6 \pm 0.068$$

So, the 95% confidence interval for the population proportion is $[0.532, 0.668]$.
Confidence intervals provide a useful way to estimate population parameters and quantify the uncertainty associated with sample estimates.

14 ANOVA Test

ANOVA (Analysis of Variance) is a statistical method used to compare the means of three or more groups to determine if there are any statistically significant differences between them.

14.1 One-Way ANOVA

One-way ANOVA tests the effect of a single factor on a response variable. The null hypothesis (H_0) states that all group means are equal.
The test statistic is:

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}}$$

If the calculated F-value is greater than the critical value from the F-distribution table, we reject the null hypothesis.

14.2 Example

Suppose we have three groups with the following means: $\bar{X}_1 = 5$, $\bar{X}_2 = 7$, $\bar{X}_3 = 6$. We perform a one-way ANOVA to test if the means are significantly different.
Calculate the F-value and compare it to the critical value to determine if there are significant differences between the group means.
ANOVA helps in identifying whether the differences in sample means are due to actual differences in the population means or just random variation.

15 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform a large set of variables into a smaller one that still contains most of the information in the large set. It is widely used in data analysis and machine learning for feature extraction and data visualization.

15.1 Steps in PCA

The steps involved in performing PCA are as follows:

1. **Standardize the Data:** Standardize the dataset to have a mean of zero and a standard deviation of one. This ensures that each variable contributes equally to the analysis.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where x_{ij} is the value of the i-th observation and j-th variable, $\bar{x}_j$ is the mean of the j-th variable, and s_j is the standard deviation of the j-th variable.

2. **Compute the Covariance Matrix:** Calculate the covariance matrix to understand how the variables in the dataset are related to each other.

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

   where $\mathbf{z}_i$ is the standardized data vector of the i-th observation, and $\bar{\mathbf{z}}$ is the mean vector of the standardized data.

3. **Compute the Eigenvalues and Eigenvectors:** Calculate the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors represent the directions of the principal components, and the eigenvalues represent the magnitude of the variance in these directions.

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$$

   where $\mathbf{v}$ is the eigenvector and λ is the eigenvalue.

4. **Sort Eigenvalues and Eigenvectors:** Sort the eigenvalues in descending order and arrange the corresponding eigenvectors accordingly. The eigenvectors with the largest eigenvalues are the principal components.

5. **Transform the Data:** Project the original data onto the principal components to obtain the transformed dataset.

$$\mathbf{Z}_{\text{PCA}} = \mathbf{Z}\mathbf{V}$$

   where $\mathbf{Z}$ is the standardized data matrix, and $\mathbf{V}$ is the matrix of eigenvectors.

15.2 Choosing the Number of Principal Components

The number of principal components to retain can be determined by examining the cumulative explained variance. The explained variance for each principal component is given by the ratio of its eigenvalue to the sum of all eigenvalues. The cumulative explained variance is the sum of the explained variances of the principal components.

$$\text{Explained Variance} = \frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$$

$$\text{Cumulative Explained Variance} = \sum_{i=1}^{k}\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$$

Typically, the number of principal components is chosen such that the cumulative explained variance exceeds a certain threshold (e.g., 95%).

15.3 Example: PCA on a Sample Dataset

Suppose we have a dataset with three variables and five observations:

Variable 1	Variable 2	Variable 3
2.5	2.4	3.5
0.5	0.7	1.2
2.2	2.9	3.1
1.9	2.2	2.9
3.1	3.0	3.7

1. **Standardize the Data:**

0.87	0.83	0.92
−1.43	−1.42	−1.36
0.63	1.18	0.68
0.28	0.13	0.48
1.65	1.72	1.72

2. **Compute the Covariance Matrix:**

$$\mathbf{C} = \begin{pmatrix} 1.0 & 0.99 & 0.98 \\ 0.99 & 1.0 & 0.97 \\ 0.98 & 0.97 & 1.0 \end{pmatrix}$$

3. **Compute the Eigenvalues and Eigenvectors:**

$$\lambda = \begin{pmatrix} 2.97 & 0.03 & 0.00 \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} 0.58 & -0.58 & 0.58 \\ 0.58 & 0.58 & -0.58 \\ 0.58 & -0.58 & -0.58 \end{pmatrix}$$

4. **Transform the Data:**

$$\mathbf{Z}_{\text{PCA}} = \mathbf{ZV} = \begin{pmatrix} 1.5 & -0.5 & 0.0 \\ -2.5 & 0.5 & 0.0 \\ 1.0 & 1.0 & 0.0 \\ 0.5 & -0.5 & 0.0 \\ 2.5 & -0.5 & 0.0 \end{pmatrix}$$

In this example, the first principal component explains most of the variance in the data, and the transformed dataset is now represented in terms of the principal components.

16 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique used for visualizing high-dimensional data. It is particularly well-suited for embedding high-dimensional data into a 2D or 3D space for visualization.

16.1 How t-SNE Works

t-SNE works by minimizing the divergence between two distributions: one that measures pairwise similarities of the input objects in the high-dimensional space and one that measures pairwise similarities of the corresponding low-dimensional points. The algorithm consists of the following steps:

1. **Compute Pairwise Similarities:** Compute pairwise similarities between data points in the high-dimensional space using a Gaussian distribution. For each pair of data points i and j, the similarity p_{ij} is given by:

$$p_{ij} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

where σ_i is the bandwidth parameter for point i, determined using a perplexity parameter.

2. **Compute Low-Dimensional Embedding:** Initialize the low-dimensional points y_i randomly and compute pairwise similarities using a Student's t-distribution. The similarity q_{ij} in the low-dimensional space is given by:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

3. **Minimize Kullback-Leibler Divergence:** Minimize the Kullback-Leibler divergence between the high-dimensional and low-dimensional similarity distributions using gradient descent. The Kullback-Leibler divergence $KL(P\|Q)$ is given by:

$$KL(P\|Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The gradient of the Kullback-Leibler divergence with respect to the low-dimensional points y_i is:

$$\frac{\partial KL}{\partial y_i} = 4 \sum_{j \neq i} \left(p_{ij} - q_{ij}\right) \left(y_i - y_j\right) \left(1 + \|y_i - y_j\|^2\right)^{-1}$$