Statistical Data Analysis Notes

Şafak Bilici

Contents

1 Introduction 2

2 Data Related Definitions 2
2.1 Median . . . 2
2.2 Moving Average Filter . . . 3
2.3 Moving Difference Filter . . . 3
2.4 Variance and Standard Deviation . . . 3
2.5 Percentiles . . . 4
2.6 Quartile . . . 4
2.7 Five-Number Summary . . . 4
2.8 Interquartile Range (IQR) . . . 5
2.9 Data Transformation . . . 5
2.10 Coefficient of Variation . . . 5
2.11 Scaling and Shifting Variables . . . 5
2.12 Correlation . . . 6
2.13 Sample Covariance . . . 6
2.14 Two Categorical Variables . . . 6

3 Probability 6
3.1 Joint vs. Marginal Probability . . . 6
3.2 Disjoint Events . . . 7
3.3 Conditional Probability . . . 7
3.4 The law of total probability . . . 7
3.5 Independent Events . . . 7
3.6 Bayes' Theorem . . . 7
3.7 Discrete probability distributions . . . 7
3.8 Bernoulli Distribution . . . 7
3.9 Binomial Distribution . . . 8
3.10 Poisson Distribution . . . 8
3.11 68-95-99.7 Rule . . . 8

4 Estimation 8
4.1 Sampling Distribution . . . 8
4.2 Confidence Intervals for the Population Mean . . . 9
4.3 z-critical Values . . . 9
4.4 Confidence Interval When the Population Variance is Unknown . . . 9

4.5 Margin of Error 10

5 Hypothesis Testing **10**

5.1 Null and Alternative Hypotheses 10

6 Statistical Inference for the Relationship Between Two Variables **11**

7 Analysis of Variance (ANOVA) **12**

8 Linear Regression **13**

8.1 One Numerical Explanatory Variable 13

8.2 Confidence Interval for Regression Coefficients 13

8.3 Hypothesis testing 14

1 Introduction

Categories of data analysis are

- Narrative (laws, arts etc.)
- Descriptive (social sciences)
- Statistical/mathematical
- Audio-optical
- Others

Descriptive Statistics: Uses sample information to explain/make abstraction of population. (non-parametrics)
Inferential Statistics: Uses sample statistics to infer some "phenomena" of population parameters. (parametric)
Data: unprocessed facts and figures without any added interpretation or analysis.
Information: data that has been interpreted so that it has meaning for the user.
Knowledge: a combination of information, experience and insight that may benefit the individual or the organisation.
Analysis: Analysis is defined as the procedure by which we break down an intellectual or substantial whole into parts.
Synthesis: Synthesis is defined as the procedure by which we combine separate elements or components in order to form a coherent whole.

2 Data Related Definitions

2.1 Median

- the median is more robust against outliers.
- The median may be given with its interquartile range (IQR).
- The 1st quartile point has the $\frac{1}{4}$ of the data below it
- The 3rd quartile point has the $\frac{3}{4}$ of the sample below it.
- The IQR contains the middle $\frac{1}{2}$ of the sample.

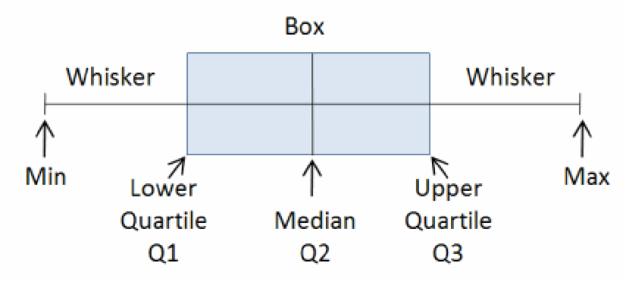


Figure 1: box and whisker plot

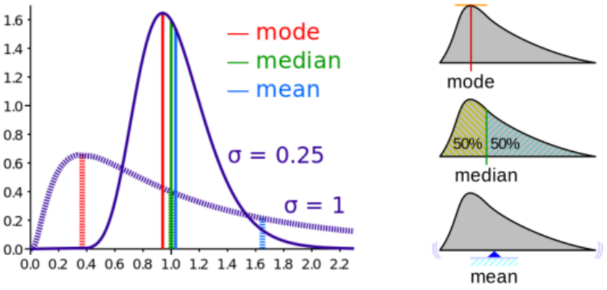


Figure 2: Geometric visualisation of the mode, median and mean of an arbitrary probability density function.

2.2 Moving Average Filter

$$y_n = \sum_{j=-k}^{k} w_j \cdot x_{n+j}$$

k is the order.

2.3 Moving Difference Filter

First order:

$$Dy_t = y_t - y_{t-1}$$

Higher order differences (2nd order):

$$D^2 y_t = D(Dy_t) = Dy_t - Dy_{t-1} = y_t - 2y_{t-1} + y_{t-2}$$

2.4 Variance and Standard Deviation

The sample variance is a common measure of dispersion based on the squared deviations:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu(x))^2}{n-1}$$

The square root of the variance is called the sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu(x))^2}{n-1}}$$

- SD indicates how much a set of values is spread around the average:
  - A range of one SD above and below the mean (abbreviated to ±1 SD) includes 68.2% of the values.
  - ±2 SD includes 95.4% of the data.
  - ± includes 99.7%.

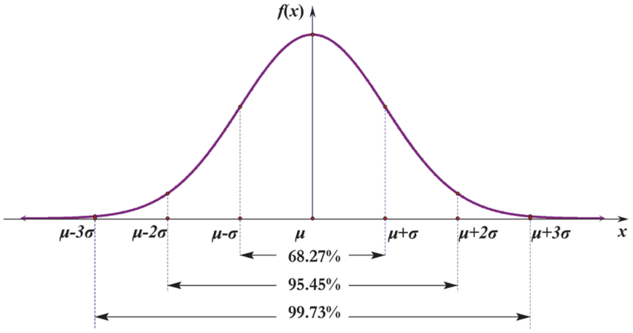


Figure 3: 3 sigma rule

2.5 Percentiles

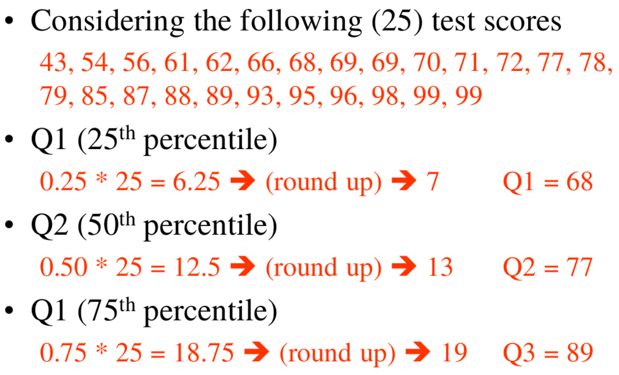
We have n sample size and we want to find kth percentile.

1. $j \leftarrow \frac{k}{100} \cdot n$
2. if j is not whole number, round up j and select the jth item.
3. else if j is whole number, average jth and $(j+1)$th item.

2.6 Quartile

For sampled data, the median is also known as the 2nd quartile, Q2. Given Q2, we can find the 1st quartile, Q1, by simply taking the median value of those samples that lie below the 2nd quartile. We can find the 3d quartile, Q3, by taking the median value of those samples that lie above the 2nd quartile.
Quartiles can also be found in terms of percentiles:

- 1st quartile is 25th percentile.
- 2nd quartile is 50th percentile.
- 3rd quartile is 75th percentile.



- Considering the following (25) test scores
  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78,
  79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99
- Q1 (25th percentile)
  0.25 * 25 = 6.25 ➔ (round up) ➔ 7 Q1 = 68
- Q2 (50th percentile)
  0.50 * 25 = 12.5 ➔ (round up) ➔ 13 Q2 = 77
- Q1 (75th percentile)
  0.75 * 25 = 18.75 ➔ (round up) ➔ 19 Q3 = 89

Figure 4: quartile example

2.7 Five-Number Summary

min, 1st quartile, median, 3rd quartile, max. This way, the five-number summary provides 0, 0.25, 0.50, 0.75, and 1 quantiles.

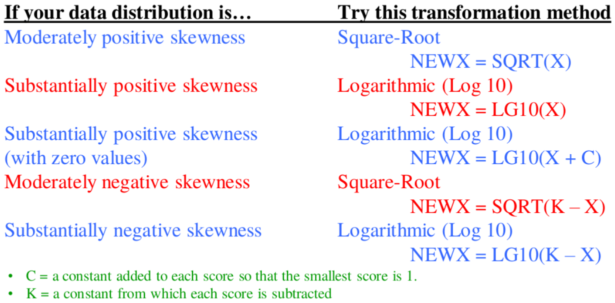
2.8 Interquartile Range (IQR)

$$IQR = Q3 - Q1$$

- If a sample is higher than $Q3 + 1.5 \cdot IQR$, then it is said to be outlier.
- If a sample is less than $Q1 - 1.5 \cdot IQR$, then it is said to be outlier.

2.9 Data Transformation

- to make the distribution of the data normal. This increases the sensitivity of statistical tests.
- logarithm or sqrt: to reduce the influence of extreme values in our analysis.
- We can create a new variable based on two or more existing variables (for example BMI).



If your data distribution is…	Try this transformation method
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

- C = a constant added to each score so that the smallest score is 1.
- K = a constant from which each score is subtracted

Figure 5: data transformation functions

2.10 Coefficient of Variation

In general, the coefficient of variation is used to compare variables in terms of their dispersion when the means are substantially different. To quantify dispersion independently from units, we use the coefficient of variation

$$CV = \frac{s}{\mu}$$

In general, when we multiply the observed values of a variable by a constant a, its mean, standard deviation, and variance are multiplied by a, $\mid a \mid$, and a^2 , respectively. The coefficient of variation is not affected.

2.11 Scaling and Shifting Variables

If we shift the observed values by b, $y = x + b$, then

- $\mu(y) = \mu(x) + b$
- $s_y = s_x$
- $s_y^2 = s_x^2$

If we multiply the observed values by the constant a and then add the constant b to the result, $y = ax + b$, then

- $\mu(y) = a\mu(x) + b$
- $s_y =\mid a \mid s_x$
- $s_y^2 = a^2 s_x^2$

2.12 Correlation

Consider a set of observed pairs of values, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. For these observed pairs of values, Pearson's correlation coefficient is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu(x)) \cdot (y_i - \mu(y))}{(n-1) \cdot s_x \cdot s_y}$$

For the two variable, s_x and s_y denote the sample standard deviations.

- The values of r are always between ± 1 and ± 1.

2.13 Sample Covariance

If the standard deviations are removed from the denominator in Pearson's correlation coefficient, the statistic is called the sample covariance,

$$v_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu(x)) \cdot (y_i - \mu(y))}{n-1}$$

2.14 Two Categorical Variables

- **difference of proportions**:
  - $p_2 - p_1$
  - We can present this difference as a percentage using the sample proportion (risk) in the p_1 group as the baseline: $\frac{p_2 - p_1}{p_1} \cdot 100\%$
- **relative proportion**
  - $\frac{p_2}{p_1}$
- It is more common to compare the sample odds:
  - $o = \frac{p}{1-p}$
- We usually compare the sample odds using the sample odds ratio
  - $OR_{21} \frac{o_2}{o_1}$
  - An odds ratio equal to 1 means that the odds are equal in both groups and is interpreted as no relationship between the two categorical variables.
  - Values of the odds ratio away from 1 (either greater than or less than 1) indicate that the relationship is strong.

3 Probability

3.1 Joint vs. Marginal Probability

- We refer to the probability of the intersection of two events, $p(E_1 \cap E_2)$ as their joint probability.
- In contrast, we refer to probabilities $p(E_1)$ and $P(E_2)$ the marginal probabilities of events:

$$p(E_1 \cup E_2) = p(E_1) + p(E_2) - p(E_1 \cap E_2)$$

3.2 Disjoint Events

Two events are called disjoint or mutually exclusive if they never occur together. Disjoint events have no elements (outcomes) in common, and their intersection is the empty set.

3.3 Conditional Probability

$$p(A \mid B) = \frac{p(A \cup B)}{p(B)}$$

$$p(A' \mid B) = 1 - p(A \mid B)$$

$$p(A \cap B \mid C) = p(A \mid C) + p(B \mid C) - p(A \cup B \mid C)$$

3.4 The law of total probability

$$p(A) = p(A \mid B_1)p(B1) + ... + p(A \mid B_k)p(B_k)$$

3.5 Independent Events

$$p(A \cap B) = p(A) \cdot p(B)$$

$$p(A \cup B) = p(A) + p(B) - p(A) \cdot p(B)$$

3.6 Bayes' Theorem

$$p(A \mid B) = \frac{p(B \mid A) \cdot p(A)}{p(B)}$$

$$p(A_i \mid B) = \frac{p(B \mid A_i) \cdot p(A_i)}{\sum_{i=0}^{k} p(A \mid B_k) \cdot p(B_k)}$$

3.7 Discrete probability distributions

The probability distribution of a discrete random variable is fully defined by the probability mass function (pmf). This is a function that specifies the probability of each possible value within range of random variable.

3.8 Bernoulli Distribution

$$X \sim Bernoulli(\theta)$$

$$p(X = x) = \begin{cases} 1 - \theta & \text{for } x = 0 \\ \theta, & \text{for } x = 1 \end{cases}$$

$$\mu = \theta$$

$$\sigma^2 = \theta \cdot (1 - \theta)$$

3.9 Binomial Distribution

A sequence of binary random variables is called Bernoulli trials.

$$X \sim Binomial(n, \theta)$$

$$p(X = x) = \binom{n}{x} p^x \cdot (1 - p)^{n-x}$$

$$\mu = n \cdot \theta$$

$$\sigma^2 = n \cdot \theta \cdot (1 - \theta)$$

3.10 Poisson Distribution

$$p(x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

3.11 68-95-99.7 Rule

The 68–95–99.7% rule for normal distributions specifies that

- 68% of values fall within 1 standard deviation of the mean:

$$P(\mu - \sigma < X \leq \mu + \sigma)$$

- 95% of values fall within 2 standard deviations of the mean:

$$P(\mu - 2\sigma < X \leq \mu + 2\sigma)$$

- 99.7% of values fall within 3 standard deviations of the mean:

$$P(\mu - 3\sigma < X \leq \mu + 3\sigma)$$

4 Estimation

4.1 Sampling Distribution

We start by assuming that the random variable of interest, X, has a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Assume that population variance σ^2 is known and we want to estimate only population mean μ. Suppose that we take a sample of size $n = 2$ from the population:

$$X_1, X_2 \sim \mathcal{N}(\mu, \sigma^2)$$

Because they are independent and identically distributed (IID), their sum is also normally distributed,

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) = \mathcal{N}(2\mu, 2\sigma^2)$$

The sample mean is $n \cdot \mu / n = \mu$ and sample variance is σ^2/n. In this case,

$$\hat{X} \sim \mathcal{N}(\mu, \sigma^2/n)$$

4.2 Confidence Intervals for the Population Mean

It is common to express our point estimate along with its standard deviation to show how much the estimate could vary if different members of population were selected as our sample.

Suppose that $\sigma^2 = 15^2$ and sample size is $n = 100$. So, the standard deviation is $\sigma / \sqrt{n} = 1.5$. Following the 68–95–99.7% rule, with 0.95 probability, the value of X is within 2 SDs from its mean, μ,

$$\mu - 2 \times 1.5 \leq \bar{X} \leq \mu + 2 \times 1.5$$

$$\mu - 3 \leq \bar{X} \leq \mu + 3$$

We are, however, interested in estimating the population mean. By rearranging the terms of the above inequality:

$$\bar{X} - 3 \leq \mu \leq \bar{X} + 3$$

We refer to this interval as our 95% confidence interval for the population mean μ. We say that the confidence level or confidence coefficient for the above interval is 0.95.

If we want to increase our confidence level to 0.997, we use the multiplier 3 since 99.7% of observations fall within 3 SDs of the mean. Therefore, our 99.7% CI for the population mean is

$$[\bar{x} - 3 \times \sigma / \sqrt{n}, \bar{x} + 3 \times \sigma / \sqrt{n}]$$

4.3 z-critical Values

In general, for a given confidence level, c, we use the standard normal distribution to find the value whose upper tail probability is $(1 - c)/2$. Then with the point estimate x, the confidence interval for the population mean at c confidence level is

$$[\bar{x} - z_{crit} \times \sigma / \sqrt{n}, \bar{x} + z_{crit} \times \sigma / \sqrt{n}]$$

4.4 Confidence Interval When the Population Variance is Unknown

To find confidence intervals for the population mean when the population variance is unknown, we use $SE = s/\sqrt{n}$ instead of $\sigma / \sqrt{n}$. t_{crit} obtained from a t-distribution with $n - 1$ degrees of freedom instead of z crit based on the standard normal distribution. The confidence interval for the population mean at c confidence level is

$$[\bar{x} - t_{crit} \times s/\sqrt{n}, \bar{x} + t_{crit} \times s/\sqrt{n}]$$

Example:
Suppose that we want to find the 95% CI for the population proportion of mothers who smoke during their pregnancy in the year 1986. Using the birthwt data set with $n = 189$, the estimate for this proportion is $x = p = 0.39$. Using p, we estimate the population variance $p(1-p) = 0.39 \times 0.61 = 0.24$. The SE for the sample mean is

$$SE = \sqrt{p(1-p)/n} = \sqrt{(0.39 \times 0.61)/183} = 3$$

The 95% CI is then

$$[0.39 - 2 \times 0.03, 0.39 + 2 \times 0.03] = [0.33, 0.45]$$

4.5 Margin of Error

In general, it is common to present interval estimates for c confidence level as

$$\text{point estimate} \quad \pm \quad \text{Margin of error}$$

When the population variance σ^2 is known, the margin of error e is calculated as

$$e = z_{crit} \times \frac{\sigma}{\sqrt{n}}$$

When the population variance is not known and we need to use the data to estimate it using the sample standard deviation, s, the margin of error is calculated as

$$e = t_{crit} \times \frac{s}{\sqrt{n}}$$

5 Hypothesis Testing

To evaluate hypotheses, we rely on estimators, their sampling distributions, their specific values from observed data.

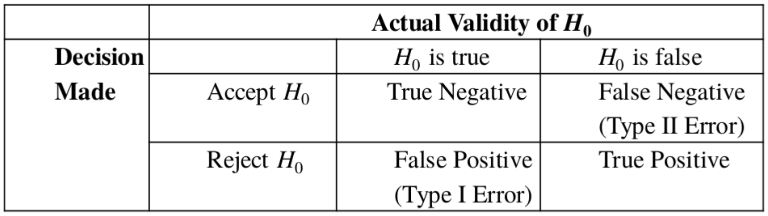
5.1 Null and Alternative Hypotheses

The null hypothesis usually reflects the “status quo” or “nothing of interest”. Denoted as H_0.
The alternative hypothesis the hypothesis we are investigating through a scientific study. Denoted as H_A.
Consider the body temperature example, where we want to examine the null hypothesis $H_0 : \mu = 98.6$ against the alternative hypothesis $H_A : \mu < 98.6$. For hypothesis testing, we focus on the null hypothesis since it tends to be simpler. To this end, we examine the evidence that the observed data provide against the null hypothesis H_0 .

- If the evidence against H_0 is strong, we reject H_0.
- If not, we state that the evidence provided by the data is not strong enough to reject H_0, and we fail to reject it.

With respect to our decision regarding the null hypothesis H_0, we might make two types of errors:

- TYPE I Error (α)
  - We reject H_0 when it is true and should not be rejected.
- TYPE II Error (β)
  - We fail to reject H_0 when it is false and should be rejected.



		Actual Validity of H_0	
Decision Made		H_0 is true	H_0 is false
	Accept H_0	True Negative	False Negative (Type II Error)
	Reject H_0	False Positive (Type I Error)	True Positive

Now suppose that we have a hypothesis testing procedure that fails to reject the null hypothesis when it should be rejected with probability β. This means that our test correctly rejects the null hypothesis with probability $1-\beta$. We refer to this probability as the power of the test. In practice, it is common to first agree on a tolerable type I error rate α, such as 0.01, 0.05, and 0.1.

Consider the body temperature example, where we want to examine the null hypothesis H_0 : $\mu = 98.6$ against the alternative hypothesis $H_A : \mu < 98.6$. To start with, suppose that $\sigma^2 = 1$ is known and we have randomly selected a sample of 25 healthy people from the population and measured their body temperature. Using the CLT, the sampling distribution of X is approximately normal as follows:

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n = 1/25)$$

If the null hypothesis is true and the population mean is $\mu = 98.6$, the sampling distribution of $\bar{X}$ becomes

$$\bar{X} \mid H_0 \sim \mathcal{N}(98.6, 0.04)$$

In reality, we have one value, x, for the sample mean. We can use this value to quantify the evidence of departure from the null hypothesis. Suppose that from our sample of 25 people we find that the sample mean is $\bar{x} = 98.4$. To evaluate the null hypothesis $H_0 : \mu = 98.6$ versus the alternative $H_A : \mu < 98.6$.

6 Statistical Inference for the Relationship Between Two Variables

In general, we can denote the means of the two groups as μ_1 and μ_2. In contrast, the alternative hypothesis is one the following:

- $H_A : \mu_1 = \mu_2$
- $H_A : \mu_1 < \mu_2$
- $H_A : \mu_1 \neq \mu_2$

By the Central Limit Theorem,

$$\bar{X}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2{}_1}{n_1}\right)$$

$$\bar{X}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma^2{}_2}{n_2}\right)$$

Therefore,

$$\bar{X}_{12} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_1}{n_2}\right)$$

We can rewrite this as

$$\bar{X}_{12} \sim \mathcal{N}(\mu_{12}, SD_{12}^2)$$

where

$$SD_{12} = \sqrt{\frac{\sigma^2{}_1}{n_1} + \frac{\sigma^2{}_1}{n_2}}$$

We want to test our hypothesis that $H_A : \mu_{12} \neq 0$ (i.e., the difference between the two means is not zero) against the null hypothesis that $H_0 : \mu_{12} = 0$. If the null hypothesis is true, then $\mu_{12} = 0$

and $\bar{X}_{12} \sim \mathcal{N}(0, SD_{12}^2)$. As before, however, it is more common to standardize the test statistic by subtracting its mean (under the null) and dividing the result by its standard deviation.

$$Z = \frac{\bar{X}_{12}}{SD_{12}} \sim \mathcal{N}(0, 1)$$

$$z = \frac{\bar{x}_{12}}{SD_{12}} \sim \mathcal{N}(0, 1)$$

Then, depending on the alternative hypothesis, we can calculate the p-value, which is the observed significance level, as:

- if $H_A : \mu_{12} > 0, \;\; p_{obs} = P(Z \leq z)$
- if $H_A : \mu_{12} < 0, \;\; p_{obs} = P(Z \leq z)$
- if $H_A : \mu_{12} \neq 0, \;\; p_{obs} = 2 \times P(Z \geq \mid z \mid)$

Example:
Suppose that our sample includes $n_1 = 25$ women and $n_2 = 27$ men. The sample mean of body temperature is $\bar{x}_1 = 98.2$ for women and $\bar{x}_2 = 98.4$ for men. Then, our point estimate for the difference between population means is $\bar{x}_1 2 = -0.2$. We assume that $\sigma^2{}_1 = 0.8$ and $\sigma^2{}_2 = 1$. The variance of the sampling distribution is (0.8/25)+(1/27) = 0.07, and the standard deviation is $SD_{12} = \sqrt{0.07} = 0.26$. The z-score is

$$z = \frac{\bar{x}}{SD_{12}} = \frac{-0.2}{0.26} = -0.76$$

$H_A : \mu_{12} \neq 0$ and $z = -0.76$. Therefore $p_{obs} = 2p(Z \geq \mid z \mid) = 2 \times 0.22 = 0.44$. For the body temperature example, $p_{obs} = 0.44$ is greater than the commonly used significance levels (e.g., 0.01, 0.05, and 0.1). Therefore, the test result is not statistically significant, and we cannot reject the null hypothesis (which states that the population means for the two groups are the same) at these levels.

7 Analysis of Variance (ANOVA)

The process of evaluating hypotheses regarding the group means of multiple populations is called the Analysis of Variance (ANOVA). ANOVA models generalize the t-test and are used to compare the means of multiple groups identified by a categorical variable with more than two possible categories. Since we are only considering one factor only, this method is specifically called one way ANOVA. In general, the **between-groups variation** is denoted as SS_B and calculated by

$$SS_B = \sum_{i=1}^{k} n_i \cdot (\bar{y}_i - \bar{y})^2$$

where k is the number of groups. The **within-groups variation** is denoted as SS_W and calculated by

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_i)^2$$

We measure the **total variation** in Y by

$$SS_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

The test statistic for examining the null hypothesis is called F- statistic (more specifically, ANOVA F -statistic) and is defined as

$$F = \frac{SS_B/(k-1)}{SS_W/(n-k)}$$

For the one-way ANOVA, the F-statistic has $F(df_1 = k-1, df_2 = n-k)$ distribution under the null hypothesis (i.e., assuming that the null hypothesis is true). The F-distribution, which is a continuous probability distribution, is very important for hypothesis testing. We refer to df_1 and df_2 as the numerator degrees of freedom and denominator degrees of freedom, respectively.

Example: As an example, we analyze the Cushings data set. The Type variable in the data set shows the underlying type of syndrome, which can be one of four categories: adenoma (a), bilateral hyperplasia (b), carcinoma (c), unknown (u). Objective is to find whether the four groups are different with respect to urinary excretion rate of Tetrahydrocortisone. We denote by Y the urinary excretion rate of Tetrahydrocortisone and by X the Type variable, where $X = 1$ for Type = a, $X = 2$ for Type = b, $X = 3$ for Type = c, and $X = 4$ for Type = u. Denote the individual observations as y_ij : the urinary excretion rate of Tetrahydrocortisone of the jth individual in group i. Total number of observations is $n = 27$, the number of observations in each group is $n_1 = 6$, $n_2 = 10$, $n_3 = 5$, and $n_4 = 6$. The overall (for all groups) observed sample mean for the response variable is $\bar{y} = 10.46$. We also find the group specific means: $\bar{y}_1 = 3.0$, $\bar{y}_2 = 8.2$, $\bar{y}_3 = 19.7$, $\bar{y}_4 = 14.0$. So, the degrees of freedom parameters are $df_1 = 4 - 1 = 3$ and $df_2 = 27 - 4 = 23$. $SS_B = 893.5$ and $SS_W = 2123.6$. The observed value of F-statistic is $f = 3.2$ given under the column labeled F value. The resulting p-value is then 0.04. Therefore, we can reject H_0 at 0.05 significance level (but not at 0.01) and conclude that the differences among group means for urinary excretion rate of Tetrahydrocortisone are statistically significant (at 0.05 level).

8 Linear Regression

8.1 One Numerical Explanatory Variable

$$y = \alpha + \beta X + \varepsilon$$

First, we find the slope of regression line using the sample correlation coefficient r between the response variable Y and and the explanatory variable X

$$\beta = r\frac{s_x}{s_y}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

As an alternative way, the least-squares estimates of slope and intercept can be obtained as follows:

$$\beta = \frac{S_{xy}}{S_{xx}}$$

$$\alpha = \bar{y} - \beta\bar{x}$$

where $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \sum_i (x_i - \bar{x})^2$

8.2 Confidence Interval for Regression Coefficients

$$[\beta - t_{crit} \times SE_\beta, b + t_{crit} \times SE_\beta]$$

SE_β is obtained from

$$SE_\beta = \frac{\sqrt{RSS/(n-2)}}{\sum_i (x_i - \bar{x})^2}$$

8.3 Hypothesis testing

To assess the null hypothesis that the population regression coefficient is zero, which is interpreted as no linear relationship between the response variable and the explanatory variable, we first calculate the t-score.

$$t = \frac{\beta}{SE_{\beta}}$$

Then, we find the corresponding p-value as follows:

- if $H_A : \beta < 0$ then $p_{obs} = p(T \leq t)$
- if $H_A : \beta > 0$ then $p_{obs} = p(T \geq t)$
- if $H_A : \beta \neq 0$ then $p_{obs} = 2p(T \geq \mid t \mid)$