University of Twente

Statistics for Engineers

Summary of the reader

W. van den Brink
Based on the 2021 version of the Intelligent Interaction Design module

December 22, 2021

Contents

Chapter 1

Descriptive statistics

1.1 Introduction

Definition 1
If a $X_1, \ldots, X_n$ is a random sample of X, or: from the distribution of X, then:

1. $X_1, \ldots, X_n$ are independent
2. $X_1, \ldots, X_n$ all have the same distribution as X (the population distribution)

1.2 Numerical measures, histogram and bar graphs

Different kinds of variables:

- **Quantitative** (or **numerical**) variables (possibly with an interval scale)
- **Qualitative** (or **categorical**) variables, with different scales:
  - **Ordinal** scale, where the order of the categories matters
  - **Nominal** scale, where the order does not matter

The mean of categorigal variables does not exist. The **sample mode**, the most frequently occuring category, does exist.

We can define **sample variables** for the whole sample. These are random variables.

For discrete variables, we represent the sample observations graphically in a bar graph. For continuous variables we use a histogram.

A bar graph can be considered as an experimental (estimated) probability function. A histogram can be considered an experimental density function.

Rule of Thumb 1 (Determining the number of intervals in a historgram)
For n observations of a continuous variable, a historgram with about $\sqrt{n}$ (equally large) intervals is constructed.

Definition 2 (Order statistics)
The **order statistics** $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ is an order of the observations $x_1, x_2, \ldots, x_n$, such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Definition 3 ((Sample) median)
The (sample) median is

$$m = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[x_{\left(\frac{1}{2}n\right)} + x_{\left(\frac{1}{2}n+1\right)}\right] & \text{if } n \text{ is even} \end{cases}$$

In words: the (sample) median is the middle observation, or, in the case of an even amount of observations, the mean of the two middle observations.

Definition 4 (k^{th} percentile)
For the k^{th} percentile of n ordered observations:

- At least k% of the observations are less than or equal to the k^{th} percentile
- At least $(100-k)$% is greater than or equal to the k^{th} percentile

To compute the k^{th} percentile of n observations $x_1, x_2, \ldots, x_n$:

1. Compute k% of n: $c = \frac{k}{100} \cdot n$.
2. If c is not integer, round c upward to the first larger integer $[c]$: the k^{th} percentile is $k_{([c])}$.
3. If c is integer, then the k^{th} percentile is $\frac{x_{(c)}+x_{(c+1)}}{2}$.

Definition 5 (Quartiles)
We define the following quartiles:

- The lower quartile Q_1 (or Q_{L} is the 25^{th} percentile.
- The second quartile Q_2 is equal to the median m.
- The upper quartile Q_3 (or Q_{U} is the 75^{th} percentile.

1.2.1 Measures for the centre

There are three measures for the centre:

- The **sample mean**: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- The **median**: Definition 3.
- The **mode**: the most frequently occuring observation.

When a graph is skewed to the right, we have $\bar{x} < m$. Otherwise, we have $\bar{x} > m$. We say that the median is resistant: it is not sensitive to extreme observations or outliers.

1.2.2 Measures for the variance

There are three measures for variance:

- The sample variance
- The standard deviation
- The Inter Quartile Range

Definition 6 (Sample variance of observations)
The sample variance of the observations $x_1, \ldots, x_n$ is

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

Definition 7
(Sample standard deviation) The sample standard deviation of $x_1, \ldots, x_n$ is $s = \sqrt{s^2}$.

s and s^2 are non-negative. They are equal to 0 iff all observed values are the same.

Definition 8 (Inter Quartile Range)

$$\text{IQR} = Q_3 - Q_1$$

Reasoning: in the interval (Q_1, Q_3) there are (about) 50% of the observations: the "middle" 50% of the sample distribution. The IQR is the width (range) of this interval.

Chebyshev's rule for all distributions and the Empirical rule for mound shaped distributions apply to both probability distributions (μ and σ^2) and data distributions ($\bar{x}$ and s^2).

Definition 9 (Chebyshev's rule)
For any set of observations $x_1, \ldots, x_n$, the proportion of observations within the interval $(\bar{x} - k \cdot s, \bar{x} + k \cdot s)$ is at least $1 - \frac{1}{k^2}$.

This inequality is informative for all integer and rational numbers $k > 1$.

Definition 10
If the sample is $x_1, \ldots, x_n$, the z-score of an observation x is $x = \frac{x - \bar{x}}{s}$.

Interpret as follows:

- A z-score -3 means that the observation x is three standard deviations less than the sample mean (quite extreme according to the empirical rule): $x = \bar{x} - 3 \cdot s$.
- $z = 1.4$ means: x is 1.4 standard deviations larger than $\bar{x}$: $x = \bar{x} + 1.4 \cdot s$.

1.3 Classical numerical summary

Definition 11 (k^{th} central moment of X)
$\mathrm{E}\,(X - \mu)^k$ is called the k^{th} central moment of X.

- The first central moment ($k = 1$) is always 0: $\mathrm{E}\,(X - \mu) = \mathrm{E}\,(X) - \mu = 0$.
- The second central moment ($k = 2$) is per definition the variance: $\mathrm{E}\,(X - \mu)^2 = \mathrm{var}\,(X)$.
- The third central moment $\mathrm{E}\,(X - \mu)^3$ gives information about the symmetry of the distribution: if the distribution is symmetric, such as the normal and the uniform distribution, this central moment is 0. If the distribution is skewed to the right, it is positive. And negative if the distribution is skewed to the left.
- The fourth central moment $\mathrm{E}\,(X - \mu)^4$ is larger if the tails of the distribution are thicker.

The third and fourth central moment depend on the chosen scale and should be divided by σ^3 and σ^4, respectively.

Definition 12 (Skewness (coefficient) of X)

$$\gamma_1 = \frac{\mathrm{E}\,(X - \mu)^3}{\sigma^3}$$

is the skewness (coefficient) of X.

Definition 13 (Kurtosis of X)

$$\gamma_2 = \frac{\mathrm{E}\,(X - \mu)^4}{\sigma^3}$$

is the kurtosis of X.

Measure for	**Centre** μ	**Variation** σ^2	**Skewness** γ_1	**Kurtosis** γ_2
$\mathrm{U}\,(a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0	1.8
$\mathrm{N}\left(\mu, \sigma^2\right)$	μ	σ^2	0	3
$\mathrm{Exp}\,(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	2	9

Table 1.1: Values of important distributions

Definition 14
The classical nummerical summary of observations $x_1, \ldots, x_n$ consists of:

Sample size	n
Sample mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Sample variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$
Sample standard deviation	$s = \sqrt{s^2}$
Sample skewness coefficient	$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}}$
Sample kurtosis	$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$

The skewness coefficient gives information about the symmetry of the distribution:

- if the distribution is symmetric, its value is 0, so for a sample taken from a symmetrical distribution it should be close to 0.
- if the distribution is skewed to the right, it is positive.
- if the distribution is skewed to the left, it is negative.

The kurtosis attains larger values if the tail (or both tails) of a distribution is thicker, that the probability of (very) large or small values is relatively large.

1.4 Outliers, Box plot and Stem-and-leaf plot

Definition 15
The 5-numbers summary of $x_1, \ldots, x_n$ is $x_{(1)}, Q_1, m, Q_3$ and $x_{(n)}$.

It is graphically presented as a so called box plot.

Definition 16 (The $1.5 \times \mathrm{IQR}$-rule for determination of outliers)
Observations outside the interval $(Q_1 - 1.5 \times \mathrm{IQR}, Q_3 + 1.5 \times \mathrm{IQR})$ are outliers.

Definition 17 (The $3 \times \mathrm{IQR}$-rule for determination of extremes)
Observations outside the interval $(Q_1 - 3 \times \mathrm{IQR}, Q_3 + 3 \times \mathrm{IQR})$ are extremes.

1.5 Q-Q plots

Definition 18
A uniform Q-Q plot is a graph of n points $\left(x_{(i)}, \mathrm{E}X_{(i)}\right)$, where the ordered observation $X_{(i)}$ is the X coordinate, and its expected value $\mathrm{E}\left(X_{(i)}\right) = \frac{i}{n+1}$ according to the $U(0,1)$-distribution is the Y coordinate $(i = 1, 2, \ldots, n)$.

Definition 19
An exponential Q-Q plot is a graph of n points $\left(x_{(i)}, \mathrm{E}X_{(i)}\right)$ of ordered observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ on the X-axis and their expected values $\mathrm{E}\left(X_{(i)}\right)$, according to the exponential distribution, on the Y-axis

Definition 20
A normal Q-Q plot is a graph of n points $\left(x_{(i)}, \mathrm{E}X_{(i)}\right)$, where the ordered observation $X_{(i)}$ is the X coordinate, and its expected value $\mathrm{E}\left(X_{(i)}\right) = \mu + \sigma \times \Phi^{-1}\left(\frac{i}{n+1}\right)$ is the Y-coordinate.

Generally, if the points do not deviate from the line $y = x$ too much, the assumption of the corresponding distribution is confirmed.

Chapter 2

Estimation

2.1 Important Results Probability Theory

If we have a random sample, taken from a population with expectation μ and variance σ^2, it follows that the summation $X_1 + X_2 + \cdots + X_n$ has an expectation $n\mu$ and variance $n\sigma^2$.

Furthermore, if the transition to other units of measurement or linear transformation of the variables is considered, then the following properties apply:

- $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$
- $\mathrm{var}(aX + b) = a^2\mathrm{var}(X)$

2.1.1 The Normal Model and Random Samples from the Normal Distribution

Many statistical techniques are based on the assumption of a normal model of variables.

$X \sim N\left(\mu, \sigma^2\right)$ means that the population shows a bell shaped distribution, symmetric around the line $x = \mu$ and having a standard deviation σ. The Empirical rule applies:

Interval	**Probability** of "variable in interval"
$(\mu - \sigma, \mu + \sigma)$	$\approx 68\%$
$(\mu - 2\sigma, \mu + 2\sigma)$	$\approx 95\%$
$(\mu - 3\sigma, \mu + 3\sigma)$	$\approx 99.7\%$

A normal distribution can be standardized:

$$X \sim N\left(\mu, \sigma^2\right) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

We can then use the $N(0, 1)$ table in the reader to find $P(Z \leq z), z \geq 0$.

Property 1
For a random sample, taken from a $N\left(\mu, \sigma^2\right)$-distribution, we have:

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right) \text{ and } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

2.1.2 The Binomial Distribution and the Normal Approximation of the Binomial Probabilities

Property 2
If $X \sim B(n, p)$, then we have approximately (CLT) for sufficiently large n $(n \geq 25, np > 5, n(1-p) > 5$:

$$X \sim N\left(np, np(1-p)\right) \text{ and } \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

2.2 Estimates, Estimators and their Properties

- The **sample mean** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a (point) estimate of the population mean or expectation $\mu = \mathrm{E}(X)$, if $x_1, x_2, \ldots, x_n$ are the observations.
- The **sample variance** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$ is an estimate of the population variance $\sigma^2 = \mathrm{var}(X)$
- The **sample standard deviation** $s = \sqrt{s^2}$ is an estimate of σ.
- The **sample proportion** $\hat{p} = \frac{x}{n}$ is an estimate of the population proportion p (success rate). x is the observed number of successes, a realization of the binomial number X, which can be written as $x = \sum_{i=1}^{n} x_i$. Here x_i is the 1-0-alternative for each Bernoulli trial.

  So: $\hat{p} = \sum_{i=1}^{n} x_i = \overline{x}$.

Definition 21
An estimator T of the population parameter θ is a statistic $T(X_1, X_2, \ldots, X_n)$.

An estimate t is the observed value $T(x_1, x_2, \ldots, x_n)$ of T.

Definition 22
T is an unbiased estimator of θ if $\mathrm{E}(T) = \theta$.

2.3 Comparing Estimators

Definition 23
The Mean Squared Error of an estimator T of the parameter θ is $\mathrm{E}(T-\theta)^2$.

Notation: MSE, $\mathrm{MSE}(T)$.

T_1 is better than T_2 if $\mathrm{MSE}(T_1) < \mathrm{MSE}(T_2)$.

Definition 24
The bias of an estimator T of the parameter θ is equal to $\mathrm{E}(T) - \theta$. It tells us whether the estimator overestimates or underestimates θ.

Property 3

$$\mathrm{MSE}(T) = (\mathrm{E}T - \theta)^2 + \mathrm{var}(T)$$

Chapter 3

Confidence intervals

3.1 Introduction

3.2 Confidence Interval for the Population Mean μ

An interval estimate is an interval in which an unknown variable lies with a certain level of confidence.

The estimation error is half the length of the interval.

The measurement error is the difference between the estimator and the real value.

3.3 Confidence Interval for the Variance σ^2

Assuming a normal distribution, we can create an interval estimate for mu. There are two ways to do this, depending on whether σ^2 is known.

3.3.1 Confidence interval for μ if σ^2 is known

We assume we have a random sample $X_1, X_2, \ldots, X_n$ taken from the $N\left(\mu, \sigma^2\right)$-distribution, with unknown μ and known σ^2.

3.4 Confidence Interval for the Population Proportion p

3.5 Overview of confidence intervals in case of one sample problem

Population model	**Parameter**	**Confidence interval**	c **from the**
$N\left(\mu, \sigma^2\right)$	μ, if o^2 is known	$\left(\overline{X} - c \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + c \cdot \frac{\sigma}{\sqrt{n}}\right)$	$N(0,1)$-table
	μ, if o^2 is unknown	$\left(\overline{X} - c \cdot \frac{S}{\sqrt{n}}, \overline{X} + c \cdot \frac{S}{\sqrt{n}}\right)$	t_{n-1}-table
	σ^2, if μ is unknown	$\left(\frac{(n-1)S^2}{c_1}, \frac{(n-1)S^2}{c_2}\right)$	χ^2_{n-1}-table
	σ, if μ is unknown	$\left(\sqrt{\frac{(n-1)S^2}{c_1}}, \sqrt{\frac{(n-1)S^2}{c_2}}\right)$	χ^2_{n-1}-table
Dichotomous, proportion p	p	$\left(\hat{p} - c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$	$N(0,1)$-table

Chapter 4

Hypothesis tests

4.1 Tests on μ for known σ^2: introduction of concepts

Definition 25
The 8-step testing procedure:

1. Give a probability model of the observed values (the statistical assumptions).
2. State the null hypothesis and the alternative hypothesis, using parameters in the model.
3. Give the proper test statistic.
4. State the distribution of the test statistic if H_0 is true.
5. Compute (give) the observed value of the test statistic.
6. State the test:
   (a) Determine the rejection region, or,
   (b) Compute the p-value.
7. State your statistical conclusion: reject or fail to reject H_0 at the given significance level.
8. Draw the conclusion in words.

Definition 26
The probability of a type I error is the probability of rejecting H_0, though it is true.

Definition 27
The probability of a type II error is the probability of accepting H_1 while H_0 should not be rejected.

Definition 28
The power of a test is the probability of a correct rejection of the null hypothesis. It is equal to 1 minus the probability of a type II error.

4.2 Tests on the population mean μ, if σ^2 is unknown

Definition 29
If we test $H_0 : \mu = \mu_0$, based on a random sample of the normal distribution with unknown expectation μ and unknown variance σ^2, the test statistic is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

Property 4
If in Definition 29 H_0 is true, the test statistic has a t-distribution: $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{n-1}$.

4.3 Tests on the variance σ^2

4.4 Test on the population proportion p

Chapter 5

Two samples problems

5.1 The difference of two population proportions

Given a probability model of two independent bionmial variables $X \sim \mathcal{B}(n_1, p_1)$ and $Y \sim \mathcal{B}(n_2, p_2)$, we can estimate the difference $p_1 - p_2$ with the estimator $\frac{X}{n_1} - \frac{Y}{n_2}$, often denoted as $\hat{p}_1 - \hat{p}_2$.

Theorem 1
The estimator $\hat{p}_1 - \hat{p}_2$ is unbiased.

Proof.

$$\mathrm{E}\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) = \mathrm{E}\left(\frac{X}{n_1}\right) - \mathrm{E}\left(\frac{Y}{n_2}\right) = p_1 - p_2$$

This follows from the property that $\frac{X}{n_1}$ is an unbiased estimator for p_1.

Likewise, $\mathrm{E}\left(\frac{Y}{n_2}\right) = p_2$. □

Furthermore, we find that

$$\begin{aligned}\mathrm{var}\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) &\overset{\text{ind.}}{=} \mathrm{var}\left(\frac{X}{n_1}\right) + \mathrm{var}\left(\frac{Y}{n_2}\right) \\ &= \frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}\end{aligned}$$

From these properties, we find that, for sufficiently large n_1 and large n_2, both $\frac{X}{n_1}$ and $\frac{Y}{n_2}$ are normally distributed. Consequently, $\frac{X}{n_1} - \frac{Y}{n_2} \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$.

Thus, approximately (using the CLT):

$$Z = \frac{\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{N}(0,1)$$

5.1.1 Construction of a confidence interval for the difference $p_1 - p_2$ of two population proportions (for large samples)

Definition 30
The estimated standard deviation of $\hat{p}_1 - \hat{p}_2$ is as follows:

$$\sqrt{\frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}}$$

Notation: $\hat{\sigma}_{\hat{p}_1-\hat{p}_2}$ or $\mathrm{SE}(\hat{p}_1-\hat{p}_2)$

Property 5 (approximate confidence interval for the difference of two population proportions)
If $X \sim \mathcal{B}(n_1, p_1)$ and $Y \sim \mathcal{B}(n_2, p_2)$ are independent, then for lage n_1 and n_2:

$$(1-\alpha)100\% - \mathrm{CI}(p_1 - p_2) = \left(\hat{p}_1 - \hat{p}_2 - c\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + c\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right)$$

where c is such that $P(Z \leq c) = 1 - \frac{\alpha}{2}$.

Rule of Thumb 2
Apply Property 5 for sufficiently "large n_1 and n_2", as before: $n \geq 25$, $n\hat{p} > 5$, and $n(1-\hat{p}) > 5$, but now for both pairs $(n_1, \hat{p}_1)$, and $(n_2, \hat{p}_2)$.

5.1.2 Test on the equality of two population proportions p_1 and p_2 (for large samples)

Consider the null hypothesis $H_0 : p_1 = p_2$. We can then use the standardized difference $\hat{p}_1 - \hat{p}_2$ to construct a test.

Theorem 2
The test statistic to test $H_0 : p_1 = p_2$ is:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0,1)$$

approximately if $H_0 : p_1 = p_2$ is true.

Proof. From $H_0 : p_1 = p_2$, it follows that $\mathrm{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2 = 0$ under H_0. Then $p_1 = p_2 = p$, and $\hat{p}_1$ and $\hat{p}_2$ estimate the same unknown p.

Using both samples we can add both the number of successes and the sample sizes: $\hat{p} = \frac{X+Y}{n_1+n_2}$.

The standard deviation is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \overset{p_1=p_2=p}{=} \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \overset{\text{est.}}{\approx} \sqrt{\hat{p}\,(left(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Together, this results in the test statistic Z. □

When applying this test, consider the following:

- If $H_1 : p_1 > p_2$, then the test is upper-tailed.
- If $H_1 : p_1 < p_2$, then the test is lower-tailed.
- If $H_1 : p_1 \neq p_2$, then the test is two-tailed.

5.2 The difference of two population means

In this section we consider two independent random samples, taken from the normally distributed variables X and Y in (two) populations.

We have a probability model of two independent random samples, drawn from normal distributions:

- $X_1, \ldots, X_n$ is a random sample of $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$.

- $Y_1, \ldots, Y_n$ is a random sample of $Y \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$.
- $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ are independent.

We use the following notations for the sample means and variances:

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad S_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2, \qquad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i, \qquad S_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(Y_i - \overline{Y}\right)^2$$

An estimator for $\mu_1 - \mu_2$ is $\overline{X} - \overline{Y}$:

$$\mathrm{E}\left(\overline{X} - \overline{Y}\right) = \mathrm{E}\left(\overline{X}\right) - \mathrm{E}\left(\overline{Y}\right) = \mu_1 - mu_2$$

Since the X_is are independent and all $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$, the sample mean is normally distributed as well:

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$

And likewise for $\overline{Y}$.

So $\mathrm{var}\left(\overline{X} - \overline{Y}\right) \overset{\text{ind.}}{=} \mathrm{var}\left(\overline{X}\right) + \mathrm{var}\left(\overline{Y}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

In conclusion:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{sigma_2^2}{n_2}\right)$$

from this it follows that:

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

5.2.1 Confidence interval for $\mu_1 - \mu_2$ and a test on $\mu_1 - \mu_2$ for equal, but unknown variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Property 6 (confidence interval and test for the difference of 2 population means)
For the probability model of two independent samples, drawn from normal distributions with equal, but unknown variance, we have:

1. The pooled sample variance

   $$S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_Y^2$$

   is the best variance estimator.

2. $$(1 - \alpha)100\% - \mathrm{CI}\left(\mu_1 - \mu_2\right) = \left(\overline{X} - \overline{Y} - c\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2},\right)}, \overline{X} - \overline{Y} + c\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2},\right)}\right)$$

   where $P\left(T_{n_1+n_2-2} \geq c\right) = \frac{\alpha}{2}$.

3. If we test on $H_0 : \mu_1 - \mu_2 = \Delta_0$, the test statistic $T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$.

5.3 Test on the equality of variances

To test whether the assumption of equal variances of two populations is justified, we choose the following hypotheses:

Test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Property 7
For a test on $H_0 : \sigma_1^2 = \sigma_2^2$ in a model of two independent random samples, drawn from normal distributions, the test statistic and its distribution are as follows:

$$F = \frac{S_X^2}{S_Y^2} \sim \mathcal{F}_{n_2-1}^{n_1-1}, \qquad \text{if } H_0 : \sigma_1^2 = \sigma_2^2 \text{ is true}$$

5.4 Paired samples

If two random samples are independent, we find:

$$\text{var}\left(\overline{X} - \overline{Y}\right) \overset{\text{ind.}}{=} \text{var}\left(\overline{X}\right) + \text{var}\left(\overline{Y}\right)$$

If the samples are independent, then:

$$\text{var}\left(\overline{X} - \overline{Y}\right) \overset{\text{ind.}}{=} \text{var}\left(\overline{X}\right) + \text{var}\left(\overline{Y}\right) - 2\text{cov}\left(\overline{X}, \overline{Y}\right)$$

with $-2\text{cov}\left(\overline{X}, \overline{Y}\right)$ a (usually unknown) covariance term. In this case, the distribution of $\overline{X} - \overline{Y}$ cannot be determined.

When we have pairs of dependent observations, we naturally switch to the difference of each of the observed pairs, which can be assumed independent. If, in addition, the normal distribution applies to the differences, we can apply the one sample t-test on the mean difference.

Chapter 6

Chi-square tests

6.1 Testing on a specific distribution with k categories

The multinomial distribution is based on independent trials with k outcomes numbered $1, 2, \ldots, k$ and with probabilities $p_1, p_2, \ldots, p_k$. The probability function is a generalization of the binomial case.

$$P\left(N_1 = n_1, \ldots, N_k = n_k\right) = \frac{n!}{n_1! \cdot \ldots n_k!} p_1^{n_1} \cdot \ldots p_k^{n_k}$$

We have the following conditions for the totals:

- $n_1 + \cdots + n_k = n$, and,
- $p_1 + \cdots + p_k = 1$.

The numbers N_i are dependent. Furthermore, $N_i \sim \mathcal{B}\left(n, p_i\right)$.

Property 8
If the numbers $N_1, \ldots, N_k$ $(k \geq 2)$ have a multinomial distribution with success rates $p_1, \ldots, p_k$, total number n and expected values $\mathrm{E}N_i = n\, p_i$, then the variable

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(N_i - \mathrm{E}N_i\right)^2}{\mathrm{E}N_i} \sim \chi^2_{k-1}$$

has an approximate Chi-squared distribution with $k - 1$ degrees of freedom.

A condition (rule of thumb) for this approximation is $\mathrm{E}N_i \leq 5$, $i = 1, \ldots, k$.

We introduce the notation $\mathrm{E}_0 N_i$: the expectation of N_i, if H_0 is true.

The test statistic for the test on $H_0 : p_i = p_{i0}$ $(i = 1, \ldots, k)$ is:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(N_i - \mathrm{E}_0 N_i\right)^2}{\mathrm{E}_0 N_i}$$

Provided that $\mathrm{E}_0 N_i \geq 5, i = 1, \ldots, k$.

Example 1 (Testing the fairness of a dice with 120 rolls)
Apply the 8-step hypothesis test:

(1) Model: $N_i =$"the number of rolls with i as result (face up)", $i = 1, \ldots, 6$.

$N_1, \ldots, N_6$ have a multinomial distribution with $n = 120$ trials and unknown probabilities $p_1, \ldots, p_6$ for the six possible outcomes.

(2) We test $H_0 : p_i = \frac{1}{6}$ for $i = 1, \ldots, 6$, versus $H_1 : p_i \neq \frac{1}{6}$ for at least one i, with $\alpha = 0.05$.

(3) The test statistic is $\chi^2 = \sum_{i=1}^{6} \frac{(N_i - \mathrm{E}_0 N_i)^2}{\mathrm{E}_0 N_i}$, where $\mathrm{E}_0 N_i = n \cdot \frac{1}{6} = 20 (i = 1, \ldots, 6)$.

(4) Distribution under H_0: $\chi^2 \sim \chi^2_{6-1}$.

(5) Observed value of χ^2:

Number face up	i	1	2	3	4	5	6	**Total**
Number of times	n_i	18	15	23	22	17	25	$120 = n$
Expactation if dice is fair	$\mathrm{E}_0(N_i) = n \cdot \frac{1}{6}$	20	20	20	20	20	20	$120 = n$

So:

$$\chi^2 = \frac{(18-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(23-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(25-20)^2}{20} = \frac{76}{20} = 3.8$$

(6) The test is: reject H_0 if $\chi^2 \geq c$, where $c = 11.1$ in the χ^2_5-table, such that $P\left(\chi^2_5 \geq c\right) = \alpha = 0.05$.

(7) $\chi^2 = 3.8 < 11.1$: χ^2 does not lie in the rejection region, so H_0 is not rejected.

(8) We could not show convincingly that die dice is not fair, at a 5% significance level.

In summary, the Chi-square test allows us to test whether a fully specified distribution applies. An important condition for applying the test is that the expected values $\mathrm{E}_0 N_i$ of all categories should be at least 5.

6.2 Chi-square tests for cross tables

Categorical values assume one of two or more categories. Given two categorical values, we can analyze the relative magnitude and the relation of the values. Observed numbers are usually presented in a cross table.

A $r \times c$ cross table contains r rows and c columns. The number n_{ij}, with i the row number ($1 \leq i \leq r$) and j the column number ($1 \leq j \leq c$), contains the observed number with a specific combination of categories.

		Variable 2		
		$\ldots$	j	$\ldots$
	$\vdots$		$\vdots$	
Variable 1	i	$\ldots$	n_{ij}	$\ldots$
	$\vdots$		$\vdots$	

The total observed number with a certain "row category" i is the row total $n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{ic}$.

Likewise, the total observed number with a certain "column category" j is the column total $n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{rj}$.

The joint probabilities for each cell of the table can be calculated with $\frac{n_{ij}}{n}$. And computing $\frac{n_{\bullet j}}{n}$ and $\frac{n_{i\bullet}}{n}$ gives the marginal distributions of the column and row categories, respectively.

These probabilities estimate the population probabilities p_{ij}.

We can also determine the conditional distributions for a row or column by normalizing the values in each row or column with the total amount of observations in the respective row or column.

From a cross table we can ask questions about the relative proportions: are they significantly different? We use a Chi-square test to verify this. To determine the expected value for every cell, we use the following formula:

$$\hat{\mathrm{E}}_0 N_{ij} = \frac{n_{i\bullet} \times n_{\bullet 1}}{n} = \frac{\text{row total} \times \text{column total}}{n}$$

Generalizing, the test is as follows:

Property 9 (Test on the independence of two variables in a $r \times c$-cross table)
If the numbers N_{ij} ($i = 1, \ldots, r$, and $j = 1, \ldots, c$) have a multinomial distribution with sample size n and unknown probabilities p_{ij}, then the test on $H_0 : p_{ij} = p_{i\bullet} \times p_{\bullet j}$ for all (i, j) against $H_1 : p_{ij} \neq p_{i\bullet} \times p_{\bullet j}$ for at least one pair (i, j) (essentially doubting the independence of the categorical variables), can be executed with the following test statistic:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(N_{ij} - \hat{\mathrm{E}}_0 N_{ij}\right)^2}{\hat{\mathrm{E}}_0 N_{ij}}$$

where

$$\hat{\mathrm{E}}_0 N_{ij} = \frac{n_{i\bullet} \times n_{\bullet 1}}{n} = \frac{\text{row total} \times \text{column total}}{n}$$

and $\chi^2 \sim \chi^2_{(r-1)(c-1)}$ (χ^2 has a Chi-square distribution with $df = (r-1)(c-1)$.

When the rows of a $2 \times c$-cross table are independent, we do not test for independence, but we test for homogeneity. This changes step 1 and 2 of the testing procedure:

(1) The model is as follows:

- The two random samples are independent.
- The numbers $N_{11}, N_{12}, \ldots, N_{1c}$ in the first row are multinomially distributed with total n_1 and probabilities $p_{11}, p_{12}, \ldots, p_{1c}$.
- The numbers $N_{21}, N_{22}, \ldots, N_{2c}$ in the second row are multinomially distributed with total n_2 and probabilities $p_{21}, p_{22}, \ldots, p_{2c}$.

(2) Test

$$H_0 : p_{11} = p_{21} \text{ and } p_{12} = p_{22} \text{ and } \ldots \text{ and } p_{1c} = p_{2c}$$

against

$$H_1 : p_{11} \neq p_{21} \text{ or } p_{12} \neq p_{22} \text{ or } \ldots \text{ or } p_{1c} \neq p_{2c}$$

Everything else stays the same.

Chapter 7

Choice of Model and Non-Parametric methods

7.1 Introduction

So far, this reader has assumed the normal distribution in most examples, theories, et cetera. Chapter 1 discussed various data analytic methods to test normality: numerical measures, histograms, and Q-Q plots. This chapter will introduce the Shapiro-Wilk test to decide on normality.

We present the difference between parametric tests (which assume normality, which we can approximate with enough samples and the CLT) and non-parametric tests. Two alternatives are presented:

- An alternative for the t-test on the expectation of a population: the **sign test**.
- An alternative for the t-test on the difference of the μs for two independent samples: **Wilcoxon's sum test**.

7.2 Large samples

If we draw random samples from a continuous, but not normal distribution, the sample mean $\overline{X}$ only has an approximate normal distribution if n is sufficiently large. So far, we used $n \geq 25$ as a rule of thumb. In this chapter we use $n \geq 40$ as a safer rule of thumb.

For a large ($n \geq 40$) sample drawn from a not-normal distribution, we have approximately:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{\text{CLT}}{\sim} \mathcal{N}(0,1)$$

From $P\left(-c < \frac{\overline{X}-\mu}{S/\sqrt{n}} < c\right) \approx 1 - \alpha$, it follows:

$$(100-\alpha)\% - \text{CI}(\mu) = \left(\overline{X} - c\frac{S}{\sqrt{n}}, \overline{X} + c\frac{S}{\sqrt{n}}\right), \text{ with } \Phi(c) = 1 - \frac{1}{2}\alpha$$

We can now conduct a test on $H_0 : \mu = \mu_0$ with test statistic $Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$. Under H_0, Z is approximately $\mathcal{N}(0,1)$-distributed.

We can use the above confidence interval and test for paired samples as well. In this case, the observed differences are considered a one-sample problem.

When comparing μ_1 and μ_2 for two independent samples from not-normal distributions, we can apply the approximate normal distribution if $n_1 \geq 40$ and $n_2 \geq 40$.

We know:

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

By estimating σ_1^2 and σ_2^2 with S_1^2 and S_2^2, respectively, we see:

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

and:

$$(100\% - \alpha) - \mathrm{CI}\,(\mu_1 - \mu_2) = \left(\overline{X} - \overline{Y} - c \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \overline{X} - \overline{Y} + c \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right), \text{ with } \Phi(c) = 1 - \frac{1}{2}\alpha$$

Finally, we can test $H_0 : \mu_1 - \mu_2 = \Delta_0$ with the test statistic

$$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\text{CLT}}{\sim} \mathcal{N}(0,1) \text{ under } H_0$$

7.3 Shapiro-Wilk's test on normality

Shapiro and Wilk designed a test on normality, with test statistic

$$W = \frac{\left(\sum_i a_i X_{(i)}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}$$

Remarks:

- $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics, given the observed sample $X_1, X_2, \ldots, X_n$.
- The numbers $a_1, a_2, \ldots, a_n$ can be found in Shapiro-Wilk's table.
- The denominator represents the "variation" of the sample set. Divide it by $n-1$ and you will find the sample variance. In symbols: $\sum_i \left(X_i - \overline{X}\right)^2 = (n-1)S^2$.
- The coefficients a_i are chosen such that the value of W is close to 1 if the normal distribution applies. We know $W \leq 1$, so a value significantly less than 1 proves that the normal distribution does not apply.

  Consequently, Shapiro-Walk's test is lower-tailed: the rejection region has shape $W \leq c$, where the critical value c is found in Shapiro-Wilk's table for a given sample size n.

7.4 The sign test on the median

The sign test is a hypothesis test. The probability model resembles the amount of observations greater or smaller than, for example, μ_0. Then, under H_0, $X =$ "the number of observations with a certain condition" $\sim \mathcal{B}(n,p)$.

An overview of the one sample tests on the "center" (mean or median) of a distribution:

Model	σ^2, n	**CI for** μ	**Test statistic**	**Find** c **in the**
$\mathcal{N}(\mu, \sigma^2)$	σ^2 known, any n	$\overline{X} \pm c \cdot \frac{\sigma}{\sqrt{n}}$	$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\mathcal{N}(0,1)$-table
	σ^2 unknown, any n	$\overline{X} \pm c \cdot \frac{S}{\sqrt{n}}$	$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	t_{n-1}-table
Not-normal	σ^2 unnown, $n \geq 40$	$\overline{X} \pm c \cdot \frac{S}{\sqrt{n}}$	$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$	$\mathcal{N}(0,1)$-table (approximation!)
	$n < 40$	...	Sign test on the median	$\mathcal{B}\left(n, \frac{1}{2}\right)$-table. When $n \geq 15$: $\mathcal{N}\left(\frac{n}{2}, \frac{n}{4}\right)$.

7.5 Wilcoxon's rank sum test

Wilcoxon's rank sum test allows us to test whether the values of X are structurally greater than the values of Y, without assuming a specific distribution for the populations. It works as follows:

- Order all observations $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ in one sequence from small to large, such that each observations is awarded a rank between 1 and $n_1 + n_2 = N$.
- Add all the ranks $R(X_i)$ of only the x-values: $W = \sum_i R(X_i)$.
- The sum of the ranks W will be (relatively) large, if the x-values are systematically (structurally) higher than the y-values.

  We reject H_0 : "no structural difference" in favor of the alternative H_1 : "the x-values are structurally greater than the y-values" if the sum of ranks of the x-values is large enough:

  Reject H_0 if $W = \sum_i R(X_i) \geq c$.

What is the distribution of W? Well:

Property 10
If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_2$ are independent random samples, drawn from unknown but equal distributions and the ranks are determined in the total sequence of $n_1 + n_2 = N$ observations, then the sum of ranks of the x-values $W = \sum_{i=1}^{n} R(X_i)$ is, for large n_1 and n_2, approximately distributed with $\mu = \mathrm{E}(W) = \frac{1}{2}n_1(N+1)$ and $\sigma^2 = \mathrm{var}(W) = \frac{1}{12}n_1 n_2(N+1)$.

We will use $n_1 > 5$ and $n_2 > 5$ as a rule of thumb for applying this normal approximation.

An overview of the two-sample problems with respect to the difference of two population means:

Model	$n_1, n_2, \sigma_1^2, \sigma_2^2$	**CI for** $\mu_1 - \mu_2$	**Test on** H_0 : $\mu_1 - \mu_2 = \Delta_0$	**Find** c **in the**
$\mathcal{N}(\mu, \sigma^2)$	$\sigma_1^2 = \sigma_2^2$, all n_1, n_2	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$t_{n_1+n_2-2}$-table
	$\sigma_1^2 \neq \sigma_2^2$, $n_1 \geq 40$ and $n_2 \geq 40$	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$\mathcal{N}(0,1)$-table
Not-normal	$n_1 \geq 40$ and $n_2 \geq 40$	$\overline{X} - \overline{Y} \pm c \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$\mathcal{N}(0,1)$-table (approx.)
	$n_1 < 40$ or $n_2 < 40$	Wilcoxon's rank sum test $W = \sum_i R(X_i)$		