The Use of Normal Distribution in Statistics

Juri Strumpflohner (juri.strumpflohner@stud-inf.unibz.it)

Advanced Statistics Project Work

Free University of Bozen-Bolzano
Academic Year 2008/2009

September 25, 2009

Contents

1 Introduction **3**

2 The Normal Distribution **3**
2.1 The Central Limit Theorem . . . 4
2.2 Shape, Parameters and Properties of the Normal Distribution Curve . . . 4
2.3 Experiment - Verifying the CLT . . . 6

3 Standard normal distribution **8**
3.1 Standardizing normal random variables X . . . 9
3.2 Unstandardizing normal random variables Z . . . 11

4 Standard Normal Null Distribution (Z-test) **11**

5 The Normality Assumptions in Parametric Tests **13**
5.1 t-test . . . 13
5.2 F-test . . . 14
5.3 χ^2-test . . . 15

6 Conclusion **16**

7 Appendix **17**

References **19**

List of Figures

1 Bell-shaped normal curve . . . 5
2 Effects of changing μ and σ . . . 6
3 Distribution of the sample . . . 7
4 distribution of the sample as n increases . . . 7
5 Standard normal distribution curve with corresponding areas . . . 8
6 Shaded area between $z = 0.33$ and $z = 1.33$ . . . 9
7 t students (red) vs. normal distribution (blue) . . . 13
8 Possible shape of an F-distribution . . . 15

1 Introduction

When observing and analyzing random phenomena in statistics we talk about random variables, usually written X, which are numerical outcomes of such a random phenomenon.

Definition 1.1. *A **random variable** is a function of an outcome,*

$$X = f(\omega)$$

where the domain of the variable is the sample space Ω and ω is its outcome. In other words, it is a quantity that depends on a chance.

There exist two main kind of random variables:

- discrete random variables
- continuous random variables

Discrete random variables have a finite range of values and are usually counts (i.e. number of children of a family). Moreover, their values can be listed, or arranged in a sequence (i.e. the number of jobs submitted to a printer).

Continuous random variables - to the contrary - can take a whole interval of values, where the interval may be bounded or unbounded. The interval is uncountable and can therefore not be listed in a sequence. Examples of continuous random variables can be physical measures (i.e. weight, height) or times (i.e. software installation time).

The actual outcome ω of X is not known prior to the experiment. Therefore the possible outcomes and the corresponding probabilities are being collected for determining the distribution of X.

Definition 1.2. *Collection of all the probabilities related to X is the distribution of X. The function*

$$P(x) = P\{X = x\}$$

is the probability mass function, or pmf. The cumulative distribution function, or cdf is defined as

$$F(x) = P\{X = x\} = \sum_{y<=x} P(y).$$

We distinguish between the families of discrete distributions (i.e. Bernoulli distribution, Binomial distribution, Poisson distribution,...) and the families of continuous distributions (i.e. Uniform distribution, Exponential distribution, Normal distribution,...).

This report will mainly focus on the family of continuous ones, especially on the Normal distribution.

2 The Normal Distribution

The normal distribution is one of the most commonly observed continuous random variables when dealing with random phenomena. According to the central limit theorem (see section 2.1) many attributes follow this distribution if a sufficiently large number of measures has been collected.

The name “normal” comes from the historical derivation of its use as a model of errors in scientific observations. It first appeared in 1733 in a work of the mathematician Demoivre. Later in 1786 it was also indipendently derived by Laplace. The most well-known association however is with the German mathematician Karl Friedrich Gauss, who derived a new formula for the curve in 1809. Therefore, the normal curve is also refered to as “Gaussian” curve.

Normal distributions play a major role in the field of statistical inference and form the base assumption of many other kind of distributions. Moreover many economic phenomena generate random variables which can be approximated very well with a normal distribution.

2.1 The Central Limit Theorem

The Central Limit Theorem - which is also known as the second fundamental theorem of probability - is defined as follows.

Theorem 2.1 (Central Limit Theorem). *Let $X_1, X_2, \ldots, X_n$ be n indipendent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Moreover let the sum of the n random variables be S_n, given by*

$$S_n = \sum_{i=1}^{n} X_i = X_1 + \ldots + X_n. \tag{1}$$

Let Z_n be the standardized sum defined by

$$Z_n = \frac{S_n - E(S_n)}{Std(S_n)} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{2}$$

then the distribution of Z_n converges to the standard normal distribution $N(0,1)$ as n approaches infinity. Assuming $\Phi(z)$ to be the cumulative distribution function of $N(0,1)$, this can be written as

$$\lim_{n \to \infty} P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z) = \Phi(z) \tag{3}$$

for all z.

Simply speaking, this means that any variable that is the sum of a sufficiently large number (a rule of thumb is $n > 30$) of indipendent factors is likely to be normally distibuted. This is the main reason why the normal distribution is used throughout natural science, statistics etc. as a quite simple model for handling complex phenomena.

2.2 Shape, Parameters and Properties of the Normal Distribution Curve

Figure 1 shows the shape of a normal curve.

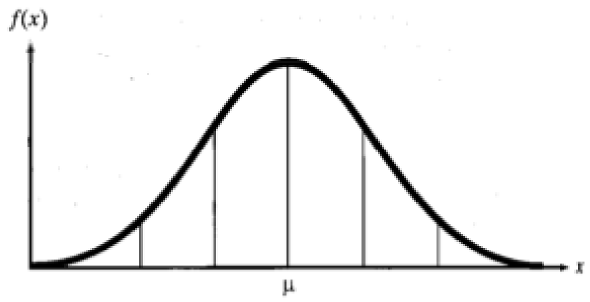


Figure 1: Bell-shaped normal curve

All of the distributions that are part of the family of normal distributions have the following properties:

- the curve given by the probability density function is bell-shaped;
- the curve is perfectly symmetric around the mean μ;
- horizontal axis is approached “asymptotically”;
- median and mode fall together, being equal to the mean μ;
- the inflection points of the curve occur at $\mu - \sigma$ and $\mu + \sigma$, one standard deviation from the mean;
- the total area under the curve is equal to 1;
- almost all the scores (0.997 of them) lie within 3 standard deviations from the mean.

The shape of the normal distribution curve is given by the according probability density function which is defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty \tag{4}$$

where

- μ = Mean of the normal random variable $x = E(X)$
- σ = Standard deviation (strictly positive)
- $\pi = 3.416$
- $e = 2.71828$

As can be seen from formula 4, there are two main parameters which influence the shape of the curve: μ and σ. μ is the mean of the random variable x and is also called the *location parameter*. Changing μ will shift the curve horizontally. σ defines the standard deviation, which measures the variation and dispersion of the probability distribution. It is therefore also called the *scale parameter* and makes the curve more concentrated or more flat.

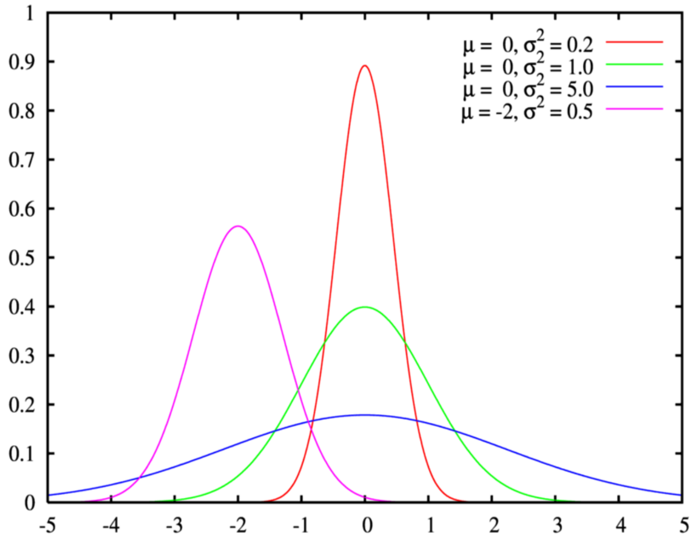


Figure 2: Effects of changing μ and σ

2.3 Experiment - Verifying the CLT

For verifying the properties given by the central limit theorem and the according normal distribution, a small experiment has been conducted, asking a random sample of people for their height and gender. This has been done by submitting a simple questionaire[1] to a random list of people. At the end, just male people have been selected for the experiment since the amount of females was too small. The data used for this experiment can be found in the appendix.

Property	**table value: mean to z**
Sample size (n)	48
Sample mean (μ)	1.799
Sample mode	1.8
Sample median	1.8

Table 1: Properties of the chosen sample for the experiment

From the table it can be immediately seen that the mean, mode and median are nearly equal. This is one of the properties of the normal distribution as already mentioned in section 2.2. Figure 3 shows the distribution of the sample data which can be seen to be approximately normally distributed.

[1]The questionaire can be found at `http://spreadsheets.google.com/viewform?hl=en&pli=1&formkey=dFFWcjVONORTRVpidDJPTFVHMONUWHc6MA.`.

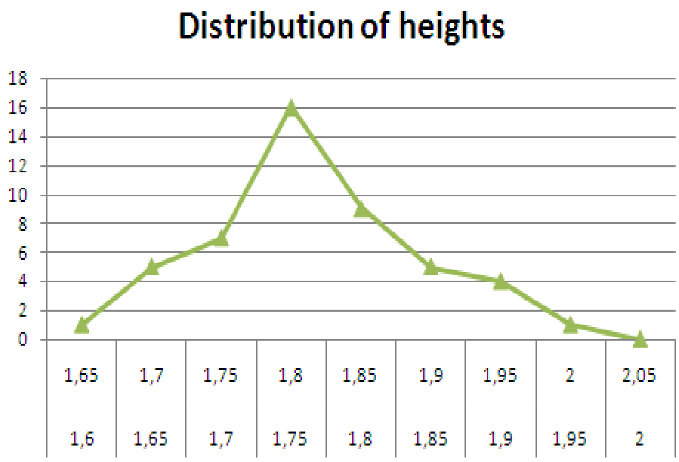
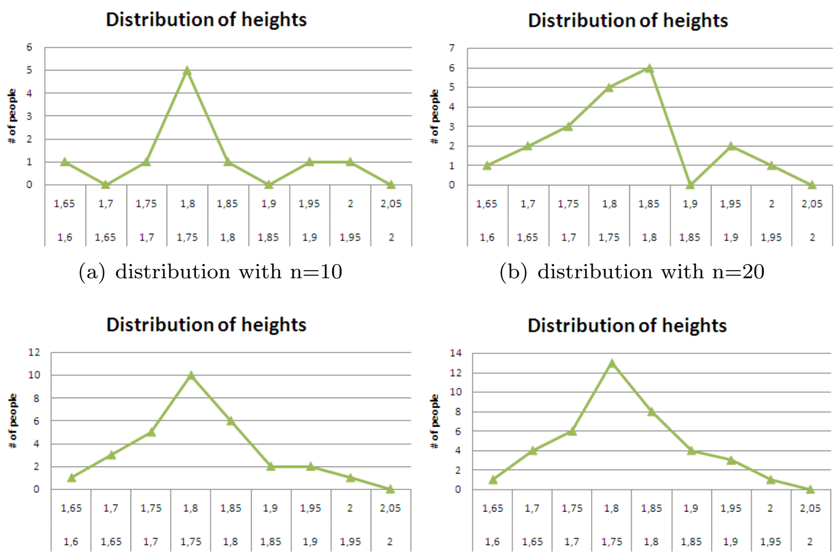


Figure 3: Distribution of the sample

The x-axis in figure 3 is divided into different ranges from a height of 1.60m to a height of 2.05m. The y-axis shows the number of people that fall with the different ranges.

Figure 4 shows how the distribution approximates the normal distribution shape as the sample size n increases. The specific values falling within the different samples has been determined by the time when the data has been entered in the questionaire. In this way it was possible to show the according changes in the shape.



(a) distribution with n=10

(b) distribution with n=20

(c) distribution with n=30

(d) distribution with n=40

Figure 4: distribution of the sample as n increases

3 Standard normal distribution

One of the reasons of the widespread usage of the normal distribution is due to its particularly useful, simple and well-known mathematical properties. Different normal distribution curves vary just in their means and standard deviations. Therefore, by standardizing their parameters, they can be brought to a common appearance and hence treated uniformly.

Definition 3.1. *The* ***standard normal distribution*** *is a normal distribution with $\mu = 0$ and $\sigma = 1$. A random variable with standard normal distribution, denoted by the symbol z, is called a standard normal random variable denoted $Z \sim N(0, 1)$.*

The probability density function of a standard normal distribution is given by the following formula:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \quad (5)$$

The following figure shows the standard normal curve centered at the mean $\mu = 0$ and with a standard deviation $\sigma = 1$.

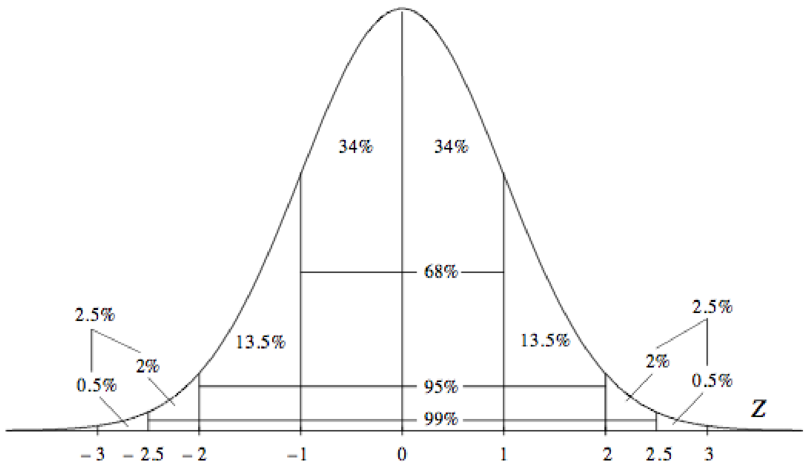


Figure 5: Standard normal distribution curve with corresponding areas

From this figure it results that

- 68% of the z-scores[2] lie within 1 standard deviation from the mean;
- 95% of the z-scores lie within 2 standard deviation from the mean;
- 99.7% of the z-scores lie within 3 standard deviation from the mean.

Example: What is the probability of selecting values with a z-score greater than 1?

[2]We distinguish between the standard random normal variable Z and its values z (scores) represented on the horizontal axis as seen on figure 5

Solution: The answer is quite simple and can be determined by looking at the areas under the curve of figure 5. The according probability of selecting a z-score greater than 1 is $0.135 + 0.025 = 0.16$ or 16%.

For calculating the probabilities of z-scores where $z = 1.23$, $z = 2.45$ etc, that is, the values of z are not whole numbers, appropriate standard normal tables can be used for obtaining the pre-calculated values (usually in the form of the value of the cumulative distribution function) for the corresponding areas. ◇

Example: What is the area between the z-scores of 0.33 and 1.33?

Solution: Figure 6 highlights the area that has to be calculated.

Figure 6: Shaded area between $z = 0.33$ and $z = 1.33$

For calculating the shaded area we first consult the standard normal distribution table for the values from the mean up to $z = 1.33$ and $z = 0.33$. We get the following result:

z	**table value: mean to z**
0.33	0.1293
1.33	0.4082

Table 2: Values retrieved from the z-tables

Consequently, we subtract the smaller area from the larger one for getting the final result of the shaded area: $0.4082 - 0.1293 = 0.2789$ or 27.89%. ◇

3.1 Standardizing normal random variables X

As described in the previous section, the areas under the probability density function of a standard normal distribution are well-documented and can be retrieved by consulting the corresponding standard normal distribution tables. However, most often, the distribution of the random normal variable X is not standardized. So to take advantage of these tables, the raw scores have to be standardized appropriately. To get the corresponding z-score, it has to be calculated how many standard deviations the raw score is from the mean.

Mathematically this can be written as

$$Z = \frac{X - \mu_x}{\sigma_x} \tag{6}$$

This can be proofed as follows. Given the cumulative density function of a normal random variable, denoted by

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(u-\mu)^2}{2\sigma^2}} du, x \in \mathbb{R}, \tag{7}$$

then the cumulative distribution function of a standard normal distribution is given by

$$\Phi_{0,1}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du, x \in \mathbb{R}. \tag{8}$$

Substituting x with the transformation formula (see formula 6) we get

$$\Phi_{0,1}(\frac{X - \mu_x}{\sigma_x}) = \int_{-\infty}^{\frac{X-\mu_x}{\sigma_x}} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du, x \in \mathbb{R}. \tag{9}$$

We can now apply the substitution $t = u\sigma + \mu$. Transforming it, we get $u = \frac{t-\mu}{\sigma}$. If we now apply the substitution we have

$$\Phi_{0,1}(\frac{X - \mu_x}{\sigma_x}) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} \frac{1}{\sigma} dt = \tag{10}$$

$$\Phi_{0,1}(\frac{X - \mu_x}{\sigma_x}) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \tag{11}$$

As can be seen, the final outcome (see formula 11) is exactly the cumulative distribution function of a normal random variable $N(\mu, \sigma^2)$. Hence we get

$$F_{N(0,1)}(\frac{x - \mu}{\sigma}) = \Phi(\frac{x - \mu}{\sigma}) = F_{N(\mu,\sigma^2)}(x). \tag{12}$$

Example: Suppose that the average household income in some country is 900 coins, and the standard deviation is 200 coins. Assuming the Normal distribution of incomes, compute the proportion of "middle class", whose income is between 600 and 1200 coins.

Solution: For a Normal ($\mu = 900$, $\sigma = 1200$) random variable X, first the normalization process has to be done in order to be able to use the standard normal distribution tables. We have

$$P\{600 < X < 1200\} = P\{\frac{600 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{1200 - \mu}{\sigma}\} = \tag{13}$$

$$P\{\frac{600 - 900}{1200} < Z < \frac{1200 - 900}{1200}\} = P\{-1.5 < Z < 1.5\} = \tag{14}$$

$$\Phi(1.5) - \Phi(-1.5) = 0.9332 - 0.0668 = 0.8664. \tag{15}$$

◇

3.2 Unstandardizing normal random variables Z

Similarly to standardizing raw values to standard normal values (as decribed in the previous section 3.1) it may also arise the situation where the probability is given and the goal is to find the corresponding value of x. The according formula can be constructed by solving the previously seen standardization formula (see formula 6) for X:

$$X = \mu_x + \sigma_x Z \tag{16}$$

Example: The government of the country in the previous example decides to issue food stamps to the poorest 3% of households. Below what income will families receive food stamps?

Solution: Using the formula 16 we have

$$x = \mu + \sigma z \tag{17}$$

We have $P\{X < x\} = 3\% = 0.03$. We have to do a lookup on the standard normal distribution tables for a z-score that matches this probability. We get $z = -1.88$. Substituting this in the formula we have

$$x = \mu + \sigma(-1.88) = 900 + 200(-1.88) = 524 \tag{18}$$

Basically families with an income below 524 coins will receive food stamps. ⋄

4 Standard Normal Null Distribution (Z-test)

Another important field in statistical inference is the verification of statements and claims, known as *hypothesis testing.*

Especially related to the normal distribution we have the so-called "Z-test". This kind of test is used when the distribution of the test statistic under the null hypothesis can be approximated with a standard normal distribution. Usually, statistics that are averages of approximately indipendent data values and given that the sample size is large enough, tend to be normally distributed. This is also guaranteed by the central limit theorem mentioned in section 2.1.

There are three different cases that have to be considered when performing a Z-test.

1. Level α test with right-tail alternative

$$\begin{cases} \text{reject } H_0 & \text{if } Z > z_\alpha \\ \text{accept } H_0 & \text{if } Z \le z_\alpha \end{cases}$$

The corresponding rejection and acceptance region is given as follows

$$R = (z_\alpha, +\infty) \text{ and } A = (-\infty, z_\alpha].$$

This means that under the null hypothesis the Z-test - also denoted Z - belongs to the acceptance region A which gets accepted with probability

$$\Phi(z_\alpha) = 1 - \alpha$$

where Φ denotes the standard cumulative distribution function and α is the probability of false rejection (type I error). The corresponding test would be defined as

$$H_0 : \mu = \mu_0$$
$$H_A : \mu > \mu_0$$

2. Level α test with left-tail alternative

$$\begin{cases} \text{reject } H_0 & \text{if } Z < -z_\alpha \\ \text{accept } H_0 & \text{if } Z \geq -z_\alpha \end{cases}$$

   The corresponding rejection and acceptance region is given as follows

$$R = (-\infty, -z_\alpha) \text{ and } A = [-z_\alpha, +\infty).$$

   Similarly as before, Z belongs to A under the null hypothesis with a probability of $\Phi(z_\alpha) = 1 - \alpha$.

   The according test would be defined as

$$H_0 : \mu = \mu_0$$
$$H_A : \mu < \mu_0$$

3. Level α test with two-sided alternative

$$\begin{cases} \text{reject } H_0 & \text{if } |Z| > z_{\alpha/2} \\ \text{accept } H_0 & \text{if } |Z| \leq z_{\alpha/2} \end{cases}$$

   In a two-sided test we have the following rejection and acceptance region

$$R = (-\infty, -z_{\alpha/2}) \bigcup (z_{\alpha/2}, +\infty) \text{ and}$$
$$A = [-z_{\alpha/2}, z_{\alpha/2}].$$

   The probability of a type I error is again equal to α.

   The according test would be defined as

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

The test statistic is given by

$$Z = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{Var(\hat{\theta})}} \tag{19}$$

where θ is the population parameter and $\hat{\theta}$ its estimator. For the test statistic to be applicable, the following conditions have to be met:

- The population estimator $\hat{\theta}$ has to be normally distributed. This is usually justified by the central limit theorem and the collection of indipendent, large samples.
- $E(\hat{\theta})$ and $Var(\hat{\theta})$ are known when the hypothesis is true

5 THE NORMALITY ASSUMPTIONS IN PARAMETRIC TESTS

In statistical inference when comparing measurements, two different kind of families have to be distinguished:

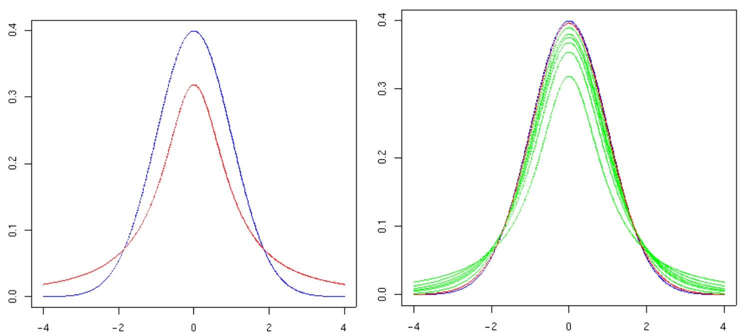
- parametric tests
- non-parametric tests

While the non-parametric tests don't assume any specific kind of distribution, parametric tests are based on some distribution that can be described, i.e. that the population of interest is approximately normally distributed. They are called "parametric" because these assumptions are on the population parameters (i.e. the population mean and variance). Parametric tests are often called to be more *robust* than non-parametric tests when violations of these assumptions occur. This robustness is given by the central limit theorem which guarantees that these tests still work quite reliably with large samples even if the population is not normally distributed.

The following sections will briefly describe some of the most commonly used parametric tests which are related to the normal distribution: the t-test, F-test and χ^2-test.

5.1 t-test

A t-test is a statistical hypothesis test where the test statistic follows a Student's t distribution. The Student's t distribution is suitable for small samples when the population is approximately normal and when the variance is unknown. Figure 7 shows the relationship between the t students and the normal distribution.



(a) t students df=1 (b) t students df=30

Figure 7: t students (red) vs. normal distribution (blue)

For using the t-test, some assumptions have to be made about the underlying population and the chosen sample:

- the population distributions are normal;

- the population distributions have the same variance (homogeneity of the variance);
- normality of the sample means;
- sample variance follows a scaled χ^2 distribution;
- the sample mean and variance are statistically indipendent.

Even though, it has been shown that, the t-test is robust against the violation of these assumptions (i.e. violations of the normality assumption as well as the homogeneity of the variance). Again, the central limit theorem plays a major role in this robustness in that it guarantees that the sample mean of moderately large samples will still be well approximated, even if the underlying data is not normally distributed.

Here are some commonly used examples of the t-tests:

- Indipendent one-sample t-test
  This is used for testing for equality of the mean to a specified value μ_0 (hypothesis testing).
- Indipendent two-sample t-test
  This kind of test is used when the number n of participants of two groups are equal and it can be assumed that they have the same variance. There are variations where either the sample size or the variance doesn't match.
- Dependent t-test for paired samples
  This test is used when there is one sample that is being tested twice. It is often also called "matched-pair t-test" and can take place when there are "before-after" tests.

5.2 F-test

As mentioned in the previous section 5.1, we may use the t-test when having the problem of comparing two population means with small (indipendent) samples. However one of the assumptions for the t-test to be performed optimally is that the homogeneity of the population variances is given. So before applying a t-test on the population, we may therefore first want to verify whether this condition actually holds. This is what the F-test may be used for.

The F-test is a statistical test where the according null hypothesis has an F-distribution. Its main purpose is the analysis of the variance. In more detail it becomes relevant when trying to calculate the ratios of variances of normally distributed statistics. This is the case when doing a comparison of variances which is actually done by comparing the ratio of two variances, s.t. if the ratio is 1, they are equal and the null hypothesis will be accepted.

The F-test is given by the ratio of two indipendent χ^2 variables divided by their respective degrees of freedom

$$F = \frac{\frac{\nu_1 * s_1^2}{\sigma_1^2}/\nu_1}{\frac{\nu_2 * s_2^2}{\sigma_2^2}/\nu_2} = \frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2} \tag{20}$$

where ν_1 and ν_2 denote the degrees of freedom.

To make an inference about the ratio σ_1^2/σ_2^2, sample data will be collected and the sample variances will be used for the test statistic:

$$F = \frac{s_1^2}{s_2^2} \tag{21}$$

In order to be able to establish a rejection region, the sampling distribution of s_1^2/s_2^2 has to be known. This distribution is based on the following assumptions

- the two sampled populations are normally distributed;
- the samples are randomly and indipendently selected from their repective populations.

If these assumptions are met and the null hypothesis H_0 is true, the underlying distribution of F is the F-distribution with $(n_1 - 1)$ numerator degrees of freedom and $(n_2 - 1)$ denominator degrees of freedom respectively.

Figure 8 shows an example of an F-distribution.

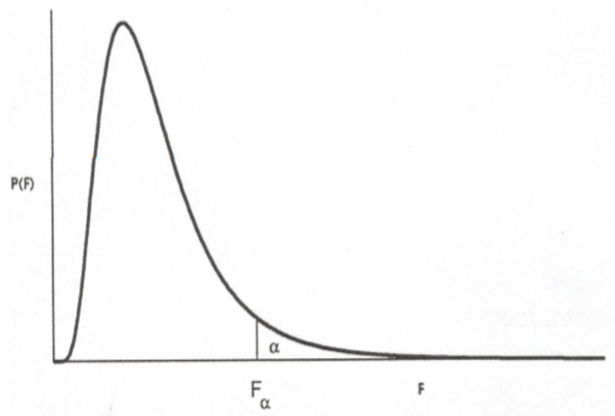


Figure 8: Possible shape of an F-distribution

5.3 χ^2-test

The χ^2-test is a statistical test in which, when the null hypothesis H_0 is true, the underlying sampling distribution is a χ^2 distribution.

The χ^2 distribution has a wide usage in inferential statistics as for instance in statistical significance tests. The most well known usage however is in combination with the χ^2-test for goodness of fit of an observed distribution to a theoretical one. It basically measures the degree of disagreement between the data and the null hypothesis. Mathematically this can be formulated as

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}. \tag{22}$$

Note, the farther the observed value is from its respective expected value, the larger will be the outcome of the above formula, that is, the larger will be χ^2.

The test is therefore relevant for this report because it can be used to test whether some sample data is likely to follow a normal distribution.

6 Conclusion

The normal distribution is one of the most used distribution in statistics but also in other areas such as in natural science and social science as a simple model for handling complex phenomena.

This report mainly described the usage of the normal distribution in statistics. It explained its main properties and parameters, showing the shape of the according distribution function as well as highlighting the relation with the central limit theorem by illustrating it on the basis of a simple experiment. Moreover the main processes of standardizing and unstandardizing of raw scores to and from standard normal values has been shown by providing appropriate examples. Furthermore the normality assumptions in parametric tests have been described by briefly mentioning the most used distributions in parametric tests like the t-test, F-test and the χ^2-test.

7 Appendix

The following table contains the sample data used for the experiment mentioned in section 2.3.

timestamp	**height**	**gender**
8/19/2009 21:29:37	1.8	male
8/19/2009 21:33:56	1.78	male
8/19/2009 22:18:27	1.78	male
8/20/2009 6:48:26	1.93	male
8/20/2009 7:38:43	1.73	female
8/20/2009 7:45:55	1.8	male
8/20/2009 8:43:09	1.75	male
8/20/2009 9:39:02	1.63	female
8/20/2009 10:21:36	1.8	male
8/20/2009 10:26:17	1.6	male
8/20/2009 10:42:35	1.83	male
8/20/2009 23:52:16	2	male
8/21/2009 7:56:23	1.71	male
8/21/2009 17:42:43	1.82	male
8/21/2009 19:36:40	1.68	female
8/22/2009 17:24:55	1.69	male
8/22/2009 22:10:09	1.7	male
8/23/2009 16:53:54	1.82	male
8/24/2009 11:30:22	1.73	male
8/25/2009 9:09:34	1.91	male
8/25/2009 12:45:00	1.82	male
8/25/2009 12:45:38	1.82	female
8/26/2009 15:53:40	1.83	male
8/26/2009 15:53:58	1.85	male
8/26/2009 15:57:21	1.88	male
8/26/2009 15:57:52	1.8	male
8/26/2009 16:01:38	1.88	male
8/26/2009 16:03:23	1.78	male
8/26/2009 16:04:07	1.78	male
8/26/2009 16:07:39	1.73	male
8/26/2009 16:07:42	1.78	male
8/26/2009 16:12:10	1.75	male
8/26/2009 16:18:06	1.7	male
8/26/2009 16:18:35	1.8	male
8/26/2009 16:20:36	1.9	male
8/26/2009 16:23:07	1.75	male
8/26/2009 16:26:08	1.87	male
8/26/2009 17:17:33	1.93	male
8/26/2009 17:24:54	1.83	male
8/26/2009 17:34:00	1.76	male
8/26/2009 17:45:08	1.78	male

8/26/2009 17:45:47	1.81	male
8/26/2009 17:47:47	1.68	male
8/26/2009 18:02:36	1.79	male
8/26/2009 20:10:10	1.93	male
8/26/2009 20:25:06	1.8	male
8/26/2009 20:25:46	1.8	male
8/26/2009 20:32:36	1.75	male
8/26/2009 21:11:20	1.7	male
8/26/2009 21:13:25	1.8	male
8/26/2009 22:39:45	1.82	male
8/28/2009 10:20:40	1.86	male

Table 3: Sample data used in the experiment

References

[1] Chi-square distribution, August 2009. http://en.wikipedia.org/wiki/Chi-square_distribution.

[2] Chi-square goodness of fit test, August 2009. http://www.stat.yale.edu/Courses/1997-98/101/chigf.htm.

[3] F-test, August 2009. http://en.wikipedia.org/wiki/F-test.

[4] Normal distribution, August 2009. http://en.wikipedia.org/wiki/Normal_distribution.

[5] Stats: F-test, August 2009. http://people.richland.edu/james/lecture/m170/ch13-f.html.

[6] Student's t-distribution, August 2009. http://en.wikipedia.org/wiki/Student%27s_t-distribution.

[7] M. Baron. *Probability and statistics for computer scientists.* Chapman & Hall/CRC, 2006.

[8] Sue Gordon. The normal distribution. University of Sydney, 2006.

[9] J.T. McClave, P.G. Benson, and T. Sincich. *A first course in business statistics.* Prentice Hall, 1998.

[10] C. Robson. *Experiment, design and statistics in psychology.* Penguin Books, 1983.