Chapter 1

Robustness and Distribution Assumptions

1.1 Introduction

In statistics, one often works with model assumptions, i.e., one assumes that data follow a certain model. Then one makes use of methodology that is based on the model assumptions.

With the above setup, choosing the methodology can be a quite delicate issue, since the performance of many methods may be very sensitive to whether the model assumption hold or not. For some methods, even very small deviations from the model may result in poor performance.

Methods that perform well, even when there are some (more or less) minor deviation from the model assumption, are called *robust*.

1.2 Distribution Assumptions in Statistics

Let X be a real-valued random variable (r.v.), which is assumed to have a certain specific distribution function $F : \mathbb{R} \to [0, 1]$. Here F is allowed to depend on a *parameter* $\theta \in \mathbb{R}^m$, so that the distribution can be written as

$$\mathbf{P}\{X \leq x\} = F(x; \theta) \qquad \text{for} \quad x \in \mathbb{R}.$$

The parameter θ is assumed to have a certain specific value, which is normally not known.

Example 1.1. *The random variable X is assumed to have normal $N(\mu, \sigma^2)$-distribution for some (unknown) selection of the parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+$.*

The above-mentioned type of scenario, or variants there of, are the frameworks for *parametric statistic methods*. One example, where the method uses a distributional assumption in a crucial manner, is analysis of variance, which assumes normal distribution, and it is not applicable when that assumption is violated.

Observe that, in practice, one can usually not uncritically accept assumptions on the distribution as valid. Hence it is important to be able to determine if the data really comes from an assumed distribution $F(x; \theta)$, for some value of the parameter θ.

Let $X_1, \ldots, X_n$ be a *random sample* of X, i.e., independent random variables with the same distribution as X (which is $F(x;\theta)$ if the assumption on the distribution holds). For the above mentioned reasons, it is often of importance to determine whether the distribution of X really is $F(x;\theta)$. This cannot be done in a completely precise manner, as we have randomness.

In fact, to test the distribution assumption $F(x;\theta)$, one has to use some statistical test, which hopefully, with a large probability of being correct, can tell whether the data obeys the assumption.

1.2.1 Graphical Test of the Distribution Assumptions

We start with stating some facts that will be of importance to us:

Let X be a random variable that has a continuous distribution function F. Then the random variable $F(X)$ has a uniform distribution over $[0,1]$. To see this, just notice that

$$\mathbf{P}\{F(X) < x\} = \mathbf{P}\{X < F^{-1}(x)\} = F(F^{-1}(x)) = x \qquad \text{for} \quad x \in [0,1]^1.$$

The above fact is very useful, because it says that if we have a sample of random variables, and we perform a transformation so that they become uniformly distributed over $[0,1]$, then the transformation should (more or less) be the distribution function!

Now, as a consequence of the *Glivenko-Cantelli theorem*[2], we have the following theorem:

Theorem 1.1. *If the sample $X_1, \ldots, X_n$ has continuous distribution function $F(x;\theta)$, then for the ordered sample $X_{(1)} \leq \ldots \leq X_{(n)}$ and a fixed $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbf{P}\{\max_{1\leq i\leq n} \left|(i-0.5)/n - F(X_{(i)};\theta)\right| \leq \epsilon\} \to 1.$$

Now, if the assumption that the sample $X_1, \ldots, X_n$ has the distribution function $F(x;\theta)$ is correct, then, according to Theorem 1.1,

$$\max_{1\leq i\leq n} \left|(i-0.5)/n - F(X_{(i)};\theta)\right| \approx 0 \qquad \text{for large} \quad n.$$

Consequently, a plot of the sequence of pairs

$$\left\{\left((i-0.5)/n, F(X_{(i)};\theta)\right)\right\}_{i=1}^{n},$$

a so-called *pp-plot*, is approximately a 45°-line. The same is then true for a so-called *qq-plot* of the sequence

$$\left\{\left(X_{(i)}, F^{-1}((i-0.5)/n;\theta)\right)\right\}_{i=1}^{n}.$$

A systematic discrepancy of a *pp*-plot or *qq*-plot from a 45°-line indicates that the $F(x;\theta)$-assumption is not true. Notice that, because of randomness, these plots never become completely straight-lined, for a finite sample size n, even when the $F(x;\theta)$-assumption holds, but always display a certain random variation around the 45°-line. The larger n, the smaller that random variation becomes.

[1] Here F^{-1} is a generalized inverse, if F is non-invertible.

[2] Glivenko-Cantelli says that the empirical distribution of a sample of a random variable converges uniformly to the distribution of the random variable, as sample size tends to infinity.

When the $F(x;\theta)$-assumption is false, an additional systematic discrepancy from the 45°-line occurs, resulting in an (in some sense) curved plot.

Normally, the value of the parameter θ is not known, and hence must be estimated by an estimator $\hat{\theta}$. Supposing that $F(x;\theta)$ is a continuous function of θ, and that the estimator $\hat{\theta}$ is *consistent*, i.e., that it converges to θ in probability when $n \to \infty$, the following *pp*- and *qq*-plots would be approximate 45°-lines

$$\left\{\left((i-0.5)/n, F(X_{(i)};\hat{\theta})\right)\right\}_{i=1}^{n} \qquad \text{and} \qquad \left\{\left(X_{(i)}, F^{-1}((i-0.5)/n;\hat{\theta})\right)\right\}_{i=1}^{n},$$

when the $F(x;\theta)$-assumption holds.

The decision whether a *pp*- or *qq*-plot displays systematic discrepancy, or only random variation discrepancy, from a 45°-line, is conveniently done by means of a comparison with a *reference plot*, without systematic discrepancy. This in turn, can be done by generating a sample $Y_1, \ldots, Y_n$ from a random variable Y that really has the distribution function $F(y;\theta)$, or $F(y;\hat{\theta})$ if θ is unknown and estimated, so that the *pp*-plot

$$\left\{\left((i-0.5)/n, F(Y_{(i)};\theta)\right)\right\}_{i=1}^{n} \qquad \text{or} \qquad \left\{\left((i-0.5)/n, F(Y_{(i)};\hat{\theta})\right)\right\}_{i=1}^{n},$$

and the *qq*-plot

$$\left\{\left(Y_{(i)}, F^{-1}((i-0.5)/n;\theta)\right)\right\}_{i=1}^{n} \qquad \text{or} \qquad \left\{\left(Y_{(i)}, F^{-1}((i-0.5)/n;\hat{\theta})\right)\right\}_{i=1}^{n}$$

display only random variation discrepancy from a 45°-line.

Of course, systematic variations from a 45°-line can be hidden by large random variations, when the sample size n is small in relation to the systematic variation. A non-significant *pp*- or *qq*-plot, without clear systematic variations from a 45°-line, does not necessarily imply that the $F(x;\theta)$-assumption is true[3]. However, one can conclude that the random variation of the data material is similar to the variations of true $F(x;\theta)$-data. This in turn, hopefully, should be enough for making practical use of the $F(x;\theta)$ assumption, at least with robust methodology.

1.2.2 Statistical Test of Distribution Assumptions

In the previous section we described how to test a distribution assumption qualitatively, by a graphical procedure. However, to do it quantitatively, we have to employ a test, which produces a statistic, and hence gives us a p-value, that may be significant or non-significant, in turn. Be aware that, like any test, we will never be able to state that the null hypothesis (in this case that the data follows a certain distribution, like Normal) is correct. What we get out of the test is a statement of the type "data may be Normal" or "the probability that data comes from a Normal distribution is small".

Chi-Square Goodness-of-Fit Test

With the *chi-square test*, given an assumed continuous or discrete distribution $F(x;\theta)$ for a sample, one can assign probabilities that a random variable has a value within an interval, or a so called

[3]In the same fashion, a non-significant outcome of a statistical hypothesis test does not necessarily imply that the null hypothesis is true.

bin. Note that the choice of how to set the bins matters, the actual value of the chi-square test statistic will depend on how the data is binned.

The *chi-square test statistic* of k bins is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed frequency for bin i, i.e., the number of observations that lies in the bin $(l_i, u_i]$, and

$$E_i = n(F(u_i; \hat{\theta}) - F(l_i; \hat{\theta}))$$

is the expected frequency of bin i, with F denoting the assumed distribution function. Here n is the sample size, as before, and $l_1 < u_1 \leq l_2 < u_2 \leq \ldots \leq l_k < u_k$, with k being the number of bins.

Under the null hypothesis, that the F-assumption is true, the test statistic χ^2 follows, approximately, a chi-square distribution, with $k - 1 - c$ degrees of freedom, where c is the number of estimated parameters.

One disadvantage with the chi-square test, is that it is an *asymptotic test* (rather than exact one), i.e., it requires a large enough sample size in order for the chi-square approximation to be valid. As a rule of thumb, each bin should contain at least 5 observations from the sample.

Kolmogorov-Smirnov Goodness-of-Fit Test

The *Kolmogorov-Smirnov test* is applicable when assuming a continuous distribution F for the sample $X_1, \ldots, X_n$. The test statistic is given by[4]

$$D = \max_{1 \leq i \leq n} \left| F(X_{(i)}; \theta) - \frac{i}{n} \right|,$$

were $X_{(1)} \leq \ldots \leq X_{(n)}$ is the ordered sample.

Observe that the Kolmogorov-Smirnov statistic D is a measure of how much a *pp*-plot deviates from a 45°-line.

To calculate the p-value for D, one makes use of the fact that $\sqrt{n}D$ is asymptotically *Kolmogorov distributed*, under the null hypothesis. The distribution function of the Kolmogorov distribution is given by

$$Q(x) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} \mathrm{e}^{-2i^2x^2}.$$

In practice, there is seldom any need to do manual computations with the Kolmogorov distribution, as the computations are handled by statistical programs. Note that when the parameters are unknown, and therefore need to be estimated, $\sqrt{n}D$ is no longer Kolmogorov distributed and we need other means of calculating a p-value for the test[5].

[4]We have previously used $(i - 0.5)/n$ instead of i/n: This choice is really a matter of taste, and of no practical importance.

[5]This is implemented in different ways in different softwares. For example, in the case of testing whether the data comes from a normal distribution (with unknown mean and variance), the modified test procedure is called the Lilliefors test and R has a specific function, called lillie.test(), for this.

Some extension of the Kolmogorov-Smirnov test has been made, to emphasize certain regions of values. One important example, is the *Kuiper test*, with test statistic

$$K = \max_{1 \le i \le n} \left(F(X_{(i)}; \theta) - \frac{i}{n} \right) + \max_{1 \le i \le n} \left(\frac{i}{n} - F(X_{(i)}; \theta) \right).$$

A Kuiper test emphasizes the importance of the *tails*, i.e., the smallest and largest observations. This is of importance in applications to assessment of risk.

1.3 Parameter Estimation

1.3.1 Maximum Likelihood Estimation

Let $x_1, \ldots, x_n$ be a random sample from a r.v. X (assumed) having density $f_X(x; \theta)$. The *likelihood function* is defined as

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f_X(x_i; \theta).$$

Note that it is a function of the parameter θ, the values $x_1, \ldots, x_n$ come from our observations! The maximum likelihood (ML) estimator of θ is

$$\hat{\theta} = \arg \max_{\theta \in \mathbb{R}^m} L(\theta; x_1, \ldots, x_n)$$

It is often easier to regard the logarithm of the likelihood function, i.e. $l(\theta, x_1, \ldots, x_n) = \log L(\theta, x_1, \ldots, x_n)$ and maximize this instead.

Example 1.2. *Let $x_1, \ldots, x_n$ be a random sample from a r.v. which is Exp(λ)-distributed, i.e. $f_X(x; \lambda) = \lambda e^{-\lambda x}$. This means that the likelihood function becomes*

$$L(\lambda; x_1, \ldots, x_n) = \lambda^n \prod_{i=1}^{n} e^{-\lambda x_i}.$$

In this case it is analytically more tractable to regard the log-likelihood function, i.e.

$$l(\lambda; x_1, \ldots, x_n) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

The ML-estimator is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}.$$

Even if in simple cases such as above the optimal value of the likelihood can, with minimal effort, be found analytically, that is not always the case. It is more common that the likelihood equation or equations (in case of several parameter being estimated simultaneously, like in regression) do not have an analytical solution. Numerical optimization has then to be applied.

1.3.2 Robust Estimation

One of the most natural illustrations of *robust estimation* techniques, is the estimation of a location parameter, of a continuous symmetric distribution: Assume that we have a sample $X_1, \ldots, X_n$ from a distribution function of the form $F(x, \theta) = F(x - \theta)$, where $\theta \in \mathbb{R}$. Here θ is called a *location parameter*.

Example 1.3. *If F is a normal distribution, then θ coincides with the expected value and the median. Further, the* sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a good estimator of θ.

Example 1.4. *If F is Cauchy distributed, with probability density function*

$$f_X(x; \theta) = \frac{1}{\pi(1 + (x - \theta)^2)},$$

then the sample mean is not a good estimator of the location parameter θ. The reason for this is that the Cauchy distribution allows, with a large probability very large values, and does in fact not even have a well-defined (finite) expected value. This means that the sample may contain some extremely small or large "non-typical" values, so called outliers*, and a few such may heavily influence the sample mean, so that it deviates significantly from θ.*

To avoid the problem indicated in Example 1.4, one may, for example, replace the sample mean with the *sample median*, as estimator of the location parameter: The latter does not display the sensitivity to outliers, as does the former. In other words, the sample median is a *robust estimator*.

Robustness can, of course, be defined in mathematical terms. That description usually is based on the *influence function*, which measures the sensitivity to outliers. However, this subject matter goes beyond the scope of this course.

An intuitive way, to view the issue of the robustness of an estimator, is to look at the *breakdown point*. This is the largest percentage of data points that can be changed arbitrarily, without causing undue influence on the estimator.

Example 1.5. *The sample median has* 50% *breakdown. This is illustrated by the fact that, for a sample of* 100 *ordered data, the first* 49 *can be changed arbitrarily, as long as their values stay smaller than the* 50*:th smallest observation, without affecting the value of the sample median at all.*

The sample mean is not a robust estimator, because changing the value of a single observation may heavily influence the value of the sample mean. This means that the sample mean has breakdown point 0%*.*

In practice, the choice of estimator is a trade-off between robustness and *efficiency*, as very robust estimators tends to be inefficient, i.e., they do not make full use of the data material.

1.4 Softwares - some hints on useful commands/routines

General hint for all program languages we are going to use in the course: before rushing into raw-coding a formula, make sure there isn't a pre-programmed function that does exactly what you need.

Matlab: Help → Product help → Search. Surprisingly effective. Also, Matlab is good at using vectors for calculations, meaning that if you can re-write your for-loop as a vector operation, do it.

R: rnorm, qnorm, pnorm, matrix, apply, postscript, dev.off() (the last two are for saving plots). The syntax is similar to Matlab, but it uses different symbols, so look up for the brackets. R is also good at using vectors for calculations, and using **apply** is a nice way to apply a function to the rows or the columns of a matrix.

We recommend you to use RStudio, an IDE (Integrated Development Environment) for R. If you do, you have access to help pages within the environment and can easily search for commands and concepts. Also, you can save the plots by clicking Export.

However, if you choose to run R from the terminal, you can find out what a command does through e.g. **?rnorm** (and leave help by pressing **q**). Alternatively, google it.

1.5 Computer assignment

The aim of this lab is not so much to teach complex statistical methods as to give an introduction to some of the different software used in the course. In Assignment 1 you are required to use different methods to check if a distribution assumption is valid and to explore how the coverage of a confidence interval for the variance estimate may be influenced by a faulty distribution assumption, while in Assignment 2 you will compare the behavior of location estimators; the sample mean, median and trimmed mean.

You should first do the lab using Matlab. Then you are required to repeat parts of the lab using R. The parts to repeat using R are Assignment 1.1 (but only part 2 with data from gamma distribution), Assignment 1.2 (but only part 2 with data from gamma distribution) and Assignment 2.3 (but only for data from the contaminated distribution).

To pass the exercise you have to present the answers to the questions marked in bold to the exercise teacher. No full report is needed on this exercise.

There are several questions that ask you to perform tests or construct confidence intervals. For tests, we want you to formally set up the null hypothesis and its alternative, and then give the value of the test statistic and/or p-value and explain what your obtained test statistic and/or p-value means in terms of accepting/rejecting the null hypothesis, and what conclusions you can draw from this acceptance/rejection. For confidence intervals, we want you to give a definition of it and explain what obtaining a certain CI tells you about the distribution parameter in question.

1.5.1 Assignment 1, Effect of Distribution Assumptions

Assignment 1.1

The task in Assignment 1.1 is to review a specific distribution assumption. In both cases below, let us say you do not know the true distribution for the data, but you think that it is the normal distribution (i.e. your distribution assumption for the data is "normal distribution").

- Generate a sample with $n = 100$ normally distributed random variables with parameters $\mu = 0$ and $\sigma = 1$. Get an overview of your data by doing a histogram of the data. Do a pp- or qq-plot of the data. Also perform a formal goodness-of-fit test (Chi-square or Kolmogorov-Smirnov, or some other relevant test). Interpret the results. *Should be done in Matlab*

- Now, generate $n = 100$ observations from a decidedly non-normal distribution, namely $Gamma(a, b)$ with $a = 2$ (shape parameter), $b = 2$ (scale parameter). We still pretend that we do not know the true distribution and that we think that it is the normal distribution. As before, make a histogram of the data to get an overview, and then do a pp- or qq-plot and perform a formal goodness-of-fit test. Interpret the results. *Should be done in Matlab and R.*

Comment: There are other probability plots than pp- and qq-plots. A popular one is the so called "normal probability plot" which the "normplot" function in Matlab does. It is similar to pp/qq and serves the same purpose. You may use it, but in this case you will have to explain what, exactly, is plotted and why the "straight line = normal distribution" argument is valid even here.

Comment for the implementation in R: The function chisq.test() might seem a bit user-unfriendly for doing a goodness-of-fit test of normality. A tip is to read the footnote on page 4 in this document and check out the functions in the R package "nortest".

Question 1.1. Are the simulated data sets normally distributed? Show histograms, pp- or qq-plots and state the null hypothesis for the goodness of fit test. Can the null hypothesis be rejected?

Assignment 1.2

Let $X_1, \ldots, X_n$ be is a random sample from a distribution with mean μ and standard deviation σ. An estimator for the variance σ^2, when μ is not known, is the sample variance

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

where $\bar{X}$ is the sample mean. Assuming that $X_i \sim N(\mu, \sigma)$, we can construct the test statistic

$$\frac{(n-1)S_X^2}{\sigma^2}$$

which is going to follow a chi-square distribution with $n - 1$ degrees of freedom. This statistic can then be used to create a confidence interval for σ^2 (if in doubt, consult any coursebook on basic statistics).

Here, you are going to examine what happens to the coverage of the confidence interval if the normal distribution assumption is fulfilled or not.

- Again, generate a sample with $n = 100$ normal distributed random variables with parameters $\mu = 0$ and $\sigma = 1$. Calculate a confidence interval for σ^2, with confidence level 0.95. Repeat 1000 times and count the number of intervals which contain the true σ^2. Also, calculate the average width of the intervals. How do the results compare with what you expect from the confidence level being 0.95? *Should be done in Matlab.*

- Repeat the task above, but instead using $n = 100$ observations from a $Gamma(a, b)$ distribution with a=2, b=2 (assume that you think the data come from a normal distribution). What percentage of the CI:s cover the true σ^2? What is the average width of the intervals? Remember that the variance of a gamma distribution is ab^2. How do the results compare with what you expect from the confidence level being 0.95? *Should be done in Matlab and R.*

Question 1.2. How many of your 1000 simulated confidence intervals contain the true σ^2 in each case? What would you expect from the confidence level being 0.95?

1.5.2 Assignment 2, Robust Estimation

An ϵ-*contaminated* normal distribution may be defined as

$$X = WY + (1 - W)Z = \begin{cases} Y & \text{with probability} \quad 1 - \epsilon \\ Z & \text{with probability} \quad \epsilon \end{cases} .$$

Here $Y \sim \mathrm{N}(\mu = 0, \sigma_Y^2)$, $Z \sim F$, and $W \sim \text{Bernoulli}(1 - \epsilon)$, with $\epsilon \in (0, 1)$ being a small number. Further, F is other distribution than the $\mathrm{N}(0, \sigma_Y^2)$ distribution that usually displays much wilder fluctuation (i.e., more extreme values).

This contaminated distribution can be viewed as that some phenomena usually is observed, at the rate of $1 - \epsilon$, but that some other phenomena is observed at the rate ϵ. In practice, this can be caused by somebody occasionally making a faulty measurement, or a sloppy computer registration of a result.

As the contaminated distribution is not normal, it can be difficult to analyze. In addition, when using a model of this kind, one is usually interested in the non-contaminated random variable Y, rather than the contaminated variable X.

One common way to handle the contaminated data, is to *remove outliers*. Notice that this is correct, if one is only interested in Y, but might be erroneous if really interested in the contaminated distribution of X.

Assignment 2.1

- Sample 100 observations from such a contaminated distribution, where $\sigma_Y = 1$, F Cauchy distributed with location parameter 0 and $\epsilon = 0.05$. Do a histogram of the data, a *pp*- or *qq*-plot of the data and a goodness-of-fit test. Repeat the simulation a couple of times. Interpret and comment on the results. Can you see outliers?

Hint: For simulation of the Cauchy distribution, one can make use of the fact that a Cauchy distribution with location parameter 0 is the same thing as a Student-t distribution, with 1 degree of freedom.

Assignment 2.2

- Sample 100 observations from the distribution in the first part and estimate the mean. Repeat 1000 times. Order the means and register the value of results number 25 and 975.[6] Also, make a histogram of the means. Do the same thing for a non-contaminated normal distribution. Compare the results. *Should be done in Matlab.*

Assignment 2.3

- Repeat the tasks in Assignment 2.2, but this time replacing the sample mean with the robust estimators, made up of the sample median, and of the α-*trimmed mean*

$$\bar{X}_\alpha = \frac{X_{(k+1)} + \ldots + X_{(n-k)}}{n - 2k} \qquad \text{with} \qquad \alpha = \frac{k}{n},$$

  respectively, where $X_{(\cdot)}$ is the ordered sample. Choose $\alpha = 0.1$. Again, order the results and register the values of results number 25 and 975. Create histograms both for the medians and the trimmed means. Comments and conclusions? *Should be done in Matlab and R (but you do not have to compare with a non-contaminated distribution when using R).*

Question: Show a histogram for 100 observations of the contaminated distribution (one simulation). Show a histogram of the means of 1000 simulations. Show the 25th and 975th value of the mean, median and alpha-trimmed mean. Which of these estimators are more robust to outliers?

[6] The range between these two values does in fact make up a *bootstrap* confidence interval for the expected value of X, with confidence level $\alpha = 0.95$: We will return to this later in the course.