# 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 $\theta$ 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 $F(\cdot; \theta)$, for some value of the parameter $\theta$.

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 $F(\cdot;\theta)$ 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 $F(\cdot;\theta)$. This cannot be done in a completely precise manner, as we have randomness.

In fact, to test the distribution assumption $F(\cdot;\theta)$, 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 have 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 direct consequence of the *Glivenko-Cantelli theorem*[2], we have the following theorem:

**Theorem 1.1.** *If the sample $X_1, \ldots, X_n$ has distribution function $F(\cdot;\theta)$, then for the ordered sample $X_{(1)} \leq \ldots \leq X_{(n)}$, we have*

$$\lim_{n\to\infty} \max_{1\leq i\leq n} \left|(i-0.5)/n - F(X_{(i)};\theta)\right| = 0.$$

Now, if the assumption that the sample $X_1, \ldots, X_n$ has the distribution function $F(\cdot;\theta)$ 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 $F(\cdot;\theta)$-assumption is not true. Notice that, because of randomness, these plots never

[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.

become completely straight-lined, for a finite sample size $n$, even when the $F(\cdot;\theta)$-assumption holds, but always display a certain random variation around the 45°-line. The larger $n$, the smaller that random variation becomes.

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

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

when the $F(\cdot;\theta)$-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 $F(\cdot;\theta)$, or $F(\cdot;\hat{\theta})$ if $\theta$ is unknown and estimated, so that the *pp*-plot

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

and the *qq*-plot

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

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 $F(\theta;\cdot)$-assumption is true[3]. However, one can conclude that the random variation of the data material is similar to the variations of true $F(\theta;\cdot)$-data. This in turn, hopefully, should be enough for making practical use of the $F(\theta;\cdot)$ 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".

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

### Chi-Square Goodness-of-Fit Test

With the *chi-square test*, given an assumed continuous or discrete distribution $F(\cdot;\theta)$ for a sample, one can assign probabilities that a random variable has a value within an interval, or a so called *bin*. Quite obviously, the actual value of the chi-square test statistic will dependent on how the data is binned.

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.

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 $l_1 < u_1 < l_2 < u_2 \leq \ldots \leq l_k < u_k$, with $k$ being the number of bins.

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

### 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]

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

were $X_{(1)} \leq \ldots \leq X_{(n)}$ is the ordered sample. As before, all unknown parameters for $F$ have to be estimated.

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.

[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.

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

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

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 $\theta$, the values $x_1, \ldots, x_n$ come from our observations! The maximum likelihood (ML) estimator of $\theta$ 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($\lambda$)-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. See Project 2 for numerical optimization in Matlab and Project 3 for optimization in R.

### 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 $\theta$ is called a *location parameter.*

**Example 1.3.** *If $F$ is a normal distribution, then $\theta$ 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 $\theta$.*

**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 $\theta$. 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 $\theta$.*

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.

It should be noted that robustness is related to the concept of *non-parametric statis-*

*tics*, i.e., statistical methodology that do not rely on distribution assumptions.

## 1.4 Software

In this, first, lab all program packages that show up in the course are going to be used. Matlab, R, Mathematica, C. General tip for all of them: before rushing into raw-coding a formula, make sure there isn't a pre-programmed function that does exactly what you need.

Useful commands/routines.

**Matlab**. Help → Product help → Search

**R**. **rnorm, qnorm, pnorm, postscript, dev.off()**

**Mathematica**. Help → Documentation Center. **ConstantArray, RandomInteger, RandomReal, NormalDistribution, CDF**

**C**. Feel free to use pre-programmed libraries like **gsl**.

## 1.5 Computer assignment

You will, hopefully, find the questions themselves to be very simple. In Task 1 you are required to calculate a common confidence interval for the variance estimate, while in Task 2 you will compare the behaviour of the mean and the median. The aim of this lab is not so much to teach complex statistical methods as to give a introduction to the different software used in the course.

The two software packages that we are going to concentrate on the most are Matlab and R. There will also be one lab in C and one in Mathematica. In this, first, lab you are required to do the two tasks using at least the first two of those. That is, each task is to be done at least twice, once in Matlab and once in R. If you use the other software packages besides those two, you will get 2 extra points for each task.

1. **Effect of Distribution Assumptions.**

   Let $X_1, \ldots, X_n$ be is a random sample from a distribution with mean $\mu$ and standard deviation $\sigma$. An estimator for the variance $\sigma^2$, when $\mu$ 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 $\sigma^2$ (if in doubt, consult any coursebook on basic statistics).

   You are going to examine what happens to the coverage of the confidence interval if the Normal distribution assumption is not met.

- Generate $n = 100$ normal distributed random variables with parameters $\mu = 0$ and $\sigma = 1$. Do a *pp* (or *qq*) plot of the data (skip the plot in C). Also perform a formal goodness-of-fit test (Chi-square or Kolmogorov-Smirnov). Comments?
- Again, generate $n = 100$ normal distributed random variables with parameters $\mu = 0$ and $\sigma = 1$. Calculate a confidence interval for $\sigma^2$, with confidence level 0.95. Repeat 1000 times and count the number of intervals which contain the true $\sigma^2$. Also, calculate the average width of the intervals. How does this compare with what you expect from the confidence level being 0.95?
- Now, generate $n = 100$ observations from a decidedly non-normal distribution, namely $Gamma(a, b)$ with $a = 10, b = 1$. Let us say that you do not know the true distribution and think that it is Normal. As before, do a plot and perform a formal goodness-of-fit test. Comments? Try different parameters. Can you always detect by looking at the plots/tests that the data actually does not come from a Normal distribution?
- Repeat the data generation process 1000 times and construct the confidence intervals. What percentage of the CI cover the true $\sigma^2$? What the average width of the intervals? Remember that the variance of a *Gamma* distribution is $ab^2$. How does the result compare with what you expect from the confidence level being 0.95?
- There exists a myth that says that everything fixes itself as long as the sample size is large enough. Repeat the previous assignment for other, larger, $n$. Comments?

2. **Robust Estimation**

An $\epsilon$-*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 $Y \sim \mathrm{N}(0, \sigma_Y^2)$, $Z \sim F$, and $W \sim \mathrm{Bernoulli}(1 - \epsilon)$, 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 what is observed, at the rate of $1 - \epsilon$, but that some other phenomena is observed, at the rate $\epsilon$. 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 $X$.

- 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 *pp*-plot of the data (skip the plot in C) and a goodness-of-fit test. Can you see the outliers?

Tip: 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.

- Sample 100 observations from the distribution and estimate the mean. Repeat 1000 times. Order the results and register the value of results number 25 and 975. [5] Make a histogram of the result.
- Repeat the above tasks, 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 and $X_{(\cdot)}$ again is the ordered sample. Choose $\alpha = 0.1$. Again, order the results and register the value of results number 25 and 975. Do a histogram. Conclusions?

To get a pass on the lab you have to complete Task 1 and 2 in both Matlab and R.

[5] 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.