Skriptum zur Vorlesung "Mathematical Statistics"

Karl Grill[1]

Version 1.2
November 4, 1998

Contents

Chapter 1

Introduction

Smoking is the principal cause of statistics
Irish Graffiti

1.1 Basic Notions

The field of statistics is usually divided into two parts: first, descriptive statistics which deals with the gathering and representation of large sets of data, and mathematical or inductive statistics whose object is the determination of properties of a large population based on what is called a random sample. The latter is the object of the present lecture.
Usually, statistical questions arise in conjunction with a (presumably large) population of individuals (these need not necessarily be human or even animate — the collection of all tv sets produced in a factory within a certain period of time makes a perfectly sound statistical population). These individuals have one or more properties which are going to be studied - e.g., length, height, weight, sex, color of eyes or hair,... These data are usually either numerical values (measurements) or some sort of classification (like sex, colours). We can represent the latter by, e.g., integer numbers starting from one, so there is no obvious loss of generality if we assume that our data are real numbers (actually, there is the possibility to allow more general spaces than the real numbers, e.g., some general metric space, but we won't go into that).
The reasoning behind all of mathematical statistics is the observation that if we select an individual from our population randomly, then the value of a certain datum associated with that individual is a random variable whose distribution is the (relative) frequency distribution of that datum

within the population. Furthermore, if we draw a sample of n individuals from the population *with replacement* then the values of the datum will be independent random variables with the same distribution. If sampling is done without replacement, there is some dependency between those random variables, some of which will be explored in the problems below. If the size of the population is large with respect to the sample size, however, the effect of non-replacement is small. Thus, from now on, we will assume that sampling is done with replacement. Furthermore, although real populations will always generate discrete distributions, we will allow any distribution for our sample random variables (which should make sense at least as a kind of approximation). This gives us

Definition 1 *A random sample of size n from distribution P is a sequence $X_1, \ldots, X_n$ of independent, identically distributed random variables with common distribution P.*

The random variables in this definition may be d-dimensional (e.g., if we study height and weight and any possible relation between them).
The job of statistics now is to make some statement about the distribution P based on the sample $X_1, \ldots, X_n$. The following theorem shows that we can expect this to work:

Theorem 1 (Glivenko-Cantelli) *Let $X_1, \ldots$ be a sequence of i.i.d.r.v. with common distribution function F and define the* <u>*empirical distribution function*</u> *as*

$$F_n(x) = \frac{\#\{i \le n : X_i \le x\}}{n}.$$

Then, with probability one as $n \to \infty$,

$$\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \to 0.$$

<u>Proof.</u> Fix $\epsilon > 0$. Choose $M > 1/\epsilon$, $x_0 = -\infty$, $x_M = \infty$, and for $k = 1, \ldots, M-1$ choose x_k in such a way that $F(x_k - 0) \le \frac{k}{M} \le F(x_k)$ and $x_k \ge x_{k-1}$. Then, there is a (random) n_0 such that for $n > n_0$ and any $k = 1, \ldots, M-1$

$$|F_n(x_k) - F(x_k)| < \epsilon$$

and

$$|F_n(x_k - 0) - F(x_k - 0)| < \epsilon.$$

by the strong law of large numbers. For any $x \in \mathbb{R}$ we can find a k with $1 \le k \le M$ and $x_{k-1} \le x < x_k$. With this k, we have, for $n > n_0$

$$F_n(x) \ge F_n(x_{k-1}) \ge F(x_{k-1}) - \epsilon \ge F(x_k - 0) - \epsilon - \frac{1}{M} \ge$$

$$F(x_k - 0) - 2\epsilon \geq F(x) - 2\epsilon.$$

Similarly

$$F_n(x) \leq F_n(x_k - 0) \leq F(x_k - 0) + \epsilon \leq F(x_{k-1}) + \epsilon + \frac{1}{M} \leq$$

$$F(x_{k-1}) + 2\epsilon \leq F(x) + 2\epsilon.$$

The last two inequalities, together with the arbitrariness of ϵ prove the theorem.
In many cases we have some information about the shape of the possible distributions P. For example, if we are only interested in whether a voter is in favor of a given party or not, we may put

$$X_i = \begin{cases} 1 & \text{if } i\text{-th person polled is in favor of party A} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, we have

$$\mathbf{P}(X_i = 1) = 1 - \mathbf{P}(X_i = 0) = p,$$

where p is the percentage of voters in favor of party A. In this case, the distribution is determined by the parameter p.
More generally, if Θ is a subset of $\mathbb{R}^d$ and for any $\theta \in \Theta$, P_θ is a probability distribution, then we call $\{P_\theta,\ \theta \in \Theta\}$ a parametric family of probability distributions. In general, we will assume that such a parametric family is dominated by a σ-finite measure μ, i.e., for each $\theta \in \Theta$ the probability measure P_θ is absolutely continuous with respect to μ. In that case, the distribution P_θ is determined by its Radon-Nikodym derivative $f_\theta = \frac{dP_\theta}{d\mu}$. In most cases of practical interest, the dominating measure is either Lebesgue measure or some counting measure.
If the assumption that the distribution of the sample comes from a given parametric family is made, we speak of parametric statistics, otherwise of nonparametric statistics.

1.2 Sample Statistics

There are lies, damned lies, and statistics.
W. Churchill

A statistic is merely something that can be calculated from a sample. A little more mathematically formulated:

Definition 2 *If f is a function from $\mathbb{R}^n$ to $\mathbb{R}^d$ and $(X_1, \dots, X_n)$ is a random sample, then we call*

$$T = f(X_1, \dots, X_n)$$

a statistic.

Important examples of statistics are the sample mean

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

and the sample variance

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

An important notion is that of a sufficient statistic. Loosely speaking, a sufficient statistic is one that already contains all the information about the parameter that can be obtained from the sample.

Definition 3 *A statistic T is called sufficient for the parameter θ if the conditional distribution of $(X_1, \dots, X_n)$ given T does not depend on θ.*

Remark. The statistic T may of course be d-dimensional.
An important criterion is the following:

Theorem 2 *Let $(P_\theta,\ \theta \in \Theta)$ be a parametric family of distributions dominated by a measure μ, and $f_\theta = \frac{dP_\theta}{d\mu}$. The statistic $T = T(X_1, \dots, X_n)$ is sufficient for θ if the so-called likelihood function*

$$L(x_1, \dots, x_n, \theta) = f_\theta(x_1) \dots f_\theta(x_n)$$

admits a decompositon

$$L(x_1, \dots, x_n, \theta) = g(T(x_1, \dots, x_n), \theta) h(x_1, \dots, x_n),$$

where $h(.)$ does not depend on θ.

Proof. We prove the theorem for the case where μ is a counting measure, i.e., for discrete P_θ. In this case,

$$\mathbf{P}_\theta(T = t) = \sum_{\{(x_1,\dots,x_n):T(x_1,\dots,x_n)=t\}} L(x_1, \dots, x_n, \theta) =$$

$$\sum_{\{(x_1,\dots,x_n):T(x_1,\dots,x_n)=t\}} g(t, \theta) h(x_1, \dots, x_n) = g(t, \theta) H(x_1, \dots, x_n),$$

where $H(.)$ does not depend on θ. Thus

$$\mathbf{P}_\theta(X_1 = x_1, \ldots, X_n = x_n | T = t)) = \begin{cases} \frac{h(x_1,\ldots,x_n)}{H(x_1,\ldots,x_n)} & \text{if } T(x_1,\ldots,x_n) = t, \\ 0 & \text{otherwise,} \end{cases}$$

which does not depend on θ.
One example for a sufficient set of statistics that is always good is the order statistics. These are defined as

$$X_{n:i} = \inf\{x : \#\{j \leq n : X_j \leq x\} \geq i\}.$$

In other words, $(X_{n:1}, \ldots, X_{n:n})$ are $X_1, \ldots, X_n$ in ascending order. It is obvious from the previous criterion that these are sufficient.

1.3 Problems

We can work it out

The Beatles

1. In an urn there are N slips of paper with numbers $x_1, \ldots, x_N$ written on them. Let
$$\mu = \frac{x_1 + \ldots + x_N}{N},$$
$$\sigma^2 = \frac{(x_1 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}.$$
Let $(X_1, \ldots, X_n)$ be the results of drawing n items with replacement. Calculate the expectation and variance of $\bar{X}_n$.

2. In the previous example, calculate the expectation of S_n^2.

3. Solve problem 1 for sampling without replacement.

4. Solve problem 2 for sampling without replacement.

5. Show that $(\bar{X}_n, S_n^2)$ is a pair of sufficient statistics for the normal distributions with densities
$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

6. Find a sufficient statistic for the family of uniform distributions with densities
$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq x \leq \theta, \\ 0 & \text{otherwise} \end{cases}$$

7. Find a pair of sufficient statistics for the family of uniform distributions with densities

$$f_\theta(x) = \begin{cases} 1/(\theta_2 - \theta_1) & \text{if } \theta_1 \le x \le \theta_2, \\ 0 & \text{otherwise} \end{cases}$$

8. Show that $\bar{X}_n$ is sufficient for the family of exponential distributions with densities

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} e^{-\theta x} & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases}$$

9. Find a pair of sufficient statistics for the family of gamma distributions with densities

$$f_{\alpha,\beta}(x) = \begin{cases} \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Chapter 2

Basic Theory of Estimation

How many raods must a man walk down
before you call him a man?
B. Dylan

2.1 Point Estimation

Throughout this chapter, we shall assume that $(X_1, \ldots, X_n)$ is a random sample whose distribution is from some parametric family $(P_\theta, \theta \in \Theta)$. We are trying to find an approximate value for θ from the sample. We define

Definition 4 *An estimator is a sequence $(\hat{\theta}_n, n \in \mathbb{N})$ of statistics, where $\hat{\theta}_n$ is a function of $(X_1, \ldots, X_n)$.*

Remark. Obviously, this definition is there so as to allow us to let the sample size tend to infinity. Apart from that, an estimator can be almost everything (identically 0, for example), so we need a few additional definitions to get a little more sense out of it.

Definition 5 *An estimator $\hat{\theta}_n$ is called*

- *(weakly) consistent if $\hat{\theta}_n \to \theta$ in probability with respect to P_θ (i.e., $\mathbf{P}_\theta(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ for any $\epsilon > 0$),*
- *strongly consistent if $\hat{\theta}_n \to \theta$ with probability one,*
- *unbiased, if $\mathbf{E}_\theta(\hat{\theta}_n) = \theta$,*

- *efficient, if it is unbiased and has minimal variance among all unbiased estimators,*
- *asymptotically unbiased, if* $\mathbf{E}_\theta(\hat{\theta}_n) \to \theta$*,*
- *asymptotically efficient, if it is asymptotically unbiased, and if the ratio of its variance and the infimum of the variances of all unbiased estimators tends to one.*

The most important notion, of course, is consistency, because it ensures that the estimated value will be close to the actual value of the parameter if we choose the sample size n big enough.
Let us consider a few examples:

1. The sample mean $\bar{X}_n$ is an unbiased estimator of the expectation μ of the underlying distribution. The laws of large numbers imply that it is also consistent.
2. The sample variance S_n^2 is an unbiased consistent estimator of the variance σ^2 - that is why we use $n-1$ in the denominator.
3. If we replace the denominator $n-1$ in the definition of S_n^2 by n, the resulting estimator will no longer be unbiased; it will, however, still be consistent and asymptotically unbiased.

Next, we present two methods for calculating estimators. The first one is the <u>method of moments</u>. This originates from the fact that by the law of large numbers, the sample mean converges to the expectation of the underlying distribution with probability one. Thus, if the expectation is a function of the parameter that has a continuous inverse f^{-1}, then $f^{-1}(\bar{X}_n)$ will be a consistent estimator of the parameter. If there is more than one parameter (as for the normal distribution), we use the higher moments $\mathbf{E}(X^k)$ which may be approximately calculated from the sample as

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

For the normal distribution, this yields the equations

$$\hat{\mu} = \bar{X}_n$$

and

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

so, finally

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

The idea behind the second method is to choose the parameter in such a way that the particular sample has the maximal probability of being observed. In other words, we choose $\hat{\theta}$ in such a way that the Likelihood function L is maximal. This is, rather surprisingly, called the maximum likelihood method, and the resulting estimator is called the maximum likelihood (ML) estimator. In most cases, the maximum can be found by using the standard method of zeroing the derivative (or the partial derivatives if there is more than one parameter) of the likelihood function. Usually, it is convenient to first take the logarithm of the likelihood function.

Under rather weak conditions, the ML estimator is consistent; we will not prove this strictly, but rather give a heuristic reasoning. Namely, suppose that θ_0 is the "real" parameter. The logarithm of the likelihood function, evaluated at some θ, is

$$l(\theta) = \sum_{i=1}^{n} \log(f_\theta(X_i)).$$

By the law of large numbers, we have

$$\frac{1}{n} l(\theta) \to \mathbf{E}_{\theta_0}(\log f_\theta(X_i)) = \int \log(f_\theta(x)) f_{\theta_0}(x) d\mu(x).$$

Using the inequality $\log x \leq x - 1$, we get

$$\int \log(f_\theta(x)) f_{\theta_0}(x) d\mu(x) - \int \log(f_{\theta_0}(x)) f_{\theta_0}(x) d\mu(x) =$$

$$\int \log(\frac{f_\theta(x)}{f_{\theta_0}(x)}) f_{\theta_0}(x) d\mu(x) \leq$$

$$\int (\frac{f_\theta(x)}{f_{\theta_0}(x)} - 1) f_{\theta_0}(x) d\mu(x) = \int f_\theta(x) d\mu(x) - \int f_{\theta_0}(x) d\mu(x) = 1 - 1 = 0.$$

Thus, asymptotically, the likelihood function should attain its maximum at θ_0.

We now have a look at efficient estimators. First, we derive a lower bound for the variance of an unbiased estimator:

Theorem 3 *(Cramér-Rao) Let X be a random variable with distribution P_θ, where $\theta \in \Theta$ is a real parameter, and Θ is supposed to be an interval.*

Furthermore, assume that the density $f_\theta(x)$ is twice differentiable with respect to θ and that both $|f'|$ and $|f''|$ (where $'$ denotes differentiation with respect to θ) are bounded above, uniformly in θ, by some integrable function. Furthermore, let $\hat{\theta}$ be an unbiased estimator of θ. Then

$$\mathbf{Var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

where

$$I(\theta) = \mathbf{E}((\frac{\partial \log f_\theta(X)}{\partial \theta})^2) = -\mathbf{E}(\frac{\partial^2 \log f_\theta(X)}{\partial \theta^2})$$

is the so called Fisher information.

Remark. If we have a random sample of size n, we may consider the whole sample as one n-dimensional random variable; it is readily obtained that the Fisher information of this n-dimensional variable is just n times the Fisher information $I(\theta)$ of one coordinate, thus we have for an unbiased estimator $\hat{\theta}_n$ based on the sample of size n:

$$\mathbf{Var}_\theta(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}.$$

This implies that, in cases where the Cramér-Rao theorem is applicable, we have a lower bound of order $1/n$ for the variance.
Proof. By assumption, we have, for all θ,

$$\int f_\theta(x) d\mu = 1$$

and

$$\int \hat{\theta} f_\theta(x) d\mu = \theta.$$

We may take derivatives on both sides of these equations and obtain (we let $'$ denote partial differentiation with respect to θ)

$$\int f'_\theta(x) d\mu = 0$$

and

$$\int \hat{\theta} f'_\theta(x) d\mu = 1.$$

Subtracting θ times the first equation from the second yields

$$\int (\hat{\theta} - \theta) \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x) d\mu = 1.$$

By the Cauchy-Schwarz inequality, we get

$$1 \leq \int (\hat{\theta} - \theta)^2 f_\theta(x) d\mu \int (\frac{f'_\theta(x)}{f_\theta(x)})^2 f_\theta(x) d\mu.$$

In the last equation, the first factor is the variance of $\hat{\theta}$, and the second one is the Fisher information.
What remains to be proven is the equality of the first and second expression for the Fisher information. This is left as an exercise to the reader.
There is, of course, the question if the Cramér-Rao bound can be attained. This is not the case in general; we give a characterization of the distributions that attain the bound in the problem section. Under some additional conditions, however, it can be shown that the ML estimator is asymptotically unbiased and that the ratio of its variance and the Cramér-Rao bound tends to 1, i.e., it is asymptotically efficient.
A final theorem tells us that, in case there is a sufficient statistic, we can limit our search for "good" estimators to those that are a function of the sufficient statistic.

Theorem 4 *If $\hat{\theta}$ is an unbiased estimator and T is a sufficient statistic, then*

$$\tilde{\theta} = \mathbf{E}_\theta(\hat{\theta}|T)$$

is also an unbiased estimator and has a variance that is not greater than that of $\hat{\theta}$.

Proof. First observe that $\tilde{\theta}$ does not depend on θ.
Then,

$$\mathbf{E}_\theta(\tilde{\theta}) = \mathbf{E}_\theta(\mathbf{E}_\theta(\hat{\theta}|T)) = \mathbf{E}_\theta(\hat{\theta}) = \theta,$$

so $\tilde{\theta}$ is unbiased. Furthermore, applying the Cauchy-Schwarz inequality to the conditional expectation, we have

$$\mathbf{E}_\theta(\tilde{\theta}^2) = \mathbf{E}_\theta((\mathbf{E}_\theta(\hat{\theta}|T))^2) \leq \mathbf{E}_\theta(\mathbf{E}_\theta(\hat{\theta}^2|T)) = \mathbf{E}_\theta(\hat{\theta}^2).$$

This, of course, implies that the variance of $\tilde{\theta}$ is at most that of $\hat{\theta}$.

2.2 Confidence Intervals

In the previous section, we studied the problem of finding reasonable estimates for a parameter. We defined some desirable properties of such estimators and gave some general principles for their construction. One question that is left open is that of the accuracy of our estimates —

sometimes we'd like to know how close to the "real" parameter our estimator may be. We are now going to tackle this problem.
We will illustrate some of the ideas of this chapter on the following model:

$$\mathbf{P}(X_i = 1) = \theta,\ \mathbf{P}(X_i = 0) = 1 - \theta,\ (0 < \theta < 1).$$

First, it is obvious that, in general, we can't give any absolute bounds on the parameter (based on a sample); in fact, in our simple example, whatever the value of θ may be, any sequence of zeros and ones has a positive probability to be the actual outcome of our sample. Therefore, we have to settle for a little bit less, and this we do by admitting a small probability of missing the right value of θ:

Definition 6 *A confidence interval with coverage probability γ is an interval $[a, b]$, where a and b are sample statistics with*

$$\mathbf{P}(a \leq \theta \leq b) \geq \gamma.$$

Remarks.

1. Popular choices for γ are 0.95 (most usual) and 0.99.

2. If we have equality in the definition of the confidence interval, then we call it an exact confidence interval. In general, we can't expect exact confidence intervals to exist. In our example and for a sample size of one, for instance, we only have four numbers a_0, b_0, a_1, and b_1 (i.e., one intervall each for the cases $X_1 = 0$ and $X_1 = 1$) to work with, and it is obvious that we can't get a constant probability γ for θ to be in one of those intervals (apart from trivial cases, of course). One way to get exact confidence intervals in cases like this, too, would be to resort to randomization. This means that we carry out an additional random experiment (like rolling a die) and use its outcome to determine the limits of our confidence interval. We won't go into this now, but we will pick this idea up again in the chapter on testing.

3. If there is more than one parameter, we may of course study each one independently and give single confidence intervals for each one of them. In some cases however, this is not quite adequate, and one is more interested in studying confidence regions that have a probability γ of containing the point whose coordinates are the individual parameters. In that case, there is the question of the shape of the confidence region. The most obvious choice is, of course, a cartesian product of one-dimensional intervals, but sometimes other choices, e.g., balls or ellipsoids, are more natural (usually because the associated distributions can be more easily calculated).

4. There are few general rules for the construction of confidence intervals, so we are going to study some cases of practical importance. Of course, if there is a sufficient statistic, we may use this as a starting point; failing that, the ML estimator is sometimes a good starting point (under sufficient regularity, it is not only asymptotically unbiased and efficient, but even has an asymptotic normal distribution with expectation θ and variance $1/nI(\theta)$).

2.2.1 The normal distribution

As the normal distribution has two parameters, there are two types of confidence intervals: one for the mean μ, and one for the variance σ^2. Starting with the mean, let us first assume that we know σ^2. In that case, we try to find the limits of our confidence interval as

$$[\bar{X}_n - c, \bar{X}_n + c]$$

with c determined by the equation

$$\mathbf{P}(\bar{X}_n - c \leq \mu \leq \bar{X}_n + c) = \gamma.$$

The latter leads to

$$\gamma = \mathbf{P}(\mu - c \leq \bar{X}_n \leq \mu + c) = \mathbf{P}(-\frac{c}{\sqrt{\sigma^2/n}} \leq \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq \frac{c}{\sqrt{\sigma^2/n}}) =$$

$$\Phi(\frac{c}{\sqrt{\sigma^2/n}}) - \Phi(-\frac{c}{\sqrt{\sigma^2/n}}) = 2\Phi(\frac{c}{\sqrt{\sigma^2/n}}) - 1.$$

So, finally

$$c = \sqrt{\frac{\sigma^2}{n}}\Phi^{-1}(\frac{1+\gamma}{2}),$$

and our confidence interval becomes

$$\left[\bar{X}_n - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma^2}{n}}, \bar{X}_n + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma^2}{n}}\right],$$

where we have replaced the notation $\Phi^{-1}(\alpha)$ by the more usual z_α. Now, let us turn to the case where σ^2 is unknown. In that case, we cannot use the confidence interval from above because it still contains the unknown variance. Of course, we have to replace it by the sample variance, and there are actually two ways to do this:

1. We may just replace σ^2 by S_n^2 and treat the resulting confidence interval as an approximate one (suitable if n is large).

2. Or, we may turn the approximate confidence interval into an exact one by making use of the following

Theorem 5 *Let $X_1, X_2, \ldots$ be i.i.d. normal random variables with mean μ and parameter σ^2. Then*

1. *$\bar{X}_n$ has a normal distribution with mean μ and variance σ^2/n.*

2. *$\frac{n-1}{\sigma^2} S_n^2$ has a χ^2-distribution with $n-1$ degrees of freedom, which has density*
$$\frac{x^{\frac{n-3}{2}}}{\Gamma(\frac{n-1}{2}) 2^{\frac{n-1}{2}}} e^{-\frac{x}{2}} \quad (x > 0).$$

3. *$\bar{X}_n$ and S_n^2 are independent.*

4. *$\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}}$ has a t- (or Student-) distribution with $n-1$ degrees of freedom, which has density*
$$\frac{1}{\sqrt{\pi(n-1)}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \left(1 + \frac{x^2}{n-1}\right)^{-\frac{n}{2}}.$$

We omit the proof.
Using this theorem, we get the confidence interval for μ:

$$[\bar{X}_n - t_{n-1,\frac{1+\gamma}{2}} \sqrt{\frac{S_n^2}{n}}, \bar{X}_n + t_{n-1,\frac{1+\gamma}{2}} \sqrt{\frac{S_n^2}{n}}],$$

where $t_{n,\alpha}$ denotes the α-quantile of the t-distribution.
Using the same ideas as above, we calculate the following confidence intervals for σ^2:

1. If μ is known:
$$[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,\frac{1+\gamma}{2}}}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,\frac{1-\gamma}{2}}}],$$

2. If μ is unknown:
$$[\frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{1+\gamma}{2}}}, \frac{(n-1)S_n^2}{\chi^2_{n-1,\frac{1-\gamma}{2}}}].$$

2.2.2 Proportions

A question that arises very often in practice is that for the proportion of individuals in the population that have a certain property (e.g., voters of a given party, defective items in some manufacturing process). We already know that this corresponds to finding the parameter θ in our standard model

$$\mathbf{P}(X=1)=\theta,\, \mathbf{P}(X=0)=1-\theta,\ (0<\theta<1).$$

We are going to calculate an approximate confidence interval for θ. Our starting point for this is the central limit theorem which tells us that $\bar{X}_n$ has an approximate normal distribution with mean θ and variance $\frac{\theta(1-\theta)}{n}$. This would give us the confidence interval

$$[\bar{X}_n - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\theta(1-\theta)}{n}}, \bar{X}_n + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\theta(1-\theta)}{n}}],$$

but, unfortunately, we don't know the value of θ. There are two ways out of this dilemma: the simpler one would be to simply replace θ by its estimator $\bar{X}_n$; if we want to be a little more exact, we can solve the equation

$$\bar{X}_n \pm z_{\frac{1+\gamma}{2}}\sqrt{\frac{\theta(1-\theta)}{n}} = \theta$$

and use the solutions of this equation (which will become quadratic if properly handled) as the limits of our confidence interval.

2.3 Problems

With a little help from my friends
The Beatles

1. Show that the distributions that attain the Cramér-Rao bound have densities of the form

$$f(x)=\exp(c(x)+a(x)A(\theta)+B(\theta)),$$

where

$$B'(\theta)=-\theta A'(\theta).$$

Besides, the efficient estimator is

$$\hat{\theta}=a(X),$$

or, for the case of a sample of size n

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} a(X_i),$$

and coincides with the ML estimator.

2. Calculate the ML estimator for the parameter λ of the exponential distribution. Calculate the expectation and find a factor that will make it unbiased.

3. Let $X_1, \ldots, X_n$ be a sample from a uniform distribution on $[0, \theta]$ $(\theta > 0)$.

   (a) Calculate the ML estimator for θ.

   (b) Calculate an estimator for θ using the moment method.

   (c) Modify the ML estimator so that it is unbiased.

   (d) Compare the variances of the estimators in (b) and (c).

4. For the exponential distribution, calculate the Cramér-Rao bound for the variance of an unbiased estimator of the parameter λ. Does the unbiased estimator from problem 2 attain this bound?

5. Let $\mathbf{P}(X_i = 1) = \theta$, $\mathbf{P}(X_i = 0) = 1 - \theta$ $(0 < \theta < 1)$. Show that $\bar{X}_n$ is an efficient estimator for θ.

6. Show that the ML estimator for the parameter λ of a Poisson distribution is efficient.

7. Let $X_1, \ldots, X_n$ be a sample from a uniform distribution on $[\theta, 2\theta]$ $(\theta > 0)$.

   (a) Calculate the ML estimator for θ.

   (b) Modify the ML estimator so that it becomes unbiased.

   (c) Show that

   $$\tilde{\theta} = \frac{1}{3}(\max(X_1, \ldots, X_n) + \min(X_1, \ldots, X_n))$$

   is another unbiased estimator, and that it has smaller variance than the one in part (b).

8. Let $(X_1, \ldots, X_n)$ be a sample from a distribution with density

$$f_\theta(x) = \frac{1}{2}\exp(-|x-\theta|).$$

Calculate the ML estimator for θ.

9. Assume that the following sample is from a normal distribution:

1.1 2.1 1.5 2.3 1.4 2.6 1.7 1.3 2.8 1.6

Calculate 95% confidence intervals for μ and σ^2.

10. In problem 9, calculate a 99% confidence interval for μ if $\sigma^2 = 0.5$ is known.

11. Use the central limit theorem to calculate an approximate confidence interval for the parameter λ of the Poisson distribution.

12. Use the central limit theorem to calculate an approximate confidence interval for the parameter λ of the exponential distribution.

13. Let $X_1, X_2, \ldots, X_n$ be a sample from an exponential distribution. Use the fact that $2n\lambda\bar{X}_n$ has a χ^2-distribution with $2n$ degrees of freedom to calculate a confidence interval for λ.

14. Let X be Poisson with parameter λ, and Y χ^2 with $2n$ degrees of freedom. Use the equality

$$\mathbf{P}(X \geq n) = \mathbf{P}(Y \leq 2\lambda)$$

to calculate a confidence interval for the parameter λ of a Poisson distribution.

(Hint: we try to find $l(x) < u(x)$ with $\mathbf{P}_\lambda(\lambda < l(X)) = \mathbf{P}(\lambda > u(X)) = \frac{1-\gamma}{2}$. Replace λ by $l(k)$ resp. $u(k)$ and use the above relation.)

15. Calculate a confidence interval for the parameter $\theta > 0$ of a uniform distribution on $[0, \theta]$.

Chapter 3

Testing

A hundred experiments canprove me right yet, one single experiment can prove me wrong.

Sir Isaac Newton

3.1 Basic Notions

The problems we are studying in this chapter are of the following kind: based on a sample, we have to decide whether the underlying distribution has a given property or not. We may illustrate this with a simple example: Suppose we want to take part in the statisticians' favorite game — coin tossing — and, in order to win for sure, we have a coin made that shows "heads" with probability 3/4 and "tails" with probability 1/4. But, as things go, nobody took care which way around the coin was struck, so we don't know which side is the one with the greater probability.

In this case, what we do, of course, is flip the coin a number of times (100, say) and choose the one that shows more often as the one with higher probability.

Using this way of making our decision, we can go wrong in either of two ways — we may decide for "heads" when "tails" would have been the right choice, and vice versa (In this example, both possibilities are symmetric, but this is not usually the case). The probability of making a wrong decision is

$$\mathbf{P}(X < 50|\theta = .75) = \Phi(\frac{50 - 75}{\sqrt{100 * .75 * .25}}) = \Phi(-5.78) = 4 * 10^{-9}.$$

In general, things are less symmetric than in this example, and for either one of the two possibilities there may be more than one choice for the underlying probability distribution. This leads us to

Definition 7 *A hypothesis is any subset H of the set of probability distributions in our statistical model.*
If we have a parametric model, a hypothesis can be expressed in terms of the parameter. In that case, we speak of parametric hypotheses.
If the hypothesis contains only one distribution, it is called a simple hypothesis, otherwise it is a composite hypothesis.
For parametric hypotheses, we can further distinguish one-sided (like $\theta < a$ or $\theta > b$) and two-sided (like $\theta \neq a$ or $\theta \notin [a, b]$) hypotheses.

With these definitions, we can restate the task of a statistical test as: decide between two hypotheses, using a random sample.
We are going to break symmetry now and call one hypothesis the "null" hypothesis H_0 and the other one the "alternative" (or sometimes the "one-hypothesis") H_1. Instead of saying which hypothesis we decide for, we are going to say that we "accept" or "reject" the null hypothesis. As the decision is based on a sample, we can define our test by the set of sample values that will make us reject the null hypothesis — the so-called "rejection region", or, equivalently, by the indicator function ϕ of the rejection region, which has the nice feature that $\phi(X_1, \ldots, X_n)$ is just the index of the hypothesis we choose as our decision. Thus, we arrive at the definition

Definition 8 *A (non-randomized) test is a function from $\mathbb{R}^n$ to $\{0, 1\}$.*
A randomized test is a function from $\mathbb{R}^n$ to $[0, 1]$, with the interpretation that $\phi(X_1, \ldots, X_n)$ is the probability that we reject H_0 (e.g., if $\phi = 0.5$, we may flip a coin and reject if it shows "heads" and accept if it shows "tails").

Now that we have defined our test, we may look at how things can go wrong:

Definition 9 *An error of the first kind occurs if we reject H_0 although it is true.*
An error of the second kind occurs if we accept H_0 although it is wrong.

Now, the probabilities of incurring an error of the first or second kind are functions of the actual distribution, so there is no fixed value for them if our hypotheses are composite. In fact, if $P \in H_0$,

$$\mathbf{P}(\text{first kind error}|P) = \mathbf{E}_P(\phi),$$

whereas for $P \in H_1$

$$\mathbf{P}(\text{second kind error}|P) = \mathbf{E}_P(1-\phi).$$

The usual way to cope with this fact is to impose an upper limit on the probability of a first kind error, which is usually denoted by α and called the "level of significance" of the test (the most common choice is $\alpha = .05$). Viewed as a function of the actual distribution (or of the actual parameter) the probability of rejecting H_0 is called the power function of the test, and the probability of accepting H_0 is called the operation characteristic of the test.
In most cases of practical interest, it is convenient to express the test in terms of a test statistic (often an estimator of the parameter, or some sufficient statistic), and to reject if the value of the test statistic exceeds (or falls short of) some critical value. So, the usual procedure in actually carrying out a test consists of the following steps:

1. State the null hypothesis and alternative.
2. Fix the level α.
3. Choose an appropriate test statistic.
4. Use the distribution of the test statistic under the null hypothesis to calculate the critical value.
5. Calculate the value of the test statistic from the sample and make a decision based on whether this value is above or below the critical value.

3.2 The Neyman-Pearson and likelihood ratio test

For a given level of significance, one would like to have the probability of an error of the second kind as small as possible. Of course, the question arises whether this can be achieved. In the case that both H_0 and H_1 are simple, this question is answered by the following

Theorem 6 *(Neyman-Pearson) If $H_0 = \{P_0\}$ and $H_1 = \{P_1\}$, then the best (randomized) test with level α (based on one observation) is the Neyman-Pearson test which has the form*

$$\phi(x) = \begin{cases} 1 & \text{if } f_1(x) > kf_0(x), \\ c & \text{if } f_1(x) = kf_0(x), \\ 0 & \text{if } f_1(x) < kf_0(x), \end{cases}$$

where

$$f_i = \frac{dP_i}{d(P_0 + P_1)},$$

(actually, any density with respect to a common dominating measure works) and $k \in [0, \infty]$ *and* $c \in [0, 1]$ *are calculated from the equation*

$$\mathbf{E}_0(\phi) = \alpha.$$

(We write $\mathbf{E}_0$ *and* $\mathbf{E}_1$ *to denote expectations with respect to* P_0 *and* P_1*, respectively)*

Remark Although the theorem is stated for sample size one, the general case can be handled by viewing a sample of size n as one observation of an n-dimensional random variable; the Radon-Nikodym derivatives in the theorem then become the respective likelihood functions.
Proof. It is quite obvious that k and c can be chosen in such a way that $\mathbf{E}_0(\phi) = \alpha$. What we have to prove, then, is that the test we construct in this way is optimal, i.e., that any other test of level α has a larger probability for an error of the second kind. In other words, we have to prove that for any test ψ with

$$\mathbf{E}_0(\psi) \leq \alpha$$

we have

$$\mathbf{E}_1(\psi) \leq \mathbf{E}_1(\phi).$$

To this end, consider

$$\mathbf{E}_1(\psi) - ki\mathbf{E}_0(\psi).$$

Letting $P = P_0 + P_1$, this becomes

$$\int \psi(x)(f_1(x) - kf_0(x))dP(x) = \int_{f_1(x) < kf_0(x)} \psi(x)(f_1(x) - kf_0(x))dP(x) +$$

$$\int_{f_1(x) = kf_0(x)} \psi(x)(f_1(x) - kf_0(x))dP(x) + \int_{f_1(x) > kf_0(x)} \psi(x)(f_1(x) - kf_0(x))dP(x) \leq$$

$$\int_{f_1(x) < kf_0(x)} 0(f_1(x) - kf_0(x))dP(x) + \int_{f_1(x) = kf_0(x)} c(f_1(x) - kf_0(x))dP(x) +$$

$$\int_{f_1(x) > kf_0(x)} 1(f_1(x) - kf_0(x))dP(x) =$$

$$\mathbf{E}_1(\phi) - k\mathbf{E}_0(\phi).$$

So,

$$\mathbf{E}_1(\psi) \leq \mathbf{E}_1(\phi) + k(\mathbf{E}_0(\psi) - \alpha) \leq \mathbf{E}_1(\phi).$$

Based on the ideas behind the Neyman-Pearson theorem is the following definition which gives us a general principle for constructing tests. Although there is no general theorem like the Neyman-Pearson theorem for it, it gives useful tests in many situations:

Definition 10 *The likelihood ratio test is given by*

$$\phi(X_1, \ldots, X_n) = \begin{cases} 0 & \textit{if } l > \lambda, \\ c & \textit{if } l = \lambda, \\ 1 & \textit{if } l < \lambda, \end{cases}$$

where the likelihood ratio statistic l is defined by

$$l = \frac{\sup_{\theta \in H_0} L(X_1, \ldots, X_n, \theta)}{\sup_{\theta \in \Theta} L(X_1, \ldots, X_n, \theta)}.$$

3.3 Special tests for the normal distribution

3.3.1 Tests for μ

In testing hypotheses for the mean μ of a normal distribution, we have to distinguish two cases, according to whether the variance σ^2 is known or not. In the easier case where σ^2 is known, we can start out from simple hypotheses and the Neyman-Pearson lemma which gives us (see the problem section), assuming $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1 > \mu_0$, rejection for

$$\bar{X}_n > \mu_0 + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}.$$

The first observation that may be made is that the actual value of μ_1 does not enter in this test. This means that this test is the best test with level α for all $\mu_1 > \mu_0$. So, we may use the same test for testing $H_0 : \mu = \mu_0$ against the composite alternative $H_1 : \mu > \mu_0$. Further investigation shows that we can replace the simple null hypothesis by $H_0 : \mu \leq \mu_0$ because for $\mu < \mu_0$, the probability of rejecting decreases. Thus, the above test is optimal for testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.
For the two-sided alternative, namely for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, we have two possible approaches. First, we may use some intuitive reasoning and use a test that rejects if the distance $\bar{X}_n - \mu$ is large in absolute value. This leads us to rejecting if

$$|\bar{X}_n - \mu_0| > z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}.$$

The second approach is to use the likelihood ratio method, which gives the same result.
Now, if σ^2 is unknown, we are going to replace it by its estimator, S_n^2. As was the case with confidence intervals, we have to use the quantiles of the t-distribution with $n-1$ degrees of freedom and get the rejection regions

$$\bar{X}_n - \mu_0 > t_{n-1;1-\alpha}\sqrt{\frac{S_n^2}{n}}$$

or

$$\bar{X}_n - \mu_0 < t_{n-1;1-\alpha}\sqrt{\frac{S_n^2}{n}}$$

for the one-sided problems and

$$|\bar{X}_n - \mu_0| > t_{n-1;1-\frac{\alpha}{2}}\sqrt{\frac{S_n^2}{n}}$$

in the two-sided case. These are usually called the t-tests.

3.3.2 Tests for σ^2

Using the ideas of the preceding section, we get the following tests for the variance of a normal distribution: indextest!normal distribution!variance
If μ is known:
For $H_0 : \sigma^2 \leq \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$: reject if

$$\sigma_0^{-2}\sum_{i=1}^{n}(X_i - \mu)^2 > \chi^2_{n;1-\alpha}.$$

For $H_0 : \sigma^2 \geq \sigma_0^2$ against $H_1 : \sigma^2 < \sigma_0^2$: reject if

$$\sigma_0^{-2}\sum_{i=1}^{n}(X_i - \mu)^2 < \chi^2_{n;\alpha}.$$

For $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$: reject if

$$\sigma_0^{-2}\sum_{i=1}^{n}(X_i - \mu)^2 > \chi^2_{n;1-\frac{\alpha}{2}}.$$

or

$$\sigma_0^{-2}\sum_{i=1}^{n}(X_i - \mu)^2 < \chi^2_{n;\frac{\alpha}{2}}.$$

If μ is unknown:
For $H_0 : \sigma^2 \leq \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$: reject if

$$\sigma_0^{-2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 > \chi^2_{n-1;1-\alpha}.$$

For $H_0 : \sigma^2 \geq \sigma_0^2$ against $H_1 : \sigma^2 < \sigma_0^2$: reject if

$$\sigma_0^{-2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 < \chi^2_{n-1;\alpha}.$$

For $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$: reject if

$$\sigma_0^{-2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 > \chi^2_{n-1;1-\frac{\alpha}{2}}.$$

or

$$\sigma_0^{-2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 < \chi^2_{n-1;\frac{\alpha}{2}}.$$

3.3.3 Two-sample tests

In some cases, one wants to compare two samples from different populations having a normal distribution. Thus, there are two samples: $(X_1, \ldots, X_n)$ from a normal distribution with mean μ_1 and variance σ_1^2 and $(Y_1, \ldots, Y_m)$ with mean μ_2 and variance σ_2^2. The most common question is: are the two means equal? Unfortunately, there is no simple test for this question without a restriction on the variance; if the two variances are equal, however, we can use the test statistic

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

which has a t-distribution with $n + m - 2$ degrees of freedom.
If we don't have any information on the relative sizes of the two variances, we can use the following test statistic:

$$\frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \frac{1}{m(m-1)} \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2}}$$

which has an approximate t distribution whose number of degrees of freedom f is obtained from the equation

$$\frac{1}{f} = \frac{c^2}{m-1} + \frac{(1-c)^2}{n-1},$$

where

$$c = \frac{\frac{1}{m(m-1)} \sum_{j=1}^{m} (Y_j - \bar{Y}_m)^2}{\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 + \frac{1}{m(m-1)} \sum_{j=1}^{m} (Y_j - \bar{Y}_m)^2}.$$

From the above it seems natural that one might want to test for the equality of two variances, too. The case of the highest practical importance is the one where both means are unknown. In that case one uses the test statistic

$$F = \frac{S_X^2}{S_Y^2}.$$

where

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

and

$$S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y}_m)^2.$$

which has an F-distribution with $(n-1, m-1)$ degrees of freedom. The F-distribution with (a, b) degrees of freedom is the distribution of bX/aY, where X and Y are independent χ^2-distributed with a resp. b degrees of freedom.

3.3.4 Unbiased tests.

Above, we have seen that the one-sided test for the mean μ of a normal distribution with known variance is optimal in the sense that any other test of the same level has at least the same probability of a second kind error for any $\mu \in H_1$. One may ask (and that's what we are doing) whether there is a similar optimality criterion for the two-sided test. We cannot expect that without further conditions, as the best test for $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ as well as the one for $H_1 : \mu < \mu_0$ can be used as tests for the two-sided alternatives. A test that is optimal for all $\mu \neq \mu_0$ should therefore have the same probability for an error of the second kind as the test for $H_1 : \mu < \mu_0$ if $\mu < \mu_0$, and the same as the test for $H_1 : \mu > \mu_0$ if $\mu > \mu_0$. This, however, is not possible. Thus, we have to add some additional restriction to our tests:

Definition 11 *A test ϕ of level α is called unbiased if for any $P \in H_1$:*

$$\mathbf{E}_P(\phi) \geq \alpha.$$

(This is a natural requirement: we wouldn't want any distribution from H_1 to gives us a higher probability of accepting H_0 than the distributions from H_0.)
For a parametric test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, and assuming that the densities $f_\theta(x)$ are sufficiently differentiable, this implies that for an unbiased test

$$\mathbf{E}_{\theta_0}(\phi) = \alpha,$$

and

$$\frac{\partial}{\partial \theta}(\mathbf{E}_\theta(\phi))|_{\theta=\theta_0} = 0.$$

For a fixed θ_1, we may copy the idea of the Neyman-Pearson lemma and try to find a test ϕ of the following form:

$$\phi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{if } L(X_1, \ldots, X_n, \theta_1) > kL(X_1, \ldots, X_n, \theta_0) + \tilde{k}L'(X_1, \ldots, X_n, \theta_0), \\ c & \text{if } L(X_1, \ldots, X_n, \theta_1) = kL(X_1, \ldots, X_n, \theta_0) + \tilde{k}L'(X_1, \ldots, X_n, \theta_0), \\ 0 & \text{if } L(X_1, \ldots, X_n, \theta_1) < kL(X_1, \ldots, X_n, \theta_0) + \tilde{k}L'(X_1, \ldots, X_n, \theta_0). \end{cases}$$

If we can find such a test (with $0 \leq c \leq 1$, $k \geq 0$ and $\tilde{k}$ determined by the equations

$$\mathbf{E}_{\theta_0}(\phi) = \alpha,$$

and

$$\frac{\partial}{\partial \theta}(\mathbf{E}_\theta(\phi))|_{\theta=\theta_0} = 0,$$

then we can prove in the same way as for the Neyman-Pearson lemma, that any other unbiased test must have a greater probability of a second kind error (it is not guaranteed, however, that we succeed); of course we hope that the actual choice of θ_1 doesn't enter in the final form of the test, so that our test will be uniformly optimal for all $\theta \neq \theta_0$.
Let us try this for the mean of a normal distribution with known variance: in that case, the likelihood function is

$$L(X_1, \ldots, X_n, \mu) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{2\sigma^2}\right).$$

For its derivative, we have

$$L'(X_1, \ldots, X_n, \mu) = \frac{n(\mu - \bar{X}_n)}{\sigma^2} \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{2\sigma^2}\right).$$

After some calculation, the inequality

$$L(X_1, \dots, X_n, \mu_1) > kL(X_1, \dots, X_n, \mu_0) + \tilde{k}L'(X_1, \dots, X_n, \theta_0)$$

turns into

$$\exp(\frac{n(2\bar{X}_n - \mu_0 - \mu_1)(\mu_1 - \mu_0)}{2\sigma^2}) < k + \tilde{k}\frac{n(\mu - \bar{X}_n)}{\sigma^2}.$$

The right hand side is linear in $\bar{X}_n$, whereas the left hand side is a convex function of $\bar{X}_n$. Thus, there are (at most) 2 points a and b where the two functions coincide, and the equality above is equivalent to

$$a < \bar{X}_n < b.$$

The distribution of $\bar{X}_n$ is continuous, so we don't need the number c in the definition of our test, and it becomes

$$\phi = \begin{cases} 0 & \text{if } a \leq \bar{X}_n \leq b, \\ 1 & \text{if } \bar{X}_n < a \text{ or } \bar{X}_n > b. \end{cases}$$

a and b are to be determined from the conditions

$$\mathbf{E}_{\mu_0}(\phi) = \alpha,$$

and

$$\frac{\partial}{\partial \mu}(E_\mu(\phi))|_{\mu=\mu_0} = 0).$$

We first calculate

$$\mathbf{E}_\mu(\phi) = \mathbf{P}_\mu(\bar{X}_n < a) + \mathbf{P}_\mu(\bar{X}_n > b) = \Phi(\frac{a - \mu}{\sqrt{\sigma^2/n}}) + 1 - \Phi(\frac{b - \mu}{\sqrt{\sigma^2/n}}).$$

taking derivatives with respect to μ, we obtain

$$\frac{\partial}{\partial \mu}(E_\mu(\phi)) = \frac{1}{\sqrt{2\sigma^2/n}}e^{-\frac{(b-\mu)^2}{2\sigma^2/n}} - \frac{1}{\sqrt{2\sigma^2/n}}e^{-\frac{(a-\mu)^2}{2\sigma^2/n}}.$$

Thus, the equation

$$\frac{\partial}{\partial \mu}(E_\mu(\phi))|_{\mu=\mu_0} = 0)$$

is equivalent to

$$(a - \mu)^2 = (b - \mu)^2,$$

and the final test turns out to be identical to the one we already know. This method works for most of the "classical" distributions, though the resulting test may not be very easy to handle (even the simple example of proportions poses some problems).

3.4 Problems

We can swing together 'cause we think we're doin' it right

Lindisfarne

1. For a sample from a normal distribution, calculate the best test for $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ if σ^2 is known.

2. For a sample from a normal distribution, calculate the likelihood ratio test for $H_0 : \mu = \mu_0, 0 < \sigma^2 < \infty$ against the two-sided alternative. Show that this test is equivalent to the t-test.

3. For a sample from a Poisson distribution, find the best test for $\lambda = \lambda_0$ against $\lambda = \lambda_1$.

4. For a sample from a uniform distribution on $[0, \theta]$ construct a test for $H_0 : \theta < \theta_0$ against $H_1 : \theta > \theta_0$.

5. If there is a sufficient statistic, show that the likelihood ratio test is always a function of that sufficient statistic.

6. Find the form of the best test for $H_0 : \mu = \mu_0, \sigma^2 = \sigma_0^2$ against $H_1 : \mu = \mu_1, \sigma^2 = \sigma_1^2$ (for a normal distribution).

7. For an exponential distribution, construct a test for testing $H_0 : \lambda = \lambda_0$ against the two-sided alternative.

8. Let $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ be two samples from normal distributions with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively. Calculate the likelihood ratio test for testing $\sigma_1^2 = \sigma_2^2$ and show that it is equivalent to the F-test.

9. In a sample of 200 products, 15 defective items were found. The manufacturer claims that the percentage of defective items is at most 5%. Test at the 5% level.

10. A sample of 60 bottles of beer showed a mean content of 470 ccm with a sample variance of 2000. Assuming a normal distribution, test $H_0 : \mu = 500$.

11. The lifetime of lightbulbs is supposed to be exponentially distributed. A certain brand is claimed to have an expected lifetime of 2000 hours. A sample showed the following lifetimes:

    900 1760 2140 2050 2820 2200 3000 600 1500 1980

    Test at the 5% level.

12. For the following sample from a normal distribution

    1.1 0.7 1.6 1.3 0.8 1.4 0.7 0.6 1.8 1.4 1.3

    test $H_0 : \mu = 1.0$ against the two-sided alternative, if $\sigma^2 = 0.5$

13. Do the previous problem for unknown variance.

14. For the sample in problem 12, test whether $\sigma^2 = 0.5$ ($\alpha = 0.1$).

15. Calculate the density of the F-distribution.

16. Use the central limit theorem to show that for large n

$$\chi^2_{n,p} \approx n + z_p\sqrt{2n}.$$

17. Use the result of the previous problem to prove that for large a and b

$$F_{a,b,p} \approx 1 + z_p\sqrt{\frac{2(a+b)}{ab}}.$$

    (hint: Replace the χ^2_n-distributed random variables by $n + U\sqrt{2n}$, where U is approximately standard normal.)

18. $((X_1, Y_1), \ldots, (X_n, Y_n))$ is a sample from a two-dimensional normal distribution. Devise a test for $H_0 : \mu_X = 2\mu_Y$.

    (hint: any linear combination of X_n and Y_n is normally distributed.)

19. Suppose that $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ are samples from normal distributions with equal variances. Construct a test for $H_0 : \mu_X = 2\mu_Y$.

    (hint: this should look similar to the test for equal means, only the denominator has to be changed, and the scaling factor needs some adjustment.)

20. The following is another sample from a normal distribution:

    1.0 0.9 0.7 1.4 1.1 0.8 0.6 0.9 1.1

    Assuming equal variances, test whether this has the same mean as the sample in problem 12

21. In the previous problem, use the F-test to check whether the variances are equal ($\alpha = .1$).

Chapter 4

Analysis of Variance (ANOVA)

Little boxes on the hillside
little boxes made of ticky-tacky
little boxes on the hillside
little boxes all the same

Pete Seeger

4.1 The Fisher-Cochran Theorem

The name "analysis of variance" is slightly misleading; in fact, what this chapter is about is a comparison of the means of several samples from different normal distributions. The variances don't enter the picture - they are even supposed to be the same for all the individual samples. There is, however, a reason why this name persists: namely, we use a decomposition of the sum of the squares of the sample values for our test.

To put things in more precise terms, let k samples $(X_{i1}, \ldots, X_{in_i})$ $(i = 1, \ldots, k)$ be given. assume that X_{ij} has a normal distribution with mean μ_i and variance σ^2 (the same for all samples). We want to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative that at least one μ_i is different from the others. This question could of course also be approached by using the likelihood ratio method and you will be pleased to know (left as an exercise) that the method we develop here is equivalent to the likelihood ratio test.

We start out from the sum

$$S_I = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_N)^2,$$

where

$$N = \sum_{i=1}^{k} n_i$$

and

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}$$

is the mean of the combined samples.
By simple calculations,

$$\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_N)^2 = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 + n_i(\bar{X}_i - \bar{X}_N)^2,$$

where

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

is the mean of the ith sample. Thus, we arrive at the decomposition

$$S_I = S_{II} + S_{III},$$

where

$$S_{II} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2,$$

and

$$S_{III} = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_N)^2.$$

The distribution of S_{II} is not affected by the actual values of $\mu_1, \ldots, \mu_k$. For S_{III}, we can expect it to be smallest if all means are equal, and to increase if we depart from this hypothesis (because $\bar{X}_i$ is close to μ_i). So, we will reject our null hypothesis of equal means if S_{III} is large with respect to S_{II}.
In order to obtain the critical value for such a test, we need the distributions of the sums S_{II} and S_{III}. To obtain those, we will do a little general study of sums of squares. As usual, we start with some definitions:

Definition 12 *Let $X_1, \ldots, X_n$ be independent standard normal random variables. Furthermore, let*

$$Y_i = \sum_{j=1^n} a_{ij} X_j \quad (i = 1, \ldots, k),$$

where (a_{ij}) are real numbers. Then we call

$$\sum_{i=1}^{k} Y_i^2$$

a sum of squares.
Furthermore, the maximum number of linearly independent random variables among the Y's (which is k minus the number of linear relations among the Y's) is called the number of degrees of freedom of the sum of squares. This is the maximum number of Y's that can be independently chosen, whereas the rest will follow from the linear relations among them.

The first thing to observe is that we have to prove that the number of degrees of freedom is well defined, i.e., it does not depend on the special choice of a_{ij}. It seems most convenient to use matrix notation for this. So, let

$$X = (X_1, \ldots, X_n)^T$$

and

$$Y = (Y_1, \ldots, Y_k)^T$$

be the column vectors whose elements are the numbers X_i resp. Y_i, and

$$A = (a_{ij})_{k \times n}$$

be the matrix whose elements are a_{ij}.
Then

$$Y = AX$$

and the sum of squares can be written as

$$S = Y^T Y = X^T A^T A X.$$

The number of degrees of freedom of S is readily identified as the rank of A. But this is also the rank of $A^T A$, which follows from the fact that their null spaces coincide. Namely, if

$$Ax = 0,$$

then also

$$A^T Ax = 0;$$

on the other hand

$$A^T Ax = 0$$

implies

$$(Ax)^T (Ax) = x^T A^T Ax = 0,$$

so we also have

$$Ax = 0.$$

Now, it is immediate that, if $\tilde{A}$ is another matrix that generates the same sum of squares, then $\tilde{A}^T \tilde{A} = A^T A$, so the ranks of A and $\tilde{A}$ must be the same.

Examples.

1. $X_1^2 + \ldots + X_k^2$ is the simplest sum of squares, and has k degrees of freedom.

2. $(X_1 - \bar{X}_n)^2 + \ldots + (X_n - \bar{X}_n)^2$ is another sum of squares. Its number of degrees of freedom is $n - 1$, because if we put

   $$Y_i = X_i - \bar{X}_n,$$

   then we have the linear relation

   $$\sum Y_i = 0.$$

   It is obvious that there are no other linear relations.

3. $5X_1^2 + 3x_2^2 + 17(3X_1 - 5X_2)^2$ has 2 degrees of freedom.

The principal tool in our treatment of sums of squares is the following

Theorem 7 *(FISHER-COCHRAN) If the sum of squares S has a χ^2-distribution with f degrees of freedom and if, for $(i = 1, \ldots, k)$, S_i is a sum of squares with f_i degrees of freedom satisfying*

$$S = S_1 + \ldots + S_k,$$

then $S_1, \ldots, S_n$ are independent and S_i is χ^2-distributed with f_i degrees of freedom if and only if

$$f = f_1 + \ldots + f_k.$$

Remark.

1. Observe that the independence of $S_1, \ldots, S_k$ is not required; instead, it will follow from the theorem if the sum of their numbers of degrees of freedom is the number of degrees of freedom of their sum.

2. It is a surprising fact that if a sum of squares is χ^2-distributed with f degrees of freedom, then f is the number of degrees of freedom of f (this can be deduced from the proof below).

Proof. The "if" part simply states that a sum of independent χ^2-distributed random variables is χ^2-distributed with a number of degrees of freedom that is the sum of the numbers of degrees of freedoms of the summands. So, what remains to be proven is the other direction. First, it is sufficient to prove the theorem for $k = 2$. Namely, it is obvious that the number of degrees of freedom of a sum of sums of squares is at most the sum of the numbers of degrees of freedom of the summand terms. Thus $S_2 + \ldots + S_k$ can have at most $f_2 + \ldots + f_k$ degrees of freedom. If it had less than that, then $S_1 + \ldots + S_k$ would have less than $f_1 + \ldots + f_k$ degrees of freedom, which is in violation of our assumptions. Thus, the general case is obtained from $k = 2$ by induction.
Assume that
$$S = X_1^2 + \ldots + X_f^2.$$
This entails no loss of generality, because any sum of squares can be written in the form
$$S = \sum_{i=1}^{k} a_i Y_i^2,$$
where Y_i are independent standard normal random variables that are linear combinations of X_i, and a_i are some positive constants. Namely, if we use matrix notation again, we have
$$S = X^T A^T A X.$$
$A^T A$ is a symmetric nonnegative definite marix, so there is an orthogonal matrix B (i.e., $B^T B = I$) such that $B^T A^T A B$ is a diagonal matrix (its diagonal elements are the eigenvalues λ_i of $A^T A$). Let
$$Y = B^T X.$$
It follows that Y is a vector of independent standard normal random variables, and we have
$$S = X^T A^T A X = Y^T B^T A^T A B Y = \sum_{i=1}^{f} \lambda_i Y_i^2.$$

It is an easy application of the theory of characteristic functions (left as an exercise to the reader) to prove that if that sum has a χ^2-distribution with f degrees of freedom, then $k = f$ and $a_i = 1$.
Now, S_1 and S_2 can be written as

$$S_i = X^T A_i^T A_i X \ (i = 1, 2),$$

where A_i are $k_i \times f$ matrices with rank f_i. As $A_i^T A_i$ is a symmetric nonnegative definite matrix, there is again an orthogonal $f \times f$ matrix B (i.e., $B^T B = I$) such that $B^T A_1^T A_1 B$ is a diagonal matrix (and the diagonal elements are the eigenvalues λ_i of $A_1^T A_1$). Let

$$Z = B^T X.$$

It follows that Z is a vector of independent standard normal random variables, and we have

$$\sum_{i=1}^{f} Z_i^2 = Z^T Z = X^T B B^T X = X^T X = \sum X_i^2 = S$$

and

$$S_1 = X^T A_1^T A_1 X = Z^T B^T A_1^T A_1 B Z = \sum_{i=1}^{f} \lambda_i Z_i^2.$$

Obviously, the number of degrees of freedom of S_1 is equal to the rank of $B^T A_1^T A_1 B$, which is the number of nonzero elements among $\lambda_1, \ldots, \lambda_f$. This gives us

$$S_2 = S - S_1 = \sum_{i=1}^{f} (1 - \lambda_i) Z_i^2.$$

The number of degrees of freedom of S_2 is the number of nonzero elements among $1 - \lambda_1, \ldots, 1 - \lambda_f$. We know that exactly f_1 among the λ_i are nonzero, so $f_2 = f - f_1$ are zero. For those f_2, $1 - \lambda_i$ is nonzero, so the rest must be zero, or otherwise the number of degrees of freedom of S_2 would be greater than f_2. Thus, exactly f_1 of the λ_i are 1, and the remaining f_2 are 0. This means that, for $i = 1, 2$, the sum S_i is the sum of the squares of f_i independent standard normal random variables, which proves the theorem.

4.2 One-way analysis of variance

Now we can go about evaluating the null distribution for our test of $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against the alternative that the means are different. Remember that our test statistic was

$$\frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2.}$$

First of all, it is clear that the distribution of this statistic does not depend on σ^2 and μ (which is the common value of $\mu_1 \dots \mu_k$), so we may assume that the X_{ij} are standard normal.
The sum

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2.$$

Has a χ^2 distribution with $n-1$ degrees of freedom (in fact, this is just the sum of the squares of the differences of n standard normal random variables and their average). The inner sum can be decomposed in the following way:

$$\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + n_1(\bar{X}_i - \bar{X})^2.$$

Summing up over i we get

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 + \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2.$$

The sum on the left-hand side has $n-1$ degrees of freedom. On the right-hand side, the first sum has $n-k$ degrees of freedom (summing $X_{ij} - \bar{X}_i$ over j gives zero for each i), and the second sum has $k-1$ degrees of freedom (here, $\sum n_i(\bar{X}_i - \bar{X}) = 0$ is the only linear relation). Thus, the number of degrees of freedom on the left-hand side is the sum of the numbers of degrees of freedom on the right-hand side, so we can apply the Fisher-Cochran theorem and get that

$$S_1 = \sum_{i=1}^{k}\sum_{j=0}^{n_i}(X_{ij} - \bar{X}_i)^2$$

has a χ^2-distribution with $n-k$ degrees of freedom, and

$$S_2 = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

has a χ^2 distribution with $k-1$ degrees of freedom, and S_1 and S_2 are independent. This implies that our test statistic

$$F = \frac{S_2/(k-1)}{S_1/(n-k)}$$

has an F distribution with $k-1$ and $n-k$ degrees of freedom. So, we reject H_0 if $F > F_{k-1,n-k,1-\alpha}$.
Once we have found out that there is a difference between the sample (by getting a rejection from the F-test), we might want to find a confidence interval for μ_i (or a difference $\mu_i - \mu_j$). This can be done using the standard formula for sample i, but this would be a little wasteful because it doesn't make use of all the information that is available. In fact

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=0}^{n_i}(X_{ij} - \bar{X}_i)^2$$

is a more accurate estimate of σ^2 than the sample variance of the ith sample alone. This leads us to the confidence interval

$$[\bar{X}_i - t_{n-k,\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\sigma}^2}{n_i}}, \bar{X}_i + t_{n-k,\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\sigma}^2}{n_i}}].$$

for μ_i, or, for $\mu_i - \mu_j$,

$$[\bar{X}_i - X_j - t_{n-k,\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\sigma}^2(n_i+n_j)}{n_i n_j}}, \bar{X}_i - X_j + t_{n-k,\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\sigma}^2(n_i+n_j)}{n_i n_j}}].$$

4.3 Two-way analysis of variance

In this section, the fun will be increased even more by allowing more than one influence on our data; in the simplest example, we may assume that some industrial product is worked upon by two different types of machines, A and B; the final length (or weight, or whatever) of the product is supposed to be the sum of the influences of both machines. Assume that we have machines $A_1, \ldots, A_n$ resp. $B_1, \ldots, B_k$ of each type. For testing purposes, we may produce one product with each possible combination of machines. Let X_{ij} be the size of the product produced with the combination A_i with B_j. Under our assumptions, X_{ij} will have a normal distribution with mean $\mu_i + \nu_j$ and variance σ^2 (yes, again we demand equal variances). We may want to test whether the influences of the machines of type A are all equal, for example. We shall see in a minute how this is done, but first observe that we get the same results for $\mu_i + \nu_j$ if we add a constant to every μ_i and subtract the same constant from every ν_i. We avoid the ambiguities that may arise from this by writing $\mu_{ij} = a_i + b_j + \mu$ and demanding that $\sum a_i = \sum b_j = 0$. With those conventions, we want to test

$$H_0 : a_i = 0, i = 1, \dots, n$$

against the alternative that at least one a_i is different from zero.
We could find the decomposition of the sum of squares

$$\sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X})^2$$

that fits our needs by simple heuristic arguments, yet it is desirable to have a general principle that could be used in a greater number of cases. This will be the likelihood principle under a new name. For the normal distribution, finding the maximum of the likelihood function is equivalent to minimizing the sum of squares

$$\sum (X_{ij} - \mu_{ij})^2.$$

Thus, if we suppose that μ depends on a parameter θ, then the ML estimator will be the number that minimizes the sum of squares above, hence it is called the least squares estimator (least squares estimators can be used for other distributions than the normal distribution, but they are at their best if there is a normal distribution to work with).
In order to study the general theory, let us replace the double index above by a single one, and let us assume that, for $i = 1 \dots, n$, each X_i has a normal distribution with mean μ_i and variance σ^2. Furthermore, let μ_i be a linear function of k parameters $\theta_1, \dots, \theta_k$. We can write this down in matrix notation by defining the column vectors

$$X = (X_1, \dots, X_n)^T,$$

$$\mu = (\mu_1, \dots, \mu_n)^T$$

and

$$\theta = (\theta_1, \dots, \theta_k)$$

and a $n \times k$ matrix A such that

$$\mu = A\theta.$$

We assume that $k < n$ and that A has rank k.
We wish to test the null hypothesis that some set of linear relations among the prameters pertains, i.e., there is some $r \times k$ matrix B such that we have $H_0 : B\theta = 0$. Again $r < k$, and the rank of B is supposed to be r.
We can first find the estimator for θ without the restriction $B\theta = 0$ by the usual method of zeroing the derivatives with respect to θ_i which gives us

$$\hat{\theta} = (A^T A)^{-1} A^T X.$$

Letting $\hat{\mu} = A\hat{\theta}$, and using some matrix algebra, we find that

$$X^T X = (X - \hat{\mu})^T (X - \hat{\mu}) + \hat{\mu}^T \hat{\mu}.$$

If the X_i have standard normal distributions, then the sum on the left-hand side has a χ^2-distribution with n degrees of freedom.
The first sum on the right-hand side has $n - k$ degrees of freedom because there are k linear relations between the components of $X - \hat{\mu}$ by virtue of the equation $A^T(X - \hat{\mu}) = 0$. The second sum on the right-hand side has k degrees of freedom, as this is the rank of the matrix $A(A^T A)^{-1} A^T$ that generates this sum of squares. Thus, by the Fisher-Cochran theorem, the distributions of the sums on the right-hand side are χ^2 with $n - k$ and k degrees of freedom, respectively.
If we take the additional condition $B\theta = 0$ into account, the least squares estimator can be obtained by Lagrange's multiplier method, giving an estimator $\tilde{\theta}$ and a vector of estimated means $\tilde{\mu}$. A little more matrix algebra brings us to the decomposition

$$\hat{\mu}^T \hat{\mu} = \tilde{\mu}^T \tilde{\mu} + (\hat{\mu} - \tilde{\mu})^T (\hat{\mu} - \tilde{\mu}).$$

The numbers of degrees of freedom are k for the left-hand side, $k - r$ for the first sum on the right, and r for the third sum on the right. Thus, by the Fisher-Cochran theorem, they have the corresponding χ^2 distributions.
Finally, we arrive at the following: The statistic

$$F = \frac{\sum(\hat{\mu}_i - \tilde{\mu}_i)^2 / r}{\sum(X_i - \hat{\mu}_i)^2 / (n - k)}$$

has a F distribution with r and $n - k$ degrees of freedom, where k is the number of degrees of freedom of our parameter space (=number of parameters- number of linear relations among them), and r is the number of independent equations in our null hypothesis.
In our example, there are nk Variables X_{ij}, $n + k + 1$ parameters $(\mu, a_1, \ldots, a_n, b_1, \ldots, b_k)$ and 2 linear relations ($\sum a_i = \sum b_i = 0$), giving

us $n+k-1$ degrees of freedom for the parameter space. Our null hypothesis $a_1 = \ldots = a_n = 0$ has only $n-1$ linearly independent equations, because the last one is implied by the condition $\sum a_i = 0$.
We see, that in the general analysis above, we have to replace n by nk, k by $n+k-1$, and r by $n-1$.
If we write $X_{.j}$ for $n^{-1}\sum_{i=1}^n X_{ij}$, and use the analogous definitions for $X_{i.}$ and $X_{..}$, we can write our test statistic as

$$F = \frac{\sum_{i,j}(X_{i.} - X_{..})^2/(n-1)}{\sum_{i,j}(X_{ij} - X_{i.} - X_{.j} + X_{..})^2/(n-1)(k-1)} =$$

$$\frac{(k-1)k\sum_i (X_{i.} - X_{..})^2}{\sum_{i,j}(X_{ij} - X_{i.} - X_{.j} + X_{..})^2},$$

and we reject H_0 if F exceeds $F_{n-1,(n-1)(k-1);1-\alpha}$.

4.4 Problems.

We shall overcome some day.
Pete Seeger

1. Modify the two-way test for the case that we have samples of sizes n_{ij} instead of single observations for each combination of i and j.
2. (Two-way ANOVA with interaction) In general, we cannot expect the effects of two different influences to just add up. This prompts us to consider the more general model

$$\mu_{ij} = \mu + a_i + b_j + c_{ij}$$

with the conditions

$$\sum_i a_i n_{ij} = \sum_j b_j n_{ij} = 0,$$

$$\sum_i c_{ij} n_{ij} = 0 \; (j = 1, \ldots, k),$$

$$\sum_j c_{ij} n_{ij} = 0 \; (i = 1, \ldots, n).$$

Here a_i and b_j are the usual linear influences, and the interaction terms c_{ij} represent the way the joint influence deviates from the simple summation of the individual influences.

In this case, there are too many parameters to work with single observations, so we have to use samples of size n_{ij} for each combination of i and j. Calculate a test for $H_0 : c_{ij} = 0$.

3. Show that for the comparison of two samples, ANOVA is equivalent to the t-test.

4. For four samples of iron ore from different providers, the metal contents were measured:

i	n_i	$\bar{X}_i$	S_i^2
1	10	5.3	0.11
2	12	5.1	0.13
3	10	5.0	0.12
4	11	5.4	0.10

Test for equality of the means.

5. in the previous example, calculate a confidence interval for μ_1.

6. Compare the confidence interval from the previous example with the one calculated from the first sample alone.

7. Construct a confidence interval for a_i in the two-way ANOVA.

8. In a factory, ball bearings are manufactured in two production stages. There are four different units for the first stage, and three different units for the second stage. One ball bearing for each combination is measured for its inner diameter:

	i			
j	1	2	3	4
1	10.5	10.4	10.1	9.9
2	10.3	10.3	10.2	10.0
3	10.2	10.1	10.4	10.3

Test whether there is a difference between the units in stage 1.

9. In the previous problem, assume that two bearings are made for each combination:

i	1	1	1	2	2	2	3	3	3	4	4	4
j	1	2	3	1	2	3	1	2	3	1	2	3
	10.6	10.4	10.3	10.5	10.4	10.2	10.2	10.3	10.5	10.0	10.1	10.4
	10.4	10.2	10.1	10.2	10.0	10.4	10.2	10.1	10.3	10.3	9.9	10.2

Again, test whether there is a difference between the units of stage 1.

Chapter 5

Linear Regression.

And you know something is happening
but you don't know what it is
do you, Mr, Jones?

B. Dylan

5.1 Simple linear regression.

In this chapter, we investigate simple functional relationships between two random variables. We assume that x and y are related through the equation

$$y = ax + b,$$

where a and b are real constants. If we have to find the values of a and b from measurements, and if all measuremants are exact, we can calculate a and b from only two observations. But, alas, things usually are not as perfect as they should be, so we have to account for measurement errors in some way. So, we rather write

$$y = ax + b + e,$$

where e is a (random) error (not necessarily =2.71828...). Thus, our task is to estimate a and b from n observations. We fix numbers $x_1, \ldots, x_n$ — those are supposed to be nonrandom; there is another branch of the theory that deals with random x's (called correlation theory), but we won't touch that here; a great part of regression theory, however, can be considered as correlation theory conditioned on $x_1, \ldots, x_n$.

Now, we have to make some assumptions on the errors e_i. We want those to have expectation zero (which means that there is no systematic error in the measurements) and a variance σ^2 that is the same for all observations. Then, as usual, the errors for different measurements are supposed to be independent. With these assumption alone, one can develop some part of the theory. In order to get confidence intervals, or for testing, we must make some assumptions on the distribution of e_i, and, of course, we pick the normal distribution. So, for us, e_i will have a normal distribution with mean 0 and variance σ^2. The estimators we use for a and b will be the maximum likelihood estimators which turn, as is customary for the normal distribution, into least squares estimators. The calculations yield

$$\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

and

$$\hat{b} = \bar{Y} - \hat{a}\bar{x}.$$

The expectations of $\hat{a}$ and $\hat{b}$ are of course the parameters a and b themselves. With a little work, one calculates the variances

$$\mathbf{Var}(\hat{a}) = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\sigma^2$$

and

$$\mathbf{Var}(\hat{b}) = (\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2})\sigma^2.$$

We still need an estimator for σ^2. This is achieved by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{a}x_i - \hat{b})^2.$$

By the Fisher-Cochran theorem, the sum in the above formula, divided by σ^2, has a χ^2 distribution with $n-2$ degrees of freedom, and is independent of $\hat{a}$ and $\hat{b}$. This directly leads us to the confidence intervals

$$[\hat{a} - t_{n-2,\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{a} + t_{n-2,\frac{1+\gamma}{2}}\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}]$$

for a and (in the following formulae, "t" always means "$t_{n-2;\frac{1+\gamma}{2}}$")

$$[\hat{b} - t\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2})}, \hat{b} + t\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2})}]$$

for b. In many cases, one wants a confidence interval for $y(x) = ax + b$:

$$[\hat{y}(x) - t\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum i(x_i - \bar{x})^2})}, \hat{y}(x) + t\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2})}].$$

Sometimes, one wants to find an interval that contains the value $Y(x) = ax + b + e$ with a given probability γ. This prediction interval is given by

$$[\hat{y}(x) - t\sqrt{\hat{\sigma}^2(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2})}, \hat{y}(x) + t\sqrt{\hat{\sigma}^2(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2})}].$$

5.2 Multiple linear regression.

Now, we consider functions of the form

$$y = a_1 x_1 + \ldots + a_k x_k.$$

(There is no b in this formula, but it can be treated by letting $x_i = 1$ for some i. We fix n sets of numbers $(x_{1l}, \ldots, x_{kl})$ $(l = 1, \ldots, n)$ and measure the related

$$Y_l = a_1 x_{1l} + \ldots a_k x_{kl} + e_l.$$

With the usual assumptions on our error terms e_i, we get the equations

$$\sum_{i=1}^{k} a_i b_{ij} = c_j (j = 1, \ldots, k),$$

where

$$b_{ij} = \sum_{l=1}^{n} x_{il} x_{jl}$$

and

$$c_j = \sum_{l=1}^{n} Y_l x_{jl}.$$

Again, $\hat{a}_i$ is an unbiased estimator of a_i, and the random variables $\hat{a}_i$ have a joint normal distribution with covariance matrix $\sigma^2 B^{-1}$, where B is the $(k \times k)$-matrix with entries b_{ij}.
Again, we have the estimator

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{k} \hat{a}_j x_{ji})^2$$

for $\hat{\sigma}^2$ which is independent of $\hat{a}_1, \ldots, \hat{a}_k$.
Thus, we arrive at the confidence intervals

$$[\hat{a}_i - t_{n-k,\frac{1+\gamma}{2}}\sqrt{\hat{\sigma}^2(B^{-1})_{ii}}\,,\ \hat{a}_i + t_{n-k,\frac{1+\gamma}{2}}\sqrt{\hat{\sigma}^2(B^{-1})_{ii}}].$$

For

$$y(x) = a_1x_1 + \ldots + a_kx_k$$

we get the estimator

$$\hat{y}(x) = \hat{a}_1x_1 + \ldots + \hat{a}_kx_k$$

and the confidence interval

$$[\hat{y}(x) - t_{n-k,\frac{1+\gamma}{2}}\sqrt{\hat{\sigma}^2x^TB^{-1}x}\,,\ \hat{y}(x) + t_{n-k,\frac{1+\gamma}{2}}\sqrt{\hat{\sigma}^2x^TB^{-1}x}],$$

where x denotes the column vector with entries x_i.
We also have a prediction interval:

$$[\hat{y}(x) - t_{n-k,\frac{1+\gamma}{2}}\sqrt{\hat{\sigma}^2(1+x^TB^{-1}x)}\,,\ \hat{y}(x) + t_{n-k,\frac{1+\gamma}{2}}\sqrt{\hat{\sigma}^2(1+x^TB^{-1}x)}],$$

5.3 Other functional relations.

In practice, there is often the question of fitting a nonlinear function to our data. Especially important are the exponential and power functions ($y = AB^x$ or $y = Ax^B$), and the polynomials ($y = a_0 + a_1x + \ldots + a_kx^k$). The polynomials can easily be treated using the methods of the previous section, by letting $x_{il} = x_l^{i-1}$. The other two types of functions present more problems; actually, we can write down the equations for the least squares estimators, but those cannot be solved in closed form (they can be solved numerically, of course); so we rather take logarithms on both sides to convert this into simple linear regression. We have to be aware, however, that by doing so we have changed our stochastic model. Now the error term does no longer come in as an additive term, but it is added in the exponent. Anyway, this enables us to calculate confidence intervals and prediction intervals, which are not easily obtained by other methods.

5.4 Problems

All together now

The Beatles

1. Suppose that n is even and that we are free to choose $x_1, \dots, x_n$ form the interval $[-A, A]$. How do we have to choose $x_1, \dots, x_n$ in order to estimate

   (a) a

   (b) b

   as accurately as possible?

2. As before, for n odd.

3. Sometimes the variance of the error term is a function $\sigma^2(x)$ of x. Show that in this case the ML estimators for a and b can be obtained by using ordinary linear regression for

$$\frac{y}{\sigma(x)} = a\frac{x}{\sigma(x)} + b\frac{1}{\sigma(x)}.$$

4. For the following data

i	1	2	3	4	5
x	0.5	1.0	2.0	2.5	3.0
Y	1.3	2.7	1.5	2.0	2.2

   fit a linear regression function and calculate a prediction interval for $x = efregrdata.5$.

5. In problem 4 , use a regression function of the form $y = ae^{bx}$.

6. In problem 4 , use a regression function of the form $y = \frac{a}{b+x}$.

7. In problem 4 , use a quadratic polynomial as the regression function.

8. In problem 4 , calculate a confidence interval for a.

9. In problem 4 , calculate a confidence interval for b.

10. In problem 4 , calculate a confidence interval for $y(1.5)$.

Chapter 6

The χ^2-Family of Tests.

Whereupon the Plumber said in tones of disgust:
"I suggest that we proceed at once to infinity.

J.L. Synge

6.1 The multinomial distribution and the χ^2-statistic

In many cases we have to deal with testing wether a certain discrete random variable has a given distribution.
In the simplest setting, assume that X can take k values — without restricting generality, we may assume that these are $1, \dots, k$ — with probabilities $p_1, \dots, p_k$, and we want to test $H_0 : p_i = p_{i0}$ $(i = 1, \dots, k)$ against the alternative that at least one p_i is different from p_{i0}. The standard way to approach this problem is to use the likelihood ratio test, of course, and the likelihood ratio statistic turns out to be

$$\ell = \prod_{i=1}^{n} \frac{p_{i0}^{Y_i}}{(Y_i/n)^{Y_i}},$$

where Y_i is the number of occurrences of i among $X_1, \dots, X_n$.
In order to calculate the critical value for this test, we need the distribution of $Y_1, \dots, Y_k$ under the null hypothesis. This is a multinomial

distribution:

$$\mathbf{P}(Y_1 = y_1, \ldots, Y_k = y_k) = \begin{cases} \frac{n!}{y_1! \ldots y_k!} p_{1,0}^{y_1} \cdots p_{k0}^{y_k} & \text{if } y_1 + \ldots + y_k = n, \\ 0 & \text{otherwise.} \end{cases}$$

For the sake of simplicity, we drop the subscript "0" from p_{i0} in the following calculations.
Now, at least in theory, we are able to use the multinomial distribution to evaluate the null distribution of ℓ. In general, however, this is too tedious to be carried out in practice. So, we rather assume that n is large and try to use some approximation. We put

$$Y_i = np_i + U_i\sqrt{n},$$

where U_i has mean zero and variance $p_i(1 - p_i)$ and is approximately normal. Furthermore, using the expansion

$$\log(1 + x) = x - \frac{x^2}{2} + o(x^2)$$

we get

$$\log \ell = -\frac{1}{2}\sum_{i=1}^{k} \frac{U_i^2}{p_i} + o(1) = -\frac{1}{2}\sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} + o(1).$$

Thus, for large n, the likelihood ratio test is approximately equivalent to the test that rejects H_0 if

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$$

is large.
What we need now is a result on the distribution of χ^2. First, let us consider the joint distribution of $U_1, \ldots, U_k$. Approximately, they have normal distributions with mean zero; the variances and covariances are

$$\mathbf{Var}(U_i) = p_i(1 - p_i)$$

and

$$\mathbf{Cov}(U_i, U_j) = -p_i p_j \ (i \neq j).$$

Let U_0 be another normal random variable with mean zero and variance one, independent from $U_1, \ldots, U_k$, and let

$$V_i = U_i + p_i U_0.$$

It is readily verified that

$$\mathbf{Var}(V_i) = p_i$$

and

$$\mathbf{Cov}(V_i, V_j) = 0.$$

So (at least as an approximation), the random variables V_i are independent normal random variables. Furthermore, we have the decomposition

$$\sum_{i=1}^{k} \frac{V_i^2}{p_i} = U_0^2 + \sum_{i=1}^{k} \frac{U_i^2}{p_i},$$

(In the last equation, we use the fact that $\sum U_i = 0$) where the left-hand side has k degrees of freedom, and the two summands on the right-hand side have one and $k-1$ degrees of freedom, respectively. This, by virtue of the Fisher-Cochran theorem, permits us to state that

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$$

has an approximate χ^2-distribution with $k-1$ degrees of freedom.
Now, let us turn to the case where H_0 is composite. Assume that H_0 states that $p_1, \ldots, p_k$ is of the form $p_i = p_i(\theta_1, \ldots, \theta_s)$, where $\theta_1, \ldots, \theta_s$ are real parameters. It can be shown, using Taylor expansions and the ideas from analysis of variance, that, if we use the ML estimators $\hat{\theta}_i$,

$$\chi^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i(\hat{\theta}_1, \ldots, \hat{\theta}_s))^2}{np_i(\hat{\theta}_1, \ldots, \hat{\theta}_s)}$$

has an approximate χ^2-distribution with $k-1-s$ degrees of freedom (assuming that the parameters are all independent; otherwise we have to use the number of degrees of freedom of the set of parameters instead, which again is the number of parameters that can be independently chosen). Be advised, however, that this result depends on some regularity conditions, similar to the ones in the Cramér-Rao theorem (which in our case amount to saying that p_i is sufficiently often differentiable with respect to θ). These conditions may be violated even in simple cases (the uniform distribution is one example). Anyway, it is very useful and very widely used, even in cases where it is not completely justified.
As the χ^2-distribution is only an asymptotic one, we need some rule for its application. The usual rule of thumb is that np_i should be at least 5 for each i.

6.2 The χ^2 goodness-of-fit test.

There is not much new in this section. We now want to test $H_0 : X$ has a given distribution. If this distribution is discrete with a finite number k of possible values, we can directly use the first result from the previous section. This means, we calculate

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

and reject if this exceeds $\chi^2_{k-1,1-\alpha}$. If the distribution has infinitely many possible values (in particular if it is continuous), or if some of the numbers np_i are less than 5, we can divide the set of possible values into classes, and let Y_i be the number of occurrences of values from class number i among $X_1, \ldots, X_n$.
In some cases, the distribution of X is not completely specified (the most common example is the normal distribution — often one wants to know whether the data are normally distributed, but doesn't know mean and variance). For that case, too, the previous section shows the way: we estimate the parameters by the maximum likelihood method (although this is not exactly what was done in the previous section) and use these to calculate the probabilities for each class (to me, it seems convenient to choose the classes in such a way that the probabilities p_i are all equal; in the case of the normal distribution, this means that the class boundaries are $\bar{X} + z_{i/k}\sqrt{S_n^2}$, $(i = 1, \ldots, k)$.) Finally, we calculate the χ^2 statistic in the usual way, and use $k - 1 -$ number of estimated parameters as the number of degrees of freedom.

6.3 The χ^2 test of independence.

In this section, we look at paired observations (or two-dimensional ones, if you prefer) $(X_i, Y_i), i = 1, \ldots, n$ and we want to test whether X and Y are independent. We assume that both X and Y have a discrete distribution (otherwise, we have to use a suitable classification); without restricting generality, we may assume that X takes values 1 to k, and that Y takes values 1 to m. If X and Y are in fact independent, then for $1 \leq j \leq k$ and $1 \leq l \leq m$, letting

$$p_j = \mathbf{P}(X = j)$$

and

$$q_l = \mathbf{P}(Y = l)$$

we have

$$\mathbf{P}(X = j, Y = l) = \mathbf{P}(X = j)\mathbf{P}(Y = l) = p_j q_l.$$

Thus, we have expressed the joint distribution of X and Y by the $k+m$ parameters $p_1, \ldots, p_k$ and $q_1, \ldots, q_m$. This almost cries out for an application of the χ^2 test. If we let

$$Z_{jl} = \#\{i : X_i = j, Y_i = l\},$$

$$Z_{.l} = \#\{i : Y_i = l\} = \sum_{j=1}^{k} Z_{jl},$$

and

$$Z_{j.} = \#\{i : X_i = j\} = \sum_{l=1}^{m} Z_{jl},$$

then the ML estimators for p_j and q_l are $Z_{j.}/n$ and $Z_{.l}/n$, respectively. The χ^2 statistic is

$$\sum_{j=1}^{k}\sum_{l=1}^{m} \frac{(Z_{jl} - Z_{j.}Z_{.l}/n)^2}{Z_{j.}Z_{.l}/n}.$$

We have to find the number of degrees of freedom of this. The number of classes, of course, is km. We had to estimate the $k+m$ parameters, but only $k+m-2$ are independent because we have the relations $\sum p_j = \sum q_l = 1$. Thus, the number of degrees of freedom of our χ^2 statistic is

$$km - 1 - (k + m - 2) = (k - 1)(m - 1).$$

6.4 The χ^2 test for homogeneity.

This time, we have k samples $(X_{i1}, \ldots, X_{in_i}), i = 1, \ldots, k$. We assume that X_{ij} can take values from 1 to m and want to check if the underlying distributions are the same for all samples. We let

$$n = \sum_{i=1}^{k} n_i,$$

$$Y_{il} = \#\{j : X_{ij} = l\},$$

$$Z_l = \#\{(i, j) : X_{ij} = l\}.$$

under the null hypothesis that all samples have the same distribution $\mathbf{P}(X_{ij} = l) = p_l$, the ML estimator for p_l is Z_l/n, and we arrive at the χ^2 statistic

$$\sum_{i,l} \frac{(Y_{il} - n_i Z_l/n)^2}{n_i Z_l/n},$$

which has

$$(m-1)(k-1)$$

degrees of freedom. (It may be noted that this is very similar to the test for independence; in fact, it is only a different way of interpreting the same data)

6.5 Problems

All my trials, Lord
soon be over

trad.

1. A die is rolled 600 times showing the following results:

1	2	3	4	5	6
95	111	89	106	88	113

   Test whether the die is fair.

2. Test whether the following numbers have a normal distribution with mean 1 and variance 0.5:

   1.95 2.13 0.74 0.05 2.70 1.84 1.25 1.00 0.70 1.02

   1.71 0.58 0.16 1.12 1.16 0.21 0.70 1.30 1.06 1.75

3. Do the previous problem for unknown mean and variance.

4. Test whether the data from problem 2 have an exponential distribution.

5. Test whether the data from problem 2 have a uniform distribution on $[0, \theta]$.

6. Test whether the following numbers have a Poisson distribution:

   0 1 2 1 3 4 3 3 2 4 3 0 3 5 3 2 3 1 2 4 3 3 3 5 2 4 1 0 3 2 2 3 2 3 4 5 6 1 0 6

7. In the previous problem, test whether the data have a binomial distribution with $n = 6$ and $p = 1/2$.

8. Do the previous problem for unknown p.

9. If $X_1, \ldots, X_n$ are uniformly distributed on $[0, \theta]$, let $Y_1, \ldots, Y_{n-1}$ be $X_1, \ldots, X_n$ without the maximum $M = \max_{i \le n} X_i$. Furthermore, let $Z_i = Y_i/M$. Show that $Z_1, \ldots, Z_{n-1}$ is a sample from a uniform distribution on $[0, 1]$. Use this fact to show that the χ^2-statistic in problem 5 has an approximate χ^2-distribution with $k-1$ (and not $k-2$!) degrees of freedom. (this is, of course, proof of the fact that the uniform distribution is just not regular enough).

10. A sample of 500 students showed the following marks in mathematics and english:

	mathematics				
english	1	2	3	4	5
1	40	10	0	0	0
2	25	30	15	20	10
3	40	35	45	30	50
4	20	10	20	30	20
5	0	0	10	20	20

Are the marks in the two subjects independent?

11. For $m = k = 2$, give a simple (?) formula for the χ^2 statistic in the test for independence.

12. 460 bolts are classified as "conforming" or "non-conforming" with respect to their measurements and their breaking strength:

	breaking strength	
meas.	conf.	non-c.
conf.	416	23
non-c.	16	5

Are the two classifications independent?

13. Three professors evaluate the same 100 tests with the following results

mark	prof. 1	prof. 2	prof. 3
1	35	30	28
2	22	24	20
3	16	20	22
4	12	20	17
5	15	6	13

Are the three professors equally strict?

Chapter 7

The Kolmogorov-Smirnov Test

a discussion of order ... has become essential to any inderstanding of the foundation of mathematics

B.Russell

7.1 The one-sample test

This is another goodness of fit test, i.e., we want to test H_0 : the distribution function of X id F against the alternative that it is different from F. Our point of departure is the Glivenko-Cantelli theorem which states that the empirical distribution function

$$F_n(x) = \frac{1}{n}\#\{j : X_j < x\}$$

converges uniformly to the actual distribution function of X. Thus, we may choose

$$D_n = \|F_n - F\| = \sup_x |F_n(x) - F(x)|$$

as our test statistic. H_0 will be rejected if D_n is greater than a critical value. The following theorem shows that — at least for continuous F — this critical value can be determined independently from F.

Theorem 8 *If the distribution function F is continuous, then the null distribution of D_n does not depend on F.*

For the proof of this theorem, we need the following

Lemma 1 1. *If the distribution function F of X is continuous, then $F(X)$ is uniformly distributed on* $[0,1]$.

2. *For any distribution function F and U uniform on* $[0,1]$, *$X = F^{-1}(U)$ has distribution function F, where*

$$F^{-1}(x) = \sup\{t : F(t) < x\}.$$

We don't need the second assertion, so we only prove the first one.
By the intermediate value theorem, for any $0 < y < 1$, there is an x such that $F(x) = y$. Thus

$$\mathbf{P}(F(X) \leq y) \geq \mathbf{P}(X \leq x) = F(x) = y$$

and

$$\mathbf{P}(F(X) < y) \leq \mathbf{P}(X < x) = F(x - 0) = y$$

This shows that $F(X)$ is uniformly distributed on $[0,1]$.
Now, let us turn to the proof of the theorem. We first calculate a formula for D_n which is a little more convenient. To this end, let

$$X_{n:1} < X_{n:2} < \ldots < X_{n:n}$$

be the order statistics (i.e., the sample $X_1, \ldots, X_n$ in ascending order). Then for $0 \leq k \leq n$ $F_n(x) = \frac{k}{n}$ if $X_{n:k} < x < X_{n:k+1}$ (let $X_{n:0} = -\infty$, $X_{n:n+1} = +\infty$). Thus,

$$D_n = \max_k \sup_{X_{n:k} < x < X_{n:k+1}} |F_n(x) - F(x)| =$$

$$\max_k \sup_{X_{n:k} < x < X_{n:k+1}} |\frac{k}{n} - F(x)| =$$

$$\max_k \max(|\frac{k}{n} - F(X_{n:k})|, |\frac{k}{n} - F(X_{n:k+1})|) =$$

$$\max_k \max(|\frac{k}{n} - F(X_{n:k})|, |\frac{k-1}{n} - F(X_{n:k})|).$$

If we let $U_i = F(X_i)$, then $U_1, \ldots, U_n$ will be a sample from a uniform distribution on $[0,1]$, and $U_{n:k} = F_(X_n : k)$. Thus, D_n is the same whether we calculate it from X or from U. This proves the theorem.
Thus, one can calculate the finite sample distribution of D_n, although the calculations may get quite intricate. For large n, there is an asymptotic result:

Theorem 9 *Let $\lambda_n = \sqrt{n} D_n$.. For $n \to \infty$, the distribution function of λ_n is approximately*

$$K(x) = \begin{cases} \sum_{k=-\infty}^{\infty} (-1)^n e^{-2n^2x^2} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}.$$

The proof of this theorem is beyond the scope of this lecture.

7.2 The Two-Sample Test

Now, we are going to compare two sample $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$. We want to test the hypothesis that X and Y have the same distribution. Using the empirical distribution functions F_n and G_m, we define our test statistic as

$$D_{n,m} = \|F_n - G_m\|.$$

If we assume that the common distribution function F of X and Y is continuous, one can prove in the same way as before that the distribution of $D_{n,m}$ does not depend on F. In addition, there is the asymptotic result:

Theorem 10

$$\lambda_{n,m} = \sqrt{\frac{nm}{n+m}} D_{n,m}$$

has a limiting distribution with distribution function K.

7.3 Problems

To everything - turn, turn, turn,
there is a season - turn, turn, turn,
and a time for ev'ry purpose under heaven.
Book of ecclesiastes, adap. P. Seeger

1. Use the Kolmogorov-Smirnov test for problem 2, chapter 6.

2. Test whether the following data are uniformly distributed on [0, 1].

   0.23 0.31 0.45 0.67 0.87 0.49 0.93 0.80

3. Use the two-sample version of the Kolmogorov-Smirnov test for problem 20, chapter 3.

4. Show that

$$\mathbf{P}(D_n < x) = \mathbf{P}(\frac{k}{n} - x < U_{n:k} < \frac{k-1}{n} + x) =$$

$$= n!\mathbf{P}(\frac{k}{n} - x < U_k < \frac{k-1}{n} + x, U_1 < U_2 < \ldots < U_n).$$

5. Use the previous problem to calculate the distribution of D_2.

6. If both X_i and Y_j are uniformly distributed on $[0,1]$, calculate the Variance of $\eta_{m,n}(x) = F_n(x) - G_m(x)$ and the covariance of $\eta_{m,n}(x)$ and $\eta_{m,n}(y)$ (This should make it plausible that the normalizing factor in our limit theorems is right).

Appendix A

Tables

The distribution function of the normal distribution:

$$\Phi(x) = \int_{-\infty}^{x} e^{-\frac{x^2}{2}}$$

	0	1	2	3	4	5	6	7	8	9
0.0	.500	.504	.508	.512	.516	.520	.524	.528	.532	.536
0.1	.540	.544	.548	.552	.556	.560	.564	.567	.571	.575
0.2	.579	.583	.587	.591	.595	.599	.603	.606	.610	.614
0.3	.618	.622	.626	.629	.633	.637	.641	.644	.648	.652
0.4	.655	.659	.663	.666	.670	.674	.677	.681	.684	.688
0.5	.691	.695	.698	.702	.705	.709	.712	.716	.719	.722
0.6	.726	.729	.732	.736	.739	.742	.745	.749	.752	.755
0.7	.758	.761	.764	.767	.770	.773	.776	.779	.782	.785
0.8	.788	.791	.794	.797	.800	.802	.805	.808	.811	.813
0.9	.816	.819	.821	.824	.826	.829	.831	.834	.836	.839
1.0	.841	.844	.846	.848	.851	.853	.855	.858	.860	.862
1.1	.864	.867	.869	.871	.873	.875	.877	.879	.881	.883
1.2	.885	.887	.889	.891	.893	.894	.896	.898	.900	.901
1.3	.903	.905	.907	.908	.910	.911	.913	.915	.916	.918
1.4	.919	.921	.922	.924	.925	.926	.928	.929	.931	.932
1.5	.933	.934	.936	.937	.938	.939	.941	.942	.943	.944
1.6	.945	.946	.947	.948	.949	.951	.952	.953	.954	.954
1.7	.955	.956	.957	.958	.959	.960	.961	.962	.962	.963
1.8	.964	.965	.966	.966	.967	.968	.969	.969	.970	.971
1.9	.971	.972	.973	.973	.974	.974	.975	.976	.976	.977
2.0	.977	.978	.978	.979	.979	.980	.980	.981	.981	.982
2.1	.982	.983	.983	.983	.984	.984	.985	.985	.985	.986
2.2	.986	.986	.987	.987	.987	.988	.988	.988	.989	.989
2.3	.989	.990	.990	.990	.990	.991	.991	.991	.991	.992
2.4	.992	.992	.992	.992	.993	.993	.993	.993	.993	.994
2.5	.994	.994	.994	.994	.994	.995	.995	.995	.995	.995
2.6	.995	.995	.996	.996	.996	.996	.996	.996	.996	.996
2.7	.997	.997	.997	.997	.997	.997	.997	.997	.997	.997
2.8	.997	.998	.998	.998	.998	.998	.998	.998	.998	.998
2.9	.998	.998	.998	.998	.998	.998	.998	.999	.999	.999

Quantiles z_p of the normal distribution:

p	z_p	p	z_p	p	z_p
.51	.025	.71	.553	.91	1.341
.52	.050	.72	.583	.92	1.405
.53	.075	.73	.613	.93	1.476
.54	.100	.74	.643	.94	1.555
.55	.126	.75	.674	.95	1.645
.56	.151	.76	.706	.96	1.751
.57	.176	.77	.739	.97	1.881
.58	.202	.78	.772	.975	1.960
.59	.228	.79	.806	.98	2.054
.60	.253	.80	.842	.99	2.326
.61	.279	.81	.878	.991	2.366
.62	.305	.82	.915	.992	2.409
.63	.332	.83	.954	.993	2.457
.64	.358	.84	.994	.994	2.512
.65	.385	.85	1.036	.995	2.576
.66	.412	.86	1.080	.996	2.652
.67	.440	.87	1.126	.997	2.748
.68	.468	.88	1.175	.998	2.878
.69	.496	.89	1.227	.999	3.090
.70	.524	.90	1.282	.9999	3.719

Quantiles $t_{n;p}$ of Student's t-distribution with n degrees of freedom:

n	.9	.95	.975	.99	.995	n	.9	.95	.975	.99	.995
1	3.078	6.314	12.706	31.821	63.675	26	1.316	1.706	2.056	2.479	2.779
2	1.886	2.920	4.303	6.965	9.725	27	1.314	1.703	2.052	2.473	2.467
3	1.638	2.353	3.183	4.541	5.841	28	1.313	1.701	2.048	2.467	2.763
4	1.533	2.132	2.776	3.747	4.604	29	1.311	1.699	2.045	2.462	2.756
5	1.476	2.015	2.571	3.365	4.032	30	1.310	1.697	2.042	2.457	2.750
6	1.440	1.943	2.447	3.143	3.707	31	1.309	1.696	2.040	2.453	2.744
7	1.415	1.895	2.365	2.998	3.499	32	1.309	1.694	2.037	2.449	2.738
8	1.397	1.860	2.306	2.896	3.355	33	1.308	1.692	2.035	2.445	2.733
9	1.383	1.833	2.262	2.821	3.250	34	1.307	1.691	2.032	2.441	2.728
10	1.372	1.812	2.228	2.764	3.169	35	1.306	1.690	2.030	2.438	2.724
11	1.363	1.796	2.201	2.718	3.106	40	1.303	1.684	2.021	2.423	2.704
12	1.356	1.782	2.179	2.681	3.055	45	1.301	1.679	2.014	2.412	2.690
13	1.350	1.771	2.160	2.650	3.012	50	1.299	1.676	2.009	2.403	2.678
14	1.345	1.761	2.145	2.624	2.977	55	1.297	1.673	2.004	2.396	2.668
15	1.341	1.753	2.131	2.602	2.947	60	1.296	1.671	2.000	2.390	2.660
16	1.337	1.746	2.120	2.583	2.921	65	1.295	1.669	1.997	2.385	2.654
17	1.333	1.740	2.110	2.567	2.898	70	1.294	1.667	1.994	2.381	2.648
18	1.330	1.734	2.101	2.552	2.878	75	1.293	1.665	1.992	2.377	2.643
19	1.328	1.729	2.093	2.539	2.861	80	1.292	1.664	1.990	2.374	2.639
20	1.325	1.725	2.086	2.528	2.845	85	1.292	1.663	1.988	2.371	2.635
21	1.323	1.721	2.080	2.518	2.831	90	1.291	1.662	1.987	2.368	2.632
22	1.321	1.717	2.074	2.508	2.819	95	1.291	1.661	1.985	2.366	2.629
23	1.319	1.714	2.069	2.500	2.807	100	1.290	1.660	1.984	2.364	2.626
24	1.318	1.711	2.064	2.492	2.797	105	1.290	1.659	1.983	2.362	2.623
25	1.316	1.708	2.060	2.485	2.787	∞	1.282	1.645	1.960	2.326	2.576

Quantiles $\chi^2_{n;p}$ of the χ^2-distribution with n degrees of freedom

n	.005	.01	.02	.025	.05	.1	.5	.9	.95	.975	.98	.99	.995
1	.000	.000	.001	.001	.004	.016	.455	2.706	3.841	5.024	5.412	6.635	7.879
2	.010	.020	.040	.051	.103	.211	1.386	4.605	5.991	7.378	7.824	9.210	10.597
3	.072	.115	.185	.216	.352	.584	2.366	6.251	7.815	9.348	9.837	11.345	12.838
4	.207	.297	.429	.484	.711	1.064	3.357	7.779	9.488	11.143	11.668	13.277	14.860
5	.412	.554	.752	.831	1.145	1.610	4.351	9.236	11.070	12.832	13.308	15.086	16.750
6	.676	.872	1.134	1.237	1.635	2.204	5.348	10.645	12.592	14.449	15.033	16.812	18.548
7	.989	1.239	1.564	1.690	2.167	2.833	6.346	12.017	14.067	16.013	16.622	18.475	20.278
8	1.344	1.646	2.032	2.180	2.733	3.490	7.344	13.362	15.507	17.535	18.168	20.090	21.955
9	1.735	2.088	2.532	2.700	3.325	4.168	8.343	14.684	16.919	19.023	19.679	21.666	23.589
10	2.156	2.558	3.059	3.247	3.940	4.865	9.342	15.987	18.307	20.483	21.161	23.209	25.188
11	2.603	3.053	3.609	3.816	4.575	5.578	10.341	17.275	19.675	21.920	22.618	24.725	26.757
12	3.074	3.571	4.178	4.404	5.226	6.304	11.340	18.549	21.026	23.336	24.054	26.217	28.300
13	3.565	4.107	4.765	5.009	5.892	7.042	12.340	19.812	22.362	24.736	25.472	27.688	29.819
14	4.075	4.660	5.368	5.629	6.571	7.790	13.339	21.064	23.685	26.119	26.873	29.141	31.319
15	4.601	5.229	5.985	6.262	7.261	8.547	14.339	22.307	24.996	27.488	28.259	30.578	32.801
16	5.142	5.812	6.614	6.908	7.962	9.312	15.338	23.542	26.269	28.845	29.633	32.000	34.267
17	5.697	6.408	7.255	7.564	8.672	10.085	16.338	24.769	27.587	30.191	30.995	33.409	35.718
18	6.265	7.015	7.906	8.231	9,390	10.835	17.338	25.909	28.869	31.526	32.346	34.805	37.156
19	6.844	7.633	8.567	8.907	10.117	11.651	18.338	27.204	30.144	32.852	33.687	36.191	38.582
20	7.434	8.260	9.237	9.591	10.851	12.443	19.337	28.412	31.410	34.170	35.020	37.566	39.997
21	8.034	8.897	9.915	10.283	11.591	13.240	20.337	29.615	32.671	35.479	36.343	38.932	41.401
22	8.643	9.542	10.600	10.982	12.338	14.041	21.337	30.813	33.924	36.781	37.659	40.289	42.796
23	9.260	10.196	11.293	11.689	13.091	14.848	22.337	32.007	35.172	38.076	38.968	41.638	44.181
24	9.886	10.856	11.992	12.401	13.848	15.659	23.337	33.196	36.415	39.364	40.270	42.980	45.559
25	10.520	11.524	12.697	13.120	14.611	16.473	24.337	34.382	37.652	40.646	41.566	44.324	46.928
26	11.160	12.198	13.409	13.844	15.379	17.292	25.336	35.563	38.885	41.923	42.856	45.642	48.290
27	11.808	12.879	14.125	14.573	16.151	18.114	26.336	36.741	40.113	43.194	44.140	46.963	49.645
28	12.461	13.565	14.847	15.308	16.928	18.939	27.336	37.916	41.337	44.461	45.419	48.278	50.993
29	13.121	14.256	15.574	16.047	17.708	19.768	28.336	39.087	42.557	45.722	46.693	49.588	52.336
30	13.787	14.953	16.306	16.791	18.493	20.599	29.336	40.256	43.773	46.979	47.962	50.892	53.672
31	14.458	15.655	17.042	17.539	19.281	21.434	30.336	41.422	44.985	48.232	49.226	52.191	55.003
32	15.134	16.362	17.783	18.291	20.072	22.271	31.336	42.585	46.194	49.480	50.487	53.486	56.328
33	15.815	17.074	18.527	19.047	20.867	23.110	32.336	43.745	47.400	50.725	51.743	54.776	57.648
34	16.501	17.789	19.275	19.806	21.664	23.952	33.336	44.903	48.602	51.966	52.995	56.061	58.964
35	17.192	18.509	20.027	20.569	22.465	24.797	34.336	46.059	49.802	53.203	54.244	57.342	60.275
40	20.707	22.164	23.838	24.433	26.509	29.051	39.335	51.805	55.758	59.342	60.436	63.691	66.766
45	24.311	25.901	27.720	28.366	30.612	33.350	44.335	57.505	61.656	65.410	66.555	69.957	73.166
50	27.991	29.707	31.664	32.357	34.764	37.689	49.335	63.167	67.505	71.420	72.613	76.154	79.490
55	31.735	33.570	35.659	36.398	38.958	42.060	54.335	68.796	73.311	77.380	78.619	82.292	85.749
60	35.534	37.485	39.699	40.482	43.188	46.459	59.335	74.397	79.082	83.298	84.580	88.397	91.952
65	39.383	41.444	43.779	44.603	47.450	50.883	64.335	79.973	84.821	89.177	90.501	94.422	98.105
70	43.275	45.442	47.893	48.758	51.739	55.329	69.334	85.527	90.531	95.023	96.388	100.425	104.215
75	47.206	49.475	52.039	52.942	56.054	59.795	74.334	91.061	96.217	100.839	102.243	106.393	110.286
80	51.172	53.540	56.213	57.153	60.391	64.278	79.334	96.578	101.879	106.629	108.069	112.329	116.321
85	55.170	57.634	60.412	61.389	64.749	68.777	84.334	102.079	107.522	112.393	113.871	118.236	122.325
90	59.196	61.754	64.635	65.647	69.126	73.291	89.334	107.565	113.145	118.136	119.649	124.116	128.299
95	63.250	65.898	68.879	69.925	73.520	77.818	94.334	113.038	118.752	123.858	125.405	129.973	134.247
100	67.328	70.065	73.142	74.222	77.929	82.358	99.334	118.498	124.342	129.561	131.142	135.806	140.169

Quantiles $F_{a,b;0.95}$ of the F-distribution with a and b degrees of freedom
Remember that $F_{a,b,1-p} = 1/F_{b,a,p}$.

b \ a	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	∞
1	161	200	216	225	230	234	237	239	241	242	246	248	250	251	252	253	254
2	18.51	19.00	19.20	19.20	19.30	19.30	19.40	19.40	19.40	19.40	19.40	19.40	19.50	19.50	19.50	19.50	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.69	8.66	8.62	8.59	8.58	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.72	5.70	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.46	4.44	4.41	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.77	3.75	3.71	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.38	3.34	3.32	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.04	3.02	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.83	2.80	2.76	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.66	2.64	2.59	2.54
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.43	2.40	2.35	2.30
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.27	2.24	2.19	2.13
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.15	2.12	2.07	2.01
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.06	2.04	1.98	1.92
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.99	1.97	1.91	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.92	1.87	1.84	1.78	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.79	1.76	1.70	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.69	1.66	1.59	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.63	1.60	1.52	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.59	1.56	1.48	1.39
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.81	1.72	1.62	1.57	1.53	1.45	1.35
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95	1.79	1.70	1.60	1.54	1.51	1.43	1.32
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94	1.78	1.69	1.59	1.53	1.49	1.41	1.30
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.68	1.57	1.52	1.48	1.39	1.28
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62	1.52	1.46	1.41	1.32	1.19

Quantiles $F_{a,b;0.99}$ of the F-distribution with a and b degrees of freedom

b \ a	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	100	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6157	6209	6261	6287	6300	6330	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	34.12	30.80	29.50	28.70	28.20	27.90	27.70	27.50	27.30	27.20	26.90	26.70	26.50	26.40	26.40	26.20	26.10
4	21.20	18.00	16.70	16.00	15.50	15.20	15.00	14.80	14.70	14.50	14.20	14.00	13.80	13.70	13.70	13.60	13.50
5	16.26	13.30	12.10	11.40	11.00	10.70	10.50	10.30	10.20	10.10	9.72	9.55	9.38	9.29	9.24	9.13	9.02
6	13.75	10.90	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.14	7.09	6.99	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.91	5.86	5.75	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.12	5.07	4.96	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.57	4.52	4.42	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.17	4.12	4.01	3.91
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.62	3.57	3.47	3.36
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.27	3.22	3.11	3.00
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10	3.02	2.97	2.86	2.75
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92	2.84	2.78	2.68	2.57
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.69	2.64	2.54	2.42
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.54	2.45	2.40	2.29	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.30	2.25	2.13	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.11	2.06	1.94	1.80
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.79	2.70	2.42	2.27	2.10	2.01	1.95	1.82	1.68
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.94	1.88	1.75	1.60
70	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.78	2.67	2.59	2.31	2.15	1.98	1.89	1.83	1.70	1.54
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55	2.27	2.12	1.94	1.85	1.79	1.66	1.49
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61	2.52	2.24	2.09	1.92	1.82	1.76	1.62	1.46
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89	1.80	1.73	1.60	1.43
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.13	2.97	1.79	1.69	1.63	1.48	1.28

Critical values $D_{n;p}$ of the Kolmogorov-Smirnov one-sample test.

n	0.95	0.99	n	0.95	0.99
			21	0.2827	0.3443
8	0.4543	0.5419	22	0.2809	0.3367
9	0.4300	0.5133	23	0.2749	0.3295
10	0.4093	0.4889	24	0.2693	0.3229
11	0.3912	0.4677	25	0.2640	0.3166
12	0.3754	0.4491	26	0.2591	0.3106
13	0.3614	0.4325	27	0.2544	0.3050
14	0.3489	0.4176	28	0.2499	0.2997
15	0.3376	0.4042	29	0.2457	0.2947
16	0.3273	0.3920	30	0.2417	0.2899
17	0.3180	0.3809	35	0.2243	0.2690
18	0.3094	0.3706	40	0.2101	0.2521
19	0.3014	0.3612	45	0.1984	0.2380
20	0.2941	0.3524	50	0.1884	0.2260

For $n > 50$ use the approximation $D_{n;p} \approx K_p/\sqrt{n}$.

For $p > 1/2$, $K_p \approx \sqrt{\log(\frac{2}{1-p})/2}$. In particular

$$K_{0.95} = 1.358, \quad K_{0.99} = 1.628.$$

Appendix B

Solutions of Problems

B.1 Problems from Chapter 1

1. $\mathbf{E}\bar{X}_n = \frac{1}{n}\sum \mathbf{E}X_i = \mu$,

$$\mathbf{Var}\bar{X}_n = (\frac{1}{n})^2 \sum \mathbf{Var}X_i = \frac{\sigma^2}{n}.$$

(As drawing is done with replacement, $X_1, \ldots, X_n$ are independent.)

2.

$$\mathbf{E}(S_n^2) = \frac{1}{n-1}\mathbf{E}(\sum_{i=1}^{n}(X_i - \bar{X})^2) = \frac{1}{n-1}\mathbf{E}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2) =$$

$$\frac{1}{n-1}(n\mathbf{E}(X_1^2) - n\mathbf{E}(\bar{X}^2)) = \frac{1}{n-1}(n((\mathbf{E}X_1))^2 + \mathbf{Var}(X_1)) - n((\mathbf{E}\bar{X})^2 + \mathbf{Var}(\bar{X}))) =$$

$$\frac{n}{n-1}(\mathbf{Var}(X_1) - \mathbf{Var}(\bar{X})) = \sigma^2.$$

3. By symmetry, each number $x_1, \ldots, x_n$ is equally likely to be the one drawn in the k-th place. This means that the distribution of X_k is the same as in problem 1. So, again, the expectation of the sample mean is μ.

$$\mathbf{Var}(\bar{X}_n) = \frac{1}{n^2}(n\mathbf{Var}X_1 + n(n-1)\mathbf{Cov}(X_1, X_2)).$$

One can easily calculate the covariance above directly, but the easiest way is to observe that for $n = N$ the variance of the sample mean is 0. Thus, the covariance term is

$$-\frac{\sigma^2}{N-1}$$

and we finally obtain

$$\mathbf{Var}(\bar{X}_n) = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

4. As in problem 2, we have

$$\mathbf{E}(S_n^2) = \frac{n}{n-1}(\mathbf{Var}(X_1) - \mathbf{Var}(\bar{X})).$$

Using the result of problem 3, we finally obtain $\frac{N}{N-1}\sigma^2$.

5.

$$L(X_1, \ldots, X_n, \mu, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp(-\frac{\sum_{i=1}^n (X_i - \mu)^2)}{2\sigma^2}) =$$

$$= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp(-\frac{(n-1)S_n^2 + n(\bar{X} - \mu)^2}{2\sigma^2}).$$

6. $T = \max(X_1, \ldots, X_n)$.

7. $T_1 = \max(X_1, \ldots, X_n)$, $T_2 = \min(X_1, \ldots, X_n)$.

8. The likelihood function is

$$L = \theta^n \exp(-\theta n \bar{X}_n) I_{[0,\infty]^n}(X_1, \ldots, X_n).$$

This is of the form $f(\theta, \bar{X}_n) g(X_1, \ldots, X_n)$.

9. $T_1 = \prod X_i$, $T_2 = \sum X_i$.

B.2 Problems from Chapter 2

1. In our proof of the Cramér-Rao theorem, we use Cauchy's inequality. For equality to hold, the two factors under the integral must be proportional. Thus,

$$\frac{\partial}{\partial\theta}(\log f_\theta(x)) = (\hat{\theta} - \theta) C(\theta)$$

where C is the proportionality factor which may depend on θ. By integrating over θ we get

$$\log f_\theta(x) = c(x) + \hat{\theta} \int C(\theta) d\theta - \int \theta C(\theta) d\theta.$$

This is already of the form we need. If we look at a sample of size n, the likelihood function must be of the same form, with the only difference that c and $\hat{\theta}$ are functions of n variables $X_1 \ldots, X_n$. $X_1 = X_2 = \ldots = X_n$, we see that f_θ has the form

$$\log f_\theta(x) = c_f(x) + a(x)A(\theta) + B(\theta),$$

where

$$c_f(x) = \frac{1}{n} c(x, \ldots, x),$$

$$a(x) = \hat{\theta}(x, \ldots, x),$$

$$A(\theta) = \frac{1}{n} \int C(\theta) d\theta,$$

and

$$B(\theta) = \frac{1}{n} \int \theta C(\theta) d\theta,$$

and comparing coefficients, we see that

$$\hat{\theta} = \frac{1}{n} \sum a(X_i).$$

On the other hand, by differentiating the equation

$$\int f_\theta(x) d\mu(x) = 1$$

with respect to θ, we find that if the densities are of this form, then $\hat{\theta}$ is an unbiased estimator and attains the Cramér-Rao bound.

2.

$$\hat{\lambda} = \frac{1}{\bar{X}_n}.$$

$\bar{X}_n$ has a Gamma distribution with density

$$\frac{x^{n-1}(n\lambda)^n e^{-n\lambda x}}{n!} \ (x > 0).$$

This yields

$$\mathbf{E}(\hat{\lambda}) = \mathbf{E}(\frac{1}{\bar{X}_n}) = \int_0^\infty \frac{x^{n-2}(n\lambda)^n e^{-n\lambda x}}{(n-1)!} = \frac{n\lambda}{n-1}.$$

The unbiased estimator is

$$\tilde{\lambda} = \frac{n-1}{n\bar{X}_n}.$$

3. (a) The likelihood function is

$$L(X_1, \ldots, X_n, \theta) = \begin{cases} \theta^{-n} & \text{if } \theta \geq \max(X_1, \ldots, X_n) \\ 0 & \text{otherwise.} \end{cases}$$

The first branch is decreasing, so the maximum is attained by choosing θ as small as possible. So, the ML estimator is

$$\hat{\theta}_a = \max(X_1, \ldots, X_n).$$

(b)

$$\hat{\theta}_b = 2\bar{X}_n.$$

(c)

$$\hat{\theta}_c = \frac{n+1}{n}\hat{\theta}_a.$$

(d)

$$\mathbf{Var}(\hat{\theta}_b) = \frac{\theta^2}{3n},$$

$$\mathbf{Var}(\hat{\theta}_c) = \frac{\theta^2}{n(n+2)},$$

(Use the fact that $\max(X_1, \ldots, X_n)$ has distribution function $(x/\theta)^n$ for $0 \leq x \leq \theta$) and the latter is smaller for $n > 1$ (for $n = 1$ the two estimators are the same).

4. $\log f_\lambda(x) = \log \lambda - \lambda x$. Differentiating twice with respect to λ gives $-\lambda^{-2}$. Thus the Cramér-Rao bound is λ^{-2}. The estimator from problem 2 does not attain this bound — its variance is larger by a factor $\frac{n-1}{n-2}$.

5. We verify that the estimator attains the Cramér-Rao bound. This is most easily verified by showing that the density

$$f_\theta(x) = \theta^x(1-\theta)^{1-x}$$

satisfies the conditions from problem 1.

6. Again, use problem 1.

7. (a) $\hat{\theta}_a = \frac{1}{2}\max(X_1, \ldots, X_n)$.

   (b) $\hat{\theta}_b = \frac{2(n+1)}{2n+1}\hat{\theta}_a$.

   (c) $\mathbf{Var}(\hat{\theta}_b) = \frac{n\theta^2}{(2n+1)^2(n+2)}$.

   The joint density of $U = \max(X_1, \ldots, X_n)$ and $V = \min(X_1, \ldots, X_n)$ is

$$f(u,v) = \begin{cases} \frac{n(n-1)(u-v)^{n-2}}{\theta^n} & \text{if } \theta < v < u < 2\theta, \\ 0 & \text{otherwise,} \end{cases}$$

   so

$$\mathbf{Var}\tilde{\theta} = \frac{n(n-1)}{9\theta^n}\int_\theta^{2\theta}\int_\theta^u (u+v)^2(u-v)^{n-2}\,dv\,du \ - \theta^2 = \frac{2\theta^2}{9(n+1)(n+2)}.$$

   For $n > 1$ this is smaller than the variance of $\hat{\theta}_b$.

8. We have to minimize

$$\sum_{i=1}^n |X_i - \theta|.$$

Unfortunately, the usual way of zeroing the derivative doesn't work perfectly in this case; but if θ is different from $X_1, \ldots, X_n$, then the derivative exists and equals

$$\sum \operatorname{sig}(\theta - X_i) = \#\{i : X_i < \theta\} - \#\{i : X_i > \theta\}.$$

Thus, if more than $n/2$ of the X_i are below θ, the sum above is increasing, otherwise it is decreasing. So, the minimum is attained if

$$\theta = X_{n:\frac{n+1}{2}}$$

if n is odd, and for any

$$X_{n:\frac{n}{2}} \le \theta \le X_{n:\frac{n}{2}+1}$$

if n is even, in other words, if θ is a median of $X_1, \ldots, X_n$.

9. $\bar{X}_n = 1.84$, $S_n^2 = .3338$.
   Confidence interval for μ: [1.47,2.21]
   Confidence interval for σ^2: [0.158,1.11]

10. $[1.264, 2.416]$.

11. $[\bar{X}_n - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\bar{X}_n}{n}}]$

12. $[\frac{1-z_{\frac{1+\gamma}{2}}n^{-1/2}}{\bar{X}_n}, \frac{1+z_{\frac{1+\gamma}{2}}n^{-1/2}}{\bar{X}_n}]$

13. $[\frac{\chi^2_{2n;\frac{1-\gamma}{2}}}{2n\bar{X}_n}, \frac{\chi^2_{2n;\frac{1+\gamma}{2}}}{2n\bar{X}_n}]$.

14. If we find $l(x) < u(x)$ such that for all λ
$$\mathbf{P}_\lambda(\lambda < l(X)) \leq \frac{1-\gamma}{2}$$
and
$$\mathbf{P}_\lambda(\lambda > u(X)) \leq \frac{1-\gamma}{2},$$
then $[l(X), u(X)]$ will be a confidence interval with coverage probability γ. Assume that both l and u are nondecreasing. For $l(k-1) < \lambda < l(k)$
$$\mathbf{P}_\lambda(\lambda < l(X)) = \mathbf{P}_\lambda(X \geq k) = \mathbf{P}(Y \leq 2\lambda),$$
where Y is chi^2_k. This probability should be at most $\frac{1-\gamma}{2}$, so we choose
$$l(k) = \frac{1}{2}\chi^2_{2k;\frac{1-\gamma}{2}}.$$
In a similar way we obtain
$$u(k) = \frac{1}{2}\chi^2_{2k+2;\frac{1+\gamma}{2}}.$$
If we have a sample of size n, $S = \sum_{i=1}^n$ has a Poisson distribution with parameter $n\lambda$, so finally, the confidence interval is
$$[\frac{1}{2n}\chi^2_{2S;\frac{1-\gamma}{2}}, \frac{1}{2n}\chi^2_{2S+2;\frac{1+\gamma}{2}}].$$

15. Let $\hat{\theta} = \max(X_1, \ldots, X_n)$. Since
$$\mathbf{P}(\hat{\theta} < x\theta) = x^n,$$
we can use the confidence interval
$$[\hat{\theta}, \hat{\theta}\frac{1}{(1-\gamma)^{1/n}}]$$

B.3 Problems from Chapter 3

1. First assume that $\mu_1 > \mu_0$. The ratio f_1/f_0 has the logarithm

$$\frac{\sum(X_i-\mu_0)^2}{2\sigma^2} - \frac{\sum(X_i-\mu_1)^2}{2\sigma^2} = \frac{\mu_1-\mu_0}{\sigma^2}\sum X_i + C,$$

where C is some constant. Thus, the Neyman-Pearson test is equivalent to

$$\phi = \begin{cases} 1 & \text{if } \sum X_i > K, \\ c & \text{if } \sum X_i = K, \\ 0 & \text{if } \sum X_i > K. \end{cases}$$

As $\sum X_i$ has a continuous distribution, we can let $c = 0$. We calculate K from

$$\alpha = \mathbf{P}_0(\sum X_i > K) = 1 - \Phi(\frac{K - n\mu_0}{\sqrt{n\sigma^2}}),$$

so

$$K = n\mu_0 + z_{1-\alpha}\sqrt{n\sigma^2}.$$

for $\mu_1 < \mu_0$, we reject if

$$\sum X_i < n\mu_0 + z_\alpha\sqrt{n\sigma^2}.$$

2. The denominator of the likelihood ratio statistic attains its maximum for $\mu = \bar{X}_n$ and

$$\sigma^2 = \frac{1}{n}\sum(X_i - \bar{X}_n)^2.$$

In the numerator, $\mu = \mu_0$ is fixed, and the maximum is attained for

$$\sigma^2 = \frac{1}{n}\sum(X_i - \mu_0)^2.$$

The likelihood ratio becomes

$$\left(\frac{\sum(X_i-\bar{X}_n)^2}{\sum(X_i-\mu_0)^2}\right)^{n/2} = \left(\frac{\sum(X_i-\bar{X}_n)^2}{\sum(X_i-\bar{X}_n)^2 + n(\bar{X}_n-\mu_0)^2}\right)^{n/2} =$$

$$(1+\frac{T^2}{n-1})^{-n/2},$$

where T is the t statistic. Thus we reject if $|T|$ is larger than a critical value; this is exactly the two-sided t test.

3. For $\lambda_1 > \lambda_0$, the Neyman-Pearson test is

$$\phi = \begin{cases} 1 & \text{if } \sum X_i > K, \\ c & \text{if } \sum X_i = K, \\ 0 & \text{if } \sum X_i < K, \end{cases}$$

where K and $0 \leq c < 1$ are determined from

$$c\frac{(n\lambda_0)^K}{K!}e^{-n\lambda_0} + \sum_{i>K} \frac{(n\lambda_0)^i}{i!}e^{-n\lambda_0} = \alpha.$$

4. For $\theta_1 > \theta_0$, the Neyman-Pearson test is

$$\phi = \begin{cases} 1 & \text{if } \theta_0 < \max(X_1, \ldots, X_n) \leq \theta_1, \\ \alpha & \text{if } \max(X_1, \ldots, X_n) \leq \theta_0, \end{cases}$$

We can modify ϕ on a set of probability 0:

$$\tilde{\phi} = \begin{cases} 1 & \text{if } \theta_0 < \max(X_1, \ldots, X_n), \\ \alpha & \text{if } \max(X_1, \ldots, X_n) \leq \theta_0. \end{cases}$$

This does no longer depend on the value of θ_1, so it is the best test for $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. For any $\theta < \theta_0$, The rejection probability is α, so it is also a test for $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

5. The likelihood function is

$$L(X_1, \ldots, X_n, \theta) = g(X_1, \ldots, X_n) f(T, \theta).$$

The factor g cancels in the likelihood ratio statistic, so it only depends on T.

6. Reject if

$$\sum \left(\frac{(X_i - \mu_0)^2}{\sigma_0^2} - \frac{(X_i - \mu_1)^2}{\sigma_1^2} \right) > t_c.$$

Unfortunately, we cannot handle the distribution of this. (Properly scaled, we obtain a so-called non-central χ^2 distribution which is the distribution of $(X_1 - a)^2 + \sum_{i=2}^n X_i^2$, where X_i are independent standard normal).

7. We can easily calculate the likelihood ratio test which rejects if

$$\lambda_0 \bar{X}_n e^{-\lambda_0 \bar{X}_n} < K,$$

which is equivalent to

$$X < \frac{A}{\lambda_0} \text{ or } X > \frac{B}{\lambda_0},$$

where A and B are determined by the equations

$$Ae^{-A} = Be^{-B}$$

and

$$\int_A^B \frac{x^{n-1}}{(n-1)!} e^{-x} dx = 1 - \alpha.$$

Another approach is to use the ideas from section 3.3 to obtain the best unbiased test. This turns out (after some calculations) to be identical to the likelihood ratio test.

8. In the likelihood ratio statistic, the denominator is maximal for $\mu_1 = \bar{X}$, $\mu_2 = \bar{Y}$,

$$\sigma_1^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

$$\sigma_2^2 = \frac{1}{m} \sum_{i=1}^{m} (Y_i - \bar{Y})^2.$$

In the numerator, we have to put

$$\sigma^2 = \frac{1}{n+m} (\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{m} (Y_i - \bar{Y})^2).$$

The likelihood ratio statistic is a function of F.

9. We test $H_0 : p = 0.05$ against $H_1 : p > 0.05$ (we don't care about p being smaller).

$\hat{p} = 0.075$. The critical value is

$$0.05 + 1.645 \sqrt{\frac{.05 * .95}{200}} = .0754$$

We accept H_0.

10. This time, test $H_0 : \mu = 500$ against $H_1 : \mu < 500$. We reject if

$$\bar{X}_n < 500 - t_{59;.95} \sqrt{\frac{2000}{60}} = 490.4$$

We reject H_0.

11. Test $H_0 : \lambda = 1/2000$ against $H_1 : \lambda > 1/2000$ (i.e., test against smaller mean lifetime) We reject if $S = \sum X_i$ is smaller than a critical value. This critical value can be obtained by making use of the fact that $2\lambda S$ has a χ^2 distribution with $2n$ degrees of freedom:

$$t_c = \frac{\chi^2_{2n;0.05}}{2\lambda} = 10851,$$

$S = 18950$ is greater than that, so we accept H_0.

12. $\bar{X}_n = 1.155$. We reject if

$$|\bar{X}_n - 1.0| > 1.96\sqrt{.5/11} = 0.417,$$

and as the left hand side is only 0.155, we accept H_0.

13. $S_n^2 = 0.4034$, and we use

$$t = \frac{1.155 - 1.0}{\sqrt{0.4034/11}} = 0.809$$

This is smaller in modulus than $t_{10:0.975} = 2.030$, so we accept H_0.

14. We reject if

$$(n-1)S_n^2 < \sigma_0^2 \chi^2_{n-1,0.05} = 1.970$$

or

$$(n-1)S_n^2 > \sigma_0^2 \chi^2_{n-1,0.95} = 24.996.$$

Neither is the case, so we accept H_0.

15.

$$f_{a,b}(x) = \frac{a}{b} \frac{\Gamma(\frac{a+b}{2})}{\Gamma(\frac{a}{2})\Gamma(\frac{b}{2})} \frac{(\frac{a}{b}x)^{\frac{a}{2}-1}}{(1+\frac{a}{b}x)^{\frac{a+b}{2}}} \ (x > 0).$$

16. As χ_n^2 is the distribution of a sum of the squares of n independent standard normal random variables, it follows from the central limit theorem that it has an approximate normal distribution. As its expectation and variance are n and $2n$, respectively, we get the desired result.

17. $F_{a,b}$ is the distribution of $\frac{X/a}{Y/b}$, where X and Y are independent χ^2 with a resp. b degrees of freedom. In the light of the previous problem, we may approximately (up to terms of smaller order)

replace X by $a + U\sqrt{2a}$, and Y by $b + V\sqrt{2b}$, where U and V are standard normal. This gives

$$F \approx \frac{1 + U\sqrt{2/a}}{1 + V\sqrt{2/b}} \approx 1 + U\sqrt{\frac{2}{a}} - V\sqrt{\frac{2}{b}} = 1 + W\sqrt{\frac{2(a+b)}{ab}},$$

where W is another standard normal random variable.

18. Under H_0, $Z_i = X_i - 2Y_i$ is normal with mean zero, so we can use the standard t test on $Z_1, \ldots, Z_n$.

19. We consider the general case of testing $H_0 : \mu_Y = a\mu_X$ under the assumption of equal variances. We easily find that

$$\bar{Y}_m - a\bar{X}_n \sim \mathcal{N}(0, \sigma^2(\frac{1}{m} + \frac{a^2}{n})),$$

$$\frac{m-1}{\sigma^2} S_Y^2 \sim \chi^2_{m-1},$$

and

$$\frac{n-1}{\sigma^2} S_X^2 \sim \chi^2_{n-1},$$

so we use the test statistic

$$T = \frac{\bar{Y}_m - a\bar{X}_n}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{nm(n+m-2)}{n + a^2 m}},$$

which has a t distribution with $m + n - 2$ degrees of freedom.

20. $\bar{Y}_m = 0.9444$, $S_Y^2 = 0.2404$

$$t = \frac{1.155 - 0.944}{\sqrt{10 * .4034 + 8 * 0.2404}} \sqrt{\frac{11 * 9 * 18}{20}} = 0.816$$

This is less than $t_{18,0.975} = 2.101$, so we accept H_0.

21. $F = \frac{.4034}{.2404} = 1.678 < F_{10,8;0.95} = 3.35$, so we accept H_0.

B.4 Problems from Chapter 4

1. We denote the variables in the sample for combination (i, j) by X_{ijl}, $i = 1, \ldots, n_{ij}$. Our model is

$$\mu_{ij} = \mu + a_i + b_j$$

with

$$\sum_{i=1}^{n} a_i n_{ij} = \sum_{j=1}^{k} b_j n_{ij} = 0$$

Let

$$n_{i.} = \sum_{j=1}^{k} n_{ij},$$

$$n_{.j} = \sum_{i=1}^{n} n_{ij},$$

$$n_{..} = \sum_{j=1}^{k} n_{.j},$$

then we get

$$\hat{\mu} = \frac{1}{n_{..}} \sum_{i,j,l} X_{ijl} = \bar{X}_{...},$$

$$\hat{a}_i = \frac{1}{n_{i.}} \sum_{j,l} X_{ijl} - \hat{\mu} = \bar{X}_{i..} - \bar{X}_{...},$$

$$\hat{b}_j = \frac{1}{n_{.j}} \sum_{i,l} X_{ijl} - \hat{\mu} = \bar{X}_{.j.} - \bar{X}_{...}.$$

In addition to the definitions of $X_{...}$ etc. as given above, let

$$\bar{X}_{ij.} = \frac{1}{n_{ij}} \sum_{l} X_{ijl}.$$

Finally, the test statistic is

$$F = \frac{\sum_i n_{i..}(\bar{X}_{i..} - \bar{X}_{...})^2}{\sum_{i,j,l}(X_{ijl} - X_{i..} - X_{.j.} + X_{...})^2},$$

and this has $k-1$ and $n_{..} - (n+k-1)$ degrees of freedom.

2. With the same notoations as in the previous problem, our test statistic is

$$F = \frac{\sum_{i,j} n_{ij}(X_{ij.} - X_{i..} - X_{.j.} + X_{...})^2/(n-1)(k-1)}{\sum_{i,j,l}(X_{ijl} - X_{ij.})^2/(n_{..} - nk)}.$$

This has $(n-1)(k-1)$ and $n_{..} - nk$ degrees of freedom.

3. We show that the F statistic is just the square of the t statistic used in the t test for the comparison of two samples of equal variance. The denominator is already of the form we need, so we only have to check the numerator. This becomes

$$n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2,$$

and as

$$\bar{X} = \frac{1}{n}(n_1\bar{X}_1 + n_2\bar{X}_2),$$

we obtain

$$(\bar{X}_1 - \bar{X}_2)^2 \frac{n_1 n_2^2 + n_2 n_1^2}{n} = (\bar{X}_1 - \bar{X}_2)^2 \frac{n_1 n_2}{n_1 + n_2},$$

which is exactly what we need.

4.

$$F = \frac{\sum_{i=1}^4 n_i(\bar{X}_i - \bar{X})^2/3}{\sum_{i=1}^4 (n_i - 1)S_i^2/(n-k)} = \frac{1.06/3}{4.5/39} = 3.0622 > 2.84 = F_{3,39;.95}$$

We reject H_0.

5. The confidence interval using all the samples is

$$[5.3 - t_{39,.975}\sqrt{\frac{4.5}{39 * 10}}, 5.3 + t_{39,.975}\sqrt{\frac{4.5}{39 * 10}}] = [5.0827, 5.5173].$$

6. From sample 1 alone we get

$$[5.0628, 5.5372].$$

7. we let

$$S^2 = \frac{1}{(n-1)(k-1)} \sum_{i,j} (X_{ij} - X_{i.} - X_{.j} + X_{..})^2.$$

with this notation, and from the fact that

$$\hat{a}_i = X_{i.} - X_{..}$$

has a normal distribution with mean a_i and variance $\frac{n-1}{nk}\sigma^2$, we get the confidence interval

$$X_{i.} - X_{..} - t_{(n-1)(k-1),\frac{1+\gamma}{2}} \sqrt{\frac{(n-1)S^2}{nk}}, X_{i.} - X_{..} + t_{(n-1)(k-1),\frac{1+\gamma}{2}} \sqrt{\frac{(n-1)S^2}{nk}}.$$

8. This is a simple two-way question with $n = 4$, $k = 3$; the F statistic is
$$F = 1.6071 < 4.76 = F_{3,6;.95}$$
we accept H_0.

9. $$F = 1.25 < 3.16 = F_{3,18;0.95},$$
we accept H_0.

B.5 Problems from Chapter 5

1. Of course, we want to make the variance of our estimators as small as possible. This means that
   (a) for $\hat{a}$, we have to make $\sum(x_i - \bar{x})^2$ as large as possible. The derivative with respect to x_i is
   $$2(x_i - \bar{x}).$$
   Thus, the sum is increasing if $x_i > \bar{x}$, and decreasing if $x_i < \bar{x}$. The maximum, therefore, can only be attained if $x_i = A$ or $x_i = -A$. Let k be the number of x's that are $= A$. Then the sum becomes
   $$A^2(n - \frac{1}{n}(n-2k)^2).$$
   This becomes a minimum if $k = n/2$.
   (b) for $\hat{b}$, the variance will be minimal if $\bar{x} = 0$. We don't need any other condition, but it might be best to use the set of x's from part (a), as this also makes the variance of $\hat{a}$ a minimum.

2. (a) The analysis from problem 1 stays valid, but now, we can't choose $k = n/2$ as n is odd. Instead, both $k = (n-1)/2$ and $k = (n+1)/2$ will minimize the variance of $\hat{a}$.
   (b) Again, the variance of $\hat{b}$ is a minimum if $\bar{x} = 0$. This time, however, the solution of part (a) does not satisfy this condition.

3. As always, we assume that the errors are normally distributed. The likelihood function is
$$\frac{1}{\sqrt{(2\pi)^n \prod \sigma(x_i)^2}} \exp\left(-\sum \frac{(Y_i - ax_i - b)^2}{2\sigma(x_i)^2}\right).$$

Maximizing this likelihood function is equivalent to finding the minimum of

$$\sum \frac{(Y_i - ax_i - b)^2}{\sigma(x_i)^2} = \sum (\frac{Y_i}{\sigma(x_i)} - a\frac{x_i}{\sigma(x_i)} - b\frac{1}{\sigma(x_i)})^2.$$

This is exactly what we have to minimize to obtain the usual least squares estimators for a and b in the modified problem.

4.

x	Y	x^2	xY	Y^2
0.5	1.3	0.25	0.65	1.69
1.0	2.7	1.00	2.70	7.29
2.0	1.5	4.00	3.00	2.25
2.5	2.0	6.25	5.00	4.00
3.0	2.2	9.00	6.60	4.84
9.0	9.7	20.50	17.95	20.07

$$\hat{a} = .114$$

$$\hat{b} = 1.735$$

$$\hat{\sigma}^2 = .398$$

$$\hat{Y}(3.5) = .114 * 3.5 + 1.735 = 2.134$$

prediction interval: $[-2.221, 6.459]$.

5. We let $Z = \log Y$ and compute the regression $Z = \log a + bx = B + Ax$:

x	Z	x^2	xZ	Z^2
0.5	0.2624	0.25	0.1312	0.0688
1.0	0.9933	1.00	0.9933	0.9866
2.0	0.4055	4.00	0.8110	0.1644
2.5	0.6931	6.25	1.7328	0.4804
3.0	0.7885	9.00	2.3655	0.6217
9.0	3.1428	20.50	6.0338	2.3419

$$\hat{A} = 0.08762 = \hat{b}$$

$$\hat{B} = 0.47085$$

$$\hat{a} = 1.6014$$

$$\hat{\sigma}^2 = 0.1111$$

$$\hat{Z}(3.5) = 0.47085 + .08762 * 3.5 = 0.7775$$
$$\hat{Y}(3.5) = 2.176$$

prediction interval for Z: $[-0.1894, 1.7444]$.

prediction interval for Y: $[0.8275, 5.7224]$.

6. We let $Z = 1/Y$ and compute the regression $Z = \frac{b}{a} + \frac{1}{a}x = B + Ax$:

x	Z	x^2	xZ	Z^2
0.5	0.7692	0.25	0.3846	0.5917
1.0	0.3704	1.00	0.3704	0.1372
2.0	0.6667	4.00	1.3333	0.4444
2.5	0.5000	6.25	1.2500	0.2500
3.0	0.4545	9.00	1.3636	0.2066
9.0	2.7608	20.50	4.7019	1.6299

$$\hat{A} = -0.06222$$
$$\hat{B} = 0.6642$$
$$\hat{a} = -16.07$$
$$\hat{b} = -10.67$$
$$\hat{\sigma}^2 = 0.0295$$
$$\hat{Z}(3.5) = 0.6642 - .06222 * 3.5 = 0.4464$$
$$\hat{Y}(3.5) = 2.2400$$

prediction interval for Z: $[-0.0517, 0.9446]$.

The prediction interval for Y would be the image of the prediction interval for Z under the function $x \to 1/x$. This doesn't make much sense.

7. $Y = a + bx + cx^2$

x	Y	x^2	xY	Y^2	x^3	x^2Y	x^4
0.5	1.3	0.25	0.65	1.69	0.125	0.325	0.0625
1.0	2.7	1.00	2.70	7.29	1.000	2.700	1.0000
2.0	1.5	4.00	3.00	2.25	8.000	6.000	8.0000
2.5	2.0	6.25	5.00	4.00	15.625	12.500	39.0625
3.0	2.2	9.00	6.60	4.84	27.000	19.800	81.0000
9.0	9.7	20.50	17.95	20.07	51.75	41.325	129.125

$5\hat{a}+$	$9\hat{b}+$	$20.5\hat{c} =$	9.7
$9\hat{a}+$	$20.5\hat{b}+$	$51.75\hat{c} =$	17.95
$20.5\hat{a}+$	$51.75\hat{b}$	$129.125\hat{c} =$	41.325

$$\hat{a} = 1.782$$

$$\hat{b} = 0.0376$$

$$\hat{c} = 0.0221$$

$$\hat{\sigma}^2 = .598$$

$$\hat{Y}(3.5) = 1.782 + 0.0376 * 3.5 + 0.0221 * 3.5^2 = 2.184$$

The variance of $\hat{Y}$ is $K\sigma^2$, where

$$K = A + 3.5B + 3.5^2 C$$

and A, B, C are the solutions of

$$\begin{array}{r|r|r|r} 5\hat{A}+ & 9\hat{B}+ & 20.5\hat{C} = & 1 \\ 9\hat{A}+ & 20.5\hat{B}+ & 51.75\hat{C} = & 3.5 \\ 20.5\hat{A}+ & 51.75\hat{B} & 129.125\hat{C} = & 3.5^2 \end{array}$$

$K = .03615$, so the variance for the prediction is $1.03615\sigma^2$, and we get the prediction interval: $[0.569, 3.799]$.

8. $[-0.509, 0.736]$.

9. $[-0.400, 3.870]$.

10. $\hat{y}(1.5) = 1.906$
    confidence interval: [0.943,2.869]

B.6 Problems from Chapter 6

1.

$$\chi^2 = \frac{(95-100)^2}{100} + \frac{(111-100)^2}{100} + \ldots + \frac{(113-100)^2}{100} = 6.16 < 11.070 = \chi^2_{5;0.95}$$

The die is fair.

2.

i	class	Y_i	np_i	$(Y_i - npi)^2/np_i$
1	$(-\infty, 0.524]$	3	5	0.8
2	$(0.524, 1]$	5	5	0.0
3	$(1, 1, 476]$	6	5	0.2
4	$(1.476, \infty)$	6	5	0.2
$\sum$	-	20	20	1.2

$$\chi^2 = 1.2 < 7.815 = \chi^2_{3;.95}$$

We accept the null hypothesis.

3. $\hat{\mu} = \bar{X} = 1.156,\ \hat{\sigma}^2 = S^2 = .4844$

i	class	Y_i	np_i	$(Y_i - npi)^2/np_i$
1	$(-\infty, 0.532]$	3	5	0.8
2	$(0.532, 1]$	5	5	0.0
3	$(1, 1, 468]$	6	5	0.2
4	$(1.468, \infty)$	6	5	0.2
$\sum$	-	20	20	1.2

$$\chi^2 = 1.2 < 3.841 = \chi^2_{1;.95}$$

We accept the null hypothesis.

4. $\hat{\lambda} = 1/\bar{X} = 0.864$

i	class	Y_i	np_i	$(Y_i - npi)^2/np_i$
1	$(0, 0.333]$	3	5	0.8
2	$(0.333, 0.802]$	4	5	0.2
3	$(0.802, 1.604]$	7	5	0.8
4	$(1.604, \infty)$	6	5	0.2
$\sum$	-	20	20	2.0

$$\chi^2 = 2.0 < 5.991 = \chi^2_{2;.95}$$

We accept the null hypothesis.

5. $\hat{\theta} = \max_i(X_i) = 2.70$

i	class	Y_i	np_i	$(Y_i - npi)^2/np_i$
1	$(0, 0.675]$	4	5	0.2
2	$(0.675, 1.35]$	10	5	5.0
3	$(1.35, 2.025]$	4	5	0.2
4	$(2.025, 2.70]$	2	5	1.8
$\sum$	-	20	20	7.2

$$\chi^2 = 7.2 > 5.991 = \chi^2_{2;.95}$$

We reject the null hypothesis.

6. $\hat{\lambda} = \bar{X} = 2.675$

i	class	Y_i	np_i	$(Y_i - np_i)^2/np_i$
1	0–1	9	10.129	0.078
2	2	8	9.861	0.351
3	3	13	8.793	2.013
4	4	5	5.880	0.132
5	5–∞	5	5.337	0.021
$\sum$	-	40	40.000	2.595

$\chi^2 = 2.595 < \chi^2_{3;.95} = 7.815$

We accept H_0.

7.

i	class	Y_i	np_i	$(Y_i - np_i)^2/np_i$
1	0–1	9	4.375	4.889
2	2	8	9.375	0.202
3	3	13	12.5	0.020
4	4	5	9.375	2.042
5	5–6	5	4.375	0.089
$\sum$	-	40	40.000	7.422

$\chi^2 = 7.422 < \chi^2_{4,.95} = 9.488$

We accept H_0.

8. $\hat{p} = \bar{X}/6 = 0.446$

i	class	Y_i	np_i	$(Y_i - np_i)^2/np_i$
1	0–2	17	9.937	5.020
2	3	13	12.065	0.072
3	4	5	11.248	3.471
4	5–6	5	6.750	0.454
$\sum$	-	40	40.000	9.017

$\chi^2 = 9.017 > \chi^2_{2,.95} = 5.991$

We reject H_0.

9. By symmetry, for $0 < z_i < 1$,

$$\mathbf{P}(Z_1 \le z_1, \ldots, Z_{n-1} \le z_{n-1} | M = m) =$$

$$\mathbf{P}(Y_1 \le mz_1, \ldots, Y_{n-1} = \le z_{n-1} | M = m) =$$

$$\mathbf{P}(Y_1 \le mz_1, \ldots, Y_{n-1} \le mz_{n-1} | X_1 < m, \ldots, X_{n-1} < m, X_n = m) =$$

$$\mathbf{P}(X_1 \le mz_1, \ldots, X_{n-1} \le y_{n-1} | X_1 < m, \ldots, X_{n-1} < m) = z_1 * \ldots * z_{n-1}.$$

Thus, $Z_1, \ldots, Z_{n-1}$ are independent, uniformly distributed on $[0,1]$.

Now, in our usual χ^2 test situation, we would divide the interval $[0, M]$ into k intervals of length M/k. Denote the number of X's in the ith interval by U_i. Then the χ^2 statistic is

$$\chi_X^2 = \frac{k}{n} \sum_{i=1}^{k} (U_i - n/k)^2.$$

On the other hand, if we denote the number of Z_i's in the interval $(\frac{i-1}{k}, \frac{i}{k}]$ by V_i, then

$$U_i = \begin{cases} V_i & \text{if } i < k \\ V_i + 1 & \text{if } i = k. \end{cases}$$

Furthermore, the chi^2 statistik for the sample $Z_1, \ldots, Z_{n-1}$ is

$$\chi_Z^2 = \frac{k}{n-1} \sum_{i=1}^{k} (V_i - (n-1)/k)^2,$$

and a little calculation yields

$$\chi_X^2 = \frac{n-1}{n} \chi_Z^2 + \frac{k}{n}(V_k - \frac{n-1}{k}) + \frac{k-1}{n}.$$

For large n, this implies that χ_X^2 and χ_Z^2 have the same limiting distribution. We know that χ_Z^2 has an approximate χ^2 distribution with $k-1$ degrees of freedom, so the same is true for χ_X^2.

10. $\chi^2 = 170.7 > 26.269 = \chi^2_{16,.95}$

    We reject the null hypothesis of independence.

11. $$\chi^2 = \frac{n(Z_{11}Z_{22} - Z_{12}Z_{21})^2}{Z_{1.}Z_{2.}Z_{.1}Z_{.2}}.$$

12. $\chi^2 = 12.090 > 3.841 = \chi^2_{1;.95}$

    We reject the hypothesis of independence.

13. $\chi^2 = 8.4727 < 15.507 = \chi^2_{8,.95}$

    We accept the null hypothesis that the three professors are equal.

B.7 Problems from Chapter 7

1.

i	$X_{n:i}$	$F(X_{n:i})$	$F_n(X_{n:i}-0)$	$F_n(X_{n:i})$	Δ_i
1	0.05	0.090	0.00	0.05	0.090
2	0.16	0.117	0.05	0.10	0.067
3	0.21	0.133	0.10	0.15	0.033
4	0.58	0.278	0.15	0.20	0.128
5	0.70	0.337	0.20	0.25	0.137
6	0.70	0.337	0.25	0.30	0.087
7	0.74	0.356	0.30	0.35	0.056
8	1.00	0.500	0.35	0.40	$\underline{0.150}$
9	1.02	0.512	0.40	0.45	0.112
10	1.06	0.532	0.45	0.50	0.087
11	1.12	0.567	0.50	0.55	0.067
12	1.16	0.587	0.55	0.60	0.037
13	1.25	0.637	0.60	0.65	0.037
14	1.30	0.663	0.65	0.70	0.037
15	1.71	0.841	0.70	0.75	0.141
16	1.75	0.855	0.75	0.80	0.105
17	1.84	0.883	0.80	0.85	0.083
18	1.95	0.918	0.85	0.90	0.068
19	2.13	0.945	0.90	0.95	0.045
20	2.70	0.992	0.95	1.00	0.042

$D_{20} = 0.150 < 0.2941$. We accept H_0.

2.

i	$X_{n:i}$	$F_n(X_{n:i}-0)$	$F_n(X_{n:i})$	Δ_i
1	0.23	0.000	0.125	$\underline{0.230}$
2	0.31	0.125	0.250	0.185
3	0.45	0.250	0.375	0.200
4	0.49	0.375	0.500	0.115
5	0.67	0.500	0.625	0.170
6	0.80	0.625	0.750	0.175
7	0.87	0.750	0.875	0.120
8	0.93	0.875	1.000	0.070

$D_8 = 0.23 < 0.4543$: We accept H_0.

3. As for the one-sample test, it can easily be shown that

$$\sup\{|G_m(x) - F_n(x)|\} =$$

$$\sup\{\max(|G_m(x)-F_n(x)|, |G_m(x-0)-F_n(x-0)|) : x \in \{X_i, i \leq n\} \cup \{Y_j, j \leq m\}\}.$$

Thus, we use the following table:

x	$F_n(x-0)$	$F_n(x)$	$G_m(x-0)$	$G_m(x)$	$\Delta(x-0)$	$\Delta(x)$
0.6	0.000	0.111	0.000	0.091	0.000	0.020
0.7	0.111	0.222	0.091	0.273	0.020	0.051
0.8	0.222	0.333	0.273	0.364	0.051	0.031
0.9	0.333	0.556	0.364	0.364	0.031	0.192
1.0	0.556	0.667	0.364	0.364	0.192	0.303
1.1	0.667	0.889	0.364	0.455	0.303	<u>0.434</u>
1.3	0.889	0.889	0.455	0.636	0.434	0.253
1.4	0.889	1.000	0.636	0.818	0.253	0.182
1.6	1.000	1.000	0.818	0.909	0.182	0.091
1.8	1.000	1.000	0.909	1.000	0.091	0.000

We don't have tables for the two-sample test, so we use the asymptotic distribution (although we shouldn't):

$$D_{nm} = 0.434$$

$$\lambda_{nm} = 0.434\sqrt{\frac{9*11}{20}} = 0.966 < 1.358,$$

so we accept H_0.

4. Let S be the set of all permutations of $\{1,\ldots,n\}$. Then

$$\mathbf{P}(D_n < x) = \mathbf{P}(|\frac{k}{n} - U_{n:k}| < x, |\frac{k-1}{n} - U_{n:k}| < x) = \mathbf{P}(\frac{k}{n} - x < U_{n:k} < \frac{k-1}{n} + x) =$$

$$\sum_{s\in S}\mathbf{P}(\frac{k}{n} - x < U_{s(k)} < \frac{k-1}{n} + x, U_{n:k} = U_{s(k)}) =$$

$$\sum_{s\in S}\mathbf{P}(\frac{k}{n} - x < U_{s(k)} < \frac{k-1}{n} + x, U_{s(1)} \le U_{s(2)} \le \ldots \le U_{s(n)}) =$$

$$= n!\mathbf{P}(\frac{k}{n} - x < U_k < \frac{k-1}{n} + x, U_1 \le U_2 \le \ldots \le U_n).$$

5.

$$\mathbf{P}(D_2 \le x) = \mathbf{P}(\frac{1}{2} - x \le U_1 \le x, 1 - x \le U_2 \le \frac{1}{2} + x, U_1 < U_2)$$

If $x < \frac{1}{4}$, $\mathbf{P}(D_2 \le x) = 0$.
If $\frac{1}{4} \le x < \frac{1}{2}$,

$$\mathbf{P}(D_2 \le x) = 2\mathbf{P}(\frac{1}{2} - x \le U_1 \le x, 1 - x \le U_2 \le \frac{1}{2} + x) = 2(\frac{1}{2} - 2x)^2.$$

If $\frac{1}{2} \leq x \leq 1$,

$$\mathbf{P}(D_2 \leq x) = \mathbf{P}(U_{2:1} \leq x, U_{2:2} \geq 1-x) =$$

$$1 - \mathbf{P}(U_{2:1} \geq x) - \mathbf{P}(U_{2:2} \leq 1-x) + \mathbf{P}(U_{1:1} \geq x, U_{2:2} \leq 1-x) =$$

$$1 - 2(1-x)^2 + 0 = 1 - 2(1-x)^2.$$

6.

$$\eta_{nm}(x) = F_n(x) - G_m(x) = \sum_{i=1}^{n}\sum_{j=1}^{m}(V_i(x) - W_j(x)),$$

where

$$V_i = \begin{cases} 1 & \text{if } X_i \leq x, \\ 0 & \text{otherwise,} \end{cases}$$

$$W_j = \begin{cases} 1 & \text{if } Y_j \leq x, \\ 0 & \text{otherwise,} \end{cases}$$

if $i_1 \neq i_2$, $V_{i_1}(x)$ and $V_{i_2}(y)$ are independent; also, if $j_1 \neq j_2$, $W_{j_1}(x)$ and $W_{j_2}(y)$ are independent; furthermore, $V_i(a)$ and $W_j(b)$ are independent for any choice of i, j, a, and b. For the remaining cases, we get the covariances

$$\mathbf{Cov}(V_i(x), V_i(y)) = \mathbf{Cov}(W_j(x), W_j(y)) = x \wedge y - xy,$$

where we have put

$$x \wedge y = \min(x, y).$$

Thus

$$\mathbf{Cov}(\eta_{nm}(x), \eta_{nm}(y)) = \sum_{i=1}^{n} \frac{1}{n^2}\mathbf{Cov}(V_i(x), V_i(y)) + \sum_{j=1}^{m} \frac{1}{m^2}\mathbf{Cov}(W_j(x), W_j(y)) =$$

$$\frac{1}{n}(x \wedge y - xy) + \frac{1}{m}(x \wedge y - xy) = \frac{n+m}{nm}(x \wedge y - xy).$$

Letting $x = y$ gives the variance of $\eta_{mn}(x)$.

Appendix C

German translations of technical terms

accept annehmen

alternative hypothesis Alternativhypothese, Gegenhypothese, Eins-Hypothese (brr)

analysis of variance Varianzanalyse

composite hypothesis zusammengesetzte Hypothese

confidence interval Konfidenzintervall

confidence region Konfidenzbereich

consistent konsistent

coverage probability Überdeckungswahrscheinlichkeit

degrees of freedom Freiheitsgrade

descriptive statistics beschreibende Statistik

distribution Verteilung

efficient effizient

empirical distribution function empirische Verteilungsfunktion

estimation Schätzung

estimator Schätzer

frequency Häufigkeit

goodness-of-fit test Anpassungstest

hypothesis Hypothese

inductive statistics schließende Statistik

level of significance Signifikanzniveau (kurz NIveau)

likelihood Likelihood (bei Leuten, die Fremdwörter nicht so sehr mögen: "Plausibilität")

likelihood ratio test Likelihoodquotiententest

linear regression lineare Regression

mathematical statistics Mathematische Statistik

maximum likelihood estimator Maximum Liklihood Schätzer (auch: plausibler Schätzer)

method of moments Momentenmethode

multiple linear regression mehrfache lineare Regression

non-parametric statistics nichtparametrische Statistik

null hypothesis Nullhypothese

one-sided hypothesis einseitige Hypothese

one-way analysis of variance einfache Varianzanalyse

operation characteristic Operationscharkteristik

parameter Parameter

parametric family parametrische Familie

parametric statistics parametrische Statistik

population Grundgesamtheit, Population

power function Macht, Mächtigkeit

prediction interval Vorhersageintervall

randomized randomisiert

reject ablehnen, verwerfen

rejection region Verwerfungsbereich

sample mean Stichprobenmittel

sample variance Stichprobenvarianz

sample Stichprobe

simple hypothesis einfache Hypothese

simple linear regression einfache lineare Regression

statistic Statistik

statistics Statistik

sufficient statistic suffiziente Statistik

test Test, Signifikanztest

two-sided hypothesis zweiseitige Hypothese

two-way analysis of variance doppelte Varianzanalyse

unbiased estimator erwartungstreuer Schätzer, unverzerrter Schätzer

unbiased test unverfälschter Test

uniform distribution Gleichverteilung

Appendix D

The Gnu General Public License

D.1 Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software—to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These

restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.
For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.
We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.
Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.
Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.
The precise terms and conditions for copying, distribution and modification follow.

D.2 TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

1. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in the term "modification".) Each licensee is addressed as "you".

   Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the

Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

2. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

   You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

3. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

   (a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.

   (b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.

   (c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

   These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

4. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

   (a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

   (b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,

   (c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

   The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

5. You may not copy, modify, sublicense, or distribute the Program except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

6. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

7. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

8. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence

you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

9. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

10. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

    Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

11. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

12. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

13. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

D.3 How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms. To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
{one line to give the program's name and an idea of what it does.}
Copyright (C) 19yy  {name of author}

This program is free software; you can redistribute it and/or
modify it under the terms of the GNU General Public License
as published by the Free Software Foundation; either version 2
of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public License
along with this program; if not, write to the Free Software
Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.
```

Also add information on how to contact you by electronic and paper mail. If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

```
Gnomovision version 69, Copyright (C) 19yy {name of author}
Gnomovision comes with ABSOLUTELY NO WARRANTY; for details
type `show w'.  This is free software, and you are welcome
to redistribute it under certain conditions; type `show c'
for details.
```

The hypothetical commands *show w* and *show c* should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than *show w* and *show c*; they could even be mouse-clicks or menu items—whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

```
Yoyodyne, Inc., hereby disclaims all copyright
interest in the program 'Gnomovision'
(which makes passes at compilers) written
by James Hacker.

{signature of Ty Coon}, 1 April 1989
Ty Coon, President of Vice
```

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.

Index