Lecture 1

Overview of some probability distributions.

In this lecture we will review several common distributions that will be used often throughtout the class. Each distribution is usually described by its probability function (p.f.) in the case of discrete distributions or probability density function (p.d.f.) in the case of continuous distributions. Let us recall basic definitions associated with these two cases.

Discrete distributions.

Suppose that a set $\mathcal{X}$ consists of a countable or finite number of points,

$$\mathcal{X} = \{a_1, a_2, a_3, \cdots\}.$$

Then a probability distribution $\mathbb{P}$ on $\mathcal{X}$ can be defined via a function $p(x)$ on $\mathcal{X}$ with the following properties:

1. $0 \leq p(a_i) \leq 1$,
2. $\sum_{i=1}^{\infty} p(a_i) = 1$.

A function $p(x)$ is called the probability function. If X is a random variable with distribution $\mathbb{P}$ then $p(a_i) = \mathbb{P}(X = a_i)$ - a probability that X takes value a_i. Given a function $\varphi : \mathcal{X} \to \mathbb{R}$, the expectation of $\varphi(X)$ is defined by

$$\mathbb{E}\varphi(X) = \sum_{i=1}^{\infty} \varphi(a_i)p(a_i).$$

(Absolutely) continuous distributions.

Continuous distribution $\mathbb{P}$ on $\mathbb{R}$ is defined via a probability density function (p.d.f.) $p(x)$ on $\mathbb{R}$ such that

$$p(X) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(X)dx = 1.$$

If a random variable X has distribution $\mathbb{P}$ then the probability that X takes a value in the interval $[a, b]$ is given by

$$\mathbb{P}(X \in [a, b]) = \int_a^b p(x)dx.$$

Clearly, in this case for any $a \in \mathbb{R}$ we have $\mathbb{P}(X = a) = 0$. Given a function $\varphi : \mathcal{X} \to \mathbb{R}$, the expectation of $\varphi(X)$ is defined by

$$\mathbb{E}\varphi(X) = \int_{-\infty}^{\infty} \varphi(x)p(x)dx.$$

Notation. The fact that a random variable X has distribution $\mathbb{P}$ will be denoted by $X \sim \mathbb{P}$.

Normal (Gaussian) Distribution $N(\alpha, \sigma^2)$. Normal distribution is a continuous distribution on $\mathbb{R}$ with p.d.f.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} \text{ for } x \in (-\infty, \infty).$$

Here $-\infty < \alpha < \infty, \sigma > 0$ are the parameters of the distribution. Let us recall some properties of a normal distribution. If a random variable X has a normal distribution $N(\alpha, \sigma^2)$ then the r.v.

$$Y = \frac{X - \alpha}{\sigma} \sim N(0, 1)$$

has a *standard normal distribution* $N(0, 1)$. To see this, we can write,

$$\begin{aligned} \mathbb{P}\Big(\frac{X-\alpha}{\sigma} \in [a, b]\Big) = \mathbb{P}(X \in [a\sigma + \alpha, b\sigma + \alpha]) &= \int_{a\sigma+\alpha}^{b\sigma+\alpha} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} dx \\ &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy, \end{aligned}$$

where in the last integral we made a change of variables $y = (x - \alpha)/\sigma$. This, of course, means that $Y \sim N(0, 1)$. The expectation of Y is

$$\mathbb{E}Y = \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 0$$

since the integrand is an odd function. To compute the second moment $\mathbb{E}Y^2$, let us first note that since $\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$ is a probability density function, it integrates to 1, i.e.

$$1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$

If we integrate this by parts, we get,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \frac{1}{\sqrt{2\pi}} y e^{-\frac{y^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi}} (-y) e^{-\frac{y^2}{2}} dy \\ &= 0 + \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \mathbb{E}Y^2. \end{aligned}$$

Thus, the second moment $\mathbb{E}Y^2 = 1$. The variance of Y is

$$\text{Var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = 1 - 0 = 1.$$

It is now easy to compute the mean and the variance of $X = \alpha + \sigma Y \sim N(\alpha, \sigma^2)$,

$$\mathbb{E}X = \alpha + \sigma \mathbb{E}Y = \alpha, \;\; \mathbb{E}X^2 = \mathbb{E}(\alpha^2 + 2\alpha\sigma Y + \sigma^2 Y^2) = \alpha^2 + \sigma^2,$$

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \alpha^2 + \sigma^2 - \alpha^2 = \sigma^2.$$

Thus, parameter α is a mean and parameter σ^2 is a variance of a normal distribution. Let us recall (without giving a proof) that if we have several, say n, independent random variables $X_i, 1 \le i \le n$, such that $X_i \sim N(\alpha_i, \sigma_i^2)$ then their sum will also have a normal distribution

$$X_1 + \ldots + X_n \sim N(\alpha_1 + \ldots + \alpha_n, \sigma_1^2 + \ldots + \sigma_n^2).$$

Normal distribution appears in one of the most important results that one learns in probability class, namely, a *Central Limit Theorem (CLT)*, which states the following. If $X_1, \ldots, X_n$ is an i.i.d. sample such that $\sigma^2 = \mathrm{Var}(X) < \infty$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}X_i) = \sqrt{n}(\bar{X} - \mathbb{E}X_1) \to^d N(0, \sigma^2)$$

converges in distribution to a normal distribution with zero mean and variance σ^2, where convergence in distribution means that for any interval $[a, b]$,

$$\mathbb{P}\Big(\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \in [a, b]\Big) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.$$

This result can be generalized for a sequence of random variables with different distributions and it basically says that the sum of many independent random variables/factors approximately looks like a normal distribution as long as each factor has a small impact on the total sum. A consequence of this phenomenon is that a normal distribution gives a good approximation for many random objects that by nature are affected by a sum of many independent factors, for example, person's height or weight, fluctuations of a stock's price, etc.

Bernoulli Distribution $B(p)$**.** This distribution describes a random variable that can take only two possible values, i.e. $\mathcal{X} = \{0, 1\}$. The distribution is described by a probability function

$$p(1) = \mathbb{P}(X = 1) = p, \;\; p(0) = \mathbb{P}(X = 0) = 1 - p \text{ for some } p \in [0, 1].$$

It is easy to check that

$$\mathbb{E}X = p, \;\; \mathrm{Var}(X) = p(1 - p).$$

Binomial Distribution $B(n, p)$**.** This distribution describes a random variable X that is a number of successes in n trials with probability of success p. In other words, X is a sum of n independent Bernoulli r.v. Therefore, X takes values in $\mathcal{X} = \{0, 1, \ldots, n\}$ and the distribution is given by a probability function

$$p(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

It is easy to check that

$$\mathbb{E}X = np, \;\; \mathrm{Var}(X) = np(1-p).$$

Exponential Distribution $E(\alpha)$. This is a continuous distribution with p.d.f.

$$p(x) = \begin{cases} \alpha e^{-\alpha x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Here, $\alpha > 0$ is the parameter of the distribution. Again, it is a simple calculus exersice to check that

$$\mathbb{E}X = \frac{1}{\alpha}, \;\; \mathrm{Var}(X) = \frac{1}{\alpha^2}.$$

This distribution has the following nice property. If a random variable $X \sim E(\alpha)$ then probability that X exceeds level t for some $t > 0$ is

$$\mathbb{P}(X \geq t) = \mathbb{P}(X \in [t, \infty)) = \int_t^\infty \alpha e^{-\alpha x} dx = e^{-\alpha t}.$$

Given another $s > 0$, the conditional probability that X will exceed level $t + s$ given that it will exceed level t can be computed as follows:

$$\begin{aligned} \mathbb{P}(X \geq t + s | X \geq t) &= \frac{\mathbb{P}(X \geq t + s, X \geq t)}{\mathbb{P}(X \geq t)} = \frac{\mathbb{P}(X \geq t + s)}{\mathbb{P}(X \geq t)} \\ &= e^{-\alpha(t+s)} / e^{-\alpha t} = e^{-\alpha s} = \mathbb{P}(X \geq s), \end{aligned}$$

i.e.

$$\mathbb{P}(X \geq t + s | X \geq t) = \mathbb{P}(X \geq s).$$

If X represent a lifetime of some object in some random conditions, then the above property means that the chance that X will "live" longer then $t + s$ given that it will "live" longer than t is the same as the chance that X will live longer than t in the first place. Or, in other words, if X is "alive" at time t then it is "like new". Therefore, some natural examples that can be described by exponential distribution are the lifetime of high quality products (or, possibly, soldiers in combat).

Poisson Distribution $\Pi(\lambda)$. This is a discrete distribution with

$$\mathcal{X} = \{0, 1, 2, 3, \ldots\},$$

$$p(k) = \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \text{ for } k = 0, 1, 2, , \ldots$$

It is an exercise to show that

$$\mathbb{E}X = \lambda, \mathrm{Var}(X) = \lambda.$$

Poisson distribution could be used to describe the following random objects: the number of stars in a random area of the space; number of misprints in a typed page; number of wrong connections to your phone number; distribution of bacteria on some surface or weed in the field. All these examples share some common properties that give rise to a Poisson distribution. Suppose that we count a number of random objects in a certain region T and this counting process has the following properties:

1. Average number of objects in any region $S \subseteq T$ is proportional to the size of S, i.e. $\mathbb{E}\text{Count}(S) = \lambda|S|$. Here $|S|$ denotes the size of S, i.e. length, area, volume, etc. Parameter $\lambda > 0$ represents the intensity of the process.

2. Counts on disjoint regions are independent.

3. Chance to observe more than one object in a small region is very small, i.e. $\mathbb{P}(\text{Count}(S) \geq 2)$ becomes small when the size $|S|$ gets small.

We will show that under these assumptions will imply that the number $\text{Count}(T)$ of objects in the region T has Poisson distribution $\Pi(\lambda|T|)$ with parameter $\lambda|T|$.

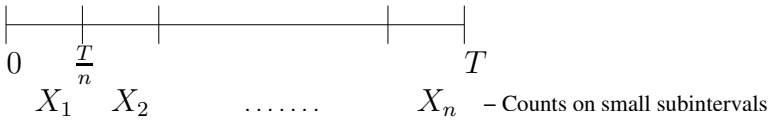


Figure 1.1: Poisson Distribution

For simplicity, let us assume that the region T is an interval $[0, T]$ of length T. Let us split this interval into a large number n of small equal subintervals of length T/n and denote by X_i the number of random objects in the ith subinterval, $i = 1, \ldots, n$. By the first property above,

$$\mathbb{E}X_i = \frac{\lambda T}{n}.$$

On the other hand, by definition of expectation

$$\mathbb{E}X_i = \sum_{k\geq 0} k\mathbb{P}(X_i = k) = 0 + \mathbb{P}(X_i = 1) + \varepsilon_n,$$

where $\varepsilon_n = \sum_{k\geq 2} k\mathbb{P}(X_i = k)$, and by the last property above we assume that ε_n becomes small with n, since the probability to observe more that two objects on the interval of size T/n becomes small as n becomes large. Combining two equations above gives, $\mathbb{P}(X_i = 1) \approx \lambda\frac{T}{n}$. Also, since by the last property the probability that any count X_i is ≥ 2 is small, i.e.

$$\mathbb{P}(\text{at least one } X_i \geq 2) \leq n\text{o}\Big(\frac{T}{n}\Big) \to 0 \text{ as } n \to \infty,$$

$\text{Count}(T) = X_1 + \cdots + X_n$ has approximately binomial distribution $B(n, \lambda|T|/n)$ and we can write

$$\begin{aligned}\mathbb{P}(\text{Count}(T) = X_1 + \cdots + X_n = k) &\approx \binom{n}{k}\Big(\frac{\lambda T}{n}\Big)^k\Big(1 - \frac{\lambda T}{n}\Big)^{n-k} \\ &\to \frac{(\lambda T)^k}{k!}e^{-\lambda T}.\end{aligned}$$

The last limit is a simple calculus exercise and this is also a famous Poisson approximation of binomial distribution taught in every probability class.

Uniform Distribution $U[0,\theta]$**.** This distribution has probability density function

$$p(x) = \begin{cases} \frac{1}{\theta}, & x \in [0,\theta], \\ 0, & otherwise. \end{cases}$$

Matlab review of probability distributions.

Matlab Help/Statistics Toolbox/Probability Distributions.

Each distribution in Matlab has a name, for example, normal distribution has a name 'norm'. Adding a suffix defines a function associated with this distribution. For example, 'normrnd' generates random numbers from distribution 'norm', 'normpdf' gives p.d.f., 'normcdf' gives c.d.f., 'normfit' fits the normal distribution for a given dataset (we will look at this last type of functions when we discuss Maximum Likelihood Estimators). Please, look at each function for its syntax, input, output, etc. Type 'help normrnd' to quickly see how the normal random number generator works. Also, there is a graphic user interface tools like 'disttool' (to run it just type disttool in the main Matlab window) that allows you to play with different distributions, or 'randtool' that generates and visualizes random samples from different distributions.

Lecture 2

Maximum Likelihood Estimators.

Matlab example. As a motivation, let us look at one Matlab example. Let us generate a random sample of size 100 from beta distribution Beta$(5, 2)$. We will learn the definition of beta distribution later, at this point we only need to know that this isi a continuous distribution on the interval $[0, 1]$. This can be done by typing 'X=betarnd(5,2,100,1)'. Let us fit different distributions by using a distribution fitting tool 'dfittool'. We try to fit normal distribution and beta distribution to this sample and the results are displayed in figure 2.1.

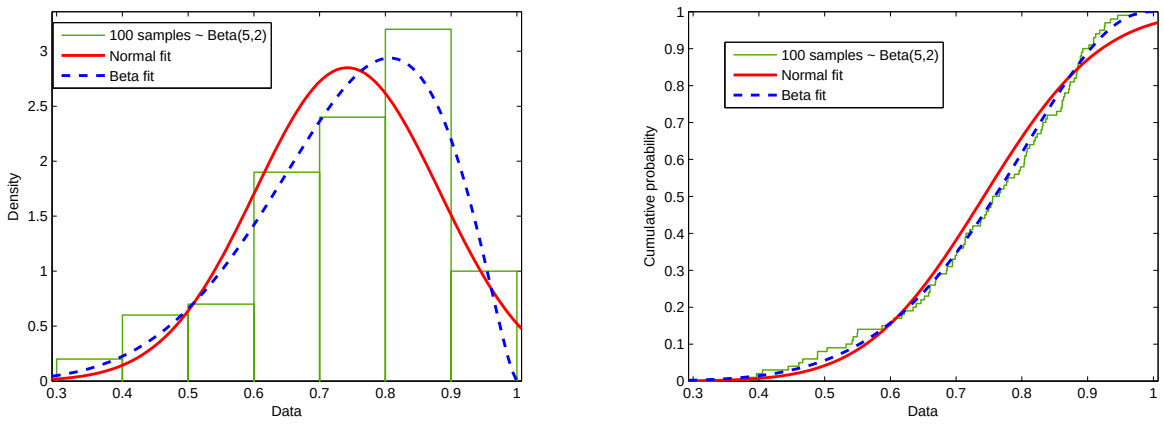


Figure 2.1: Fitting a random sample of size 100 from Beta$(5, 2)$. (a) Histogram of the data and p.d.f.s of fitted normal (solid line) and beta (dashed line) distributions; (b) Empirical c.d.f. and c.d.f.s of fitted normal and beta distributions.

Besides the graphs, the distribution fitting tool outputs the following information:

```
Distribution:    Normal
Log likelihood:  55.2571
```

```
Domain:           -Inf < y < Inf
Mean:             0.742119
Variance:         0.0195845

Parameter  Estimate  Std. Err.
mu         0.742119   0.0139945
sigma      0.139945  0.00997064

Estimated covariance of parameter estimates:
       mu            sigma
mu     0.000195845   6.01523e-020
sigma  6.01523e-020  9.94136e-005


Distribution:    Beta
Log likelihood:  63.8445
Domain:          0 < y < 1
Mean:            0.741371
Variance:        0.0184152

Parameter  Estimate  Std. Err.
a          6.97783    1.08827
b          2.43424   0.378351

Estimated covariance of parameter estimates:
   a         b
a   1.18433  0.370094
b  0.370094  0.143149
```

The value 'Log likelihood' indicates that the tool uses the maximum likelihood estimators to fit the distribution, which will be the topic of the next few lectures. Notice the 'Parameter estimates' - given the data 'dfittool' estimates the unknown parameters of the distribution and then graphs the p.d.f. or c.d.f. corresponding to these parameters.

Since the data was generated from beta distribution, it is not surprising that beta distribution fit seems better than normal distribution fit, which is particularly clear from figure 2.1 (b), that compares how estimated c.d.f. fits the empirical c.d.f. Empirical c.d.f. is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

where $I(X_n \leq x)$ is the indicator that X_i is $\leq x$. In other words, $F_n(x)$ is the proportion of observations below level x.

One can ask several questions about this example:

1. How to estimate the unknown parameters of a distribution given the data from this distribution?

2. How good are these estimates, are they close to the actual 'true' parameters?

3. Does the data come from a particular type of distribution, for example, normal or beta distribution?

In the next few lectures we will study the first two questions and we will assume that *we know what type of distribution the sample comes from, so we only do not know* the parameters of the distribution. In the context of the above example, we would be told that the data comes from beta distribution, but the parameters $(5, 2)$ would be unknown. Of course, in general we might not know what kind of distribution the data comes from - we will study this type of questions later when we look at the so called *goodness-of-fit hypotheses tests*. In particular, we will see graphs like 2.1 (b) again when we study the Kolmogorov-Smirnov goodness-of-fit test.

□

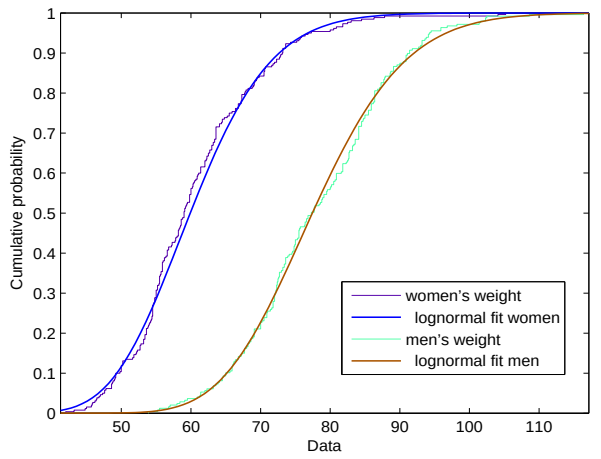
Example. We consider a dataset of various body measurements from [1] (dataset can be dowloaded from journal's website), including weight, height, waist girth, abdomen girth, etc. First, we use Matlab fitting tool to fit weight and waist girth of men and women (separately) with lognormal distribution, see figure 2.2 (a) and (b). Wikipedia article about normal distribution gives a reference to a 1932 book "Problems of Relative Growth" by Julian Huxley for the explanation why the sizes of full-grown animals are approximately log-normal. One short explanation is consistency between linear and volume dimensions - if linear dimensions are lognormal and volume dimensions are proportional to cube of linear dimensions then they also are lognormal. Assumption that sizes are normal would violate this consistency, since the cube of normal is not normal. We observe, hovewer, that the fit of women's waist with lognormal is not very accurate. Later in the class we will learn several statistical tests to decide if the data comes from a certain distribution or a family of distributions, but here is a preview of what's to come. Chi-squared goodness-of-fit test rejects the hypothesis that the distribution of logarithms of women's waists is normal:

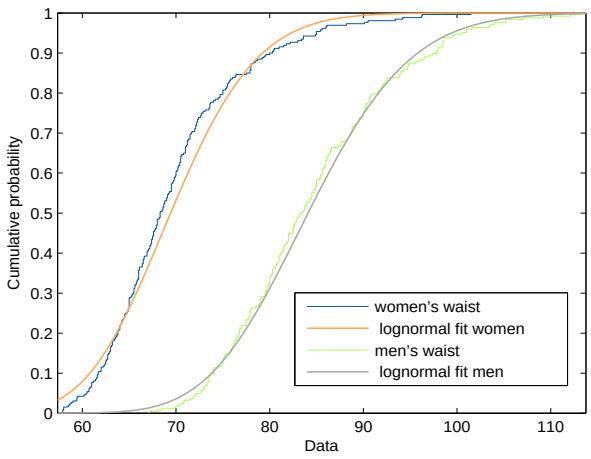
```
[h,p,stats]=chi2gof(log_women_waist)

h = 1, p = 5.2297e-004
stats =  chi2stat: 22.0027
         df: 5
         edges: [1x9 double]
         O: [21 44 67 60 28 18 12 10]
         E: [1x8 double]
```

and so does Lilliefor's test (adjusted Kolmogorov-Smirnov test):

```
[h,p,stats]=lillietest(log_women_waist)

h = 1, p = 0, stats = 0.0841.
```

The same tests accept the hypotheses that other variables have lognormal distribution. Author's in [1] suggest that we can fit women's waist with Gamma distribution. Since Gamma





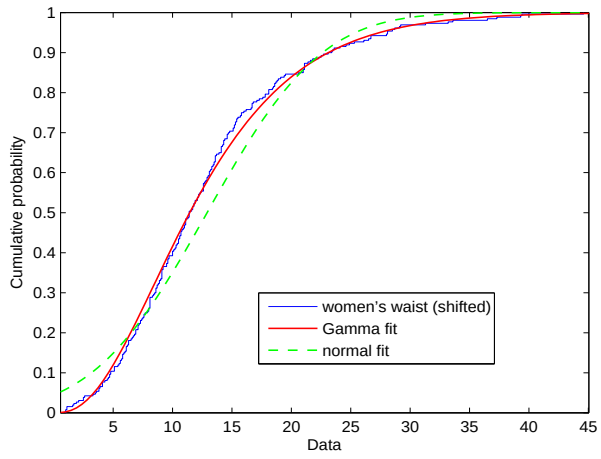


Figure 2.2: Fitting weight (upper left) and waist girth (upper right) with lognormal distribution. Lower left: fitting women's waist with shifted Gamma and normal distributions.

does not have a translation (shift) parameter, when we fit Gamma distribution we can either add to it a shift parameter or instead shift all data to start at zero. In figure 2.2 (c) we fit Gamma and, for the sake of illustration, normal distribution, to women's waist sample. As we can see, Gamma fits the data better than lognormal and much better than normal. To find the parameters of fitted Gamma distribution we use Matlab 'gamfit' function:

```
param=gamfit(women_waist_shift)

param = 2.8700    4.4960.
```

Chi-squared goodness-of-fit test for a *specific* (fitted) Gamma distribution:

```
[h,p,stats]=chi2gof(women_waist_shift,'cdf',@(z)gamcdf(z,param(1),param(2)))
```

```
h = 0, p = 0.9289, stats =  chi2stat: 2.4763, df: 7
```

accepts the hypothesis that the sample has Gamma distribution $\Gamma(2.87, 4.496)$. This test is not 'accurate' in some sense, which will be explained later. One can also check that Gamma distribution fits well other variables - men's waist girth, weight of men and weight of women.

□

Let us consider a family of distributions $\mathbb{P}_\theta$ indexed by a parameter (which could be a vector of parameters) θ that belongs to a set Θ. For example, we could consider a family of normal distributions $N(\alpha, \sigma^2)$ in which case the parameter would be $\theta = (\alpha, \sigma^2)$ - the mean and variance of the distribution. Let $f(X|\theta)$ be either a probability function (in case of discrete distribution) or a probability density function (continuous case) of the distribution $\mathbb{P}_\theta$. Suppose we are given an i.i.d. sample $X_1, \ldots, X_n$ with unknown distribution $\mathbb{P}_\theta$ from this family, i.e. parameter θ is unknown. A *likelihood function* is defined by

$$\varphi(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta).$$

We think of the sample $X_1, \ldots, X_n$ as given numbers and we think of φ as a function of the parameter θ only. The likelihood function has a clear interpretation. For example, if our distributions are discrete then the probability function

$$f(x|\theta) = \mathbb{P}_\theta(X = x)$$

is the probability to observe a point x and the likelihood function

$$\varphi(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta) = \mathbb{P}_\theta(X_1) \times \ldots \times \mathbb{P}_\theta(X_n) = \mathbb{P}_\theta(X_1, \ldots, X_n)$$

is the probability to observe the sample $X_1, \ldots, X_n$ when the parameters of the distribution are equal to θ. In the continuous case the likelihood function $\varphi(\theta)$ is the probability density function of the vector $(X_1, \ldots, X_n)$.

Definition: (*Maximum Likelihood Estimators.*) Suppose that there exists a parameter $\hat{\theta}$ that maximizes the likelihood function $\varphi(\theta)$ on the set of possible parameters Θ, i.e.

$$\varphi(\hat{\theta}) = \max_{\theta \in \Theta} \varphi(\theta).$$

Then $\hat{\theta}$ is called the Maximum Likelihood Estimator (MLE).

When finding the MLE it sometimes easier to maximize the log-likelihood function since

$$\varphi(\theta) \to \text{ maximize } \Leftrightarrow \log \varphi(\theta) \to \text{ maximize}$$

maximizing φ is equivalent to maximizing $\log \varphi$. Log-likelihood function can be written as

$$\log \varphi(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$

Let us give several examples of computing the MLE.

Example 1. Bernoulli distribution $B(p)$.

$$\mathcal{X} = \{0, 1\},\ \mathbb{P}(X = 1) = p,\ \mathbb{P}(X = 0) = 1 - p,\ p \in [0, 1].$$

Probability function in this case is given by

$$f(x|p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases} = p^x (1 - p)^{1-x}.$$

Likelihood function is

$$\begin{aligned} \varphi(p) &= f(X_1|p) f(X_2|p) \dots f(X_n|p) \\ &= p^{\#\text{ of 1's}} (1 - p)^{\#\text{ of 0's}} = p^{X_1 + \dots + X_n} (1 - p)^{n - (X_1 + \dots + X_n)} \end{aligned}$$

and the log-likelihood function is

$$\log \varphi(p) = (X_1 + \dots + X_n) \log p + (n - (X_1 + \dots + X_n)) \log(1 - p).$$

To maximize this over $p \in [0, 1]$ let us find the critical point $(\log \varphi(p))' = 0$,

$$(X_1 + \dots + X_n) \frac{1}{p} - (n - (X_1 + \dots + X_n)) \frac{1}{1 - p} = 0.$$

Solving this for p gives,

$$p = \frac{X_1 + \dots + X_n}{n} = \bar{X}$$

and, therefore, the proportion of successes $\hat{p} = \bar{X}$ in the sample is the MLEstimator of the unknown true probability of success, which is a very natural and intuitive estimator. For example, by law of large numbers, we know that

$$\bar{X} \to \mathbb{E} X_1 = p$$

in probability (we will recall this definition in the next lecture), which means that our estimate will approximate the unknown parameter p well when we get more and more data.

Remark. In each example, once we compute the estimate of parameters, we can try to prove directly, using the explicit form of the estimate, that it approximates well the unkown parameters, as we did in Example 1. However, in the next lecture we will describe in a general setting that MLE has 'good properties'.

Example 2. Normal distribution $N(\alpha, \sigma^2)$. The p.d.f. of normal distribution is

$$f(X|(\alpha, \sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X - \alpha)^2}{2\sigma^2}}.$$

and, therefore, likelihood function is

$$\varphi(\alpha, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i - \alpha)^2}{2\sigma^2}}.$$

and log-likelihood function is

$$\begin{aligned}\log\varphi(\alpha,\sigma^2) &= \sum_{i=1}^{n}\Big(\log\frac{1}{\sqrt{2\pi}}-\log\sigma-\frac{(X_i-\alpha)^2}{2\sigma^2}\Big)\\ &= n\log\frac{1}{\sqrt{2\pi}}-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\alpha)^2.\end{aligned}$$

We want to maximize the log-likelihood with respect to $-\infty<\alpha<\infty$ and $\sigma^2>0$. First, obviously, for any σ we need to minimize $\sum(X_i-\alpha)^2$ over α. The critical point condition is

$$\frac{d}{d\alpha}\sum_{i=1}^{n}(X_i-\alpha)^2=-2\sum_{i=1}^{n}(X_i-\alpha)=0$$

and solving this for α we get that $\hat\alpha=\bar X$. We can plug this estimate in the log-likelihood and it remains to maximize

$$n\log\frac{1}{\sqrt{2\pi}}-n\log\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\bar X)^2$$

over σ. The critical point condition reads,

$$-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum(X_i-\bar X)^2=0$$

and solving this for σ we obtain that the MLE of σ^2 is

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

The normal distribution fit in figure 2.1 corresponds to these parameters $(\hat\alpha,\hat\sigma^2)$.

Exercise. Generate a normal sample in Matlab and fit it with a normal distribution using 'dfittool'. Then plot a p.d.f. or c.d.f. corresponding to MLE above and compare this with 'dfittool'.

Let us give one more example of MLE.

Uniform distribution $U[0,\theta]$ on the interval $[0,\theta]$. This distribution has p.d.f.

$$f(x|\theta)=\begin{cases}\frac{1}{\theta}, & 0\le x\le\theta,\\ 0, & \text{otherwise.}\end{cases}$$

The likelihood function

$$\begin{aligned}\varphi(\theta)=\prod_{i=1}^{n}f(X_i|\theta) &= \frac{1}{\theta^n}I(X_1,\dots,X_n\in[0,\theta])\\ &= \frac{1}{\theta^n}I(\max(X_1,\dots,X_n)\le\theta).\end{aligned}$$

Here the indicator function $I(A)$ equals to 1 if event A happens and 0 otherwise. What the indicator above means is that the likelihood will be equal to 0 if at least one of the factors is 0 and this will happen if at least one observation X_i will fall outside of the 'allowed' interval $[0, \theta]$. Another way to say it is that the maximum among observations will exceed θ, i.e.

$$\varphi(\theta) = 0 \text{ if } \theta < \max(X_1, \ldots, X_n),$$

and

$$\varphi(\theta) = \frac{1}{\theta^n} \text{ if } \theta \geq \max(X_1, \ldots, X_n).$$

Therefore, looking at the figure 2.3 we see that $\hat{\theta} = \max(X_1, \ldots, X_n)$ is the MLE.

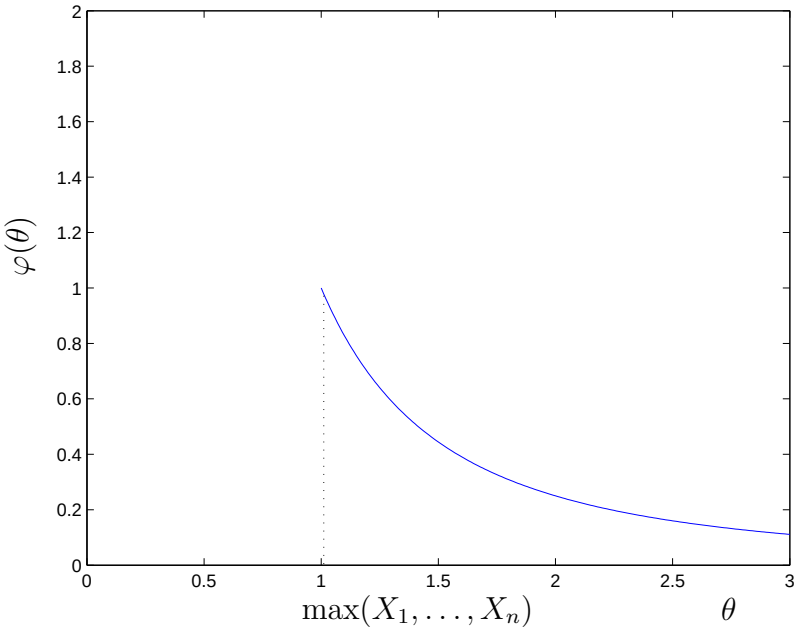


Figure 2.3: MLE for the uniform distribution.

Sometimes it is not so easy to find the maximum of the likelihood function as in the examples above and one might have to do it numerically. Also, MLE does not always exist. Here is an example: let us consider uniform distribution $U[0, \theta)$ and define the density by

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x < \theta, \\ 0, & \text{otherwise.} \end{cases}$$

The difference is that we 'excluded' the point θ by setting $f(\theta|\theta) = 0$. Then the likelihood function is

$$\varphi(\theta) = \prod_{i=1}^{n} f(X_i|\theta) = \frac{1}{\theta^n} I(\max(X_1, \ldots, X_n) < \theta)$$

and the maximum at the point $\hat{\theta} = \max(X_1, \ldots, X_n)$ is not achieved. Of course, this is an artificial example that shows that sometimes one needs to be careful.

References:

[1] Grete Heinz, Louis J. Peterson, Roger W. Johnson, Carter J. Kerk, (2003) "Exploring Relationships in Body Dimensions". *Journal of Statistics Education*, Volume 11, Number 2.

Lecture 3

Properties of MLE: consistency, asymptotic normality. Fisher information.

In this section we will try to understand why MLEs are 'good'.

Let us recall two facts from probability that we be used often throughout this course.

- **Law of Large Numbers (LLN):**

  If the distribution of the i.i.d. sample $X_1, \ldots, X_n$ is such that X_1 has finite expectation, i.e. $|\mathbb{E}X_1| < \infty$, then the sample average

  $$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n} \to \mathbb{E}X_1$$

  converges to its expectation *in probability* , which means that for any arbitrarily small $\varepsilon > 0$,

  $$\mathbb{P}(|\bar{X}_n - \mathbb{E}X_1| > \epsilon) \to 0 \text{ as } n \to \infty.$$

  Note. Whenever we will use the LLN below we will simply say that the average converges to its expectation and will not mention in what sense. More mathematically inclined students are welcome to carry out these steps more rigorously, especially when we use LLN in combination with the Central Limit Theorem.

- **Central Limit Theorem (CLT):**

  If the distribution of the i.i.d. sample $X_1, \ldots, X_n$ is such that X_1 has finite expectation and variance, i.e. $|\mathbb{E}X_1| < \infty$ and $\sigma^2 = \text{Var}(X) < \infty$, then

  $$\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \to^d N(0, \sigma^2)$$

  converges in distribution to normal distribution with zero mean and variance σ^2, which means that for any interval $[a, b]$,

  $$\mathbb{P}\Big(\sqrt{n}(\bar{X}_n - \mathbb{E}X_1) \in [a, b]\Big) \to \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx.$$

In other words, the random variable $\sqrt{n}(\bar{X}_n - \mathbb{E}X_1)$ will behave like a random variable from normal distribution when n gets large.

Exercise. Illustrate CLT by generating 100 Bernoulli random varibles $B(p)$ (or one Binomial r.v. $B(100, p)$) and then computing $\sqrt{n}(\bar{X}_n - \mathbb{E}X_1)$. Repeat this many times and use 'dfittool' to see that this random quantity will be well approximated by normal distribution.

We will prove that MLE satisfies (usually) the following two properties called *consistency* and *asymptotic normality*.

1. **Consistency.** We say that an estimate $\hat{\theta}$ is consistent if $\hat{\theta} \to \theta_0$ in probability as $n \to \infty$, where θ_0 is the 'true' unknown parameter of the distribution of the sample.

2. **Asymptotic Normality.** We say that $\hat{\theta}$ is asymptotically normal if

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{\theta_0})$$

where $\sigma^2_{\theta_0}$ is called the asymptotic variance of the estimate $\hat{\theta}$. Asymptotic normality says that the estimator not only converges to the unknown parameter, but it converges fast enough, at a rate $1/\sqrt{n}$.

Consistency of MLE.

To make our discussion as simple as possible, let us assume that a likelihood function is smooth and behaves in a nice way like shown in figure 3.1, i.e. its maximum is achieved at a unique point $\hat{\theta}$.

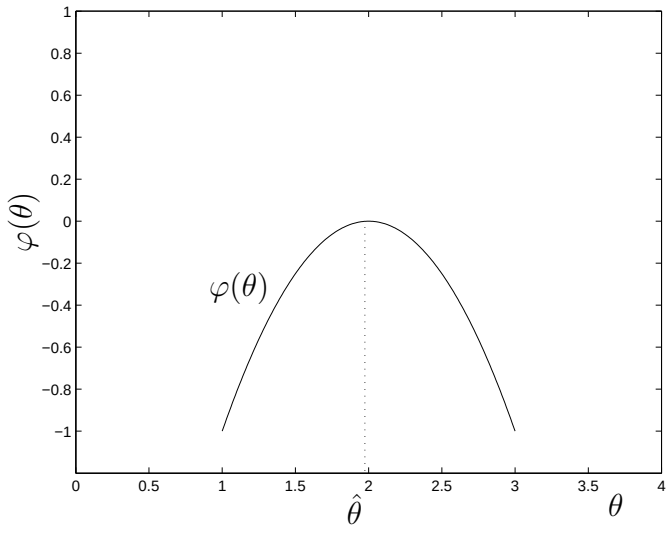


Figure 3.1: Maximum Likelihood Estimator (MLE)

Suppose that the data $X_1, \ldots, X_n$ is generated from a distribution with unknown parameter θ_0 and $\hat{\theta}$ is a MLE. Why $\hat{\theta}$ converges to the unknown parameter θ_0? This is not immediately obvious and in this section we will give a sketch of why this happens.

First of all, MLE $\hat{\theta}$ is the maximizer of

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i|\theta)$$

which is a log-likelihood function normalized by $\frac{1}{n}$ (of course, this does not affect maximization). Notice that function $L_n(\theta)$ depends on data. Let us consider a function $l(X|\theta) = \log f(X|\theta)$ and define

$$L(\theta) = \mathbb{E}_{\theta_0} l(X|\theta),$$

where $\mathbb{E}_{\theta_0}$ denotes the expectation with respect to the true uknown parameter θ_0 of the sample $X_1, \ldots, X_n$. If we deal with continuous distributions then

$$L(\theta) = \int (\log f(x|\theta)) f(x|\theta_0) dx.$$

By law of large numbers, for any θ,

$$L_n(\theta) \to \mathbb{E}_{\theta_0} l(X|\theta) = L(\theta).$$

Note that $L(\theta)$ does not depend on the sample, it only depends on θ. We will need the following

Lemma. *We have that for any* θ,

$$L(\theta) \le L(\theta_0).$$

Moreover, the inequality is strict, $L(\theta) < L(\theta_0)$, *unless*

$$\mathbb{P}_{\theta_0}(f(X|\theta) = f(X|\theta_0)) = 1.$$

which means that $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$.

Proof. Let us consider the difference

$$L(\theta) - L(\theta_0) = \mathbb{E}_{\theta_0}(\log f(X|\theta) - \log f(X|\theta_0)) = \mathbb{E}_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)}.$$

Since $\log t \le t - 1$, we can write

$$\begin{aligned} \mathbb{E}_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)} &\le \mathbb{E}_{\theta_0}\Big(\frac{f(X|\theta)}{f(X|\theta_0)} - 1\Big) = \int \Big(\frac{f(x|\theta)}{f(x|\theta_0)} - 1\Big) f(x|\theta_0) dx \\ &= \int f(x|\theta) dx - \int f(x|\theta_0) dx = 1 - 1 = 0. \end{aligned}$$

Both integrals are equal to 1 because we are integrating the probability density functions. This proves that $L(\theta) - L(\theta_0) \le 0$. The second statement of Lemma is also clear.

□

We will use this Lemma to sketch the consistency of the MLE.

Theorem: *Under some regularity conditions on the family of distributions, MLE $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to \theta_0$ as $n \to \infty$.*

The statement of this Theorem is not very precise but but rather than proving a rigorous mathematical statement our goal here is to illustrate the main idea. Mathematically inclined students are welcome to come up with some precise statement.

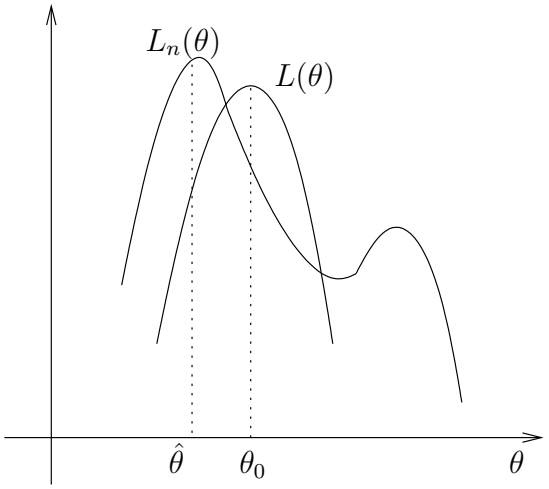


Figure 3.2: Illustration to Theorem.

Proof. We have the following facts:

1. $\hat{\theta}$ is the maximizer of $L_n(\theta)$ (by definition).

2. θ_0 is the maximizer of $L(\theta)$ (by Lemma).

3. $\forall \theta$ we have $L_n(\theta) \to L(\theta)$ by LLN.

This situation is illustrated in figure 3.2. Therefore, since two functions L_n and L are getting closer, the points of maximum should also get closer which exactly means that $\hat{\theta} \to \theta_0$.

□

Asymptotic normality of MLE. Fisher information.

We want to show the asymptotic normality of MLE, i.e. to show that

$$\sqrt{n}(\hat{\theta} - \theta_0) \to^d N(0, \sigma^2_{MLE}) \text{ for some } \sigma^2_{MLE}$$

and compute σ^2_{MLE}. This asymptotic variance in some sense measures the quality of MLE. First, we need to introduce the notion called Fisher Information.

Let us recall that above we defined the function $l(X|\theta) = \log f(X|\theta)$. To simplify the notations we will denote by $l'(X|\theta), l''(X|\theta)$, etc. the derivatives of $l(X|\theta)$ **with respect to** θ.

Definition. *(Fisher information.)* Fisher information of a random variable X with distribution $\mathbb{P}_{\theta_0}$ from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is defined by

$$I(\theta_0) = \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 \equiv \mathbb{E}_{\theta_0}\Big(\frac{\partial}{\partial\theta}\log f(X|\theta)\Big|_{\theta=\theta_0}\Big)^2.$$

Remark. Let us give a very informal interpretation of Fisher information. The derivative

$$l'(X|\theta_0) = (\log f(X|\theta_0))' = \frac{f'(X|\theta_0)}{f(X|\theta_0)}$$

can be interpreted as a measure of how quickly the distribution density or p.f. will change when we slightly change the parameter θ near θ_0. When we square this and take expectation, i.e. average over X, we get an averaged version of this measure. So if Fisher information is large, this means that the distribution will change quickly when we move the parameter, so the distribution with parameter θ_0 is 'quite different' and 'can be well distinguished' from the distributions with parameters not so close to θ_0. This means that we should be able to estimate θ_0 well based on the data. On the other hand, if Fisher information is small, this means that the distribution is 'very similar' to distributions with parameter not so close to θ_0 and, thus, more difficult to distinguish, so our estimation will be worse. We will see precisely this behavior in Theorem below.

Next lemma gives another often convenient way to compute Fisher information.

Lemma. *We have,*

$$\mathbb{E}_{\theta_0} l''(X|\theta_0) \equiv \mathbb{E}_{\theta_0} \frac{\partial^2}{\partial\theta^2} \log f(X|\theta_0) = -I(\theta_0).$$

Proof. First of all, we have

$$l'(X|\theta) = (\log f(X|\theta))' = \frac{f'(X|\theta)}{f(X|\theta)}$$

and

$$(\log f(X|\theta))'' = \frac{f''(X|\theta)}{f(X|\theta)} - \frac{(f'(X|\theta))^2}{f^2(X|\theta)}.$$

Also, since p.d.f. integrates to 1,

$$\int f(x|\theta)dx = 1,$$

if we take derivatives of this equation with respect to θ (and interchange derivative and integral, which can usually be done) we will get,

$$\int \frac{\partial}{\partial\theta} f(x|\theta)dx = 0 \text{ and } \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx = \int f''(x|\theta)dx = 0.$$

To finish the proof we write the following computation

$$\begin{aligned}
\mathbb{E}_{\theta_0} l''(X|\theta_0) &= \mathbb{E}_{\theta_0} \frac{\partial^2}{\partial\theta^2} \log f(X|\theta_0) = \int (\log f(x|\theta_0))'' f(x|\theta_0)dx \\
&= \int \Big(\frac{f''(x|\theta_0)}{f(x|\theta_0)} - \Big(\frac{f'(x|\theta_0)}{f(x|\theta_0)} \Big)^2 \Big) f(x|\theta_0)dx \\
&= \int f''(x|\theta_0)dx - \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 = 0 - I(\theta_0 = -I(\theta_0).
\end{aligned}$$

□

We are now ready to prove the main result of this section.

Theorem. *(Asymptotic normality of MLE.) We have,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \to N\Big(0, \frac{1}{I(\theta_0)}\Big).$$

As we can see, the asymptotic variance/dispersion of the estimate around true parameter will be smaller when Fisher information is larger.

Proof. Since MLE $\hat{\theta}$ is maximizer of $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$, we have

$$L_n'(\hat{\theta}) = 0.$$

Let us use the Mean Value Theorem

$$\frac{f(a) - f(b)}{a - b} = f'(c) \text{ or } f(a) = f(b) + f'(c)(a - b) \text{ for } c \in [a, b]$$

with $f(\theta) = L_n'(\theta), a = \hat{\theta}$ and $b = \theta_0$. Then we can write,

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\hat{\theta}_1)(\hat{\theta} - \theta_0)$$

for some $\hat{\theta}_1 \in [\hat{\theta}, \theta_0]$. From here we get that

$$\hat{\theta} - \theta_0 = -\frac{L_n'(\theta_0)}{L_n''(\hat{\theta}_1)} \text{ and } \sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta}_1)}. \tag{3.0.1}$$

Since by Lemma in the previous section we know that θ_0 is the maximizer of $L(\theta)$, we have

$$L'(\theta_0) = \mathbb{E}_{\theta_0} l'(X|\theta_0) = 0. \tag{3.0.2}$$

Therefore, the numerator in (3.0.1)

$$\begin{aligned} \sqrt{n}L_n'(\theta_0) &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n l'(X_i|\theta_0) - 0\Big) \qquad (3.0.3)\\ &= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n l'(X_i|\theta_0) - \mathbb{E}_{\theta_0} l'(X_1|\theta_0)\Big) \to N\Big(0, \mathrm{Var}_{\theta_0}(l'(X_1|\theta_0))\Big) \end{aligned}$$

converges in distribution by Central Limit Theorem.

Next, let us consider the denominator in (3.0.1). First of all, we have that for all θ,

$$L_n''(\theta) = \frac{1}{n}\sum l''(X_i|\theta) \to \mathbb{E}_{\theta_0} l''(X_1|\theta) \text{ by LLN.} \tag{3.0.4}$$

Also, since $\hat{\theta}_1 \in [\hat{\theta}, \theta_0]$ and by consistency result of previous section, $\hat{\theta} \to \theta_0$, we have $\hat{\theta}_1 \to \theta_0$. Using this together with (10.0.3) we get

$$L_n''(\hat{\theta}_1) \to \mathbb{E}_{\theta_0} l''(X_1|\theta_0) = -I(\theta_0) \text{ by Lemma above.}$$

Combining this with (3.0.3) we get

$$-\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\hat{\theta}_1)} \to^d N\Big(0, \frac{\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0))}{(I(\theta_0))^2}\Big).$$

Finally, the variance,

$$\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0)) = \mathbb{E}_{\theta_0}(l'(X|\theta_0))^2 - (\mathbb{E}_{\theta_0}l'(x|\theta_0))^2 = I(\theta_0) - 0$$

where in the last equality we used the definition of Fisher information and (3.0.2).

□

Let us compute Fisher information for some particular distributions.

Example 1. The family of Bernoulli distributions $B(p)$ has p.f.

$$f(x|p) = p^x(1-p)^{1-x}$$

and taking the logarithm

$$\log f(x|p) = x\log p + (1-x)\log(1-p).$$

The second derivative with respect to parameter p is

$$\frac{\partial}{\partial p}\log f(x|p) = \frac{x}{p} - \frac{1-x}{1-p}, \;\; \frac{\partial^2}{\partial p^2}\log f(x|p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

Then the Fisher information can be computed as

$$I(p) = -\mathbb{E}\frac{\partial^2}{\partial p^2}\log f(X|p) = \frac{\mathbb{E}X}{p^2} + \frac{1-\mathbb{E}X}{(1-p)^2} = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}.$$

The MLE of p is $\hat{p} = \bar{X}$ and the asymptotic normality result states that

$$\sqrt{n}(\hat{p} - p_0) \to N(0, p_0(1-p_0))$$

which, of course, also follows directly from the CLT.

Example. The family of exponential distributions $E(\alpha)$ has p.d.f.

$$f(x|\alpha) = \begin{cases} \alpha e^{-\alpha x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

and, therefore,

$$\log f(x|\alpha) = \log\alpha - \alpha x \Rightarrow \frac{\partial^2}{\partial\alpha^2}\log f(x|\alpha) = -\frac{1}{\alpha^2}.$$

This does not depend on X and we get

$$I(\alpha) = -\mathbb{E}\frac{\partial^2}{\partial\alpha^2}\log f(X|\alpha) = \frac{1}{\alpha^2}.$$

Therefore, the MLE $\hat{\alpha} = 1/\bar{X}$ is asymptotically normal and

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \to N(0, \alpha_0^2).$$

□

Lecture 4

Multivariate normal distribution and multivariate CLT.

We start with several simple observations. If $X = (x_1, \dots, x_k)^T$ is a $k \times 1$ random vector then its expectation is

$$\mathbb{E}X = (\mathbb{E}x_1, \dots, \mathbb{E}x_k)^T$$

and its covariance matrix is

$$\mathrm{Cov}(X) = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T.$$

Notice that a covariance matrix is always symmetric

$$\mathrm{Cov}(X)^T = \mathrm{Cov}(X)$$

and nonnegative definite, i.e. for any $k \times 1$ vector a,

$$a^T \mathrm{Cov}(X) a = \mathbb{E}a^T(X - \mathbb{E}X)(X - \mathbb{E}X)^T a^T = \mathbb{E}|a^T(X - \mathbb{E}X)|^2 \geq 0.$$

We will often use that for any vector X its squared length can be written as $|X|^2 = X^T X$. If we multiply a random $k \times 1$ vector X by a $n \times k$ matrix A then the covariance of $Y = AX$ is a $n \times n$ matrix

$$\mathrm{Cov}(Y) = \mathbb{E}A(X - \mathbb{E}X)(X - \mathbb{E}X)^T A^T = A\mathrm{Cov}(X)A^T.$$

Multivariate normal distribution. Let us consider a $k \times 1$ vector $g = (g_1, \dots, g_k)^T$ of i.i.d. standard normal random variables. The covariance of g is, obviously, a $k \times k$ identity matrix, $\mathrm{Cov}(g) = I$. Given a $n \times k$ matrix A, the covariance of Ag is a $n \times n$ matrix

$$\Sigma := \mathrm{Cov}(Ag) = AIA^T = AA^T.$$

Definition. *The distribution of a vector Ag is called a (multivariate) normal distribution with covariance Σ and is denoted $N(0, \Sigma)$.*

One can also shift this disrtibution, the distribution of $Ag + a$ is called a normal distribution with mean a and covariance Σ and is denoted $N(a, \Sigma)$. There is one potential problem

with the above definition - we assume that the distribution depends only on covariance matrix Σ and does not depend on the construction, i.e. the choice of g and a matrix A. For example, if we take a $m \times 1$ vector g' of i.i.d. standard normal random variables and a $n \times m$ matrix B then the covariance of Bg' is a $n \times n$ matrix

$$\mathrm{Cov}(Bg') = BB^T.$$

It is possible that $\Sigma = AA^T = BB^T$ so both constructions should give a normal distribution $N(0, \Sigma)$. This is, indeed, true - the distribution of Ag and Bg' is the same, so the definition of normal distribution $N(0, \Sigma)$ does not depend on the construction. It is not very difficult to prove that Ag and Bg' have the same distribution, but we will only show the simplest case.

Invertible case. Suppose that A and B are both square $n \times n$ invertible matrices. In this case, vectors Ag and Bg' have density which we will now compute. Since the density of g is

$$\prod_{i \le n} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x_i^2\Big) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2}|x|^2\Big),$$

for any set $\Omega \in \mathbb{R}^n$ we can write

$$\mathbb{P}(Ag \in \Omega) = \mathbb{P}(g \in A^{-1}\Omega) = \int_{A^{-1}\Omega} \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2}|x|^2\Big) dx.$$

Let us now make the change of variables $y = Ax$ or $x = A^{-1}y$. Then

$$\mathbb{P}(Ag \in \Omega) = \int_{\Omega} \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2}|A^{-1}y|^2\Big) \frac{1}{|\det(A)|} dy.$$

But since

$$\det(\Sigma) = \det(AA^T) = \det(A)\det(A^T) = \det(A)^2$$

we have $|\det(A)| = \sqrt{\det(\Sigma)}$. Also

$$|A^{-1}y|^2 = (A^{-1}y)^T(A^{-1}y) = y^T(A^T)^{-1}A^{-1}y = y^T(AA^T)^{-1}y = y^T\Sigma^{-1}y.$$

Therefore, we get

$$\mathbb{P}(Ag \in \Omega) = \int_{\Omega} \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \frac{1}{\sqrt{\det(\Sigma)}} \exp\Big(-\frac{1}{2}y^T\Sigma^{-1}y\Big) dy.$$

This means that a vector Ag has the density

$$\frac{1}{\sqrt{\det(\Sigma)}} \exp\Big(-\frac{1}{2}y^T\Sigma^{-1}y\Big)$$

which depends only on Σ and not on A. This means that Ag and Bg' have the same distributions.

□

It is not difficult to show that in a general case the distribution of Ag depends only on the covariance Σ, but we will omit this here. Many times in these lectures whenever we want to represent a normal distribution $N(0,\Sigma)$ constructively, we will find a matrix A (not necessarily square) such that $\Sigma = AA^T$ and use the fact that a vector Ag for i.i.d. vector g has normal distribution $N(0,\Sigma)$. One way to find such A is to take a matrix square-root of Σ. Since Σ is a symmetric nonnegative definite matrix, its eigenvalue decomposition is

$$\Sigma = QDQ^T$$

for an orthogonal matrix Q and a diagonal matrix D with eigenvalues $\lambda_1, \ldots, \lambda_n$ of Σ on the diagonal. In Matlab, '[Q,D]=eig(Sigma);' will produce this decomposition. Then if $D^{1/2}$ represents a diagonal matrix with $\lambda_i^{1/2}$ on the diagonal then one can take

$$A = QD^{1/2} \quad \text{or} \quad A = QD^{1/2}Q^T.$$

It is easy to check that in both cases $AA^T = QDQ^T = \Sigma$. In Matlab $QD^{1/2}Q^T$ is given by 'sqrtm(Sigma)'. Let us take, for example, a vector $X = QD^{1/2}g$ for i.i.d. standard normal vector g which by definition has normal distribution $N(0,\Sigma)$. If $q_1, \ldots, q_n$ are the column vectors of Q then

$$X = QD^{1/2}g = (\lambda_1^{1/2}g_1)q_1 + \ldots + (\lambda_n^{1/2}g_n)q_n.$$

Therefore, in the orthonormal coordinate basis $q_1, \ldots, q_n$ a random vector X has coordinates $\lambda_1^{1/2}g_1, \ldots, \lambda_n^{1/2}g_n$. These coordinates are independent with normal distributions with variances $\lambda_1, \ldots, \lambda_n$ correspondingly. When $\det\Sigma = 0$, i.e. Σ is not invertible, some of its eigenvalues will be zero, say, $\lambda_{k+1} = \ldots = \lambda_n = 0$. Then the random vector will be concentrated on the subspace spanned by vectors $q_1, \ldots, q_k$ but it will not have density on the entire space $\mathbb{R}^n$. On the subspace spanned by vectors $q_1, \ldots, q_k$ a vector X will have a density

$$f(x_1, \ldots, x_k) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\Big(-\frac{x_i^2}{2\lambda_i}\Big).$$

Linear transformation of a normal random vector.

Suppose that Y is a $n \times 1$ random vector with normal distribution $N(0,\Sigma)$. Then given a $m \times n$ matrix M, a $m \times 1$ vector MY will also have normal distribution $N(0, M\Sigma M^T)$. To show this, find any matrix A and i.i.d. standard normal vector g such that Ag has normal distribution $N(0,\Sigma)$. Then, by definition, $M(Ag) = (MA)g$ also has normal distribution with covariance

$$(MA)(MA)^T = MAA^TM^T = M\Sigma M^T.$$

□

Orthogonal transformation of an i.i.d. standard normal sample.

Throughout the lectures we will often use the following simple fact. Consider a vector $X = (X_1, \ldots, X_n)^T$ of i.i.d. random variables with standard normal distribution $N(0,1)$. If V is an orthogonal $n \times n$ matrix then the vector $Y := VX$ also consists of i.i.d. random

variables $Y_1, \ldots, Y_n$ with standard normal distribution. A matrix V is orthogonal when one of the following equivalent properties hold:

1. $V^{-1} = V^T$.
2. The rows of V form an orthonormal basis in $\mathbb{R}^n$.
3. The columns of V form an orthonormal basis in $\mathbb{R}^n$.
4. For any $x \in \mathbb{R}^n$ we have $|Vx| = |x|$, i.e. V preserves the lengths of vectors.

Below we will use that $|\det(V)| = 1$. Basically, orthogonal transformations represent linear transformations that preserve distances between points, such as rotations and reflections. The joint p.d.f of a vector X is given by

$$f(x) = f(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(\sqrt{2\pi})^n} e^{-|x|^2/2},$$

where $|x|^2 = x_1^2 + \ldots + x_n^2$. To find the p.d.f. of a vector $Y = VX$, which is an linear transformation of X, we can use the change of density formula from probability or the change of variables formula from calculus as follows. For any set $\Omega \subseteq \mathbb{R}^n$,

$$\begin{aligned} \mathbb{P}(Y \in \Omega) &= \mathbb{P}(VX \in \Omega) = \mathbb{P}(X \in V^{-1}\Omega) \\ &= \int_{V^{-1}\Omega} f(x)dx = \int_{\Omega} \frac{f(V^{-1}y)}{|\det(V)|} dy. \end{aligned}$$

where we made the change of variables $y = Vx$. We know that $|\det(V)| = 1$ and, since $|V^{-1}y| = |y|$, we have

$$f(V^{-1}y) = \frac{1}{(\sqrt{2\pi})^n} e^{-|V^{-1}y|^2/2} = \frac{1}{(\sqrt{2\pi})^n} e^{-|y|^2/2} = f(y).$$

Therefore, we finally get that

$$\mathbb{P}(Y \in \Omega) = \int_{\Omega} f(y)dy$$

which proves that a vector Y has the same joint p.d.f. as X.

□

Multivariate CLT.

We will state a multivariate Central Limit Theorem without a proof. Suppose that $X = (x_1, \ldots, x_k)^T$ is a random vector with covariance Σ. We assumed that $\mathbb{E}x_i^2 < \infty$. If $X_1, X_2, \ldots$ is a sequence of i.i.d. copies of X then

$$S_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \to^d N(0, \Sigma),$$

where convergence in distribution $\to^d$ means that for any set $\Omega \in \mathbb{R}^k$,

$$\lim_{n\to\infty} \mathbb{P}(S_n \in \Omega) = \mathbb{P}(Y \in \Omega)$$

for a random vector Y with normal distribution $N(0, \Sigma)$.

□

Lecture 5

Confidence intervals for parameters of normal distribution.

Let us consider a Matlab example based on the dataset of body temperature measurements of 130 individuals from the article [1]. The dataset can be downloaded from the journal's website. This dataset was derived from the article [2]. First of all, if we use 'dfittool' to fit a normal distribution to this data we get a pretty good approximation, see figure 5.1.

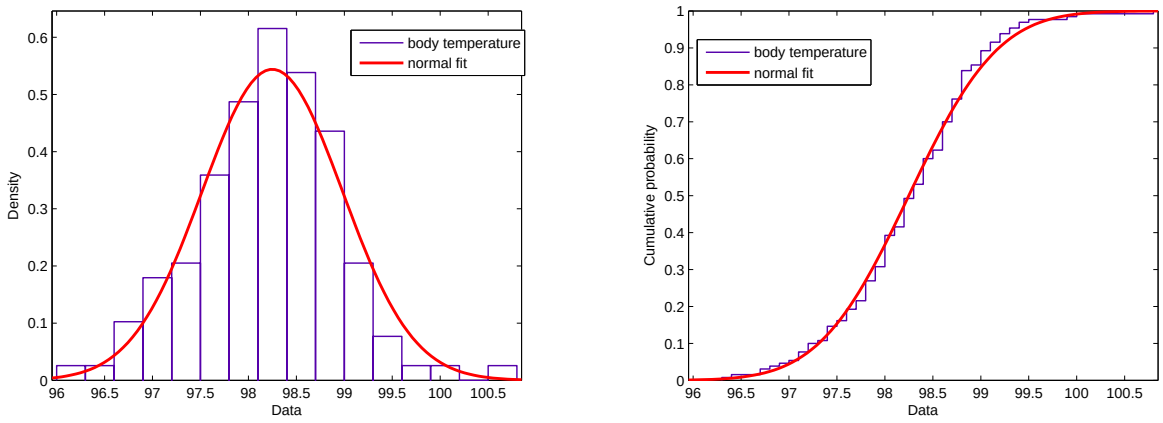


Figure 5.1: Fitting a body temperature dataset. (a) Histogram of the data and p.d.f. of fitted normal distribution; (b) Empirical c.d.f. and c.d.f. of fitted normal distribution.

The tool also outputs the following MLEstimates $\hat{\mu}$ and $\hat{\sigma}$ of parameters μ, σ of normal distribution:

```
Parameter  Estimate  Std. Err.
mu          98.2492  0.0643044
sigma      0.733183  0.0457347.
```

Also, if our dataset vector name is 'normtemp' then using the matlab function 'normfit' by typing '[mu,sigma,muint,sigmaint]=normfit(normtemp)' outputs the following:

```
mu = 98.2492, sigma = 0.7332,
muint = [98.122, 98.376], sigmaint = [0.654, 0.835].
```

The last two intervals here are 95% *confidence intervals* for parameters μ and σ. This means that not only we are able to estimate the parameters of normal distribution using MLE but also to garantee with confidence 95% that the 'true' unknown parameters of the distribution belong to these confidence intervals. How this is done is the topic of this lecture. Notice that conventional 'normal' temperature 98.6 does not fall into the estimated 95% confidence interval $[98.122, 98.376]$.

Distribution of the estimates of parameters of normal distribution.

Let us consider a sample

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$$

from normal distribution with mean μ and variance σ^2. MLE gave us the following estimates of μ and σ^2 - $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \bar{X^2} - (\bar{X})^2$. The question is: how close are these estimates to actual values of the unknown parameters μ and σ^2? By LLN we know that these estimates converge to μ and σ^2,

$$\bar{X} \to \mu, \bar{X^2} - (\bar{X})^2 \to \sigma^2, n \to \infty,$$

but we will try to describe precisely how close $\bar{X}$ and $\bar{X^2} - (\bar{X})^2$ are to μ and σ^2. We will start by studying the following question:

What is the joint distribution of $(\bar{X}, \bar{X^2} - (\bar{X})^2)$ *when* $X_1, \ldots, X_n$ *are i.i.d from* $N(0,1)$*?*

A similar question for a sample from a general normal distribution $N(\mu, \sigma^2)$ can be reduced to this one by renormalizing $Z_i = (X_i - \mu)/\sigma$. We will need the following definition.

Definition. *If* $X_1, \ldots, X_n$ *are i.i.d. standard normal then the distribution of*

$$X_1^2 + \ldots + X_n^2$$

is called the χ^2_n*-distribution (chi-squared distribution) with* n *degrees of freedom.*

We will find the p.d.f. of this distribution in the following lectures. At this point we only need to note that this distribution does not depend on any parameters besides degrees of freedom n and, therefore, could be tabulated even if we were not able to find the explicit formula for its p.d.f. Here is the main result that will allow us to construct confidence intervals for parameters of normal distribution as in the Matlab example above.

Theorem. *If* $X_1, \ldots, X_n$ *are i.d.d. standard normal, then sample mean* $\bar{X}$ *and sample variance* $\bar{X^2} - (\bar{X})^2$ *are independent,*

$$\sqrt{n}\bar{X} \sim N(0,1) \text{ and } n(\bar{X^2} - (\bar{X})^2) \sim \chi^2_{n-1},$$

i.e. $\sqrt{n}\bar{X}$ *has standard normal distribution and* $n(\bar{X^2} - (\bar{X})^2)$ *has* χ^2_{n-1} *distribution with* $(n-1)$ *degrees of freedom.*

Proof. Consider a vector Y given by a specific orthogonal transformation of X:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = VX = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \vdots & ? & \vdots \\ \cdots & \cdots & \cdots \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

Here we choose a first row of the matrix V to be equal to

$$v_1 = \Big(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Big)$$

and let the remaining rows be any vectors such that the matrix V defines orthogonal transformation. This can be done since the length of the first row vector $|v_1| = 1$, and we can simply choose the rows $v_2, \ldots, v_n$ to be any orthogonal basis in the hyperplane orthogonal to vector v_1.

Let us discuss some properties of this particular transformation. First of all, we showed above that $Y_1, \ldots, Y_n$ are also i.i.d. standard normal. Because of the particular choice of the first row v_1 in V, the first r.v.

$$Y_1 = \frac{1}{\sqrt{n}} X_1 + \ldots + \frac{1}{\sqrt{n}} X_n = \sqrt{n}\bar{X}$$

and, therefore,

$$\bar{X} = \frac{1}{\sqrt{n}} Y_1. \tag{5.0.1}$$

Next, n times sample variance can be written as

$$\begin{aligned} n(\bar{X^2} - (\bar{X})^2) &= X_1^2 + \ldots + X_n^2 - \Big(\frac{1}{\sqrt{n}}(X_1 + \ldots + X_n)\Big)^2 \\ &= X_1^2 + \ldots + X_n^2 - Y_1^2. \end{aligned}$$

The orthogonal transformation V preserves the length of X, i.e. $|Y| = |VX| = |X|$ or

$$Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2$$

and, therefore, we get

$$n(\bar{X^2} - (\bar{X})^2) = Y_1^2 + \ldots + Y_n^2 - Y_1^2 = Y_2^2 + \ldots + Y_n^2. \tag{5.0.2}$$

Equations (5.0.1) and (5.0.2) show that sample mean and sample variance are independent since Y_1 and $(Y_2, \ldots, Y_n)$ are independent, $\sqrt{n}\bar{X} = Y_1$ has standard normal distribution and $n(\bar{X^2} - (\bar{X})^2)$ has χ^2_{n-1} distribution since $Y_2, \ldots, Y_n$ are independent standard normal.

□

Let us write down the implications of this result for a general normal distribution:

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2).$$

In this case, we know that

$$Z_1 = \frac{X_1 - \mu}{\sigma}, \cdots, Z_n = \frac{X_n - \mu}{\sigma} \sim N(0,1)$$

are independent standard normal. Theorem applied to $Z_1, \ldots, Z_n$ gives that

$$\sqrt{n}\bar{Z} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

and

$$\begin{aligned} n(\bar{Z^2} - (\bar{Z})^2) &= n\Big(\frac{1}{n}\sum\Big(\frac{X_i - \mu}{\sigma}\Big)^2 - \Big(\frac{1}{n}\sum\frac{X_i - \mu}{\sigma}\Big)^2\Big) \\ &= n\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{X_i - \mu}{\sigma} - \frac{1}{n}\sum\frac{X_i - \mu}{\sigma}\Big)^2 \\ &= n\frac{\bar{X^2} - (\bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}. \end{aligned}$$

We proved that MLE $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \bar{X^2} - (\bar{X})^2$ are independent and

$$\boxed{\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0,1), \;\; \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}.}$$

Confidence intervals for parameters of normal distribution.

We know that by LLN a sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$ converge to mean μ and variance σ^2:

$$\hat{\mu} = \bar{X} \to \mu, \hat{\sigma}^2 = \bar{X^2} - (\bar{X})^2 \to \sigma^2.$$

In other words, these estimates are consistent. Based on the above description of the joint distribution of the estimates, we will give a precise quantitative description of how close $\hat{\mu}$ and $\hat{\sigma}^2$ are to the unknown parameters μ and σ^2.

Let us start by giving a definition of a *confidence interval* in our usual setting when we observe a sample $X_1, \ldots, X_n$ with distribution $\mathbb{P}_{\theta_0}$ from a parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, and θ_0 is unknown.

Definition: Given a *confidence level* parameter $\alpha \in [0,1]$, if there exist two statistics

$$S_1 = S_1(X_1, \ldots, X_n) \text{ and } S_2 = S_2(X_1, \ldots, X_n)$$

such that probability

$$\mathbb{P}_{\theta_0}(S_1 \le \theta_0 \le S_2) = \alpha \quad (\text{ or } \ge \alpha)$$

then we will call $[S_1, S_2]$ a *confidence interval* for the unknown parameter θ_0 with the confidence level α.

This definition means that we can garantee with probability/confidence α that our unknown parameter lies within the interval $[S_1, S_2]$. We will now show how in the case of a normal distribution $N(\mu, \sigma^2)$ we can construct confidence intervals for unknown μ and σ^2. Let us recall that in the last lecture we proved that if

$$X_1, \ldots, X_n \text{ are i.d.d. with distribution } N(\mu, \sigma^2)$$

then

$$A = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0,1) \text{ and } B = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

and the random variables A and B are independent. If we recall the definition of χ^2-distribution, this means that we can represent A and B as

$$A = Y_1 \text{ and } B = Y_2^2 + \ldots + Y_n^2$$

for some $Y_1, \ldots, Y_n$ - i.d.d. standard normal.

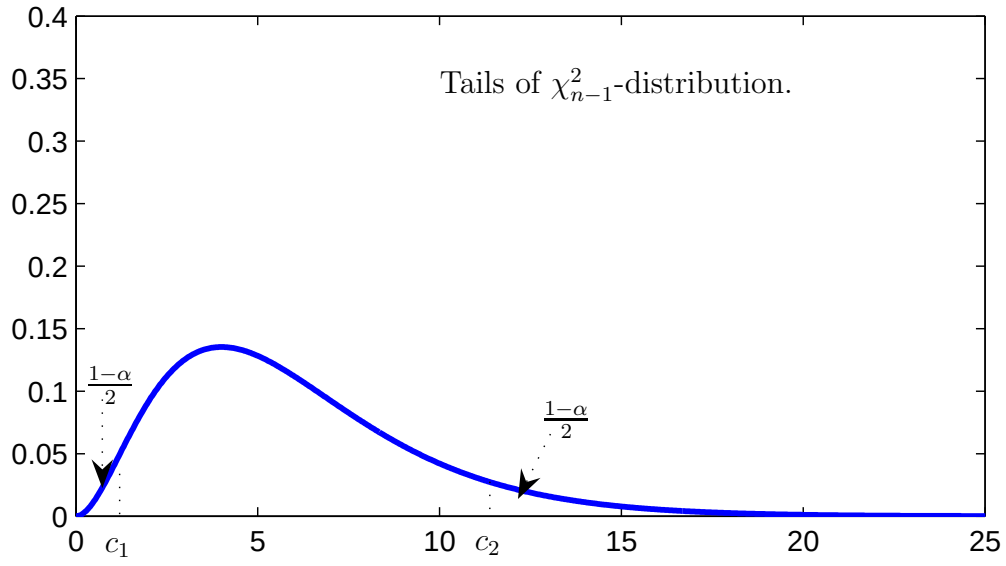


Figure 5.2: p.d.f. of χ^2_{n-1}-distribution and α-confidence interval.

First, let us consider p.d.f. of χ^2_{n-1} distribution (see figure 5.2) and choose points c_1 and c_2 so that the area in each tail is $(1-\alpha)/2$. Then the area between c_1 and c_2 is α which means that

$$\mathbb{P}(c_1 \leq B \leq c_2) = \alpha.$$

Therefore, we can 'garantee' with probability α that

$$c_1 \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq c_2.$$

Solving this for σ^2 gives

$$\frac{n\hat{\sigma}^2}{c_2} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{c_1}.$$

This precisely means that the interval

$$\Big[\frac{n\hat{\sigma}^2}{c_2}, \frac{n\hat{\sigma}^2}{c_1}\Big]$$

is the α-confidence interval for the unknown variance σ^2.

Next, let us construct the confidence interval for the mean μ. We will need the following definition.

Definition. If $Y_0, Y_1, \ldots, Y_n$ are i.i.d. standard normal then the distribution of the random variable

$$\frac{Y_0}{\sqrt{\frac{1}{n}(Y_1^2 + \ldots + Y_n^2)}}$$

is called (Student) t_n-distribution with n degrees of freedom.

We will find the p.d.f. of this distribution in the following lectures together with p.d.f. of χ^2-distribution and some others. At this point we only note that this distribution does not depend on any parameters besides degrees of freedom n and, therefore, it can be tabulated. Consider the following expression:

$$\frac{A}{\sqrt{\frac{1}{n-1}B}} = \frac{Y_1}{\sqrt{\frac{1}{n-1}(Y_2^2 + \ldots + Y_n^2)}} \sim t_{n-1}$$

which, by definition, has t_{n-1}-distribution with $n-1$ degrees of freedom. On the other hand,

$$\frac{A}{\sqrt{\frac{1}{n-1}B}} = \sqrt{n}\frac{(\hat{\mu}-\mu)}{\sigma}\Big/\sqrt{\frac{1}{n-1}\frac{n\hat{\sigma}^2}{\sigma^2}} = \frac{\sqrt{n-1}}{\hat{\sigma}}(\hat{\mu}-\mu).$$

If we now look at the p.d.f. of t_{n-1} distribution (see figure 5.3) and choose the constants $-c$ and c so that the area in each tail is $(1-\alpha)/2$, (the constant is the same on each side because the distribution is symmetric) we get that with probability α,

$$-c \le \frac{\sqrt{n-1}}{\hat{\sigma}}(\hat{\mu}-\mu) \le c$$

and solving this for μ, we get the confidence interval

$$\hat{\mu} - c\frac{\hat{\sigma}}{\sqrt{n-1}} \le \mu \le \hat{\mu} + c\frac{\hat{\sigma}}{\sqrt{n-1}}.$$

Example. (Textbook, Section 7.5, p. 411)) Consider a sample of size $n = 10$ from normal distribution with unknown parameters:

$$0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58.$$

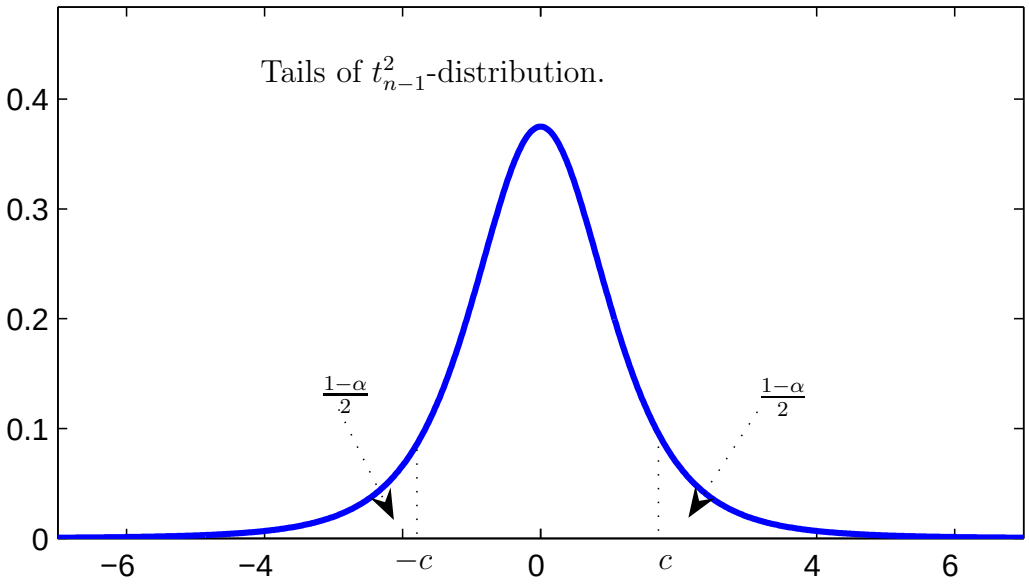


Figure 5.3: p.d.f. of t_{n-1} distribution and confidence interval for μ.

We compute the estimates

$$\hat{\mu} = \bar{X} = 1.379 \text{ and } \hat{\sigma}^2 = \bar{X^2} - (\bar{X})^2 = 0.0966.$$

Let us choose confidence level $\alpha = 95\% = 0.95$. We have to find c_1, c_2 and c as explained above. Using the table for t_9-distribution we need to find c such that

$$t_9(-\infty, c) = 0.975$$

which gives us $c = 2.262$. To find c_1 and c_2 we have to use the χ^2_9-distribution table so that

$$\chi^2_9([0, c_1]) = 0.025 \Rightarrow c_1 = 2.7$$

$$\chi^2_9([0, c_2]) = 0.975 \Rightarrow c_2 = 19.02.$$

Plugging these into the formulas above, with probability 95% we can garantee that

$$\begin{aligned} \bar{X} - c\sqrt{\frac{1}{9}(\bar{X^2} - (\bar{X})^2)} \leq \quad & \mu \quad \leq \bar{X} + c\sqrt{\frac{1}{9}(\bar{X^2} - (\bar{X})^2)} \\ 1.1446 \leq \quad & \mu \quad \leq 1.6134 \end{aligned}$$

and with probability 95% we can garantee that

$$\frac{n(\bar{X^2} - (\bar{X})^2)}{c_2} \leq \sigma^2 \leq \frac{n(\bar{X^2} - (\bar{X})^2)}{c_1}$$

or

$$0.0508 \leq \sigma^2 \leq 0.3579.$$

These confidence intervals may not look impressive but the sample size is very small here, $n = 10$.

References.

[1] Allen L .Shoemaker (1996), "What's Normal? - Temperature, Gender, and Heart Rate". *Journal of Statistics Education,* v.4, n.2.

[2] Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich". *Journal of the American Medical Association*, 268, 1578-1580.

Lecture 6

Gamma distribution, χ^2-distribution, Student t-distribution, Fisher F-distribution.

Gamma distribution. Let us take two parameters $\alpha > 0$ and $\beta > 0$. Gamma function $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

If we divide both sides by $\Gamma(\alpha)$ we get

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} dy$$

where we made a change of variables $x = \beta y$. Therefore, if we define

$$f(x|\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

then $f(x|\alpha,\beta)$ will be a probability density function since it is nonnegative and it integrates to one.

Definition. The distribution with p.d.f. $f(x|\alpha,\beta)$ is called Gamma distribution with parameters α and β and it is denoted as $\Gamma(\alpha,\beta)$.

Next, let us recall some properties of gamma function $\Gamma(\alpha)$. If we take $\alpha > 1$ then using integration by parts we can write:

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} e^{-x} dx = \int_0^\infty x^{\alpha-1} d(-e^{-x}) \\ &= x^{\alpha-1}(-e^{-x})\Big|_0^\infty - \int_0^\infty (-e^{-x})(\alpha-1)x^{\alpha-2} dx \\ &= (\alpha-1)\int_0^\infty x^{(\alpha-1)-1} e^{-x} dx = (\alpha-1)\Gamma(\alpha-1). \end{aligned}$$

Since for $\alpha = 1$ we have

$$\Gamma(1) = \int_0^\infty e^{-x} dx = 1$$

we can write

$$\Gamma(2) = 1 \cdot 1,\ \Gamma(3) = 2 \cdot 1,\ \Gamma(4) = 3 \cdot 2 \cdot 1,\ \Gamma(5) = 4 \cdot 3 \cdot 2 \cdot 1$$

and proceeding by induction we get that $\Gamma(n) = (n-1)!$

Let us compute the kth moment of gamma distribution. We have,

$$\begin{aligned}
\mathbb{E}X^k &= \int_0^\infty x^k \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha+k)-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} \underbrace{\int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)} x^{\alpha+k-1} e^{-\beta x} dx}_{\text{p.d.f. of } \Gamma(\alpha+k,\beta) \text{ integrates to } 1} \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\beta^k} = \frac{(\alpha+k-1)\Gamma(\alpha+k-1)}{\Gamma(\alpha)\beta^k} \\
&= \frac{(\alpha+k-1)(\alpha+k-2)\ldots\alpha\Gamma(\alpha)}{\Gamma(\alpha)\beta^k} = \frac{(\alpha+k-1)\cdots\alpha}{\beta^k}.
\end{aligned}$$

Therefore, the mean is

$$\mathbb{E}X = \frac{\alpha}{\beta}$$

the second moment is

$$\mathbb{E}X^2 = \frac{(\alpha+1)\alpha}{\beta^2}$$

and the variance

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{\beta^2} - \Big(\frac{\alpha}{\beta}\Big)^2 = \frac{\alpha}{\beta^2}.$$

Below we will need the following property of Gamma distribution.

Lemma. *If we have a sequence of independent random variables*

$$X_1 \sim \Gamma(\alpha_1, \beta), \ldots, X_n \sim \Gamma(\alpha_n, \beta)$$

then $X_1 + \ldots + X_n$ *has distribution* $\Gamma(\alpha_1 + \ldots + \alpha_n, \beta)$

Proof. If $X \sim \Gamma(\alpha, \beta)$ then a moment generating function (m.g.f.) of X is

$$\begin{aligned}
\mathbb{E}e^{tX} &= \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x} dx \\
&= \frac{\beta^\alpha}{(\beta-t)^\alpha} \underbrace{\int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\beta-t)x} dx}.
\end{aligned}$$

The function in the last (underbraced) integral is a p.d.f. of gamma distribution $\Gamma(\alpha, \beta - t)$ and, therefore, it integrates to 1. We get,

$$\mathbb{E}e^{tX} = \Big(\frac{\beta}{\beta - t}\Big)^{\alpha}.$$

Moment generating function of the sum $\sum_{i=1}^n X_i$ is

$$\mathbb{E}e^{t\sum_{i=1}^n X_i} = \mathbb{E}\prod_{i=1}^n e^{tX_i} = \prod_{i=1}^n \mathbb{E}e^{tX_i} = \prod_{i=1}^n \Big(\frac{\beta}{\beta - t}\Big)^{\alpha_i} = \Big(\frac{\beta}{\beta - t}\Big)^{\sum \alpha_i}$$

and this is again a m.g.f. of Gamma distibution, which means that

$$\sum_{i=1}^n X_i \sim \Gamma\Big(\sum_{i=1}^n \alpha_i, \beta\Big).$$

□

χ^2_n-distribution. In the previous lecture we defined a χ^2_n-distribution with n degrees of freedom as a distribution of the sum $X_1^2 + \ldots + X_n^2$, where X_is are i.i.d. standard normal. We will now show that which χ^2_n-distribution coincides with a gamma distribution $\Gamma(\frac{n}{2}, \frac{1}{2})$, i.e.

$$\chi^2_n = \Gamma\Big(\frac{n}{2}, \frac{1}{2}\Big).$$

Consider a standard normal random variable $X \sim N(0,1)$. Let us compute the distribution of X^2. The c.d.f. of X^2 is given by

$$\mathbb{P}(X^2 \le x) = \mathbb{P}(-\sqrt{x} \le X \le \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

The p.d.f. can be computed by taking a derivative $\frac{d}{dx}\mathbb{P}(X \le x)$ and as a result the p.d.f. of X^2 is

$$\begin{aligned} f_{X^2}(x) &= \frac{d}{dx}\int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\sqrt{x})^2}{2}} (\sqrt{x})' - \frac{1}{\sqrt{2\pi}} e^{-\frac{(-\sqrt{x})^2}{2}} (-\sqrt{x})' \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}} e^{-\frac{x}{2}} = \frac{1}{\sqrt{2\pi}} x^{\frac{1}{2}-1} e^{-\frac{x}{2}}. \end{aligned}$$

We see that this is p.d.f. of Gamma Distribution $\Gamma(\frac{1}{2}, \frac{1}{2})$, i.e. we proved that $X^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2})$. Using Lemma above proves that $X_1^2 + \ldots + X_n^2 \sim \Gamma(\frac{n}{2}, \frac{1}{2})$.

□

Fisher F-distribution. Let us consider two independent random variables,

$$X \sim \chi^2_k = \Gamma\Big(\frac{k}{2}, \frac{1}{2}\Big) \quad \text{and} \quad Y \sim \chi^2_m = \Gamma\Big(\frac{m}{2}, \frac{1}{2}\Big).$$

Definition: *Distribution of the random variable*

$$Z = \frac{X/k}{Y/m}$$

is called a Fisher distribution *with degrees of freedom k and m, is denoted by $F_{k,m}$.*

First of all, let us notice that since $X \sim \chi^2_k$ can be represented as $X_1^2 + \ldots + X_k^2$ for i.i.d. standard normal $X_1, \ldots, X_k$, by law of large numbers,

$$\frac{1}{k}(X_1^2 + \ldots + X_k^2) \to \mathbb{E}X_1^2 = 1$$

when $k \to \infty$. This means that when k is large, the numerator X/k will 'concentrate' near 1. Similarly, when m gets large, the denominator Y/m will concentrate near 1. This means that when both k and m get large, the distribution $F_{k,m}$ will concentrate near 1.

Another property that is sometimes useful when using the tables of F-distribution is that

$$F_{k,m}(c, \infty) = F_{m,k}\Big(0, \frac{1}{c}\Big).$$

This is because

$$F_{k,m}(c,\infty) = \mathbb{P}\Big(\frac{X/k}{Y/m} \geq c\Big) = \mathbb{P}\Big(\frac{Y/m}{X/k} \leq \frac{1}{c}\Big) = F_{m,k}\Big(0, \frac{1}{c}\Big).$$

Next we will compute the p.d.f. of $Z \sim F_{k,m}$. Let us first compute the p.d.f. of

$$\frac{k}{m}Z = \frac{X}{Y}.$$

The p.d.f. of X and Y are

$$f(x) = \frac{(\frac{1}{2})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{1}{2}x} \quad \text{and} \quad g(y) = \frac{(\frac{1}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} e^{-\frac{1}{2}y}$$

correspondingly, where $x \geq 0$ and $y \geq 0$. To find the p.d.f of the ratio X/Y, let us first write its c.d.f. Since X and Y are always positive, their ratio is also positive and, therefore, for $t \geq 0$ we can write:

$$\mathbb{P}\Big(\frac{X}{Y} \leq t\Big) = \mathbb{P}(X \leq tY) = \int_0^\infty \Big(\int_0^{ty} f(x)g(y)dx\Big)dy$$

since $f(x)g(y)$ is the joint density of X, Y. Since we integrate over the set $\{x \leq ty\}$ the limits of integration for x vary from 0 to ty.

Since p.d.f. is the derivative of c.d.f., the p.d.f. of the ratio X/Y can be computed as follows:

$$\begin{aligned}
\frac{d}{dt}\mathbb{P}\Big(\frac{X}{Y} \leq t\Big) &= \frac{d}{dt}\int_0^\infty \int_0^{ty} f(x)g(y)dxdy = \int_0^\infty f(ty)g(y)ydy \\
&= \int_0^\infty \frac{(\frac{1}{2})^{\frac{k}{2}}}{\Gamma(\frac{k}{2})}(ty)^{\frac{k}{2}-1}e^{-\frac{1}{2}ty}\frac{(\frac{1}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})}y^{\frac{m}{2}-1}e^{-\frac{1}{2}y}ydy \\
&= \frac{(\frac{1}{2})^{\frac{k+m}{2}}}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}\underbrace{\int_0^\infty y^{(\frac{k+m}{2})-1}e^{-\frac{1}{2}(t+1)y}dy}
\end{aligned}$$

The function in the underbraced integral almost looks like a p.d.f. of gamma distribution $\Gamma(\alpha,\beta)$ with parameters $\alpha=(k+m)/2$ and $\beta=1/2$, only the constant in front is missing. If we miltiply and divide by this constant, we will get that,

$$\begin{aligned}\frac{d}{dt}\mathbb{P}\Big(\frac{X}{Y}\le t\Big) &= \frac{(\frac{1}{2})^{\frac{k+m}{2}}}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}\frac{\Gamma(\frac{k+m}{2})}{(\frac{1}{2}(t+1))^{\frac{k+m}{2}}}\int_0^\infty \frac{(\frac{1}{2}(t+1))^{\frac{k+m}{2}}}{\Gamma(\frac{k+m}{2})}y^{(\frac{k+m}{2})-1}e^{-\frac{1}{2}(t+1)y}dy\\ &= \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}(1+t)^{\frac{k+m}{2}},\end{aligned}$$

since the p.d.f. integrates to 1. To summarize, we proved that the p.d.f. of $(k/m)Z=X/Y$ is given by

$$f_{X/Y}(t)=\frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}t^{\frac{k}{2}-1}(1+t)^{-\frac{k+m}{2}}.$$

Since

$$\mathbb{P}(Z\le t)=\mathbb{P}\Big(\frac{X}{Y}\le\frac{kt}{m}\Big)\implies f_Z(t)=\frac{\partial}{\partial t}\mathbb{P}(Z\le t)=f_{X/Y}\Big(\frac{kt}{m}\Big)\frac{k}{m},$$

this proves that the p.d.f. of $F_{k,m}$-distribution is

$$\begin{aligned}f_{k,m}(t) &= \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}\frac{k}{m}\Big(\frac{kt}{m}\Big)^{\frac{k}{2}-1}\Big(1+\frac{kt}{m}\Big)^{-\frac{k+m}{2}}.\\ &= \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})}k^{k/2}m^{m/2}t^{\frac{k}{2}-1}(m+kt)^{-\frac{k+m}{2}}.\end{aligned}$$

Student t_n-distribution. Let us recall that we defined t_n-distibution as the distribution of a random variable

$$T=\frac{X_1}{\sqrt{\frac{1}{n}(Y_1^2+\cdots+Y_n^2)}}$$

if $X_1,Y_1,\ldots,Y_n$ are i.i.d. standard normal. Let us compute the p.d.f. of T. First, we can write,

$$\mathbb{P}(-t\le T\le t)=\mathbb{P}(T^2\le t^2)=\mathbb{P}\Big(\frac{X_1^2}{(Y_1^2+\cdots+Y_n^2)/n}\le t^2\Big).$$

If $f_T(x)$ denotes the p.d.f. of T then the left hand side can be written as

$$\mathbb{P}(-t\le T\le t)=\int_{-t}^t f_T(x)dx.$$

On the other hand, by definition,

$$\frac{X_1^2}{(Y_1^2+\ldots+Y_n^2)/n}$$

has Fisher $F_{1,n}$-distribution and, therefore, the right hand side can be written as

$$\int_0^{t^2} f_{1,n}(x)dx.$$

We get that,

$$\int_{-t}^{t} f_T(x)dx = \int_{0}^{t^2} f_{1,n}(x)dx.$$

Taking derivative of both side with respect to t gives

$$f_T(t) + f_T(-t) = f_{1,n}(t^2)2t.$$

But $f_T(t) = f_T(-t)$ since the distribution of T is obviously symmetric, because the numerator X has symmetric distribution $N(0,1)$. This, finally, proves that

$$f_T(t) = f_{1,n}(t^2)t = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}\frac{1}{\sqrt{n}}(1+\frac{t^2}{n})^{-\frac{n+1}{2}}.$$

□

Section 7

Testing hypotheses about parameters of normal distribution. T-tests and F-tests.

We will postpone a more systematic approach to hypotheses testing until the following lectures and in this lecture we will describe in an ad hoc way T-tests and F-tests about the parameters of normal distribution, since they are based on a very similar ideas to confidence intervals for parameters of normal distribution - the topic we have just covered.

Suppose that we are given an i.i.d. sample from normal distribution $N(\mu, \sigma^2)$ with some unknown parameters μ and σ^2. We will need to decide between two hypotheses about these unknown parameters - *null hypothesis* H_0 and *alternative hypothesis* H_1. Hypotheses H_0 and H_1 will be one of the following:

$$H_0 : \mu = \mu_0, \;\; H_1 : \mu \neq \mu_0,$$
$$H_0 : \mu \geq \mu_0, \;\; H_1 : \mu < \mu_0,$$
$$H_0 : \mu \leq \mu_0, \;\; H_1 : \mu > \mu_0,$$

where μ_0 is a given 'hypothesized' parameter. We will also consider similar hypotheses about parameter σ^2. We want to construct a *decision rule*

$$\delta : \mathcal{X}^n \to \{H_0, H_1\}$$

that given an i.i.d. sample $(X_1, \ldots, X_n) \in \mathcal{X}^n$ either accepts H_0 or rejects H_0 (accepts H_1). Null hypothesis is usually a 'main' hypothesis in a sense that it is expected or presumed to be true and we need a lot of evidence to the contrary to reject it. To quantify this, we pick a parameter $\alpha \in [0, 1]$, called *level of significance*, and make sure that a decision rule δ rejects H_0 when it is actually true with probability $\leq \alpha$, i.e.

$$\mathbb{P}(\delta = H_1 | H_0) \leq \alpha.$$

The probability on the left hand side is understood as a worse case scenario given that the null hypothesis is true, i.e.

$$P(\delta = H_1 | H_0) = \sup_{(\mu, \sigma^2) \in H_0} \mathbb{P}_{\mu, \sigma^2}(\delta = H_1).$$

Level of significance α is usually small, for example, $\alpha = 0.05$.

Example. Let us consider a Matlab example about normal body temperature from the lecture about confidence intervals. If a vector 'normtemp' represents body temperature measurements of 130 people then typing the following command in Matlab

```
[H,P,CI,STATS] = ttest(normtemp,98.6,0.05,'both')
```

produces the following output:

```
H = 1, P = 2.4106e-007, CI = [98.1220, 98.3765]
STATS = tstat: -5.4548, df: 129, sd: 0.7332.
```

Here $\mu_0 = 98.6$, $\alpha = 0.05$, 'both' means that we consider a null hypothesis $\mu = \mu_0$ in which case the alternative $\mu \neq \mu_0$ is a *two-sided* hypothesis. The alternative $\mu > \mu_0$ corresponds to parameter 'right', and $\mu < \mu_0$ corresponds to parameter 'left'. H = 1 means that we reject null hypothesis and accept H_1, P=2.4106e-007 is a *p-value* that we will discuss below, CI is a 95% confidence interval for μ_0 that we constructed in the previous lecture. If we want to test the hypothesis $\mu \geq 98.6$ then typing

```
[H,P,CI,STATS] = ttest(normtemp(1:130,1),98.6,0.05,'left')
```

outputs

```
H = 1, P = 1.2053e-007, CI = [-Inf, 98.3558],
STATS = tstat: -5.4548, df: 129, sd: 0.7332.
```

Notice that CI and P are different in this case. The fact that (in both cases) we rejected H_0 means that there is a significant evidence against it. In fact, we will see below that a p-value quantifies in some sense how unlikely it is to observe this dataset assuming that the null hypothesis is true. p-value of order 10^{-7} is a strong evidence against the hypothesis that a normal body temperature is $\mu_0 = 98.6$.

□

Let us explain how these tests are constructed. They are based on the result that we proved before that for MLE $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \bar{X^2} - \bar{X}^2$ satisfy

$$A = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \sim N(0,1) \text{ and } B = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

and the random variables A and B are independent.

Hypotheses about mean of one normal sample. We showed that a random variable

$$\sqrt{n-1}\frac{\hat{\mu} - \mu}{\hat{\sigma}} \sim t_{n-1}$$

has t_{n-1}-distribution with $n-1$ degrees of freedom. Let us consider a *t-statistic*

$$T = \sqrt{n-1}\frac{\hat{\mu} - \mu_0}{\hat{\sigma}}.$$

This statistic behaves differently depending on whether the 'true' unknown mean $\mu = \mu_0$, $\mu < \mu_0$ or $\mu > \mu_0$. First of all, if $\mu = \mu_0$ then $T \sim t_{n-1}$. If $\mu < \mu_0$ then we can rewrite

$$T = \sqrt{n-1}\frac{\hat{\mu} - \mu}{\hat{\sigma}} + \sqrt{n-1}\frac{\mu - \mu_0}{\hat{\sigma}} \to -\infty$$

since the first term has proper distribution t_{n-1} and the second term goes to infinity. Similarly, when $\mu > \mu_0$ then $T \to +\infty$. Therefore, we can make a decision about our hypotheses based on this information about the behavior of T.

I. ($H_0 : \mu = \mu_0$.) In this case, the indication that H_0 is not true would be if $|T|$ becomes too large, i.e. $T \to \pm\infty$. Therefore, we consider a decision rule

$$\delta = \begin{cases} H_0, & \text{if } -c \le T \le c \\ H_1, & \text{if } |T| > c. \end{cases}$$

The choice of the threshold c depends on the level of significance α. We would like to have

$$\mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(|T| > c | H_0) \le \alpha.$$

But given that $\mu = \mu_0$, we know that $T \sim t_{n-1}$ and, therefore, we can choose c from a condition

$$\mathbb{P}(|T| > c | H_0) = t_{n-1}(|T| > c) = 2t_{n-1}((c, \infty)) = \alpha$$

using the table of t_{n-1}-distribution. Notice that this decision rule is equivalent to finding the $(1-\alpha)$-confidence interval for unknown parameter μ and making a decision based on whether μ_0 falls into this interval.

□

II. ($H_0 : \mu \ge \mu_0$.) In this case, the indication that H_0 is not true, i.e. $\mu < \mu_0$, would be if $T \to -\infty$. Therefore, we consider a decision rule

$$\delta = \begin{cases} H_0, & \text{if } T \ge c \\ H_1, & \text{if } T < c. \end{cases}$$

The choice of the threshold c depends on the condition

$$\mathbb{P}(\delta = H_1 | H_0) = \mathbb{P}(T < c | H_0) \le \alpha.$$

Since we know that

$$T' = T - \sqrt{n-1}\frac{\mu - \mu_0}{\hat{\sigma}} \sim t_{n-1}$$

we can write

$$\mathbb{P}(T < c | H_0) = \sup_{\mu \ge \mu_0} \mathbb{P}\Big(T' \le c - \sqrt{n-1}\frac{\mu - \mu_0}{\hat{\sigma}}\Big) = \mathbb{P}(T' \le c) = t_{n-1}((-\infty, c]) = \alpha$$

and we can find c using the table of t_{n-1}-distribution.

□

III. ($H_0 : \mu \leq \mu_0$.) Similar to the previous case, the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } T \leq c \\ H_1, & \text{if } T > c. \end{cases}$$

and we find c from the condition $t_{n-1}([c, +\infty)) = \alpha$.

□

p-value. Figure 7.1 illustrates the definition of p-value for all three cases above. p-value can be understood as a probability, given that the null hypothesis H_0 is true, to observe a value of T-statistic equally or less likely than the one that was observed. Thus, the small p-value means that the observed T-statistic is very unlikely under the null hypothesis which provides a strong evidence against H_0. The confidence level α defines what we consider as 'unlikely enough' to reject the null hypothesis.

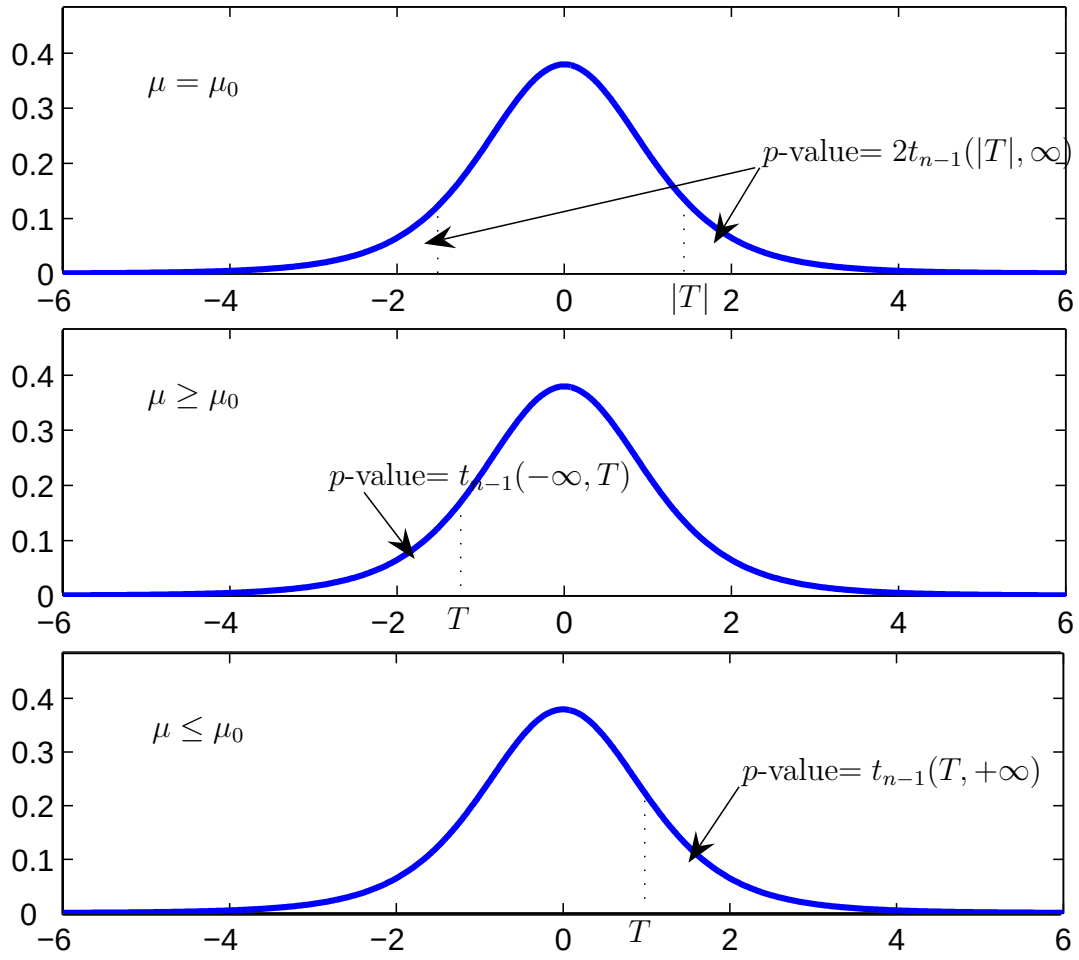


Figure 7.1: p-values for different cases.

Hypotheses about variance of one normal sample. Next we will test similar two-sided or one sided hypotheses about the variance, for example, $H_0 : \sigma = \sigma_0$ vs. $H_1 : \sigma \neq \sigma_0$, etc. We will use the fact that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$-distribution and as a result the test will be

based of the following statistic:

$$Q = \frac{n\hat{\sigma}^2}{\sigma_0^2}.$$

Since we can write

$$Q = \frac{n\hat{\sigma}^2}{\sigma^2}\frac{\sigma^2}{\sigma_0^2} \sim \frac{\sigma^2}{\sigma_0^2}\chi^2_{n-1},$$

then, clearly, Q will behave differently depending on whether $\sigma = \sigma_0$, $\sigma > \sigma_0$ or $\sigma < \sigma_0$.

I. ($H_0 : \sigma = \sigma_0$.) In this case the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } c_1 \le Q \le c_2 \\ H_1, & \text{if } Q < c_1, c_2 < Q. \end{cases}$$

Thresholds c_1, c_2 should satisfy the condition

$$\mathbb{P}(\delta = H_1|H_0) = \mathbb{P}(Q < c_1|\sigma = \sigma_0) + \mathbb{P}(Q > c_2|\sigma = \sigma_0) = \chi^2_{n-1}(0, c_1) + \chi^2_{n-1}(c_2, \infty) = \alpha.$$

For example, we can take

$$\chi^2_{n-1}(0, c_1) = \frac{\alpha}{2} \text{ and } \chi^2_{n-1}(c_2, \infty) = \frac{\alpha}{2}.$$

□

II. ($H_0 : \sigma \le \sigma_0$.) In this case the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } Q \le c \\ H_1, & \text{if } Q > c. \end{cases}$$

Threshold c should satisfy the condition

$$\mathbb{P}(\delta = H_1|H_0) = \sup_{\sigma \le \sigma_0} \mathbb{P}(Q > c) = \sup_{\sigma \le \sigma_0} \mathbb{P}\Big(\frac{n\hat{\sigma}^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}c\Big) = \mathbb{P}\Big(\frac{n\hat{\sigma}^2}{\sigma^2} > c\Big) = \chi^2_{n-1}(c, \infty) = \alpha.$$

□

III. ($H_0 : \sigma \ge \sigma_0$.) In this case the decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } Q \ge c \\ H_1, & \text{if } Q < c. \end{cases}$$

Threshold c is determined by

$$\mathbb{P}(\delta = H_1|H_0) = \sup_{\sigma \ge \sigma_0} \mathbb{P}(Q < c) = \mathbb{P}\Big(\frac{n\hat{\sigma}^2}{\sigma^2} < c\Big) = \chi^2_{n-1}(0, c) = \alpha.$$

□

Comparing means of two normal samples. In the normal body temperature dataset first 65 observations correspond to men and last 65 observations correspond to women. We

would like to test the hypothesis that normal body temperature of men and women is the same. There are several way to do this.

Paired t-test. First, we can perform the so called *paired t-test*. Since the number of observations is the same in both groups, we can pair them together and assume that their differences $Z_i = X_i - Y_i$ will also be normal. This sounds like a reasonable assumption since X_i and Y_i should be independent if the measurements were taken independently. Since $\mu_z = \mu_x - \mu_y$, hypothesis $\mu_x = \mu_y$ is equivalent to $\mu_z = 0$ which means that we can do the usual t-test for one sample $Z_1, \ldots, Z_n$. Running

```
[H,P,CI,STATS]=ttest(men,women,0.05,'both')
```

outputs

```
H = 1, P = 3.9773e-019, CI = [-0.3348,-0.2437]
STATS = tstat: -12.6858, df: 64, sd: 0.1838.
```

We reject null hypothesis that the means are equal and, in fact, p-value of order 10^{-19} is a strong evidence against it. However, it seems rather suspicious that there is such a strong evidence against H_0, especially after we perform a two sample t-test below which also rejects H_0 but with a much higher p-value of 0.0239. When we examine the data file more closely we notice that the body temperatures were arranged in an increasing order both for men and women. This means that the assumption that our samples are i.i.d. is not longer true. To restore this, we randomly permute both vectors and denote their difference by 'z'. (To permute 'men' type 'men(randperm(65))', the same for women.) Then performing t-test for the difference 'z'

```
[H,P,CI,STATS]=ttest(z,0,0.05,'both')
```

we get

```
H = 1, P =0.0442, CI = [0.0078, 0.5707]
STATS = tstat: 2.0528, df: 64, sd: 1.1359
```

which is a more reasonable (and correct) outcome.

□

Two sample t-test assuming equal variances. If we run the following command in Matlab:

```
[H,P,CI,STATS]=ttest2(men,women,0.05,'both','equal')
```

we get the following output:

```
H = 1, P = 0.0239, CI = [-0.5396, -0.0388],
STATS = tstat: -2.2854, df: 128, sd: 0.7215.
```

We again reject the hypothesis that $\mu_x = \mu_y$ at the level of significance $\alpha = 0.05$ but this time p-value is equal to 0.0239. Here 'both' means that we test two-sided hypothesis $\mu_x = \mu_y$, and 'equal' means that we assume that the 'true' unknown variances of the distributions of two samples σ_x^2 and σ_y^2 are equal, i.e.

$$\sigma_x^2 = \sigma_y^2 = \sigma^2$$

Let n and m be the number of observations in the first sample (Xs) and second sample (Ys) correspondingly. We proved that

$$A_x = \frac{\sqrt{n}(\hat{\mu}_x - \mu_x)}{\sigma_x} \sim N(0,1) \text{ and } B_x = \frac{n\hat{\sigma}_x^2}{\sigma_x^2} \sim \chi^2_{n-1}$$

and

$$A_y = \frac{\sqrt{m}(\hat{\mu}_y - \mu_y)}{\sigma_y} \sim N(0,1) \text{ and } B_y = \frac{m\hat{\sigma}_y^2}{\sigma_y^2} \sim \chi^2_{m-1}$$

and A_x, B_x, A_y, B_y are independent. Using the properties of normal distribution we get

$$A = \Big(\frac{(\hat{\mu}_x - \mu_x)}{\sigma_x} - \frac{(\hat{\mu}_y - \mu_y)}{\sigma_y}\Big) \Big/ \Big(\frac{1}{n} + \frac{1}{m}\Big) \sim N(0,1)$$

and by definition of χ^2-distributions,

$$B = \frac{n\hat{\sigma}_x^2}{\sigma_x^2} + \frac{m\hat{\sigma}_y^2}{\sigma_y^2} \sim \chi^2_{n+m-2}.$$

Therefore,

$$A \Big/ \sqrt{\frac{1}{n+m-2} B} \sim t_{n+m-2}.$$

Notice that because σ_x^2 and σ_y^2 are unknown, in general, we can not compute this expression. However, if we assume that variances are equal then all σ_x and σ_y will cancel out and we will get

$$\Big(\frac{nm(n+m-2)}{n+m}\Big)^{1/2} \frac{(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)}{(n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2)^{1/2}} \sim t_{n+m-2.}$$

Since this expression depends only on the difference of means $\mu_x - \mu_y$, we can test hypotheses about this difference based on the statistic

$$T = \Big(\frac{nm(n+m-2)}{n+m}\Big)^{1/2} \frac{\hat{\mu}_x - \hat{\mu}_y}{(n\hat{\sigma}_x^2 + m\hat{\sigma}_y^2)^{1/2}}$$

For example, if we want to test $H_0 : \mu_x = \mu_y$ or, equivalently, $\mu_x - \mu_y = 0$ we can consider a decision rule

$$\delta = \begin{cases} H_0, & \text{if } -c \le T \le c \\ H_1, & \text{if } |T| > c \end{cases}$$

and find c from the condition $2t_{n+m-2}(c, \infty) = \alpha$. Notice that in the Matlab output above we have $df = 128$, i.e. $n + m - 2 = 65 + 65 - 2 = 128$ degrees of freedom. One-sided tests are also similar to the case of one sample test.

t-test with unequal variances. Assuming that variances are equal could be unjustified. There is a version of t-test which does not make this assumption. However, we can no compute exactly the distribution of t-statistic above (since variances do not cancel out), and we can only construct 'approximate' tests. For example, running in Matlab:

```
[H,P,CI,STATS]=ttest2(men,women,0.05,'both','unequal')
```

gives

```
H = 1, P = 0.0239, CI =[-0.5396,-0.0388],
STATS =  tstat: -2.2854, df: 127.5103, sd: [2x1 double].
```

Notice non integer value for degrees of freedom 127.5103. To construct the test for this general case we can start with (using properties of normal distribution)

$$(\hat{\mu}_x - \mu_x) - (\hat{\mu}_y - \mu_y) \sim N\Big(0, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\Big)$$

or

$$((\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))/\Big(\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\Big)^{1/2} \sim N(0,1)$$

We do not know the variances σ_x^2 and σ_y^2 but we know by law of large numbers that their estimates $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ converge and, therefore,

$$(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))\Big/\Big(\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}\Big)^{1/2} \approx N(0,1)$$

will have approximately normal distribution when n and m are large. We can now construct all the tests as above, only now they will be approximate. However, usually a different (supposedly better) approximation is used, called *Satterthwaite approximation*, also used by Matlab. First of all, instead of $\hat{\sigma}_x^2$ we will use *unbiased* estimated of variance:

$$\sigma_x'^2 = \frac{n\hat{\sigma}_x^2}{n-1}$$

which will give us a slightly different expression

$$((\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))\Big/\Big(\frac{\hat{\sigma}_x^2}{n-1} + \frac{\hat{\sigma}_y^2}{m-1}\Big)^{1/2} \approx N(0,1).$$

Unbiased estimate $\sigma_x'^2$ is different from MLE $\hat{\sigma}_x^2$ only by a fraction $n/(n-1)$ and we can see that this makes very small difference between two expressions above. More important difference is that instead of using normal approximation $\approx N(0,1)$ we will use a t_ν-distribution approximation

$$((\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y))\Big/\Big(\frac{\hat{\sigma}_x^2}{n-1} + \frac{\hat{\sigma}_y^2}{m-1}\Big)^{1/2} \approx t_\nu \tag{7.0.1}$$

where the number of degrees of freedom ν is determined from the following consideration. We know from the definition of t_n-distribution and properties of χ^2_n-distribution that (using informal notations)

$$t_n = \frac{N(0,1)}{\sqrt{\frac{1}{n}\chi^2_n}} = \frac{N(0,1)}{\sqrt{\frac{1}{n}\Gamma\left(\frac{n}{2},\frac{1}{2}\right)}}.$$

This could be used as a definition of t_n-distribution even when degrees of freedom parameter n is not integer. To find a good approximation in (7.0.1), we need to find a good approximation

$$\Big(\frac{\hat\sigma_x^2}{n-1}+\frac{\hat\sigma_y^2}{m-1}\Big)\Big/\Big(\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{n}\Big) \approx \frac{1}{\nu}\Gamma\Big(\frac{\nu}{2},\frac{1}{2}\Big).$$

It is easy to check that the expectations of both sides are equal, so we will choose ν from the condition that the variances of both sides are also equal, which will give

$$\nu = \frac{\left(\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{\sigma_x^2}{n}\right)^2+\frac{1}{y-1}\left(\frac{\sigma_y^2}{m}\right)^2}.$$

Finally, since the variances are unknown we will replace them by their unbiased estimates and take

$$\nu = \frac{\left(\frac{\hat\sigma_x^2}{n-1}+\frac{\hat\sigma_y^2}{m-1}\right)^2}{\frac{1}{n-1}\left(\frac{\hat\sigma_x^2}{n-1}\right)^2+\frac{1}{y-1}\left(\frac{\hat\sigma_y^2}{m}\right)^2}.$$

Therefore, we obtain approximation (7.0.1) which is supposedly better than a simple normal approximation. The degrees of freedom *df* : 127.5103 in the Matlab output is precisely given by this formula.

□

Comparing the variances of two normal distributions: F-test. Suppose that we want to test whether the variances of two normal distributions are equal. For example, in the first two sample t-test we assumed that $\sigma_x^2 = \sigma_y^2$. We can test this in Matlab:

```
[H,P,CI,STATS]=vartest2(men,women,0.05,'both')
```

and we get the following output:

```
H = 0, P = 0.6211, CI = [0.5388, 1.4481]
STATS = fstat: 0.8833, df1: 64, df2: 64.
```

We accept the two-sided null hypothesis $H_0 : \sigma_x = \sigma_y$. The high p-value 0.6211 means that there is no evidence against null hypothesis. This test is constructed as follows. Since we know that

$$B_x = \frac{n\hat\sigma_x^2}{\sigma_x^2} \sim \chi^2_{n-1} \text{ and } B_y = \frac{m\hat\sigma_y^2}{\sigma_y^2} \sim \chi^2_{m-1}$$

the ratio

$$\frac{B_x/(n-1)}{B_y/(m-1)} = \frac{n(m-1)\hat{\sigma}_x^2}{m(n-1)\hat{\sigma}_y^2}\frac{\sigma_y^2}{\sigma_x^2} \sim F_{n-1,m-1}$$

has $F_{m-1,n-1}$-distribution with $(n-1, m-1)$ degrees of freedom. Let us consider a statistic

$$F = \frac{n(m-1)\hat{\sigma}_x^2}{m(n-1)\hat{\sigma}_y^2} \sim \frac{\sigma_x^2}{\sigma_y^2}F_{n-1,m-1}.$$

When $\sigma_x^2 = \sigma_y^2$, we have $F \sim F_{n-1,m-1}$, when $\sigma_x^2 > \sigma_y^2$, F will tend to be above the 'typical range' of $F_{n-1,m-1}$ distribution, and when $\sigma_x^2 < \sigma_y^2$, F will tend to be below the 'typical range' of $F_{n-1,m-1}$ distribution. As a result, we get the following tests.

I. $(H_0 : \sigma_x = \sigma_y.)$ The decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } c_1 \le F \le c_2 \\ H_1, & \text{if } F < c_1, c_2 < F. \end{cases}$$

Thresholds c_1, c_2 should satisfy the condition

$$\begin{aligned} \mathbb{P}(\delta = H_1|H_0) &= \mathbb{P}(F < c_1|\sigma_x = \sigma_y) + \mathbb{P}(F > c_2|\sigma_x = \sigma_y) \\ &= F_{n-1,m-1}(0, c_1) + F_{n-1,m-1}(c_2, \infty) = \alpha. \end{aligned}$$

For example, we can take

$$F_{n-1,m-1}(0, c_1) = \frac{\alpha}{2} \text{ and } F_{n-1,m-1}(c_2, \infty) = \frac{\alpha}{2}.$$

□

II. $(H_0 : \sigma_x \le \sigma_y.)$ The decision rule will be

$$\delta = \begin{cases} H_0, & \text{if } F \le c \\ H_1, & \text{if } F > c. \end{cases}$$

Thresholds c should satisfy the condition

$$\mathbb{P}(\delta = H_1|H_0) = \mathbb{P}(F > c|\sigma_x = \sigma_y) = F_{n-1,m-1}(c, \infty) = \alpha.$$

The test for $H_0 : \sigma_x \ge \sigma_y$ is similar.

□

Section 8

Testing simple hypotheses. Bayes decision rules.

Let us consider an i.i.d. sample $X_1, \dots, X_n \in \mathcal{X}$ with unknown distribution $\mathbb{P}$ on $\mathcal{X}$. Suppose that the distribution $\mathbb{P}$ belongs to a set of k specified distributions, $\mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_k\}$. Then, given a sample $X, \dots, X_n$ we have to decide among k simple hypotheses:

$$\begin{cases} H_1: & \mathbb{P} = \mathbb{P}_1 \\ H_2: & \mathbb{P} = \mathbb{P}_2 \\ \quad \vdots & \\ H_k: & \mathbb{P} = \mathbb{P}_k. \end{cases}$$

In other words, we need to construct a decision rule

$$\delta : \mathcal{X}^n \to \{H_1, \dots, H_k\}.$$

Let us note that sometimes this function δ will be random because when several hypotheses are 'equally likely' it will make sense to pick among them randomly. This idea of a *randomized* decision rule will be explained more precisely in the following lectures, but for now we will simply think of δ as a function of the sample.

Suppose that the ith hypothesis is true, i.e. $\mathbb{P} = \mathbb{P}_i$. Then the probability that a decision rule δ will make an error is

$$\mathbb{P}(\delta \neq H_i | H_i) = \mathbb{P}_i(\delta \neq H_i),$$

which is called the *error of type* i or type i error. When $k = 2$, i.e. we consider two hypotheses H_1 and H_2, the error of type 1

$$\alpha_1 = \mathbb{P}_1(\delta \neq H_1)$$

is also called the *size* or *level of significance* of the decision rule δ and

$$\beta = 1 - \alpha_2 = 1 - \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2(\delta = H_2)$$

is called the *power* of δ.

In order to construct a good decision rule, we need to decide how to compare decision rules. Ideally, we would like to make the errors of all types as small as possible. However,

typically there is a trade-off among errors and it is impossible to minimize them simultaneously. A decision rule δ will create a partition of space $\mathcal{X}^n$ into k disjoint subsets $A_1, \ldots, A_k$ such that

$$\delta(X_1, \ldots, X_n) = H_j \quad \text{if and only if} \quad (X_1, \ldots, X_n) \in A_j.$$

Increasing a set A_j will decrease the error of type j since

$$\alpha_j = \mathbb{P}(A_j^c) = 1 - \mathbb{P}(A_j)$$

and, therefore, in this sense k simple hypotheses compete with each other. Of course, it is possible to give an example in which all errors are zero. For example, if all distributions $\mathbb{P}_1, \ldots, \mathbb{P}_k$ concentrate on disjoint subsets of $\mathcal{X}$ then one observation is enough to predict the correct hypothesis with no error.

One way to compare decision rules would be to assign weights $\xi(1), \ldots, \xi(k)$ to the hypotheses and consider a *weighted error*

$$\xi(1)\alpha_1 + \ldots \xi(k)\alpha_k = \xi(1)\mathbb{P}(\delta \neq H_1|H_1) + \ldots + \xi(k)\mathbb{P}(\delta \neq H_k|H_k).$$

In the next section we will construct decision rules that minimize this weighted error.

In the case of two simple hypotheses H_1 and H_2 it is more common to construct 'good' decision rules based on a different criterion. Before we describe this criterion, let us first see that in many practical problems different types of errors have very different meanings.

Example. Suppose that a medical test is done to determine if a patient is sick. Then based on the data from the test we have to decide between two hypotheses:

$$H_1 : \text{ positive}; H_2 : \text{ negative}.$$

Then the error of type one $\mathbb{P}(\delta = H_2|H_1)$ means that we determine that the patient is sick when he is not and the error of type two $\mathbb{P}(\delta = H_1|H_2)$ means that we determine that a patient is not sick when he is. Clearly, these errors are of a very different nature. In the first case a patient will not get a necessary treatment. In the second case a patient might get unnecessary and potentially harmful treatment. However, in the second case additional tests can be done whereas in the first case the sickness may be completely overlooked. This means that it may be more important to control the error of type 1 in this case.

Example. Radar missile detection/recognition. Suppose that based on a radar image we decide between a missile and a passenger plane:

$$H_1 : \text{ missile}, H_2 : \text{ not missile}.$$

Then the error of type one $\mathbb{P}(\delta = H_2|H_1)$, means that we will ignore a missile and error of type two $\mathbb{P}(\delta = H_2|H_1)$, means that we will possibly shoot down a passenger plane (which happened before). It depends on the situation to decide which error is more important to control.

Another example could be when 'guilty' or 'not guilty' verdict in court is decided based on some data. Presumption of innocence means that 'no guilty' hypothesis is a more important *null hypothesis* and the error of type $\mathbb{P}$('guilty'|'not guilty') should be controlled. When

a drug company comes up with a new drug, it is their responsibility to prove that a drug works significantly better than a sugar pill, so a 'more important' null hypothesis in this case is that a drug does not work better.

These examples illustrate that in many situations a particular hypothesis is more important in a sense that the error corresponding to this hypothesis should be controlled. We will assume that H_1 is this hypothesis. Let $\alpha \in [0,1]$ be the largest possible error of type one that we are willing to accept, which means that we will only consider decision rules in the class

$$K_\alpha = \{\delta : \alpha_1 = \mathbb{P}_1(\delta \neq H_1) \leq \alpha\}.$$

It now makes sense that among all decision rules in this class we should try to find a decision rule that makes the error of type two, $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$, as small as possible. We will show how to construct such decision rules in the following lectures but, first, we will construct decision rules that minimize the weighted error.

Bayes decision rules.

Given hypotheses $H_1, \ldots, H_k$ let us consider k nonnegative weights $\xi(1), \ldots, \xi(k)$ that add up to one $\sum_{i=1}^k \xi(i) = 1$. We can think of weights ξ as a priori probability on the set of k hypotheses that represent their relative importance. Then the *Bayes error* of a decision rule δ is defined as

$$\alpha(\xi) = \sum_{i=1}^k \xi(i)\alpha_i = \sum_{i=1}^k \xi(i)\mathbb{P}_i(\delta \neq H_i),$$

which is simply a weighted error. We would like to make the Bayes error as small as possible.

Definition: *Decision rule δ that minimizes $\alpha(\xi)$ is called a* Bayes decision rule.

Next theorem constructs Bayes decision rules in terms of p.d.f. or p.f. of $\mathbb{P}_i, 1 \leq i \leq k$.

Theorem. *Assume that each distribution $\mathbb{P}_i$ has p.d.f or p.f. $f_i(x)$. A decision rule δ that predicts H_j when*

$$\xi(j)f_j(X_1)\ldots f_j(X_n) = \max_{1\leq i\leq k} \xi(i)f_i(X_1)\ldots f_i(X_n)$$

is a Bayes decision rule.

In other words, we choose hypotheses H_j if it maximizes the weighted likelihood function

$$\xi(i)f_i(X_1)\ldots f_i(X_n)$$

among all hypotheses. If this maximum is achieved simultaneously on several hypotheses we can pick any one of them, or at random.

Proof. Let us rewrite the Bayes error as follows:

$$\begin{aligned}
\alpha(\xi) &= \sum_{i=1}^k \xi(i)\mathbb{P}_i(\delta \neq H_i) \\
&= \sum_{i=1}^k \xi(i) \int I(\delta \neq H_i) f_i(x_1)\ldots f_i(x_n)dx_1\ldots dx_n
\end{aligned}$$

$$
\begin{aligned}
&= \int \sum_{i=1}^{k} \xi(i) f_i(x_1) \dots f_i(x_n) \Big(1 - I(\delta = H_i)\Big) dx_1 \dots dx_n \\
&= \sum_{i=1}^{k} \xi(i) \underbrace{\int f_i(x_1) \dots f_i(x_n) dx_1 \dots dx_n}_{\text{this joint density integrates to 1 and } \sum \xi(i) = 1} \\
&\quad - \int \sum_{i=1}^{k} \xi(i) f_i(x_1) \dots f_i(x_n) I(\delta = H_i) dx_1 \dots dx_n \\
&= 1 - \int \sum_{i=1}^{k} \xi(i) f_i(x_1) \dots f_i(x_n) I(\delta = H_i) dx_1 \dots dx_n.
\end{aligned}
$$

To minimize this Bayes error we need to maximize this last integral, but we can actually maximize the sum inside the integral

$$
\xi(1) f_1(x_1) \dots f_1(x_n) I(\delta = H_1) + \dots + \xi(k) f_k(x_1) \dots f_k(x_n) I(\delta = H_k)
$$

by choosing δ appropriately. For each $(x_1, \dots, x_n)$ decision rule δ picks only one hypothesis which means that only one term in this sum will be non zero, because if δ picks H_j then only one indicator $I(\delta = H_j)$ will be non zero and the sum will be equal to

$$
\xi(j) f_j(x_1) \dots f_j(x_n).
$$

Therefore, to maximize the integral δ should simply pick the hypothesis that maximizes this expression, exactly as in the statement of the Theorem. This finishes the proof.

□

Let us write down a Bayes decision rule in the case of two simple hypotheses H_1, H_2. For simplicity of notations, given a sample $X = (X_1, \dots, X_n)$ we will denote the joint p.d.f. or p.f. by

$$
f_i(X) = f_i(X_1) \dots f_i(X_n).
$$

Then the Bayes decision rule that minimizes the weighted error

$$
\alpha = \xi(1) \mathbb{P}_1(\delta \neq H_1) + \xi(2) \mathbb{P}_2(\delta \neq H_2)
$$

is given by

$$
\delta = \begin{cases} H_1: & \xi(1) f_1(X) > \xi(2) f_2(X) \\ H_2: & \xi(2) f_2(X) > \xi(1) f_1(X) \\ H_1 \text{ or } H_2: & \xi(1) f_1(X) = \xi(2) f_2(X). \end{cases}
$$

Equivalently,

$$
\delta = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} > \frac{\xi(2)}{\xi(1)} \\ H_2: & \frac{f_1(X)}{f_2(X)} < \frac{\xi(2)}{\xi(1)} \\ H_1 \text{ or } H_2: & \frac{f_1(X)}{f_2(X)} = \frac{\xi(2)}{\xi(1)}. \end{cases} \tag{8.0.1}
$$

(Here $\frac{1}{0} = +\infty$, $\frac{0}{1} = 0$.) This type of test is often called a *likelihood ratio test* because it is expressed in terms of the ratio $f_1(X)/f_2(X)$ of likelihood functions.

Example. Suppose we have one observation X_1 and two simple hypotheses

$$H_1 : \mathbb{P} = N(0,1) \quad \text{and} \quad H_2 : \mathbb{P} = N(1,1).$$

Take equal weights

$$\xi(1) = \frac{1}{2} \text{ and } \xi(2) = \frac{1}{2}.$$

Then a Bayes decision rule δ that minimizes

$$\frac{1}{2}\mathbb{P}_1(\delta \neq H_1) + \frac{1}{2}\mathbb{P}_2(\delta \neq H_2)$$

is given by

$$\delta(X_1) = \begin{cases} H_1 : & \frac{f_1(X)}{f_2(X)} > 1 \\ H_2 : & \frac{f_1(X)}{f_2(X)} < 1 \\ H_1 \text{ or } H_2 : & \frac{f_1(X)}{f_2(X)} = 1. \end{cases}$$

This decision rule has a very intuitive interpretation. If we look at the graphs of these p.d.f.s (figure 8.1) the decision rule picks the first hypothesis when the first p.d.f. is larger, $x \leq 0.5$, and otherwise picks the second hypothesis, $x > 0.5$

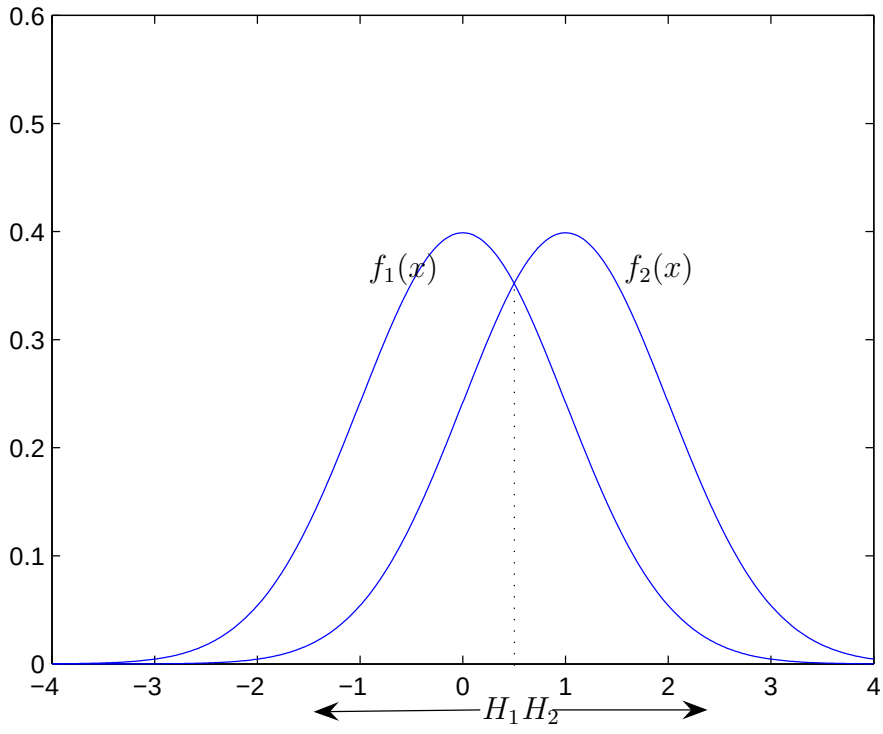


Figure 8.1: Bayes decision rule.

Example. Let us a general example case of n observations $X = (X_1, \ldots, X_n)$, two simple hypotheses $H_1 : \mathbb{P} = N(0,1)$ and $H_2 : \mathbb{P} = N(1,1)$, and arbitrary a priori weights $\xi(1), \xi(2)$. Then Bayes decision rule is given by (8.0.1). The likelihood ratio can be simplified:

$$\begin{aligned} \frac{f_1(X)}{f_2(X)} &= \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n X_i^2} \Big/ \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n (X_i-1)^2} \\ &= e^{\frac{1}{2}\sum_{i=1}^n ((X_i-1)^2 - X_i^2)} = e^{\frac{n}{2} - \sum_{i=1}^n X_i}. \end{aligned}$$

Therefore, the decision rule picks the first hypothesis H_1 when

$$e^{\frac{n}{2} - \sum X_i} > \frac{\xi(2)}{\xi(1)} \quad \text{or, equivalently,} \quad \sum X_i < \frac{n}{2} - \log \frac{\xi(2)}{\xi(1)}.$$

Similarly, we pick the second hypothesis H_2 when

$$\sum X_i > \frac{n}{2} - \log \frac{\xi(2)}{\xi(1)}.$$

In case of equality, we pick either H_1 or H_2.

□

Section 9

Most powerful tests for two simple hypotheses.

Now that we learned how to construct the decision rule that minimizes the Bayes error we will turn to our next goal - constructing the decision rule with controlled error of type 1 that minimizes error of type 2. Given $\alpha \in [0,1]$ we consider the class of decision rules

$$K_\alpha = \{\delta : \mathbb{P}_1(\delta \neq H_1) \leq \alpha\}$$

and we will try to find $\delta \in K_\alpha$ that makes the type 2 error $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$ as small as possible.

Theorem. *Assume that there exists a constant c, such that*

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha. \tag{9.0.1}$$

Then the decision rule

$$\delta = \begin{cases} H_1: & \frac{f_1(X)}{f_2(X)} \geq c \\ H_2: & \frac{f_1(X)}{f_2(X)} < c \end{cases} \tag{9.0.2}$$

is the most powerful in class K_α.

Proof. Take $\xi(1)$ and $\xi(2)$ such that

$$\xi(1) + \xi(2) = 1, \;\; \frac{\xi(2)}{\xi(1)} = c,$$

i.e.

$$\xi(1) = \frac{1}{1+c} \;\; \text{and} \;\; \xi(2) = \frac{c}{1+c}.$$

Then the decision rule δ in (9.0.2) is the Bayes decision rule corresponding to weights $\xi(1)$ and $\xi(2)$ which can be seen by comparing it with (8.0.1), only here we break the tie in favor of H_1. Therefore, this decision rule δ minimizes the Bayes error which means that for any other decision rule δ',

$$\xi(1)\mathbb{P}_1(\delta \neq H_1) + \xi(2)\mathbb{P}_2(\delta \neq H_2) \leq \xi(1)\mathbb{P}_1(\delta' \neq H_1) + \xi(2)\mathbb{P}_2(\delta' \neq H_2). \tag{9.0.3}$$

By assumption (9.0.1) we have

$$\mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha,$$

which means that δ comes from the class K_α. If $\delta' \in K_\alpha$ then

$$\mathbb{P}_1(\delta' \neq H_1) \leq \alpha$$

and equation (9.0.3) gives us that

$$\xi(1)\alpha + \xi(2)\mathbb{P}_2(\delta \neq H_2) \leq \xi(1)\alpha + \xi(2)\mathbb{P}_2(\delta' \neq H_2)$$

and, therefore,

$$\mathbb{P}_2(\delta \neq H_2) \leq \mathbb{P}_2(\delta' \neq H_2).$$

This exactly means that δ is more powerful than any other decision rule in class K_α.

□

Example. Suppose we have a sample $X = (X_1, \ldots, X_n)$ and two simple hypotheses $H_1 : \mathbb{P} = N(0,1)$ and $H_2 : \mathbb{P} = N(1,1)$. Let us find most powerful δ with the error of type 1

$$\alpha_1 \leq \alpha = 0.05.$$

According to the above Theorem if we can find c such that

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha = 0.05$$

then we know how to find δ. Simplifying this equation gives

$$\mathbb{P}_1\Big(\sum X_i > \frac{n}{2} - \log c\Big) = \alpha = 0.05$$

or

$$\mathbb{P}_1\Big(\frac{1}{\sqrt{n}}\sum X_i > c' = \frac{1}{\sqrt{n}}(\frac{n}{2} - \log c)\Big) = \alpha = 0.05.$$

But under the hypothesis H_1 the sample comes from standard normal distribution $\mathbb{P}_1 = N(0,1)$ which implies that the random variable

$$Y = \frac{1}{\sqrt{n}}\sum X_i$$

is standard normal. We can use the table of normal distribution to find

$$\mathbb{P}(Y > c') = \alpha = 0.05 \Rightarrow c' = 1.64.$$

Therefore, the most powerful test with level of significance $\alpha = 0.05$ is:

$$\delta = \begin{cases} H_1 : & \frac{1}{\sqrt{n}}\sum X_i \leq 1.64 \\ H_2 : & \frac{1}{\sqrt{n}}\sum X_i > 1.64. \end{cases}$$

What will the error of type 2 be for this test?

$$\begin{aligned}\alpha_2 &= \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2\Big(\frac{1}{\sqrt{n}}\sum X_i \leq 1.64\Big) \\ &= \mathbb{P}_2\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1) \leq 1.64 - \sqrt{n}\Big).\end{aligned}$$

The reason we subtracted 1 from each X_i is because under the second hypothesis X's have distribution $N(1,1)$ and random variable

$$Y = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 1)$$

will be standard normal. Therefore, the error of type 2 for this test will be equal to

$$\mathbb{P}(Y < 1.64 - \sqrt{n}) = N(0,1)(-\infty, 1.64 - \sqrt{n}).$$

For example, when the sample size $n = 10$, $\alpha_2 = \mathbb{P}(Y < 1.64 - \sqrt{10}) = 0.087$.

□

Randomized most powerful test.

Next we will show how to get rid of the assumption (9.0.1) which, unfortunately, does not always hold as will become clear from examples below.

If we examine carefully the proof of Theorem we notice that condition (9.0.1) was necessary to ensure that the likelihood ratio test has error of type 1 *exactly*! equal to α. Also, the test was designed to be a Bayes test and in Bayes tests we have a freedom of breaking a tie in an arbitrary way. In the following version of previous theorem we will show that the most powerful test in class K_α can always be constructed by breaking a tie randomly in a way that makes error of type 1 exactly equal to α.

Theorem. *Given any $\alpha \in [0,1]$ we can always find $c \in [0,\infty)$ and $p \in [0,1]$ such that*

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) + (1-p)\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} = c\Big) = \alpha. \tag{9.0.4}$$

In this case, the most powerful test $\delta \in K_\alpha$ is given by

$$\delta = \begin{cases} H_1 : & \frac{f_1(X)}{f_2(X)} > c \\ H_2 : & \frac{f_1(X)}{f_2(X)} < c \\ pH_1 + (1-p)H_2 : & \frac{f_1(X)}{f_2(X)} = c, \end{cases}$$

where in the case of equality we break the tie randomly by picking H_1 with probability p and H_2 with probability $1-p$. This test δ is called a randomized most powerful *test for two simple hypotheses at the level of significance α.*

Proof. Let us first assume that we can find c and p such that (9.0.4) holds. Then the error of type 1 for the randomized test δ is

$$\alpha_1 = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) + (1-p)\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} = c\Big) = \alpha, \tag{9.0.5}$$

since δ does not pick H_1 exactly when the likelihood ratio is less than c or when it is equal to c in which case H_1 is not picked with probability $1-p$. This means that the randomized test $\delta \in K_\alpha$. The rest of the proof repeats the proof of the previous theorem. We only need to point out that our randomized test will still be a Bayes test since in the case of equality

$$\frac{f_1(X)}{f_2(X)} = c$$

the Bayes test allows us to break the tie arbitrarily and we choose to break it randomly in a way that ensures that the error of type one will be equal to α, as in (9.0.5).

The only question left is why we can always choose c and p such that (9.0.4) is satisfied. If we look at the function

$$F(t) = \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} < t\Big)$$

as a function of t, it will increase from 0 to 1 as t increases from 0 to ∞. Let us keep in mind that, in general, $F(t)$ might have jumps. We can have two possibilities: either (a) at some point $t = c$ the function $F(c)$ will be equal to α, i.e.

$$F(c) = \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} < c\Big) = \alpha$$

or (b) at some point $t = c$ it will jump over α, i.e.

$$F(c) = \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} < c\Big) < \alpha$$

but

$$\mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} \le c\Big) = F(c) + \mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} = c\Big) \ge \alpha.$$

In both cases we find p by solving (9.0.4) with this value of c. In the case (a) we get $p = 1$ and for (b) we get

$$1 - p = (\alpha - F(c))\Big/\mathbb{P}\Big(\frac{f_1(X)}{f_2(X)} = c\Big).$$

□

Example. Suppose that we have one observation $X \sim B(p)$ from Bernoulli distribution with probability of success $p \in (0,1)$. Consider two simple hypotheses

$$\begin{aligned} H_1 : p = 0.2, \text{ i.e. } f_1(x) &= 0.2^x 0.8^{1-x}, \\ H_2 : p = 0.4, \text{ i.e. } f_2(x) &= 0.4^x 0.6^{1-x}. \end{aligned}$$

Let us take the level of significance $\alpha = 0.05$ and find the most powerful $\delta \in K_{0.05}$. In figure 9.1 we show the graph of the function

$$F(c) = \mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big).$$

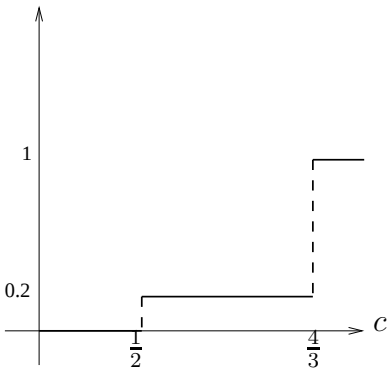


Figure 9.1: Graph of $F(c)$.

Let us explain how this graph is obtained. First of all, the likelihood ratio can take only two values:

$$\frac{f_1(X)}{f_2(X)} = \begin{cases} 1/2, \text{ if } & X = 1 \\ 4/3, \text{ if } & X = 0. \end{cases}$$

If $c \le 1/2$ then the set

$$\Big\{\frac{f_1(X)}{f_2(X)} < c\Big\} = \emptyset \text{ is empty and } F(c) = \mathbb{P}_1(\emptyset) = 0.$$

If $1/2 < c \le 4/3$ then the set

$$\Big\{\frac{f_1(X)}{f_2(X)} < c\Big\} = \{X = 1\} \text{ and } F(c) = \mathbb{P}_1(X = 1) = 0.2.$$

Finally, if $4/3 < c$ then the set

$$\Big\{\frac{f_1(X)}{f_2(X)} < c\Big\} = \{X = 0 \text{ or } 1\} \text{ and } F(c) = \mathbb{P}_1(X = 0 \text{ or } 1) = 1.$$

The function $F(c)$ jumps over the level $\alpha = 0.05$ at the point $c = 1/2$. To determine p we have to make sure that the error of type one is equal to 0.05, i.e.

$$\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} < c\Big) + (1-p)\mathbb{P}_1\Big(\frac{f_1(X)}{f_2(X)} = c\Big) = 0 + (1-p)0.2 = 0.05$$

which gives that $p = 3/4$. Therefore, the most powerful test of size $\alpha = 0.05$ is

$$\delta = \begin{cases} H_1 : & \frac{f_1(X)}{f_2(X)} > \frac{1}{2} \text{ or } X = 0 \\ H_2 : & \frac{f_1(X)}{f_2(X)} < \frac{1}{2} \text{ or never} \\ \frac{3}{4}H_1 + \frac{1}{4}H_2 : & \frac{f_1(X)}{f_2(X)} = \frac{1}{2} \text{ or } X = 1. \end{cases}$$

□

In the example above one could also consider two or more observations. Another example would be to consider, let's say, two observations from Poisson distribution $\Pi(\lambda)$ and test two simple hypotheses $H_1 : \lambda = 0.1$ vs. $H_2 : \lambda = 0.3$.

Section 10

Chi-squared goodness-of-fit test.

Example. Let us start with a Matlab example. Let us generate a vector X of 100 i.i.d. uniform random variables on $[0,1]$:

```
X=rand(100,1).
```

Parameters $(100,1)$ here mean that we generate a 100×1 matrix or uniform random variables. Let us test if the vector X comes from distribution $U[0,1]$ using χ^2 goodness-of-fit test:

```
[H,P,STATS]=chi2gof(X,'cdf',@(z)unifcdf(z,0,1),'edges',0:0.2:1)
```

The output is

```
H = 0, P = 0.0953,
STATS = chi2stat: 7.9000
        df: 4
        edges: [0 0.2 0.4 0.6 0.8 1]
        O: [17 16 24 29 14]
        E: [20 20 20 20 20]
```

We accept null hypothesis $H_0 : \mathbb{P} = U[0,1]$ at the default level of significance $\alpha = 0.05$ since the p-value 0.0953 is greater than α. The meaning of other parameters will become clear when we explain how this test works. Parameter 'cdf' takes the handle @ to a fully specified c.d.f. For example, to test if the data comes from $N(3,5)$ we would use '@(z)normcdf(z,3,5)', or to test Poisson distribution $\Pi(4)$ we would use '@(z)poisscdf(z,4).'

It is important to note that when we use chi-squared test to test, for example, the null hypothesis $H_0 : \mathbb{P} = N(1,2)$, the alternative hypothesis is $H_0 : \mathbb{P} \neq N(1,2)$. This is different from the setting of t-tests where we would assume that the data comes from normal distribution and test $H_0 : \mu = 1 \quad \text{vs.} \quad H_0 : \mu \neq 1$.

□

Pearson's theorem.

Chi-squared goodness-of-fit test is based on a probabilistic result that we will prove in this section.

Figure 10.1:

Let us consider r boxes $B_1, \ldots, B_r$ and throw n balls $X_1, \ldots, X_n$ into these boxes independently of each other with probabilities

$$\mathbb{P}(X_i \in B_1) = p_1, \ldots, \mathbb{P}(X_i \in B_r) = p_r,$$

so that

$$p_1 + \ldots + p_r = 1.$$

Let ν_j be a number of balls in the jth box:

$$\nu_j = \#\{\text{balls } X_1, \ldots, X_n \text{ in the box } B_j\} = \sum_{l=1}^{n} I(X_l \in B_j).$$

On average, the number of balls in the jth box will be np_j since

$$\mathbb{E}\nu_j = \sum_{l=1}^{n} \mathbb{E} I(X_l \in B_j) = \sum_{l=1}^{n} \mathbb{P}(X_l \in B_j) = np_j.$$

We can expect that a random variable ν_j should be close to np_j. For example, we can use a Central Limit Theorem to describe precisely how close ν_j is to np_j. The next result tells us how we can describe the closeness of ν_j to np_j simultaneously for all boxes $j \leq r$. The main difficulty in this Thorem comes from the fact that random variables ν_j for $j \leq r$ are not independent because the total number of balls is fixed

$$\nu_1 + \ldots + \nu_r = n.$$

If we know the counts in $n-1$ boxes we automatically know the count in the last box.

Theorem.(*Pearson*) *We have that the random variable*

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} \to^d \chi^2_{r-1}$$

converges in distribution to χ^2_{r-1}-distribution with $(r-1)$ degrees of freedom.

Proof. Let us fix a box B_j. The random variables

$$I(X_1 \in B_j), \ldots, I(X_n \in B_j)$$

that indicate whether each observation X_i is in the box B_j or not are i.i.d. with Bernoulli distribution $B(p_j)$ with probability of success

$$\mathbb{E}I(X_1 \in B_j) = \mathbb{P}(X_1 \in B_j) = p_j$$

and variance

$$\mathrm{Var}(I(X_1 \in B_j)) = p_j(1-p_j).$$

Therefore, by Central Limit Theorem the random variable

$$\begin{aligned} \frac{\nu_j - np_j}{\sqrt{np_j(1-p_j)}} &= \frac{\sum_{l=1}^n I(X_l \in B_j) - np_j}{\sqrt{np_j(1-p_j)}} \\ &= \frac{\sum_{l=1}^n I(X_l \in B_j) - n\mathbb{E}}{\sqrt{n\mathrm{Var}}} \to^d N(0,1) \end{aligned}$$

converges in distribution to $N(0,1)$. Therefore, the random variable

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to^d \sqrt{1-p_j}N(0,1) = N(0, 1-p_j)$$

converges to normal distribution with variance $1-p_j$. Let us be a little informal and simply say that

$$\frac{\nu_j - np_j}{\sqrt{np_j}} \to Z_j$$

where random variable $Z_j \sim N(0, 1-p_j)$.

We know that each Z_j has distribution $N(0, 1-p_j)$ but, unfortunately, this does not tell us what the distribution of the sum $\sum Z_j^2$ will be, because as we mentioned above r.v.s ν_j are not independent and their correlation structure will play an important role. To compute the covariance between Z_i and Z_j let us first compute the covariance between

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \text{ and } \frac{\nu_j - np_j}{\sqrt{np_j}}$$

which is equal to

$$\begin{aligned} &\mathbb{E}\frac{\nu_i - np_j}{\sqrt{np_i}}\frac{\nu_j - np_j}{\sqrt{np_j}} = \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - \mathbb{E}\nu_i np_j - \mathbb{E}\nu_j np_i + n^2p_ip_j) \\ &= \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - np_inp_j - np_jnp_i + n^2p_ip_j) = \frac{1}{n\sqrt{p_ip_j}}(\mathbb{E}\nu_i\nu_j - n^2p_ip_j). \end{aligned}$$

To compute $\mathbb{E}\nu_i\nu_j$ we will use the fact that one ball cannot be inside two different boxes simultaneously which means that

$$I(X_l \in B_i)I(X_l \in B_j) = 0. \tag{10.0.1}$$

Therefore,

$$\begin{aligned}\mathbb{E}\nu_i\nu_j &= \mathbb{E}\Big(\sum_{l=1}^n I(X_l\in B_i)\Big)\Big(\sum_{l'=1}^n I(X_{l'}\in B_j)\Big) = \mathbb{E}\sum_{l,l'} I(X_l\in B_i)I(X_{l'}\in B_j)\\ &= \underbrace{\mathbb{E}\sum_{l=l'} I(X_l\in B_i)I(X_{l'}\in B_j)}_{\text{this equals to 0 by (10.0.1)}} + \mathbb{E}\sum_{l\neq l'} I(X_l\in B_i)I(X_{l'}\in B_j)\\ &= n(n-1)\mathbb{E}I(X_l\in B_j)\mathbb{E}I(X_{l'}\in B_j) = n(n-1)p_ip_j.\end{aligned}$$

Therefore, the covariance above is equal to

$$\frac{1}{n\sqrt{p_ip_j}}\Big(n(n-1)p_ip_j - n^2p_ip_j\Big) = -\sqrt{p_ip_j}.$$

To summarize, we showed that the random variable

$$\sum_{j=1}^r \frac{(\nu_j - np_j)^2}{np_j} \to \sum_{j=1}^r Z_j^2.$$

where normal random variables $Z_1,\dots,Z_n$ satisfy

$$\mathbb{E}Z_i^2 = 1-p_i \text{ and covariance } \mathbb{E}Z_iZ_j = -\sqrt{p_ip_j}.$$

To prove the Theorem it remains to show that this covariance structure of the sequence of (Z_i) implies that their sum of squares has χ^2_{r-1}-distribution. To show this we will find a different representation for $\sum Z_i^2$.

Let $g_1,\dots,g_r$ be i.i.d. standard normal random variables. Consider two vectors

$$\boldsymbol{g} = (g_1,\dots,g_r)^T \text{ and } \boldsymbol{p} = (\sqrt{p_1},\dots,\sqrt{p_r})^T$$

and consider a vector $\boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p}$, where $\boldsymbol{g}\cdot\boldsymbol{p} = g_1\sqrt{p_1}+\ldots+g_r\sqrt{p_r}$ is a scalar product of $\boldsymbol{g}$ and $\boldsymbol{p}$. We will first prove that

$$\boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p} \text{ has the same joint distribution as } (Z_1,\dots,Z_r). \tag{10.0.2}$$

To show this let us consider two coordinates of the vector $\boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p}$:

$$i^{th}: g_i - \sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_i} \quad \text{and } j^{th}: g_j - \sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_j}$$

and compute their covariance:

$$\begin{aligned}&\mathbb{E}\Big(g_i - \sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_i}\Big)\Big(g_j - \sum_{l=1}^r g_l\sqrt{p_l}\sqrt{p_j}\Big)\\ &= -\sqrt{p_i}\sqrt{p_j} - \sqrt{p_j}\sqrt{p_i} + \sum_{l=1}^n p_l\sqrt{p_i}\sqrt{p_j} = -2\sqrt{p_ip_j} + \sqrt{p_ip_j} = -\sqrt{p_ip_j}.\end{aligned}$$

Similarly, it is easy to compute that

$$\mathbb{E}\Big(g_i - \sum_{l=1}^{r} g_l\sqrt{p_l}\sqrt{p_i}\Big)^2 = 1 - p_i.$$

This proves (10.0.2), which provides us with another way to formulate the convergence, namely, we have

$$\sum_{j=1}^{r}\Big(\frac{\nu_j - np_j}{\sqrt{np_j}}\Big)^2 \to^d |\boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{p})\boldsymbol{p}|^2.$$

But this vector has a simple geometric interpretation. Since vector $\boldsymbol{p}$ is a unit vector:

$$|\boldsymbol{p}|^2 = \sum_{l=1}^{r}(\sqrt{p_i})^2 = \sum_{l=1}^{r} p_i = 1,$$

vector $\boldsymbol{V_1} = (\boldsymbol{p}\cdot\boldsymbol{g})\boldsymbol{p}$ is the projection of vector $\boldsymbol{g}$ on the line along $\boldsymbol{p}$ and, therefore, vector $\boldsymbol{V_2} = \boldsymbol{g} - (\boldsymbol{p}\cdot\boldsymbol{g})\boldsymbol{p}$ will be the projection of $\boldsymbol{g}$ onto the plane orthogonal to $\boldsymbol{p}$, as shown in figure 10.2.

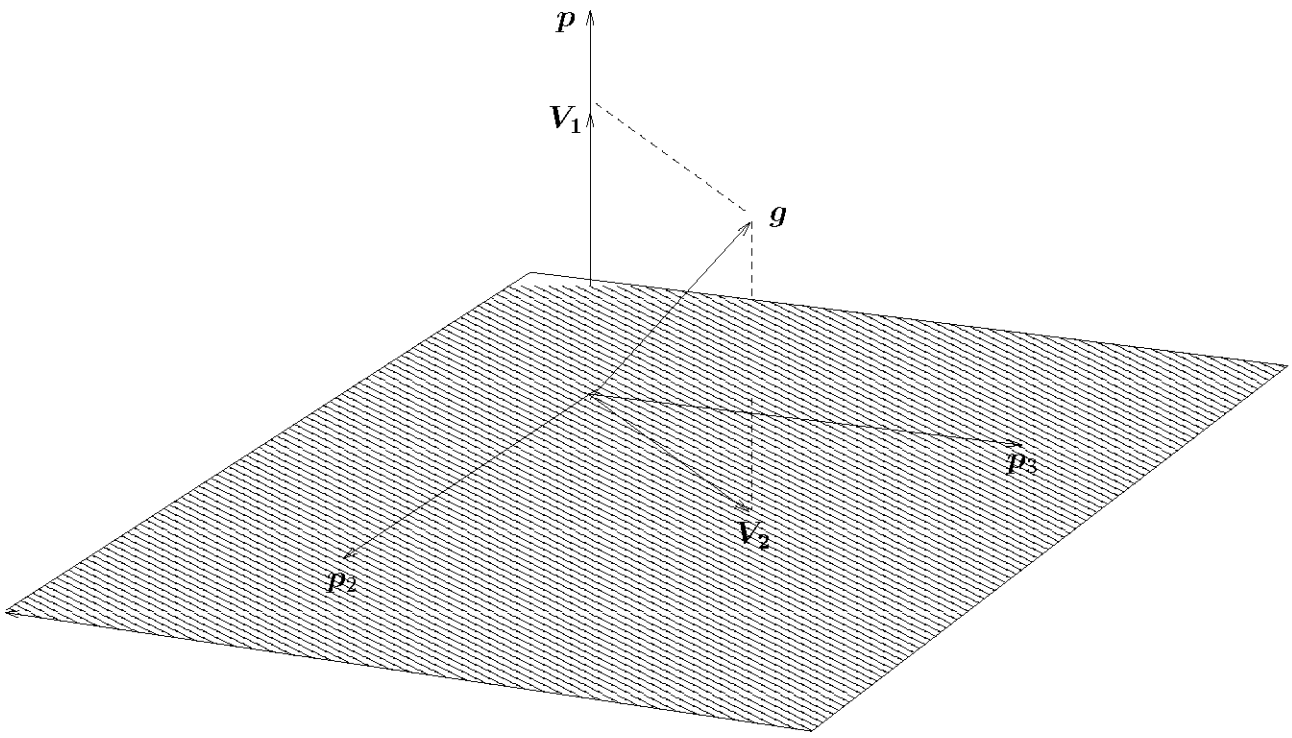


Figure 10.2: New coordinate system.

Let us consider a new orthonormal coordinate system with the first basis vector (first axis) equal to $\boldsymbol{p}$. In this new coordinate system vector $\boldsymbol{g}$ will have coordinates

$$\boldsymbol{g}' = (g_1', \ldots, g_r') = V\boldsymbol{g}$$

obtained from $\boldsymbol{g}$ by orthogonal transformation

$$V = (\boldsymbol{p}, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_r)$$

that maps canonical basis into this new basis. But we proved in Lecure 4 that in that case $g_1', \ldots, g_r'$ will also be i.i.d. standard normal. From figure 10.2 it is obvious that vector $\boldsymbol{V_2} = \boldsymbol{g} - (\boldsymbol{p} \cdot \boldsymbol{g})\boldsymbol{p}$ in the new coordinate system has coordinates

$$(0, g_2', \ldots, g_r')^T$$

and, therefore,

$$|\boldsymbol{V_2}|^2 = |\boldsymbol{g} - (\boldsymbol{p} \cdot \boldsymbol{g})\boldsymbol{p}|^2 = (g_2')^2 + \ldots + (g_r')^2.$$

But this last sum, by definition, has χ^2_{r-1} distribution since $g_2', \cdots, g_r'$ are i.i.d. standard normal. This finishes the proof of Theorem.

□

Chi-squared goodness-of-fit test for simple hypothesis.

Suppose that we observe an i.i.d. sample $X_1, \ldots, X_n$ of random variables that take a finite number of values $B_1, \ldots, B_r$ with unknown probabilities $p_1, \ldots, p_r$. Consider hypotheses

$$\begin{aligned} H_0 &: \quad p_i = p_i^\circ \text{ for all } i = 1, \ldots, r, \\ H_1 &: \quad \text{for some } i, p_i \neq p_i^\circ. \end{aligned}$$

If the null hypothesis H_0 is true then by Pearson's theorem

$$T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to^d \chi^2_{r-1}$$

where $\nu_i = \#\{X_j : X_j = B_i\}$ are the observed counts in each category. On the other hand, if H_1 holds then for some index i, $p_i \neq p_i^\circ$ and the statistics T will behave differently. If p_i is the true probability $\mathbb{P}(X_1 = B_i)$ then by CLT

$$\frac{\nu_i - np_i}{\sqrt{np_i}} \to^d N(0, 1 - p_i).$$

If we rewrite

$$\frac{\nu_i - np_i^\circ}{\sqrt{np_i^\circ}} = \frac{\nu_i - np_i + n(p_i - p_i^\circ)}{\sqrt{np_i^\circ}} = \sqrt{\frac{p_i}{p_i^\circ}} \frac{\nu_i - np_i}{\sqrt{np_i}} + \sqrt{n} \frac{p_i - p_i^\circ}{\sqrt{p_i^\circ}}$$

then the first term converges to $N(0, (1 - p_i)p_i/p_i^\circ)$ and the second term diverges to plus or minus ∞ because $p_i \neq p_i^\circ$. Therefore,

$$\frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} \to +\infty$$

which, obviously, implies that $T \to +\infty$. Therefore, as sample size n increases the distribution of T under null hypothesis H_0 will approach χ^2_{r-1}-distribution and under alternative hypothesis H_1 it will shift to $+\infty$, as shown in figure 10.3.

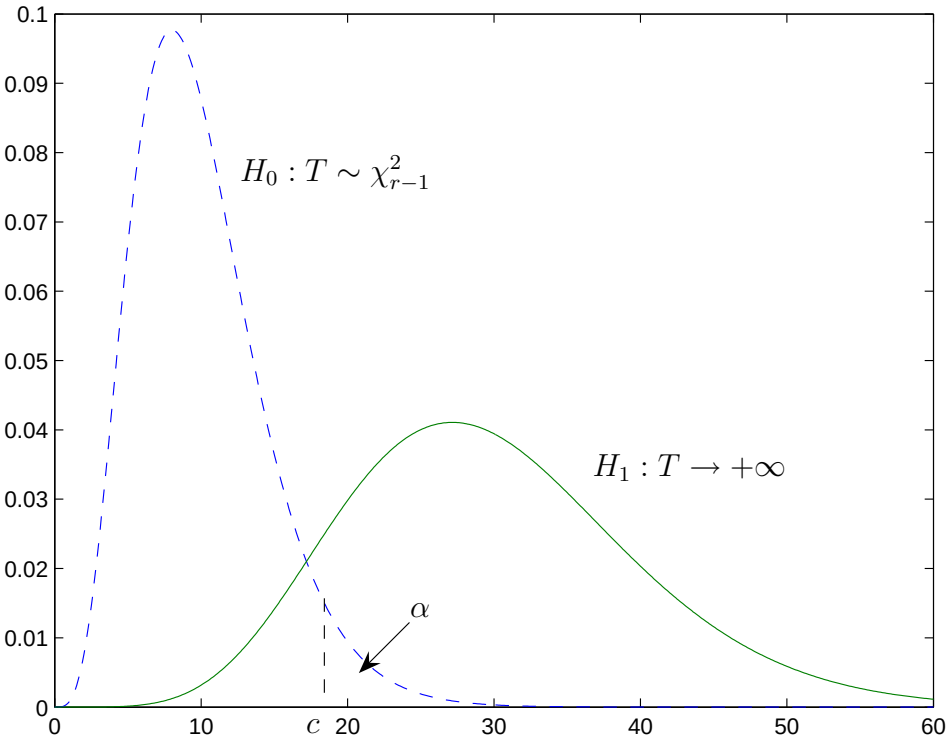


Figure 10.3: Behavior of T under H_0 and H_1.

Therefore, we define the decision rule

$$\delta = \begin{cases} H_1 : & T \leq c \\ H_2 : & T > c. \end{cases}$$

We choose the threshold c from the condition that the error of type 1 is equal to the level of significance α :

$$\alpha = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1(T > c) \approx \chi^2_{r-1}(c, \infty)$$

since under the null hypothesis the distribution of T is approximated by χ^2_{r-1} distribution. Therefore, we take c such that $\alpha = \chi^2_{r-1}(c, \infty)$. This test δ is called the *chi-squared goodness-of-fit* test.

□

Example. (*Montana outlook poll.*) In a 1992 poll 189 Montana residents were asked (among other things) whether their personal financial status was worse, the same or better than a year ago.

Worse	Same	Better	Total
58	64	67	189

We want to test the hypothesis H_0 that the underlying distribution is uniform, i.e. $p_1 = p_2 = p_3 = 1/3$. Let us take level of significance $\alpha = 0.05$. Then the threshold c in the chi-squared

test

$$\delta = \begin{cases} H_0: & T \le c \\ H_1: & T > c \end{cases}$$

is found from the condition that $\chi^2_{3-1=2}(c, \infty) = 0.05$ which gives $c = 5.9$. We compute chi-squared statistic

$$T = \frac{(58 - 189/3)^2}{189/3} + \frac{(64 - 189/3)^2}{189/3} + \frac{(67 - 189/3)^2}{189/3} = 0.666 < 5.9$$

which means that we accept H_0 at the level of significance 0.05.

□

Goodness-of-fit for continuous distribution.

Let $X_1, \ldots, X_n$ be an i.i.d. sample from unknown distribution $\mathbb{P}$ and consider the following hypotheses:

$$\begin{cases} H_0: & \mathbb{P} = \mathbb{P}_0 \\ H_1: & \mathbb{P} \neq \mathbb{P}_0 \end{cases}$$

for some particular, possibly continuous distribution $\mathbb{P}_0$. To apply the chi-squared test above we will group the values of Xs into a finite number of subsets. To do this, we will split a set of all possible outcomes $\mathcal{X}$ into a finite number of intervals $I_1, \ldots, I_r$ as shown in figure 10.4.

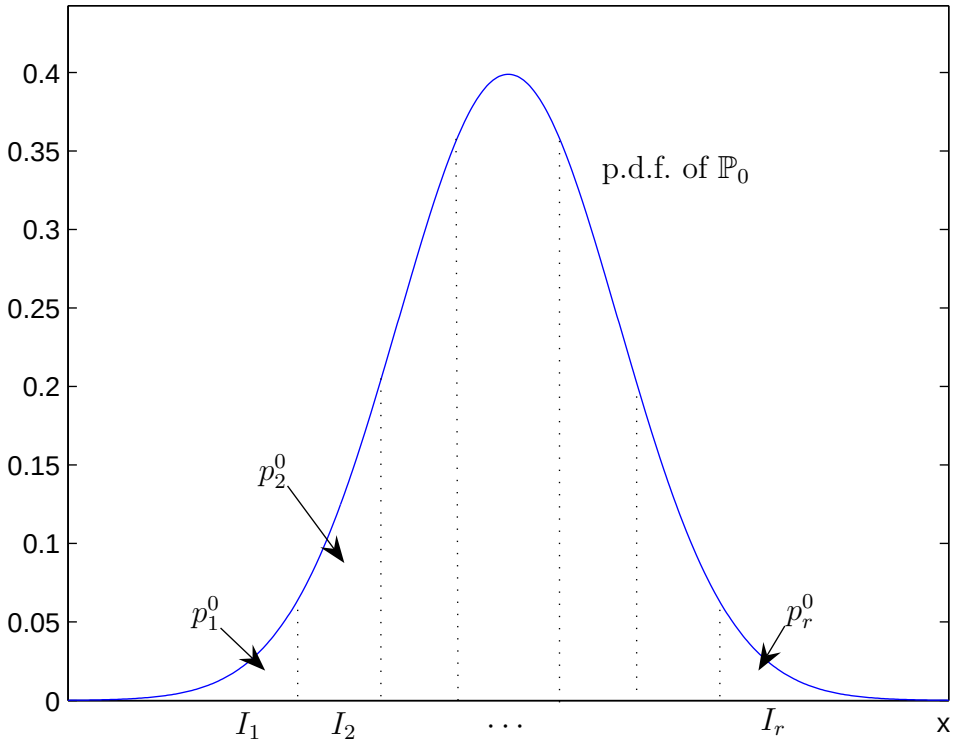


Figure 10.4: Discretizing continuous distribution.

The null hypothesis H_0, of course, implies that for all intervals

$$\mathbb{P}(X \in I_j) = \mathbb{P}_0(X \in I_j) = p_j^0.$$

Therefore, we can do chi-squared test for

$$\begin{aligned} H_0' &: \ \mathbb{P}(X \in I_j) = p_j^0 \text{ for all } j \le r \\ H_1' &: \ \text{otherwise.} \end{aligned}$$

Asking whether H_0' holds is, of course, a weaker question that asking if H_0 holds, because H_0 implies H_0' but not the other way around. There are many distributions different from $\mathbb{P}$ that have the same probabilities of the intervals $I_1, \ldots, I_r$ as $\mathbb{P}$. On the other hand, if we group into more and more intervals, our discrete approximation of $\mathbb{P}$ will get closer and closer to $\mathbb{P}$, so in some sense H_0' will get 'closer' to H_0. However, we can not split into too many intervals either, because the χ^2_{r-1}-distribution approximation for statistic T in Pearson's theorem is asymptotic. The rule of thumb is to group the data in such a way that the expected count in each interval

$$np_i^0 = n\mathbb{P}_0(X \in I_i) \ge 5$$

is at least 5. (Matlab, for example, will give a warning if this expected number will be less than five in any interval.) One approach could be to split into intervals of equal probabilities $p_i^0 = 1/r$ and choose their number r so that

$$np_i^0 = \frac{n}{r} \ge 5.$$

Example. Let us go back to the example from Lecture 2. Let us generate 100 observations from Beta distribution $B(5, 2)$.

```
X=betarnd(5,2,100,1);
```

Let us fit normal distribution $N(\mu, \sigma^2)$ to this data. The MLE $\hat{\mu}$ and $\hat{\sigma}$ are

```
mean(X) = 0.7421, std(X,1)=0.1392.
```

Note that 'std(X)' in Matlab will produce the square root of unbiased estimator $(n/n-1)\hat{\sigma}^2$. Let us test the hypothesis that the sample has this fitted normal distribution.

```
[H,P,STATS]= chi2gof(X,'cdf',@(z)normcdf(z,0.7421,0.1392))
```

outputs

```
H = 1, P = 0.0041,
STATS = chi2stat: 20.7589
        df: 7
        edges: [1x9 double]
        O: [14 4 11 14 14 16 21 6]
        E: [1x8 double]
```

Our hypothesis was rejected with p-value of 0.0041. Matlab split the real line into 8 intervals of equal probabilities. Notice 'df: 7' - the degrees of freedom $r - 1 = 8 - 1 = 7$.

□

Section 11

Goodness-of-fit for composite hypotheses.

Example. Let us consider a Matlab example. Let us generate 50 observations from $N(1, 2)$:

```
X=normrnd(1,2,50,1);
```

Then, running a chi-squared goodness-of-fit test 'chi2gof'

```
[H,P,STATS]= chi2gof(X)
```

outputs

```
H = 0, P = 0.8793,
STATS = chi2stat: 0.6742
        df: 3
        edges: [-3.7292 -0.9249 0.0099 0.9447 1.8795 2.8142 5.6186]
        O: [8 7 8 8 9 10]
        E: [8.7743 7.0639 8.7464 8.8284 7.2645 9.3226]
```

The test accepts the hypothesis that the data is normal. Notice, however, that something is different. Matlab grouped the data into 6 intervals, so chi-squared test from previous lecture should have $r - 1 = 6 - 1 = 5$ degrees of freedom, but we have 'df: 3'! The difference is that now our hypothesis is not that the data comes from a *particular given* distribution but that the data comes from a *family* of distributions which is called a *composite* hypothesis. Running

```
[H,P,STATS]= chi2gof(X,'cdf',@(z)normcdf(z,mean(X),std(X,1)))
```

would test a simple hypothesis that the data comes from a particular normal distribution $N(\hat{\mu}, \hat{\sigma}^2)$ and the output

```
H = 0, P = 0.9838
STATS =   chi2stat: 0.6842
```

```
df: 5
edges: [-3.7292 -0.9249 0.0099 0.9447 1.8795 2.8142 5.6186]
O: [8 7 8 8 9 10]
E: [8.6525 7.0995 8.8282 8.9127 7.3053 9.2017]
```

has 'df: 5.' However, we **can not** use this test because we estimate the parameters $\hat{\mu}$ and $\hat{\sigma}^2$ of this distribution using the data so this is not a particular given distribution; in fact, this is the distribution that fits the data the best, so the T statistic in Pearson's theorem will behave differently.

□

Let us start with a discrete case when a random variable takes a finite number of values $B_1, \dots, B_r$ with probabilities

$$p_1 = \mathbb{P}(X = B_1), \dots, p_r = \mathbb{P}(X = B_r).$$

We would like to test a hypothesis that this distribution comes from a family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$. In other words, if we denote

$$p_j(\theta) = \mathbb{P}_\theta(X = B_j),$$

we want to test

$$\begin{aligned} H_0 :&\ \ p_j = p_j(\theta) \text{ for all } j \le r \text{ for some } \theta \in \Theta \\ H_1 :&\ \ \text{otherwise.} \end{aligned}$$

If we wanted to test H_0 for one particular fixed θ we could use the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)},$$

and use a simple chi-squared goodness-of-fit test. The situation now is more complicated because we want to test if $p_j = p_j(\theta), j \le r$ at least for some $\theta \in \Theta$ which means that we have many candidates for θ. One way to approach this problem is as follows.

(Step 1) Assuming that hypothesis H_0 holds, i.e. $\mathbb{P} = \mathbb{P}_\theta$ for some $\theta \in \Theta$, we can find an estimate θ^* of this unknown θ and then

(Step 2) try to test if, indeed, the distribution $\mathbb{P}$ is equal to $\mathbb{P}_{\theta^*}$ by using the statistics

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)}$$

in chi-squared goodness-of-fit test.

This approach looks natural, the only question is what estimate θ^* to use and how the fact that θ^* also depends on the data will affect the convergence of T. It turns out that if we let θ^* be the maximum likelihood estimate, i.e. θ that maximizes the likelihood function

$$\varphi(\theta) = p_1(\theta)^{\nu_1} \dots p_r(\theta)^{\nu_r}$$

then the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)} \to^d \chi^2_{r-s-1} \tag{11.0.1}$$

converges to χ^2_{r-s-1} distribution with $r-s-1$ degrees of freedom, where s is the dimension of the parameter set Θ. Of course, here we assume that $s \leq r-2$ so that we have at least one degree of freedom. Very informally, by dimension we understand the number of free parameters that describe the set

$$\Big\{(p_1(\theta), \ldots, p_r(\theta)) : \theta \in \Theta\Big\}.$$

Then the decision rule will be

$$\delta = \begin{cases} H_1: & T \leq c \\ H_2: & T > c \end{cases}$$

where the threshold c is determined from the condition

$$\mathbb{P}(\delta \neq H_0 | H_0) = \mathbb{P}(T > c | H_0) \approx \chi^2_{r-s-1}(c, +\infty) = \alpha$$

where $\alpha \in [0,1]$ is the level of sidnificance.

Example 1. Suppose that a gene has two possible alleles A_1 and A_2 and the combinations of these alleles define three genotypes A_1A_1, A_1A_2 and A_2A_2. We want to test a theory that

$$\begin{gathered} \text{Probability to pass } A_1 \text{ to a child } = \theta \\ \text{Probability to pass } A_2 \text{ to a child } = 1-\theta \end{gathered}$$

and that the probabilities of genotypes are given by

$$\begin{aligned} p_1(\theta) &= \mathbb{P}(A_1A_1) = \theta^2 \\ p_2(\theta) &= \mathbb{P}(A_1A_2) = 2\theta(1-\theta) \\ p_3(\theta) &= \mathbb{P}(A_2A_2) = (1-\theta)^2. \end{aligned} \tag{11.0.2}$$

Suppose that given a random sample $X_1, \ldots, X_n$ from the population the counts of each genotype are ν_1, ν_2 and ν_3. To test the theory we want to test the hypothesis

$$\begin{aligned} H_0: &\ \ p_1 = p_1(\theta),\ p_2 = p_2(\theta),\ p_3 = p_3(\theta) \text{ for some } \theta \in [0,1] \\ H_1: &\ \ \text{otherwise.} \end{aligned}$$

First of all, the dimension of the parameter set is $s=1$ since the distributions are determined by one parameter θ. To find the MLE θ^* we have to maximize the likelihood function

$$p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3}$$

or, equivalently, maximize the log-likelihood

$$\begin{aligned} \log p_1(\theta)^{\nu_1} p_2(\theta)^{\nu_2} p_3(\theta)^{\nu_3} &= \nu_1 \log p_1(\theta) + \nu_2 \log p_2(\theta) + \nu_3 \log p_3(\theta) \\ &= \nu_1 \log \theta^2 + \nu_2 \log 2\theta(1-\theta) + \nu_3 \log (1-\theta)^2. \end{aligned}$$

If we compute the critical point by setting the derivative equal to 0, we get

$$\theta^* = \frac{2\nu_1 + \nu_2}{2n}.$$

Therefore, under the null hypothesis H_0 the statistic

$$\begin{aligned} T &= \frac{(\nu_1 - np_1(\theta^*))^2}{np_1(\theta^*)} + \frac{(\nu_2 - np_2(\theta^*))^2}{np_2(\theta^*)} + \frac{(\nu_3 - np_3(\theta^*))^2}{np_3(\theta^*)} \\ &\to^d \chi^2_{r-s-1} = \chi^2_{3-1-1} = \chi^2_1 \end{aligned}$$

converges to χ^2_1-distribution with one degree of freedom. Therefore, in the decision rule

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c \end{cases}$$

threshold c is determined by the condition

$$\mathbb{P}(\delta \neq H_0 | H_0) \approx \chi^2_1(T > c) = \alpha.$$

For example, if $\alpha = 0.05$ then $c = 3.841$.

□

Example 2. A blood type O, A, B, AB is determined by a combination of two alleles out of A, B, O and allele O is dominated by A and B. Suppose that p, q and $r = 1 - p - q$ are the population frequencies of alleles A, B and O correspondingly. If alleles are passed randomly from the parents then the probabilities of blood types will be

Blood type	Allele combinations	Probabilities	Counts
O	OO	r^2	$\nu_1 = 121$
A	AA, AO	$p^2 + 2pr$	$\nu_2 = 120$
B	BB, BO	$q^2 + 2pr$	$\nu_3 = 79$
AB	AB	$2pq$	$\nu_4 = 33$

We would like to test this theory based on the counts of each blood type in a random sample of 353 people. We have four groups and two free parameters p and q, so the chi-squared statistics T under the null hypotheses will have $\chi^2_{4-2-1} = \chi^2_1$ distribution with one degree of freedom. First, we have to find the MLE of parameters p and q. The log likelihood is

$$\begin{aligned} &\nu_1 \log r^2 + \nu_2 \log(p^2 + 2pr) + \nu_3 \log(q^2 + 2qr) + \nu_4 \log(2pq) \\ &= 2\nu_1 \log(1 - p - q) + \nu_2 \log(2p - p^2 - 2pq) + \nu_3 \log(2q - q^2 - 2pq) + \nu_4 \log(2pq). \end{aligned}$$

Unfortunately, if we set the derivatives with respect to p and q equal to zero, we get a system of two equations that is hard to solve explicitly. So instead we can minimize log likelihood numerically to get the MLE $\hat{p} = 0.247$ and $\hat{q} = 0.173$. Plugging these into formulas of blood type probabilities we get the estimated probabilities and estimated counts in each group

	O	A	B	AB
$\hat{p}_i$	0.3364	0.3475	0.2306	0.0855
$n\hat{p}_i$	118.7492	122.6777	81.4050	30.1681

We can now compute chi-squared statistic $T \approx 0.44$ and the p-value $\chi^2_1(T, \infty) = 0.5071$. The data agrees very well with the above theory.

□

We could also use a similar test when the distributions $\mathbb{P}_\theta, \theta \in \Theta$ are not necessarily supported by a finite number of points $B_1, \ldots, B_r$, for example, continuous distributions. In this case if we want to test the hypothesis

$$H_0 : \mathbb{P} = \mathbb{P}_\theta \text{ for some } \theta \in \Theta$$

we can group the data into r intervals $I_1, \ldots, I_r$ and test the hypothesis

$$H_0 : p_j = p_j(\theta) = \mathbb{P}_\theta(X \in I_j) \text{ for all } j \leq r \text{ for some } \theta.$$

For example, if we discretize normal distribution by grouping the data into intervals $I_1, \ldots, I_r$ then the hypothesis will be

$$H_0' : p_j = N(\mu, \sigma^2)(I_j) \text{ for all } j \leq r \text{ for some } (\alpha, \sigma^2).$$

There are two free parameters μ and σ^2 that describe all these probabilities so in this case $s = 2$. Matlab function 'chi2gof' tests for normality by grouping the data and computing statistic T in (11.0.1) - that is why it uses χ^2_{r-s-1} distribution with

$$r - s - 1 = r - 2 - 1 = r - 3$$

degrees of freedom and, thus, 'df: 3' in the example above.

Example. Let us test if the data 'normtemp' from normal body temperature dataset fits normal distribution.

```
[H,P,STATS]= chi2gof(normtemp)
```

gives

```
H = 0, P = 0.0504
STATS =  chi2stat: 9.4682
         df: 4
         edges: [1x8 double]
         O: [13 12 29 27 35 10 4]
         E: [9.9068 16.9874 27.6222 31.1769 24.4270 13.2839 6.5958]
```

and we accept null hypothesis at the default level of significance $\alpha = 0.05$ since p-value $0.0504 > \alpha = 0.05$. We have $r = 7$ groups and, therefore, $r - s - 1 = 7 - 2 - 1 = 4$ degrees of freedom.

□

In the case when the distributions $\mathbb{P}_\theta$ are continuous or, more generally, have infinite number of values that must be grouped in order to use chi-squared test (for example, normal or Poisson distribution), it can be a difficult numerical problem to maximize the "grouped" likelihood function

$$\mathbb{P}_\theta(I_1)^{\nu_1} \cdot \ldots \cdot \mathbb{P}_\theta(I_r)^{\nu_r} \to \max_\theta \to \theta^*.$$

It is tempting to use a usual non-grouped MLE $\hat{\theta}$ of θ instead of the above θ^* because it is often easier to compute, in fact, for many distributions we know explicit formulas for these MLEs. However, if we use $\hat{\theta}$ in the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \tag{11.0.3}$$

then it will no longer converge to χ^2_{r-s-1} distribution. A famous result in [1] proves that typically this T will converge to a distribution "in between" χ^2_{r-s-1} and χ^2_{r-1}. Intuitively this is easy to understand because θ^* specifically fits the grouped data $\nu_1, \ldots, \nu_r$ so the expected counts

$$np_1(\theta^*), \ldots, np_r(\theta^*)$$

should be a better fit compared to the expected counts

$$np_1(\hat{\theta}), \ldots, np_r(\hat{\theta}).$$

On the other hand, these last expected counts should be a better fit than simply using the true expected counts

$$np_1(\theta_0), \ldots, np_r(\theta_0)$$

since the MLE $\hat{\theta}$ fits the data better than the true distribution. So typically we would expect

$$\sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)} \leq \sum_{j=1}^{r} \frac{(\nu_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \leq \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta_0))^2}{np_j(\theta_0)}.$$

But the left hand side converges to χ^2_{r-s-1} and the right hand side converges to χ^2_{r-1}. Thus, if the decision rule is based on the statistic (11.0.3):

$$\delta = \begin{cases} H_1: & T \leq c \\ H_2: & T > c \end{cases}$$

then the threshold c can be determined conservatively from the tail of χ^2_{r-1} distribution since

$$\mathbb{P}(\delta \neq H_0 | H_0) = \mathbb{P}(T > c) \leq \chi^2_{r-1}(T > c) = \alpha.$$

□

References:

[1] Chernoff, Herman; Lehmann, E. L. (1954) The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Ann. Math. Statistics* **25**, pp. 579-586.

Section 12

Tests of independence and homogeneity.

In this lecture we will consider a situation when our observations are classified by two different features and we would like to test if these features are independent. For example, we can ask if the number of children in a family and family income are independent. Our sample space $\mathcal{X}$ will consist of $a \times b$ pairs

$$\mathcal{X} = \{(i,j) : i = 1, \dots, a,\ j = 1, \dots, b\}$$

where the first coordinate represents the first feature that belongs to one of a categories and the second coordinate represents the second feature that belongs to one of b categories. An i.i.d. sample $X_1, \dots, X_n$ can be represented by a *contingency table* below where N_{ij} is the number all observations in a cell (i,j).

Table 12.1: Contingency table.

	Feature 2			
Feature 1	1	2	$\cdots$	b
1	N_{11}	N_{12}	$\cdots$	N_{1b}
2	N_{21}	N_{22}	$\cdots$	N_{2b}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a	N_{a1}	N_{a2}	$\cdots$	N_{ab}

We would like to test the independence of two features which means that

$$\mathbb{P}(X = (i,j)) = \mathbb{P}(X^1 = i)\mathbb{P}(X^2 = j).$$

If we introduce the notations

$$\mathbb{P}(X = (i,j)) = \theta_{ij},\ \mathbb{P}(X^1 = i) = p_i \text{ and } \mathbb{P}(X^2 = j) = q_j,$$

then we want to test that for all i and j we have $\theta_{ij} = p_i q_j$. Therefore, our hypotheses can be formulated as follows:

$$\begin{aligned} H_0 &: \ \theta_{ij} = p_i q_j \text{ for all } (i,j) \text{ for some } (p_1, \ldots, p_a) \text{ and } (q_1, \ldots, q_b) \\ H_1 &: \ \text{otherwise.} \end{aligned}$$

We can see that this null hypothesis H_0 is a special case of the composite hypotheses from previous lecture and it can be tested using the chi-squared goodness-of-fit test. The total number of groups is $r = a \times b$. Since p_is and q_js should add up to one

$$p_1 + \ldots + p_a = 1 \text{ and } q_1 + \ldots + q_b = 1$$

one parameter in each sequence, for example p_a and q_b, can be computed in terms of other probabilities and we can take $(p_1, \ldots, p_{a-1})$ and $(q_1, \ldots, q_{b-1})$ as free parameters of the model. This means that the dimension of the parameter set is

$$s = (a-1) + (b-1).$$

Therefore, if we find the maximum likelihood estimates for the parameters of this model then the chi-squared statistic:

$$T = \sum_{i,j} \frac{(N_{ij} - n p_i^* q_j^*)^2}{n p_i^* q_j^*} \to \chi^2_{r-s-1} = \chi^2_{ab-(a-1)-(b-1)-1} = \chi^2_{(a-1)(b-1)}$$

converges in distribution to $\chi^2_{(a-1)(b-1)}$ distribution with $(a-1)(b-1)$ degrees of freedom. To formulate the test it remains to find the maximum likelihood estimates of the parameters. We need to maximize the likelihood function

$$\prod_{i,j} (p_i q_j)^{N_{ij}} = \prod_i p_i^{\sum_j N_{ij}} \prod_j q_j^{\sum_i N_{ij}} = \prod_i p_i^{N_{i+}} \prod_j q_j^{N_{+j}}$$

where we introduced the notations

$$N_{i+} = \sum_j N_{ij} \quad \text{and} \quad N_{+j} = \sum_i N_{ij}$$

for the total number of observations in the ith row and jth column. Since p_is and q_js are not related to each other, maximizing the likelihood function above is equivalent to maximizing $\prod_i p_i^{N_{i+}}$ and $\prod_j q_j^{N_{+j}}$ separately. Let us maximize $\prod_{i=1}^a p_i^{N_{i+}}$ or, taking the logarithm, maximize

$$\sum_{i=1}^a N_{i+} \log p_i = \sum_{i=1}^{a-1} N_{i+} \log p_i + N_{a+} \log(1 - p_1 - \ldots - p_a),$$

since the probabilities add up to one. Setting derivative in p_i equal to zero, we get

$$\frac{N_{i+}}{p_i} - \frac{N_{a+}}{1 - p_1 - \ldots - p_{a-1}} = \frac{N_{i+}}{p_i} - \frac{N_{a+}}{p_a} = 0$$

or $N_{i+}p_a = N_{a+}p_i$. Adding up these equations for all $i \le a$ gives

$$np_a = N_{a+} \Longrightarrow p_a = \frac{N_{a+}}{n} \Longrightarrow p_i = \frac{N_{i+}}{n}.$$

Therefore, we get that the MLE for p_i:

$$p_i^* = \frac{N_{i+}}{n}.$$

Similarly, the MLE for q_j is:

$$q_j^* = \frac{N_{+j}}{n}.$$

Therefore, chi-square statistic T in this case can be written as

$$T = \sum_{i,j} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n}$$

and the decision rule is given by

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c \end{cases}$$

where the threshold is determined from the condition

$$\chi^2_{(a-1)(b-1)}(c, +\infty) = \alpha.$$

Example. In 1992 poll 189 Montana residents were asked whether their personal financial status was worse, the same or better than one year ago. The opinions were divided into three groups by income range: under 20K, between 20K and 35K, and over 35K. We would like to test if opinions were independent of income.

Table 12.2: Montana outlook poll.

	$b = 3$			
$a = 3$	Worse	Same	Better	
$\le$ 20K	20	15	12	47
(20K, 35K)	24	27	32	83
$\ge$ 35K	14	22	23	59
	58	64	67	189

The chi-squared statistic is

$$T = \frac{(20 - 47 \times 58/189)^2}{47 \times 58/189} + \ldots + \frac{(23 - 67 \times 59/189)^2}{67 \times 59/189} = 5.21.$$

If we take level of significance $\alpha = 0.05$ then the threshold c is:

$$\chi^2_{(a-1)(b-1)}(c, +\infty) = \chi^2_4(c, \infty) = \alpha = 0.05 \Rightarrow c = 9.488.$$

Since $T = 5.21 < c = 9.488$ we accept the null hypothesis that opinions are independent of income.

□

Test of homogeneity.

Suppose that the population is divided into R groups and each group (or the entire population) is divided into C categories. We would like to test whether the distribution of categories in each group is the same.

Table 12.3: Test of homogeneity

	Category 1	$\cdots$	Category C	$\sum$
Group 1	N_{11}	$\cdots$	N_{1C}	N_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group R	N_{R1}	$\cdots$	N_{RC}	N_{R+}
$\sum$	N_{+1}	$\cdots$	N_{+C}	n

If we denote

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_{ij}$$

so that for each group $i \leq R$ we have

$$\sum_{j=1}^{C} p_{ij} = 1$$

then we want to test the following hypotheses:

$$\begin{aligned} H_0 &: \ p_{ij} = p_j \text{ for all groups } i \leq R \\ H_1 &: \ \text{otherwise} \end{aligned}$$

If observations $X_1, \ldots, X_n$ are sampled independently from the entire population then homogeneity over groups is the same as independence of groups and categories. Indeed, if have homogeneity

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = \mathbb{P}(\text{Category}_j)$$

then we have

$$\mathbb{P}(\text{Group}_i, \text{Category}_j) = \mathbb{P}(\text{Category}_j|\text{Group}_i)\mathbb{P}(\text{Group}_i) = \mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i)$$

which means the groups and categories are independent. Another way around, if we have independence then

$$\begin{aligned} \mathbb{P}(\text{Category}_j|\text{Group}_i) &= \frac{\mathbb{P}(\text{Group}_i, \text{Category}_j)}{\mathbb{P}(\text{Group}_i)} \\ &= \frac{\mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i)}{\mathbb{P}(\text{Group}_i)} = \mathbb{P}(\text{Category}_j) \end{aligned}$$

which is homogeneity. This means that to test homogeneity we can use the test of independence above.

Interestingly, the same test can be used in the case when the sampling is done not from the entire population but from each group separately which means that we decide a priori about the sample size in each group - $N_{1+}, \dots, N_{R+}$. When we sample from the entire population these numbers are random and by the LLN N_{i+}/n will approximate the probability $\mathbb{P}(\text{Group}_i)$, i.e. N_{i+} reflects the proportion of group i in the population. When we pick these numbers a priori one can simply think that we artificially renormalize the proportion of each group in the population and test for homogeneity among groups as independence in this new artificial population. Another way to argue that the test will be the same is as follows. Assume that

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_j$$

where the probabilities p_j are all given. Then by Pearson's theorem we have the convergence in distribution

$$\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_j)^2}{N_{i+}p_j} \to \chi^2_{C-1}$$

for each group $i \le R$ which implies that

$$\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}p_j)^2}{N_{i+}p_j} \to \chi^2_{R(C-1)}$$

since the samples in different groups are independent. If now we assume that probabilities $p_1, \dots, p_C$ are unknown and plug in the maximum likelihood estimates $p_j^* = N_{+j}/n$ then

$$\sum_{i=1}^{R}\sum_{j=1}^{C} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n} \to \chi^2_{R(C-1)-(C-1)} = \chi^2_{(R-1)(C-1)}$$

because we have $C-1$ free parameters $p_1, \dots, p_{C-1}$ and estimating each unknown parameter results in losing one degree of freedom.

Example (Textbook, page 560). In this example, 100 people were asked whether the service provided by the fire department in the city was satisfactory. Shortly after the survey, a large fire occured in the city. Suppose that the **same** 100 people were asked whether they thought that the service provided by the fire department was satisfactory. The result are in the following table:

	Satisfactory	Unsatisfactory
Before fire	80	20
After fire	72	28

Suppose that we would like to test whether the opinions changed after the fire by using a chi-squared test. However, the i.i.d. sample consisted of pairs of opinions of 100 people

$$(X_1^1, X_1^2), \dots, (X_{100}^1, X_{100}^2)$$

where the first coordinate/feature is a person's opinion before the fire and it belongs to one of two categories

$$\{\text{“Satisfactory“, ”Unsatisfactory”}\},$$

and the second coordinate/feature is a person's opinion after the fire and it also belongs to one of two categories

$$\{\text{“Satisfactory“, ”Unsatisfactory”}\}.$$

So the correct contingency table corresponding to the above data and satisfying the assumption of the chi-squared test would be the following:

	Sat. before	Uns. before
Sat. after	70	10
Uns. after	2	18

In order to use the first contingency table, we would have to poll 100 people after the fire independently of the 100 people polled before the fire.

□

Section 13

Kolmogorov-Smirnov test.

Suppose that we have an i.i.d. sample $X_1, \ldots, X_n$ with some unknown distribution $\mathbb{P}$ and we would like to test the hypothesis that $\mathbb{P}$ is equal to a particular distribution $\mathbb{P}_0$, i.e. decide between the following hypotheses:

$$H_0 : \mathbb{P} = \mathbb{P}_0, \quad H_1 : \mathbb{P} \neq \mathbb{P}_0.$$

We already know how to test this hypothesis using chi-squared goodness-of-fit test. If distribution $\mathbb{P}_0$ is continuous we had to group the data and consider a weaker discretized null hypothesis. We will now consider a different test for H_0 based on a very different idea that avoids this discretization.

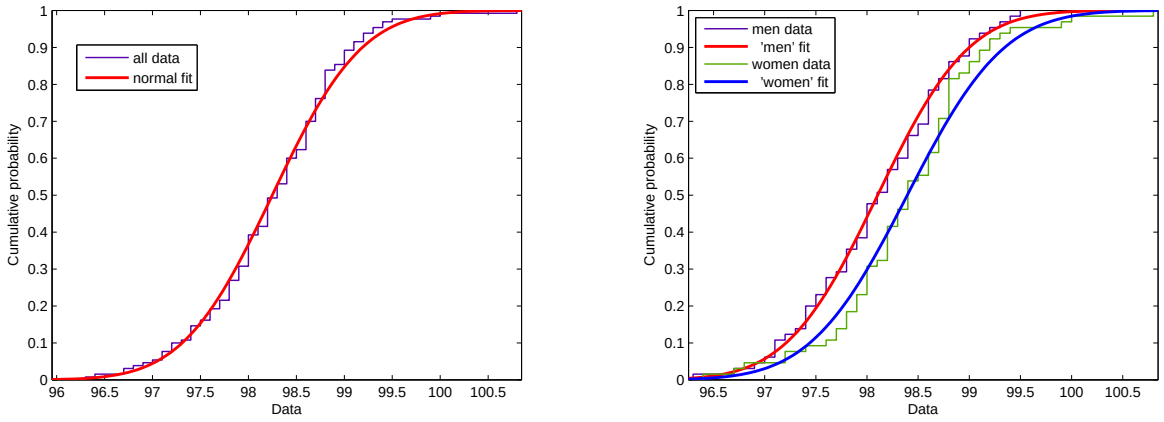


Figure 13.1: (a) Normal fit to the entire sample. (b) Normal fit to men and women separately.

Example.(*KS test*) Let us again look at the normal body temperature dataset. Let 'all' be a vector of all 130 observations and 'men' and 'women' be vectors of length 65 each corresponding to men and women. First, we fit normal distribution to the entire set 'all'. MLE $\hat{\mu}$ and $\hat{\sigma}$ are

```
mean(all) = 98.2492, std(all,1) = 0.7304.
```

We see in figure 13.1 (a) that this distribution fits the data very well. Let us perfom KS test that the data comes from this distribution $N(\hat{\mu}, \hat{\sigma}^2)$. To run the test, first, we have to create a vector of $N(\hat{\mu}, \hat{\sigma}^2)$ c.d.f. values on the sample 'all' (it is a required input in Matlab KS test function):

```
CDFall=normcdf(all,mean(all),std(all,1));
```

Then we run Matlab 'kstest' function

```
[H,P,KSSTAT,CV] = kstest(all,[all,CDFall],0.05)
```

which outputs

```
H = 0, P = 0.6502, KSSTAT = 0.0639, CV = 0.1178.
```

We accept H_0 since the p-value is 0.6502. 'CV' is a critical value such that H_0 is rejected if statistic 'KSSTAT'>'CV'.

□

Remark. KS test is designed to test a simple hypothesis $\mathbb{P} = \mathbb{P}_0$ for a *given specified* distribution $\mathbb{P}_0$. In the example above we estimated this distribution, $N(\hat{\mu}, \hat{\sigma}^2)$ from the data so, formally, KS is inaccurate in this case. There is a version of KS test, called Lilliefors test, that tests normality of the distribution by comparing the data with a fitted normal distribution as we did above, but with a correction to give a more accurate approximation of the distribution of the test statistic.

Example. (*Lilliefors test.*) We use Matlab function

```
[H,P,LSTAT,CV] = lillietest(all)
```

that outputs

```
H = 0, P = 0.1969, LSTAT = 0.0647, CV = 0.0777.
```

We accept the normality of 'all' with p-value 0.1969.

□

Example. (*KS test for two samples.*) Next, we fit normal distributions to 'men' and 'women' separately, see figure 13.1 (b). We see that they are slightly different so it is a natural question to ask whether this difference is statistically significant. We already looked at this problem in the lecture on t-tests. Under a reasonable assumption that body temperatures of men and women are normally distributed, all t-tests - paired, with equal variances and with unequal variances - rejected the hypothesis that the mean body temperatures are equal $\mu_{men} = \mu_{women}$. In this section we will describe a KS test for two samples that tests the hypothesis $H_0 : \mathbb{P}_1 = \mathbb{P}_2$ that two samples come from the same distribution. Matlab function 'kstest2'

```
[H,P,KSSTAT] = kstest2(men, women)
```

outputs

```
H = 0, P =   0.1954, KSSTAT = 0.1846.
```

It accepts the null hypothesis since p-value $0.1954 > 0.05 = \alpha$ - a default value of the level of significance. According to this test, the difference between two samples is not significant enough to say that they have different distribution.

□

Let us now explain some ideas behind these tests. Let us denote by $F(x) = \mathbb{P}(X_1 \le x)$ a c.d.f. of a true underlying distribution of the data. We define an *empirical* c.d.f. by

$$F_n(x) = \mathbb{P}_n(X \le x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

that counts the proportion of the sample points below level x. For any fixed point $x \in \mathbb{R}$ the law of large numbers implies that

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) \to \mathbb{E}I(X_1 \le x) = \mathbb{P}(X_1 \le x) = F(x),$$

i.e. the proportion of the sample in the set $(-\infty, x]$ approximates the probability of this set. It is easy to show from here that this approximation holds uniformly over all $x \in \mathbb{R}$:

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

i.e. the largest difference between F_n and F goes to 0 in probability. The key observation in the Kolmogorov-Smirnov test is that the distribution of this supremum does not depend on the 'unknown' distribution $\mathbb{P}$ of the sample, if $\mathbb{P}$ is continuous distribution.

Theorem 1. *If $F(x)$ is continuous then the distribution of*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$$

does not depend on F.

Proof. Let us define the inverse of F by

$$F^{-1}(y) = \min\{x : F(x) \ge y\}.$$

Then making the change of variables $y = F(x)$ or $x = F^{-1}(y)$ we can write

$$\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le t) = \mathbb{P}(\sup_{0\le y\le 1} |F_n(F^{-1}(y)) - y| \le t).$$

Using the definition of the empirical c.d.f. F_n we can write

$$F_n(F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le F^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n} I(F(X_i) \le y)$$

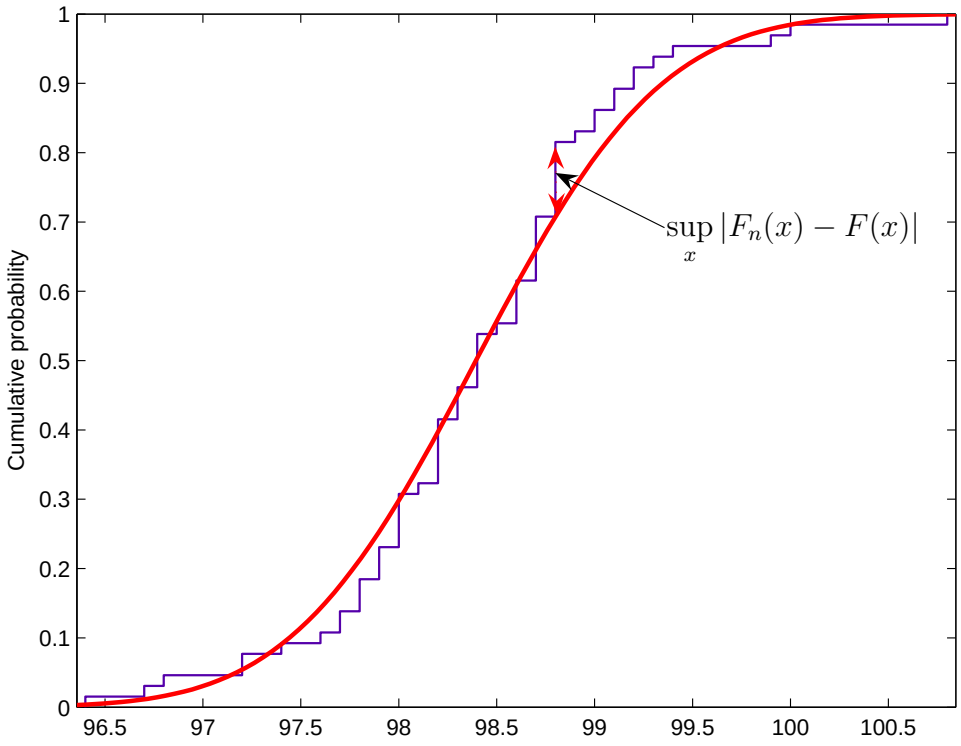


Figure 13.2: Kolmogorov-Smirnov test statistic.

and, therefore,

$$\mathbb{P}(\sup_{0\le y\le 1} |F_n(F^{-1}(y)) - y| \le t) = \mathbb{P}\Big(\sup_{0\le y\le 1} \Big|\frac{1}{n}\sum_{i=1}^{n} I(F(X_i)\le y) - y\Big| \le t\Big).$$

The distribution of $F(X_i)$ is uniform on the interval $[0,1]$ because the c.d.f. of $F(X_1)$ is

$$\mathbb{P}(F(X_1)\le t) = \mathbb{P}(X_1 \le F^{-1}(t)) = F(F^{-1}(t)) = t.$$

Therefore, the random variables

$$U_i = F(X_i) \text{ for } i \le n$$

are independent and have uniform distribution on $[0,1]$, so we proved that

$$\mathbb{P}(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le t) = \mathbb{P}\Big(\sup_{0\le y\le 1} \Big|\frac{1}{n}\sum_{i=1}^{n} I(U_i\le y) - y\Big| \le t\Big)$$

which is clearly independent of F.

□

To motivate KS test, we will need one more result which we will formulate without proof. First of all, let us note that for a fixed point x the CLT implies that

$$\sqrt{n}(F_n(x) - F(x)) \to^d N\Big(0, F(x)(1-F(x))\Big)$$

because $F(x)(1-F(x))$ is the variance of $I(X_1 \le x)$. If turns out that if we consider

$$\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)|$$

it will also converge in distribution.

Theorem 2. *We have,*

$$\mathbb{P}\Big(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \le t\Big) \to H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}$$

where $H(t)$ is the c.d.f. of Kolmogorov-Smirnov distribution.

□

Let us reformulate the hypotheses in terms of cumulative distribution functions:

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0,$$

where F_0 is the c.d.f. of $\mathbb{P}_0$. Let us consider the following statistic

$$D_n = \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F_0(x)|.$$

If the null hypothesis is true then, by Theorem 1, we distribution of D_n can be tabulated (it will depend only on n). Moreover, if n is large enough then the distribution of D_n is approximated by Kolmogorov-Smirnov distribution from Theorem 2. On the other hand, suppose that the null hypothesis fails, i.e. $F \neq F_0$. Since F is the true c.d.f. of the data, by law of large numbers the empirical c.d.f. F_n will converge to F and as a result it will not approximate F_0, i.e. for large n we will have

$$\sup_x |F_n(x) - F_0(x)| > \delta$$

for some small enough δ. Multiplying this by $\sqrt{n}$ implies that

$$D_n = \sqrt{n}\sup_{x\in\mathbb{R}}|F_n(x) - F_0(x)| > \sqrt{n}\delta.$$

If H_0 fails then $D_n > \sqrt{n}\delta \to +\infty$ as $n \to \infty$. Therefore, to test H_0 we will consider a decision rule

$$\delta = \begin{cases} H_0 : & D_n \le c \\ H_1 : & D_n > c \end{cases}$$

The threshold c depends on the level of significance α and can be found from the condition

$$\alpha = \mathbb{P}(\delta \neq H_0 | H_0) = \mathbb{P}(D_n \ge c | H_0).$$

Since under H_0 the distribution of D_n can be tabulated for each n, we can find the threshold $c = c_\alpha$ from the tables. In fact, most statistical table books have these distributions for n up to 100. Seems like Matlab has these tables built in the 'kstest' but the distribution of D_n is not available as a separate function. When n is large then we can use KS distribution to find c since

$$\alpha = \mathbb{P}(D_n \geq c | H_0) \approx 1 - H(c).$$

and we can use the table for H to find c.

□

KS test for two samples.

Kolmogorov-Smirnov test for two samples is very similar. Suppose that a first sample $X_1, \ldots, X_m$ of size m has distribution with c.d.f. $F(x)$ and the second ssmple $Y_1, \ldots, Y_n$ of size n has distribution with c.d.f. $G(x)$ and we want to test

$$H_0 : F = G \quad \text{vs.} \quad H_1 : F \neq G.$$

If $F_m(x)$ and $G_n(x)$ are corresponding empirical c.d.f.s then the statistic

$$D_{mn} = \Big(\frac{mn}{m+n}\Big)^{1/2} \sup_x |F_m(x) - G_n(x)|$$

satisfies Theorems 1 and 2 and the rest is the same

□

Example. Let us consider a sample of size 10:

$$0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64$$

and let us test the hypothesis that the distribution of the sample is uniform on $[0, 1]$ i.e. $H_0 : F(x) = F_0(x) = x$. The figure 13.3 shows the c.d.f. F_0 and empirical c.d.f. $F_n(x)$. To compute D_n we notice that the largest difference between $F_0(x)$ and $F_n(x)$ is achieved either before or after one of the jumps, i.e.

$$\sup_{0 \leq x \leq 1} |F_n(x) - F(x)| = \max_{1 \leq i \leq n} \begin{cases} |F_n(X_i^-) - F(X_i)| & \text{- before the } i\text{th jump} \\ |F_n(X_i) - F(X_i)| & \text{- after the } i\text{th jump.} \end{cases}$$

Writing these differences for our data we get

before the jump	after the jump
$\lvert 0 - 0.23 \rvert$	$\lvert 0.1 - 0.23 \rvert$
$\lvert 0.1 - 0.33 \rvert$	$\lvert 0.2 - 0.33 \rvert$
$\lvert 0.2 - 0.42 \rvert$	$\lvert 0.3 - 0.42 \rvert$
$\lvert 0.3 - 0.43 \rvert$	$\lvert 0.4 - 0.43 \rvert$
$\ldots$	

The largest value will be achieved at $|0.9 - 0.64| = 0.26$ and, therefore,

$$D_n = \sqrt{n} \sup_{0 \leq x \leq 1} |F_n(x) - x| = \sqrt{10} \times 0.26 = 0.82.$$

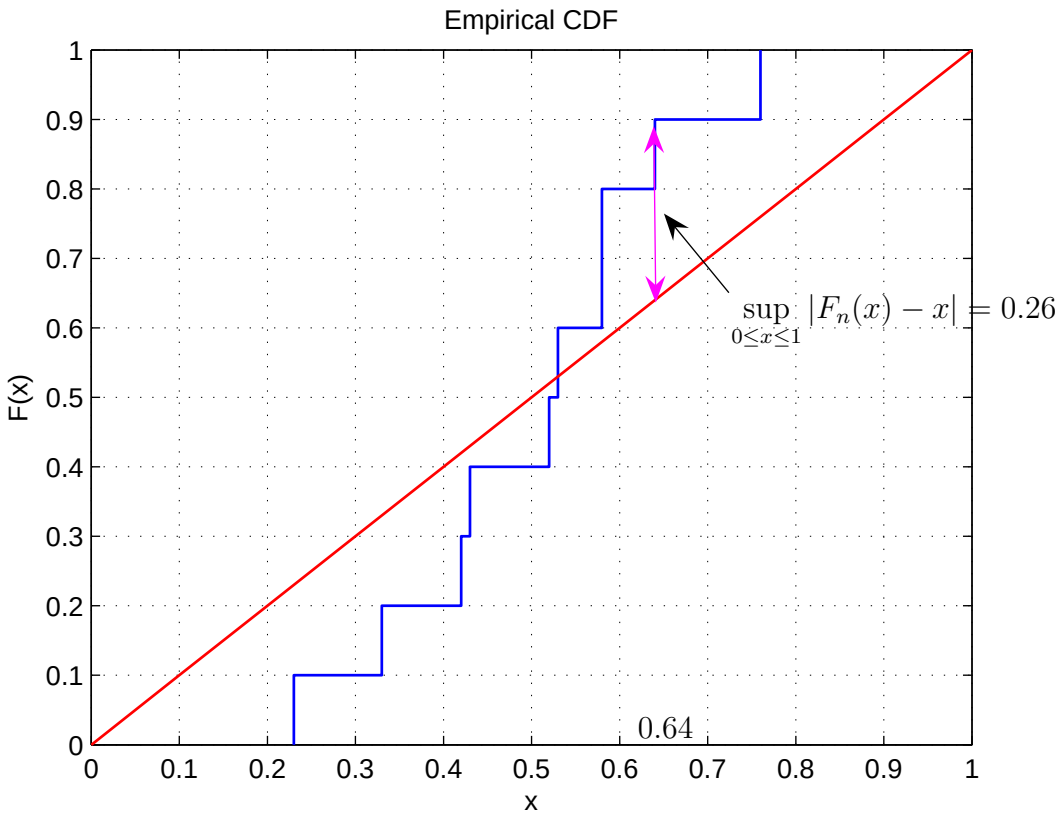


Figure 13.3: F_n and F_0 in the example.

If we take the level of significance $\alpha = 0.05$ and use KS approximation of Theorem 2 to find threshold c:

$$1 - H(c) = 0.05 \Rightarrow c = 1.35,$$

then according to KS test

$$\delta = \begin{cases} H_1: & D_n \le 1.35 \\ H_2: & D_n > 1.35 \end{cases}$$

we accept the null hypothesis H_0 since $D_n = 0.82 < c = 1.35$.

However, we have only $n = 10$ observations so the approximation of Theorem 2 might be inaccurate. We could use the advanced statistical tables to find the distibution of D_n for $n = 10$ or let Matlab do it. Running

```
[H,P,KSSTAT,CV] = kstest(X,[X,X],0.05)
```

(remark[1]) outputs

```
H = 0, P = 0.4466, KSSTAT = 0.2600, CV = 0.4093.
```

[1]Here the second input of 'kstest' should be a $n \times 2$ matrix where the first column is the data X and the second column is the corresponding values of c.d.f. $F_0(x)$. But since we test with $F_0(x) = x$, the second column is equal to X and, thus, we input '[X,X]'

Since Matlab function 'kstest' does not scale the statistic by $\sqrt{n}$ since it is using the exact distribution of $\sup_x |F_n(x) - F(x)|$ instead of approximation of Theorem 2, the critical value 'CV' mupliplied by $\sqrt{n}$, i.e. $\sqrt{10} \times 0.4093 = 1.294$ will be exactly our threshold such that

$$\mathbb{P}(D_n > c|H_0) = \alpha = 0.05.$$

It is slightly different from $c = 1.35$ given by the approximation of Theorem 2. So for small sample sizes it is better to use the exact distribution of D_n.

□

Section 14

Simple linear regression.

Let us look at the 'cigarette' dataset from [1] (available to download from journal's website) and [2]. The cigarette dataset contains measurements of tar, nicotine, weight and carbon monoxide (CO) content for 25 brands of domestic cigarettes. We are going to try to predict CO as a function of tar and nicotine content. To visualize the data let us plot each of these variable against others, see figure 14.1. Since the variables seem to have a linear relationship we fit a least-squares line, which we will explain below, to fit the data using Matlab tool 'polytool'. For example, if our vectors are 'nic' for nicotine, 'tar' for tar and 'carb' for CO then, for example, using

```
polytool(nic,carb,1)
```

will produce figure 14.1 (a), etc. We can also perform statistical analysis of these fits, in a sense that will gradually be explained below, using Matlab 'regress' function. For carbon monoxide vs. tar:

```
[b,bint,r,rint,stats]=regress(carb,[ones(25,1),tar]);

b =  2.7433     bint = 1.3465  4.1400
     0.8010            0.6969  0.9051

stats = 0.9168  253.3697  0.000  1.9508,
```

for carbon monoxide vs. nicotine

```
[b,bint,r,rint,stats]=regress(carb,[ones(25,1),nic]);

b = 1.6647      bint =  -0.3908   3.7201
    12.3954               10.2147  14.5761

stats = 0.8574  138.2659  0.000  3.3432
```

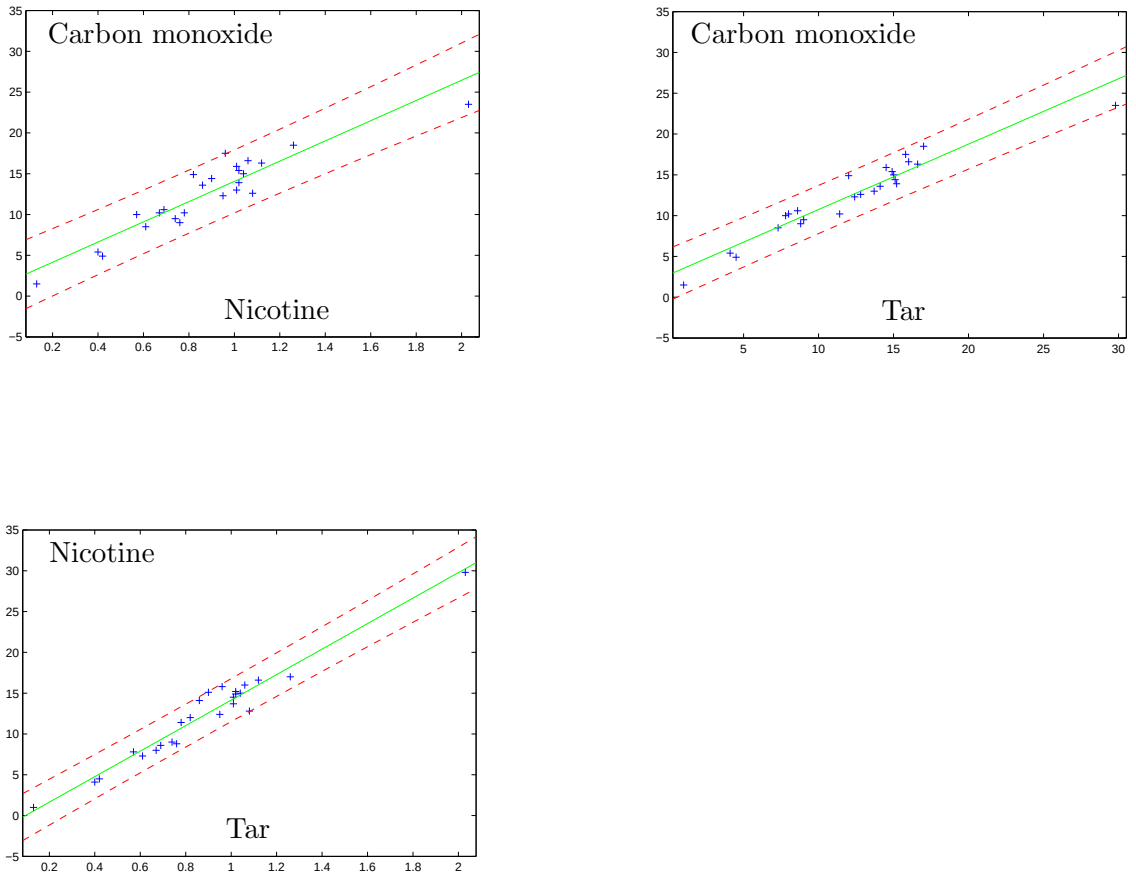


Figure 14.1: Least-squares line (solid line). (a) Carbon monoxide content (mg.) vs. nicotine content (mg.). (b) Carbon monoxide vs. tar content. (c) Tar content vs. nicotine content.

and for nicotine vs. tar

```
[b,bint,r,rint,stats]=regress(tar,[ones(25,1),nic]);

b = -1.4805     bint = -2.8795    -0.0815
    15.6281              14.1439    17.1124

stats = 0.9538  474.4314  0.000  1.5488
```

The output of 'regress' gives a vector 'b' of parameters of a fitted least-squares line, 95% confidence intervals 'bint' for these parameters, and 'stats' contains in order:

R^2 statistic, F statistic, p-value of F statistic, MLE $\hat{\sigma}^2$ of the error variance.

All of these will be explained below.

□

Simple linear regression model.

Suppose that we have a pair of variables (X, Y) and a variable Y is a linear function of X plus random noise:

$$Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon,$$

where a random noise ε is assumed to have normal distribution $N(0, \sigma^2)$. A variable X is called a predictor variable, Y - a response variable and a function $f(x) = \beta_0 + \beta_1 x$ - a linear regression function.

Suppose that we are given a sequence of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ that are described by the above model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We have three unknown parameters - β_0, β_1 and σ^2 - and we want to estimate them using a given sample. The points $X_1, \ldots, X_n$ can be either random or non random, but from the point of view of estimating linear regression function the nature of Xs is in some sense irrelevant so we will think of them as fixed and non random and assume that the randomness comes from the noise variables ε_i. For a fixed X_i, the distribution of Y_i is equal to $N(f(X_i), \sigma^2)$ with p.d.f.

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - f(X_i))^2}{2\sigma^2}}$$

and the likelihood function of the sequence $Y_1, \ldots, Y_n$ is:

$$\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - f(X_i))^2} = \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2}.$$

Let us find the maximum likelihood estimates of β_0, β_1 and σ^2 that maximize this likelihood function. First of all, it is obvious that for any σ^2 we need to minimize

$$L := \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

over β_0, β_1. The line that minimizes the sum of squares L is called the *least-squares line*. To find the critical points we write:

$$\begin{aligned} \frac{\partial L}{\partial \beta_0} &= -\sum_{i=1}^n 2(Y_i - (\beta_0 + \beta_1 X_i)) = 0 \\ \frac{\partial L}{\partial \beta_1} &= -\sum_{i=1}^n 2(Y_i - (\beta_0 + \beta_1 X_i)) X_i = 0 \end{aligned}$$

If we introduce the notations

$$\bar{X} = \frac{1}{n}\sum X_i,\ \bar{Y} = \frac{1}{n}\sum Y_i,\ \bar{X^2} = \frac{1}{n}\sum X_i^2,\ \bar{XY} = \frac{1}{n}\sum X_i Y_i$$

then the critical point conditions can be rewritten as

$$\beta_0 + \beta_1 \bar{X} = \bar{Y} \text{ and } \beta_0 \bar{X} + \beta_1 \bar{X^2} = \bar{XY}.$$

Solving for β_0 and β_1 we get the MLE

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ and } \hat{\beta}_1 = \frac{\bar{XY} - \bar{X}\bar{Y}}{\bar{X^2} - \bar{X}^2}.$$

These estimates are used to plot least-squares regression lines in figure 14.1. Finally, to find the MLE of σ^2 we maximize the likelihood over σ^2 and get:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

The differences $r_i = Y_i - \hat{Y}_i$ between observed response variables Y_i and the values predicted by the estimated regression line

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

are called the *residuals.* The R^2 statistic in the examples above is defined as

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

The numerator in the last sum is the sum of squares of the residuals and the numerator is the variance of Y and R^2 is usually interpreted as the proportion of variability in the data explained by the linear model. The higher R^2 the better our model explains the data. Next, we would like to do statistical inference about the linear model.

1. Construct confidence intervals for parameters of the model β_0, β_1 and σ^2.

2. Construct prediction intervals for Y given any point X (dotted lines in figure 14.1).

3. Test hypotheses about parameters of the model. For example, F-statistic in the output of Matlab function 'regress' comes from a test of the hypothesis $H_0 : \beta_0 = 0, \beta_1 = 0$ that the response Y is not 'correlated' with a predictor variable X.

In spirit all these problems are similar to statistical inference about parameters of normal distribution such as t-tests, F-tests, etc. so as a starting point we need to find a joint distribution of the estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$.

To compute the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is very easy because they are linear combinations of Y_is which have normal distributions and, as a result, $\hat{\beta}_0$ and $\hat{\beta}_1$ will have normal distributions. All we need to do is find their means, variances and covariance, which is a straightforward computation. However, we will obtain this as a part of a more general computation that will also give us joint distribution of all three estimates $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}^2$. Let us denote the sample variance of Xs by

$$\sigma_x^2 = \bar{X^2} - \bar{X}^2.$$

Then we will prove the following:

1. $\hat{\beta}_1 \sim N\Big(\beta_1, \dfrac{\sigma^2}{n\sigma_x^2}\Big)$, $\hat{\beta}_0 \sim N\Big(\beta_0, \Big(\dfrac{1}{n} + \dfrac{\bar{X}^2}{n\sigma_x^2}\Big)\sigma^2\Big) = N\Big(\beta_0, \dfrac{\sigma^2}{n\sigma_x^2}\overline{X^2}\Big)$,
   $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\dfrac{\bar{X}\sigma^2}{n\sigma_x^2}$.
2. $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.
3. $\dfrac{n\hat{\sigma}^2}{\sigma^2}$ has χ^2_{n-2} distribution with $n-2$ degrees of freedom.

Remark. Line 1 means that $(\hat{\beta}_0, \hat{\beta}_1)$ have jointly normal distribution with mean (β_0, β_1) and covariance matrix

$$\Sigma = \frac{\sigma^2}{n\sigma_x^2}\begin{pmatrix} \overline{X^2} & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}.$$

Proof. Let us consider two vectors

$$a_1 = (a_{11}, \ldots, a_{1n}) = \Big(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\Big)$$

and

$$a_2 = (a_{21}, \ldots, a_{2n}) \text{ where } a_{2i} = \frac{X_i - \bar{X}}{\sqrt{n\sigma_x^2}}.$$

It is easy to check that both vectors have length 1 and they are orthogonal to each other since their scalar product is

$$a_1 \cdot a_2 = \sum_{i=1}^n a_{1i}a_{2i} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i - \bar{X}}{\sqrt{n\sigma_x^2}} = 0.$$

Let us choose vectors $a_3, \ldots, a_n$ so that $a_1, \ldots, a_n$ is orthonormal basis and, as a result, the matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ a_{12} & \cdots & a_{n2} \\ \vdots & \vdots & \vdots \\ a_{1n} & \cdots & a_{nn} \end{pmatrix}$$

is orthogonal. Let us consider vectors

$$Y = (Y_1, \ldots, Y_n), \mu = \mathbb{E}Y = (\mathbb{E}Y_1, \ldots, \mathbb{E}Y_n)$$

and

$$Y' = (Y_1', \ldots, Y_n') = \frac{Y - \mu}{\sigma} = \Big(\frac{Y_1 - \mathbb{E}Y_1}{\sigma}, \ldots, \frac{Y_n - \mathbb{E}Y_n}{\sigma}\Big)$$

so that the random variables $Y_1', \ldots, Y_n'$ are i.i.d. standard normal. We proved before that if we consider an orthogonal transformation of i.i.d. standard normal sequence:

$$Z' = (Z_1', \ldots, Z_n') = Y'A$$

then $Z'_1, \ldots, Z'_n$ will also be i.i.d. standard normal. Since

$$Z' = Y'A = \Big(\frac{Y-\mu}{\sigma}\Big)A = \frac{YA - \mu A}{\sigma}$$

this implies that

$$YA = \sigma Z' + \mu A.$$

Let us define a vector

$$Z = (Z_1, \ldots, Z_n) = YA = \sigma Z' + \mu A.$$

Each Z_i is a linear combination of Y_is and, therefore, it has a normal distribution. Since we made a specific choice of the first two columns of the matrix A we can write down explicitly the first two coordinates Z_1 and Z_2 of vector Z. We have,

$$Z_1 = \sum_{i=1}^{n} a_{i1} Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i = \sqrt{n}\bar{Y} = \sqrt{n}(\hat{\beta}_0 + \hat{\beta}_1 \bar{X})$$

and the second coordinate

$$\begin{aligned} Z_2 &= \sum_{i=1}^{n} a_{i2} Y_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})Y_i}{\sqrt{n\sigma_x^2}} \\ &= \sqrt{n\sigma_x^2} \sum_{i=1}^{n} \frac{(X_i - \bar{X})Y_i}{n\sigma_x^2} = \sqrt{n\sigma_x^2}\hat{\beta}_1. \end{aligned}$$

Solving these two equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ we can express them in terms of Z_1 and Z_2 as

$$\hat{\beta}_1 = \frac{1}{\sqrt{n\sigma_x^2}} Z_2 \text{ and } \hat{\beta}_0 = \frac{1}{\sqrt{n}} Z_1 - \frac{\bar{X}}{\sqrt{n\sigma_x^2}} Z_2.$$

This easily implies claim 1. Next we will show how $\hat{\sigma}^2$ can also be expressed in terms of Z_is.

$$\begin{aligned} n\hat{\sigma}^2 &= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} \Big((Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X}) \Big)^2 \ \{\text{since } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}\} \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - 2\hat{\beta}_1 n\sigma_x^2 \underbrace{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{n\sigma_x^2}}_{\hat{\beta}_1} + \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \hat{\beta}_1^2 n\sigma_x^2 = \sum_{i=1}^{n} Y_i^2 - \underbrace{n(\bar{Y})^2}_{Z_1^2} - \underbrace{\hat{\beta}_1^{\,2} n\sigma_x^2}_{Z_2^2} \\ &= \sum_{i=1}^{n} Y_i^2 - Z_1^2 - Z_2^2 = \sum_{i=1}^{n} Z_i^2 - Z_1^2 - Z_2^2 = Z_3^2 + \cdots + Z_n^2. \end{aligned}$$

In the last line we used the fact that $Z = YA$ is an orthogonal transformation of Y and since orthogonal transformation preserves the length of a vector we have,

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} Y_i^2.$$

If we can show that $Z_3, \dots, Z_n$ are i.i.d. with distribution $N(0, \sigma^2)$ then

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \Big(\frac{Z_3}{\sigma}\Big)^2 + \ldots + \Big(\frac{Z_n}{\sigma}\Big)^2 \sim \chi^2_{n-2}$$

has χ^2-distribution with $n-2$ degrees of freedom, because $Z_i/\sigma \sim N(0,1)$. Since we showed above that

$$Z = \mu A + \sigma Z' \Rightarrow Z_i = (\mu A)_i + \sigma Z'_i,$$

the fact that $Z'_1, \dots, Z'_n$ are i.i.d. standard normal implies that Z_is are independent of each other and $Z_i \sim N((\mu A)_i, \sigma^2)$. Let us compute the mean $\mathbb{E}Z_i = (\mu A)_i$:

$$\begin{aligned}
(\mu A)_i &= \mathbb{E}Z_i = \mathbb{E}\sum_{j=1}^n a_{ji}Y_j = \sum_{j=1}^n a_{ji}\mathbb{E}Y_j = \sum_{j=1}^n a_{ji}(\beta_0 + \beta_1 X_j) \\
&= \sum_{j=1}^n a_{ji}(\beta_0 + \beta_1\bar{X} + \beta_1(X_j - \bar{X})) \\
&= (\beta_0 + \beta_1\bar{X})\sum_{j=1}^n a_{ji} + \beta_1 \sum_{j=1}^n a_{ji}(X_j - \bar{X}).
\end{aligned}$$

Since the matrix A is orthogonal its columns are orthogonal to each other. Let $a_i = (a_{1i}, \dots, a_{ni})$ be the vector in the ith column and let us consider $i \geq 3$. Then the fact that a_i is orthogonal to the first column gives

$$a_i \cdot a_1 = \sum_{j=1}^n a_{j1}a_{ji} = \sum_{j=1}^n \frac{1}{\sqrt{n}} a_{ji} = 0$$

and the fact that a_i is orthogonal to the second column gives

$$a_i \cdot a_2 = \frac{1}{\sqrt{n\sigma_x^2}} \sum_{j=1}^n (X_j - \bar{X})a_{ji} = 0.$$

This show that for $i \geq 3$

$$\sum_{j=1}^n a_{ji} = 0 \text{ and } \sum_{j=1}^n a_{ji}(X_j - \bar{X}) = 0$$

and this proves that $\mathbb{E}Z_i = 0$ for $i \geq 3$ and $Z_i \sim N(0, \sigma^2)$ for $i \geq 3$. As we mentioned above this also proves that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$.

Finally, $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$ because $\hat{\sigma}^2$ can be written as a function of $Z_3, \dots, Z_n$ and $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as functions of Z_1 and Z_2.

□

Statistical inference in simple linear regression. Suppose now that we want to find the confidence intervals for unknown parameters of the model β_0, β_1 and σ^2. This is

straightforward and very similar to the confidence intervals for parameters of normal distribution. For example, using that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$, if we find the constants c_1 and c_2 such that

$$\chi^2_{n-2}(0, c_1) = \frac{1-\alpha}{2} \text{ and } \chi^2_{n-2}(c_2, +\infty) = \frac{1-\alpha}{2}$$

then with probability α we have $c_1 \leq n\hat{\sigma}^2/\sigma^2 \leq c_2$. Solving this for σ^2 we find the α confidence interval:

$$\frac{n\hat{\sigma}^2}{c_2} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{c_1}.$$

Similarly, we find the α confidence interval for β_1. Since

$$(\hat{\beta}_1 - \beta_1)\Big/\sqrt{\frac{\sigma^2}{n\sigma_x^2}} \sim N(0,1) \text{ and } \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

then

$$(\hat{\beta}_1 - \beta_1)\sqrt{\frac{n\sigma_x^2}{\sigma^2}}\Big/\sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} \sim t_{n-2}$$

has Student t_{n-2}-distribution with $n-2$ degrees of freedom. Simplifying, we get

$$(\hat{\beta}_1 - \beta_1)\sqrt{\frac{(n-2)\sigma_x^2}{\hat{\sigma}^2}} \sim t_{n-2}. \tag{14.0.1}$$

Therefore, if we find c such that $t_{n-2}(-c, c) = \alpha$ then with probability α:

$$-c \leq (\hat{\beta}_1 - \beta_1)\sqrt{\frac{(n-2)\sigma_x^2}{\hat{\sigma}^2}} \leq c$$

and solving for β_1 we obtain the α confidence interval:

$$\hat{\beta}_1 - c\sqrt{\frac{\hat{\sigma}^2}{(n-2)\sigma_x^2}} \leq \beta_1 \leq \hat{\beta}_1 + c\sqrt{\frac{\hat{\sigma}^2}{(n-2)\sigma_x^2}}.$$

Similarly, to find the confidence interval for β_0 we use that

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{n\sigma_x^2}\right)\sigma^2}}\Big/\sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} = (\hat{\beta}_0 - \beta_0)\Big/\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1 + \frac{\bar{X}^2}{\sigma_x^2}\right)} \sim t_{n-2} \tag{14.0.2}$$

and α confidence interval for β_0 is:

$$\hat{\beta}_0 - c\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1 + \frac{\bar{X}^2}{\sigma_x^2}\right)} \leq \beta_0 \leq \hat{\beta}_0 + c\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1 + \frac{\bar{X}^2}{\sigma_x^2}\right)}.$$

□

We can now construct various t-tests based on t-statistics (14.0.1) and (14.0.2).

Linear combinations of parameters. More generally, let us compute the distribution of a linear combination

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1$$

of the estimates. This will allow us to construct confidence intervals and t-tests for linear combinations of parameters $c_0\beta_0 + c_1\beta_1$. Clear, the distribution of this linear combination will be normal with mean

$$\mathbb{E}(c_0\hat{\beta}_0 + c_1\hat{\beta}_1) = c_0\beta_0 + c_1\beta_1.$$

We compute its variance:

$$\begin{aligned}
\mathrm{Var}(c_0\hat{\beta}_0 + c_1\hat{\beta}_1) &= \mathbb{E}(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c_0\beta_0 - c_1\beta_1)^2 = \mathbb{E}(c_0(\hat{\beta}_0 - \beta_0) + c_1(\hat{\beta}_1 - \beta_1))^2 \\
&= c_0^2 \underbrace{\mathbb{E}(\hat{\beta}_0 - \beta_0)^2}_{\text{variance of } \hat{\beta}_1} + c_1^2 \underbrace{\mathbb{E}(\hat{\beta}_1 - \beta_1)^2}_{\text{variance of } \hat{\beta}_0} + 2c_0c_1 \underbrace{\mathbb{E}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)}_{\text{covariance}} \\
&= c_0^2\Big(\frac{1}{n} + \frac{\bar{X}^2}{n\sigma_x^2}\Big)\sigma^2 + c_1^2\frac{\sigma^2}{n\sigma_x^2} - 2c_0c_1\frac{\bar{X}\sigma^2}{n\sigma_x^2} \\
&= \sigma^2\Big(\frac{c_0^2}{n} + \frac{(c_0\bar{X} - c_1)^2}{n\sigma_x^2}\Big).
\end{aligned}$$

This proves that

$$c_0\hat{\beta}_0 + c_1\hat{\beta}_1 \sim N\Big(c_0\beta_0 + c_1\beta_1, \sigma^2\Big(\frac{c_0^2}{n} + \frac{(c_0\bar{X} - c_1)^2}{n\sigma_x^2}\Big)\Big). \tag{14.0.3}$$

Using $(c_0, c_1) = (1, 0)$ or $(0, 1)$, will give the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$.

□

Prediction Intervals. Suppose now that we have a new observation X for which Y is unknown and we want to predict Y or find the confidence interval for Y. According to simple regression model,

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

and it is natural to take $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ as the prediction of Y. Let us find the distribution of their difference $\hat{Y} - Y$. Clearly, the difference will have normal distribution so we only need to compute the mean and the variance. The mean is

$$\mathbb{E}(\hat{Y} - Y) = \mathbb{E}\hat{\beta}_0 + \mathbb{E}\hat{\beta}_1 X - \beta_0 - \beta_1 X - \mathbb{E}\varepsilon = \beta_0 + \beta_1 X - \beta_0 - \beta_1 X - 0 = 0.$$

Since a new pair (X, Y) is independent of the prior data we have that Y is independent of $\hat{Y}$. Therefore, since the variance of the sum or difference of independent random variables is equal to the sum of their variances, we get

$$\mathrm{Var}(\hat{Y} - Y) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y) = \sigma^2 + \mathrm{Var}(\hat{Y}),$$

where we also used that $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon) = \sigma^2$. To compute the variance of $\hat{Y}$ we can use the formula above with $(c_0, c_1) = (1, X)$

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}(\hat{\beta}_0 + X\hat{\beta}_1) = \sigma^2\Big(\frac{1}{n} + \frac{(\bar{X} - X)^2}{n\sigma_x^2}\Big).$$

Therefore, we showed that

$$\hat{Y} - Y \sim N\Big(0, \sigma^2\Big(1 + \frac{1}{n} + \frac{(\bar{X} - X)^2}{n\sigma_x^2}\Big)\Big).$$

As a result, we have:

$$\frac{\hat{Y} - Y}{\sqrt{\sigma^2\Big(1 + \frac{1}{n} + \frac{(\bar{X}-X)^2}{n\sigma_x^2}\Big)}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} \sim t_{n-2}$$

and the $1 - \alpha$ prediction interval for Y is

$$\hat{Y} - c\sqrt{\frac{\sigma^2}{n-2}\Big(n + 1 + \frac{(\bar{X} - X)^2}{\sigma_x^2}\Big)} \le Y \le \hat{Y} + c\sqrt{\frac{\sigma^2}{n-2}\Big(n + 1 + \frac{(\bar{X} - X)^2}{\sigma_x^2}\Big)}.$$

These are the dashed curves created by Matlab 'polytool' function.

□

Simultaneous confidence set for (β_0, β_1) and F-test. We will now construct a statistic that will allow us to give a confidence set for both parameters β_0, β_1 at the same time and test the hypothesis of the type

$$H_0 : \beta_0 = 0 \text{ and } \beta_1 = 0. \tag{14.0.4}$$

The values $(0, 0)$ could be replaced by any other predetermined values. Looking at the proof of the joint distribution of the estimates, as an intermediate step we showed that estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can be related to

$$Z_1 = \sqrt{n}(\hat{\beta}_0 + \hat{\beta}_1 X) \quad \text{and} \quad Z_2 = \sqrt{n\sigma_x^2}\hat{\beta}_1$$

where normal random variables Z_1, Z_2 are independent of each other and independent of

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Also, Z_1 and Z_2 have variance σ^2. Standardizing these random variables we get

$$A = \frac{\sqrt{n}}{\sigma}((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)\bar{X}) \sim N(0, 1) \quad \text{and} \quad B = \frac{\sqrt{n\sigma_x^2}}{\sigma}(\hat{\beta}_1 - \beta_1) \sim N(0, 1)$$

which implies that $A^2 + B^2 \sim \chi^2_2$-distribution. By definition of F-distribution,

$$\frac{n-2}{2}(A^2 + B^2) \Big/ \frac{n\hat{\sigma}^2}{\sigma^2} \sim F_{2,n-2}.$$

Simplifying the left-hand side we get

$$F := \frac{n-2}{2\hat{\sigma}^2}\Big((\hat{\beta}_0 - \beta_0)^2 + \bar{X}^2(\hat{\beta}_1 - \beta_1)^2 + 2\bar{X}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\Big) \sim F_{2,n-2}.$$

This allows us to obtain a joint confidence set (ellipse) for parameters β_0, β_1. Given a confidence level $\alpha \in [0,1]$ is we define a threshold c by $F_{2,n-2}(0,c) = \alpha$ then with probability α we have

$$F := \frac{n-2}{2\hat{\sigma}^2}\Big((\hat{\beta}_0 - \beta_0)^2 + \bar{X}^2(\hat{\beta}_1 - \beta_1)^2 + 2\bar{X}(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\Big) \leq c.$$

This inequality defines an ellipse for (β_0, β_1). To test the hypothesis (14.0.4), we use the fact that under H_0 the statistic

$$F := \frac{n-2}{2\hat{\sigma}^2}(\hat{\beta}_0^2 + \bar{X}^2\hat{\beta}_1^2 + 2\bar{X}\hat{\beta}_0\hat{\beta}_1) \sim F_{2,n-2}$$

and define a decision rule by

$$\delta = \begin{cases} H_0: & F \leq c \\ H_1: & F > c, \end{cases}$$

where $F_{2,n-2}(c, \infty) = \alpha$ - a level of significance.

F-statistic output by Matlab 'regress' function will be explained in the next section.

□

References.

[1] "Using Cigarette Data for An Introduction to Multiple Regression." by Lauren McIntyre, *Journal of Statistics Education* v.2, n.1 (1994).

[2] Mendenhall, W., and Sincich, T. (1992), Statistics for Engineering and the Sciences (3rd ed.), New York: Dellen Publishing Co.

Section 15

Multiple linear regression.

Let us consider a model

$$Y_i = \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

where random noise variables $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$. We can write this in a matrix form

$$Y = X\beta + \varepsilon,$$

where Y and ε are $n \times 1$ vectors, β is $p \times 1$ vector and X is $n \times p$ matrix. We will denote the columns of matrix X by $X_1, \ldots X_p$, i.e.

$$X = (X_1, \ldots, X_p)$$

and we will assume that these columns are linearly independent. If they are not linearly independent, we can not reconstruct parameters β from X and Y even if there is no noise ε. In simple linear regression this would correspond to all Xs being equal and we can not estimate a line from observations only at one point. So from now on we will assume that $n > p$ and the rank of matrix X is equal to p. To estimate unknown parameters β and σ we will use maximum likelihood estimators.

Lemma 1. *The MLE of β and σ^2 are given by:*

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{n}|Y - X\hat{\beta}|^2 = \frac{1}{n}|Y - X(X^T X)^{-1} X^T Y|^2.$$

Proof. The p.d.f. of Y_i is

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{2\sigma^2}(x - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2\Big)$$

and, therefore, the likelihood function is

$$\begin{aligned} \prod_{i=1}^n f_i(Y_i) &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2\Big) \\ &= \Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n \exp\Big(-\frac{1}{2\sigma^2}|Y - X\beta|^2\Big). \end{aligned}$$

To maximize the likelihood function, first, we need to minimize $|Y - X\beta|^2$. If we rewrite the norm squared using scalar product:

$$\begin{aligned} |Y - X\beta|^2 &= (Y - \sum_{i=1}^{p} \beta_i X_i, Y - \sum_{i=1}^{p} \beta_i X_i) \\ &= (Y,Y) - 2\sum_{i=1}^{p} \beta_i (Y, X_i) + \sum_{i,j=1}^{p} \beta_i \beta_j (X_i, X_j). \end{aligned}$$

Then setting the derivatives in each β_i equal to zero

$$-2(Y, X_i) + 2\sum_{j=1}^{p} \beta_j (X_i, X_j) = 0$$

we get

$$(Y, X_i) = \sum_{i=1}^{p} \beta_j (X_i, X_j) \quad \text{for all} \quad i \le p.$$

In matrix notations this can be written as $X^T Y = X^T X\beta$. Matrix $X^T X$ is a $p \times p$ matrix. Is is invertible since by assumption X has rank p. So we can solve for β to get the MLE

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

It is now easy to minimize over σ to get

$$\hat{\sigma}^2 = \frac{1}{n}|Y - X\hat{\beta}|^2 = \frac{1}{n}|Y - X(X^T X)^{-1} X^T Y|^2.$$

□

To do statistical inference we need to compute the joint distribution of these estimates. We will prove the following.

Theorem. *We have*

$$\hat{\beta} \sim N\Big(\beta, \sigma^2 (X^T X)^{-1}\Big), \;\; \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$$

and estimates $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

Proof. First of all, let us rewrite the estimates in terms of random noise ε using $Y = X\beta + \varepsilon$. We have

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= (X^T X)^{-1}(X^T X)\beta + (X^T X)^{-1} X^T \varepsilon = \beta + (X^T X)^{-1} X^T \varepsilon \end{aligned}$$

and since

$$\begin{aligned} Y - X(X^T X)^{-1} X^T Y &= X\beta + \varepsilon - X(X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= X\beta + \varepsilon - X\beta - X(X^T X)^{-1} X^T \varepsilon = (I - X(X^T X)^{-1} X^T)\varepsilon \end{aligned}$$

we have

$$\hat{\sigma}^2 = \frac{1}{n}|(I - X(X^TX)^{-1}X^T)\varepsilon|^2.$$

Since $\hat{\beta}$ is a linear transformation of a normal vector ε it will also be normal with mean

$$\mathbb{E}\hat{\beta} = \mathbb{E}(\beta + (X^TX)^{-1}X^T\varepsilon) = \beta$$

and covariance matrix

$$\begin{aligned}
\mathbb{E}(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T &= \mathbb{E}(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\mathbb{E}\varepsilon\varepsilon^TX(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.
\end{aligned}$$

This proves that $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$. To prove that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent and to find the distribution of $n\hat{\sigma}^2/\sigma^2$ we will use the following trick. This trick can also be very useful computationally since it will relate all quantities of interest expressed in terms of $n \times p$ matrix X to quantities expressed in terms of a certain $p \times p$ matrix R which can be helpful when n is very large compared to p. We would like to manipulate the columns of matrix X to make them orthogonal to each other, which can be done by Gram-Schmidt orthogonalization. In other words, we want to represent matrix X as

$$X = X_0R$$

where X_0 is $n \times p$ matrix with columns $X_0^1, \ldots, X_0^p$ that are orthogonal to each other and, moreover, form an orthonormal basis, and matrix R is $p \times p$ invertible (and upper triangular) matrix. In Matlab this can be done using economy size QR factorization

```
[X0, R]=qr(X,0).
```

The fact that columns of X_0 are orthonormal implies that

$$X_0^TX_0 = I$$

- a $p \times p$ identity matrix. Let us replace X by X_0R everywhere in the estimates. We have

$$(X^TX)^{-1}X^T = (R^TX_0^TX_0R)^{-1}R^TX_0^T = (R^TR)^{-1}R^TX_0^T = R^{-1}(R^T)^{-1}R^TR^{-1}X_0^T,$$

$$X(X^TX)^{-1}X^T = X_0R(R^TX_0^2X_0R)^{-1}R^TX_0^T = X_0RR^{-1}(R^T)^{-1}R^TX_0^T = X_0X_0^T.$$

As a result

$$\hat{\beta} - \beta = R^{-1}X_0^T\varepsilon \quad \text{and} \quad n\hat{\sigma}^2 = |(I - X_0X_0^T)\varepsilon|^2. \tag{15.0.1}$$

By construction p columns of X_0, which are also the rows of X_0^T, are orthonormal. Therefore, we can choose the last $n - p$ rows of a $n \times n$ matrix

$$A = \begin{pmatrix} X_0^T \\ \cdots \end{pmatrix}$$

to make A an orthogonal matrix, we just need to choose them to complete, together with rows of X_0^T, the orthonormal basis in $\mathbb{R}^n$. Let us define a vector

$$g = A\varepsilon, \text{ i.e. } \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix} = \begin{pmatrix} X_0^T \\ \cdots \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Since ε is a vector of i.i.d. standard normal, we proved before that its orthogonal transformation g will also be a vector of independent $N(0,\sigma^2)$ random variables $g_1,\dots,g_n$. First of all, since

$$\hat{g} := \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix} = X_0^T \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

we have

$$\hat{\beta} - \beta = R^{-1} X_0^T \varepsilon = R^{-1} \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix} = R^{-1}\hat{g}. \tag{15.0.2}$$

Next, we will prove that

$$|(I - X_0 X_0^T)\varepsilon|^2 = g_{p+1}^2 + \ldots + g_n^2. \tag{15.0.3}$$

First of all, orthogonal transformation preserves lengths, so $|g|^2 = |A\varepsilon|^2 = |\varepsilon|^2$. On the other hand, let us write $|\varepsilon|^2 = \varepsilon^T\varepsilon$ and break ε into a sum of two terms

$$\varepsilon = X_0 X_0^T \varepsilon + (I - X_0 X_0^T)\varepsilon.$$

Then we get

$$|g|^2 = |\varepsilon|^2 = \varepsilon^T\varepsilon = \Big(\varepsilon^T X_0 X_0^T + \varepsilon^T(I - X_0 X_0^T)\Big)\Big(X_0 X_0^T\varepsilon + (I - X_0 X_0^T)\varepsilon\Big).$$

When we multiply all the terms out we will use that $X_0^T X_0 = I$ since the matrix $X_0^T X_0$ consists of scalar products of columns of X_0 which are orthonormal. This also implies that

$$X_0 X_0^T (I - X_0 X_0^T) = X_0 X_0^T - X_0 I X_0^T = 0.$$

Using this we get

$$\begin{aligned} |g|^2 = |\varepsilon|^2 &= \varepsilon^T X_0 X_0^T \varepsilon + \varepsilon^T (I - X_0 X_0^T)(I - X_0 X_0^T)\varepsilon \\ &= |X_0^T\varepsilon|^2 + |(I - X_0 X_0^T)\varepsilon|^2 = |\hat{g}|^2 + |(I - X_0 X_0^T)\varepsilon|^2 \end{aligned}$$

because $\hat{g} = X_0^T\varepsilon$ so we finally proved that

$$|(I - X_0 X_0^T)\varepsilon|^2 = |g|^2 - |\hat{g}|^2 = g_1^2 + \ldots + g_n^2 - g_1^2 - \ldots - g_p^2 = g_{p+1}^2 + \ldots + g_n^2$$

which is (15.0.3). This proves that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ and it is also independent of $\hat{\beta}$ which depends only on $g_1,\dots,g_p$ by (15.0.2).

□

Let us for convenience write down equation (15.0.2) as a separate result.

Lemma 2. *Given a decomposition $X = X_0 R$ with $n \times p$ matrix X_0 with orthonormal columns and invertible (upper triangular) $p \times p$ matrix R we can represent*

$$\hat{\beta} - \beta = R^{-1}\hat{g} = R^{-1} \begin{pmatrix} g_1 \\ \vdots \\ g_p \end{pmatrix}$$

for independent $N(0, \sigma^2)$ random variables $g_1, \ldots, g_p$.

Confidence intervals and t-tests for linear combination of parameters β. Let us consider a linear combination

$$c_1\beta_1 + \ldots + c_p\beta_p = c^T\beta$$

where $c = (c_1, \ldots, c_p)^T$. To construct confidence intervals and t-tests for this linear combination we need to write down a distribution of $c^T\hat{\beta}$. Clearly, it has a normal distribution with mean $\mathbb{E}c^T\hat{\beta} = c^T\beta$ and variance

$$\mathbb{E}(c^T(\hat{\beta} - \beta))^2 = \mathbb{E}c^T(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T c = c^T \mathrm{Cov}(\hat{\beta})c = \sigma^2 c^T (X^T X)^{-1} c.$$

Therefore,

$$\frac{c^T(\hat{\beta} - \beta)}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}} \sim N(0, 1)$$

and using that $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ we get

$$\frac{c^T(\hat{\beta} - \beta)}{\sqrt{\sigma^2 c^T (X^T X)^{-1} c}} \Big/ \sqrt{\frac{1}{n-p}\frac{n\hat{\sigma}^2}{\sigma^2}} = c^T(\hat{\beta} - \beta)\sqrt{\frac{n-p}{n\hat{\sigma}^2 c^T (X^T X)^{-1} c}} \sim t_{n-p}.$$

To obtain the distribution of one parameter $\hat{\beta}_i$ we need to choose a vector c that has all zeros and 1 in the ith coordinate. Then we get

$$(\hat{\beta}_i - \beta_i)\sqrt{\frac{n-p}{n\hat{\sigma}^2((X^T X)^{-1})_{ii}}} \sim t_{n-p}.$$

Here $((X^T X)^{-1})_{ii}$ is the ith diagonal element of the matrix $(X^T X)^{-1}$. This is a good time to mention how the quality of estimation of β depends on the choice of X. For example, we mentioned before that the columns of X should be linearly independent. What happens if some of them are nearly collinear? Then some eigenvalues of $(X^T X)$ will be 'small' (in some sense) and some eigenvalues of $(X^T X)^{-1}$ will be 'large'. (Small and large here are relative terms because the size of the matrix also grows with n.) As a result, the confidence intervals for some parameters will get very large too which means that their estimates are not very accurate. To improve the quality of estimation we need to avoid using collinear predictors. We will see this in the example below.

□

Joint confidence set for β and F-test. By Lemma 2, $R(\hat{\beta}-\beta)=\hat{g}$ and, therefore,

$$g_1^2+\ldots+g_p^2=|\hat{g}|^2=\hat{g}^T\hat{g}=(\hat{\beta}-\beta)^TR^TR(\hat{\beta}-\beta)=(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta).$$

Since $g_i\sim N(0,\sigma^2)$ this proves that

$$\frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)}{\sigma^2}\sim\chi^2_p.$$

Using that $n\hat{\sigma}^2/\sigma^2\sim\chi^2_{n-p}$ gives

$$\frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)}{p\sigma^2}\Big/\frac{n\hat{\sigma}^2}{(n-p)\sigma^2}=\frac{(n-p)}{np\hat{\sigma}^2}(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)\sim F_{p,n-p}.$$

If we take c such that $F_{p,n-p}(0,c_\alpha)=\alpha$ then

$$\frac{(n-p)}{np\hat{\sigma}^2}(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)\le c_\alpha \tag{15.0.4}$$

defines a joint confidence set for all parameters β simultaneously with confidence level α.

Suppose that we want to test a hypothesis about all parameters simultaneously, for example,

$$H_0:\beta=\beta_0.$$

Then we consider a statistic

$$F=\frac{(n-p)}{np\hat{\sigma}^2}(\hat{\beta}-\beta_0)^TX^TX(\hat{\beta}-\beta_0), \tag{15.0.5}$$

which under null hypothesis has $F_{p,n-p}$ distribution, and define a decision rule by

$$\delta=\begin{cases} H_0: & F\le c \\ H_1: & F>c, \end{cases}$$

where a threshold c is determined by $F_{p,n-p}(c,\infty)=\alpha$ - a level of significance. Of course, this test is equivalent to checking if vector β_0 belongs to a confidence set (15.0.4)! (We just need to remember that confidence level = 1 - level of significance.)

□

Simultaneous confidence set and F-test for subsets of β. Let

$$s=\{i_1,\ldots,i_k\}\subseteq\{1,\ldots,p\}$$

be s subset of size $k\le p$ of indices $\{1,\ldots,p\}$ and let $\beta_s=(\beta_{i_1},\ldots,\beta_{i_k})^T$ be a vector that consists of the corresponding subset of parameters β. Suppose that we would like to test the hypothesis

$$H_0:\beta_s=\beta_s^0$$

for some given vector β_s^0, for example, $\beta_s^0=0$. Let $\hat{\beta}_s$ be a corresponding vector of estimates. Let

$$\Sigma_s=\Big((X^TX)^{-1}_{i,j}\Big)_{i,j\in s}$$

be a $k \times k$ submatrix of $(X^T X)^{-1}$ with row and column indices in the set s. By the above Theorem, the joint distribution of $\hat{\beta}_s$ is

$$\hat{\beta}_s \sim N(\beta_s, \sigma^2 \Sigma_s).$$

Let $A = \Sigma_s^{1/2}$, i.e. A is a symmetric $k \times k$ matrix such that $\Sigma_s = AA^T$. As a result, a centered vector of estimates can be represented as

$$\hat{\beta}_s - \beta_s = Ag,$$

where $g = (g_1, \ldots, g_k)^T$ are independent $N(0, \sigma^2)$. Therefore, $g = A^{-1}(\hat{\beta}_s - \beta_s)$ and the rest is similar to the above argument. Namely,

$$\begin{aligned} g_1^2 + \ldots + g_k^2 &= |g|^2 = g^T g = (\hat{\beta}_s - \beta_s)^T (A^{-1})^T A^{-1} (\hat{\beta}_s - \beta_s) \\ &= (\hat{\beta}_s - \beta_s)^T (AA^T)^{-1} (\hat{\beta}_s - \beta_s) = (\hat{\beta}_s - \beta_s)^T \Sigma_s^{-1} (\hat{\beta}_s - \beta_s) \sim \sigma^2 \chi_k^2. \end{aligned}$$

As before we get

$$F = \frac{(n-p)}{nk\hat{\sigma}^2} (\hat{\beta}_s - \beta_s)^T \Sigma_s^{-1} (\hat{\beta}_s - \beta_s) \sim F_{k,n-p}$$

and we can now construct a simultaneous confidence set and F-tests.

□

Remark. Matlab regression function 'regress' assumes that a matrix X of explanatory variables will contain a first column of ones that corresponds to an "intercept" parameter β_1. The F-statistic output by 'regress' corresponds to F-test about all other "slope" parameters:

$$H_0 : \beta_2 = \ldots = \beta_p = 0.$$

In this case $s = \{2, 3, \ldots, p\}$, $k = p - 1$ and

$$F = \frac{(n-p)}{n(p-1)\hat{\sigma}^2} \hat{\beta}_s^T \Sigma_s^{-1} \hat{\beta}_s \sim F_{p-1,n-p}.$$

□

Example. Let us take a look at the 'cigarette' dataset from previous lecture. We saw that tar, nicotine and carbon monoxide content are positively correlated and any pair is well described by a simple linear regression. Suppose that we would like to predict carbon monoxide as a linear function of both tar and nicotine content. We create a 25×3 matrix X:

```
X=[ones(25,1),tar,nic];
```

We introduce a first column of ones to allow an intercept parameter β_1 in our multiple linear regression model:

$$\mathrm{CO}_i = \beta_1 + \beta_2 \mathrm{Tar}_i + \beta_3 \mathrm{Nicotin}_i + \varepsilon_i.$$

If we perform a multiple linear regression:

```
[b,bint,r,rint,stats] = regress(carb,X);
```

We get the estimates of parameters and 95% confidence intervals for each parameter

```
b = 3.0896      bint = 1.3397   4.8395
    0.9625               0.4717   1.4533
   -2.6463             -10.5004   5.2079
```

and, in order, R^2-statistic, F-statistic from (15.0.5), p-value for this statistic

$$F_{p,n-p}(F,+\infty) = F_{3,25-3}(F,+\infty)$$

and the estimate of variance $\hat{\sigma}^2$:

```
stats = 0.9186  124.1102  0.000  1.9952.
```

First of all, we see that high R^2 means that linear model explain most of the variability in the data and small p-value means that we reject the hypothesis that all parameters are equal to zero. On the other hand, simple linear regression showed that carbon monoxide had a positive correlation with nicotine and now we got $\hat{\beta}_3 = -2.6463$. Also, notice that the confidence interval for β_3 is very poor. The reason for this is that tar and nicotine are nearly collinear. Because of this the matrix

$$(X^T X)^{-1} = \begin{pmatrix} 0.3568 & 0.0416 & -0.9408 \\ 0.0416 & 0.0281 & -0.4387 \\ -0.9408 & -0.4387 & 7.1886 \end{pmatrix}$$

has relatively large last diagonal diagonal value. We recall that Theorem gives that the variance of estimate $\hat{\beta}_3$ is $7.1886\sigma^2$ and we also see that the estimate of σ^2 is $\hat{\sigma}^2 = 1.9952$. As a result the confidence interval for β_3 is rather poor.

Of course, looking at linear combinations of tar and nicotine as new predictors does not make sense because they lose their meaning, but for the sake of illustrations let us see what would happen if our predictors were not nearly collinear but, in fact, orthonormal. Let us use economic QR decomposition

```
[X0,R]=qr(X,0)
```

a new matrix of predictor X_0 with orthonormal columns that are some linear combinations of tar and nicotine. Then regressing carbon monoxide on these new predictors

```
 [b,bint,r,rint,stats] = regress(carb,X0);
```

we would get

```
b = -62.6400     bint = -65.5694  -59.7106
    -22.2324             -25.1618  -19.3030
     0.9870              -1.9424    3.9164
```

all confidence intervals of the same relatively better size.

□

Example. The following data presents per capita income of 20 countries for 1960s. Also presented are the percentages of labor force employed in agriculture, industry and service for each country. (Data source: lib.stat.cmu.edu/DASL/Datafiles/oecdat.html)

```
COUNTRY     PCINC AGR IND SER

CANADA       1536  13  43  45
SWEEDEN      1644  14  53  33
SWITZERLAND  1361  11  56  33
LUXEMBOURG   1242  15  51  34
U. KINGDOM   1105  4   56  40
DENMARK      1049  18  45  37
W. GERMANY   1035  15  60  25
FRANCE       1013  20  44  36
BELGUIM      1005  6   52  42
NORWAY       977   20  49  32
ICELAND      839   25  47  29
NETHERLANDS  810   11  49  40
AUSTRIA      681   23  47  30
IRELAND      529   36  30  34
ITALY        504   27  46  28
JAPAN        344   33  35  32
GREECE       324   56  24  20
SPAIN        290   42  37  21
PORTUGAL     238   44  33  23
TURKEY       177   79  12  9
```

We can perform simple linear regression of income on each of the other explanatory variables or multiple linear regression on any pair of the explanatory variables. Fitting simple linear regression of income vs. percent of labor force in agriculture, industry and service:

```
polytool(agr,income,1),
```

etc., produces figure 15.1. Next, we perform statistical inference using 'regress' function. Statistical analysis of linear regression fit of income vs. percent of labor force in agriculture:

```
[b,bint,r,rint,stats]=regress(income,[ones(20,1),agr])

b = 1317.9    bint = 1094.7   1541.1
     -18.9             -26.0   -11.7

stats =   0.6315  30.8472  2.8e-005  74596.
```

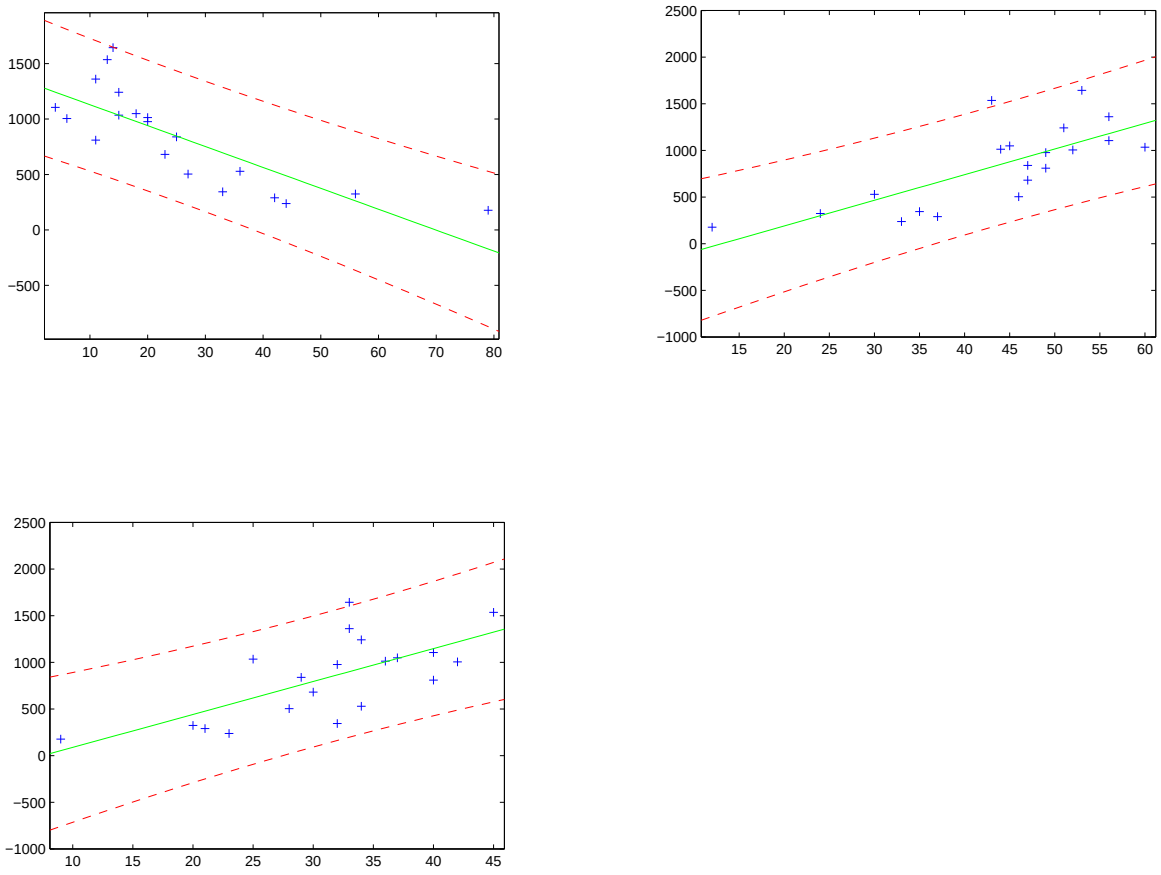


Figure 15.1: Linear regression of income on percent of labor force in agriculture, industry and service.

For income vs. percent of labor force in industry

```
[b,bint,r,rint,stats]=regress(income,[ones(20,1),ind]);

b = -359.3115    bint = -907.1807   188.5577
    27.4905                15.3058   39.6751

stats = 0.5552   22.4677   0.0002   90042
```

and for income vs. labor force in service

```
[b,bint,r,rint,stats]=regress(income,[ones(20,1),serv]);

b = -264.5199     bint =  -858.0257  328.9858
      35.3024                16.8955   53.7093
```

```
stats =   0.4742  16.2355   0.0008   106430.
```

We see that in all three cases, the hypotheses that parameters of least-squares line are both zero can be rejected at conventional level of significance $\alpha = 0.05$. Looking at the confidence intervals for the estimates of slopes we observe that the correlation of income with percent of labor force in agriculture is negative, and other two correlations are positive.

We can also perform a multiple regression on any two explanatory variables. We can not perform multiple linear regression with all three explanatory variables because they add up to 100%, i.e. they are linearly dependent. If we create a predictor matrix

```
X=[ones(20,1),agr,ind];
```

and perform multiple linear regression

```
[b,bint,r,rint,stats]=regress(income,X);
```

we get

```
b = 1272.1    bint = -632.6   3176.9
     -18.4            -39.1      2.3
       0.8            -31.4     32.9

stats =  0.6316  14.5703  0.0002  78972
```

Of course, one can find many shortcomings of this model. For example, having the entire population in agriculture results in prediction of $1272.1 - 1840 < 0$ negative income per capita.

□

Section 16

Linear constraints in multiple linear regression. Analysis of variance.

Multiple linear regression with general linear constraints. Let us consider a multiple linear regression $Y = X\beta + \varepsilon$ and suppose that we want to test a hypothesis given by a set of s linear equations. In a matrix form:

$$H_0 : A\beta = c,$$

where A is a $s \times p$ matrix and c is a $s \times 1$ vector. We will assume that $s \leq p$ and the matrix A has rank s. This generalizes two types of hypotheses from previous lecture, when we considered only one linear combination of parameters ($s = 1$ case) or tested hypothesis about all parameters simultaneously ($s = p$ case).

To test this general hypothesis, a natural idea is to compare how far $A\hat{\beta}$ is from c and to do this we need to find the distribution of $A\hat{\beta}$. Clearly, this distribution is normal with mean $A\beta$ and covariance

$$\mathbb{E}A(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T A^T = A\mathrm{Cov}(\hat{\beta})A^T = \sigma^2 A(X^T X)^{-1}A^T = \sigma^2 D$$

where we introduced a notation

$$D := A(X^T X)^{-1}A^T.$$

A matrix D is a symmetric positive definite invertible $s \times s$ matrix and, therefore, we can take its square root $D^{1/2}$. It is easy to check that the covariance of $D^{-1/2}A(\hat{\beta} - \beta)$ is $\sigma^2 I$. This implies that

$$\frac{1}{\sigma^2}|D^{-1/2}A(\hat{\beta} - \beta)|^2 = \frac{1}{\sigma^2}(A(\hat{\beta} - \beta))^T D^{-1}A(\hat{\beta} - \beta) \sim \chi^2_s.$$

Under null hypothesis, $A\beta = c$, we get

$$\frac{1}{\sigma^2}(A\hat{\beta} - c)^T D^{-1}(A\hat{\beta} - c) \sim \chi^2_s. \tag{16.0.1}$$

Since $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$ is independent of $\hat{\beta}$, we get

$$\frac{1}{s\sigma^2}(A\hat{\beta}-c)^T D^{-1}(A\hat{\beta}-c)\Big/\frac{n\hat{\sigma}^2}{(n-p)\sigma^2}$$
$$= \frac{n-p}{ns\hat{\sigma}^2}(A\hat{\beta}-c)^T D^{-1}(A\hat{\beta}-c) \sim F_{s,n-p}. \tag{16.0.2}$$

This is enough to test hypothesis H_0. However, in a variety of applications a different equivalent representation of (16.0.1) is more useful. It is given in terms of MLE $\hat{\beta}_A$ of β that satisfies the constraint in H_0. In other words, $\hat{\beta}_A$ is obtained by solving:

$$|Y - X\beta|^2 \to \min_{\beta} \quad \text{subject to the constraint} \quad A\beta = c. \tag{16.0.3}$$

Lemma. *If $\hat{\beta}_A$ is solution of (16.0.3) then the left hand side of (16.0.1) is equal to*

$$\frac{1}{\sigma^2}|X(\hat{\beta}_A - \hat{\beta})|^2. \tag{16.0.4}$$

Proof. First, let us find the constrained MLE $\hat{\beta}_A$ explicitly. By method of Lagrange multipliers we need to solve a system of two equations:

$$A\beta = c, \;\; \frac{\partial}{\partial\beta}\Big(|Y - X\beta|^2 + (A\beta - c)^T\lambda\Big) = 0,$$

where λ is a $s \times 1$ vector. The second equation is

$$-2X^TY + 2X^TX\beta + A^T\lambda = 0.$$

Solving this for β gives

$$\hat{\beta}_A = (X^TX)^{-1}X^TY - \frac{1}{2}(X^TX)^{-1}A^T\lambda = \hat{\beta} - \frac{1}{2}(X^TX)^{-1}A^T\lambda.$$

Since $\hat{\beta}_A$ must satisfy the linear constraint, we get

$$c = A\hat{\beta}_A = A\hat{\beta} - \frac{1}{2}A(X^TX)^{-1}A^T\lambda = A\hat{\beta} - \frac{1}{2}D\lambda.$$

Solving this for λ, $\lambda = 2D^{-1}(A\hat{\beta} - c)$, we get

$$\hat{\beta}_A = \hat{\beta} - (X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c).$$

and, therefore,

$$X(\hat{\beta}_A - \hat{\beta}) = -X(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c).$$

We can use this formula to compute

$$|X(\hat{\beta}_A - \hat{\beta})|^2 = (X(A\hat{\beta} - \hat{\beta}))^T X(\hat{\beta}_A - \hat{\beta})$$
$$= (A\hat{\beta} - c)^T(X(X^TX)^{-1}A^TD^{-1})^T X(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c)$$
$$= (A\hat{\beta} - c)^T D^{-1}A(X^TX)^{-1}X^TX(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c).$$
$$= (A\hat{\beta} - c)^T D^{-1}A(X^TX)^{-1}A^TD^{-1}(A\hat{\beta} - c)$$
$$= (A\hat{\beta} - c)^T D^{-1}DD^{-1}(A\hat{\beta} - c)$$
$$= (A\hat{\beta} - c)^T D^{-1}(A\hat{\beta} - c).$$

Comparing with (16.0.1) proves Lemma.

□

Using (16.0.2) and Lemma, we get that under null hypothesis H_0:

$$\frac{n-p}{ns\hat{\sigma}^2}|X(\hat{\beta}_A - \hat{\beta})|^2 \sim F_{s,n-p}. \tag{16.0.5}$$

There are many different models that are special cases of a multiple linear regression and many hypotheses about these model can be written as a general linear constraints. We will describe one such model in detail - one-way layout in analysis of variance. Then we will describe a couple of other models without going into details since the idea will become clear.

Analysis of variance: one-way layout. Suppose that we are given p independent samples

$$\begin{gathered} Y_{11}, \ldots, Y_{1n_1} \sim N(\mu_1, \sigma^2) \\ \vdots \\ Y_{p1}, \ldots, Y_{pn_p} \sim N(\mu_p, \sigma^2) \end{gathered}$$

of sizes $n_1, \ldots, n_p$ correspondingly. We assume that the variance of all distributions are equal. We would like to test the hypothesis that the means of all distributions are equal,

$$H_0 : \mu_1 = \ldots = \mu_p.$$

This problem is in fact a special case of a multiple linear regression and testing hypothesis given by linear equations. We can write

$$Y_{ki} = \mu_k + \varepsilon_{ki}, \quad \text{where} \quad g_{ki} \sim N(0, \sigma^2), \quad \text{for} \quad k = 1, \ldots, p, \;\; i = 1 \ldots, n_i.$$

Let us consider $n \times 1$ vector, where $n = n_1 + \ldots + n_p$,

$$Y = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{p1}, \ldots, Y_{pn_p})^T$$

and $p \times 1$ parameter vector

$$\mu = (\mu_1, \ldots, \mu_p)^T.$$

Then we can write all the equations in a matrix form

$$Y = X\mu + \varepsilon,$$

where X is the following $n \times p$ matrix:

$$X = \left(\begin{array}{cccc} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 \\ \hline 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 \end{array}\right).$$

The blocks have $n_1, \ldots, n_p$ rows. Basically, the predictor matrix X consists of indicators to which group the observation belongs to. The hypothesis H_0 can be written in a matrix form as $A\mu = 0$ for $(p-1) \times p$ matrix

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 & -1 \\ 0 & 1 & \ldots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & -1 \end{pmatrix}.$$

We need to compute the statistic in (16.0.5) that will have distribution $F_{p-1,n-p}$. First of all,

$$X^T X = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & n_r \end{pmatrix}.$$

Since $\hat{\mu} = (X^T X)^{-1} X^T Y$ it is easy to see that for each $i \leq p$,

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik} = \bar{Y}_i \text{ - the average of } i\text{th sample.}$$

We also get

$$\hat{\sigma}^2 = \frac{1}{n} |Y - X\mu|^2 = \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_i)^2.$$

To find the MLE $\hat{\mu}_A$ under the linear constraints $A\mu = 0$ we simply need to minimize $|Y - X\mu|^2$ over vectors $\mu = (\mu_1, \ldots, \mu_1)^T$ with all equal coordinates. But, obviously, $X\mu$ is a vector $(\mu_1, \ldots, \mu_1)^T$ of size $n \times 1$, so we need to minimize

$$\sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - \mu_1)^2 \min_{\mu_1}$$

and we get

$$\mu_1 = \frac{1}{n} \sum_{i=1}^{p} \sum_{k=1}^{n_i} Y_{ik} = \bar{Y} \text{ - overall average of all samples.}$$

Therefore,

$$\hat{\mu}_A - \hat{\mu} = (\bar{Y} - \bar{Y}_1, \ldots, \bar{Y} - \bar{Y}_p)^T$$

and

$$|X(\hat{\mu}_A - \hat{\mu})|^2 = \sum_{i=1}^{p} \sum_{k=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2.$$

By (16.0.5), under the null hypothesis H_0,

$$F := \frac{n-p}{p-1} \frac{\sum_{i=1}^{p} n_i (\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^{p} \sum_{k=1}^{n_i} (Y_{ik} - \bar{Y}_i)^2} \sim F_{p-1,n-p}. \tag{16.0.6}$$

In order to test H_0, we define a decision rule

$$\delta = \begin{cases} H_0, & F \le c_\alpha \\ H_1, & F > c_\alpha \end{cases}$$

where $F_{p-1,n-p}(c_\alpha, +\infty) = \alpha$. The sum in the numerator in (16.0.6) represents the total variation of the sample means $\bar{Y}_i$ of each population around the overall mean $\bar{Y}$. The sum in the numerator represent the total variation of the observations Y_{ik} around their particular sample means $\hat{Y}_i$. This interpretation of the test statistic explains the name - analysis of variance, or anova.

□

Example. Let us again consider normal body temperature dataset and perform anova test to compare the mean body temperature for men and women. Previously we have tested this using t-tests and KS test for two samples. We use Matlab function

```
[p,tbl,stats]=anova1([men, women])
```

where 'men' and 'women' are 65×1 vectors. For unequal groups 'anova1' requires a second argument with group labels. The output produces a table 'tbl':

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Columns'	[2.7188]	[1]	[2.7188]	[5.2232]	[0.0239]
'Error'	[66.6262]	[128]	[0.5205]		
'Total'	[69.3449]	[129]			

'SS' gives the sum of squares in the numerator of (16.0.6) ('Columns'), denominator ('Error'), and their total sum. Degrees of freedom 'df' represent degrees of freedom $p-1$ and $n-p$. 'MS' represents the normalized sums of squares by corresponding degrees of freedom. 'F' is a statistic in (16.0.6) and 'Prob>F' is a p-value corresponding to this F-statistic. This means that at the level of significance $\alpha = 0.05$ we reject the null hypothesis that the means are equal.

□

Analysis of variance: two-way layout. Suppose that we again have samples from different groups only now the groups will have two categories defined by two factors. For example, if we want to compare SAT scores in different states but also separate public and private schools then we will have groups defined by two factors - state and school type. We consider the following model of the data:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

for $i = 1, \dots, a, j = 1, \dots, b$ and $k = 1, \dots, n_{ij}$, i.e. we have a categories of the first factor, b categories of the second factor and n_{ij} observations in group (i, j). This model is not any different from one-way anova, simply the groups are indexed by two parameters/factors, but the estimation of parameters can be carried out as in the one-way anova. However, to test various hypotheses about the effects of these two factors it is more convenient to write the model in an equivalent way as follow:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where we assume that

$$\sum_{i=1}^{a} \alpha_i = 0, \ \sum_{j=1}^{b} \beta_j = 0, \ \sum_{i=1}^{b} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0.$$

These constraints define all parameters uniquely from original parameters μ_{ij}. Parameter μ is called the *overall mean.* The reason we separate *additive effects* α_i and β_j of two factors from the most general *interaction effect* γ_{ij} is because it is easier to formulate various hypotheses in terms of these parameters. For example:

- $H_0 : \alpha_1 = \ldots = \alpha_a = 0$ - the *additive* effect of the first factor is insignificant;
- $H_0 : \beta_1 = \ldots = \beta_b = 0$ - the *additive* effect of the second factor is insignificant;
- $H_0 :$ all $\gamma_{ij} = 0$ - the effect of the *interaction* of both factors is insignificant, i.e. the effect of factors is additive.

Matlab function 'anova2' performs two-way layout of anova if the sizes of all groups n_{ij} are equal, i.e. the data is *balanced*. If the sizes of groups are different one should use 'anovan' - a general N-way anova.

□

Analysis of covariance. This is another special case of multiple regression when all groups of data have a continuous predictor variable. The model is:

$$Y_{ik} = \alpha + \alpha_i + (\beta + \beta_i)X_{ik} + \varepsilon_{ik}$$

for $i = 1, \ldots, a$ and $k = 1, \ldots, n_i$. We have a groups and n_i observations in ith group. To determine the parameters uniquely we assume that

$$\sum_{i=1}^{a} \alpha_i = 0, \ \sum_{i=1}^{a} \beta_i = 0.$$

Example. (*Fruitfly dataset*) We consider a dataset from [1] (available on the journal's website) and [2]. The experiment consisted of five groups of male fruitflies, 25 male fruitflies in each group. The males in each group were supplied with different number of either receptive or non receptive females each day.

Group 1: 8 newly inseminated non-receptive females per day;

Group 2: no females;

Group 3: 1 newly inseminated non-receptive female per day;

Group 4: 1 receptive female per day;

Group 5: 8 receptive females per day.

The experiment was designed to test if the increased reproduction results in decreased longevity, so the lifespan of each male fruitfly was the response variable Y.

One-way anova. Let us start with a one-way anova, i.e. we consider a model

$$Y_{ij} = \mu_i + \varepsilon_{ik}, \quad \text{where } i = 1, \ldots, 5, \ \ k = 1, \ldots, 25$$

and test the hypothesis $H_0 : \mu_1 = \ldots = \mu_5$. Suppose that 'lifespan1' is a 25×5 matrix such that each column contains observations from one group. Then running

```
[p,tbl,stats]=anova1(lifespan1);
```

produces the boxplot in figure 16.1 and a table 'tbl':

```
'Source'     'SS'            'df'   'MS'            'F'        'Prob>F'
'Columns'    [1.1939e+004]   [  4]  [2.9848e+003]  [13.6120]  [3.5156e-009]
'Error'      [2.6314e+004]   [120]  [   219.2793]
'Total'      [3.8253e+004]   [124]
```

p-value suggests how unlikely hypothesis H_0 is. The boxplot suggests that the last group's

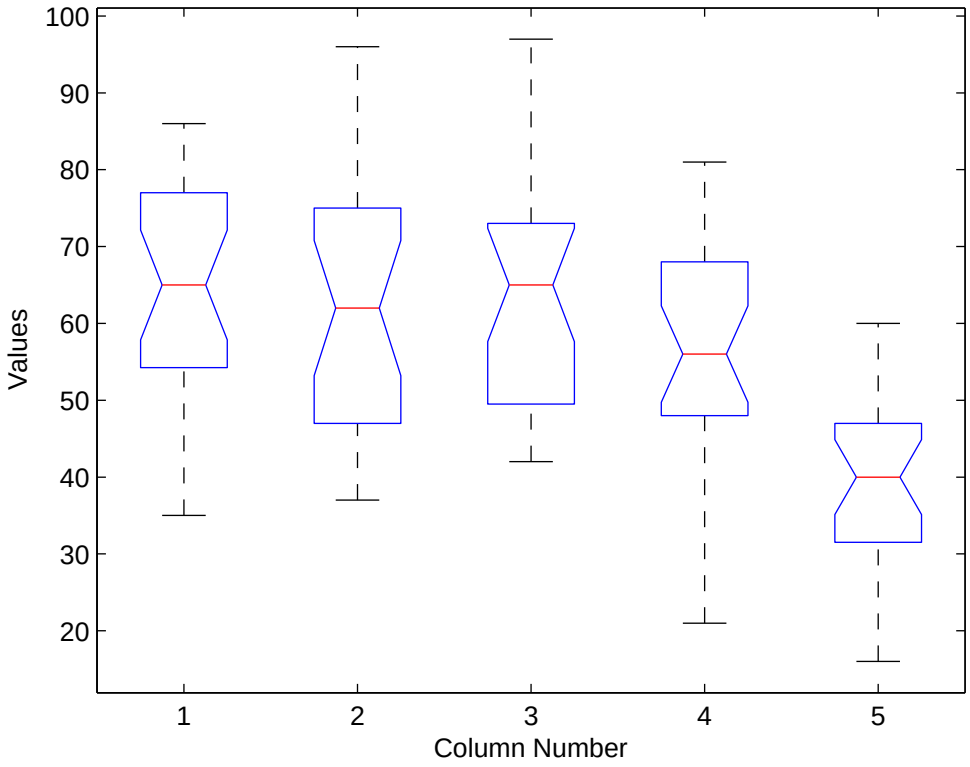


Figure 16.1: Boxplot for one-way ANOVA.

lifespan is most different from the other four groups. As a result, we might want to test the hypothesis $H_0 : \mu_1 = \ldots = \mu_4$ that the means of the first four groups are equal,

```
[p,tbl,stats]=anova1(lifespan1(:,1:4));
```

we get the following table

```
'Source'   'SS'           'df' 'MS'        'F'       'Prob>F'
'Columns'  [   988.0800] [ 3] [329.3600] [1.3869] [0.2515]
'Error'    [2.2798e+004] [96] [237.4842]
'Total'    [2.3787e+004] [99]
```

and we see that the p-value is 0.2515, so we accept H_0 if the level of significance $\alpha \leq p$-value.

□

Two-way anova. Let us now consider four groups without the second group (no females) and test the effects of two factors:

- Factor A: 'receptive' or 'non-receptive';
- Factor B: '1' or '8'.

This means that we consider a model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

for $i = 1, \ldots, 2, j = 1, \ldots, 2$ and $k = 1, \ldots, 25$. To use Matlab function 'anova2' we arrange the data into a 50×2 matrix 'lifespan2' such that two columns represent two categories of Factor A, the first 25 rows represent group '1' in Factor B and rows 26 through 50 represent group '8' in Factor B. Then

```
[p,tbl,stats]=anova2(lifespan2,25)
```

produces (here 25 indicates the number of replicas in one cell) the table

```
'Source'      'SS'           'df' 'MS'           'F'       'Prob>F'
'Columns'     [6.6749e+003] [ 1] [6.6749e+003] [32.3348] [1.3970e-007]
'Rows'        [1.7223e+003] [ 1] [1.7223e+003] [ 8.3430] [     0.0048]
'Interaction' [2.3717e+003] [ 1] [2.3717e+003] [11.4890] [     0.0010]
'Error'       [1.9817e+004] [96] [   206.4308]
'Total'       [3.0586e+004] [99]
```

p-values in the last column correspond to three hypotheses:

- $H_0 : \alpha_1 = \alpha_2 = 0$, i.e. the effect of Factor A is insignificant;
- $H_0 : \beta_1 = \beta_2 = 0$, i.e. the effect of Factor B is insignificant;
- $H_0 : \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$, i.e. the effect of the 'interaction' between Factors A and B is insignificant.

Small p-values suggest that all these hypotheses should be rejected.

□

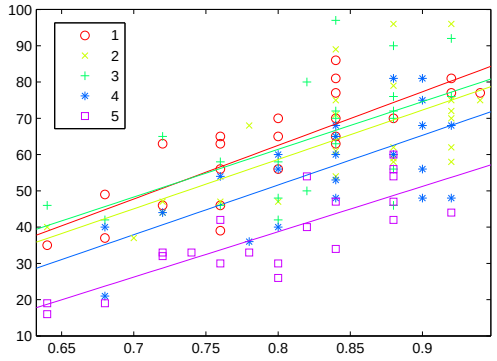
Analysis of covariance. Besides reproduction factors A and B, another continuous explanatory variable for longevity was used - the length of thorax (a division of a body between the head and the abdomen - chest). We are now in the setting of ancova:

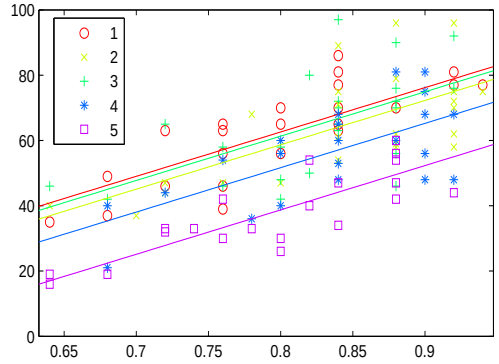
$$Y_{ik} = \alpha + \alpha_i + (\beta + \beta_i)X_{ik} + \varepsilon_{ik}$$

for $i = 1, \ldots, 5$ and $k = 1, \ldots, 25$. Analysis of covariance tool in Matlab

```
aoctool(thorax,lifespan,groups);
```

produces the following output, figure 16.2:





Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	-53.788	10.5457	-5.1	0
1	-1.912	19.7963	-0.1	0.9232
2	3.546	19.8575	0.18	0.8586
3	10.063	22.6154	0.44	0.6572
4	-4.204	22.703	-0.19	0.8534
5	-7.492	20.2771	-0.37	0.7124
Slope	135.473	12.7783	10.6	0
1	12.316	24.3219	0.51	0.6136
2	0.653	23.7582	0.03	0.9781
3	-4.024	27.3217	-0.15	0.8832
4	1.528	27.1208	0.06	0.9552
5	-10.473	25.0536	-0.42	0.6767

Coefficient Estimates

Term	Estimate	Std. Err.	T	Prob>|T|
Intercept	-54.062	10.2551	-5.27	0
1	8.006	1.8898	4.24	0
2	4.077	1.8894	2.16	0.0329
3	6.73	1.881	3.58	0.0005
4	-2.94	1.8914	-1.55	0.1227
5	-15.873	1.8981	-8.36	0
Slope	135.819	12.439	10.92	0

ANOVA Table

Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
groups	4	9611.5	2402.9	21.09	0
thorax	1	13168.9	13168.9	115.59	0
groups*thorax	4	42.5	10.6	0.09	0.9844
Error	115	13102.1	113.9		

ANOVA Table

Source	d.f.	Sum Sq	Mean Sq	F	Prob>F
groups	4	9611.5	2402.9	21.75	0
thorax	1	13168.9	13168.9	119.22	0
Error	119	13144.7	110.5		

Figure 16.2: Left column top to bottom: graph of fitted line for each group, estimates of coefficients, ancova test table. Right column: same under assumption that all slopes are equal.

We see that the p-value of 'groups*thorax' interaction, corresponding to the hypothesis that all $\beta_i = 0$, is equal to 0.9844, which means that we can accept this hypothesis. As a result, we fit the model with equal slopes for all groups, figure 16.2, right column. The p-values for 'groups' and 'thorax', corresponding to the hypotheses all $\alpha_i = 0$ and $\beta = 0$, are almost 0 and we should reject these hypotheses.

□

References.

[1] Hanley, J. A., and Shapiro, S. H. (1994), "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention," *Journal of Statistics Education*, Volume 2, Number 1.

[2] Linda Partridge and Marion Farquhar (1981), "Sexual Activity and the Lifespan of Male Fruitflies," *Nature*, 294, 580-581.

Section 17

Classification problem. Boosting.

Suppose that we have the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ that consist of pairs (X_i, Y_i) such that X_i belongs to some set $\mathcal{X}$ and Y_i belongs to a set $\mathcal{Y} = \{+1, -1\}$. We will think of Y_i as a label of X_i so that all points in the set $\mathcal{X}$ are divided into two classes corresponding to labels ± 1. For example, X_is can be images or representations of images and Y_is classify whether the image contains a human face or not. Given this data we would like to find a classifier

$$f : \mathcal{X} \to \mathcal{Y}$$

which given a point $X \in \mathcal{X}$ would predict its label Y. This type of problem is called classification problem. In general, there may be more than two classes of points which means that the set of labels may consist of more than two points but, for simplicity, we will consider the simplest case when we have only two labels ± 1.

We will take a look at one approach to this problem called boosting and, in particular, prove one interesting property of the algorithm called AdaBoost.

Let us assume that we have a family of classifiers

$$\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}.$$

Suppose that we can find many classifiers in $\mathcal{H}$ that can predict labels Y_i better than "tossing a coin" which means that they predict the correct label at least half of the time. We will call $\mathcal{H}$ a family of *weak classifiers* because we do not require much of them, for example, all these classifiers can make mistakes on, let's say, 30% or even 45% of the sample.

The idea of boosting consists in trying to combine these weak classifiers so that the combined classifier predicts the label correctly most of the time. Let us consider one particular algorithm called Adaboost.

Given weights $w(1), \ldots, w(n)$ that add up to one we define the weighted classification error of the classifier h by

$$w(1)I(h(X_1) \neq Y_1) + \ldots + w(n)I(h(X_n) \neq Y_n).$$

AdaBoost algorithm. We start by assigning equal weights to the data points:

$$w_1(1) = \ldots = w_1(n) = \frac{1}{n}.$$

Then for $t = 1, \ldots, T$ we repeat the following cycle:

1. Find $h_t \in \mathcal{H}$ such that weighted error

$$\varepsilon_t = w_t(1)I(h_t(X_1) \neq Y_1) + \ldots + w_t(n)I(h_t(X_n) \neq Y_n)$$

   is as small as possible.

2. Let $\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$ and update the weights:

$$w_{t+1}(i) = w_t(i)\frac{e^{-\alpha_t Y_i h_t(X_i)}}{Z_t},$$

   where

$$Z_t = \sum_{i=1}^{n} w_t e^{-\alpha_t Y_i h_t(X_i)}$$

   is the normalizing factor to ensure that updated weights add up to one.

After we repeat this cycle T times we output the function

$$f(X) = \alpha_1 h_1(X) + \ldots + \alpha_T h_T(X)$$

and use $\text{sign}(f(X))$ as the prediction of label Y.

First of all, we can assume that the weighted error ε_t at each step t is less than 0.5 since, otherwise, if we make a mistake more than half of the time we should simply predict the opposite label. For $\varepsilon_t \leq 0.5$ we have,

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} \geq 0.$$

Also, we have

$$Y_i h_t(X_i) = \begin{cases} +1 & \text{if } h_t(X_i) = Y_i \\ -1 & \text{if } h_t(X_i) \neq Y_i. \end{cases}$$

Therefore, if h_t makes a mistake on the example (X_i, Y_i) which means that $h_t(X_i) \neq Y_i$ or, equivalently, $Y_i h_t(X_i) = -1$ then

$$w_{t+1}(i) = \frac{e^{-\alpha_t Y_i h_t(X_i)}}{Z_t} w_t(i) = \frac{e^{\alpha_t}}{Z_t} w_t(i).$$

On the other hand, if h_t predicts the label Y_i correctly then $Y_i h_t(X_i) = 1$ and

$$w_{t+1}(i) = \frac{e^{-\alpha_t Y_i h_t(X_i)}}{Z_t} w_t(i) = \frac{e^{-\alpha_t}}{Z_t} w_t(i).$$

Since $\alpha_t \geq 0$ this means that we increase the relative weight of the ith example if we made a mistake on this example and decrease the relative weight if we predicted the label Y_i

correctly. Therefore, when we try to minimize the weighted error at the next step $t+1$ we will pay more attention to the examples misclassified at the previous step.

Theorem: *The proportion of mistakes made on the data by the output classifier* $\text{sign}(f(X))$ *is bounded by*

$$\frac{1}{n}\sum_{i=1}^{n} I(\text{sign}(f(X_i))) \neq Y_i) \leq \prod_{t=1}^{T} \sqrt{4\varepsilon_t(1-\varepsilon_t)}.$$

Remark: If the weighted errors ε_t will be strictly less than 0.5 at each step meaning that we predict the labels better than tossing a coin then the error of the combined classifer will decrease exponentially fast with the number of rounds T. For example, if $\varepsilon_t \leq 0.4$ then $4\epsilon_t(1-\epsilon_t) \leq 4(0.4)(0.6) = 0.96$ and the error will decrease as fast as 0.96^T.

Proof. Using that $I(x \leq 0) \leq e^{-x}$ as shown in figure 17.1 we can bound the indicator of making an error by

$$I(\text{sign}(f(X_i)) \neq Y_i) = I(Y_i f(X_i) \leq 0) \leq e^{-Y_i f(X_i)} = e^{-Y_i \sum_{t=1}^{T} \alpha_t h_t(X_i)}. \tag{17.0.1}$$

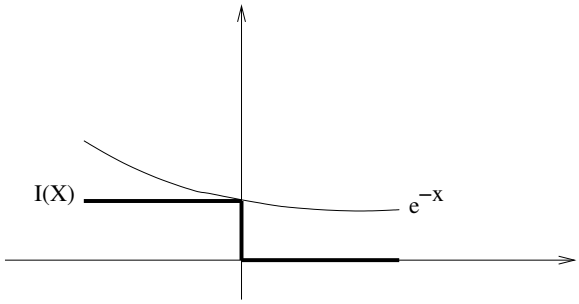


Figure 17.1: Example.

Next, using the step 2 of AdaBoost algorithm which describes how the weights are updated we can express the weights at each step in terms of the weights at the previous step and we can write the following equation:

$$\begin{aligned} w_{T+1}(i) &= \frac{w_T(i)e^{-\alpha_T Y_i h_T(X_i)}}{Z_T} = \frac{e^{-\alpha_T Y_i h_T(X_i)}}{Z_T}\frac{w_{T-1}(i)e^{-\alpha_{T-1} Y_i h_{T-1}(X_i)}}{Z_{T-1}} \\ &= \text{repeat this recursively over } t \\ &= \frac{e^{-\alpha_T Y_i h_T(X_i)}}{Z_T}\frac{e^{-\alpha_{T-1} Y_i h_{T-1}(X_i)}}{Z_{T-1}} \cdots \frac{e^{-\alpha_1 Y_i h_1(X_i)}}{Z_1} w_1(i) = \frac{e^{-Y_i f(X_i)}}{\prod_{t=1}^{T} Z_t}\frac{1}{n}. \end{aligned}$$

This implies that

$$\frac{1}{n}e^{-Y_i f(X_i)} = w_{T+1}(i)\prod_{t=1}^{T} Z_t.$$

Combining this with (17.0.1) we can write

$$\frac{1}{n}\sum_{i=1}^{n} I(\text{sign}(f(X_i) \neq Y_i)) \leq \sum_{i=1}^{n}\frac{1}{n}e^{-Y_i f(X_i)} = \prod_{t=1}^{T} Z_t \sum_{i=1}^{n} w_{T+1}(i) = \prod_{t=1}^{T} Z_t. \tag{17.0.2}$$

Next we will compute

$$Z_t = \sum_{i=1}^{n} w_t(i) e^{-\alpha_t Y_i h_t(X_i)}.$$

As we have already mentioned above, $Y_i h_t(X_i)$ is equal to -1 or $+1$ depending on whether h_t makes a mistake or predicts the label Y_i correctly. Therefore, we can write,

$$\begin{aligned} Z_t &= \sum_{i=1}^{n} w_t(i) e^{-\alpha_t Y_i h_t(X_i)} = \sum_{i=1}^{n} w_t(i) I(Y_i = h_t(X_i)) e^{-\alpha_t} + \sum_{i=1}^{n} w_t(i) I(Y_i \neq h_t(X_i)) e^{\alpha_t} \\ &= e^{-\alpha_t} (1 - \underbrace{\sum_{i=1}^{n} w_t(i) I(Y_i \neq h_t(X_i))}_{\varepsilon_t}) + e^{\alpha_t} \underbrace{\sum_{i=1}^{n} w_t(i) I(Y_i = h_t(X_i))}_{\varepsilon_t} \\ &= e^{-\alpha_t}(1 - \varepsilon_t) + e^{\alpha_t} \varepsilon_t. \end{aligned}$$

Up to this point all computations did not depend on the choice of α_t but since we bounded the error by $\prod_{t=1}^{T} Z_t$ we would like to make each Z_t as small as possible and, therefore, we choose α_t that minimizes Z_t. Simple calculus shows that we should take $\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$ which is precisely the choice made in AdaBoost algorithm. For this choice of α_t we get

$$Z_t = (1 - \varepsilon_t) \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} + \varepsilon_t \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} = \sqrt{4 \varepsilon_t (1 - \varepsilon_t)}$$

and plugging this into (17.0.2) finishes the proof of the bound.

□