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.

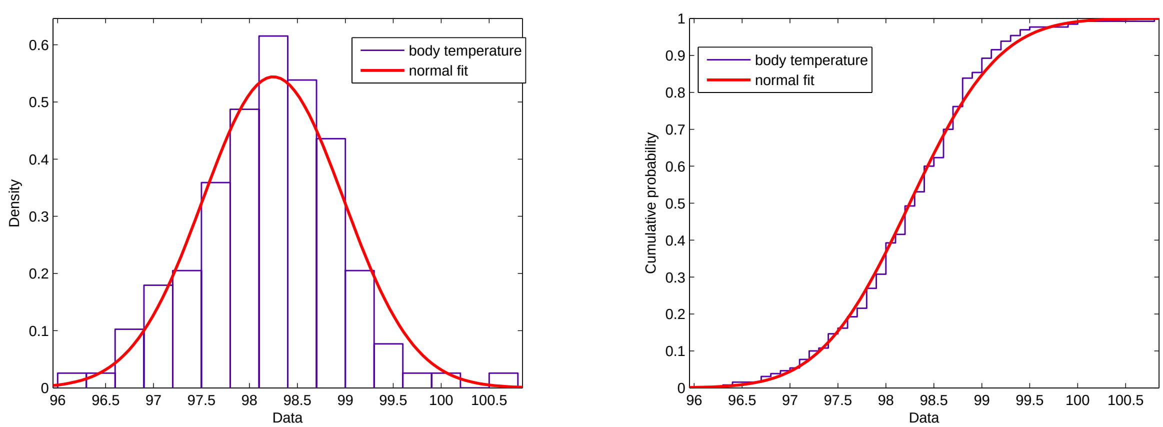

Figure 5.1: Fitting a body temperature dataset. (a) Histogram of the data and p.d.f. of fitted normal distribution; (b) Empirical c.d.f. and c.d.f. of fitted normal distribution.

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_n^2-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,*

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

i.e. $\sqrt{n}\bar{X}$ has standard normal distribution and $n(\bar{X^2} - (\bar{X})^2)$ has χ_{n-1}^2 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

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

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

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

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

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

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.

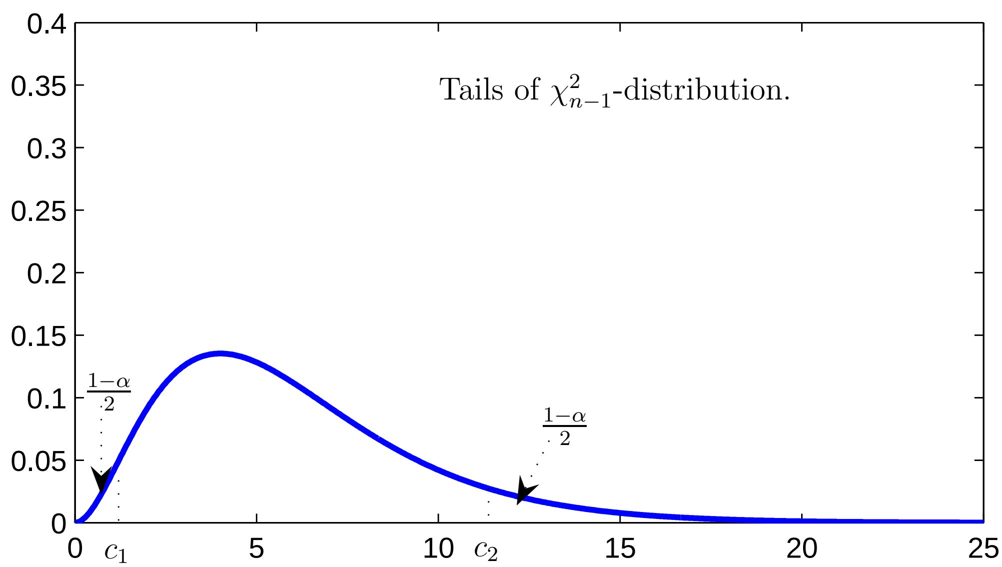

Figure 5.2: p.d.f. of χ^2_{n-1}-distribution and α-confidence interval.

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:

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

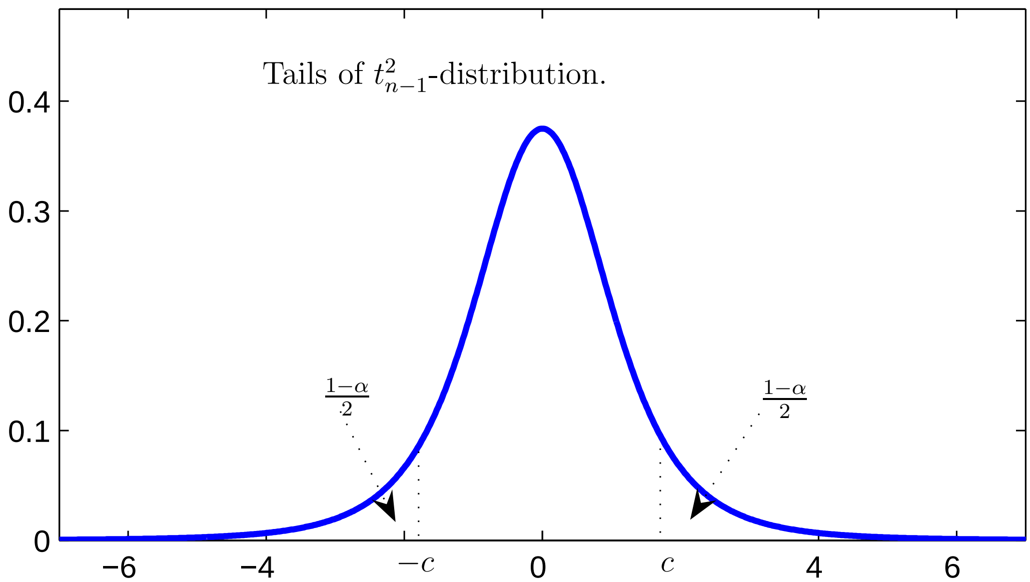

Figure 5.3: p.d.f. of t_{n-1} distribution and confidence interval for μ.

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.