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.

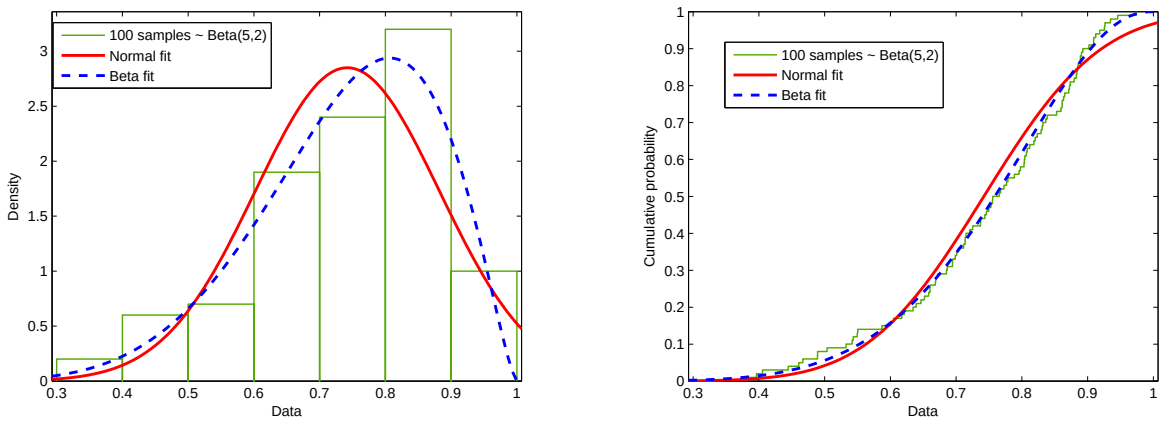


Figure 2.1: Fitting a random sample of size 100 from Beta$(5, 2)$. (a) Histogram of the data and p.d.f.s of fitted normal (solid line) and beta (dashed line) distributions; (b) Empirical c.d.f. and c.d.f.s of fitted normal and beta distributions.

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.

□

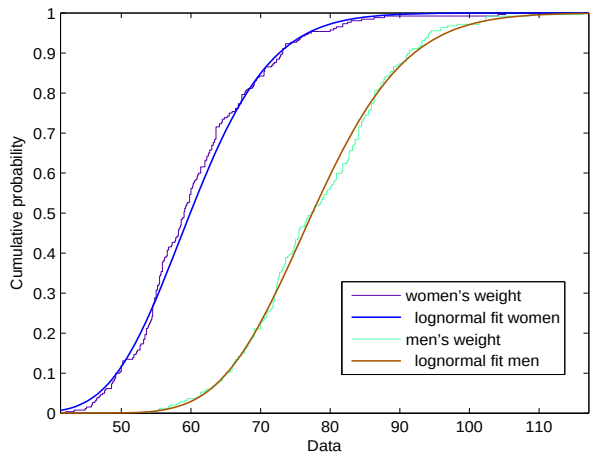
Example. We consider a dataset of various body measurements from [1] (dataset can be dowloaded from journal's website), including weight, height, waist girth, abdomen girth, etc. First, we use Matlab fitting tool to fit weight and waist girth of men and women (separately) with lognormal distribution, 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:

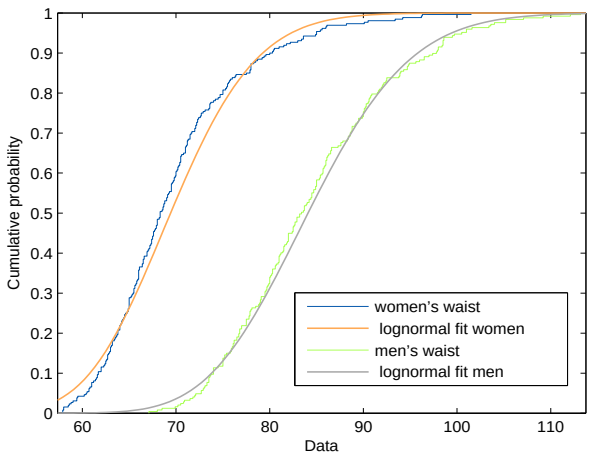
```
[h,p,stats]=chi2gof(log_women_waist)

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





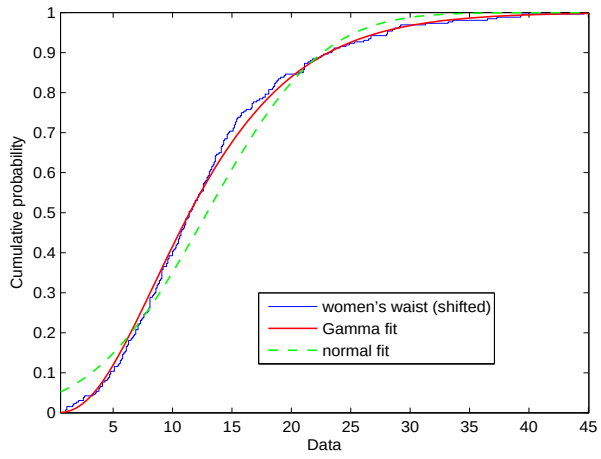


Figure 2.2: Fitting weight (upper left) and waist girth (upper right) with lognormal distribution. Lower left: fitting women's waist with shifted Gamma and normal distributions.

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.

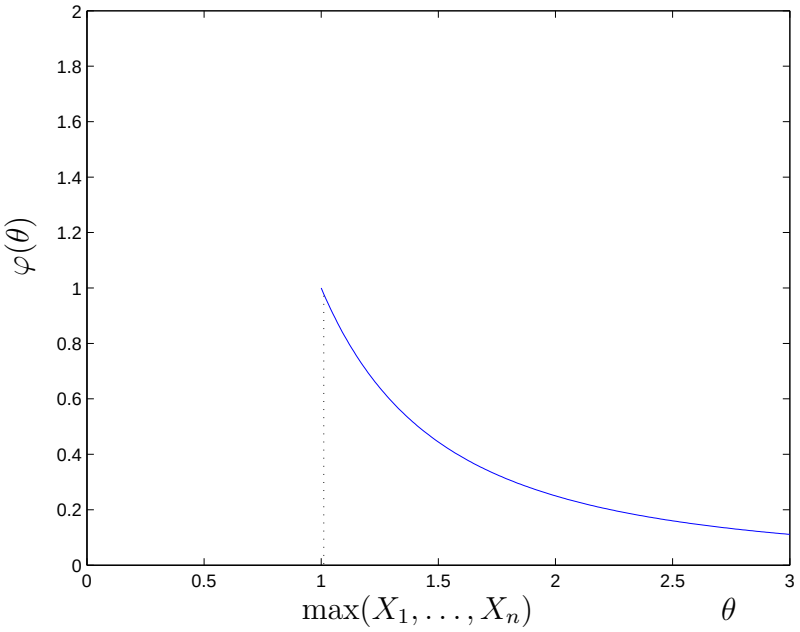


Figure 2.3: MLE for the uniform distribution.

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.