

Lecture 3. Fitting Distributions to data - choice of a model.

Igor Rychlik

Chalmers

Department of Mathematical Sciences

Probability, Statistics and Risk, MVE300 • Chalmers • April 2013. Click on **red text** for extra material.

Random variables and cdf.

Random variable is a numerical outcome X , say, of an experiment. To describe its properties one needs to find probability distribution $F_X(x)$. Three approaches will be discussed:

- I Use only the observed values of X (data) to model the variability of X , i.e. normalized histogram, empirical cdf, see Lecture 2.
- II Try to find the proper cdf by means of reasoning. For example a number of heads in 10 flips of a fair coin is $\text{Bin}(10, 1/2)$.
- III Assume that F_X belongs to a class of distributions $b + aY$, for example Y standard normal. Then choose values of parameters a, b that best "fits" data.

Case II - Example:

Let roll a fair die. Sample space $\mathcal{S} = \{1, \dots, 6\}$ and let random variable K be the number shown. All results are equally probable hence $p_k = P(K = k) = 1/6$.

In 1882, R. Wolf rolled a die $n = 20\,000$ times and recorded the number of eyes shown

Number of eyes k	1	2	3	4	5	6
Frequency n_k	3407	3631	3176	2916	3448	3422

Was his die fair?

The χ^2 test, proposed by Karl Pearson' (1857-1936), can be used to investigate this issue.

Pearson' χ^2 test:

Hypothesis H_0 : We claim that

$$P(\text{"Experiment results in outcome } k\text{"}) = p_k, \quad k = 1, \dots, r.$$

In our example $r = 6$, $p_k = 1/6$.

Significance level α : Select the probability (risk) of rejecting a true hypothesis. Constant α is often chosen to be 0.05 or 0.01. Rejecting H_0 with a lower α indicates stronger evidence against H_0 .

Data: In n experiments one observed n_k times outcome k .

Test: Estimate p_k by $p_k^* = n_k/n$. Large distances $p_k - p_k^*$ make hypothesis H_0 questionable. Pearson proposed to use the following statistics to measure the distance:

$$Q = \sum_{k=1}^r \frac{(n_k - np_k)^2}{np_k} \left(= n \sum_{k=1}^r \frac{(p_k^* - p_k)^2}{p_k} \right) \quad (1)$$

Details of the χ^2 test

How large Q should be to reject the hypothesis? Reject H_0 if $Q > \chi_{\alpha}^2(f)$, where $f = r - 1$. Further, in order to use the test, as a rule of thumb one should check that $np_k > 5$ for all k .

Example 1

For Wolf's data Q is

$$Q = 1.6280 + 26.5816 + 7.4261 + 52.2501 + 3.9445 + 2.3585 = 94.2$$

Since $f = r - 1 = 5$ and the quantile $\chi_{0.05}^2(f) = 11.1$, we have $Q > \chi_{0.05}^2(5)$ which leads to **rejection of the hypothesis of a fair dice**.¹

Example 2

Are children birth months uniformly distributed? **Data**,

Matlab code:

¹Not rejecting the hypothesis does not mean that there is strong evidence that H_0 is true. It is recommendable to use the terminology "reject hypothesis H_0 " or "not reject hypothesis H_0 " but not to say "accept H_0 ".

Case III - parametric approach to find F_X .

Parametric estimation procedure of F_X contains three main steps:
choice of a model; finding the parameters; analysis of error:

- ▶ Choose a model, i.e. select one of the standard distributions $F(x)$ (normal, exponential, Binomial, Poisson ...). Next postulate that

$$F_X(x) = F\left(\frac{x - b}{a}\right).$$

- ▶ Find estimates (a^*, b^*) such that $F_n(x) \approx F((x - b^*)/a^*)$ ($F_X(x) \approx F_n(x)$), here first **method of moments** to estimates parameters will be presented. Then more advanced and often more accurate **maximum likelihood** method will be presented on the next lecture.

Moments of a rv. - Law of Large Numbers (LLN)

- ▶ Let X_1, \dots, X_k be a sequence of iid variables all having the distribution $F_X(x)$. Let $E[X]$ be a constant, called the **expected value** of X ,

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx, \text{ or } E[K] = \sum_k k p_k$$

- ▶ If the expected value of X exists and is finite then, as k increases (we are averaging more and more variables), the average

$$\frac{1}{k}(X_1 + X_2 + \dots + X_k) \approx E[X]$$

with equality when k approaches infinity.

- ▶ Linearity property $E[a + bX + cY] = a + bE[X] + cE[Y]$.

Example 3

Other moments

- ▶ Let X_i be iid all having the distribution $F_X(x)$. Let us also introduce constants called the **moments** of X , defined by

$$E[X^n] = \int_{-\infty}^{+\infty} x^n f_X(x) dx \quad \text{or} \quad E[K^n] = \sum_k k^n p_k.$$

- ▶ If $E[X^n]$ exists and is finite then, as k increases, the average

$$\frac{1}{k}(X_1^n + X_2^n + \cdots + X_k^n) \approx E[X^n].$$

- ▶ The same is valid for other functions of r.v.

Variance, Coefficient of variation

- ▶ The **variance** $V[X]$ and **coefficient of variation** $R[X]$

$$V[X] = E[X^2] - E[X]^2, \quad R[X] = \frac{\sqrt{V[X]}}{E[X]}.$$

- ▶ IF X, Y are independent then

$$V[a + bX + cY] = b^2V[X] + c^2V[Y].$$

Example 4

- ▶ Note that $V[X] \geq 0$. If $V[X] = 0$ then X is a constant.

Example: Expectations and variances.

Example 5

Expected yearly wind energy production, on blackboard.

Distribution		Expectation	Variance
Beta distribution, Beta(a, b)	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, 0 < x < 1$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Binomial distribution, Bin(n, p)	$p_k = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \dots, n$	np	$np(1-p)$
First success distribution	$p_k = p(1-p)^{k-1}, k = 1, 2, 3, \dots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Geometric distribution	$p_k = p(1-p)^k, k = 0, 1, 2, \dots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Poisson distribution, Po(m)	$p_k = e^{-m} \frac{m^k}{k!}, k = 0, 1, 2, \dots$	m	m
Exponential distribution, Exp(a)	$F(x) = 1 - e^{-x/a}, x \geq 0$	a	a^2
Gamma distribution, Gamma(a, b)	$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, x \geq 0$	a/b	a/b^2
Gumbel distribution	$F(x) = e^{-e^{-(x-b)/a}}, x \in \mathbb{R}$	$b + \gamma a$	$a^2 \pi^2 / 6$
Normal distribution, N(m, σ^2)	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}, x \in \mathbb{R}$ $F(x) = \Phi((x-m)/\sigma), x \in \mathbb{R}$	m	σ^2
Log-normal distribution, $\ln X \in N(m, \sigma^2)$	$F(x) = \Phi\left(\frac{\ln x - m}{\sigma}\right), x > 0$	$e^{m+\sigma^2/2}$	$e^{2m+2\sigma^2} - e^{2m+\sigma^2}$
Uniform distribution, U(a, b)	$f(x) = 1/(b-a), a \leq x \leq b$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
Weibull distribution	$F(x) = 1 - e^{-\left(\frac{x-b}{c}\right)^a}, x \geq b$	$b + a\Gamma(1 + 1/c)$	

$$a^2 \left[\Gamma\left(1 + \frac{2}{c}\right) - \Gamma^2\left(1 + \frac{1}{c}\right) \right]$$

Method of moments to fit cdf to data:

- ▶ When a cdf $F_X(x)$ is specified then one can compute the expected value, variance, coefficient of variation and other moments $E[X^k]$.
- ▶ If cdf $F_X(x) = F\left(\frac{x-b}{a}\right)$, i.e. depends on two parameters a, b then also moments are function of the parameters.

$$E[X^k] = m_k(a, b)$$

- ▶ LLN tells us that having independent observations x_1, \dots, x_n of X the average values

$$\bar{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k \rightarrow E[X^k], \quad \text{as } n \rightarrow \infty.$$

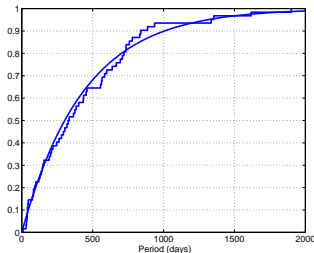
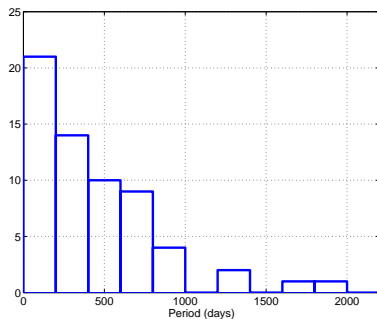
- ▶ **Methods of moments** recommends to estimate the parameters a, b by a^*, b^* that solve the equation system

$$m_k(a^*, b^*) = \bar{m}_k, \quad k = 1, 2.$$

Periods in days between serious earthquakes:

Example 6

By experience we choose exponential family
 $F_X(x) = 1 - e^{-x/a}$. Since $a = E[X]$ we choose $a^* = \bar{x} = 437.2$ days.



Left figure - histogram of 62 observed times between earthquakes. Right figure - comparison of the fitted exponential cdf to the earthquake data compared with ecdf - we can see that the two distributions are very close.

Is $a = a^*$, i.e. is error $e = a - a^* = a - 437.2 = 0$?

Example 7

Poisson cdf The following data set gives the number of killed drivers of motorcycles in Sweden 1990-1999:

39 30 28 38 27 29 38 33 33 36.

Assume that the number of killed drivers per year is modeled as a random variable $K \in \text{Po}(m)$ and that numbers of killed drivers during consecutive years, are independent and identically distributed.

From the table we read that $E[K] = m$ hence methods of moments recommends to estimate parameter m by the average number $m^* = \bar{k}$, viz. $m^* = (39 + \dots + 36)/10 = 33.1$.

Is $m = m^*$, i.e. is error $e = m - m^* = m - 33.1 = 0$?

Gaussian model

Example 8

Since $V[X] = E[X^2] - E[X]^2$ LLN gives the following estimate of the variance

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow V[X], \quad \text{as } n \text{ tends to infinity.}$$

We proposed to model weight of newborn baby X by normal (Gaussian) cdf $N(m, \sigma^2)$. Since $E[X] = m$ and $V[X] = \sigma^2$ hence the method of moments recommends to estimate m, σ^2 by $m^* = \bar{x}$, $(\sigma^2)^* = s_n^2$. For the data $m^* = 3400$ g, $(\sigma^2)^* = 570^2$, g².

Are $m = m^*$ and $\sigma^2 = s_n^2$, i.e. are errors

$$e_1 = m - m^* = m - 33.1 = 0, \quad e_2 = \sigma^2 - (\sigma^2)^* = \sigma^2 - 570^2 = 0?$$

Weibull model

For environmental variables often Weibull cdf fits well data. Suppose that

$$F_X(x) = 1 - \exp\left(-\left(\frac{x}{a}\right)^c\right),$$

a is scale parameter, c shape parameter. Using the table we have that

$$E[X] = a\Gamma(1 + 1/c), \quad R[X] = \frac{\sqrt{\Gamma(1 + 2/c) - \Gamma(1 + 1/c)^2}}{\Gamma(1 + 1/c)}.$$

Method of moments: estimate the coefficient of variation by $\sqrt{s_n^2/\bar{x}}$, solve numerically the second equation for c^* , see Table 4 on page 256, then $a^* = \bar{x}/\Gamma(1 + 1/c^*)$.

Example 9

Fitting Weibull cdf to bearing lifetimes

Example 10

Fitting Weibull cdf to wind speeds measurements

Estimation error:

In for the exponential, Poisson and Gaussian models the unknown parameter θ were $E[X]$ and has been estimated by $\theta^* = \bar{x}$. The estimation error $e = \theta - \theta^*$ is unknown (θ is not known). We want to describe the possible values of e by finding the distribution of the estimation error $\mathcal{E} = \theta - \theta^*$!

Let X_1, X_2, \dots, X_n be a sequence of n iid random variables each having finite values of expectation $m = E[X_1]$ and variance $V[X_1] = \sigma^2 > 0$. The **central limit theorem** (CLT) states that as the sample size n increases, the distribution of the sample average \bar{X} of these random variables approaches the normal distribution with a mean m and variance σ^2/n irrespective of the shape of the original distribution. ²

²"The first version of CLT was postulated by the French-born mathematician Abraham de Moivre who, in a remarkable article published in 1733, used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin."

Computation of $m_{\mathcal{E}}$, $\sigma_{\mathcal{E}}^2$.

Using **Central Limit Theorem** we can approximate cdf $F_{\mathcal{E}}(e)$ by normal distribution $N(m_{\mathcal{E}}, \sigma_{\mathcal{E}}^2)$, where $m_{\mathcal{E}} = E[\mathcal{E}]$, $\sigma_{\mathcal{E}}^2 = V[\mathcal{E}]$.

It is easy to demonstrate (see blackboard) that for the studied cases $E[\Theta^*] = \theta$ and hence $m_{\mathcal{E}} = E[\mathcal{E}] = 0$. Estimators having $m_{\mathcal{E}} = 0$ are called **unbiased**.

Similarly one can show that $\sigma_{\mathcal{E}}^2 = V[\mathcal{E}] = V(X)/n$ (see blackboard). Using the table we have that:

- ▶ $\sigma_{\mathcal{E}}^2 = m/n$ if X is Poisson $Po(m)$
- ▶ $\sigma_{\mathcal{E}}^2 = a^2/n$ if X is $\text{Exp}(a)$
- ▶ $\sigma_{\mathcal{E}}^2 = \sigma^2/n$ if X is $N(m, \sigma^2)$ ³

³Problem, variance $\sigma_{\mathcal{E}}^2$ depends on unknown parameters! Since $\theta^* \rightarrow \theta$ as $n \rightarrow \infty$ one is estimating $\sigma_{\mathcal{E}}^2$ by replacing θ by θ^* and denote the approximation by $(\sigma_{\mathcal{E}}^2)^*$.

In this lecture we met following concepts:

- ▶ χ^2 -test.
- ▶ Method of moments to fit(cdf) to data.
- ▶ Examples of data described using exponential, Poisson, Gaussian (normal) and Weibull cdf.
- ▶ Central Limit Theorem, giving normal distribution of estimation errors.

Examples in this lecture "click"