Lecture 2. Distributions and Random Variables

Igor Rychlik

Chalmers

Department of Mathematical Sciences

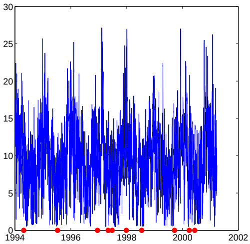
Probability, Statistics and Risk, MVE300/MVE395 • Chalmers
• March 2014. • Click on red text for extra material.

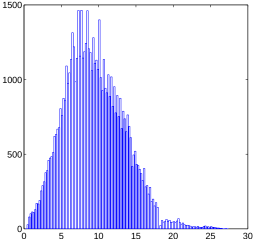
Wind Energy production:

Available Wind Power

$$p = 0.5\rho_{air} A_r\, v^3$$

, ρ_{air} air density, A_r area swept by rotor, v - hourly wind speed.





Left: 7 years data.
Right: Histogram
wind "distribution".

Location of wind mill depends on expected yearly production:

$$p_{yr} = \frac{1}{7} 0.5\rho_{air} A_r \sum_{i=1}^{61354} v_i^3 = 116678, \quad \text{[some units]}.$$

p_{yr} could be estimated using the histogram (wind speed distribution).

Random variables:

Often in engineering or the natural sciences, outcomes of random experiments are numbers associated with some physical quantities. Such experiments, called **random variables**, will be denoted by capital letters, e.g., U, X, Y, N, K.

The set $\mathcal{S}$ of possible values of a random variable is a sample space which can be all real numbers, all integer numbers, or subsets thereof.

Example 1 For the experiment flipping a coin, let to the outcomes "Tails" and "Heads" assign the values 0 and 1 and denote by X. One say that X is **Bernoulli distributed**. What does it mean "distributed"?

Uniformly distributed random variables:

Experiment: Roll a die. Number shown on the die N is a random variable.

If die is fair N is uniformly distributed.

Example 2 Examples of random variables: N_{2005} - number of children born in Stockholm year 2005. K a month a child (selected at random) was born. Data. Is K uniformly distributed?

How to get uniformly distributed number in $[0, 1]$, U, say?

Use binary representation of the numbers $u = 011\dots$

$$u = \frac{0}{2} + \frac{1}{4} + \frac{1}{8} + \dots.$$

Probability distribution function:

A statement of the type "$X \leq x$" for any fixed real value x, e.g. $x = -2.1$ or $x = 5.375$, plays an important role in computation of probabilities for statements on random variables and a function

$$F_X(x) = \mathrm{P}(X \leq x), \quad x \in \mathbb{R},$$

is called the **probability distribution**, **cumulative distribution function**, or **cdf** for short.

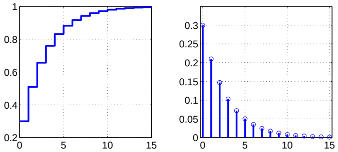
Example 3 Distribution of K. Figures

Example 4 The probability of any statement about the random variable X is computable (at least in theory) when $F_X(x)$ is known.

Probability mass function

If K takes a finite or (countable) number of values it is called **discrete** random variables and the distribution function $F_K(x)$ is a "stair" looking function that is constant except the possible jumps. The size of a jump at $x = k$, say, is equal to the probability $\mathrm{P}(K = k)$, denoted by p_k, and called the **probability-mass function**.

Example 5 Probability mass function of K. Pmf



Geometrical distribution with
$p_k = 0.70^k \cdot 0.30$, for $k = 0, 1, 2, \ldots$.

Left: Distribution function.

Right: Probability-mass function.

Counting variables

Geometric probability-mass function:

$$\mathrm{P}(K = k) = p\,(1-p)^k, \quad k = 0, 1, 2, \ldots$$

Binomial probability-mass function:

$$\mathrm{P}(K = k) = p_k = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

Poisson probability-mass function:

$$\mathrm{P}(K = k) = \mathrm{e}^{-m} \frac{m^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

Example 6 Binomially distributed random variable.

Ladislaus Bortkiewicz

Ladislaus Bortkiewicz
(1868-1931)

Important book published in 1898:

Das Gesetz der kleinen Zahlen

Law of Small Numbers

If an experiment is carried out by n independent trials and the probability for "success" in each trial is p, then the number of successes K is given by the binomial distribution:

$$K \in \mathrm{Bin}(n, p).$$

If $n \to \infty$ and $p \to 0$ so that $m = n \cdot p$ is constant, we have approximately that

$$K \in \mathrm{Po}(np).$$

(The approximation is satisfactory if $p < 0.1$ and $n > 10$.)

Example 7 Let p be probability that accident occurs during one year, n be number of structures (years) then number of accidents during one year $K \in \mathrm{Po}(np)$, example of accident.

Example 8 How good is Poisson approximation?

CDF - defining properties:

Any function $F(x)$ satisfying the following three properties is a distribution of some random variable:

- The distribution function $F_X(x)$ is non-decreasing function.
- $F_X(-\infty) = 0$ while $F_X(+\infty) = 1$.
- $F_X(x)$ is right continuous.

If $F_X(x)$ is continuous then $\mathrm{P}(X = x) = 0$ for all x and X is called **continuous**. The derivative $f_X(x) = F_X'(x)$ is called **probability density function** (pdf) and

$$F_X(x) = \int_{-\infty}^{x} f_X(z)\,dz.$$

Hence any positive function that integrates to one defines a cdf.

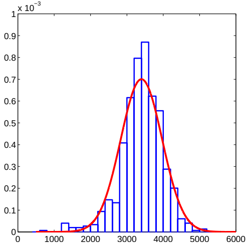
Example 8

Normal pdf- and cdf-function:

The cdf of standard normal cdf is defined through its pdf-function:

$$\mathsf{P}(X \le x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\xi^2/2}\, \mathrm{d}\xi.$$

The class of normal distributed variables $Y = m + \sigma X$, where $m,\ \sigma > 0$ are constants is extremely versatile. From a theoretical point of view, it has many advantageous features; in addition, variability of measurements of quantities in science and technology are often well described by normal distributions.



Example 10 Normalized histogram of weights of 750 newborn children in Malmö.

Solid line the normal pdf with $m = 3400$ g, $\sigma = 570$ g.

Is this a good model? Have girls and boys the same weights variability?

Standard Distributions

In this course we shall meet many classes of discrete cdf: Binomial, Geometrical, Poisson, ...; and continuous cdf: uniform, normal (Gaussian), log-normal, exponential, χ^2, Weibull, Gumbel, beta ...

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

Quantiles

The α *quantile* x_α, $0 \leq \alpha \leq 1$, is a generalization of the concepts of median and quartiles and is defined as follows:

The **quantile** x_α for a random variable X is defined by the following relations:

$$\mathrm{P}(X \leq x_\alpha) = 1 - \alpha, \quad x_\alpha = F^{-}(1 - \alpha).$$

Finding quantiles of normal cdf *Example 11*

Empirical probability distribution

- Suppose experiment was repeated n times rending in a sequence of X values, $x_1, \ldots, x_n$. The fraction $F_n(x)$ of the observations satisfying the condition "$x_i \leq x$"

$$F_n(x) = \frac{\text{number of } x_i \leq x,\ i = 1, \ldots, n}{n}$$

  is called the **empirical cumulative distribution function** (ecdf).

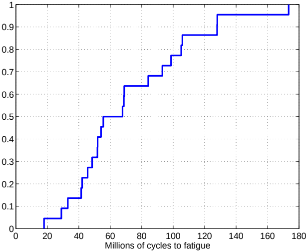
- The **Glivenko–Cantelli Theorem** states that the maximal distance between $F_n(x)$ and $F_X(x)$ tends to zero when n increases without bounds, viz. $\max_x |F_X(x) - F_n(x)| \to 0$ as $n \to \infty$.

- Assuming that $F_X(x) = F_n(x)$, means that the uncertainty in the future (yet unknown) value of X is model by means of drawing a lot from an urn, where lots contain only the observed values x_i. By Glivenko-Cantelli th. this is a good model when n is large.

Example: lifetimes for ball bearings

Data:

17.88,	28.92,	33.00,	41.52,	42.12,	45.60,	48.48,
51.84,	51.96,	54.12,	55.56,	67.80,	68.64,	68.88,
84.12,	93.12,	98.64,	105.12,	105.84,	127.92,	128.04,
173.40.						



ECDF of ball bearings life time.

Example wind speed data

In this lecture we met following concepts:

- Random variables (rv).
- Probability distribution (cdf), mass function (pmf), density (pdf).
- Law of small numbers.
- Quantiles.
- Empirical cdf.
- You should read how to generate uniformly distributed random numbers.
- How to generate non-uniformly distributed random numbers by just transforming uniform random numbers.

Examples in this lecture "click"