

Biostatistics

O.O.Bogomolets National Medical University

March 4, 2013

Crash course into Probability theory

- Probability, Random variables
- Mean and Variance
- Probability density function, Probability distributions
- Central limit theorem

- Probability

- Probability

- 1 The probability of a random event denotes the relative frequency of occurrence of an experiment's outcome, when repeating the experiment.
- 2 Probability is a way to represent an individual's degree of belief in a statement.

- Probability

- 1 The probability of a random event denotes the relative frequency of occurrence of an experiment's outcome, when repeating the experiment.
- 2 Probability is a way to represent an individual's degree of belief in a statement.

- Randomness are not the same things!

Probability, now in plain English...

Probability is the measure of how likely an event is.

The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.

– *Pierre-Simon Laplace, A Philosophical Essay on Probabilities*

Probability is the measure of how likely an event is.

$$P(A) = \frac{|A|}{|\Omega|}$$

Probability, now in plain English...

- The **sample space**, Ω , is the collection of possible outcomes of an experiment

Example: die roll $\Omega = \{1, 2, 3, 4, 5, 6\}$

- An **event**, say A , is a subset of Ω

Example: die roll is even $A = \{2, 4, 6\}$

- An **elementary** or **simple** event is a particular result of an experiment

Example: die roll is a four, $\omega = 4$

- \emptyset is called the **null event** or the **empty set**

Probability is the measure of how likely an event is.

$$P(A) = \frac{|A|}{|\Omega|}$$

$|A|$ – The number of ways event A can occur

$|\Omega|$ – The total number of possible outcomes

$$P(A) = \frac{|A|}{|\Omega|}$$

$$\Omega = \{heads, tails\}$$

$$A = heads$$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{1}{2}$$

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

$$A = \{5, 6\}$$

$$P(A) = \frac{|A|}{|\Omega|} = \frac{2}{6} = \frac{1}{3}$$

Random variables

- **Random variable** is a function, which maps events or outcomes to real numbers

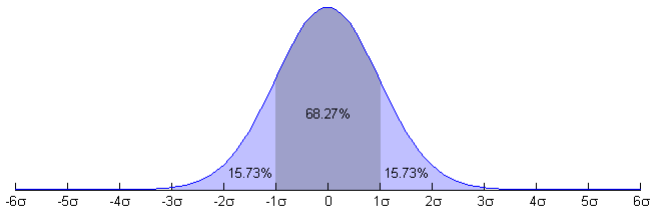
$$\xi : \Omega \rightarrow \mathbb{R}$$

- A random variable's possible values might represent the possible outcomes of a yet-to-be-performed experiment, or the potential values of a quantity whose already-existing value is uncertain. Intuitively, a random variable can be thought of as a quantity whose value is not fixed, but which can take on different values.
- Example:

$$Y(\omega) = \begin{cases} 1, & \text{if } \omega = \text{heads,} \\ 0, & \text{if } \omega = \text{tails.} \end{cases}$$

Probability density function

- **Probability density function** of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point.
- The probability for the random variable to fall within a particular region is given by the integral of this variables density over the region. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one



Mathematical Expectation

- **Mathematical expectation** of a random variable is the weighted average of all possible values that this random variable can take on. The weights used in computing this average correspond to the probabilities in case of a discrete random variable, or densities in case of a continuous random variable.
- Discrete random variable: $E[X] = \sum_i x_i p_i$
- Continuous random variable: $E[X] = \int_{-\infty}^{\infty} x p(x) dx$
- Example:
Let X represent the outcome of a roll of a six-sided dice. More specifically, X will be the number of pips showing on the top face of the die after the toss. The possible values for X are 1, 2, 3, 4, 5, 6, all equally likely. The expectation of X is

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

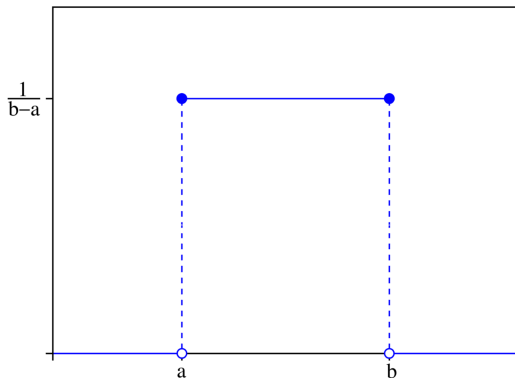
- **Variance** is a measure of how far a set of numbers are spread out from each other.
- If a random variable ξ has the **expected value** $\mu = E[\xi]$, then the variance of ξ is given by:

$$\text{Var}(\xi) = E[(\xi - \mu)^2]$$

- Discrete random variable: $\text{Var}(X) = \sum_{i=1}^n p_i \cdot (x_i - \mu)^2$, where $\mu = \sum_{i=1}^n p_i \cdot x_i$
- Let X represent the outcome of a roll of a six-sided dice. The variance of X is $\sum_{i=1}^6 \frac{1}{6} (i - 3.5)^2 = \frac{1}{6} \sum_{i=1}^6 (i - 3.5)^2 = \frac{1}{6} ((-2.5)^2 + (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 + 2.5^2) = \frac{1}{6} \cdot 17.50 = \frac{35}{12} \approx 2.92$

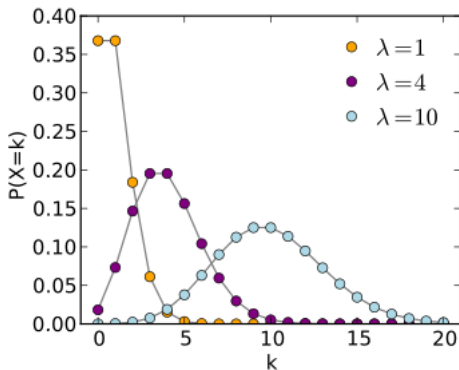
Uniform distribution

- Probability density function:
$$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$
- $E[\xi] = \frac{1}{2}(a + b)$, $\text{Var}[\xi] = \frac{1}{12}(b - a)^2$



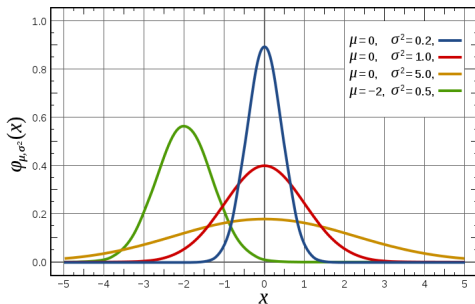
Poisson distribution

- Probability mass function: $p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- $E[X] = \lambda$, $\text{Var}[X] = \lambda$
- Expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event.



Normal distribution

- Probability density function: $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $E[\xi] = \mu$, $\text{Var}[\xi] = \sigma$



Central limit theorem

- Conditions under which the mean of a sufficiently large number of independent¹ random variables, each with finite mean and variance, will be approximately normally distributed.

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} N(0, 1), \mu = E x_i, \sigma = \text{Var } x_i$$

- Since real-world quantities are often the balanced sum of many unobserved random events, this theorem provides a partial explanation for the prevalence of the normal probability distribution.
- The CLT also justifies the approximation of large-sample statistics to the normal distribution in controlled experiments.

¹and identically distributed

Let's have a **break!**

- Statistics and Biology
- Experimental and observational studies
- Sampling error
- Correlation and Regression analysis
- Statistical hypothesis testing

- **Statistics** is the science of the collection, organization, and interpretation of data.
- It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.
- **Biostatistics** is the application of statistics to a wide range of topics in biology. The science of biostatistics encompasses the design of biological experiments, especially in medicine and agriculture; the collection, summarization, and analysis of data from those experiments; and the interpretation of, and inference from, the results.

Lies, damned lies, and **statistics**

Biostatistics is a theory and methodology for the acquisition and use of quantitative evidence in biomedical research. Biostatisticians develop innovative designs and analytic methods targeted at increasing available information, improving the relevance and validity of statistical analyses, making best use of available information and communicating relevant uncertainties.

From the Johns Hopkins Department of Biostatistics 2007 self study

Applications of biostatistics

- Public health, including epidemiology, health services research, nutrition, and environmental health
- Design and analysis of clinical trials in medicine
- Population genetics, and statistical genetics
- Analysis of genomics data
- Ecology, ecological forecasting
- Biological sequence analysis
- Systems biology for gene network inference or pathways analysis

Random variable and Samples. Statistical Population.

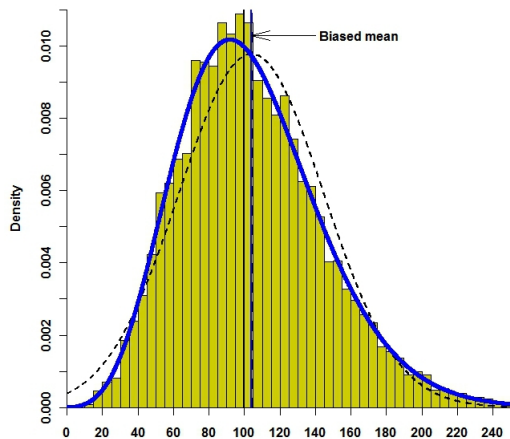
- A **statistical population** is a set of entities concerning which statistical inferences are to be drawn
- **Sample** is a subset of a population
- A sample concretely represents n experiments in which we measure the same quantity.
- Example:
We want to measure the height of our patients. In this case statistical population – set of all possible heights of patient. If X represents the height of an individual and we measure N individuals, X_i will be the height of the i -th individual.

Sample Mean, Mode, Median. Sample Variance, Standart Deviation

- **Sample Mean:** $\bar{x} = \frac{1}{N} \sum_{k=1}^N x_k$
- **Mode** is the value that occurs most frequently in a data set or a probability distribution
- **Median** is the numeric value separating the higher half of a sample from the lower half.
- **Sample Variance:** $\sigma_N^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \bar{x})^2$
- **Standart Deviation:** $\sigma_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$

Histogram

$$p_n(x) = \sum_{k=1}^N \frac{1}{|\Delta_k|} v_k I_{\Delta_k}(x), v_k = \frac{1}{n} \sum_{j=1}^n I_{\Delta_k}(X_j)$$



Sampling error, standard error

- **Sampling error** or estimation error is the error caused by observing a sample instead of the whole population.
- The likely size of the sampling error can generally be controlled by taking a large enough random sample from the population
- If the observations are collected from a random sample, statistical theory provides probabilistic estimates of the likely size of the sampling error for a particular statistic or estimator. These are often expressed in terms of its **standard error**:

$$m = \frac{\sigma}{\sqrt{N}}$$

here σ is standart deviation, N – number of observations

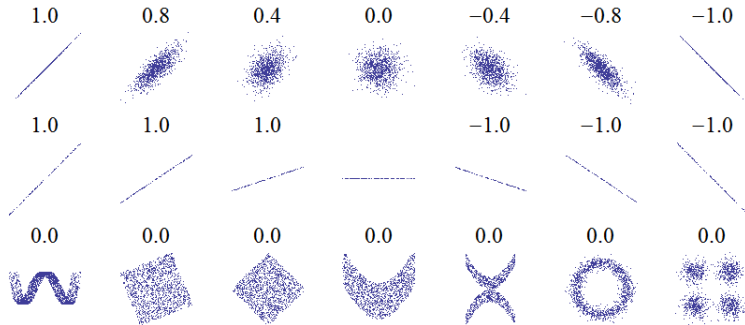
Correlation

- In statistics, **correlation** and dependence are any of a broad class of statistical relationships between two or more random variables or observed data values.
- The most familiar measure of dependence between two quantities is the **Pearson product-moment correlation coefficient**, or "Pearson's correlation."

$$\mathbb{R}_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sigma_X \cdot \sigma_Y}$$

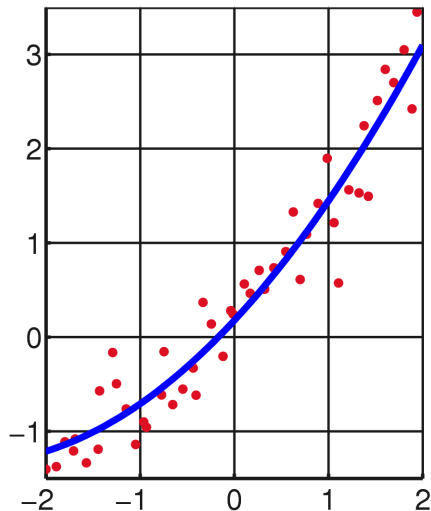
Correlation	$ \mathbb{R}_{X,Y} $
None	0.0 to 0.09
Small	0.1 to 0.3
Medium	0.3 to 0.5
Large	0.5 to 1.0

Correlation

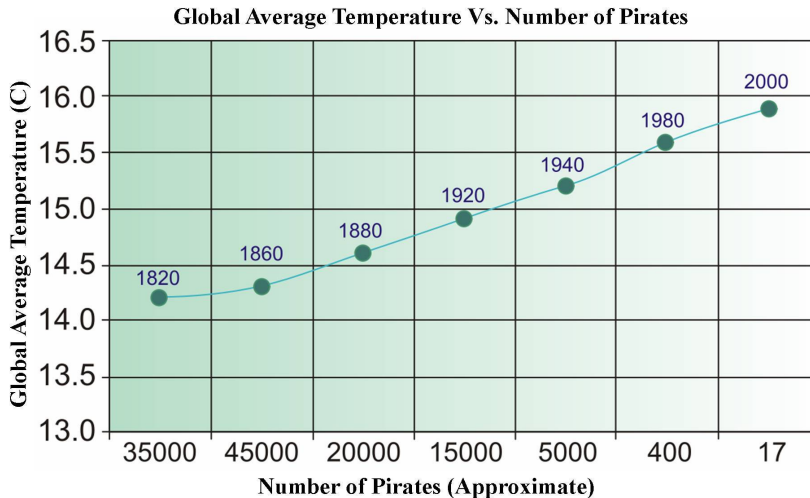


- A method for fitting a curve (not necessarily a straight line) through a set of points using some goodness-of-fit criterion.
- The most common method for goodness-of-fit criteria – **Least Squares Fitting**
- The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model.

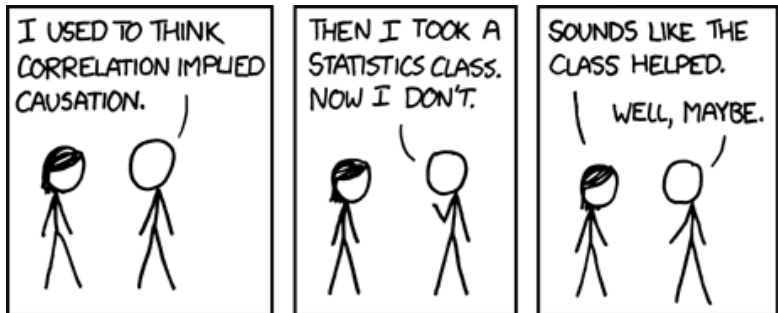
Regression analysis - example



Post hoc ergo propter hoc



Post hoc ergo propter hoc



Dixi