Probabilities and distributions

Johan A. Elkink
School of Politics & International Relations
University College Dublin

8 October 2018

Outline

Probability: definition

Frequentist approach: a probability is the proportion of times a particular event will occur given a long run of repeated experiments.

Subjective (**Bayesian**) approach: a probability is the formal quantification of the subjective belief about how likely a certain event will occur when an experiment is performed.

Probability: properties

For any event A, $0 \leq Pr(A) \leq 1$

If S contains all possible events, $Pr(S) = 1$

If B is "not A", $Pr(B) = 1 - Pr(A)$

If A and C are independent events, $Pr(AC) = Pr(A)Pr(C)$

Probability: example

$$Pr(\text{head}) = \frac{1}{2}$$

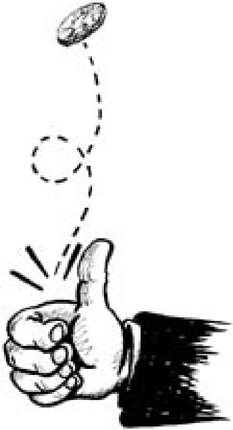
$$Pr(\text{tail}) = \frac{1}{2}$$

Sum of all possible events:

$$Pr(\text{head}) + Pr(\text{tail}) = \frac{1}{2} + \frac{1}{2} = 1$$

Probability of head and then tail:

$$Pr(\text{head})Pr(\text{tail}) = \frac{1}{4}$$

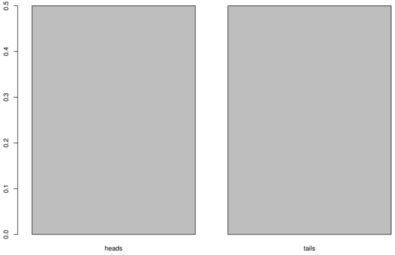


Outline

Probability mass function

For nominal or ordinal variables, the **probability distribution** looks like a relative frequency graph (histogram); "probability mass function" (p.m.f.).

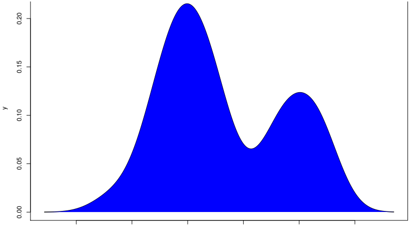
All probabilities together **sum** to 1.



Probability density function

For interval and ratio variables, the probability distribution is a continuous function, the "probability density function" (p.d.f.).

The total **area** under the graph is 1.

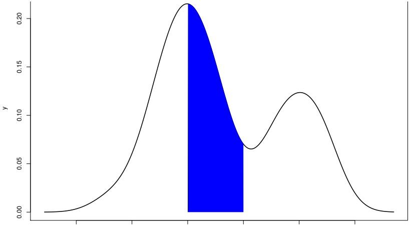


For a continuous variables, you can only speak of the probability that the observation will be within a certain **range** of values, not of the probability of a specific value. The probability is represented by the area under the graph over that range.

Probability density function

For interval and ratio variables, the probability distribution is a continuous function, the "probability density function" (p.d.f.).

The total **area** under the graph is 1.



For a continuous variables, you can only speak of the probability that the observation will be within a certain **range** of values, not of the probability of a specific value. The probability is represented by the area under the graph over that range.

Bernoulli trial

A Bernoulli distribution is a distribution with two possible outcomes (e.g. heads and tails) and a probability attached to each outcome.

A **Bernoulli trial** is "an experiment in which s trials are made of an event, with probability p of success in any given trial."

(Weisstein, Eric W. "Bernoulli Trial." `http://mathworld.wolfram.com/BernoulliTrial.html`)

Binomial distribution

The **binomial distribution** is "the (...) probability distribution (...) of obtaining exactly n successes out of N Bernoulli trials."

$$P(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

(Weisstein, Eric W. "Binomial Distribution." http://mathworld.wolfram.com/BinomialDistribution.html)

Normal distribution

$$P(n|N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

$$\lim_{N\to\infty} p(n) = \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{(n-Np)^2}{2Npq}}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(n-\bar{n})^2}{2\sigma^2}}$$

i.e. the **limiting distribution** of the binomial distribution is the **normal distribution**, with $\sigma^2 \equiv Npq$.

Also called **Gaussian distribution**, but Gauss did not invent it.

(Weisstein, Eric W. "Binomial distribution." http://mathworld.wolfram.com/binomialdistribution.html) and (Davidson and MacKinnon, 1999, 130–135)

χ^2-distribution

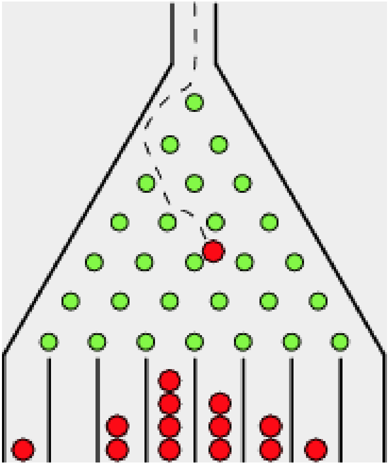
The sum of squares of r independent normal distributions, is distributed chi-squared with r degrees of freedom:

$$\chi^2(r) \equiv \sum_i^r x_i^2$$

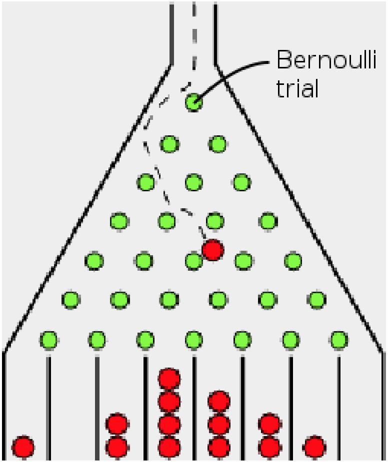
(Weisstein, Eric W. "Chi-squared distribution." http://mathworld.wolfram.com/chi-squareddistribution.html)

Outline

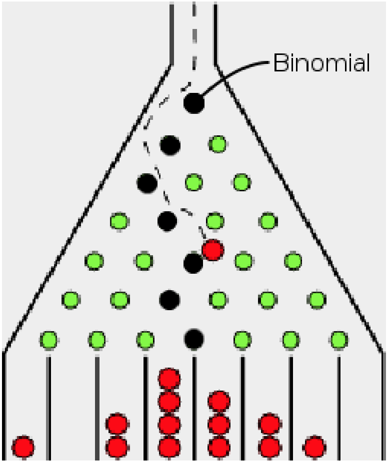
Galton board



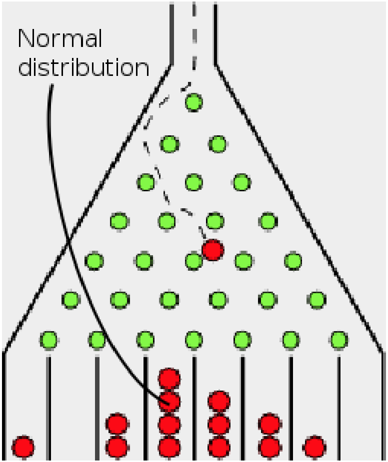
Galton board



Galton board



Galton board



Normal distribution

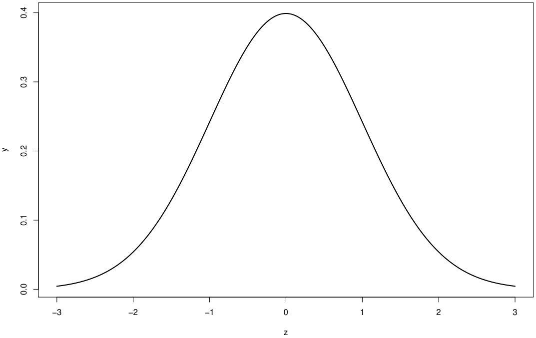


Figure: Normal distribution

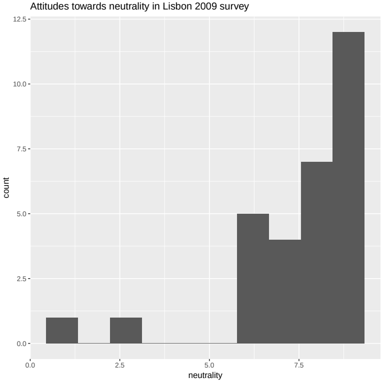
Normal distribution

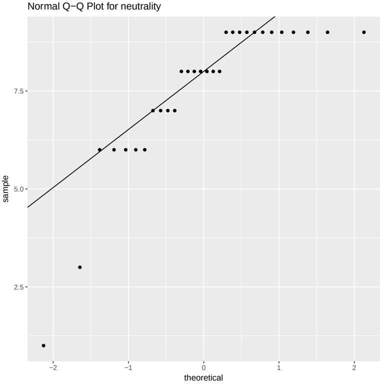
$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Thus entirely determined by the mean μ and the standard deviation σ.

Notation: $y_i \sim N(\mu, \sigma^2)$, meaning Y is normally distributed with mean μ and variance σ^2.

Check for normality





Example from the Lisbon 2009 post-referendum survey: Attitude towards Irish neutrality, on a scale from 1 "Ireland should be willing to accept limitations on its neutrality" to 9 "Ireland should do everything it can to strengthen its neutrality".

z-scores

If a variable follows a normal distribution, we can **standardize** it, converting the values of the variable to z-**scores**.

Definition: the z-**score** of a value y_i for variable Y is the number of standard deviations between y_i and the mean of Y, $\bar{y}$.

$$z_i = \frac{y_i - \bar{y}}{s_y} \qquad z_i \sim N(0, 1)$$

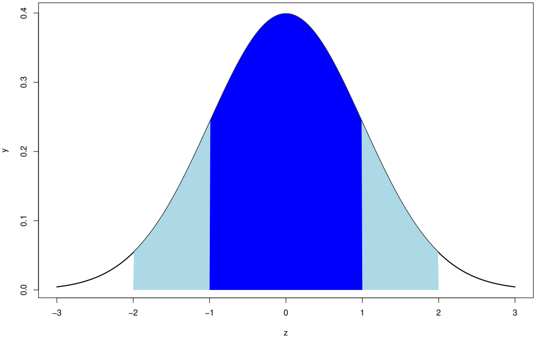
The $N(0, 1)$ distribution is called the **standard normal distribution**.

Calculating z-scores for a variable is called **standardizing** the variable.

Rules of thumb

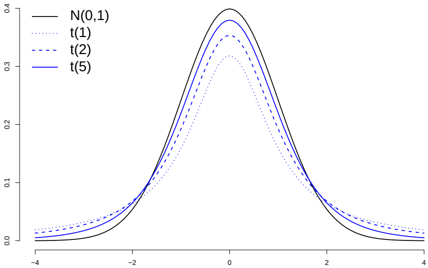
The surface under the probability distribution between $z = -1$ and $z = 1$ is approximately 68%.

Between $z = -2$ and $z = 2$ approximately 95%.



t-distribution

When σ is not known, but estimated, we use the t-distribution. When N is sufficiently large, the t-distribution approximates the normal distribution.



Outline

1. "The local police force in Shinbone, Kansas, gives all applicants an entrance exam and accepts only those applicants who score in the top 15% on this test. If the mean score this year is 87 and the standard deviation is 8, would an individual with a score of 110 be accepted?"
2. "A scale measuring prejudice has been administered to a large sample of respondents. The distribution of scores is approximately normal with a mean of 31 and a standard deviation of 5. If a score of 40 or more is considered 'highly prejudiced,' what is the probability that a person selected at random will have a score in that range?"

(Healey, 1996, 128–129)

"A 25-item scale measuring political conservatism has been administered to a sample of 500 respondents. The scores of eight respondents are listed below."

score	z-score	% below
19		
10		
14		
15		
15		
18		
20		
22		

"Assume that the distribution of a college entrance exam is normal with a mean of 500 and a standard deviation of 100."

score	z-score	% above
650		
400		
375		
586		
437		
526		
621		

(Healey, 1996, 127–128)

Davidson, Russell and James G. MacKinnon. 1999. *Econometric theory and methods*. Oxford: Oxford University Press.

Healey, Joseph F. 1996. *Statistics: a tool for social research*. Wadsworth.