Basic Statistics I

Lecture 4

Definitions

Data: The word "data" refers to the information that has been collected from an experiment, a survey, a historical record, etc.

Types of statistics:

Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand.

Inferential statistics is generalizing from our data to another set of cases. This involved mathematical procedures whereby we convert information about the sample into intelligent guesses about the population.

Populations and samples

Sample is a small subset of a larger set of data used to draw inferences about the larger set. The larger set is known as the population from which the sample is drawn.

A sample is typically a small subset of the population.

Example: You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Whom will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. This is also an example of inferential statistics.

Simple Random Sampling: Such sampling requires every member of the population to have an equal chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. Random samples, especially if the sample size is small, are not necessarily representative of the entire population.

Random Assignments: This is random division of the sample into two groups. Random assignment is critical for the validity of an experiment.

Stratified Random Sampling

In stratified random sampling: the population is divided into a number of subgroups (or strata). Random samples are then taken from each subgroup with sample sizes proportional to the size of the subgroup in the population. For instance, if a population contained equal numbers of men and women, and the variable of interest is suspected to vary by gender, one might conduct stratified random sampling to insure a representative sample

Example: A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In this example, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example: A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In this example, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer!

Sample Size: Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population.

Variable: There are dependent and independent variables. For example, if you want to correlate the number of hours students study as a function of the success rate of the students. The number of hours is the independent variable and the success in the exam is the dependent variable.

Qualitative and Quantitative variables: Qualitative variables may constitute those that cannot associated with numbers- e.g. religion, hair color etc. Quantitative variables are those that can be expressed by numbers- e.g. number of students in a class.

Discrete or Continuous variables: values such as the number of children in kindergarten are discrete variables- e.g. there are 3, 4 or 5 children but not 3.2 children. The change in temperature is a continuous variable e.g. 24.4 or 23.2 deg C.

Percentile

Definition: the 65% percentile (for example) is the lowest score that is higher than 65% of the scores.

How to estimate the percentiles: We aim to measure 25th percentile for the 8 numbers in this Table. Numbers are given ranks from 1 to 8.

1. Compute the rank ® for 25th percentile: $R = P/100 \times (N+1)$

Where P is the percentile (25) and N the total number 8.

R= 25/100 x (8+1) = 9/4 = 2.25

If R is an integer, the Pth percentile is the number with rank R. Where R is not an integer, we compute the Pth percentile with interpolation

2. Define the integer part of R (IR). Here it is IR=2. Define the fractional part of r (FR). Here it is FT=0.25.

4. Find the score with rank I_R and I_R+1. From the table this is the score with rank 2 and with rank 3. These are 5 and 7

5. Interpolate by multiplying the difference between the scores by FR and add the result to the lower score

(0.25)(7-5) + 5 = 5.5 - The 25th percentile is 5.5.

Test Score Example

Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

Distributions

A **probability distribution** is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. The probability distribution is a description of a random phenomenon in terms of the probabilities of events. For example, if random variable X denotes the probability of a coin toss, the probability distribution of X would take the value 0.5 for X=heads and 0.5 for X tails.

Probability distributions are divided into two classes: Discrete and Continuous.

Discrete probability distribution is when the set of outcomes are discrete- like tossing a coin. This can be encoded by a discrete list of the probabilities of the outcomes known as a probability mass function.

Continuous probability distribution is applicable where the set of the possible outcomes can take on values in a continuous range- like the temperature in a given day. This is described by probability density function (PDF).

A probability distribution whose sample is one-dimensional (i.e. real numbers) is called univariate while a distribution whose sample is a vector space is called multivariate. A univariate distribution gives the probability of a single random variable. A multivariate distribution (a joint probability distribution) gives the probability of a random vector- a list of two or more random vector- a list of two or more random variables taking on various combinations of values.

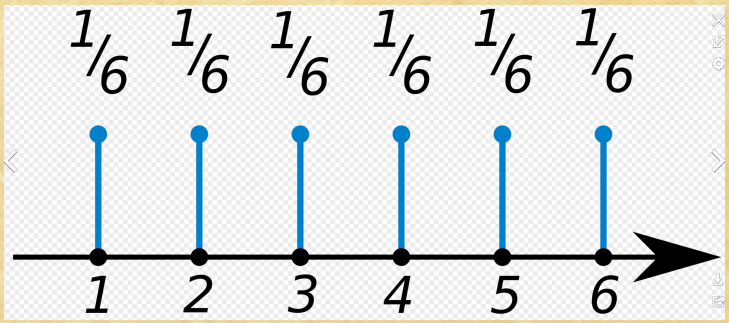
Discrete probabilities

Discrete random variables

We specify a probability mass function, p, assigning a probability to each possible outcome. When throwing a dice, each of the six values 1 to 6 has the probability 1/6. The probability of an even is them defined as the sum of the probabilities of the outcomes that satisfy the event. For example, the probability that dice rolls an even value is:

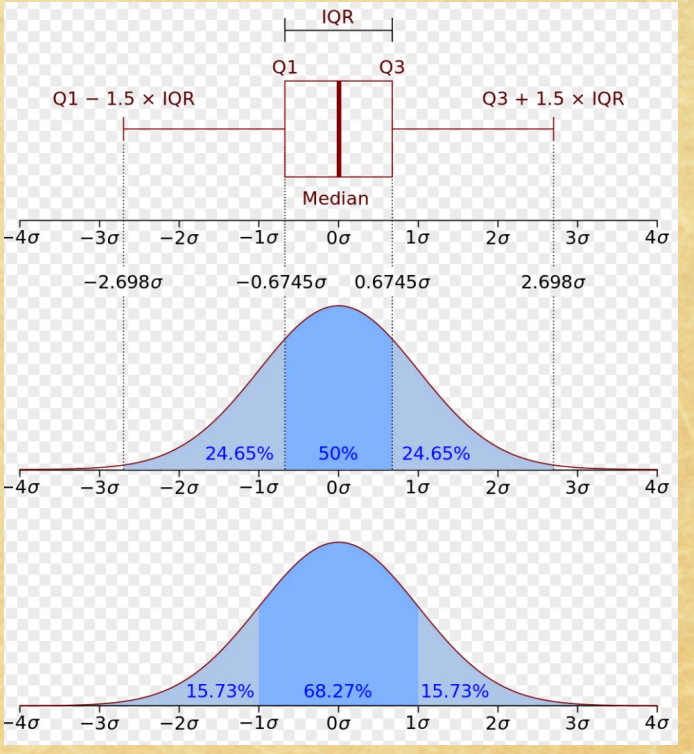
$P(2)+p(4)+p(6) = 1/6+1/6+1/6 = 1/2$

Example of throwing a dice



Probability Density Function

A probability density function (PDF) of a continuous random variable is a function whose value at any given sample (or point) in the sample space is interpreted as relative likelihood that the value of the random variable would equal that sample. While the absolute likelihood of a continuous random variable is zero (since there are an infinite set of possible values to begin with), the value of the PDF at two different samples can be used to infer how much more likely it is that the random variable would equal one sample compared to other.



Cumulative Distribution Functions

A random variable *X* with density f_x, where f_x is non-negative integrable function

$$\Pr[a \leq X \leq b] = \int_a^b f_X(x)\,dx.$$

The cumulative distribution function, F(x) is

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du,$$

And if f_x is continuous at *x*

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Unlike the probability, a probability distribution function could have values greater than one

Normal Distribution

A normal distribution function is expressed as

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

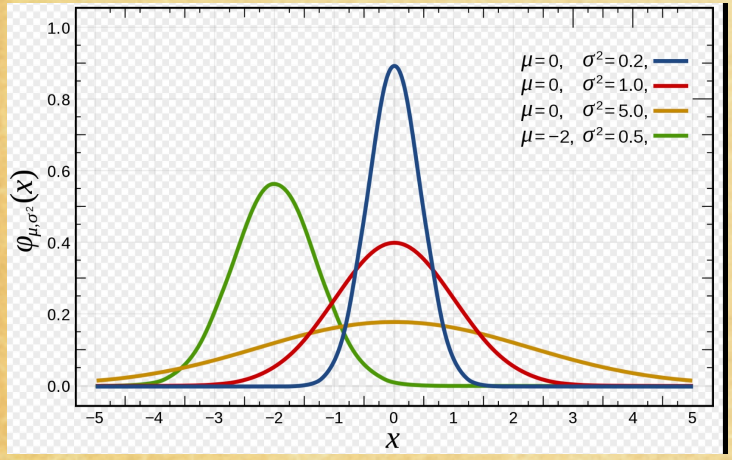
The expected value of a random variable X given its probability distribution is

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx.$$

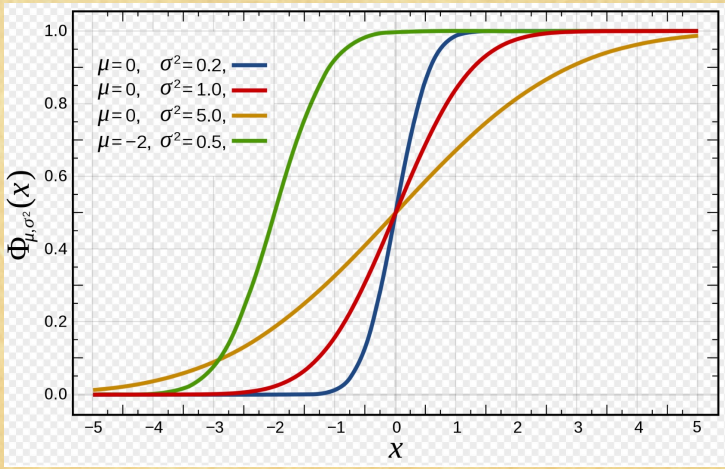
The normal distribution is parametrised in terms of the mean (μ) and variance (σ^2)

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

Normal Distributions



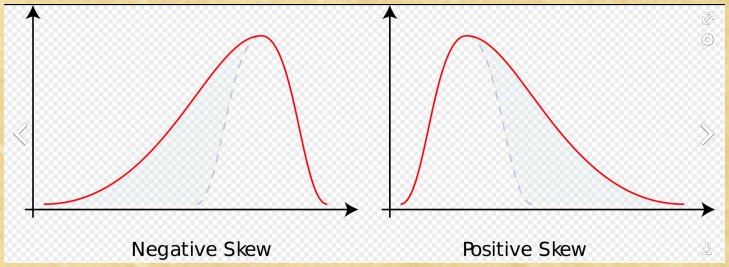
Cumulative distribution



Shapes of Distributions

Skewness

Skewness is a presentation of the asymmetry of the probability distribution about its mean. The skewness value could be positive, negative or undefined. It is estimated as:



3 (mean – median)/s + O(skewness2)

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

Kurtosis

Kurtosis is a measure of the "tailedness" of the probability distribution of a real valued random variable. It describes the shape of the distribution. Kurtosis is estimated as:

$$\operatorname{Kurt}[X] = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2},$$

Moments

A **standardized moment** of a probability distribution is a moment (of higher degree central moment) that is normalized by division by the standard deviation which makes the moment invariant. This has the advantage that such normalized moments differ only in other properties than variability, facilitating comparison of shapes of different distributions.

Let *X* be a random variable with a probability distribution *P* and mean $\mu = E[X]$ (the first moment about zero). The operator E depending the expected value of X. The standard moment of degree k is μ_k/σ_k, that is the ratio of the *kth* moment about the mean

$$\mu_k = \mathrm{E}\left[(X-\mu)^k\right] = \int_{-\infty}^{\infty} (x-\mu)^k P(x)\,dx,$$

to the *k*th power of the standard deviation,

$$\sigma^k = \left(\sqrt{\mathrm{E}[(X-\mu)^2]}\right)^k.$$

The First Four Standardaised Moments

Degree k		Comment
1	$\tilde{\mu}_1 = \frac{\mu_1}{\sigma^1} = \frac{\mathrm{E}\left[(X-\mu)^1\right]}{(\mathrm{E}[(X-\mu)^2])^{1/2}} = \frac{\mu-\mu}{\sqrt{\mathrm{E}[(X-\mu)^2]}} = 0$	The first standardized moment is zero, because the first moment about the mean is always zero.
2	$\tilde{\mu}_2 = \frac{\mu_2}{\sigma^2} = \frac{\mathrm{E}\left[(X-\mu)^2\right]}{(\mathrm{E}[(X-\mu)^2])^{2/2}} = 1$	The second standardized moment is one, because the second moment about the mean is equal to the variance σ^2.
3	$\tilde{\mu}_3 = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}[(X-\mu)^2])^{3/2}}$	The third standardized moment is a measure of skewness.
4	$\tilde{\mu}_4 = \frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}\left[(X-\mu)^4\right]}{(\mathrm{E}[(X-\mu)^2])^{4/2}}$	The fourth standardized moment refers to the kurtosis.

Properties of Normal Distributions

- It is symmetric around the point $x=\mu$, which is at the same time mode, mean and median of the distribution.
- It is unimodal. Its first derivative is positive for $x < \mu$ and negative for $x > \mu$ and zero at $x=\mu$.
- The area under the curve and along the x axis is one.
- Its density has two inflection points where the second derivative of f is zero and changes sign, located one standard deviation away from the mean, namely at $x= \mu - \sigma$ and $x = \mu + \sigma$

Binomial Distribution

The **binomial distribution** with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking yes/no question and each with outcome: success/yes/true/one (with probability p) or failure/no/false/zero with probability q=1-p. A single success/failure experiment is also called a Bernoulli trial. For a single trial (n=1), the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test. The random variable X follows the binomial distribution with parameters n ε N and p ε {0,1}, we have

X ~ B(n,p). The probability of getting exactly k successes in n trials is given by the probability mass function

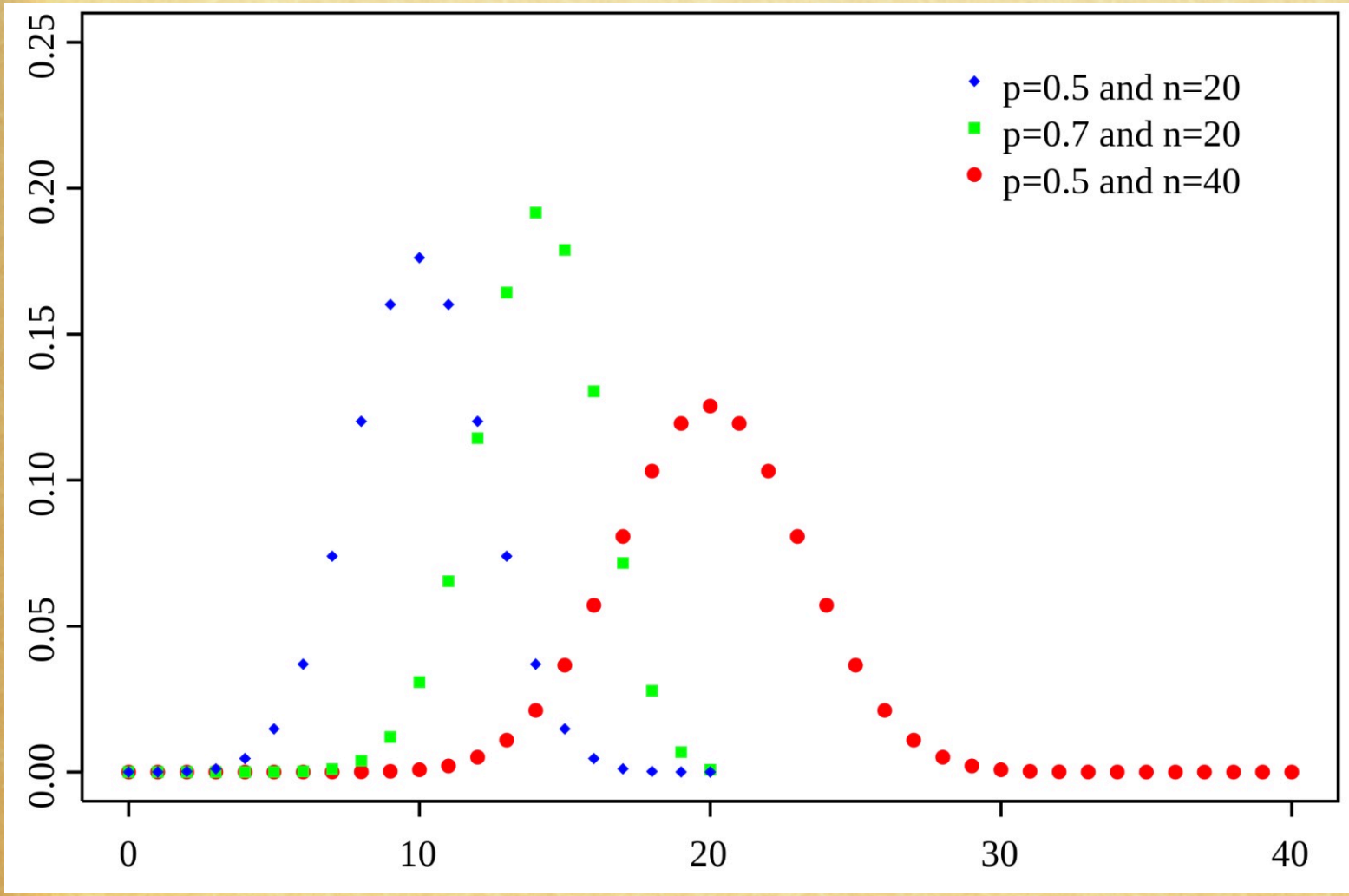
$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Is the binomial coefficient. This can be understood as follows. K successes occur with probability p_k and n-k failures occur with probability $(1-p)^{n-k}$. However, the k successes can occur anywhere among the n trials and there ar $\binom{n}{k}$ different ways of distributing k successes in n trials.

Binomial Distributions



Standard Deviation

The standard deviation of a distribution is defined as

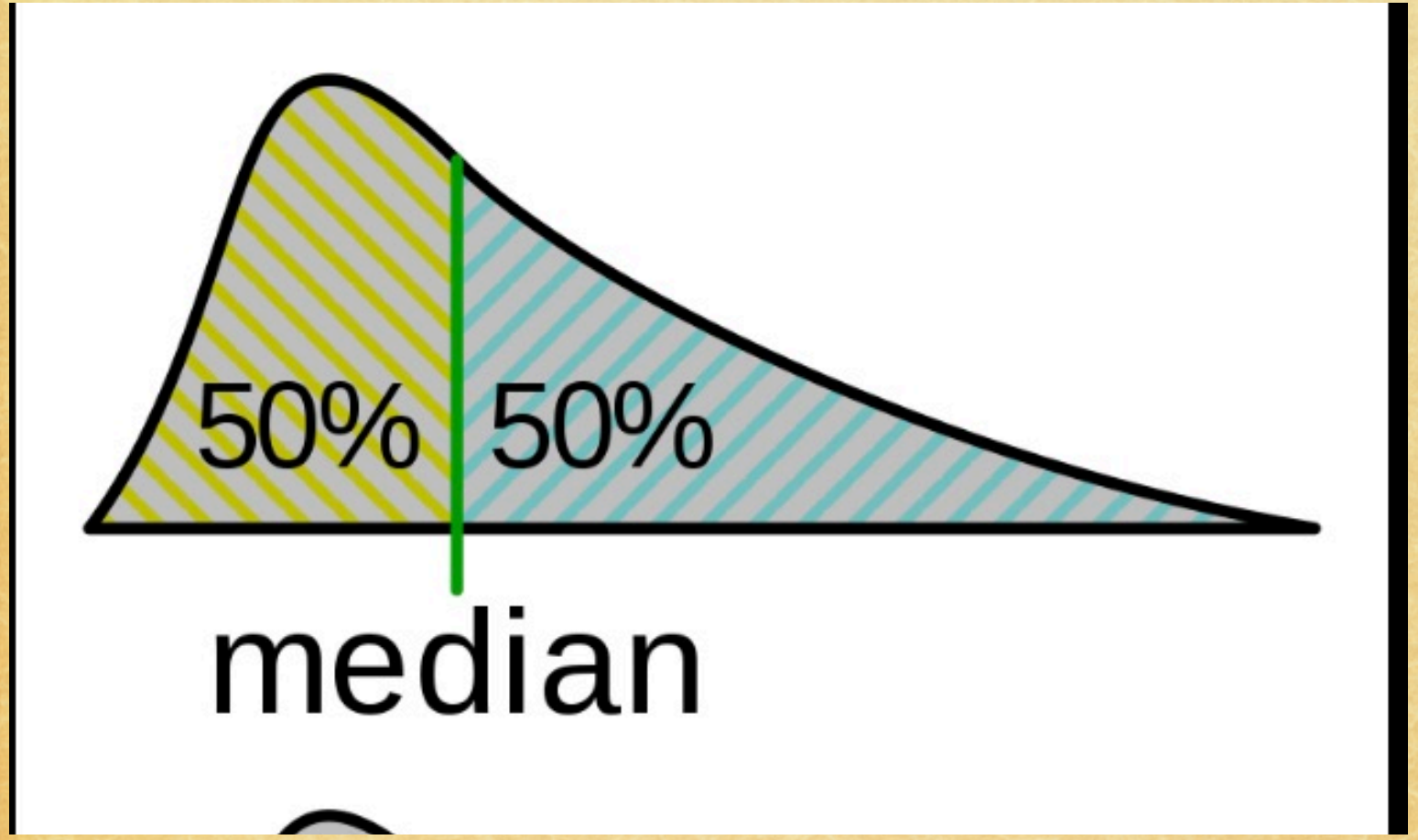
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}},$$

The mode of a sample is the element that occurs more often in the distribution. The mode of the vector {2,3,6,6,6,7,7,8,8, 9, 100} is 6.

The mean is the sum of all the elements divided by the number of the elements.

The median of a sample is, when the sample is sorted, the element in the middle.

Mode, Median and Mean



Standard Deviation

In the case where X takes random values from a finite data set $x_1, x_2, \ldots, x_N$, with each value having the same probability, the standard deviation is

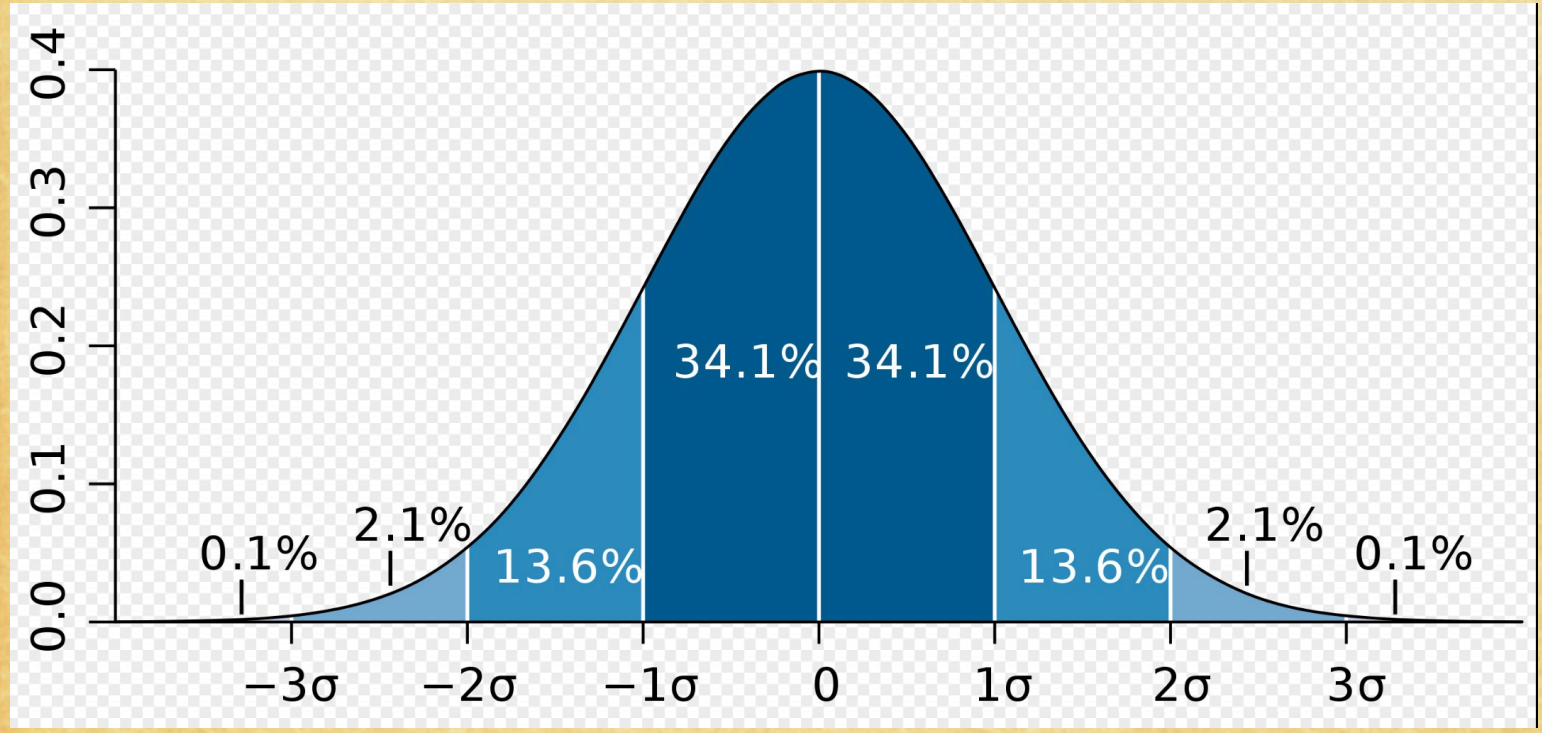
$$\sigma = \sqrt{\frac{1}{N}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2\right]}, \text{ where } \mu = \frac{1}{N}(x_1 + \cdots + x_N),$$

or, using summation notation,

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

If, instead of having equal probabilities, the values have different probabilities, let x_1 have probability p_1, x_2 have probability p_2, ..., x_N have probability p_N. In this case, the standard deviation will be

$$\sigma = \sqrt{\sum_{i=1}^{N} p_i(x_i-\mu)^2}, \text{ where } \mu = \sum_{i=1}^{N} p_i x_i.$$


0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

Regressions

This is about predictions. Knowing one independent variable, we predict other variables using defined relations. This is called regression. In simple linear regression the predictor Y is expressed as a function of criterion variable X by a linear relation. Linear regression consists of finding the best fitting straight line through these points. The straight line in the figure is the regression with the vertical lines being the errors corresponding to each point from the mean line. The best-fitting line is the line that minimizes the sum of the square errors of prediction. The formula for regression line is

$Y' = bX + A$

Where Y' is the predicted number, b is the slope of the line and A is the Y intercept. Consider the points in the Table.Mx is the mean of X, My is the Mean of Y, sX is the standard deviation of X, sY is the standard deviation of Y and r is the correlation coefiicient. The slope is:

$B = r\, s_Y/S_x$

The intercept A is

$A = M_y - bM_x$

Regression line

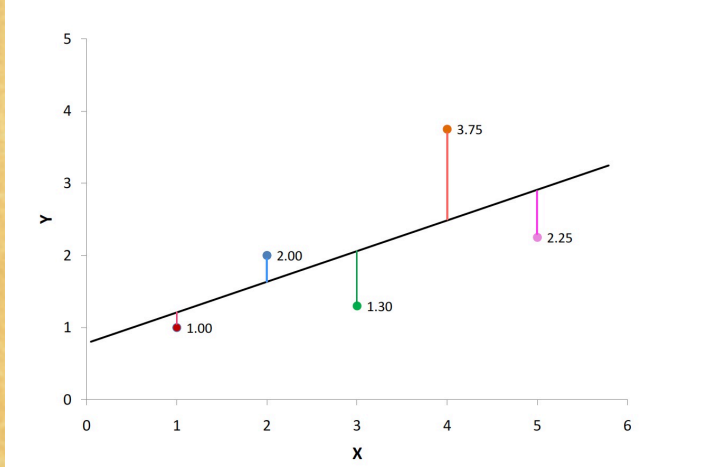


Table 2. Example data.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436