DEFINITIONS

Statistics

Collecting and summarising data, random variables, distributions and hypothesis tests.

73 definitions **7** subtopics By Lucas Potrykus

CONTENTS

1. Types of data 5
2. Averages and spread 15
3. Statistical diagrams 6
4. Populations and sampling 12
5. Random variables and distributions 19
6. Correlation and regression 6
7. Hypothesis testing 10

HOW TO USE THIS

Each definition gives the exact meaning of a word or symbol. Where there is a symbol, it is shown next to the name, with how to say it out loud. Most definitions have a short example.

Learn definitions word for word: in proofs and in exams, the precise wording is what earns the marks.

Searchable online, with links to every topic, at lemmamathematics.com/definitions.

TYPES OF DATA

More in **Data presentation and interpretation**

Qualitative and quantitative data

Qualitative (categorical) data are descriptions, such as colour; quantitative data are numbers.

Discrete and continuous data

Discrete data take separate values (often counts); continuous data can take any value in a range (measurements).

> EXAMPLE
> Number of siblings is discrete; height is continuous.

Variable (in statistics)

A characteristic that is measured or observed and varies between individuals.

Grouped data and class boundaries

Data recorded in classes rather than as exact values. The class boundaries are the exact limits of each class, and the class width is their difference.

Bivariate data

Data with two variables recorded for each individual, such as height and mass.

AVERAGES AND SPREAD

More in **Data presentation and interpretation**

Mean

The sum of the values divided by how many there are: $\bar{x} = \frac{\sum x}{n}$, or $\frac{\sum fx}{\sum f}$ from a frequency table.

Sample mean

$\bar{x}$

SAY *x bar*

The mean of a set of data: $\bar{x} = \frac{\sum x}{n}$. As a random variable, $\bar{X}$, the mean of a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, has $\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

> EXAMPLE
> The mean of $2, 4, 9$ is $\bar{x} = 5$.

Median

The middle value when the data are in order (the mean of the middle two if n is even). It is not affected by extreme values.

Mode

The most common value (or the modal class, for grouped data).

Range

Largest value minus smallest value: a simple, outlier-sensitive measure of spread.

Quartiles

Q_1, Q_2, Q_3

SAY *lower quartile; median; upper quartile*

The values a quarter, half and three quarters of the way through ordered data. The interquartile range is $Q_3 - Q_1$; a common rule calls a value an outlier if it is more than $1.5\times$ IQR beyond a quartile.

Interquartile range

$Q_3 - Q_1$

SAY *Q three minus Q one*

$Q_3 - Q_1$, the spread of the middle half of the data. Not affected by extreme values.

Percentile

The value below which a given percentage of the data lies. The median is the 50th percentile.

Summary statistics

S_{xx}, S_{xy}

SAY *S x x; S x y*

$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$ and $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$. They feed into variance, correlation and regression.

> EXAMPLE
> For $1, 2, 3$: $S_{xx} = 14 - \frac{36}{3} = 2$.

Sample standard deviation and variance

s, s^2

SAY *s; s squared*

The spread of a sample. Books differ: some divide S_{xx} by n, others by $n - 1$ (which gives an unbiased estimate of σ^2). Check which one is meant.

> EXAMPLE
> For $1, 2, 3$: dividing by $n - 1$ gives $s^2 = 1$.

Standard deviation

σ, s

SAY *sigma; s*

The square root of the variance: roughly the typical distance of a value from the mean, in the same units as the data.

Coding

Transforming data with $y = \frac{x-a}{b}$ to simplify calculations. Then $\bar{x} = a + b\bar{y}$ and $\sigma_x = b\sigma_y$: adding a constant does not change the spread.

Linear interpolation (grouped data)

Estimating the median or a quartile of grouped data by assuming the values are spread evenly through each class.

Outlier

A value that is unusually far from the rest. A common rule: more than $1.5 \times \text{IQR}$ below Q_1 or above Q_3, or more than 2 or 3 standard deviations from the mean.

Skewness

Lack of symmetry. Positive skew has a long tail to the right (typically mean > median > mode); negative skew a long tail to the left.

STATISTICAL DIAGRAMS

More in **Data presentation and interpretation**

Histogram

A diagram for continuous grouped data in which the **area** of each bar is proportional to the frequency. With unequal classes, the height is the frequency density.

Frequency density

Frequency divided by class width: the height of a histogram bar.

Box plot

A diagram showing the minimum, Q_1, median, Q_3 and maximum, with outliers often marked separately.

Cumulative frequency

The running total of the frequencies. A cumulative frequency graph plots it against the upper class boundaries and is used to estimate the median and quartiles.

Stem-and-leaf diagram

A display that keeps every value: the leading digits form the stem and the last digit is a leaf. It needs a key.

Scatter diagram

A plot of bivariate data as points, used to see whether the variables are related.

POPULATIONS AND SAMPLING

More in **Statistical sampling**

Population

The whole set of items or individuals of interest.

Sample

A selection from the population, used to draw conclusions about it.

Census

Observing or measuring every member of the population. Accurate, but expensive and impossible if testing destroys the item.

Sampling unit and sampling frame

A sampling unit is a single member of the population; the sampling frame is the list of all sampling units, from which the sample is chosen.

Simple random sampling

Every sample of size n has the same chance of being chosen, for instance by numbering the frame and using random numbers.

Systematic sampling

Choosing every kth item from an ordered list, starting from a random point among the first k.

Stratified sampling

Splitting the population into groups (strata) and sampling randomly from each in proportion to its size.

Quota sampling

An interviewer selects people to fill quotas for each group. Not random, so open to bias, but quick and needs no frame.

Opportunity sampling

Taking whoever is available at the time. Easy and cheap, but unlikely to be representative.

Cluster sampling

Dividing the population into naturally occurring groups (clusters), choosing some clusters at random, and sampling within them.

Bias

A systematic tendency for a sample or method to give results that are wrong in the same direction.

Statistic and parameter

A statistic is a quantity calculated from a sample (such as $\bar{x}$); a parameter is the corresponding value for the population (such as μ). Statistics estimate parameters.

RANDOM VARIABLES AND DISTRIBUTIONS

More in **Statistical distributions**

Random variable

A quantity whose value depends on chance, written with a capital letter such as X. Its particular values are written with lower case, x.

Probability distribution

A description of all the values a random variable can take and their probabilities. For a discrete variable, $\sum P(X = x) = 1$.

Has the distribution

$X \sim$

SAY *X is distributed as; X follows*

Says which probability distribution the random variable X follows. Capital X is the random variable; lower case x is a particular value it takes.

> EXAMPLE
> $X \sim \mathrm{B}(10, 0.3)$; then $\mathrm{P}(X = x)$ is the probability that X takes the value x.

Expected value (mean)

$\mathrm{E}(X)$

SAY *E of X; the expectation of X*

The long-run average value of X: $\mathrm{E}(X) = \sum x\,\mathrm{P}(X = x)$. Also written μ. Expectation is linear: $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

> EXAMPLE
> For a fair die, $\mathrm{E}(X) = 3.5$.

Variance

$\mathrm{Var}(X)$

SAY *var X; the variance of X*

The average squared distance from the mean: $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$. Also written σ^2. Note $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$.

> EXAMPLE
> For a fair die, $\mathrm{Var}(X) = \frac{91}{6} - 3.5^2 = \frac{35}{12}$.

Population mean and standard deviation

μ, σ

SAY *mu; sigma*

Greek letters describe the **population** (usually unknown); Roman letters such as $\bar{x}$ and s describe a **sample** you have measured.

More in **Data presentation and interpretation**

Cumulative distribution function

$F(x)$

$F(x) = P(X \leq x)$.

Probability density and cumulative distribution

$f(x)$, $F(x)$

SAY *f of x; capital F of x*

For a continuous random variable, the area under the density $f(x)$ gives probabilities, and $F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$.

> EXAMPLE
> For X uniform on $[0, 2]$: $f(x) = \frac{1}{2}$ and $F(x) = \frac{x}{2}$ for $0 \leq x \leq 2$.

Discrete uniform distribution

Each of n values is equally likely, with probability $\frac{1}{n}$.

> EXAMPLE
> The score on a fair dice.

Binomial distribution

$\mathrm{B}(n, p)$

SAY *B n p; binomial n p*

The number of successes in n independent trials, each with probability p of success. Mean np, variance $np(1-p)$.

> EXAMPLE
> If $X \sim \mathrm{B}(10, 0.3)$ then $\mathrm{E}(X) = 3$ and $\mathrm{Var}(X) = 2.1$.

Conditions for a binomial model

A fixed number of trials n, each with two outcomes (success or failure), a constant probability of success p, and independent trials.

Poisson distribution

$\mathrm{Po}(\lambda)$

SAY *Po lambda; Poisson lambda*

The number of events in a fixed interval when they happen independently at a constant average rate λ. Its mean and variance are both λ.

> EXAMPLE
> Calls arrive at 4 per hour on average: $X \sim \mathrm{Po}(4)$, $\mathrm{P}(X = 0) = \mathrm{e}^{-4}$.

Geometric distribution

$\mathrm{Geo}(p)$

The number of trials up to and including the first success, when each trial succeeds with probability p: $P(X = r) = (1-p)^{r-1}p$.

Normal distribution

$\mathrm{N}(\mu, \sigma^2)$

SAY *N mu sigma squared; normal*

The bell-shaped distribution with mean μ and **variance** σ^2. The second number is the variance, not the standard deviation: a classic exam trap.

> EXAMPLE
> If $X \sim \mathrm{N}(50, 16)$, the standard deviation is 4.

Standard normal distribution

Z

$Z \sim N(0, 1)$. Any normal variable can be standardised with $z = \frac{x-\mu}{\sigma}$.

Standard normal and its cumulative distribution

Z, $\Phi(z)$

SAY *Z; capital phi of z ("fie")*

$Z \sim \mathrm{N}(0, 1)$, and $\Phi(z) = \mathrm{P}(Z \leq z)$. Any normal variable can be standardised: $Z = \frac{X-\mu}{\sigma}$.

> EXAMPLE
> $\Phi(1.96) \approx 0.975$ and $\Phi(-z) = 1 - \Phi(z)$.

Continuity correction

Adjusting by 0.5 when a continuous distribution approximates a discrete one, such as $P(X \leq 10) \approx P(Y < 10.5)$.

Normal approximation to the binomial

If n is large and p is close to $\frac{1}{2}$, $B(n, p) \approx N(np, np(1-p))$, with a continuity correction.

Central limit theorem

For a large sample from any distribution with mean μ and variance σ^2, the sample mean is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$.

CORRELATION AND REGRESSION

More in **Data presentation and interpretation**

Correlation

A linear relationship between two variables: positive if they tend to increase together, negative if one tends to decrease as the other increases.

Correlation coefficient

r, ρ

SAY *r; rho ("row")*

The product moment correlation coefficient, $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$, measures linear correlation in a sample, with $-1 \leq r \leq 1$.

Its population version is ρ.

> EXAMPLE
> $r = 0.95$ means strong positive linear correlation, not that one variable causes the other.

More in **Hypothesis testing**

Correlation and causation

Correlation does not imply causation: two variables can be correlated because of a third factor, or by chance.

Explanatory and response variables

The explanatory (independent) variable is the one set or measured first; the response (dependent) variable is the one that may depend on it.

Regression line

The line of best fit $y = a + bx$ found by least squares; it is used to predict y from x. b is the change in y for each unit increase in x.

Interpolation and extrapolation

Interpolation predicts within the range of the data and is usually reliable; extrapolation predicts outside it and may not be.

HYPOTHESIS TESTING

More in **Hypothesis testing**

Hypothesis test

A procedure for deciding, from sample data, whether there is enough evidence to reject a claim about a population parameter.

Null and alternative hypotheses

H_0, H_1

SAY *H nought; H one*

H_0 is the default claim you assume is true; H_1 is what you look for evidence of. Both are statements about a population parameter, never about the sample.

> EXAMPLE
> $H_0 : p = 0.3$, $H_1 : p > 0.3$ is a one-tailed test.

One-tailed and two-tailed tests

A one-tailed test has H_1 in one direction ($p > 0.3$); a two-tailed test has $H_1 : p \neq 0.3$, and the significance level is split between the two tails.

Test statistic

The value calculated from the sample that is used to decide the outcome of the test.

Significance level

α

SAY *alpha*

The probability of rejecting H_0 when it is actually true, fixed before the test (often 5%). The critical region is chosen so that this probability is at most α.

Critical region

The set of values of the test statistic that lead to rejecting H_0. Its boundary is the **critical value**; the rest is the acceptance region.

p-value

The probability, assuming H_0 is true, of a result at least as extreme as the one observed. Reject H_0 if it is less than the significance level.

Actual significance level

For a discrete distribution, the true probability of the critical region under H_0, which is usually a little below the nominal level.

Type I and Type II errors

A Type I error rejects H_0 when it is true (probability = the significance level); a Type II error fails to reject H_0 when it is false.

Conclusion in context

The last line of a test, stated in words about the situation and without certainty: "there is (in)sufficient evidence to suggest …".

Spotted a mistake, or a definition that should be here? Email lpotry31@gmail.com. More at lemmamathematics.com.