Overview: Statistical Notation

Flora Green

Summary

An overview of some of the notation you may encounter whilst studying statistics.

How to use

This is a concise reference guide for statistics notation. It contains the following sections:

- Types of data
- Data collection
- Notation relating to the entire population vs. a sample
- Data analysis
- Probability
- Hypothesis testing
- Distributions
- Set notation

Types of data

The types of data which you might encounter can be split into two categories: **quantitative data** and **qualitative data**.

Name	Definition	Example
Qualitative data	Data which is expressed with words.	Eye colour.

Name	Definition	Example
Quantitative data	Data which can be expressed numerically.	Distance.

Types of qualitative data

Name	Definition	Example
Nominal data	Data which can be categorized.	Colour.
Ordinal data	Data which can be ordered.	Rankings on a scale of "strongly disagree" to "strongly agree".

Types of quantitative data

Name	Definition	Example
Discrete data	Data which can be counted, because each possible value which the data could take is distinct.	The number of items in a stack.
Continuous data	Data which is measured within a range, rather than counted.	Height.

Data collection

These terms are often used in connection with data collection.

Name	Definition
Explanatory / independent variable	Something which you observe and measure.

Name	Definition
Response variable	Something which changes because of changes within the explanatory variable(s).

Notation relating to the entire population vs. a sample

This table distinguishes the symbols used for mean and variance based on the data from which they are calculated.

Symbol	Population	Sample
Mean	μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$

Data analysis

This table describes some terms which may be useful when analyzing a set of data.

Name	Value	Definition
Expected value	$\mathbb{E}[X] = \mu$	The average result of the random variable X.

Name	Value	Definition
Population variance	$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$	The average squared difference from the population mean. It is a measure of the spread of the data.
Sample variance	$s^2 = Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$	The average squared difference from the sample mean. It is an unbiased estimate of the population variance.

Name	Value	Definition
Population covariance	$\operatorname{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)}{n}$	For a data set of size n, this is the average difference from each data point and its mean. It measures the strength of the correlation between two random variables.

Name	Value	Definition
Sample covariance	$\operatorname{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{n-1}$	For a set of n observations, this is the average difference from each data point and its mean. It measures the strength of the correlation between two random variables.
Sum of squared deviations	$SS_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$	The sum of the squared differences between each data point and its mean.

The expected value can be calculated in the following ways:

Formula	Type of data
$\mathbb{E}[X] = \sum_{i=1}^{n} x_i P(X = x_i)$	Discrete.
$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$	Continuous.

Probability

These terms are often used in connection with probability.

Name	Value	Definition
Sample space	S	The set of all possible outcomes you could observe.
Random variable	X	A variable whose possible values are the outcomes of some trial.
Observed value	x	A real-life outcome of a trial.
	$\mathbb{P}(X = x)$	The probability that the random variable X is equal to a possible value x.
	$\mathbb{P}(X \leq x)$	The probability that the random variable X is less than or equal to a possible value x.

Name	Value	Definition
	$\mathbb{P}(X \geq x)$	The probability that the random variable X is greater than or equal to a possible value x.

Hypothesis testing

These terms are often used in connection with hypothesis tests.

Name	Value	Definition
Distributed	$\sim$	The random variable follows the given distribution.
Null hypothesis	H_0	The hypothesis which is assumed to be correct during a hypothesis test.
Alternative hypothesis	H_1	The hypothesis which you are testing during a hypothesis test.
Significance level	α	The probability at which you would reject the null hypothesis H_0 during a hypothesis test.

> **Tip**
>
> You will see these symbols when working with probability distributions, like $X \sim Bin(n, p)$ and $\mathbb{P}(X = x)$.

Distributions

This table defines some functions which are often associated with distributions.

Given a random variable X,

Name	Value	Definition	Discrete	Continuous
Support		The set of values across which X is defined.		
Probability mass function (PMF)	$f_X(x) = \mathbb{P}(a \leq x \leq b) = \frac{d}{dx}(F_X(x))$	A function which can be used to find the probability that a random variable X is equal to a given number, x.		

Name	Value	Definition	Discrete	Continuous
Probability density function (PDF)	$f_X(x) = \mathbb{P}(X = x)$	A function which can be used to find the probability that a random variable X is equal to a given number, x.		
Cumulative distribution function (CDF)	$F_X(x) = \mathbb{P}(X \leq x)$	A function which can be used to find the probability that a random variable X is less than or equal to a given number, x.		

Name	Value	Definition	Discrete	Continuous
Probability generating function (PGF)	$G_X(s) = \mathbb{E}(s^X)$	A power function where the coefficient of s^i corresponds to the probability of that x_i value occurring.		
Moment generating function (MGF)	$M_X(t) = \mathbb{E}(e^{tX})$			

Set notation

Name	Symbol	Example	Definition
Set			A collection of elements.
Empty set	ϕ		A set with no elements.
Element of	$\in$	$a \in A$	a is an element within the set A.
Union	$\cup$	$A \cup B$	The set of elements which are in A or B.
Intersection	$\cap$	$A \cap B$	The set of elements which are in A and B.

Name	Symbol	Example	Definition
Subset	$\subseteq$	$A \subseteq B$	Set A contains fewer elements than set B and all of the elements in set A are also in set B.
Proper subset	$\subset$	$A \subset B$	All of the elements in set A are also in set B and sets A and B are not equal.
Superset	$\supset$	$B \supset A$	Set B contains more elements than set A, but all of the elements in set A are also in set B.
Difference	$\setminus$	$B \setminus A$	The set of all elements which are within set B but not set A.

Further reading

For more information on probability, please see Guide: Introduction to probability.

For more information on hypothesis test, please see Guide: Introduction to hypothesis testing.

For more information on mean, expected value, variance, and standard deviation, please see Guide: Expected value, variance, standard deviation.

For more information on PMFs, PDFs, and CDFs, please see Guide: PMFs, PDFs, and CDFs.

Version history

v1.0: initial version created 12/25 by Flora Green as part of a University of St Andrews VIP project.