

Introduction to Sampling & Hypothesis Testing

John Pinney

October 2019

Graduate School, Imperial College London

Random Variables

Distribution functions

Discrete vs continuous

Expectation & variance

Sampling

Sampling distributions

Sampling methods

Parameter Estimation

Confidence intervals

Hypothesis Testing

The testing process

Parametric tests

Type I / type II errors

Random Variables

Random Variables — Definitions

In statistics, a **random variable** is any variable whose value depends on some random phenomenon.

Examples

coin toss, dice roll, choosing a card from a deck,
time of a radioactive decay, ...

The set of all possible outcomes of a random variable is called the **sample space**, Ω , for that variable.

A **probability distribution** describes the probability of each possible outcome in the sample space.

Random Variables — Discrete vs continuous

A **discrete random variable** can take only a finite set of values:

Example

rolling 2 dice: $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

A **continuous random variable** can take an infinite number of values within a given interval (or set of intervals).

Example

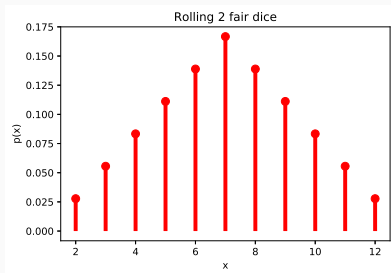
angle of a spinner, in degrees: $\Omega = [0, 360)$

Random Variables — pmf

The probability distribution, $p(x)$, for a discrete random variable X is called a **probability mass function**, pmf.

Example

rolling 2 fair dice



The pmf gives the probability of a particular outcome:

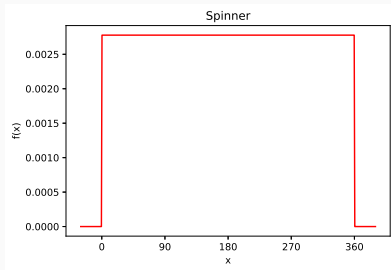
$$\mathbb{P}(X = x) = p(x)$$

Random Variables — pdf

The probability distribution, $f(x)$, for a continuous random variable X is called a **probability density function**, pdf.

Example

angle of a spinner



Integrating the pdf between two values gives the probability of an outcome within that interval:

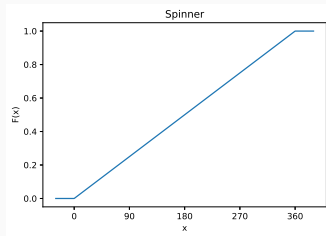
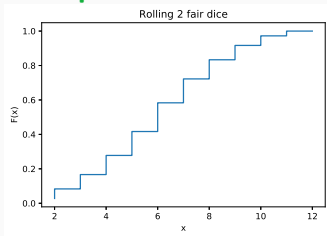
$$\mathbb{P}(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx$$

Random Variables — cdf

For convenience in calculating probabilities over intervals, we define the **cumulative distribution function**, cdf, as

$$F(x) = \mathbb{P}(X \leq x)$$

Examples



Random Variables — Expectation

The **expected value**, $\mathbb{E}X$, of a random variable X is the probability-weighted average of all its possible values, also known as its **mean**, μ .

$$\mu = \mathbb{E}X = \begin{cases} \sum_x xp(x), & X \text{ discrete.} \\ \int xf(x)dx, & X \text{ continuous.} \end{cases}$$

Examples

rolling 2 fair dice: $\mathbb{E}(X) = 7$

spinner: $\mathbb{E}(X) = 180$

Random Variables — Variance

We can also describe the degree to which the values taken by X are spread out from the mean.

The **variance**, $\text{Var}X$ of a random variable X is defined as

$$\text{Var}X = \mathbb{E}(X - \mu)^2$$

The **standard deviation**, σ is the square root of the variance.

Together, the mean and standard deviation of a random variable give us a simple summary of its distribution.

Random Variables — Examples in Python

See the python notebook
`random_variables.ipynb`
for some examples of commonly encountered discrete and
continuous random variables.

Sampling

Sampling — Sample statistics

When we take a **finite sample** of size n from a random variable, the distribution of the sample is *not* the same as the underlying theoretical distribution.

We use \bar{x} and s^2 to represent the sample mean and variance:

$$\bar{x} = \frac{\sum x}{n}$$

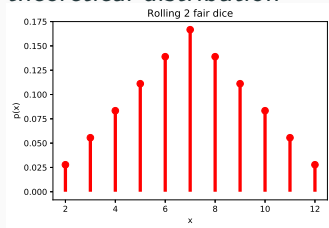
$$s^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Sampling — Sample statistics

Example

rolling 2 fair dice

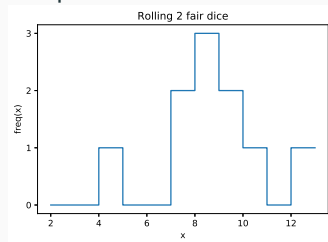
theoretical distribution



$$\mu = 7$$

$$\sigma^2 = 5.83$$

sample distribution over 10 trials



$$\bar{x} = 8.2$$

$$s^2 = 3.96$$

Sampling — Law of large numbers

The **law of large numbers** states that as we take larger and larger samples of a random variable, the sample mean gets closer to the theoretical (or *population*) mean, μ .

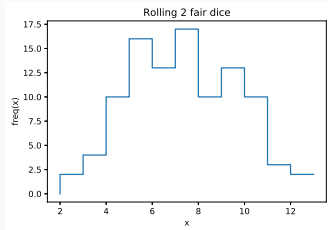
This also implies that the sample variance s^2 approaches the population variance σ^2 as n increases.

Sampling — Law of large numbers

Example

rolling 2 fair dice

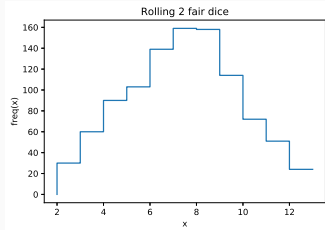
sample distribution: 100 trials



$$\bar{x} = 6.87$$

$$s^2 = 5.29$$

sample distribution: 1000 trials



$$\bar{x} = 6.92$$

$$s^2 = 5.74$$

Sampling — Sampling distribution of the mean

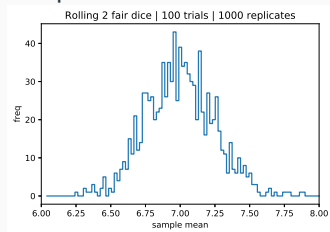
\bar{x} becomes a more precise estimate of μ as we gather more data.

We can see this by repeating the sampling process many times and plotting histograms of \bar{x} .

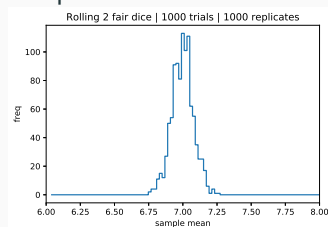
Example

rolling 2 fair dice — 1000 replicates

sample mean: 100 trials



sample mean: 1000 trials



Sampling — Sampling distribution of the mean

When we only have access to a finite sample of size n , it is helpful to know how precise our estimate of the population mean will be.

The observed sample mean, \bar{x} behaves as if it is drawn from a continuous random variable \bar{X} with mean μ and a variance that decreases as n increases.

\bar{X} is called the **sampling distribution of the mean**.

Sampling — Central limit theorem

For a sample of size n , the **central limit theorem** states that \bar{X} *converges* to a **normal distribution**:

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ for large } n$$

Note that this is true *regardless* of the distribution of X itself.

The central limit theorem is the theoretical justification for many statistical procedures.

Sampling — Standard deviation vs. standard error

The **population standard deviation**: σ (unknown)

The **sample standard deviation**: s (calculated from observed data)

The **standard error of the mean**: $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ for large n

Sampling — Unbiased estimator for population variance

When n is small (say $n < 75$), the sample variance s^2 is not a good approximation for the population variance.

In fact, it is a *biased estimator*, which tends to consistently under-predict the value of σ^2 .

We can improve our estimate by using the **unbiased sample variance**:

$$s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

In many practical applications, the population of interest is not infinite, just very large (e.g. the population of the UK).

There are a variety of ways to try to obtain a **representative** sample of a finite population, so that the conclusions from the sample are generalisable to the population as a whole.

Sampling — Random sampling

Simple random: Each individual is chosen randomly and entirely by chance.

Systematic: Every k th individual is sampled from an ordered list.

Stratified: Partition population into heterogenous subpopulations and draw a sample from each one.

Cluster: Total population is split into homogenous clusters, and a subset of clusters is sampled.

Sampling — Non-random sampling

Quota: Interviewers told to sample a certain number of a targeted population.

Convenience: The sample is drawn from the most accessible part of the population.

Snowball: Existing study subjects recruit future subjects from their acquaintances.

Voluntary: Study subjects are self-selected.

Parameter Estimation

Parameter Estimation — Point estimates

We have seen how to derive an estimated mean and variance for a population, based on a sample.

$$\hat{\mu} = \bar{x} = \frac{\sum x}{n}$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$$

These are examples of **point estimates**, where we quote a single value for a population parameter without an associated uncertainty.

Parameter Estimation — Confidence intervals

However, it is often more helpful to be able to give a plausible range of values for a parameter, based on the data collected. This is known as a **confidence interval**.

The python notebook
`confidence_intervals.ipynb`
shows how the central limit theorem can be used to derive confidence intervals for the mean of a population.

Hypothesis Testing

Hypothesis Testing — Terminology

The **null hypothesis** H_0 and **alternative hypothesis** H_1 are always two *rival* hypothesis, e.g.

$$H_0: \mu = 0;$$

$$H_1: \mu \neq 0$$

Test statistic: A quantity derived from the sample, used in hypothesis testing.

Hypothesis Testing — Terminology

P-value: The probability of obtaining an observation as extreme or more extreme than the test statistic, assuming that the null hypothesis is true.

e.g. $p = 0.03$

The smaller the p-value is, the more unlikely the observation would be to occur if H_0 were true.

Hypothesis Testing — Terminology

The **significance level** α is how we assess the p-value, and it must be selected in advance of the hypothesis test. We will reject H_0 when $p < \alpha$. From the definition of the p-value, α is the probability of incorrectly rejecting H_0 if it is true. By choosing a smaller α , we can specify a more conservative test.

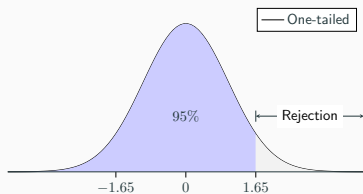
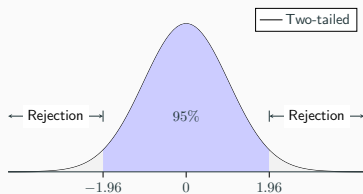
e.g. If $\alpha = 0.05$, $p = 0.03 < \alpha$, reject H_0

Hypothesis Testing — Procedure

- 1 Propose a research question
- 2 Formulate null hypothesis H_0 and alternative hypothesis H_1
- 3 Calculate the test statistic
- 4 Find the p-value
- 5 Compare p-value with the required significance level α
- 6 Make a decision about whether to reject H_0

Hypothesis Testing — One-tailed vs two-tailed

- Two-tailed test: $H_0: \mu = 0, H_1: \mu \neq 0$
 - For z-test with a given significance level $\alpha = 0.05$, H_0 is rejected when $P(|X| > z_{\alpha/2}) < \alpha/2$, where $z_{\alpha/2} = 1.96$
- One-tailed test: $H_0: \mu = 0, H_1: \mu > 0$
 - For z-test with a given significance level $\alpha = 0.05$, H_0 is rejected when $P(X > z_{\alpha}) < \alpha$, where $z_{\alpha} = 1.65$



Hypothesis Testing — Parametric tests

Parametric tests rely on a probability distribution of known form as a model for the null hypothesis.

The python notebook

`hypothesis_testing.ipynb`

contains some worked examples of commonly encountered parametric tests.

Hypothesis Testing — Type I & type II errors

Rejection Table		
	H_0 True	H_0 False
Reject H_0	Type I Error (α)	Correct Decision ($1 - \beta$)
Fail to Reject H_0	Correct Decision	Type II Error (β)

Two possible errors can be made when using p-values to make a decision

- **Type I error:** reject the null hypothesis when it is true
- **Type II error:** not reject the null hypothesis when it is false

Hypothesis Testing — Type I & type II errors

Probability of Type I and Type II errors:



- α : the significance level α is the probability of Type I error.
- β : the probability of Type II error relative to the H_1 is called β

The **statistical power** of a test is given by $1 - \beta$, i.e. the probability that H_0 is rejected when H_1 is true.

Methods to reduce errors:

- $\alpha \downarrow \longrightarrow \beta \uparrow$
- $\beta \downarrow \longrightarrow \alpha \uparrow$
- Increase the sample size, n

References

-  William Mendenhall, Terry Sincich. *Statistics for Engineering and the Sciences*. Pearson/Prentice Hall, Upper Saddle River, New Jersey, 2007.
-  Douglas G. Altman. *Practical Statistics for Medical Research*. CRC press, 1990.

Acknowledgments

Many thanks to Yuan Qin, who developed the original version of this course.