

MAX PLANCK INSTITUTE FOR GRAVITATIONAL PHYSICS
IMPRS LECTURE SERIES

Making sense of data: introduction to statistics for gravitational wave astronomy

Lecturer: Jonathan Gair

Winter, 2019–2020

This course will provide a general introduction to statistics, which will be useful for researchers working in the area of gravitational wave astronomy. It will start with some of the basic ideas from classical (frequentist) and Bayesian statistics then show how some of these ideas are or will be used in the analysis of data from current and future gravitational wave (GW) detectors. The final section of the course will introduce some advanced topics that are also relevant to GW observations. These topics will not be expounded in great depth, but some of the key ideas will be described to provide familiarity with the concepts. The aim of the course will be to establish sufficient grounding in statistics that students will be able to understand research seminars and papers, and know where to begin if carrying out research in these areas.

The lectures will be supported by a number of computer practicals. Statisticians typically use the community software package `R` and this is also commonly used by researchers in other disciplines. Most new statistical methods that are developed are implemented as `R` packages and so familiarity with `R` will enable the user to carry out fairly sophisticated analyses straightforwardly. However, in physics it is more common these days to use `PYTHON` and there are a number of libraries of statistical functions and methods available for `PYTHON` as well. Therefore, the practicals will use `PYTHON`.

Course outline

1. (weeks 1–2) Classical (frequentist) statistics.
 - Random variables: definition, properties, some useful probability distributions, central limit theorem.
 - Statistics: definition, estimators, likelihood, desirable properties of estimators, Cramer-Rao bound.
 - Hypothesis testing: definition, Neyman-Pearson lemma, power and size of tests, type I and type II errors, ROC curves, confidence regions, uniformly-most-powerful tests.
2. (weeks 3–4) Bayesian statistics.
 - Bayes' theorem, conjugate priors, Jeffrey's prior.
 - Bayesian hypothesis testing, hierarchical models, posterior predictive checks.
 - Sampling methods for Bayesian inference.
3. (weeks 5–6) Statistics in gravitational wave astronomy.
 - Stochastic processes, optimal filtering, signal-to-noise ratio, sensitivity curves.
 - Frequentist statistics in GW astronomy: false alarm rates, Fisher Matrix, PSD estimation.
 - Bayesian statistics in GW astronomy: parameter estimation, population inference, model selection.
4. (weeks 7–8) Advanced topics in statistics.
 - Time series analysis: auto-regressive processes, moving average processes, ARMA models.
 - Nonparametric regression: kernel density estimation, smoothing splines, wavelets.
 - Gaussian processes, Dirichlet processes.

1 Random variables

In classical physics most things are deterministic. There are physical laws governing the evolution of a system which can be solved and used to predict the state of the system in the future. In reality there are many situations in which things are not (or effectively not) deterministic, and so the outcome of an experiment cannot be predicted with certainty. However, if the experiment is repeated many times some outcomes will occur more frequently than others. This notion of in-deterministicity in measurements is encoded in the concept of a *random variable*. A random variable, X , is a quantity that, when observed, can take one of a (possibly infinite) number of values. Prior to making a measurement the value of the random variable cannot be predicted, but the relative frequency of the outcomes over many experiments are described by a *probability distribution*. The value that X takes in a particular observation (or experiment), x_i , say is called a *realisation* of the random variable.

Random variables can be *discrete*, in which case the values that the variable takes are drawn from a countable set of discrete possibilities, or *continuous* in which case the random variable may take on any value within one or more ranges.

1.1 Discrete random variables

A discrete random variable X can take on any of a (possibly infinite but countable) set of possible values, $\{x_1, x_2, \dots\}$, which together comprise the *sample space*. The probability that X takes any particular value is represented by a *probability mass function* (pmf), which is a set of numbers $\{p_i\}$ with the properties $0 \leq p_i \leq 1$ for all i and $\sum p_i = 1$. The probability that X takes the value x_i is p_i .

1.2 Examples of discrete random variables

1.2.1 Binomial and related distributions

The Binomial distribution is the distribution of the number of success in n trials for which the probability of success in one trial is p . We write $X \sim B(n, p)$ and

$$P(X = k) = p_k = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \{1, \dots, n\}, \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

When $n = 1$ this is the *Bernoulli distribution*. The binomial distribution is the distribution of the sum of n Bernoulli trials, i.e., the number of “successes” in n trials. A related distribution is the *negative binomial distribution* which has pmf

$$P(X = k) = p_k = \begin{cases} \binom{k+r-1}{k} p^k (1-p)^r & \text{if } k \in \{0, 1, \dots\}, \\ 0 & \text{otherwise} \end{cases} . \quad (2)$$

This is the distribution of the number of successes in a sequence of Bernoulli trials that will be observed before r failures have been observed. Setting $r = 1$ and $p \rightarrow (1-p)$ this is the *geometric distribution*, which is the distribution of the number of trials required before the first success.

Another generalisation of the Binomial distribution is the *multinomial distribution*. In this case the outcome of a trial is not a binary ‘success’ or ‘fail’, but it is one of k possible outcomes. The probability of each outcome is denoted p_i with $\sum_{i=1}^k p_i = 1$ and the multinomial distribution describes the probability of seeing n_1 occurrences of outcome 1, n_2 occurrences of outcome 2 etc. in n trials. The pmf is

$$P(\{n_1, \dots, n_k\}) = \begin{cases} \frac{n!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} & \text{if } n_i \geq 0 \forall i \text{ and } \sum_{i=1}^k n_i = n \\ 0 & \text{otherwise} \end{cases} . \quad (3)$$

Applications: counting problems, e.g., distribution of events in categories or time, trials factors.

1.2.2 Poisson distribution

This is the distribution of the number of occurrences of some event in a certain time interval if that event occurs at a *rate* λ . The quantity X follows a Poisson distribution, $X \sim P(\lambda)$ if

$$P(X = k) = p_k = \begin{cases} \lambda^k e^{-\lambda} / k! & \text{if } k \in \{0, 1, \dots\}, \\ 0 & \text{otherwise} \end{cases} . \quad (4)$$

The Poisson distribution is the limiting distribution of $B(n, p)$ as $n \rightarrow \infty$, $p \rightarrow 0$ with $np = \lambda$ fixed.

Applications: distribution of number of events in a population, e.g., gravitational wave sources.

1.3 Continuous random variables

A continuous random variable can take any (usually real, but the extension to complex RVs is straightforward) value within some continuous range, or some set of ranges, which together comprise the *sample space* \mathcal{X} . The probability that X takes a particular value is characterised by the *probability density function* (pdf), $p(x)$. The probability that X takes a value in the range x to $x + dx$ is $p(x)dx$. The pdf has the properties $0 \leq p(x) \leq 1$ for all $x \in \mathcal{X}$ and

$$\int_{x \in \mathcal{X}} p(x) dx = 1. \quad (5)$$

For single valued random variables with non-disjoint sample spaces continuous random variables may also be characterised by the *cumulative density function* or CDF, defined as

$$P(X \leq x) = \int_{-\infty}^x p(x) dx. \quad (6)$$

1.3.1 Uniform distribution

X is uniform on an interval (a, b) , denoted $X \sim U[a, b]$ if the pdf is constant on the interval $[a, b]$

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} . \quad (7)$$

X takes values only in the range $[a, b]$.

Applications: often used as an “uninformative” prior in parameter estimation.

1.3.2 Normal distribution

X is Normal with *mean* μ and *variance* σ^2 , denoted $X \sim N(\mu, \sigma^2)$ if the pdf has the form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (8)$$

X takes all values in the range $(-\infty, \infty)$. If $\mu = 0$ and $\sigma^2 = 1$ we say that X follows a *standard Normal distribution*.

Applications: distribution of noise fluctuations in a gravitational wave detector, priors on mass distribution, most common distribution to assume in parametric statistics.

1.3.3 Chi-squared distribution

X is chi-squared with k degrees of freedom, denoted $X \sim \chi^2(k)$ or χ_k^2 is the pdf has the form

$$p(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad (9)$$

Here $\Gamma(n)$ is the Gamma function, defined by

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx \quad (10)$$

and such that $\Gamma(n+1) = n!$. X takes non-negative real values only, $x \in [0, \infty)$. This is the distribution of the sum of the squares of n independent standard normal distributions.

There is also a *non-central chi-square distribution* which depends on two parameters — degrees of freedom, $k > 0$, as before plus a *non-centrality parameter*, $\lambda > 0$. This has the pdf

$$p(x) = \frac{1}{2} e^{-\frac{(x+\lambda)}{2}} \left(\frac{x}{\lambda}\right)^{\frac{k}{4}-\frac{1}{2}} I_{\frac{k}{2}-1}(\sqrt{\lambda x}) \quad (11)$$

where $I_\nu(y)$ is the modified Bessel function of the first kind. The non-central chi-square distribution again takes non-negative values only and arises as the distribution of the sum of k independent normal distributions with equal (unit) variance, but non-zero means, denoted μ_i . The non-centrality parameter is then $\lambda = \sum_{i=1}^k \mu_i^2$.

Applications: used to test for deviations from normality, e.g., in noise fluctuations in a gravitational wave detector.

1.3.4 Student's t-distribution

X follows Student's t-distribution with $n > 0$ degrees of freedom, $X \sim t_n$, if it has pdf

$$p(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \quad (12)$$

The Student t -distribution arises in hypothesis testing as the distribution of the ratio of a standard Normal distribution to the square root of an independent χ_n^2 distribution, normalised by the degrees of freedom. Specifically if $X \sim N(0, 1)$ and $Y \sim \chi_n^2$ then $X/\sqrt{Y/n}$ follows a t_n distribution.

Applications: used for statistical test on significance of parameters in linear models, used as a “heavy-tailed” distribution for robust parameter estimation, arises naturally when marginalising over uncertainty in power-spectral density estimation.

1.3.5 F-distribution

X follows an F-distribution with degrees of freedom $n_1 > 0$ and $n_2 > 0$ if it has pdf

$$p(x) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} \quad (13)$$

where $B(a, b)$ is the beta function, which is given by

$$B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx \quad (14)$$

and is related to the Gamma function through $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. The F-distribution arises as the ratio of two independent chi-squared distributions with n_1 and n_2 degrees of freedom.

Applications: arises primarily in analysis of variance to test differences between groups.

1.3.6 Exponential distribution

X is exponential with *rate* $\lambda > 0$, $X \sim \mathcal{E}(\lambda)$ if it has pdf

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

X takes positive real values only, $x \in (0, \infty)$. The exponential distribution is the distribution of the time that elapses between successive events of a Poisson process.

Applications: distribution of time lag between events, e.g., gravitational wave signals.

1.3.7 Gamma distribution

X is Gamma with parameters $n > 0$ and $\lambda > 0$, $X \sim \text{Gamma}(n, \lambda)$, if it has pdf

$$p(x) = \begin{cases} \frac{1}{\Gamma(n)} \lambda^n x^{n-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

X takes positive real values only, $x \in (0, \infty)$. The Gamma distribution is the distribution of the sum of n exponential distributions with parameter λ .

Applications: conjugate distribution to the Poisson distribution, so useful in Bayesian analysis of rates. Useful as prior distribution whenever variable has support on $[0, \infty)$.

1.3.8 Beta distribution

X is Beta with parameters $a > 0$ and $b > 0$, $X \sim \text{Beta}(a, b)$, if it has pdf

$$p(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

X takes values in the range $x \in (0, 1)$ only.

Applications: conjugate to binomial distribution. Useful as prior when variable has support on $[0, 1]$, e.g., for probabilities.

1.3.9 Dirichlet distribution

The Dirichlet distribution is a multivariate extension of the Beta distribution. A realisation of a Dirichlet random variable is a set of K values, $\{x_i\}$, satisfying the constraints $0 < x_i < 1$ for all i and $\sum_{i=1}^K x_i = 1$. The Dirichlet distribution is characterised by a vector of *concentration parameters* $\vec{\alpha} = (\alpha_1, \dots, \alpha_K)$ satisfying $\alpha_i > 0$ for all i and has pdf

$$p(x) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^K x_i^{\alpha_i-1}, \quad \text{where } B(\vec{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{j=1}^K \alpha_j\right)}. \quad (18)$$

Applications: infinite dimensional generalisation is a Dirichlet process which is used as a distribution on probability distributions. Very important in Bayesian nonparametric analysis.

1.3.10 Cauchy distribution

X follows a Cauchy distribution (also known as a Lorentz distribution) with *location parameter* x_0 and *scale parameter* $\gamma > 0$, if it has pdf

$$p(x) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}. \quad (19)$$

X takes any real value $x \in (-\infty, \infty)$. The Cauchy distribution arises as the distribution of the x intercept of a ray issuing from the point (x_0, γ) with a uniformly distributed angle. It is also the distribution of the ratio of two independent zero-mean Normal distributions.

Applications: used to model distributions with sharp features. In a gravitational wave context it is used as a model for lines in the spectral density of gravitational wave detectors, for example in BayesLine (and hence BayesWave).

1.4 Properties of random variables

The pdf (or pmf) of a random variable tells us everything about the random variable. However, it is often convenient to work with a smaller number of quantities that summarise the properties of the distribution. These characterise the ‘average’ value of a random variable and the spread of the random variable about the average. We summarise a few of these quantities here. They all rely on the notion of an *expectation value*, denoted \mathbb{E} . The expectation value of a function, $T(X)$, of a discrete random variable X is defined by

$$\mathbb{E}(T(X)) = \sum_{i=1}^{\infty} p_i t(x_i). \quad (20)$$

A similar definition holds for continuous random variables by replacing the sum with an integral

$$\mathbb{E}(T(X)) = \int_{-\infty}^{\infty} p(x)t(x)dx. \quad (21)$$

1.4.1 Quantities representing the average value of a random variable

- **Mean** The mean, often denoted μ , is the expectation value of X , $\mu = \mathbb{E}(X)$.
- **Median** The median, m , is the central value of the distribution in probability, i.e., a value such that the probability of obtaining a value smaller than that or larger than that is (roughly) equal. For discrete random variables $m = x_k$, where

$$\sum_{i:x_i < x_k} p_i < 0.5 \quad \text{and} \quad \sum_{i:x_i \leq x_k} p_i \geq 0.5. \quad (22)$$

For continuous random variables m is the value such that

$$\int_{-\infty}^m p(x) dx = \int_m^{\infty} p(x) dx = \frac{1}{2}. \quad (23)$$

- **Mode** The mode, M , is the ‘most probable’ value of the random variable. For discrete random variables

$$M = \operatorname{argmax}_{i \in \mathcal{X}} p_i \quad (24)$$

and for continuous random variables

$$M = \operatorname{argmax}_{x \in \mathcal{X}} p(x). \quad (25)$$

The mode may not be unique.

1.4.2 Quantities representing the spread of a random variable

- **Variance** The variance, often denoted σ^2 , is the expectation value of the squared distance from the mean, i.e.,

$$\operatorname{Var}(X) = \mathbb{E} [(X - \mathbb{E}(X))^2]. \quad (26)$$

- **Standard deviation** The standard deviation is simply the square root of the variance, usually denoted σ .
- **Covariance** When considering two random variables, X and Y say, the covariance is defined as the expectation value of the product of their distance from their respective means, i.e.,

$$\operatorname{cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]. \quad (27)$$

Here the expectation value is taken with respect to the joint distribution (see section on independence below).

- **Skewness** Given the mean, μ , and variance, σ^2 , defined above, the skewness of a distribution is

$$\gamma_1 = \mathbb{E} \left[\left(\frac{x - \mu}{\sigma} \right)^3 \right]. \quad (28)$$

- **Kurtosis** In a similar way, kurtosis is defined as

$$\text{Kurt}(X) = \mathbb{E} \left[\left(\frac{x - \mu}{\sigma} \right)^4 \right]. \quad (29)$$

This measures the heaviness of the tails of the distribution of the random variable. The kurtosis of the Normal distribution is 3, so it is common to quote *excess kurtosis*, which is the kurtosis minus 3, i.e., the excess relative to the Normal distribution.

- **Higher moments** Higher moments can be defined in a similar way. The n 'th moment about a reference value c of a probability distribution is

$$\mathbb{E} [(X - c)^n]. \quad (30)$$

Moments are usually defined with c taken to be the mean, μ , as in the definition of skewness and kurtosis above.

1.4.3 Moment generating functions

A useful object for computing summary quantities of a probability distribution is the *moment generating function*, $M_X(t)$, which is defined as

$$M_X(t) = \mathbb{E} [e^{tX}] \quad t \in \mathbb{R}. \quad (31)$$

It is clear that derivatives of this function with respect to t , evaluated at $t = 0$, give successive moments about zero of the distribution. Moment generating functions (MGFs) are defined in the same way for both discrete and continuous random variables.

In Table 1 we list these various summary quantities for the probability distributions listed earlier. Where quantities are not known in closed form they are omitted from this table.

| Distribution | Mean | Median | Mode | Variance | Skewness | Excess kurtosis | MGF |
|---------------------------------|--|--|--|---|--|-----------------------------|---|
| Binomial(n, p) | np | $\lfloor np \rfloor$ | $\lfloor (n+1)p \rfloor$ | $np(1-p)$ | $\frac{1-2p}{\sqrt{np(1-p)}}$ | $\frac{1-6p(1-p)}{np(1-p)}$ | $(1-p+pe^t)^n$ |
| Poisson(λ) | λ | $\approx \lfloor \lambda + \frac{1}{3} - \frac{0.02}{\lambda} \rfloor$ | $\lfloor \lambda \rfloor - 1, \lfloor \lambda \rfloor$ | λ | $\lambda^{-\frac{1}{2}}$ | λ^{-1} | $\exp[\lambda(e^t - 1)]$ |
| Uniform $[a, b]$ | $\frac{1}{2}(a+b)$ | $\frac{1}{2}(a+b)$ | all | $\frac{1}{12}(b-a)^2$ | 0 | $-\frac{6}{5}$ | $\frac{e^{tb}-e^{ta}}{t(b-a)}$ |
| Normal(μ, σ^2) | μ | μ | μ | σ^2 | 0 | 0 | $\exp[\mu t + \frac{1}{2}\sigma^2 t^2]$ |
| χ_n^2 | n | $\approx n(1 - \frac{2}{9n})^3$ | $\max(n-2, 0)$ | $2n$ | $\sqrt{\frac{8}{n}}$ | $\frac{12}{n}$ | $(1-2t)^{-k/2}$ |
| Student's t_n | 0 | 0 | 0 | $\frac{n}{n-2}$ | 0 for $n > 3$ | $\frac{6}{n-4}$ for $n > 4$ | — |
| F(n_1, n_2) | $\frac{n_1}{n_2-2}$ | — | $\frac{n_2(n_1-2)}{n_1(n_2+2)}$ | $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ | $\frac{(2n_1+n_2-2)\sqrt{8(n_2-4)}}{(n_2-6)\sqrt{n_1(n_1+n_2-2)}}$ | see caption | — |
| $\mathcal{E}(\lambda)$ | $\frac{1}{\lambda}$ | $\frac{\ln 2}{\lambda}$ | 0 | $\frac{1}{\lambda^2}$ | 2 | 6 | $\frac{\lambda}{\lambda-t}$ |
| Gamma(n, λ) | $\frac{n}{\lambda}$ | — | $\frac{n-1}{\lambda}$ | $\frac{n}{\lambda^2}$ | $\frac{2}{\sqrt{n}}$ | $\frac{6}{n}$ | $(1-\frac{t}{\lambda})^{-n}$ |
| Beta(a, b) | $\frac{a}{a+b}$ | $I_{\frac{1}{2}}^{[-1]}(a, b)$ | $\frac{a-1}{a+b-2}$ | $\frac{ab}{(a+b)^2(a+b+1)}$ | $\frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}$ | see caption | see caption |
| Dirichlet ($K, \vec{\alpha}$) | $\frac{\alpha_i}{\sum_{j=1}^K \alpha_j}$ | — | $\frac{\alpha_i-1}{\sum_{j=1}^K \alpha_j - K}$ | $\frac{\bar{\alpha}_i(1-\bar{\alpha}_i)}{\alpha_0+1}$ | — | — | — |
| Cauchy (x_0, γ) | undefined | x_0 | x_0 | undefined | undefined | undefined | does not exist |

Table 1: Summary of important properties of common probability distributions. The excess kurtosis of the F distribution is $12n_1(5n_2 - 22)(n_1 + n_2 - 2) + (n_2 - 4)(n_2 - 2)^2/[n_1(n_2 - 6)(n_2 - 8)(n_1 + n_2 - 2)]$. For the Beta(a, b) distribution, the excess kurtosis is $6[(a - b)^2(a + b + 1) - ab(a + b + 2)]/[ab(a + b + 2)(a + b + 3)]$ and the MGF is $1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{a+r}{a+b+r} \right) \frac{t^k}{k!}$. For the Dirichlet distribution, the mean and variance are quoted for one component of the distribution, x_i , the parameters $\alpha_0 = \sum_{j=1}^K \alpha_j$ and $\bar{\alpha}_i = \alpha_i / \sum_{j=1}^K \alpha_j$ and the covariance $\text{cov}(x_i, x_j) = -\bar{\alpha}_i \bar{\alpha}_j / (1 + \alpha_0)$.

1.5 Independence

Most of the random variables described above are single valued, but a few of them, e.g., the multinomial and Dirichlet distributions, return multiple values. In other situations, several random variables might be evaluated simultaneously, or sequentially, or the same random variable might be observed multiple times. When dealing with multiple random variables, covariance as introduced above is an important concept, as is *independence*. A set of random variables $\{X_1, \dots, X_N\}$ are said to be *independent* if

$$P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_N \leq x_N) = P(X_1 \leq x_1)P(X_2 \leq x_2) \dots P(X_N \leq x_N) \quad \forall x_1, x_2, \dots, x_N. \quad (32)$$

In terms of the pdf (or pmf) the random variables are independent if their joint distribution $p(x_1, \dots, x_N)$ can be separated

$$p(x_1, \dots, x_N) = p_{X_1}(x_1)p_{X_2}(x_2) \dots p_{X_N}(x_N). \quad (33)$$

Independence of two random variables implies that the covariance is 0, but the converse is not true except in certain special cases, for example for two Normal random variables.

A set of variables $\{X_i\}$ is called *independent identically distributed* or IID if they are independent and all have the same probability distribution. This situation arises often, for example when taking multiple repeated observations with an experiment.

1.6 Linear combinations of random variables

Suppose X_1, \dots, X_N are (not necessarily independent) random variables and consider a new random variable Y defined as

$$Y = \sum_{i=1}^N a_i X_i. \quad (34)$$

For any set of random variables

$$\mathbb{E}(Y) = \sum_{i=1}^N a_i \mathbb{E}(X_i), \quad \text{Var}(Y) = \sum_{i=1}^N a_i^2 \text{Var}(X_i) + \sum_{i \neq j} a_i a_j \text{cov}(X_i, X_j). \quad (35)$$

If the random variables are *independent* then the variance expression simplifies to

$$\text{Var}(Y) = \sum_{i=1}^N a_i^2 \text{Var}(X_i) \quad (36)$$

and the moment generating function of Y can be found to be

$$M_Y(t) = \prod_{i=1}^N M_{X_i}(a_i t). \quad (37)$$

A commonly used linear combination of random variables is the *sample mean* of a set of IID random variables, defined as

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i \quad (38)$$

for which

$$\mathbb{E}(\hat{\mu}) = \mathbb{E}(X_1), \quad \text{Var}(\hat{\mu}) = \frac{1}{n} \text{Var}(X_1), \quad M_{\hat{\mu}}(t) = \left(M_{X_1} \left(\frac{t}{N} \right) \right)^N. \quad (39)$$

1.7 Laws of large numbers

Suppose that X_1, \dots, X_n are a sequence of IID random variables, each having finite mean μ and variance σ^2 . We denote the sum of the random variables by

$$S_n = \sum_{i=1}^n X_i, \quad \text{which implies } \mathbb{E}(S_n) = n\mu, \quad \text{Var}(S_n) = n\sigma^2. \quad (40)$$

Laws of large numbers tells us that the sample mean becomes increasingly concentrated around the mean of the random variable as the number of samples tends to infinity.

1.7.1 Weak law of large numbers

The *weak law of large numbers* states that, for $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (41)$$

1.7.2 Strong law of large numbers

The *strong law of large numbers* states simply

$$P\left(\frac{S_n}{n} \rightarrow \mu\right) = 1. \quad (42)$$

1.7.3 Central limit theorem

In many applications, people assume that the data generating process is Normal. This is partially because the Normal distribution is convenient to work with and has many nice properties, but also because regardless of the distribution large samples of random variables tend to look quite Normally distributed. This fact is encoded in the *Central Limit Theorem*, which states that the standardized sample mean, S_n^* , is approximately standard Normal in the limit $n \rightarrow \infty$

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}. \quad (43)$$

Formally the statement of the central limit theorem is

$$\lim_{n \rightarrow \infty} P(a \leq S_n^* \leq b) = \Phi(b) - \Phi(a) = \lim_{n \rightarrow \infty} P(n\mu + a\sigma\sqrt{n} \leq S_n \leq n\mu + b\sigma\sqrt{n}). \quad (44)$$

2 Frequentist statistics

In the last section we discussed the notion of a random variable. When observing phenomena in nature or performing experiments we would like to deduce the distribution of the random variable, i.e., the probability distribution from which realisations of that random variable are drawn. In **parametric inference** we assume that the distribution of the random variable takes a particular form, i.e., it belongs to a known family of probability distributions. All of the distributions that were described in the previous section are characterised by one or more parameters and so inference about the form of the distribution reduces to inference about the values of those parameters.

In *frequentist statistics* we assume that the parameters characterising the distribution are *fixed* but *unknown*. Statements about the parameters, for example *significance* and *confidence* are statements about multiple repetitions of the same observation, with the parameters fixed. Key frequentist concepts are *statistics*, *estimators* and *likelihood*.

A **statistic** is a random variable or random vector $T = t(\mathbf{X})$ which is a function of \mathbf{X} but does not depend on the parameters of the distribution, θ . Its realised value is $t = t(\mathbf{x})$. In other words a statistic is a function of observed data only, not the unknown parameters.

An **estimator** is a statistic used to estimate the value of a parameter. Typically the random vector would be a set of IID random variables, X_1, \dots, X_n with pdf $p(x|\theta)$. A function $\hat{\theta}(X_1, \dots, X_n)$ of X_1, \dots, X_n used to infer the parameter values is called an **estimator** of θ ; note that $\hat{\theta}$ is a random variable with a sampling distribution in this latter context. The value of the estimator at the observed data $\hat{\theta}(x_1, \dots, x_n)$ is called an **estimate** of θ .

A statistic might also be used to provide an upper or lower limit for a *confidence interval* on the value of a parameter, or to evaluate the validity of a hypothesis in *hypothesis testing*.

2.1 Likelihood

Likelihood is central to the theory of frequentist parametric inference.

If an event E has probability which is a specified function of parameters $\vec{\theta}$, then the likelihood of E is $\mathbb{P}(E|\vec{\theta})$, regarded as a function of $\vec{\theta}$.

The likelihood, denoted $L(\vec{\theta}; \mathbf{x})$, is functionally the same as the pdf of the data generating process, the difference is that the likelihood is regarded as a function of the parameters $\vec{\theta}$ while the pdf is regarded as a function of the observed data, \mathbf{x} . It is often convenient to work with the **log likelihood**

$$l(\theta; \mathbf{x}) = \ln[L(\theta; \mathbf{x})] = \ln[p(\mathbf{x}|\theta)] \quad (\theta \in \Theta)$$

Another useful quantity is the **score**

$$\frac{\partial l}{\partial \theta_i}$$

which is a vector that is also regarded as a function of $\vec{\theta}$ with the data fixed at the observed values.

One interpretation of likelihood is that, given data \mathbf{x} , the relative plausibility of or support for different values $\vec{\theta}_1, \vec{\theta}_2$ of $\vec{\theta}$ is expressed by

$$\frac{L(\vec{\theta}_1; \mathbf{x})}{L(\vec{\theta}_2; \mathbf{x})} \quad \text{or} \quad l(\vec{\theta}_1; \mathbf{x}) - l(\vec{\theta}_2; \mathbf{x}).$$