Principle of Data Science

Xinyu Zhong
Queens' College

January 7, 2024

Contents

Abstract

Abstract of this course

Abstract

Abstract of this course

1 Understanding Data

1.1 Visualising Data

1.1.1 Histograms

1.1.2 Scatter Plots

1.2 Measuring the moments

1.2.1 Average

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

1.2.2 Variance

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

1.2.3 Higher Moments

nth central moments

$$\mu_n = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^n$$

nth algebraic moments

$$\alpha_n = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

1.3 Covariance and Correlation

Covariance:

$$V_{xy} = cov(x, y) = \overline{xy} - \bar{x}\bar{y} = \frac{1}{N}\sum_{i}^{N} (x_i - \bar{x})(y_i - \bar{y}) \tag{1}$$

Highly positive covariance means that one variance tends high when the other high, whereas a highly negative covariance means that one variable tends low while the other tends high. A low magnitude of absolute covariance indicates that two variable are more independent of each other.

Covariance Matrix:

$$\boldsymbol{V} = E\left[(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})^{\mathrm{T}}\right] = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{1k}\sigma_1\sigma_n & \dots & \dots & \sigma_n^2 \end{bmatrix} \tag{2}$$

Correlation:

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x\sigma_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x\sigma_y} \tag{3}$$

Correlation is just covariance normalised by the standard deviation product. It is also known as R, where R^2 is the coefficient of determination.

1.4 Learning from Data

1.4.1 Typical Structure of Data

Common data structure include NumPy array or Pandas dataframes.
Note that we have *features* and *events* for a set of data. In pandas, the features correspond to columns whereas event correspond to rows.

1.4.2 Exploiting correlation in data

See Machine Learning for more information.

- Recall: true positive rate. Sometimes also called signal efficiency or sensitivity. The fraction of all positive or signal events that are correctly classified:

$$TPR = \frac{TP}{TP + FN}.$$

- Specificity: true negative rate. Sometimes also called background efficiency. The fraction of all negative or background events that are correctly classified:

$$TNR = \frac{TN}{TN + FP}.$$

- False positive rate. The fraction of all negative or background events that incorrectly classified:

$$FPR = \frac{FP}{FP + TN}.$$

- False negative rate. The fraction of all positive or signal events that incorrectly classified:

$$FNR = \frac{FN}{FN + TP}$$

- Note the relationships between TPR and FNR, as well as TNR and FPR:

$$TPR + FNR = \frac{TP}{TP + FN} + \frac{FN}{FN + TP} = \frac{TP + FN}{TP + FN} = 1,$$
$$TNR + FPR = \frac{TN}{TN + FP} + \frac{FN}{FP + TN} = \frac{TN + FP}{TN + FP} = 1,$$

so that $FPR = 1 - TNR, FNR = 1 - TPR$.

1.4.3 Performance criteria and metrics

- True positive (TP). Correct prediction of positive or signal outcome.
- False positive (FP). Incorrect prediction of positive or signal outcome.
- True negative (TN). Correct prediction of negative or background outcome.
- False negative (FN). Incorrect prediction of negative or background outcome.
- All positive or signal events are given by $P = TP + FN$.
- All negative or background events are given by $N = TN + FP$.
- All events classified as positive or signal-like are given by $C_P = TP + FP$.

- All events classified as negative or signal-like are given by $C_N = TN + FN$.
- Recall(or signal efficiency in particle physics): true positive rate. Sometimes also called signal efficiency or sensitivity. The fraction of all positive or signal events that are correctly classified:
$$TPR = \frac{TP}{TP + FN}.$$
- Specificity(or): true negative rate. Sometimes also called background efficiency. The fraction of all negative or background events that are correctly classified:
$$TNR = \frac{TN}{TN + FP}.$$
- False positive rate. The fraction of all negative or background events that incorrectly classified:
$$FPR = \frac{FP}{FP + TN}.$$
- False negative rate. The fraction of all positive or signal events that incorrectly classified:
$$FNR = \frac{FN}{FN + TP}$$
- Note the relationships between TPR and FNR, as well as TNR and FPR:
$$TPR + FNR = \frac{TP}{TP + FN} + \frac{FN}{FN + TP} = \frac{TP + FN}{TP + FN} = 1,$$
$$TNR + FPR = \frac{TN}{TN + FP} + \frac{FN}{FP + TN} = \frac{TN + FP}{TN + FP} = 1,$$
- Accuracy. The fraction of all events that correctly classified:
$$\alpha = \frac{TP + TN}{P + N}.$$
- Error rate. The fraction of all events that are incorrectly classified:
$$\varepsilon = \frac{FP + FN}{P + N}.$$
- Purity. The fraction of all events classified positively that are correctly classified:
$$\rho_P = \frac{TP}{TP + FP} \quad \text{and} \quad \rho_N = \frac{TN}{TN + FN}.$$
- Significance. For a counting experiment quantifies the statistical significance:
$$\sigma = \frac{TP}{\sqrt{TP + FP}}.$$
- Signal-to-noise, sometimes also called signal-to-background ratio:
$$SNR = \frac{TP}{FP}.$$
- F-score, sometimes also called F-measure:
$$F = 2\frac{\text{precision } \times \text{ recall}}{\text{precisoin } + \text{ recall}} = \frac{2TPR}{2TPR + FPR + FNR}.$$

Concepts of Type I and Type II errors.
Usually, one algorithm will prioritise one thing at the cost of the other, for example maximise TPR will results in a high FPR as well.
A good metric to optimise is ROC (receiver-operating-characteristic), which is the curve in FPR vs TPF graph, and we would like to push the curve to the top left.

1.4.4 Data challeneges

2 Probability

2.1 Definition of Probability

2.1.1 Frequentist

- The true probability in practise is never obtainable (because we cannot perform infinite experiments). We can only estimate the probability given the sample size we have. However, it is always possible to perform more experiments than we can keep doing so until we reach the desired accuracy (and any accuracy is in principle achievable).
- Frequentist probability can only be applied to repeatable experiments. For example, I cannot use it to determine the probability it will rain the day after tomorrow. I need a system in which I can keep relevant conditions stable to perform repeatable experiments.
- One does not need to have any prior beliefs about outcomes, the probability is purely determined from observations.

2.1.2 Bayesian

- The true probability is always obtainable (because we can always update our beliefs given new information). However, in practise we may not have enough information to determine the true probability.
- Finetti's coherent bet:If you are willing to bet on the outcome of a random experiment, then you should be willing to bet on the outcome of any exchangeable random experiment.
- Bayesian takes into account the prior belief of the outcome, and update the probability based on the new information.

2.2 Property of Probability

Property of probability is based on Kolmogorov's axioms

- Addition
- Conditional Probability
- Independence A and B are independent if P(A, B) = P(A)P(B)

2.2.1 Monty Hall Problem

Check Example Sheet 1

2.3 Probability mass/density Function

- $P(X)$ to denote probability mass function (discrete probability)
- $p(X)$ to denote probability density function (continuous probability)

2.4 Change of variables

We use

$$X \sim f(x)$$

to denotes that X is distributed like f(x)

$$y = h(x)$$

to denotes that y is a function of X To change variables, probability must be conserved. We use CDF to change variables.

2.4.1 Jacobian Matrix

Jacobian Matrix is the matrix of all first-order partial derivatives of a vector-valued function. It is used to transform the probability density function from one set of variables to another.

2.4.2 The cumulative distribution

Cumulative distribution function:

$$F(X) = \int_{-\infty}^{X} f\left(X'\right) dX' \tag{4}$$

which is the integrated p.d.f.

It is defined so that

$$F\left(X_{\min}\right) = 0,$$
$$F\left(X_{\max}\right) = 1,$$

with some very useful properties

$$P\left(X < X'\right) = \int_{X_{\min}}^{X'} f(X)dX = F\left(X'\right),$$
$$P\left(X' < X < X''\right) = \int_{X'}^{X''} f(X)dX = F\left(X''\right) - F(X).$$

The integral of p.d.f is the difference in c.d.f.
The c.d.f of a Gaussain distribution is mapped to Chi^2 distribution.

2.4.3 The Joint, Marginal and conditional distribution

Joint Probability:

$$f(X,Y) = g(X)h(Y) \tag{5}$$

Marginal Distribution:

$$g(X) = \int f(X,Y)dY \text{ the marginal distribution in } X \tag{6}$$
$$h(Y) = \int f(X,Y)dX \text{ the marginal distribution in } Y. \tag{7}$$

Conditional distribution:

$$g(X \mid Y) = \frac{f(X,Y)}{h(Y)} = \frac{f(X,Y)}{\int f(X,Y)dX} theprobabilityofXgivenY, \tag{8}$$
$$h(Y \mid X) = \frac{f(X,Y)}{g(X)} = \frac{f(X,Y)}{\int f(X,Y)dY} theprobabilityofYgivenX. \tag{9}$$

2.4.4 Bayes' theorem for continuous variables

Bayes' theorem:

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)} = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)d\theta} \tag{10}$$

The terms in the equation

- $p(\theta \mid X)$ - the *posterior distribution* - our probability distribution for the parameter θ given the data we have observed X. 50
- $p(X \mid \theta)$ - the *likelihood function* - the likelihood we observe the data X given a particular value of θ. This is of vital importance across all of statistics and machine learning. We will discuss the likelihood in much more detail in Sec. 3.2.
- $p(\theta)$ - the *prior distribution* - encompassing our prior beliefs about θ, this can be based on previous measurements, previous beliefs or indeed be flat (although note that a change of variables or basis will not necessarily maintain flatness). The prior influences the outcome of Bayesian inference, which can be seen as an advatange and a disadvantage. I will leave discussion of priors to the other stats course.
- $p(X)$ - the *evidence* - this is just a normalisation factor that ensures the posterior is a p.d.f. For Bayesian inference the evidence can often be ignored as the posterior is proportional to the numerator. However, the evidence can be quite a useful quantity for goodness of fit tests. I will leave discussion of the evidence to the other stats course.

.

2.5 Properties of Distributions

2.5.1 Expectation, mean and variance

Expectation:

$$E[g(X)] = \int g(X)f(X)dX \tag{11}$$

where $f(x)$ is the probability density function of X and $g(x)$ is a function of X Expectation is a linear operator, is the true mean of the distribution

Variance:

$$\begin{aligned} V(x) = \sigma^2 &= E\left[(X-\mu)^2\right] && (12)\\ &= E\left[X^2 - 2\mu X + \mu^2\right] && (13)\\ &= E\left[X^2\right] - \mu^2 && (14)\\ &= \sigma^2 && (15) \end{aligned}$$

We can also define the moments of the distribution using expectation values. The moments are defined as

$$\begin{array}{ll} \mu_\ell = E\left[X^\ell\right] & \text{the } \ell^{\text{th}} \text{ algebraic moment} \\ \alpha_\ell = E\left[(X - E[X])^\ell\right] & \text{the } \ell^{\text{th}} \text{ moment about the mean.} \end{array}$$

In our notation the mean is $\mu = \mu_1$ and the variance is $\sigma^2 = \alpha_2$. The skew is $\gamma_1 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$. The kurtosis is $\gamma_2 = \beta_2 - 3 = \mu_4/\mu_2^2 - 3$.

2.5.2 Covariance and Correlation

We can also define the covariance between two random variables X and Y,

$$V(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E[XY] - E[X]E[Y],$$

and then the correlation

$$\rho(X,Y) = \frac{V(X,Y)}{\sigma_X \sigma_Y}$$

Note the math behind derivation.

Example Assume $x \sim f(x) = Ne^{-x^2}$
Find $N \Rightarrow$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow N \Rightarrow \frac{1}{\sqrt{\pi}}$$

Find $E[x] \Rightarrow$

$$\int_{-\infty}^{\infty} xf(x)dx = 0$$

as its add function.
Find $E\left[(x-\mu)^2\right] \Rightarrow$

$$E\left[x^2\right] - (E[x])^2$$

$$E\left[x^2\right] = \int_{-\infty}^{\infty} x^2 f(x)dx \quad \Rightarrow \quad \sigma^2 = \frac{1}{2}$$

To get $\sigma = 1$. $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \rightarrow$ stanclond model shift x by μ and scale by $\frac{1}{\sigma}$. ie. $z = \frac{x-\mu}{\sigma} \Rightarrow$ Now Model

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

$$\frac{1}{N} = \sigma\sqrt{2\pi}, \text{ Mean } = \mu \quad \text{ variance } = \sigma^2$$

2.5.3 The characteristic function

characteristic function:

$$\varphi(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itX} f(X)dX \tag{16}$$

which means that $f(X)$ is completely defined by the characteristic functions

$$f(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) e^{-iXt} dt$$

The usefulness of the characteristic function is shown in proof for central limit theorem, or algebraic moments

$$\mu_n = E\left[X^n\right] = \int_{-\infty}^{\infty} X^n f(X)dX$$

which can be obtained by differentiating the characteristic function n times at point $t = 0$

$$\varphi_n(t) = \frac{d^n \varphi(t)}{dt^n} = i^n \int_{-\infty}^{\infty} x^n e^{itX} f(X)dX$$

such that $\varphi_n(0) = i^n \mu_n$.

2.6 Common Distribution

This section will discuss the common distribution, including the binomial distribution, poisson distribution, normal distribution, chi-squared distribution, exponential distribution, polynomial distribution, and the multi-variate normal distribution.

p.d.f.s depend on one random variable x and parameters θ, write as

$$p(x;\theta)$$

where ';' distinguish rvs and parameters

2.6.1 Binomial Distribution

Binomial distribution:

$$P(k;p,n) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \tag{17}$$

given n trials,$p(success) = p$, $p(fail) = q = 1-p$ and the total probability of k triads are success is

$$p^k(1-p)^{n-k}$$

2.6.2 Poisson Distribution

Poisson distribution:

$$P(k,\lambda) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{18}$$

Derivation of Poisson Distribution from binomial

$$P(k;\lambda/n,n) = \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \frac{n!}{k!(n-k)!}$$

as $n \to \infty$, $\lambda \to \infty$ such that λ/n is finite, then

$$\begin{aligned} P(k;\lambda/n,n) &\to \frac{\lambda^k}{k!}e^{-\lambda} \\ &\to P(k;\lambda) \end{aligned}$$

2.6.3 Normal Distribution

Normal Distribution:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{19}$$

2.6.4 Multi-variate Normal Distribution

for independent r.v.s. $x_1, x_2, \ldots, x_n$

$$p(\vec{\boldsymbol{x}}; \vec{\boldsymbol{\mu}}, \vec{\boldsymbol{\sigma}}) = \prod_{i=1}^{n} p(x_i; \mu_i, \sigma_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}$$

with dependent r.v.s. have a correlation term: $\vec{\boldsymbol{\sigma}} =$

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix}$$

Terms in exp becomes $(x-\mu)^T V^{-1}(x-\mu)$

2.6.5 The Exponential Distributions

Exponential Distribution:

$$p(x; \lambda) = \lambda e^{-\lambda x} \tag{20}$$

The exponential distribution is the continuous analogue of the geometric distribution. It describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate.

Exponential distribution is often concerned with the amount of time until some event occurs. For example, the amount of time (beginning now) until an earthquake occurs has an exponential distribution. Other examples include the length, in minutes, of long distance business telephone calls, and the amount of time, in months, a car battery lasts.

2.6.6 Polynomial distributions

2.6.7 Chi-squared Distribution

Chi-squared Distribution:

$$P(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \tag{21}$$

Chi-squared distribution with k degree of freedom gives the distribution of the sum of squares of k independent standard normal variables
Example of Chi-squared distribution is the distribution of the sum of squares of k independent standard normal variables(by definition).

$$\chi^2 = \sum_{i=1}^{k} x_i^2$$

where x_i are independent standard normal variables, k is the degree of freedom, in this case, k is the number of independent standard normal variables.

Another example is the ratio of negative log likelihoods for two models: according to Wilks' theorem, the ratio of negative log likelihoods for two models is distributed as a chi-squared distribution with the difference in the number of parameters as the degree of freedom.

Another example is chi-squared test, which is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Table 1: Common Probability Distributions

Distribution	PDF $f(x)$	CDF $F(x)$	Mean μ	Variance σ^2	Characteristic Function $\phi(t)$
Gaussian	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\frac{1}{2}\left[1+\text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$	μ	σ^2	$e^{i\mu t-\frac{\sigma^2t^2}{2}}$
Exponential	$\lambda e^{-\lambda x}$	$1-e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda-it}$
Uniform	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$	$e^{-\lambda}\sum_{k=0}^{x}\frac{\lambda^k}{k!}$	λ	λ	$e^{\lambda(e^{it}-1)}$
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$\sum_{k=0}^{x}\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(pe^{it}+1-p)^n$

2.6.8 Convolution of distributions

Convolution is used when we want to find the distribution of the sum of two independent random variables.
Note that the convolution of two distributions is the same as the product of their characteristic functions.

2.6.9 Generating samples from distributions and the inverse c.d.f.

Using inverse c.d.f. to generate samples from distributions or use accept and reject method

2.7 Limit Theorems

2.7.1 Convergence

The law of large numbers states that the sample mean converges to the true mean as the sample size increases. The central limit theorem states that the distribution of the sample mean is Gaussian as the sample size increases. The central limit theorem is a consequence of the law of large numbers

2.7.2 Central Limit Theorem

The sum of a random variable, S, where

$$S = \sum X_i \text{ which has } \mu = \sum \mu_i \text{ and } \sigma^2 = \sum \sigma_i^2.$$

The central limit theorem states that the distribution of S will tend to a normal distribution with $\mu = \mu$ and $\sigma = \sigma$ as $N \to \infty$. In particular if I redefine the random variable so that it is shifted by the mean and scaled by the standard deviation so that,

$$S \to S' = \frac{S - \sum_{i=1}^{N}\mu_i}{\sqrt{\sum_{i=1}^{N}\sigma_i^2}}$$

2.7.3 Errors

Imagine some linear function, $Z = aX + b$, where a and b are constants and X is a random variable with a measured or known variance. We can compute the variance of Z using linear expectation properties:

$$\begin{aligned} V(Z) &= E\left[Z^2\right] - E[Z]^2 \\ &= E\left[(aX+b)^2\right] - E[aX+b]^2 \\ &= a^2 E\left[X^2\right] + abE[X] + b^2 - a^2 E[X]^2 - 2abE[X] - b^2 \\ &= a^2\left(E\left[X^2\right] - E[X]^2\right) \\ &= a^2 V(X). \end{aligned}$$

We can now extend this to any generic function $f(X)$ if we only consider small errors and the linear term in a Taylor expansion of X around some point X_0,

$$f(X) \approx f\left(X_0\right) + \left(X - X_0\right)\left(\frac{df}{dX}\right)\Big|_{X=X_0} + \mathcal{O}\left(X^2\right).$$

We can then write the variance and standard deviation as

$$\begin{aligned} V(f) &\approx \left(\frac{df}{dX}\right)^2 V(X) \\ \sigma_f &\approx \left|\frac{df}{dX}\right| \sigma_X. \end{aligned}$$

We can now extend this formalism to functions of many random variables. For example if we have two random variables, X and Y, and a generic function $f(X, Y)$ then,

$$\begin{aligned} V(f) &= \left(\frac{df}{dX}\right)^2 V(X) + \left(\frac{df}{dY}\right)^2 V(Y) + 2\left(\frac{df}{dX}\right)\left(\frac{df}{dY}\right)\mathrm{cov}(X,Y) \\ \sigma_f^2 &= \left(\frac{df}{dX}\right)^2 \sigma_X^2 + \left(\frac{df}{dY}\right)^2 \sigma_Y^2 + 2\left(\frac{df}{dX}\right)\left(\frac{df}{dY}\right)\rho\sigma_X\sigma_Y \end{aligned}$$

3 Classical Statistics, Estimation and Uncertainties

3.1 Frequentist evaluation of estimators

3.1.1 Consistency, bias and efficiency of estimates

Consistency A consistent estimator will converge to true value as datasize increases:

$$\lim_{N\to\infty} \hat{\theta}(\vec{x}) = \theta$$

Bias The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated:

$$\mathrm{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

It is generally preferrable to have consistent estimator as opposed to biased estimators.

Efficiency The efficiency of an estimator is the ratio of the variances of two unbiased estimators. The estimator with the lower variance is said to be more efficient.

The minimun variance bound states that the variance of an unbiased estimator is always greater than or equal to the inverse of the Fisher information.

3.1.2 Bias-Variance Trade-off

Bias-variance trade-off applies to both ML and statistical modelling. In order to minimise the variance, we want to take in more data, however, the more data we take in, the more bias we introduce. This is the bias-variance trade-off.

- High bias, low variance: Simple model that do not depend on the data much, but do not fit the data well. For example, fitting a straight line to a quadratic function. This leads to high bias (fails to capture the true shape of the data), but low variance (the line will always be roughly in the same place relative to the data points).
- Low bias, high variance: Complex model that fit the data well, but do not generalise well.
- In a nutshell, complexity scales with variance, and simplicity scales with bias.

3.1.3 Estimation of the mean, variance and standard deviation

1. mean: the arithmetic mean of a sample provides a consistent estimator of the true mean. i.e. $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$
2. the variance of the estimate of the mean is given by $\sigma^2_{\hat{\mu}} = \frac{\sigma^2}{N}$, where σ^2 is the variance of the distribution. This is a result of the central limit theorem. The sample mean is the most efficient estimator of the true mean, as the variance is the minimum variance bound, assuming Gaussian distribution.

3.2 The likelihood function

The likelihood function:

$$p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \tag{22}$$

The likelyhood function is a function of θ only, and is the probability of observing the data given the parameter θ

3.2.1 Minimum Variance Bound

There is another useful property of the likelihood which relates to efficiency of estimators. It can be shown that there is a limit to the efficiency of an estimator called the minimum variance bound. Providing the estimator is unbiased then the minimum variance bound states that

$$\begin{aligned} V(\hat{\theta}) &\geq \left(E\left[\left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right] \right)^{-1} \\ &\geq -\left(E\left[\left(\frac{\partial^2 \ln L}{\partial \theta^2} \right) \right] \right)^{-1}. \end{aligned}$$

An estimator is determined to be efficient if the variance of the estimator, $V(\hat{\theta})$, is the equal to the minimum variance bound. Otherwise an estimators efficiency is quantified by the ratio of the minimum variance bound to the variance.

3.3 Maximum Likelihood Estimation

Maximise L is equivalent to maximise $\ln(L)$.In practise it is much easier to work with the natural logarithm of the likelihood, because then the product turns into a sum, and at least for computers it is

easier to sum small numbers than find their product. The $\hat{\theta}$ which maximises L will also be the value that maximise $\ln L$ so it is more common to the see ML condition written as

$$\left.\frac{\partial \ln L(\theta)}{\partial \theta}\right|_{\theta=\theta\hat{\theta}} = \frac{\partial}{\partial \theta}\sum_{i=1}^{N} \ln p\left(X_i \mid \theta\right) = 0.$$

It can be shown that the variance of the estimate is equal to minimum variance bound as N is large.

3.4 Least-square method

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_i^2}$$

σ_{y_i} is the associated uncertainty of y_i. We will see that it is liked to the log likelihood function by

$$\chi^2 = -2\ln L + C$$

note that both side of the equation is a function of θ (estimated ?) only.
This allows us to obtain the relationship in difference:

$$\Delta\chi^2 = -2\Delta \ln L$$

Wilks' theorem:

As $N \to \infty$, the test-statistic which is twice the negative log likelihood ratio approached the χ^2 distribtion. This is an example of hypothesis test where the null hypothesis is the valule of the parameter at the best fit and the alternative hypothesis is the value of the parameter where we read off the log likelihood difference.

3.4.1 Fisher information

Firstly, we need to define the score function, which is the derivative of the log likelihood function with respect to the parameter θ.

Score function:

$$S(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta} \tag{23}$$

The score function is the derivative of the log likelihood function with respect to the parameter θ.
It can be shown that the expectation value of score is zero.

Fisher information:

$$I(\theta) = E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right], \tag{24}$$

$$= -E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right] \tag{25}$$

The Fisher information is the expected value of the squared score function, i.e. variance of the score function. It is then implified by the second derivative of the log likelihood function.

This fisher information is related to the minimum variance bound by an inverse

3.5 Method of Moments

Method of moments is a method of estimating the parameters of a statistical model. Foe example, give a distribution of known form but unknown parameters,

$$f(x;\vec{\theta})$$

We can estimate the parameters by equating the sample moments to the theoretical moments.

$$1 = \int_{-\infty}^{\infty} f(x;\vec{\theta})dx \tag{26}$$

$$\mu = \int_{-\infty}^{\infty} xf(x;\vec{\theta})dx = \text{some funciton of theta}, i.e. g(\mu,\sigma) \approx \hat{\mu} \tag{27}$$

$$\mu^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x;\vec{\theta})d = \text{some function of theta} \approx \hat{\mu}^2 \tag{28}$$

$$... \tag{29}$$

Here $\vec{\theta}$ is a vector of parameters, hence we need the right amount (number of independent parameters, i.e. μ,σ) of moments to estimate the parameters. We can estimate the parameters by equating the sample moments $\hat{\mu}$ to real moments μ, and solve simultaneous equations to get the parameters.

3.5.1 Uncetainties from method of moments estimates

3.6 Goodness-of-Fit Tests

This section will discuss the question of "how good was my fit in the first place", the answer to which is the goodness-of-fit test.

Test Statistics:

$$\text{Quantify the agreement between the data and the model} \tag{30}$$

i.e. from the probability distribution, we can compute the probabilitay that we got the value we did, which in term gives us info on what we think the quality of the fit. Example of Test Statistics include χ^2

3.6.1 χ^2 Tests

χ^2 tests is chosen as we expect $\chi^2/d.o.f = 1$, as the expectation of the χ^2 for k degree of freedom is k.

The χ^2 probability, i.e. $1-F(\chi^2)$ (1 - c.d.f. of the χ^2 distribution of the appropriate degree of freedom). What the value p means is the probability that a function that does describe the data gives a value of the χ^2 larger than the one you find (A larger χ^2 means you did a bad job fitting).

For p-value:

- if p-value is small, the model do not agree well
- if p-value is large, the model and data agree too well, over fitting the data.

χ^2 test is very good at spotting the small p-value.

This test is very dependent on the binning that us used and it is bad if you are trying to assess the compatibility of a model with may only contain a discrepancy in one bin. Example in Higgs Boson discovery, the χ^2 test is not good at spotting the small discrepancy in the 125 GeV bin.

3.6.2 Residual and Pull

Residual is the difference between the data and the model, i.e. $r_i = y_i - f(x_i)$, where y_i is the data and $f(x_i)$ is the model.
Pull is residual weighted by the error, i.e. $p_i = \frac{r_i}{\sigma_i}$, where σ_i is the error of the data. It is also known as *student's t distribution.*
Pull should be Gaussian distributed, with mean 0 and variance 1. If the pull is not Gaussian distributed, it means that the model is not a good fit to the data.
Note that by summing the pull in quadrature, we get the χ^2 value.

3.6.3 Kolmogorov-Smirnoff KS Tests

KS test can be used to test the compatibility of two distributions. It is an unbinned test, so it can be an alternative to the χ^2 test when the number of the event is small. KS score is determined by finding the maximum difference between the two cumulative distributions, multiplying the square root of the sample size.
KS score:

$$P_{\text{KS}} = \max |F_1(x) - F_2(x)| \sqrt{N} \tag{31}$$

which is the maximum deviation between the two distributions CDF and multiplying by the square root of the sample size.

3.7 Confidence Intervals

Whenever we qant to quote a point estimate, we should also quote an interval estimate. e.g. The common practise is quote 68.3% or 95.4% confidence interval.

3.7.1 Bayesian Intervals

A Bayesian credible interval of μ corresponding to some confidence β is constructed by requiring that

$$\beta = \int_{\mu_1}^{\mu_2} p(\mu \mid X) d\mu.$$

3.7.2 Classical Intervals (Neyman-Pearson intervals)

Classical intervals are based on the Neyman-Pearson lemma, which states that the likelihood ratio test is the most powerful test for a given significance level β For a frequentist the confidence interval, $[\mu_1, \mu_2]$, is a member of a set such that

$$p\left(\mu \in [\mu_1, \mu_1]\right) = \beta.$$

3.7.3 Flip-flopping

Flip-flopping is the problem of the Neyman-Pearson intervals, where the upper and lower limits of the interval flip-flop as the data changes.

3.7.4 Feldman-Cousins intervals

Feldman-Cousins intervals are a frequentist method for constructing confidence intervals. It effectively solves the problem of flip-flopping between the upper and lower limits. It is a generalisation of the Neyman-Pearson lemma. For a given confidence level β, the Feldman-Cousins interval is the interval that maximises the probability of the interval containing the true value of the parameter, μ. Firstly, for each X, with μ fixed, we can define the likelihood ratio $R = \frac{p(X|\mu)}{p(X|\hat{\mu})}$, and we add values of X until the

cumulative likelihood ratio is equal to β. Going back to our simple example of the normal distribution then $\hat{\mu} = 0$ when $X < 0$ and $\hat{\mu} = X$ when $X \geq 0$ so that the likelihood ratio will be

$$R = \frac{p(X \mid \mu)}{p(X \mid \hat{\mu})} \begin{cases} \frac{\exp\left[-(X-\mu)^2/2\right]}{1} & \text{if } X \geq 0 \\ \frac{\exp\left[-(X-\mu)^2/2\right]}{\exp[-X^2/2]} & \text{if } X < 0 \end{cases}.$$

3.8 Hypothesis Testing

Type I error is the probability of rejecting the null hypothesis when it is true,α

$$\alpha = P(X > X_0 | H_0)$$

Type II error is the probability of accepting the null hypothesis when it is false. i.e. $(1 - \beta)$

$$\beta = P(X > X_0 | H_1)$$

3.8.1 Neyman-Pearson Lemma

Neyman-Pearson Lemma:

The likelihood ratio test is the most powerful test

$$T = -2 \ln \frac{L(\theta_0)}{L(\theta_1)}$$

Note here that T follows a χ^2 distribution with x degree of freedom, where x is the difference in the number of parameters between the two models. The Neyman-Pearson Lemma addresses the case where you want to construct a test with the maximum power subject to a specified Type I error rate (α). In other words, it helps you design the most powerful test for detecting a particular effect, given a constraint on the probability of making a Type I error.

3.8.2 Quoting a significance

We usually quote the p-value as the significance of the result. The p-value is the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true. Note that a larger test statistic means a smaller p-value.
The smaller the p-value, the more significant the result is. Consequently, when converting the p-value to a Z-score I need to incorporate both sides of the normal distribution tail. This is equivalent to the conversion for **two-sided**,

$$Z = \Phi^{-1}(1 - p/2)$$

where Φ^{-1} is the inverse c.d.f. of the normal distribution. Notice that this is equivalent to the computation we have often used from the χ^2 distribution with one degree of freedom,

$$Z = \sqrt{\Psi^{-1}(1 - p)}$$

where Ψ^{-1} is the inverse c.d.f. of the χ^2 distribution.
For **one-sided** test, we can use the conversion

$$Z = \Phi^{-1}(1 - p),$$

using the inverse c.d.f. of the normal distribution and

$$Z = \sqrt{\Psi^{-1}(1 - 2p)}$$

using the inverse c.d.f. of the χ^2 distribution. Below is a snippet of code which demonstrates this

4 Limit Setting

CL_{sb} is the confidence level for the signal + background hypothesis.

$$CL_{sb} = P(X < X_0|H_1)$$

Assuming the signal + background hypothesis is true, the probability of obtaining a value of the test statistic X less than X_0 is CL_{sb}.
The limit of CL_{sb} is that it has very low sensitivity to the signal, as the signal is usually very small compared to the background.
The improvement is to use the CL_s method, which is the confidence level for the signal hypothesis.

$$CL_s = \frac{p_{sb}}{1 - pb}$$

where p_{sb} is the probability of obtaining a value of the test statistic X more than X_0 assuming the signal + background hypothesis is true, and p_b is the probability of obtaining a value of the test statistic X less than X_0 assuming the background only hypothesis is true.
Note that p-value is defined differently in the two hypothesis.

1. If two test are indistinguishable, then $p_{sb} = 1 - p_b$.
2. If two test are perfectly distinguishable, then $1 - p_b = 1$ and $CL_s = p_{sb} = CL_{sb}$

4.1 Resamplings

Resampling is the process of drawing data from a sample to produce additional samples. Resample is done to estimate the distribution of the estimate $\hat{\theta}$.

4.1.1 Permutation Test

$T = \Delta\bar{X} = \bar{X}_A - \bar{X}_B$ If the events are from same distribution, then random swap will not affect the difference in the mean.

4.1.2 Non-Parametrix Bootstrapping

- Bootstrap generates the toys by sampling from the original dataset with **replacement**.
- Non-parametric bootstrap is more flexible and doesn't assume a particular distribution, making it robust but potentially less efficient than the parametric counterpart.

4.1.3 Parametric Bootstrap

- The samples are generated based on the parameters we have fitted, hence the name parametric bootstrap.
- Parametric bootstrap relies on the assumption that the data follows a specified parametric model, making it more efficient if the model is correct.

4.1.4 Jackknife

- Jackknife is a resampling technique used to estimate the bias and standard error of a statistic.
- Jackknife is a special case of the leave-one-out cross-validation technique.

- Give N data points, we can generate N samples by leaving out one data point each time.

$$\begin{aligned}\hat{\theta}_1 &= f(X_2, X_3, \ldots, X_N)\\ \hat{\theta}_2 &= f(X_1, X_3, \ldots, X_N)\\ &\vdots\\ \hat{\theta}_N &= f(X_1, X_2, \ldots, X_{N-1}).\end{aligned}$$

- The Jackknife average is the average of the N samples.

$$\hat{\theta}_{\mathrm{J}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$$

- We use this to estimate the bias with is

$$\mathrm{Bias}(\hat{\theta}) = (N-1)\left(\hat{\theta}_{\mathrm{J}} - \hat{\theta}\right)$$

- This results in a bias-corrected estimator

$$\hat{\theta'}_{\mathrm{J}} = \hat{\theta}_{\mathrm{J}} - \mathrm{Bias}(\hat{\theta}) = N\hat{\theta}_{\mathrm{J}} - (N-1)\hat{\theta}$$

4.1.5 BCA

Bias corrected and accelerated (BCA) bootstrap is a method for estimating the accuracy of a parameter of a distribution. It improves on the basic bootstrap by adjusting for bias and skewness in the distribution of the sample statistic. It also has better coverage than the percentile bootstrap.

5 Advanced Topics

5.1 Measurement errors and forward modelling

See ADS notes for more details.

5.2 Optimisation

5.2.1 Optimisation Algorithms

We have a few algorithms to find the minimum of a function, including the gradient descent, conjugate gradient, and the Newton-Raphson method.

5.3 Regularisation

Regularisation is a technique used to prevent overfitting in complex models. It is a form of regression, that constrains/regularises or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
See ADS notes for more details.

5.4 Density Estimation

This chapter uses the Kernal Density Estimation (KDE) method to estimate the probability density function of a random variable,which replaces the step function.

5.4.1 Kernal Density Estimation

Kernal Density Estimation:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x - x_i}{h}\right) \tag{32}$$

where K is the kernal function, h is the bandwidth, and N is the number of data points.

5.4.2 Expectation Maximisation and GMM

See ADS notes