Principles of Data Science

Xander Byrne

Michaelmas 2023

1 Fondamentali

1.1 Kolmogorov Axioms

Consider a set of exclusive outcomes X_i, with probabilities $P(X_i)$. The Kolmogorov axioms state that these probabilities must satisfy:

- $P(X_i) \geq 0 \, \forall \, i$. Probabilities can't be negative.
- $\sum_i P(X_i) = 1$, where the sum is over all possible outcomes. Something has to happen.
- For mutually exclusive probabilities (where it is impossible for both to happen), $P(X_i \,\text{or}\, X_j) = P(X_i) + P(X_j)$

1.2 Properties of Probabilities

Consider two outcomes $X_1 = A$ and $X_2 = B$.

- If A and B are *not* mutually exclusive, then the above formula overcounts the possibility that both can happen. In this case, we need to subtract this off to get $P(A \,\text{or}\, B)$:

  $$P(A \,\text{or}\, B) = P(A) + P(B) - P(A \,\text{and}\, B)$$

  If mutually exclusive, this reduces to the above.

- If A and B are *independent*, then

  $$P(A \,\text{and}\, B) = P(A) \times P(B)$$

- The probability of A, assuming that it is known that B occurs, is called the *conditional probability.* It is given by:

  $$P(A|B) = \frac{P(A \,\text{and}\, B)}{P(B)}$$

  If mutually exclusive, this will naturally be 0. If independent, this will reduce to $P(A)$.

- The above naturally leads to *Bayes' Theorem*, relating the conditional probabilities $P(A|B)$ and $P(B|A)$:

  $$P(B|A) = \frac{P(B)}{P(A)} P(A|B)$$

- Often there are several outcomes B_i which correspond to the outcome B (e.g. rolling a 4 corresponds to rolling an even number), and outcome A always occurs with one of the outcomes B_i. For example, A could be Arsenal winning, and the B_i could be the starting goalkeeper being David Raya, Aaron Ramsdale or Karl Hein. The following law - the *law of total probability* - is then sometimes useful:

  $$P(A) = \sum_i P(A|B_i)P(B_i)$$

  Sometimes it is useful to substitute this into the denominator of Bayes' Theorem.

1.3 Probability Distributions

For a random variable X which can take a range of values, we can define a function p which yields the probability of the outcome taking a particular value.

If the random variable can take any of a set of discrete values x_i, then the function $p = p(x_i)$, such that of course $\sum_i p(x_i) = 1$, is called a *probability mass function* (pmf).

If the random variable can take any real value between $x = L$ and $x = U$, then it makes no sense to speak of the probability of X taking a particular real value – that probability would be zero. It only makes sense to speak of the probability that X takes a value between x and $x + \mathrm{d}x$, which is written $p(x)\,\mathrm{d}x$, defining the *probability distribution function* (pdf) $p(x)$. We then have $\int_L^U p(x)\,\mathrm{d}x = 1$.

The cumulative distribution function (cdf) $F(x)$ is the probability that X takes a value less than x:

$$F(x) \equiv \int_L^x p(x)\,\mathrm{d}x\,; \qquad F(L) = 0; \qquad F(U) = 1; \qquad P(a < X < b) = F(b) - F(a)$$

Consider a joint pdf of two different random variables X and Y. If X and Y are discrete, the probability of outcome $(X, Y) = (x, y)$ is written $p(x, y)$. If X and Y are continuous, then $p(x, y)\,\mathrm{d}x\,\mathrm{d}y$ is the probability of X being between x and $x + \mathrm{d}x$ and Y being between y and $y + \mathrm{d}y$.

If X and Y are independent, then the joint pdf will factor out into a pdf for each variable: $p(x, y) = p_X(x)p_Y(y)$.

More generally, to access the overall probability distribution for one of the variables, say X, we need to account for all the possible values of Y:

$$p_X(x) = \int_{L_Y}^{U_Y} p(x, y)\,\mathrm{d}y$$

If independent, this reduces to the above case. This is referred to as the *marginal distribution* of X, and integrating out a subset of the random variables under investigation is called *marginalisation.* Analogously to the conditional probability, we can then write the probability of Y given X as the probability of both, out of the probability of X.

$$p(y|x) = \frac{p(x, y)}{\int p(x, y)\,\mathrm{d}y}$$

1.3.1 Change of variables

Suppose we know that $X \sim p(x)$, i.e. X is distributed with some known pdf $p_X(x)$. Suppose also that there is another random variable $M = M(X)$ which is a known function of X. How is M distributed?

Suppose first that $M(X)$ is one-to-one. In this case, if X is between x and $x + \mathrm{d}x$ (which has a probability of $p(x)\,\mathrm{d}x$), then M will be between $m = M(x)$ and $m + \mathrm{d}m = M(x + \mathrm{d}x) = m + (\mathrm{d}M/\mathrm{d}x)\,\mathrm{d}x$. So the probability of M being between m and $m + \mathrm{d}m$ is $p_X(x)\,\mathrm{d}x$. Thus

$$p_M(m)\,|\mathrm{d}m| = p_X(x)\,\mathrm{d}x \qquad \Rightarrow \qquad p_M(m) = p_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}m}\right| = \frac{p_X(x)}{|\mathrm{d}M/\mathrm{d}X\,|}$$

where the absolute sign accounts for the possibility of M being monotonically decreasing as well as increasing.

If $M(X)$ is *not* one-to-one, then for M to take the value m, X might take several values x. If $M(X) = X^2$, then both $X = \pm 2$ will lead to $M = 4$. We therefore need to account for all of those regions, and we instead have

$$p_M(m) = \sum_i \frac{p(x_i)}{|\mathrm{d}M/\mathrm{d}X\,|_{x_i}}$$

where the sum is over all the different values of x_i for which $M(x_i) = m$.

Extending the monotonic case to multiple dimensions, if we have multiple independent variables $\mathbf{X} \sim p(\mathbf{X})$, and multiple transformation functions $\mathbf{Y}(\mathbf{X})$, then the joint pdf for U and V will be

$$p(\mathbf{Y}) = \left|J\left(\frac{\mathbf{X}}{\mathbf{Y}}\right)\right| p(\mathbf{X}), \qquad J\left(\frac{\mathbf{X}}{\mathbf{Y}}\right)_{i,j} = \frac{\partial X_i}{\partial Y_i}$$

where we are using the inverse Jacobian of the transformation.

1.4 Properties of Distributions

1.4.1 Mean, Variance, and Moments

Consider the random variable $X \sim p(x)$. The expectation value of X, written μ or $E[X]$ or $\langle X \rangle$ depending on the context, is the mean of the distribution:

$$\mu \equiv \int x p(x)\,\mathrm{d}x$$

More generally, the expectation of any function $M(X)$ is

$$E[M(X)] = \int M(x) p(x)\,\mathrm{d}x$$

From this and the properties of integration we see that the expectation is a linear operator: $E[aX + bY] = aE[X] + bE[Y]$.

The spread of a distribution can be quantified by the variance $V[X] = \sigma^2$, or the standard deviation σ. $V[X]$ is the expectation of the squared deviation from the mean

$$V[X] \equiv E\big[(X - E[X])^2\big] = E\big[X^2\big] - E[X]^2 = E\big[X^2\big] - \mu^2$$

$$= \int (x-\mu)^2\, p(x)\,\mathrm{d}x = \int x^2\, p(x)\,\mathrm{d}x - \left(\int x\, p(x)\,\mathrm{d}x\right)^2$$

We can more generally define *algebraic and central moments* of the distribution by:

$$\mu_\ell \equiv E\big[X^\ell\big] \qquad\qquad \alpha_\ell \equiv E\Big[(X - E[X])^\ell\Big]$$

We see that $\mu_1 = \mu$ and $V[X] = \alpha_2$. Higher-order quantities include the *skew* γ and the *kurtosis* κ:

$$\gamma \equiv E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\alpha_3}{\alpha_2^{3/2}}; \qquad\qquad \kappa \equiv E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\alpha_4}{\alpha_2^2}$$

The *excess kurtosis* $\kappa - 3$ quantifies the tailedness compared to a Gaussian distribution, which has $\kappa = 3$.

1.4.2 Covariance and Correlation

For two random variables X and Y, we define the *covariance* of the two variables as

$$\operatorname{cov}[X,Y] \equiv E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Note that $\operatorname{cov}[X,X] = V[X]$. If we construct a *covariance matrix* $V_{xy} = \operatorname{cov}[X,Y]$, the diagonal terms will therefore be the variances of the individual random variables. If X and Y are independent, then the integral used to calculate $E[XY]$ will factor out into $E[X]E[Y]$, the covariance will be 0, and the covariance matrix will be diagonal.

A de-dimensionalised version of the covariance is the *correlation*, defined by:

$$\rho[X,Y] = \frac{\operatorname{cov}[X,Y]}{\sigma_X \sigma_Y}$$

The elements of the *correlation matrix* are then clear, including the diagonal components which will all be 1. The correlation is always between -1 and 1, ultimately due to the Cauchy-Schwarz inequality:

$$\begin{aligned}
\operatorname{cov}[X,Y]^2 &= \left(\iint (x-\mu_X)(y-\mu_Y)p(x,y)\,\mathrm{d}x\,\mathrm{d}y\right)^2 \\
&\le \int (x-\mu_X)^2 p(x,y)\,\mathrm{d}x \int (y-\mu_Y)^2 p(x,y)\,\mathrm{d}y = V[X]V[Y] \\
\Rightarrow |\operatorname{cov}[X,Y]| &\le \sigma_X\sigma_Y \qquad \Rightarrow \qquad |\rho[X,Y]| \le 1
\end{aligned}$$

where the relevant inner product is defined as $\langle X, Y\rangle \equiv E[X,Y]$.

1.4.3 Propagating Errors

If we know the variance of some random variables X_i, how can we calculate the variance of some function $Z(X_i)$? It can be shown that the variance is not a linear operator, but rather:

$$V[aX + bY + c] = a^2 V[X] + b^2 V[Y] + 2ab \operatorname{cov}[X,Y]$$

What if we have a more complicated, nonlinear function of the X_i? If we assume small deviations from some value (likely the *mode*), we can Taylor expand to approximate the function as linear about that point. We then have

$$V[Z(X,Y)] \approx \left.\frac{\partial Z}{\partial X}\right|_{X_0}^2 V[X] + \left.\frac{\partial Z}{\partial Y}\right|_{Y_0}^2 V[Y] + 2\left.\frac{\partial Z}{\partial X}\right|_{X_0}\left.\frac{\partial Z}{\partial Y}\right|_{Y_0} \operatorname{cov}[X,Y]$$

1.4.4 Convolutions

Consider two random variables X and Y which are independent: $p(x,y) = p_X(x)p_Y(y)$. We now wish to find the pdf of their sum $Z = X + Y$. It is easiest to approach this by first looking at the cumulative distribution for Z:

$$F(Z) \equiv P(X + Y < z) = \int_{-\infty}^{\infty} p_X(x)\,\mathrm{d}x \int_{-\infty}^{z-x} p_Y(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} p_Y(y)\,\mathrm{d}y \int_{-\infty}^{z-y} p_X(x)\,\mathrm{d}x$$

The pdf is then found by differentiating with respect to z:

$$p_Z(z) \equiv \frac{\mathrm{d}F}{\mathrm{d}z} = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)\,\mathrm{d}x = \int_{-\infty}^{\infty} p_Y(y)p_X(z-y)\,\mathrm{d}y$$

which is just the convolution: $p_Z = p_X \otimes p_Y$.

1.4.5 Characteristic Function

For a continuous random variable $X \sim p(X)$, the *characteristic function* $\varphi(t)$ is simply the Fourier transform of the probability distribution function:

$$\varphi(t) \equiv \int_{-\infty}^{\infty} e^{itx}\, p(x)\,\mathrm{d}x = E\big[e^{itX}\big]$$

The probability distribution $p(x)$ can be recovered using the inverse transform: $p(x) = \frac{1}{2\pi}\int e^{-itx}\varphi(t)\,\mathrm{d}t$.

The characteristic function is useful for several reasons. For example, it encodes all of the moments of the distribution:

$$\varphi(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!} E[X^n] = \sum_{n=0}^{\infty} \frac{(it)^n}{n!}\mu_n$$

which can be individually extracted by differentiation at $t = 0$:

$$\left.\frac{\mathrm{d}^n\varphi}{\mathrm{d}t^n}\right|_0 = i^n \mu_n$$

This is useful because sometimes it is easier to find the moments of a distribution by finding the characteristic function and differentiating, than by integrating $x^n p(x)$.

1.5 Common Probability Distributions

1.5.1 Binomial

Consider a fixed number of trials n, where in each trial the outcome is either success (with probability p) or failure (with probability $q = 1 - p$). The binomial distribution gives the probability of achieving exactly k successes out of these n trials. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ different ways of achieving k successes out of n, each of which has probability $p^k(1-p)^{n-k}$, so the probability that there are k successes is given by:

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

The sum of these probabilities is the binomial expansion of $(p + (1-p))^n$, which is 1 as required. The mean number of successes can be shown to be $E[k] = np$, and we can also find $E[k^2] = n(n-1)p + np$, and hence $V[k] = np(1-p)$ and $\sigma_k = \sqrt{np(1-p)}$.

1.5.2 Poisson

The Poisson distribution counts the number of independent events occurring in a given time period, assuming a constant mean number of occurrences λ. The distribution is therefore defined for any non-negative integer up to infinity. An example is the number of cosmic rays reaching a detector in an hour: it is unlikely but technically possible for 10^{90} cosmic rays to hit the detector in an hour; it is possible for 0 cosmic rays to hit it; it is impossible for -1 cosmic rays to hit it.

The Poisson distribution is derived from the binomial distribution. We can think of each n infinitesimal time intervals $\delta t = T/n$ each as a trial, which can either be a success (cosmic ray detected) with a very small probability $p = \lambda/n$, or a failure with probability $1-p$. Letting n tend to infinity but $np = \lambda T$ remain finite, the binomial distribution becomes

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

from which we obtain the Poisson distribution

$$P(k;\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

It can be shown that the mean of the Poisson distribution is $E[k] = \lambda$, and $E[k^2] = \lambda(\lambda+1) \Rightarrow V[k] = \lambda \Rightarrow \sigma = \sqrt{\lambda}$.

An important use case of the Poisson distribution is in histograms. Each bin of a histogram counts the number of events that go into that bin, and if these are independent and come in at a uniform rate then the number in each bin is Poisson-distributed, with λ in that case being the number in that bin. The error bars on histogram bin heights (σ) should therefore be the square root of the height of the bin.

1.5.3 Normal

For the *standard* normal distribution, $p(x) = e^{-x^2/2}/\sqrt{2\pi}$. It can easily be shown that this distribution has $E[X] = 0$, $E[X^2] = 1$, and hence $V[X] = 1$ and $\sigma = 1$. By performing the transformation $x \mapsto (x-\mu)/\sigma$, we obtain the general normal distribution with a mean μ and standard deviation σ:

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

The cdf of the normal distribution, $\Phi(x)$, is usually expressed in terms of the error function $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\,\mathrm{d}t$. We find

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$$

It is useful to extend the normal distribution to many variables $\mathbf{X} = \{X_i\}$, some of which may be correlated. Generalising the above, we find:

$$p(\mathbf{x};\boldsymbol{\mu},\mathbf{V_X}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{V_X}|}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal\,\mathbf{V}_\mathbf{X}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right]$$

where $\mathbf{V_X}$ is the covariance matrix of the X_i.

1.5.4 Chi-squared

The chi-squared distribution with k degrees of freedom is the distribution of the sum of k independent standard normal random variables. That is, if $X_i \sim N(\mu = 0, \sigma^2 = 1)$, then the overall random variable $X = \sum_i X_i^2$ will be χ^2-distributed. This distribution turns out to be:

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

It can be shown that $E[X] = k$ and $V[X] = 2k$.

Suppose we have 2 degrees of freedom, X_1 and X_2, both of which are standard normal distributed. For each of these, there is a 68.3% probability that they are between -1 and 1, i.e. within 1σ of μ (where σ and μ refer to the individual normal distributions here). However, it is not true to say that the probability that $\sqrt{X_1^2 + X_2^2}$ is within σ of $\boldsymbol{\mu}$ is 68.3%. It is not even true to say that the probability is $0.683^2 = 0.466$, as one might intuitively think. We are integrating over a disk $X_1^2 + X_2^2 \leq 1$, so the probability is instead

$$P(X_1^2 + X_2^2 < 1) = \int_{-1}^{1} p(x_1)\,\mathrm{d}x_1 \int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} p(x_2)\,\mathrm{d}x_2 = \sqrt{\frac{2}{\pi}} \int_0^1 e^{-x_1^2/2}\,\mathrm{erf}\left(\sqrt{\frac{1-x_1^2}{2}}\right) \mathrm{d}x_1$$

which comes out to about 0.393. Replacing the 1s in the integration bounds with a^2 tells us the probability of $\sqrt{X_1^2 + X_2^2}$ being within $a\sigma$ of the origin $X_1 = X_2 = 0$.

1.6 Generating Samples

There are two main ways of generating samples from a known distribution.

1.6.1 Accept-Reject

The simplest method is the *accept-reject* method. First, find the maximum value of the pdf $p_{\max}$; we do not require the distribution to be normalised for this method to work. Then, generate two random numbers, x_i and y_i, from the uniform distributions $[L, U]$ and $[0, p_{\max}]$. Calculate the value of the distribution at x_i, $p(x_i)$. The sample x_i is *accepted* if $p(x_i) < y_i$, and rejected otherwise. This means that x_i is proportionally more likely to be accepted if it is in regions where the pdf is higher, and as a result the accepted x_i are samples selected from the distribution. Figure 1 demonstrates the accept-reject method for a binomial distribution, though this method works equally well for continuous distributions.

1.6.2 Inverse CDF

The cdf for a random variable X is a mapping from the range which X can come from, to the interval $[0, 1]$. Conversely, the *inverse* cdf, known as the *percentage point function* (ppf), is a mapping from $[0, 1]$ to the range of possible outcomes for X. The gradient of the cdf is the pdf, so the cdf will be steeper for more probable x. The ppf will therefore be flattest where it takes the distribution's most probable values. Figure 2 shows this for a normal distribution with $\mu = 2, \sigma = 3$. The ppf is flattest around a value of 2, which is the most likely region of outcomes in the pdf.

We can generate samples from a distribution in the following way. Sample first from a *uniform* distribution on $[0, 1]$. Then feed those uniform samples into the ppf of the distribution

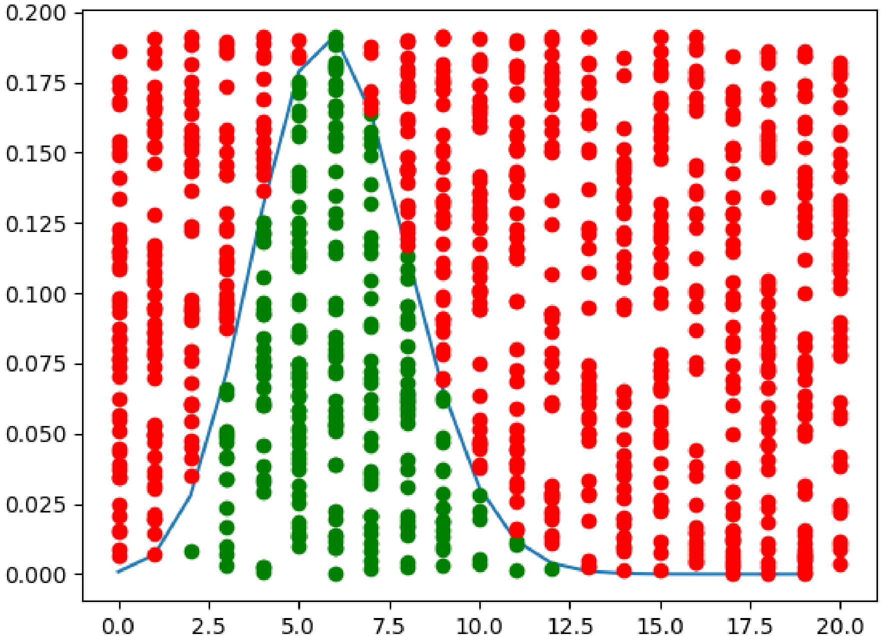

Figure 1: The accept-reject method for generating samples from a binomial distribution, with $n = 20$ and $p = 0.3$.

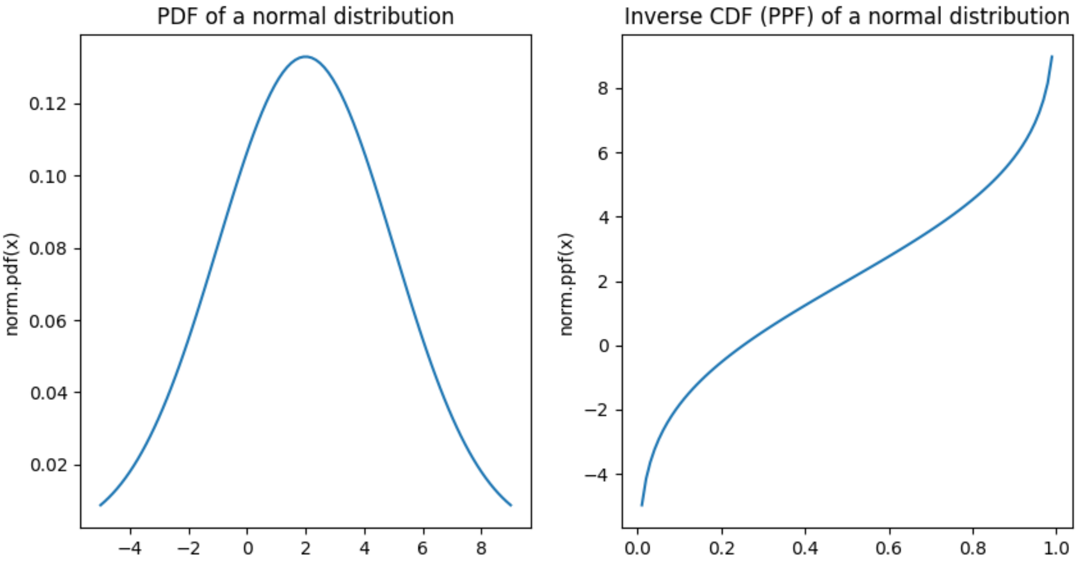

Figure 2: The cdf and pdf of a normal distribution, and the corresponding ppf.

in question. This will yield random samples from the distribution, because more of the uniform samples will be around the flatter regions of the cdf – that is, the most likely regions of the pdf.

This method does however require knowing the ppf, which is sometimes difficult to obtain, especially analytically.

1.7 Central Limit Theorem

For a series of samples of size N independently drawn from any distribution, the distribution of the sample sum, and hence the sample mean, tends towards a normal distribution as $N \to \infty$..

Proof. Consider a set of N random variables $\{X_i\}$ drawn from some distribution with mean μ and standard deviation σ. The mean is $\bar{X} = \sum_i X_i/N$. Consider the variable

$$Y = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_i - \mu}{\sigma}$$

It can easily be shown that $E[Y] = 0$ and $V[Y] = 1$. The characteristic function of Y is $\varphi_Y(t) = E\big[e^{itY}\big]$. and

$$e^{itY} = \exp\left(i\frac{t}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_i - \mu}{\sigma} \right) = \prod_{i=1}^{N} \exp\left(i\frac{t}{\sqrt{N}} \frac{X_i - \mu}{\sigma} \right) = \exp\left(i\frac{t}{\sqrt{N}} \frac{X - \mu}{\sigma} \right)^N$$

where in the final step we use the fact that the X_i are identically distributed. The exponential is then the characteristic function of $Z = (X - \mu)/\sigma$, stretched by a factor $\sqrt{N}$. Now for this variable, clearly $E[Z] = 0$ and $V[Z] = 1$, so $E[Z^2] = 1$. As such, the characteristic function for this variable is given by $\varphi_Z(t) = 1 + it(0) + \frac{1}{2}(it)^2(1) + \mathcal{O}(t^3) = 1 - \frac{1}{2}t^2 + \mathcal{O}(t^3)$. The characteristic function of Y is then given by:

$$\varphi_Y(t) = \left(1 - \frac{t^2}{2N} + \mathcal{O}\left(\frac{t^3}{N^{3/2}}\right)\right)^N = 1 - \frac{1}{2}t^2 + \mathcal{O}\big(N^{-1/2}\big) = e^{-t^2/2} + \mathcal{O}\big(N^{-1/2}\big)$$

which of course tends to $e^{-t^2/2}$ as $N \to \infty$. Now for the standard normal distribution, the characteristic function is

$$\int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx - \frac{1}{2}x^2}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-it)^2}\,\mathrm{d}x = e^{-t^2/2}$$

Now two distributions with the same characteristic function are the same, because the characteristic function can be inverted to give the original distribution. Thus Y is standard-normal distributed. Thus $\bar{X}$ is normal distributed with mean μ and variance σ^2/N. ■

2 Estimating Parameters

When repeating an experiment, we are sampling from an unknown distribution, about which we can only make inferences based on the sample. This chapter looks at how to do that.

Estimators are functions used to estimate parameters θ of a distribution $p(x;\theta)$. Estimators are functions of a sample $\{X_i\}$ drawn from that distribution: $X_i \sim p(x;\theta)$. An estimator of a parameter θ is denoted $\hat{\theta}$. Ultimately, an estimator can be *any* function of the sample, but some functions will be better than others. Three ideal properties of an estimator are:

- **Consistency**. An estimator is consistent if, as the sample size approaches infinity, the estimator converges on the true value:
$$\lim_{N\to\infty} \hat{\theta} = \theta$$
- **Unbiasedness**. The bias of an estimator is the deviation of the expectation of the estimator from the true value: $b[\hat{\theta}] = E[\hat{\theta} - \theta]$. An estimator is unbiased if $b = 0$ for any N. An estimator can be biased yet consistent if $\lim_{N\to\infty} b[\hat{\theta}] = 0$
- **Efficiency**. If we use $\hat{\theta}$ to estimate θ on several samples, it is likely that $\hat{\theta}$ will vary around the true value. An estimator is said to be efficient if its variance is as low as possible. More precisely, there is a particular value of the minimum variance of an estimator (called the *minimum variance bound*, discussed later), and an estimator is called efficient if its variance is equal to that minimum possible value

A good estimator of the mean of a distribution is to use the arithmetic mean of a sample:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

This estimator is consistent[1], unbiased ($E[\hat{\mu}] = \mu$), and turns out to be efficient, as its variance $V[\hat{\mu}] = \sigma^2/N$ turns out to be the minimum possible variance.

Estimating the variance of the sample, σ^2, is more nuanced. Consider first the case where we know the true mean μ. The estimator $\hat{V} = \frac{1}{N}\sum(X_i - \mu)^2$ is then unbiased, as $E[\hat{V}] = \frac{1}{N}\sum E\big[(X_i - \mu)^2\big] = \sigma^2$.

What if we do not know the true mean – how can we then estimate the sample variance? We might think that a good estimator is $\hat{V} = \frac{1}{N}\sum(X_i - \hat{\mu})^2 = \frac{1}{N}\sum(X_i^2 - \hat{\mu}^2)$, where we now use our estimator for the mean since we don't know the true mean. However, this estimator underestimates the true variance, as $\hat{\mu}$ is estimated from the sample itself, so the values in the sample will naturally be unduly close to $\hat{\mu}$, by design of $\hat{\mu}$:

$$\begin{aligned} E[\hat{V}] &= \frac{1}{N}\sum\big(E\big[X_i^2\big] - E[\hat{\mu}^2]\big) \\ &= \frac{1}{N}\sum\big(V[X_i] + E[X_i]^2 - V[\hat{\mu}] - E[\hat{\mu}]^2\big) \\ &= \frac{1}{N}\sum\left(\sigma^2 + \mu^2 - \frac{\sigma^2}{N} - \mu^2\right) = \sigma^2\left(1 - \frac{1}{N}\right) \end{aligned}$$

[1] Strictly, this assumes *convergence in distribution* – that is, as $N \to \infty$ the sample approximates the true distribution

which underestimates the true variance by a factor $(N-1)/N$, and is hence biased. Note that as $N \to \infty$, this factor tends to 1, so this estimator is at least consistent. We can make an unbiased estimator by simply dividing by this factor:

$$\hat{V} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu})^2$$

The correction from N to $N-1$ is sometimes called *Bessel's correction*, and this formula is often called the sample variance $\hat{s}^2$.

This estimate will not be completely accurate every time, and until $N \to \infty$ will itself have some variance $V[\hat{V}]$. This can be derived assuming that the deviations $X_i - \hat{\mu}$ are normally distributed about 0 with standard deviation σ. In this case, $\hat{V}$ is proportional to a sum of squares of normal random variables, and therefore follows a scaled χ^2 distribution, with $N-1$ degrees of freedom (since we have used the sample to derive the quantity $\hat{\mu}$). χ^2 distributions from *standard*-normally distributed variables have a variance of $2k$ for k degrees of freedom. Hence the sum here has a variance of $2(N-1)\sigma^4$. Accounting for the factor of $1/(N-1)$ at the front of the sum, the variance of this estimator is thus:

$$V[\hat{V}] = \frac{2\sigma^4}{N-1}$$

That this goes to 0 as $N \to \infty$ makes $\hat{V}$ a consistent estimator.

An estimator for the standard deviation is $\hat{\sigma} = \sqrt{\hat{V}}$. The variance of this estimator can be calculated using the error propagation formula:

$$V[\hat{\sigma}] = \left(\frac{\partial \hat{\sigma}}{\partial \hat{V}}\right)^2 V[\hat{V}] = \left(\frac{1}{2}\,\hat{V}^{-1/2}\bigg|_{\hat{\sigma}=\sigma}\right)^2 \frac{2\sigma^4}{N-1} = \frac{1}{4}\frac{1}{\sigma^2}\frac{2\sigma^4}{N-1} = \frac{\sigma^2}{2(N-1)}$$

and so the error on the estimated standard deviation is $\sigma/\sqrt{2(N-1)}$.

2.1 Likelihood

When inferring parameters $\boldsymbol{\theta}$ of a distribution $p(x;\boldsymbol{\theta})$ from a sample $\{X_i\}$, it is natural to think about how likely that sample was to occur for various different values of $\boldsymbol{\theta}$. This will be a function of the parameters $\boldsymbol{\theta}$, and is called the *likelihood function*. It is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(X_i;\boldsymbol{\theta}) = p(\mathbf{X};\boldsymbol{\theta})$$

where $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$. The likelihood is not technically a pdf, as the parameters $\boldsymbol{\theta}$ are not, strictly speaking, random variables.

A natural way of estimating the parameters $\boldsymbol{\theta}$ is to find the $\boldsymbol{\theta}$ which maximises the likelihood function $\mathcal{L}$ – this is called the *maximum likelihood method*. Dropping to one parameter θ for notational convenience, this means that the estimator $\hat{\theta}$ is the value of θ where $\partial\mathcal{L}/\partial\theta = 0$. The product above sometimes makes the differentiation awkward (and computationally can result in stupidly small numbers for large samples), so we typically instead minimise $\ln \mathcal{L}$.

2.1.1 Minimum Variance Bound

The *Fisher score* $S(\theta)$ is the partial derivative

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

The score is therefore 0 at the maximum-likelihood estimate $\theta = \hat{\theta}_{\mathrm{ML}}$. We can show that $\hat{\theta}_{\mathrm{ML}}$ is an unbiased estimator by showing that the expected value of $S(\theta)$ is 0:

$$\begin{aligned} E[S(\theta)] &= \int \frac{\partial \ln p(\mathbf{X};\theta)}{\partial \theta} p(\mathbf{X};\theta)\,\mathrm{d}\mathbf{X} = \int \frac{1}{p(\mathbf{X};\theta)} \frac{\partial p(\mathbf{X};\theta)}{\partial \theta} p(\mathbf{X};\theta)\,\mathrm{d}\mathbf{X} \\ &= \frac{\partial}{\partial \theta} \int p(\mathbf{X};\theta)\,\mathrm{d}\mathbf{X} = \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

where we assume that the distribution is normalised for all values of θ.

The *Fisher information* $I(\theta) = V[S(\theta)]$ is the variance of the Fisher score. We just showed that $E[S] = 0$, so we have

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

which, incidentally, is non-negative. This can be rewritten using the lemma:

$$\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2 = \frac{1}{\mathcal{L}} \frac{\partial^2 \mathcal{L}}{\partial \theta^2} - \frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}$$

and the fact that

$$E\left[\frac{1}{\mathcal{L}} \frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right] = \int \frac{1}{p(\mathbf{X};\theta)} \frac{\partial^2 p(\mathbf{X};\theta)}{\partial \theta^2} p(\mathbf{X};\theta)\,\mathrm{d}\mathbf{X} = \frac{\partial^2}{\partial \theta^2} \int p(\mathbf{X};\theta)\,\mathrm{d}\mathbf{X} = \frac{\partial^2}{\partial \theta^2} 1 = 0$$

Hence

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

The Fisher information turns out to be the inverse of the minimum variance of $\hat{\theta}_{\mathrm{ML}}$:

$$\Rightarrow V\left[\hat{\theta}_{\mathrm{ML}}\right] \geq I(\theta)^{-1} = E\left[\left(\frac{\partial \ln \mathcal{L}}{\partial \theta}\right)^2\right]^{-1} = -E\left[\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2}\right]^{-1}$$

Apparently as $N \to \infty$, this tends to an equality. A general estimator $\hat{\theta}$ is said to be *efficient* if it has a variance equal to this minimum. Also apparently, in the $N \to \infty$ limit, the estimate will tend to a normal distribution, centred on the true value θ_*, and with the above variance:

$$\hat{\theta}_{\mathrm{ML}} \sim \frac{1}{\sqrt{2\pi V_{\mathrm{MVB}}}} \exp\left[-\frac{\left(\hat{\theta}_{\mathrm{ML}} - \theta_*\right)^2}{2V_{\mathrm{MVB}}}\right]$$

where V_{MVB} is the minimum variance bound above. As such, if we calculate $\hat{\theta}_{\mathrm{ML}}$ for lots of samples from the same distribution, we will find a normal distribution. Typically we instead look at $-2\ln\mathcal{L}$, which will look like:

$$-2\ln\mathcal{L} = \ln\left(2\pi V_{\mathrm{MVB}}\right) + \frac{\left(\hat{\theta}_{\mathrm{ML}} - \theta_*\right)^2}{V_{\mathrm{MVB}}}$$

which is a quadratic in $\hat{\theta}_{\mathrm{ML}}$. This quantity – rather than $\mathcal{L}$ itself – is useful for calculating uncertainties: to find the 1σ error, one can simply find where $-2\ln\mathcal{L}$ goes above its minimum value by 1; to find the 2σ error, look at where it goes above its minimum value by 4; etc.

For multiple dimensions, the covariance of two ML-estimated parameters tends to

$$\operatorname{cov}(\hat{\theta}_{\mathrm{ML},i}, \hat{\theta}_{\mathrm{ML},j}) \to -E\left[\frac{\partial^2 \ln\mathcal{L}}{\partial\theta_i \partial\theta_j}\right]$$

and tends to a multivariate normal distribution as $N \to \infty$.

2.2 Variations on the Likelihood

2.2.1 Profile Likelihood

If we want a 1-dimensional likelihood function for a single parameter λ in a multiparameter distribution, for example to quote 1D errors, this becomes more difficult if the likelihood covaries with multiple parameters, i.e. if the above is non-zero. In such a case, one *could* slice through the multidimensional Gaussian (to get the conditional probability), or integrate over the other dimensions (marginal probability), but apparently to preserve the property of $-2\ln\mathcal{L}$ going up by 1 at $\pm 1\sigma$, up by 4 at $\pm 2\sigma$ etc., one should instead take the *profile* likelihood function $\mathcal{L}^*(\lambda)$. For each value of λ, one minimises the full likelihood $\mathcal{L}$ with respect to all the other parameters, subject to this value of λ being fixed. Naturally, for the maximum-likelihood value of $\lambda = \lambda_{\mathrm{ML}}$, the profile likelihood $\mathcal{L}^*$ will be at a maximum, otherwise we will have found another maximum to the overall likelihood function $\mathcal{L}$. And for the value of λ where $-2\ln\mathcal{L}^* = \min(-2\ln\mathcal{L}^*) + n^2 = \min(-2\ln\mathcal{L}) + n^2$, we will have found a value of λ which is $\pm n\sigma_\lambda$ away from the maximum-likelihood value but where all the other parameters nonetheless maximise the likelihood as far as possible.

2.2.2 Extended Maximum Likelihood

It is possible, particularly in particle-physics-based counting experiments, that the number of observations N is a (Poisson-distributed) random variable that is effectively part of the experiment – we don't know it in advance, and the mean number of events ν is another parameter. We have $p(N;\nu) = \nu^N e^{-\nu}/N!$, so the total likelihood $p(\mathbf{X}, N; \boldsymbol{\theta})$ becomes:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{\nu^N e^{-\nu}}{N!} \prod_{i=1}^{N} p(X_i; \boldsymbol{\theta})$$

where the parameters $\boldsymbol{\theta}$ now include ν.

2.2.3 Binned Maximum Likelihood

For a very large sample size, the product in the definition of $\mathcal{L}$ (or the sum in $\ln\mathcal{L}$) will get outrageously large. In such cases, it is quicker to bin the samples X_i, into bins $[x_{Lb}, x_{Ub}]$, where b runs from 0 to B, the number of bins. In this way, the "samples" are no longer the X_i, but the N_b, the number of X_i which are in the bin b; we have $\sum_b N_b = N$. These samples are Poisson distributed, about a mean λ_b given by the expected number of events in that bin:

$$\lambda_b = \int_{x_{Lb}}^{x_{Ui}} p(X; \boldsymbol{\theta})\, \mathrm{d}X = F(x_{Ui}) - F(x_{Li})$$

which we note depends on the parameters $\boldsymbol{\theta}$. The binned likelihood function is then:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{b=1}^{B} \frac{\lambda_b^{N_b} e^{-\lambda_b}}{N_b!}$$

where the dependence on $\boldsymbol{\theta}$ lies in λ_b. The advantage of the binned likelihood function is that the product is now over only B terms rather than $N \gg B$, and calculation of maximum likelihood parameters can be sped up by significant factors. A disadvantage is that it can be difficult to calculate λ_b, as cdfs can be hard to calculate. One can estimate the cdf by numerically integrating, or estimating the integral as the pdf's value at the bin centre multiplied by the bin width, though this will lead to a bias.

2.3 Least-Squares Estimation

The least-squares method of estimation applies to fitting a parametrised curve to some (X_i, Y_i) data. For some model $y = f(x; \boldsymbol{\theta})$, we wish to minimise:

$$\chi^2 = \sum_{i=1}^{N} (Y_i - f(X_i; \boldsymbol{\theta}))^2$$

It is likely that the Y_i values have some uncertainties σ_i, and to be more accepting of deviations from the model value we weight each contribution by σ_i^{-2}, minimising instead:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{Y_i - f(X_i; \boldsymbol{\theta})}{\sigma_i} \right)^2$$

We can show that minimising χ^2 is the same as maximising the likelihood function of some normally-distributed data $Y_i \sim N(f(X_i; \boldsymbol{\theta}), \sigma_i^2)$. For such data, the likelihood is given by:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2} \left(\frac{Y_i - f(X_i; \boldsymbol{\theta})}{\sigma_i} \right)^2 \right]$$

$$\Rightarrow -2 \ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(\frac{Y_i - f(X_i; \boldsymbol{\theta})}{\sigma_i^2} \right)^2 + \text{const.} = \chi^2 + \text{const.}$$

Now because if we are adding lots of data together the resulting quantity tends to a Gaussian distribution, this means that for *any* large dataset, $-2\ln\mathcal{L}(\boldsymbol{\theta})$ approaches a χ^2 distribution (albeit offset by a constant). The number of degrees of freedom is the number of observations in the χ^2 sum minus the number of free parameters being fitted ($\dim \boldsymbol{\theta}$):

$$k = N - \dim \boldsymbol{\theta}$$

Now the mean of a χ^2 distribution with k degrees of freedom is k, so we expect a χ^2 value of roughly k. If much higher, the model is a poor fit to data; if much lower, we are overfitting.

An broad example use case is where we have a large sample $\{X_i\}$ which we put into bins b, fitting to a distribution $y_b = f(x_b; \boldsymbol{\theta})$, where x_b is the x-value in the centre of bin b, for example. In this case, the data Y_b are the number of events X_i that are in bin b, and because binning is effectively a Poisson process the errors $\sigma_b = \sqrt{Y_b}$.

2.4 Method of Moments

The method of moments may be a quicker approach than maximum-likelihood or least-squares estimation. It estimates parameters using the fact that, for large samples, sample statistics will tend to their true values. For example, the mean of a distribution with given parameters $\boldsymbol{\theta}$ is $\mu(\boldsymbol{\theta}) = \int X f(X;\boldsymbol{\theta})\,\mathrm{d}X$. For a large sample size, we can calculate the sample mean $\bar{X}$ and say that this will be close to $\mu(\boldsymbol{\theta})$. If the model only has one parameter ($\boldsymbol{\theta} = \theta$), then we can simply invert this to estimate

$$\hat{\theta}_{\mathrm{MOM}} = \mu^{-1}(\bar{X})$$

where the function $\mu(\theta)$ comes from the parametrised integral and $\bar{X}$ comes from the sample.

For the case of multiple parameters, we include higher-order moments. We find $\mu_2(\boldsymbol{\theta}) = \int X^2 f(X;\boldsymbol{\theta})\,\mathrm{d}X$, and the sample estimate $\hat{\mu}_2$ (for example, $\bar{X^2}$), giving two equations in $\boldsymbol{\theta}$; if $\dim\boldsymbol{\theta} = 2$, these can be solved to give θ_1 and θ_2. If $\boldsymbol{\theta}$ is more parameters, we keep going to higher orders until we find an invertible set of equations to find $\boldsymbol{\theta}_{\mathrm{MOM}}$.

Alternatively, rather than using higher order moments, we can use higher order *central* moments, for example σ^2.

2.5 Confidence Intervals

If the probability that the true value of θ lies between θ_L and θ_U is γ, then $[\theta_L, \theta_U]$ is a *confidence interval* of θ with confidence level γ. Often we choose $\gamma = 0.683$ or 0.954, corresponding to 1σ or 2σ from the mean of a Gaussian distribution. In a frequentist sense, we are saying that the true value θ_0 (which exists, somewhere!) is not known to us but we know where it could be and with what probability. If the true value is really within the confidence interval in a fraction γ of experiments, the interval is said to *cover*; otherwise to *overcover* or *undercover.* Overcoverage is a confidence interval which is too large, being more conservative, but failing to claim discoveries where they might exist and hence losing statistical power. Undercoverage is more dangerous, as it can lead to false discoveries.

Confidence intervals can be constructed in many different ways. Perhaps the most natural is the *Neyman-Pearson interval.* Suppose we are estimating a parameter θ from a single datum X. We first construct the likelihood, which here is just $p(X|\theta)$. For each θ, we then find the values of X_L and X_U such that

$$p(X < X_L|\theta) \equiv \int_L^{X_L} p(X|\theta)\,\mathrm{d}X = \quad \frac{1-\gamma}{2} \quad = \int_{X_U}^{U} p(X|\theta)\,\mathrm{d}X \equiv p(X > X_U|\theta)$$

where L and U are the lowest and highest values that X can take. In other words, assuming that value of θ, X is equally likely to be below X_L as above X_U, and has a probability γ of being in between. In the space of X and θ, we can scan along the θ axis and construct this $[X_L, X_U]$ range (called the *acceptance region*) for each value of θ, drawing out a *confidence belt* in X-θ space. Now, when we get some data X_0, we flip this over, and the confidence interval for θ is the intersection of $X = X_0$ and the confidence belt constructed above.

2.5.1 Intervals near Physical Boundaries: the Feldman-Cousins Interval

Suppose θ is some sort of mass. We know $\theta > 0$, but if we come to the conclusion that $\theta = 2 \pm 3$, this is kind of unsatisfying at the lower end. Conversely, if we cut things off at 0, we might write $\theta = 2^{+3}_{-2}$, but then this does not cover: θ will be between 0 and 5 in less than a fraction γ of experiments. This might emerge from a likelihood function $\mathcal{L}(\theta)$ being not very different at 0 than at $\hat{\theta}_{\mathrm{ML}}$, such that the parabola of $-2\ln\mathcal{L}$ around $\hat{\theta}_{\mathrm{ML}}$ might not reach 1 at all between $\theta = 0$ and $\theta = \hat{\theta}_{\mathrm{ML}}$. How would we then give a lower 1σ error on θ?

The *Feldman-Cousins* interval resolves this, constructing a confidence belt as follows:

1. For each X, calculate the maximum-likelihood value of the parameter, $\hat{\theta}_{\mathrm{ML}}$. It may be that this value is outside the boundary, in which case set the best estimate to the boundary edge: if we require $\theta > 0$ but for some X we have $\hat{\theta}_{\mathrm{ML}} < 0$, set instead $\hat{\theta} = 0$. This will draw out a maximum-likelihood-curve in X-$\hat{\theta}$ space, which may hug the boundary for some values of X.

2. Flipping things over, consider a fixed value of θ. Along that line, calculate as a function of X, the quantity
$$R = \frac{p(X|\theta)}{p(X_{\theta=\hat{\theta}}|\theta)}$$
where $X_{\theta=\hat{\theta}}$ is the value of X at which, for this value of θ, $\theta = \hat{\theta}$: that is, the point on the maximum-likelihood-curve at this value of $\hat{\theta}$. This quantity R will of course be equal to 1 at $X = X_{\theta=\hat{\theta}}$, but will generally be less than 1, as other values of X are by definition less probable. Indeed, R should monotonically decrease in both directions, away from $X = X_{\theta=\hat{\theta}}$.

3. A range in X is constructed around $X_{\theta=\hat{\theta}}$ for a given θ as follows. It is most intuitive to think in the discrete-X case, but the continuous case naturally follows.

   (a) Take the two X-values on either side of $X_{\theta=\hat{\theta}}$, as the initial interval.

   (b) Then, check the two points just outside the interval: expand the interval to include whichever has the highest R.

   (c) Then check the two points just outside the new interval (one of which will have previously been rejected), and again include whichever has the higher R.

   (d) Continue expanding the interval in this way, until we achieve the required coverage, finding X_L and X_U such that
   $$\int_{X_L}^{X_U} p(X|\theta)\,\mathrm{d}X = \gamma$$

In this way, a confidence belt is again constructed in X-θ space, and hence confidence intervals can be calculated for any data X that come in. The result is a smooth confidence belt, which covers the required confidence interval essentially by definition. The Feldman-Cousins intervals are, however, usually quite slow to calculate.

3 Goodness of Fit and Hypothesis Testing

Once we have fit the parameters of a model to our data, we wish to *evaluate* the fit, quantifying how well the model fits the data, perhaps in comparison to another model. This is the job of *test statistics*, which are functions T of the model and the data. An example of a test statistic is the χ^2 statistic, which is smaller the better-fitting the model is.

Although T provides a quantifier of the goodness of fit of a model to data, the quantity itself is essentially meaningless if it is not known how T is distributed. For any test statistic T, another equally good test statistic is $T + 17$, but neither really *means* anything in isolation. If it is known how T is distributed, we can work out how likely it was that our data and model gave the value it did (or of something "more extreme"), converting the arbitrary scale of T's distribution into a probability. If we have some hypothesis H, corresponding to a particular model attempting to describe a dataset, then if H is true we can write the probability of obtaining a test statistic $T = T_0$ or more extreme as:

$$p = \int_{T_0}^{\infty} P(T|H)\,\mathrm{d}T$$

where we assume "more extreme" for this test statistic corresponds to larger values of T. This is called the p value. If p is very small, then the model poorly describes the data, and a different model is probably better. If p is very close to 1, we're probably overfitting.

Take $T = \chi^2$ as an example, suppose we have a model and dataset which give $\chi^2 = \chi_0^2$. How well does our model fit the data? In other words, what would be the probability of obtaining this good a fit, or worse? Well it would be

$$p = \int_{\chi_0^2}^{\infty} p(\chi^2; k)\,\mathrm{d}\chi^2 = 1 - F(\chi_0^2; k)$$

If p is very low, then the χ_0^2 is quite high, suggesting that the fit is not very good. If p is close to 1, then χ_0^2 is close to 0, and the fit is suspiciously good. If we have a model H which accurately describes the data, we would have $\chi_0^2 \approx k$, so we would have a p-value of $1 - F(k; k)$, which seems to be between about 0.3 and 0.5 depending on the number of degrees of freedom. We see that for a χ^2 test, even though H accurately describes the data the p-value is not particularly low: certainly not low enough to claim some kind of discovery. This speaks to the low *statistical power* of the χ^2 test statistic – even for a model which accurately describes the data, the natural tendency for data to vary (and hence each contribution to the χ^2 to be on average 1) means that we would probably not be able to detect a real effect. Better statistical power comes from comparison of two hypotheses.

3.1 Hypothesis Tests

This typically involves considering a *null hypothesis* H_0, which hypothesises that the data occurred due to chance alone: the "default" option. This is in contrast to the *alternative hypothesis* H_1, which states that the data occurred due to some new, interesting effect. To judge which of these hypotheses to accept or reject, we employ a test statistic T whose distribution $P(T)$ is known under both H_0 and H_1, being $P(T|H_0)$ and $P(T|H_1)$: the null and alternate distributions. We can draw out these two distributions, which will generally have some overlap. When we get some data X, we feed that into the test statistic under the assumption of H_0, to obtain a test statistic T_0.

Suppose that the alternate hypothesis corresponds to larger test statistics, so that the alternate distribution lies to larger T than the null distribution. We can then find the probability of having obtained a test statistic at least as extreme as T_0 under the null hypothesis, which would be:

$$\alpha = \int_{T_0}^{\infty} P(T|H_0)\,\mathrm{d}T$$

If the data are entirely consistent with H_0, we will find $\alpha \approx 1$; if something very un-H_0-y has happened, then we will see a very small α. α is also the probability that we would unjustly reject H_0, i.e. the probability that we see a very high test statistic and think that the null hypothesis is false (a Type I error; loss).

Instead, we might see quite a low T_0 and unjustly *accept* H_0 despite it not being true and H_1 being a better description of the data. The probability of this happening is

$$1 - \beta = \int_{-\infty}^{T_0} P(T|H_1)\,\mathrm{d}T$$

(This is a Type II error; contamination).

What makes an ideal test statistic? We want a test statistic to be able to distinguish between two hypotheses, so in an ideal world there would be no overlap between the null and alternate distributions, and the measured value of T_0 would tell us instantly which hypothesis is true. In reality, the measured value of T_0 will probably somewhere in the overlap between $P(T|H_0)$ and $P(T|H_1)$, but to reduce the probability of this happening it would be nice if our two distributions had as little overlap as possible. The best test statistics separate the null and alternate distributions as much as possible. For the case where H_0 is a special case of the alternate hypothesis space (for example, with one of the H_1 parameters set equal to 0), the *Neyman-Pearson lemma* states that the most powerful test statistic between two hypotheses is the log-likelihood-ratio:

$$T = -2\ln\left(\frac{\mathcal{L}(X|H_0)}{\mathcal{L}(X|H_1)}\right)$$

The numerator must be lower than the denominator, as H_0 is a special case of H_1, so we have strictly $T > 0$. If the data are very far away from H_0, then the numerator will be low, the fraction will be much less than 1 and T_0 will be very large. For similar reasons to the relation between $-2\ln\mathcal{L}$ and χ^2 seen in section 2.3 (*Wilks' Theorem*), this test statistic is distributed as a χ^2 distribution with 1 degree of freedom (assuming there is only one parameter which differentiates H_0 and H_1). When we get some data, we can evaluate the test statistic T_0 with the data, the null hypothesis, and the alternative hypothesis (in particular, the alternative hypothesis with the best-fitting value of the parameter), and find the probability that T_0 would be that large; this probability is

$$p = \int_{T_0}^{\infty} P(\chi^2;1)\,\mathrm{d}\chi^2 = 1 - F(T_0;1)$$

This probability p is the probability that the test statistic could be so high under the null hypothesis. If p is very low, we can reject the null hypothesis with a confidence of $1-p$; there is only a probability p that doing so would be mistaken.

For some reason, people sometimes convert the p-value into a Z-score, corresponding to the number of standard deviations away from the mean of a normal distribution outside which we would get a probability p. A higher Z-score means our data were more unlikely under the null hypothesis. However, one should be careful because sometimes the Z-score is being taken as a one-sided test (for example, if we are testing whether or not a signal exists) rather than a two-sided test (like testing the unconstrained value of some parameter), which would give a lower Z-score.

3.2 Limit Setting

Imagine we are searching for a signal against a background. It may turn out that our signal is not very big, and has a size consistent with 0 at some confidence level, so the signal may not actually exist. We can then set an *upper limit* for the size of the signal, at some confidence level.

The procedure for obtaining an upper limit at a particular confidence level is as follows. For a given value of the signal size, one calculates the log-likelihood ratio (or some other test statistic, though this will be the most powerful) of there being a signal of *at most* that size, compared to the null hypothesis of there being no signal at all. One then calculates the p-value of this test statistic; if the log-likelihood ratio is used then this will be $\sim \chi^2$, with the number of degrees of freedom depending on how many extra dof the inclusion of a signal gives to the probability distribution. The signal size is then adjusted up or down until the p-value corresponds to the confidence level desired: if we want an upper limit at 95% confidence, we vary the signal size until we reach the test statistic with a p-value of 0.05.

If the overall sensitivity of an experiment is quite low, there might be no possible data which would cause us to reject the null hypothesis. This might occur when the test statistic is distributed similarly under alternate hypothesis and the null hypothesis, so no value of the test statistic would be able to distinguish between the two at a given confidence level. A way of mitigating this is the CLs method, which alters the critical p-value by an amount depending on the null-hypothesis test statistic distribution:

$$p_{\mathrm{CLs}} = \frac{P(T > T_0 | H_1)}{P(T > T_0 | H_0)}$$

whereas previously we would just be looking at the numerator. By dividing by $P(T > T_0|H_0)$, we are increasing the p-value we are reporting, making this method deliberately conservative.

3.3 Resampling Methods

We will always want a bigger sample, but taking large samples directly from the distribution is sometimes infeasible and we have to make the best of the sample we have. Resampling consists of augmenting that sample, or constructing multiple samples from it.

3.3.1 Jackknife resampling

Suppose we are estimating a parameter θ from some data $\{X_1, X_2, \ldots, X_N\}$. We might have come up with an estimator function $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_N)$. Using a leave-one-out procedure, we can obtain N different estimates for θ by calculating $\hat{\theta}_{-1} = \hat{\theta}(X_2, X_3, \ldots, X_N)$, then $\hat{\theta}_{-2} = \hat{\theta}(X_1, X_3, \ldots, X_N)$, right the way up to $\hat{\theta}_{-N} = \hat{\theta}(X_1, X_2, \ldots, X_{N-1})$. Rather than using the estimator function $\hat{\theta}$ on all N data, we can instead make a *jackknifed estimate* $\hat{\theta}_J$ by taking the mean of these leave-one-out estimates:

$$\hat{\theta}_J = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_{-i}$$

Slight detour: if we estimate a true parameter θ with a consistent estimator $\hat{\theta}_N$ on N data, then $\lim_{N\to\infty}(\hat{\theta}_N - \theta) = 0$. We can thus write the expectation of $\hat{\theta}_N$ as a power series in $1/N$:

$$E[\hat{\theta}_N] = \theta + \frac{a_1}{N} + \frac{a_2}{N^2} + \dots$$

where the coefficients a_i will depend on the estimator $\hat{\theta}_N$. If we are only using $N-1$ data points, as with the $\hat{\theta}_{-i}$ above, then the expectation will be

$$E[\hat{\theta}_{N-1}] = \theta + \frac{a_1}{N-1} + \frac{a_2}{(N-1)^2} + \dots = E[\hat{\theta}_{-i}]$$

Now the jackknifed estimate $\hat{\theta}_J$ is $1/N$ times the sum of N of these $\hat{\theta}_{-i}$ quantities, all of which will have the same expectation and thus $E[\hat{\theta}_J]$ is also given by the above. We can therefore construct a quantity which is only biased to order N^{-2} rather than N^{-1}:

$$\begin{aligned} E\Big[N\hat{\theta}_N - (N-1)\hat{\theta}_J\Big] &= \Big(N\theta + a_1 + \frac{a_2}{N} + \dots\Big) - \Big((N-1)\theta + a_1 + \frac{a_2}{N-1} + \dots\Big) \\ &= \theta + \frac{a_2}{N^2} + \mathcal{O}(N^{-1}) \end{aligned}$$

We therefore write the *unbiased jackknife estimator* as:

$$\hat{\theta}_{\mathcal{J}} = N\hat{\theta}_N - (N-1)\hat{\theta}_J$$

3.3.2 Bootstrapping

Having taken a sample of size N from the underlying distribution, we can generate further samples of size N by sampling with replacement *from the original sample.* For instance, if the original sample was $\{1, 5, 2, 3, 4\}$, a bootstrapped sample might be $\{5, 2, 5, 3, 3\}$. The representativeness of the bootstrapped samples are, however, entirely dependent on the representativeness of the original sample.

One reason this is useful is if there is a parameter that we are estimating from the sample, we can generate a distribution for that parameter by estimating it for many bootstrapped samples. *Parametric bootstrapping* is useful where we have some model for the underlying distribution, and consists of the following procedure. We get a sample, and fit the model to the data in that sample. This gives us a point estimate for each of the parameters of the model. We then generate lots of bootstrap samples from the original sample, and for each of these bootstraps we re-fit the model, obtaining a new set of parameters. This gives a distribution for each of the parameters, where the mean should be roughly centred on the point estimates obtained by fitting the model to the original sample, and the spread gives an indication of how sensitive we are to that parameter.