Lecture Notes on

Statistics

Kevin Zhou

kzhou7@gmail.com

These notes cover basic statistics, with a focus on particle physics experiments. I have included discussion of statistical methodology, but much of it is just my personal opinion. The primary sources were:

- Richard Weber's Statistics IB lecture notes. A very clear set of notes, covering a limited set of topics well; examples are drawn from the softer sciences. The notes are a bit dry and brief, due to an artificial limitation of one A5 sheet of paper per lecture. Don't miss the problems and digressions on the course website.
- Cowan, *Statistical Data Analysis*. A gentle introduction clearly outlining the basics. Very good for a first pass to get intuition. Also see the course lecture slides, which go into more detail on modern topics.
- Lista, *Statistical Methods for Data Analysis in Particle Physics*.
- Preneau, *Data Analysis Techniques for Physical Scientists*.

The most recent version is here; please report any errors found to kzhou7@gmail.com.

Contents

1 Introduction

1.1 Random Variables

We begin by establishing notation and context.

- Our data is a vector of random variables, $\mathbf{X} = (X_1, \ldots, X_n)$, which are often independent and identically distributed.
- The data are drawn picked from a family of distributions (e.g. normal distributions with some μ and σ) where typically the parameters θ are unknown. The goal of parameter estimation is to estimate these parameters.
- Often the parameters θ occupy a subset of a vector space. We will only cover "parametric" statistics, where this space is finite-dimensional. "Nonparametric" statistics is significantly more complicated.
- In the frequentist picture, we do not specify a distribution of θ. Instead, we should think of θ as fixed but unknown, with all randomness coming from the random variables X.
- In general, we write parameters separated from variables; for example, the pdf of a general normal distribution would be written $f(x|\mu, \sigma^2)$. The pdf of a random variable X is written $f_X(x)$. When there is no possibility of confusion, we will just write $f(x)$ or $g(x)$, and so on.
- A statistic $T(\mathbf{X})$ is any function of the data, and is hence also a random variable. Sometimes the distribution of a statistic is called a sampling distribution.
- Specific values are indicated by lowercase. For example, we can have a specific value $t(\mathbf{x})$ where $\mathbf{x} = (x_1, \ldots, x_n)$. (This distinction is rarely made in practice in physics, where both the random variables and their values would have the same letter.)
- An estimator is a statistic we use to estimate a parameter θ. It is unbiased if

  $$\mathbb{E}(T) = \theta$$

  where the expectation is taken over possible values of X given fixed θ. Estimators may also be written as the corresponding parameter with a hat, e.g. $\hat{\theta}$.

Note. Generally, unbiased estimators need not exist. For example, consider a coin flip with probability p. Then for any estimator,

$$\mathbb{E}\,\hat{\theta} = (1-p)\hat{\theta}(0) + p\hat{\theta}(1).$$

This means we can only estimate linear functions of p without bias. More generally, unbiased estimators can be extremely poor, with arbitrarily high spread; a lack of bias is a good property, but far from the only important thing. There is often a tradeoff between the bias of an estimator and its variance.

Now we review some useful facts about distributions.

- It is often useful to combine random variables by addition or multiplication. Let X and Y have pdfs $g(x)$ and $h(y)$. Then the pdf of $Z = X + Y$ is

$$f(z) = \int_{-\infty}^{\infty} g(z-y)h(y)\,dy$$

which is the Fourier convolution of g and h. The pdf of $Z = XY$ is

$$f(z) = \int_{-\infty}^{\infty} g(z/y)h(y)\,\frac{dy}{|y|}$$

which is the Mellin convolution of g and h. In practice these integrals are hard to perform; instead one can perform a Fourier or Mellin transform, which converts these convolutions to multiplication.

- When performing a change of variables from n random variables $\mathbf{X}$ to n other independent random variables $\mathbf{A}$, the joint pdf is multiplied by the Jacobian,

$$g(a_1, \ldots, a_n) = f(x_1, \ldots, x_n)\,|\det J|, \quad J_{ij} = \frac{\partial x_i}{\partial a_j}$$

- We will write the expected value of a random variable X as

$$\mathbb{E}(X) = \mu_X$$

again suppressing the subscript when there is no chance of confusion. The algebraic moments are $\mathbb{E}(X^n)$, while the central moments are

$$\mathbb{E}((X-\mu)^n) = \mu_{X,n}.$$

The variance is the second central moment, and is written σ_X^2.

- The covariance matrix for a set of random variables $\mathbf{X}$ is denoted V, where

$$V_{ij} = \mathbb{E}((X_i - \mu_i)(X_j - \mu_j)) = \mathbb{E}(X_i X_j) - \mu_i \mu_j$$

using the notation $\mu_i = \mu_{X_i}$. The diagonal elements of the covariance matrix are just the variances. The correlation coefficients are defined as

$$\rho_{ij} = \frac{V_{ij}}{\sigma_i \sigma_j} \in [-1, 1].$$

If the random variables are independent, the off-diagonal elements of the covariance matrix vanish, though the converse is not true.

- In physics, one often uses a heuristic procedure known as "error propagation" (where "error" just means what we call standard deviation). The point of error propagation is that if one only knows the means and covariance matrix of a set of random variables $\mathbf{X}$, it is possible to approximate the means and covariance of functions $\mathbf{Y}(\mathbf{X})$ of those random variables. This approximation is good as long as the low-order Taylor expansion of $\mathbf{Y}$ in $\mathbf{X}$ is good.

- Specifically, by evaluating $\mathbf{Y}(\mathbf{X})$ up to first order, we have

$$\mathbb{E}(\mathbf{Y}) = \mathbf{y}(\mu_\mathbf{X}) + \text{second order in } (\mathbf{X} - \mu_\mathbf{X}).$$

  Similarly, the covariance matrix U of the new variables is

$$U_{k\ell} = \sum_{ij} \left(\frac{\partial y_k}{\partial x_i} \frac{\partial y_\ell}{\partial x_j} \right)_{\mathbf{x}=\mu_\mathbf{X}} V_{ij} + \text{second order in } (\mathbf{X} - \mu_\mathbf{X}).$$

  Note that if we define the matrix of partial derivatives $A_{ij} = \partial y_i / \partial x_j$, then

$$U = AVA^T.$$

- For example, if $Y = X_1 + X_2$, then

$$\sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$$

  while if $Y = X_1 X_2$, then

$$\frac{\sigma_Y^2}{y^2} = \frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} + \frac{2V_{12}}{\mu_1 \mu_2}.$$

  If X_1 and X_2 are independent, this is just the usual addition in quadrature of absolute and relative error, respectively. Note that the first result is exact, while the second relies on the higher-order corrections to the Taylor series being small.

- Sometimes, it is useful to diagonalize the covariance matrix using a linear transformation,

$$\mathbf{Y} = A\mathbf{X}.$$

  The new covariance matrix is $U = AVA^T$ and V is symmetric, so this can be achieved for some orthogonal matrix A. This is called principal component analysis. The resulting principal components can be less intuitive than the original variables; for instance, if elements of $\mathbf{X}$ represent qualitatively different things (e.g. gender and age) then combining them requires an arbitrary choice of relative normalization, which affects the principal components.

1.2 Examples of Distributions

We now review some important distributions.

- The most important case is when X is standard normal, $X \sim N(0, 1)$, which has

$$f(x) = \varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

  The cdf of X is called Φ. More generally, for $X \sim N(\mu, \sigma^2)$ we have

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

  Even more generally, $\mathbf{X}$ is multivariate normal, $\mathbf{X} \sim N(\boldsymbol{\mu}, V)$, if

$$f(\mathbf{x}) = \frac{1}{\sqrt{|\det(2\pi V)|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T V^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

- We now consider some discrete distributions. We say X is binomial with parameters (N, p) if

$$f(n) = \binom{N}{n} p^n (1-p)^{N-n}, \quad \mathbb{E}(X) = Np, \quad \operatorname{var} X = Np(1-p).$$

  This represents, for example, the number of successes in N independent trials with probability of success p. Note that we are using the same notation for probabilities and probability densities.

- The multinomial distribution generalizes the binomial distribution to more than two outcomes. Let the outcomes have probabilities p_i, where $i \in \{1, \ldots, m\}$. Then

$$f(n_1, \ldots, n_m) = \frac{N!}{n_1! \cdots n_m!} p_1^{n_1} \cdots p_m^{n_m}.$$

  The expected values are clearly $\mu_i = Np_i$. To compute the covariance matrix, it is useful to group the outcomes into "outcome i, outcome j, and anything else", so we never have to deal with more than three options, and note that the N trials are independent. This gives

$$V_{ij} = N(\delta_{ij} p_i - p_i p_j).$$

  The multinomial distribution could represent, e.g. the occupancies of bins in a histogram with a fixed total count. The fluctuations of the bins are anticorrelated, because having more counts in one bin leaves fewer for the other bins.

- We say X is Poisson with parameter λ if

$$f(n) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad \mathbb{E}(X) = \operatorname{var} X = \lambda.$$

  One can show that this is the limit of a binomial distribution as $N \to \infty$ and $p \to 0$, fixing $Np = \lambda$, and hence can be thought of as the counts due to a memoryless process occurring in continuous time. This is a simple example of how the limit $N \to \infty$ alone doesn't guarantee a normal distribution, instead one needs enough variance to "smear out" the distribution, $\lambda \to \infty$.

- We say X is geometric with parameter p if

$$f(n) = p(1-p)^{n-1}, \quad n = 1, 2, \ldots.$$

  One example is the number of independent flips needed to get a head on a coin with heads probability p.

- The continuous analogue of the geometric distribution is the exponential. If $X \sim \mathcal{E}(\lambda)$, then

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad \mathbb{E}(X) = \frac{1}{\lambda}, \quad \operatorname{var} X = \frac{1}{\lambda^2}.$$

- The gamma distribution is the sum of n independent exponentials. If $X \sim \text{gamma}(n, \lambda)$, then

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}, \quad x > 0, \quad \mathbb{E}(X) = \frac{n}{\lambda}, \quad \operatorname{var} X = \frac{n}{\lambda^2}.$$

  The sum of gamma distributed variables is also gamma distributed, and of course in the limit $n \to \infty$, it approaches a normal distribution. One example of a gamma distributed quantity is the time until the n^{th} event in a memoryless process. It can also be used as a generic pdf for a quantity known to be positive.

- If $Y \sim N(\mu, \sigma^2)$, then $X = e^Y$ is log-normal distributed,

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

  It is the result of multiplying many independent random variables, and

$$\mathbb{E}(X) = e^{\mu+\sigma^2/2}, \quad \operatorname{var} X = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

- For completeness, X is uniformly distributed on $[\alpha, \beta]$, $X \sim U(\alpha, \beta)$, if

$$f(x) = \frac{1}{\beta - \alpha} \mathbb{1}(x \in [\alpha, \beta]), \quad \mathbb{E}(X) = \frac{\alpha + \beta}{2}, \quad \operatorname{var} X = \frac{(\beta - \alpha)^2}{12}.$$

  An important example is that the cdf of any random variable, treated as a function of that random variable, is uniform on $[0, 1]$.

- We say $X \sim \text{beta}(a, b)$ if

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}, \quad 0 < x < 1, \quad B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

  By some exhausting use of gamma function identities, we have

$$\mathbb{E}(X) = \frac{a}{a+b}, \quad \operatorname{var}(X) = \frac{ab}{(a+b+1)(a+b)^2}.$$

  For example, the a^{th} smallest of b numbers independently sampled from $U(0, 1)$ is distributed this way. The beta distribution can also be used as a generic pdf for a quantity known to be bounded in a finite interval.

- We say X is χ^2-distributed with n degrees of freedom, $X \sim \chi_n$, if

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \quad x > 0, \quad \mathbb{E}(X) = n, \quad \operatorname{var} X = 2n.$$

  It is the distribution of the sum of the squares of n independent standard normal random variables, and it will be useful for "goodness of fit" tests. For a general multivariate normal distribution, the quantity $(\mathbf{X} - \boldsymbol{\mu})^T V^{-1} (\mathbf{X} - \boldsymbol{\mu})$ is χ^2-distributed. (The square root of this quantity is "chi distributed".) The exponential and χ^2 distributions are special cases of the gamma distribution.

- The Cauchy/Breit–Wigner/Lorentz distribution is

$$f(x) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

  where x_0 represents the location of the peak, and Γ represents the full width at half maximum. Physically, X could be the energy of a decaying particle, in which case Γ is its decay rate.

- It is important to note that the Cauchy distribution has a heavy tail, so that its expected value does not exist; in fact, none of its moments exist. Distributions such as these, which are regarded as pathological in mathematics, are quite common and useful in physics. Of course, the *real* energy distribution sampled by a particle decay does have moments, since it is bounded; however, the Cauchy distribution is a very useful approximation. Another example of such a distribution is the Landau distribution, which represents the probability distribution of the energy loss of a charged particle when traversing a layer of matter.

Note. It can be surprising to have the expected value of a random variable not exist. After all, can't one always measure the expected value by just sampling it many times and taking the average? Actually, if the expected value doesn't exist, this procedure doesn't work! The average will never "settle down" as the number of trials go up, since large deviations will occur often enough to swing the whole average.

Note. Stable distributions are those which are closed under taking linear combinations of independent random variables with that distribution. The normal distribution is the classic one, but there is an entire family of stable distributions, called the Levy alpha-stable distributions, which are distinguished by heavier tails.

One can also consider the asymptotic distribution of statistics besides the sum. For example, we mentioned above that for most common distributions, the asymptotic distribution of the product is the log-normal distribution. One can also consider the highest value. Up to normalization, the Fisher–Tippett/extreme value theorem states that there are only three possible asymptotic distributions: the Gumbel distribution, the Frechet distribution, and the Weibull distribution.

Example. Computing a conditional distribution. Let $X \sim \text{Pois}(\lambda)$ and $R \sim \text{Pois}(\mu)$ independently, and let $Y = X + R$. Then

$$f_{X|Y}(x|y) = \frac{\lambda^x e^{-\lambda}}{x!}\frac{\mu^{y-x}e^{-\mu}}{(y-x)!}\Bigg/\sum_{x'+r'=y}\frac{\lambda^{x'}e^{-\lambda}}{x'!}\frac{\mu^{r'}e^{-\mu}}{r'!} = \binom{y}{x}\lambda^x\mu^{y-x}\Bigg/\sum_{x'+r'=y}\binom{y}{x'}\lambda^{x'}\mu^{r'}$$

where we canceled the normalizing factors and multiplied by $y!/y!$. The denominator is $(\mu+\lambda)^y$ by the binomial theorem. Simplifying,

$$f_{X|Y}(x|y) = \binom{y}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(\frac{\mu}{\lambda+\mu}\right)^{y-x}.$$

That is, the distribution is binomial, with total y and $p = \lambda/(\lambda+\mu)$. This makes sense, since we can think of X and R as counts of memoryless processes occurring independently during a fixed time interval, and p is just the chance that a particular count belonged to X.

Example. Another complex example. Let $X_1 \sim \text{gamma}(n_1, \lambda)$ and $X_2 \sim \text{gamma}(n_2, \lambda)$, independently. Their joint distribution is

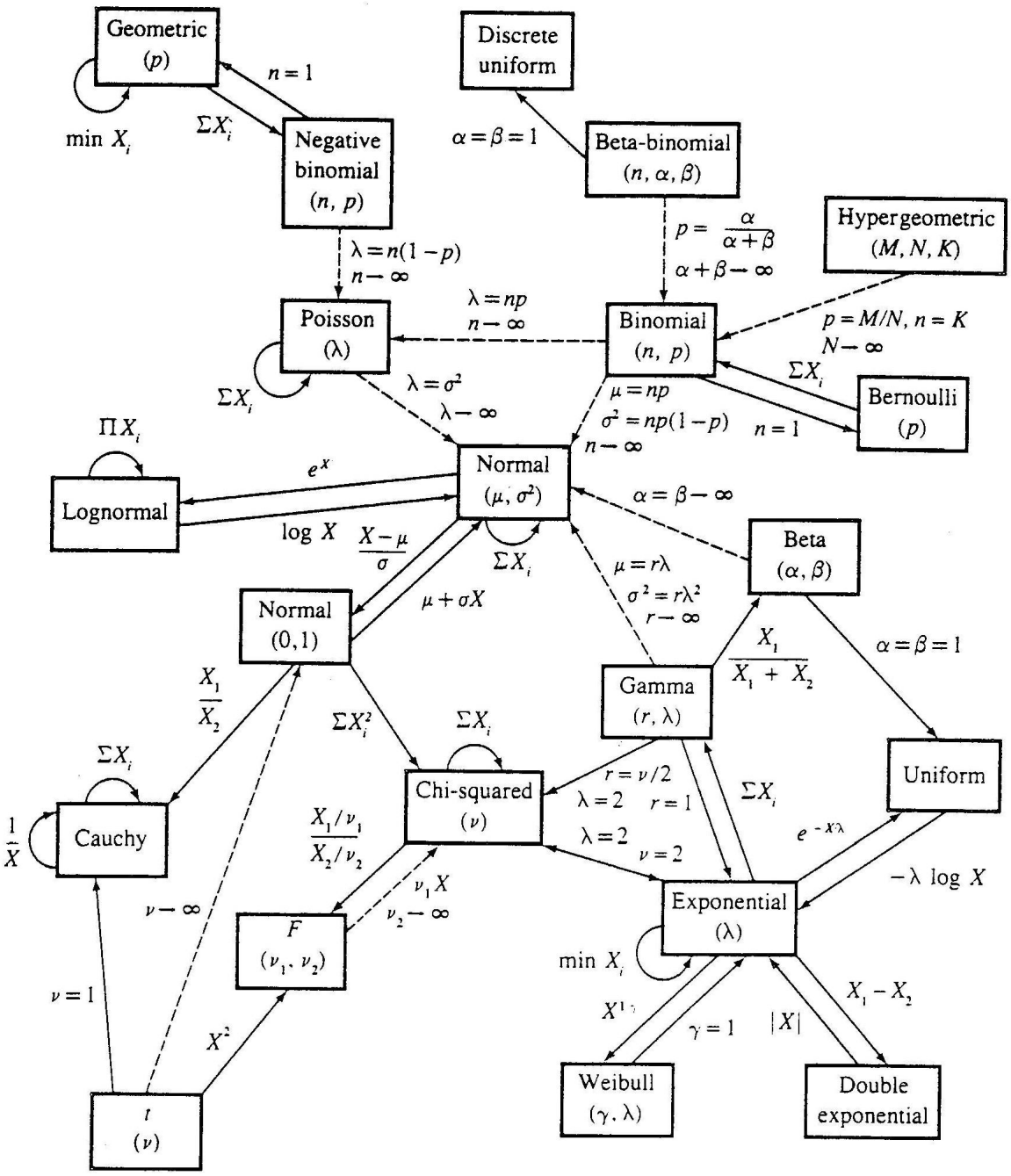
$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\lambda^{n_1+n_2}x_1^{n_1-1}x_2^{n_2-1}}{(n_1-1)!(n_2-1)!}e^{-\lambda x_1}e^{-\lambda x_2}.$$

Consider the distribution of $Y = X_1/(X_1+X_2)$. Using a routine double integral change of variables,

$$f(y) = \int_0^\infty dx_1\,\frac{x_1}{y^2}\,f_{\mathbf{X}}(x_1, x_2)\Big|_{x_2=x_1/y-x_1}$$

and it turns out that $Y \sim \text{beta}(n_1, n_2)$. This is compatible with our earlier intuitive descriptions of the beta and gamma distributions. We can think of X_1 as the time to wait for n_1 counts of a memoryless process, and $X_1 + X_2$ as the time to wait for $n_1 + n_2$ counts. For each possible value of the total time, we can think of Y as the result of picking $n_1 + n_2$ random points in that time interval and taking the n_1^{th} smallest. And since this holds for every possible value of the total time, Y is beta distributed.

A reference chart of common distributions and relations between them is below.



1.3 Characteristic Functions

(todo)

2 Parameter Estimation

2.1 Maximum Likelihood

Suppose X has distribution $f(x|\theta)$. Then the likelihood of θ given the observed value x is

$$\text{lik}(\theta) = p(x|\theta).$$

For multiple independent observations,

$$\text{lik}(\theta) = \prod_i f(x_i|\theta)$$

and so we often instead consider the log likelihood, which adds. The maximum likelihood estimator (MLE) $\hat{\theta}(x)$ is the one that maximizes the likelihood.

Example. Suppose candies come in k equally frequent colors. We examine three candies and find that they are red, green, and red. Then the likelihood is

$$\text{lik}(k) = p(x|k) = \frac{k-1}{k}\frac{1}{k}.$$

The value of k that maximizes this is $\hat{k} = 2$.

Example. Suppose $X \sim B(n,p)$, where n is known and p is to be estimated. Then

$$\log p(x|n,p) = \log \binom{n}{x} p^x (1-p)^{n-x} \sim x \log p + (n-x)\log(1-p).$$

This is maximized for $\hat{p} = X/n$. Since $\mathbb{E}(X/n) = p$, the MLE is unbiased.

Example. Consider $X_1, \ldots, X_n$ geometric with parameter p to be estimated. Then

$$\log f(x|p) = \log \prod_i (1-p)^{x_i - 1} p = \left(\sum_i x_i - n \right) \log(1-p) + n \log p.$$

Maximizing this yields $\hat{p} = \overline{X}^{-1}$. This is a biased estimate.

Example. Consider $X_1, \ldots, X_n$ exponential with parameter λ to be estimated. Then

$$\log f(x|\lambda) = \sum_i (\log \lambda - \lambda x_i) = n(\log \lambda - \lambda \overline{x}).$$

By similar logic to the previous example the MLE is $\hat{\lambda} = \overline{x}^{-1}$, which is biased,

$$\mathbb{E}[\hat{\lambda}] = \frac{n}{n-1}\lambda.$$

We can also parametrize the exponential by the lifetime τ,

$$f(x|\tau) = \frac{1}{\tau} e^{-x/\tau}$$

in which case the MLE of τ is $\hat{\tau} = \overline{x}$, which is unbiased. The lesson is that generally the MLE is not unbiased, and that even if the MLE is unbiased, it generally won't stay that way under a reparametrization. To show this more generally, if $\hat{\theta}$ is an unbiased MLE for θ, the MLE estimate of $f(\theta)$ is $f(\hat{\theta})$. This is only unbiased if $\mathbb{E}[f(\theta)] = f(\mathbb{E}[\theta])$, which is not generally true. On the other hand, the bias of most "reasonable" estimators goes to zero as $n \to \infty$.

Next, we introduce sufficient statistics, which contain all the inference information in the data.

- The statistic $T(X)$ is sufficient for θ if, for each t, the conditional distribution of X does not depend on θ. Therefore, knowing anything besides T is of no help for estimating θ.
- To rephrase this, note that we can always write

$$f(x|\theta) = \mathbb{P}_\theta(X = x) = \mathbb{P}_\theta(T(X) = t)\,\mathbb{P}_\theta(X = x|T(X) = t).$$

  The statement that T is sufficient means that the second factor doesn't depend on θ, so

$$f(x|\theta) = g(T(x), \theta)h(x).$$

  This is called the factorization criterion.
- Any invertible function of a sufficient statistic is also sufficient, so they are not unique.
- The maximum likelihood estimator maximizes $g(T(x), \theta)h(x)$. Since $h(x)$ is fixed, the MLE is a function of the sufficient statistic.
- As a trivial example, the full data set x itself is a sufficient statistic. More generally, the likelihood will have terms combining x and θ, and our goal is to rearrange these terms to contain only a few functions of x.
- A sufficient statistic is minimal if any sufficient statistic is a function of it. Minimal sufficient statistics give the greatest possible data reduction without losing parameter information.
- One can show that T is minimal if and only if

$$f(x|\theta)/f(y|\theta) \text{ is independent of } \theta \text{ iff } T(x) = T(y).$$

  That is, when T is minimal, if x and y are indistinguishable from the point of view of parameter estimation, then their statistics should be the same. (The converse is just the criterion to be a sufficient statistic in the first place.)

Example. Suppose $X_1, \ldots, X_n \sim \text{Pois}(\lambda)$, with λ to be estimated. Then

$$f(x|\lambda) = \prod_i \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_i x_i} e^{-n\lambda}}{\prod x_i!}.$$

We thus recognize the sample mean $T(x) = \sum_i x_i$ as a sufficient statistic for λ; the factors are the numerator and the denominator. The MLE is t/n, and this estimate is unbiased.

Example. Now suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$ to be estimated. Then

$$f(x|\mu, \sigma^2) = \frac{e^{-\sum_i (x_i - \mu)^2/2\sigma^2}}{(2\pi\sigma^2)^{n/2}}.$$

The sum in the exponential can be rearranged as

$$\sum_i (x_i - \mu)^2 = \sum_i (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2.$$

We have thus written the right-hand side in terms of the sample statistics

$$\overline{X} = \frac{1}{n}\sum_i X_i, \quad S_{XX} = \sum_i (X_i - \overline{X})^2$$

where $\overline{X}$ is the sample mean, and S_{XX} is proportional to the sample variance. These are a set of sufficient statistics, where the factor $h(x) = 1$. It's fairly clear that

$$\overline{X} \sim N(\mu, \sigma^2/n), \quad n(\overline{X} - \mu)^2 \sim \sigma^2 \chi_1^2$$

but the distribution of S_{XX} is a bit more subtle, as we'll discuss below.

Example. Let $X_1, \ldots, X_n \sim U(0, \theta)$ with θ to be estimated. Then

$$f(x|\theta) = \frac{1(\max x_i \le \theta)}{\theta^n}$$

so that a sufficient statistic for θ is $\max x_i$. This is also the MLE. However, using the usual trick that the cdf of uniform random variables is a power,

$$\mathbb{E} \max x_i = \frac{n}{n+1}\theta$$

so the MLE is biased. However, $\mathbb{E}\,\hat{\theta} \to \theta$ as $n \to \infty$, so we say $\hat{\theta}$ is asymptotically unbiased. One can show that under some mild assumptions, the MLE is always asymptotically unbiased, which helps justify its use.

2.2 Rao–Blackwell Theorem

We can quantify the quality of an estimator with the mean squared error,

$$\text{MSE} = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

Note that if $\hat{\theta}$ is unbiased, this is just the variance of $\hat{\theta}$.

Example. Consider $X_1, \ldots, X_n \sim B(1, p)$ with p to be estimated. Then the MLE is $\hat{p} = \overline{X}$, and this estimate is unbiased. The MSE is

$$\text{var}(\hat{p}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

By contrast, if we instead took $\hat{p} = X_1$, we would still have an unbiased estimate, but the MSE would be higher by a factor of n.

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2)$ to be estimated. The log likelihood is

$$\log(f|\mu, \sigma^2) = \log \prod_i \frac{e^{-(x_i-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \sim -n \log \sigma - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}.$$

Setting the derivatives to zero, the MLEs are

$$\hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{S_{XX}}{n}.$$

As expected, these are functions of the sufficient statistics found in the previous section. However, $\hat{\sigma}^2$ is biased, as

$$S_{XX} = \sum_i (X_i - \overline{X})^2 = \sum_i (X_i - \mu + \mu - \overline{X})^2 = \sum_i (X_i - \mu)^2 - n(\mu - \overline{X})^2$$

so upon taking the expectation value of both sides,

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2.$$

Intuitively, this is just the parallel axis theorem. The point is that the spread of a sample of points will be smaller relative to its own center than to the true distribution center. As expected, though, the MLE is asymptotically unbiased.

Note. Perhaps unintuitively, the estimator $S_{XX}/(n+1)$ has smaller MSE than either the MLE S_{XX}/n or the unbiased estimator $S_{XX}/(n-1)$, which is also called the standard error. Intuitively, this is because σ is squared, so we prefer a lower estimate because overestimates are penalized more. In general, MSEs will vary under reparametrizations, so the minimum MSE estimator need not be unbiased or a MLE generically.

Note. By similar logic, one can show that

$$\hat{V}_{XY} = \frac{n}{n-1}\left(\overline{XY} - \overline{XY}\right)$$

is an unbiased estimator for the covariance matrix of a pair of random variables. One can then normalize by the standard errors to get an estimator for the correlation coefficient, $\hat{R} = \hat{V}_{XY}/\sqrt{S_{XX}S_{YY}}$.

Theorem (Rao–Blackwell)**.** Let $\hat{\theta}$ be an estimator of θ with finite MSE. Suppose that T is sufficient for θ and let $\theta^* = \mathbb{E}(\hat{\theta}|T)$. Then for all θ,

$$\mathbb{E}[(\theta^* - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - \theta)^2].$$

Equality is achieved when $\hat{\theta}$ is a function of T.

Proof. We have

$$\mathbb{E}[(\theta^* - \theta)^2] = \mathbb{E}_T\left[\mathbb{E}_X[(\hat{\theta} - \theta)|T]^2\right] \leq \mathbb{E}_T\left[\mathbb{E}_X[(\hat{\theta} - \theta)^2|T]\right] = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

In the inequality, we used the fact

$$\mathrm{var}(W) = \mathbb{E}(W^2) - (\mathbb{E}\,W)^2 \geq 0$$

for the random variable $W = (\hat{\theta} - \theta|T)$. Equality occurs when $\mathrm{var}(W) = 0$, i.e. when $\hat{\theta} - \theta$ is uniquely determined by T, so that $\hat{\theta}$ is a function of T.

Note. The intuition behind the Rao–Blackwell theorem is that all the information is already contained in the sufficient statistic, so that whenever $\hat{\theta}$ varies for fixed T, we are simply picking up extra noise which adds to the MSE. This reiterates that our statistics should be functions of the sufficient statistic. Conversely, the averaging procedure used to define θ^* can often produce a good estimator starting from a very naive one. Also note that if $\hat{\theta}$ is unbiased, then θ^* is also unbiased.

Example. Let $X_1, \dots X_n \sim \text{Pois}(\lambda)$ with λ to be estimated. We know that $t = \sum_i x_i$ is a sufficient statistic. Now start with the unbiased estimator $\widetilde{\lambda} = X_1$. Then 'Rao–Blackwellization' gives

$$\lambda^* = \mathbb{E}\left[X_1 \middle| \sum_i X_i = t\right] = \frac{1}{n}\,\mathbb{E}\left[\sum_i X_i \middle| \sum_i X_i = t\right] = \frac{t}{n}.$$

This recovers our MLE estimator.

Example. Suppose we instead want to estimate $\theta = e^{-\lambda} = \mathbb{P}(X_i = 0)$. A simple unbiased estimator is $\hat{\theta} = 1(X_1 = 0)$. Then

$$\theta^* = \mathbb{E}\left[1(X_1 = 0) \middle| \sum_i X_i = t\right] = \mathbb{P}\left[X_1 = 0 \middle| \sum_i X_i = t\right] = \left(\frac{n-1}{n}\right)^t.$$

This is a much better estimator, better than any decent estimator we might have guessed.

Example. Let $X_1, \dots, X_n \sim U(0, \theta)$ with θ to be estimated. Starting with the unbiased estimator $\widetilde{\theta} = 2X_1$ and the sufficient statistic $t = \max_i x_i$,

$$\theta^* = \mathbb{E}\left[2X_1 | \max_i X_i = t\right] = 2\left(\frac{t}{n} + \frac{n-1}{n}(t/2)\right) = \frac{n+1}{n}t.$$

This is an unbiased estimate; the MLE estimate $\hat{\theta} = t$ is biased.

Definition. We say $\widetilde{\theta}$ is consistent if $\mathbb{P}(|\widetilde{\theta} - \theta| > \epsilon) \to 0$ as $n \to \infty$.

Theorem. The Cramer–Rao bound states that

$$\text{var}(\widetilde{\theta}) \geq \frac{(1 + \partial b/\partial\theta)^2}{nI(\theta)}$$

where $b = \mathbb{E}[\widetilde{\theta}] - \theta$ is the bias, and

$$I(\theta) = \mathbb{E}\left[-\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right]$$

is the Fisher information for each observation. An estimator is efficient if it saturates this bound.

Definition. We say an estimator $\widetilde{\theta}$ is asymptotically efficient if

$$\lim_{n\to\infty} \frac{\text{var}(\widetilde{\theta})}{1/nI(\theta)} = 1.$$

The MLE is both consistent and asymptotically efficient.

Example. The sample average $\overline{X}$ is an unbiased estimator of $\mathbb{E}[X_i]$, assuming this quantity exists at all. The weak law of large numbers states that this estimator is also consistent, if the X_i have finite variance. Hence it does not apply to the Cauchy distribution. In fact, in this case averaging doesn't do anything at all: the sample average $\overline{X}$ has exactly the same distribution as X_i, for any number of samples!

Example. Suppose you have been waiting for a time t for your friend to show up, and you think their arrival time is exponentially distributed with parameter λ. Given that they haven't shown up yet, the MLE estimate is $\hat{\lambda} = 0$, so we expect them to never show up.

This example exposes a weakness of point estimates; a confidence interval would be better. It also shows that parameter estimation is not sufficient for making decisions; a complete calculation should account for the costs of waiting and leaving early.

2.3 Confidence Intervals

We now switch from point estimation to interval estimation.

Definition. Let $a(X) \leq b(X)$ be two statistics. The interval $[a(X), b(X)]$ is called a $100\gamma\%$ confidence interval for θ if

$$\mathbb{P}\left(a(X) \leq \theta \leq b(X)\right) = \gamma$$

where γ is independent of θ. We say that $[a(X), b(X)]$ has $100\gamma\%$ coverage. In particle physics, the coverage is also called the "confidence level" CL.

Note. As in point estimation, we regard θ as fixed but unknown, so the probability in the definition above is taken over possible data X, not over possible θ. However, the equality must be true for all θ. Also, note that γ is not the probability that a *specific* confidence interval $[a(x), b(x)]$ contains θ. Every specific confidence interval either does or does not contain θ. Furthermore, as for estimation, there are multiple criteria for defining confidence intervals; as we'll see, some reasonable intervals don't have well-defined coverage at all.

Example. We can construct a 95% confidence interval by returning $(-\infty, \infty)$ with 95% probability, and the empty interval with 5% probability. However, once we see the specific interval we get, we know for sure whether or not it contains θ.

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known and μ to be estimated. We know that

$$\sqrt{n}(\overline{X} - \mu)/\sigma \sim N(0, 1).$$

Therefore, if $\mathbb{P}(\xi \leq N(0,1) \leq \eta) = \gamma$, then

$$\mathbb{P}(\xi \leq \sqrt{n}(\overline{X} - \mu)/\sigma \leq \eta) = \mathbb{P}(\overline{X} - \eta\sigma/\sqrt{n} \leq \mu \leq \overline{X} - \xi\sigma/\sqrt{n}) = \gamma.$$

This is not sufficient to determine the confidence interval; we need an additional principle that tells us which values are "more extreme" and hence should be excluded from the interval. For example, if higher values are more extreme, then we set the lower bound to $-\infty$, yielding a one-sided confidence interval. These issues get more subtle for multi-dimensional "confidence regions".

In this case, we'll choose to construct a confidence intervals as narrow as possible, in this parametrization. This implies a symmetric confidence interval, $\eta = -\xi$, giving

$$\Phi(\eta) = 1 - \gamma/2.$$

For example, for a 95% confidence interval, $\xi = 1.96$, and for 99%, $\xi = 2.58$. More realistically, we wouldn't know σ^2, so we would have to estimate it using the sample variance. It turns out that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1}$$

where t_{n-1} is the "Student's t-distribution on $n-1$ degrees of freedom". Then confidence intervals can be constructed using tables of t_{n-1} values. We'll give examples of this when we cover the t-test.

Example. Opinion polls. Suppose we want to estimate some proportion of people p. We interview n people and get a sample mean $\hat{p}$. For high n, it is a good approximation that

$$\hat{p} \sim N(p, p(1-p)/n).$$

Suppose we want the poll to have a $100\eta\%$ error margin. This means we want

$$\mathbb{P}(\hat{p} - \eta \leq p \leq \hat{p} + \eta) \geq 95\%.$$

Using the approximate distribution for $\hat{p}$, this probability is equal to

$$\Phi(\eta\sqrt{n/p(1-p)}) - \Phi(-\eta\sqrt{n/p(1-p)}) \geq \Phi(\eta\sqrt{4n}) - \Phi(-\eta\sqrt{4n}).$$

Then we require $\eta\sqrt{4n} \geq 1.96$. For a typical 3% error margin, $n \geq 1068$. Typical opinion polls use $n \approx 1100$. Alternatively we can simply replace p with $\hat{p}$, which is a fairly good approximation.

Note. The easy way to construct a confidence interval is to find some simple function $f(X, \theta)$ whose distribution doesn't depend on θ. In the first example, this function was $\sqrt{n}(\overline{X} - \mu)/\sigma$. We can then bound $\mathbb{P}(\xi \leq f(X, \theta) \leq \eta)$, and hopefully rearrange the inequality to the desired form $a(X) \leq g(\theta) \leq b(X)$.

This can't be done for the opinion poll case, where we had to either determine the interval by looking at all possible p, or naively set $\hat{p} = p$. Generally, the easy method works whenever the parameters are "location" or "scale" parameters. The parameter p in the opinion poll controls both simultaneously.

Example. A confidence interval with a scale parameter. Let $X_1, \ldots, X_n \sim \mathcal{E}(\theta)$. Then $T(X) = \sum_i X_i$ is sufficient for θ, and

$$T \sim \text{gamma}(n, \theta).$$

It is useful to consider the statistic

$$S = 2\theta T, \quad S \sim \text{gamma}(n, 1/2) = \chi^2_{2n}.$$

where the right-hand side is the χ^2 distribution with $2n$ degrees of freedom. The cdf of this distribution is conventionally written F_{2n}. By looking up its values, we can construct confidence intervals for θ.

Note. As we've seen, $S_{XX}/(n-1)$ is an unbiased estimator of σ^2. We now infer the distribution of S_{XX}. Since S_{XX} is unaffected by a shift, let's shift μ to zero without loss of generality. Each element of the vector $\mathbf{X}$ has distribution $N(0, \sigma^2)$, and the elements are independent, so $\mathbf{X}$ is distributed as a multivariate normal with diagonal covariance matrix $\sigma^2 I$.

For an orthogonal matrix A, the vector $\mathbf{Y} = A\mathbf{X}$ has the same diagonal covariance matrix. We can use this to separate out the sample mean and sample variance. We choose A so that its first row has elements $1/\sqrt{n}$, so that $Y_1 = \sqrt{n}\,\overline{X}$, and Y_1 is independent of the other elements of $\mathbf{Y}$. But we also have

$$\sum_{i \neq 1} Y_i^2 = \sum_i Y_i^2 - Y_1^2 = \sum_i X_i^2 - n\overline{X}^2 = S_{XX}$$

from an earlier example. Therefore, S_{XX} is independent of $\overline{X}$, and

$$S_{XX} \sim \sigma^2 \chi^2_{n-1}$$

which should be compared to

$$\sum_i (X_i - \mu)^2 \sim \sigma^2 \chi^2_n.$$

The intuition is just that one of the "directions" is taken up by $\overline{X}$, so there is one fewer degree of freedom. We can then use the distribution of S_{XX} to construct confidence intervals for σ^2.

2.4 Bayesian Estimation

Bayesian estimation is a totally different approach to parameter estimation.

- We think of the parameter as a random variable, with a prior distribution $p(\theta)$. In reality, it might not make sense to think of θ as varying, so we instead interpret this distribution as representing our beliefs in what θ is. This is a valid interpretation of probability because it satisfies the same axioms as the frequentist interpretation of probability as a long-term average in repeated experiments.
- Given data x_i, we update the distribution to a posterior distribution $p(\theta|x_i)$ by Bayes' rule,

  $$p(\theta|x_i) = \frac{f(x_i|\theta)p(\theta)}{\int f(x_i|\phi)p(\phi)d\phi}.$$

  Of course, Bayes' rule is just a simple mathematical fact that is also true for frequentist probabilities; it just happens to be much more important in Bayesian statistics.
- The denominator is just the probability of observing the data x_i in the first place, assuming the prior distribution. Since it is just a normalizing constant, we typically write

  $$p(\theta|x_i) \propto p(\theta)f(x_i|\theta)$$

  with manual normalization of the posterior. Thus, the prior is updated by multiplication by the likelihood.
- In our notation, the conditional distribution is written as

  $$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

  To avoid singularities, we define it to be zero when $f_Y(y)$ is zero. However, in practice this shouldn't happen because the prior distribution should have support over all possibilities.
- Estimation in Bayesian statistics is done with loss functions.
  - To give a point estimator, we return $\hat{\theta}$ which minimizes the expected loss $\mathbb{E}[L(\theta, \hat{\theta})]$.
  - If we use quadratic error loss $L(\theta, a) = (a - \theta)^2$, the result is the posterior mean.
  - If we use absolute error loss $L(\theta, a) = |a - \theta|$, the result is the posterior median.
  - In the case of delta-function error loss and a uniform prior, we get the MLE.
- It's worth elaborating on the previous point. In any parametrization, the MLE matches with the Bayesian result for delta-function error loss and a uniform prior *in that parametrization*. If we change the parametrization and transform the prior appropriately, then they will no longer coincide. This can be used to argue that either the MLE or the Bayesian result is pathological, though which one depends on taste.
- Confidence intervals can be constructed easily in Bayesian statistics: to form a $100(1-\alpha)\%$ Bayesian confidence interval, we just give an interval that covers $100(1-\alpha)\%$ of the posterior probability. However, such confidence intervals do not necessarily satisfy the frequentist definition of a confidence interval, i.e. they do not have the so-called "coverage" property. As such, they are alternatively called "credible intervals".

- This illustrates an important point in the foundations of statistics: unlike many other fields, the final results that one reports depend on the foundations! As we go on, we'll see some examples where the difference can be large.

Example. A biased coin is tossed n times, giving t heads. Suppose the prior distribution on the heads probability is uniform, $p(\theta) \sim U(0,1)$. Then

$$p(\theta|x_i) \propto \theta^t(1-\theta)^{n-t}$$

so the posterior is beta distributed.

Example. Let $X_1, \ldots, X_n \sim \mathcal{E}(\lambda)$ with prior $\lambda \sim \mathcal{E}(\mu)$. Then

$$p(\lambda|x_i) \propto \mu e^{-\lambda\mu} \prod_i \lambda e^{-\lambda x_i} \propto \lambda^n e^{-\lambda(\mu+\sum_i x_i)}$$

which is gamma$(n+1, \mu + \sum_i x_i)$.

The most common objection to Bayesian statistics is the prior dependence of the predictions.

- There are several ways to reply to this objection. First, it's worth noting that frequentist results are also implicitly prior dependent, in the sense of depending upon beliefs. For instance, many statistical tests we will consider below depend on beliefs about, e.g. the independence and distribution of the data. Moreover, many frequentist procedures are equivalent to Bayesian ones under a flat prior in an arbitrary parametrization ("Laplace's principle of insufficient reason"), which isn't any better justified.

- Another reply is that, in the limit of a large amount of data, the likelihood function will become very sharply peaked. So as long as the initial prior was not very unreasonable (e.g. sharply peaked about a wrong value), then the effect of the prior should "wash out".

- This idea leads us to the idea of "objective" Bayesian statistics. The idea is that if we have no preexisting opinions, the prior should be "maximally uninformative". This vague notion becomes well-defined if we fix what kind of measurements we can make. A prior is maximally uninformative for those measurements if it is expected to change as much as possible upon making a measurement. (For example, a delta function prior would never change at all, so it is informative.) Another way of phrasing this is that the set of measurements one can perform gives a preferred parametrization, upon which we can take the uniform prior.

- This particular idea leads to the Jeffreys prior,

  $$p(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})}$$

  where $I(\boldsymbol{\theta})$ is the Fisher information.

- The Jeffreys prior is known to have pathologies in higher dimensions. A common alternative in this case is the "reference prior" of Bernardo and Berger, which is constructed along the idea of having the posterior be as dominated by the likelihood as possible.

- Note that all "objective" priors need some external input, because parametrization is arbitrary. It would be more accurate to call them "priors selected by formal rules".

- There is also an opposing "subjective" school of Bayesian statistics. Philosophically, the prior should represent a degree of belief, and real beliefs shouldn't be set by what measurement apparatuses happen to be available. Instead one should infer a prior using "expert elicitation", which involves cornering an expert and demanding their best guess.
- These two schools of thought can conflict in particle physics. For example, a more "objective" prior for the mass m of a new particle might be uniform in m, while often theory subjectively favors a distribution uniform in $\log m$. In general, one should try to run the analysis with multiple reasonable priors to see how sensitive the results are to them.
- In some cases the Bayesian and frequentist ideas of probability overlap, in which case they coincide. For example, in an image classification task where half the pictures are dogs, $p(\text{dog}) = 1/2$ is meaningful in both pictures.
- In general, it is difficult to construct priors with many parameters. For example, with a prior that is reasonably uniform, almost all the probability density gets pushed to the edge of the space, by the "curse of dimensionality". This is a deep issue, which also affects other fields, such as machine learning.

Example. Consider a coin that is heads with probability $\theta \in [0, 1]$. The only measurement we can perform is to flip the coin and see whether it is heads or tails. Then the Fisher information is

$$I(\theta) = \mathbb{E}\left[\left(\frac{d}{d\theta}\log p(x|\theta)\right)^2\right] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

so the Jeffreys prior is

$$p(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}.$$

We could also flip the coin multiple times and count the heads, but this just multiplies the Fisher information by a constant, so it doesn't change the answer.

3 Hypothesis Testing

A statistical hypothesis is an assertion about the distribution of random variables, and a test of a statistical hypothesis is a rule that decides whether to reject that assertion.

3.1 Definitions

We will use the Neyman–Pearson framework.

- We take data $\mathbf{x} = (x_1, \ldots, x_n)$ drawn from a density f. (Note that f is the joint density for $\mathbf{x}$, i.e. we are not assuming that the x_i are iid, though we often will in practice.) We work with two hypotheses about f, the null hypothesis H_0 and the alternative hypothesis H_1.
- Our test will reject either the null hypothesis or the alternative hypothesis. The null hypothesis is meant to be a conservative default.
- Often, it is assumed that f belongs to a specific parametric family $\theta \in \Theta$. Then the hypothesis are of the form
$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$$
where Θ_0 and Θ_1 are disjoint, but need not contain the whole space Θ. We say the hypothesis are simple if the Θ_i contain a single value θ_i. Otherwise, they are composite.
- Another type of hypothesis is
$$H_0 : f = f_0, \quad H_1 : f \neq f_0$$
where f_0 is a specified density; this is a goodness-of-fit test. A third alternative is that we may test $H_0 : f = f_0$ against $H_1 : f = f_1$.
- A test is defined by a critical region C, so that
$$\text{if } \mathbf{x} \in C, H_0 \text{ is rejected}, \quad \text{if } \mathbf{x} \in \overline{C}, H_0 \text{ is accepted/not rejected.}$$
In particle physics, the boundary of C is called the "cut", or decision boundary.
- When H_0 is rejected when it is true, we make a type I error; when H_0 is accepted when it is false, we make a type II error. We generally consider type I errors to be more serious, as H_0 is chosen to be conservative.
- As a result, we often characterize tests by their (maximum) probability of type I error. Define
$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(\mathbf{X} \in C | \theta).$$
We call α the size, or significance level of the test. For a simple null hypothesis, this is simply the probability of type I error. Typically, once α is fixed, we define C to minimize the probability of type II error, which is quantified by the "power" of the test.
- We also call C the rejection region and $\overline{C}$ the acceptance region. That is, we focus on how we view the null hypothesis H_0. We usually do not say we accept H_1, because there is a great deal of arbitrariness in H_1, though this depends on the context.

Note. The way hypotheses are chosen varies significantly between different fields and even within a field. In many cases, H_0 is qualitatively different from H_1 and definitely conservative: it could be the hypothesis that no new particle is present, or that a drug or psychological intervention has no effect on a disorder. But note that the former case is quite different, because it is very difficult to make H_1 the complement of H_0, as a new particle can manifest in many different ways.

Within particle physics, there are many layers of hypothesis testing. For example, one could apply hypothesis testing to the problem of distinguishing anomalous events, where H_0 is the Standard Model (SM) expectation for each event; such tests must thus be run millions of times per second. And even within one event, one needs a test to distinguish, e.g. different charged particles that fly through a part of a detector. For example, H_0 could represent a pion, H_1 an electron, and so on.

This is quite different from the other cases because for each particle or event, exactly one of the hypotheses is true (up to some mild idealizations). Depending on the situation, the null hypothesis might be a boring background, in which case we think of the alternative hypothesis as representing "signal", or neither might be "more conservative" than the other. And since this is a repeatable process, there is a well-defined notion of "the probability for H_i to be true" even for a frequentist. If we think of the H_1 as representing signal events and H_0 as background, the goal of maximizing the power for fixed test size translates into maximizing the signal purity for fixed background rejection. The point here is that there is some overlap between classification and hypothesis testing.

Note. Philosophical differences in hypothesis testing. Within frequentist statistics, there are the Fisher and Neyman–Pearson frameworks. In the Fisher framework, one does not specify an alternative hypothesis at all; one only computes p-values, which require only a null hypothesis. Thus the emphasis is on type I error. As the p-value is lowered, the evidence against the null hypothesis continuously increases. In the Neyman–Pearson framework, one thinks of an alternative hypothesis, and designs the test so that, for a fixed type I error, there is a minimum possibility of type II error. This leads to tests where one decides to reject or fail to reject the null hypothesis depending on whether the p-value is below some threshold.

These two frameworks optimize for different things. To oversimplify a bit, the Fisher framework is suited for exploration, while the Neyman–Pearson framework is suited for decision making. The sharp p-value threshold in Neyman–Pearson seems artificial for exploratory studies, but it is necessary for, e.g. classifying individual particles as they fly by. The concrete choice of H_1 in the Neyman–Pearson framework allows for more powerful tests tailored to H_1, while the Fisher framework is more agnostic (in particle physics language, "model-independent"). However, it is worth emphasizing that no test can be *truly* model-independent. The definition of a p-value involving results "at least as extreme" as those observed depends on what one expects can happen.

The same philosophical divides also occur in the Bayesian framework, which we'll describe in more detail later. One can treat the Bayesian framework as an instrument for exploration, where one merely accumulates evidence and updates probabilities accordingly. We have

$$\frac{p(H_0|x)}{p(H_1|x)} = \frac{p(x|H_0)}{p(x|H_1)} \frac{p(H_0)}{p(H_1)}$$

where the likelihood ratio above is called the Bayes factor B. Jeffreys' scale describes the strength of Bayesian evidence,

$$B > 150 : \text{very strong}, \quad B > 20 : \text{strong}, \quad B > 3 : \text{substantial}, \quad 3 > B > 1 : \text{barely worth mentioning}.$$

On the other hand, one can also take the Bayesian framework as the foundation for a *decision theory*, which just means any mathematical framework for making decisions. A typical framework for

decision theory involves maximizing some expected utility, where the expectation value is calculated with respect to the Bayesian probabilities. One can then use this theory to decide whether to accept or reject hypotheses.

3.2 Likelihood Ratio Tests

Many tests are based on likelihood ratios.

- The likelihood of a hypothesis $H : \theta \in \Theta$ is

$$L_{\mathbf{x}}(H) = \sup_{\theta \in \Theta} f_{\mathbf{X}}(\mathbf{x}|\theta)$$

  which reduces to the usual likelihood in the case of a simple hypothesis.

- The likelihood ratio of two hypotheses is

$$L_{\mathbf{x}}(H_0, H_1) = L_{\mathbf{x}}(H_1)/L_{\mathbf{x}}(H_0).$$

  Note that if $T(\mathbf{x})$ is sufficient for θ, then the likelihood ratio is simply a function of T.

- A likelihood ratio test is one with a critical region of the form

$$C = \{\mathbf{x} : L_{\mathbf{x}}(H_0, H_1) > k\}$$

  where k is determined by the size α.

Lemma (Neyman–Pearson)**.** For simple hypotheses, likelihood ratio tests are optimal, in the sense that they minimize type II error for a given maximum type I error.

Proof. For simple hypotheses, minimizing the type II error for fixed test size is equivalent to saying

$$\text{maximize} \int_C f(\mathbf{x}|\theta_1)\, d\mathbf{x} \text{ such that } \int_C f(\mathbf{x}|\theta_0)\, d\mathbf{x} = \alpha$$

where we defined $f_i(\mathbf{x}) = f(\mathbf{x}|\theta_i)$ for brevity. Stated this way, it is clear that a likelihood ratio test is optimal, because it just takes the region that gives the most of the former integrand per unit of the latter integrand.

Let's show this more formally. Consider any test with size α with critical region D. The likelihood ratio test is $C = \{\mathbf{x} : f_1(\mathbf{x})/f_0(\mathbf{x}) > k\}$ where k is determined by the size α. Now note that

$$0 \leq (\phi_C(\mathbf{x}) - \phi_D(\mathbf{x}))(f_1(\mathbf{x}) - kf_0(\mathbf{x}))$$

since the product terms always have the same sign. Integrating, we get four terms,

$$0 \leq \mathbb{P}(\mathbf{X} \in C|H_1) - P(\mathbf{X} \in D|H_1) - k\left[\mathbb{P}(\mathbf{X} \in C|H_0) - \mathbb{P}(\mathbf{X} \in D|H_0)\right].$$

The last two terms cancel, as they are simply the size α, giving the result.

Most of the time, we'll be working with non-simple hypotheses, so the Neyman–Pearson lemma does not apply. But the intuition it gives will lead us to favor likelihood ratio tests in general.

Example. The one-tailed z-test. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with σ^2 known. We test

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1, \quad \mu_1 > \mu_0.$$

The likelihood ratio is

$$\frac{f(\mathbf{x}|\mu_1, \sigma^2)}{f(\mathbf{x}|\mu_0, \sigma^2)} = \exp\left[\sum_i ((x_i - \mu_0)^2 - (x_i - \mu_1)^2)/2\sigma^2\right] = \exp\left[n(2\overline{x}(\mu_1 - \mu_0) + (\mu_0^2 - \mu_1^2))/2\sigma^2\right].$$

The likelihood ratio is a monotone function of the sufficient statistic $\overline{x}$, so the test is simply

$$\text{reject } H_0 \text{ if } \overline{x} > c, \quad \mathbb{P}(\overline{X} > c|H_0) = \alpha.$$

More explicitly, we can compute the statistic

$$Z = \sqrt{n}(\overline{X} - \mu_0)/\sigma \sim N(0, 1)$$

so the test is

$$\text{reject } H_0 \text{ if } z > z_\alpha, \quad z_\alpha = \Phi^{-1}(1 - \alpha).$$

For example, for $\alpha = 0.05$, we take $z_\alpha = 1.645$. Notice that since the test is defined by its size, it only depends on μ_0. That's why, at this stage, it makes no sense to talk about accepting or rejecting μ_1. The only dependence on μ_1 was that it was above μ_0, e.g. we would have gotten the same result with the composite hypothesis $H_1 : \mu > \mu_0$. In an extreme case, the alternative hypothesis could have been $\mu_1 = 10^{10^{100}}$, and it certainly wouldn't have made sense to accept that.

We can also two-tailed tests; these could arise from a composite hypothesis $H_1 : \mu \neq \mu_0$. In this case, the test changes to

$$\text{reject } H_0 \text{ if } |z| > z_{\alpha/2}.$$

We will see this more explicitly below.

Example. Suppose we want to test if a coin is fair. Let p be the probability of heads, and we test $H_0 : p = 0.5$ against $H_1 : p > 0.5$. For $n \gg 1$ flips, the distribution of the number of heads is

$$X \sim B(n, p) \approx N(np, np(1 - p)).$$

Under H_0, the distribution is $N(0.5n, 0.25n)$. We can thus construct a size α test by just using the criterion from the z-test, i.e. that $z > z_\alpha$. However, this is not the same thing as the z-test, because in the z-test the alternative hypothesis is a normal with the same variance as the null hypothesis. In fact, this test is not a likelihood ratio test either. This is another example of how the way we construct tests can be independent of the alternative hypothesis.

3.3 Properties of Tests

We now define a few more useful quantities.

- For a likelihood ratio test, the p-value of an observation is

$$p^* = \sup_{\theta \in \Theta_0} P_\theta(L_{\mathbf{X}}(H_0, H_1) \geq L_{\mathbf{x}}(H_0, H_1)).$$

  For simple hypotheses, assuming that H_0 holds, the p-value is the probability of seeing something at least as extreme as $\mathbf{x}$ against the null hypothesis.

- Recall that the rejection region C for a likelihood ratio test is the region with total probability α under H_0 that maximizes the likelihood ratio, while p^* counts the amount of probability weight with higher likelihood ratio. Then H_0 is rejected exactly when $p^* \leq \alpha$.
- Since p^* is more informative than the binary accept/reject, we often report p^* without specifying α. The p-value is also called the significance level of the data $\mathbf{x}$.
- More generally, we can define tests where C is defined by the value of some statistic T. Then the p-value is defined as the chance under H_0 of seeing a value of T "at least as extreme". Of course, this is a subjective notion that is vulnerable to p-hacking, as we discuss further below.
- The power of a test is defined as

$$B(\theta) = \mathbb{P}(\mathbf{X} \in C|\theta).$$

  Note that $\alpha = \sup_{\theta \in \Theta_0} B(\theta)$, while for $\theta \in \Theta_1$, $1 - B(\theta) = \mathbb{P}(X \in \overline{C}|\theta) = \mathbb{P}(\text{type II error}|\theta)$. That is, we would like the power to be low over Θ_0 and high over Θ_1. When we talk about find an "optimal" or "most powerful" test, we mean maximizing the power over Θ_1.
- Given Θ_0 and Θ_1, a uniformly most powerful (UMP) test of size α is one with size α and the maximum possible power for all $\theta \in \Theta_1$. UMP tests do not necessarily exist, but likelihood ratio tests are often UMP.
- Tests are closely related to confidence intervals. For example, in a test of size α for $H_0 : \theta = \theta_0$, the region $\overline{C}$ is a $100(1-\alpha)\%$ confidence interval for θ_0. In other words, if the confidence interval includes θ_0, then there's a related test in which H_0 is not rejected.

Note. As always, there are philosophical differences on how to use p-values. If one's goal is to use a statistical test to yield a binary accept/reject, then p-values are a distraction. (For example, they allow one to "bargain" if the p-value is close to the threshold.) But if one's goal is to weigh evidence in a more exploratory manner, p-values are useful. For example, in the Bayesian approach, the p-value is related to how we should update our prior probability for the null hypothesis.

Example. Recall the one-tailed z-test. By the Neyman–Pearson lemma, it is optimal for any $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$. The test does not depend on the value of μ_1, so it is an UMP test for $H_1 : \mu > \mu_0$.

Note. There is typically no UMP test for two-tailed hypotheses: a two-tailed test has decent power on both sides, but will get beaten in power on one side by each of the one-tailed tests. In fact, zooming out, we should not expect UMP tests to exist for anything but the most restricted of hypotheses. For example, in particle physics, alternative hypotheses might include supersymmetry, a Z' boson, and so on. The most powerful tests for each of these options are necessarily tailored towards that option. Just as this two-tailed example shows, the more "model independent" a test is, the more power it typically sacrifices for each specific option.

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where μ is known, and we wish to test $H_0 : \sigma^2 \leq 1$ against $H_1 : \sigma^2 > 1$. It's easier to first consider some alternative, simple hypotheses, $H_0' : \sigma = \sigma_0$ and $H_1' : \sigma = \sigma_1 > \sigma_0$. The likelihood ratio is

$$\frac{f(\mathbf{x}|\mu, \sigma_1^2)}{f(\mathbf{x}|\mu, \sigma_0^2)} = (\sigma_0/\sigma_1)^n \exp\left[\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)\sum_i (x_i - \mu)^2\right]$$

which, as expected, only depends on $T = \sum_i (x_i - \mu)^2$. The likelihood is monotonic in T, so the optimal test involves an upper cutoff on T.

Now replace H_0' with H_0. In this case, the size α of the test is set by the case $\sigma_0 = 1$, for which

$$\sum_i (x_i - \mu)^2 \sim \chi_n^2$$

which means the optimal test of H_0 against H_1' is

$$\text{reject } H_0 \text{ if } t > F_\alpha^{(n)}$$

where $F_\alpha^{(n)}$ is the upper α point of χ_n^2. Since this doesn't depend on σ_1 beyond the fact that $\sigma_1 > \sigma_0$, it is also the UMP test of H_0 against H_1.

Note. As a final warning, note that p-values can depend on the experimental procedures, even if the data is identical, because they may vary the sample space. For example, suppose that we test if a coin is biased with $p > 1/2$, $H_0 : p = 1/2$. The coin is flipped five times, giving H, H, H, H, T.

If the experimental procedure was to flip the coin five times and count the number of heads, then clearly the best test is to reject H_0 if the number of heads H is large. Then

$$p^* = \mathbb{P}(H \geq 4) = 0.1875.$$

Alternatively, if the procedure was to flip the coin until a tail is seen, then the best test is to reject H_0 if the number of tosses N is large. Then we get the very different p-value,

$$p^* = \mathbb{P}(N \geq 5) = 0.0625.$$

In both cases the null hypothesis is the same, but we need to know how the data was collected to evaluate the p-value. This is a serious issue in experiments which naturally go on continuously in time: one must have a concrete stopping rule, and use it when computing the p-value, or else the results will be biased.

The conceptual reason for the difference is that the two different experiments have different notions of data being "at least as extreme" as that observed. This was famously criticized by Jeffreys, who stated:

> [The use of the p-value implies that] a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.

A benefit of Bayesian hypothesis testing is that it avoids this issue: as long as we know we have all the data that was collected, and know that it was collected properly, we just update our odds by its likelihood ratio. That is, the Bayesian approach obeys the "likelihood principle", that the only thing that matters is the likelihood of the data. (No matter what the stopping rule is, the expected proportion of heads for a fair coin is exactly $1/2$, by the linearity of expectation.) Note that one needs to have *all* the data to do this properly; it would be thrown off by publication bias.

3.4 Generalized Likelihood Tests

With the background now in place, we introduce a slew of practical tests. In these tests, the null hypothesis is a *subset* of the alternative hypothesis; in other words, the purpose of the test is solely to see if we can reject the null hypothesis, without any specific alternative in mind. Explicitly,

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta, \quad \Theta_0 \subset \Theta.$$

Working in the parametric framework, suppose that Θ_0 is a submanifold of Θ, with dimension lower by p. Then we have the following theorem.

Theorem (Wilks)**.** If $\Theta_0 \subset \Theta$ satisfies certain conditions, and has dimension lower by p, then as $n \to \infty$ with iid data $\mathbf{X} = (X_1, \dots, X_n)$, then

$$2 \log L_{\mathbf{X}}(H_0, H_1) \sim \chi^2_p$$

if H_0 is true. If H_0 is not true, the left-hand side tends to be larger. We can reject H_0 if $2 \log L_{\mathbf{x}} > c$, where $\alpha = \mathbb{P}(\chi^2_p > c)$ to give a test of size approximately α.

Example. The two-tailed z-test again. Let $X_1, \dots X_n \sim N(\mu, \sigma^2)$ be independent, where σ^2 is known, and

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \text{ unconstrained.}$$

The likelihood is maximized over H_1 when μ is equal to the sample mean, so

$$L_{\mathbf{x}}(H_0, H_1) = \frac{\sup_\mu f(x|\mu, \sigma^2)}{f(x|\mu_0, \sigma^2)} = \frac{\exp(-\sum_i (x_i - \overline{x})^2/2\sigma^2)}{\exp(-\sum_i (x_i - \mu_0)^2/2\sigma^2)} = \exp(n(\overline{x} - \mu_0)^2/2\sigma^2).$$

In other words, we reject the null hypothesis if the sample mean $\overline{X} \sim N(\mu_0, \sigma^2/n)$ is far from μ_0, which makes sense. The point of this example is that in this case, the theorem above holds exactly. Note that we can write

$$\log L_{\mathbf{x}}(H_0, H_1) = \frac{1}{2} \left(\frac{\sqrt{n}(\overline{x} - \mu_0)}{\sigma} \right)^2 = \frac{Z^2}{2}$$

where $Z \sim N(0, 1)$ is standard normal. Then $2 \log L_{\mathbf{X}}(H_0, H_1) = Z^2 \sim \chi^2_1$.

Example. Let $X_1, \dots X_n \sim N(\mu, \sigma^2)$ be independent again, but now suppose μ is known, and

$$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 \text{ unconstrained.}$$

The likelihood is maximized over H_1 when

$$\sigma^2 = \sigma^2_{\text{opt}} = \frac{1}{n} \sum_i (x_i - \mu)^2$$

which gives the likelihood ratio

$$L_{\mathbf{x}}(H_0, H_1) = \frac{\sup_{\sigma^2} f(x|\mu, \sigma^2)}{f(x|\mu, \sigma_0^2)} = \frac{(2\pi\sigma^2_{\text{opt}})^{-n/2} \exp(-n/2)}{(2\pi\sigma_0^2)^{-n/2} \exp(-\sum_i (x_i - \mu)^2/2\sigma_0^2)}.$$

This can be written simply in terms of the test statistic $t = \sigma^2_{\text{opt}}/\sigma_0^2$, which gives

$$2 \log L_{\mathbf{x}}(H_0, H_1) = n(t - 1 - \log t).$$

This is another case where the theorem above is exact, because $T \sim \chi^2_n$ by the definition of the χ^2 distribution.

We should reject H_0 when t is far from 1. There is not a clearly best choice of the critical region, because the alternative hypothesis is not simple. However, a sensible prescription is to define the critical region symmetrically on the distribution of T. That is,

$$\text{reject } H_0 \text{ if } t > F^{(n)}_{\alpha/2} \text{ or } t < F^{(n)}_{1-\alpha/2}.$$

Note that this is just the two-tailed version of a test in the previous section.

Note. The above test is called a χ^2-test because the test statistic is χ^2-distributed. In other words, to perform the test in practice, one would go to a table of χ^2 inverse cdf values. However, sometimes people say misleadingly that "χ^2-tests are only for categorical data". There is no deep meaning to this statement; people only say it because the most common tests for continuous data involve comparing means, where χ^2 distributions are not involved.

Example. The two-sample two-tailed z-test. Consider the independent data points

$$X_1, \dots, X_m \sim N(\mu_1, \sigma^2), \quad Y_1, \dots, Y_n \sim N(\mu_2, \sigma^2)$$

where σ^2 is known, and

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_i \text{ unconstrained.}$$

The likelihood must be maximized for both H_0 and H_1, and in both cases results in setting the means to the relevant sample means. If $\overline{w}$ is the mean of the combined samples,

$$L_{\mathbf{x}}(H_0, H_1) = \frac{\sup_{\mu_1,\mu_2} f(\mathbf{x}|\mu_1, \sigma^2) f(\mathbf{y}|\mu_2, \sigma^2)}{\sup_{\mu} f(\mathbf{x}|\mu, \sigma^2) f(\mathbf{y}|\mu, \sigma^2)} = \frac{f(\mathbf{x}|\overline{x}, \sigma^2) f(\mathbf{y}|\overline{y}, \sigma^2)}{f(\mathbf{x}|\overline{w}, \sigma^2) f(\mathbf{y}|\overline{w}, \sigma^2)}$$

Simplifying this expression leads to

$$2 \log L_{\mathbf{x}}(H_0, H_1) = \frac{mn}{m+n} \frac{(\overline{x} - \overline{y})^2}{\sigma^2}.$$

We should reject H_0 when $|\overline{x} - \overline{y}|$ is large. On the other hand,

$$Z = \left(\frac{1}{m} + \frac{1}{n} \right)^{-1/2} \frac{\overline{X} - \overline{Y}}{\sigma} \sim N(0, 1)$$

which tells us the right-hand side above is distributed as χ^2_1, so the theorem is again exact. Again, it is ambiguous how to define the critical region, but a sensible choice is to reject H_0 if $|z| > z_{\alpha/2}$.

Example. "The" χ^2-test. Consider n iid trials, each with k possible outcomes. The data is a histogram of outcomes $(x_1, \dots, x_k)$ where $\sum_i x_i = n$. Let p_i be the probability of outcome i. Then we can test

$$H_0 : p_i = p_i(\theta) \text{ for } \theta \in \Theta_0, \quad H_1 : p_i \text{ unconstrained.}$$

This is called a goodness-of-fit test, since H_0 specifies the whole distribution, while under the assumption that the trials are iid, H_1 is completely general. In any case, as long as the trials are iid, the outcomes follow a multinomial distribution,

$$\mathbb{P}(\mathbf{x}|p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

so the log likelihood is

$$\log f(\mathbf{x}|p) = \text{const} + \sum_i x_i \log p_i.$$

For the alternative hypothesis, this must be maximized under the constraint $\sum_i p_i = 1$, and using Lagrange multipliers straightforwardly yields $\hat{p}_i = x_i/n$, which is intuitive. For the null hypothesis, this is maximized over Θ_0.

Let $\hat{\theta}$ be the MLE of θ under H_0. Then

$$2 \log L_{\mathbf{x}}(H_0, H_1) = 2 \sum_i \log(\hat{p}_i / p_i(\hat{\theta}))$$

This can be recast in a more simply computable form. Define

$$o_i = x_i, \quad e_i = np_i(\hat{\theta}), \quad \delta_i = o_i - e_i$$

where o_i represents the outcomes, and e_i the expected outcomes under H_0. Then

$$2 \log L_{\mathbf{x}}(H_0, H_1) = 2 \sum_i o_i \log(o_i/e_i) = 2 \sum_i (\delta_i + e_i) \log(1 + \delta_i/e_i).$$

Assuming the deviations are small, $\delta_i/e_i \ll 1$, we can expand the logarithm to first order, giving

$$2 \log L_{\mathbf{x}}(H_0, H_1) \approx 2 \sum_i (\delta_i + e_i)(\delta_i/e_i) = \sum_i \frac{\delta_i^2}{e_i} = \sum_i \frac{(o_i - e_i)^2}{e_i} = \left(\sum_i \frac{o_i^2}{e_i} \right) - n.$$

This is the Pearson χ^2 statistic, in several forms. Of course, if the deviations were not small this approximation wouldn't work, but if n is large and the deviations are not small, it's clear that H_0 should be rejected, without running a test at all. A rule of thumb is that the null hypothesis will be rejected if the χ^2 statistic is significantly greater than the number of degrees of freedom.

Now we can see our above theorem at work. For a large sample size, the deviations can be approximated as continuous, and δ_i/e_i becomes normally distributed, so the χ^2 statistic indeed follows a χ^2 distribution. For H_1, there are $k-1$ parameters to choose, since the probabilities sum to one. For H_0, suppose there are p parameters to choose. Then if H_0 is true,

$$2 \log L_{\mathbf{x}}(H_0, H_1) \sim \chi^2_{k-p-1}$$

We will give some more specific examples below.

Note. The advantage of "the" χ^2-test is that it is easy to use. However, for smaller sample sizes the approximations made above are less accurate. In these cases one should use the more general "multinomial test", working directly with the unapproximated likelihood ratio and its more complicated distribution. Also, in some applications the trials are not independent, but are rather drawn from a fixed population without replacement. If the population is not much larger than the sample size, then one should use the hypergeometric test.

Example. The χ^2-test of homogeneity. Consider a rectangular array X_{ij} with m rows and n columns. Define the row, column, and overall sums by

$$X_{i\cdot} = \sum_j X_{ij}, \quad X_{\cdot j} = \sum_i X_{ij}, \quad X_{\cdot\cdot} = \sum_{ij} X_{ij}.$$

Suppose that each row i is described with a multinomial distribution, where entry j has probability p_{ij}, and the row sum $X_{i\cdot}$ is fixed. Then one can take

$$H_0 : p_{ij} = p_j \text{ for all } i, \quad H_1 : p_{ij} \text{ unconstrained.}$$

For example, rows may indicate outcomes for patients, with or without a certain intervention, in which case the null hypothesis is that the intervention has no effect.

We must now pick probabilities $\hat{p}_j$ and $\hat{p}_{ij}$ that maximize the likelihoods for H_0 and H_1, where we each hypothesis we have likelihood

$$\log f(x) = \text{const} + \sum_{ij} x_{ij} \log p_{ij}.$$

Using Lagrange multipliers again, one can show that this is achieved when the probabilities match the empirically observed ones,

$$\hat{p}_j = \frac{x_{\cdot j}}{x_{\cdot\cdot}}, \quad \hat{p}_{ij} = \frac{x_{ij}}{x_{i\cdot}}.$$

Plugging these in, we have

$$2 \log L_x(H_0, H_1) = 2 \sum_{ij} x_{ij} \log(x_{ij} x_{\cdot\cdot} / x_{i\cdot} x_{\cdot j}).$$

Defining the observed and expected counts

$$o_{ij} = x_{ij}, \quad e_{ij} = \hat{p}_j x_{i\cdot}$$

and using the same approximations as in the previous example, we find

$$2 \log L_x(H_0, H_1) = \sum_{ij} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

so the test statistic has the same basic form. The hypothesis H_0 has $n - 1$ parameters while H_1 has $m(n-1)$ parameters, so there are $(m-1)(n-1)$ degrees of freedom.

Example. Consider the same setup as above, but the row sums are not fixed; instead the overall sum is fixed, and all of the elements of the array are part of one common multinomial distribution. We take the hypotheses

$$H_0 : p_{ij} = p_i q_j, \quad H_1 : p_{ij} \text{ unconstrained}$$

where the p_i and q_j are row and column probability distributions. In this case, the rows and columns can represent two independent variables, measured in a fixed number of trials, and the null hypothesis is that these variables are independent.

The same approach as above gives

$$\hat{p}_i = \frac{x_{i\cdot}}{x_{\cdot\cdot}}, \quad \hat{q}_j = \frac{x_{\cdot j}}{x_{\cdot\cdot}}, \quad \hat{p}_{ij} = \frac{x_{ij}}{x_{\cdot\cdot}}.$$

Upon defining $o_{ij} = x_{ij}$ and $e_{ij} = \hat{p}_i \hat{q}_j x_{\cdot\cdot}$ and applying the same approximations, the test statistic has the same form as above. The hypothesis H_0 has $m + n - 2$ parameters while H_1 has $mn - 1$, so there are $(m-1)(n-1)$ degrees of freedom. Notice that this is exactly the same result as the previous test: ultimately the only difference is whether we fix the row sums or only the overall sum, and it turns out that it doesn't matter.

Note. Frequentist tests can lead to unintuitive results. For example, consider an experiment that tests for a relationship between sex and eye color, with the following results.

	blue	brown
male	20	10
female	10	20

One statistician may hold the null hypothesis H_0 that sex and eye color are independent; applying a test above we find that H_0 is rejected at the 1% level. Another statistician may hold the null hypothesis H_0' that all combinations of sex and eye color have 25% probability; applying a test above we find H_0' is not rejected even at the 5% level. Thus we should disbelieve H_0 and believe H_0', even though H_0' logically implies H_0!

This result makes sense if one invokes some Bayesian ideas about how hypotheses are evaluated. From this perspective, H_0 is a set of hypotheses (parametrized by the overall fraction of males, and fraction of blue eyed people), and H_0' is a subset of H_0 (the case where these fractions are both 1/2). We begin with some overall prior $p(H_0)$, which itself includes a prior distribution over the fractions. The effect of the data is to strongly penalize H_0 as a whole. However, the part of H_0 that gets penalized the least is H_0'. In other words, after seeing the data, we will of course still have $p(H_0) \geq p(H_0')$, but $p(H_0')/p(H_0)$ has increased.

We can also explain the result in another way. The test statistics in the two cases are identical; the only difference is the number of degrees of freedom. The hypothesis H_0 uses more degrees of freedom to explain the data, but they don't help at all. Holding predictivity equal, we should prefer simpler theories, so it makes sense to reject H_0 but not H_0'.

This highlights another issue with hypothesis testing: if somebody were planning on testing H_0' and saw that it wasn't significant upon seeing the data, they could make their result significant by changing their null hypothesis to H_0. In this way, a "significant" p-value can be found out of essentially any data. In fact, this can happen innocently, without any active "p-hacking" or "fishing expedition" on the part of the researcher, if the statistical test chosen is conditional on the data. For some cautionary examples, see *The Statistical Crisis in Science*.

3.5 The t and F Tests

In this section, we introduce some more practical tests.

- If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ independently of X, then

$$Z = \frac{X}{(Y/n)^{1/2}} \sim t_n, \quad \mathbb{E}(Z) = 0, \quad \operatorname{var} Z = \frac{n}{n-2}$$

  where t_n is the Student's t-distribution on n degrees of freedom. Its pdf is

$$f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}.$$

  In the limit $n \to \infty$ the t-distribution approaches $N(0,1)$, but for finite n it has heavier tails.

- As mentioned earlier, this distribution is useful when we have normally distributed data with unknown variance, which is estimated from the sample variance, because

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1}$$

  where $S_{XX}/(n-1)$ is the unbiased estimator of the variance. The point here is that the distribution of T does not depend on σ^2 which may be unknown.

- Therefore, given independent $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and σ^2 unknown, we can construct a $100(1-\alpha)\%$ confidence interval for μ by defining T as above, and taking

$$-t_{\alpha/2}^{(n-1)} \leq t \leq t_{\alpha/2}^{(n-1)}$$

  where $t_{\alpha/2}^{(n-1)}$ is the upper $\alpha/2$ point of t_{n-1}.

- Note that for low n, the t-distribution is pathological; for example, for $n = 1$ it reduces to the Cauchy distribution, and for $n \leq 2$ its variance does not exist.

Example. The one sample t-test for testing a given mean. Starting from the same assumptions as above, we test

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \text{ unconstrained}$$

where σ^2 is unknown. It is a "nuisance parameter" which must be accounted for, but which is not directly relevant. The likelihood ratio is

$$L_{\mathbf{x}}(H_0, H_1) = \frac{\max_{\mu,\sigma^2} f(x|\mu, \sigma^2)}{\max_{\sigma^2} f(x|\mu_0, \sigma^2)}.$$

We have found in a previous example that the likelihood in the denominator is maximized when

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2.$$

Meanwhile the likelihood in the numerator is maximized when

$$\mu = \overline{x}, \quad \sigma^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2 = \frac{s_{xx}}{n}.$$

By plugging these in and simplifying the likelihood ratio, we arrive at

$$L_{\mathbf{x}}(H_0, H_1) = \left(1 + \frac{n(\overline{x} - \mu_0)^2}{\sum_i (x_i - \overline{x})^2}\right)^{n/2}.$$

In other words, the likelihood ratio depends only on the statistic T as defined above, where under H_0 we have $T \sim t_{n-1}$. Thus, the size α test constructed by rejecting H_0 if $|t| > t_{\alpha/2}^{(n-1)}$ is also a generalized likelihood ratio test.

Example. The two sample t-test for testing equality of means. Consider two independent samples of independent variables, where $X_1, \dots, X_m \sim N(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_n \sim N(\mu_2, \sigma^2)$. We test the hypotheses

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_i \text{ unconstrained}$$

where σ^2 is unknown, but assumed to be common between the samples. A similar argument to above shows that the likelihood ratio only depends on the quantity $(\overline{x} - \overline{y})/(s_{xx} + s_{yy})$, which motivates us to look at its distribution under H_0. This is a case where the framework of the likelihood ratio pays off, since it's not intuitively obvious what combination of s_{xx} and s_{yy} should go in the denominator.

Under H_0, we have

$$(\overline{X} - \overline{Y})\left(\frac{1}{m} + \frac{1}{n}\right)^{-1/2} \frac{1}{\sigma} \sim N(0, 1), \quad \frac{S_{XX} + S_{YY}}{\sigma^2} \sim \chi^2_{m+n-2}.$$

Therefore, by the definition of the Student's t-distribution,

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{(1/m + 1/n)(S_{XX} + S_{YY})/(m+n-2)}} \sim t_{m+n-2}.$$

Thus, we can construct a size α test by rejecting H_0 if $|t| > t^{(m+n-2)}\alpha/2$.

Note. Choosing a test can be subtle. For example, consider an intervention meant to reduce resting heart rate. One can either use the two-sample t-test for the heart rates of the participants before and after, or apply the one-sample t-test to the set of changes in heart rates, where the null hypothesis is $\mu = 0$. The latter is called a paired samples t-test and is a much better choice here. First, there's no reason to believe that the initial and final pulse rates are even close to normally distributed; the differences stand a better chance. Second, the two-sample test will be much less powerful, because it doesn't know about the pairing, and so is thrown off by the wide spread of initial pulse rates. On the other hand, the paired samples t-test only makes sense if the "before" and "after" quantities can be usefully compared; for example, it would be useless when considering the result of an intervention on an acute disease.

Example. Testing a given variance with unknown mean. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with μ unknown and

$$H_0 : \sigma^2 = \sigma_0^2, \quad H_1 : \sigma^2 \text{ unconstrained.}$$

In other words, we have reversed which parameter is the nuisance parameter. Using earlier results,

$$L_{\mathbf{x}}(H_0, H_1) = \frac{\max_{\mu,\sigma^2} f(\mathbf{x}|\mu,\sigma^2)}{\max_{\mu} f(\mathbf{x}|\mu,\sigma_0^2)} \propto (s_{xx})^{-n/2} e^{s_{xx}/2\sigma_0^2}.$$

This only depends on s_{xx}, which motivates us to consider the test statistic $T = S_{XX}/\sigma_0^2$, where H_0 will be rejected if it is far from 1. Under H_0, $T \sim \chi^2_{n-1}$, so one possible test of size α is

$$\text{reject } H_0 \text{ if } T \notin [F^{-1}_{n-1}(\alpha/2), F^{-1}_{n-1}(1-\alpha/2)].$$

Note. Nuisance parameters are another case where frequentist and Bayesian approaches differ. In our simple examples above, we have constructed test statistics which are independent of the nuisance parameter. **(finish)**

Next, we introduce the F-test and analysis of variance.

- Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent. Then

  $$Z = \frac{X/m}{Y/n} \sim F_{m,n}$$

  has the F-distribution on m and n degrees of freedom. The probability distribution is

  $$f(x) = \frac{1}{B(m/2, n/2)} \left(\frac{m}{n}\right)^{m/2} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}.$$

- If $T \sim F_{m,n}$ then $1/T \sim F_{n,m}$. Statistical tables usually only give upper percentage points for the F-distribution, but we can find $\mathbb{P}(T < x)$ since it is equal to $\mathbb{P}(1/T > 1/x)$, which is listed elsewhere in the table.

- If $X \sim t_n$, then $X^2 \sim F_{1,n}$.

Example. The two-sample comparison of variances, also known as "the" F-test. Consider two independent samples of independent variables, where $X_1, \dots, X_m \sim N(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_n \sim N(\mu_2, \sigma_2^2)$ where

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2$$

where the μ_i are unknown nuisance parameters. Using either the likelihood ratio or common sense, we consider the statistic

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma_2}^2} = \frac{S_{XX}/(m-1)}{S_{YY}/(n-1)} \sim \frac{\sigma_1^2 \chi_{m-1}^2/(m-1)}{\sigma_2^2 \chi_{n-1}^2/(n-1)} = \frac{\sigma_1^2}{\sigma_2^2} F_{m-1,n-1}.$$

Therefore, under H_0 we have $F \sim F_{m-1,n-1}$. Since the alternative hypothesis is one-tailed, we get the greatest power if we reject H_0 if $f > F_\alpha^{(m-1,n-1)}$.

4 Applications in Particle Physics

4.1 Classification

In this section, we discuss how statistics is actually used in particle physics, departing from the clean mathematical formalism above. We will be a little sloppier with notation, e.g. not always distinguishing a random variable and its value. First, we consider the relatively straightforward case of classification.

- Above, we have discussed the likelihood ratio as a test statistic. However, in a particle physics experiment where one needs to classify particles and events, this is not very useful because the data $\mathbf{x}$ in each event is extremely high-dimensional. (Elements of $\mathbf{x}$ might include the number of muons, the p_T of the hardest jet, the missing energy, and so on.) Furthermore, the likelihoods cannot be computed from first principles, and instead require Monte Carlo simulation.

- Therefore, it is practical to assume that the test statistic $T(\mathbf{x})$ has a simple prescribed form, then optimize it given that form. One simple option is a linear function,

  $$t(\mathbf{x}) = \mathbf{a}^T\mathbf{x}.$$

  The goal is to choose $\mathbf{a}$ to maximize the separation between $g(t|H_0)$ and $g(t|H_1)$, where we are restricting to simple hypotheses. Of course, there is not a unique definition of "separation", so the result will depend on the definition.

- Under each hypothesis, the data $\mathbf{x}$ have the mean values and covariance matrix

  $$\boldsymbol{\mu}_k = \int \mathbf{x} f(\mathbf{x}|H_k)\, d\mathbf{x}, \quad V_k = \int (\mathbf{x}-\boldsymbol{\mu}_k)(\mathbf{x}-\boldsymbol{\mu}_k)^T f(\mathbf{x}|H_k)\, d\mathbf{x}.$$

  Thus, each hypothesis gives a mean and variance for the test statistic,

  $$\tau_k = \mathbf{a}^T\boldsymbol{\mu}_k, \quad \Sigma_k^2 = \mathbf{a}^T V_k \mathbf{a}.$$

- We choose to define the separation by

  $$J(\mathbf{a}) = \frac{(\tau_0-\tau_1)^2}{\Sigma_0^2+\Sigma_1^2}$$

  on the grounds that it behaves reasonably under shifts and overall scaling. Some routine calculation shows that this is maximized for

  $$\mathbf{a} \propto (V_0+V_1)^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1).$$

  The resulting test statistic is called Fisher's linear discriminant function.

- In order to use this classifier, the quantities V_k and $\boldsymbol{\mu}_k$ must still be found by Monte Carlo simulations, but there are much fewer relevant quantities. Specifically, if $\mathbf{x}$ is n-dimensional, we need to compute $O(n^2)$ quantities.

- For comparison, the full likelihood ratio is a function on n-dimensional space, so it is technically infinite-dimensional. Of course, we always perform binning to render everything finite-dimensional. Suppose that every dimension gets m bins; then a likelihood ratio test would require computing $O(m^n)$ quantities, which is far greater than $O(n^2)$. This is worsened by the fact that we generally want bins as fine as possible, since coarse bins throw away information.

We now motivate some more complex test statistics.

- Suppose that the hypotheses H_0 and H_1 are both multivariate normal distributions with the same covariance matrix V, so that $\mathbf{a} \propto V^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$.
- In this case, the likelihood ratio is

$$r \equiv L_{\mathbf{x}}(H_0, H_1) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T V^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T V^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right) \propto e^{t(\mathbf{x})}.$$

  In other words, a test based on Fisher's linear discriminant is a likelihood ratio test, which is another one of its nice properties.
- Now, let π_0 and π_1 be the probabilities that H_0 and H_1 are true. (This is usually not allowed in hypothesis testing, but makes sense for classification tasks.) Then by Bayes' theorem,

$$p(H_0|\mathbf{x}) = \frac{1}{1+\pi_1/\pi_0 r} = \frac{1}{1+e^{-t(\mathbf{x})}} \equiv s(t)$$

  as long as we add a suitable constant to t. In other words, the probability is a sigmoid function in the test statistic.
- If we fed the result of this classifier into another classifier, which satisfied the same assumptions, then the net result of these classifiers would be an alternating composition of sigmoids and linear functions. This motivates the choice of using more general test statistics in terms of such functions, since we will "effectively" get them anyway upon composition.
- Specifically, we could define a test statistic of the form

$$t(\mathbf{x}) = s\left(a_0 + \sum_i a_i h_i(\mathbf{x})\right), \quad h_i(\mathbf{x}) = s\left(w_{i0} + \sum_i w_{ij} x_j\right).$$

  This is a simple feed-forward neural network. We think of the h_i as "neurons", connected to the inputs x_i with "weights" w_{ij}. They are then connected to the output by the weights a_i. Cutting on the value of t can now give a nonlinear boundary for C in parameter space.

Note. In the past ten years, there has been great interest in using "jet substructure" to identify the particles produced in a collision. The classic example is a top quark, which almost always decays to a b-quark and a W boson. The W boson usually decays of pairs of quarks, so that three jets are produced. However, at the high energies of the LHC, the top quark is produced with relativistic energy, causing the jets to be collimated into a single "fat" jet, with a three-pronged substructure.

In the early 2010s, various physics-motivated approaches were invented to identify top jets. For example, one can tag the b-jet, since it has a characteristic displaced vertex, and demand the other two jets have an invariant mass near m_W. Alternatively, one can run a clustering algorithm to see whether three subcomponents exist in the jet, as quantified by a variable called the N-subjettiness. These are human-engineered "high-level" features, but in the late 2010s, physicists began throwing deep neural networks at the problem, feeding them "low-level" features like the four-momenta of all the hadrons in the jet. They are trained using simulated data from Monte Carlo programs, such as Pythia, and typically have about a million parameters. A vast number of neural network architectures have been tried, with the state of the art outperforming the high-level features by roughly a factor of 3 in background rejection for order-one signal efficiencies.

4.2 Signal and Exclusion

We now discuss the basics of evaluating the significance of a signal.

- Suppose that one is counting events of a particular type. This can be modeled as a Poisson process, with contributions from both signal and background. Suppose there are n_{obs} observed events, and the background has a known mean ν_b. The null hypothesis is that the signal has zero mean, $\nu_s = 0$.

- The p-value is then the probability of seeing at least this many events,

$$p = \mathbb{P}(n \geq n_{\text{obs}}) = \sum_{n=n_{\text{obs}}}^{\infty} \frac{\nu_b^n}{n!} e^{-\nu_b n}.$$

  For example, if $\nu_b = 0.5$ and $n_{\text{obs}} = 5$, then $p = 1.7 \times 10^{-4}$.

- Note that the total number of counts is Poisson distributed. Thus, if we used the naive rules of error analysis, we might write

$$\nu_b + \nu_s = n_{\text{obs}} \pm \sqrt{n_{\text{obs}}}$$

  which in the case $n_{\text{obs}} = 5$ would only be "two sigma" away from the background rate. However, this is severely misleading. When we talk about a result being "$n\sigma$", we only mean that its p-value is lower than the threshold $1 - \Phi(n)$,

$$1\sigma : p \leq 0.16, \quad 2\sigma : p \leq 0.023, \quad 3\sigma : p \leq 0.0013, \quad 4\sigma : p \leq 0.000032, \quad 5\sigma : p \leq 0.00000029$$

  even if the normal distribution is completely irrelevant. We take 5σ to denote discovery, and 3σ and above to denote a "hint". The range $n_{\text{obs}} \pm \sqrt{n_{\text{obs}}}$ above does make sense, but it's relevant for bounding ν_s once we know it is nonzero, not for establishing that it is nonzero.

- It's important not to take extremely small p-values too seriously. In the Bayesian perspective, one's belief in a signal should increase as the p-value decreases, but it should eventually saturate, because of the possibility of systematic errors, which are not treated by anything in these notes.

- Another issue is that of the "look elsewhere effect". In practice, most searches will involve looking for a bump on a histogram. This is a job for the χ^2-test, where the null hypothesis is the SM expectation. However, if the histogram has 10^3 bins, then we expect a 3σ deviation in some bin even if there is no signal, just by random chance. Instead, the p-value should be computed "globally", by finding the probability for seeing fluctuations as severe as those observed *anywhere* across the histogram. One could also argue that an additional correction is necessary if one looks at many histograms, but this is harder to make precise. The standard methodology for correction for the look elsewhere effect is given in the article *Trial Factors for the Look Elsewhere Effect in High Energy Physics*.

- Incidentally, the look elsewhere effect is precisely the reason that 5σ is used as a discovery standard today. In the 1960s, several claimed discoveries did not pan out, despite low p-values by the standards of other sciences. The 5σ criterion was proposed as a crude way to compensate for the uncorrected look elsewhere effect. It has been argued that the 5σ criterion be allowed to vary between experiments, accounting for the degree of surprise of the result, the amount of uncompensated look elsewhere effect, the risk of systematics, and so on.

- Histogram bins should in theory be as narrow as possible, to reduce the amount of information that is thrown out. However, if they are too narrow, the fluctuations become large and the plot becomes impossible to read. A general rule of thumb is to bin finely enough so that any expected peaks will be resolved by several bins. If the global correction is applied properly, binning more finely shouldn't have much effect on the p-value. In the case of very small data samples, it's better to not bin at all, and use goodness of fit tests designed for continuous variables, such as Kolmogorov–Smirnov or Smirnov–Cramer–von Mises.
- The bin counts in a one-dimensional histogram follow a multinomial distribution regardless of how the underlying continuous variable is distributed; we say the test is distribution-free. However, in general the situation is more complicated, and the p-value can't be computed analytically or with standard lookup tables. Instead, a common procedure is to use Monte Carlo to infer the distribution of the test statistic.

We now briefly discuss how parameter exclusion is done.

- Parameter exclusion refers to rejecting some subset of the *alternative* hypothesis H_1. This isn't covered as much in traditional statistics courses, because the focus there is on excluding null hypotheses. However, in particle physics the null hypothesis (the Standard Model) works an overwhelming majority of the time, so it would be impractical to only publish results that reject it. Publishing exclusions is also very useful to keep track of what we have ruled out.
- For concreteness, suppose we reduce a statistical analysis down to a single test statistic T, and suppose that deviations from the null hypothesis can only increase the value t. (This could apply, e.g. to counts in a histogram.) The test statistic has distributions $f(t|H_0)$ and $f(t|H_1)$ under each hypothesis, and we define

  $$p_0 = \mathbb{P}(T \geq t|H_0), \quad p_1 = \mathbb{P}(T \leq t|H_1).$$

  We decide whether the data indicates a discovery by looking at p_0, as described above.
- We can define exclusion by looking at p_1, and a conventional choice is to set $p_1^* = 0.05$. This is much less stringent than the cutoff for discovery because false alarms have greater costs. Alternatively, it is because the SM is an excellent null hypothesis with a very high prior, while individual BSM theories have low prior probabilities due to their great number.
- However, this can raise the possibility of spurious exclusion. For example, suppose an experiment has no sensitivity to H_1 at all, i.e. the distributions $f(t|H_0)$ and $f(t|H_1)$ are the same. Then with some low probability, such an experiment may exclude H_1, though it seems unreasonable that any exclusions could be set at all. One ad hoc fix is to exclude when $p_1/(1-p_0) < 0.05$. Note that if H_0 and H_1 are point hypotheses, this is just the Bayesian update factor for $p(H_1)/p(H_0)$.
- In general we deal with a family of alternative hypotheses, parametrized by new couplings. The procedure above thus allows us to exclude a range of these couplings.

Note. In Bayesian hypothesis testing, one could report the Bayesian update factor for $p(H_0)/p(H_1)$. However, this runs into issues such as Lindley's paradox. Suppose we are testing whether a coin is fair. A reasonable Bayesian prior on the probability of heads p_h might be

$$p(p_h) = \frac{1}{2}\,\delta\left(p_h - \frac{1}{2}\right) + \frac{1}{2}$$

representing a 50/50 chance of a fair coin, and a rigged coin with uniform probability. This is a toy model for new physics searches, where the fair coin corresponds to the SM (a simple null hypothesis) and the biased coin corresponds to new physics (a composite alternative hypothesis).

Now suppose the coin is flipped 10^5 times, registering $(10^5/2) + \sqrt{10^5}$ heads. Under frequentist hypothesis testing, $H_0 : p_h = 1/2$ is rejected at the 2σ level. But after performing the Bayesian update, the probability for the coin to be fair becomes much higher! The reason is that the frequentist method tests H_0 in itself, while the Bayesian method is effectively comparing it against a particular alternative, i.e. a uniform distribution on p_h, and most values of $p_h \in [0, 1]$ are strongly disfavored by the data.

Therefore, there is no logical paradox in Lindley's paradox, but it does highlight an important difference between the two methods. The Bayesian method gives a penalty to H_1 for having a broad prior, most of which is incompatible with the data. From a model builder's perspective I think this is right: it's just the fine tuning penalty in a different form. In other words, the Bayesian method accurately tracks how theorists think about evidence. More discussion of the Bayesian approach to theory evaluation and its relation to naturalness is given in my dissertation.

The issue that Lindley's paradox reveals is essentially a more subtle aspect of prior dependence. For simple hypotheses, the prior dependence isn't too important because one can just publish Bayesian update factors. For composite hypotheses, the detailed prior distribution can significantly affect the effect of the evidence on the overall probabilities for the hypotheses – and many realistic hypotheses are not just composite as in Lindley's paradox, but multi-dimensional. In these cases, evidence cannot be summarized in a single number, making the Bayesian method impractical for reporting experimental results; this is why experimentalists generally prefer p-values. But one could just as well argue that Lindley's paradox shows that p-values give incomplete information.

Luckily, these caveats don't apply to exclusion, which is less confusing. When excluding parameter space in the Bayesian picture, we don't have to compare point hypotheses to composite hypotheses. A point in parameter space can just be counted as excluded at 2σ if its probability density goes down by at least 95%.

Many existing resolutions of Lindley's paradox do not work in particle physics. For example, one could criticize putting a finite prior weight on a point, because in the softer sciences effectively every intervention can have some (small) effect. But in particle physics, the point represents the SM, which really could be true to a precision much greater than that of our experiments. More precisely, Lindley's paradox can occur if H_0 is much narrower than the experimental resolution, which in turn is much narrower than H_1, and this is often the case in particle physics. Another proposed resolution from the softer sciences is to start with $p(H_0) \ll 1/2$, because "the point of an experiment is to rule out H_0". But it would be odd to treat the venerable SM in this way; if anything, in most searches we should take $p(H_1) \ll 1/2$. Yet another approach is to declare that any effect small enough to be susceptible to Lindley's paradox is practically irrelevant. However, in particle physics *any* deviation from the SM is incredibly important. Finally, one could remove the arbitrariness in the Bayesian approach by fixing an objective prior such as Jeffreys' prior; however, it seems absurd to have our beliefs about the fundamental laws of nature decided by our measurement apparatus. Indeed, in particle physics the situation is the reverse: we must use our beliefs about these laws to decide which apparatuses to build! For a comprehensive review of arguments like these, see the article *The Jeffreys–Lindley Paradox and Discovery Criteria in High Energy Physics*.

Lindley's paradox reveals another issue. Suppose we were considering a parameter which was not bounded, like the probability of heads was. The resulting Bayesian update factor is *zero* if H_1 is given a uniform prior on an infinite interval, while introducing a cutoff makes the Bayesian update

factor explicitly cutoff-dependent. Such problems occur whenever the hypotheses H_0 and H_1 are defined on unbounded parameter spaces of differing dimension. Since the cutoff choice essentially determines the final reported result, one must think carefully about it.

4.3 Confidence Intervals

A separate issue is the practical construction of confidence intervals.

- Previously, we have described the frequentist definition for a confidence interval. However, in practice this definition is difficult to use beyond the simplest problems. Instead, we introduce some alternative procedures for constructing such intervals, which approximate the frequentist definition in certain limits.
- If we are using the MLE $\hat{\theta}$ to report the central value, then it is useful to report its variance (or equivalently standard deviation). This is reasonable, because it turns out that in the limit of many samples, the MLE has a Gaussian distribution, and furthermore the bias of the MLE usually goes to zero. In this case, the interval $[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% confidence interval in the frequentist sense, in addition to indicating the typical spread of results $\hat{\theta}$ in repeated experiments.
- There is a clear problem here, because $\hat{\theta}$ is a random variable which depends on the unknown true value θ, and thus $\sigma_{\hat{\theta}}(\theta)$ is a function of θ. We can avoid this problem by just *plugging in* the particular value of $\hat{\theta}$, which yields the estimate $\hat{\sigma}_{\hat{\theta}}(\hat{\theta})$.
- As long as the MLE doesn't have too high of a spread, this is a reasonable procedure. More concretely, $\hat{\sigma}_{\hat{\theta}}(\hat{\theta})$ is a random variable, and this procedure makes sense as long as *its* standard deviation $\sigma_{\hat{\sigma}_{\hat{\theta}}(\hat{\theta})}(\theta)$ is much smaller than its value.
- For example, if θ were the expected count of a Poisson process and $\theta = 10^4$, then

  $$\sigma_{\hat{\theta}}(\theta) = \sqrt{10000} = 100.$$

  Suppose that in a particular run, we got a value within this uncertainty, $\hat{\theta} = 10^4 - 10^2$. Then our estimate of the MLE standard deviation would be

  $$\hat{\sigma}_{\hat{\theta}}(\hat{\theta}) = \sqrt{10000 - 100} = 99.5$$

  which is definitely close enough; nobody would care about a small "uncertainty in uncertainty".
- In practice, we cannot compute $\sigma_{\hat{\theta}}(\theta)$ analytically. Instead, we can use Monte Carlo simulation to numerically compute the distribution of $\hat{\theta}$. Again we run into the problem that this distribution depends on the unknown true value of θ, and we circumvent it by plugging in $\hat{\theta}$. This will not give an accurate estimate of the distribution of $\hat{\theta}$ (it will be centered on the particular value $\hat{\theta}$ instead of $E[\hat{\theta}]$), but it will usually give a good estimate of the distribution's width.
- An even quicker way, used in many numeric programs, is to use the Cramer–Rao bound. We simply note that the MLE is asymptotically efficient, so

  $$\operatorname{var}(\hat{\theta}) \approx \left(\mathbb{E}\left[-\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right] \right)^{-1}$$

  where we defined the likelihood $L(\theta) = f(\mathbf{x}|\theta)$.

- Again, this requires knowledge of the true value θ, so we simply plug it in. Calculating the expectation value requires a potentially expensive Monte Carlo simulation, so we just replace it with the observed value. That is, we report

$$\widehat{\sigma^2}_{\hat{\theta}} = \left(-\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} \right)^{-1}.$$

  More generally, we have the estimated covariance matrix

$$(\widehat{V^{-1}})_{ij} = -\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

  Note that when such approximations apply, the covariance of multiple independent experiments can be combined in quadrature, using the standard rules of error analysis. The real issue is disentangling possible correlations between the experiments, a subtlety we won't discuss here.

- The log-likelihood can be expanded about its maximum in a Taylor series,

$$\log L(\theta) = \log L(\hat{\theta}) + \frac{1}{2}\frac{\partial^2 \log L}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \dots.$$

  In the large sample limit, the likelihood becomes a sharply peaked Gaussian, so these terms are an accurate approximation for the log-likelihood. In this case, the previous estimate for the variance is equivalent to the estimate

$$\log L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) = \log L(\hat{\theta}) - \frac{1}{2}.$$

  In cases where the likelihood isn't a Gaussian, we can use this as yet another definition of $\hat{\sigma}_{\hat{\theta}}$. This is especially useful when estimating multiple variables, where it allows the straightforward numeric construction of confidence regions.

Note. The general procedure of estimating parameters directly from the data, and then "plugging them in" to find a confidence interval, is called a "Wald interval". (One might also call this the rules of high school error analysis.) As we've seen, this yields a reasonable result if there is a large amount of good data, but it lacks the theoretical guarantees that genuine confidence intervals do. The classic pathological example is a rare Poisson process with zero observed counts; in this case the "plug in" mean and variance are zero, so all confidence intervals are $[0, 0]$, and all other values are excluded to infinite precision! As such, the Wald interval is generally regarded as obsolete in the statistical literature.

Next, we return to the frequentist definition of a confidence interval.

- It is useful to generalize our earlier definition of a confidence interval, so that

$$\mathbb{P}(\theta < a(X)) = \alpha, \quad \mathbb{P}(\theta > b(X)) = \beta.$$

  In this case we say $[a(X), b(X)]$ has coverage or confidence level $1 - \alpha - \beta$. Usually, two-sided confidence intervals are taken to be central confidence intervals, obeying $\alpha = \beta = \gamma/2$.

- When we include an estimator, confidence intervals need not be symmetric about it. In this case, we write $\hat{\theta}^{+d}_{-c}$ to mean the confidence interval is $[\hat{\theta} - c, \hat{\theta} + d]$.

- Previously, we considered simple examples of confidence intervals, where the dependence on the unknown variable θ could be eliminated. In more general situations we can use the Neyman construction. Let us define the functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ implicitly by

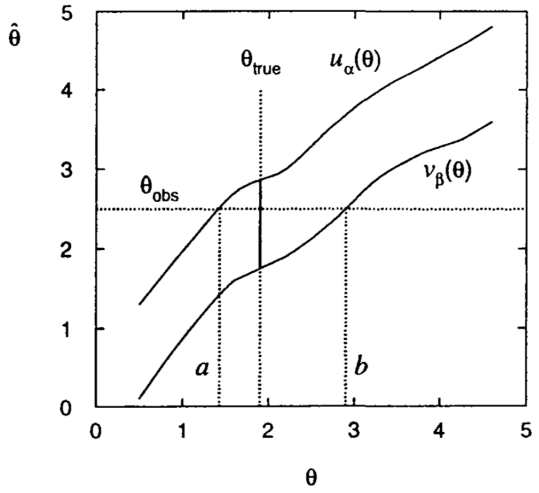
$$\alpha = \mathbb{P}(\hat{\theta} \geq u_\alpha(\theta)), \quad \beta = \mathbb{P}(\hat{\theta} \leq v_\beta(\theta))$$

  where $\hat{\theta}$ is some estimator. As long as this estimator is reasonable, the functions are monotonic, so we can invert them. The bounds for the confidence interval are then

$$a(\hat{\theta}) = u_\alpha^{-1}(\hat{\theta}), \quad b(\hat{\theta}) = v_\beta^{-1}(\hat{\theta}).$$

  Note that this doesn't work for discrete random variables, because the functions $u_\alpha(\theta)$ and $v_\beta(\theta)$ won't exist.

- The Neyman construction can be visualized graphically as shown.



  This method is quite general, but it's often intractable. In the large-sample limit where $\hat{\theta}$ is normally distributed, the result coincides with the approximate methods we gave above.

- Confidence intervals are often described as "$n\sigma$". This means that we have a confidence interval with $\alpha = \beta = \gamma/2$ where

$$1\sigma : 1 - \gamma = 0.6827, \quad 2\sigma : 1 - \gamma = 0.9544, \quad 3\sigma : 1 - \gamma = 0.9973.$$

  This language is used regardless of whether $\hat{\theta}$ is normally distributed. When a confidence interval is given without context, it's usually 1σ.

- One-sided confidence intervals have either α or β equal to zero, and correspond to parameter limits. For example, suppose we have a particular confidence interval $[-\infty, b(x)]$. Then by our earlier naive definition of exclusion, this excludes parameters $\theta > b(x)$ at the $100(1-\beta)\%$ level.

Example. Extended maximum likelihood is a tweak on maximum likelihood. In this case, we have an iid sample $x_1, \ldots, x_n$, but n itself is also a random variable. It is often the case that $n \sim \text{Pois}(\nu)$, since e.g. n could be the number of events passing cuts. In this case, we have the extended likelihood function

$$L(\nu, \boldsymbol{\theta}) = \frac{e^{-\nu}}{n!} \prod_i \nu f(x_i|\boldsymbol{\theta}).$$

If ν is independent of $\boldsymbol{\theta}$, then it is just a nuisance parameter and the MLE gives $\hat{\nu} = n$, along with the same $\hat{\theta}$ if we hadn't thought of n as a random variable at all. However, it is often the case that ν depends on $\boldsymbol{\theta}$ (for example, $\boldsymbol{\theta}$ may contain a cross section), in which case including it in the MLE leads to tighter confidence intervals. Also note that the MLE works just as well with binned data, though it doesn't actually require binning at all.

Example. A comparison of several different confidence intervals for the mean of a Poisson distribution. **(todo)**

Bayesian confidence intervals may have very little coverage, though under typical conditions all the confidence intervals above will match in the large sample limit. Incidentally, one can show that the Bayesian confidence interval calculated using Jeffreys' prior converges to the frequentist confidence interval the fastest.

Note. Estimators and confidence intervals near a physical limit. Suppose we have some quantity which is known on physical grounds to be positive, such as $m^2 = E^2 - p^2$, or the rate or scattering cross section for a process. If the true value is small, our estimators will often end up negative. For example, a downward fluctuation may cause us to fit a "valley" into a histogram when the signal can only look like a bump. If we are unlucky, our entire confidence interval can be negative, which sounds absurd.

It is tempting in these cases to set the estimator to zero or truncate the confidence interval, but this will lead to bias if the result is averaged with other experiments. (If we can never report downward fluctuations, then we'll only see upward fluctuations, leading to false discovery.)

These issues can also be handled with Bayesian confidence intervals. From the Bayesian perspective, a physical limit just corresponds to a region where the prior is zero. The Bayesian confidence interval can be constructed the same way as usual, and the results of multiple experiments can be combined by multiplying their Bayesian update factors.

5 Regression Models