Statistical Methods for
Astrophysics and Cosmology

Andrea Lapi

SISSA, Trieste, Italy
INFN-TS, Italy
INAF-OATS, Italy

room: A-509
phone: 470
email: lapi@sissa.it
Website: https://lapi.jimdo.com
PART II: Statistical inference

- Samples, statistics, and estimators
- Point estimation and Fisher information
- Interval estimation
- Resampling
- Hypothesis testing
- Kolmogorov-Smirnov nonparametric testing
- Regression and correlation
- Sufficiency
- Information, entropy, and priors
- Luminosity functions, counts, and biases
**Decision and inference**

Science is, at least in a broad sense, a decision process, often based on small-data samples. For example:

- measure a quantity (parameter estimation) → given some data, what is the best estimate (including uncertainty) of a particular parameter?
- search for correlations → are two measured variables correlated with each other, implying a possible physical connection?
- test a model (hypothesis testing) → given data and one or more models, are our data consistent with the models? Which model best describes the data?

**Statistics** are summarizing properties of the data, constructed from them, that constitute the basis toward decision-making.
So far, we have focused on considering given probability distributions for a random variable $X$, and on using them to compute various quantities. Now we are going to consider the case where the probability distribution for some random variable $X$ is at least partially unknown (e.g., its parameters); then we will focus on the prescriptions to make some statement about the unknown quantities given a particular set of measured values $X=x$. In other words, the goal of statistical inference is to take the outcome of an experiment to infer something about the validity of one or more hypothesis.

If we combine the random variables related to observables into a vector $\mathbf{X}$ and the parameters into a vector $\mathbf{\theta}$ then we have a pdf $f_{\mathbf{X}}(\mathbf{x};\mathbf{\theta})$. The idea is that when we perform an experiment, we get specific values $\mathbf{X}=\mathbf{x}$, and we can use in a way $f_{\mathbf{X}}(\mathbf{x};\mathbf{\theta})$ to infer something on $\mathbf{\theta}.$
A case of particular interest is when the variables $X_i$ in $X$ are independent and identically distributed, i.e., they form a random sample. Then

$$f_X(x;\theta) = \prod_i f(x_i;\theta)$$

To illustrate the different approaches to the problem, in the sequel we will consider the following example: suppose we have $n$ data points $x_i$, drawn from independent Gaussian distributions with the same unknown mean $\mu$ and different known variances $\sigma_i^2$. We are thus basically making $n$ independent measurements of the same unknown $\mu$, each with its own error $\sigma_i^2$. How can we use the measured values $x_i$ to say something about the $\mu$?
Bayesian point of view: posterior pdf

The Bayesian perspective treats the inherently random realization \( x \) and the unknown parameters \( \theta \) on the same footing, both representing uncertainty or lack of knowledge. The sampling function is thus better written as

\[
f = f(x|\theta)
\]

a pdf for \( x \) values given \( \theta \). If your apriori knowledge about \( \theta \) is reflected by a pdf \( f(\theta) \), then the joint pdf of having parameters \( \theta \) and observed data \( x \) is

\[
f(x,\theta) = f(x|\theta) f(\theta)
\]
Actually we are interested in the \textit{a-posteriori} knowledge of $\theta$ given that we made the measurements $x$, which is the \textit{conditional} probability

$$
f (\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta) f(\theta)}{f(x)}
$$

This is a form of Bayes’s theorem. Naming conventions:

- $f(\theta)$ is the prior pdf
- $f(\theta|x)$ is the posterior pdf
- $f(x|\theta)$ is the sampling pdf, or likelihood pdf when considering its $\theta$ dependence
- $f(x) = \int d\theta \ f(x|\theta) f(\theta)$ is a normalization factor

\textbf{Caveats} on the Bayesian approach: conceptual subtlety to assigning probabilities to unknowns which are not repeatable experiments; some arbitrariness in the prior pdf.
Coming to our working example, the likelihood is

\[ f(\{x_i\} | \mu) = \prod_{i=1}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left( -\frac{(x_i - \mu)^2}{2\sigma_i^2} \right) = \frac{1}{(2\pi)^{n/2} \prod_{i=1}^{n} \sigma_i} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma_i^2} \right) \]

Now when we substitute into Bayes’s theorem, we will get a pdf for \(\mu\) that is Gaussian too. We can write this transparently by completing the square and writing

\[ \chi^2(\{x_i\}; \mu) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma_i^2} = \frac{[\mu - \mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} + \chi_0^2(\{x_i\}) \]

and solving for \(\mu_0(\{x_i\}), \sigma_\mu^2(\{x_i\}),\) and \(\chi_0^2(\{x_i\}),\) the names of which have been deliberately chosen.
Expanding both sides and equalizing, we find

\[
\frac{1}{\sigma^2_\mu(\{x_i\})} = \sum_{i=1}^{n} \frac{1}{\sigma^2_i}
\]

which is actually independent of the \( \{x_i\} \),

\[
\chi^2_0(\{x_i\}) = \sum_{i=1}^{n} \frac{x_i^2}{\sigma^2_i} - \left[ \frac{\mu_0(\{x_i\})}{\sigma^2_\mu(\{x_i\})} \right]^2
\]

which is irrelevant here, and

\[
\mu_0(\{x_i\}) = \sigma^2_\mu(\{x_i\}) \sum_{i=1}^{n} \frac{x_i}{\sigma^2_i} = \frac{\sum_{i=1}^{n} \sigma^{-2}_i x_i}{\sum_{i=1}^{n} \sigma^{-2}_i}
\]

which is a sort of weighted mean.
Then the likelihood function can be written as

\[ f(\{x_i\}|\mu) = \frac{e^{-\chi^2_0(\{x_i\})/2}}{(2\pi)^{n/2} \prod_{i=1}^{n} \sigma_i} \exp \left( - \frac{[\mu - \mu_0(\{x_i\})]^2}{2\sigma^2_\mu} \right) \]

and the posterior pdf reads

\[ f(\mu|\{x_i\}) \propto \exp \left( - \frac{[\mu - \mu_0(\{x_i\})]^2}{2\sigma^2_\mu} \right) f(\mu) \]

where the proportionality constant can be worked out from the normalization condition on the posterior pdf. As you can see the essential information about the outcome of the experiment is **encoded** in the weighted mean \( \mu_0(\{x_i\}) \).
Samples, statistics and estimators

**Frequentist point of view: optimal estimates**

In the frequentist approach, we are just limited to the use of the likelihood $f(x|\theta)$; this often reduces to try getting an estimate of our unknowns.

Coming to our working example, the best thing we can do is trying to estimate the value of $\mu$, and to consider what is the best estimator $\widehat{\mu}$ to use. For simplicity let’s restrict ourselves to linear combination of the random variables

$$\widehat{\mu} = \sum_i a_i X_i$$

We would like that the expectation value of the estimate $<\widehat{\mu}> = \sum_i a_i <X_i> = \mu \sum_i a_i$ coincides with the true value $\mu$ (unbiased estimate), so when $\sum_i a_i = 1$. 
The variance of the estimate is

$$\text{Var}(\mu) = \sum_i \sum_j a_i a_j \text{Cov}(X_i X_j) = \sum_i a_i^2 \sigma_i^2$$

The optimal estimate would be the unbiased estimate with the lowest variance. We can use the method of Lagrange multipliers, by minimizing $$\sum_i a_i^2 \sigma_i^2 + \lambda (\sum_i a_i - 1)$$ with respect to \{a_i\} and to \(\lambda\). This gives \(a_i = -\lambda/2 \sigma_i^2\) and \(\sum_i a_i = 1\), so yielding \(\lambda = -2/\sum_i \sigma_i^{-2}\). The optimal estimate is found to be

$$\mu = \sum_i \sigma_i^{-2} X_i / \sum_i \sigma_i^{-2}$$

with variance \(\text{Var}(\mu) = 1/ \sum_i \sigma_i^{-2}\). Note that this is precisely the weighted mean that showed up in the posterior pdf for the Bayesian approach, and its variance is the width of the posterior pdf in the case of a uniform prior.
Samples, statistics and estimators

Statistics and estimators

Given a random vector $\mathbf{X}$, a statistics $T(\mathbf{X})$ is any function of the random variables. The statistics $T(\mathbf{X})$ is itself a random variable, that in a particular realization where $\mathbf{X}=\mathbf{x}$ takes on values $T(\mathbf{x})$. Often a statistics is used to get an estimate of a parameter $\theta$; in that case we say that $T(\mathbf{X})$ is an estimator for $\theta$.

A statistics $T(\mathbf{X})$ is called an unbiased estimator for $\theta$ if

$$<T(\mathbf{X})> = \theta$$

otherwise $B=<T> - \theta$ is called the bias of $T$. For example, consider a random sample $\mathbf{X}$ drawn from any distribution with $<X_i> = \mu$. Then we know that the statistics $T(\mathbf{X}) = \mathbf{X} = \sum_i X_i / n$ satisfies $<\mathbf{X}> = \mu$. Thus $\mathbf{X}$ is an unbiased estimator of $\mu$. 
Consider now an exponential pdf with unknown parameter $\lambda$:

$$f(x; \lambda) = \lambda \exp(-\lambda x)$$

We know that $1/\lambda$ is the expectation value for an exponential pdf, so $X$ is an unbiased estimator if $1/\lambda$. However, one can show that $1/X$ is a biased estimator of $\lambda$, because in general

$$<1/X> \neq 1/<X>$$

E.g., for a random sample of size $n$ following an exponential pdf we have

$$<1/X> = n \lambda^n \int dx_1 \ldots \int dx_n \exp[-\lambda(x_1 + \ldots + x_n)] / (x_1 + \ldots + x_n)$$
Change to variables $y_i$

$$y_i = \lambda^n \sum_{j=1}^{i} x_j$$

so that the limits of the integrals become $0 < y_i < y_{i+1}$ for $i < n$ and $0 < y_n < \infty$

We get

$$<1/X> = n \lambda \int dy_n \exp[-y_n] / y_n^{n-2} \Gamma(n) = n \lambda \Gamma(n-1)/ \Gamma(n) = n \lambda/(n-1)$$

Of course, in this case we can just multiply by a constant to get an unbiased estimator for $\lambda$, i.e.

$$(1-1/n) \times (1/X)$$
On the other hand, it is instructive to look at the average squared error for an estimator $T$ of a parameter $\theta$:

$$<[T(X) - \theta]^2> = <[(T-<T>)+(<T>-\theta)]^2> = <(T-<T>)^2> + 2 (<T>-\theta)<(T-<T>)> + (<T>-\theta)^2 =$$

$$= \text{Var}[T] + B^2$$

This is the square of the bias plus its variance. So it is reasonable to look for an unbiased estimator with the smallest possible variance!

The square root of the variance for an estimator in known as its standard error

$$\sigma_T = \sqrt{\text{Var}[T]}$$
**Consistency**

A sample of statistics \( T_n \) is called a consistent estimator for \( \theta \) if and only if it converges in probability to it, i.e.,

\[
T_n \xrightarrow{} \theta
\]

For example, the quantity \( (n-1) \frac{S^2}{n} = \sum_i (X_i - X_n)^2 \) is a biased estimator of the variance since \( \langle (n-1) \frac{S^2}{n} \rangle = (n-1) \frac{\sigma^2}{n} \), but it is still consistent since \( (n-1) \frac{S^2}{n} \xrightarrow{} \sigma^2 \) as \( n \xrightarrow{} \infty \).

On the other hand the sequence \( T_n = X_n \) is obviously an unbiased estimator of the mean but does not converge to anything, so it is not a consistent estimator.
**Estimating with moments**

One method for estimating the parameters $\theta$ of the likelihood is to express them as a combination of its moments

$$<X^k> = \int dx f(x|\theta) x^k$$

E.g., for a Gaussian likelihood we have the two parameters $\mu$ and $\sigma$. The moments of the likelihood are $<x> = \mu$ and $<x^2> = \mu^2 + \sigma^2$; solving for $\mu$ and $\sigma$ we can thus define the method-of-moments estimates of $\mu$ and $\sigma$ as

$$\mu = <X> \quad \sigma^2 = (<X^2> - <X>)^2 = (n-1)/n * S^2$$

Note that $<\mu> = \mu$ while $<\sigma^2> = (n-1)/n * \sigma$, so $\mu$ is unbiased while $\sigma$ is biased.
Point estimation and Fisher information

**Maximum likelihood estimator**

One method for obtaining an estimator of the parameter $\theta$ is to pick up the value which **maximizes** the likelihood $f(x|\theta)$, where $x$ are the observed data. This seems reasonable because it gives the parameter value for which the actual data were most likely. However, there is still something a bit fishy, since the likelihood function is not a probability distribution for $\theta$, but rather for $x$ with a dependence from $\theta$. The mle is defined as

$$\frac{df(x|\theta)}{d\theta}|_{\theta=\hat{\theta}} = 0$$

The Bayesian approach tells why the mle is a good idea. Since $f(\theta|x) \propto f(x|\theta) f(\theta)$ from Bayes, we see that if the prior pdf $f(\theta)$ for $\theta$ is **uniform**, then the mle $\hat{\theta}$ also maximizes the posterior pdf $f(\theta|x)$. However, generally the posterior maximum will be affected by a change of variable, while the mle will not, since $f(x|\theta)$ is not a density in $\theta$. 
Point estimation and Fisher information

Since the max likelihood estimate $\theta$ depends on $x$, it can be thought as a random variable, hence as a statistics constructed from the random sample $X$. In this respect two considerations give support to the use of the maximum likelihood estimate:

1. Suppose $\theta_0$ is the true value of a parameter $\theta$, and $X$ is a random sample of size $n$, then the true parameter maximizes the likelihood in the limit $n \to \infty$, i.e.,

$$\lim_{n \to \infty} f(x|\theta_0) > f(x|\theta)$$

2. The maximum likelihood estimator converges in probability to the true parameter, i.e., it is a consistent estimator, i.e.,

$$\theta \xrightarrow{p} \theta_0$$

So essentially in the large $n$ limit, the true value becomes the mle and viceversa.
Log-likelihood

In practice, it is easier to compute the maximum likelihood estimate using the logarithm of the likelihood function and considering it as a function of $\theta$, i.e.,

$$L(\theta) = \ln f(x|\theta)$$

For example, consider the exponential distribution. The loglikelihood function is

$$L(\lambda) = n \ln(\lambda) - \lambda \sum x_i$$

Differentiating and setting this to zero gives the (biased) estimator

$$\lambda = n / \sum x_i = 1/X$$
**Point estimation and Fisher information**

**Method of least squares**

Suppose you have two random variables $X$ and $Y$ related by a known functional dependence $g$ such that

$$ Y = g(X, \theta) $$

where $\theta$ is a vector of parameters. Assume that the variables follow a normal distribution with $\sigma_X << \sigma_Y$. The method of least squares is based on the loglikelihood

$$ L(\theta) = \sum_{i=1}^{n} \frac{[y_i - g(x_i, \theta)]^2}{\sigma_i^2} + \text{cost.} $$

If the variance are equal, the maximum likelihood method tells us that it is enough to minimize the squares of residuals given by $\sum_{i=1}^{n} [y_i - g(x_i, \theta)]^2$.
As a relevant example, suppose to fit the data with a straight line. In this case \( Y = aX + b \). Find the parameters of the best fitting ("regression") line. Assuming the same variance \( \sigma^2 \) for the \( y \) values, one has

\[
\partial_a \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = 0 \quad \Rightarrow \quad a \sum_i x_i^2 + b \sum_i x_i - \sum_i x_i y_i = 0
\]

\[
\partial_b \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = 0 \quad \Rightarrow \quad a \sum_i x_i + nb - \sum_i y_i = 0
\]

whose solution is the following (where the average is defined as \( x = (1/n) \sum_i x_i \))

\[
a = \frac{\sum_i xy - \sum_i x y}{\sum_i x^2 - \sum_i x^2}
\]

\[
b = \frac{\sum_i y^2 - \sum_i x y}{\sum_i x^2 - \sum_i x^2}
\]

\[
\sigma_a^2 = \frac{\sigma^2}{\sum_i x^2 - \sum_i x^2}
\]

\[
\sigma_b^2 = \frac{\sigma^2}{\sum_i x^2 - \sum_i x^2}
\]
Point estimation and Fisher information

If the variances are not equal, we have to minimize \( \sum_{i=1}^{n} \frac{[y_i-g(x_i,\theta)]^2}{\sigma_i^2} \)

\[ \partial_a \sum_{i=1}^{n} \frac{[y_i - (ax_i+b)]^2}{\sigma_i^2} = 0 \]
\[ \partial_b \sum_{i=1}^{n} \frac{[y_i - (ax_i+b)]^2}{\sigma_i^2} = 0 \]

whose solution is the following

\[ a = \frac{[xy-xy]}{[x^2-x^2]} \]
\[ b = \frac{[y x^2 - x xy]}{[x^2 - x^2]} \]

where the averages are weighted with a weight \( \sigma_i^{-2} \) and the variances are

\[ \sigma_a^2 = \frac{1}{\sum_i \sigma_i^{-2}} \frac{[x^2 - x^2]}{[x^2 - x^2]} \]
\[ \sigma_b^2 = \frac{x^2}{\sum_i \sigma_i^{-2}} \frac{x^2 - x^2}{x^2 - x^2} \]
One can also try to fit the data with other shapes, e.g., a polynomial by assuming
\[ Y = a + b \, X + c \, X^2 + d \, X^3. \] One finds

\[
\sum_{i=1}^{n} [y_i - (a + bx_i + cx_i^2 + dx_i^3)]^2 \frac{x_i^k}{\sigma_i^2} = 0 \quad k=0,1,2,3
\]

whose solution is given by the system

\[
a + b \, x + c \, x^2 + d \, x^3 - y = 0
\]

\[
a x + b \, x^2 + c \, x^3 + d \, x^4 - y x = 0
\]

\[
a x^2 + b \, x^3 + c \, x^4 + d \, x^5 - y x^2 = 0
\]

\[
a x^3 + b \, x^4 + c \, x^5 + d \, x^6 - y x^3 = 0
\]
Other estimator algorithms modifying the least square method that are often used:

**Least absolute residuals (LAR)** → minimize the absolute difference of residuals, rather than the square difference, implying that extreme values have a lesser influence.

**Bisquare weights** → minimize a weighted sum of squares, where the weight to each point depends on how far the point is from the fitted line. Points farther from the line get reduced weight, while points that are farther from the line than would be expected by random chance get zero weight.

Typically, LAR is preferable for data with less outliers/anomalies, while the bisquare for data with outliers that you do not want to affect your fit because you think are due to disturbances.
Point estimation and Fisher information

- Least Square fit
- Least Absolute Residual fit
- Bisquare fit
Quality of an estimator: Cramer-Rao bound

The variance of an unbiased estimator $\text{Var}(Y) = \langle [Y - \theta]^2 \rangle$ can be interpreted as a loss function, measuring the risk of making a poor estimate. So choosing an estimator that minimizes the variance is highly desirable. An estimator $T(X)$ is called an unbiased minimum variance estimator (mve) of $\theta$ if it is unbiased, i.e. $\langle T(X) \rangle = \theta$, and if the variance of $T$ is less than or equal to the variance of every other unbiased estimator of $\theta$.

Looking for mve among all estimator is a complex issue. A first look at the problem comes by introducing the concept of Fisher information in terms of the log-likelihood $L(\theta) = \ln f(x|\theta)$ as

$$I(\theta) = \text{Var}[L'(\theta)]$$
First of all note that

\[ <L'(\theta)> = d_\theta \int dx \, f(x,\theta) = d_\theta 1 = 0 \]

Then the Fisher information is actually

\[ I(\theta) = <[L'(\theta)]^2> \]

The following identity holds

\[ 0 = d_\theta <L'(\theta)> = d_\theta \int dx \, f(x,\theta) \, \partial_{\theta} \ln f(x,\theta) = \int dx \, \partial_{\theta} f(x,\theta) \, \partial_{\theta} \ln f(x|\theta) + \int dx \, f(x|\theta) \, \partial_{\theta \theta}^2 \ln f(x|\theta) \]

so that

\[ I(\theta) = - <L''(\theta)> \]
Point estimation and Fisher information

For example, consider the normal distribution with mean $\theta$ and variance $\sigma^2$. Then

$$L(\theta) = -(x - \theta)^2 / 2\sigma^2 - \ln(2\pi\sigma^2)/2$$

so taking derivatives we have $L'(\theta) = (x - \theta) / \sigma^2$ from which we can easily cross-check that $<L'(\theta)> = 0$ and

$$L''(\theta) = -1 / \sigma^2$$

from which we compute the Fisher information

$$I(\theta) = 1 / \sigma^2$$
Point estimation and Fisher information

Interestingly, for a random sample of size \( n \) one has

\[
L_X(\theta) = \ln f(x|\theta) = \ln \prod_i f(x_i;\theta) = \sum_i \ln f(x_i;\theta)
\]

and so

\[
I_X(\theta) = -\langle L_X''(\theta) \rangle = -\sum_i \langle L''(\theta) \rangle = n I(\theta).
\]

Now for a general estimator \( T(X) \) of a parameter \( \theta \), the variance is given by

\[
\text{Var}(T|\theta) = \langle [T - \langle T \rangle(\theta)]^2 | \theta \rangle
\]

This has a similar structure to the Fisher information for the sample, since both are of the form

\[
\langle [a(X)]^2 | \theta \rangle = a \cdot a,
\]

where we have defined the dot product

\[
a \cdot b = \int dx \ f(x, \theta) \ a(x) \ b(x)
\]

From the Cauchy-Schwarz algebraic inequality \((a \cdot a) (b \cdot b) \geq | a \cdot b |^2\) one has
Point estimation and Fisher information

\[
\text{Var}(T|\theta) \cdot n \cdot I(\theta) \geq \left[ <T - <T>(\theta)] L'(\theta)|\theta> \right]^2 = \left[ <T L'(\theta)| \theta > - <T>(\theta) <L'(\theta)> \right]^2 = \\
= \left[ \int dx \ T(x) \partial_{\theta} f(x, \theta) \right]^2 = \left[ \partial_{\theta} \int dx \ T(x) f(x, \theta) \right]^2 = \left[ d_{\theta} <T> \right]^2 = \left[ <T(\theta)>' \right]^2
\]

So we have obtained the so-called Cramer-Rao bound

\[
\text{Var}(T|\theta) \geq \left[ <T (\theta)>' \right]^2 / n I(\theta)
\]

In the special case when \( T \) is an unbiased estimator of \( \theta \) then \( <T(\theta)> = \theta \) and

\[
\text{Var}(T|\theta) \geq 1 / n I(\theta)
\]

The variance of an estimator can be no less than the inverse of the Fisher information of the sample.
Point estimation and Fisher information

For a Gaussian likelihood with mean $\theta$ and variance $\sigma^2$, and an estimator like the sample mean $X$ the bound tells us that

$$\text{Var}(X|\theta) \geq \frac{1}{n} I(\theta) = \frac{\sigma^2}{n}$$

In this case we know that the bound is saturated, and the sample mean is the lowest variance unbiased estimator of the distribution mean $\theta$.

When the Cramer-Rao bound is saturated, the estimator $T$ is said to be efficient. Otherwise the ratio of the bound to the actual variance is called efficiency of the estimator. Another selling point of the maximum likelihood estimate is that it becomes efficient in the limit of infinite sample size. Also, the distribution of the mle converges to a Gaussian with zero mean and minimum variance specified by the bound:

$$\sqrt{n} \left[ \hat{\theta} - \theta \right] \rightarrow \text{N}(0, 1/I(\theta))$$
**Fisher information matrix**

We can now ask how the likelihood or the posterior pdf perform in the vicinity of the maximum likelihood estimate of the parameters. One trick would be to Taylor expand $f(x|\theta)$ near its maximum, but since $f(x|\theta) \geq 0$ this can cause problems if we extrapolate too much. So instead expand the log-likelihood

$$L(\theta) \sim L(\theta) + L'(\theta) (\theta - \theta) + \frac{1}{2} L''(\theta) (\theta - \theta)^2 + ...$$

Taking the expectation value, now we know that $<L'(\theta)> = 0$ and $<L''(\theta)> < 0$. If we truncate the expansion at the first nontrivial order we get

$$f(x|\theta) \sim f(x|\theta_0) \exp[- \frac{1}{2} <|L''(\theta)|> (\theta - \theta)^2]$$
Point estimation and Fisher information

which is a Gaussian with width \(|L''(\theta)|^{-1/2}\). The second derivative \(|L''(\theta)|\) is the one dimensional version of what is known as the Fisher information matrix. More rigorously, if one has multiple parameters \(\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}\) it is true that

\[
<\partial L(\theta)/\partial \theta_i> = \int dx \partial f(x|\theta) / \partial \theta_i = (\partial / \partial \theta_i) \int dx f(x|\theta) = 0
\]

Since \(\partial L(\theta)/\partial \theta_i\) is a random vector, we can construct the (symmetric) variance-covariance matrix, known as Fisher matrix, with elements

\[
F_{ij} = <\partial L(\theta)/\partial \theta_i \partial L(\theta)/\partial \theta_j>
\]

Now we can differentiate with respect to \(\partial \theta_j\) to get the identity

\[
0 = (\partial / \partial \theta_j) \int dx \partial f(x|\theta) / \partial \theta_i = \int dx \partial f(x|\theta)/\partial \theta_j \partial \ln f(x|\theta) / \partial \theta_i + \]

\[
+ \int dx f(x|\theta) \partial^2 \ln f(x|\theta) / \partial \theta_i \partial \theta_j = <\partial^2 L(\theta)/\partial \theta_i \partial \theta_j> + F_{ij}
\]
So an equivalent way of writing the Fisher information matrix is

\[ F_{ij} = - \langle \partial^2 L(\theta)/\partial \theta_i \partial \theta_j \rangle \]

The Fisher matrix gives an estimate of the uncertainties of the parameters. Recall indeed from the Cramer-Rao bound that the mle \( \hat{\theta} \) from a sample of size \( n \gg 1 \) tends to a normal distribution; in the multiparametric case this is a multivariate normal

\[ \sqrt{n} [\hat{\theta} - \theta] \rightarrow N[0, F(\theta)^{-1}] \]

The matrix inverse \( F(\theta)^{-1} \) of the Fisher matrix is the variance-covariance matrix of the limiting distribution. The multiparameter version of the Cramer-Rao bound is

\[ \text{Var}[T_i(X)] \geq (1/n) \left[ F(\theta)^{-1} \right]_{ii} \]
Thus the diagonal elements of the inverse Fisher matrix provide lower bounds on the variance of the unbiased estimators of the parameters. This is the justification for the common practice of quoting such quantity as one-sigma uncertainty on the parameter estimate.

Note that if the Fisher matrix has off-diagonal elements, then it is mandatory to take the diagonal elements of the inverse Fisher and not the inverse of the diagonal elements of the Fisher matrix, since typically

$$[\mathbf{F}^{-1}]_{ii} \geq 1/ F_{ii}$$

This is thus a stronger bound than one would get by assuming all of the “nuisance” parameters to be known and applying the usual Cramer-Rao bound to that of interest.
Interval estimation

Beyond finding the most likely parameter value and describing the shape of the likelihood around that value, an important task in parameter estimation is to provide an interval that we associate quantitatively with likely values of the parameter.

**Bayesian plausible intervals**

From the Bayesian point of view, where a particular sample instance $x$ results in a posterior pdf $f(\theta|x)$, one would have a straightforward definition for a plausible interval in which we think $\theta$ is to lie with some probability $1-\alpha$

$$\int_{\theta_l}^{\theta_u} d\theta \ f(\theta|x) = 1-\alpha$$

So this means the probability that the estimates are under the posterior pdf, between $\theta_l$ and $\theta_u$ is $1-\alpha$. 
This leaves the freedom to choose where the interval begins. Some convenient choices are:

- **Lower** limit $P(\theta_l < \theta) = 1 - \alpha$
- **Upper** limit $P(\theta < \theta_u) = 1 - \alpha$
- A symmetric two-sided interval $P(\theta > \theta_i) = P(\theta_u > \theta) = \alpha / 2$
- A interval **centered** on the mode $\theta$ of the posterior pdf $P(\theta - \Delta\theta/2 < \theta < \theta + \Delta\theta/2) = 1 - \alpha$
- The **narrowest** possible intervals, i.e., of all the intervals with $P(\theta_i < \theta < \theta_u) = 1 - \alpha$ pick up the one that minimizes $\theta_u - \theta_i$

Note that the prescription above is unique for a given posterior pdf, but recall that to obtain that pdf we need to use whatever information we have about the problem to construct the appropriate **prior** pdf $f(\theta)$.
Interval estimation

Frequentist confidence intervals

In the frequentist picture, we can only design a procedure to generate an interval such that if you collect many random datasets and make such an interval from each, some fraction of those intervals will contain the true parameter value: this is known as a (frequentist) confidence interval. It is defined as a pair of statistics $L=L(\mathbf{X})$ and $U=U(\mathbf{X})$ chosen so that the probability that the parameter $\theta$ lies between them is $1-\alpha$

$$P(L< \theta<U) = 1-\alpha$$

It is important to stress that the probability is referred to the randomness of the statistics $L$ and $U$, and not to $\theta$; from the frequentist perspective we cannot talk about probabilities for $\theta$; it has some specific value, even if it is unknown. Given a particular realization $x$ of the sample $\mathbf{X}$, we have a confidence intervals between $L(x)$ and $U(x)$. 
Interval estimation

One method to construct the confidence intervals is to choose a statistics $T(X;\theta)$ dubbed pivot variable, whose pdf is a known function of the parameters, and find an interval using

$$P(a < T(X;\theta) < b) = 1-\alpha$$

By solving algebraically the inequalities $a < T(X;\theta)$ and $T(X;\theta) < b$ for $\theta$, one should be able to write

$$P(L(X) < \theta < U(X)) = 1-\alpha$$

However, note that this construction is not unique; different choice of the pivot variable will give different confidence intervals.
Interval estimation

Now to illustrate the pivot variable method, consider the case where $\mathbf{X}$ is a sample of size $n$ drawn from a normal distribution with both mean $\mu$ and variance $\sigma$ unknown, where we want a confidence interval on $\mu$.

The pivot variable should depend on $\mu$ and $\mathbf{X}$ but not on $\sigma$, so taking

$$Z = \frac{(\mathbf{X} - \mu)}{(\sigma/\sqrt{n})}$$

will not work. Fortunately, we know from Student’s theorem that the quantity

$$T = \frac{(\mathbf{X} - \mu)}{\sqrt{(S^2/n)}}$$

obeys a t-distribution with $n-1$ degrees of freedom, i.e., $f_T(t; n-1)$. 
This works as a pivot variable, since $S^2 = \sum_i (X_i - \bar{X})^2 / n$ depends only on the sample, and requires no knowledge of $\mu$ or $\sigma$.

In general, the $(1-\alpha)$ 100th percentile of the distribution $t_{\alpha,n}$ is easy to compute

$$1-\alpha = P(T < t_{\alpha,n}) = \int_{-\infty}^{t_{\alpha,n}} dt f_T(t; n)$$

Since we want a two-side confidence interval, actually we need $t_{\alpha/2,n}$ and $t_{1-\alpha/2,n}$, but because of the symmetry in the t-distribution we can use that

$$t_{1-\alpha/2,n} = - t_{\alpha/2,n}$$
Interval estimation

Turning back to our pivot variable $T$, which is $t$-distributed with $n-1$ degrees of freedom, we find

$$1 - \alpha = P(-t_{\alpha/2, n-1} < T < t_{\alpha/2, n-1}) = P(-t_{\alpha/2, n-1} < (\bar{X} - \mu) / \sqrt{S^2/n} < t_{\alpha/2, n-1})$$

which, solving for $\mu$ the two inequalities gives the confidence intervals

$$P[\bar{X} - t_{\alpha/2, n-1} \sqrt{S^2/n} < \mu < \bar{X} + t_{\alpha/2, n-1} \sqrt{S^2/n}] = 1 - \alpha$$

Thus typically the Student distribution is exploited to give confidence intervals to Gaussian estimates, especially in the cases of a sample with few data points.
Resampling

The best way to establish the properties of a population is to examine each and every member, make the appropriate measurements, and record their values.

The alternative is to select many samples, and to estimate their statistical properties. Plainly, by taking different samples we might record quite different values. Finding out how different, it is the best way to determine how precise our estimates are.

There are two practical procedures to do this known as bootstrap and jacknife; these can be used not only to get an estimate of a parameter but also to estimate its variance when the latter is difficult to get (e.g., correlated data), and check for biases.
Bootstrap

The bootstrap treats the original sample of values as a stand-in for the population and resamples from it repeatedly, with replacement, computing the desired estimate each time:

• observe a sample $X = x = \{x_1 \ldots x_n\}$ and compute the estimate $\theta(X)$
• for $i=1\ldots s$ where $s$ is the number of bootstrap samples to be generated:
  o generate a bootstrap sample $X^i = x^i = \{x^i_1 \ldots x^i_n\}$ by sampling with replacement from the observed dataset
  o compute $\theta^i = \theta(X^i)$ in the same way that you calculated the original estimate
• compute the bootstrap mean and variance

$$\bar{\theta}^* = (1/s) \sum_{i=1\ldots s} \theta^i$$

$$S^* = \frac{1}{(s-1)} \sum_{i=1\ldots s} (\theta^i - \bar{\theta}^*)^2$$
Suppose that the measured data come from a cumulative distribution function $F(x)$. In an ideal world, we would just get more samples from $F$ and then evaluate the variance of our estimate with these new samples. In reality we do not know this and so we cannot. What bootstrap does is to replace $F$ with an empirical distribution which is an estimate of $F$. The number of bootstrap sample $s$ needs to be large but since the error in evaluating the variance decreases like $1/\sqrt{s}$ there is no point in making $s$ too big. A good technique is to plot the estimate of variance against $s$ to see whether it has settled down to some value.
Resampling

*Example:* Bhavsar (1990) estimated via bootstrap the uncertainty in measuring the slope of the galaxy angular two-point correlation function $w(\theta)$ for radio galaxies.

*Left:* a least square fit to the slope gives $-0.19$. *Right:* the distribution of slopes obtained by bootstrapping the sample with 1000 trials; the slope is less than zero (signal is present) for 96.8 percent of the trials.
Resampling

**Jacknife**

The jacknife is a more orderly version of the bootstrap.

- observe a sample \( X = x = \{x_1 \ldots x_n\} \) and compute the estimate \( \theta(X) \)
- for \( i = 1 \ldots n \):
  - generate a jacknife sample \( X^{-i} = x^{-i} = \{x_1 \ldots x_{i-1} \ x_{i+1} \ldots x_n\} \) by leaving out the i-th observation
  - compute \( \theta^{-i} = \theta(X^{-i}) \) in the same way that you calculated the original estimate
- compute the jacknifed mean and variance

\[
\bar{\theta}^* = \frac{1}{n} \sum_{i=1}^{n} \theta^{-i} \quad \quad \quad \quad S^* = \frac{(n-1)}{n} \sum_{i=1}^{n} (\theta^{-i} - \bar{\theta}^*)^2
\]
The jackknife is a good method to compute not only the variance of an estimate, but also its bias. The jackknife estimate of the bias is given by \( B = (n-1) \left( \bar{\theta}^* - \bar{\theta} \right) \) and the bias corrected jackknife estimate reads

\[
\theta^+ = \bar{\theta} - B = n \bar{\theta} - (n-1) \theta^*
\]

This comes about because, for many statistics, \( B(\theta) \sim a/n + b/n^2 \). Then \( B(\bar{\theta}^*) \sim a/(n-1) + b/(n-1)^2 \) and likewise for \( B(\theta^*) \). Thus

\[
<B> = (n-1) \langle \bar{\theta}^* - \bar{\theta} \rangle \sim (n-1) \left[ a[1/(n-1)-1/n] + b \left[ 1/(n-1)^2-1/n^2 \right] \right] = a/n + b(2n-1)/n^2 (n-1) \sim B(\theta)
\]

so the jackknife bias estimates the true bias up to order \( n^{-2} \). This implies:

\[
B(\theta^+) = B(\theta) - <B> \sim a/n + b/n^2 - a/n - b(2n-1)/n^2 (n-1) \sim -b/n(n-1)
\]
Example: Draw 10 samples from a normal distribution and calculate variance. Do this for 1000 trials, and the result is the blue histogram on the top panel. The mean variance is 0.89999 less than the true 1.0, because of the infamous $1/n$ vs $1/(n-1)$ issue...

Now calculate variances for each of 1000 jackknife trials, resulting in the red histogram of the bottom panel. The peak has shifted to larger values; in fact, the mean variance is 0.9834, much closer to the true value 1.0.

Bias has been removed (reduced)!
Hypothesis testing

So far we have considered how to get handle of parameters in a probability distribution given that we had a sample drawn from that, and found a particular realization. Now we wish to consider how to use the realization of the sample to distinguish between two competing hypotheses about what the underlying distribution is. For the sake of simplicity we will assume there is only one family $f(x, \theta)$ parameterized by $\theta$ which lies somewhere in a region $\Omega$ and then take the hypotheses to be

$$H_0 \rightarrow \text{the distribution is } f(x, \theta) \text{ where } \theta \in \omega_0$$
$$H_a \rightarrow \text{the distribution is } f(x, \theta) \text{ where } \theta \in \omega_a$$

Typically $H_0$ represents the absence of the effect we are looking for, and is known as null hypothesis, while $H_a$ represents the presence of the effect, and is known as alternative hypothesis.
Hypothesis testing

A hypothesis test is simply a rule for choosing between the two hypotheses depending on the realization of the sample $X = x$. Stated more generally we construct a critical region $C$ which is a subset of the n-dimensional sample space $\Omega$. If $X \in C$ we reject the null hypothesis and favor $H_a$; if $X \in \neg C$, we accept the null hypothesis and favor $H_0$.

Of course, since $X$ is random, there will be some probability $P(X \in C ; \theta)$ that we will reject the null hypothesis, which depends on the value of $\theta$. If the test were perfect, that probability would be 0 if $H_0$ were true (i.e., for any $\theta \in \omega_0$) and 1 if $H_a$ were true (i.e., for any $\theta \in \omega_a$) but we would not be doing statistics. So in practice there is some chance we will choose the wrong hypothesis, i.e., some probability that, given a value of $\theta \in \omega_0$ associated to $H_0$, the realization of our data will cause us to reject $H_0$, and some probability that, given a value of $\theta \in \omega_a$ associated to $H_a$, the realization of our data will cause us to accept $H_0$. 
Hypothesis testing

We say that:

If $H_0$ is true and we reject $H_0$ $\rightarrow$ Type I error or false positive

If $H_a$ is true and we reject $H_0$ $\rightarrow$ true positive

If $H_0$ is true and we accept $H_0$ $\rightarrow$ true negative

If $H_a$ is true and we accept $H_0$ $\rightarrow$ Type II error or false negative

Typically, a false positive is considered worse than a false negative, so we decide how high a false positive probability we can live with, and then try to find the test which gives the lowest false negative probability.
Hypothesis testing

Given a critical region $C$, we would like to talk about the associated false positive probability $\alpha$ and the associated false negative probability $1-\gamma$ (while $\gamma$ is called power of the test), but we have to be careful since $H_0$ and $H_1$ are in general composite hypotheses. This means each of them does not correspond to a single parameter value $\theta$ and thus to a single distribution, but rather to a range of values $\theta \in \omega_0$ or $\theta \in \omega_a$. So both $\alpha$ and $\gamma$ may depend on the value of $\theta$. We take the false alarm probability $\alpha$ to be the worst case scenario within the null hypothesis

$$\alpha = \max_{\theta \in \omega_0} P(X \in C ; \theta)$$

This is called the size of the critical region $C$, and sometimes also significance of the test, which is a bit counterintuitive; a low value of $\alpha$ means the probability of a false positive is low, which means a positive result is more significant than if $\alpha$ were higher. It is the probability that we will falsely reject the null hypothesis $H_0$, …
Hypothesis testing

...maximized over any parameters within the range associated to $H_0$. On the other hand, since the alternative hypothesis almost always has a definite parameter $\theta$ associate with it, we define the probability of correctly rejecting the null hypothesis as

$$\gamma_C = P(X \in C ; \theta) \quad \text{for} \quad \theta \in \omega_a$$

We explicitly write this as a function of the critical region $C$, since we might want to compare different tests with the same false alarm probability $\alpha$ (critical regions with the same size $\alpha$) to see which one is more powerful.

As an example, let $X$ be a random sample of size $n$ from a normal distribution, where the null hypothesis $H_0$ is $\mu=0$, and the alternative hypothesis $H_a$ is $\mu>0$. For simplicity, let us assume that the variance $\sigma^2$ is known; alternatively, when the sample is large we can use $S^2$ as an estimate.
Hypothesis testing

If \( H_a \) were true, the test statistics \( Z = \frac{X}{\sigma/\sqrt{n}} \) would be distributed like a Gaussian with unit variance and \emph{positive} mean. Thus large measured values of \( Z \) will indicate that the null hypothesis can be rejected. In this vein, we can define a critical region

\[
C \rightarrow Z = \frac{X}{\sigma/\sqrt{n}} > z_\alpha
\]

that will correspond to a test with false alarm rate \( \alpha \), set by the condition

\[
P(\text{reject } H_0 \rightarrow X \in C | H_0 \text{ is true } \rightarrow \mu=0) = P[\frac{X}{\sigma/\sqrt{n}} > z_\alpha] = \alpha
\]

The \textbf{power} of the test for a given true value of \( \mu \) is easy computable in terms of the error function \( \Phi \) :

\[
\gamma_C (\mu) = P[\frac{X}{\sigma/\sqrt{n}} > z_\alpha] = P[\frac{(X-\mu)}{(\sigma/\sqrt{n})} > \frac{z_\alpha - \mu}{(\sigma/\sqrt{n})}] = 1 - \Phi[\frac{z_\alpha - \mu}{(\sigma/\sqrt{n})}] = \Phi[\mu/(\sigma/\sqrt{n}) - z_\alpha]
\]
Hypothesis testing

For definiteness, let’s choose a significance $\alpha = 0.05$; then using the Gaussian distribution

$$P[Z = \frac{X}{\sigma/\sqrt{n}} > z_\alpha] = \alpha$$

will imply $z_\alpha = 1.645$. And the power of the test is $\gamma(\mu) = \Phi[\mu/(\sigma/\sqrt{n}) - 1.645]$.

The power of the test as a function of $\mu$ is illustrated on the right. The power goes from the value of the significance $\alpha$ when $\mu = 0$, and then increases with $\gamma(\mu) > \alpha$ when $\mu > 0$. This is a very desiderable property, since the test should be more likely to reject $H_0$ when it is false than when it is true! A test with this property is called unbiased.
Hypothesis testing

*p-values*

In this example, given a data realization $x$, and specifically a sample mean $\bar{x}$, we will reject the null hypothesis if $\bar{x} > z_\alpha \sigma / \sqrt{n}$. This means there will be some values of the false alarm probability $\alpha$ for which we reject $H_0$ and some for which we do not. One convenient way to proceed is quoting the most stringent significance level $\alpha$ for which $H_0$ would be rejected. In other words, we ask, given a measurement (in this case $x$) how likely is that we would find a measurement at least as extreme, just by accident, if the null hypothesis were true.

$$p = P[X > x; \mu = 0] = 1 - \Phi[\frac{x}{\sigma / \sqrt{n}}] = \Phi[-\frac{x}{\sigma / \sqrt{n}}]$$

This is known as *p-value*. A lower p-value means that the results were less likely to have occurred by chance in the absence of a real effect. Typically $p<0.05$ is good!
Hypothesis testing

*Hypothesis testing: Pearson’s Chi-squared tests*

Both the false alarm rate and the p-values are ways of talking about how inconsistent the data are with the null hypothesis. Another similar test is the goodness-of-fit or Pearson’s chi-squared test: we measure how closely the actually collected data come to the most likely values, and how likely we were to deviate by that much or more, if the model were actually correct.

As an example, consider the case where the null hypothesis tells us that our random data vector $\mathbf{X}$ is made up of $n$ independent random variables with means $\mu_i$ and variances $\sigma_i$. Then we know that the test statistics

$$Z_i = (X_i - \mu_i)/\sigma_i$$

is a standard normal.
Hypothesis testing

Thus

\[ Y = \sum_i Z_i^2 = \sum_i \left( \frac{X_i - \mu_i}{\sigma_i} \right)^2 \]

is a \( \chi^2(n) \) random variable. If the measured data gives a \( Y = y \) much larger than expected for a \( \chi^2(n) \) distribution, this casts doubts on the null hypothesis. In other words, we construct a test with false alarm rate probability \( \alpha \) by the condition

\[ P(Y > \chi^2_{n,\alpha}) = \alpha \]

where \( \chi^2_{n,\alpha} \) is the 100 \( (1 - \alpha) \)th percentile of the \( \chi^2(n) \) distribution.

Other tests based on this very same approach are the likelihood ratio, Wald, and Rao score tests. We will now analyse them in some detail.
Hypothesis testing

More in general, consider a likelihood \( f(x|\theta) \). A test often used is the so called likelihood ratio test, based on the quantity \( \Lambda(x) = f(x|\theta) / f(x|\theta_0) \) with \( \theta_0 \) is the mle of the parameter. It can be demonstrated that asymptotically it converges in distribution to

\[
\chi^2_L = -2 \ln \Lambda(x) \to \chi^2(1)
\]

For finite \( n \) we assume this is approximately true and compare \( \chi^2_L \) to the 100(1-\( \alpha \))th percentile of the \( \chi^2(1) \) distribution to get a test with false alarm probability \( \alpha \).

E.g., consider a sample of size \( n \) drawn from a Gaussian with mean \( \theta \) and variance \( \sigma^2 \), for which the mle is the sample mean \( \bar{x} \). Then the likelihood ratio is \( \Lambda(x) = \exp[-n (\theta-\bar{x})^2 / 2\sigma^2] \). Since the sample mean \( \bar{X} \) is distributed as a Gaussian with mean \( \theta \) and variance \( \sigma^2/n \) we have

\[
-2 \ln \Lambda(x) = (\bar{X} - \theta)^2 / (\sigma^2 / n) \sim \chi^2(1)
\]
Hypothesis testing

Another possibility is to consider the result that the quantity $\sqrt{n} [\theta(X) - \theta]$ converges in distribution to a Gaussian with mean zero and variance $1/I(\theta)$, and construct the statistics

$$\chi^2_W = n I(\theta) [\theta(X) - \theta]^2$$

Comparing this to the percentiles of $\chi^2(1)$ is known as **Wald test**.

E.g., for a Gaussian we have $I(\theta) = 1/\sigma^2$ and Wald test statistics

$$\chi^2_W = (n/ \sigma^2) [X - \theta]^2 \sim \chi^2_L \sim \chi^2(1)$$

being of the same form of the likelihood ratio test.
Hypothesis testing

Another possibility is to consider the log-likelihood $L(\theta) = \ln f(x|\theta)$. We know that $<L'(\theta)> = 0$ and that $\text{Var}[L'(\theta)] = n I(\theta)$. So we can construct a test statistics which, in the limit that the sum of scores is normally distributed, becomes a $\chi^2(1)$, i.e.,

$$\chi^2_R = [L'(\theta)]^2/n I(\theta)$$

Comparing this to the percentiles of $\chi^2(1)$ is known as Rao scores test.

E.g., for a Gaussian we have $L'(\theta) = n (\bar{x} - \theta)/\sigma^2$ and Rao scores test statistics

$$\chi^2_R = (n/\sigma^2) [\bar{x} - \theta]^2 \sim \chi^2_L \sim \chi^2_w \sim \chi^2(1)$$

which is again of the same form of the preceding tests, and exactly $\chi^2(1)$ distributed.
Hypothesis testing

**Most powerful test: Neyman-Pearson lemma**

The most powerful test between two one-point hypotheses $H_0$ and $H_1$ can be constructed from the likelihood ratio $\Lambda(x) = f(x|H_0) / f(x|H_1)$ in the following way. We define the critical region $C$ so that $x \in C$ if and only if $\Lambda(x) < k$ where $k$ is defined by

$$P[\Lambda(x) < k|H_0] = \int_C dx \ f(x|H_0) = \alpha$$

which ensures that the critical region is of size given by the significance $\alpha$. Now we prove that the power of this test

$$\gamma_C = P[\Lambda(x) < k|H_1] = \int_C dx \ f(x|H_1)$$

is greater than for any other test with the same significance.
Hypothesis testing

**Proof**

If A is some other critical region of size $\alpha$ than we have to prove that $\gamma_C > \gamma_A$. To this purpose split A and C in terms of their overlap region $C \cap A$ as

$$C = (C \cap \neg A) \cup (C \cap A) \quad \quad \quad A = (\neg C \cap A) \cup (C \cap A)$$

The contribution to $C \cap A$ cancels out of any comparison, so we have just to prove that $\gamma_{C \cap \neg A} > \gamma_{\neg C \cap A}$. Now by definition $\Lambda(x) < k$ in C and $\Lambda(x) > k$ in $\neg C$ implying that

$$\gamma_{C \cap \neg A} = \int_{C \cap \neg A} dx \ f(x|H_1) / k > \int_{C \cap \neg A} dx \ f(x|H_0) / k$$
$$\gamma_{\neg C \cap A} = \int_{\neg C \cap A} dx \ f(x|H_1) / k < \int_{\neg C \cap A} dx \ f(x|H_0) / k$$

Now $\alpha = \int_C dx \ f(x|H_0) = \int_{C \cap \neg A} + \int_{C \cap A}$ and $\alpha = \int_A dx \ f(x|H_0) = \int_{\neg C \cap A} + \int_{C \cap A}$ so that

$$\int_{C \cap \neg A} dx \ f(x|H_0) = \int_{\neg C \cap A} dx \ f(x|H_0)$$ implying $\gamma_{C \cap \neg A} > \gamma_{\neg C \cap A}$, hence proving $\gamma_C > \gamma_A$. 
Suppose now that $H_0$ is a one-point hypothesis while $H_1$ is a composite hypothesis which allows for a range of values of the model parameters $\theta$ but comes with a prior $f(\theta|H_1)$. If we define a test which rejects $H_0$ when

$$\Lambda(x) = \frac{f(x|H_0)}{f(x|H_1)} = \frac{f(x|H_0)}{\int d\theta f(x|\theta) f(\theta|H_1)} < k$$

the Neyman-Pearson lemma tells us that this is the most powerful test of $H_0$ over $H_1$. This is most powerful in the sense of maximizing the power function

$$\gamma_C(H_1) = P[\Lambda(x) < k|H_1] = \int_C dx f(x|H_1) = \int_C dx \int d\theta f(x|\theta) f(\theta|H_1) = \int d\theta \gamma(\theta) f(\theta|H_1)$$

We note that the above is precisely the test statistics based on posteriors from Bayes

$$[f(x|H_0)/f(x|H_1)] * [P(H_0)/P(H_1)] * [f(x)/f(x)] = P(H_0|x)/P(H_1|x)$$
Hypothesis testing

**Reduced Chi-square**

Consider now $H_0$ to be a composite hypothesis corresponding to a family of models, such that the $X_i$ are to follow normal distributions with means $\mu_i(\theta)$ and variances $\sigma_i(\theta)$ depending on a m-dimensional vector of parameters $\theta$. Then the $\chi^2$-statistics corresponding to a particular set of parameter values is

$$Y(\theta) = q(X, \theta) = \sum_i [X_i - \mu_i(\theta)]^2 / \sigma_i^2(\theta)$$

If $H_0$ is true and $\theta$ are the actual parameter values, $Y(\theta)$ is $\chi^2(n)$ distributed. But suppose that we do not know the parameter values, and we have collected a data vector $X=x$. We can choose as our best estimate of the parameters the values $\hat{\theta}$

$$\partial \theta_j Y(\theta) = 0 \quad \text{for} \quad j=1,2,\ldots,m$$
Hypothesis testing

If you put these values back into $Y$, we get the minimized chi-squared values

$$\bar{Y} = Y[\theta(X)]$$

It turns out that, under many circumstances, this statistics obeys a $\chi^2(n-m)$ distribution, which is called reduced chi-squared. If a model predicts that $X$ are to follow a Gaussian distribution, we can collect a realization of these data $x=x$ and construct a test with significance $\alpha$ from

$$P(\bar{Y} > \chi^2_{n-m, \alpha}) = \alpha$$

As an example of reduced $\chi^2$, suppose $H_0$ says that, for a parameter $\theta$, the data are a normal random sample with each mean $\mu_i = \theta$ and each variance $\sigma_i = \sigma$. Then the chi-square statistics is $Y = \sum_i (X_i - \theta)^2 / \sigma^2$. The minimum chi-squared is
Hypothesis testing

\[ 0 = \frac{2}{\sigma^2} \sum_i (-x_i + \theta) = \frac{2}{\sigma^2} (n\theta - \sum_i x_i) \]

Thus the corresponding statistics is

\[ \theta (X) = \overline{X} \]

i.e., the sample mean. Note that \( \theta \) is also the maximum likelihood estimate, since the likelihood function \( L(\theta) \propto \exp(-Y) \). Substituting \( \theta \) into the chi-square statistic, we get the minimized chi-square

\[ Y = Y(\theta = \overline{X}) = \sum_i (X_i - \overline{X})^2 / \sigma^2 = (n-1) \frac{S^2}{\sigma^2} \]

where \( S^2 \) is the sample variance. But from Student’s theorem this follows a \( \chi^2(n-1) \) distribution; so we have confirmed that the chi-squared statistics minimized over m=1 parameter follows a \( \chi^2(n-m) \) distribution.
Bayesian hypothesis testing

In the Bayesian framework, the joint pdf for a sample $X$ drawn from a distribution with parameter $\theta$ can be written $f(x, \theta)$, and we can talk about the posterior probability $P(H_1|x)$ that a hypothesis is true, given an observed sample $x$. There are few caveats, however.

Typically, our hypothesis allows to describe the joint pdf for a random sample $X$ collected in the presence of that hypothesis $f(x|H_1)$, and then we can use Bayes’s theorem to construct the desired probability

$$P(H_1|x) = \frac{f(x|H_1) P(H_1)}{f(x)}$$

But this expression depends on $P(H_1)$, which is the prior probability that the…
Hypothesis testing

...hypothesis $H_1$ is true, which we might have difficulty stating. Another issue in the Bayesian framework is that the joint pdf $f(x|H_1)$ and $f(x|H_0)$ require to specify the hypothesis a little more precisely than we have done so far (a range like $\theta \in \omega_1$ is not enough). In particular, if one or both are composite hypothesis, we need to state the pdf associated with the hypothesis for $\theta$. For instance, marginalizing over $\theta$ yields

$$f(x|H_1) = \int_{\omega_1} d\theta \ f(x|\theta) \ f(\theta|H_1)$$

Finally, the denominator $f(x)$ is the overall pdf for the sample, marginalized over a complete set of mutually exclusive models, i.e.,

$$f(x) = f(x|H_0) \ P(H_0) + f(x|H_1) \ P(H_1) + f(x|H_2) \ P(H_2) + \ldots$$

which we are even less likely to have an handle on.
Hypothesis testing

**Odds ratio, Bayes factor, and Occam principle**

The usual way to circumvent these issues is to calculate not $P(H_1|x)$ but the odds ratio $O_{10}$, the ratio of the posterior probabilities of the competing models $H_1$ and $H_0$

$$O_{10} = P(H_1|x)/P(H_0|x) = \frac{f(x|H_1) P(H_1)}{f(x|H_0) P(H_0)} = B_{10} P(H_1)/P(H_0)$$

where the ratio of the “evidences” $B_{10} = f(x|H_1)/f(x|H_0)$ is known as **Bayes factor**. The Bayes factor allows to overcome one of the problems about using a frequentist chi-square test to assess the validity of a model: adding more parameters to the model make the fit better, but clearly an overtuned model is scientifically unsatisfying. Suppose model $H_1$ has one parameter and model $H_0$ has none. Then

$$B_{10} = \int d\theta f(x|\theta,H_1) f(\theta|H_1) / f(x|H_0)$$
Let’s assume that the likelihood can be approximated as a Gaussian near the mle:

\[ f(x|\theta, H_1) \sim f(x|\theta, H_1) \exp\left(-\frac{(\theta - \theta)^2}{2\sigma^2}\right) \]

We will also assume that this is sharply peaked compared to the prior \( f(\theta|H_1) \) and therefore we can replace \( \theta \) with \( \theta \) in the argument of the prior:

\[
\int d\theta \ f(x|\theta, H_1) \ f(\theta|H_1) \sim f(x|\theta, H_1) \ f(\theta|H_1) \ \int d\theta \ \exp\left(-\frac{(\theta - \theta)^2}{2\sigma^2}\right) = f(x|\theta, H_1) \ f(\theta|H_1) \ (2\pi\sigma^2)^{\frac{1}{2}}
\]

Then the Bayes factor can be approximated as

\[
B_{10} = \left[ \frac{f(x|\theta, H_1)}{f(x|H_0)} \right] \times (2\pi\sigma^2)^{\frac{1}{2}} f(\theta|H_1)
\]

The first factor is the ratio of the likelihood between the best-fit version of model with parameters and that with no parameters.
Hypothesis testing

That’s the end of the story in the frequentist approach, and we can see that if $H_0$ is included as a special case of $H_1$ then the ratio will be always greater or equal than 1, i.e., the tunable model will be able to find a higher likelihood than the model with no parameter. But in the Bayesian version, there is also a second factor, called the “Occam factor”, because it implements the Occam razor, the principle that, all else being equal, the simpler explanation must be favoured over more complicated ones. In fact, because the prior $f(\theta|H_i)$ is normalized, then $[f(\theta|H_i)]^{-1}$ is a measure of the width of the prior, i.e., how much parameter space the tunable model is available to it. For example, if the prior is uniform the Occam factor becomes $(2\pi\sigma_\theta^2)^{1/2} / (\theta_{\text{max}} - \theta_{\text{min}})$

Because we have assumed that the likelihood function was narrowly peaked compared to the prior, the Occam factor is always less than one, and the tunable model must have a large enough increase in likelihood over the simpler model in order to overcome this effect.
Hypothesis testing

*Example:* modelling of radio-source spectra. Left: data, corresponding to a powerlaw $S(f) = k f^{-1}$. Right: same data but with an offset error of $b = 0.4$ units as well as random noise $\varepsilon = 10\%$ Gaussian added.

First take the original data (left hand side) and test the model $S(f) = k f^{-\gamma}$. Each term in the likelihood product is proportional to

$$
[2\pi(\varepsilon k f_i^{-\gamma})^2]^{1/2} \exp \left[-\frac{(S_i - k f_i^{-\gamma})^2}{2(\varepsilon k f_i^{-\gamma})^2}\right]
$$
Hypothesis testing

The log-likelihood is shown in the top left panel.

The Gaussian approximation to likelihood (via Fisher matrix expansion) is in the bottom-left panel.

Marginalized posteriors are shown in the right panels (one parameter regardless of the other).
Hypothesis testing

Now take the data with offset and random noise added (right-hand) and test the original model without offset and another possible model $S(f) = b + k f^{-\gamma}$ with offset $b$. Each likelihood term is

$$[2\pi (\varepsilon k f_i^{-\gamma})^2]^{1/2} \exp [- (S_i - b - k f_i^{-\gamma})^2 / 2 (\varepsilon k f_i^{-\gamma})^2]$$

and we pose a Gaussian prior on $b$ of mean 0.4 (we are well informed! 😊). Then marginalize $b$ out to find the likelihood reported in figure (solid: model without offset; dashed: model with offset). The model with offset does a better job in recovering the true parameter.

But now suppose we are not informed... is a model with offset (1) a good a priori bet over one without (0)?

To answer, compute the odds ratio [assume $P(1)=P(0)$, i.e., agnostic state] by integrating likelihood over parameters:

$$O_{10} = B_{10} P(1)/ P(0) = B_{10} = 8 \rightarrow \text{definitely a good bet!}$$
Hypothesis testing

Note that up to now we have adopted a Gaussian prior on the offset $b$, with known mean 0.4 and stddev $\varepsilon$, i.e.,

$$[2\pi \varepsilon^2]^{1/2} \exp\left[-(b-0.4)^2/2\varepsilon^2\right]$$

...a very strong assumption! Suppose we know the stddev $\varepsilon$ but not the mean $\langle b \rangle$, and take the prior on $\langle b \rangle$ as uniform. This is called an hyperparameter, and is useful to get a posterior distribution for the interesting parameters (in our case $k$ and $\gamma$) which includes a range of models. Recomputing the log-likelihood and marginalizing over both $b$ and $\langle b \rangle$ we get the solid contours on the right (dashed line is model with known $\langle b \rangle$): the figure shows there is a tendency for estimating flatter powerlaws if we do not know much about $\langle b \rangle$. 
Finally, we can allow a separate offset $b_i$ at each frequency, so that each term in the likelihood product takes the form

$$
\exp \left[ -\frac{(b_i - <b>)^2}{2 \varepsilon^2} \right] \cdot \left[ 2\pi (\varepsilon k f_i \gamma)^2 \right]^{1/2} \exp \left[ -\frac{(S_i - b_i - k f_i \gamma)^2}{2 (\varepsilon k f_i \gamma)^2} \right]
$$

and assume again the weak uniform prior on $<b>$.

The log-likelihood contours (dashed: model above; solid: original model without offset) tell that by allowing a range of models via hyperparameters the solution has moved away from the well-defined (but wrong) parameter of the non-offset model… but this is at the cost on much wider error bounds on the parameters.
Kolmogorov-Smirnov nonparametric testing

Non parametric test are used to compare a sample with a reference probability distribution (or two samples), without assumption about the true distribution. Let $F(x) = P(X_1 \leq x)$ a cdf of a true underlying distribution of data. We define an empirical cdf by

$$F_n(x) = \frac{1}{n} \sum I(x_i \leq x)$$

where $I$ is an histogram that counts the proportion of the data below level $x$. The law of large numbers implies

$$F_n(x) \rightarrow \langle I (X_1 \leq x) \rangle = P(X_1 \leq x) = F(x)$$

It is easily recognized that this approximation holds uniformly for all $x$, so that

$$\max_x |F_n(x) - F(x)| \rightarrow 0$$
The crucial observation in the KS test is that the distribution of

$$\max_x |F_n(x) - F(x)|$$

does not depend on F. For given x, the central limit theorem implies that

$$\sqrt{n} |F_n(x) - F(x)| \Rightarrow N(0, F(x) [1-F(x)])$$

since $F(x)[1-F(x)]$ is the variance of $I[X_1 \leq x]$. Now it turns out that (a statement difficult to be proved)

$$P[\sqrt{n} \max_x |F_n(x) - F(x)| > t] \Rightarrow H(t) = 1 - 2 \sum_{i=1...\infty} (-1)^{i-1} e^{-2i^2 t}$$

where $H(t)$ is the cdf of the Kolmogorov-Smirnov distribution.
Kolmogorov-Smirnov nonparametric testing

Let us reformulate the hypothesis in terms of cdf

$$H_0 : F = F_0 \quad \text{vs} \quad H_1 : F \neq F_0$$

Now we can consider the statistics

$$D_n = \sqrt{n} \max_x | F_n(x) - F_0(x) |$$

If the null hypothesis is true, then $D_n$ is well approximated by the KS distribution. On the other hand, suppose the null hypothesis fails. Since $F$ is the true cdf of the data, by law of large numbers the empirical cdf $F_n$ will converge to $F$ and it will not approximate $F_0$. For large $n$ we will have $\max_x | F_n(x) - F_0(x) | > \delta$ for small enough $\delta$. Then one finds

$$D_n = \sqrt{n} \max_x | F_n(x) - F_0(x) | > \delta \sqrt{n}$$
Kolmogorov-Smirnov nonparametric testing

If $H_0$ fails $D_n > \delta \sqrt{n} \to \infty$ for $n \to \infty$. Therefore, to test $H_0$ we will consider the decision rule

$$\delta = \begin{cases} H_0 : D_n \leq c & \text{vs} & H_1 : D_n > c \end{cases}$$

The threshold $c$ depends on the level of significance $\alpha$ and can be found from the condition

$$\alpha = P[\delta 
eq H_0 | H_0] = P[D_n > c | H_0]$$

When $n$ is large we can use the KS distribution to find $c$ since

$$\alpha = P[D_n > c | H_0] \sim 1 - H(c)$$
Kolmogorov-Smirnov nonparametric testing

Example: KS tests on subsamples of ellipticals from the Disney-Wall (1977) sample of bright ellipticals.

Top: distribution in minor-to-major axis, for radio undetected and detected ellipticals. KS test rejects the hypothesis that the subsamples are drawn from the same true distribution, at a significance of <1%.

Bottom: same for ellipticals closer and more distant than 30 Mpc. KS test indicates no significant difference between the intrinsic distributions of the two subsamples.
Regression and correlation

**Pearson correlation coefficient**

We are now interested in testing whether two variables are stochastically independent, and to define statistics that measure the association between two variables. Given measured datasets \{x_i\} and \{y_i\} we consider the sample covariance

\[
s(x,y) = \frac{1}{(n-1)} \sum_i (x_i - \bar{x})(y_i - \bar{y})
\]

and the *Pearson* correlation coefficient

\[
r(x,y) = \frac{s(x,y)}{s(x)s(y)}
\]

The correlation satisfies the following trivial properties

\[
r(cx,y) = \pm r(x,y) \quad \text{if } c > 0 \text{ or } c < 0 \quad \text{and} \quad r(x+c,y+c)=r(x,y)
\]
Regression and correlation

Now suppose we have a cloud of measure points in the \((x,y)\) plane. Fix \(x\) as the predictor and \(y\) as the response variable, and try to find the best line \(y=a x+b\) passing through the points: this is a typical regression problem. We know from the least square method that the regression line minimizing the squared errors is given by

\[
y = y + s(x,y) \frac{x-x}{s^2(x)}
\]

and the minimized square error is

\[
s^2(y) [1 - r^2(x,y)]
\]

\(r(x,y)=\pm 1\) if and only if points lie on a line with positive or negative slope.
Regression and correlation

The correlation measures the degree of linearity of the sample points. If the data vectors are uncorrelated then $x$ has no value as predictor of $y$, and the regression line is the horizontal line $y=y$, $r(x,y)=0$ and the minimized square errors is $s^2(y)$. 

Note that the sample regression line with predictor variable $x$ and response variable $y$ is NOT the same as the regression line with predictor variable $y$ and response variable $x$, except in the extreme cases where all points lie on a line.

If we model the joint distribution of $x$ and $y$ as a bivariate Gaussian, then the correlation probability tests the significance of a nonzero value of $r(x,y)$. It can be demonstrated (it is a hard work) that the probability followed by $r$ is a Student’s $t$-statistics with $n-2$ degrees of freedom

$$t = r \sqrt{(n-2) / (1-r^2)}$$
Regression and correlation

Some examples:

- \( r = 0.005 \pm 0.031 \)
  - prob. null hypothesis \( \sim 0.8 \)

- \( r = 0.861 \pm 0.008 \)
  - prob. null hypothesis \( \sim 10^{-8} \)
Regression and correlation

Caveat 1:
$r(x,y)=0$ does NOT imply that the variables are uncorrelated, but only that there is not a linear relation.

Caveat 2:
beware of summary statistics!

Anscombe’s quartet comprises four datasets with nearly identical simple statistical properties: same mean and sample variance in $x$ and $y$, same regression line and same $r=0.816$. 
Regression and correlation

**Spearman rank correlation coefficient**

The Pearson coefficient assumes that the relation is linear and that the variances on x and y account for the errors on the data and their intrinsic scatter. However, in general these conditions are not satisfied if either we have no prior information about the parent distribution or when statistical errors are different for each datapoint. The solution is to adopt nonparametric test, typically the Spearman (or Kendall) correlation: analyse the correlation in the ranks (ordinal number of sorted values) where variables are distributed uniformly an no prior is needed. The **Spearman rank correlation coefficient** is defined as

\[
r_s = 1 - 6 \frac{(n^3-n)}{n} \sum_i [v_i-u_i]^2
\]

where \([v_i-u_i]\) is the difference in ranks in the i-th observations. The probability followed by the Spearman rank correlation coefficient is again a Student’s t with n-1 d.o.f.
Regression and correlation

A Spearman coefficient equal to 1 results when the two variables are monotonically related, even if the relationship is nonlinear.

When data are roughly elliptically distributed and there are no prominent outliers, Spearman and Pearson coefficients are similar.

Spearman coefficient is less sensitive than Pearson to strong outliers in the tails of the sample, because Spearman limits the outlier to the value of its rank.
Partial correlation

Given two variables $X$ and $Y$, if you know they both depend on a third variable $Z$ in the same way, then you should study how strong is the correlation between $X$ and $Y$ considering that it can be partially induced by $Z$ (the lurking variable).

It may be of interest to know if there is a correlation between $X$ and $Y$ that it is NOT due to their common relation with $Z$. To do this one computes the partial correlation

$$r(x,y|z) = \frac{r(x,y) - r(x,z)r(y,z)}{\sqrt{1-r^2(x,z)} \sqrt{1-r^2(y,z)}}$$

that measures the relationship between $X$ and $Y$ while mathematically controlling the influence of $Z$ by holding it constant. The above correlation can be Pearson or Spearman.
Regression and correlation

Example: churches vs. serious crimes (lurking variable is city dimension)
Regression and correlation

**Principal component analysis**

Principal component analysis (PCA) is the ultimate searcher for correlation when many variables are present. It can answer the questions: given a sample of N objects with n measured parameters for each, what is correlated with what? What variables produce primary correlations, and what produce secondary via the lurking third (or n-2) variables? PCA is an algorithm belonging to the family of multivariate statistics, that operates as follows. It finds a new set of variables

\[
\xi_i = \sum_{j=1}^{n} a_{ij} x_j
\]

with values of \(a_{ij}\) such that the smallest number of new variables account for as much of the variance as possible. The \(\xi_i\) are the principal components. If most of the variance involves just a few of the n new variables, we have found a simplified description of the data.
Regression and correlation

**Example**: sample of QSO by Francis & Wills 1999

<table>
<thead>
<tr>
<th>PG name</th>
<th>$L_{1216}$</th>
<th>$\alpha_\times$</th>
<th>logFWM</th>
<th>H$\beta$/</th>
<th>logEW</th>
<th>logFWM</th>
<th>logEW</th>
<th>CIV/</th>
<th>logEW</th>
<th>SiIII/</th>
<th>NV/</th>
<th>$\lambda$1400/</th>
</tr>
</thead>
<tbody>
<tr>
<td>0947+396</td>
<td>45.66</td>
<td>1.51</td>
<td>3.684</td>
<td>0.23</td>
<td>1.18</td>
<td>3.520</td>
<td>2.08</td>
<td>1.78</td>
<td>0.45</td>
<td>1.24</td>
<td>0.306</td>
<td>0.179 0.143</td>
</tr>
<tr>
<td>0953+414</td>
<td>45.83</td>
<td>1.57</td>
<td>3.496</td>
<td>0.25</td>
<td>1.26</td>
<td>3.432</td>
<td>2.19</td>
<td>1.78</td>
<td>0.40</td>
<td>1.24</td>
<td>0.164</td>
<td>0.189 0.093</td>
</tr>
<tr>
<td>1114+445</td>
<td>44.99</td>
<td>0.88</td>
<td>3.660</td>
<td>0.20</td>
<td>1.23</td>
<td>3.654</td>
<td>2.27</td>
<td>1.85</td>
<td>0.42</td>
<td>1.48</td>
<td>0.222</td>
<td>0.175 0.092</td>
</tr>
<tr>
<td>1115+407</td>
<td>45.41</td>
<td>1.89</td>
<td>3.236</td>
<td>0.54</td>
<td>0.78</td>
<td>3.403</td>
<td>1.90</td>
<td>1.51</td>
<td>0.33</td>
<td>1.14</td>
<td>0.385</td>
<td>0.228 0.134</td>
</tr>
<tr>
<td>1116+215</td>
<td>46.00</td>
<td>1.73</td>
<td>3.465</td>
<td>0.47</td>
<td>1.00</td>
<td>3.446</td>
<td>2.14</td>
<td>1.71</td>
<td>0.34</td>
<td>1.20</td>
<td>0.440</td>
<td>0.254 0.126</td>
</tr>
<tr>
<td>1202+281</td>
<td>44.77</td>
<td>1.22</td>
<td>3.703</td>
<td>0.29</td>
<td>1.56</td>
<td>3.434</td>
<td>2.72</td>
<td>2.41</td>
<td>0.69</td>
<td>1.87</td>
<td>0.164</td>
<td>0.154 0.098</td>
</tr>
<tr>
<td>1216+069</td>
<td>46.03</td>
<td>1.36</td>
<td>3.725</td>
<td>0.20</td>
<td>1.00</td>
<td>3.514</td>
<td>2.12</td>
<td>1.95</td>
<td>0.54</td>
<td>1.20</td>
<td>0.037</td>
<td>0.121 0.056</td>
</tr>
<tr>
<td>1226+023</td>
<td>46.74</td>
<td>0.94</td>
<td>3.547</td>
<td>0.57</td>
<td>0.70</td>
<td>3.477</td>
<td>1.64</td>
<td>1.44</td>
<td>0.45</td>
<td>1.00</td>
<td>0.280</td>
<td>0.174 0.018</td>
</tr>
<tr>
<td>1309+355</td>
<td>45.55</td>
<td>1.51</td>
<td>3.468</td>
<td>0.28</td>
<td>1.28</td>
<td>3.406</td>
<td>2.01</td>
<td>1.68</td>
<td>0.41</td>
<td>1.15</td>
<td>0.303</td>
<td>0.131 0.064</td>
</tr>
<tr>
<td>1322+659</td>
<td>45.42</td>
<td>1.69</td>
<td>3.446</td>
<td>0.59</td>
<td>0.90</td>
<td>3.351</td>
<td>2.19</td>
<td>1.85</td>
<td>0.41</td>
<td>1.30</td>
<td>0.291</td>
<td>0.135 0.097</td>
</tr>
<tr>
<td>1352+183</td>
<td>45.34</td>
<td>1.52</td>
<td>3.556</td>
<td>0.46</td>
<td>1.00</td>
<td>3.548</td>
<td>2.14</td>
<td>1.80</td>
<td>0.41</td>
<td>1.29</td>
<td>0.357</td>
<td>0.203 0.116</td>
</tr>
<tr>
<td>1402+261</td>
<td>45.74</td>
<td>1.93</td>
<td>3.281</td>
<td>1.23</td>
<td>0.30</td>
<td>3.229</td>
<td>1.91</td>
<td>1.59</td>
<td>0.39</td>
<td>1.09</td>
<td>0.568</td>
<td>0.227 0.161</td>
</tr>
<tr>
<td>1415+451</td>
<td>45.08</td>
<td>1.74</td>
<td>3.418</td>
<td>1.25</td>
<td>0.30</td>
<td>3.434</td>
<td>2.32</td>
<td>1.78</td>
<td>0.29</td>
<td>1.40</td>
<td>0.688</td>
<td>0.210 0.142</td>
</tr>
<tr>
<td>1427+480</td>
<td>45.54</td>
<td>1.41</td>
<td>3.405</td>
<td>0.36</td>
<td>1.76</td>
<td>3.300</td>
<td>2.03</td>
<td>1.82</td>
<td>0.49</td>
<td>1.21</td>
<td>0.265</td>
<td>0.126 0.117</td>
</tr>
<tr>
<td>1440+356</td>
<td>45.23</td>
<td>2.08</td>
<td>3.161</td>
<td>1.19</td>
<td>1.00</td>
<td>3.192</td>
<td>2.14</td>
<td>1.54</td>
<td>0.21</td>
<td>1.05</td>
<td>0.747</td>
<td>0.141 0.092</td>
</tr>
<tr>
<td>1444+407</td>
<td>45.92</td>
<td>1.91</td>
<td>3.394</td>
<td>1.45</td>
<td>0.30</td>
<td>3.479</td>
<td>1.99</td>
<td>1.34</td>
<td>0.21</td>
<td>1.06</td>
<td>0.809</td>
<td>0.335 0.164</td>
</tr>
<tr>
<td>1512+370</td>
<td>46.04</td>
<td>1.21</td>
<td>3.833</td>
<td>0.16</td>
<td>1.76</td>
<td>3.546</td>
<td>2.02</td>
<td>2.05</td>
<td>0.75</td>
<td>1.28</td>
<td>0.228</td>
<td>0.182 0.050</td>
</tr>
<tr>
<td>1626+554</td>
<td>45.48</td>
<td>1.94</td>
<td>3.652</td>
<td>0.32</td>
<td>0.95</td>
<td>3.631</td>
<td>2.14</td>
<td>1.80</td>
<td>0.39</td>
<td>1.36</td>
<td>0.197</td>
<td>0.217 0.118</td>
</tr>
</tbody>
</table>
Regression and correlation

For each of the 13 variables, normalize by subtracting mean and dividing by stddev.

<table>
<thead>
<tr>
<th>qso\data:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.14</td>
<td>-0.14</td>
<td>1.014</td>
<td>-0.80</td>
<td>0.39</td>
<td>0.636</td>
<td>-0.13</td>
<td>0.09</td>
<td>0.215</td>
<td>-0.069</td>
<td>-0.252</td>
<td>-0.170</td>
<td>1.005</td>
</tr>
<tr>
<td>2</td>
<td>0.52</td>
<td>0.04</td>
<td>-0.061</td>
<td>-0.75</td>
<td>0.57</td>
<td>-0.103</td>
<td>0.39</td>
<td>0.09</td>
<td>-0.157</td>
<td>-0.069</td>
<td>-0.934</td>
<td>0.022</td>
<td>-0.300</td>
</tr>
<tr>
<td>3</td>
<td>-1.35</td>
<td>-2.03</td>
<td>0.877</td>
<td>-0.88</td>
<td>0.50</td>
<td>1.760</td>
<td>0.77</td>
<td>0.38</td>
<td>-0.008</td>
<td>1.175</td>
<td>-0.655</td>
<td>-0.246</td>
<td>-0.326</td>
</tr>
<tr>
<td>4</td>
<td>-0.42</td>
<td>0.99</td>
<td>-1.548</td>
<td>-0.04</td>
<td>-0.55</td>
<td>-0.346</td>
<td>-0.99</td>
<td>-1.06</td>
<td>-0.678</td>
<td>-0.588</td>
<td>0.128</td>
<td>0.771</td>
<td>0.770</td>
</tr>
<tr>
<td>5</td>
<td>0.89</td>
<td>0.51</td>
<td>-0.238</td>
<td>-0.21</td>
<td>-0.03</td>
<td>0.015</td>
<td>0.15</td>
<td>-0.21</td>
<td>-0.604</td>
<td>-0.276</td>
<td>0.392</td>
<td>1.270</td>
<td>0.561</td>
</tr>
<tr>
<td>6</td>
<td>-1.84</td>
<td>-1.01</td>
<td>1.123</td>
<td>-0.66</td>
<td>1.27</td>
<td>-0.086</td>
<td>2.90</td>
<td>2.77</td>
<td>2.001</td>
<td>3.197</td>
<td>-0.934</td>
<td>-0.649</td>
<td>-0.170</td>
</tr>
<tr>
<td>7</td>
<td>0.96</td>
<td>-0.59</td>
<td>1.192</td>
<td>-0.88</td>
<td>-0.03</td>
<td>0.585</td>
<td>0.06</td>
<td>0.81</td>
<td>0.885</td>
<td>-0.276</td>
<td>-1.544</td>
<td>-1.283</td>
<td>-1.266</td>
</tr>
<tr>
<td>8</td>
<td>2.54</td>
<td>-1.85</td>
<td>0.231</td>
<td>0.03</td>
<td>-0.73</td>
<td>0.275</td>
<td>-2.22</td>
<td>-1.36</td>
<td>0.215</td>
<td>-1.313</td>
<td>-0.377</td>
<td>-0.265</td>
<td>-2.258</td>
</tr>
<tr>
<td>9</td>
<td>-0.11</td>
<td>-0.14</td>
<td>-0.221</td>
<td>-0.68</td>
<td>0.62</td>
<td>-0.321</td>
<td>-0.47</td>
<td>-0.34</td>
<td>-0.083</td>
<td>-0.536</td>
<td>-0.266</td>
<td>-1.091</td>
<td>-1.057</td>
</tr>
<tr>
<td>10</td>
<td>-0.40</td>
<td>0.40</td>
<td>-0.347</td>
<td>0.08</td>
<td>-0.27</td>
<td>-0.782</td>
<td>0.39</td>
<td>0.38</td>
<td>-0.083</td>
<td>0.242</td>
<td>-0.324</td>
<td>-1.014</td>
<td>-0.196</td>
</tr>
<tr>
<td>11</td>
<td>-0.57</td>
<td>-0.11</td>
<td>0.282</td>
<td>-0.24</td>
<td>-0.03</td>
<td>0.870</td>
<td>0.15</td>
<td>0.17</td>
<td>-0.083</td>
<td>0.190</td>
<td>-0.007</td>
<td>0.291</td>
<td>0.300</td>
</tr>
<tr>
<td>12</td>
<td>0.31</td>
<td>1.11</td>
<td>-1.291</td>
<td>1.64</td>
<td>-1.66</td>
<td>-1.805</td>
<td>-0.94</td>
<td>-0.72</td>
<td>-0.232</td>
<td>-0.847</td>
<td>1.007</td>
<td>0.752</td>
<td>1.475</td>
</tr>
<tr>
<td>13</td>
<td>-1.15</td>
<td>0.54</td>
<td>-0.507</td>
<td>1.69</td>
<td>-1.66</td>
<td>-0.086</td>
<td>1.00</td>
<td>0.09</td>
<td>-0.976</td>
<td>0.760</td>
<td>1.584</td>
<td>0.425</td>
<td>0.979</td>
</tr>
<tr>
<td>14</td>
<td>-0.13</td>
<td>-0.44</td>
<td>-0.582</td>
<td>-0.48</td>
<td>1.74</td>
<td>-1.210</td>
<td>-0.37</td>
<td>0.26</td>
<td>0.513</td>
<td>-0.225</td>
<td>-0.449</td>
<td>-1.187</td>
<td>0.326</td>
</tr>
<tr>
<td>15</td>
<td>-0.82</td>
<td>1.56</td>
<td>-1.977</td>
<td>1.55</td>
<td>-0.03</td>
<td>-2.116</td>
<td>0.15</td>
<td>-0.94</td>
<td>-1.571</td>
<td>-1.054</td>
<td>1.867</td>
<td>-0.899</td>
<td>-0.326</td>
</tr>
<tr>
<td>16</td>
<td>0.72</td>
<td>1.05</td>
<td>-0.644</td>
<td>2.18</td>
<td>-1.66</td>
<td>0.292</td>
<td>-0.56</td>
<td>-1.79</td>
<td>-1.571</td>
<td>-1.002</td>
<td>2.165</td>
<td>2.824</td>
<td>1.553</td>
</tr>
<tr>
<td>17</td>
<td>0.98</td>
<td>-1.04</td>
<td>1.867</td>
<td>-0.97</td>
<td>1.74</td>
<td>0.854</td>
<td>-0.42</td>
<td>1.23</td>
<td>2.448</td>
<td>0.138</td>
<td>-0.626</td>
<td>-0.112</td>
<td>-1.423</td>
</tr>
<tr>
<td>18</td>
<td>-0.26</td>
<td>1.14</td>
<td>0.831</td>
<td>-0.58</td>
<td>-0.15</td>
<td>1.567</td>
<td>0.15</td>
<td>0.17</td>
<td>-0.232</td>
<td>0.553</td>
<td>-0.775</td>
<td>0.560</td>
<td>0.352</td>
</tr>
</tbody>
</table>
Regression and correlation

Construct variance covariance matrix (13 by 13)

\[
\begin{pmatrix}
1.0000 & -0.1530 & 0.1135 & -0.0414 & -0.1420 & 0.0627 & -0.7656 & -0.4387 & 0.0620 & -0.6803 & -0.0962 & 0.1764 & -0.3794 \\
-0.1530 & 1.0000 & -0.6775 & 0.6117 & -0.5099 & -0.4853 & -0.0647 & -0.4348 & -0.6603 & -0.3460 & 0.6255 & 0.4159 & 0.6514 \\
0.1135 & -0.6775 & 1.0000 & -0.7000 & 0.5029 & 0.7748 & 0.2860 & 0.6694 & 0.7656 & 0.5151 & -0.7008 & -0.2118 & -0.4287 \\
-0.0414 & 0.6117 & -0.7000 & 1.0000 & -0.7829 & -0.5306 & -0.1602 & -0.5832 & -0.6826 & -0.3701 & 0.9295 & 0.5139 & 0.5182 \\
-0.1420 & -0.5099 & 0.5029 & -0.7829 & 1.0000 & 0.1549 & 0.3013 & 0.6476 & 0.6979 & 0.3944 & -0.6505 & -0.5894 & -0.4519 \\
0.0627 & -0.4853 & 0.7748 & -0.5306 & -0.1602 & 1.0000 & 0.1207 & 0.2595 & 0.2923 & 0.3465 & -0.4627 & 0.1881 & -0.1898 \\
-0.7656 & -0.0647 & 0.2860 & -0.1602 & 0.3013 & 0.1207 & 1.0000 & 0.7653 & 0.2489 & 0.8897 & -0.1574 & -0.1864 & 0.1630 \\
-0.4387 & -0.4348 & 0.6694 & -0.5306 & 0.6476 & 0.2595 & 0.7653 & 1.0000 & 0.7925 & 0.8609 & -0.6196 & -0.4830 & -0.2307 \\
0.0620 & -0.6603 & 0.7656 & -0.6826 & 0.6979 & 0.2923 & 0.2489 & 0.7925 & 1.0000 & 0.5117 & -0.7328 & -0.4608 & -0.5046 \\
-0.6803 & -0.3460 & 0.5151 & -0.3701 & 0.3944 & 0.3465 & 0.8897 & 0.8609 & 0.5117 & 1.0000 & -0.3930 & -0.2054 & 0.0287 \\
-0.0962 & 0.6255 & -0.7008 & 0.9295 & -0.6505 & 0.4627 & -0.1574 & -0.6196 & -0.7328 & -0.3930 & 1.0000 & 0.5622 & 0.5626 \\
0.1764 & 0.4159 & -0.2118 & 0.5139 & -0.5894 & 0.1881 & -0.1864 & -0.4830 & -0.4608 & -0.2054 & 0.5622 & 1.0000 & 0.6198 \\
-0.3794 & 0.6514 & -0.4287 & 0.5182 & -0.4519 & -0.1898 & 0.1630 & -0.2307 & -0.5046 & 0.0287 & 0.5626 & 0.6198 & 1.0000 \\
\end{pmatrix}
\]

Solve the eigenvalue problem to find eigenvalues, which are the variances.

**Eigenvalues:** 6.451  2.820  1.589  0.624  0.565  0.343  0.261  0.172  0.122  0.023  0.019  0.010  0.002

It is seen that the first 3 components contribute about 84% of the total variance. For eigenvalues to be significant, they must be larger than unity.
The eigenvectors are the PCs in terms of linear combination of the normalized original variables; e.g., $PC_1 = -0.055 \times x_1 - 0.294 \times x_2 + 0.330 \times x_3 + \ldots$ and so on so forth.

Now study the partial correlation between the PCs and the original variables; to test contribution of each variable to each PC, redo the analysis without that variable and then check the significance of the correlation between it and the new PC.
Regression and correlation

Residual analysis in regression

The appropriateness of a given regression model can be checked by examining residual plots, i.e., observed-predicted values vs. predicted values.

Ideally, your plot should look symmetrically distributed, and clustered toward the middle of the plot, with no clear pattern.
Regression and correlation

This requirement is not met when:

\textbf{y-axis unbalanced} \rightarrow \text{majority of points are below zero line, but a few points are very above the line. This is a typical problem induced by outliers, observations that lie abnormally distant from other values in a random sample from a population. Usually solved by transforming one variable (e.g., with a log-transform)}
Regression and correlation

This requirements is not met when:

**heteroscedasticity** → residuals get larger as the prediction moves from small to large (or vice versa). Usually indicates that a variable is missing, and the model can be improved.
Regression and correlation

This requirements is not met when:

nonlinearity $\rightarrow$ a linear model does not accurately represent the relation between variables. Solution is to create a nonlinear model.
Regression and correlation

This requirement is not met when:

**strong outliers** → the regression has clear outliers. Remove or use appropriate estimators (e.g., bilinear).
Sufficiency

We now come back to the issue of searching for the minimum variance unbiased estimator (mve). We have seen that an estimator saturating the Cramer-Rao bound is a minimum variance one. But how to search for the mve among all the unbiased estimator of a parameter? In this context it is extremely relevant the following concept. An estimator $T(X)$ of a parameter $\theta$ is a *sufficient* statistics for it if

$$f_X(x, \theta)/f_T(t, \theta) = H(x)$$

i.e., the likelihood for the whole sample is a $\theta$-independent *multiple* of the likelihood for $T$. The terminology is appropriate because the statistics $T$ exhausts all the information about $\theta$ present in the sample; e.g., given $T=t$, the conditional joint pdf of $X$ and hence of any other statistics $G(X)$ does not depend on $\theta$, and so cannot be used for statistical inference on that parameter.
We now demonstrate the so called factorization theorem: T is a sufficient statistics for \( \theta \) if and only if the likelihood for \( X \) can be factorized as follows

\[
f_X(x, \theta) = k_1[t(x), \theta] k_2(x)
\]

in terms of two nonnegative functions \( k_{1,2} \)

**Proof.**

Assume the factorization above. Then change variable from \( x \) to \( y(x) \), with \( y_1(x) = t(x) \) and the Jacobian of the transformation being \( |J| \). Then

\[
f_Y(y, \theta) = k_1(y_1, \theta) k_2(y) |J|
\]

Now we compute the pdf for \( y_1 = t(x) \)
Sufficiency

\[ f_T(y_1, \theta) = \int dy_2 \ldots dy_n f_Y(y, \theta) = \int dy_2 \ldots dy_n k_1(y_1, \theta) k_2(y) |J| = k_1(y_1, \theta) \int dy_2 \ldots dy_n k_2(y) |J| \]

Now the integral does not depend on \( \theta \), so it is only a function \( m(y_1) \). Thus

\[ f_T(y_1, \theta) = k_1(y_1, \theta) m(y_1) \]

If \( m(y_1) = 0 \) then \( f_Y(y_1, \theta) = 0 \). If \( m(y_1) > 0 \) then \( k_1(y_1, \theta) = f_T(y_1, \theta)/m(y_1) \) and the assumed factorization becomes

\[ f_X(x, \theta) = f_T(t, \theta) \times k_2(x)/m[t(x)] \]

so proving sufficiency. Conversely, if \( T \) is a sufficient statistics, then the factorization can be readily found by taking \( k_1 = f_T \). The theorem is proved.
Rao-Blackwell theorem

We now prove the so called Rao-Blackwell theorem: let $X$ and $Y$ two random variables such that $Y$ has a mean $\mu$ and positive variance $\sigma_Y^2$, and let $<Y|X>_Y = \varphi(X)$; then $<\varphi(X)>_X = \mu$ and $\sigma_{\varphi(X)}^2 \leq \sigma_Y^2$.

Proof.

We have

$$\varphi(X) = <Y|X>_Y = \int dy \ y \ f(y|X) = \int dy \ y \ f(x,y) / f(X)$$

Then

$$<\varphi(X)>_X = \int dx \ f(x) \ \varphi(x) = \int dx \int dy \ f(x,y) \ y = \int dy \ y \int dx \ f(x,y) = \int dy \ y \ f(y) = \mu$$

proving the first part of the theorem.
Sufficiency

Now consider

\[ \sigma^2_Y = \langle (Y - \mu)^2 \rangle_Y = \int dy \, f(y) \, [Y - \mu]^2 = \int dx \int dy \, f(x,y) \, [Y - \varphi(x) + \varphi(x) - \mu]^2 = \]

\[ = \langle [Y - \varphi(x)]^2 \rangle_{XY} + \langle [\varphi(x) - \mu]^2 \rangle_X + 2 \langle [Y - \varphi(x)] \, [\varphi(x) - \mu] \rangle_{XY} \]

But the last term is zero because

\[ \langle [(Y - \varphi(x)) \, [\varphi(x) - \mu]] \rangle_{XY} = \int dx \int dy \, f(x,y) \, [y - \varphi(x)] \, [\varphi(x) - \mu] = \]

\[ = \int dx \, [\varphi(x) - \mu] \, f(x) \int dy \, f(y|x) \, [y - \varphi(x)] = 0 \]

given that \( \varphi(x) = \langle Y|x \rangle = \int dy \, f(y|x) \, y. \)
Sufficiency

Accordingly

\[ \sigma_Y^2 = <[Y - \varphi(x)]^2 >_X + \sigma_{\varphi(x)}^2 \geq \sigma_{\varphi(x)}^2 \]

and this completes the proof of the theorem.

Putting together our results on sufficiency and the Rao-Blackwell theorem, we have that if \( Y_1 \) is a sufficient statistics for \( \theta \), and \( Y_2 \) is an unbiased estimator of \( \theta \) based on some other combination of data, we can make another unbiased estimator

\[ \varphi(Y_1) = < Y_2 | Y_1 > \]

which has an equal or lesser variance (Since \( Y_1 \) is sufficient, \( f(y_2|y_1) \) does not depend on \( \theta \), so \( \varphi(Y_1) \) is independent of \( \theta \) too). So when looking for the minimum variance estimator, we can confine our attention to sufficient statistics.
Sufficiency

Completeness and uniqueness

The Rao-Blackwell theorem tells us that if there is a sufficient statistics $Y$ for $\theta$, one function of it is a minimum variance estimator. But we could imagine a case where, even though we have constructed an unbiased estimator $\varphi(Y)$, there might be another unbiased estimator $\psi(Y)$ which could have a lower variance. One thing we do know about such a pair of unbiased estimator is that

$$<\varphi(Y) - \psi(Y)> = \theta - \theta = 0$$

Now in general this does not mean that $\varphi(Y) = \psi(Y)$. But suppose that the pdf $f_Y(y, \theta)$ of $Y$ has the property that any function $u(Y)$ which satisfies $<u(Y)> = 0$ for all $\theta$ also satisfies $u(Y) = 0$. Then we know that the function $\varphi(Y)$ is unique, and there is only one unbiased estimator that can be built from the sufficient statistic, and this is the mve.
Sufficiency

This property is known as completeness. We say that \( f_X(x, \theta) \) is a complete family of distributions if, for any \( u(x) \) satisfying \( \langle u(X) \rangle = 0 \) for all \( \theta \), then \( \Pr[u(X)=0]=1 \).

As an example, consider the exponential distribution with rate parameter \( \theta \), i.e.,

\[
f_X(x, \theta) = \theta \exp[- \theta x]
\]

We can write the likelihood for a sample of size \( n \) in exponential form as

\[
f_X(x, \theta) \propto \exp[- n \theta x +n \ln \theta]
\]

Now it is easy to show that \( Y = \sum_i X_i \) is a sufficient statistics for \( \theta \) by the factorization

\[
f_Y(y, \theta) = N(y) \exp[- \theta y +n \ln \theta] \propto f_X(x, \theta)
\]
This is also a complete statistics, since the only function with $<u(y)>=0$ is equal to zero. This can be demonstrated on considering that

$$<u(y)> = \exp(n \ln \theta) \int dy \ u(y) \ N(y) \ \exp[- \ \theta \ y ] = 0$$

implies $u(y) = 0$ because $N(y) > 0$ and $\exp[- \ \theta \ y ]>0$. Now it is easy to show that

$$<Y> = n/\theta$$

If we can make an unbiased estimator of $\theta$ out of this, we know it will be the mve. We already know that $<1/Y> = \theta/(n-1)$ so that the unique mve is

$$\phi(Y) = (n-1)/Y$$
In the Bayesian approach to parameter estimation, we use some data $D$ to make an inference based on Bayes’s theorem to construct the posterior pdf

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

where the prior distribution $f(\theta)$ is expected to reflect the plausibility we assign to different values of $\theta$, given our information on the process. Unfortunately, sometimes it is difficult to convert this information into a prior pdf. Some tricks can be used for constructing useful prior given seemingly incomplete information.

**Jeffreys prior**

The simplest prior one may think of is a uniform distribution over the allowed range of $\theta$. However, since $f(\theta)$ is a pdf, this causes an issue with any change of variables…
Information, entropy, and priors

If, instead of $\theta$, we use a parameter $\tau(\theta)$, then

$$f(\tau) = f(\theta) \left| d\theta/d\tau \right|$$

so if $f(\theta)$ is a constant, $f(\tau)$ will not be, in general. The Jeffrey prior is a prescription for generating a prior pdf from a likelihood function, in a way which is invariant under reparametrization. Given a log-likelihood $L(\theta)$, we construct

$$f(\theta) \propto \sqrt{I(\theta)}$$

in terms of the Fisher information $I(\theta) = < -L''(\theta) > = < [L'(\theta)]^2 >$. Changing variable yields

$$I(\tau) = I(\theta) \left| d\tau/d\theta \right|^2$$

which gives just the Jacobian needed to transform $f(\theta)$ in $f(\tau)$. 
As an example, for an exponential distribution

\[ f(x, \lambda) = \lambda \exp(-\lambda x) \]

the loglikelihood is \( L(\lambda) = \ln \lambda - \lambda x \), the Fisher information is \( I(\lambda) = 1/\lambda^2 \) and the Jeffreys prior reads

\[ f(\lambda) = 1/\lambda \]

For a Gaussian, the Jeffreys prior for \( \sigma \) is \( 1/\sigma \), but for the mean \( \mu \) it is uniform.

There is a sort of odd property of the Jeffreys prior, however. To construct it, you need to know the likelihood, and in some cases this is not simple.
**Maximum entropy principle**

A way around the issue of constructing a prior when we have incomplete information is to use the maximum entropy prescription. It suggests to choose the prior which maximizes the (Shannon) information entropy subject to any constraint

\[ S = - \sum_i p_i \ln p_i \]

where the summation is over all the states of a system with probability \( p_i \). The motivation for this definition is as follows. Suppose there are a number \( K \) of discrete states, each with probability \( p_i \). The frequentist interpretation of probability tells us that if we run the random experiment \( N \) times, it will be found in the \( i \)-th state \( N_i = N p_i \) times. The number of different ways to choose which \( N_1 \) of the \( N \) experiments are …
... in state 1, \( N_2 \) are in state 2, etc. is

\[
\Omega = \frac{N!}{N_1! N_2! \ldots N_K!}
\]

The (thermodynamic) entropy is defined as \( S \propto \ln \Omega \), the logarithm of the number of equivalent states that can be constructed. If \( N_i \) is large we can use the Sterling approximation \( \ln N_i! \sim N_i \ln N_i - N_i \) so

\[
\ln \Omega \sim N \ln N - N - \sum_i N_i \ln N_i - \sum_i N_i = -N \sum_i (N_i / N) \ln N_i / N = -N \sum_i p_i \ln p_i
\]

Since we are looking for \( S \) up to a constant we can define

\[
S \propto (\ln \Omega) / N = -\sum_i p_i \ln p_i
\]
Suppose now that only the minimal constraint of a normalized prior is present, i.e.,

\[ \Sigma_i p_i = 1 \]

Let's use the method of Lagrange multipliers and minimize

\[ S_{\text{eff}} = -\Sigma_i p_i \ln p_i + \lambda (\Sigma_i p_i - 1) \]

Differentiating with respect to \( p_j \) and \( \lambda \) one finds

\[ p_j = \exp(\lambda - 1) \quad \text{and} \quad \Sigma_{i=1}^K p_i = 1 \]

so that the distribution is uniform

\[ p_i = 1/K \]
Information, entropy, and priors

In the continuous case, one defines the information entropy as

\[ S = - \int dx \ f(x) \ln \left[ \frac{f(x)}{m(x)} \right] \]

where \( m(x) \) is a measure, called Lebesgue measure, which acts like a density in \( x \). The above definition is invariant under reparameterization, since changing variable from \( x \) to \( y \) one has the same transformation for \( f \) and \( m \).

The other added complication is that varying \( S \) with respect to \( f(x) \) is a functional derivatives. As an example, one can show that if \( m(x) = 1 \), the maximum entropy distribution with mean \( \mu \) and variance \( \sigma^2 \) is a Gaussian…

\[
S_{\text{eff}} = \int_{-\infty}^{\infty} f(x) \ln f(x) \, dx + \lambda_0 \left( \int_{-\infty}^{\infty} f(x) \, dx - 1 \right) \\
+ \lambda_1 \left( \int_{-\infty}^{\infty} xf(x) \, dx - \mu \right) + \lambda_2 \left( \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx - \sigma^2 \right)
\]
Luminosity functions, counts, and biases

Detection

The concept of detection is preliminary to whatever else occurs in astronomy. The meaning is generally to locate and perform a confident measurement on some feature of an object or a population of objects.

Two relevant points:

- Elusive objects close to the limit of detectability are often very interesting!
- Nondetections are also very important, they contain information!

Let $S$ be a flux density of a survey with flux limit $S_{\text{lim}}$. Two important properties of the survey are:

- Reliability: it is $1-F$ where $F$ is the false alarm rate, i.e., the chance that pure noise will produce a detection.
- Completeness: chance that a measurement of a real source is above flux limit.
Luminosity functions, counts, and biases

Example: suppose our measurement is \( s \) and the noise on the measurement is Gaussian, with unit standard deviation. The source has a true flux \( S \), measured in unit of the std dev. Then the pdf of the data given the source is

\[
P(s|S) = \exp[-(s-S)^2 / 2] / \sqrt{2\pi}
\]

and the pdf of the data when there is no source is

\[
P(s|S=0) = \exp[-s^2 / 2] / \sqrt{2\pi}
\]

Integrating from 0 to \( S \) makes easy to plot up completeness against reliability, taking the flux limit as a parameter.
Luminosity functions, counts, and biases

E.g., a $2\sigma$ flux limit and $3\sigma$ source yields a reliability of 100%-7%=93% and a completeness of 80%.

High completeness goes in hand with low reliability and vice versa.
Now we turn to discuss detection from the Bayesian viewpoint. Take a prior probability that a source with flux \( S \) is present in the measured area to be \( \varepsilon N(S) \), where \( N(S) \) is the pdf that a single source will have flux \( S \). The probability of no source is \( (1 - \varepsilon) \delta(S) \). Then the posterior probability \( P(\text{source present, flux } S | \text{data}) \) is

\[
\varepsilon P(\text{data}|S) N(S) / [\varepsilon \int P(\text{data}|S)N(S)dS + (1-\varepsilon) \int P(\text{data}|S=0)dS ]
\]

and integrating over \( S \) yields probability of a source present in the data.

E.g., take the noise distribution to be Gaussian and a flat or a power-law prior \( N(S) \). The value of \( \varepsilon \) reflects our initial confidence that the source is present, so typically will be small. Its role is to suppress our confidence of a detection in low S/N cases.
Luminosity functions, counts, and biases

*Left:* detection as a function of flux for $1,2,3,4\sigma$ with $\varepsilon=0.05$. *Right:* detection at any positive flux density as a function of data values for $\varepsilon=0.5,0.05,0.005$. Using a power-law prior (bottom) gives similar results with respect to a flat prior (top), but recognizes the possibility that no source may be present; rarity of bright sources in this prior means we need a better S/N to achieve same confidence of detection.
Luminosity functions, counts, and biases

*Luminosity function and number counts*

A property of an astronomical objects can have a distribution function; the most basic is the luminosity function. This is a differential density: $\phi(L)dL$ is the number density of objects with luminosity in the range $dL$ (although usually logarithmic bins are used). We can integrate the distribution function to get the total number of objects per unit volume, and divide by this to get a pdf:

$$p(L) = \frac{\phi(L)}{\int dL \phi(L)}$$

Suppose objects are distributed with a constant density. Consider some area of solid angle $\Omega$ and limiting flux density $S$: for each luminosity, the objects can be seen out to a distance $D$, where

$$S = \frac{L}{4\pi D^2}$$
Luminosity functions, counts, and biases

so \( D = (L/4\pi S)^{1/2} \), and the corresponding volume is \( V = \Omega D^{3/3} \). The total number of objects brighter than \( S \) is then given by multiplying by the luminosity function and integrating:

\[
N(>S) = \int dL \phi \quad V = \Omega \quad \int dL \phi \frac{L^{3/2}}{3} (4\pi S)^{3/2}
\]

which comes out as \( N \propto S^{-3/2} \), independent of the form of \( \phi \). The differential distribution function (or number counts) is thus

\[
dN/dS \propto S^{-\gamma} \quad \text{with} \quad \gamma \sim 5/2
\]

the so-called Euclidean number counts. A shallower slope arises if we are seeing to a point where the density falls, or where the flux-distance relation in cosmology becomes non-Euclidean. A larger slope implies more objects at large distances.
**Luminosity functions, counts, and biases**

*Maximum likelihood fitting of number counts*

Given observations with flux density $S$ above a known measurement limit $S_0$, what is the best estimate for $\gamma$? The model pdf for $S$ is

$$p(S) \, dS = (\gamma - 1) \, S_0^{-\gamma} \, S^{-\gamma} \, dS$$

where the factor in front arises because of the normalization condition $\int_{S_0}^{\infty} p(S) \, dS = 1$. The log-likelihood for $n$ observations reads

$$L(\gamma) = \sum_{i=1}^{n} [\ln(\gamma - 1) + (\gamma - 1) \ln S_0 - \gamma \ln S_i]$$

The maximum likelihood estimate is

$$\gamma = 1 + n / \sum_{i=1}^{n} \ln(S_i / S_0)$$
Luminosity functions, counts, and biases

**Malquist and Eddington bias**

Unfortunately, the determination of the number counts may be affected by biases. The most important are:

**Eddington bias** → due to observational errors, some fainter fluxes will be measured as being slightly brighter than they really are, and vice versa. Because there are more faint than bright sources, there can be a tendency to increase the number counts over the true value.

**Malquist bias** → Because there are more intrinsically faint than bright sources (in terms of luminosity), the objects found at given measured flux density will have their flux overestimated.
Luminosity functions, counts, and biases

Suppose $x$ is the observable, and $f(x)$ the pdf for number of objects in $dx$. Let $p(\varepsilon)$ be the pdf for getting an error $\varepsilon$ in $x$, so the true value is $x_{\text{obs}} - \varepsilon$. The pdf observed in $x$ is

$$g(x_{\text{obs}}) = \int d\varepsilon \, f(x_{\text{obs}} - \varepsilon) \, p(\varepsilon)$$

The bias in $x_{\text{obs}}$ is simply

$$\langle \varepsilon \rangle = \int d\varepsilon \, f(x_{\text{obs}} - \varepsilon) \, p(\varepsilon) \, \varepsilon / g(x_{\text{obs}})$$

Now Taylor expand and recognize that the average of $\varepsilon$ over the pdf is null by definition

$$g(x_{\text{obs}}) \sim \int d\varepsilon \left[ f(x_{\text{obs}}) - f'(x_{\text{obs}}) \, \varepsilon + \frac{1}{2} f''(x_{\text{obs}}) \, \varepsilon^2 \right] p(\varepsilon) = f(x_{\text{obs}}) + \frac{1}{2} f''(x_{\text{obs}}) \sigma^2$$

where $\sigma$ is the rms error. The Eddington bias depends mainly on the curvature of the pdf.
Luminosity functions, counts, and biases

Similarly, to lowest order, the Malquist bias is

\[
< \varepsilon > = \int d\varepsilon \left[ f(x_{\text{obs}}) - f'(x_{\text{obs}}) \varepsilon \right] p(\varepsilon) \varepsilon / f(x_{\text{obs}}) = - f'(x_{\text{obs}}) \sigma^2 / f(x_{\text{obs}})
\]

and depends mainly on the first derivative of the pdf.

In terms of number counts, one has the fractional Eddington and Malquist bias

\[
1 + \frac{1}{2} f'' \sigma^2 / f = 1 + \frac{1}{2} (\gamma - 1) \sigma^2
\]

\[
\Delta \ln S = (\gamma - 1) \sigma^2
\]

The term Malquist bias tends to be used quite broadly, especially in conjunction with the fact that more luminous objects are seen to greater distances. So the mean luminosity of a flux limited sample will be higher than the mean luminosity just averaging over the luminosity function.
Luminosity functions, counts, and biases

1/$V_{\text{max}}$ method

One of the best methods to determine the luminosity function from data is to use the 1/$V_{\text{max}}$ estimator. Suppose you have a catalog of objects with measured luminosities $L_i$. Let $V_{\text{max}}(L_i)$ the maximum volumes within which an object with luminosity $L_i$ could lie and still be in the catalog. This depends on survey limits, distribution of objects, and dependence of detectability on distance. The estimate of the binned luminosity function is given by

$$N(L_{j-1} < L < L_j) = \sum_{L_{j-1} < L_i < L_j} 1 / V_{\text{max}}(L_i)$$

This is a maximum likelihood estimator; moreover, errors are uncorrelated in each bin, and take on value $1/\sqrt{N_j}$ where $N_j$ is the number of objects in each bin. A better estimate can be obtained with bootstrap methods.
Luminosity functions, counts, and biases

Example: SFR functions from far-IR (*Herschel*) and UV (*HST*) data
Luminosity functions, counts, and biases

**Upper and lower limits: survival analysis**

In many cases only upper and lower limits are available in catalogues (censored data). Statistics dealing with limits is called survival analysis. This is a terminology coming from medicine: at the end of experiment some subjects have survived, some others do not.

The key assumption is that the chance of having a limit for some property is independent of the true value of that property: censoring is random. This is often met for flux-limited sample: e.g., for an object of true luminosity $L$ and distance $D$, the condition for censoring upper limit is that $L/D^2 < S$, where $S$ is the flux limit of the survey. If $D$ is a random variable independent of $L$ and $S$ is fixed, then the chance of censoring is independent of $L$.

Typically one uses the Kaplan-Meier estimator, which is a maximum likelihood estimator. It is defined as
Luminosity functions, counts, and biases

\[ K(L_k) = 1 - \prod_{i=1}^{k-1} (1 - d_i/n_i)^{\delta_i} \]

For lower limits: arrange data in increasing order
\( d_i \) is the number of observations with \( L_i \)
\( n_i \) is the number of observations with \( \geq L_i \)
\( \delta_i = 1 \) for detection and 0 for lower limit

For upper limits: arrange data in decreasing order
\( d_i \) is the number of observations with \( L_i \)
\( n_i \) is the number of observations with \( \leq L_i \)
\( \delta_i = 1 \) for detection and 0 for upper limit

In general results from survival analysis and \( 1/V_{\text{max}} \) method agree.
Luminosity functions, counts, and biases

Example: High-redshift UV luminosity functions
Luminosity functions, counts, and biases

Confusion limit and $P(D)$ analysis

In many astronomical instances, faint objects are more numerous than bright ones. Faint objects crowd together, and ultimately they start to be unresolved and our signal becomes a mixture of objects with different intensities, blended together by the resolution of our instrument. Here it is an example: simulation of a 1D scan of sources obeying an input euclidean source counts $(N>S) \propto S^{-1.5}$ analyzed with a Gaussian beam, with an average of 1 source per beam. The maximum likelihood slope of the source counts inferred from counting the peaks in the record is -1.8, substantially steeper than the input.
Luminosity functions, counts, and biases

Statistics dealing with confusion-limited data is known a $P(D)$ or probability of deflection, the deflections being that of a pen on a chart recorder. The method derives the pdf of a measurement from the underlying source counts; its benefits are that information is obtained from sources that are much too faint to be detected as individuals, and the correct form of the counts are derived in an unbiased way.

The derivation is extremely complex and requires conditional pdf, autocorrelation functions, Fourier transforms. The net result is that the Fourier transform $P(\omega)$ of the $p(D)$ distribution contains the source counts in the form

$$P(\omega) = \exp[R(\omega) - R(0)]$$

$$r(s) = \int dx \frac{N[s/\Omega(x)]}{\Omega(x)}$$

where $R(\omega)$ is the Fourier transform of $r(s)$. Of course to recover $N(s)$ is a complex procedure that requires modeling and simulations, taking into account realistic noise.
Luminosity functions, counts, and biases

Example: radio counts