Point Estimation

The goal of Point Estimation is
to find the point in μ-space which gives
the "best" estimate (measurement) of the parameter μ.

We assume, as always, $P(\text{data}|\text{hypothesis}) = P(X|\mu)$ known.

What we mean by the "best" estimate depends very much on whether we will use a Frequentist or Bayesian method.

Historically, the Bayesian was the first method, so we start there.

Bayes' Theorem for Parameter Estimation

For estimation of the parameter μ, we can rewrite Bayes' Theorem:

$$P(\mu|\text{data}) = \frac{P(\text{data}|\mu)P(\mu)}{P(\text{data})}$$

Evaluating $P(\text{data}|\mu)$ at the observed data is the likelihood function, so we have:

$$P(\mu|\text{data}) = \frac{\mathcal{L}(\mu)P(\mu)}{P(\text{data})}$$

which is a **probability density function** in the unknown μ.

$P(\text{data})$ is just a constant, which can be determined from the normalization condition: $\int_\Omega P(\mu|\text{data}) = 1$

Note that the above cannot be Frequentist probabilities, because hyp and μ are not random variables.
They determine the degree of belief in different values of μ.

Priors and Posteriors

Assigning names to the different factors, we get:

$$\text{Posterior pdf}(\mu) = \frac{\mathcal{L}(\mu) \times \text{Prior pdf}(\mu)}{\text{normalization factor}}$$

The Prior pdf represents your belief about μ before you do any experiments. If you already have some experimental knowledge about μ (for example from a previous experiment), you can use the posterior pdf from the previous expt. as the prior for the new one. But this implies that somewhere in the beginning there was a prior which contained no experimental evidence [Glen Cowan calls this the Ur-prior].

The usual case is that we don't want to include information from previous measurements or any other ideas we may have about the parameter value. Hence the search for a prior that expresses prior ignorance, the uninformative prior.

The Search for the Uninformative Prior

The most obvious candidate is the uniform prior, with density uniform over the range of μ which is usually $(0, \infty)$ or $(-\infty, \infty)$. The uniform (or flat) prior is the limit of a Gaussian density with infinite variance, and it simplifies calculations since it is just $= 1$. Unfortunately, in this case:

1. The prior is not a proper density, since it cannot be normalized.
2. It puts most of the belief at infinity, hardly what we want.
3. It is not invariant under nonlinear change of variables $\mu \to f(\mu)$.
4. It is not really uninformative since you will get a different answer depending on whether the prior is uniform in μ, $1/\mu$, $\ln \mu$, etc.

Two of the four problems given above can be resolved by moving the endpoint(s) from infinity to a large but finite value, the cutoff C, so the density is now uniform over $(0, C)$ or $(-C, C)$. The prior then becomes normalizable, and if C is big enough, you can usually show that the posterior is (almost) independent of C.

Alternatives to the Uniform Prior

Faced with the problems of uninformative priors,
Harold Jeffreys (1891-1989) English mathematician, astronomer, geophysicist and Bayesian statistician, took a somewhat different view. He tried to find the "right" prior, like a law of physics.

For the Poisson case, he proposed the prior $P(\mu) = 1/\mu$, based on the argument of scale invariance. This was the first of the Jeffreys priors.

The $1/\mu$ prior has some attractive features: It is uniform in ln μ, and it goes to zero smoothly for large μ, so it could really represent belief. However, it is not normalizable (an improper pdf).

There is much more to be said about priors, but it is better to leave that until we consider Interval Estimation, and we have seen how Frequentist Point Estimation works.

From Posterior Density to Point Estimate

In the true Bayesian spirit, the posterior density $P(\mu \,|\, \text{dat})$ represents all our knowledge and belief about μ, so there is no need to process this pdf any further, but if we want a point estimate we must take another step.

Given the Bayesian posterior density for μ, the most common ways to make the Bayesian point estimate $\hat{\mu}$ are:

1. The posterior expectation of μ, $E(\mu) = \int_\Omega \mu P(\mu \,|\, \text{data}) d\mu$
2. The posterior mode of $P(\mu \,|\, \text{data})$, the value of μ for which the posterior pdf takes on its maximum value.

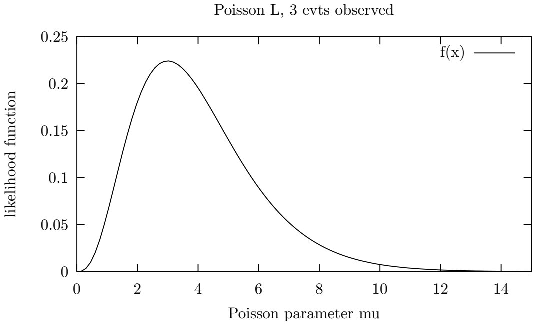
The latter is sometimes called incorrectly the most probable value.
A better name is the highest posterior density, or HPD point.

If the Prior probability density is taken as a uniform density,
then the HPD point will occur at the maximum of the likelihood $\mathcal{L}(\mu)$.

Example of Bayesian Posterior

Example: In a Poisson process, we observe 3 events.

$$\mathcal{L}(\mu) = P(3|\mu) = \frac{e^{-\mu}\mu^3}{3!}$$



If the prior $P(\mu)$ is flat, the peak in the posterior occurs at $\mu = 3$.
For n events observed, with flat prior, the peak occurs at $\mu = n$.

.

Bayesian alternatives for Estimation of Poisson Mean

We have considered two possible priors and two choices of point estimate. Applying these four combinations to our example of the Poisson mean, we obtain four possible Bayesian estimates:

Bayesian Point Estimates of Poisson Mean

Estimate taken as	Uniform Prior	$1/\mu$ Prior
Expectation $E(\mu)$	$\hat{\mu} = n + 1$	$\hat{\mu} = n$
Posterior Mode	$\hat{\mu} = n$	$\hat{\mu} < n$

There are of course many other possible choices.

Point Estimation – Bayesian Summary

Bayesian point estimation is considered by statisticians as a mathematically coherent methodology.

However, for physicists, it requires some things which are hard to accept:

1. The definition of probability as degree of belief is necessarily subjective and hard to measure.
2. The prior pdf is required, although it is usually unknown, subjective, or arbitrary.

In practice, both these problems become less important as the amount of data increases, because asymptotically:

- the data should dominate the prior and
- the Posterior pdf should tend toward a Gaussian.

However, in this limit, almost any statistical method would give the same result, so this is not a good argument in favor of Bayesian methods.

The modern justification for using Bayesian methods is always given in terms of Decision Theory, which will come later in chap. 6.

Point Estimation - from Bayesian to Frequentist

Up to the early 1900's, the only statistical theory was Bayesian.

In fact, frequentist methods were already being used:
Linear least-squares fitting of data had been in use for many years, although its statistical properties were unknown.

And in 1900, Karl Pearson published the Chi-square test
to be treated later under *goodness-of-fit*.

About the same time, another English biologist, R. A. Fisher, was one of several people looking for a statistical theory that would not require as input prior belief and would not be based on subjective probabilities.

He succeeded in making a frequentist theory of point estimation,
(but was unable to produce an acceptable theory of interval estimation).

Point Estimation - Frequentist

An Estimator $\mathcal{E}_\theta$ is a function of the data X which can be used to estimate (measure) the unknown parameter θ to produce the estimate $\hat{\theta}$.

$$\hat{\theta} = \mathcal{E}_\theta(X)$$

The goal:
Find that function $\mathcal{E}_\theta$ which gives estimates $\hat{\theta}$
closest to the true value of θ.

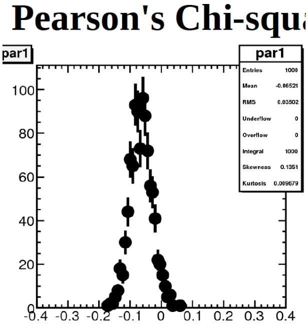
As usual, we know $P(X|\theta)$
and because the estimate is a function of the data,
we also know the distribution of $\hat{\theta}$, for any given value of θ:

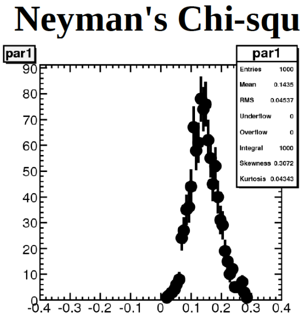
$$P(\hat{\theta}|\theta) = \int_X \mathcal{E}_\theta(X) P(X|\theta) dX$$

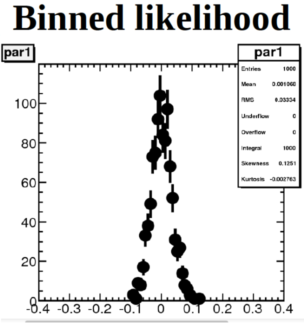
.

Frequentist Estimates

For some trial estimators $\mathcal{E}_\theta$, where the true value of $\theta = 0$, the distributions of estimates $\hat{\theta}$ might look something like this:







Now we can see whether these estimators have the desired properties. Are they (1) consistent, (2) unbiased, (3) efficient, and (4) robust?

Consistency

Let $\mathcal{E}_\theta$ be an estimator producing estimates $\hat{\theta}_n$, where n is the number of observations entering into the estimate.

Given any $\varepsilon > 0$ and any $\eta > 0$, $\mathcal{E}_\theta$ is a consistent estimator of θ if an N exists such that

$$P(|\hat{\theta}_n - \theta_0| > \varepsilon) < \eta$$

for all $n > N$, where θ_0 is the assumed true value.

That is, if $\mathcal{E}_\theta$ is a consistent estimator of θ,
the estimates $\hat{\theta}_n$ converge (in probability) to the true value of θ.

Since all reasonable Frequentist estimators are consistent, I thought this property was only of theoretical interest, until I discovered that Bayesian estimators are not in general consistent in many dimensions.

Bias

We define the bias b of the estimate $\hat{\theta}$ as the difference between the expectation of $\hat{\theta}$ and the true value θ_0,

$$b_N(\hat{\theta}) = E(\hat{\theta}) - \theta_0 = E(\hat{\theta} - \theta_0)\,.$$

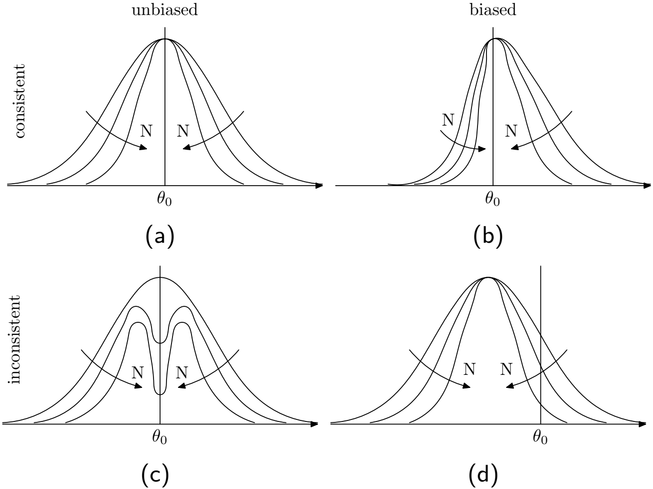
Thus, an estimator is unbiased if, for all N and θ_0,

$$b_N(\hat{\theta}) = 0$$

or

$$E(\hat{\theta}) = \theta_0\,.$$

Bias vs Consistency



(a) (b)

(c) (d)

Figure: examples of distributions of estimates with different properties. The arrows show increasing amount of data.

Efficiency

Among those estimators that are consistent and unbiased, we clearly want the one whose estimates have the smallest spread around the true value, that is, estimators with a small variance.

We define the efficiency of an estimator in terms of the variance of its estimates $V(\hat{\theta})$:

$$\text{Efficiency} = \frac{V_{\text{min}}}{V(\hat{\theta})}$$

where V_{min} is the smallest variance of any estimator.

The above definition is possible because, as we shall see, V_{min} is given by the Cramér-Rao lower bound.

Fisher Information

Let the pdf of the data X be denoted by f or by L:

$$P(\text{data}|\text{hypothesis}) = f(X|\theta) = L(X|\theta)$$

depending on whether we are primarily interested in the dependence on X or θ.

The amount of information given by an observation X about the parameter θ is defined by the following expression (if it exists)

$$\begin{aligned} I_X(\theta) &= E\left[\left(\frac{\partial \ \ln \ L(X|\theta)}{\partial\theta}\right)^2\right] \\ &= \int_{\Omega_\theta}\left(\frac{\partial \ \ln \ L(X|\theta)}{\partial\theta}\right)^2 L(X|\theta)dX \,. \end{aligned}$$

Fisher Information cont.

If $\boldsymbol{\theta}$ has k dimensions, the definition becomes

$$\left[\mathcal{I}_X(\boldsymbol{\theta}) \right]_{ij} = E\left[\frac{\partial \ \ln \ L(X|\boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial \ \ln \ L(X|\boldsymbol{\theta})}{\partial \theta_j} \right]$$

$$= \int_{\Omega_\theta} \left[\frac{\partial \ \ln \ L(X|\boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial \ \ln \ L(X|\boldsymbol{\theta})}{\partial \theta_j} \right] L(X|\boldsymbol{\theta}) dX \, .$$

Thus, in general, $\mathcal{I}_X(\boldsymbol{\theta})$ is a $k \times k$ matrix. Assuming certain regularity conditions, the same matrix can be expressed as the expectation of the second derivative matrix see next slide:

$$\left[\mathcal{I}_X(\boldsymbol{\theta}) \right]_{ij} = -E\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ \ln \ L(X|\boldsymbol{\theta}) \right] .$$

from $E(\partial \ln L)^2$ to $E(\partial^2 \ln L)$

Since $L(x1, x2 \ldots |\theta) = \prod_i f(x_i|\theta)$ is the joint density function of the data, it must be normalized:

$$\int_\Omega L\, dX = 1, \quad \text{so} \quad \int_\Omega \frac{\partial\ L}{\partial\theta}\, dX = 0$$

Multiply and divide by L:

$$\int_\Omega \left(\frac{1}{L}\frac{\partial\ L}{\partial\theta}\right) L\, dX \quad = \quad E\left(\frac{\partial\ \ln\ L}{\partial\theta}\right) \quad = \quad 0$$

Differentiate again, and again move ∂ into the $\int$:

$$\int_\Omega \left\{ \left(\frac{1}{L}\frac{\partial\ L}{\partial\theta}\right)\frac{\partial\ L}{\partial\theta} + L\, \frac{\partial}{\partial\theta}\left(\frac{1}{L}\frac{\partial\ L}{\partial\theta}\right)\right\} dX \quad = \quad 0$$

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

Fisher Information cont.

So the Fisher information in the sample X about the parameter(s) $\boldsymbol{\theta}$ is

$$\left[\underset{\sim}{I}_X \, (\boldsymbol{\theta}) \right]_{ij} = -E \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ \ln \ L(X|\boldsymbol{\theta}) \right] .$$

It can be seen that $\underset{\sim}{I}_X \, (\boldsymbol{\theta})$ has the additive property: If I_N is the information in N events, then $I_N(\theta) = N I_1(\theta)$.

We will also see that information about θ is related to the minimum variance possible for an estimator of *theta*.

But first we introduce the concept of Sufficient Statistics

Sufficiency

Any function of the data is called a statistic.

A sufficient statistic for θ is a function of the data that contains all the information about θ.

A statistic $T(X)$ is sufficient for θ if the conditional density function for X given T, $f(X|T)$ is independent of θ.

Sufficient statistics are clearly important for data reduction.

Cramér-Rao Inequality

Let the estimator $\hat{\theta}$ be an unbiassed estimator of θ with sampling distribution $q(\hat{\theta}|\theta)$.

Then the variance of the sampling distribution,

$$V(\hat{\theta}) = \int [\hat{\theta} - E(\hat{\theta})]^2 q(\hat{\theta}|\theta) d\hat{\theta} \, ,$$

is related to the information by the Cramér–Rao inequality:

$$V(\hat{\theta}) \geq \frac{1}{I_{\hat{\theta}}} = \frac{1}{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]} \, .$$

The Usual Estimators

The most common general-purpose estimators are:

- Maximum likelihood is the most important method, mostly because it can be shown to be **asymptotically efficient**.
- Least Squares is **asymptotically efficient** for fitting data in histograms, and is generally easier to apply than M.L.
- Binned or Poisson Likelihood for histograms is more efficient, less biased and less sensitive to binning than Least Squares.

Maximum Likelihood

The likelihood of a set of N independent observations $\mathbf{X}$ is

$$L(\mathbf{X}|\theta) = \prod_{i=1}^{N} f(X_i, \theta)\,,$$

where $f(X, \theta)$ is the p.d.f. of any observation X.
The maximum likelihood estimate of the parameter θ is that value $\hat{\theta}$ for which $L(\mathbf{X}|\theta)$ has its maximum, given the particular observations $\mathbf{X}$.
Note that maximizing $\ln L$ or L gives the same result.
The likelihood equation is

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{N} \ln\, f(X_i, \theta) = \frac{\partial}{\partial\theta} \ln\, L(\mathbf{X}, \theta) = 0\,.$$

since that is the analytic way to find the maximum, but in practice we will usually find the maximum numerically.

Asymptotic Properties of Maximum Likelihood

Asymptotically (for very large data samples), the M. L. estimator has optimal properties:

- It is consistent.
- It is efficient, the variance $V(\hat{\theta})$ being given by the Cramer–Rao lower bound
$$V(\hat{\theta}) \underset{N\to\infty}{\longrightarrow} \left\{ E\left[\left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right] \right\}^{-1} .$$
- The estimates $\hat{\theta}$ are Normally distributed.
- Since it is consistent, it is asymptotically unbiased.

Asymptotic Properties of Maximum Likelihood 2

If the range of the data is independent of the parameters θ, then the variance $V(\hat{\theta})$ may be estimated by

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

The estimate $\sqrt{N}(\hat{\theta} - \theta)$ is distributed as $N[0, I_1^{-1}(\theta)]$.
(Estimates are asymptotically Gaussian-distributed.)

We will give an example where the range of the data depends on the parameter, and the above properties do not hold..

Finite Sample Properties of Maximum Likelihood

- For finite samples, M.L. estimates are efficient only when there exist sufficient statistics for the parameter(s) being evaluated, and that can be shown only for the exponential family, consistent with the Darmois Theorem.
- Although the estimates are in general biased, they have a more important property, invariance, which is incompatible with unbiasedness because the definition of bias is not invariant.

Least Squares

Consider a set of observations $Y_1, \ldots, Y_N$ from a distribution with expectations $E(Y_i, \boldsymbol{\theta})$ and covariance matrix $\underset{\sim}{V}$. The $\boldsymbol{\theta}$ are unknown parameters and the $E(Y_i, \boldsymbol{\theta})$ and $V_{ij}(\theta)$ are known functions of $\boldsymbol{\theta}$.
In the method of least squares the estimates of the θ_k are those values $\hat{\theta}_k$ which minimize

$$\begin{aligned} Q^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N}[Y_i - E(Y_i, \boldsymbol{\theta})](\underset{\sim}{V}^{-1})_{ij}[Y_j - E(Y_j, \boldsymbol{\theta})] \\ &= [\mathbf{Y} - E(\mathbf{Y}, \boldsymbol{\theta})]^{\mathrm{T}} \, \underset{\sim}{V}^{-1} \, [\mathbf{Y} - E(\mathbf{Y}, \boldsymbol{\theta})] \, . \end{aligned}$$

Least Squares 2

When the observations Y_i are independent, it follows that they are uncorrelated, and the covariance matrix is diagonal, with elements

$$V_{ii} = \sigma_i^2(\boldsymbol{\theta})\,.$$

The covariance form then simplifies to the familiar sum of squares

$$Q^2 = \sum_{i=1}^{N} \frac{[Y_i - E(Y_i, \boldsymbol{\theta})]^2}{\sigma_i^2(\boldsymbol{\theta})}$$

The $\hat{\boldsymbol{\theta}}$ are found by solving the Normal equations

$$\partial Q^2 / \partial \boldsymbol{\theta} = 0\,,$$

Linear Least Squares

The method of linear least squares is applicable when the variances σ_i^2 are independent of the r parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, and the expectations $E(Y_i, \boldsymbol{\theta})$ are linear in the θ_j's,

$$E(Y_i, \boldsymbol{\theta}) = \sum_{j=1}^{r} a_{ij}\theta_j \,, \quad i = 1, \ldots, N$$

or in matrix notation

$$E(\mathbf{Y}, \boldsymbol{\theta}) = \underset{\sim}{A}\, \boldsymbol{\theta}$$

The elements a_{ij} of the *design matrix* $\underset{\sim}{A}$ are given by a model.

In the linear case, the solution of the Normal equations is

$$\widehat{\boldsymbol{\theta}} = (\underset{\sim}{A}^{\mathrm{T}} \underset{\sim}{V}^{-1} \underset{\sim}{A})^{-1}\, \underset{\sim}{A}^{\mathrm{T}} \underset{\sim}{V}^{-1}\, \mathbf{Y}\,.$$

Linear Least Squares cont.

Since the linear least squares solution is found by matrix inversion and multiplication (no minimization needed), one often solves the non-linear problem by linearization, setting:

$$a_{ij} = \frac{\partial E(Y_i, \theta)}{\partial \theta_j}$$

Example of linear least squares: fitting a curve to a polynomial.

$$Y_i = Y(X_i) = \theta_0 + \theta_1 X_i + \theta_2 X_i^2 + \theta_3 X_i^3$$

is clearly of the linear form. To find the matrix $\underset{\sim}{A}$ one only needs to evaluate the $(j-1)^{\mathrm{th}}$ power of X_i.
Solving the Normal equations $\partial Q^2 / \partial \boldsymbol{\theta} = 0$, we find:

$$\widehat{\boldsymbol{\theta}} = (\underset{\sim}{A}^{\mathrm{T}} \underset{\sim}{V}^{-1} \underset{\sim}{A})^{-1} \; \underset{\sim}{A}^{\mathrm{T}} \underset{\sim}{V}^{-1} \, \mathbf{Y} \, .$$

which is exact and unique as long as $\underset{\sim}{A}^{\mathrm{T}} \underset{\sim}{V}^{-1} \underset{\sim}{A}$ is non-singular.

Least Squares

For fitting data in histograms, the asymptotic properties of least squares are the same as for maximum likelihood, and in fact the two methods are often identical. When they are different, it is believed that M.L. generally approaches the asymptotic limit faster than L.S.

The biggest difference is largely practical.

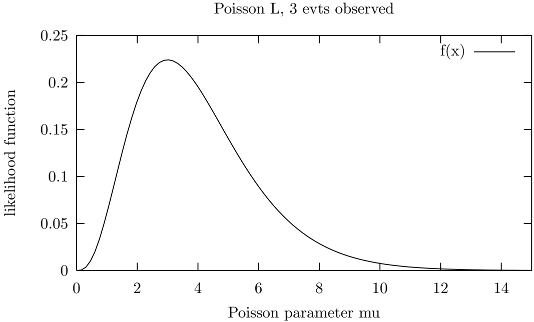
If the data are already grouped into bins or points, L.S. is more convenient and there is no advantage in using M.L.

The subject of M.L. and L.S. fitting will treated in more detail in the second half of the course.

Point Estimation: Example: Poisson data

Example: In a Poisson process, we observe 3 events.

$$\mathcal{L}(\mu) = P(3|\mu) = \frac{e^{-\mu}\mu^3}{3!}$$



The peak in the likelihood occurs at $\mu = 3$.
Generalizing from 3 to n, we get the expected result:
with n events observed, $\hat{\mu} = n$

Example: Weighted Average

Suppose we have Normally-distributed observations X_i of a quantity μ, each X_i being distributed with standard deviation σ_i:

$$f(X_i|\mu) = N(\mu, \sigma_i^2) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(X_i-\mu)^2}{\sigma_i^2}\right] .$$

We wish to use this data to estimate μ. The likelihood function is the product of the $f(X_i|\mu)$, and its logarithm is:

$$\ln \mathcal{L}(\mu) = k - \sum_i \frac{1}{2}\frac{(X_i-\mu)^2}{\sigma_i^2}$$

where k is a constant. It is clear that in this case, maximizing the log likelihood is equivalent to minimizing χ^2.
In both cases, the solution is the familiar weighted average:

$$\hat{\mu} = \frac{\sum_i \frac{X_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

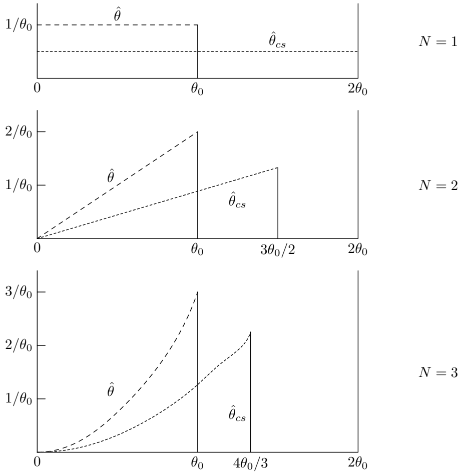
Example: A Poor M. L. Estimator

Suppose that one observes N events X_i chosen randomly from a uniform distribution between 0 and θ, where the upper bound θ is the unknown parameter.
This is a case where the range of the data depends on the value of the parameter θ.
Since $\theta \geq X_i$ for all i, the likelihood function $L = \theta^{-N}$ will have its maximum at $\hat{\theta} = X_{\mathsf{max}}$, where X_{max} is the largest observed value of X.
It is clear that this estimator (almost) always gives a result which is too small, and the obvious correction is to use the common-sense estimate:

$$\hat{\theta}_{\mathrm{cs}} = X_{\mathsf{max}} + \frac{X_{\mathsf{max}}}{N} \,.$$

This estimate in fact turns out to be unbiased, as can easily be verified.

Example: A Poor M. L. Estimator cont.



Distribution of maximum-likelihood estimates $\hat{\theta}$ and common-sense estimates $\hat{\theta}_{\text{cs}}$ for $N = 1, 2, 3$.

Robustness

Suppose we wish to estimate of the centre of an unknown, symmetric distribution. The centre of a distribution is defined by a location parameter. Some examples are:

- The mean is the expectation of the variable X.
- The median is that value X for which the cumulative distribution has $F(X) = 0.5$.
- The mode is that value of X for which the p.d.f. has a maximum.
- The midrange is defined when the possible values of X are limited to the range $[X_{\min}, X_{\max}]$. Then the *midrange* is $(X_{\min} + X_{\max})/2$.

Robustness cont.

For any particular sample of (finite) data X_i, we can define:

- The sample mean is the mean or average of the X_i.
- The sample median is the value X such that half the X_i lie above it and half below. If the number of data values is odd, it is the central value. If the number is even, it is usually taken as halfway between the two central values.
- The sample mode is the value of X halfway between the two nearest values of X_i.
- The sample midrange is halfway between the smallest and largest values of X_i, that is $(X_{i\,\mathrm{min}} + X_{i\,\mathrm{max}})/2$.

Robustness 3

The sample mean is the most obvious and most often used estimator of location, because

- it is consistent whenever the variance of the underlying distribution is finite (law of large numbers);
- it is optimal (minimum variance, unbiased) when the underlying distribution is Normal.

However, if the distribution of X is not Normal, the sample mean is not the best estimator of the mean of the distribution, even when the mean of the distribution exists. Below we list the best estimator of location for some important distributions:

Distribution	Minimum-variance location estimator
Normal	sample mean
Uniform	midrange (mean of extreme values)
Cauchy	maximum-likelihood estimate
Double-exponential	median (middle value)

Robustness 4

A robust estimator is one which, although not optimally efficient for any one distribution, has a high efficiency over a broad range of distributions. Let us define:

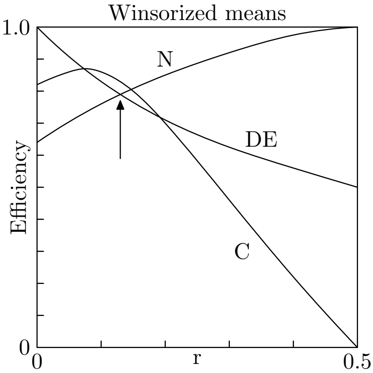
- The trimmed mean of the total of N observations: remove the $n/2$ highest values and the $n/2$ lowest values, and compute the mean of the remaining $N - n$ observations.
- The Winsorized mean of the total of N observations: replace the $n/2$ highest values by the highest remaining value, and the $n/2$ lowest by the lowest remaining value, and compute the mean of the new sample of N values.

For both estimators there is one free parameter, usually taken as half the fraction of remaining (unchanged, or not rejected) values,

$$r = \frac{N - n}{2N} .$$

Note that for $r = 0.5$, both estimators are in fact the sample mean, and as $r \to 0$, both become equivalent to the sample median.

Robustness 5



Asymptotic efficiencies of trimmed and Winsorized means for Normal (N), double-exponential (DE) and Cauchy (C) distributions. Arrows indicate minimax point for optimum robustness.

N-dependence of errors

Physicists distinguish between two sources of error:
statistical errors and systematic errors.

Statisticians distinguish two different sources of error:
The bias and variance of an estimator.

We will distinguish three different sources of experimental error, each with its own dependence on the number of observations n.

- The systematic error usually does not decrease with n.
- The bias of an asymptotically unbiased estimator typically decreases like n^{-1}.
- The statistical error, or square root of the variance of an estimator, typically decreases like $n^{-1/2}$, but there are exceptions: (e.g. midrange in slide 46.)

Why least squares?

1. When the data are Gaussian-distributed :

   Maximum Likelihood reduces to Least Squares.

   But Least Squares is older than M.L. So why was it used? Probably because:

   - In the Decision-Theoretic Approach, it follows from a quadratic penalty or cost function.
   - When the model is linear, the solution is a linear function of the data.

2. So what happens if we try minimizing the sums of different powers of the residuals?

Alternatives to Least Squares

Should one sometimes use least cubes, least absolute values, etc.?

We call a_p the L_p estimate of location if a_p minimizes the quantity

$$L_p(a) = \sum_{i=1}^{N} |X_i - a|^p \, .$$

For all $p \geq 1$, the L_p estimate is well-defined, and the properties of these estimates have been studied [Rice and White, 1964].

1. For $p = 1$, a_p is the sample median.
2. For $p = 2$, a_p is the sample mean, the least-squares estimator.
3. For $p = \infty$, a_p is the sample midrange, the average of the lowest and highest values. L_∞ is the Chebyshev Norm.
4. As $p \rightarrow -\infty$, a_p tends to the sample mode, the point of highest density, corresponding to the maximum of the pdf.

So small (or negative) values of p give more weight to points near the middle of the distribution,
and large values give more weight to the tails.

Why least squares?

Asymptotic variances of L_p location estimates.
for some symmetric distributions

	L_1	L_2	L_∞
Distribution	(median)	(mean)	(midrange)
Uniform	$1/4N$	$1/12N$	$1/(2N^2+6N+4)$
Triangular		$1/6N$	$(4-\pi)/4N$
Normal	$\pi/2N$	$1/N$	$\pi^2/12\ \log\ N$
Double-exponential	$1/2N$	$2/N$	$\pi^2/12$
Cauchy	$\pi^2/4N$	∞	∞

This means that when fitting a set of points to a hypothesis (for example, fitting a track to measurements in a detector) the usual least-squares estimator, based on the L_2 Norm, is optimal only if the measurements are Normally distributed. For some detectors, another L_p norm may be more efficient.

Examples where Least Squares is not Optimal

1. Fitting Points to a curve
when the point measurements are not Normally distributed.

When the distribution of measurements has longer tails than a Gaussian, use the L_p norm with $p < 2$.

The opposite extreme (distributions with no tails at all) may be attained with some detectors based on discrete elements (strips or wires) such that a hit defines a window through which the track must pass.
In this case, the Chebyshev Norm may provide much superior accuracy compared with least squares.

See F. James, *Fitting Tracks in Wire Chambers using the Chebyshev Norm instead of Least Squares*, NIM 211(1983)145

2nd Example where Least Squares is not Optimal

2. Fitting data to a Histogram
when there are not many events in some bins.

Then the Poisson distribution of events in each bin is not approximately Gaussian, and it is better to use the binned likelihood.

See Baker and Cousins, *Clarification of the Use of Chi-square and Likelihood Functions in Fits to Histograms* NIM 221 (1984) 437

2nd Example where Least Squares is not Optimal

from: Baker and Cousins, *Clarification of the Use of Chi-square and Likelihood Functions in Fits to Histograms* NIM 221 (1984) 437

The likelihood function for the Poisson-distributed histogram contents is

$$L(\theta) = \prod_i e^{-\mu_i(\theta)} \mu_i^{n_i}(\theta)/n_i!$$

where $\mu_i(\theta)$ is the content of the i^{th} bin predicted by the model, and n_i is the observed contents of the bin.
It is convenient to work with the likelihood ratio λ which is the above likelihood divided by the likelihood for data without errors. Then the quantity $-2\ln\lambda$ asymptotically obeys a Chi-square distribution, and the quantity to be minimized reduces to

$$\chi^2_\lambda = -2\ln\lambda = 2\sum_i [\mu_i(\theta) - n_i + n_i \ln(n_i/\mu_i(\theta))]$$