

Statistics

Lecture 1

August 4, 2000

Frank Porter

Caltech

The plan for these lectures:

The Fundamentals; Point Estimation

Maximum Likelihood, Least Squares and All That

What is a Confidence Interval?

Interval Estimation

Monte Carlo Methods

Additional topics will be covered by Roger Barlow and Norman Graf

Probability

Familiar concept, but give definition for the record:

Probability: A **probability**, $P(E)$, is a real additive set function defined on sets E in sample space S satisfying the properties:

- 1) If E is a subset (event) in S , then $P(E) \geq 0$.
- 2) $P(S) = 1$.
- 3) $P(E_1 \cup E_2 \cup \dots) = P(E_1) + P(E_2) + \dots$ for any sequence (finite or infinite) of disjoint events E_1, E_2, \dots in S .

Or, for short:

Probability: A **probability** is a measure in which the measure of the entire sample space is equal to one.

We give probability a “physical meaning” in terms of the **Frequency Interpretation**: If we draw an element from our sample space S many times, we will obtain event E in a fraction $P(E)$ of the samplings.

Exercises

Rule of Complementation: Prove:

$$P\left(\bigcap_{i=1}^n \tilde{E}_i\right) = 1 - P\left(\bigcup_{i=1}^n E_i\right).$$

Application: What is the probability of losing at least one track in a six track event, if the single-track inefficiency is 5%? State any assumptions.

Arbitrary union: Prove:

$$\begin{aligned} P\left(\bigcup_{i=1}^n E_i\right) &= \sum_{i=1}^n P(E_i) - \sum_{j>i}^n P(E_i \cap E_j) \\ &\quad + \sum_{k>j>i}^n P(E_i \cap E_j \cap E_k) \\ &\quad \dots \\ &\quad + (-)^{n-1} P(E_1 \cap E_2 \dots \cap E_n). \end{aligned}$$

Random Variables

We find it useful to map abstract sample spaces into real numbers:

Random Variable: A **Random Variable** (RV) is a variable that takes on a distinct value for each element of the sample space.

A random variable may vary over a discrete or continuous spectrum (or a combination).

Probability Functions

If x is a discrete RV, we say that $p(x) \equiv P[E(x)]$ is the probability of x , where $E(x)$ is the inverse mapping of x onto the sample space and we have:

$$\sum_{\text{all } x} p(x) = 1.$$

If x is a continuous RV, we take the appropriate continuum limit of the above notion, with

$$p(x)dx \equiv P[E(y : x \leq y < x + dx)]$$

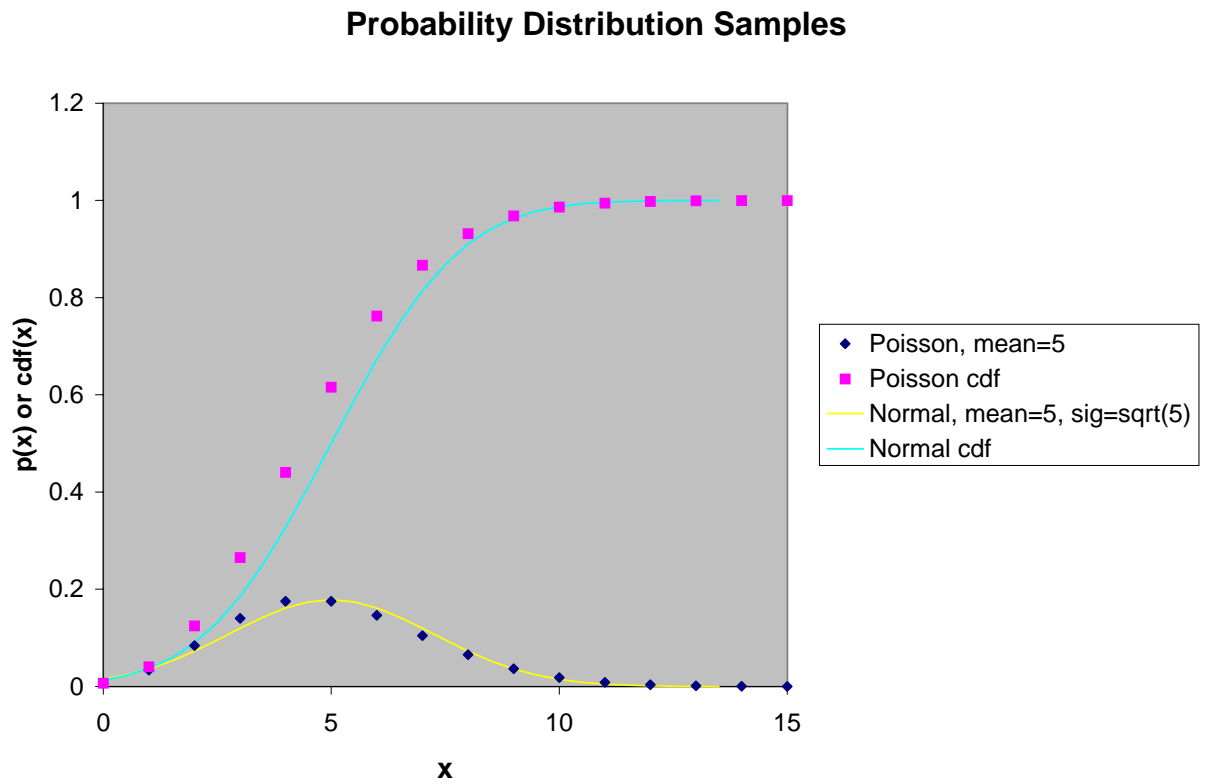
(in the limit). Here $p(x)$ is called a “probability density function” (pdf).

The “cumulative distribution function” (cdf) for RV x is the probability of not exceeding a value x .

Note: I will be sloppy in the distinction between a random variable and a particular value in the space of RVs.

Exercise: Turn all my statements into respectable mathematics.

Sample of Discrete and Continuous Probabilities



Joint Distributions

A “**joint probability distribution**” is one in which the abstract sample space has been mapped into a multidimensional RV space (natural if the sample space is describable as a product space). In this case we often collect the RVs into a vector \mathbf{x} .

Suppose we have a joint pdf, $p(x, y)$, in random variables x, y , and let:

$$q(x) \equiv \int_{-\infty}^{\infty} p(x, y) dy$$
$$r(y) \equiv \int_{-\infty}^{\infty} p(x, y) dx.$$

We define a **Conditional Probability**, $s(x|y)$ or $t(y|x)$, according to:

$$p(x, y) = s(x|y)r(y)$$
$$= t(y|x)q(x).$$

We read $s(x|y)$ as telling us the “probability of x , given y .”

Bayes' Theorem

We have, e.g.,

$$\begin{aligned} s(x|y) &= \frac{p(x, y)}{r(y)} \\ &= \frac{t(y|x)q(x)}{\int_{-\infty}^{\infty} p(x, y)dx}. \end{aligned}$$

This important result in probability theory is known as **Bayes' Theorem**.

It is used in a fundamental way in “Bayesian statistics”.

Application – Exercise: There are two radioactive sources. One emits gamma rays in 75% of the decays, and beta rays in the other 25%. The other emits gammas 1/3 of the time, and betas 2/3. A source is chosen at random, and the first decay observed is a gamma. What is the probability that a gamma will be observed on the second decay? How about the 83rd decay?

Statistical Independence

Independence: Two random variables, x and y , are **statistically independent** iff:

$$p(x, y) = q(x)r(y).$$

Expectation Values

The **Expectation Value** of a function, f , of a random variable x , is defined by:

$$\langle f(x) \rangle = \int_{\text{all } x} f(x)p(x)dx,$$

with the obvious generalization to joint pdfs.

Theorem: (Independence) If x and y are two statistically independent RVs, then

$$\langle f(x)g(y) \rangle = \langle f(x) \rangle \langle g(y) \rangle.$$

Proof: Trivial.

Mean and Variance

Mean: The **mean** of a random variable is its expectation value.

Variance: The **variance** of a random variable x is the square of the **standard deviation**, and is the expectation value:

$$\begin{aligned}\text{var}(x) &= \sigma_x^2 = \langle (x - \langle x \rangle)^2 \rangle \\ &= \langle x^2 \rangle - \langle x \rangle^2.\end{aligned}$$

The variance generalizes in the multivariate case to the **Moment Matrix**, with elements:

$$\begin{aligned}M_{ij} &= \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \\ &= \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle.\end{aligned}$$

Note that the diagonal elements are simply the individual variances. The off-diagonal elements are called **covariances**.

The **covariance coefficients**, measuring the degree of linear correlation, are given by:

$$\rho_{ij} = \frac{M_{ij}}{\sqrt{M_{ii}M_{jj}}}.$$

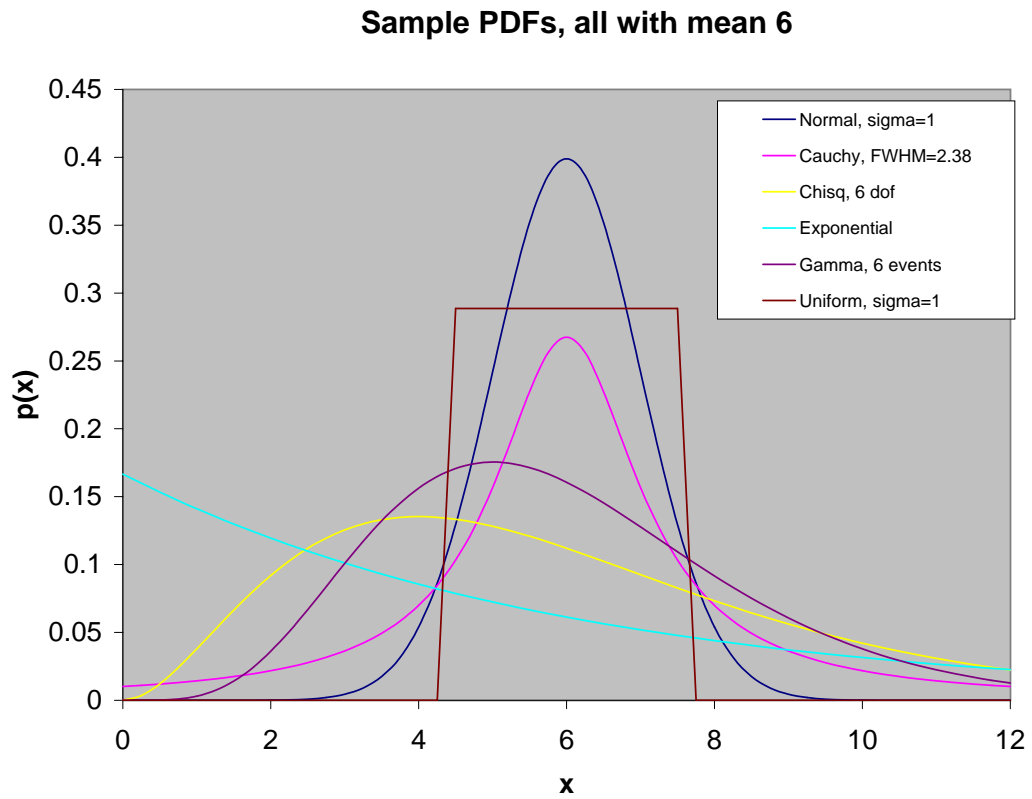
Some Important Probability Distributions

Name	space	$p(x)$	$\langle x \rangle$	$\sigma(x)$
Binomial	$\{0, 1, \dots, n\}$	$\binom{n}{x} \theta^x (1 - \theta)^{n-x}$	$n\theta$	$\sqrt{n\theta(1 - \theta)}$
Poisson	$\{0, 1, 2, \dots\}$	$\theta^x e^{-\theta} / x!$	θ	$\sqrt{\theta}$
Cauchy ⁽¹⁾	real numbers	$\frac{2}{\pi\Gamma} \frac{1}{1+4(x-\theta)^2/\Gamma^2}$	θ	∞
Chisquare	positive real	$\frac{e^{-x/2} x^{n/2-1}}{\Gamma(n/2) 2^{n/2}}$	n	$\sqrt{2n}$
Exponential	positive real	$\theta e^{-\theta x}$	$1/\theta$	$1/\theta$
Gamma ⁽²⁾	positive real	$\frac{\theta^n x^{n-1} e^{-\theta x}}{\Gamma(n)}$	n/θ	\sqrt{n}/θ
Normal	real numbers	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$	θ	σ
Uniform	reals in $\{0, \theta\}$	$1/\theta$	$\theta/2$	$\theta/\sqrt{12}$

(1) Also known as Breit-Wigner.

(2) x = time to observe n events in a Poisson process.

Comparison of Common PDFs



Transformations

We are often interested in the probability distribution for quantities $\mathbf{y} = (y_1, y_2, \dots, y_n) = \mathbf{f}(\mathbf{x})$, given the probability distribution for the (perhaps measured) quantities $\mathbf{x} = (x_1, x_2, \dots, x_n)$. If the y 's are linearly independent, the new pdf for \mathbf{y} is simply found by:

$$\begin{aligned} q(\mathbf{y})d^n(\mathbf{y}) &= q[\mathbf{f}(\mathbf{x})] \left| \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right| d^n(\mathbf{x}) \\ &= p(\mathbf{x})d^n(\mathbf{x}). \end{aligned}$$

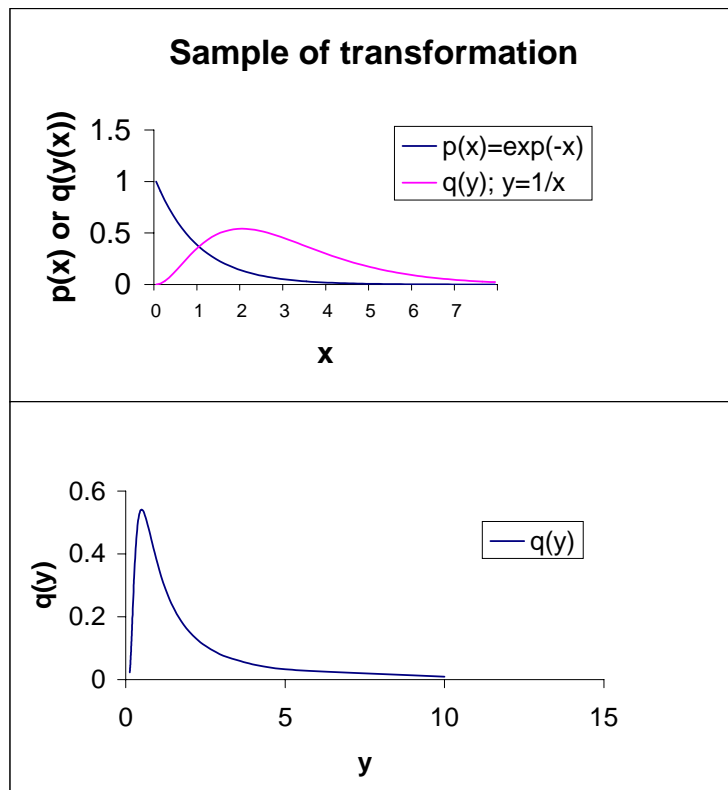
Hence,

$$q(\mathbf{y}) = \frac{p[\mathbf{f}^{-1}(\mathbf{y})]}{\left| \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right| [\mathbf{f}^{-1}(\mathbf{y})]}.$$

Exercise: If $p(x) = \theta e^{-x/\theta}$, what is the pdf for $y = 1/x$?

Rather than determining the entire transformation, we are often content to learn the new moment matrix.

Transformation Example



Propagation of Errors

If $\mathbf{y} = (y_1, y_2, \dots, y_k)$ is linearly dependent on $\mathbf{x} = (x_1, x_2, \dots, x_n)$,
i.e.,

$$\mathbf{y} = T\mathbf{x} + \mathbf{a},$$

where T is a $k \times n$ transformation matrix, then it is easily shown that the moment matrix for \mathbf{y} is given by:

$$M_y = TM_x T^\dagger.$$

If \mathbf{y} is non-linearly dependent on \mathbf{x} , we often make the linear approximation anyway, letting

$$T_{ij} = \left. \frac{\partial y_i}{\partial x_j} \right|_{\mathbf{x} \sim \langle \mathbf{x} \rangle}.$$

It should be kept in mind though, that this corresponds to taking the first term in a Taylor series expansion, and may not be a good approximation for some transformations, or far away from $\langle \mathbf{x} \rangle$.

Propagation of Errors - Special Case

Example: Suppose $k = 1$. Then, in the linear approximation:

$$\begin{aligned} M_y &= \sigma_y^2 = T M_x T^\dagger \\ &= \sum_{i=1}^n \sum_{j=1}^n \frac{\partial y}{\partial x_i} \Big|_{\mathbf{x} \sim \langle \mathbf{x} \rangle} \frac{\partial y}{\partial x_j} \Big|_{\mathbf{x} \sim \langle \mathbf{x} \rangle} (M_x)_{ij}. \end{aligned}$$

If the x_i 's are statistically independent, then

$$(M_x)_{ij} = \sigma_{x_i}^2 \delta_{ij},$$

and hence,

$$\sigma_y^2 = \sum_{i=1}^n \left(\frac{\partial y}{\partial x_i} \Big|_{\mathbf{x} \sim \langle \mathbf{x} \rangle} \right)^2 \sigma_{x_i}^2.$$

This is our most commonly-used form for propagating errors.

Just remember the assumptions of linearity and independence,

as well as the typically approximate knowledge of $\langle \mathbf{x} \rangle$!

Characteristic Functions

An important tool in probability theory is the **characteristic function** (or the related moment generating function), which is simply the Fourier transform of the pdf:

$$\begin{aligned}\phi_x(k) &= \int_{-\infty}^{\infty} e^{ikx} p(x) dx \\ &= \langle e^{ikx} \rangle.\end{aligned}$$

Theorem: (Uniqueness) If two random variables have the same characteristic function, then the cdf's are identical.

Proof: Exercise. (Hint: Consider a step function of x , and show that the derivative of the characteristic function of this step function, evaluated at $k = 0$, is related to the cdf.)

The pdf's may not be identical, but it doesn't matter, since they differ only on a set of measure zero.

Central Limit Theorem

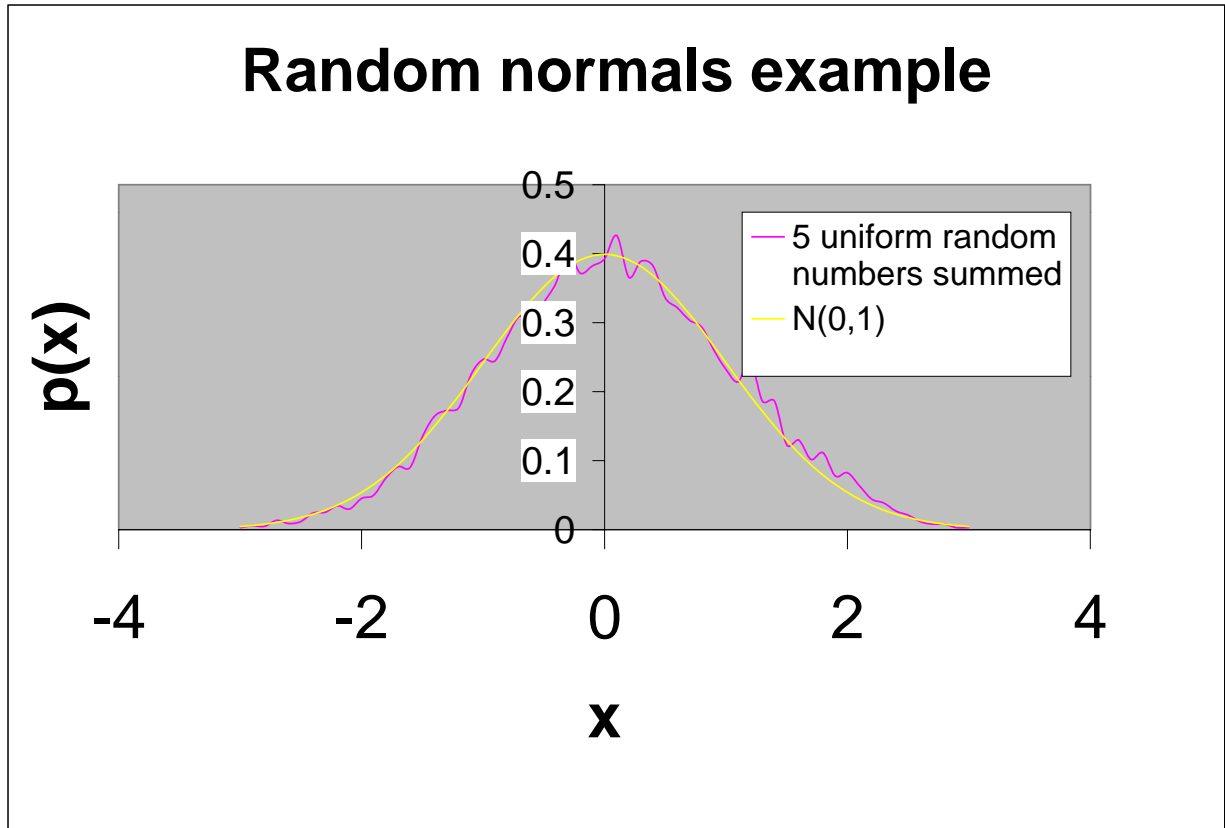
We are ready now for the celebrated:

Theorem: (Central Limit Theorem) Let (x_1, x_2, \dots, x_n) be a random sample of size n from a distribution having mean μ and variance σ^2 . Then, if $S/n = \frac{1}{n} \sum_{i=1}^n x_i$ is the **sample mean**, the distribution of S/n approaches the normal distribution as $n \rightarrow \infty$, with mean $\langle S/n \rangle = \mu$ and variance $\langle (S/n - \langle S/n \rangle)^2 \rangle = \sigma^2/n$.

Proof: Exercise. (Hint: Show that the characteristic function approaches the characteristic function of the normal distribution in this large n limit.)

Exercise: Generation of Normal Distribution: Suppose you have a source of uniformly distributed random numbers on the interval $(0, 1)$. Using the Central Limit Theorem, give an algorithm to produce random numbers which are approximately normally distributed, with mean 0 and variance 1 .

Example – Application of Central Limit Theorem



Statistics

Statistics: (Webster) “a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data”.

At first sight, a vague definition; after more thought, an excellent definition!

Statistics is not a single-purpose field. You (yes, you!) must decide what you want to accomplish **BEFORE** you apply a statistical procedure!

Goals: IMPORTANT!

I want to concentrate on two major, distinct, goals that we physicists have in quoting results:

Information: We wish to summarize the information content of a set of measurements.

Belief: We wish to summarize our “degree of belief” concerning the truth of some physical statement, incorporating available information.

Don't expect to accomplish both goals simultaneously!

Elaboration: Information

All of the information from a measurement is specified by giving the result of the measurement along with everything that is knowable about the sampling distribution (i.e., about the apparatus and measurement procedure).

Example: In a counting experiment, we observe n events. This is the result of the measurement. We also require the sampling distribution. For example, perhaps it is Poisson, depending on an unknown signal parameter with calculable efficiency, and on some determinable background rate:

$$p(n) = \frac{(\theta_s + \theta_b)^n e^{-(\theta_s + \theta_b)}}{n!}.$$

Elaboration: Belief

Ultimately, interested in the truth of physical statements, e.g., concerning the value of some fundamental parameter.

Or, may be planning an experiment, where “success” depends on some imperfectly known quantity (e.g., SUSY mass scale).

In these cases, we are faced with the need to make a “decision”. It is not enough to provide information.

To make a decision, we may make use of whatever experimental information is available, and any theoretical constraints and possibly prejudices.

Poisson example again: We may believe, e.g., that $\theta_s > 0$, referring to the previous slide. That is, our description above is correct in its implication that the signal and background processes add incoherently. To decide where we think θ_s probably lies, we wish to fold this constraint into the analysis. **This isn't as trivial an idea as it may seem! – How do you define “Belief”?**

Point Estimation

In the context of an experiment to measure the value of some unknown parameter (e.g., the mass of a particle), we would like to quote some sort of “best” estimate for that parameter, given the experimental result. This is the problem of **Point Estimation**.

What does “best” mean? Several possible criteria:

Consistent Unbiased
Efficient Sufficient
Robustness “Physical”
Tractable

Note that, as a function of random variable(s), our parameter estimator is itself a random variable, drawn from some probability distribution.

Notation: We’ll use a “hat” over a symbol for a parameter to indicate an estimator for the parameter instead of the parameter itself.

Consistency

Consistent: An estimator, $\hat{\theta}$, is **consistent** if

$$\lim_{n \rightarrow \infty} \hat{\theta}(x_1, x_2, \dots, x_n) = \theta.$$

That is, an estimator is **consistent** if it converges to the parameter in the limit of large statistics.

Note the distinction between consistency, and bias, next.

Example: The quantity

$$s = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2,$$

where $m = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean, is a consistent estimator for the variance of a normal distribution, with samples $\{x_1, x_2, \dots, x_n\}$.

Bias

Bias: Given an estimator, $\hat{\theta}$, for a parameter θ , the **bias**, $b(\theta)$, of the estimator is:

$$b(\theta) = \langle \hat{\theta} \rangle - \theta.$$

Note that bias is often thought of as a “systematic error”.

Example: The quantity s defined in the previous slide is a biased estimator for the variance, with $b(\theta) = -\sigma^2/n$. We may remove the bias in this case simply by multiplying by $n/(n - 1)$.

Efficiency

An estimator, $\hat{\theta}_a$, is said to be more efficient than another estimator $\hat{\theta}_b$, if its variance is smaller.

Note that the goal of good efficiency, by itself, is readily achieved with useless estimators. For example, we spend millions of dollars to measure $\sin 2\beta$. The estimator we use has a non-zero variance, improving as the data size increases. However, we could avoid all this if we only want an efficient estimator. Forget the experiment, and use $\sin 2\hat{\beta} = 0.5$. You can't get more efficient than zero variance!

Next time we'll give a useful theorem for how well we can do.

Example: As we'll be able to prove later, the sample mean is an optimally efficient unbiased estimator for the mean of a normal distribution.

Sufficient Statistics

Sufficient: A statistic $S = S(\mathbf{x})$ is **sufficient** for parameter θ if the conditional probability for \mathbf{x} , given a value of S , is independent of θ :

$$\frac{\partial p(\mathbf{x}|S)}{\partial \theta} = 0.$$

Intuition: A sufficient statistic contains all of the information in the data concerning the parameter of interest. Once S is specified, there is no additional information in the \mathbf{x} concerning θ .

Exercise: Show that the sample mean is a sufficient statistic for the true mean of a normal distribution of known variance.

Robust Statistics

In general, we don't know exactly the probability distribution from which we are sampling when we do an experiment. In particular, there are often extended "tails" above our approximate forms (e.g., non-Gaussian tails on an approximately Gaussian distribution).

A robust statistic is one which is relatively insensitive to the existence of these tails.

Example: The median of a distribution is typically a more robust estimator for a location parameter than the mean.

Aside: Location Parameter

Location parameter: If the probability density function is of the form:

$$p(x; \theta) = p(x - \theta),$$

then θ is called a **location parameter** for x .

Example: In the normal distribution:

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

θ is a location parameter for x .

“Physical Statistics”

Here, I simply mean that it may be desirable to have an estimator which is guaranteed to be in some restricted range, corresponding to theoretically allowed values for the parameter of interest.

Note that, if all you are trying to do is to summarize the information content of a measurement, as opposed to making some statement about the true value of a parameter, this is not an interesting property to require.

Tractableness

Theoretically unimportant, but crucial in practice! We are willing to sacrifice other goals to get an answer at all, as long as we can get something “good enough”.

Next Time

We'll consider three common methods for parameter estimation, and discuss their properties in terms of the above “desirable features”.

We'll also give the Rao-Cramer-Frechet theorem for efficiency, and illustrate its application.