BABAR Analysis Document #318, Version 1

Recommended Statistical Procedures for *BABAR*

Roger Barlow, Robert Cahn, Glen Cowan, Francesca Di Lodovico, William Ford, Gautier Hamel de Monchenault, David Hitlin, David Kirkby, Francois Le Diberder, Gerry Lynch, Frank Porter, Soeren Prell, Art Snyder, Mike Sokoloff, Roland Waldi

May 7, 2002

Contents

Chapter 1

Overview

1.1 Purpose

The *BABAR* Statistics Working Group was formed in order to address questions of statistics that come up in the analysis and presentation of results from the experiment. The idea is to prepare recommended statistical methodologies in response to these questions. This report is the result.

The report is organized into two major parts: A pedagogical part which provides the language and background of statistics, and a recommendations part which gives the recommended procedures from the working group. The charge to the working group is included as Appendix A.

In addition to this report, there is a Statistics Working Group hypernews forum for discussions, and a web page. The hypernews forum is:

http://babar-hn.slac.stanford.edu:5090/HyperNews/get/Statistics.html

The Statistics Working Group home page is at URL:

http://www.slac.stanford.edu/BFROOT/www/Statistics/

From that page, a "Statistics Bibliography" may be reached, which includes links to other discussions within HEP, as well as more general references.

Part I

Pedagogy

Chapter 2

Probability

2.1 Probability

Probability is an axiomatic theory in mathematics defined by three axioms.
Probability: A **probability**, $P(E)$, is a real additive set function defined on sets E in sample space S satisfying the properties (axioms):

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

Each real-world quantity that fulfills these axioms can be regarded as a probability. There are in fact two such use-cases, which are related, but differ and should be kept in mind as separate entities:

One "physical meaning" is given to probability 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. It is generally only in our minds that we do this repitition, so this is a conceptual description in typical practice, although in the context of frequentist statistics a more concrete description can be taken.

There is another common interpretation, which applies to a single event, when an element is drawn once and cannot be repeated many times. This interpretation is often called "subjective probability" and is associated with the "degree of belief" in event E. It summarises the expectation on the unknown occurrence of this event, and is likewise expressed in terms of a probability.

Rolling dice can be used as an example for both representations.

The “frequency” probability is in the statement: “the probablilty that a rolled die shows 6 is 1/6”.

The “subjective” probability is in the statement: “the probablilty that *this* rolled die will show 6 is 1/6”.

The difference in the two statements is, that the first talks about any die, *i.e.*, about a large sample of rolled dice, where indeed the frequency converges to 1/6 (if the die is unbiased). The second statement talks about one single event. Here, the frequency is in fact either 100% or 0. But it is not (yet) known which of the two is true, hence one can only express a “degree of belief” about what will happen, or has already happened but is unknown.

2.2 Probabilities and PDFs

Theorem: (**Rule of Complementation**)

$$P(\cap_{i=1}^{n} \widetilde{E}_i) = 1 - P(\cup_{i=1}^{n} E_i).$$

Theorem: (**Arbitrary union**)

$$\begin{aligned} P(\cup_{i=1}^{n} E_i) &= \sum_{i=1}^{n} P(E_i) - \sum_{j>i}^{n} P(E_i \cap E_j) && (2.1) \\ &+ \sum_{k>j>i}^{n} P(E_i \cap E_j \cap E_k) && (2.2) \\ &\cdots && (2.3) \\ &+(-)^{n-1} P(E_1 \cap E_2 \cdots \cap E_n). && (2.4) \end{aligned}$$

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).

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.

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[1]:

$$q(x) \equiv \int_{-\infty}^{\infty} p(x, y)dy \tag{2.5}$$

$$r(y) \equiv \int_{-\infty}^{\infty} p(x, y)dx. \tag{2.6}$$

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

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

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

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

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:

$$\text{var}(x) = \sigma_x^2 = \langle (x - \langle x\rangle)^2\rangle \tag{2.7}$$

$$= \langle x^2\rangle - \langle x\rangle^2. \tag{2.8}$$

[1]For simplicity, we'll often treat our random variables as continuous, but the generalization to discrete RVs is not difficult

The variance generalizes in the multivariate case to the **Covariance Matrix** (alternatively known as the **Moment Matrix**, the **Error Matrix**, the **Variance Matrix**, or the **Dispersion Matrix**) with elements:

$$\begin{aligned} M_{ij} &= \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle & (2.9) \\ &= \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle. & (2.10) \end{aligned}$$

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

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

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

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

$$\begin{aligned} q(\mathbf{y}) d^n(\mathbf{y}) &= q[\mathbf{f}(\mathbf{x})] \left| \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right| d^n(\mathbf{x}) & (2.11) \\ &= p(\mathbf{x}) d^n(\mathbf{x}). & (2.12) \end{aligned}$$

Hence,

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

Rather than determining the entire transformation, we are often content to learn the new moment matrix. If $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ is linearly dependent on $\mathbf{x} = (x_1, x_2, \ldots x_n)$, *i.e.*,

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

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

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

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 $\langle \mathbf{x} \rangle$.
Example: Suppose $k = 1$. Then, in the linear approximation:

$$M_y = \sigma_y^2 = TM_xT^\dagger \tag{2.13}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \left.\frac{\partial y}{\partial x_i}\right|_{\mathbf{x}\sim\langle\mathbf{x}\rangle} \left.\frac{\partial y}{\partial x_j}\right|_{\mathbf{x}\sim\langle\mathbf{x}\rangle} (M_x)_{ij}. \tag{2.14}$$

If the x_i's are statistically independent, then

$$(M_x)_{ij} = \sigma_{x_i}^2 \delta_{ij}, \tag{2.15}$$

and hence,

$$\sigma_y^2 = \sum_{i=1}^{n} \left(\left.\frac{\partial y}{\partial x_i}\right|_{\mathbf{x}\sim\langle\mathbf{x}\rangle} \right)^2 \sigma_{x_i}^2. \tag{2.16}$$

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 $\langle \mathbf{x} \rangle$!

2.3 Bayes Theorem

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

$$p(x, y) = s(x|y)r(y) \tag{2.17}$$

$$= t(y|x)q(x). \tag{2.18}$$

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

We have, *e.g.*,

$$s(x|y) = \frac{p(x, y)}{r(y)} \tag{2.19}$$

$$= \frac{t(y|x)q(x)}{\int_{-\infty}^{\infty} p(x, y)dx}. \tag{2.20}$$

This important result in probability theory is known as **Bayes' Theorem**. It is used in a fundamental way in "Bayesian statistics".

2.4 Central Limit Theorem

Theorem: (**Central Limit Theorem**) Let $(x_1, x_2, \ldots, x_n)$ be a set of n independent random variables (*e.g.*, measurement results) from in general different distributions with means $(\mu_1, \mu_2, \ldots, \mu_n)$ and (finite) variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$. Then, if $S = \sum_{i=1}^n x_i$ is the sum of these numbers, the distribution of S approaches the normal distribution as $n \to \infty$, with mean $\langle S \rangle = \sum_{i=1}^n \mu_i$ and variance $\langle (S - \langle S \rangle)^2 \rangle = \sum_{i=1}^n \sigma_i^2$.

A special case is the **sample mean** $S/n = \frac{1}{n} \sum_{i=1}^n x_i$ of a sample of values from **one** distribution with mean μ and variance σ^2, where $\langle S/n \rangle = \mu$ and $\langle (S/n - \langle S/n \rangle)^2 \rangle = \sigma^2/n$.

We make frequent use of this theorem in statistical analysis. The normal (or "Gaussian") distribution is well-understood and is "well-behaved". While we often sample from non-normal distributions, the properties of the normal distribution are so desirable that the approximation by a normal distribution is often used. The underlying justification for this is the central limit theorem. But it is important to understand that it is an approximation, and not to push it outside its realm of validity. The expectations for the sample mean and variance, however, are exact results even for non-Gaussian distributions, as long as those quantities exist. Hence, for example, "errors" can be added quadratically even if the distributions are not Gaussian.

Chapter 3

Basic Statistical Notions

The previous chapter summarized the elements of probability theory of relevance to the purposes of this report. We turn now to the subject of "statistics", having to do, for present purposes, with the extraction of information from a random sampling process. This is sometimes referred to as the "inverse probability problem": Instead of addressing the sampling of a random variable from some specified probability distribution, we imagine that we are interested in learning something about a (not completely specified) probability distribution by taking random samples from it. In this chapter are collected several notions basic to the practice of statistics. Here are defined also some of the basic terms in common use.

3.1 Goals of Statistics: Bayesian and Frequentist Statistics

Because of the substantial confusion among particle physicists concerning the subjects of "Bayesian" and "frequentist" (or "classical"[1]) statistics, a brief discussion is appropriate.

An intuitive way to think about these two branches of statistics is to regard the frequentist approach as being directed at summarizing relevant information content in a dataset; whereas the goal of the Bayesian is to infer something about the underlying sampling distribution. Thus, it is useful to

[1]Beware that some authors use "classical" to refer to Bayesian statistics. After all, it preceded frequentist statistics!

connect the notion of "information" with the frequentist, and the notion of "decision" with the Bayesian. Of course, like any good decision-maker, the Bayesian uses whatever information is available, but the difference is that the Bayesian doesn't stop there. Instead, s/he proceeds to a conclusion corresponding to a "degree-of-belief" concerning the "truth" of some statement.

It may be argued that the first goal of an experimenter is to summarize the information contained in the data, as objectively as possible. The (frequentist) approach to achieving this goal consists of making some statement which has the property that it would be true in some specified fraction of trials (repetitions of the experiment), and false in the remaining fraction. It is important to realize that the frequentist typically doesn't know *or care* what the truth value is for any given sampling.

As the experimenter happens to also be a physicist, a second goal is to summarize what s/he thinks the "truth" really is, based on the information in the experiment, and perhaps other input. This is the domain of Bayesian statistics. While the first goal should be a requirement in any publication of results, the second is optional – it can be left to the readers to form their own conclusions concerning the physical implications of the measurement, but the reader can't divine the information, so that must be presented.

It may be noted that people have sometimes proposed methodologies which attempt to satisfy Bayesian urges (*e.g.*, never make a statment that is known to be in the "false" category), while maintaining frequentist validity. This may be a noble goal, but the resulting methodologies are not especially attractive for various reasons, and such algorithms are not advocated here. Instead, it is simply admitted that there is more than one goal, and that different techniques may be optimal for each.

A more expansive discussion of this distinction between the classical (information theory) and the Bayesian (decision theory) statistical methodologies may be found in a paper by F. James and M. Roos[1]. They address in particular the example of neutrino mass observations near the "physical" boundary.

3.2 Confidence Level

We shall use the terms "significance level" and "confidence level" interchangeably. The meaning depends on the context, as follows:

Confidence Level: In frequency statistics, the term "confidence level" is

> used to refer to the probability that the statement made will be correct, in the frequency sense. Thus, for a "68% confidence interval", the confidence level is 68%, because the interval includes the true value of the parameter in 68% of the trials. Likewise, for a hypothesis test, the "size of the test" may be referred to as a confidence level, because that is the probability with which the null hypothesis is accepted, assuming the null hypothesis is correct.
>
> In Bayesian statistics, the term "confidence level" is used to refer to the degree of belief in the statement made, that is, to the size of an integral for some region over the posterior probability distribution.

In this document, the symbol α will usually be used when referring to a confidence level. Other terms that often appear in the literature are "confidence coefficient", "probability content", "degree of confidence", *etc.* There is not a lot of uniformity in usage. Note also that some authors define this quantity to be one minus the definition here.

3.3 Likelihood Function

An experimental measurement is considered to be a sampling from a probability distribution, describing the probability of observing any given result in the sample space of possible outcomes. This distribution is often referred to as the "sampling distribution". Thus, the complete description of the experiment consists in giving the result of the measurement ("sampling"), together with the sampling distribution. Typically, the sampling distribution depends on the values of unknown parameter(s), which we are trying to learn about.

The probability distribution may be contrasted with the "likelihood function":

Likelihood Function If an experiment has been performed resulting in a measurement x, drawn from some probability distribution with population parameter θ, the **likelihood function** for that experiment is defined as the probability distribution function evaluated at the observed value of x.

Such a likelihood function may be denoted by $\mathcal{L}(\theta; x)$, and is treated as a function of θ in algorithms concerned with summarizing information relevant to θ, or in making inferences about the value of θ.

Chapter 4

Point Estimation

The problem of point estimation is to arrive at a "best" estimate of the value of an unknown parameter governing the sampling distribution. The choice of estimator, that is, the definition of "best," depends on various possible criteria. We describe these below in a heuristic way. More careful definitions, in terms of ϵs and δs can be found in texts such as Kendall and Stuart, but are generally no more essential than are the ϵs and δs of calculus in doing day-to-day calculations.

4.1 Consistency, Bias, Efficiency, and Sufficiency

Our task then is to determine some unknown parameter or parameters associated with a probability distribution function (PDF) whose form is known. We are to use some data - random variables - and some rule associating an estimate of the parameter(s) θ with the data $t_1, t_2 \ldots$.

Consider some PDF and ask how we might find its mean μ and variance (σ^2). Given n data points we might take as estimators[1] for μ and σ

$$\begin{aligned} \mathcal{E}_\mu &= \frac{1}{n}\sum t_i \\ \mathcal{E}_{\sigma^2} &= \frac{1}{n}\left(\sum t_i^2 - \frac{1}{n}(\sum t_j)^2\right) \end{aligned} \tag{4.1}$$

[1]We use here the notation $\mathcal{E}_\theta$ to stand for an estimator for quantity θ. Another common notation for the same quantity is $\hat{\theta}$, which we shall also use elsewhere.

Now take expectation values with respect to the true distribution, whatever it is. Then

$$\begin{aligned} \langle t \rangle &= \mu \\ \left\langle (t-\mu)^2 \right\rangle &= \left\langle t^2 \right\rangle - \mu^2 = \sigma^2 \end{aligned} \tag{4.2}$$

so we find

$$\begin{aligned} \langle \mathcal{E}_\mu \rangle &= \frac{1}{n} \sum \langle t_i \rangle = \mu \\ \langle \mathcal{E}_{\sigma^2} \rangle &= \frac{1}{n} \left\langle \sum (t_i^2 - \frac{1}{n} (\sum t_j)^2) \right\rangle \\ &= \left\langle \frac{n-1}{n^2} \sum (t_i^2) - \frac{1}{n^2} \sum_{i \neq j} t_i t_j \right\rangle \\ &= \frac{n-1}{n} (\left\langle t^2 \right\rangle - \mu^2) = \frac{n-1}{n} \sigma^2 \end{aligned} \tag{4.3}$$

What we see is two different behaviors. The estimator for the mean has an expectation value that is equal to the true mean for all n. On the other hand, our estimator for the variance on average underestimates the true variance for all n. We say that the estimator for the mean is **unbiased** because its expectation value is the true value for any sample size. The estimator for the variance is biased. However, the estimator for the variance is **consistent** because asymptotically it converges to the true value. It might be thought that we should always use an unbiased estimator, but this is not really so important in practice because a bias of order $1/n$ is likely to be swamped by a statistical uncertainty of order $1/\sqrt{n}$.

What is important for sure is that we use a method that doesn't introduce any unnecessary uncertainty into the determination of the physical parameters of interest. That is, we would like an estimator of a parameter that has the least variance, *i.e.*, the least expected uncertainty. We state without proof (see, *e.g.*, Kendall and Stuart, v. 2, pp. 8-9 of 3rd Edition) the following:

Theorem: (**Exponential Form**) Suppose the PDF is $f(t, \theta)$ and the observed data are $t_1, t_2, \ldots t_n$. Form

$$\ln \mathcal{L} = \sum_i \ln f(t_i, \theta) \tag{4.4}$$

and calculate its derivative with respect to θ. If it can be written in the form

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = A(\theta)[\mathcal{E}(t_1, t_2, \dots t_n) - g(\theta)] \tag{4.5}$$

then $\mathcal{E}(t_1, t_2, \dots t_n)$ is an unbiased estimator of $g(\theta)$ with the minimum possible variance and that variance is

$$\mathrm{var}(\mathcal{E}) = \frac{(dg/d\theta)^2}{A(\theta)} \tag{4.6}$$

where

$$A(\theta) = n \int dt \frac{1}{f} \left(\frac{\partial f}{\partial \theta} \right)^2 \tag{4.7}$$

Such an estimator is called, imaginatively, a **Minimum Variance Bound** (MVB) estimator.

As a pertinent example take $f = e^{-t/\tau}/\tau$ so that we are trying to find the lifetime. Now

$$\frac{\partial \ln \mathcal{L}}{\partial \tau} = n\tau^{-2}[\frac{1}{n}\sum_i t_i - \tau] \tag{4.8}$$

so the mean of the measured decay times is an unbiased estimate of true lifetime with variance

$$\sigma^2 = \tau/\sqrt{n} \tag{4.9}$$

a satisfyingly familiar result. Note that if we ask whether we can get an unbiased estimate of minimal variance for the decay rate $\Gamma = 1/\tau$ we find we cannot! The reason is that $[(1/n)\sum_i t_i]^{-1}$ is a biased estimator of Γ. This underlines the relative lack of importance to be attached to having unbiased estimators.

Usually we are not so concerned with corrections of order $1/n$ and are happy enough to have estimators that asymptotically have minimal variance. These are called **efficient** estimators of $g(\theta)$ and they have variance

$$\mathrm{var}(\mathcal{E}) = \frac{\left(\frac{dg}{d\theta}\right)^2}{n \int dt \, \frac{1}{f} \left(\frac{\partial f}{\partial \theta}\right)^2} \tag{4.10}$$

In particular, an efficient estimator $\mathcal{E}$ of θ itself has variance

$$\mathrm{var}(\mathcal{E}) = \frac{1}{n \int dt \, \frac{1}{f} \left(\frac{\partial f}{\partial \theta}\right)^2} \tag{4.11}$$

Below we explain that the maximum likelihood method gives us asymptotically efficient estimators.

Given many measurements, t_i, we can imagine forming many combinations of them, each of which carries some information. For example, we could form the estimators for the mean and variance as described above. Now if we want to determine some particular parameter, the various combinations we can form of the data may or may not be useful. Suppose there is one combination that carries all the useful information, that is once that combination is available, nothing else can be learned from the data no matter what combination we form. We then call this particular statistic **sufficient**. It turns out that this concept is closely related to the question of finding an estimator whose variance is the minimum allowed by the general theorem above. If there is an MVB estimator (and we've given the condition to determine whether there is one), then that estimator is a sufficient statistic. In the example above, we see that $(\sum t_i)/n$ is a sufficient statistic for τ. It is clear that if we find a sufficient statistic for θ, then we have a sufficient statistics for any function of θ. For example, we found the the mean of the decay times is a sufficient statistic for the lifetime and therefore for the decay rate, as well.

4.2 Maximum Likelihood Method

The **Maximum Likelihood Method** is a convenient means of determining parameters in a fit to data. We suppose we are to determine a quantity θ by fitting with a dataset $\{t_i\}$. For simplicity, we assume that there is a single distribution $f(t, \theta)$. It is normalized so that

$$\int dt\, f(t, \theta) = 1 \tag{4.12}$$

independent of A. The log-likelihood function for this dataset is

$$\ln \mathcal{L} = \sum_i \ln f(t_i, \theta) \tag{4.13}$$

We will treat θ as a variable here, and indicate the true value of the parameter by $\overline{\theta}$. As the number of data points increases, the likelihood function converges to

$$\ln \mathcal{L} \to n \int dt\, f(t, \overline{\theta}) \ln f(t, \theta) \tag{4.14}$$

since $f(t,\overline{\theta})$ gives the actual distribution of events in t.

The likelihood function is maximized when

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} = 0 \tag{4.15}$$

provided, of course, that

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} < 0 \tag{4.16}$$

The log-likelihood estimator of $\overline{\theta}$ is the value of θ obtained by solving for the maximum of the log-likelihood.

Now as the number of events increases, we see that

$$\frac{\partial \ln \mathcal{L}}{\partial \theta} \to n \int dt\, f(t,\overline{\theta}) \frac{1}{f(t,\theta)} \frac{\partial f(t,\theta)}{\partial \theta} \tag{4.17}$$

If we evaluate this at $\theta = \overline{\theta}$ we find it vanishes as a consequence of

$$\frac{\partial}{\partial \theta} \int dt\, f(t,\theta) = 0 \tag{4.18}$$

Thus asymptotically the maximum likelihood occurs at the true value of the parameter.

For each dataset we find a value of θ that maximizes the log-likelihood. The values of θ we find will be (asymptotically) distributed in a normal distribution around $\overline{\theta}$. With $\delta\theta = \theta - \overline{\theta}$, the probability distribution is

$$\frac{dP}{d\theta} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\delta\theta^2}{2\sigma^2}} \tag{4.19}$$

where

$$\frac{1}{\sigma^2} = -\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \tag{4.20}$$

Asymptotically

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \theta^2} \to n \frac{\partial^2}{\partial \theta^2} \int dt\, f(t,\overline{\theta}) \ln f(t,\theta) = -n \int dt\, \frac{1}{f} \left(\frac{\partial f}{\partial \theta}\right)^2 \tag{4.21}$$

where everything is evaluated at $\overline{\theta}$. From this we find the famous asymptotic relation

$$\sigma^{-2} = n \int dt\, \frac{1}{f} \left(\frac{\partial f}{\partial \theta}\right)^2 \tag{4.22}$$

which shows that the log-likelihood estimator is efficient.

Often we have events that fall into separate categories, each of which has its own distribution. Suppose there are two categories with distributions $f_1(t,\theta)$ and $f_2(t,\theta)$. If these are normalized so that

$$\int dt\, f_1(t,\theta) + \int dt\, f_2(t,\theta) = 1 \tag{4.23}$$

then everything will work as before. The log-likelihood function is simply the sum of two pieces, where each data point contributes to the appropriate sum:

$$\ln \mathcal{L} = \sum_i \ln f_1(t_i,\theta) + \sum_j \ln f_2(t_j,\theta) \tag{4.24}$$

The corresponding σ is given by

$$\sigma^{-2} = n\left[\int dt\,\frac{1}{f_1}\left(\frac{\partial f_1}{\partial\theta}\right)^2 + \int dt\,\frac{1}{f_2}\left(\frac{\partial f_2}{\partial\theta}\right)^2\right] \tag{4.25}$$

As a pertinent example consider the time distribution in $B \to J/\psi K_S$. If the opposite B is tagged as a B^0 the distribution is

$$f_+(DA;t) = \frac{\Gamma}{4}e^{-\Gamma|t|}(1 + DA\sin\Delta m t) \tag{4.26}$$

whereas if the tag is a $\overline{B}^0$, the sign in front of DA is negative. We see that these distributions are properly normalized:

$$\int dt\, f_+(t,DA) + \int dt\, f_-(t,DA) = 1 \tag{4.27}$$

We compute directly

$$\begin{aligned}\sigma^{-2} &= n\int_0^\infty dt\,\Gamma e^{-\Gamma t}\frac{D^2\sin^2\Delta m t}{1 - D^2A^2\sin^2\Delta m t}\\ &= nD^2\left[\frac{2x^2}{1+4x^2} + \frac{24x^4}{(1+4x^2)(1+16x^2)}D^2A^2\ldots\right]\end{aligned} \tag{4.28}$$

If there are several tagging categories i, each with a fraction ϵ_i of the events and with dilution $D_i = 1 - 2w_i$, there is a contribution to σ^{-2} from each:

$$\sigma^{-2} = n\sum_i \epsilon_i D_i^2\left[\frac{2x^2}{1+4x^2} + \frac{24x^4}{(1+4x^2)(1+16x^2)}D_i^2A^2\ldots\right] \tag{4.29}$$

With the convenient approximation $x \approx 1/\sqrt{2}$ and dropping all but the lowest order term, we obtain the rule of thumb:

$$\sigma(A) = \sqrt{\frac{3}{n \sum_i \epsilon_i D_i^2}} \tag{4.30}$$

Suppose we have two categories but that we don't know what fraction of the events will fall into each. We can introduce a second variable, ϵ, representing the fraction in the first distribution and maximize the likelihood with respect to both θ and ϵ. Here we require unit normalization for the two distributions separately:

$$\int dt\, f_1(t, \theta) = 1; \qquad \int dt\, f_1(t, \theta) = 1 \tag{4.31}$$

and take the two functions to be

$$\epsilon f_1; \qquad (1 - \epsilon) f_2 \tag{4.32}$$

Now when we maximize with respect to ϵ we find

$$\frac{\partial}{\partial \epsilon} \left[\sum_{i=1}^{m} \ln[\epsilon f_1(t_i, \theta)] + \sum_{j=1}^{n} \ln[(1 - \epsilon) f_2(t_j, \theta)] \right] = 0 \tag{4.33}$$

the result is simply

$$\frac{m}{\epsilon} = \frac{n}{1 - \epsilon} \tag{4.34}$$

In other words, ϵ is determined simply to reproduce the number of events in the two categories. The equation for θ will be just as if we hadn't worried about ϵ. Thus, if f_1 and f_2 have unit normalization, we can forget about ϵ.

When, as in the case above, there is more than one parameter to be determined, say θ_i, $i = 1, ..q$, the likelihood equations are simply

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_i} = 0 \tag{4.35}$$

For this to be a maximum we need the (negative of the) matrix of second derivatives

$$L_{ij} = -\frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j} \tag{4.36}$$

to be positive definite. Asymptotically the distribution of output θ_i for a set of true values $\overline{\theta}_i$ has a Gaussian shape in $d\theta_i = \theta_i - \overline{\theta}_i$

$$dP = \sqrt{\det L}(2\pi)^{-q/2} e^{-\frac{1}{2}\delta\theta_i \delta\theta_j L_{ij}} d\theta_1 \dots d\theta_q \tag{4.37}$$

From this we can find not just the expected variance in θ_i, $\langle \delta\theta_i^2 \rangle$ but more generally $\langle \delta\theta_i \delta\theta_j \rangle$. A direct computation shows that

$$\langle \delta\theta_i \delta\theta_j \rangle = [L^{-1}]_{ij} = M_{ij} \tag{4.38}$$

The inverse of L, the covariance matrix M, is thus the central quantity. In particular

$$\sigma_i = M_{ii}^{-1/2} \tag{4.39}$$

and the correlation matrix is

$$\rho_{ij} = \frac{M_{ij}}{\sigma_i \sigma_j} \tag{4.40}$$

The elements of L are given asymptotically by

$$L_{ij} = N \int dt \, \frac{1}{f} \frac{\partial f}{\partial \theta_i} \frac{\partial f}{\partial \theta_j} \tag{4.41}$$

If, due to fluctuations, the output θ_i comes out one sigma high, then we expect θ_j to come out ρ_{ij} units of sigma high, as well.

In practice, the CERN workhorse Minuit is generally used to find the minimum of the negative log-likelihood. Since Minuit is looking for χ^2 we need to convert log-likelihood to this quantity. If it is the (positive) loglikelihood that is computed, probabilities vary as (in the Gaussian approximation) $e^{\Delta \ln \mathcal{L}}$ while for χ^2 we have $e^{-\chi^2/2}$ so the conversion is $-2 \ln \mathcal{L} \sim \chi^2$. Minuit produces both the covariance matrix and the correlation matrix as part of the regular output.

Chapter 5

Confidence Intervals

5.1 Definition of a Confidence Interval

The definition of a confidence interval was originally proposed by Neyman [4]:

> "If the functions θ_ℓ and θ_u possess the property that, whatever be the possible value ϑ_1 of the parameter θ_1 and whatever be the values of the unknown parameters $\theta_2, \theta_3, \ldots, \theta_s$, the probability
>
> $$P\{\theta_\ell \leq \vartheta_1 \leq \theta_u | \vartheta_1, \theta_2, \ldots, \theta_s\} \equiv \alpha, \tag{5.1}$$
>
> then we will say that the functions θ_ℓ and θ_u are the lower and upper confidence limits of θ_1, corresponding to the confidence coefficient α."

This definition is often misunderstood, so further exposition is in order: First, to clarify some terminology, what Neyman refers to as a "confidence coefficient" will be referred to here as a "confidence level". The interval (θ_ℓ, θ_u) is called the **confidence interval** for θ_1. Note that some authors, including Kendall and Stuart, use $1-\alpha$; present usage is intended to be consistent with the dominant convention in high energy physics. The "functions θ_ℓ and θ_u" is to be understood as meaning that these quantities are funtions of whatever random variable is being sampled. Thus, θ_ℓ and θ_u are themselves random variables.

A simple example should help: Suppose the sampling distribution is a uniform distribution on $(\theta, \theta + 1)$:

$$p(x; \theta) = \begin{cases} 1 & x \in (\theta, \theta + 1) \\ 0 & \text{Otherwise.} \end{cases} \tag{5.2}$$

Suppose further that we have sampled a value X from this distribution, and wish to obtain an $\alpha = 90\%$ confidence interval on the parameter θ. We search for an interval of the form: $\theta_\ell(x) = x - a$, $\theta_u(x) = x - a + b$, with $b > 0$. Note that other choices are possible; the choice here satisifies the desirable property of sufficiency, as well as "simplicity" (being linear functions). The condition on a and b for a confidence interval is:

$$\begin{aligned} \alpha = 0.9 &= \text{Prob}[\theta_\ell(x) < \theta < \theta_u(x)] & (5.3) \\ &= \text{Prob}(x - a < \theta < x - a + b) & (5.4) \\ &= \text{Prob}(x < \theta + a) - \text{Prob}(x < \theta + a - b) & (5.5) \\ &= \int_{\theta+a-b}^{\theta+a} p(x;\theta)dx. & (5.6) \end{aligned}$$

The solution to this equation for a and b is not unique – other criteria must be invoked to decide which choice of confidence interval to use. For example, we might wish to quote a symmetric interval about the point estimator $\hat{\theta} = X - 1/2$. In this case, we give the confidence interval $(X - 0.95, X - 0.05)$. What we mean when we quote such an interval, is that 90% of the time we apply this prescription (*e.g.*, in many repetitions of the experiment), the quoted interval will include θ, and 10% of the time it will exclude θ. That is, a confidence interval as defined above is to be understood in terms of frequentist statistics.

To understand the definition better, and how to apply it, consider (following paraphrased from [3]) a distribution $p(x;\theta)$ dependent on a single unknown parameter θ and suppose that there is a random sample of n values $x_1, x_2, ...x_n$ from the population. For any given confidence level α, we seek two quantities θ_ℓ and θ_u such that:

$$P(\theta_\ell < \theta < \theta_u) = \alpha, \tag{5.7}$$

for any value of θ. The quantities θ_ℓ and θ_u depend only on α and the sample values, plus any further criteria needed to arrive at a unique prescription.

For any fixed α, the totality of confidence intervals for different data samples $\{x\}$ determines a field within which θ is asserted to lie. This field is called the **confidence belt**. We can graphically represent the confidence belt in the plane of the parameter θ and the data x (condensed into a single quantity here for illustration, see Fig. 5.1).

The confidence belt for an experiment and significance level α may be constructed as follows, referring to Fig. 5.1: For any value of θ, say θ_0, we find values of the random variable $x_\ell(\theta_0)$ and $x_u(\theta_0)$ such that

$$\begin{aligned} \alpha &= \text{Prob}[x_\ell(\theta_0) < x < x_u(\theta_0)] & (5.8) \\ &= \int_{x_\ell(\theta_0)}^{x_u(\theta_0)} p(x;\theta_0)dx. & (5.9) \end{aligned}$$

The interval $(x_\ell\theta_0), x_u(\theta_0))$ is graphed as a horizontal line in Fig. 5.1. We carry out this procedure for all values of θ, and plot the band as shown in the figure. This will define an area in the x, θ-plane which is the confidence belt. Note that the construction may require including values of θ which are considered "unphysical" in order to include the entire sample space. In the present context of frequency statistics, this should not be regarded as a difficulty, as no claim is being made concerning possible values of θ.

This belt is now used to define intervals in θ that correspond to vertical lines in Fig. 5.1. These intervals in θ are the *confidence intervals*. That is, once the confidence belt has been constructed, we may use it to read off the desired confidence interval, for any given sampled value X. We simply go to the value of X on the horizontal axis, and read off the values of θ in the belt at that value of X.

Any method which gives confidence intervals that contain the true value with probability α is said to have a "correct coverage", which is a frequentist concept. The intervals as constructed above have the correct coverage by definition.

5.1.1 One and two sided Confidence Intervals

One sided confidence intervals ("limits") are often quoted when a measured value is close to a physical boundary, otherwise we typically quote two sided ("central") confidence intervals. For example, a mass or branching fraction measurement which results in a value not significantly (in the statistical sense) above zero may be summarized as a limit. In HEP, limits are usually quoted at the $\alpha = 90\%$ confidence level, while central intervals are nearly always approximately 68% confidence intervals.

There is, however, a slight pitfall here: Often, the decision to quote a limit or a central interval is based on the result itself. This introduces an additional dependence in the PDF which is usually not modeled. The result may acquire a bias. For example, it is more likely that a central interval ("positive result") will be reported if an upward fluctuation has occurred. Thus, first reports of branching fractions tend to be biased high.

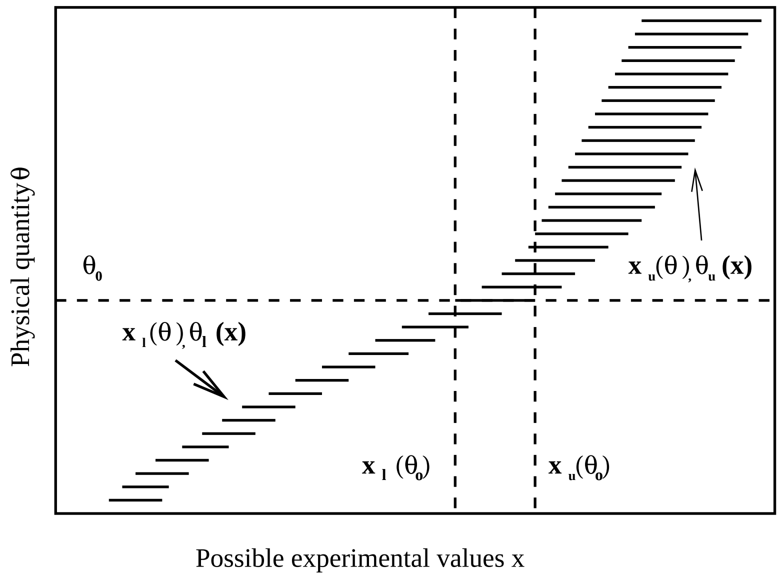

Figure 5.1: Confidence intervals for a single unknown parameters θ, for the possible values of x.

There are various ways of mitigating this difficulty. One recent proposal combines a solution to this problem with a means of quoting an interval which can satisfy some boundary constraint. This is the method of Feldman and Cousins [5]. The approach consists in building the confidence belt using the Likehood Ratio ordering principle. This means building the confidence interval for a particular value of θ adding values of x to the acceptance region according to the order given by the Likehood Ratio, $R(x)$. The two major examples of how to build confidence belts with the unified approach and how these compare to the frequentistic approach are given for Gaussian data close to the boundary and for Poisson data for small samples in Ref. [5]. Roe and WoodRoofe worried about the properties of this method in the case of small statistics, and produced alternative recommendations [6, 7].

5.2 Confidence Regions in Multidimensional Cases

The extension of the notion of a Confidence Interval to multidimensional cases, resulting in "Confidence Regions" is straightforward in principle. However, some discussion of the common two-dimensional case of an "error ellipse" will help to avoid some common confusion.

Figure 5.2 shows an "error ellipse" for a two parameter (θ_1, θ_2) problem. The center of the ellipse is given by estimators $(\hat{\theta}_1, \hat{\theta}_2)$, and the outline of the ellipse depends on the moment matrix for these estimators as shown. Typically, such an ellipse is obtained either as the locus of points where the χ^2 increases by one from its minimum (at least-squares estimators $(\hat{\theta}_1, \hat{\theta}_2)$, or where the ln of the likelihood decreases by 1/2 from its maximum (again, at $(\hat{\theta}_1, \hat{\theta}_2)$, but this time standing for maximum likelihood estimators). The area enclosed by this ellipse is not, in general, a 68% confidence region. Instead, if the estimators $(\hat{\theta}_1, \hat{\theta}_2)$ are sampled from a normal distribution, it is the one-dimesional projections of the ellipse which are 68% confidence intervals. For example, the interval $(\hat{\theta}_1 - \sqrt{(H^{-1})_{11}}, \hat{\theta}_1 + \sqrt{(H^{-1})_{11}})$ is a 68% confidence interval (in the normal case). In fact, the ellipse corresponds to approximately a 39% confidence region.

5.3 Bayesian Intervals

Interval estimation in Bayesian statistics, yielding **Bayes Intervals**, is based on assumptions concerning the prior distributions (more details can be found in Section 8.2). The method presented in the Review of Particle Properties up to the 1997 version [8] to explain how to set upper limits for Poisson processes in presence of background, is described by the formula:

$$\alpha = 1 - \frac{e^{-(\mu_B+N)} \sum_{n=0}^{n=n_0} \frac{(\mu_B+N)^n}{n!}}{e^{-\mu_B} \sum_{n=0}^{n=n_0} \frac{\mu_B^n}{n!}}. \tag{5.10}$$

N is the upper limit on the unknown mean μ_S for the signal with confidence coefficient ϵ. μ_B is the background mean and n_0 is the observed number of events in the Poisson process under investigation.

The formula was derived by Helene [9] using Bayesian statistics with uniform prior. Such formulas can be investigated in frequency terms, but

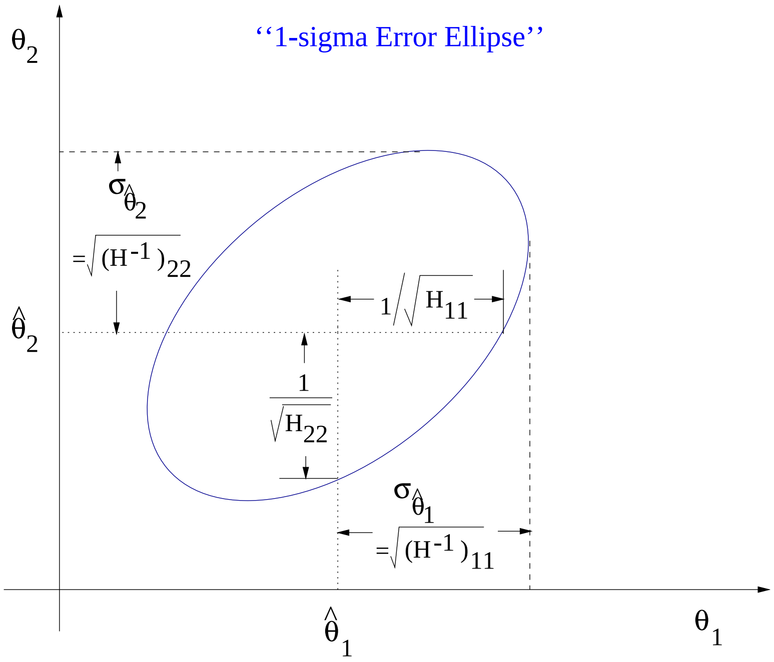

Figure 5.2: The "error ellipse" in (θ_1, θ_2) parameter space. H^{-1} is the covariance matrix for the estimators $(\hat{\theta}_1, \hat{\theta}_2)$.

the frequencies will typically be only approximately given by α. In the case of discrete distributions such as the Poisson, exact confidence intervals are difficult (though not impossible) to achieve, and one usually accepts intervals which may over-cover. In the particular example cited here, the frequency properties have suffered some confusion in the literature, see references [10], [11].

Chapter 6

Testing Hypotheses

Statistical tests are no substitute for common sense. If a fit does not look right, it probably is not. If a result looks too good to be true, it probably is. With these caveats, one can use statistics to determine whether data is consistent with any one hypothesis, and one can use statistics to determine how well data discriminate between competing hypotheses.

6.1 How Does One Test a Hypothesis?

Is a set of data consistent with a hypothesis? We construct tests based on expected outcomes. Various statistical measures are calculated and compared to those we expect. Our tests depend on both the data and the model(s), and interpreting the results can require some care. We will focus our attention here on using χ^2 and likelihood ($\mathcal{L}$) calculations and their associated significance levels.

Measuring the chi-square (χ^2) for ν degrees of freedom is a very commonly used test of a hypothesis. In the general case, for a series of measurements x_i where a hypothesis predicts values ξ_i, we define the difference vector $\epsilon_i = x_i - \xi_i$ and use the weight matrix W_{ij} [the inverse of the covariance matrix, $W_{ij} = (M^{-1})_{ij}$], to calculate

$$\chi^2 = \epsilon_i W_{ij} \epsilon_j \ .$$

In the special case where the measurements are expected to be uncorrelated, so that the covariance matrix is diagonal, we denote the expected standard deviation of x_i as $\sigma_i = \sqrt{M_{ii}}$. The weight matrix becomes diagonal with non-zero elements $W_{ii} = 1/\sigma_i^2$ and

$$\chi^2 = \sum \frac{(x_i - \xi_i)^2}{\sigma_i^2} \ .$$

If the predicted values ξ_i are independent of the observed values x_i, then the number of degrees of freedom, ν, is the number of terms in the sum. If the predicted values ξ_i depend on fit parameters which depend in turn on the observed values x_i, then ν will be reduced.

6.1.1 Testing a Vertexing Hypothesis

A χ^2 calculation can be used to discriminate between signal and background on an event-by-event basis. In this case we accept candidates for which χ^2 indicates a high enough probability of belonging to the signal distribution and reject those for which it indicates a relatively low probability of belonging to the signal distribution. As an example, we can consider the case where we want to isolate a sample of $K_S^0 \to \pi^-\pi^+$ decays. A key feature of the signal is that both pion tracks emerge from a common vertex. Given the errors on the measured trajectories, a vertexing algorithm calculates a vertex position, an error matrix for the derived vertex position which depends on the trajectories and their error matrices, and from these χ^2.

BABAR's `BtaOpFastVtxV0` class calculates a χ^2 for one degree of freedom. If the calculation of the error matrix is correct, the distribution of χ^2 for signal events should be the same as that which would be observed for a Gaussian distribution of width unity: 68.3% of the events will have $\chi^2 \leq 1.0$; 95.5% will have $\chi^2 \leq 2.0$, etc. One may then choose to accept candidates with $\chi^2 < 2.5$ and reject those with greater values of χ^2, for example. One often converts a χ^2 measurement into what is commonly called a *probability* measurement assuming that the observed value of χ^2 is derived from a normal distribution. We will define the significance level (α) for the hypothesis that the two tracks do emerge from a common vertex as

$$\alpha(\chi^2_A) \ \equiv \ 1 - \int_{\chi^2 < \chi^2_A} \mathcal{P}(x)\, dx \ = \ \int_{\chi^2 > \chi^2_A} \mathcal{P}(x)\, dx$$

where $\mathcal{P}(x)$ is the gaussian probability for a normal distribution in one dimension:

$$\mathcal{P}(x) \ \equiv \ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\chi^2(x)} \ .$$

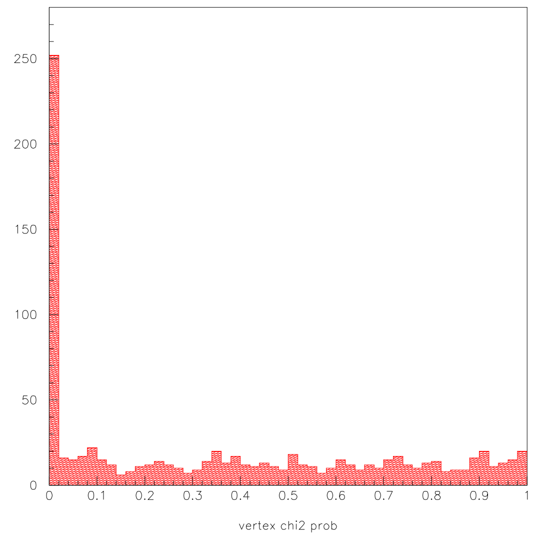

Figure 6.1: χ^2 Confidence Level distribution for vertex fit example.

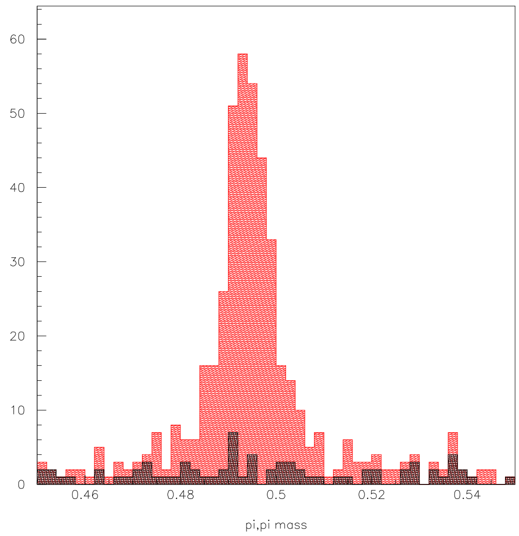

Figure 6.2: $\pi\pi$ mass distribution for data in Fig. 6.1. The dark shaded events are those with $\alpha < 0.01$.

For signal events, the significance level should be *uniformly* distributed between 0 and 1.

Fig. 6.1 shows the significance level distribution of $K_S^0 \to \pi^-\pi^+$ candidates which have been selected requiring that they point back to a primary vertex candidate from which they are separated by 0.5 - 6.0 cm, that the K_S^0 vertex is inside the beam pipe, and that the candidate mass lies in the range 450 MeV $< m(\pi^-\pi^+) <$ 550 MeV. In addition, the vertexing significance level for each candidate is greater than 0.001, a selection made at the NTUPLE level. The shape appears to consist of a peak at low values, presumably due to background, superposed on a relatively uniform distribution, which we hope is associated with signal. This qualitative understanding is borne out in Fig. 6.2. Here, the $m(\pi^-\pi^+)$ distribution is shown for all the candidates, and with that for candidates falling in the first bin of Fig. 6.1 (those with $\alpha < 0.01$) shaded in the darker color. Several conclusions can be drawn from these plots:

- the vertex χ^2, or the corresponding significance level, discriminates well between the hypothesis that two tracks emerge from a common point in space and the complementary hypothesis that they do not, at least for K_S^0 candidates in this sample.

- the significance level distribution is uniform between 2% and 100%, indicating that the tracks' error matrices and the algorithm for combining them are correct.

- in the range above a few percent, the significance level (or the corresponding value of χ^2) *does not discriminate* significantly between signal and background. The signal is uniformly distributed in this range, and the background is essentially absent. A candidate with significance level 95% is only marginally more likely to be signal than one with significance level 35%.

- in the range below 2%, the significance level (or the corresponding value of χ^2) may or may not discriminate between hypotheses. To address this question one would need to determine the significance level distribution for both signal and background in this range.

6.1.2 Fitting a Distribution – a Related Example

We can work with the same set of data and ask more questions. The data from Fig. 6.2 which remain after requiring signficance level greater than 2% are shown in Fig. 6.3. We can fit this data as the sum of a linear background and a Gaussian signal two different ways using HFIT in PAW: the default algorithm uses a bin-by-bin χ^2 minimization and a somewhat more sophisticated algorithm maximizes likelihood. (Note that HFIT uses MINUIT.) Both fit the data to the functional form

$$f(x) = p_1 + p_2 x + p_3 e^{-\frac{1}{2}\left(\frac{(x-p_4)}{p_5}\right)^2} .$$

The central values and the reported errors for each of the parameters for the two fits are listed in Table 6.1. In each case, the fit reports χ^2 values for 43 points, the number of bins with non=zero entries. What do these numbers mean? Can they be interpreted correctly using simple statistical arguments?

The reported values of χ^2 appear to be reasonable, 32.2 in the first fit and and 37.2 in the second fit. The number of points used in calculating χ^2 is 43 for each fit, and the number of parameters 5, so one might calculate

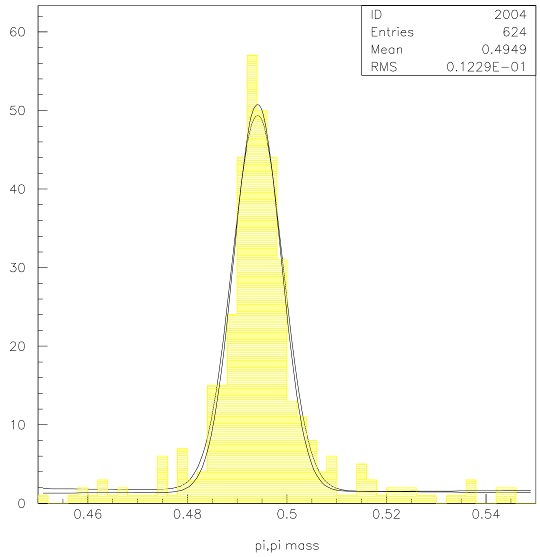

Figure 6.3: Data described in the text with PAW χ^2 minimization and binned maximum likelihood fits superposed. The parameters for both fits are reported in Table 6.1.

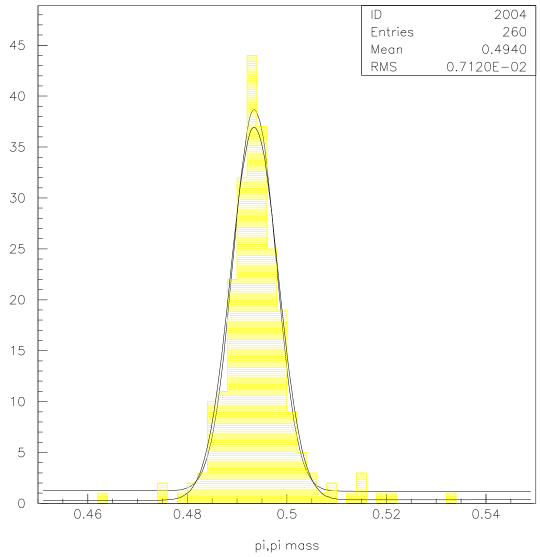

Figure 6.4: Data of Fig. 6.3 with helicity angle cut. Again, both χ^2 minimization and binned maximum likelihood fits are superposed

χ^2/ν for the two fits to be 0.85 and 0.98 using $\nu = 38$. From a statistical perspective, what are the possible problems with this approach? First, the χ^2 calculations (and the χ^2 minization fit) ignore bins with zero entries. Second, the interpretation of χ^2 is well-defined only when the parent distribution is expected to be Gaussian; when the number of entries in a bin is zero, one, or two, this cannot be the case. Nonetheless, the fits look good to the eye, and the values of χ^2 probably indicate that the fits describe the distribution of the data points accurately. This can be tested by generating Monte Carlo samples of similar size from known distributions of signal and background, fitting these samples, and observing the results. Do the central values of the fits reproduce the generated values? Do the fitted values of the error estimates reproduce the the fluctuations in the central values? Do the χ^2 distributions from these fits include the experimentally observed values with reasonable probability? The central values of the K_S^0 mass are the same in the two fits, but the central values for the widths differ by more than 1 σ. And this for precisely the same data. The tests described must therefore exclude at least one of the the two fitting techniques. The only way to certify the validity of the fitting process is to test it in detail.

parameter	from maximum likelihood fit	from χ^2 minimization
p_1	4.2 ± 0.5	-0.7 ± 3.5
p_2	-5.2 ± 0.9	3.3 ± 7.1
p_3	47.8 ± 3.8	49.5 ± 4.3
p_4	0.4941 ± 0.0003	0.4941 ± 0.0003
p_5	0.0051 ± 0.0003	0.0047 ± 0.0003

Table 6.1: These are the parameters determined by the binned maximum likelihood fit and the χ^2 minimization fit to the data of Fig. 6.3 described in the text.

Even though both fits give the same central value for the K_S^0 mass, and even though the reported values of χ^2/ν are good for both fits, and even though both fits appear to locate the center of the peak correctly, there is a substantial problem with the fit – it does not provide a correct measurement of the mass of the K_S^0. Both fits report the mass to be 494.1 ± 0.3 MeV/c^2. The PDG value is 497.67 ± 0.03 MeV/c^2. The difference between the fitted values and the PDG value is approximately 3.6 ± 0.3 MeV/c^2, a

12 σ difference. What if our hypothesis is wrong? What if the background is not uniform? Could there be a reflection under the peak which distorts the apparent shape? To eliminate background from $\Lambda \rightarrow p\pi^-$ reflections and e^+e^- pairs, one may require that the magnitude of the cosine of the polar angle of the pions in the $\pi^-\pi^+$ center-of-momentum, taken with respect to the line of flight, be less than 0.70. This produces the cleaner signal seen in Fig. 6.4. The two fits of this data set find K_S^0 mass values of 493.5 ± 0.0003 MeV/c^2. Furthermore, if one further divides this data set into three momentum ranges ($p < 0.5\,\mathrm{GeV}/c$, $0.5\,\mathrm{GeV}/c < p < 0.9\,\mathrm{GeV}/c$, and $p > 0.9\,\mathrm{GeV}/c$) with roughly equal populations, the K_S^0 mass values range from 493.5 to 493.7 MeV/c^2 with statistical errors of 0.6 - 0.7 MeV/c^2. There is no significant systematic variation which even hints that the central values from the fits are wrong. The hypothesis that the distribution can be described by a linear background and a Gaussian signal is consistent with the data; however, the fitted K_S^0 mass is not correct within the statistical error. Simple statistical tests do not test the hypothesis that the derived K_S^0 mass is correct. [For the curious - the wrong magnetic field value was used in a tcl file, and there is no statistical test to find such a mistake.]

6.1.3 Fitting a Data Set with Low Statistics

Calculating χ^2 or χ^2/ν for data sets with low statistics can be very misleading. The χ^2 test is only valid if the underlying sampling distribution is normal. For example, in the low statistics case, this assumption is incorrect. To compare data with two (or more) hypotheses, one should calculate the likelihood ratio(s) for these hypotheses. We will consider here the simplest case where one does a simple counting experiment and the observations can be described by Poisson statistics. That is, in any set of observations where the expected number of events is μ, the probability to observe n is

$$\mathcal{P}(n) = \frac{\mu^n e^{-\mu}}{n!}$$

To be concrete, let us consider two hypotheses A and B with measurements n_i in bins $i = 1, 2, \cdots$. The predicted number of entries in each bin will be denoted $\mu_{A,i}$ and $\mu_{B,i}$ for the two hypotheses. Of course, these predicted values are the mean values for ensembles of of measurements and will be real numbers; the observed values will be integer numbers of events observed. The ratio of the likelihoods will be

$$\frac{\mathcal{L}_A}{\mathcal{L}_B} = \prod_i \frac{(\mu_{A,i})^{n_i}\, e^{-\mu_{A,i}}}{(\mu_{B,i})^{n_i}\, e^{-\mu_{B,i}}}$$

Usually, we consider the difference of the logarithms rather than the ratio itself, and fold in an additional factor of -2:

$$-2\Delta \ln \mathcal{L} = -2\left(\ln \mathcal{L}_A - \ln \mathcal{L}_B\right) = -2 \ln\left(\frac{\mathcal{L}_A}{\mathcal{L}_B}\right).$$

This makes comparisons with $\Delta\chi^2$ more intuitive – a lower value of this difference will correspond to a better hypothesis. In the limit of high statistics where the Poisson may be approximated with a Gaussian ($\mathcal{P}_{\mathcal{H}} \propto e^{-\chi^2_{\mathcal{H}}/2}$), this difference of log likelihoods corresponds to

$$-2\Delta \ln \mathcal{L} = -2\left[-\frac{1}{2}\chi^2_A + \frac{1}{2}\chi^2_B\right] = \chi^2_A - \chi^2_B,$$

with $\chi^2_B = 0$.

A binned maximum likelihood fit is used when the predictions for each bin are functions of some parameter α or a fixed number of parameters α_i $i = 1, 2, \cdots$. The most common statistical errors assigned to such parameters are the variations which correspond to changes in $-2\Delta \ln \mathcal{L}$ of one unit which correspond to changes in χ^2 of one unit. Similarly, 2σ error bars are assigned to variations which lead to changes of 4 units. However, the absolute value of the maximum likelihood does not lend itself to simple interpretation regarding the quality of the fit in the same way that χ^2/ν does for fits to high statistics samples. But judging the quality of a fit remains important.

We wish to determine whether a particular fit is consistent with the data. We can think of this as a hypothesis test, in which the null hypothesis is that the data follows the model in the fit, and the alternative hypothesis is everything else. We may implement this test by forming a likelihood ratio statistic, in which the numerator likelihood is the maximized likelihood according to the fit, and the denominator likelihood is the maximized likelihood over "all other" possibilities.

In our present case of the Poisson distribution, the best possible fit to the data will exactly predict the numbers of entries observed in each and every bin. One may thus compare the likelihood of the actual fit to this best possible fit:

$$\frac{\mathcal{L}_{\text{fit}}}{\mathcal{L}_{\text{best}}} = \prod_i \frac{\mu_i^{n_i}\, e^{-\mu_i}}{n_i^{n_i}\, e^{-n_i}}$$

for which

$$-2\Delta \ln \mathcal{L} = -2 \sum_i \left[n_i \ln \mu_i - n_i \log n_i - \mu_i + n_i \right] .$$

(Note that $n_i \ln n_i \to 0$ for $n_i = 0$ as $0^0 \equiv 1$.) If, as often is the case, the fit constrains the total number of events to the observed value, then this statistic tests the shape of the distribution. Models with external constraints may not satisfy this condition, in which case $-2\Delta \ln \mathcal{L}$ will describe the overall normalization of the model as well as the shape.

Intuitively, we see that the goodness-of-fit is best if the likelihood ratio is near one, and poorest if near zero. However, to make a quantitative test, the distribution of this test statistic must be known. To address this question, one may produce a series of Monte Carlo simulations and study the distribution of likelihood ratios. What is the probability that in an ensemble of Monte Carlo simulations one will observe a likelihood ratio better (or worse) than that observed in the data?

Some care must be taken in how the experiment is conducted and how the results are reported. One should allow the size of the Monte Carlo sample to fluctuate statistically, typically this may be done using the observed size of the sample as the mean value. If trying to describe a combination of background plus signal, both must be allowed to fluctuate independently. If the Monte Carlo data sets are then compared with the model based on the fit, the range of likelihood ratios (or $\Delta \log \mathcal{L}$) will account for the statistical fluctuations in sample size as well as for the shape. This procedure answers the question of whether the data is consistent with one particular hypothesis.

Similarly, one might ask whether the data is inconsistent with some null hypothesis. In this case one can create Monte Carlo ensembles of null hypothesis data and ask how often the signal hypothesis gives a better likelihood than does the null hypothesis. If the model has a free parameter, then one must also allow this parameter to be determined in fitting the Monte Carlo data samples. For example, if one is searching for a new state such as a charm meson or charm baryon resonance whose mass and width are not known *a priori*, the appropriate question will be how often one spuriously generates a signal with a reasonable mass and width. Assigning a statistical significance to the result of such a study is precise only if all the parameters to be considered, and the ranges considered acceptable, are specified before the search for the signal and the study of spuriously generated signals begins. Suppose, for example, one decides *posteriori* to consider only possible resonances with widths less than 40 MeV/c^2 rather than those with widths less than 60 MeV/c^2 because the possible signal was observed with a width 20$\pm$5 MeV/c^2. The evalutation of the number of number of expected spurious signals according to the 40 MeV/c^2 cut is problematic in such a procedure. This is because the sampling PDF may also have the possibility that one could have observed a potential signal with a width of 45 ± 11 MeV/c^2, in which case the cut might have been made differently. This lack of specificity at the beginning of the analysis will confound the statistical interpretation of the results.

6.2 Likelihoods in Particle Identification

Likelihood ratios are often used to identify reconstructed tracks as e , μ, π, K, or p. In this case, we have a discrete set of hypotheses, but the data may or may not discriminate between them. The goal of using likelihood ratios in this case is to provide the best discrimination possible between hypotheses and to quantify the relative consistency of detector response with each of the hypotheses. The following discussion has been extracted from *BABAR* Note 422 and modified slightly.

Probability Density Functions

The response of a detector to each particle species is given by a probability density function (PDF). The PDF, written as $\mathcal{P}(x; p, H)$ describes the probability that a particle of species $H = e^\pm$, $\mu^\pm$, π^+, π^-, K^+, K^-, p, $\overline{p}$; γ, K_L, n, $\overline{n}$ leaves a signature x described by a vector of measurements (dE/dx, E/p, DIRC angle $\ldots$). $\mathcal{P}(x; p, H)dx$ is the probability for the detector to respond to a track of momentum p and type H with a measurement in the range $(x, x + dx)$. As with any PDF, the integral over all possible values is unity, $\int \mathcal{P}(x; p, H)\, dx = 1$. Note that the momentum is treated as part of the hypothesis for the PDF and therefore is placed to the right of the semicolon. Drift chamber momentum measurements are usually of sufficient precision that they can be treated as a given quantity. In borderline cases when the precision is almost sufficient, it is sometimes treated by assuming that momentum is perfectly measured and smearing the PDF. For example, a Čerenkov threshold would become a Gaussian convoluted with the actual turn-on function.

The vector x may describe a single measurement in one detector, several measurements in one detector, or several measurements in several detectors. The measurements may be correlated for a single hypothesis. An example of correlated measurements within a single device is E/p and shower shape of electrons in the electromagnetic calorimeter. An example of correlated measurements in separate detectors is the energy deposited by charged pions in the electromagnetic calorimeter and in the instrumented magnetic flux return. In many cases of interest the correlations will be reasonably small and the overall PDF can be determined as a product of the PDFs for individual detectors. For example, the specific ionization deposited by a charged track as it traverses the drift chamber has almost no influence on the Čerenkov angles observed by the DIRC.

The difficult part of PID analysis is determining the PDFs, their correlations (if any) and understanding the uncertainties for these distributions.

Likelihood

Given the relevant PDFs, the likelihood that a track with measurement vector x is a particle of species H is denoted by $\mathcal{L}(H; p, x)$. The functional forms of PDFs and the corresponding likelihood functions are the same:

$$\mathcal{L}(H; p, x) \equiv \mathcal{P}(x; p, H) \tag{6.1}$$

The difference between $\mathcal{L}(H; p, x)$ and $\mathcal{P}(x; p, H))$ is subtle: probability is a function of the measurable quantities (x) for a fixed hypothesis (p, H); likelihood is a function of the particle type (H) for a fixed momentum p and the measured value (x). Therefore, an observed track for which x has been measured has a likelihood for each particle type. Competing particle type hypotheses should be compared using the ratio of their likelihoods. Other variables having a one-to-one mapping onto the likelihood ratio are equivalent. Two commonly used mappings of the likelihood ratio are the difference of log-likelihoods and a 'normalized' likelihood ratio, sometimes called 'likelihood fraction'. For example, to distinguish between the K^+ and π^+ hypotheses for a track with measurements x_{obs}, these three quantities would be written as:

$$\mathcal{L}(K^+; p_{obs}, x_{obs}) / \mathcal{L}(\pi^+; p_{obs}, x_{obs}) \tag{6.2}$$

$$\ln(\mathcal{L}(K^+;p_{obs},x_{obs})) - \ln(\mathcal{L}(\pi^+;p_{obs},x_{obs})) \tag{6.3}$$

$$\mathcal{L}(K^+;p_{obs},x_{obs})/(\mathcal{L}(K^+;p_{obs},x_{obs}) + \mathcal{L}(\pi^+;p_{obs},x_{obs})) \tag{6.4}$$

It can be shown rigorously that the likelihood ratio [Eqn. (2)] and its equivalents [Eqns. (3) & (4) & any other 1-to-1 mapping] discriminate between hypotheses most powerfully. For any particular cut on the likelihood ratio there exists no other set of cuts or selection procedure which gives a higher signal efficiency for the same background rejection.

There has been an implicit assumption made so far that there is perfect knowledge of the PDF describing the detector. In the real world, there are often tails on distributions due to track confusion, nonlinearities in detector response, and many other experimental sources which are imperfectly described in PDFs. While deviations from the expected distribution can be determined from *control samples* of real data and thereby taken into account correctly, the tails of these distributions are often associated with fake or badly reconstructed tracks. This is one reason why experimentalists should include an additional consistency test.

Goodness-of-fit

A statistical test for goodness-of-fit does not try to distinguish between competing hypotheses;[1] it addresses how well the measured quantities accord with those expected for a particle of type H. The question is usually posed, *"What fraction of genuine tracks of species H look less H-like than does this track?"* This is the prescription for a significance level. For a device measuring a single quantity and a Gaussian response function, a track is said to be consistent with hypothesis at the 31.7% (4.55%) significance level if the measurement falls within 1 (2) σ of the peak value. If the PDF is a univariate Gaussian,

$$\mathcal{P}(x;p,H) = \frac{1}{\sqrt{2\pi}\,\sigma(p,H)} \exp\left[-\frac{1}{2}\left(\frac{(x-\mu(p,H))}{\sigma(p,H)}\right)^2\right] , \tag{6.5}$$

the significance level (SL) for hypothesis H of a measured track with $x = x_{\rm obs}$ is defined by

$$\mathrm{SL}(x_{\rm obs};H) \equiv 1 \; - \; \int_{\mu_H - x_{\rm obs}}^{\mu_H + x_{\rm obs}} \mathcal{P}(x;H)\,dx. \tag{6.6}$$

[1] At least, this is how physicists usually think of goodness-of-fit. It is, however, sometimes useful to think even here in the language of alternative hypotheses, in which the goodness-of-fit test is between the model of interest, and some broader model space. An example of this was given in Section 6.1.3.

Notice that the integration interval is defined to have symmetric limits around the central value. This is an example of a two-sided test. Mathematically, one may also define a one-sided test where the integration interval ranges from $x_{\rm obs}$ to $+\infty$ or from $-\infty$ to $x_{\rm obs}$. However, for a physicst establishing consistency, it is only sensible to talk about the symmetric, two-sided significance levels defined in the last equation when presented with a Gaussian PDF. This definition is equally sensible for other symmetric PDFs with a single maximum.

Nature is not always kind enough to provide Gaussian or monotonic PDFs. For example, asymmetric PDFs are encountered when making specific ionization (dE/dx) measurements. Multiple peaks in a PDF might be encountered when considering the energy deposited by a 1.2 GeV π^- in an electromagnetic calorimeter. Although the π^- will typically leave a minimum ionizing signature, some fraction of the time there will be a charge exchange reaction ($\pi^- + p \to n\pi^0$) which deposits most of the π^- energy electromagnetically. A particularly useful generalization of the significance level of an observation $x_{\rm obs}$ given the hypothesis H is defined to be;

$$\mathrm{SL}(x_{\rm obs}; H) \;=\; 1 - \int_{\mathcal{P}(x;H) > \mathcal{P}(x_{\rm obs};H)} \mathcal{P}(x;H)\,dx \tag{6.7}$$

Although we define the consistency in terms of an integral over the PDF of x, note that the range(es) is(are) specified in terms of the PDF, not in terms of x. This allows a physically meaningful definition. While other definitions of significance level are possible mathematically, we recommend that *BABAR* use only this definition.

Note that because the PDF is normalized to 1, the significance level can be defined equivalently as

$$\mathrm{SL}(x_{\rm obs}; H) \;=\; \int_{\mathcal{P}(x;H) < \mathcal{P}(x_{\rm obs};H)} \mathcal{P}(x;H)\,dx \tag{6.8}$$

All significance levels derived from smooth distributions of the true hypothesis are *uniformly* distributed between 0 and 1 (as are confidence levels). This can be used to test the correctness of the underlying PDF using a pure control sample.

Using significance levels to remove tracks which are inconsistent with all hypotheses takes a toll on the efficiency (presumably small), and may also discriminate between hypotheses. In general, if a cut is made requiring SL $> \alpha$, the *false negative* rate, or *type-1 error*, is α. This is identical to the statement that the *efficiency* of this cut is equal to $1 - \alpha$. The *false positive* rate, or *type-2 error*, $\beta(H)$ can depend on the definition of the SL, *i.e.*, on the design of the test, and is identical to the *misidentification probability.* The *background fraction* in a sample is the sum $\sum \beta_i \mathcal{P}_{\mathcal{A}i}$, where $\mathcal{P}_{\mathcal{A}i}$ is the fraction of particle i in the sample.

Consistencies control only the efficiency. Minimizing background, however, depends on the type of sample. A fixed cut on consistency will produce very different background rates depending on whether we are looking at $\tau^+\tau^-$ or $B\overline{B}$ events, if we are just interested in a certain momentum range, if we cut on multiplicity, or if we consider only tracks from a secondary vertex. That's why each analysis will have its own optimum way to do PID.

Mis-using Consistencies

Any procedure for combining measures of consistency such as *confidence levels* or *significance levels* must be arbitrary with an infinite number of equally valid alternatives. For example, the method proposed in reference [12] is mathematically equivalent to the following recipe:

- use the inverse of $CL = P(\chi^2|2)$ to convert each of n probabilities CL_i into a χ^2_i
- add them up, i.e. $\chi^2 = \sum_{i=1}^n \chi^2_i$
- use $CL = P(\chi^2|2n)$ to convert χ^2 into a new "combined" CL.

The arbitrariness of the method is immediately seen in this recipe, since equally "sensible" results would be obtained if 2 degrees of freedom were replaced by k dof for any integer k, and $2n$ were replaced by kn in the last step.

Probabilities

In the case (such as particle identification) where the *a priori* probabilities of the competing hypotheses are known numbers, $\mathcal{P}_{\mathcal{A}}(H)$, likelihoods can be used to calculate the expected purities of given selections. Consider the case where 7 pions are produced for each kaon. Then the fraction of kaons in a sample with measurement vector x is given by:

$$\mathcal{F}(K;x) = \frac{\mathcal{L}(K;x)\cdot\mathcal{P}_{\mathcal{A}}(K)}{\mathcal{L}(\pi;x)\cdot\mathcal{P}_{\mathcal{A}}(\pi)+\mathcal{L}(K;x)\cdot\mathcal{P}_{\mathcal{A}}(K)} = \frac{\mathcal{L}(K;x)}{\mathcal{L}(\pi;x)\cdot 7+\mathcal{L}(K;x)}\ . \tag{6.9}$$

This can be considered as a weighted likelihood ratio where the weighting factors are *a priori* probabilities. The $\mathcal{F}(K;x)$ are also called *posteriori probabilities*, *relative probabilities*, or *conditional probabilities*, and their calculation according to Eq. 6.9 is an application of Bayes' theorem 2.20. The purity, *i.e.*, the fraction of kaons in a sample selected with, say, $\mathcal{F}(K;x) > 0.9$, is determined by calculating the number of kaons observed in the relevant range of values of $\mathcal{F}$ and normalizing to the total number of tracks observed there, *e.g.*,

$$\text{fraction}(\mathcal{F}_H > 0.9) \;=\; \frac{\int_{0.9}^{1}\frac{dN}{d\mathcal{F}(H;x)}\mathcal{F}(H;x)d\mathcal{F}(H;x)}{\int_{0.9}^{1}\frac{dN}{d\mathcal{F}(H;x)}d\mathcal{F}(H;x)} \tag{6.10}$$

where the integration variable is the value of $\mathcal{F}(H;x)$.

Using likelihoods, consistencies, and probabilities

If PDFs [and *a priori* probabilities] were perfectly understood, using likelihood ratios [and the probabilities calculated above] to discriminate between hypotheses would suffice. However, the tails of distributions are likely to be unreliable. Some tracks will have signatures in the detectors that are very unlikely for any hypothesis. Others will have inconsistent signatures in different detectors, not in accord with any single hypothesis. We do not want to call something a K rather than π when the observed value of some parameter is extremely improbable for either hypothesis, even if the likelihood ratio strongly favors the K hypothesis. Extremely improbable events indicate detector malfunctions and glitches more reliably than they indicate particle species; they should be excluded. For many purposes, this can be done conveniently by cutting on the consistency of the selected hypothesis. If the PDFs are reasonably well understood, this has the additional advantage that it provides the efficiency of the cut.

Only in the case of a single Gaussian distributed variable do consistencies contain all the information to calculate the corresponding likelihood functions. There is a two-to-one mapping from the variable to the consistency and a one-to-one mapping from the PDF to the consistency. One can compute probabilities directly from likelihoods only because they are proportional to PDFs. To compare relative likelihoods, one must either retain the likelihoods or have access to the PDFs used to compute consistencies. If there is more than one variable involved, or the distribution is non-Gaussian, even this possibility evaporates; any consistency corresponds to a surface in the parameters space, and one cannot recover the values of the parameters or the likelihood, even in principle.

A simple example

Let us consider a toy system that is moderately realistic. Imagine a ring-imaging Čerenkov detector in which Čerenkov angles θ_C are measured with Gaussian resolution σ_C which we assume to be constant. If all the incident particles are known to be either kaons or pions of some fixed momentum, then the distribution of Čerenkov angles θ_C will consist of the superposition of two Gaussian distributions, one centered at the central value for kaons, θ_K, and one centered at the central value for pions, θ_π. Assuming the pion:kaon ratio is 7:1, and the separation between θ_K and θ_π is $4\,\sigma_C$, the θ_C distribution will look like that in Fig. 6.5. The PDF for the pion hypothesis is the normalized probability function

$$\mathcal{P}(\theta_C;\pi) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_C}\exp\left[-\frac{1}{2}\left(\frac{(\theta_C-\theta_\pi)}{\sigma_C}\right)^2\right]\;. \tag{6.11}$$

Similarly, the PDF for the kaon hypothesis is

$$\mathcal{P}(\theta_C;K) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_C}\exp\left[-\frac{1}{2}\left(\frac{(\theta_C-\theta_K)}{\sigma_C}\right)^2\right]\;. \tag{6.12}$$

The curves in Fig. 6.5 were produced by multiplying the pion PDF by 7 and adding it to the kaon PDF (and normalizing the total area under the curve to unity).

Using the observed Čerenkov angle, it is possible to calculate the relative probabilities of kaons and of pions at any measured θ_C:

$$\mathcal{F}(\pi;\theta_C) = \frac{7\cdot\mathcal{P}(\theta_c;\pi)}{7\cdot\mathcal{P}(\theta_c;\pi)+\;1\cdot\mathcal{P}(\theta_c;K)}\;. \tag{6.13}$$

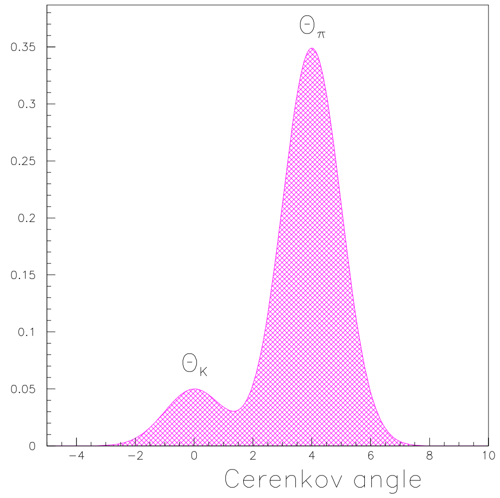

Figure 6.5: A possible Čerenkov angle distribution constructed assuming 7 times as many pions as kaons at a momentum where the nominal Čerenkov angles are separated by $4\,\sigma$.

Similarly, the kaon purity is

$$\mathcal{F}(K;\theta_C) = \frac{\mathcal{P}(\theta_c;K)}{7\cdot\mathcal{P}(\theta_c;\pi) + \ \mathcal{P}(\theta_c;K)}\ . \tag{6.14}$$

By construction, $\mathcal{F}(\pi;\theta_C)+\mathcal{F}(K;\theta_C) \ = \ 1$. To the extent that we truly know the *a priori* probabilities (given the other selection criteria used), and we have included all possible hypotheses in our analysis, the weighted likelihood ratios in this problem, $\mathcal{F}(K;\theta_c)$ and $\mathcal{F}(\pi;\theta_c)$ give the relative probabilities that a particle with the observed value of θ_c will be a kaon or a pion. Imperfections in these conditions must be evaluated as part of the systematic error.

The relative probability calculations for this problem are illustrated in Fig. 6.6. The blue curve shows the expected kaon fraction at each Čerenkov angle and the red curve shows the expected pion fraction. Several features of the distributions are worth noting. The pion and kaon PDFs are equal for θ_C exactly half way between θ_π and θ_K, but the pion probability is 7/8 while the kaon probability is 1/8. In effect, the Čerenkov detector has provided

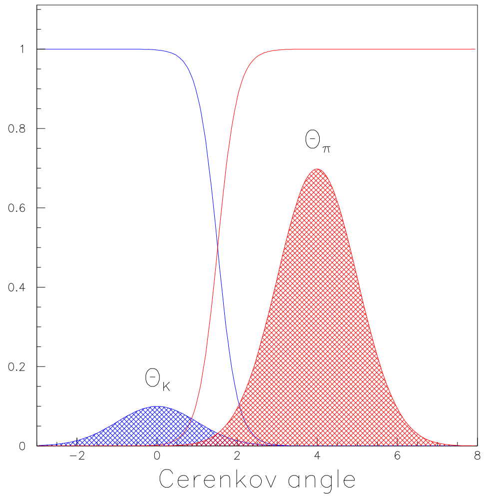

Figure 6.6: An illustration of the relative probability calculation for the Čerenkov angle distribution of Fig. 6.5.

no information and the final probability depends only on the *a priori* probabilities, despite the fact that the significance level for the kaon hypothesis is 4.55%, a value which is often acceptable.

6.3 Consistency of Correlated Analyses

We are sometimes confronted with the following issue: Suppose that we have taken a dataset consisting of a set of events. We do an analysis on this dataset, and obtain a result of interest. We then take the same dataset, and repeat the analysis, possibly with differences, to again obtain a result on the same matter of interest. How do we determine whether our two results are "consistent"?

Let us first attempt to ensure that we understand the question. Consider the limit in which we repeat the identical analysis. In this case, the result should be identical. If it isn't, there is an "inconsistency". Clearly, "inconsistency" here means "mistake" – the assertion that the analyses are identical must be incorrect. Now consider the limit in which the first analysis includes a random selection of one-half of the events, and the second analysis includes a selection of precisely the half not used in the first analysis. The two analyses are otherwise identical. In this case the difference between the two results should be of purely statistical origin. "Consistency" in this case is somewhat less trivial to determine, because the results can differ due to statistical fluctuations. However, a conclusion of "inconsistency" is still an assertion that a "mistake" was made, up to whatever small probability is permitted for a large statistical fluctuation.

The key point is that we are trying to evaluate whether any observed difference in the two analyses is an indication that there is something wrong with one or both analyses, or whether the difference is consistent with being due to statistical fluctuations.

6.3.1 How to Evaluate Consistency

Let θ stand for a physical parameter of interest. Let $\hat{\theta}_1$ be the estimated value of θ in the first analysis, and $\hat{\theta}_2$ be the estimated value in the second analysis. The most obvious statistic for evaluating consistency between the two analyses is $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$. We need the pdf, $p(\Delta\theta)$, for this difference in order to make a test for consistency. Given this, we simply compute:

$$\alpha = 1 - \int_{-\Delta\theta}^{\Delta\theta} p(x)dx.$$

α is the probability that we would observe a difference greater than the observed difference, due to statistical fluctuations. If α is bigger than some number, then we conclude all is well, if not, we worry that something is wrong.

Actually, we have made a simplifying assumption here: If $p(x)$ depends on other parameters, including θ itself, then the analysis is complicated by the need to know these parameters. In the case that θ is a location parameter for $\hat{\theta}_1$ and $\hat{\theta}_2$, then there is no dependence of $p(x)$ on θ. As this is often at least approximately the case, our simplifying assumption is of interest. We'll stick to this assumption here, while recognizing that in practice additional complications may need to be treated.

The question we wish to address for now is: What is $p(x)$? In the first limiting case of the introduction, $p(x) = \delta(x)$. In the second limiting case, denote $q_i(\hat{\theta}_i;\theta)$, $i = 1, 2$ as the pdf's for the results of the two analyses. By statistical independence, we have $p(\Delta\theta) = \int_{-\infty}^{\infty} q_1(x)q_2(x + \Delta\theta)\,dx$. More generally, there may be correlation between the samplings, with a joint pdf $q(\hat{\theta}_1, \hat{\theta}_2; \theta)$. In this case,

$$p(\Delta\theta) = \int_{-\infty}^{\infty} q(x, x + \Delta\theta)\, dx.$$

6.3.2 Example: Normal Distribution

Suppose that our samplings are from a bivariate normal distribution with common mean:

$$q(x_1, x_2; \theta) = A \exp\left\{-\frac{1}{2}\left[(x_1 - \theta)^2 W_{11} + (x_2 - \theta)^2 W_{22} + 2(x_1 - \theta)(x_2 - \theta) W_{12}\right]\right\},$$

where W is the inverse of the covariance matrix. If the covariance matrix is known, then the consistency may be evaluated by a simple χ^2 goodness-of-fit test:

$$\chi^2(1 \text{ dof}) = (x_1 - \hat{\theta})^2 W_{11} + (x_2 - \hat{\theta})^2 W_{22} + 2(x_1 - \hat{\theta})(x_2 - \hat{\theta}) W_{12},$$

where $\hat{\theta}$ is the value of θ which minimizes the χ^2.

On the other hand, the covariance matrix may not be fully known, or may be a lot of trouble to estimate. In particular, we may not know the correlation coefficient (ρ). We can still make a test for consistency, though at a cost of power (probability of correctly deciding that the results are not the same) and/or at the cost of significance level (probability of incorrectly deciding that the results are different) due to the uncertainty in ρ, the correlation coefficient. The difference $x_2 - x_1$ is distributed according to a normal distribution with mean zero and variance $\sigma_{x_1}^2 + \sigma_{x_2}^2 - 2\rho\sigma_{x_1}\sigma_{x_2}$. Since ρ is bounded by $\rho \in (-1, 1)$, the variance of $x_2 - x_1$ is in the region $(\sigma_{x_1} - \sigma_{x_2})^2$ to $(\sigma_{x_1} + \sigma_{x_2})^2$.

If a test of the observed distance using the smaller variance gives consistency, then we may be confident that the results are consistent. If a test of the observed distance using the larger variance gives inconsistency, then we conclude that the results are inconsistent. This is as much as we can do in the absence of knowledge concerning ρ. However, we might at least know that the correlation is not negative, and thence be able to tighten the test. In this case, the maximum variance of the difference is $\sigma_{x_1}^2 + \sigma_{x_2}^2$.

6.3.3 Example: Branching Fraction

Assume that we have a dataset corresponding to N B decays. We wish to determine the branching fraction θ to a particular final state. We do two analyses, yielding samples of N_1 and N_2 events, with efficiencies ϵ_1 and ϵ_2, respectively. For simplicity, we assume that the uncertainties in N and ϵ_i are negligible, and that there is no background contribution. The mean number of events expected in each of these analyses is $\langle N_i\rangle = \theta\epsilon_i N$. The two estimates of the branching fraction are given by:

$$\hat{\theta}_i = \frac{N_i}{N\epsilon_i}, \quad i = 1, 2.$$

If we treat N as fixed, we may model our samplings according to the multinomial distribution. Let N_{12} be the number of events which are common to both samples N_1 and N_2, with corresponding efficiency ϵ_{12}. Thus, $\langle N_{12}\rangle = \theta\epsilon_{12}N$. Note that $\max(0, \epsilon_1+\epsilon_2-1) \leq \epsilon_{12} \leq \min(\epsilon_1, \epsilon_2)$, and that statistically independent sampling corresponds to $\epsilon_{12} = \epsilon_1\epsilon_2$. Define $\bar{N}_i \equiv N_i - N_{12}$, *i.e.*, the number of events that are selected in analysis "i", but not in the other analysis. Also let $\bar{\epsilon}_i$ be the efficiency for an event to be selected in analysis i but not in the other analysis. We thus divide our entire sample of N events into four disjoint sets, with a partitioning given by:

$$P(\bar{N}_1, \bar{N}_2, N_{12}; \theta) = \frac{N}{\bar{N}_1!\bar{N}_2!N_{12}!(N - \bar{N}_1 - \bar{N}_2 - N_{12})!}(\bar{\epsilon}_1\theta)^{\bar{N}_1}(\bar{\epsilon}_2\theta)^{\bar{N}_2}(\bar{\epsilon}_{12}\theta)^{\bar{N}_{12}}[1 - (\epsilon_1 + \epsilon_2 + \epsilon_{12})\theta]^{(N-\bar{N}_1-\bar{N}_2-N_{12})}.$$

We typically may use the Poisson limit:

$$P(\bar{N}_1, \bar{N}_2, N_{12}; \theta) = \frac{\mu_1^{\bar{N}_1}}{\bar{N}_1!}\frac{\mu_2^{\bar{N}_2}}{\bar{N}_2!}\frac{\mu_{12}^{N_{12}}}{N_{12}!}e^{-\mu_1-\mu_2-\mu_{12}},$$

where

$$\mu_i \equiv \bar{\epsilon}_i\theta\langle N\rangle, \qquad \mu_{12} \equiv \epsilon_{12}\theta\langle N\rangle.$$

Our difference test statistic is

$$\Delta\theta = \frac{1}{N}\left(\frac{N_2}{\epsilon_2} - \frac{N_1}{\epsilon_1}\right) = \frac{1}{N}\left[\frac{\bar{N}_2}{\epsilon_2} - \frac{\bar{N}_1}{\epsilon_1} + N_{12}\left(\frac{1}{\epsilon_1} - \frac{1}{\epsilon_2}\right)\right].$$

Note that $\epsilon_i = \bar{\epsilon}_i+\epsilon_{12}$. The distribution of this test statistic may be evaluated according to the above probability distribution with Monte Carlo methods.

The observed difference may then be compared with the predicted distribution in order to evaluate consistency. As in the Gaussian case, it is possible that the "correlation" parameter, ϵ_{12}, may not be known, and a similar treatment to that for the Gaussian example will be necessary. In many cases, it will probably be reasonable to assume that $\epsilon_1\epsilon_2 < \epsilon_{12} < \min(\epsilon_1, \epsilon_2)$, *i.e.*, that the selection is not anti-correlated.

It may be remarked that the test here checks for consistency between the two results. It doesn't check for other possible problems, such as whether we have consistency with the expected overlap. Other tests could be devised to address such questions.

It is possible that, in the case where we don't know the correlation parameter, we can estimate it from other data available to us. That is, we may have sets of signal-like events and background-like events from our two analyses which can be used to estimate the relative values of $\bar{\epsilon}_1$, $\bar{\epsilon}_2$, and ϵ_{12}. These can then be used to evaluate whether the observed signal numbers show consistent behavior.

6.3.4 Example: Two Analyses on the Same Events

A case that causes some confusion is when the event selection is identical, but two analyses are performed on the selected sample. Because different information may be used in the two analyses, some variation in the results may be expected. As the measured information is in the form of random variables, this is still a problem amenable to statistical analysis.[17] The question being asked is still whether the observed difference is consistent with statistical fluctuations, versus the possibility that there is an inconsistency (mistake).

A concrete illustration is the following situation: Suppose that we have a set of events consisting of mass measurements, $\{m_1, \ldots, m_n\}$, of some resonance. Let the resonance mass be denoted θ. Assume for simplicity that the measurements are all made with Gaussian resolution functions, with possibly different widths, but all unbiased, and assume that the natural resonance width is negligible. We may form an estimate of the resonance mass by taking the sample mean of all the measurements:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} m_i.$$

This is an unbiased estimator for θ, since $\langle\hat{\theta}_1\rangle = \theta$.

Now suppose that we actually have, for each measurement, the resolution, σ_i, with which it is made. This additional information does not invalidate our estimator $\hat{\theta}_1$, but we can incorporate this information into another estimator:

$$\hat{\theta}_2 = \sum_{i=1}^{n} \frac{m_i}{\sigma_i^2} / \sum_{i=1}^{n} \frac{1}{\sigma_i^2}.$$

Again, we have an unbiased estimator for θ, since $\langle \hat{\theta}_2 \rangle = \theta$.

Both $\hat{\theta}_1$ and $\hat{\theta}_2$ are normally distributed, with moment matrix:

$$M = \begin{pmatrix} \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 & 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2} \\ 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2} & 1/\sum_{i=1}^{n}\frac{1}{\sigma_i^2} \end{pmatrix}.$$

Note that the form of this matrix is indicative of the fact that all of the information in the first analysis is used in the second analysis. According to our above analysis of the bivariate normal, the difference between the two estimators is distributed according to the normal distribution with standard deviation:

$$\sigma_{\Delta\theta} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 - \frac{1}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}}.$$

A simple χ^2 test can thus be used for checking consistency between the two results.

6.3.5 Dealing with Systematic Uncertainties

Typically our results have "systematic" uncertainties in addition to the statistical uncertainties we have so far been dealing with. In some cases, the systematic effect will be identical for both results. For example, both results might be based on the same estimate of the integrated luminosity. In this situation, there is no additional uncertainty in $\Delta\theta$ from this source.

On the other hand, it is possible that a systematic effect may be different in the two analyses. For example, the efficiency estimates in the two analyses may be made differently. This could lead to a "systematic" uncertainty in $\Delta\theta$ affecting our criteria for consistency. If possible, any common systematic should be separated out, leaving only the "independent" systematics, call them s_{x_1} and s_{x_2}. Then it is reasonable to assign a systematic uncertainty of $s_{\Delta\theta} = \sqrt{s_{x_1}^2 + s_{x_2}^2}$ to the difference, similarly to the result for uncorrelated statistical uncertainties. If it is too difficult to separate out the common systematics, then the best one can do is embark on a treatment similar to the discussion for the unknown correlation in the statistical errors.

Finally, a comment on what to do with the observed difference in the results of the two analyses. In general, the existence of two such results is a "fortuitous" circumstance – doing the two analyses is not a part of the experiment design. In particular, there is no plan that the purpose of doing the two analyses is in order to evaluate a "systematic uncertainty". The systematic uncertainties should be evaluated as appropriate in each analysis. The existence of more than one analysis may be used as a "check" for mistakes, but no new systematic uncertainty should be assigned to cover the difference between the two analyses.

Chapter 7

Systematic Errors

7.1 Introduction

The term 'systematic error' is generally taken as signifying any error not due to statistical fluctuations in data (although see Section 7.5.2.) This clearly covers a very large range of possibilities. Some attempt is made to categorise them here, to explain the different types that need different handling. In order to do this we have included many examples and illustrations. Some are invented but where we can we have used real instances from particle physics, especially from the types of analysis being done with *BABAR*.

7.1.1 Errors and Mistakes

A 'systematic error' is an *error*, as opposed to a *mistake*. We regard the expressions 'systematic error' and 'systematic uncertainty' as equivalent, in the same way that 'statistical error' and 'statistical uncertainty' are equivalent. One could argue that the term 'error' is ambiguous and should be dropped in favour of the unambiguous alternatives 'deviation' and 'uncertainty', as appropriate, but this is unrealistic.

We therefore disagree with an otherwise reputable statistics textbook[23] when it presents, as a paradigm of 'systematic error', measurements made with a steel rule by an experimenter who forgets to allow for thermal expansion. This is a mistake: a systematic deviation. When the experimenter realises that the measurements are faulty (either because they remember about thermal expansion, or because they find inconsistencies) then they apply a systematic correction, and a systematic error to cover the uncertainty in that correction.

So if results are suspect (because they disagree with the Standard Model, or the World Average, or amongst themselves...) then do not blame an 'unknown systematic error' but an 'unknown systematic effect'. This is further considered in section 7.2.3. (Speakers often cover themselves in such cases by talking about an 'unknown systematic', supplying the adjective but leaving the noun to be filled in by the audience. If you do this, you should at least be very clear in your own mind which you mean.)

7.1.2 The philosophical status of systematic errors

In many cases (not all) a systematic error is not a good frequentist concept. Usually there are exact values, we just don't know them.

> **Example 7.1:** V_{cb} **from the** b **lifetime**
> The lifetime of B hadrons clearly depends on the magnitude of the CKM element V_{cb}. It also depends on many other things, for example the mass of the b quark. This uncertainty can be folded into the value to give an estimate and error for the matrix element. However the b quark mass has no 'error' in the frequentist sense – all b quarks have the same mass, we just don't know exactly what it is.

In such an analysis even the most virulent frequentist [1] will bite the bullet and go Bayesian [18]. This is done with reservations, and the proviso that the extent of the uncertainty in the quantity be small, or at least make a small difference to the final result, so that changes to the prior distribution do not affect the result.

7.1.3 Combination of Systematic Errors

When you take several measurements, then any systematic effect shifts them all in the same way. Averaging has no benefit. This is perhaps responsible for the widespread and erroneous belief that "you can't add systematics in quadrature". You can add systematic errors using the standard undergraduate formula for the combination of errors, in exactly the same way as you add statistical errors, remembering to include the correlation term

[1] Probably a reference to some of the authors.

$$Cov(x_i, x_j) = \sum_k \sum_\ell \left(\frac{\partial x_i}{\partial y_k}\right)\left(\frac{\partial x_j}{\partial y_\ell}\right) Cov(y_k, y_\ell) \tag{7.1}$$

i.e. you use Equation 2.14 but you cannot simplify it to 2.16.

Another version of this myth says that that you can't add systematics because they're not Gaussian. In fact there is nothing in the combination of errors formula that requires a Gaussian distribution – variances add for any convolution – and indeed the Central Limit Theorem ensures that the final distribution is likely to be Gaussian, whatever the ingredients. (A Gaussian distribution is only necessary when you match deviations to probabilities: 68% within 1 σ and so on.) Another standard text [14] misleadingly recommends that systematic errors be added linearly, on the grounds that if you know the values are somewhere in the two ranges, you know the final value is within the sum of the two ranges: this is true for tolerances (as the engineers use) but inappropriate for errors as understood and used in physics.

Where the combination of errors formula does come unstuck, for statistical and systematic errors, is if the errors are large, in the sense that the first term in the Taylor expansion is not enough. With $\theta = 1.1$ it is valid to say $\sigma_{tan\theta} = \sigma_\theta / \cos^2\theta$ for $\sigma_\theta = 0.05$ but not for $\sigma_\theta = 0.50$! Such examples do occur, and are consided in section 7.5.3.

7.2 Finding systematic errors

Systematic errors can be divided into 3 categories depending on the techniques needed to find and evaluate them.

7.2.1 Straightforward Errors

Type 1 ('good') errors arise from clear causes, and can be evaluated. Calibration errors are a typical example: various calibration constants are applied to the data (affecting everything in the same way), and the errors on those constants are determined as part of the calibration procedure.

Example 7.2:Luminosity errors
In calculating the luminosity in any period, there is a statistical error from fluctuations in the numbers of small-angle Bhabhas measured by the luminosity monitor, and, as the cross section varies rapidly with angle, a systematic error from ignorance of the exact (effective) position of the luminosity monitor, which can be evaluated from survey errors and Monte Carlo simulation.

Example 7.3:Background distributions
In studying rare B decays [15] systematic errors arise from the accuracy with which the signal and background probability functions are known. We obtain these functions, such as the shape of the ΔE or resonance mass distributions, from auxiliary measurments taken on simulated or real data.

Their evaluation is discussed in section 7.4.

7.2.2 Uncertain errors

Type 2 ('bad') errors arise from clear causes but can not be evaluated.

In such cases there is no unambiguously correct procedure to determine the error. It is important to state clearly what values have been used, so they can be revised in the light of later developments if necessary, but one must acknowledge the imperfect nature of the result. This is not as bad as it may sound, because such errors are often a small component of the total systematic error, especially when added in quadrature. Don't sweat the small stuff.

Theory Errors

'Theory errors' are a typical example. A theoretical prediction is not precise: its accuracy is limited by simplifications in the model used, or in the number of terms taken in an expansion, or other mathematical approximations used.

Example 7.4:Luminosity errors again
Luminosity measurements also have a systematic error due to the accuracy of the theory: Bhabha scattering has only been computed to a certain order of diagram.

One tactic here is to ask some appropriate experts how accurate they think a calculation is. A consensus may emerge – or it may not. Theorists can be touchy: someone who has spent 6 months performing a complicated calculation within the framework of a particular model does not respond well to questions about the validity of the model. If they do produce a value it can be very conservative – a tolerance rather than a standard deviation.

Asking several theorists for an assessment of the accuracy of a prediction is an interesting and enlightening experience, but probably won't help you get the error estimate you wanted.

Another method is to take several typical but different values or models and compare the results. The standard deviation of the results then gives you an estimate of the error.

Unfortunately circumstances may restrict you to a few results (even two).

Two theories

If you have two cases which you know are extreme (e.g. no mixing or full mixing) with the true value anywhere between them, then it is allowable to take the mean of the two results as the central value, and the difference over $\sqrt{12}$ as the error.

> **Example 7.5:A 24 hour window**
> If a daily check shows that a channel was working yesterday but was dead at the same time today, then (in the absence of other information) the best estimate for the dead time is 12 ± 6.9 hours.

This is a popular technique as $\sqrt{12}$ is quite a large number, but it should not be abused. Its validity is restricted to cases where you really have two extreme models (not just any two models) and where the truth can plausibly be said to lie anywhere in between. If the two models are extreme, and with equal (subjective) probability either one or the other is true, then the error would be the difference over 2.

If you believe them to be 'typical', i.e. random samples from a space of models, then you should use the difference divided by $\sqrt{2}$, i.e. the rms difference from the mean is corrected for bias by the usual $\sqrt{\frac{N}{N-1}}$ factor.

Example 7.6:ρ polarisation

In the measurement of the branching ratio $B \to J/\psi\rho$ [24] the acceptance depends on the polarisation with which the ρ is produced. This could be completely transverse or completely longitudinal or anything in between. The efficiency used is the average of those for transverse and longitudinal polarisation, and the error the difference divided by $\sqrt{12}$

Example 7.7:The mass of the (1700)

ALEPH [21] measure the mass of the ρ' in $\tau^{\pm} \to \pi^{\pm}\pi^{0}\nu$ decay. Using the Kühn and Santamaria model gives a mass of 1363 ± 15 MeV. Using the Gounaris and Sakurai model gives 1400 ± 16. They quote a combined result of 1380 ± 24.

We do not recommend this, preferring to quote the two values separately. [20].

7.2.3 Errors from unknown causes: Checks

Type 3 ('ugly') errors arise from sources that have been overlooked. Because the causes are unknown the errors are unquantifiable. However the existence of such causes may reveal itself through consistency checks. These often take the form of changing something which should in principle not affect the result, and also of evaluating a similar result where the value is known. Satisfactory checks will give confidence in the result, both for the authors and others.

Example 7.8:Expanding steel ruler

Measurement was made with a steel rule, without realising that the temperature is different from that at which it was calibrated. The experimenter checks that measurements in the morning and afternoon are consistent. They find a difference – and realise that mornings are cool and afternoons are warm.

Some checks are specific to an analysis, for example measuring the CP a symmetries in channels where it must be zero. Others are standard: vary histogram binning, fit backgrounds with different parametrisations, and use Maximum Likelihood fits instead of counting techniques (or vice versa).

The purpose of these checks is *not* to find systematic errors. It is to find mistakes, (or, to look at the other side of the coin, to give confidence in the absence of mistakes). Traditional practice has confused them with the estimation of known systematics, perhaps because the methods used are similar: run an analysis, change something, run it again and see how much the result shifts. It is important to distinguish between cases where one legitimately expects the change to affect the result ('educated checks') and cases where the result should be stable against the change ('blind checks').

It would obviously not be correct to adjust the measurement and its error for every consistency check that has been made, certainly if the inconsistencies are within the statistical uncertainties of the checks. Adding many systematic errors of this type will increase the total systematic error of the measurement unjustifiably, and penalise the virtuous, especially since many consistency checks are statistically limited. One should correct for (untraceable) mistakes, but not for statistical fluctuations.

Illustrative and totally fictitious example

To elaborate, and highlight the pitfalls of the common (ab)usage, consider the following example:

Suppose BaBar measures an asymmetry of 0.50 ± 0.02 in some channel. As a sensible check, it does the same analysis for a channel for which the result must be zero. Unfortunately it comes out (despite thorough re-checking) as 0.10 ± 0.01. Clearly there is a 'mistake' at work.

It is decided (for political reasons) not to suppress this measurement; it has to be used.

This involves TWO stages

First the *value* is adjusted. By how much depends on the judgement of the experts and is not a question for statistics.

- It could be that these channels are so similar that a bias of 0.1 in one means there is a bias of 0.1 in the other, and the measured asymmetry is shifted to 0.40.

- It could be that a mistake of +0.1 in one channel could manifest itself as -0.1 in another channel, and we leave the value at 0.50.

- There could be some sort of compromise, perhaps at 0.45. The bias is negative, but 0.1 is judged excessive.

Secondly the *error* is adjusted. This is linked to the value adjustement.

- If we believe the mistake is nothing but a bias measurement, our answer is $0.40 \pm 0.02 \pm 0.01$ (though one would probably inflate that 0.01 a bit!)

- If the anomaly of 0.1 is taken as an estimate of the magnitude of a mistake one would probably ascribe an error of that value, giving $0.50 \pm 0.02 \pm 0.10$. (Even though the correction in this case is 0, it is still there)

- If an adjustment of half the anomaly was made as a compromise, that would probably deserve a largish error, perhaps equal to the shift, giving $0.45 \pm 0.02 \pm 0.05$.

The 0.10 error in the second instance happens to be equal to the original 'safety check' measurement value of 0.10, but as you can see in this drawn-out example they are NOT inevitably the same. They are different quantities: one is the mistake (or, at least, the evidence for it). The other is the (systematic) error on the correction applied to compensate for that mistake. With another method of correction they can be different. This is what is glossed over in the usual 'Make lots of checks and fold the discrepancies into the systematic errors' graduate thesis approach.

Evaluating the result of a check

When a check is done the result is examined for a possible effect. In the above artifical example there is a 10σ effect and so no question about whether the 'mistake' really existed. In real life we have effects at the 1 to 3 σ level and have to decide what to do.

At what point action is triggered depends not only on the significance of the effect but also on the number of checks being done (if you work hard and do 20 checks, one should give a 2σ deviation.) It also depends on the *a priori* likelihood of an effect being present. If one has good argument that a potential bias may lurk behind a well defined effect, and even if the check leads to a small apparent bias, but with a large error, it is still sound to correct for the small bias (to be nice) and to add a large(r) uncertainty. For blind checks, if nothing significant is seen, you should just forget it: in that case 'significant' could mean a 3 sigmas effect. For educated checks, you correct for the bias whatever the statistical precision is, and increase accordingly the systematics.

If the analysis passes a blind check, then all you do is tick the box and move on. (You may wish to add a line to the publication “We have checked that the result is not sensitive to...”) If it fails then the problem has got to be studied. Possible results are (in a rough order of preference)

- A bug in the check is found and fixed. The analysis now passes the test.
- A bug in the analysis is found and fixed. The result shifts. The analysis now passes the test.
- Mature consideration convinces you that the result could depend on this change after all. This now becomes an educated check
- You convince yourself that this is a statistical fluctuation and declare that the blind check was passed after all
- The result is not published
- An error is ascribed to ‘unknown effects’ is guessed at and incorporated in the systematic error.

The significance of a check

A check which consists of a different analysis of the data (for example, using different histogram bins) involves comparing the two results. The difference between these results will not be zero (as the techniques are different, after all) but should be ‘small’. Here ‘small’ cannot be defined with respect to the statistical error, as the results use the same data. The deviation can be compared with the difference in quadrature between the errors from the two methods, which gives reasonably accurately the expected error on the deviation.

Actually the error on the difference is given by the limits:

$$\sqrt{\sigma_1^2 - \sigma_0^2} - \sqrt{\sigma_2^2 - \sigma_0^2} \leq \sigma_{diff} \leq \sqrt{\sigma_1^2 - \sigma_0^2} + \sqrt{\sigma_2^2 - \sigma_0^2} \tag{7.2}$$

Where σ_1 is the larger of the two errors, σ_2 is the smaller, and σ_0 is the Minimum Variance Bound. If the better technique saturates the Minimum Variance Bound then the range decreases to nothing and $\sigma_{diff}^2 = \sigma_1^2 - \sigma_2^2$

as given. Care and precision are needed in this check, and it may be necessary/useful to give errors to more significant figures than appears strictly necessary. If the better measurement does not saturate the MVB then the range widens quite rapidly and this method may not be useful. Simulation with a toy Monte Carlo will work in such cases, though it is a significant amount of work.

Full simulation: a special check and a special case

A classic example of a 'consistency check' is the running of the whole analysis chain on Monte Carlo data to ensure that the values input for the parameter in question are returned as output. This should really be done for several values, covering the area of interest. Usually this is a pure consistency check (eg for $\sin 2\beta$ measurements) but there are some counterexamples when this is used as input; for example in the measurement of the W mass from hadronic decays, there are significant distortions to the value due to kinematic effects such as the assignment of particles to jets; the Monte Carlo has to be used to give this correction (the uncertainty on the correction is a systematic error).

If, as is generally the case, the analysis method is believed to be unbiassed then even if the result is of moderate significance(0-2 σ) there is a practice of correcting the final result for this bias and increasing the systematic error accordingly (by the Monte Carlo statistical error). This is an insurance to cover the possibility of an unknown mistake in a (often complex) measurement technique. This practice has the implicit assumption that the cause of the bias in the simulation is also present in data. It does not assume that there are no additional biases in data that are not present in the simulation. This 'consistency check' is special, because it tests the whole measurement technique and not just a small part of it. The corresponding error does normally not contribute much to the total error since usually the statistics of the simulation is significantly larger than the signal sample in data.

This is a blind check (of a special kind admittedly) so if it does not show any significant effect (3σ, say) consistency with the logic of the preceding section says says it may be dropped. However, consistency is "the hobgoblin of tiny minds ", as Emerson puts it, to take it into account is like taking a kind of insurance against mistakes: it is costly, but far less than being caught without insurance ! However this special status should be granted only to a simulation of the full analysis, including background effects. Other partial simulations may be more practical (due to eg lack of large background

MC samples) but are on a level with the other blind checks. We also hope this advice will encourage people to do their full check with large MC data samples so the uncertainty that gets added is small.

Example 7.9:First $\sin 2\beta$
In our first $\sin 2\beta$ publication an analysis is done on simulated data and shows that the value inserted into the Monte Carlo is then produced by the analysis and there is not evidence for bias. Nevertheless we assign a contribution of 0.014 to the systematic error.

7.3 Correlations

Correlations are especially significant in the handling of systematic errors for two reasons:

- Systematic errors affect all data in the same way, so the effects are correlated.
- The errors themselves are often correlated.

Example 7.10:Double Gaussian
A typical example of this is the fitting of a resolution by a double Gaussian

$$P(x) = \frac{1-\alpha}{\sqrt{2\pi}\sigma_c} e^{-x^2/2\sigma_c^2} + \frac{\alpha}{\sqrt{2\pi}\sigma_t} e^{-x^2/2\sigma_t^2} \tag{7.3}$$

where the 3 parameters: the core resolution, the tail resolution, and the fraction of the peak which is tail, are all strongly correlated. If these correlations are not included, the errors will be overestimated.

The correlation matrix between two independent variables is the unit matrix , between two completely correlated variables it is all ones. The covariance matrices have the same structure with appropriate values of σ.

$$\rho = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad V = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \qquad \text{independent} \tag{7.4}$$

$$\rho = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \qquad V = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \qquad \text{completely correlated} \tag{7.5}$$

Example 7.11:Average leptons
A leptonic efficiency is an average over efficiencies for muons and electrons. Two efficiencies η_μ and η_e are obtained from independent MC samples, with errors σ_μ and σ_e. They are then multipled by a factor $\eta_T \pm \sigma_T$ to account for tracking efficiency losses not in the MC.

$$E = \frac{E_\mu + E_e}{2} = \frac{\eta_\mu \eta_T + \eta_e \eta_T}{2} \tag{7.6}$$

The error can be evaluated as

$$\sigma_E^2 = (\frac{\eta_T}{2}, \frac{\eta_T}{2}, \frac{\eta_\mu + \eta_e}{2}) \begin{pmatrix} \sigma_\mu^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & \sigma_T^2 \end{pmatrix} \begin{pmatrix} \frac{\eta_T}{2} \\ \frac{\eta_T}{2} \\ \frac{\eta_\mu + \eta_e}{2} \end{pmatrix}$$

or equivalently as

$$\sigma_E^2 = (\frac{\eta_T}{2}, \frac{\eta_T}{2}, \frac{\eta_\mu}{2}, \frac{\eta_e}{2}) \begin{pmatrix} \sigma_\mu^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_T^2 & \sigma_T^2 \\ 0 & 0 & \sigma_T^2 & \sigma_T^2 \end{pmatrix} \begin{pmatrix} \frac{\eta_T}{2} \\ \frac{\eta_T}{2} \\ \frac{\eta_\mu}{2} \\ \frac{\eta_e}{2} \end{pmatrix}$$

An analysis can often be broken down into quantities which are either completely independent or completely correlated, with matrices accordingly diagonal or uniform, correlations zero or 1. When they are combined, then the partial corelation structure emerges.

Example 7.12:Using systematic errors in fits
Consider fit through two sets of x, y points. The values were measured with two different sets of apparatus (or the same apparatus under two different conditions) so all y measurements have individual error σ, however the first set of points share a common systematic error S. So do the second set of points

$$Cov(y_i, y_j) = \sigma^2 + S^2 \quad \text{if } i = j \tag{7.7}$$
$$= S^2 \quad \text{if } i \text{ and } j \text{ are in the same set} \tag{7.8}$$
$$= 0 \quad \text{otherwise} \tag{7.9}$$

This matrix is inverted and used in the χ^2 minimisation

7.4 Incorporating Errors

Errors can also be divided into:

- Those that can be incorporated algebraically (e.g. selection efficiency used to get from a measured branching ratio to the true value.)
- Those that are incorporated in the fit, though not in any simple algebraic form.
- Those that are incorporated in the fit and selection.

For these latter two the systematic error has to be found by changing value used in the fit (and, if necessary, selection) and ascertaining the change in the result. Usually one changes the value by $\pm 1\sigma$ and reads off the result: if the deviation to either side is the same this is a simple error, otherwise it can be quoted as an asymmetric error – see section 7.5.1. However there may be cases where one changes a different amount and then scales (e.g. change by $\pm 5\sigma$ and then take 0.2 times the shift, if the effect is small and susceptible to rounding errors) and/or take several values and observe the shape of the effect.

Example 7.13: $Sin2\beta$

For example, in the $\sin 2\beta$ meansurement, the Beam Spot Size and SVT alignment affect both the selection and the fitting. The B lifetime and the mass difference Δm_d affect the fitting but not the selection.

7.4.1 Errors on the errors

Suppose you want to evaluate the systematic uncertainty in a result R due to the uncertainty σ_p in some parameter p. You evaluate $\frac{1}{2}|R(p+\sigma_p)-R(p-\sigma_p)|$, having satisfied yourseklf that the deviations are reasonbly symmetric.

However the results $R(p)$ etc have themselves got statistical errors σ_R. How should these be incorporated.

You need to establish carefully what you're doing. If these 3 analyses are all performed on the same data, and σ_R comes from the statistical errors on this data, then these errors are totally (or strongly correlated) and do not need to be considered. Changing the value of a branching ratio in an analysis is an evaluation that fits this desctiprion.

On the other hand if the σ_R errors are due to finite Monte Carlo statistics on separate Monte Carlo ssamples, then this does represent a real uncertainty.

In principle this situation should never be important. If the MC statistics errors are not small compared to the data statistics errors then you need more Monte Carlo. (Or you might use intelligent reweighting rather than regenerating).

If this is not the case then as a final resort the uncertainty should be folded in in quadrature.

$$\sigma^2_{R:p} = \left(\frac{1}{2}\left(R(p+\sigma_p) - R(p-\sigma_p)\right)\right)^2 + \sigma^2_R$$

7.5 Quoting Errors

It is a general practice in Particle Physics (though not outside) to quote statistical and systematic errors separately, and there are excellent reasons for this.

In some cases the analysis may usefully give a total error by adding the two in quadrature, and there is no reason not to do this if it is done in addition to the standard format (which should always be the prominently presented result.)

Separate systematic errors

In some cases it may be helpful to a present contributions from different sources of systematic error separately, for example where there is a large effect from an unknown intermediate branching ratio which will probably soon be known more accurately, or where this result is expected to be combined with other results from this experiment or others, with correlated errors.. Such presentation will enable the error to be readily recomputed when the better numbers are available. However this should not be done as a general procedure, only with good reason.

7.5.1 Asymmetric errors

It often does happen that the error obtained from the likelihood or from changing an analysis parameter is not symmetric (Of course statistical errors may also be asymmetric, especially for small statistics.)

> **Example 7.14:Asymmetric Error**
> Changing the value of parameter a by $+\sigma_a$ changes the result from 10.0 to 10.5; Changing it by $-\sigma_a$ changes it to 9.8. The result is quoted as $10.0^{+0.5}_{-0.2}$.

When combining results with asymmetric errors add each side seperately in quadrature. There is no real theoretical justification for this - what you should do depends on the shape of the likelihood curves and all you really know about them is that they're not Gaussian - but there is no better alternative and it is a convention.

> **Example 7.15:Combining results**
> if $a = 10.0 \pm 1.0^{+0.3}_{-0.5}$ and $b = 20.0 \pm 1.0^{+0.4}_{-1.2}$ then $c = a + b$ has the value $30.0 \pm 1.4^{+0.5}_{-1.3}$

The sense of the shift should be reflected in the error quoted. This is needed in determining errors if this result is later combined with others where this systematic error is correlated between them .

> **Example 7.16:Another Asymmetric Error**
> Changing the value of parameter a by $+\sigma_a$ changes the result from 10.0 to 9.8; Changing it by $-\sigma_a$ changes it to 10.5. The result is quoted as $10.0^{-0.2}_{+0.5}$.

7.5.2 Systematics entangled in the data

While many systematic errors are established independently of the data, in some cases the data sample itself contains useful information about the systematic effect – and in extreme cases this may be the only source of information. In this last case the systematic error behaves essentially like a statistical error, the difference being that it is of no interest.

There are (at least) 3 ways used by practising particle physicists to define what is meant by a 'systematic error'

A An error not due to statistical fluctuations in the data.

B An error not due to statistical fluctuations in the data sample being studied

C An error not associated with statistical fluctuations in the measurement data for the parameter of interest in the data sample being studied.

Example 7.17:Mixed statistical and systematic errors
A sample of 100 events in a peak in a mass window comprises signal on top of an estimated 30 ± 6 background events.

If the background estimate is external (e.g. from a previous experiment, or from Monte Carlo) it can certainly be termed 'systematic' as it satisfies all 3 definitions. The overall estimate of the signal is just $70 \pm 10 \pm 6$

If the background is estimated by interpolating the level in the sidebands in this experiment, this satisfies B and C but not A. Some physicists would call it a systematic error and some would call it statistical. Our view is that, although this is not our recommendation, provided it is clear what is happening, there is no harm in calling this a 'systematic' error. The overall estimate of the signal is still validly given by $70 \pm 10 \pm 6$.

If the background and signal are estimated together using the data in the mass window (and some other discriminator variable) this satisfies (C) but not the other two. Analogy with the previous example would suggest that it can also be called a systematic error. However the signal estimate can no longer be written as $70 \pm 10 \pm 6$.

In a case like this, a combined error from a combined fit, the division into 'statistical' and 'systematic' is not particularly meaningful. We should recommend that in such cases the quoted result should be the combined error; if desired a form of words should be added, *e.g.*,

We observe 70 ± 15.9 *events; if the background were known exactly the error would reduce to 10.0.*

Example 7.18:Combined systematics and statistical

People have asked where the 15.9 came from. The actual value depends, of course, on the analysis. Suppose that as well as the mass we have some tagging variable x. A cut has been optimised on x and it turns out that 67.4% of signal events lie above the cut, and 32.6% below, while for background events the split is (by a coincidence) exactly the other way, 32.6:67.4.

Suppose further we actually observe 57 events above the cut and 43 below. Arithmetic then gives the signal and background numbers as

$$57 = 70 * 0.674 + 30 * 0.326 \qquad 43 = 70 * .326 + 30 * .674$$

The result is actually obtained from

$$N_S = \frac{57 * 0.674 - 43 * 0.326}{0.674^2 - 0.326^2}$$

These fractions are obtained from Monte Carlo and hence (!) known with complete certainty. The 57 and 43 are independent Poisson variables, so the total error on N_S is given by

$$\sigma_S^2 = \frac{57 * 0.674^2 + 43 * 0.326^2}{(0.674^2 - 0.326^2)^2} = 15.9^2$$

For example, in the $\sin 2\beta$ measurement, the Beam Spot Size ,SVT alignment, B lifetime and mass difference Δm_d are known from other parts of the data. The CP background and the tag misidentification probabilities are determined from the fit (to the combined CP+Mixing dataset) but are of no interest, and the Δt resolution is partly determined from other sources and partly determined from the fit to this data set.

7.5.3 Systematic Errors and Likelihoods

In many interesting and relevant cases with low statistics, a result cannot be expressed as a value with an error, but the full information is only expressible as a likelihood function L. The maximum likelihood estimator may well be chosen as a quoted value, but the one σ limits where the likelihood falls by $\frac{1}{2}$ are asymmetric, and the two σ limits where L falls by 2 are not twice the one σ limit. In such cases the full likelihood function is the only way to combine results from different experiments.

The location of the maximum of the likelihood, and the exploring of the function around it, is done numerically, for example with `MINUIT`[25].

When dealing with errors on more than one quantity, with the distribution given by a multidimensional Gaussian, with non-zero correlation, the error on a single quantity is found by projecting the distribution onto the appropriate axis – not, as might be thought at first sight, taking a slice through it (see section 5.2). The same is true for non Gaussian distributions: to get the likelihood as a function of one patrameter one has to integrate out all tghe othersi, not take a particular set of values.. Suppose that the probability of a result x is a function of the parameter(s) of interest a and parameter(s) of no interest b. Then the log likelihood is given by

$$\ln L(\vec{x}|a, b) = \sum \ln P(x_i|a, b) \tag{7.10}$$

For example, a could be the branching ratio to a rare decay channel, and b could be a property of the background distribution. a could be $\sin 2\beta$ and b could be the resolution in z.

If the probability distribution for the b parameters is known, typically as Gaussians with certain means and standard deviations, they can, in a Bayesian treatment7.1.2, be integrated over[15].

$$L(\vec{x}|a) = \int L(\vec{x}|a, b)P(b)\, db \tag{7.11}$$

The integral can be done by Monte Carlo techniques, generating N values of b with the right properties, $\{b_1 \dots b_N\}$, and using

$$L(\vec{x}|a) = \frac{1}{N} \sum_{k=1}^{N} L(\vec{x}|a, b_k) \tag{7.12}$$

as the likelihood whose properties are studied. This nicely includes our *a priori* information about the values b with the knowledge gained from the actual fit. If the systematic errors are known (from their determination) to be correlated, then the N values are generated by finding the rotation which diagonalises the covariance matrixV_b, and then applying that to a set of independent Gaussians to generate values with the correct correlation. (In some cases the σ values may be altered in the light of superior information, while maintaining the correlation matrix.)

An alternative technique is to include in the likelihood the χ^2 contribution for deviation(s) of the b parameter(s) from their central value(s) b_0 [22].

$$\ln L' = \sum_i \ln P(x_i|a,b) - \frac{1}{2}(b-b_0)^T V_b^{-1}(b-b_0) \qquad (7.13)$$

(which includes the case where b is known only from the data if V is zero).

In such cases the fit gives a set of likelihood contours, and the error(s) on the quantities of interest a are read off by projecting the distribution(s) in the uninteresting quantities b onto their axes.

It may be desired to separated this combined error (which is quite satisfactory) into statistical and systematic parts. This is a slightly questionable procedure. Neverthless the question 'What would the errors be on a if I knew b?' has (even though this may be impossible to achieve technically) a perfectly good answer: these are given by the intersection of the likelihood contours with the b axes, and can be found by fixing the b values. [22].

It is then possible to express the 'systematic error' as the difference in quadrature between the total error and the reduced error obtained by this means. However this is not a very useful quantity, and is of dubious validity as the errors are typically 'large' , invalidating the linear combination of errors formula. [19] and it is probably better to emphasise the total error in giving results, rather than present it as the sum of such parts.

7.6 Incorporating systematic uncertainty into a limit

It often happens that a limit is desired, but the presence of a systematic uncertainty complicates the analysis. We recommend taking an approach which is consistent with the more general recommendations of this chapter. Thus, we do not recommend some of the ad hoc approaches that have sometimes been employed (*e.g.*, evaluating the limit using a correction term with a $\pm 1\sigma$ in the systematic). While not recommended, such methods are not necessarily "bad", and indeed can be used as a check on the robustness of the result – if the result shows great sensitivity to the methodology, then this should be understood.

In the typical, well-behaved (*i.e.*, approximately normal) case, the statistical and systematic uncertainties are added in quadrature, according to the general recommendation at the beginning of this chapter. The resulting uncertainty is then treated as a single normal error in obtaining the limit.

A more difficult, but still common situation is the low statistics case, where the statistical uncertainty should be treated with Poisson errors. This uncertainty will typically dominate over the systematic uncertainty, but the systematic should be incorporated into the limit, unless it is completely negligible. Again assuming that a normal approximation for the systematic is reasonable, the recommended procedure is to fold the Poisson distribution with a normal distribution for the systematic uncertainty.

This is a (see section 7.1.2) mix of frequentist and Bayesian argument. It is the approach taken by Cousins and Highland [16] who expand the gaussian integrals and obtain

$$L' = L(1 + \frac{L-N}{2}\sigma_r^2)$$

where L is the upper limit obtained without considering systeamatic errors, N is the number of events on which it is based, and σ_r is the *relaitive* systematic uncertainty in the efficiency/acceptance factor.

Example 7.19:Cousins and Highland in use

If you observe 3 events then the 90% upper limit has the value 6.68. If the branching ratio to be obtained from this limit through $B = N/S$ is also uncertain due to a 10% uncertainty in the scaling factor S then according to the formula the limit value increases to 6.81.

For example, suppose that we are estimating a branching fraction by counting decays into the desired channel (*e.g.*, $B^0 \to \gamma\gamma$):

$$\hat{B} = An, \tag{7.14}$$

where n is the number of observed decays, and A is the appropriate factor for efficiency and number of parent particles. A is evaluated with some (systematic) uncertainty. Presuming a normal approximation is a reasonable model for this uncertainty, the overall pdf for random variables A and n is:

$$p(n, A; B) = e^{-B/A_0}\frac{(B/A_0)^n}{n!}\frac{1}{\sqrt{2\pi}\sigma_A}\exp\left[-\frac{1}{2}\left(\frac{A-A_0}{\sigma_A}\right)^2\right], \tag{7.15}$$

where B is the actual branching fraction, and A is sampled from an $N(A_0, \sigma_A)$ distribution.

A prescription for obtaining a 90% confidence level upper limit on B is then

1. Pick a B (arbitrarily, but near the expected limit).
   (a) Generate a value A_0 according to $N(A, \sigma_A)$.
   (b) Generate n according to a Poisson with mean B/A_0.
2. Ask how often the n thus sampled is larger than the observed value of n in the experiment.
3. If this probability is 0.10, averaged over many samplings for A_0, then quote B as the 90% confidence level upper limit on the branching fraction.
4. Iterate, until the desired value of B is obtained.

This can be achieved through a Java program on the Statistics Group web page.

Chapter 8

Bayesian Statistics

8.1 Introduction

In the previous chapters, we have largely been concerned with the analysis and presentation of the information content in the data, of relevance to any chosen physics topic. Of equal importance is the issue of how to use this information to make inferences concerning physically interesting conclusions. This is the subject of the current chapter, and a shift in thinking is required to address the transition in goal. Thus, while frequentist statistics has been appropriate so far, as a means of summarizing information content, it is inadequate to the task of making decisions based on that information. This new enterprise is the domain of Bayesian statistics, in which it is attempted to describe what we think to be the degree of "truth" of a statement about physics.

For example, we are taking data which contains information concerning the value of $\sin 2\beta$. We might summarize such data with a central value and confidence interval (leaving off issues of systematic uncertainties for simplicity here): $\sin 2\beta = A \pm B$. Perhaps we'll get the result that $\sin 2\beta = 1.1 \pm 0.4$ in our experiment. A question of physical interest, given the result, is: "Is $\sin 2\beta$ different from 0?" It is procedures to answer questions of this sort that the present chapter addresses.

It can be argued that these issues needn't be addressed here. The point-of-view could be taken that the duty is done once the information in the measurement is presented. It is left to the reader to decide on the physical implications. This is not an unreasonable attitude, and can be taken as defining the minimum which should be presented. However, the reader is probably interested in knowing the conclusions of the experimenters, who are in many ways best-equipped to provide the interpretation of the results. In any event, they can't resist doing it, hence it is important to try to do it with methodology that has desirable properties.

We note that we are adopting the "subjective Bayesian" interpretation. It is also possible (perhaps discarding the notion of degree-of-belief), to add some additional principles to guide the choice of prior distribution. This is the "objective Bayesian" approach, which we do not discuss.

8.2 General Methodology

The basic tool to be applied to the decision problem is the Bayesian posterior probability, formed from the likelihood function and the Bayesian prior probability. If the problem concerns the value of an unknown parameter θ, this takes the form:

$$P(\theta;\{x\}) = \frac{L(\theta;\{x\})P(\theta)}{\int L(\theta;\{x\})P(\theta)d\theta},$$

where $\{x\}$ represents the measurements from the experiment. The quantity $P(\theta)$, called the "prior", is the Bayesian prior probability, *i.e.*, is the degree-of-belief distribution prior to the new results (in the subjective Bayesian interpretation).

8.3 Choice of Priors

The posterior degree-of-belief depends on both the new information and on the prior distribution. If there is previous relevant data, or reliable theoretical constraints, these can be included in the prior distribution.

For example, we may be making a new measurement of a parameter, which has been measured previously by sampling from a normal distribution for which the parameter of interest was the mean. If this constitutes effectively the entire prior knowledge, the prior distribution is simply (actually, it really isn't quite so simple, but for practical purposes this is the thing to do):

$$P(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{(\theta - X_0)^2}{2\sigma_0^2}\right], \tag{8.1}$$

if the previous measurement was with standard deviation σ_0, and sampled a value X_0. If the new experiment also samples a result from a normal distribution, with mean equal to θ, and standard deviation σ_1, then, for a result X_1, the posterior distribution is:

$$P(\theta; X_1) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\theta - X)^2}{2\sigma^2}\right], \tag{8.2}$$

where X is given by the usual weighted mean:

$$X = \left(\frac{X_0}{\sigma_0^2} + \frac{X_1}{\sigma_1^2}\right) \Big/ \frac{1}{\sigma^2}, \tag{8.3}$$

and:

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma_1^2}. \tag{8.4}$$

It may be noted in this example that the posterior distribution is the same as the posterior distribution obtained by regarding the measurements in a different way: We could view the two experiments, with results X_0 and X_1 as instead a single experiment, sampling from a two-dimensional joint pdf. With a prior distribution which assigns a uniform probability on θ, the posterior distribution of Eqn. 8.2 is again obtained.

The example above gives a glimpse of one of the most controversial aspects of the Bayesian methodology. How do we specify the prior representing the degree of belief prior to any relevant data? How do we express "complete ignorance" of the value of a parameter? In the example, we chose a uniform prior in θ, assigning equal prior probabilities to each possible value (in this example, the result is an improper probability distribution, but this is not an essential complication, since the normalization is divided out).

One apparent problem with this choice is that it is not parameterization-independent: If we make this choice, but then decide we are interested in $1/\theta$, the prior for $1/\theta$ is not uniform. Another apparent difficulty is that our true prior degree of belief will typically incorporate other knowledge in a not-well-defined (and probably subjective) way, but so that a uniform prior is not an accurate representation. For example, we may be quite confident that θ is a "small" number compared with 10^{324}.

There is a large literature which deals with these issues, and some algorithms have been proposed, based on additional principles. However, these algorithms tend to have the difficulty that they are not distribution-independent: The prior distribution to be assumed depends on the sampling distribution of the subsequent measurement. This is philisophically troublesome, because it suggests that our prior degree of belief describing ignorance depends on the way we are going to do an experiment.

One way to address this difficulty is to recognize that we (always?) have some information or prejudice which shapes our prior degree of belief differently from "complete ignorance". Thus, a suitable group of individuals could get together and consider the implications of what is known to arrive at a "consensus prior". This is certainly not a rigorous procedure, but is perhaps the best that realistically can be done.

It may finally be remarked that, if the choice of prior within the realm of this discussion actually makes a significant difference in the posterior distribution, it may be an indication that the new data is of little relevance (*i.e.*, is non-informative) to the question.

8.4 When Do We Have a Signal?

The question: Is there a signal? is a yes/no question which can be phrased in terms of an hypothesis test ($H1$: There is a signal; *vs.* $H0$: There is no signal.). But it is important to recognize that the question is being asked here in the decision sense, *i.e.*, we are trying to give our degree-of-belief concerning the answer. Hence, a Bayesian treatment is appropriate.

If it is supposed that the possibility of mistakes has been dealt with, the question of significance in terms of statistics may be asked. The available tool is the Bayes posterior distribution, describing our degree-of-belief as a function of (for example) possible parameter values. Hence, we may evaluate a "Bayesian acceptance region", at significance level α, by finding the set A such that

$$\alpha = \int_A P(\theta; \{x\}) d\theta, \tag{8.5}$$

where

$$P(\theta \in A; \{x\}) > P(\theta \notin A; \{x\}). \tag{8.6}$$

It is then a matter of policy to choose a significance level at which a signal is to be claimed. Note that we may have chosen not to fold the degree of prior belief in a non-null result into the prior, although we could attempt to do this. Thus, the choice of significance level will depend on how surprising the result is. The point of this is that for a more startling result the chance that it is due to a statistical fluctuation is greater.

Part II

Recommended Procedures

Chapter 9

General Recommendations

9.1 General Principles

There are a number of general guidelines and principles which we list here:

1. Whatever methodology is used, it should be clearly presented, so that the reader can understand what was done. If the technique is based on some published material, a reference should be given.

2. It should be anticipated that someone may wish to compare and/or combine the result with other information. Thus, enough information must be given to permit this, *e.g.*, systematic uncertainties may need to be quoted separately.

3. Related to the preceeding item, is the recommendation that whatever "physically relevant" quantity is quoted as a measurement, it is usually desirable to also give as close as possible the "actually measured" quantity. For example, the number of events in some mass peak (and in the background underneath) should be quoted, as well as the branching fraction derived therefrom. For another example, the raw numbers of events (and estimated backgrounds, and potentially with and without efficiency corrections) used for the extraction of V_{ub} should be given, as well as the procedure used in going from these numbers to the value of V_{ub}.

4. It should be kept in mind that the sampling distribution is in general only approximately known, and there can be long "tails". Thus, robust

methodologies are desirable, even at the cost of sufficiency or power.

5. It is useful, at least in one's thinking, to separate the presentation of "results", from the discussion of their "implications". By "results" is meant the information content of whatever measurement has been performed. By "implications" is meant the relevance to statements about physics. See the next section for an expanded discussion.

6. An important general principle, but one which is sometimes difficult to follow, is that the "experiment design should precede the experiment". For example, the criteria for claiming a signal should be decided before looking at the mass plot. Further, the temptation must be resisted to "tune" one's analysis until some desired result is achieved. While this principle is sometimes nearly impossible to practice, it should be kept in mind that not following it has sometimes led to disastrous results. An example of experiment design is the blind analysis.[2]

7. A corollary to the previous point is that we should avoid the potential bias introduced by basing the decision to publish on the result of an analysis. Again, there are practical issues in implementation, but the concern is real (the effect has been observed of published branching fractions starting out high).

9.2 Bayesian and Frequentist Statistics

Because of the substantial confusion among particle physicists concerning the subjects of "Bayesian" and "frequentist" (or "classical") statistics, a brief discussion is appropriate.

An intuitive way to think about these two branches of statistics is to regard the frequentist approach as being directed at summarizing relevant information content in a dataset; whereas the goal of the Bayesian is to infer something about the underlying sampling distribution. Thus, it is useful to connect the notion of "information" with the frequentist, and the notion of "decision" with the Bayesian. Of course, like any good decision-maker, the Bayesian uses whatever information is available, but the difference is that the Bayesian doesn't stop there. Instead, s/he proceeds to a conclusion corresponding to a "degree-of-belief" concerning the "truth" of some statement.

It may be argued that the first goal of an experimenter is to summarize the information contained in the data, as objectively as possible. The (frequentist) approach to achieving this goal consists of making some statement which has the property that it would be true in some specified fraction of trials (repetitions of the experiment), and false in the remaining fraction. It is important to realize that the frequentist typically doesn't know *or care* what the truth value is for any given sampling.

As the experimenter happens to also be a physicist, a second goal is to summarize what s/he thinks the "truth" really is, based on the information in the experiment, and perhaps other input. This is the domain of Bayesian statistics. While the first goal should be a requirement in any publication of results, the second may be regarded as optional – it can be left to the readers to form their own conclusions concerning the physical implications of the measurement, but the reader can't divine the information, so that must be presented.

It may be noted that people have sometimes proposed methodologies which attempt to satisfy Bayesian urges (*e.g.*, never make a statment that is known to be in the "false" category), while maintaining frequentist validity. This may be a noble goal, but the resulting methodologies are not especially attractive for various other reasons, and such algorithms are not advocated here. Instead, it is simply admitted that there is more than one goal, and that different techniques may be optimal for each.

The Statistics Working Group has tried to keep these notions in mind in the recommendations that follow. Thus, the next several chapters deal primarily in frequentist statistics, and the final section, on interpretation of results, deals more in Bayesian statistics.

9.3 Notation

The notation "χ^2" should be reserved for random varibales which are distributed according to the χ^2 probability distribution, or for the quantity which is minimized in a least-squares fit, if that quantity is distributed to a reasonable approximation by the χ^2 distribution. Use of the χ^2 notation in other instances should be avoided. In particular, using this notation as a substitute for a general $-2\ln\mathcal{L}$ quantity should be avoided. Note that a χ^2, as a sum of squares, can never be negative.

9.4 Number of Significant Places

When quoting a result, it can be cumbersome and even misleading to give too many significant digits. The following are some recommended procedures in this area. Where the Particle Data Group has a policy, we are adopting it below [26].

1. Results and errors should be quoted to a consistent number of places.

2. Error estimates should be given to two significant digits if less than 0.355 (up to the appropriate factor of ten), and one digit otherwise. For examples, we should quote:

$$1.97 \pm 0.33$$

   and also:

$$2.0 \pm 0.6$$

   . If the errors are asymmetric, they should be quoted to consistent places, and the smaller error prevails in determining how many. For example:

$$1.97^{+0.61}_{-0.33}.$$

3. Limits should be quoted to one or two digits based on the same criteria.

Chapter 10

Analysis Design – Recommendations

10.1 Introduction

Being able to rely on statistical procedures depends on good design of an analysis. The design of an analysis can also have a large impact on how informative a result may be obtained. Thus, the purpose of the present chapter is to look at some of the issues in obtaining good analysis design.

Currently, this chapter is a place holder. The Statistics Working Group felt that recommendations in this area would be useful, but the initial effort was aimed at areas which were more directly included in its charge.

Chapter 11

Confidence Intervals–Recommendations

Interval estimation may serve two purposes:

(a) to summarize the information in the experimental result, or

(b) to summarize the physical interpretation of the result.

We recommend to stick completely to the frequentist methods of interval estimation in case (a), and to Bayesian interval estimation in case (b).

As a basic example, a fit result of $\sin 2\beta = 1.12 \pm 0.13$ should define a 68% confidence interval for $\sin 2\beta$ in the sense that among all experiments measuring this quantity, for 68% the true value is within the quoted interval, and for 32% it is outside. A 95% CL interval for the same result would (approximately) state $0.86 < \sin 2\beta < 1.38$.

On the other hand, we may want to give a 95% CL interval with a lower limit on $\sin 2\beta$ which includes the mathematical constraint that $-1 \leq \sin 2\beta \leq +1$. This is already an interpretation of the pure measurement, since an external constraint is added to the experimental information. Here, we suggest to use Bayesian methodology and calculate a lower bound l on $x = \sin 2\beta$ by

$$0.95 = \frac{\int_l^1 L(x|1.12, 0.13)dx}{\int_{-1}^1 L(x|1.12, 0.13)dx}$$

where $x = \sin 2\beta$ and L is the likelihood function evaluated for the fit result and error. We then give $l < \sin 2\beta \leq 1$ with 95% CL. Note that in this case, we have chosen to use a uniform prior, within the interval $[-1, 1]$. This methodology is elaborated in chapters 8 and 15.

11.1 Specific Recommendations

We list now a number of simple recommendations for determining and quoting confidence intervals:

1. If an interval is quoted without explicit definition, it should be a 68% confidence interval (or good enough approximation), in the frequency sense. Note that a one standard deviation interval is not necessarily a 68% confidence interval (the sampling distribution may not be normal).

2. "Two-sided" intervals should be quoted whenever possible. One-sided intervals ("limits") may optionally be quoted in addition. The confidence level of a one-sided interval should always be given, but common practice of quoting a 90% confidence limit should usually be followed.

3. If the Gaussian approximation for the sampling distribution is valid, evaluating confidence intervals either by finding the points where $\Delta\chi^2 = 1$ or where $\Delta \ln \mathcal{L} = -1/2$ is recommended. If the Gaussian approximation is not valid, then:

   (a) After consideration, it may be determined that the Gaussian approximation is good enough, and the above methods used.

   (b) After consideration, it may be concluded that the Gaussian approximation is not good enough. In this case, a calculation based on the sampling pdf should be made (perhaps with a Monte Carlo) in order to understand how to estimate an interval. A possible outcome is that the above methods may be applied, but with different values for the changes (*e.g.*, instead of a change of $1/2$ in $\ln \mathcal{L}$, perhaps the change corresponding to a 68% confidence interval will be found to be 0.7). Of course, a description should be given.

4. The "parabolic" errors produced by programs such as MINUIT may be used to estimate a 68% confidence interval if one is confident that they are a sufficiently accurate approximation. The same statement applies to use of the first-order formula for propagation of errors. Incorporation of possible correlations in multi-variate situations should be followed.

5. The method of finding an interval containing a given fraction (*e.g.*, 68%) of the area under the likelihood function is generally not recommended, though it can produce valid (in the frequency sense) confidence intervals in some circumstances.[27] This method is, of course, the recommended procedure if a Bayes interval is desired.

11.2 Discussion of Case of "Physical Boundaries"

Often we are interested in some physical parameter, such as a mass, which cannot take on certain values, *e.g.*, cannot be negative. Because of detector resolution, it may happen that the sampling distribution used in the measurement has a non-zero probability for values outside this physical region to occur. Of course, there is nothing "unphysical" really happening, but people sometimes get very concerned about quoting the result of a measurement in the "unphysical" region. In particular, general methods have been proposed (*e.g.*, Ref. [5]) which generate confidence intervals which lie entirely within the "physical" region. We do not recommend the use of such methodologies, but because of their acceptance by some researchers, we here explain our reasons.

First, the motivation to quote an interval in the so-called physical region is actually irrelevant in frequentist statistics, as it has to do with interpretation of the results rather than summarizing the information content. The desire to keep the interval in the physical region, as a reflection of what we believe to represent a probability statement about the "true" value of the parameter, leads us naturally into Bayesian statistics. If that is the goal, we should just admit it, and be Bayesian. It should be noted that the intervals obtained by the Feldman-Cousins (and similar) methodology will not, typically, be the same as the intervals obtained in a Bayesian analysis, even though the "physical" constraint is satisfied.

Second while the proposed methods do yield valid (frequentist) confidence intervals, they are in fact completely equivalent to the confidence intervals as we usually calculate them, in the sense that the information content is the same (and there exists a 1:1 mapping between them). However, as a means of summarizing the information, these proposed intervals are rather less straightforward. For example, if an observation is far into the "unphysical" region (it's not really unphysical, of course), then it seems counterproductive to obscure that fact.

Another objection that has been made to, *e.g.*, the Feldman-Cousins method, is simply that it is not, in general, especially easy to calculate. Finally, the objection has been made that this methodology raised further difficulties with respect to combining results.

Chapter 12

Hypothesis Tests–Recommendations

12.1 Significance of a "Signal"

The significance of a signal is best quoted as a probability for the observation to be consistent with no signal. It takes some judgement to decide what probability to evaluate for this, but in general the idea is to evaluate the probability that the no-signal model can give a "fluctuation" as large as, or greater than, that observed in the data.

We note that simply giving a number for a "significance level" may result in confusion. Thus, whenever a "significance" is quoted, the definition of the quantity used should be given.

If the normal distribution is a reasonable implication, then quoting significance as a "number of sigma" (deviation from the no signal hypothesis) can also be used. On the other hand, if the normal distribution is not a good approximation, then this should not be done, as it will tend to be misleading.

A common case is when there is a small number of events, where the Poisson distribution should be used. The significance should be calculated as the Poisson probability for the background to fluctuate to the observed number of events, or more. But this should not be applied blindly – the question should first be asked whether the background estimate is reliable or not. If the background estimate is deemed reliable, but the estimated value has a significant statistical uncertainty, then the probability distribution describing the background estimate should be folded into the computation of the significance of the "signal". Sometimes a Monte Carlo is the most convenient way to perform the calculation.

Another common case is when changes in a χ^2 variable, or changes in $-2\ln\mathcal{L}$, are used to determine significance. To use these methods properly, either it must be known that the normal approximation is valid, or alternatively, a caculation (*e.g.*, by Monte Carlo) must be performed to understand the distribution expected for these variables under the null (no-signal) hypothesis.

Methods involving integrals of the likelihood function should similarly be first understood probabilistically, or alternatively relegated to the domain of an explicitly Bayesian analysis (see Chapter 15).

12.2 Fits

Doing a fit to some data typically involves varying parameters until a "best fit" is obtained. A discussion of this is in place under the heading of testing hypotheses, because a model is involved, which may or may not be a valid description of the data sample.

1. When performing a least-squares fit to binned data, beware of programs such as PAW when fitting to low-statistics data. PAW's default calculation is incorrect; it ignores bins with no events. When performing such a fit to data which may have low bin contents, a safe procedure is to combine bins with especially low event counts, until a minimum count is reached. A reasonable minimum for Poisson data is to require at least seven events per bin. If this procedure is followed, then the χ^2 goodness-of-fit statistic will generally be a good approximation, making sure the number of degrees of freedom is properly computed.

2. A maximum likelihood fit to binned event data is preferable to a least-squares fit if the statistics are low.

3. If it is important not to lose the information from combining bins (or from using bins which are wide compared to interesting structure), the least-squares method is not appropriate (or possibly even a binned likelihood fit). Instead, an unbinned fit to the individual event distribution should be performed. The maximum likelihood method is typically an

excellent way to proceed in this case. Goodness-of-fit may be evaluated using a likelihood ratio statistic [see 6.1.3]. A Monte Carlo may be required to understand the sampling distribution of this statistic, however.

4. A graphical display of a fit result (*i.e.*, a fitted curve superimposed on data) is highly recommended. This provides a way to visually check that a sensible result is being obtained.

5. When using the results of a fit, it is important not to lose sight of assumptions which might have been made, such as restrictions on background shape. A corresponding evaluation of systematic uncertainties should be made.

6. When quoting a χ^2 goodness-of-fit statistics, the χ^2 and the number of degrees of freedom should both be given, so that the reader can evaluate the confidence level. Quoting χ^2/n_{DOF} is not recommended. Alternatively, the χ^2 confidence level may be given.

12.3 Consistency of Correlated Analyses

It often happens that two analyses for the same result, on overlapping datasets, exist. The results will usually not be identical, and the question arises as to whether the difference is consistent with being due to statistical fluctuations, or whether there is evidence for a problem with one or both of the analyses. We address this issue in the following recommendations (see section 6.3 for further discussion):

1. Often the results of the two analyses are expressed as estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. In this case, a suitable statistic to use in testing for consistency may be the difference, $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$. The test for consistency is based on the joint pdf for the two results, $q(\hat{\theta}_1, \hat{\theta}_2; \theta)$. The sampling distribution for $\Delta\theta$ is evaluated according to (there may be additional dependence of the distribution on parameters, including θ):
$$p(\Delta\theta) = \int_{-\infty}^{\infty} q(x, x + \Delta\theta)\, dx.$$

Then the consistency test, for any given observed value of $\Delta\theta$ is performed by computing:

$$\alpha = 1 - \int_{-\Delta\theta}^{\Delta\theta} p(x)dx.$$

α is the probability that we would observe a difference greater than the observed difference, due to statistical fluctuations. If α is bigger than some number then we conclude all is well, if not, we worry that something might be wrong. A value of 0.05 is suggested for this test. If multiple tests are being made, then that should be properly accounted for in the evaluation of overall significance.

2. In testing for consistency, proper accounting for any correlations should be performed. If this is deemed too difficult, or not worth the effort, then limiting cases might be considered as a crude check. For example, if the sampling distributions are approximately Gaussian and the correlation coefficient, while unknown, is believed to be non-negative, then the difference can be compared with the range of standard deviations permitted: $|\sigma_{\hat{\theta}_1} - \sigma_{\hat{\theta}_2}|$ to $\sqrt{\sigma^2_{\hat{\theta}_1} + \sigma^2_{\hat{\theta}_2}}$.

3. It may be possible to estimate the correlation by looking at the behavior in a related "control sample".

4. The systematic uncertainties should be evaluated as appropriate in each analysis. The existence of more than one analysis may be used as a "check" for mistakes, but no new systematic uncertainty should be assigned to cover the difference between the two analyses.

5. If there are systematic uncertainties which differ for the two analyses, these should be incorporated into the consistency test. If possible, any common systematic should be separated out, leaving only the "independent" systematics, call them s_{x_1} and s_{x_2}. Then it is reasonable to assign a systematic uncertainty of $s_{\Delta\theta} = \sqrt{s^2_{x_1} + s^2_{x_2}}$ to the difference, similarly to the result for uncorrelated statistical uncertainties. If it is too difficult to separate out the common systematics, then the best one can do is embark on a treatment similar to the discussion for the unknown correlation in the statistical errors.

Chapter 13

Graphical Presentation–Recommendations

This section deals with the graphical presentation of data and the results of an analysis, and in particular, provides recommendations for the presentation of statistical content. We also offer some general guidelines for clear presentation that go beyond statistical considerations.

13.1 How to Display Data

By "data", we mean a set of measurements $\{\vec{x_k}\}_{k=1}^{n}$ whose distribution is modeled in an analysis. The elements of this set often correspond to selected events, but could equally be anything countable. For example, candidates of some type (with possibly more than one per event), or even sets of events (*e.g.*, runs).

13.1.1 Histogram Errors

The most common graphical presentation of data is a histogram of the distribution of one variable $y = f(\vec{x})$. Each bin, labeled by the index j, has a full width, Δy_j, centered on y_j, and contains a non-negative integer number, n_j, of data elements. Figure 13.1 illustrates the following recommendations:

1. The horizontal axis label should describe the variable y (e.g., "Decay-Time Difference, Δt") and specify the units in which it is measured (*e.g.*, "ps").

2. The vertical axis label should specify the basic unit of the data (e.g., "Events") and how the number of events in each bin, n_j, is normalized in the plot (*e.g.*, "/ 0.1 ps"). In the case of equal-width bins, the normalization Δy, should generally be to the common bin width.

3. The contents of each bin should be represented with a point at $(y_j, n_j \cdot (\Delta y/\Delta y_k))$ and an error bar that represents the "1-sigma" (68% confidence level) interval of the expected number of events given the observed number.

4. Horizontal error bars are discouraged for equally-spaced histograms, but recommended for unequal bin spacing (this convention clearly signals the different approaches). Horizontal error bars should cover the full width of the bin when used. In some cases, it may be useful to plot the point at the horizontal center-of-gravity of the data, with an error bar given by the rms variation. If this is done, it should be clearly explained in the caption.

The calculation of an appropriate error interval, as recommended in point 3 above, deserves more attention. We assume that the probability density for observing n events in a bin as a function of the expected number, ν, is described by the Poisson distribution

$$P(n\,;\nu) = \frac{e^{-\nu}\,\nu^n}{n!}\ .$$

The left-hand side of Figure 13.2 shows this distribution for $n = 5$ and compares it with the Gaussian distribution of mean 5 and r.m.s. $\sqrt{5}$ that is often (incorrectly) assumed instead.

An error bar extending from n to $n \pm \delta n$ corresponds to a one-sided interval of confidence level

$$\alpha_{\pm} = \pm \int_n^{n\pm\delta n} P(n\,;x)\,dx = \frac{\pm 1}{n!}\,\left(\Gamma(n+1, n\pm\delta n) - \Gamma(n+1, n)\right)\ .$$

The right-hand side of Figure 13.2 compares the confidence levels of Poisson and Gaussian distributions for $n = 5$ as a function of $x = n \pm \delta n$. Note that the Gaussian confidence levels are symmetric but that the Poisson confidence levels are larger above than below n at any given δn.

Ideally, we would like $\alpha_- = \alpha_+ \simeq 34\%$. This generally leads to error bars that extend further on the low side than the high side, as shown in

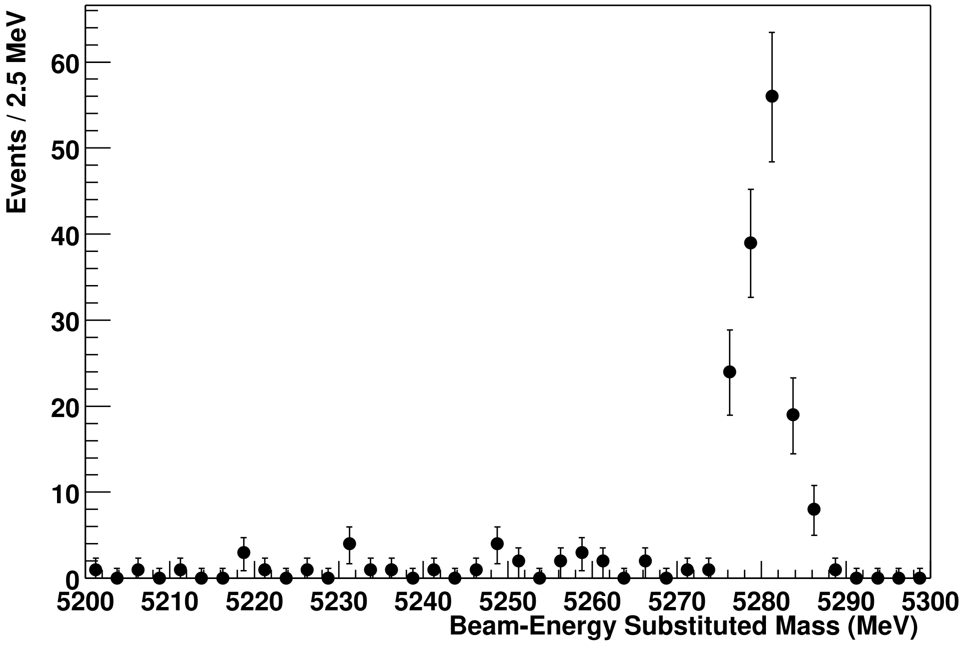

Figure 13.1: An example histogram demonstrating the recommendations for labelling axes and representing bin contents.

Figure 13.3. However, for $n = 0$ or 1, this is not possible since the total cummulative probability below n is less than 34%. In this case, we settle for $\alpha_- + \alpha_- \simeq 68\%$ by fixing the lower bound of the error interval at zero and integrating up to the desired total confidence level. This leads to error bars for $n = 0$ and 1 that extend further on the high side than the low side, as shown in Figure 13.3.

Although the correct pdf for calculating binned errors is Poisson, it is common practice to assume Gaussian statistics which yield symmetric errors and are equivalent to Poisson errors at large n. In order to quantify the discrepancy between a Gaussian error interval (m_1, m_2) and the corresponding Poisson error interval (n_1, n_2), we define a metric

$$\Delta \equiv \frac{|n_1 - m_1| + |n_2 - m_2|}{(n_2 - n_1) + (m_2 - m_1)}$$

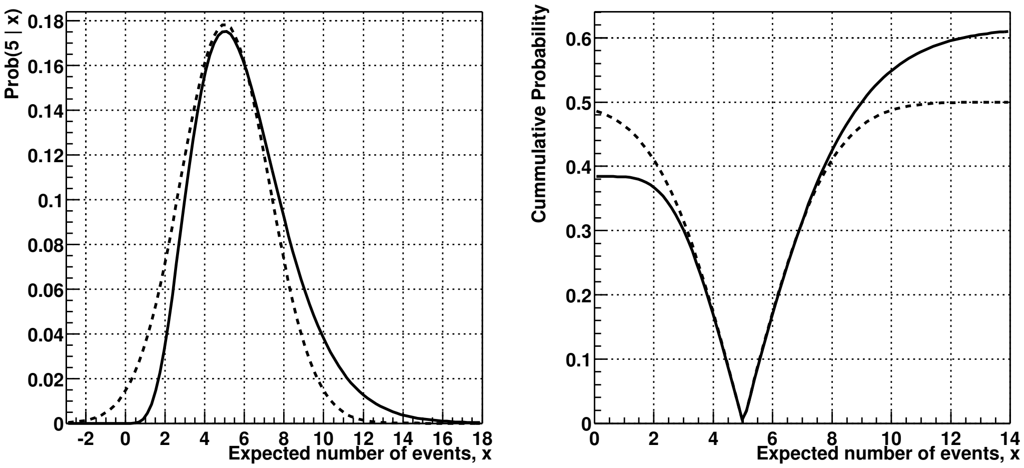

Figure 13.2: Comparison of Poisson (solid curves) and Gaussian (dashed curves) probability density functions for the expected number of events given 5 observed events. The left-hand plot compares the normalized pdf's. The right-hand plot shows the cummulative probabilities integrated above or below 5.

which is plotted as a function of the observed number of events, n, in the right-hand side of Figure 13.3. If we specify a maximum discrepancy 2%, then we require that Poisson errors be calculated for bins containing fewer than 10 entries. This should be considered as a minimum standard, and plots should ideally use Poisson errors for all bins. Table 13.1 lists the 68% Poisson error intervals corresponding to $n < 10$, for reference.

The following fragment from a ROOT macro (using the *BABAR* RooFitTools package) provides a practical example of how to present data following the recommendations made here, and was used to generate Figure 13.1:

```
RooRealVar mass("mass","Beam-Energy Substituted Mass",
                5200,5300,"MeV");
RooDataSet *data= RooDataSet::read("events.dat",mass);
data->Plot(mass)->Draw();
```

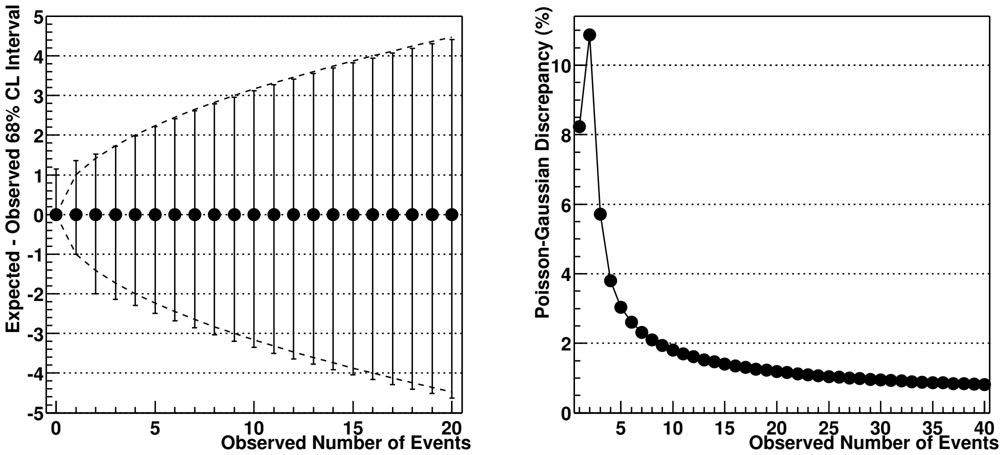

Figure 13.3: A comparison of Poisson and Gaussian confidence levels used for calculating binned statistical errors. The left-hand plot compares Poisson error bars (calculated as described in the text) with symmetric errors of size $\sqrt{n}$ (dashed curves) that are associated with a Gaussian pdf. The right-hand plot shows the discrepancy between the Poisson and Gaussian confidence levels (defined in the text) as a function of the number of entries in a bin.

13.2 How to Display a Fit Model

A graphical presentation of the model used in analysis is most useful when superimposed on a representation of the data used to determine the model's parameters. This combination then provides a visual impression of the goodness of a fit.

5. A presentation that combines data and a fit model should be chosen to clearly and accurately convey the statistical goodness of fit in the parameters of interest.

n	$n - \delta n$	$n + \delta n$
0	0	1.15
1	0	2.36
2	0	3.52
3	0.86	4.73
4	1.70	5.98
5	2.51	7.21
6	3.32	8.42
7	4.14	9.61
8	4.97	10.8
9	5.81	12.0

Table 13.1: Poisson 68% CL error intervals tabulated for different numbers of observed events.

13.2.1 Projecting a Model onto One Dimension

13.2.2 Displaying a Model in Several Dimensions

13.3 How to Display a Likelihood Function

13.3.1 Displaying a Likelihood Curve

6. A likelihood curve should be displayed as $\log(L/L_{max})$ where L_{max} is the maximum likelihood value corresponding to the best fit of the parameters to the data. This choice eliminates the somewhat arbitrary value of L_{max} from the plot and allows several related likelihood curves (with different values of L_{max}) to be directly compared.

7. The likelihood ratio should not be scaled by a factor of -2 since there is no reason to prefer a χ^2-style presentation when an unbinned likelihood fit was used.

8. In the case of a fit to several parameters, the likelihood ratio should be calculated as $\log(L_{max}(p)/L_{max})$ where $L_{max}(p)$ is the maximum likelihood obtained in a fit with p fixed and all other parameters varying.

13.3.2 Displaying Likelihood Contours

Provide a table of $n-$sigma $\log(L/L_{max})$ values calculated for different numbers of dimensions.

Refer to the MINUIT `mncont` routine.

Chapter 14

Systematic Errors–Recommendations

- Results should always be quoted with a separate statistical and systematic error. For presentational purposes these may also be combined in quadrature and given as a single error. In some cases, especially if this result will be combined with others, it may be appropriate to separate systematic errors into components, but this should not be general practice (see section 7.5).

- As many *BABAR* measurements will be limited by systematic errors, their calculation and manipulation should be done on the basis of sound practice, not folklore. Be able to justify the procedures used,

- Errors in a result due to factors which contain uncertainties should be considered as systematic even if those uncertainties are in themselves basically statistical, provided the sample on which they are based is independent of the sample from which the result is obtained (see sections 7.2.1, 7.5.2).

- Errors in a result due to factors which contain statistical uncertainties, derived from the same data sample but not of primary interest, should typically be considerd as statistical (see section 7.5.2.)

- Errors in a result due to factors which contain theoretical uncertainties may be manipulated in the usual way, but their evaluation is more uncertain (see section 7.2.2). There will often be no universally-agreed 'correct' value. A full explanation of the methodology you choose to use should be given, with justifications as appropriate; critical readers may then apply alternatives if they so wish. The following techniques may be used in evaluating such 'theory errors'.

  - Errors cited by theorists may be used. At least two estimates should be obtained, and the experimenter should then use their own judgement in selecting a value, and should be free to increase or decrease it if it is believed there is reason to do so.
  - Several different models may be used, and the spread of results used to give an error. How you do this depends on your judgement of the model credibilities- for examples see section 7.2.2. It may be appropriate to quote different results for different models.
  - Taking the error as the difference between the results of two models divided by $\sqrt{12}$ can only be done if the two models represent opposite extremes of the effect in question, and if the effect can reasonably be expected to lie anywhere in between them (see section 7.2.2.)

- In combining uncertainties, the correct covariance matrix must be used, or, if unavailable, best possible estimate (see section 7.3).

- Uncertainties in parameters that are incorporated in the fit in a complicated way are not amenable to treatment by algebraic manipulation. The errors that arise from them must be evaluated by altering the parameter values and repeating the fit. This should generally be done by taking half the difference between the result values from plus and minus one standard deviation of the parameter, unless there is reason or evidence to believe that the errors are asymmetric.

- Asymmetric errors should be given with their sign, so that correlations can be followed.

- Asymmetric errors should be combined by adding the positive and negative values, separately, in quadrature.

- The Likelihood function may be a good way to explore the effects of systematic errors, especially if event numbers are small (see section 7.5.3). The errors quoted should be those at which the log likelihood falls by

0.5 (for 1 dimension); if these are inconsistent with those obtained by other means (e.g. half the values for which it falls by 2.0, or those obtained from a 68% integral) then the full likelihood function should be given.

- The robustness of any result should be established by many checks (see Section 7.2.3). These would typically include

  - Inclusion/exclusion of appropriate parts of the data, depending on running conditions.
  - Variation of important cuts
  - Different fitting techniques
  - Performing a similar analysis for which the result is known beforehand, on logical grounds or from previous experiments.
  - A full consistency check on a simulated data sample.

  It is important to distinguish between such *blind checks*, for which no effect is expected and which are used to detect mistakes, and the *educated checks* used to detect biases / estimate corrections. (For illustrative examples see section 7.2.3.)

  The result of a blind check should be considered carefully to ascertain whether it is significant. If it is not, then no further action need be taken – it should not contribute to the systematic error. If it is, this indicates a mistake in the analysis, which should be searched for and (ideally) located and corrected. Only if this is unsuccessful should the systematic error should be inflated as a last and desperate resort,

- The full consistency check on simulated data will usually be a blind check, and the above argument applies to it equally, nevertheless we advocate that any discrepancy be added in quadrature to the systematic error (see section 7.2.3).

- To incorporate a systematic error into a limit in the typical, well-behaved (*i.e.*, approximately normal) case, the statistical and systematic uncertainties are added in quadrature, according to the general recommendation at the beginning of this chapter. The resulting uncertainty is then treated as a single normal error in obtaining the limit.

- To incorporate a small systematic uncertainty into a low-statistics limit the Cousins and Highland formula can be used.

  $$L' = L(1 + \frac{L - N}{2}\sigma_r^2)$$

  For definitions see section 7.6.

- To incorporate a large systematic uncertainty into a low-statistics limit, a toy Monte Carlo should be used, as described in Section 7.6.

Chapter 15

Interpretation of Results

15.1 Introduction

When it is desired to interpret a measurement in the context of a statement about physical "truth", we recommend adopting the Bayesian methodology.

As discussed in section 8.2, the basic formula for the Bayesian methodology, as it pertains to the determination of an unknown parameter θ, is:

$$P(\theta;\{x\}) = \frac{L(\theta;\{x\})P(\theta)}{\int L(\theta;\{x\})P(\theta)d\theta}.$$

15.2 Choice of Priors

For perhaps the majority of measurements in *BABAR*, it will be appropriate to simply use the frequentist results numerically as also the interpretation. This is often equivalent to choosing a prior probability distribution ("prior") which is uniform in the unknown parameter of interest. This is readily justified for those results where the information from *BABAR* is good, that is, any conclusions are not highly dependent on the choice of smooth prior expressing ignorance, and the measurement is substantially better than previous results.

However, there will undoubtably be situations where the conclusion is not so straightforward:

1. The measured result may be near a physical boundary.
2. The measurement may be of comparable (or poorer) precision with previous knowledge.

3. The measurement may have so little information that the choice of smooth prior makes an important difference.

Dealing with the case of a physical boundary is straightforward: The prior should be zero for unphysical values of the parameter(s).

The case where previous knowledge is comparable with, or better than, the information from *BABAR* should also be typically straightforward: A prior should be used which encapsulates the previous knowledge. There may be difficulty if it is not clear how to do this, *e.g.*, if a simple normal approximation may not be valid, if the systematic uncertainties are substantial, if the information provided by the earlier experiment is insufficient to formulate a prior, or if there is disagreement with the methodology used by the earlier experiment.

Finally, if there is little information content, and the choice of smooth prior is important, this should be pointed out in the discussion.

The general recommendation on the choice of prior is as follows:

- Express ignorance with a prior which is zero in an "unphysical" region, and a constant elsewhere.
- If the choice of how ignorance is expressed makes a difference to the result, say so, with an example.
- Prior experimental information should be included according to the posterior from such experiments. Typically, this will take on the form of a normal distribution with standard deviation obtained by adding statistical and systematic errors in quadrature as done by the Particle Data Group, including a scale factor if appropriate. Care must be taken, of course, to separate out common systematic uncertainties.

15.3 When do We Have a Signal?

The question "Is there a signal?" may not be only a "statistical" one: We must be open to the possibility that we have made a mistake (an a priori more likely prospect the more startling the result). A critical examination of the experiment and analysis design, and of the available cross checks must be made to evaluate this possibility. A large amount of judgement is required, and only general guidelines can be given:

- Was the analysis pre-determined? That is, were the cuts established prior to looking at the data? It not, there may have been tuning of cuts giving a biased result.
- How much "searching" was involved? If a lot of distributions were looked at, the signal may be just a large fluctuation that is not so improbable given the extent of the search. This again, is better evaluated (*i.e.*, statistical methods can be used) if a careful design was originally carried out. If it was not pre-determined how much searching would go on, considerable judgement is required to evaluate the significance.
- Does the purported signal show expected behavior? One check is to look at what is cut out, and see whether there is structure in the opposite direction from the signal in the discards. If so, the signal may be an artifact.

If it is supposed that the question of mistakes has been dealt with, the question of significance in terms of statistics may be asked. The significance may be computed either in frequentist terms or in Bayesian terms, but if the ultimate question is whether to claim a "significant" effect, some policy is required an answer to be given.

The statistics working group has grappled with possibilities for such a policy. The current stance is not to make such a recommendation, *i.e.*, not to specify cuts on significance level for discriminating between claim of a signal or not.

Instead, it is recommended that the situation be quoted in such a way that the reader can decide. For example, instead of saying "We observe a signal for $B \to 3\pi$", say "We observe $B \to 3\pi$ at the 1% significance level." Or, instead of "We have observed CP violation in B decays", say "The measured value of $\sin 2\beta$ is different from zero at the 6% significance level." A corollary is that language in titles of the form "Observation of…" is discouraged, as well as adjectives such as "first".

Acknowledgments

We have benefitted from many discussions with and queries from many members of the *BABAR* collaboration in producing this report. We are also grateful for comments received from a few people outside of *BABAR*: Gary Feldman, Frederick James, and Louis Lyons.

Appendix A

Charge to the *BaBar* Statistics Working Group

September 8, 2000
DRAFT 2

The BaBar experiment has begun presenting results at conferences, and will soon be preparing papers for publication. It is important that the results be presented as objectively as possible and with sound statistical procedures. Thus, the "Statistics Working Group" is charged with establishing a set of recommended statistical methods for use in quoting the results of BaBar analyses.

Questions that should be addressed include:

I. Confidence Intervals:

  A. What are recommended procedures for computing confidence intervals?
  B. What are the criteria for choosing to quote one-sided or two-sided confidence intervals?
  C. How should we incorporate systematic uncertainties? [See also below.]

II. Hypothesis Testing:

  A. What are recommended procedures for evaluating "goodness-of-fit"?

B. How do we determine the "significance" of a new result, *e.g.*, of a possible signal? [Including, how to deal with systematic uncertainties? See also below.]

III. Graphical Presentation:

A. What are recommendations concerning when and how to display statistical information graphically? [For example, plots of likelihood, chisq, contours.]

B. How should we deal with systematic uncertainties in graphical presentations? [See also below.]

IV. Systematic Errors:

A. What types of uncertainties should be identified as "systematic"?

B. How should we quote results with systematic uncertainties? [Note: A key consideration here is that people should be able to eliminate irrelevant common systematics in comparing quantities such as branching fractions.]

C. What are the recommended procedures for combining results, with proper attention to common sources of uncertainty?

V. Interpretation of Results:

A. Given a measurement, what procedures are recommended for arriving at a physical interpretation? [That is, given the relevant information content from the experiment, how do we formulate a physics conclusion? The point of this item is that the working group should consider the distinction between summarizing information, and arriving at physical conclusions. For example, a Bayesian methodology may be appropriate in the latter case, while not in the former.]

B. What are the recommendations for prior distributions?

C. When do we claim a "signal"? [Note: Criteria may differ depending on whether it amounts to a measurement of an "expected" effect *vs.* the announcement of a new "unexpected" effect.]

These questions should be addressed, as far as possible, with explicit practical procedures. These may differ depending on issues such as sample size, presence of background, and whether the measurement is "statistics-" or "systematics-" limited.

The discussions of the working group should be open to the collaboration. The report should be available from the web. A first draft should be available for comment by the December collaboration meeting.

Bibliography

[1] F. James and M. Roos, *Phys. Rev.* **D44** (1991) 299.

[2] T. Champion *et al.*, "Blind Analysis in *BABAR*", *BABAR* Anaysis Document 91, `http://www.slac.stanford.edu/babar-internal/BAD/-doc/detail.html?docNum=91`.

[3] M. G. Kendal and A. Stuart, "The Advanced Theory of Statistics", Vol.2. Charles Griffin and Company Limited, London.

[4] J. Neyman, Phil. Trans. Royal Soc. London, Series A, **236**, 333 (1937)

[5] G.J. Feldman and R.D. Cousins, *Phys. Rev.* **D57**, 3873 (1998).

[6] B.P. Roe, M.B. Woodroofe, *Phys. Rev.* **D60** 053009 (1999).

[7] B.P. Roe, M.B. Woodroofe, *Phys. Rev.* **D63** 013009 (2001).

[8] D.E. Groom et al, *The European Physical Journal* **C15**, 1 (1997).

[9] O. Helene, Nucl. Instr. and Meth. **212**, 319 (1983).

[10] G. Zech, Nucl. Instr. and Meth. **A277**, 608 (1989).

[11] V. Highland, Nucl. Instr. and Meth. **A398**, 429 (1997), followed by reply by G. Zech.

[12] P. Janot and F. Le Diberder, "Optimally combined confidence limits", Nucl. Instrum Meth. **A411**, 449 (1998).

[13] CLEO Collaboration, Phys.Rev.Lett. **84**, 5283 (2000).

[14] John R. Taylor "An Introduction to Error Analysis", University Science Books (1982).

[15] W. T. Ford, “Systematic Errors with Maximum Likelihood Fits in rare B decay measurements”, *BABAR* Note #529 (2000).

[16] R.D. Cousins and V L. Highland, *Nucl. Inst & Meth.* **A320**, 331 (1992).

[17] Art Snyder has discussed the problem in the context of two maximum likelihood analyses: http://www.slac.stanford.edu/~snyder/shifts.ps.

[18] See the section by Glen Cowan in T. Hurth *et al.*, “Prospects for CP violation”, proceedings of the Durham workshop to be published in J. Phys. G.

[19] R. Cahn, “Errors Induced by the Resolution Function”, *BABAR* internal document.

[20] J. Fullwood et al. ‘A study of $\tau^- \to \pi^-\pi^0\nu_\tau$ using the BaBar detector. *BABAR* Analysis document 185.

[21] ALEPH (R. Barate et al), Z. Phys. **C 76**, 15 (1997)

[22] S. Prell, “Systematic Error Estimation Strategy for $\sin 2\beta$ in $20\,\mathrm{fb}^{-1}$”, talk on November 3, 2000, `http://www.-slac.stanford.edu/BFROOT/www/Physics/CP/beta/Meetings/-03Nov00/session5/prell.pdf`.

[23] P. R. Bevington “Data Reduction and Analysis for the Physical Sciences”, McGraw Hill, New York, 1969.

[24] J. Boyd. ‘A study of $B^0 \to J/\psi\rho^0$ *BABAR* Analysis document 135.

[25] MINUIT manual/write up.

[26] Private communications with T. Trippe, M. Barnett, and D. Groom (2001).

[27] F. C. Porter, “Interval Estimation Using the Likelihood Function”, *Nucl. Inst. and Meth. in Phys. Res. A*, **368** (1996) 793-803; Electronic version available from URL: `http://www.slac.stanford.edu/-BFROOT/www/Statistics/bibliography.html`