Basic Statistics for Experimental Physics

It is important for both experimental and theoretical physicists to have a good understanding of probability and statistics in order to evaluate experimental results. A number of techniques exist for making such evaluations although a relatively small number of simple techniques usually suffice. Although knowledge of probability and statistics theory is important, one should never use these as a crutch. It is much more important to skillfully design and carry out an experiment than it is to exhaustively statistically analyze the results of a poor experiment. It soon becomes apparent to anyone with even a small amount of experience in analyzing data that statistical analysis merely quantifies what one's intuitive impressions are. There is a certain amount of calibration of intuition that goes on when one simultaneously observes data and imposes statistical tests on the data and this is perhaps one of the more important uses of probability and statistics to the beginning physicist. There will be many occasions when decisions will, of necessity, be based on impressions of likelyhood. Often there is simply not the time to perform detailed statistical calculations.

Certain notions of probability theory are implicit for application of statistical techniques. The following sequence of mathematically oriented assertions sets the stage for our brief study of statistics. I direct the attention of anyone wishing a more detailed exposition to the excellent introductory text by Larson (Introduction to Probability Theory and Statistical Inference, Wiley, 1969) from which much of the following is taken.

Definition: An experiment is any operation whose outcome cannot be predicted.

Definition: The sample space S of an experiment is the set of all possible outcomes for the experiment.

Definition: An event is a subset of the sample space. Every subset is an event.

Definition: An event occurs if any one of its elements is the outcome of the experiment.

Definition: A probability function is a real valued set function defined on the class of all subsets of the sample space S; the value that is associated with a subset A is denoted by P(A).

Probability theory is based on the following three axioms:

1. $P(S) = 1$
2. $P(A) \geq 0$ for all subsets A of S
3. $P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$ if $A_i \cap A_j = \phi$ for $i \neq j$

In the above we have used standard set theory notation for unions, intersections and the null set. Below we will use the symbol $\overline{A}$ to denote the complement of the subset A. The following theorems, which are easily proved from the axioms, are stated without proof.

Theorem: $P(\phi) = 0$

Theorem: $P(\overline{A}) = 1 - P(A)$

Theorem: $P(\overline{A}\cap B) = P(B) - P(A\cap B)$

Theorem: $P(A\cup B) = P(A) + P(B) - P(A\cap B)$

Two more definitions, related to the intuitively obvious technique for calculating the probability of the simultaneous occurrence of two independent events, are:

Definition: Two events, A and B, are independent if and only if $P(A\cap B) = P(A)P(B)$.

Definition: The conditional probability of B occurring given that A has occurred (written $P(B|A)$) is $P(B|A) = P(B\cap A)/P(A)$ if $P(A) > 0$. If $P(A) = 0$, we define $P(B|A) = 0$.

The above definitions and theorems embody much of that which is usually taken as intuitive notions of probability. To extend their utilization to statistics, one requires the concept of random variables and distribution functions.

Definition: A <u>random variable X</u> is a real-valued function of the elements of a sample space S.

Definition: A random variable X is <u>discrete</u> if its range forms a discrete (countable) set of real numbers. A random variable X is <u>continuous</u> if its range forms a continuous (uncountable) set of real numbers and the probability of X equalling any single value is 0.

Definition: The <u>probability function</u> for X is a function, $p_X(x)$, of a real variable x and is defined to be $p_X(x) = P(X(\omega)=x)$ where ω denotes a generic element of the sample space.

Definition: The <u>distribution function</u> for a random variable X (denoted $F_X(x)$) is a function of a real variable x such that: (1) the domain of definition of F_X is the whole real line, and (2) for any real x, $F_X(x) = P(X\leq x)$.

Definition: <u>The probability density function</u> for a continuous random variable Y (denoted $f_Y(y)$) is a function of a real variable y such that (1) the domain of f_Y is the whole real line and (2) for any real numer t

$$F_Y(t) = \int_{-\infty}^{t} f_Y(y)\,dy$$

Definition: (1) If X is a discrete random variable with probability function $p_X(x)$, the <u>expected value of H(X)</u> is defined to be:

$$E[H(X)] = \Sigma H(x)p_X(x)$$

(2) If X is a continuous random variable with probability density function $f_X(x)$, the expected value of H(X) is defined to be

$$E[H(X)] = \int_{-\infty}^{\infty} H(x)f_X(x)dx$$

Definition: The average value of a random variable X is defined to be $\mu_X = E[X]$.

Definition: The varience of a random variable X is defined to be: $\sigma_X^2 = E[(X-\mu_X)^2]$. The square root of the varience is the standard deviation of X.

There are only several probability distributions that are of interest to us. These are the binomial, Poisson, exponential and normal (or Gaussian) distributions.

Binomial Random Variable

We first define a Bernoulli trial as an experiment which has two possible outcomes: success and failure. The probability for success is denoted by p and the probability for failure by q=1-p. If X denotes the number of successes in n repeated independent Bernoulli trials, then X is called the Binomial random variable with parameters n and p. It is straightforward to calculate the Binomial probability function. There are n!/[(n-x)!x!] ways of ordering x successes in a sequence of n trials. Each of these orderings has a probability $p^x q^{n-x}$ of occurring. Thus:

$$p_X(x) = \binom{n}{x} p^x q^{n-x} \quad \text{where}$$

$\binom{n}{x}$ = n!/[(n-x)!x!] is the Binomial coefficient.

It is easy to show that for the Binomial distribution $\mu_X = np$ and $\sigma_X^2 = npq$.
$= \mu_x q$

Poisson Random Variable

The Poisson random variable is a limiting case of the binomial random variable. This is so because ~~because~~ a Poisson process with parameter λ is defined as a sequence of intervals for which the occurrence of events are possible. For a Poisson process the intervals can be made short enough so that the probability of two or more events occurring per interval is negligible. The intervals are usually in units of space or time and the Poisson parameter λ is the probability per unit interval for the occurrence of an event. For an interval of length s, and for a Poisson process with parameter λ, the Poisson random variable X is the number of successes k within this interval. We shall now show that

$$p_X(k) = \frac{(\lambda s)^k e^{-\lambda s}}{k!} .$$

As mentioned before, the Poisson process is a limiting case of the Binomial distribution. Each Bernoulli trial is an interval of length s/n and there are n such trials, where n is large enough so that the probability of two events per trial is negligible. Thus

$$p_X(k) = \binom{n}{k}(\lambda s/n)^k(1-\lambda s/n)^{n-k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(\frac{\lambda s/n}{1-\lambda s/n}\right)^k(1-\lambda s/n)^n$$

$$= \frac{n^k}{k!}\,\frac{(\lambda s)^k}{n^k}\,(1-1/n)(1-2/n)\cdots(1-(k-1)/n)\ (1-\lambda s/n)^n/(1-\lambda s/n)^k$$

In the limit as n → ∞

$$p_X(k) \rightarrow \frac{(\lambda s)^k e^{-\lambda s}}{k!} \quad \text{which is the desired result.}$$

The mean and varience for the Poisson random variable can be shown to be $\mu_X = \lambda s$ and $\sigma_X^2 = \lambda s$.

Exponential Random Variable

This is our first example of a random variable with a continuous probability density. The exponential random variable T with parameter λ is defined to be the interval between an arbitrary starting point and the point of occurrence for the first event for a Poisson process with parameter λ. To derive the probability density function, $f_T(s)$, we note that the probability of no successes from $s = 0$ to s is $\exp(-\lambda s)$(this is the case for k=0 for the Poisson distribution). The probability of an event within ds is λds so that

$$f_T(s) = \lambda e^{-\lambda s}$$

It can be shown that $\mu_T = 1/\lambda$ and $\sigma_T^2 = 1/\lambda^2$.

Normal Random Variable

This is undoubtedly the most important random variable to the scientist. This is due to the fact that the majority of experimental distributions are at least approximately determined by normal variations. There is a theoretical explanation for this fact which is embodied in the central limit theorem. We will come back to this theorem after we have first discussed the properties of the normal, or as it is also called, the Gaussian probability density function. A random variable is said to be normally distributed if and only if its probability density function is:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The parameter μ can be any real number while σ must be positive. As the symbols suggest, the mean and varience of the normal distribution are $\mu_X = \mu$ and $\sigma_X^2 = \sigma^2$.

We now define the moment generating function which is quite useful in obtaining information relating to probability distributions.

Definition: The moment-generating function, $m_X(t)$, for a random variable X is defined to be $m_X(t) = E[\exp(tX)]$.

The name moment-generating function comes from the easily proved fact that

$$m_X^{(k)}(0) \equiv \frac{d^k}{dt^k} m_X(t)\Big|_{t=0} = m_k$$

where m_k is the kth moment of the distribution, i.e. $m_k = E[X^k]$. It is intuitively obvious that a probability density function is determined by its moments. The first moment determines an average value, the second moment is related to the peak width, the third to the assymmetry etc. This fact can be proven rigorously if we assume that only positive values of a random variable are allowed. In this case $m_X(t)=m_Y(t)=\int_0^\infty e^{tx} f_X(x)\,dx = \int_0^\infty e^{ty} f_Y(x)\,dx$

or $\mathcal{L}_X(-t) = \mathcal{L}_Y(-t)$ where $\mathcal{L}$ denotes the Laplace transform. Since Laplace transforms are invertable this means that $f_X(x) = f_Y(x)$. The proof of the general case can be found in advanced texts on probability. We state the theorem below.

Theorem: Assume that X and Y are random variables with moment generating functions $m_X(t)$ and $m_Y(t)$ respectively. Then $m_X(t) = m_Y(t)$ if and only if $f_X(t) = f_Y(t)$ for all t.

We will now introduce the concept of jointly distributed random variables:

Definition: Given an experiment, the pair (X,Y) is called a <u>two dimensional random variable</u> if each of X and Y associates a real number with every element of S.

This definition is easily extended to n dimensional random variables, which are n-tuples $(X_1,X_2,\cdots X_n)$. The random variables $X_1,X_2,\cdots X_n$ are said to be independent if

$$p_{\underline{X}}(\underline{\boldsymbol{x}}) = \prod_{i=1}^{n} p_{X_i}(x_i) \quad \text{where}$$

the underlines denote n-tuples. We now state and prove the following useful theorem:

Theorem: If $X_1,X_2,\cdots X_n$ are independent, identically distributed random variables and

$$Y = \sum_{i=1}^{n} X_i \text{ then } m_Y(t) =[m_X(t)]^n.$$

Proof: $m_X(t) = E[e^{tY}] = E[e^{tX_1+tX_2+\cdots tX_n}] = E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}]$
$= [m_X(t)]^n$

A very useful theorem makes use of this result.

Theorem: Let $X_1,X_2,\cdots X_n$ be independent, identically distributed normal random variables with parameters μ and σ. The $Y = X_1+X_2+\cdots X_n$ is a normal random variable with parameters $n\mu$ and $n\sigma^2$.

Proof: We first evaluate $m_X(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-(x-\mu)^2/2\sigma^2}dx$.

Using the relation $\int_{-\infty}^{\infty} e^{-ax^2+bx}dx = \sqrt{\pi/a}\ e^{b^2/(4a)}$ we find

$m_X(t) = e^{t\mu+t^2\sigma^2/2}$. From the previous theorem,

$m_Y(t) = e^{n\mu t + n\sigma^2t^2/2}$. Thus, from the uniqueness property of generating functions Y is normally distributed with mean $n\mu$ and varience $n\sigma^2$.

The usefullness of the above result will be apparent when we consider sampling theory and statistics.

We quote below a theorem which is independent of the distribution. Proof will be left as an exercise.

Theorem: Let $X_1, X_2, \cdots X_n$ be independent random variables with means $\mu_1, \mu_2, \cdots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$ respectively. If $Y = \sum_{i=1}^{n} a_i X_i$ (a_i's are constants) then $\mu_Y = \sum_{i=1}^{n} a_i \mu_i$ and

$$\sigma_y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2 \,.$$

It follows from this theorem that $\overline{X} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i$ has mean $\mu_{\overline{X}} = \mu$ (if $\mu_{X_i} = \mu$ for all i) and variance $\sigma_{\overline{X}}^2 = \sigma^2/n$. The variance of the distribution of a mean is smaller than the variance of the single distribution by a factor $1/n$. This fact is of considerable importance for statistical inference.

We now state and give a heuristic proof of the Central Limit Theorem:

Theorem: Suppose $X_1, X_2, X_3 \cdots$ is a sequence of indepedent, identically distributed random variables, each with mean μ and variance σ^2. Define the sequence $Z_1, Z_2, Z_3, \cdots$ by $Z_n = (\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ $n = 1, 2, 3, \cdots$ where $\overline{X}_n = \sum_{i=1}^{n} X_i/n$. Then for all real t, $\lim_{n\to\infty} F_{Z_n}(t) = N_Z(t)$ where $N_Z(t)$ is the standard normal distribution function (i.e. the random variable Z has a mean value of 0 and a variance of 1).

Proof: We have seen above that $E[\overline{X}_n] = \mu$ and $E[(X_n - \mu)^2] = \sigma^2/n$. Thus $E[Z_n] = 0$ and $E[Z_n^2] = 1$. The moment generating function for Z_n is:

$$m_{Z_n}(t) = E[e^{tZ_n}] = E[e^{t(\overline{X}_n - \mu)/(\sigma/\sqrt{n})}] = E[\prod_{i=1}^{n} e^{t(X_i - \mu)/(\sigma\sqrt{n})}]$$

$$= \prod_{i=1}^{n} E[e^{t(X_i - \mu)/(\sigma\sqrt{n})}]$$

$$m_{Z_n}(t) = [\, m_{(X-\mu)/\sigma}(t/\sqrt{n})]^n$$

$$\ln m_{z_n}(t) = n \ln[m_{(X-\mu)/\sigma}(t/\sqrt{n})]$$

The moment generating function can be written as a power series with moments as coefficients. Since $\overline{(X-\mu)/\sigma} = 0$ and $\sigma^2_{(X-\mu)/\sigma} = 1$ we have

$$m_{(X-\mu)/\sigma}(t) = 1 + t^2/2 + m_3 t^3/3! + \cdots$$

and

$$m_{(X-\mu)/\sigma}(t/\sqrt{n}) = 1 + t^2/2n + m_3 \frac{(t/\sqrt{n})^3}{3!} + m_4 \frac{(t/\sqrt{n})^4}{4!} + \cdots$$

$$= 1 + a(t)$$

$$\ln m_{Z_n}(t) = n\ \ln(1+a(t)) = n\ (a(t)-a^2(t)/2+a^3(t)/3-\cdots) \quad \text{for } |a| < 1$$

As $n \to \infty$, $na \to t^2/2$, $na^k \to 0$ if $k=2,3,4,\cdots$

So $$\lim_{n\to\infty} \ln m_{Z_n}(t) = t^2/2 \quad \text{or} \quad \lim_{n\to\infty} m_{Z_n}(t) = e^{t^2/2}$$

But this is just the moment generating function for a normal random variable with mean 0 and variance 1. This completes the proof.

Thus, the arithmetic average of a large number of independent, identically distributed random variables is distributed according to the normal distribution in the limit of large sample size, regardless of the nature of the individual random variable distribution. This explains the fact that the normal distribution is a good approximation for a large number of probability laws. It is impossible to determine in general how large n must be for a given probability law to be approximated by a normal distribution to a given degree of accuracy. An important example of this will be seen later for the various energy straggling distributions. Two familiar distributions which are readily approximated by the normal distribution are the binomial and Poisson distributions. As a rule of thumb, the normal approximation to the binomial distribution is good if the number of trials $n \geqslant 30$ and $np \geqslant 5$ (since the Poisson distribution is in fact a limiting case of the binomial distribution as $n \to \infty$, we see that a Gaussian distribution approximates a Poisson distribution if $\lambda s \geqslant 5$).

Statistical Inference

Up to this point we have assumed well defined sample spaces, experiments and probability laws. Prediction of expected values followed in a straightforward manner. Each probability law had well defined parameters which characterized the law.

The experimentalist operates in a shadier world. He is doing experiments to unravel the parameters of an unknown probability law. There are often competing theories to be decided amongst on the basis of experimental results, and each of these theories has a different probability law. The subject of determining the best probability law and the best estimates of the parameters which characterize that law is known as statistical inference. The notion of sampling is central to this subject. A sequence of experiments on the random variable X is known as a random sample. A random sample of size n is an n-tuple of random variables $X_1, X_2, \cdots X_n$. Any function of the elements of a random sample is called a statistic. Thus $\sum X_i/n$ is a statistic and is referred to as the sample mean. The k'th sample moment is defined to be $M_k = (\sum X_i^k)/n$. Note that means and moments take on different meanings than in probability theory. In sampling theory, moments are statistics and as such are random variables. One can repeat a sample many times, each time obtaining a different value for a sample moment. A straightforward but important theorem regarding sample moments follows:

Theorem: Let $X_1, X_2, \cdots X_n$ be a random sample of X. Then $E[M_k] = m_k$ where M_k is the k'th moment of the sample and m_k is the k'th moment of X.

Proof: $E[M_k] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^k\right] = \frac{1}{n}\sum m_k = m_k$ where we have used the fact that the elements of a random sample are identically distributed.

Thus, if we let $\overline{X} = M_1$, then $E[\overline{X}] = \mu$. An important corollary follows the definition of the sample variance.

Definition: If $X_1, X_2, \cdots, X_n$ are a random sample of X, the <u>sample variance</u> is defined to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \quad \text{where } \overline{X} \text{ is the sample mean.}$$

This is probably familiar to most of you as an estimate of the variance of the parent distribution. The justification for this follows from the fact that $E[s^2] = \sigma^2$ as can be seen below:

$$E[s^2] = \frac{1}{n-1} E[\Sigma(X_i-\overline{X})^2] = \frac{1}{n-1} E[nM_2 - n\overline{X}^2] = \frac{n}{n-1}[m_2 - E[\overline{X}^2]]$$

But,

$$E[\overline{X}^2] = E[(\overline{X}-\mu)^2] + \mu^2 = \sigma^2/n + \mu^2 \quad \text{so}$$

$$E[s^2] = \frac{n}{n-1}[\sigma^2 + \mu^2 - \mu^2 - \sigma^2/n] = \sigma^2$$

The most important distribution function in the study of statistics and in the analysis of experimental data is that of the chi-squared (χ^2) random variable. It is defined in the following way:

Definition: If $X_1, X_2, \cdots X_n$ is a random sample of a <u>normal</u> random variable X with mean μ and variance σ^2, then

$$Y = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$$

is a <u>χ^2 random variable</u> with n degrees of freedom.

Theorem: The probability density function for the χ^2 random variable with n degrees of freedom is

$$f_Y(y) = \frac{1}{\Gamma(n/2)} \frac{1}{2^{n/2}} y^{(n/2)-1} e^{-y/2}, \quad y > 0$$

$$= 0 \text{ otherwise.}$$

Proof:

$$m_Y(t) = E[e^{tY}] = \prod_{i=1}^{n} E[e^{t(X_i-\mu)^2/\sigma^2}] = [m_{(X-\mu)^2/\sigma^2}(t)]^n$$

$$= [m_{\chi_1^2}(t)]^n \qquad \chi_1^2 = (X-\mu)^2/\sigma^2$$

$$m_{\chi_1^2}(t) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{t(x-\mu)^2/\sigma^2} e^{-(x-\mu)^2/2\sigma^2} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-(x-\mu)^2(1-2t)/2\sigma^2} dx; \text{ if } y = x-\mu,$$

$$m_{\chi_1^2}(t) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-y^2(1-2t)/2\sigma^2} dx \text{ and if } x^2 = y^2(1-2t)/\sigma^2$$

$$m_{\chi_1^2}(t) = \frac{1}{\sqrt{2\pi}\sigma} \frac{\sigma}{(1-2t)^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = (1-2t)^{-\frac{1}{2}}$$

so $$m_Y(t) = (1-2t)^{-n/2}$$

But $$I \equiv \int_0^{\infty} \frac{1}{\Gamma(n/2)} \frac{1}{2^{n/2}} y^{n/2-1} e^{-y/2} e^{ty} dy = \frac{1}{\Gamma(n/2)2^{n/2}} \int_0^{\infty} y^{n/2-1} e^{y(t-\frac{1}{2})} dy$$

Let $x = -y(t-\frac{1}{2})$

$$I = \frac{1}{\Gamma(n/2)2^{n/2}} \int_0^{\infty} x^{n/2-1} \frac{1}{(-t+\frac{1}{2})^{n/2-1}} e^{-x} \frac{(+dx)}{(-t+\frac{1}{2})}$$

$$I = \frac{1}{\Gamma(n/2)2^{n/2}} \frac{1}{(-t+\frac{1}{2})^{n/2}} \Gamma(n/2) = \frac{1}{(1-2t)^{n/2}}$$

This completes the proof.
It can be shown that $\mu_{\chi^2} = n$ and $\sigma^2_{\chi^2} = 2n$.

Theorem: If Y and Z are independent χ^2 random variables with m and n degrees of freedom respectively, then Y + Z is a χ^2 random variable with m+n degrees of freedom.

Proof: $$m_{Y+Z}(t) = E[e^{tY}e^{tZ}] = m_Y(t)m_Z(t) = (1-2t)^{-m/2}(1-2t)^{-n/2}$$
$$= (1-2t)^{-(m+n)/2}$$

It is easy to show that $$\frac{\Sigma(X_i-\mu)^2}{\sigma^2} = \frac{\Sigma(X_i-\overline{X})^2}{\sigma^2} + \frac{(\overline{X}-\mu)^2}{\sigma^2/n}$$

The term on the left is a χ^2 random variable with n degrees of freedom and the second term on the right is a χ^2 random variable with 1 degree of freedom. From the above theorem, it is plausible, and in fact can be proved that:

$\frac{\Sigma(X_i-\overline{X})^2}{\sigma^2}$ is a χ^2 random variable with n-1 degrees of freedom.

The meaning of the term degrees of freedom becomes clear: one degree of freedom has been lost in determining the mean value $\overline{X}$. Effectively, there are only n-1 independent elements in the sample. One has been used up to determine $\overline{X}$.

If we are sampling a single random variable, the χ^2 distribution can be used to evaluate a confidence level for an estimate of σ if the random variable is known to be Gaussian. Most probability texts and compilations of mathematical functions have tables of the chi-squared distribution which can be used for this purpose. A simple extention of the chi-squared concept also allows one to determine the goodness of fit of a hypothesized function

$\bar{y}(x)$ to a collection of data $y_i(x)$, where x is assumed to be known. As an example, x may be a channel on a pulse height analyzer and y_i may be the number of observed counts in the i'th channel, while $\bar{y}(x)$ would be the expected number of counts based on some theory. It can then be shown that the statistic:

$$\chi^2 = \sum_{i=1}^{n} \frac{[y_i-\bar{y}(x_i)]^2}{\sigma_i^{\,2}} ,$$

(where i runs over the number of values of x_i and where σ_i is the known standard deviation for the normally distributed random variable $y(x_i)$) is a chi-squared random variable with ν degrees of freedom where $\nu=n-m$, m being the number of parameters used to estimate the function $\bar{y}(x)$. The method of maximum likelyhood is used to evaluate the m functional parameters. In this method, parameters are selected which maximize the theoretical probability of observing the actual outcome of an experiment. For the case of normally distributed random variables at each x_i, it can be shown that the method of maximum likelyhood is equivalent to the least squares method whereby the statistic χ^2 is minimized by adjusting the parameters of $\bar{y}(x_i)$. Goodness of fit is then checked by looking up the minimized value of chi-squared in a mathematical table. If the chi-squared table yields the result that the probability for observing the actual result is less than 0.1%, one says the discrepancy is very highly significant, if less than 1% it is highly significant and if less than 5% it is significant. Inapproprriate usage of these words could cause misunderstanding so it would serve you well to memorize them. Discrepancy, in the present context, is taken to imply that the assumed form of the function $\bar{y}(x)$ is not correct. It is important to realize that a chi-squared test never can be used to prove that a given function is the correct one, it can only be used to reject hypothetical functions.

Having lapsed momentarily into a somewhat sloppy discussion of chi-squared in the hopes of striking a familiar note we will now try to regain our composure with the following definition:

Definition: Supose that X is a random variable whose probability law depends on an unknown parameter θ. Given a random sample of $X_1,X_2,\cdots,X_n$, the two statistics L_1 and L_2 form a $100(1-\alpha)\%$ confidence interval for θ if $P(L_1\leqslant\theta\leqslant L_2) \geqslant 1-\alpha$.

The concept of a confidence interval is quite important to an experimentalist since this is as close as he or she can generally come to the truth. Measurements are always of the nature of random samples and the unknowns are the parameters of probability distributions. Thus an experimentalist cannot hope to determine θ but it is within his or her grasp to determine the statistics L_1 and L_2 which bracket θ. The goal of the experimentalist is to determine L_1 and L_2 which are very nearly equal to each other for values of α close to zero. As an example, suppose we take a random sample of a normal random variable with unknown μ but of known variance σ^2. We know that $\bar{X}$ is a normal random variable with mean μ and varience σ^2/n so that it is clear that $(\bar{X}-\sigma/\sqrt{n},\ \bar{X}+\sigma/\sqrt{n})$ forms a 68.3% confidence interval for μ. What if we do not know what σ is however? There is something called the t-distribution, which is quite close to the normal distribution and, which enables us to form confidence intervals in this case. This follows from the theorem below which we will state without proof:

Theorem: Let $X_1, X_2, \cdots X_n$ be a random sample of a normal random variable with mean μ and variance σ^2. Then $(\bar{X}-\mu)/(s/\sqrt{n})$ has the t-distribution with n-1 degrees of freedom where s is the sample variance and the t-distribution density function with m degrees of freedom is:

$$f_T(t) = \frac{\Gamma((m+1)/2)}{\Gamma(m/2)\sqrt{m\pi}} (1+t^2/m)^{-(m+1)/2} .$$

By consulting a table of t-distributions we can find the 68.3% confidence intervals for μ and we see they have the form:

$(\bar{X}-ks/n^{\frac{1}{2}}, \bar{X}+ks/n^{\frac{1}{2}})$ where k is given below for various n:

n	k
2	2.26
5	1.22
10	1.12
20	1.08

We see that as n gets larger the distinction between the t-distribution and the Gaussian gets smaller (A Gaussian distribution would have k=1).

We remind you that you already know how to form confidence intervals for an unknown varience σ^2 of a normal random variable by taking a random sample $X_1, X_2, \cdots X_n$ and forming the sample mean $\bar{X}$. This is so since $\sum(X_i-\bar{X})^2/\sigma^2$ is a chi-squared random variable with n-1 degrees of freedom. Thus, if $\chi^2_{\alpha/2}$ denotes the point for which $P(\chi^2<\chi^2_{\alpha/2}) = \alpha/2$ and $P(\chi^2<\chi^2_{1-\alpha/2}) = 1-\alpha/2$ then

$$L_1 = \sum(X_i-\bar{X})^2/\chi^2_{1-\alpha/2} \quad \text{and} \quad L_2 = \sum(X_i-\bar{X})^2/\chi^2_{\alpha/2} \quad \text{form}$$

a $100(1-\alpha)\%$ confidence interval.

Two other distributions exist which are useful in statistical inference and we will close our discussion by briefly touching on these. One of these is the F distribution. If two statistics χ^2_1 and χ^2_2 are determined which follow the chi-squared distribution, the ratio

$(\chi^2_1/\nu_1)/(\chi^2_2/\nu_2)$ (ν_1 and ν_2 are the number of degrees of

freedom for χ^2_1 and χ^2_2 respectively) is distributed according to the F distribution (F stands for R.A. Fisher, one of the most influential workers in the field of statistics). Like the normal, t- and χ^2 distributions, tables of the F distribution can be found in most probability and statistics texts. The most common application of the F-test is in determining if a fit of data to a multi-parameter function is significantly improved by adding another parameter. To simplify tabulations it is convenient to use the statistic

$$F = \frac{\chi^2(m-1) - \chi^2(m)}{\chi^2(m)/(N-m-1)}$$

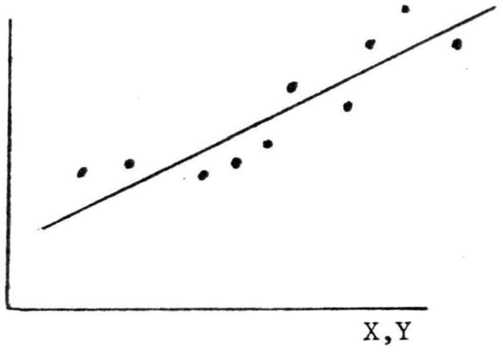
where m is the number of parameters used. The additivity feature of χ^2 random variables implies that the numerator is a chi-squared random variable with 1 degree of freedom. Thus instead of tabulating F distributions for a plane (ν_1,ν_2) we merely need to tabulate for various values of ν_2. If the added parameter is useful, the observed value of F will correspond to a small probability for such an observation, as computed from tabulated values of the F distribution. This is so because the statistic F does not actually correspond to an F random variable unless the data is fit by the proper curve.

This is of course the same reasoning that applies to the chi-squared test. The statistic $\sum(X-\overline{X})^2/\sigma^2$ is only a χ^2 random variable if $\overline{X}$ is the sample mean. For exactly the same reason, $\sum(y_i-\overline{y}(x_i))^2/\sigma_i^2$ is a χ^2 random variable only if $\overline{y}(x_i)$ is the maximum likelyhood estimate of the actual, albeit unknown, average value of the random variable $y(x_i)$.

The chi-squared and F tests for goodness of fits are known as "distribution dependent tests." In order to use chi-squared one needs to know the variances at each x_i and one needs to know that the distributions $y(x_i)$ are normal. The F test is something of an improvement in that the variances cancel out although it still only works for Gaussian distributions and it only allows one to decide if one fit is better than another and not if either one is consistent with the correct function (if one does not know the variance at x_i it is impossible to asses goodness of fit since it is conceivable within the realm of statistics that a set of scattered data points actually reflect a "scattered" parent function). There exist a number of tests which are referred to as distribution free tests. These can be applied to a data set for which we have complete ignorance regarding the nature of the parent distribution. In general, they are all based on arranging observations and ranking them in some way. One such test is the run test. Let $X_1,X_2,\cdots,X_m$ be an ordered sample and let $Y_1,Y_2,\cdots Y_n$ be a second ordered sample (by ordered we mean $X_1\leqslant X_2\leqslant\cdots\leqslant X_m$). If we let $Z_1,Z_2,\cdots,Z_{m+n}$ denote the ordered sampl~~ing~~e consisting of the X's and Y's then we define a "run" as a sequence of one or more X's or one or more Y's. Let R denote the number of runs. For $m\geqslant 11$ and $n\geqslant 11$ it can be shown that

$$\mu_R = \frac{2mn}{m+n} + 1 \qquad \text{and} \qquad \sigma_R^2 = \frac{2mn(2mn-m-n)}{(m+n)^2(m+n+1)}$$

and that if $|\mu_R-R|/\sigma_R \geqslant 1.645, 1.96, 2.33$ or 2.58 we can reject the hypothesis that the samples come from the same population at the 5, 2.5, 1.0 or 0.5% level respectively. One possible use for this test is to determine whether or not a distribution is actually random. Consider the data below and the best fit straight line.



Let $x_1,x_2,\cdots,x_m$ denote the abscissas of those points which lie below the line and let $y_1,y_2,\cdots,y_n$ denote the abscissas of the points above the line. If the parent distribution is truly random, the ordered sample of x's and y's will pass the run test. Note that we do not require <n> = <m>, i.e. it is not necessary for the parent distribution to even be symmetrically distributed about the line, let alone Gaussian. This is a very powerful way to test the hypothesis that a line (it does not have to be a linear fit) is an acceptable fit of randomly distributed data. If the test fails we can reject either the fit or the assumption of randomness.

In closing we again caution you against using statistics as a crutch. It is very rare for the probability distributions encountered in the real world to be as clean and well defined as those found in mathematical theorems. Even if one is fairly confident that an experimental distribution is Gaussian, it is nearly impossible to eliminate the possibility that a few scattered points far from the peak are not part of a tail to the distribution. Tails are just one problem to contend with. Among the many other problems is the fact that in a very large fraction of instances one knows for sure that he is dealing with a situation in which the requirements for the standard tests for hypothesis are not met. For example,suppose one is testing the hypothesis that a histogram is fit by a Gaussian function. Suppose the i'th bin of the histogram has N_i counts. Let E_i be the expected number of counts in this bin based on the assumed Gaussian function:

$$E_i = K \exp\,[-(i-i_p)^2/(2\sigma^2)]$$

where i_p is the average bin (note that while i is an integer, i_p does not need to be an integer). Estimates of the parameters i_p and σ can be obtained via the relations:

$$i_p = \Sigma i N_i/\Sigma N_i$$

$$\sigma^2 = \Sigma(i-i_p)^2 N_i/(\Sigma N_i-1)$$

To determine K we must minimize the statistic

$$\chi^2 = \sum \frac{(N_i-E_i)^2}{\sigma_i^2}$$

But what do we use for σ_i^2? Since the process of having points fall into a given bin in a histogram is binomial, the variance , σ_i^2, of the number of counts in the bin is $pq\Sigma N_i$ where p is the probability that a bin is added to for any given event and q=1-p. If there are many populated bins or if a given bin population is very small compared to that of other bins then q is nearly equal to 1 and the varience is simply $\sim p\Sigma N_i$ or E_i. This is the variance one expects from a Poisson distribution which follows from the fact that a Poisson distribution is a good approximation to a binomial distribution for a small value of p. Thus we can write:

$$\chi^2 = \sum \frac{(N_i-E_i)^2}{E_i}$$

This statistic is minimized to determine K. Can we then use the minimized χ^2 to determine goodness of fit? Strictly speaking the answer is no. Since the data is distributed according to Poisson statistics, the use of the chi-squared test, which requires normal variations, is not valid. However it should be recalled that for $E_i \gtrsim 5$, a normal approximation to the Poisson distribution is justified. Thus if bins of unequal width are arranged so that $E_i \gtrsim 5$, we can , in an approximate sort of way, use the χ^2 distribution with ΣN_i-3 degrees of freedom for a test of goodness of fit. In a similar crude manner, one often simply uses N_i as an estimate of σ_i^2 rather than the actual expected number of counts in bin i. What then has happened to mathematical precision? It has simply evolved. The fact of the matter is that the theory of statistics would be useless unless some consideration is given to the fact that the rigorous mathematical requirements are almost never satisfied in the real world. To cover themselves, mathematicians have come up with a word which precisely defines how one copes with imprecision. The word is robustness. A procedure is robust if it still works fairly well even if the assumptions are not quite satisfied. Of course, physicists have long known about robustness. Experiments are never perfect, but good ones work fairly well.