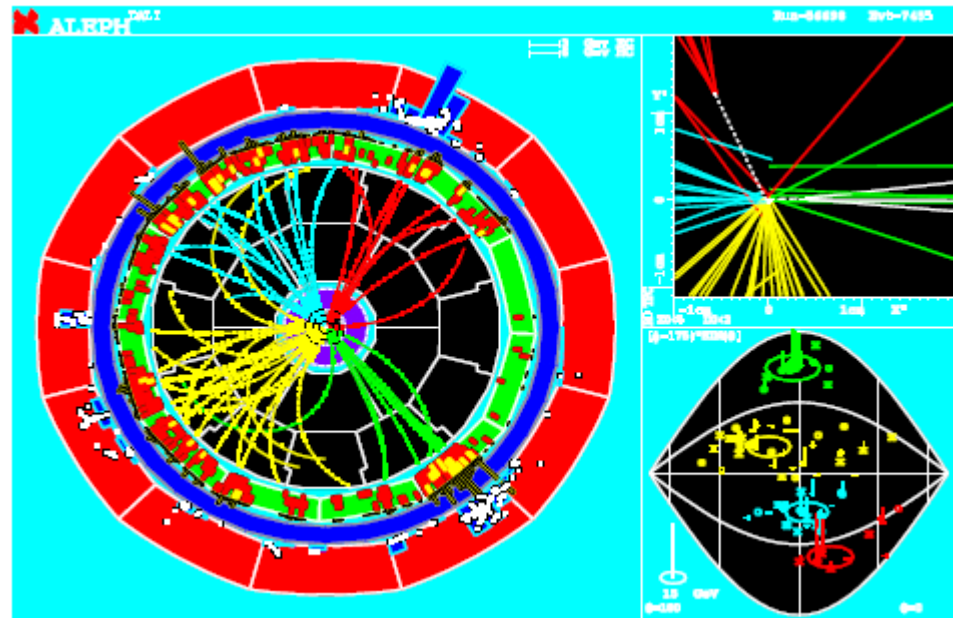


Statistics and Data Analysis (HEP)

- **Back to basics**
- **The Gaussian Limit**
- **Fitting and Hypothesis Testing**
- **Dark Arts**

ALEPH (LEP detector) event display



Follow the course/slides from M. A. Thomson lectures at Cambridge University

Lectures Synopsis

Lecture 1: Back to basics

Introduction, Probability distribution functions, Binomial distributions, Poisson distribution

Lecture 2: The Gaussian Limit

The central limit theorem, Gaussian errors, Error propagation, Combination of measurements, Multi-dimensional Gaussian errors, Error Matrix

Lecture 3: Fitting and Hypothesis Testing

The χ^2 test, Likelihood functions, Fitting, Binned maximum likelihood, Unbinned maximum likelihood

Lecture 4: Dark Arts

Bayesian statistics, Confidence intervals, systematic errors.

Experimental Physics

- ★ Experimental science concerned with two types of experimental measurement:
 - ♦ Measurement of a quantity : **parameter estimation**
 - ♦ Tests of a theory/model : **hypothesis testing**
- ★ For **parameter estimation** we usually have some data (a set of measurements) and from which we want to obtain
 - ♦ The **best estimate** of the true parameter; “the measured value”
 - ♦ The **best estimate** of how well we have measured the parameter; “the uncertainty”
- ★ For **hypothesis testing** we usually have some data (a set of measurements) and one or more theoretical models, and want
 - ♦ A measure of how consistent our data are with the model; “a probability”
 - ♦ Which model best describes our data; “a relative probability”

To address the above questions we need to use and understand **statistical** techniques

- ★ In these **4** lectures we will cover most aspects of statistics as applied to experimental high energy physics:
 - ♦ Nothing will be stated without proof (or at least justification).
 - ♦ Understanding the derivations will help you to understand the basis behind the statistical techniques

The path to enlightenment:

- If you measure something always quote an **uncertainty**
- Understand what you are doing and why
- Don't forget that you are usually *estimating* the uncertainty
 - e.g. don't worry too much about whether an effect is 2.9σ and 3.1σ
unlikely you can estimate the uncertainty that well
- Don't worry too much about the difference between Bayesian and Frequentist approaches
 - often give same results
 - if the results are different – usually means data are weak
– so do another experiment

Three Types of Errors

Statistical Uncertainties:

- ★ Random fluctuations
 - ♦ e.g. shot noise, measuring small currents, how many electrons arrive in a fixed time
 - ♦ Tossing a coin N times, how many heads

The main topic of these lectures

Systematic Uncertainties:

- ★ Biases
 - ♦ e.g. energy calibration wrong
 - ♦ Thermal expansion of measuring device
 - ♦ Imperfect theoretical predications

Discussed in the last lecture

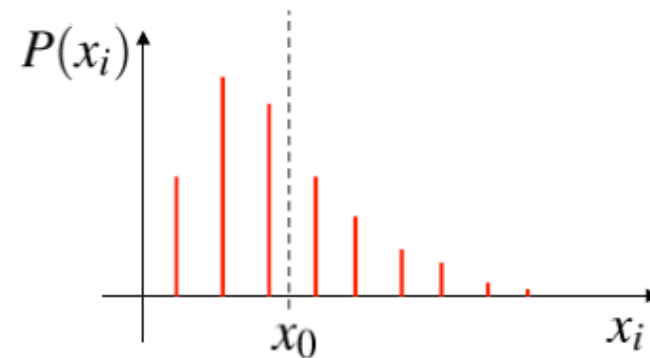
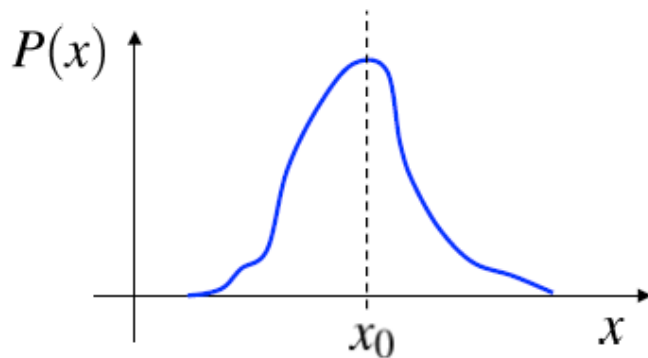
Blunders, i.e. **errors**:

- ★ Mistakes
 - ♦ Forgot to include a particular background in analysis
 - ♦ Bugs in analysis code

Not discussed, never happen...

Probability Distributions

- ★ Suppose we are trying to measure some quantity with true value x_0 the result of a single measurement follows a probability density function (PDF) which may or may not be of a known form.



- ★ Normalised:

$$\int_{-\infty}^{+\infty} P(x) dx = 1$$

$$\sum_{i=0}^{\infty} P(x_i) = 1$$

- ★ In general, can parameterise the PDF by its moments α_n

$$\alpha_n = \int x^n P(x) dx$$

$$\alpha_n = \sum x^n P_i$$

Note: $\alpha_n \equiv \langle x^n \rangle$

Mean and Variance

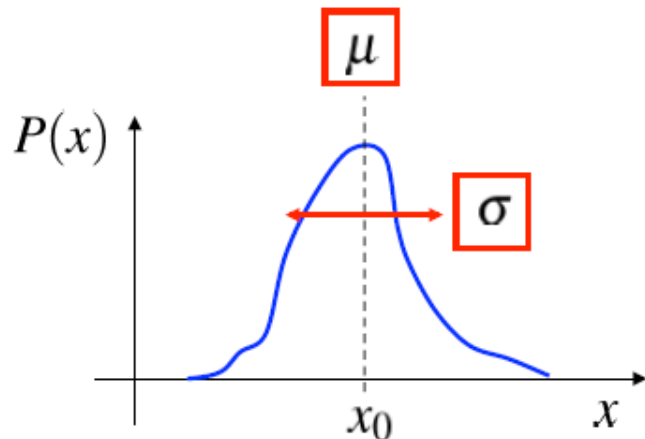
★ Can now define a few important properties of the PDF

Mean: $\mu \equiv \langle x \rangle = \int xP(x)dx$ “average of many measurements”

Mean of squares: $\langle x^2 \rangle = \int x^2P(x)dx$

Variance: $Var(x) \equiv \sigma^2 \equiv \langle (x - \mu)^2 \rangle = \int (x - \mu)^2P(x)dx$

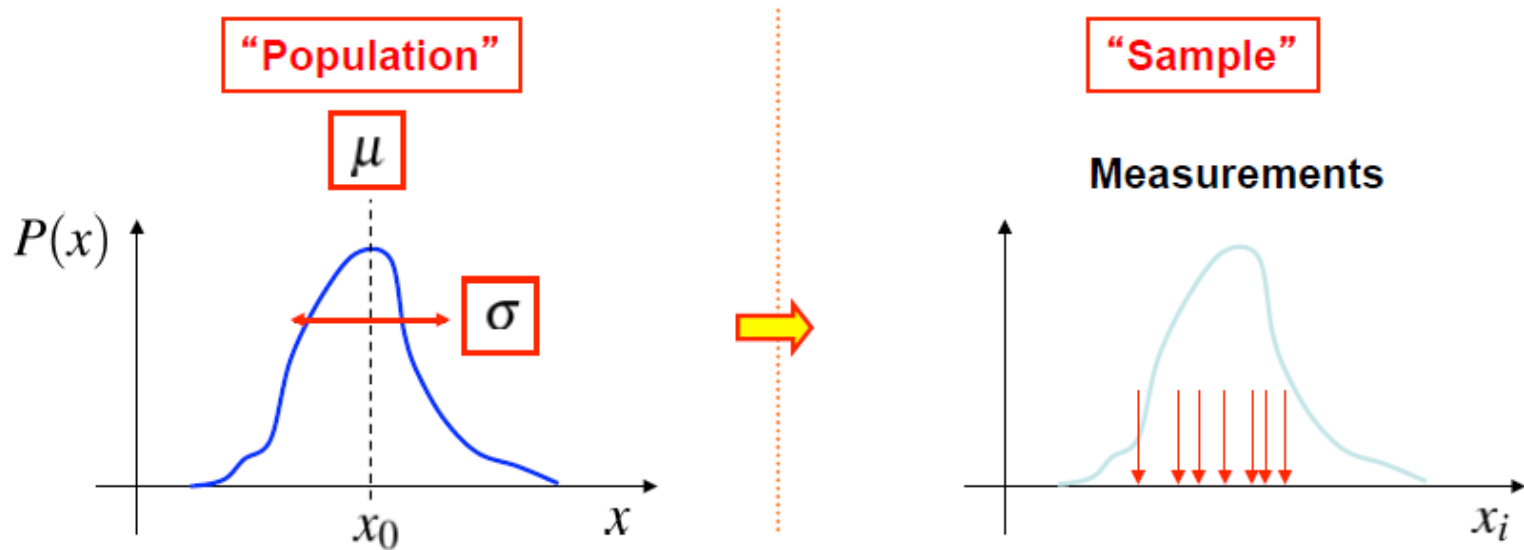
- The variance represents the width of the PDF about the mean
- Convenient to express this in terms of the **standard deviation** σ
- μ and σ describe the mean and “width” of a PDF
- Sometimes you will see the 3rd and 4th moments used (skewness, kurtosis) (these are not particularly useful)



$$\begin{aligned}\sigma^2 \equiv \langle (x - \mu)^2 \rangle &= \langle x^2 - 2\mu x + \mu^2 \rangle \\ &= \langle x^2 \rangle - 2\mu \langle x \rangle + \mu^2 \\ &= \langle x^2 \rangle - 2\mu^2 + \mu^2 \\ &= \langle x^2 \rangle - \mu^2\end{aligned}$$

Estimating the Mean and Variance

- ★ In general do not know the PDF – instead have a number of measurements distributed according to the PDF
- ★ Unless one has a infinite number of measurements cannot fully reconstruct the PDF (not a particularly useful thing to do anyway)
- ★ But can obtain unbiased estimates of the mean and variance



- ★ **Best estimate** of mean of distribution is the mean of the sample

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

- ★ Can also define **sample variance**

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

- ★ How does sample variance s^2 relate to true variance σ^2 ?
- ★ Can calculate average value of variance

$$\begin{aligned} \langle s^2 \rangle &= \langle (x_i - \bar{x})^2 \rangle \\ &= \langle x_i^2 \rangle - 2 \langle x_i \frac{1}{n} \sum_j x_j \rangle + \frac{1}{n^2} \langle [\sum_j x_j]^2 \rangle \\ &= \langle x_i^2 \rangle - \frac{2}{n} \langle x_i^2 + \sum_{j \neq i} x_i x_j \rangle + \frac{1}{n^2} (n \langle x_i^2 \rangle + n(n-1) \langle x_i x_j \rangle_{i \neq j}) \\ &= \langle x^2 \rangle - \frac{1}{n} \langle x^2 \rangle + \frac{(n-1)}{n} \langle x_i x_j \rangle_{i \neq j} \\ &= \frac{(n-1)}{n} (\langle x^2 \rangle - \langle x_i x_j \rangle_{i \neq j}) \\ &= \frac{(n-1)}{n} (\langle x^2 \rangle - \mu^2) = \frac{n-1}{n} \sigma^2 \end{aligned}$$

Question 1: prove

$$\langle x_i x_j \rangle_{i \neq j} = \mu^2$$

what assumption have you made?

- ★ Hence, on average, the sample variance is a factor $\frac{n-1}{n}$ smaller than the true variance
- ★ For an **unbiased estimate** of the true variance for a single measurement use:

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- ★ For the best **unbiased estimate** of the true mean use the sample mean:

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

- ★ What is the “error” (i.e. square root of the variance) on the sample mean ?

$$\begin{aligned} \text{Var}(\bar{x}) \equiv \sigma_{\bar{x}}^2 &= \langle (\bar{x} - \mu)^2 \rangle \\ &= \langle \left(\frac{1}{n} \sum_i x_i - \mu \right)^2 \rangle \\ &= \frac{1}{n^2} n \langle x^2 \rangle + \frac{n(n-1)}{n^2} \langle x_i x_j \rangle_{i \neq j} - 2\mu \langle \bar{x} \rangle + \mu^2 \\ &= \frac{\langle x^2 \rangle}{n} + \frac{n-1}{n} \mu^2 - \mu^2 \\ &= \frac{\langle x^2 - \mu^2 \rangle}{n} = \frac{\sigma^2}{n} \end{aligned}$$

- ★ Hence the uncertainty on the mean is \sqrt{n} smaller than the uncertainty on a single measurement

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- ★ Note: this is general result – doesn't rely on distribution
- ★ Of course we only have an **estimate** of σ , so our **best (unbiased) estimate** of the uncertainty on the mean is:

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{n}} s_{n-1}$$

- ★ There is one final question we can ask... what is the uncertainty on our estimate of the uncertainty. The answer to this question depends on the form of the PDF.
 - We'll come back to this in the context of a Gaussian distribution.....

QUESTION 2

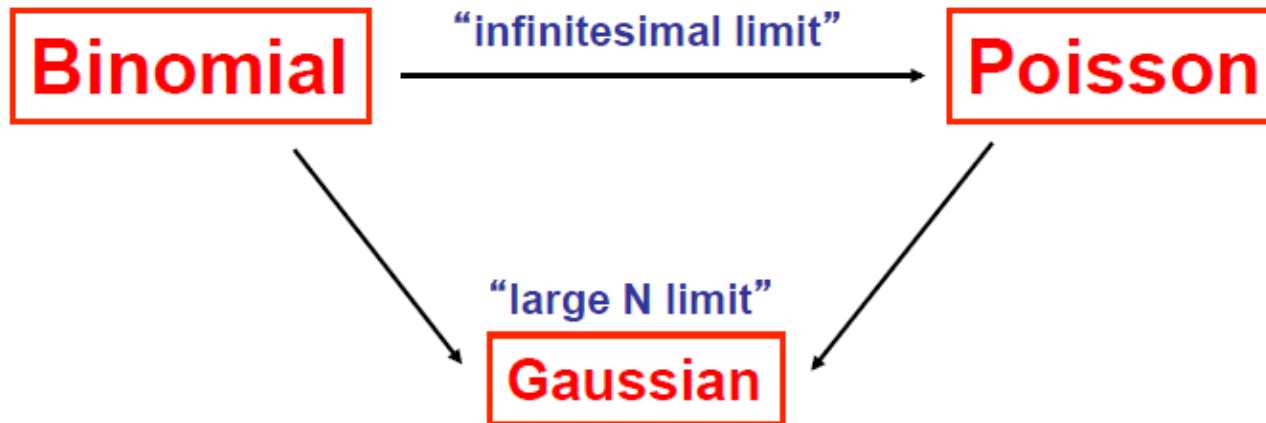
Given 5 measurements of a quantity x : 10.2, 5.5, 6.7, 3.4, 3.5

What is the **best estimate** of x and what is the **estimated** uncertainty?

For later, **how well do you know the uncertainty?**

Special Probability Distributions

- ★ So far, dealt in generalities
- ★ Now consider some special distributions...
- ★ Simplest case “Binomial distribution”
 - ♦ Random process with two outcomes with probabilities p and $(1-p)$
 - ♦ Repeat process a fixed number of times \Rightarrow distribution of outcomes
- ★ Next simplest, “Poisson distribution”
 - ♦ Discrete random process with fixed mean
- ★ Then, “Gaussian distribution”
 - ♦ Continuous “high statistics” limit



Binomial Distribution

- ★ Applies for a fixed number of trials when there are two possible outcomes, e.g.
 - ♦ Toss an unbiased coin ten times, how many heads ?

$$P(r; n) = {}^n C_r p^r (1-p)^{n-r}$$

$$\begin{aligned}\bar{x} &= \frac{\sum_{r=0}^n rP(r)}{\sum_0^n P(r)} = \sum_0^n rP(r) \\ &= \sum_{r=0}^n r p^r (1-p)^{n-r} \frac{n!}{r!(n-r)!} \\ &= np \sum_{r=1}^n p^{(r-1)} (1-p)^{(n-r)} \frac{(n-1)!}{(r-1)!(n-r)!} && \text{(n=0 term is zero)} \\ &= np \sum_{r'=0}^{n-1} p^{r'} (1-p)^{(n-1-r')} \frac{(n-1)!}{r'!(n-1-r')!} && \text{(let } r' = r-1\text{)} \\ &= np \sum_{r=0}^{n-1} P(r; n-1) \leftarrow \text{normalised to unity} \\ &= np\end{aligned}$$

★ Hence

$$\bar{x} = np$$

(hardly a surprising result)

Variance of the binomial distribution

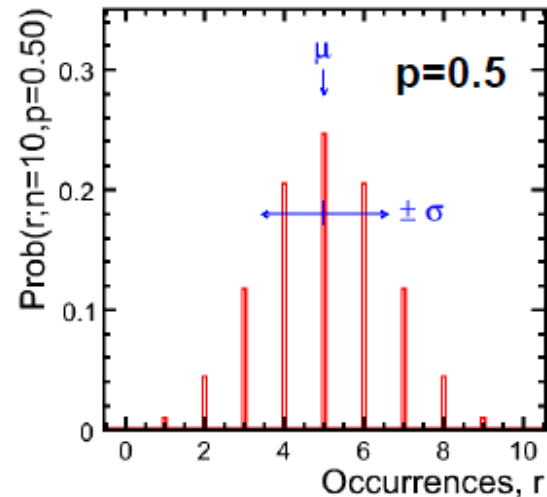
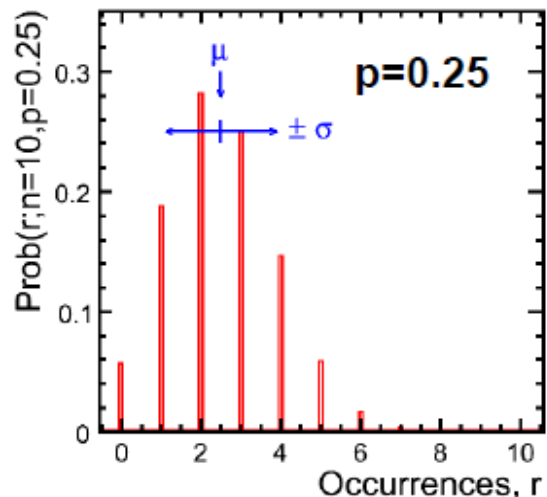
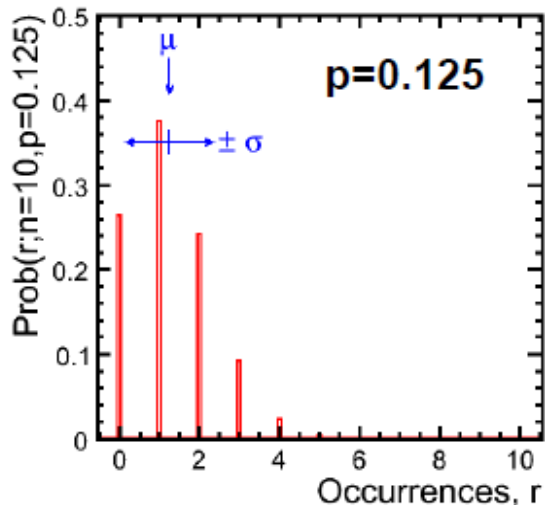
$$\begin{aligned} \text{Var}(r) &= \langle (r - \mu)^2 \rangle = \langle r^2 \rangle - \mu^2 \\ \langle r^2 \rangle &= \frac{\sum r^2 P(r;n)}{P(r;n)} = \sum_{r=0}^n r^2 p^r (1-p)^{n-r} \frac{n!}{r!(n-r)!} \\ &= np \sum_{r=1}^n r p^{r-1} (1-p)^{n-r} \frac{(n-1)!}{(r-1)!(n-r)!} \\ &= np \sum_{r'=0}^{n-1} (r'+1) p^{r'} (1-p)^{n-1-r'} \frac{(n-1)!}{r'!(n-1-r')!} \\ &= np \sum_{r=0}^{n-1} P(r;n-1) + np \sum_{r=0}^{n-1} r P(r;n-1) \\ &= np + np \times (n-1)p \\ \langle r^2 \rangle &= np(np - p + 1) \end{aligned}$$



$$\begin{aligned} \text{Var}(r) &= \langle r^2 \rangle - \mu^2 = np(np - p + 1) + np - (np)^2 \\ &= np(1 - p) \end{aligned}$$

$$\boxed{\text{Var}(r) = np(1 - p)}$$

e.g. $n=10$



★ What is the meaning of σ ?

- By definition, σ , is root of the mean square (rms) deviation from the mean

$$\sigma \equiv \langle (r - \mu)^2 \rangle^{\frac{1}{2}}$$

- For a binomial distribution $\sigma = \sqrt{np(1-p)}$
- It provides a well-defined **measure** of the spread about the mean
- For above values: 62 %, 57 %, and 66 % of distribution within $\pm 1 \sigma$ of mean
Answer depends on n and p , but roughly ~55-70%

Example: Efficiency Uncertainty

- ★ Suppose you use MC events to determine a selection efficiency
 - ♦ **m** out of **n** events pass some selection, what is the efficiency and uncertainty
- ★ This is a binomial process (**fixed number of trials**). Hence the number of events passing the selection will be distributed as:

$$P(m; n) = {}^n C_m \varepsilon^m (1 - \varepsilon)^{n-m}$$

- ★ Want to quote **best estimate** of the efficiency and the **best estimate** of the uncertainty (i.e. square root of the variance).

- ★ Best estimate of efficiency is “clearly”: $\varepsilon_e = \frac{m}{n}$

- ★ From properties of binomial distribution expect

$$\sigma^2 = \langle \varepsilon^2 \rangle = n\varepsilon(1 - \varepsilon) \times \frac{1}{n^2}$$

$$\sigma^2 = \frac{\varepsilon(1 - \varepsilon)}{n} \quad \left(= \frac{m(n - m)}{n^3} \right)$$

e.g. 90 out of 100 events pass trigger requirements,

$$\varepsilon = 0.90 \pm 0.03$$

A more advanced analysis

- ★ Asserted that our best estimate of the true efficiency ε is $\varepsilon_e = \frac{m}{n}$

Suppose we repeated the experiment many times

$$\langle \varepsilon_e \rangle = \frac{\langle m \rangle}{n} = \frac{n\varepsilon}{n} = \varepsilon$$

so on average this procedure gives an **unbiased estimate** of ε

GOOD

- ★ What about our estimate for the variance ?

$$\sigma_e^2 = \frac{\varepsilon_e(1 - \varepsilon_e)}{n} = \frac{m(n - m)}{n^3}$$

Again suppose we repeated the experiment many times

$$\begin{aligned} \langle \sigma_e^2 \rangle &= \frac{n\langle m \rangle}{n^3} - \frac{\langle m^2 \rangle}{n^3} \\ &= \frac{n^2\varepsilon}{n^3} - \frac{n^2\varepsilon^2 - n\varepsilon^2 + n\varepsilon}{n^3} \\ &= \frac{\varepsilon(1 - \varepsilon)}{n} + \frac{\varepsilon(1 - \varepsilon)}{n^2} = \frac{n + 1}{n^2} \varepsilon(1 - \varepsilon) \\ &= \frac{n + 1}{n} \sigma^2 \end{aligned}$$

GOOD ENOUGH

a problem ...

$$\sigma^2 = \frac{\varepsilon(1 - \varepsilon)}{n}$$

- ★ Suppose you want to estimate a trigger efficiency based on 100 MC events
- ★ If all the MC events pass the trigger selection...
 - best estimate of efficiency is 100 %
 - but what about the uncertainty on the efficiency ?
 - the above equation would suggest **zero**
 - this is clearly nonsense
 - **so what's wrong ?**

We'll come back to this in lecture 4...

The Poisson Distribution

- ★ Probably the most important distribution for experimental particle physicists
- ★ Appropriate for discrete counts at a **fixed rate**
 - e.g. in time t , on average expect μ events

$$p(n; \mu) = \frac{\mu^n e^{-\mu}}{n!}$$

- ★ The form of this equation is not immediately obvious (unlike that of the binomial distribution) – so (for completeness) derive the Poisson Distribution...
- ★ In time t , on average expect μ events. Now divide t into N intervals of δt
 - Probability of **one event** on δt is δp

$$\delta p = \mu \frac{\delta t}{t} = \frac{\mu}{N}$$

- Probability of getting two events is negligibly small
- Hence the problem has been transformed into N trials each with two discrete outcomes, i.e. a **binomial distribution**

$$p(n; \mu) = \lim_{N \rightarrow \infty} \delta p^n (1 - \delta p)^{N-n} \frac{N!}{n!(N-n)!}$$

The Poisson Distribution

$$P = (\delta p)^n (1 - \delta p)^{N-n} \frac{N!}{n!(N-n)!}$$

$$\ln P = n \ln \delta p + (N-n) \ln(1 - \delta p) + \ln N! - \ln n! - \ln(N-n)!$$

First consider:

$$\begin{aligned} (N-n) \ln(1 - \delta p) &= (N-n) [-\delta p + \cancel{(\delta p)^2/2} + \dots] \\ &\approx -N\delta p + n\delta p \\ &= -\mu + \frac{n}{N}\mu \end{aligned}$$

hence

$$\lim_{N \rightarrow \infty} \{(N-n) \ln(1 - \delta p)\} = -\mu$$

Stirling's approx

Now consider:

$$\begin{aligned} \ln \frac{N!}{(N-n)!} &= N \ln N - N - (N-n) \ln(N-n) + (N-n) \\ &= N \ln N + n - (N-n) \ln \left(1 - \frac{n}{N}\right) - (N-n) \ln N \\ &\approx n \ln N + n + (N-n) \frac{n}{N} \\ &= \ln N^n + \cancel{\frac{n^2}{N}} \end{aligned}$$

hence

$$\lim_{N \rightarrow \infty} \left\{ \frac{N!}{(N-n)!} \right\} = N^n$$

So finally,

$$P(n; N) = (\delta p)^n (1 - \delta p)^{N-n} \frac{N!}{n!(N-n)!}$$

becomes:

$$P(n; \mu) = (\delta p)^n e^{-\mu} \frac{N^n}{n!} = \left(\frac{\mu}{N}\right)^n e^{-\mu} \frac{N^n}{n!}$$

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

★ Check that the Poisson distribution is normalised...

$$\begin{aligned} \sum_{n=0}^{\infty} P(n; \mu) &= e^{-\mu} \left(1 + \frac{\mu}{1!} + \frac{\mu^2}{2!} + \dots\right) \\ &= e^{-\mu} e^{+\mu} = 1 \end{aligned}$$

Properties of the Poisson Distribution

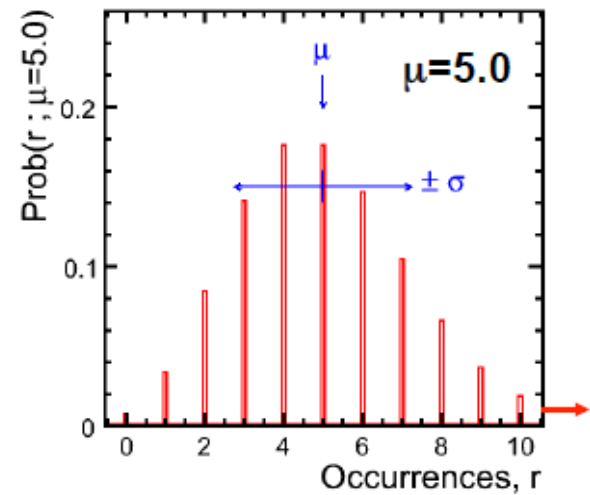
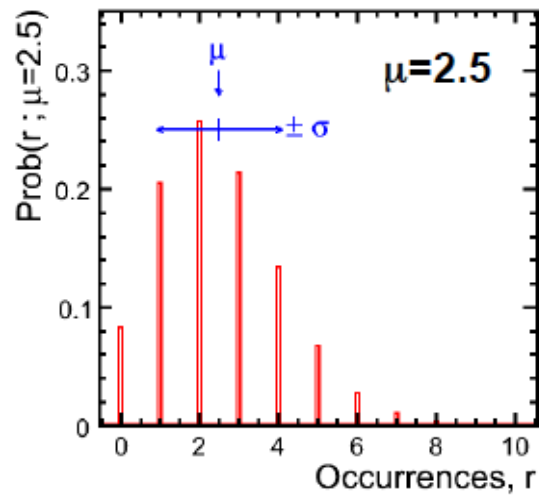
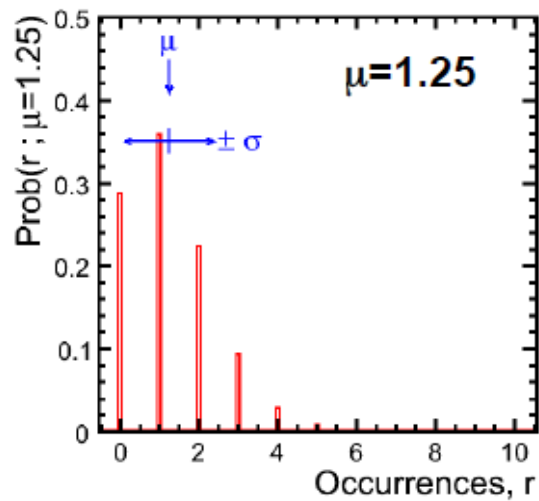
$$\begin{aligned}\langle n \rangle &= \sum_{n=0}^{\infty} nP(n; \mu) = \sum_{n=0}^{\infty} n \frac{\mu^n e^{-\mu}}{n!} \\ &= \sum_{n=1}^{\infty} n \frac{\mu^n e^{-\mu}}{n!} \\ &= \mu \sum_{n=1}^{\infty} \frac{\mu^{n-1} e^{-\mu}}{(n-1)!} \\ &= \mu \sum_{n'=0}^{\infty} \frac{\mu^{n'} e^{-\mu}}{n'!} \\ &= \mu \sum_{n=0}^{\infty} P(n; \mu) \\ &= \mu\end{aligned}$$

$$\langle n \rangle = \mu$$

$$\begin{aligned}\langle n^2 \rangle &= \sum_{n=0}^{\infty} n^2 P(n; \mu) = \sum_{n=0}^{\infty} n^2 \frac{\mu^n e^{-\mu}}{n!} \\ &= \sum_{n=1}^{\infty} n^2 \frac{\mu^n e^{-\mu}}{n!} \\ &= \mu \sum_{n=1}^{\infty} n \frac{\mu^{n-1} e^{-\mu}}{(n-1)!} \\ &= \mu \sum_{n'=0}^{\infty} (n'+1) \frac{\mu^{n'} e^{-\mu}}{n'!} \\ &= \mu \left\{ \sum_{n=0}^{\infty} nP(n; \mu) + \sum_{n=0}^{\infty} P(n; \mu) \right\} \\ &= \mu^2 + \mu \\ \sigma^2 = \text{Var}(n) &= \langle n^2 \rangle - \langle n \rangle^2 \\ &= \mu\end{aligned}$$

$$\sigma^2 = \mu$$

e.g. $\mu=1.25, 2.5, 5.0$



$$\langle N \rangle = \mu \quad \sigma = \sqrt{\mu}$$

Example I

★ Suppose I am trying to measure a cross section for a process

- observe N events for an integrated luminosity of \mathcal{L}
- for this luminosity the expected number of events is

$$\mu = \sigma \mathcal{L}$$

- observed number of events will be Poisson distributed according to μ
- our best unbiased estimate of μ is simply the number of observed events

$$\mu_e = N$$

- for a Poisson distribution the variance is equal to the mean
- hence we can **estimate** the uncertainty on the **estimated mean** as \sqrt{N}

$$\mu_e = N \pm \sqrt{N}$$

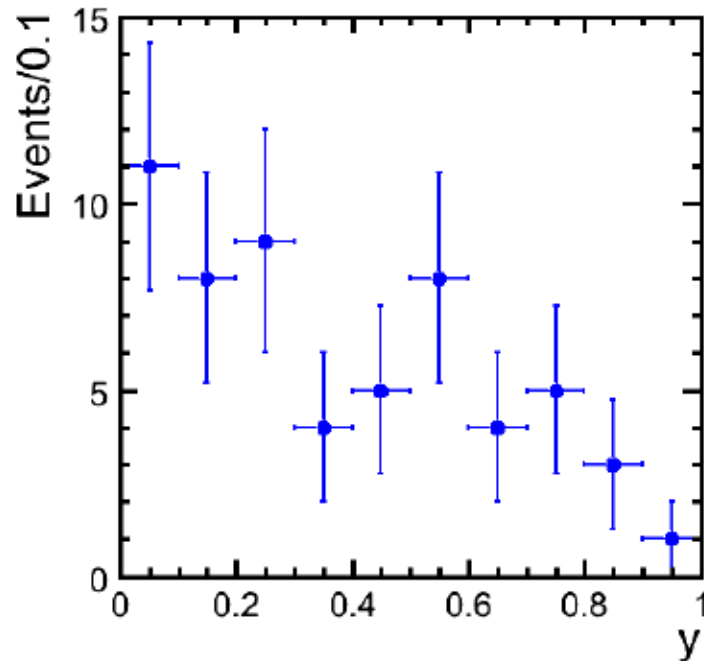
$$\sigma = \frac{1}{\mathcal{L}}(N \pm \sqrt{N})$$

NOTE: if you observe N events, the **estimated** uncertainty on the **mean of the underlying Poisson distribution** is \sqrt{N}
: it is not the “error” on N – there is no uncertainty on what you counted

★ Poisson fluctuations are the ultimate limit to any counting experiment

Example II

- ★ Suppose a colleague makes a histogram of event counts as a function of y
 - the histogram includes errors bars (made by root)



- ★ How should you interpret the error bars
 - If symmetric then probably \sqrt{N}
 - i.e. they indicate the expected "spread" assuming the mean expected counts in that bin are equal to the observed value
 - For large N this is not unreasonable
 - But for small N this doesn't make much sense...

High Statistics Limit of Poisson Distribution

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

$$\begin{aligned}\text{let } f(x) &= \ln P(x; \mu) \\ &= -\mu - \ln x! + x \ln \mu \\ &\approx -\mu + x \ln x - x + x \ln \mu\end{aligned}$$

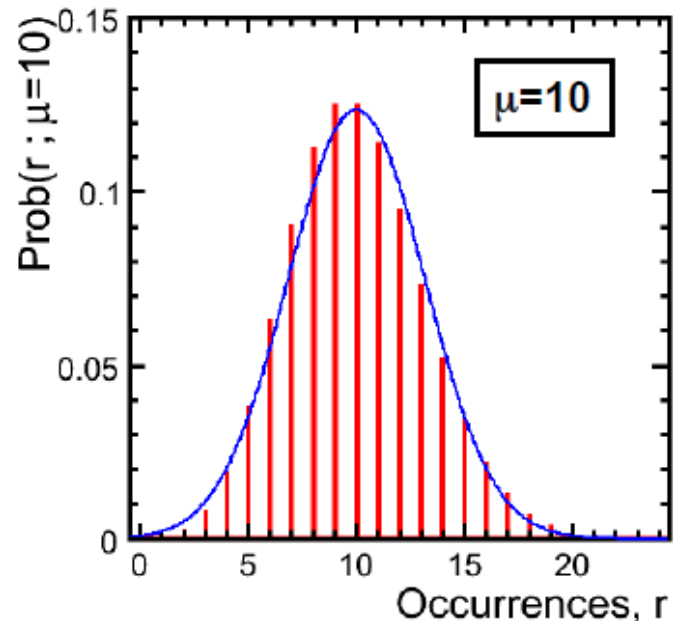
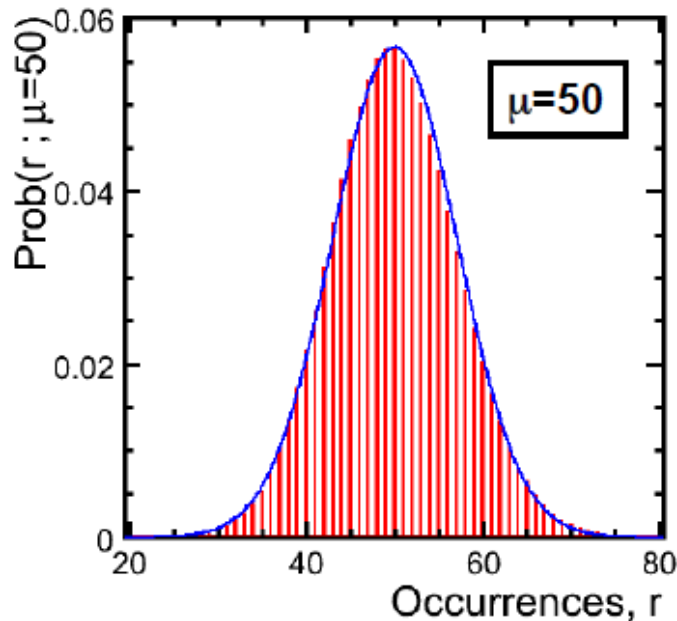
$$\begin{aligned}\text{hence } f'(x) &= \ln \mu - \ln x \\ f''(x) &= -1/x\end{aligned}$$

Taylor expansion about mean:

$$\begin{aligned}f(x) &= f(\mu) + (x - \mu)f'(\mu) + \frac{1}{2!}(x - \mu)^2 f''(\mu) + \frac{1}{3!}(x - \mu)^3 f'''(\mu) \dots \\ &= f(\mu) - \frac{(x - \mu)^2}{2\mu} + \frac{(x - \mu)^3}{6\mu^2} + \dots\end{aligned}$$

$$P(x; \mu) \approx ke^{-\frac{(x-\mu)^2}{2\mu}}$$

$$P(x; \mu) \approx ke^{-\frac{(x-\mu)^2}{2\mu}}$$



★ Even for relatively small μ , (apart from in the extreme tails), a **Gaussian Distribution** is a pretty good approximation

Next lecture

- ★ Investigate the treatment of statistics in the **Gaussian Limit**
 - The central limit theorem
 - Gaussian errors
 - Error propagation
 - Combination of measurements
 - Multi-dimensional Gaussian errors
 - Error Matrix**