Counting Statistics

Introduction

Th. Cocolios, N. Severijns (KU Leuven)
E.Cortina (UCLouvain)

Outline

Section 1

Introduction

1. Introduction

Counting Statistics

1. The study of radiation detection is subjected to inherent fluctuations
   - Basics processes based in QM
   - The number of interactions is random by nature
   - Unavoidable sources of uncertainty
2. Statistics it's a tool that will allow us to connect
   - Theory. In principle no deal with statistics
   - Measurement. By intrinsic construction, data is described with statistical methods
3. Radiation detection is based largely in the so called counting experiments → counting statistics.
4. There are excellent books:
   - Bevington, Data Reduction and Error Analysis
   - Barlow, Statistics, A Guide to the use of Statistical Methods in the Physical Sciences
   - Lyons, Statistics for Nuclear and Particle Physics

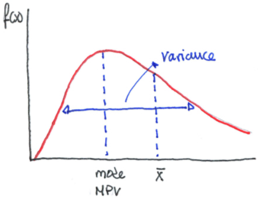
Section 2

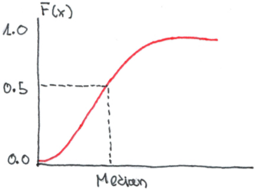
Random Variables

2. Random Variables

Random Variables

- A random variable is a variable whose value is not known.
  - It can take various values (continuous or discrete)
  - A probability distribution (called pdf) gives the probability to obtain the different values
  - The integral (sum) of pdf gives the cumulative distribution function (cdf)





mean	$E[x]$	$\mu = \frac{\int x f(x) dx}{\int f(x) dx}$	$\mu = \frac{\sum x f(x)}{\sum f(x)}$
variance	$E[(x-\mu)^2]$	$\sigma^2 = \frac{\int (x-\mu)^2 f(x) dx}{\int f(x) dx}$	$\sigma^2 = \frac{\sum (x-\mu)^2 f(x)}{\sum f(x)}$

Data Description: AVERAGE

Mean $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Geometric Mean $\overline{x} = \sqrt[N]{x_1 \cdot x_2 \ldots x_N}$

Harmonic Mean $\overline{x} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}}$

Root Mean Square $\overline{x} = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_N^2}{N}}$

Mode Most probable value

Median Half way point in cdf

Weighted Mean $\overline{x} = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$

Data Description: SPREAD

Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2 = V(x)$
Unbiased Variance	$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$
Standard deviation	$\sigma = \sqrt{V(x)}$
Mean Absolute dev.	$\sigma = \frac{1}{N}\sum \lvert x_i - \overline{x}\rvert$
Range	$\sigma = x_{max} - x_{min}$
Interquantile	$\sigma = x_{75\%} - x_{25\%}$
FWHM	Full Width at Half Maximum

Localize the mode

Draw an horizontal line at half heigth

FWHW = distance between intersection points

In a gaussian FWHM=2.35σ

Data Description: SPREAD more than one variable

- 1 variable

$$V(x) = \frac{1}{N}\sum_i (x_i - \overline{x})^2$$

- 2 variables

$$V(x,y) = \frac{1}{N}\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\,\overline{y}$$

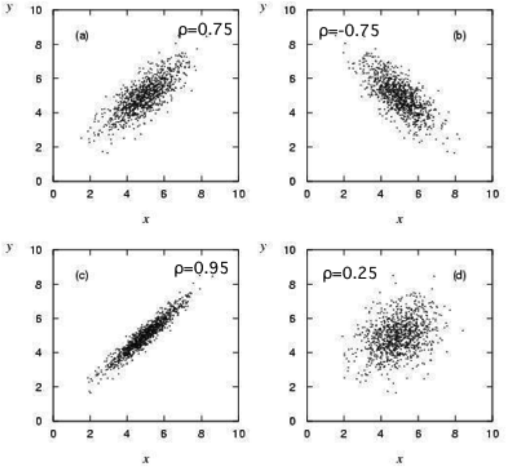
$$V(x,x) = V(x) = \sigma_x^2$$
$$V(y,y) = V(y) = \sigma_y^2$$
$$V(x,y) = V(y,x) = cov(x,y)$$

$$\begin{pmatrix} \sigma_x^2 & cov(x,y) \\ cov(x,y) & \sigma_y^2 \end{pmatrix}$$

correlation $\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

- Easy generalization for >2 variables

Correlation



Section 3

Important pdf

3. Important pdf
   - Binomial Distribution
   - Poisson Distribution
   - Gaussian Distribution
   - χ^2 Distribution
   - Beta Distribution
   - Uniform distribution

Binomial Distribution

- Bernoulli process:
  1. N trials (integer and finite number of trials)
  2. Each trial has a binary outcome: success or failure
  3. Probability of success (p) is constant from trial to trial
  4. The trials are independent
- Given N trials of a Bernoulli process with probability of success p, the pdf is given by the Bernoulli distribution

$$B(n)=\binom{N}{n}p^n(1-p)^{N-n}$$

$$\binom{N}{n}=\frac{N!}{n!(N-n)!}$$

$A)\ \sum_{n=0}^{N} B(n)=1$

$B)\ \mu=\sum_{n=0}^{N} nB(n)=Np$

$C)\ \sigma^2=\sum_{n=0}^{N}(n-\mu)^2 B(n)=Np(1-p)$

http://www.distributome.org/tools.html
http://www.distributome.org/V3/sim/BinomialSimulation.html

Bernoulli process example: Radioactivity

The number of disintegrations of N radioactive atoms in a time t is a bernoulli process with a probability success of

$$p = 1 - e^{-\lambda t}$$

	Bernoulli conditions	Radioactivity
1)	N trials	N atoms
2)	Binary outcome	Each atom can decay or not
3)	p constant for all trials	True if λ small
4)	Trials independent	ok

Example: Binomial distribution

Consider a ^{42}K source with an activity of 37 Bq.

$\lambda = 1.55 \times 10^{-5}\, s^{-1} \rightarrow p = 1 - e^{-\lambda t} \simeq \lambda = 1.55 \times 10^{-5}$

(a) Mean disintegration rate

$$\begin{aligned} \mu = Np &= N(1 - e^{-\lambda t}) \\ &= N[1 - (1 - \lambda t)] \\ &= N\lambda t \xrightarrow{t=1s} = N\lambda = A \\ &= 37\, s^{-1} \end{aligned} \quad \Rightarrow \quad N = \frac{\mu}{\lambda} = 2.39 \times 10^{6}$$

(b) Standard deviation

$$\sigma = \sqrt{Npq} = 6.09\, s^{-1}$$

(c) Probability of counting exactly 40 counts in 1 second

$$B(n) = \binom{N}{n} p^{n} (1-p)^{N-n}$$

$$\Big\downarrow \; n = 40, \; N = 2.39 \times 10^{6}$$

$$\binom{N}{n} \simeq \frac{N^{n}}{n!}$$

$$B(40) = \frac{(2.39 \times 10^{6})^{40}}{40!} 0.0000155^{40} 0.9999845^{2.36 \times 10^{6} - 40}$$

Using logs to solve $\rightarrow$ $B(40) = 0.0561$

Poisson Distribution

- If $N >> 1$, $N >> n$, and $p << 1$, instead of the binomial distribution we can use the Poisson distribution:

$$P(n) = \frac{\mu^n e^{-\mu}}{n!} \quad \rightarrow \quad \begin{aligned} \text{mean} &= \mu \\ \sigma^2 &= \mu \end{aligned}$$

http://www.distributome.org/V3/sim/PoissonSimulation.html

For the example used in binomial distribution:

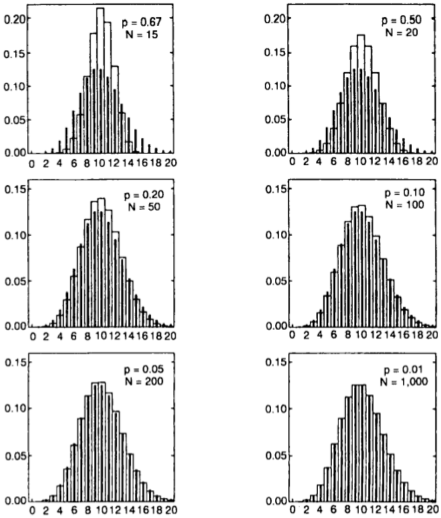
	40 events	σ
Binomial	B(40) = 0.0561	6.09
Poisson	P(40) = 0.0559	6.08

Poisson Distribution

- Poisson distribution can also be deduced from three postulates
  1. The number of successes in any interval is INDEPENDENT of the number of successes in any other disjoint time interval.
  2. The probability if a single success in a very short interval is proportional to the length of the interval.
  3. The probability of more than one success in a very short interval is negligible.

- Examples of Poisson processes:
  - Number of accidents/catastrophes
  - Number of eggs pond by a brood of hens
  - Number of cosmic rays or background events registered by a counter

Comparison between Binomial and Poisson



Gaussian Distribution

- In case of $p \to 0$ and $N \to \infty$ both Binomial and Poisson tends to a Gaussian

$$G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

http://www.distributome.org/V3/sim/NormalSimulation.html

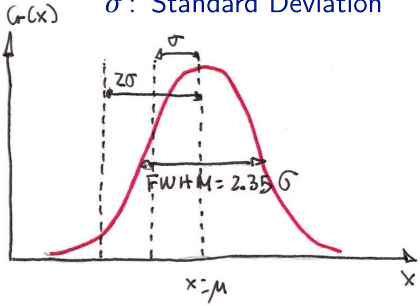
- Gaussian is a continuous distribution with two parameters:

μ : Mean

σ : Standard Deviation

Best Estimators $\rightarrow$ $\mu = \frac{1}{N}\sum_i x_i$

$\sigma^2 = \frac{1}{N-1}\sum_i (x_i - \mu)^2$



ϵ_0	$\int_{-\epsilon_0}^{\epsilon_0} G(x)dx$
0.674σ	0.500
1.00σ	0.683
1.64σ	0.900
1.96σ	0.950
2.00σ	0.955
2.58σ	0.990
3.00σ	0.997

Gaussian Distribution

- Reduced gaussian: gaussian with $\mu = 0$ and $\sigma = 1$

$$z = \frac{x-\mu}{\sigma} \quad \rightarrow \quad g(z) = \frac{1}{2\pi} e^{-\frac{z^2}{2}}$$

http://davidmlane.com/hyperstat/z_table.html

- In case of counting statistics: $\sigma^2 = \mu$

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

  - This function is defined only for integer values of x
  - Normalization constraint

$$\sum_i f(x_i) = 1$$

χ^2 Distribution

- Distribution closely related with data modeling.
- Given a set of N measurements (x_i, y_i)
  - with errors in the y_i measurements of σ_i,
  - a model for the $y_i = f(x_i;\theta)$ where $\theta = (\theta_1, \theta_2, ..., \theta_m)$ is a set of m parameters,

  one can evaluate the goodness of fit based on the χ^2 function:

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i;\theta)]^2}{\sigma_i^2}$$

- χ^2 is also a random variable with pdf:

$$p_{\chi^2}(z;n) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{-\frac{z}{2}}$$

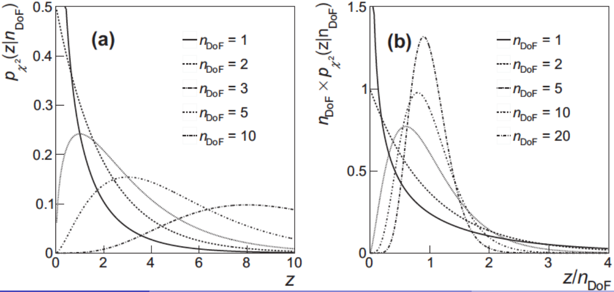
  where $z = \chi^2$ and $n = N - m$ is an integer called number of degrees of freedom and it correspond of the number of independent variables in a system or model.

χ^2 Distribution: Properties

$$E[z] = \int_0^\infty z \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} dz = n$$

$$Var[z] = \int_0^\infty (z-n)^2 \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} z^{\frac{n}{2}-1} e^{-\frac{z}{2}} dz = 2n$$

- χ^2 distribution peaks around n
- The scaled variable χ^2 per degrees of freedom (χ^2/n_{dof}) peaks at unity



Beta Distribution

- The beta distribution is defined as

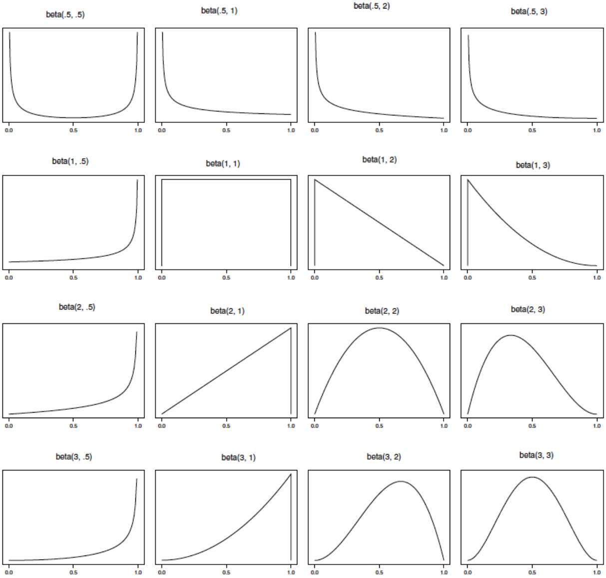
$$\mathrm{Beta}(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \qquad \begin{array}{r} x \in [0,1] \\ \alpha,\beta \in \mathbb{R}^{+} \end{array}$$

$$E[x] = \frac{\alpha}{\alpha+\beta}$$

$$Var[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- This distribution may be regarded as an "extension" of the binomial distribution ($\alpha = n+1,\ \beta = n$)
- Various α and β generate a wide range of shapes.
- Very useful and convenient to be used as "prior" in Bayesian statistics.

Beta distribution



Uniform distribution

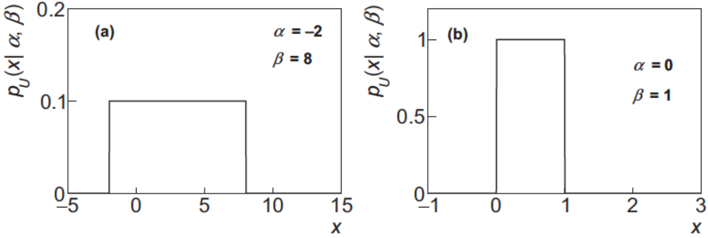
- In a uniform distribution all possible values of the random variable are equally probable

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\alpha+\beta} & \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \frac{\alpha+\beta}{2}$$

$$Var[x] = \frac{(\beta-\alpha)^2}{12}$$

- If $\alpha = 0$ and $\beta = 1$ → Standard Uniform Distribution



Section 4

Applications

Error Propagation

- Direct consequence of Gaussian distribution. If $u = u(x, y, z, \ldots)$

$$\sigma_u^2 = \left(\frac{\partial u}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial u}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial u}{\partial z}\right)^2 \sigma_x^2 + \cdots$$
$$2\left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial u}{\partial y}\right) cov(x,y) + 2\left(\frac{\partial u}{\partial x}\right)\left(\frac{\partial u}{\partial z}\right) cov(x,z) + \cdots$$

$$\sigma_u^2 = \sum_i \left(\frac{\partial u}{\partial x_i}\right)^2 \sigma_{x_i}^2 + 2\sum_i \sum_{j>i} \left(\frac{\partial u}{\partial x_i}\right)\left(\frac{\partial u}{\partial x_j}\right) cov(x_i, x_j)$$

- In case of independent variables ($cov(x_i, x_j) = 0$)

$$\sigma_u = \sqrt{\sum_i \left(\frac{\partial u}{\partial x_i}\right)^2 \sigma_{x_i}^2} < \sum_i \left|\frac{\partial u}{\partial x_i}\right| \sigma_{x_i}$$

Error Propagation: Examples

- $u = x \pm y$

$$\left.\begin{array}{l}\dfrac{\partial u}{\partial x} = +1 \\ \dfrac{\partial u}{\partial y} = \pm 1\end{array}\right\} \rightarrow \quad \begin{aligned} cov(x,y) = 0 &\rightarrow \sigma_u^2 = \sigma_x^2 + \sigma_y^2 \\ cov(x,y) \neq 0 &\rightarrow \sigma_u^2 = \sigma_x^2 + \sigma_y^2 \pm 2cov(x,y) \\ &\qquad = \sigma_x^2 + \sigma_y^2 \pm 2\rho_{xy}\sigma_x\sigma_y \end{aligned}$$

$$\text{If } \sigma_x = \sigma_y \rightarrow \left.\begin{array}{l}\rho_{xy} = -1 \text{ addition} \\ \rho_{xy} = +1 \text{ subtraction}\end{array}\right\} \quad \sigma_u = 0$$

- $u = Ax \rightarrow \sigma_u = A\sigma_x$
  - BEWARE with errors in count rates!!!
  - If x=100 counts in T=10s

$$\sigma_x = \sqrt{x} = 10$$

$$u = \frac{x}{T} = 10\, s^{-1} \rightarrow \sigma_u = \frac{\sigma_x}{T} = 1 \neq \sqrt{10}$$

Error Propagation: Examples

- $u = xy$

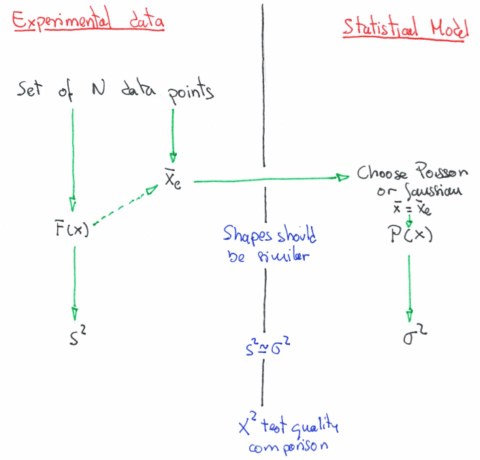
$$\sigma_u^2 = y^2\sigma_x^2 + x^2\sigma_y^2 \xrightarrow[\text{div by } u^2=(xy)^2]{} \left(\frac{\sigma_u}{u}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$$

- $u = x/y$

$$\sigma_u^2 = \frac{1}{y^2}\sigma_x^2 + \left(\frac{-x}{y^2}\right)^2\sigma_y^2 \xrightarrow[\text{div by } u^2=(x/y)^2]{} \left(\frac{\sigma_u}{u}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$$

Applications: Expected Fluctuations

- Are statistical fluctuations observed when a measurement is repeated N times compatible with the fluctuations expected from counting statistics?



χ^2 test

- Chi-square distribution (χ^2) is a function defined as:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma_i^2} \xrightarrow[\mu=\sigma^2=\overline{x}]{\text{Counting Statistics}} \chi^2 = \sum_{i=1}^{N} \frac{(x_i - \overline{x})^2}{\overline{x}}$$

- Chi-square is closely related to sample variance

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

- And both are related as:

$$\chi^2 = (N-1)\frac{s^2}{\overline{x}}$$

- $N-1$: Number of degrees of freedom $= \nu$
- $\frac{s^2}{\overline{x}} \simeq 1$. Any deviation from unity = comparison between the observed variance (s^2) and the predicted one ($\overline{x}$).

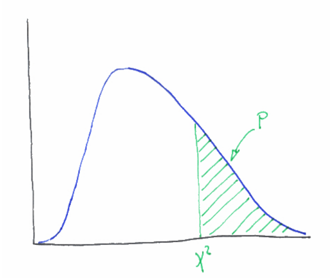
- Distribution of χ^2 depends on the number of degrees of freedom

http://www.distributome.org/V3/sim/ChiSquareSimulation.html_Distrib

χ^2 test

- The probability that a random sample from a Poisson distribution would have a value χ^2 greater than the one calculated with our data set as:

$$p = \int_{\chi^2}^{\infty} \chi^2(x;\nu)dx$$

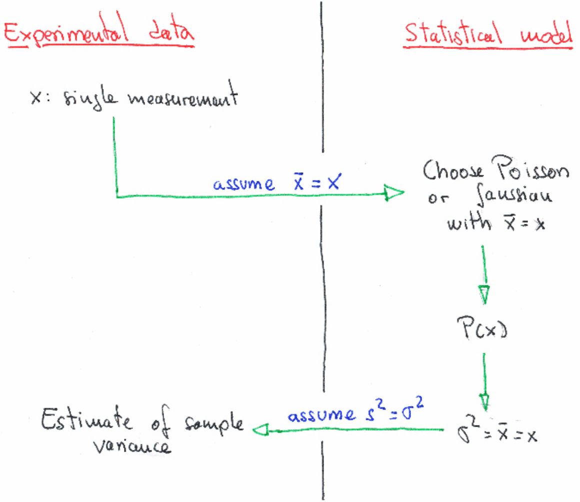


- $p < 0.05 \rightarrow$ Small fluctuations
- $p > 0.95 \rightarrow$ Large fluctuations
- Lots of "calculators" on internet

http://www.distributome.org/V3/calc/ChiSquareCalculator.html

Precision of a Single Measurement

- In most experiments we have just one measurement.
- Statistics allows us to estimate the uncertainty of this measurement.



Applications: Precision of a Single Measurement

- The best estimate of the deviation from the true mean of a single measurement is:

$$\sigma^2 = x \quad \rightarrow \quad \sigma = \sqrt{x}$$

- The relative error due the counting is:

$$\varepsilon_r = \frac{\sigma}{x} = \frac{\sqrt{x}}{x} = \frac{1}{\sqrt{x}}$$

  - The total number of events fix the relative error
  - To reduce ε_r a factor 2 we need 4 times more statistics
  - To reduce ε_r a factor 10 we need 100 times more statistics

- Be aware! This conclusion applies only to DIRECT count measurements. It cannot be applied to:
  - Counting rates
  - Sums or differences of counts
  - Averages of independent counts

In case of any derived quantity, error propagation should be applied

Optimization of Counting Experiments

- In general one measurement consists in two steps:
  - Measure $N_1 = S + B$ during a time T_1
  - Measure $N_2 = B$ during a time T_2
- The rate due to the source is

$$S + B = \frac{N_1}{T_1} \qquad B = \frac{N_2}{T_2} \quad \rightarrow \quad S = \frac{N_1}{T_1} - \frac{N_2}{T_2}$$

- Applying error propagation we obtain:

$$\sigma_S^2 = \left(\frac{\sigma_{N_1}}{T_1}\right)^2 + \left(\frac{\sigma_{N_2}}{T_2}\right)^2 = \left(\frac{S+B}{T_1}\right) + \left(\frac{B}{T_2}\right)$$

- If the total time $T = T_1 + T_2$ is fixed, which is the optimal fraction of the time measurement in order to minimize the error?

Optimization of Counting Experiments

- Differentiating the expression above we obtain

$$2\sigma_S d\sigma_S = -\frac{S+B}{T_1^2} dT_1 - \frac{B}{T_2^2} dT_2$$

- And applying the constraint that

$$T = T_1 + T_2 = \text{constant} \quad \rightarrow \quad dT = dT_1 + dT_2 = 0$$

- We obtain the optimum division of time as:

$$\left.\frac{T_1}{T_2}\right|_{\text{opt}} = \sqrt{\frac{S+B}{B}}$$

- And combining the equations above:

$$\frac{1}{T} = \left(\frac{\sigma_S}{S}\right)^2 \frac{S^2}{(\sqrt{S+B} + \sqrt{B})^2}$$

  where the total time is related to the relative error and the rates S and B

Optimization of Counting Experiments

$$\frac{1}{T} = \left(\frac{\sigma_S}{S}\right)^2 \frac{S^2}{(\sqrt{S+B}+\sqrt{B})^2}$$

- In case $S >> B$

$$\frac{1}{T} \simeq \left(\frac{\sigma_S}{S}\right)^2 S$$

  - Background has no statistical influence
  - Experiment should be designed to maximize S
- In case $S << B$

$$\frac{1}{T} \simeq \left(\frac{\sigma_S}{S}\right)^2 \frac{S^2}{4B}$$

  - In this case we should optimize S^2/B
  - We improve (optimize) if changing the conditions we obtain

$$\frac{(S^2/B)_{\text{new}}}{(S^2/B)_{\text{old}}} > 1$$

Limits of detectability

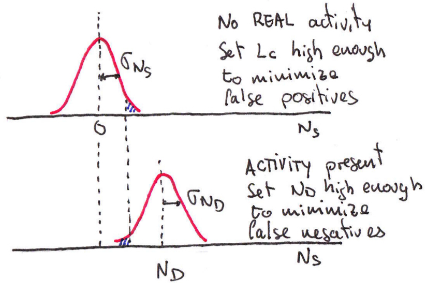
- What is the smallest signal that a system can detect?
  - MDA=Minimum Detectable Amount

$$N_S = N_T - N_B$$
$$\sigma_S^2 = \sigma_{N_T}^2 + \sigma_{N_B}^2$$

  - Let's compare N_S to a critical level L_C
  - L_C takes into account statistical fluctuations and instrumental variations
  - Simple rule: $N_S < L_C \rightarrow$ compatible with background
    $N_S > L_C \rightarrow$ activity is present

Limits of detectability: No real activity is present

$$\langle N_T \rangle = \langle N_B \rangle \rightarrow \langle N_S \rangle = 0$$
$$\sigma^2_{N_S} = \sigma^2_{N_T} + \sigma^2_{N_B} = 2\sigma^2_{N_B}$$
$$\sigma_{N_S} = \sqrt{2}\sigma_{N_B}$$



- ❒ We set L_C in such a way that only a fraction of the upper gaussian is above L_C
- ❒ Minimize the number of false-positive
- ❒ For $\pm 1.645\sigma$, 90% of data is within this interval
- ❒ Only 5% of measurements will be above L_C

$$L_C = 1.645\,\sigma_{N_S} = 1.645\sqrt{2}\sigma_{N_B}$$

$$L_C = 2.326\,\sigma_{N_B}$$

Limits of detectability: Real activity is present

- Any conclusion that there is no activity is a false-negative
- MDA: source strength needed to reduce false-negatives

- For 5% then: $N_D = L_C + 1.645\sigma$
- $N_D = N_S \rightarrow N_D = N_T - N_B$

$$\sigma^2_{N_D} = \sigma^2_{N_T} + \sigma^2_{N_B} = N_T + N_B$$
$$\sigma^2_{N_D} = N_D + 2N_B$$
$$\sigma_{N_D} = \sqrt{N_D + 2N_B}$$

- $L_C = 1.645\sigma_{N_B} = 1.645\sqrt{2N_B}$
- Substituting we get

$$N_D = 1.645(\sqrt{2N_B} + \sqrt{N_D + 2N_B})$$

$$\boxed{N_D = 4.65\sqrt{N_B} + 2.71}$$