Practical Statistics for Particle Physicists
Lecture 1

Harrison B. Prosper
Florida State University

European School of High-Energy Physics

Anjou, France

6 – 19 June, 2012

Outline

- Lecture 1
  - Descriptive Statistics
  - Probability
  - Likelihood
  - The Frequentist Approach – 1

- Lecture 2
  - The Frequentist Approach – 2
  - The Bayesian Approach

- Lecture 3 – Analysis Example

Practicum

I shall place some files (toy data and code) at

http://www.hep.fsu.edu/~harry/ESHEP12

e.g.,

topdiscovery.tar (already there)

contactinteractions.tar

just download and unpack

Descriptive Statistics

Descriptive Statistics – 1

Definition: A **statistic** is any function of the data ***X***.

Given a sample ***X*** $= x_1, x_2, \ldots x_N$, it is often of interest to compute statistics such as

the **sample average**

$$\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

and the **sample variance**

$$\boxed{S^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

In any analysis, it is good practice to study **ensemble averages**, denoted by < … >, of relevant statistics

Descriptive Statistics – 2

Ensemble Average	$< x >$
Mean	μ
Error	$\varepsilon = x - \mu$
Bias	$b = < x > - \mu$
Variance	$V = < (x - < x >)^2 >$
Mean Square Error	$\text{MSE} = < (x - \mu)^2 >$

Descriptive Statistics – 3

$$\text{MSE} = \langle (x - \mu)^2 \rangle$$
$$= V + b^2$$

Exercise 1: Show this

The **MSE** is the most widely used measure of **closeness** of an ensemble of statistics {**x**} to the **true value** $\boldsymbol{\mu}$

The **root mean square** (RMS) is

$$\text{RMS} = \sqrt{\text{MSE}}$$

Descriptive Statistics – 4

Consider the *ensemble* average of the *sample* **variance**

$$< S^2 > = < \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 >$$

$$= < \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \frac{2}{N} \sum_{i=1}^{N} x_i \bar{x} + \frac{1}{N} \sum_{i=1}^{N} \bar{x}^2 >$$

$$= \frac{1}{N} \sum_{i=1}^{N} < x_i^2 > - < \bar{x}^2 >$$

$$= < x^2 > - < \bar{x}^2 >$$

Descriptive Statistics – 5

The ensemble average of the sample variance

$$< S^2 > = < x^2 > - < \bar{x}^2 >$$

$$= < x^2 > - \frac{< x^2 >}{N} - \left(\frac{N-1}{N} \right) < x >^2$$

$$= V - \frac{V}{N}$$

has a negative bias of $-V / N$

Exercise 2: Show this

Descriptive Statistics – 6

Now, consider the variance of the sample average

$$< \Delta\overline{x}^2 > = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} < \Delta x_i \Delta x_j >$$

$$= \frac{1}{N^2} \left(\sum_{i=1}^{N} < \Delta x_i^2 > + \sum_{i=1}^{N} \sum_{j \neq i}^{N} < \Delta x_i \Delta x_j > \right)$$

where

$$\Delta\overline{x} \equiv \overline{x} - < x > \quad \text{and} \quad \Delta x_i \equiv x_i - < x >$$

Descriptive Statistics – 7

Suppose that the data are correlated as follows

$$< \Delta x_i \Delta x_j > = \rho V$$

then

$$< \Delta \bar{x}^2 > = \frac{1}{N^2}\left(\sum_{i=1}^{N} < \Delta x_i^2 > + \sum_{i=1}^{N}\sum_{j \neq i}^{N} < \Delta x_i \Delta x_j > \right)$$

$$= \frac{V}{N}\left(1 + (N-1)\rho\right)$$

Descriptive Statistics – Summary

The **sample average**
is an unbiased estimate
of the ensemble average

$$\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

The **sample variance**
is a biased estimate
of the ensemble variance

$$\boxed{S^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

The **variance of the sample average** decreases like **1/*N***
until we reach a limit imposed
by the degree of correlation in the data

$$\boxed{V_{\bar{x}} = \frac{V}{N}\left[1 + (N-1)\rho\right]}$$

Probability

Probability – 1

Basic Rules

1. $P(A) \geq 0$
2. $P(A) = 1$ if A is true
3. $P(A) = 0$ if A is false

Sum Rule

4. $P(A+B) = P(A) + P(B)$ if AB is false *

Product Rule

5. $P(AB) = P(A|B)\, P(B)$ *

***A+B = A or B, AB = A and B, A|B = A given that B is true**

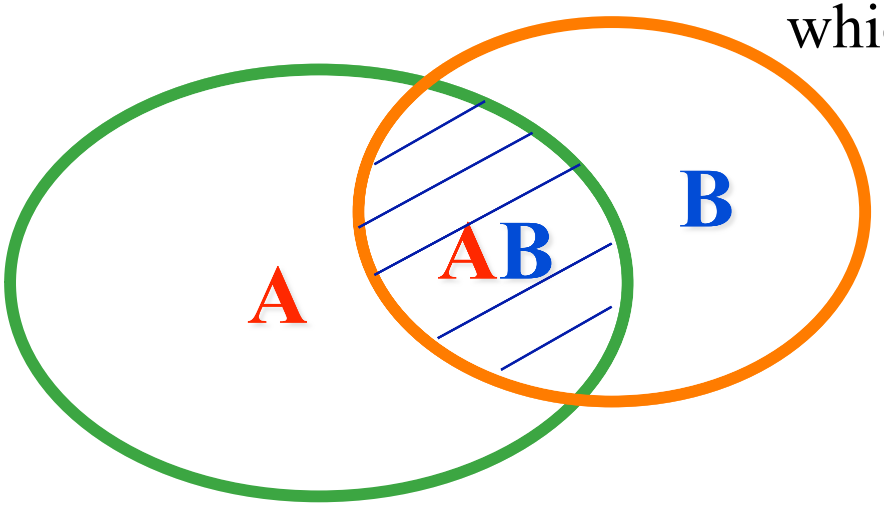
Probability – 2

By definition, the **conditional probability** of A *given* B is

$$\boxed{P(A \mid B) = \frac{P(AB)}{P(B)}}$$

P(A) is the probability of A *without restriction*.

P(A|B) is the probability of A when we *restrict* to the conditions under which B is true.



$$\boxed{P(B \mid A) = \frac{P(AB)}{P(A)}}$$

Probability – 3

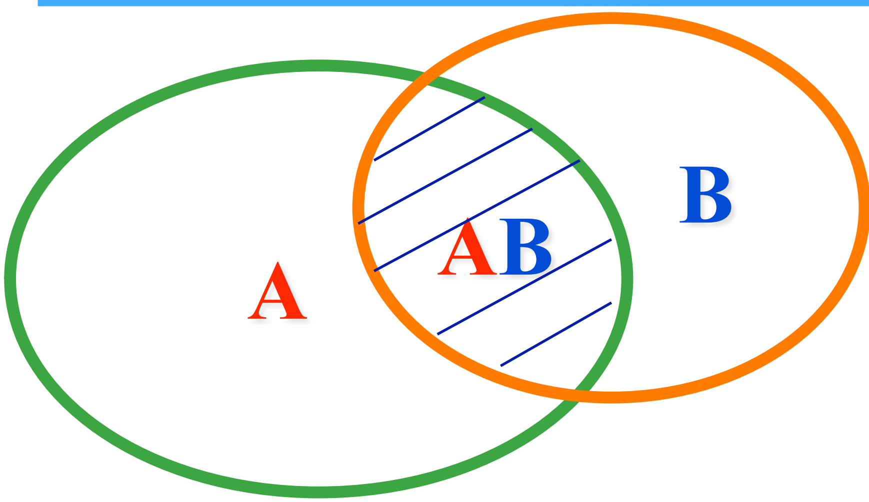
From $P(AB) = P(B \mid A)P(A)$

$= P(A \mid B)P(B)$

we deduce

***Bayes' Theorem*:**

$$\boxed{P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}}$$



Probability – 4

A and B are mutually exclusive if

$$P(AB) = 0$$

A and B are exhaustive if

$$P(A) + P(B) = 1$$

Theorem

$$P(A+B) = P(A) + P(B) - P(AB)$$

$\boxed{\textbf{Exercise 3}\text{: Prove theorem}}$

Probability

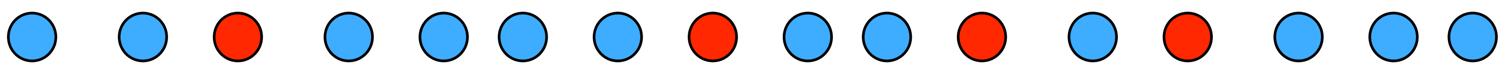
Binomial & Poisson Distributions

Binomial & Poisson Distributions – 1

A **Bernoulli** trial has two outcomes:

S = success or ***F*** = failure.

Example: Each collision between protons at the LHC is a Bernoulli trial in which something interesting happens (***S***) or does not (***F***).



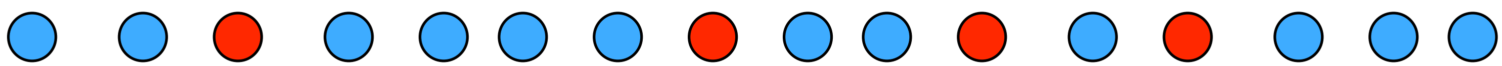
Let $p = P(S)$ be the probability of a success, assumed to be the ***same at each trial***. Since ***S*** and ***F*** are ***exhaustive***, the probability of a failure is $1 - p$. For a given order O of n trails, the probability $\Pr(k,O|n)$ of ***exactly k*** successes and $n - k$ failures is

$$\Pr(k,O,n) = p^k(1-p)^{n-k}$$

Binomial & Poisson Distributions – 2

If the order ***O*** of successes and failures is irrelevant, we can eliminate the order from the problem by ***marginalizing***, that is summing over all possible orders

$$\Pr(k,n) = \sum_{O} \Pr(k,O,n) = \sum_{O} p^k (1-p)^{n-k}$$



This yields the **binomial distribution**

$$\text{Binomial}(k,n,p) \equiv \binom{n}{k} p^k (1-p)^{n-k}$$

Sometimes this is written as $k \sim \text{Binomial}(n,p)$, where "~" means "is distributed as"

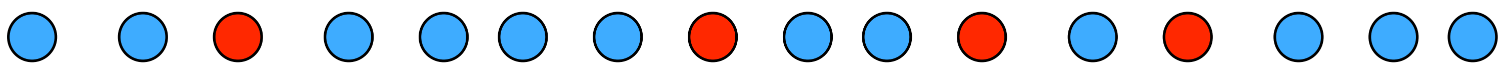
Binomial & Poisson Distributions – 3

We can prove that the mean number of successes $\boldsymbol{a}$ is

$\boldsymbol{a} = \boldsymbol{p}\ \boldsymbol{n}$.

Exercise 4: Prove it

Suppose that the probability, $\boldsymbol{p}$, of a success is very small,



then, in the limit $\boldsymbol{p} \rightarrow 0$ and $\boldsymbol{n} \rightarrow \infty$, such that $\boldsymbol{a}$ is *constant*,

Binomial$(k, n, p) \rightarrow$ **Poisson**(k, a).

The Poisson distribution is generally regarded as a good model for a **counting experiment**

Exercise 5: Show that Binomial$(k, n, p) \rightarrow$ Poisson(k, a)

Common Distributions and Densities

$\mathrm{Uniform}(x,a)$	$1/a$
$\mathrm{Binomial}(k,n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$
$\mathrm{Poisson}(k,a)$	$a^k \exp(-a)/k!$
$\mathrm{Gaussian}(x,\mu,\sigma)$	$\exp(-(x-\mu)^2/2\sigma^2)/\sigma\sqrt{2\pi}$
$\mathrm{Chisq}(x,n)$	$x^{n/2-1}\exp(-x/2)/2^{n/2}\Gamma(n/2)$
$\mathrm{Gamma}(x,a,b)$	$x^{b-1}a^b \exp(-ax)/\Gamma(b)$
$\mathrm{Exp}(x,a)$	$a\exp(-ax)$

Probability – What is it exactly?

There are *at least* two interpretations of probability:

1. **Degree of belief** in, or plausibility of, a proposition

   Example:

   the world will end on December 21, 2012

2. **Relative frequency** of outcomes in an *infinite* sequence of *identically repeated* trials

   Example:

   trials: proton-proton collisions at the LHC

   outcome: a jet in a given rapidity and p_T bin

Likelihood

Likelihood – 1

The ***likelihood function*** is proportional to the probability, or probability density function (**pdf**), of observables evaluated at the observed data.

Example:

$p(D|\, d) = \text{Poisson}(D\,|d)$ ***probability*** of observables D

$p(\mathbf{17}|d) = \text{Poisson}(\mathbf{17}|d)$ ***likelihood*** of observation $D = 17$

$\text{Poisson}(D|d) = \exp(-d)\; d^D \,/\, D!$

Likelihood – 2

Given the likelihood function we can answer questions such as:

1. How do I estimate a parameter?
2. How do I quantify its accuracy?
3. How do I test an hypothesis?
4. How do I quantify the significance of a result?

Writing down the likelihood function requires:

1. Identifying all that is *known*, e.g., the data
2. Identifying all that is *unknown*, e.g., the parameters
3. Constructing a probability model *for both*

Likelihood – 3

Example: Top Quark Discovery (1995), D0 Results

knowns:

$D = 17$ events

$B = 3.8 \pm 0.6$ background events

unknowns:

b	expected background count
s	expected signal count
$d = b + s$	expected event count

Note: we are uncertain about *unknowns*, so **17 ± 4.1** is a statement about d, *not about the observed count* **17**!

Likelihood – 4

The likelihood is a fundamental ingredient in the two most important approaches to inference:

Frequentist

1. Fundamental idea: **frequentist principle**.
2. Use the likelihood function ***only***.

Bayesian

1. Fundamental idea: ***all*** uncertainty can be modeled using probabilities.
2. Use Bayes theorem ***always***.

The Frequentist Approach – 1

The Frequentist Approach

The Frequentist Principle (Neyman, 1937)

Construct statements such that a fraction $f \geq p$ of them will be true over an (infinite) ensemble of statements. The fraction f is called the ***coverage probability*** and p is called the ***confidence level*** (CL).

Note: The confidence level is a property of the ***ensemble*** to which the statements are presumed to belong. In general, the confidence level will change if the ensemble changes.

Neyman's construction of ***confidence intervals*** is the classic example of the frequentist principle in action.

The Frequentist Approach
Maximum Likelihood

Maximum Likelihood – 1

Example: Top Quark Discovery (1995), D0 Results

$D \quad = 17$ events

$B \quad = 3.8 \pm 0.6$ events

Likelihood

$$p(D \mid s, b) = \text{Poisson}(D, s+b)\ \text{Gamma}(k, b, Q+1)$$

$$= \frac{(s+b)^D e^{-(s+b)}}{D!} \frac{(bk)^Q e^{-bk}}{\Gamma(Q+1)}$$

where

$B \quad = Q / k$ $\qquad Q = (B / \delta B)^2 = (3.8 / 0.6)^2 = 41.11$

$\delta B \quad = \sqrt{Q} / k$ $\qquad k = B / \delta B^2 = 3.8 / 0.6^2 = 10.56$

Maximum Likelihood – 2

knowns:

$D = 17$ events

$B = 3.8 \pm 0.6$ background events

unknowns:

b	expected background count
s	expected signal count

Find maximum likelihood estimates (MLE):

$$\frac{\partial \ln p(17 \mid s,b)}{\partial s} = \frac{\partial \ln p(17 \mid s,b)}{\partial b} = 0 \quad \Rightarrow \hat{s},\ \hat{b}$$

$$\hat{s} = D - B, \quad \hat{b} = B$$

Maximum Likelihood – 3

The **Good**

- Maximum likelihood estimates (MLE) are **consistent**: RMS goes to zero as more and more data are acquired
- If an unbiased estimate for a parameter exists, maximum likelihood will find it
- Given the MLE for $\mathbf{s}$, the MLE for $y = g(\mathbf{s})$ is just $\hat{y} = g(\hat{s})$

The **Bad** (from a frequentist point of view!)

- In general, MLEs are biased **Extra Exercise**: Show this

The **Ugly**

- Correcting for bias, however, can waste data and sometimes yield absurdities

The Frequentist Approach

Confidence Intervals

Confidence Intervals – 1

Consider a counting experiment that observes ***D*** events with expected signal s and no background. Its likelihood is

$$p(D \mid s) = \text{Poisson}(D \mid s)$$

Neyman devised a way to make statements of the form

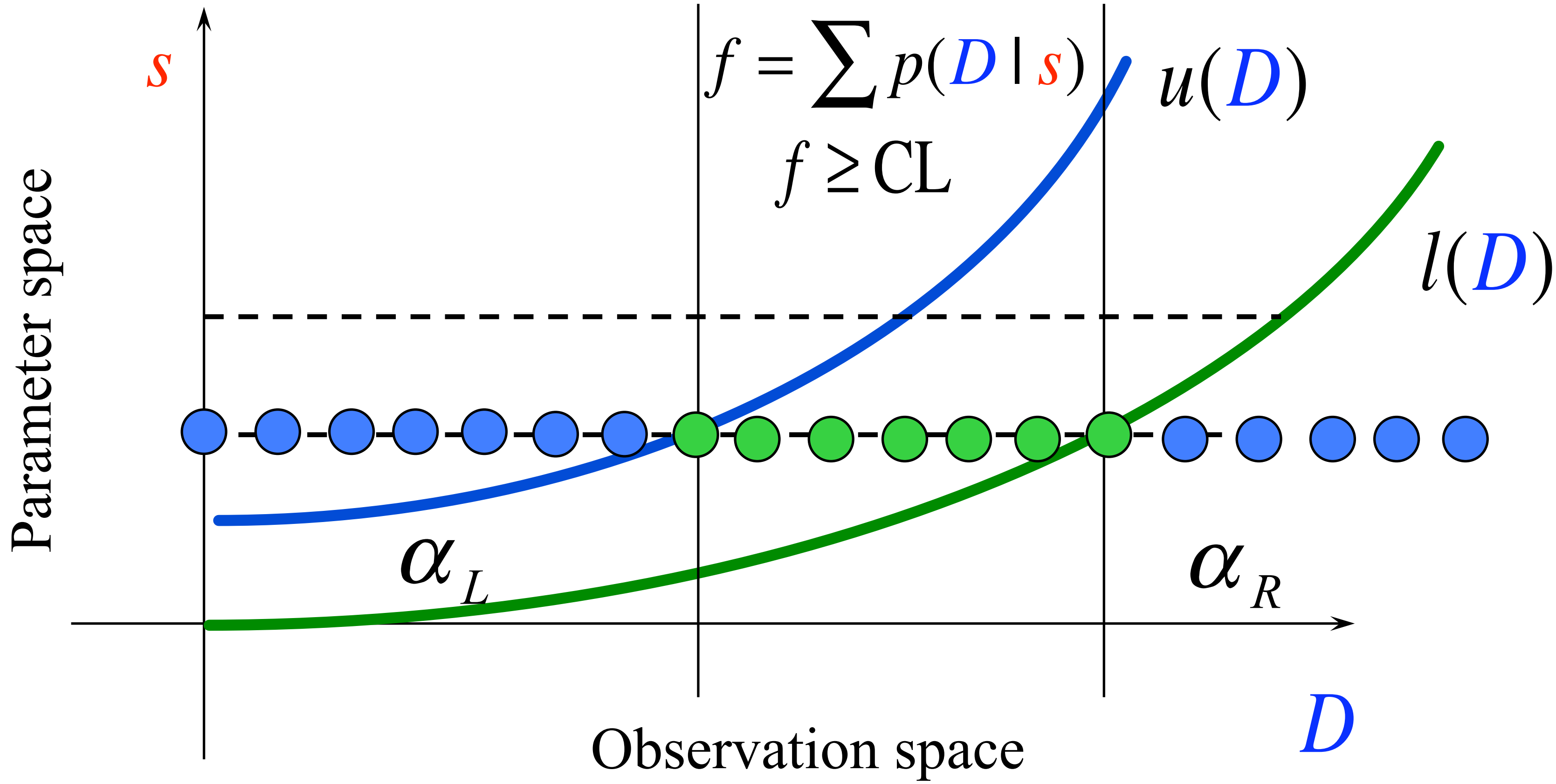
$$s \in [l(D), u(D)]$$

with the guarantee that at least a fraction ***p*** of them are true.

s is presumed to be a *constant*. But, since we don't know ***s***, this criterion needs to hold whatever the value of ***s***.

Confidence Intervals – 2

For each value s find a region in the *observation space* with probability content $f \geq p$ = CL



Confidence Intervals – 3

- **Central Intervals (Neyman)**
  Has equal probabilities on either side

- **Feldman – Cousins Intervals**
  Contains largest values of the ratios $p(D|\, s) \,/\, p(D|D)$

- **Mode – Centered Intervals**
  Contains largest probabilities $p(D|\, s)$

By construction, all these intervals satisfy the frequentist principle: ***coverage probability*** $\geq$ ***confidence level***

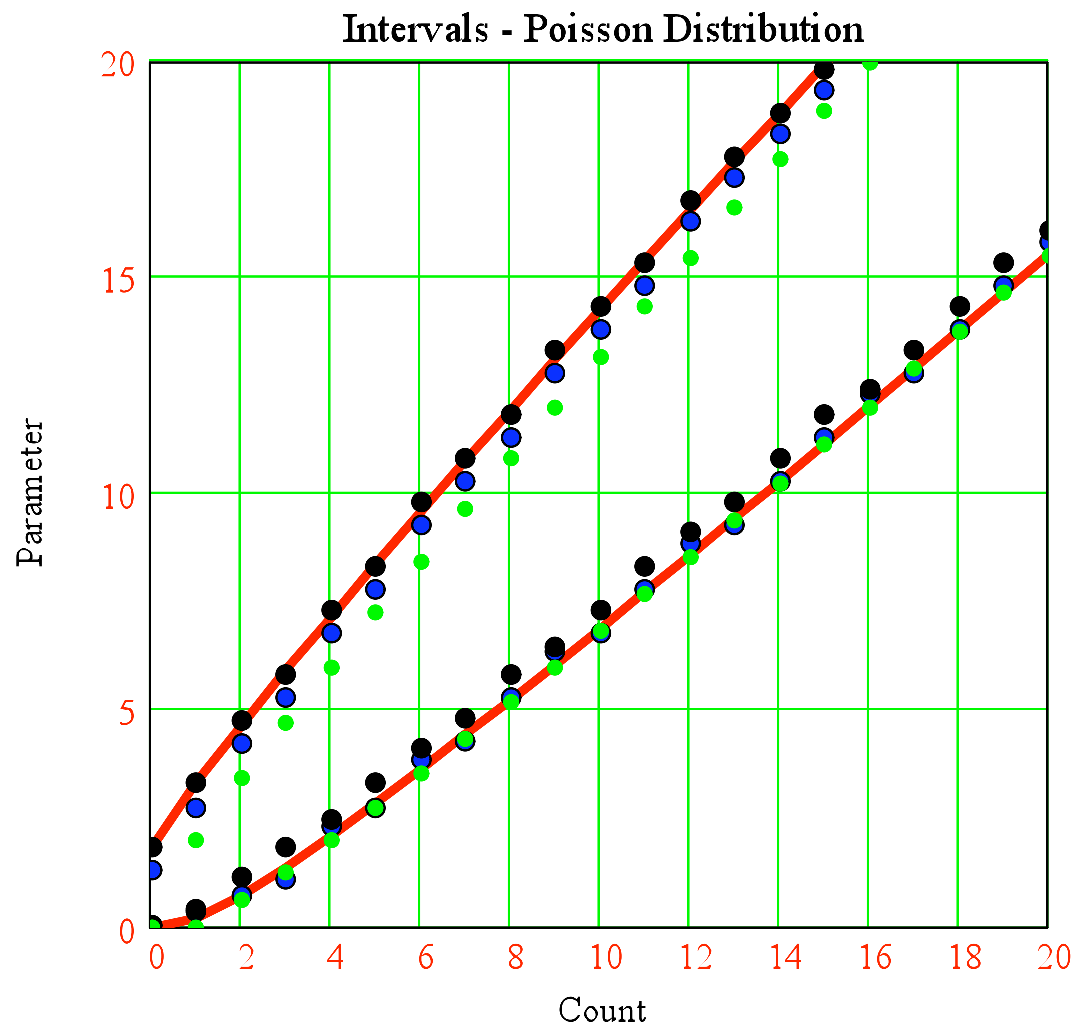
Confidence Intervals – 4

Central

Feldman-Cousins

Mode-Centered

D ± √D



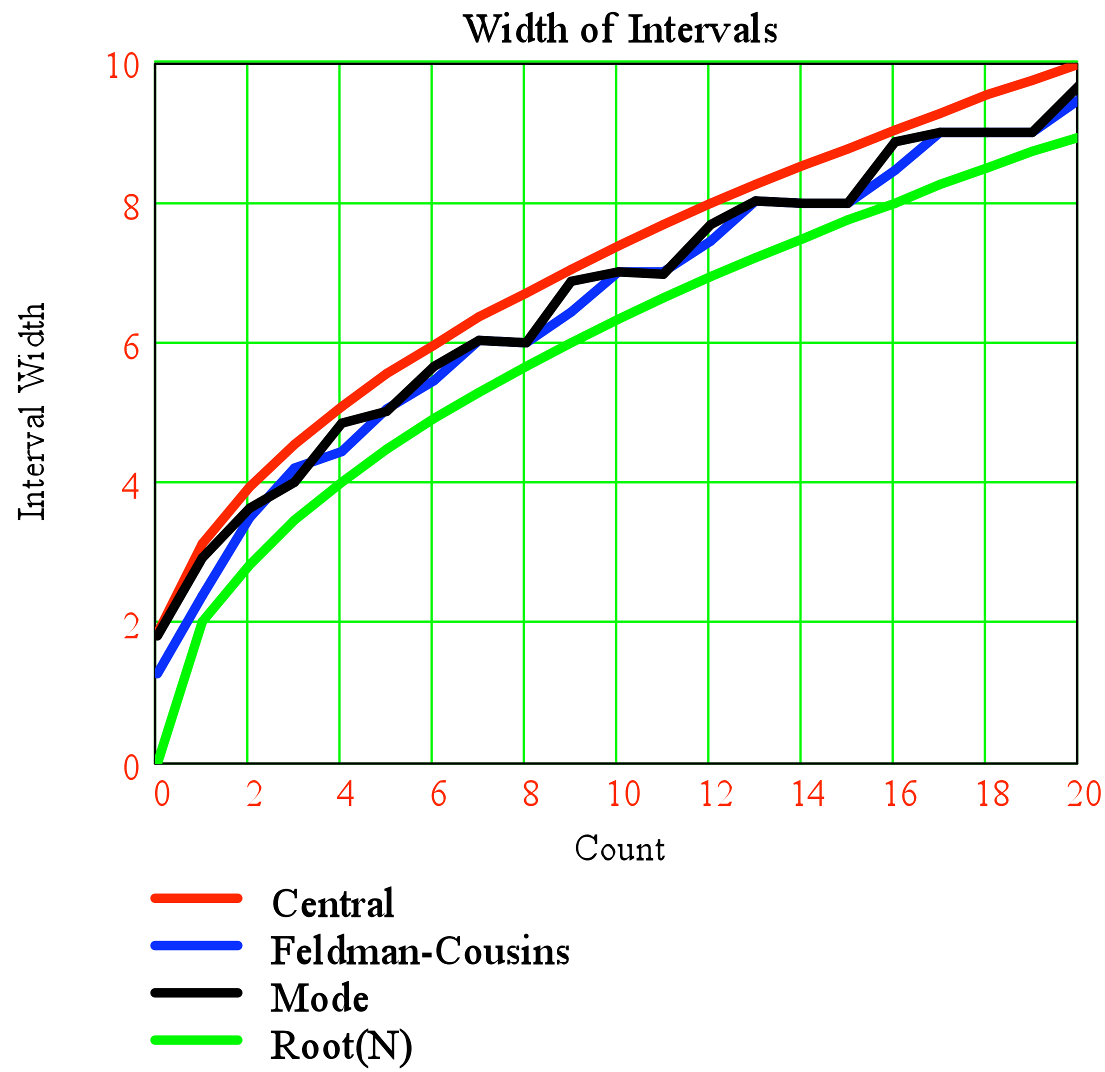
Confidence Intervals – 5

Central

Feldman-Cousins

Mode-Centered

$D \pm \sqrt{D}$



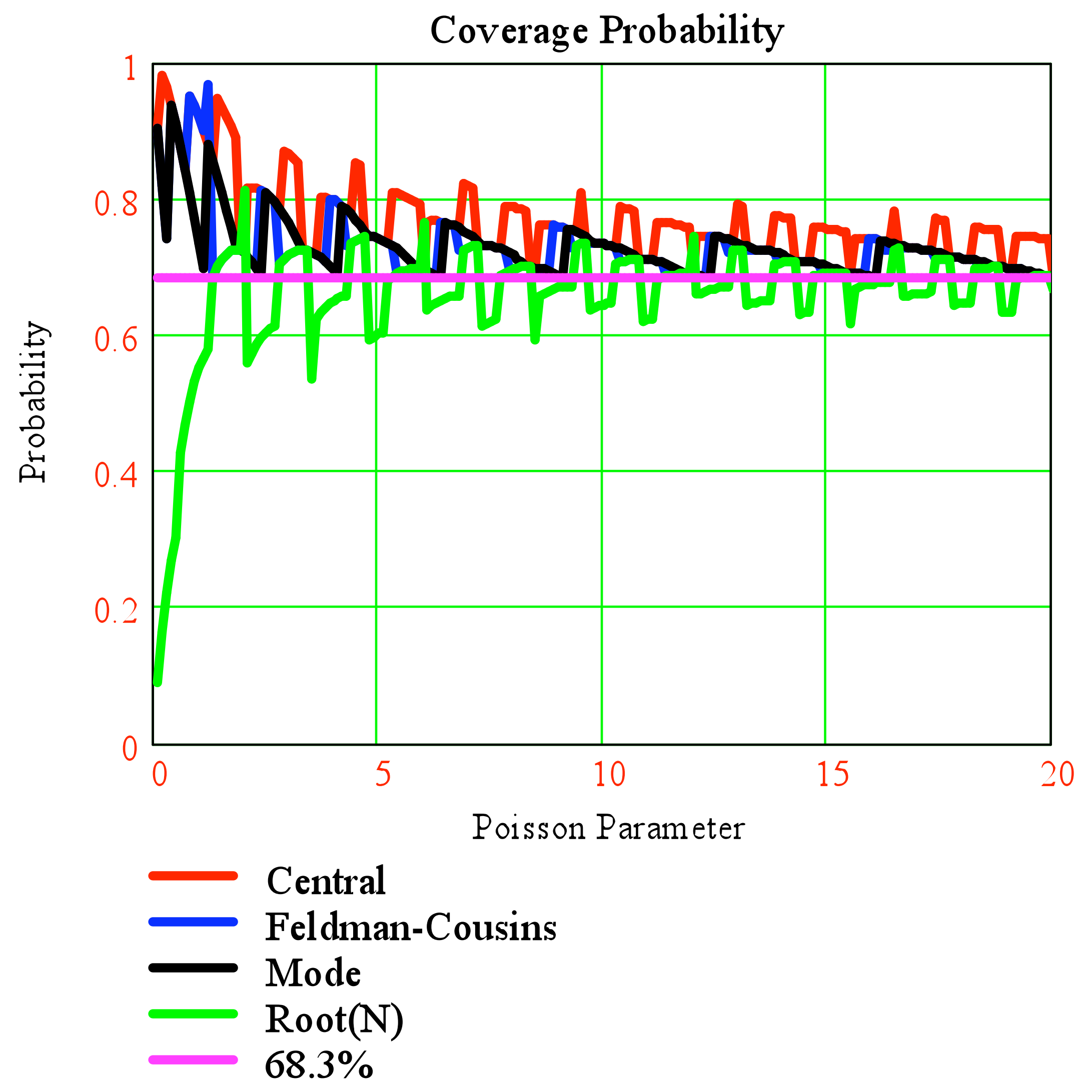
Confidence Intervals – 6

Central

Feldman-Cousins

Mode-Centered

D ± √D



Summary

Probability

Probability is an abstraction that must be interpreted.

Likelihood Function

This is the critical ingredient in any non-trivial statistical analysis.

Frequentist Principle

Construct statements such that a given (minimum) fraction of them are true over a given ensemble of statements.