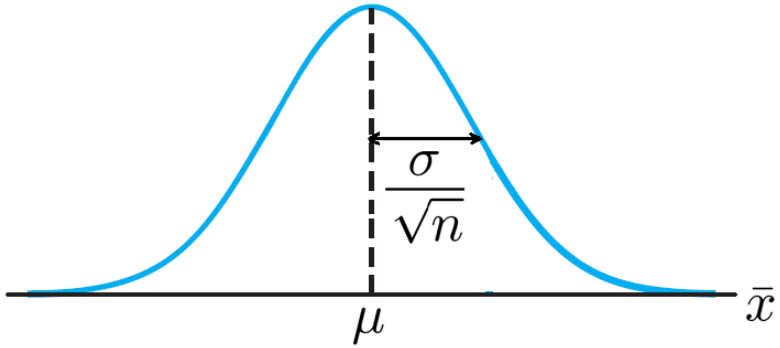
MIDDLE EAST TECHNICAL UNIVERSITY

Particle Data Analysis in High Energy Physics

Lecture 10

Statistics 1



Ahmet Bingül

METU, Physics

Mar 2026

Content

In this chapter, we'll see the following concepts at introductory level:

Sources of Errors

Sampling

Maximum Likelihood Method

Error Propagation

Measurement Errors

Measurement

Measurement is the assignment of numbers to objects or events.

All measurements consist of three parts:

- magnitude,
- dimensions (units) and
- uncertainty

For example, in PDG

π^0 MASS 134.9768 ± 0.0005 MeV

Sources of Errors

Measurements of any kind, in any experiment, are always subject to uncertainties or errors, as they are more often called.

There are two fundamentally different types of experimental error.

- **Statistical (or Random) errors**

- **Systematic errors**

Statistical errors are random in nature.

Repeated measurements will differ from each other and from the true value by amounts which are not individually predictable, although the average behaviour over many repetitions can be predicted.

- *Scale reading errors belong to this class:*
  if we get 50 people to measure our glass block, we expect to get a range of (slightly) different values.
- *Intrinsically random processes like radioactive decay also belong in this category.*

Statistical errors may be reduced by repeating the same experiment many times

Systematic errors arise from problems in the design of the experiment.

They are not random, and affect all measurements in some well-defined way.

Sometimes, it is not easy to predict the systematical errors.

The systematic errors may be reduced by calibration of the device.

Accuracy & Precision

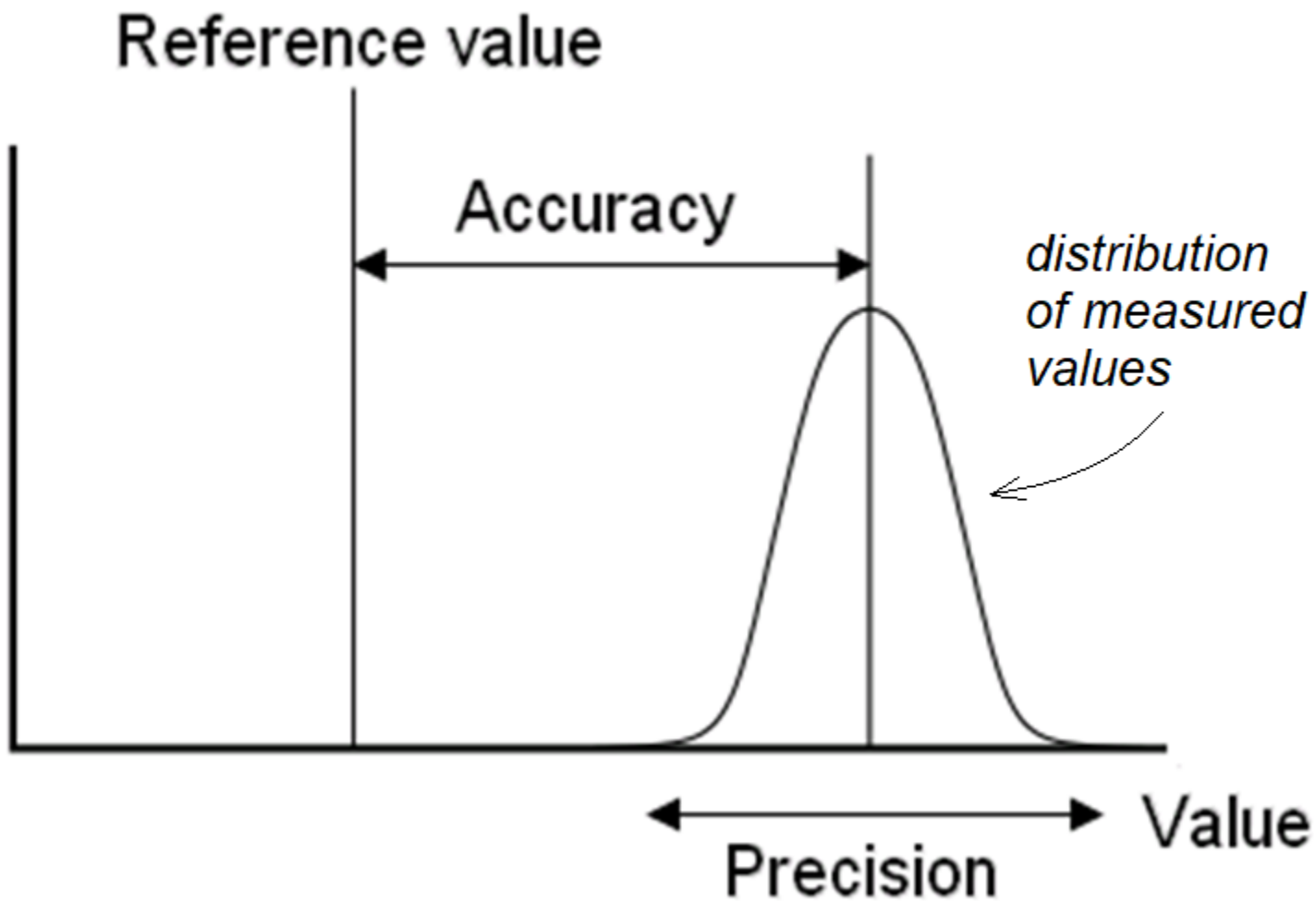
Accuracy

The accuracy indicates proximity of measurement results to the true (actual) value.

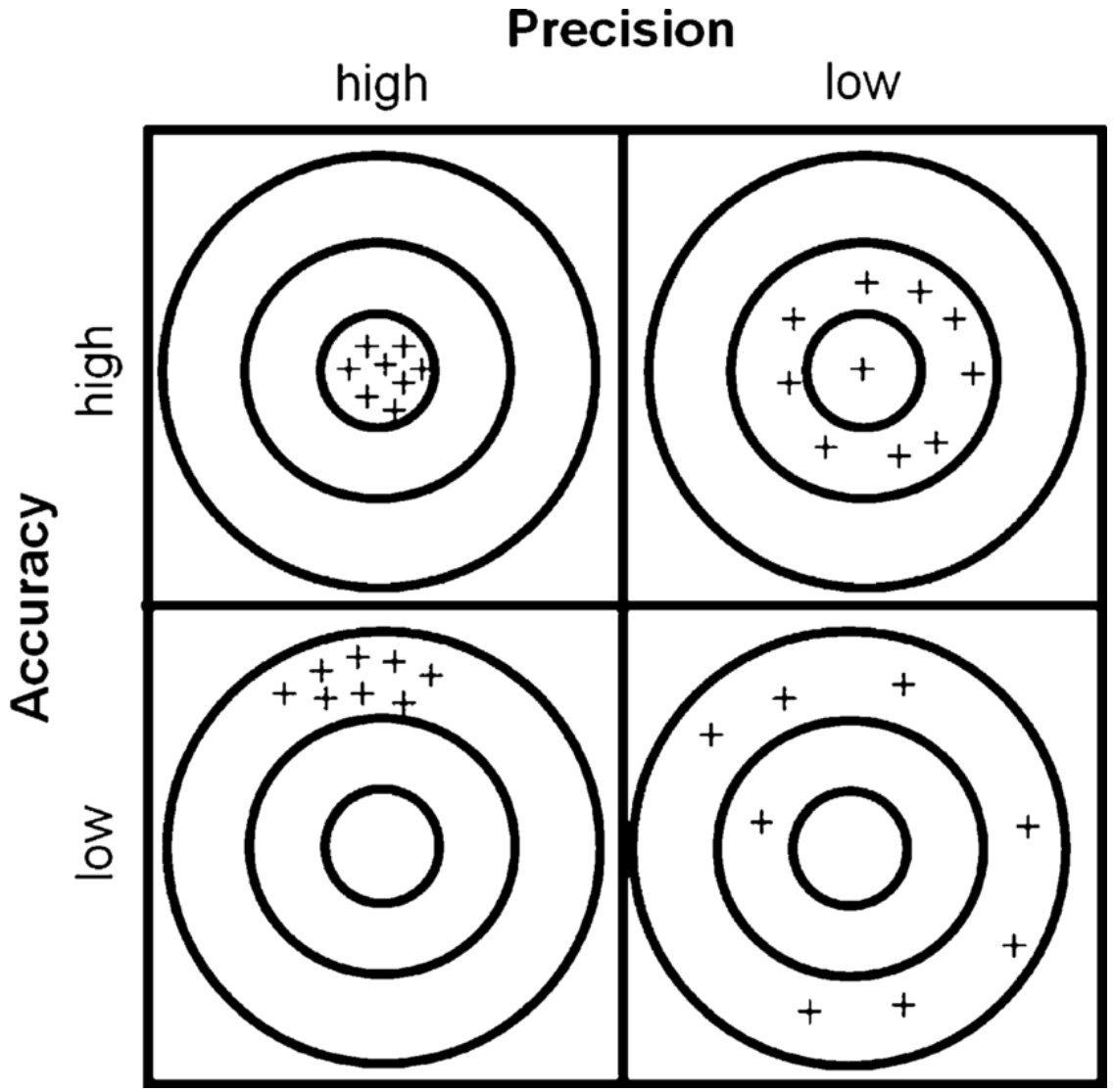
Precision

The precision is the repeatability or reproducibility of the measurement

A measurement system can be accurate but not precise, precise but not accurate. This can be represented by an analogy to the grouping of arrows in a target.


Reference value
Accuracy
distribution
of measured
values
Value
Precision

A target analogy for the comparison of accuracy and precision.

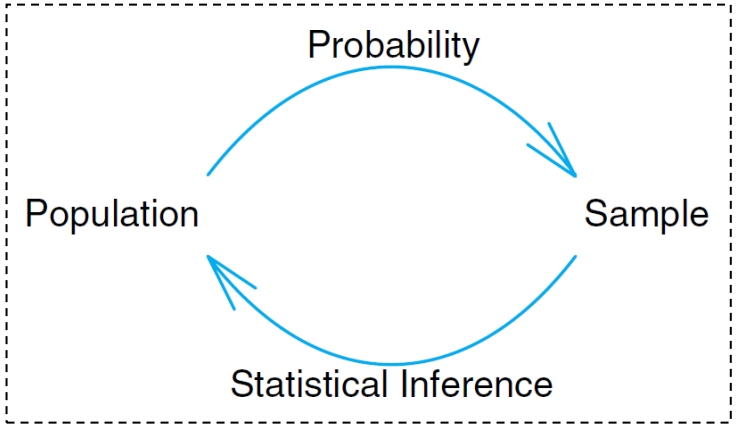


Sampling

Statistics

Statistics is the body of procedures and techniques used to collect, present, and analyze data on which to base decisions in the presence of uncertainty.

Statistics is subdivided into descriptive and inferential:



Descriptive statistics

summarizes a body of data with one or two pieces of information that characterize the whole data.

Inferential statistics

are techniques that allow us to infer generalizations about **populations** by measuring the properties of **samples** of the populations.

Sampling

Sampling is the experimental method by which information can be obtained about the parameters of an unknown distribution or population.

- A **population** consists of the totality of the observations with which we are concerned.
- A **sample** is a subset of a population.

Example

Population: All Higgs bosons that have ever been or will ever be produced in the universe.
The population parameters would be the true physical mass and the decay width of Higss boson.

Sample: The few thousand Higgs boson candidate events observed and reconstructed by the ATLAS and CMS detectors at the LHC.

Sample Moments

Let $x_1, x_2, \ldots, x_n$ be a sample of size n from a distribution whose theoretical (population) mean is μ and variance σ^2. Then we define

Sample mean $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

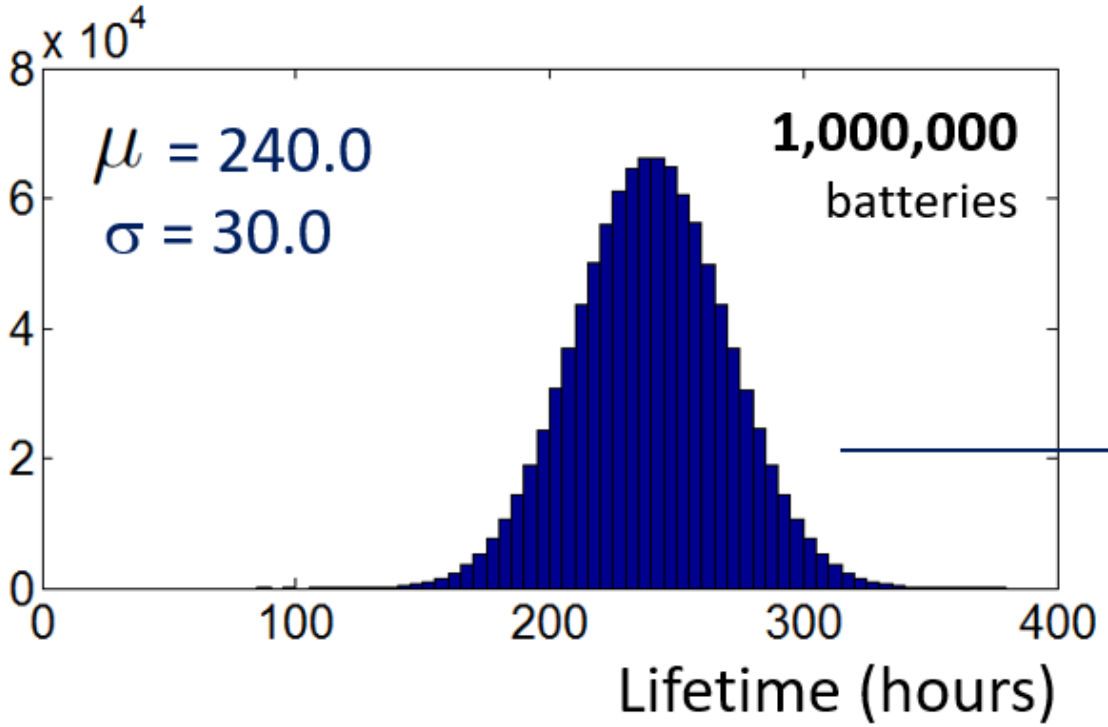
Sample variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

When $n \to \infty$, we obtain population parameters: $\mu = \lim_{n\to\infty} \bar{x}$ and $\sigma^2 = \lim_{n\to\infty} s^2$

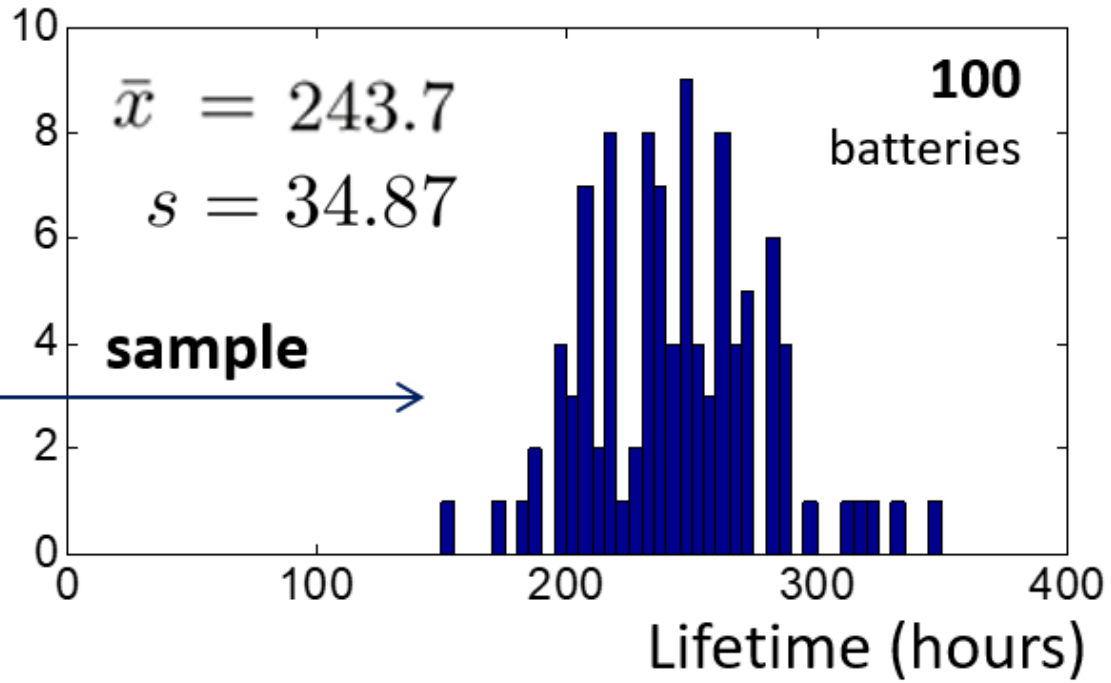
For multivariate samples $(x_1, y_1), (x_2, y_2)$, ..., sample covariance can be calulated by

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})$$

Distribution of battery lifetime for a *normal* population of 1,000,000 batteries

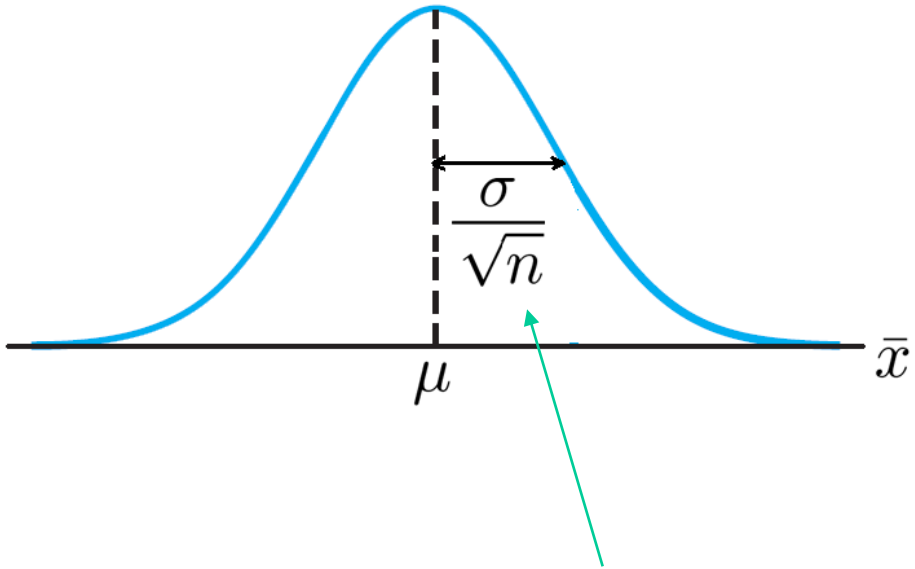


The distribution of battery lifetimes for a sample of size n=100 from the population.



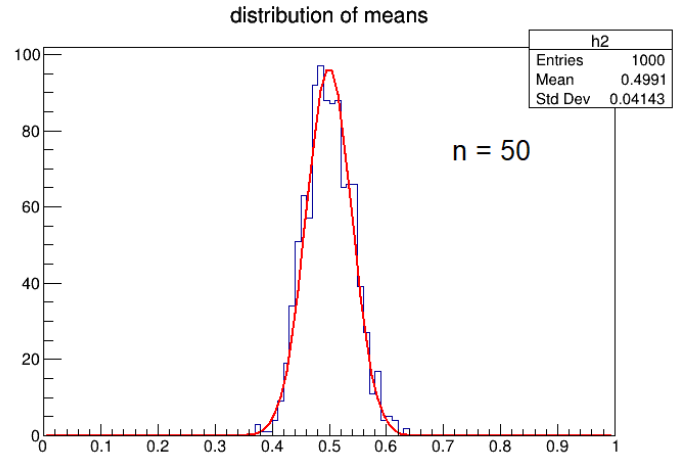
Sampling Distribution of Means

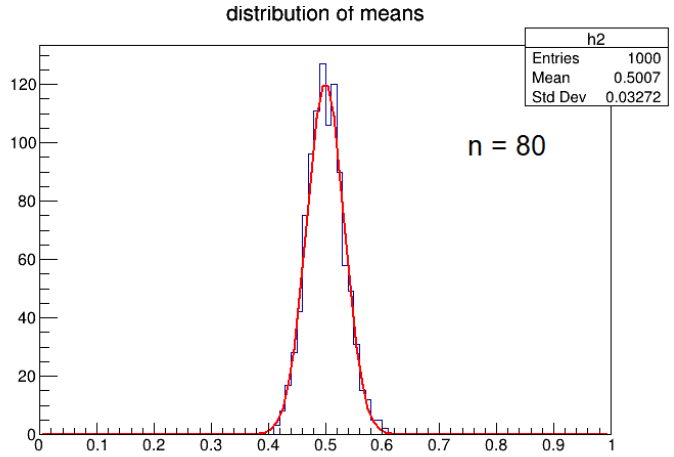
The literature tells us that the distribution of the means of a collection of samples each containing *n* observations is clustered around the population mean with a standard deviation $\sigma/\sqrt{n}$.

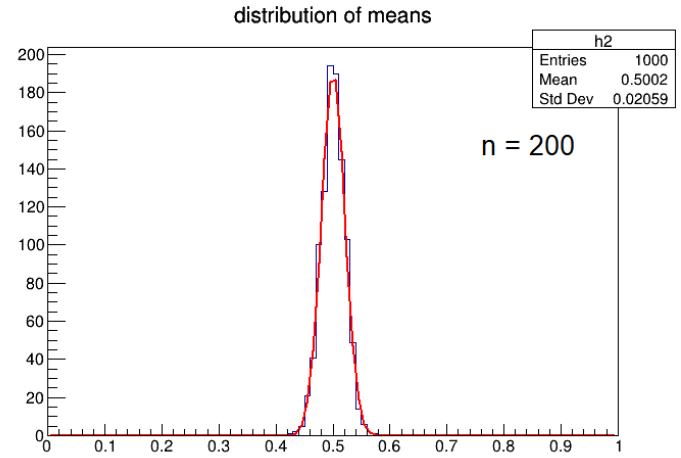


this is called the **standard error** in the sample mean.

```
############################################################
# Proof of centeral limit theorem
# using uniform distribution function
############################################################
import ROOT
import time
from numpy import sqrt

# set seed of random number generator from timer
seed = int(time.time())
rnd = ROOT.TRandom(seed)

h1 = ROOT.TH1F("h1","sample;;",100,0,1)
h2 = ROOT.TH1F("h2","distribution of means;;",100,0,1)

m = 1000 # number of samples
n = 50   # size of one sample
std = 1/sqrt(12)  # theoretical std of uniform distribution
clt = std/sqrt(n) # std predicted by central limit theorem

for j in range(m):
   for i in range(n):
      h1.Fill(rnd.Uniform(0,1))
   mean = h1.GetMean()
   print("mean of sample ",j, " is ",mean)
   h2.Fill(mean)
   h1.Reset()
# fit and draw distribution of means
h2.Fit("gaus")
h2.Draw()

print("Mean of means = %.5f" % h2.GetMean())
print("Std  of means = %.5f" % h2.GetStdDev())
print("Central Limit = %.5f" % clt)
input("press Enter so quit")
```

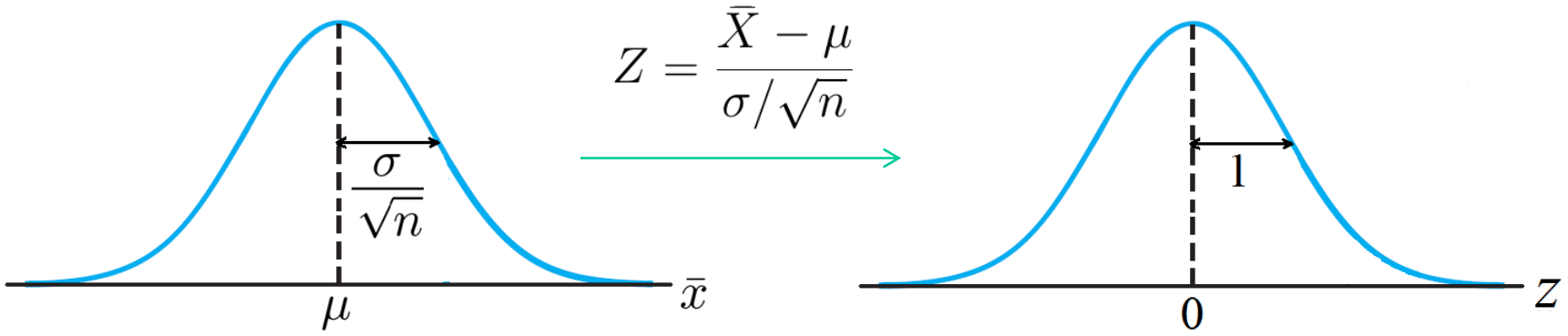






The Central Limit Theorem

For **large n**, the difference between a sample mean and the true population mean, divided by the standard error, follows a **standard normal distribution**.

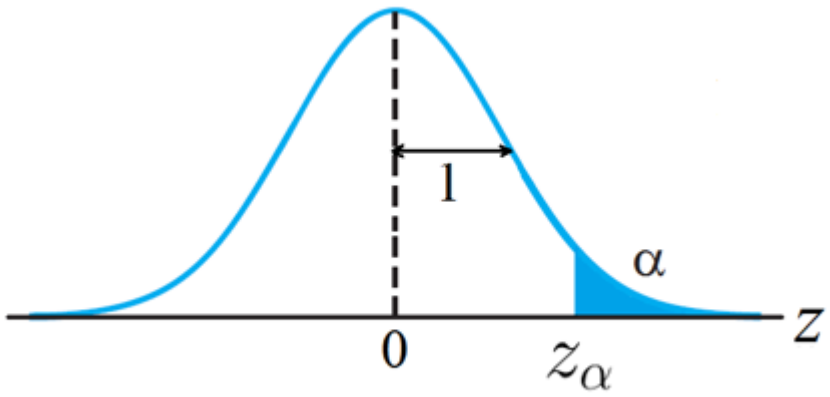


The distribution of $\bar{x}$ with large n.

The Z transformed distribution.

Sampling Probability

α is the probability = the area of standard normal curve between $[z_\alpha, \infty]$.

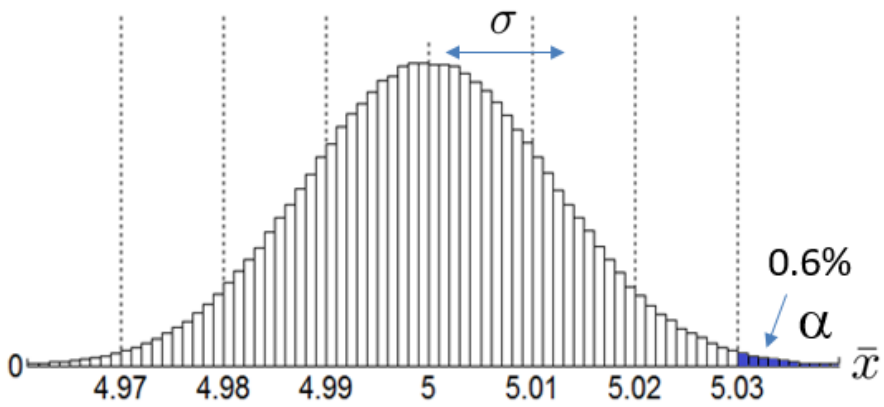


$$P(z > z_\alpha) = \alpha$$
$$= 1 - \text{normal\_cdf}(z_\alpha)$$

Example

A factory manufactures cylindrical components with a diameter that is approximately normally distributed with a mean of 5 millimeters and a standard deviation of 0.12 millimeters.

(a) Sketch the sampling distribution for the mean for a sample size of 100.



(b) What is the probability of obtaining a sample mean greater than 5.03 mm?

$$z_\alpha = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{5.03 - 5}{0.12/\sqrt{100}} = \frac{0.03}{0.012} = 2.5$$

$$\alpha = P(Z > z_\alpha) = P(Z > 2.5)$$

$$= 1 - \text{normcdf}(2.5) = 1 - 0.9938 = 0.0062$$

With a sample size of 100, only **0.6%** of samples would yield a sample mean greater than 5.03 mm.

Parameter Estimation
(using Maximum Likelihood Method)

The Maximum Likelihood Method

The method of maximum likelihood is only applicable if the form of the theoretical distribution from which the sample is taken is known. Suppose we have a sample of n independent observations $x_1, x_2, \ldots, x_n$, from a theoretical distribution $f(x|\theta)$ where θ is the parameter to be estimated. The method then consists of calculating the **likelihood function**,

$$L(\theta|x) = f(x_1|\theta) f(x_2|\theta) \ldots f(x_n|\theta)$$

which can be recognized as the probability for observing the sequence of values $x_1, x_2, \ldots, x_n$. The principle now states that this probability is a maximum for the observed values. Thus, the parameter θ must be such that L is a maximum. So, we need to solve

$$\frac{dL}{d\theta} = 0$$

Depending on the form of L, it may also be easier to maximize the logarithm of L rather than L itself.

$$\frac{d(\ln L)}{d\theta} = 0$$

The solution, $\hat{\theta}$, is known as the **maximum likelihood estimator** for the parameter θ.

It should be realized now that $\hat{\theta}$ is a random variable. What is the error on $\hat{\theta}$? i.e What's $\sigma(\hat{\theta})$? This is given by the standard deviation of the estimator distribution.

General formula: $$\sigma^2(\hat{\theta}) = \int (\hat{\theta} - \theta)^2 L(\theta|x)\, dx_1\, dx_2 \ldots dx_n$$

Approximate formula: $$\sigma^2(\hat{\theta}) \simeq -\left(\frac{d^2 \ln L}{d\theta^2}\right)^{-1}$$

If there is more than one parameter, the matrix of the second derivatives must be formed, i.e.,

$$U_{ij} = -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}$$

The diagonal elements of the inverse matrix then give the approximate variances,

$$\sigma^2(\hat{\theta}_i) \simeq (U^{-1})_{ii}$$

Estimator for Poisson Distribution

Suppose we have n measurements of samples, $x_1, x_2, x_3, \ldots, x_n$ from a Poisson distribution with mean μ. The likelihood function for this case is then

$$L(\mu|x) = \prod_{i=1}^{n} \frac{\mu^{x_i}}{x_i!} \exp(-\mu) = \exp(-n\mu) \prod_{i=1}^{n} \frac{\mu^{x_i}}{x_i!} .$$

To eliminate the product sign, we take the logarithm

$$L^* = \ln L = -n\mu + \sum x_i \ln \mu - \sum \ln x_i! .$$

Differentiating and setting the result to zero, we then find

$$\frac{dL^*}{d\mu} = -n + \frac{1}{\mu} \sum x_i = 0 ,$$

which yields the solution

$$\hat{\mu} = \frac{1}{n} \sum x_i = \bar{x} .$$

This is of no great surprise, it is a validation.

Estimator for Poisson Distribution

The variance of $\bar{x}$ is

$$\sigma^2(\bar{x}) = E[(\bar{x}-\mu)^2] = \frac{\sigma^2}{n}$$

For a Poisson distribution, $\sigma^2 = \mu$, so that the error on the estimated Poisson mean is

$$\sigma(\hat{\mu}) = \sqrt{\frac{\mu}{n}} \simeq \sqrt{\frac{\hat{\mu}}{n}} = \sqrt{\frac{\bar{x}}{n}}$$

where have substituted the estimated value $\hat{\mu}$ for the theoretical μ.

Estimator for Gaussian Distribution

For a sample of n points, all taken from the same Gaussian distribution, the likelihood function is

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right].$$

Once again, taking the logarithm,

$$L^* = \ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum\frac{(x_i-\mu)^2}{\sigma^2}.$$

Taking the derivatives with respect to μ and σ^2 and setting them to 0, we then have

$$\frac{\partial L^*}{\partial \mu} = \sum\frac{x_i-\mu}{\sigma^2} = 0$$

$$\frac{\partial L^*}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum\left(\frac{x_i-\mu}{\sigma}\right)^2\frac{1}{\sigma^2} = 0.$$

Solving

$$\hat{\mu} = \frac{\sum x_i}{n} = \bar{x} \qquad \sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

The best estimate of the theoretical mean for a Gaussian is thus the sample mean

This is the standard error of the mean.

For the moment, however, σ is still unknown. Solving (4.48) for σ^2 yields the estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \mu)^2 \simeq \frac{1}{n} \sum (x_i - \bar{x})^2 = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

This, of course, is just the sample variance.

the standard deviation of $\hat{\sigma}$

$$\sigma(\hat{\sigma}) = \frac{\sigma}{\sqrt{2(n-1)}} \simeq \frac{\hat{\sigma}}{\sqrt{2(n-1)}}$$

The Weighted Mean

We have a sample $x_1, x_2, \ldots, x_n$ where each value is from a Gaussian distribution having the same mean μ but a different standard deviation σ_i. The likelihood function is:

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right].$$

$$L^* = \ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\sum \frac{(x_i-\mu)^2}{\sigma_i^2} .$$

Maximizing this we then find the weighted mean and associated error:

$$\hat{\mu} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} \qquad \sigma^2(\hat{\mu}) = \frac{1}{\sum 1/\sigma_i^2}$$

Example

Consider the simple experiment to measure the length of an object.

The following results are from such a measurement is given below.

What isthe best estimate for the length of this object?

17.62	17.62	17.615	17.62	17.61
17.61	17.62	17.625	17.62	17.62
17.61	17.615	17.61	17.605	17.61

Solution

Since the errors in the measurement are instrumental, the measurements are Gaussian distributed.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 17.61533$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 5.855 \times 10^{-3}$$

$$\sigma(\bar{x}) = s/\sqrt{15} = 0.0015$$

The best value for the length of the object is:

$$x = 17.616 \pm 0.002$$

Example

Given three independent the measurements for the gravitational acceleration data:

$$9.77 \pm 0.14\ \mathrm{m/s^2}, \quad 9.82 \pm 0.10\ \mathrm{m/s^2}, \quad 9.86 \pm 0.20\ \mathrm{m/s^2}$$

Find the combined result of *g*.

Solution

We use weigthed mean formula:

$$g = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} = 9.811$$

$$\sigma(g) = \sqrt{\frac{1}{\sum 1/\sigma_i^2}} = 0.075$$

Combined result is:

$$g = 9.811(75)\ \mathrm{m/s^2}$$

Example

Consider the following series of measurements of the counts per minute from a detector viewing a ^{22}Na source,

2201 2145 2222 2160 2300

What is the decay rate and its uncertainty?

Solution

Radioactive decay is described by a Poisson distribution, we use the estimators for this distribution

$$\hat{\mu} = \bar{x} = 2205.6 \quad \text{and}$$

$$\sigma(\hat{\mu}) = \sqrt{\frac{\bar{x}}{n}} = \sqrt{\frac{2205.6}{5}} = 21$$

$$\text{Count Rate} = 2206 \pm 21 \text{ counts/min}$$

What if we count 5 minutes without stopping? We observe a total of 11208 counts.

The error is $\sigma = \sqrt{11208} = 106$. Hence, $\text{Count Rate} = 11208 \pm 106$ or dividing by five

$$\text{Count Rate} = 2206 \pm 21 \text{ counts/min}$$

Error Propagation

Error Propagation

How does the sum, multiplication, etc., of two uncertain measurement results affect the outcome?
In statistical data analysis, error propagation is related to the effect of measurement uncertainties on the outcome function.

Consider a quantity $u = f(x, y)$ where x and y are quantities having errors σ_x and σ_y, respectively.
The variance σ_u^2 can be defined as: $\sigma_u^2 = E[(u - \bar{u})^2]$
Let us expand $(u - \bar{u})$ to first order:

$$(u - \bar{u}) \simeq (x - \bar{x}) \left.\frac{\partial f}{\partial x}\right|_{\bar{x}} + (y - \bar{y}) \left.\frac{\partial f}{\partial y}\right|_{\bar{y}}$$

$$E[(u - \bar{u})^2] \simeq E\left[(x - \bar{x})^2 \left(\frac{\partial f}{\partial x}\right)^2 + (y - \bar{y})^2 \left(\frac{\partial f}{\partial y}\right)^2 + 2(x - \bar{x})(y - \bar{y}) \frac{\partial f}{\partial x} \frac{\partial f}{\partial y}\right]$$

$$\sigma_u^2 \simeq \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2 \operatorname{cov}(x, y) \frac{\partial f}{\partial x} \frac{\partial f}{\partial y}$$

Error of a Sum: $u = x + y$

$$\sigma_u^2 = \sigma_x^2 + \sigma_y^2 + 2\,\mathrm{cov}\,(x, y)$$

Error of a Difference: $u = x - y$

$$\sigma_u^2 = \sigma_x^2 + \sigma_y^2 - 2\,\mathrm{cov}\,(x, y)\,.$$

Error of a Product: $u = xy$

$$\frac{\sigma_u^2}{u^2} \approx \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} + 2\,\frac{\mathrm{cov}\,(x, y)}{xy}$$

Error of a Ratio: $u = x/y$

$$\frac{\sigma_u^2}{u^2} \approx \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} - 2\,\frac{\mathrm{cov}\,(x, y)}{xy}$$

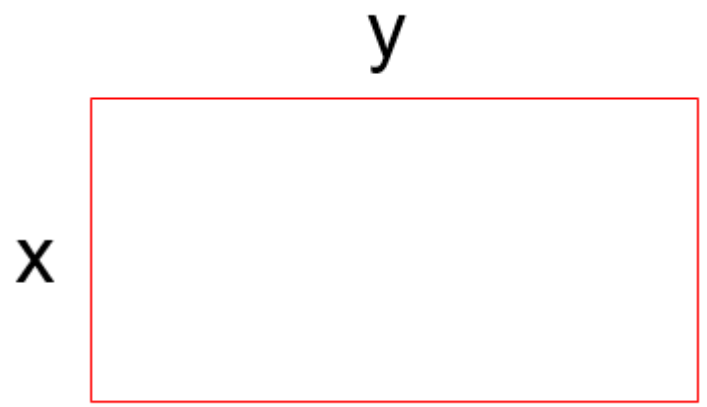
Example

What is the area of the rectangle

if x and y are uncorrelated and measured values are:

$$x = 1.0 \pm 0.1\ \mathrm{m}$$

$$y = 2.0 \pm 0.2\ \mathrm{m}$$



Answer:

$$A = 2.00 \pm 0.28\ \mathrm{m}^2$$

For multi variable function, $u = f(x_1, x_2, \ldots, x_n)$, if variables are indelendent, error propagation formula is given by:

$$\sigma_u^2 = \left(\frac{\partial f}{\partial x_1}\right)^2 \sigma_1^2 + \left(\frac{\partial f}{\partial x_2}\right)^2 \sigma_2^2 + \cdots + \left(\frac{\partial f}{\partial x_n}\right)^2 \sigma_n^2$$

Exercises

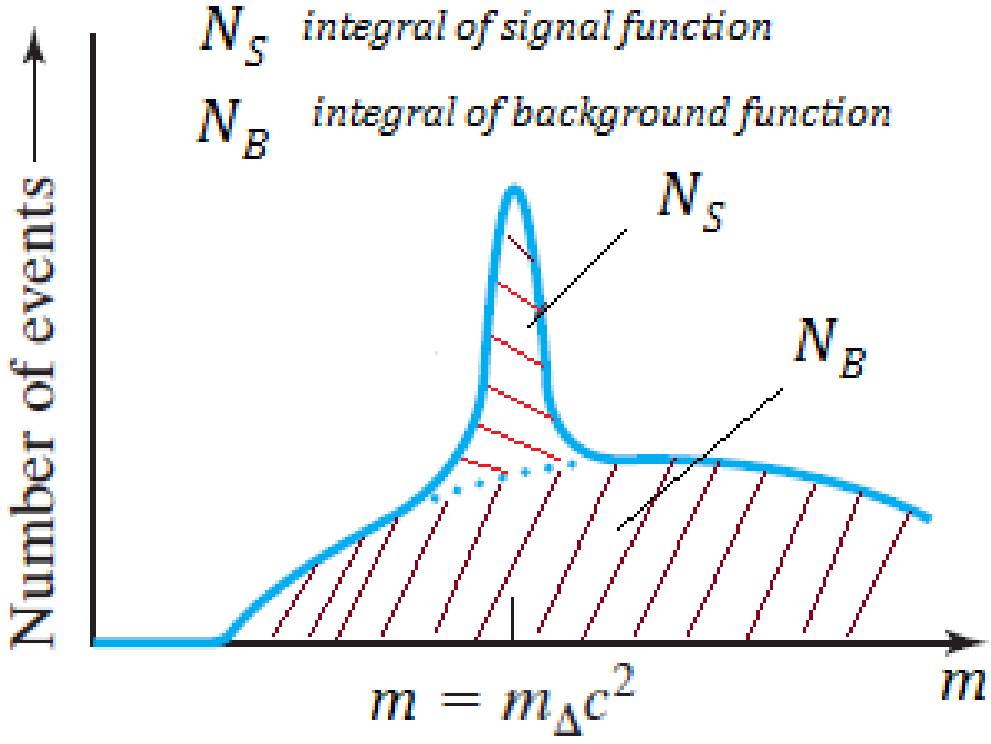
1. Repeat first example for n = 20 and n = 200 samples.

2. Cross section formula is given by:

$$\sigma = \frac{N_S}{\mathcal{L} \times \varepsilon \times \mathcal{B}}$$

$N_S = 2000 \pm \sqrt{2000}$ (number of signals)
$\mathcal{L} = 1.3 \pm 0.2\ \text{fb}^{-1}$ (integrated Luminosity)
$\varepsilon = 0.7 \pm 0.1$ (efficiency)
$\mathcal{B} = 0.45 \pm 0.03$ (branching ratio)

Find σ and its uncertainty.
Hint all values are independent.



3. A resistance is connected in a circuit as shown in Figure.

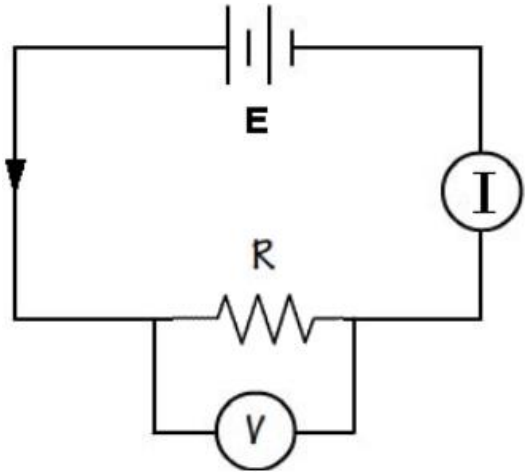
The potential across the resistor and the current passing thorough the resistor are measured to be:

$V = 101.3\,V \pm 0.5\,\mathrm{V}$

$I = 10.7\,A \pm 0.1\,\mathrm{A}$

$\rho = 0.1$ (Correlation coefficient)

Determine the resistance and its uncertainty.



4. It is necessary to use the lifetime of the muon in a calculation. However, in searching through the literature, 7 values are found from different experiments:

2.198 ± 0.001 μs	2.203 ± 0.004 μs	2.202 ± 0.003 μs
2.197 ± 0.005 μs	2.198 ± 0.002 μs	2.1966 ± 0.0020 μs
2.1948 ± 0.0010 μs		

Determine the combined value of lifetime and its error.

5. A population has a density function given by $f(x) = Ax^2\exp(-kx^2)$ where $-\infty < x < \infty$.

a) Determine the normalization constant A.

b) For n obervations, $x_1, x_2, \ldots, x_n$, made from this this population, find the maximum likelihood estimate of k.

6. A population has a density function given by $f(x) = Ax^k$ where $0 < x < 1$.

a) Determine the normalization constant A.

b) For n obervations, $x_1, x_2, \ldots, x_n$, made from this this population, find the maximum likelihood estimate of k.

c) For $n = 3$ evalulate $\hat{k}$ if $x_1 = 0.1, x_2 = 0.2, \text{and } x_3 = 0.7$