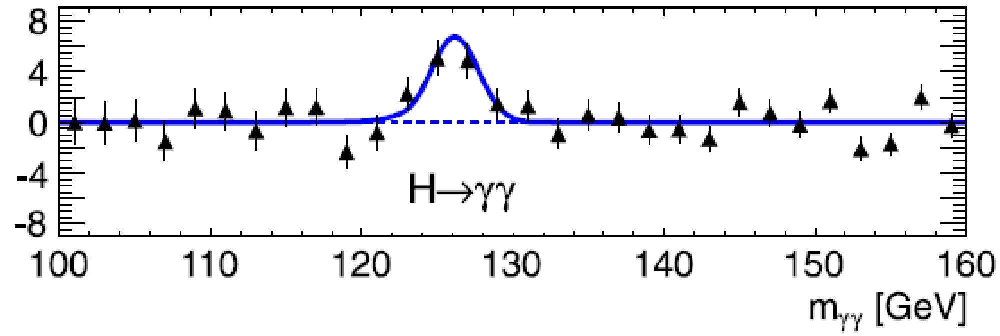
EP122
Measurement Techniques and Calibration

Topic 4

Introductory Statistics



http://www.gantep.edu.tr/~bingul/ep122

Department of Engineering Physics

University of Gaziantep

Feb 2016

Content

Content

PART I

BASIC OCTAVE COMMANDS

Scalars

Arithmetic works as expected.

Note that the result is given the name "ans" each time

```
>> 2 + 3
ans = 5
```

```
>> 1234/5786
ans = 0.2133
```

```
>> 2^5
ans = 32
```

You can choose your own names

```
>> a = sqrt(2)
a = 1.4142
```

OneDim Arrays (Vectors)

```
>> x = [0  0.25  0.5  0.75  1]

x = 0    0.2500    0.5000    0.7500    1.0000
```

```
>> x = 0:0.25:1

x = 0    0.2500    0.5000    0.7500    1.0000
```

```
>> dizi = 1:7

dizi = 1      2      3      4      5      6      7
```

```
>> dizi = -5:2:5

dizi = -5     -3      -1      1      3      5
```

```
>> v = [1 2 3]    % row vector
v = 1   2   3
```

```
>> v = [1 2 3]'   % transpose of a row vector
v =
     1
     2
     3
```

```
>> x = [1 2 3];
>> y = [5 6 7];
>> x .* y
ans = 5 12 21
```

TwoDim Arrays (Matrices)

```
>> A = [1 1 1; 2 2 2]

A =   1      1      1

      2      2      2
```

```
>> A = [1 1 1

        2 2 2]

A =   1      1      1

      2      2      2
```

```
>> B = A'

B =   1      2

      1      2

      1      2
```

Visualizing and Analysing Data

To visualize data

- **plot(x,y)** X-Y graph
- **hist(x)** Histogram
- **pie(x)** Pie chart
- **. . .**

To analyze data

- **mean(x)** Average value
- **std(x)** Standard deviation
- **max(x)** Maximum value
- **min(x)** Minimum value
- **. . .**

Example 1

Exam Scores of 20 students:

55 42 65 68 64 72 75 58 87 89

77 66 91 39 44 57 69 75 68 81

```
octave:> x=[55 42 65 68 64 72 75 58 87 89 ...
            77 66 91 39 44 57 69 75 68 81];
```

Example 1: Basic Analysis

```
octave:> x=[55 42 65 68 64 72 75 58 87 89 ...
            77 66 91 39 44 57 69 75 68 81];
octave:> mean(x)
ans =  67.100
octave:> std(x)
ans =  14.853
octave:> length(x)
ans =  20
octave:> std(x) / sqrt(length(x))
ans =  3.3213
octave:> max(x)
ans =  91
octave:> min(x)
ans =  39
```

Reading Text Files

Consider **gravity.txt** (*) contains 1000 measurements of gravitational acceleartion (g) on the Earth surface at sea level.
Find the <u>mean</u>, <u>maximum</u> and <u>minimum</u> values of the data.

We can read this data directly into a vector and process it as follows:

```
octave:> g = textread('gravity.txt');
octave:> mean(g)
ans = 9.8120
octave:> max(g)
ans = 10.4856
octave:> min(g)
ans = 9.0473
```

(*) Download the file at:
`http://www1.gantep.edu.tr/~bingul/ep122/data/gravity.txt`

PART II

BASIC STATISTICS

Statistics

Statistics is a way of extracting information from a data.

Wikipedia says:

Statistics is the study of the collection, organization, analysis, interpretation, and presentation of data.

http://en.wikipedia.org/wiki/Statistics

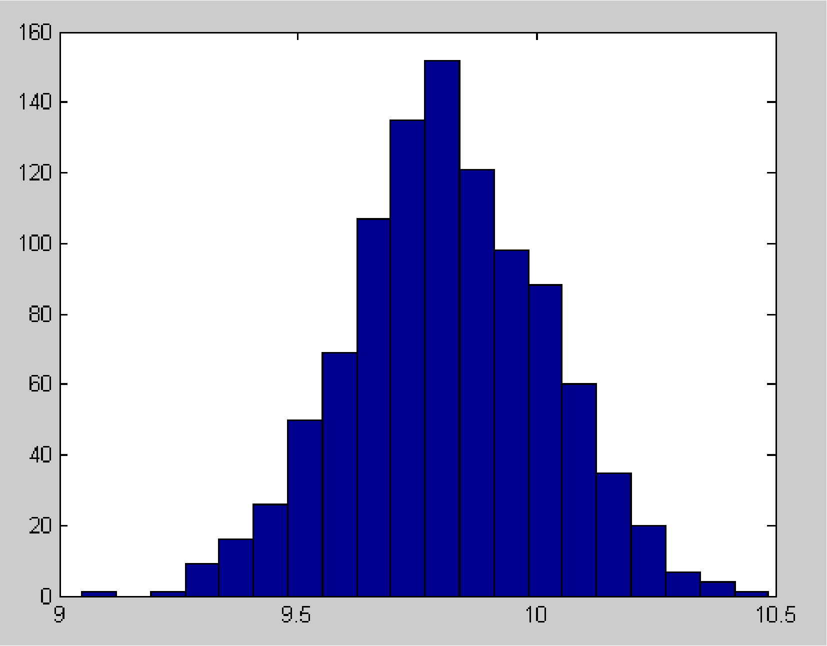
Data Analysis

Data analysis is a very broad subject covering many techniques and types of data. In this lecture we will study some basic calculations that are commonly performed on sampled data.

Consider again the file **gravity.txt**. We can plot the histogram of the data:

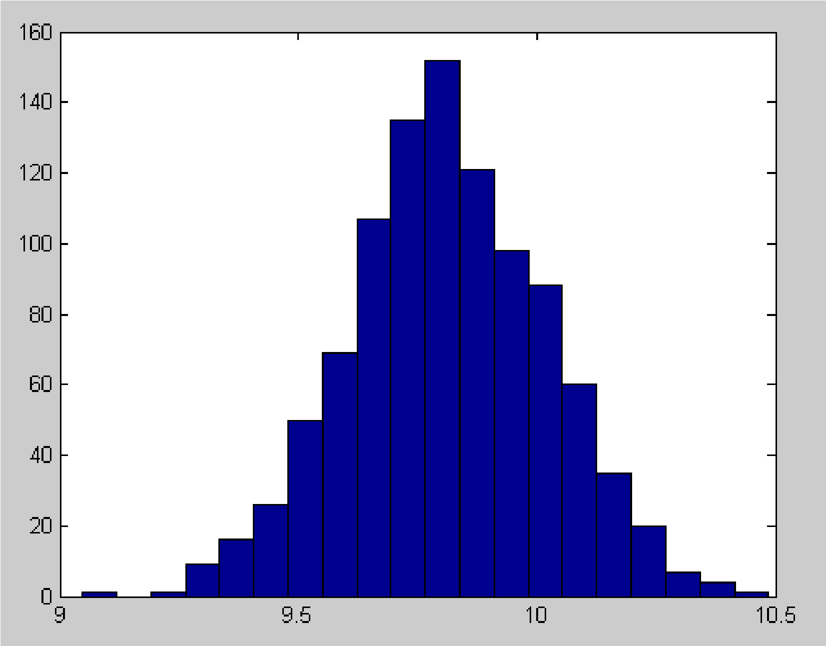
```
>> g = textread('gravity.txt');

>> hist(g,20)
```



The ***mean*** $\bar{x}$ of the sample is given by:

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



The ***standard deviation*** σ of the sample is defined as:

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

```
>> g = textread('gravity.txt');
>> mean(g)
ans = 9.8120
>> std(g)
ans = 0.2077
```

For this data, σ is the size of the variation of *g* about the mean.

For bi-variate data (two variables) the correlation coefficient (ρ) is a measure of the linear dependence between one variable and the other.

Given a sample (size *n*) of bi-variate data,

$Z = \{ (x_1,y_1), (x_2,y_2), (x_3,y_3), \ldots , (x_n,y_n) \}$

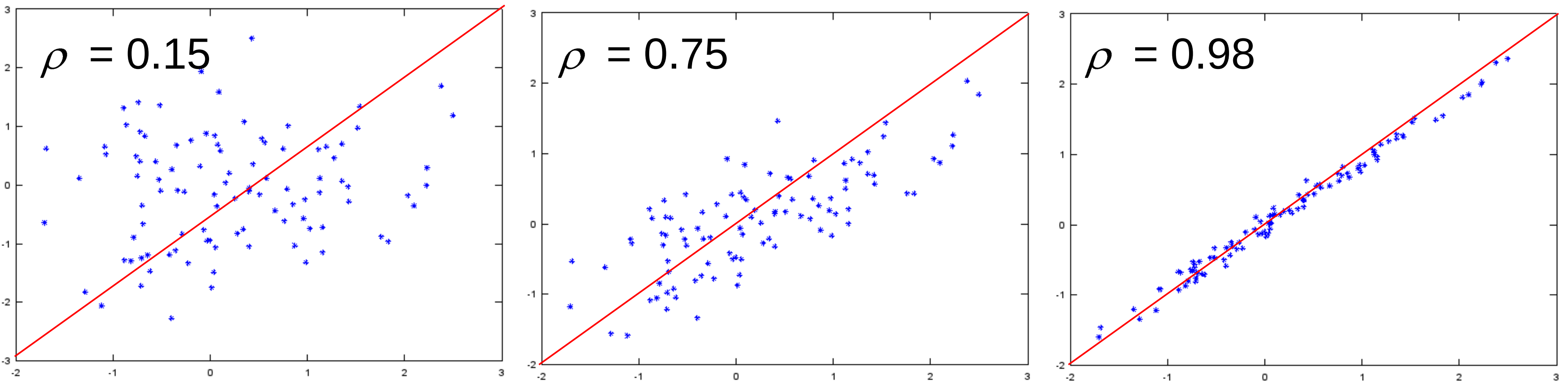
$$\boxed{\rho = \frac{\overline{xy} - \bar{x}.\bar{y}}{\sigma_x \sigma_y}}$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

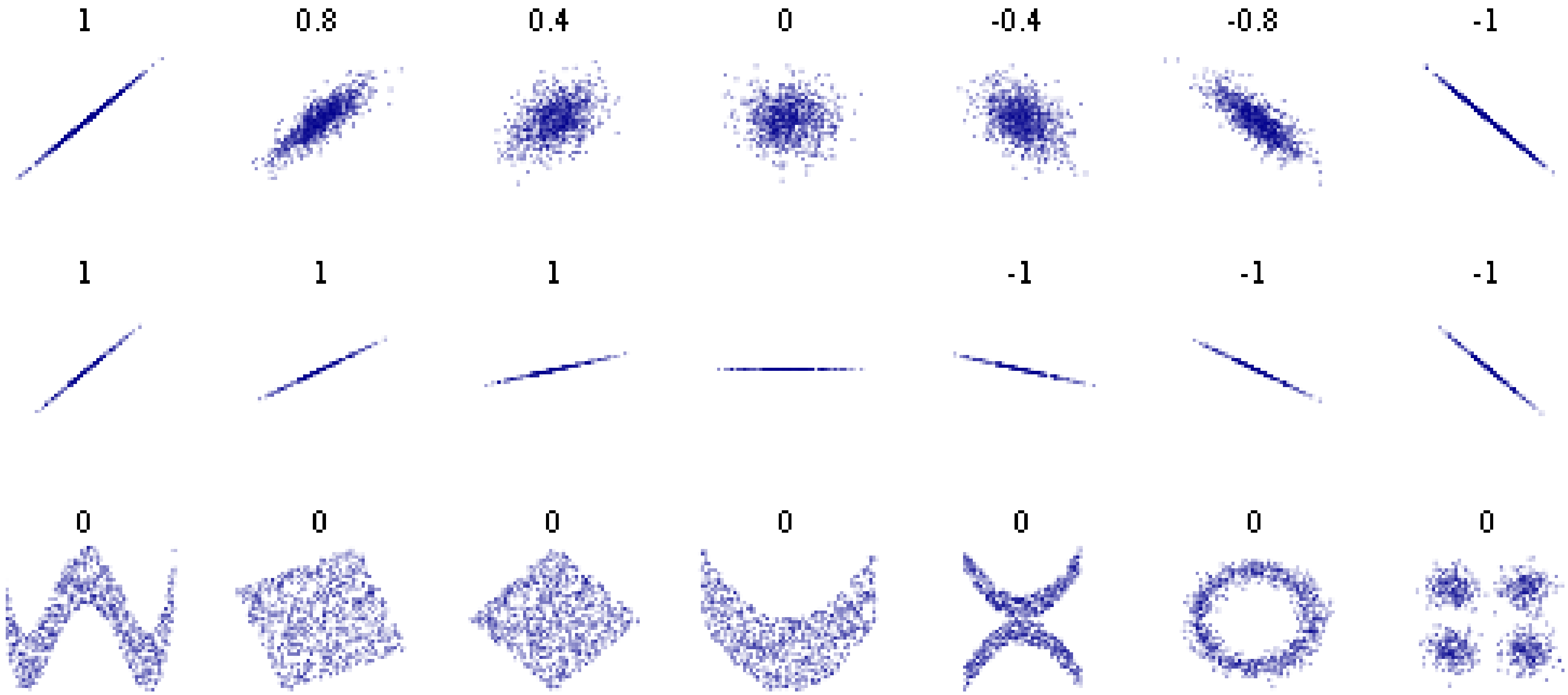
$$\rho = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sigma_x \sigma_y}$$

$$-1 \le \rho \le 1$$



$\rho = 0$ if there is no correlation

$\rho = \pm 1$ if X and Y are fully correlated


1
0.8
0.4
0
-0.4
-0.8
-1
1
1
1
-1
-1
-1
0
0
0
0
0
0
0

Example Using the following data, calculate the correlation coefficient of X and Y.

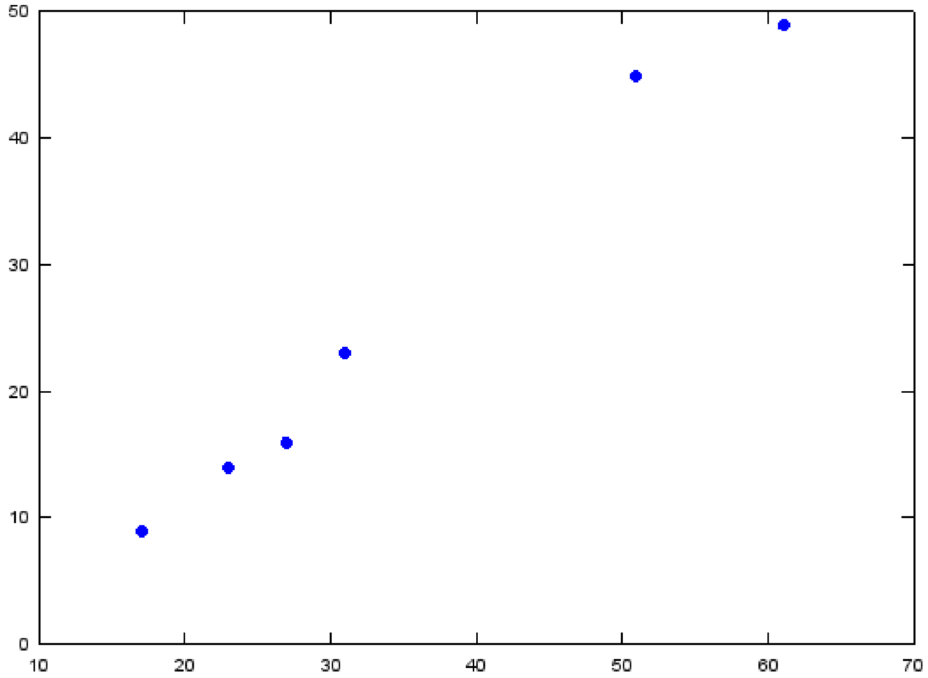
$X = \{17, 23, 27, 31, 51, 61\}$

$Y = \{9, 14, 16, 23, 45, 49\}$

Answer (rho = 0.9)

To solve the problem we can use also Octave.

```
octave:> X = [ 17 23 27 31 51 61 ];
octave:> Y = [  9 14 16 23 45 49 ];
octave:> plot(X,Y,'*')
octave:> rho=(mean(X.*Y)-mean(X)*mean(Y)) / (std(X)*std(Y))
rho =  0.82658
```



Median and Mod

The median is the number in the middle and
the mode is the most frequent number in a data set.

For the data set:
A = {3, 4, 4, 5, 6, 8, 8, 8, 10} => median = 6, mod = 8.

For the data set:
B = {5, 6, 7, 9, 11,12,18,18} => median=(9+11)/2=10, mod=18.

C = {2, 2, 5, 9, 9, 9, 10, 10, 11,12,18} => mod is 9 (*unimodal*)
D = {2, 3, 4, 4, 4, 5, 7, 7, 7, 9} => mod 4 and 7 (*bimodal*)
E = {1, 2, 3, 8, 9, 10, 12, 14, 18} => mod is unknown

PART II

CONFIDENCE INTERVAL

Population & Sample

In statistics it is very important to distinguish between population and sample.

A **population** is defined as all members of a specified group.

A **sample** is a part of a population.

- The sample is used to describe the characteristics (e.g. mean or standard deviation) of the whole population.
- The size of a sample may be 1%, or 10%, or 50% of the population, but it is never the whole population.

If the mean is measured using the whole population then this would be the population mean and is represented by μ.

$$Population\ mean\ (\mu) = \frac{sum\ of\ the\ population\ data}{population\ size\ N} = \frac{\sum x_i}{N}$$

N is the number of items in the population.

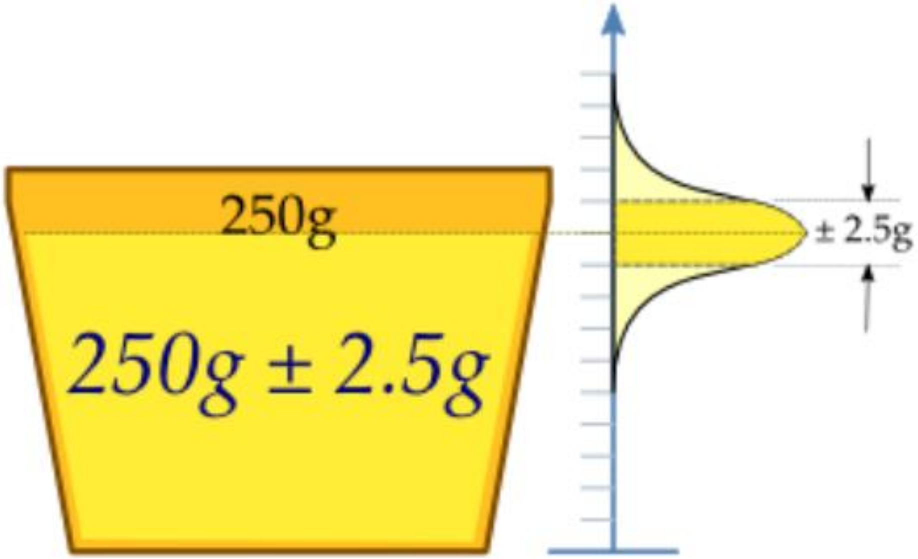
The mean of a sample is called as **sample mean** and is represented by $\bar{x}$.

$$Sample\ mean\ (\bar{x}) = \frac{sum\ of\ the\ sample\ data}{sample\ size\ n} = \frac{\sum x_i}{n}$$

n is the number of items in the sample.

Example

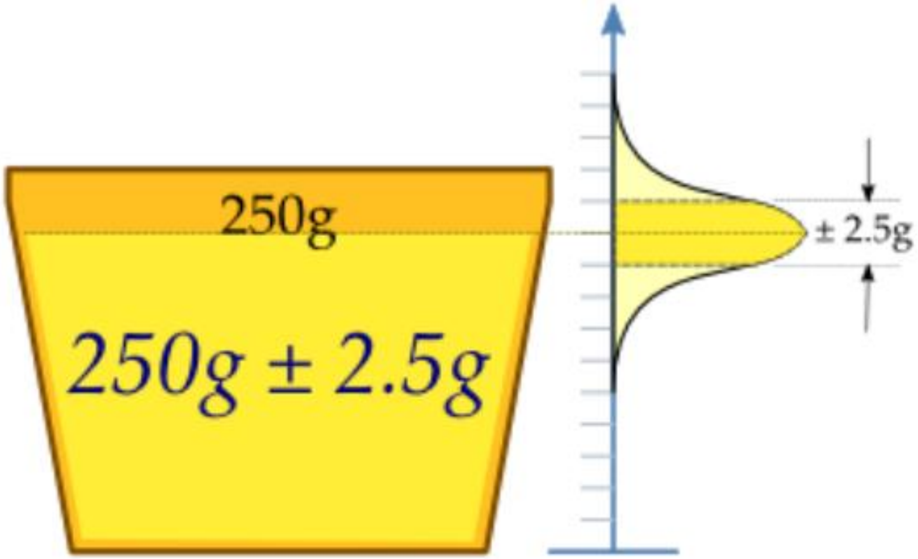
A machine fills cups with a liquid, and is supposed to be adjusted so that the content of the cups is 250 g of liquid. As the machine cannot fill every cup with exactly 250 g, the content added to individual cups shows some variation, and is considered a random variable *X*.



This variation is assumed to be normally distributed around the desired values: μ = 250 g and σ = 2.5 g.

To determine if the machine is adequately calibrated, a sample of n = 20 cups of liquid are chosen at random and the cups are weighed. The values are:

X={247.1, 250.0, 250.1, 249.8, 246.7,
254.4, 249.2, 249.4, 247.0, 247.0,
245.0, 253.3, 251.2, 250.7, 250.6,
247.3, 248.5, 248.0, 243.6, 250.2}



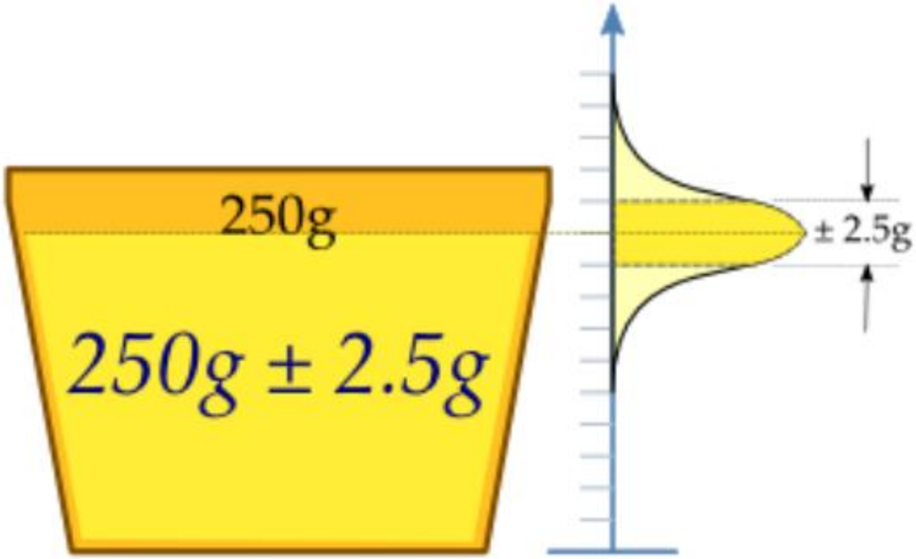
The **sample mean** and **sample standard deviation**:

```
>> x = [247.1, 250.0, 250.1, 249.8, 246.7, ...
        254.4, 249.2, 249.4, 247.0, 247.0, ...
        245.0, 253.3, 251.2, 250.7, 250.6, ...
        247.3, 248.5, 248.0, 243.6, 250.2];

>> mean(x)
ans =  248.9550
octave:> std(x)
ans =    2.6015
```

Is the result consistent with μ = 250 g and σ = 2.5 g?

Is the machine calibrated adequately?



Confidence Interval (=Güvenlik Aralığı)

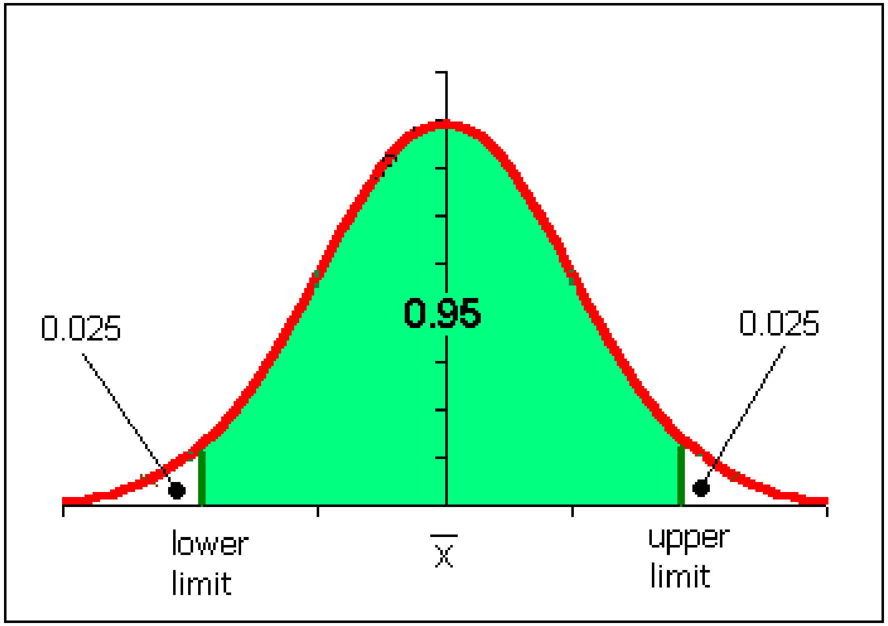
In statistics, confidence interval (CI)
is a type of interval estimate of a population parameter
and is used to indicate the reliability of an estimate.

How frequently the observed interval contains the parameter is determined by the confidence level (CL) or confidence coefficient.

http://en.wikipedia.org/wiki/Confidence_level

Confidence Levels are defined as follows:

CL	+- sigma
0.800	1.28σ
0.900	1.65σ
0.950	1.96σ
0.990	2.58σ
0.999	3.29σ



We usually use 95% CL which corresponds approximately +- 2σ region under the standard normal curve.

Simplified Expression for a 95% Confidence Interval

There is a constant multiplier, usually a constant around 2 or a little higher, that comes from the distribution being used and the degree of confidence required.

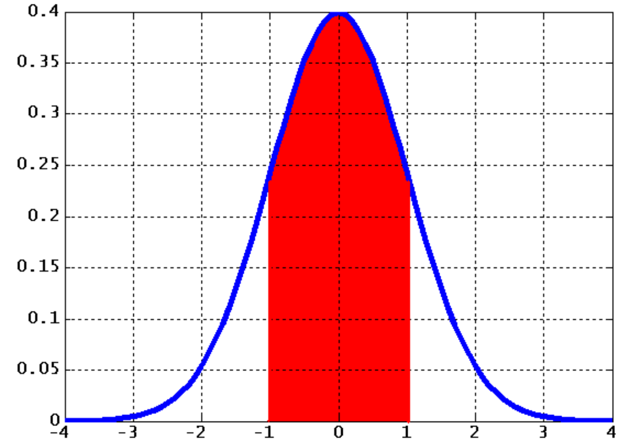
The "margin of error" is some multiplier times the standard error, and it is added to and subtracted from the mean to get the endpoints of the interval.

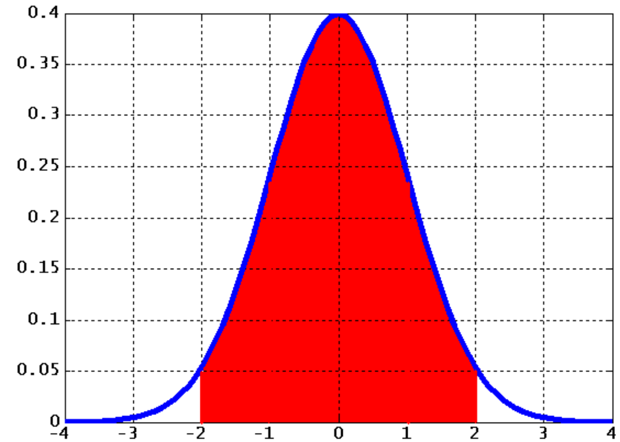
$$\overline{X} \pm 2\frac{\sigma}{\sqrt{n}}$$

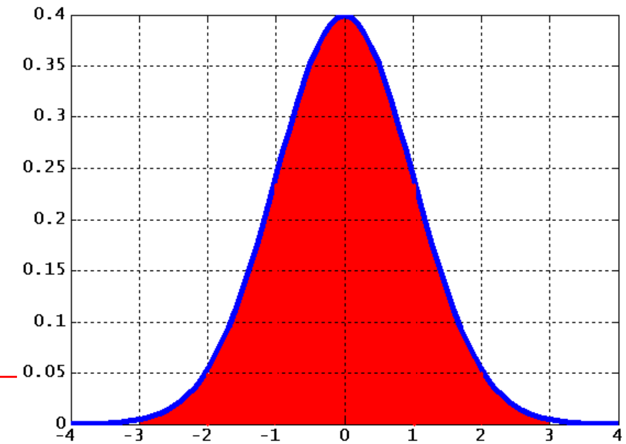
The sample mean is the best point estimate and so it is the center of the confidence interval.

The standard error of the mean, which is the standard deviation of the sampling distribution, is $\sigma/\sqrt{n}$.

CL	Area under the curve
%68	$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.6827$
%95	$\int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.9545$
%99.7	$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.9973$







Return back to our example.

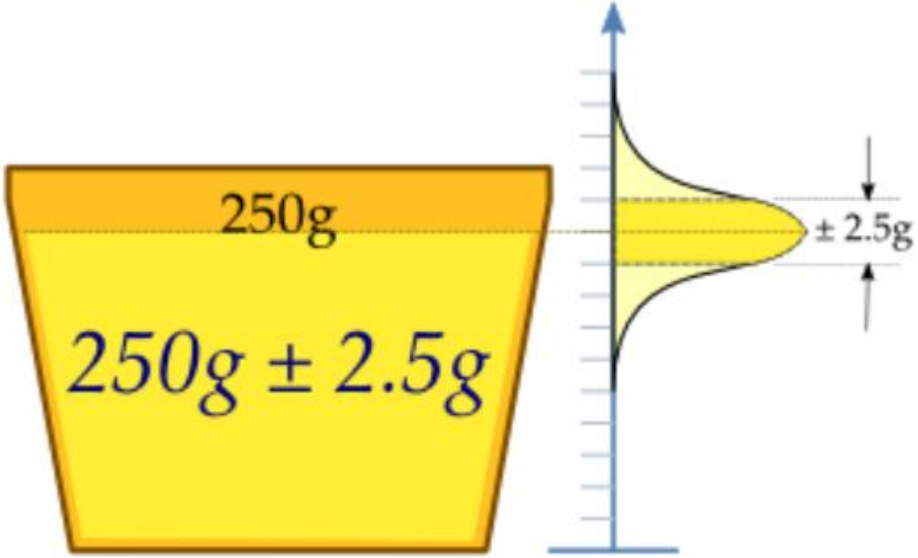
To get an impression of the expectation μ, it is sufficient to give an estimate.

The sample mean, standard deviation and standard error:

$$\bar{x} = \frac{\sum x_i}{20} = 248.955\ \mathrm{g}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{20-1}} = 2.6015\ \mathrm{g}$$

$$\sigma_E = \frac{\sigma}{\sqrt{n}} = \frac{2.6015}{\sqrt{20}} = 0.5817\ \mathrm{g}$$



with 95% confidence level, the population mean lies between the interval:

$$\bar{x} - 2\sigma_E \leq \mu \leq \bar{x} + 2\sigma_E$$

$$248.955 - 2(0.5817) \leq \mu \leq 248.955 + 2(0.5817)$$

$$247.7916 \leq \mu \leq 250.1184$$

That is: the true mean is somewhere between [247.7916, 250.1184] with 95% probability.

Remember, the (claimed) true mean is μ = 250 g.

- *Therefore our sample mean is consistent with the true mean.*
- *There is no reason to believe the machine is wrongly calibrated.*

Example

A coin is thrown 30 times.

(a) Calculate the mean (expected) number heads.

(b) Imagine you observed 20 heads.
Compute how many standard deviations your observation differ from the mean value. Is the coin fair?

(c) Imagine you observed 30 heads.
Compute how many standard deviations your observation differ from the mean value. Is the coin fair?

Binomial Distribution:

(a) $$P_{\text{binom}} = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu = np = 30 \times 0.5 = 15$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{30 \times 0.5 \times (1-0.5)} = 2.74$$

(b) $$(20-15)/2.74 = 1.83$$

1.83 sigma difference

20 heads is consistent with 15 => the coin is fair

(c) $$(30-15)/2.74 = 5.47$$

5.47 sigma difference

30 heads is <u>not</u> consistent with 15 => discovery, the coin is <u>not</u> fair.

For any distribution, if your result is

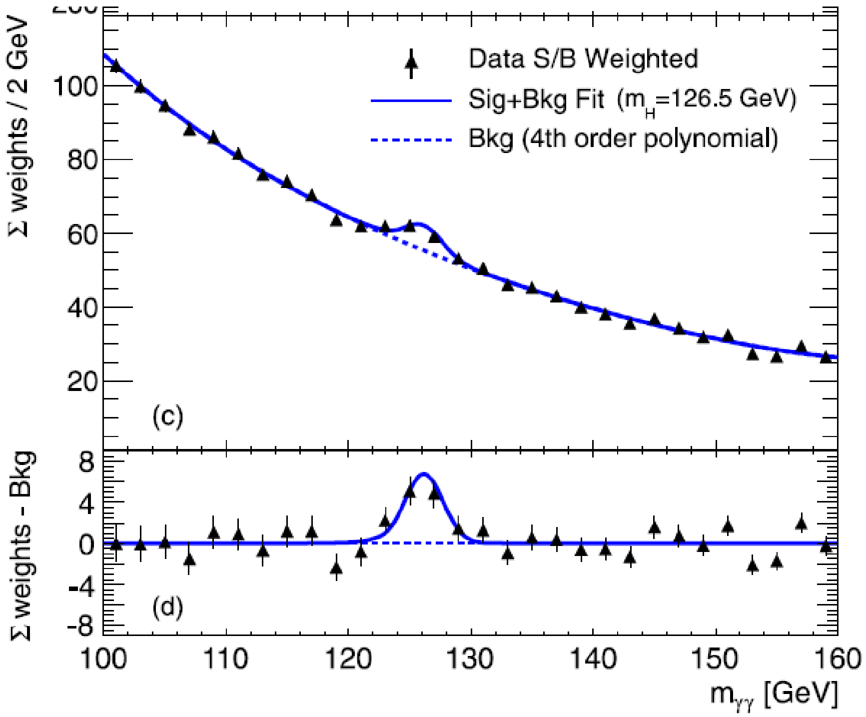
+-1σ away from true mean -> in *good agreement*

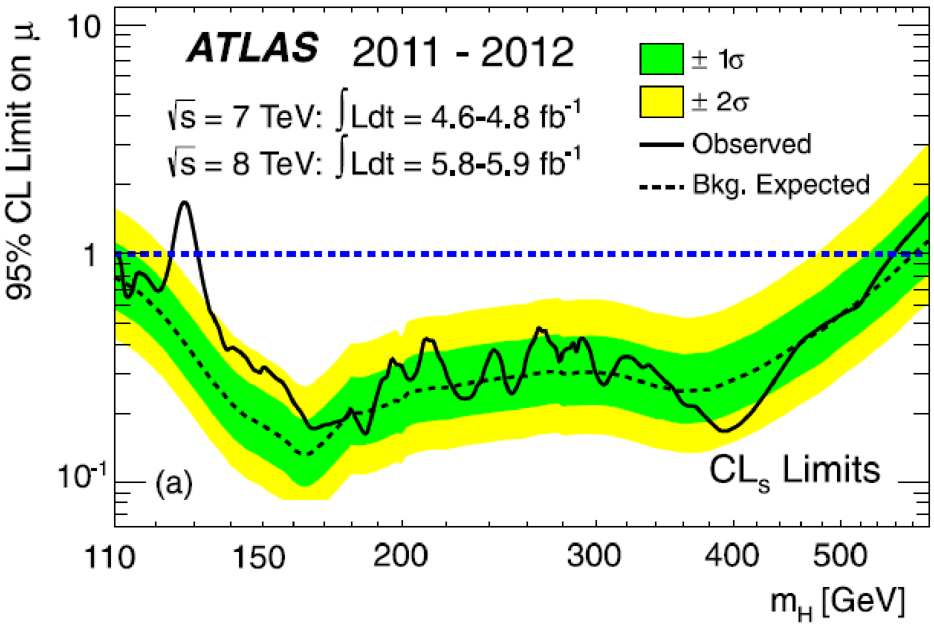
+-2σ away from true mean -> *consistent*

+-3σ away from true mean -> there is a *signal* for something

+-5σ away from true mean -> you *discover* something

Discovery of the Higgs Boson at CERN





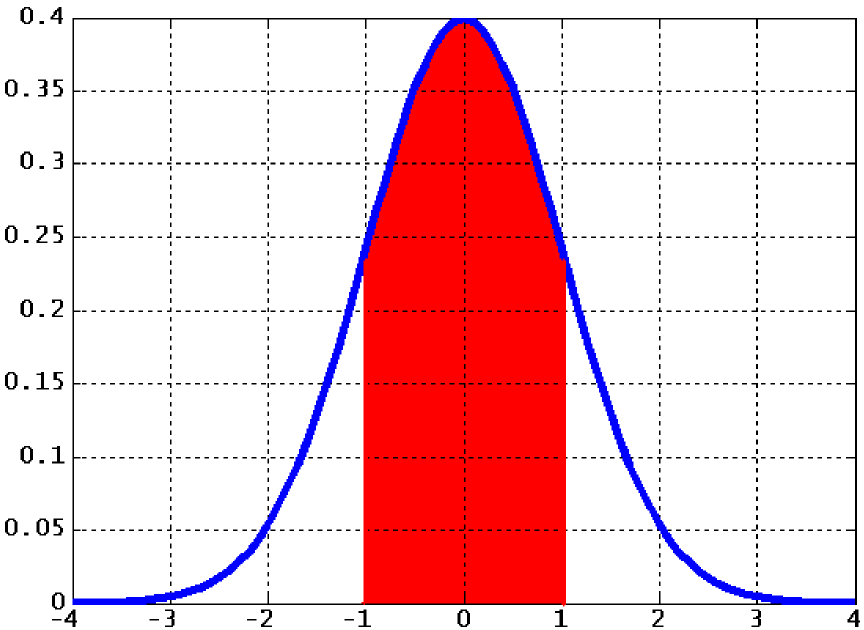
Important Functions for Measurement & Calibration

1. Gaussian Function

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2 / 2\sigma^2]$$

Mean: μ

Std.Dev: σ

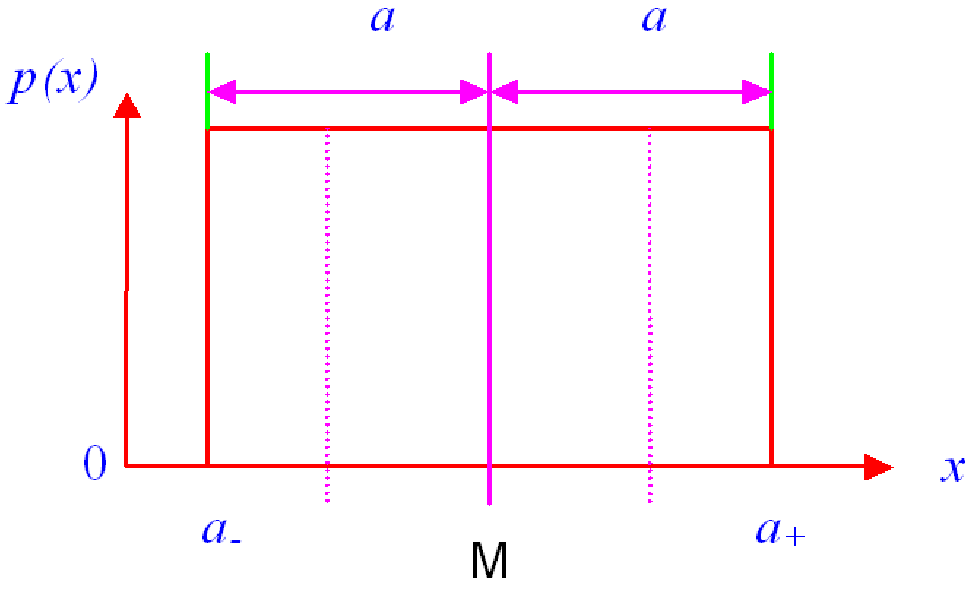


Important Functions for Measurement & Calibration

2. Rectangular Distribution

Mean: M

Std.Dev: $\sigma = \dfrac{a}{\sqrt{3}}$

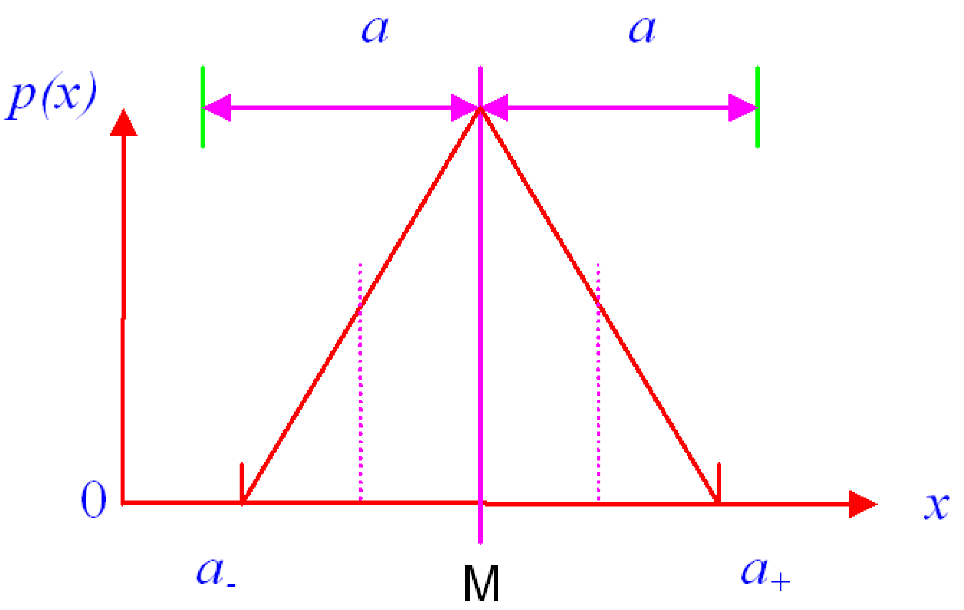


Important Functions for Measurement & Calibration

3. Triangular Distribution

Mean: M

Std.Dev: $\sigma = \dfrac{a}{\sqrt{6}}$



Questions

1. Time of execution of a computer program in seconds are given by:
   `T = {358,353,357,358,362,364,358,361,360,355}`
   Calculate mean, median and mod of the data.

2. In a hospital, the mean values of weights of 250 babies as a function of month are obtained*. Determine the correlation coefficient between:

(a) month and boy

(b) month and girl

(c) boy and girl

	Weight (kg)	
Month	**Boy**	**Girl**
0	3.4	3.2
1	4.4	4.1
2	5.5	5.0
3	6.4	5.7
4	7.1	6.4
5	7.7	6.9
6	8.3	7.5
9	9.4	8.6
12	10.2	9.4

* Data is taken from: http://dergiler.ankara.edu.tr/dergiler/36/854/10838.pdf

3. An industrial refrigerator is used to cool food in a processing factory. An experiment is performed to test the effect of wind speed of the refrigerator on the temperature in the refrigerator. The results are given in the table.

(a) Plot the data

(a) Calculate the correlation coefficient

(b) Comment on the result

W (m/s)	T(^{o}C)
3.8	1.69
8.4	1.34
7.3	1.45
3.9	1.75
1.7	1.87
9.6	1.05
4.5	1.60
6.4	1.45
0.4	2.02
8.7	1.15
8.8	1.15
0.7	1.99

4. Repat the example, by using the rectangular distribution function for a = **2.5 g**.

5. Repat the example, by using the triangular distribution function for a = **2.5 g**.

Answer:

	Mean	Sigma	Interval with 95% CL
Gaussian	248.9550	2.6015	[247.7916, 250.1184]
Rectangular	248.9550	1.4434	[246.0682, 251.8418]
Triangular	248.9550	1.0206	[247.9344, 249.9756]

References

1. “Data Analysis for Physical Science Students”, L. Lyons, Cambridge University Press
2. “Probability and Statistics” - M. Spiegel et. al., Shaum
3. “Probability” - S. Lipshutz, Shaum
4. “Radiation Detection and Measurement”, G.F.Knoll, Wiley