Introduction to Statistics

Lecture for CHE 1148

Sirish L. Shah
Department of Chemical and Materials Engineering
University of Alberta, Canada
Visiting Professor @ Hiroshima University, Japan

Organization

- Basic introduction of statistics
  - Random variables and their distributions
  - The central limit theorem
  - Autocorrelation
  - Covariance and Correlation
  - Laws of probability
  - Bayes Theorem

Normal Distribution

- Repeated observations that differ because of experimental error often vary about some central value in a roughly symmetrical distribution in which small deviations occur much more frequently than large ones. A continuous probability distribution, which is valuable for representing this situation is the **Gaussian or Normal Distribution**. It is given by

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \qquad -\infty < y < \infty$$

where μ is the mean, and σ is the standard deviation.

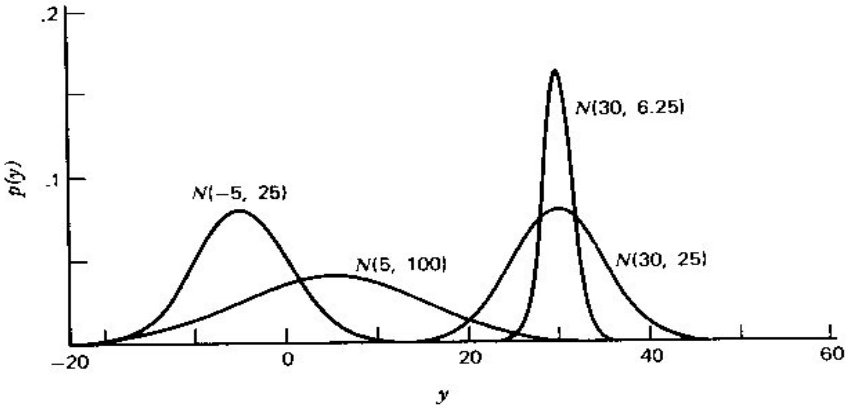


Figure 1: Normal distribution with different means and variances

- Denoted by $N(\mu, \sigma^2)$

Characterizing a Normal Distribution

- Once the mean μ and the variance σ^2 of a normal distribution are given, the entire distribution is characterized. Area under a normal distribution=1.

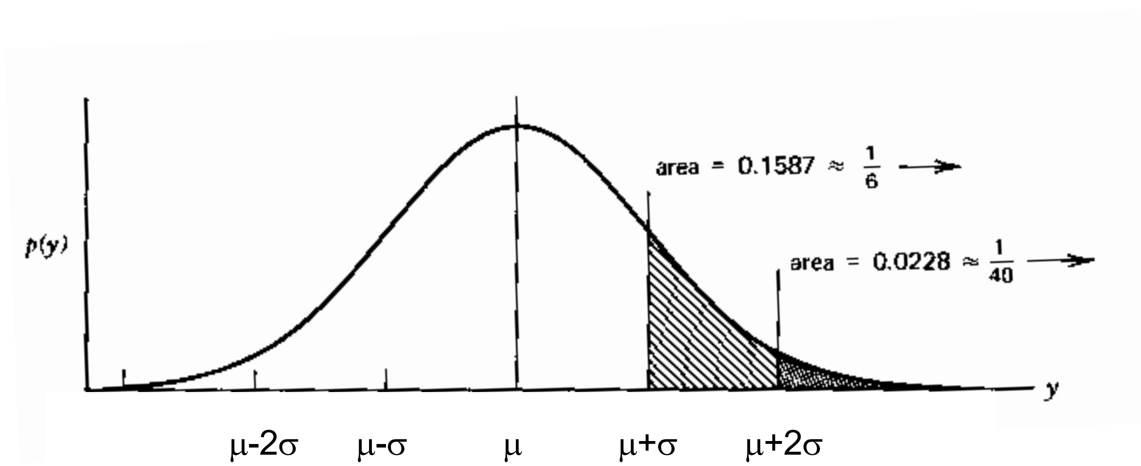
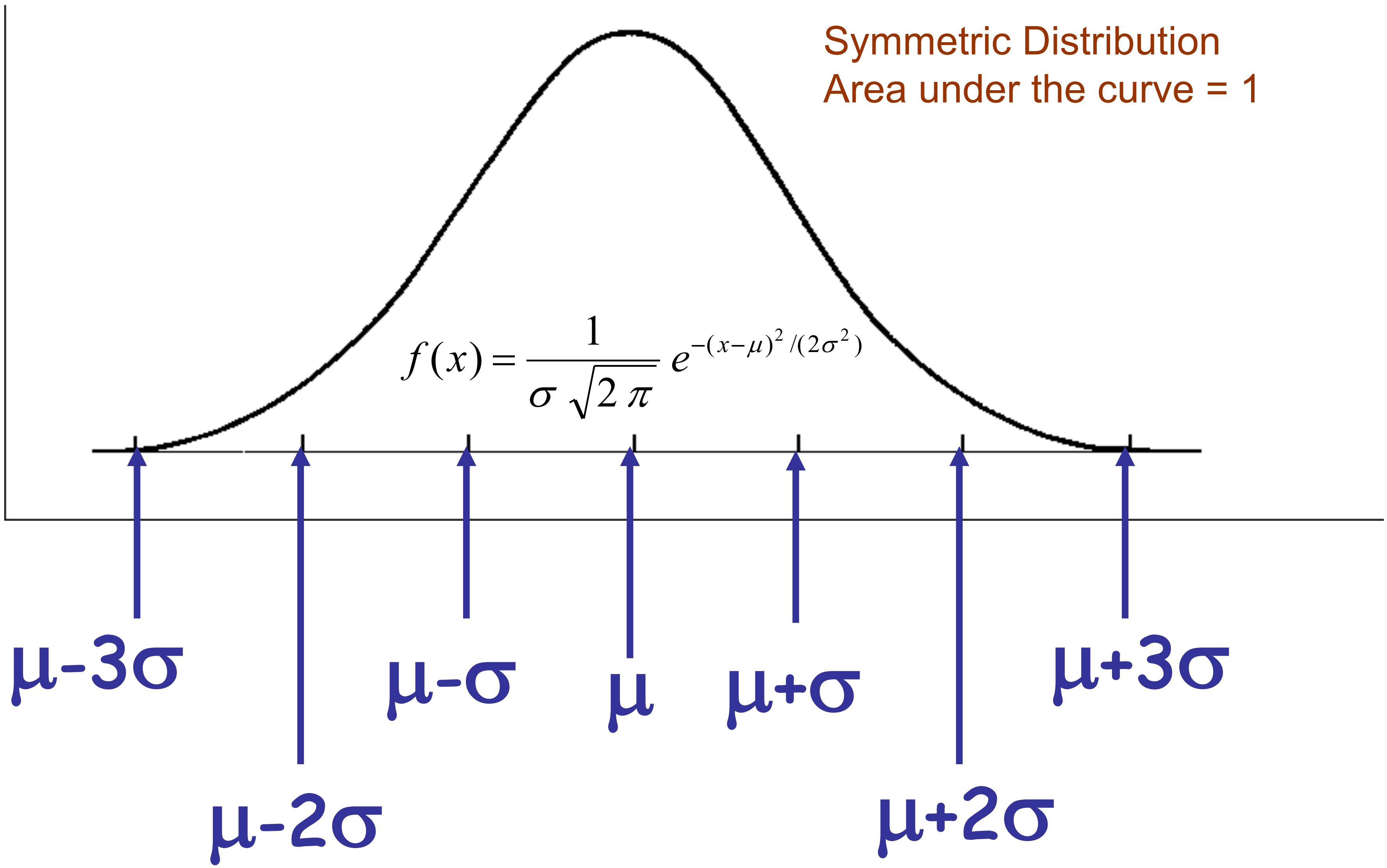
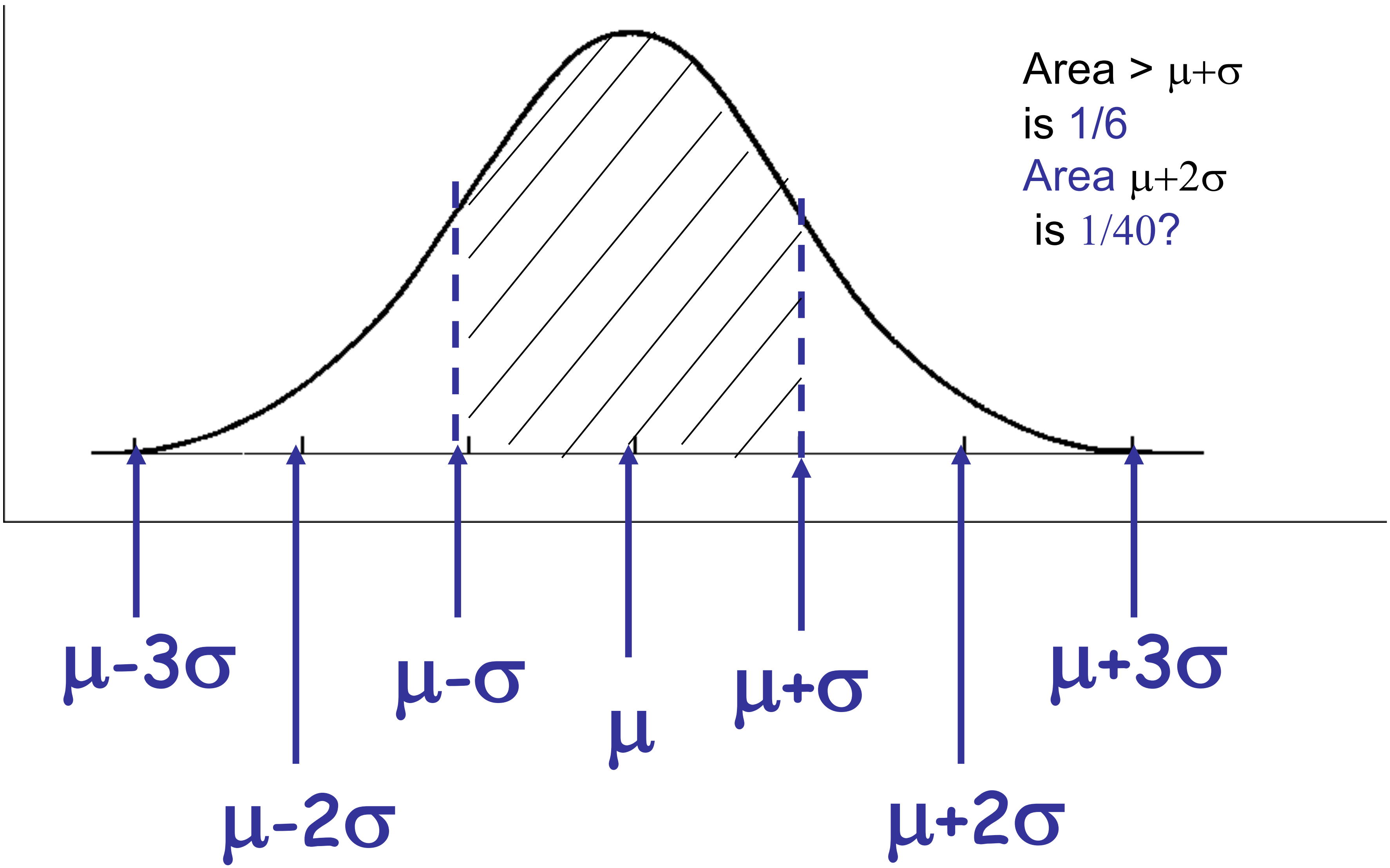


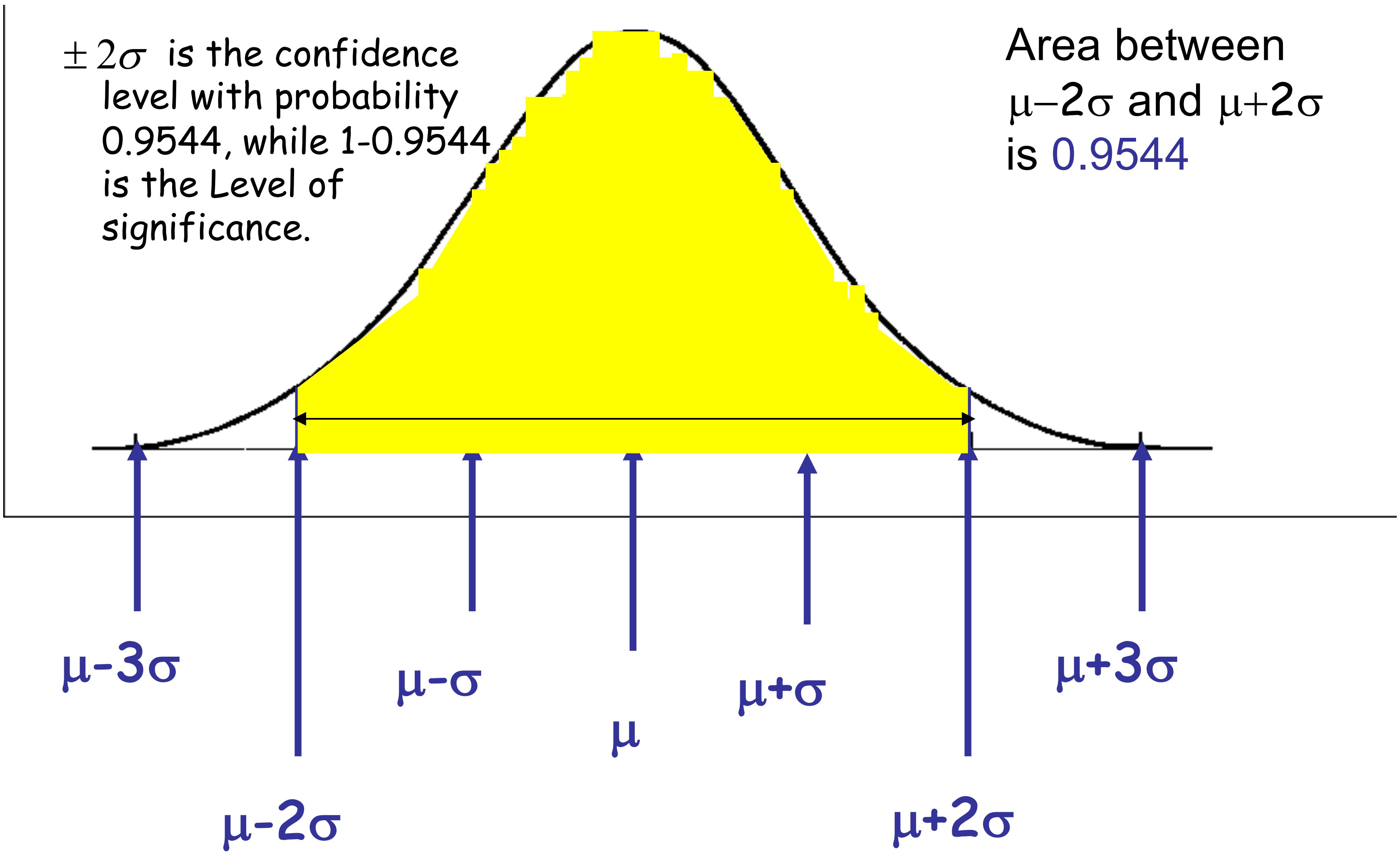
Figure 2: Tail areas of the normal distribution

Normal Distribution

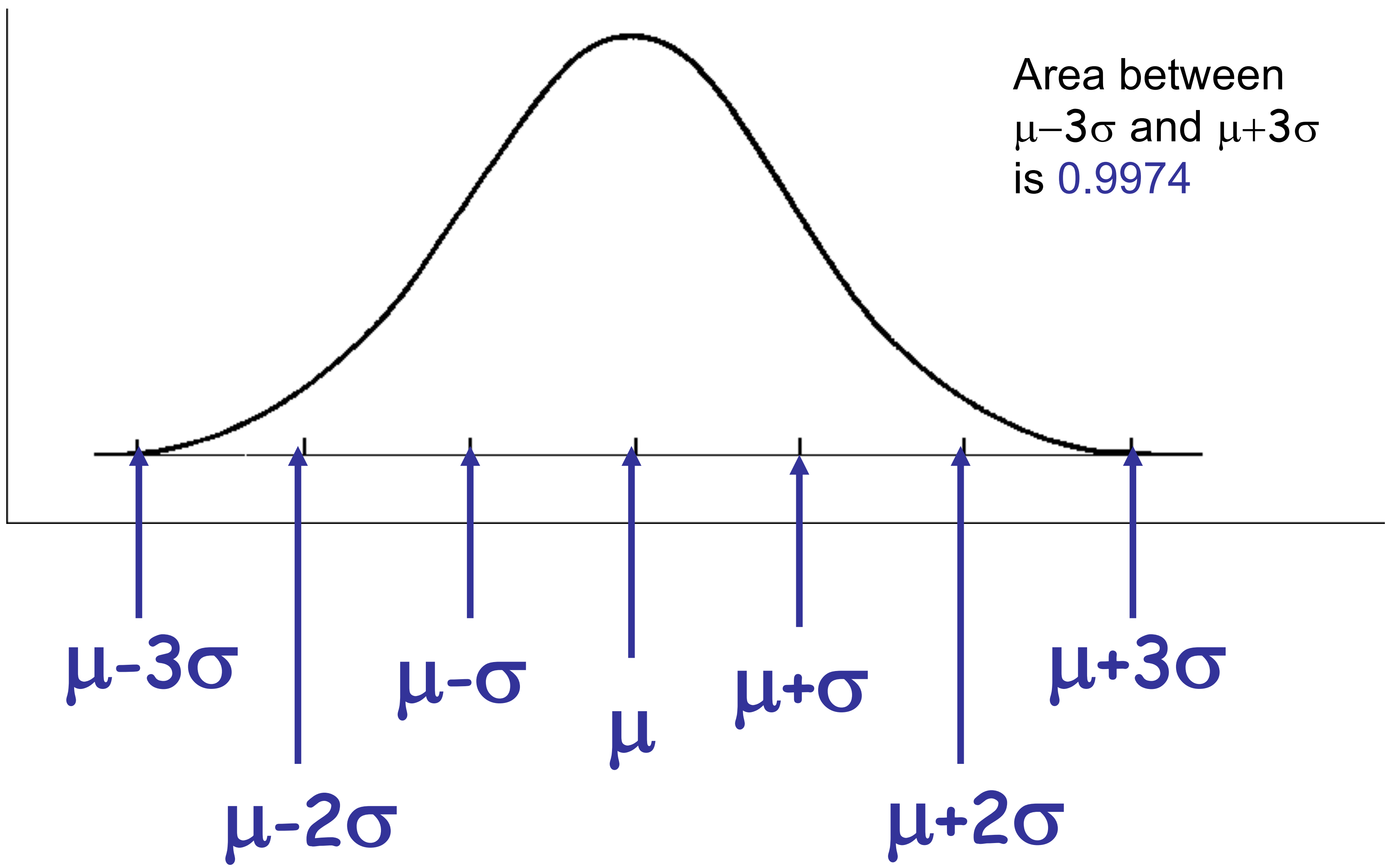




Area > μ+σ
is 1/6
Area μ+2σ
is 1/40?
μ-3σ
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
μ+3σ



Area between $\mu-2\sigma$ and $\mu+2\sigma$ is 0.9544



Area between
μ−3σ and μ+3σ
is 0.9974
μ-3σ
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
μ+3σ

Standard Normal Distribution

- If y is a normally distributed variable, then it can be expressed in terms of a standardized normal deviate or a unit normal deviate,

$$z = \frac{y - \mu}{\sigma}$$

The distribution of *z* is *N(0,1).* Tables for unit normal deviate are given in most statistics books.

Using Tables of Normal Distribution

Example 1

- Suppose the daily level of an impurity in a reactor feed is known to be approximately normally distributed with a mean of μ = 4.0 and a standard deviation of σ = 0.3. What is the probability that the impurity level on a *randomly chosen* day will exceed 4.4 ?

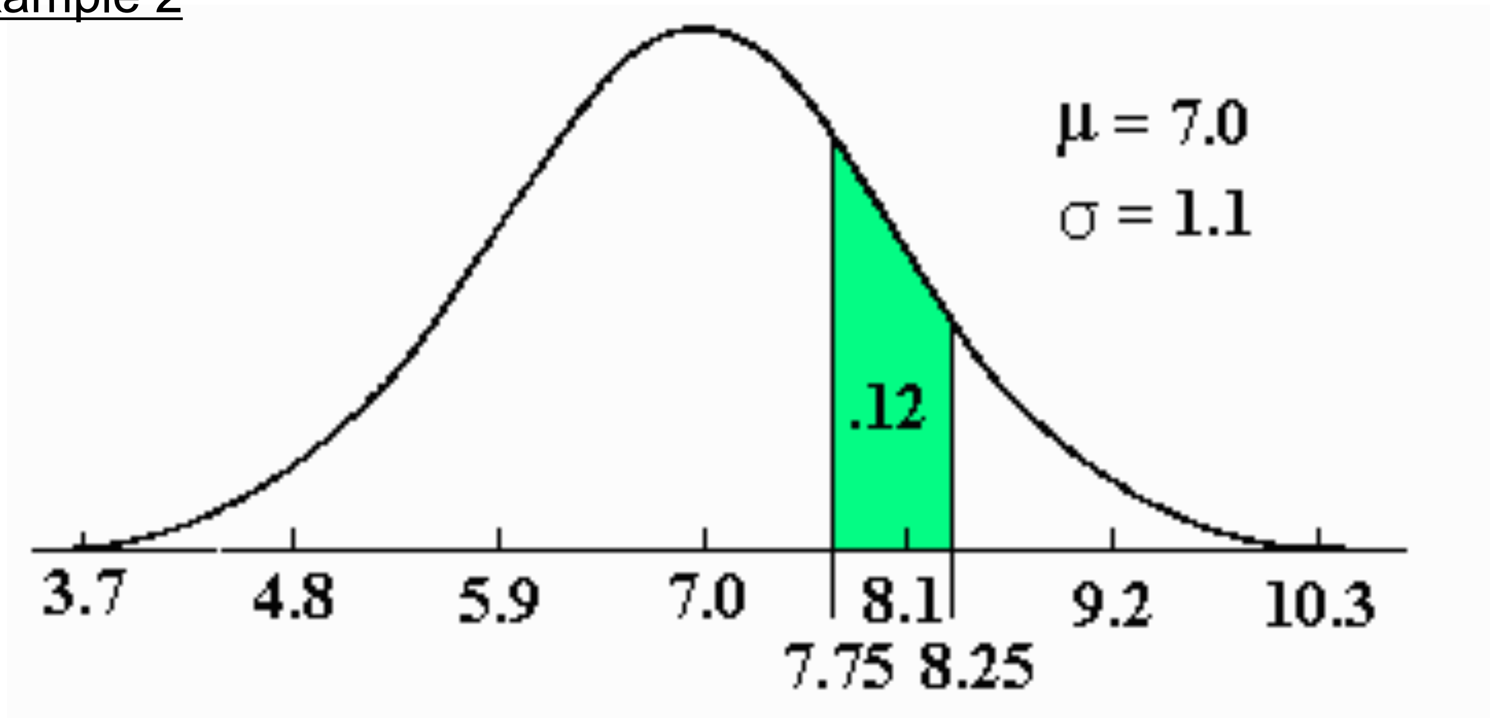
$$\Pr(y > 4.4) = \Pr(y-\mu > 4.4-\mu) = \Pr\left(\frac{(y-4.0)}{0.3} > \frac{(4.4-4.0)}{0.3}\right)$$

From the Table of unit normal distribution we find that $\Pr(z > 1.33) = 0.0918$, *i.e.* there is 9% probability that the impurity level on a randomly chosen day will exceed 4.4.

Or

Out of 100 randomly chosen days, there will be 9 days when the impurity level will exceed 4.4.

Example 2



$$P(7.75 \leq x \leq 8.25) = \int_{7.75}^{8.25} \frac{1}{1.1\sqrt{2\pi}} e^{-\frac{(x-7)^2}{2\times 1.1^2}} dx$$

= cdf('Normal',8.25,7,1.1)-cdf('Normal',7.75,7,1.1)

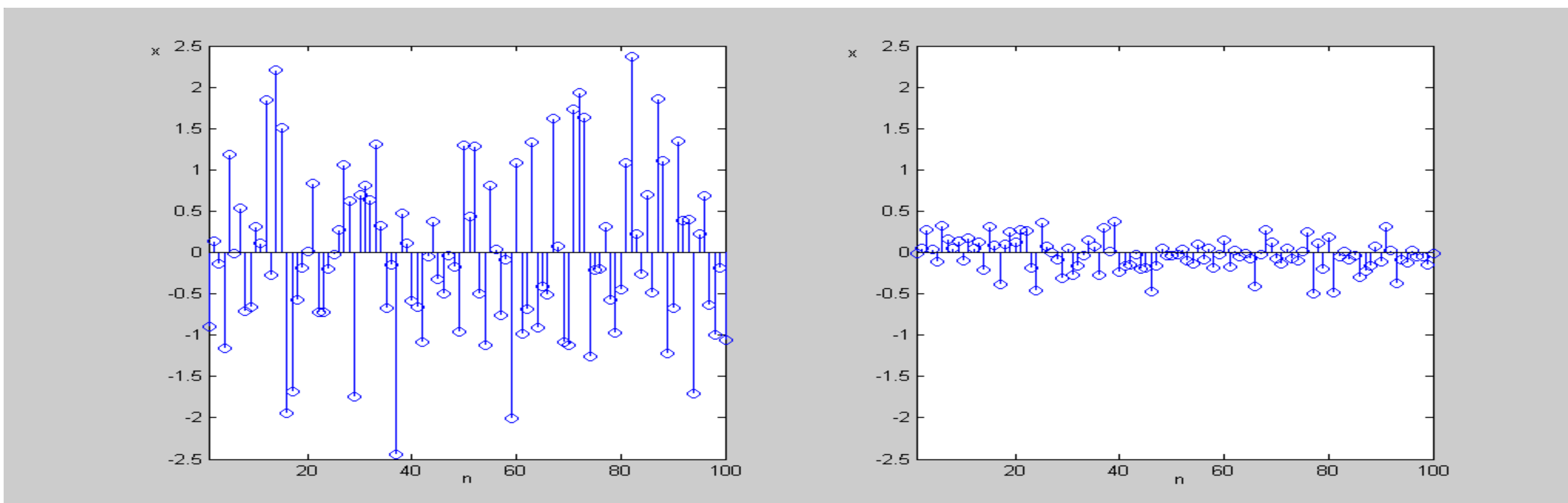
The (population) variance of a RV gives an idea of how widely spread the values of the RV are likely to be. It is the second moment of the distribution, indicating how closely concentrated round the expected value of the distribution is. Variance is defined by

$$\mathrm{Var}(x) = E(x)^2 - (E(x))^2 = \sigma^2.$$

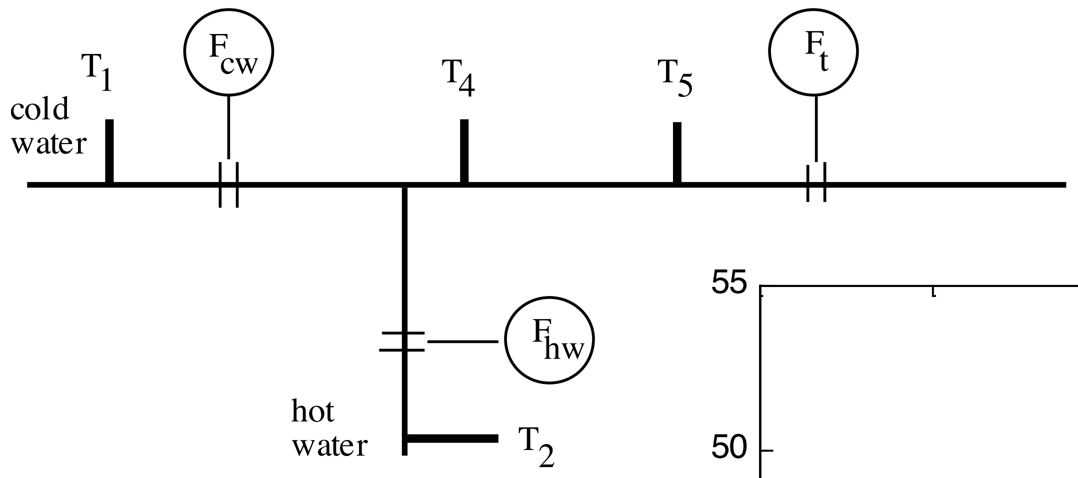
$\sigma = \sqrt{Var(X)}$ is referred to as the standard deviation.

Example 3

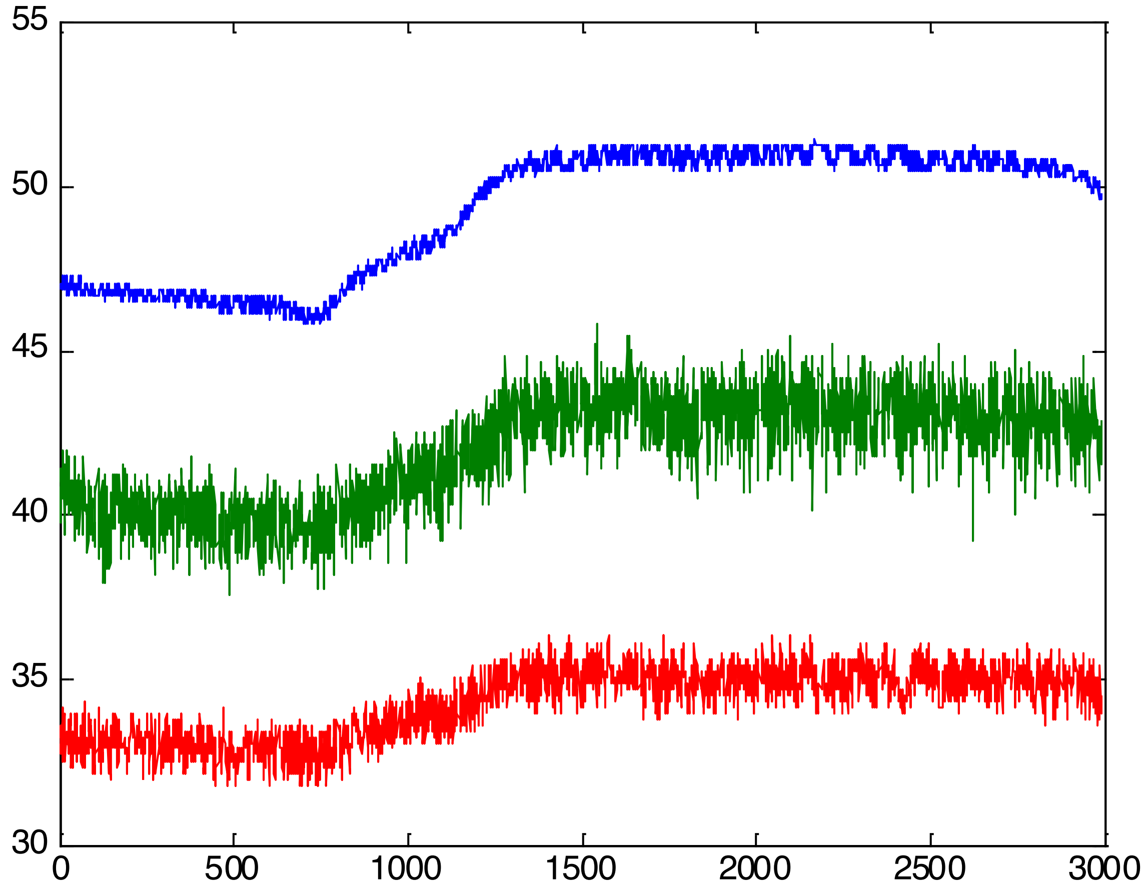
$x \sim N(0,1)$ $\qquad\qquad$ $x \sim N(0, 0.2^2)$



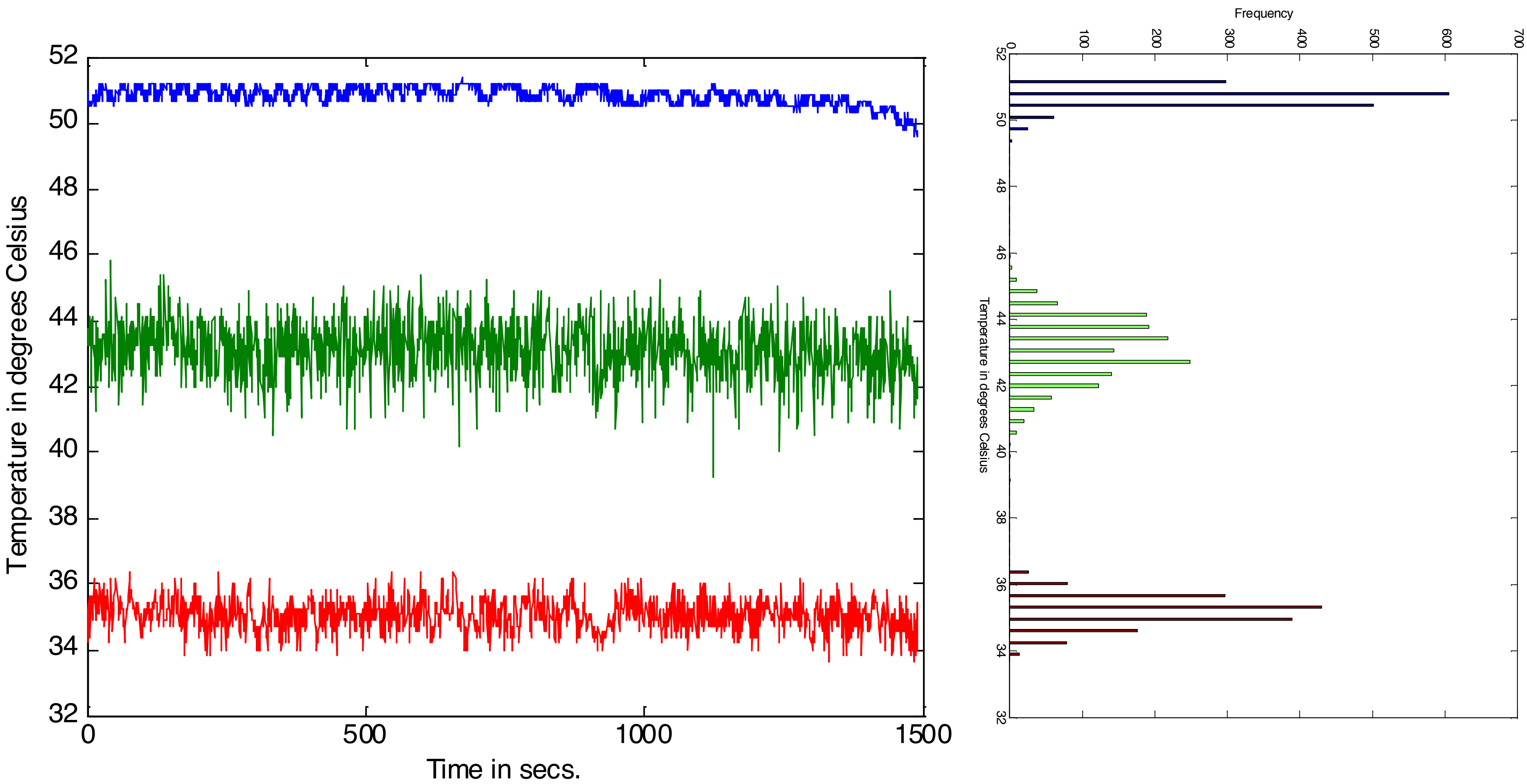
A snapshot of real data



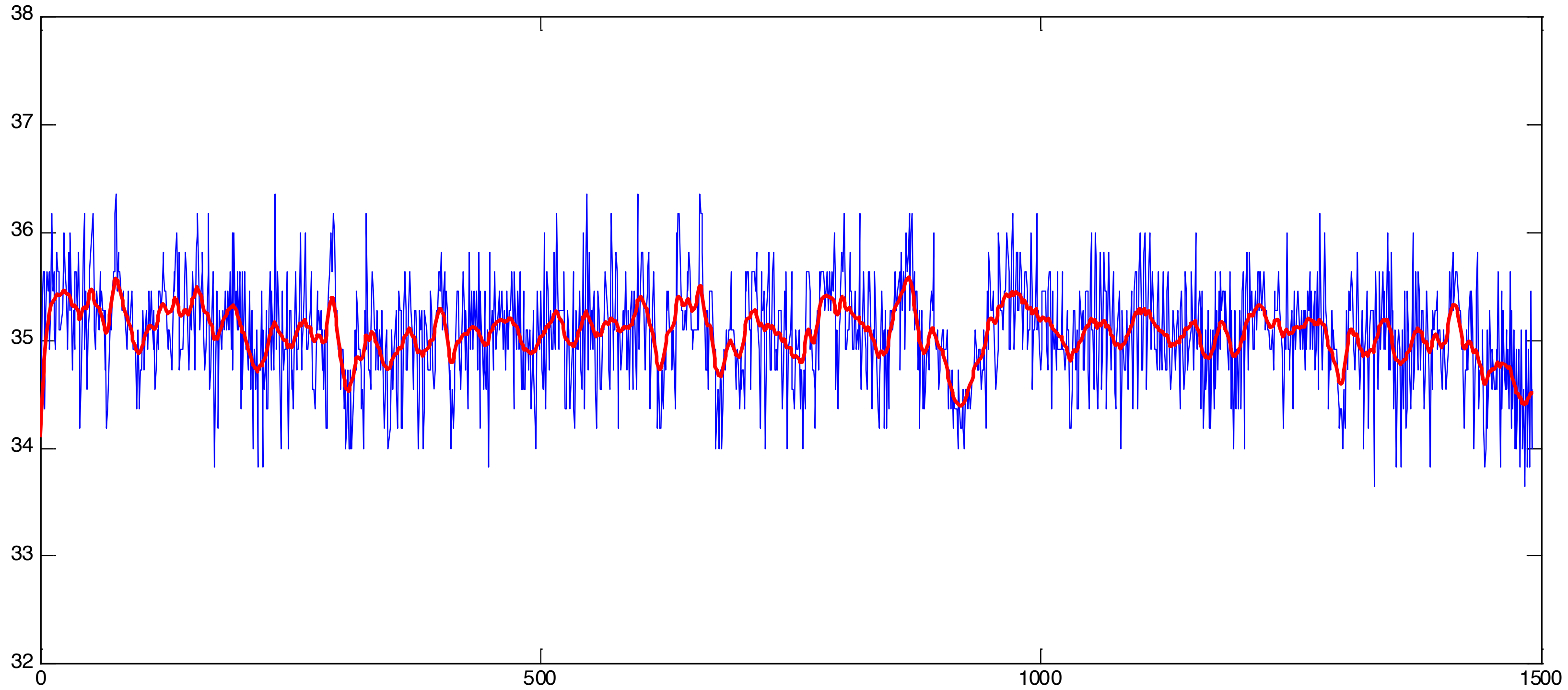
Schematic of a mixing process. “F” stands for flowrate, and “T” stands for temperature



Distribution of real data



Look at T5



Raw data:

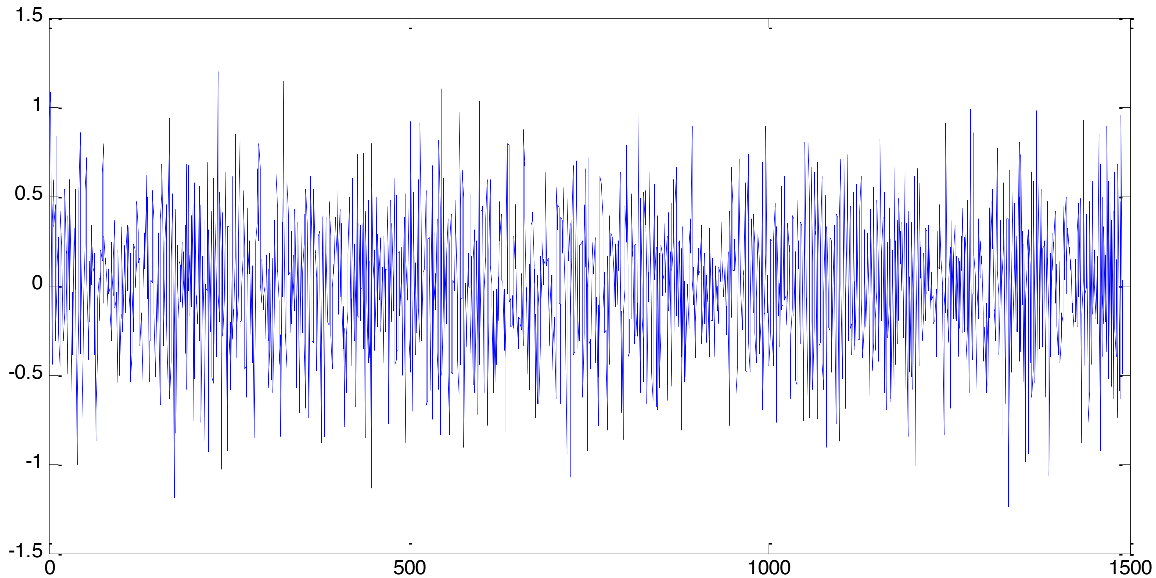
```
>> mean(T5)
ans =
  35.0581
>> std(T5)
ans =
   0.4864
```

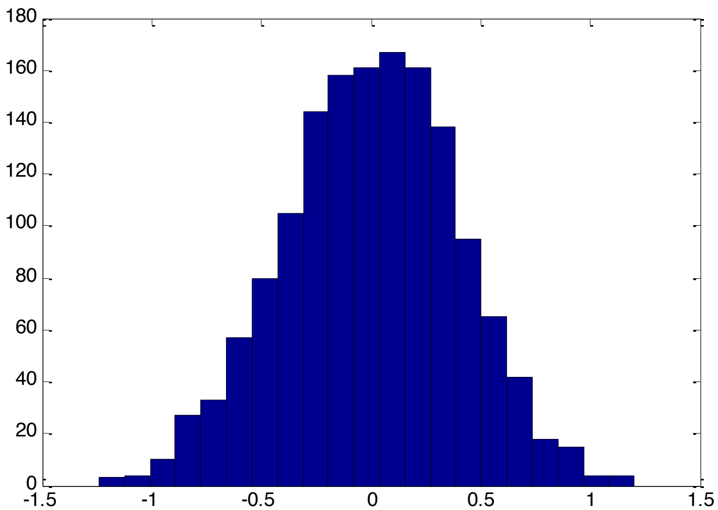
Filtered data:

```
>> Tff=filtfilt(0.2, [1 -0.8], T5);
>> figure; plot(T5); hold;
>> plot(Tff, 'r')
>> mean(Tff)
ans =
  35.0549
>> std(Tff)
ans =
   0.2175
```

Distribution of the error

```
>> e=T5-Tff;
>> hist(e, 21)
>> figure; plot(e)
>> mean(e)
ans =
   0.0032
>> std(e)
ans =
   0.3963
```





Sample

- Sample: The usually few observations that have actually occurred as some kind of sample from the population.

  For example, ages of students in this class is a sample from the theoretical population of ages of all engineering students in the world.

Population vs. Sample

	Population	Sample
Definition	A hypothetical set of N observations from which the sample of observations can be imagined to come (typically N is very large)	A set of n observations actually obtained. Typically n is relatively small
Measure of Location	Population mean: $\mu = E(y) = \frac{\sum y_i}{N} = \int_{-\infty}^{+\infty} yp(y)dy$	Sample mean: $\bar{y} = \frac{\sum y_i}{n}$
Measure of spread	Population variance: $\sigma^2 = \frac{\sum (y_i - \mu)^2}{N} = \int_{-\infty}^{+\infty} (y-\mu)^2 p(y)dy$	Sample variance: $s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1}$
	Population standard deviation: $\sigma = \sqrt{\sigma^2}$	Sample standard deviation: $s = \sqrt{s^2}$

“E” denotes the expectation operator.

In the previous slide: *E* denotes the expectation operator and is used to mean:

$$\underset{N \to \infty}{Limit} \frac{1}{N} \sum_{i=1}^{N}$$

For example, to find the population mean of { *y* }, one can use the expectation operator as:

$$E(y) = \underset{N \to \infty}{Limit} \frac{1}{N} \sum_{i=1}^{N} y_i = \mu$$

It is seldom possible to have an infinitely large sample to work with. Thus, one has to work with descriptive statistical measures of a finite sample or a finite number of observations

Sample Mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Sample Variance

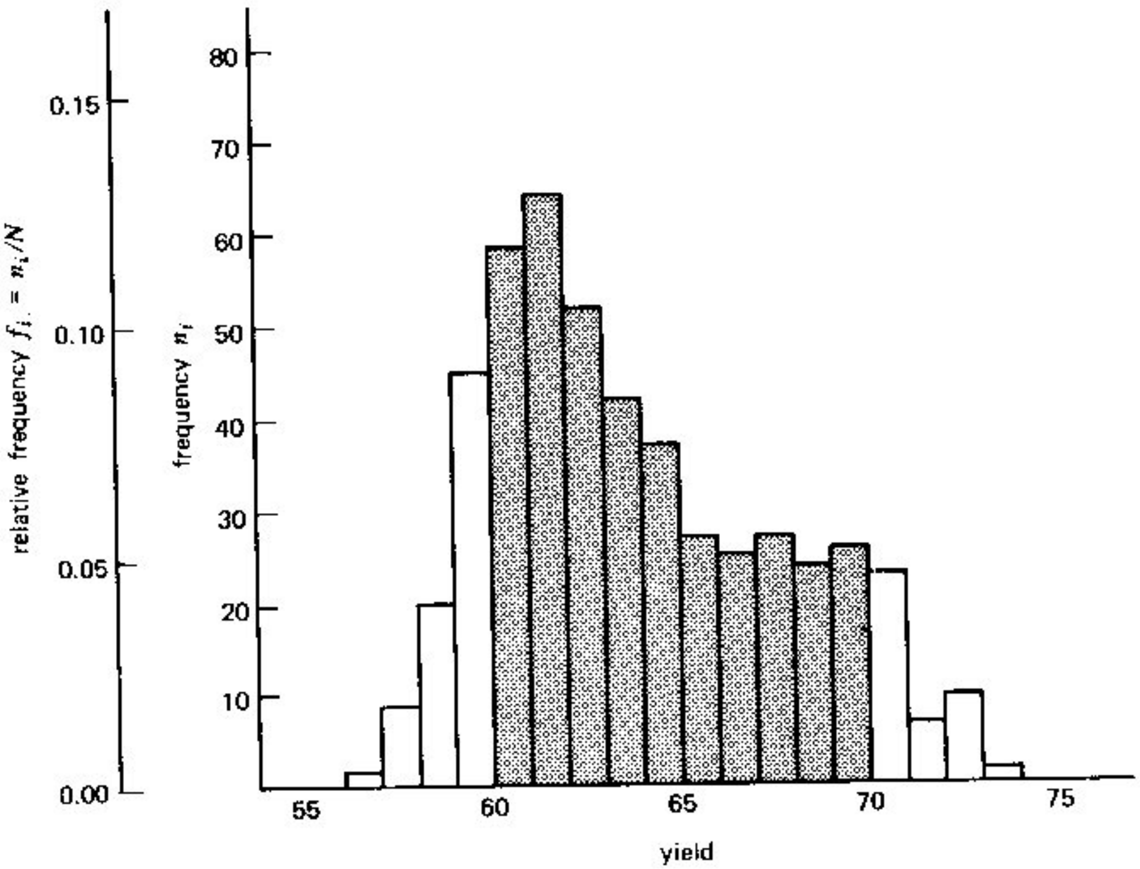
$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Sample Std. Dev.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

The n-1 in the denominator of s^2 is due to the loss of one degree of freedom (DOF) in replacing m by $\bar{y}$.

Data Representation: Frequency Distribution/Histogram



When large number of observations, then frequency distribution

Fig. 3: Frequency distribution of a sample of 500 hypothetical plant yields

Frequency Histogram

- Gives a vivid representation of the distribution of observations.
  - Gives us an idea about the location and spread
  - Also tells us, for example that about 4/5 of the observations lie between 60 and 70. This fraction, more precisely 382/500, is represented by the shaded area under the frequency diagram between values 60 and 70.
  - Also, note that if we make the area of the rectangle erected on the i^{th} interval of the histogram equal to the relative frequency n_i/N of the values occurring in that interval, this is equivalent to choosing the vertical scale so that the area under the whole histogram is equal to unity.

Definition of Population

- Population of observations: The total aggregate of observations that conceptually might occur as the result of performing a particular operation in a particular way is referred to as the population of observations. Population is assumed to have size N, where N is very large
- For example: Population of ages of engineering students in the world.

Frequency Distribution of Population

For a large population, the bumps in the frequency diagram due to sampling variation would disappear, and we might obtain a histogram having the regular appearance as below. Note that the ordinate is relative frequency

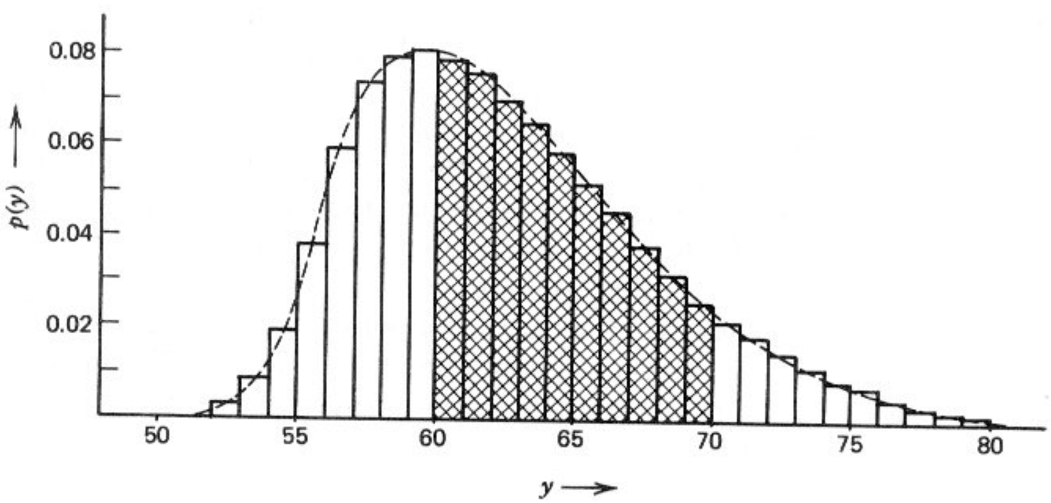


Figure 4: Probability distribution for a population of yield values

Probability Distribution

- The continuous (without bumps) relative frequency distribution for a population is known as the probability distribution function, with the properties that:

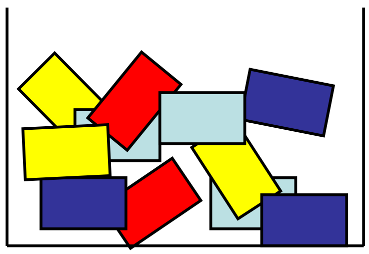
$$1 \geq p(y) \geq 0$$ The relative frequency is between 0 and 1

$$\int_{-\infty}^{+\infty} p(y)dy = 1$$ The total area under the curve is 1

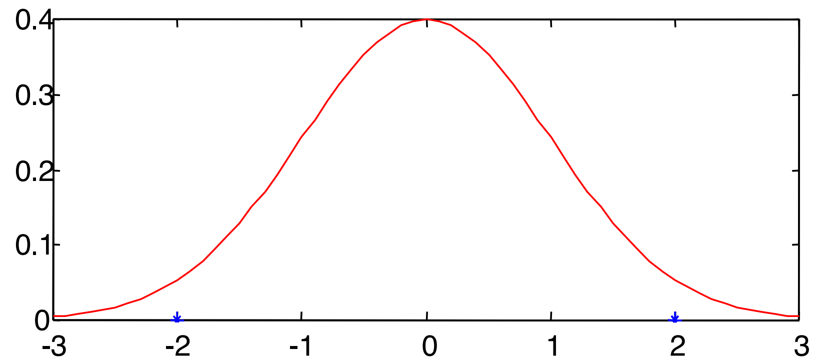
$$\int_{a}^{b} p(y)dy = \Pr(a \leq y \leq b)$$ By definition

So, for example the Pr(y<60)=0.361 in Figure 4

Central Limit Tendency for Averages

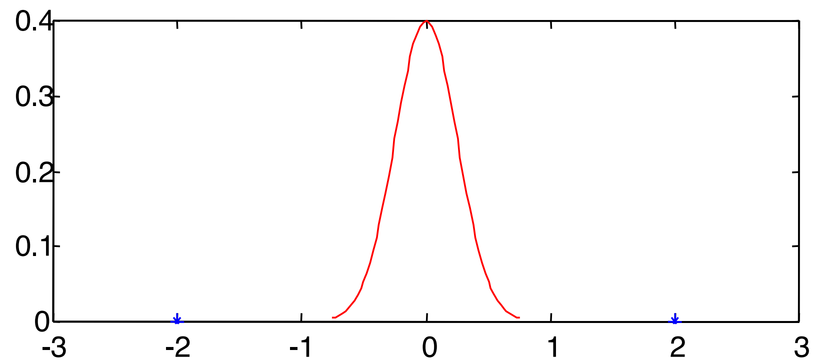


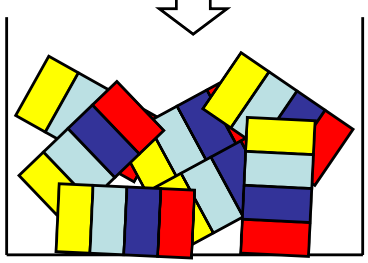
Parent Distribution



Take ‘n’ **random** samples. Find average. Repeat this procedure many times and make a distribution of these samples.

Sampled Distribution





- Sampled distribution is more nearly normal than parent distribution.
- Variance of randomly sampled distribution is smaller than parent distribution.

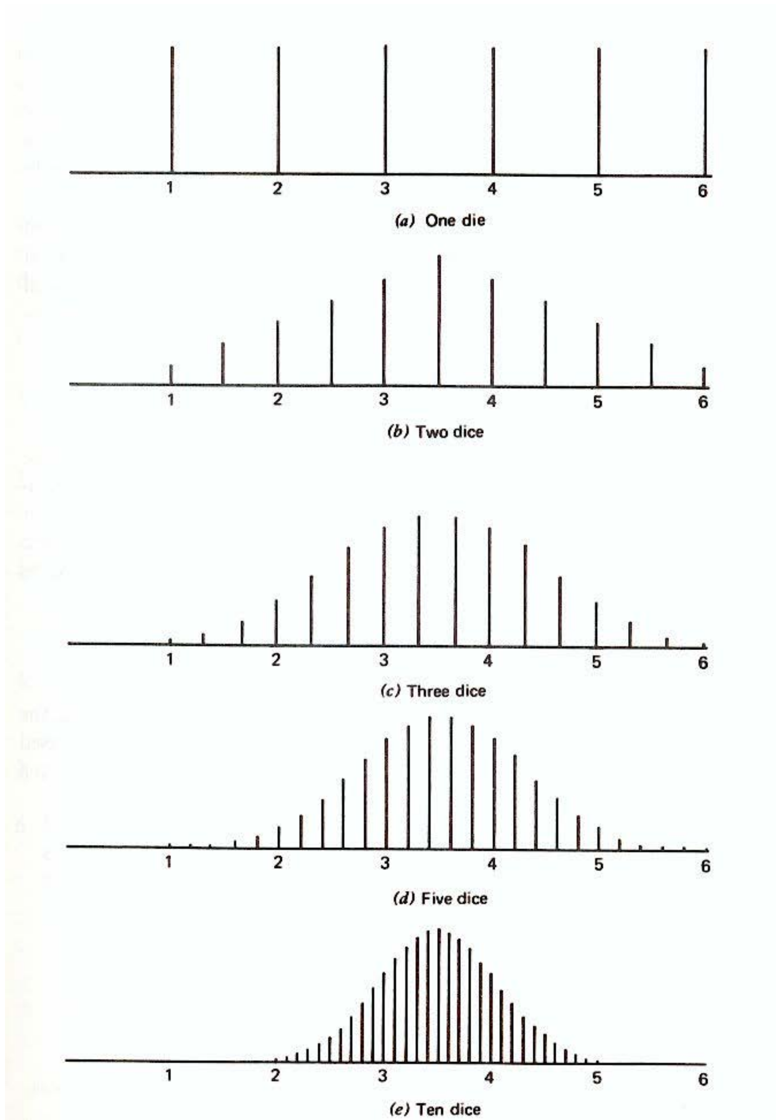
Central Limit Theorem (cont.)

- If the sampling is random (so that the errors are independent and uncorrelated), then we have the simple rule that $\bar{y}$ varies about the population mean η with variance $\frac{\sigma^2}{n}$. Thus

$$E(\bar{y}) = \mu, \qquad V(\bar{y}) = \frac{\sigma^2}{n}$$

Where n is the number of random samples collected at each time.

Central Limit Tendency for Averages: Dice Example



The Central Limit Theorem

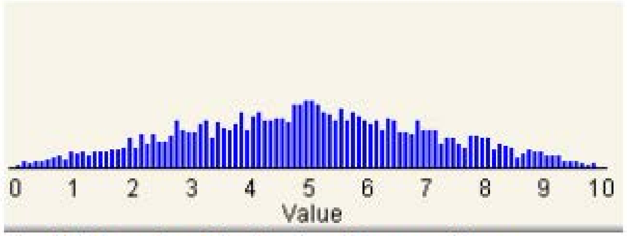


Fig. 1B) Sampling Distribution (from a uniform population) n = 2: number of samples = 4000; mean = 5.01; std dev 2.048; std error = 2.037

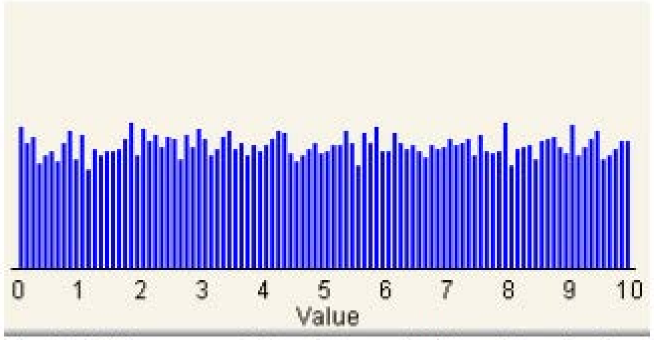


Fig. 1A) Histogram of Population - Uniform Distribution: all values in the population are randomly determined but all equally likely. This approximates a uniform distribution. Data points in population = 16,000; mean = 4.996 std dev 2.882

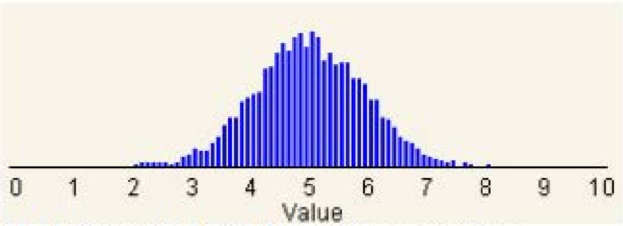


Fig. 1C) Sampling Distribution (from a uniform population) n = 10: number of samples = 2010; mean = 5.015; std dev 0.906; std error = 0.911

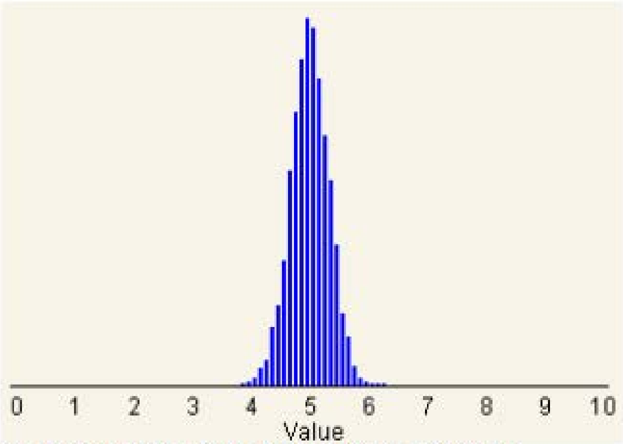


Fig. 1D) Sampling Distribution (from a uniform population) n = 80: number of samples = 4000; mean = 4.989; std dev 0.322; std error = 0.322

CLT-Applet

Central Limit Theorem Applet

The attached applet simulates a population by generating 16,000 floating point random numbers between 0 and 10. Each time the "New Population" button is pressed it generates a new set of random numbers. The plot labeled Population Distribution shows a histogram of the 16,000 data points.

Link: http://www.intuitor.com/statistics/CentralLim.html

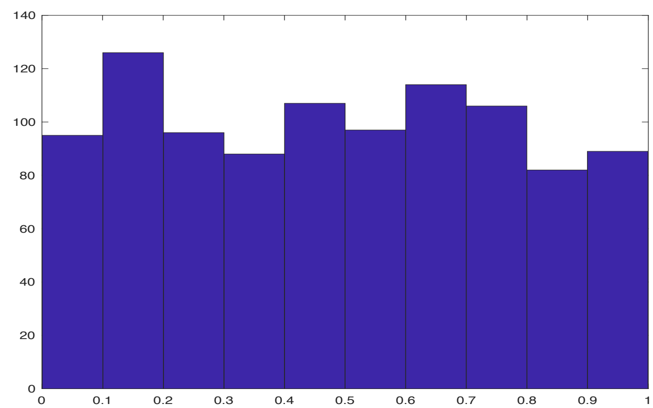
Use of ‘rand’ and ‘ceil’ in Matlab

>> help rand

'rand’ gives Uniformly distributed pseudorandom numbers in the range 0 to 1.

>> x=rand(1000,1); %generates a vector with1000 numbers

>> hist(x)



>> x(1:10)'

ans = 0.8147   0.9058   0.1270   0.9134   0.6324   0.0975   0.2785   0.5469   0.9575   0.9649

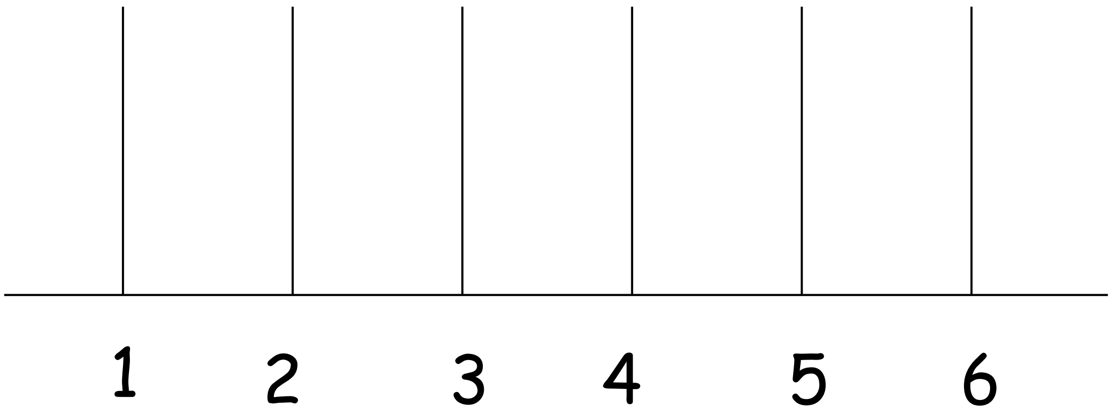
>> ceil(x(1:10)'*6)

ans = 5   6   1   6   4   1   2   4   6   6

% ceil(x) rounds the elements of x to the nearest integers towards infinity.

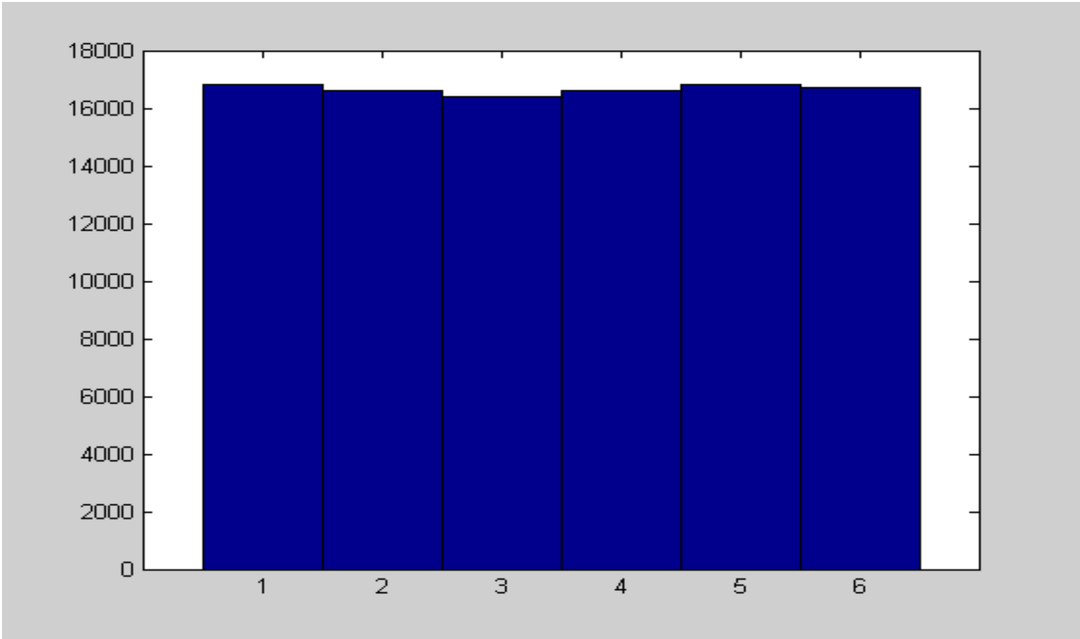
Example: Central Limit Tendency for Averages

1 Die:



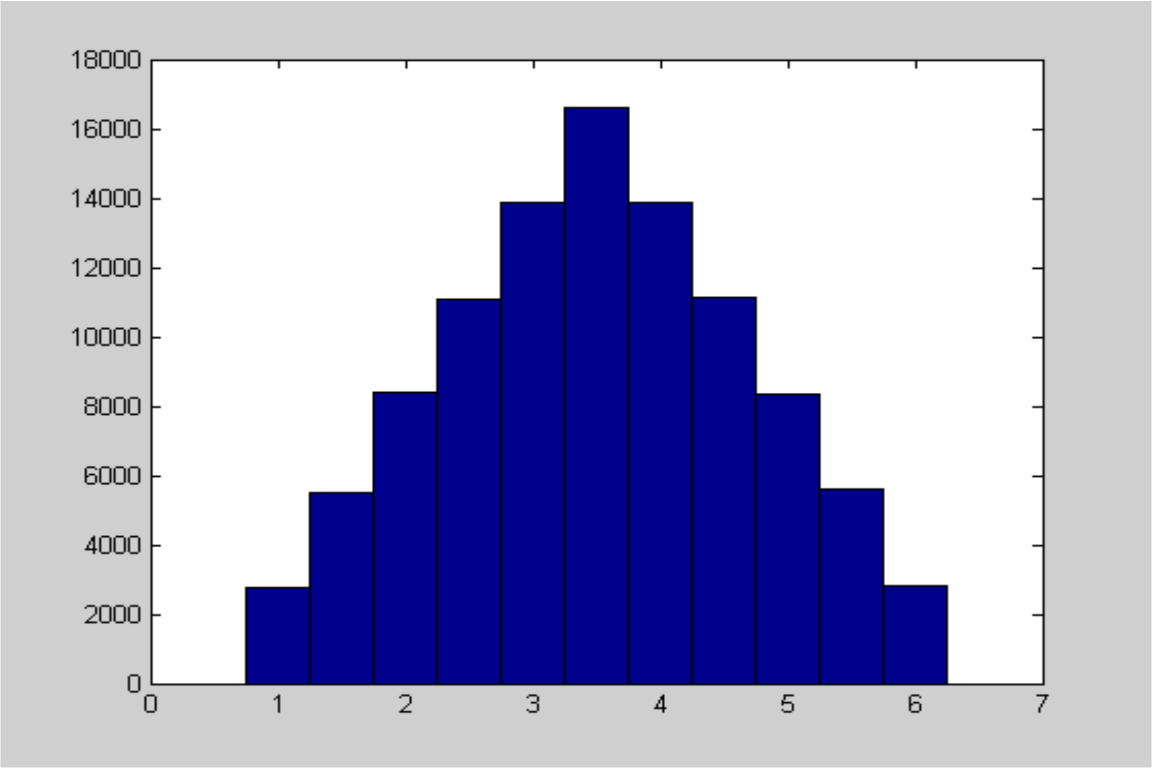
>> b=ceil(rand(100000,1)*6);

>> hist(b,1:1:6)



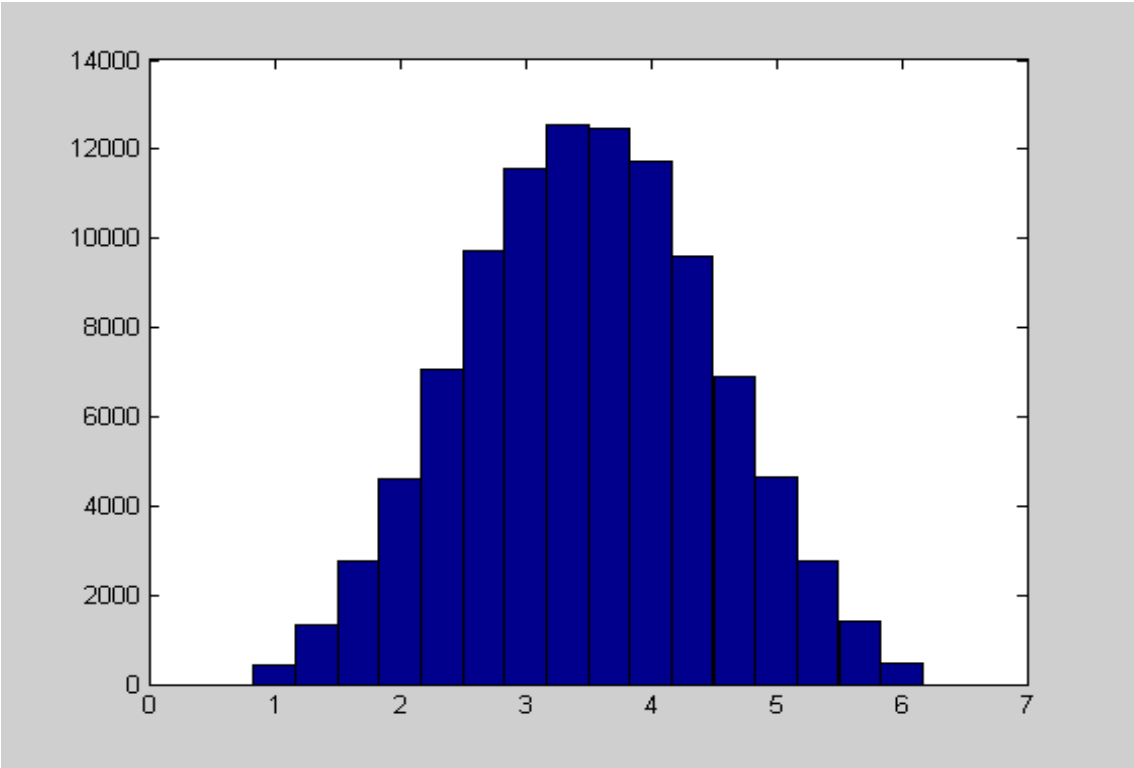
Two dice are thrown. Average is computed. Plot histogram

```
>> b=ceil(rand(100000,2)*6);
>> c=mean(b');
>> n=unique(c);
>> hist(c,n)
```



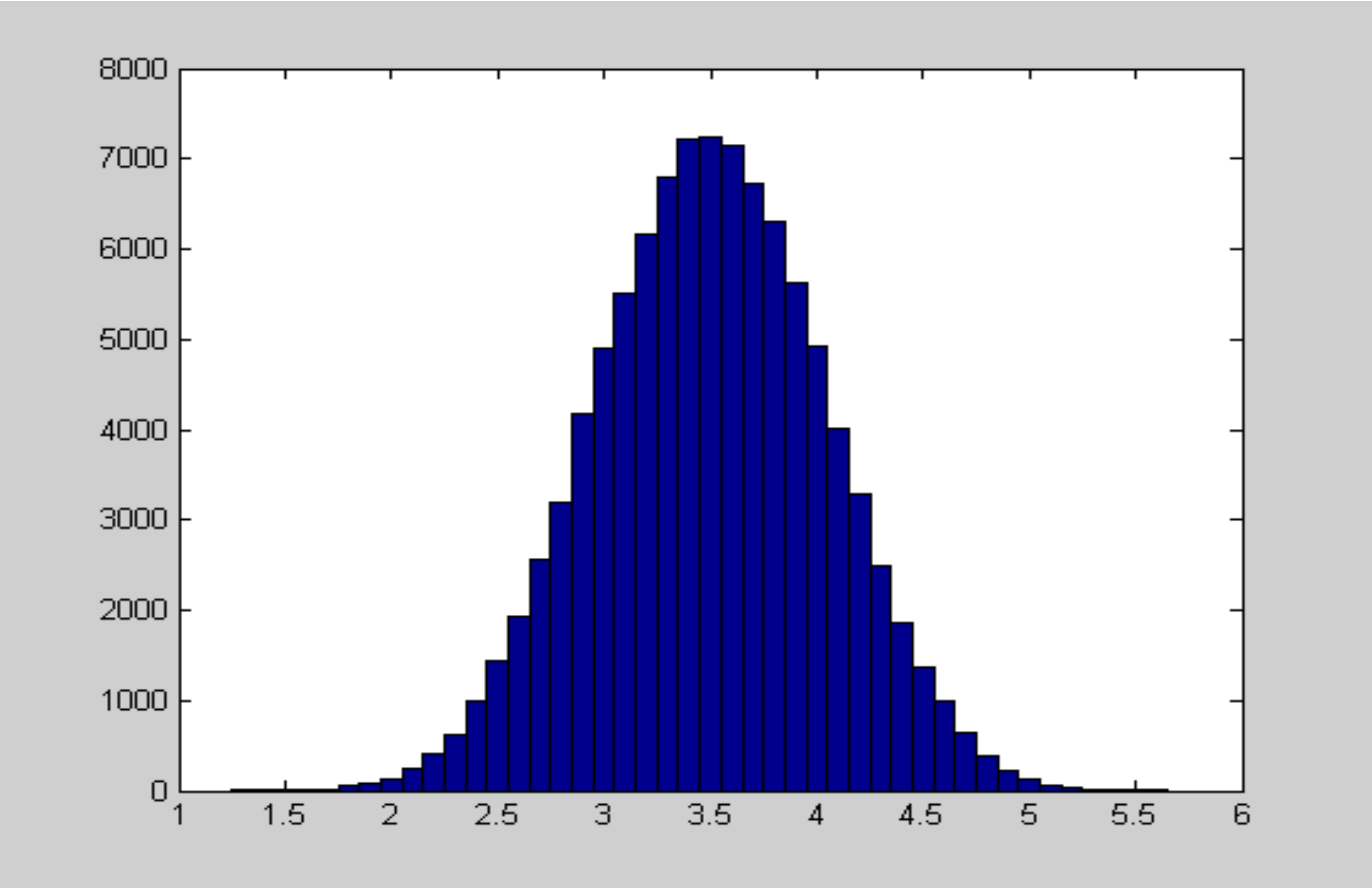
Three dice are thrown. Average is computed. Plot histogram

```
>> b=ceil(rand(100000,3)*6);

>> c=mean(b');

>> n=unique(c);

>> hist(c,n)
```



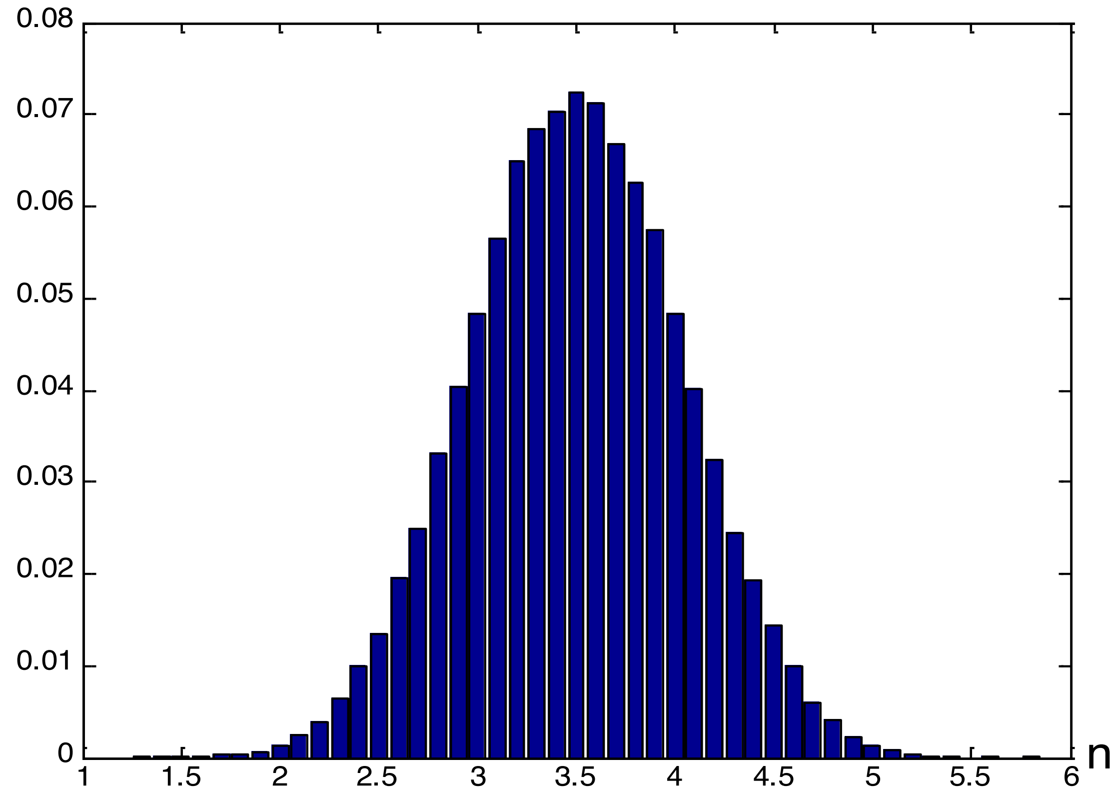
10 Dice Average

```
>> b=ceil(rand(100000,10)*6);

>> c=mean(b');  >> n=unique(c);   >> hist(c,n)
```



```
>> b=ceil(rand(100000,10)*6);
>> c=mean(b');
>> n=unique(c);
>> N0=histc(c,n);
>> N1=N0/10000;
>> bar(N1,n);
```

Even though the individual observations do not come from a normal distribution, the distribution of the average value tends to be normally distributed. (even if averages of five individual observations are considered)



The central limit theorem explains that no matter what distribution a RV follows, the sampled averages of this RV tend to be close to the normal distribution.

The Central Limit Theorem

Let X_1, X_2,...be the independently, identically distributed (i.i.d.) RVs having mean μ and finite variance σ^2. Then

$$\frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}} \sim \aleph(0,1)$$

Or

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \aleph(0,1)$$

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

The Central Limit Theorem

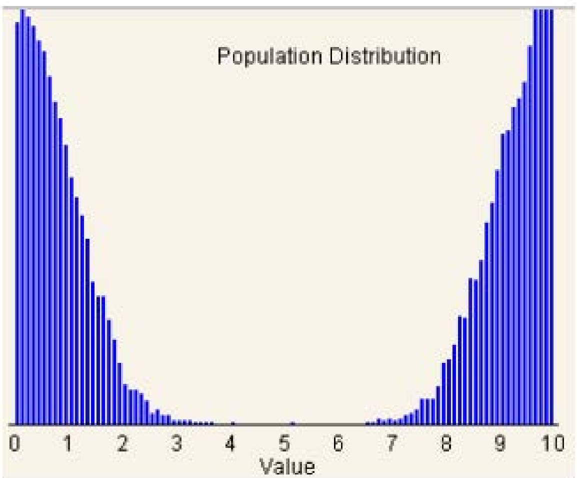


Fig. 2A) Histogram of Population - Bimodal Distribution: population = 16,000; mean = 5.002 std dev 4.242

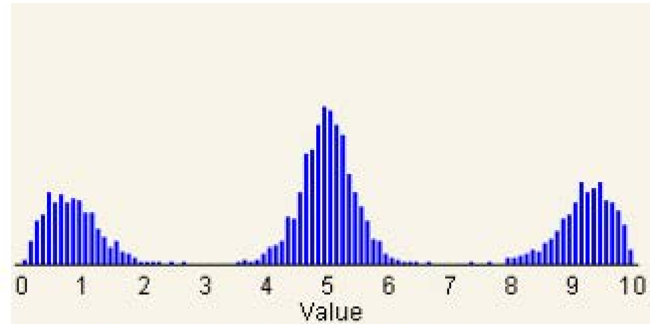


Fig. 2B) Sampling Distribution (from a bimodal population) n = 2: number of samples = 4000; mean = 4.977; std dev 3.017; std error = 2.999

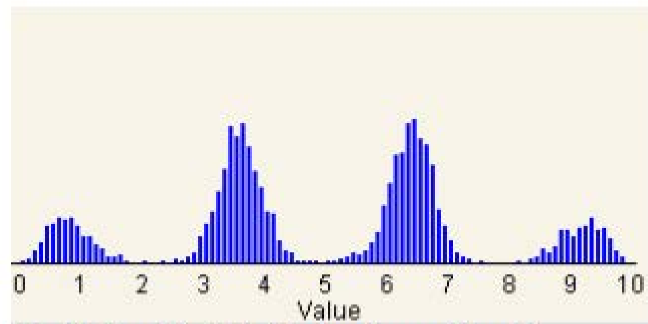


Fig. 2C) Sampling Distribution (from a bimodal population) n = 3: number of samples = 4000; mean = 4.946; std dev 2.425; std error = 2.449

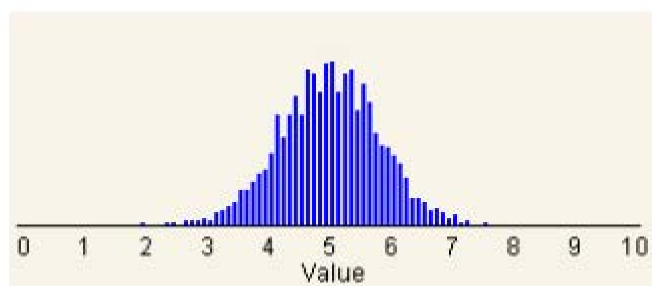
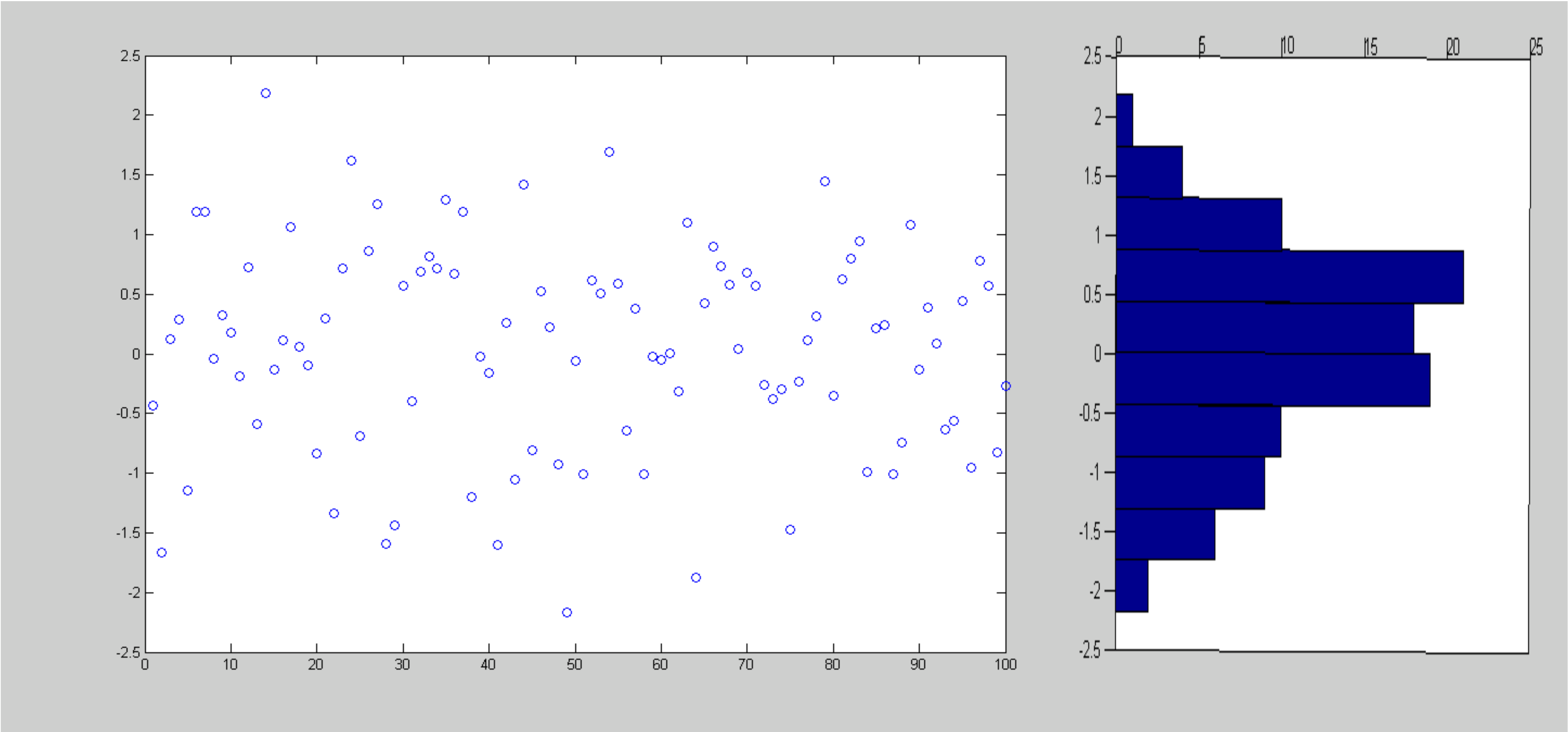


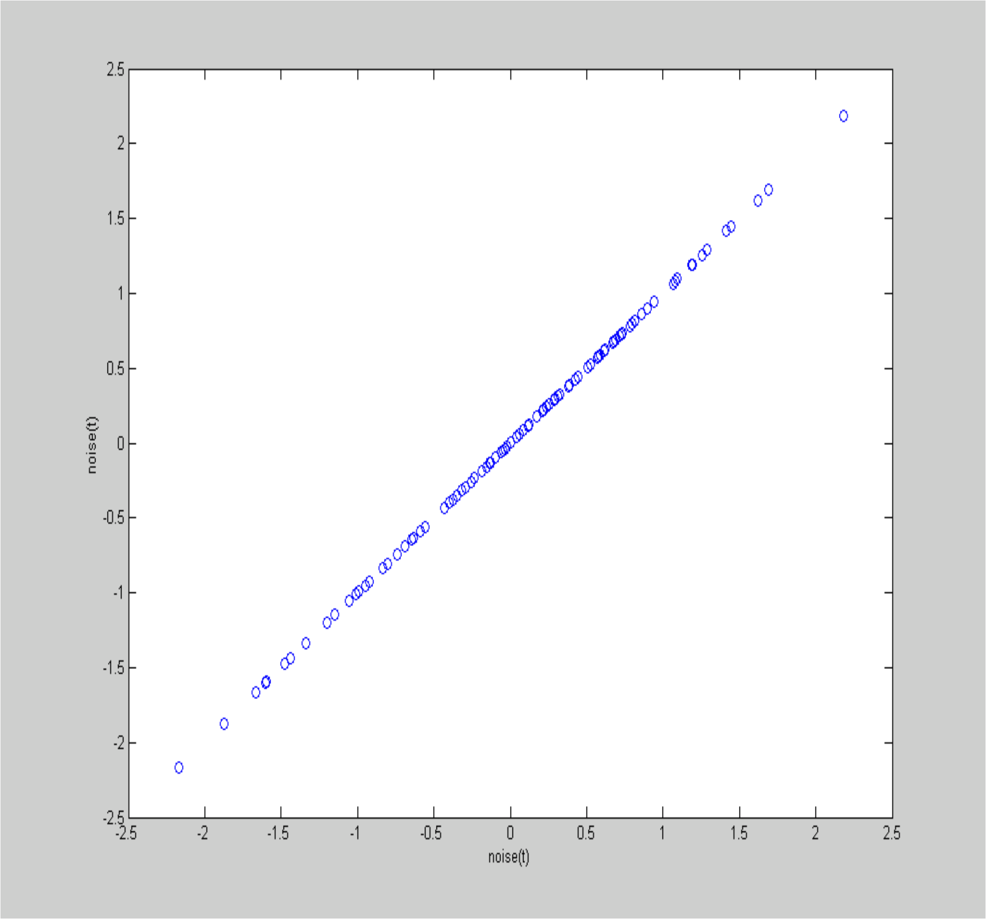
Fig. 2D) Sampling Distribution n = 30: number of samples = 4000; mean = 5.032; std dev 0.722; std error = 0.722

Time series and distribution of machine generated random noise

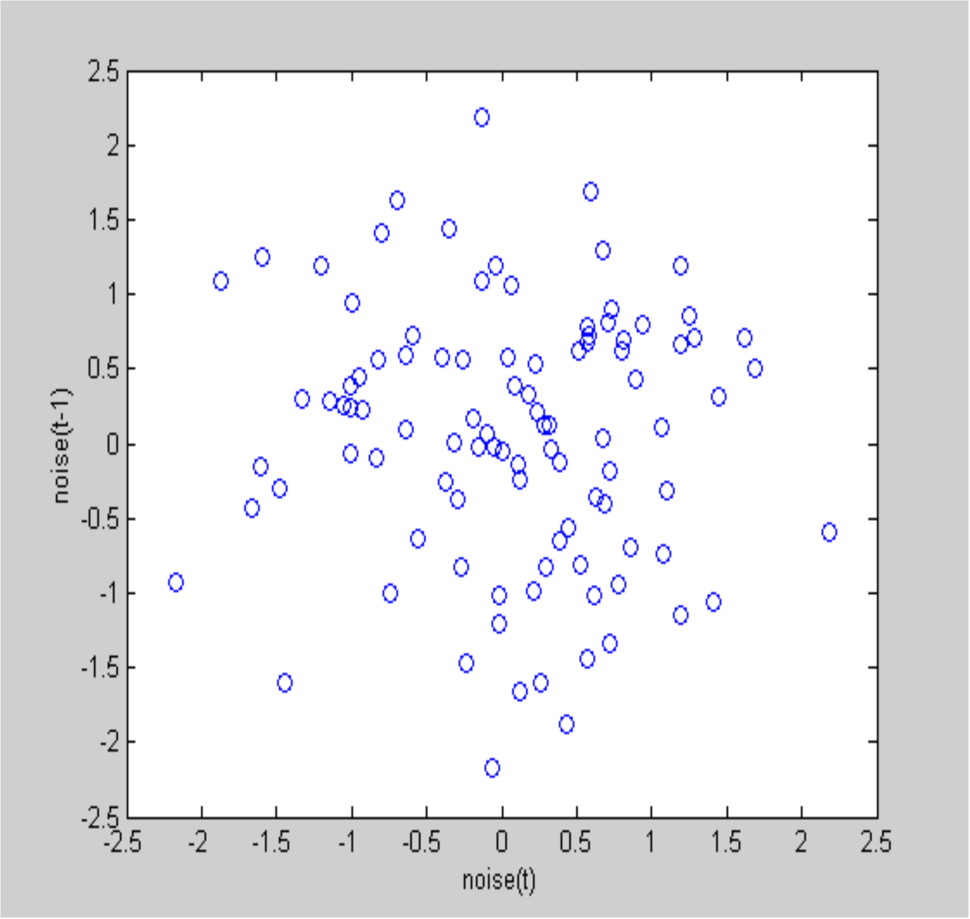


Serial Dependence of White noise with different lags

0 Lag!

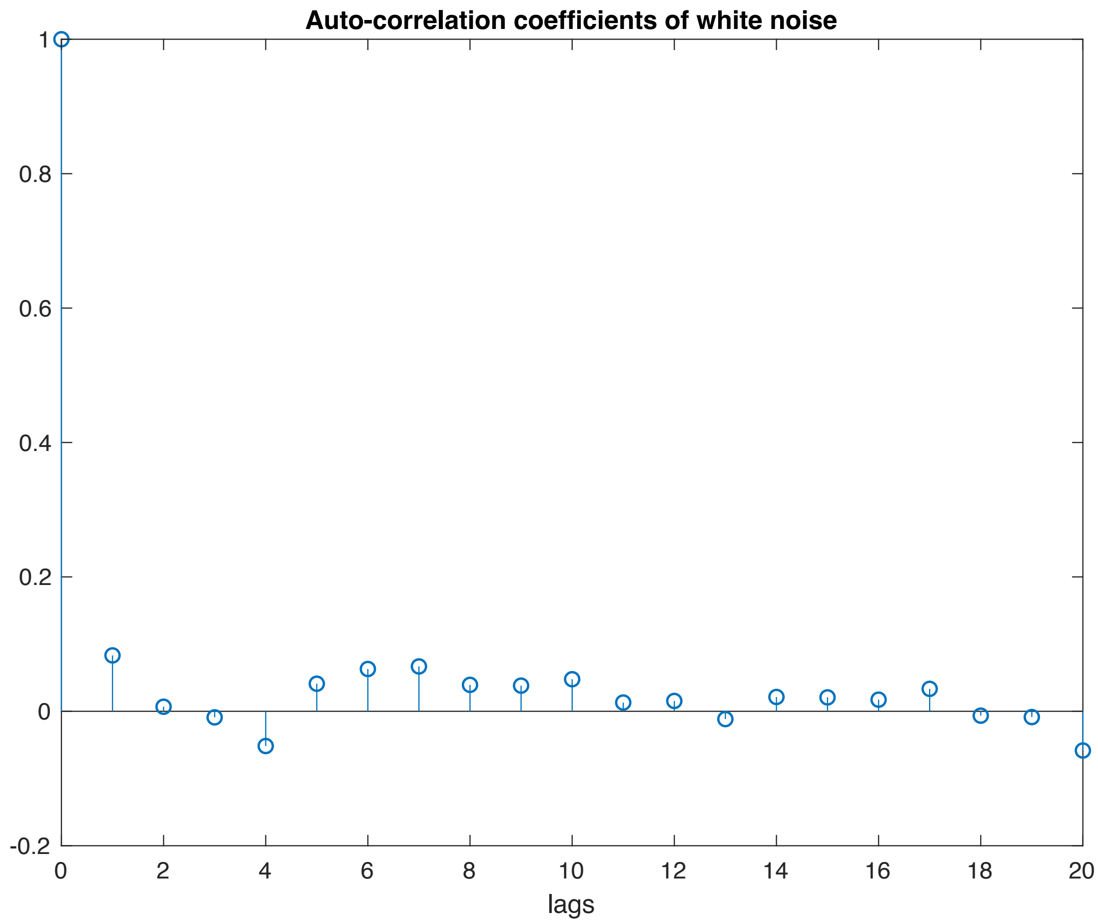


1 Lag



Autocorrelation of White noise

```
>>ww = randn(1000,1);
] >>[c,lags]= xcorr(ww,20,'coeff');
>>stem(0:20,c (21:41))
```



Covariance and correlation

- In probability theory and statistics, the mathematical descriptions of **covariance and correlation** are very similar. Both describe the degree of similarity between two random variables or sets of random variables.

Covariance: $$\gamma_{xy}(m) = E[(x_n - E(x))(y_{n+m} - E(y))]$$

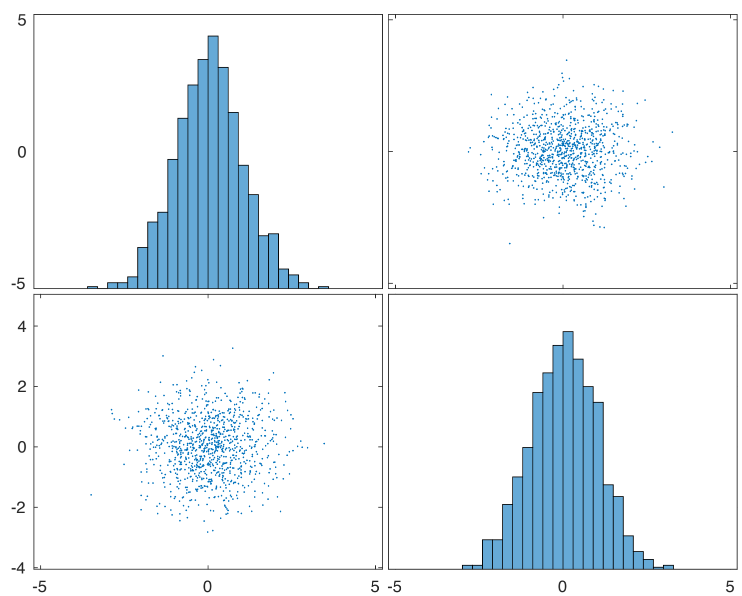
Correlation: $$\phi_{xy}(m) = \frac{E[(x_n - E(x))(y_{n+m} - E(y))]}{\sigma_x \sigma_y}$$

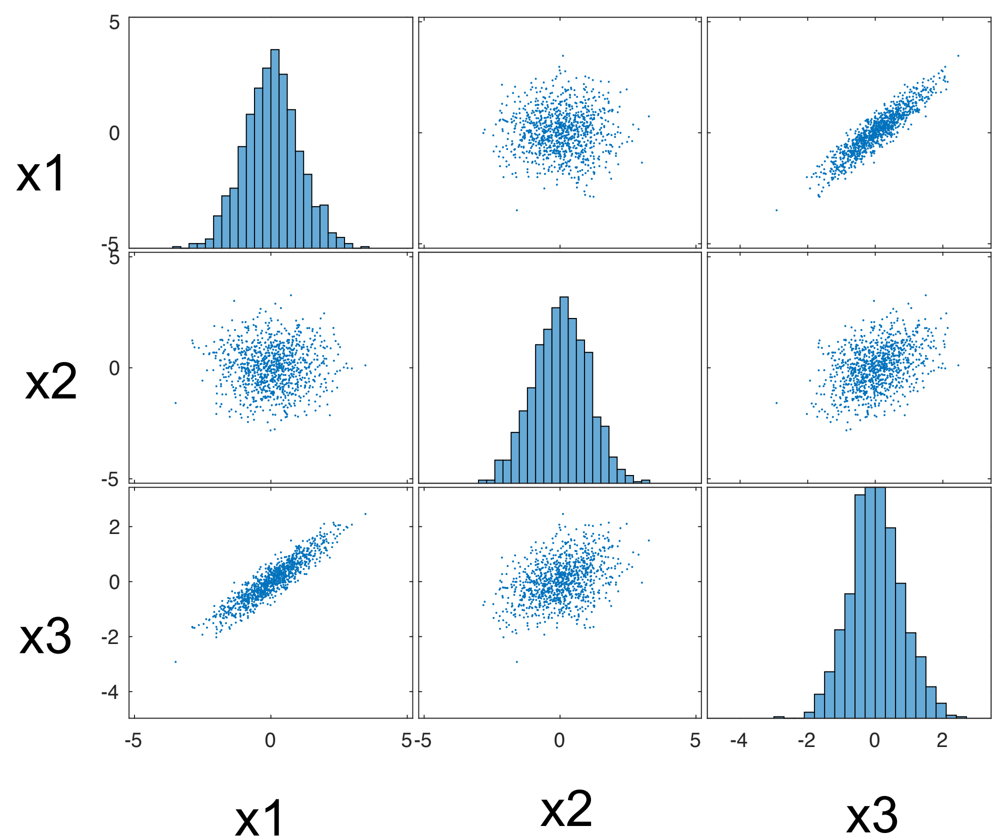
Auto-correlation: $$\phi_{xx}(m) = \frac{E[(x_n - E(x))(x_{n+m} - E(x))]}{\sigma_x^2}$$

Covariance and correlation

```
>> x=randn(2000,2);
>> plotmatrix(x)
```

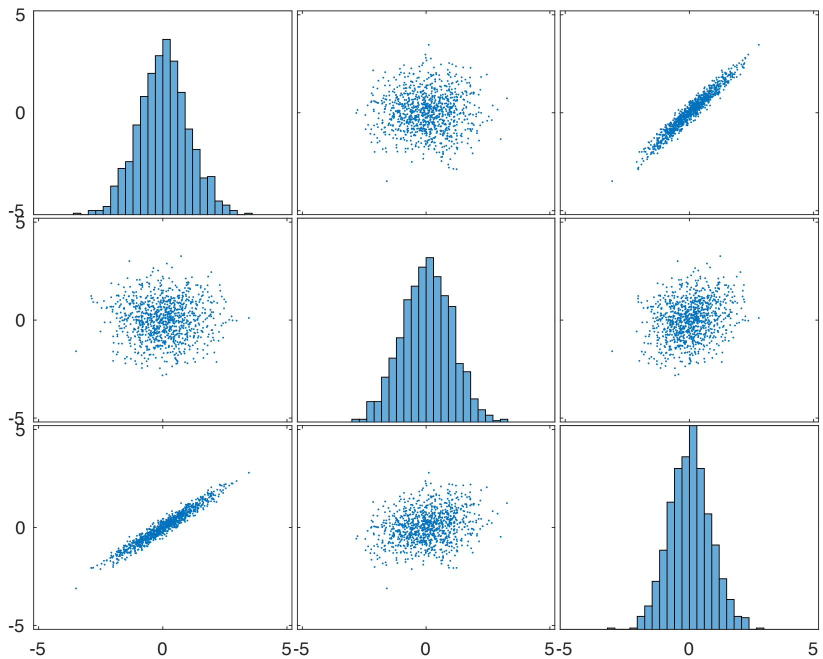
```
>> x3=x*[0.7; 0.3];
>> figure;
>> plotmatrix([x x3])
```



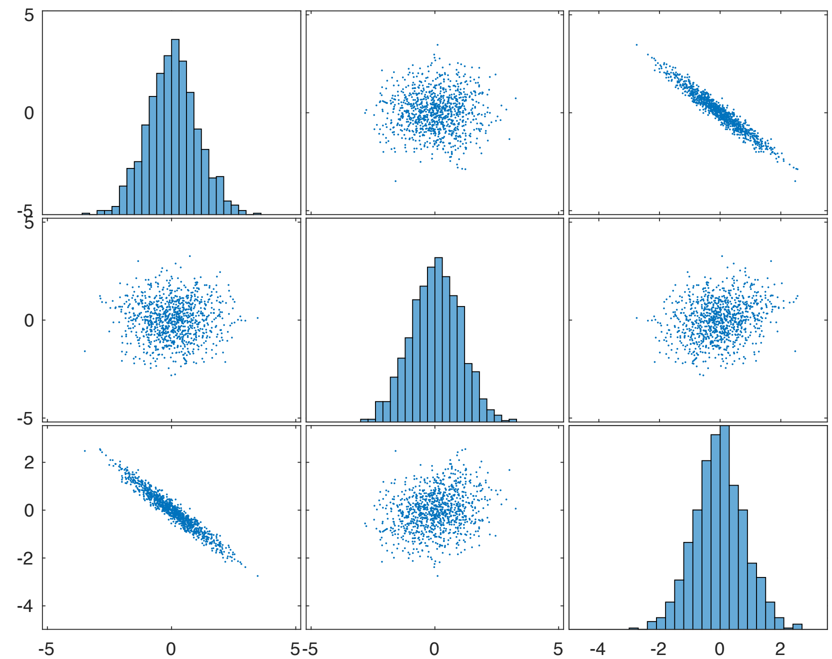


Covariance and Correlation

```
>> x3=x*[0.8; 0.2];
>> figure;
>> plotmatrix([x x3])
```



```
>> x3=x*[-0.8; 0.2];
>> figure;
>> plotmatrix([x x3])
```



Laws of Probability

- **Independent Events**
- Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other.
- In probability theory we say that two events, A and B, are independent if the probability that they both occur is equal to the product of the probabilities of the two individual events, i.e.

$$P(A \cap B) = P(A) \bullet P(B)$$

Probability of two independent events

Example

- Suppose that a man and a woman each have a pack of 52 playing cards. Each draws a card from his/her pack. Find the probability that they each draw the ace of clubs.
  - We define the events:
  - A = probability that man draws ace of clubs = 1/52
  - B = probability that woman draws ace of clubs = 1/52
  - Clearly events A and B are independent so:
  - $P(A \cap B) = P(A) \bullet P(B)$ = 1/52 . 1/52 = 0.00037
  - That is, there is a very small chance that the man and the woman will both draw the ace of clubs.

Mutually Exclusive Events

- Two events are mutually exclusive (or disjoint) if it is impossible for them to occur together.
- Formally, two events A and B are mutually exclusive if and only if $A \cap B = 0$
- If two events are mutually exclusive, they cannot be independent and vice versa.
- *Examples*
  - Experiment: Rolling a die once
  - Sample space S = {1,2,3,4,5,6}
  - Events A = 'observe an odd number' = {1,3,5}
  - B = 'observe an even number' = {2,4,6}
  - $A \cap B = 0$ = the empty set, so A and B are mutually exclusive.

Addition Rule

- The addition rule is a result used to determine the probability that event A or event B occurs or both occur.
- The result is often written as follows, using set notation:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

  - where:
  - P(A) = probability that event A occurs
  - P(B) = probability that event B occurs
  - $P(A \cup B)$= probability that event A or event B occurs
  - $P(A \cap B)$= probability that event A and event B both occur
  - For mutually exclusive events, that is events which cannot occur together:
  - $P(A \cap B)$ = 0
  - The addition rule therefore reduces to
  - $P(A \cup B)$ = P(A) + P(B)
  - For independent events, that is events which have no influence on each other
  - : $P(A \cap B) = P(A) \bullet P(B)$
  - The addition rule therefore reduces to

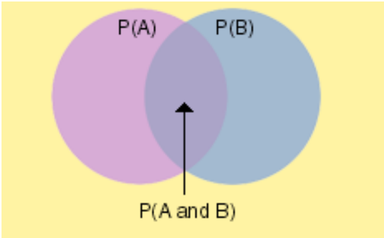
$$P(A \cup B) = P(A) + P(B) - P(A) \bullet P(B)$$

Addition Rule

- Example
  - Suppose we wish to find the probability of drawing either a king or a spade in a single draw from a pack of 52 playing cards.
  - We define the events A = 'draw a king' and B = 'draw a spade'
  - Since there are 4 kings in the pack and 13 spades, but 1 card is both a king and a spade, we have:
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ = 4/52 + 13/52 - 1/52 = 16/52
  - So, the probability of drawing either a king or a spade is 16/52 (= 4/13).

Conditional Probability

- In many situations, once more information becomes available, we are able to revise our estimates for the probability of further outcomes or events happening.
- The usual notation for "event A occurs given that event B has occurred" is "A | B" (A given B). The symbol | is a vertical line and does not imply division. P(A | B) denotes the probability that event A will occur given that event B has occurred already.
  - A rule that can be used to determine a conditional probability from unconditional probabilities is:



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

  - where:
  - P(A | B) = the (conditional) probability that event A will occur given that event B has occurred already
  - $P(A \cap B)$ = the (unconditional) probability that event A and event B both occur
  - P(B) = the (unconditional) probability that event B occurs

Conditional Probability

- Example
  - Given two dice, what is the probability that both will show '6'?

$$P(A \cap B) = P(A) \bullet P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

  - Now if the first die is thrown and shows a '6' what is the probability that the second die will show a '6'?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = {}^{1/36}\!\big/_{1/6} = \frac{1}{6}$$

Bayes' Theorem

- Bayes gave a special case involving continuous prior and posterior probability distributions and discrete probability distributions of data, but in its simplest setting involving only discrete distributions, Bayes' theorem relates the conditional and marginal probabilities of events *A* and *B*, where *B* has a non-vanishing probability:

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

- Each term in Bayes' theorem has a conventional name:
  - P(*A*) is the prior probability or marginal probability of *A*. It is "prior" in the sense that it does not take into account any information about *B*.
  - P(*A*|*B*) is the conditional probability of *A*, given *B*. It is also called the posterior probability because it is derived from or depends upon the specified value of *B*.
  - P(*B*|*A*) is the conditional probability of *B* given *A*. It is also called the likelihood.
  - P(*B*) is the prior or marginal probability of *B*, and acts as a normalizing constant.
  - Bayes' theorem in this form gives a mathematical representation of how the conditional probability of event *A* given *B* is related to the converse conditional probability of *B* given *A*.

Example to illustrate application of Bayes' Theorem

- 1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?
- P(woman has BC given that she had a + mammograph) ?

Example to illustrate application of Bayes' Theorem

- 100 out of 10,000 women at age forty who participate in routine screening have breast cancer. 80 of every 100 women with breast cancer will get a positive mammography. 950 out of 9,900 women without breast cancer will also get a positive mammography. If 10,000 women in this age group undergo a routine screening, about what fraction of women with positive mammographies will actually have breast cancer?
- P(woman has BC given that she had a + mammograph)=80/(80+950)=.078

Example to illustrate application of Bayes' Theorem

- P(woman has BC| knowing that she had a + mammograph) :

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{\dfrac{\overset{\text{likelihood}}{80}}{100} \times \dfrac{\overset{\text{prior}}{100}}{10000}}{\dfrac{(80+950)}{10000}} = \frac{80}{1030} = 0.078$$

Statistical inferencing

In a nutshell: The outcome of an analytics exercise is related to the concept of Statistical Inferencing

What is statistical inferencing?

Given prior information, a model, past and current data then how does one make the best decision in the presence of uncertainty (such as disturbances, model errors, sensor noise and other imprecise information).

Concluding remarks

- Statistical reasoning is paramount in deriving the correct inference.
- Be aware that the inference will nevertheless have a degree of uncertainty, usually expressed as confidence bounds.
- In many instances, it makes sense to use Bayes' rule since we do have prior and likelihood information.

Linear Algebra Concepts in Regression Methods

The Problem

$$\underline{\underline{A}}\,\underline{x} = \underline{b}$$

Solve for the unknown vector $\underline{x}$

In words: is it possible to have a linear combination of column vectors of *A* equal to *b* ?

i.e. $$\begin{bmatrix} \underline{a_1} & \underline{a_2} & \cdots\cdots & \underline{a_n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \underbrace{x_1\underline{a_1} + x_2\underline{a_2} + \cdots\cdots + x_n\underline{a_n}}_{\textit{linear combination of columns of } A} = \underline{b}$$

The Solution

An exact solution exists iff $b \in R(A)$

(b lies in the range space of A)

i.e. if b can be expressed as a linear combination of the columns of A, then an exact solution exists.

Corollary:

If A is of full rank then an exact solution exists.

The Solution (cont..)

$$\underline{\underline{A}}\,\underline{x} = \underline{b}$$

The Solution:

a) If $A \in R^{n \times n}$ and A is of full rank (i.e., $b \in R(A)$)

then:

$$\underline{\hat{x}} = \underline{\underline{A^{-1}}}b$$

To verify the solution, check if the RHS=LHS:

$$\underline{\underline{A}} \cdot \underline{\hat{x}} - b = \underline{\underline{A}} \cdot \underline{\underline{A^{-1}}} \cdot b - b = 0$$

i.e., a <u>UNIQUE</u> and an <u>EXACT</u> solution exists.

The Solution (cont..)

b) If $A \in R^{n \times m}$ and <u>A is of full rank</u> then an <u>exact</u> but <u>Non-unique</u> solution exists. If n <m i.e., an underspecified system with fewer equations than unknowns.

Example: $$\begin{bmatrix} 1 & 0 & 2\cdot 1 \\ 0 & 1 & 2\cdot 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$b \in R(A)$ since A is of full rank.

This yields many solutions- i.e., non-unique solutions but each one is exact.

The Solution (cont..)

e.g. $\underline{\hat{x}_1} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$ or $\underline{\hat{x}_2} = \begin{bmatrix} -0.1 \\ 0.9 \\ 1 \end{bmatrix}$

check if solutions are exact i.e.,

$\underline{\underline{A}}\underline{\hat{x}_1} - b = \begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \underline{0}$ and $\underline{\underline{A}}\underline{\hat{x}_2} - b = \begin{bmatrix} 2 \\ 3 \end{bmatrix} - \begin{bmatrix} 2 \\ 3 \end{bmatrix} = \underline{0}$

Exact but <u>non-unique</u> solutions.

The Solution (cont..)

c) $\underline{\underline{A}} \cdot \underline{x} = b$ with $A \in R^{n \times m}$ and n>m

i.e., over-specified set of equations or more equations than unknowns.

Solution: $\underline{\underline{A^T}} \cdot \underline{\underline{A}} \cdot \underline{x} = \underline{\underline{A^T}}\, \underline{b}$

Now $\left(A^T \cdot A\right) \in R^{m \times m}$ and is most likely invertible $\therefore \underline{\hat{x}} = \left(A^T A\right)^{-1} A^T b$

But solution is generally <u>not exact</u>, i.e., $\underline{\underline{A}}\, \underline{\hat{x}} \neq b$

However, $\left(\underline{\underline{A}}\, \underline{\hat{x}} - b\right)^T \left(\underline{\underline{A}}\, \underline{\hat{x}} - b\right)$ is smallest in the least square sense.

Note that in MATLAB $A \cdot x = B$ $\rightarrow$ Solution: x = A\B

This operation gives actual or generalized inverse.

F-test

***F*-test of equality of variances**

Source: Wikipedia, the free encyclopedia

In statistics, an ***F*-test of equality of variances** is a test for the null hypothesis that two normal populations have the same variance. Notionally, any *F*-test can be regarded as a comparison of two variances, but the specific case being discussed in this article is that of two populations, where the test statistic used is the ratio of two sample variances. This particular situation is of importance in mathematical statistics since it provides a basic exemplar case in which the F-distribution can be derived.[2] For application in applied statistics, there is concern that the test is so sensitive to the assumption of normality that it would be inadvisable to use it as a routine test for the equality of variances. In other words, this is a case where "approximate normality" (which in similar contexts would often be justified using the central limit theorem), is not good enough to make the test procedure approximately valid to an acceptable degree.

F Distribution

Suppose that a sample of n_1 observations is randomly drawn from a normal distribution having variance $\sigma^2{}_1$, a second sample of n_2 observations drawn from a second normal distribution having variance $\sigma^2{}_2$. Then what can we say about $s^2{}_1/\ s^2{}_{2:}$

Ans:

where ν_1=n_1-1, and ν_2=n_2-1, are the degrees of freedom.

$$\frac{s_1^2}{s_2^2} \sim \frac{\sigma_1^2}{\sigma_2^2} F_{\nu_1,\nu_2}$$

F-test….contd.

Let $X_1, ..., X_n$ and $Y_1, ..., Y_m$ be independent and identically distributed samples from two populations which each have a normal distribution. The expected values for the two populations can be different, and the hypothesis to be tested is that the variances are equal. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

be the sample means. Let

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \text{ and } S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2$$

be the sample variances. Then the test statistic

$$F = \frac{S_X^2}{S_Y^2}$$

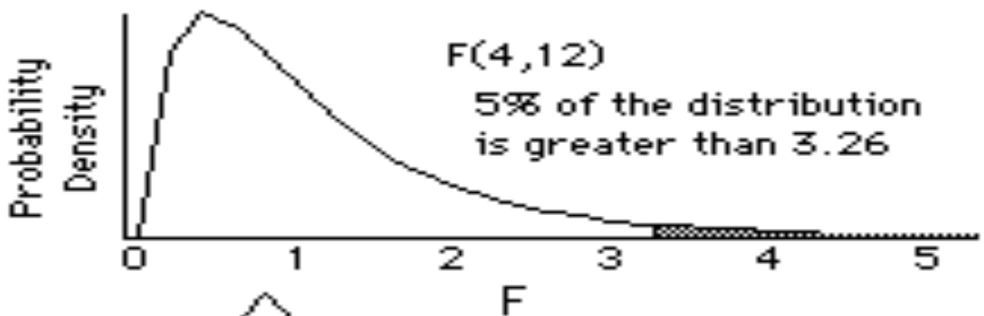
has an F-distribution with $n - 1$ and $m - 1$ degrees of freedom if the null hypothesis of equality of variances is true. Otherwise it follows an F-distribution scaled by the ratio of true variances. The null hypothesis is rejected if F is either too large or too small.

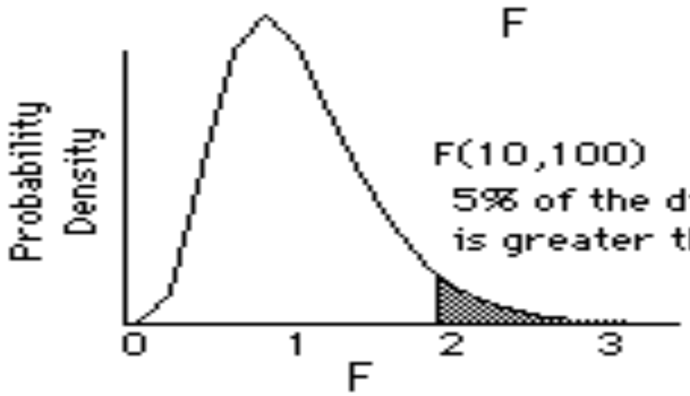
Source: Wikipedia, the free encyclopedia

F Distribution

Let $\chi^2(m)$ and $\chi^2(n)$ be independent variables distributed as chi-squared with *m* DOF and *n* DOF.

The ratio $\dfrac{\chi^2(n)/n}{\chi^2(m)/m}$ is distributed as a F-distribution over the domain [0, ∞) with [n, m] DOF.





finv(0.95,4,12) = 3.2592

finv(0.95,10,100) = 1.9267

cdf(‘f’ ,3.26,4,12) = 0.95

cdf(‘f’ ,1.93,10,100) = 0.95