Statistics Part II — Basic Theory

Joe Nahas

University of Notre Dame

UNIVERSITY OF
NOTRE DAME

Department of Applied and Computational Mathematics and Statistics (ACMS)

- **ACMS courses that may be useful**
  - **ACMS 30440. Probability and Statistics**
    - **An introduction to the theory of probability and statistics, with applications to the computer sciences and engineering**
  - **ACMS 30600. Statistical Methods & Data Analysis I**
    - **Introduction to statistical methods with an emphasis on analysis of data**

Population versus Sample

- **Consider ND Freshman SAT Scores:**
  - **Well defined population**
  - **We could obtain all the 2023 Freshman records and determine the statistics for the full population.**
    - μ **– the Mean SAT score**
    - σ **- the Standard Deviation of the SAT scores**
  - **We could obtain a sample of say 100 Freshman records and determine estimates for the statistics.**
    - **m or** $\bar{x}$ **– Estimated Mean or Average SAT score in the sample**
    - **s – the Estimate of the Standard Deviation**

Notation

Measure	Population Greek Letters		Sample Roman Letters	
Location	Mean	μ	Estimate of the Mean, Average	m $\overline{x}$
Spread	Variance	σ^2	Sample Variance	s^2
	Standard Deviation	σ	Sample Standard Deviation	s
Correlation	Correlation Coefficient	ρ	Sample Correlation Coefficient	r

Statistic Outline

1. Background:
   - A. Why Study Statistics and Statistical Experimental Design?
   - B. References
2. Basic Statistical Theory
   - A. Basic Statistical Definitions
     - i. Distributions
     - ii. Statistical Measures
     - iii. Independence/Dependence
       - a. Correlation Coefficient
       - b. Correlation Coefficient and Variance
       - c. Correlation Example
   - B. Basic Distributions
     - i. Discrete vs. Continuous Distributions
     - ii. Binomial Distribution
     - iii. Normal Distribution
     - iv. The Central Limit Theorem
       - a. Definition
       - b. Dice as an example

Statistic Outline (cont.)

3. Graphical Display of Data
   - A. Histogram
   - B. Box Plot
   - C. Normal Probability Plot
   - D. Scatter Plot
   - E. MatLab Plotting
4. Confidence Limits and Hypothesis Testing
   - A. Student's t Distribution
     - i. Who is "Student"
     - ii. Definitions
   - B. Confidence Limits for the Mean
   - C. Equivalence of two Means

Statistic Outline

1. Background:
   - A. Why Study Statistics and Statistical Experimental Design?
   - B. References
2. Basic Statistical Theory
   - A. Basic Statistical Definitions
     - i. Distributions
     - ii. Statistical Measures
     - iii. Independence/Dependence
       - a. Correlation Coefficient
       - b. Correlation Example
       - c. Correlation Coefficient and Variance
   - B. Basic Distributions
     - i. Discrete vs. Continuous Distributions
     - ii. Binomial Distribution
     - iii. Normal Distribution
     - iv. The Central Limit Theorem
       - a. Definition
       - b. Dice as an example

Basic Statistical Definitions

- **Distribution:**
  - The pattern of variation of a variable.
  - It records or defines the numerical values of the variable and how often the value occurs.
  - A distribution can be described by shape, center, and spread.
- **Variable – x:**
  - A characteristic that can assume different values.
- **Random Variable:**
  - A function that assigns a numerical value to each possible outcome of some operation/event.
- **Population:**
  - The total aggregate of observations that conceptually might occur as a result of performing a particular operation in a particular way.
  - The universes of values.
  - Finite or Infinite

P. Nahas

Basic Definitions (cont.)

- **Sample:**
  - A collection of some of the outcomes, observations, values from the population.
  - A subset of the population.
- **Random Sample:**
  - Each member of the population has a equal chance of being chosen as a member of the sample.
- **Bias:**
  - The tendency to favor the selections of units with certain characteristics.
- **Active Data Collection:**
  - Planned data collection with specific goals in mind to maximize the information.
- **Passive Data Collection:**
  - Data that just comes our way – that "just is."
  - We do not always know how it was obtained.

P. Nahas

Basic Definitions (cont.)

- **Measurement:**
  - **The assignment of numerals to objects and events according to rules.**
- **Data:**
  - **Can be numerical or textual in form.**
  - **Can be sources of information.**

In God we trust, all others bring data!

Stu Hunter

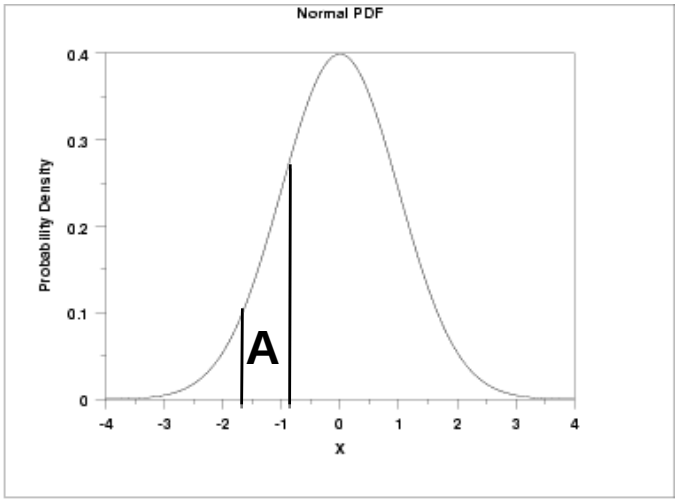
P. Nahas

Probability Distribution Function

- **Probability Distribution Functions:**
  - **Described by the probability density function f(x) where**

$$\int_R f(x)dx = 1$$

$$f(x) \geq 0, x \in R$$



where R is the range of x.

$$P(x \in A) = \int_A f(x)dx$$

$$P(x = b) = \int_b f(s)dx = 0$$

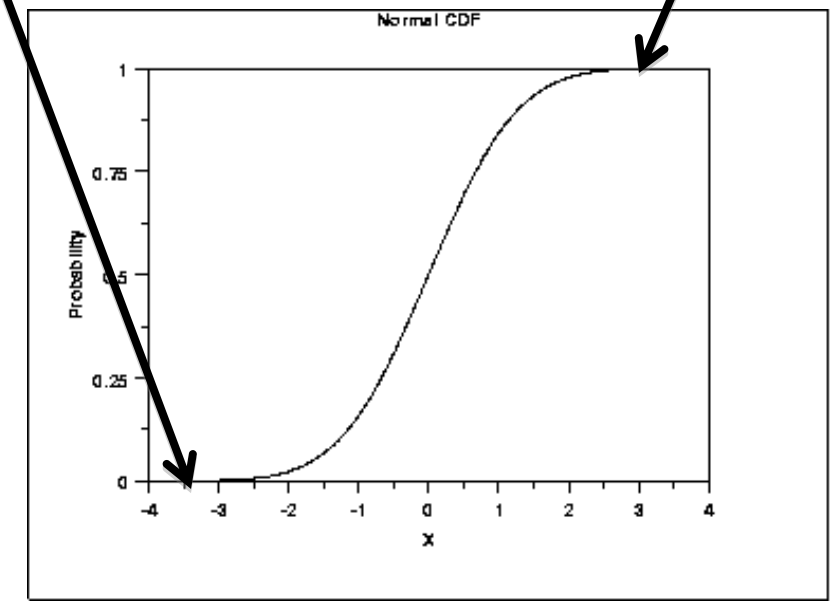
P. Nahas

Cumulative Distribution Function

- **Cumulative Distribution Function (CDF) F(x) where**

$$f(x) = F'(x)$$

$$\lim_{x \to -\infty} F(x) = 0 \qquad \lim_{x \to +\infty} F(x) = 1$$



P. Nahas

A measure of Location: Mean

- **Mean**
  - **Also known as The First Moment.**

$$\mu = E(X) = \int_{x \in R} x f(x) dx \qquad \mu = E(X) = \sum_{x \in R} x f(x)$$

  - **The mean of a sum = the sum of the means.**

$$\mu_{x+y} = \mu_x + \mu_y$$

  - **The mean is a linear function.**

$$\mu_{a+bx} = a + b\mu_x$$

  - **The estimate of the mean (sample average):**

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

where n is the number of items in the sample.

P. Nahas

Other Measures of Location

- **Median**
  - **The midpoint of the distribution.**
    - **There are as many point above and below the median.**
- **Mode**
  - **The peak value of the distribution.**

Measures of Spread

- **Variance**
  - **Also known as The Second Moment.**

$$\sigma^2 = E[(X-\mu)^2] = \int_{x \in f} (x-\mu)^2 f(x) dx$$

  - **If and only if x and y are <u>independent</u>:**

$$\sigma^2_{x+y} = \sigma^2_{x-y} = \sigma^2_x + \sigma^2_y$$

  - **s^2 is the variance <u>estimate</u> (<u>sample</u> variance):**

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

where n is the number of items in the sample.

- **Standard Deviation**
  - **σ is the standard deviation.**
  - **s is the standard deviation <u>estimate</u> (<u>sample</u> standard deviation).**

P. Nahas

Correlation

Let Y and X_i, i= 1 to n be random variables, and

$$Y = \sum_{i=1}^{n} a_i X_i$$

where μ_i is the mean of X_i and σ_i^2 is the variance of X_i

then the variance of y= σ_y^2

$$= E[(Y - \mu_y)^2]$$

$$= E[\sum_{i=1}^{n} (a_i x_i - a_i \mu_i)^2]$$

$$= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i<j} \sum_{j} a_i a_j \rho_{ij} \sigma_i \sigma_j$$

where ρ_{ij} is the correlation coefficient for the population $X_i X_j$.

P. Nahas

The Correlation Coefficient

The correlation coefficient ρ, is a statistical moment that gives a measure of *linear dependence* between two random variables. It is *estimated* by:

$$r = \frac{s_{xy}}{s_x s_y}$$

where s_x and s_y are the square roots of the estimates of the variance of x and y, while s_{xy} is an estimate of the *covariance* of the two variables and is estimated by:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{n-1}$$

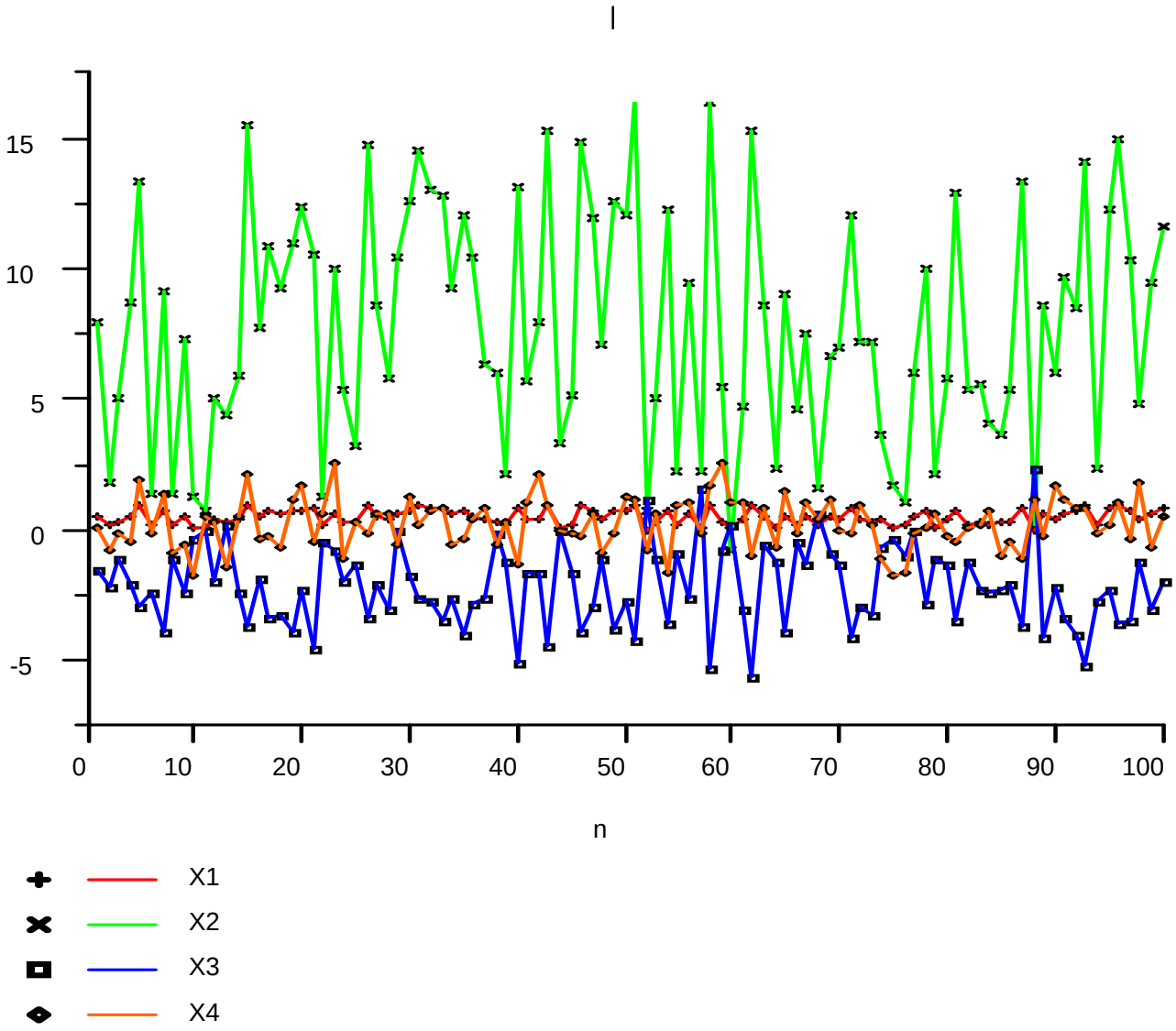
Poolla & Spanos

Correlation

- If ρ = 1, two reandom variables are correlated.
- If ρ = 0, two random variables are not correlated.
- If ρ = -1, two random variables are inversely correlated.

- Example:
  - The height and weight of a sample of the population of people.
    - You would expect a positive correlation.

Correlation Example



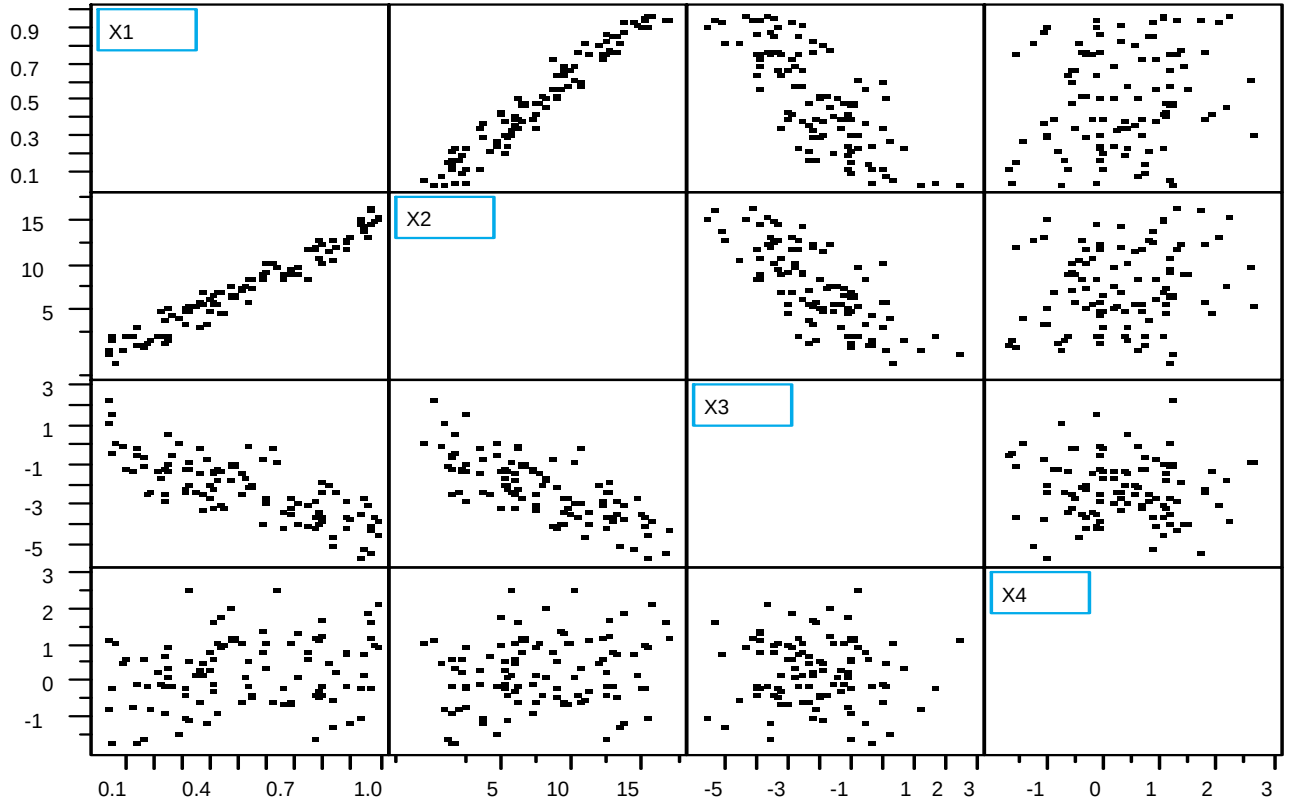
Poolla & Spanos

Correlation Example

Correlations

Variable	X1	X2	X3	X4
X1	1.0000	0.9719	-0.7728	0.2089
X2	0.9719	1.0000	-0.7518	0.2061
X3	-0.7728	-0.7518	1.0000	-0.0753
X4	0.2089	0.2061	-0.0753	1.0000

Scatter Plot Matrix



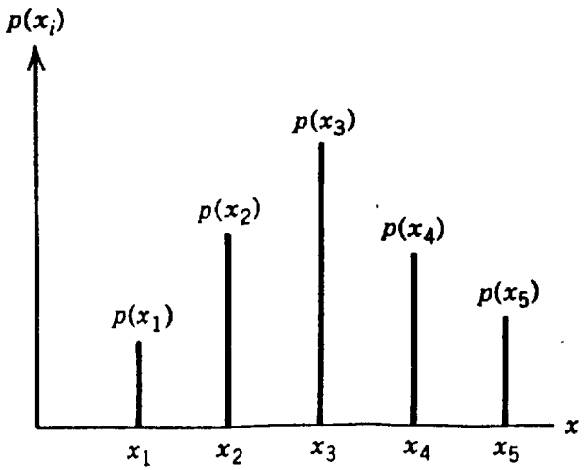
Poolla & Spanos

Statistic Outline

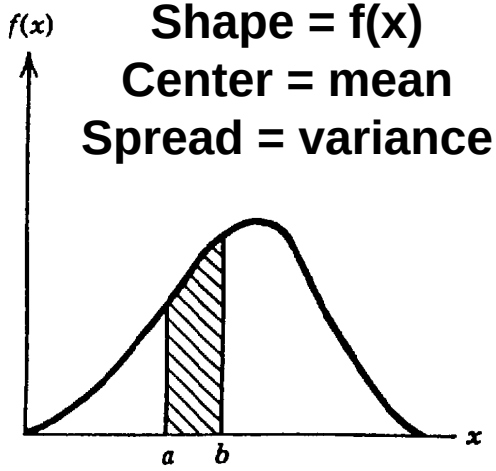
1. Background:
   - A. Why Study Statistics and Statistical Experimental Design?
   - B. References
2. Basic Statistical Theory
   - A. Basic Statistical Definitions
     - i. Distributions
     - ii. Statistical Measures
     - iii. Independence/Dependence
       - a. Correlation Coefficient
       - b. Correlation Example
       - c. Correlation Coefficient and Variance
   - B. Basic Distributions
     - i. Discrete vs. Continuous Distributions
     - ii. Binomial Distribution
     - iii. Normal Distribution
     - iv. The Central Limit Theorem
       - a. Definition
       - b. Dice as an example

The Distribution of x

Discrete Distributions



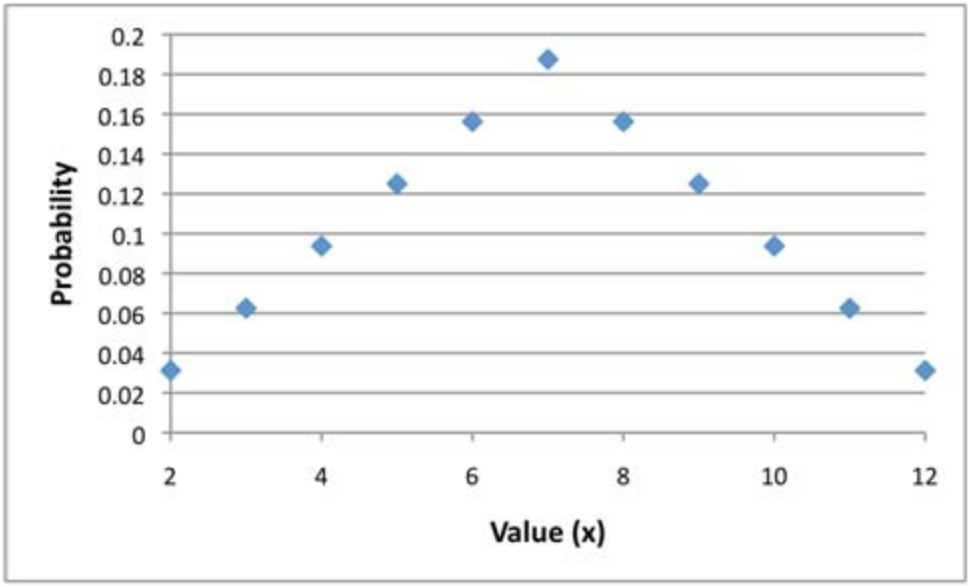
Continuous Distributions



- Can calculate the probability of a randomly chosen observation of the population falling within a given range – so it is a probability distribution.
- Vertical ordinate, P(x) is called the probability density
- Can we find a mathematical function to describe the probability distribution?

Poolla & Spanos

A Discrete Distribution



What discrete distribution is this?

A Discrete Distribution: the Binomial

- **The binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial.**
  - **These outcomes are appropriately labeled "success" and "failure".**
- **The binomial distribution is used to obtain the probability of observing x successes in n trials, with the probability of success on a single trial denoted by p.**
  - **The binomial distribution assumes that p is fixed for all trials.**

$$P(x, p, n) = \binom{n}{x} (p)^x (1-p)^{(n-x)} \qquad \text{for } x = 0, 1, 2, \cdots, n$$

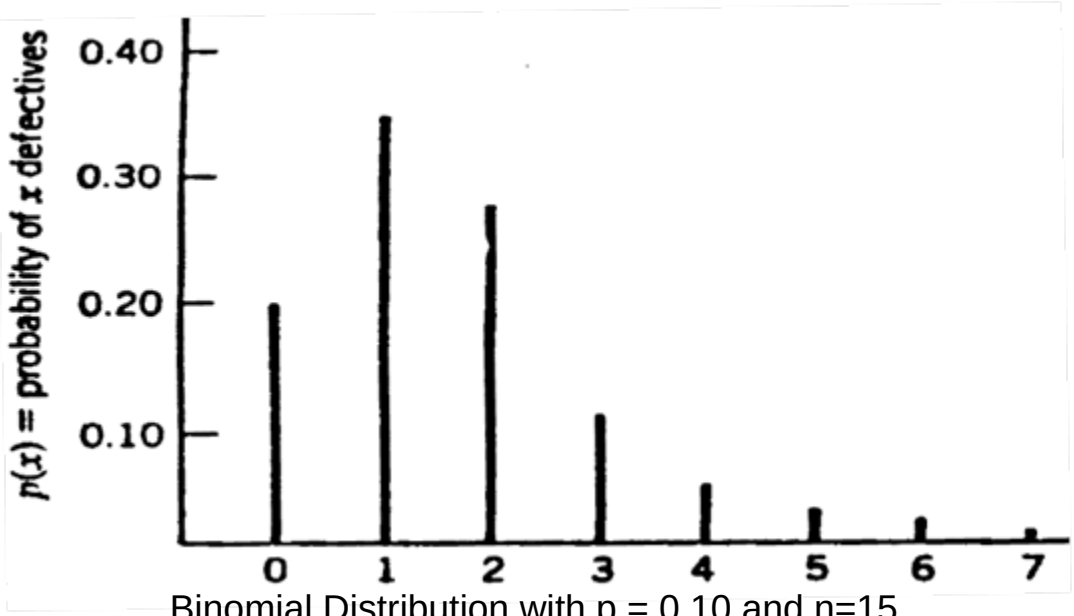
where:

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

Mean = np

Standard Deviation = $\sqrt{np(1-p)}$

A Discrete Distribution: the Binomial



Binomial Distribution with p = 0.10 and n=15.

- **Simple Examples:**
  - **Number of heads in 10 coin flips: p = 0.5, n = 10**
  - **Number of ones in 5 rolls of a die: p = 1/6, n = 5**

$$P(x,p,n) = \binom{n}{x} (p)^x (1-p)^{(n-x)} \qquad \text{for } x = 0, 1, 2, \cdots, n$$

Example: Using the Binomial Distribution

- The probability that a memory cell fails is 10^{-9}.
- In a 64 Mbit memory array what is the probability that:
  - All cells are OK?
  - 1 cell failed?
  - 5 cells failed?
  - More than 5 cells failed?

Binomial Solution

64 Mb Memory = n = 2^{26} = 6.7E+7

$$P(X=0)=\binom{n}{x}p^x(1-p)^{n-x}$$

$$=\binom{n}{0}(10^{-9})^0(1-(10^{-9}))^{n-0}$$

$$=1\bullet 1\bullet(1-(10^{-9}))^n$$

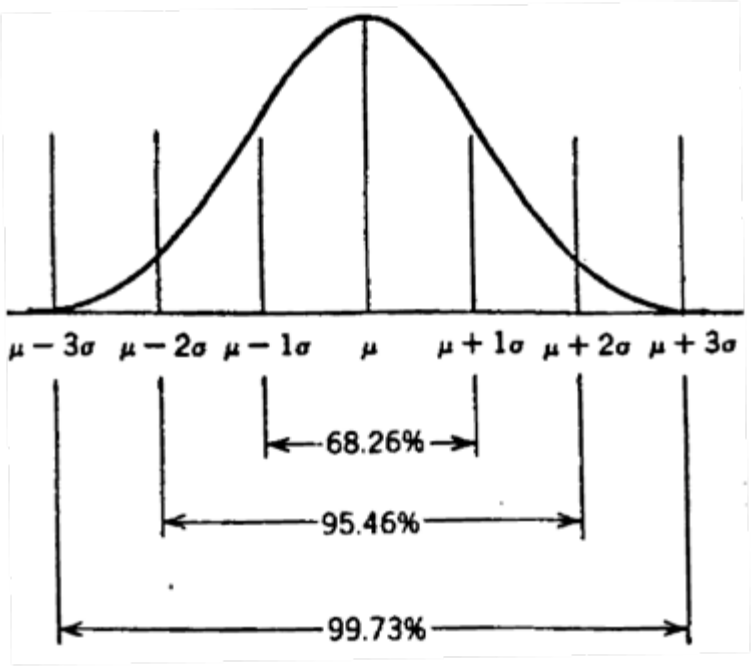
$$\cong 1-n(10^{-9})+\cdots$$

$$\cong 1-67\bullet 10^6\bullet 10^{-9}$$

$$\cong 0.933$$

P =	1.00E-09
n =	6.71E+07
Failures	
0	93.51%
1	6.28%
2	0.21%
3	4.7102E-05
4	7.9025E-07
5	1.0607E-08
> 3 failures	0.004790%
> 5 failures	1.198E-10

A Continuous Distribution: the Normal



$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \qquad -\infty < x < \infty$$

Notation: $x \sim N(\mu, \sigma)$
i.e.: x is Normally Distributed with a mean of μ and a standard deviation of σ.

Also called a Gaussian Distribution

Standard Normal Distribution

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

is the Standard Normal Distribution: $z \sim N(0,1)$

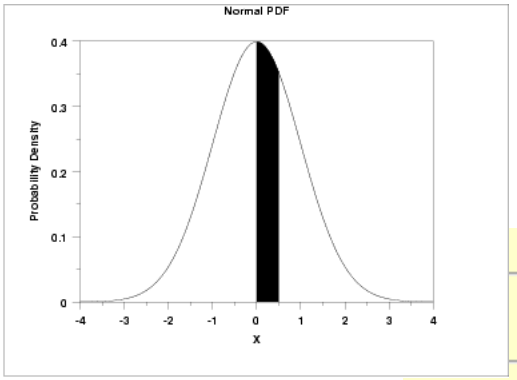
i.e. $\mu = 0$, $\sigma = 1$

so if $x \sim N(\mu,\sigma)$ **, then** $z = \frac{x-\mu}{\sigma} \sim N(0,1)$

There are tables of $\Phi(z) = \int_0^z \frac{e^{-\omega^2/2}}{\sqrt{2\pi}} d\omega$

P. Nahas

Normal Distribution Table from ESH



f(z) = f(-z)

Note: Area from 0 to +∞ = 0.5

Area under the Normal Curve from 0 to X

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899

Example: Table Lookup for Normal Distribution

- The wafer-to-wafer thickness of a poly layer is distributed normally around 500nm with a ? of 20nm:
  - P_{th} ~ N (500 nm, 20 nm)

- What is the probability that a given wafer will have polysilicon thicker than 510nm?

- ... thinner than 480nm?

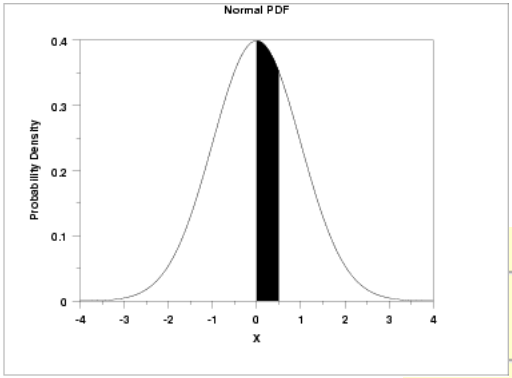
- ... between 490 and 515nm?

Example: Table Lookup for Normal Distribution

- The wafer-to-wafer thickness of a poly layer is distributed normally around 500nm with a □ of 20nm:
  - Pth ~ N (500 nm, 20 nm)
- What is the probability that a given wafer will have polysilicon thicker than 510nm?
  - 510 – 500 nm = 10 nm = 0.5 □ from mean
  - From table 0 to 0.5 □ = 0.19 for between 500 and 510 nm.
  - Greater than 510 nm = 0.5 – 0.19 = 0.31
- ... thinner than 480nm?
  - 500 – 480 nm = 20 nm = 1 □ from mean
  - From table 0 to 1 □ = 0.34 for between 480 and 500 nm
  - Thinner than 480 = 0.5 – 0.34 = 0.26
- ... between 490 and 515nm?
  - 500 – 490 nm = 0.5 □ from mean; 515 – 500 = 0.75 □ from mean
  - From table: 0.19 + 0.27 = 0.46

Poolla & Spanos

Normal Distribution Table from ESH



Area under the Normal Curve from 0 to X

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899

The Additivity of Variance

- IF $y = a_1x_1 + a_2x_2 + \ldots + a_nx_n$
  - then $\mu_y = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$
  - and $\sigma_y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2$
  - This applies applies under the assumption that the parameters x_i are ***independent***..

Examples:

- The thickness variance of a layer defined by two consecutive growths:
  - $\mu_t = \mu_{g1} + \mu_{g2}$
  - $\sigma_t^2 = \sigma_{g1}^2 + \sigma_{g2}^2$
- The thickness variance of a growth step followed by an etch step:
  - $\mu_t = \mu_g - \mu_e$
  - $\sigma_t^2 = \sigma_g^2 + \sigma_e^2$

Example: How to Combine Consecutive Steps

- The thickness of a SiO2 layer is distributed normally around 600nm with a □ of 20nm:
  - G_{ox} ~ N (600nm, 20nm)
- During a polysilicon removal step with limited selectivity, some of the oxide is removed. The removed oxide is:
  - R_{ox} ~N (50nm, 5nm)
- What is the probability that the final oxide thickness is between 540 and 560nm?

Example: How to Combine Consecutive Steps

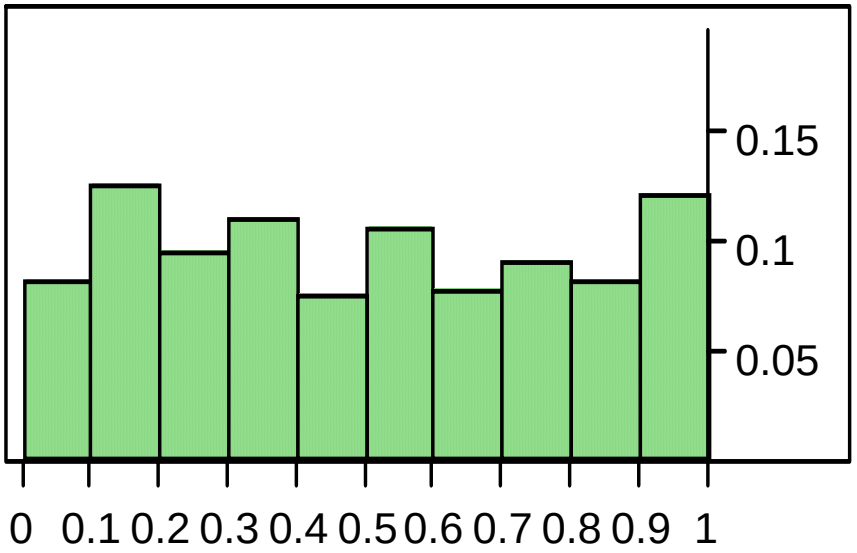
- The thickness of a SiO2 layer is distributed normally around 600nm with a □ of 20nm:
  - G_{ox} ~ N (600 nm, 20 nm)
- During a polysilicon removal step with limited selectivity, some of the oxide is removed. The removed oxide is:
  - R_{ox} ~N (50nm, 5nm)
- What is the probability that the final oxide thickness is between 540 and 560nm?
- Calculations:
  - $\mu_E = \mu_G - \mu_R = 600 - 50$ nm = 550 nm
  - $□_E^2 = □_G^2 + □_R^2 = 20^2 + 5^2$ nm^2 = 425 nm^2; $□_E$ = 20.6 nm
  - E_{ox} ~N(550 nm, 20.3 nm)
  - 540 nm = $\mu_E - 0.49\ □_E$; 560 = $\mu_E + 0.49\ □_E$
  - 0.19 + 0.19 = 0.38

Poolla & Spanos

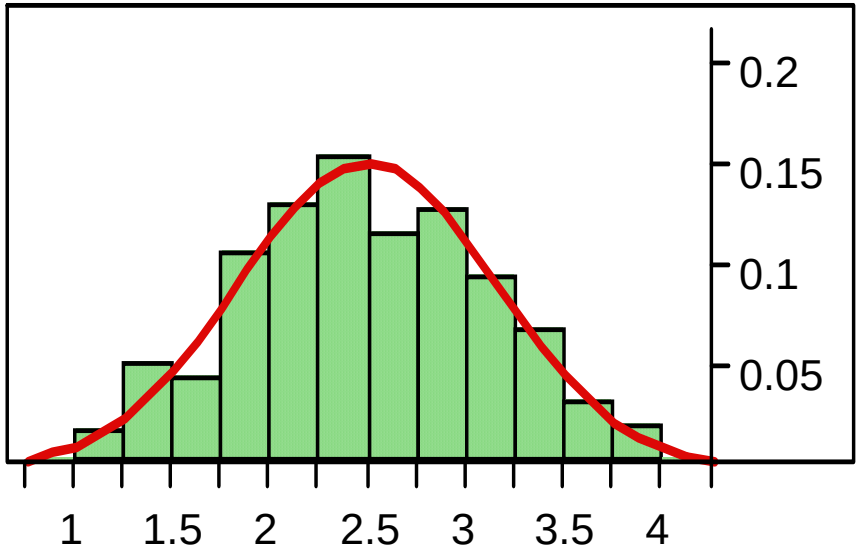
The Central Limit Theorem:

- **The distribution of a sum or average of many random variables is close to normal.**
  - **This is true even if the variable are not independent and even if they have different distributions.**
- **More observations are needed if the distribution shape is far from normal.**
- **No distribution should be dominant.**

Uniformly distributed number



Sum of 5 unif. distr. numbers:



P Nahas and Poolla & Spanos

Dice and the Central Limit Theorem

One Die	Count	Probability
1	1	0.167
2	1	0.167
3	1	0.167
4	1	0.167
5	1	0.167
6	1	0.167
	6	
Two Die	Count	Probability
2	1	0.028
3	2	0.056
4	3	0.083
5	4	0.111
6	5	0.139
7	6	0.167
8	5	0.139
9	4	0.111
10	3	0.083
11	2	0.056
12	1	0.028
	36	
Three Die	Count	Probability
3	1	0.005
4	3	0.014
5	6	0.028
6	10	0.046
7	15	0.069
8	21	0.097
9	25	0.116
10	27	0.125
11	27	0.125
12	25	0.116
13	21	0.097
14	15	0.069
15	10	0.046
16	6	0.028
17	3	0.014
18	1	0.005
	216