Statistics Part I – Introduction

Joseph J. Nahas

University of Notre Dame

UNIVERSITY OF NOTRE DAME

A Very Simple Example: A Pair of Die

- A pair of six sided die
  - Values for each die: 1, 2, 3, 4, 5, 6.
  - Values for the pair: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.
  - When tossed, each side has a probability of 1/6.
- What is the Mean value of a throw of the dice?
- What is the Variance/Standard Deviation?

Experiment with Die

Collect data from fifteen throws of a pair of die.

Dice Experiment – 15 throws

Trial	a	b	x=a+b	x-μ	(x-μ)^2	
1	6	5	11	3.133	9.818	
2	6	6	12	4.133	17.084	
3	6	2	8	0.133	0.018	
4	2	6	8	0.133	0.018	
5	3	4	7	-0.867	0.751	
6	2	6	8	0.133	0.018	
7	4	2	6	-1.867	3.484	
8	2	1	3	-4.867	23.684	
9	3	1	4	-3.867	14.951	
10	6	3	9	1.133	1.284	
11	4	5	9	1.133	1.284	
12	5	2	7	-0.867	0.751	
13	6	6	12	4.133	17.084	
14	6	1	7	-0.867	0.751	
15	2	5	7	-0.867	0.751	
						"Standard
	Count		"Mean"		"Variance"	Deviation"
	15		7.867		6.552	2.560

Dice Experiment – 100 throws

Trial	a	b	x=a+b	x-μ	(x-μ)^2	
1	2	6	8	1.050	1.103	
2	3	5	8	1.050	1.103	
3	1	1	2	-4.950	24.503	
4	4	3	7	0.050	0.002	
5	6	1	7	0.050	0.002	
95	1	6	7	0.050	0.002	
96	2	1	3	-3.950	15.603	
97	6	4	10	3.050	9.303	
98	5	6	11	4.050	16.403	
99	6	4	10	3.050	9.303	
100	6	4	10	3.050	9.303	
						"Standard
	Count		"Mean"		"Variance"	Deviation"
	100		6.950		5.806	2.409

Link: A Excel Spreadsheet Experiment

A Very Simple Example: A Pair of Die

- What is the Mean value of a throw of the dice?
  - Theoretically?
- What is the Variance and Standard Deviation of the the values from a pair of thrown die?
  - Theoretically?

Dice Theoretical Analysis

Value	Probability	Mean μ			Variance σ^2	Standard Deviation
x	P	x*P	x - μ	$(x - \mu)^2$	$P*(x - \mu)^2$	σ
2	1/36	2/36	-5	25	1*25/36	
3	2/36	6/36	-4	16	2*16/36	
4	3/36	12/36	-3	9	3*9/36	
5	4/36	20/36	-2	4	4*4/36	
6	5/36	30/36	-1	1	5*1/36	
7	6/36	42/36	0	0	6*0/36	
8	5/36	40/36	1	1	5*1/36	
9	4/36	36/36	2	4	4*4/36	
10	3/36	30/36	3	9	3*9/36	
11	2/36	22/36	4	16	2*16/36	
12	1/36	12/36	5	25	1*25/36	
Sum	36/36=1	μ=252/36=7			210/36=5.83	2.42

A Very Simple Example: A Pair of Die

- What is the Mean value of a throw of the dice?
  - Experimentally?
  - Theoretically?
  - What is the difference?
- What is the Variance and Standard Deviation of the the values from a pair of thrown die?
  - Experimentally?
  - Theoretically?
- What is the difference between the theoretical result and the experimental result?

Ideal vs Experiment

- Measures from theory and experiment have different names.
- Measures of Location:
  - μ - Mean – Measure of the center of the distribution based on the full population of the random variable.
  - $\bar{x}$ - Estimate of the Mean (Average) - Measure of the center of the distribution based on a sample of the random variable.
  - Very seldom is the Mean actually calculated because the full population is usually unknown; it is the Estimate of the Mean or Average that is calculated.
- Measures of Spread:
  - σ^2 – Variance – Measure of the spread of the distribution based on the full population.
  - σ – Standard Deviation
  - s^2 – Variance Estimate – Based on a sample of the population.
  - s – the Standard Deviation Estimate

Notation

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

A Look Ahead: What can we learn about the mean from the average?

- Confidence Limits
  - Sometimes called Error Bars

$$\mu \subset \bar{x} \pm t_{1-\alpha/2,N-1}\frac{s}{\sqrt{N}}$$

  - CL ∝ s
    - i.e. proportional to the estimate of the standard deviation.
  - CL ∝ $1/N^{1/2}$
    - proportional to the inverse square root of the number of samples.
    - If you want better estimates, take more samples.
  - CL ∝ $t_{1-\alpha/2, N-1}$
    - where $t_{1-\alpha/2, N-1}$ is Student's t distribution. (Student, Biometrika 1908)
    - where α is the desired significance level, e.g. 95%
    - where N is the number of samples and N-1 is the degrees of freedom for the distribution.
- Question: What does Guinness Beer have to do with Confidence Limits?

15 Samples

Trial	a	b	x=a+b	x-μ	(x-μ)^2	
1	5	6	11	3.200	10.240	
2	5	2	7	-0.800	0.640	
3	4	1	5	-2.800	7.840	
4	6	6	12	4.200	17.640	
5	5	5	10	2.200	4.840	
6	6	2	8	0.200	0.040	
7	6	6	12	4.200	17.640	
8	4	5	9	1.200	1.440	
9	3	1	4	-3.800	14.440	
10	3	1	4	-3.800	14.440	
11	2	1	3	-4.800	23.040	
12	4	4	8	0.200	0.040	
13	5	1	6	-1.800	3.240	
14	5	4	9	1.200	1.440	
15	5	4	9	1.200	1.440	
	Count		"Mean"		"Variance"	"Standard Deviation"
	15		7.800		8.457	2.908
Confidence		95%	Mean	7.800	+/-1.610	

100 Samples

Trial	a	b	x=a+b	x-μ	(x-μ)^2	
1	6	5	11	4.120	16.974	
2	1	5	6	-0.880	0.774	
3	5	2	7	0.120	0.014	
4	4	3	7	0.120	0.014	
5	1	6	7	0.120	0.014	
95	4	6	10	3.120	9.734	
96	3	3	6	-0.880	0.774	
97	2	3	5	-1.880	3.534	
98	3	5	8	1.120	1.254	
99	2	5	7	0.120	0.014	
100	6	2	8	1.120	1.254	
	Count		"Mean"		"Variance"	"Standard Deviation"
	100		6.880		5.339	2.311
	Confidence	95%	Mean	6.880	+/-0.458	

Note smaller confidence range

Link: A Excel Spreadsheet Experiment

What Have We Learned

- We most often do not know the distribution of the underlying population of the random variable.
- We can estimate parameters such as mean and standard deviation from a sample of data but we cannot know their actual values.
- We can estimate a range of possibilities for the parameters with a certain degree of confidence.

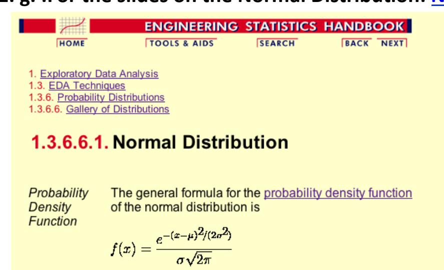
Why Statistics?

Why do we need to study statistics to do good research?

- Variability of process parameters is prevalent.
- Ultimate determination of variability is through experimental measurements.
- Patterns of variability are complex.
  - i.e. contribution of various sources to the total variation.
- Separate signal from noise.
  - e.g. the contribution of the measurement system.
- To be able to make informed decisions in the presence of uncertainty.
- To properly interpret, model, and use the data, we need an understanding of formal statistical techniques.

Online References

- NIST/SEMATECH Engineering Statistics Handbook (NIST ESH)
  - http://www.itl.nist.gov/div898/handbook/index.htm
  - Slides will contain section references in the handbook
    - E. g. :For the slides on the Normal Distribution: NIST ESH 1.3.6.6.2



  - Click on Detailed Contents on Home page to find the pages.

References on Slides

References and links to the
NIST Engineering Statistics Handbook
will be placed here.

Sources

- Lectures contain material from:
  - Prof. Michael Orshansky, ECE, UT Austin
  - Profs. Kameshwar Poolla, and Costas J. Spanos, EECS, UC Berkeley
  - Patricia A. Nahas

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