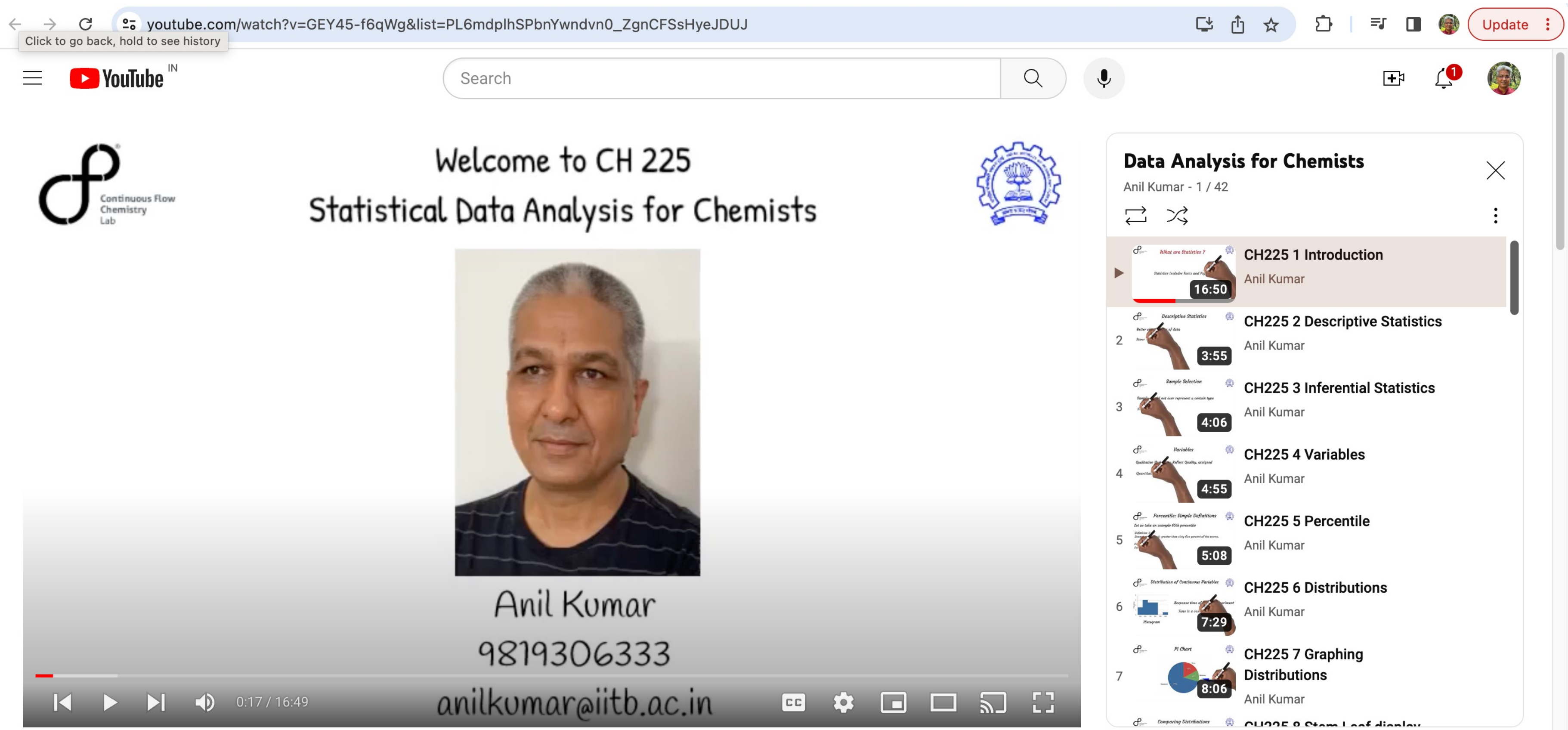
CH 225: Data Analysis for Chemists

The role of statistics, Graphical and numerical methods for describing and summarizing data, Population distributions, Sampling variability and sampling distributions, Estimation using a single sample, Hypothesis testing a single sample, Comparing two populations or treatments, Simple linear regression and correlation, ANOVA.

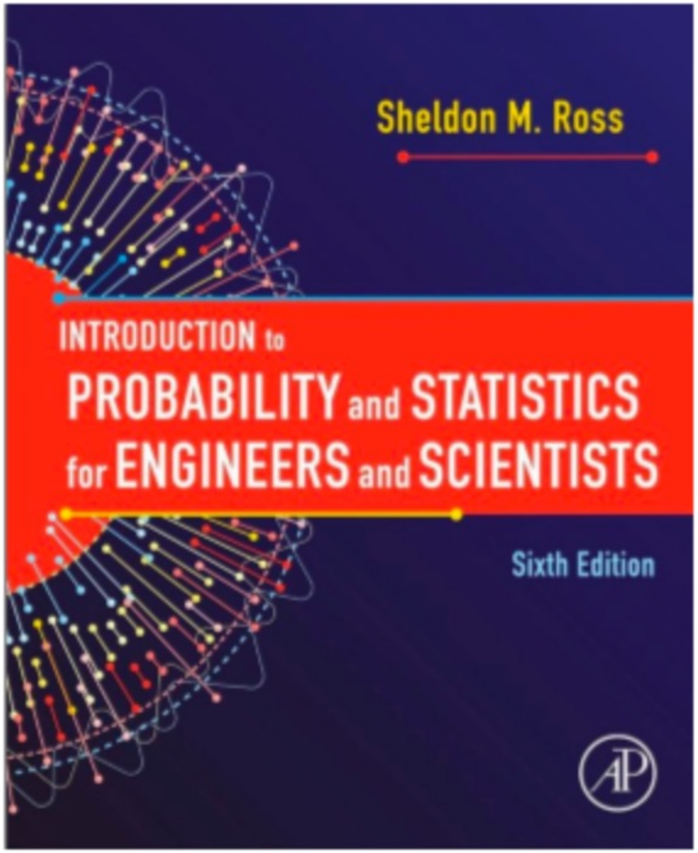
CH 225: Data Analysis and Interpretation

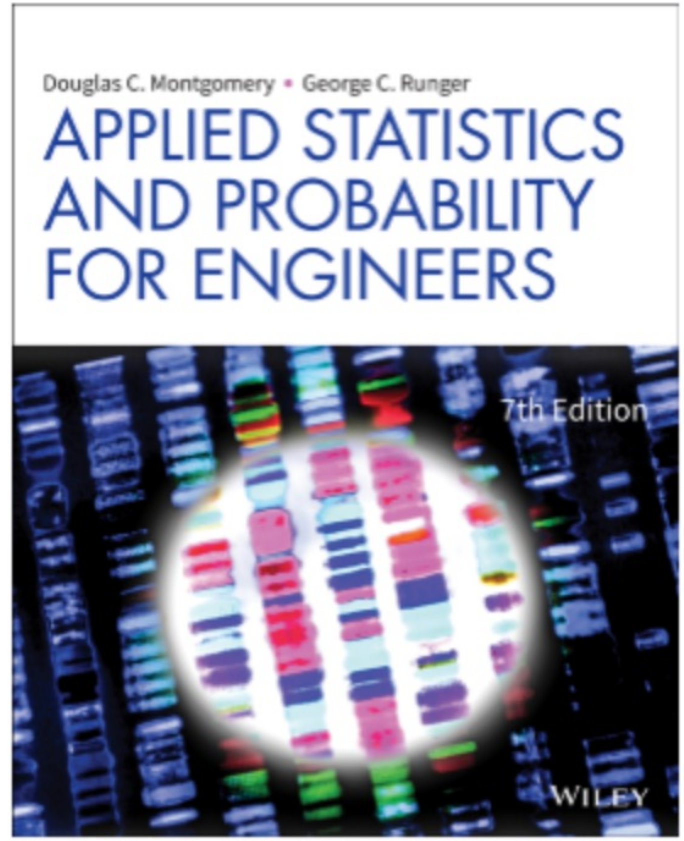
- **80% Attendance (absence with medical certificate from IITB)**
- **10 Marks for Class Participation in terms of Attendance (minus one mark per absence, max of 10)**
- **10 Marks for a quiz before the midsem**
- **30 Marks for the midsem**
- **10 Marks for a quiz after the midsem**
- **40 Marks for the final exam**



https://www.youtube.com/watch?v=GEY45-f6qWg&list=PL6mdplhSPbnYwndvn0_ZgnCFSsHyeJDUJ

Watch Quickly up to Video 20 (First two weeks)


Sheldon M. Ross
INTRODUCTION to
PROBABILITY and STATISTICS for ENGINEERS and SCIENTISTS
Sixth Edition
AP


Douglas C. Montgomery • George C. Runger
APPLIED STATISTICS AND PROBABILITY FOR ENGINEERS
7th Edition
WILEY

Descriptive Statistics: ***Part of statistics concerned with the description and summarization of data***

Inferential Statistics: ***Part of statistics concerned with the drawing of conclusions from data***

Population: ***The total collection of all the elements that we are interested in***

Sample: ***A subgroup of the population that will be studied in detail***

Random and Stratified Random Sample

Bar Charts

Pi Charts

Histograms

Frequency (relative) Graphs

Stem-and-Leaf Plots

Paired Data (Scattered, Correlated)

STATISTICS

MEAN:	Arithmetic Mean
MEDIAN:	Middle Value
MODE:	Most Frequent Value
VARIANCE:	Spread (Range, IQR etc)
STANDARD DEVIATION:	Spread

MEAN

Mean does not itself to be one of the data point

$$y_i = x_i + c \quad i = 1, 2 \ldots\ldots n \quad \text{Then } \bar{y} = \bar{x} + c$$

$$y_i = cx_i \quad i = 1, 2 \ldots\ldots n \quad \text{Then } \bar{y} = c\bar{x}$$

Symmetrical Distributions: Means and Mode are close

MEAN

Sample mean of a frequency table

Value	Frequency
3	2
4	1
5	3

$$\bar{x} = (f1/n)x1 + (f2/n)x2 + \ldots\ldots. + (fk/n)xk$$

$$\bar{x} = (2/6)3 + (1/6)4 + (3/6)5 = 25/6$$

Weighted Average of the disntinct values

Variability

Variance: Average Squared difference of scores from Mean

$$\text{Variance} = S^2 = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

$$\text{Standard Deviation} = s = \sqrt{s^2}$$

Variability does not change when you add a constant to each data value

If one multiplies the data set by a constant c, then

$y_i = cx_i \qquad i = 1, \ldots, n \quad \text{Then} \qquad s^2 = c^2 s_x^2$

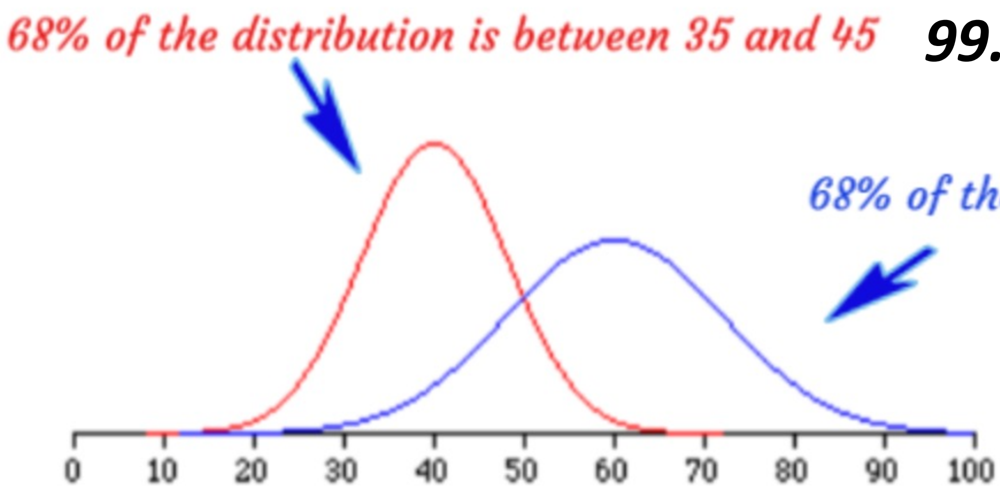
Variability

Normal Distributions

68% Distribution = Mean ± 1 SD

95% Distribution = Mean ± 2 SD

99.7% Distribution = Mean ± 3σ



Normal Distributions with Std Dev of 5 and 10

Variables

Independent and Dependent

Qualitative and Quantitative

Discrete and Continuous

Statistic	Sample	Population
Mean	$\overline{x}$	μ
Variance	s^2	σ^2
Standard Deviation	s	σ

Expectation of Random Variable

$$Mean = Expected\ Value = \mathcal{E}\,[X]$$

Experiment : Flip a fair coin twice

Sample Space $\mathcal{S} = \{HH, HT, TH, TT\}$

Random Variable X : Number of Heads

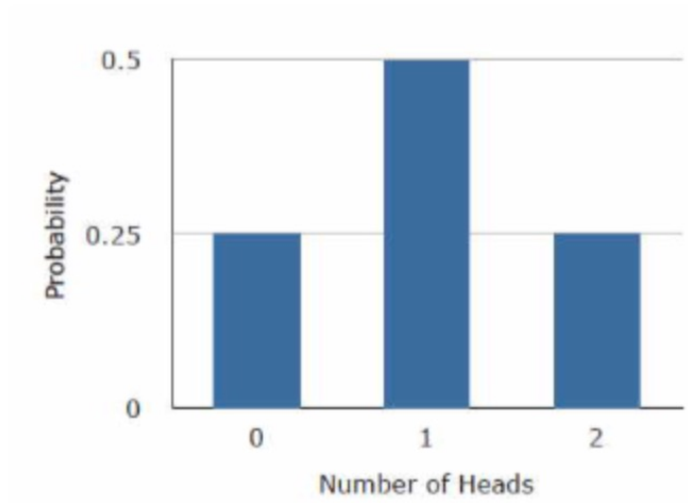
Random Variable X : 0, 1, and 2

Continuous Flow Chemistry Lab

Expectation of Random Variable

Experiment : Flip a fair coin twice

$\mathcal{P}(X = 0) = \mathcal{P}(TT) = 1/4,$
$\mathcal{P}(X = 1) = \mathcal{P}(TH) + \mathcal{P}(HT) = 2/4,$
$\mathcal{P}(X = 2) = \mathcal{P}(HH) = 1/4,$



$Mean = \mu = \mathcal{E}(X) = 0(1/4) + 1(2/4) + 2(1/4) = 1$

$$E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$$

Expectation of Random Variable

Expected Value for Continuous Variable

$$E\left[X\right] = \int_{-\infty}^{\infty} x f(x) dx$$

Variance of Random Variable

If X is a random variable with mean μ, then the variance of X, denoted by $Var(X)$, is defined by

$$Var(X) = \mathcal{E}[(X-\mu)^2]$$

Variance of Random Variable

$$Var(X) = \mathcal{E}[(X-\mu)^2]$$

$$Var(X) = \mathcal{E}[(X-\mu)^2]$$

$$= \mathcal{E}[X^2 - 2\mu X + \mu^2]$$

$$= \mathcal{E}[X^2] - \mathcal{E}[2\mu X] + \mathcal{E}[\mu^2]$$

$$= \mathcal{E}[X^2] - 2\mu\mathcal{E}[X] + \mu^2$$

$$Var(X) = \mathcal{E}[X^2] - \mu^2$$

Sum of Random Variables

If two random variables are independent:

$$\mathcal{E}[X + Y] = \mathcal{E}[X] + \mathcal{E}[Y]$$

$$\mathcal{E}[X - Y] = \mathcal{E}[X] - \mathcal{E}[Y]$$

$$Var[X + Y] = Var[X] + Var[Y]$$

$$Var[X - Y] = Var[X] + Var[Y]$$

Sum of Random Variables

If two random variables are dependent:

$$\mathcal{E}[X + Y] = \mathcal{E}[X] + \mathcal{E}[Y]$$

$$\mathcal{E}[X - Y] = \mathcal{E}[X] - \mathcal{E}[Y]$$

$$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y + 2\rho\,\sigma_x\,\sigma_y$$

$$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2\rho\,\sigma_x\,\sigma_y$$

ρ = Correlation between X and Y

Bivariate Data: Correlation

$$Correl = r = \frac{Covar(X, Y)}{[Var(X)\ Var(Y)]^{1/2}}$$

$$Covar(X, Y) = \mathcal{E}\,[(X - \mu_x)\ (Y - \mu_y)]$$

ρ For Population

r For Sample