Population (Probability) Distribution

Satishkumar L. Varma
Professor, Department of Computer Engineering
PCE, New Panvel
www.sites.google.com/view/vsat2k
www.vsat2k.wordpress.com
www.vsat2k.moodlecloud.com

Outline

- Probability Distribution
- Theoretical:
  - Binomial and normal distributions
  - Poisson and Exponential distributions
  - Hyper geometric and uniform distributions
- Tests for equality of mean and variances of two populations
- Confidence interval
- Probability value (p-value)
- Significant Level
- Degree of Freedom
- **Type I and II error**

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

- Understand the idea of probability distribution used in concept in statistics
- How Joint and conditional probabilities are used to analyze corpus data
- Know the different Statistical Measures used in hypothesis testing
- How probability plays an important role in statistical hypothesis testing

Probability

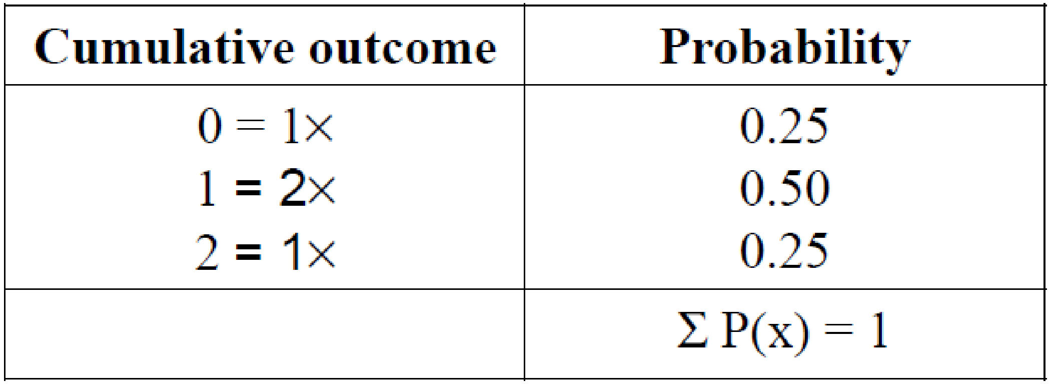
- Two reasons why probability is important for the analysis of linguistic data:
  - Joint and conditional probabilities are used to analyze corpus data
  - Probability plays an important role in statistical hypothesis testing
- **Simple probability**
  - Eg: If you toss a dice with six number (i.e. 1,2,3,4,5,6) what is the probability that you will toss a 6?
  - Ans: P(6) = 1/6 = 0.1666
- Values range from 0 to 1 and total probabilities of the sample is 1
- If two events are independent, the probability is the sum of their individual probabilities
- Two events A and B are independent if knowing that the occurrence of A does not change the probability of the occurrence of B

Statistical hypothesis testing

- **Joint probability**
  - P(A,B) = P(A) x P(B)
  - Example: P(5,6) = (0.166) x (0.166) = 0.0277
- **Conditional probability**
  - P(A | B) = P(A ^ B) / P(B)
  - Eg: A corpus including 2000 nouns & 500 adjectives, 50 adjectives precede a noun
  - What is the likelihood that a noun occurs after an adjective?
  - What is the likelihood that an adjective precedes a noun?
  - P(ADJ | N) = P(ADJ ^ N) / P(N)
  - P(ADJ | N) = P(50) / P(2000) = 0.1666
  - P(N | ADJ) = P(50) / P(500) = 0.5714

Probability distribution

- What is the probability that you get two heads if you toss a coin twice?
  - 0 heads = HH 25%
  - 1 head = HT + TH 50%
  - 2 heads = TT 25%
  - Sample space
  - Random variable



Cumulative outcome	Probability
0 = 1×	0.25
1 = 2×	0.50
2 = 1×	0.25
	Σ P(x) = 1

Population (Probability) Distributions

- One of the most important concepts in statistics is the idea of a probability distribution
- **Discrete probability distribution**
  - This is the binomial probability distribution
  - Which we can plot as a histogram
  - Prominent in statistics but not much in the geosciences
- **Continuous probability distribution**
  - Most common continuous distribution by far is the normal of Gaussian distribution
  - Normal distribution is the well-known bell shaped distribution
  - Feature of Normal Distribution
    - It is analytical
    - It often describes populations quite well
    - It works quite well if the populations has a lot of outliers
    - Central limit theorem:
      - whatever the probability distribution of x, the probability distribution of means of x for repeated samples of n random samples tends to become normally distributed as n increases with

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Population (Probability) Distributions

- Measures of central tendency and dispersion use two key terms
    - Population
    - Sample
- Information is used to make inferences about population from which sample was drawn
- Characteristics of the population are referred to as parameters,
- while characteristics of the sample are referred to as statistics.

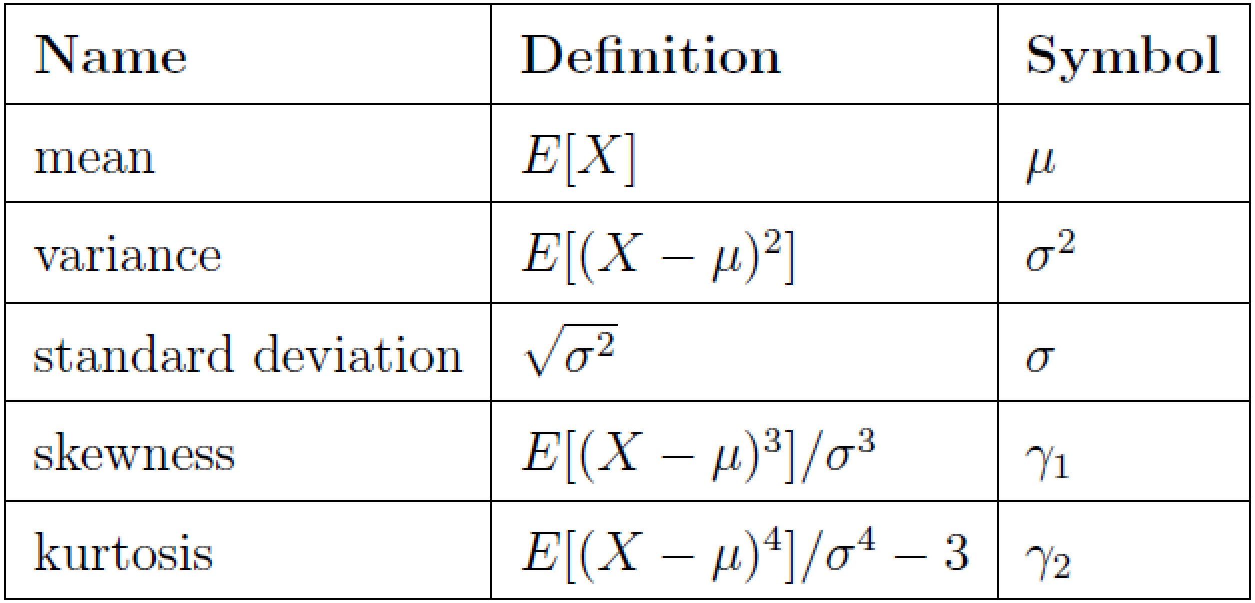
Population (Probability) Distributions

- **Central tendency: mean and median**
  - **Central tendency** refers to ways of specifying where the "middle" of a probability distribution lies. Examples include the mean and median parameters.
  - **Mean** (expected value) of a random variable can be thought of as the "balance point" of the distribution if the PDF is cut out of cardboard.
  - **Median** is the value that splits the distribution in half so that there is a 50/50 chance of a random value from the distribution occurring above or below the median.

- **Spread: variance, standard deviation, and interquartile range**
  - Population **variance** is the mean squared distance of any value from the mean of the distribution, but you only need to think of it as a measure of spread on a different scale from standard deviation.
  - **Standard deviation** is defined as the square root of the variance

Parameters Describing Distributions

- Common parameters and their definitions as expected values



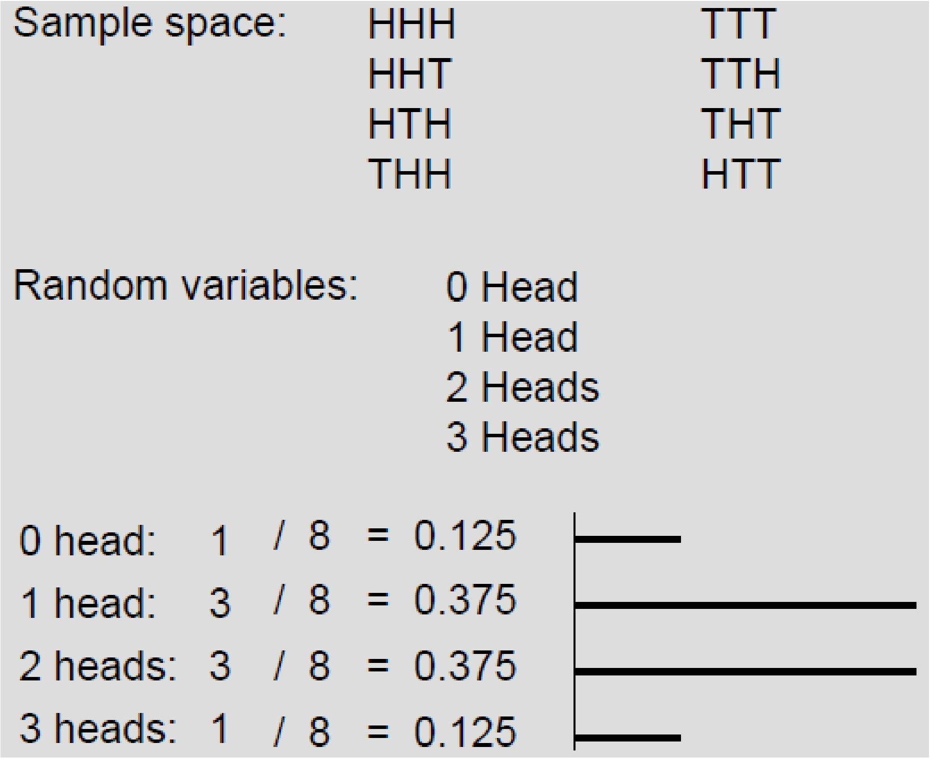
Name	Definition	Symbol
mean	$E[X]$	μ
variance	$E[(X-\mu)^2]$	σ^2
standard deviation	$\sqrt{\sigma^2}$	σ
skewness	$E[(X-\mu)^3]/\sigma^3$	γ_1
kurtosis	$E[(X-\mu)^4]/\sigma^4-3$	γ_2

Binomial Distribution

- Bernoulli trail:
  - two possible outcomes on each trail
  - the outcomes are independent of each other
  - the probability ratio is constant across trails
- Binomial distribution has the following properties:
  - It is based on categorical / nominal data.
  - There are exactly two outcomes for each trail.
  - All trials are independent.
  - The probability of the outcomes is the same for each trail.
  - A sequence of Bernoulli trails gives us the binomial distribution.

Binomial Distribution

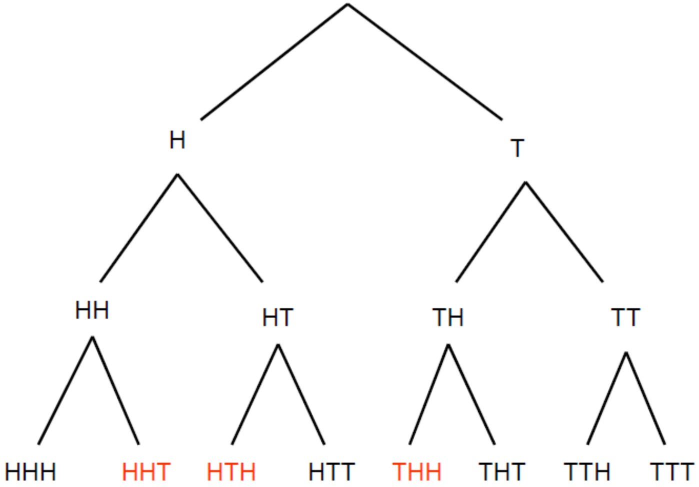
- Example: A coin is tossed three times. What is the probability of obtaining two heads?



Sample space: HHH TTT
HHT TTH
HTH THT
THH HTT

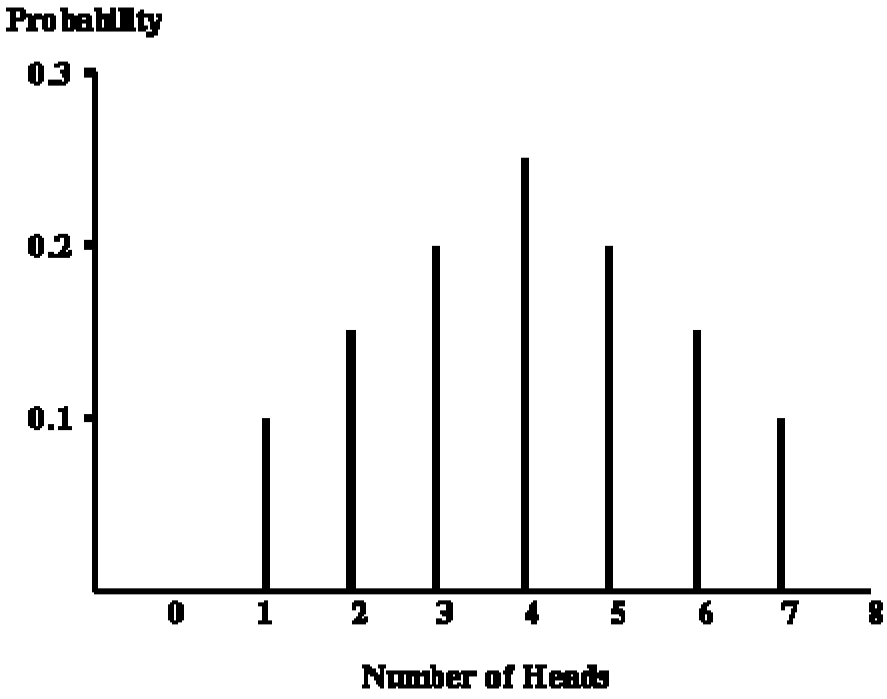
Random variables: 0 Head
1 Head
2 Heads
3 Heads

0 head:	1	/	8	=	0.125
1 head:	3	/	8	=	0.375
2 heads:	3	/	8	=	0.375
3 heads:	1	/	8	=	0.125



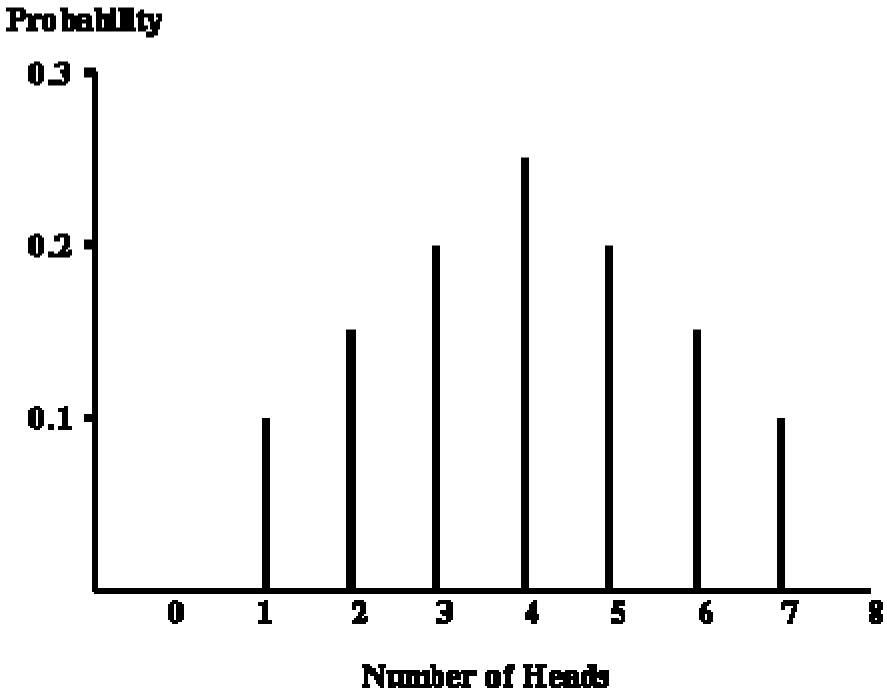
Binomial Distribution

- Example: If you toss a coin 8 times what is the probability of obtaining a score of:
  - 0 heads
  - 1 head
  - 2 heads
  - 3 heads
  - 4 heads
  - 5 heads
  - 6 heads
  - 7 heads
  - 8 heads



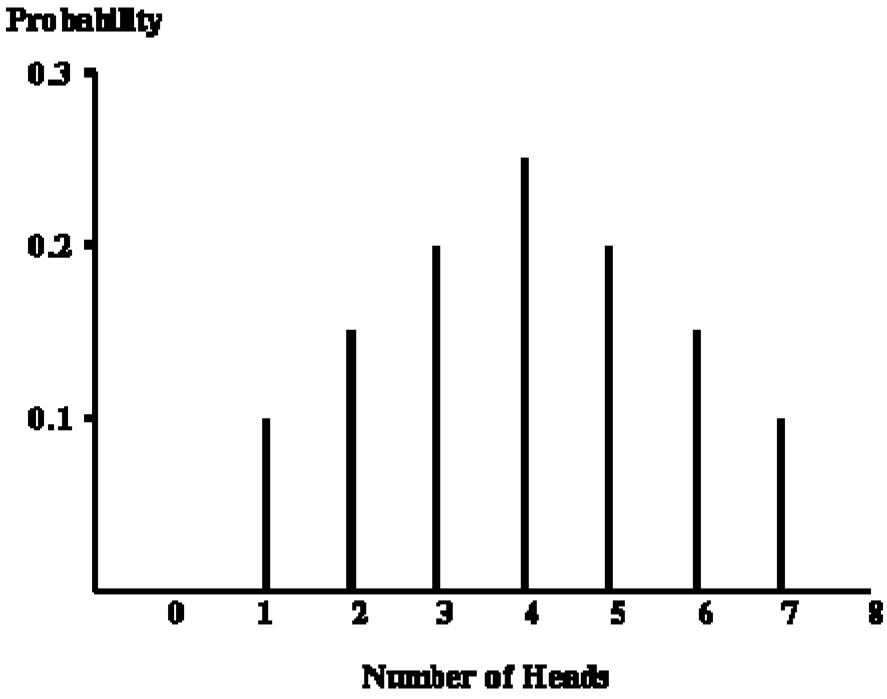
Binomial Distribution

- Example: Tossing a coin one 100 times, yielded 42 heads and 58 tails. Is this a fair coin?
  - Heads: 42
  - Tails: 58
-
-



Binomial Distribution

- Example: Tossing a coin one 100 times, yielded 42 heads and 58 tails. Is this a fair coin?
  - Heads: 42
  - Tails: 58
- Expected: 50% - 50%
- Sample error?



Population (Probability) Distributions

- **Example: Research on Child Language**
  - Q: Find if there is difference in MLU of any Language-Speaking Boys/Girls
- **Answer**
  - Collect sample of utterances (say 100) from boys and girls
  - Calculate Mean length of utterance (MLU) = # Morphemes / # Utterances
  - Higher MLU indicate a higher level of language proficiency
  - Boys MLU = 3.4 and Girls MLU = 2.9
  - The sample mean is different
- **Question: Can we conclude that boys and girls produce different MLUs**
  - We need a probability model to answer this
  - We need to inspect dataset to find the right model
- **Important Aspect to Note**
  - MLU is interval data
  - Data of boys and girls groups is centered around a mean

Population (Probability) Distributions

Example 2: Research on Height and Weight

Sex	Weight	Height
Male	178	75
Male	196	100
Male	145	60
Male	170	71
Male	180	80
Male	175	69
Male	185	78
Male	190	90
Male	183	70
Male	182	85

Sex	Weight	Height
Female	163	60
Female	142	51
Female	150	55
Female	165	64
Female	160	53
Female	175	50
Female	182	72
Female	169	65
Female	162	62
Female	182	80

Population (Probability) Distributions

- Population skewness of a distribution is a measure of asymmetry (zero is symmetric) and
- Population kurtosis is a measure of peakedness or flatness compared to a Gaussian distribution
- If a distribution is "pulled out" towards higher values (to the right), then it has positive skewness
- If it is pulled out toward lower values, then it has negative skewness.
- A symmetric distribution, e.g., the Gaussian distribution, has zero skewness
- The population kurtosis of a distribution measures how far away a distribution is from a Gaussian distribution in terms of peakedness vs. flatness.
- Compared to a Gaussian distribution, a distribution with negative kurtosis has "rounder shoulders" and "thin tails", while a distribution with a positive kurtosis has more a more sharply shaped peak and "fat tails".

Population (Probability) Distributions

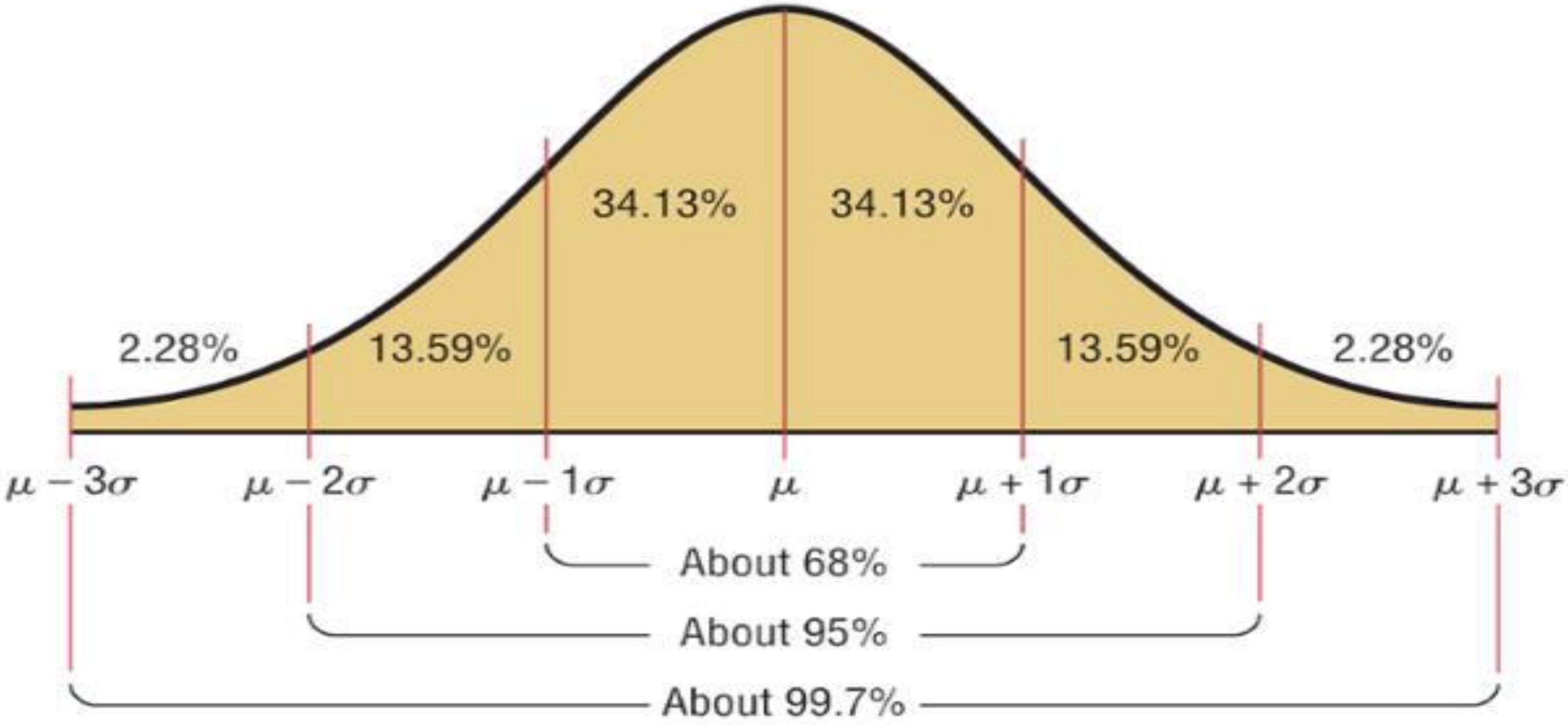
- Family of **Gaussian** (also called **Normal**) **distributions** is indexed by the mean and variance (or standard deviation) of the distribution.
- The **t-distributions**, which are all centered at 0, are indexed by a single parameter called the degrees of freedom.
- The **chi-square family of distributions** is also indexed by a single degree of freedom value.
- The **F distributions** are indexed by two degrees of freedom numbers designated numerator and denominator degrees of freedom.

Normal Distribution

- It a continuous, symmetric, and bell-shaped distribution of a variable
- Gives model for all kinds of real world phenomena having sets of data items
- Examples
  - heights and weights of humans,
  - intelligence quotients,
  - life spans, test scores, etc..
- Center of the curve represents the mean, median, and mode
- Curve is symmetrical around the mean
- Mean, Median, & Mode are equal and are located at center of the distribution
- Normal distribution has only one mode
- The tails meet the x-axis in infinity
- The total under the curve is equal to 1 or or 100% (by definition)

Normal Distribution

- The 68-95-99 Rule for the Normal Distribution
  - Appx. 68% of the data values lie within 1 st.dev. of the mean in both directions
  - Appx. 95% of the data values lie within 2 st.dev. of the mean in both directions
  - Appx. 99.7% of the data values lie within 3 st.dev. of the mean in both directions



Normal Distribution

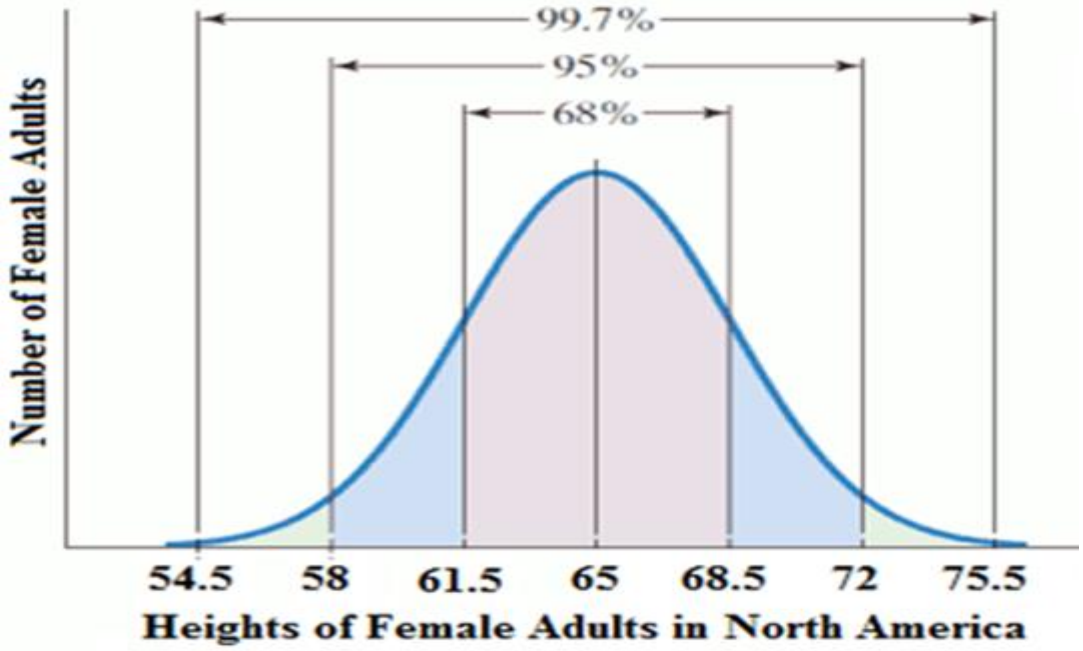
- **Computing z - Scores**
  - z - score tells us how many standard deviations an arbitrary data value in a normal distribution lies above or below the mean.
  - The formula for the **z - score**:

$$z\text{-score} = \frac{\text{data item} - \text{mean}}{\text{standard deviation}}$$

- **Percentiles**
  - Besides z - scores, there is another measure of a data value's position from the mean. It is called **percentile**
  - Percentiles are often used in standardized tests
  - Eg., if an Exam score is in the 45th percentile,
    - this means that 45% of the scores are less than this score
- **Quartiles**
  - Three popular percentiles are the quartiles
  - They indicate a division of a data set into four equal parts
  - The 25th percentile is called the **first quartile**
  - The 50th percentile is called the **second quartile**
  - The 75th percentile is called the **third quartile**

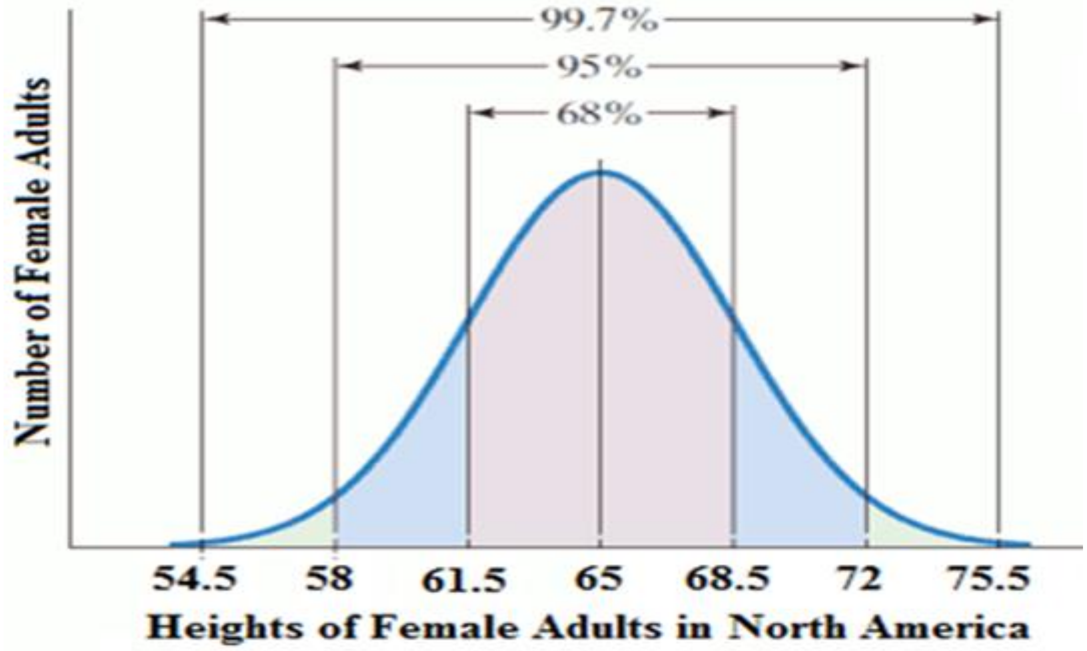
Normal Distribution

- Example: Adult female heights in a country are said to be ND with
    - a mean of 65 inches and
    - a standard deviation of 3.5 inches
- Que 1: Find the height that is 2 standard deviations above and below the mean.
    - above the mean: 65 + 2(3.5) = 72 inches
    - below the mean: 65 - 2(3.5) = 58 inches
- Que 2: Use the normal distribution to find the percentage of women in a country with the following heights:
    - between 61.5 and 68.5 inches
    - between 65 and 72 inches
    - between 54.5 and 58 inches
    - above 72 inches



Normal Distribution

- **Between 61.5 and 68.5 inches = 68%**
- **Between 65 and 71 inches = 95%   2 = 47.5%**
  - % of women with heights between 65 & 72 inches is not directly given in Figure
  - However, because a normal distribution is symmetric about the mean,
    - Half of 95% of women, or 47.5%, have heights between 65 and 72 inches.
- **Between 54.5 and 58 inches = (99.7%   2) - (95 %   2) = 2.35%**
  - % of women with heights between 54.5 & 58 inches is not directly given in Figure
  - Here we need the difference of with half of 99.7% and half of 95%.
- **Above 72 inches = (100% - 95%)   2 = 2.5%**



Normal Distribution

- **Find the z - score for a height of 60 inches, 65 inches, and 73 inches**

$$z - score = \frac{60 - 65}{3.5} \approx -1.43$$

- 60 inches:
  - z - score of a data value less than the mean is always negative
  - A female who is 60 inches tall is 1.43 standard deviations below the mean

$$z - score = \frac{65 - 65}{3.5} = 0$$

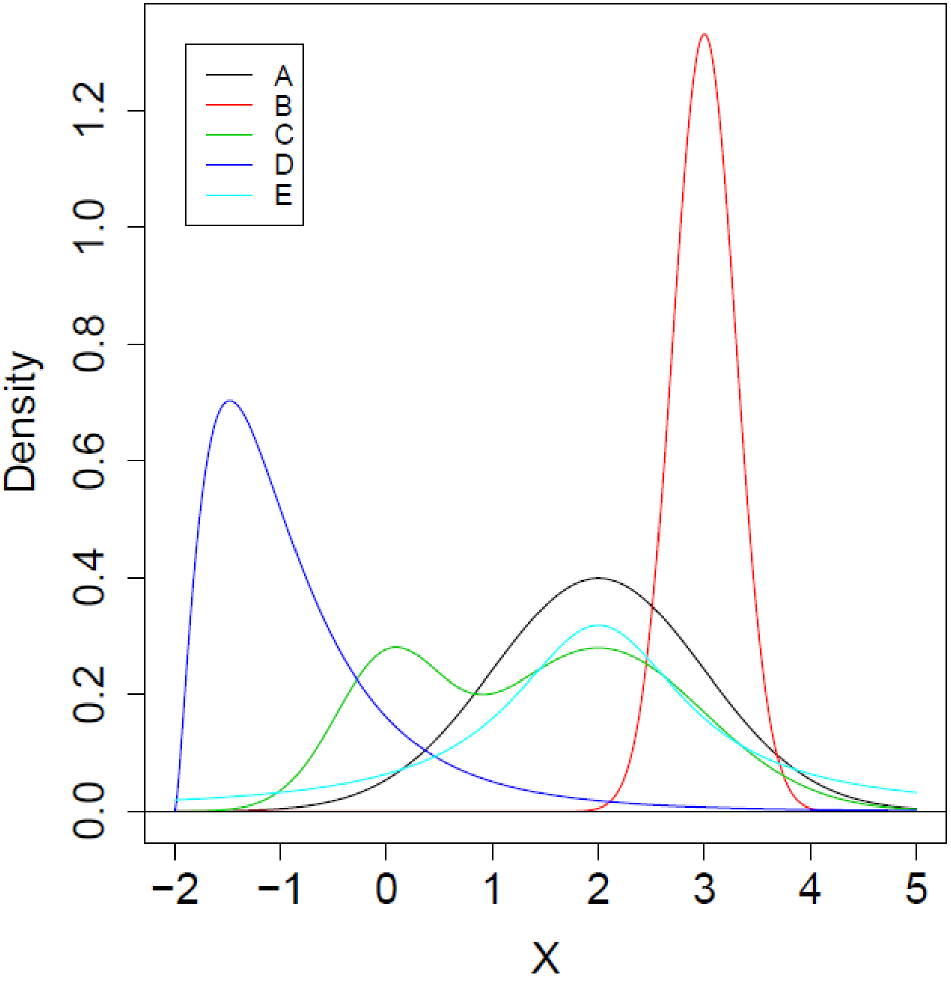
- 65 inches: A z - score for the mean is always 0.

$$z - score = \frac{73 - 65}{3.5} \approx 2.29$$

- 73 inches:
  - z - score of a data value greater than the mean is always positive
  - A female who is 73 inches tall is 2.29 standard deviations above the mean

Population (probability) distributions

- Population (probability) distributions of 5 different continuous random variables.
- **Distribution A** is a unimodal (one peak) symmetric distribution, centered around 2.0.
- It has perfect bell-shape of a Gaussian distribution.
- **Distribution B** is also Gaussian in shape, has a different central tendency (shifted higher or rightward), and has a smaller spread.
- **Distribution C** is bimodal (two peaks) so it cannot be a Gaussian distribution.
- **Distribution D** has the lowest center and is asymmetric (skewed to the right), so it cannot be Gaussian.
- **Distribution E** appears similar to a Gaussian distribution, but while symmetric and roughly bell-shaped, it has "tails" that are too fat to be a true bell-shaped, Gaussian distribution.



Probability Distributions

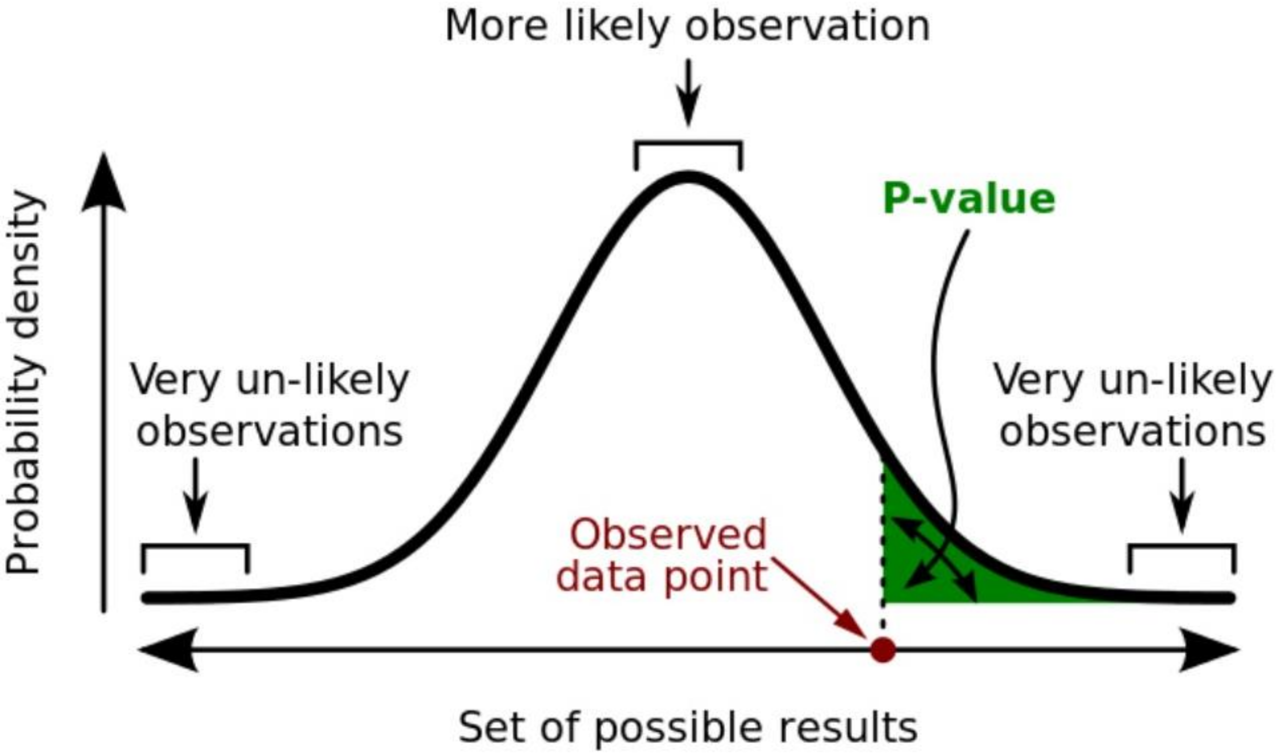
- These distributions work for several hypothesis tests
- Test of Hypothesis
  - Test whether a population parameter is less than, equal to, or greater than a specified value
- Remember an inference without a measure of reliability is little more than a guess
- Next: Parameters Describing Distributions

Probability value (p-value)

- **Probability value (p-value)**
  - Indicates the probability that we will obtain the distribution in our sample given that there is no relationship between x and y in the true population
  - It is conditional on the hypothesis that the null hypothesis is true
  - If there is **no relationship** (difference) between X and Y in the true population,
  - Then there is a **less than a 5% chance** (i.e. 1 out of 20 chance) that
    - We obtain the distribution in a given sample
- **p-value indicates**
  - The probability of the null hypothesis to be true is 5% [True/False]
  - The probability of the alternative hypothesis to be true is 95% [True/False]
  - Given that the null hypothesis is true, there is a 5% chance of obtaining the distribution in a particular sample [True/False]
- **The probability of the experimental hypothesis is not measured at all!**

Probability value (p-value)

- **Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)**
- **p-value is probability of an observed result (assume the null hypothesis is true)**
- **Probability density of each outcome is computed under null hypothesis**

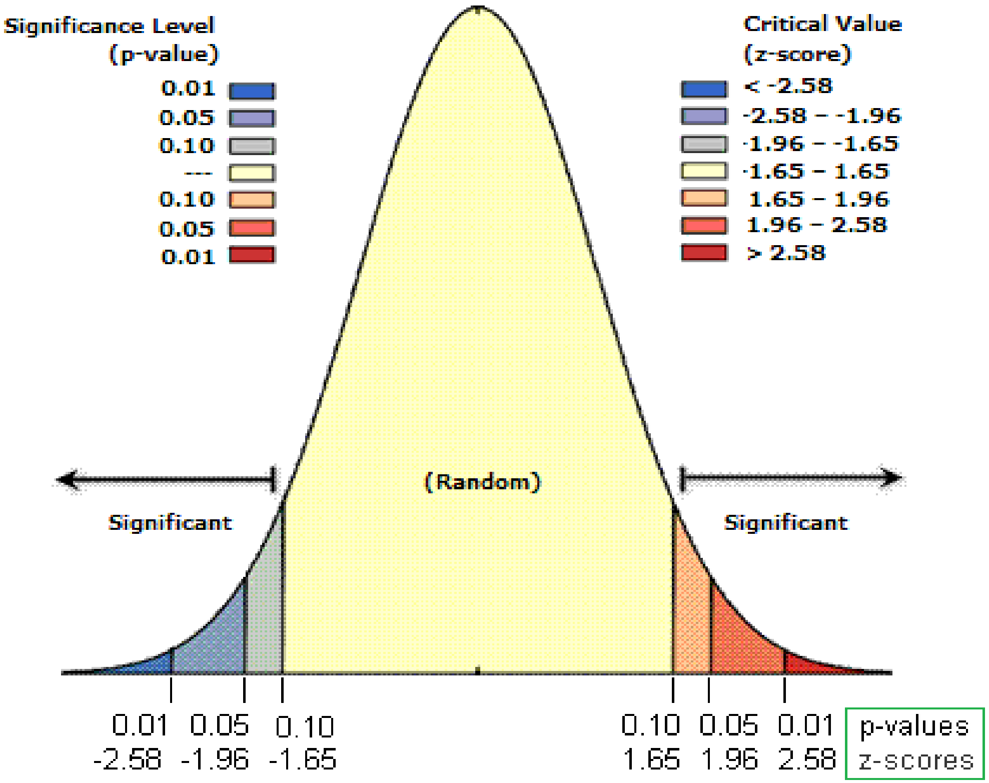


Statistical Measures

- How to decide whether to reject or accept the null hypothesis (H0)
- p-value

Condition	Decision
p > 0.05	Not Significant
p = 0.01 to 0.05	Significant
p = 0.001 to 0.01	Very Significant
p < 0.001	Extremely Significant

- Z score: Test of statistical significance
- Helps to decide whether or not to reject H0



Other Statistical Measures

- **Why Other Statistical Measures**
  - Recently the p-value was taken as a ‘holy’ cut-off point
  - But the value .05 is of course arbitrary
- So, researchers now consider p-value in conjunction with other statistical measures:
  - Z Score
  - Significant Level (alpha or α)
  - Effect size
  - Confidence intervals
- Confidence intervals: indicate the range within which the mean (or other statistical measure) must lie, assuming a particular degree of certainty (e.g. 95%) in true population
- Effect size: indicates the degree to which difference in dep.variable are due to changes in indep.variable

Other Statistical Measures

- **Other Statistical Measures**

Term	Definition	Relationship to α
p-value	It is calculated probability after A/B test that A and B are actually equal (we want this to be as low as possible)	α is threshold value that we measure p-values against
Significant Level	Used to describe a result indicating strong evidence of a difference between A and B	P-value must be < α to declare our result significant
Confidence Level	% of time (typically 95%) we anticipate our test results to be correct	Confidence Level = 1- α

Other Statistical Measures

- DF is am important idea n Statistics, Prob.distributions, Hypo.tests, and Regres.analysis
- Definition of Degrees of Freedom (DF):
  - # indep.values that a statistical analysis can estimate
  - i.e # values that are free to vary as you estimate parameters
  - available amount of indep.inform limits # parameters that we can estimate
  - DF = Sample size minus # parameters we need to calculate during an analysis
  - DF is usually a positive whole number
  - DF indicates how much indep.info goes into a parameter estimate
- DF is a combination of:
  - how much data you have and
  - how many parameters we need to estimate
- We want a lot of info. to go into parameter estimates to obtain
  - more precise estimates and
  - more powerful hypothesis tests
- So, we want many degrees of freedom!

> - DF are # observations in a sample that are free to vary while estimating stat. parameters
> - i.e the amount of independent data that we can use to estimate a parameter

Example to Illustrate DF concept

- To understand why, we need to talk about the freedom to vary
- Suppose we collect the random sample of observations
- Imagine we know mean (2.8) but don’t know the value of an observation (x in the table)
  - Table Values = [2, 1, 5, 2, x]
  - mean is 2.8 and it is based on 5 values
  - sum = 2.8 * 5 = 14 (based on the equation for the mean)
  - 10 + x = 14 then x = 4
- Observations
  - So, the last number has no freedom to vary
  - x is not an independent piece of information because it cannot be any other value
  - here mean imposes a constraint on the freedom to vary
  - x and the mean are entirely dependent on each other
- So, affter estimating mean, we have only 4 indep.pieces of info. for a sample size of 5

- DF are # observations in a sample that are free to vary while estimating stat. parameters
- i.e the amount of independent data that we can use to estimate a parameter

Degrees of Freedom and Probability Distributions

- Family of distributions (define the distributions) to determine statistical significance
  - t-distribution
  - F-distribution
  - chi-square distribution
  - DF
- DF also define the distributions for the test statistics of various hypothesis tests
- Hypothesis tests use these to determine statistical significance
- DF (among these family of distributions) define the shape
- Hypothesis tests use these distributions to calculate p-values
- So, the DF are directly linked to p-values through these distributions!

Type 1 and Type 2 Error

- **Type 1 and Type 2 error**
    - Type 1 error: we reject a true null hypothesis
    - Type 2 error: We accept a false null hypothesis
- p-value indicates the probability of making a type 1 error
- p-value does not say anything about making a type 2 error!

Model-> Prediction	**True (Null Hypothesis)**	**False (Null Hypothesis)**
$P > 0.05$	Correct	Error [Type 2]
$P < 0.05$	Error [Type 1]	Correct

Summary

- For solved examples:
  - https://goo.gl/MSrab2

Bibliography

[1] Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, 3rd edition Publisher: Morgan Kaufmann; 3 edition

[2] Business Intelligence, 2/E; Efraim Turban, Ramesh Sharda, Dursun Delen, David King; pearson Education

[3] Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with Xlminer; 2nd edition, Galit Shmueli, Nitin R. Patel and Peter C. Bruce; John Wiley

[4] Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Author: Berry, Gordon S. Linoff, Format: Paperback, 648 pages, Edition: 3; Publisher: John Wiley & Sons Inc.

[5] Robert Groth, Data Mining: Building Competitive Advantage, Prentice Hall, 2000.

[6] P. N. Tan, M. Steinbach, Vipin Kumar, “Introduction to Data Mining”, Pearson Education

[7] Alex Berson and Smith, “Data Mining and Data Warehousing and OLAP”, McGraw Hill Publication

[8] E. G. Mallach, “Decision Support and Data Warehouse Systems”, Tata McGraw Hill.

[9] Michael Berry and Gordon Linoff “Mastering Data Mining- Art & science of CRM”, Wiley Student Edition

Thank You.