STAT6171001
Basic Statistics

Normal Distribution & Estimation
Session 8

Raymond Bahana
rbahana@binus.edu

Session Learning Outcomes

Upon completion of this session, students are expected to be able to

- LO 2. Analyze a problem by using the basic concept of descriptive and inferential statistics
- LO 3. Design a descriptive and inferential statistics solution to meet a given set of computing requirements in the context of computer science
- LO4. Produce descriptive and inferential statistics solutions

Topics

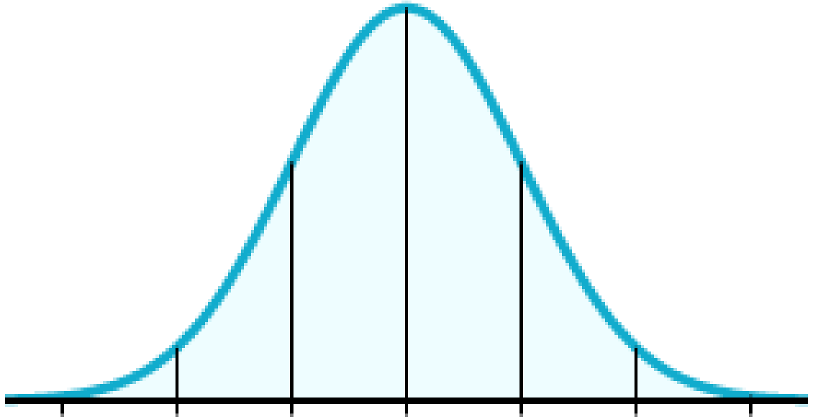
- Normal Distribution
- Sampling Distributions
- Estimation

Normal Distribution

Introduction

- The familiar bell-shaped normal curve describes many observed frequency distributions

- Most of the statistical analyses presented are based on the bell-shaped or normal distribution



Normal Distributions

- The normal distribution is the most important and most widely used distribution in statistics.
- It is sometimes called the “bell curve,” although the tonal qualities of such a bell would be less than pleasing.
- It is also called the “Gaussian curve” after the mathematician Karl Friedrich Gauss.

Standard Deviation

- The Standard Deviation is a measure of how spread-out numbers are.
- The standard deviation is simply the square root of the variance.
- The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal.

Variance

- The average of the squared differences from the Mean.
- To calculate the variance, follow these steps:
  1. Work out the Mean
  2. Then for each number: subtract the Mean and square the result
  3. Then work out the average of those squared differences.

Variance - Example

- You and your friends have just measured the heights of your dogs (in millimeters):
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

- Find out:
  - the Mean
  - the Variance
  - the Standard Deviation

Source: https://www.mathsisfun.com/data/standard-deviation.html

Variance - Example

- First step is to find the Mean:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 30}{5} = \frac{1970}{5} = 394$$

- So, the mean (average) height is 394 mm

- Now we calculate each dog's difference from the Mean
- The difference are: 206, 76, −224, 36, −94

Source: https://www.mathsisfun.com/data/standard-deviation.html

Variance - Example

- To calculate the Variance, take each difference, square it, and then average the result.

$$\text{Variance} = \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$\sigma^2 = \frac{42436 + 5776 + 50176 + 1296 + 8836}{5} = 21704$$

- And the Standard Deviation is just the square root of Variance, so:

$$\sigma = \sqrt{21704}$$

$= 147.32277$

$= 147$ (to the nearest mm)

Source: https://www.mathsisfun.com/data/standard-deviation.html

Standard Deviation

- Now we can show which heights are within one Standard Deviation (147mm) of the Mean.
- So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.
- **But ... there is a small change with Sample Data**

Source: https://www.mathsisfun.com/data/standard-deviation.html

Standard Deviation

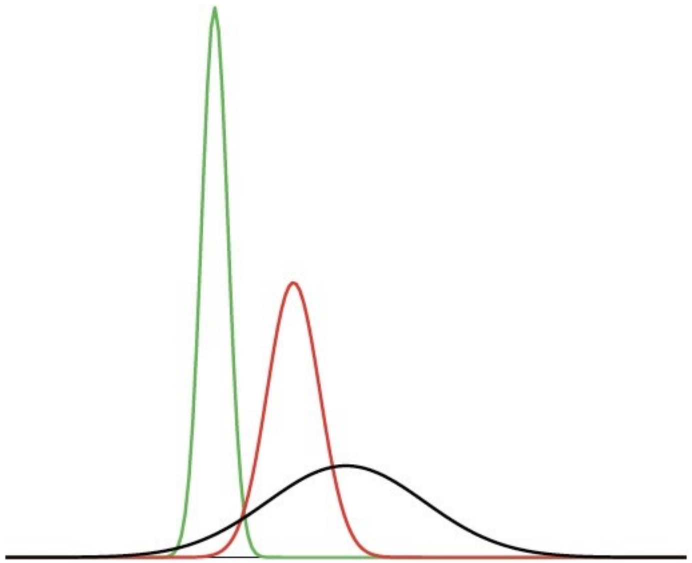
- That example is for a Population (the 5 dogs are the only dogs we are interested in).
- But if the data is a Sample (a selection taken from a bigger Population), then the calculation **changes**!

- The Population: divide by N when calculating Variance
- A Sample: divide by N-1 when calculating Variance

Source: https://www.mathsisfun.com/data/standard-deviation.html

Normal Distributions

- Normal distributions can differ in their means and in their standard deviations.
- The following figure shows three normal distributions.
  - The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5,
  - The distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1
  - The distribution in black (right-most) has a mean of 2 and a standard deviation of 3.



Normal distributions differing in mean and standard deviation.

These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

Binomial Distributions

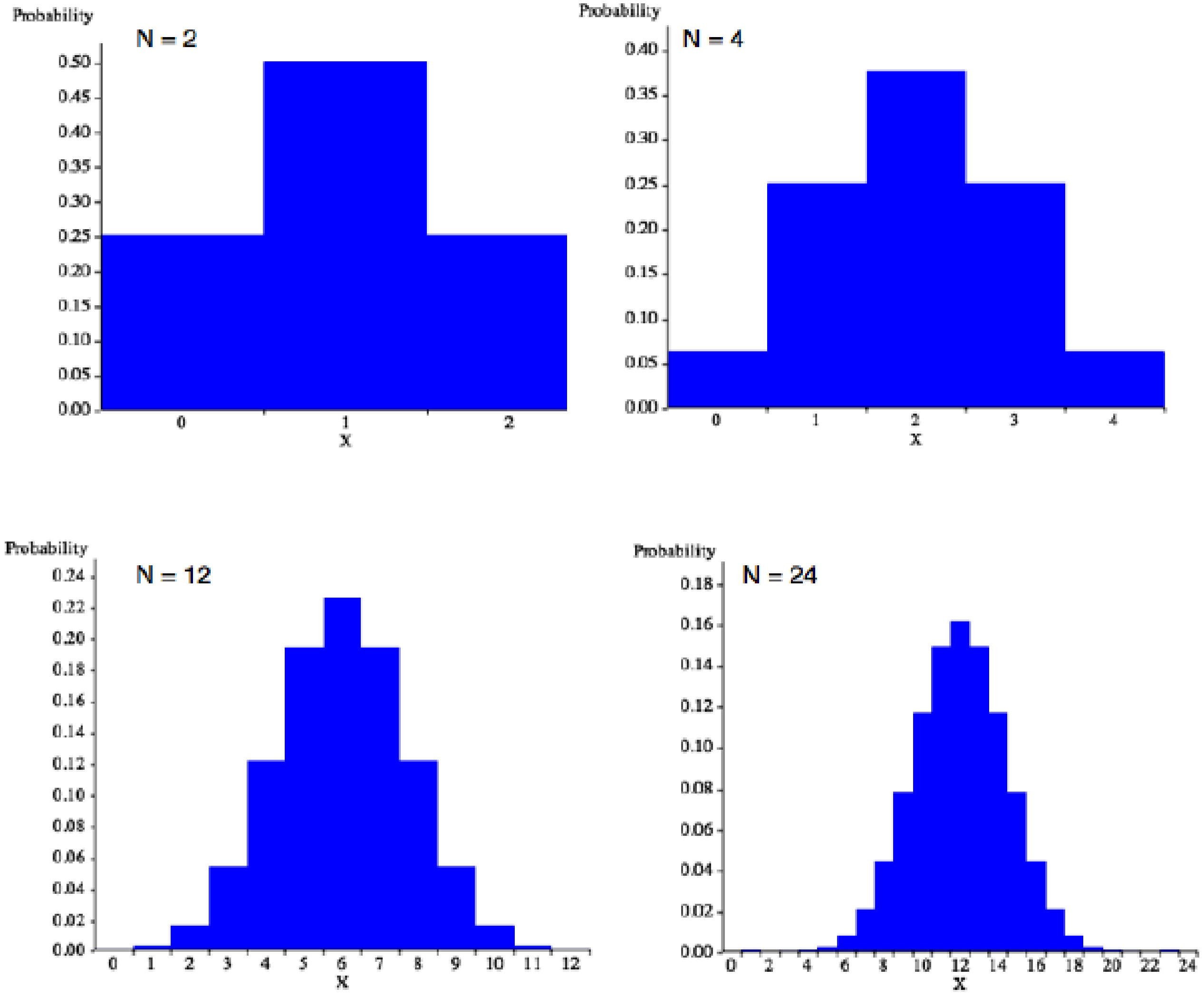
- If a fair coin is flipped 100 times, what is the probability of getting 60 or more heads?
- The probability of exactly x heads out of N flips is computed using the formula:

$$P(x) = \frac{N!}{x!\,(N-x)!}\pi^{x}(1-\pi)^{N-x}$$

- x is the number of heads (60), N is the number of flips (100), and π is the probability of a head (0.5).
- Therefore, to solve this problem, you compute the probability of 60 heads, then the probability of 61 heads, 62 heads, etc.

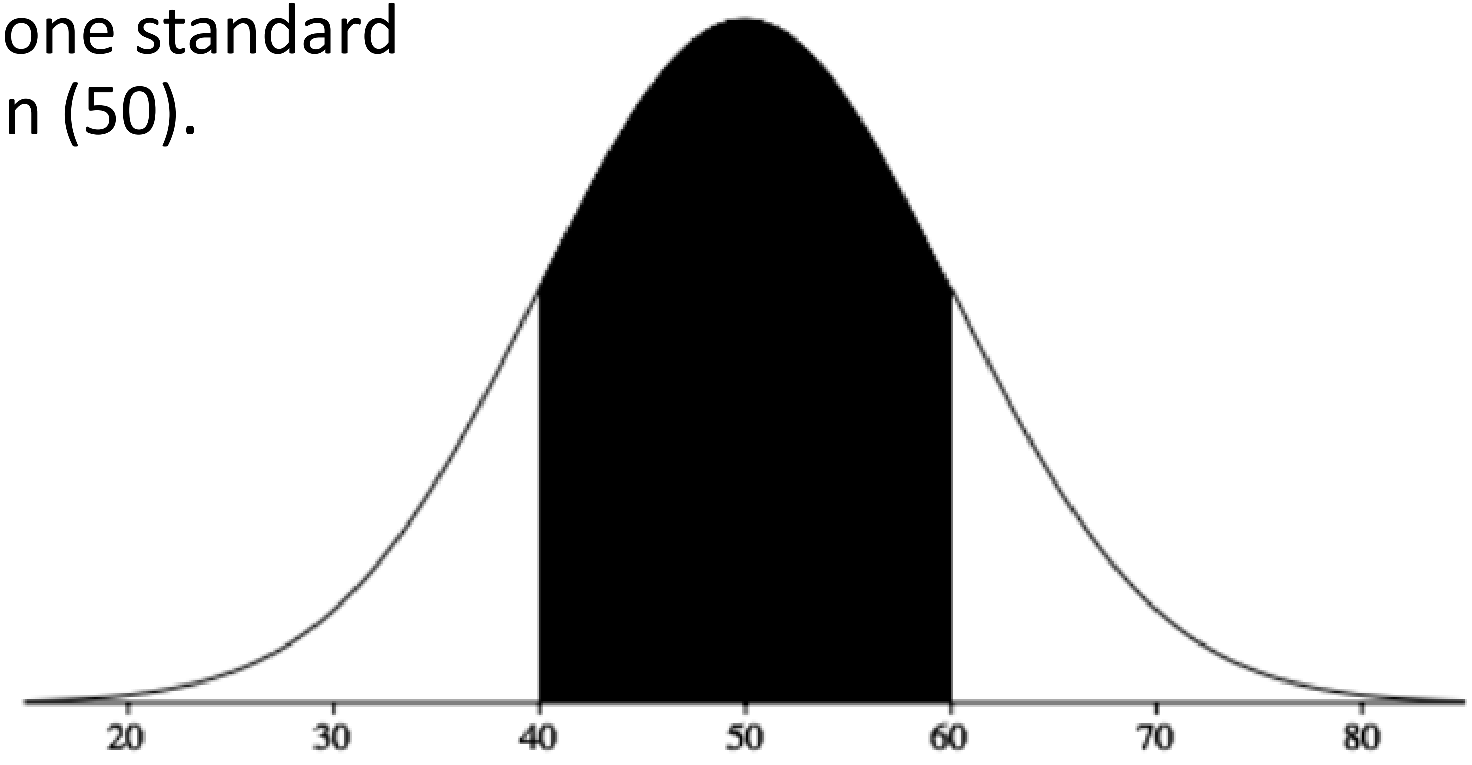
Binomial Distributions

- Abraham de Moivre, an 18th century statistician, was often called upon to make these lengthy computations.
- When the number of events increased, the shape of the binomial distribution approached a very smooth curve.
- Called the “normal curve.”



Areas Under Normal Distributions

- Normal distribution with a mean of 50 and standard deviation of 10.
- 68% of the area is within one standard deviation (10) of the mean (50).



Standard Normal Distribution

- A normal distribution with a mean of 0 and a standard deviation of 1 is called a **standard normal distribution**.
- A value from any normal distribution can be transformed into its corresponding value on a standard normal distribution using the following formula:

$$Z = (X - \mu)/\sigma$$

- As a simple application, what portion of a normal distribution with a mean of 50 and a standard deviation of 10 is below 26?
- Applying the formula:

$$Z = (26 - 50)/10 = -2.4$$

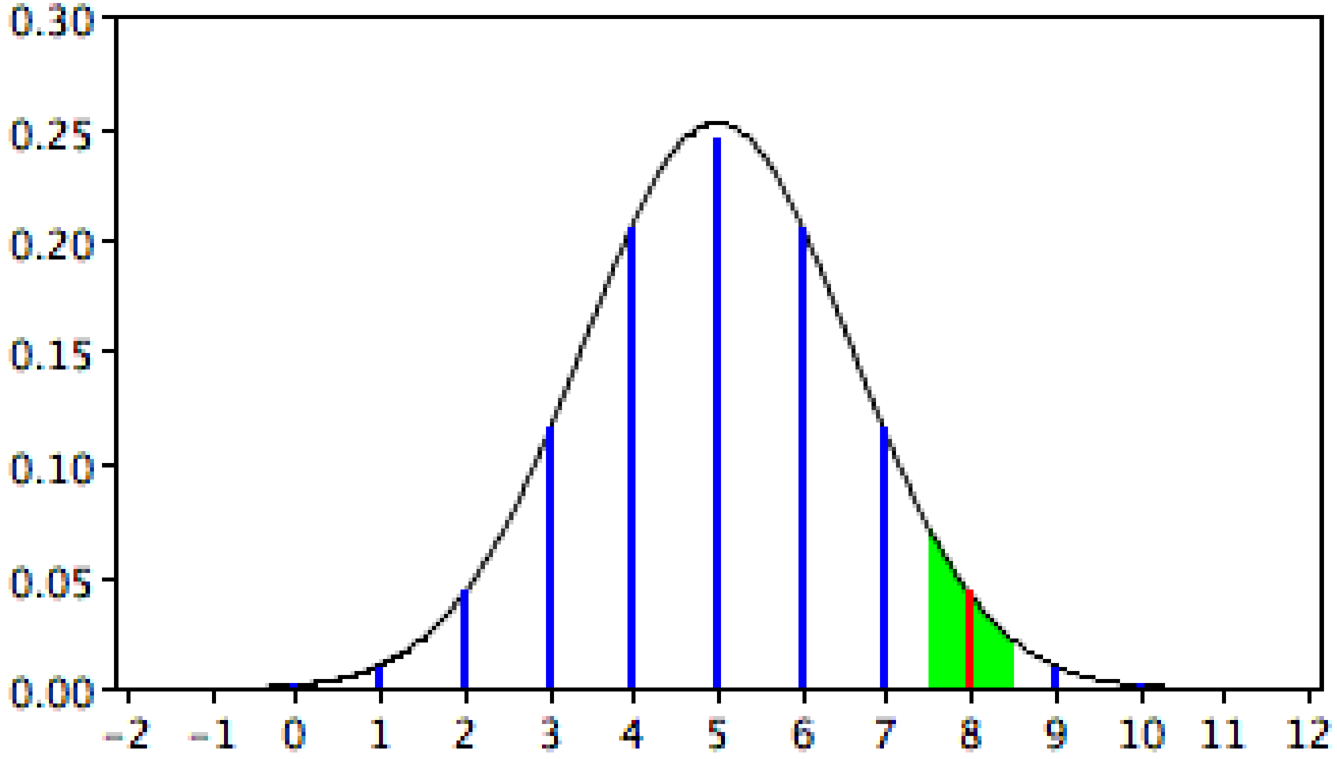
Normal Approximation to the Binomial

- Assume you have a fair coin and wish to know the **probability** that you would get **8 heads out of 10 flips**.
- The binomial distribution has a mean of $\mu = N\pi = (10)(0.5) = 5$ and a variance of $\sigma^2 = N\pi(1-\pi) = (10)(0.5)\ (0.5) = 2.5$
- The standard deviation is therefore $\sigma = 1.5811$

- **Binomial distribution** is a **discrete probability distribution**, whereas the **normal distribution** is a **continuous distribution**.
- The solution is to consider any value from 7.5 to 8.5 to represent an outcome of 8 heads.

Normal Approximation to the Binomial

- Using this approach, figure out the area under a normal curve from 7.5 to 8.5.



Normal Approximation to the Binomial

- You could find the solution using a table of the standard normal distribution (a Z table) as follows:
  1. Find a Z score for 8.5 using the formula Z = (8.5 - 5)/1.5811 = 2.21.
  2. Find the area below a Z of 2.21 = 0.98645.
  3. Find a Z score for 7.5 using the formula Z = (7.5 - 5)/1.5811 = 1.58.
  4. Find the area below a Z of 1.58 = 0.94295.
  5. Subtract the value in step 4 from the value in step 2 to get 0.0435

Sampling Distributions

Introduction

- The concept of a sampling distribution is perhaps the most basic concept in inferential statistics.
- It’s also a difficult concept because a sampling distribution is a theoretical distribution rather than an empirical distribution.

Introduction

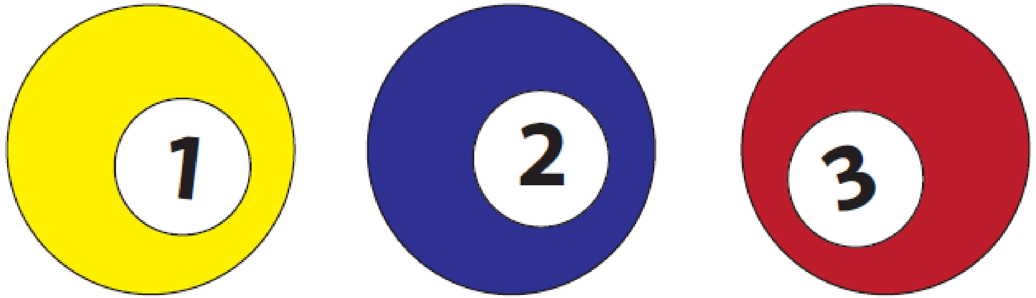
- Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample.
- You would not expect your sample mean to be equal to the mean of all women in Houston, might be lower or it might be higher.
- Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.
- A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population parameter.

Discrete Distributions

- A discrete distribution has a range of values that are countable

Example

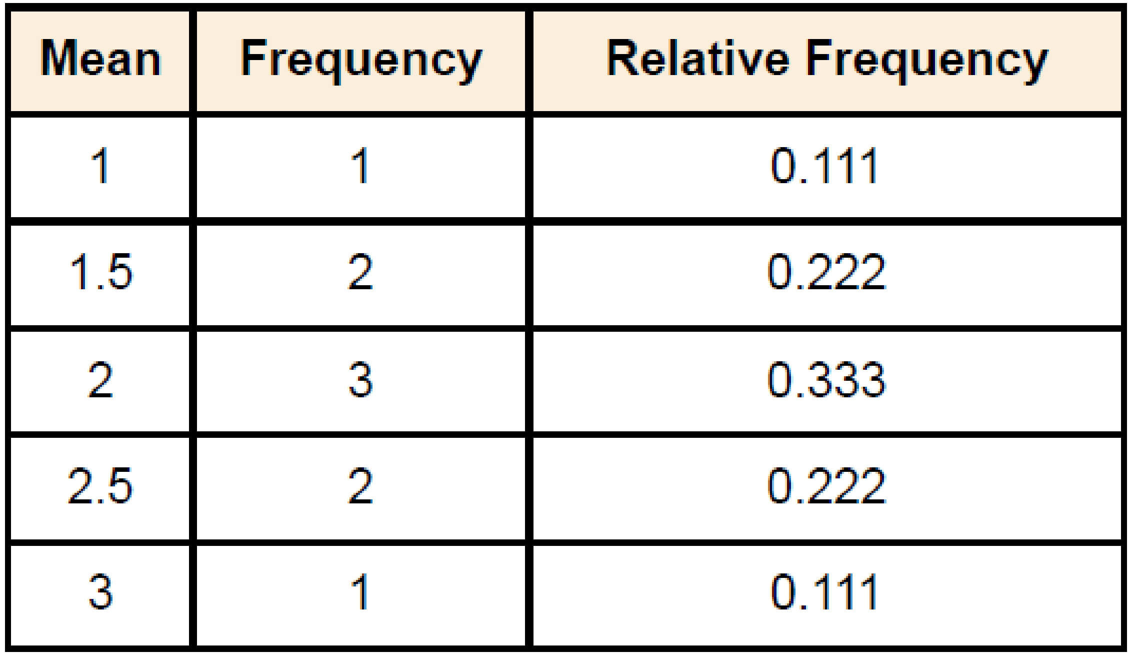
- 3 pool balls, each with a number on it. Suppose 2 of the balls are selected randomly (with replacement) and the average of their numbers is computed.



Outcome	Ball 1	Ball 2	Mean
1	1	1	1
2	1	2	1.5
3	1	3	2
4	2	1	1.5
5	2	2	2
6	2	3	2.5
7	3	1	2
8	3	2	2.5
9	3	3	3

Discrete Distributions

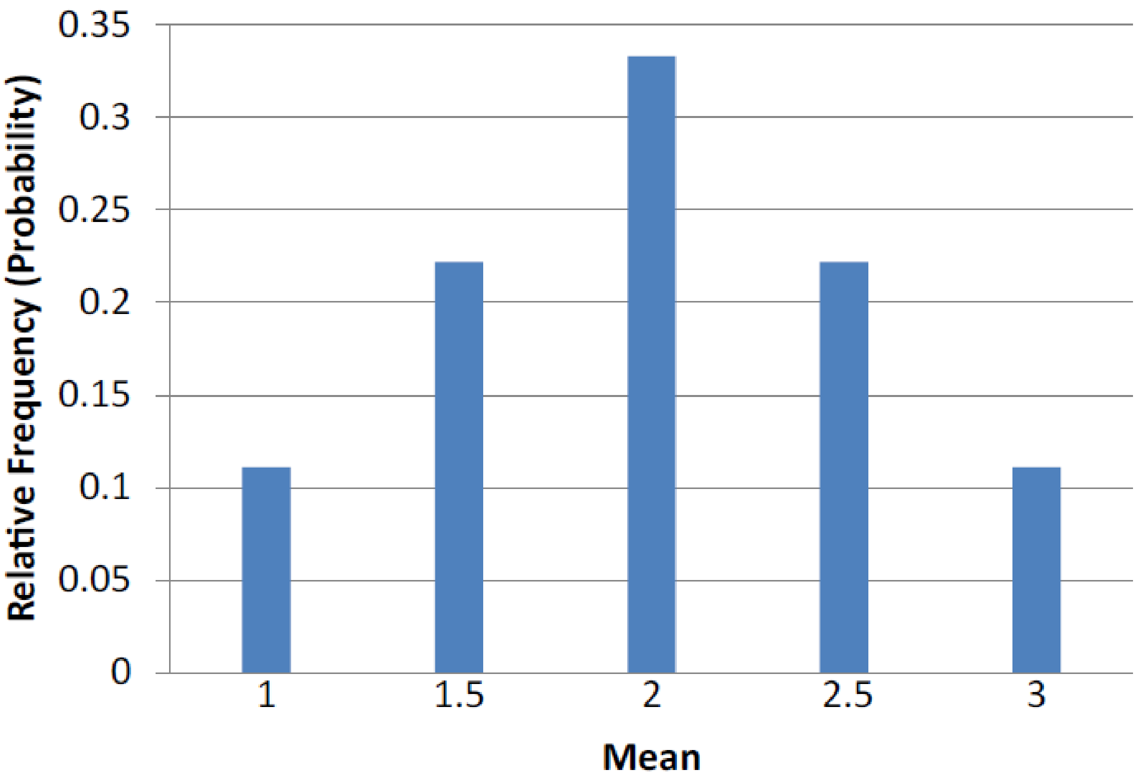
- Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0
- The frequencies of these means are shown in the next table.
- Frequencies of means for N = 2.
- The relative frequencies are equal to the frequencies divided by 9 because there are 9 possible outcomes.



Mean	Frequency	Relative Frequency
1	1	0.111
1.5	2	0.222
2	3	0.333
2.5	2	0.222
3	1	0.111

Discrete Distributions

- Distribution of means for N = 2.
- The distribution shown is called the sampling distribution of the mean.
- For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions.
- The pool balls have only the values 1, 2, and 3, and a sample mean can have 1 of only 5 values.



Continuous Distributions

- What if we had a 1000 pool balls with numbers ranging from 1 to 1.000 in equal steps? (Although this distribution isn’t really continuous, it’s close enough to be considered continuous for practical purposes.) and we sampled 2 balls and computed the mean of these two balls.
- In the previous example, we started by computing the mean for each of the nine possible outcomes.
- This would get a bit tedious for this example since there are 1,000,000 possible outcomes (1,000 for the 1st ball x 1,000 for the 2nd).
- Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency distributions.

Sampling Distribution of the Mean

- The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled.
- Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ.
- The symbol μ_M is used to refer to the mean of the sampling distribution of the mean

$$\mu_M = \mu$$

Sampling Distribution of the Mean

- The variance of the sampling distribution of the mean is computed as follows:

$$\sigma_m^2 = \frac{\sigma^2}{N}$$

- The variance of the sampling distribution of the mean is the population variance divided by N, the sample size.
- Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

Sampling Distribution of the Mean

- The standard error of the mean is the standard deviation of the sampling distribution of the mean. It is therefore the square root of the variance of the sampling distribution of the mean and can be written as:

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

Estimation

Introduction

- One of the major applications of statistics is estimating population parameters from sample statistics.

Example:

- For example, a poll may seek to estimate the proportion of adult residents of a city that support a proposition to build a new sports stadium.
- Out of a random sample of 200 people, 106 say they support the proposition.
- Thus, in the sample, 0.53 of the people supported the proposition.
- This value of 0.53 is called a point estimate of the population proportion.
- It is called a point estimate because the estimate consists of a single value or point.

Point Estimates

- Point estimates are usually supplemented by interval estimates called confidence intervals.

- Confidence intervals are intervals constructed using a method that contains the population parameter a specified proportion of the time.

Point Estimates

Example:

- If the pollster used a method that contains the parameter 95% of the time it is used, he or she would arrive at the following 95% confidence interval: $0.46 < \pi < 0.60$.
- The pollster would then conclude that somewhere between 0.46 and 0.60 of the population supports the proposal.
- The media usually reports this type of result by saying that 53% favor the proposition with a margin of error of 7%.

Degrees of Freedom

- Some estimates are based on more information than others.

- Example: an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5.

- The degrees of freedom (df) of an estimate is the number of independent pieces of information on which the estimate is based.

Degrees of Freedom

Example:

- Let's say the mean height of Martians is 6 and wish to **estimate the variance** of their heights.
- We randomly sample one Martian and find that its height is 8.
- The variance is defined as the mean squared deviation of the values from their population mean.
- This single squared deviation from the mean $(8-6)^2 = 4$ is an estimate of the mean squared deviation for all Martians.
- Therefore, based on this sample of one, we would estimate that the population variance is 4.
- This estimate is based on a single piece of information and therefore has 1 df.

Degrees of Freedom

Example:

- If we sampled another Martian and obtained a height of 5, then we could compute a 2nd estimate of the variance, $(5-6)^2 = 1$.
- We could then average our 2 estimates (4 and 1) to obtain an estimate of 2.5.
- Since this estimate is based on two independent pieces of information, it has 2 df.
- The two estimates are independent because they are based on two independently and randomly selected Martians.
- The estimates would not be independent if after sampling one Martian, we decided to choose its brother as our second Martian.

Estimates

- It’s rare that we know the population mean when we are estimating the variance.
- Instead, we must first estimate the population mean (μ) with the sample mean (M).

Estimates

- Let’s assume we don’t know the population mean and therefore we have to estimate it from the sample.
- We have sampled 2 Martians and found that their heights are 8 and 5.
- Therefore M, our estimate of the population mean, is

$$M = (8+5)/2 = 6.5$$

- We can now compute two estimates of variance:

$$\text{Estimate 1} = (8-6.5)^2 = 2.25$$

$$\text{Estimate 2} = (5-6.5)^2 = 2.25$$

Estimates

- Are these 2 estimates independent?
- The answer is **NO** because each height contributed to the calculation of M.
- Since the first Martian’s height of 8 influenced M, it also influenced Estimate 2.
- The important point is that the 2 estimates aren’t independent and therefore we don’t have 2 degrees of freedom.

Characteristics of Estimators

- 2 important characteristics of statistics used as point estimates of parameters: **bias** and **sampling variability**.
  - Bias refers to whether an estimator tends to either over or underestimate the parameter.
  - Sampling variability refers to how much the estimate varies from sample to sample.

Characteristics of Estimators

- Have you ever noticed that some scales give you very different weights each time you weigh yourself?
- Let's compare two scales.
- **Scale 1** is a very high-tech digital scale and gives essentially the same weight each time you weigh yourself; it varies by at most 0.02 pounds from weighing to weighing. But it’s calibrated incorrectly and, on average, overstates your weight by 1 pound.
- **Scale 2** is a cheap scale and gives very different results from weighing to weighing. Sometimes it vastly overestimates and underestimates it. However, the average of many measurements would be your actual weight.

Characteristics of Estimators

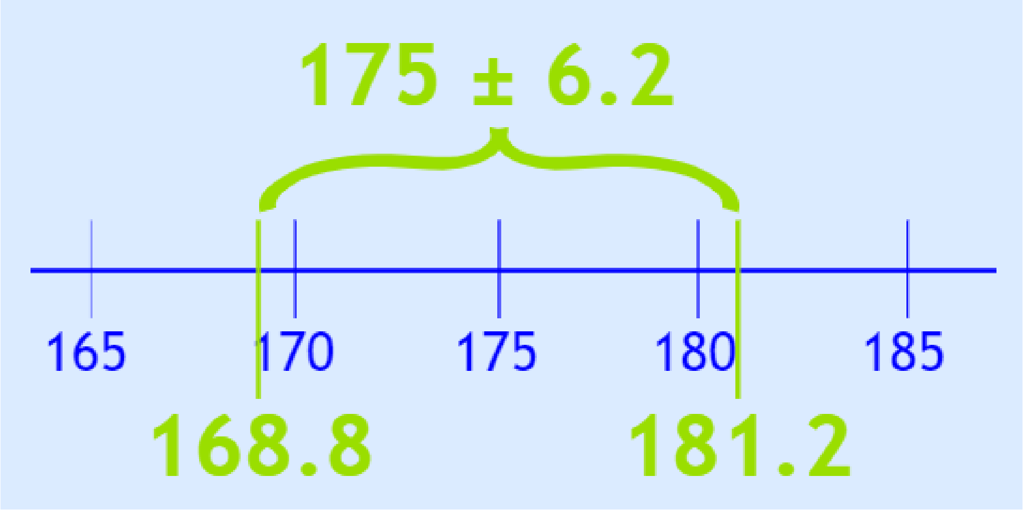
- Scale 1 is biased since, on average, its measurements are 1 pound higher than your actual weight.
- Scale 2, by contrast, gives unbiased estimates of your weight. However, Scale 2 is highly variable, and its measurements are often very far from your true weight.
- Scale 1, despite being biased, is fairly accurate. Its measurements are never more than 1.02 pounds from your actual weight.

Confidence Intervals

- A **Confidence Interval** is a range of values we are fairly sure our true value lies in.

Example:

- We measure the heights of 40 randomly chosen men, and get a mean height of 175cm, the standard deviation of men's heights is 20cm.
- The 95% Confidence Interval is →



Confidence Intervals

- Result says the true mean of ALL men (if we could measure all their heights) is likely to be between 168.8cm and 181.2cm

- But it might not be!
- The "95%" says that 95% of experiments like we just did will include the true mean, but 5% won’t.
- So, there is a 1-in-20 chance (5%) that our Confidence Interval does NOT include the true mean.

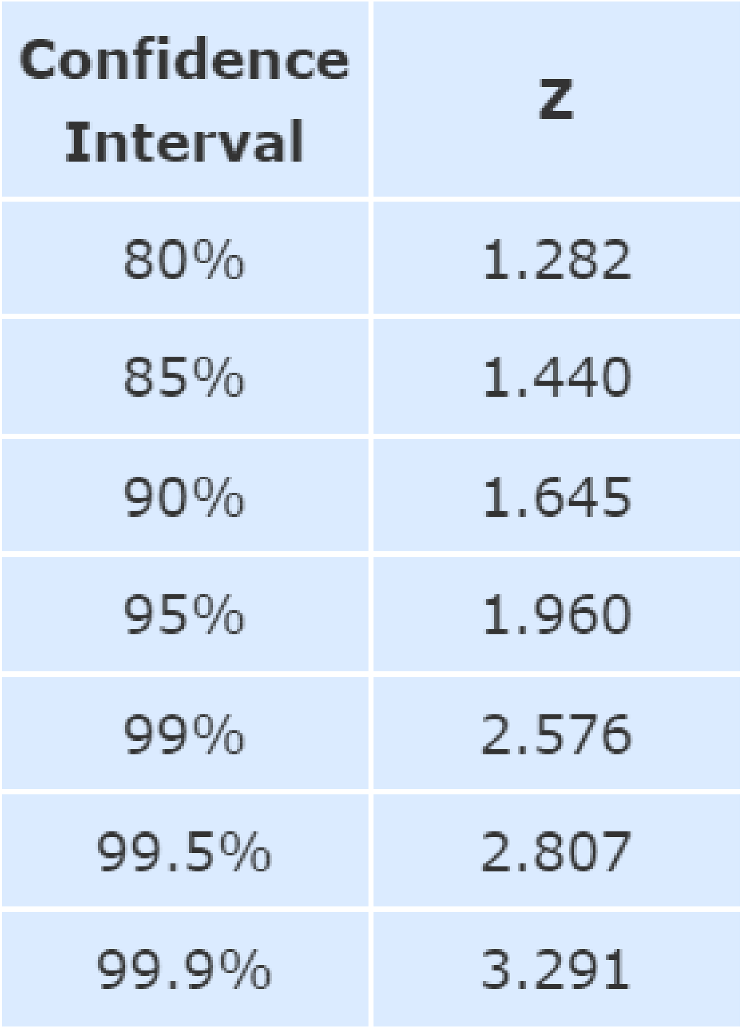
Calculating the Confidence Interval

- **Step 1**: start with
  - the number of observations n
  - the mean X
  - and the standard deviation s

- Using our example:
  - number of observations n = 40
  - mean X = 175
  - standard deviation s = 20

Calculating the Confidence Interval

- **Step 2:** decide what Confidence Interval we want: 95% or 99% are common choices.
- Then find the "Z" value for that Confidence Interval
- For 95% the Z value is 1.960



Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Calculating the Confidence Interval

- **Step 3**: use that Z value in this formula for the Confidence Interval

$$\overline{X} \pm Z\frac{s}{\sqrt{n}}$$

- Where:
  - X is the mean
  - Z is the chosen Z-value from the table above
  - s is the standard deviation
  - n is the number of observations

- And we have: $175 \pm 1.960 \times \frac{20}{\sqrt{40}}$ → 175cm ± 6.20 cm

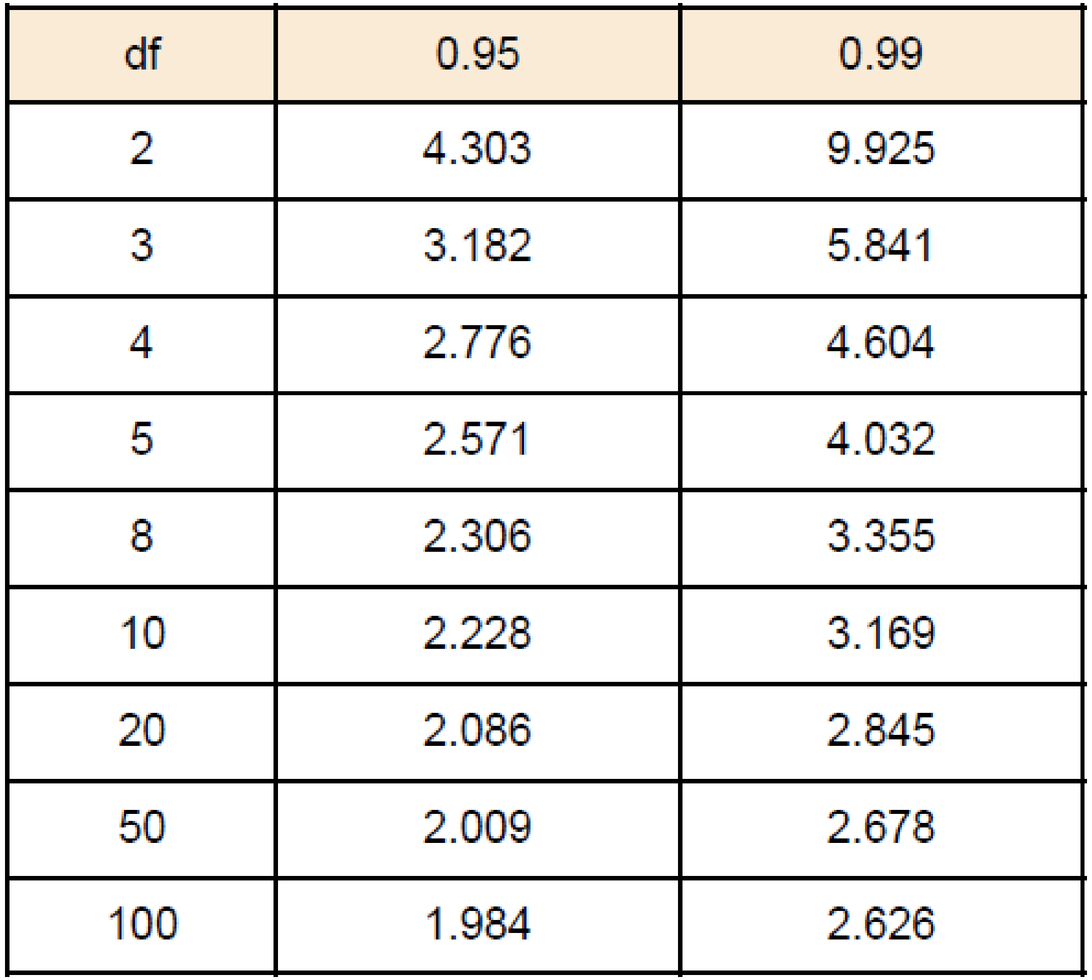
- In other words: from 168.8cm to 181.2 cm

T-Distribution

- The t-Distribution also called the student’s t-distribution and is used while making assumptions about a mean when we don’t know the standard deviation.
- In probability and statistics, the normal distribution is a bell-shaped distribution whose mean is μ and the standard deviation is σ.
- The t-distribution is similar to normal distribution but flatter and shorter than a normal distribution.
- As high as the degrees of freedom (df), the closer this distribution will approximate a standard normal distribution with a mean of 0 and a standard deviation of 1.

T-Distribution

- T-Distribution Table



df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

References

- Witte, R.S.&Witte, J.S. (2017). Statistics (11th ed.). Wiley. ISBN: 978-1119386056.
- Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D.& Zimmer, H. (2003). Introduction to Statistics. Online edition at https://open.umn.edu/opentextbooks/textbooks/459
- Levine, D.M., Stephan, D.F. & Szabat, K.A. (2017). Statistics for Managers Using Microsoft Excel (8th ed.). Pearson. ISBN: 978-0134566672

Thank you