SAMPLE: Early-Semester Content

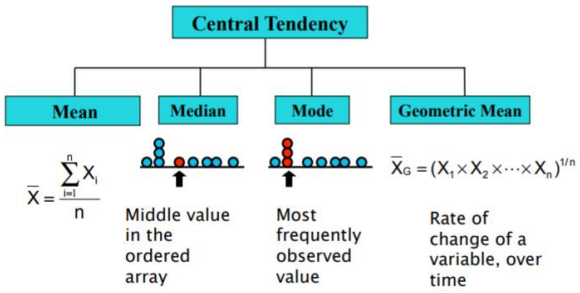
Week 3 - Numerical descriptive measures:

Week 3 LO's:

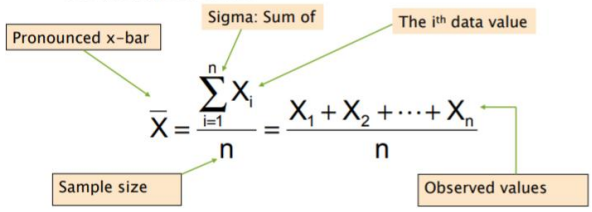
- Central tendency
  - Mean
  - Median
  - Mode
  - Geometric Mean
- Variation and shape
  - Range
  - Variance (Sample)
  - Standard Deviation (Sample)
  - Coefficient of Variation
  - Z-Scores
  - Shape: Skewness and Kurtosis
  - Descriptive Measures Using Excel
  - Quartiles and Boxplot
  - Boxplots using Statcrunch
- Population Measures
  - Mean and Variance
- Rules and Ethical Considerations
  - The Empirical Rule
  - Chebyshev's Rule
  - Ethical Considerations

- Summary of definitions:
  - Central tendency – the extent to which the data values group together around a typical or central value
  - Variation – the amount of dispersion or degree of scattering of values around the central value
  - Shape – the pattern in the distribution of values from the lowest value to the highest value

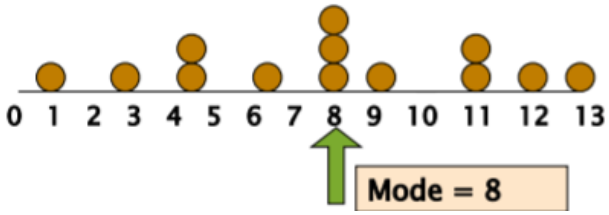
Central tendency:

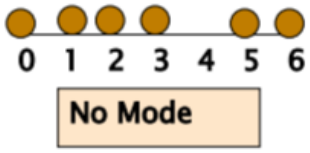


- Mean (arithmetic) – the most common measure of central tendency
  - For a sample of size n:



    - $\sum_{i=1}^{n} X_i$ represents the sum of values of X_i, starting at X_1 and ending with X_n
    - If i=3, it would be sum of values of X starting at X_3 to X_n
  - Can be affected by extreme values (e.g. outliers)
  - Only useful for numerical data
- Median – in an ordered array, the median is the "middle" number (50% above, 50% below)
  - Median position = $\frac{n+1}{2}$ position in the ordered data
    - If the number of values is odd, the median is the middle number
    - If the number of values is even, the median is taken as the average of the two middle numbers
    - Note: this is to determine the position of the median in the ordered array, not the value of the median
  - Useful for all data that has an order
  - Not affected by (a few) extreme values

- Mode – value that has "highest likelihood of occurring"/occurs the most frequently
  - 

  - 

    - There may be no mode or several modes
  - Not affected by extreme values
  - Used for both numerical and categorical data
- Geometric mean – often used to measure the rate of change of a variable over time
  - $$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$
- Geometric mean rate of return – measures the status of an investment over time
  - $$\overline{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$
    - R_t is the rate of return in time period t
  - Example: An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two

$$X_1 = \$100{,}000 \quad X_2 = \$50{,}000 \quad X_3 = \$100{,}000$$

50% decrease 100% increase

Arithmetic mean rate of return: $\overline{X} = \frac{(-.5)+(1)}{2} = .25 = 25\%$ **Misleading result**

Geometric mean rate of return:

$$\begin{aligned}\overline{R}_G &= [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1 \\ &= [(1+(-.5))\times(1+(1))]^{1/2} - 1 \\ &= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More representative result

Which measure to choose:

- Mean:
  - Advantage – most commonly used and easy to calculate, most commonly used
  - Disadvantage – sensitive to outliers and can't be used for categorical data
- Median:
  - Advantage – less sensitive to outliers so it is the next most popular measure
  - Disadvantage – does not consider all the information (extreme values are important)
- Sometimes both the mean and the median are used.
- Mode:
  - Advantage – easy to answer the 'popularity' question
  - Disadvantage – usually reported for discrete or categorical data only, and it is not applicable in all the cases
- Geometric mean (return):
  - Advantage – useful to measure and track percentage changes
  - Disadvantage – cannot be applied to numbers with different signs

SAMPLE: Middle-Semester Content

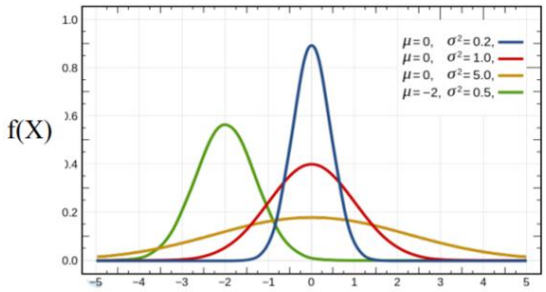
Week 6 – Continuous probability distributions

Week 6 LO's:

- Continuous Probability Distributions Introduction
- Normal Distribution
- Uniform Distribution
- Exponential Distribution

Continuous probability distributions:

- A continuous random variable can assume any value on a continuum/range, e.g.:
  - Thickness, height, weight, volume of an item
  - Time required to complete a task, time between events
  - Temperature
  - Financial return, percentage change
- A continuous random variable can potentially take on any value, depending on the ability to accurately and finely measure it
- Continuous random variables relevant to business:
  - Weight of cans, packets, boxes
  - Download times, web query times
  - Financial return, volume of trades, time between trades
- Continuous probability density:
  - Instead of probabilities for each value of X, a continuous r.v. has a probability density function

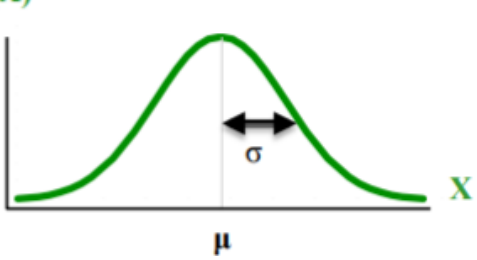


  - Exam note: f(X) it is the relative probability of one point in that area compared to another point in another area vs. discrete distributions where the height of the function represents the true probability
  - Each density curve represents the relative likelihood of each X value
    - The area of each shaded region is the probability that X is in that region
  - We only consider the probability of X being in a certain range/region of values)
    - $$P(a < X < b) = \int_a^b f(x)dx$$
  - This is because there are essentially infinite possible outcomes for a continuous r.v., so the probability of any individual outcome is 0 → area under a single value for X is 0 i.e.:
    - $$P(X = a) = P(a \le X \le a) = \int_a^a f(x)dx = 0$$
  - Total area under the entire density curve is always equal to 1 i.e. if P(a < X < b) = 1, then (a, b) covers all possible values of X
    - $$1 \approx \int_{-\infty}^{\infty} f(x)dx$$

Normal/Gaussian probability distribution:

Definitions:

- Normal probability distribution is a bell-shaped density curve
  - Distribution is therefore symmetric (skewness = 0)
  - Mean = median = mode
  - The random variable has an infinite theoretical range
- Characteristics:
  - The mean parameter is mu (μ)
  - The standard deviation parameter is sigma (σ)
- Shape of the normal distribution curve:
  - Location of distribution is given by μ i.e. changing μ shifts the distribution left or right
  - Spread of distribution is given by σ i.e. changing σ increases or decreases the spread



- Note that most real data are NOT normally distributed
  - Some exceptions of normal distributions:
    - I.Q (constructed to be normal, but very slight right skew)
    - Positions of particles in fluid
    - Sum of many random variables (central limit theorem)
  - A normal probability distribution is mostly assumed as an approximation, but this is sometimes disastrous e.g. CDO pricing → Global Financial Crisis
- Normal density function (not examinable):

The formula for the normal **probability density function** is

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2\right]$$

Where $\exp(1) = e$ = mathematical constant ≈ 2.71828

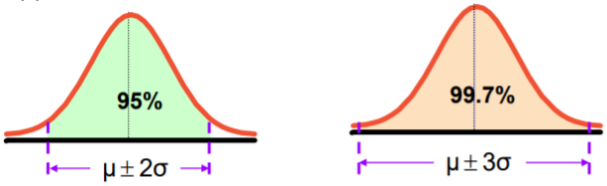
π = mathematical constant ≈ 3.14159

μ = the population mean $E(X)$

σ = the population standard deviation $\text{Var}(X) = \sigma^2$

X = a value of the continuous variable

  - If a value of X is further from the mean, f(X) moves closer to zero
- The 'normal' rules → the normal distribution is exactly bell-shaped, so it follows the "empirical rule":
  - Approx. 68% of the data in a normal distribution is within ± one standard deviation of the mean, i.e. $\mu \pm 1\sigma$
  - Approx. 95% of the data in a normal distribution lies within ± two standard deviations of the mean, or $\mu \pm 2\sigma$
  - Approx. 99.7% of the data in a normal distribution lies within $\mu \pm 3\sigma$



SAMPLE: Late-Semester Content

Week 9 – Hypothesis testing (one sample tests)

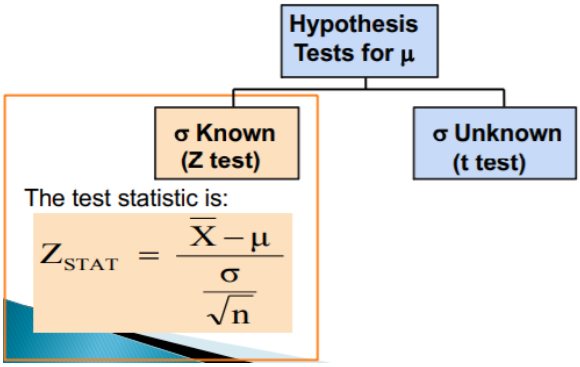
Week 9 LO's:

- Introduction to Hypothesis Testing (last week) – Chapter 9
- Hypothesis tests for the mean – One Sample Tests
  - σ Known, Two-Tail Tests
    - Critical value approach
    - p-value approach
    - Comparison to confidence intervals
  - σ unknown, Two-Tail Tests
    - Critical value and p-value approaches
    - Comparison to confidence intervals
  - One-Tail Tests
    - Critical value approach
    - p-value approach
- Hypothesis tests for the Proportion – One Sample Tests
  - Critical value approach
  - p-value approach
- Two Sample Tests - Chapter 10

DCOVA

Hypothesis testing for the mean when σ is known (two-tail test):

- Use Z tests
- Convert sample statistic ($\bar{X}$) to a Z_{STAT} test statistic

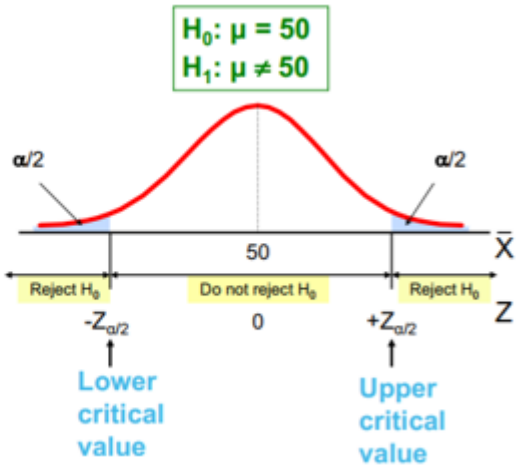


- Assumptions:
  - 1. Population standard deviation (σ) is known
  - 2. Population is normally distributed, or if the population is not normal→ the sample size for our sample means distribution must be sufficiently "large" ($n \geq 30$ for CLT)
  - 3. Also assume that all samples are randomly selected (SRS)

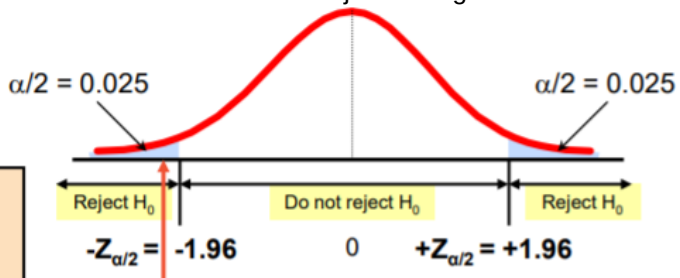
Critical value approach to testing:

- For a two-tail test:

There are two cutoff values (critical values), defining the regions of rejection



- 6 steps in critical value hypothesis testing:
  - 1. State the null and alternative hypotheses, H_0 and H_1
  - 2. Choose level of significance (α) and sample size (n)
  - 3. Determine appropriate test statistic and sampling distribution i.e. appropriate technique
  - 4. Determine critical values and identify rejection and non-rejection regions
    - This is determined by our chosen level of significance (α) by using the Standardized (use table or computer)
  - 5. Collect data and compute test statistic value
    - Determine the test statistic by converting (i.e. standardizing) the sample statistic ($\bar{X}$) to a test statistic (Z_{STAT})
  - 6. Make the statistical decision and state the managerial conclusion
    - If the test statistic falls into the non-rejection region → do not reject (NOT 'accept') the null hypothesis H_0
    - If the test statistic falls into the rejection region: reject the null hypothesis
    - Express the managerial conclusion in the context of the business problem
- E.g. – test the claim that the true mean diameter of a manufactured bolt is 30mm. (Given σ = 0.8):
  - 1. State the null and alternative hypotheses, H_0 and H_1
    - H_0: μ = 30
    - H_1: $\mu \neq 30$ → this is a two-tail test
  - 2. Specify the desired level of significance and the sample size
    - Suppose α = 0.05 and n = 100
  - 3. Determine the appropriate technique
    - σ is known so this is a Z test
  - 4. Determine the critical values
    - For α = 0.05, the critical Z values (Z_{CRIT}) are $\pm$ 1.96
  - 5. Collect the data and compute the test statistic
    - Suppose the sample results are n = 100, $\bar{X}$ = 29.84 (σ = 0.8 is known)
    - The test statistic would be $Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{29.84 - 30}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{0.08} = -2.0$
  - 6. Make the statistical decision and state the managerial conclusion
    - We need to determine if the test statistic is in the rejection region



Reject H_0 if $Z_{STAT} < -1.96$ or $Z_{STAT} > 1.96$; otherwise do not reject H_0

Here, Z_{STAT} = -2.0 < -1.96, so the test statistic is in the rejection region

- Since Z_{STAT} = -2.0 < -1.96, reject the null hypothesis and conclude there is sufficient evidence, at the 5% significance level, that the mean diameter of the manufactured bolts is not equal to 30