Stat319: Probability and Statistics for Engineers and Scientists

Chapter 1 –Descriptive Statistics

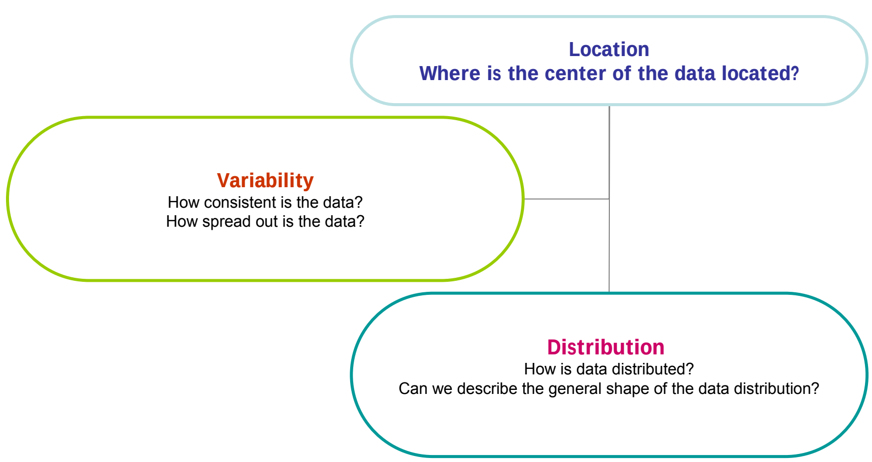
Chapter 1 Topics

- Must Read lab manual chapter 1.
- What is descriptive statistics?
- Measures of Location (Mean, Median, Mode)
  - Definition
  - What they represent?
  - How to compute?
- Percentiles & Quartiles
  - Definition
  - What they represent?
  - How to compute?
- Relationship between Mean & Median
  - Mean = Median → distribution is symmetrical
  - Mean > Median → distribution is skewed (not symmetrical) to the right
  - Mean < Median → distribution is skewed (not symmetrical) to the left

What is descriptive statistics?

- Descriptive statistics
  - Describing data with summary information
- Main focus:
  - How to describe data
- How to describe:
  - Data distribution
  - Data central location
  - Data spread

Descriptive Chapter Overview



Getting to know your data

- How to know much about your data?
- Quick way
  - Do stem-and-Leaf Plot
- Data like a tree
  - Can be broken up into
    - Stem
    - Leaves

Stem and Leaf Plot

Steps to create a Stem & Leaf plot (lab manual,

1) Divide observations into stem and leaf
2) List stems in one column (ascending order of stems)
3) List leaf of each observation in appropriate stem or row
4) Count occurrence of each leaf and tally in "frequency" column

Example for No nitrogen data

observation		
0.32	Step 1)	Stem = First two digits, Leaf = last digit
0.53	Step 2)	place stem in one colum
0.28	Step 3)	place leaf in next colum in the corresponding row for appropriate stem
0.37	Step 4)	Count occurrence of each leaf and tally in "frequency" column
0.47		
0.43		
0.36		
0.42		
0.38		
0.43		

Stem	Leaf	Frequency
0.2	8	1
0.3	2 6 7 8	4
0.4	2 3 3 7	4
0.5	3	1
Total		10

Stem & Leaf Example- Nitrogen Data (Walpole Data from Ex 1.2)

- Steps
  1. Stem= first decimal Leaf=last digit
  2. Place stem in one column in ascending order
  3. Place Leaf in next column in the corresponding row for appropriate Stem
  4. Count occurrence of each Leaf & tally in 'Frequency' column

Observation	
0.26	√
0.43	√
0.47	√
0.49	√
0.52	√
0.75	√
0.79	√
0.86	√
0.62	√
0.46	√

Stem	Leaf	Frequency
0.2	6	1
0.3		
0.4	3 6 7 9	4
0.5	2	1
0.6	2	1
0.7	5 9	2
0.8	6	1
Total		10

Engineering Probability & statistics: A decision making approach

Measures of Location

- Where is the data center located for the sample we are trying to describe?
- Mean = arithmetic average (numerical Average, p.9) $\bar{y} = \frac{1}{n} \sum y$
- Median = the middle of ordered observations (uninfluenced center, p.9)
- Mode= the most frequent observation (Lab p. 19)

Example 1.2 p.9 of Walpole		ordered data	
No nitrogen	Nitrogen	No nitrogen	Nitrogen
X	X	X	X
0.32	0.26	0.28	0.26
0.53	0.43	0.32	0.43
0.28	0.47	0.36	0.46
0.37	0.49	0.37	0.47
0.47	0.52	0.38	0.49
0.43	0.75	0.42	0.52
0.36	0.79	0.43	0.62
0.42	0.86	0.43	0.75
0.38	0.62	0.47	0.79
0.43	0.46	0.53	0.86

Measures of Location

- Where is the data center located for the sample we are trying to describe?
- Mean = arithmetic average (numerical Average, p.9) $\bar{y} = \frac{1}{n}\sum y$
- Median = the middle of ordered observations (uninfluenced center, p.9)
- Mode= the most frequent observation (Lab p. 19)

Example 1.2 p.9 of Walpole			ordered data			
	No nitrogen	Nitrogen		No nitrogen	Nitrogen	
	X	X		X	X	
	0.32	0.26		0.28	0.26	
	0.53	0.43		0.32	0.43	
	0.28	0.47		0.36	0.46	
	0.37	0.49		0.37	0.47	
	0.47	0.52		0.38	0.49	
	0.43	0.75		0.42	0.52	
	0.36	0.79		0.43	0.62	
	0.42	0.86		0.43	0.75	
	0.38	0.62		0.47	0.79	
	0.43	0.46		0.53	0.86	
Total	3.99	5.65	Total	3.99	5.65	
Mean = Total/n	0.399	0.565				
Mode = ?				0.43	None	
Need to order data to get Median:		Median = (X(n/2)+X(n/2+1))/2=		0.400	0.505	for even data
		Median = X(n+1)/2=				for odd data

Sum these values for numerator of Mean

The middle values

No Mode. All values occur once only.

More Example

Table 1.1 The life of 40 car batteries recorded to the nearest tenth of a year.

TABLE 1.1 Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

(Walpole et.al. 2002, 16)

Any value belonging to $\left[2.20-\frac{0.10}{2},\ 2.20+\frac{0.10}{2}\right) = [2.15,\ 2.25)$ is recorded as 2.2

Engineering Probability & statistics: A decision making approach

More Example

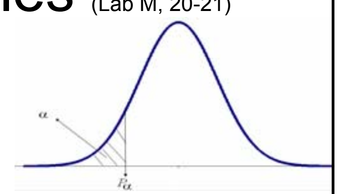
Stem-and-Leaf Plot

Stem	Leaf	f	f/n
1	69	2	
2	25669	5	
3	0011112223334445567778899	25	
4	11234577	8	

Mode for grouped data

Class Interval	Class midpoint	f	f/n
[1, 2)	1.5	2	0.050
[2, 3)	2.5	5	0.125
[3, 4)	3.5	25	0.625
[4, 5)	4.5	8	0.200

Calculating Percentiles (Lab M, 20-21)

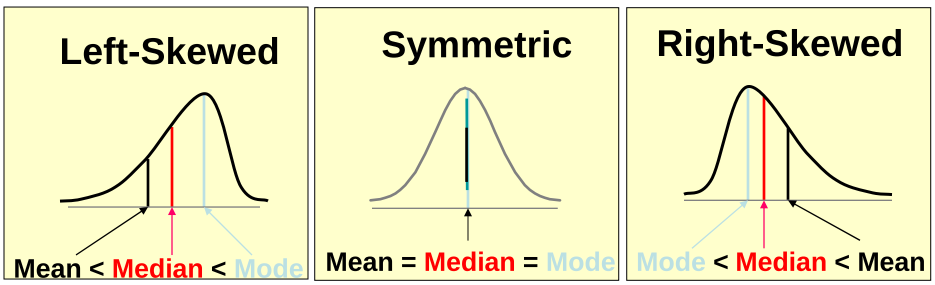


- P_α = value that exceeds α% of data
  - Data position: $R_\alpha = \alpha \dfrac{1+n}{100} = i + d, \quad \alpha = 1, 2, \ldots, 99;$
  - αth Percentile: $P_\alpha = (1-d)y_{(i)} + dy_{(i+1)}$
- Special percentiles (True for any distribution)
  - P_{25}= 25th percentile = 1st quartile (Q_1)
  - P_{50}= 50th percentile = 2st quartile (Q_2) = Median
  - P_{75}= 75th percentile = 3st quartile (Q_3)
- Computing P_{25} (no Nitrogen data)
  - Step 1: Order the observations in ascending order (see below)
  - Step 2. Determine R_{25} = 25(n+1)/100= 25(10+1)/100 = 2.75
  - Step 3.Separate R_{25} = 2.75 into integer (i) and decimals (d):
    - R_{25} = 2.75 = 2 + 0.75
  - Step 4: The 25th percentile is given by
    - $P_{25} = x_{(2)} + 0.75(x_{(3)}-x_{(2)})$
    - P_{25} = 0.32 + 0.75(0.36-0.32)
    - P_{25} = 0.35

ordered data	
	No nitrogen
	X
	0.28
	0.32
	0.36
	0.37
	0.38
	0.42
	0.43
	0.43
	0.47
	0.53

Shape of a Distribution

- Describes how data is distributed
- Symmetric or skewed

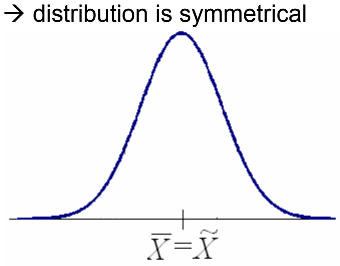


(Longer tail extends to left)

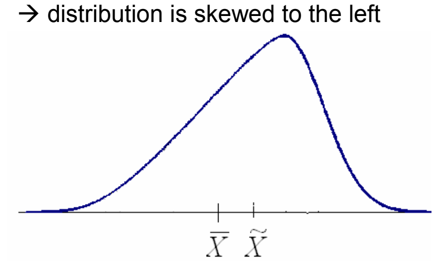
(Longer tail extends to right)

Mean versus Median (symmetric vs skewed)

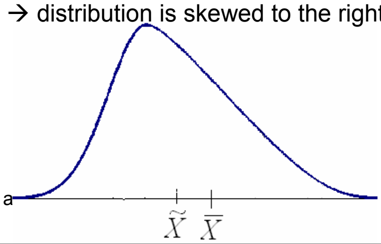
Mean = Median
→ distribution is symmetrical



Mean < Median
→ distribution is skewed to the left



Mean > Median
→ distribution is skewed to the right



Further about location indices

House Prices:
$2,000,000
500,000
300,000
100,000
100,000
Sum 3,000,000

- **Mean:** ($3,000,000/5)
  = **$600,000**
- **Median:** middle value of ranked data
  = **$300,000**
- **Mode:** most frequent value
  = **$100,000**

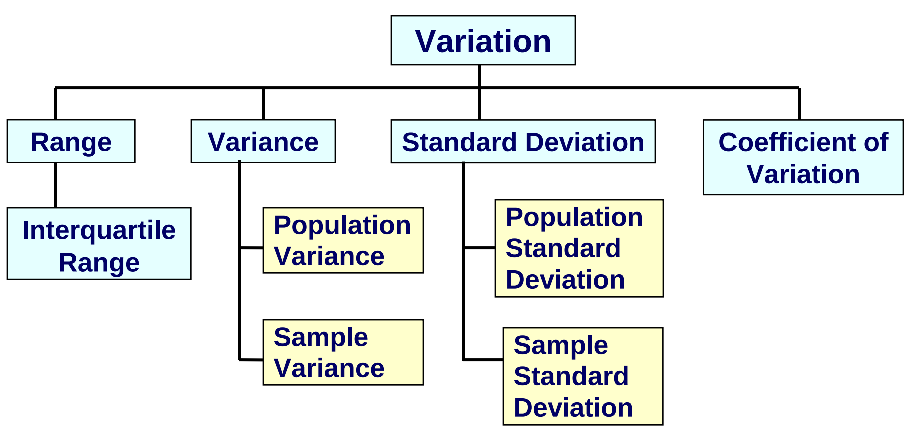
Which measure of location is the "best"?

- **Mean** is generally used, unless extreme values (outliers) exist
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Chapter 1 Topics (cont.)

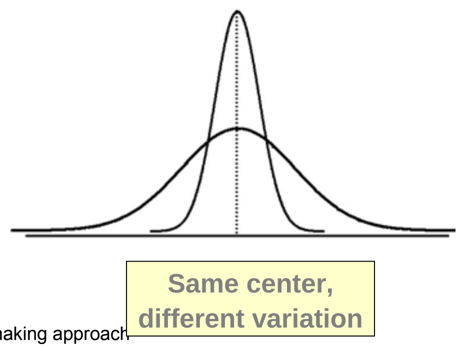
- Measures of Variability (range, variance, Standard Deviation, Interquartile range)
  - Companies want products that are consistent in quality – good for business
    - Profit for manufactured products is a function of process variability
    - Process engineers are responsible for controlling process variability
    - In Chapters 8-15, variability indices play a major role. Very important to remember how to obtain indices, why, and what they represent
  - **Definition**
  - **What they represent?**
  - **How to compute? (Book Example 1.3)**
    - Degrees of freedom = # of **Independent** pieces of **data information** available for computing variability
  - **Why compute? Which variability index is more important?**
    - Depends on situation
      - Inference on variance : variance is important
      - Inference on mean: standard deviation is important

Measures of Variation



Variation

- Measures of variation give information on the **spread** or **variability** of the data values.



Measures of Variability

- Measures of data spread
  - How spread out is the data?
- Range (R) = Max-Min
- Variance = average squared deviation from the mean

$$s^2 = \frac{TSS}{n-1} \text{ where } TSS = \sum (y-\overline{y})^2 = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

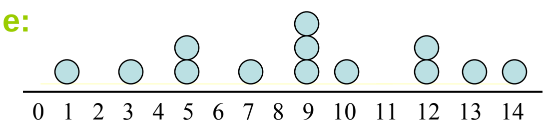
- Standard Deviation (s) = Square root of Variance

Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = x_{maximum} - x_{minimum}$$

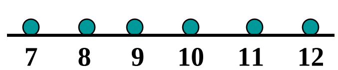
Example:



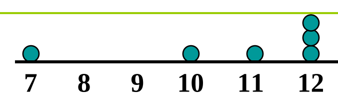
Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

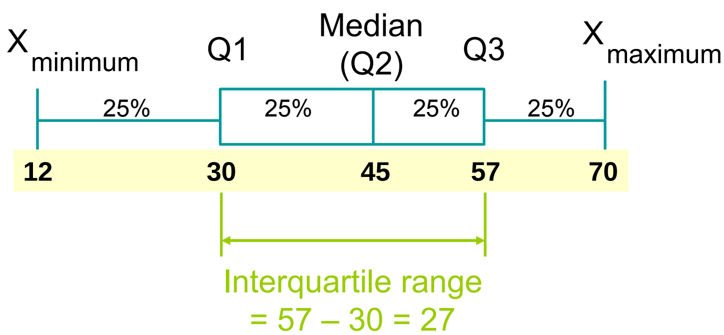
Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high-and low-valued observations and calculate the range from the remaining values.
- Interquartile range = 3rd quartile – 1st quartile

Engineering Probability & statistics: A decision making approach

Interquartile Range

Box-Whiskers Plot (or Box-Plot) Example:



Engineering Probability & statistics: A decision making approach

Variance

- Average of squared deviations of values from the mean
  - **Sample** variance:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1}$$

  - **Population** variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N}$$

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data
  - **Sample** standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

  - **Population** standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \mu)^2}{N}}$$

Calculation Example: Sample Standard Deviation

Sample Data (Y_i) : 10 12 14 15 17 18 18 24

n = 8 Mean = $\bar{y}$ = 16

$$s = \sqrt{\frac{(10-\bar{y})^2 + (12-\bar{y})^2 + (14-\bar{y})^2 + \cdots + (24-\bar{y})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{126}{7}} = 4.2426$$

$$s = \sqrt{\frac{TSS}{n-1}} \text{ where } TSS = \sum (y-\bar{y})^2 = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

Engineering Probability & statistics: A decision making approach

Measures of Variability

Example 1.2 p.9 of Walpole

	No nitrogen					Nitrogen			
	X	X - Mean	$(X - Mean)^2$	X^2		X	X - Mean	$(X - Mean)^2$	X^2
	0.32	-0.079	0.006241	0.1024		0.26	-0.305	0.093025	0.0676
	0.53	0.131	0.017161	0.2809		0.43	-0.135	0.018225	0.1849
	0.28	-0.119	0.014161	0.0784		0.47	-0.095	0.009025	0.2209
	0.37	-0.029	0.000841	0.1369		0.49	-0.075	0.005625	0.2401
	0.47	0.071	0.005041	0.2209		0.52	-0.045	0.002025	0.2704
	0.43	0.031	0.000961	0.1849		0.75	0.185	0.034225	0.5625
	0.36	-0.039	0.001521	0.1296		0.79	0.225	0.050625	0.6241
	0.42	0.021	0.000441	0.1764		0.86	0.295	0.087025	0.7396
	0.38	-0.019	0.000361	0.1444		0.62	0.055	0.003025	0.3844
	0.43	0.031	0.000961	0.1849		0.46	-0.105	0.011025	0.2116
Total	3.99	0.0000	0.047690	1.639700		5.65	0.0000	0.313850	3.506100
Mean = Total/n	0.399					0.565			
Total Sum of Squares (TSS) or S_{xx}									
variance = [total $(X - Mean)^2$]/(n-1)									
standard deviation = square root of variance									
Range = Max - Min									

Engineering Probability & statistics: A decision making approach

Measures of Variability

Example 1.2 p.9 of Walpole

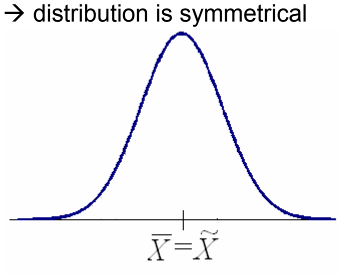
	No nitrogen					Nitrogen			
	X	X - Mean	(X - Mean)^2	X^2		X	X - Mean	(X - Mean)^2	X^2
	0.32	-0.079	0.006241	0.1024		0.26	-0.305	0.093025	0.0676
	0.53	0.131	0.017161	0.2809		0.43	-0.135	0.018225	0.1849
	0.28	-0.119	0.014161	0.0784		0.47	-0.095	0.009025	0.2209
	0.37	-0.029	0.000841	0.1369		0.49	-0.075	0.005625	0.2401
	0.47	0.071	0.005041	0.2209		0.52	-0.045	0.002025	0.2704
	0.43	0.031	0.000961	0.1849		0.75	0.185	0.034225	0.5625
	0.36	-0.039	0.001521	0.1296		0.79	0.225	0.050625	0.6241
	0.42	0.021	0.000441	0.1764		0.86	0.295	0.087025	0.7396
	0.38	-0.019	0.000361	0.1444		0.62	0.055	0.003025	0.3844
	0.43	0.031	0.000961	0.1849		0.46	-0.105	0.011025	0.2116
Total	3.99	0.0000	0.047690	1.639700		5.65	0.0000	0.313850	3.506100
Mean = Total/n	0.399	0.04769/(10-1)				0.565			
Total Sum of Squares (TSS) or S_{xx}				0.04769					0.31385
variance = [total (X - Mean)^2]/(n-1)			0.00529889			3.506100-(5.65)^2/10		0.034872222	
standard deviation = square root of variance			0.07279347					0.186741057	
Range = Max - Min			0.25					0.6	

Chapter 1 Topics (cont.)

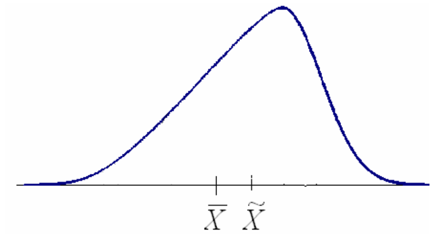
- Concept of relative frequency distribution
  - “ a picture is worth a thousand words (data)”
  - Shape of distribution
    - Symmetrical vs
    - Skewed
      - to the right
      - to the left
  - Number of Modes for distribution
    - One mode –Unimodal
    - Two modes - Bimodal
    - Multiple mode - Multimodal
  - Special distribution – Bell-shaped curve (Normal Curve)
- Empirical Rule
- z-scores
- Coefficient of Variation (C.V.)
- Coefficient of Skewness (C.S.)

Mean versus Median (symmetric vs skewed)

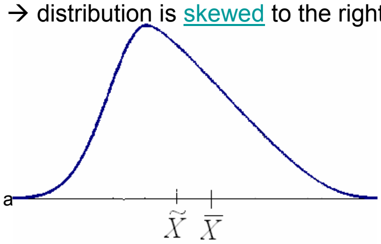
Mean = Median
→ distribution is symmetrical



Mean < Median
→ distribution is skewed to the left



Mean > Median
→ distribution is skewed to the right

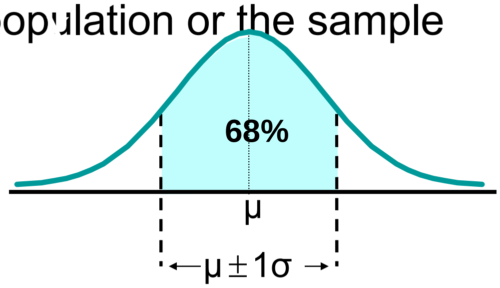


Empirical rule

- Special unimodal symmetrical distribution: Bell shaped (Normal curve)
- Rule is used to determine if data might at a first glance follow the normal distribution
- Rule:
  - Approx 68% of measurement will lie within 1 standard deviation of their mean
  - Approx 95% of measurement will lie within 2 standard deviation of their mean
  - Almost all measurements will lie within 3 standard deviation of their mean
- A population/sample satisfying all 3 properties above is said to satisfy the empirical rule.
  - This however, doesn't guarantee that data come from a normal distribution. (coz: Rule does not mention anything about the mode)

The Empirical Rule

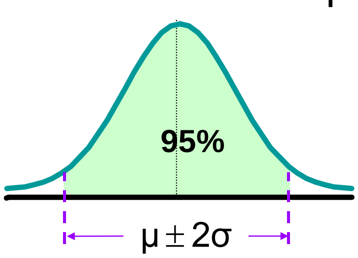
- If the data distribution is bell-shaped, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample

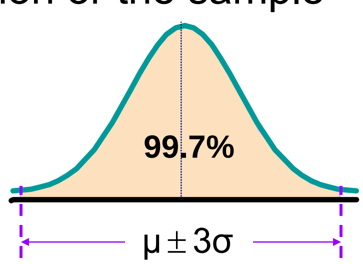


Engineering Probability & statistics: A decision making approach

The Empirical Rule

- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains about 99.7% of the values in the population or the sample





Engineering Probability & statistics: A decision making approach

Example for Empirical Rule

Example 1.2 p.9 of Walpole

	No nitrogen		1	2	3		Nitrogen		1	2	3
	X	$(X-Mean)^2$	Rule1	Rule2	Rule3		X	$(X-Mean)^2$	Rule1	Rule2	Rule3
	0.32	0.006241	Out	In	In		0.26	0.093025	Out	In	In
	0.53	0.017161	Out	In	In		0.43	0.018225	In	In	In
	0.28	0.014161	Out	In	In		0.47	0.009025	In	In	In
	0.37	0.000841	In	In	In		0.49	0.005625	In	In	In
	0.47	0.005041	In	In	In		0.52	0.002025	In	In	In
	0.43	0.000961	In	In	In		0.75	0.034225	In	In	In
	0.36	0.001521	In	In	In		0.79	0.050625	Out	In	In
	0.42	0.000441	In	In	In		0.86	0.087025	Out	In	In
	0.38	0.000361	In	In	In		0.62	0.003025	In	In	In
	0.43	0.000961	In	In	In		0.46	0.011025	In	In	In
Total	3.99	0.047690					5.65	0.313850			
Mean = Total/n	0.399		7/10 or 70%	10/10 or 100%	10/10 or 100%		0.565		7/10 or 70%	10/10 or 100%	10/10 or 100%
Total Sum of Squares (TSS) or S_{xx}											
variance = [total $(X-Mean)^2$]/(n-		0.00529889						0.034872222			
standard deviation = square root		0.07279347						0.186741057			
Mean-k*s	0.399-0.07279=		0.3262	0.2534	0.1806				0.3783	0.1915	0.0048
Mean+k*s	0.399+0.07279=		0.4718	0.5446	0.6174				0.7517	0.9385	1.1252

Z-scores

- Z = (x-Mean)/(standard Deviation)
- Transforms observations into standard deviation units
- Negative z scores: data below mean
- Positive z scores: data above mean
- Magnitude of z score: how far away data is from mean

Measures of Variability

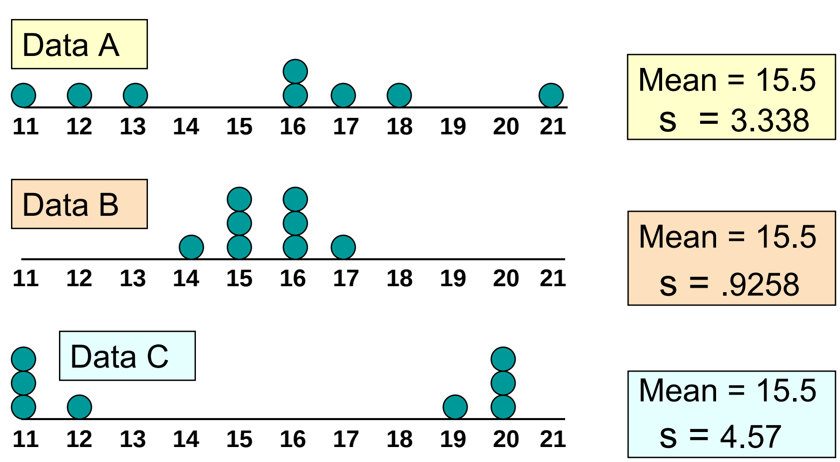
Example 1.2 p.9 of Walpole

	No nitrogen					Nitrogen			
	X	X - Mean	(X - Mean)^2	Z		X	X - Mean	(X - Mean)^2	Z
	0.32	-0.079	0.006241	-1.085		0.26	-0.305	0.093025	-1.633
	0.53	0.131	0.017161	1.800		0.43	-0.135	0.018225	-0.723
	0.28	-0.119	0.014161	-1.635		0.47	-0.095	0.009025	-0.509
	0.37	-0.029	0.000841	-0.398		0.49	-0.075	0.005625	-0.402
	0.47	0.071	0.005041	0.975		0.52	-0.045	0.002025	-0.241
	0.43	0.031	0.000961	0.426		0.75	0.185	0.034225	0.991
	0.36	-0.039	0.001521	-0.536		0.79	0.225	0.050625	1.205
	0.42	0.021	0.000441	0.288		0.86	0.295	0.087025	1.580
	0.38	-0.019	0.000361	-0.261		0.62	0.055	0.003025	0.295
	0.43	0.031	0.000961	0.426		0.46	-0.105	0.011025	-0.562
Total	3.99	0.0000	0.047690	0.000		5.65	0.0000	0.313850	0.000
Mean = Total/n	0.399					0.565			
Total Sum of Squares (TSS) or S_{xx}									
variance = [total (X - Mean)^2]/(n-1)			0.00529889					0.034872222	
standard deviation = square root of variance			0.07279347					0.186741057	
Range = Max - Min			0.25					0.6	

(0.43-0.399)/0.07279

Engineering Probability & statistics: A decision making approach

CV: Comparing Standard Deviations



Engineering Probability & statistics: A decision making approach

Coefficient of Variation

- Measures relative variation
- Sometimes in percentage (%)
- Shows variation relative to mean
- Is used to compare two or more sets of data measured in different units

Population

$$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100\%$$

Sample

$$CV = \left(\frac{s}{\bar{x}}\right) \cdot 100\%$$

Coefficient of variation (CV)

- Relates variability in sample to the mean

$$CV = s / \bar{y}$$

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Coefficient of Skewness (CS)

- Indicates direction of the relative frequency distribution either
  - Skewed to lower values (left)
  - Skewed to higher values (right)
  - Symmetrical

$$CS = \frac{\bar{y} - \tilde{y}}{s/3}$$

- **Negative value of CS: Negative skewed/Skewed Left/left tailed distribution**
- **Positive value of CS: Positive skewed/Skewed Right/Right tailed distribution**
- **CS = 0 : Symmetrical distribution**

Examples

	X	X - Mean	$(X - Mean)^2$			X	X - Mean	$(X - Mean)^2$	
	0.32	-0.079	0.006241			0.26	-0.305	0.093025	
	0.53	0.131	0.017161			0.43	-0.135	0.018225	
	0.28	-0.119	0.014161			0.47	-0.095	0.009025	
	0.37	-0.029	0.000841			0.49	-0.075	0.005625	
	0.47	0.071	0.005041			0.52	-0.045	0.002025	
	0.43	0.031	0.000961			0.75	0.185	0.034225	
	0.36	-0.039	0.001521			0.79	0.225	0.050625	
	0.42	0.021	0.000441			0.86	0.295	0.087025	
	0.38	-0.019	0.000361			0.62	0.055	0.003025	
	0.43	0.031	0.000961			0.46	-0.105	0.011025	
Total	3.99	0.0000	0.047690			5.65	0.0000	0.313850	
median	0.4					0.505			
Mean = Total/n	0.399					0.565			
Total Sum of Squares (TSS) or S_{xx}									
variance = [total $(X - Mean)^2$]/(n-1)			0.00529889					0.034872222	
standard deviation = square root of variance			0.07279347					0.186741057	
Range = Max - Min			0.25					0.6	
Coeff of variation = (std deviation)/mean				0.182					0.331
Coeff of Skewness =(Mean-Median)/(std devation)/3				-0.0412					0.9639

Stem & Leaf Example- Nitrogen Data (Walpole Data from Ex 1.2 -Review)

- Steps
  1. Stem= first decimal Leaf=last digit
  2. Place stem in one column in ascending order
  3. Place Leaf in next column in the corresponding row for appropriate Stem
  4. Count occurrence of each Leaf & tally in 'Frequency' column

Observation	
0.26	√
0.43	√
0.47	√
0.49	√
0.52	√
0.75	√
0.79	√
0.86	√
0.62	√
0.46	√

Stem	Leaf	Frequency
0.2	6	1
0.3		
0.4	3 6 7 9	4
0.5	2	1
0.6	2	1
0.7	5 9	2
0.8	6	1
Total		10

Stem-and Leaf Information

- Gives the shape of the distribution
- No nitrogen data
  - Skewed right distribution

Chapter 1 Topics (cont.)

- Graphical Methods and Data Description
  - " a picture is worth a thousand words (data)"
  - Stem and leaf plot (p16-17)
  - Frequency distributions
    - Frequency tables (p. 18 & Lab M, pp.10-12)
    - Graphical displays
      - Frequency Histogram (p. 12 & Lab M, pp. 18-19)
      - Frequency plots (Lab M, pp.13-15)
        - » plot
        - » Polygon
        - » Smoothed frequency curves (p. 19)
      - Cumulative Frequency plot (Lab M, pp.13-15)
      - And Relative Frequency equivalents
    - Box-plot (lab M, p. 24) & Outlier detection (Inner & Outer fences)
    - Other graphs
      - Bar Chart (for discrete & Qualitative data, Lab M, pp.15-17)
      - Pie chart (for qualitative data, Lab M, pp. 17-18)
      - Scatterplot (for ordered bivariate data, X and Y, p352): will be discussed further in chap 11

Why Use Frequency Distributions?

- A frequency distribution is a way to summarize data
- The distribution condenses the raw data into a more useful form...
- and allows for a quick visual interpretation of the data

Engineering Probability & statistics: A decision making approach

Frequency Distribution: Discrete Data

- Discrete data: possible values are countable

Example: An advertiser asks 200 customers how many days per week they read the daily newspaper.

Number of days read	Frequency
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30
Total	**200**

Engineering Probability & statistics: A decision making approach

Relative Frequency

Relative Frequency: What proportion is in each category?

Number of days read	Frequency	Relative Frequency
0	44	.22
1	24	.12
2	18	.09
3	16	.08
4	20	.10
5	22	.11
6	26	.13
7	30	.15
Total	**200**	**1.00**

$$\frac{44}{200} = .22$$

22% of the people in the sample report that they read the newspaper 0 days per week

Engineering Probability & statistics: A decision making approach

Frequency Distributions

Example 1.2 p.9 of Walpole Frequency Distribution

No nitrogen				Nitrogen			
X	count	frequency	cum. Freq	X	count	f	cum. Freq
0.32				0.26			
0.53				0.43			
0.28				0.47			
0.37				0.49			
0.47				0.52			
0.43				0.75			
0.36				0.79			
0.42				0.86			
0.38				0.62			
0.43				0.46			
Total							

STOP! Wrong! Data MUST be SORTED in increasing order first

Engineering Probability & statistics: A decision making approach

Frequency Distributions for No Nitrogen Data

0.32
0.53
0.28
0.37
0.47
0.43
0.36
0.42
0.38
0.43

Relative Frequency
= Frequency/n

X	Tally	Frequency	Cumulative Frequency	Relative Frequency	Relative Cumulative Frequency
0.28	/	1	1	0.10	0.10
0.32	/	1	2	0.10	0.20
0.36	/	1	3	0.10	0.30
0.37	/	1	4	0.10	0.40
0.38	/	1	5	0.10	0.50
0.42	/	1	6	0.10	0.60
0.43	//	2	8	0.20	0.80
0.47	/	1	9	0.10	0.90
0.53	/	1	10	0.10	1.00
Total		**10**		1.00	

If n >30 data, we may have too many rows in the frequency distribution. We need to do something to improve our frequency distribution. We need grouped frequency distributions.

Engineering Probability & statistics: A decision making approach

Grouped Frequency Distributions

Example 1.2 p.9 of Walpole Grouped Frequency Distribution

No nitrogen X	count	frequency	cum. Freq	relative frequency	relative cum. Freq	Nitrogen X	count	f	F	rf	rF
(0.25-0.30]	1	1	1	0.10	0.10	(0.25-0.30]	1	1	1	0.10	0.10
(0.30-0.35]	1	1	2	0.10	0.20	(0.30-0.35]		0	1	0.00	0.10
(0.35-0.40]	111	3	5	0.30	0.50	(0.35-0.40]		0	1	0.00	0.10
(0.40-0.45]	111	3	8	0.30	0.80	(0.40-0.45]	1	1	2	0.10	0.20
(0.45-0.50]	1	1	9	0.10	0.90	(0.45-0.50]	111	3	5	0.30	0.50
(0.50-0.55]	1	1	10	0.10	1.00	(0.50-0.55]	1	1	6	0.10	0.60
						(0.55-0.60]		0	6	0.00	0.60
						(0.60-0.65]	1	1	7	0.10	0.70
						(0.65-0.70]		0	7	0.00	0.70
						(0.70-0.75]		0	7	0.00	0.70
						(0.75-0.80]	11	2	9	0.20	0.90
						(0.80-0.85]		0	9	0.00	0.90
						(0.85-0.90]	1	1	10	0.10	1.00
Total		10						10			

Even without calculating variance, the nitrogen data is more variable.

$$\text{Class Width} \simeq \frac{\text{Range}}{\text{Number of Classes}}$$

But, the best number of classes for a set of data is √n. That is for this data square root of 10 = 3.16228 or 3.

So class width = 0.08

keeping same class width for the nitrogen data and having only 3 classes for the no nitrogen data gives us the following.

Example 1.2 p.9 of Walpole Grouped Frequency Distribution

No nitrogen X	count	frequency	cum. Freq	relative frequency	relative cum. Freq	Nitrogen X	count	f	F	rf	rF
(0.25-0.33]	11	2	2	0.20	0.20	(0.25-0.33]	1	1	1	0.10	0.10
(0.33-0.41]	111	3	5	0.30	0.50	(0.33-0.41]		0	1	0.00	0.10
(0.41-0.49]	1111	4	9	0.40	0.90	(0.41-0.49]	1111	4	5	0.40	0.50
(0.49-0.57]	1	1	10	0.10	1.00	(0.49-0.57]	1	1	6	0.10	0.60
						(0.57-0.65]	1	1	7	0.10	0.70
						(0.65-0.73]		0	7	0.00	0.70
						(0.73-0.81]	11	2	9	0.20	0.90
						(0.81-0.89]	1	1	10	0.10	1.00
Total		10						10			

Engineering Probability & statistics: A decision making approach

General Guidelines

- Lab Manual (p.10):

$$\text{number of classes} = \sqrt{n}$$

- Distributions with numerous observations are more likely to be smooth and have gaps filled since data are plentiful
- Class Width
  - Class widths can typically be reduced as the number of observations increases

$$\text{Class Width} \simeq \frac{\text{Range}}{\text{Number of Classes}}$$

Battery Life Example

Table 1.1 The life of 40 car batteries recorded to the nearest tenth of a year.

TABLE 1.1 Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

(Walpole et.al. 2002, 16)

Any value belonging to $\left[2.20 - \frac{0.10}{2},\ 2.20 + \frac{0.10}{2}\right) = [2.15,\ 2.25)$ is recorded as 2.2

Grouped Frequency Example for Battery Life data

Range $= 4.7 - 1.6 = 3.1$, No. of Classes: $\sqrt{40} \approx 6.32$

Class Width: $3.1/6 \approx 0.52$, (Walpole et.al. 2002, 16)

1.5-1.9 is to the nearest 0.1 year. So rounded battery lives of 1.45 to 1.95 will be included in this interval

Interval	Midpoint	f	f/n	F	F/n
1.5—1.9	1.7	2	0.05	2	0.050
2.0—2.4	2.2	1	0.025	3	0.075
2.5—2.9	2.7	4	0.100	7	0.175
3.0—3.4	3.2	15	0.375	22	0.550
3.5—3.9	3.7	10	0.250	32	0.800
4.0—4.4	4.2	5	0.125	37	0.925
4.5—4.9	4.7	3	0.075	40	1.000

80th percentile is 3.9 years, that is **80% of the batteries** have lifetimes **less than 3.95** years (since the lifetimes are recorded to the nearest 10th). **Total lifetimes** of the batteries that have lifetime between 2.95 years and 3.45 years is **3.2 ×15 =48** approximately.

Chapter 1 Topics (cont.)

- Mean, Variance, and Percentiles of Grouped Data
  - Approximate: lose precision
  - But sometimes when you don’t have any other information or choice, losing some precision is a small price to pay

Mean, Variances and Percentiles of Grouped data

Mean** for **grouped data:

$$\bar{y} = \frac{1}{n}\sum yf$$

Variance** for **grouped data:

$$s^2 = \frac{TSS}{n-1}, \quad TSS = \sum y^2 f - \frac{1}{n}\left(\sum yf\right)^2$$

Percentiles** for **grouped data:

Can obtain from the relative cumulative frequency column directly or by the following modifications to the percentile formula

$$d = \frac{\frac{\alpha}{100} - rF_j}{rF_{j+1} - rF_j}, \quad \alpha = 1, 2, \ldots, 99;$$

rF_j = relative cumulative frequency for the j^{th} class

$$P_\alpha = (1-d)y_{(j)} + dy_{(j+1)}$$

Mean & Variances from Grouped Data

Interval*	Midpoint	f	f/n
1.5—2	1.75	2	0.05
2.0—2.5	2.25	1	0.025
2.5—3	2.75	4	0.100
3.0—3.5	3.25	15	0.375
3.5—4	3.75	10	0.250
4.0—4.5	4.25	5	0.125
4.5—5	4.75	3	0.075

*Lower limit included

The following quantities are calculated from the above frequency distribution:

$$\sum y = (1.75)(2) + (2.25)(1) + \cdots + (4.75)(3) = 138.5,$$
$$\bar{y} = 138.5/n \approx 3.4625,$$
$$\sum y^2 f = (1.75)^2(2) + (2.25)^2(1) + \cdots + (4.75)^2(3) = 498.5,$$
$$TSS = 498.5 - (138.5)^2/n = 18.94375,$$
$$s^2 \approx 0.485737179,$$
$$s \approx 0.696948476$$

	Original	Grouped
$\bar{y}$	3.4125	3.4625
s	0.7028	0.6969

Determining Percentiles from Relative Cumulative Frequency Distributions

Determination of Percentiles from CRF Distribution

To derive 75[th] percentile, we proceed as follows: **Let q = 75[th] percentile**

rF	x
0.55	3.4
0.75	q
0.80	3.9

$$\frac{q-3.4}{0.75-0.55}=\frac{3.9-3.4}{0.80-0.55},$$

$$q=3.8,$$

Graphical Methods

Stem and Leaf Plot: NoNitro (Exam1p2)
NoNitro
one leaf=1 case

stem°leaf (leaf unit=.1000000, e.g., 6°5 = .6500000)	Class n	Percentiles
2° 8	1	
3° 2	1	
3° 678	3	25%
4° 233	3	median
4° 7	1	75%
5° 3	1	
5°	0	
min = .2800000 max = .5300000 Total N:	10	

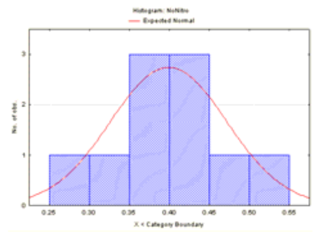
Frequency table: NoNitro (Exam1p2)

From To	Count	Cumulative Count	Percent	Cumulative Percent
.2500000<=x<.3000000	1	1	10.00000	10.0000
.3000000<=x<.3500000	1	2	10.00000	20.0000
.3500000<=x<.4000000	3	5	30.00000	50.0000
.4000000<=x<.4500000	3	8	30.00000	80.0000
.4500000<=x<.5000000	1	9	10.00000	90.0000
.5000000<=x<.5500000	1	10	10.00000	100.0000
.5500000<=x<.6000000	0	10	0.00000	100.0000
Missing	0	10	0.00000	100.0000

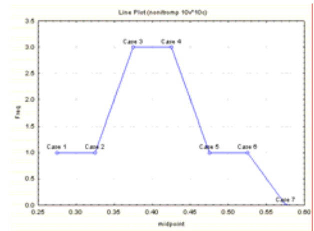
- Stem and Leaf Plot (STATISTICA)
- Frequency Tables (STATISTICA)

Graphical Methods

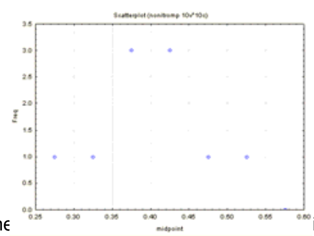
Histogram



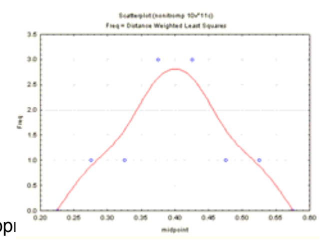
Frequency Polygon



Frequency Plot

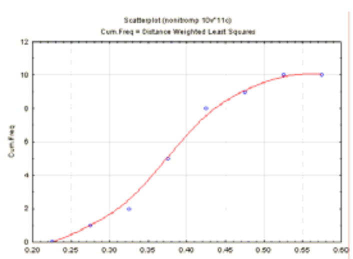


Smoothed Frequency Curve (Distribution)



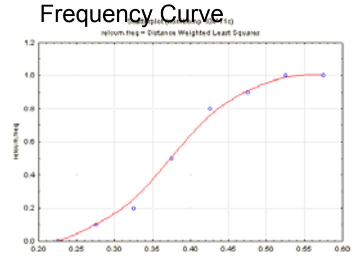
Graphical Methods

Smoothed Cumulative Frequency Curve

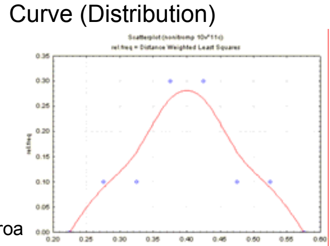


The Relative frequency equivalents of the previous plots can also be used.

Smoothed Relative Cumulative Frequency Curve

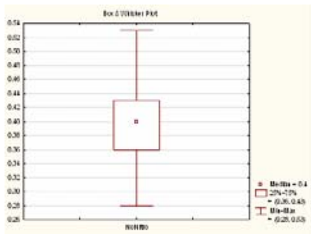


Smoothed Relative Frequency Curve (Distribution)

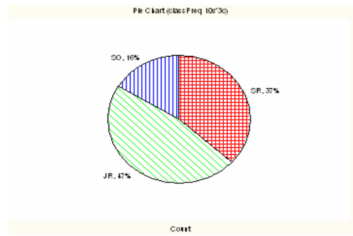


Graphical Methods

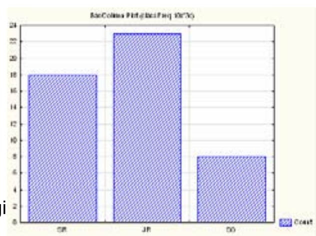
Box and Whisker plot



Pie Chart (for qualitative data)



Bar Chart (for discrete and qualitative data)



Scatterplot (for bivariate data - X,Y data) – only in Final exam (chap 11)