**Stats 2B03 | September 6, 2016 | Lecture 1**

Prof: Aaron Childs

* Google him for class website
* Password is 14-digit student barcode, not student number
* Labs begin on Monday, September 12

Overview & Types of Data [1.1 & 1.2]

* The population is the complete collection of all subjects that are being studied
  + Always use population data sets
    - i.e. Survey of Mac students
* A sample is a group of subjects, selected from the population
  + Falls under descriptive statistics: consists of organizing, summarizing, and presenting data
    - More info. in chapter 2
  + Inferential Statistics: consists of using a sample to draw conclusions about a population
    - More info. in chapter 6
* How to use probability
  + Chapter 3 – 5
* Parameter: A measurement describing some characteristic of a population
* Statistic: A measurement describing some characteristic of a sample
  + Same thing as parameter but instead of a population, it’s a sample
    - i.e. Population of all Mac students
      * Average age is a parameter (μ)
      * Take a sample of 10 students and find average = (X-Bar)
        + This is a statistic
* Quantitative Data: Consists of numbers representing counts or measurements
  + i.e. Weight, age, temperature
* Qualitative (Or Categorical) Data: Cannot be measured, but can only be separated into different categories
  + i.e. Eye color, exercise/do not exercise
* Levels of measurement
  + The nominal level of measurement is categorized by data that consists of names, labels, or categories only – where the categories can’t be ordered
    - i.e. Eye color: Green, blue, brown
      * No order, just names
  + Data are at the ordinal level of measurement if they can be arranged in some order, but differences between data values either can’t be determined or are meaningless
    - i.e. Course letter grades
    - i.e. Exercise / Do not exercise
  + Interval level is like ordinal but differences are meaningful. But there is not a natural zero starting point
    - i.e. Temperature
      * 0 degrees doesn’t mean no temperature, it’s just a level of temperature. Temperature doesn’t start anywhere
        + However, things like height and time do have natural starting points.

**Stats 2B03 | September 8th, 2016 | Lecture 2**

* Continued From Last Lecture…
  + The ratio level is the interval level but with a natural zero starting point
    - i.e. Weight, age, distance traveled to work
      * Refer to table 1.1 (pg. 9 of textbook)

Frequency Distributions [2.2]

* Example: 55, 63, 72, 41, 87, 75, 64, 60
  + Small data sets don’t need to be organized, but large sets do
* Frequency Table
  + Grouping: Freq.: Cumulative Freq.: Relative Freq.: Cumulative %:
  + 40 – 50 1 1 1/8 = 12.5 12.5
  + 50 – 60 1 2 12.5 25
  + 60 – 70 3 5 3/8 = 37.5 62.5
  + 70 – 80 2 7 2/8 = 25 87.5
  + 80 – 90 1 8 12.5 100
    - Convention: When the class boundaries overlap, the right class boundary is not included in the interval
      * i.e. In the freq. table above, 60 will not be in the [50-60] group, but it will be in the [60-70] group.
    - Cumulative Frequency: The cumulative frequency for the data set [60-70] means that 5 people got a mark of 69 or less (<70)
    - Relative Frequency: Divide the freq. of the grouping by total number of frequencies
    - Cumulative Percent: Write down relative frequency as a percent, but add the previous relative frequency to the next one

Visualizing Data [2.3]

* Histogram: Effective because you don’t lose data and data can be easily inferred
* Stem & Leaf Plots:
  + Example: 55, 63, 72, 41, 87, 75, 64, 60
    - Groupings: [40-50], [50-60], etc.

4 | 1 Median = (63 + 64) / 2 = 63.5

5 | 5

6 | 034 🡨 This is an example of a steam & leaf plot. If you were to rotate this

7 | 25 on it’s side (CCW) by 90 degrees, it would look like a histogram.

8 | 7 CCW = Counter-Clockwise

* The Median X is the middle value of an ordered data set
  + Median = {The middle, if n is an odd value

OR average of the two numbers, if n is an even value

* + - i.e. 5, 8, 10, 21, 25 (n=5 [odd])
      * X = 10
    - i.e. 6, 12, 15, 18, 26, 21 (n=6 [even])
      * X = (15+18) / 2 = 16.5

4 | 1 1 🡨 Leaf unit = 1.0 [leaf unit is the number you multiply the data set with]

5 | 5 2 🡨 Number of observations/data entries in this row and lower

6 | 034 (3) 🡨 # of observations in the row containing the median; symbolized by ( )

7 | 25 3 🡨 Number of observations/data entries in this row and lower

8 | 7 1 🡨 Depths (pg.21 of lab manual)

* Leaf unit allows you to recreate the data from the plot. It’s kind of useless with whole numbers, but comes in handy with decimals

Reading this steam & leaf plot, as is, implies that the data set is 30 and 32. The leaf unit is added so you have to multiply it by the data set. 0.1 x 30 = 3.0… 0.1 x 32 = 3.2

* Data value = Value from plot [x] leaf unit
  + i.e. 2.8, 2.9, 3.0, 3.2, 5.4, 6.7, 6.9
    - Leaf Unit = 0.1

2 2 |89

(2) 3| 02

3 4|

3 5| 4

2 6 | 79

Pro tip: If the data set has double decimal digits, then the leaf unit is 0.01. Furthermore, on the stem & leaf plot, you would separate the numbers with a space. And for numbers with only one decimal place, you add a zero at the end.

i.e. 2.8 🡪 2.80

i.e.

* i.e. 55, 63, 72, 41, 87, 75, 64, 60
  + Grouping: [40-45], [45-50], [50-55], etc.

**Stats 2B03 | September 9th, 2016 | Lecture 3**

1 4 | 1

1 4 |

1 5 |

2 5 | 5 🡨 Number of entities in this row and below is 2

(3) 6 | 034 🡨 Number of entities in this row is 3

3 6 | 🡨 Number of entities in this row and everything below it is 3

3 7 | 2 🡨 Same with this row

2 7 | 5

1 8 |

1 8 | 7

Measure of Center [2.4]

* The average of a sample x1,x2,… xn is called the sample mean (or just the mean), and is denoted by (X with a line over it; X-Bar]
  + i.e. X = (X1 + X2 + … + Xn) / n = ∑ Xi / n} i=1
* The average of the population is called the population mean and is denoted by μ (mu)
  + i.e. μ = ∑ Xi / N} where i=1
    - Note that N is the population size
* The mode of a data set is the value that occurs most often
  + i.e. 7, 8, 8, 8, 9, 9, 10
    - 8 is the mode because it occurs the most often
  + i.e. 7, 8, 8, 8, 9, 9, 9, 10 ,10, 11
    - Modes: 8 & 9
      * This data set is bimodal
* Mean from a frequency distribution
  + i.e. Estimate the mean
    - Grouping Frequency

40 – 50 1 🡨 The average of 40 and 50 is 45

50 – 60 1 🡨 (50 + 60) / 2 = 55

60 – 70 3 🡨 And so on, and so forth

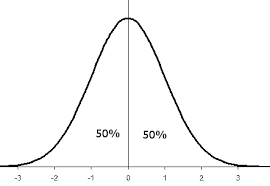
70 – 80 2 🡨 75

80 – 90 1 🡨 85

* + You can estimate the original data set using the middle of the classes:
    - i.e. 45, 55, 65, 65, 65, 75, 85
      * Then find the mean:
      * = (45 + 55 + [65x3] + [2x75] + 85) / 8 = 66.25
        + Note that:

≈ ∑ (f **.** x) / ∑f

Where x = midpoint and f = frequency

* Distribution Shapes
  + Can be normal or bell-shaped
    - i.e. 
  + Right skewed
    - i.e. 
  + Left skewed
    - i.e. 

Measures of Variation [2.5]

* The range R of a data set is the largest value (XL) minus the SA, smallest value (XS)
  + i.e. (R = XL - XS)
* The sample variance (or just variance) of a data set X1, X2, Xn is defined by:
  + S2 = (X1 – )2 + (X2 – )2 +…+ (Xn – )2 / n – 1

= ∑ (Xi – )2 / (n-1)

NEVER EVER ROUND

* + And is also equal to:

Use this for calculations

* + - S2 = n (X12 + X22 +…+ Xn2) – (X1 + X2 +…+ Xn)2

= n ∑Xi2 – (∑Xi)2 / n(n – 1)

* The population variance of a population of values X1, X2…XN is defined by
  + S2 = ∑ (xi – μ)2 / N, where i=1
    - i.e. 55, 63, 72, 41, 87, 75, 64, 60
      * R = 87 – 41 = 49 🡨 Range is the largest value – smallest
      * = 64.625

S2 has no real interpretation. The units are totally different from the data set; they’re squared.

* + - S2 = (55-64.625)2 + (63 – 64.625)2 + (60 – 64.625)2 / 8 – 1

= 191.125

* + - S2 = 8 (552 + 632 +…+ 602) – (55 + 63 + … + 60)2 / 8(8 – 1)

= 191.125

* The sample standard deviation (or just standard deviation) is defined:
  + S = √S2  🡨 S is equal to the square root of S2
    - And in our example, S = √191.125 = 13.82
      * This give us the “average” amount that the data values differ from (X-Bar)
        + For instance, if is average mark of a test, then the typical student is 13.82% away from the average/

**Stats 2B03 | September 13th, 2016 | Lecture 4**

* Research Ethics Lecture
  + Presented by McMaster Research Ethics Board [MREB]
    - Dr. Karen Szala-Meneok
  + Responsible for reviewing all non-medical research involving human participants carried out by McMaster faculty, students, and staff to ensure the safety and well-being of human participants
    - Established due to the unethical of scientists and doctors in the past
      * i.e. Nazi scientists using Jewish people, and exposing them to extreme circumstances, ending in death
      * i.e. CIA funding multiple researches on mind control. McGill university was given thousands of dollars
  + All researches and studies must follow three golden rules:
    - 1. All patients must be treated with the utmost respect
      * The participants must be seen as humans, and not numbers
      * Everyone partaking in the study must sign and agree to the terms
    - 2. Complete anonymity must be maintained throughout the research
      * Any information linking data to the participants or researchers must be removed and stripped
    - 3. Researchers must stick to the original guidelines approved by MREB
      * Modifications to the experiment during the study is strictly prohibited. Any modification must be approved by the board

**Stats 2B03 | September 15th, 2016 | Lecture 5**

* Practical applications of
  + Population 1 VS Population 2
    - for population 1 🡪 gives a large S2
      * is not a good estimator of μ
    - for population 2 🡪 gives a small S2
      * is a good estimate of μ
  + When using μ to estimate a population’s average, it is best to take a population sample where gives a small S2 value
    - S2 is a measure of how good is an estimator of μ
* Range rule of thumb
  + S ≈ range/4
    - i.e. For test 1
      * S ≈ 80/4 ≈ 20
* The coefficient of variation is a relative measure of variation that can be used to compare the variation of two different data sets, possibly measured in different units
  + It is defined by:
    - CV = (S/) x 100% 🡨 This is not a percentage
  + i.e.
    - (A) Ages in years: 3, 4, 5, 6
      * 1 = 4.5, S2 = 5/3
    - (B) The same ages in months: 36, 48, 60, 72
      * 2 = 54, S2 = 240
    - Which data set as more variation? The data set is the same, it’s just their age in months instead of years
      * A) C.V. = (5/3 / 4.5)0.5 = 28.69
      * B) C.V. = (240/54)0.5 = 28.69
        + Note: The C.V. is independent of the scale of measurement
* Epiricical Rule: For a distribution that is roughly bell-shaped, approximately 68% of the observations are within one standard deviation of the mean (See figure 2-13, pg. 62)
  + i.e. Are between – S and + S
    - i.e. If average of a test is 70 and S = 20 (as calculated above), 68% of the marks are between 50 and 90
* ≈ 95% are within 2 standard deviations of the mean
  + i.e. Between – 2S and + 2S
    - i.e. Same as above but 95% of the marks are between 110 and 30
* ≈ 99.7% between – 3S and + 3S

**Refer to image below and**

**Note:**

(95-68)/2 = 27/2 = 13.5

(99.7-95)/2 = 4.7/2 = 2.35 (not 2.4%)

(100-99.7)/2 = 0.3/2 = 0.15% (not 0.1%)



* Example: IQ’s are known to be bell-shaped mean 100 and standard deviation is 15
  + A) Approximately 95% of people have an IQ between \_\_\_\_\_ and \_\_\_\_\_?
    - 100 – 2(15) and 100 + 2(15)
      * 70 and 130
  + B) What percentage have an IQ between 85 and 130
    - 85 = – S
    - 130 = + 2S
      * 34 + 34 + 13.5 = 81.5%

Measures of Relative Standing Percentiles & Quartiles [2.6]

* The Kth percentile is the number Pk with the property that approximately K percent of the data is less Pk and (100 – K) % of the data is bigger than Pk
  + i.e. The 50th percentile is the median
* Method for finding the Kth percentile:
  + 1. Put the data in increasing order
  + 2. Find L = (K/100) x n
    - L is the locator or position
  + 3. A) If L is not a whole number, then round L up. Then Pk is the Lth value in the data set.
  + 3. B) Otherwise, Pk is the average of the Lth and (L+1)st value
    - When L is a whole number
  + See figure 2-15 on p.73

**Stats 2B03 | September 16th, 2016 | Lecture 6**

* The quartiles [Q1, Q2, Q3] divide the data into roughly four equal parts
  + Q1 is the 25th percentile (P25)
  + Q2 is the 50th percentile (P50)
  + Q3 is the 75th percentile (P75)
    - Refer to image below



* i.e. 15, 13, 6, 5, 12, 50, 22, 18, 17, 10, 5, 6, 10, 12, 13, 15, 17, 18, 22, 50
  + Q1 = P25
  + L = (25/100) x 10 = 2.5 🡨 Since L is not a whole number, you round up

= 3 🡨 2.5 rounds to 3

* + Q1 = 10
  + Q2 = P50
  + L = (50/100) x 10 = 5
  + Q2 = (13 + 15)/2 = 14
  + Q3 = P75
  + L = (75/100) x 10 = 7.5 = 8
  + Q3 = 18
* Quartiles are for big data sets to calculate the spread of data
  + i.e. For test marks, if Q3 is 82, it means that 75% of the class got a mark less than 82% and 25% got a mark above 82%
* The interquartile range is defined by: IQR = Q3 – Q1
  + This is the range for the middle 50% of the data
* Outlier: An unusually large or small data value compared with the rest of the data values
  + Any data value smaller than Q1 – 1.5[IQR] or larger than Q3 + 1.5[IQR] is considered to be an outlier
    - Outliers help determine if the data is a real observation or a mistake
      * i.e. If everyone got 80% on a test, and one guy got 2%, he’s an outlier, because he’s far far far from the rest ot the data
      * i.e. If some guy inputs his height at 10 feet tall, he’s an outlier, and the data can be completely removed because its obviously a mistake on his behalf
  + Example (Refers to previous data set. Found below Quartile image)
    - IQR = 18 – 10 = 8
    - Q1 – 1.5[IQR] = 10 – 1.5(8) = -2
    - Q3 + 1.5[IQR] = 18 + 1.5(8) = 30
      * Therefore, 50 is an outlier, because it is very large

Exploratory Data Analysis (EDA) [2.7]

* Boxplots: The five number summary consists of the values: min, Q1, Q2, Q3, max
  + Min, Q1, Q2, Q3, Max
    - You add all these numbers up and take the average
      * Refer to image below for an example of a boxplot



* + If quartiles are more far apart, there’s more variation in the data
  + When quartiles are close together, there’s less variation in the data
* Modified Boxplot
  + Let a1 be the largest observation that is NOT an outlier
    - i.e. a1 = 22
  + Let a2 be the smallest value that is NOT an outlier
    - i.e. a2 = 5
      * Data set is found below the quartiles image
  + If you have any outliers, you draw stars around the boxplot, like so:

See figure on p.81

* + - Modified boxplots are more informative because they don’t account for outliers, hence the data is clean and undisrupted
      * i.e. If only one kid got a really high mark on a test, a boxplot would extend the arms to that arm, giving an inconsistent reading. A modified boxplot removes this error by removing outliers

Fundamentals & Addition Rules [3.2 + 3.3]

* Definition: An event is any collection of results or outcomes of an experiment
  + i.e. In a family of 3 children, let A be the event “exactly two boys”
    - A = {BBG, BGB, GBB}
  + A simple event is an outcome or event that cannot be broken down into simpler components
    - B = “All Girls” = {GGG}
      * This is a simple event because it cannot be broken down further
        + Event “A” is not a simple event because there are 3 possibilities
  + The sample space (S) for an experiment is the set of all simple events
    - i.e. In a family of three children, the sample space is:
      * S = {BBB, GBB, BGB, BBG, GGB, GBG, BGG, GGG}
    - The following is not a simple event because it can be broken down further; S = {0, 1, 2, 3} 🡨 # of boys
      * This can be broken down further. If the family had two boys, then the order of those boys, in which they are conceived is variable
* Basic Rule Of Probability:
  + If the outcomes in “S” are equally likely then the probability of A is:
    - P(A) = [# of outcomes in A] / [# of outcomes in S]

= [# of ways A can occur] / [total # of possible outcomes]

* + - i.e. (Data is found above)
      * P (exactly 2 boys) = 3/8
      * P (all girls) = 1/8
        + However, this cannot be used for all data sets because some data sets are to big. Hence, we need some rules…
  + Some rules:
    - Rule P1: 0 ≤ P(A) ≤ 1
    - Rule P2: P(A) = 1 – P()
      * Where , “A complement” is the event that “A” does not occur
        + Note that is the opposite of “A”

**Stats 2B03 | September 20th, 2016 | Lecture 7**

* Example [Continued from last lecture]
  + P (at least one boy)

= 1 – P (no boys)

= 1 – P (all girls)

= 1 – 1/8 = 7/8

* Definition: Two events are disjoint or mutually exclusive if they cannot both occur at the same time
  + i.e. A family cannot have all boys or all girls at the same time
* Definition: A∪B (read as “A union B”) is the event that A occurs or B occurs (or both)
  + Rule P3: If A and B are mutually exclusive then P(A∪B) = P(A) + P(B)

|  |  |  |  |
| --- | --- | --- | --- |
| Parental Handedness  (father x mother) | Handedness Of Student | | Total Number  Of Entities |
| Left  Handed | Right  Handed |
| L X L | 0 | 3 | 3 |
| L X R | 5 | 34 | 39 |
| R X L | 7 | 22 | 29 |
| R X R | 32 | 419 | 451 |
| Total | 44 | 478 | 522 |

* If a person is selected at random from the above 522 people, find the probability that their parents were RR or LR
  + P (RR∪LR) = P(RR) + P(LR)

= 451/522 + 39/522

* Definition A∩B (read as “A intersects B”) is the event that A occurs and B occurs
  + Rule P4: P (A∩B) = P(A) + P(B) – P(A∩B)
    - i.e. P (person is left handed or that their parents are RR)
      * = P(L∪RR) = P(L) + P(RR) – P(L∩RR)

= 44/522 + 451/522 – 32/522

* + - OR…
      * (5 + 7 + 32 + 419)/522
        + Just add up the total number of people that satisfy our conditions and divide it by n (total population)
* When to use P3 vs P4?
  + P3 is used for mutually exclusive events
    - Events that cannot occur together
  + P4 is used for events that are NOT mutually exclusive
    - It’s never wrong to use P4 because in the likelihood that the events are mutually exclusive, you’ll just be subtracting zero from your answer
      * P(A) + P(B) – P(A∩B) 🡨 P(A∩B) = 0
  + See figure 3 – 6 on pg. 104

Multiplication Rules & Conditional Probability [3.4 + 3.5]

* Rule P5: If A and B are independent, then:
  + P(A∩B) = P(A) x P(B)
    - i.e. Suppose that each child has probability 1/5 of inheriting a certain disease. In a family of 2 children, the probability that both children inherit the disease is P(D∩D) = P(D) x P(D)

= 1/5 x 1/5 = 1/25

* + - * There is a 1/25 chance that both children inherit the disease
    - i.e. In a family of 4 children…
      * A) P (none inherit the disease)

= P (NNNN) = P(N) x P(N) x P(N) x P(N)

= 4/5 x 4/5 x 4/5 x 4/5

= 16/20

B) P (exactly one inherits the disease)

= P (DNNN or NDNN or NNDN or NNND)

= P (DNNN) + P (NDNN) + P (NNDN) + P (NNND)

= 1/5 x 4/5 x 4/5 x 4/5 🡨 P(DNNN)

+ 4/5 x 1/5 x 4/5 x 4/5 🡨 P(NDNN)

+ 4/5 x 4/5 x 1/5 x 4/5 🡨 P(NNDN)

+ 4/5 x 4/5 x 4/5 x 1/5 🡨 P(NNND)

C) P (at least one of the children inherits the disease)

= 1 – P (none inherit the disease)

= 1 – (4/5)4

* Definition: The conditional probability of B given A
  + P(B|A) is the probability that event ‘B’ occurs given that ‘A’ has already occurred
    - Read as “B given A”
  + Rule P6: P(A∩B) = P(A) P(B|A)
    - When to use P5 vs. P6?
      * P5: If A and B are independent
        + Sometimes wrong to use this
      * P6: if A and B are NOT independent
        + Never wrong to use this

You can always use it

* + - * See figure 3 – 9 on page 112
* Key things to note:
  + OR = Add
    - When you see (∪), you have to add the probabilities
  + AND = Multiply
    - When you see (∩), you have to multiply the probabilities

**Stats 2B03 | September 22th, 2016 | Lecture 8**

* Continued From Last Lecture
  + i.e. Two people are selected at random from a group of 15 men and 30 women. Find the probability that the first is a man and the second us a woman
    - P (M∩W) = P(M) x P(W|M)

= (15/45) x (30/44)

= 90/396

* + i.e. (Continued)
    - A) If 4 people are selected, find the probability that
      * A) All 4 are men
        + P (MMMM) = P(M) x P(M) x P(M) x P(M)

= (15/45) x (14/44) x (13/43) x (12/42)

= ANS

* + - B) Exactly one of them is a man
      * P (MWWW or WMWW or WWMW or WWWM)

= (15/45) x (30/44) x (29/43) x (28/42) 🡨 MWWW

+ + +

(30/45) x (15/44) x (29/43) x (28/42) 🡨 WMWW

+ + +

(30/45) x (29/44) x (15/43) x (28/42) 🡨 WWMW

+ + +

(30/45) x (29/44) x (28/43) x (15/42) 🡨 WWWM

= 4 x (15/45) x (30/44) x (29/43) x (28/42)

= ANS

* + - C) At least three are men
      * P (MMMW or MMWM or MWMM or WMMM or MMMM)

= 4 x (15/45) x (14/44) x (13/43) x (30/42)

+ + +

(15/45) x (14/44) x (13/43) x (12/42) 🡨 MMMM

= ANS

* Rule P7: P(B|A) = [P(A∩B) / P(A)]
  + i.e. See example below

|  |  |  |  |
| --- | --- | --- | --- |
| Parental Handedness  (father x mother) | Handedness Of Student | | Total Number  Of Entities |
| Left  Handed | Right  Handed |
| L X L | 0 | 3 | 3 |
| L X R | 5 | 34 | 39 |
| R X L | 7 | 22 | 29 |
| R X R | 32 | 419 | 451 |
| Total | 44 | 478 | 522 |

* Referring to the table above, if a person is randomly selected, find the probability that
  + A) They are right handed, given that both parents are right-handed
    - P (R|RR) = [P(R∩RR) / P(RR)]

= (419/522) / (451/522)

OR

= 419/451 = 0.9290

* + B) They are right handed, given that at least one of their parents is left handed
    - P (R|LR or RL or LL)

= (3 + 22 + 34) / (3 + 29 + 39)

= 0.8310

* + C) P (RR|R)

= (419/478)

= 0.8765

* + D) P (RR|L)

= 32/44

= 0.7273

* The data suggests that you have a higher chance of being left handed if one of your parents is left handed

Random Variables [4.2]

* Chapter problem: Is this gender selection method effective?
  + Similar to tossing a coin
    - Probability is not always concrete, there is random variability
      * See pg. 157
* A random variable (r.v.) is a variable that has a single numerical value, determined by chance, for each outcome of an experiment
  + i.e. X = Number of boys in a family of three children
  + i.e. X = adult height
* A r.v. is *discrete* if we can list all of its possible values
  + i.e. X = Number of boys (0, 1, 2, 3)
  + i.e. You cannot list all the possible values for height. Even if you started at 4 feet, what’s next? Height is extremely variable
* A *probability distribution* of a discrete r.v. X is a listing of the values of X along with their probability
  + i.e. Among 14 newborn babies, let X represent the number of girls

**Stats 2B03 | September 23th, 2016 | Lecture 9**

* Example (Continued From Last Lecture)
  + i.e. Among 14 newborn babies, let x be the number of girls
    - See table 4 – 1 on pg. 159 for the probability distribution
      * Note that: ∑ P(x) and 0 ≤ P(X) ≤ 1
* Identifying unusual results
  + The rare event rule:
    - If under a given assumption, the probability of an observed event is extremely small (typically, 0.05 or less) we conclude that the assumption is probably not correct
  + Chapter problem
    - In a preliminary test of the MicroSort technique, 14 couples who wanted girls are found.
      * After using the technique, 13 of them had girls. Is the P(Girl) > ½ for this technique?
        + If we assume P(Girl) = ½

Then P(x ≥ 13)

= P (X=13) + P(X=14)

= 0.001 + 0.00

= 0.001

0.001 < 0.05

* + The chances of 14 selected couples having 13 girls is a very rare event. The chances of it occurring are 1/1000
    - So we can conclude that the technique probably works
      * However, it is possible that a really unlikely event occurred
* Mean, Variance, and Standard Deviation
  + Definition: The mean or expected value μ of a random variable is the long-term (theoretical) average of the random variable if the experiment is repeated a large number of times
    - i.e. If you have 14 kids, on average you would expect 7 girls and 7 boys
      * μ = E(x) = ∑ x **.** P(x)
        + i.e. See table 4 – 3

If 14 couples have children and x is the number of girls, then μ = E(x) = 6.993 (really 7.0)

* The variance, σ2 measures the amount of variation in the possible values of the random variable X, and is given by:
  + σ2 = ∑ (x – μ)2 P(x)

= ∑ x2 P(x) - μ2

* + - * + i.e. σ2 = 52.467 – 72 = 3.564951
* Standard deviation
  + σ = √σ2
    - i.e. σ = √3.564951
* Range rule of thumb (for identifying unusual results)
  + Maximum usual value = μ + 2σ
  + Minimum usual value = μ – 2σ
    - Most of them time, the values will be between the min/max values
      * Based off of empirical table
  + i.e. (Continuation)
    - μ – 2σ = 7 – 2(1.9) = 3.2
    - μ + 2σ = 7 + 2(1.9) = 10.8
      * Since X = 13 > 10.8
        + 13 girls is an unusual value, if the P(Girl) = ½
      * Therefore, we can conclude that P(Girl) is probably not ½

The Binomial Distribution [4.3 + 4.4]

* Definition: A Bernoulli (1654-1705) trial is a random process with only two possible outcomes of interest called “success” or “failure”
  + i.e. Gender of a newborn, flipping a coin, etc.
  + i.e. Presence or absence of a disease
* A Bernoulli process (or binomial experiment) consults of:
  + 1. ‘n’ independent Bernoulli trials
  + 2. The probability of success, P, is the same for each trial
    - i.e. The birth of 3 children
      * (n=3, p=1/2)
    - i.e. The number of children in a family of 6, who inherits a certain disease where the probability of inheriting the disease for each child is 1/5
      * (n=6, p=1/5)
    - If X is the number of successes in a Bernoulli process, then X is a binomial random variable with parameters n and p
* Counting result
  + The number of arrangements of ‘n’ letters consisting of x number of S’s and n-x F’s is:
    - How many ways can we rearrange these letters?
      * SS….S OR FF…F
      * nCX = (n\_x) = (n!) / (x! (n-x)!)
        + n choose x
      * where n! = n (n-1) x 3 x 2 x 1
        + i.e. 5! = 5 x 4 x 3 x 2 x 1
      * ! = factorial
    - Example: In how many ways can 3 S’s and 2 F’s be rearranged
      * S S S F F
        + n = S, X = 3
      * 5C3 = (5\_3) = (5!) / (3! (5-3)!)
        + = (5 x 4 x 3 x 2 x 1) / (3 x 2 x 1 x 2 x 1)
        + = 10
  + i.e.
    - A multiple choice test has 20 questions with 5 choices for each question. Suppose that a student guesses every question. Find the probability of:
      * A) Getting exactly 6 correct
        + Let x be the number of correct answers
        + Then x is binomial with n=20 and p=1/5
        + P(x=6) = P(cc…c, ww….w)

6 correct [C] and 14 wrongs [W]

In any order

* + - * + = (20\_6) (1/5)6(4/5)14

20\_6 (20 choose 6) = 38,760

* + - * + = 0.1091
      * B) Passing the test
        + P(x ≥ 10) = P(x=10) + P(x=11) +…+ P(x=20)

= (20\_10)(1/5)10(4/5)10 +

(20\_11)(1/5)11(4/5)9 +

(20\_20)(1/5)20(4/5)0

= 0.0026

Note that P(failing) = 1 – 0.0026 = 0.9974

* + - * C) How many would you expect to get right?
        + 20 x 1/5 = 4
  + Result
    - If x is binomial with parameters ‘n’ and ‘p’ then
      * P(x) = P(x successes) = (n\_x) Px (1 – P) n-x (x = 0, 1 , n)
        + μ = E(X) = np
        + σ2 = E(X) = np (1 – p)
      * Note: This formula can only be used when you have a binomial random variable

The Poisson Distribution [4.5]

* Let x be the number of occurrences of an event per unit of “time”, where there are a large number of possible occurrences of the event, but the probability of each occurrence is small
  + i.e. Number of accidents at a particular intersection
    - There is a possibility for a large number of accidents, but each drivers possibility of crashing is very small
* Let λ be the average number of occurrences of the vent per unit of “time”
* The X is called a Poisson [1781 – 1840] r.v.
  + i.e. X = number of people who acquire a certain rare disease in a given year
  + Result:
    - If X is a Poisson r.v. then:
      * P(x) = P(x occurrences) = (e- λ **.** λx) / x!
        + x = 0, 1, …
        + 0! = 1

This is a definition. Zero factorial equal to one

* + - Example:
      * A hospital emergency room admits an average of 5 new patients per hour. Find the probability that
        + A) 8 people will be admitted in the next hour

Let X be the number of people admitted in the next hour. Then x is Poisson with λ = 5

P(8) = (e- 5 **.** 58) / 8!

= 0.653

* + - * + B) At least 3 people will be admitted in the next hour

P(x ≥ 3) = 1 – P(x < 3)

= 1 – [P(0) + P(1) + P(2)]

= 1 – [(e- 5 **.** 50) / 0! +

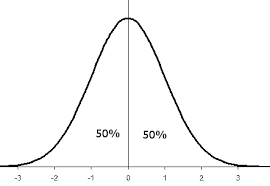
[(e- 5 **.** 51) / 1! +

[(e- λ **.** 52) / 2!]

= 0.8753

The Standard Normal Distribution [5.2]

* A random variable [r.v.] is continuous if it can assume all values in a given interval
  + i.e. Time, temperature, weight, age
    - A function f(x) is a probability density function (pdf) for a continuous random variable, X, if
      * I) f(x) ≥ 0
      * II) The total area under f(x) is 1
      * III) P(a ≤ x ≤ b) = area under f(x) between a and b (=Sf(x)dx)
* X is called a normal r.v. with mean μ and variance σ2 if its PDF is:
  + f(x) = 1 / (2(pi))0.5 **.** e –(x – μ)^2 / 2 σ σ
* If μ = 0 and σ = 1, then X is called a standard normal r.v. and is denoted by Z
  + f(Z) = 1 / (2(pi))0.5 **.** e –(Z)^2 / 2
    - Range: Negative infinity < Z < Positive infinity
* Areas under the standard normal curve
  + Table A – 2 (pages 632 – 633) gives areas under the standard normal curve to the left of Z



* + - i.e. Find the area under the standard normal curve
      * A) To the left of Z = 1.83
        + AreaLEFT = 0.9664

Value is achieved by referring to the table

* + - * B) To the right of Z = -2.17
        + AreaLEFT = 0.150
        + AreaRIGHT = 1 – 0.150 = 0.9850
      * C) Between Z = 0.87 and Z = 3.15
        + Area = 0.9992 – 0.8078 = 0.914

First, find the area for 3.15, and then subtract it with the area for 0.87

* Probabilities
  + If Z is a standard normal r.v. then P(a < Z < b) = Area under the standard normal curve between ‘a’ and ‘b’
* Note: For any continuous r.v. P(a < X < b) = P(a ≤ X ≤ b)

= P(a < X ≤ b)

= P(a ≤ X < b)

* + Examples
    - A) Find:
      * P(-1.25 < Z ≤ 2.46)

= 0.9931 – 0.1056 = 0.8876

* + - B) Z, such that:
      * P(Z1, < Z, < 1.87) = 0.8340

= Area1.87 = 0.9693

Z1 = 0.9693 – 0.834 = 0.1353 🡨 Look at the table and find the

Z1 = -1.10 closest value to this number

and that’s what Z1 is

Applications Of Normal Distributions [5.3]

* Converting to standard normal
  + Result: The area under the standard normal curve between ‘a’ and ‘b’ is the same as the area under the standard normal curve between

(a – μ) / σ and (b – μ) / σ

* + - Example: The length of human pregnancies are normally distributed with mean 268 days and standard deviation of 15 days
  + A) Find the probability that a randomly selected woman has a pregnancy between 260 and 280 days
    - Let x be the length of the pregnancy
      * P(260 ≤ X ≤ 280)
        + We can convert to a standard normal table by using the formula: (a – μ) / σ and (b – μ) / σ

(a – μ) / σ = (260 – 268) / 15 = -0.53

(b – μ) / σ = (280 – 268) / 15 = 0.80

* + - * + Find the area under the normal curve using the values calculated above

0.53 = 0.2981

0.80 = 0.7881

* + - * + 0.7881 – 0.2981 = 0.49
  + B) What percentage of pregnancies last longer than 300 days?
    - P(X > 300)
      * (300 – 268) / 15 = 2.13
        + Look on the table for the left value of 2.13

But we want the right so subtract 1 from it

* + - * + 2.13 = 0.9834 (Retrieved from table)
      * 1 – 0.9834 = 0.0166
        + 1.66% of pregnancies last longer than 300 days
  + C) Only 10% of pregnancies last less than *(how many)* days?
    - Convert to a standard normal graph
      * (a – 268) / 15 = -1.28
        + On the distribution table, find the Z value that corresponds with 0.1 (because 10% of pregnancies last how many days, is the question)
      * a = -1.28(15) + 268

= 248.8

* Example
  + Suppose that marks on a test are normally distributed with μ = 75. Find σ if 15% of the class got over 90%
    - 15% of the area is dedicated to students who got above 90%
      * (90 – 75) / σ ≈ 1.04
      * σ = 15/1.04 = 14.423

Central Limit Theorem [5.4]

* Histogram for normal population with mean μ and variance σ2
  + Normally distributed histogram. Bell shape



* Population with mean μ and variance σ2
  + Not normal
    - Left skewed histogram
      * A lot of people failed the test
  + Imagine you take a couple of samples from this population, average the samples, and then take the average of those averages
    - The average of those averages, would be very close to the average of the total population
      * A histogram of this would be normally distributed
* Histogram of will be approximately normal with mean μX-bar = μ
  + And σX-bar = σ/(sqrt.n)
    - is approximately normal with the above equations
      * The bigger your sample size is, the less variation in your sample average
  + If the shape of the original distribution is not not normal (or unknown) then the approximation should only be used if n > 30. If the original distribution is normal, then it can be used for any sample size (and is exact)
    - See tables 5 – 6 and Figures 5 – 19 and 5 – 20 ON pages 223, 224
* Example:
  + The average cholesterol level for women aged 20 – 29 is 183 with standard deviation 37. In a sample of 40 women aged 20 – 29, find the probability that the average cholesterol level is greater than 195
    - Can use normal distribution because the population is greater than 30
      * P( > 195)
      * is approximately normal with
        + μX-bar = 183
        + σX-bar = 37/sqrt.40
      * Calculate the area by converting to standard normal
        + (195 – 183) /37/sqrt.40 = 2.05

Look on the distribution table for the LEFT

2.05 = 0.9798

1 – 0.9798 = 0.0202

* + If you have sample, and you want the probability for the average of that sample, you need to use ∑x-bar
    - When to divide by σ/sqrt.n???
      * When calculating the probability for the average of a sample of size n

Assessing Normality [5.7]

* If you have the complete population data set, you draw a histogram, and if it is bell-shaped, then you know it is normal
  + Given a population of values, we can draw a histogram to see if the population follows a normal distribution
  + Given a sample of values, how do we determine if the population follows a normal distribution???
    - Procedure
      * 1. Put the data values in increasing order
      * 2. Identify Z-values corresponding to areas to the left of 1/2n, 3/2n,…, (2n-1) / 2n
      * 3. Plot the (x,y) pairs where X is an ordered data value, and y is the corresponding Z-value (AKA: Z-scores)
    - Example:
      * 55, 63, 72, 41, 87, 75, 64, 60
        + 1. Order the data first

41, 55, 60, 63, 64, 72, 75, 87

* + - * + 2. (1/2n) = 1/2x8 = 1/16 = 0.625

We need to find the area to the left for the Z value of 0.625

Z1 = -1.53

* + - * + 2. (3/2x8) = 0.1875

The corresponding Z value is: Z2 = -0.89

* + - You need as many Z values as the number of observations
      * -1.53, -0.89, -0.49, -0.16, 0.16, 0.49, 0.89, 1.53
        + 3. Plot (41, -1.53), (55, -0.89), etc.

PLOT ON A GRAPH

* + The idea is that:
    - Z scores are spread out in the same way that a normal data set should be
      * Most of the data is in the middle and less of it at the ends
        + Therefore, if the data set is spread out in the same way as the Z-scores, then it is reasonable to assume that the population is normal
      * If the original data set is a perfect data set, then the Z scores data set produces a straight line graph
* Interpreting the normal probability plot
  + If the data set is a random sample from a population that is normally distributed, then the points on the normal probability plot will roughly form a straight line
    - The points should be randomly scattered about the straight line with no discernible no-linear pattern among the points

Estimating A Population Proportion [6.2]

* Chapter Problem [see p.255]
  + Mendel (1822 – 1884)
    - Cross fertilized pure green peas and pure yellow peas
      * According to his theory: the new pea would have one gene from the green pea and one from the yellow pea
    - Then he cross-fertilized these G.Y. peas
      * Breeding G.Y. peas together holds 4 possible outcomes
        + G.G., G.Y., Y.G., Y.Y.,
      * If the green is dominant, then the proportion of yellow peas should 0.25 or 25%
        + He observed that the offspring consisted of 580 peas, 428 of which were green and 152 were yellow

152/580 = 26.2%

Does this mean his theory was wrong???

Where do you draw the line???

* Confidence interval
  + Definition: A confidence interval (C.I.) for a population parameter is a range of values used to estimate the parameter based on a sample.
  + The confidence level is the proportionate of time that the C.I. will actually contain the parameter, if the process is repeated a large number of times
* Critical value
  + Definition: A critical value is a number of the borderline of separating sample statistics that are likely to occur from those that are unlikely to occur
* Definition: Zα/2 is a critical value that is a Z-score with the property that the area under the standard normal curve to the right of Zα/2 is equal to Z α/2
  + Example: Find Zα/2 
    - A) If α = 0.1
      * Zα/2 = Z0.1/2 = Z0.05
      * Z0.05 = (1.64 + 1.65)/2 = 1.645
    - B) If α = 0.25
      * Zα/2 = Z0.25/2 = Z0.025
      * Z0.025 = 1.96
    - C) If α = 0.01
      * Zα/2 = Z0.01/2 = Z0.005
      * Z0.005 = 2.575
  + Let P denote the population proportion
    - i.e. P could be…
      * P = The proportion of peas that should be yellow
      * P = The proportion of people who have a certain disease
    - Let denote the sample proportion, and let = 1 -
      * Result:
        + A (100(1 – α)% confidence interval for P is

+ E

Where E = (Zα/2)(sqrt.(/n))

Is called the margin or error

* + - * + This should only be used if n > 5 and n > 5
  + Chapter problem
    - Find α 95%
      * C.I. interval for the true proportion
    - α = 0.05 [So that 100(1 – α) = 100 x 0.95 = 95]
      * (152/580) +/- Z0.025 (sqrt.((152/580) (1-152/580)/580)))
      * 0.262 +/- 0.358
      * 0.226 +/-0.298
        + i.e. We estimate 22.6% < P < 29.8%

This is consistent with Mendel’s theory

* + Interpretation
    - If the experiment is repeated a large number of times, then the confidence interval [C.I.] will contain the true proportion, P, 100 (1 – α)% of those times
      * So, for example, if a 100 different people repeated the above, Mendel’s experiment, and if Mendel’s theory were true, only 95 of those experiments would produce a confidence interval that contained the true value of P, P = 0.25
        + So 5 of those people would falsely conclude that Mendel’s theory is wrong

See figure 6.3 on pg.265

* + - * + Historically, Mendel’s data was lined up too nicely. Statisticians believe that Mendel altered the experiment to have better, cleaner results
      * i.e. If you flipped a coin ten times, in 100 different trials, there is a possibility that you would get all ten tails in a trial. That one trial would make you falsely conclude that the coin is unfair and weighted. This is just due to random variation [r.v.]
  + The bigger the sample size, the smaller the margin of error
* Determining the sample size
  + Suppose that we want to estimate the proportion of Canadians who are left handed, accurate to within 3% with a 90% confidence interval
    - How large a sample should be used?
      * (Zα/2)(sqrt.(/n)) = 0.3 WITH α = 0.1
      * (Z0.05)2(/n) = (0.3)2
      * ((1.645)2 / (0.3)2)(/n) = n
    - In general, n = ((Zα/2)2())/ E2
      * Case 1: can be roughly estimated from previous data. For example from Table 1, = 44/522
        + Then n = (1.645)2 / (0.3)2 x (44/522) (1 – 44/522)

n = 232.07

Always round UP

n = 233

* + - * Case 2: No prior information about is known
        + Here we can use the fact that (1 - ) cannot be bigger than ¼. So replace by 1/4

n = (1.645/0.3)2 x ¼ = 751.67

n = 752 [ROUND UP]

Estimating A Population Mean μ With σ Known [6.3]

* Recall that μ and σ are the average (mean) and standard deviation, respectively, of the population
  + i.e. μ = Average body temperature of a healthy adult
  + i.e. μ = Average duration of relief of a pain reliever
    - Result: If σis known, and either the population is normal or n > 30 then a 100 (1 – α) %
      * Confidence interval for μ is +/- E
        + Where E = Zα/2 σ/sqrt.n is the margin of error
    - Example:
      * A sample of 106 healthy adults gave a mean body temperature of 98.2 (See Data Set 2 in Appendix B)
    - Suppose that σ is known to be 0.62
      * Estimate the mean body temperature of a healthy adult using a 99% confidence interval
        + α = 0.01

98.2 +/- Z0.005 (-62/sqrt.106)

* 1. +/- 0.155
  2. o 98.36
     + We are 99% confident that this interval contains the true mean body temperature of healthy adults
       - i.e. It is very unlikely that μ = 98.6
* Determining the sample size
  + Example: How many students should be sampled if we want to estimate the class average, μ, on Test #1 accurate to within 3 marks with a 98% confidence interval. Assume that σ = 16.17
    - Zα/2 σ/sqrt.n = 3
      * Where α = 0.02
    - Sqrt.n = (Z0.01 (16.17)) / 3
      * Z0.01 = 2.33
        + You have the area (0.01) find the corresponding value
    - n = ((2.33)2 (16.17)2) / 32

= 157.72 🡪 158 (Always round UP)

* + - So, we need 158 people to write the test early to estimate the total class average
* What if σ is unknown?
  + 1. Estimate σ based on previous data
  + 2. Do a pilot study, and use S, from that that study to estimate σ
    - A pilot study is a small scale preliminary study conducted in order to evaluate feasibility, time, cost, adverse events, and effect in an attempt to predict an appropriate sample size
  + 3. Use the range rule of thumb, i.e. σ ≈ range/4
    - i.e. σ ≈ 80/4 = 20
      * The highest value is 100 and the lowest is 20
        + These are hypothetical test marks
      * Almost always gives a value of σ that is too big, which is okay because it’ll give you a bigger sample size
  + Note: The methods of chapter 6 on should only be used (in practice) if the population size is at least 20 times the sample size

Estimating μ With σ Unknown [6.4]

* Result: If σ is unknown and either the population is normal or n > 30 then a 100(1 – α) % C.I for μ is +/- E
  + Where E = tn – 1, α/2 (S/sqrt.n)
    - i.e. If n = 106 and α = 0.01, find tn – 1, α/2
    - t106 – 1, 0.01/2
    - t105, 0.01/2
      * t105 is called degrees of freedom
      * t105 = 2.626
        + Look up this number on the T-table
        + There are two numbers, depending if you divide α by 2 or not. Always divide α by 2 and use the first column
* Example: A sample of 106 healthy adults gave a mean body temperature of = 98.2 with S = 0.62. Estimate the mean body temperature, μ, for a healthy adult, using a 99% confidence interval
  + C.I. = x-bar +/- (p.hat)(s/sqrt.n)

= 98.2 +/- Z(0.005(0.62/sqrt.106)

α

μ

σ

* Replace “equations” with fonts
* Organize notes (grammar, etc.)
* Format titles (& or +)
* Make a legend for equations/alpha numeric symbols
* Fix the dates and lecture number!!!
  + Remove dates?