AP Statistics Part 1 Practice

Rishi Salwi and Eric Zheng

September 26, 2018

1. The following plots were created:

   (a) Stem and leaf plot:
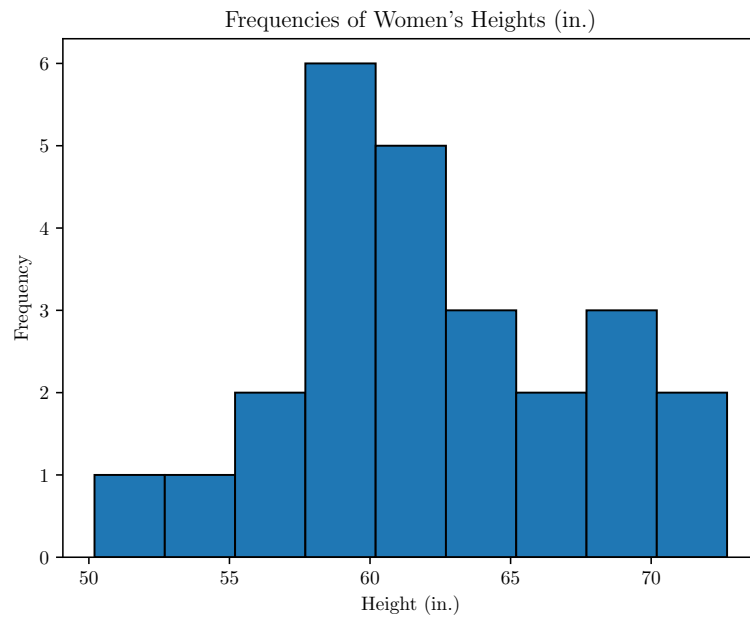
   Figure 1: Stem and Leaf Plot of Women's Heights (in.)

Stem	Leaf			
50	2			
51				
52				
53	2			
54				
55				
56	7			
57	4	9		
58	6	9		
59	0	3	3	
60				
61	4			
62	1	1	4	4
63	1	1	3	
64				
65	4	6		
66				
67	7			
68	5	5		
69				
70	4	5		

   (b) Frequency table:

   Table 1: Frequency Table of Women's Heights (in.)

Interval	Frequency
$[50.2, 52.7)$	1
$[52.7, 55.2)$	1
$[55.2, 57.7)$	2
$[57.7, 60.2)$	6
$[60.2, 62.7)$	5
$[62.7, 65.2)$	3
$[65.2, 67.7)$	2
$[67.7, 70.2)$	3
$[70.2, 72.7)$	2

(c) Histogram:

Frequencies of Women's Heights (in.)

(d) Pie chart:

Pie Chart of Women's Heights (in.)

(e) Box plot:

Box Plot of Women's Heights (in.)

Min = 50.2
Q1 = 58.9
Med = 62.1
Q3 = 65.4
Max = 70.5

Min   Q1   Med   Q3   Max

50.0   52.5   55.0   57.5   60.0   62.5   65.0   67.5   70.0
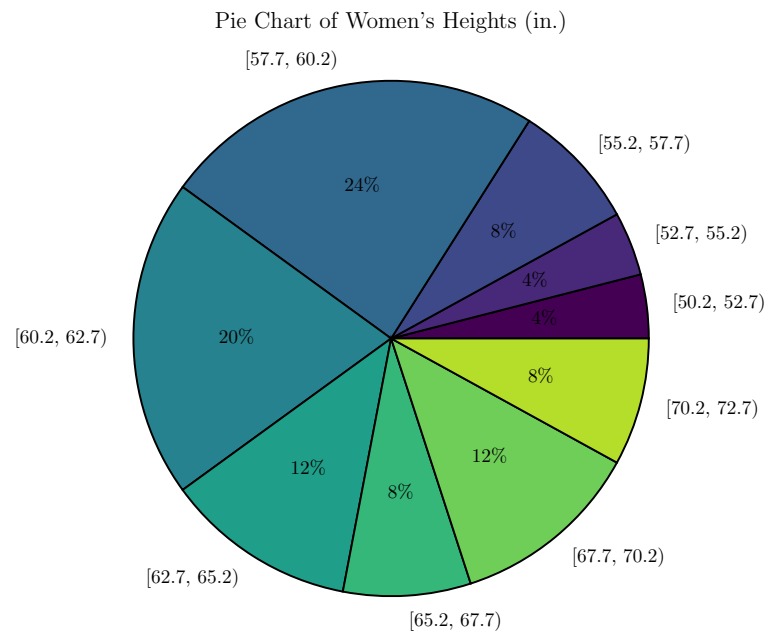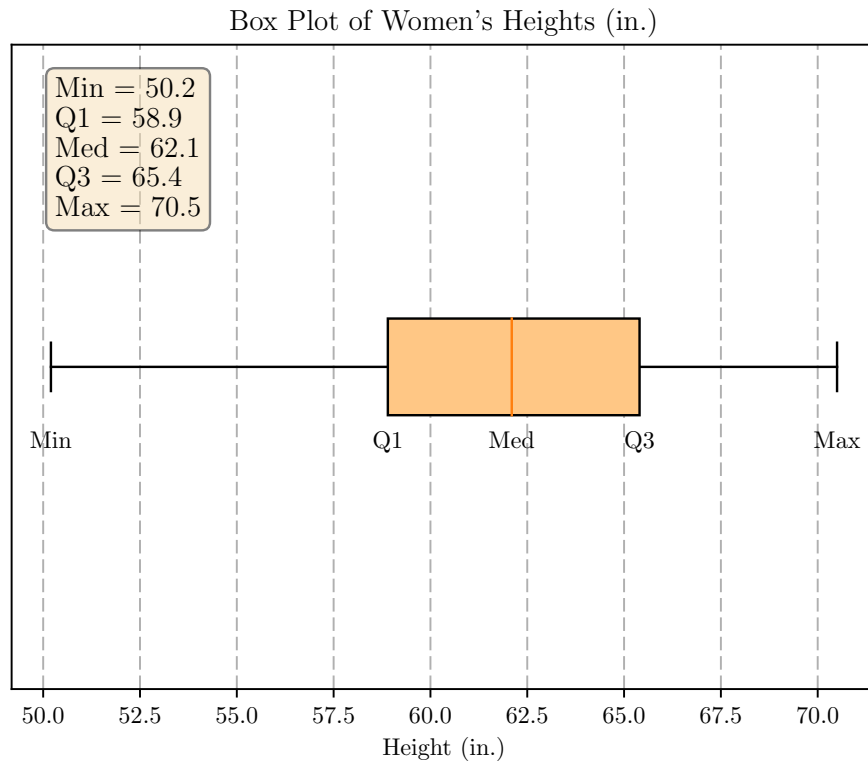
Height (in.)

2. These data were used to compute the following sample summative statistics:

Table 2: Summative statistics about women's heights (in inches)

Statistic	Value
Mean	61.88
Median	62.1
Midrange	60.4
Mode	68.5, 63.1, 62.4, 62.1, 59.3
Range	20.3
Standard Deviation	5.07
Variance	25.68

3. The data set possesses a left tail. This implies that is is not normally distributed; rather than being symmetric, it has a left skew.

4. In this scenario:

(a) The mean increased. Let $\bar{x} = (\Sigma x)/n$ represent the original mean, where x are the original data and n is the original class size. Then $\bar{x}' = (100 + \Sigma x)/(n + 1)$ is the new mean. Then we have:

$$\bar{x}' - \bar{x} = \frac{100n + \Sigma x}{n(n+1)} \tag{1}$$

This difference is positive (indicating increase in mean) if $100 > (\Sigma x)/n = \bar{x}$; moreover, the mean of a data set cannot exceed its maximum value, so $\bar{x} \leq 92$. Since $100 > 92$, the mean must have increased.

(b) For a sufficiently large class size, the median stays roughly the same. For very small classes (consider, say, a class of simply $\{71, 92\}$), the addition of a high score will appreciably change the median. However, for a class of large enough size for a "normal distribution" to be meaningful, this change is negligible. Loosely speaking, the new score, being greater than the original maximum, will shift the median up by half a value. Since the data are normally distributed, it is expected that the data are dense around the median, so this shift has very little impact. (It is even possible for this shift to have no impact if the median value is repeated.)

(c) The range increased. The original range was $92 - 71 = 21$. The new range is $100 - 71 = 29$, an obvious increase.

(d) The interquartile should stay roughly the same. Q1 and Q3 are both shifted slightly to the right by the new score, so their difference does not change by much, if at all. (A similar argument can be made that the interquartile range represents the "middle 50%" of the data. Adding a single extreme value is unlikely to change this.)

(e) The standard deviation increases. Intuitively, the standard deviation expresses how "spread out" the data are, so adding an extreme value will increase it.

5. We examine the hypothesis:

Null: Age and music preference are independent.
Alternate: Music preference is dependent on age.

With the data:

Table 3: Music Preference by Age Group

	Age Group			
	15-24	25-34	35-44	45-54
Classical	12	37	72	85
	10.7%	18.9%	40.4%	38.1%
Rock	89	106	64	100
	79.5%	54.1%	36.0%	44.8%
Country	11	53	42	38
	9.8%	27.0%	23.6%	17.0%
Total	112	196	178	223
	100%	100%	100%	100%

There are a few noticeable trends in the data; for instance, older participants were significantly more likely to prefer classical music then their younger peers, while younger participants were much more likely to enjoy rock music. As such, we conclude:

Reject the null hypothesis. Support the alternate hypothesis. The data do support that music preference is dependent on age.

6. For extra credit, the "68-96-99.7 rule" applies to normally distributed curves. 68% of the data fit in between one standard deviation of the mean. 95% of the data fit in between two standard deviations of the mean. 99.7% of the data fit in between three standard deviations of the mean. This is called the empirical rule.