**1.42. Sleeping in college.**

A recent article in a college newspaper stated that college students get an average of 5.5 hour of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night. Identify which value represents the sample mean and which value represents the claimed population mean.

* Sample mean 🡺 6.25 hours
* Claimed Population Mean 🡺 5.5 hours

**1.46. Medians and IQRs**

For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

1. (1) 3, 5, 6, 7, 9

(2) 3, 5, 6, 7, 20

1. 🡺 Mean : 6, Standard Deviation : 2.23
2. 🡺 Mean : 8.2, Standard Deviation: 6.76

* #2 has higher mean since 20 is greater than 9,
* #2 has higher deviation since 20 is farther away from the rest of the data compared to 9.

1. (1) 3, 5, 6, 7, 9
2. 3, 5, 7, 8, 9

(1) 🡺 Mean: 6 Standard Deviation: 2.23

(2) 🡺 Mean: 6.4 Standard Deviation: 2.40

* #2 has higher mean since it contains elements 7 and 8 which are greater than 6 and 7.
* #2 has higher standard deviation since it contains elements 7 and 8 which are farther than 6 and 7.

1. (1) 1, 2, 3, 4, 5
2. 6, 7, 8, 9, 10

(1) 🡺 Mean: 3 Standard Deviation: 1.58

(2) 🡺 Mean: 8 Standard Deviation: 1.58

* #2 mean is higher than that of #1 as each of the values in distribution #2 is greater than #1
* Both distribution have the same standard deviation since they are equally distanced around their respective means.

1. (1) 0, 10, 50, 60, 100

(2) 0, 100, 500, 600, 1000

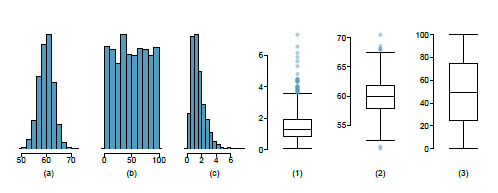
(1) 🡺 Mean: 44 Standard Deviation: 40.37

(2) 🡺 Mean: 440 Standard Deviation: 403.73

* #2 Mean and standard deviation are 10 times that of #1, since #2 contains elements which are 10 times that of #1.

**1.50 Mix-and-match.**

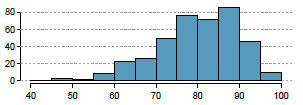
Describe the distribution in the histograms below and match them to the box plots.



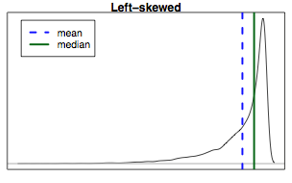
1. 🡺 (2). Histogram is Unimodal, with single prominent peak.
2. 🡺 (3). Histogram is uniform and no single peak which stands out.
3. 🡺 (1). Histogram is Right skewed, and is Unimodal with 2 prominent peaks. One peak is slightly higher than the other.

**1.52 Median vs. mean.**

Estimate the median for the 400 observations shown in the histogram, and note whether you expect the mean to be higher or lower than the median.

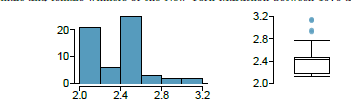


* Median is probably around 80
* Mean would be lower than median because
  + Histogram is left skewed.
  + Typically in left skewed histogram mean is less than median **(mean < median)**



**1.54. Marathon winners.**

The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1970 and 1999.



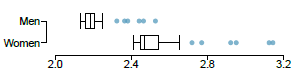
1. What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

* The histogram is bimodal with 2 peaks not exactly same one been higher than the other.
* Histogram is right skewed
* Median and outliers are pretty much evident from the box plot.

1. What may be the reason for the bimodal distribution? Explain.

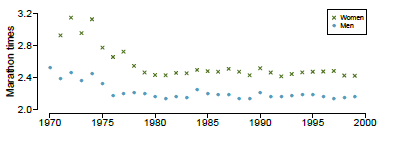
* 2 Peaks are finishing times of men and women.
* 2 Peaks represent majority of men and women finishing time and not all.

1. Compare the distribution of marathon times for men and women based on the box plot shown below.



* Median for men for timing of completing the marathon is less than that of women.
* Overall all men are completing much lesser time than that of women.
* There is considerable variable number in the women completing the race compared to men. This is evident from the difference between the lower and upper whisker in both box plots.
* Outlier’s distribution for women is more compared to men.

1. The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.



* One prominent pattern is the marathon time completion for men and women over the years starting from year 1970 to 2000.
* For both men and women marathon completion time has reduced from 1970 to bit greater than 1975 and has remained close to same range till year 2000.
* Women completion time is higher than men.

**1.56. Distributions and appropriate statistics, Part II.**

For each of the following, state whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

1. Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

* **Right Skewed**. There are considerable number of houses in first 2 quartiles than that of 3rd quartile. Density of number of houses in first 2 quartiles is higher than third. If we have to make the chart the x-axis would be increase in house price and y-axis would be number of increase in number of houses. Tail would have less number of houses due higher price, so number of families buying those houses.
* **Median** would be best representation of the value since the mean would factor in extreme values of $6,000,000. Value of $6,000,000 is significantly higher which would impact the mean.
* **IQR** would be best to represent the variability of the data since standard deviation considers extreme value of $6,000,000 which has significant impact.

1. Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

* **Symmetric**. There are few houses costing more than $1,200,000, also the higher value is not exorbitantly high. If we have to make the chart the x-axis would be increase in house price and y-axis would be number of increase in number of houses. It would have peak and center and the tails on both sides.
* **Mean** can be used to represent the data, as there is uniform distribution. Extreme values are not outliers and can be considered. **Median** would produce similar results as well.
* Since data is uniform **standard deviation** can used to represent variability. **IQR** can will also produce similar outcome.

1. Number of alcoholic drinks consumed by college students in a given week. Assume that most of these students don't drink since they are under 21 years old, and only a few drink excessively.

* **Left Skewed.** There are very few students who are drinking. If we have to plot the chart x-axis would be increase in age and y-axis would be increase in the number of alcoholic drinks consumed by students. Below 21 most of the students do not drink so the value of the bars would be very minimal or none. Also majority of the students fall under the category of age < 21. Peaks would be at age and 21 and above.
* **Mean** would be best representation of the data.
* **Standard deviation** would be good to represent the variability.
* **Reason for choosing median and standard deviation is,** it factors in extreme values, where most of the analysis data lies. If we take IQR, we will be ignoring the 4 quartile data as outliers as most of the data related to number of alcoholic drinks consume by student lies in that range.

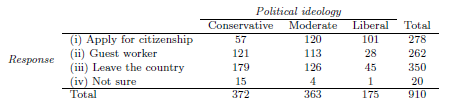
1. Annual salaries of the employees at a Fortune 500 company where only a few high level executives earn much higher salaries than all the other employees.

* **Right skewed.** If we have to plot the chart that x-axis would in increase in salary and y-axis would be increase in number of employees. As high level executives are less in number, peak height would be less.
* **Median** would be best representation of the data. Executive salaries are outliers so they need to be discarded as they do not represent the salary distribution and if we factor in those they will skew up the numbers.
* **IQR** would represent the variability of the data. Executive data would be in the 4th quartile and need to be discarded in this case.

**1.66. Views on immigration.**

910 randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be

1. allowed to keep their jobs and apply for US citizenship,
2. allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or
3. Lose their jobs and have to leave the country. The results of the survey by political ideology are shown below



Percentage wise table

1. What percent of these Tampa, FL voters identify themselves as conservatives?

**40.88%**

1. What percent of these Tampa, FL voters are in favor of the citizenship option?

**30.55%**

1. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

**6.26%**

1. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

**Conservatives: 15.32%.**

57/372 = 0.1532

**Moderates: 33.06%**

120/363= 0.3306

**Liberals: 57.71%**

101/175=0.5771

1. Do political ideology and views on immigration appear to be independent? Explain your reasoning.

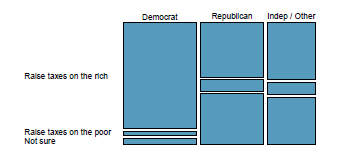
* Below table indicates percentages of among conservatives, moderates and liberals opinion on 4 options Apply for citizenship, Guest worker, Leave the country, not sure

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Conservative** | **Moderate** | **Liberal** |
| (i) | **Apply for citizenship** | 15.32% | 33.06% | 57.71% |
| (ii) | **Guest worker** | 32.53% | 31.13% | 16.00% |
| (iii) | **Leave the country** | 48.12% | 34.71% | 25.71% |
| (iv) | **Not sure** | 4.03% | 1.10% | 0.57% |

* It is pretty much evident form the table that political ideology and views on immigration are linked to each other.
* Conservatives are harsh on immigration, as 15.32% only agree on giving possible path to citizenship. While 48.12% want the illegal immigrants to leave the country and 35% agree on keeping as guest worker. Leave the country is extreme view against immigration and guest worker as well is moderate view against immigration.
* On the contrary to conservatives, % of liberals on same views are at opposite spectrum. 57.71% of liberals are in favor of allowing illegal workers to path to citizenship. While 25.71% of liberals are of opinion that they should leave the country and less percentage of 16% of liberals want illegal immigrants as guest workers.

**1.68. Raise taxes**

A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. The survey also collected information on the political party affiliation of the respondents. Based on the mosaic plot shown below, do views on raising taxes and political affiliation appear to be independent? Explain your reasoning?



1. Views on raising taxes on different economy classes based political party ideology **definitely differs** and is very much evident from the mosaic plot. Raising taxes and political affiliation are definitely dependent on each other and are not independent.
2. Very large proportion of the Democrats are of the opinion on increasing the taxes on the Rich people compared to Republican and Independent/others.
3. Republican on raising taxes with rich and not sure are almost in close proportions with very few inclining to raise taxes on poor.
4. As far as raising taxes on poor are concerned, more republicans are in favor compared to Democrats. But in both cases number is not that significant.
5. Independent/other have similar opinion as that of republicans, with slight variation in each of the raising taxes on rich or raise taxes on poor.