Murray Lynch   February 16, 2015

WWS 200
Precept: P03

Homework 1

<u>1) BASIC CONCEPTS</u>

(a)

Standard deviation is 0, because standard deviation measures the typical distance of a
data point from the mean (all data points are 0 which is the mean)

(b)

mean(y)   = sum(y)/n
                 = (14 + 11 + 10 + 16 + 14 + 13) / 6
                 = 13

median(y)   = mid-point of data (for even numbers of data, the median is the mid-point
                 between the two middle observations)
                 = (13 + 14) / 2
                 = 13.5

for standard deviation:

deviation(yx) = yx - mean(y) (where x demarks a specific data point)

sq(dev.(yx)) = (yx - mean(y))^2

deviation 1   = (10 - 13)
               = -3
sq(dev. 1)   = (-3)^2
               = 9

deviation 2   = (11 - 13)
               = -2
sq(dev. 2)   = (-2)^2
               = 4

deviation 3   = (13 - 13)
               = 0
sq(dev. 3)   = 0^2
               = 0

```
deviation 4      = (14 - 13)
                 = 1
sq(dev. 4)       = 1^2
                 = 1

deviation 5      = (14 - 13)
                 = 1
sq(dev. 5)       = 1^2
                 = 1

deviation 6      = (16 - 13)
                 = 3
sq(dev. 6)       = 3^2
                 = 9


sum of deviations     = (-3) + (-2) + 0 + 1 + 1 + 3


variance              = sum(sq. devaitions) / (n-1)
                      = (9 + 4 + 0 + 1 + 1 + 9 ) / (6 – 1)
                      = 4.8

standard deviation (s) = ( variance )^0.5
                       = ( 4.8 )^0.5
                       = 2.19089
```

(c)

This would work if the maximum value for team 1 is less than both the median and mean value for team 2.

For example:

team 1 members record the following times: 11.0, 11.5, 12.0, 12.5, 13.0
mean is 12.0
median is 12.0

team 2 members record the following times: 13.0, 13.5, 14.0, 14.5, 15.0
mean is 14.0
median is 14.0

Joe recorded a time of 13.0s for team 1

After Joe switches to team 2, the data looks like:

team 1 members record the following times: 11.0, 11.5, 12.0, 12.5
mean is 11.75
median is 11.75

team 2 members record the following times: 13.0, 13.0, 13.5, 14.0, 14.5, 15.0
mean is 13.83
median is 13.75

Both mean and median for both teams drops.

(d)

A SRS (simple random sample) of the US population means that each person in the US has an equal chance of being selected for the sample.

When comparing two groups of people, in this case Native Americans versus other Americans, a disproportional stratified random sample would be better. This is because the Native American population is too small relative to the population of other Americans to make a comparison with a proportional stratified random sample.

(e)

new sample mean $= (\text{(old sample mean} * (n-2)) + 1.1 + 1.01) / n$
$= ((1 * (29 - 2) + 1.1 + 1.01) / 29$
$= (29.11) / 29$
$= 1.0038$

Since both new values deviate from the mean by less than 0.2, and the new mean does not significantly change the deviations of the other observations, the standard deviation would decrease.

(f)

It would be multimodal because the first peak would represent the mode of women's heights while the second would represent the mode of men's heights.

3.2) R EXERCISE

(a)

Commands (with results) to load dataset and calculate mean:

> setwd("~/wws200/Homework 1")
set working directory to the homework folder

> load("injury.Rdata")
load dataset from working directory

> mean(injury$durat)
[1] 9.922203
the average time away from work for workers is this sample is 9.9 weeks

(b)

See R Script for full set of commands and explanatory comments.

> dim(injury)
[1] 7150   25
provides the dimensions, i.e. the number of observations and variables
of the object "injury"

Results for the dimension function show that there are 7150 observations and 25
variables for the object "injury".

Using the max and min values provided by the summary function, we find that the oldest
person is 98 y.o. and the youngest is 12 y.o.

> summary (injury$age)
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.   NA's
  12.00   24.00   32.00   34.71   43.00   98.00   4
provides basic descriptive measurements of the "age" variable

Using the mean value provided by the summary function, we find that 26.18% of the
observed sample had a lower back injury.

> summary (injury$lowback)
Min. 1st Qu. Median   Mean 3rd Qu.   Max.
 0.0000 0.0000 0.0000 0.2618 1.0000 1.0000
provides basic descriptive measurements of the "age" variable

Using the mean value of the summary function, we find that 78.69% of the observed sample is in KY, while 21.31% are in MI.

```
> summary (injury$mi)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2131  0.0000  1.0000
> summary (injury$ky)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0      1.0000  1.0000  0.7869  1.0000  1.0000
```

(c)

Male is categorical/ nominal.

Durat is quantative/ordinal and continuous.
Potential range is the length of the period in which observations were made.
Observed range is max – min, which is $182 – 0.25 = 181.75$
Mean is 9.922203
Standard deviation is 24.49754

Highearm is categorical/nominal.
Potential range is 1 $(1 – 0)$
Observed range is max – min, which is $1 – 0 = 1$
Mean is 0.3988811
Standard deviation is 0.4897025

Afchnge is categorical/nominal.
Potential range is 1 $(1 – 0)$
Observed range is max – min, which is $1 – 0 = 1$
Mean is 0.4732867
Standard deviation is 0.4320899

Age is ordinal and discrete.
Potential range is the age of the youngest working adult minus the age of the oldest working adult
Observed range is max – min, which is $98 - 12 = 86$
Mean is 34.71

Injtype is categorical and discrete.
Potential range is how many type categories the researcher decides on.

Observed range is max – min, which is 8 – 1 = 7
Mean is 4.450951
Standard Devation is 1.516924

(d)

IQR for age is 43 – 24 = 19
Standard deviation is

IQR for durat is 8 -2 = 6
Standard deviation is 24.49754
Therefore, IQR is less than sd