

2 Basic Computations and Summary Statistics

2.1 Measures of the Center of the Data

- Outlier – a data value that is not consistent with the bulk of the data
 - Outliers can sometimes lead to complications in data analysis and erroneous conclusions if not treated appropriately
 - Possible reasons and recommended actions
 - * The outlier is a legitimate value that represents natural variability for the group and should remain in the data set
 - * A mistake is made, and the value should be corrected or removed from the data set
 - * The observation in question belongs to a different group than those measured, and can be removed if desired
 - Only remove outliers with proper justification
- A statistic is said to be robust/resistant if the presence of outliers does not cause it to change very much
- Mean
 - Sample mean – the mathematical average of a data set, defined by

$$\bar{x} = \frac{\sum x_i}{n}$$

where n represents the sample size

- Population mean – the mathematical average of a population, defined by

$$\mu = \frac{\sum x_i}{N}$$

where N represents the population size

- Properties
 - * Likely not a value in the data set
 - * Not robust/resistant
 - * Most commonly used metric for quantitative data
 - * Cannot be used with qualitative data
- Median
 - Found by arranging the original data set in increasing (or decreasing) order and taking the middle number (averaging if needed)
 - Properties
 - * Possibly a value in the data set
 - * Very robust/resistant
 - * Cannot be used with qualitative data

- Mode
 - The data value that occurs with the greatest frequency
 - Properties
 - * Data sets can be unimodal, bimodal, multimodal or have no mode
 - * If it exists, must be in the data set
 - * Highly variable from data set to data set
 - * Used mostly with qualitative data

2.2 Skewness and the Mean, Median, and Mode

- Distribution shapes
 - Right skewed – a data set that has a longer right tail, where usually the mode is less than the median which is less than the mean
 - Left skewed – a data set that has a longer left tail, where usually the mean is less than the median which is less than the mode
 - Symmetric – a data set that has the same value for the mean and median, and, if unimodal, the mode

2.3 Measures of the Spread of the Data

- Range – the difference between the maximum data value and the minimum data value
- Standard deviation – a measure of spread that can be thought of as the average distance that a “typical” data point differs from the mean
- Variance – a measure of spread that is the square of standard deviation
- Sample standard deviation – denoted as s
- Population standard deviation – denoted as σ
- Sample variance – denoted as s^2
- Population variance – denoted as σ^2

2.4 Measures of the Location of the Data

- Percentile – the value such that a certain percent of the data values are less than it
- Quartile – values that divide the data set into quarters such that $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$
- Standardized values (z scores) – a unitless measure of location that explains how extreme a data point is relative to the mean (or rather, how many standard deviations above or below the mean a data point is), defined by

$$z = \frac{x - \bar{x}}{s} \text{ or } z = \frac{x - \mu}{\sigma}$$

Examples

- Example 1

- I created the variable “SLG” (slugging percentage) for the baseball data. The formula for this is $\frac{1B + 2 * 2B + 3 * 3B + 4 * HR}{AB}$. The summary statistics of mean, median, and standard deviation are shown in the table. Also shown are the same summary stats broken down by those making the major league minimum and those making more than the minimum. Note that this presentation is more appealing than the software output, which I’ve included with the code for completeness.

Slugging Percentage (SLG)			
	Median	Mean	Standard Deviation
Overall	0.424	0.426	0.573
Min	0.423	0.426	0.514
Not Min	0.425	0.426	0.573

- I then created a new variable called “Speed Rating”. This assigns slow, moderate, fast, or very fast to players based on the number of stolen bases (SB) they had. There 57 slow players, 18 moderate players, 17 fast players, and 15 very fast players.

- Code

- In SAS the format statement tells us how to display a created variable. It’s not needed, but something I encourage you to play with so that you can get the output displayed as you’d like. SAS requires you to wrap variables that start with a number with the single quotes, and I follow that up with an “n” to tell it that it’s a numeric variable. Notice I don’t have to do that with “hr” and “ab”. Note that I’m sorting by the variable “min” – that’s because any time you run something like a proc means and want output by group, it has to be sorted by that group (which is the step I do right before).

```
proc import out=baseball datafile="C:\Users\rsides\Dropbox\TWU\Stat Programming + SAS\Data Sets\Baseball.xlsx" dbms=xlsx replace;
  getnames=yes;
run;

*create new variable;
data baseball;
  set baseball;
  format slg 4.3;
  slg = ('1b'n + 2*'2b'n + 3*'3b'n + 4*hr) / ab;
run;

*summary statistics;
proc means mean median std;
  var slg;
run;

proc sort;
  by min;
run;

proc means mean median std;
  by min;
  var slg;
run;

*new variable;
data baseball;
  set baseball;
  length speedrating $10;
  if sb < 5 then speedrating = 'slow';
  else if sb < 10 then speedrating = 'moderate';
  else if sb < 20 then speedrating = 'fast';
  else speedrating = 'very fast';
run;

proc tabulate;
  class speedrating;
  table speedrating;
run;
```

The MEANS Procedure		
Analysis Variable : slg		
Mean	Median	Std Dev
0.4259767	0.4243323	0.0721917

..... Page Break

The MEANS Procedure		
MIN=NO		
Analysis Variable : slg		
Mean	Median	Std Dev
0.4259342	0.4253769	0.0757488

..... MIN=YES

Analysis Variable : slg		
Mean	Median	Std Dev
0.4261505	0.4233577	0.0568826

..... Page Break

speedrating			
fast	moderate	slow	very fast
N	N	N	N
17	18	57	15

..... Page Break

- Note that in R, the code shows up in the output. Thus, it's really only necessary to include the output.

```

> #Creating a new variable
> SLG <- round((Data$`1B` + 2*Data$`2B` + 3*Data$`3B` + 4*Data$HR) / Data$AB, 3)
> Data <- cbind(Data, SLG) #append the new variable to the data set
>
> #Break into groups by binary variable (mine is "MIN" which indicates if a player made the major league minimum or more)
> Min <- Data[which(Data$MIN == "YES"), ]
> More <- Data[which(Data$MIN == "NO"), ]
>
> #Summary statistics of home runs (count) and slugging percentage (measurement) overall and by salary split
> #summary() gives the five number summary (includes median) and mean
> summary(Data$SLG)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2490 0.3760  0.4240  0.4259  0.4910  0.5730
> sd(Data$SLG)
[1] 0.0721869
> summary(Min$SLG)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.3000 0.3940  0.4230  0.4261  0.4910  0.5140
> sd(Min$SLG)
[1] 0.0568307
> summary(More$SLG)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.2490 0.3713  0.4250  0.4259  0.4913  0.5730
> sd(More$SLG)
[1] 0.07575229
>
> #Add categorical variable
> SpeedRating <- numeric(nrow(Data))
> for(i in 1:nrow(Data)) {
+   if(Data$SB[i] < 5) SpeedRating[i] <- 'slow' else
+   if(Data$SB[i] < 10) SpeedRating[i] <- 'Moderate' else
+   if(Data$SB[i] < 20) SpeedRating[i] <- 'Fast' else SpeedRating[i] <- 'Very Fast'
+ }
> length(which(SpeedRating == "slow")) #This is to show how many players I put into each category
[1] 57
> length(which(SpeedRating == "Moderate"))
[1] 18
> length(which(SpeedRating == "Fast"))
[1] 17
> length(which(SpeedRating == "Very Fast"))
[1] 15
> Data <- cbind(Data, SpeedRating)

```