Descriptive Statistics with R

Advanced Psychological Research Methods

Dr Christopher Wilson

Questions about last week’s session?

Let’s import the data

```
1 library("tidyverse")
2 Album_Sales <- read_csv("Datasets/album_sales.csv")
```

Let’s look at the data

```
1 head(Album_Sales)
```

```
# A tibble: 6 × 5
  Adverts Sales Airplay Attract Genre
    <dbl> <dbl>   <dbl>   <dbl> <chr>
1    10.3   330      43      10 Country
2   986.    120      28       7 Pop
3  1446.    360      35       7 HipHop
4  1188.    270      33       7 HipHop
5   575.    220      44       5 Metal
6   569.    170      19       5 Country
```

Let’s make sure our data types are correct #1

- This variable is currently stored as charcters, not as a factor / category variable

```
1 str(Album_Sales$Genre)
```

```
chr [1:200] "Country" "Pop" "HipHop" "HipHop" "Metal" "Country" "Pop" ...
```

Let’s make sure our data types are correct #2

- We can save it as a factor

```
1 Album_Sales$Genre <- as.factor(Album_Sales$Genre)
2
3 str(Album_Sales$Genre)
```

```
Factor w/ 4 levels "Country","HipHop",..: 1 4 2 2 3 1 4 4 3 2 ...
```

Summarising data: Central tendency

Measures of central tendency

The main measures of central tendency are:

- Mean
- Median
- Mode

Mean

“What is the mean of album sales?”

```
1  mean(Album_Sales$Sales)
```

```
[1] 193.2
```

Trimmed mean

- The trimmed mean is used to reduce the influence of outliers on the summary

```
1 mean(Album_Sales$Sales, trim = 0.05)
```

```
[1] 192.6667
```

Median

“What is the median amount of Airplay?”

```
1 median(Album_Sales$Airplay)
```

```
[1] 28
```

Mode

“What is the most common attractiveness rating of bands?”

- The easiest way to get the mode in R is to generate a frequency table

```
1  table(Album_Sales$Attract)
```

```
 1  2  3  4  5  6  7  8  9 10
 3  1  1  4 17 44 73 44 12  1
```

- We can then look for the most frequently occuring response

Measures of dispersion or variance

Range

The range is the difference between the lowest and highest values

- You can calculate it using these values

```
1 max(Album_Sales$Airplay) - min(Album_Sales$Airplay)
```

```
[1] 63
```

- Or you can use the range command to get the min and max values in one go

```
1 range(Album_Sales$Airplay)
```

```
[1]  0 63
```

Interquartile range

- We know that the median is the “middle” of the data = 50th percentile
- The interquatile range is the difference between the values at the 25th and 75th percentiles

```
1 quantile( x = Album_Sales$Airplay, probs = c(.25,.75) )
```

```
  25%   75%
19.75 36.00
```

- Interquartile range = 36 - 19.75 = 16.25

Sum of squares

- The difference between each value and the mean value, squared, and then summed together

```
1  sum( (Album_Sales$Adverts - mean(Album_Sales$Adverts))^2 )
```

```
[1] 46936335
```

Variance

- Variance: Sum of sqaures divided by n-1

```
# variance calculation
varianceAdverts <- sum( (Album_Sales$Adverts - mean(Album_Sales$Adverts))^2 ) / 199
```

Standard deviation

- Standard deviation is square root of the variance

```
# sd calculation


sqrt(varianceAdverts)
```

```
[1] 485.6552
```

- Can be calculated using the sd() command

```
sd(Album_Sales$Adverts)
```

```
[1] 485.6552
```

The *psych* package includes a lot of useful descriptive stats

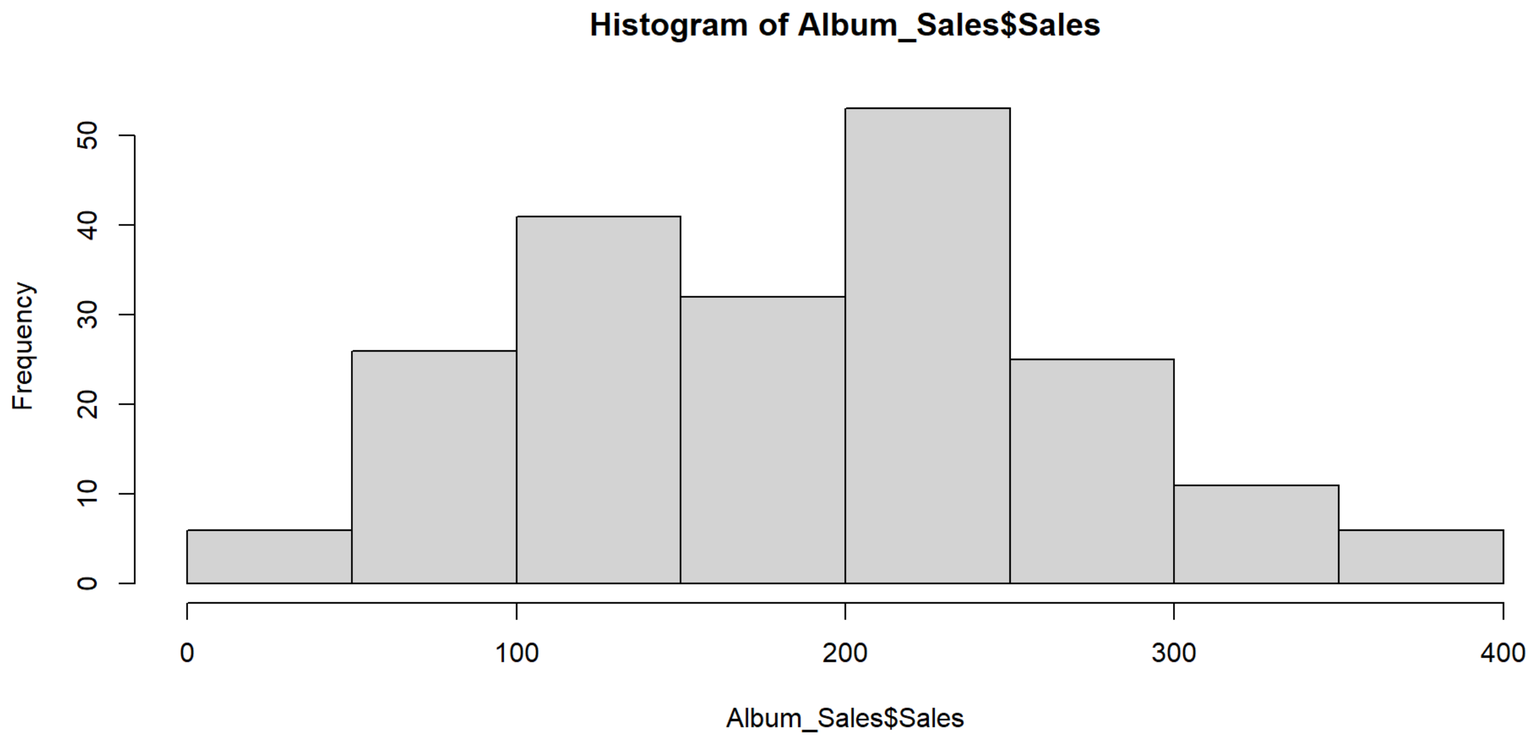
```
1 library("psych")
```

Skewness and Kurtosis

Assessing skewness of distribution #1

- It is possible to use graphs to view the distribution
- We will focus on graphic presentation of data next week

```
1 hist(Album_Sales$Sales)
```



Assessing skewness of distribution #2

- We can check raw skewness value using the *skew()* command in the **psych** package

```
1  skew(Album_Sales$Sales)
```

```
[1] 0.0432729
```

Kurtosis

informal term	technical name	kurtosis value
“too flat”	platykurtic	negative
“just pointy enough”	mesokurtic	zero
“too pointy”	leptokurtic	positive

```
1  kurtosi(Album_Sales$Sales)
```

```
[1] -0.7157339
```

Assessing normality of distribution

- We can use the shapiro-wilk test of normality
- This is part of “base” r (no package needed)

```
1  shapiro.test(Album_Sales$Sales)
```

```
    Shapiro-Wilk normality test

data:  Album_Sales$Sales
W = 0.98479, p-value = 0.02965
```

Getting an overall summary

summary() - in “base R”

```
1  summary(Album_Sales)
```

```
   Adverts            Sales            Airplay         Attract      
Min.   :   9.104   Min.   : 10.0   Min.   : 0.00   Min.   : 1.00  
1st Qu.: 215.918   1st Qu.:137.5   1st Qu.:19.75   1st Qu.: 6.00  
Median : 531.916   Median :200.0   Median :28.00   Median : 7.00  
Mean   : 614.412   Mean   :193.2   Mean   :27.50   Mean   : 6.77  
3rd Qu.: 911.226   3rd Qu.:250.0   3rd Qu.:36.00   3rd Qu.: 8.00  
Max.   :2271.860   Max.   :360.0   Max.   :63.00   Max.   :10.00  
    Genre    
Country:46  
HipHop :53  
Metal  :48  
Pop    :53  
```

describe() - in the “psych” package #1

```
1 describe(Album_Sales)
```

```
          vars   n   mean     sd median trimmed    mad  min     max   range  skew
Adverts      1 200 614.41 485.66 531.92  560.81 489.09  9.1 2271.86 2262.76  0.84
Sales        2 200 193.20  80.70 200.00  192.69  88.96 10.0  360.00  350.00  0.04
Airplay      3 200  27.50  12.27  28.00   27.46  11.86  0.0   63.00   63.00  0.06
Attract      4 200   6.77   1.40   7.00    6.88   1.48  1.0   10.00    9.00 -1.27
Genre*       5 200   2.54   1.12   3.00    2.55   1.48  1.0    4.00    3.00 -0.02
          kurtosis    se
Adverts      -0.17 34.34
Sales        -0.72  5.71
Airplay      -0.09  0.87
Attract       3.56  0.10
Genre*       -1.37  0.08
```

describe() - in the “psych” package #2

- We can describe by factor variables

```
1 describeBy(Album_Sales, group = Album_Sales$Genre)
```

```
 Descriptive statistics by group
group: Country
         vars  n   mean     sd median trimmed    mad  min     max   range  skew
Adverts     1 46 656.22 507.96 574.14  620.40 581.96  9.1 1985.12 1976.01  0.51
Sales       2 46 201.74  73.64 210.00  200.79  66.72 60.0  360.00  300.00  0.03
Airplay     3 46  29.07  10.53  28.00   28.50  11.12  9.0   54.00   45.00  0.44
Attract     4 46   6.52   1.63   7.00    6.71   1.48  1.0   10.00    9.00 -1.49
Genre*      5 46   1.00   0.00   1.00    1.00   0.00  1.0    1.00    0.00   NaN
         kurtosis    se
Adverts     -0.65 74.89
Sales       -0.52 10.86
Airplay     -0.10  1.55
Attract      3.54  0.24
Genre*        NaN  0.00
```

Questions?