PSY 3307

Frequency Distributions

Jonathan A. Pedroza, MS, MA

Cal Poly Pomona

2021-08-24

Straight into Terms

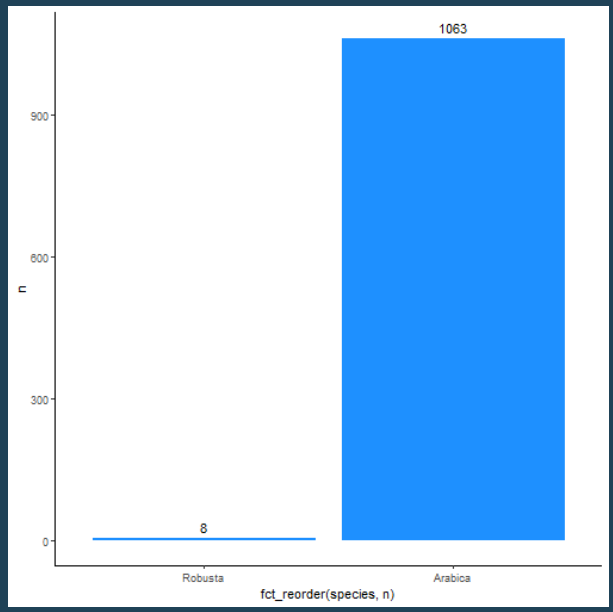
- **raw score** is the score given to a participant
- **Frequency** denoted as f; number of times a score occurs/is counted

Note. Not F, that is something completely different.

- **Frequency Distribution** is a distribution of each score and the number of times the score has occurred/is counted

Example

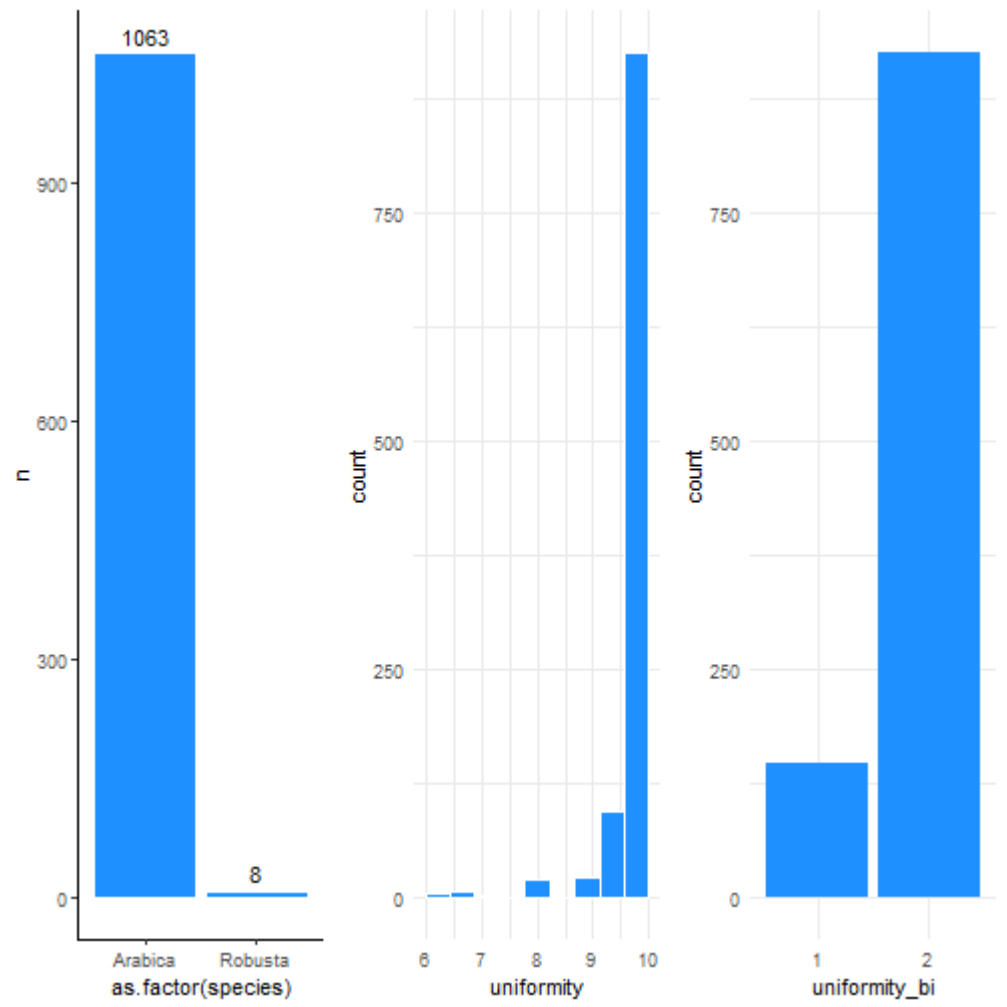
species	n
Arabica	1063
Robusta	8



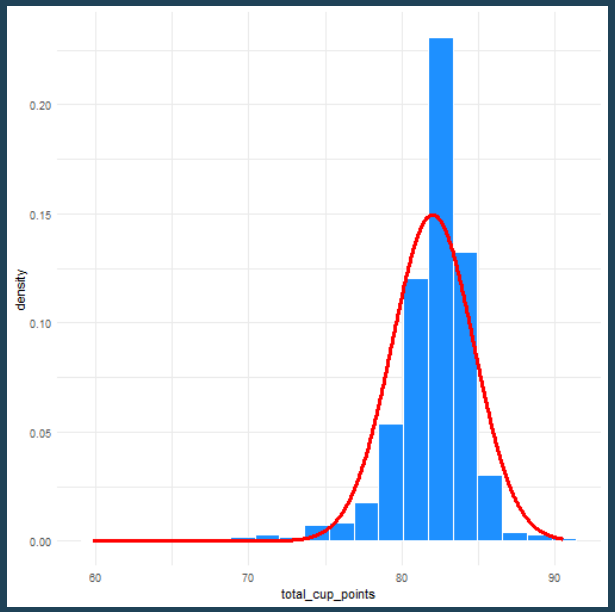
Visualizing Frequencies

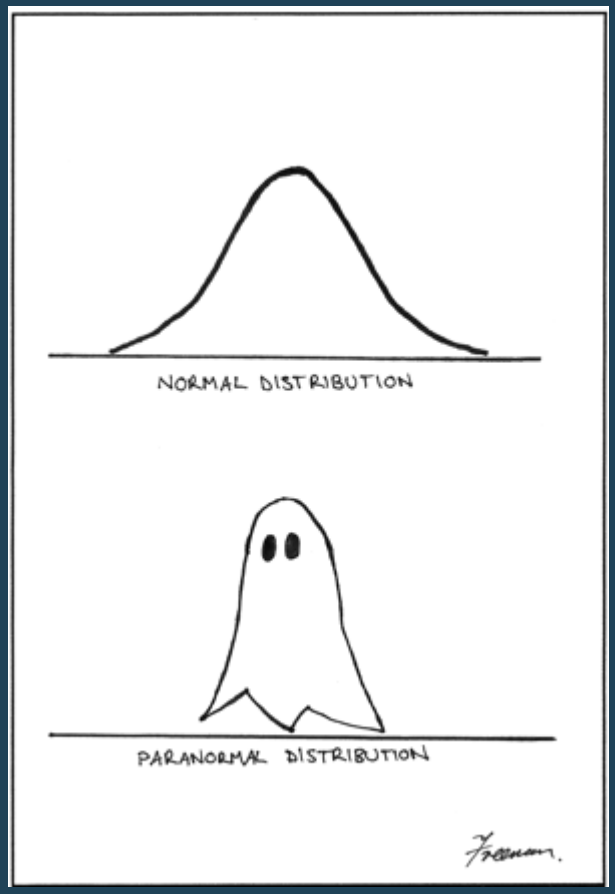
- Best way of visualizing frequencies is by using a bar graph
- **Bar graph** graph with vertical bar over each nominal/ordinal category
- **Side Note** A pie graph will always be inferior to a bar graph/any other visual
- **Histogram** is a frequency graph used for interval or ratio scores
- **Frequency Polygon** similar to a histogram, which shows data points connected with straight lines
- **Grouped Distributions** put continuous data into categories
- The next slide has data ranging from 6-10 in coffee uniformity; I could lump them as anything below a perfect 10 (6-9) and then 10 as a different category

Bar Graph ≠ Histogram



"Normal" Distribution



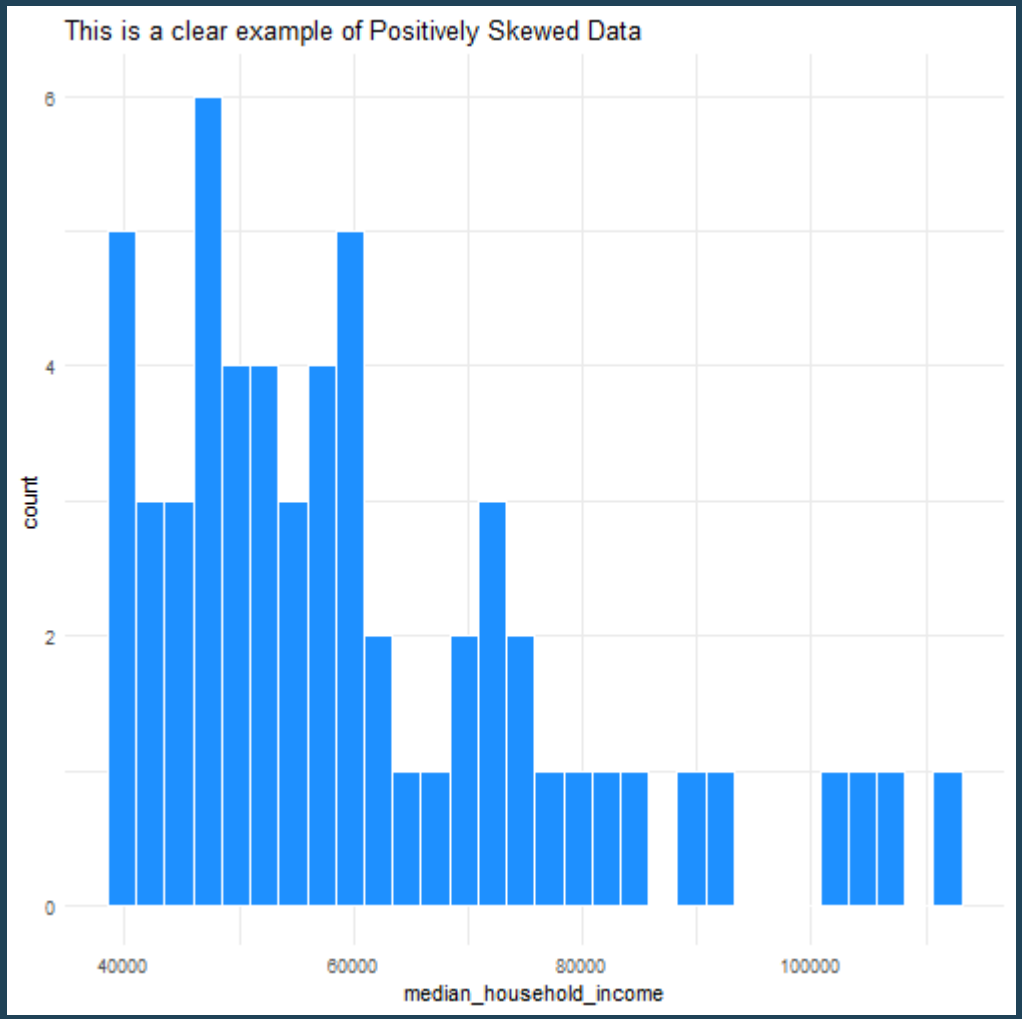


Why Is It a Normal Distribution?

- **Normal curve** is often called the bell-shaped curve; is symmetrical
- **Normal Distribution** same thing as normal curve; represents the population because if you have enough data you will get a normal distribution (central limit theorem); if your data looks like this in a histogram, you're in good shape
- **Distribution Tail** has two tails; these will be more important for statistics

Skewed Distributions

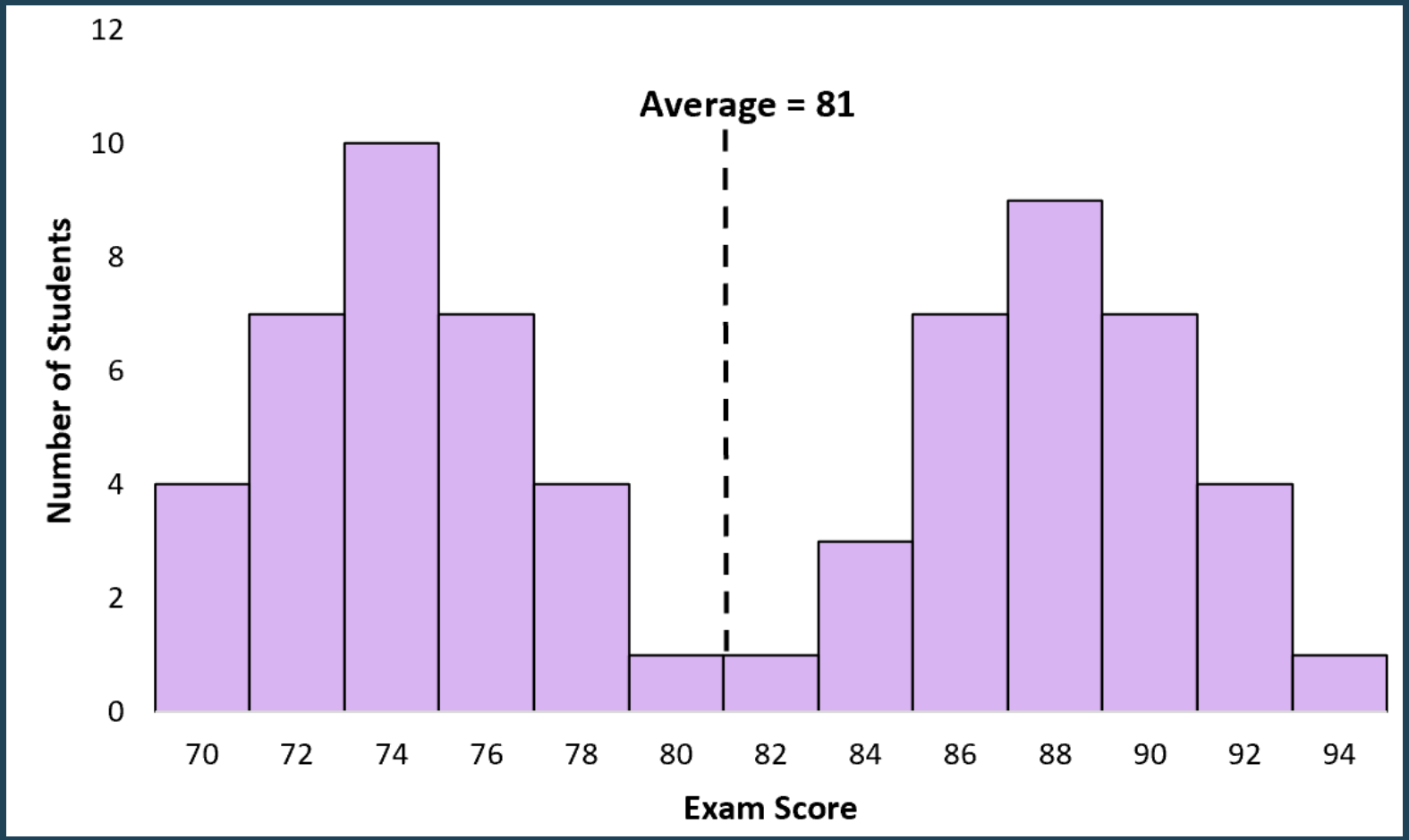
- **Negative Skew** is like a finger pointing left; not normal and is asymmetrical; indicates higher frequency of middle and higher scores; no low frequency in higher scores
- **Positive Skew** is like a finger pointing right; not normal and is asymmetrical; indicates higher frequency of low and middle scores; no low frequency in lower scores
- Some thresholds are that if you have a skewness value of +-2 or +-3 then you're good to use that variable like it is.
- **Kurtosis** is when your frequency are really skinny and tall or really flat and wide


This is a clear example of Positively Skewed Data
count
median_household_income
0
2
4
6
40000
60000
80000
100000

```
##    skew
## X1 1.08
```

Bimodal Distribution

- **Bimodal Distribution** is when your distribution has two humps with a valley in the middle; high frequencies both below and above the middle of the plot


Average = 81
Number of Students
Exam Score
0
2
4
6
8
10
12
70
72
74
76
78
80
82
84
86
88
90
92
94

Some Notes From JP

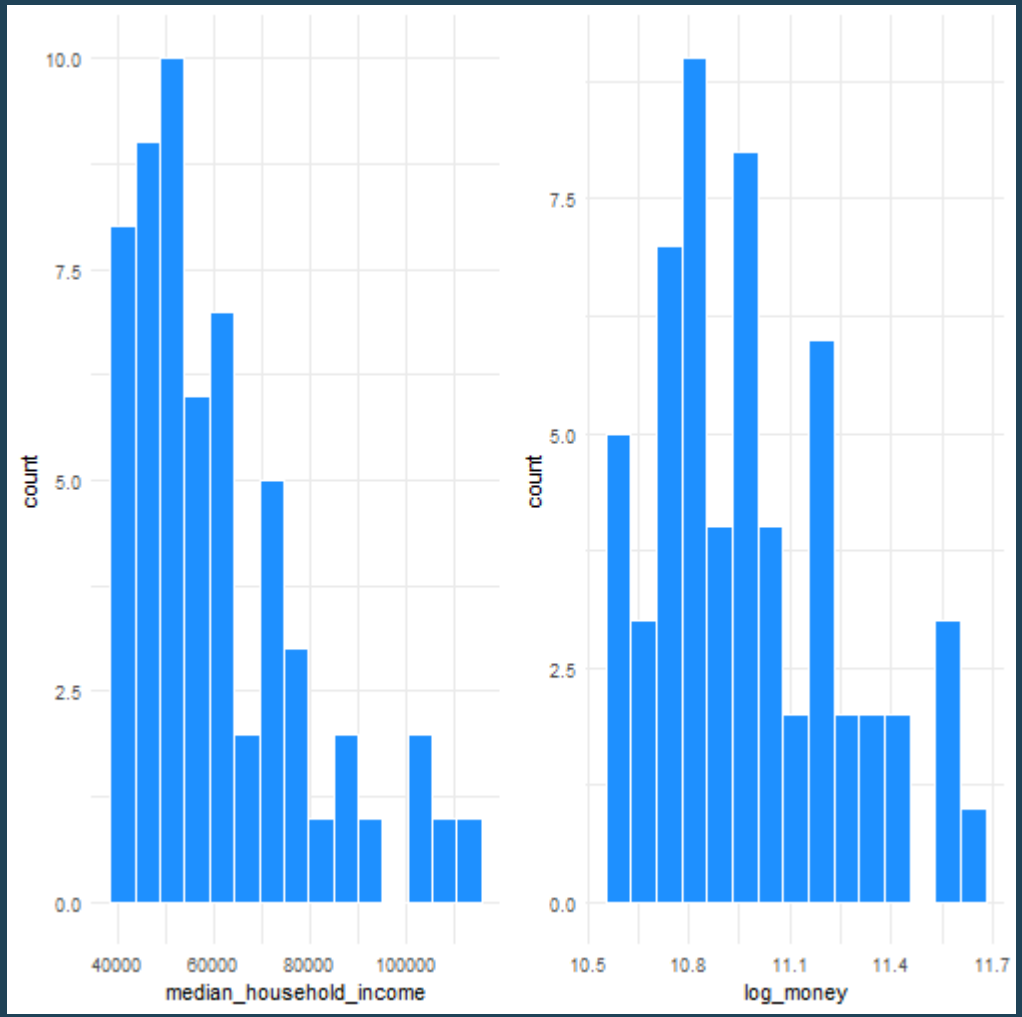
Deciding what is positively and negatively skewed from visuals alone is not good enough

You'll want to look into your descriptive statistics and look at skewness and kurtosis to make sure they are within +-3

Some statistics can handle some skewness and kurtosis

Other times you'll have to transform the variable using fancy methods that we will not talk about in this class.

Example of Transformations



Relative Frequency

- **Relative Frequency** the proportion of times a score occures/is counted in the distribution

$$Relative\ frequency = f/N$$

Here, f is the frequency of one category of a nominal variable. N is the total number of observations for all the categories of that variable.

```
f = 37
N = 2000

relative_frequency = f/N
relative_frequency
```

```
## [1] 0.0185
```

```
percent = relative_frequency*100
percent
```

```
## [1] 1.85
```

Can Also Calculate Simple Frequency

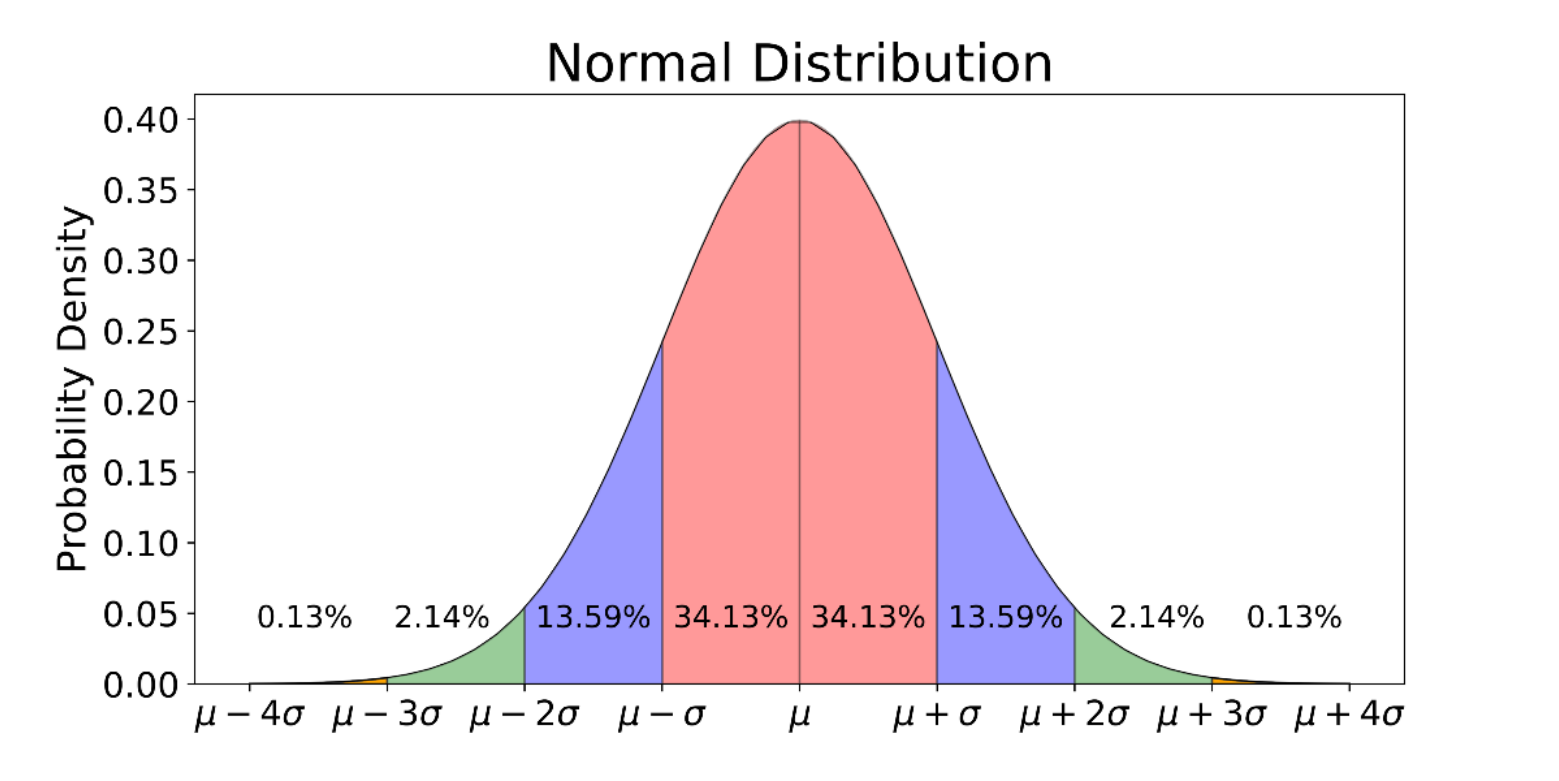
```
simple_frequency = relative_frequency*N
simple_frequency
```

```
## [1] 37
```

Let's Use Some Real Data

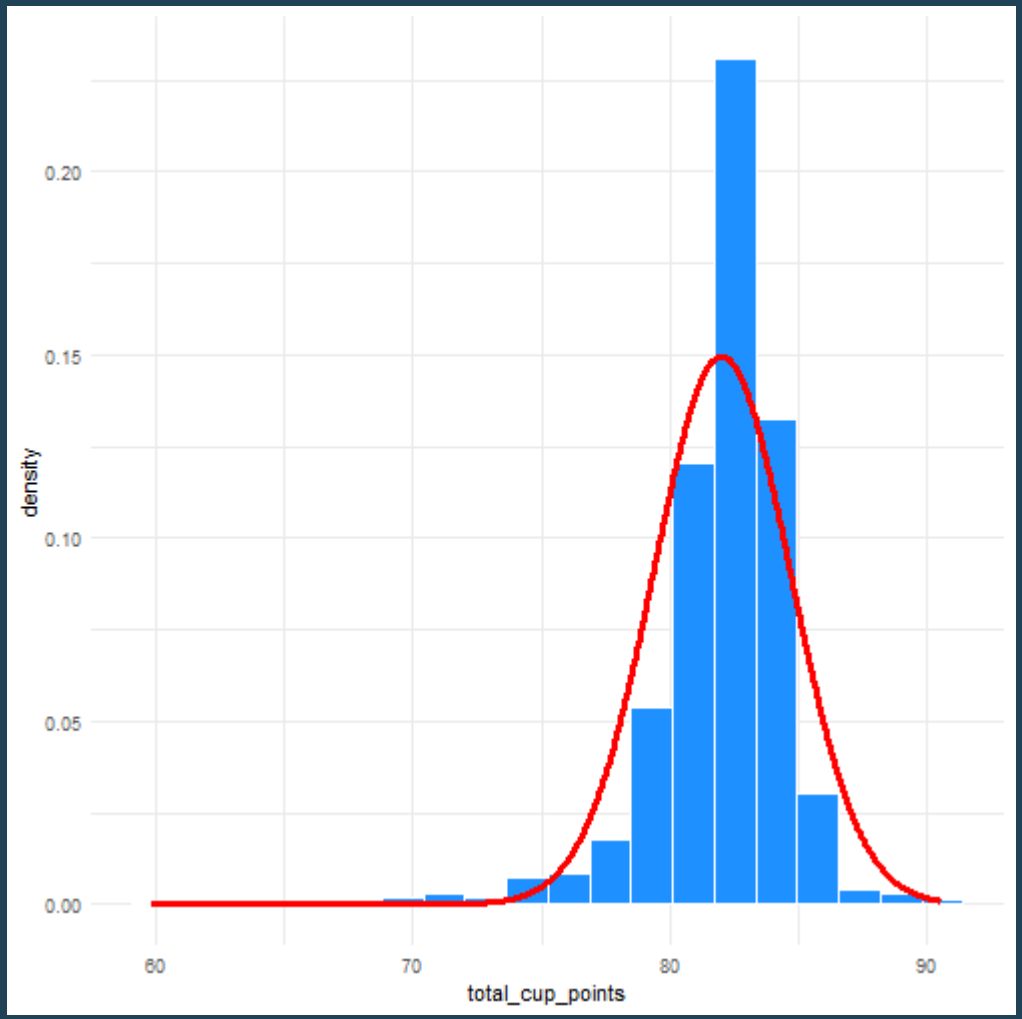
species	n	total	rel_freq	percent
Arabica	1063	1071	0.9925303	99.2530345
Robusta	8	1071	0.0074697	0.7469655

Relative Frequency Using Normal Curve



Relative Frequency Using Normal Curve

- **Proportion of Area under the Curve** is the proportion of total area under the normal curve
- **Percentile** is the percentage of all scores in the sample below a particular score
- **Cumulative Frequency** is the number of scores in the data at or below a particular score

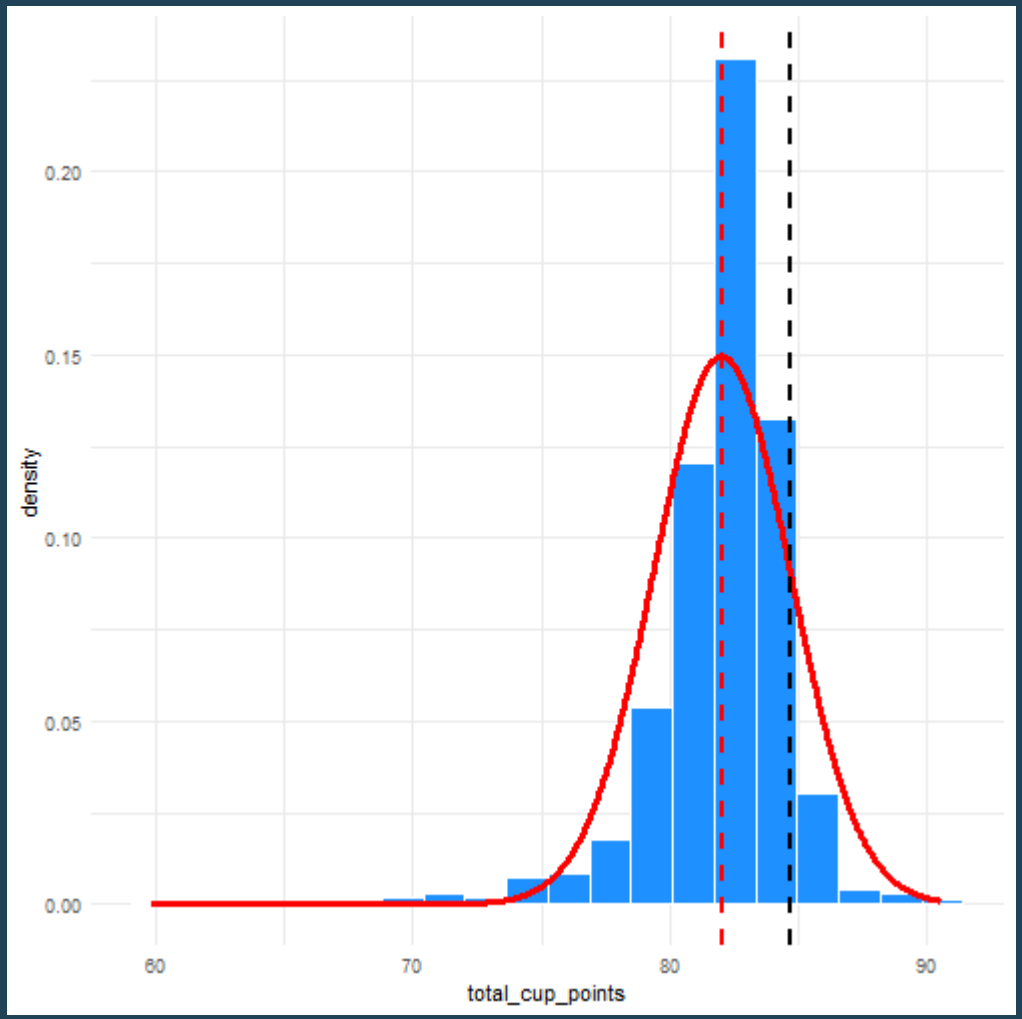

0.20
0.15
0.10
0.05
0.00
density
60
70
80
90
total_cup_points

```
psych::describe(coffee$total_cup_points, na.rm = TRUE)
```

```
##    vars    n  mean   sd median trimmed  mad   min   max range  skew kurtosi
## X1    1 1071 82.03 2.67  82.42    82.3 1.85 59.83 90.58 30.75 -2.11    10.5
##      se
## X1 0.08
```

```
sd_plus1 = 82.03 + 2.67
sd_plus1
```

```
## [1] 84.7
```


0.20
0.15
0.10
0.05
0.00
density
60
70
80
90
total_cup_points

```
coffee %>%
  filter(total_cup_points < 84.7) %>%
  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   994
```

```
cummulative_freq = 994/1071
cummulative_freq
```

```
## [1] 0.9281046
```

```
percentile = cummulative_freq*100
percentile
```

```
## [1] 92.81046
```