Stats Intro:

Population Vs Sample

Parameters Vs Statistics

Variables: Independent and Dependent

Variables:

Qualitative: Nominal vs Ordinal

Quantitative: Discrete vs Continuous

Descriptive Stats- summarize, organize and simply data from sample.

Outliers

Measures of Central Tendency (units must be same) -Python will do this for you.

Mean- (Con: influenced by outliers)- use median if outlier exists.

Median- (Cons: may not capture important fluctuations in data)

If total value is even: to find mid points: n/2 and n/2+1

Mode: Most occurring value (best with categorical data)

-uni, bi, tri, multi

-Pro: works for both qaunt and qual, not affected by outliers

-Con: Not useful measure

Measure of Variability

Range- Max – min=Range (not a good measure)

Percentile- How a value compares to other values. (python)

**Step 1:** order the values from low to high

**Step 2:** Count the number of values in the dataset (n)

**Step 3:** calculate the value of k \* n / 100 (index)

**Step 4:** If the index is not an integer  then round to the nearest whole number, then go to Step 6. If the index is an integer, then go to Step 5.

**Step 5:** Count the values in your data set from left to right until you reach the number. Then find the mean for that corresponding number and the next number. The resultant value is the kth percentile of your data set.

**Step 6:** Count the values in your data set from left to right until you reach the number. The obtained value will be the kth percentile of your data set.

Ex: In a college, a list of grades of 25 students has been declared. Their grades are given as:

43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. Find the 25th percentile?

Step 1: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99

Step 2: n =25

Step 3: 25\*25/100 = 625/100 = 6.25

Step 4: 6

25th percentile is 66%

Quartiles Q1-2-3-4

Interquartile Range (not affected by outliers)

IQR= Q3-Q1= Upper Q – Lower Q

Median= Q2 =50%

Q1 and Q3 will be on other sides of Q2

Variance (how far the point is away from mean)- tells you how spread out the data is.

It is average of squared distances from the mean. To find the variance, follow these five steps:

Calculate the variance for these final exam scores.

24, 58, 61, 61, 67, 73, 76, 79, 82, 83, 85, 87, 88, 88, 92, 93, 94, 95 - > n =18

**Step 1:** Calculate the mean

Step 1: mean = 1386/18 = 77

**Step 2:** Subtract the mean from each value individually to get the deviation from the mean

Step 2: -53, -19, -16, -16, -10, -3, -1, 3, 5, 6, 8, 10, 11, 11, 15, 16, 17, 18

**Step 3:** Square each deviation

Step 3: 2809, 361, 256, 256, 100, 9, 1, 9, 25, 36, 64, 100, 121, 121, 225, 256, 289, 324

**Step 4:** Add all deviations.

Step 4: 5362

**Step 5:** Divide the addition by (n -1) where n is the number of samples in your dataset

Step 5: 5362/(18-1) = 5362/17 = 315.4

Standard dev (on average how far the point is away from mean) (impacted by outliers)

* If SD is high, then data is spread out from each other
* Id SD is low, then data is close to each other

Multiple data sets?

Coefficient of Variation (unit free). (higher the better) (POPULAR)

-Higher the value, the greater of dispersion. Meaning less variation in the set.

CV= Standard Deviation/Mean

Do this for all Sample sets and compare the values. Use the bigger value. Meaning less variation in the set.

-never compare SD of multi sample sets. Only compare DV.

Box plot:

**Minimum**: The minimum value in a dataset

**Q1**: The first quartile

**Median:** The median is the middle value of the dataset. It is also called second quartile

**Q3:** The third quartile

**Maximum:** The maximum value in a dataset

**IQR:** The difference between the third quartile and first quartile

**Outlier:** The data that falls on the far left or right side of the ordered data. Need min and max Q1-1.5 x IQR and Q3+1.5x IRQ resp to calculate

Ex: A sample of  10  boxes of raisins has these weights (in grams): 25, 28, 29, 29, 30, 34, 35, 35, 37, 38.

Solution:

25, 28, 29, 29, 30, 34, 35, 35, 37, 38.

Minimum = 25

Q1 = 29

Median = (30+34)/2 = 32

Q3 = 35

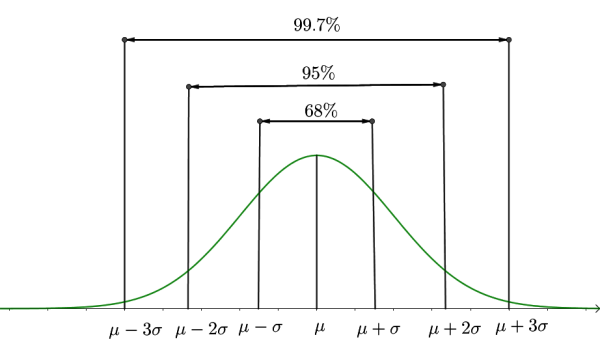
Maximum = 38

IQR = Q3 - Q1 = 35 - 29 = 6

Normal distribution: - most data point cluster in mid. Other left or right. the [mean](https://www.w3schools.com/statistics/statistics_mean.php) = μ and the standard deviation= σ

* Symmetric bell shape
* Mean, median and mode are equal. All located at the center of the distribution
* Approximately 68% of the data falls within 1 standard deviation of the mean
* Approximately 95% of the data falls within 1 standard deviation of the mean
* Approximately 99.7% of the data falls within 1 standard deviation of the mean

Empirical Rule in norm distributio: 68/95/99. If you can see where 99.7% of data. Then you can calculate outliers.



Skewness- helps to measure symmetry of dataset. (tails) 0 skewness = data is symmetrical. Mean=median=mode

Chart, line chart

Description automatically generated

-Negative skew: left tail longer= most data is on right side. Mode>median>mean

A picture containing diagram

Description automatically generated

-Positive skew: right tail is longer= most data is on the left side : Mean>median> mode

A picture containing chart

Description automatically generated

Negative People are Mean

Finding Outliers

* Outliers are greater than μ + 3\*σ or less than μ - 3\*σ (think of out side of the 99.7%)
* **(Tuckey’s method)** Outliers are greater than Q3+ 1.5 \* IQR or less than Q1- 1.5 \* IQR

If not all the outliers are found, increase the numbers. Ex Q3+2\*IRQ, Q3+2.5\*IRQ

Or μ + 4\*σ

Find the **outliers** in the following data points: 5, 93,  94, 95, 96, 99, 104, 105, 110, 199

* μ (mean) = 100
* σ (standard deviation)= 46.06

* μ + 3\*σ =  238.18
* μ - 3\*σ = -38.18

Since there are no points in the data that are greater than these points, that means there are no outliers.

1B

2D

3B

4A

5B

67,6

7C

860

9116. 700 is the out lier

10How wide the data points are. Max -min

11B

13B

1456.57, SD 7.5

153

164

16Min-3, Q1 4, Med- 10.5, Q3 15, IQR 11

18Min- 2, Q1 5, Med- 12, Q3 19, IQR- 14

19A.

20B

21C

2232,00 ,28,00

23None

2442