STATISTICS FUNDAMENTALS

LEARNING OBJECTIVES

PART I

- Stats about a given dataset: mean, median, standard deviation and correlation
- Visualize: Boxplot and Histograms
- Create a notebook, analyze a dataset, use git to commit and push the notebook

PART II: TRANSFORMATIONS

- Normal Distributions
- Dealing with categorical variables

PRE WORK & REVIEW

LAST LESSON REVIEW

- Frame Good Questions with S.M.A.R.T
- Study types: Cross sectional vs longitudinal
- Numpy and Pandas

QUESTIONS?

ANY QUESTIONS FROM LAST CLASS?

QUESTIONS FROM EXIT TICKET

STATISTICS FUNDAMENTALS

MEAN / AVERAGE

- Sum of all values divided by number of values $\bar{X} = \frac{1}{N} \sum X_i$
- Sensitive to outliers
- Other means: Harmonic mean $H = \frac{n}{\sum \frac{1}{x_i}}$
- In probability Expectation of X: $\mathbb{E}[X]$

MEDIAN

The median refers to the midpoint in a series of numbers.

Arrange the numbers in order smallest to largest.

- If there is an odd number of values, the middle value is the median.
- If there is an even number of values, the average of the middle two values is the median.

MEDIAN EXAMPLES

- $a = [1, 2, 3, 4, 5] => median(a) = 3$
- $a = [1, 2, 3, 4, 5, 6] => median(a) = \frac{3 + 4}{2}$
- $a = [10, 12, 18, 28, 32, 34, 36, 40, 1000, 10000]$

  $$median(a) = 33$$

  but

  $$\bar{a} = 1121$$

More robust to outliers than mean

STANDARD DEVIATION

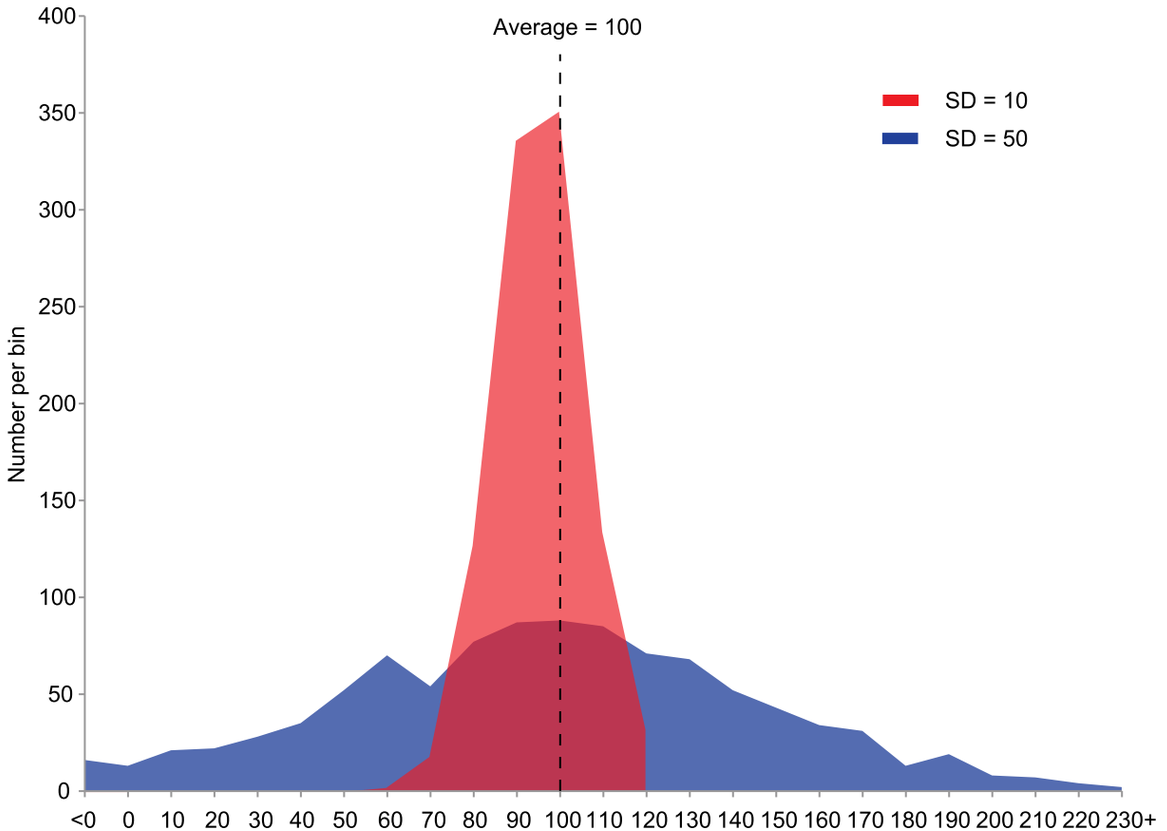
Standard deviation σ measures the amount of variation of a set of data values around its mean.

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2\right]}, \text{ where } \mu = \frac{1}{N}(x_1+\cdots+x_N)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N}x_i.$$

STANDARD DEVIATION

Both datasets have a mean of 100 but different SDs.



BIAS VS UNBIASED STANDARD DEVIATION

https://en.wikipedia.org/wiki/Bias_of_an_estimator

Biased: $\hat{\sigma} = \frac{1}{N} \sum (x_i - \bar{x})^2$

Unbiased: $\hat{\sigma} = \frac{1}{N-1} \sum (x_i - \bar{x})^2$

CORRELATION

Correlation is a measure of the dependence between 2 variables X and Y.

- X increases => Y increases
  - positive correlation, think children's age => height
- X increases => Y decreases
  - negative correlation, think adult's age => eyesight

INDEPENDENT VARIABLES

CORRELATION

Different ways to calculate correlation.

Most common one is Pearson Correlation

$$X = x_1, \ldots . x_N \qquad Y = y_1, \ldots, y_N$$

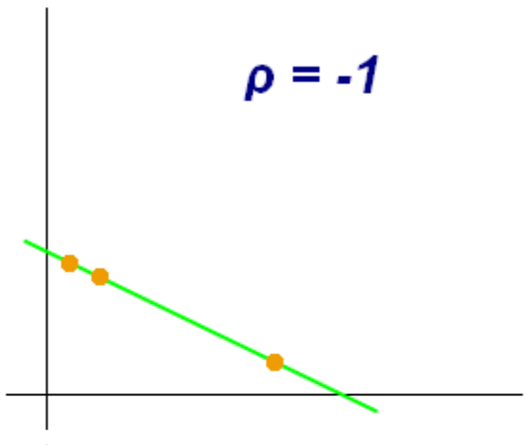
Correlation between X and Y:

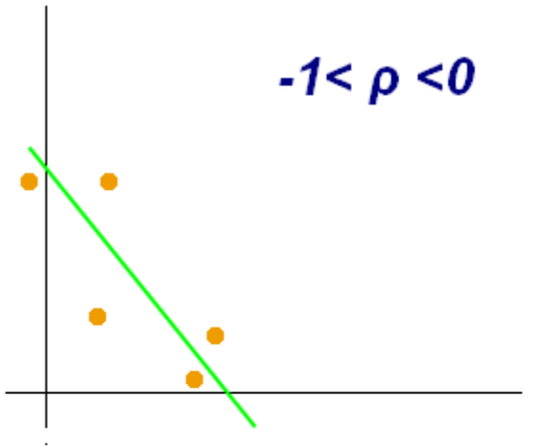
$$\rho_{X,Y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

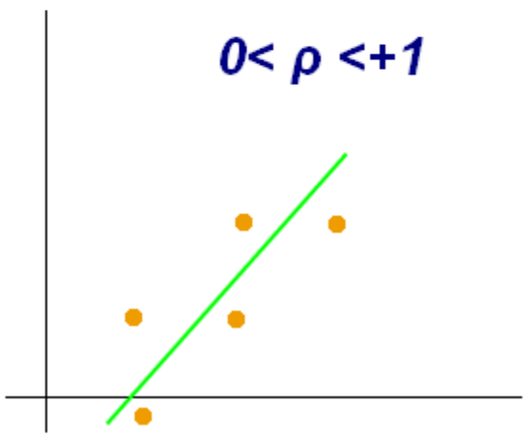
where

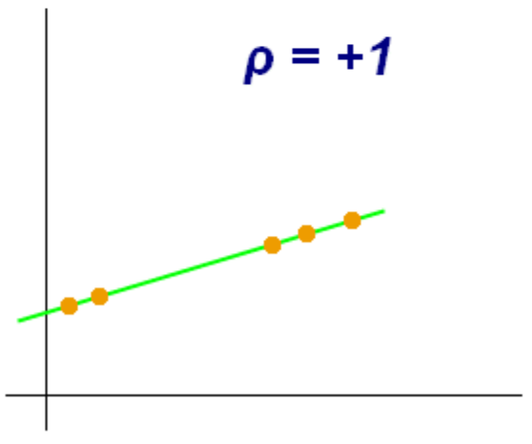
$\bar{x} = \frac{1}{N}\sum x_i$ *and* $\bar{y} = \frac{1}{N}\sum y_i$

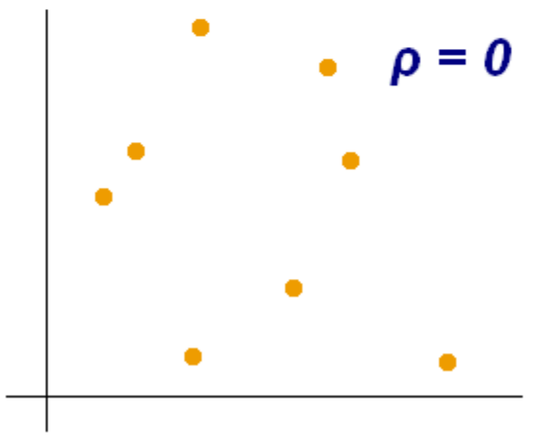
CORRELATION









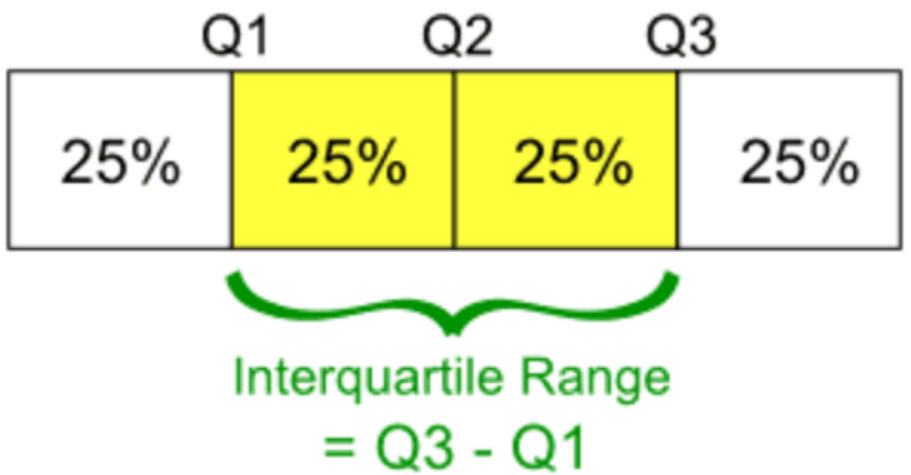


QUARTILES AND INTERQUARTILE RANGE

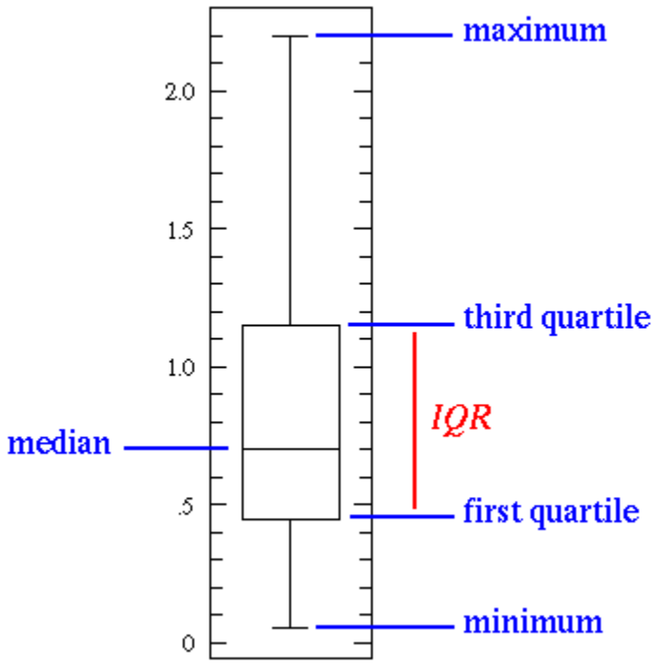
Quartiles divide a rank-ordered data set into four equal parts.

The values that divide each part are called first, second, and third quartiles, denoted Q1, Q2, and Q3, respectively.

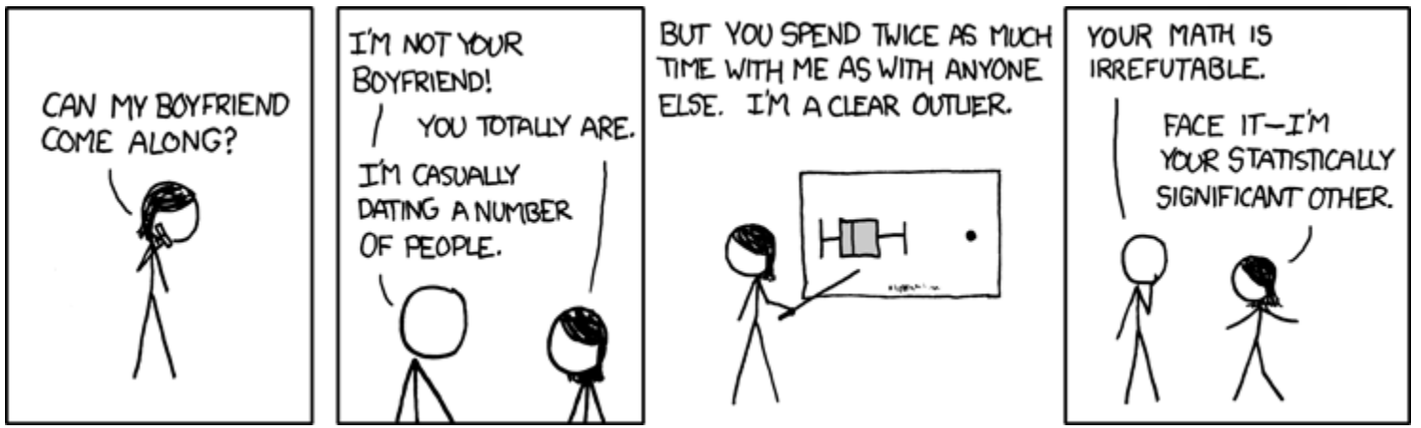
The interquartile range (IQR) is Q3 - Q1, a measure of variability.



BOXPLOT


maximum
2.0
1.5
third quartile
1.0
IQR
median
.5
first quartile
0
minimum

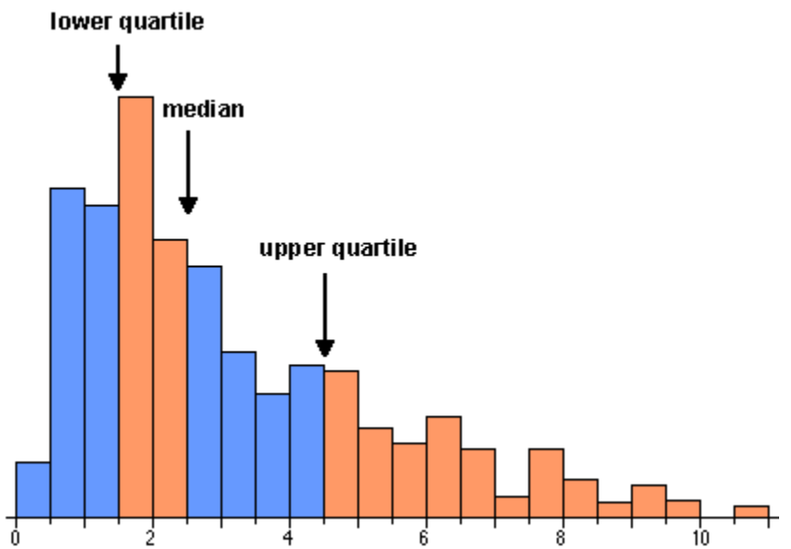
BOXPLOT


CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

HISTOGRAMS

- slice your data into N range
- for each subrange you count the number of samples in that range

Approximation of the distribution of the variable.



YOUR TURN!

- jupyter notebook

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

Load the Online Retail dataset
https://github.com/alexperrier/gads/blob/master/03_statistics_fundamentals/data/online_

LAB: CLEAN THE DATA

Let's clean the data

1. Remove rows with:
    - Negative Quantities
    - Unit Prices of 0
    - Rows with Missing values

```
How many remaining rows?
```

2. Stats: Mean, median, std, ... for Quantity and UnitPrice (df.describe())

```
Anything striking ?
```

3. More cleaning maybe ?

```
limit to the 95% quantile: df.UnitPrice.quantile(0.95) & df.Quantity.quantile(0.95
```

LAB: SOME VISUALIZATIONS

1. Plot the histograms and boxplot for Quantity and UnitPrice

```
label your axes with plt.xlabel('...')
```

2. Interpret correlation graph using sns

```
import seaborn as sns
f, ax = plt.subplots(figsize=(16, 16))
sns.corrplot(df.sample(1000)[['Quantity','UnitPrice']], ax=ax)
```

GIT

GIT

GIT

GIT

Part I

1. Create new repo on github
2. Initialize, first commit and push

```
git init
touch README.md
git add .
git commit -m "First commit"
```

3. Save the notebook in the local folder where you have your git repo
4. Commit and push

```
git add .
git commit -m "Data Exploration"
git push origin master
```

5. Check it is on github

Part II

1. Clone your neighbors repo

```
# move up
cd ..
git clone .....
```

2. add a file to it

```
# Create a new file
touch new_file.txt
# Add the file
git add .
```

3. commit and push

```
git commit -m "Your commit message"
git push origin master
```

4. Repo owner: pull your repo and check the file is in there

```
git pull origin master
ls -al
```

CATEGORICAL VS CONTINUOUS

CONTINOUS

Numeric variables can take on a large range of non-predetermined, quantitative values. These are things such as height, income, etc.

CATEGORICAL

Categorical variables can take on a specific set of variables.

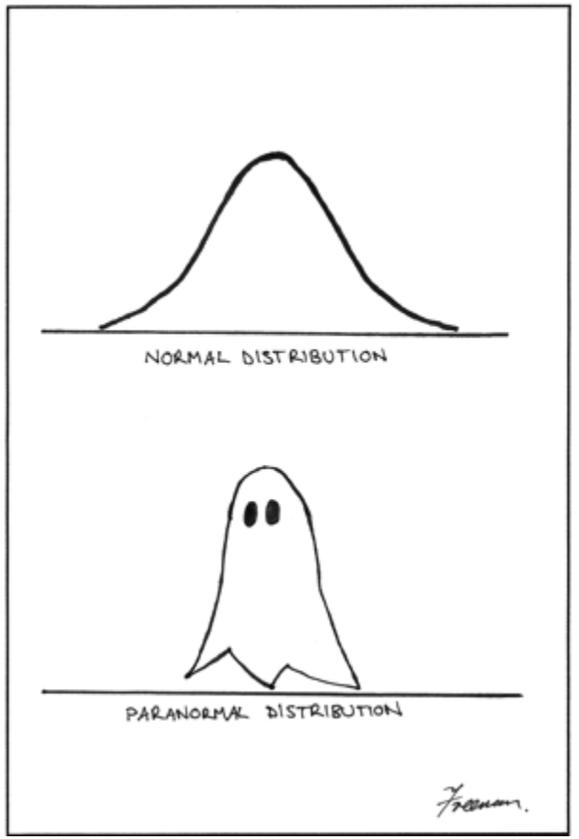
These are things such as gender, colors, countries, fare level, courses, music genre, housing types, ...

CONTINOUS TO CATEGORICAL

- Age [0 to 120] becomes Age Group [0-18, 18-25, 25-35, ...]
- Free text becomes categories

NORMAL DISTRIBUTION

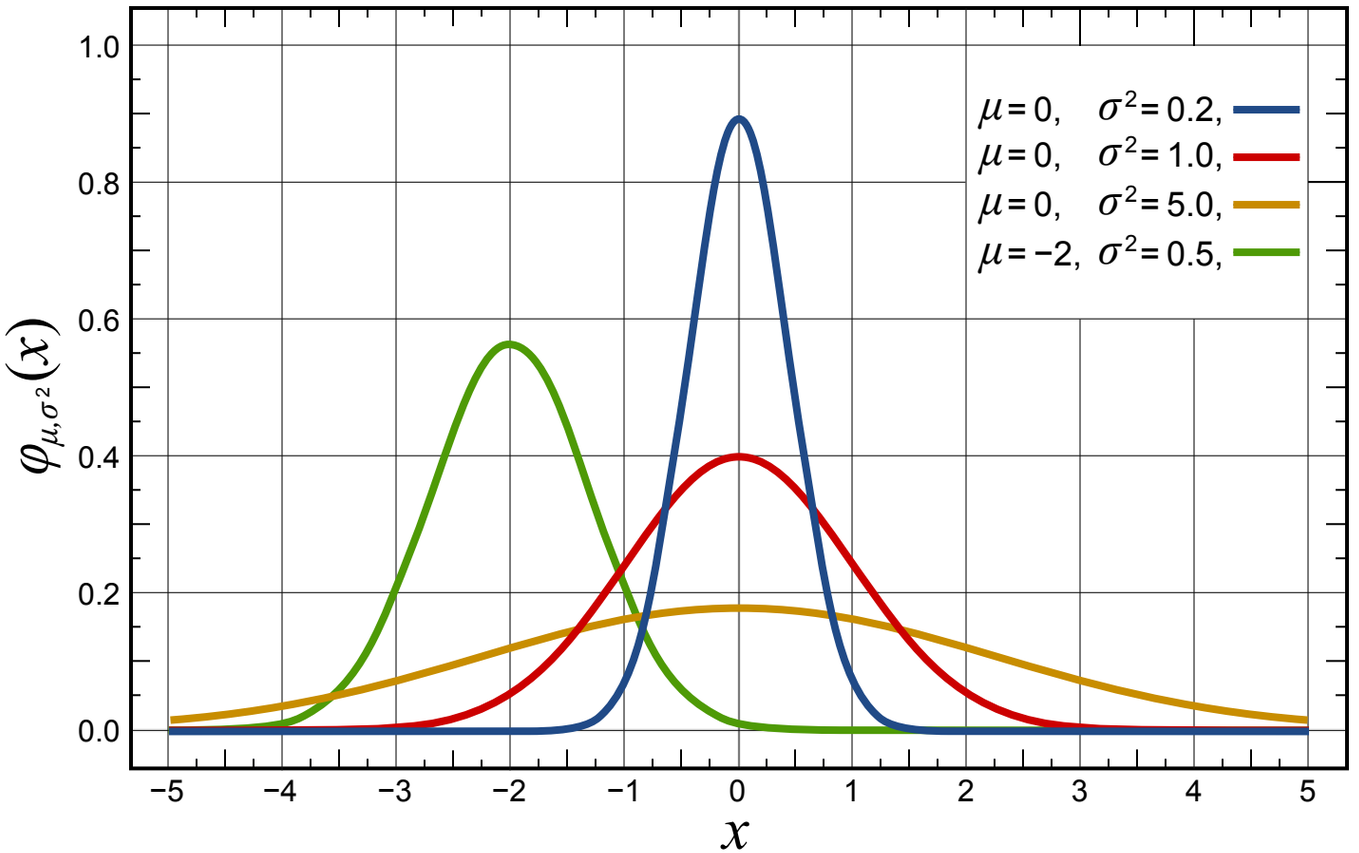
NORMAL DISTRIBUTION



NORMAL DISTRIBUTION

$\mathcal{N}(\mu, \sigma^2)$

- Mean μ determines the center of the distribution.
- Standard deviation σ^2 determines the height and width of the distribution.

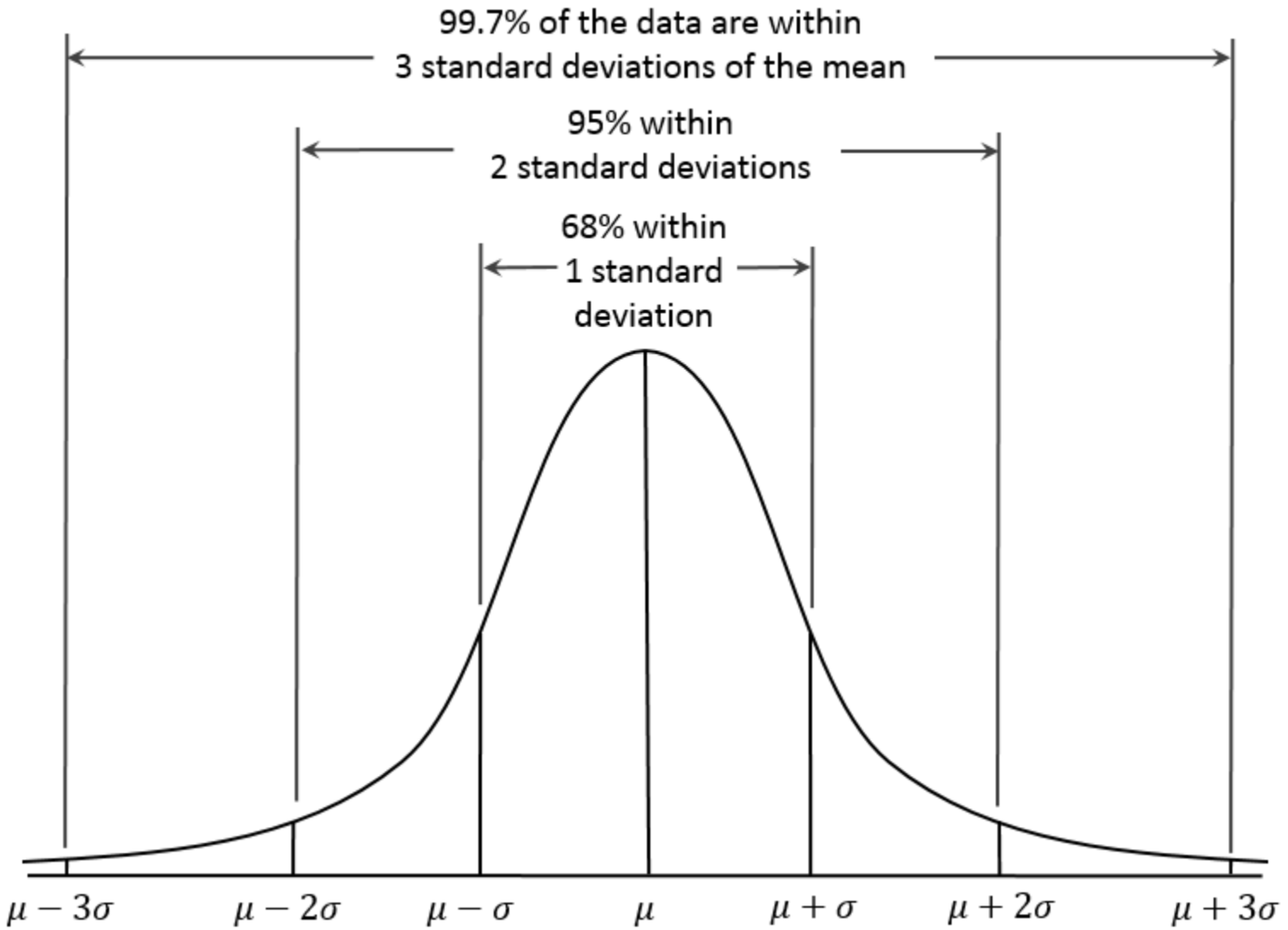

1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
$\varphi_{\mu,\sigma^2}(x)$
$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$

CENTRAL LIMIT THEOREM

- Large number of samples
- Each sample is independent on the values of the other observations
- The mean of the samples follows a Normal Distribution with the same mean
- and variance equal to the variance of the samples divided by the sample size.

http://blog.vctr.me/posts/central-limit-theorem.html

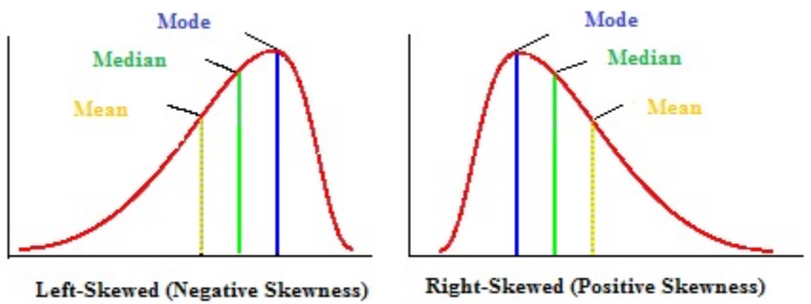
NORMAL DISTRIBUTION



SKEWNESS

Two metrics are commonly used to describe your distribution: skewness and kurtosis.

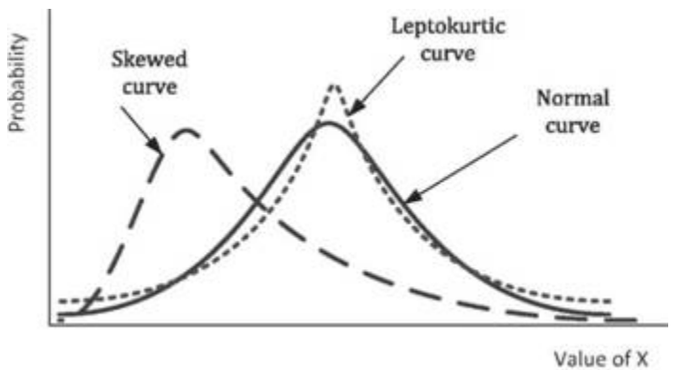
Skewness a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.



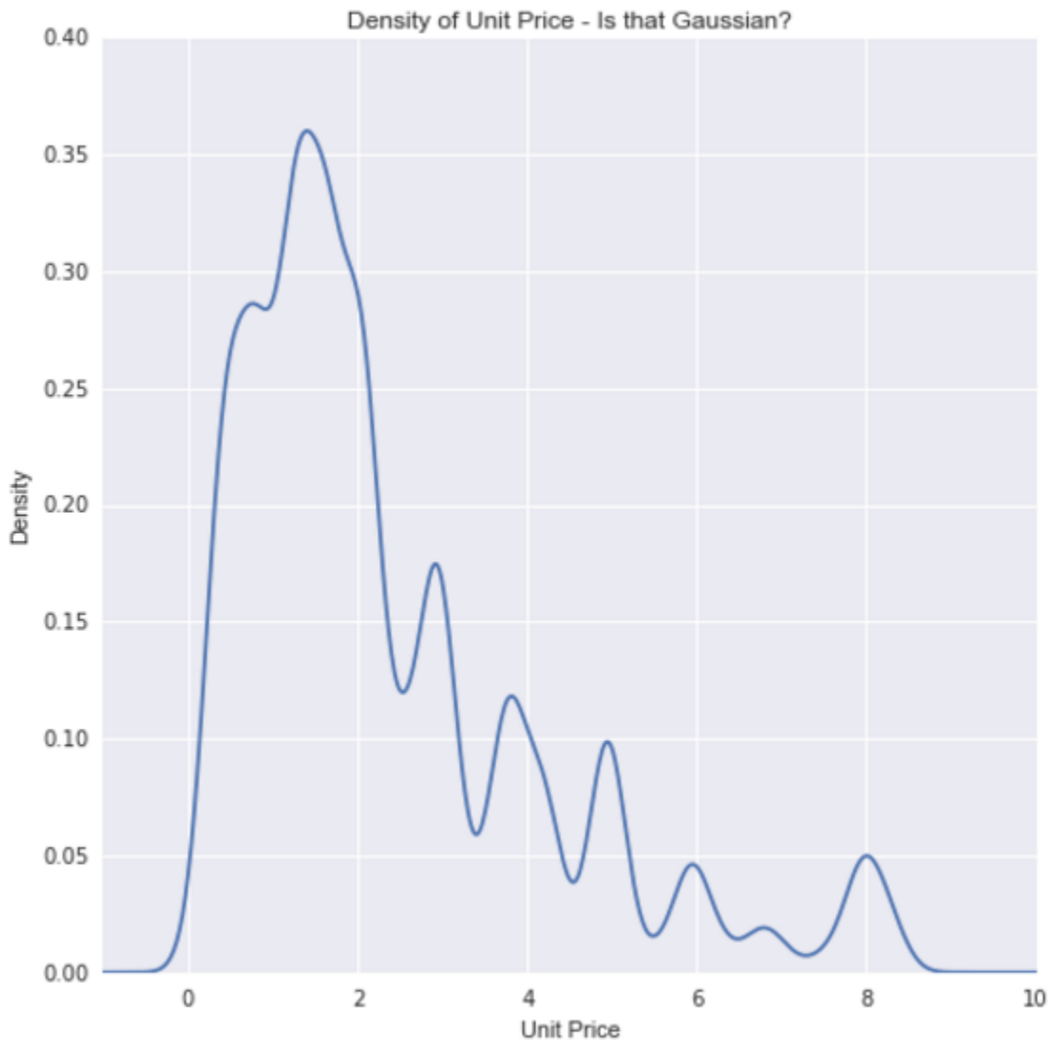
KURTOSIS

Kurtosis

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails.



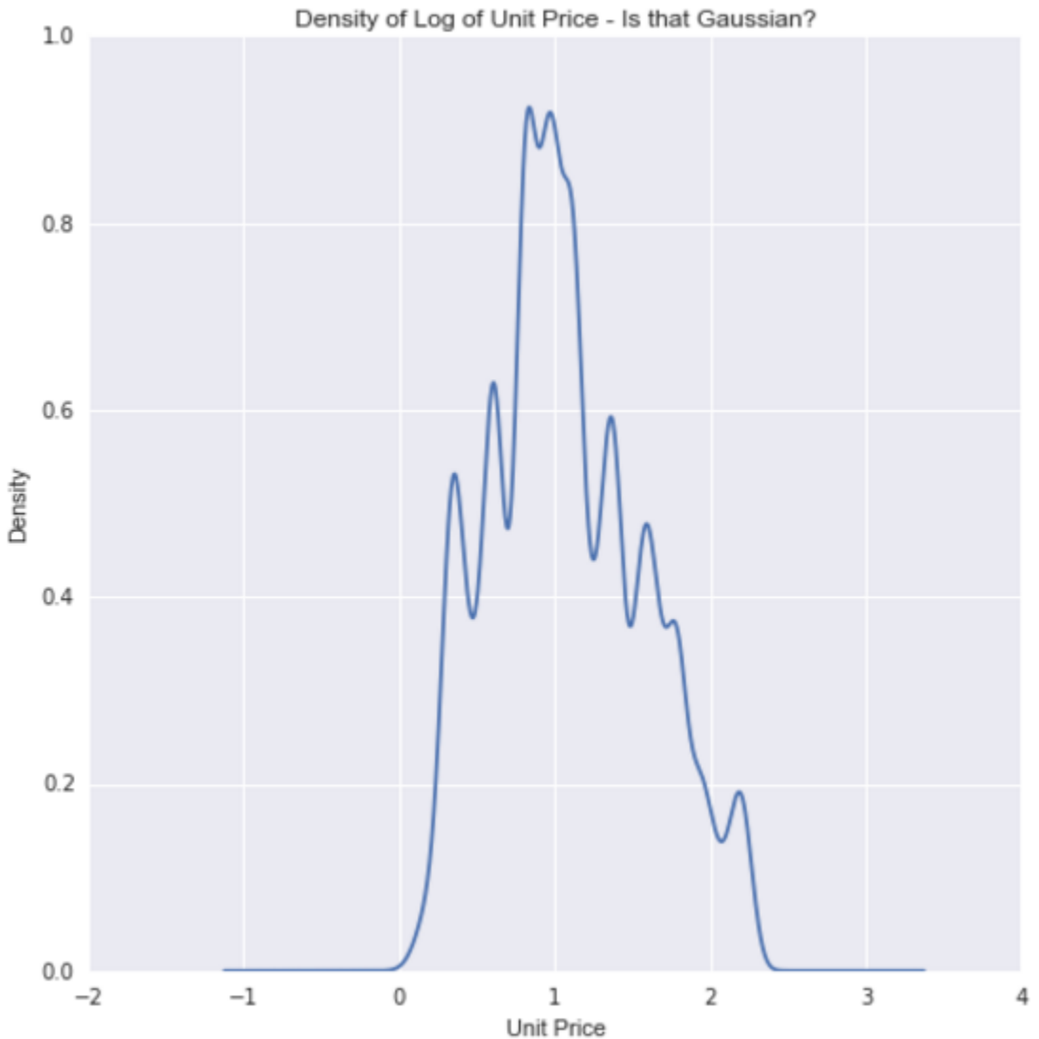
IS YOUR DATA NORMALLY DISTRIBUTED?



CORRECT SKEWNESS

- Normality of data is an assumption in certain linear regression models
- Also impacts Confidence Intervals

CORRECT SKEWNESS



NORMAL DISTRIBUTION

Codealong:
https://github.com/alexperrier/gads/blob/master/03_statistics_fundamentals/py/Normal%

Launch a jupyter notebook

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

BOX COX TRANSFORMATION

$$y = \frac{x^{\lambda} - 1}{\lambda} \qquad if \qquad \lambda ! = 0$$
$$y = log(x) \qquad if \qquad \lambda = 0$$

Ref: Box-Cox Transformations

DUMMY VARIABLES - ONE HOT ENCODING

CATEGORICAL - WHAT'S THE PROBLEM?

- Need numbers not strings

```
country : [US, UK, FR, CA] => [0,1,
```

- Coded as 1,2,3 induces order among categories

CATEGORICAL

CONVERT CATEGORY INTO N VARIABLES

Country: [US, UK, FR, CA]

	Is US?	Is UK ?	Is FR ?	Is CA ?
No	0	0	0	0
Yes	1	1	1	1

IN FACT N-1 NEEDED

Since all categories are known and exclusive

	Is US?	Is UK ?	Is FR ?
No	0	0	0
Yes	1	1	1

000 => means: Is CA? = 1

IN PANDAS

```
pd.get_dummies()
```

LESSON REVIEW

BEFORE NEXT CLASS

HOMEWORK

Study and reproduce the following Notebooks, tutorials and read the articles

- https://github.com/alexperrier/gads/blob/master/03_statistics_fundamentals/py/lesson-3-homework.ipynb
- Dataset exploration: Boston house pricing
- Visualizing the distribution of a dataset
- Data Types and Formats
- How does skewness impacts regression model?

5 QUESTIONS ABOUT TODAY

EXIT TICKET