STATISTICAL THINKING!

PART 1

Bee-Swarm Plot, ECDF
Mean, Median, Outliers
Percentiles & Boxplot
Variance & Standard Deviation
Covariance & Correlation
PMF, CDF, PDF
Binomial & Poisson Distribution

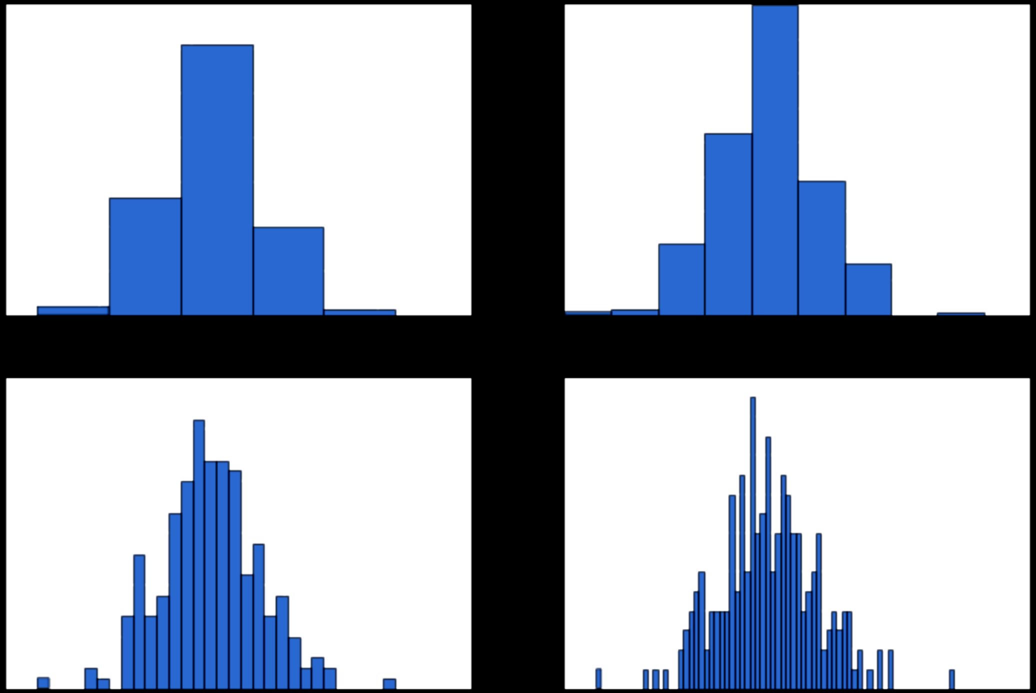
PART 2

Anscombe’s Quartet
Bootstrapping
Confidence Intervals
Hypothesis Testing
Null & Alternate Hypothesis
P-value(ugh..)
A/B Testing

ROHAN KOKKULA

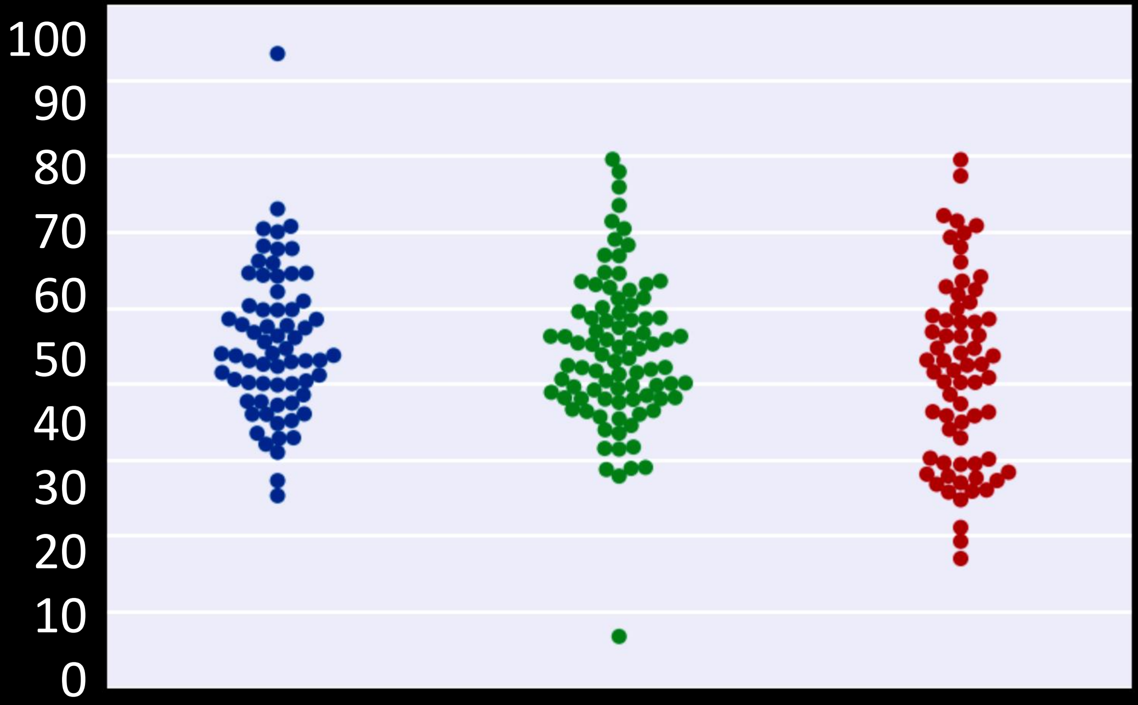
PART 1

BEE SWARM PLOT



Binning Bias (Drawback of Histogram)

i.e different plot shapes for different bins

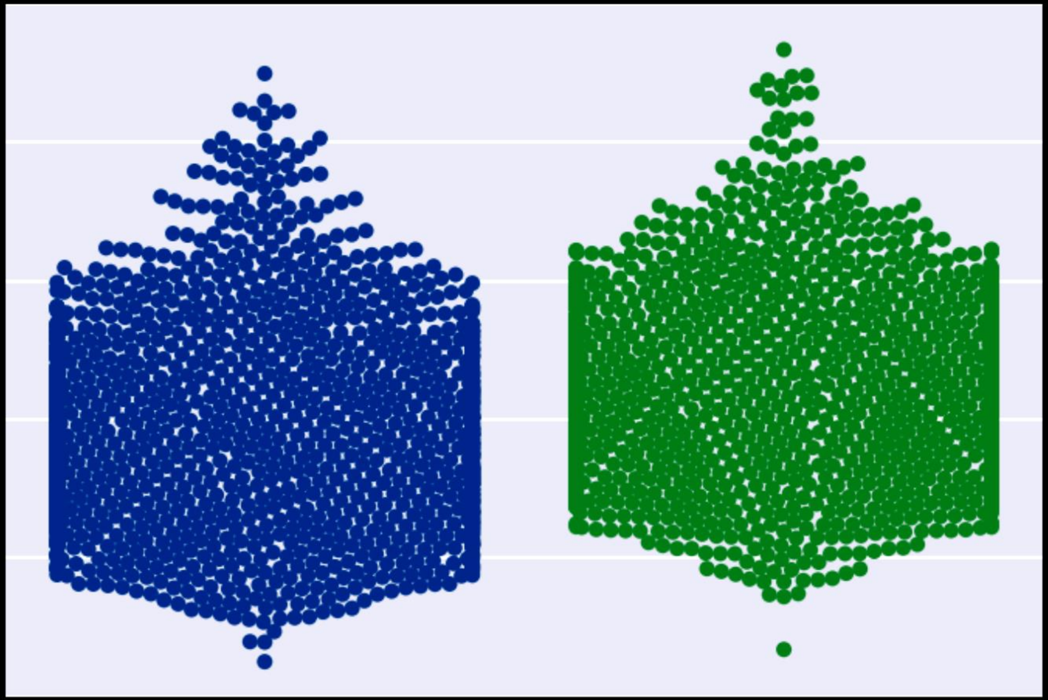


BEE SWARM PLOT

To overcome Binning Bias, each data point is plotted w.r.t its percent contribution

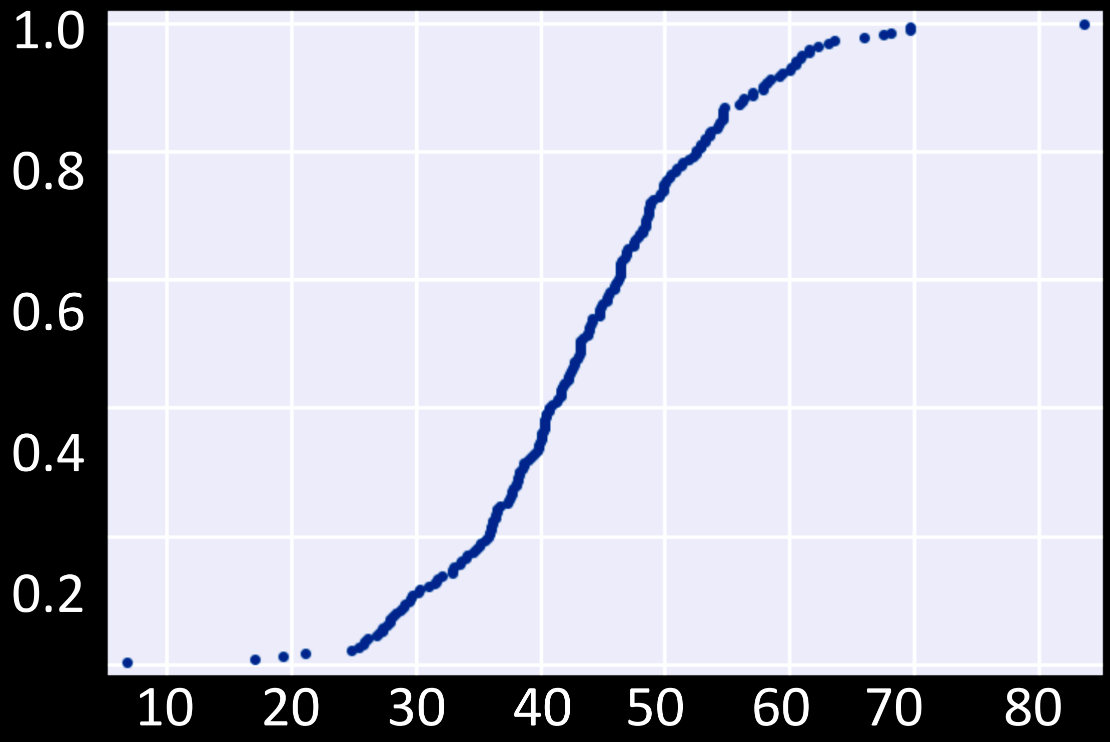
PART 1

ECDF – EMPIRICAL CUMULATIVE DISTRIBUTION FUNCTION



Limiting Extreame Points(Drawback of Swarm)

i.e if there are too many data points at a percent then those are pushed to the limits.



ECDF

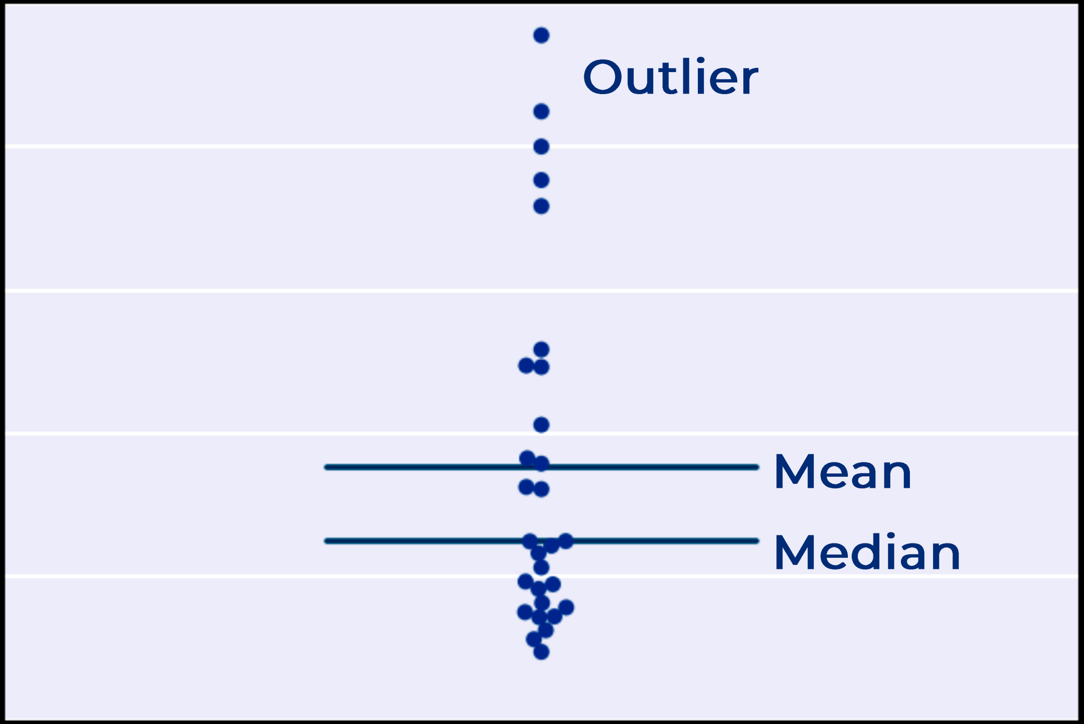
To overcome that, ECDF plots all points having percent wrt percentile(scale 0-1)

PART 1

Mean, Median & Outliers

We Know that.!!! Right?

Well, taking Mean will include outliers and could affect result and taking Median will neglect the importance of outliers.
i.e mean changes and median doesn’t in case of outliers.
But what are outliers? They’re just extreme points from the rest data which could have caused by an error or does have a meaning to it.

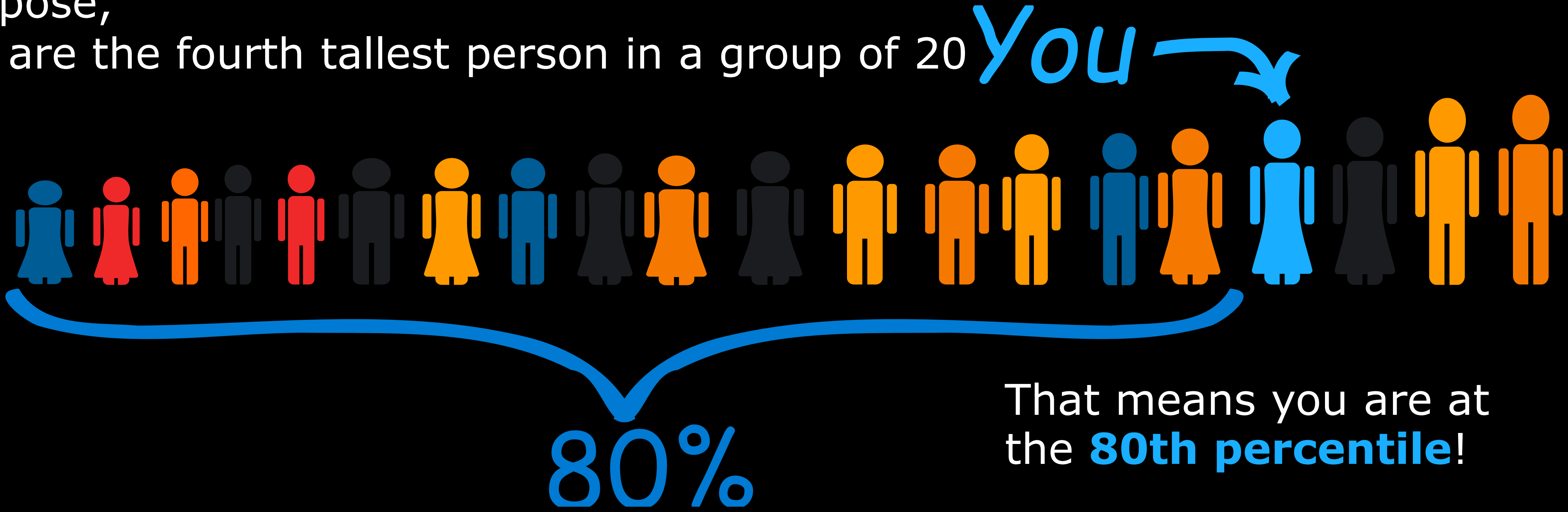


PART 1

Percentiles

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls.

Suppose,
you are the fourth tallest person in a group of 20

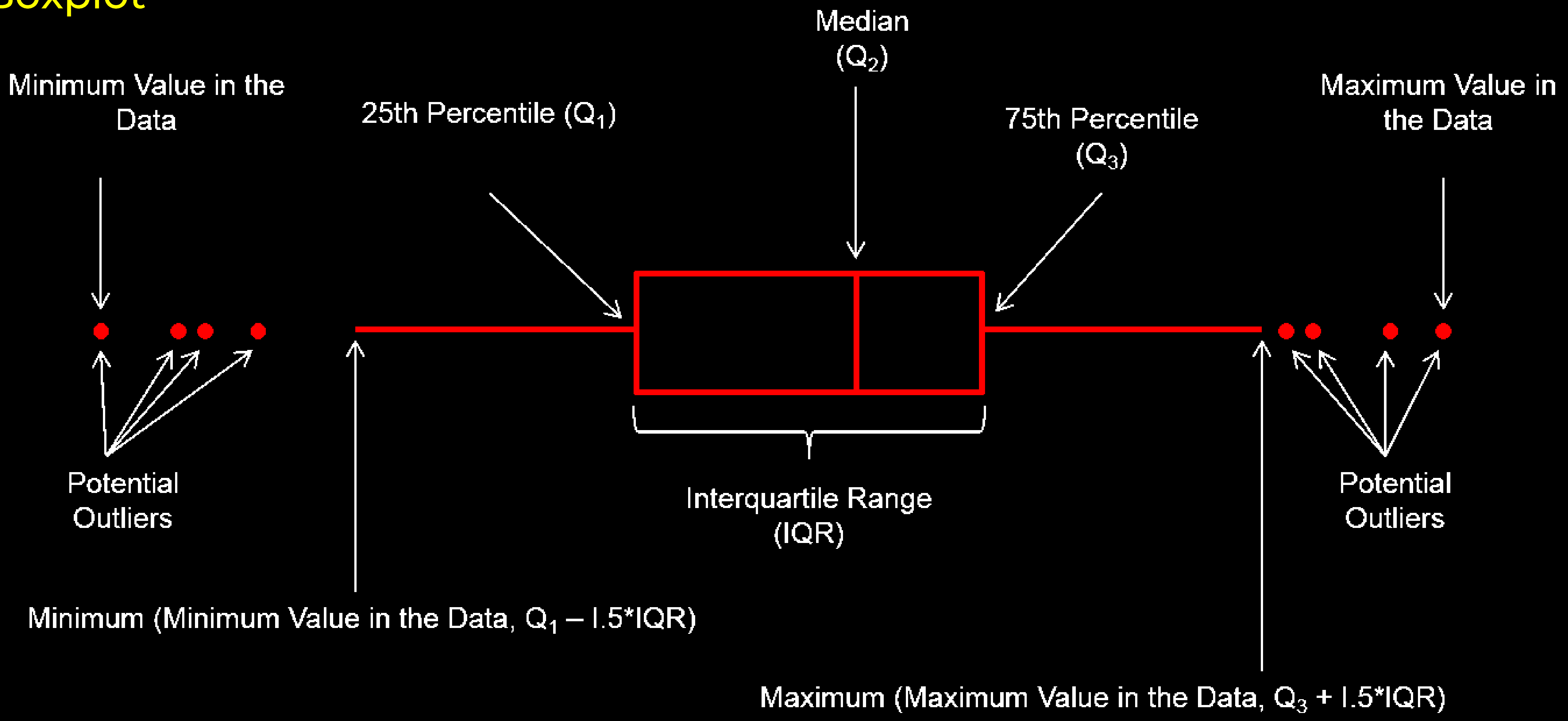


That means you are at the **80th percentile**!

ROHAN KOKKULA

PART 1

Boxplot



Median
(Q_2)
Minimum Value in the Data
25th Percentile (Q_1)
75th Percentile (Q_3)
Maximum Value in the Data
Potential Outliers
Interquartile Range (IQR)
Potential Outliers
Minimum (Minimum Value in the Data, Q_1 – 1.5*IQR)
Maximum (Maximum Value in the Data, Q_3 + 1.5*IQR)

ROHAN KOKKULA

PART 1

Variance & Standard Deviation

The Variance is a way to measure how far a set of numbers is spread out. It describes how much a random variable differs from its mean. It is defined as the average of the squares of the differences between the individual point and the mean.

$$\text{Var} = \frac{\Sigma(\mathrm{X}-\bar{x})^2}{n} \qquad \text{S.D} = \sqrt{\frac{\Sigma(\mathrm{X}-\bar{x})^2}{n}}$$

X = observed value
$\bar{x}$ = mean
n = no. of observations

The Standard Deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

ROHAN KOKKULA

PART 1

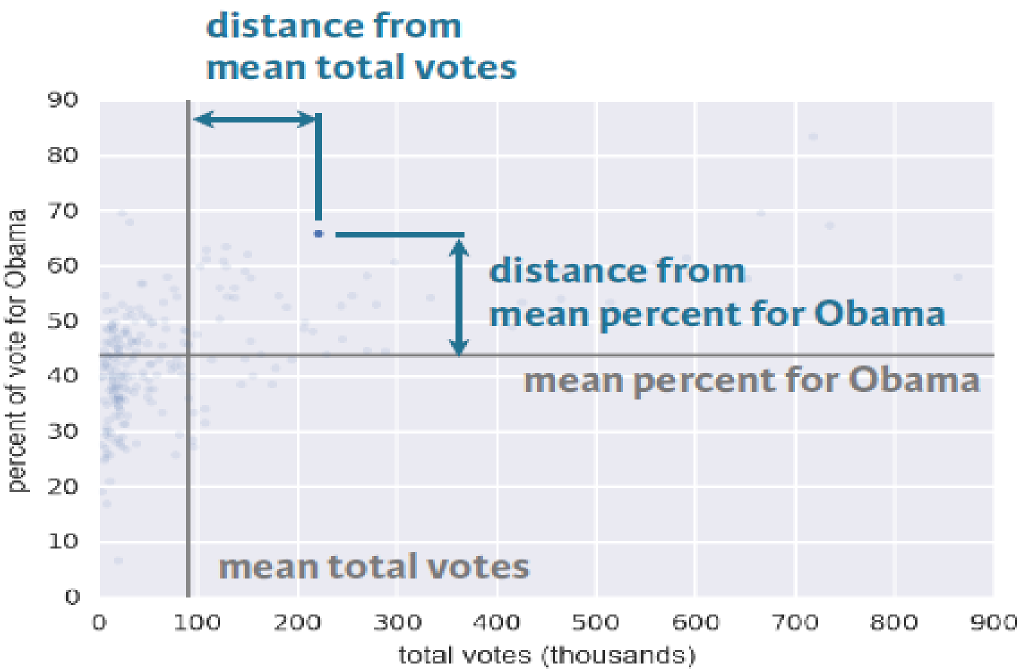
Covariance & Correlation

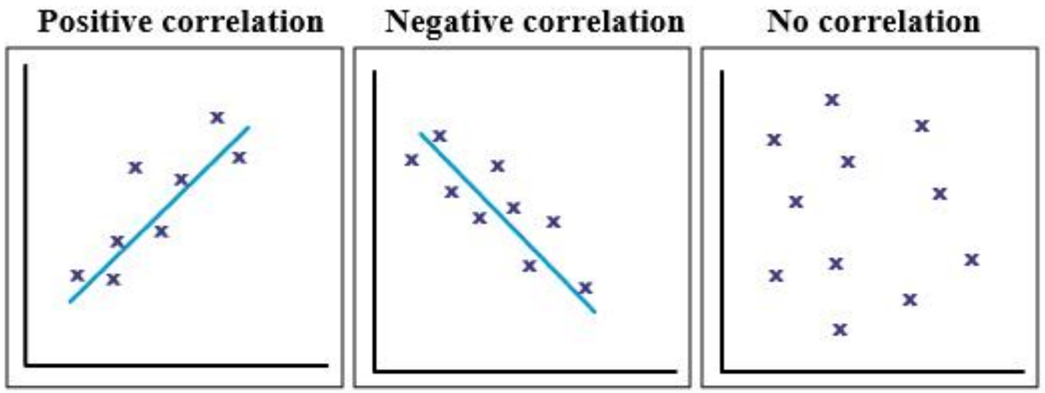
Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together.

"Covariance" indicates the direction of the linear relationship between variables. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.

Correlation Coefficient is obtained by dividing the covariance of the two variables by the product of their standard deviations.

While correlation coefficients lie between -1 and +1, covariance can take any value between -∞ and +∞.





The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

The points lie close to a straight line, which has a negative gradient.

This shows that as one variable **increases**, the other **decreases**.

There is no pattern to the points.

This shows that there is **no connection** between the two variables.

PART 1

PMF, CDF, PDF

PMF – Probability Mass Function

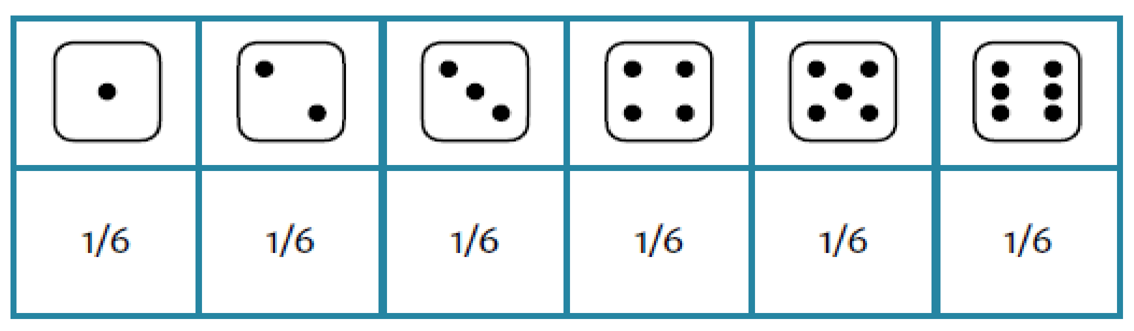
A function that gives the probability that a discrete random variable is exactly equal to some value.
Eg. Rolling a dice.

CDF – Cumulative Distribution Function

A Function that gives the probability that a random variable X with a given probability distribution will be found at a value less than or equal to x.
Or, it’s just the cumulative summation of PMF or PDF.
The last value is always equal to 1 as summation of probabilistic possibilities is 1

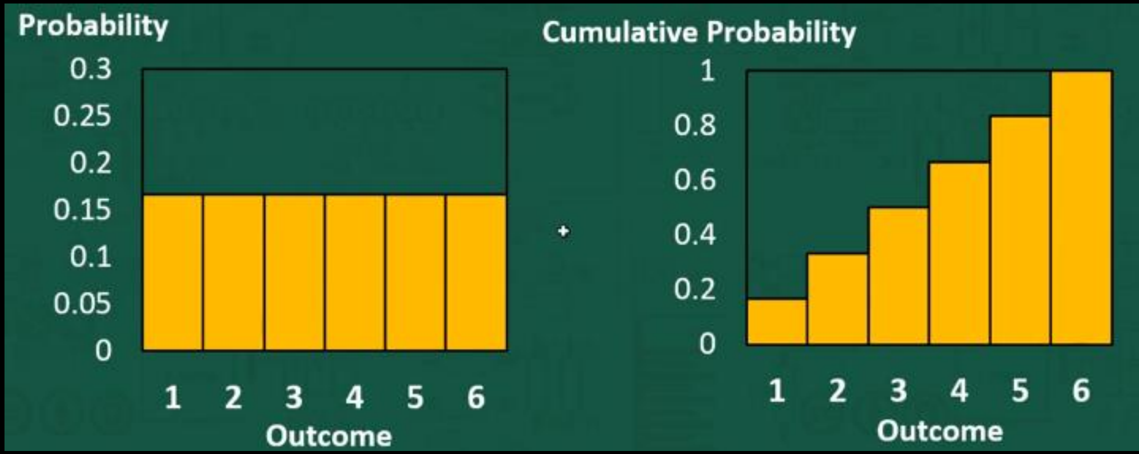
PDF – Probability Density Function

A function that gives the probability that a continuous random variable is exactly equal to some value.
Eg. Height

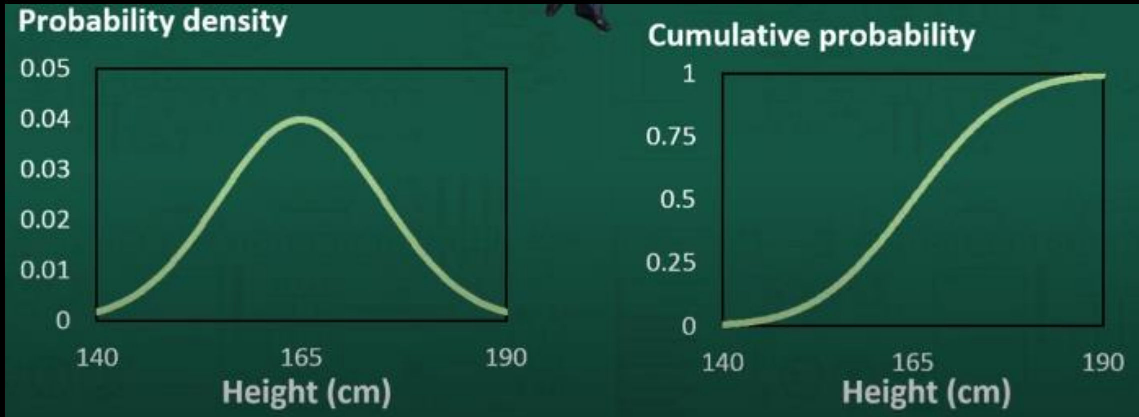


⚀	⚁	⚂	⚃	⚄	⚅
1/6	1/6	1/6	1/6	1/6	1/6

Discrete PMF & CDF



Continuous PDF & CDF



PART 1

Binomial Distribution

It describes discrete data. i.e situations where there can be only two results in a random experiment.

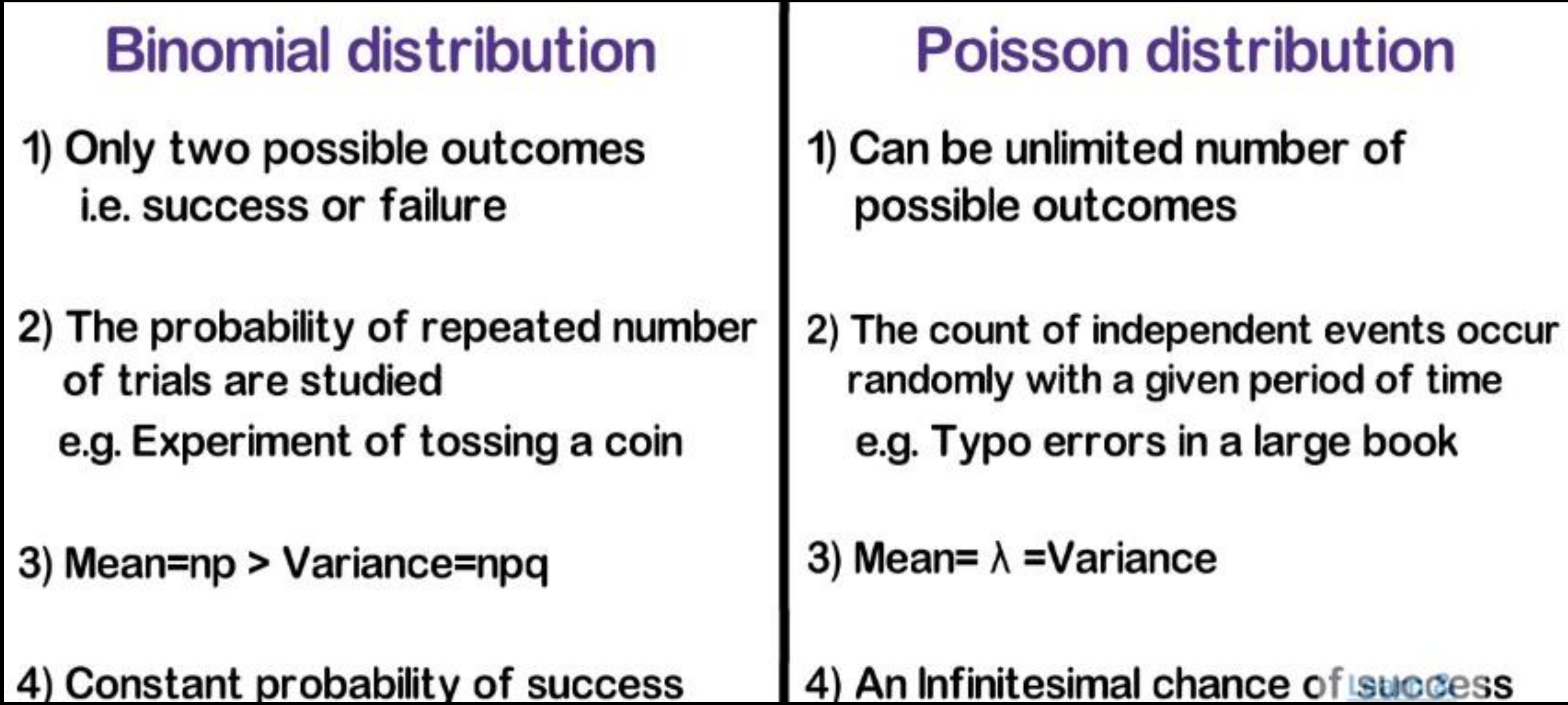
Eg. Pass or Failure, Yes or No, Rainy or Summer

The binomial distribution tells us the probability distribution of getting r successful outcomes out of n trials. Here, the probability of each successful outcome is p and the probability of a loss is 1-p.

Poisson Distribution

It describes discrete data. i.e situations where random variable can take integer values.
Eg. Number of patients arriving at a physician’s office, Number of cars arriving at a toll booth

The number r of arrivals of a Poisson process in a given time interval with average rate of λ arrivals per interval is Poisson distributed.



Binomial distribution	Poisson distribution
1) Only two possible outcomes i.e. success or failure	1) Can be unlimited number of possible outcomes
2) The probability of repeated number of trials are studied e.g. Experiment of tossing a coin	2) The count of independent events occur randomly with a given period of time e.g. Typo errors in a large book
3) Mean=np > Variance=npq	3) Mean= λ =Variance
4) Constant probability of success	4) An Infinitesimal chance of success

PART 2

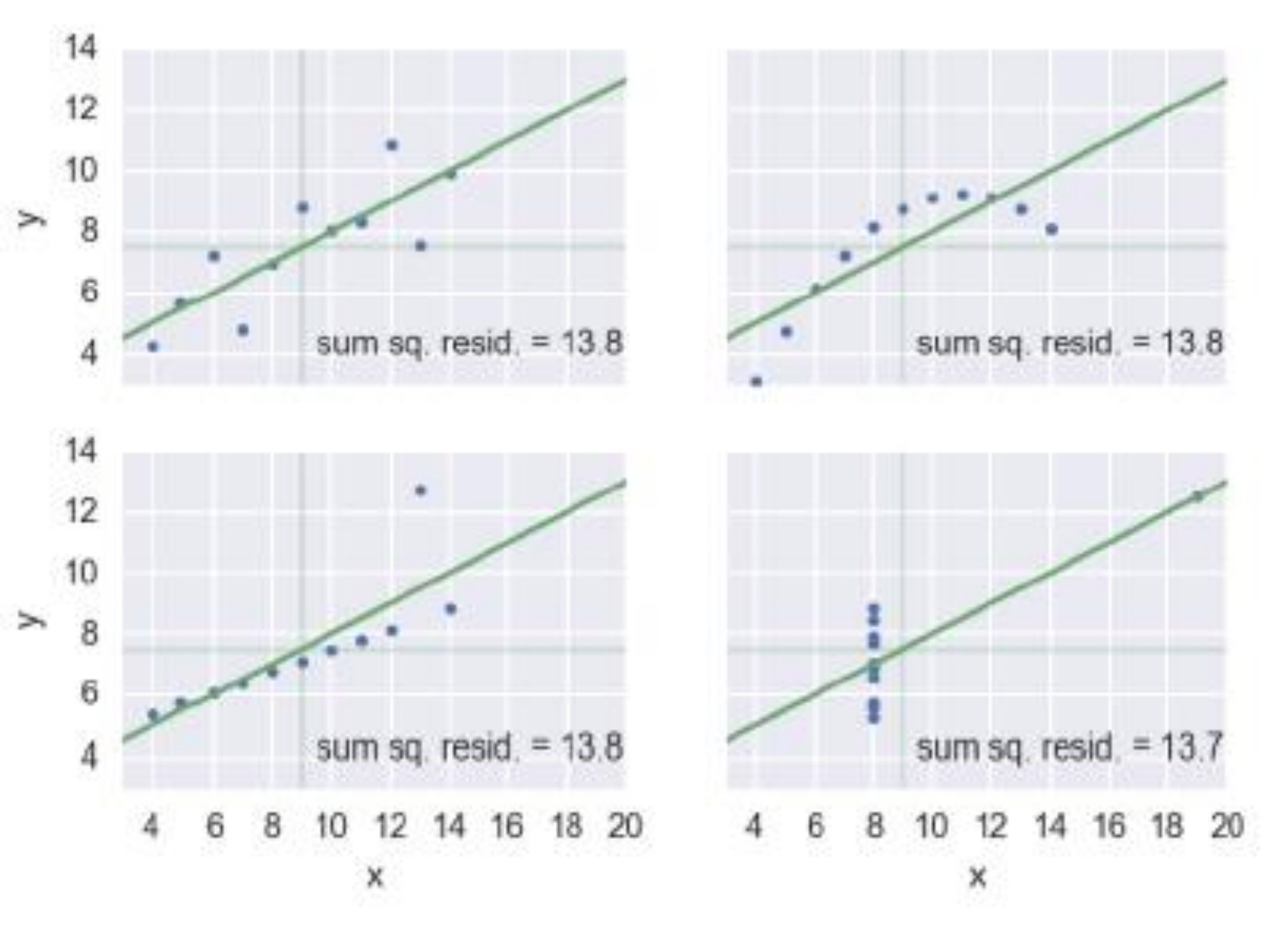
Anscombe’s Quartet

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

The following parameters are same for all of the 4 different graphs.

Mean of x-axes and y-axes
Slope, Intercept
Sum of squared residuals

Applying Linear regression will not be ideal for all the cases.



PART 2

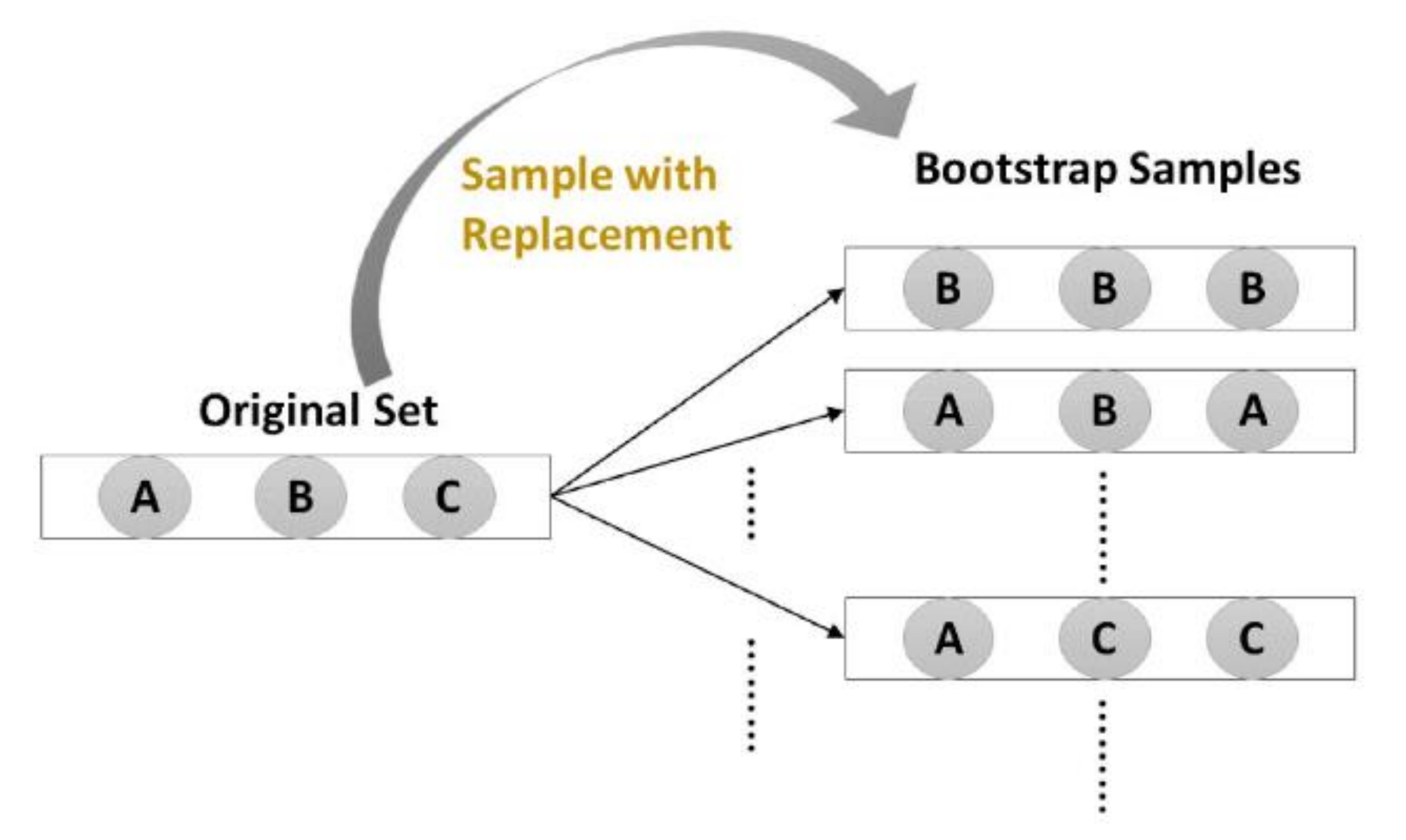
Bootstrapping

Bootstrapping is any test or metric that relies on random sampling with replacement.
It allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates.

It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution, such as percentile points, proportions, odds ratio, and correlation coefficients.

Bootstrap is also an appropriate way to control and check the stability of the results.

This can be implemented by constructing a number of resamples with replacement, of the observed data set (and of equal size to the observed data set).



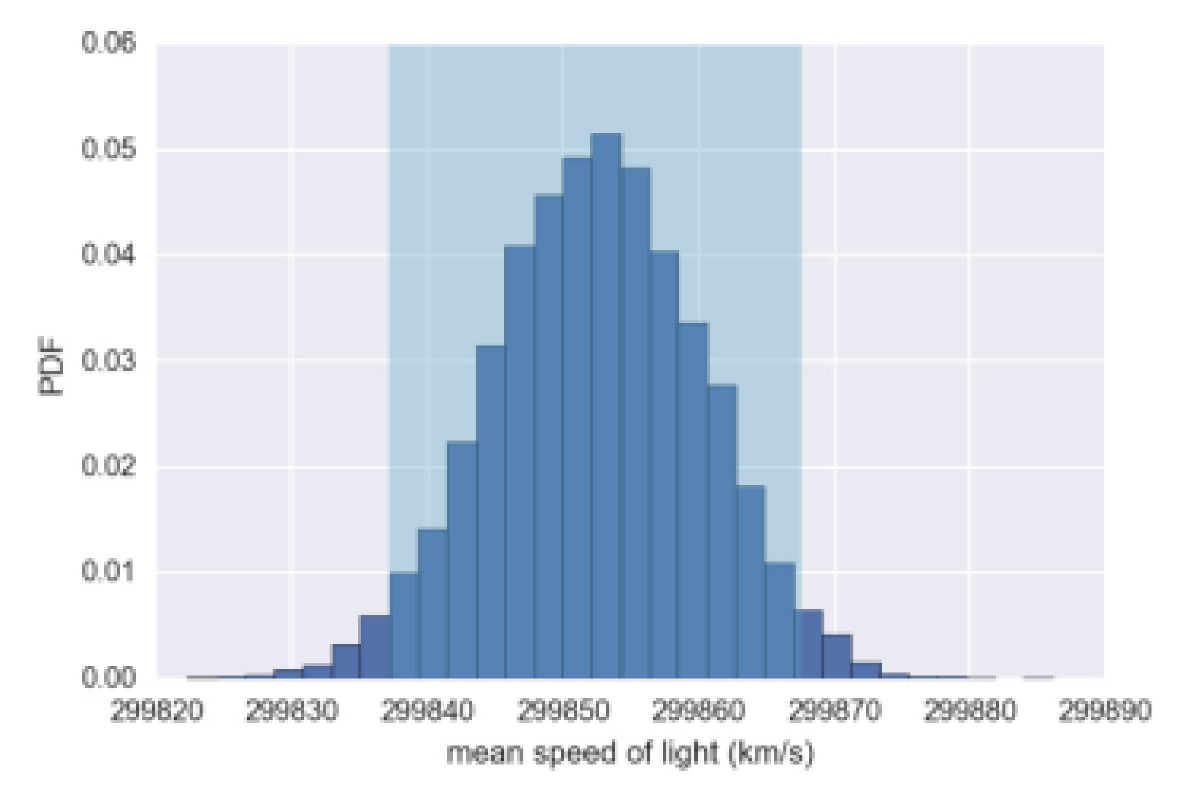
PART 2

Confidence Interval

A Confidence Interval (CI) is a type of estimate computed from the statistics of the observed data. This proposes a range of plausible values for an unknown parameter (for example, the mean). The interval has an associated confidence level that the true parameter is in the proposed range.

Factors affecting the width of the confidence interval include the size of the sample, the confidence level, and the variability in the sample.
A larger sample will tend to produce a better estimate of the population parameter, when all other factors are equal.

A higher confidence level will tend to produce a broader confidence interval.



PART 2

Hypothesis Testing

What is Hypothesis?

A hypothesis is an educated guess about something in the world around you.
It should be testable, either by experiment or observation.
For example:
You might run an experiment and find that a certain drug is effective at treating headaches.
But if you can't repeat that experiment, no one will take your results seriously.

What is Hypothesis Statement?

"If I...(do this to an independent variable)....then (this will happen to the dependent variable)."
For example:
If I (decrease the amount of water given to herbs) then (the herbs will increase in size).
If I (give exams at noon instead of 7) then (student test scores will improve).
If I (look in this certain location) then (I am more likely to find new species).

What is Hypothesis Testing?

Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.
You're basically testing whether your results are valid by figuring out the odds that your results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

PART 2

Hypothesis Testing

What is Null Hypothesis?

If you trace back the history of science, the null hypothesis is always the accepted fact.
With the null hypothesis, you get what you expect, from a historical point of view.
Simple examples of null hypotheses that are generally accepted as being true are:
DNA is shaped like a double helix.
There are 8 planets in the solar system (excluding Pluto).
Taking Vioxx can increase your risk of heart problems (a drug now taken off the market).

Rejecting the Null Hypothesis

Ten or so years ago, we believed that there were 9 planets in the solar system.
Pluto was demoted as a planet in 2006. The null hypothesis of “Pluto is a planet” was replaced by Alternate Hypothesis : “Pluto is not a planet.”

What is Alternate Hypothesis?

The alternate hypothesis is just an alternative to the null. Basically, you’re looking at whether there’s enough change (with the alternate hypothesis) to be able to reject the null hypothesis.
Significance level (α), a probability threshold below which the null hypothesis will be rejected.
Common values are 5% and 1%.
Reject the null hypothesis, in favor of the alternative hypothesis, if and only if the p-value is less than (or equal to) the significance level (α) threshold).

PART 2

P - value

A p value is used in hypothesis testing to help you support or reject the null hypothesis.
The p value is the evidence against a null hypothesis. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

P-value is NOT the probability that the null hypothesis is true
It does not provide the probability that either hypothesis is correct

For example, there is a 2.54% chance your results could be random (i.e. happened by chance). That’s pretty tiny. On the other hand, a large p-value of .9(90%) means your results have a 90% probability of being completely random and not due to anything in your experiment.

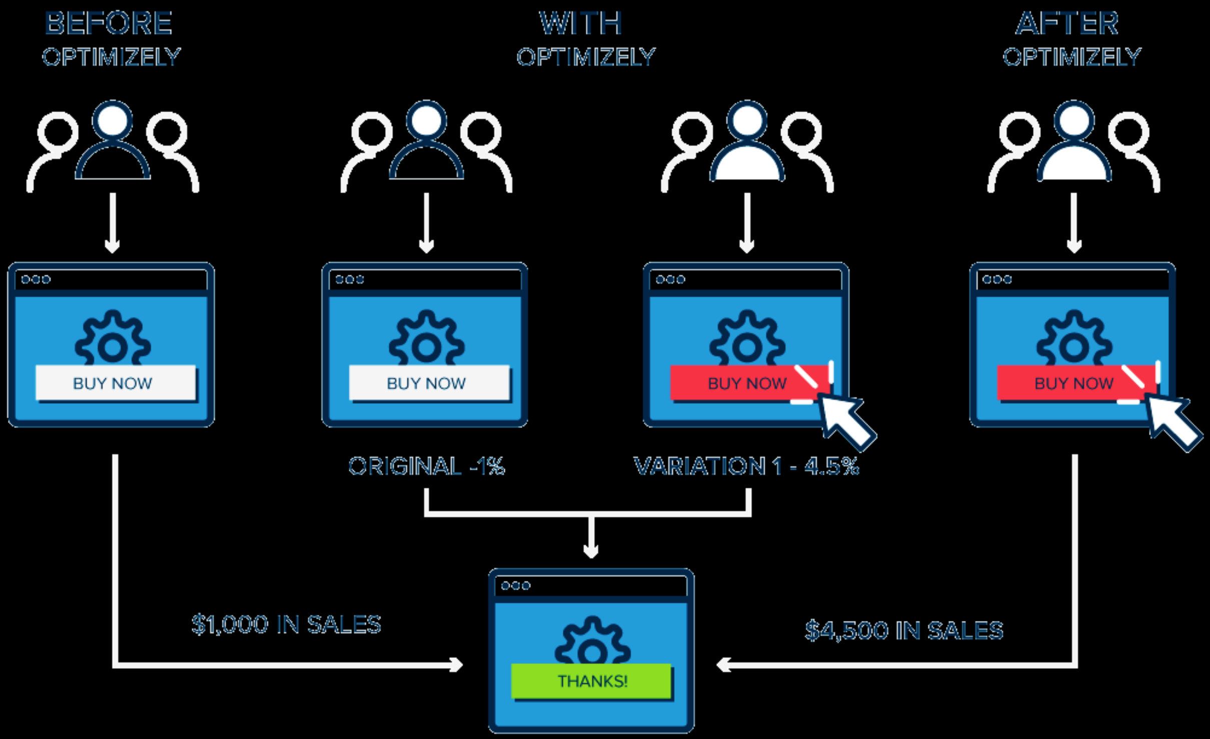
A small p (≤ 0.05), reject the null hypothesis. This is strong evidence that the null hypothesis is invalid.
A large p (> 0.05) means the alternate hypothesis is weak, so you do not reject the null.

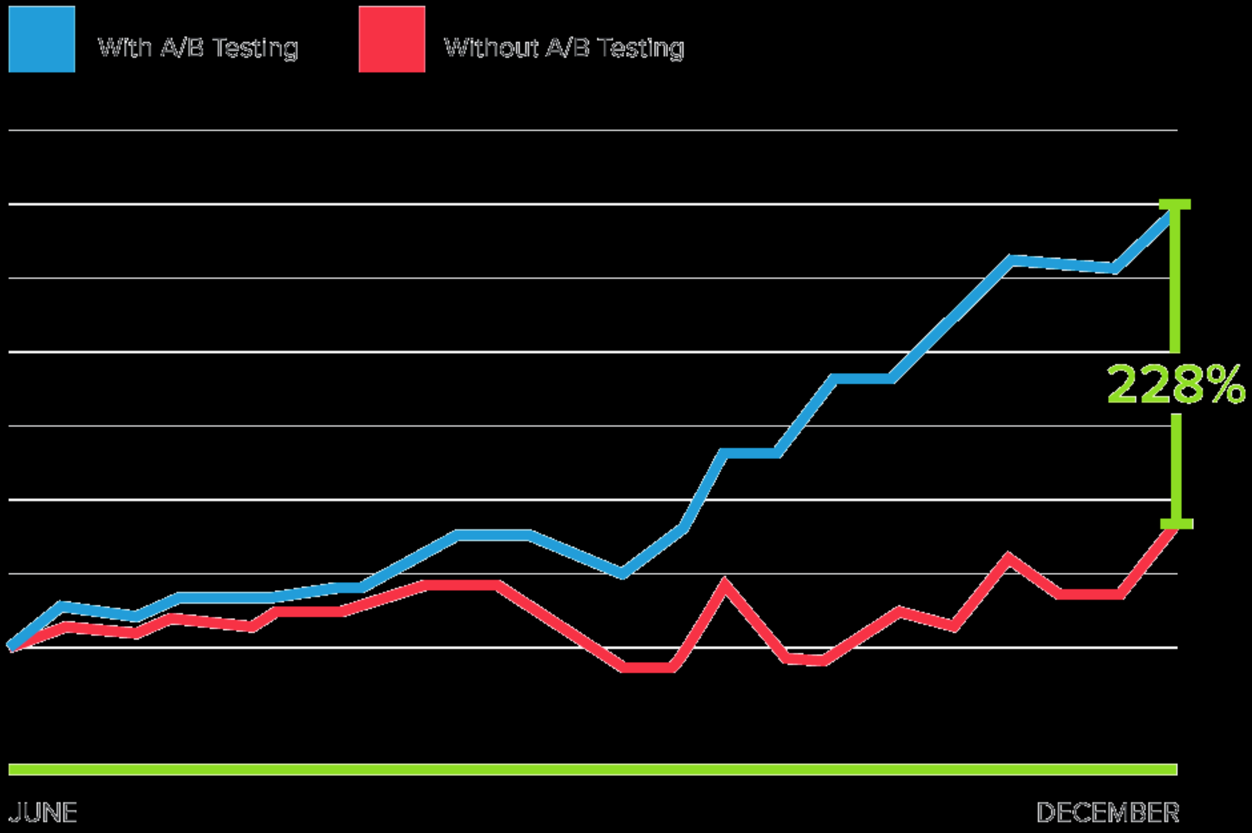
For example, say that a fair coin is tested for fairness (the null hypothesis). At a significance level of 0.05, the fair coin would be expected to (incorrectly) reject the null hypothesis in about 1 out of every 20 tests.

PART 2

A/B Testing

A/B testing is a user experience research methodology.
A/B tests consist of a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics.
A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.





ROHAN KOKKULA