Statistics in AI

Session 3

Importance of Statistics in AI / ML

Why Statistics?

Statistics is a field of mathematics that is universally agreed to be a prerequisite for a deeper understanding of machine learning.

Although statistics is a large field with many esoteric theories and findings, the nuts and bolts tools and notations taken from the field are required for machine learning practitioners.

Random Numbers

Random Number Generator is a device that can generate one or many random numbers within a defined scope.

Random number generators can be hardware based or pseudo-random number generators.

Hardware based random-number generators can involve the use of a dice, a coin for flipping, or many other devices.

A simple way to generate random numbers:

Take a four digit **number** - multiply it by itself - producing a 7 or 8 digit **number**.

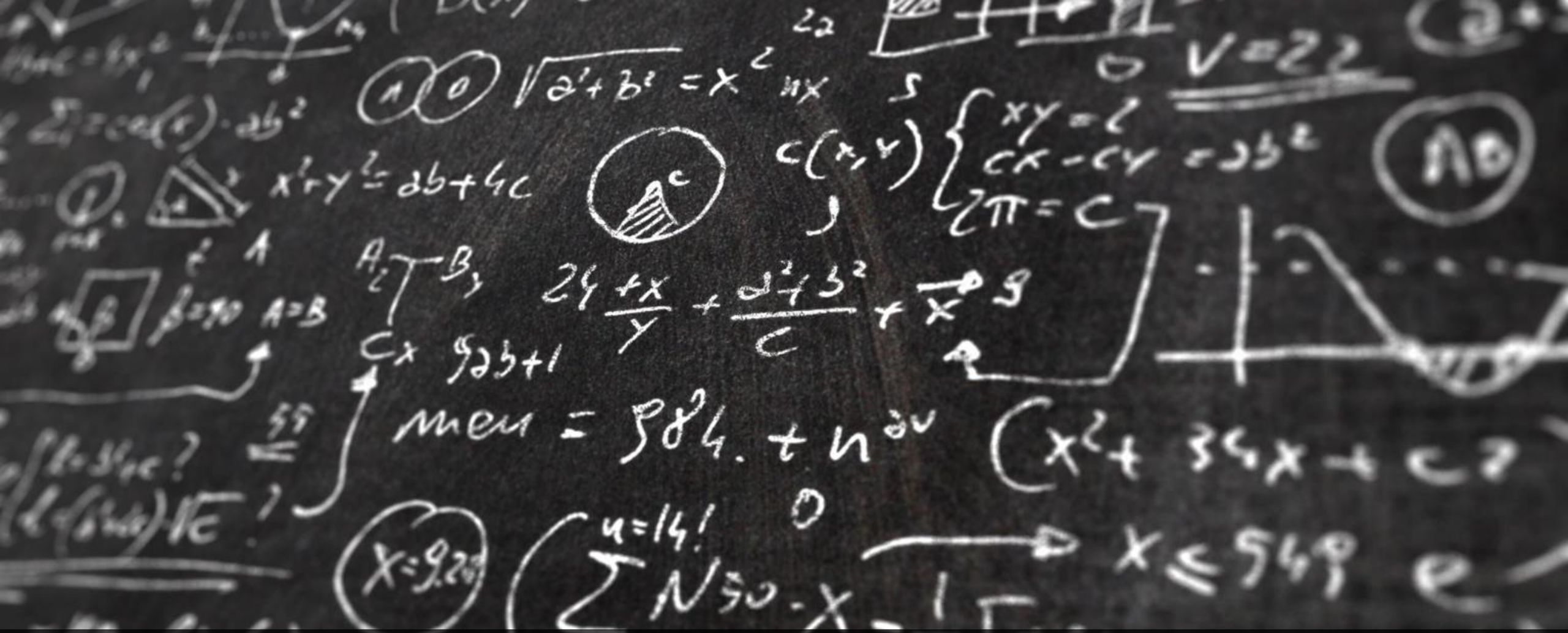
Take the middle four digits of that - and use that to **generate** the next **random number**.

There are some **numbers** (such as 0000) which make terrible starting points.

Probability

Joint and Conditional Probability

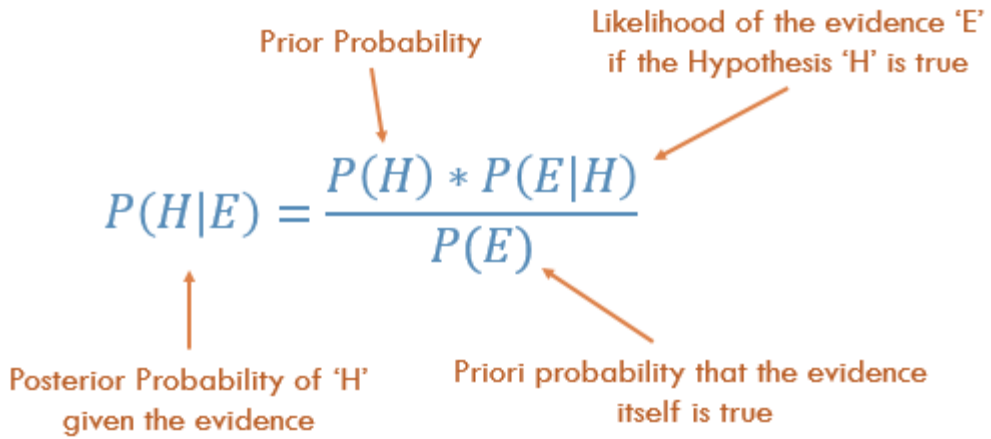
- **Joint Probability**
  - Probability of events A and B denoted by**P(A and B) or P(A ∩ B)**is the probability that events A and B both occur. **P(A ∩ B) = P(A). P(B) .** This only applies if **A** and **B**are independent, which means that if **A** occurred, that doesn't change the probability of **B**, and vice versa.
- **Conditional Probability**
  - Let us consider A and B are not independent, because if A occurred, the probability of B is higher. When A and B are not independent, it is often useful to compute the conditional probability, P (A|B), which is the probability of A given that B occurred: **P(A|B) = P(A ∩ B)/ P(B).**



Bayes' Theorem

Bayes's theorem is a relationship between the conditional probabilities of two events.

For example, if we want to find the probability of selling ice cream on a hot and sunny day, Bayes' theorem gives us the tools to use prior knowledge about the likelihood of selling ice cream on any other type of day (rainy, windy, snowy etc.).



Covid 19 and Bayes Theorem

- Suppose that during a medical examination, doctor informs patient that he has tested positive for a rare disease. You are also aware that there is some uncertainty in the results of these tests. Assuming we have a **Sensitivity** (also called the **true positive rate)** result for 95% of the patients with the disease, and a **Specificity** (also called the **true negative rate**) result for 95% of the healthy patients.

- If we let "+" and "–" denote a positive and negative test result, respectively, then the test accuracies are the conditional probabilities : $P(+|disease) = 0.95$, $P(-|healthy) = 0.95$,

		Real Situation	
		Sick	Healthy
Test	Sick	True result	False result
	Healthy	False result	True result

How to evaluate *P(+)*, all positive cases ?

We have to consider two possibilities, P(+|disease) and P(+|healthy). The probability of a false positive, P(+|healthy), is the complement of the P(-|healthy). Thus P(+|healthy) = 0.05.

$$P(\text{disease}|+) = \frac{P(+|\text{disease})P(\text{disease})}{P(+|\text{disease})P(\text{disease}) + P(+|\text{healthy})P(\text{healthy})}$$

Importantly, Bayes' theorem reveals that in order to compute the conditional probability that you have the disease given the test was positive, you need to know the "prior" probability you have the disease$P_{(\text{disease})}$, given no information at all. That is, you need to know the overall incidence of the disease in the population to which you belong. Assuming these tests are applied to a population where the actual disease is found to be 0.5%, $P_{(\text{disease})} = 0.005$ which means $P_{(\text{healthy})} = 0.995$.

So, $P(\text{disease}|+) = 0.95 * 0.005 / (0.95 * 0.005 + 0.05 * 0.995) = 0.088$

In other words, despite the apparent reliability of the test, the probability that the patient actually has the disease is still less than 9%. Getting a positive result increases the probability you have the disease. But it is incorrect to interpret the 95 % test accuracy as the probability you have the disease.

Descriptive Statistics

Descriptive statistics refers to methods for summarizing and organizing the information in a data set.

We will use below table to describe some of the statistical concepts

Characteristics of 10 loan applicants

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
1	Single	y	38,000	2	2009	Good
2	Married	y	32,000	7	2010	Good
3	Other	n	25,000	9	2011	Good
4	Other	n	36,000	3	2009	Good
5	Other	y	33,000	4	2010	Good
6	Other	n	24,000	10	2008	Bad
7	Married	y	25,100	8	2010	Good
8	Married	y	48,000	1	2007	Good
9	Married	y	32,100	6	2009	Bad
10	Married	y	32,200	5	2010	Good

Elements: The entities for which information is collected are called the elements. In the above table, the elements are the 10 applicants. Elements are also called cases or subjects.

Variables: The characteristic of an element is called a variable. It can take different values for different elements.e.g., marital status, mortgage, income, rank, year, and risk. Variables are also called attributes.

Variables can be either **qualitative** or **quantitative**.

- **Qualitative:** A qualitative variable enables the elements to be classified or categorized according to some characteristic. The qualitative variables are marital status, mortgage, rank, and risk. Qualitative variables are also called **categorical** variables.

- **Quantitative:** A quantitative variable takes numeric values and allows arithmetic to be meaningfully performed on it. The quantitative variables are income and year. Quantitative variables are also called **numerical** variables.

- **Discrete Variable**: A numerical variable that can take either a finite or a countable number of values is a discrete variable, for which each value can be graphed as a separate point, with space between each point. ‘year’ is an example of a discrete variable..

- **Continuous Variable**: A numerical variable that can take infinitely many values is a continuous variable, whose possible values form an interval on the number line, with no space between the points. ‘income’ is an example of a continuous variable.

Characteristics of 10 loan applicants

Applicant	Marital Status	Mortgage	Income ($)	Income Rank	Year	Risk
1	Single	y	38,000	2	2009	Good
2	Married	y	32,000	7	2010	Good
3	Other	n	25,000	9	2011	Good
4	Other	n	36,000	3	2009	Good
5	Other	y	33,000	4	2010	Good
6	Other	n	24,000	10	2008	Bad
7	Married	y	25,100	8	2010	Good
8	Married	y	48,000	1	2007	Good
9	Married	y	32,100	6	2009	Bad
10	Married	y	32,200	5	2010	Good

Measures of Center

Mean

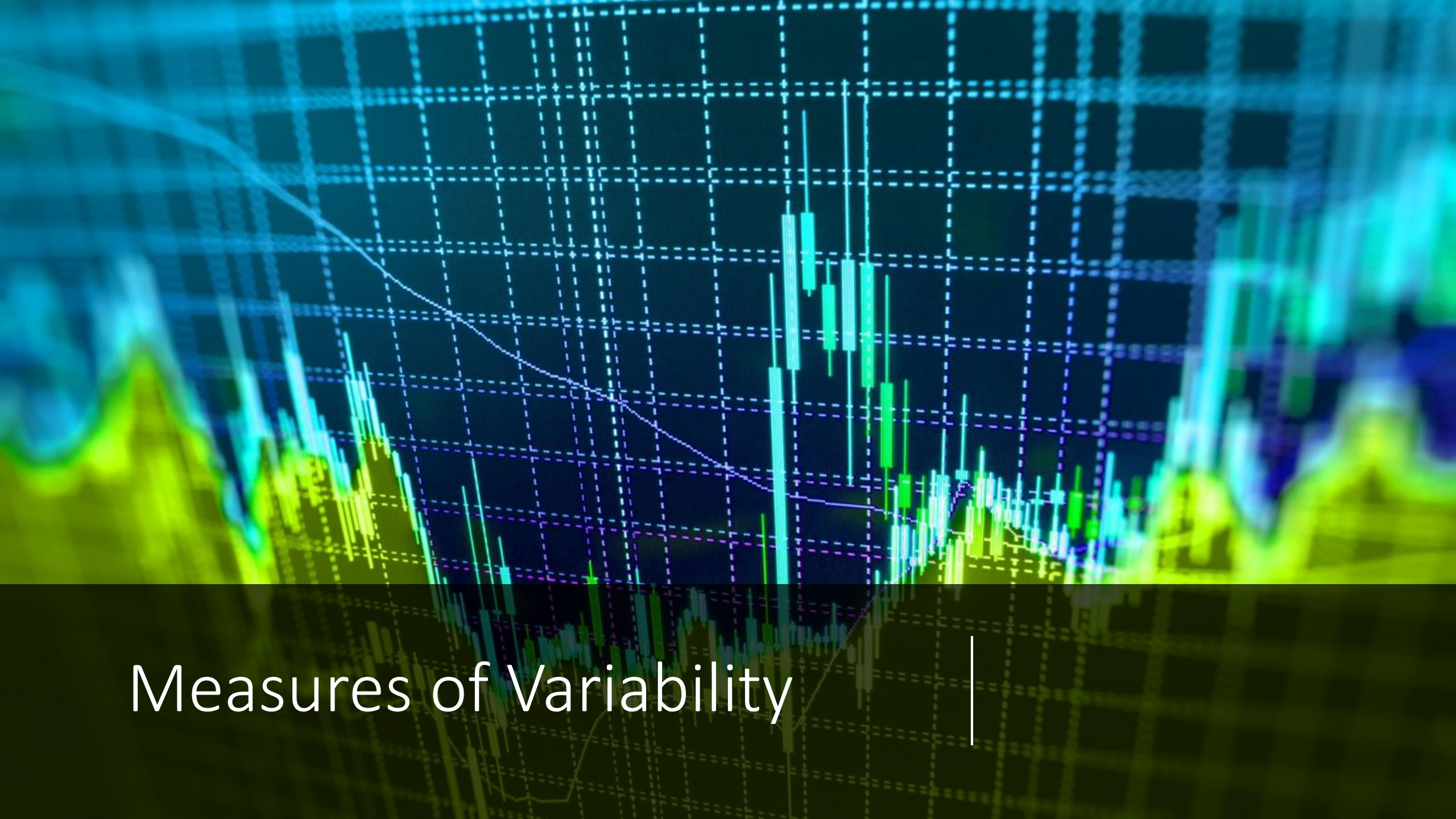
The mean is the arithmetic average of a data set. To calculate the mean, add up the values and divide by the number of values.The sample mean is the arithmetic average of a sample, and is denoted $\bar{x}$ (“x-bar”). The population mean is the arithmetic average of a population, and is denoted μ (“myu”, the Greek letter for m).

Median

The median is the middle data value, when there is an odd number of data values and the data have been sorted into ascending order. If there is an even number, the median is the mean of the two middle data values. When the income data are sorted into ascending order, the two middle values are $32,100 and $32,200, the mean of which is the median income, $32,150.

Mode

The mode is the data value that occurs with the greatest frequency. Both quantitative and categorical variables can have modes, but only quantitative variables can have means or medians. Each income value occurs only once, so there is no mode. The mode for year is 2010, with a frequency of 4.



Measures of Variability

Range

The range of a variable equals the difference between the maximum and minimum values. The range of income is:

```
range(income) = max (income) – min (income) = 48,000 – 24,000 =$24000
```

Range only reflects the difference between largest and smallest observation, but it fails to reflect how data is centralized.

Variance

Population variance is defined as the average of the squared differences from the Mean, denoted as σ^2 (“sigma-squared”):

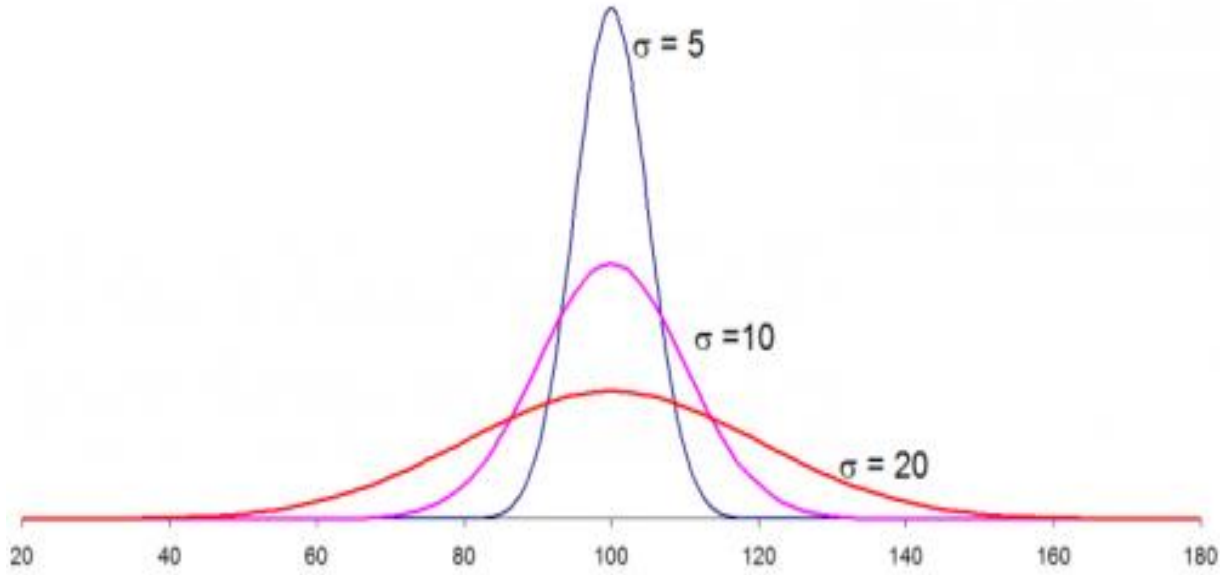
$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Standard Deviation

The *standard deviation* or `sd` of a bunch of numbers tells you how much the individual numbers tend to differ from the mean.

The sample standard deviation is the square root of the sample variance: $sd = \sqrt{s^2}$. For example, incomes deviate from their mean by $7201.

The population standard deviation is the square root of the population variance: $sd = \sqrt{\sigma^2}$.



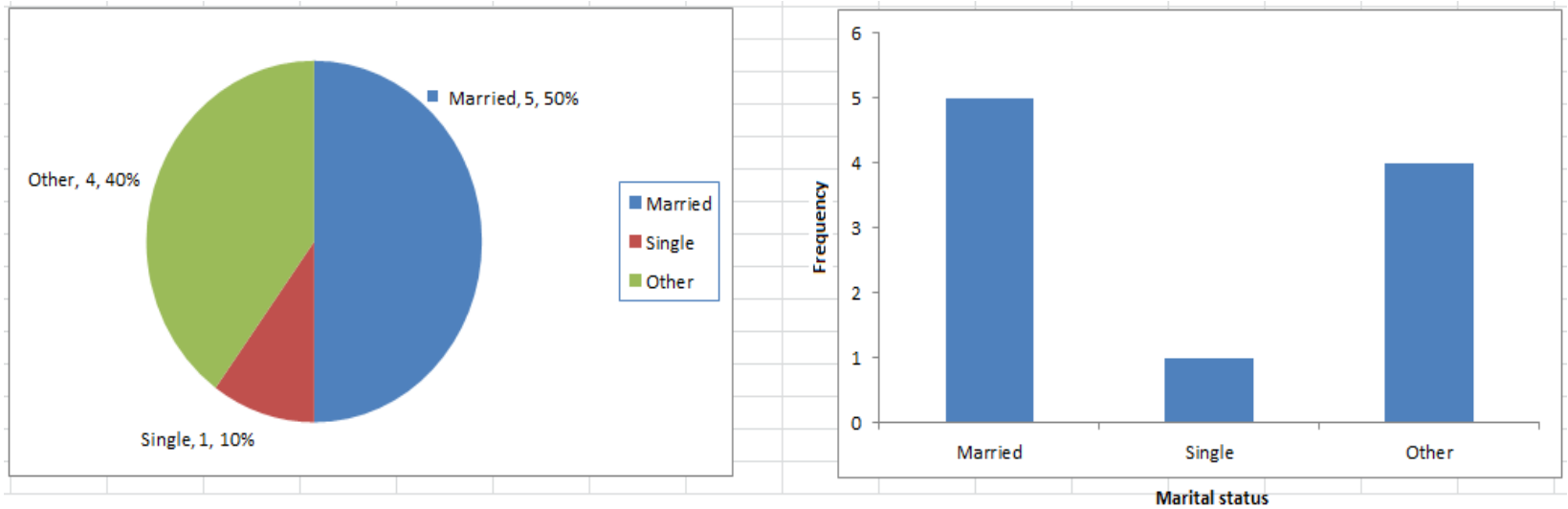
Three different data distributions with same mean (100) and different standard deviation (5,10,20)

The smaller the standard deviation, narrower the peak, the data points are closer to the mean. The further the data points are from the mean, the greater the standard deviation.

Uni-variate and Bivariate Descriptive Statistics

Uni-variate Descriptive Statistics

Different ways you can describe patterns found in uni-variate data include central tendency : mean, mode and median and dispersion: range, variance, maximum, minimum, quartiles , and standard deviation.



Pie chart [left] & Bar chart [right] of Marital status from loan applicants table.

The various plots used to visualize uni-variate data typically are Bar Charts, Histograms, Pie Charts. etc.

Bi-variate Descriptive Statistics

Bi-variate analysis involves the analysis of two variables for the purpose of determining the empirical relationship between them. The various plots used to visualize bi-variate data typically are scatter-plot, box-plot

Correlation

Correlation

A correlation is a statistic intended to quantify the strength of the relationship between two variables. The **correlation coefficient** `r` quantifies the strength and direction of the linear relationship between two quantitative variables. The correlation coefficient is defined as:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y}$$

where `sx` and `sy` represent the standard deviation of the x-variable and the y-variable, respectively. `-1 ≤ r ≤ 1`.

If r is positive and significant, we say that x and y are **positively correlated**. An increase in x is associated with an increase in y.

If r is negative and significant, we say that x and y are **negatively correlated**. An increase in x is associated with a decrease in y

Covariance in stocks

Calculating a stock's covariance starts with finding a list of previous prices or "historical prices" as they are called on most quote pages. Typically, you use the closing price for each day to find the return.

To begin the calculations, find the closing price for both stocks and build a list. For example:

Daily Return for Two Stocks Using the Closing Prices

Day	ABC Returns	XYZ Returns
1	1.1%	3.0%
2	1.7%	4.2%
3	2.1%	4.9%
4	1.4%	4.1%
5	0.2%	2.5%

Next, we need to calculate the average return for each stock:

- For ABC, it would be (1.1 + 1.7 + 2.1 + 1.4 + 0.2) / 5 = 1.30.
- For XYZ, it would be (3 + 4.2 + 4.9 + 4.1 + 2.5) / 5 = 3.74.
- Then, we take the difference between ABC's return and ABC's average return and multiply it by the difference between XYZ's return and XYZ's average return.
- Finally, we divide the result by the sample size and subtract one. If it was the entire population, you could divide by the population size.

$$\text{Covariance} = \frac{\sum(\text{ReturnABC} - \text{AverageABC}) * (\text{ReturnXYZ} - \text{AverageXYZ})}{(\text{Sample Size}) - 1}$$

Using our example of ABC and XYZ above, the covariance is calculated as:
= [(1.1 - 1.30) x (3 - 3.74)] + [(1.7 - 1.30) x (4.2 - 3.74)] + [(2.1 - 1.30) x (4.9 - 3.74)] + …
= [0.148] + [0.184] + [0.928] + [0.036] + [1.364]
= 2.66 / (5 - 1)
= 0.665

In this situation, we are using a sample, so we divide by the sample size (five) minus one.

The covariance between the two stock returns is 0.665. Because this number is positive, the stocks move in the same direction. In other words, when ABC had a high return, XYZ also had a high return.

Thank you