Probability - Distributions

Juan Arellano

KSM Consulting

jarellano@ksmconsulting.com

November 9, 2020

Overview

What is a distribution?

A probability distribution is the possible values a variable can take and how frequently they occur

- Y is the actual outcome of an event
- y is one of the possible outcomes

So for instance, Y could be the number of red marbles we draw out of a bag, while y could be 5

Key Definitions

Distributions have two characteristics

- mean - average value
- variance - how spread out the data is
- The mean of a distribution is often denoted as μ
- The variance of a distribution is often noted as σ^2

Key Definitions

Distributions have two characteristics

- mean - average value
- variance - how spread out the data is
- The mean of a distribution is often denoted as μ
- The variance of a distribution is often noted as σ^2

More Key Definitions

Population data is the term we use when are talking about "all of the data", while sample data is only part of "all of the data". Sample and population notation is different.

- Sample mean: $\bar{x}$
- Sample variance: s^2

More Key Definitions

Standard Deviation is the positive square root of variance, or σ when dealing with population data or s when dealing with a sample. A constant relationship exists between mean and variance

$$\sigma^2 = E((Y - \mu)^2)$$

Types of Distributions

There are various types of probability distributions. We will examine the characteristics of the most common distributions.
Notation: X, variable, and , Tilde, N, type, and (μ, σ^2), characteristics.

Discrete distributions:

- Uniform distributions
- Bernoulli distribution
- Binomial distribution
- Poisson distribution

Continuous distributions:

- Normal distribution
- Student's T distribution
- Chi-Squared distribution
- Exponential distribution
- Logistic distribution

Types of Distributions

There are various types of probability distributions. We will examine the characteristics of the most common distributions. Notation: X, variable, and , Tilde, N, type, and (μ, σ^2), characteristics.

Discrete distributions:

- Uniform distributions
- Bernoulli distribution
- Binomial distribution
- Poisson distribution

Continuous distributions:

- Normal distribution
- Student's T distribution
- Chi-Squared distribution
- Exponential distribution
- Logistic distribution

Key Characteristics of Types of Distributions

Discrete Distributions

- Finitely many distinct outcomes
- Every unique outcome must have a probability assigned to it
- One peculiarity, $P(Y <= y) = P[Y < (+1)]$

Continuous Distributions

- Sample space is infinte, therefore we cannot record the frequency of each distinct value
- The probability for any individual value is equal to 0. Therefore, P(X ¡ 6) = P(X ¡= 6)
- The relationship between the PDF and the CDF of a continuous distribution is that the CDF is the area under the curve of the PDF

Key Characteristics of Types of Distributions

Discrete Distributions

- Finitely many distinct outcomes
- Every unique outcome must have a probability assigned to it
- One peculiarity, $P(Y <= y) = P[Y < (+1)]$

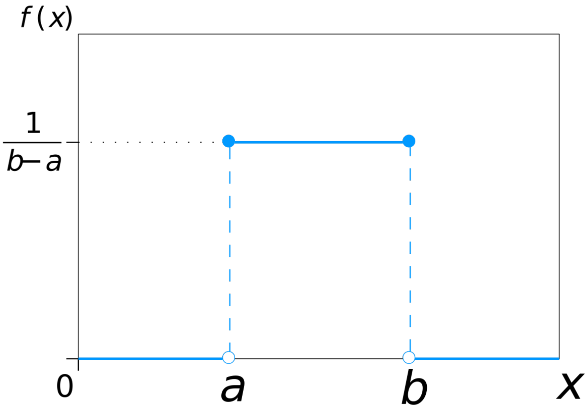
Continuous Distributions

- Sample space is infinte, therefore we cannot record the frequency of each distinct value
- The probability for any individual value is equal to 0. Therefore, P(X ¡ 6) = P(X ¡= 6)
- The relationship between the PDF and the CDF of a continuous distribution is that the CDF is the area under the curve of the PDF

Uniform Distribution

- We use the letter U(a,b) where a,b are the end points
- All outcomes must have equal probability
- One drawback: the expected value provides us no relevant information
- We can still calculate mean and variance, however they are completely uninterpreted and possess on predictive power

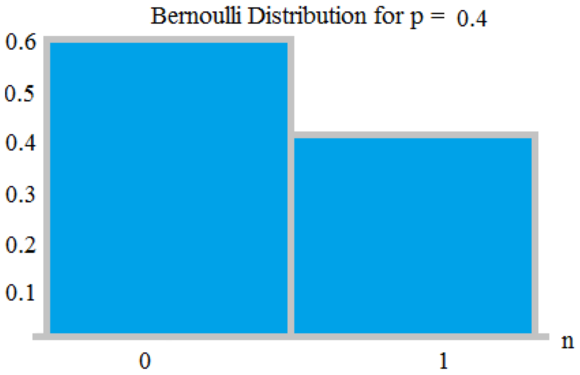
Uniform Distribution



Bernoulli Distribution

- We use Bern(p) to note our Bernoulli distribution, where p is the probability of success
- Events with 1 trial and 2 possible outcomes follow this distribution
- Usually have to assign which outcome is 0 and which one is 1
- Expected value will either be p or 1-p
- Conventionally we usually assign 1 with the probability equal to p
- Variance: p(1 - p), regardless of what the expected value is

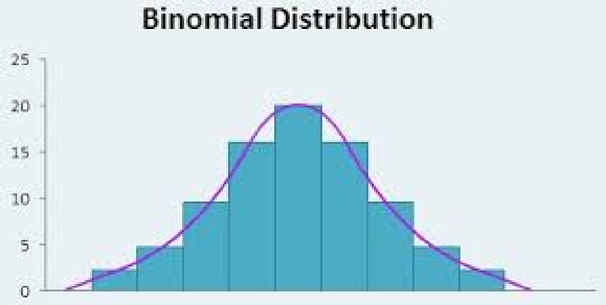
Bernoulli Distribution



Binomial Distribution

- A sequence of identical Bernoulli events
- Notation: B(n, p) where n is the number of trials and p is the probability of success of each individual trial
- A quiz with 10 T/F questions, guessing on one question is a Bernoulli event, the whole quiz is a Binomial event
- If we wish to find out the number of ways in which 4 out of 6 trials can be successful, we would need to find the number of combinations
- Expected value: the sum of all values in the sample space multiplied by their probabilities

Binomial Distribution

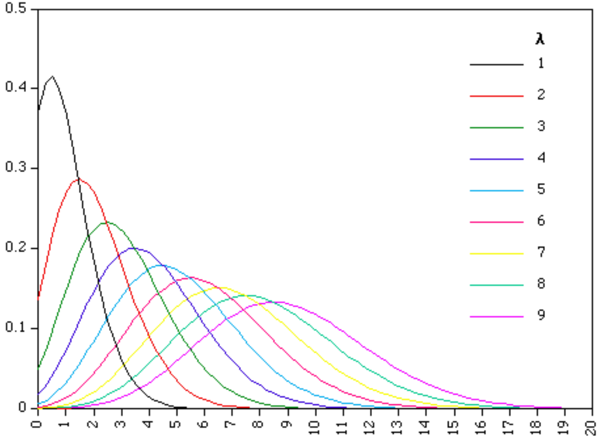


Poisson Distribution

- Notation: Po(λ)
- Deals with the frequency with which an event occurs instead of a probability it requires how often an event occurs
- For instance, a firefly will light up 3 times in 10 seconds, we can use a Poisson distribution to see how many times it lights up in 20 seconds
- No event can happen a negative amount of times so the graph starts at 0
- Poisson distribution is wildly different, it's probability function is just as different $P(Y) = \frac{\lambda^y * e^y}{y!}$

Poisson Distribution

The Poisson Distribution



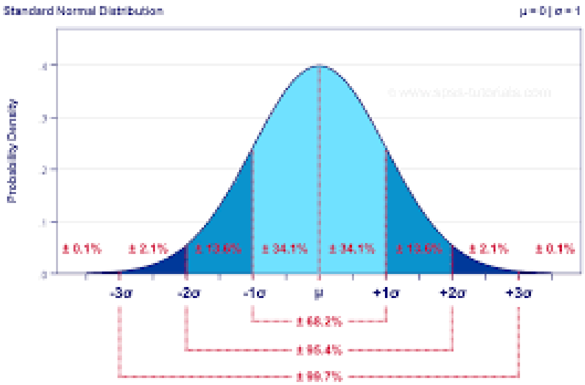
Normal Distribution

- We define a normal distribution using $N(\mu, \sigma^2)$
- Often appears in nature and life, the size of a full grown male lion follows a normal distribution

Distinct characteristics:

- Graph follows a bell shape
- Graph is symmetric - A lion is equally likely to weight 350 pounds or 450 pounds since they are equally away from the mean, 400
- $\mathrm{E(X)} = \mu$
- We can deduce variance from the expected value
- "68, 95, 99.7," 68 percent of all outcomes fall 68 percent within one standard deviation from the mean, 95 percent fall within two standard deviations, and 99.7 within three
- Outliers are extremely rare in normal distributions

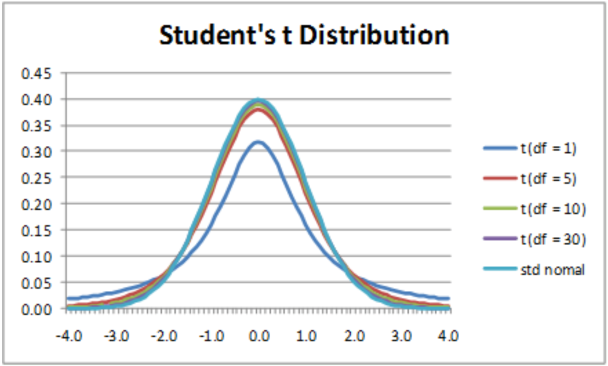
Normal Distribution



Student's T Distribution

- Notation t(k) where k is the degrees of freedom
- We use the student's t distribution when we don't have enough a sample size
- Curve is also symmetric and bell shaped
- The tales are bigger to accomodate the occurence of extreme values away from the mean
- If k ¿ 2, the expected value of a t distribution is the mean
- Frequently used when conducting statistical analysis
- Useful for hypothesis testing with limited data

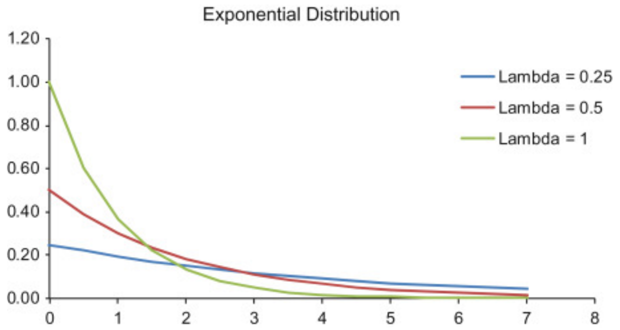
Student's T Distribution



Exponential Distribution

- Defined by Exp(λ)
- Start off highs, initially decreases, then eventually plateauing
- Think of a YouTube vlog where it increases in views at the start but eventually plateaus
- The curve resembles a sort of boomerang
- λ is our rate parameters and is how fast our curve plateaus
- The expected value is $\frac{1}{\lambda}$

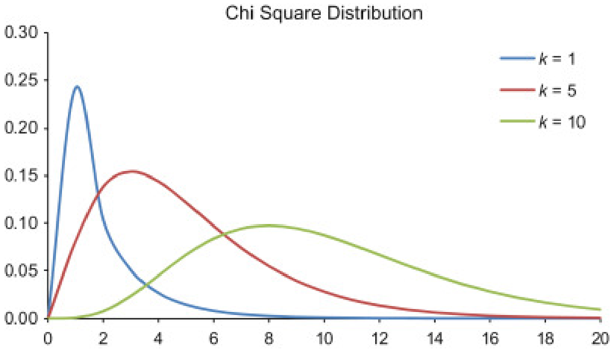
Exponential Distribution



Chi-Squared Distribution

- We denote it with $\chi^2(k)$ where k is the degrees of freedom
- Very few events follow such a distribution
- Useful for hypothesis testing and computing confidence intervals
- Distribution is asymmetric and highly skewed to the right and graph starts at 0 rather than a negative number
- Contains a table of known values like the normal and student's t distribution

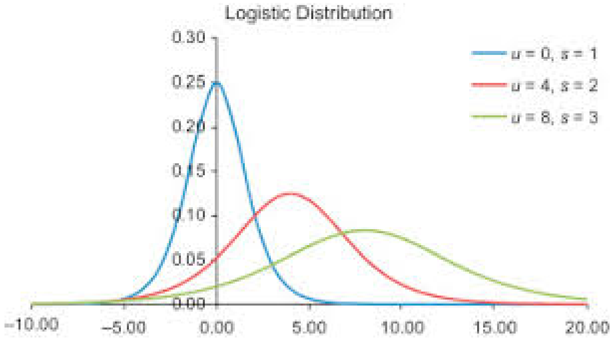
Chi-Squared Distribution



Logistic Distribution

- Defined by Logistic(μ, S), where μ is the location and S is the scale
- The graph of the Logistic distribution resembles that of the Normal distribution
- Used to help us forecast binary outcomes

Logistic Distribution

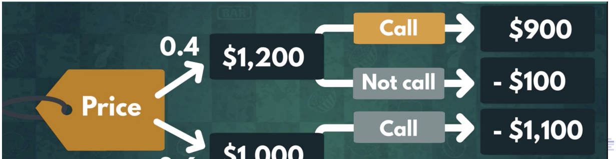


Probability in Finance

An option is an agreement between two parties for the price of a stock or item at a future point in time. Since this pact puts one of the parties at a disadvantage, the person at an advantage pays for an option. For instance, say you are offered to buy 10 stocks of Google for 1,100 in an option, of which you are paying a 100 premium.The stock price has a .40 likelihood of increasing to 1,200 while there's a .60 chance of it dropping to 1,000. We can make a decision tree.

Probability in Finance

If the price drops to 1,000 and you do not call the option, you lose 100 and if you do call the option you lose 1,100 dollars. If the price increases and you do not call the option, you still lose 100 but if you call the option you make 900. We can calculate the expected payoff. If it's less than 0, it's disadvantageous to make the deal. If it's 0, it's a fair deal. If the payoff is positive, then the deal is favourable. The expected payoff can be calculated by the following. Assuming you buy the option, if the price increases, you have a payoff of 900, while if it decreases you lose 100. Expected value $=$ 0.6(-100) $+$ 0.4(900) $=$ 300. This means the deal is favourable. Investors can charge a higher premium to make it a fair deal.

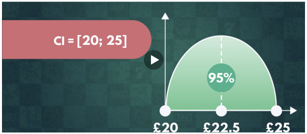


Probability in Statistics

Statistics focuses predominantly on samples and incomplete data. We then must focus highly on expected values and prediction intervals. Statistics builds on probability, and data science builds on statistics. The three crucial requirements for hypothesis testing in statistics are mean, variance, and type of distribution. In the field of statistics, we are often provided sample data without knowing the type. When determining the distribution, any distribution predicts a value for all points within our dataset. The distribution anticipate the actual data point. The more distributions we know the easier it will be to determine which one we are dealing with for any problem. What statisticians call as mathematical modeling, data scientists call supervised machine learning.

Probability in Statistics

Imagine you are on a trip to London. You want to decide how much money you should bring to eat. You use Yelp to get data on average price of food at 10 restaurants. Using standard deviation, we can calculate our Confidence Interval and it will tell us we should carry between 20 and 25 dollars as that's 95 percent confidence interval of prices of food.



Probability in Data Science

In data analysis, we usually try to analyze past data, find insights, to then make predictions about the future. In mathematical modeling, we run artificial simulations to see how well our predictions match up to various future scenarios. In mathematic modeling, we often can run a Monte Carlo simulation, where we generate artificial data to test the predictive power of our mathematical models. We don't use completely random data, but we must follow certain restrictions. In essence, data science is the expansion of probability, statistics, and programming that implements computational technology to solve more advanced questions. However, even when using it to predict future outcomes, there is still great uncertainty.