STA-1380 – Elementary Statistics
Week #3

Hello! Welcome to the additional online **Weekly Resources** for the course of **STA-1380**. Following a traditional calendar semester, these will be some of the topics your professors will go over. If you do not see material your section is going over for the week, please look at the other resources listed for this course. In addition to these resources, there might be **Group Tutoring** for this course, please see our website for more details. These sessions will go over these materials in more detail as well as any questions about the material.

Any additional help or services can be found through the **Baylor Tutoring Website.** Visit to schedule a free 30-minute private tutoring session, drop-in times for your course, the Baylor Tutoring YouTube channel, or any additional tutoring resources.

Contacts: Sid Rich M-Th 9am-8pm (Fall and Spring class days) Office Phone: 254-710-4135

Topic of the Week:
"Discrete Probability Distributions"

Key Points:

- Probability Mass Function
- Binomial Distribution
- Cumulative Distribution Functions

One of the most important aspects of Statistics is the data that backs up a study. Using the raw data, distributions are often computed as a visual representation of the probabilities seen in the data to help analysts interpret the data and form meaningful conclusions. These are called Probability Distributions. Depending on the nature of a variable, there are two distribution families that can represent data and form the basis of Statistics: Discrete and Continuous Distributions. During this week, we will go over the nature and example of a Discrete Distribution.

Discrete Distributions: These are centered around data pools of Discrete Random Variables. In which the areas that the data on the graph can occupy are *finite*. If the data is often presented as a whole number or if the presence of half of a unit is impossible, a discrete model is needed.

> Think of throwing darts. Say you plan to run an experiment that records how many bullseyes a student can hit. If they have 10 darts, they have 10 chances to make a

bullseye. Since there is zero probability of landing half of a dart on a bullseye, a Discrete Distribution would be the best fit model for this experiment.

Another example would be to record how many swings (or trials) a batter must take before they hit the ball. The batter cannot take half a swing, making the variable finite. The probability of each swing being the final attempt is quantifiable, therefore, the data is discrete.

In many cases, the data that will be interpreted will match the characteristics of an existing distribution. Some of the most common distributions that this course will focus on include the Binomial, Normal, Student T, and Chi-Squared distributions. Note that there are many more distributions with varying complexity to come. For now, look out for these distributions, it is wise to become very familiar with each of them.

Highlight #1
"Probability Mass Function"

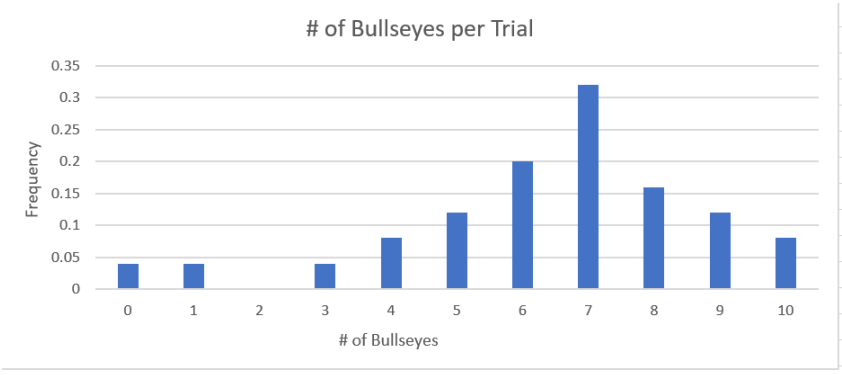
Definition: A distribution of each possible value a discrete random variable can take along with the probability associated with each value.

Notation: $f(x) = P(X = x)$

Examples: Function Tables, Histograms (Bar graph), or Function Notations

$$f(x) = P(X = x) \begin{cases} 1/4, \text{ if } x = 1 \\ 1/4, \text{ if } x = 2 \\ 1/4, \text{ if } x = 3 \\ 1/4, \text{ if } x = 4 \\ 0, \quad \text{otherwise} \end{cases}$$

* A piecewise function depicting a 25% chance evenly spread out amongst 4 possible values



* 30 simulated trials for the "Darts" example

X	1	2	3	4	5	6	7
f(x)	0.7	0.21	0.063	0.0189	0.00567	0.001701	0.00051

*A Table depicting the probability of how many swings it takes for a batter to hit a pitch.

The PMF is one of the most important tools a statistician can use when dealing with discrete data. The 'P' in PMF can help you remember that the data is trying to show the likelihood of the data's occurrence. In the batting example, we can see that 70% of the probability distribution is on the first swing. This means that for P(X=1) will have .7 as the probability. Remember that for a PMF, the value of P(X=x) will have an exact value. That is why we use the word 'Mass' to describe the quantifiable amount of probability each potential 'x' value could have.

Later in the course, data points with lower probability will be the center focus of all statistical inferences. Data like the chance of only throwing 1 dart with a frequency of .033, or the batter taking up to 5 swings before striking his first ball are both instances that show a lower chance of happening naturally.

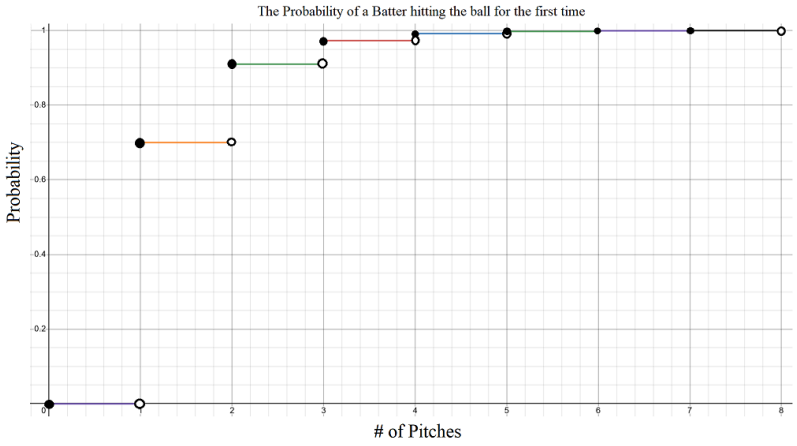
Highlight #2
"Cumulative Distribution Function"

Definition: The individual PMF values of a data distribution combined and added to each increasing 'X' value to depict a cumulative probability.

Notation: $F(x) = P(X \leq x)$ or $F(x) = \sum f(x)$

Example: Piecewise Function Graphs, Formulas, and Tables

X	F(x)
0	0.03333
1	0.06667
2	0.06667
3	0.1
4	0.16667
5	0.26667
6	0.43333
7	0.7
8	0.83333
9	0.93333
10	1



*A CDF Table for the Darts experiment * A CDF Graph of the Batting Experiment

The CDF is another visual aid used for both discrete and continuous distributions alike. The only difference is that for a discrete CDF the graph is a piecewise or 'step' function rather than a continuous line. CDFs for the Continuous Functions are not shown here and will be covered in next week's lecture notes.

Unlike the PMF, the CDF's purpose is to show the distribution of the probabilities and to show how the entire data set reaches closer and closer to 1. Remember, any probability distribution can only equal up to 1. The gaps between the piecewise steps are exactly equal to the individual PMFs of each 'X' value. Take the figure showing the batting probabilities. We know from the previous table that the likelihood of the batter hitting the first ball is .7, and the gap between 1 and 0 on this graph is .7 as well. A PMF can be derived from a discrete CDF by subtracting two values and finding the gaps between them.

It is also important to remember inequalities when dealing with CDFs. $P(X \leq 5)$ and $P(X < 5)$ do not produce the same values for a discrete CDF. Using the Darts table, the value of $P(X \leq 5)$ is equal to .2667 because it includes 5. However, $P(X < 5)$ is only .1667, lacking the additional .1 that the PMF of $X = 5$ would offer. If there is a bar below the inequality sign, include that value. Graphically, this is shown by the open and closed circles. For the batting experiment, $P(X \leq 2)$ is .91 because the closed circle is included and therefore the value includes the additional .21 that spans the entire $2 \leq X < 3$ area.

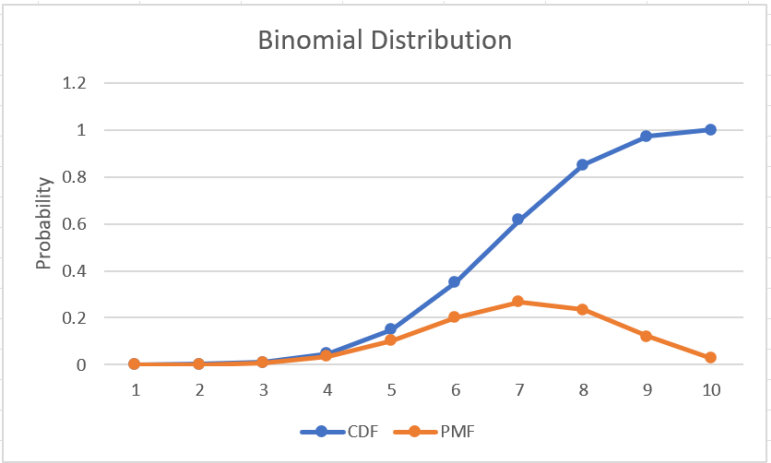
Highlight #3
"Binomial Distribution"

Definition: A discrete distribution with 4 requirements: Independence, Probability of Success, Number of Trials, and Distinct Outcomes.

Notation: $X \sim Bin(n, p)$ n = number of trials, p = success rate, q = (1- p) or failure rate
Mean = $n*p$ Variance = $n*p*q$ Std. Dev = $(n*p*q)^{(1/2)}$

Example: Tables and Graphs

X	F(x)	f(x)
1	0.000143686	0.000137781
2	0.001590386	0.001446701
3	0.010592078	0.009001692
4	0.047348987	0.036756909
5	0.150268333	0.102919345
6	0.350389282	0.200120949
7	0.617217214	0.266827932
8	0.850691654	0.233474441
9	0.971752475	0.121060821
10	1	0.028247525



*A table and Graph depicting the PMF and CDF of a Binomial Distribution with 10 trials and a probability of success being .7

Binomial Distributions are one of the first true distributions that all statistians become familiar with. This distribution's PDF is is unimodal which means it has one peak and two tails. This shape is also often refered to as a 'Bell-shaped' curve. The peak of the distribution will always center around the mean of the distribution. The mean of a Binomial Distribution is the expected value which can be calculated by ($n * p$). If there are 10 possible chances to throw a dart into a bullseye and a thrower has a known success rate of making 70% of their throws, then the expected value is logically going to be 7 darts out of 10.

A strong pneomonic device to remember what defines a Binomial model is the Acronym "BINS". B for Bernoulli. I for Independent. N for Number of trials. S for Success rate.

A Bernoulli trial is a term to describe an event with only two potential outcomes, either a success or a failure. In our example, the success would be throwing a dart on a bullseye and a failure would mean that the dart lands anywhere else. A success or failure can be any sort of random variable so long as there are only two outcomes and they are mutually exclusive. Independence is also important for statistical inference.

If a Dart thrower takes all of their dart shots in 1 sitting, they can quickly become warmed up and have the motions and practice of a previous throw impact the next. Therefore, each throw or 'trial' must be independent of each other to remove any chance of impact.

The number of trials must also be clearly established in order to compute a Binomial distribution. Since the number of trials impacts the mean, standard deviation, and variance, it is important to check and see if there are a set number of trials before continuing. (A Binomial Distribution with an unset number of trials such as the batting example is called a Geometric Distribution, which is not covered in this course).

Success is also important because it is the base probability that the model revolves around. A coin has a constant 50% probability of rolling heads or tails. If the probability were to be inconsistant, than the mean, variance, and standard deviation would also be incomputatable.

Important things to remember about any statistical model is how to use or identify the tools of inference from the given details. For a Binomial model, trials and success rates are the basis for the statstitical data. Practice turning means and variances back into the base forms of trials and successes.

Binomial Distribution PMFs have a set function called a combination. Which can be written as:

$$f(x) = P(X=x) = \binom{n}{x} p^x q^{n-x} \quad X = 0,1,2 \ldots . n$$

The symbol '*k*' is used to show the specific (X= x) being identified. The combination is read as "(*n* choose *k*)" which means that for any given number of trials, the PMF will be looking for the exact probability that *k* number of trials are a success. The variable *k* will equal the '*x'* value you are trying to identify. It is important to know how data is computed so that word problems can be deciphered and completed. Sometimes a question may require you to understand how the equation works in order to properly use the data.

The CDF of the Binomial adds all the previous PDFs together, meaning that the traditional notation is $P(X \leq x)$. If you want to reverse the notation, you can take the anti-probability (1-P$(X \leq x)$) Remember that this value will include 'x', so make sure you identify which trials you are accounting for.

Check Your Learning

1. A carnival has implemented a new dice game where a weighted d6 (6-sided die) is rolled. The price to play the game is worth 3 tokens. There is an equal likelihood of rolling a 1 through 3, a 10% chance of rolling a 4, a 5% chance of rolling a 5, and a 1% chance of rolling a 6. If 1-3 is rolled, the player gets nothing. If a 4 is rolled, they get 5 tokens. If they roll a 5, they get 20 tokens. If they roll a 6, they get 75 tokens.
    a. Create a PMF and CDF table of the game's probable outcomes.
    b. Find the xpected outcome of playing the game. Is it worth playing?
2. There are 7 Cats and 3 dogs in a shelter. The shelter wants to find a combination of 5 different animals to show in a picture catalogue.
    a. What is the probability of picking more than 3 cats for the catalogue.
    b. Find the Mean and Standard Deviation of this set?

Things Students Struggle With

1. PMF vs CDF
    a. With acronyms sharing multiple letters, it can be difficult to remember the rules of each distribution and how to classify each one. One way to remember a PMF versus a CDF is that a PMF is a Partial Distribution. It only shows the probability associated with each observed value. A CDF is a Cumulative function. It is the

process of combining all of the previous probabilities until it sums to 1.

2. How to read a Binomial Problem:

a. It is very common to be confused about the phrasing of questions, especially if notation is not used. Here is a written conversion of what each phrase may mean. Recall that the symbol '*X*' means the data result or 'variable' the graph is representing, and the '*x'* is the exact values that the probability notation is discussing.

At least (Greater than and including the value specified) $P(X \geq x)$
At Most (less than and including the value specified) $P(X \leq x)$
More than (Greater than but not including the value) $P(X > x)$
Less than (fewer than but not including the value) $P(X < x)$
Between (The accumulated probabilities (CDF)) of $P(X = a < x < b)$

Concluding Comments

That's it for this week! Please reach out if you have any questions and don't forget to visit the Tutoring Center website for further information at https://www.baylor.edu/tutoring.

Answers to CYL

1. **a.**

X	1	2	3	4	5	6
PDF	0.28	0.28	0.28	0.1	0.05	0.01
CDF	0.28	0.56	0.84	0.94	0.99	1

b. E(x) = (.28)(1) + (.28)(2) + (.28)(3)+(.1)(4)+(.05)(5)+(.01)(6) = 2.39. Because the Expected value is less than the base amount to play (3 tokens), it is not viable to play the game because there is a negative pay out.

2. **a.** 0.528 recall that 'more than' does not include the .309 probability of picking exactly 3 cats.
b. Probability: 7/10 or .7 5 'trials'. Mean = .7*5 or 3.5. Std Dev = 3.5*.3 = 1.05