Activity #8: Discrete Distributions

Last unit, we derived probability models for discrete random variables – variables that can take a countable sequence of values. In this activity, we're going to learn about some special discrete distributions.

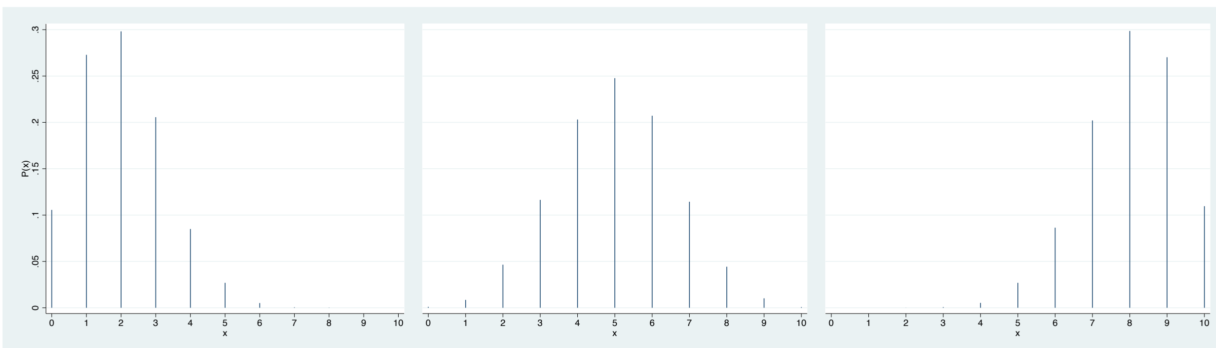
We're already familiar with the Binomial Distribution:

Binomial Distribution

Conditions:
- A series of n independent trials
- Each trial has two possible outcomes (success/failure)
- The probability of a success (p) remains constant over trials
- X = the number of successes in n trials
- E[x] = np   Var[x] = np(1-p)

PMF: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ = binompdf(n, p, x)   CDF: P(X ≤ x) = binomcdf(n, p, x)

Below, you can see the probability mass functions for three binomial distributions with n = 10 independent trials.



p=0.20   p=0.50   p=0.80

1) Suppose you're the manager of a baseball team looking for a free agent that can hit .300. You give that free agent 25 at-bats in spring training and decide that you will give him a contract if he gets at least 9 hits. Suppose this free agent really does have a batting average of .300. What's the probability that you will make a mistake and not give him a contract? How can you reduce this probability?

2) SAU has 3,281 students enrolled in the Fall 2011 semester. Suppose 150 of these students are satisfied with the parking situation on campus. If you select 30 students at random, what's the probability you'll find more than one student who is satisfied with parking? What is the expected number of satisfied students you will find? Have the conditions for a binomial distribution been met in this example? Note: 150/3281 = 0.0457

Scenario: A bored statistics professor plays solitaire during his office hours. Of the 444 games he has played, he has won 74 times without cheating. Winning percentage: 74/444 = 1/6 = 0.167

3) Let X = the number of games the professor must play until his next win? Is X a discrete or continuous random variable?

4) What's the probability the professor wins on his first game? What's the probability his first win is the 2nd game he plays?

5) Calculate the probability of the professor's first winning game occurring on the 3rd or 4th attempt.

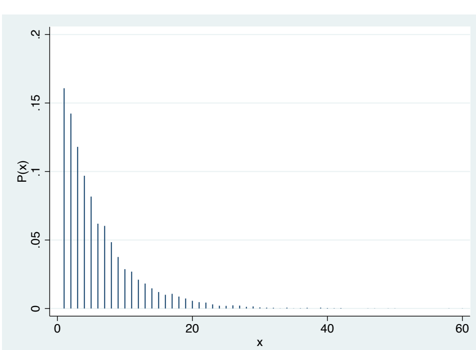
6) Derive a general expression for calculating P(X = k) = the probability that the first win comes on the kth trial.

Geometric Distribution

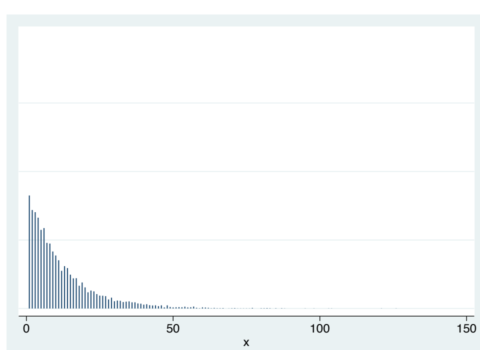
Conditions:
- A series of independent trials
- Each trial has two possible outcomes (success/failure)
- The probability of a success (*p*) remains constant over trials
- P(X = k) = P(the first success comes on the kth trial)
- E[x] = 1/p Var[x] = (1-p) / p^2

PMF: $P(X = k) = p(1-p)^{k-1}$ = geometpdf(p, k)

CDF: P(X ≤ k) = geometcdf(p, k)



Geometric distribution with p = 1/6



Geometric distribution with p = 1/12

7) Recall that we're assuming 4.57% of SAU students are satisfied with the parking situation on campus. Suppose you start randomly calling SAU students to ask if they are satisfied with parking on campus. How many students do you expect to call before finding the first student who is satisfied? What's the probability that you must call 10 or more people before finding the first person who is satisfied?

8) Suppose the probability of an engine malfunctioning during a 1-hour period is 0.02. Find the probability that the engine will survive at least 2 hours. What is the expected value in this scenario and what does it represent?

9) A student sneaks into a professor's office in order to steal the answers to next week's test. The student has only 10 minutes to get the answers, but finds they are safely hidden inside a safe. To open the safe, the student must guess a 4-digit code (using the digits 0-9). If the student only has time to guess 100 codes at random, what's the probability the student will open the safe?

Scenario: Recall the bored statistics professor who has won 1/6 = 0.167 = 16.7% of his solitaire games.

10) What's the probability the professor's 2nd win is on his 3rd attempt? What's the probability his 3rd win is on the 4th attempt? What's the probability his rth win is on the xth game?

Negative Binomial Distribution

Conditions:
- A series of independent trials
- Each trial has two possible outcomes (success/failure)
- The probability of a success (*p*) remains constant over trials
- P(X = r) = P(the rth success comes on the kth trial)
- E[x] = r/p   Var[x] = r(1-p) / p^2

PMF: $P\left(X = r\right) = \begin{pmatrix} x-1 \\ r-1 \end{pmatrix} p^r (1-p)^{x-r}$

11) Suppose a couple has an equal probability of having a male or female child. What's the probability that a couple's 5th child is their 2nd daughter? What's the expected number of children they will have in order to get 2 sons?

12) A geological study indicates that an exploratory oil well drilled in a particular region should strike oil with probability 0.2. Find the probability that the 3rd oil strike is on the 5th well drilled.

Scenario: A jury of 12 people is selected at random from a pool of 16 men and 18 women.

13) In how many different ways could we select 7 women from the 18? How many different groups of 5 men could we select from the pool of 16? How many different groups of 12 people can we select from 34? Calculate the probability of selecting exactly 5 men and 7 women at random in this scenario.

Hypergeometric Distribution

Conditions:
- A pool of objects (N) is dichotomized into two subgroups: M and N-M
- We randomly select a sample of n objects without replacement
- We are asked P(selecting x of the M and n-x of the N-M objects)
- E[x] = nM/N   Var[x] = (nM/N)(1-M/N)((N-n)/(N-1))

PMF: $$\frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} = \frac{\begin{pmatrix} \text{\# of objects in group A} \\ \text{\# of group A objects selected} \end{pmatrix}\begin{pmatrix} \text{\# of objects in group B} \\ \text{\# of group B objects selected} \end{pmatrix}}{\begin{pmatrix} \text{Total \# of objects} \\ \text{Total \# of objects selected} \end{pmatrix}}$$

14) Suppose we want to find the probability distribution for the number of automobile accidents at the corner of Gaines and Locust in a single week. How could we derive this probability distribution?

Let's split the week into n subintervals. In fact, let's make the subintervals so small that *at most, one accident could occur in each subinterval*. Then, if we let p = P(an accident in any subinterval), we have:

P(no accidents occur in a subinterval) = 1 – p
P(1 accident occurs in a subinterval) = p
P(more than 1 accident occurs in a subinterval) = 0 (the subinterval is too small for this to happen)

We can now think of the number of accidents during the week as the *total number of subintervals that contain one accident*. If the occurrence of accidents from subinterval to subinterval can be assumed to be independent, we now have:

- A series of independent trials (subintervals) with two possible outcomes (accident or no accident)
- A constant probability of an accident occurring (p)
- P(X = x) = the probability of observing x accidents in a certain number of trials

Which probability distribution can be used with the above conditions?

15) How do we divide the week into those small subintervals? How many subintervals do we need? Honestly, I don't know. It seems reasonable, though, that as we divide the week into a greater number of n subintervals, the probability p of an accident occurring in one of those shorter subintervals will decrease.

Recall the probability mass function for a binomial distribution: $P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$

If we let the number of trials (subintervals) approach infinity, we have:

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x}=\lim_{n\to\infty}\frac{n(n-1)(n-2)\cdots(n-x)!}{x!(n-x)!}p^x(1-p)^{n-x}=\lim_{n\to\infty}\frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}p^x(1-p)^{n-x}$$

If we let $\lambda = np$, we can write $\lim_{n\to\infty}\frac{n(n-1)(n-2)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$ and rearrange terms:

$$\lim_{n\to\infty}\left(\frac{\lambda^x}{x!}\right)\left(1-\frac{\lambda}{n}\right)^n\frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^{-x}$$

$$=\left(\frac{\lambda^x}{x!}\right)\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x}$$

$$=\left(\frac{\lambda^x}{x!}\right)\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)$$

$$=\left(\frac{\lambda^x}{x!}\right)\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n(1)$$

How do we evaluate this limit? Since we have an indeterminate form, we need to use L'Hopital's Rule:

$$\ln(y) = \lim_{n\to\infty} n\ln\left(1-\frac{\lambda}{n}\right) = \lim_{n\to\infty} \ln\left(\frac{1-\frac{\lambda}{n}}{\frac{1}{n}}\right) = \lim_{n\to\infty}\left(\frac{\frac{\lambda}{n^2}/1-\frac{\lambda}{n}}{\frac{-1}{n^2}}\right) = \lim_{n\to\infty}\left(\frac{-\lambda}{1-\frac{\lambda}{n}}\right) = -\lambda$$

So we now know: $\ln(y) = \lim_{n\to\infty} n\ln\left(1-\frac{\lambda}{n}\right) = -\lambda.$ Solving for y, we find: $y = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$

Plugging this into where we left off on the last page, we get:

$$\lim_{n\to\infty}\left(\frac{\lambda^x}{x!}\right)\left(1-\frac{\lambda}{n}\right)^n \frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x}\left(1-\frac{\lambda}{n}\right)^{-x} = \left(\frac{\lambda^x}{x!}\right)\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \left(\frac{\lambda^x}{x!}\right)e^{-\lambda}$$

So when we use the Binomial Distribution while allowing the number of trials to approach infinity, we can calculate probabilities with:

$$P(X=x) = \left(\frac{\lambda^x}{x!}\right)e^{-\lambda}$$

In our scenario, that's the formula we would use to calculate the probability of observing x automobile accidents in a week.

Poisson Distribution Conditions:

- P(# number of events occur in a fixed interval of time, space, distance, volume)
- The events occur with a known average rate
- The events occur independently of the time since the last event
- $E[x] = Var[x] = \lambda$

PMF: $P(X=x) = \left(\frac{\lambda^x}{x!}\right)e^{-\lambda}$

16) Suppose there are 15 automobile accidents each year on the corner of Gaines and Locust. Calculate the probability that we will have no accidents during a one week period. What's the expected value in this scenario?

Scenario: Entomologists estimate the average person inadvertently consumes almost a pound of bug parts each year. Title 21, Part 110.110 of the *Code of Federal Regulations* allows the Food and Drug Administration to establish "Food Defect Action Levels" -- the maximum level of natural or unavoidable defects in food for human use that present no health hazard.

If you have some time (and the stomach for it), check out your favorite food on the FDA website:
http://www.fda.gov/food/guidancecomplianceregulatoryinformation/guidancedocuments/sanitation/ucm056174.htm

Some examples of the Food Defect Action Levels include:

- Raspberries: Avg. mold count is 60% or more; 4+ larvae per 500 grams; 10+ whole insects per 500 grams
- Chocolate: Avg. 60+ insect fragments per 100 grams; 1 rodent hair per 100 grams
- Macaroni: Avg. 225 insect fragments or 4.5 rodent hairs per 225 grams
- Peanut Butter: Avg. 30 insect fragments per 100 grams

17) Suppose you buy crackers from a vending machine that are spread with 20 grams of peanut butter. Let's further suppose the peanut butter does average 30 insect fragments per 100 grams. What would be the expected value (lambda) for our 20 grams of interest?

18) What are the chances that your vending machine crackers will have no insect fragments?

19) Calculate the probability of finding at least 5 insect fragments in those vending machine crackers.

20) Industrial accidents occur, on average, 3 times per month at a certain plant. In the last 2 months, the plant has experienced 10 accidents. What was the likelihood of this happening?