Activity #8: Special discrete distributions (Binomial, Hypergeometric, Geometric, Negative Binomial, Poisson)

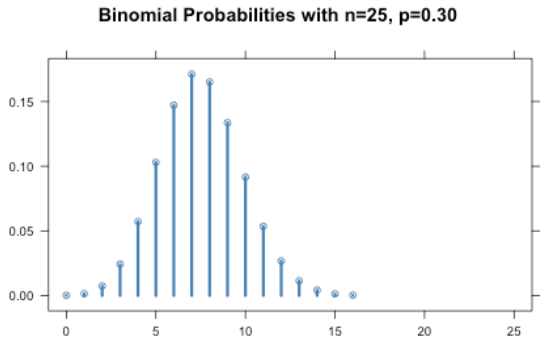
1. You're the manager of a baseball team looking for a free agent that can hit .300. You find one potential free agent and give him 25 at-bats in spring training. You've decided to give him a contract if he gets at least 9 hits. Suppose this free agent really does have a batting average of .300. What's the likelihood you will make a mistake and <u>not</u> give him a contract? How can you reduce this likelihood?

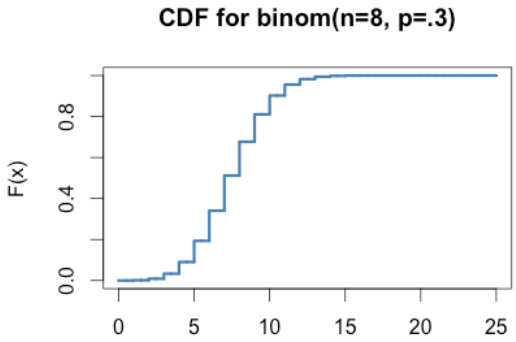
   Given: n = number of trials = _____ Independent trials? **YES** **NO**

   Each trial is dichotomous? **YES** **NO** p = P(success on any trial) = _____

   Calculate: E(x) = _____

   P(X ≥ 9) = _____





2. Suppose I create two versions of the next unit exam. Version A is easy – you could correctly answer 80% of the questions, while Version B is difficult – you could only answer 20% of the items correctly. I give you a sample of 5 test questions from one of the tests and you find that you're able to answer 3 of them.

   It looks like these test questions come from the easy version of the test, but you're not sure. Calculate how much more likely it is that these sample questions came from the easy version.

   Calculate: P(X = 3 | easy version) = _____

   P(X = 3 | difficult version) = _____

   It is _____ times more likely that these questions came from the easy version

Questions 1-2 were examples of the **Binomial Distribution**.

Conditions:
- A series of *n* independent trials, with each trial having two possible outcomes (0 / 1)
- A constant probability of success, *p*, on each trial
- X = the number of successes over those *n* trials

Expected Value: E[X] = np Variance: Var[X] = np(1-p)

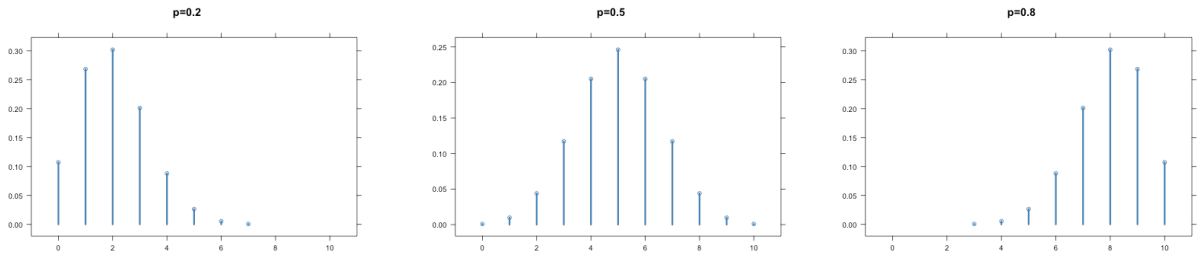
Calculating:

Hand calculator: $P(X = \#) = \binom{n}{\#} p^{\#} (1-p)^{n-\#}$

TI-8x: DISTR menu
PMF: P(X = #) = binompdf(n, p, #)
CDF: P(X ≤ #) = binomcdf(n, p, #)
CDF: P(X > #) = 1 – binomcdf(n, p, #)

R: PMF: P(X = #) = `dbinom(#, n, p, log=FALSE)`
CDF: P(X ≤ #) = `pbinom(#, n, p, lower.tail=TRUE, log.p=FALSE)`
CDF: P(X > #) = `pbinom(#, n, p, lower.tail=FALSE, log.p=FALSE)`
Quantiles: X_q = `qbinom(q, n, p, lower.tail=TRUE, log.p=FALSE)`
Simulating # random values: `rbinom(#, n, p)`

Online: http://stattrek.com/online-calculator/binomial.aspx



3. St. Ambrose had 3,281 students enrolled in the Fall 2011 semester. Suppose 150 (or 4.57%) of these students were satisfied with parking on campus. If you select 30 of these students at random, what's the probability you would find more than one student who was satisfied with parking? Have the conditions for a binomial distribution been satisfied?

4. A jury of 12 people is selected at random from a pool of 16 men and 18 women. Calculate the probability this jury will have exactly 5 men and 7 women.

Questions 3-4 were examples of the **Hypergeometric Distribution**.

Conditions:
- A pool of objects (N) consists of two subgroups of size M and N–M
- We randomly select a sample of *n* objects without replacement
  - Notice the probability of selecting a particular type of object changes each trial
- We're interested in calculating P(selecting x of the M and n–x of the N–M objects)

$$E[X] = n\frac{M}{N} = np \qquad Var[X] = \left[\frac{N-n}{N-1}\right] n\frac{M}{N}\left[1-\frac{M}{N}\right] = \left[\frac{N-n}{N-1}\right] np(1-p)$$

Notice the expected value is the same as that for the Binomial Distribution. The variance only differs by the finite population correction factor (N–n) / (N–1). If our sample *n* is relatively small compared to a large population *N*, then the binomial distribution provides a good approximation of the Hypergeometric.

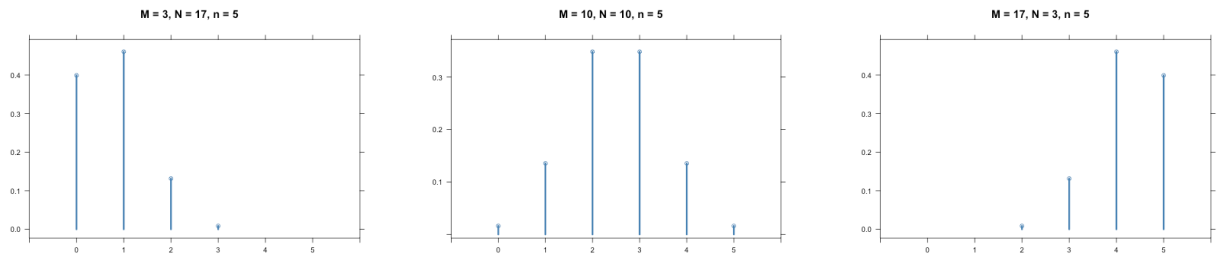
Calculating:

Hand calculator: $P(X=\#) = \dfrac{\binom{M}{\#}\binom{N-M}{n-\#}}{\binom{N}{n}}$

TI-8x: No direct method; use nCr to calculate combinations (MATH -> PRB)

R: PMF: P(X = #) = `dhyper(#, M, N, n, log=FALSE)`
CDF: P(X ≤ #) = `phyper(#, M, N, n, lower.tail=TRUE, log.p=FALSE)`
CDF: P(X > #) = `phyper(#, M, N, n, lower.tail=FALSE, log.p=FALSE)`
Quantiles: X_q = `qhyper(q, M, N, n, lower.tail=TRUE, log.p=FALSE)`
Simulating # random values: `rhyper(#, M, N, n)`

Online: http://stattrek.com/online-calculator/hypergeometric.aspx



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. That gives him a winning percentage of 0.167 (or 1/6).

5. Calculate the following for this professor:

- P(win on first game) = _____
- P(1st win on 2nd game) = _____
- P(1st win on 3rd game) = _____
- P(1st win on 4th game) = _____
- P(1st win on k^{th} game) = _____

Question #5 is an example of the **Geometric Distribution**.

Conditions:
- A series of *n* independent trials, with each trial having two possible outcomes (0 / 1)
- A constant probability of success, *p*, on each trial
- P(X = k) = probability that the first success comes on the kth trial

Expected Value: E[X] = 1/p Variance: Var[X] = (1 – p) / p^2

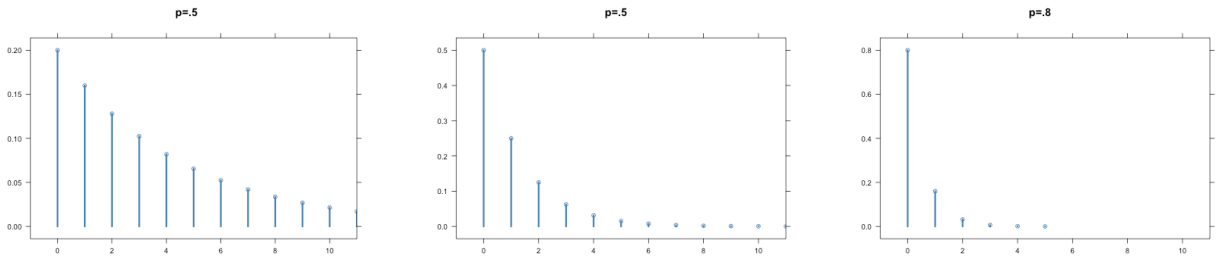
Calculating:

Hand calculator: $P(X = \#) = p(1-p)^{k-1}$

TI-8x: DISTR menu
PMF: P(X = #) = geometcdf(p, #)
CDF: P(X ≤ #) = geometcdf(p, #)
CDF: P(X > #) = 1 – geometcdf(p, #)

R: PMF: P(X = #) = `dgeom(#–1, p, log=FALSE)`
CDF: P(X ≤ #) = `pgeom(#–1, n, p, lower.tail=TRUE, log.p=FALSE)`
CDF: P(X > #) = `pgeom(#–1, n, p, lower.tail=FALSE, log.p=FALSE)`
Simulating # random values: `rgeom(#, p)`

Online: http://www.stat.berkeley.edu/~stark/Java/Html/ProbCalc.htm



6. Recall that 4.57% of St. Ambrose students are satisfied with parking on campus. Suppose you start randomly calling SAU students to ask if they are satisfied with parking.

   - E[X] = expected number of students you will call before finding 1st student who is satisfied = _____
   - P(Call 10 or more students before finding first person who is satisfied) = _____

7. A student sneaks into a professor's office to steal the answers to next week's test, but finds those answers secure in a safe. To open the safe, the student must guess a 4-digit code (using 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?

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. Evaluate whether the conditions for a geometric distribution have been met.

Scenario: Recall the bored statistics professor who wins 0.167 of the time in solitaire.

9. Calculate the following:

   - P(2nd win on 3rd attempt) = _____
   - P(3rd win on 4th attempt) = _____
   - P(rth win on kth attempt) = _____

Question #9 is an example of the **Negative Binomial Distribution**.

Conditions:
- A series of *n* independent trials, with each trial having two possible outcomes (0 / 1)
- A constant probability of success, *p*, on each trial
- P(X = k) = probability that the first success comes on the kth trial

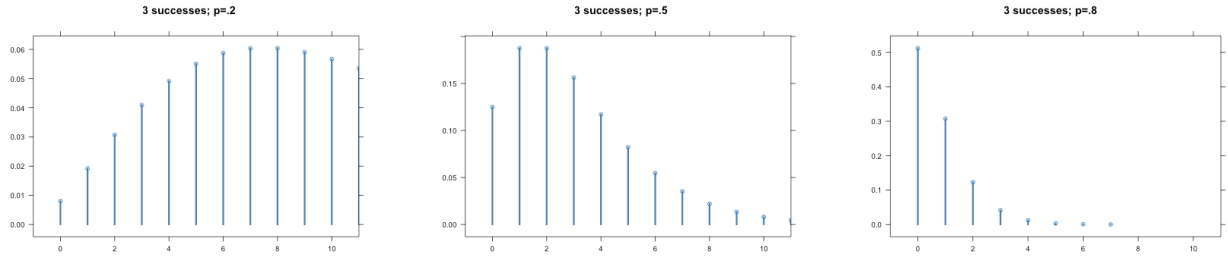
Expected Value: E[X] = r/p Variance: Var[X] = r(1 – p) / p^2

Calculating:

Hand calculator: $P(\text{rth win on kth trial}) = \binom{k-1}{r-1} p^r (1-p)^{k-r}$

R: PMF: P(rth win on kth trial) = `dnbinom(k–r, r, p, log=FALSE)`
CDF: P(rth win ≤ kth trial) = `pnbinom(k–r, r, p, lower.tail=TRUE, log.p=FALSE)`
CDF: P(rth win ≤ kth trial) = `pnbinom(k–r, r, p, lower.tail=FALSE, log.p=FALSE)`
Simulating # random values: `rgeom(#, r, p)`

Online: http://stattrek.com/online-calculator/negative-binomial.aspx



10. Suppose a couple has an equal probability of having a male or female child.

- P(5th child is 2nd daughter) = _____

- Expected # of children to get 2 sons) = _____

11. 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.

Suppose we want to model the probability of the number of automobile accidents at the corner of Gaines and Locust in a year. How could we derive this probability distribution?

Let's split the week into *n* subintervals (trials). We could choose for our trials to be measured in months, weeks, days, hours, minutes, or seconds, but we ultimately decide to choose subintervals to be so small that only one accident could possibly occur in any subinterval. Then, if we define **p = P(an accident in any subinterval)**, we have:

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

The number of accidents during the year can now be thought of 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 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

12. Which probability distribution can be used with the above conditions?

How do we divide the year into those small subintervals? How many subintervals do we need? I don't know, but it seems reasonable that as we divide the year into a greater number of *n* (smaller) subintervals, the probability of an accident occurring in an interval will decrease.

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

If we let the number of subintervals (trials) 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:

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

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

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

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

We have an indeterminate form, so we can use L'Hopital's Rule to evaluate this limit…

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

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:

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

To calculate probabilities under a Binomial Distribution with a near-infinite number of trials, we use:

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

That's the formula we would use to calculate the probability of observing x automobile accidents in a year.

Poisson Distribution.

Conditions:
- Events occur independently with a known average rate
- P(# of events occur in a fixed interval of time, space, distance, volume)
- Expected value = variance

Expected Value = Variance = λ.

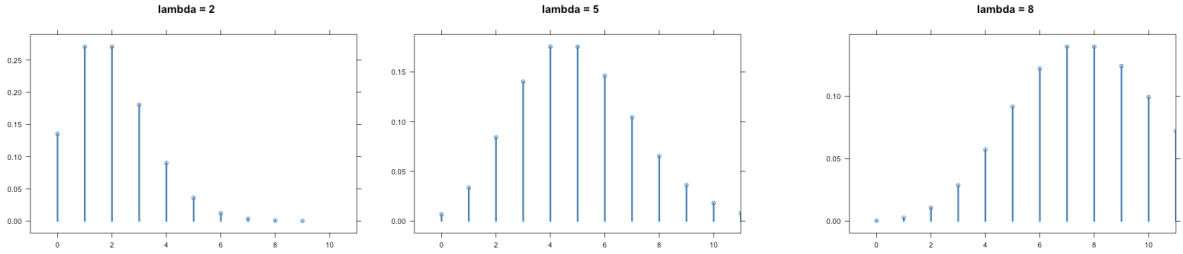
Calculating:

Hand calculator: $P(X=\#) = \frac{\lambda^x}{x!}e^{-\lambda}$

TI-8x: DISTR menu
PMF: P(X = #) = poissonpdf(λ, #)
CDF: P(X ≤ #) = poissoncdf(λ, #)
CDF: P(X > #) = 1 – poissoncdf(λ, #)

R: PMF: P(X = #) = `dpois(#, λ, log=FALSE)`
CDF: P(X ≤ #) = `ppois(#, λ, lower.tail=TRUE, log.p=FALSE)`
CDF: P(X > #) = `ppois(#, λ, lower.tail=FALSE, log.p=FALSE)`
Simulating # random values: `rpois(#, λ)`

Online: http://stattrek.com/online-calculator/poisson.aspx



13. Suppose there are 15 automobile accidents each year on the corner of Gaines and Locust.

- P(no accidents during a one week period) = _____
- Expected value = _____

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 500g; 10+ whole insects per 500g
- 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

14. 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?

15. What are the chances your vending machine crackers will have no insect fragments?

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

17. According to the factory manager, industrial accidents occur, on average, 3 times per month. An OSHA inspector believes that rate is actually higher; perhaps as high as 8 times per month.

    In the last 2 months, the factory has experienced 10 accidents. Calculate the likelihood of this under both the beliefs of the manager and the OSHA inspector.