BIOS 4120: Introduction to Biostatistics
Breheny

Lab #7

I. Binomial Distribution

$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

RCode: dbinom(x, size, prob)
binom.test(x, n, p = 0.5)

$P(X < K) = P(X = 0) + P(X = 1) + \ldots + P(X = k-1)$
$P(X \geq 1) = 1 - P(X = 0)$

Assumptions:

- The number of trials n must be fixed in advance
- The probability that the event occurs, p, must be the same from trial to trial
- The trials must be independent
- Only two possible outcomes

II. Practice Problems

1) An agent sells life insurance policies to five equally aged, healthy people. According to recent data, the probability of a person living in these conditions for 30 years or more is 2/3. Calculate the probability that after 30 years:

Use $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ formula
N = 5, p = 2/3

a. All five people are still living.
K = 5…. P(X = 5) = 0.135
b. at least three people are still living.
1 – P(X = 0) – P(X=1) – P(X=2) = 0.795

c. Exactly two people are still living.
P(X=2) = 0.16

2) A pharmaceutical lab states that a drug causes negative side effects in 3 of every 100 patients. To confirm this affirmation, another laboratory chooses 5 people at random who have consumed the drug. What is the probability of the following events?

Use P(X = k) = $\binom{n}{k}p^k(1-p)^{n-k}$ formula
N = 5, p = 0.03

a. None of the five patients experience side effects.
P(X = 0) = 0.86

b. At least two had side effects.
1 – P(X = 0) – P(X=1) = 0.008

c. It is highly plausible that Hispanic people experience side effects more often than Caucasian patients. Suppose of the 5 people; three are Caucasian and two are Hispanic. Is this a problem for the previous two situations? Explain.

- Yes because the assumption of a constant probability for the binomial distribution will be violated, since Hispanic people will have a higher probability for a bad effect.

3) Let X = the number of 65- to 74-year-olds who suffer from diabetes in the sample of size 7. X is a Bin(7, 0.125) random variable.
Use P(X = k) = $\binom{n}{k}p^k(1-p)^{n-k}$ formula
N = 7, p = 0.125

a. If you wish to make a list of the seven persons chosen, how many ways can they be ordered?
7! = 5040

b. Without regard to order, in how many ways can you select four individuals from this group of 7?
7!/(4!*3!) = 35

c. What is the probability that two of the seven people have diabetes?
P(X = 2) = 0.17

d. What is the probability that four of the seven people have diabetes?
P(X = 4) = 0.006

4) Suppose you are interested in monitoring air pollution in LA over a one-week period. Let X be a random variable that represents the number of days out of seven on which the concentration of carbon monoxide surpasses a specified level. Do you believe X has a binomial distribution? Explain.

- No, because if one of the seven days has a concentration of CO above a certain level, then more than likely the probability of it still being above that level is higher for the next few days. Independence would be violated.

III. Quiz Review

$\hat{\beta} = r\frac{\sigma_y}{\sigma_x}$

$y = \alpha + \beta x$ What is α? What is β?

The correlation coefficient says that if you go up in *x* by one standard deviation, you can expect to go up in *y* by r standard deviations (standard units).

Predicting y with x

1. $z_x = \frac{x-\bar{x}}{SD_x}$
2. $z_y = rz_x$
3. $y = \bar{y} + z_y SD_y$

Plots and Descriptive Measures

Be familiar with: histograms, boxplots, barcharts, standard deviations (+/- 1, +/- 2), mean, median, percentiles, skewness.

Probability

Intersections, unions, complements

Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multiplication rule: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$

$P(A^C) = 1-P(A)$

$P(A) = P(A \cap B) + P(A \cap B^C)$

Bayes' Theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(\overline{B}|A)+P(A^C)P(B|A^C)}$$

Diagnostic Tests
Sensitivity: P(T|D)
Specificity: $P(T^-|D^-)$
Prevalence: P(D)

IV. Practice Problems

1. What does the Pearson correlation coefficient measure?
-- the linear association between two continuous variables

2. It is hypothesized that there are fluctuations in norepinephrine (NE) levels which accompany fluctuations in affect with bipolar affective disorder (manic-depressive illness; low affect scores represents increased mania). Let's say the regression line looks like:

NE = 39 – 0.017*Affect

a. What is the relationship between norepinephrine levels and affect test score?
-- Since the slope is negative we can say on the average, as affect score increases then NE levels decrease. Low affect scores result in higher NE levels.

b. Interpret the slope coefficient.
-- For a one unit increase in affect score, we can expect the NE level to decrease by 0.017 units.

c. Find the correlation coefficient if the standard deviation for NE and Affect is 8.43 and 384.9, respectively.
-0.017*384.9/8.43 = -0.78

3. Given a dataset:
3.21 3.38 4.19 4.37 4.71 4.76 4.79 5.06 5.23 5.36 5.50 5.56 5.64 5.76

a. Find the 25th and 75th percentiles.
25th :14*.25 = 3.5…position 4 is 4.37 ; 75th : 14*.75 = 10.5…position 11 is 5.5

b. find the mean and median.
4.82 ; 4.93

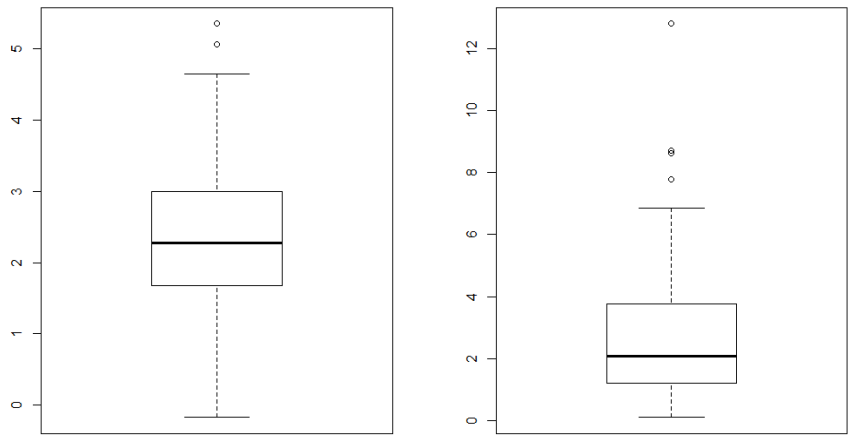
c. Is this data skewed or symmetric?
Slightly skewed left, however it is not by much. It is roughly symmetric.

4. The prevalence of colon cancer is 40%. A colonoscopy can test for colon cancer, and it has a sensitivity of ___ and a specificity of ___. The predictive value positive (PVP) of this test is about ____.

	Colon Cancer	No Colon Cancer
Positive Test	30	20
Negative Test	10	40

Sensitivity = 30/40 = 0.75; specificity = 40/60 = 0.67
PVP = (.75*.4)/(.75*.4 + .33*.6) = 0.6

5. Examine the following boxplots:



Which boxplot has the higher median? Has the most outliers? Has the most variability? Are both data sets symmetric? What are the components of the boxplot? Explain.

The boxplot on the left; the boxplot on the right; the boxplot on the right; the left boxplot is roughly symmetric and the boxplot on the right is right skewed; 25th and 75th percentiles are the 'boxes', and the thick bar is the median. The circles are 'outliers'

6. The probability of event A occurring is 47%. The probability of event B occurring is 18%. The probability of both events occurring at the same time is 10%.

a. Is event A independent of event B?

no; .47*.18 = 0.08 which does not equal 0.1

b. Find P(B|A) and P(A|B).

.1/.47 = 0.21; .1/.18 = 0.56

c. Find P(A U B).

.47 + .18 - .1 = 0.55

The rest of the lab is open for questions.