BiostatisticsMidterm

Tanvir Khan (tk2886)

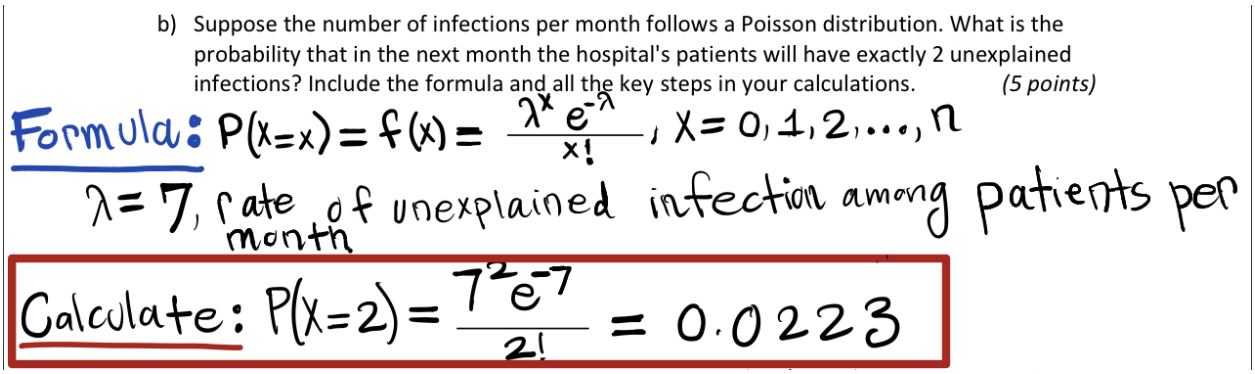
11/04/2021

Problem 1

a). Assumptions that must be true to use the Poisson Distribution to model the number of infections per month:

1. Events occur one at a time; two or more events cannot occur exactly at the same time and location;
2. The occurrence of an event in a given period is independent of the occurrence of an event in a non-overlapping period;
3. The expected number of events during any period is constant.

b).



We may use R code to get the same value.

```
prob = dpois(2, 7)
prob
```

```
## [1] 0.02234111
```

The probability that in the next month the hospital's patients will have exactly 2 unexplained infections is: **0.0223411** *or 2.23%.*

Problem 2

Formula: Bayes Theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_i)P(B_1) + P(A|B_2)P(B_2) + \dots P(A|B_k)P(B_k)}$$

D = developed CHD

C = Initial serum cholesterol levels above 200

$P(D) = 0.25$, $P(D^c) = 1 - 0.25 = 0.75$

$P(C|D) = 0.60$

$P(C|D^c) = 0.16$

Interested in:

$$P(D^c|C^c) = \frac{P(D^c \cap C^c)}{P(C^c)} = \frac{P(C^c|D^c)P(D^c)}{P(C^c|D^c)P(D^c) + P(C^c|D)P(D)}$$

$P(C^c|D^c) = 1 - P(C|D^c) = 0.84$

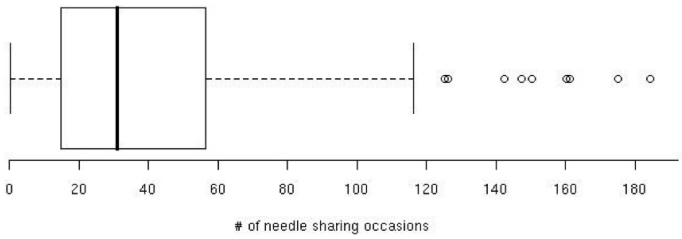
$P(C^c|D) = 1 - P(C|D) = 0.40$

$$\frac{(.84)(.75)}{(.84)(.75) + (.40)(.25)} = \frac{.63}{.63 + .1} = 0.8630 \approx 0.86$$

The probability that a random chosen subject will not develop CHD, given that he had an initial serum cholesterol level below or equal 200 is: **0.8630** *or 86%.*

Problem 3

Problem 3:



Researchers from New York, studying needle sharing behavior among intravenous drug users, randomly selected 500 patients from the city's two major drug detoxification clinics. At admission, each participant reported the number of their needle sharing occasions in the last year. The researchers summarized the observed data for 'number of needle sharing occasions in the last year' in the box-plot above.

a) **[Circle only one correct answer.]** Based on the box-plot of the data, the proportion of subjects whose number of needle sharing occasions is less than the sample average number of needle sharing occasions is: *(4 points)*

i. Less than 50%
ii. More than 50%
iii. Exactly 50%
iv. Exactly 100%

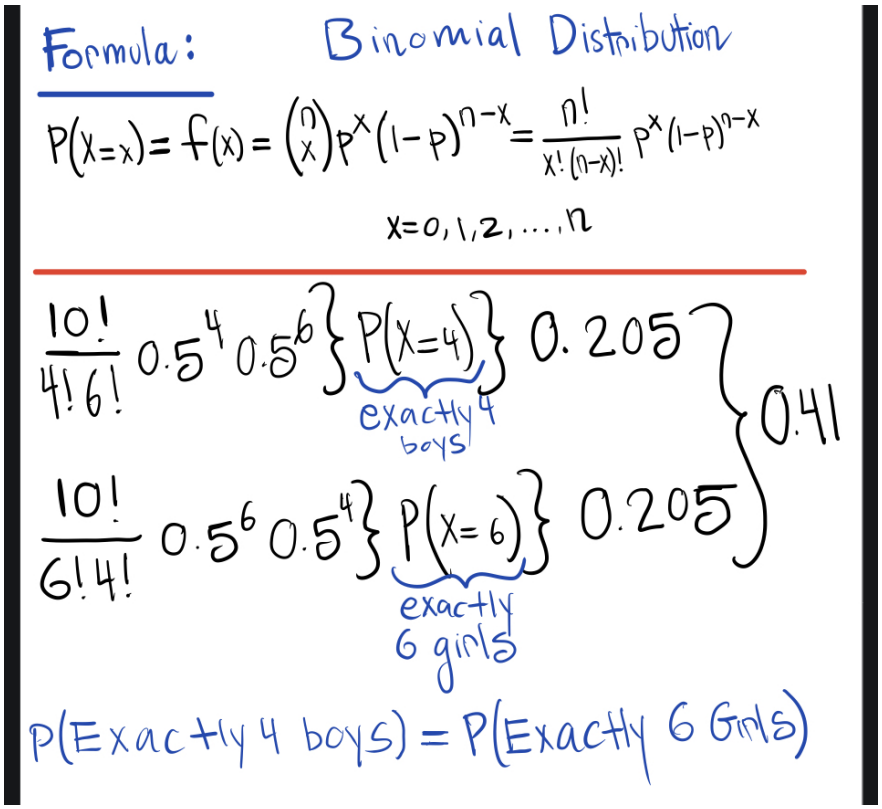
b) **[Circle only one correct answer.]** What is the probability that the number of needle sharing occasions for a randomly selected patient from the study is smaller than the third quartile (Q3) and larger than the first quartile (Q1)? *(4 points)*

i. About 25%
ii. About 50%
iii. About 75%
iv. About 100%

c) **[Circle only one correct answer.]** In order to report the number of needle sharing occasions per WEEK, the researchers divided all the observed data by 52 and computed new sample statistics. What is the relationship between the sample statistics for ONE WEEK and the sample statistics for ONE YEAR? *(4 points)*

i. The 'one week' mean, median and standard deviation are 52 times larger than the 'yearly' mean, median and standard deviation
ii. The 'one week' mean, median and standard deviation are 52 times smaller than the 'yearly' mean, median and standard deviation
iii. The 'one week' mean and median are 52 times larger than the 'yearly' mean and median. The standard deviation will stay the same.
iv. The 'one week' mean and median are 52 times smaller than the 'yearly' mean and median. The standard deviation will stay the same.

Problem 4

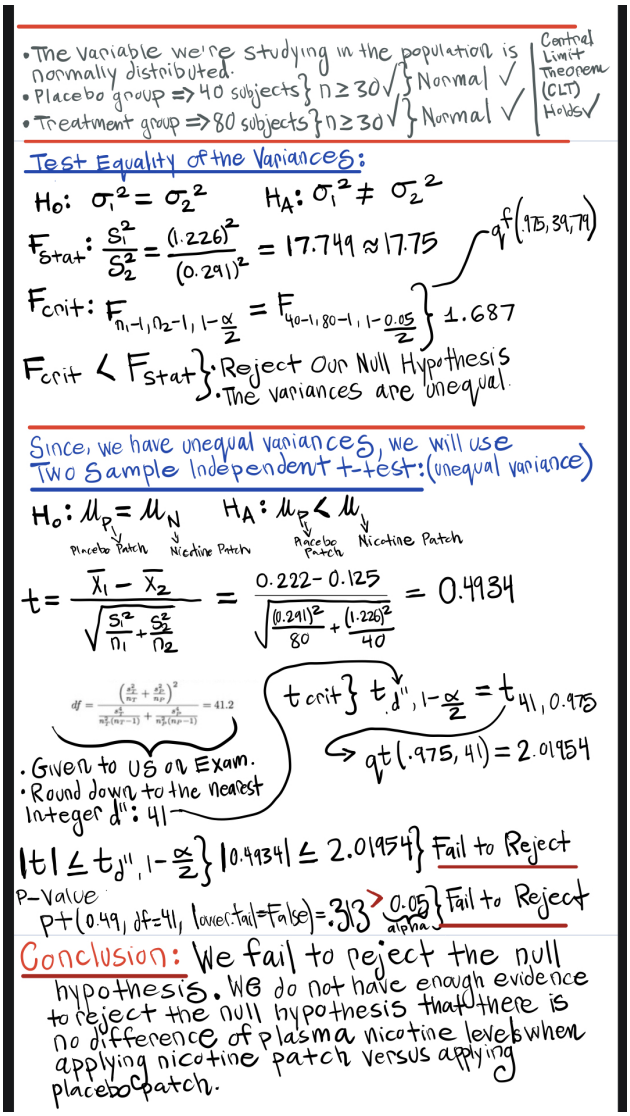


We also may use R code to get the value:

```
binom_data = dbinom(4, 10, 0.5) + dbinom(6, 10, 0.5)
```

A couple has 10 children and the probability of the event of having exactly 4 boys or exactly 6 boys or exactly 6 girls is: **0.4101562** *or 41%.*

Problem 5a



Problem 5b

b) **[Circle only one correct answer.]** What is the proper definition of the p-value in the above hypothesis test? *(4 points)*

i. The probability of rejecting the null hypothesis, assuming that the alternative hypothesis is true.

ii. The probability of rejecting the null hypothesis, assuming that the null hypothesis is true.

iii. The probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming that the alternative hypothesis is true.

iv. The probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming that the null hypothesis is true.

Problem 6

$n \geq 30$ ✓
↳ $n = 40$ soldiers } Since n is 40 and is greater than 30, the sampling distribution of the mean is normal distribution ✓

Central Limit Theorem Holds ✓

$n = 40$

$\bar{x} = 45$

$\sigma^2 = 20$ } Variance

alpha (α) $= 0.05$

Formula: $\bar{x} \pm t_{n-1,\,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$

$$45 \pm t_{45-1,\,1-\frac{0.05}{2}} \frac{\sqrt{20}}{\sqrt{40}}$$

$$45 \pm t_{44,\,0.975} \frac{\sqrt{20}}{\sqrt{40}}$$

$qt(.975, 44)$

2.0153

$$45 \pm 2.0153 \frac{\sqrt{20}}{\sqrt{40}}$$

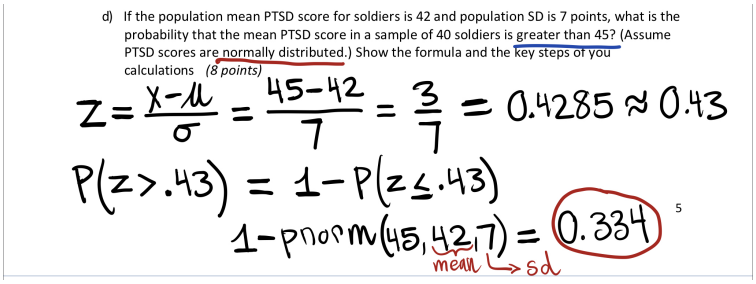
$(43.575, 46.425)$

Interpretation: We are 95% confident that the mean Post traumatic stress disorder (PTSD) score of the population is between 43.58 and 46.43.

b) **[True/False]**
If the number of soldiers in part a) increases, and all other quantities stay the same, the confidence interval will be more narrow. *(4 points)*
i. True *ii. False*

c) **[True/False]**
If we created a 90% confidence interval for part a), it would be wider than the 95% confidence interval. *(4 points)*
i. True *ii. False*

Problem 6d



For problem 6d, we may use Z-table, or R code.

```
ptsd_prob = 1-pnorm(45, mean = 42, sd = 7)
```

Interpretation: *The probability that the mean PTSD score in a sample of 40 soldiers is greater than 45 is 0.3341176 or 33.4%.*

Problem 7a

a) Fill in the ANOVA table. *(5 points)*

Analysis of Variance Table
Response: Effectiveness of the vaccine

	Sum Sq	DF	Mean Sq	F value
Vaccine	**140**	3 ← 4−1	46.67 = 140/3	46.67/17.39 = 2.68
Error	1600 ← 1740 − 140	92 ← 96−4	17.39 = 1600/92	
Total	**1740**	95 ← 96−1		

k−1 (DF of Vaccine); n−k (DF of Error); n−1 (DF of Total)

Problem 7b

One-Way ANOVA: Overall F-Test

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: at least two means are not equal

Test statistics:

$$F = \frac{\text{Between SS}/(k-1)}{\text{Within SS}/(n-k)} \sim F_{k-1,\,n-k}$$ distribution under H_0

$$\frac{46.67}{17.39} = 2.68$$

F-Critical:

$$F_{k-1,\,n-k,\,1-\alpha} \Big\} F_{4-1,\,96-4,\,1-0.05}$$

$$F_{3,\,92,\,0.95}$$

qf(0.95, 3, 92) = 2.7036

F-Value ≤ F-Critical : Fail to Reject H_0

2.68 ≤ 2.70 : Fail to Reject H_0

Conclusion: At 0.05 significance level, we fail to reject the Null hypothesis and conclude the four vaccines have same effectiveness. There is no significant differences among the population means for the four vaccines.

Problem 8

For problem 8, I'll be using a ***Chi-Squared: Test of Independence****. I chose to use this statistical test because the problem states we need to determine if there is association between the treatment group and dropout rates and we may use Chi-Squared: Test of Independence to test for independence or association of the rows and column variables. For this problem, I am interested if one variable's value provides any information about the value of the other variable, i.e. are the two variables independent or dependent/associated? Also the problem indicates to create "table of expected values" (contingency table) and we have learned these tables are created for Chi-Square test.*