

Biostatistics I - Midterm (Fall 2021)

Tanvir Khan (UNI - 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).

- b) Suppose the number of infections per month follows a Poisson distribution. What is the probability that in the next month the hospital's patients will have exactly 2 unexplained infections? Include the formula and all the key steps in your calculations. (5 points)

Formula: $P(x=x) = f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x=0,1,2,\dots,n$

$\lambda = 7$, rate of unexplained infection among patients per month

Calculate: $P(x=2) = \frac{7^2 e^{-7}}{2!} = 0.0223$

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

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

D = developed CHD

C = initial serum cholesterol levels above 200

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

$$P(C|D) = 0.60$$

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

Interested in:

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

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

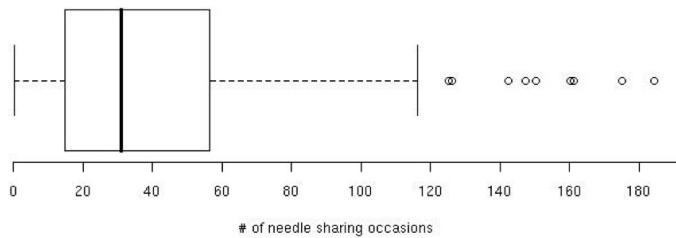
$$P(C^c|D) = 1 - P(C|D) = 0.40$$

$$\frac{(0.84)(0.75)}{(0.84)(0.75) + (0.40)(0.25)} = \frac{0.63}{0.63 + 0.1} = 0.8630 \approx 0.86$$

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

Formula: Binomial Distribution

$$P(X=x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$x=0, 1, 2, \dots, n$

$$\left. \begin{aligned} & \frac{10!}{4!6!} 0.5^4 0.5^6 \} P(X=4) \\ & \frac{10!}{6!4!} 0.5^6 0.5^4 \} P(X=6) \end{aligned} \right\} 0.205 \quad \left. \begin{aligned} & \text{exactly 4 boys} \\ & \text{exactly 6 girls} \end{aligned} \right\} 0.41$$

$$P(\text{Exactly 4 boys}) = P(\text{Exactly 6 Girls})$$

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 girls is: **0.4101562** or 41%.

Problem 5a

- The Variable we're studying in the population is normally distributed.
 - Placebo group $\Rightarrow 40$ subjects $n \geq 30 \checkmark$ Normal ✓
 - Treatment group $\Rightarrow 80$ subjects $n \geq 30 \checkmark$ Normal ✓
- Central Limit Theorem (CLT) Holds ✓

Test Equality of the Variances:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_A: \sigma_1^2 \neq \sigma_2^2$$

$$F_{\text{stat}}: \frac{s_1^2}{s_2^2} = \frac{(1.226)^2}{(0.291)^2} = 17.749 \approx 17.75$$

$$F_{\text{crit}}: F_{n_1-1, n_2-1, 1-\alpha/2} = F_{40-1, 80-1, 1-\frac{0.05}{2}} \approx 1.687$$

$F_{\text{crit}} < F_{\text{stat}}$ Reject Our Null Hypothesis
The variances are unequal.

Since we have unequal variances, we will use Two Sample Independent t-test: (unequal variance)

$$H_0: \mu_p = \mu_n \quad H_A: \mu_p < \mu_n$$

Placebo Patch Nicotine Patch Placebo Patch Nicotine Patch

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{0.222 - 0.125}{\sqrt{\frac{(0.291)^2}{80} + \frac{(1.226)^2}{40}}} = 0.4934$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^2}{n_1(n_1-1)} + \frac{s_2^2}{n_2(n_2-1)}} = 41.2$$

- Given to us on Exam.
- Round down to the nearest integer $d.f. = 41$

$$t_{\text{crit}} \} t_{d.f., 1-\alpha/2} = t_{41, 0.975}$$

$$\rightarrow qt(.975, 41) = 2.01954$$

$|t| \leq t_{d.f., 1-\alpha/2} \} |0.4934| \leq 2.01954 \} \text{Fail to Reject}$

P-value $p(t > 0.49, df=41, \text{lower.tail=False}) = 0.313 > 0.05 \} \text{Fail to Reject}$

Conclusion: We fail to reject the null hypothesis. We do not have enough evidence to reject the null hypothesis that there is no difference of plasma nicotine levels when applying nicotine patch versus applying placebo patch.

Problem 5b

- b) [Circle only one correct answer.] What is the proper definition of the p-value in the above hypothesis test? (4 points)
- The probability of rejecting the null hypothesis, assuming that the alternative hypothesis is true.
 - The probability of rejecting the null hypothesis, assuming that the null hypothesis is true.
 - The probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming that the alternative hypothesis is true.
 - The probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming that the null hypothesis is true.

Problem 6a

$n \geq 30 \checkmark$ } Since n is 40 and is greater than
 $\hookrightarrow n = 40$ soldiers } 30, the sampling distribution of
 the mean is normal distribution ✓
 Central Limit Theorem Holds ✓

$$n = 40$$

$$\bar{x} = 45$$

$$\sigma^2 = 20 \} \text{Variance}$$

$$\alpha \text{ or } (\alpha) = 0.05$$

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

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

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

$$qt(0.975, 44)$$

$$45 \pm 2.02 \frac{\sqrt{20}}{\sqrt{40}}$$

$$(43.572, 46.428)$$

Interpretation: We are 95% confident that the mean Post traumatic stress disorder (PTSD) score of the population is between 43.572 and 46.428.

Problem 6b and 6c

- 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

- d) If the population mean PTSD score for soldiers is 42 and population SD is 7 points, what is the probability that the mean PTSD score in a sample of 40 soldiers is greater than 45? (Assume PTSD scores are normally distributed.) Show the formula and the key steps of your calculations. (8 points)

$$Z = \frac{X - \mu}{\sigma} = \frac{45 - 42}{7} = \frac{3}{7} = 0.4285 \approx 0.43$$

$$P(z > .43) = 1 - P(z \leq .43)$$

$$1 - pnorm(45, 42, 7) = 0.334$$

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 $\leftarrow 4-1$	46.67 $= \frac{140}{3}$	$\frac{46.67}{17.39} = 2.68$
Error	1600 $\leftarrow 1740 - 140$	92 $\leftarrow 96-4$	17.39 $= \frac{1600}{92}$	
Total	1740	95 $\leftarrow 96-1$		

$$\begin{array}{l} k-1 \\ \downarrow \\ 3 \end{array}$$

$$\begin{array}{l} n-k \\ \uparrow \\ 95 \end{array}$$

$$\begin{array}{l} n-1 \\ \uparrow \\ 96-1 \end{array}$$

Problem 7b

One-Way ANOVA: Overall F-Test

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

$$H_1: \text{at least two means are not equal}$$

Test statistics:

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

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

F-Critical:

$$\begin{aligned} F_{k-1, n-k, 1-\alpha} &\} F_{4-1, 96-4, 1-0.05} \\ &\} F_{3, 92, 0.95} \\ &\} q_f(0.95, 3, 92) \\ &= 2.7036 \end{aligned}$$

$F\text{-value} \leq F\text{-Critical}$: Fail to Reject H_0

$2.68 \leq 2.70$: 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.

		Drop Out	Non-Drop Out	Expected	Total
Treatment A	15	$\frac{37.50}{180} = 0.208$	$\frac{143.50}{180} = 34.72$	35	50
Treatment B	10	$\frac{37.70}{180} = 0.208$	$\frac{143.70}{180} = 55.61$	60	70
Placebo	12	$\frac{37.60}{180} = 0.208$	$\frac{143.60}{180} = 47.67$	48	60
Total		37	143		180

H_0 : Treatment groups and rates of dropout are independent ($P_1 = P_2 = P_3$)

H_1 : Treatment groups and rates of dropout are associated/dependent.

$$\chi^2 = \sum_{i=1}^3 \sum_{j=1}^2 = \frac{(15-0.208)^2}{0.208} + \frac{(35-34.72)^2}{34.72} + \frac{(10-0.208)^2}{0.208}$$

$$\rightarrow \frac{(60-55.61)^2}{55.61} + \frac{(12-0.208)^2}{0.208} + \frac{(48-47.67)^2}{47.67}$$

$$= 2.167 + 0.5608 + 1.339 + 0.3465 + 0.0088 + 0.00228$$

$$= 4.424 \approx 4.42$$

Under the null hypothesis: $\chi^2 \sim \chi^2_2$.

$$\chi^2_{(R-1) \times (C-1)} \quad \chi^2_{(B-1) \times (Z-1)} \quad \chi^2_{(2)(1)} \quad \chi^2_{2, 0.95}$$

qchisq(0.95, 2)
5.99

Decision Rule: Fail to Reject H_0 .

Interpretation: At 0.05 significance level, $\chi^2 < \chi^2_{2, 0.95} = 5.99$, we fail to reject the null hypothesis and conclude that there is evidence that Treatment groups and rates of dropout are independent.