Practice Test for Exam 1

Stats I

Barboza-Salerno

1. Binomial problem from class.
2. (probability; 21 points). Suppose you are a social worker interested in learning more about bias in the child welfare system. You know that children can be substantiated or not for child abuse and neglect. Suppose 40% of your clients are substantiated for child maltreatment and that 90% of the children substantiated are from low-income neighborhoods, while children who are not substantiated are not from low-income neighborhoods 80% of the time. Hint: Draw a tree diagram and carefully apply the formulas we used in class. Follow this schema:

Let = the probability that your client is substantiated and = the probability of being from a low-income neighborhood. Before answering the following questions, write out the probability of NOT being substantiated, the probability of NOT being in a low-income neighborhood, and the formulas for the intersection and union of two events, S and LI. Then draw a tree diagram with the first branch being the probability of being substantiated and not substantiated and then the second branch being the probabilities that substantiated and not substantiated children are from low-income families. Now, answer the following questions:

(a) Find the probability that a randomly selected customer is substantiated and is from a low-income neighborhood.

(b) Find the probability that a randomly selected customer is from a low-income neighborhood.

(c) Find the probability that a randomly selected customer is substantiated or from a low-income neighborhood.

(d) Find the probability that a randomly selected client is not from a low-income neighborhood.

(e) Find the probability that a randomly selected client is not substantiated and not from a low-income neighborhood.

(f) Find the conditional probability that a client is substantiated given that the client is from a low-income neighborhood.

(g) Find the conditional probability that a client is not substantiated given that the client is not from a low-income neighborhood.

1. (z-scores; probability; 20 points). The CBCL score is a checklist given by practitioners or researchers to measure behavioral and emotional problems among children. Let *X* be the Child Behavior Checklist (CBCL) score of a randomly selected child and assume . Find the following probabilities using the formula and then use R to check your results. *Hint:* draw the picture for each one.
2. Find the 40th percentile of *X*, that is the CBCL score that X is less than or equal to with probability = .40. Then interpret what that means substantively.
3. Find the median of X, i.e., the CBCL score that X is less than or equal to with probability = .50.

4. (hypothesis testing, sampling distributions; 20 points): 100 OSU PhD social work students were asked to name their favorite course. The distribution of responses is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subj | Stats I | Stats II | Research Methods | Social Justice in Social Work | Social Work Theories |
| Freq | 10 | 15 | 50 | 20 | 5 |

The administration wishes to test two different hypotheses.

1. The first hypothesis is that the distribution of responses is independent.
2. The second hypothesis is about the faculty’s impression about how popular these courses are. The faculty expected the following distribution, and what to test their impression.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subj | Stats I | Stats II | Research Methods | Social Justice in Social Work | Social Work Theories |
| Freq | 10 | 10 | 56 | 14 | 10 |

State the sampling distribution, any assumptions, and compute the test statistic at the level.

5. (probability/odds/chi-square; 20 points): Suppose you run an analysis of data that queried *N* = 43 people about the type of county they live in and the number of encounters they had with the police. You ran a contingency table and received the following output. Answer the following questions:

|  |  |  |
| --- | --- | --- |
|  | No contact | Contact |
| Rural | 6 | 9 |
| Urban | 6 | 22 |

1. Suppose you want to select one person from among the total number of people you asked, what is the probability that the person lives in an urban county?
2. What is the probability that the person you selected lives in an urban setting and had police contact?
3. Given that the person you selected had police contact, what is the probability that the person lives in an urban setting?
4. Are the events that the person selected had police contact and that the person selected lives in an urban setting independent?
5. Test the hypothesis that the two events are independent. Under the null hypothesis that having police contact and living in an urban setting are independent, the following expected counts were computed. Calculate thetest statistics and compare it to the critical value of 3.81 for 1 degree of freedom. What do you conclude?

|  |  |  |
| --- | --- | --- |
|  | No contact | Contact |
| Rural | 4.186 | 10.814 |
| Urban | 7.814 | 20.186 |

Now, we would like to make some relative comparisons so we further understand how much more likely someone from an urban or rural county is more or less likely to have police contact. Compute the following statistics and draw a conclusion:

1. The odds of police contact for someone living in urban versus rural county?
2. The relative risk of police contact for someone living in a urban versus rural county?
3. The odds ratio of police contact for people living in rural versus urban counties. Compare the odds ratio of police contact for people living in urban versus rural. Are those the same? What is the relationship between them?

6) (z-scores; probability; Use intuition along w/R to check answers; 20 points): Suppose a population has values that are Normally distributed. *Hint:* recall that the functions pnorm and qnorm compute the probability (i.e., area) under the curve for a given Z-score and the Z-score for a probability under the curve respectively. Also note that the probability and Z-score returned is P(Z < z), i.e., the area under the curve to the left of the Z-score. Then:

1. How many standard deviations should you go above and below the mean to in order to capture 1/3 of the values?
2. The top 10% of the values are all at least how many standard deviations above the mean?
3. What proportion of the values lie within 3 standard deviations of the mean?
4. What is the distance between the median and the 75th percentile in terms of the number of standard deviations?
5. At what percentile is a value that is 2 standard deviations above the mean?
6. The bottom 5% of values lie at least how many standard deviations below the mean?

7. (working with contingency; probability)

Consider a box of fruits, 19 are good, 25 bad, and 6 are rotten. Moreover, suppose 9 of the good fruits are oranges, 5 of the bad fruits are oranges, and 3 of the rotten fruits are oranges, and the rest of the marbles are tangerines. Suppose you draw one fruit from the bag and let G, B, and R be the events that the fruit is good, bad, and rotten, respectively, and let O be the event that it is orange.

1. Are the events R and O independent?
2. (challenge): provide a chi-square analysis to test your hypothesis that R and O are independent.

Conceptual problems (T/F)

1. Independence of two events implies that they are not dependent.
2. You can have two mutually exclusive events that are dependent (challenge: use the rules of probability to demonstrate your answer).
3. A larger sample size increases the variance of the sample mean (Hint: what is the formula?)

Applied Work (100 points)

Use the National Victimization Survey data to answer the following questions.

You are interested in the types of places where people are victimized and whether different demographic groups are more or less likely to be victimized in specific places. One place where the research suggests people are highly likely to be victimized is in public housing. Of course, I suggest you think of the reasons for why this may be the case before doing such a study. In any event, we will first focus on testing the hypothesis that people who live in public housing are more likely to be polyvictimized compared to those who do not. Follow these steps

1. Download the National Crime Victimization Survey located here
2. Filter the data for year = 2020.
3. Identify the following variables in the dataset: household income, public housing, urbanicity, sex (original), and type of victimization. Rename and save these variables as a subset of the original data called NCVS\_subset.
4. Rename the variables as follows: urban, hhincome, pubhouse, sex, victim
5. Run the descriptive statistics on all of the variables. Identify any issues with the variable coding, address those issues, and re-run the descriptive statistics.

At the end of that exercise, you should have something that looks like this

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **URBANICITY (START 2020 Q1)** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Urban | 25121 | 9.3 | 9.3 | 9.3 |
| Suburban | 186620 | 69.0 | 69.0 | 78.3 |
| Rural | 58825 | 21.7 | 21.7 | 100.0 |
| Total | 270566 | 100.0 | 100.0 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **HOUSEHOLD INCOME** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Less than $5,000 | 3271 | 1.2 | 1.6 | 1.6 |
| $5,000-$7,499 | 1228 | .5 | .6 | 2.2 |
| $7,500-$9,999 | 1939 | .7 | 1.0 | 3.2 |
| $10,000-$12,499 | 2896 | 1.1 | 1.4 | 4.6 |
| $12,500-$14,999 | 2816 | 1.0 | 1.4 | 6.0 |
| $15,000-$17,499 | 2775 | 1.0 | 1.4 | 7.4 |
| $17,500-$19,999 | 3014 | 1.1 | 1.5 | 8.9 |
| $20,000-$24,999 | 7534 | 2.8 | 3.7 | 12.6 |
| $25,000-$29,999 | 7366 | 2.7 | 3.6 | 16.3 |
| $30,000-$34,999 | 9307 | 3.4 | 4.6 | 20.9 |
| $35,000-$39,999 | 10540 | 3.9 | 5.2 | 26.1 |
| $40,000-$49,999 | 17479 | 6.5 | 8.7 | 34.7 |
| $50,000-$74,999 | 35858 | 13.3 | 17.7 | 52.5 |
| $75,000-$99,999 | 29744 | 11.0 | 14.7 | 67.2 |
| $100,000-$149,999 | 33515 | 12.4 | 16.6 | 83.8 |
| $150,000-$199,999 | 15956 | 5.9 | 7.9 | 91.7 |
| $200,000 or more | 16814 | 6.2 | 8.3 | 100.0 |
| Total | 202052 | 74.7 | 100.0 |  |
| Missing | Residue | 68514 | 25.3 |  |  |
| Total | | 270566 | 100.0 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PUBLIC HOUSING** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Yes (public housing) | 4366 | 1.6 | 5.9 | 5.9 |
| No (not public housing) | 69661 | 25.7 | 94.1 | 100.0 |
| Total | 74027 | 27.4 | 100.0 |  |
| Missing | Residue | 119 | .0 |  |  |
| Out of universe | 196420 | 72.6 |  |  |
| Total | 196539 | 72.6 |  |  |
| Total | | 270566 | 100.0 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SEX (ORIGINAL)** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Male | 130231 | 48.1 | 48.2 | 48.2 |
| Female | 140219 | 51.8 | 51.8 | 100.0 |
| Total | 270450 | 100.0 | 100.0 |  |
| Missing | Residue | 98 | .0 |  |  |
| Out of universe | 18 | .0 |  |  |
| Total | 116 | .0 |  |  |
| Total | | 270566 | 100.0 |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TYPE OF VICTIMIZATION** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Violent crime | 1091 | .4 | 16.8 | 16.8 |
| Property crime | 5148 | 1.9 | 79.3 | 96.1 |
| Both V and P | 254 | .1 | 3.9 | 100.0 |
| Total | 6493 | 2.4 | 100.0 |  |
| Missing | No incident | 264073 | 97.6 |  |  |
| Total | | 270566 | 100.0 |  |  |

Now, test the hypothesis that public housing and type of victimization are independent. Then do the same for urbanicity and type of victimization. Go to analyze → descriptives → crosstabs and enter the variables as follows

A screenshot of a computer

Description automatically generated

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | | |
|  | | | TYPE OF VICTIMIZATION | | | Total |
| Violent crime | Property crime | Both V and P |
| URBANICITY (START 2020 Q1) | Urban | Count | 155 | 947 | 44 | 1146 |
| % within TYPE OF VICTIMIZATION | 14.2% | 18.4% | 17.3% | 17.6% |
| Suburban | Count | 744 | 3515 | 175 | 4434 |
| % within TYPE OF VICTIMIZATION | 68.2% | 68.3% | 68.9% | 68.3% |
| Rural | Count | 192 | 686 | 35 | 913 |
| % within TYPE OF VICTIMIZATION | 17.6% | 13.3% | 13.8% | 14.1% |
| Total | | Count | 1091 | 5148 | 254 | 6493 |
| % within TYPE OF VICTIMIZATION | 100.0% | 100.0% | 100.0% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 20.683a | 4 | <.001 |
| Likelihood Ratio | 20.500 | 4 | <.001 |
| Linear-by-Linear Association | 14.795 | 1 | <.001 |
| N of Valid Cases | 6493 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 35.72. | | | |

**PUBLIC HOUSING \* TYPE OF VICTIMIZATION**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | | |
|  | | | TYPE OF VICTIMIZATION | | | Total |
| Violent crime | Property crime | Both V and P |
| PUBLIC HOUSING | Yes (public housing) | Count | 57 | 145 | 20 | 222 |
| % within TYPE OF VICTIMIZATION | 11.1% | 6.7% | 13.2% | 7.8% |
| No (not public housing) | Count | 455 | 2033 | 132 | 2620 |
| % within TYPE OF VICTIMIZATION | 88.9% | 93.3% | 86.8% | 92.2% |
| Total | | Count | 512 | 2178 | 152 | 2842 |
| % within TYPE OF VICTIMIZATION | 100.0% | 100.0% | 100.0% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymptotic Significance (2-sided) |
| Pearson Chi-Square | 17.904a | 2 | <.001 |
| Likelihood Ratio | 16.309 | 2 | <.001 |
| Linear-by-Linear Association | 1.770 | 1 | .183 |
| N of Valid Cases | 2842 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 11.87. | | | |

Write a paragraph summarizing these results.

Compute the following:

* Odds ratio coding the effect of being polyvictimized versus not for individuals living in public housing
* Risk ratio for polyvictimized for those living in public housing and those not living in public housing

*Note*: this requires condensing the polyvictimization data into 0 = polyvictim and 1 = not polyvictim. This is called recoding a variable. After recoding, be sure to properly label the variables. The odds ratio is below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PH \* polyvictim Crosstabulation** | | | | | |
|  | | | polyvictim | | Total |
| Polyvictim | Not Polyvictim |
| PH | Public Housing | Count | 20 | 202 | 222 |
| % within PH | 9.0% | 91.0% | 100.0% |
| Not Public Housing | Count | 132 | 2488 | 2620 |
| % within PH | 5.0% | 95.0% | 100.0% |
| Total | | Count | 152 | 2690 | 2842 |
| % within PH | 5.3% | 94.7% | 100.0% |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Estimate** | | | |
|  | Value | 95% Confidence Interval | |
| Lower | Upper |
| Odds Ratio for PH (Public Housing / Not Public Housing) | 1.866 | 1.141 | 3.051 |
| For cohort polyvictim = Polyvictim | 1.788 | 1.140 | 2.804 |
| For cohort polyvictim = Not Polyvictim | .958 | .918 | 1.000 |
| N of Valid Cases | 2842 |  |  |

The odds of being a polyvictim among those in public housing are 1.866 times higher than for those living in public housing compared to those who do not. As for risk, polyvictims are 1.788 times as likely to live in public housing versus not live in public housing and those who are not polyvictims are .958 times as likely to live in public housing compared to not living in public housing. Challenge: the 95% confidence interval suggests possible ranges of the variable in the population. You may guess that the 95% interval assumes we are wrong 5% of the time, or that alpha = .05. I know we have not covered this yet, but given the confidence interval noted in the output, do you suppose the odds ratio and risk estimates are statistically significant (i.e., unlikely to occur by chance)? Why or why not?

Now, it is your turn.

Use the variable “ATTACK, THREAT, THEFT: LOCATION CUES” which is measured as follows

A screenshot of a computer

Description automatically generated

Also use the variable “HOW LONG AT THIS ADDRESS (YEARS)” which is measured as

A screenshot of a computer

Description automatically generated

Recode the variable years lived at address as follows

1-5 years = 0

6-11 years = 1

All other values is system missing

Test the hypothesis that individuals who lived less than 6 years at their address are more likely to be attacked compared to those who lived at their address for 6 or more years. Compute the chi-square, odds and risk ratio, and relevant descriptives and then write the results.