Rates and Proportions - Homework 1

Matt Isaac

September 8, 2017

1.2 Which scale of measurement is most appropriate for the following variables - nominal, or ordinal?

a. Political party affiliation (Democrat, Republican, unaffiliated) - nominal.

b. Highest degree obtained (none, high school, bachelor's, master's, doctorate) - ordinal.
c. Patient condition (good, fair, serious, critical) - ordinal.
d. Hospital location (London, Boston, Madison, Rochester, Toronto) - nominal.
e. Favorite beverage (beer, juice, milk, soft drink, wine, other) - nominal.
f. How often feel depressed (never, occasionally, often, always) - ordinal.

1.3

Each of 100 muiltiple-choice questions on an exam has four possible answers, but one correct response. For each question, a student randomly selects one response as an answer.

a. Specify the distribution of the student's number of correct answers on the exam:

Let X count the number of correct answers on the exam. Then $X \sim Binomial(100, 0.25)$.

b. Based on the mean and standard deviation of that distribution, would it be surprising if the student made at least 50 correct responses? Explain.

We can approximate the binomial distribution with a normal distribution, using the mean and variance as the parameters of the normal distribution.

$E(X) = np = 100 * 0.25 = 25$ $Var(X) = np(1-p) = 100 * 0.25(1-0.25) = 18.75$

So, $X \sim N(25, 18.75)$. To answer this question, I will find the probability of the student getting a score of fifty or greater.

$P(X \geq 50) = 1 - P(X \leq 50) = 1 - P(Z \leq \frac{50-25}{\sqrt{18.75}} = 1 - P(Z \leq 5.773) = 1 - 1 = 0.$

There is essentially a 0% chance that this student will score 50 or higher on the exam. I would be shocked and amazed if this ever occurred.

1.6

Genotypes AA, Aa, and aa occur with probabilities $(p1, p2, p3)$. For $n = 3$ independent observations the observed frequencies are $(n1, n2, n3)$.

a. Explain how you can determine n3 from knowing n1 and n2. Thus, the multinomial distribution of $(n1, n2, n3)$ is actually two-dimensional.

As far as we've been told, AA, Aa, and aa are the only three possible genotypes. In other words, everyone will fall into one of those three categories. Thus, $p1 + p2 + p3 = 1$. If $p1$ and $p2$ are fixed, $p3$ s also fixed to equal $1 - (p1 + p2)$.

b. Show the set of all possible observations, $(n1, n2, n3)$ with $n = 3$.

$(3,0,0), (0,3,0), (0,0,3), (2,1,0), (2,0,1), (1,2,0), (0,2,1), (1,0,2), (0,1,2), (1,1,1)$

c. Suppose $(\pi_1, \pi_2, \pi_3) = (0.25, 0.50, 0.25)$. Find the multinomial probability that $(n1, n2, n3) = (1, 2, 0)$.

$P(1,2,0) = (\frac{3!}{1!2!0!})0.25^1 0.50^2 0.25^0 = .1875$

d. Refer to (c). What probability distribution does $n1$ alone have? Specify the values of the sample size index and parameter for that distribution.

$n1$ will have a binomial distribution. Each observation will either be $n1$ ('success') or not $n1$ ('failure'). Each trial is independent and identical.

Thus, $n1 \sim Binomial(0.25, 3)$.

1.7

a. Find the probabililty of playing Russian Rouletete six times and never having the bullet fire.

P(Playing 6 times with no bullet fired) $= \frac{5}{6}^6 = .3349$

b. Suppose one kept playing this game until the bullet fires. Let Y denote the number of the game on which the bullet fires. Argue that the probability of the outcome y equals $\frac{5}{6}^{y-1} * \frac{1}{6}$ for $y = 1, 2, 3, ..$

Playing 6 times, we had $\frac{5}{6}^6$ for probability of the bullet never firing. Let us assume that the bullet fires on the seventh game. The probability of this happening would be $\frac{5}{6}^6 * \frac{1}{6}$. We multiply the probabilities of each sub event together to find the probability of the overall event happening. In the example just given, $y = 7$. So $\frac{5}{6}^6 * \frac{1}{6} = \frac{5}{6}^{y-1} * \frac{1}{6}$. This will hold for any value of y, where y = 1, 2, 3,.

1.8 When the 2000 General Social Survey asked subjects whether they would be willing to accept cuts in their standard of living to protect the environment, 344 of 1170 subjects said "yes."

a. Estimate the population proportion who would say "yes."

An unbiased estimate for p is $\hat{p} = \frac{344}{1170} = 0.29$.

b. Conduct a significance test to determine whether a majority or minority of the population would say "yes." Report and interpret the p-value.

$H_0 : p \geq 0.5$ vs. $H_A : p < 0.5$.

$z = \frac{0.29 - 0.5}{\sqrt{\frac{0.29(1-0.29)}{1170}}} = -15.83$ $P(Z < -15.83) = 0$

The P-value \$ = 0 < 0.05\$. There is no chance that we would have gotten the value of $\hat{p}$ that we did if p is greater than or equal to 0.5. If we surveyed the entire population, a minority would say 'yes'.

c. Construct and interpret a 99% confidence interval for the population proportion who would say 'yes'.

A 99% confidence interval for p is $0.29 \pm z_{.005}\sqrt{\frac{0.29(1-0.29)}{1170}}$ The interval is $(0.256, 0.324)$. We are 99% confident that this interval covers the true proportion of the population that would say 'yes'.

2.17.

a. Find the P-value for testing that the incidence of heart attacks is independent of aspirin intake using χ^2. Interpret results.

Table of Expected Counts:

Group	Yes	No	Total
Placebo	146.48	10887.52	11034
Aspirin	146.52	10890.48	11037

$\chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$

$\chi^2 = \frac{(189-146.48)^2}{146.48} + \frac{(104-146.52)^2}{146.52} + \frac{(10845-10887.52)^2}{10887.52} + \frac{(10933-10890.48)^2}{10890.48} = 25.014$

With one degree of freedom, and $\chi^2 = 25.014$, the P-value is \$ <0.001\$. We fail to reject our null hypothesis that occurrance of Myocardial Infarction and aspirin use are independent events.

b. Find the P-value for testing that the incidence of heart attacks is independent of aspirin intake using G^2. Interpret results.

$G^2 = 2\sum n_{ij} \log \frac{n_{ij}}{\mu_{ij}}$
$G^2 = 2[189 \log(\frac{189}{146.48}) + 104 \log(\frac{104}{146.52}) + 10845 \log(\frac{10845}{10887.52}) + 10933 \log(\frac{101933}{10890.48})] = 25.372$

Again, with one degree of freedom, and $G^2 = 25.372$, the P-value is < 0.001. We fail to reject the null hypothesis that occurrance of Myocarial Infarction and aspirin use are independent events.

2.18.

a. Show how to obtain the estimated expected cell count of 35.8 for the first cell.

$n = 21 + 159 + 110 + 53 + 372 + 221 + 94 + 249 + 83 = 1362$ row 1 total $= 21 + 159 + 110 = 290$ column 1 total $= 21 + 53 + 94 = 168$

expected count $=$ (row total $*$ column total)$/n = (290 * 168)/1362$

b. For testing independence, $\chi^2 = 73.4$. Report the *df* value and the P-value, and interpret.

$df = (3-1)(3-1) = 4$ P-value < 0.001 We reject H_0. There is evidence that income and happiness level are not independent.

c. Interpret the standardized residuals in the corner cells having counts 21 and 83.

For cell 1,1, (count of 21), the standardized residual is -2.973. This cell residual provides marginal evidence that the variables are not truly independent. There were less people with Above average income who are 'Not too happy' than we'd expect if the variables were really independent.

For cell 3,3 (count of 83), the standardized residual is -5.907. This cell provides substantial evidence that income and happiness are not independent. There were less people with Below average income who are 'Very happy' than we'd expect if the variables were really independent.

d. Interepret the standardized residuals in the corner cells having counts 110 and 94

For cell 1,3, (count of 110), the standardized residual is 3.144. This cell residual provides evidence that the variables are not truly independent. There were more people with above average income who are 'Very happy' than we'd expect if the variables were really independent.

For cell 3,1 (count of 94), the standardized residual is 7.368. This cell provides substantial evidence that income and happiness are not independent. There were more people with Below average income who are 'Not too happy' than we'd expect if the variables were really independent.

2.21

a. No, it would not be valid to apply the chi-squared test to this table. Subjects were allowed to choose more than one response, so cell counts do not add up to 100.

b. We can construct three tables of 'Yes/No' data: one for each of the three responses. For example, if we know that 60 of the 100 men chose A, we know that 40 men did not choose A. Likewise, if 75 women chose A, 25 did not. This can be repeated for choices B and C.

Gender	A	Not A
Men	60	40
Women	75	25