Stat 525 Homework 2

Kenny Flagg

September 16, 2016

1. *The data below give population and death rate information by age groups for males and females in a town (Rateboro) by age and sex. Also given is the age-specific population in thousands for the US (the standard population).*

   (a) *Confirm the crude (total) death rates for Rateboro males and females.*

   Total death rate for Rateboro males: $\dfrac{46}{2{,}200} = 0.0209$

   Total death rate for Rateboro females: $\dfrac{54}{2{,}600} = 0.0208$

   (b) *Confirm the male 18-34 age group death rate of 0.0067. This is the only one you need to do.*

   Death rate for Rateboro males aged 18-34: $\dfrac{6}{900} = 0.0067$

   (c) *Find the directly-adjusted death rates for Rateboro males and females separately using the US population as the standard population.*

   Males:

$$\begin{aligned} DAR_{male} &= \frac{0.0067 \times 60{,}000 + 0.0037 \times 45{,}000 + 0.0500 \times 20{,}000 + 0.1100 \times 15{,}000}{60{,}000 + 45{,}000 + 20{,}000 + 15{,}000} \\ &= 0.0230 \end{aligned}$$

   Females:

$$\begin{aligned} DAR_{female} &= \frac{0.0013 \times 60{,}000 + 0.0063 \times 45{,}000 + 0.0200 \times 20{,}000 + 0.0760 \times 15{,}000}{60{,}000 + 45{,}000 + 20{,}000 + 15{,}000} \\ &= 0.0135 \end{aligned}$$

2. *Based on the results in Question 1:*

   (a) *Based on the crude rates do males or females have a more favorable mortality experience in Rateboro?*

   Based on the crude death rates, females in Rateboro have a (barely) lower death rate (0.0208) than males do (0.0209).

   (b) *Based on the adjusted rates do males or females have a more favorable mortality experience in Rateboro?*

   Based on the directly-adjusted death rates, females in Rateboro have a much lower mortality rate (0.0135) than males (0.0230).

   (c) *Which of the rates (crude versus adjusted) are more appropriate in comparing the overall death rate of males and females and why?*

   The adjusted rates are more appropriate because they are not affected by differences in age distribution. When using the crude rates, the difference that is a apparent in the adjusted rates is masked because a larger proportion of the female population is in the two oldest age groups compared to the male population, and these groups have higher death rates than the younger age groups.

   (d) *An experienced epidemiologist had this to say about the data: "the Rateboro data are generally consistent with the typical finding of a more favorable mortality experience of U.S. females; the anomolous result for the 35-59 year-old group, with the high death rate among females (more than 50% greater than the rate for males) is evidence that the Rateboro environment is more suitable for males in the age range 35-59 than for females." Comment on the epidemiologist's comment. (Hint: How many deaths is the epidemiologist basing the latter part of the conclusion on?)*

   The epidemiologist made a statement about relative risk ($RR = 0.0063/0.0038 = 1.66$ means the rate for females is "more than 50% greater"). This is misleading because the death rates are very low (6.3 per thousand for females and 3.8 per thousand for males) in both groups. The fact that few individuals in the 35-59 age group died seems more pertinent than the difference in death rate between males and females of that age group.

3. *Tuberculosis (TB) was once a major problem in the US. Improvements in physical conditions of living and the development of drugs such as isoniazid helped reduce the impacts of the disease and even led to some public health officials dreaming of eradicating it in the US (AIDS helped change that). TB has always been more of a problem in nonwhite populations than in white populations. Data from 3 North Carolina counties are shown below. The first table shows the cases of TB in these 3 counties during the 5 year period from January 1, 1986 to December 31, 1990. The second table shows the mean population size in the 3 counties during the same time period. The third table shows the mean annual incidence of TB in the US over the same time period.*

   (a) *The table below gives age and sex specific incidence rates for TB for each county and overall. Actually the rates are approximations based on the mean population sizes above. Confirm the rate of 7 for white males in Johnston County (this is the only one you need to compute). In computing the rate note that there are 5 years with an average population of 31,721 for each year. Compare the rates discussing any potential reasons for differences.*

   The approximate annual incidence rate for white males in Johnston county during the five years is

$$\frac{11\text{ cases}}{31{,}721\text{ people}\times 5\text{ years}} = 0.0000694\text{ cases / person-year}$$

   or about 7 cases per 100,000 people per year.

   In all three counties, nonwhites have higher TB incidence than whites. In Johnston and Wilson Counties, nonwhite males have much higher TB incidence than nonwhite females, while white males and white females have similar incidence. In Orange County, TB incidence in each group is lower than the incidence for the corresponding group in the other counties, and males and females of the same racial group have similar TB incidence rates. These differences could be explained by economic and infrastructure differences affecting individuals' income and the accessibility and affordability of healthcare in each county.

   (b) *The table below gives SMRs comparing each county to the national TB rates. Confirm the row entries for Johnston County. Compare the 3 rates and discuss.*

   Total observed cases in Johnston County: $11 + 8 + 43 + 13 = 75$

   Expected cases in Johnston County:

$$\frac{7.4\times 31{,}721\times 5}{100,000} + \frac{3.6\times 33{,}955\times 5}{100,000} + \frac{39.2\times 6{,}910\times 5}{100,000} + \frac{19.8\times 8{,}078\times 5}{100,000} = 39.39$$

   SMR for Johnston County: $SMR = \dfrac{75}{39.39} = 1.90$

   (c) *There are 3 possible pairwise comparisons of county-level SMRs we can make. Which of them, if any, is reasonable and which are problematic? Justify your answer.*

   It is reasonable to compare SMRs for Johnston and Orange Counties because these have similar sex and race distributions. Wilson county has a higher proportion of nonwhites, so its SMR is not comparable to the SMRs of the other counties.

4. *A simple random of 5,000 is taken from a population and none of them have rare illness of interest. Give an approximate 95% CI for the proportion who have the disease using the Rule of Three and the exact interval method.*

   The Rule of Three suggests an upper bound of $\dfrac{3}{5000} = 0.0006$ for an interval of (0, 0.0006).

```
Exact binomial test

data:  0 and 5000
number of successes = 0, number of trials = 5000, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.0000000000 0.0007375038
sample estimates:
probability of success
                     0
```

   Using the binomial distribution, the exact interval is (0, 0.00074).

5. *Italy played Bulgaria in a semifinal match in the 1994 World Cup (soccer). A newspaper reported that the odds Italy wins are 11/10 whereas the odds Bulgaria wins are 3/10. Based on the given odds find the probability each wins this semifinal match. Do the given odds make sense? Justify your answer.*

   The probability that Italy wins is $\dfrac{11}{21} = 0.524$. The probability that Bulgaria wins is $\dfrac{3}{13} = 0.231$. These values sum to 0.755 so this doesn't make any because it implies a 0.245 probability that something happens other than one of these teams winning the semifinal.

6. *Problem 4.2 on page 42 in Jewell.*

   Odds of death (men): $\dfrac{682}{161} = 4.24$

   Odds of death (women): $\dfrac{127}{339} = 0.375$

   Odds ratio for death: $\dfrac{4.24}{0.375} = 11.3$

   Odds ratio for survival: $\dfrac{1/4.24}{1/0.375} = 1/11.3 = 0.0884$

   Relative risk of death: $\dfrac{682/843}{127/466} = 2.97$

   Relative risk of survival: $\dfrac{161/843}{339/466} = 0.263$

   Relative risk is the best way to summarize the difference between men and women because it incorporates the comparison and is easy to interpret on a meaningful scale (the probability scale): "Men on the Titanic were 2.97 times as likely to die as women on the Titanic. Men on the Titanic were 0.263 times as likely to survive as women on the Titanic."

7. *Derive the probability distribution for D under the specific heterogeneity assumption given on page 19 of the notes and find the mean and variance of D.*

   Let $Y_i = 1$ if individual i is diseased and 0 otherwise. Then

$$
\begin{aligned}
P(D=0) &= P(Y_1=0)P(Y_2=0)P(Y_3=0)P(Y_4=0) \\
&= 0.80 \times 0.90 \times 0.95 \times 0.99 \\
&= 0.67716, \\
P(D=1) &= P(Y_1=1)P(Y_2=0)P(Y_3=0)P(Y_4=0) \\
&\quad + P(Y_1=0)P(Y_2=1)P(Y_3=0)P(Y_4=0) \\
&\quad + P(Y_1=0)P(Y_2=0)P(Y_3=1)P(Y_4=0) \\
&\quad + P(Y_1=0)P(Y_2=0)P(Y_3=0)P(Y_4=1) \\
&= 0.20 \times 0.90 \times 0.95 \times 0.99 \\
&\quad + 0.80 \times 0.10 \times 0.95 \times 0.99 \\
&\quad + 0.80 \times 0.90 \times 0.05 \times 0.99 \\
&\quad + 0.80 \times 0.90 \times 0.95 \times 0.01 \\
&= 0.28701, \\
P(D=2) &= P(Y_1=1)P(Y_2=1)P(Y_3=0)P(Y_4=0) \\
&\quad + P(Y_1=1)P(Y_2=0)P(Y_3=1)P(Y_4=0) \\
&\quad + P(Y_1=1)P(Y_2=0)P(Y_3=0)P(Y_4=1) \\
&\quad + P(Y_1=0)P(Y_2=1)P(Y_3=1)P(Y_4=0) \\
&\quad + P(Y_1=0)P(Y_2=1)P(Y_3=0)P(Y_4=1) \\
&\quad + P(Y_1=0)P(Y_2=0)P(Y_3=1)P(Y_4=1) \\
&= 0.20 \times 0.10 \times 0.95 \times 0.99 \\
&\quad + 0.20 \times 0.90 \times 0.05 \times 0.99 \\
&\quad + 0.20 \times 0.90 \times 0.95 \times 0.01 \\
&\quad + 0.80 \times 0.10 \times 0.05 \times 0.99 \\
&\quad + 0.80 \times 0.10 \times 0.95 \times 0.01 \\
&\quad + 0.80 \times 0.90 \times 0.05 \times 0.01 \\
&= 0.03451, \\
P(D=3) &= P(Y_1=1)P(Y_2=1)P(Y_3=1)P(Y_4=0) \\
&\quad + P(Y_1=1)P(Y_2=1)P(Y_3=0)P(Y_4=1) \\
&\quad + P(Y_1=1)P(Y_2=0)P(Y_3=1)P(Y_4=1) \\
&\quad + P(Y_1=0)P(Y_2=1)P(Y_3=1)P(Y_4=1) \\
&= 0.20 \times 0.10 \times 0.05 \times 0.99 \\
&\quad + 0.20 \times 0.10 \times 0.95 \times 0.01 \\
&\quad + 0.20 \times 0.90 \times 0.05 \times 0.01 \\
&\quad + 0.80 \times 0.10 \times 0.05 \times 0.01 \\
&= 0.00131,
\end{aligned}
$$

$$\begin{aligned} P(D=4) &= P(Y_1=1)P(Y_2=1)P(Y_3=1)P(Y_4=1) \\ &= 0.20 \times 0.10 \times 0.05 \times 0.01 \\ &= 0.00001. \end{aligned}$$

So the probability mass function is $f_D(d) = \begin{cases} 0.67716 & \text{if } d=0 \\ 0.28701 & \text{if } d=1 \\ 0.03451 & \text{if } d=2 \\ 0.00131 & \text{if } d=3 \\ 0.00001 & \text{if } d=4 \\ 0 & \text{otherwise} \end{cases}$.

The mean is

$$\begin{aligned} E(D) &= 0 \times 0.67716 + 1 \times 0.28701 + 2 \times 0.03451 + 3 \times 0.00131 + 4 \times 0.00001 \\ &= 0.36 \end{aligned}$$

and the variance is

$$\begin{aligned} Var(D) &= E\left[(D-E(D))^2\right] \\ &= E\left(D^2\right) - (E(D))^2 \\ &= 0 \times 0.67716 + 1 \times 0.28701 + 2^2 \times 0.03451 + 3^2 \times 0.00131 + 4^2 \times 0.00001 - 0.36^2 \\ &= 0.3074. \end{aligned}$$