Stat 525 Homework 1

Kenny Flagg

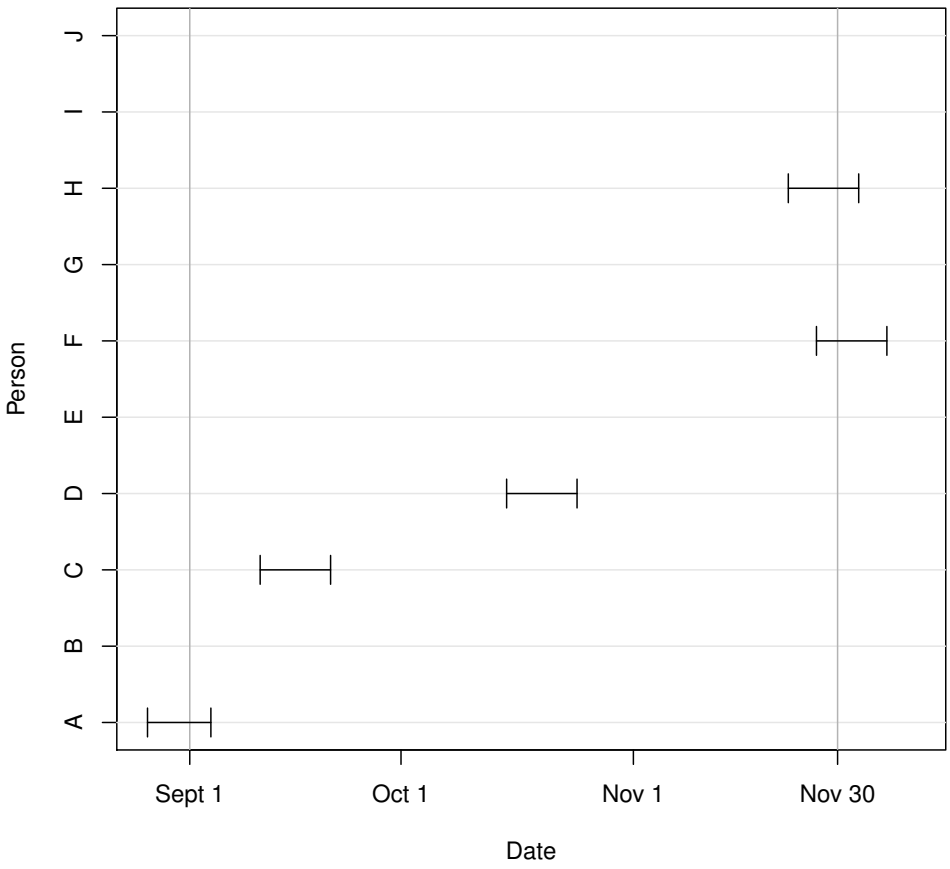
September 9, 2016

1. *We are studying upper respiratory infections (URI) under the following assumptions:*

   - *Each infection lasts 10 days and the person is immune once he/she recovers.*
   - *Infections begin at 12:01 AM on the indicated date.*
   - *No one dies during the study period.*
   - *September and November have 30 days, August and October have 31 days.*
   - *A person with a URI is not at risk of other URI infections during the 10 days he/she is sick.*

   The time period of interest is September 1 - November 30 inclusive. The plot below shows the results for each of the 10 people. The following histories are observed with the dates indicating the date of onset of URI: Person A got sick on August 25, Person C got sick on September 10, Person D got sick on October 15, Person F got sick on November 28, and Person H got sick on November 24. The other individuals did not get sick. The disease histories are shown below.

   (Plot and answers begin on the next page.)



(a) *Compute point prevalence on September 1.*

On September 1, one individual had the URI and nine were at risk, so

$$\text{point prevalence} = \frac{1}{9+1} = 0.1$$

(b) *Compute point prevalence on November 30.*

On November 30, two individuals had the URI and five were at risk, so

$$\text{point prevalence} = \frac{2}{5+2} = 0.286$$

(c) *Compute incidence proportion over the interval [Sept 1, Nov 30].*

On September 1, nine individuals were at risk. Four individuals contracted the URI between September 1 and November 30. Therefore

$$\text{incidence proportion} = \frac{4}{9} = 0.444$$

(d) *Compute person-days at risk for the interval.*

The table below shows how many days of each month each person was at risk. There were a total of 679 person-days at risk in the interval [Sept 1, Nov 30].

Person	September	October	November	Total
A	0	0	0	0
B	30	31	30	91
C	9	0	0	9
D	30	14	0	44
E	30	31	30	91
F	30	31	26	87
G	30	31	30	91
H	30	31	23	84
I	30	31	30	91
J	30	31	30	91
Total	249	231	199	679

(e) *Compute incidence rate for the interval.*

The incidence rate is

$$\frac{4 \text{ people}}{679 \text{ person-days}} = 0.005891\text{day}$$

2. *On page 9 of the notes I give Incidence Proportions and Point Prevalences for CHD for the data in Table 2.1 on page 11. Verify these values.*

- High Cholesterol Incidence: $\frac{85}{85+462} = 0.156$
- High Cholesterol Prevalence: $\frac{38}{38+371} = 0.093$
- Low Cholesterol Incidence: $\frac{28}{28+516} = 0.051$
- Low Cholesterol Prevalence: $\frac{33}{33+347} = 0.087$

3. *Given the definition we have for the incidence proportion we have in Jewell, it could, stangely enough, theoretically exceed 1. What type of disease process would be implied by such a result?*

This would imply that the disease recurs—that is individuals recover quickly, become at-risk again, and contract the disease again multiple times during the interval. Note that this may or may not be possible depending on how a "new case" is defined.

4. *Problem 3.1 on page 29 in Jewell. Read over the problem carefully. Note that the problem asks for CIs for two groups: one who ate no fish and one group who did eat fish. The (approximate) 95% CIs you will be computing are for incidence proportions of CHD related deaths in a 25 year time period. Give me the following for both groups.*

   - *Approximate continuity corrected Wald intervals.*
   - *Approximate continuity corrected Score intervals (use* `prop.test`*).*
   - *The exact intervals (use* `binom.test`*).*

   Based on the intervals does there appear to be evidence of a relationship between CHD related deaths and fish consumption? Justify your answer. (Note: the clearest way to answer this question is to construct an interval for the difference between two incidence proportions, but just address the question by looking at the intervals from the two groups separately. Care is needed and we will all admit to a bit of speculation here.)

   For the men who reported no fish consumption:

   - Proportion who died from CHD:

$$\widehat{p} = \frac{42}{205} = 0.2049$$

   - Wald interval:

$$0.2049 \pm \left(1.96 \sqrt{\frac{0.2049 \times 0.7951}{205}} + \frac{0.5}{205} \right) = (0.1472, 0.2626)$$

   - Score interval:

```
prop.test(42, 205, alternative = 'two.sided', conf.level = 0.95, correct = TRUE)

#
#  1-sample proportions test with continuity correction
#
# data:  42 out of 205, null probability 0.5
# X-squared = 70.244, df = 1, p-value < 2.2e-16
# alternative hypothesis: true p is not equal to 0.5
# 95 percent confidence interval:
#  0.1531417 0.2679440
# sample estimates:
#        p
# 0.204878
```

   The interval is $(0.1531, 0.2679)$

- Exact interval:

```
binom.test(42, 205, alternative = 'two.sided', conf.level = 0.95)

#
#  Exact binomial test
#
# data:  42 and 205
# number of successes = 42, number of trials = 205, p-value < 2.2e-16
# alternative hypothesis: true probability of success is not equal to 0.5
# 95 percent confidence interval:
#  0.1518281 0.2666734
# sample estimates:
# probability of success
#               0.204878
```

The interval is $(0.1518, 0.2667)$

For the men who reported over 35 g of daily fish consumption:

- Proportion who died from CHD:

$$\widehat{p} = \frac{34}{261} = 0.1303$$

- Wald interval:

$$0.1303 \pm \left(1.96\sqrt{\frac{0.1303 \times 0.8697}{261}} + \frac{0.5}{261} \right) = (0.0875, 0.173)$$

- Score interval:

```
prop.test(34, 261, alternative = 'two.sided', conf.level = 0.95, correct = TRUE)

#
#  1-sample proportions test with continuity correction
#
# data:  34 out of 261, null probability 0.5
# X-squared = 141.24, df = 1, p-value < 2.2e-16
# alternative hypothesis: true p is not equal to 0.5
# 95 percent confidence interval:
#  0.09310052 0.17865683
# sample estimates:
#         p
# 0.1302682
```

The interval is $(0.0931, 0.1787)$

- Exact interval:

```
binom.test(43, 261, alternative = 'two.sided', conf.level = 0.95)

#
#  Exact binomial test
#
# data:  43 and 261
# number of successes = 43, number of trials = 261, p-value < 2.2e-16
# alternative hypothesis: true probability of success is not equal to 0.5
# 95 percent confidence interval:
#  0.1218703 0.2154320
# sample estimates:
# probability of success
#               0.164751
```

The interval is $(0.0919, 0.1773)$

Comparison:

The observed proportions of men who died of CHD are 0.2049 in the no fish group and 0.1303 in the more than 35 g of fish group. Since we're just waving our hands around at this point, I'll compare the narrowest intervals (the exact intervals). The exact 95% confidence interval for the true proportion of men who reported consuming no fish and died of CHD is (0.1518, 0.2667). The exact interval for the true proportion of men who reported consuming more than 35 g of fish daily and died of CHD is (0.0919, 0.1773). These intervals overlap, suggesting the observed difference could be due to random chance. There is little evidence that the true proportion of CHD deaths differs between the two groups.

5. *Problem 3.2 on pages 29 and 30 in Jewell.*

- $P(\text{child with single parent}) = \dfrac{65085}{986342} = 0.0660$
- $P(\text{child died}) = \dfrac{664}{986342} = 0.000673$
  - $P(\text{child died}|\text{single parent}) = \dfrac{56}{65085} = 0.000860$
  - $P(\text{child died}|\text{two parents}) = \dfrac{608}{921257} = 0.000660$
- $P(\text{suicide}) = \dfrac{115}{986342} = 0.000117$
  - $P(\text{suicide}|\text{single parent}) = \dfrac{19}{65085} = 0.000292$
  - $P(\text{suicide}|\text{two parents}) = \dfrac{96}{921257} = 0.000104$

The proportions of children who died and committed suicide are small (less than one in 1,000 in both the singe-parent and two-parent groups). Both of these proportions are larger for the single-parent group than for the two-parent group, with the suicide rate more than twice as

high for children in single-parent households than for children in two-parent households. This leaves the possibility that mortality rate and/or suicide rate are associated with living situation. Note that these data cover the entire population of interest and are not representative of children in other countries or time periods, so frequentist hypothesis tests and confidence intervals for the difference in proportions are not approporiate.

6. *STAT Graduate Students:*

   (a) *Recalling that*

$$h(t) = \frac{d}{dt}\left(-\log S(t)\right)$$

   show that, if S(0) = 1 then

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}$$

   Using the fundamental theorem of calculus,

$$\begin{aligned}
\int_0^t \frac{d}{dt}\left(-\log S(u)\right)du &= \int_0^t h(u)du \\
-\log S(u)\big|_{u=0}^t &= \int_0^t h(u)du \\
-\log S(t) + \log S(0) &= \int_0^t h(u)du \\
-\log S(t) + \log 1 &= \int_0^t h(u)du \\
-\log S(t) &= \int_0^t h(u)du \\
S(t) &= \exp\left\{-\int_0^t h(u)du\right\}
\end{aligned}$$

   (b) *Use the result in (a) to find the survival function if*

$$h(t) = \frac{\alpha t^{\alpha-1}}{\beta^\alpha}$$

   The survival function is

$$\begin{aligned}
S(t) &= \exp\left\{-\int_0^t \frac{\alpha u^{\alpha-1}}{\beta^\alpha}du\right\} \\
&= \exp\left\{-\frac{u^\alpha}{\beta^\alpha}\Big|_{u=0}^t\right\} \\
&= \exp\left\{-\frac{t^\alpha}{\beta^\alpha} + \frac{0^\alpha}{\beta^\alpha}\right\} \\
&= \exp\left\{-\left(\frac{t}{\beta}\right)^\alpha\right\}
\end{aligned}$$

(c) *Find the probability density function for survival time given the survival function you found in part (b).*

The cumulative probability density function for survival time is

$$\begin{aligned} F(t) &= P(\text{dies before time } t) \\ &= 1 - P(\text{survives past time } t) \\ &= 1 - S(t) \\ &= 1 - \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\} \end{aligned}$$

for $t > 0$, so the probability density function is

$$\begin{aligned} f(t) &= \frac{d}{dt}F(t) \\ &= \frac{\alpha t^{\alpha-1}}{\beta^{\alpha}} \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\} \end{aligned}$$

for $t > 0$, and zero elsewhere.