2018 Prelim

Will Gertsch

8/19/2020

1

a

$P(D|+) = 170/200 = 0.85$. We will use the normal approximation confidence interval for the proportion

$$\hat{p} \pm Z_{.975}\sqrt{Var(\hat{p})}$$

where $Var(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{200}$. We use 200 in the denominator since we are only considering the sample of women who test positive. Therefore the confidence interval is

$$0.85 \pm 1.96\sqrt{.85(.15)/200} \implies (0.8005124, 0.8994876)$$

b Assuming $P(D|-) = 0$.

i

Use the law of total probability

$$P(D) = P(D|+)P(+) + P(D|-)P(-) = P(D|+)P(+) = \frac{170}{200} * \frac{200}{5000} = 0.034$$

ii

$\hat{p} = 0.034$. Using the normal approximation again

$$0.034 \pm 1.96\sqrt{.034(1-.034)/5000} \implies (0.02997658, 0.04002342)$$

c Assuming $P(D|-) = 0.06$. Use the law of total probability again.

$$P(D) = P(D|+)P(+) + P(D|-)P(-) = 0.034 + (0.06)(4800/5000) = 0.0916$$

d

We start with the normal CI and solve for n.

$$1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.02$$

Rearranging, we get

$$n = \frac{\hat{p}(1-\hat{p})}{0.0001041233}$$

Since we do not have an estimate of $\hat{p}$, we could use the critics proposed proportion. However, I would prefer to use $\hat{p} = 1/2$ since that value will maximize n and give us an upper bound for the sample size we need. Therefore $n = 2401$. This is actually less than what the investigators thought they needed.

2

a

$$\hat{\beta}_1 = r_{XY}\frac{SD(Y)}{SD(X)} = 0.366 * \frac{1.09}{1.125} = 0.3546133$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 1.953 - 0.3546133 * 1.108 = 1.560088$$

b Every additional attack is associated with a $10^{0.3546133} = 2.262629$ increase in the level of IgE (no units were given for this).

c

I chose to think of this as an overall F-test first.

$$F^* = \frac{MSR}{MSE} = \frac{\frac{5.73}{1}}{\frac{37.06}{35}} = 5.411495$$

F^* has a $F(1, 35)$ distribution under the null. $F_{.95}(1, 35) = 4.121338$ so reject the null hypothesis. Therefore is evidence that the number of attacks has a significant linear relationship with log10(IgE).

If we didn't want to use the F distribution, we could note that $F^* = T^2$ and that the equivalent t distribution is $t(35)$ with critical value $t_{.95}(35) = 1.689572$.

d

These models are not nested since the SLR model is not a constrained version of the dummy variable model. This means that we cannot use the F-test to compare the models. However since the problem is obviously set up for a partial F-test, I will do a partial F-test.

$$F^* = \frac{\frac{6.76-5.73}{2}}{\frac{36.02}{33}} = 0.4718212$$

The distribution of F^* is $F(2, 33)$ and the critical value is $F_{.95}(2, 33) = 3.284918$. Therefore, fail to reject the null hypothesis. The dummy variable model is not significantly better at explaining the variance of log10(IgE).

e

I would recommend a Poisson regression model with link function $g(\lambda) = \log\lambda$. In other words, $Y|X = x_i \sim$ Poisson(λ_i). Thus the model is

$$\log\lambda_i = \beta_0 + \beta_1 X_i$$

Every 10-fold unit increase in IgE is associated with a e^{β_1} increase in the number of attacks.

3

a

We can fit the model by noting that $e^{\beta_0+\beta_1}$ e^{β_0} are odds. Let $p_0 = 32/(32 + 81) = 0.2831858$ and $p_1 = 14/(14 + 93) = 0.1308411$ be the probabilities of AIDS symptoms for the standard and new treatments

respectively. These correspond to odds of $\frac{p_0}{1-p_0} = 0.3950616$ and $\frac{p_1}{1-p_1} = 0.1505376$. Since $e^{\hat{\beta}_0}$ are the baseline odds, we can set it equal to the odds for the standard treatment.

$$e^{\hat{\beta}_0} = 0.3950616 \implies \hat{\beta}_0 = -0.9287136$$

To get $\hat{\beta}_1$, we note that that $e^{\hat{\beta}_1}$ is the OR of the treatments.

$$e^{\hat{\beta}_1} = \frac{0.1505376}{0.3950616} \implies \hat{\beta}_1 = -0.9648289$$

b The easiest test we can conduct is the Wald test.

$$\frac{\hat{\beta}}{SE(\hat{\beta})} \dot{\sim} N(0,1)$$

What is the SE of the log-odds ratio? Recall from 200A that

$$SE(\hat{OR}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Therefore, $SE(\hat{\beta}) = 0.3546504$. Therefore the value of the test statistic is $-0.9648289/0.3546504 = -2.720507$ which is less than the critical value of -1.96. Therefore reject the null hypothesis and conclude that the treatment significantly decreases the odds of AIDS symptoms during follow-up.

c

Use the Wald CI

$$\hat{\beta}_1 \pm Z_{.975} SE(\hat{\beta}_1) \implies -0.7195 \pm 1.96\sqrt{0.07783}$$

Therefore the confidence interval is (-1.266301, -0.1726986). Since the interval does not include 0, we conclude that the treatment decreases the odds of AIDS symptoms at follow up while controlling for race. This is a stronger argument in favor of the treatment than (b) since the model is now accounting for race.

d

$$-2\log\lambda = -2(-167.576 + 166.884) = 1.384$$

The test statistic follows an approximate χ^2_1 distribution with the critical value 3.84. Therefore, the interaction term does not significantly improve the model. This means that the treatment effect is the same for both races. This makes sense since the estimated treatment effect does not change very much from (c) to (b).

4

Thanks to John for providing this solution.

a

Let $Y_i = F(X_i)$. The CDF of Y_i is

$$P(Y_i \leq y) = P(F(X_i) \leq y) = P(X_i \leq F^{-1}(y)) = F(F^{-1}(y)) = y$$

Therefore $Y_i \sim U(0,1)$ since the CDF of a uniform is $\frac{y-a}{b-a}$.

Thus the pdf of Y_i is 1 and the joint pdf of all the $Y_i's$ is 1 for $0 < y_i < 1$.

b

$$P(F(X_{(1)}) \leq 0.5) = P(X_{(1)} \leq F^{-1}(0.5)) = 1 - P(X_{(1)} > F^{-1}(0.5))$$

Now break up the inequality using the defintion of the minimum.

$$1 - P(X_{(1)} > F^{-1}(0.5)) = 1 - P\left(X_1 > F^{-1}(0.5), \ldots, X_n > F^{-1}(0.5)\right)$$

Since the $X_i's$ are iid, we can rewite as

$$= 1 - \prod_{i=1}^{n} P(X_i > F^{-1}(0.5)) = 1 - P(X_i > F^{-1}(0.5))^n = 1 - \left[1 - F(F^{-1}(0.5))\right]^n$$

So the final answer is

$$= 1 - (1/2)^n$$

I had originally though that the answer would be 1 since $P(X_{(1)} \leq F^{-1}(0.5))$ seems to be asking about the min being less than the median. The trick is that $X_{(1)}$ is the min of the sample while $F^{-1}(0.5)$ is the theoretical median.

I also ran a simulation that confirms the result.

```
n = 2 # sample size
S = 1e6 # number of simulations
ind = numeric(S) # whether inequality holds
mins = numeric(S)
for (i in 1:S) { # takes a little while to run if n is big
  X <- rnorm(n) # sample from continuous distr
  mins[i] = min(X)
  ind[i] = pnorm(mins[i]) <= 0.5
}

mean(ind) # prob that min is less than median
```

```
## [1] 0.750593
```

```
1-(1/2)^n # theoretical value
```

```
## [1] 0.75
```

c

The joint distribution of $X_{(1)}$ and $X_{(n)}$ is

$$f_{X_{(1)},X_{(n)}}(u,v) = \frac{n!}{(n-1)!(n-2)!} \cdot 1 \cdot 1 \cdot (v-u)^{n-2}(1-v)^{n-1} = \frac{n}{(n-2)!}(v-u)^{n-2}(1-v)^{n-1}$$

for $0 < u, v < 1$.

5

a

By the central limit theorem

$$\sqrt{n}(\bar{X} - \theta^{-1}(1-\theta)) \to^D N(0, \theta^{-2}(1-\theta))$$

b We will use the continuous mapping theorem.

$$\tilde{\theta} = \frac{n + (1+\bar{X})}{(n+1)(1+\bar{X})} = \frac{n}{n+1}\frac{1}{1+\bar{X}} + \frac{1}{n+1}$$

As $n \to \infty$, $n/(n+1) \to 1$ and $1/(n+1) \to 0$. By the continuous mapping theorem, $1/(1+\bar{X}) \to^p 1/(1+\mu)$. Therefore,

$$\tilde{\theta} \to^p \frac{1}{1 + \frac{1-\theta}{\theta}} = \theta$$

To find the limiting distribution, use the delta method. First, let

$$\tilde{\theta} = g(\bar{X}) = \frac{n}{n+1}\frac{1}{1+\bar{X}} + \frac{1}{n+1}$$

$$g'(x) = -\frac{n}{n+1}\frac{1}{(1+x)^2}$$

$$g(\mu) = g(\theta^{-1}(1-\theta)) = \frac{n}{n+1}\theta + \frac{1}{n+1}$$

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \dot{\sim} N\left(0, \left[-\frac{n}{n+1}\theta^2\right]^2 \theta^{-2}(1-\theta)\right)$$

Therefore the limiting distribution is

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \to^D N\left(0, \theta^2(1-\theta)\right)$$

c

Likelihood:

$$L(\theta) = \theta^n (1-\theta)^{\sum x_i}$$

Log-likelihood:

$$\ell(\theta) = n\log\theta + \sum x_i \log(1-\theta)$$

$$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{n}{\theta} - \frac{\sum x_i}{1-\theta} = 0$$

Solving this equation, $\theta = \frac{n}{\sum x_i + n}$. Check to see if $L(\theta)$ is concave.

$$\ell''(\theta) = -\frac{n}{\theta^2} - \frac{\sum x_i}{(1-\theta)^2} < 0$$

Therefore the mle is

$$\hat{\theta} = \frac{n}{\sum x_i + n} = \frac{n}{n(\bar{X}+1)} = \frac{1}{\bar{X}+1}$$

To find the limiting distribution, we note that the mle is a function of $\bar{X}$. We can use this function for the delta method.

$$g(x) = \frac{1}{x+1}, g'(x) = -\frac{1}{(x+1)^2}$$

$$g(\mu) = g\left(\frac{1-\theta}{\theta}\right) = \theta$$

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \to^D N\left(0, [-\theta^2]^2\theta^{-2}(1-\theta)\right)$$

Therefore the limiting distribution is

$$N\left(0, \theta^2(1-\theta)\right)$$

d

(i)

$$p(\theta|X) \propto P(X|\theta)P(\theta) \propto \theta^{(a+n)-1}(1-\theta)^{\left(\sum x_i + b\right)-1}$$

Therefore the posterior distribution is Beta$(a+n, \sum x_i + b)$.

(ii)

The mean of a beta distribution is $a/(a+b)$ so the mean of the posterior is

$$\frac{a+n}{a+n+\sum x_i + b}$$

To compute this expected value

$$EX = \int_0^\infty x \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} dx = \frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b}$$