2019-prelim

Will Gertsch

8/14/2020

1

a

$$OR = e^{\beta_3} = e^{.071293} = 1.073896$$

The odds of a miss for males is 1.07 the odds of a miss if female, holding all else constant.

b

Solving for the probability, we get

$$p_i = \frac{e^{\beta_0+\beta_1 t_i+\beta_2 s_i+\beta_3 sex}}{1+e^{\beta_0+\beta_1 t_i+\beta_2 s_i+\beta_3 sex}}$$

Plugging the problem values

$$e^{\beta_0+\beta_1\cdot 2+\beta_2\cdot 0+\beta_3\cdot 1} = e^{-.715349+2(.0040216)+.0712193} = 0.52936$$

Therefore,

$$p_i = \frac{0.52936}{1+0.52936} = 0.3461317$$

c

The coefficient of s is negative with a confidence interval that does not include 0. This signifies that the period after the program began showed a significant decline in misses. We could test this with

$$\frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \dot{\sim} N(0,1)$$

which is what the output is showing. We could also use a deviance based test where we compare the model with and without s.

d

We could test this hypothesis using the ANOVA contrast:

$$L = \frac{\beta_1+\beta_2+\beta_3+\beta_4}{4} - \frac{\beta_6+\beta_7+\beta_8+\beta_9}{4}$$

The hypotheses would be $H_0 : L = 0$ and $H_A : L \neq 0$. This test will compare the average effect of these two time periods but it will not tell you if there is a difference in trend. The model in c) says that there is a negative trend in the later period and gives an month-by-month estimate of the decrease.

2

a

$$P(X=k) = \binom{k+r-1}{k}(1-p)^k p^r = \exp\left[\log\binom{k+r-1}{k} + k\log(1-p) + r\log p\right]$$

$$\exp\left[\frac{k\log(1-p) - (-r\log p)}{1} + \log\binom{k+r-1}{k}\right]$$

Therefore $\theta = \log(1-p)$, $b(\theta) = -r\log p = -r\log(1-e^\theta)$, $a(\phi) = 1$.

b

Recall the mean of a exponential distribution is $b'(\theta)$ and the variance is $a(\phi)b''(\theta)$.

$$b'(\theta) = \frac{-r}{\log(1-e^\theta)} \cdot -e^\theta = \frac{re^\theta}{1-e^\theta}$$

Plugging in for θ:

$$b'(\theta) = \frac{r(1-p)}{p}$$

Therefore, $EX = \frac{r(1-p)}{p}$. Now take the 2nd derivative to find the variance.

$$b''(\theta) = \frac{re^\theta(1-e^\theta) + re^\theta(e^\theta)}{(1-e^\theta)^2} = \frac{re^\theta(1-e^\theta+e^\theta)}{(1-e^\theta)^2} = \frac{re^\theta}{(1-e^\theta)^2}$$

Plug in for *theta*

$$b''(\theta) = \frac{r(1-p)}{p^2}$$

Therefore, $\operatorname{Var} X = \frac{r(1-p)}{p^2}$

c

Modeling the number of malaria episodes for a clinical trial of a malaria drug. The link function must meet the requirement such that $\theta = g(\mu)$. Therefore, we have

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

Therefore

$$g(\mu) = \log\mu - \log r + \log p = \log\frac{\mu p}{r}$$

Is it okay that the link function depends on the parameter of the modeled distribution?

d

Could use either χ^2-goodness of fit or deviance test.

e

If $\lambda \sim \gamma(\alpha, \beta)$ in $Y \sim Poisson(\lambda)$, then Y is negative binomial. The mean of Poisson and negative binomial are the same, but the variances are different. Specifically, the variance of the negative binomial is allowed to be larger. this means that if the data was really Poisson, the negative binomial model would over estimate the variance and we would have underdispersion.

3

a

Use the CDF method:

$$P(Y \leq y) = P(|X| \leq y) = P(-y \leq X \leq y) = F_X(y) - F_X(-y)$$

Therefore, the pdf is

$$f_Y(y) = f_X(y) + f_X(-y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} + \frac{1}{\sqrt{2\pi}} e^{-(-y)^2/2} = \frac{2}{\sqrt{2\pi}} e^{-y^2/2}$$

for $0 < y < \infty$.

b

$$M_Y(t) = E(e^{tY}) = E(e^{t|X|}) = \int_{-\infty}^{\infty} e^{t|X|} f_X(x) dx$$

We can break the integral into two pieces based the absolute value.

$$\int_{-\infty}^{\infty} e^{t|X|} f_X(x) dx = \int_{-\infty}^{0} e^{-tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx + \int_{0}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Now we combine the exponential terms by completing the square like so.

$$-tx - x^2/2 = -\frac{1}{2}(x^2 + 2tx + t^2 - t^2) = -\frac{1}{2}(x+t)^2 + t^2/2$$

$$tx - x^2/2 = -\frac{1}{2}(x^2 - 2tx + t^2 - t^2) = -\frac{1}{2}(x-t)^2 + t^2/2$$

The integrals are now

$$e^{t^2/2} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+t)^2} dx + e^{t^2/2} \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx$$

Note that the kernels are $N(-t, 1)$ and $N(t, 1)$.

Now let's apply the transformation $z = x + t$, $dz = dx$ to the first integral to get

$$e^{t^2/2} \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Apply the transformation $z = x - t$ to the second integral to get

$$e^{t^2/2} \int_{-t}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Note that both kernels have been standardized. We can use the fact that the cdf of X is

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

to rewrite the integrals.

$$e^{t^2/2} \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz + e^{t^2/2} \int_{-t}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = e^{t^2/2} F_X(t) + e^{t^2/2}(1 - F_X(-t))$$

Therefore, the moment generating function is

$$M_Y(t) = e^{t^2/2} \left(F_X(t) - F_X(-t) + 1\right)$$

c

$$M_Y'(t) = te^{t^2/2}\left(F_X(t) - F_X(-t) + 1\right) + e^{t^2/2}\left(2f_X(t)\right)$$

Therefore the mean is

$$EY = M_Y'(t)|_{t=0} = 0 + 1 \cdot 2f_X(0) = \frac{2}{\sqrt{2\pi}}e^0 = \sqrt{\frac{2}{\pi}}$$

4

a

Whatever 95% CI we end up using, we can use this probability statement

$$P(\text{CI includes } \mu) = .95$$

because this will be true if we let the number of simulations go to infinity. Therefore, the probability of 5 CI including μ is $(.95)^5 = .77$.

b

We must figure out the probability that the CI includes the median and then compute the probability that 5 simulated CI's will include the median. Because we know the population standard deviation, we can use the CI

$$\bar{X} \pm Z_{.975}\frac{\sigma}{\sqrt{n}}$$

In the problem, $Z_{.975}\frac{\sigma}{\sqrt{n}} = 1.96 * (16/10) = 3.136$. Therefore

$$P(\text{CI includes } m) = P(\bar{X} - 3.136 < 40 < \bar{X} + 3.136) = P(36.864 < \bar{X} < 43.136)$$

Apply the central limit theorem, $\bar{X} \dot{\sim} N(36, 2.56)$.

$$= P\left(\bar{X} < \frac{43.136 - 36}{\sqrt{2.56}}\right) - P\left(\bar{X} < \frac{36.864 - 36}{\sqrt{2.56}}\right) = \Phi(4.46) - \Phi(0.54)$$

Therefore, $P(\text{CI includes } m) = .29$. Finally, the probability that all the CI contains the median in all 5 simulations is $.29^5 = 0.002$.

c

This question is equivalent to finding the sample size required such that

$$P(\hat{p} - .0075 < p < \hat{p} + .0075) = .99$$

We will need to select a CI for the proportion. I chose to use the normal approximation

$$\hat{p} \pm Z_{.995}\sqrt{\frac{\hat{p}(1-\hat{p})}{S}}$$

where S is the number of simulations. Usually, we could use $\hat{p}$ to find the variance, but we can't since we don't know $\hat{p}$. Instead, we can find the value that maximizes $\hat{p}(1-\hat{p})$ and use that to establish a worst case simulation count. $\hat{p}_{max} = 0.5$ so we can solve for S.

$$2.57\sqrt{\frac{1/4}{S}} = .0075 \implies S = 29,355.1$$

Therefore, I would recommend running the simulation 29,356 times.

d

I will also use the normal approx. here. (rule of 5 doesn't hold, but it's best I can do and I get the same conclusion when doing an exact test in R)

The hypotheses are $H_0 : p = 0.9$, $H_A : p \neq 0.9$. The test statistic is

$$Z^* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/s}} = \frac{1 - 0.9}{\sqrt{0.9 * 0.1/10}} = 1.054093$$

The p-value is

$$2(1 - \Phi(1.054093)) = 0.2918404$$

e

Once again, I use the normal approx.

$$\hat{p} \pm Z_{.975}\sqrt{p_0(1-p_0)/s} \implies 1 \pm 1.96\sqrt{.9 * .1/10}$$

Therefore the confidence interval is (0.8140581, 1.185942), which we collapse to (0.8140581, 1).

f

No, she is accepting the null hypothesis. Suppose we used 0.8 as the null value. If we perform the analysis, then we also get a non-significant p-value. Therefore, by her logic, the true value would also be 0.8. The confidence interval is better evidence, but it is much too wide to say that the true coverage is 90%.

for fun

I also did a Bayesian analysis of this for fun.

Suppose we model the data as binomial and use the conjugate prior $p \sim Beta(1,1) = U(0,1)$. The posterior distribution is $p|Y \sim Beta(11,1)$. The mode of the posterior is 1 since the prior is uninformative and all the data are 1. The 95% credible interval is (0.7277416, 0.9970675). This supports my conclusion that she probably shouldn't claim that the true coverage is 90%.

5

a

Likelihood:

$$L(a,b) = \left(\frac{b^a}{\Gamma(a)}\right)^n \left(\prod x_i\right)^{a-1} e^{-b\sum x_i}$$

Log-likelihood:

$$\ell(a,b) = na\log(b) - n\log(\Gamma(a)) + (a-1)\sum \log x_i - b\sum x_i$$

Take the first derivative and set equal to zero to obtain likelihood equations:

$$\frac{\partial}{\partial a}\ell(a,b) = n\log b - \frac{n\Gamma'(a)}{\Gamma(a)} + \sum \log x_i = 0$$

$$\frac{\partial}{\partial b}\ell(a,b) = \frac{na}{b} - \sum x_i = 0$$

b

This is were I needed help. I found this problem was assigned when Dr. Telesca taught the course.

$$I_n(a,b) = -E\begin{bmatrix} \frac{\partial^2 \ell}{\partial a^2} & \frac{\partial^2 \ell}{\partial a \partial b} \\ \frac{\partial^2 \ell}{\partial b \partial a} & \frac{\partial^2 \ell}{\partial b^2} \end{bmatrix} = -E\begin{bmatrix} -n\psi'(a) & \frac{n}{b} \\ \frac{n}{b} & -\frac{na}{b^2} \end{bmatrix}$$

where $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$. Therefore

$$I_n(a,b) = n\begin{bmatrix} \psi'(a) & -1/b \\ -1/b & a/b^2 \end{bmatrix}$$

The mles follow the joint asymptotic distribution

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \text{MVN}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \frac{1}{n}\begin{bmatrix} \psi'(a) & -1/b \\ -1/b & a/b^2 \end{bmatrix}^{-1}\right)$$

c

$\mu = a/b$, therefore $\hat{\mu} = \hat{a}/\hat{b}$ by the invariance property of MLEs. We can use the 2nd likelihood equation from before to say that $\hat{\mu} = \bar{X}$. Therefore

$$\hat{\mu} \dot{\sim} N\left(\mu, \frac{VarX}{n}\right) = N\left(\mu, \frac{a}{nb^2}\right)$$

d

By Bayes theorem

$$P(b|X) \propto P(X|b)P(b) \propto (b^a)^n e^{-b\sum x_i} b^{\alpha-1} e^{-\beta b} = b^{na+\alpha-1} e^{-b(\sum x_i + \beta)}$$

Therefore the posterior is

$$b|X \sim \text{Gamma}\left(na+\alpha, \sum x_i + \beta\right)$$

The mean of a gamma is a/b, therefore the mean of the posterior is

$$\tilde{b}_n = \frac{na+\alpha}{\sum x_i + \beta}$$

To show consistency, we will use the continuous mapping theorem.

$$\tilde{b}_n = g(\bar{X}) = \frac{na+\alpha}{n\bar{X}+\beta} = \frac{na}{n\bar{X}+\beta} + \frac{\alpha}{n\bar{X}+\beta} = \frac{a}{\bar{X}+\beta/n} + \frac{1}{n}\frac{a}{\bar{X}+\beta/n}$$

As $n \to \infty$, $g(\bar{X}) \to^p g(\mu)$ and the second term will go to zero while the β drops out of the first term. $\mu = a/b$.

$$g(\bar{X}) \to^p g(\mu) = \frac{a}{a/b} = b$$

e

The hypotheses are $H_0 : a_x = a_y, b_x = b_y$ and $H_A : a_x \neq x_y, b_x \neq b_y$. We will use a likelihood ratio test. The likelihood for the alternative model is

$$L(\hat{\theta}) = \prod_i^n \frac{\hat{b}_x^{\hat{a}_x}}{\Gamma(\hat{a}_x)} x_i^{\hat{a}_x - 1} e^{-\hat{b}_x x_i} \prod_j^m \frac{\hat{b}_y^{\hat{a}_y}}{\Gamma(\hat{a}_y)} y_i^{\hat{a}_y - 1} e^{-\hat{b}_y y_i}$$

$L(\theta_0)$ will be the same, but with $\hat{a}_0, \hat{b}_0$ as the estimates for the parameters of both distributions. The test statistic is the log-likelihood ratio statistic

$$2 \log \frac{L(\hat{\theta})}{L(\theta_0)} \dot{\sim} \chi^2_2$$

The alternative model has 4 parameters while the null model has 2 parameters. Hence the difference in degrees of freedom is 2.