Prediction:

total samples - 306

81 - ¬Survived with avg of 7.5 positive ax nodes

225 - Survived with avg of 2.8 positive ax nodes

positive nodes follows geometric distribution so:

$\rightarrow \mu = \frac{1}{p}$

We need to find: P(survived | nodes=2)

Using bayes theorem:

#1 $P(\text{Survived} \mid \text{nodes}=2) = \frac{P(\text{nodes}=2 \mid \text{survived}) \cdot P(\text{survived})}{P(\text{nodes}=2)}$

#2 $P(\text{nodes}=x \mid \neg\text{Survived}) \sim \text{geo} \rightarrow 7.5 = \bar{x} = \frac{1}{\hat{p}_1} \Rightarrow \hat{p}_1 = \frac{2}{15}$

$\Rightarrow P(\text{nodes}=x \mid \neg\text{Survived}) = \left(1-\frac{2}{15}\right)^{x-1} \cdot \frac{2}{15}$

#3 $P(\text{nodes}=x \mid \text{Survived}) \sim \text{geo} \rightarrow 2.8 = \bar{x} = \frac{1}{\hat{p}_2} \Rightarrow \hat{p}_2 = \frac{5}{14}$

$\Rightarrow P(\text{nodes}=x \mid \text{Survived}) = \left(1-\frac{5}{14}\right)^{x-1} \cdot \frac{5}{14}$

#4 $P(\text{survived}) = \frac{225}{306} = \frac{25}{34} \Rightarrow P(\neg\text{Survived}) = 1 - \frac{25}{34} = \frac{9}{34}$

#5 $P(\text{nodes}=2) = \underset{\text{law of total probability}}{=} P(\text{nodes}=2 \mid \text{Survived}) \cdot P(\text{Survived}) + P(\text{nodes}=2 \mid \neg\text{Survived}) \cdot P(\neg\text{Survived}) =$

law of total probability

by #2, #3, #4

$= \left(1-\frac{5}{14}\right)^1 \cdot \frac{5}{14} \cdot \frac{25}{34} + \left(1-\frac{2}{15}\right)^1 \frac{2}{15} \cdot \frac{9}{34} = \frac{1125}{6664} + \frac{13}{425} = 0.1994$

$\Rightarrow P(\text{survive} \mid \text{nodes}=2) = \frac{\frac{1125}{6664}}{0.1994} = \underline{\underline{0.8466}}$

Likelihood:

1. PDF: $f(x,\theta) = \begin{cases} \frac{2x}{\theta} \cdot e^{-\frac{x^2}{\theta}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$

$$L(\theta) = P(\vec{x}|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Using log-likelihood we get:

$$LL(\theta) = \log \prod_{i=1}^{n} P(x_i|\theta) = \sum_{i=1}^{n} \log(f(x_i|\theta)) =$$

$$= \sum_{i=1}^{n} \log\left(\frac{2x}{\theta} \cdot e^{-\frac{x^2}{\theta}}\right) = \sum_{i=1}^{n} \left(\log(2x) - \log(\theta) - \frac{x^2}{\theta}\right)$$

We use the dirivative of $LL(\theta)$ to find the maximum likelihood and the θ estimator:

$$0 = \frac{\partial LL(\theta)}{\partial \theta} = \sum_{i=1}^{n} -\frac{1}{\theta} + \frac{x^2}{\theta^2}$$

$\xrightarrow[\text{by } \theta^2]{\text{multiply}}$ $$0 = \sum_{i=1}^{n} x_i^2 - \hat{\theta}$$

$$n\hat{\theta} = \sum_{i=1}^{n} x_i^2$$

The estimator is: $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i^2}{n}$

2. $\hat{\theta} = \frac{0.5^2}{3} + \frac{0.5^2}{3} + \frac{1^2}{3} = \frac{1}{2}$

Hypothesis:

$X \sim Bernoulli(p_x)$ $\qquad$ $Y \sim Bernoulli(p_y)$

$n_x = 100$ $\qquad$ $n_y = 150$

$\hat{p}_x = \frac{60}{100}$ $\qquad$ $\hat{p}_y = \frac{70}{150}$

The hypothesis we test:

$$H_0: p_x = p_y \quad vs. \quad H_x: p_x > p_y$$

We assume that the sample sizes are large enough to use normal approximation, therefor the test statistic is:

$$\hat{z} = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{100}+\frac{1}{150}\right)}} = \frac{0.6 - \frac{7}{15}}{\sqrt{\frac{13}{25}\cdot\frac{12}{25}\cdot\frac{1}{60}}} = 2.067$$

$$\hat{p} = \frac{\hat{p}_x \cdot n_x + \hat{p}_y \cdot n_y}{n_x + n_y} = \frac{100 \cdot 0.6 + 150 \cdot \frac{7}{15}}{250} = \frac{13}{25}$$

Since our test is one sided (H_0 vs H_x) and ~~and~~ our $\alpha = 0.01$ (1% significance level) we need to compare $\hat{z} = 2.067$ with

$$Z_{1-\alpha} = Z_{0.99} = 2.326 > 2.067$$

$\longrightarrow$ We cannot reject our null hypothesis (with 1% confidence level).