

Final Exam  
David Yang

1. The conditional distribution of  $\theta \mid Y_1, \alpha$  satisfies

$$\Pr(\theta \mid Y_1, \alpha) = \frac{L(Y_1 \mid \theta, \alpha) \Pr(\theta \mid \alpha)}{\Pr(Y_1 \mid \alpha)}.$$

Since  $L(Y_1 \mid \theta, \alpha) = f(y_1 \mid \theta)$ ,  $\Pr(\theta \mid \alpha) = g(\theta \mid \alpha)$ , and

$$\Pr(Y_1 \mid \alpha) = \int \Pr(Y_1, \theta \mid \alpha) d\theta = \int f(y_1 \mid \theta, \alpha) g(\theta \mid \alpha) d\theta,$$

we substitute to get that

$$\begin{aligned} \Pr(\theta \mid Y_1, \alpha) &= \frac{L(Y_1 \mid \theta, \alpha) \Pr(\theta \mid \alpha)}{\Pr(Y_1 \mid \alpha)} \\ &= \frac{f(y_1 \mid \theta) g(\theta \mid \alpha)}{\int f(y_1 \mid \theta, \alpha) g(\theta \mid \alpha) d\theta}. \end{aligned}$$

We conclude that

$$\boxed{\theta \mid Y_1, \alpha \sim \frac{f(y_1 \mid \theta) g(\theta \mid \alpha)}{\int f(y_1 \mid \theta, \alpha) g(\theta \mid \alpha) d\theta}}.$$

Similarly, note that

$$\Pr(Y_2 \mid Y_1, \alpha) = \int L(Y_2 \mid Y_1, \theta, \alpha) \Pr(\theta \mid Y_1, \alpha) d\theta.$$

Since  $Y_1$  and  $Y_2$  are independent, it follows that  $L(Y_2 \mid Y_1, \theta, \alpha) = L(Y_2 \mid \theta, \alpha)$ . Consequently, substituting  $L(Y_2 \mid Y_1, \theta, \alpha) = L(Y_2 \mid \theta, \alpha) = f(y_2 \mid \theta)$  and our result  $\Pr(\theta \mid Y_1, \alpha) = \frac{f(y_1 \mid \theta) g(\theta \mid \alpha)}{\int f(y_1 \mid \theta, \alpha) g(\theta \mid \alpha) d\theta}$  from above, we get that

$$\begin{aligned} \Pr(Y_2 \mid Y_1, \alpha) &= \int L(Y_2 \mid \theta, \alpha) \Pr(\theta \mid Y_1, \alpha) d\theta \\ &= \int f(y_2 \mid \theta) \frac{f(y_1 \mid \theta) g(\theta \mid \alpha)}{\int f(y_1 \mid \theta, \alpha) g(\theta \mid \alpha) d\theta} d\theta \end{aligned}$$

and we conclude that

$$\boxed{Y_2 \mid Y_1, \alpha \sim \int f(y_2 \mid \theta) \frac{f(y_1 \mid \theta) g(\theta \mid \alpha)}{\int f(y_1 \mid \theta, \alpha) g(\theta \mid \alpha) d\theta} d\theta}.$$

2. a) Let  $p_1$  represent the proportion of respondents indicating preference for immunity from criminal prosecution for version 1 of the survey question, and let  $p_2$  represent the proportion of respondents indicating preference for immunity from criminal prosecution for version 2 of the survey question.

Our null and alternative hypotheses are as follows:

$$H_0: p_1 - p_2 = 0, H_1: p_1 - p_2 \neq 0.$$

We test using a 2-proportion z-test for the differences in responses between these two surveys. Our test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{(\hat{p}(1 - \hat{p})) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where  $\hat{p}_1 = \frac{73}{384}$ ,  $\hat{p}_2 = \frac{97}{324}$ ,  $n_1 = 384$ ,  $n_2 = 324$ , and  $\hat{p} = \frac{73+97}{384+324}$ . Plugging in these values and solving, we get that  $z \approx -3.39$ .

Using  $R$ , we find that  $P(|z| < 3.39) \approx \boxed{0.00069}$ , which is a statistically significant result. We conclude that there is a statistically significant difference between the proportion of respondents indicating a preference for presidential immunity in the two question versions.

All of these results match those from the *prop.test()* command in  $R$  (see attached code).<sup>1</sup>

- b) A statistically significant difference does imply evidence for a causal effect. This national survey is an experiment; there is randomization in terms of which “subjects” receive the two distinct survey versions, which ensures that the responses are representative of the surveyed population. Furthermore, the randomization removes the effect of other confounding variables. Consequently, since the result is statistically significant (hence not due to random chance alone) and the randomization of our survey removes the effect of confounding variables, we conclude that a statistically significant difference does imply evidence for a causal effect.

---

<sup>1</sup>as we discussed, the plus-four test can also be used for this problem.

3. a) The linear model I fit was the log of the price as the explanatory variable with the log of the carat size and the color of the diamond as the response variables (I found that leaving out the clarity as a response variable helped the model satisfy the Normal linear regression assumptions better in terms of the homoscedasticity and normality of residuals).

This model satisfies the Normal linear regression assumptions relatively well; the residuals are approximately normal distributed (QQPlot roughly linear), there seems to be a linear relationship in the data, independence and normality of residuals, and the homoscedasticity assumption is satisfied (roughly equal variance across fitted values).

- b) The diamond that appears to be priced the lowest, given its characteristics is the one that has the largest negative residual in our linear model, which is Diamond 11 (see R output). The diamonds that appear to be priced the lowest given their characteristics include diamonds 11, 18, 22, 14, and 3 under this model.
- c) I calculated the 95% prediction interval for the price of a 0.5 carat ideal-cut, internally flawless diamond with color D in two different ways (see R code). They match and give a 95% prediction interval around  $[1350, 3586]$ .

4. a) For  $X_1, \dots, X_n$  i.i.d  $\text{Unif}(0, \theta)$ , we have that

$$\begin{aligned} L(X_1, \dots, X_n | \theta) &= \prod_{i=1}^n \frac{1}{\theta} I(0 < X_i \leq \theta) \\ &= \left(\frac{1}{\theta}\right)^n \prod_{i=1}^n I(0 < X_i \leq \theta) \\ &= \frac{1}{\theta^n} I(\max(X_i) \leq \theta). \end{aligned}$$

The log-likelihood is thus equal to

$$l(\theta) = \begin{cases} -n \ln \theta & \text{if } \max(X_i) \leq \theta \\ \text{undefined} & \text{otherwise} \end{cases},$$

Since  $-n \ln \theta$  is a decreasing function with respect to  $\theta$ , this is maximized for smaller values of  $\theta$ , so

$$\boxed{\hat{\theta} = \max(X_i)}$$

is the maximum likelihood estimate.

To find the distribution of  $\frac{\hat{\theta}}{\theta}$ , we consider its pdf. Note that

$$\Pr\left(\frac{\hat{\theta}}{\theta} \leq x\right) = \Pr(\hat{\theta} \leq x\theta) = \Pr(\max(X_i) \leq x\theta).$$

Note that each  $X_i \sim \text{Unif}(0, \theta)$ , so we have that

$$\begin{aligned} \Pr(\max(X_i) \leq x\theta) &= I(0 < x \leq 1) \left(\frac{x\theta}{\theta}\right)^n \\ &= I(0 < x \leq 1) x^n. \end{aligned}$$

Taking the derivative of this CDF to get the pdf, we get that

$$f_{\frac{\hat{\theta}}{\theta}}(x) = I(0 < x \leq 1) n x^{n-1}$$

which is precisely the pdf for a  $\text{Beta}(n, 1)$  random variable (as  $\frac{\Gamma(n+1)}{\Gamma(n)\Gamma(1)} = n$ ). Thus,  $\frac{\hat{\theta}}{\theta} \sim \text{Beta}(n, 1)$ , as desired.

- b) The significance level of the test of  $H_0: \theta = 100$  vs.  $H_1: \theta \neq 100$  that rejects when  $\hat{\theta} > 100$  is

$$\alpha = \Pr(\text{reject } H_0 \mid H_0 \text{ true}) = \Pr(\hat{\theta} > 100 \mid \theta = 100) = \boxed{0}$$

(logically, the maximum of the  $X_i$ 's cannot be greater than 100 if each  $X_i$  is between 0 and 100).

On the other hand, the power of the test is

$$\begin{aligned} \Pr(\text{reject } H_0 \mid H_1 \text{ true}) &= \Pr(\hat{\theta} > 100 \mid \theta = 200) \\ &= 1 - \prod_{i=1}^n \Pr(X_i < 100 \mid \theta = 200) \\ &= \boxed{1 - \frac{1}{2^n}}. \end{aligned}$$

- c) The power of the test for  $H_0: \theta = 100$  vs.  $H_1: \theta > 100$ , with rejection when  $\hat{\theta} > 100$ , cannot be computed as it depends on the true value of the parameter  $\theta$ ; if  $\hat{\theta} > \theta$ , then the power is 0, otherwise it would be  $1 - \left(\frac{100}{\theta}\right)^n$ .

However, we do know that the test is the most powerful for every simple alternative in  $H_1$ ; thus, the Neyman-Pearson Lemma can be extended and we know that the test, even though its power cannot be computed, is the uniformly most powerful test of level  $\alpha$ .

- d) For the test of  $H_0: \theta = 100$  vs.  $H_1: \theta \neq 100$ , the generalized likelihood ratio test statistic  $\Lambda$  is

$$\begin{aligned}\Lambda &= \frac{\max_{\theta \in H_0} L(X_1, \dots, X_n | \theta)}{\max_{\theta \in H_1} L(X_1, \dots, X_n | \theta)} \\ &= \frac{\left(\frac{1}{100}\right)^n I(\max(X_i) \leq 100)}{\left(\frac{1}{\hat{\theta}}\right)^n}\end{aligned}$$

where  $\hat{\theta} = \max(X_i)$  and the numerator comes from the fact that  $\theta = 100$  under  $H_0$ . Simplifying, we get that

$$\Lambda = \left(\frac{\max(X_i)}{100}\right)^n I(\max(X_i) \leq 100).$$

- e) Note that if  $X \sim \text{Unif}(0, \theta)$ , then  $f_X(x) = \frac{1}{\theta} I(0 < x \leq \theta)$ . This does not follow the form of the exponential family of distributions, due to the indicator variable (equivalently, because the support depends on the parameter  $\theta$ ). Consequently,  $-2 \log \Lambda$  may not have null sampling distribution approximately  $\chi^2_{(1)}$ .

We will show that  $W = -2 \log \Lambda \sim \chi^2_{(2)}$  by identifying the CDF of  $W$ . Note that this is only defined when  $\max(X_i) \leq 100$ , in which case  $\Lambda \leq 1$ , so  $\log \Lambda \leq 0$  and  $W \geq 0$ .

Let  $Y = \frac{\hat{\theta}}{\theta} = \frac{\max(X_i)}{\theta}$ . Consider  $\Pr(W \leq w) = \Pr(-2 \log \Lambda \leq w)$ . Since  $\theta = 100$  under the null hypothesis, and  $\Lambda = \left(\frac{\max(X_i)}{100}\right)^n$ , we know that  $\Lambda = Y^n$ . Plugging this in, we get that

$$\begin{aligned}\Pr(W \leq w) &= \Pr(-2 \log \Lambda \leq w) \\ &= \Pr(-2 \log(Y^n) \leq w) \\ &= \Pr\left(\log(Y^n) \geq -\frac{1}{2}w\right) \\ &= \Pr(Y^n \geq e^{-\frac{1}{2}w}) \\ &= \Pr(Y \geq e^{-\frac{w}{2n}}) \\ &= 1 - \Pr(Y \leq e^{-\frac{w}{2n}}).\end{aligned}$$

From part (a), we know that  $Y \sim \text{Beta}(n, 1)$ , so the Beta CDF tells us that  $\Pr(Y \leq$

$y) = y^n I(0 < y \leq 1)$ . It follows that

$$\begin{aligned}\Pr(W \leq w) &= 1 - \Pr(Y \leq e^{-\frac{w}{2n}}) \\ &= 1 - (e^{-\frac{w}{2n}})^n \\ &= 1 - e^{-\frac{1}{2}w}.\end{aligned}$$

This is precisely the CDF for an  $\text{Expo}(\frac{1}{2})$  or  $\chi^2_{(2)}$  random variable, so we know that  $W = -2 \log \Lambda \sim \chi^2_{(2)}$  as desired.

*This is my last ever undergraduate assignment. Thanks for a great semester, Phil!*