Estimation & inference: more examples

Lecture 17b (STAT 24400 F24)

Review example

Example (review of various inference procedures and tests)

A movie producer wants to determine the effectiveness of a movie trailer.
Will a person pay to see the movie after watching the trailer?

To study this question, you recruit 230 Amazon Prime subscribers from Illinois:

- 100 from urban areas
- 80 from suburban areas
- 50 from rural areas

Suppose that you have chosen these subjects randomly from the population of Amazon Prime members in each type of region.

The measured data

When watching the trailer —

- Among the 100 individuals sampled from urban areas,
  - Time on the Amazon website (in minutes): mean 11, SD 8
  - 28 individuals pay to buy the movie
- Among the 80 individuals sampled from suburban areas,
  - Time on the Amazon website (in minutes): mean 17, SD 8
  - 24 individuals pay to buy the movie
- Among the 50 individuals sampled from rural areas,
  - Time on the Amazon website (in minutes): mean 6, SD 9
  - 30 individuals pay to buy the movie
- In total among all 100+80+50=230 individuals,
  - Time on the Amazon website (in minutes): mean 12, SD 9
  - 82 individuals pay to buy the movie

Question 1: mean time on website and C.I.

What is a 90% confidence interval for the mean time spent on the website for the population of suburban Amazon Prime members?

- The 80 observations from suburban areas are an i.i.d. sample from this population (of suburban Amazon Prime members)
- Within this sample, we have $\bar{X} = 17$ and $S = 8$

A 90% confidence interval for the true mean μ:

$$\bar{X} \pm t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

$\leftarrow S = 8$

$\leftarrow n = 80$

$\bar{X} = 17$ $\quad t_{79,0.05} = 1.664$

$$\rightsquigarrow \quad 17 \pm 1.664 \cdot \frac{8}{\sqrt{80}} = 17 \pm 1.489 = [15.511, 18.489]$$

Question 2: probability of buying the movie

What is a 90% confidence interval (based on Fisher information) for the proportion of urban Amazon Prime members who would pay to watch the movie?

- The 100 observations from urban areas are an i.i.d. sample from this population
- The distribution: $X_1, \ldots, X_{100} \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, where p = true prob.
- MLE $= \hat{p} = \frac{28}{100}$, the sample proportion who bought the movie
- Fisher information (for one obs.) is $\mathcal{I}(p) = \frac{1}{p(1-p)}$

A 90% confidence interval for the true parameter p:

$$\hat{p} \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n\mathcal{I}(\hat{p})}}$$

$\hat{p} = 0.28$ $z_{0.05} = 1.645$ $n = 100$ $\mathcal{I}(\hat{p}) = \frac{1}{0.28(1-0.28)}$

$$\rightsquigarrow \quad 0.28 \pm 1.645 \cdot \sqrt{\frac{0.28(1-0.28)}{100}} = 0.28 \pm 0.074 = [0.206, 0.354]$$

Question 2: probability of buying the movie (cont.)

Remarks

In the above we used the asymptotic distr. of the MLE $\hat{p}$ to construct a C.I.

Recall if i.i.d. $X_1, \cdots, X_n \sim f(\cdot|\theta_0)$ and $\hat{\theta}$ is the MLE, then for large n,

$$\hat{\theta} \approx N\left(\theta_0, \frac{1}{n\mathcal{I}(\theta_0)}\right), \qquad \sqrt{n\mathcal{I}(\theta_0)}(\hat{\theta} - \theta_0) \to N(0,1)$$

where the convergence is in distribution. More usefully,

$$\sqrt{n\mathcal{I}(\hat{\theta})}(\hat{\theta} - \theta_0) \to N(0,1)$$

which gives us an asymptotic level α C.I. for θ_0,

$$\hat{\theta} \pm z_{\alpha/2} \frac{1}{\sqrt{n\mathcal{I}(\hat{\theta})}}$$

Question 3: comparing urban & rural

Suppose that we gathered this data with the aim of testing

H_0 : rural Amazon Prime members are three times as likely as urban Amazon Prime members to pay for the movie

- The observed counts:

	Yes	No	Total
Rural	30	20	50
Urban	28	72	100

- The probabilities (3 degrees of freedom):

	Yes	No	
Rural	p_{RY}	p_{RN}	$\leftarrow$ let $r = p_{\text{RY}} + p_{\text{RN}}$ = proportion that's rural
Urban	p_{UY}	p_{UN}	$\leftarrow$ $1 - r = p_{\text{UY}} + p_{\text{UN}}$ = proportion that's urban

- According to H_0: (2 degrees of freedom)

$$\frac{p_{\text{RY}}}{p_{\text{RY}} + p_{\text{RN}}} = 3 \cdot \frac{p_{\text{UY}}}{p_{\text{UY}} + p_{\text{UN}}} = 3p \quad \rightsquigarrow$$

	Yes	No
Rural	$3pr$	$(1-3p)r$
Urban	$p(1-r)$	$(1-p)(1-r)$

Question 3: comparing urban & rural (cont.)

Finding the MLE under H_0:

$$\text{Likelihood} = \frac{150!}{30!20!28!72!} \cdot (p_{RY})^{30}(p_{RN})^{20}(p_{UY})^{28}(p_{UN})^{72}$$

$$= \frac{150!}{30!20!28!72!} \cdot (3pr)^{30}((1-3p)r)^{20}(p(1-r))^{28}((1-p)(1-r))^{72}$$

$$= (\text{constant}) \cdot \underbrace{p^{30+28}(1-3p)^{20}(1-p)^{72}}_{\text{maximize over } p} \cdot \underbrace{r^{30+20}(1-r)^{28+72}}_{\text{maximize over } r}$$

maximize $58\log(p) + 20\log(1-3p) + 72\log(1-p)$

deriv. $= \frac{58}{p} - \frac{60}{1-3p} - \frac{72}{1-p} = 0 \quad \rightsquigarrow \hat{p} = 0.2182$

maximize $50\log(r) + 100\log(1-r)$

deriv. $= \frac{50}{r} - \frac{100}{1-r} = 0 \quad \rightsquigarrow \hat{r} = \frac{1}{3}$

Question 3: comparing urban & rural (cont.)

The MLE under H_0:

$$\hat{p} = 0.2182, \qquad \hat{r} = \frac{1}{3}$$

	Yes	No
Rural	$\hat{p}_{\mathrm{RY}} = 3\hat{p}\hat{r} = 0.2182$	$\hat{p}_{\mathrm{RN}} = (1-3\hat{p})\hat{r} = 0.1151$
Urban	$\hat{p}_{\mathrm{UY}} = \hat{p}(1-\hat{r}) = 0.1419$	$\hat{p}_{\mathrm{UN}} = (1-\hat{p})(1-\hat{r}) = 0.5248$

The expected counts under H_0:

	Yes	No
Rural	$150 \cdot 0.2182 = 32.73$	$150 \cdot 0.1151 = 17.26$
Urban	$150 \cdot 0.1419 = 21.29$	$150 \cdot 0.5248 = 78.72$

Question 3: comparing urban & rural (cont.)

- The observed counts:

	Yes	No	Total
Rural	30	20	50
Urban	28	72	100

- The expected counts under H_0:

	Yes	No
Rural	32.73	17.26
Urban	21.29	78.72

Run Pearson's χ^2 test at level $\alpha = 0.05$:

$$X^2 = \frac{(30-32.72)^2}{32.72} + \frac{(20-17.26)^2}{17.26} + \frac{(28-21.29)^2}{21.29} + \frac{(72-78.72)^2}{78.72} = 3.3495$$

$\rightsquigarrow$ p-value $= 1 - F_{\chi^2_{3-2}}(3.3495) = 0.0672 \Rightarrow$ do not reject H_0 (at level $\alpha = 0.05$)

Question 3: comparing urban & rural (cautionary note)

Suppose that instead we had invented this hypothesis after examining the gathered data. What might be the issue?

- It is not valid to test hypotheses that were chosen as a result of the outcomes of the experiment/study —
  we must decide on our testing procedure (hypotheses/questions being asked, which test to run, etc) before observing the data
- This is related to the multiple testing problem — there are many possible questions we might choose to ask after observing the data

Question 4: exponential distribution of time

Assume that the time spent by an individual on the Amazon site is exponentially distributed, with rate λ_{U}, λ_{S}, λ_{R} for the urban, suburban, or rural populations, respectively.

Are these rates all the same?

$$H_0 : \lambda_{\mathrm{U}} = \lambda_{\mathrm{S}} = \lambda_{\mathrm{R}} \quad \text{vs} \quad H_1\text{: these equalities don't all hold}$$

We will run a generalized LRT at level $\alpha = 0.05$.

$$\text{Likelihood} = \prod_{i=1}^{100} \lambda_{\mathrm{U}} e^{-\lambda_{\mathrm{U}} X_i^{\mathrm{U}}} \cdot \prod_{i=1}^{80} \lambda_{\mathrm{S}} e^{-\lambda_{\mathrm{S}} X_i^{\mathrm{S}}} \cdot \prod_{i=1}^{50} \lambda_{\mathrm{R}} e^{-\lambda_{\mathrm{R}} X_i^{\mathrm{R}}}$$
$$= \lambda_{\mathrm{U}}^{100} e^{-\lambda_{\mathrm{U}} \cdot 100 \cdot 11} \cdot \lambda_{\mathrm{S}}^{80} e^{-\lambda_{\mathrm{S}} \cdot 80 \cdot 17} \cdot \lambda_{\mathrm{R}}^{50} e^{-\lambda_{\mathrm{R}} \cdot 50 \cdot 6}$$

Question 4: exponential distribution of time (cont.)

$$\Lambda = \frac{\text{Maximum likelihood under the constraint } \lambda_U = \lambda_S = \lambda_R}{\text{Maximum likelihood over any } \lambda_U, \lambda_S, \lambda_R > 0}$$

- Maximize the likelihood under constraint $\lambda_U = \lambda_S = \lambda_R$:

$$\text{Likelihood} = \underbrace{\lambda^{100} e^{-\lambda \cdot 100 \cdot 11} \cdot \lambda^{80} e^{-\lambda \cdot 80 \cdot 17} \cdot \lambda^{50} e^{-\lambda \cdot 50 \cdot 6}}_{\text{MLE } \hat{\lambda} = \frac{1}{12}}$$

- Maximize likelihood over any $\lambda_U, \lambda_S, \lambda_R > 0$:

$$\text{Likelihood} = \underbrace{\lambda_U^{100} e^{-\lambda_U \cdot 100 \cdot 11} \cdot \lambda_S^{80} e^{-\lambda_S \cdot 80 \cdot 17} \cdot \lambda_R^{50} e^{-\lambda_R \cdot 50 \cdot 6}}_{\text{MLE } \hat{\lambda}_U = \frac{1}{11},\ \hat{\lambda}_S = \frac{1}{17},\ \hat{\lambda}_R = \frac{1}{6}}$$

$\Lambda = 1.87 \times 10^{-7} \rightsquigarrow -2\log(\Lambda) = 30.99$

$\rightsquigarrow$ p-value $= 1 - F_{\chi^2_{3-1}}(30.99) = 1.87 \times 10^{-7} \quad \Rightarrow$ reject H_0

Question 5: Bayesian inference

Let θ = population proportion of suburban Amazon Prime members that would buy the movie after watching the trailer.

Suppose that we placed a prior distribution on θ with the density

$$g(t) = \frac{\pi}{2}\sin(\pi t), \quad 0 \le t \le 1$$

- Calculate the posterior density of θ after observing the data
- Calculate the posterior mean

(No need to simplify—we will have integral expressions etc)

Question 5: Bayesian inference (cont.)

Hierarchical model:

$$\begin{cases} \theta \sim \text{distrib. with density } g(t) = \frac{\pi}{2}\sin(\pi t) \text{ on } t \in [0,1] \\ X \mid \theta \sim \text{Binomial}(80, \theta) \end{cases}$$

The posterior distribution of θ given the data $X = 24$:

$$h_{\theta|X}(t \mid 24) = (\text{normalizing const.}) \cdot \underbrace{\frac{\pi}{2}\sin(\pi t)}_{\text{prior}} \cdot \underbrace{\binom{80}{24} t^{24}(1-t)^{56}}_{\text{likelihood}},\ 0 \le t \le 1$$

We can calculate the constant since posterior density must integrate to 1:

$$h_{\theta|X}(t \mid 24) = \frac{\sin(\pi t) \cdot t^{24}(1-t)^{56}}{\int_{s=0}^{1} \sin(\pi s) \cdot s^{24}(1-s)^{56}\, ds}$$

$$\Rightarrow \text{ posterior mean } = \mathbb{E}(\theta \mid X = 24) = \int_{t=0}^{1} t \cdot \frac{\sin(\pi t) \cdot t^{24}(1-t)^{56}}{\int_{s=0}^{1} \sin(\pi s) \cdot s^{24}(1-s)^{56}\, ds}\, dt$$