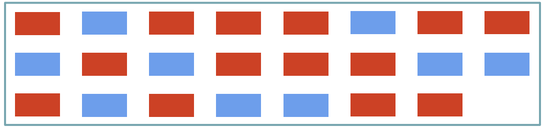
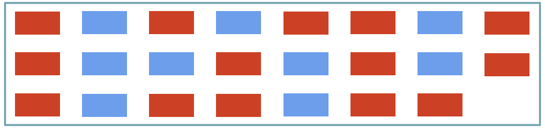
Analyzing categorical data

Chapter 13 of Rice

07/20/2021

In the previous lecture,





If F=G, every assignment of ranks to the pooled observations is equally likely,

- Comparing two independent samples:
  - Under Normal assumption, we can use *t* test.
  - Without any distributional assumption, use Mann-Whitney test:

$$U_1 = mn + \frac{n(n+1)}{2} - R_1$$
$$U_2 = mn + \frac{m(m+1)}{2} - R_2$$

U statistic is $U = \min\{U_1, U_2\}$, and rejection region is $R = \{U \le c\}$.

- Comparing paired samples:
  - Under Normal assumption, we can use *t* test.

$$\text{Under } H_0,\ T(\mathbf{X}_n) = \frac{\bar{D}_n}{\sqrt{\frac{1}{n} S_D^2}} \sim t_{n-1}.$$

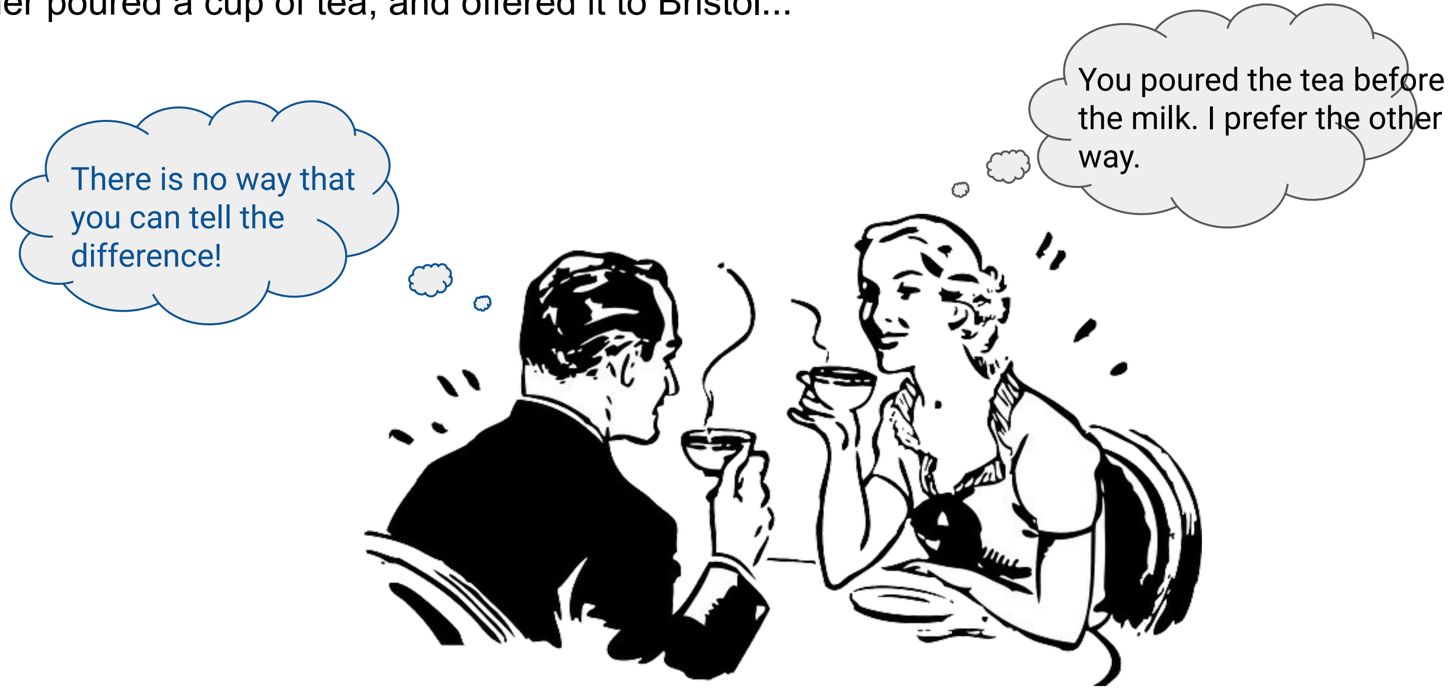
  - Without any distributional assumption, use Wilcoxon ranked sum test:

$$\begin{aligned} W_+ &= \text{positive rank sum}, \\ W_- &= \text{negative rank sum}, \\ \mathrm{R} &= \{W_+ \le c \text{ or } W_- \le c\} \text{ for two-sided hypothesis.} \end{aligned}$$

  - `wilcox.test(x,y, paired=TRUE)`

Tea-tasting experiment

In 1919, Fisher poured a cup of tea, and offered it to Bristol...



Ronald Fisher Dr. Muriel Bristol

H_0: Bristol does not know how to choose the cups prepared by the 2nd method.

Tea-tasting experiment

4 prepared by first pouring the tea, then adding milk

4 prepared by first pouring the milk, then adding tea

Randomly sort the cups and serve to Bristol, and let her choose 4 that were prepared by the second method.

$$\binom{8}{4} = \underbrace{70 \text{ combinations}}$$

Tea-tasting experiment

H_0 : Bristol has no skills in determining the order.

→ She has correct skill

→ She has a skill resulting in totally opposite classification

Success count	Combinations of selection	Number of Combinations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
	Total	70

$\binom{4}{4}$

$\binom{4}{3}\binom{4}{1}$

$\binom{4}{2}\binom{4}{2}$

$\binom{4}{1}\binom{4}{3}$

$\binom{4}{4}$

Hypergeometric distribution:

$$P(X=k)=\frac{\binom{n}{k}\binom{m}{r-k}}{\binom{n+m}{r}},\ k=0,\ \ldots,\ r.$$

Bristol correctly selected out all 4 cups of the second method.

$$p\text{-value} = P(\text{count} \geq 4 \,|\, H_0) + P(\text{count} \leq 0 \,|\, H_0) = \frac{2}{70} \approx 0.028.$$

< 0.05

Binomial distribution

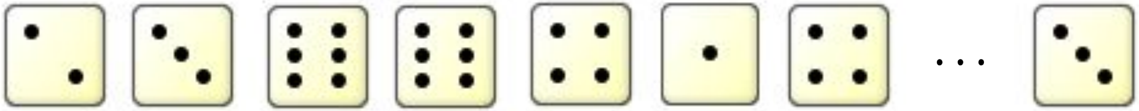
n coin flips

p captures the fairness of the coin

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

number of heads

Multinomial distribution



n tosses

p_i is the probability that i comes up

$p_1 + \cdots + p_6 = 1$

$$P(Y_1 = k_1, \ \ldots, \ Y_6 = k_6) = \frac{n!}{k_1! \cdots k_6!} p_1^{k_1} \cdots p_6^{k_6}, \ \text{with } k_1 + \cdots + k_6 = n.$$

number of i in the n tosses

$EY_i = np_i$, $i = 1, 2, \cdots 6$.

LRT for multinomial distribution

9.5 of Rice

07/20/2021

LRT for multinomial distribution

Example 1. Let $\theta = (p_1, \ldots, p_6)$, and $X_1, \ldots, X_n$ are the results of n tosses. Consider testing

$$H_0 : p_1 = p_2,\ p_3 = p_4 = p_5 = p_6 \ \text{ versus } \ H_1 : H_0 \text{ is not true.}$$

Solution. Denote $Y_j = \#$ of observations equal to j.

We know that $(Y_1, \cdots Y_6) \sim \text{multinomial}(p_1, \cdots p_6)$

$$L(p_1, \cdots p_6 \mid Y_1, \cdots Y_6) = \frac{n!}{Y_1! \cdots Y_6!} p_1^{Y_1} \cdots p_6^{Y_6}.$$

$$\Theta_0 = \left\{ p_1 = p_2,\ p_3 = p_4 = p_5 = p_6,\ \sum_{i=1}^{6} p_i = 1 \right\}, \quad \Theta = \Theta_0 \cup \Theta_1 = \left\{ \sum_{i=1}^{6} p_i = 1 \right\}$$

$$2p_1 + 4p_3 = 1, \quad p_3 = \frac{1 - 2p_1}{4}. \qquad p_6 = 1 - p_1 - \cdots - p_5$$

$\dim \Theta_0 = 1 \qquad \dim \Theta = 5$

$\nu = \dim \Theta - \dim \Theta_0 = 4.$

$$\sup_{\Theta} L(p_1, \cdots p_6 \mid Y_1, \cdots Y_6) = \sup_{p_1, \cdots p_5} \frac{n!}{Y_1! \cdots Y_6!} \underbrace{p_1^{Y_1} \cdots p_5^{Y_5} (1 - p_1 - \cdots - p_5)^{Y_6}}_{\text{denote } L_0(p_1, \cdots p_5)}$$

Then $\ell_0(p_1, \cdots p_5) = \log L_0(p_1, \cdots p_5) = Y_1 \log p_1 + \cdots + Y_5 \log p_5 + Y_6 \log(1 - p_1 - \cdots - p_5)$

LRT for multinomial distribution

Solution cont'd. Take the derivatives with respect to $p_1, \dots, p_5$:

$$
\left\{
\begin{aligned}
&\frac{\partial l_0}{\partial p_1} = \frac{Y_1}{p_1} - \frac{Y_6}{1-p_1-\dots-p_5} = 0 \quad \longrightarrow \quad Y_1(1-p_1-\dots-p_5) - Y_6 \cdot p_1 = 0 \\
&\quad \vdots \\
&\frac{\partial l_0}{\partial p_5} = \frac{Y_5}{p_5} - \frac{Y_6}{1-p_1-\dots-p_5} = 0 \quad \longrightarrow \quad Y_5(1-p_1-\dots-p_5) - Y_6 \, p_5 = 0
\end{aligned}
\right.
$$

$$\hat{p}_1 = \frac{Y_1(1-p_1-\dots-p_5)}{Y_6} = \frac{Y_1 p_6}{Y_6} = \frac{Y_1}{Y_6} \cdot \frac{Y_6}{Y_1+\dots+Y_6} = \frac{Y_1}{Y_1+\dots+Y_6}$$

$$\hat{p}_2 = \frac{Y_2}{Y_1+\dots+Y_6}$$

$$\hat{p}_5 = \frac{Y_5}{Y_1+\dots+Y_6}$$

Sum over these 5 equations:

$$(Y_1+\dots+Y_5)(1-p_1-\dots-p_5) - Y_6(p_1+\dots+p_5) = 0$$

$$\Longleftrightarrow \; Y_1+\dots+Y_5 - (p_1+\dots+p_5)(Y_1+\dots+Y_5) - (p_1+\dots+p_5)Y_6$$

$$= (Y_1+\dots+Y_5) - (p_1+\dots+p_5)(Y_1+\dots+Y_6) = 0 \quad \Longleftrightarrow \quad p_1+\dots+p_5 = \frac{Y_1+\dots+Y_5}{Y_1+\dots+Y_6}$$

$$\Longleftrightarrow \; \hat{p}_6 = 1 - \frac{Y_1+\dots+Y_5}{Y_1+\dots+Y_6} = \boxed{\frac{Y_6}{Y_1+\dots+Y_6}}$$

Then $\sup_{H_1} L(p_1, \dots, p_6 \mid Y_1, \dots, Y_6) = \dfrac{n!}{Y_1! \cdots Y_6!} \hat{p}_1^{Y_1} \cdots \hat{p}_6^{Y_6}$.

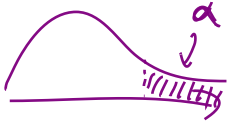
$$\sup_{H_0} L(p_1,\dots,p_6 \mid Y_1\dots Y_6) = \sup_{\substack{p_1=p_2\\ p_3=\dots=p_6\\ 2p_1+4p_3=1}} L(p_1,\dots,p_6 \mid Y_1\dots Y_6) = \sup_{p_1} \frac{n!}{Y_1!\cdots Y_6!}\, p_1^{Y_1} p_1^{Y_2}\, p_3^{Y_3} p_3^{Y_4} p_3^{Y_5} p_3^{Y_6}$$

$$p_3 = \frac{1-2p_1}{4}$$

$$= \sup_{p_1} \frac{n!}{Y_1!\cdots Y_6!}\, \underbrace{p_1^{Y_1+Y_2}\left(\frac{1-2p_1}{4}\right)^{Y_3+\dots+Y_6}}_{L_0(p_1)}$$

$$\ell_0(p_1) = \log L_0(p_1) = (Y_1+Y_2)\log p_1 + (Y_3+\dots+Y_6)\log\left(\frac{1-2p_1}{4}\right)$$

$$\frac{\partial \ell_0(p_1)}{\partial p_1} = \frac{Y_1+Y_2}{p_1} - 2\,\frac{Y_3+\dots+Y_6}{\frac{1-2p_1}{4}} = 0 \Rightarrow 2\tilde{p}_1 = \frac{Y_1+Y_2}{Y_1+\dots+Y_6}$$



$$\tilde{p}_1 = \frac{(Y_1+Y_2)/2}{Y_1+\dots+Y_6}$$

$$\sup_{H} L(p_1,\dots,p_6 \mid Y_1\dots Y_6) = \frac{n!}{Y_1!\cdots Y_6!}\, \tilde{p}_1^{\,Y_1+Y_2}\, \tilde{p}_3^{\,Y_3+\dots+Y_6}$$

$$\tilde{p}_3 = \frac{(Y_3+\dots+Y_6)/4}{Y_1+\dots+Y_6}$$

$$\lambda(X_n) = \frac{\sup_{H_0} L}{\sup_{H} L} = \left(\frac{\tilde{p}_1}{\hat{p}_1}\right)^{Y_1}\left(\frac{\tilde{p}_1}{\hat{p}_2}\right)^{Y_2}\cdots\left(\frac{\tilde{p}_3}{\hat{p}_6}\right)^{Y_6}$$

$$R = \{\lambda(X_n) \le c\} = \{-2\log\lambda(X_n) \ge \chi^2_4(\alpha)\}$$

$$-2\log\lambda(X_n) = 2\sum_{i=1}^{6} -\log\left(\frac{\tilde{p}_i}{\hat{p}_i}\right)^{Y_i} = 2\sum_{i=1}^{6} Y_i \log\frac{\hat{p}_i}{\tilde{p}_i} \xrightarrow{d} \chi^2_{\nu} = \chi^2_4$$

$$-2\log\lambda(\Sigma_n) = 2\sum_{i=1}^{6} Y_i \log\frac{\hat{p}_i}{\tilde{p}_i} = 2\sum_{i=1}^{6} Y_i \log\frac{n\hat{p}_i}{n\tilde{p}_i} = 2\sum_{i=1}^{6} O_i \log\frac{O_i}{E_i} \leftarrow$$

$$Y_1 + \cdots + Y_6 = n$$

observed count

$$\hat{p}_i = \frac{Y_i}{Y_1+\cdots+Y_6} = \frac{Y_i}{n} \Rightarrow n\hat{p}_i = Y_i = O_i$$

expected count under H_0

$$\tilde{p}_1 = \frac{(Y_1+Y_2)/2}{Y_1+\cdots+Y_6} \Rightarrow n\tilde{p}_1 = \frac{Y_1+Y_2}{2} = E_1$$

$$\tilde{p}_3 = \frac{(Y_3+\cdots+Y_6)/4}{Y_1+\cdots+Y_6} \Rightarrow n\tilde{p}_j = \frac{Y_3+\cdots+Y_6}{4} = E_3$$

$$E_1 = E_2, \quad E_3 = \cdots = E_6$$

LRT for multinomial distribution

Example 2. Let $\theta = (p_1, \ldots, p_4)$, and $X_1, \ldots, X_n$ are the outcomes of n experiments. Consider testing

$$H_0 : p_1 = 2\theta,\ p_2 = \theta,\ p_3 = 3\theta,\ p_4 = 1 - 6\theta \ \text{ versus }\ H_1 : H_0 \text{ is not true.}$$

in which $\theta \in (0, 1/6)$.

Solution. Denote $Y_j = \#$ of observations equal to j.

$$L(p_1, \ldots p_4 | Y_1, \ldots Y_4) = \frac{n!}{Y_1! \cdots Y_4!} p_1^{Y_1} \cdots p_4^{Y_4}$$

$$\Theta_0 = \{\underbrace{p_1 = 2\theta, p_2 = \theta, p_3 = 3\theta, p_4 = 1-6\theta}_{\dim \Theta_0 = 1}\}$$

$$\Theta = \Theta_0 \cup \Theta_1 = \{\sum_{i=1}^{4} p_i = 1\}$$

$$\rightarrow \nu = \dim \Theta - \dim \Theta_0 = 2 \leftarrow \dim \Theta = 4 - 1 = 3.$$

$$\sup_{\Theta} L(p_1, \ldots p_4 | Y_1, \ldots Y_4) = \frac{n!}{Y_1! \cdots Y_4!} \hat{p}_1^{Y_1} \cdots \hat{p}_4^{Y_4} \quad \text{where } \hat{p}_i = \frac{Y_i}{n}, \ i = 1, 2, 3, 4.$$

$$\sup_{\Theta_0} L(p_1, \ldots p_4 | Y_1, \ldots Y_4) = \sup_{\theta \in (0, 1/6)} \frac{n!}{Y_1! \cdots Y_4!} (2\theta)^{Y_1} \theta^{Y_2} (3\theta)^{Y_3} (1-6\theta)^{Y_4}$$

$$= \sup_{\theta \in (0, 1/6)} \frac{n!\, 2^{Y_1} 3^{Y_1}}{Y_1! \cdots Y_4!} \underbrace{\theta^{Y_1+Y_2+Y_3} (1-6\theta)^{Y_4}}_{L_0(\theta)}$$

$$\ell_0(\theta) = \log L_0(\theta) = (Y_1+Y_2+Y_3)\log(\theta) + Y_4 \log(1-6\theta)$$

$$\frac{\partial \ell_0}{\partial \theta} = \frac{Y_1+Y_2+Y_3}{\theta} - 6\frac{Y_4}{1-6\theta} = 0 \Rightarrow \hat{\theta} = \frac{(Y_1+Y_2+Y_3)/6}{n}. \leftarrow$$

$$\sup_{\Theta_0} L(P_1,\dots P_4 \mid Y_1,\dots Y_4) = \underbrace{\frac{n!}{Y_1!\cdots Y_4!}}\ \underbrace{(2\hat{\theta})}_{P_1(\hat{\theta})}{}^{Y_1}\ \underbrace{\hat{\theta}}_{P_2(\hat{\theta})}{}^{Y_2}\ \underbrace{(3\hat{\theta})}_{P_3(\hat{\theta})}{}^{Y_3}\ \underbrace{(1-6\hat{\theta})}_{P_4(\hat{\theta})}{}^{Y_4}$$

$$\lambda(\bar{x}_n) = \frac{\sup_{\Theta_0} L}{\sup_{\Theta} L} = \left(\frac{P_1(\hat{\theta})}{\hat{P}_1}\right)^{Y_1} \cdots \left(\frac{P_4(\hat{\theta})}{\hat{P}_4}\right)^{Y_4}$$

$$-2\log\lambda(\bar{x}_n) = 2\sum_{i=1}^{4} Y_4 \log\frac{\hat{P}_1}{P_1(\hat{\theta})} = 2\sum_{i=1}^{4} O_i \log\frac{O_i}{E_i}.$$

$$\xrightarrow{d} \chi^2_\nu = \chi^2_2.$$

$$O_i = Y_i$$

$$E_i = n\,P_i(\hat{\theta})$$

LRT for multinomial distribution

Theorem A. Let $X_1, \ldots, X_n$ be the results of n experiments. Each experiment has m possible outcomes. Consider testing

$$H_0 : (p_1, \ldots, p_m) = (p_1(\theta), \ldots, p_m(\theta)) \text{ versus } H_1 : H_0 \text{ is not true.}$$

Then the likelihood ratio under the null hypothesis satisfies

$$-2\log\lambda(\mathbf{X}_n) = 2\sum_{i=1}^{m} O_i \log\frac{O_i}{E_i} \xrightarrow{d} \chi^2_{m-2}, \text{ as } n \to \infty.$$

$O_j = \#$ of observations equal to j.

$\hat{\theta}$ = the MLE in H_0.

$E_i = np_i\left(\hat{\theta}\right)$

$\Theta = \{ \sum_{i=1}^{m} p_i = 1 \}$ $\quad$ $\dim \Theta = m-1$

$\Theta_0 = \{ p_i = p_i(\theta) \}$ $\quad$ $\dim \Theta_0 = 1$

LRT for multinomial distribution

Corollary A. Let $X_1, \ldots, X_n$ be the results of n experiments. Each experiment has m possible outcomes. Consider testing

$$H_0 : (p_1, \ldots, p_m) = (p_1(\theta), \ldots, p_m(\theta)) \text{ versus } H_1 : H_0 \text{ is not true.}$$

Then the likelihood ratio under the null hypothesis satisfies $\rightarrow$ χ^2 statistic

$$2\sum_{i=1}^{m} O_i \log\frac{O_i}{E_i} \approx \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{d} \chi^2_{m-2}, \text{ as } n \to \infty.$$

$O_j = \#$ of observations equal to j.

$\hat{\theta}$ = the MLE in H_0.

$E_i = np_i(\hat{\theta})$

$$f(x) = \underbrace{x \log\frac{x}{x_0}}, \quad f'(x) = 1 + \log\frac{x}{x_0}, \quad f''(x) = \frac{1}{x}$$

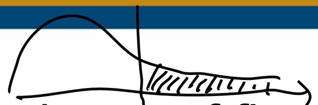
$$\underbrace{f(x) - f(x_0)}_{=0} \approx \underbrace{f'(x_0)}_{=1}(x - x_0) + \underbrace{\frac{f''(x_0)}{2}}_{=\frac{1}{2x_0}}(x - x_0)^2$$

$$\sum_{i=1}^{m} O_j = n$$

$$\sum_{i=1}^{m} E_i = \sum_{i=1}^{m} n p_i(\hat{\theta}) = n \underbrace{\sum_{i=1}^{m} p_i(\hat{\theta})}_{=1} = n$$

$$\Rightarrow \underbrace{f(O_i)}_{= O_i \log\frac{O_i}{E_i}} - \underbrace{f(E_i)}_{=0} \approx (O_i - E_i) + \frac{1}{2E_i}(O_i - E_i)^2$$

$$\Rightarrow 2\sum_{i=1}^{m} O_i \log\frac{O_i}{E_i} \approx 2\underbrace{\sum_{i=1}^{n}(O_i - E_i)}_{=0} + \cancel{2}\sum_{i=1}^{n}\frac{(O_i - E_i)^2}{\cancel{2}E_i} \Rightarrow 2\sum_{i=1}^{n} O_i \log\frac{O_i}{E_i} \approx \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$$



0, 1, ..., n, ∞ all pos integers

Application: Goodness of fit test

Example 3. Researcher observed the number of emissions of α particles in many 10-sec intervals. If we fit the data using a Poisson(λ) model, what is the best parameter estimate for λ?
How good is the model assumption? Are the observations really Poisson distributed?

n	Observed	
0–2	18	$e^{-\lambda}\left(\frac{\lambda^0}{0!}+\frac{\lambda^1}{1!}+\frac{\lambda^2}{2!}\right) \leftarrow P_1(\lambda)$
3	28	$e^{-\lambda}\frac{\lambda^3}{3!} \leftarrow P_2(\lambda)$
4	56	$e^{-\lambda}\frac{\lambda^4}{4!} \leftarrow P_3(\lambda)$
5	105	
6	126	
7	146	$\vdots$
8	164	
9	161	
10	123	
11	101	
12	74	
13	53	
14	23	
15	15	
16	9	$e^{-\lambda}\frac{\lambda^{16}}{16!} \leftarrow$
17+	5	$1-e^{-\lambda}\left(\frac{\lambda^0}{0!}+\cdots+\frac{\lambda^{16}}{16!}\right)$ $P_{16}(\lambda)$

Average count $\bar{X}_n = 8.392.$

$\hat{\lambda}_{MLE} = \bar{X}_n$

H_0: Obs subject to Poisson(λ) vs. H_1: H_0 is not true

$(P_1, \cdots P_{16}) = (P_1(\lambda), \cdots P_{16}(\lambda))$

$$2\sum_{i=1}^{16} O_i \log\frac{O_i}{E_i} \xrightarrow{d} \chi^2_{m-2} = \chi^2_{14}.$$

O_i = in the table

$E_i = n P_i(\hat{\lambda}) = n P_i(\bar{X}_n)$

Application: Goodness of fit test

p-value $= P(-2\log\lambda(\underline{X}_n) \geq 8.70854 \mid H_0)$
$= \text{pchisq}(8.70854,\ df = 14,\ \text{lower.tail} = \text{FALSE})$
$= 0.84926$

Example 3. Researcher observed the number of emissions of α particles in many 10-sec intervals. If we fit the data using a Poisson(λ) model, what is the best parameter estimate for λ?
How good is the model assumption? Are the observations really Poisson distributed?

n	Observed		Expected
0–2	18	$n e^{-\hat\lambda}\left(\frac{\hat\lambda^0}{0!}+\frac{\hat\lambda^1}{1!}+\frac{\hat\lambda^2}{2!}\right)$	12.2
3	28	$n e^{-\hat\lambda}\frac{\hat\lambda^3}{3!}$	27.0
4	56	$n e^{-\hat\lambda}\frac{\hat\lambda^4}{4!}$	56.5
5	105		94.9
6	126		132.7
7	146	⋮	159.1
8	164		166.9
9	161		155.6
10	123		130.6
11	101		99.7
12	74		69.7
13	53		45.0
14	23		27.0
15	15		15.1
16	9	$n e^{-\hat\lambda}\frac{\hat\lambda^{16}}{16!}$	7.9
17+	5	$n\left[1-e^{-\hat\lambda}\left(\frac{\hat\lambda^0}{0!}+\cdots+\frac{\hat\lambda^{16}}{16!}\right)\right]$	7.1

Average count $\bar{X}_n = 8.392$.

$\alpha = 0.05$

H_0 : Observations are Poisson distributed.

$$-2\log\lambda(\mathbf{X}_n) = 2\sum_{i=1}^{m} O_i \log\frac{O_i}{E_i} = 8.70854.$$

$R = \{-2\log\lambda(\underline{X}_n) \geq \chi^2_{14}(\alpha)\}$

$\chi^2_{14}(\alpha)$: qchisq(α, df = 14, lower.tail = FALSE) $= 23.685$

We fail to reject $H_0 \Rightarrow$ Poisson(λ) is a good fit.

Application: Goodness of fit test

$R = \{ -2\log\lambda(\vec{x}_n) \geq \underbrace{\chi^2_1(\alpha)}_{=3.841} \}$

We fail to reject H_0

H_0 model is a good fit.

Example 4. If the gene frequencies are in equilibrium, the genotypes AA, Aa, aa occur in the population with probability θ^2, $2\theta(1-\theta)$, $(1-\theta)^2$, according to the Hardy-Weinberg equilibrium model.

$$H_0: \text{ HW model is true.} \quad \text{versus} \quad H_1: \text{ HW model is not true.}$$

p-value = pchisq(0.03250, df=1, lower.tail = FALSE) = 0.8569

$H_0: (p_1, p_2, p_3) = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)$

vs. H_1: H_0 is not true.

$\nu = 3-2 = 1$.

We observed the phenotypes:

	AA	Aa	aa
O_i	342	500	187
E_i	$n\hat{\theta}^2 = 340.587$	$n \cdot 2\hat{\theta}(1-\hat{\theta}) = 502.826$	$n(1-\hat{\theta})^2 = 185.587$

$$\sup_{\Theta_0} L(p_1, p_2, p_3 \mid H_0) = \sup_{\theta} \frac{n!}{r_1!\, r_2!\, r_3!} (\theta^2)^{r_1} (2\theta(1-\theta))^{r_2} ((1-\theta)^2)^{r_3}$$

$$= \sup_{\theta} \frac{n!\, 2^{r_i}}{r_1!\, r_2!\, r_3} \underbrace{\theta^{2r_1+r_2} (1-\theta)^{r_2+2r_3}}_{\text{Denote } L_0(\theta)}$$

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = 0.03250$$

$\ell_0(\theta) = \log L_0(\theta) = (2r_1 + r_2)\log\theta + (r_2 + 2r_3)\log(1-\theta)$

$$2\sum_{i=1}^{m} O_i \log \frac{O_i}{E_i} = 0.03249$$

$\frac{\partial \ell_0}{\partial \theta} = 0 \Rightarrow \hat{\theta} = \frac{2r_1 + r_2}{2n} = \frac{2\times 342 + 500}{2 \times 1029} = 0.57532.$

Application: Goodness of fit test

1. The population is a categorical variable so that the grouped cells subject to a multinomial distribution;

2. Tickets in each cell are independent; ✓ iid assumptions

3. Large sample size n so that no more than 20% of expected counts less than 5

Fisher’s exact test

13.2 of Rice

07/20/2021

Dependencies between row and column classifications

Example 5. During phase 3 trial, some vaccine recipients were asked to complete diaries of their symptoms during the 7 days after vaccination.

	Pfizer / *BNT162b2*	Placebo
Fever ≥ 38.0°C	331	10
No fever	1,767	2,093
Total	2,098	2,103

* Systemic reactions in persons aged 18–55 years

Are occurrences of fever related to the vaccine/placebo treatment?

Could randomization result in such an imbalance?

H_0 : There is no relation. Any imbalance is due to randomization.

Dependencies between row and column classifications

Example 5. During phase 3 trial, some vaccine recipients were asked to complete diaries of their symptoms during the 7 days after vaccination.

	Treatment	Control	
Symptom	n_{11}	n_{12}	$n_{1.}$
No symptom	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

$n_{11} + n_{12} = n_{1.}$

H_1 : There is a relation.

H_0 : There is no relation. Any imbalance is due to randomization.

⇓ *induces*

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{1.}-n_{11}}}{\binom{n_{..}}{n_{1.}}}$$

$$\frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{12}}}{\binom{n_{..}}{n_{1.}}} \qquad n_{12} \to n_{1.} - n_{11}$$

$R = \{ N_{11} \le c_1 \text{ or } N_{11} \ge c_2 \}$

Tea-tasting experiment

H_0 : Bristol has no skills in determining the order.

	Milk first	Tea first	
Milk first	*4*	*0*	*4*
Tea first	*0*	*4*	*4*
	4	*4*	*8*

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{4}{n_{11}}\binom{4}{4-n_{11}}}{\binom{8}{4}}$$

N_{11}	0	1	2	3	4
p	0.014	0.229	0.514	0.229	0.014

$R = \{ N_{11} \le c_1 \text{ or } N_{11} \ge c_2 \}$

$P(N_{11} \le 0 \text{ or } N_{11} \ge 4 \mid H_0) = 0.014 + 0.014 = 0.028$

Dependencies between row and column classifications

Example 6. A group of supervisors each examined a personnel file to decide whether to promote the employee or not. The files are identical except for the gender label.

H_0 : There is no gender bias. Any imbalance is due to randomization.

	Male	Female	
Promote	21	14	35
Hold file	3	10	13
	24	24	

* From 13.2 of Rice

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{24}{n_{11}}\binom{24}{35-n_{11}}}{\binom{24+24}{35}}$$

⇕

```
dhyper(n11, m=24, n=24, k=35)
```

```
p-value = 2*phyper(21-1, m=24, n=24, k=35,
lower.tail = FALSE)
```

```
[1] 0.04899141
```

Dependencies between row and column classifications

Example 5 *cont'd*. During phase 3 trial, some vaccine recipients were asked to complete diaries of their symptoms during the 7 days after vaccination.

H_0 : There is no relation. Any imbalance is due to randomization.

	Pfizer / *BNT162b2*	Placebo	
Fever ≥ 38.0°C	331	10	341
No fever	1,767	2,093	3860
	2,098	2,103	

* Systemic reactions in persons aged 18–55 years

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{2098}{n_{11}}\binom{2103}{341-n_{11}}}{\binom{2098+2103}{341}}$$

$\Updownarrow$

```
dhyper(n11, m=2098, n=2103, k=341)
```

Benefits of more trials and repeated tests
$\Rightarrow$ *More significant results*

```
p-value = 2*phyper(331-1, m=2098, n=2103, k=341,
lower.tail = FALSE)
```

```
[1] 2.548842e-90
```

Dependencies between row and column classifications

Example 7. Phase 3 trial was a large, randomized, double-blind, placebo-controlled clinical trial:

H_0 : Infection rate is not related to vaccine/placebo treatment. $\leftrightarrow$ H_1 : It is related.

	Pfizer / *BNT162b2*	Placebo	
SARS-CoV-2 infected	9	169	178
No infection	21,711	21,559	43,439
	21,720	21,728	

* *Age ≥ 16, infections observed with onset at least 7 days after the second dose*

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{21720}{n_{11}}\binom{21728}{178-n_{11}}}{\binom{21720+21728}{178}}$$

$\Updownarrow$

```
dhyper(n11, m=21720, n=21728, k=178)
```

```
p-value = 2*phyper(9, m=21720, n=21728, k=178,
lower.tail = TRUE)
```

```
[1] 1.702187e-39
```

Dependencies between row and column classifications

Example 7 *cont'd*. Phase 3 trial was a large, randomized, double-blind, placebo-controlled clinical trial:

H_0 : Infection rate is not related to vaccine/placebo treatment. $\leftrightarrow$ H_1 : Infection rate is $\boxed{\text{lower}}$ in the vaccine group.

	Pfizer / *BNT162b2*	Placebo	
SARS-CoV-2 infected	9	169	178
No infection	21,711	21,559	43,439
	21,720	21,728	

* *Age ≥ 16, infections observed with onset at least 7 days after the second dose*

Hypergeometric distribution :

$$P(N_{11} = n_{11} \mid H_0) = \frac{\binom{21720}{n_{11}}\binom{21728}{178-n_{11}}}{\binom{21720+21728}{178}}$$

$\Updownarrow$

```
dhyper(n11, m=21720, n=21728, k=178)
```

```
p-value = phyper(9, m=21720, n=21728, k=178,
lower.tail = TRUE)
```

```
[1] 8.510933e-40
```

Tomorrow ...

$I \times J$ table

- Chi-squared test of homogeneity
- Chi-squared test of independence