Ch. 14 Analysis of Categorical Data

In many cases, the data we collect is categorical in nature i.e. the observations fall into one of several distinct categories and we record the count or proportion

e.g. smoker / former smoker / non-smoker

e.g. SA / A / N / D / SD (5-pt. Likert scale for surveys)

The no. of observations that fall into each category follows a multinomial dist. Recall the multinomial experiment (p. 279)

① experiment has n identical trials

② outcome of each trial falls into one of k classes or cells.

③ $p_i = P(\text{outcome falls in cell } i)$ $i = 1$ to k

$p_1 + p_2 + \cdots + p_k = 1$

④ Trials are independent (i.e. p_i constant for every trial)

⑤ Y_i = no. of outcomes that fall in cell i $i = 1$ to k

$Y_1 + Y_2 + \cdots + Y_k = n$

The jt. prob. fct. (from ch. 4)

$$p(y_1, \ldots, y_k) = \frac{n!}{y_1! y_2! \cdots y_k!} p_1^{y_1} \cdot p_2^{y_2} \cdots p_k^{y_k}$$

Of course, we are generally interested in conducting inference on p_i (proportion of smokers / former smokers; proportion of defectives / seconds, etc.)

Exact calculations involving the multinomial dist. are very cumbersome ⇒ Karl Pearson developed the χ^2 test which

is an approximate test!

14.2 The Chi-square Test

- As we have shown before, $Z \sim N(0,1)$ then $W = Z^2 \sim \chi^2_{(1)}$
- We also have the CLT . . .

  Let $Y_1 \cdots Y_n$ be iid w/ $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2 < \infty$

$$Y_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim N(0,1)$$

Question: If $Y \xrightarrow{D} N(0,1)$ would $X = Y^2 \xrightarrow{D} \chi^2_{(1)}$?

1st, recall the definition of convergence in dist.

Def

Let $\{Y_n\}$ be a sequence of RVs and let Y be a RV. Let $C[F(y)]$ denote the set of all pts where $F(y)$ is continuous. We say that Y_n converges in dist. to Y if

$$\lim_{n\to\infty} F_{Y_n}(y) = F(y) \qquad \forall y \in C[F(y)]$$

denoted $Y_n \xrightarrow{D} Y$

CLT says that $Y_n = \frac{\sum Y_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1)$

$\Rightarrow \lim_{n\to\infty} F_{Y_n}(y) = \Phi(y)$, $\Phi(y)$ is the cdf of a $N(0,1)$

Let $X_n = (Y_n)^2$ and Let $H_n(x) = P(X_n \le x)$ be the cdf of X_n

$\Rightarrow H_n(x) = P(X_n \le x) = P(Y_n^2 \le x) = P(-\sqrt{x} \le Y_n \le \sqrt{x})$

$= F_n(\sqrt{x}) - F_n(-\sqrt{x})$

$\Rightarrow \lim_{n\to\infty} H_n(x) = \lim_{n\to\infty} F_n(\sqrt{x}) - \lim_{n\to\infty} F_n(-\sqrt{x})$

$= \Phi(\sqrt{x}) - \Phi(-\sqrt{x})$ note: $\Phi(y)$ cont. everywhere

Symmetry $= 2 \cdot \int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$

$u = y^2 \Rightarrow du = 2y\,dy \Rightarrow du = 2\sqrt{u}\,dy$

$= 2 \cdot \int_0^{x} \frac{1}{\sqrt{2\pi}} e^{-u/2} \cdot \frac{du}{2\sqrt{u}}$

$= \int_0^{x} \frac{1}{\Gamma(1/2)\cdot 2^{1/2}} u^{1/2-1} e^{-u/2}\, du$

cdf of $X \sim GAM(\alpha = 1/2, \beta = 2) \sim \chi^2_{(1)}$

$\Rightarrow X_n = (Y_n)^2 \xrightarrow{D} \chi^2_{(1)}$ ■

Consider a binomial dist. (2 categories S/F)

Let $Y_1 \sim Bin(n, p_1)$ and $Y_2 = n - Y_1 \Rightarrow Y_2 \sim Bin(n, p_2)$

If $W = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}}$ then $W \xrightarrow{D} N(0,1)$ (CLT)

Then $Q = W^2 = \frac{(Y_1 - np_1)^2}{np_1(1-p_1)} \xrightarrow{D} \chi^2_{(1)}$

$$Q = \frac{(Y_1 - np_1)^2 \cdot (1-p_1) + (Y_1 - np_1)^2 \cdot p_1}{np_1(1-p_1)}$$

$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_1 - np_1)^2}{n(1-p_1)}$$

$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(n - Y_2 - n(1-p_2))^2}{n \cdot p_2}$$

$$= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

What do we have here?

$Y_1 = O_1 \quad np_1 = E_1 \quad Y_2 = O_2 \quad E_2 = np_2$

$$\Rightarrow Q = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(1)} \; !!$$

note: This is a large-sample approximation using CLT
It is not an exact test!

We have shown this for $k=2$, and it can be shown in

general for k categories that

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)}$$

, $E_i = n \cdot p_i$ (marginal EVs)

note: K cells of data, subject to one linear constraint ($p_1 + \cdots + p_k = 1$)

14.3 Goodness-of-fit test $\Rightarrow$ d.f. = k-1

e.g. Mendel Genetics

R: round pea Y: yellow note: round and yellow

r: wrinkled pea y: green carry the dominant genes

If we cross two heterogenous parents: ($Rr Yy \times Rr Yy$)

we get the classic 9:3:3:1 ratio

round/wrinkled genes from each parent	Yellow/green genes from each parent: YY	Yy	yY	yy
RR	RY	RY	RY	Ry
Rr	RY	RY	RY	Ry
rR	RY	RY	RY	Ry
rr	rY	rY	rY	ry

Our hypothesis then is for every 16 plants:

9 round/yellow 3 round/green 3 wrinkled/yellow 1 wrinkled green

① H_0: $p_1 = 9/16$ $p_2 = p_3 = 3/16$ $p_4 = 1/16$

H_a: at least one $p_i \neq p_0$

Spse the observed counts are:

$O_1 = 219 \quad O_2 = 81 \quad O_3 = 64 \quad O_4 = 31 \quad (n = 400)$

Is this data consistent w/ our hypothesis?

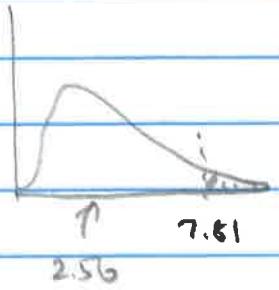
$E_1 = 400(9/16) = 225 \qquad E_2 = E_3 = \frac{3}{16}(400) = 75$

$E_4 = \frac{1}{16}(400) = 25$

$$Q = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(219-225)^2}{225} + \frac{(81-75)^2}{75}$$

$$+ \frac{(64-75)^2}{75} + \frac{(31-25)^2}{25} = 2.56 \sim \chi^2_{(3)}$$

$\chi^2_{3,.05} = 7.81$



do not reject H_0

i.e. the observed data does not give us any evidence against the 9:3:3:1 ratio

note: $Q \approx 0$ means the observed counts matched the expected counts too closely (Fisher / Mendel controversy)

e.g. Death by horse kick (Prussian calvary)

- Ten army corps observed over 20 yrs (1875 - 1894)
- no. of deaths by horse kick were recorded each year for each corps, giving us 200 observations

no. of deaths	count	p_i	E_i
0	109	.54335	108.67
1	65	.33145	66.29
2	22	.10110	20.22
3	3	.02055	4.11
4	1	.00315	.63
5 or more	0	.0004	.08

Collapse

Does this data follow a poisson dist?

$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad E(Y) = \lambda \quad V(Y) = \lambda$$

Use $\bar{Y}$ to estimate λ ...

$$\bar{Y} = \frac{109(0) + 65(1) + 22(2) + 3(3) + 1(4)}{200} = .61$$

i.e. on avg, we observed .61 deths per year

$$p_i = \frac{\lambda^i e^{-\lambda}}{i!}, \quad E_i = 200 \cdot \frac{(.61)^i e^{-.61}}{i!}$$

$$\chi^2 = \frac{(109-108.67)^2}{108.67} + \cdots + \frac{(4-4.82)^2}{4.82} = .3223$$

$\chi^2_{(3),.05} = 7.81 \Rightarrow$ don't reject H_0, p-value = .9558

note: data almost matches too well!

14.4 Contingency Tables

In many cases we are interested in determining if there is a relationship between 2 variables. When the variables are numerical, we can use linear regression techniques.

e.g. X = wt. of car Y = MPG

X = SAT Y = College GPA

When our variables are categorical in nature, we need to rely again on counts are proportions in each cell

Example 14.3

Spse we want to evaluate the effectives of the flu vaccine in a small community. Data is collected and summarized in a contingency table

DV	IV: No vaccine	IV: One shot	IV: Two shots
Flu	24	9	13
No Flu	289	100	565

In this case, we would like to know if the no. of shots has an influence on a patient's chance of getting flu.

We can think of this as a test of independence (between flu / no. of shots) or as a test for homogeniety of proportions i.e. the proportion of people who get the flu should be similar for no shot, 1 shot, or 2 shots if the flu vaccine is ineffective.

Observed proportions of flu for each level of shots:

	no vaccine	one shot	two shots
Flu	.077	.083	.0225
no Flu	.923	.917	.9775

By inspection, we might surmise that 2 shots significantly reduces someone's chance of contracting the flu. How do we know if these differences are too large to be attributable to random chance?

Consider an $r \times c$ contingency table

rows \ columns	1	2	...	c	
1					$n_{1\cdot}$
2					$n_{2\cdot}$
.		n_{ij}			.
.					.
r					$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot c}$	n

n_{ij} = count in cell (i,j) → row i, col. j

$n_{i\cdot}$ = i^{th} row total

$$= \sum_{j=1}^{c} n_{ij}$$

n = grand total

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j}$$

$n_{\cdot j}$ = j^{th} col. total

$$= \sum_{i=1}^{r} n_{ij}$$

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \qquad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \qquad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

$$= \sum_{j=1}^{c} \hat{p}_{ij} \qquad = \sum_{i=1}^{r} \hat{p}_{ij}$$

The joint dist. of $(n_{11}, n_{12}, \ldots, n_{rc})$ is multinomial with p_{ij} for $i = 1$ to r and $j = 1$ to c

We want to test that rows are independent of columns (i.e. $p_{ij} = p_{i\cdot} \times p_{\cdot j}$)

note: rc cell subject to one linear constraint $\sum_i \sum_j p_{ij} = 1 \Rightarrow$ d.f. $= rc - 1$

from before ...

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - n \cdot p_{ij})^2}{n p_{ij}} \xrightarrow{D} \chi^2_{(rc-1)}$$

The reason we don't use this test statistic, is that we don't know p_{ij}. However, under H_0: $p_{ij} = p_{i\cdot} \times p_{\cdot j}$

↳ independence $P(A \cap B) = P(A) \cdot P(B)$

So we need to estimate $p_{i\cdot}$ ($i = 1$ to r) and $p_{\cdot j}$ ($j = 1$ to c)

Since $\sum_{i=1}^{r} p_{i\cdot} = 1$ and $\sum_{j=1}^{c} p_{\cdot j} = 1$, we only need to estimate

$(r-1) + (c-1)$ parameters.

Find the MLE estimates of p_{ij} under H_0

$$L(p_{11}, p_{12}, \ldots, p_{rc}) = \binom{n}{n_{11} \cdots n_{rc}} \cdot p_{11}^{n_{11}} \cdots p_{rc}^{n_{rc}}$$ → multinomial prob. fct

$$\ln L = \ln \binom{n}{n_{11} \cdots n_{rc}} + n_{11} \ln p_{11} + \cdots + n_{rc} \ln p_{rc}$$

Subject to constraint: $\sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} = 1$

Under H_0: $p_{ij} = p_{i\cdot} \times p_{\cdot j}$

$$\Rightarrow \ln L_{\Omega_0} = \ln \binom{n}{n_{11} \cdots n_{rc}} + n_{11}[\ln p_{1\cdot} + \ln p_{\cdot 1}] + n_{12}[\ln p_{1\cdot} + \ln p_{\cdot 2}]$$

$$+ \cdots + n_{rc}[\ln p_{r\cdot} + \ln p_{\cdot c}]$$

Subject to constraints: $\sum_{i=1}^{r} p_{i\cdot} = 1$, $\sum_{j=1}^{c} p_{\cdot j} = 1$

To max $\ln L_{\Omega_0}$ w.r.t. $p_{i\cdot}$'s and $p_{\cdot j}$'s which are subject to constraints, use LaGrange multipliers.

Let $k = \ln L_{\Omega_0} + \lambda \underbrace{\left(1 - \sum_{i=1}^{r} p_{i\cdot}\right)}_{0} + \mu \underbrace{\left(1 - \sum_{j=1}^{c} p_{\cdot j}\right)}_{0}$

Take partials of k w.r.t. all $p_{i\cdot}$'s, $p_{\cdot j}$'s, and λ and μ

$$\frac{\partial k}{\partial p_{i\cdot}} = \frac{n_{i\cdot}}{p_{i\cdot}} - \lambda \overset{\text{SET}}{=} 0 \qquad \frac{\partial k}{\partial p_{\cdot j}} = \frac{n_{\cdot j}}{p_{\cdot j}} - \mu \overset{\text{SET}}{=} 0$$

$$\frac{\partial k}{\partial \lambda} = 1 - \sum_{i=1}^{r} p_{i\cdot} \overset{\text{SET}}{=} 0 \qquad \frac{\partial k}{\partial \mu} = 1 - \sum_{j=1}^{c} p_{\cdot j} \overset{\text{SET}}{=} 0$$

Solve r+c+2 eqns. simultaneously

- $n_{i\cdot} = \lambda \cdot p_{i\cdot}$ $(i = 1 \text{ to } r)$

$$\sum_{i=1}^{r} n_{i\cdot} = \sum_{i=1}^{r} \lambda \cdot p_{i\cdot} \Rightarrow n = \lambda \cdot \left(\sum_{i=1}^{r} p_{i\cdot}\right) \Rightarrow n = \lambda$$

$$\Rightarrow \boxed{\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}}$$

- $n_{\cdot j} = \mu p_{\cdot j}$ $(j = 1 \text{ to } c)$

$$\sum_{j=1}^{c} n_{\cdot j} = \sum_{j=1}^{c} \mu p_{\cdot j} \Rightarrow n = \mu \left(\sum_{j=1}^{c} p_{\cdot j}\right) \Rightarrow n = \mu$$

$$\Rightarrow \boxed{\hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}}$$

so under H_0, $\hat{p}_{ij} = \hat{p}_{i\cdot} \times \hat{p}_{\cdot j} = \dfrac{n_{i\cdot} \times n_{\cdot j}}{n^2}$

E_{ij} = "expected count in row i, col j under H_0"

$$= n \cdot \hat{p}_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n} = \boxed{\frac{(\text{row total}) \times (\text{col. total})}{(\text{grand total})}}$$

Let O_{ij} = "observed count in row i, col j" $(O_{ij} = n_{ij})$

$$\Rightarrow Q = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n \cdot \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j})^2}{n \cdot \hat{p}_{i\cdot} \cdot \hat{p}_{\cdot j}}$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(rc-1-(r-1)-(c-1))}$$

note:

$rc - 1 - (r-1) - (c-1) = rc - r - c + 1 = (r-1) \times (c-1)$

This is the χ^2 test statistic from 2040/2050!

It is based on the CLT and is an approximation!

Back to the flu example:

We compute the expected counts (in parentheses)

using $E_{ij} = \frac{r_i \times c_j}{n}$

	no vaccine	1 shot	2 shots	
Flu	24 (14.4)	9 (5.0)	13 (26.6)	46
no Flu	289 (298.6)	100 (104.0)	565 (551.4)	954
	313	109	578	1000

H_0: Independance between vaccine and contracting the Flu

H_a: Relationship between vaccine and contracting the Flu

$$\chi^2 = \frac{(24-14.4)^2}{14.4} + \frac{(9-5)^2}{5} + \cdots + \frac{(565-551.4)^2}{551.4}$$

$$= 17.35$$

$$\chi^2_{(2),.05} = 5.991$$

$17.35 > 5.991 \Rightarrow$ Reject H_0 What do we conclude?

going back to the observed proportions I would conclude ---

	no vaccine	1 shot	2 shots
Flu	.077	.083	.0225
no Flu	.923	.917	.9775

Taking 2 flu shots significantly reduces your chance of contracting the flu. Taking the flu shot only once shows no evidence of being effective.

Relationship to large-sample z-test for p_1 vs. p_2

e.g. Is there a relationship between smoking and lung cancer?

		Lung problem: Yes	Lung problem: No	
Smoking	Yes	(22.5) 35	(22.5) 10	45
	No	(27.5) 15	(27.5) 40	55
		50	50	100

note: $\hat{p}_S = \frac{35}{45} = .778$, $\hat{p}_{NS} = \frac{15}{55} = .273$

χ^2 test : Ho: independence between smoking and lung problems

Ha: relationship exists between smoking and lung problems

$$\chi^2 = \frac{(35-22.5)^2}{22.5} + \frac{(10-22.5)^2}{22.5} + \frac{(15-27.5)^2}{27.5}$$

$$+ \frac{(40-27.5)^2}{27.5} = 25.2525 \qquad \text{p-value} = .0000005$$

⇒ Reject H_0 i.e. lung problems and smoking are related

z-test for p_1 vs. p_2

$H_0: p_S = p_{NS}$

$H_a: p_S \neq p_{NS}$

$\hat{p}_S = \frac{35}{45} = .778 \qquad \hat{p}_{NS} = \frac{15}{55} = .273$

$\hat{p} = \frac{35+15}{45+55} = .50$

$$z = \frac{(.778 - .273)}{\sqrt{(.5)(.5)\left(\frac{1}{45} + \frac{1}{55}\right)}} = 5.025189 \qquad \text{p-value} = .0000005$$

$z^2 = (5.025189)^2 = 25.2525$

These are the same tests! $z^2 \sim \chi^2_{(1)}$

14.5 r x c Tables with Fixed Row or Column Totals

In many applications, we might want to fix the no. we have in each column.

e.g. R.S. a fixed no. of people who had no vaccine, one shot, or two shots

What does this do to our test statistic? Keep in mind that we still want to test if the proportion of those who get flu is the same as we move from one column to the next.

In this case, we could consider the no. who get the flu, n_{ij}, to be binomial.

	No vaccine	one shot	Two shots
flu	n_{11}	n_{12}	n_{13}
no flu			
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$

fixed

$n_{11} \sim Bin(n_{\cdot 1}, p_{11})$ $n_{12} \sim Bin(n_{\cdot 2}, p_{12})$ $n_{13} \sim Bin(n_{\cdot 3}, p_{13})$

Under H_0: $p_{11} = p_{12} = p_{13} = p$

If we extend this to more than 2 rows, then we have the counts in each column following a multinomial dist.

In this case,

$H_0: p_{1i} = p_{1\cdot}, \ldots, p_{ri} = p_{r\cdot}$

cols

rows	1	2	...	c	
1	n_{11}	n_{12}		n_{1c}	
2	n_{21}			n_{2c}	test if $p_{ij} = p_{i\cdot}$
⋮	⋮			⋮	
r	n_{r1}			n_{rc}	
	$n_{\cdot 1}$			$n_{\cdot c}$	
	multinomial dist.			multinomial dist.	

The likelihood under H_0 (same proportions for each category)

13

$$L_{\Omega_0} = \left[\binom{n_{\cdot 1}}{n_{11} \cdots n_{r1}} p_{1\cdot}^{n_{11}} \cdots p_{r\cdot}^{n_{r1}} \right] \cdots \left[\binom{n_{\cdot c}}{n_{1c} \cdots n_{rc}} p_{1\cdot}^{n_{1c}} \cdots p_{r\cdot}^{n_{rc}} \right]$$

Subject to constraint $\sum_{i=1}^{r} p_{i\cdot} = 1$

(use Lagrange multipliers again)

$$k = L_{\Omega_0} - \lambda \left(1 - \sum_{i=1}^{r} p_{i\cdot} \right)$$

$$\ln k = \sum_{j=1}^{c} \ln \binom{n_{\cdot j}}{n_{1j} \cdots n_{rj}} + \left(n_{11} \ln p_{1\cdot} + \cdots + n_{r1} \ln p_{r\cdot} \right)$$

$$+ \left(n_{1c} \ln(p_{1\cdot}) + \cdots + n_{rc} \ln(p_{r\cdot}) \right) + \lambda \left(1 - \sum_{i=1}^{r} p_{i\cdot} \right)$$

$$\frac{\partial \ln k}{\partial p_{i\cdot}} = \sum_{j=1}^{c} n_{ij} / p_{i\cdot} - \lambda = \frac{n_{i\cdot}}{p_{i\cdot}} - \lambda \overset{\text{SET}}{=} 0$$

$$\frac{\partial \ln k}{\partial \lambda} = 1 - \sum_{i=1}^{r} p_{i\cdot} \overset{\text{SET}}{=} 0 \quad \Rightarrow \quad \sum_{i=1}^{r} p_{i\cdot} = 1$$

Solving $(r+1)$ eqns. simultaneously ....

$$n_{i\cdot} = \lambda \cdot p_{i\cdot} \quad (r \text{ eqns.})$$

$$\sum_{i=1}^{r} n_{i\cdot} = \lambda \cdot \sum_{i=1}^{r} p_{i\cdot} \quad \Rightarrow \quad n = \lambda$$

$$\Rightarrow \hat{p}_{i\cdot} = \frac{\hat{n}_{i\cdot}}{\lambda} = \boxed{\frac{n_{i\cdot}}{n}}$$

$$\Rightarrow E_{ij} \overset{H_0}{=} n_{\cdot j}\left(\frac{n_{i\cdot}}{n}\right) = \boxed{\frac{n_{i\cdot} \times n_{\cdot j}}{n}} = \frac{(\text{row total}) \times (\text{col. total})}{\text{grand total}}$$

This is the same statistic!

---

d.f. = (rc cells) − (c constraints) − (r − 1)

$p_{1j} + \cdots + p_{rj} = 1$ for $j = 1$ to c

(r − 1): estimated parameters

$= c(r-1) - 1\cdot(r-1) = (c-1)\times(r-1)$

Example 14.4

Survey of voter sentiment in 4 midcity political wards

r.s. of 200 voters in each ward

Is there sufficient evidence to indicate that voting preferences differ in the 4 wards?

	Ward 1 (.38)	2 (.265)	3 (.295)	4 (.24)		
Favor A	76 (59)	53 (59)	59 (59)	48 (59)	236	(.295)
don't favor A	124 (141)	147 (141)	141 (141)	152 (141)	564	(.705)
	200	200	200	200	800	

note: Observed proportions: .38 vs. .265 vs. .295 vs. .24
Are these differences too large to be attributable to chance alone?

$H_0: p_1 = p_2 = p_3 = p_4$

H_a: at least one p_i is different

$$\chi^2 = \frac{(76-59)^2}{59} + \frac{(53-59)^2}{59} + \ldots + \frac{(152-141)^2}{141}$$

$$= 10.72$$

$\chi^2_{3,.05} = 7.815 \Rightarrow$ Reject H_0 i.e. We believe there are real differences in voter preferences between wards.

16.2 Bayesian Priors, Posteriors and Estimators

To this point, we have assumed we have no information regarding the unknown parameter Θ, until we collect sample information.

In many instances this is probably not correct.

e.g. We want to estimate the proportion of crystal meth addicts in Utah County. We know this should be small.

The idea of Bayesian statistics, is that if we make an initial assumption regarding Θ, (i.e. utilize "prior info on θ"), then our estimator after collecting sample data will be better than if we had assumed no prior info at all on θ was available.

Bayesian process:

Let $L(y_1, y_2, \ldots, y_n | \theta)$ denote the likelihood fct for a r.s. of size n. This represents the density of our sample data given Θ.

In Bayesian analysis, Θ is viewed as a RV itself with prior dist. or prior density fct $g(\theta)$.

$\Rightarrow f(y_1, \ldots, y_n, \theta) = L(y_1, y_2, \ldots, y_n | \theta) * g(\theta)$ is the joint density fct. of $y_1 \cdots y_n, \theta$

Recall: $f(y_2|y_1) = \dfrac{f(y_1,y_2)}{f(y_1)}$, $f(y_1) = \int_{-\infty}^{\infty} f(y_1,y_2)\,dy_2$

⇒ the marginal density of our sample data $(y_1 \cdots y_n)$ is

$$m(y_1 \cdots y_n) = \int_{-\infty}^{\infty} \underbrace{L(y_1 \cdots y_n|\theta) \times g(\theta)}_{\text{jt. density}}\,\underbrace{d\theta}_{\text{integrate out } \theta}$$

Therefore, the posterior density of $\theta \mid y_1 \cdots y_n$ is

$$g^*(\theta \mid y_1, y_2, \cdots, y_n) = \frac{L(y_1, \cdots, y_n|\theta) \times g(\theta)}{\int_{-\infty}^{\infty} L(y_1, \cdots, y_n|\theta) \times g(\theta)\,d\theta}$$

note: The posterior density contains all pertinent info about θ: (1) prior info on θ (2) sample info on θ

Example 16.1

We would like to estimate p = proportion of pop. afflicted with a certain disease. We need to assume a prior dist. for p, and proportions are often modeled with a Beta dist.

Recall: $Y \sim \text{Beta}(\alpha, \beta)$

$$f(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}\, y^{\alpha-1}(1-y)^{\beta-1}, \quad 0<y<1$$

0)

$$E(Y) = \frac{\alpha}{\alpha+\beta} \qquad V(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

note: we will worry about the choice of α, β later

Our sample data: $Y_1 \cdots Y_n$ is a r.s. from $Y_i \sim Bin(1, p)$

i.e. each person either has the disease or they don't.

$$p(y_i|p) = p^{y_i}(1-p)^{1-y_i}, \quad y_i = 0, 1$$

$$\Rightarrow L(y_1 \cdots y_n | p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

$$= p^{\sum y_i}(1-p)^{n-\sum y_i}$$

prior dist for p: $g(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$

$\Rightarrow$ jt. density of $Y_1 \cdots Y_n, p$ is:

$$f(y_1 \cdots y_n, p) = p^{\sum y_i}(1-p)^{n-\sum y_i} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$\Rightarrow$ marginal density of $Y_1 \cdots Y_n$ is:

$$m(y_1 \cdots y_n) = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \underbrace{p^{\sum y_i + \alpha - 1}(1-p)^{n-\sum y_i+\beta-1}}\, dp$$

Let $\alpha' = \sum y_i + \alpha$, $\beta' = n - \sum y_i + \beta$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot \frac{\Gamma(\Sigma y_i+\alpha)\cdot\Gamma(n-\Sigma y_i+\beta)}{\Gamma(n+\alpha+\beta)}$$

note: does not depend on p

The posterior density of p given our sample data:

$$g^*(p \mid y_1 \cdots y_n) =$$

$$\frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} p^{\Sigma y_i+\alpha-1} (1-p)^{n-\Sigma y_i+\beta-1}}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\Sigma y_i+\alpha)\cdot\Gamma(n-\Sigma y_i+\beta)}{\Gamma(n+\alpha+\beta)}}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(\Sigma y_i+\alpha)\cdot\Gamma(n-\Sigma y_i+\beta)} \cdot p^{(\Sigma y_i+\alpha)-1} (1-p)^{(n-\Sigma y_i+\beta)-1}, \quad 0<p<1$$

$$0 \quad , \quad -$$

$\sim$ beta (α', β')

$\alpha^* = \Sigma y_i + \alpha$

$\beta^* = n - \Sigma y_i + \beta$

i.e. the posterior dist. of $p \mid y_1 \cdots y_n$ is also a beta dist. This means that the beta dist. is a conjugate prior dist. for a bernoulli or binomial dist.

posterior mean $E(p \mid y_1 \cdots y_n) = \frac{\Sigma y_i + \alpha}{n+\alpha+\beta}$

prior mean $E(p) = \frac{\alpha}{\alpha+\beta}$

If $\sum y_i$ is high i.e. a lot of people in our sample had the disease, the expected value of $p \uparrow$ (vice versa if $\sum y_i$ small)

Spse we think $p \approx .25$ and use $p \sim \text{beta}(\alpha=1, \beta=3)$

$\Rightarrow E(p) = \frac{1}{1+3} = .25$, $V(p) = \frac{1(3)}{(1+3)^2(1+3+1)} = .0375$

- If $n=25$, $\sum y_i = 10$ (i.e. $\hat{p} = .40$)

$\Rightarrow \alpha^* = 10+1 = 11$, $\beta^* = (25-10)+3 = 18$

posterior mean $= \frac{11}{29} = .379$ posterior variance $= .0078$

---

note: $\hat{p} = 10/25 = .40$ $V(\hat{p}) = \frac{.4(.6)}{25} = .0096$

- If $n = 100$, $\sum y_i = 40$

$\Rightarrow \alpha^* = 40+1 = 41$ $\beta^* = (100-40)+3 = 63$

posterior mean $= \frac{41}{41+63} = .394$ posterior variance $= .0023$

---

note: as $n \uparrow$, more wt. is applied to sample data

Spse we think $p \approx .25$ but use $p \sim \text{beta}(\alpha=10, \beta=30)$

$\Rightarrow E(p) = \frac{10}{10+30} = .25$, $V(p) = \frac{(10)(30)}{(10+30)^2(10+30+1)} = .00457$

- If $n=25$, $\sum y_i = 10$

$\Rightarrow \alpha^* = 10+10 = 20$, $\beta^* = 25-10+30 = 45$

posterior mean $= \frac{20}{20+45} = .307$ posterior variance $= .00323$

- If $n=100$, $\sum y_i = 40$

$\Rightarrow \alpha^* = 40+10 = 50$, $\beta^* = (100-40)+30 = 90$

posterior mean $= \frac{50}{50+90} = .357$ posterior variance $= .0016$

note: using a prior dist. of beta $(10, 30)$ means that the posterior estimate for p will not be as "reactive" to the sample data. However, it will move closer to the sample estimate as $n \uparrow$.

see graphs in R

Def 16.2 Let $Y_1 \cdots Y_n$ be a r.s. w/ likelihood fct $L(y_1, y_2, \ldots, y_n|\theta)$, and let θ have prior density $g(\theta)$. The posterior Bayes estimator $t(\theta)$ is given by

$$\hat{t(\theta)}_B = E(t(\theta) \mid y_1 \cdots y_n)$$

The posterior Bayes estimator of θ is

$$\hat{\theta}_B = E(\theta \mid y_1 \cdots y_n)$$ which is the posterior mean