# Multinomial Distribution



Bernoulli : $p(y) = \pi^{y}(1-\pi)^{1-y}$ , $y = 0, 1$

Binomial : $p(y) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y}$

$y = 0, 1, \ldots, n$

$Y \sim B(n, \pi)$

## Multinomial

Statistical experiment with $c$ outcomes.

Repeat independently $n$ times

$\Pr(\text{Outcome } j) = \pi_j$ , for $j = 1, \ldots, c$

# of times outcome $j$ occurs is $n_j$

for $j = 1, \ldots, c$

Integer-valued multivariate distribution

joint distribution of

$n_1, n_2, \ldots n_c$

## Multinomial coefficient

Have $n$ objects. # of ways to label $n_1$ as type 1
$n_2$ as type 2
$\vdots$
$n_c$ as type $c$ is

$$\binom{n}{n_1\, n_2 \cdots n_c} = \frac{n!}{n_1!\, n_2! \cdots n_c!}$$

Ex Of 30 graduating students, how many ways are there for 15 to be employed in a job related to their field of study 10 employed in unrelated field, & 5 unemployed

$$\binom{30}{15\ 10\ 5} = \frac{30!}{15!\,10!\,5!}$$

$$= 465{,}817{,}912{,}560$$

Multinomial distribution

Denote by $M(n, \underset{\sim}{\pi})$

$\pi = (\pi_1, \pi_2 -- \pi_c)$

$\underset{\sim}{n} = (n_1, n_2, -- n_c)$ $\quad \underset{\sim}{n} \sim M(n, \underset{\sim}{\pi})$

$$P(n_1, \ldots, n_c) = \binom{n}{n_1\, n_2 -- n_c} \pi_1^{n_1} \pi_2^{n_2} -- \pi_c^{n_c}$$

where

$$\sum_{j=1}^{c} \pi_j = 1 \quad \text{and} \quad \sum_{j=1}^{c} n_j = n$$

<u>Example</u> Of 10 people at a graduation party what is the prob that 1 year later

4 ~~Married~~ Single
3 Married ~~Divorced~~
2 Divorced
1 Widowed

$$P(SSSSMMMDDW) = \pi_1^4 \pi_2^3 \pi_3^2 \pi_4^1$$

Ten slots to put letters in

So

$$P(4,3,2,1) = \binom{10}{4\ 3\ 2\ 1} \pi_1^4 \pi_2^3 \pi_3^2 \pi_4$$

# Multinomial Distribution

Statistical Experiment with $c$ outcomes

Repeated $n$ times

$Pr(\text{outcome}) \; \pi_j \; , \; j = 1, \ldots, c$

# of times outcome $j$ happens is $n_j \; , \; j = 1, \ldots, c$

$$P(n_1, \ldots, n_c) = \binom{n}{n_1 \; n_2 \; \cdots \; n_c} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c} \; ,$$

$$0 \le n_j \le n \; , \; 0 \le \pi_j < 1 \; , \; \sum_{j=1}^{c} \pi_j = 1 \; , \; \sum_{j=1}^{c} n_j = n$$

Or

$$P(n_1, n_2, \ldots n_{c-1}) = \frac{n!}{n_1! \, n_2! \cdots n_{c-1}! \, (n - \sum_{j=1}^{c-1} n_j)!}$$

$$\times \; \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_{c-1}^{n_{c-1}} \left(1 - \sum_{j=1}^{c-1} \pi_j\right)^{n - \sum_{j=1}^{c-1} n_j}$$

Can combine categories, adding probabilities, & result is still multinomial.

For 2 categories Get Binomial.

$$P(n_1, n_2) = \frac{n!}{n_1! \, \underset{\substack{\uparrow \\ (n - n_1)}}{n_2!}} \pi_1^{n_1} (1 - \pi_1)^{n - n_1}$$

Binomial

Expected value

$n \pi_1$

Variance

$n \pi_1 (1 - \pi_1)$

# Sample problem

Suppose that for recent university graduates

- P(Job related to field of study) = 0.6
- P(Job unrelated to field of study) = 0.3
- P(No job) = 0.1

Of 30 randomly chosen students, what is the probability that

a) 15 are employed in a job related to field of study, 10 are employed in a job unrelated to their field of study, and 5 are unemployed?

$$\frac{30!}{15!\ 10!\ 5!}(0.6)^{15}(0.3)^{10}(0.1)^{5} \doteq 0.0129$$

dmultinom(c(15,10,5), prob=c(.6,.3,.1))

b) Exactly 5 are unemployed? Binomial

$$\binom{30}{5}(.1)^5(.9)^{25} = 0.1023.$$

Conditional probabilities are also multinomial

- Given that a student finds a job, what is the probability that it is in her field of study?

$$P(\text{Field} \mid \text{Job}) = \frac{P(\text{Field} \cap \text{Job})}{P(\text{JOB})}$$

$$= \frac{.6}{.6 + .3} = \frac{2}{3}$$

- Suppose we choose 50 students at random from those who found jobs. What is the probability that exactly $y$ of them will be employed in their field of study?

$$P(y) = \binom{50}{y} \left(\frac{2}{3}\right)^{y} \left(\frac{1}{3}\right)^{50-y}$$

for $y = 0, \ldots 50$

# Estimation

1 = Field
2 = Out of Field
3 = No job

Hypothetical Data File

| Case | Job | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 3 | 0 | 0 | 1 |
| 3 | 2 | 0 | 1 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| n | 2 | 0 | 1 | 0 |
| | | $n_1$ | $n_2$ | $n_3$ |

$\sum_{i=1}^{n} y_{ij} = n_j$ Add down columns

$(\pi_1, \pi_2, \pi_3)$

Cases (n of them) are independent $M(1, \underset{\sim}{\pi})$

$E(y_j) = \pi_j$

Column totals count # of time each outcome occurs in n independent trials $(\pi_1, \pi_2, \pi_3)$

Their joint distribution is $M(n, \underset{\sim}{\pi})$

If you make a frequency table ...

$n_j$ are cell frequencies

| Field | Out | Unemp |
|---|---|---|
| 32 | 69 | 8 |

Cell frequencies: Joint dist is multinomial

Each individual frequency is Binomial $B(n, \pi_i)$

| Job Category | Freq | % |
|---|---|---|
| Employed in Field | 106 | 53 |
| Employed outside Field | 74 | 37 |
| Unemployed | 20 | 10 |
| Total | 200 | 100 |

Likelihood Function

$$\ell(\underset{\sim}{\pi}) = \prod_{i=1}^{n} P(Y_1 = y_1, Y_2 = y_2, \ldots Y_c = y_c ; \underset{\sim}{\pi})$$

$$= \prod_{i=1}^{n} \pi_1^{y_{i1}} \pi_2^{y_{i2}} \ldots \pi_c^{y_{ic}}$$

$$= \pi_1^{\sum_{i=1}^{n} y_{i1}} \pi_2^{\sum_{i=1}^{n} y_{i2}} \ldots \pi_c^{\sum_{i=1}^{n} y_{ic}}$$

$$= \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

Likelihood for multinomial



$$l(\underset{\sim}{\pi}) = l = \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

Want MLE of $(\pi_1, \ldots \pi_c)$

$$\frac{\partial}{\partial \pi_k} \log l = \frac{\partial}{\partial \pi_k} \log\left( \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_{c-1}^{n_{c-1}} \left(1 - \sum_{j=1}^{c-1} \pi_j\right)^{n_c} \right)$$

$$= \frac{\partial}{\partial \pi_k} \left( \sum_{j=1}^{c-1} n_j \log \pi_j + n_c \log \left(1 - \sum_{j=1}^{c-1} \pi_j \right) \right)$$

$$= \frac{n_k}{\pi_k} + \frac{n_c}{1 - \sum_{j=1}^{c-1} \pi_j} (-1) \underbrace{\frac{\partial \pi_k}{\partial \pi_k}}_{=1} \overset{\text{set}}{=} 0$$

$$\frac{n_k}{\pi_k} = \frac{n_c}{\pi_c} \quad \text{for } k = 1, \ldots, c-1$$

$$\frac{n_1}{\pi_1} = \frac{n_2}{\pi_2} = \cdots = \frac{n_c}{\pi_c} \quad \text{so}$$

$$\frac{n_1}{\pi_1} = \frac{n_2}{\pi_2} \Rightarrow n_1 \pi_2 = n_2 \pi_1 \Rightarrow \pi_2 = \frac{n_2}{n_1} \pi_1$$

$$\pi_3 = \frac{n_3}{n_1} \pi_1$$

$$\pi_4 = \frac{n_4}{n_1} \pi_1$$

$$\vdots$$

$$\pi_c = \frac{n_c}{n_1} \pi_1$$

$\pi_1 + \pi_2 + \cdots + \pi_c = 1$ so

$$\frac{n_1}{n_1}\pi_1 + \sum_{j=2}^{c} \frac{n_j}{n_1}\pi_1 = 1$$

$$= \frac{\pi_1}{n_1} + n_1 + \sum_{j=2}^{c} n_j = n$$

$$\Rightarrow \pi_1 = \frac{n_1}{n} = p_1$$

$$\pi_2 = \frac{n_2}{n_1}\pi_1 = \frac{n_2}{\cancel{n_1}}\frac{\cancel{n_1}}{n} = \frac{n_2}{n} = p_2$$

$$\pi_3 = \frac{n_3}{n_1}\pi_1 = \frac{n_3}{n_1}\frac{n_1}{n} = \frac{n_3}{n} = p_3$$

$$\vdots$$

$$\pi_c = \frac{n_c}{n_1}\pi_1 = \frac{n_c}{\cancel{n_1}}\frac{\cancel{n_1}}{n} = \frac{n_c}{n} \quad p_c$$

$$n_k \sim B(n, \pi_k) \quad p_k = \frac{n_k}{n} = \bar{y}_k$$

CLT says $\bar{y} \mathrel{\dot\sim} N(\mu, \frac{\sigma^2}{n})$, so

$$p_k \mathrel{\dot\sim} N\left(\pi_k, \frac{\pi_k(1-\pi_k)}{n}\right)$$

use for tests, confidence intervals

| Job Category | Frequency | Percent |
|---|---|---|
| Employed in Field | 106 | 53 |
| Employed outside Field | 74 | 37 |
| Unemployed | 20 | 10 |
| Total | 200 | 100.0 |

Find a 95% confidence interval for proportion unemployed.

using $p_k \dot{\sim} N\left(\pi_k, \frac{\pi_k(1-\pi_k)}{n}\right)$, CI is

$$p_k \pm 1.96\sqrt{\frac{p_k(1-p_k)}{n}} = 0.1 \pm 1.96\sqrt{\frac{0.1(1-0.1)}{200}}$$

$$= 0.1 \pm 0.042 \text{ or } (0.058, 0.142)$$

## Hypothesis tests

- Likelihood ratio tests
- Pearson Chi-squared tests

# Likelihood Ratio tests in general



Model $y_1, \ldots, y_n \overset{iid}{\sim} F_\beta$, $\beta \in \mathcal{B}$ ← parameter space

Parameter space is ~~the~~ the set of values the parameter (or parameter vector) can take.

Bernoulli: $\beta = \pi$, $\mathcal{B} = (0,1) = \{\pi : 0 < \pi < 1\}$

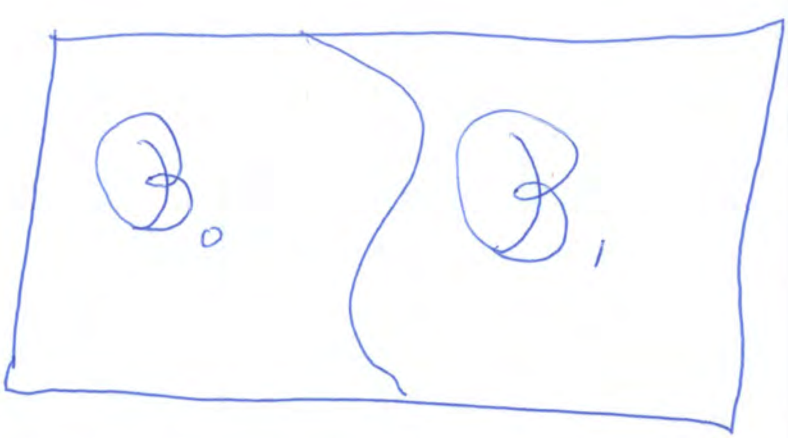
Poisson$(\lambda)$: $\beta = \lambda$, $\mathcal{B} = (0, \infty) = \{\lambda : \lambda > 0\}$

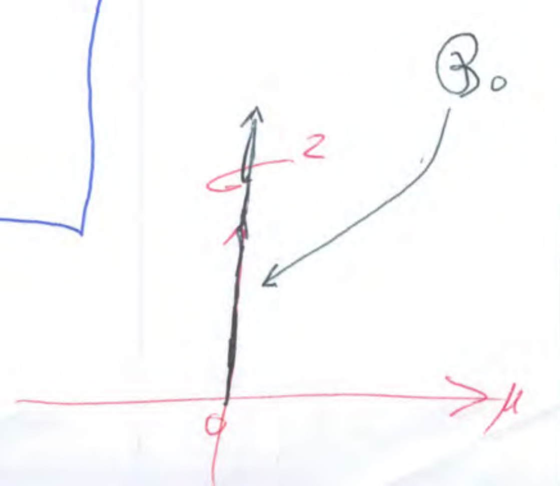
Normal$(\mu, \sigma^2)$ $\beta = (\mu, \sigma^2)$, $\mathcal{B} = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$

Null Hypothesis $H_0 : \beta \in \mathcal{B}_0$ vs $H_1 : \beta \in \mathcal{B}_1$

$$\mathcal{B}_0 \cup \mathcal{B}_1 = \mathcal{B}, \quad \mathcal{B}_0 \cap \mathcal{B}_1 = \emptyset$$

$\mathcal{B}$




Example: Normal data $H_0 : \mu = 0$

$\mathcal{B}_0 = \{(\mu, \sigma^2) : \mu = 0, \sigma^2 > 0\}$

$\mathcal{B} =$ 


$H_0: \beta \in \mathcal{B}_0$ Do 2 ML problems

Max $\ell(\beta)$ over whole parameter space $\mathcal{B}$

Max $\ell(\beta)$ over just $\mathcal{B}_0$

Form the ratio $\dfrac{\text{Max}_{\beta\in\mathcal{B}_0} \ell(\beta)}{\underset{\beta\in\mathcal{B}}{\text{Max}}\ \ell(\beta)} \le 1$ $\ell(\hat{\beta})$ MLE

If a <u>Lot</u> less than one, $H_0$ is questionable

test statistic $G^2 = -2 \log \dfrac{\underset{\beta\in\mathcal{B}_0}{\text{Max}}\ \ell(\beta)}{\underset{\beta\in\mathcal{B}}{\text{Max}}\ \ell(\beta)}$

Small LR, log is big neg #, $-2$ makes it large positive #.

Under some conditions when $H_0$ is true,

$$G^2 \overset{\cdot}{\sim} \chi^2(df = \dots)$$

$df$ = # of (non-redundant) constraints on $\beta$ imposed by $H_0$.

Suppose $y_1, \ldots, y_n \sim N(\mu, \sigma^2)$ $H_0: \mu = \mu_0$
One constraint

Regression $\beta_1 = \beta_2 = \beta_3 = 0$ 3 constraints

For ex $M(1, (\pi_1\ \pi_2\ \pi_3\ \pi_4))$

$$H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{1}{4}$$

One factor ANOVA (3 treatments)

$H_0$ $\mu_1 = \mu_2 = \mu_3$ 2 constraints

$$\mu_1 = \mu_2, \ \mu_1 = \mu_3, \ \mu_2 = \mu_3$$

$$\mu_1 = \frac{1}{2}(\mu_2 + \mu_3)$$

Still 2 constraints

Regression $H_0: L\beta = h$

$L$: $r \times p$, $\beta$: $p \times 1$, $h$: $r \times 1$

$r$ constraints

**Example**

University administrators recognize that the percentage of students who are unemployed after graduation will vary depending upon economic conditions, but they claim that still, about twice as many students will be employed in a job related to their field of study, compared to those who get an unrelated job. To test this hypothesis, they select a random sample of 200 students from the most recent class, and observe 106 employed in a job related to their field of study, 74 employed in a job unrelated to their field of study, and 20 unemployed. Test the hypothesis using a large-sample likelihood ratio test and the usual 0.05 significance level State your conclusions in symbols and words.

What is the model

$y_1, \dots, y_{200} \overset{iid}{\sim} M(1, (\pi_1, \pi_2, \pi_3))$

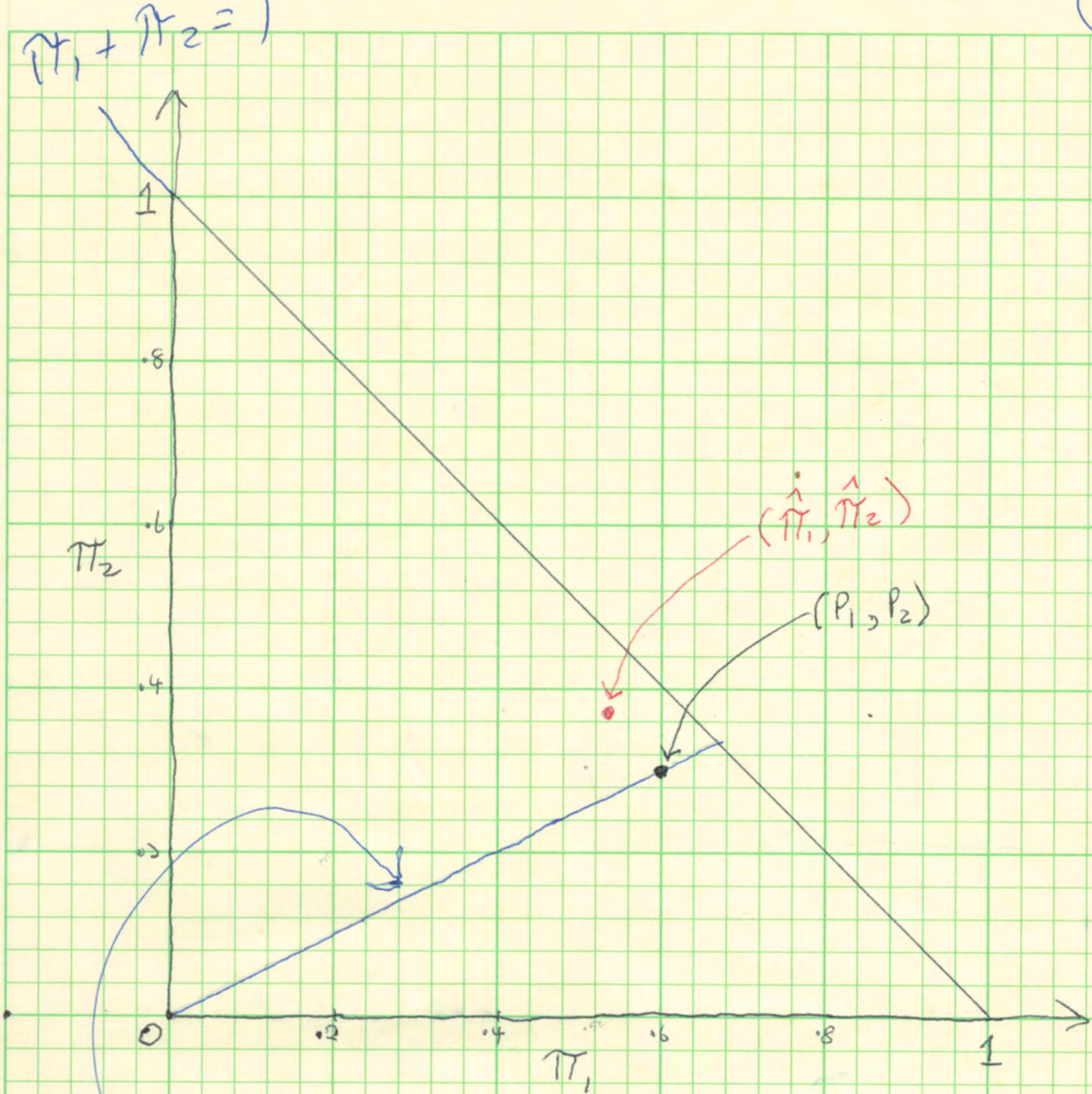
What is $H_0$? $H_0: \pi_1 = 2\pi_2$

df? $\underline{1}$

What is the parameter space $\mathcal{B}$?
Remember, one parameter is redundant.
there are 2 unknown parameters.

$$\mathcal{B} = \{(\pi_1, \pi_2): 0<\pi_1<1,\ 0<\pi_2<1,\ \pi_1+\pi_2<1\}$$

Draw it.




$$\mathcal{B}_0 = \left\{ (\pi_1, \pi_2) : \begin{array}{l} 0 < \pi_1 < 1 \\ 0 < \pi_2 < 1 \end{array},\ \pi_1 = 2\pi_2 \right\}$$

$$\pi_2 = \frac{1}{2}\pi_1$$

What is the unrestricted MLE. Just write it down.

17

$$\underset{\sim}{p} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}\right) = \left(\frac{106}{200}, \frac{74}{200}, \frac{20}{200}\right)$$

$$= (0.53, 0.37, 0.10)$$

Derive the restricted MLE. Your answer is a symbolic expression. It's a vector. Show your work.

There's really only one unknown. In $\mathcal{B}_0$,

$$\ell(\underset{\sim}{\pi}) = (2\pi_2)^{n_1} \pi_2^{n_2} (1 - 2\pi_2 - \pi_2)^{n_3}$$

$$\frac{\partial}{\partial \pi_2}\left(n_1 \log(2\pi_2) + n_2 \log \pi_2 + n_3 \log(1 - 3\pi_2)\right)$$

$$\frac{\partial}{\partial \pi_2}\left(n_1 \log 2 + n_1 \log \pi_2 + n_2 \log \pi_2 + n_3 \log(1 - 3\pi_2)\right)$$

$$= \frac{\partial}{\partial \pi_2}\left(n_1 \log 2 + (n_1 + n_2) \log \pi_2 + n_3 \log(1 - 3\pi_2)\right)$$

$$= 0 + \frac{n_1 + n_2}{\pi_2} + \frac{n_3}{1 - 3\pi_2}(-3) \overset{set}{=} 0$$

$$\frac{n_1 + n_2}{\pi_2} = \frac{3 n_3}{1 - 3\pi_2} \iff n_1 + n_2 - 3(n_1 + n_2)\pi_2 = 3 n_3 \pi_2$$

$$\Rightarrow n_1 + n_2 = 3\pi_2(n_1 + n_2 + n_3) \Rightarrow \hat{\pi}_2 = \frac{n_1 + n_2}{3n}$$

$$\pi_2 = \frac{n_1+n_2}{3n}, \quad \pi_1 = 2\pi_2 = \frac{2(n_1+n_2)}{3n}$$ 18

$$\pi_3 = 1 - \pi_1 - \pi_2 = 1 - \frac{n_1+n_2}{3n} - \frac{2(n_1+n_2)}{3n}$$

$$= 1 - \frac{\cancel{3}(n_1+n_2)}{\cancel{3}n} = \frac{n}{n} - \frac{n_1+n_2}{n}$$

$$= \frac{n-n_1-n_2}{n} = \frac{n_3}{n} = p_3$$

$$\hat{\underset{\sim}{\pi}} = \left( \overset{0.6}{\frac{2(n_1+n_2)}{3n}}, \overset{0.3}{\frac{n_1+n_2}{3n}}, \overset{0.1}{\frac{n_3}{n}} \right)$$

Calculate $G^2$. Show your work

$$G^2 = -2 \log \frac{\hat{\pi}_1^{n_1} \hat{\pi}_2^{n_2} \cancel{\hat{\pi}_3^{n_3}}}{p_1^{n_1} p_2^{n_2} \cancel{p_3^{n_3}}}$$

$$= -2 \log \left( \left(\frac{\hat{\pi}_1}{p_1}\right)^{n_1} \left(\frac{\hat{\pi}_2}{p_2}\right)^{n_2} \right)$$

$$= -2 \left( n_1 \log \frac{\hat{\pi}_1}{p_1} + n_2 \log \frac{\hat{\pi}_2}{p_2} \right)$$

$$= -2 \left( 106 \log \frac{0.6}{0.53} + 74 \log \frac{0.3}{0.37} \right)$$

$$= 4.739$$ qchisq(0.95,1) = 3.841 Reject $H_0$

| Job | $n_j$ | $p_j$ | $\hat{\pi}_j$ | $\hat{\mu}_j$ |
|---|---|---|---|---|
| Related | 106 | 0.53 | 0.60 | 120 |
| Unrelated | 74 | 0.37 | 0.30 | 60 |
| Unemployed | 20 | 0.10 | 0.10 | 10 |
| | 200 | | | |

$H_0: \pi_1 = 2\pi_2$ Rejected with $G^2 = 4.74 > 3.84$

Draw directional conclusion based on expected frequencies

Each $n_j \sim B(n, \pi_j)$ so $E(n_j) = n\pi_j = \mu_j$

Estimate $\mu_j$ with $\hat{\mu}_j = n\hat{\pi}_j$

| Job | Observed Freq | Expected Freq |
|---|---|---|
| Related | 106 | 120 |
| Unrelated | 74 | 60 |
| Unemployed | 20 | 20 |

| Job | Obs Freq | Exp Freq | Residual |
|---|---|---|---|
| Related | 106 | 120 | −14 |
| Unrelated | 74 | 60 | 14 |
| Unemp | 20 | 20 | 0 |

Job related to field of study is less likely than prediction from theory!

This derivation is cleaner than the way I did it in lecture

---

Express $G^2$ in terms of observed and expected frequencies.

From the formula sheet

$$G^2 = -2 \log \frac{\ell(\hat{\beta}_0)}{\ell(\hat{\beta})} = -2 \log \frac{\ell(\hat{\underset{\sim}{\pi}})}{\ell(\underset{\sim}{p})}$$

$$= 2 \log \left( \frac{\ell(\hat{\pi})}{\ell(p)} \right)^{-1} = 2 \log \frac{\ell(p)}{\ell(\hat{\pi})}$$

$$= 2 \log \frac{p_1^{n_1} p_2^{n_2} \cdots p_c^{n_c}}{\hat{\pi}_1^{n_1} \hat{\pi}_2^{n_2} \cdots \hat{\pi}_c^{n_c}} = 2 \log \left( \left( \frac{p_1}{\hat{\pi}_1} \right)^{n_1} \cdots \left( \frac{p_c}{\hat{\pi}_c} \right)^{n_c} \right)$$

$$= 2 \sum_{j=1}^{c} \log \left( \frac{p_j}{\hat{\pi}_j} \right)^{n_j} = 2 \sum_{j=1}^{c} n_j \log \frac{p_j}{\hat{\pi}_j}$$

$$= 2 \sum_{j=1}^{c} n_j \log \frac{n_j}{n \hat{\pi}_j} = 2 \sum_{j=1}^{c} n_j \log \frac{n_j}{\hat{\mu}_j}$$

Pearson Chisquared

$$X^2 = \sum_{j=1}^{c} \frac{(n_j - \hat{\mu}_j)^2}{\hat{\mu}_j} \quad \text{where } \hat{\mu}_j = n\hat{\pi}_j$$

$$G^2 = 2\sum_{j=1}^{c} n_j \log\left(\frac{n_j}{\hat{\mu}_j}\right)$$

Same df: Determined by $H_0$

For jobs data

$$X^2 = \frac{(106-120)^2}{120} + \frac{(74-60)^2}{60} + 0$$

$$= 4.9 \quad (\text{compare } G^2 = 4.74)$$

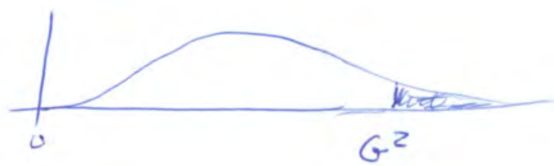
| Job Category | Observed Frequency | Expected Frequency |
|---|---|---|
| Employed in field | 106 | 120 |
| Employed outside field | 74 | 60 |
| Unemployed | 20 | 20 |

**Likelihood Ratio Test**

$$G^2 = 2\sum_{j=1}^{c} n_j \log\left(\frac{n_j}{\widehat{\mu}_j}\right)$$

```
> rm(list=ls())
> n_j = c(106,74,20); n = sum(n_j)
> pihat_j = c(0.6,0.3,0.1); muhat_j = n * pihat_j
> G2 = 2 * sum(n_j*log(n_j/muhat_j)); G2
[1] 4.739477
> critval = qchisq(0.95,df=1); critval # Also on the formula sheet
[1] 3.841459
> pval = 1-pchisq(G2,df=1); pval
[1] 0.02947803
```

$< 0.05$




**Pearson Chi-squared Test**

$$X^2 = \sum_{j=1}^{c} \frac{(n_j - \widehat{\mu}_j)^2}{\widehat{\mu}_j}$$

```
> X2 = sum((n_j-muhat_j)^2/muhat_j); X2
[1] 4.9
> pval = 1-pchisq(X2,df=1); pval
[1] 0.0268567
```

Rules of thumb

$G^2$ is OKAY if all expected freq are at least 5

$X^2$ " " " " " 1

23

# Is a die fair? Sample questions

Roll the die 300 times and observe these frequencies:

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 72 | 39 | 54 | 44 | 44 | 47 |

1. State a reasonable model for these data.

$y_1, \ldots, y_{300} \stackrel{iid}{\sim} M(1, (\pi_1, \ldots, \pi_6))$

2. Without any derivation, estimate the probability of rolling a 1. Your answer is a number.

$P_1 = \frac{72}{300} = 0.24$

3. Give an approximate 95% confidence interval for the probability of rolling a 1. Your answer is a set of two numbers. using $p \pm 1.96\sqrt{\frac{P(1-P)}{n}}$

$= 0.24 \pm 1.96\sqrt{\frac{0.24 * 0.76}{300}} = 0.24 \pm 0.048$

$= (0.192, 0.288)$

4. What is the null hypothesis corresponding to the *main question*, in symbols?

$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6$

5. What is the parameter space $\mathcal{B}$? Remember, there are really 5 unknown parameters.

$\mathcal{B} = \{(\pi_1, \ldots \pi_5) : 0 < \pi_j < 1, \sum_{j=1}^{5} \pi_j < 1\}$

6. What is the restricted parameter space $\mathcal{B}_0$?

$\mathcal{B}_0 = \{(\pi_1, \ldots \pi_5) : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \frac{1}{6}\}$

7. What are the degrees of freedom? The answer is a number.

5

8. What is the critical value of the test statistic at $\alpha = 0.05$? The answer is a number. Use R.

qchisq(0.95, 5) = 11.705

9. What are the estimated expected frequencies under $H_0$? Give 6 numbers.

$\hat{\mu}_j = n\hat{\pi}_j = 300 \cdot \frac{1}{6} = 50, 50$

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 50 | 50 | 50 | 50 | 50 | 50 |

|  | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $n_i$ OBS | 72 | 39 | 54 | 44 | 44 | 47 |
| $\hat{\mu}_i$ EXP | 50 | 50 | 50 | 50 | 50 | 50 |

24

10. Carry out the likelihood ratio test.

(a) What is the value of the test statistic? Your answer is a number. Show some work.

$$G^2 = 2 \sum_{j=1}^{c} n_j \log \frac{n_j}{\hat{\mu}_j} = 2\left(72 \log \frac{72}{50} + 3 \log \frac{39}{50} + \cdots + 47 \log \frac{47}{50}\right)$$

$$= 13.12 > 11.705$$

(b) Do you reject $H_0$ at $\alpha = 0.05$? Answer Yes or No.

(c) Using R, calculate the $p$-value.

1 - pchisq(13.12, 5) = 0.022

(d) Do the data provide convincing evidence against the null hypothesis?

Yes

11. Carry out Pearson test. Answer the same questions you did for the likelihood ratio test.

(a) $$X^2 = \sum_{j=1}^{c} \frac{(n_j - \hat{\mu}_j)^2}{\hat{\mu}_j} = \frac{(72-50)^2}{50} + \frac{(39-50)^2}{50} + \cdots + \frac{(47-50)^2}{50} = 14.05$$

(b) Yes, Reject $H_0$

(c) 1 - pchisq(14.05, df = 5) = 0.015

(d) Yes

25

12. Does the confidence interval for $\pi_1$ allow you to reject $H_0: \pi_1 = \frac{1}{6}$ at $\alpha = 0.05$? Answer Yes or No. CI was $(0.192, 0.288)$, $\frac{1}{6} = 0.1667$ outside CI, so Yes

13. In plain language, what do you conclude from the test corresponding to the confidence interval? (You need not actually carry out the test.)

There are more ones than expected ~~order~~ if die is fair.

Chances of a one are greater than 1/6.

14. Is there evidence that the chances of getting 2 through 6 are unequal?

(a) What is the null hypothesis?

$$H_0: \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6$$

(b) What is the restricted parameter space $\mathcal{B}_0$? It's convenient to make the first category the residual category.

$$\mathcal{B}_0 = \{(\pi_2, \ldots \pi_6): 0 < \pi_j < 1, \sum_{j=2}^{6} \pi_j < 1, \pi_2 = \pi_3 = \cdots = \pi_6\}$$

(c) Write the likelihood function for the restricted model. How many free parameters are there in this model?

$$l_0 = (1-5\pi_2)^{n_1} \pi_2^{n_2} \pi_2^{n_3} \pi_2^{n_4} \pi_2^{n_5} \pi_2^{n_6}$$

(d) Obtain the restricted MLE $\widehat{\boldsymbol{\pi}}$. Your final answer is a set of 6 numbers.

$$\frac{\partial}{\partial \pi_2} \log l_0 = \frac{\partial}{\partial \pi_2} \log\left((1-5\pi_2)^{n_1} \pi_2^{n-n_1}\right)$$

$$= \frac{\partial}{\partial \pi_2}\left(n_1 \log(1-5\pi_2) + (n-n_1)\log \pi_2\right)$$

$$= \frac{n_1(-5)}{1-5\pi_2} + \frac{n-n_1}{\pi_2} \stackrel{set}{=} 0$$

$$\Longleftrightarrow \frac{5n_1}{1-5\pi_2} = \frac{n-n_1}{\pi_2}$$

$$\Longleftrightarrow 5n_1\pi_2 = n-n_1 - 5\pi_2(n-n_1)$$

$$\Longleftrightarrow 5n_1\pi_2 + 5(n-n_1)\pi_2 = n-n_1$$

$$\Longleftrightarrow 5\pi_2\left(\cancel{n_1} + n \cancel{-n_1}\right) = n-n_1$$

$$\Longleftrightarrow \pi_2 = \frac{n-n_1}{5n}$$

$$\pi_1 = 1 - 5\pi_2 = 1 - \cancel{5}\,\frac{n-n_1}{\cancel{5}n}$$

$$= \cancel{1-1} + \frac{n_1}{n} = \frac{n_1}{n}$$

$$\hat{\underset{\sim}{\pi}} = \left(\frac{n_1}{n}, \frac{n-n_1}{5n}, \frac{n-n_1}{5n}, \cdots \frac{n-n_1}{5n}\right)$$

$$= \left(\frac{72}{300}, \frac{228}{\underset{1500}{\cancel{300}}}, \cdots, \frac{228}{\underset{1500}{\cancel{300}}}\right)$$

$$= (0.24, 0.152, 0.152, 0.152, 0.152, 0.152)$$

|  | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Obs | 72 | 39 | 54 | 44 | 44 | 47 |
| Exp | 72 | 45.6 | 45.6 | 45.6 | 45.6 | 45.6 |

27

(e) Give the estimated expected frequencies $(\widehat{\mu}_1, \ldots, \widehat{\mu}_6)$.

$$n\hat{\underset{\sim}{\pi}} = 300(0.24, 0.152 \; -- \;, 0.152$$

$$= (72, 45.6, 45.6, 45.6 \; -- \; 45.6)$$

(f) Calculate the likelihood ratio test statistic. Your answer is a number.

$$G^2 = 2\sum_{j=1}^{c} n_j \log \frac{n_j}{\hat{\mu}_j} = 2\left(\cancel{72 \log \frac{72}{72}} + 39 \log \frac{39}{45.6} + \cdots + 47 \log \frac{47}{45.6}\right)$$

$$= 2.62$$

(g) Do you reject $H_0$ at $\alpha = 0.05$? Answer Yes or No.

Crit value with df = 4 is 9.488, so no

(h) Using R, calculate the $p$-value.

1 - pchisq(2.62, 4) = 0.62 > 0.05.

(i) Do the data provide convincing evidence against the null hypothesis?

No

(j) What do you conclude, in words

There is no evidence that the chances of 2 through 6 are unequal.