Lecture 5

Proportion data: chi-square tests

比例数据：卡方检验

卡方检验=比例检验

Chi-square tests = proportion tests

- Sometimes data are presented as proportions:

  有时数据以比例表示

- ***One-sample proportion*** 单样本比例
  - Is there a difference in the proportion of women and men among UK millionaires?

  这里只有一个“独立”比例，即加起来是100%

  (**NOTE**: only one 'independent' proportion here: if women are 52%, men must be 48% - proportions must add up to 100%)
  - Proportion of children vs. adults in a village

- ***Two independent proportions*** 两个独立的比例
  - Difference in proportion of married people in London vs. Glasgow?
  - (**NOTE**: here, proportions do not need to add up to 100% 这里加起来不需要是100%

- To analyse proportions, we use *proportion tests* (aka *chi-square tests*)

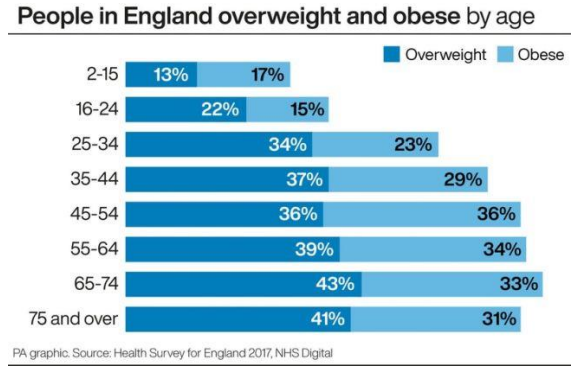
  为了分析比例，我们使用比例检验(卡方检验)

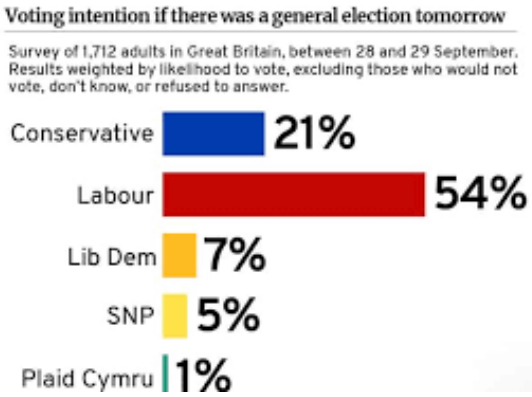
Applications

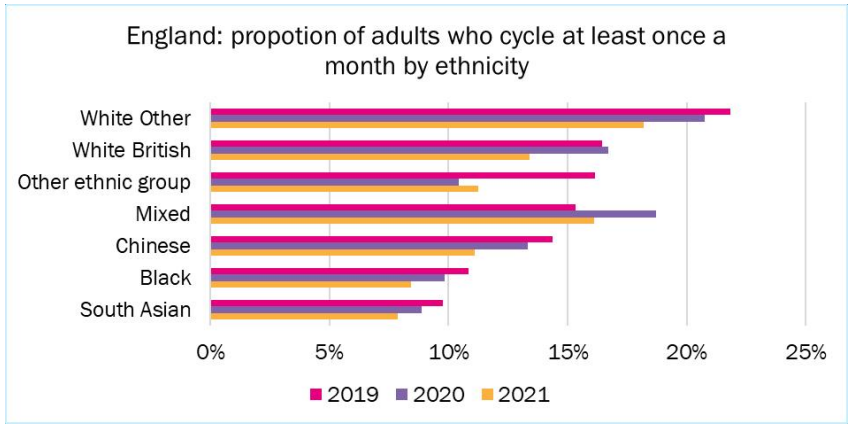
比例分析在各个领域都有广泛的应用

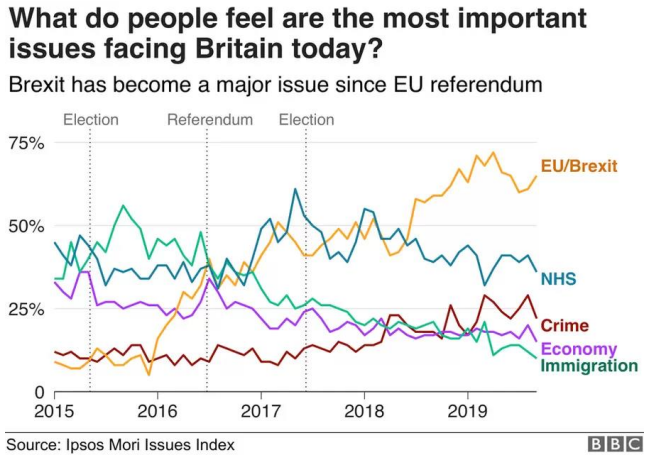
- Proportion analysis is widely applied in all fields!

- Market research
- Opinion polls
- Behavioural trends over time
- Public health
- etc.



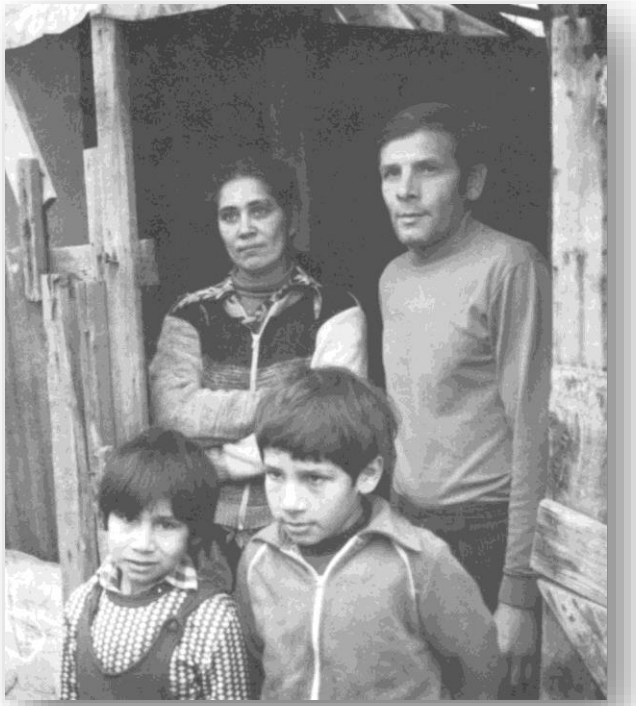






Proportion test, one sample: sex ratios

- As a rule, sex ratio at birth is ~1 in humans 通常，人类的出生性别比例约为1
  - very similar number of newborn girls and boys
  - slightly male-based ratio at ~1.05, to compensate for higher male infant mortality



- But Berenczi and Dunbar (1997) identified an exception: a female-biased ratio at birth among Hungarian gypsies

- Their explanation:
  - rural and urban gypsies are poorer than neighbour Hungarians
  - daughters often marry richer Hungarians; sons very unlikely to marry out

- Their prediction: natural selection should produce an excess of females *at birth*
  - (note: sex ratio *at birth* is unaffected by infanticide etc.)

- But let's re-examine the evidence presented in the study

- In Hungarian gypsies, is sex ratio at birth significantly different from p=0.5?

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

本讲中p为比例

note: in this lecture, *p* is proportion; *P* is significance value
(careful: *R* output does not make the distinction!)

单样本比例检验

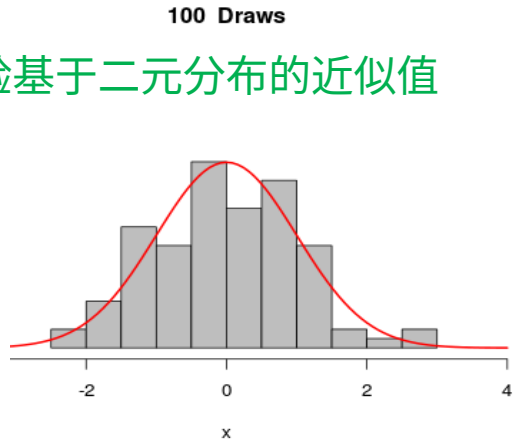
One-sample proportion test

- =tests the likelihood of a proportion p deviating from a test value 测试比例p偏离测试值的可能性
  - if there are two proportions, predicted proportion of each is p = 0.5
  - for 3 proportions, predicted p for each is p = 0.33

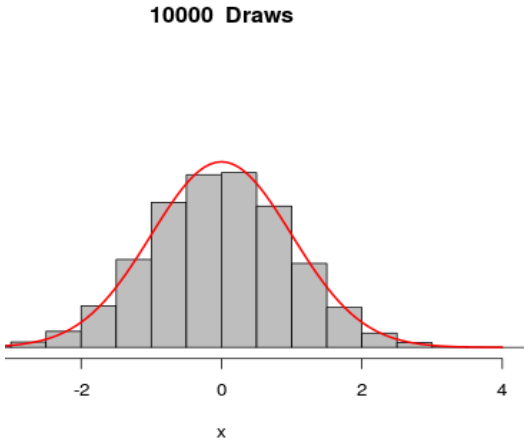
如果有两个比例，每个比例的预测比例为p=0.5 对于三个比例，每个是p=0.33

- Proportion test is based on an approximation to the binary distribution estimating probability of x positives out of n attempts
  - e.g. coin tossing: binary distribution estimates probability of x heads, each one with p(head)=0.5, in n tosses
  - expected mean: np (=10*0.5)=5; you expect 5 heads in 10 tosses); variance: np(1-p)

比例检验基于二元分布的近似值



- Test statistic: u= $\frac{x-np}{\sqrt{np(1-p)}}$ 检验统计量u
  - u statistic similar to t statistic: a standardised difference between observed proportion and a test value
  - u distribution approaches normal with large n 当n较大时，u分布趋近于正态分布
  - (alternative u^2, which has a chi-square distribution also approaching normality with large n)



To test the likelihood of say x=6 heads in n=15 tosses:

> prop.test(x=positives, n=total, p=proportion tested)

(default p: p=0.5)

x是pos，n是总数，p是测试比例

Gypsy sex ratios

- Are there fewer gypsy boys than girls at birth?

1) Rural gypsies, *all babies*:

- sex ratio (=boys/girls)=89.3%=0.893, n=254
  =>120 boys, 134 girls, (120/134=0.893)
  - Proportion of boys (=boys/total)= 120/254 = 0.47
  - **Question: is p=0.47 different from 0.5?**

```
> prop.test(120, 254, 0.5)
1-sample proportions test with continuity correction
data:  120 out of 254, null probability 0.5
X-squared = 0.6654, df = 1, p-value = 0.4147
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4099875 0.5357417
sample estimates:
      p
0.4724409
```

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

零假设：比例p=0.5

Null hypothesis: proportion p=0.5

- P=0.41　与p=0.55和平衡的1:1性别比例无显著差异
- =>no significant deviation from p=0.55 and from balanced 1:1 sex ratio
- accept null hypothesis　接收零假设
- test value *p*=0.5 is included in 95% CI

测试值p=0.5包含在95%的CI里

- ***no evidence of fewer boys than girls in rural Roma***

Note: calculating boys (b) and girls (g)

- Table shows:
  - Sample size n = b + g
  - (males/females) x 100 (i.e. shown as percentage)

- Rural gypsy: n=254 and male/female = 89.3; then
  - n = b + g = 254
  - b/g = 0.893

therefore

girl = 总数 / (1 + 男/女)

- g = n/(1 + sex ratio)

- Urban gypsies

Exercise:

Run the same test with urban gypsies, all children (n=239, ratio=0.897)

Does the sex ratio at birth differ from p=0.5?

一胎

First-borns only

- But decision to have a second baby or abort may depend on sex of previous offspring
  - if 1st child is a boy, Roma parents are more likely to try again (until child is female);
  - if it is a girl, they are more likely to stop

- Solution: analysing only *1st-born* rural babies:
  - sex ratio=0.813, n=87 => boys=39, girls=48
  - *p*(boys)=39/87=0.448

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

```
> prop.test(39, 87, 0.5)
1-sample proportions test with continuity correction
data:  39 out of 87, null probability 0.5
X-squared = 0.7356, df = 1, p-value = 0.3911
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3427920 0.5583697
sample estimates:
      p
0.4482759
```

- P=0.39
- 95% CI includes test value *p*=0.5
- No evidence of fewer boys in rural Gypsy 1st-borns

- ***One-sample proportion tests do not show any evidence that rural gypsies have a female-biased sex ratio***

Exercise:

Run the same test with urban gypsies, first-born children only (n=77, ratio=0.943)

Does the sex ratio at birth differ from p=0.5?

卡方检验用于检验观察到的类别变量的分布是否与期望的不同

1.单因素卡方检验(卡房拟合度检验)：一个分类变量的预期频率与观察到的频率相比是否存在显著差异

Break

2.二因素卡方检验(独立性卡方检验)：检验两个类别变量之间是否存在关系

两个独立的比例

Two independent proportions

- But proportion of boys in rural gypsies vs. Hungarians could still differ *from each other*
  - i.e. we can compare the two independent proportions of boys in rural Gypsies vs. Hungarians
  - (maybe proportion of Roma boys is 48%, but of Hungarian boys it is 55%)

- Now we need *two-sample proportion tests*
  - based on *u*-statistic (=difference between proportions divided by pooled variance) and chi-square distribution

基于u统计量(=比例之间的差异除以合并方差)和卡方分布

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

Syntax:

```
> prop.test(c(x1, x2…), c(n1, n2…))
```

n是总数

- *xi*= positive cases (=boys) in group 1 and group 2
- *ni*=total cases (=boys+girls) in group 1 and group 2

Rural Gypsies vs. Hungarians, all babies

```
> prop.test(c(120,114), c(254,216))

2-sample test for equality of proportions with continuity correction

data:  c(120, 114) out of c(254, 216)
X-squared = 1.2171, df = 1, p-value = 0.2699
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.15018442  0.03951076
sample estimates:
   prop 1    prop 2
0.4724409 0.5277778
```

零假设是两方是否有显著差异

All rural babies:

- Rural gypsies: 254 total, 120 boys; *p*(boys)=0.47
- Rural Hungarians: ratio 1.118, 216 total -> 114 boys, 102 girls; *p*(boys)=0.527

Result:

- P>0.26
- 95% CI includes zero (no difference)

95%CI包括零(无差异)

Conclusion: no significant difference in sex ratios between rural populations

Exercise:

Run the same test two independent proportions test on urban children

Does the sex ratio at birth differ between Gypsies and Hungarians?

仅限一胎

First births only

- But now take only first births:
  - Rural gypsies: 39 boys, 48 girls, n=87
  - Rural Hungarians: 52 boys, 33 girls, n=85

```
> prop.test(c(39,52), c(87,85))

2-sample test for equality of proportions
with continuity correction

data:  c(39, 52) out of c(87, 85)
X-squared = 3.9795, df = 1, p-value = 0.04606
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.322272741  -0.004704946
sample estimates:
   prop 1          prop 2
0.4482759    0.6117647
```

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

差异显著

- Finally a significant difference!
  - P<0.05 (just about)
  - 95% CI excludes zero
  - difference between p(boys)=0.448 (Gypsies) and p(boys)=0.61 (Hungarians) is significant

- Conclusion: ***rural gypsies show lower proportion of first-born boys than rural Hungarians***

低于

Exercise:

Run the same test two independent proportions test on urban children, first-born only

Does the sex ratio at birth differ between Gypsies and Hungarians?

```
> prop.test(c(37,58),c(77,102))

 2-sample test for equality of proportions with continuity correction

data:  c(37, 58) out of c(77, 102)
X-squared = 1.0367, df = 1, p-value = 0.3086
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.24678311  0.07056717
sample estimates:
   prop 1    prop 2
0.4805195 0.5686275
```

Conclusions

- The only test that shows a difference between sex ratios in Roma and Hungarians is a two-independent sample comparing rural, first-born children

- But in that test, the 'abnormal' population seems to be rural Hungarians, with a very high number of boys among first-borns!
  - they have 57.6% more boys than girls!!!
  - (but is this ratio significant?)

- Very biased conclusion from the authors of the study!
  - They ran test after test until they obtained the significant result they "wanted" to find
  - (we'll see how to fix this later)

Table 2. *Sex ratios at birth for each population*

	number of sons per 100 daughters			
	rural populations		urban populations	
	Gypsy	Hungarian	Gypsy	Hungarian
A. all children				
sample size	254	216	239	224
males/100 females	89.3	111.8	89.7	113.3
B. first-born children only				
sample size	87	85	77	102
males/100 females	81.3	157.6	94.3	131.8

其他情况：两个以上独立比例

Other cases: more than two independent proportions

- If you want to compare more than two independent proportions
  - e.g. proportion of boys among rural gypsies, rural Hungarian, urban gypsies, urban Hungarians ***all at the same time***)

延长

- Just extend *prop.test* to four populations

这里的n指总数

```
> prop.test(c(x1, x2, x3, x4), c(n1, n2, n3, n4))
```

以矩阵形式输入

- Function *chisq.test* is similar to *prop.test*, but you enter it in matrix form

两个计算出来的p什么的都是一样的

Syntax:

```
> chisq.test(matrix(c(x1, x2, x3, x4, n1, n2, n3, n4), nrow=m))
```

Enter all positives (boy numbers) first, then all negatives (girl numbers)

Important! Here n = negatives (in our case, girls), not total!!!

Matrix is read by column (default: byrow=F)

这里的n在例子中指女孩数量，而不是总数

(m is the number of groups or rows in the matrix; in the example above, m=4)

m是矩阵中的组数或者行数，在例子中是4

Exercise:

1) Run prop.test on the four proportions: rural gypsies, rural Hungarian, urban gypsies, urban Hungarians

2) Run a chi-square test on the same four proportions

其他情况：一个样本n个比例

Other cases: one sample, n proportions

- You may want to test whether a die (instead of a coin) is loaded
  - now there are six proportions (1/6 for each side) that add up to 100%
  - tested proportion is now p=1/6=0.17

- This test can be done with function *chisq.test*

```
> chisq.test(matrix(c(x1, x2, x3, x4, x5, x6), nrow=1))
```

x1是掷骰子得到1的次数等

- x1 = number of times you got a 1 rolling the die, etc.
- nrow= 1 (meaning all 6 values are from the same die; since only one goup here, parameter not necessary)

nrow=1表明所有6个值来自同一个骰子，因为只有一个组
所以参数不是必须的

二项检验

Binomial test

二项分布是放回抽取(独立重复)

- The binomial test is equivalent to a *prop.test*, except that it is based on the binomial distribution itself 他是基于二项分布本身

要注意是单侧还是双侧检验

观测值，样本总量，(检验比率)

```
> prop.test(120, 254, 0.5)
1-sample proportions test with continuity correction
data:  120 out of 254, null probability 0.5
X-squared = 0.6654, df = 1, p-value = 0.4147
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4099875 0.5357417
sample estimates:  p
0.4724409
```

有些人喜欢这个因为它估计一个精确的p值

- Some prefer *binom.test* because it estimates an exact P value
  - contrary to *prop.test* that calculates P value from a normal approximation to the binomial

```
> binom.test(120,254)
        Exact binomial test
data:  120 and 254
number of successes = 120, number of trials = 254, p-
value = 0.4147
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.4097151 0.5358173
sample estimates: probability of success
          0.4724409
```

prop是从正态分布近似值计算p的值

费歇尔精确检验

Fischer exact test

- Fisher exact test also calculates exact *P* value

  输入positive和negative，而不是总数

- Now you enter the positive *x* cases (e.g. boys) and the ***negative*** cases (e.g girls), ***instead of total n***
- Test is based on odds-ratios not proportions 测试是基于赔率而不是比例
  - 95% CI looks different
  - if odds ratio (of boys to girls) is different from 1, proportions differ

如果优势比不等于1，则比例不同

Syntax:

>fisher.test(matrix(c(pos1, pos2,…, neg1, neg2,…), *m*))

m= number of compared groups

m是比较组的数量

fisher精确检验也计算精确的p值

Using data from urban gypsies vs. Hungarians

```
> fisher.test(matrix(c(39, 52, 48, 33), 2))

Fisher's Exact Test for Count Data

data:  matrix(c(39, 52, 48, 33), 2)
p-value = 0.03396
alternative hypothesis: true odds ratio is not
equal to 1
95 percent confidence interval:
 0.2684001 0.9885450
sample estimates:
odds ratio
 0.5176459
```

一般化

Generalisation

卡房或者fisher精确检验可以同时推广到任意数量的样本和比例

- Chi-square or Fisher exact tests can be generalised for any number of samples and proportions *at the same time*

- For example, does choice of degree (Archaeology, Biology, Engineering) at UCL affect the final grade (1st, 2.1, 2.2, 3rd class degree)?
  - results per degree (number of 1st, 2.1, 2.2, 3rd for each degree, which add up to 1)
  - number of degrees *nrow* (in this case, nrow=3)

- Syntax:

>chisq.test(matrix(c($n_{1st,arc}$, $n_{1st,bio}$, $n_{1st,eng}$, $n_{2.1,arc}$, $n_{2.1,bio}$, $n_{2.1,eng}$, $n_{2.2,bio}$, $n_{2.2,arc}$, $n_{2.2,eng}$, $n_{3rd,arc}$, $n_{3rd,bio}$, $n_{3rd,eng}$), nrow=3))

fisher检验的优点是可以计算出精确的p值

- advantage of Fisher test is to calculate exact *P* value