107-1 Statistics

LAB13: RELATIONSHIP OF CATEGORICAL VARIABLES

助教：廖皓宇、吳家禎、賴冠宇

2018/12/21

1221實習：類別資料的相關性

- 使用提供的資料(heartatk.csv)
- 相關性分析
  - Statistics about risk
  - Chi-square test for two-way table
  - Goodness-of-fit test
- R程式練習題作業講解

12/21 LAB

1. Statistics about risk
   ① By manual calculation
   ② By functions:
      - *library(epitools)*
2. χ^2 test – homogeneity, independence
   ① By manual calculation
   ② By function:
      - *chisq.test()*
3. Goodness-of-fit test
   ① Testing uniform distribution
   ② Testing binomial distribution ← **HW: R程式練習題2**
   ③ Testing normal distribution ← **HW: R程式練習題3**

Goodness-of-fit test: Testing uniform distribution

- Signi$ficance\ level$: $\alpha = 0.05$

Step 1.
- $\boldsymbol{H_0}$: $\boldsymbol{x}$ 服從均一分配
- $\boldsymbol{H_1}$: $\boldsymbol{x}$ 不服從均一分配

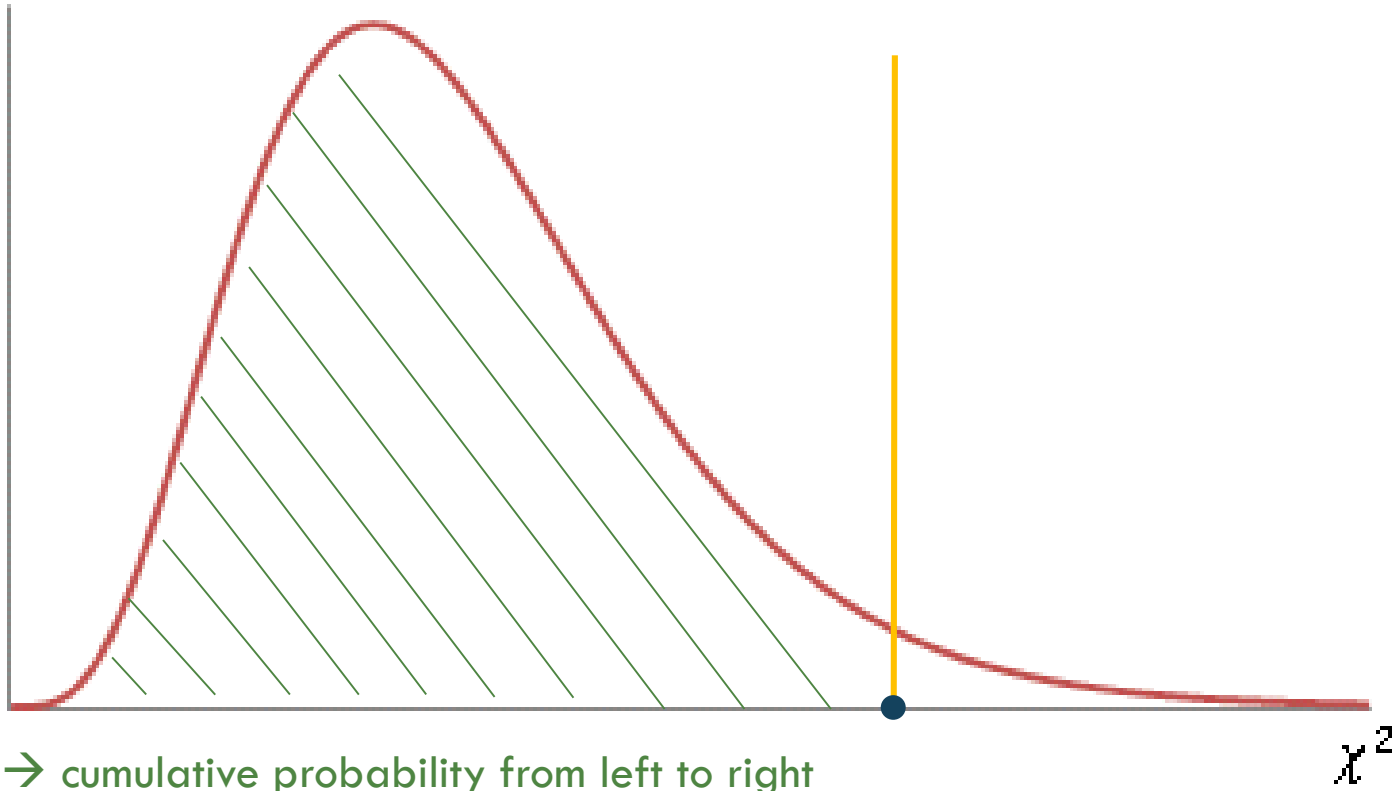
Step 2.
- $Checking\ conditions$
- $\chi^2 = 6.04$

Step 3.
- $p\ value = 0.7359083$

Step 4.
- $p\ value > \alpha.\ \ Do\ not\ Reject\ H_0.$

Step 5.
- $\boldsymbol{x}$ 服從均一分配。換言之，每個數字被抽中的機率相同。

R functions for Chi-square distribution



pchisq(): χ^2 and df → cumulative probability from left to right

qchisq: cumulative probability, df → χ^2

dchisq(): probability

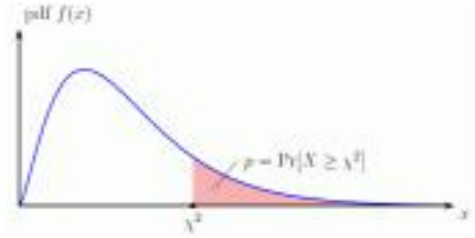
rchisq(): sampling

查表

表頭：cumulative probability (α)

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38



自由度

χ^2值

R程式練習題 1

1. 自訂函數MyChiSq
   - Function input: a two-way table
   - Function output: χ^2 , df, p-value

   - 可以heartatk.csv作為函數之測試資料
   - Table中的某一個數值：

```
for (i in 1:nrow(table)) {
  for (j in 1:ncol(table)) {
    table[i,j] =
  }
}
```

R程式練習題 2

Testing binomial distribution

Goodness-of-fit test 1

- 過去90週，每週會發生降雨的日數如下：

每週降雨日數	0	1	2	3	4	5	6	7
週數	5	13	26	19	20	7	0	0

- 定義隨機變數X：一週內的降雨日數，在顯著水準α=0.05，檢定X是否服從binomial distribution (二項分配)?
- 提示：dbinom(x, size, prob)

R程式練習題 2

Testing binomial distribution

- The probability when x = __
  - **dbinom(x = , size = , prob =)**
- The expected value for each x = __
  - **n × probability**
- Checking conditions in step 2
  - Combining rows in order to satisfy the conditions

```
#combining 6,7,8 rows
tbl.678 = tbl[6,] + tbl[7,] + tbl[8,] #add up 6,7,8 rows
tbl = tbl[-c(6,7,8),] #remove 6,7,8 rows
tbl = rbind(tbl, tbl.678) #add the combined row
```

- Degree of freedom: k-1-r
  - K: 組數
  - r: 估計的參數數量 ← **二項試驗中的p（從樣本資料推估而來）**
    = k-1-1

To review binomial distribution, see Stat_lab4_1005.pdf

QUICK REVIEW: Binomial distribution

Binomial conditions:

① N trials
② Only 2 choices (eg. Win vs. lose, sick vs. not sick)
③ Each trial is independent to others
④ The probability of 2 choices: p and (1-p)

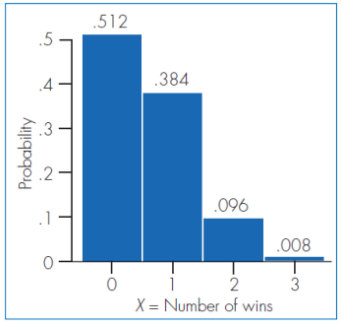
↓

Random variable (x): in N trials, the count of a certain choice

↓ *With multiple times…*

The distribution of random variable (x)

- The probability of $x = 1$



R程式練習題 3

Testing normal distribution

Goodness-of-fit test 2

- 環保署某測站連續監測40個月，每月平均PM2.5濃度值(μg/m³)如下：

18.8	14.6	14.0	15.8	12.4	13.2	16.1	13.8	16.2
16.1	17.8	18.7	15.8	13.3	13.6	16.4	13.8	16.6
15.3	19.0	18.4	15.0	18.8	18.1	17.3	16.3	17.5
18.1	14.2	18.0	13.0	13.3	12.4	16.6	14.1	20.6
16.8	13.3	18.2	16.9					

- 定義隨機變數X：每月平均PM2.5濃度值，在顯著水準α=0.05，檢定X是否服從normal distribution (常態分配)?
- 提示：pnorm(q, mean, sd); qnorm(p, mean, sd)

R程式練習題 3

Testing normal distribution

- 將連續資料離散化以進行卡方檢定

- Degree of freedom: k-1-r
  - K: 組數
  - r: 估計的參數數量　←計算期望值時的平均值及標準差（從樣本資料推估而來）
    = k-1-2

作業12 類別資料的相關性

- 練習題6題(Ch. 4+15)
  - 4.12; 4.26; 4.36; 15.18; 15.26; 15.40
- R程式練習題(繳交程式碼與執行結果)

1. – 自訂函數MyChiSq，使用heartatk.csv資料，檢定SEX與DIED間的獨立性，並與內建函數的計算結果相互驗證
2. – Goodness-of-fit test 1：如下頁投影片
3. – Goodness-of-fit test 2：如下下頁投影片