Lecture 9. Tests For Categorical Data

CONTENTS

1. Chi-square goodness of fit
2. Two sample test for binomial proportions
3. Fisher's exact test
4. Two sample test for binomial proportions for matched data
5. R×C contingency tables
6. Correlation analysis of categorical variable data

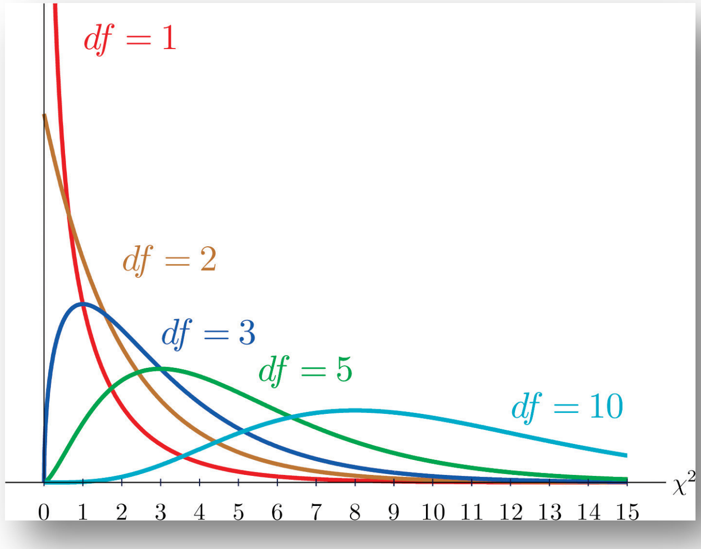
Recall

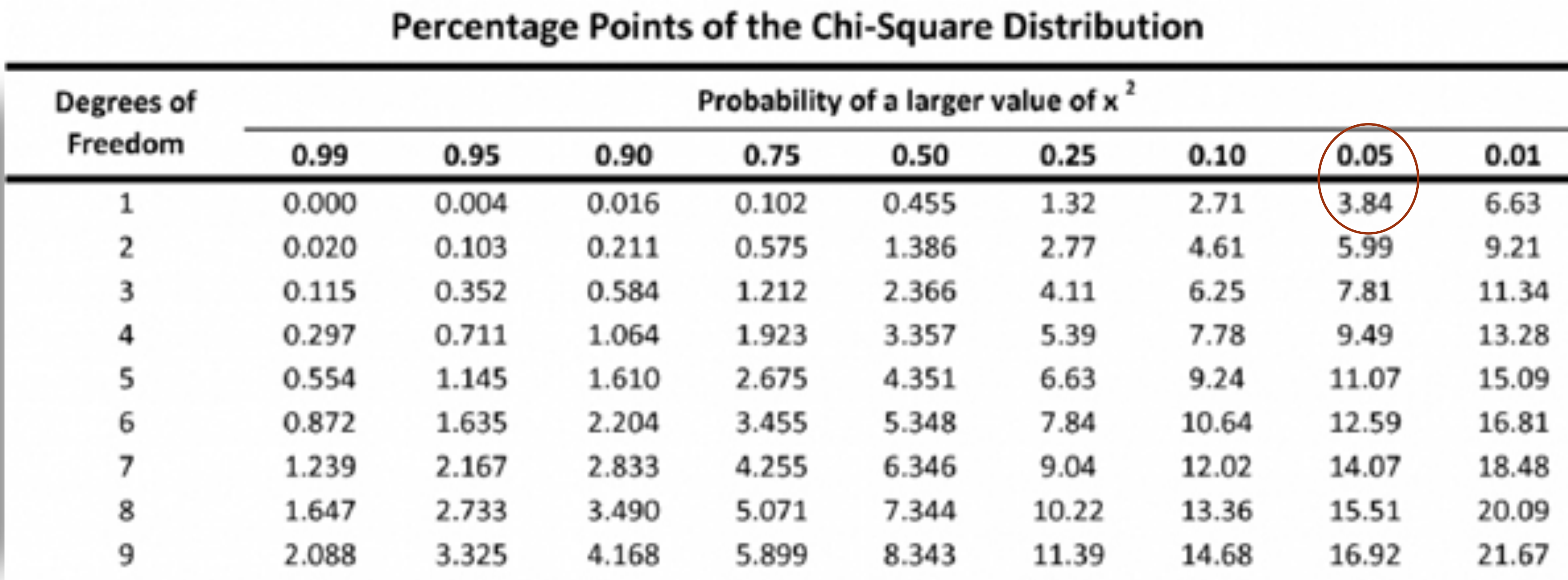
Sampling from k independent standard normal distributions and then square and sum the values, you'll produce a chi-square distribution with k degrees of freedom.

$$X_k^2 = (z_1)^2 + (z_2)^2 + ... + (z_k)^2$$

$$Y \sim x_n^2$$

$$x^2_{0.05,1} = 3.84 = 1.96^2$$



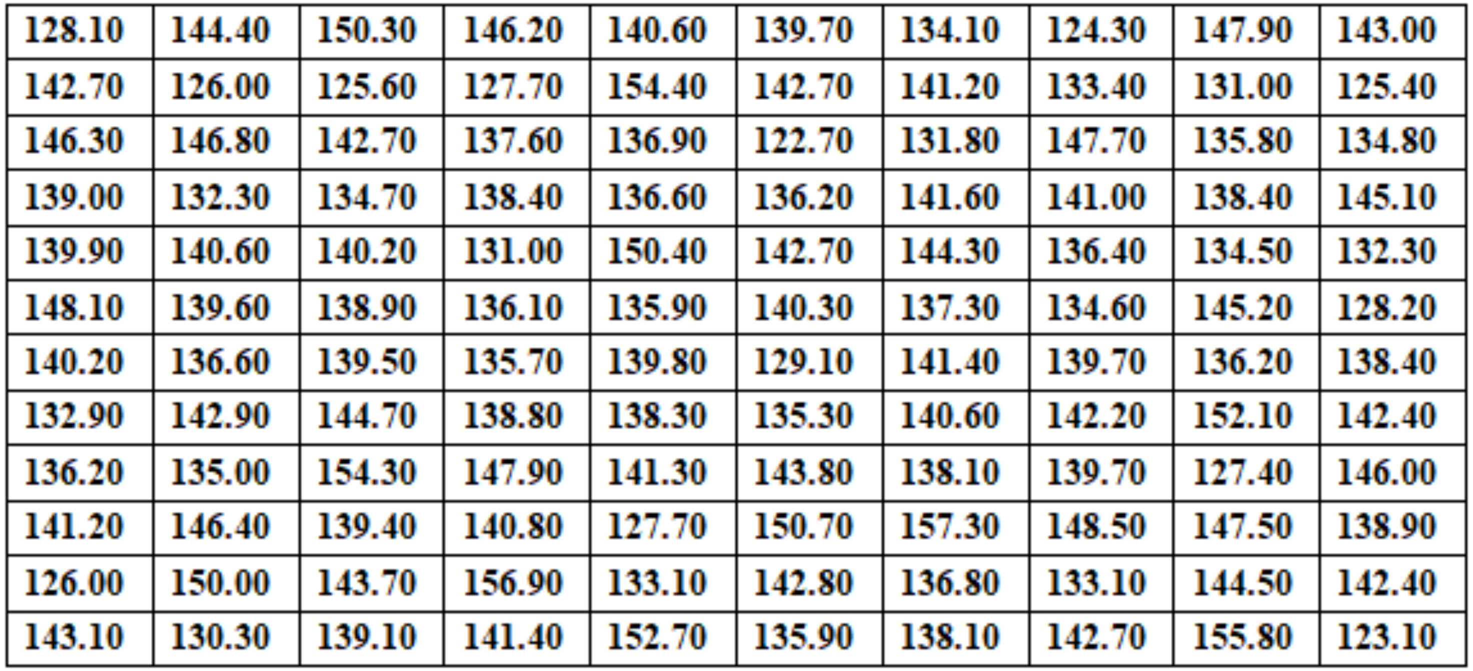


Percentage Points of the Chi-Square Distribution

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

1. CHI-SQUARE GOODNESS OF FIT

- Example 1: We already known that the height (in cm) of 12-year-old boy in Suzhou follows a normal distribution N(139.48,7.3^2). Suppose we randomly selected 120 boys aged 12 from Suzhou Industrial Park, and created a frequency table, as follows,



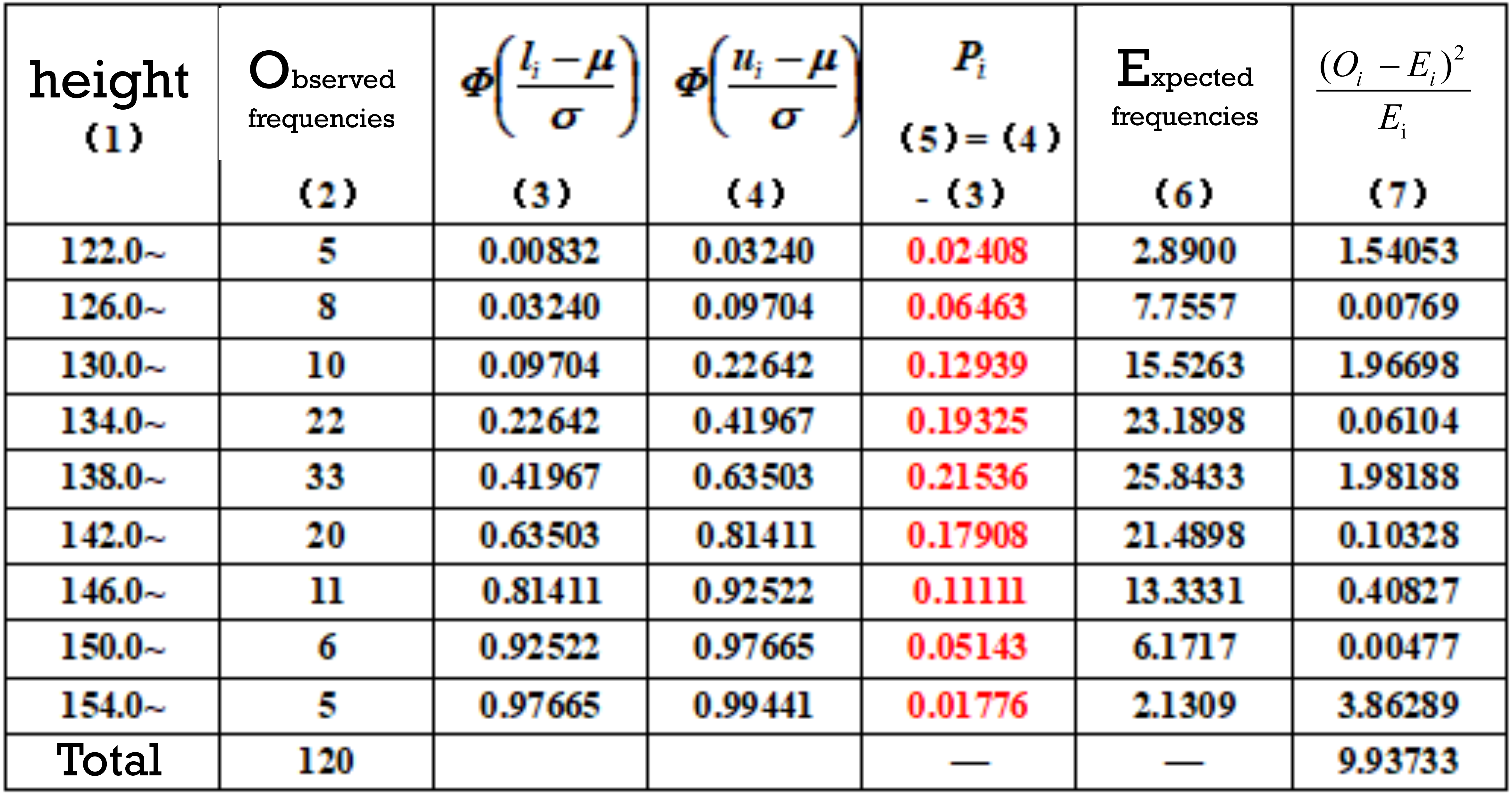
128.10	144.40	150.30	146.20	140.60	139.70	134.10	124.30	147.90	143.00
142.70	126.00	125.60	127.70	154.40	142.70	141.20	133.40	131.00	125.40
146.30	146.80	142.70	137.60	136.90	122.70	131.80	147.70	135.80	134.80
139.00	132.30	134.70	138.40	136.60	136.20	141.60	141.00	138.40	145.10
139.90	140.60	140.20	131.00	150.40	142.70	144.30	136.40	134.50	132.30
148.10	139.60	138.90	136.10	135.90	140.30	137.30	134.60	145.20	128.20
140.20	136.60	139.50	135.70	139.80	129.10	141.40	139.70	136.20	138.40
132.90	142.90	144.70	138.80	138.30	135.30	140.60	142.20	152.10	142.40
136.20	135.00	154.30	147.90	141.30	143.80	138.10	139.70	127.40	146.00
141.20	146.40	139.40	140.80	127.70	150.70	157.30	148.50	147.50	138.90
126.00	150.00	143.70	156.90	133.10	142.80	136.80	133.10	144.50	142.40
143.10	130.30	139.10	141.40	152.70	135.90	138.10	142.70	155.80	123.10

Height	observed value
122-	5
126-	8
130-	10
134-	22
138-	33
142-	20
146-	11
150-	6
154-	5
total	**120**

- We would like to test whether these measurements came from N(139.48,7.3^2).

This assumption can be tested by first computing what the expected frequencies would be in each group if the data did come from an underlying normal distribution and by then comparing these expected frequencies with the corresponding observed frequencies.

Remember how to calculate the expected frequencies for a normal distribution?



height (1)	Observed frequencies (2)	$\Phi\left(\frac{l_i-\mu}{\sigma}\right)$ (3)	$\Phi\left(\frac{u_i-\mu}{\sigma}\right)$ (4)	P_i (5)=(4) - (3)	Expected frequencies (6)	$\frac{(O_i-E_i)^2}{E_i}$ (7)
122.0~	5	0.00832	0.03240	0.02408	2.8900	1.54053
126.0~	8	0.03240	0.09704	0.06463	7.7557	0.00769
130.0~	10	0.09704	0.22642	0.12939	15.5263	1.96698
134.0~	22	0.22642	0.41967	0.19325	23.1898	0.06104
138.0~	33	0.41967	0.63503	0.21536	25.8433	1.98188
142.0~	20	0.63503	0.81411	0.17908	21.4898	0.10328
146.0~	11	0.81411	0.92522	0.11111	13.3331	0.40827
150.0~	6	0.92522	0.97665	0.05143	6.1717	0.00477
154.0~	5	0.97665	0.99441	0.01776	2.1309	3.86289
Total	120			—	—	9.93733

GOODNESS-OF-FIT TEST

- A Pearson **goodness-of-fit test,** in general, refers to measuring how well do the observed data correspond to the fitted (assumed) model.
- X_1,X_2,X_3,…X_n is a random sample from unknow distribution F(X), $F_0(X)$ is a known distribution.

H0: $F(X)= F_0(X)$; H1: $F(X) \neq F_0(X)$.

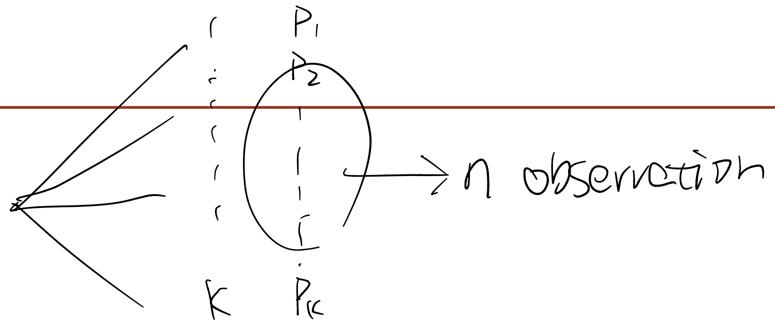
Or we can test whether F(X) follow a certain type of distribution(i.e., normal distribution)

H0: $F(X)= F_0(X;\theta)$; H1: $F(X) \neq F_0(X ;\theta)$.

Chi-square goodness-of-fit test

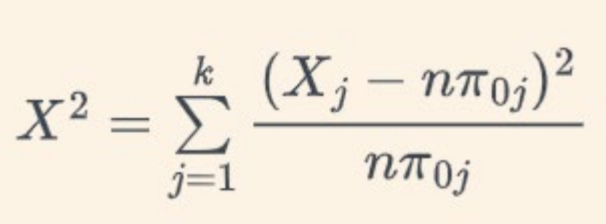
- A **Chi-square goodness-of-fit test,** allows us the test if the sample data from a categorical variable fits the pattern of expected probabilities for the variable.

Suppose a categorical variable has k possible outcomes (categories) with probabilities p1,p2,...,pk. Suppose n independent observations are taken from this categorical variable.



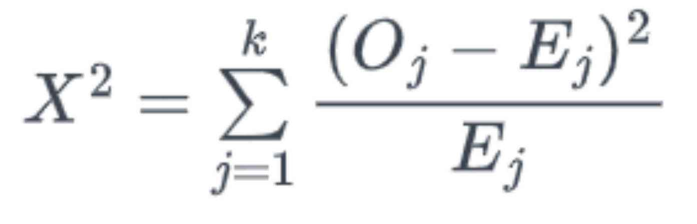
H0: $p_1 = \pi_{01}$, $p_2 = \pi_{02}$, $p_3 = \pi_{03}$, ………. $p_k = \pi_{0k}$

Pearson Goodness-of-fit Test Statistic



$$X^2 = \sum_{j=1}^{k} \frac{(X_j - n\pi_{0j})^2}{n\pi_{0j}}$$

An easy way to remember it is



$$X^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

(Equation 9.1)

where $O_j = X_j$ is the **observed count** in cell j, and $E_j = E(X_j) = n\pi_{0j}$ is the **expected count** in cell j under the assumption that null hypothesis is true.

X2 measure how closely the model, have an approximate chi-square distribution with k−1 degrees of freedom when H0 is true. $X^2 \sim x^2_{k-1}$

If the sample proportions $\hat{\pi}_j$ are close to the model's π_{0j}, then $O_j \rightarrow E_j$, and $X^2 \rightarrow 0$.

Proof for df=1

Suppose a binomial distribution has 2 outcomes (1,2), of which the probabilities are p1 and p2, respectively. $P_2=1-p_1$

Suppose we draw a sample from this binomial distribution, the sample size is n, the number of outcome 1 and 2 is v1 and v2, respectively.

According to **Pearson** Goodness-of-fit Test Statistic

$$X^2=\frac{(v_1-np_1)^2}{np_1}+\frac{(v_2-np_2)^2}{np_2}=\frac{(v_1-np_1)^2}{np_1}+\frac{[(n-v_1)-(n-np_1)]^2}{n(1-p_1)}$$

$$=\frac{(v_1-np_1)^2}{np_1(1-p_1)}=\left[\frac{v_1-np_1}{\sqrt{np_1(1-p_1)}}\right]^2$$

De Moivre-Laplace

If n→+∞, it tends to follow N(0,1)

Extension， H0: pi=πi(θ), i=1,,,k, where θ is unknown parameter

Pearson Goodness-of-fit Test Statistic

Use Maximum Likelihood Estimation to estimate θ

Pearson-Fisher theorem

$$X^2=\sum_{i=1}^{k}\frac{(X_j-n\hat{\pi}_{0j})}{n\hat{\pi}_{0j}},\ \text{and X}^2\sim x^2_{\boxed{k-r-1}}\ \text{if n}\rightarrow\infty$$

(Equation 9.2)

where k= the number of groups and r = the number of parameters estimated from the data to compute the expected frequencies

For example 1: x^2=9.93733;

df=k-1=9-1=8, and $x_{0.05,8}^2$=15.51;

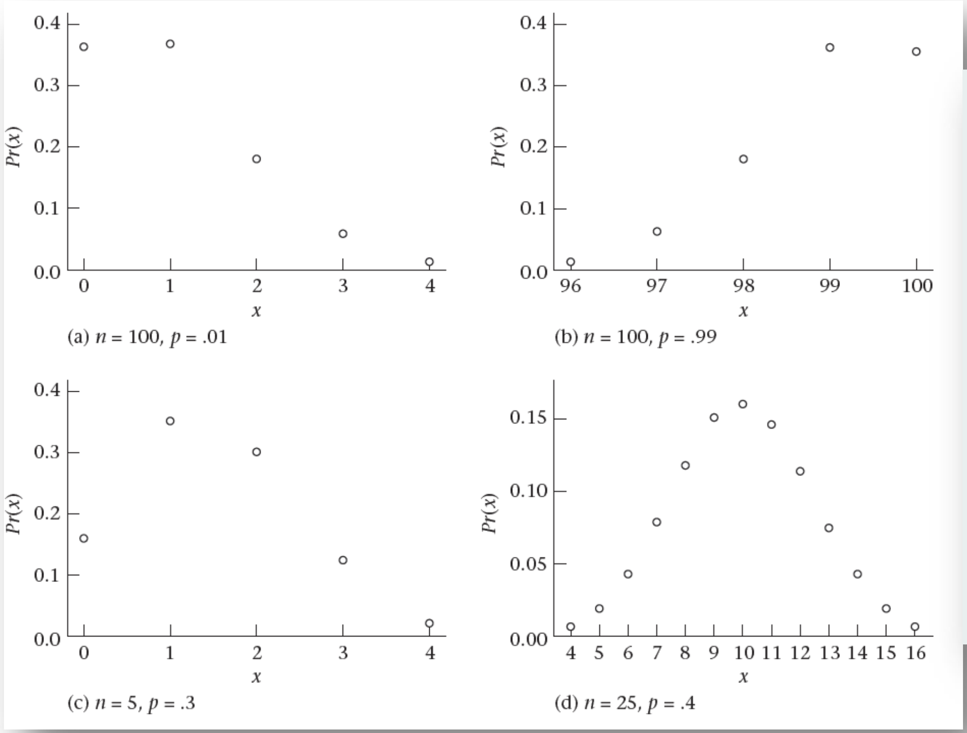
Then accept H_0

2. Two sample test for binomial proportions

Recall: normal-theory method

Normal Approximation to the Binomial Distribution (npq>5)

$$X \sim B(\mathrm{n,p})$$

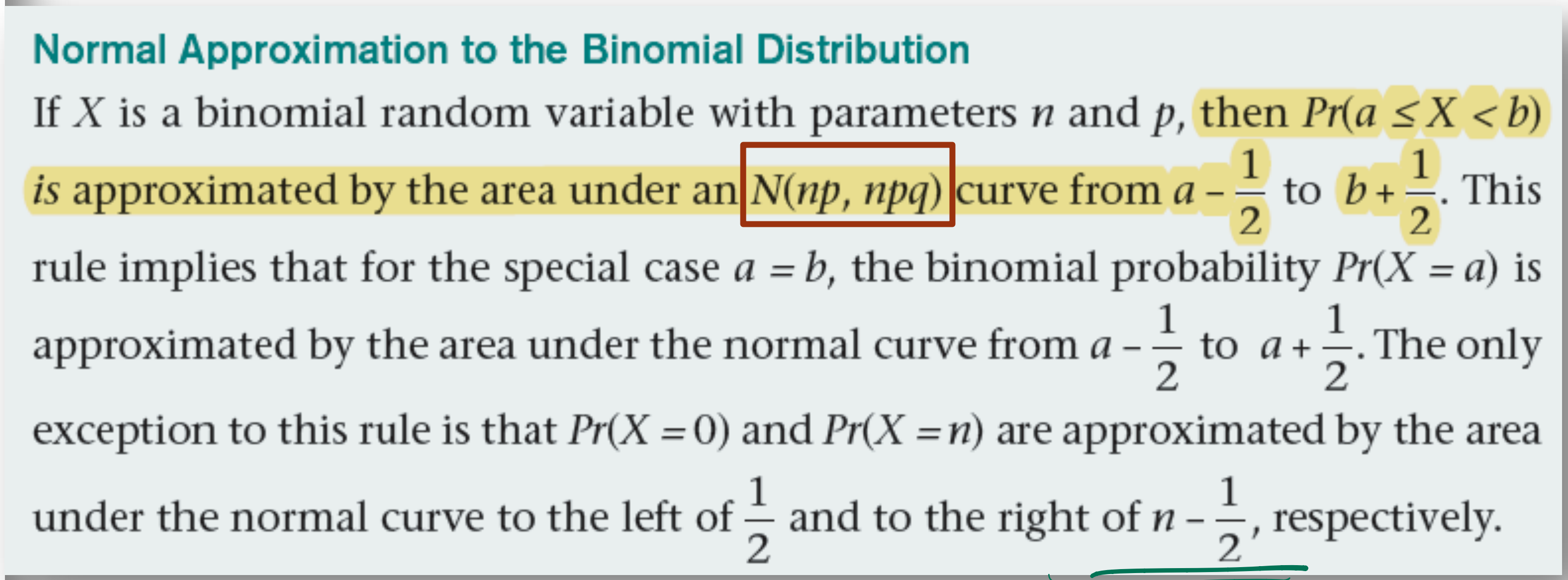


(a) $n = 100$, $p = .01$

(b) $n = 100$, $p = .99$

(c) $n = 5$, $p = .3$

(d) $n = 25$, $p = .4$



Normal Approximation to the Binomial Distribution

If X is a binomial random variable with parameters n and p, then $Pr(a \le X < b)$ *is* approximated by the area under an $\boxed{N(np,\, npq)}$ curve from $a - \frac{1}{2}$ to $b + \frac{1}{2}$. This rule implies that for the special case $a = b$, the binomial probability $Pr(X = a)$ is approximated by the area under the normal curve from $a - \frac{1}{2}$ to $a + \frac{1}{2}$. The only exception to this rule is that $Pr(X = 0)$ and $Pr(X = n)$ are approximated by the area under the normal curve to the left of $\frac{1}{2}$ and to the right of $\underline{n - \frac{1}{2}}$, respectively.

- **Example 1: One doctor wants to study whether or not there is a difference for the treatment effectiveness of chronic pharyngitis between drug A group and drug B group. He randomly divided 80 homogenous patients suffering from chronic pharyngitis into drug A and B groups, and then observed the treatment effectiveness. The data is shown in table 1.**

Table 1 Data of effect treated by drug A and drug B

Drug group	effective		total	Rate of effectiveness(%)	
	yes	no			
A	41 $X1$	4	45 (n1)	91.1	$B(n1,\pi1)$
B	24 $X2$	11	35 (n2)	68.8	$B(n2,\pi2)$
Total	65	15	80	81.3	

$H0: \pi1=\pi2$

$H1: \pi1\neq\pi2$

Normal-Theory Method （samples large enough, npq>5 ）

Under H0:π1=π2=p

$$p1 \sim N(\text{p,pq/n1});\ \text{p2} \sim N(\text{p,pq/n2});\ \text{samples are indepedent}$$

$$p1 - p2 \sim N(0, \frac{pq}{n_1} + \frac{pq}{n_2})$$

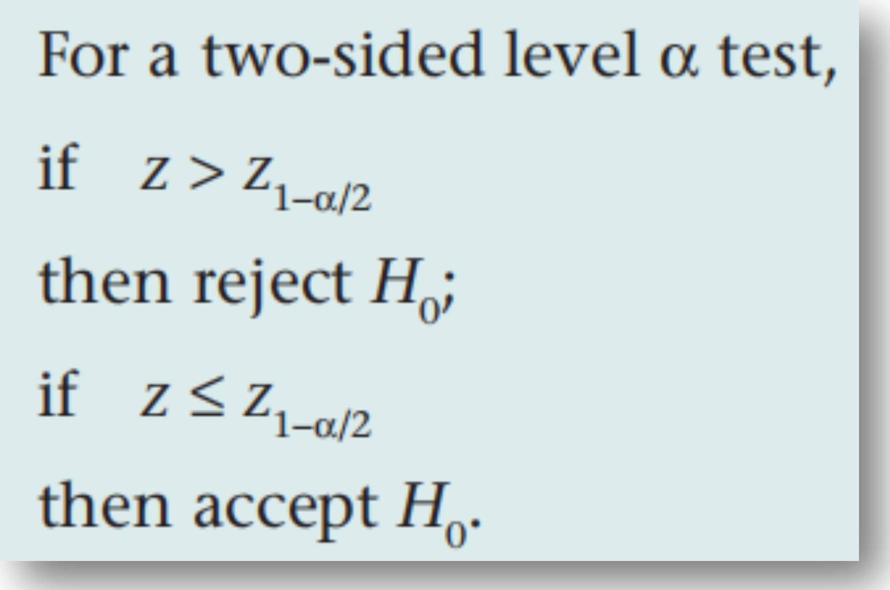
then under H0,

$$\text{z=}\frac{p1 - p2}{\sqrt{pq(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0,1)$$

The best estimator for *p* is based on a weighted average of the sample proportions p1 and p2.

$$\hat{p} = \frac{n1p1 + n2p2}{n1 + n2} = \frac{X1 + X2}{n1 + n2}$$

$$z = \frac{|\,p1 - p2\,| - \left(\dfrac{1}{2n_1} + \dfrac{1}{2n_2}\right)}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}, \text{ where } \hat{\text{p}} = \frac{X_1 + X_2}{n_1 + n_2}$$



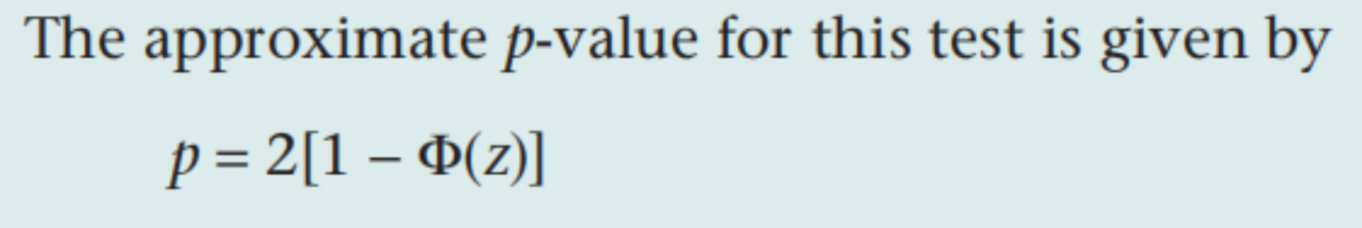
For a two-sided level α test,

if $\quad z > z_{1-\alpha/2}$

then reject H_0;

if $\quad z \le z_{1-\alpha/2}$

then accept H_0.



The approximate *p*-value for this test is given by

$$p = 2[1 - \Phi(z)]$$

Solution1 for example 1

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{15}{80} = 81.3\%,\ \hat{q} = 1 - \hat{p} = 18.7\%$$

$$note, n_1 \hat{p}\hat{q} = 45 \times 0.813 \times 0.187 = 6.84 > 5$$

$$n_2 \hat{p}\hat{q} = 35 \times 0.813 \times 0.187 = 5.32 > 5$$

$$z = \frac{|p1 - p2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{pq(\frac{1}{n_1} + \frac{2}{n_2})}}$$

$$= \frac{|0.911 - 0.688| - (\frac{1}{2 \times 45} + \frac{1}{2 \times 35})}{\sqrt{0.813 \times 0.187 \times (\frac{1}{35} + \frac{1}{45})}} = 25.47994$$

$Z > Z_{a/2} = 1.96$, then reject H0

Contingency-Table Method

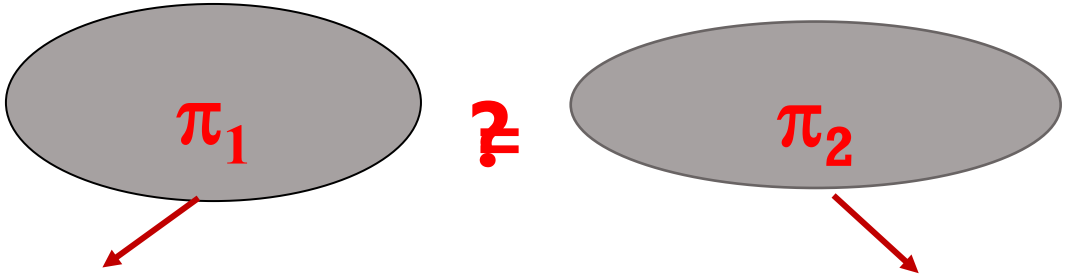
Table 1.1 Data of effect treated by drug A and drug B

Drug group	effective		total	Rate of effectiveness(%)
	yes	no		
A	41	4	45 (row margins)	91.1
B	24	11	35 (row margins)	68.8
Total	65 (column margins)	15 (column margins)	80 (grand total)	81.3

is called a 2 × 2 contingenc table or four-fold table because it has two categories for effective status and two categories for drug groups.

Step 1 .Hypothesis

- $\pi 1$: the probability or population proportion of effectiveness treated by drug A
- $\pi 2$: the probability or population proportion of effectiveness treated by drug B



Sample1:

p_1=(41/45)×100%=91.1%

Sample2:

p_2=(24/35)×100%=68.6%

homogeneity of binomial proportions

H_0: $\pi_1 = \pi_2$, the difference between p_1 and p_2 is caused by sampling error

H_1: $\pi_1 \neq \pi_2$, the difference between p_1 and p_2 is caused by sampling error+ treatment effect

Table 2 Expected table for the data in example 1

Drug group	effective		total	Rate(%)
	yes	no		
A	41(36.6)	4(8.4)	45	91.1
B	24(28.4)	11(6.6)	35	68.8
Total	65	15	80	81.3

➢Assume H_0:π_1= π_2 =π is true, the best estimate of common rate π is p_c, which is 81.3%.

E_{11}=n_1p_1=45×81.3%=45×(65/80)=36.6

E_{12}=$n_1(1-p_1)$=45×(100%-81.3%)=45×(15/80)=8.4

E_{21}= n_2p_2=35×81.3%=35×(65/80)=28.4

E_{22}=$n_2(1-p_2)$=35×(100%-81.3%)=35×(15/80)=6.6

$$E_{rc} = \frac{n_r n_c}{N}$$ **(Equation 9.3)**

n_r correpsonding row margin

n_c correpsonding column margin

How to compare the difference between actual number (or observed number) and theoretical number(or expected number)?

$$\sum \frac{(O-E)^2}{E} \sim \chi^2_{\alpha/2,\nu}$$ **(Equation 9.1)**

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

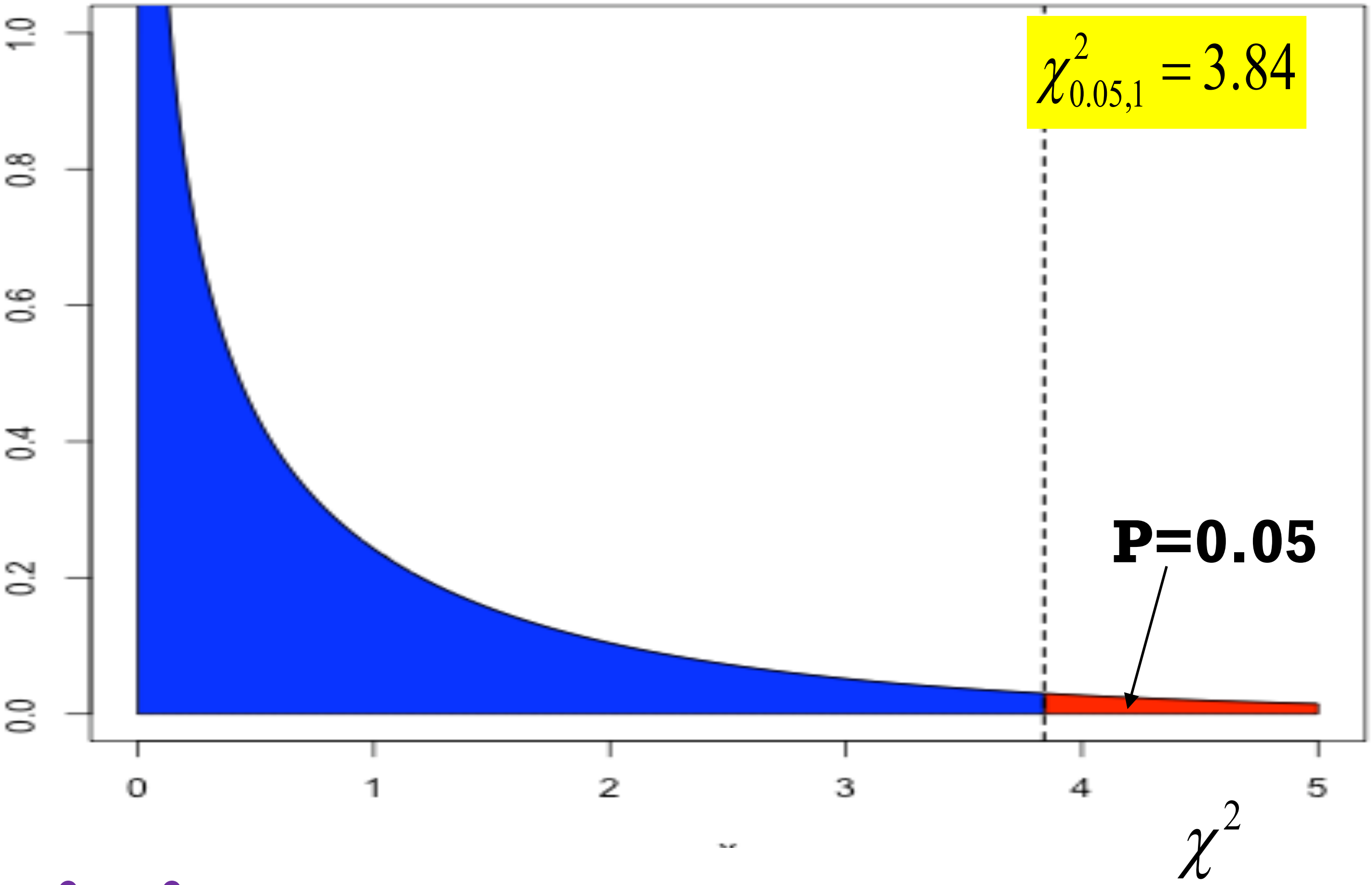
(Only for 2×2 case)

***df*=ν=(r-1)(c-1), or df=k-1-r,where k=rc** ? **(Equation 9.2)**

- If this sum is small, H_0 is not rejected, because small value of this sum means good agreement between the actual cells and the theoretical cells.

- **If this sum is large, H_0 is rejected.**

CHI-SQUARE (χ^2) DISTRIBUTION



- If $\chi^2 \geq \chi^2_{0.05,1}$, $P \leq 0.05$, then reject H_0
- If $\chi^2 < \chi^2_{0.05,1}$, $P > 0.05$, then can not reject H_0

Step 2 Calculate the χ^2 value

$$\chi^2=\sum\frac{(A-T)^2}{T}=\frac{(41-36.6)^2}{36.6}+\frac{(4-8.4)^2}{8.4}+\frac{(24-28.4)^2}{28.4}+\frac{(11-6.6)^2}{6.6}=6.565$$

$$\nu=(r\text{-}1)(c\text{-}1)=(2\text{-}1)(2\text{-}1)=1$$

Step 3 Estimate the P value, and draw a conclusion

$$\chi^2_{0.05,1}=3.84,\ 6.565>3.84,\ P<0.05$$

- Because P<0.05, according to the significance level α=0.05, we have evidence to refuse the H_0, and accept the H_1.

- We can draw a conclusion that the population rate of treatment effectiveness for drug A does not equal to the population rate for drug B.

BASIC PRINCIPLES OF CHI-SQUARE TEST

✵ **Chi-square value is the difference between O and E values. It represents the agreement degree between the O and E.**

✵ **If H_0 is true, O is close to E, and the chi-square value will not be very big in most conditions, the chance of small chi-square value is big, and the chance of the large chi-square value appeared to be small.**

✵ **If we have a big chi-square value, the agreement between the actual and the theoretical cells is poor, so we reject the null hypothesis.**

Short computational form for chi-square test for 2×2 contingency table

Table 3 General contingency table

group	outcome		total
	+	-	
A	***a***	***b***	*a*+*b*
B	***c***	***d***	*c*+*d*
total	*a*+*c*	*b*+*d*	*n*=*a*+*b*+*c*+*d*

$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$ **(Equation 9.4)**

- We use the short computational form to calculate chi-square in example 1:

Drug Group	effective		total
	yes	no	
A	41	4	45
B	24	11	35
Total	65	15	80

$$\chi^2 = \frac{(41\times 11 - 4\times 24)^2 \times 80}{45\times 35\times 65\times 15} = 6.565$$

Yates-corrected Chi-square test for a 2×2 contingency table

- **If any of four cells is :1≤T<5 and n≥40 , we should compute the Yates-corrected chi-square**

$$\chi^2 = \sum \frac{(|O - E|\text{-}0.5)^2}{E}$$

(Equation 9.5)

$$\chi^2 = \frac{(\,|\,ad - bc\,|\,\text{-}\frac{n}{2})^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

(Equation 9.6)

- **Example 2 Suppose we want to investigate the relationship between high salt intake and death from cardiovascular disease(CVD). A retrospective study is done among men ages 50-54 in a specific county who died over a 1-month period. The investigators try to include approximately an equal number of men who died from CVD (the cases) and men who died from other causes(the controls).**

Of 35 people who died from CVD, 5 were on a high-salt diet, whereas of 25 people who died from other causes, 2 were on such a diet.

Table 4 data concerning the possible association between cause of death and high salt intake

Cause of death	Type of diet		Total
	High salt	Low salt	
Non-CVD	2(2.92)	23(22.08)	25
CVD	5(4.08)	30(30.92)	35
Total	7	53	60

- ➢ $T_{11}=25\times7/60=2.92$
- ➢ $T_{21}=35\times7/60=4.08$

We find T_{11} and T_{21} are between 1 and 5, so, the Yates-corrected Chi-square test should be used to compare difference between two groups.

(1) Hypothesis:

H_0: $\pi_1=\pi_{2,}$ the population proportion of high-salt intake among people who died from CVD equals to proportion high-salt intake among people who died from other diseases

H_1: $\pi_1\neq\pi_{2,}$ two population proportions don't equal to each other

$\alpha=0.05$

(2) Calculate the corrected chi-square

$$\chi^2=\frac{(|2\times 30-23\times 5|-60/2)^2\times 60}{25\times 35\times 7\times 53}=0.116$$ $\nu=(2-1)(2-1)=1$

(3) Conclusion

$\chi^2=0.116<3.84$, $P>0.05$, we can't reject H_0.

We Have Several Requirements For The Fourfold Table

- **(1) If T≥5 and *n* ≥40**

$$\chi^2 = \sum \frac{(A-T)^2}{T} \qquad \textbf{or} \qquad \chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

- **(2) If any 1≤T<5 and *n* ≥40**

$$\chi^2 = \sum \frac{(|A-T|\boxminus 0.5)^2}{T} \qquad \textbf{or} \qquad \chi^2 = \frac{(|ad-bc|\text{-}\frac{n}{2})^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

- **(3) If any T<1 or *n*<40, we should use Fisher's exact test to compute the P value**

3. FISHER'S EXACT TEST

FISHER’S EXACT TEST

- When any of theoretical value is less than 1(T<1) or *n*<40, we need to use fisher’s exact method to estimate the *P* value and make the statistical inference.
- We can use R, SPSS or SAS to get the *P* value.

Table 5 Comparison of mouse death rate for poison A and B

Poison group	Death outcome		total	Rate of death(%)
	Yes	No		
A	1	9	10	10.0
B	5	5	10	50.0
Total	6	14	20	30.0

4. Two Sample Test For Binomial Proportions For Matched Data

Two-sample Test for Matched-pair Data (McNemar’s Test) (McNemar chi-square Test)

♦Matched pair design patterns :

❄Two subjects in each pair receive different treatments(different subjects match), it needs to match the pair for the subjects by some conditions such as same age, sex, weight or others before randomizing

❄One subject receives two different treatments (within-subjects design)

ASSUMPTIONS

- You must have one nominal variable with two categories (i.e. dichotomous variables) and one independent variable with two connected groups.
- The two groups in your the dependent variable must be mutually exclusive. In other words, participants cannot appear in more than one group.
- Your sample must be a random sample.

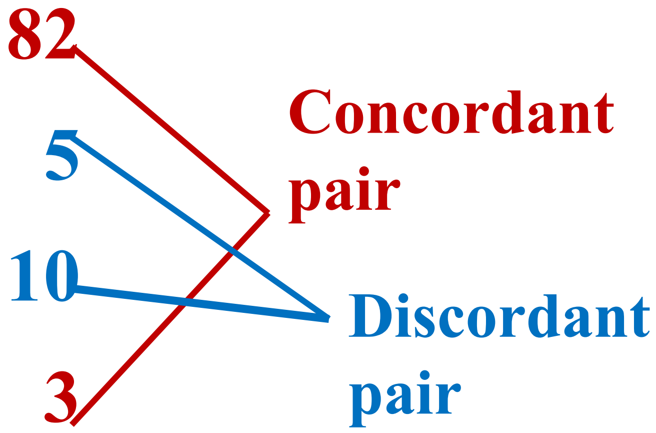
➢Matched pair design

Example 4 The standard screening test for Down's syndrome is based on a combination of maternal age and the level of serum cases can be identified, while 5% of the normal are detected as positive.

A new test is proposed that may be better or worse than the standard test. To assess their relative efficacy, both tests are used on the same subjects and compared with the true diagnosis. Let + equals to correct assessment, - equals to incorrect assessment. The results are given in table 6.

Table 6 Comparison of two screening tests for Down’s syndrome

Standard test	New test	N
+	+	82
+	-	5
-	+	10
-	-	3


Concordant pair

Discordant pair

- ➢There’re four different combinations

85(82+3) pairs have the same outcomes, and 15(5+10) pairs have different outcomes.

Table 7 Comparison of two screening tests for Down's syndrome

Standard test	New test		total
	+	-	
+	82(a)	5(b)	87
-	10(c)	3(d)	13
total	92	8	100

- The same outcomes don't contribute to comparing the difference between two methods. If the new test and standard test are equally effective, an equal number of new test and standard test discordant pairs would be expected.

(McNemar chi-square Test)

1.Hypothesis

Let the capital letter "B" denote population parameter of b, and C denote population parameter of c.

H0: B=C, H1: B≠C, α=0.05

2. Compute test statistic

✌ **If H_0 is true, the theoretical value=(b+c)/2**

(1)When b+c≥40

$$\frac{(113-15)^2}{128}$$

$$\chi^2 = \frac{\left[b-\frac{b+c}{2}\right]^2}{\frac{b+c}{2}} + \frac{\left[c-\frac{b+c}{2}\right]^2}{\frac{b+c}{2}} = \frac{(b-c)^2}{b+c}$$

(Equation 9.7)

(2)When b+c<40

$$\chi^2=\frac{\left[|b-\frac{b+c}{2}|-\frac{1}{2}\right]^2}{\frac{b+c}{2}}+\frac{\left[|c-\frac{b+c}{2}|-\frac{1}{2}\right]^2}{\frac{b+c}{2}}=\frac{(|b-c|-1)^2}{b+c}$$

(Equation 13.8)

✐For this example, b+c<40

$$\chi^2=\frac{(|b-c|-1)^2}{b+c}=\frac{(|5-10|-1)^2}{5+10}=1.07$$

ν=(2-1)(2-1)=1

3. Conclusion

χ^2=1.07<3.84, P>0.05, H_0 can not be rejected.

5. R×C CONTINGENCY TABLE

- An R×C contingency table is a table with R rows and C columns. It displays the relationship between two variables, where the variable in the rows has R categories and the variable in the columns has C categories. The number of R or C is more than 2.

Example 5 Obesity is an important risk factor for many diseases. However, in studying the effects of obesity, it is important to be aware of other risk factors that may be potentially related to obesity. One commonly used measure of obesity is body mass index (BMI) (kg/m^2),which is often categorized as follows：

normal=BMI<25, overweight=BMI 25.0-29.9， and obesity=BMI≥30.0.

The data in table 8 were presented in a study relating education to BMI category. What test can be used to compare the percentages of individuals with at least a high-school education in the three BMI groups?

Table 8 Relationship between BMI category and education level (n=261)

BMI category	≥High-school education	<High-school education	Total	%≥High-school education
Normal	70	7	77	90.9
Overweight	105	15	120	87.5
Obese	53	11	64	77.1
total	228	33	261	87.36

- **Generalizing our experience from the 2×2 situation, the expected number for every cell can be formed in the same way as 2×2 table**

1.Hypothesis

H_0: $\pi_1=\pi_2=\pi_{3,}$ Population percentages of people with at least high school education are same among three BMI groups

H_1: At least two groups' π are different

$\alpha=0.05$

2. Compute test statistic

- $E_{11}=(77\times228)/261=67.26$
- $E_{12}=(77\times33)/261=9.74$
- $E_{21}=(120\times228)/261=104.83$
- $E_{22}=(120\times33)/261=15.17$
- $E_{31}=(64\times228)/261=55.91$
- $E_{32}=(64\times33)/261=8.09$

➢Basic formula

$$\chi^2 = \frac{(70-67.26)^2}{67.26} + \frac{(7-9.74)^2}{9.74} + \frac{(105-104.83)^2}{104.83} + \frac{(15-15.17)^2}{15.17} + \frac{(53-55.91)^2}{55.91} + \frac{(11-8.09)^2}{8.09}$$

$$= 2.08$$

➢the special formula

$$\chi^2 = n[\sum \frac{O^2}{n_r n_c} - 1] = 261[\frac{70^2}{77\times 228} + \frac{7^2}{77\times 33} + \frac{105^2}{120\times 228}$$

$$= \frac{15^2}{120\times 33} + \frac{53^2}{64\times 228} + \frac{11^2}{64\times 33} \text{-1}] = 2.08$$

➢ν=(3-1)(2-1)=2

3. Conclusion

χ^2=2.08<$\chi^2_{0.05,2}$=5.99, P>0.05, H_0 can not be rejected.

➢**Example 6** **One doctor wants to know whether there is difference in the proportions distribution of blood type of acute lymphatic leukemia(ALL) patients between children and adults. He collected some data in table 9.**

Table 9 Blood type distribution between children and adults

group	**A(%)**	**B(%)**	**O(%)**	**AB(%)**	**Total(%)**
children	30(26.8)	38(33.9)	32(28.6)	12(10.7)	112(100.0)
adults	19(24.7)	30(39.0)	19(24.7)	9(11.7)	77(100.0)
total	49	68	51	21	189

- **H_0: the population proportions of the blood type are equal for children and adults ALL patients**
- **H1: the population proportions of the blood type are not equal for children and adults ALL patients**

α=0.05

$$\chi 2 = n(\sum \frac{O^2}{n_r \cdot n_c} - 1) = 189 \times (\frac{30^2}{112 \times 49} + \frac{38^2}{112 \times 68} + \frac{32^2}{112 \times 51} + \frac{12^2}{112 \times 21}$$

$$+ \frac{19^2}{77 \times 49} + \frac{30^2}{77 \times 68} + \frac{19^2}{77 \times 51} + \frac{9^2}{77 \times 71} - 1) = 0.695$$

$$\nu = (r-1)(c-1) = (4-1)(2-1) = 3$$

Conclusion $\chi^2_{0.05,3} = 7.81, 0.695 < 7.81, P > 0.05$

We can’t reject H_0, there is no significant difference in proportions of blood between children and adults ALL patients, which means the blood type does not relate to age of ALL patients.

★Cautions in Chi-square test for R×C table

- Use this test only if both of the following two conditions are satisfied:

(1) No more than 1/5 of the cells have expected values <5.

(2)No cell has a theoretical or an expected value<1.

➤If these two conditions are not satisfied, there is no continuity correction, we can use Fisher's exact test to compare differences among groups.

6. Correlation analysis of categorical variable data

DEGREE OF ASSOCIATION

Example 4 **The standard screening test for Down’s syndrome is based on a combination of maternal age and the level of serum cases can be identified, while 5% of the normal are detected as positive.**

Table 7 Comparison of two screening tests for Down’s syndrome

Standard test	New test		total
	+	-	
+	82(a)	5(b)	87
-	10(c)	3(d)	13
total	92	8	100

- Phi coefficient

$$r_n = \frac{(ad-bc)}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$$

$$= \frac{82\times3-5\times10}{\sqrt{87\times13\times92\times8}} = 0.116181$$

□Hypothesis test, H0: r_n=0

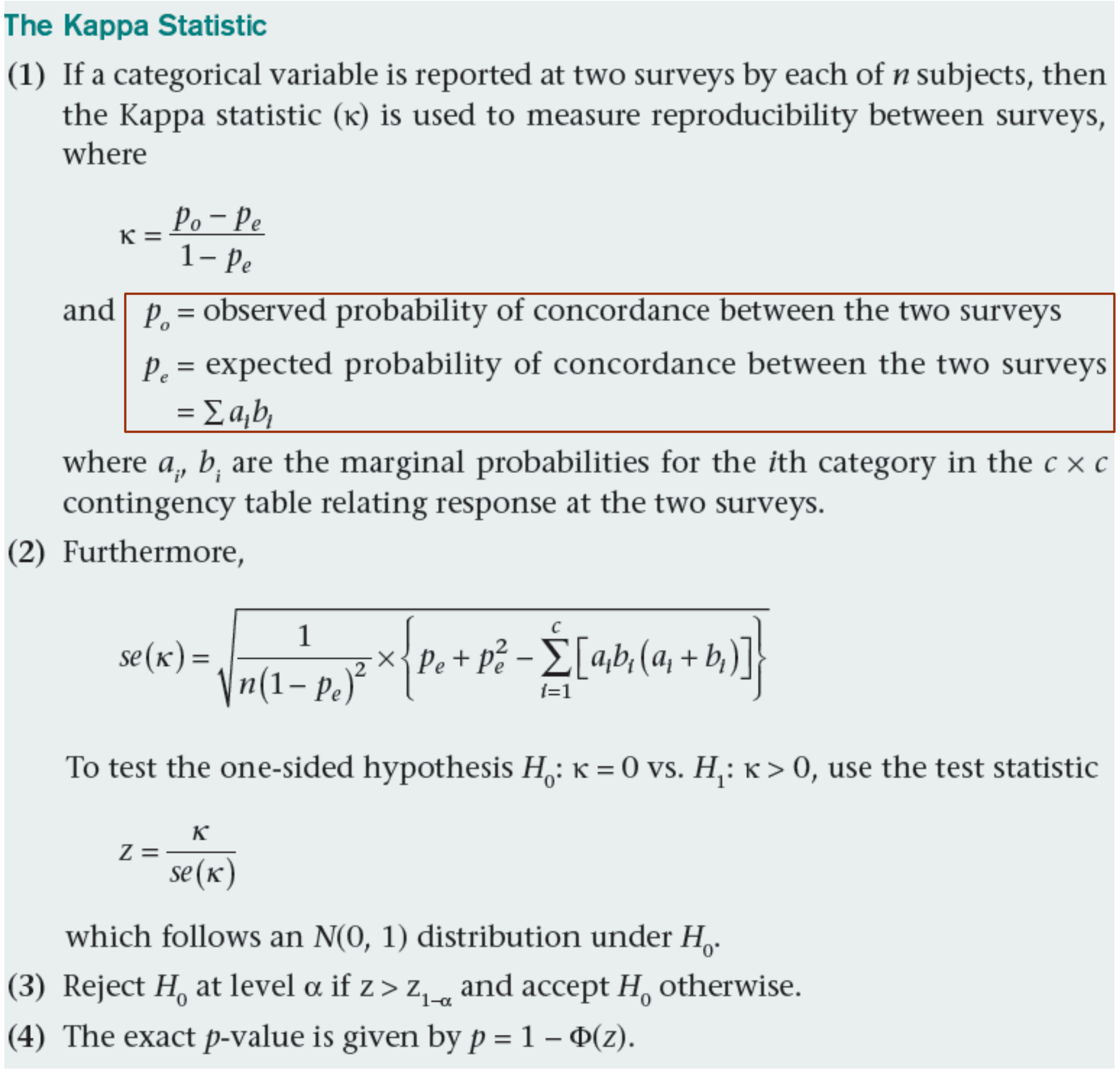
$$\chi^2 = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

(Equation 9.4)

$$= \frac{(82\times3-5\times10)^2\times100}{87\times13\times92\times8} = 4.615$$

χ^2=4.615>3.84, therefore the results of two screening tests is connected

KAPPA STATISTIC



The Kappa Statistic

(1) If a categorical variable is reported at two surveys by each of n subjects, then the Kappa statistic (κ) is used to measure reproducibility between surveys, where

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

and p_o = observed probability of concordance between the two surveys

p_e = expected probability of concordance between the two surveys $= \sum a_i b_i$

where a_i, b_i are the marginal probabilities for the ith category in the $c \times c$ contingency table relating response at the two surveys.

(2) Furthermore,

$$se(\kappa) = \sqrt{\frac{1}{n(1-p_e)^2} \times \left\{ p_e + p_e^2 - \sum_{i=1}^{c} \left[a_i b_i (a_i + b_i) \right] \right\}}$$

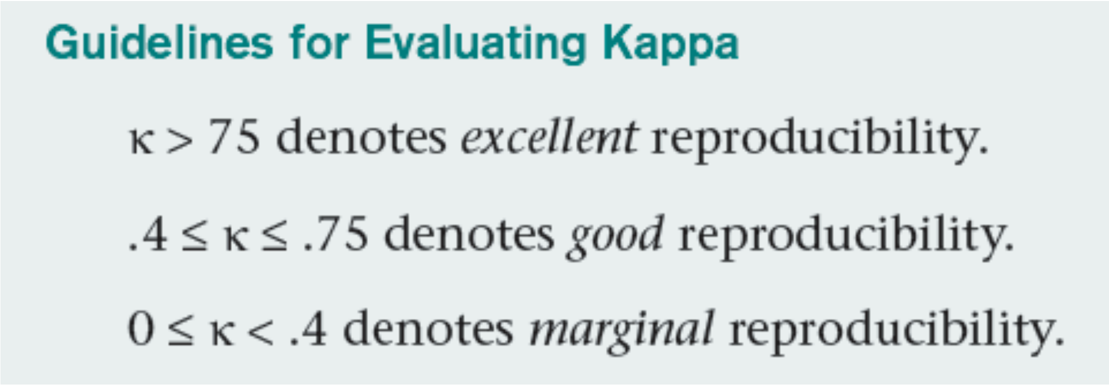
To test the one-sided hypothesis H_0: $\kappa = 0$ vs. H_1: $\kappa > 0$, use the test statistic

$$z = \frac{\kappa}{se(\kappa)}$$

which follows an $N(0, 1)$ distribution under H_0.

(3) Reject H_0 at level α if $z > z_{1-\alpha}$ and accept H_0 otherwise.

(4) The exact p-value is given by $p = 1 - \Phi(z)$.



Guidelines for Evaluating Kappa

$\kappa > 75$ denotes *excellent* reproducibility.

$.4 \le \kappa \le .75$ denotes *good* reproducibility.

$0 \le \kappa < .4$ denotes *marginal* reproducibility.

$$p_a = \sum_{i=1}^{k} O_{ii} / N = \frac{82+3}{100} = 0.85$$

$$p_e = \sum_{i=1}^{k} n_{i+} n_{+i} / N^2$$

$$= \frac{92 \times 87 + 8 \times 13}{100^2} = 0.7238$$

$$K = \frac{p_a - p_e}{1 - p_e} = 0.4569$$

Summary for this chapter

1. This chapter discussed the most widely used techniques for analyzing qualitative or categorical data.

2. For large sample, we can use chi-square test to compare binomial proportions or rates from two independent samples. For the small-sample case, Fisher’s exact test is used to compare binomial proportions in two independent samples.

Summary

3. To compare binomial proportions in the paired samples, McNemar’s test for proportions should be used.

4. For R×C contingency table, a chi-square test is a direct generalization of 2 ×2 contingency table test.