

Hypothesis Testing: Categorical Data

Outline


1. Two sample test for Binomial proportions
 - normal theory approach
 - contingency table approach
2. Fisher's exact test
3. McNemar's test
4. Sample size and power estimation
5. $R \times C$ contingency tables
6. Goodness of Fit
7. The Kappa statistic

Two-sample test

For Binomial proportions

Example:

Breast Cancer

- Suppose we are interested in the **association** between the incidence of breast cancer and the age at first childbirth.
- Breast-cancer cases were identified among women in selected hospitals in the United States, Greece, Yugoslavia, Brazil, Taiwan, and Japan.  **case**
- **Controls** were chosen from women of comparable age who were in the hospital at the same time as the cases but who did **not** have breast cancer.
- These women are divided into two categories:
 - women whose age at first birth was ≤ 29 years
 - women whose age at first birth was ≥ 30 years

Outcomes

Age at first birth			
Status	≥ 30	≤ 29	Total
Case	<u>683</u>	<u>2537</u>	<u>3220</u>
Control	<u>1498</u>	<u>8747</u>	<u>10245</u>
Total	2181	11284	13465

Comparing two proportions

- Similar to the two-sample t-test
- Consider testing $H_0: p_1 = p_2$ vs $H_1: p_1 \neq p_2$
- The test statistic is

$$TS = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$ is the estimate of the common proportion under the null hypothesis.
- This statistic is normally distributed for large n_1 and n_2 .

Back to Breast Cancer

- $\hat{p}_1 = \frac{683}{3220} = 0.212, \hat{p}_2 = \frac{1498}{10245} = 0.146$
- $\hat{p} = \frac{(683+1498)}{(3220+10245)} = \frac{2181}{13465} = 0.162$
- Note that
 - $n_1\hat{p}(1 - \hat{p}) = 3220 \times 0.162 \times (1 - 0.162) = 437 > 5$
 - $n_2\hat{p}(1 - \hat{p}) = 10245 \times 0.162 \times (1 - 0.162) = 1391 > 5$We can use the normal approximation to binomial distribution.
- The test statistic is given by
$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p} \times (1 - \hat{p}) \times (1/3220 + 1/10245)}} = 8.85$$
- The p-value = $2 \times (1 - \Phi(8.85)) < .001 \Rightarrow$ highly significant

Contingency table approach

contingency-table

- Setup: 2×2 contingency table

	Exposed	Unexposed	Total
Disease	O_{11}	O_{12}	n_1
No Disease	O_{21}	O_{22}	n_2
Total	m_1	m_2	n

- O_{ij} = observed cell counts in the (i, j) cell
- n_i = row marginal total (row margin)
- m_j = column marginal total (column margin)
- n = grand total
- E_{ij} = expected cell counts under $H_0 = \frac{n_i m_j}{n}$

Hypothesis of homogeneity

- Is the chance of Disease different for Exposed as compared to the Unexposed?
- Is the proportion of Exposed different in those who get the Disease as compared to those who do not?
- Test if two proportions are equal: one set of margins is fixed
- Chi-square test of homogeneity

Hypothesis of independence

- Two binary random variables
- $X \rightarrow \{\text{Exposed, Unexposed}\} = \{0, 1\}$
- $Y \rightarrow \{\text{Disease, No Disease}\} = \{0, 1\}$
- Population/Sample $(X_i, Y_j) \Rightarrow$ Summarize in a two-by-two table
- Test if X is independent of Y: both sets of margins (tea-tasting example) or the total sample size (cross-sectional study) is assumed to be fixed
- Chi-square test of independence/association

Tea-tasting example

- 8 cups of tea: 4 prepared by first adding milk and 4 prepared by first adding the tea
- A lady's guess, both margins are fixed

Truth (milk/tea first)			
Guess (milk/tea first)	2(4)	2(0)	4
	2(0)	2(4)	4
	4	4	8

Cross-sectional example

- Maternal age versus birth-weight
- The total sample size is fixed

Mat. Age	Birth-weight		Total
	< 2500 g	≥ 2500g	
< 20y	20	80	100
≥ 20y	30	270	300
Total	50	350	400

- H_0 : MA is independent of BW
- H_1 : MA is not independent of BW

Chi-squared testing

- Under H_0 (if $E_{ij} > 5$ for all i, j):

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Follows χ^2 distribution with **1 df**.
- Reject if the statistic is too large.
- exactly the square of the two-proportion test statistic
- Yates' continuity correction:

$$\chi^2 = \sum_{i,j} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

Back to Breast Cancer

Observed table

Age at first birth			
Status	≥ 30	≤ 29	Total
Case	683	2537	3220
Control	1498	8747	10245
Total	2181	11284	13465

Expected table

Age at first birth			
Status	≥ 30	≤ 29	Total
Case	521.6	2698.4	3220
Control	1659.2	8585.6	10245
Total	2181	11284	13465

Contd.

- Note that the $E_{ij} > 5$ for all i, j .
- The χ^2 test statistic is

$$\chi^2 = \frac{(683-521.6)^2}{521.6} + \frac{(2537-2698.4)^2}{2698.4} + \frac{(1498-1659.4)^2}{1659.4} + \frac{(8747-8585.6)^2}{8585.6} = 78.37$$

- The p-value $< .001 \Rightarrow$ highly significant

```
> pchisq(78.33, 1, lower.tail = FALSE)
[1] 8.543992e-19
```


R codes

```
> dat <- matrix(c(683,2537,1498,8747), 2)
> chisq.test(dat, correct = FALSE)
```

Pearson's Chi-squared test

```
data: dat
X-squared = 78.3698, df = 1, p-value < 2.2e-16
```

- with Yates' continuity correction:

```
> chisq.test(dat)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: dat
X-squared = 77.8851, df = 1, p-value < 2.2e-16
```

Alternative form

	Exposed	Unexposed	Total
Disease	n_{11}	n_{12}	n_{1+}
No Disease	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- Alternative computational form for the χ^2 statistic

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{+1}n_{+2}n_{1+}n_{2+}}$$

- This statistic does not change if you transpose the rows and the columns of the table

What's more

- The χ^2 statistic can be used
 - the rows are fixed (binomial)
 - the columns are fixed (binomial)
 - the total sample size is fixed (multinomial)
 - none are fixed (Poisson)
- For a given set of data, any of these assumptions results in the same value for the statistic
- Degree of freedom = (number of cells) - 1 - (number of parameters estimated)
- The Chi-squared test is equivalent to the comparison of proportions

Fisher's Exact Test

Motivation example:

Cardiovascular Disease

- Suppose we want to investigate the relationship between high salt intake and death from cardiovascular disease (CVD).

Data concerning the possible association between cause of death and high salt intake			
Type of diet			
Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

- Note that $E_{11} = 7 \times 25/60 = 2.92 < 5$ and $E_{12} = 7 \times 35/60 = 4.08 < 5$.
- How should the possible association between cause of death and type of diet be assessed?

Hyper-geometric distribution

- X : number of high salt intake for the non-CVD
- Y : number of high salt intake for the CVD
- $H_0: p_1 = p_2 = p$
- Under H_0
 - $X \sim \text{Bin}(n_1; p)$
 - $Y \sim \text{Bin}(n_2; p)$
 - $X + Y \sim \text{Bin}(n_1 + n_2; p)$

$$\Rightarrow P(X = x | X + Y = z) = \frac{\binom{n_1}{x} \binom{n_2}{z-x}}{\binom{n_1+n_2}{z}} \quad (\text{Proof!})$$

\Rightarrow the conditional distribution = hypergeometric

Fisher's exact test

- To test $H_0: p_1 = p_2$ vs $H_1: p_1 \neq p_2$ ($H_2: p_1 < p_2, H_3: p_1 > p_2$)
- General procedure:
 - ① Enumerate all possible tables with the same row and column margins as the observed table.
 - ② Compute the exact probability of each table enumerated in step 1.
 - ③ Suppose the observed table is the a table and the last table enumerated is the k table.
 - H_1 : p-value = $2 \times \min[P(0) + P(1) + \dots + P(a), P(a) + P(a + 1) + \dots + P(k), .5]$.
 - H_2 : p-value = $P(0) + P(1) + \dots + P(a)$
 - H_3 : p-value = $P(a) + P(a + 1) + \dots + P(k)$

Back to Cardiovascular Disease

Cause of death	High salt	Low salt	Total
Non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

- Consider $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$
- Recall we are fixing the row and column totals
- p-value requires table s as **extreme or more extreme** (under H_1) than the one observed.
=> See next slide

Tables supporting H_1

0	25	1	24	2	23	3	22
7	28	6	29	5	30	4	31
.017		.105		.252		.312	

4	21	5	20	6	19	7	18
3	32	2	33	1	34	0	35
.214		.082		.016		.001	

- $P(\text{table1}) = \frac{\binom{25}{0}\binom{35}{7}}{\binom{60}{7}} = 0.017$
- R code: `dhyper(1,25,35,7)`
- $p\text{-value} = 2 \times \min(0.017 + 0.105 + 0.252, 0.252 + 0.312 + 0.214 + 0.082 + 0.016 + 0.001, 0.5)$
 $= 0.749$

R codes

```
> dat <- matrix(c(2,23,5,30), 2)
> fisher.test(dat)
```

Fisher's Exact Test for Count Data

```
data: dat
p-value = 0.6882
alternative hypothesis: true odds ratio is not equal
to 1
95 percent confidence interval:
 0.04625243 3.58478157
sample estimates:
odds ratio
 0.527113
```

McNemar's Test

Two-Sample Test for Binomial
Proportions for Matched-Pair Data

Motivation example

Cancer

- Suppose we want to compare two different chemotherapy regimens for breast cancer after mastectomy.
- The patients are assigned to pairs matched on age (within 5 years) and clinical condition.

Outcome			
Treatment	Survive for 5 years	Die within 5 years	Total
A	526	95	621
B	515	106	621
Total	1041	201	1242

	Outcome (Y=1/0)	
	A	B
Pair 1	1	1
Pair 2	1	0
Pair 3	0	1
Pair 4	0	0
...		

Contd.

- Notice the small difference in 5-year survival between the two treatment groups: treatment A group = $526/621 = .847$ vs. treatment B group = $515/621 = .829$.
- Indeed, the Yates-corrected chi-square statistic is 0.59 with 1 *df*, which is not significant.
- The two samples are *not* independent because the two members of each matched pair are similar in age and clinical condition.
- the Yates-corrected chi-square test *cannot* be used in this situation because the *p*-value will not be correct.

Dependence

- Matched binary can arise from
 - Measuring a response at two occasions
 - Matching on case status in a retrospective study
 - Matching on exposure status in a prospective or cross-sectional study
- The pairs on binary observations are dependent, so our existing methods do not apply

Outcome of treatment A patient	Outcome of treatment B patient		
	Survive for 5 years	Die within 5 years	Total
Survive for 5 years	510	16	526
Die within 5 years	5	90	95
Total	515	106	621

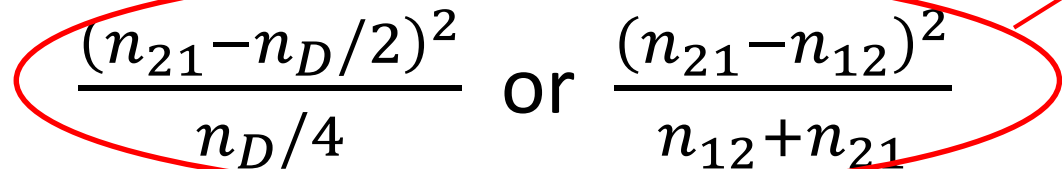
Notation

Outcome of treatment B patient			
Outcome of treatment A patient	Survive for 5 years	Die within 5 years	Total
Survive for 5 years	$n_{11} (\pi_{11})$	$n_{12} (\pi_{12})$	$n_{1+} (\pi_{1+})$
Die within 5 years	$n_{21} (\pi_{21})$	$n_{22} (\pi_{22})$	$n_{2+} (\pi_{2+})$
Total	$n_{+1} (\pi_{+1})$	$n_{+2} (\pi_{+2})$	n

- p = the probability that a discordant pair is of type A
- If A and B are equally effective, then n_{12} and n_{21} would be expected to be equal, and p should be $1/2$
- If A is more effective than B, p is expected to be $< 1/2$; if B is more effective than A, p should be $> 1/2$
- Then the test hypothesis $H_0: p = 1/2$ vs. $H_1: p \neq 1/2$

McNemar's test

- $n_D = n_{12} + n_{21}$
- Under H_0 , $n_{21} \sim \text{Binomial}(n_D, 1/2)$
- Normal approximation


$$\frac{(n_{21} - n_D/2)^2}{n_D/4} \quad \text{or} \quad \frac{(n_{21} - n_{12})^2}{n_{12} + n_{21}}$$

McNemar's test

follows an asymptotic χ^2 distribution with 1 *df*

- Notice that only the discordant cells enter into the test
- Exact p-value

$$p = 2 \times \min \left\{ \sum_{k=0}^{n_{21}} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D}, \sum_{k=n_{21}}^{n_D} \binom{n_D}{k} \left(\frac{1}{2}\right)^{n_D} \right\}$$

Back to Cancer

- Test statistic $X^2 = (16 - 5)^2 / (16 + 5) = 5.76$
- p-value = 0.016
- Hence we reject the null hypothesis
 - conclude that the treatment A member of the pair is significantly more likely to survive for 5 years than the treatment B member.
 - all other things being equal (such as toxicity, cost, etc.), treatment A would be the treatment of choice.

```
> mcnemar.test(matrix(c(510,16,5,90), 2), correct=F)
```

McNemar's Chi-squared test

```
data: matrix(c(510, 16, 5, 90), 2)
```

```
McNemar's chi-squared = 5.7619, df = 1, p-value = 0.01638
```

The `correct` option applies a continuity correction

- With the default `correct=TRUE`

```
> mcnemar.test(matrix(c(510,16,5,90), 2))
```

McNemar's Chi-squared test with continuity correction

```
data: matrix(c(510, 16, 5, 90), 2)
```

```
McNemar's chi-squared = 4.7619, df = 1, p-value = 0.0291
```

- McNemar's test is equivalent to one-sample binomial test

```
> prop.test(5,21,correct=T) # p-value = 0.0291
```

```
> prop.test(5,21,correct=F) # p-value = 0.01638
```

```
> binom.test(5,21) # p-value = 0.0266
```

	Outcome (Y=1/0)	
	A	B
Pair 1	1	1
Pair 2	1	0
Pair 3	0	1
Pair 4	0	0
...		

McNemar's test

	Outcome (Y is continuous)	
	A	B
Pair 1	y11	y12
Pair 2	y21	y22
Pair 3	y31	y32
Pair 4	y41	y42
...		

???

Sample size and power estimation

For comparing two binomial proportions

Motivation example

Cancer, Nutrition

- Suppose we know from Connecticut tumor-registry data that the incidence rate of breast cancer over a 1-year period for initially disease-free women ages 45–49 is 150 cases per 100,000.
- Interested: whether ingesting large doses of vitamin E in capsule form will prevent breast cancer?
- The study is set up with
 - a control group of 45- to 49-year-old women who are mailed placebo pills and are expected to have the same disease rate as indicated in the Connecticut tumor-registry data
 - a study group of similar-age women who are mailed vitamin
- E pills and are expected to have a 20% reduction in risk.
- How large a sample is needed if a two-sided test with a significance level of .05 is used and a power of 80% is desired?

Two proportions comparison

- Consider testing $H_0: p_1 = p_2$ vs $H_1: p_1 \neq p_2$
- the test statistic is

$$TS = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- where $\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$ is the estimate of the common proportion under the null hypothesis.

Independent samples

- Test $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$
- The power is Achieved with sample sizes n_1 and n_2

$$\text{Power} = \Phi \left[\frac{\Delta}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}} - z_{1-\alpha/2} \frac{\sqrt{\bar{p} \bar{q} (1/n_1 + 1/n_2)}}{\sqrt{p_1 q_1 / n_1 + p_2 q_2 / n_2}} \right]$$

where p_1, p_2 = projected true probabilities of success in the two groups

$$q_1, q_2 = 1 - p_1, 1 - p_2$$

$$\Delta = |p_2 - p_1|$$

$$\bar{p} = \frac{p_1 + k p_2}{1 + k}$$

$$\bar{q} = 1 - \bar{p}$$

Derived by Normal approximation

- With a significance level α and power $1 - \beta$, the following sample size is required

$$n_1 = \left[\sqrt{\bar{p} \bar{q} \left(1 + \frac{1}{k} \right)} z_{1-\alpha/2} + \sqrt{p_1 q_1 + \frac{p_2 q_2}{k}} z_{1-\beta} \right]^2 / \Delta^2$$

$$n_2 = k n_1$$

Back to Cancer

- $p_1 = 150 \text{ per } 100,000 \text{ or } 150/10^5 = .00150$
- $q_1 = 1 - .00150 = .99850$
- $p_2 = 0.8 p_1 = .00120$ and $q_2 = 1 - .00120 = .99880$
- $1 - \beta = 0.8, \alpha = .05$
- $k = 1$ (because $n_1 = n_2$)
- $\bar{p} = (.00150 + .00120)/2 = .00135, \bar{q} = .99865$
- Thus
- $n_1 = \dots = 234,881 = n_2$
- or about 235,000 women in each group.

Paired samples

- If McNemar's test is used, test whether or not $p = 1/2$, where p is the probability of one type of discordant pair (A) among all discordant pairs
- with a significance level α and power $1 - \beta$, then the sample size is determined by

$$n = \frac{(z_{1-\alpha/2} + 2z_{1-\beta}\sqrt{p_A q_A})^2}{4(p_A - .5)^2 p_D} \text{ matched pairs}$$

- And the power is Achieved with n matched pairs

$$\text{Power} = \Phi \left[\frac{1}{2\sqrt{p_A q_A}} (z_{\alpha/2} + 2|p_A - .5|\sqrt{np_D}) \right]$$

where p_D = projected proportion of discordant pairs among all pairs
 p_A = projected proportion of discordant pairs of type A among discordant pairs

One-sample
binomial test

Example

Cancer

- we want to compare two different regimens of chemotherapy (A, B) for treatment of breast cancer .
- Matched-pair design: matched on age and clinical stage of disease
- Based on previous work, it is estimated that patients in a matched pair will respond similarly to the treatments in 85% of matched pairs (i.e., both will either die or have a recurrence or both will be alive and not have a recurrence).
- Furthermore, for matched pairs in which there is a difference in response, it is estimated that in two-thirds of the pairs the treatment A patient will either die or have a recurrence, and the treatment B patient will not.

Contd.

- How many participants (or matched pairs) need to be enrolled in the study to have a 90% chance of finding a significant difference using a two-sided test with type I error = .05?
- We have $\alpha = .05$, $1 - \beta = .90$, $p_D = 1 - .85 = .15$, $p_A = \frac{2}{3}$, $q_A = \frac{1}{3}$.

$$\begin{aligned} n \text{ (pairs)} &= \frac{\left[z_{.975} + 2z_{.90} \sqrt{(2/3)(1/3)} \right]^2}{4(2/3 - 1/2)^2 (.15)} \\ &= \frac{[1.96 + 2(1.28)(.4714)]^2}{4(1/6)^2 (.15)} = \frac{3.1668^2}{.0167} = 602 \text{ matched pairs} \end{aligned}$$

- Therefore, 1204 women in 602 matched pairs must be enrolled. This will yield approximately $.15 \times 602 = 90$ discordant pairs.

Contd.

- If 400 matched pairs are enrolled, how much power would such a study have?
- We have $\alpha = .05$, $p_D = .15$, $p_A = \frac{2}{3}$, $n = 400$.

$$\begin{aligned}\text{Power} &= \Phi \left\{ \frac{1}{2\sqrt{(2/3)(1/3)}} \left[z_{.025} + 2 \left| \frac{2}{3} - .5 \right| \sqrt{400(.15)} \right] \right\} \\ &= \Phi \{ 1.0607 [-1.96 + 2(1/6)(7.7460)] \} \\ &= \Phi [1.0607 (0.6220)] = \Phi (0.660) = .745\end{aligned}$$





- Therefore, the study would have 74.5% power, or a 74.5% chance of detecting a statistically significant difference.

Compliance

- In practice, the compliance is not perfect.
- How the sample size and power estimation change under this circumstance?
- In medicine, compliance describes the degree to which a patient correctly follows medical advice. (from Wikipedia)
- A clinical trial comparing an active treatment vs. placebo
- Two types of noncompliance to consider:
 - dropout rate
 - drop-in rate

Example

Cardiovascular Disease

- A 5-year randomized clinical trial : assess the effect of aspirin in preventing myocardial infarction (MI).
-  **dropin rate**
•  10% of the participants in the aspirin group were not complying (that is, were not taking their study [aspirin] capsules).
-  **dropin rate**
•  5% of the participants in the placebo group were taking aspirin regularly on their own outside the study protocol.
- How does this lack of compliance affect the sample size and power estimates for the study?

Notations

- λ_1 = dropout rate, λ_2 = drop-in rate
- under the assumption of perfect compliance:
- p_1 = incidence of MI among physicians who actually take aspirin
- p_2 = incidence of MI among physicians who don't take aspirin
- Assuming that compliance is not perfect:
- p_1^* = observed rate of MI in the aspirin groups
- p_2^* = observed rate of MI in the placebo groups
- With the help of the total probability rule!!!

Adjustment

$$\begin{aligned} p_1^* &= P(\text{MI} | \text{assigned to aspirin group}) \\ &= P(\text{MI} | \text{aspirin-group complier}) \times Pr(\text{compliance in the aspirin group}) \\ &\quad + P(\text{MI} | \text{aspirin-group noncomplier}) \times P(\text{noncompliance in the aspirin group}) \\ &= p_1(1 - \lambda_1) + p_2\lambda_1 \end{aligned}$$

$$\begin{aligned} p_2^* &= P(\text{MI} | \text{assigned to placebo group}) \\ &= P(\text{MI} | \text{placebo-group complier}) \times Pr(\text{compliance in the placebo group}) \\ &\quad + P(\text{MI} | \text{placebo-group noncomplier}) \times P(\text{noncompliance in the placebo group}) \\ &= p_2(1 - \lambda_2) + p_1\lambda_2 \end{aligned}$$

- In the presence of noncompliance, sample size and power estimates should be based on the compliance-adjusted rates (p_1^*, p_2^*) rather than on the perfect compliance rates (p_1, p_2) .

Back to Cardiovascular Disease

- Suppose we assume that the incidence of MI is .005 per year among participants who actually take placebo and that aspirin prevents 20% of MIs (i.e., relative risk = $p_1/p_2 = 0.8$)
- How many participants need to be enrolled in each group to achieve 80% power using a two-sided test with significance level = .05?

Solution

- This is a 5-year study, so the 5-year incidence of MI among participants who actually take placebo $\approx 5(.005) = .025 = p_2$.
- Because the risk ratio = 0.8, we have $p_1/p_2 = 0.8$ or $p_1 = .020 =$ 5-year incidence of MI among participants who actually take aspirin.
- $p_1^* = p_1(1 - \lambda_1) + p_2\lambda_1 = .9(.020) + .1(.025) = .0205$
- $p_2^* = p_2(1 - \lambda_2) + p_1\lambda_2 = .95(.025) + .05(.020) = .02475$

Contd.

- Plug p_1^*, p_2^* in equation of slide 38, we have $n_1 = n_2 = 19,196$, total 38,392
- Plug p_1, p_2 in equation of slide 38, we have $n_1 = n_2 = 13,794$, total 27,588
- $(p_1, p_2) = (0.020, 0.025)$ vs $(p_1^*, p_2^*) = (0.0205, 0.02475)$

=> Increase the required sample size by $100\% \times (38,392 - 27,588)/27,588 = 39\%$.

$R \times C$ Contingency Tables

Motivation example

Breast Cancer

- Suppose we want to study further the relationship between age at first birth and development of breast cancer, as the first example in this class.
- We would like to know whether the effect of age at first birth follows a consistent trend, that is,
 - more protection for women whose age at first birth is <20 than for women whose age at first birth is 25–29
 - higher risk for women whose age at first birth is ≥ 35 than for women whose age at first birth is 30–34.
- We want to test for a **relationship** between age at first birth and case–control status.

Outcomes

- Data from the international study investigating the possible association between age at first birth and case-control status

Age at first child-birth						
Status	<20	20-24	25-29	30-34	≥ 35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465
% cases	.184	.214	.259	.298	.351	.239

Chi-Square Test

- Under the following conditions:
 - No more than 1/5 of the cells have expected values < 5 .
 - No cell has an expected value < 1 .

- The test statistic

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Follows χ^2 distribution with $(R - 1) \times (C - 1)$ df
- Reject if the statistic is too large.

Back to Breast Cancer

Observed table

Age at first child-birth						
Status	<20	20-24	25-29	30-34	≥ 35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465

Expected table

Age at first child-birth						
Status	<20	20-24	25-29	30-34	≥ 35	Total
Case	416.6	1348.3	933.6	371.9	149.7	3220
Control	1325.4	4289.7	2970.4	1183.1	476.3	10,245
Total	1742	5638	3904	1555	626	13,465

Contd.

- Note that the $E_{ij} > 5$ for all i, j .

- The χ^2 test statistic is

$$\chi^2 = \frac{(320-416.6)^2}{416.6} + \frac{(1206-1348.3)^2}{1348.3} + \dots + \frac{(406-476.3)^2}{476.3} = 130.3$$

- Under H_0 , χ^2 follows a chi-square distribution with $(2 - 1) \times (5 - 1)$, or 4, df .
- The p-value $< .001 \Rightarrow$ highly significant

```
> pchisq(130.3, 4, lower.tail = FALSE)
[1] 3.359266e-27
```

R codes

```
> dat <- t(matrix(c(320,1206,1011,463,220,  
+                  1422,4432,2893,1093,406),5))  
> chisq.test(dat)
```

Pearson's Chi-squared test

```
data: dat  
X-squared = 130.172, df = 4, p-value <  
2.2e-16
```

Notes

- Association and independence test yield the same results
- Same test results if
 - row totals are fixed
 - column totals are fixed
 - total count is fixed
 - none are fixed
- Note that this is common in statistics; mathematically equivalent results are applied in different settings, but result in different interpretations

Recalled Breast Cancer

- The “age at first child-birth” is an ordinal variable.
- Moreover, we notice an increasing *trend* in the proportion of women with breast cancer in each succeeding column.

Age at first child-birth						
Status	<20	20-24	25-29	30-34	≥ 35	Total
Case	320	1206	1011	463	220	3220
Control	1422	4432	2893	1092	406	10,245
Total	1742	5638	3904	1555	626	13,465
% cases	.184	.214	.259	.298	.351	.239

Score

- The score variable can represent some particular numeric attribute of the group.
 - a quantitative exposure category
 - => use the average exposure within the category as the score.
 - the exposure level is not easily quantified
 - => either mid-ranks or consecutive integers are reasonable choices for scores.
- For example in the previous example

Age at first child-birth				
<20	20-24	25-29	30-34	≥ 35
17.5	22.5	27.5	32.5	37.5

Chi-Square Test for Trend in Binomial Proportions

$2 \times k$	Score/categories			
	S_1	...	S_k	Total
frequency	x_1	...	x_k	$x = \sum x_i$
total	n_1	...	n_k	$n = \sum n_i$

$\bar{p} = \frac{x}{n}$

- From regression approach we get (if $np(1 - p) \geq 5$) :

$$X_{trend}^2 = \frac{(\sum x_i S_i - x \bar{S})^2}{\bar{p}(1 - \bar{p})(\sum n_i S_i^2 - (\sum n_i S_i)^2)/n}$$

under H_0 approximately follows a chi-square distribution with 1 *df*.

Back to Breast Cancer

- Scores: $S_i = 1, 2, 3, 4, 5$
- $X_{trend}^2 = 129.01 \sim \chi_1^2$ under H_0 p-value $< .001$
- **Conclusion:** a significant trend in the proportion of breast-cancer cases among age-at-first-birth groups.
- Learning to program R is a worthwhile endeavor as it allows you to program non-standard test procedures yourself.

```
> x <- c(320,1206,1011,463,220)
> n <- c(1742,5638,3904,1555,626)
> scores <- c(1,2,3,4,5)
> Ptrend(x,n,scores) # self program
```

Test for Trend in Binomial Proportions

=====

A = 567.16

Chi-square Statistic = 129.012

p-value = 0

Goodness-of-Fit

Using Chi-Square Test

Motivation example

Hypertension

- Diastolic blood-pressure measurements were collected at home in a community-wide screening program of 14,736 adults ages 30–69 in East Boston, Massachusetts.
- Whether these measurements came from an underlying normal distribution?
- Frequency distribution of mean diastolic blood pressure

Group (mm,Hg)	< 50	≥ 50, < 60	≥ 60, < 70	≥ 70, < 80	≥ 80, < 90	≥ 90, < 100	≥ 100, < 110	≥ 110	Total
Freq	57	330	2132	4584	4604	2119	659	251	14,736

Chi-Square Test

- Under the following conditions:
 - No more than 1/5 of the cells have expected values <5.
 - No cell has an expected value <1.

- The test statistic

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- Follows χ^2 distribution with $g - 1 - k$ df, where g=the number of groups and k=number of parameters to be estimated
- For the example **Hypertension**
 - $k = 2, g = 8, g - 1 - k = 5$
 - $X^2 = \dots = 350.2$
 - p-value < .001

Kappa statistic

Measures of Association
for nominal data

Motivation example

Nutrition

- A diet questionnaire was mailed to 537 female American nurses on two separate occasions several months apart. The questions asked included the quantities eaten of more than 100 separate food items.

	Survey 2		
Survey 1	≤ 1 serving/week	> 1 serving/week	Total
≤ 1 serving/week	136	92	228
> 1 serving/week	69	240	309
Total	205	332	537

- Notice that the responses on the two surveys are the same only for $136 + 240 = 376$ out of 537 (70.0%) women.
- How can **reproducibility** of response for the beef-consumption data be quantified?

Contd.

- Let p_o = observed probability of concordance between the two surveys
 - In the previous example, $p_o = (136+240)/537 = 0.7$
 - What is the expected probability of concordance if the survey responses are independent?
 - Let
 - a_i : the marginal probabilities for the i th category in survey 1
 - b_j : the marginal probabilities for the j th category in survey 2
 - Then the expected probability of concordance $p_e = \sum a_i b_i$
 - $a_1 = \frac{228}{537} = .425, a_2 = \frac{309}{537} = .575, b_1 = \frac{205}{537} = .382, b_2 = \frac{332}{537} = .618$
 - $p_e = (.425 \times .382) + (.575 \times .618) = .518$
- Cohen's Kappa

Kappa statistic

- The Cohen's Kappa statistic (κ) measures the agreement/association between two variables
- defined by

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

- Where
 - p_o = observed probability of concordance between the two surveys
 - p_e = expected probability of concordance between the two surveys = $\sum a_i b_i$, a_i, b_i is the marginal probabilities of the two surveys.

Kappa statistic

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

$$z = \frac{\kappa}{se(\kappa)} \sim \mathcal{N}(0,1)$$

- where the Standard Error

$$se(\kappa) = \frac{1}{(1 - p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum a_i b_i (a_i + b_i)}$$

- Note that we are customarily interested in one-tailed tests because negative values for Kappa usually have no **biological significance**.

Back to Nutrition

- $p_o = 0.700, p_e = 0.518$
- $\kappa = \frac{0.700 - 0.518}{1 - 0.518} = 0.378$
- $se(\kappa) = \dots = 0.0430$
- $z = \frac{0.378}{0.0430} = 8.8 \sim \mathcal{N}(0,1)(\text{under } H_0)$
- $p\text{-value} = 1 - \Phi(8.8) < 0.001$
- Conclusion: **highly significant** reproducibility between the first and second surveys for beef consumption.

R codes

```
> library("vcd")
> dat <- matrix(c(136,69,92,240),2)
> k.dat <- Kappa(dat)
> summary(k.dat)
```

	value	ASE
Unweighted	0.3781906	0.04100635
weighted	0.3781906	0.05504449

weights:

	[,1]	[,2]
[1,]	1	0
[2,]	0	1

```
> confint(k.dat)
```

Kappa	lwr	upr
Unweighted	0.2978196	0.4585616
weighted	0.2703054	0.4860758

Guidelines

- Guidelines for Evaluating Kappa

Kappa	Measure of reproducibility
> 0.75	Excellent
$0.4 \leq \kappa \leq 0.75$	Good
$0 \leq \kappa < 0.4$	marginal

- In general, reproducibility is not good for many items in dietary surveys, indicating the need for multiple dietary assessments to reduce variability.