PIB2: Data Analysis

Part I

Two-Sample Z Test for P

Inferences About Proportions of Two Groups

Part II

- χ^2(chi square) distribution
- χ^2(chi square) test

PIB2: Data Analysis

Part II

- x^2 (chi square) distribution
- x^2 (chi square) test

PIB2: Data Analysis – χ^2 distribution/test

χ^2 **(chi square) distribution**

Preliminary Idea: Sum of n values of a random variable

We know that **if** a random variable X is **normal**, then:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

has a Student′s t−distribution with $(n - 1)$ degrees of freedom. Since:

$$x_1 + x_2 + x_3 + \ldots + x_n = n\bar{x},$$

It follows that the sum of n values of a normal random variable X also follows a Student′s t-distribution with $(n - 1)$ degrees of freedom.

Data Analysis –χ2 distribution/test

χ^2 (chi square) distribution

Sum of Squares of random numbers

Question: If $x_1 + x_2 + x_3 + \ldots + x_n$ follows a Student′s t-distribution with $(n-1)$ degrees of freedom, what can we say about the **sum of the squares** of these values?

In other words, what is the distribution of → $x_1^2 + x_2^2 + \cdots + x_n^2$?

Answer:
Statisticians have shown that **<u>if *X* is normal</u>**, then the sum of squares of *n* values of x_i, namely,

$$x_1^2 + x_2^2 + \cdots + x_n^2$$

has a χ^2 distribution with $(n-1)$ degrees of freedom.

Data Analysis –χ2 distribution/test

χ^2 (chi square) distribution

1. It is called the chi-square distribution.

2. “Chi” rhymes with “High” – and the “ch” is pronounced like “k”.

3. It is a continuous random variable.

4. It has n – 1 degrees of freedom.

5. It’s values are non-negative (i.e. $\geq$ 0).

6. It is always skewed to the right.

7. It becomes more symmetrical as n increases.

8. It approximates a normal distribution for large values of n.

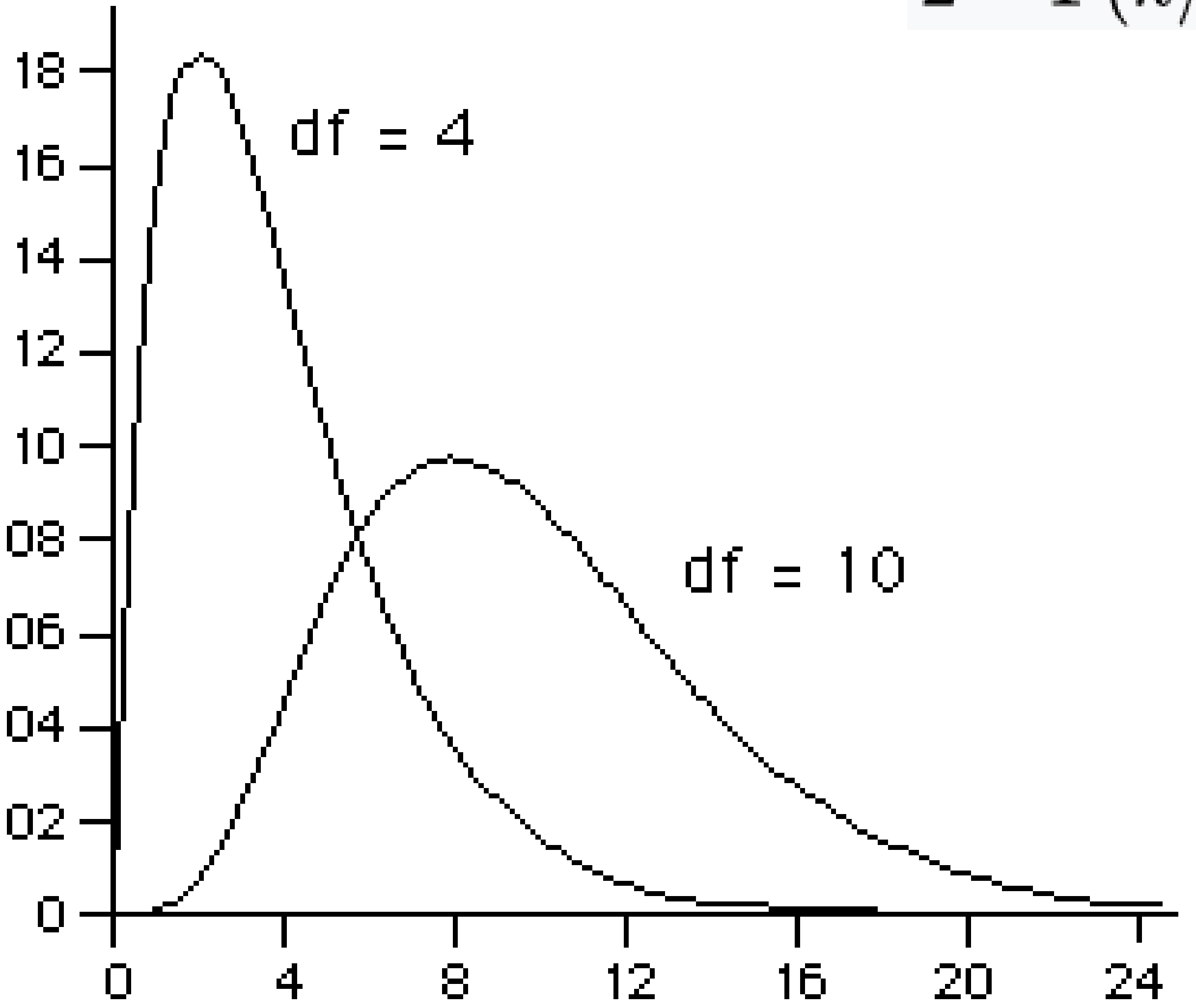
Data Analysis –χ2 distribution/test

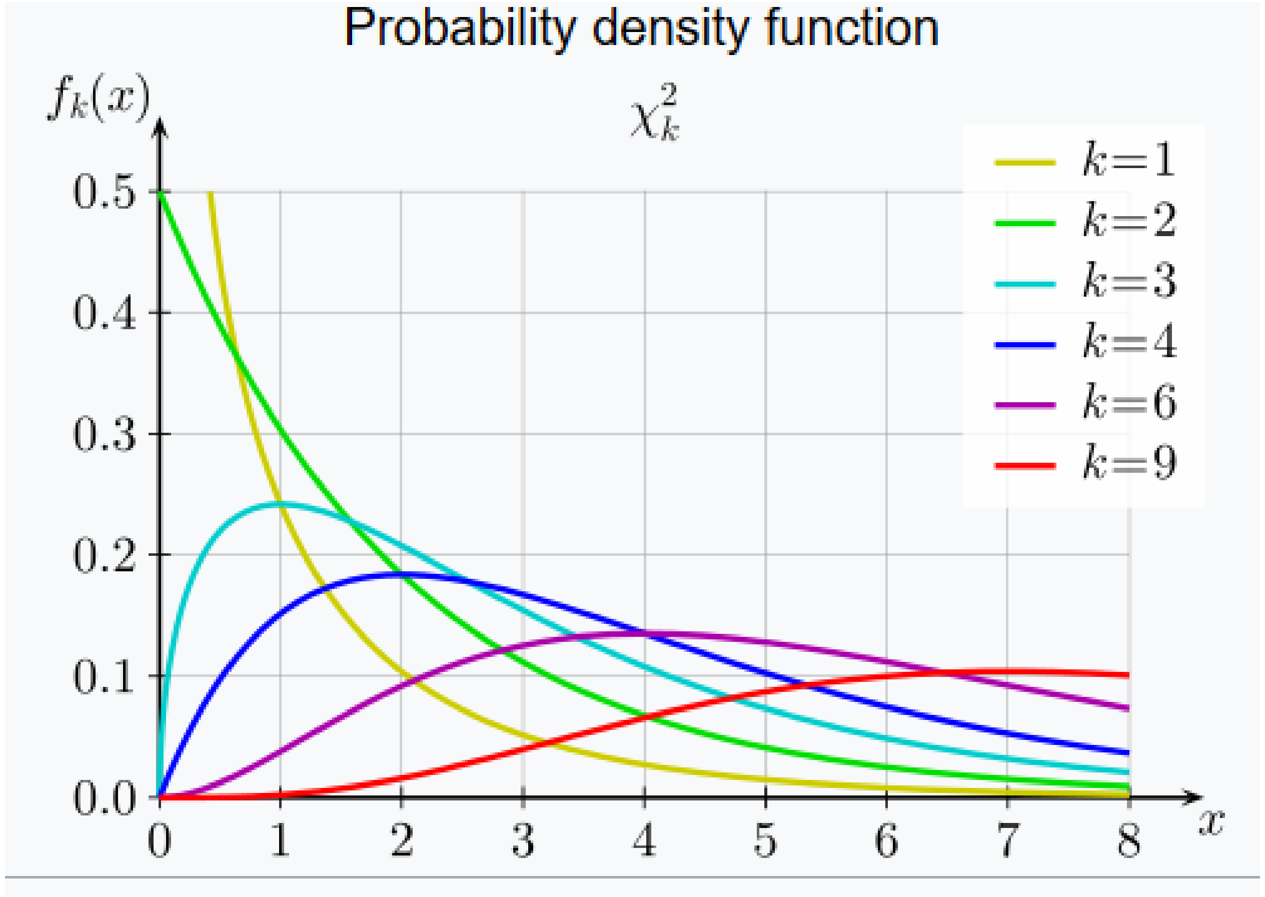
χ^2 (chi square) distribution

$$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$$

$$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$$

Two Chi-square distributions

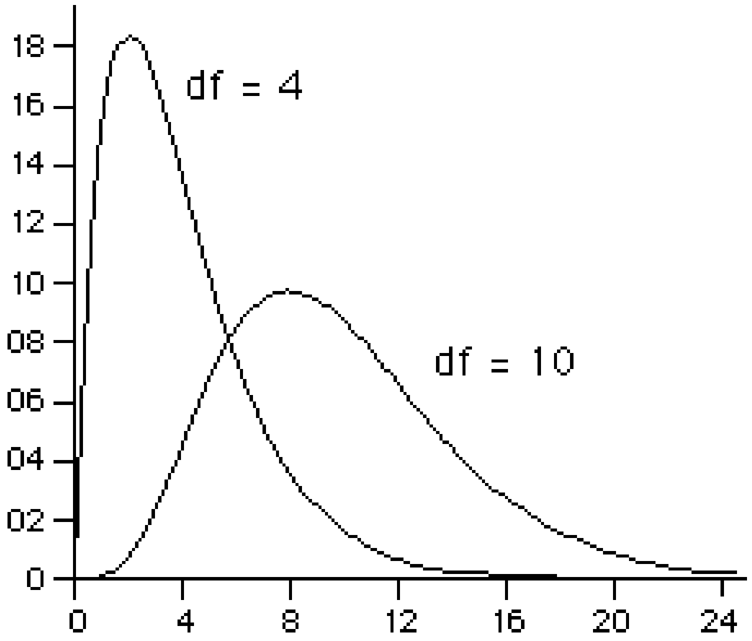




Data Analysis –χ2 distribution/test

χ^2 (chi square) distribution

The sample variance s^2 follows a chi-square distribution



The **sample variance** is defined by:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

It follows that:

$$(n-1)s^2 = \sum(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2.$$

Since the right-hand-side of this expression is a sum of squares it follows that, when X is normal, $(n-1)s^2$ has a χ^2 distribution with $(n-1)$ degrees of freedom.

Data Analysis –χ2 distribution/test

χ^2 (chi square) distribution

Standardizing the Test Statistic

In a test of hypothesis for a population variance σ^2, the test statistic is the sample variance s^2. The standardized test statistic is denoted by $\boldsymbol{\chi}^{\mathbf{2}*}$ and is defined by:

$$\chi^{2*} = \frac{(n-1)s^2}{\sigma_0^2}$$

Note: The standardized values are found in the standard chi-square tables.

Chi-square table characteristics

- The chi-square tables are *not* symmetrical.
- Therefore lower-tail values and upper-tail values must be listed separately.
- In the extract of the chi-square tables shown in the next slide, lower-tail areas are shaded in yellow, upper tail areas are shaded in blue.

Data Analysis –χ2 distribution/test

χ^2 (chi square) distribution

Chi-square table

df	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.0000393	0.000157	0.000982	0.00393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0001	0.020	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.003	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.018	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.056	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.126	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.228	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	0.36	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	0.53	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	0.73	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	0.95	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	1.20	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
25	6.08	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	6.55	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	7.03	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.65
28	7.50	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	8.00	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	8.50	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

Basic Logic

- Chi Square is a test of significance based on **bivariate tables**.
- We are looking for significant differences between the **actual cell frequencies** in a table (**f_o**) and those that would be **expected by random chance** (**f_e**).
- The data are often presented in a **table format**. If starting with raw data on two variables, a **bivariate table** must be created first.

Data Analysis –χ2 distribution/test

χ^2 **(chi square) test**

Bivariate Tables:

- Must have a title.
- Cells are intersections of columns and rows.
- Subtotals are called marginals.
- *N* is reported at the intersection of row and column marginals.
- Columns are scores of the independent variable.
- There will be as many columns as there are scores on the independent variable.
- Rows are scores of the dependent variable.
- There will be as many rows as there are scores on the dependent variable.
- There will be as many cells as there are scores on the two variables combined.
- Each cell reports the number of times each *combination* of scores occurred.

Data Analysis –$\chi 2$ distribution/test

χ^2 (chi square) test

What your table should look like:

	Title		
Rows	Columns →		Total
Row 1	cell a	cell b	Row Marginal 1
Row 2	cell c	cell d	Row Marginal 2
Total	Column Marginal 1	Column Marginal 2	N

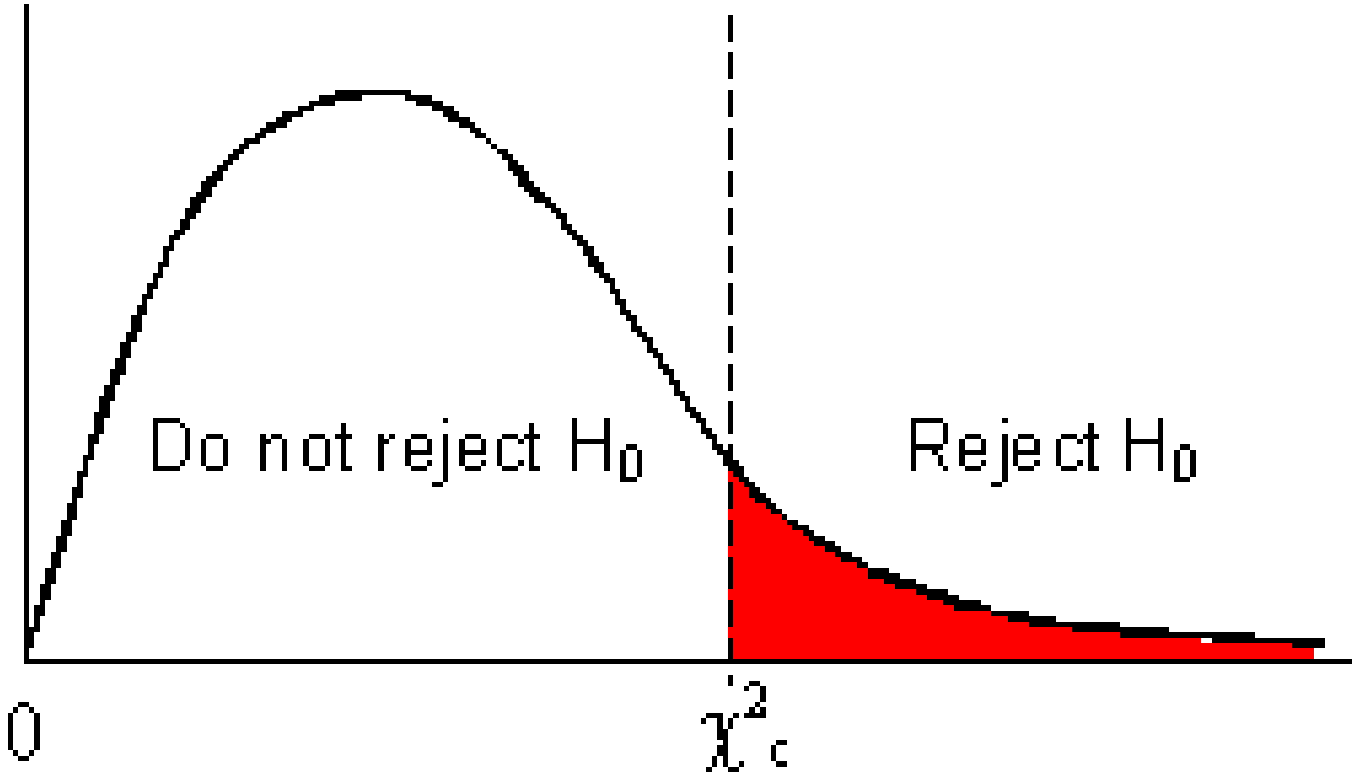
Data Analysis –χ2 distribution/test

χ^2 (chi square) test

The Chi Square Distribution

- The chi square distribution is asymmetric and its values are always positive.
- Degrees of freedom are based on the table and are calculated as:

$$df = (rows - 1) \times (columns - 1)$$



Data Analysis –χ2 distribution/test

χ^2 (chi square) test

Example

Question: Are the homicide rate and volume of gun sales related for a sample of 25 cities?

	HOMICIDE RATE		
GUN SALES	Low	High	Totals
High	8	5	13
Low	4	8	12
Totals	12	13	N = 25

The bivariate table showing the relationship between homicide rate (columns) and gun sales (rows). This 2 × 2 table has 4 cells.

Data Analysis –χ2 distribution/test

χ^2 **(chi square) test** (Solution Using 5-Step Method)

Step 1 Make Assumptions and Meet Test Requirements

- Independent random samples.
- Level of measurement is nominal.
- Note that no assumption is made about the shape of the sampling distribution.
- When the distribution is normal, a **parametric test** (Z- or t-test, ANOVA) can be used.
- The chi square test is **non-parametric**. It can be used when **normality is not assumed**.

Step 2 State the Null and Alternate Hypothesis

H_0: The variables are independent

You can also say: $H_0\colon\ f_0 = f_e$

H_1: The variables are dependent

Or: $H_1\colon f_0 \neq f_e$

Data Analysis –χ2 distribution/test

χ^2 **(chi square) test** (Solution Using 5-Step Method)

Step 3 Select the Sampling Distribution and Establish the Critical Region

- Because normality is not assumed and our data are in tabular form, our sampling distribution → χ^2
- $\alpha = 0.05$
- $df \;=\; (r-1)(c-1) \;=\; 1$
- $\chi^2\,(critical) \;=\; 3.841$

	HOMICIDE RATE		
GUN SALES	Low	High	Totals
High	8	5	13
Low	4	8	12
Totals	12	13	N = 25

Step 4 Calculate the Test Statistic

Formula:

$$\chi^2\,\text{(obtained)} = \sum \frac{(f_o - f_e)^2}{f_e}$$

Method:

1. Find expected frequencies for each cell.
2. Complete computational table to find χ^2 (obtained)

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

1. Find expected frequencies for each cell.

- To find $f_e = \dfrac{\text{row marginal} \times \text{column marginal}}{N}$

	HOMICIDE RATE		
GUN SALES	Low	High	Totals
High	8	5	13
Low	4	8	12
Totals	12	13	N = 25

- Multiply column and row marginals for each cell and divide by N.
    - (13 · 12)/25 = 156/25 = 6.24
    - (13 · 13)/25 = 169/25 = 6.76
    - (12 · 12)/25 = 144/25 = 5.76
    - (12 · 13)/25 = 156/25 = 6.24

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

$$f_e = \frac{\text{row marginal} \times \text{column marginal}}{N}$$

Multiply column and row marginals for each cell and divide by N.

$(13 \cdot 12)/25 = 156/25 = 6.24$
$(13 \cdot 13)/25 = 169/25 = 6.76$
$(12 \cdot 12)/25 = 144/25 = 5.76$
$(12 \cdot 13)/25 = 156/25 = 6.24$

Observed and Expected Frequencies for each cell (Note that totals are unchanged):

	HOMICIDE RATE		
GUN SALES	Low	High	Total
High	$f_o = 8$ $f_e = 6.24$	$f_o = 5$ $f_e = 6.76$	13
Low	$f_o = 4$ $f_e = 5.76$	$f_o = 8$ $f_e = 6.24$	12
Total	12	13	N = 25

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

2. Complete Computational Table

- A table like this will help organize the computations:
- (a) Add values for **f_o** and **f_e** for each cell to table.

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
8	6.24			
5	6.76			
4	5.76			
8	6.24			
Total 25	25			

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

Computational Table (cont.)

- (b) Subtract each f_e from each f_o. The total of this column *must* be zero.

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
8	6.24	1.76		
5	6.76	-1.76		
4	5.76	-1.76		
8	6.24	1.76		
Total 25	25	0		

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

Computational Table (cont.)

- (c) Square each of these values

f_o	f_e	f_o - f_e	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
8	6.24	1.76	3.10	
5	6.76	-1.76	3.10	
4	5.76	-1.76	3.10	
8	6.24	1.76	3.10	
Total 25	25	0		

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

Computational Table (cont.)

- (d) Divide each of the squared values by the f_e for that cell.
- (e) The sum of this column is chi square.

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2 /f_e$
8	6.24	1.76	3.10	.50
5	6.76	-1.76	3.10	.46
4	5.76	-1.76	3.10	.54
8	6.24	1.76	3.10	.50
Total 25	25	0		χ^2 = 2.00

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

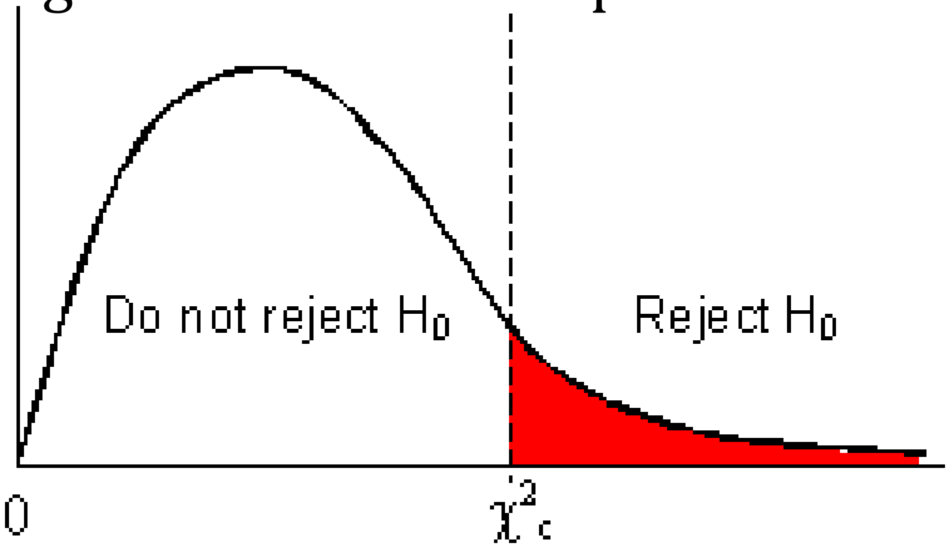
Step 5 Make a Decision and Interpret the Results of the Test

χ^2 (critical) = 3.841
χ^2 (obtained) = 2.00

The **test statistic is not** in the **Critical Region**.

Fail to reject the H_0 (→ The variables are independent).

Conclusion: There is no significant relationship between homicide rate and gun sales (based on these data).



Data Analysis –χ2 distribution/test

χ^2 **(chi square) test**

Interpreting Chi Square

- The chi square test tells us ***only*** if the variables are **independent** or not.
- It does not tell us the pattern or nature of the relationship.
- To investigate the pattern, compute % within each column and compare across the columns.

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

Interpreting Chi Square

- Cities low on homicide rate were high in gun sales and cities high in homicide rate were low in gun sales.
- As homicide rates increase, gun sales decrease. This relationship is not significant but does have a clear pattern.

	HOMICIDE RATE		
GUN SALES	Low	High	Total
High	8 **(66.7%)**	5 **(38.5%)**	13
Low	4 **(33.3%)**	8 **(61.5%)**	12
Total	12 **(100%)**	13 **(100%)**	N = 25

Data Analysis –χ2 distribution/test

χ^2 (chi square) test

The Limits of Chi Square

- Like all tests of hypothesis, chi square is sensitive to **sample size**.
  - As N increases, obtained chi square increases.
  - With large samples, trivial relationships may be significant.

    To correct for this, when $N > 1000$, set your $\alpha = 0.01$.

- Remember: **significance** is not the same thing as **importance**.

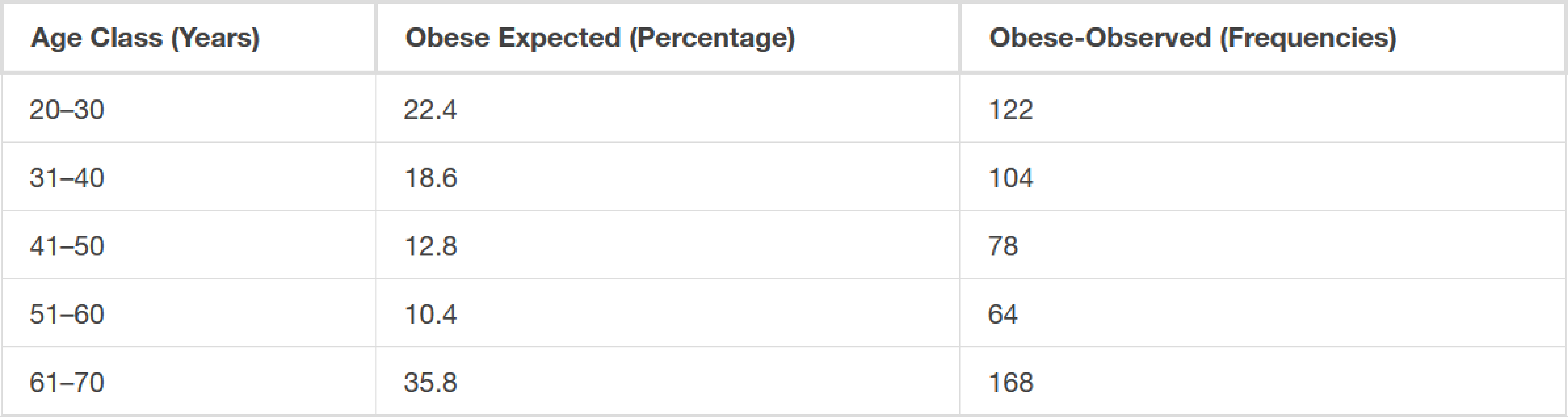
Note:

- If you use this approach on an exam, you may also want to mention **why this approach is appropriate**.
- Specifically, the approach is appropriate because the sampling method was simple random sampling, the variables under study were categorical, and the expected **frequency count** was at **least 5** in each cell of the contingency table.

PIB2: Data Analysis

Classroom activity

The table below contains information from a survey among 499 participants classified according to their age groups. The second column shows the percentage of obese people per age class among the study participants. The last column comes from a different study at the national level that shows the corresponding percentages of obese people in the same age classes in the USA. Perform a hypothesis test at the 5% significance level to determine whether the survey participants are a representative sample of the USA obese population.



Age Class (Years)	Obese Expected (Percentage)	Obese-Observed (Frequencies)
20–30	22.4	122
31–40	18.6	104
41–50	12.8	78
51–60	10.4	64
61–70	35.8	168

Data Analysis –χ2 distribution/test

Video Session

- **An Introduction to the Chi-Square Distribution**: https://www.youtube.com/watch?v=hcDb12fsbBU
- **What is a Chi-Square?** https://www.youtube.com/watch?v=ZjdBM7NO7bY

Data Analysis –χ2 distribution/test

Homework

Data Analysis - **HWA07** – Group Exercise Sheet: 7

- Two-Sample Z Test for P
- χ^2(chi square) test

Data Analysis –χ2 distribution/test

Hands-on Activity – R session

Practice Exercises

Data Analysis - **HWA08** – Group Exercise Sheet: 8

`PIB2_HWA08_Chi_Test_R.Rmd`