STAT S251F: Unit 5

Analysis of Variance

&

Chi Square Test

Tony CHAN, Ph.D.

1

Part I

Multi-sample Inference:

Analysis of Variance (ANOVA)

2

5.1 Introduction

- In Unit 3 (Hypothesis Testing), we learned to test the significance of ***two population means*** using ***two-sample unpaired or paired t test***.
- The ***number of populations*** (***k***) considered in Unit 3 is ***only two*** and the *t* tests mentioned above are only suitable for testing ***two population means***.

How to perform this kind of significance test for $k \geq 3$?

- One intuitive and plausible way of testing is to test ***two population means at a time***. If this method works, we have to perform C_2^k such *t* tests.
- The workload is very heavy if the tests are performed manually. In fact, this ***seemingly workable approach*** will lead to ***erroneous conclusion (involving complicated probability calculation of Type I error)***.

3

5.1 Introduction (Con'td)

- ANOVA is a technique that ***partitions*** the total sum of squares of deviations (***SST***) of the observations about their mean into ***portions*** associated with ***independent variables in the experiment*** and a ***portion associated with error.***
- The ***ANOVA table*** was previously discussed in the Unit 4: "Linear Correlation & Regression Models" with ***quantitative independent variables***. In this unit, the focus will be on ***nominal independent variables*** (called ***factors*** in ANOVA).

4

5.2 Terminologies in ANOVA

Factor/Independent Variable

- A ***factor*** refers to a ***nominal/qualitative/categorical variable*** under examination in an ***experiment*** as a possible ***cause of variation*** in the ***response variable, y***. One or more factors may be involved in a given study.

Levels/Factor levels

- ***Levels*** refer to the ***categories***, ***measurements***, or ***strata*** of a factor of interest in the experiment.

e.g.1 Education attainment is a factor. Its levels may include:
kindergarten, primary, secondary, post-secondary,
post-graduate (5 levels)
e.g.2: Marital Status: Single, married, divorced (3 levels)
e.g.3: Gender: male, female (2 levels)

- Each specific level of a factor (or, in multi-factor experiments, the intersection of a level of one factor with a level of another factor) is referred to as a ***treatment***.

5

5.2 Terminologies in ANOVA (Cont'd)

Experiment

- A study or investigation designed for the purpose of examining the effect that one variable has on the value of another variable.

Experimental Unit

- A unit such as a person, a tree, a pig that receives a treatment (e.g. training method, amount of fertilizer or feed) is called an experimental unit.

Dependent Variable

- We measure or observe values from this dependent variable. In *ANOVA*, the dependent variable will be a ***quantitative variable.*** E.g. soft drink consumption, examination score, or the time required to type a document.

6

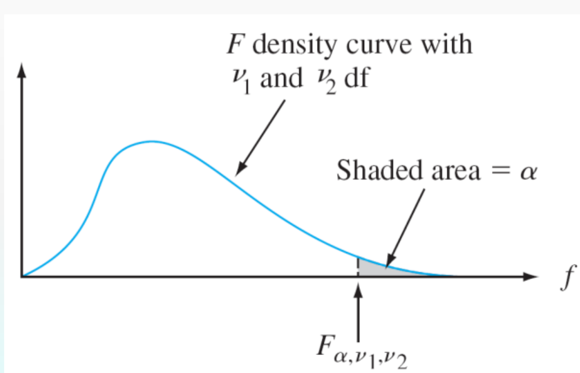
5.3 F-distribution

- ANOVA procedures rely on a statistical distribution called the F-distribution.
- A variable is said to have an F-distribution if its distribution has the shape of a special type of a right skewed curve, F-curve.
- There are infinitely many F-distributions. An F-distribution can be identified by its two numbers of degrees of freedom, v_1 and v_2: $F(v_1, v_2)$ or F_{v_1,v_2}.
- v_1= d.f. for the numerator; v_2= d.f. for the denominator.
  (Since the F-distribution is a ratio of 2 independent chi-squares distributions.)

7

5.3 F-distribution (Cont'd)

The graph of a typical *F* density function is shown below:



As can be seen from the F-curve above, a random variable that has an *F* distribution cannot assume a negative value.

8

5.3 F-distribution (Cont'd)

- Analogous to the notation $t_{\alpha;v}$, we use $F_{\alpha;v_1,v_2}$ for the value on the horizontal axis that captures α of the area under the F density curve with v_1 and v_2 df in the upper tail.

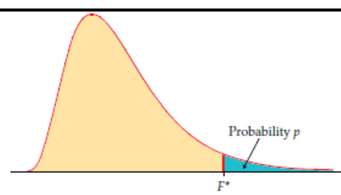
- The table on Slide 11 gives $F_{\alpha;v_1,v_2}$ for α = 0.10, 0.05, 0.01, and 0.001, and various values of v_1 (in different ***columns*** of the table) and v_2 (in different groups of ***rows*** of the table).

- For example, from the table on Slide 11, we obtain the following:

$$F_{.05,3,5} = 5.41$$
$$F_{.05,5,3} = 9.01.$$

9

Table entry for p is the critical value F^* with probability p lying to its right.



F critical values

Degrees of freedom in the denominator	p	1	2	3	4	5	6	7	8	9
		Degrees of freedom in the numerator								
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24

5.4 Completely Randomized Design (CRD)

A CRD is a completely random allocation of p treatments to n experimental units.

1. Experimental units (subjects) are assigned randomly to treatments
   - Subjects are assumed homogeneous

2. One factor or independent variable
   - 2 or more treatment levels or groups

3. Analyzed by one-way ANOVA

5.4.1 One-Way ANOVA F-test

- Tests the equality of 2 or more (p) population means: $\mu_1, \mu_2, \cdots, \mu_p$.

- Variables
  - One nominal independent variable
  - One continuous dependent variable

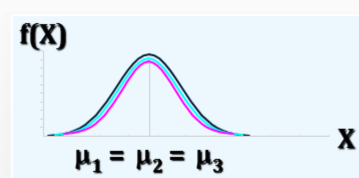
5.4.2 Assumptions for One-Way ANOVA

- Simple random samples
  - The samples taken from the populations under study are simple random samples
- Independent samples:
  - The samples taken from the populations under consideration are independent of one another.
- Normality
  - For each population, the variable under consideration is normally distributed.
- Constant variance
  - The variances of the variable under study are the same for all the populations

13

5.4.3 One-Way ANOVA F-Test Hypotheses

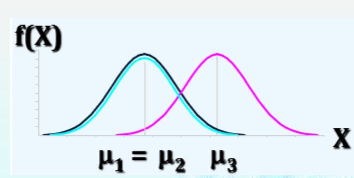
$H_0: \mu_1 = \mu_2 = \cdots = \mu_p$



In other words,

- H_0: All pop means are equal
- H_0: No treatment effect

$H_1: \mu_i \neq \mu_j, i \neq j, 1 \leq i, j \leq n$



In other words,

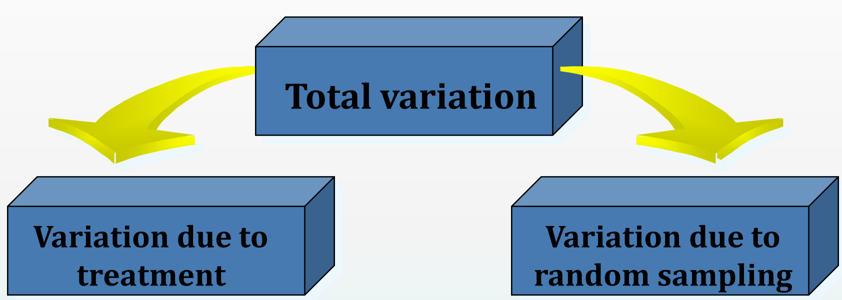
- H_1: At least 1 pop mean is different
- H_1: Treatment effect exists
- H_1: Not H_0

14

5.4.4 Basic Idea of One-way ANOVA

- Compares 2 types of variation: treatment variation and random variation, to test equality of means.
- If treatment variation is significantly greater than random variation, then the p means are not equal.
- Variation measures can be obtained by partitioning the total variation.

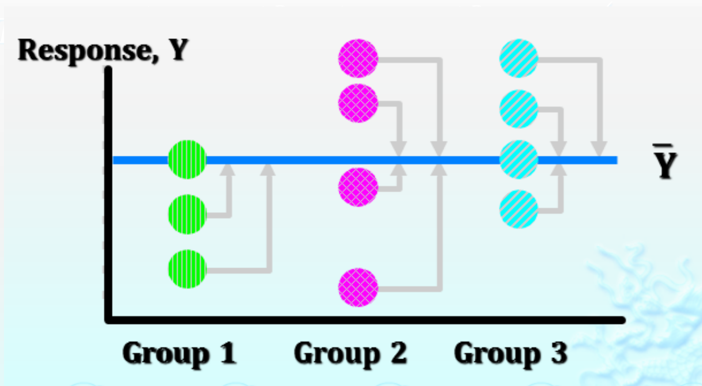
5.4.5 One-Way ANOVA: Partitions Total Variation



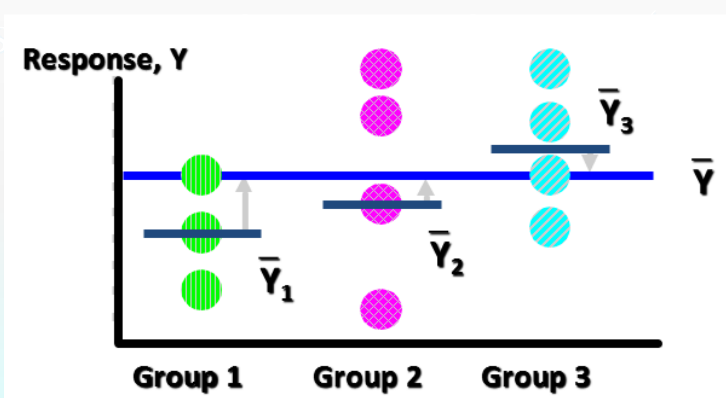
Sum of Squares Among
Sum of Squares Between
Sum of Squares Treatment (SSt)
Among Groups Variation

Sum of Squares Within
Sum of Squares Error (SSE)
Within Groups Variation

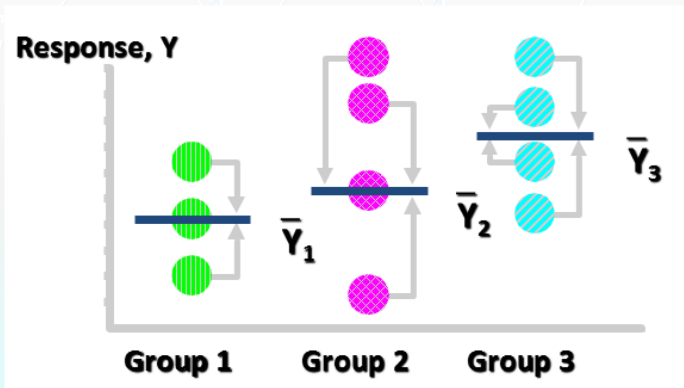
Total Variation



Treatment Variation



Random (Error) Variation



5.4.6 Typical Data Layout for a CRD

Treatment (Level)	Observation				Total	Average
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\overline{y_{1.}}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\overline{y_{2.}}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	
p	y_{p1}	y_{p2}	$\cdots$	y_{pn}	$y_{p.}$	$\overline{y_{p.}}$
					$y_{..}$	$\bar{y}_{..}$

where $y_{..} = grand\ total = \sum_{i=1}^{p}\sum_{j=1}^{n} y_{ij}$;

$\bar{y}_{..} = grand\ mean = \frac{y_{..}}{a \times p} = \frac{y_{..}}{N}$;

$N = total\ number\ of\ observations.$

5.4.7 Calculation of Sums of Squares

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$$

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{p} y_{i.}^2$$

$$SSt = SST - SSE$$

5.4.8 One-Way ANOVA Test Statistic

- Test statistic
  - $\boxed{F = \frac{MSt}{MSE} = \frac{SSt/(p-1)}{SSE/(n-p)}}$
    - MSt = mean square for treatment
    - MSE = mean square for error
- Degrees of freedom
  - $v_1 = p - 1$
  - $v_2 = n - p$
    - p = # of populations, groups, or levels
    - n = overall sample size

5.4.9 ANOVA Table

Source of Variation	Degree of Freedom	Sum of Squares	Mean Squares (Variance Estimate)	F-ratio F_0
Treatment	p - 1	SSt	$MSt = \frac{SSt}{p-1}$	$F_0 = \frac{MSt}{MSE}$
Error	n - p	SSE	$MSE = \frac{SSE}{n-p}$	-
Total	n - 1	SST	-	-

23

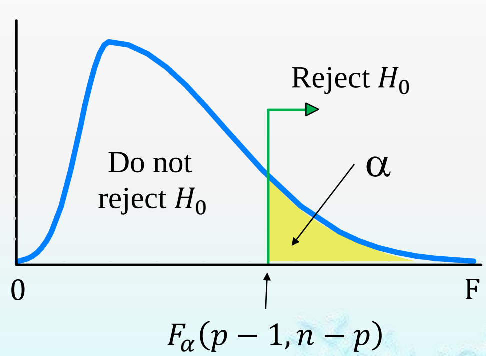
5.4.10 One-Way ANOVA F-Test Critical Value

If all the means are equal, then

$$F = \frac{MSt}{MSE} \approx 1.$$

Decision: do not reject H_0.

We reject H_0 only for large values of F.



always a one-tailed test!

24

Example 5.1

A vet epidemiologist wants to see if 3 food supplements have different mean milk yields. He assigns 15 cows randomly to the 3 supplements and 5 cows per supplement. At the 5% level, test if there is a difference in mean yields.

Food Supplement 1	Food Supplement 2	Food Supplement 3
25.40	23.40	20.00
26.31	21.80	22.20
24.10	23.50	19.75
23.74	22.75	20.60
25.10	21.60	20.40

25

Solution

	Food Supplement $i = 1$	Food Supplement $i = 2$	Food Supplement $i = 3$
j=1	y_{11} =25.40	y_{21} =23.40	y_{31} =20.00
j=2	y_{12} =26.31	y_{22} =21.80	y_{32} =22.20
j=3	y_{13} =24.10	y_{23} =23.50	y_{33} =19.75
j=4	y_{14} =23.74	y_{24} =22.75	y_{34} =20.60
j=5	y_{15} =25.10	y_{25} =21.60	y_{35} =20.40
	$y_{1.}$ =124.65 (summing up all j's from j=1 to j=5)	$y_{2.}$ = 113.05 (summing up all j's from j=1 to j=5)	$y_{3.}$ = 102.95 (summing up all j's from j=1 to j=5)

26

Numerical Calculation of SST

CASIO fx-50FH/CASIO fx-50FHII	CASIO fx-2650p/CASIO fx-3950p
1) Enter Statistical mode (SD mode): mode mode **4** → SD mode	1) Enter Statistical mode (SD mode): mode mode **1** → SD mode
2) Data input: 25.4 M+ 26.31 M+ to 20.6 M+ 20.4 M+	2) Data input: 25.4 M+ 26.31 M+ to 20.6 M+ 20.4 M+
SST: shift 2 2 ANS x^2 $\times$ 15 $=$ SST = 58.2172	SST: shift 2 2 ANS x^2 $\times$ 15 $=$ SST = 58.2172

27

Numerical Calculation of SSE

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n}\sum_{i=1}^{p} y_{i.}^2$$

1st part: " $\sum_{i=1}^{p}\sum_{j=1}^{n} y_{ij}^2$ = 7794.3787" can be obtained as follows:

CASIO scientific calculator: shift 1 =

2nd part (refer to Slide 27 for $y_{1.}, y_{2.}$ *and* $y_{3.}$):

$\frac{1}{n}\sum_{i=1}^{p} y_{i.}^2 = \frac{1}{5}(124.65^2 + 113.05^2 + 102.95^2)$ =7783.3255

SSE = 7794.3787 - 7783.3255 = 11.0532

28

Numerical Calculation of SS(treatment)

$\because SST = SSt + SSE$

$\therefore SSt = SST - SSE = 58.2172 - 11.0532$
$= 47.1640$

Solution

ANOVA Table

SV	DF	SS	MS (Variance Estimate)	F-ratio F_0
Treatment (Food)	3-1 = 2	47.1640	$\frac{47.1640}{2} = 23.5820$	25.60
Residual/ Error	15-3 = 12	11.0532	$\frac{11.0532}{12} = 0.9211$	-
Total	15-1 = 14	58.2172	-	-

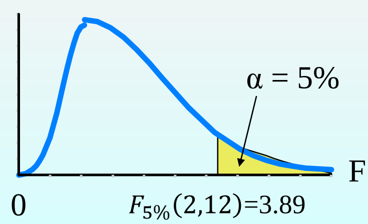
Solution

$H_0: \mu_1 = \mu_2 = \mu_3 \leftrightarrow H_1$: Not H_0.
$\alpha = 5\%$,

$v_1 = 3 - 1 = 2$,

$v_2 = 15 - 3 = 12$.

Critical Value:



Test Statistic:

$$F = \frac{MSt}{MSE} = \frac{23.5820}{0.9211} = 25.6$$

$\because$ 25.6 > 3.89,

$\therefore$ reject H_0 at the 5% level

Conclusion:

There is evidence

that the population means are

different.

Exercise 5.1

1. The Energy Information Administration gathers data on residential energy consumption and expenditures. The data are shown below:

Northeast	Midwest	South	West
13	15	5	8
8	10	11	10
11	16	9	6
12	11	5	5
11	13	7	7

Test, at the 5% level of significance, if a difference exists in mean annual energy consumption by households among the 4 U.S. regions.

Solution

Let μ_1, μ_2, μ_3 and μ_4 be the last years' mean energy consumptions by households in the 4 regions , respectively.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 \leftrightarrow \quad H_1$: $not\ H_0$

The row totals, $y_{i.}$'s are calculated as follows:

Northeast	Midwest	South	West
13	15	5	8
8	10	11	10
11	16	9	6
12	11	5	5
11	13	7	7
$y_{1.} = 55$	$y_{2.} = 65$	$y_{3.} = 37$	$y_{4.} = 36$

Solution (Cont'd)

SST = 202.55 (For data input data, refer to Slide 28.)

$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n} \sum_{i=1}^{p} y_{i.}^2$

$\sum_{i=1}^{p} \sum_{j=1}^{n} y_{ij}^2$ = 2065

$\frac{1}{n} \sum_{i=1}^{p} y_{i.}^2 = \frac{1}{5}(55^2 + 65^2 + 37^2 + 36^2) = 1983$

SSE = 2065 – 1983 = 82

SSt = SST – SSE = 202.55 – 82 = 120.55

Solution (Cont'd)

ANOVA Table

SV	DF	SS	MS (Variance Estimate)	F-ratio F_0
Treatment (Region)	4-1 = 3	120.55	$\frac{120.55}{3} = 40.1833$	$\frac{40.1833}{5.125} = 7.84$
Residual/ Error	20-4 = 16	82	$\frac{82}{16} = 5.125$	-
Total	20-1 = 19	202.55	-	-

Take $\alpha = 5\%$. $F_{5\%;3,16} = 3.24 < F_0$.

Reject H_0 at the 5% level and conclude that a difference exists in mean annual energy consumption by households among the 4 regions.

35

2. A researcher analyzed bacteria-culture data. Five strains of cultured bacteria that caused staph infections were observed for 24 hours at 27°C. The following table reports bacteria counts, in millions, for different cases from each of the 5 strains.

Strain A	Strain B	Strain C	Strain D	Strain E
9	3	10	14	33
27	32	47	18	43
22	37	50	17	28
30	45	52	29	59
16	12	26	20	31

At the 5% significance level, do the data provide sufficient evidence to conclude that a difference exists in mean bacteria counts among the 5 strains of bacteria?

36

Solution

Strain A	Strain B	Strain C	Strain D	Strain E
9	3	10	14	33
27	32	47	18	43
22	37	50	17	28
30	45	52	29	59
16	12	26	20	31
$y_{1.} = 104$	$y_{2.} = 129$	$y_{3.} = 185$	$y_{4.} = 98$	$y_{5.} = 194$

Let $\mu_i =$ the population mean bacteria counts of the i-th strain, $i = A, B, \ldots, E$.

H_0: $\mu_A = \mu_B = \cdots = \mu_E \quad \leftrightarrow \quad H_1: not\ H_0$

37

Solution (Cont'd)

SST = 5260 (For data input data, refer to Slide 28.)

$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n} y_{ij}^2 - \frac{1}{n} \sum_{i=1}^{p} y_{i.}^2$

$\sum_{i=1}^{p} \sum_{j=1}^{n} y_{ij}^2 = 25424$

$\frac{1}{n} \sum_{i=1}^{p} y_{i.}^2 = \frac{1}{5}(104^2 + 129^2 + 185^2 + 98^2 + 194^2) = 21784.4$

SSE = 25424 – 21784.4 =3639.6

SSt = SST – SSE = 5260 – 3639.6 =1620.4

38

Solution (Cont'd)

SV	DF	SS	MS (Variance Estimate)	F-ratio F_0
Treatment (Strain)	5-1 = 4	1620.4	$\frac{1620.4}{4}$ =405.1	$\frac{405.1}{181.98}$ =2.23
Residual/ Error	25-5 = 20	3639.6	$\frac{3639.6}{20} = 181.98$	-
Total	25-1 = 24	5260	-	-

Take $\alpha = 5\%. F_{5\%;4,20} = 2.87 > F_0$.

Do not reject H_0 at the 5% level and conclude that no difference exists in mean bacteria counts among the 5 strains of bacteria.

39

Part II

Chi Square Test

40

5.5 Parametric Statistics

- A major branch of statistics
  - Assuming that data follow a type of probability distribution (e.g. normal distribution)
  - Making inferences about the parameters of the distribution (e.g. population mean, population variance, etc.)
  - They are not distribution-free. In other words, they require a probability distribution.

41

5.6 Non-parametric Statistics

- They are also called ***distribution-free statistics***.
  - ***They do not rely on assumptions that the data are drawn from a given probability distribution*** (data model is not specified).
  - It was widely used for studying populations that take on a ranked order (e.g. movie reviews, opinions about hotel ranking on a 5-point Likert scale). They are appropriate for studying ***ordinal data***.
  - Data with ***frequencies or percentage***
    - # of kids in different grades
    - The % of people receiving social security

42

5.7 Chi Square (χ^2) Distribution

Definition

- If X_i ($i = 1,2,\ldots, k$) are k independent, normally distributed random variables with mean 0 and variance 1, then the random variable is distributed as a chi-square distribution with k degrees of freedom. This can be written as

$$X_i \sim N(0,1) \;\;\Rightarrow\;\; Q = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$$

- The chi-square distribution has one parameter: k - a positive integer that specifies the # of d.f. (i.e. the # of X_i)

5.7 χ^2 Distribution (Cont'd)

- The probability density function (pdf) of the chi-square distribution is given by

$$f(x;k) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma(k/2)}\, x^{(k/2)-1} e^{-x/2} & \text{for } x > 0, \\ 0 & \text{for } x \le 0, \end{cases}$$

where Γ denotes the Gamma function.

- The gamma function is defined by

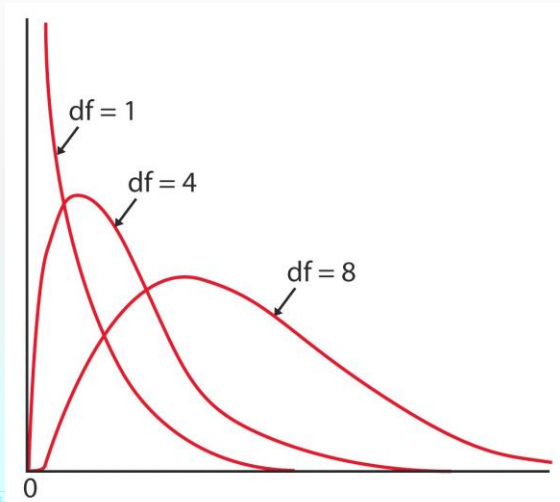
$$\Gamma(s) = \int_0^{+\infty} x^{s-1} e^{-s}\, dx$$

- If n is a positive integer, then

$$\Gamma(n+1) = n!$$

5.7 χ^2 Distribution (Cont'd)

The curves of chi-square distributions:



5.8 Features of a Chi-square Distributions

- Total area under a chi-square curve is equal to 1.
- It is never symmetric; Rather, it is ***skewed to the right.***
- The shape of the chi-square distribution depends on the degrees of freedom (just like t-distribution).
- As the # of df, *k* increases, the chi-square distribution becomes more symmetric.
- The ***values of χ^2 are nonnegative***, i.e. values of χ^2 are always ≥ 0. The χ^2 value increases to a peak and then asymptotically tends to 0.
- The table in the next slide gives a part of the critical values of chi-square distributions

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06
38	19.29	20.69	22.88	24.88	27.34	49.51	53.38	56.90	61.16
42	22.14	23.65	26.00	28.14	30.77	54.09	58.12	61.78	66.21
46	25.04	26.66	29.16	31.44	34.22	58.64	62.83	66.62	71.20
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15
55	31.73	33.57	36.40	38.96	42.06	68.80	73.31	77.38	82.29
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38
65	39.38	41.44	44.60	47.45	50.88	79.97	84.82	89.18	94.42
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.84	106.39
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33
85	55.17	57.63	61.39	64.75	68.78	102.08	107.52	112.39	118.24
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12
95	63.25	65.90	69.92	73.52	77.82	113.04	118.75	123.86	129.97
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81

5.9 Uses of Chi-Square Distribution

Goodness of Fit Test

How close are ***sample results ("observed frequencies")*** to the ***expected results ("expected frequencies")***?

<u>Example 5.2</u>

Around half number of heads (H) and half number of tails (T) are expected in tossing a coin. Assume that a coin is tossed 100 times. We expect 50 H's and 50T's (***expected frequencies***). The sample outcomes (***observed frequencies***) are 48 heads and 52 tails.

Can you arrive at the conclusion that the coin is ***fair or unbiased*** based on 48H's and 52 T's?

5.9 Uses of Chi-Square Distribution (Cont'd)

Test of Independence

Are 2 variables of interest (row variable and column variable) independent of each other?

Examples

- Are ***starting salaries (1st variable)*** of fresh graduates independent of graduates' ***fields of study (2nd variable)***?
- Is ***beer preference (1st variable)*** independent of the ***gender (2nd variable)*** of the beer drinker?

49

5.10 Chi-square Test

- One-sample chi-square test includes only ***1 dimension (One-way table)***.
  - Whether the # of respondents is ***uniformly/equally distributed*** across all levels of ***educational attainment (1 dimension)***.
  - Whether the # of absent employees in any one week has a ***uniform pattern*** across the working days of a week.
- Two-sample chi-square test includes ***2 dimensions (Two-way table)***.
  - Whether number of absences from work is independent of ***job position (1st dimension)*** and ***gender (2nd dimension).***

50

Chi-square Goodness-of-fit Test

51

5.11 Chi-Square Goodness-of-Fit Test

- A Chi-square goodness-of-fit test is used to test whether a ***frequency distribution*** fits an ***expected distribution***.
- The ***observed frequency O*** of a category is the frequency for the category ***observed*** in the sample data.
- The ***expected frequency E*** of a category is the ***calculated frequency*** for the category. Expected frequencies are obtained assuming the ***specified (or hypothesized) distribution***. The expected frequency for the i-th category is

$$E_i = np_i$$

where n is the number of trials (the sample size) and p_i is the assumed probability of the i-th category.

52

5.11.1 Observed and Expected Frequencies

Example 5.3
200 teenagers are randomly selected and asked what their favorite pizza topping is. The results are shown below:

Topping	Results (n = 200)	% of teenagers
Cheese	78	41%
Pepperoni	52	25%
Sausage	30	15%
Mushrooms	25	10%
Onions	15	9%

Find the observed frequencies and the expected frequencies.

Observed Frequency	Expected Frequency
78	200 × 41% = 82
52	200 × 25% = 50
30	200 × 15% = 30
25	200 × 10% = 20
15	200 × 9% = 18

5.11.2 Assumptions for χ^2 Goodness-of-fit Test

The following must be true:

1. The ***observed frequencies*** must be obtained by using a ***random sample***.
2. Each ***expected frequency*** must be **≥ 5**.

If the above assumptions are satisfied, then the ***sampling distribution*** for the goodness-of-fit test is ***approximated*** by a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of categories. The test statistic for the chi-square goodness-of-fit test is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

The test is always a right-tailed test.

where O represents the observed frequency of each category and E represents the expected frequency of each category.

5.11.3 Performing a χ^2 Goodness-of-fit Test

1. Identify the claim. State H_0 and H_1 .
2. Specify the level of significance, α .
3. Identify the degrees of freedom, k-1.
4. Determine the critical value, $\chi^2_{\alpha;k-1}$.
5. Determine the rejection region, $\chi^2 > \chi^2_{\alpha;k-1}$.

5.11.3 Performing a χ^2 Goodness-of-fit Test (Cont'd)

6. Calculate the test statistic, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.
7. Decide to reject or fail to reject H_0 .
   - If χ^2 is in the rejection region, reject H_0.
   - Otherwise, do not to reject H_0.
8. Draw conclusion based on the context of the original claim.

Example 5.4
A researcher claims that the distribution of favorite pizza toppings among teenagers is as shown below. 200 randomly selected teenagers are surveyed.

Topping	Frequency, f
Cheese	39%
Pepperoni	26%
Sausage	15%
Mushrooms	12.5%
Onions	7.5%

Using $\alpha = 0.01$, and the observed and expected values previously calculated, test the researcher's claim using a chi-square goodness-of-fit test.

57

Solution

H_0: The distribution of pizza toppings: 39% cheese, 26% pepperoni, 15% sausage, 12.5% mushrooms, and 7.5% onions.

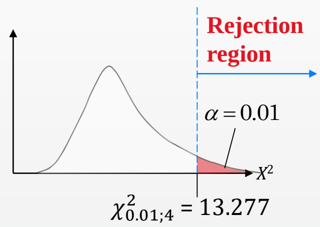
H_1: *not* H_0

Because there are 5 categories, the chi-square distribution has $k - 1 = 5 - 1 = 4$ df.

With d.f. = 4 and $\alpha = 0.01$, $\chi^2_{0.01;4} = 13.277$.

58

Solution (Cont'd)



Topping	Observed Frequency	Expected Frequency
Cheese	78	82
Pepperoni	52	50
Sausage	30	30
Mushrooms	25	20
Onions	15	18

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(78-82)^2}{82} + \frac{(52-50)^2}{50} + \frac{(30-30)^2}{30} + \frac{(25-20)^2}{20} + \frac{(15-18)^2}{18}$$

$$\approx 2.025$$

$\because 2.025 < 13.277$, $\therefore$ do not reject H_0.

There is not enough evidence at the 1% level to reject the researcher's claim.

59

Exercise 5.3

1. It is thought that each of the 8 outcomes of an experiment is equally likely to occur. The experiment is performed 400 times. The following table displays the observed frequencies:

Observed Frequency	45	42	55	53	40	62	47	56
Expected Frequency								

Perform a test at the 1% level to investigate the validity of the theory.

60

Solution

Observed Frequency	45	42	55	53	40	62	47	56
Expected Frequency	50	50	50	50	50	50	50	50

If the theory is correct, then each of the 8 outcomes must be

$$\frac{400}{8} = 50$$

$$\chi^2 = \frac{(45-50)^2}{50} + \frac{(42-50)^2}{50} + \frac{(55-50)^2}{50} + \frac{(53-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(62-50)^2}{50} + \frac{(47-50)^2}{50} + \frac{(56-50)^2}{50} = 9.04$$

d.f. = 8 – 1 = 7. At α = 1%, $\chi^2_{0.01;7}$ = 18.475

∵ 9.04 < 18.475,

∴ do not reject H_0 at the 1% level.

There is no sufficient evidence that the theory is correct.

Chi-square Test of Independence

5.12 Contingency Tables

An ***r* × *c* contingency table** shows the ***observed frequencies*** for 2 variables. The observed frequencies are arranged in ***r* rows and *c* columns**. The ***intersection*** of a ***row*** and a ***column*** is called a ***cell***.

The following contingency table shows a random sample of 321 fatally injured passenger vehicle drivers by age and gender.

	Age					
Gender	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 and older
Male	32	51	52	43	28	10
Female	13	22	33	21	10	6

5.12.1 Expected Frequency

Assuming the 2 variables are ***independent***, we can use the contingency table to find the ***expected frequency*** for each cell.

Finding the Expected Frequency for Contingency Table Cells

The expected frequency, E_{ij} for the cell located at *i*-th row and *j*-th column in a contingency table is

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n},$$

where n = the overall sample size;

$n_{i.}$ = *i*-th row total;

$n_{.j}$ = *j*-th column total.

Example 5.5

Find the expected frequency for each "Male" cell in the contingency table for the sample of 321 fatally injured drivers. Assume that the variables, age and gender, are independent.

	Age						
Gender	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 & older	Total
Male	32	51	52	43	28	10	216
Female	13	22	33	21	10	6	105
Total	45	73	85	64	38	16	321

Expected Frequencies for "Male"

	Age						
Gender	16 – 20 $j = 1$	21 – 30 $j = 2$	31 – 40 $j = 3$	41 – 50 $j = 4$	51 – 60 $j = 5$	61 and older, $j = 6$	Total
Male	32	51	52	43	28	10	$n_{1.}$ =216
Female	13	22	33	21	10	6	$n_{2.}$ =105
Total	$n_{.1}$ =45	$n_{.2}$ =73	$n_{3.}$ =85	$n_{.4}$ =64	$n_{.5}$ =38	$n_{.6}$ = 16	n =321

$$E_{11} = \frac{n_{1.} \times n_{.1}}{n} = \frac{216\times45}{321}=30.28$$

$$E_{14} = \frac{n_{1.} \times n_{.4}}{n} = \frac{216\times64}{321}=43.07$$

$$E_{12} = \frac{n_{1.} \times n_{.2}}{n} = \frac{216\times73}{321}=49.12$$

$$E_{15} = \frac{n_{1.} \times n_{.5}}{n} = \frac{216\times38}{321}=25.57$$

$$E_{13} = \frac{n_{1.} \times n_{.3}}{n} = \frac{216\times85}{321}=57.20$$

$$E_{16} = \frac{n_{1.} \times n_{.6}}{n} = \frac{216\times16}{321}=10.77$$

Expected Frequencies for "Female"

	Age						
Gender	16 – 20 $j = 1$	21 – 30 $j = 2$	31 – 40 $j = 3$	41 – 50 $j = 4$	51 – 60 $j = 5$	61 and older, $j = 6$	Total
Male	32	51	52	43	28	10	$n_{1.}$ =216
Female	13	22	33	21	10	6	$n_{2.}$ =105
Total	$n_{.1}$ =45	$n_{.2}$ =73	$n_{3.}$ =85	$n_{.4}$ =64	$n_{.5}$ =38	$n_{.6}$ = 16	n =321

$$E_{21} = \frac{n_{2.} \times n_{.1}}{n} = \frac{105 \times 45}{321} = 14.72 \qquad E_{24} = \frac{n_{2.} \times n_{.4}}{n} = \frac{105 \times 64}{321} = 20.93$$

$$E_{22} = \frac{n_{2.} \times n_{.2}}{n} = \frac{105 \times 73}{321} = 23.88 \qquad E_{25} = \frac{n_{2.} \times n_{.5}}{n} = \frac{105 \times 38}{321} = 12.43$$

$$E_{23} = \frac{n_{2.} \times n_{.3}}{n} = \frac{105 \times 85}{321} = 27.80 \qquad E_{26} = \frac{n_{2.} \times n_{.6}}{n} = \frac{105 \times 16}{321} = 5.23$$

67

5.12.2 Chi-square Independence Test

A ***chi-square independence test*** is used to test the ***independence*** of 2 variables. Using a chi-square test, we can determine if the occurrence of ***one variable affects the probability of the occurrence of the other variable***.

For the chi-square independence test to be used, the following ***conditions*** must be true.

1. The ***observed frequencies*** must be obtained by using a ***random sample***.
2. Each ***expected frequency*** must be ≥ 5.

68

5.12.2 Chi-square Independence Test

If the above conditions are satisfied, then the sampling distribution for the chi-square independence test is ***approximated by a chi-square distribution*** with degrees of freedom:

$$(r-1)(c-1)$$

where r and c are the number of rows and columns, respectively, of a contingency table. The test statistic for the chi-square independence test is

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(r-1)(c-1)$$

The test is always a right-tailed test.

where O represents the observed frequencies and E represents the expected frequencies.

69

5.12.3 Performing a χ^2 Independence Test

1. Identify the claim. State H_0 and H_1.
2. Specify the level of significance, α.
3. Identify the degrees of freedom, $(r-1)(c-1)$.
4. Determine the critical value, $\chi^2_{\alpha;(r-1)(c-1)}$.
5. Determine the rejection region, $\chi^2 > \chi^2_{\alpha;(r-1)(c-1)}$.

70

5.12.3 Performing a χ^2 Independence Test (Cont'd)

6. Calculate the test statistic, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.

7. Decide to reject or not to reject H_0.
   - If χ^2 is in the rejection region, reject H_0.
   - Otherwise, fail to reject H_0.

8. Draw conclusion based on the context of the original claim.

71

Simplified Chi-square Test Formula

Instead of using $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$, we can employ the following simplified formula:

$$\chi^2_{Calc} = n\left(\sum \frac{O_i^2}{n_{i.}n_{.j}} - 1\right)$$

When using the above formula, we do not need to calculate expected frequencies.

72

Example 5.6

In order to investigate whether the distribution of blood types in Europe is the same as in the United States, information was collected on 200 randomly picked people in Europe and 300 randomly picked people in the U.S. From the data, which are summarized below, is it true that the distributions of blood types in Europe and the U.S. are significantly different?

Blood Type	LOCATION		Total
	Europe	U.S.	
A	95	125	220
B	50	70	120
O	45	90	135
AB	10	15	25
Total	200	300	500

73

Solution

H_0 : The distribution of blood types in Europe is the same as that in the U.S.

versus

H_1 : Not H_0.

Blood Type	LOCATION		Total
	Europe	U.S.	
A	95	125	220
B	50	70	120
O	45	90	135
AB	10	15	25
Total	200	300	500

$$\chi^2_{Calc} = n\left(\sum \frac{O_i^2}{n_{i.}n_{.j}} - 1\right) = 500\left(\frac{95^2}{220\times 200} + \frac{50^2}{120\times 200} + \cdots + \frac{15^2}{25\times 300} - 1\right) = 3.57$$

$$\chi^2_{calc} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(95-88)^2}{88} + \frac{(50-48)^2}{48} + \Lambda + \frac{(15-15)^2}{15}$$

$$= 3.56$$

74

Solution (Cont'd)

$$df = (r-1)(c-1) = (4-1)(2-1) = 3.$$
$$\alpha = 5\% \Rightarrow \chi^2(3; 0.05) = 7.815.$$

$$\Theta\ \chi^2_{calc} = 3.56 < 7.815,$$
$$\therefore \text{ we do not reject } H_0.$$

Conclusion:

There is no reason to believe that the distribution of blood types in Europe is any different from the distribution in the U.S.

75

5.12.4 Yate's Continuity Correction for 2×2 Contingency Table

- The way it is typically used – compare to critical value - is based on ***large sample theory***. However, we may not always have large samples.
- Yate's correction is a correction of Pearson chi-square that adjusts the chi-square for ***small samples*** or ***df =1***.
- In practice, there is no universal agreement as to whether this adjustment should be used.
- The literature seems to indicate that the ***correction for continuity*** should be used when there are ***two rows and two columns.***
- Yate's continuity correction is only applied to ***2×2 tables***.

76

5.12.4 Yate's Continuity Correction for 2×2 Contingency Table

Test statistic formula for Yate's continuity correction is given by

$$\chi^2_{calc} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(|O_{ij} - E_{ij}| - \frac{1}{2}\right)^2}{E_{ij}} = \sum_{all\ cells} \frac{\left(|O_{ij} - E_{ij}| - \frac{1}{2}\right)^2}{E_{ij}}$$

Or

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

77

5.12.5 When to Use Yate's Continuity Correction?

- When df = 1 for a 2×2 contingency table, where

  df = $(r-1)(c-1) = (2-1)(2-1) = 1$

- The most conservative recommendation says that ***all expected frequency counts, E_{ij}*** should be ≥ 5.

- Therefore, when $1 \le E_{ij} \le 5$, we need to apply Yate's continuity correction.

78

Example 5.7

The following table shows the result of treating a certain disease using medicines A and B. The numbers in brackets are the expected frequencies calculated using the following formula:

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n}, \quad i = 1, 2; j = 1, 2.$$

Medicine	Effective	Non-effective	Total
A	40 (36.75)	2 (5.25)	42
B	16 (19.25)	6 (2.75)	22
Total	56	8	64

Compute the value of the chi-square test statistic.

79

Solution (Cont'd)

$\because$ $E_{22} = 2.75 < 5$,

$\therefore$ we need to use Yate's continuity correction.

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

$$= \frac{64\left(|40 \times 6 - 16 \times 2| - \frac{64}{2}\right)^2}{56 \times 8 \times 22 \times 42} = 4.789$$

Medicine	Effective	Non-effective	Total
A	$a = 40$	$b = 2$	42
B	$c = 16$	$d = 6$	22
Total	56	8	$n = 64$

80

Example 5.8

The following contingency table shows a ***random*** sample of 321 fatally injured passenger vehicle drivers by age and gender. The expected frequencies are displayed in parentheses. At $\alpha = 0.05$, can you conclude that the drivers' ages are related to gender in such accidents?

	Age						
Gender	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 and older	Total
Male	32 (30.28)	51 (49.12)	52 (57.20)	43 (43.07)	28 (25.57)	10 (10.77)	216
Female	13 (14.72)	22 (23.88)	33 (27.80)	21 (20.93)	10 (12.43)	6 (5.23)	105
	45	73	85	64	38	16	321

Solution

Because each expected frequency is ≥ 5 and the drivers were randomly selected, the chi-square independence test can be used to test whether the variables are independent.

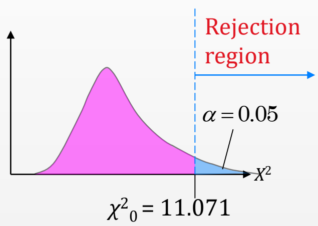
H_0: The drivers' ages are independent of gender.

H_1: The drivers' ages are dependent on gender.

d.f. = $(r - 1)(c - 1) = (2 - 1)(6 - 1) = (1)(5) = 5$

With d.f. = 5 and $\alpha = 0.05$, $\chi^2_{0.05;5} = 11.071$.

Solution



$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.84$$

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
32	30.28	1.72	2.9584	0.0977
51	49.12	1.88	3.5344	0.072
52	57.20	−5.2	27.04	0.4727
43	43.07	−0.07	0.0049	0.0001
28	25.57	2.43	5.9049	0.2309
10	10.77	−0.77	0.5929	0.0551
13	14.72	−1.72	2.9584	0.201
22	23.88	−1.88	3.5344	0.148
33	27.80	5.2	27.04	0.9727
21	20.93	0.07	0.0049	0.0002
10	12.43	−2.43	5.9049	0.4751
6	5.23	0.77	0.5929	0.1134

$\because 2.84 < 11.071$,

$\therefore$ do not reject H_0 at the 5 % level.

There is not enough evidence at the 5% level to conclude that age is dependent on gender in such accidents.

83

Exercise 5.4

1. A driving school examined the results of 100 candidates who were taking driving test for the first time. They found that, of the 40 men, 28 passed and out of the 60 women, 34 passed.

(a) Construct a contingency table summarizing the information in the above question.

(b) Compute the expected frequencies.

(c) Write down the hypotheses to be tested.

(d) Determine the critical value at the 5% level. Hence, write down the decision rule.

(e) Compute the test statistic.

(f) Make decision and draw conclusion.

84

(a)

		Result of first-time Candidates		
		Pass	Fail	Total
Gender	Male	28 (24.8)	12 (15.2)	40
	Female	34 (37.2)	26 (22.8)	60
	Total	62	38	100

(b) The expected frequencies are in brackets of the above 2×2 contingency table.

(c) H_0: Candidate's gender and the ability to pass in the first time is independent.
H_1: not H_0.

(d) df = $(2-1)(2-1)$=1, $\chi^2_{5\%;1} = 3.841$. The decision rule is: $\chi^2 > 3.841$.

(e) As df =1, we use Yate's continuity correction when calculating test statistic.

$$\chi^2 = \frac{(|28-24.8|-0.5)^2}{24.8} + \frac{(|12-15.2|-0.5)^2}{15.2} + \frac{(|34-37.2|-0.5)^2}{37.2} + \frac{(|26-22.8|-0.5)^2}{22.8} = 1.29$$

85

Solution

(f) $\because$ 1.29 < 3.841,
$\therefore$ do not reject H_0 at the 5% level.
Therefore, gender of candidate and the ability to pass in the first time are independent.

Alternative Method to Compute Test Statistic:

$$\chi^2_{Calc} = \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{100\left(|28\times 26 - 12\times 34| - \frac{100}{2}\right)^2}{(28+12)(34+26)(28+34)(12+26)} = 1.29$$

		Result of first-time Candidates		
		Pass	Fail	Total
Gender	Male	a = 28	b = 12	40
	Female	c = 34	d = 26	60
	Total	62	38	100

86