Chapter 11

Chi-Square Tests

Contingency Tables

- Useful in situations comparing multiple population proportions
- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.

Textbook : P410 : Section 11.1 Paragraph 1&2 / Table 11.1

The properties of the chi square distribution.

- Is continuous distribution.
- Positive skewed curve (skewed to the right curve).
- It is not symmetric curve.

Type of χ^2 tests :

1. Chi-Square Test for the difference between two proportions.
2. Chi-Square Test for the difference among more than two proportions.
3. Chi-Square Test of independence.

Chi-Square Test for the difference between two proportions.

Step (1): State the null and alternate hypotheses :
H_0: The two proportions should be the same
H_1: The two proportions should not be the same

Step (2): Select the level of significance (α)
Step (3): The test statistic

$$\boxed{\chi^2_{STAT} = \sum_{all\ cells} \frac{(f_o - f_e)^2}{f_e}}$$

- where:
  f_o : observed frequency in a particular cell .
  f_e : expected frequency in a particular cell if H_0 is true

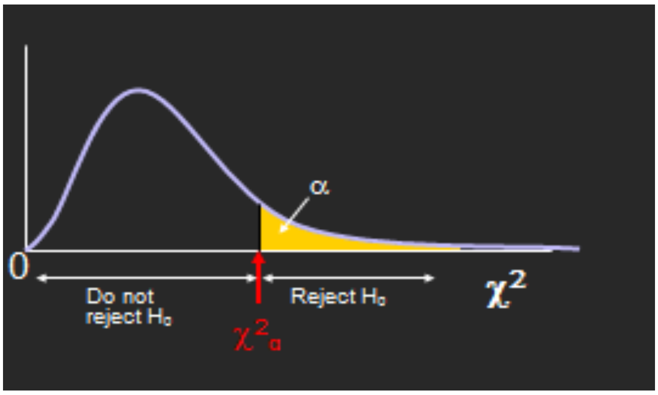
(Assumed: each cell in the contingency table has expected frequency of at least 5)
"Slide 7"

$$f_e = \frac{rowtotal \times columntotal}{n}$$ **or**

$\bar{P} *$ Observed frequency in a particular cell .

$$\bar{P}: \text{the overall } Porportion = \frac{X_1 + X_2}{n_1 + n_2}$$

(Slide 9)
Step (4): The critical value (Textbook : P547 : χ^2 table)
df = (r-1)(c-1) = (2-1)(2-1) =1
$\chi^2_{(\alpha,1)}$
(Slide 7)
Step (5) : The χ^2_{Stat} test statistic approximately follows a chi-squared distribution
Reject H_o If $\chi^2_{Stat} > \chi^2_{(\alpha,1)}$ otherwise, do not reject H_0



Example (1): (Slide 8)
Suppose we examine a sample of 300 children Left-Handed vs. Gender. Sample results organized in a contingency table: "Slide 4-9"

Gender	Hand Preference		
	Left	Right	Total
Female	12	108	120
Male	24	156	180
Total	36	264	300

Test where the Proportion of females who are left handed is equal to the proportion of males who are left handed (α=0.05).

Solution :

Step (1):
$H_0: \pi_1 = \pi_2$ (Proportion of females who are left handed is equal to the proportion of males who are left handed)
$H_1: \pi_1 \neq \pi_2$ (The two proportions are not the same)
Step (2): The level of significance (α=0.05).

$$\chi^2_{(2-1)(2-1),0.05} = \chi^2_{1\times1,0.05} = \chi^2_{1,0.05} = 3.841$$

Step (3): The test statistic

Gender	Hand Preference		
	Left	Right	Total
Female	12 / 14.4	108 / 105.6	120
Male	24 / 21.6	156 / 158.4	180
Total	36	264	300

$$f_e = \frac{rowtotal \times columntotal}{n}$$

$f_{row,column} = f_{r,c}$

For example: $f_{11} = \frac{120\times36}{300} = 14.4$, $f_{12} = \frac{120\times264}{300} = 105.6$,

$f_{21} = \frac{180\times36}{300} = 21.6$,$f_{22} = \frac{180\times264}{300} = 158.4$

Totals for the observaed and expected frequencies are the same

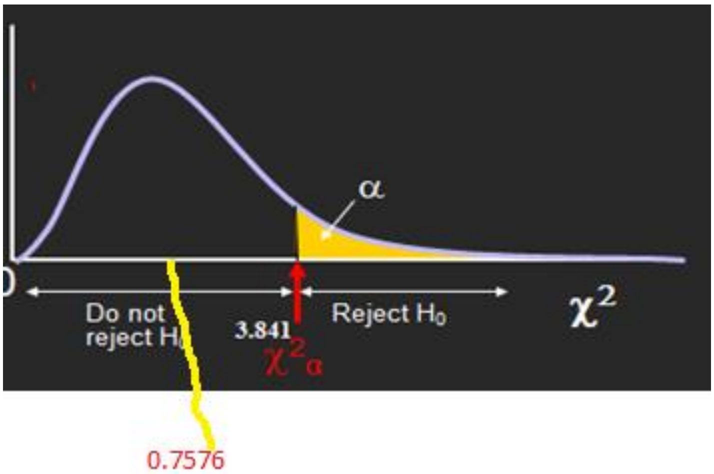
Totals for the observaed frequencies =12+108+24+156=300

Totals for the expected frequencies =14.4+105.6+21.6+158.4=300

$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = 0.7576$$

Step (4): Rule: If $\chi^2_{stat>3.841}$, Reject H_0, otherwise, do not reject H_0

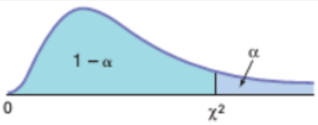


Step (5): Decision: Do not reject H_0

So we do not reject H_0 and conclude that there is insufficient evidence that the two proportions are different at $\alpha = 0.05$

Critical Values of χ^2

For a particular number of degrees of freedom, entry represents the critical value of χ^2 corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).



Degrees of Freedom	Cumulative Probabilities											
	0.005	0.01	0.025	0.05	0.10	0.25	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas (α)											
	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548

Example (2): (Textbook : P411-415 / Example11.1)

The following table is the contingency table for the hotel guest satisfaction study. The contingency table has two rows, indicating whether the guest would return to the hotel or would not return to the hotel, and two columns, one for each hotel. The cells in the table indicate the frequency of each row-and-column combination. The row totals indicate the number of quests who would return to the hotel and the number of guests who would not return to the hotel. The column totals are the sample sizes for each hotel location.

Choose Hotel Again	Hotel		
	Beachcomber	Windsurfer	Total
Yes	163	154	317
No	64	108	172
Total	227	262	489

Test where the Proportion of guests who would return Beachcomber, π_1 , is equal to the population proportion of guests who would return to the Windsurfer, π_2 ,you can use the Chi-Square test for the difference between two proportion.(α=0.05)

Solution

Step (1):

H_0: $\pi_1 = \pi_2$ (There is no difference between the two population Proportion)

H_1: $\pi_1 \neq \pi_2$ (The population Proportion are not the same)

If H_0 is true, there is no difference between the proportions of guests who are likely to choose either of these hotels again

Step (2): level of significance (α=0.05)

$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{(2-1)(2-1),0.05} = \chi^2_{1\times1,0.05} = \chi^2_{1,0.05} = 3.841$

Step (3): The test statistic

Choose Hotel Again	Hotel		
	Beachcomber	Windsurfer	Total
Yes	163 147.16	154 169.84	317
No	64 79.84	108 92.16	172
Total	227	262	489

$$f_e = \frac{row total \times column total}{n}$$

$f_{row,column} = f_{r,c}$

For example: $f_{11} = \frac{317\times227}{489} = 147.16$, $f_{12} = \frac{317\times262}{489} = 169.84$,

$f_{21} = \frac{172\times227}{489} = 79.84$,$f_{22} = \frac{172\times262}{489} = 92.16$

Totals for the observaed and expected frequencies are the same

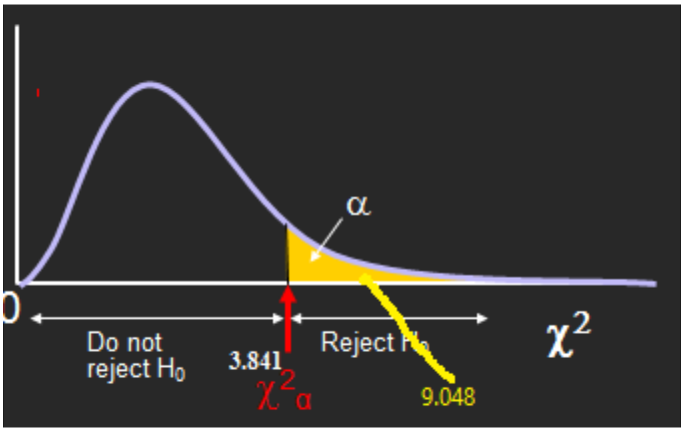
Totals for the observaed frequencies =163+154+64+108=489

Totals for the expected frequencies =147.16+169.84+79.84+92.16=489

$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(163-147.16)^2}{147.16} + \frac{(154-169.84)^2}{169.84} + \frac{(64-79.84)^2}{79.84} + \frac{(108-92.16)^2}{92.16} = 9.048$$

Step (4): Decision Rule: If $\chi^2_{stat>3.841}$ Reject H_0

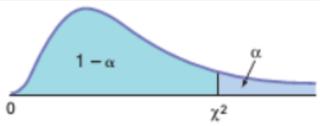


Step (5): Decision: Reject H_0

Since the test statistic is greater than the critical value, there is sufficient evidence to conclude there is a significant difference between the proportions of guests who would return to Beachcomber is different from the proportion of guests who would return to the Windsurfer

Critical Values of χ^2

For a particular number of degrees of freedom, entry represents the critical value of χ^2 corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).



	Cumulative Probabilities											
	0.005	0.01	0.025	0.05	0.10	0.25	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas (α)											
Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548

Chi-Square Test for the difference among more than two proportions.

Step (1): State the null and alternate hypotheses :

H_0: The proportions should be the same

H_1: The proportions should not be the same

Step (2): Select the level of significance (α)

Step (3): The test statistic

$$\chi^2_{STAT} = \sum_{all\ \text{cells}} \frac{(f_o - f_e)^2}{f_e}$$

- where:
  f_o = observed frequency in a particular cell of the 2*c table.
  f_e = expected frequency in a particular cell if H_0 is true

(Assumed: each cell in the contingency table has expected frequency of at least 1)

"Slide 15"

$$f_e = \frac{rowtotal \times columntotal}{n}$$

Or

$$\bar{P} * \text{Observed frequency in a particular cell}\,.$$

$$\bar{P}\text{: the overall } Porportion = \frac{X_1 + X_2 + \cdots \ldots + X_c}{n_1 + n_2 + \cdots \ldots . . + n_c} = \frac{X}{n}$$

"Slide 16"

Step (4): The critical value: df = (r-1)(c-1) = (2-1)(c-1) = (c-1)

$\chi^2_{(\alpha,(c-1))}$

Step (5) : The χ^2_{Stat} test statistic approximately follows a chi-squared distribution

Reject H_o If $\chi^2_{\text{Stat}} > \chi^2_{(\alpha,(c-1))}$ otherwise, do not reject H_0

Example (3):

Most companies consider big data analytics critical to success .However, is there a difference among small (<100 employees), mid-size (100-999 employees), and large (1000+ employees) companies in the proportion of companies that have already deployed big data project? A study showed the results for the different company size.

(Data extracted from 2014 big data outlook:big data is transformative –where is your company?)

Have already deployed big data projects	Company Size			
	Small	Mid-sized	large	Total
Ye	18	74	52	144
No	182	126	148	456
Total	200	200	200	600

Assume that 200 decision makers involved in big data purchases within each company size were surveyed. At <u>the 0.05 level of significance, is there evidence of a difference</u> among companies of different sizes with respect to the proportion of companies that have already deployed big data projects?

Solution

Step (1):

$H_0: \pi_1 = \pi_2 = \pi_3$

H_1:At least one proportion differs where $\pi_1 = \text{small}, \pi_2 = \text{medium}, \pi_3 = \text{large}$

If H_0 is true, there is no difference between the three proportions.

Step (2): The level of significance (α=0.05)

$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{(2-1)(3-1),0.05}=\chi^2_{1\times2,0.05}=\chi^2_{2,0.05}=5.991$

Step (3):

Have already deployed big data projects	Company Size			
	Small	Mid-sized	large	Total
Ye	18 / 48	74 / 48	52 / 48	144
No	182 / 152	126 / 152	148 / 152	456
Total	200	200	200	600

where:

f_o = observed frequency in a particular cell

f_e = expected frequency in a particular cell if H_0 is true

$$f_e = \frac{row total \times column total}{n}$$

$f_{row,column} = f_{r,c}$

For example: $f_{11} = \frac{144\times200}{600} = 48$, $f_{12} = \frac{144\times200}{600} = 48$, $f_{13} = \frac{144\times200}{600} = 48$,

$f_{21} = \frac{456\times200}{600} = 152$, $f_{22} = \frac{456\times200}{600} = 152$, $f_{23} = \frac{456\times200}{600} = 152$

Totals for the observaed and expected frequencies are the same

Totals for the observaed frequencies =18+74+52+182+126+148=600

Totals for the expected frequencies =48+48+48+152+152+152=600

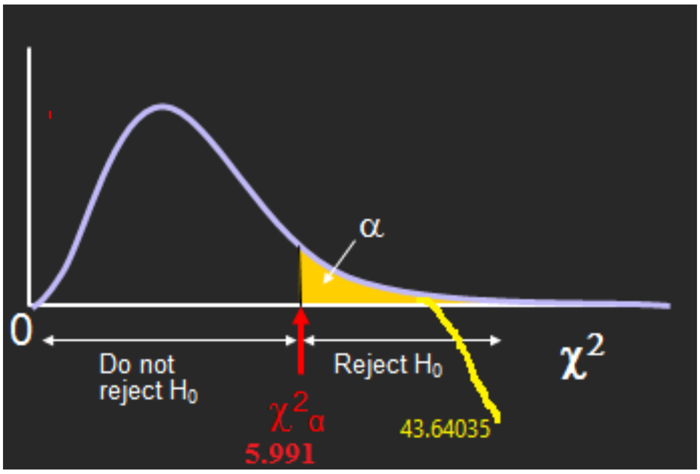
$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(18-48)^2}{48} + \frac{(74-48)^2}{48} + \frac{(52-48)^2}{48} + \frac{(182-152)^2}{152} + \frac{(126-152)^2}{152} + \frac{(148-152)^2}{152} = 43.64035$$

Step (4): Decision Rule: If $\chi^2_{stat>5.991}$ Reject H_0, otherwise, do not reject H0

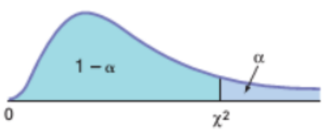
Step (5):

Since $\chi^2_{stat} = 43.64035$ is greater than the upper critical value of 5.991, reject H0 .There is evidence among the groups with respect to the proportion of companies that have already deployed big data projects



Critical Values of χ^2

For a particular number of degrees of freedom, entry represents the critical value of χ^2 corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).



	Cumulative Probabilities											
	0.005	0.01	0.025	0.05	0.10	0.25	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas (α)											
Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750

Chi-Square Test of independent.

Similar to the χ^2 test for equality of more than two proportions, but extends the concept to contingency tables with r rows and c columns (Slide 17)

The test is applied when you have two categorical variables from a single population, and it is used to determine whether there is a significant association between the two variables.

Step (1): State the null and alternate hypotheses :

H_0: The two categorical variables are independent

(i.e., there is no relationship between them)

H_1: The two categorical variables are dependent

(i.e., there is a relationship between them)

Step (2): Select the level of significance (α)

Step (3): The test statistic

$$\boxed{\chi^2_{STAT} = \sum_{all\ \text{cells}} \frac{(f_o - f_e)^2}{f_e}}$$

- where:
  f_o = observed frequency in a particular cell of th r*c table
  f_e = expected frequency in a particular cell if H_0 is true

(Assumed: each cell in the contingency table has expected frequency of at least 1)
"Slide 18"

$$f_e = \frac{row\,total \times column\,total}{n}$$

Step (4): The critical value:

df = (r-1)(c-1) $\chi^2_{(\alpha,((r-1)(c-1))}$

Step (5) : The χ^2_{Stat} test statistic approximately follows a chi-squared distribution

Reject H_o If $\chi^2_{Stat} > \chi^2_{(\alpha,((r-1)(c-1))}$

Example (4): Slide(21-25)
The meal plan selected by 200 students is shown below:

Class Standing	Number of meals per week			Total
	20/week	10/week	none	
Fresh.	24	32	14	70
Soph.	22	26	12	60
Junior	10	14	6	30
Senior	14	16	10	40
Total	70	88	42	200

At the 0.05 level of significance, is there evidence that meal plan and class standing are independent(i.e., there is no relationship between them)

Solution

Step (1):

H_0: Meal plan and class standing are independent
(i.e., there is no relationship between them)

H_1: Meal plan and class standing are dependent
(i.e., there is a relationship between them)

Step (2): The level of significance (α =0.05)

$\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{(4-1)(3-1),0.05} = \chi^2_{3\times2,0.05} = \chi^2_{6,0.05} = 12.592$

Degree of freedom= (4-1)(3-1)=6

Step (3):

Class Standing	Number of meals per week			Total
	20/week	10/week	none	
Fresh.	24 24.5	32 30.8	14 14.7	70
Soph.	22 21	26 26.4	12 12.6	60
Junior	10 10.5	14 13.2	6 6.3	30
Senior	14 14	16 17.6	10 8.4	40
Total	70	88	42	200

where:

f_o = observed frequency in a particular cell

f_e = expected frequency in a particular cell if H_0 is true

$$f_e = \frac{rowtotal \times columntotal}{n}$$

$$f_{row,column} = f_{r,c}$$

For example: $f_{11} = \frac{70\times70}{200} = 24.5$, $f_{12} = \frac{70\times88}{200} = 30.8$, $f_{13} = \frac{70\times42}{200} = 14.7$,

$f_{21} = \frac{60\times70}{200} = 21$, $f_{22} = \frac{60\times88}{200} = 26.4$, $f_{23} = \frac{60\times42}{200} = 12.6$,

$f_{31} = \frac{30\times70}{200} = 10.5$, $f_{32} = \frac{30\times88}{200} = 13.2$, $f_{33} = \frac{30\times42}{200} = 6.3$,

$f_{41} = \frac{40\times70}{200} = 14$, $f_{42} = \frac{40\times88}{200} = 17.6$, $f_{43} = \frac{40\times42}{200} = 8.4$,

Totals for the observaed and expected frequencies are the same

Totals for the observaed frequencies =24+32+14+22+26+12+10+14+6+14+16+10=200

Totals for the expected frequencies

=24.5+30.8+14.7+21+26.4+12.6+10.5+13.2+6.3+14+17.6+8.4=200

$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(24-24{,}5)^2}{24.5} + \frac{(32-30.8)^2}{30.8} + \frac{(14-14.8)^2}{14.8} + \frac{(22-21)^2}{21} + \frac{(26-26.4)^2}{26.4} + \frac{(12-12.6)^2}{12.6}$$

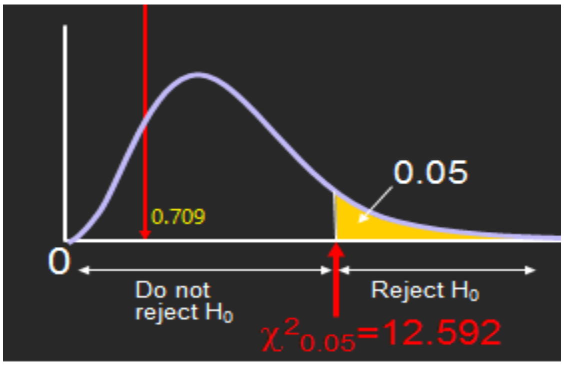
$$+ \frac{(10-10.5)^2}{10.5} + \frac{(14-13.2)^2}{13.2} + \frac{(6-6.3)^2}{6.3} + \frac{(14-14)^2}{14} + \frac{(16-17.6)^2}{17.6} + \frac{(10-8.4)^2}{8.4} = 0.709$$

Step (4): Decision Rule: If $\chi^2_{stat>12.592}$ Reject H_0, otherwise, do not reject H0

Step (5): Here, $\chi^2_{stat=0.709<\chi^2_{0.05}=12.592}$,

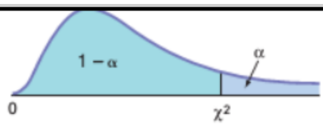
so do not reject H_0

Conclusion: there is insufficient evidence that meal plan and class standing are related at $\alpha = 0.05$



Critical Values of χ^2

For a particular number of degrees of freedom, entry represents the critical value of χ^2 corresponding to the cumulative probability $(1 - \alpha)$ and a specified upper-tail area (α).



Degrees of Freedom	Cumulative Probabilities											
	0.005	0.01	0.025	0.05	0.10	0.25	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas (α)											
	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548