Statistical Inference Theory
Lesson 43
The Chi-Square Distribution

43.1- What is the Chi-square distribution?

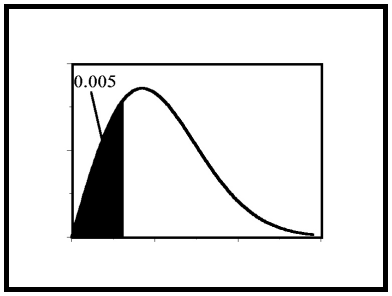
43.1 Problem 1:
Step 1: Since N = 21, the number of degrees of freedom is d = 21.

Step 2: Since $\alpha = 0.005$, select from the first row of the table $\chi^2_{0.005.}$.

Step 3: Select from the first column the row for d = 21.

Step 4: Match this row with the column of $\chi^2_{0.005.}$ The intersection of this column and row is $\chi^2_{0.005} = 41.4$.

1.

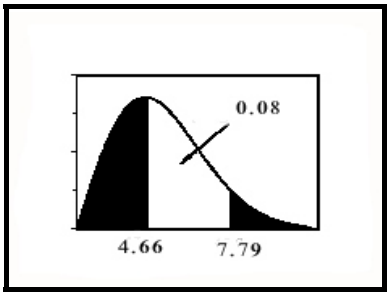


43.1 Problem 2:
Step 1: Since N = 10, the number of degrees of freedom is d = 10.

Step 2: Since the shaded area is 0.005, we need to compute $\alpha = 1 - 0.005 = 0.995$.

Step 3: Using $\alpha = 0.995$, we use $\chi^2_{0.995}$. From the table we find $\chi^2_{0.995} = 6.91$.

3.



43.1 - Problem 3:
Step 1:Since N = 14, the number of degrees of freedom is d = 14.

Step 2: For d = 14, $\chi^2_{0.90} = 7.79$.

Step 3: For d = 14, $\chi^2_{0.99} = 4.66$.

Step 4: From step 2 the right-hand shaded area in the figure is 0.90.

Step 5: From step 3 the right-hand shaded area in the figure is 1 - 0.99 = 0.01.

Step 6: Therefore, the non-shaded area is 1 - (0.01 + 0.90) = 0.08.

43.1 - Problem 4:

Since our chi-square table only exists for N ≤ 30, we use the formula $\chi^2 \approx \frac{1}{2}(Z + \sqrt{2d - 1})^2$.

where Z is standard normal distribution with mean 0 and standard deviation 1.
.

Step 1: Since $\alpha = 0.90$, rewrite the above formula as

$\chi^2_{0.90} \approx \frac{1}{2}(Z_{0.90} + \sqrt{2d - 1})^2$.

Step 2: d = 35

Step 3: Since Z is a standard normal distribution random variable, look up the area in the standard normal distribution table for 0.90 - 0.50 = 0.40.

Step 4: From step 3, $Z_{0.90} = -1.28$.

Step 5: $\chi^2_{0.05} \approx \frac{1}{2}(Z_{0.90} + \sqrt{2d - 1})^2 = \frac{1}{2}(-1.28 + \sqrt{2(35) - 1})^2 \approx 24.7$.

43.2- Estimating σ^2 and σ using a χ^2 confidence interval.

43.2 - Problem 1:
➤(a).
Step 1: Since we have a 99% confidence interval, $\alpha = (1 - 0.99)/2 = 0.005$, $1 - \alpha = 0.995$.

Step 2: For N - 1 = 25 - 1 = 24 degrees of freedom, from the table we have

$\chi^2_{0.05} = 36.4$,

$\chi^2_{0.995} = 9.89$.

Step 3: From the above formula: $\frac{25(0.0096)}{36.4} \le \sigma^2 \le \frac{25(0.0096)}{9.89}$,

$0.0066 \le \sigma^2 \le 0.0.024.$

Step 4: Take the square root of each number in the inequality: $0.08 \le \sigma \le 0.15.$.

➤(b).

Since N = 50, we use the formula: $\chi^2 \approx \frac{1}{2}(Z + \sqrt{2(N) - 1})^2$.

Step 1: $\chi^2_{0.005} \approx \frac{1}{2}(Z_{0.005} + \sqrt{2(50) - 1})^2 = (1/2)(2.58 + 9.95)^2 \approx 78.5$

Step 2: $\chi^2_{0.995} \approx \frac{1}{2}(Z_{0.995} + \sqrt{2(50) - 1})^2 = (1/2)(-2.58 + 9.95)^2 \approx 27.38$

Step 3: From the above formula: $\frac{50(0.0092)}{78.5} \le \sigma^2 \le \frac{50(0.0092)}{27.38}$,

$0.006 \le \sigma^2 \le 0.017.$

Step 4: Taking the square root of both sides gives $0.08 \le \sigma \le 0.13.$

43.3- Hypothesis testing for and a population standard deviation σ.

43.3 - Problem 1:

➤(a).

The concern is that the standard deviation will be larger than σ = 1. Therefore,

H_o: σ = 1

H_a: σ > 1

➤(b).

Step 1: Since the alternative hypothesis is σ > 0.1, we use the right-hand side of the chi-square table for α = 0.10.

Step 2: N = 10 , s = 1.8, σ = 0.10

Step 3: $\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{(10)(1.8^2)}{1^2} = 32.4$

Step 4: For d = 10 - 1 = 9, and α = 0.10, the Chi-square table gives $\chi^2_{0.10} = 14.7$.

Step 5: Therefore, c* = 14.7.

Step 6: The decision rule is : If $\chi^2 \ge 14.7$ then reject the claim.

➤(c).
Since $\chi^2 = 32.4 > 14.7$, the value s = 1.8 minutes cannot not be explained as caused by random variation. Therefore, the claim is rejected.

43.3 - Problem 2:
For these problems, we use the following formula:

$$\chi^2 = \frac{(x_1 - e_1)^2}{e_1} + \frac{(x_2 - e_2)^2}{e_2} + \frac{(x_3 - e_3)^2}{e_3} + \ldots + \frac{(x_N - e_N)^2}{e_N}$$,where

x_k is the observed frequencies,

e_k is the expected frequencies,

d = N - 1, the degrees of freedom.

➤(a).
H_o: The machines were down no more than 5% of the time.
H_a: The machines were down more than 5% down time.

➤(b).
To complete the above table, we need to compute each expected values.

Step 1: To compute the expected values we assume H_o is true.

Step 2: Records of down-time was taken over 1,000 hours. .

Step 3: The claim is that at most each machine will be down 5% of the time.

Since 5%X 1,000 = 50, we have

Step 4:

Machine	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Number hours down	57	60	55	48	59	61	58
Expected number of hours down	50	50	50	50	50	50	50

➤(c).
Step 1: For the above formula we set

$x_1 = 57, x_2 = 60, x_3 = 55, x_4 = 48, x_5 = 59, x_6 = 61, x_7 = 58$

$e_1 = 50, e_2 = 50, e_3 = 50, e_4 = 50, e_5 = 50, e_6 = 50, e_7 = 50$, d = 7 - 1 = 6

Step 2: $\chi^2 = \frac{(57 - 50)^2}{50} + \frac{(60 - 50)^2}{50} + \frac{(55 - 50)^2}{50} + \frac{(48 - 50)^2}{50} +$

$\frac{(59 - 50)^2}{50} + \frac{(61 - 50)^2}{50} + \frac{(58 - 50)^2}{50} = 8.88$

Step 3: For d = 7 - 1 = 6, the chi-square table give $\chi^2_{0.10} = 10.60$.

Step 4: Since 8.88 < 10.6 we reject H_a and at $\alpha = 0.10$ level of significance, we conclude that we have no statistical basis for rejecting the leasing company's claim.

43.3 - Problem3:

➤(a).

H_o: The percentage of voters concern on issues has not changed.

H_a: The percentage of voters concern on issues has changed.

➤(b).

To complete the above table, we need to compute each expected values.

Step 1: To compute the expected values we assume H_o is true.

Step 2: Since we assume there is no change from the first survey, we multiply each cell in the first table times 1,500:

Issues	Percentage of voters survey
Crime	52%X 1500 = 780
Decline in family values	31%X1500 = 465
Federal Balanced Budget	11%X1500 = 165
Environment	6%X1500 = 90

Step 3: We now have the following table:

Issues	Number of voters in survey	Expected Number of voters in survey
Crime	810	780
Decline in family values	501	465
Federal Balanced Budget	144	165
Environment	45	90

➤(c).
Step 1: For the above formula we set

$x_1 = 810$, $x_2 = 501$, $x_3 = 144$, $x_4 = 45$,

$e_1 = 780$, $e_2 = 465$, $e_3 = `165$, $e_4 = 90$,

Step 2: $\chi^2 = \frac{(810 - 780)^2}{780} + \frac{(501 - 465)^2}{465} + \frac{(144 - 165)^2}{165} + \frac{(45 - 90)^2}{90} \approx 29.75$

Step 3: For d = 5 - 1 = 4, the chi-square table give $\chi^2_{0.05} = 9.49$.

Step 4: Since 9.49 < 20.33 we reject H_0 and conclude that at a $\alpha = 0.05$ level of significance, we have reason to believe the percentages of defective hard drives reported by the five companies are not correct.

➤(d).
Step 4: Since 29.75 > 7.81, reject H_o. For $\alpha = 0.05$, there has been a significant change of opinion.

43.3 - Problem 4:

➤(a).
H_o: Her vacancy rate is 35%.
H_a. Her vacancy rate is not 35%.

➤(b).
Step 1: Assume H_o is true.

Step 2: It is reasonable to assume that the distribution of of vacancies is a binomial distribution:

$P\{X = k\} = \binom{6}{k}(0.35)^k(0.65)^{6-k}$

Number of rooms vacant per day k	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Number of days vacant	266	880	1244	750	305	52	3
Expected Number of days vacant $3500\binom{6}{k}(0.35)^k(0.65)^{6-k}$	262.5	854	1148	822.5	332.5	70	7

➤(c).
Step 1: For the above formula we set

$x_1 = 266$, $x_2 = 880$, $x_3 = 1244$, $x_4 = 750$, $x_5 = 305$, $x_6 = 52$, $x_7 = 3$

$e_1 = 262.5$, $e_2 = 854$, $e_3 = 1148$, $e_4 = 822.5$, $e_5 = 332.5$, $e_6 = 70$, $e_7 = 7$

Step 2: $\chi^2 = \frac{(266 - 262.5)^2}{262.5} + \frac{(880 - 854)^2}{854} + \frac{(1244 - 1148)^2}{1148} + \frac{(750 - 822.5)^2}{822.5} +$

$\frac{(305 - 332.5)^2}{332.5} + \frac{(52 - 70)^2}{70} + \frac{(3 - 7)^2}{7} \approx 24.45$

➤(d).
For d = 7 - 1 = 6, the chi-square table give $\chi_{0.05} = 12.6$.

Since 24.45 > 12.6, we reject H_o and have a statistical basis for reject her claim of a vacancy rate of 35%.

43.5 - Contingency Tables

43.5 - Problem 1:
➤(a).
H_o: The association between these additives and gasoline mileage performance are statistically independent.

H_a: The association between these additives and gasoline mileage performance are statistically dependent.

➤(b).
Assuming H_o is true, we have

35% $(\frac{35}{100} = 35\%)$ of the cars tested used additive A.

41% $(\frac{41}{100} = 41\%)$ of the cars tested used additive B.

24% $(\frac{24}{100} = 24\%)$ of the cars tested used additive C.

	0% - 4.99% Mileage increase	5% - 9.99% mileage increase	10% - 14.99% mileage increase	15% mileage increase or more	Total
Additive A	(0.35)(27) = 9.45	(0.35)(25) = 8.75	(0.35)(28) = 9.8	(0.35)(20) = 7	35
Additive B	(0.41)(27) = 11.07	(0.41)(25) = 10.25	(0.41)(28) = 11.48	(0.41)(20) = 8.2	41
Additive C	(0.24)(27) = 6.48	(0.24)(25) = 6	(0.24)(28) = 6.72	(0.24)(20) = 4.8	24
Total	27	25	28	20	100

➤(c).
Step 1: For the formula,

$$\chi^2 = \frac{(x_1 - e_1)^2}{e_1} + \frac{(x_2 - e_2)^2}{e_2} + \frac{(x_3 - e_3)^2}{e_3} + \ldots + \frac{(x_N - e_N)^2}{e_N} .$$

Step 2: $x_1 = 12$, $x_2 = 11$, $x_3 = 8$, $x_4 = 4$, $x_5 = 11$, $x_6 = 9$, $x_7 = 12$, $x_8 = 9$,

$x_9 = 7$, $x_{10} = 5$, $x_{11} = 8$, $x_{12} = 4$

$e_1 = 66.78$, $e_2 = 51.1$, $e_3 = 22.12$, $e_4 = 124.974$, $e_5 = 95.63$, $e_6 = 41.396$,

$e_7 = 191.277$, $e_8 = 146.365$, $e_9 = 63.358$, $e_{10} = 93.969$, $e_{11} = 71.905$, $e_{12} = 31.126$

Step 3: $$\chi^2 = \frac{(12 - 9.45)^2}{9.45} + \frac{(11 - 8.75)^2}{8.75} + \frac{(8 - 9.8)^2}{9.8} + \frac{(4 - 7)^2}{7} +$$

$$\frac{(11 - 11.07)^2}{11.07} + \frac{(9 - 10.25)^2}{10.25} + \frac{(12 - 11.48)^2}{11.48} + \frac{(9 - 8.2)^2}{8.2} +$$

$$\frac{(7 - 6.48)^2}{6.48} + \frac{(5 - 6)^2}{6} + \frac{(8 - 6.72)^2}{6.72} + \frac{(4 - 4.8)^2}{4.8} \approx 3.72$$

Step 4:d = (r - 1)(c - 1) = (3-1)(4-1) = 6

➤(d).
For $\alpha = 0.05$, and 6 degree of freedom, $\chi^2_{0.05} = 12.6$.

Since 3.72 < 12.6, reject H_a and accept H_o Therefore, we conclude that there is no statistical dependent relationship between these additives and improved gasoline mileage.

Supplementary Problems

1.

From the equation $\frac{Ns^2}{\sigma^2} = \chi^2$, we solve for s^2 as follows:

Step 1: Multiply both sides by σ^2: $Ns^2 = \sigma^2\chi^2$,

Step 2: Divide both sides by N: $s^2 = \frac{\chi^2\sigma^2}{N}$,

To solve for s, take the square root of both sides: $s = \frac{\chi\sigma}{\sqrt{N}}$.

2.

For a given sample size N and α, find a general confidence interval formula for s^2 and s.

From section 2, of the book we have: $\frac{Ns^2}{\chi^2_{\alpha}} \leq \sigma^2 \leq \frac{Ns^2}{\chi^2_{1-\alpha}}$.

Using a sequence of algebraic steps on the above formula, we get the following confidence interval:

$$\frac{\sigma^2\chi^2_{1-\alpha}}{N} \leq s^2 \leq \frac{\sigma^2\chi^2_{\alpha}}{N}.$$

Taking the square root of the above inequality gives

$$\frac{\sigma\chi_{1-\alpha}}{\sqrt{N}} \leq s \leq \frac{\sigma\chi_{\alpha}}{\sqrt{N}} .$$

3.

In question 2, we can use the following confidence interval for s^2:

$$\frac{\sigma^2\chi^2_{1-\alpha}}{N} \leq s^2 \leq \frac{\sigma^2\chi^2_{\alpha}}{N}.$$

Step 1: N = 100, the sample size of bottles taken.

$\sigma^2 = 0.1$

Step 2: For a 95% confidence interval, we have (1 - 0.95)/2 = 0.025 = α.

Step 3: Since N exceeds 30, we use the formula:

$\chi^2 \approx \frac{1}{2}(Z + \sqrt{2d - 1})^2$, where

Z is normally distributed with mean 0 and s.d. 1,

d is the number of degrees of freedom.

Since we want a 95% confidence, we have from the normal distribution table, Z = ± 1.96.

$\chi^2_{0.975} = \frac{1}{2}(Z + \sqrt{2d - 1})^2 = \frac{1}{2}(-1.96 + \sqrt{2(100) - 1})^2 = 73.81$

$\chi^2_{0.025} = \frac{1}{2}(Z + \sqrt{2d - 1})^2 = \frac{1}{2}(1.96 + \sqrt{2(100) - 1})^2 = 129.07$

Substituting these numbers into the above inequality gives:

$$\frac{(0.1)(73.81)}{100} \leq s^2 \leq \frac{(0.1)(129.07)}{100}.$$

$0.074 \leq s^2 \leq 0.13$,

$0.272 \leq s \leq 0.36$

4.
Since our chi-square table only exists for $N \leq 30$, we use the formula

$$\chi^2 \approx \frac{1}{2}(Z + \sqrt{2d - 1})^2.$$

where Z is standard normal distribution with mean 0 and standard deviation 1.

Step 1: Since $\alpha = 0.05$, rewrite the above formula as

$\chi^2_{0.05} \approx \frac{1}{2}(Z_{0.05} + \sqrt{2d - 1})^2$.

Step 2: d = 350

Step 3 :Since Z is a standard normal distribution random variable, look up the area in the standard normal distribution table for $0.50 - \alpha = 0.45$.

Step 4: From step 3 $Z_{0.05} = 1.64$.

Step 5: $\chi^2_{0.05} \approx \frac{1}{2}(Z_{0.05} + \sqrt{2d - 1})^2 = \frac{1}{2}(1.64 + \sqrt{2(350) - 1})^2 \approx 394.2$

5.

For the formula we use d = 50. Substituting the values in each cell in the second row into the above formula gives the cell values in the first row. For example

$$z = \sqrt{2\chi^2} - \sqrt{2d - 1} = \sqrt{2(10)} - \sqrt{2(50) - 1} \approx -5.48 .$$

Following this procedure we have

z	-5.48	-3.63	-2.20	-1.01	1
χ^2	10	20	30	40	60

A machine drills holes in metal plates. The diameter tolerance of each hole is σ = 0.001 millimeters. Each hour 50 plates are tested for drill accuracy by computing s.

6.

H_o: σ = 0.001

H_a: σ ≠ 0.001

7.

Rewriting the above formula, we get the confidence interval:

$$\frac{\sigma\chi_{1-\alpha}}{\sqrt{N}} \leq s \leq \frac{\sigma\chi_{\alpha}}{\sqrt{N}}.$$

N = 50

σ = 0.001

$$\chi^2_{0.02} \approx \frac{1}{2}(Z_{0.02} + \sqrt{2d - 1})^2 = \frac{1}{2}(2.06 + \sqrt{2(50) - 1})^2 \approx 72.12$$

$$\chi^2_{0.90} \approx \frac{1}{2}(Z_{0.98} + \sqrt{2d - 1})^2 = \frac{1}{2}(-2.06 + \sqrt{2(50) - 1})^2 \approx 31.12$$

$$\frac{(0.001)(31.12)}{\sqrt{50}} \leq s \leq \frac{(0.001)(72.12)}{\sqrt{50}}$$

$0.004 \leq s \leq 0.01$

Since an s = 0 is the best tolerance we state the decision rule as:

If $0 \leq s \leq 0.01$ allow the machine to continue; otherwise stop the machine.

8.
Since $s^2 = 0.0009$, then $s = 0.03$, the s value is outside the interval. Therefore, the machine will be stopped.

9.
The confidence interval changes form

$$\frac{(0.001)(31.12)}{\sqrt{50}} \leq s \leq \frac{(0.001)(72.12)}{\sqrt{50}}$$

to

$$\frac{(0.0015)(31.12)}{\sqrt{50}} \leq s \leq \frac{(0.0015)(72.12)}{\sqrt{50}}.$$

$0.007 \leq s \leq 0.015$

If $s^2 = 0.0019$, then $s \approx 0.044$ and since $0.044 \geq 0.015$ the machine should be shut down.

10.
H_o: There is no association for a recovery between the drug, the placebo and gender.

H_a: There is a association for a recovery between the drug, the placebo and gender.

11.
From the table below

134/250 = 53.6% recovered and

116/250 = 46.4% did not recover.

Patients	**Recovered**	**Did not recover**	**Total**
Drug (males)	34	16	50
Drug (female)	39	36	75
Placebo (male)	23	27	50
Placebo (female)	38	37	75
Total	134	116	250

The expectation is therefore:

Patients	**Recovered**	**Did not recover**	**Total**
Drug (males)	(0.536)(50) = 26.8	(0.464)(50) = 23.2	50
Drug (female)	(0.536)(75) = 40.2	(0.464)(75) = 34.8	75
Placebo (male)	(0.536)(50) = 26.8	(0.464)(50) = 23.2	50
Placebo (female)	(0.536)(75) = 40.2	(0.464)(75) = 34.8	75
Total	134	116	250

$x_1 = 34$, $x_2 = 16$, $x_3 = 39$, $x_4 = 36$, $x_5 = 23$, $x_6 = 27$, $x_7 = 38$, $x_8 = 37$

$e_1 = 26.8$, $e_2 = 23.2$, $e_3 = 40.2$, $e_4 = 34.8$,

$e_5 = 26.8$, $e_6 = 23.2$, $e_7 = 40.2$, $e_8 = 34.8$

$$\chi^2 = \frac{(34 - 26.8)^2}{26.8} + \frac{(16 - 23.2)^2}{23.2} + \frac{(39 - 40.2)^2}{40.2} + \frac{(36 - 34.8)^2}{34.8} +$$

$$\frac{(23 - 26.8)^2}{26.8} + \frac{(27 - 23.2)^2}{23.2} + \frac{(38 - 40.2)^2}{40.2} + \frac{(37 - 34.8)^2}{34.8} \approx$$

$1.93 + 2.23 + 0.04 + 0.04 + 0.54 + 2.32 + 0.12 + 0.14 = 7.32$

12.
Since N = 250, we use the formula,

$$\chi^2_{0.05} \approx \frac{1}{2}(Z_{0.05} + \sqrt{2d - 1})^2 = \frac{1}{2}(1.64 + \sqrt{2(250) - 1})^2 \approx 287.$$

Since $7.32 < 287$, H_o would not be rejected.