13.1, 13.2 Chi-Squared Dist, Goodness of Fit

GOALS:

1. Learn the properties of the χ^2 Distribution.
2. Understand how the shape of the χ^2 Distribution changes as the df increases.
3. Be able to find p-values.
4. Recognize that χ^2 tests are right-tailed only.
5. Use the "Goodness of Fit" χ^2 test to compare samples to known or expected distributions.

Study Ch. 13.1, # 1-4 all

Study Ch. 13.2, # 9-15, 25, 27, 31

[# 11-17, ~27, 29, ~33]

link to geogebra demo

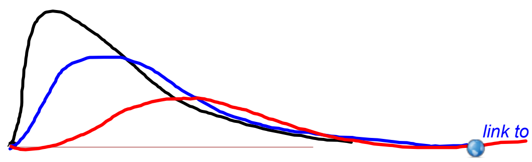
Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Chi-Squared distribution, χ^2

1. Not symmetrical: Right-skewed.
   On left, starts at 0 on x-axis.
   On right, approaches x-axis as asymptote.
2. Area under χ^2 Curve = 1
3. Different curves for different df
   As df increases, χ^2 --> normal curve



link to geogebra demo

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

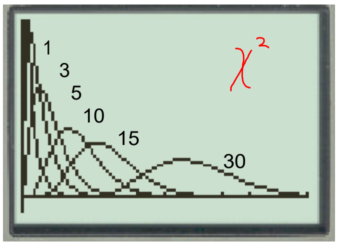
χ^2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Graph Chi-Squared distribution, χ^2, on Calculator

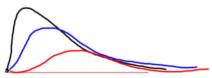
link to geogebra demo

1. Enter into L1:
   1, 3, 5, 10, 15, 30
2. WINDOW:
   xmin: -0.02
   xmax: 50
   xscale: 5
   ymin: -0.02
   ymax 0.25
3. y= 2nd DISTR χ^2pdf(x,L1)
4. GRAPH



df from 1 to 30

Different curves for different df
As df increases, χ^2 --> normal curve

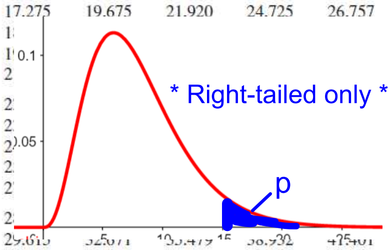


Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

χ^2

$\chi^2_{0.10}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	df
2.706	3.841	5.024	6.635	7.879	1
4.605	5.991	7.378	9.210	10.597	2
6.251	7.815	9.348	11.345	12.838	3
7.779	9.488	11.143	13.277	14.860	4
9.236	11.070	12.833	15.086	16.750	5
10.645	12.592	14.449	16.812	18.548	6
12.017	14.067	16.013	18.475	20.278	7
13.362	15.507	17.535	20.090	21.955	8
14.684	16.919	19.023	21.666	23.589	9
15.987	18.307	20.483	23.209	25.188	10
17.275	19.675	21.920	24.725	26.757	11
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	13
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	14
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	15
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	16
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	17
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	18
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	19
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	20
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	21
30.813	33.924	36.781	40.290	42.796	22
32.007	35.172	38.076	41.638	44.181	23
33.196	36.415	39.364	42.980	45.559	24
34.382	37.653	40.647	44.314	46.928	25
35.563	38.885	41.923	45.642	48.290	26
36.741	40.113	43.195	46.963	49.645	27
37.916	41.337	44.461	48.278	50.994	28
39.087	42.557	45.722	49.588	52.336	29
40.256	43.773	46.979	50.892	53.672	30
51.805	55.759	59.342	63.691	66.767	40
63.167	67.505	71.420	76.154	79.490	50
74.397	79.082	83.298	88.381	91.955	60
85.527	90.531	95.023	100.424	104.213	70
96.578	101.879	106.628	112.328	116.320	80
107.565	113.145	118.135	124.115	128.296	90
118.499	124.343	129.563	135.811	140.177	100



Can use the tables to estimate the p-value:
If Given:
test χ^2 = 13.094, df = 6

$0.025 < p < 0.05$

OR,

Can compute the p-value on the calculator:
$p = 1 - \chi^2\text{cdf}(0, 13.094, 6)$
$= 0.0416$

Using
$p = 1 - \chi^2\text{cdf}(0, \text{test } \chi^2, df)$

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Goodness-of-Fit Test

Used to compare one distribution to another.

Requires:

1. simple random sample,
2. adequate sample size.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

In 1990 the distribution of cars, by type, was: 32.8 % small, 44.8% medium size, 9.4% large, and 13.0% larger (eg. SUV). For a recent simple random sample of 500 cars, car type is listed below. Has the distribution of car type changed?

	Obs	Given Probability:
S	133	.328
M	249	.448
L	47	.094
Lx	71	.130
Total	500	

Need to compare observed values with **the expected distribution**, based on assumptions (prior data, general knowledge, etc.)

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Chi-Square Goodness-of-Fit *only Right-tailed*

Step 1A: Calculate the expected frequencies $E = np$

where n = sample size, p = rel freq or probability; compare to assumptions

Assumptions:
1. All expected frequencies ≥ 1
2. At most 20% of the expected frequencies are less than 5
3. SRS

Step 1B: H_0: The variable has the specified distribution.

H_a: The variable does not have the specified distribution.

Step 2: Decide α

Step 3: Compute the test statistic, using a table of values:

$$\chi^2_T = \sum\left[\frac{(O-E)^2}{E}\right]$$

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
				Σ

Step 4 : Find **df = k - 1** where k = number of categories

Step 5 : Find CV(s) **on Table VII.**

Step 6: Decide: reject H_0 or not?
Reject if test statistic is in rejection region (tail).

Step 5: p-value from calculator
$p = 1 - \chi^2\text{cdf}(0, \chi^2_T, df)$

Step 6: Decide: reject H_0 or not?
Reject if $p < \alpha$

Step 7: Verbal interpretation

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

In 1990 the distribution of cars, by type, was: 32.8 % small, 44.8% medium size, 9.4% large, and 13.0% larger (eg. SUV). For a recent simple random sample of 500 cars, car type is listed below. At the 5% s.l., has the distribution of car type changed?

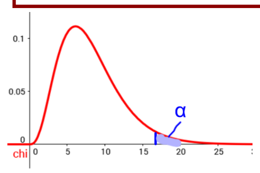
	Obs	Given Probability:
S	133	.328
M	249	.448
L	47	.094
Lx	71	.130
Total	500	

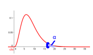
H_0: distribution is same as 1990
H_a: distribution is different from 1990

Need to compare observed values with **the expected distribution**, based on assumptions (prior data, general knowledge, etc.)

Calculator: χ^2 GOF Test
1. Enter observed data into L1
2. Enter expected probabilities into L2
3. Find Σx, the sum of observed values
4. In the header for L3, compute *np*:
   L2 x (Σx) (expected values)
5. STAT/TESTS/ χ^2 GOF-Test
   Observed: L1
   Expected: L3
   df: #categories - 1
   CALCULATE

Calculator: χ^2 **w/o** GOF Test
1. Enter observed data into L1
2. Enter expected probabilities into L2
3. Find Σx, the sum of observed values
4. In the header for L3, compute *np*:
   L2 x (Σx) (expected values)
5. In header for L4, compute indiv χ^2
   $(O - E)^2 / E$, or
   (L1-L3)² /L3 using column headings
6. STAT/CALC/1-Variable Stats/ **L4**
   Find Σx
   This is the **test statistic**, χ^2
7. Use critical value or find p from table.
   $p = 1 - \chi^2\text{cdf}(0, \chi^2_T, df)$





Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

In 1990 the distribution of cars, by type, was: 32.8 % small, 44.8% medium size, 9.4% large, and 13.0% larger (eg. SUV). For a recent simple random sample of 500 cars, car type is listed below. At the 5% s.l., has the distribution of car type changed?

H_0: distribution is same as 1990
H_a: distribution is different from 1990

	Obs	Exp	O-E	$(O-E)^2$	$(O-E)^2 / E$
S	133	164	-31	961	5.860
M	249	224	25	625	2.790
L	47	47	0	0	0.000
Lx	71	65	6	36	0.554
Total	500	500			9.204

Given Probability:
.328 E= np = 500*.328 = 164
.448
.094
.130 E= np = 500*.130 = 65

all E $\geq$ 1
none $\leq$ 5
∴ assumptions met

Step 1A: Calculate the expected frequencies $E = np$
where n = sample size, p = rel freq or probability; compare to assumptions
Assumptions: 1. All expected frequencies $\geq$ 1
2. At most 20% of the expected frequencies are less than 5
3. SRS
Step 1B: H_0: The variable has the specified distribution.
H_a: The variable does not have the specified distribution.

Step 2: Decide α

Step 3: Compute the test statistic, using a table of values:

$$\chi^2_T = \sum\left[\frac{(O-E)^2}{E}\right]$$

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$

Step 4: Find **df = k - 1** where k = number of categories

Step 5: Find CV(s) **on Table VII.**
Step 6: Decide: reject H_0 or not? Reject if test statistic is in rejection region (tail).

Step 5: p-value from calculator $p = 1 - \chi^2cdf(0, \chi^2_T, df)$
Step 6: Decide: reject H_0 or not? Reject if $p < \alpha$

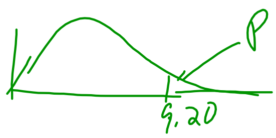
Step 7: Verbal interpretation

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

In 1990 the distribution of cars, by type, was: 3
medium size, 9.4% large, and 13.0% larger (e
recent simple random sample of 500 cars, car
At the 5% s.l., has the distribution of car type c



$p = 1-\chi^2cdf(0,9.204,3)$
$= 0.0267$

(1) H_0: distribution is same as 1990
H_a: distribution is different from 1990

L1 ✓ L3 ✓ np ✓ p L2

	Obs	Exp	O-E	$(O-E)^2$	$(O-E)^2 / E$
S	133	164	-31	961	5.860
M	249	224	25	625	2.790
L	47	47	0	0	0.000
Lx	71	65	6	36	0.554
Total	n 500	500			9.204

Given Probability:
.328
.448
.094
.130

p < 0.05 rej. H_0: concl. that distr. of vehicles has changed 1990

$p=0.0267 < 0.05 = \alpha$ ∴ Rej H_o
Conclude that the distribution of car types (sizes) has changed.

Step 1A: Calculate the expected frequencies $E = np$
where n = sample size, p = rel freq or probability; compare to assumptions
Assumptions: 1. All expected frequencies $\geq$ 1
2. At most 20% of the expected frequencies are less than 5
3. SRS
Step 1B: H_0: The variable has the specified distribution.
H_a: The variable does not have the specified distribution.

Step 2: Decide α

Step 3: Compute the test statistic, using a table of values:

$$\chi^2_T = \sum\left[\frac{(O-E)^2}{E}\right]$$

Step 4: Find **df = k - 1** where k = number of categories

Step 5: Find CV(s) **on Table VII.**
Step 6: Decide: reject H_0 or not? Reject if test statistic is in rejection region (tail).

Step 5: p-value from calculator $p = 1 - \chi^2cdf(0, \chi^2_T, df)$
Step 6: Decide: reject H_0 or not? Reject if $p < \alpha$

Step 7: Verbal interpretation

Calculator: χ^2 w/o GOF Test
1. Enter observed data into L1
2. Enter expected probabilities into L2
3. Find Σx, the sum of observed values
4. In the header for L3, compute *np*:
   L2 x (Σx) (expected values)
5. In header for L4, compute indiv χ^2
   $(O - E)^2 / E$, or
   $(L1-L3)^2 / L3$ using column headings
6. STAT/CALC/1-Variable Stats/ **L4**
   Find Σx
   This is the **test statistic,** χ^2
7. Use critical value or find p from table.
   $p = 1 - \chi^2cdf(0, \chi^2_T, df)$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Given

122

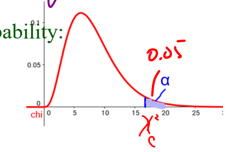
H_0: distribution is same as 1990

H_a: distribution is different from 1990

$p = 0.0267$

	L1 Obs	L3 Exp	O-E	$(O-E)^2$	L4 $(O-E)^2/E$	L2 Given Probability:
S	133	164	-31	961	5.860	.328
M	249	224	25	625	2.790	.448
L	47	47	0	0	0.000	.094
Lx	71	65	6	36	0.554	.130
Total	500	500			9.204	

$9.204 = \chi^2_T$



CNTRB = {5.859

df = c - 1 = 4 - 1 = 3

Calculator: χ^2 GOF Test

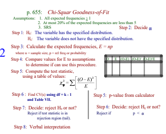
1. Enter observed data into L1
2. Enter expected probabilities into L2
3. Find Σx, the sum of observed values
4. In the header for L3, compute *np*:
   L2 x (Σx) (expected values)
5. STAT/TESTS/ χ^2 GOF-Test
   Observed: L1
   Expected: L3
   df: #categories - 1
   CALCULATE

Calculator: χ^2 w/o GOF Test

1. Enter observed data into L1
2. Enter expected probabilities into L2
3. Find Σx, the sum of observed values
4. In the header for L3, compute *np*:
   L2 x (Σx) (expected values)
5. In header for L4, compute indiv χ^2
   $(O - E)^2 / E$, or
   $(L1-L3)^2 / L3$ using column headings
6. STAT/CALC/1-Variable Stats/ **L4**
   Find Σx
   This is the **test statistic,** χ^2
7. Use critical value or find p from table.
   $p = 1 - X^2cdf(0, X^2_T, df)$

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2



13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Calculator: χ^2 w/o GOF Test

1. Enter observed data into L1
2. Enter expected probabilities into L2
3. Find Σx, the sum of observed values
4. In the header for L3, compute *np*:
   L2 x (Σx) (expected values)
5. In header for L4, compute indiv χ^2
   $(O - E)^2 / E$, or
   $(L1-L3)^2 / L3$ using column headings
6. STAT/CALC/1-Variable Stats/ **L4**
   Find Σx
   This is the **test statistic,** χ^2
7. Use critical value or find p from table.
   $p = 1 - X^2cdf(0, X^2_T, df)$

CNTRB = {XXXX.XX XXXX ...} ARROW RIGHT
each cell's contribution to test statistic

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Geographical Distribution in US Population
At the 5% s.l., has the distribution changed?

H_0: geographical distribution in US is as given
H_a: distribution is different from given

	Freq	Obs	Exp	$(O-E)^2/E$
NE	0.190	45		
MW	0.229	42		
S	0.356	92		
W	0.225	71		
		250		

all E ≥ 1
none ≤ 5 ?

χ^2 = ______

Use χ^2 GOF test, or p = 1 – χ^2cdf (0,______,__)
= _________

p=_______ ? 0.05 = α
Reject Ho or not?

Conclude: ___________________

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

Geographical Distribution in US Population
At the 5% s.l., has the distribution changed?

H_0: geographical distribution in US is as given
H_a: distribution is different from given

df=4-1=3

all E ≥ 1
none ≤ 5 !
assumptions met

	Freq	Obs	Exp	$(O-E)^2/E$
NE	0.190	45	47.5	0.132
MW	0.229	42	57.25	4.062
S	0.356	92	89.0	0.101
W	0.225	71	56.25	3.868
		250		χ^2 = 8.163

Use χ^2 GOF test, or p = 1 – χ^2cdf (0,8.163,3)
= 0.0427

p=0.0427 < 0.05 = α
Reject Ho

Conclude: population distribution in US has changed.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

A traffic commissioner wanted to know if Road Rage occurred more often on some days of the week. She randomly selected 70 incidents of Road Rage and examined the days of the week.on which they occurred.

F: At the 5% significance level, does the incidence of Road Rage occur more often on some days of the week than on other days?

Day	freq
Su	5
Mo	5
Tu	11
We	12
Th	12
Fr	18
Sa	7

How is this a GOF problem?

What is the null hypothesis?

What are the expected values?

H_0: Road Rage **is not** associated with the day of the week.
H_a: Road Rage **is** associated with the day of the week.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

A traffic commissioner wanted to know if Road Rage occurred more often on some days of the week. She randomly selected 70 incidents of Road Rage and examined the days of the week.on which they occurred.

F: At the 5% significance level, does the incidence of Road Rage occur more often on some days of the week than on other days?

Day	freq
Su	5
Mo	5
Tu	11
We	12
Th	12
Fr	18
Sa	7

How is this a GOF problem?

What is the null hypothesis?

No difference in days

What are the expected values?

Equal numbers ∴ 70*(1/7)

E = n p

H_0: Road Rage **is not** associated with the day of the week.

H_a: Road Rage **is** associated with the day of the week.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

A traffic commissioner wanted to know if Road Rage occurred more often on some days of the week. She randomly selected 70 incidents of Road Rage and examined the days of the week.on which they occurred.

F: At the 5% significance level, does the incidence of Road Rage occur more often on some days of the week than on other days?

Day	Obs freq	Exp =np
Su	5	10.0
Mo	5	10.0
Tu	11	10.0
We	12	10.0
Th	12	10.0
Fr	18	10.0
Sa	7	10.0
n=70		all >1 none < 5

How is this a GOF problem?

What is the null hypothesis?

No difference in days

What are the expected values?

Equal numbers ∴ 70*(1/7)=10

H_0: Road Rage **is not** associated with the day of the week.
H_a: Road Rage **is** associated with the day of the week.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

A traffic commissioner wanted to know if Road Rage occurred more often on some days of the week. She randomly selected 70 incidents of Road Rage and examined the days of the week.on which they occurred.

F: At the 5% significance level, does the incidence of Road Rage occur more often on some days of the week than on other days?

Day	Obs freq	Exp =np	$(O-E)^2/E$
Su	5	10.0	2.5
Mo	5	10.0	2.5
Tu	11	10.0	.1
We	12	10.0	.4
Th	12	10.0	.4
Fr	18	10.0	6.4
Sa	7	10.0	.9
n=70		all >1 none < 5	13.2

How is this a GOF problem?

What is the null hypothesis?

No difference in days

What are the expected values?

Equal numbers ∴ (1/7)*69

Find: test statistic

$$\chi^2_T = \sum\left[\frac{(O-E)^2}{E}\right]$$

= 13.2

H_0: Road Rage **is not** associated with the day of the week.
H_a: Road Rage **is** associated with the day of the week.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2

13.1, 13.2 Chi-Squared Dist, Goodness of Fit

A traffic commissioner wanted to know if Road Rage occurred more often on some days of the week. She randomly selected 70 incidents of Road Rage and examined the days of the week.on which they occurred.

F: At the 5% significance level, does the incidence of Road Rage occur more often on some days of the week than on other days?

How is this a GOF problem?
What is the null hypothesis?
No difference in days
What are the expected values?
Equal numbers ∴ (1/7)*69

Day	Obs freq	Exp =np	$(O-E)^2/E$
Su	5	10.0	2.5
Mo	5	10.0	2.5
Tu	11	10.0	.1
We	12	10.0	.4
Th	12	10.0	.4
Fr	18	10.0	6.4
Sa	7	10.0	.9
n=70		all >1 none < 5	13.2

Find: test statistic

$$\chi^2_T = \sum\left[\frac{(O-E)^2}{E}\right] = 13.2$$

Need p:

$p = 1 - \chi^2\text{cdf}(0, \text{test } \chi^2, \text{df})$

Need df: df=c-1=7-1=6

$p = 1 - \chi^2\text{cdf}(0, 13.2, 6)$
$= 0.040$

$p = 0.040 < 0.05 = \alpha$ rej. H_o

Conclude: Road Rage is different for different days of the week.

H_0: Road Rage **is not** associated with the day of the week.
H_a: Road Rage **is** associated with the day of the week.

Class Notes: Prof. G. Battaly, Westchester Community College, NY

Statistics Home Page Class Notes Homework1 Homework2