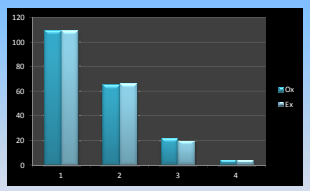
STATS 101
Introductory Statistics

CHI SQUARE TESTS



- **CHI-SQUARE TESTS** are used for

1. Testing goodness of fit
2. Testing independence of two ATTRIBUTES in a table
3. Testing homogeneity of proportions in a table

In this lecture, you will learn how to run chi-square tests in EXCEL (see chi square tests.xlsx)

The Chi-Square Test

In each application of the chi-square test, we will be given –
O_j = observed frequency of cell j
E_j = frequency of cell j computed assuming null hypothesis to be true

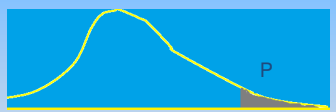
The idea behind the chi-square test is simple – if all of the expected frequencies E_j are 'close' to the corresponding observed frequencies O_j then the null hypothesis should be true. The chi-square test, therefore, has the following form:

Reject H_0 if the *DISTANCE* between the observed and expected frequencies is *LARGE*. The chi-square distance between the observed and expected frequencies (O and E) given below is used:

$$DIST(O-E) = \chi^2_{obs} = \sum \frac{(O_j - E_j)^2}{E_j}$$

In each application of the chi-square test, the null is rejected if the P-value:

$P = P(\chi^2_{df} > \chi^2_{obs}) < .05$ if using a test of size .05



χ^2_{obs}

The degree of freedom (df) of the chi-square test depends on the chi-square problem.

1. **Testing Goodness of Fit (GOF)** of a specified probability

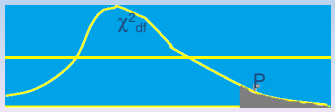
- If data is continuous, calculate a histogram (this step is not needed for discrete data); this gives us OBSERVED FREQUENCIES O_I for each class interval.
- Calculate EXPECTED FREQUENCY E_I for Cell (class interval) i by the formula
  E_I = N x P(an observation falls in Cell i if H_0 is true).
- Calculate $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Here k = # of cells or class intervals

Reject H_0 (conclude data does not follow hypothesized probability distribution) if

$$\chi^2_{obs} > \chi^2_{df}$$
$$df = k - \#\ of\ parameters\ estimated - 1$$

Alternatively, reject H_0 if the P-value as shown in following figure is < .05



TESTING GOODNESS OF FIT EXAMPLE (Discrete data)

Ex 1: Use the data (x, O) to test if the die is fair:

x	O_i	E_i	chi-sqr
1	28	30	0.13
2	36	30	1.2
3	36	30	1.2
4	30	30	0
5	27	30	0.3
6	23	30	1.63
SUM	180	180	4.46

In this table:
Ox = observed frequency of value x
Ex = EXPECTED frequency of value x

Ex = frequency expected under the null hypothesis H_0

H_0 : f(x) = 1/6, x = 1, 2, …, 6

Ex = N x f(x) = 180 (1/6) = 30 for each x

χ^2 column = $(Ox - Ex)^2 / Ex$

$$H_0: P(X = x) = 1/6, x = 1,2,\ldots,6$$
$$H_1: H_0 - false$$

FACT: If the null hypothesis is true,

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df}$$
$$df = k - 1 - \#(\text{parameters estimated})$$

df = k – 1 = 6-1 = 5 for the above example, as the null hypothesis specified the probabilities and no parameters had to be estimated (i.e., k = 0)

$$\chi^2_{calc} = \frac{(-2)^2 + (6)^2 + (6)^2 + (0)^2 + (-3)^2 + (-7)^2}{30} = \frac{134}{30} = 4.47$$
$$df = 6 - 1 = 5$$
$$\chi^2_{5,.95} = 11.07$$

We can get the P-value $P(\chi^2_5 > 4.47)$ from EXCEL in 2 ways:
1) =chidist(.47,5) will return the P-value = 0.48
2) chitest(observed range, expected range) will only return the P-value, and not χ^2_{OBS}.

We do not reject the null hypothesis, and conclude that the die is fair.

NOTE: For the above chi-square test to be valid, each E_i >= 5. If this is not the case, then combine 2 or more adjacent cells.

Using EXCEL to run Chi-square Goodness of Fit (GOF) test for Example 1

1) Type observed frequencies in column A , cells 2 – 7. Label the column Ox (use cell A1).
2) In cell C2, type =180/6 to get Ex.
3) Copy formula in cell B2 to cells B3 – B7.
4) In cell C1, type 'Chi-sqr' to label the chi-square column.
5) In cell C2, type =(A2 – B2)^2/B2 to calculate the chi-square contribution from the 1st cell. Copy formula in cell C2 to cells C3 – C7. You now have the table on the next slide.

Using EXCELto run GOF test for Example 1

EXCEL COLUMN

	A	B	C
	Ox	Ex	Chi-square
	28	30	0.13
	36	30	1.20
	36	30	1.20
	30	30	0.00
	27	30	0.30
	23	30	1.63
TOTAL	180	180	4.47

=sum(A2:A7)

To get the P-value for the observed chi-square value, type =CHIDIST(4.47,5) in a blank cell in EXCEL. This will return P = 0.48

Since P > .05, we conclude that die is fair.

Using EXCELto run GOF test for Example 1

We can also get the P-value from EXCEL directly, WITHOUT CALCULATING the χ^2_{obs} value.

Go to a blank cell in EXCEL, and type

=chitest(A2:A7,B2:B7)

This will just yield the P-value of 0.48 (same as on the previous slide).

2. CHI-SQUARE TEST OF INDEPENDENCE

Example 2: Test if AGE and GAME TYPE are independent, at test size $\alpha = 0.01$.

	AGE			
GAME	21-25	26-50	Over 50	TOTAL
Multi-Line Slots	15	37	16	68
Video Poker	25	25	17	67
Wheel of Fortune	14	40	27	81
Sports Book	11	4	1	16
Blackjack	9	23	14	46
Megabucks	3	8	1	12
TOTAL	77	137	76	290

Recall that A and B are independent if $P(A \cap B) = P(A)P(B)$

If H_0 is true:

P(Multi line slot, Age 21 – 25) =

P(Multi line slot)) x P(Age 21 – 25)

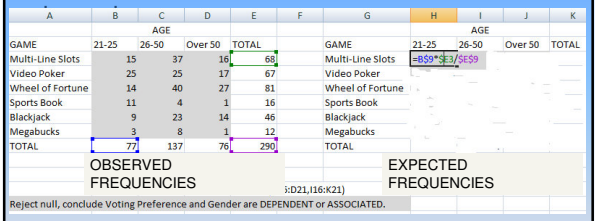
= (68/290)(77/290)

$E_{1,1}$ = 290 x (68/290)(77/290)

Hence = 68 x 77/290

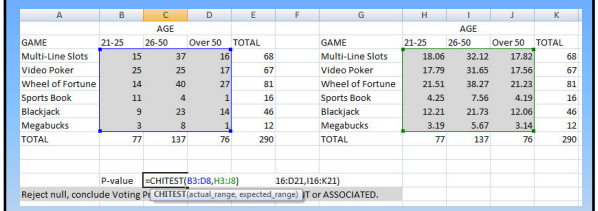
$E_{i,j}$ = Total of Row-i x Total of Column j/ Grand Total

To run the chi-square test of independence, type the data table and copy row and column headings next to data table as



If you typed data exactly as shown, type =B$9*$E3/$E$9 in cell H3, and then copy this formula in cells I3, J3, …, H8, I8, J8. This should result in EXPECTED FREQUENCIES table.

Chi-square test of independence in EXCEL for Example 2 (contd.)



In Excel, type =chitest(B3:D8,H3:J8), this will result in the P-value of 0.0015 < .05, and we conclude that GAME PREFERENCE does depend on AGE GROUP.

3. CHI-SQUARE TEST OF HOMOGENEITY

Example 3. In a telephone survey, respondents were asked to indicate their level of agreement with the statement "Cigarette smoking should be banned in public places". The results are shown in the table below: SA = strongly agree, A = agree, N = neutral, D = disagree, SD = strongly disagree.

	SA	A	N	D	SD	TOTAL
F	40	38	16	37	5	136
M	16	25	11	25	11	88
TOTAL	56	63	27	62	16	224

Test if there is no difference in Males and Females with respect to their levels of agreement on the banning of smoking in public places. Let $\alpha = 0.05$.

H_0: $p_{Fi} = p_{Mi}$, i = 1, 2, 3, 4, 5

H_1: H_0 is false

In words: when H_0 is true, there is no difference in Males and Females with respect to their levels of agreement on the banning of smoking in public places.

When H_0 is TRUE:

$p_{F1} = p_{M1}$ = P(Strong Agreement, M or F) = 56/224

$O_{1,1}$ = # of females expected to strongly agree when H_0 TRUE

= Total Females x P(Strong Agreement, M or F) = 136 x 56/224

$O_{2,1}$ = # of Males expected to strongly agree when H_0 TRUE

= Total Males x P(Strong Agreement, M or F) = 88 x 56/224,

etc.

Hence

$O_{i,j}$ = Total of Row-i x Total of Column j/ Grand Total

which is exactly the same formula as in the case of chi-square test for independence. We can therefore run the test of homogeneity in exactly the same we ran the chi-square test of independence.

EXAMPLE 3 - CHI-SQUARE TEST OF HOMOGENEITY in EXCEL

To run the chi-square test of homogeneity, type the data table and copy row and column headings next to data table as shown here.

A	B	C	D	E	F	G
OBSERVED FREQUENCIES						
Gender	SA	A	N	D	SD	TOTAL
F	40	38	16	37	5	136
M	16	25	11	25	11	88
TOTAL	56	63	27	62	16	224
CIES = ROW TOTAL x COLUMN TOTAL/GRAND TOTAL						
Gender	SA	A	N	D	SD	TOTAL
F						
M						
TOTAL						

In cell B9, type =B$5*$G3/$G$5, copy this formula in all other cells to get EXPECTED FREQUENCIES, calculate ROW TOTAL and COLUMN TOTALS in EXCEL. TYPE =CHITEST(B3:F4,B9:F10) in cell B14, which gives P=.072652.

A	B	C	D	E	F	G
OBSERVED FREQUENCIES						
Gender	SA	A	N	D	SD	TOTAL
F	40	38	16	37	5	136
M	16	25	11	25	11	88
TOTAL	56	63	27	62	16	224
CIES = ROW TOTAL x COLUMN TOTAL/GRAND TOTAL						
Gender	SA	A	N	D	SD	TOTAL
	=B$5*$G3/$G$5		16.39	37.64	9.71	136
M	22.00	24.75	10.61	24.36	6.29	88
TOTAL	56	63	27	62	16	224
P	0.072652	> .05				
	DO NOT REJECT NULL					

Since P > .05, Null is not rejected. CONCLUSION: No difference in Males and Females.