Test for Independence for Two Categorical Variables

- **Observed** counts are given in the contingency table.
- **Expected counts** for each cell are computed as follows:

$$\text{exp counts} = \frac{\text{row tol} \times \text{col tol}}{\text{big tol}}$$

Our procedure for testing independence is as follows:

- **Hypotheses:** Ho: Two categorical variables are independent (NOT ASSOCIATED)

  Ha: . - - - - . — are NOT independent (ASSOCIATED)

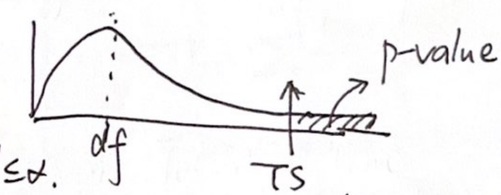
- **Test Statistic:**

$$TS, \; X^2 = \sum \frac{(obs - exp)^2}{exp} \quad (\chi^2)$$

- **p-value:**

  Chi-squared distribution w/ df = (# rows −1)×(# cols−1)

  We can use Chi-squared table to find critical value.



- **Conclusions:** Decision: Rej Ho if ~~α~~ p-val≤α.

  → VERY STRONG/STRONG/SOME/ NOT ENOUGH EVIDENDE of association.

- **DF**

  (# rows −1)×(# cols −1)

Cautions:

- Small p-value -> **strong evidence** of association, NOT evidence of a **strong association**
- Large p-value -> not enough evidence of association, NOT proof the variables are independent
- Just because quantitative data can be turned into categories does NOT mean you should use χ^2 tests for everything!

  (1 2 3 - - . 100

Pearson

Example: Let's test for independence between happiness and family income.

- State the hypotheses. H_0: Happiness INDEP Income H_a: Happines NOT IND Income (associated)
- Obtain the expected counts for each cell.

Happiness

Income	NOT Too	Pretty	Very	
Above	$\frac{423 \times 315}{1993}$ 66.86	$\frac{423 \times 1087}{1993}$ 231.13	125.01	423
Avg	139.56	482.48	260.96	883
Below	108.58	375.29	203.03	687
	315	1087	589	1993

- Compute the contribution to the test statistic for each cell and add up all the contributions to obtain the test statistic. $\frac{(obs - exp)^2}{exp}$

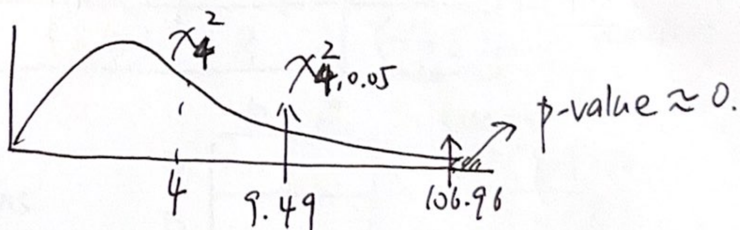
$\frac{(26-66.86)^2}{66.86}$ 24.97	$\frac{(233-231.13)^2}{231.13}$ 0.02	12.16
3.65	0.186	3.93
37.04	0.15	24.85

$$TS = \sum \frac{(obs-exp)^2}{exp}$$

$$X^2 = 24.97 + 0.02 + 12.16 + \cdots + 24.85 = \boxed{106.96}$$

- Find the p-value and state your conclusion.

$df = (\#\text{rows}-1)(\#\text{cols}-1) = (3-1)\times(3-1) = 4.$



Decision: Rej H_0 at all usual α's (sig levels)

Conclusion: Very strong evidence that there is an association between Happiness and Income.

Three ways to see the Pattern of the Relationship

- **Conditional Probabilities**

compare % NOT TOO HAPPY for
- Above Avg Income: $\frac{26}{\cancel{315}\ 423} = 6\%$
- Below Avg Income: $\frac{172}{\cancel{315}\ 687} = \cancel{25\%}\ 25\%$

compare % VERY HAPPY for
- Above Avg Income: $\frac{164}{423} = 39\%$
- Below Avg Income: $\frac{132}{687} = 19\%$

It appears that higher income is associated with higher level of happiness.

(Money makes ppl happy — TOO FAR
Association is NOT causation)

- **Residuals**

$\underline{= obs - exp}$

Where are big diff?
What is the pattern?

Income \ Happiness	not too	pretty	very
Above	26−66.86 ≈ −41	2	39
Avg	−23	−9	32
Below	63	8	−71

Higher income associated with higher levels of happiness

- **Relative Risk**

$RR = \frac{\hat{p}_1}{\hat{p}_2}$ ratio of proportions

	E	$\bar{E}$
D		
$\bar{D}$		

want: $\frac{P(D|E)}{P(D|\bar{E})}$

observed: $P(E|D)$ $P(\bar{E}|D)$
$P(E|\bar{D})$ $P(\bar{E}|\bar{D})$

RR of saying NOT TOO HAPPY for Below Avg Inc group compared to Above Avg Inc group.

$RR = \frac{25\%}{6\%} = 4.17$

Below Avg Income ppl are 4X more likely to say NOT ~~VER~~ TOO HAPPY than Above Avg Income ppl.

Degrees of Freedom

- df = (#rows – 1) x (#columns -1)
- In a contingency table, the degrees of freedom represent the number of cells of the table that are "free" to be any number, given the row and column totals.
- There are certain restrictions as to where these "free" numbers can go, and you may get negative numbers sometimes if the totals are not large enough.
- Examples:

57	43	100
93	107	200
150	150	300

$df = 1. \quad (2-1)\times(2-1)$

4	8	18	5	3	40
6	12	-8	15	7	30
10	20	10	20	10	70

$df = (2-1)\times(5-1) = 4$

2	3		10
5	1		10
1	2		10
0	4		10
			20
10	20	30	60

$df = (5-1)\times(3-1) = 8$

2	3	4	9
5	6	7	18
1	8	1	10
			10
			20
10	27	30	67

Not unique

⇒ uniquely determined
(No superfluous data)