12.2 Contingency Testing

Recall a contingency table is a 2-Dim grid displaying categories vs categories.

We now apply the techniques we just learned from 12.1, "Goodness of fit", to Contingency Tables.

Ex Heart Attacks

age \ gender	M	F	Tot
18-44	26,828	9,265	36,093
45-64	166,340	68,666	235,006
65-84	155,707	124,289	279,996
85+	35,524	57,785	93,309
Tot	384,399	260,005	644,404

Review

(a) $P(F) = \frac{\text{specific}}{\text{generic}} = \frac{\# F}{\text{Tot} \#} = \frac{260{,}005}{644{,}404}$

(b) $P(45\text{-}64) = \frac{235{,}006}{644{,}404}$

(c) $P(F \text{ and } 45\text{-}64) = \frac{68{,}666}{644{,}404}$

(d) $P(F \text{ or } 45\text{-}64) = \frac{260{,}005 + 235{,}006 - 68{,}666}{644{,}404}$

(e) $P(F \mid 45\text{-}64) = \frac{68{,}666}{235{,}006}$

(f) $P(45\text{-}64 \mid F) = \frac{68{,}666}{260{,}005}$

⊛ Test for Independence (12.2A)

Q: How can we compare counts of one group to another's?

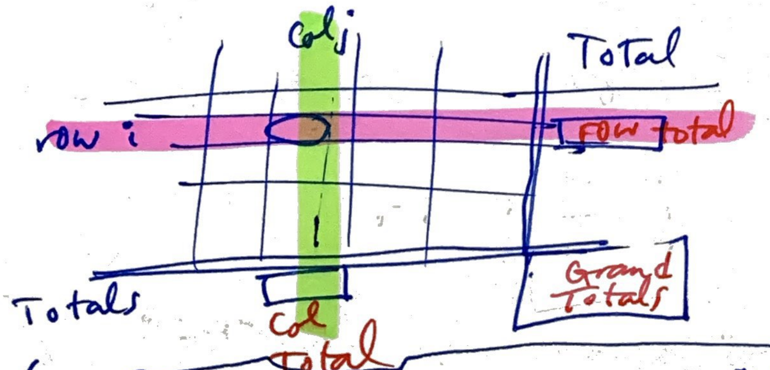
> In 12.1 (Goodness-of-Fit) we asked if a count distribution varied from the expected.

- Here in 12.2A we treat each group's counts to see if they vary from each other.

EX

> In the opening contingency table we may desire to compare Male counts to Female
>
> Q: Is the breakdown of age of male heart attacks different than the female's?

⊛ Expected Cell values:



$$\boxed{E_{ij} = \frac{(\text{row } i \text{ tot})(\text{col } j \text{ tot})}{\text{grand Total}}}$$

- $\boxed{\text{Expected count}_{\text{row } i,\, \text{col } j} = \text{G.Tot} \cdot P(\text{Row } i \mid \text{Col } j)}$.
- Expected count for row i & col j $= \left(\cancel{\text{Grand Total}}\right)\left(\frac{\text{row } i \text{ total}}{\cancel{\text{G.Tot.}}}\right)\left(\frac{\text{col } j \text{ total}}{\text{G.Tot}}\right)$

cancels out

EX Are study hrs /week distributed the same across major's?

OBSERVED

Weekly hrs \ Major	Humanities	Soc. Sci	Business	STEM	Totals
0-10	68	106	131	40	345
11-20	119	103	127	81	430
20+	70	52	51	52	225
Totals	257	261	309	173	1000

Q: Is the breakdown of study hrs the same accross majors.

Note: $P(\text{0-10hrs \& Humanities}) = \left(\frac{345}{1000}\right) \cdot \left(\frac{257}{1000}\right)$

$$\boxed{\underset{\text{row } i \text{ col } j}{\text{Expected frequency}_{ij}} = \text{totals} \cdot P(\text{row } i) \cdot P(\text{col } j)}$$

$$E_{11} = \cancel{1000}\left(\frac{345}{\cancel{1000}}\right) \cdot \left(\frac{257}{1000}\right) = 88.7$$

for cell #1

Q: Is the 68 we see in the observed data a statistical variation?

To answer these for all cells we build a table of expected cell values.

	Hum	SSci	Bus.	STEM
0-10	88.665	90.045	106.605	59.685
11-20	110.510	112.23	132.87	74.390
21+	57.825	58.725	69.525	38.925

We next compare these expected values with the observed values.

$$\boxed{\chi^2_{test} = \sum_{\substack{\text{all}\\ \text{cells}}} \frac{(O-E)^2}{E}}$$

For this example

$$\chi^2_{test} = \frac{(68-88.665)^2}{88.665} + \frac{(119-110.510)^2}{110.510}$$

$$+ \frac{(70-57.825)^2}{57.825} + \frac{(106-90.045)^2}{90.045} + \cdots$$

$$\frac{(40-59.685)^2}{59.685} + \frac{(81-74.390)^2}{74.390} + \frac{(52-38.925)^2}{38.925}$$

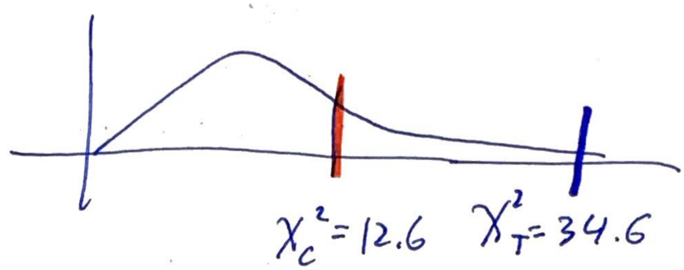
$$\boxed{\chi^2_{test} = 34.635}$$

- χ^2_{crit} : DOF = (rows − 1)(cols − 1)

$$= (3-1)(4-1) = 2 \cdot 3 = \boxed{6}$$

row 6 and $\alpha = 0.05 \Rightarrow \chi_c^2 = 12.592$

- compare $\chi^2_{test} = 34.6$ $\chi^2_c = 12.6$



- Conclude "the variation we see between the observed and the expected values is statistically significant (an anomoly)

$$\boxed{\text{So our conclusion is that the Breakdown of study hours per major varies significantly and is therefore independent of each other.}}$$

Step 7: Stat disk.

- Analys → Contingency Tables

→ populate columns with study category
((each colum is a major))

$\boxed{\text{Evaluate}}$

Results:
- pvalue 0.0000
- $\chi^2 = 34.638$ (like ours)
- Critical value χ^2_c : 12.5957
{reject if our $\chi^2 > \chi^2_c$ }

CW 12.2 ((#16 but do it on stat disk)

⊛ Tests of Homogeneity

This test is a different interpretation of the Independence Test

We now want to compare seperate populations

We record that populations as counts in a row.

Pop A in Row 1
Pop B in Row 2
etc.

Ex We want to see if the results of blood pressure drugs $\boxed{\text{effect heart attack rates}}$

Q: Does Telmisartan produce differate rates of heart attacks than the drug Ramipril?

Consider the daily dosages:

Research → Group 1: took 2 pills of Telemisartan
Control → Group 2: took 2 pills of Ramipril
$\frac{1}{2}$ Dosage → Group 3: took one of each.

Results after 24 month

Group	Drug	Fatal	non-Fatal	No attacks	TOT
1	Telmisartan	598	431	7513	8542
2	Ramipril	603	400	7573	8576
3	both	620	424	7458	8502
	TOT	1821	1255	22,544	25,620

- The expected Counts $E = \frac{(\text{row TOT})(\text{Col TOT})}{\text{Grand Tot}}$

	Fatal	non-fatal	no H.A
Pop 1	607.14	418.43	7516.43
Pop 2	609.56	420.10	7546.34
Pop 3	604.30	416.47	7481.27

- Test statistic

$$\chi^2 = \frac{(598 - 607.14)^2}{607.14} + \cdots + \frac{(7458 - 7481.27)^2}{7481.27}$$

$$\boxed{\chi^2_{test} = 2.259}$$

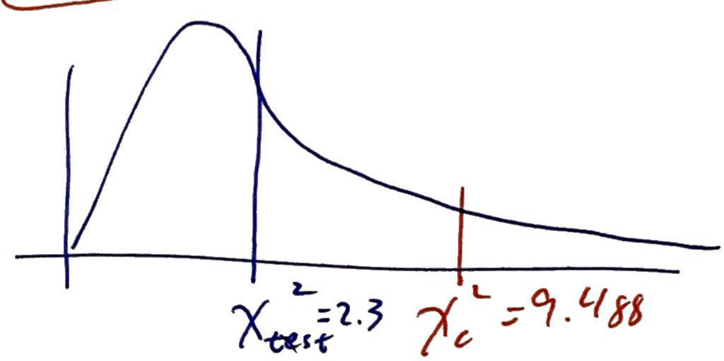
- Critical Value

$$DOF = \underset{col}{(3-1)}\underset{row}{(3-1)} = 2 \cdot 2 = \boxed{4}$$

χ^2-table : row 4 $\longrightarrow$ 9.488 ($\alpha = 0.5$)

$$\boxed{\chi^2_{crit} = 9.488}$$

- reject?



we fail to reject the null that there are no diff'ce in Heart Attack Category for these regimens.

Step 7 Stat disk

Analysis → Contingency tables

Data: But now put the comparisons of the experiment down the columns instead of along the rows.

- Select columns desiring to compare

$\boxed{\text{Evaluate}}$

Col 1	Col 2	Col 3
↓	↓	↓
598	603	620
431	400	424
7513	7573	7458

- Results

DOF = 4

$\chi_t^2 = 2.259$

$\chi_c^2 = 9.48772$

p-value = 0.6882

- plot