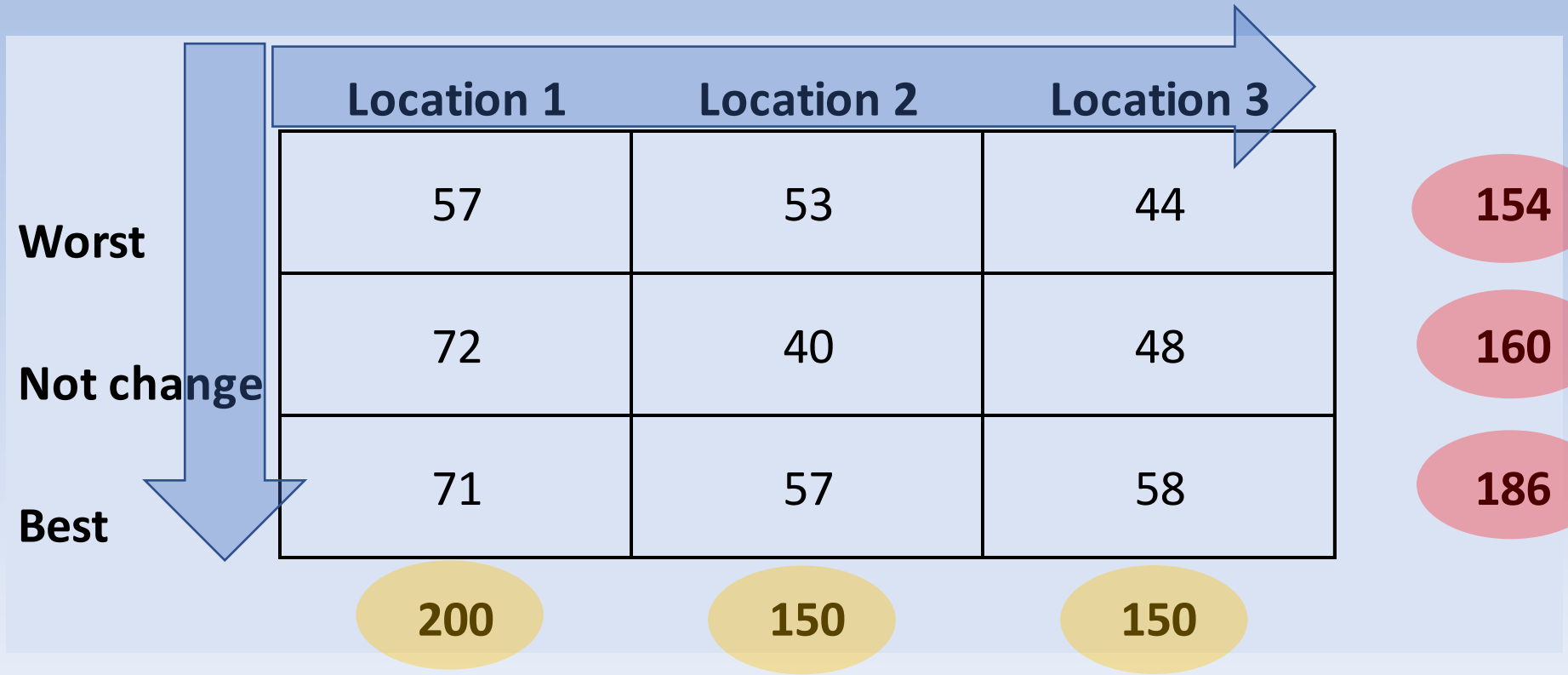
Chi - Square

Chi - Square

- allows us to decide if the differences observed between two variables can be attributed to the opportunity.
- Does not tell you the type of relationship that exists between both variables, but only that a relationship exists.
- The chi-square statistic is represented as X^2 .

Chi - Square

- First type of problem.



Chi - Square

- Second type of problem.

		Performance			
		Bad	Not change	Good	
IQ	Low	67	64	25	156
	Average	42	76	56	174
	High	10	23	37	70
		119	163	118	400

Chi - Square

- Second type of problem.

		Performance			
		Bad	Not change	Good	
IQ	Low	67	64	25	156
	Average	42	76	56	174
	High	10	23	37	70
		119	163	118	400

$$\frac{\text{Row total} \quad * \quad \text{Column total}}{\text{Grand total}}$$

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi - Square

❑ Second type of problem.

		Performance			
		Bad	Not change	Good	
IQ	Low	67 (46.41)	64 (63.57)	25 (46.02)	156
	Average	42 (51.77)	76 (70.91)	56 (51.32)	174
	High	10 (20.82)	23 (28.52)	37 (20.66)	70
		119	163	118	400

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$X^2 = \frac{(67 - 46.41)^2}{46.41} + \frac{(64 - 63.57)^2}{63.57} + \frac{(25 - 46.02)^2}{46.02}$$

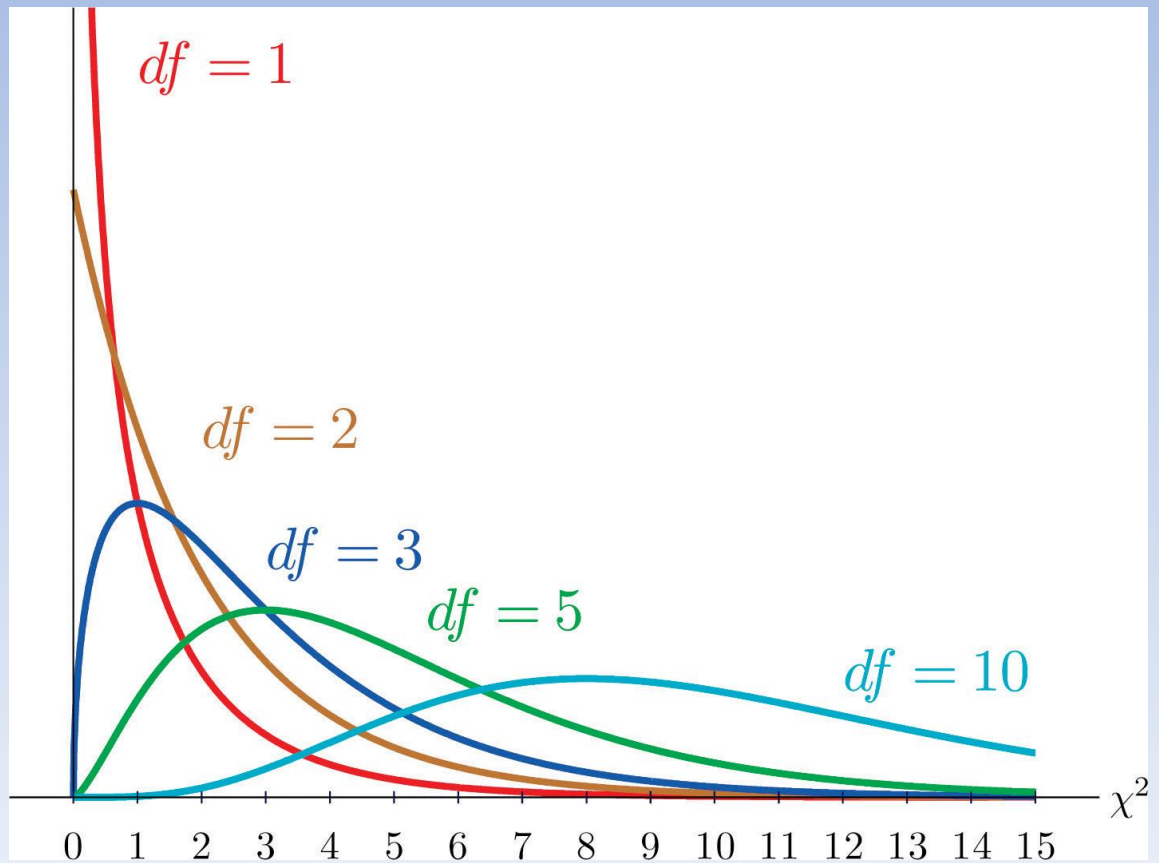
$$+ \frac{(42 - 51.77)^2}{51.77} + \frac{(76 - 70.91)^2}{70.91} + \frac{(56 - 51.32)^2}{51.32}$$

$$+ \frac{(10 - 20.82)^2}{20.82} + \frac{(23 - 28.52)^2}{28.52} + \frac{(37 - 20.66)^2}{20.66}$$

$$= \mathbf{40.89}$$

Chi - Square

- degrees of freedom.



Chi - Square

❑ degrees of freedom.

df = (r – 1) X (c – 1)
Where:
r = rows
c = columns

df = (3 – 1) X (3 – 1) = **4**

X^2 (40.89) > 13.277. We conclude that there is a relationship between the IQ and the performance of these young people in their jobs.

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1

Chi - Square

- Pearson residual.

	Thumbnail 1	Thumbnail 2	Thumbnail 3	
Click	14 (14.67)	9 (14.67)	21 (14.66)	**44**
No-click	986 (985.33)	991 (985.33)	979 (985.34)	**2,956**
	1,000	**1,000**	**1,000**	**3,000**

Chi - Square

- Pearson residual.

$$\mathbf{R} = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

	Thumbnail 1	Thumbnail 2	Thumbnail 3
Click	$\frac{14 - 14.67}{\sqrt{14.67}}$	$\frac{9 - 14.67}{\sqrt{14.67}}$	$\frac{21 - 14.66}{\sqrt{14.66}}$
No-click	$\frac{986 - 985.33}{\sqrt{985.33}}$	$\frac{991 - 985.33}{\sqrt{985.33}}$	$\frac{979 - 985.34}{\sqrt{985.34}}$

Chi - Square

- Pearson residual.

$$\mathbf{R} = \frac{\text{Observed} - \text{Expected}}{\sqrt{\text{Expected}}}$$

	Thumbnail 1	Thumbnail 2	Thumbnail 3
Click	-0.1741	-0.8660	2.5981
No-click	0.0212	0.0954	-0.2863

Chi - Square

- X^2 from Pearson residual.

$$\sum_{i}^{r} \sum_{j}^{c} R^2$$

$$\mathbf{X^2} = -0.1741^2 + -0.8660^2 + 2.5981^2 + 0.0212^2 + 0.0954^2 + -0.2863^2 = \mathbf{5.0283}$$

Chi - Square

- degrees of freedom.

$$df = (2 - 1) \times (3 - 1) = 2$$

X^2 (5.0283) < 5.991, we accept the null hypothesis. We conclude that there is no evidence to think that the click rates differ to an extent greater than chance might cause.

Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

***INTRODUCTION TO POPULATION GENETICS*, Table D.1**

Chi - Square

❑ Applications to data science.

- Multi-arm bandit algorithm.
- Determining appropriate sample sizes for web experiments.
- Are used in research by investigators in search of the statistically significant p-value.
- Resampling simulations.

Chi - Square

❑ Resampling algorithm.

Thumbnail for a YouTube channel video example:

1. Constitute a box with 44 ones (clicks) and 2,956 zeros (no clicks).
2. Shuffle, take three separate samples of 1,000, and count the clicks in each.
3. Find the squared differences between the shuffled counts and the expected counts and sum them.
4. Repeat steps 2 and 3, say 1,000 times.
5. How often does the resampled sum of squared deviations exceed the observed?. That's the p-value.

Chi - Square in Python

- scipy dot statistics package.

Test	Thumbnail 1	Thumbnail 2	Thumbnail 3
Click	14	9	21
No-click	986	991	979

```
from scipy import stats
```

```
stats.chi2_contingency(df_data, correction = True)
```

```
(5.028293763070488,
 0.0809319272850542,
 2,
 array([[ 14.66666667,  14.66666667,  14.66666667],
        [985.33333333, 985.33333333, 985.33333333]]))
```

statistics test

p-value

degrees of freedom

Expected values

p-value > 0.05 (significance level alpha chosen). We can accept the null hypothesis.