Foundations of Data Science: Hypothesis testing - eT sting for goodness-of-fit

Multiple categories

American Civil Liberties Union report in '06 jury selection in Alameda County, Ca (2010)

	Caucasian	Black/AA	Hispanic	Asian/PI	Other	Total
Population %	54	18	12	15	1	100
Observed panel numbers	780	117	114	384	58	1453
Expected panel numbers	784.62	261.54	174.36	217.95	14.53	1453.00
$\frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$	0.03	79.88	20.90	126.51	130.05	357.36

H_0: The panels were chosen by random selection from the population

H_a: The panels were chosen by some other, unspecified method.

k - groups

p_i - population proportion in the ith group

n - total size of population $\quad n = \sum_i n_i$

n_i - number in ith group

np_i - expected number in each group.

e.g. $np_1 = 100$ (5, 5%) $n_1 = 95$

$np_2 = 10$ (5, 50%) $n_2 = 5$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

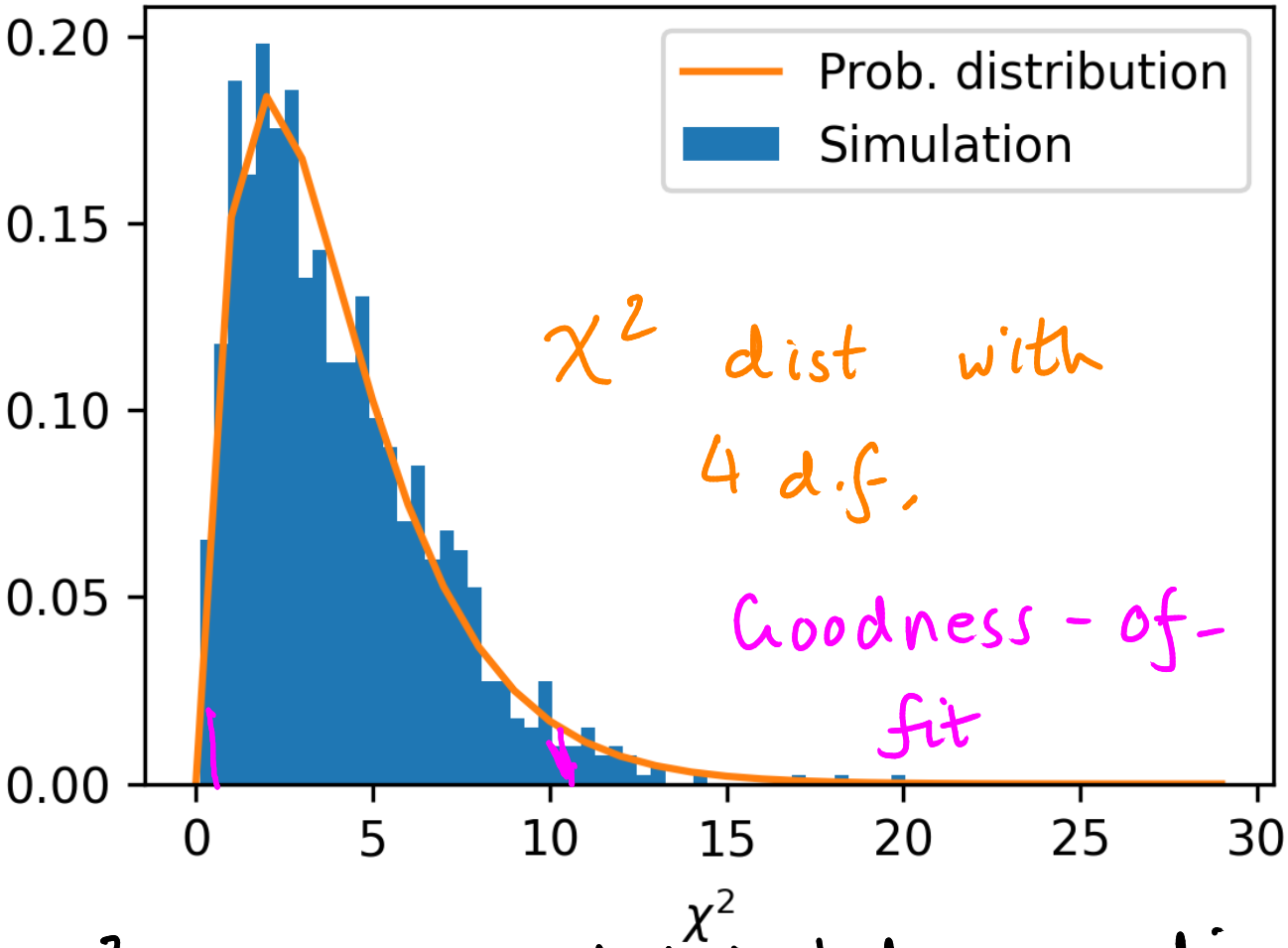
"chi - squared" In e.g. $\chi^2 = 357 - 36$

Statistical simulations

$n = 1543$

p_i

$p \simeq 0$



→ 357

χ^2 stat is distributed according to a χ^2 distribution with $k-1$ d.o.f.

2 - way contingency tables

	Female	Male	Total
Depressed	30	12	42
Not depressed	2048	1663	3711
Total	2078	1675	3753

	Population 1	Population 2	Total
Category 1	n_{11}	n_{12}	$n_{1\bullet}$
Category 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

H_0 : Being severely depressed is independent of being female or male

H_a : Some other, unspecified hypothes.

$$H_0 : P(X = x \mid Y = y) = P(X = x)\, P(Y = y)$$

$$p_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \qquad p_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

$$\Rightarrow \quad \hat{e}_{ij} = n_{\bullet\bullet}\, p_{i\bullet}\, p_{\bullet j} = \frac{n_{i\bullet}\, n_{\bullet j}}{n_{\bullet\bullet}}$$

Multiway contingency tables

$J = 2$

$I = 2$

	Female	Male	Total
Depressed	30	12	42
Not depressed	2048	1663	3711
Total	2078	1675	3753

	Population 1	Population 2	Total
Category 1	n_{11}	n_{12}	$n_{1\bullet}$
Category 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

	Female	Male
Depressed	23.25	18.75
Not depressed	2054.75	1656.25

	Population 1	Population 2
Category 1	$\hat{e}_{11}$	$\hat{e}_{12}$
Category 2	$\hat{e}_{21}$	$\hat{e}_{22}$

$$\chi^2 = \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$$

$$= 4.433$$

$$\#\ \text{d.f} = (I-1)(J-1) + 1 = 1$$

$$p = 0.035$$