Probaliy Theory and Statistics

Lecture 8: Chi-squared tests for homogeneity and independence

22 May 2025

Lecturer: Batashov Ruslan

Email: rusbatashov@gmail.com

Chi-squared test for homogeneity (two-way χ^2-test)

Suppose there are several discrete populations $X^{(1)}, X^{(2)}, \ldots, X^{(r)}$ sampled independently. We want to check if their distributions are equal.

Example

The table below shows investment preferences of 100 young (30 years old or less) and 150 adult (above 30) individual investors. Do young and adult investors have the same investment preferences? Use 5% significance level.

	Stocks	Bonds	ETFs
Young	39	28	33
Adult	55	63	32

Null and alternative hypotheses

There are several populations $X^{(1)}, X^{(2)}, \ldots, X^{(r)}$ sampled independently. Each random variable $X^{(i)}$ can take on values

$$a_1, a_2, \ldots, a_c.$$

We want to test the null hypothesis that all $X^{(i)}$ have the same distribution:

H_0: $P(X^{(1)} = a_j) = P(X^{(2)} = a_j) = \ldots = P(X^{(r)} = a_j)$ for each j.

In other words,

H_0 :

	a_1	a_2	$\ldots$	a_c
$X^{(1)}$	p_1	p_2	$\ldots$	p_c
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X^{(r)}$	p_1	p_2	$\ldots$	p_c

The alternative hypothesis is that at least one equality does not hold.

A contingency table

The table below, which represents the data, is called a contingency table.

	Category 1	Category 2	...	Category c
Group 1	O_{11}	O_{12}	...	O_{1c}
Group 2	O_{21}	O_{22}	...	O_{2c}
...	...	...	...	...
Group r	O_{r1}	O_{r2}	...	O_{rc}

r rows (groups *or* populations), c columns (categories *or* values)

O_{ij} = number of observed elements in group i with value j

A contingency table

The table below, which represents the data, is called a contingency table.

Observed (Expected)	Category 1	Category 2	...	Category c
Group 1	$O_{11}(E_{11})$	$O_{12}(E_{12})$	...	$O_{1c}(E_{1c})$
Group 2	$O_{21}(E_{21})$	$O_{22}(E_{22})$	...	$O_{2c}(E_{2c})$
...	...	...	...	...
Group r	$O_{r1}(E_{r1})$	$O_{r2}(E_{r2})$	...	$O_{rc}(E_{rc})$

r rows (groups *or* populations), c columns (categories *or* values)

O_{ij} = number of observed elements in group i with value j

E_{ij} = number of expected elements in group i with value j

How to find E_{ij}?

Explanation of the formula for E_{ij}

If H_0 is true, then

Observed	a_1	$\dots$	a_j	$\dots$	
$X^{(1)}$	O_{11}	$\dots$	O_{1j}	$\dots$	
$X^{(2)}$	O_{21}	$\dots$	O_{2j}	$\dots$	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
$X^{(i)}$	$\dots$	$\dots$	$O_{ij}(E_{ij})$	$\dots$	$O_{i\bullet}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
$X^{(r)}$	O_{r1}	$\dots$	O_{rj}	$\dots$	
			$O_{\bullet j}$		

The χ^2-statistic

$$\chi^2_{st} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

n = total number of elements in the samples

$E_{ij} = \dfrac{O_{i\bullet} O_{\bullet j}}{n}$ – estimates of expected counts

$O_{i\bullet} = \sum_{j=1}^{c} O_{ij}$ observed counts of elements in row i

$O_{\bullet j} = \sum_{i=1}^{r} O_{ij}$ observed counts of elements in column j

Theorem

If H_0 is true, then

$$\chi^2_{st} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

has approximately the χ^2-distribution with $(r-1)(c-1)$ degrees of freedom.

Condition: each $E_{ij} \geqslant 10$.

Rejection region: reject H_0 if $\chi^2_{st} > \chi^2_\alpha((r-1)(c-1))$.

$p - value = P(V > \chi^2_{st})$, where $V \sim \chi^2((r-1)(c-1))$

Why do we have $(r-1)(c-1)$ degrees of freedom?

Why do we have $(r-1)(c-1)$ degrees of freedom?

	a_1	$\dots$	a_{c-1}	a_c
$X^{(1)}$	$p_{11} \approx \widehat{p}_1$	$\dots$	$p_{1,c-1} \approx \widehat{p}_{c-1}$	p_{1c}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$X^{(r)}$	$p_{r1} \approx \widehat{p}_1$	$\dots$	$p_{r,c-1} \approx \widehat{p}_{c-1}$	p_{rc}

$$r(c-1) - (c-1) = (r-1)(c-1)$$

Chi-squared test for independence

Suppose each object in a population has two characteristics, X and Y. We want to check whether X and Y are independent.

We have two random variables X, Y with possible values

$$X : a_1, \ldots, a_r, \qquad Y : b_1, \ldots, b_c$$

and a sample

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

H_0: X and Y are independent
H_1: X and Y are not independent

Data is represented by a contingency table.

$X \backslash Y$	$Y = b_1$	$Y = b_2$	$\ldots$	$Y = b_c$
$X = a_1$	O_{11}	O_{12}	$\ldots$	O_{1c}
$X = a_2$	O_{21}	O_{21}	$\ldots$	O_{2c}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X = a_r$	O_{r1}	O_{r2}	$\ldots$	O_{rc}

Because the independence means

$$P(Y = b_j \mid X = a_i) = P(Y = b_j) \qquad \text{for all } a_i, b_j$$

the test for independence should probably look very similar to the test for homogeneity...

Theorem

If H_0 is true, then

$$\chi^2_{st} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

has approximately the $\chi^2((r-1)(c-1))$ distribution.

Here, same as above, $E_{ij} = \dfrac{O_{i\bullet} O_{\bullet j}}{n}$.

Condition: each $E_{ij} \geqslant 10$.

Rejection region: reject H_0 if $\chi^2_{st} > \chi^2_{\alpha}((r-1)(c-1))$.

Intuition behind n_{ij} and $(r-1)(c-1)$

If $X \perp\!\!\!\perp Y$, then the joint distribution is $p_{ij} = p_i^X p_j^Y$.

- There are $k = rc$ possible values of the pair (x_i, y_j)
- We test the hypothesis that $H_0 \colon p_{ij} = \widehat{p}_i^X \cdot \widehat{p}_j^Y$

It should be $rc - (r - 1 + c - 1) - 1 = (r-1)(c-1)$ d.o.f.

Example

Is there dependence between regions and incomes at 5% significance level?

	Income, thousand \$				
Region	< 5	5 – 10	10 – 15	> 15	Total
South	28	42	30	24	124
North	44	78	78	76	276
Total	72	120	108	100	400

H_0: region (X) and income (Y) are independent

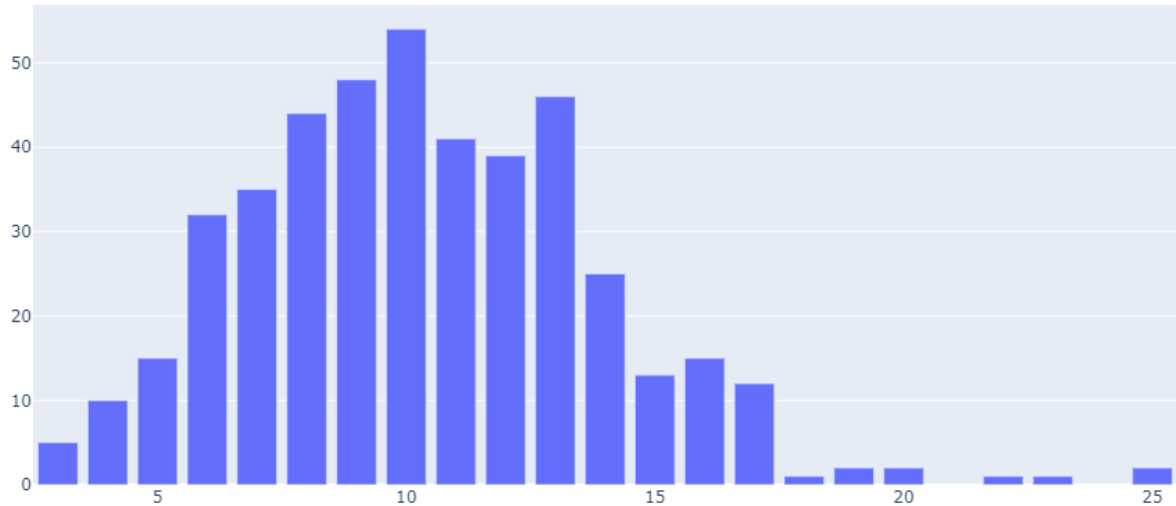
H_1: region (X) and income (Y) are dependent

Example: Blockchain

https://ergoscan.io/blocks

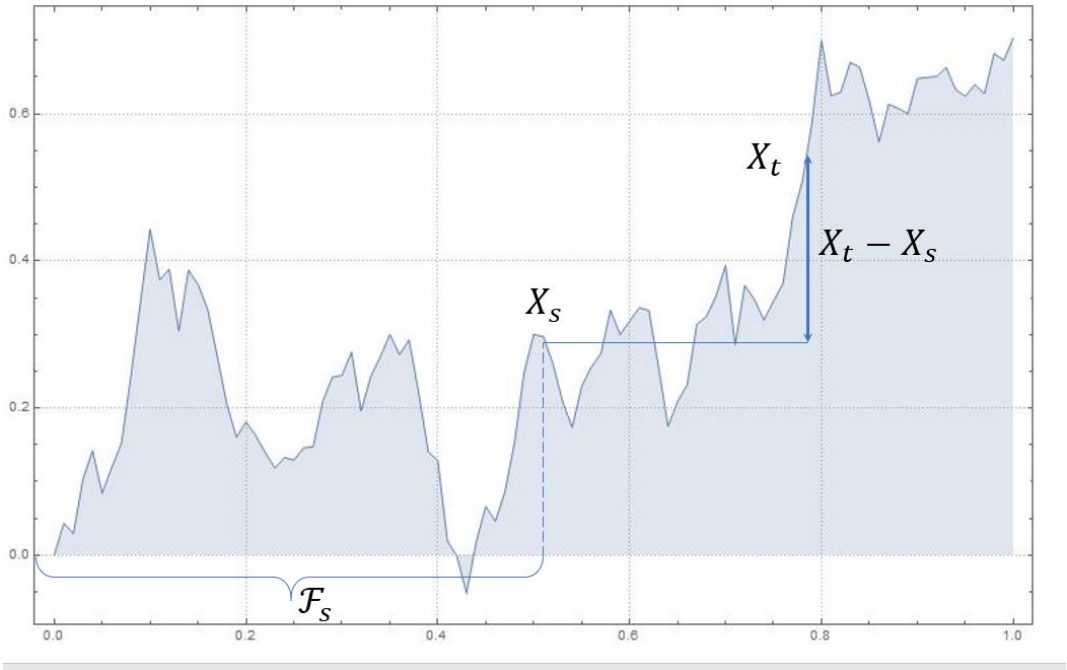
For each block implemented miners get a fixed reward.

More blocks implemented $\Rightarrow$ higher total reward.



What will be the distribution of number of blocks in an hour for Ergo coin?

Example: Independent increments of stock price



Suppose we've built a model $(X_t)_{t\geqslant 0}$ of stock price. One important property of the model is independent increments, i.e. for fixed times $t > s$ we need to know whether increment $X_t - X_s$ independent of X_s. How to check this?

Reading

Newbold, Carlson, Thorne: § 14.3

Mann: § 11.3