NON-PARAMETRIC STATISTICAL TESTS
DISTRIBUTION TESTS

The problem:

Our RV X belongs to a statistical population and has an unknown (cumulative) distribution $f(x)$ $\big(F(x)\big)$. How may we find it?
We have to construct a histogram — a vertical bar chart constructed in a special way — we form a sample with the size n:

$$x_1, x_2, \ldots, x_n$$

(it's "better" to have $n \geq 30$).
The values of the RV are grouped into several non-overlapping intervals or *class intervals* — usually all the intervals are of equal width.
All the values which belong to a given class are assigned with the value corresponding to the midpoint of the class (class mark).
Such a table is called the *contingency table*.
The number of classes k — ,,empirical formulae":

$$k \leq 5 \ln n, \quad k = 1 + 3,322 \ln n, \quad k = \sqrt{n}, \quad 2^{k-1} \leq n \leq 2^k$$

or the following table may be of some help:

the following table may be of some help:

sample size n	No of classes, k
30– 60	6– 8
60– 100	7–10
100– 200	9–12
200– 500	11–17
500–1500	16–25

The **class** limits (boundaries) must be designed in such a way that a given value x_i can be clearly **class**ified in a univocal manner, e.g. the classes can be constructed as intervals of the $[a,b)$ type – that is

$$(a,b] = \{x:\ a < x \leq b\}.$$

Or we may use a system which takes into account the accuracy (errors) of the data (cf. the following example).

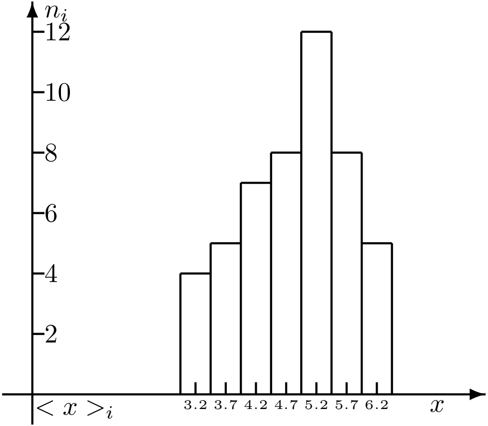
The *(class) frequencies* n_i are the numbers of the values that are in given class intervals. We have

$$\sum_{i=1}^{k} n_i = n$$

We may also use the *relative frequencies*: $f_i = \frac{n_i}{n}$.

The pairs: $\{<x>_i, n_i\}$ or $\{<x>_i, f_i\}$ form the HISTOGRAM — a vertical bar chart, with the heights of the bars proportional to n_i (or f_i).

A histogram:



the case:

49 values, $x_{min} = 3.0$, $x_{max} = 6.4$, measured with accuracy $\Delta = 0.1$.
We choose $a_0 = x_{min} - \frac{1}{2}\Delta =$ 2.95; $k =$ 7; class width $= \frac{6.4-3.0}{7} \approx 0.5$

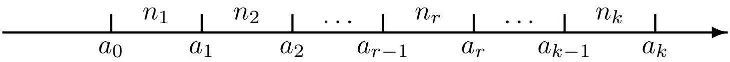
$a_1 = 3.45 \quad a_2 = 3.95 \quad a_3 = 4.45 \quad a_4 = 4.95 \quad a_1 = 5.45 \quad a_6 = 5.95 \quad a_7 = 6.45$

The contigency table will be:

class No.	1	2	3	4	5	6	7
class mark	3.45	3.95	4.45	4.95	5.45	5.95	6.45
class frequency	4	5	7	8	12	8	5

On the basis of the histogram (or other conjectures) we form a hypothesis H_0: {The RV follows a cumulative distribution $F_0(x)$} or – shortly: $H_0 : \;\; F_0(x)$.
We have at our disposal:

1. empirical frequencies (from the sample): $n_1, n_2, \ldots, n_k$



2. theoretical frequencies $(n_k)_{theor}$:

$$\mathcal{P}(X \in < \text{class} >_r) = \mathcal{P}(a_{r-1} \leq x < a_r) = F_0(a_r) - F_0(a_{r-1}) \equiv \pi_r$$

$$(n_r)_{theoretical} = n\pi_r$$

TEST STATISTIC T —

— the sum of squares of the differences between the empirical (measured) and theoretical (calculated **on the basis of the validity of** H_0) frequencies properly normalised (divided by theoretical frequencies):

$$T \equiv \chi^2_{exp} = \sum_{r=1}^{k} \frac{(n_r - n\pi_r)^2}{n\pi_r}$$

The T RV has the χ^2 distribution with $k-1$ degrees of freedom. With α fixed, the critical region is determined by the appropriate quantile of χ^2 distribution:

$$\left[\chi^2(1-\alpha, k-1), +\infty\right)$$

If the calculated $T = \chi^2_{exp}$ enters this region the hypothesis H_0 is to be rejected.

Example: see the next page

Example:

The famous father of engineering genetics, G. Mendel classified $n = 556$ peas according to two traits: shape (round versus wrinkled) and colour (green versus yellow). Each seed was put into one of $k = 4$ categories: $C_1 = ry$, $C_2 = rg$, $C_3 = wy$, and $C_1 = wg$, where r, w, g, and y denote *round*, *wrinkled*, *green*, and *yellow*, respectively. Mendel's results can be summarised in the table:

Seed type	**Frequency**
ry	315
wy	101
rg	108
wg	32

Mendel's theory predicted that the frequency counts of the seed types ry:rg:wy:wg, should occur in the ratio 9:3:3:1. Test this theory against the experimental data using the Pearson's test.

Solution: see the next page

Example – solution:

Seed type	**Frequency** experiment	**Frequency** theory
ry	315	313
wy	101	104
rg	108	104
wg	32	35

$$\chi^2_{exp} = \sum_{r=1}^{4} \frac{(\mathsf{F}_{exp} - \mathsf{F}_{th})^2}{\mathsf{F}_{th}} = \ldots = 0,51.$$

The critical region: $\chi^2_{0.95}$ for $4-1=3$ degrees of freedom is 7.8!! There is a strong suspicion that Mendel, unfamiliar with statistics, tried to improve his experimental data. They are simply too good to be true.

Case if the probabilities are not completely specified

Problem: we try to assert the statement: the number of packets X per unit time arriving at a computer network node has a Poisson distribution, with parameter λ.
We have $P(X = i) = e^{-\lambda}\lambda^i/i!$.

Validation: we record numbers of arriving packets for each of 150 time intervals. But — we don't know λ (yet). We shall use the data analyzed to find it!!!

We collect data from 150 time-intervals;
the total number of packets that arrived during those 150 time-units was 600.
Therefore we put: $\lambda = 600/150 = 4$.

The results can be arranged in a form of a . . .

No. of packets per unit time i	observed n_{obs}	total number of packets	Poisson $\mathcal{P}(X = i)$ for $\lambda = 4$	theory n_{th}
0	2	0	0,018315639	2,75
1	10	10	0,073262556	10,99
2	25	50	0,146525111	21,98
3	25	75	0,195366815	29,31
4	29	156	0,195366815	29,31
5	15	75	0,156293452	23,44
6	14	84	0,104195635	15,63
7	13	91	0,059540363	8,93
8	5	40	0,029770181	4,47
9	1	9	0,013231192	1,98
≥ 10	1	10	0,008132243	1,22
Total	150	600	1	150

...table, cntd:

No. of packets per unit time i	observed n_{obs}	total number of packets	Poisson $\mathcal{P}(X=i)$ for $\lambda=4$	theory n_{th}
0 or 1	12	10	0,091541694	13,74
2	25	50	0,146525111	21,98
3	25	75	0,195366815	29,31
4	29	156	0,195366815	29,31
5	15	75	0,156293452	23,44
6	14	84	0,104195635	15,63
7	13	91	0,059540363	8,93
≥ 8	7	59	0,051133616	7,67
Total	150	600	1	150

We combined the first two and the last three rows – a rule of thumb says that we should not have cell contents less than (or equal to) five.

Thus we have:

$$\chi^2_{exp} = \sum_{r=1}^{8} \frac{(n_{obs} - n_{th})^2}{n_{th}} = \ldots = 9.6$$

The critical region: $\chi^2_{0.95}$ for $8 - 1 - 1 = 6$ degrees of freedom is 12.59. Mark: **we subtract 2 from 8** (number of cells) not one (as in the former case), because we lost one of the degrees of freedoms of our sample calculating λ.

The Poisson distribution looks as a quite adequate fit fo the data!