Metody probabilistyczne i statystyka, 2021
informatyka algorytmiczna, WIiT PWr

7-Statistical Inference 2

Testing a distribution

Problem:
How to check that a source A has probability distribution D?

So far we have been only learning unknown parameters while distribution was known.

This is a crucial question! E.g. in

- **understanding social networks**
- **anomaly detection in cybersecurity**

Testing a distribution

Chi Square test

Goal: testing whether a source is distributed according to a hypothesis H0

Method:

- ❑ split the population to N bins
- ❑ count how many samples fall into each bin
- ❑ compute the expected value for each bin under H0
- ❑ calculate statistics
- ❑ make a decision

Chi Square test

Chi Square test

Statistics used:

$$\chi^2 = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}.$$

Obs(k) **= number of samples in bin** ***k***
Exp(k) **= expected number of samples in bin** ***k***

One sided statistics: low values for accepting H0
high value – for rejection

$$R = [\chi^2_\alpha, +\infty),$$

Chi Square test

Chi Square background

Pearson's Theorem

χ^2 **distribution for N bins converges to the chi-square distribution with N-1 degrees of freedom**

Rule of thumb: each bin should contain at least 5 samples

Chi Square background

Chi Square application example

testing a die to be unbiased:

- **6 bins corresponding to 6 possible outcomes**
- **90 samples**
- **Exp(i)=90/6=15**
- **Counts observed: 20,15,12,17,9,17**
- **Statistics:**

$$\chi^2_{\text{obs}} = \sum_{k=1}^{N} \frac{\{Obs(k) - Exp(k)\}^2}{Exp(k)}$$

$$= \frac{(20-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(17-15)^2}{15} + \frac{(9-15)^2}{15} + \frac{(17-15)^2}{15} = 5.2.$$

Chi Square application example

interpretation

ν (d.f.)	α, the right-tail probability													
	.999	.995	.99	.975	.95	.90	.80	.20	.10	.05	.025	.01	.005	.001
1	0.00	0.00	0.00	0.00	0.00	0.02	0.06	1.64	2.71	3.84	5.02	6.63	7.88	10.8
2	0.00	0.01	0.02	0.05	0.10	0.21	0.45	3.22	4.61	5.99	7.38	9.21	10.6	13.8
3	0.02	0.07	0.11	0.22	0.35	0.58	1.01	4.64	6.25	7.81	9.35	11.3	12.8	16.3
4	0.09	0.21	0.30	0.48	0.71	1.06	1.65	5.99	7.78	9.49	11.1	13.3	14.9	18.5
5	0.21	0.41	0.55	0.83	1.15	1.61	2.34	7.29	9.24	11.1	12.8	15.1	16.7	20.5

Chi Square test for independence

Goal: test the hypothesis that two parameters of the sample are stochastically independent

Application: eliminating false claims about dependence
example: eating cabbage influence cancer risk

Chi Square test for independence

- split population A to some number of bins: A_0, A_1,…, A_k
- split population B to some number of bins B_0, B_1,…, B_m
- collect samples
- calculate statistics based on a sample

Chi Square test for independence

Chi Square test for independence

	B_1	B_2	$\cdots$	B_m	row total
A_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	n_{k1}	n_{k2}	$\cdots$	n_{km}	$n_{k\cdot}$
column total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	$n_{\cdot\cdot} = n$

$$\widehat{P}\{x \in A_i \cap B_j\} = \frac{n_{ij}}{n}, \qquad \widehat{P}\{x \in A_i\} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}, \qquad \widehat{P}\{x \in B_j\} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{\cdot j}}{n}.$$

Chi Square test for independence

$$\widehat{P}\{x \in A_i \cap B_j\} = \frac{n_{ij}}{n}, \quad \widehat{P}\{x \in A_i\} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n}, \quad \widehat{P}\{x \in B_j\} = \sum_{i=1}^{k} \frac{n_{ij}}{n} = \frac{n_{\cdot j}}{n}.$$

$$\widetilde{P}\{x \in A_i \cap B_j\} = \left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right)$$

$$\widehat{\mathrm{Exp}}(i,j) = n\left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right) = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$$

$$\chi^2_{\mathrm{obs}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left\{Obs(i,j) - \widehat{\mathrm{Exp}}(i,j)\right\}^2}{\widehat{\mathrm{Exp}}(i,j)}.$$

Chi Square test for independence

$$\chi^2_{\text{obs}} = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left\{ Obs(i,j) - \widehat{\text{Exp}}(i,j) \right\}^2}{\widehat{\text{Exp}}(i,j)}.$$

Number of degrees of freedom for Chi-square distribution:

$$km - (k + m - 1) = (k-1)(m-1)$$

motivation: ***k*** **equations for computing** $n_{i\cdot}$
m **equations for computing** $n_{\cdot j}$
but only ***m+k-1*** **independent**

Chi Square test for independence

Bootstrapping

Problem:
Given a sample:

- ❑ **we know how to estimate the variation of the population – estimator for variation**
- ❑ **but how to estimate variation of this estimator?**

Bootstrapping

- Straightforward way: repeat sampling, get many values of the estimator and treat them as samples

- Requires resampling many times!

Bootstrapping

solution by Bradley Efron, based on principle from a tale story about Baron Münchausen

- **applies for any function *h* computed for the population and calculated in the same way from a sample**

- **example: variance, median**

Bootstrapping

Problem with variance of variance estimator:

Computing it on the whole population is computationally infeasible

e.g.: for a population of 100 objects and sample of size 10

there are $\binom{100}{10} \approx +\infty$ cases

Bootstrapping

method:

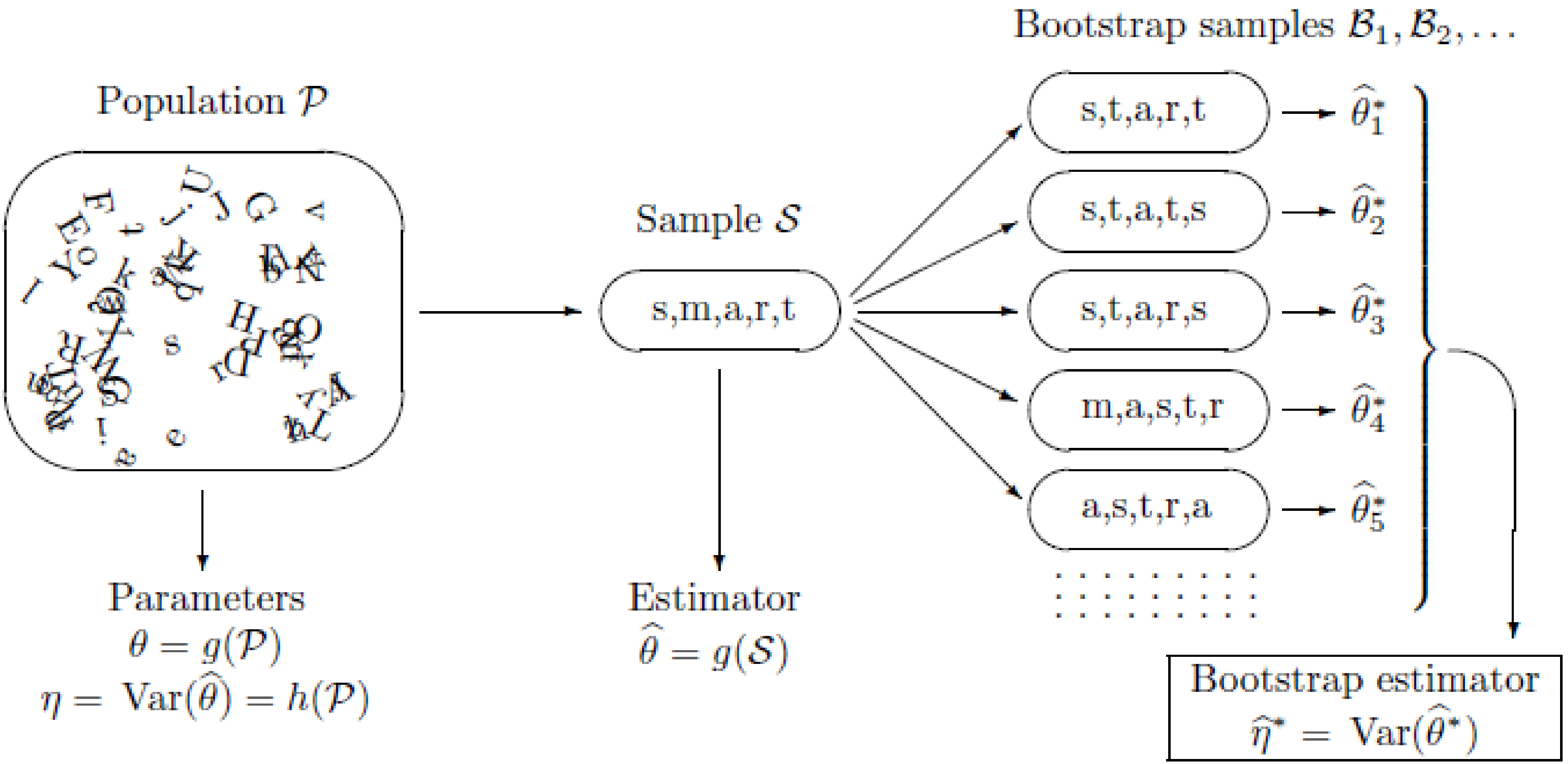
1. **take a sample of size *N*: $X_1, \ldots, X_n$**
2. **k times sample from $X_1, \ldots, X_n$ with replacement:**

$$X_{ij}^* = \begin{cases} X_1 & \text{with probability } 1/n \\ X_2 & \text{with probability } 1/n \\ \cdots & \cdots \quad \cdots \quad \cdots \quad \cdots \\ X_n & \text{with probability } 1/n \end{cases}$$

and compute the estimator from the sample obtained

Bootstrapping

Example:



Bootstrapping

Example: variance of median variance estimator

Small sample: 2,5,7, median 5

i	$\mathcal{B}_i$	$\widehat{M}_i$	i	$\mathcal{B}_i$	$\widehat{M}_i$	i	$\mathcal{B}_i$	$\widehat{M}_i$
1	$(2,2,2)$	2	10	$(5,2,2)$	2	19	$(7,2,2)$	2
2	$(2,2,5)$	2	11	$(5,2,5)$	5	20	$(7,2,5)$	5
3	$(2,2,7)$	2	12	$(5,2,7)$	5	21	$(7,2,7)$	7
4	$(2,5,2)$	2	13	$(5,5,2)$	5	22	$(7,5,2)$	5
5	$(2,5,5)$	5	14	$(5,5,5)$	5	23	$(7,5,5)$	5
6	$(2,5,7)$	5	15	$(5,5,7)$	5	24	$(7,5,7)$	7
7	$(2,7,2)$	2	16	$(5,7,2)$	5	25	$(7,7,2)$	7
8	$(2,7,5)$	5	17	$(5,7,5)$	5	26	$(7,7,5)$	7
9	$(2,7,7)$	7	18	$(5,7,7)$	7	27	$(7,7,7)$	7

Bootstrapping

Example: variance of median variance estimator

Small sample: 2,5,7, median 5

$$
\begin{aligned}
\widehat{\mathrm{Var}}^*(\widehat{M}) &= h(\mathcal{S}) = \sum_x x^2 P^*(x) - \left(\sum_x x P^*(x)\right)^2 \\
&= (4)\left(\frac{7}{27}\right) + (25)\left(\frac{13}{27}\right) + (49)\left(\frac{7}{27}\right) - \left\{(2)\left(\frac{7}{27}\right) + (5)\left(\frac{13}{27}\right) + (7)\left(\frac{7}{27}\right)\right\}^2 \\
&= \underline{3.303}. \qquad \Diamond
\end{aligned}
$$