Sequential pattern recognition. Wald's analysis.

Petr Pošík

Czech Technical University in Prague
Faculty of Electrical Engineering
Dept. of Cybernetics

Motivation: Statistical sampling plans(Statistická přejímka) **2**
Sampling plans . . . 3
CUSUM diagrams . . . 5

Sequential analysis **6**
Sequential analysis . . . 7
Sequential decision problem . . . 8
SPRT (Sequential Probability Ratio Test) . . . 9
SPRT: Thresholds A and B . . . 10
SPRT: Final suggestions . . . 11

Conclusions **12**
Summary . . . 13
Reference . . . 14

Motivation: Statistical sampling plans
(Statistická přejímka)

Sampling plans

Example situation:

- Our company builds machines. From our subcontractor, we buy spindles with the nominal diameter 7.5 mm in batches containing 10 thousands pieces. The real diameters of the components are surely different from the nominal value. How can we decide whether the batch is of acceptable quality, or that we decline to accept it and return it to the manufacturer?

Options:

- 100% control: not economical, impossible when destructive tests needed.
- Statistical sampling plans:
  - Measure (test) only a limited number of pieces.
  - Induce the quality of the whole batch from these measurements.
  - We save time, labour, and money.
  - Fundamental question question: How to determine the required number sample size to test, to be sufficiently sure when accepting/declining a batch?
  - Two possible errors:
    - We decline a good batch (error of the 1st kind, probability α)
    - We accept a bad batch (error of the 2nd kind, probability β)

Sampling plans (cont.)

Classic hypotheses testing puts in relation

- the effect size, i.e. the difference between
  - the null hypothesis H_0 (agreement with the specifications, e.g. $D = 7.5$ mm) and
  - the alternative hypothesis H_1 (unacceptable difference, e.g. $D = 7.505$ mm),
- the acceptable probabilities of errors of the 1st and 2nd kind (α a β) and
- the sample size N.

The relations:

- The lower number of errors we require, the larger sample size we need (to be 100% sure, we would need to check the whole batch).
- The smaller the difference we want to detect, the larger sample size is required.

Fixed sampling plan:

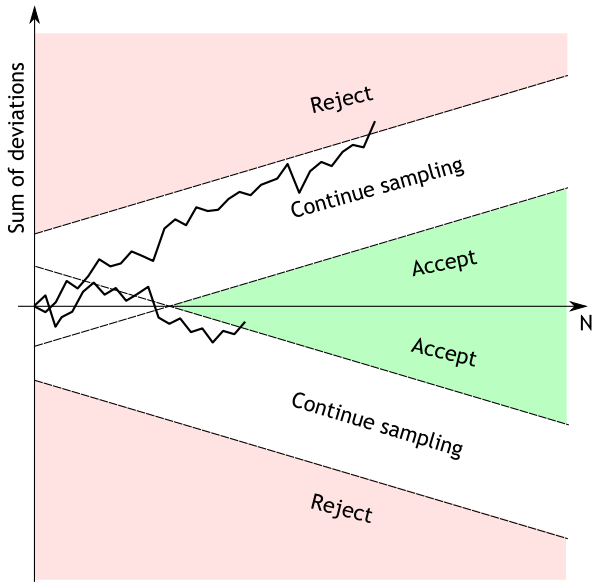
- Based on classic hypothesis testing.
- Given the differnce from specification we want to detect, and the probabilities α (and possibly β), it is possible to derive the sample size N we need to test.
- Based on N measurements we decide, whether the batch is acceptable or not.

Sequential sampling plan:

- A single measurement is carried out.
- If the measurements accumulated so far provide sufficient "proofs", accept/decline the batch, otherwise, obtain a measurement for another piece.

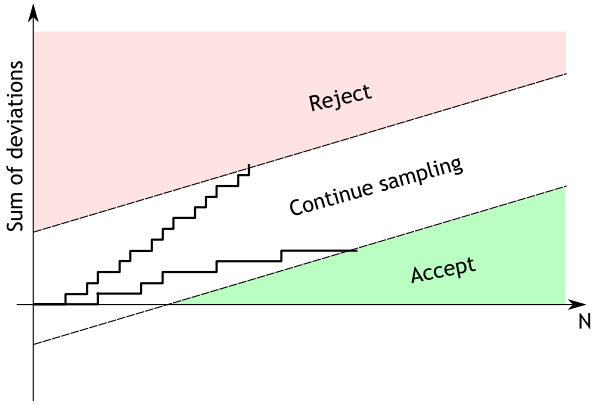
CUSUM diagrams

Two-sided sequential test of hypothesis about the mean:



- The cumulative sum (CUSUM) of the differences can increase and decrease, i.e. the differences from specifications may be both positive and negative.

One-sided sequential test of hypothesis about the population propability:



- The cumulative sum of the differences may only grow (the number of non-conforming units).

Sequential analysis 6 / 14

Sequential analysis

- Subfield of statistics and machine learning.
- The way the analysis is carried out depends in some way on the results of previous steps:
  1. selection of the experiment, measurement, or test that should be carried out next,
  2. terminate/continue with the analysis

Advantages:

- Testing can be shorter than using the classic analysis of the whole sample.
- The individual tests do not have to be of the same type!
- The diagnostic plan may be modified depending on the results of preceding steps. Compare:
  - We would like to determine if a patient suffers from cancer. We carry out all the following tests: blood analysis, X-ray scan, CT, magnetic resonance, ultrasound scan, ... Based on the results of all these measurements we decide if it is the cancer or not.
  - We would like to determine if a patient suffers from cancer. We carry out the blood analysis. If there are no abnormalities, we decide the patient does not have cancer. Otherwise, depending on the type of abnormality, we perform either a X-ray scan, CT, magnetic resonance, ...[a]

[a] This plan is not correct from the medicine point of view; it is just used as an example.

Sequential decision problem

- Object $\boldsymbol{x}$ belongs to one of two classes $\{-1, +1\}$.
- We are given an ordering of measurements $(x_1, \ldots, x_m)$ on object $\boldsymbol{x}$.
- A sequential decision strategy is a set of decision functions $S = \{S_1, \ldots, S_m\}$.
- Each decision function $S_i : \{x_1, \ldots, x_i\} \to \{-1, +1, \#\}$.

Sequential decision strategy S:

- In step i, it uses the function S_i which on the basis of measurements $x_1, \ldots, x_i$ either decides that $\boldsymbol{x}$ belongs to one of the two classes $\{-1, +1\}$, or provides decision # ("I do not know yet").
- If S_i decides for #, the strategy carries out the next measurement x_{i+1} and uses the next function S_{i+1}.
- It is characterized by the errors of the 1st and 2nd kind (α_S and β_S) and by the *expected time to decision*

$$\bar{T}_S = E(T_S(\boldsymbol{x})),$$

where the time to decision T_S is given as

$$T_S(\boldsymbol{x}) = \arg\min_i (S_i(\boldsymbol{x}) \neq \#).$$

Optimal sequential decision strategy:

$$S^* = \arg\min_S \bar{T}_S \qquad (1) \qquad \text{subject to} \quad \alpha_S \leq \alpha \quad \text{and} \quad \beta_S \leq \beta.$$

SPRT (Sequential Probability Ratio Test)

- Basic and the most important method of sequential analysis [Wal47].
- Object $\boldsymbol{x}$ has a hidden state (class) $y \in \{-1, +1\}$; this state is not known and we shall estimate it based on a sequence of measurements $x_1, \ldots, x_m$.
- We know all joint probability distributions $p(x_1, \ldots, x_m | y = c)$.
- Let's specify the hypotheses and their corresponding errors: $H_0 : y = +1 \quad \alpha = P(S(\boldsymbol{x}) = -1 | y = +1)$, $H_1 : y = -1 \quad \beta = P(S(\boldsymbol{x}) = +1 | y = -1)$
- Let's define a likelihood ratio R_m:

$$R_m(\boldsymbol{x}) = \frac{p(x_1, \ldots, x_m | y = -1)}{p(x_1, \ldots, x_m | y = +1)} \qquad (2)$$

- SPRT is the following sequential strategy:

$$S^*_m(\boldsymbol{x}) = \begin{cases} -1, & \text{if } R_m(\boldsymbol{x}) \geq A, \\ +1, & \text{if } R_m(\boldsymbol{x}) \leq B, \\ \#, & \text{if } B < R_m(\boldsymbol{x}) < A. \end{cases} \qquad (3)$$

- The thresholds A and B are parameters of the test and are determined using the required errors α and β.
- It can be shown that *SPRT with optimal thresholds A^* and B^* is the optimal sequential test in the sense of criterion (1).*

SPRT: Thresholds A and B

Upper boound for A^*:

Assume that SPRT decides $S_m^*(\boldsymbol{x}) = -1$. Thus, the following must have been fulfilled:

$$R_m(\boldsymbol{x}) = \frac{p(x_1,\ldots,x_m|y=-1)}{p(x_1,\ldots,x_m|y=+1)} \geq A^*, \text{ or}$$
$$p(x_1,\ldots,x_m|y=-1) \geq A^* \cdot p(x_1,\ldots,x_m|y=+1).$$

Because the above holds for all sequences $\boldsymbol{x}$ such that $S_m^*(\boldsymbol{x}) = -1$, for the sum over all such $\boldsymbol{x}$ the following must hold:

$$\int_{\boldsymbol{x}:S_m^*(\boldsymbol{x})=-1} p(x_1,\ldots,x_m|y=-1)d\boldsymbol{x} \geq A^* \int_{\boldsymbol{x}:S_m^*(\boldsymbol{x})=-1} p(x_1,\ldots,x_m|y=+1)d\boldsymbol{x}$$
$$\underbrace{P(S_m^*(\boldsymbol{x})=-1|y=-1)}_{1-\beta} \geq A^* \cdot \underbrace{P(S_m^*(\boldsymbol{x})=-1|y=+1)}_{\alpha}$$

The upper bound A' for A^* is

$$\frac{1-\beta}{\alpha} = A' \geq A^*. \tag{4}$$

Lower bound for B^*:

Assume that SPRT decides $S_m^*(\boldsymbol{x}) = +1$. Thus, the following must have been fulfilled:

$$R_m(\boldsymbol{x}) = \frac{p(x_1,\ldots,x_m|y=-1)}{p(x_1,\ldots,x_m|y=+1)} \leq B^*, \text{ or}$$
$$p(x_1,\ldots,x_m|y=-1) \leq B^* \cdot p(x_1,\ldots,x_m|y=+1).$$

Because the above holds for all sequences $\boldsymbol{x}$ such that $S_m^*(\boldsymbol{x}) = +1$, for the sum over all such $\boldsymbol{x}$ the following must hold:

$$\int_{\boldsymbol{x}:S_m^*(\boldsymbol{x})=+1} p(x_1,\ldots,x_m|y=-1)d\boldsymbol{x} \leq B^* \int_{\boldsymbol{x}:S_m^*(\boldsymbol{x})=+1} p(x_1,\ldots,x_m|y=+1)d\boldsymbol{x}$$
$$\underbrace{P(S_m^*(\boldsymbol{x})=+1|y=-1)}_{\beta} \leq B^* \cdot \underbrace{P(S_m^*(\boldsymbol{x})=+1|y=+1)}_{1-\alpha}$$

The lower bound B' for B^* is

$$\frac{\beta}{1-\alpha} = B' \leq B^*. \tag{5}$$

SPRT: Final suggestions

- Optimal thresholds A^* and B^* are not easy to find. In practice, Wald suggests to use the bounds instead of the optimal values,

  $$A = A' = \frac{1-\beta}{\alpha} \qquad\qquad B = B' = \frac{\beta}{1-\alpha}$$

  in other words, to use stricter thresholds than needed.
- It can be shown that by using such thresholds the error probabilities change to α' and β' such that $\alpha' + \beta' \leq \alpha + \beta$, i.e. only one kind of error my get worse, one kind of error must get better.

Not solved by Wald:

1. Optimal ordering of individual measurements:
   - In Wald's applications, all measurements are i.i.d. (independent and identically distributed), so that their ordering do not matter.
2. Estimation of the likelihood ratio from the training data:
   - again, thanks to the i.i.d. assumption

     $$p(x_1,\ldots,x_m|y=c) = \prod_{i=1}^{m} p(x_i|y=c), \tag{6}$$

     so that R_m can be computed incrementally from 1D distributions.

Summary

Sequential analysis

- allows to decide the object class using a smaller number of measurements (features) than other methods,
- has plenty of applications, especially where
  - the time to decision must be minimized, and/or
  - individual measurements are not cheap.

Applications:

- Sequential sampling plans: measurements are homogeneous, their ordering does not matter.
- Clinical studies of drug effectivity: sequential testing by groups – if a measurement on a sample of n people is sufficient, the study is terminated, otherwise another batch of n people is used ...
- Real-time face detection in photos and videos [Š09]:
  - Individual measurements are differences in brightness in various places of the picture.
  - Time to decision is a critical factor, the classifier shall be usable in real-time!!!
  - Measurements are not independent!!!
  - AdaBoost: creates the ordering of measurements and estimates the likelihood ratio.
  - The face/no face/# decision is made by SPRT.
  - The resulting combination: WaldBoost.
- ...

Reference

[Fu68] K. S. Fu. *Sequential methods in pattern recognition and machine learning (Mathematics in science and engineering)*. Academic Press, 1st edition, 1968.

[Š09] Jan Šochman. *Learning for Sequential Classification*. PhD thesis, Czech Technical University in Prague, Prague, Czech Republic, December 2009. Available online at http://cyber.felk.cvut.cz/phd/completed/Sochman-PhD12_2009.pdf.

[Wal47] Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947.