International Journal of Intelligent Technology and Applied Statistics
Vol.1, No.1 (2008) pp.1-32

Statistical Hypothesis Testing Under Interval Uncertainty: An Overview

Vladik Kreinovich[1], Hung T. Nguyen[2*] and Sa-aat Niwitpong[3]

[1] *Department of Computer Science, University of Texas,
El Paso, TX 79968, USA*
[2] *Department of Mathematical Sciences, New Mexico State University,
Las Cruces, NM 88003, USA*
[3] *Department of Applied Statistics, King Mongkut's University of Technology,
North Bangkok, Thailand*

ABSTRACT

An important part of statistical data analysis is hypothesis testing. For example, we know the probability distribution of the characteristics corresponding to a certain disease, we have the values of the characteristics describing a patient, and we must make a conclusion whether this patient has this disease. Traditional hypothesis testing techniques are based on the assumption that we know the exact values of the characteristic(s) x describing a patient. In practice, the value $\widetilde{x}$ comes from measurements and is, thus, only known with uncertainty: $\widetilde{x} \neq x$. In many practical situations, we only know the upper bound Δ on the (absolute value of the) measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$. In such situation, after the measurement, the only information that we have about the (unknown) value x of this characteristic is that x belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

In this paper, we overview different approaches on how to test a hypothesis under such interval uncertainty. This overview is based on a general approach to decision making under interval uncertainty, approach developed by the 2007 Nobelist L. Hurwicz.

Keywords: Statistical hypothesis testing; Interval uncertainty; Hurwicz criterion

1. Formulation of the problem

Statistical hypothesis testing is important. An important part of statistical data analysis is hypothesis testing.

Examples. For example, we know the probability distribution of the characteristics corresponding to a certain disease, we have the values of the characteristics describing a patient, and we must make a conclusion whether this patient has this disease.

Another example is when we want to check whether a newly proposed treatment is effective against a disease. In this case, we have a distribution corresponding to un-treated patients, and we want to check whether the values corresponding to the treated patients fit within the same distribution.

[*] Corresponding author: hunguyen@nmsu.edu

Traditional approach to statistical hypothesis testing. Traditional hypothesis testing techniques are based on the assumption that we know the exact values of the characteristic(s) x describing a tested object. These techniques will be briefly described in the following text.

Need to take measurement uncertainty into account. In practice, the value $\widetilde{x}$ comes from measurements and are, thus, only known with uncertainty: $\widetilde{x} \neq x$. In other words, there is usually non-zero *measurement error* $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$; see, e.g., [15].

It is therefore desirable to take into account the measurement error when testing statistical hypotheses.

Case of probabilistic uncertainty. Traditional approach to handling measurement uncertainty in science and engineering is to assume that we know the exact probability distribution of the measurement errors. Usually, we assume that the measurement errors are normally distributed, with 0 mean and known standard deviation σ.

Statistical hypothesis testing techniques have been extended to situations in which we have such a probabilistic information about measurement uncertainty. This extension will also be briefly discussed in the following text.

Case of interval uncertainty: description. In many practical situations, we do not know the probabilities of different values of measurement error Δx. Instead, we only know the upper bound Δ on the (absolute value of the) measurement error Δx.

In such situation, after the measurement, the only information that we have about the (unknown) value x of this characteristic is that x belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

It is therefore desirable to extend the existing statistical hypothesis testing techniques to such interval situations.

Case of interval uncertainty: what is known and what we do in this paper. There exist several approaches to statistical hypothesis testing under interval uncertainty; see, e.g., [1, 3, 4, 6, 10, 12, 13] and references therein. Some of these approaches are formally derived from reasonable assumptions, others are based on semi-heuristic ideas.

In this paper, we provide a general overview of these approaches. We show that all these approaches can be formally justified within a general approach to decision making under interval uncertainty, approach developed by the 2007 Nobelist L. Hurwicz [7].

Comment 1. Our emphasis is on the *foundations* of the corresponding approaches. Readers interested in the corresponding algorithms should consult the corresponding papers.

Comment 2. Many of the above papers go beyond interval uncertainty, to the more general case of *fuzzy* uncertainty, when we can have different intervals **x** corresponding to different degrees of confidence.

On the methodological level, once we know how to process interval uncertainty, we can also process fuzzy uncertainty – by processing the corresponding intervals level-by-level. Of course, algorithmically, this may not be the best approach. However, as we have mentioned, our main objective is to concentrate on the foundational issues. In view of this objective, in this paper, we will only concentrate on interval uncertainty, and we refer readers interesting in the fuzzy algorithms to the corresponding papers.

2. Statistical hypothesis testing: formulation of the problem

Hypothesis testing: a practical problem. In many practical situations, we need to check whether a given object satisfies a given property. For example, based on the results of medical test(s), we needs to decide whether a person is healthy or has a certain disease which requires treatment. Another example is that, based on the test results, we must decide whether a mechanical system (e.g., a bridge) is stable and ready-to-use, etc.

The tested "object" may be more complicated than a single person or a single bridge. For example, when we check how efficient is a given treatment for a disease, we may want to consider the whole group of patients who undertook this treatment as a single object

The property that is normally satisfied is called a *null hypothesis* H_0. In medical testing, a null hypothesis is that a person is healthy, and that the treatment is no effective. In engineering testing, a hull hypothesis is that the tested structure is stable.

Statistical approach to hypothesis testing: main idea. To be able to check whether a given object satisfies the hull hypothesis, we collect the data about the objects which are known to satisfy this hypothesis. Based on this data, we find the probability distribution of the measured characteristic(s) for all the objects which satisfy the hypothesis H_0.

For each tested object with value(s) x of these characteristics, we thus get a probability (density) that this x satisfies the null hypothesis. If this probability is reasonably high, then we conclude that for this object the null hypothesis holds: a person is healthy, a bridge is stable, etc. If this probability is low, then we conclude that the null hypothesis does not hold – and the *alternative* hypothesis holds: a person is not healthy, the bridge is not stable, etc.

In order to translate this somewhat informal idea about hypothesis testing decisions into a precise criterion, let us recall how general decisions can be described.

3. Decision making: general approach

Decision making: general idea. It is known (see, e.g., [8, 11, 16]) that a reasonable description of human decision making comes from the *utility theory.*

Specifically, we need to select between one of several decisions $d_1, \ldots, d_k$.

Simplest case: when we know the exact consequences of each decision. In situations in which we know the exact situations resulting from each of these decisions, we can simply compare these situations and decide which of them we prefer.

General case. In practice, often, we can only predict the probabilities of different situations.

Example. For example, suppose that, based on the body temperature, we must make a decision on whether a person has a certain fever-inducing disease (and thus, whether we should start an appropriate treatment – or maybe whether we should perform further tests).

If the temperature is high enough (e.g., 38.5), then it is reasonable to conclude that this person has a disease. In this case, if we make a decision that this person has a disease, we improve this person's health; on the other hand, if we decide not to classify this person as sick, his or her disease may worsen.

However, medium temperatures are not that definite: a person with a temperature of 37.2 is most probably healthy, but this person may also have a starting stage of the disease.

- If we classify the person as sick and he or she is sick, we improve this person's health.
- On the other hand, if we classify the person as sick and in reality the person is healthy, we may unnecessarily damage his or her health by possible side effects of the (unnecessary) treatment.

We can determine the probabilities of different situations. Based on the past observations, we can determine the probabilities of different situations under different decisions. Let $s_1, \ldots, s_n$ denote possible situations, and let p_{ij} denote the probability that the decision d_i leads to a situation s_j. How can we describe the overall effect of each decision d_i?

Utility theory approach. To describe the overall effect of a decision, let us select two special situations:

- We select a very beneficial situation S_1 which is better than all the situations s_j; for example, as S_1, we can select the situation in which I win a million dollars.

- We also select a very bad situation S_0, a situation which is worse than all the situations $s_1, \ldots, s_n$.

(In the following text, we will see that the resulting selection of the best decision does not depend on our choice of these situations.)

For every probability p from the interval $[0, 1]$, we can consider a "lottery" $L(p)$ in which the situation S_1 occurs with probability p and the situation S_0 occurs with the remaining probability $1 - p$.

When $p = 1$, we get $L(1) = S_1$. When $p = 0$, we get $L(0) = S_0$. When p continuously increases, the benefit of the lottery $L(p)$ continuously increases, from S_0 to S_1. Since every situation s_j is in between S_0 and S_1, we thus conclude that there exists a probability u_j for which s_j is equivalent to the lottery $L(u_j)$. This value u_j is called the *utility* of the situation s_j.

How the effect of each decision is described in utility theory. Now, each decision d_i leads:

- to s_1 with probability p_{i1},

- $\ldots$,

- to s_n with probability p_{in}.

Since:

- s_1 is equivalent to a lottery $L(u_1)$,

- $\ldots$,

- s_n is equivalent to a lottery $L(u_n)$,

the consequences of the decision d_i are equivalent to a composite lottery L in which:

- with probability p_{i1}, we get a lottery $L(u_1)$,

- $\ldots$,

- with probability p_{in}, we get a lottery $L(u_n)$.

In each of the lotteries $L(u_j)$, the outcomes are S_0 and S_1. So, in our composite lottery L, we also get either S_1 or S_0. Due to the formula of full probability, the probability of having S_1 in the composite lottery L is equal to

$$E_i \stackrel{\rm def}{=} p_{i1} \cdot u_1 + \ldots + p_{in} \cdot u_n.$$

Thus, each decision d_i is equivalent to a lottery $L(E_i)$ in which E_i is equal to the expected value $E_i = \sum\limits_{j=1}^{n} p_{ij} \cdot u_j$ of the utility. This value E_i is called the *expected utility* of the decision d_i.

How to select the best decision. Now, the consequences of a decision d_i are equivalent to the appropriate lottery $L(E_i)$. Lotteries $L(u)$ are easy to compare: the larger the probability u of the favorable situation S_1, the better. Thus, we must select the decision d_i with the largest value of expected utility E_i.

Comment. The numerical value of the utility depends on the choice of the events S_0 and S_1. One can easily check that if we replace these events with another pair S'_0 and S'_1, then the new values of utility u'_j are related to the old ones u_j by a linear transformation $u'_j = a \cdot u_j + b$ for some constants $a > 0$ and b. A similar relation occurs between expected utilities. Thus, as expected, the selection of the best decision does not depend on the choice of the events S_0 and S_1,

In view of this re-scaling possibility, it is reasonable to consider utilities not only as taking values from the interval $[0, 1]$, but as attaining all possible real values.

4. Decision making approach to statistical hypothesis testing

How the general decision making approach relates to statistical hypothesis testing. To apply the traditional decision making approach, we must know:

- all the probabilities, and
- all the utility values.

Probabilities. Let us first describe all related probabilities.

- Let π_0 be the probability that a randomly selected object satisfies the hypothesis H_0.
- Then, $\pi_1 = 1 - \pi_0$ is the probability that a randomly selected object satisfies the alternative hypothesis H_1.

Let $\rho_0(x)$ be the probability density of x for the objects which satisfy the null hypothesis H_0, and let $\rho_1(x)$ be the probability density of x for the objects which satisfy the hypothesis H_1.

In this case, for a given x, the probability $p_0(x) \stackrel{\text{def}}{=} P(H_0 \,|\, x)$ that an object with this value satisfies the null hypothesis can be determined by using the Bayes formula

$$p_0(x) = \frac{P(x \,|\, H_0) \cdot P_0(H_0)}{P(x \,|\, H_0) \cdot P_0(H_0) + P(x \,|\, H_1) \cdot P_1(H_1)} = \frac{\rho_0(x) \cdot \pi_0}{\rho_0(x) \cdot \pi_0 + \rho_1(x) \cdot \pi_1}.$$

The probability $p_1(x) = P(H_1 \,|\, x)$ that an object with the value x satisfies the alternative hypothesis can be determined as

$$p_1(x) = 1 - p_0(x) = \frac{\rho_1(x) \cdot \pi_1}{\rho_0(x) \cdot \pi_0 + \rho_1(x) \cdot \pi_1}.$$

Utilities. Let us now describe possible situations and their utilities. In the case of hypothesis testing, there are 2 possible original situations:

- the situation when the null hypothesis holds, and
- the situation when the null hypothesis does not hold (and thus, the alternative hypothesis holds).

Each of these original situations generates two possible situations:

- when we decide that the null hypothesis H_0 holds, and
- when we decide that the alternative hypothesis H_1 holds.

We therefore have 4 possible situations. Let us use the following notations for the utilities of these situations:

- By u_{00}, we will denote the utility of the situation in which the object actually satisfies the null hypothesis H_0, and we (correctly) classify this object as satisfying the null hypothesis H_0.
- By u_{01}, we will denote the utility of the situation in which the object actually satisfies the null hypothesis H_0, and we (incorrectly) classify this object as satisfying the alternative hypothesis H_1.
- By u_{10}, we will denote the utility of the situation in which the object actually satisfies the alternative hypothesis H_1, and we (incorrectly) classify this object as satisfying the null hypothesis H_0.
- By u_{11}, we will denote the utility of the situation in which the object actually satisfies the alternative hypothesis H_1, and we (correctly) classify this object as satisfying the alternative hypothesis H_1.

Usually, correct classification is better that the incorrect one, so $u_{00} > u_{01}$ and $u_{11} > u_{10}$.

From the general decision theory formulas to statistical decision making. Once we know the measured value x, we can make two possible decisions:

- We can make the decision d_0 that the object satisfies the null hypothesis H_0.
- We can also make the decision d_1 that the object satisfies the alternative hypothesis H_1.

We know that the object satisfies the null hypothesis H_0 with probability $p_0(x)$ and satisfies the alternative hypothesis H_1 with the probability $p_1(x) = 1 - p_0(x)$. Thus, the expected utility E_0 of the decision d_0 is equal to

$$E_0 = p_0(x) \cdot u_{00} + p_1(x) \cdot u_{10},$$

and the expected utility of the decision d_1 is equal to

$$E_1 = p_0(x) \cdot u_{01} + p_1(x) \cdot u_{11}.$$

In accordance with the general idea of decision making, we select the decision with the largest value of expected utility. In other words, we select the null-hypothesis when $E_0 \geq E_1$, i.e., when

$$p_0(x) \cdot u_{00} + p_1(x) \cdot u_{01} \geq p_0(x) \cdot u_{01} + p_1(x) \cdot u_{11}.$$

Since $u_{00} > u_{01}$ and $u_{11} > u_{10}$, we can move the term proportional to u_{01} to the left-hand side and the term proportional to p_{10} to the right-hand side and come up with an equivalent inequality

$$p_0(x) \cdot (u_{00} - u_{01}) \geq p_1(x) \cdot (u_{11} - u_{10}).$$

Since $p_1(x) > 0$ (it is a probability) and $u_{00} - u_{01} > 0$, we can divide both sides of this inequality by $p_1(x)$ and by $u_{00} - u_{01}$ and conclude that

$$\frac{p_0(x)}{p_1(x)} \geq \frac{u_{11} - u_{10}}{u_{00} - u_{01}}.$$

Substituting the Bayes expressions for $p_0(x)$ and $p_1(x)$ into this formula, we conclude that

$$\frac{\rho_0(x) \cdot p_0}{\rho_1(x) \cdot p_1} \geq \frac{u_{11} - u_{10}}{u_{00} - u_{01}},$$

i.e., that

$$\frac{\rho_0(x)}{\rho_1(x)} \geq r_0,$$

where $r_0 \stackrel{\text{def}}{=} \dfrac{u_{11} - u_{10}}{u_{00} - u_{01}} \cdot \dfrac{p_1}{p_0}$.

Thus, we arrive at the following criterion:

Statistical hypothesis testing: resulting criterion. There exists a threshold r_0 – depending on the probabilities of different hypotheses and on the utilities of different situations – for which:

- we select the null hypothesis H_0 if the ratio $\dfrac{\rho_0(x)}{\rho_1(x)}$ exceeds this threshold r_0, and
- we select the alternative hypothesis H_1 if the ratio $\dfrac{\rho_0(x)}{\rho_1(x)}$ is below this threshold r_0.

In statistical hypothesis testing, this criterion is known as the *Neyman-Pearson* criterion.

5. Towards traditional statistical approach to hypothesis testing

From the general decision making approach to the traditional statistical approach to hypothesis testing. Up to now, we discussed the general case of statistical hypothesis testing, when in principle, both hypotheses H_0 and H_1 can be equally frequent.

In practice, we mostly encounter situations in which most objects satisfy the null hypothesis H_0. In such situations, as we will see, statistical hypothesis testing can be simplified. This simplified case is, in effect, what is usually described as the traditional statistical approach to hypothesis testing; see, e.g., [17].

Let us describe how in this case, the general decision making approach leads to the known statistical hypothesis testing formulas.

Type I and type II errors: reminder. As we have mentioned, there are two possible errors in decision making:

- It is possible that the object satisfies the null hypothesis H_0, but we erroneously classify it as satisfying the alternative hypothesis H_1. In statistical hypothesis testing, such an error is called *false positive*, or *type I error.*
- It is also possible that the object satisfies the alternative hypothesis H_1, but we erroneously classify it as satisfying the null hypothesis H_0. In statistical hypothesis testing, such an error is called *false negative*, or *type II error.*

Situations when type I errors are prevailing: general description. In many practical situations, the overwhelming majority of objects satisfy the null hypothesis. In such situations, the effect of type I errors is overwhelming.

Medical example. Let us give a typical example. Suppose that we want to detect a (reasonably rare) disease which is curable if caught early. Usually, there is an easy-to-implement (and reasonably cheap) procedure that provides a rough check of this disease.

If the results of this rough check are suspicious – i.e., if we classify the patient as (probably) satisfying the alternative hypothesis H_1 – then we can apply a more detailed test to check whether indeed a person has this disease.

For example, every woman over a certain age is recommended to regularly take a mammogram test. If anything suspicious is found on a mammogram, she is advised to take more complex, more expensive, and more time consuming tests such as ultrasound testing etc.

These more sophisticated tests provide a much more reliable test of the disease. In the ideal world, if we want to have a 100% reliable detection of the tested disease, we should apply this more sophisticated test to everyone. However, this more sophisticated test is usually much more expensive and therefore, realistically, we cannot afford to apply this test to everybody. Since we are operating within a given overall budget, we can thus deduce a portion p of the population to which we can afford to apply the more sophisticated test. This portion can be, e.g., 5%, 1%, 0.1%.

Since we consider the case when the number of objects satisfying the null hypothesis H_0 is prevailing, the actual portion of the objects which satisfy the alternative hypothesis H_1 is much smaller than p. Thus, the fact that we can only afford to check p-th portion of the population means that the probability of type I error cannot exceed p.

This probability should not be made smaller than p – because then we miss a portion of the population for which we could afford secondary (more expensive) testing.

Medical example: conclusion. Thus, we conclude that we should select the test in such a way that out of the population which satisfies the null hypothesis, exactly the p-th portion is classified as satisfying the alternative hypothesis H_1.

Engineering example. The above simple argument can be repeated for engineering testing. For example, to test a mechanical structure such as a bridge or an airplane, we can perform some easy-to-implement (comparatively) inexpensive tests to make sure that everything is OK with this structure. If something suspicious is detected – i.e., if, based on this general test, we classify this object as satisfying the alternative hypothesis – we then have a chance to apply a more expensive (and more time-consuming) test to get a more reliable picture of the structure's safety.

In the ideal world, we should apply this more sophisticated test to all the structures, but in reality, we cannot afford it. For example, a Space Shuttle undergoes extensive (and expensive) tests every time it flies. As a result, the Space Shuttle is reasonably safe – but every flight costs millions and billions of dollars. We cannot afford such detailed testing every time a normal passenger airplane flies.

Therefore, based on the available budget, we must limit detailed tests to a certain proportion p of the planes.

General conclusion. In situations in which the objects satisfying the null hypothesis are prevailing,

- we determine the value p (based on the budget restrictions), and
- we design a criterion for distinguishing between the null hypothesis H_0 and the alternative hypothesis H_1 in such a way that for objects satisfying the null hypothesis H_0, the probability of misclassifying them as H_1 is exactly p.

Observation: in this formulation, we do not need to have a detailed information about H_1. In the general decision making approach, we needed to know:

- the utilities u_{ij} of different situations,
- the probabilities π_0 and π_1 of satisfying hypotheses H_0 and H_1, and
- the probability densities $\rho_0(x)$ and $\rho_1(x)$ corresponding to the two hypotheses.

In the case when the objects satisfying the null hypothesis are prevailing, we do not now need to know the utilities u_{ij}, the probabilities π_0 and π_1, or the probability density $\rho_1(x)$ corresponding to the alternative hypothesis. In this case, we only need to know the probability density corresponding to the null hypothesis.

Since we do not need to have any detailed information about the alternative hypothesis H_1, we can simply describe it as a negation of H_0. For example:

- If the null hypothesis H_0 means that a person is healthy, then H_1 means that a person is not healthy.
- If the null hypothesis H_0 means that a mechanical structure is stable, then the alternative hypothesis H_1 simply means that this structure is not stable.

This is exactly the situation which is considered in the traditional statistical hypothesis testing.

Traditional statistical approach to hypothesis testing: a general description. In the traditional statistical approach to hypothesis testing, we formulate a single hypothesis – a null hypothesis H_0. For this hypothesis, we know the probability density $\rho_0(x)$ of the population of all the objects which satisfy this hypothesis. We are also given the required probability p of the type I error.

We then select a hypothesis testing criterion in which the probability of the type I error is exactly p.

Important practical case: unimodal distributions. In many practical situations, the actual distribution is normal (Gaussian) – or close to normal, e.g., unimodal; see, e.g., [17]. It is therefore reasonable to consider the statistical hypothesis testing situations in which the distribution $\rho_0(x)$ is unimodal (e.g., Gaussian).

This is the main case considered in the traditional statistical hypothesis testing. In the following text, we will therefore mainly concentrate on this case; the main ideas can be naturally extended to a more general case.

One-sided situations. In most practical situations, the intuitive notion of abnormality is one-sided. For example, suppose that to screen for certain diseases, we measure the pulse rate, the blood pressure, and/or the cholesterol level of different feeling-well people – to make sure that we catch any sign of possible heart diseases early. If a person has a blood pressure or cholesterol level smaller than average – there is nothing wrong with that, this person may be in very good physical health. On the other hand, if one of these characteristics is much higher than average, then this is a reason to be worried.

In principle, there exist two-sided situations, but since most practical cases involve one-sided situations, these are the situations on which we will concentrate in this paper.

In one-sided situations, all the values below the mode of H_0 should be classifies as satisfying the null hypothesis. Sine the distribution corresponding to the null hypothesis is unimodal, the further we go above the mode, the smaller the probability that the corresponding object satisfies the null hypothesis H_0. So, if we classify a value x as belonging to H_1, then every larger value – with an even smaller probability of the null hypothesis – should also be classified as H_1. Thus, a reasonable idea is to set up a *threshold* t such that:

- all the values x below t are classified as H_0 ("normal"); and
- all the values x above t are classified as H_1 ("abnormal", "outliers").

In other words, we divide the real line – the set of all possible values of x – into two zones:

- the "accept" zone $A = (-\infty, t)$ in which the null hypothesis H_0 is accepted, and
- the "reject" zone $R = (t, \infty)$ in which the null hypothesis H_0 is rejected (and thus, the alternative hypothesis H_1 is accepted).

The value of the threshold t can be uniquely determined from the condition that for objects satisfying the hypothesis H_0, the probability of rejection is exactly equal to p: $\int_t^\infty \rho_0(x)\,dx = p$.

Example: normal distribution. To illustrate this idea, let us consider the case when for the “normal” objects (i.e., objects which satisfy the null hypothesis), the distribution of the measured quantity x is Gaussian, with mean a and standard deviations σ.

We know that for the Gaussian distribution:

- the probability of being outside the “two sigma” interval $[a-2\sigma, a+2\sigma]$ is approximately 10%, and

- the probability to be outside the “three sigma” interval $[a-3\sigma, a+3\sigma]$ is approximately 0.1%.

Since Gaussian distribution is symmetric:

- the probability of exceeding $a+2\sigma$ is exactly half of the probability to be outside the interval $[a-2\sigma, a+2\sigma]$ – i.e., $\approx 5\%$, and

- the probability of exceeding $a+3\sigma$ is exactly half of the probability to be outside the interval $[a-3\sigma, a+3\sigma]$ – i.e., $\approx 0.05\%$.

Thus:

- for $p \approx 5\%$, the corresponding threshold is $t = a + 2\sigma$;

- for $p \approx 0.05\%$, the corresponding threshold is $t = a + 3\sigma$.

In this case, if $x < t$, we classify the object as normal (= accept the null hypothesis), and if $x > t$, we classify the object as abnormal (i.e., reject the null hypothesis).

Need to take into account measurement uncertainty. The above criterion works well if we know the exact value x of the measured quantity.

In practice, measurements are never absolutely accurate. As a result, instead of the exact value x, we only know the measurement result $\widetilde{x}$ which is only approximately equal to x. Based on this approximate value $\widetilde{x}$, how can we then make a statistical decision?

What is known. This problem was actively researched for the situation in which we know the probabilities of different values of measurement error $\Delta x = \widetilde{x} - x$. There have also been several papers in which statistical hypothesis testing was extended to the interval case when we only know the upper bound Δ on the measurement error.

In the following sections, we will overview these results.

6. Statistical hypothesis testing under probabilistic uncertainty

Probabilistic uncertainty: a brief description. In this section, we describe how statistical hypothesis testing criteria should be modified if, instead of the knowing the exact value x of the desired quantity, we only know this value with probabilistic uncertainty. In precise terms, we only know the measurement result $\widetilde{x}$, and we know the probability distribution of the measurement error $\Delta x = \widetilde{x} - x$.

Traditionally in science and engineering, it is assumed that this measurement error is normally distributed, with 0 mean and known standard deviation σ_m. This is the case on which we will concentrate in this paper.

How to modify traditional statistical hypothesis testing technique under probabilistic uncertainty: main idea. Let us start with the simplest case of traditional statistical hypothesis testing.

In the traditional approach to statistical hypothesis testing, we assume that large values of x indicate abnormality. Thus, similar to the above description, a reasonable idea is to select a threshold t and classify an object as normal if $\widetilde{x} < t$ and as abnormal if $\widetilde{x} > t$.

First seemingly natural idea: let us select the same threshold as before. At first glance, it may sound reasonable to select the same threshold as before. For example, for $p = 5\%$, we select $t = a + 2\sigma$.

However, as we will see, this is not always a good idea.

Analysis of the situation. As we have discussed, the threshold t must be selected in such a way that for "normal" objects, the probability of misclassification is exactly p. In other words, the probability that the measured value $\widetilde{x}$ satisfies the inequality $\widetilde{x} > t$ must be equal to p.

Thus, to find the corresponding threshold t, we must find out the probability distribution for the measured values $\widetilde{x}$ corresponding to normal objects.

We know that for normal objects, the actual value x is normally distributed with mean a and standard deviation σ. The measured value $\widetilde{x}$ differs from x by the measurement error: $\widetilde{x} = x + \Delta x$. We know that the measurement error Δx is also normally distributed, with 0 mean and standard deviation σ_m. It is also usually assumed that the measurement error is independent on the measured quantity. Thus, the measured value $\widetilde{x} = x + \Delta x$ is the sum of two independent normally distributed random variables.

It is known that such a sum is also normally distribution, with the mean equal to the sum of the corresponding means and the variance equal to the sum of the corresponding variances. Thus, we conclude that the measured values $\widetilde{x}$ are normally distributed with mean a and standard deviation $\sqrt{\sigma^2 + \sigma_m^2}$. So, we arrive at the following conclusion.

Resulting criterion.

- For $p \approx 5\%$, the corresponding threshold is $t = a + 2 \cdot \sqrt{\sigma^2 + \sigma_m^2}$.
- For $p \approx 0.05\%$, the corresponding threshold is $t = a + 3 \cdot \sqrt{\sigma^2 + \sigma_m^2}$.

Comment. By comparing these formulas with the formulas corresponding to the exact values ($\sigma_m = 0$), we see that the threshold changes when we take measurement uncertainty into account.

If we keep the same threshold value as before, then the probability of exceeding the threshold will become higher than p – so, the above seemingly natural idea does not work.

More complex situations. We described the main idea on a simple example. A more detailed description of how hypothesis testing should be changed under probabilistic measurement uncertainty can be found, e.g., in [2, 5, 18].

7. Statistical hypothesis testing under probabilistic uncertainty: preliminary analysis of the problem

Interval uncertainty: a brief reminder. In many practical situations, we do not know the probabilities of different values of the measurement error $\Delta x = \widetilde{x} - x$. In many such situations, we only know the upper bound Δ on the (absolute value of the) measurement error. In this case, after the measurement, the only information that we have about the (unknown) actual value x is that x belongs to the interval $\mathbf{x} = [\underline{x}, \overline{x}]$, where $\underline{x} = \widetilde{x} - \Delta$ and $\overline{x} = \widetilde{x} + \Delta$.

Hypothesis testing under interval uncertainty: a problem. If we only know x with such interval uncertainty, then how shall we classify the corresponding object?

Let us start with a simple case of hypothesis testing, when we classify an object as normal or abnormal by comparing the value x characterizing this object with a threshold t.

If we knew the value x exactly, then we could classify the object

- as normal if $x < t$ and
- as abnormal if $x > t$.

Under interval uncertainty, we do not know the exact value x, we only know the interval $[\underline{x}, \overline{x}]$ which is guaranteed to contain x.

Cases when classification is easy. There are two cases when classification under interval uncertainty is easy:

- If $\overline{x} < t$, this means that all possible values of x from the interval $[\underline{x}, \overline{x}]$ belong to the accept set. In this case, we know that the corresponding object is normal.

- If $t < \underline{x}$, this means that all possible values of x from the interval $[\underline{x}, \overline{x}]$ belong to the reject set. In this case, we know that the corresponding object is abnormal.

Case when classification is difficult. The difficult remaining case is when some values from the interval $\mathbf{x}$ are below the threshold t, and some values are above the threshold, i.e., when the threshold is inside the interval $\mathbf{x}$.

How do we then classify an object?

Possible solution: withhold decision. A possible solution is to withhold decision, i.e., to say that based on the measurement result, we do not have enough information to accept or to reject the null hypothesis.

In many practical situations, this idea makes perfect sense; see, e.g., [4]. However, in other practical situations, we do need to make a decision: e.g., we need to decide whether to further test a patient or whether to further test a mechanical structure. How shall we make this decision?

Interval uncertainty is more difficult to handle than a probabilistic one. In the case of probabilistic uncertainty, we modified the traditional statistical approach to hypothesis testing. This was possible because this approach is based on the requirement that the probability of type I error is equal to a given value p. When we know the probability distribution of measurement error, we can still find the probability of type I error.

In the case of interval uncertainty, however, we do not know the probabilities of different values of the measurement error. As a result, we do not know the exact probability of type I error, we only know that this probability is somewhere within the corresponding interval. Thus, we cannot make definite decision based on the assumption that the probability of type I error is equal to p.

Since we cannot easily modify the traditional approach to the statistical decision making, we have to go back to the more general (and more complex) decision making situation.

General decision making approach to statistical hypothesis testing: a brief reminder. In the general decision making approach, we consider two possible decisions: a decision d_0 to proclaim the object normal, and a decision d_1 to proclaim the object abnormal. For a given value x, we compute the expected utilities E_0 and E_1 of these decisions as

$$E_0 = p_0(x)\cdot u_{00} + p_1(x)\cdot u_{10} = p_0(x)\cdot u_{00} + (1 - p_0(x))\cdot u_{10} = p_0(x)\cdot(u_{00} - u_{10}) + u_{10}$$

and

$$E_1 = p_0(x)\cdot u_{01} + p_1(x)\cdot u_{11} = p_0(x)\cdot u_{01} + (1 - p_0(x))\cdot u_{11} = p_0(x)\cdot(u_{01} - u_{11}) + u_{11},$$

where

$$p_0(x) = \frac{p_0\cdot\rho_0(x)}{p_0\cdot\rho_0(x) + p_1\cdot\rho_1(x)} = \frac{1}{1 + \dfrac{\rho_1(x)}{\rho_0(x)}\cdot\dfrac{p_1}{p_0}}.$$

Then, we select a decision with the largest value of the expected utility.

Decision making about hypotheses under interval uncertainty. If we know the exact value of x, then:

- we can find the exact value of the Neyman-Pearson ratio $r \stackrel{\text{def}}{=} \dfrac{\rho_0(x)}{\rho_1(x)}$;
- based on this ratio r, we can find the exact value of $p_0(x)$;
- and finally, based on the value $p_0(x)$, we find the exact values of the expected utilities E_0 and E_1.

Under interval uncertainty, we only know an interval $[\underline{x}, \overline{x}]$ of possible values of x. Thus:

- we can only find the range $[\underline{r}, \overline{r}]$ of possible values of the ratio r;
- based on this range, we can find the range $[\underline{p}_0, \overline{p}_0]$ of possible values of $p_0(x)$;
- and finally, based on the range of values for $p_0(x)$, we can find the ranges $[\underline{E}_0, \overline{E}_0]$ and $[\underline{E}_1, \overline{E}_1]$ of possible values of the expected utilities E_0 and E_1.

Let us derive the explicit formulas for these ranges.

First step: range of the Neyman-Pearson ratio. Let us first find the range of possible values of the ratio $r = \dfrac{\rho_0(x)}{\rho_1(x)}$. Since we only know that $x \in [\underline{x}, \overline{x}]$, we can thus conclude that this range is equal to $[\underline{r}, \overline{r}]$, where

$$\underline{r} = \min_{x\in[\underline{x},\overline{x}]} \frac{\rho_0(x)}{\rho_1(x)}; \quad \overline{r} = \max_{x\in[\underline{x},\overline{x}]} \frac{\rho_0(x)}{\rho_1(x)}.$$

A reasonable case is when $\rho_0(x)$ is a unimodal distribution, and $\rho_1(x)$ is also a unimodal distributions. In this case:

- values below the mode of $\rho_0(x)$ should be clearly classified as normal,
- values above the mode of $\rho_1(x)$ should be clearly classified as abnormal, and
- the only values which need to be classified are the values between these two modes.

In between these two modes, the density $\rho_0(x)$ is decreasing (since the distribution $\rho_0(x)$ is unimodal), the density function $\rho_1(x)$ is increasing, and thus, the ratio $r = \dfrac{\rho_0(x)}{\rho_1(x)}$ is decreasing. So, in this case,

- The ratio r of the two probability densities attains its smallest value $\underline{r}$ on the interval $[\underline{x}, \overline{x}]$ when the value of x is the largest possible, i.e., when $x = \overline{x}$.
- Similarly, the ratio r attains its largest value $\underline{r}$ on the interval $[\underline{x}, \overline{x}]$ when the value of x is the smallest possible, i.e., when $x = \underline{x}$.

In other words, in this case,

$$\underline{r} = \frac{\rho_0(\overline{x})}{\rho_1(\overline{x})}; \quad \overline{r} = \frac{\rho_0(\underline{x})}{\rho_1(\underline{x})}.$$

Second step: range of the Bayesian probability $p_0(x)$. Based on r, we compute $p_0(x)$ as

$$p_0(x) = \frac{1}{1 + \dfrac{1}{r} \cdot \dfrac{p_1}{p_0}}.$$

When r increases, the ratio $\dfrac{1}{r}$ decreases, hence the denominator decreases and thus, the ratio $p_0(x)$ increases. Thus:

- The smallest value $\underline{p}_0$ of $p_0(x)$ is attained when r is the smallest, i.e., when $r = \underline{r}$.
- The largest value $\underline{p}_0$ of $p_0(x)$ is attained when r is the largest, i.e., when $r = \overline{r}$.

In other words,

$$\underline{p}_0 = \frac{1}{1 + \dfrac{1}{\underline{r}} \cdot \dfrac{\pi_1}{\pi_0}} \quad \overline{p}_0 = \frac{1}{1 + \dfrac{1}{\overline{r}} \cdot \dfrac{\pi_1}{\pi_0}}.$$

Final step: ranges of the expected utilities. Finally, let us find the ranges of possible values of E_0 and E_1. We know that $E_0 = p_0(x) \cdot (u_{00} - u_{10}) + u_{10}$. Since correct classification is more beneficial, we conclude that $u_{00} > u_{10}$ and thus, E_0 is an increasing function of $p_0(x)$. Hence,

$$\underline{E}_0 = \underline{p}_0 \cdot (u_{00} - u_{10}) + u_{10}$$

and

$$\overline{E}_0 = \overline{p}_0 \cdot (u_{00} - u_{10}) + u_{10}.$$

Similarly, we know that $E_1 = p_0(x) \cdot (u_{01} - u_{11}) + u_{11}$. Since the correct classification is more beneficial, we conclude that $u_{11} > u_{01}$ and thus, E_1 is a decreasing function of $p_0(x)$. Hence,

$$\underline{E}_1 = \overline{p}_0 \cdot (u_{01} - u_{11}) + u_{11}$$

and

$$\overline{E}_1 = \underline{p}_0 \cdot (u_{01} - u_{11}) + u_{11}.$$

How to make a decision under interval uncertainty? Up to now, we only considered decision making for situations when we know the exact values of the expected utility. To test statistical hypotheses under interval uncertainty, we must therefore figure out how, in general, we can make decisions under interval uncertainty – i.e., how to make decisions in situations in which we only know the interval of possible values of expected utility.

This general problem was solved in the early 1950s by L. Hurwicz [7, 11], who received a 2007 Nobel prize in economics for this research. So, before applying his results to statistical hypothesis testing, let us briefly recall the main idea behind Hurwicz's approach.

8. Decision making under interval uncertainty: Hurwicz approach

Formulation of the problem. Let us assume that for some situation s, we do not know the exact value of its utility u, we only know the interval $[\underline{u}, \overline{u}]$ of possible values of this utility. How can we then make decisions based on this interval?

Main idea underlying utility theory: a brief reminder. As we have mentioned, the main idea behind utility theory is that to gauge the quality of each situation, we compare it with "lotteries" $L(p)$ – characterized by exactly known probability values p. In these terns, a situation with a utility u is a situation which is equivalent to a lottery $L(u)$.

How this general idea can be applied to decision making under interval uncertainty. In line with this general idea, to gauge the quality of a situation described by an interval $[\underline{u}, \overline{u}]$, we should find a lottery $L(u)$ which is (in some reasonable sense) equivalent to this situation. In other words, for each interval $[\underline{u}, \overline{u}]$, we must find a utility value u which is (in some reasonable sense) equivalent to this interval.

Up to now, the problem sounds similar to the classical utility theory. The main difference is that in the classical utility theory, we ask the decision maker to tell us what is the probability u for which the given situation s is equivalent to the lottery $L(u)$. For the case of an interval-valued utility, the decision maker clearly is unable to narrow down this interval to a single value. Thus, to find a value which is equivalent to an interval, we can no longer rely on the decision maker: we have to find this value ourselves.

Idea: invariance. Our objective is to develop a mapping $e(\underline{u}, \overline{u})$ that maps every interval $[\underline{u}, \overline{u}]$ into a single equivalent value $u = e(\underline{u}, \overline{u})$. What properties should this mapping have?

As we have mentioned, the numerical values of the utility depend on the choice of the two basic situations S_0 and S_1. Different choices of these two situations lead to different scales for representing utility. Different scales $u(s)$ and $u'(s)$ are related to each other by linear transformations $u'(s) = a \cdot u(s) + b$ for some $a > 0$ and b.

It is therefore reasonable to require that the desired mapping does not change under such re-scalings. Let us formulate this property in precise terms. Suppose that we start in the original scale. In this case, we have the interval $[\underline{u}, \overline{u}]$. Based on this interval, we find the equivalent value $u = e(\underline{u}, \overline{u})$.

Suppose now that we use a different scale to represent the same situation, a scale which is related to the original one by a linear transformation $u'(s) = a \cdot u(s) + b$. In this new scale, the endpoints $\underline{u}$ and $\overline{u}$ of the interval take new numerical values $\underline{u}' = a \cdot \underline{u} + b$ and $\overline{u}' = a \cdot \overline{u} + b$. When we apply the combination function e to these new values $\underline{u}'$ and $\overline{u}'$, we get an equivalent value $u' = e(\underline{u}', \overline{u}')$, i.e.,

$$u' = e(a \cdot \underline{u} + b, a \cdot \overline{u} + b).$$

It is reasonable to require that this new value represent the exact same equivalent utility u as before, but expressed in the new scale, i.e., that $u' = a \cdot u + b$ for $u = e(\underline{u}, \overline{u})$.

Substituting the expressions $u' = e(a \cdot \underline{u} + b, a \cdot \overline{u} + b)$ and $u = e(\underline{u}, \overline{u})$ into the formula $u' = a \cdot u + b$, we conclude that for every $\underline{u} < \overline{u}$, $a > 0$, and b, we have

$$e(a \cdot \underline{u} + b, a \cdot \overline{u} + b) = a \cdot e(\underline{u}, \overline{u}) + b.$$

Let us show that this natural invariance condition leads to a very specific expression for the combination function u.

Consequences of invariance. Let us pick one possible interval, e.g., an interval $[0, 1]$. This means that the actual utility of a situation is somewhere between 0 and 1.

Let us denote the utility value $e(0, 1)$ equivalent to this interval by α. From the common sense viewpoint, this value cannot be negative – since every possible value $u \in [0, 1]$ is greater than any negative number. Similarly, this equivalent value cannot be larger than 1. Thus, we must have $\alpha \in [0, 1]$.

Let $[u^-, u^+]$ be an arbitrary non-degenerate interval. One can easily check that this interval can be obtained from the interval $[0, 1]$ by an appropriate linear re-scaling: namely, from the conditions that $[a \cdot 0 + b, a \cdot 1 + b] = [u^-, u^+]$ we conclude that $a \cdot 0 + b = b = u^-$. Then, from $a \cdot 1 + b = a + b = u^+$, we conclude that $a = u^+ - b = u^+ - u^-$. For the resulting values $\underline{u} = 0$, $\overline{u} = 1$, $a = u^+ - u^-$, and $b = u^-$, the above invariance implies that $e(u^-, u^+) = (u^+ - u^-) \cdot \alpha + u^-$. By combining terms proportional to u^- and to u^+, we conclude that $u = \alpha \cdot u^+ + (1 - \alpha) \cdot u^-$. This is exactly the formula derived by L. Hurwicz. So, we arrive at the following solution to the problem of decision making under uncertainty:

Decision making under uncertainty: Hurwicz solution. When we only know an interval $[\underline{u}, \overline{u}]$ of possible values of utility corresponding to a given situation (or a given decision), then we characterize this situation (decision) by a single equivalent utility value

$$u = \alpha \cdot \overline{u} + (1 - \alpha) \cdot \underline{u},$$

and we select a decision for which the equivalent value u is the largest.

How do we select α: Hurwicz interpretation. The above approach requires that we fix the value of the parameter α. This parameter must be selected in such a way as to best represent the user's preferences. To help with this selection, L. Hurwicz provided the following reasonable interpretation of this parameter.

Let us recall that in case of the interval uncertainty, we do not know the exact value of the utility characterizing each decision, we only know the interval $[\underline{u}, \overline{u}]$ of possible values characterized by this utility.

- In the most optimistic case, we get the largest possible value $\overline{u}$ of this utility.
- In the most pessimistic case, we get the smallest possible value $\underline{u}$ of this utility.
- In reality, we will most probably get some value which is strictly between $\underline{u}$ and $\overline{u}$.

It turns out that these cases are directly related to the choice of the parameter α:

- When $\alpha = 1$, this means the equivalent utility value is equal to $u = \overline{u}$. In other words, we judge each decision by it most optimistic outcome.

- When $\alpha = 1$, this means the equivalent utility value is equal to $u = \underline{u}$. In other words, we judge each decision by it most pessimistic outcome.

- When $0 < \alpha < 1$, this means the equivalent utility value u is strictly in between the pessimistic value $\underline{u}$ and the optimistic value $\overline{u}$.

In view of this relation, the general Hurwicz criterion for decision making under interval uncertainty is also called *optimism-pessimism criterion* – because to make a decision, it uses a linear combination of the optimistic and pessimistic estimates.

Geometric interpretation of Hurwicz criterion. An interesting geometric interpretation of Hurwicz criterion is described in [9, 12, 13].

Let us assume that we want to check whether a given situation characterized by the utility interval $\mathbf{u} = [\underline{u}, \overline{u}]$ is better or worse than a standard one, with the utility value u_0. In terms of hypothesis testing, we can say that we have a null hypothesis that the standard situation (characterized by the value u_0) is better.

When the utility interval is degenerate, i.e., when $[\underline{u}, \overline{u}] = [u, u]$ and a given situation is characterized by the exact value u of the utility, then the answer to this question is straightforward:

- When the value u belongs to the "accept" set $A \stackrel{\text{def}}{=} (-\infty, u_0]$, then we accept the hypothesis – and thus claim that the standard situation u_0 is better than the given one.

- When the value u belongs to the "reject" set $R \stackrel{\text{def}}{=} (u_0, \infty)$, then we reject the hypothesis – and thus claim that the standard situation u_0 is worse than the given one u.

What happens in the non-degenerate case, when $\underline{u} < \overline{u}$? When the entire utility interval is inside the accept set A, then we accept the null hypothesis; when the entire interval is inside the reject set, then we reject the null hypothesis. The problem is when the interval contains points both from the accept set and from the reject set.

A reasonable idea is to find out what proportion of the interval $\mathbf{u}$ is in the accept set, i.e., to estimate the ratio $r \stackrel{\text{def}}{=} \dfrac{|\mathbf{u} \cap A|}{|\mathbf{u}|}$, where $|\mathbf{u}|$ denotes the width of the interval $\mathbf{u}$.

- If this ratio is sufficiently high – i.e., if it exceeds a certain threshold r_0 – then we accept the hypothesis.

- If this ratio is too small – i.e., if it is below a threshold r_0 – then we reject the hypothesis.

Let us show that this reasonable idea is indeed equivalent to the Hurwicz criterion. Indeed, here $|\mathbf{u}| = |[\underline{u}, \overline{u}]| = \overline{u} - \underline{u}$. Also, $\mathbf{u} \cap A = [\underline{u}, u_0]$, hence $|\mathbf{u} \cap A| = u_0 - \underline{u}$ and $r = \dfrac{u_0 - \underline{u}}{\overline{u} - \underline{u}}$. Thus, the condition $r \geq r_0$ is equivalent to $u_0 - \underline{u} \geq r_0 \cdot (\overline{u} - \underline{u})$, i.e., to $u_0 \geq r_0 \cdot \overline{u} + (1 - r_0) \cdot \underline{u}$. Thus, according to this reasonable idea, we accept the hypothesis if $u_0 \geq u$, where the "equivalent utility" u of the interval $[\underline{u}, \overline{u}]$ is equal to $r_0 \cdot \overline{u} + (1 - r_0) \cdot \underline{u}$. One can see that this is exactly the Hurwicz criterion, with the optimism-pessimism coefficient equal to r_0.

A similar idea is to find out what proportion of the interval is in the reject set, i.e., to estimate the ratio $r \stackrel{\text{def}}{=} \dfrac{|\mathbf{u} \cap R|}{|\mathbf{u}|}$

- If this ratio is sufficiently low – i.e., if it is does not exceed a certain threshold r_0 – then we accept the hypothesis.
- If this ratio is sufficiently high – i.e., if it is exceeds a threshold r_0 – then we reject the hypothesis.

One can check that this idea is also equivalent to the Hurwicz criterion – with $\alpha = 1 - r_0$.

9. Statistical hypothesis testing under interval uncertainty: applications of Hurwicz approach

Let us show how the Hurwicz approach to decision making under interval uncertainty can help in statistical hypothesis testing.

General case. In the general case, for each of the two possible decisions d_0 and d_1, we have intervals $[\underline{E}_0, \overline{E}_0]$ and $[\underline{E}_1, \overline{E}_1]$ of possible values of the corresponding expected utility. In accordance withe the general Hurwicz approach, we select the null hypothesis is

$$\alpha \cdot \overline{E}_0 + (1 - \alpha) \cdot \underline{E}_0 > \alpha \cdot \overline{E}_1 + (1 - \alpha) \cdot \underline{E}_1,$$

i.e., if

$$(\alpha \cdot \overline{p}_0 + (1 - \alpha) \cdot \underline{p}_0) \cdot (u_{00} - u_{10}) + u_{10} > (\alpha \cdot \underline{p}_0 + (1 - \alpha) \cdot \overline{p}_0) \cdot (u_{01} - u_{11}) + u_{11},$$

where

$$\underline{p}_0 = \frac{1}{1 + \dfrac{1}{\underline{r}} \cdot \dfrac{\pi_1}{\pi_0}}; \quad \overline{p}_1 = \frac{1}{1 + \dfrac{1}{\overline{r}} \cdot \dfrac{\pi_1}{\pi_0}}; \quad \underline{r} = \frac{\rho_0(\overline{x})}{\rho_1(\overline{x})}; \quad \overline{r} = \frac{\rho_0(\underline{x})}{\rho_1(\underline{x})}.$$

This is the interval version of the Neyman-Pearson criterion.

Extreme cases. Let us consider the extreme cases $\alpha = 1$ (optimism) and $\alpha = 0$ (pessimism).

Optimism case. In the optimism case, when $\alpha = 1$, we make decisions based on the best-case scenario. In this case, we select the null hypothesis when

$$\overline{p}_0 \cdot (u_{00} - u_{10}) + u_{10} > \underline{p}_0 \cdot (u_{01} - u_{11}) + u_{11}.$$

Pessimism case. In the pessimism case, when $\alpha = 0$, we make decisions based on the worst-case scenario. In this case, we select the null hypothesis when

$$\underline{p}_0 \cdot (u_{00} - u_{10}) + u_{10} > \overline{p}_0 \cdot (u_{01} - u_{11}) + u_{11}.$$

Case corresponding to the traditional statistical approach: reminder. Let us describe how the above criterion can be simplified in the cases corresponding to the traditional statistical approach.

When we know the exact value x, then the classification depends on whether the probability for a normal object to exceed x is smaller than or great then the given fraction p. This probability is equal to $\int_x^\infty \rho_0(t)\,dt$ and is can therefore be described in terms of the cumulative distribution function $F_0(x) \stackrel{\text{def}}{=} \int_{-\infty}^x \rho(t)\,dt$, as $1 - F_0(x)$. Thus:

- If $1 - F_0(x) > p$, then we cannot classify x as abnormal, because then, we would have to classify all objects exceeding x as abnormal, and the resulting expenses for additional checking would be too high. So, in this case, we classify the object x as normal.
- If $1 - F_0(x) \le p$, then we can afford checking this object and all the objects with higher value x, so we can afford to classify this object as abnormal.

These probabilities can be translated into benefits (utility values). If we classify an object with the value x as abnormal, this means that all the objects for which the value is x or higher will be thoroughly checked. The benefit of doing this is proportional to the number of objects who will be thus checked, i.e., to $1 - F_0(x)$. If the resulting benefit (utility) does not exceed p, we can afford to perform all these checks. If the benefit exceeds p, this means that we cannot afford so much checking – and thus, we have to classify the object as normal.

How to generalize the traditional statistical approach to the case of interval uncertainty. In case of interval uncertainty, we do not know the exact value x, we only know an interval $[\underline{x}, \overline{x}]$ which contains x. Different values x from this interval leads to different utility values $1 - F_0(x)$. When x increases, the probability $1 - F_0(x)$ of exceeding x decreases. Thus:

- the largest possible value of $1 - F_0(x)$ corresponds to the smallest possible x, i.e., to $x = \underline{x}$, and
- the smallest possible value of $1 - F_0(x)$ corresponds to the largest possible x, i.e., to $x = \overline{x}$.

So, the interval $[\underline{u}, \overline{u}]$ of possible values of utility u is proportional to

$$[1 - F_0(\overline{x}), 1 - F_0(\underline{x})].$$

According to the Hurwicz criterion, this interval is equivalent to the utility value u for which $u = \alpha \cdot (1 - F_0(\underline{x})) + (1 - \alpha) \cdot (1 - F_0(\overline{x}))$, i.e., $u = 1 - (\alpha \cdot F_0(\underline{x}) + (1 - \alpha) \cdot F_0(\overline{x}))$. Thus, we select the null hypothesis if $u \le p$, and we reject it if $u > p$.

So, we arrive at the following criterion:

Resulting criterion for statistical hypothesis testing under interval uncertainty. Suppose that we have a one-sided statistical hypothesis testing situation. Suppose also that the probability distribution of objects which satisfy the null hypothesis H_0 is described by a probability density function $\rho_0(x)$ and by the cumulative distribution function $F_0(x)$. Suppose also that we have an object for which we do not know the exact value of the quantity x, we only know the range $[\underline{x}, \overline{x}]$ of possible values of this quantity.

Suppose also that we describe the user's decision making by an optimism-pessimism value $\alpha \in [0, 1]$, and that desired type I error is p. In this case:

- we accept the null hypothesis if $1 - (\alpha \cdot F_0(\underline{x}) + (1 - \alpha) \cdot F_0(\overline{x})) \le p$, and
- we reject the null hypothesis if $1 - (\alpha \cdot F_0(\underline{x}) + (1 - \alpha) \cdot F_0(\overline{x})) > p$.

Extreme cases. Let us consider the extreme cases $\alpha = 1$ (optimism) and $\alpha = 0$ (pessimism).

Optimism case. In the optimism case $\alpha = 1$, we make our decision based on the value $\underline{x}$:

- we accept the null hypothesis if $1 - F_0(\underline{x}) \le p$, and
- we reject the null hypothesis if $1 - F_0(\underline{x}) > p$.

Pessimism case. In the optimism case $\alpha = 0$, we make our decision based on the value $\overline{x}$:

- we accept the null hypothesis if $1 - F_0(\overline{x}) \le p$, and
- we reject the null hypothesis if $1 - F_0(\overline{x}) > p$.

In the pessimism case, we are making a decision in such a way as to guarantee that for all possible values x from the interval $[\underline{x}, \overline{x}]$, the probability of exceeding x is p or smaller; see, e.g., [1].

10. Case when we also know distributions with interval uncertainty

Motivations. In the previous text, we assumed that we only know the value x (characterizing a given object) with interval uncertainty, but that the probabilities of normal and abnormal populations are known exactly. In practice, these probabilities also come from measurements and estimates and are, thus, also only know with uncertainty.

Let us therefore consider the case when, in addition to knowing x with interval uncertainty, we also know the probabilities with interval uncertainty.

Traditional statistical approach to hypothesis testing: case of interval uncertainty. Let us start with the simplest case of the traditional statistical approach to hypothesis testing. In this approach, we assume that we know the cumulative distribution function (cdf) $F_0(x)$. Interval uncertainty means that instead of the exact values of the cdf, for each x, we only know the bounds $[\underline{F}_0(x), \overline{F}_0(x)]$ on the cdf. Such an interval-valued cdf is known as a *probability box*, or *p-box*, for short.

In general, the benefit of accepting x (and larger values) is proportional to $1 - F_0(x)$, where $F_0(x)$ is an increasing function. In our case, we also know that $x \in [\underline{x}, \overline{x}]$, and that $F_0(x) \in [\underline{F}_0(x), \overline{F}_0(x)]$. Thus, the smallest possible value of the utility is attained:

- when x attains the largest possible value, and
- when $F_0(x)$ attains the largest possible value.

So, $\underline{u} = 1 - \overline{F}_0(\overline{x})$. Similarly, the largest possible value $\overline{u}$ of the corresponding utility is attained:

- when x attains the smallest possible value, and
- when $F_0(x)$ attains the smallest possible value.

So, $\overline{u} = 1 - \underline{F}_0(\underline{x})$.

Thus, the interval $[\underline{u}, \overline{u}]$ of possible values of utility u is proportional to

$$[1 - \overline{F}_0(\overline{x}), 1 - \underline{F}_0(\underline{x})].$$

According to the Hurwicz criterion, this interval is equivalent to the utility value u for which $u = \alpha\cdot(1-\underline{F}_0(\underline{x}))+(1-\alpha)\cdot(1-\overline{F}_0(\overline{x}))$, i.e., to $1-(\alpha\cdot\underline{F}_0(\underline{x})+(1-\alpha)\cdot\overline{F}_0(\overline{x})$. Thus, we select the null hypothesis if $u \le p$, and we reject it if $u > p$.

So, we arrive at the following criterion:

Resulting criterion for statistical hypothesis testing under interval uncertainty. Suppose that we have a one-sided statistical hypothesis testing situation. Suppose that we known the bounds $[\underline{F}_0(x), \overline{F}_0(x)]$ on the (unknown) cumulative distribution function which characterizes all objects that satisfy the null hypothesis H_0. Suppose that we have an object for which we do not know the exact value of the quantity x, we only know the range $[\underline{x}, \overline{x}]$ of possible values of this quantity.

Suppose also that we describe the user's decision making by an optimism-pessimism value $\alpha \in [0,1]$, and that desired type I error is p. In this case:

- we accept the null hypothesis if $1 - (\alpha \cdot \underline{F}_0(\underline{x}) + (1-\alpha) \cdot \overline{F}_0(\overline{x})) \le p$, and
- we reject the null hypothesis if $1 - (\alpha \cdot \underline{F}_0(\underline{x}) + (1-\alpha) \cdot \overline{F}_0(\overline{x})) > p$.

Extreme cases. Let us consider the extreme cases $\alpha = 1$ (optimism) and $\alpha = 0$ (pessimism).

Optimism case. In the optimism case $\alpha = 1$, we make our decision based on the values $\underline{x}$ and $\underline{F}_0(x)$:

- we accept the null hypothesis if $1 - \underline{F}_0(\underline{x}) \le p$, and
- we reject the null hypothesis if $1 - \underline{F}_0(\underline{x}) > p$.

Pessimism case. In the optimism case $\alpha = 0$, we make our decision based on the values $\overline{x}$ and $\overline{F}_0(x)$:

- we accept the null hypothesis if $1 - \overline{F}_0(\overline{x}) \le p$, and
- we reject the null hypothesis if $1 - \overline{F}_0(\overline{x}) > p$.

In the pessimism case, we are making a decision in such a way as to guarantee that for all possible values x from the interval $[\underline{x}, \overline{x}]$ and for all possible cdfs $F_0(x) \in [\underline{F}_0(x), \overline{F}_0(x)]$, the probability of exceeding x is p or smaller.

General case. In the general case, the uncertainty comes from not know the exact value of the expression

$$p_0(x) = \frac{1}{1 + \dfrac{1}{r} \cdot \dfrac{\pi_1}{\pi_0}},$$

where $r \stackrel{\text{def}}{=} \dfrac{\rho_0(x)}{\rho_1(x)}$. In the previous text, we assumed that we the only uncertainty is in x; in other words, we assume that instead of the exact value x, we only the interval $[\underline{x}, \overline{x}]$ of possible values of x. In addition to this, instead od knowing the exact values of the probabilities π_0, π_1, $p_0(x)$, and $p_1(x)$, we only know the intervals $[\underline{\pi}_0, \overline{\pi}_0]$, $[\underline{\pi}_1, \overline{\pi}_1]$, $[\underline{p}_0(x), \overline{p}_0(x)]$, and $[\underline{p}_1(x), \overline{p}_1(x)]$ containing these values. In this case, we have

$$\underline{p}_0 = \frac{1}{1 + \dfrac{1}{\underline{r}} \cdot \dfrac{\overline{\pi}_1}{\underline{\pi}_0}}; \quad \overline{p}_0 = \frac{1}{1 + \dfrac{1}{\overline{r}} \cdot \dfrac{\underline{\pi}_1}{\overline{\pi}_0}},$$

where – under the previous assumption that $\rho_0(x)$ increases with x and $\rho_1(x)$ decreases with x – we conclude that

$$\underline{r} = \frac{\underline{\rho}_0(\overline{x})}{\overline{\rho}_1(\overline{x})}; \quad \overline{r} = \frac{\overline{\rho}_0(\underline{x})}{\underline{\rho}_1(\underline{x})}.$$

11. A similar problem in which we actually observe interval ranges

A general problem that we considered so far: brief reminder. The main objective of this paper is to overview different approaches to hypothesis testing under interval uncertainty. Up to now, we considered the situations in which the quantity used in the classification has the exact value. For example, a patient has a certain count of white blood cells.

Case of interval uncertainty that we considered so far. Traditional hypothesis testing deals with the cases in which we know the exact value of this quantity. In the above text, we considered situations in which we do not know the exact value x of the quantity, we only know the interval $\mathbf{x} = [\underline{x}, \overline{x}]$ of possible values of this quantity. Based on this interval, we need to make a decision.

Another case of interval data. In practice, there are other types of situations in which we only observe intervals. Namely,

- so far, we assumed that the quantity has the exact value, and the interval uncertainty comes from the fact that we do not know this exact value;
- in many practical situations, the quantity does not have the exact value, it changes and it has a range of possible values.

Such situations are typical in many medical measurements. For example, such frequently used characteristics as the pulse rate, the body temperature, the blood pressure do not have the exact value: they change from moment to moment, they change during the day, they change from one activity to another, they simply change because of the stress of being in a doctor's office. It is therefore not reliable to use a single measured value of such a characteristic to make a medical diagnosis. A more reliable way is to measure, e.g., blood pressure throughout the day, and to report the corresponding *range* of the values – i.e., the interval $[\underline{x}, \overline{x}]$ formed by the corresponding measurement results.

This case when we actually observe the actual interval *range* of a changing quantity is different from the above case – when observe an interval that contains the actual (unknown) *value* of the un-changing quantity.

Hypothesis testing in situations in which we observe the actual ranges: formulation of the problem. In such situations, we have the following problem:

- based on the previous observations, we know the probability distribution of the intervals corresponding to the test hypothesis – and maybe the probability distribution of the intervals corresponding to the alternative hypothesis;
- we then observe an interval $[\underline{x}, \overline{x}]$ corresponding to the tested object;
- based on this interval, we must decide whether the object satisfies the tested hypothesis.

Hypothesis testing in situations in which we observe the actual ranges: how to solve this problem. The above problem is, in effect, the standard statistical testing problem. The only difference from the simplified version of statistical testing that we considered earlier is that in that version, we had only *one* observed quantity x, while here, in effect, we have *two* observed quantities: $\underline{x}$ and $\overline{x}$.

In effect, instead of 1-D random variable x, we now have a 2-D random variable $(\underline{x}, \overline{x})$. Thus, instead of 1-D distribution(s) and 1-D observations, here we have 2-D distribution(s) and 2-D observations. We can still use the standard statistical techniques to handle this situation; see, e.g., [3, 6].

Comment. In our description, we characterized an interval $[\underline{x}, \overline{x}]$ as a pair $(\underline{x}, \overline{x})$ of its endpoints. From the purely *computational* viewpoint, this makes perfect sense, because in the computer, the natural way to represent the interval $[\underline{x}, \overline{x}]$ is by describing its lower endpoint $\underline{x}$ and its upper endpoint $\overline{x}$.

However, from the viewpoint of *understanding*, an interval $[\underline{x}, \overline{x}]$ is the set of all possible values – and it is thus different from a pair of its endpoints. From this viewpoint, we have a distribution of *sets*, and based on a new observation set, we need to check whether the observed set belongs to this distribution. Thus reformulated problem becomes a particular case of problems related to *random sets*; see, e.g., [14].

Need to combine two types of interval uncertainty. The above case is mainly developed for situations in which we know the *exact* range $[\underline{x}, \overline{x}]$ [3, 6]. In practice, the range comes from measurement, and measurements are never 100% accurate. As a result, the measured values are, in general, different from the actual values of the measured quantity – and hence, the range estimation based on these measured values is, in general, different from the actual range.

If we knew the exact values $x_1, \ldots, x_n$, then we could simply compute the endpoints of the range as $\underline{x} = \min(x_1, \ldots, x_n)$ and $\overline{x} = \max(x_1, \ldots, x_n)$. If we measure the values x_i with an accuracy $\varepsilon > 0$, then instead of the actual (unknown) values $x_1, \ldots, x_n$ we get the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ for which $|\widetilde{x}_i - x_i| \le \varepsilon$ for all i. Based on these measurement results, we compute the estimates $\widetilde{\underline{x}} = \min(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and $\widetilde{\overline{x}} = \max(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

From $|\widetilde{x}_i - x_i| \le \varepsilon$, we conclude that $|\widetilde{\underline{x}} - \underline{x}| \le \varepsilon$ and $|\widetilde{\overline{x}} - \overline{x}| \le \varepsilon$. Thus, we only know both endpoints with uncertainty ε.

In set terms, we can say that instead of the exact range interval $\mathbf{x} = [\underline{x}, \overline{x}]$, we only know an interval $\widetilde{\mathbf{x}} = [\widetilde{\underline{x}}, \widetilde{\overline{x}}]$ for which the Hausdorff distance $d_H(\mathbf{x}, \widetilde{\mathbf{x}}) \le \varepsilon$, i.e., for which

$$[\widetilde{\underline{x}} + \varepsilon, \widetilde{\overline{x}} - \varepsilon] \subseteq [\underline{x}, \overline{x}] \subseteq [\widetilde{\underline{x}} - \varepsilon, \widetilde{\overline{x}} + \varepsilon].$$

Combining two types of interval uncertainty: a problem. So, we arrive at the following problem:

- based on the previous observations, we know the probability distribution of the intervals corresponding to the test hypothesis – and maybe the probability distribution of the intervals corresponding to the alternative hypothesis;

- we then observe an interval $[\widetilde{\underline{x}}, \widetilde{\overline{x}}]$ which is ε-close to actual (unknown) range corresponding to the tested object;

- based on this interval, we must decide whether the object satisfies the tested hypothesis.

Combining two types of interval uncertainty: towards practically useful algorithms. From the *methodological* viewpoint, we know how to solve this problem: we can use, e.g., the above Hurwicz approach.

The remaining *practical* problem is to transform this general methodology into efficient algorithms.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grants HRD-0734825, EAR-0225670, and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, by the Max Planck Institut für Mathematik, and by the Japan Advanced Institute of Science and Technology (JAIST) International Joint Research Grant 2006-08. This work was partly performed during Sa-aat Niwitpong's visit to Las Cruces and El Paso in Fall 2007. The authors are thankful to T. Augustin, T. Denoeux, M. A. Gil,

H. Kutterer, I. Neumann, and to many others for useful references and valuable discussions.

References

[1] T. Augustin (2002). Neyman-Pearson testing under interval probability by globally least favorable pairs – Reviewing Huber-Strassen theory and extending it to general interval probability, *Journal of Statistical Planning and Inference*, 105, 149-173.

[2] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Errors in Nonlinear Models: A Modern Perspective*, Chapman & Hall/CRC, Boca Raton, Florida.

[3] R. Coppi, M. A. Gil, and H. A. L. Kiers (2006). The fuzzy approach to statistical analysis, *Computational Statistics and Data Analysis*, 51, 1-14.

[4] T. Denoeux, M. -H. Masson, and P. -A. Hébert (2005). Nonparametric risk-based statistics and significance tests for fuzzy data, *Fuzzy Sets and Systems*, 153, 1-28.

[5] W. A. Fuller (1987). *Measurement Error Models*, Wiley, New York.

[6] M. A. Gil, M. López-Díaz, and D. A. Ralescu (2006). Overview on the development of fuzzy random variable, *Fuzzy Sets and Systems*, 157, 2546-2557.

[7] L. Hurwicz (1951). Optimality Criteria for Decision Making Under Ignorance, Cowles Commission Discussion Paper, *Statistics*, 370.

[8] R. L. Keeney and H. Raiffa (1976). *Decisions with Multiple Objectives*, John Wiley and Sons, New York.

[9] H. Kutterer and I. Neumann (2006). Multidimensional statistical tests for imprecise data, *In: Proceedings of the 6th Hotine-Marussi Symposium of Theoretical and Computational Geodesy*, Wuhan, China.

[10] V. P. Kuznetsov (1991). *Interval Statsitical Models*, Moscow, Radio i Svyaz Publ. (in Russian).

[11] R. D. Luce and H. Raiffa (1989). *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[12] I. Neumann and H. Kutterer (2007). Congruence tests and outlier detection in deformation analysis with respect to observation imprecision, *International Journal of Applied Geodesy*, 1, 1-7.

[13] I. Neumann, H. Kutterer, and S. Schön (2006). Outlier detection in geodetic applications with respect to observation imprecision, In: R. L. Muhanna and R. L. Mullen (Eds.), *Proceedings of NSF Workshop on Reliable Engineering Computing*, Savannah, Georgia, 75-90.

[14] H. T. Nguyen (2006). *An Introduction to Random Sets*, Chapman & Hall/CRC, Bocar Raton, Florida.

[15] S. Rabinovich (2005). *Measurement Errors and Uncertainties: Theory and Practice*, Springer-Verlag, New York.

[16] H. Raiffa (1970). *Decision Analysis*, Addison-Wesley, Reading, Massachusetts.

[17] D. Sheskin (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Bocar Raton, Florida.

[18] T. W. F. Stroud (1974). Comparing regressions when measurement error variances are known, *Psychometrika*, 39, 53-68.