II THE NEYMAN-PEARSON THEORY OF TESTING A STATISTICAL HYPOTHESIS [4]

The principles of statistical inference as developed in the last two decades by R.A.Fisher, Neyman and Pearson deal with the problem of testing a hypothesis and with the problem of estimation but not with the general problem of statistical inference as it has been formulated in the foregoing pages. A further restriction in these theories is that they deal only with the case that Ω is a k-parameter family of distribution functions, i.e., that the true but unknown distribution function F is known to be an element of a k-parameter family of functions

$$F(x_1, x_2, \dots, x_n, \theta_1, \theta_2, \dots, \theta_k)$$

where $\theta_1, \dots, \theta_k$ are parameters. In this case the specification of the values of the parameters specifies completely the distribution function F.

A set of parameter values can be represented by a point in a k-dimensional Euclidean space called a <u>parameter space</u>. Because of the one-to-one correspondence between elements of Ω and points of the parameter space we can identify Ω with the parameter space. If for example, $X_1, \dots, X_n$ are normally and independently distributed, each having the same distribution (equation(2)), then the parameter space is a half plane where $\theta_1 = \mu =$ mean value, and $0 \leq \theta_2 = \sigma =$ standard deviation.

A hypothesis concerning F is expressed by the statement that the true parameter point lies in a certain subset ω of the parameter space Ω. As we have done before, we shall call the hypothesis a simple one if ω consists of a single point.

[4] See, in this connection, references 12,13 and 14

Otherwise, it is called a composite hypothesis. In the above example the statement that $\mu = 0$, $\sigma = 1$ is a simple hypothesis, while merely stating that $\mu = 0$ without specifying σ is a composite hypothesis.

For the sake of simplicity we shall confine ourselves to the case of a single unknown parameter since this suffices to illustrate the basic ideas of the theories of Fisher, Neyman and Pearson. First, we shall deal with the Neyman-Pearson theory of testing a statistical hypothesis.

We assume that the unknown distribution function is known to be an element of a one-parameter family $F(x_1, x_2, \ldots, x_n, \theta)$ and we wish to test the hypothesis $\theta = \theta_0$.

A simple example for this case is the following: Let it be known that $X_1, \ldots, X_n$ are independently and normally distributed with the same mean and unit variances, i.e., Ω is the one-parameter family of distributions

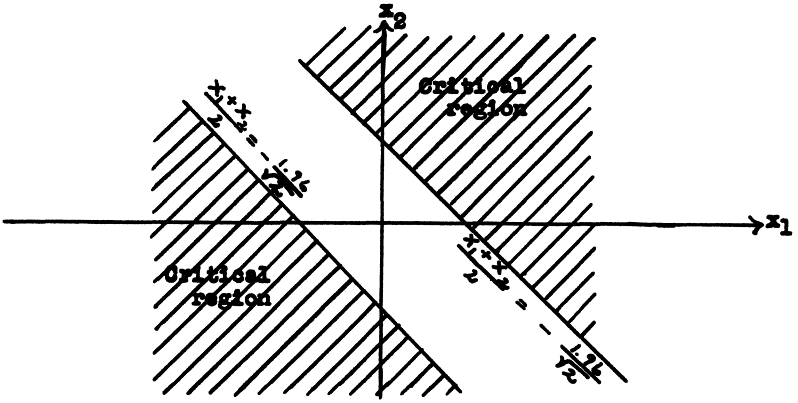
$$F(x_1, \ldots, x_n, \theta) = \frac{1}{(2\pi)^{n/2}} \int_{-\infty}^{x_1} e^{\frac{-(v-\theta)^2}{2}} dv \ldots \int_{-\infty}^{x_n} \frac{-(v-\theta)^2}{2} dv,$$

and assume that we wish to test the hypothesis that $\theta = 0$. According to the classical theory we reject this hypothesis if and only if

$$|\bar{x}| \geq c; \quad (\bar{x} = \frac{x_1 + \ldots + x_n}{n})$$

where c denotes a certain constant. The value of c is chosen in such a way that the probability of $|\bar{x}| > c$ under the assumption that the hypothesis $\theta = 0$ is true, is so small that we are willing to reject the hypothesis. If we want this probability to be 5 percent, then $c = \frac{1.96}{\sqrt{n}}$.

If, in the same example, we have made only two observations x_1, x_2, so that the sample space is the Euclidian plane, the critical region consists of all points for which $\frac{1}{2}(x_1+x_2) > \frac{1.96}{\sqrt{2}}$ and all points for which $\frac{1}{2}(x_1+x_2) < \frac{-1.96}{\sqrt{2}}$. If the point representing the observations falls within the critical region (i.e., if the arithmetic mean of the two observations is larger than $\frac{1.96}{\sqrt{2}}$ or smaller than $\frac{-1.96}{\sqrt{2}}$) we shall reject the hypothesis that the mean value is zero.


x_2
x_1
Critical region
Critical region
$\frac{x_1+x_2}{2} = -\frac{1.96}{\sqrt{2}}$
$\frac{x_1+x_2}{2} = -\frac{1.96}{\sqrt{2}}$

But the classical theory does not suggest why this critical region should be used. It merely proves that the probability for the observation point to fall within the critical region is five percent when the initial hypothesis is fulfilled. But there are infinitely many regions which enjoy the same property, and the classical theory does not give any reasons why just the one region mentioned should be chosen.

In order to arrive at a distinction between various critical regions, Neyman and Pearson advance the following considerations. In making a statement of acceptance or rejection of a hypothesis, we may commit two types of errors: rejecting the hypothesis although it is true (*error of type I*), or failing to reject it although it is false (*error of type II*). If the hypothesis consists in saying that the unknown parameter θ has a given value θ_0, the situation may be summarized as follows:

Truth or Falsehood of Statement
Concerning the Hypothesis $\theta = \theta_0$

True Situation	Statement Advanced	
	$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta = \theta_0$	Correct	Type I error
$\theta \neq \theta_0$	Type II error	Correct

By *size of the critical region* we mean the probability that the point representing the observations will fall within the critical region, where the probability in question is calculated under the assumption that the hypothesis is true. (Thus, in the example used before, the size of the critical region was five percent.) This may be expressed by saying that the size of the critical region is equal to the probability of commitclub a type I error.

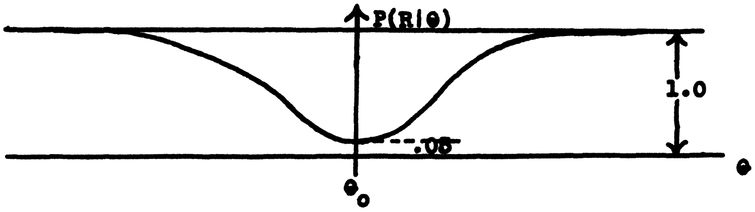
The general idea underlying the theory of Neyman and Pearson is to *minimize the probability of type II errors* while keeping the probability of type I errors constant.

If R is any region in the sample space, and E is the point of the sample space which represents the observations, we shall denote by $P(R|\theta_1)$ the probability of E lying in R calculated under the assumption that θ_1 is the true value of the unknown parameter θ, that is to say, $P(R|\theta_1)$ is equal to the Stieltjes integral $\int_R dF(x_1, \dots, x_n, \theta_1)$ over the region R. Thus, if we make the hypothesis $\theta = \theta_0$ and choose R as a critical region for this hypothesis, the size of the critical region will be given by the expression $P(R|\theta_0)$. If the hypothesis is wrong and the true value of θ is θ_1, then the probability of *avoiding* an error of type II is $P(R|\theta_1)$.

The expression $P(R|\theta_1)$, i.e., one minus the probability of an error of type II, is called the *power of the critical region R with respect to the alternative hypothesis $\theta = \theta_1$*.

The expression $P(R|\theta)$ is a function of θ. It may be plotted as a curve, the ordinate of which is equal to the size of R if the abscissa is θ_0, and equal to the power of R with respect to the alternative $\theta = \theta_1$ if the abscissa is any value $\theta_1 \neq \theta_0$. This curve is called the *power curve of the region R*.

In the former example, in which the distribution was normal with unknown mean and unit variance, and the critical region chosen was $|\bar{x}| > \frac{1.96}{\sqrt{n}}$ (where $\bar{x}$ is the arithmetic mean of the observations $x_1, x_2, \dots, x_n$), the power curve can easily be calculated and has the form shown below:



In order to compare the test $|\bar{x}|>\frac{1.96}{\sqrt{n}}$ with other possible tests, we have to compare the above power curve with the power curves of other critical regions which have the same size, five percent.

In general, if we have two critical regions R and R', both of which have the desired size, and if the power curve of R' is above that of R for the value $\theta = \theta_1$, then the critical region R' is better than R for testing the hypothesis _if_ the true value of θ happens to be θ_1. For the probability of committing a type I error is the same whether R or R' is used, while the probability of committing a type II error when using R' is smaller than when using R. If the power curve of R' is above that of R for each θ (except θ_0 for which the two curves coincide by assumption), then R' will be called _uniformly more powerful than R_. The test using the critical region R is called _non-admissible_ because its use is, under all circumstances, less favorable than the use of R'.

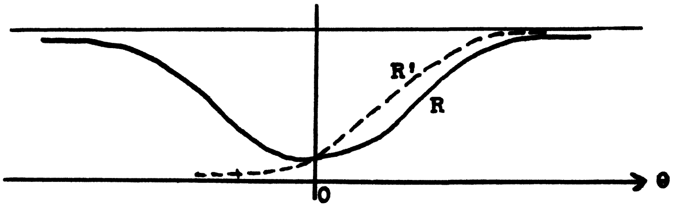
In order to make this clear, let us assume that a large number of samples is drawn, each of which consists of N individual observations. Let M be the number of such samples and let two statisticians, whom we will call S and S', test the same hypothesis, using each of the M samples. Assume that S uses the critical region R for testing while S' bases his tests on the region R'. S and S' will each obtain M answers to the question as to whether the null hypothesis (the hypothesis to be tested) should be rejected. Some of these answers will be right, others will be wrong. Let us compare the records of S and S'. We have to distinguish between the case that the null hypothesis is true and the case that it is false. a)In the first case, the answers obtained by each statistician may either be that the hypothesis is to be accepted - these answers are right; or that it should be rejected - these answers are errors of type I. The probability of committing a type I error by testing the null hypothesis from a sample drawn at random is equal to the size of the critical region used in testing. If M is large, it is practically certain that the relative frequency of type I errors will be approximately equal to their probability, i.e., to the size of the critical region. Since R and R' have, by assumption, equal size, each of the two statisticians will commit approximately the same number of errors. b)If the null hypothesis is false, some of the M answers obtained by each statistician will correctly reject it, while others will accept it, thus committing errors of type II. If M is large, the relative frequency of correct answers will be approximately equal to the power of the test used which we have pointed out is the probability of avoiding a type II error. By assumption, the power of R' is greater than that of R, regardless of what the true value of θ is, provided only that θ is different from θ_0. Therefore, the relative frequency of wrong answers obtained by S will tend to be greater than the relative frequency of wrong answers obtained by S'. Thus, if the null hypothesis is false (no matter what the true value of θ is), it is practically certain that S will make more false statements; while if the null hypothesis is true, S and S' will commit an approximately equal number of false statements. The method used by S', i.e., the application of the critical region R', is therefore superior to the method used by S, i.e., the application of the critical region R.

These considerations decide the choice between two critical regions of equal size if one of them is uniformly more powerful than the other, i.e., if the power curve of the former is above that of the latter for all values of θ except θ_0 (for which the power curves coincide). On the other hand, if the power curve of R' is above that of R for some values of θ, but below it for other values of θ, then we cannot choose one of the two regions without introducing further principles on which to base the choice.

If, for all values of θ, the power curve of a region R is never below that of any other region R' of equal size, then R is called a *uniformly most powerful region*, and the test corresponding to R a uniformly most powerful test.

The first principle for selecting a test is this: whenever we can find a uniformly most powerful test, we shall prefer it to all other tests using regions of the same size. Unfortunately, uniformly most powerful tests do not exist in most cases.

In the example which we have used on page 11 let us consider the region R' determined by the inequality $\bar{x} > \frac{1.64}{\sqrt{n}}$. It can easily be shown that R' (like the region R considered before) has the size .05. The power curves of R and R' are shown below:



We can see that for all $\theta > 0$, R' is more powerful than R, and vice versa for $\theta < 0$. In such cases further principles have to be formulated on which the choice should be based. It is clear that the choice we make will depend on our a priori degree of belief in the truth of the different possible values of θ. For instance, if we know a priori that θ cannot be negative, then we shall prefer R'.

Moreover, it can be shown that R' is uniformly most powerful if the parameter space is restricted to non-negative values of θ. If negative and positive values of θ are considered a priori as equally possible we will most likely prefer R to R'.

This example shows also that the choice of the critical region depends essentially on Ω. If Ω consists of all non-negative values of θ then the region R' is a uniformly most powerful test. If Ω consists of all non-positive values θ, then the region R'' given by $\bar{x} < \frac{-1.64}{\sqrt{n}}$ is a uniformly best region. Finally, if Ω consists of all real values θ, then the use of the region R seems to be more reasonable than that of R' or R''.

Since uniformly most powerful regions rarely exist, Neyman and Pearson introduced a further principle on which the choice of the critical region should be based, namely, the principle of unbiasedness. A test is called <u>unbiased</u> if the power function of the test has a relative minimum at the value $\theta = \theta_0$ where θ_0 is the hypothesis to be tested.

Some rationalization of this principle can be given: Suppose a test is biased, then for some value θ_1, in the neighborhood of θ_0, the power of the test is less than the size of the region. But this means that the probability of rejecting the hypothesis $\theta = \theta_0$ is larger if θ_0 is true than if θ_1 is true, which is not a desirable situation.

In general, an infinity of unbiased tests exist, hence we need a further principle in order to select a proper test from among them. We define as a uniformly most powerful unbiased test one which is at least as powerful or more powerful, with respect to all alternate hypotheses, than any other unbiased region of equal size. If a uniformly most powerful unbiased test exists, and if we accept the principle of unbiasedness, then it is obvious that it is the most advantageous test to use. Neyman and Pearson called a critical region corresponding to a uniformly most powerful unbiased test a critical region of type A_1.

Referring to the example previously considered, the critical region given by $|\bar{x}| > c$ is a region of type A_1 for testing the hypothesis in question. Another example of a region of type A_1 is the following: Let $X_1,\dots,X_n$ be independently and normally distributed with zero means and a common variance. Then, for testing the hypothesis that the common variance σ^2 is equal to $\sigma_0{}^2$, the critical region consisting of all points of the sample space which satisfy at least one of the inequalities

$$x_1{}^2 + \dots + x_n{}^2 > c_1 \quad \text{or} \quad x_1{}^2 + \dots + x_n{}^2 < c_2 ,$$

is a critical region of type A_1 if the constants c_1 and c_2 are properly chosen.

The region of type A_1 exists in an important, but very restricted, class of cases; there are many instances in which it does not exist. Therefore, Neyman and Pearson have introduced a third type of region, known as a region of type A. The region R is said to be of type A if its power function $P(W/\theta)$ is such that

$$1) \quad \left.\frac{\partial P(R|\theta)}{\partial\theta}\right|_{\theta = \theta_0} = 0$$

and

$$2) \quad \left.\frac{\partial^2 P(R|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0} \geq \left.\frac{\partial^2 P(R'|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}$$

for all regions R' which satisfy 1) and have the same size as R. The first condition restricts the region to be unbiased. The second requires the power function of a region of type A to have a greater curvature than that of any other unbiased region of the same size. To put it crudely, it means that the region is most powerful in the neighborhood of θ_0.

A critical region of type A exists under very weak conditions which are fulfilled in most of the practical cases. However, the objection can be raised against a region of type A that we are much more concerned with the behavior of the power function for alternatives θ which are far from θ_0 than for those in the neighborhood of θ_0. In spite of this, as we will see, a good justification of the use of a type A region can be given in the light of some recent results.