ERROR: FIRST READ OF FISHER'S 1926 PAPER

Overview

In his seminal 1926 paper, “the arraignment of field experiments”, Fisher clarified many fundamental issues in experimental design.

What is amazing about this paper is that there is not a single equation in it! Yet it laid the foundation of experimental design.

We will come back to this paper a few times throughout this quarter.

In this first read, we will focus on “error”:

- What is error?
- What did Fisher mean by “a valid estimate of error”?
- What are the roles of replication and randomization in obtaining a valid estimate of error?

In his seminal 1926 paper, “the arraignment of field experiments”, Fisher clarified many fundamental issues in experimental design.

What is amazing about this paper is that there is not a single equation in it! Yet it laid the foundation of experimental design.

We will come back to this paper a few times throughout this quarter.

In this first read, we will focus on “error” and ask the following questions:

What is error?

What did Fisher mean by “a valid estimate of error”?

What are the roles of replication and randomization in obtaining a valid estimate of error?

Fisher's note on error

In the opening of his 1926 paper, Fisher pointed out that field experiments have two aims—**efficiency** and **validity**:

> "From about 1923 onwards the Statistical Department at Rothamsted had been much concerned with the precision of field experiments in agriculture, and with modifications in their design, having the dual aim of **increasing the precision** and of **providing a valid estimate of error**.
>
> These two desiderata had been somewhat confused in the minds of experimenters, and the present paper was the author's first attempt at setting out the rational principles on which he might proceed."

Fisher almost dedicated this entire paper to explain what is "a valid estimate of error". Fisher was probably the first to realize that one of the aims and tasks of experimental design is to provide a valid estimate of error.

In the opening of his 1926 paper, Fisher pointed out that field experiments have two aims—efficiency and validity:

"From about 1923 onwards the Statistical Department at Rothamsted had been much concerned with the precision of field experiments in agriculture, and with modifications in their design, having the dual aim of increasing the precision and of providing a valid estimate of error."

"These two desiderata had been somewhat confused in the minds of experimenters, and the present paper was the author's first attempt at setting out the rational principles on which he might proceed."

Fisher almost dedicated this entire paper to explain what is "a valid estimate of error". Fisher was probably the first to realize that one of the aims and tasks of experimental design is to provide a valid estimate of error.

Fisher's note on error

Later in the paper, Fisher continued …

> "… The conception which has made it possible to develop a new and critical technique of plot arrangement is that an estimate of field errors derived from any particular experiment may or may not be a valid estimate, and in actual field practice is usually not a valid estimate, of the actual errors affecting the averages or differences of averages of which it is required to estimate the error. …"

Here again, Fisher emphasized that an important goal of an experiment is to provide a valid estimation of error.

Later in the paper, Fisher continued …

"… The conception which has made it possible to develop a new and critical technique of plot arrangement is that an estimate of field errors derived from any particular experiment may or may not be a valid estimate, and in actual field practice is usually not a valid estimate, of the actual errors affecting the averages or differences of averages of which it is required to estimate the error. …"

Here again, Fisher emphasized that an important goal of an experiment is to provide a valid estimation of error.

What is error?

Fisher described error this way:

> "For more than fifteen years the attention of agriculturalists has been turned to the errors of field experiments. During this period, experiments of the uniformity trial type have demonstrated the magnitude and ubiquity of **that class of error which cannot be ascribed to carelessness in measuring the land or weighing the produce, and which is consequently described as due to 'soil heterogeneity'"**

In uniformity trials, field workers tried to make the fields as similar as possible, yet there is natural variation ("heterogeneity") among the fields.

Fisher clearly pointed out here that errors are not mistakes: they are not due to "careless in measuring …"

In statistics, error means the variation in experimental outcomes. There is natural variation in experiment units, and more importantly, in how they respond to treatments. Even when the same treatment is applied to all experimental units, there still will be variation in outcomes. This is the error that we need to estimate.

Fisher described error this way:

"For more than fifteen years the attention of agriculturalists has been turned to the errors of field experiments. During this period, experiments of the uniformity trial type have demonstrated the magnitude and ubiquity of **that class of error which cannot be ascribed to carelessness in measuring the land or weighing the produce, and which is consequently described as due to 'soil heterogeneity'"

In uniformity trials, field workers tried to make the fields as similar as possible, yet there is natural variation or heterogeneity among the fields.

Fisher clearly pointed out here that errors are not mistakes: they are not due to "careless in measuring …"

In statistics, error means the variation in experimental outcomes. There is natural variation in experiment units, and more importantly, in how they respond to treatments. Even when the same treatment is applied to all experimental units, there still will be variation in outcomes. This is the error that we need to estimate.

The word "error" is a bit overloaded. In general, error means random variation due to unknown sources.

Error refers to the variation in measurements (about a mean value) that cannot be attributable to a specific source.

Depending on context, it could mean the error term in a statistical model or the residual from a fitted model, which are deviation from an expected value or from the sample mean.

It could also mean the magnitude of the error terms (e.g. when we say reducing the error).

When is a result significant?

At Fisher's time, researchers understood that if they want to compare two treatments, they have to account for the natural variation (heterogeneity) of the field (i.e., *the error*).

One possible approach is described as follows:

> "… From these he can calculate a **standard error**, or rather an estimate of the standard error, to which the experiment is subject; and, if the observed difference is many times greater than this standard error, he claims that it is significant. …"

Basically, Fisher is describing the two-sample *t*-test for comparing two group means: we take the difference of the two group means, and see how many times greater it is than the estimated standard error. Fisher continued,

> "… If we thus put our trust in the theory of errors, all the calculation necessary is to find the standard error."

At Fisher's time, researchers understood that if they want to compare two treatments, they have to account for the natural variation or heterogeneity of the field.

One possible approach is described as follows:

"… From these he can calculate a standard error, or rather an estimate of the standard error, to which the experiment is subject; and, if the observed difference is many times greater than this standard error, he claims that it is significant. …"

Basically, Fisher is describing the two-sample t-test for comparing two group means: we take the difference of the two group means, and see how many times greater it is than the estimated standard error. Fisher continued,

"… If we thus put our trust in the theory of errors, all the calculation necessary is to find the standard error."

Standard error is the squared root of the variance of a statistic.

Fisher explained "If we thus put our trust in the theory of errors, all the calculation necessary is to find the standard error. In the simple case chosen above (in which, for simplicity, it is assumed that each of the two acres beats the other equally often) all that is necessary is to multiply each of the ten errors by itself, thus forming its square, to find the average of the ten squares and to find the square root of the average. The average of the ten squares is called the variance, and its square root is called the standard error. The procedure outlined above, relying upon the theory of errors, involves some assumptions about the nature of field errors; but these assumptions are not in fact disputed, and have been extensively verified in the examination of the results of uniformity trials."

What did Fisher mean by a valid estimate of error?

> "The error of which an estimate is required is that in the difference in yield between the area marked A and the area marked B, i.e., it is an error in the difference between plots treated differently in respect of the manure tested. The estimate of error afforded by the replicated trial depends upon differences between plots treated alike. An estimate of error so derived will only be valid for its purpose if we make sure that, in the plot arrangement, **pairs of plots treated alike are not nearer together, or further apart than, or in any other relevant way, distinguishable from pairs of plots treated differently.**"

This is a mouthful, let's try to summarize it:

- For testing the difference between two group means, we need to estimate the "error in the difference between two plots treated differently".
- However, our estimation of the (standard) error is based on "differences between plots treated alike".
- For this error estimation to be valid for comparing two group means, any two units chosen from the same group must be no more and no less similar than any two units chosen from different groups.

What did Fisher mean by a valid estimate of error?

"The error of which an estimate is required is that in the difference in yield between the area marked A and the area marked B, i.e., it is an error in the difference between plots treated differently in respect of the manure tested. The estimate of error afforded by the replicated trial depends upon differences between plots treated alike. An estimate of error so derived will only be valid for its purpose if we make sure that, in the plot arrangement, pairs of plots treated alike are not nearer together, or further apart than, or in any other relevant way, distinguishable from pairs of plots treated differently."

This is a mouthful, let's try to summarize it:

For testing the difference between two group means, we need to estimate the "error in the difference between two plots treated differently".

However, our estimation of the (standard) error is based on "differences between plots treated alike".

For this error estimation to be valid for comparing two group means, any two units chosen from the same group must be no more and no less similar than any two units chosen from different groups.

Randomization

Fisher clarified why randomization can help us obtain an valid estimation of error:

> "One way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random, **so that no distinction can creep in between pairs of plots treated alike and pairs treated differently**; in such a case an estimate of error, derived in the usual way from the variations of sets of plots treated alike, may be applied to test the significance of the observed difference between the averages of plots treated differently."

In what follows, Fisher is describing the test of significance from a permutation-test perspective:

> "The estimate of error is valid, because, if we imagine a large number of different results obtained by different random arrangements, the ratio of the real to the estimated error, calculated afresh for each of these arrangements, will be actually distributed in the theoretical distribution by which the significance of the result is tested."

An important point Fisher trying to make here is that we can analyze the test results from a completely randomized experiments using the "theoretical distribution"—the normal theory distributions where we assume that the error terms in the model are i.i.d. normal with a constant variation.

Fisher clarified why randomization can help us obtain an valid estimation of error:

"One way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random, so that no distinction can creep in between pairs of plots treated alike and pairs treated differently; in such a case an estimate of error, derived in the usual way from the variations of sets of plots treated alike, may be applied to test the significance of the observed difference between the averages of plots treated differently."

In what follows, Fisher is describing the test of significance from a permutation-test perspective:

"The estimate of error is valid, because, if we imagine a large number of different results obtained by different random arrangements, the ratio of the real to the estimated error, calculated afresh for each of these arrangements, will be actually distributed in the theoretical distribution by which the significance of the result is tested."

An important point Fisher trying to make here is that we can analyze the test results from a completely randomized experiments using the "theoretical distribution"—the normal theory distributions where we assume that the error terms in the model are i.i.d. normal with a constant variation.

Here "real error" means the population parameter and "estimated"error" is the statistic used to estimate the parameter.

We will briefly discuss permutation test later in the term.

Generalizability: sample distribution versus population distribution

One thing that Fisher did not mention in this paper is the **generalizability** of the experimental conclusions.

The variability in a sample may not reflect the variability in a target population: for the sample to be representative of the population, we need to make sure that the sample (the experimental units) is selected from the target population completely at random.

With randomization, Fisher obtained valid test for his field experiments, but the conclusions from the experiments can only be generalized to the fields in the Rothamsted station where the experiments were carried out.

Similar experiments have to be repeated in other places for the conclusions to hold true more generally.

One thing that Fisher did not mention in this paper is the generalizability of the experimental conclusions.

The variability in a sample may not reflect the variability in a target population: for the sample to be representative of the population, we need to make sure that the sample (the experimental units) is selected from the target population completely at random.

With randomization, Fisher obtained valid test for his field experiments, but the conclusions from the experiments can only be generalized to the fields in the Rothamsted station where the experiments were carried out.

Similar experiments have to be repeated in other places for the conclusions to hold true more generally.

Spacial versus temporal variation

One subtle point is that Fisher seems to equate the between-year (year-to-year) variation of the same field (temporal variation) to the between-plot (plot-to-plot) variation within a year (spatial variation).

This, as pointed out by Terry Speed, requires a lot of additional assumptions.[1]

One subtle point is that Fisher seems to equate the between-year (year-to-year) variation of the same field (temporal variation) to the between-plot (plot-to-plot) variation within a year (spatial variation).

This, as pointed out by Terry Speed, requires a lot of additional assumptions.

1. Speed, “Introduction to `The arrangement of field experiments.’”.

Summary: the roles of replication and randomizaiton

Replication has two aims:

1. We need replication to estimate the error.
2. We can typically reduce the magnitude of the standard error (of a comparison) by increase the sample sizes (increase the **precision** of an experiment).

Randomization safeguards the **validity** of the error estimation:

1. If experimental units are completely randomized to different treatment groups, then two units within a group will not more or less similar than two units from different groups.

An error estimation is only valid when it matches the design. When thinking about the errors, we want to **follow the randomization**.

The three basic principles introduced in Fisher's 1926 paper are: replication, randomization and blocking.

Replication has two aims:

1. We need replication to estimate the error.

1. We can typically reduce the magnitude of the standard error, say of a comparison, by increase the sample sizes, and thus increase the precision of an experiment.

Randomization safeguards the validity of the error estimation:

1. If experimental units are completely randomized to different treatment groups, then two units within a group will not more or less similar than two units from different groups.

An error estimation is only valid when it matches the design. When thinking about the errors, we want to follow the randomization.

We will discuss the role of blocking later in the quarter.