1. Introduction

Welcome to the series of E-learning modules on Randomized tests, Nonrandomized tests and P values. In this module we are going cover the concept of randomized tests nonrandomized tests and P values, its interpretation and role in testing of hypothesis.

By the end of this session, you will be able to:

- Explain randomized and nonrandomized tests
- Compare these tests with Parametric tests
- Interpret P values and explain its importance
- Explain the advantages and misunderstandings about P values

Hypothesis testing is generally used when some comparison is to be made.
This comparison may be a single observed value versus some hypothesized quantity.
For example, the number of babies born in a single delivery to mothers undergoing fertility treatment as compared with typical singleton birth.

Or it may be a comparison of two or more groups. For example: mortality rates in intensive care unit patients who require renal replacement therapy versus those who do not.
The choice of which statistical test to use depends on the format of the data and the study design.

Randomized test
A randomized test T is the one in which no test statistic is used. The decision about the rejection of the null hypothesis H naught is taken, if it satisfies some predefined criterion. For example if it is decided that H naught will be rejected if on tossing of a coin it falls with the head on the upper side and will be accepted if it falls with tail on the upper side .

Non-Randomized test
A test T of a hypothesis H is said to be non randomized if the hypothesis H naught is rejected on the basis that a test statistic belongs to the critical region C that is Phi of (x one, x two, etc till xn) belongs to C.

It will be recalled that for hypothesis testing problems involving discrete distributions, it is usually not possible to choose a critical region consisting of realizable values of the statistic of size exactly apha, where alpha is some prescribed value.

In the hypothesis testing procedures considered so far, the sample space of observations Omega is partitioned into two regions, C and A . We can express this in terms of a function Phi as follows. Let Phi of small x be equal to Probability of, reject H naught when capital X is equal to small x

For a non-randomized test with rejection region C, Phi for a region C is just its indicator function. That is, Phi of x equal to one if x belongs to C, zero if x does not belong to C

We will extend this, to allow for some different action, if the outcome x is on the boundary of

the critical region.
The other action effectively is performing an auxiliary experiment such as tossing a coin with probability of (heads) equal to p; if heads results, reject H naught.
If tails results, h naught is accepted.
The value of p is chosen to make the Probability of (rejecting H naught) the desired value.

More formally, for a test with critical region C and a value of X equal to x_0 on the boundary, we may define Phi of x equal to one if x belongs to C, p if x is equal to x zero, zero if x is not equal to x zero and x does not belong to C. Where p is appropriately chosen.

Not many statisticians like randomized tests, in practice, because the use of them means that two statisticians could make the same assumptions, observe the same data, apply the same test, yet come to different conclusions.
This is perhaps another reason why it is desirable to report a P value for an experiment rather than arbitrarily choose a value of alpha, and use a somewhat contrived method to achieve a test with exactly this alpha.

2. Comparison with Parametric Tests

Comparison with Parametric tests
The primary goal of both parametric tests and randomized tests is to test some null hypothesis, although the null hypothesis is distinctly different from what it would be with a parametric test.

In parametric tests we randomly sample from one or more populations. We make certain assumptions about those populations, most commonly that they are normally distributed with equal variances. We establish a null hypothesis that is framed in terms of parameters.

We use our sample statistics as estimates of the corresponding population parameters, and calculate a test statistic.
We then refer that test statistic to the tabled sampling distribution of the statistic, and reject the null if our test statistic is extreme relative to the tabled distribution.
Randomization tests differ from parametric tests in almost every respect.

There is no requirement that we have random samples from one or more populations, in fact we usually have not sampled randomly.
We rarely think in terms of the populations from which the data came, and there is no need to assume anything about normality or homoscedasticity.

Our null hypothesis has nothing to do with parameters, but is phrased rather vaguely, as, for example, the hypothesis that the treatment has no effect on the how participants perform.
That might be phrased a bit more precisely by saying that, under the null hypothesis, the score that is associated with a participant is independent of the treatment that person received.

Since we are not concerned with populations, we are not concerned with estimating or even testing characteristics of those populations.
We do calculate some sort of test statistic; however we do not compare that statistic to tabled distributions.

Instead, we compare it to the results we obtain when we repeatedly randomize the data across the groups, and calculate the corresponding statistic for each randomization.
Even more than parametric tests, randomization tests emphasize the importance of random assignment of participants to treatments.

The randomization test of independence is used when you have two nominal variables.
A data set like this is often called an "R into C table," where R is the number of rows and C is the number of columns.
The randomization test is more accurate than the chi-squared test or G-test of independence when the expected numbers are small.

Fisher's exact test would be just as good as a randomization test, but there may be situations where the computer program you're using can't handle the calculations required for the Fisher's test.

Randomization (or permutation) tests free the experimenter from the constraints of random sampling, a known error distribution and equal variances,
These tests give a direct answer to the question “how likely is such a large or small result, if the applied treatments had no effect?”
The “result” may be the difference in mean responses, a correlation coefficient or any other value of interest.

3. Advantages

Advantages:
A randomization test is not a different statistical test but a different, and always valid, method of determining statistical significance.
The familiar ***t***-test and ***F***-test can be carried out by data permutation without any parametric assumptions being fulfilled.
A particular advantage of this method is that unbalanced designs and missing values are easily accommodated.

Even with only a small number of subjects the number of permutations will be large and a computer is necessary if the randomization test is to be of practical value.
To make this method of determining statistical significance generally available, an interactive microcomputer program, forming a comprehensive package for the design and analysis of experiments, has been prepared.

The procedure for a randomization test is:
Step 1: Devise a test statistic which is large if your hypothesized process is strong, and small if it is weak
Step 2: Define your null hypothesis
Step 3: Create a new data set consisting of your data, randomly rearranged. How exactly it is rearranged depends on your null hypothesis

Step 4: Calculate your test statistic for this data set, and compare it to your true value.
Step 5: Repeat steps three and four many times preferably several hundred
Step 6: your true test statistic is greater than ninety five percent of the random values, then you can reject the null hypothesis at p less than zero point zero five.

Many people are now promoting the use of randomization tests even when parametric and nonparametric tests exist.
Statistical Educators are beginning to use randomization tests as the introduction to statistics, because in many ways it is easier to grasp.
There are two general models to testing.
One of them is based on the assumption of random sampling from a population, and is usually called the "population model".

In many situations, any model actually leads to the same conclusion.
In a way, the approach based on the population model can be seen as an approximation to the randomization test.
Interestingly, Fisher was the one who proposed the randomization model and suggested that it should be the basis for our inferences.

To test the hypotheses means to choose one hypothesis or the other; that is, to make a decision.
We have a sample X from the relevant family of distributions and a statistic T of (X).
A nonrandomized test procedure is a rule delta of X that assigns two decisions to two disjoint subsets, C and A, of the range of T of X.

We equate those two decisions with the real numbers zero and one.
so delta of (X) is a real-valued function,
Delta of x is equal to zero for T of (x) belongs to C; one for T of (x) belongs to A.
If delta (X) takes the value zero, the decision is to reject
If delta (X) takes the value one, the decision is not to reject
If the range of delta of (X) is {zero, one}, the test is a nonrandomized test.

Sometimes it is useful to choose the range of delta of x as some other set of real numbers, such as {d zero, d one} or even a set with cardinality greater than two.
If the range is taken to be the closed interval [zero, one], we can interpret a value of delta of x as the probability that the null hypothesis is rejected.
If it is not the case that delta of x equals zero or one almost surely, we call the test a randomized test.

4. P-value

P value
Another quantitative measure for reporting the result of a test of hypothesis is the *p*-value.
The *p*-value is the probability of the test statistic being at least as extreme as the one observed given that the null hypothesis is true.

A small *p*-value is an indication that the null hypothesis is false.
Although there is often confusion, the p-value is not the probability of the hypothesis being true, nor is the p-value the same as the Type I error rate.

Traditionally, one rejects the null hypothesis if the p-value is less than or equal to the significance level **.**
The connection is that a hypothesis test that rejects the null hypothesis for all samples that have a p-value less than alpha will have a Type one error of alpha.

A significance level of zero point zero five would deem extraordinary any result that is within the most extreme five percent of all possible results under the null hypothesis.
In this case a p-value less than zero point zero five would result in the rejection of the null hypothesis at the five percent significance level.

When we ask whether a given coin is fair, often we are interested in the deviation of our result from the equality of numbers of heads and tails.
In this case, the deviation can be in either direction, favouring either heads or tails.

Thus, in the example of fourteen heads and six tails, we may want to calculate the probability of getting a result deviating by at least four from parity in either direction (two-sided test).
This is the probability of getting at least fourteen heads or at least fourteen tails.

In the above example we thus have:
Null hypothesis is that it is a fair coin with probability of (heads) equal to zero point five.
Observation is that we have fourteen heads out of twenty flips.
P-value of observation of given null hypothesis is **equal to** Probability of (greater than or equal to fourteen heads or greater than or equal to fourteen tails).
This is equal to two into (one minus Probability of less than fourteen) which is equal to zero point one one five.

The calculated p value exceeds zero point zero five. Hence the observation is consistent with the null hypothesis.
The observed result of fourteen heads out of twenty flips can be ascribed to chance alone, as it falls within the range of what would happen ninety five percent of the time were the coin in fact fair.

In our example, we fail to reject the null hypothesis at the five percent level. Although the coin did not fall evenly, the deviation from expected outcome is small enough to be consistent with chance.

However, had one more head been obtained, the resulting p-value (two-tailed) would have been zero point zero four one four.
This time the null hypothesis, that the observed result of fifteen out of twenty flips can be ascribed to chance alone, is rejected when using a five percent cut off.

Misunderstandings: The data obtained by comparing the p-value to a significance level will yield one of two results.
Either the null hypothesis is rejected, or the null hypothesis cannot be rejected at that significance level (which however does not imply that the null hypothesis is true).
A small p-value that indicates statistical significance does not indicate that an alternative hypothesis is correct.

There are several common misunderstandings about p-values.
The p-value is not the probability that the null hypothesis is true.
In fact, frequentist statistics does not, and cannot, attach probabilities to hypotheses.
The p-value is not **the** probability that a finding is "merely a fluke."

As the calculation of a p-value is based on the assumption that a finding is the product of chance alone, it patently cannot also be used to gauge the probability of that assumption being true.
This is different from the real meaning which is that the p-value is the chance of obtaining such results if the null hypothesis is true.
The p-value is *not* the probability of falsely rejecting the null hypothesis.
The significance level of the test is not determined by the p-value.

The significance level of a test is a value that should be decided upon before the data are viewed, and is compared against the p-value or any other statistic calculated after the test has been performed.
The p-value does not indicate the size or importance of the observed effect.
The two do vary together however – the larger the effect, the smaller sample size will be required to get a significant p-value.

5. Interpretation of P-value and Conversion of Z statistic to P-value

Interpretation of P value
It was common in the past for researchers to classify results as statistically 'significant' or 'non-significant'.
The results were based on whether the *P* value was smaller than some prespecified cut point, commonly zero point zero five.
This practice is now becoming increasingly obsolete, and the use of exact *P* values is much preferred.

This is partly for practical reasons, because the increasing use of statistical software renders calculation of exact *P* values increasingly simple as compared with the past when tabulated values were used.
The presentation of exact *P* values allows the researcher to make an educated judgment as to whether the observed effect is likely to be due to chance
This, taken in the context of other available evidence, will result in a far more informed conclusion being reached.

P-value answers the question: What is the probability of the observed test statistic when null hypothesis is true?
Thus, smaller P-values provide stronger evidence against null hypothesis.

Points to be noted:
A P value does not provide any measure of the size of an effect, and cannot be used in isolation to inform clinical judgement.
P values are affected both by the magnitude of the effect and by the size of the study from which they are derived, and should therefore be interpreted with caution.

In particular, a large P value does not always indicate that there is no association and, similarly, a small P value does not necessarily signify an important clinical effect.
The test statistic is converted to a conditional probability called a P-value.
The P- value answers the question “If the null hypothesis were true, what is the probability of observing the current data or data that is more extreme?”

We apply the following conventions:
When p value greater than zero point one zero implies the observed difference is “not significant”
When p value less than or equal to zero point one zero implies the observed difference is “marginally significant”
When p value less than or equal to zero point zero five implies the observed difference is “significant”
When p value less than or equal to zero point zero one implies the observed difference is

“highly significant”.
Use of “significant” in this context means “the observed difference is not likely due to chance.”

Conversion of Z statistic to P value

For H one: mu greater than mu naught implies P equal to probability of Z greater than Z statistic equal to right-tail beyond Z statistic
For H one: mu less than mu naught implies P equal to probability of Z less than Z statistic equal to left-tail beyond Z statistic
For H one: mu not equal to mu naught implies P equal to twice probability of one-tailed P-value.

Here’s a summary of our learning in this session where we have understood the following concepts :

- Randomized and nonrandomized tests
- Comparison with Parametric tests
- P -value , its interpretation , misunderstandings and importance
- Conversion of Z statistic to P-value