Two Way ANOVA

Block design tests

Paweł Polak

April, 2016

STAT W4413: Nonparametric Statistics - Lecture 16

Block design tests

Example

- It is well-known that applying nitrogen to wheat plants has significant effect on its productivity.
- Therefore, several different methods of applying nitrogen to wheat have been proposed.
- Suppose that we would like to compare these different methods and test which one is more effective.
- Clearly, the growth of wheat depends on the quality of the soil as well.
- In particular, the moisture and fertility of the soil may affect the productivity.

Block design tests

- In the k-sample test, our strategy was to control the quality of the soil and ensure that the moisture/fertility is the same everywhere. Then apply nitrogen to wheat (with different methods) and obtain k independent samples.

- However, as we mentioned before the main disadvantage is that any difference in the soil of different experiments may affect our conclusion.

- It is hence very important to carefully control the soil quality. This might not be straightforward in many situations.

控制变量，s.t.H0同分布

Block design tests

- Suppose that we partition the entire field (on which the wheat plants are growing) into homogenous blocks.
- Since the blocks are much smaller than the entire field we believe that the quality of soil is almost the same in each block.
- This means that each block of soil has very similar fertility and moisture.
- Then for the plants in each block we apply the different methods of applying nitrogen and after a while we measure their productivity.

Block design tests

The result we obtain is shown in the following table:

	Block			
Treatment	1	2	...	b
1	X_{11}	X_{12}	...	X_{1b}
2	X_{21}	X_{22}	...	X_{2b}
⋮	⋮	⋮	⋮	⋮
k	X_{k1}	X_{1k}	...	X_{kb}

X_{ij} is an indicator of the productivity of the wheat plant in a certain block j (fixed fertility and moisture) for a certain nitrogen treatment i.

We want to test if different treatments are different or the same.

Note that we expect

- X_{ij} to be different from $X_{i,j+1}$, since they are different blocks and have been exposed to different fertilizer/moisture level.
- X_{ij} to be different from $X_{i+1,j}$ since they have received different treatments (nitrogen method).

Block design tests

We model all these differences in the following equation:

$$X_{ij} = \mu + t_i + \beta_j + \epsilon_{ij}. \tag{1}$$

- ϵ_{ij} is considered as a noise in measuring the productivity, $\epsilon_{ij} \overset{iid}{\sim} F$ and $\mathbb{E}(\epsilon_{ij}) = 0$. it is important to note that the statistics of the noise do not change from one block/treatment to another block/treatment.
- t_i represents the impact of different treatments on the productivity. In other words, the effect is considered to be a shift in the distribution, and the shape remains unchanged.
- β_j represents the impact of fertility/moisture on the growth of the wheat. Again as is clear, we model the effect as a shift in the distribution.
- μ is the overall mean.

Block design tests

$$X_{ij} = \mu + t_i + \beta_j + \epsilon_{ij}. \tag{2}$$

Let's discuss several aspects of this model:

(1) Interactions are ignored:

- Suppose that the treatment is fixed and we are considering different blocks under the same treatment.
- As you can see the impact of different blocks is modeled by β_j.
- But it is conceivable that the soil with higher moisture may help the nitrogen treatment and boost its performance. In other words, instead of having the shift t_i for all different blocks, for some blocks the shift is higher like $2t_i$. This is known as interaction between different effects.
- In case we would like to model interactions we should also include terms of the form $g(t_i, \beta_j)$. But to simplify the problem these terms are not considered.
- In many examples there is a good reason to believe that the interaction is zero.

Block design tests

(2) Identifiability:

- Assume that there is no noise, i.e., $\epsilon_{ij} = 0$. Can we calculate μ, t_i and β_j?
- The answer to this question is in the scope of what is known as the identifiability of the parameters.
- Unfortunately, the answer is NO in our model.
- To see the problem, suppose that we can calculate the solution and it is give by μ^*, t_i^* and β_j^*.
- Then it is straightforward to prove that for every α, μ^*, $t_i^* - \alpha$ and $\beta_j^* + \alpha$ will satisfy all equations in.[1]
- To make the model identifiable we assume that

$$\sum_{i=1}^{k} t_i = 0, \qquad \sum_{i=1}^{b} \beta = 0.$$

- Under these two conditions you can easily show that we can uniquely identify μ, t_i and β_j from X_{ij}(when $\epsilon_{ij} = 0$).

[1] We have only $t + b + 1$ unknowns and bt linear equations for them. So, we must be able to solve the set of equations. However, most of the linear equations that we have are linearly dependent, e.g., the four equations $X_{11} = \mu + t_1 + \beta_1$, $X_{12} = \mu + t_1 + \beta_2$, $X_{21} = \mu + t_2 + \beta_1$ and $X_{22} = \mu + t_2 + \beta_2$ are dependent. You can easily check that $X_{11} - X_{12} + X_{21} - X_{22} = 0$. There are many other linear dependencies that you can build based on this simple example.

Block design tests

In the original problem, we want to test whether different methods of applying nitrogen are essentially the same, i.e., we would like to test

$$H_0 : t_1 = t_2 = \ldots = t_k \text{ vs. } H_1 : \text{ at least one of the equalities is violated.}$$

We again study this problem under the parametric and nonparametric settings.

Parametric setting and two-way ANOVA

Suppose that $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. We follow a similar approach to what we did in one-way ANOVA. Again we try to break the total variance of our sample into its pieces and will then interpret those pieces and obtain a statistic that can distinguish H_0 from H_1. Define

$$\begin{aligned}
\bar{X}_{i\cdot} &\triangleq \frac{1}{b}\sum_{j=1}^{b} X_{ij} \\
\bar{X}_{\cdot j} &\triangleq \frac{1}{k}\sum_{i=1}^{k} X_{ij} \\
\bar{X}_{\cdot\cdot} &\triangleq \frac{1}{kb}\sum_{i=1}^{k}\sum_{j=1}^{b} X_{ij} \qquad (3)
\end{aligned}$$

Parametric setting and two-way ANOVA

Based on this definition, we can now break the variance term to

$$
\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\bar{X})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot}+\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}+\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})^2+\sum_{i=1}^{k}b(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2+\sum_{j=1}^{b}k(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})^2.
\end{aligned}
\tag{4}
$$

Try to prove the last equality for yourself. The three terms in the second line have very interesting interpretations.

- The first term is the *within block/ treatment variation*. In other words, the difference between blocks and treatments will not affect this term. It is straightforward to confirm that

$$X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot}=\epsilon_{ij}-\frac{1}{b}\sum_{j}\epsilon_{ij}-\frac{1}{k}\sum_{i}\epsilon_{ij}+\frac{1}{kb}\sum_{i}\sum_{j}\epsilon_{ij}. \tag{5}$$

Parametric setting and two-way ANOVA

As you can see this expression does not capture any difference in treatments or blocks. In fact the other two terms in (4) capture those effects. In particular as is clear the term $\sum_{i=1}^{k} b(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})$ is capturing the difference between treatments. Finally, the term $\sum_{j=1}^{b} k(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$ captures the difference between different blocks. To understand these claims, you should try to prove:

$$\mathbb{E}\left(\sum_{i=1}^{k} b(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})\right)^2 = (k-1)\sigma^2 + b\sum_{i=1}^{k}(t_i - \frac{1}{k}\sum_{i=1}^{k} t_i)^2. \qquad (6)$$

Parametric setting and two-way ANOVA

Note that since we have assumed $\sum_{i=1}^{k} t_i = 0$, we can simplify this formula to

$$\mathbb{E}\left(\sum_{i=1}^{k} b(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})\right)^2 = (k-1)\sigma^2 + b\sum_{i=1}^{k} t_i^2.$$

Therefore, we define the following statistic that can differentiate between H_0 and H_1

$$\boxed{F = \frac{b\sum_{i=1}^{k}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2/(k-1)(b-1)}.} \tag{7}$$

Parametric setting and two-way ANOVA

Try to prove the following properties for this statistics:

1. $\sum_{i=1}^{k}\sum_{j=1}^{b}\mathbb{E}(X_{ij}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_{\cdot\cdot})^2 = (k-1)(b-1)\sigma^2.$
2. $\mathbb{E}\left(\sum_{i=1}^{k} b(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})\right)^2 = (k-1)\sigma^2 + b\sum_{i=1}^{k}(t_i - \frac{1}{k}\sum_{i=1}^{k} t_i)^2.$
3. $\boxed{\text{Under } H_0, F \sim F_{k-1,(k-1)(b-1)}.}$

As you can see as the variation between different treatments is increasing, so does the value of F (on average). Therefore, this leads us to the following criteria for rejecting the null hypothesis:

$$\text{If } F > c, \text{ reject } \; H_0.$$

The approach we described above is known as two-way analysis of variance (ANOVA).

Nonparametric tests: Permutation test

In many applications the Gaussian assumption on ϵ_{ij} is not valid.

Therefore, we would like to avoid that assumption.

Let $\epsilon_{ij} \overset{iid}{\sim} G$, where G is a CDF whose median is zero.

Under this setting, as you can see the F statistic that we introduced above is still capable of distinguishing between H_0 and H_1.

However, it is not correct to claim that under the null hypothesis it has F-distribution, as it was a direct implication of the Gaussian assumption.

As before, we design a permutation test that can use this statistic and provide an accurate test.

Nonparametric tests: Permutation test

Recall,

$$X_{ij} = \mu + t_i + \beta_j + \epsilon_{ij}. \tag{8}$$

$H_0 : t_1 = t_2 = \ldots = t_k$ vs. $H_1 :$ at least one of the equalities is violated.

Suppose that the null hypothesis is true.

If we look at the samples in each block j they are drawn from the same distribution.

Therefore, if we start permuting the samples in each block we obtain a new sample from the same distribution (Can we permute the entire samples? Not only the samples in each block!).

In other words we can permute the samples in each block and assign the samples to different treatments and still have a new set of samples from the same distribution.

Nonparametric tests: Permutation test

This observation leads us to the following permutation test for the block design problem:

1. Permute observations within each block. Doing this in all blocks can generate $(k!)^b$ new samples from the same distribution.
2. For each new sample calculate the F statistic as above.
3. Calculate the p-value of the permutation test as $\boxed{p_{perm} = \frac{\# F_i \geq F_{obs}}{(k!)^b}.}$

Nonparametric tests: Permutation test

Example

Consider a a very simple example with two treatments and two different blocks. Let's assume that $X_{11} = 1, X_{21} = 1.5, X_{12} = 1.7, X_{22} = 1.9$.

Since we would like to permute the data in each block we can also consider $x_{11}^{new} = 1.5$, and $x_{21}^{new} = 1$. Also, regarding the samples in the second block we can consider $X_{12} = 1.9$ and $X_{22} = 1.7$. Considering all the permutations we can have 4 different permutation samples.

Nonparametric tests: Permutation test

- The procedure is exactly the same as the permutation test that we had before
- the only difference is the way we generate new samples from the observed samples.
- the F statistic that is employed in the second step of the permutation test, can be replaced with any other statistic that is capable of distinguishing between H_0 and H_1.
- Another statistic that is more popular in nonparametric setting is the Friedman's statistic that will be discussed next.

Nonparametric tests: Friedman's statistic

As we discussed, usually in nonparametric settings we are interested in designing tests whose distribution under H_0 is free of the distribution G.

As usual we would like to use the rank statistics. Here is a step by step procedure to calculate Friedman's statistic.

1. Rank the data in each block and let R_{ij} denote the rank of X_{ij} in the set $\{X_{i1}, X_{i2}, \ldots, X_{ik}\}$. Clearly, under the null hypothesis $\mathbb{E}(R_{ij}) = \frac{k+1}{2}$.
2. For each treatment calculate the average rank $\bar{R}_i = \frac{1}{b}\sum_{j=1}^{B} R_{ij}$. Under H_0 we have $\mathbb{E}(\bar{R}_i) = \frac{k+1}{2}$.
3. Calculate the Friedman's statistic which is defined as

$$FM = \frac{12B}{k(k+1)} \sum_{i=1}^{k} \left(\bar{R}_i - \frac{k+1}{2} \right)^2 .$$

Nonparametric tests: Friedman's statistic

$$FM = \frac{12B}{k(k+1)} \sum_{i=1}^{k} \left(\bar{R}_i - \frac{k+1}{2} \right)^2 .$$

Under H_0 all $\bar{R}_i$ tend to be around $\frac{k+1}{2}$ and therefore, FM tends to be small.

Under H_1, on the other hand, FM tends to be larger.

Note that under the asymptotic settings FM has χ^2 distribution with $k-1$ degrees of freedom. (That is why we have $\frac{12B}{k(k+1)}$.)

We will not prove and use this fact in this course. We should only note that we can easily plug this statistic into the permutation test and obtain a test that outperforms ANOVA F-test in many cases.

J. Higgins, Introduction to modern nonparametric statistics.