k-Sample Test

Paweł Polak

April, 2016

STAT W4413: Nonparametric Statistics - Lecture 15

K-Sample tests

In the two sample tests, we considered

$$X_1, X_2, \ldots, X_n \sim F \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \sim G,$$

and asked questions regarding the relation of F and G.

Example

What if we would like to show that under the same education / family income, the crime rate is independent of the race.

Start with collecting statistics of the crime rates in different states/neighborhoods for the people that have the same education/family income, and construct tables of the form:

Race	Statistics
Hispanic	$X_{11}, X_{12}, \ldots, X_{1n_1}$
White	$X_{21}, X_{22}, \ldots, X_{2n_2}$
African American	$X_{31}, X_{32}, \ldots, X_{3n_3}$
Asian	$X_{41}, X_{42}, \ldots, X_{4n_4}$

Each X_{ij} in this table presents the crime rate of a certain race in a certain neighborhood. Based on this data we would like to test our hypothesis that says the crime rates are the same for different races.

k-sample parametric tests

One way to cast this problem as a testing problem is the following:

- Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma^2), \ldots, X_{k1}, X_{k2}, \ldots, X_{kn_k} \sim N(\mu_k, \sigma^2)$.
- Note that all the samples have the same variance.
- If our claim (that different races have the same crime rate) is true, we will have $\mu_1 = \mu_2 = \ldots = \mu_k$.
- Therefore, we test for

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad \text{vs.} \quad H_1 : \text{at least for one } i, j, \mu_i \neq \mu_j.$$

How do we test this hypothesis?

Is it any different from the two sample test?

Can we use the two-sample test to address this problem as well?

k-sample parametric tests

Let's start with a simple example that we have only three groups.

group$_1$	1	2	3
group$_2$	1.5	0.5	2.1
group$_3$	1.2	2.3	2.9

The first idea that we would like to explore is inspired by the T-test.

- The fact that $\mu_1 = \mu_2 = \mu_3$, gives us three two-sample hypotheses, each of which can be checked with a two-sample T-test.
- These three hypotheses are $H_0' : \mu_1 = \mu_2, H_0'' : \mu_2 = \mu_3$, and $H_0''' : \mu_1 = \mu_3$. If, we accept all these three hypotheses, then we can accept H_0 as well.
- Otherwise we should reject H_0.
- Therefore, we can perform the two-sample T-test to evaluate the validity of H_0', H_0'', and H_0'''.

Is this a good approach?

k-sample parametric tests

We can characterize the significance level of this test.

- Since H_0', H_0'', and H_0''' are similar we set the significance level to p for all of them.
- Now we can characterize the significance level of testing H_0 with this pairwise tests.

$$
\begin{aligned}
&\mathbb{P}(\mathit{rejecting}\ H_0 \mid H_0) = 1 - \mathbb{P}(\mathit{accepting}\ H_0|H_0) \\
&= 1 - \mathbb{P}(\mathit{accepting}\ H_0' \cap \mathit{accepting}\ H_0'' \cap \mathit{accepting}\ H_0''' \mid H_0) \\
&\overset{a}{=} 1 - \mathbb{P}(\mathit{accepting}\ H_0' \mid H_0)\mathbb{P}(\mathit{accepting}\ H_0'' \mid H_0)\mathbb{P}(\mathit{accepting}\ H_0''' \mid H_0) \\
&= 1 - (1-p)^3.
\end{aligned}
$$

Clearly, equality (a) is not exactly true. To obtain that we assume the independence of the three events. However the independence is not true. Despite this inaccurate assumption the result that we obtain from this assumption is not necessarily very far from the truth and it helps us understand the problem more clearly.

k-sample parametric tests

This result can be extended to the k-group setting. For the k-sample situation we have to consider $\binom{k}{2}$ pairwise comparisons.

If we set the significance level of the test to p, and assume that the pairwise tests are independent of each other (even though they are not), we obtain the following significance level for the k-sample test:

$$\mathbb{P}(\textit{rejecting } H_0 \mid H_0) = 1 - (1-p)^{\frac{k(k-1)}{2}}. \tag{1}$$

To understand the final results, let's look at two examples.

k-sample parametric tests

Example

Let $K = 10$ and $p = 0.01$. Then the last result shows that $\mathbb{P}(\textit{rejecting } H_0 \mid H_0) = 0.4$. In other words, even though the significance level for each pairwise test is pretty low, the significance level of the final test is very high. This is also very intuitive. Even though the chance that we make a mistake in each test is low, but since we do testing multiple times, the chance that we make mistake in at least one of them is going to be high.

To address this issue, one may use lower values of p as we do in the next example.

k-sample parametric tests

Example

Let $K = 10$ and $p = 0.001$. Then (1) shows that $\mathbb{P}(\textit{rejecting } H_0 \mid H_0) = 0.05$. This test has an acceptable significance level. But, it suffers from another problem and that is, since the significance level for each test is extremely low, the test does not have good power. In other words, it may accept the null even if the null hypothesis is violated by one of the μ_is.

In conclusion, using pairwise T-test is not a good method for testing H_0 versus H_1.[1]

Instead one should use the ANOVA.

[1] Please note that this statement is not necessarily true. In some situations this test is optimal. The understanding of the situations in which this test is optimal and proving its optimality are beyond the scope of this course.

One-way analysis of variance (ANOVA)

To understand ANOVA, consider the table we had before.

group$_1$	1	2	3
group$_2$	1.5	0.5	2.1
group$_3$	1.2	2.3	2.9

If we look at the numbers in each group, we see certain variation. If we calculate the variance of the data samples in each row, the result is an estimate for σ^2. Define

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

Then according to our discussion

$$\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

is an approximation for $(n_i - 1)\sigma^2$. Now define the average of the entire sample as

$$\bar{X} = \frac{1}{n_1 + n_2 + \ldots + n_k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}.$$

and consider the entire variance of the data.

One-way analysis of variance (ANOVA)

We do some calculations to simplify the total variance.

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i+\bar{X}_i-\bar{X})^2 \\
&\overset{(b)}{=} \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})^2 \quad (2) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2 \quad (3)
\end{aligned}$$

Prove equality (b) for yourself. The last line of (3) has two interesting terms. The first term as we discussed before reflects the variations within the groups. The second term on the other hand is reflecting the variation between different groups. We will describe this in details later. Another interesting fact that you will prove in the homework is that the two terms in (3) are independent. Based on this discussion we define the ANOVA test statistic as

One-way analysis of variance (ANOVA)

$$F^* = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(N-k)}.$$

In order to see how this statistic works, let us simplify the problem a little bit and assume that all the sample sets have the same size, i.e., $n_1 = n_2 = \ldots = n_k$.

Therefore, we drop the subscript and consider n as the size of the group.

The first thing that we would like to show is that the numerator captures the variations between the groups.

Here is our first claim.

One-way analysis of variance (ANOVA)

Lemma

If the size of each sample set is n and we have k independent sample sets, then

$$\mathbb{E}\left(\sum_{i=1}^{k} n(\bar{X}_i - \bar{X})^2\right) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\left(\mu_i - \frac{1}{k}\sum_{j=1}^{k}\mu_j\right)^2.$$

I.e.,

The numerator has captured the within group variations.

When the null hypothesis is true the variation that we see is essentially due to σ^2 that is present in the data.

But, as soon as the group become in-homogenous we see extra term $\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2$ that reflects the difference between groups.

One-way analysis of variance (ANOVA)

Proof.

First note that $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ are independent samples and $\bar{X}_i \sim N(\mu_i, \frac{\sigma^2}{n})$. Second, note that

$$\bar{X} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} X_{ij} = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{n}\sum_{j=1}^{n} X_{ij}\right) = \frac{1}{k}\sum_{i=1}^{k}\bar{X}_i. \tag{4}$$

Define $\bar{Z}_i \triangleq \bar{X}_i - \mu_i$. Clearly, $\bar{Z}_i \sim N(0, \frac{\sigma^2}{n})$.
If we define $\bar{\mu} \triangleq \frac{1}{k}\sum_{i=1}^{k}\mu_i$ and $\bar{Z} \triangleq \frac{1}{k}\sum_{i=1}^{k}\bar{Z}_i$, then

$$\mathbb{E}\left(\sum_{i=1}^{k} n(X_i - \bar{X})^2\right) = n\mathbb{E}\left(\sum_{i=1}^{k}(\bar{Z}_i + \mu_i - \bar{Z} - \bar{\mu})^2\right) = n\sum_{i=1}^{k}\mathbb{E}(\bar{Z}_i - \bar{Z})^2 + n\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2. \tag{5}$$

Note that $\bar{Z}_i \overset{iid}{\sim} N(0, \sigma^2/n)$. Hence, $\frac{\sum_{i=1}^{k}(\bar{Z}_i - \bar{Z})^2}{\sigma^2/n}$ is a χ^2_{k-1} and

$$\mathbb{E}\sum_{i=1}^{k}(\bar{Z}_i - \bar{Z})^2 = \frac{(k-1)\sigma^2}{n} \tag{6}$$

Combining (5) and (6) completes the proof. □

One-way analysis of variance (ANOVA)

Now, let's discuss the denominator of F^*.

Lemma

The expected value of the variation within group is given by

$$\mathbb{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2/(N-k)\right)=\sigma^2.$$

The proof of this result is straightforward and is left to the reader.

One-way analysis of variance (ANOVA)

- Under the null hypothesis the expected value of numerator is the same as the expected value of the denominator.
- However, once we violate the null hypothesis the expected value of denominator does not change while the expected value of numerator is getting larger.
- Hence F^* increases.
- This discussion leads us to the following test:

$$\text{reject } H_0 \text{ if } F^* > c$$

In the next slides we prove that under the null hypothesis $F^* \sim F_{k-1,N-k}$ distribution.

ANOVA and F-distribution

We will prove that $F^* \sim F_{k-1,N-k}$.

- We first prove this result in the simple setting that all the groups have the same size n.
- Then we consider a more general setting and prove that the result holds under unequal group sizes as well.
- Our proof uses the following strategy. We first prove that numerator and the denominator are two independent χ^2 random variables and hence their ratio has F distribution.
- Since H_0 holds all groups have the same mean, i.e., $\mathbb{E}(X_{ij}) = \mu$.

ANOVA and F-distribution

- Our first claim is that if we prove the result for $\mu = 0$, then it will easily extend to the more general setting where μ is an arbitrary number.
- To prove this claim suppose that $\mu \neq 0$. Then define a new set of random variables $Z_{ij} = X_{ij} - \mu$.
- Clearly, under H_0, $Z_{ij} \overset{iid}{\sim} N(0, \sigma^2)$. Furthermore, we have

$$\begin{aligned} X_{ij} - \bar{X}_i &= \mu + Z_{ij} - \mu - \bar{Z}_i = Z_{ij} - \bar{Z}_i, \\ \bar{X}_i - \bar{X} &= \mu + \bar{Z}_i - \mu - \bar{Z}, \end{aligned}$$

  where $\bar{Z}_i = \frac{1}{n_i} \sum_{j=1}^{n} Z_{ij}$, and $\bar{Z} = \frac{1}{n_1 + n_2 + \ldots + n_k} \sum_i \sum_j Z_{ij}$.
- Therefore, the F-statistic for X_{ij} is the same as the F-statistic for Z_{ij}, which is a zero mean random variable.

For the notational simplicity we assume that $\mu = 0$. But using this argument you know how you can extend it for the case $\mu \neq 0$.

ANOVA and F-distribution

We use the following lemma that is easy to prove and you have seen it several times:

Lemma

If $W_1 \sim \chi^2_{k_1}$ and is independent of $W_2 \sim \chi^2_{k_2}$, then $W_1 + W_2 \sim \chi^2_{k_1+k_2}$.

Try to prove this result for yourself by using the characteristic function of the χ^2 distribution.

Lemma

Under the null hypothesis $\bar{X}_\ell - \bar{X}$ is independent of $X_{ij} - \bar{X}_i$ for $\ell \neq i$.

ANOVA and F-distribution

Since both of the random variables are Gaussian, proving independence is equivalent to proving that

$$\mathbb{E}((\bar{X}_\ell - \bar{X})(X_{ij} - \bar{X}_i)) - \mathbb{E}((\bar{X}_\ell - \bar{X}))\mathbb{E}(X_{ij} - \bar{X}_i) = 0. \tag{7}$$

Since $\mathbb{E}((\bar{X}_\ell - \bar{X}))\mathbb{E}(X_{ij} - \bar{X}_i) = 0$, we should prove that $\mathbb{E}((\bar{X}_\ell - \bar{X})(X_{ij} - \bar{X}_i)) = 0$. We have

$$\begin{aligned}
\mathbb{E}((\bar{X}_\ell - \bar{X})(X_{ij} - \bar{X}_i)) &= \mathbb{E}(\bar{X}_\ell X_{ij}) - \mathbb{E}(\bar{X} X_{ij}) - \mathbb{E}(\bar{X}_\ell \bar{X}_i) + \mathbb{E}(\bar{X}\bar{X}_i) \\
&\overset{(a)}{=} -\mathbb{E}(\bar{X} X_{ij}) + \mathbb{E}(\bar{X}\bar{X}_i) = -\mathbb{E}(\frac{1}{N}\sum_p \sum_q X_{pq} X_{ij}) + \mathbb{E}\frac{1}{N}\sum_{p=1}^{k} n_p \bar{X}_p \bar{X}_i \\
&\overset{(b)}{=} \frac{-\sigma^2}{N} + \frac{n_i}{N}\mathbb{E}(\bar{X}_i^2) = 0.
\end{aligned} \tag{8}$$

To obtain Equality (a) we use the fact that $\bar{X}_\ell$ only depends on $X_{\ell j}$ for $j = 1, \ldots, n_\ell$ and hence is independent of X_{ij} and $\bar{X}_i$.

ANOVA and F-distribution

Lemma

Under the null hypothesis $\bar{X}_i - \bar{X}$ is independent of $X_{ij} - \bar{X}_i$.

Proof.

Again, clearly these two random variables are jointly Gaussian and hence we should prove that

$$\mathbb{E}((\bar{X}_i - \bar{X})(X_{ij} - \bar{X}_i)) = 0.$$

We have

$$\mathbb{E}((\bar{X}_i - \bar{X})(X_{ij} - \bar{X}_i)) = \mathbb{E}(\bar{X}_i X_{ij}) - \mathbb{E}(\bar{X}_i)^2 + \mathbb{E}(\bar{X} X_{ij}) - \mathbb{E}(\bar{X}\bar{X}_i) \overset{(c)}{=} \frac{\sigma^2}{n_i} - \frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - \frac{\sigma^2}{N} = 0.$$

We leave the validation of Equaility (c) as an exercise to you. □

ANOVA and F-distribution

Recall,

$$F^* = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(N-k)}.$$

Hence, combining the above two results we can easily prove that the following lemma holds.

Lemma

The numerator and denominator of F^* are independent.

We leave the proof of this lemma as an exercise as well.

As the last step we have to prove that both the numerator and the denominator have χ^2 distributions. Let's start with the denominator...

ANOVA and F-distribution

Lemma

Let $X_{ij} \sim N(\mu_i, \sigma^2)$*. Then,* $\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/\sigma^2 \sim \chi^2_{N-k}$*, where* $N = n_1 + n_2 + \ldots + n_k$*.*

Proof.

First consider $W_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/\sigma^2$. You have seen that $W_i \sim \chi^2_{n_i - 1}$. Next, note that

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2/\sigma^2 = \sum_{i=1}^{k} W_i.$$

Therefore, according to Lemma for the sum of independent chi-square r.v., $\sum_{i=1}^{k} W_i \sim \chi^2_{n_1-1+n_2-1+\ldots+n_k-1} = \chi^2_{N-k}$. □

The last step is to show that the numerator is also χ^2. Since the general proof is slightly more involved, we start with the simpler setting in which $n_1 = n_2 = \ldots = n_k$, and then as an optional reading we provide the proof for the general settings.

ANOVA and F-distribution

Lemma

Let $n \triangleq n_1 = n_2 = \ldots = n_k$. *Then* $\sum_{i=1}^k n_i(\bar{X}_i - \bar{X})^2/(\sigma^2) \sim \chi^2_{k-1}$.

Proof.

Note that $\bar{X} = \frac{1}{k}\sum_{j=1}^k \bar{X}_j$. We also have $\bar{X}_i \overset{iid}{\sim} N(\mu, \sigma^2/n)$. Therefore,

$$\sum_{i=1}^k n(\bar{X}_i - \bar{X})^2/(\sigma^2) = \sum_{i=1}^k (\bar{X}_i - \bar{X})^2/(\sigma^2/n) \sim \chi^2_{k-1}.$$

□

ANOVA and F-distribution

Combining the last three Lemmas we obtain the following theorem:

Theorem

Let $n \triangleq n_1 = n_2 = \ldots = n_k$ *and define* $N = n_1 + n_2 + \ldots + n_k$*. Then*

$$F^* \sim F_{k-1,N-k}.$$

ANOVA and F-distribution (optional)

不考

Here we would like to extend the result of Theorem 1 to the more general setting where different groups have different samples sizes. As is clear most of the results are proved in the general settings except the last Lemma in which we assumed that $n_1 = n_2 = \ldots = n_k$.

Therefore, we would like to extend the result of that lemma only.

Lemma

Under the null hypothesis we have

$$Z = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2/\sigma^2 \sim \chi^2_{k-1}.$$

ANOVA and F-distribution (optional)

Without loss of generality assume that $\mu_i = 0$. Define $Y_i = \frac{\sqrt{n_i} X_i}{\sigma}$. Then $Y_1, \ldots, Y_k \overset{iid}{\sim} N(0,1)$. We then have

不考

Y=PHI(X)

Y'AY

[V,D]=eig(A)

Y*V'~N(0,I)

$$\frac{\sqrt{n_i}\bar{X}_i - \sqrt{n_i}\bar{X}}{\sigma} = Y_i - \sum_{j=1}^{k} \frac{\sqrt{n_i n_j}}{N} Y_j.$$

yi=Xi_bar*sqrt(ni)/sigma

Therefore,

$$\begin{bmatrix} \frac{\sqrt{n_1}\bar{X}_1 - \sqrt{n_1}\bar{X}}{\sigma} \\ \vdots \\ \frac{\sqrt{n_k}\bar{X}_k - \sqrt{n_k}\bar{X}}{\sigma} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 - \frac{\sqrt{n_1 n_1}}{N} & \frac{\sqrt{-n_1 n_2}}{N} & \ldots & \frac{-\sqrt{n_1 n_k}}{N} \\ -\frac{\sqrt{n_1 n_2}}{N} & 1 - \frac{\sqrt{n_2 n_2}}{N} & \ldots & -\frac{\sqrt{n_2 n_k}}{N} \\ \vdots & \vdots & \vdots & \vdots \\ -\frac{\sqrt{n_1 n_k}}{N} & \frac{\sqrt{n_2 n_k}}{N} & \ldots & 1 - \frac{\sqrt{n_k^2}}{N} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix}}_{Y}.$$

Z

Note that the matrix A is symmetric. Also note that $Z = Y^T A^T A Y$.

Cholesky

ANOVA and F-distribution (optional)

不考

sum of squared iid N(0,1)

不考

We next show that the A matrix has one zero eigenvalue and the rest of its eigenvalues are equal to 1. Let's assume that is the case and see how we can prove that Y^TA^TAY has χ^2_{k-1} distribution. It is straightforward to prove that $A = Q\Lambda Q^T$, where the columns of Q are the eigenvectors of A and Λ is a diagonal matrix whose diagonal elements are the eigenvalues of A. Q is a unitary matrix and satisfies $Q^TQ = I$. If we define $V = Q^TY$, it is straightforward to prove that $V_1, V_2, \ldots, V_k \overset{iid}{\sim} N(0,1)$. We have

eig(A)$val=eitheror(1,0)

$$Z = V^T\Lambda V = \sum_{i=1}^{k-1} V_i^2.$$

The last inequality is due to the fact that, I assumed all the eigenvalues are equal to 1 except the last eigenvalue. Clearly, $Z \sim \chi^2_{k-1}$.

ANOVA and F-distribution (optional)

We now return to the proof of the fact that 不考 A has only one zero eigenvalue and the rest of eigenvalues are equal to 1. To find the eigenvectors that correspond to eigenvalue 1, note we should solve

$$A\alpha = \alpha.$$

This is equivalent to saying that

$$\alpha_i - \sum_{j=1}^{k} \frac{\sqrt{n_i n_j}}{N}\alpha_j = \alpha_i \Rightarrow \frac{\sqrt{n_i}}{N}\sum_{j=1}^{k}\sqrt{n_j}\alpha_j = 0 \Rightarrow \sum_{j=1}^{k}\sqrt{n_j}\alpha_j = 0$$

The last equation specifies a $k-1$ dimensional hyperplane. Therefore, A has $k-1$ eigenvalues of 1. The last eigenvalue is zero since $A\alpha = 0$ has a nontrivial solution: $\alpha_i = \sqrt{n_i}$. This completes our proof.

Kruskal-Wallis Statistic

https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php

Online Test Link http://vassarstats.net/kw3.html

In many applications the Gaussianity assumption does not hold. Therefore the F-test is not very useful for such cases.

The empirical research that has been done in the literature has shown that under the null hypothesis the probability of type I error is robust to deviations of the null distribution from Gaussian.

However, once you violate the Gaussianity, the power of the one-way ANOVA F test is not high enough.

Therefore, we would like to design tests that are free of distributional assumptions, i.e., they are nonparametric.

```
R
# Kruskal Wallis Test One Way Anova by Ranks
kruskal.test(y~A) # where y1 is numeric and A is a factor
```

Kruskal-Wallis Statistic

As the first step, we model the problem in the following way:

$$\begin{array}{rcl} X_{11}, X_{12}, \cdots, X_{1n_1} & \sim & G(x-\mu_1), \\ X_{21}, X_{22}, \cdots, X_{2n_2} & \sim & G(x-\mu_2), \\ & \vdots & \\ X_{k1}, X_{k2}, \cdots, X_{kn_k} & \sim & G(x-\mu_k), \end{array} \tag{9}$$

where F is a continuous CDF whose form is not know to us.

Kruskal-Wallis Statistic

Based on this model we would like to test the hypothesis

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs. H_1 : *at least one of the equalities is violated*

We would like to design a test statistic for this problem that has the following two properties:

1. It can distinguish between H_0 and H_1.
2. The distribution of the statistic under the null hypothesis (that all the races have the same crime rate) is free of the distribution.[2]

[2] As you may remember from the two sample tests, the second condition is to ensure that we can calculate the p-value of the test and its significance level. We will later remove this constraint by using permutation tests.

Kruskal-Wallis Statistic

To construct such statistics we use the ranks again. Let R_{ij} denote the rank of X_{ij} in the entire dataset. We construct the following table based on the ranks of each sample:

Table : Rank table for different groups

Race	Statistics
$group_1$	$R_{11}, R_{12}, \cdots, R_{1n_1}$
$group_2$	$R_{21}, R_{22}, \cdots, R_{2n_2}$
$\vdots$	$\vdots$
$group_k$	$R_{k1}, R_{k2}, \cdots, R_{kn_k}$

Based on these ranks we calculate the group averages of the ranks as

$$\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}.$$

Kruskal-Wallis Statistic

Suppose that the null hypothesis is true. Then,

$$\mathbb{E}_{H_0}(R_{ij}) = \frac{n_1 + n_2 + \ldots + n_k + 1}{2}.$$

Can you explain why? The subscript H_0 just emphasizes the fact that these expectations are taken under the assumption that the null hypothesis is true and all samples are drawn from the same distribution.

Let $N \triangleq n_1 + n_2 + \ldots + n_k$ denote the total number of samples. We conclude that

$$\mathbb{E}_{H_0}(\bar{R}_i) = \frac{N + 1}{2}.$$

Kruskal-Wallis Statistic

Now let's see what happens under the alternative.

Assume that $\mu_k > \mu_1 = \mu_2 = \ldots = \mu_{k-1}$.

Then what we expect is that $\mathbb{E}(\bar{R}_k) > \frac{N+1}{2}$, while

$$\mathbb{E}(\bar{R}_1) = \mathbb{E}(\bar{R}_2) = \ldots = \mathbb{E}(\bar{R}_{k-1}) < \frac{N+1}{2}.$$

Based on these intuitions we define the Kruskal-Wallis statistic as

$$KW \triangleq \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2$$

(Under H0)=E[1/ni*sum((j = 1 : ni), Rij)]=(N+1)/2

Kruskal-Wallis Statistic

The constant $\frac{12}{N(N+1)}$ might seem strange.

Ignoring this constant the rest of the statistic seems intuitive.

We will discuss why such constants are chosen for the statistic later.

Before then, note the following important property of KW:

1. Under H_0, KW statistic is free of the distribution F. Why?

But why have we chosen the constant $\frac{12}{N(N+1)}$?

The answer to this question is mainly due to the asymptotic analysis of KW. Under proper asymptotic settings, it can be proved that

$$KW \xrightarrow{d} \chi^2_{k-1}.$$

The constants are chosen to make this happen.(*)

(*) This is not a basic result, and proving it is out of the scope of our course. Those of you who

Permutation test - k-Sample Test

- As you may remember permutation test offered a lot of flexibility in the two sample setting.
- So, it would be nice to extend the idea of permutation test to the k-sample setting as well.
- Clearly, under the null hypothesis all the samples are drawn from the same distribution and hence are exchangeable. Therefore, as before we can argue that any permutation of the data is a sample from the same distribution (under H_0).

Therefore, we can design the permutation test as before...

Permutation test - k-Sample Test

1. Obtain the F-statistic (or Kruskal-Wallis or any other statistic that can distinguish between H_0 and H_1) for the original data and call it F_{obs}.
2. Obtain all possible permutations of the $N = n_1 + n_2 + \ldots + n_k$ observations.

$$\binom{N}{n_1 n_2 \cdots n_k} = \frac{N!}{n_1! n_2! \cdots n_k! (N - n_1 - n_2 - \ldots - n_k)!}.$$

   For each permuted sample calculate the corresponding statistic and call it F_i.

3. Calculate the p-value of the test according to $p_{perm} = \frac{1}{\binom{N}{n_1 \cdots n_k}} \sum_{i=1}^{\binom{N}{n_1 \cdots n_k}} \mathbb{I}(F_i \geq F_{obs})$.

Permutation test

Here we are using the ANOVA F-statistic.

But as we discussed before, we can replace it with any other statistic that distinguishes H_0 from H_1.

Note that the number of permutations in this case is in general much more than the number of permutations in the two-sample problem.

Therefore, the demand for Monte Carlo method is even more evident than the two-sample problem.

Since, it must be done in exactly the same way I skip this.

Comparison of KW and F-test

There has been some numerical comparative studies between F and KW tests. The overall suggestion would be:

- If distribution G is Gaussian, ANOVA outperforms KW.
- If G is not Gaussian (specially when it has heavy tails), then KW outperforms one-way ANOVA test.

Clearly, these are some rule of thumbs to help you choose the right statistic.

But they are not very accurate statements as there is no theoretical results on comparing these things.

How should I check the Gaussianity of my data set?

References

- J. Higgins, Introduction to modern nonparametric statistics.
- A. W. van der Vaart, "Asymptotic statistics".