10 Tests for Comparing k Populations (Cont'd)

MTH 3240 Environmental Statistics

Spring 2020

Notes

Objectives

Objectives:

- Obtain and interpret fitted values and residuals.
- Use plots to check the normality and common population standard deviation assumptions required by the ANOVA F test.
- Write out the group means and treatment effects versions of the ANOVA model, including any assumptions about the random error term ϵ. (**Optional for Spring 2020**)
- Carry out a Kruskal-Wallis test for differences among k population means. (**Optional for Spring 2020**)
- Decide which test (the ANOVA F test or Kruskal-Wallis test) is more appropriate for a given set of data.

Notes

Fitted Values and Residuals

- The **group means** $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k$ are sometimes called ***fitted values***.

Fitted Values:

$$\text{Fitted Value for } i\text{th Group} \;=\; \bar{Y}_i$$

- Statistical software reports $\mathbf{n}$ **duplicates** of the **fitted value** for **each group**, one duplicate for each of the n individuals in the group.

Notes

- A ***residual***, denoted e_{ij}, is the **deviation** of an individual's observed Y_{ij} value away from the **fitted value** for that individual.

Residuals:

$$e_{ij} \;=\; Y_{ij} - \bar{Y}_i$$

- Statistical software reports the values of all $\mathbf{N}$ **residuals**, one for each individual in the study.

Notes

Checking the ANOVA Assumptions

- The **ANOVA F test** requires that the k groups (samples) are from **normal** populations (or that their sample sizes are **large**) whose **standard deviations** are all **equal**.

Notes

- Two ways to check the **normality assumption**:
  - Make k *separate* **histograms** or **normal probability plots**, one for each of the k **groups**.
  - Make a *single* **histogram** or **normal probability plot** plot of the N **residuals** e_{ij}.

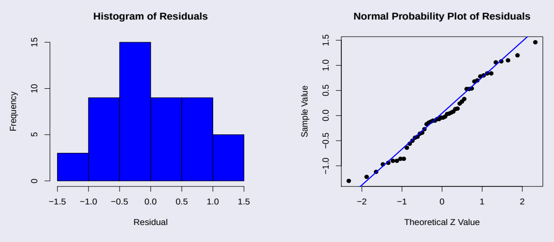
Notes

Example

For the lead measurements at five labs, the **ANOVA F test** showed statistically significant differences among the means for the five labs.

To justify this test result, we check the **normality assumption** using the plots of the **residuals** on the next slide.

Notes



The plots indicate that the **normality assumption** appears to be met.

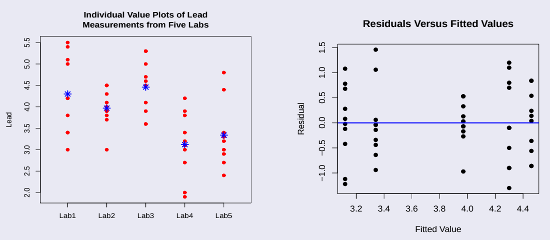
Notes

- A few ways to check the **equal population standard deviation** assumption:
  - An ***individual value plot*** of the k samples.
  - A plot of the ***residuals*** (y-axis) versus **fitted values** (group means, x-axis).
- In both plots, we look for roughly **equal amounts of within-group (vertical) spread** across the k groups.

Notes

Example

For the lead measurements at the five labs, we check the **equal standard deviation assumption** using either of these plots.



Notes

Because the amount of (vertical) spread in the points is roughly the same from one group to the next, the plots indicate that the **equal standard deviation** assumption appears to be met.

Notes

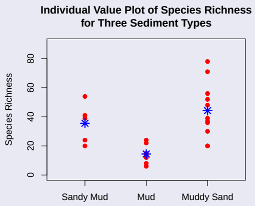
- The reason we why plot the residuals versus *fitted values* (group means) is that usually, when the **equal standard deviation assumption** is ***violated***, the groups with **bigger means** (**fitted values**) are usually the ones with **bigger standard deviations**.

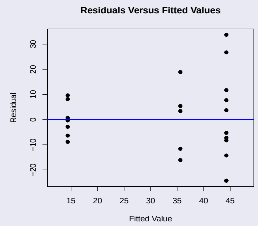
  So it's easier to detect violations of the equal standard deviation assumption by ordering the groups from left to right by their means (fitted values).

Notes

Example

The **common standard deviation assumption** is violated in the plots below, but it's easier to detect in the right plot because the groups with larger means (fitted values) have larger standard deviations.





Notes

Statistical Models (Optional for Spring 2020)

- A common approach to detecting patterns in ”noisy” data is to first think of variation in the data in terms of a ***statistical model*** that has two parts:
  1. A part representing systematic ***nonrandom variation*** in the data.
  2. Another part representing ***random variation***.

  Patterns are then detected by **estimating** or **testing hypotheses** about the **nonrandom** components in the model.

Notes

(Optional for Spring 2020)

- An example of a simple **statistical model** is the one used to describe a **measurement** $\boldsymbol{Y}$ with **measurement error** $\boldsymbol{\epsilon}$,

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu}$ is the **true (unknown) concentration** being measured,

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$$

is the difference between the measurement and the true concentration, and

$$\epsilon \sim \mathrm{N}(0, \sigma).$$

This model is **equivalent** to saying that

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}).$$

Notes

One-Factor ANOVA Model (Optional for Spring 2020)

- Recall that for **one-factor ANOVA**, we suppose the groups (samples) are from $\boldsymbol{k}$ **normal** populations whose means are $\boldsymbol{\mu_1, \mu_2, \ldots, \mu_k}$ and whose **standard deviations** are all the **same**, $\boldsymbol{\sigma}$.

Group 1: $Y_{11}, Y_{12}, \ldots, Y_{1n}$ are a sample from a $\mathrm{N}(\mu_1, \sigma)$ distribution.

Group 2: $Y_{21}, Y_{22}, \ldots, Y_{2n}$ are a sample from a $\mathrm{N}(\mu_2, \sigma)$ distribution.

$\vdots$

Group k: $Y_{k1}, Y_{k2}, \ldots, Y_{kn}$ are a sample from a $\mathrm{N}(\mu_k, \sigma)$ distribution.

Notes

(Optional for Spring 2020)

- This is written more succinctly as the so-called ***group means version*** of the ***one-factor ANOVA model***.

> **One-Factor ANOVA Model (Group Means Version)**: A statistical model for describing samples from k normal populations is
> $$Y_{ij} = \mu_i + \epsilon_{ij}\,, \tag{1}$$

Notes

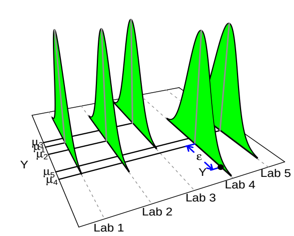
(Optional for Spring 2020)

> where
>
> Y_{ij} is the jth observation ($j = 1, 2, \ldots, n$) in the ith group ($i = 1, 2, \ldots, k$).
>
> μ_i is the mean of the ith population, called the ***ith group population mean***.
>
> ϵ_{ij} is a random error term following a $\mathbf{N(0,\ \sigma)}$ distribution.

Notes

(Optional for Spring 2020)

One−Factor Analysis of Variance Model



Notes

(Optional for Spring 2020)

- In practice, the ***model parameters*** $\mu_1, \mu_2, \ldots, \mu_k$, and σ will be **unknown**, but they can be **estimated** from the data.

Notes

(Optional for Spring 2020)

- Sometimes the model is written in terms of the **effects** of the treatments in an experiment (rather than in terms of the group population means $\mu_1, \mu_2, \ldots, \mu_k$).

Notes

(Optional for Spring 2020)

One-Factor ANOVA Model (Treatment Effects Version): Another version of the statistical model for describing samples from k normal populations is

$$Y_{ij} \;=\; \mu + \alpha_i + \;\epsilon_{ij}\,,$$

Notes

(Optional for Spring 2020)

where Y_{ij} and ϵ_{ij} are as described before, and

μ is an ***overall population mean*** (for all k populations combined).

α_i is the ***effect*** of the ith treatment.

(More formal definitions of μ and the α_i's are on the next slide.)

Notes

(Optional for Spring 2020)

- More formally, $\boldsymbol{\mu}$ and the $\boldsymbol{\alpha_i}$'s are defined as follows:

$\boldsymbol{\mu}$ = The average of the groups' population means $\mu_1, \mu_2, \ldots, \mu_k$, that is,

$$\mu \;=\; \frac{1}{k}\sum_{i=1}^{k}\mu_i\,.$$

$\boldsymbol{\alpha_i}$ = The discrepancy between the ith group's population mean μ_i and the overall mean μ, that is,

$$\alpha_i \;=\; \mu_i - \mu\,.$$

Notes

(Optional for Spring 2020)

- With these definitions, we can write the ith group's mean, μ_i, as the *overall mean plus a treatment effect*:

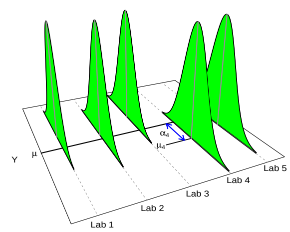
$$\begin{aligned} \mu_i &= \mu + (\mu_i - \mu) \\ &= \mu + \alpha_i \,. \end{aligned}$$

This says the **two versions** of the **one-factor ANOVA model** are **equivalent**.

Notes

(Optional for Spring 2020)

One−Factor Analysis of Variance Model



Notes

(Optional for Spring 2020)

- In terms of the two **ANOVA models**, the hypotheses are:

	Hypothesis About the μ_i's	**Equivalent Hypothesis About the α_i's**
Null	$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$	$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$
Alternative	H_a : The μ_i's aren't all equal	H_a : The α_i's aren't all 0

In either case, the **null hypothesis** says there are **no differences** among the k population means (or among the mean responses to the k treatments).

The **alternative hypothesis** says there's a **difference** among **at least one pair** of the means.

Notes

Parameter Estimates (Optional for Spring 2020)

- The **estimators** of the (unknown) **model parameters**, based on the data, are listed below along with an alternative notation for each estimator.

Model Parameter Estimators

Model Parameter	Estimator	Alternate Notation for the Estimator
μ_i	$\bar{Y}_i$	$\hat{\mu}_i$
μ	$\bar{Y}$	$\hat{\mu}$
$\alpha_i = \mu_i - \mu$	$\bar{Y}_i - \bar{Y}$	$\hat{\alpha}_i$
σ	$\sqrt{\text{MSE}}$	$\hat{\sigma}$

Notes

(Optional for Spring 2020)

- By adding and subtracting $\bar{Y}_i$ to the right side of

$$Y_{ij} = Y_{ij},$$

we get

$$Y_{ij} = \bar{Y}_i + (Y_{ij} - \bar{Y}_i).$$

Using the alternative notation $\hat{\mu}_i$ and the definition of a residual e_{ij}, this says we can write an observation Y_{ij} as

$$Y_{ij} = \hat{\mu}_i + e_{ij},$$

which resembles the **group means** version of the **one-factor ANOVA model** (1).

Notes

(Optional for Spring 2020)

- It's clear that:

1. The **fitted value** $\hat{\boldsymbol{\mu}}_i$ approximates the model's **nonrandom part**.
2. The **residual** $\boldsymbol{e_{ij}}$ approximates the model's **random error** term $\boldsymbol{\epsilon_{ij}}$.

The **square root** of the **mean squared error**, $\sqrt{\textbf{MSE}}$, estimates the **standard deviation** $\boldsymbol{\sigma}$ of the N$(0, \sigma)$ error distribution.

Notes

Dealing With Non-Normal Data

- The ANOVA F test requires that the samples were drawn from k **normal** populations (or that $\boldsymbol{n}$ is **large** for all k samples).

If this **normality** assumption isn't met (and n isn't large for all k samples), there are two possible remedies:

1. **Transform** the data to normality before carrying out the ANOVA.
2. Carry out a ***nonparametric*** test (which doesn't require normality).

We'll look at these two approaches one at a time.

Notes

Transforming Data To Normality

- The first approach to approach to testing hypotheses with **non-normal** samples is to ***transform*** the data (**all** $\boldsymbol{k}$ samples) to normality first.

Notes

Carrying Out a Nonparametric Test (Optional for Spring 2020)

- The second approach to testing hypotheses with **non-normal** samples is to use a ***nonparametric*** test procedure, i.e. one that **doesn't** rely on a normality assumption.

  The ***Kruskal-Wallis test*** described next is a **nonparametric** alternative to the *ANOVA F test*.

Notes

Kruskal-Wallis Test (Optional for Spring 2020)

- The ***Kruskal-Wallis test*** is a **nonparametric** test for differences among k population **means** $\boldsymbol{\mu_1, \mu_2, \ldots, \mu_k}$.

Notes

(Optional for Spring 2020)

- We'll want to test the hypotheses:

Null and Alternative Hypothesis:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$
H_a: The μ_i's aren't all equal.

(Same hypotheses as for the ANOVA F test.)

Notes

(Optional for Spring 2020)

- The ***test statistic***, denoted $\boldsymbol{K_w}$, is obtained after combining the k groups and **ranking** the observations (smallest to largest).
- As before, we'll let

$\boldsymbol{Y_{ij}}$ = The jth observation in the ith group
$\boldsymbol{N}$ = The total number of observations in the overall combined sample.

Notes

(Optional for Spring 2020)

- (cont'd)

$\boldsymbol{R_{ij}}$ = The **rank** of Y_{ij} in the overall combined sample.
$\boldsymbol{\bar{R}_i}$ = The **mean rank** of the observations from the ith group.
$\boldsymbol{\bar{R}}$ = The **overall mean rank** of the N observations in the overall combined sample. Thus

$$\bar{R} = \frac{1}{N}(1 + 2 + \cdots + N).$$

It can be shown that

$$\bar{R} = \frac{N+1}{2}.$$

Notes

(Optional for Spring 2020)

Kruskal-Wallis Test Statistic:

1. **Combine** the k groups, keeping track of which group each observation originally belonged to, **sort** the observations, and ***rank*** them from smallest to largest. If two or more are **tied**, assign to each of them the **average** of the **ranks** they would've been assigned if they hadn't been tied.
2. Compute the **group mean ranks** $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ and the **overall mean rank** $\bar{R}$.

Notes

(Optional for Spring 2020)

3. The **test statistic** is

$$K_w = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2.$$

Notes

(Optional for Spring 2020)

Example

The table below shows **aluminum** (Al) concentrations (μg/g wet weight) measured in carp in the **Colorado**, **Columbia**, and **Mississippi** River basins.

Notes

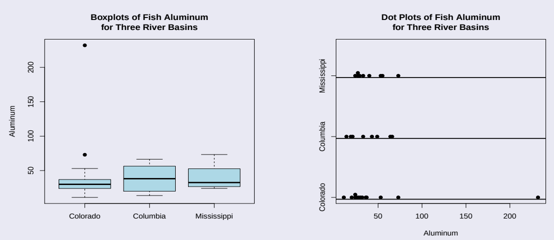
(Optional for Spring 2020)

Al in Fish

Colorado River Basin	Columbia River Basin	Mississippi River Basin
32	18.9	54.9
232	64.1	24.1
36	33.2	40.2
73	48.8	52.7
53	13.7	28.9
20	21.1	73.3
28	43.2	26.7
24	66.5	26.6
24		32.6
11		
37		
30		
26		

Notes

(Optional for Spring 2020)



Notes

(Optional for Spring 2020)

We want to know if there are any differences among the Al concentrations for the three river basins.

The plots show a (slight) indication that the samples are from right-skewed populations, so we'll use a **Kruskal-Wallis test**.

The hypotheses are

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3$$
$$H_a\text{: The } \mu_i\text{'s aren't all equal.}$$

where $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, and $\boldsymbol{\mu_3}$ are the (unknown) population **mean Al concentrations** in carp for the **three river basins**.

Notes

(Optional for Spring 2020)

Notice from the graphs that there's a lot of **overlap** among the three groups (suggesting no differences in Al concentrations).

The sample sizes are $n_1 = 13$, $n_2 = 8$, and $n_3 = 9$, so

$$N = 13 + 8 + 9 = 30.$$

The overall combined sample, sorted and ranked, is shown below (Sample $1 =$ **Colorado**, $2 =$ **Columbia**, and $3 =$ **Mississippi**).

Notes

(Optional for Spring 2020)

Sample	1	2	2	1	2	1	1	3	1
Observation	**11.0**	13.7	18.9	**20.0**	21.1	**24.0**	**24.0**	24.1	**26.0**
Rank	**1**	2	3	**4**	5	**6.5**	**6.5**	8	**9**

3	3	1	3	1	1	3	2	1	1	3
26.6	26.7	**28.0**	28.9	**30.0**	**32.0**	32.6	33.2	**36.0**	**37.0**	40.2
10	11	**12**	13	**14**	**15**	16	17	**18**	**19**	20

2	2	3	1	3	2	2	1	3	1
43.2	48.8	52.7	**53.0**	54.9	64.1	66.5	**73.0**	73.3	**232.0**
21	22	23	**24**	25	26	27	**28**	29	**30**

Different font shades indicate river basins. Notice that the three groups are **evenly "intermingled"**.

Notes

(Optional for Spring 2020)

The three **group mean ranks** are:

$$\begin{aligned}\bar{R}_1 &= \frac{1}{13}(1+4+6.5+6.5+9+12+14+15+18+19+24\\ &\quad +28+30)\\ &= \mathbf{14.4}.\end{aligned}$$

$$\begin{aligned}\bar{R}_2 &= \frac{1}{8}(2+3+5+17+21+22+26+27)\\ &= \mathbf{15.4}.\end{aligned}$$

$$\begin{aligned}\bar{R}_3 &= \frac{1}{9}(8+10+11+13+16+20+23+25+29)\\ &= \mathbf{17.2}.\end{aligned}$$

Notes

(Optional for Spring 2020)

The **overall mean rank** is

$$\bar{R} = \frac{N+1}{2} = \frac{30+1}{2} = \mathbf{15.5}.$$

The **Kruskal-Wallis test statistic** K_w is

$$\begin{aligned}K_w &= \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i(\bar{R}_i-\bar{R})^2\\ &= \frac{12}{30(30+1)}\left[13(14.4-15.5)^2+8(15.4-15.5)^2\right.\\ &\qquad \left.+\,9(17.2-15.5)^2\right]\\ &= \mathbf{0.54}.\end{aligned}$$

ranks are so similar.)

Notes

(Optional for Spring 2020)

- **If H_0 was true**, $\mu_1, \mu_2, \ldots, \mu_k$ would all be equal, ...

  and the k groups would be **evenly "intermingled"** when combined and sorted.

  In this case, the **group mean ranks** $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ would all be roughly equal, and therefore roughly equal to $\bar{R}$, so $\boldsymbol{K_w}$ would be **close to zero**.

- But **if H_a was true**, the k groups would be **"segregated"** when combined and sorted.

  In this case $\bar{R}_1, \bar{R}_2, \ldots, \bar{R}_k$ would differ substantially from each other, and therefore also from $\bar{R}$, and $\boldsymbol{K_w}$ **would be large**.

Notes

(Optional for Spring 2020)

Large values of K_w provide evidence in favor of H_a: The μ_i's aren't all equal.

Notes

(Optional for Spring 2020)

- Now suppose the samples are from **any** k (continuous) populations that have (roughly) the same shape and whose means are $\mu_1, \mu_2, \ldots, \mu_k$.

  In this case, the **null distribution** is as follows.

Notes

(Optional for Spring 2020)

Sampling Distribution of K_w Under H_0: If K_w is the Kruskal-Wallis test statistic, then when

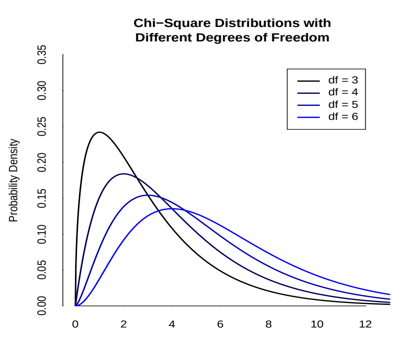
$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

is true, K_w follows a distribution called the ***chi-square distribution*** with $k-1$ ***degrees of freedom***, denoted $\chi^2(k-1)$. We write this as

$$K_w \sim \chi^2(k-1).$$

Notes

(Optional for Spring 2020)



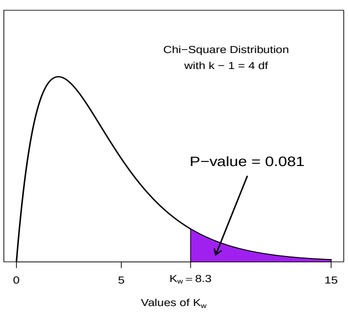
Notes

(Optional for Spring 2020)

- **P-values** and **rejection regions** are obtained from the **right** tail of the $\chi^2(k-1)$ **(chi-square) distribution**.
- The next slide shows the **p-value** when the observed K_w value is $K_w = 8.3$.

Notes

(Optional for Spring 2020)



Notes

(Optional for Spring 2020)

Kruskal-Wallis Test for $\mu_1, \mu_2, \ldots, \mu_k$

Assumptions: The data are independent random samples from k continuous populations that differ, if at all, by their means $\mu_1, \mu_2, \ldots, \mu_k$ but not their shapes, and the sample sizes $n_1, n_2, \ldots, n_k$ are all large.*

Null hypothesis: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$.

Test statistic value: $K_w = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i(\bar{R}_i - \bar{R})^2$.

Decision rule: Reject H_0 if p-value $< \alpha$ or K_w is in rejection region.

Notes

(Optional for Spring 2020)

Alternative hypothesis	P-value = area under χ^2 distribution with $k-1$ d.f.:	Rejection region = K_w values such that:**
$H_a : \mu_i \neq \mu_j$ for some i and j	to the right of K_w	$K_w \geq \chi^2_{\alpha,k-1}$

* The sample sizes are considered to be large when they're all 5 or larger if $k > 3$, and all 6 or larger if $k = 3$. For smaller sample sizes, the test statistic K_w can be compared to a table of tail areas or critical values of the exact sampling distribution of K_w, found, for example, in [**?**] or [**?**].

** $\chi^2_{\alpha,k-1}$ is the $100(1-\alpha)$th percentile of the χ^2 distribution with $k-1$ d.f.

Notes

(Optional for Spring 2020)

Example

Continuing from the previous example, we got $K_w = 0.54$, and from a table of tail areas of the chi-square distribution, using $k - 1 = 2$ df, the **p-value** is **greater than 0.100**.

Therefore, there's **no statistically significant evidence** for differences among the **mean Al concentrations** in carp for the **three river basins**.

Notes

Notes

Notes

Notes