

Statistical Methods

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 10, 2019

Nels Grevstad

Topics

- 1 One-Factor ANOVA for Population Means $\mu_1, \mu_2, \dots, \mu_I$ (Cont'd)

Nels Grevstad

Objectives

Objectives:

- State the group means version of the one-factor ANOVA model and the treatment effects version.
- Interpret residuals and fitted values.
- Use residuals to check normality and constant standard deviation assumptions.

Nels Grevstad

The One-Factor ANOVA Model (Two Versions)

- A **statistical model** is a mathematical representation of data, with components corresponding to the **two sources of variation**:
 - Deterministic variation in the data
 - Random variation in the data

Nels Grevstad

Notes

Notes

Notes

Notes

Example

Suppose $X \sim N(\mu, \sigma)$.

Let

$$\epsilon = X - \mu.$$

Then ϵ is a linear function of X , so ϵ is normally distributed (Slides 1). More precisely,

$$\epsilon \sim N(0, \sigma).$$

We can write X in the form of a **statistical model** as

$$X = \mu + \epsilon,$$

where

μ is the **true mean** of X (i.e. **expected value**).
 ϵ is a $N(0, \sigma)$ **random error** term.

- Recall that the *assumptions* for the **ANOVA F test** are that the I samples are drawn independently from $N(\mu_1, \sigma)$, $N(\mu_2, \sigma) \dots, N(\mu_I, \sigma)$ populations.
- We can write the assumptions in the form of a **statistical model** (next slide).

One-factor ANOVA Model (Group Means Version):

$$X_{ij} = \mu_i + \epsilon_{ij}, \tag{1}$$

where

μ_i is the **true mean** response for the i th treatment group
 ϵ_{ij} are iid $N(0, \sigma)$ **random errors**

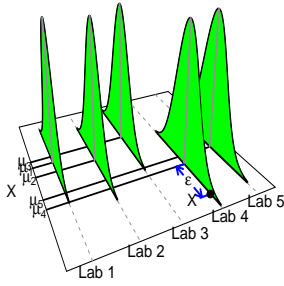
Notes

Notes

Notes

Notes

One-Factor Analysis of Variance Model



- Sometimes a different (but equivalent) **statistical model** is used to describe data in a **one-factor ANOVA** context (next slide).

One-factor ANOVA Model (Treatment Effects Version):

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad (2)$$

where

μ is a constant called the **true grand mean**

α_i is the **treatment effect** for i th treatment

ϵ_{ij} are iid $N(0, \sigma)$ **random errors**

Proposition

If we define

$$\mu = \frac{\sum_{i=1}^I \mu_i}{I} \quad \text{and} \quad \alpha_i = \mu_i - \mu,$$

then

1. The two models (1) and (2) are equivalent.
2. $\sum \alpha_i = 0$.

The first result holds because we can write μ_i in (1) as

$$\mu_i = \mu + (\mu_i - \mu) = \mu + \alpha_i.$$

The second holds because deviations away from a mean always sum to zero.

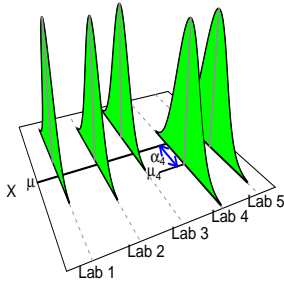
Notes

Notes

Notes

Notes

One-Factor Analysis of Variance Model



Nels Grevstad

- In terms of the **treatment effects version** of the **ANOVA model**, the hypotheses are:

$$\begin{aligned}
 H_0: & \alpha_1 = \alpha_2 = \dots = \alpha_I = 0 \\
 H_a: & \text{Not all } \alpha_i \text{'s equal zero} \quad (3)
 \end{aligned}$$

which are equivalent to the hypotheses about the μ_i 's stated previously.

Nels Grevstad

Estimating Model Parameters

- Recall that the **group means version** of the ANOVA model is

$$X_{ij} = \mu_i + \epsilon_{ij},$$

and the **treatment effects version** is

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

Nels Grevstad

Model Parameter Estimators: The unknown model parameters $\mu_i, \mu, \alpha_i,$ and σ are estimated by $\hat{\mu}_i, \hat{\mu}, \hat{\alpha}_i,$ and $\hat{\sigma}$ defined as:

Model Parameter	Estimator
μ_i	$\hat{\mu}_i = \bar{X}_i.$
μ	$\hat{\mu} = \bar{X}..$
$\alpha_i = \mu_i - \mu$	$\hat{\alpha}_i = \bar{X}_i. - \bar{X}..$
σ	$\hat{\sigma} = \sqrt{MSE}$

Nels Grevstad

Notes

Notes

Notes

Notes

Predicted Values and Residuals

- The **fitted value** (or **predicted value**) for the j th individual in the i th treatment group, \hat{X}_{ij} , is defined as:

$$\begin{aligned}\hat{X}_{ij} &= \hat{\mu} + \hat{\alpha}_i \\ &= \bar{X}_{..} + (\bar{X}_{i.} - \bar{X}_{..}) \\ &= \bar{X}_{i.}\end{aligned}$$

This is the value we'd predict, based on the data, for the response of the j th individual in the i th treatment group.

Nels Grevstad

- The **residual** for the j th observation in the i th group, e_{ij} , is defined as

$$\begin{aligned}e_{ij} &= X_{ij} - \hat{X}_{ij} \\ &= X_{ij} - \bar{X}_{i.}\end{aligned}$$

This is the deviation of the observed response X_{ij} away from the value model-predicted value.

Nels Grevstad

Proposition

The residuals sum to zero within each treatment group, i.e.

$$\sum_j e_{ij} = 0 \quad \text{for each } i = 1, 2, \dots, I.$$

Therefore they sum to zero across all groups:

$$\sum_i \sum_j e_{ij} = 0.$$

This is because the **residuals** are just **deviations** away from the **group means**.

Nels Grevstad

- Comment:** We can write

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

which resembles the **one-factor ANOVA model**.

In particular, the **residual** e_{ij} corresponds to the **random error** term ϵ_{ij} in the model.

Nels Grevstad

Notes

Notes

Notes

Notes

Checking the Model Assumptions

- For the **ANOVA F test**, we assume the ϵ_{ij} 's are iid $N(0, \sigma)$.

Note that σ is assumed to be **constant** from one group to the next.

- **Checking the Normality Assumption:** Use a **histogram** or **normal probability plot** of the **residuals**.
- **Checking the Constant σ Assumption:** Plot the **residuals** versus the **fitted values**.

Usually, when σ *isn't* constant, it increases with the group mean (second plot, next slide).

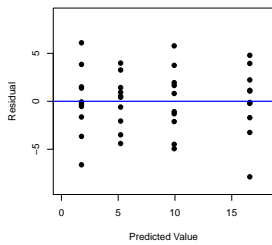
Poisson data are an example.

Nels Grevstad

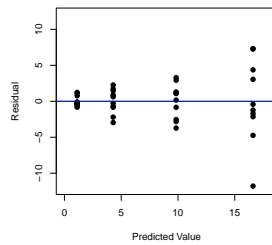
One-Factor ANOVA for Population Means $\mu_1, \mu_2, \dots, \mu_J$ (Cor

Notes

Residuals vs Predicted Values



Residuals vs Predicted Values



Nels Grevstad

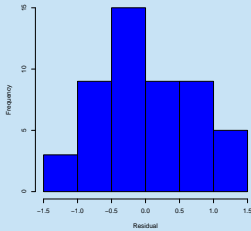
One-Factor ANOVA for Population Means $\mu_1, \mu_2, \dots, \mu_J$ (Cor

Notes

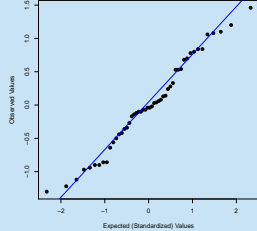
Example

For the data on lead measurements from five labs, a **histogram** and **normal probability plot** of the **residuals** are below.

Histogram of Residuals



Normal Probability Plot of Residuals



Nels Grevstad

One-Factor ANOVA for Population Means $\mu_1, \mu_2, \dots, \mu_J$ (Cor

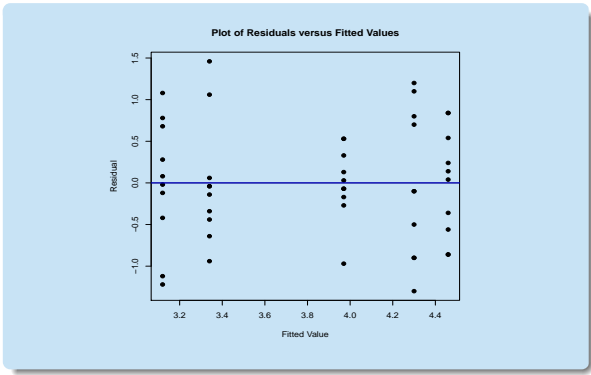
Notes

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

A scatterplot of the **residuals** versus the **fitted values** is on the next slide.

Nels Grevstad

Notes



Nels Grevstad

The plot supports the **constant σ assumption**.
 Therefore, the results of the **ANOVA F test** are valid.

Nels Grevstad

Notes

Notes

Notes

Notes
