Statistik för bioteknik sf2911
Föreläsning 15: Variansanalys

TK

14.12.2017

Learning Outcomes

- The problem of multiple comparisons
- One-way Analysis of Variance ($=$ ANOVA)
  - ANOVA table
  - F-distribution

Nationalencyklopedin

variansanalys, statistisk metod att analysera, i betydelsen dela upp, den totala variationen i ett datamaterial i ett antal komponenter som svarar mot intressanta variationsorsaker.

Analysis of Variance

Analysis of variance (ANOVA) *is a method of testing the equality of three or more population means by analyzing sample variances.*

Typical applications:

- We treat one group (1) with two aspirin tablets each day, and second group (2) with one aspirin tablet a day, while a third group (3) is given a placebo each day. We want to determine if there is sufficient evidence to support that the three groups have different mean blood pressure levels.

KTH Matematik

The statistical problem is to test whether all population means of blood pressure are equal under these three treatments.

$$H_o: \ \mu_1 = \mu_2 = \mu_3$$

We could use two-sample t-test for difference of means several times.

- H_{oA}: $\mu_1 = \mu_2$ tested to see whether there is a difference between taking two aspirin tablets and taking one asipirin tablet.
- H_{oB}: $\mu_1 = \mu_3$ tested to see whether there is a difference between taking one asipirin tablet and a placebo.
- H_{oC}: $\mu_2 = \mu_3$ tested to see whether there is a difference between taking two asipirin tablets and a placebo.

There is a problem in this: suppose we find

$$H_{oA} \text{ is rejected in favour of } \mu_1 < \mu_2$$

and

$$H_{oB}\ \mu_1 = \mu_3 \text{ is not rejected}$$

Then it "should" follow in H_{oC} that $\mu_2 > \mu_3$. But the comparison C has not been tested independently and is essentially determined when the two first tests have been decided. This becomes only worse when we compare a higher number of multiple hypotheses.

Problem of multiple comparisons

Statistical methods for dealing with multiple comparisons usually have two steps:

- An **overall test** to see if there is good evidence of **any** differences among the parameters that we want to compare.
- A detailed **follow-up analysis** to decide which of the parameters differ and to estimate how large the differences are.

Analysis of Variance

Analysis of variance (ANOVA) *is a method of testing the equality of three or more population means by analyzing sample variances.*

Typical applications:

- The analysis of microarray gene expression data typically tries to identify differential gene expression patterns in terms of differences of the population means between groups of arrays (e.g. treatments or biological conditions).

Analysis of Variance and e.g., Gene Expression Microarray Data

One question is how to make valid estimates of the relative expression for genes. Recognizing that there is inherent "noise" in microarray data, how does one estimate the error variation associated with an estimated change in expression, i.e., how does one construct the error bars? ANOVA methods can be used to normalize microarray data and provide estimates of changes in gene expression that are corrected for potential confounding effects.

Kerr, M. K., Martin, M. and Churchill, G.A. : Analysis of variance for gene expression microarray data, Journal of computational biology, *pp. 819-837, 2000.*

Analysis of Variance

We deal with **one-way analysis of variance** *or* **single factor analysis of variance**, *there is only one property describing the population.*

Definition

A **treatment** *or* **factor** *is a property that allows us to distinguish different populations from each other.*

One-Way Analysis of Variance

We have data that have been obtained so that a *factor* A is varied at k different levels $A_1,\ A_2,\ldots,A_k$. At level A_i we have n_i data values, $y_{i1}, y_{i2},\ldots,y_{in_i}$.
Our statistical model is that these are outcomes of random variables $Y_{i1}, Y_{i2},\ldots,Y_{in_i}$. We assume that all have $N(\mu_i,\sigma)$ distribution.
The quantities $\mu_1,\mu_2\ldots,\mu_k$ are thus means at the different levels.
Our goal is to compare these means.

One-Way Analysis of Variance: The Hypothesis

$\mu_1, \mu_2, \dots, \mu_k$ and σ^2 are unknown parameters. The statistical problem is to test whether all means are equal.

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

$$H_1 : \quad \mu_1, \ \mu_2, \ \cdots \ \text{and } \mu_k \text{ are not all equal}$$

H_1 simply says that H_o is not true. H_1 is not one-sided or two-sided, it is undirectional.

One-Way Analysis of Variance: data table

Level	Observations				Mean	Sample variance
A_1	y_{11}	y_{12}	$\dots$	y_{1n_1}	$\bar{y}_{1.}$	s_1^2
A_2	y_{21}	y_{22}	$\dots$	y_{2n_2}	$\bar{y}_{2.}$	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
A_k	y_{k1}	y_{k2}	$\dots$	y_{kn_k}	$\bar{y}_{k.}$	s_k^2

One-Way Analysis of Variance: notations

Here the notation is admittedly a complex one:

Level	Observations				Mean	S Sample variance
A_1	y_{11}	y_{12}	$\ldots$	y_{1n_1}	$\bar{y}_{.1}$	s_1^2

$y_{1.}$

means that we have summed over the second index:

$$y_{1.} = \frac{1}{n_1}\left(y_{11} + y_{12} + \ldots + y_{1n_1}\right)$$

Thus

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(y_{1j} - y_{1.}\right)^2$$

One-Way Analysis of Variance: notations

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$$

is the grand mean, $N = n_1 + n_2 + \cdots + n_k$ is total number of samples.

$$\bar{y}_{..} = \frac{1}{N} \left(y_{11} + y_{12} + \ldots + y_{1n_1} \right.$$
$$+ \ldots +$$
$$\left. \ldots + y_{k1} + y_{k2} + \ldots + y_{kn_k} \right)$$

One-Way Analysis of Variance

ANOVA estimates three sample variances: a **total variance** based on all the **observation deviations from the grand mean**, an **error variance** based on all the **observation deviations from their appropriate treatment means** $(y_{i.})$ and a treatment variance. The **treatment variance** is based on the **deviations of treatment means from the grand mean**, the result being multiplied by the number of observations in each treatment to account for the difference between the variance of observations and the variance of means.

One-Way Analysis of Variance

- Total SS $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2$
- SST$=$ Sum of squares for treatments (levels) or between samples $= \sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$
- SSE$=$ sums of squares for error within samples $= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2$

Total SS$=$ SST$+$ SSE

One-Way Analysis of Variance: The Hypothesis

$\mu_1, \mu_2, \ldots, \mu_k$ and σ^2 are unknown parameters. The statistical problem is to test whether all means are equal.

$$H_o : \; \mu_1 = \mu_2 = \cdots = \mu_k$$

This is the claim that all treatments, instruments e.t.c. are of equal quality.

One-Way Analysis of Variance: ANOVA Table

Source	df	SS	MSS
SST, Between samples	$k-1$	$\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	SS/df
SSE, Within samples	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$\widehat{\sigma}^2$=SS/df
Total SS	$N-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	

Source = source of variation, df= degrees of freedom, SS= sum of squares, MSS= mean sum of squares. and

$N = n_1 + n_2 + \cdots + n_k$, total number of samples.

ANOVA Table: treatment variance

$$\sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

measures the dispersion of the means. If this sum is large, then we may suspect that the factor levels are systematically different.

ANOVA Table: error variance

The second sum of squares can be written as

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2=\sum_{i=1}^{k}s_i^2(n_i-1)$$

where s_i^2 is the sample variance for level i. These are measurements of random variation, i.e., of σ^2.

ANOVA Table: the sums of squares

$$\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{k} s_i^2(n_i - 1)$$

By comparing these two sums of squares with each other we can decide if the levels are of equal value.

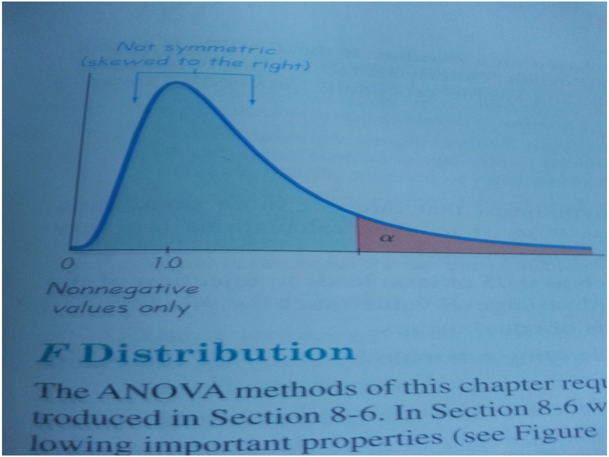
ANOVA Test Statistic

We should compare the two sums of squares by aid of their ratio so that the **test statistic** is taken as

$$F_A \overset{\text{def}}{=} \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\widehat{\sigma}^2} = \frac{\text{MSS between}}{\text{MSS within}}$$

It can be shown by an extensive exercise in mathematical statistics that F_A has an F-**distribution**, if H_o is true. (F for Sir R.A. Fisher).

ANOVA F-distribution


Not symmetric
(skewed to the right)
α
0
1.0
Nonnegative
values only

F Distribution

The ANOVA methods of this chapter requ
troduced in Section 8-6. In Section 8-6 w
lowing important properties (see Figure

ANOVA Test Statistic

Test statistic:

$$F_A = \frac{\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2/(k-1)}{\widehat{\sigma}^2} = \frac{\text{MSS between}}{\text{MSS within}}$$

The hypothesis H_A is rejected if $F_A > F_p(k-1, N-k)$. The critical value $F_p(k-1, N-k)$ gives the level of significance p, if H_o is true, and is found in a table for percentiles of the F-distribution.

F- distribution quantiles $F_{0.05}(f_2, f_1)$ %

f_2 / f_1	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242

KTH Matematik

ANOVA computations

These computations are simple and can in principle be done by hand. This is very time consuming and software is preferably used. There is statistical software, and even Excel can handle ANOVA tables.

The Millionaire calculating machine designed for ANOVA.

ANOVA Table

The lowermost row in the ANOVA table has not been given any role so far. It is there for computational reasons: this row is often easier to calculate and and the other rows are obtainable by subtraction. When relying on computers and calculators this makes little difference.

ANOVA: Example

Four instruments of measurement of length are compared. One operator measured one and the same length with each of the instruments. In the table we see the results.

Instrument	Observations			
A_1	1236	1238	1239	
A_2	1235	1234		
A_3	1236	1237	1238	
A_4	1233	1235	1234	1236

ANOVA: Example

Source	df	SS	MSS
Between instruments	3	$24\frac{3}{4}$	$33/4$
Within instruments	8	$12\frac{1}{6}$	$\widehat{\sigma}^2 = 73/48$
Total	11	$36\frac{11}{12}$	

Test statistic $F = \dfrac{33/4}{73/48} = 5.42 > 4.07 = F_{0.05}(3, 8)$. Hypothesis that the instruments are of equal value is thus rejected.

F-distribution, critical values

f_2 / f_1	1	2	3	4	5	6
1	161	200	216	225	230	234
2	18.5	19	19.2	19.2	19.3	19.3
3	10.1	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.1	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.2	3.09

Conditions for applying ANOVA

- We have k independent random samples, one form each k populations.
- Each of the k populations has a normal distribution with an unknown mean. μ_i is the unknown mean of the ith population. The means may be different in the different populations.
- All the populations have the same standard deviation σ, whose value is unknown.

Sir Ronald A. Fisher the founder of biostatistics (computing critical values of the F-distribution)