NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (6)

Normal sample mean and variance [2.5]
One-way ANOVA [2.6]

Mette Langaas

To be lectured: January 21, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

ANalysis Of VAriance (ANOVA)

Bingham and Fry (2010): Chapter 2

- — The Chi-square distribution and the F-distribution [2.1, 2.3]
- — Orthogonality and multivariate transformation formula [2.2, 2.4]
- — Normal sample mean and variance [2.5]
- — One-way ANOVA [2.6]
- — Two-way ANOVA [2.7-2.8]

Today

Primary aim chapter 1: ANalysis Of VAriance – comparing means by studying variability.

Normal sample: proof about important properties of sample mean $\bar{X}$ and sample variance S^2.

One-way ANOVA : use proof above to show properties of sample sums of squares.

Normal samples [2.5]

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Further, let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.
Then the following holds:

- $\bar{X}$ and S^2 are independent.
- $\bar{X} \sim N(\mu, \sigma^2/n)$.
- $nS^2/\sigma^2 \sim \chi^2_{n-1}$.

$\bar{X}$ end S^2 in normal samples [2.5]

identically

$X_1, X_2, \ldots, X_n$ i.i.d RV from $N(\mu, \sigma^2)$.

independent distributed

Estimators: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

i) $\bar{X}$ and S^2 are independent.

a) $Z_i = \dfrac{X_i - \mu}{\sigma} \Leftrightarrow X_i = \sigma Z_i + \mu$ $\quad Z = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$

Z_i i.i.d $N(0,1)$

b) $\bar{Z} \sim N(0, \frac{1}{n})$

$\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i \quad E(\bar{Z}) = \frac{1}{n} \sum E(Z_i) = 0$

$Var(\bar{Z}) = \frac{1}{n^2} \sum Var(Z_i) = \frac{1}{n^2} \sum_{i=1}^{n} 1 = \frac{1}{n}$

c) $\frac{n \cdot S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2}$ using $X_i = Z_i \sigma + \mu$

$\bar{X} = \bar{Z} \sigma + \mu$

$X_i - \bar{X} = Z_i \sigma + \mu - \bar{Z}\sigma - \mu = \sigma(Z_i - \bar{Z})$

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n} (Z_i - \bar{Z})^2$$

d) $$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - 2\bar{Z} \overbrace{\sum_{i=1}^{n} Z_i}^{n \cdot \bar{Z}} + n\bar{Z}^2$$

$$= \sum Z_i^2 - n\bar{Z}^2$$

$$\sum_{i=1}^{n} Z_i^2 = \underbrace{\sum_{i=1}^{n} (Z_i - \bar{Z})^2}_{\frac{nS^2}{\sigma^2}} + \underbrace{n\bar{Z}^2}_{n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2}$$

e) Let $\underset{n \times n}{P}$ be an orthogonal matrix with first row $\left[\frac{1}{\sqrt{n}}\ \frac{1}{\sqrt{n}}\ \dots\ \frac{1}{\sqrt{n}}\right]$.

This is called a Helmert transform. Many such matrices are possible, and can be found e.g. by the Gram-Schmidt procedure. One such matrix may be

$$\begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \\ \vdots & & & & \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 \cdots & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \end{bmatrix} = P$$

Why orthogonal? $P^T = P^{-1}$

$P^T P = \cancel{I}\ I$

Let $\underset{n\times 1}{W} = \underset{n\times n}{P}\ \underset{n\times 1}{Z}$

$$W_1 = [P]_{row\,1}\, Z = \frac{1}{\sqrt{n}}\Big(\underbrace{Z_1 + Z_2 + \dots + Z_n}_{n\bar{Z}}\Big) = \sqrt{n}\cdot\bar{Z}$$

$$W_1^2 = n\cdot\bar{Z}^2$$

$$\sum_{i=1}^{n} W_i^2 = W^T W = (PZ)^T PZ = Z^T \underbrace{P^T P}_{I} Z = Z^T Z = \sum_{i=1}^{n} Z_i^2$$

$$\sum_{i=2}^{n} W_i^2 = \sum_{i=1}^{n} W_i^2 - W_1^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2$$

f) $Z_1, Z_2, \dots, Z_n$ are i.i.d $N(0,1)$

$W = PZ$

$W_1, W_2, \dots, W_n$ are i.i.d $N(0,1)$ due to orthogonality theorem

Therefore W_1 will be independent of $W_2, W_3, \dots, W_n$

And W_1^2 will be independent of $W_2^2, W_3^2, \dots, W_n^2$, and of $\sum_{i=2}^{n} W_i^2$

Since $W_1 = \sqrt{n}\,\bar{Z} = \sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$, it is independent of $\sum_{i=2}^{n} W_i^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \frac{n}{\sigma^2} \cdot S^2$

$\Rightarrow$ This means that $\bar{X}$ and S^2 are independent

ii) $\bar{X} \sim N(\mu, \sigma^2/n)$

$$M_X(t) = \exp\left(\mu t + \tfrac{1}{2}\sigma^2 t^2\right)$$

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n = \exp\left(n \cdot \frac{\mu t}{n} + n \cdot \frac{1}{2}\sigma^2 \frac{t^2}{n^2}\right)$$

$$= \exp\left(\mu t + \frac{1}{2}\frac{\sigma^2}{n} \cdot t^2\right)$$

We recognise this as $N\left(\mu, \frac{\sigma^2}{n}\right)$

iii) $\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$

a) We have $\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\bar{Z}^2$

b) We know that $Z_i \sim N(0,1)$ and $Z_1, \dots, Z_n$ are independent, so that

$Z_i^2 \sim \chi^2_1$ and $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$

c) Further $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\,\bar{Z} \sim N(0,1)$ and

$n\bar{Z}^2 \sim \chi^2_1$

d) Subtraction property χ^2:

$$X = X_1 + X_2$$

χ^2_n

?

χ^2_1

$$\sum z_i^2 = \frac{nS^2}{\sigma^2} + n\bar{z}^2$$

Independent since $\bar{X}$ and S^2 are independent

χ^2_{n-1}

$$\Rightarrow \frac{nS^2}{\sigma^2} \text{ is } \chi^2_{n-1}$$

Note: If $\bar{X}$ and S^2 are independent $\Rightarrow$ data are normal

Teo 2.6: Fisher's lemma

$X_1, \dots, X_n$ iid $N(0, \sigma^2)$

$$Y_i = \sum_{j=1}^{n} c_{ij} X_j \qquad i = 1, \dots, p \qquad p < n$$

where the row of $C = \{c_{ij}\}$ are orthogonal for $i = 1, \dots, p$.

If $$S^2 = \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{p} Y_i^2$$

then

(i) S^2 is independent of $Y_1, \dots, Y_p$

(ii) $$\frac{S^2}{\sigma^2} \sim \chi^2_{n-p}$$

typo in book

Rothamsted Experimental Station

- founded in 1843 by John Bennet Lawes on his inherited 16t century estate, Rothamsted Manor,
  - wanted to investigate the impact of inorganic and organic fertilizers on crop yield
  - had founded a fertilizer manufacturing company in 1842
- Lawes appointed the chemist Joseph Henry Gilbert to the directorship of the chemical laboratory
- the two began a series of field experiments to examine the effects of inorganic fertilizers and organic manures on the nutrition and yield of a number of important crops

http://www.stats.uwo.ca/faculty/bellhouse/stat499lecture13.pdf

The Broadbalk Field Trial at Rothamsted

- this was the first field trial started by Lawes and Gilbert
- began in 1843
- purpose was to investigate the relative importance of different plant nutrients (N, P, K, Na, Mg) on grain yield of winter wheat
- weeds were controlled by hand hoeing and fallowing
  - now some herbicides are used
- The experiment continues to this day

http://www.stats.uwo.ca/faculty/bellhouse/stat499lecture13.pdf

Concrete aggregates example

— Aggregates are inert granular materials such as sand, gravel, or crushed stone that, along with water and portland cement, are an essential ingredient in concrete.

— For a good concrete mix, aggregates need to be clean, hard, strong particles free of absorbed chemicals or coatings of clay and other fine materials that could cause the deterioration of concrete.

— We could like to examine 5 different aggregates, and measure the absorption of moisture after 48hrs exposure (to moisture).

— A total of 6 samples are tested for each aggregate.

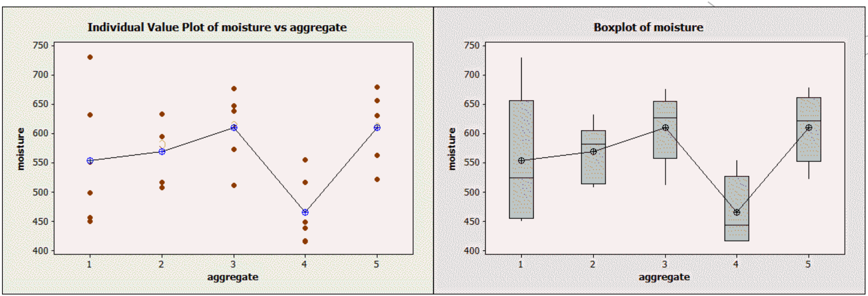
— Research question: Is there a difference between the aggregates with respect to absorption of moisture?

Concrete aggregates data

Aggregate:	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Table 13.1 of WMMY.

Concrete aggregates example



One-way ANOVA [2.6]

Aim: study the relationship between

a treatment (factor) with r levels and

aggregates

absorption of moisture.

a continuous response.

factor

Assume: $X_{ij} \sim N(\mu_i, \sigma^2)$ $\quad i = 1, \ldots, r$

independent

$j = 1, \ldots, n_i$

$n = \sum_{i=1}^{r} n_i$

We want to investigate (test) if the μ_i's are different.

$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$ vs. H_1: at least two different.

Let $\mu = \dfrac{\sum n_i \mu_i}{n}$ be called population grand mean.

Notation (sample): $X_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$ group mean

$$X_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^{r} n_i \cdot X_{i\cdot}$$

grand mean

bullet $X_{\cdot}$ = averaged out

SS =

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (X_{ij} - X_{\cdot\cdot})^2 = \sum\sum \left(\underbrace{X_{ij} - X_{i\cdot}} + \underbrace{X_{i\cdot} - X_{\cdot\cdot}} \right)^2$$

Homework = continue here