

Multivariate Analysis of Variance

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague



<http://cw.felk.cvut.cz/wiki/courses/b4m36san/start>

Agenda

- Bivariate statistical tests and their multivariate generalizations,
- relationship between continuous variables and a categorical variable
 - categorical variable = treatment, factor,
 - lots of methods, we will proceed from the most simple to most general,
- Review t-test for two groups
 - single continuous variable, binary factor/treatment,
 - non-parametric alternative,
 - multiple comparisons problem for more groups,
- Explain ANOVA
 - posthoc tests to find out which groups contributed most,
- Generalize towards MANOVA
 - two-way modification, non-parametric

Bivariate statistical models and tests

- assess strength of relationship between a pair of variables
 - independent (causal) and dependent (effect) variable,
 - rejection of null hypothesis does not imply causal relationship,
- all of them can be generalized towards multivariate statistics.

		dependent variable	
		categorical	continuous
independent variable	categorical	contingency table chi-square test	analysis of variance
	continuous	LDA logistic regression	correlation regression

Independence test for two categorical variables

- categorical variable
 - takes one of a limited (and fixed) number of possible values,
- contingency table
 - table showing observed (multivariate) joint frequency distribution,
 - for the moment concern two-way contingency tables only,
 - a pair of variables with r and c categories captured in a $r \times c$ table,
 - its elements represent frequency counts for the individual events,
 - an example: two binary variables $X_1 = \textit{gender}$ and $X_2 = \textit{disease}$

	X_{21}	\dots	X_{2c}	Σ
X_{11}	N_{11}		N_{1c}	$N_{1\bullet}$
\dots				
X_{1r}	N_{r1}		N_{rc}	$N_{r\bullet}$
Σ	$N_{\bullet 1}$		$N_{\bullet 2}$	N

	healthy	diseased	total
women	216	72	288
men	279	342	621
total	495	414	909

Independence test for two categorical variables

- independence assumption
 - H_0 : two categorical variables are independent,
 - H_a : they have an association or relationship (of an unknown structure),
 - the frequency distribution does not change with the table rows,
- compare the observed frequencies with the expected ones
 - the expectations are derived from the marginal frequencies under the independence assumption, MLE approach is taken,
 - $E_{ij} = N\bar{p}_{i.}\bar{p}_{.j} = N\frac{N_{i.}}{N}\frac{N_{.j}}{N} = \frac{N_{i.}N_{.j}}{N}$

O_{ij}	healthy	diseased	total
women	216	72	288
men	279	342	621
total	495	414	909

E_{ij}	healthy	diseased	total
women	157	131	288
men	338	283	621
total	495	414	909

Independence test for two categorical variables

- let us measure the discrepancy between the observed counts and the estimated expected counts under the null,
- Pearson's χ^2 is one of the options

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

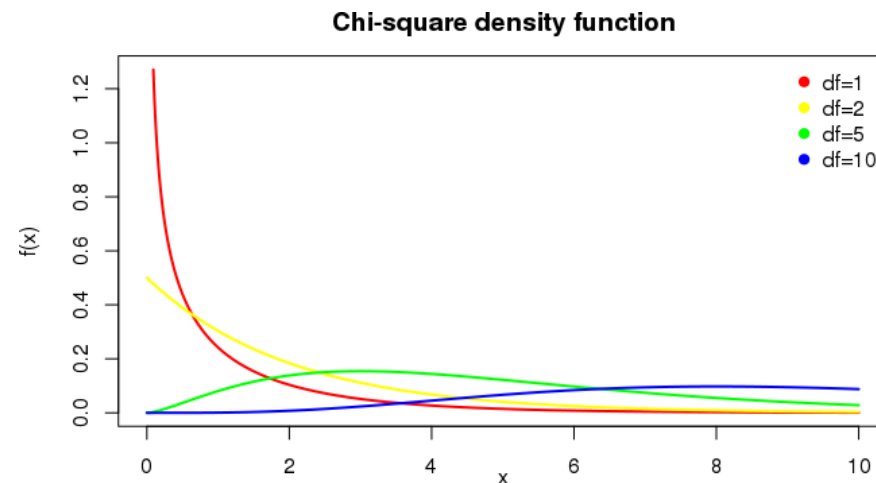
- a cumulative test statistic,
- it asymptotically approaches a χ^2 distribution
 - with $(r - 1)(c - 1)$ degrees of freedom,
- assumptions
 - non-parametric test, robust wrt distribution of the data,
 - one observation per subject, sufficient sample size ($E_{ij} \geq 5$).

Independence test for two categorical variables

- for the *gender* and *disease* relationship

$$X^2 = \frac{(216 - 157)^2}{157} + \frac{(72 - 131)^2}{131} + \frac{(279 - 338)^2}{338} + \frac{(342 - 283)^2}{283} = 71.3$$

- choose a significance level $\alpha = 0.01$ (type I error control),
- compare with the table value $\chi_{\alpha=0.01, df=1}^2 = 6.635$,
- since $X^2 > \chi_{df=1}^2$ reject H_0 ,
- the corresponding p-value: $p = 1 - F_{\chi^2(1)}(71.3) = 1.09e - 17$.



Categorical independent vs continuous dependent variable

- Review t-test for two groups

- a test in which the test statistic follows a Student's t-distribution ...
- under the null hypothesis,

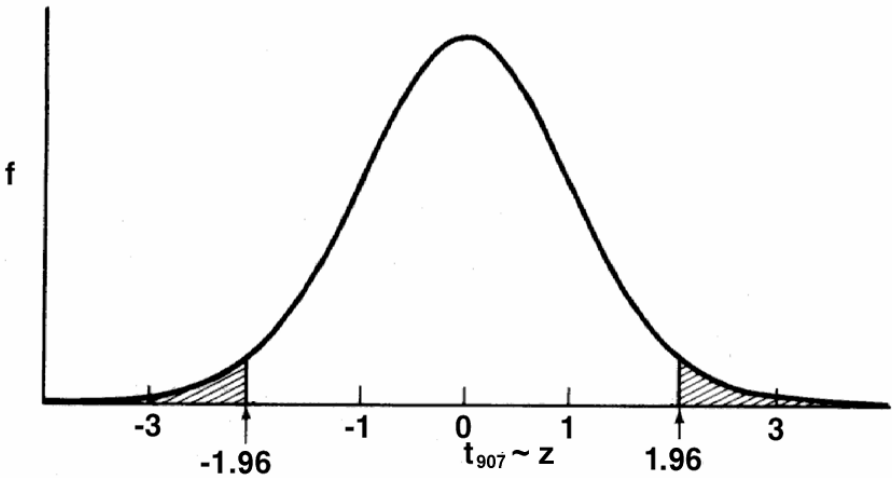
- consider a two sample t-test, $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 \neq \mu_2$

- the two populations should follow a normal distribution,
- variances of the two populations assumed equal → Student's t-tests,
- variances can differ → **Welch's test** (see below),

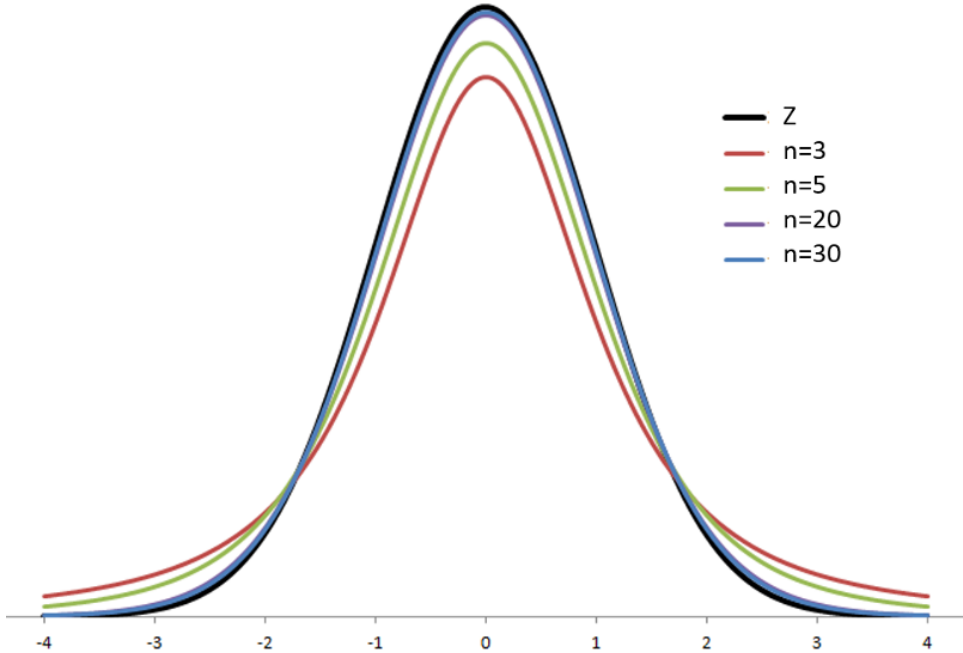
$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$$

- \bar{X}_i , s_i^2 and n_i ... sample means, variances and sizes,
- $df \leq n_1 + n_2 - 2$, the exact formula complicated,
- reject H_0 if $|t_{obs}| \geq t_{df, 1-\alpha/2}$.

t-distribution



Statlect: The Digital Textbook



Statlect: The Digital Textbook

T-test for multiple groups

- Concern a categorical variable with many levels → multiple groups
 - the hypotheses of interest
 - * $H_0 : \mu_1 = \mu_2 = \dots = \mu_g,$
 - * $H_a : \mu_i \neq \mu_j$ for at least one $i \neq j.$
- conduct a two-sample t-test for a difference in means for each pair of groups
 - the number of comparisons grows quadratically with the number of groups/levels,
- for $\alpha = 0.05$ for each comparison
 - there is a 5% chance that each comparison will falsely be called significant,
 - the overall probability of Type I error is elevated above 5%,
 - we falsely reject at least one of the partial null hypothesis with probability

$$1 - (1 - \alpha)^{\binom{g}{2}}$$

- e.g., for $g=4$ (four groups) it makes $0.26 \gg \alpha,$
- **multiple comparisons** must be corrected.

Multiple comparisons must be corrected

- often we control **family-wise error rate** (FWER)
 - the probability of making one or more false discoveries (type I errors) when performing multiple hypotheses tests,
 - the most simple FWER control is the **Bonferroni correction**,
 - test each hypothesis at level $\alpha_{indiv} = \alpha_{overall}/m$,
 - * m stands for the number of individual pair tests,
 - * follows from Bonferroni inequality for independent tests

$$\alpha_{overall} = 1 - (1 - \alpha)^m \leq m\alpha_{indiv}$$

- * in our case with 4 groups $m = \binom{4}{2} = 6$,
- * the B. inequality obviously holds

$$0.26 = 1 - 0.95^6 < 0.05 * 6 = 0.3$$

- however, this adjustment may be too conservative
 - * insufficient **power**, often does not reject H_0 although H_a is true.

Analysis of variance (ANOVA)

- compares means for multiple (usually $g \geq 3$) independent populations
 - parametric and unpaired, one-way,
 - relationship between a categorical factor F and a continuous outcome Y ,
 - extends a two sample t-test to multiple groups,

Subject	F	Y
1	f_1	y_1
2	f_2	y_2
...		
N	f_N	y_N

		1	...	g
Subject	1	y_{11}	...	y_{g1}
	2	y_{12}	...	y_{g2}

	n_i	y_{1n_1}	...	y_{gn_g}

- y_{ij} ... observation for subject j in group i ,
- n_i ... number of subjects in group i ,
- $N = n_1 + n_2 + \dots + n_g$... total sample size.

Analysis of variance (ANOVA)

■ assumptions

- the subjects are **independently sampled**
 - * employ repeated measures ANOVA otherwise,
- the data are **normally distributed** in each group
 - * $E(Y_{i.}) = \mu_i$, e.g., no group sub-populations with different means,
 - * residuals of the model below show the normal distribution
$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}$$
 - * employ non-parametric Kruskal-Wallis test otherwise,
- the data are **homoscedastic**
 - * the variability in the data does not depend on group membership,
 - * there is a common variance $var(Y_{ij}) = \sigma^2$,

■ the hypotheses of interest

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_g$,
- $H_a : \mu_i \neq \mu_j$ for at least one $i \neq j$.

Analysis of variance (ANOVA)

■ method

- partition SS_{total} , the total variation in a response variable,
- distinguish **within groups variability** SS_{error} ,
- and **between groups variability** SS_{treat} ,

$$\begin{aligned}SS_{total} &= \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \\&= \sum_{i=1}^g \sum_{j=1}^{n_i} ((y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..}))^2 = \\&= \underbrace{\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}_{SS_{error}} + \underbrace{\sum_{i=1}^g n_i (\bar{y}_{i.} - \bar{y}_{..})^2}_{SS_{treat}}\end{aligned}$$

* $\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$... group i sample mean,

* $\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^g \sum_{j=1}^{n_i} y_{ij}$... grand mean.

Analysis of variance (ANOVA)

- method

- in a similar manner, partition the number of degrees of freedom that stand behind the observed sums of the squared deviations

$$DF_{total} = N - 1 = DF_{error} + DF_{treat} = (N - g) + (g - 1) = N - 1$$

- decide whether group averages differ more than based on random variability observed in the dependent variable under the null hypothesis,
- employ **mean square** variability, both within groups and between groups

$$MS_{error} = \frac{SS_{error}}{DF_{error}} = \frac{SS_{error}}{N - g} \quad MS_{treat} = \frac{SS_{treat}}{DF_{treat}} = \frac{SS_{treat}}{g - 1}$$

Analysis of variance (ANOVA)

- method

- compare the variance between the groups and within the groups,

$$F_{obs} = \frac{MS_{treat}}{MS_{error}} \sim F_{g-1, N-g}$$

- if F_{obs} is small (close to 1), then variability between groups is negligible compared to variation within groups and the grouping does not explain much variation in the data,
- if F_{obs} is large, then variability between groups is large compared to variation within groups and the grouping explains a lot of the variation in the data

- decision rule based on F_{obs}

- reject H_0 if $F_{obs} \geq F_{\alpha, g-1, N-g}$,
- fail to reject H_0 if $F_{obs} < F_{\alpha, g-1, N-g}$.

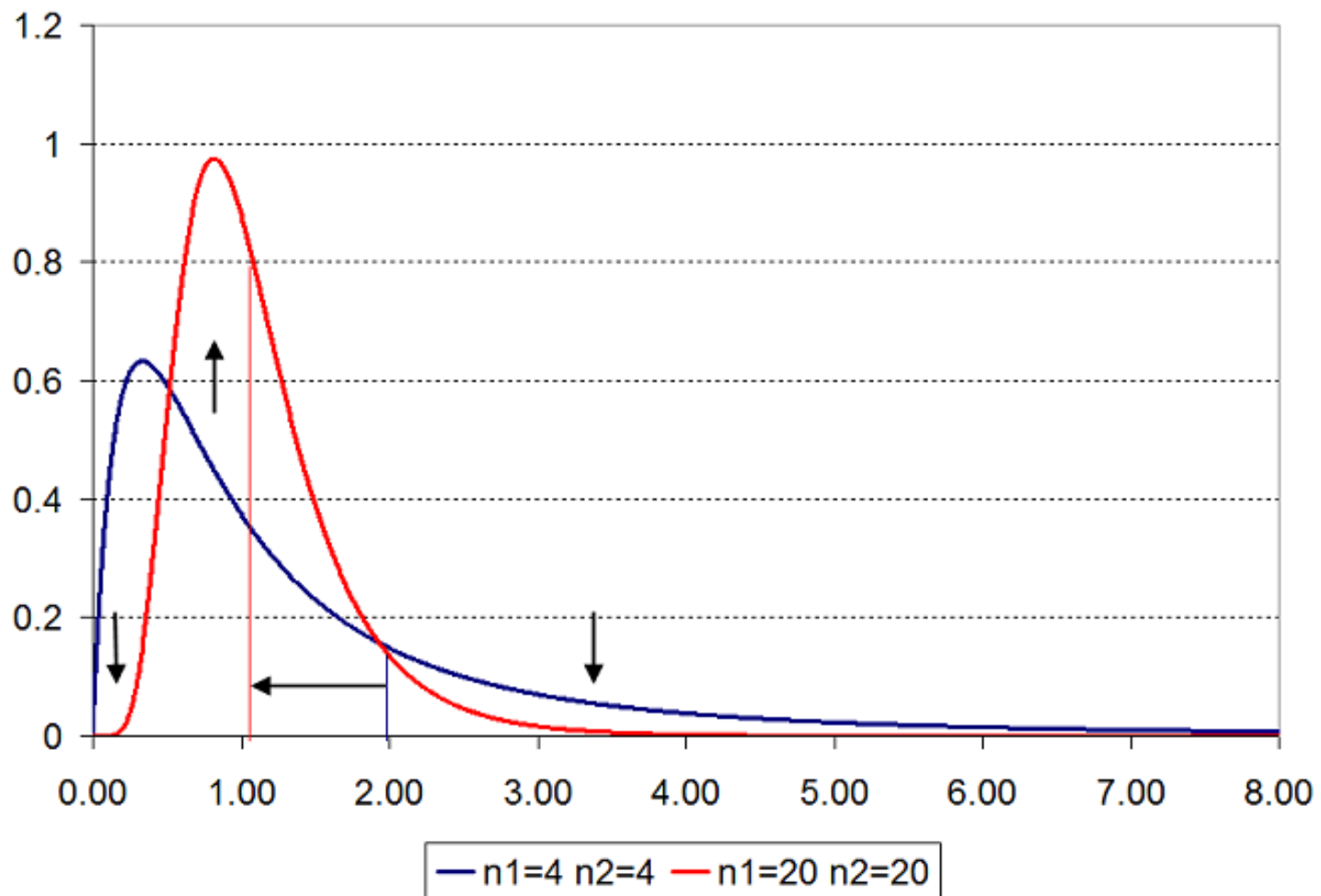
F-distribution

- F-distribution is any distribution obtained by taking the quotient of two χ^2 distributions divided by their respective degrees of freedom,
- consequently, any F-distribution has two parameters corresponding to the degrees of freedom for the two χ^2 distributions
- given $X_1 \sim \chi_{df_1}^2$ and $X_2 \sim \chi_{df_2}^2$

$$\frac{X_1/df_1}{X_2/df_2} \sim F_{df_1,df_2}$$

- F-distribution in R
 - find the value of $F_{\alpha,g-1,N-g}$:
`qf(alpha, df1, df2, lower.tail = F)`,
 - find the ANOVA p-value when knowing F_{obs} :
`pf(Fobs, df1, df2, lower.tail = F)`.

F-distribution



Statlect: The Digital Textbook

Post-hoc ANOVA tests

- after performing ANOVA (and rejecting the null hypothesis)
 - we only assume that there is some difference in group means,
- a post-hoc test identifies which particular groups stand behind the test outcome,
- Tukey's HSD (honest significant difference) test
 - a t-test that controls for family-wise error rate (FWER),
 - compares all pairs of group means,
 - identifies all pairs whose difference is larger than expected standard error,
 - observed test statistics related to the studentized range distribution,

$$q_{obs} = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{\frac{MS_{error}}{n^*}}} \sim q_{g, N-g}$$

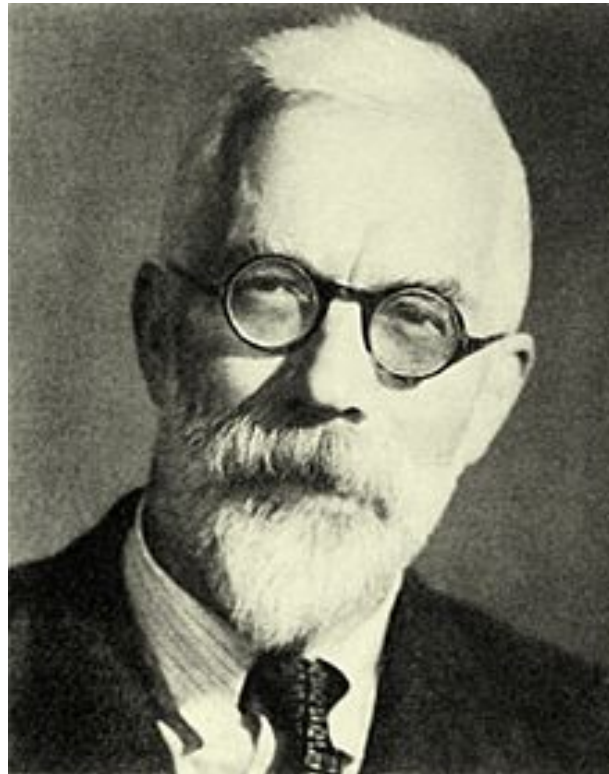
- n^* ... number of observations per group (their harmonic mean if not equal),
- always positive, sort the means before its application.

ANOVA extensions/alternatives

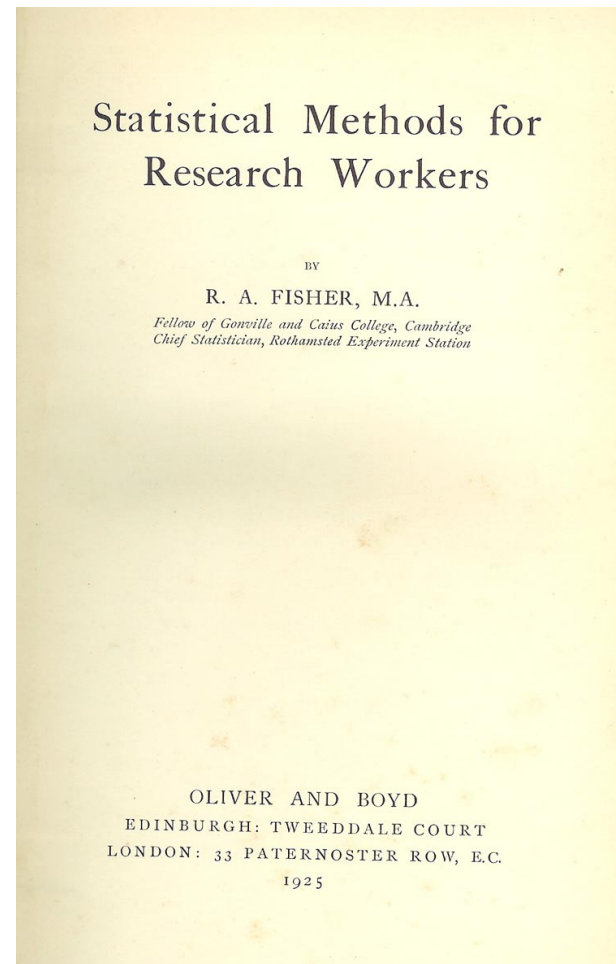
- up to now we talked about ANOVA that
 - is parametric,
 - deals with independent measurements,
 - is one-way (with a single factor),
 - concerns a single target variable only,
- other options
 - non-parametric analysis (Wilcoxon test → Kruskal-Wallis analysis),
 - compares all possible group means (repeated measures ANOVA, Friedman test if non-parametric too),
 - main effects ANOVA and factorial ANOVA,
 - multivariate ANOVA (MANOVA).



There is a big name behind every test . . .



Sir Ronald Fisher (1890-1962), evolutionary biologist and statistician



His work considered to define modern statistics

Multivariate analysis of variance (MANOVA)

- p variables measured on each subject, objects categorized into g disjoint groups.
- y_{ijk} ... an observation for variable k from subject j in group i ,
- \mathbf{y}_{ij} ... a vector of dependent variables for subject j in group i ,
- assumptions
 - the subjects are **independently sampled**,
 - the data are **multivariate normally distributed** in each group,
 - the data from all groups have common covariance matrix Σ ,
 - the data from group i has common mean vector μ_i of length p ,
- the hypotheses of interest
 - $H_0 : \mu_1 = \mu_2 = \dots = \mu_g$,
 - $H_a : \mu_{ik} \neq \mu_{jk}$ for at least one $i \neq j$ and at least one variable k .

Multivariate analysis of variance (MANOVA)

- method

- the analogy of SS_{total} in ANOVA is a $p \times p$ **cross products matrix** \mathbf{T} ,
- similarly to ANOVA, it can be decomposed into the **Error Sum of Squares and Cross Products** \mathbf{E} , and the **Hypothesis Sum of Squares and Cross Products** \mathbf{H} .

$$\begin{aligned}
 \mathbf{T} &= \sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{..})' = \\
 &= \sum_{i=1}^g \sum_{j=1}^{n_i} \{(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) + (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{..})\} \{(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) + (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{..})\}' = \\
 &= \underbrace{\sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'}_{\mathbf{E}} + \underbrace{\sum_{i=1}^g n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{..})'}_{\mathbf{H}}
 \end{aligned}$$

* $\bar{\mathbf{y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$... sample mean vector for group i ,

* $\bar{\mathbf{y}}_{..} = \frac{1}{N} \sum_{i=1}^g \sum_{j=1}^{n_i} \mathbf{y}_{ij}$... grand mean vector of length p .

Multivariate analysis of variance (MANOVA)

- explanation of the elements of **T**, **E** and **H**

- the element $t_{k,l}$ is

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ijk} - \bar{y}_{..k})(y_{ijl} - \bar{y}_{..l})$$

- for $k = l$ it is the total sum of squares for variable k , and measures the total variation in the k th variable, for $k \neq l$, this measures the dependence between variables k and l across all of the observations,

- the element $e_{k,l}$ is

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ijk} - \bar{y}_{i.k})(y_{ijl} - \bar{y}_{i.l})$$

- for $k = l$ it is the error sum of squares for variable k , and measures the within treatment variation for the k th variable, for $k \neq l$ it measures the dependence between variables k and l after taking into account the treatment,

Multivariate analysis of variance (MANOVA)

- explanation of the elements of **T**, **E** and **H**

- the element $\mathbf{h}_{k,l}$ is

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

- for $k = l$ it is the treatment sum of squares for variable k , and measures the between treatment variation for the k th variable, for $k \neq l$, this measures dependence of variables k and l across treatments.
- consequently, if the hypothesis sum of squares and cross products **H** is large relative to the error sum of squares and cross products matrix **E** we wish to reject H_0 .

Multivariate analysis of variance (MANOVA)

- Wilk's lambda test statistics for MANOVA (several other statistics exist too)
 - the determinant of the error matrix \mathbf{E} is divided by the determinant of the total matrix $\mathbf{T} = \mathbf{H} + \mathbf{E}$, we will reject the null hypothesis if Wilk's lambda is small/close to zero as then \mathbf{H} is large relative to \mathbf{E} too.

$$\Lambda^* = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

- can also be computed using the eigenvalues $\hat{\lambda}$ of $\mathbf{E}^{-1}\mathbf{H}$ ($s = \min(p, g-1)$)

$$\Lambda^* = \prod_{i=1}^s \frac{1}{1 + \hat{\lambda}_i}$$

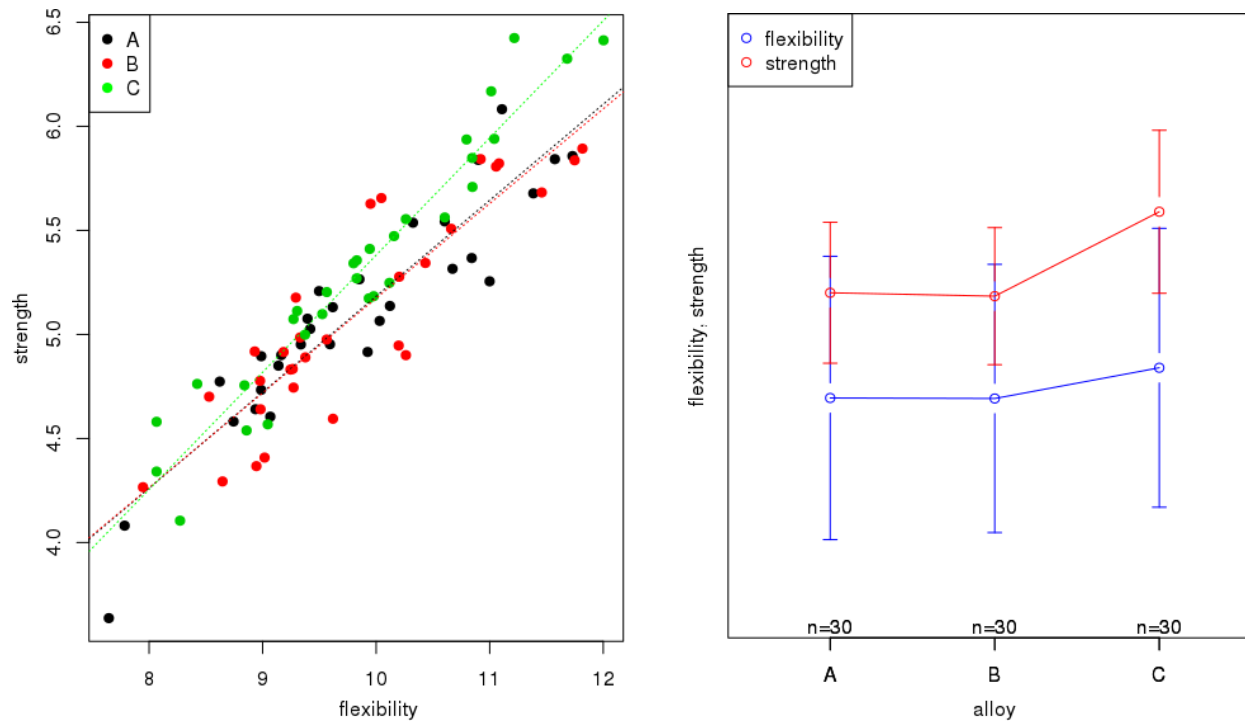
- the distribution of Λ^* is not tractable, we can only have approximations,
- e.g., Bartlett's approximation can be used if N is large

$$-\left(N - 1 - \frac{p + g}{2}\right) \ln \Lambda^* > \chi_{p(g-1), \alpha}^2$$

A typical case in which MANOVA helps

■ Mechanical engineering domain

- 90 samples of three different alloys (A, B, C),
- samples differ in flexibility and strength,
- flexibility and strength correlated, strength in C slightly increased,
- goal: decide (detect) the influence of alloy on flexibility and strength.



A typical case in which MANOVA helps

- ANOVA outcome:

```
> summary(aov(flexibility ~ alloy, alloys))
              Df Sum Sq Mean Sq F value Pr(>F)
alloy          2   0.14  0.0712   0.068  0.935
Residuals     87  91.69  1.0539

> summary(aov(strength ~ alloy, alloys))
              Df Sum Sq Mean Sq F value Pr(>F)
alloy          2   1.051  0.5254   1.759  0.178
Residuals     87 25.989  0.2987
```

- MANOVA outcome:

```
> summary(manova(cbind(flexibility, strength) ~ alloy, alloys))
              Df  Wilks approx F num Df den Df  Pr(>F)
alloy          2  0.8577   3.4313     4    172 0.00998 **
Residuals     87
```

Summary

- MANOVA compares multivariate sample means
 - it deals with multiple dependent variables at the same time,
- MANOVA advantages over ANOVA
 - better chance to discover which factor is truly important,
 - protects against Type I errors in multiple independent ANOVA runs,
 - increased power, it can reveal differences not discovered by ANOVA tests,
- MANOVA cautions
 - a complicated design, more difficult to disambiguate,
 - one degree of freedom is lost for each dependent variable that is added,
 - unsuitable if the dependent variables are perfectly correlated or uncorrelated,
- typically followed by significance tests on individual dependent variables.

The main references

:: Resources (slides, scripts, tasks) and reading

- STAT 505 course on Applied Multivariate Statistical Analysis, PennState University, <https://onlinecourses.science.psu.edu/stat505/>.
- G. James, D. Witten, T. Hastie and R. Tibshirani: **An Introduction to Statistical Learning with Applications in R**. Springer, 2014.
- A. C. Rencher, W. F. Christensen: **Methods of Multivariate Analysis**. 3rd Edition, Wiley, 2012.
- T. Hastie, R. Tibshirani and J. Friedman: **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2009.