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

- Explain ANOVA
  - relationship between continuous variables and a categorical variable
    - categorical variable = treatment, factor,
  - relationship with t-test for two groups,
  - posthoc tests to find out which groups contributed most,
  - relationship with linear regression,
- Generalize towards MANOVA
  - procedure for comparing multivariate sample means,
  - two-way modification, non-parametric.

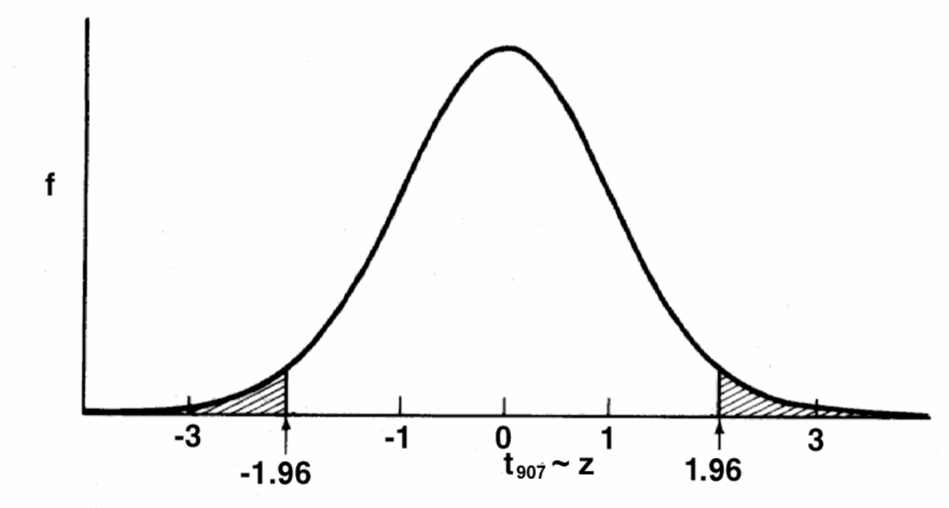
Categorical dependent vs continuous independent variable

- 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 $\rightarrow$ Student's t-tests,
  - variances can differ $\rightarrow$ **Welch's test** (see below),

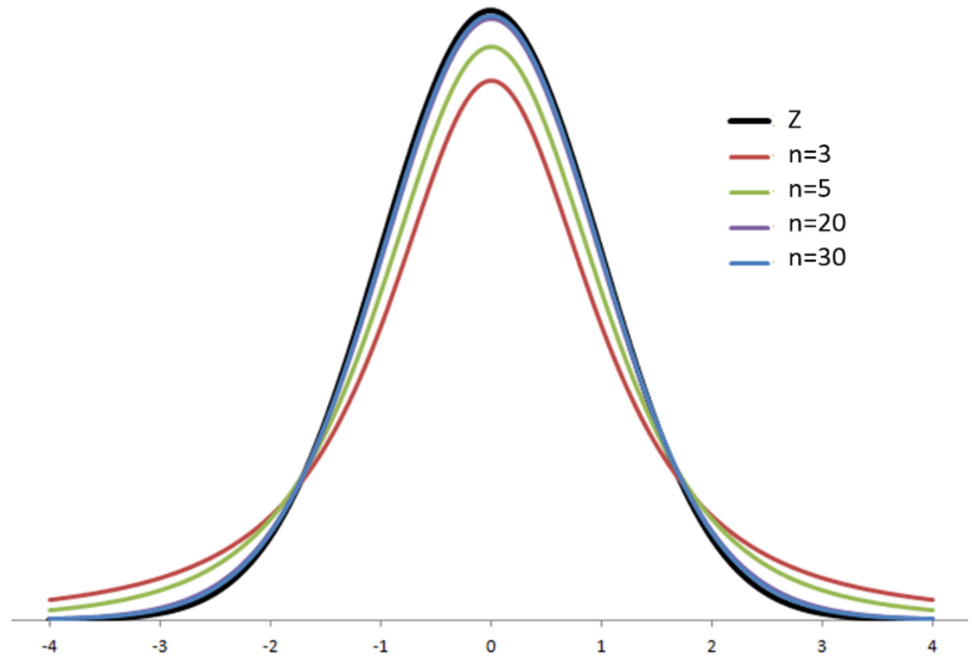
$$t_{obs} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \sim \quad t_{df}$$

  - $\bar{X}_i$, 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 $\rightarrow$ multiple groups
  - the hypotheses of interest
    * $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$,
    * $H_a : \mu_i \neq \mu_j$ 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 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
$\ldots$		
N	f_N	y_N

		1	$\ldots$	g
Subject	1	y_{11}	$\ldots$	y_{g1}
	2	y_{12}	$\ldots$	y_{g2}
	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	n_i	y_{1n_1}	$\ldots$	y_{gn_g}

- y_{ij} ...observation for subject j in group i,
- n_i ...number of subjects in group i,
- $N = n_1 + n_2 + ... + n_g$ ...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
  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$,
  - $H_a : \mu_i \neq \mu_j$ 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},

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

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

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

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 $F_{obs} < F_{\alpha,g-1,N-g}$.

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,

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

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

ANOVA vs linear regression

- Is there any link between testing of linear models and ANOVA?

$$F_{ANOVA} = \frac{SS_{treat}/(g-1)}{SS_{error}/(N-g)}$$

$$F_{LR} = \frac{(TSS - RSS)/p}{RSS/(m-p-1)}$$

ANOVA vs linear regression

- Is there any link between testing of linear models and ANOVA?

$$F_{ANOVA} = \frac{SS_{treat}/(g-1)}{SS_{error}/(N-g)}$$

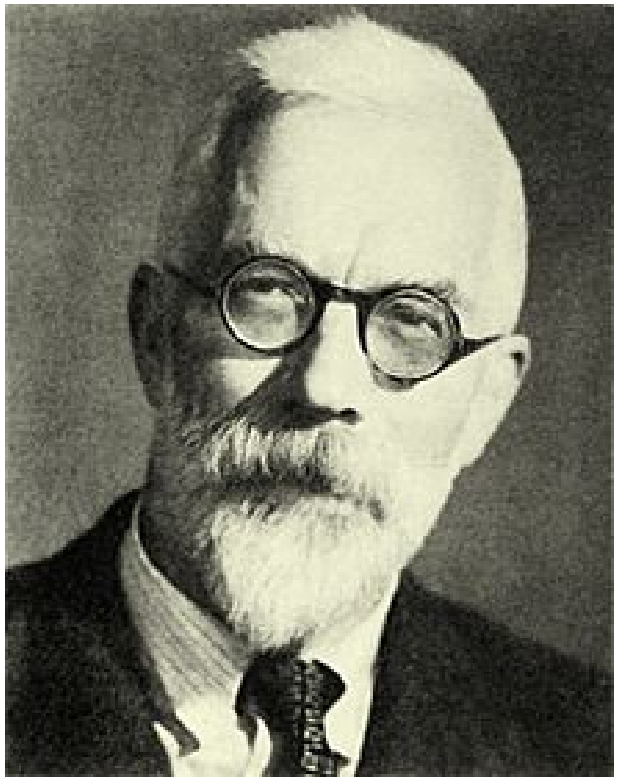
$$F_{LR} = \frac{(TSS - RSS)/p}{RSS/(m-p-1)}$$

  - the same principle,
  - exactly the same outcome for single independent categorical variable.

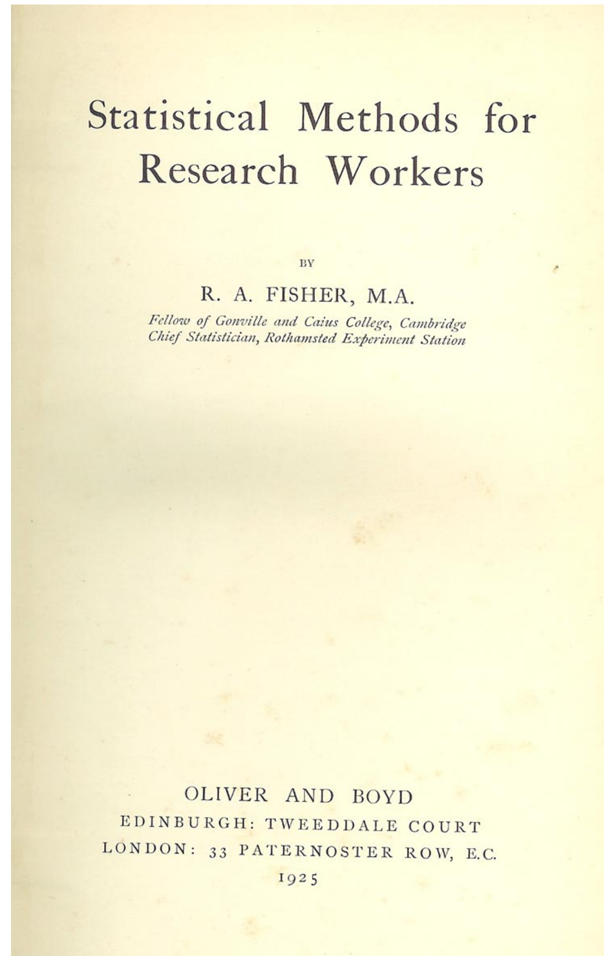
ANOVA extensions/alternatives

- 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 $\rightarrow$ 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



Statistical Methods for
Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

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 $\mu_{\mathbf{i}}$ of length p,
- the hypotheses of interest
  - $H_0 : \mu_{\mathbf{1}} = \mu_{\mathbf{2}} = \cdots = \mu_{\mathbf{g}}$,
  - $H_a : \mu_{\mathbf{ik}} \neq \mu_{\mathbf{jk}}$ 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** $\mathbf{H}$.

$$\mathbf{T} = \sum_{\mathbf{i=1}}^{\mathbf{g}} \sum_{\mathbf{j=1}}^{\mathbf{n_i}} (\mathbf{y_{ij}} - \bar{\mathbf{y}}_{..})(\mathbf{y_{ij}} - \bar{\mathbf{y}}_{..})' =$$

$$= \sum_{\mathbf{i=1}}^{\mathbf{g}} \sum_{\mathbf{j=1}}^{\mathbf{n_i}} \{(\mathbf{y_{ij}} - \bar{\mathbf{y}}_{\mathbf{i}.}) + (\bar{\mathbf{y}}_{\mathbf{i}.} - \bar{\mathbf{y}}_{..})\}\{(\mathbf{y_{ij}} - \bar{\mathbf{y}}_{\mathbf{i}.}) + (\bar{\mathbf{y}}_{\mathbf{i}.} - \bar{\mathbf{y}}_{..})\}' =$$

$$= \underbrace{\sum_{\mathbf{i=1}}^{\mathbf{g}} \sum_{\mathbf{j=1}}^{\mathbf{n_i}} (\mathbf{y_{ij}} - \bar{\mathbf{y}}_{\mathbf{i}.})(\mathbf{y_{ij}} - \bar{\mathbf{y}}_{\mathbf{i}.})'}_{\mathbf{E}} + \underbrace{\sum_{\mathbf{i=1}}^{\mathbf{g}} \mathbf{n_i}(\bar{\mathbf{y}}_{\mathbf{i}.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{\mathbf{i}.} - \bar{\mathbf{y}}_{..})'}_{\mathbf{H}}$$

$*$ $\bar{\mathbf{y}}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$ ... 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 $\mathbf{T}$, $\mathbf{E}$ and $\mathbf{H}$
  - the element $\mathbf{t}_{k,l}$ 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 kth variable, for $k \neq l$, this measures the dependence between variables k and l across all of the observations,
  - the element $\mathbf{e}_{k,l}$ 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 kth 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 $\mathbf{T}$, $\mathbf{E}$ and $\mathbf{H}$
  - 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 kth 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 $\mathbf{H}$ is large relative to the error sum of squares and cross products matrix $\mathbf{E}$ 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 $\mathbf{E}^{-1}\mathbf{H}$ ($s = min(p, g-1)$)

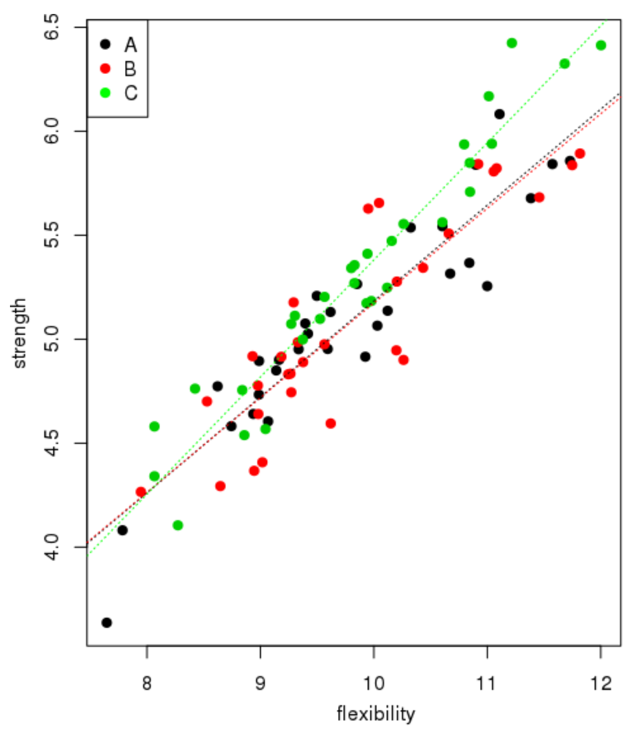
$$\Lambda^* = \prod_{i=1}^{s} \frac{1}{1 + \hat{\lambda}_i}$$

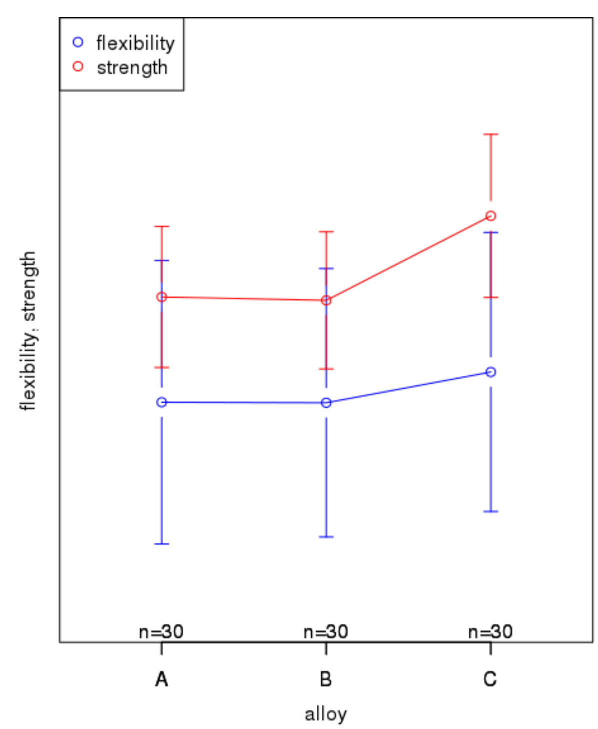
  - the distribution of Λ^* is not tractable, we can only have approximations,
  - e.g., Bartlett's approximation can be used if N is large

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

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.