Vergleich mehrerer multiv. Mittelwerte

Es gibt 3 verschiedene Situationen:

- Vergleich von Mittelwerten aus gepaarten Stichproben aus wiederholten Messungen

  (kann durch Änderung der T^2-Statistik gelöst werden)

- Vergleich von Mittelwerten aus zwei Stichproben von verschiedenen Grundgesamtheiten

  (kann durch Änderung der T^2-Statistik gelöst werden)

- Vergleich mehrerer Mittelwerte aus Stichproben von verschiedenen Grundgesamtheiten

  (MANOVA)

Gepaarte Stichproben

Angenommen es werden p Variablen am gleichen Objekt zweimal gemessen, z.B.:

- Kontroll Bedingungen vs. Behandlung
- Vor Einnahme eines Medikamentes und nach der Einnahme
- Qualitätskontrolle:
  Zwei Labore untersuchen die gleichen Objekte und ihre Ergebnisse werden verglichen.

Wichtig:
Es handelt sich immer um die gleiche Instanz, an der gemessen bzw. beobachtet wurde.

Notation

$j = 1, \ldots, n$ Beobachtungen:

$$
\begin{aligned}
X_{1j1} &= Variable\ 1\ unter\ Behandlung\ 1 \\
&\vdots \\
X_{1jp} &= Variable\ p\ unter\ Behandlung\ 1
\end{aligned}
$$

$$
\begin{aligned}
X_{2j1} &= Variable\ 1\ unter\ Behandlung\ 2 \\
&\vdots \\
X_{2jp} &= Variable\ p\ unter\ Behandlung\ 2
\end{aligned}
$$

Die *neue Stichprobe*, $\mathbf{D}_j = (D_{j1}, D_{j2}, \ldots, D_{jp})'$, ergibt sich aus

$$
\begin{aligned}
D_{j1} &= X_{1j1} - X_{2j1} \\
D_{j2} &= X_{1j2} - X_{2j2} \\
&\vdots \\
D_{jp} &= X_{1jp} - X_{2jp}
\end{aligned}
$$

...

Sei $\boldsymbol{\delta}$ der Mittelwert der (Differenz der) Grundgesamtheit, und angenommen die Stichprobe $\mathbf{D}_j$ ist unabhängig $N_p(\boldsymbol{\delta}, \Sigma_d)$ verteilt, dann ist der Test zum Niveau α für $H_o : \boldsymbol{\delta} = \mathbf{0}$ gegen $H_A : \boldsymbol{\delta} \neq \mathbf{0}$ verwerfe H_o falls:

$$T^2 = n\bar{\mathbf{D}}'\mathbf{S}_d^{-1}\bar{\mathbf{D}} > \frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}(\alpha)$$

Damit ist die $100(1-\alpha)\%$ Konfidenz Region für $\boldsymbol{\delta}$ die Menge aller $\boldsymbol{\delta}$ die

$$(\bar{\mathbf{D}} - \boldsymbol{\delta})'\mathbf{S}_d^{-1}(\bar{\mathbf{D}} - \boldsymbol{\delta}) \leq \frac{(n-1)p}{n(n-p)}\mathcal{F}_{p,n-p}(\alpha)$$

erfüllt.

Konfidenzregionen

Die simultanen Konfidenz Intervalle für die einzelnen Mittelwert-differenzen, $\delta_i, \quad i = 1, \ldots, p$ sind gegeben durch:

$$\bar{D}_i \pm \sqrt{\frac{(n-1)p}{n-p} \mathcal{F}_{p,n-p}(\alpha)} \sqrt{\frac{s^2_{D_i}}{n}}$$

und, die Bonferroni $100(1-\alpha)$% simultanen Konfidenz Intervalle sind gegeben durch:

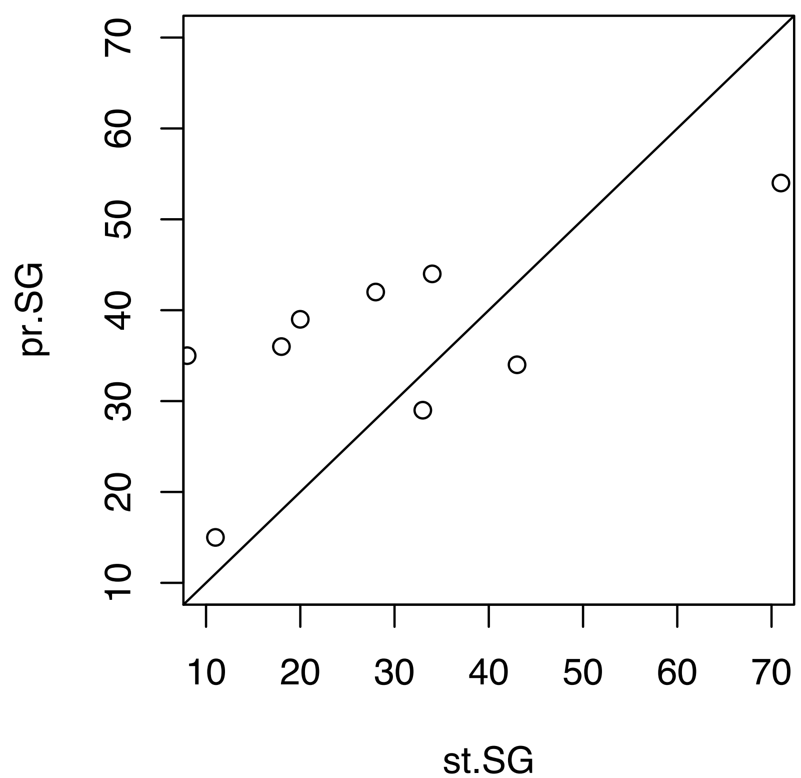
$$\bar{D}_i \pm t_{n-1}(\frac{\alpha}{2p}) \sqrt{\frac{s^2_{D_i}}{n}}$$

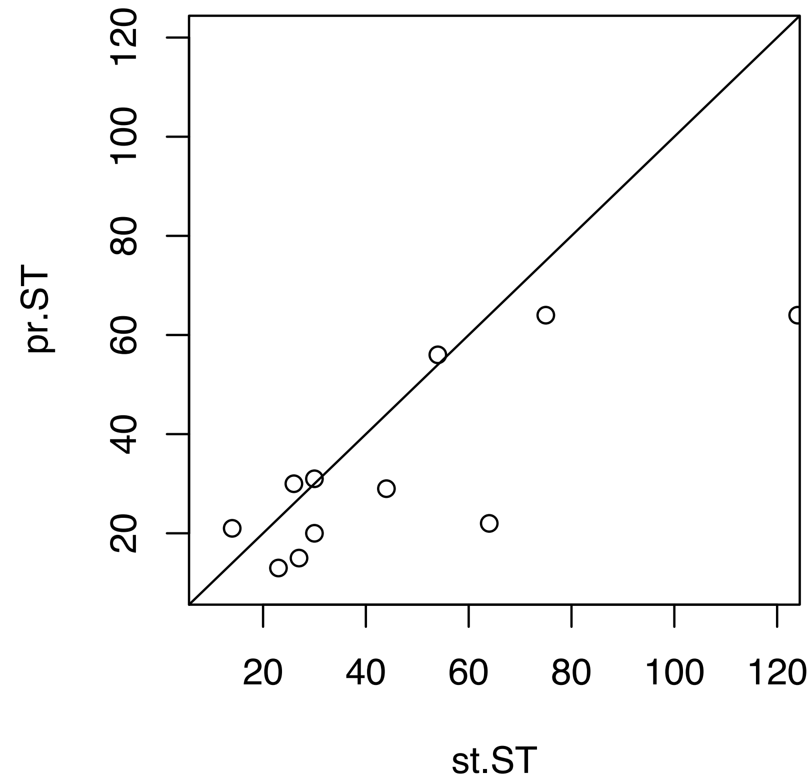
Beispiel: Wasserproben

- Gemessen wurden
  - Sauerstoffgehalt (SG)
  - Schwebeteilchen (ST)
- 2 Labore
  - staatlich
  - privat
- Frage:
  Sind die Messergebnisse als Ganzes unterschiedlich?

Probe	SG	ST	SG	ST	SG	ST
1	6	27	25	15	-19	12
2	6	23	28	13	-22	10
3	18	64	36	22	-18	42
4	8	44	35	29	-27	15
5	11	30	15	31	-4	-1
6	34	75	44	64	-10	11
7	28	26	42	30	-14	-4
8	71	124	54	64	17	60
9	43	54	34	56	9	-2
10	33	30	29	20	4	10
11	20	14	39	21	-19	-7

...





$$\bar{\mathbf{D}} = \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} \quad \mathbf{S}_d = \begin{bmatrix} 199.3 & 88.3 \\ 88.3 & 418.6 \end{bmatrix}$$

$$T^2 = 11[-9.36 \;\; 13.27] \begin{bmatrix} 0.0055 & -0.0012 \\ -0.0012 & 0.0026 \end{bmatrix} \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} = 13.6$$

$2 \times 10/9\mathcal{F}_{2,9}(0.05) = 9.47 \Rightarrow$ Verwerfe H_o.

MANOVA

Situation:

- **Daten**
  - g Stichproben der Größe $n_1, \ldots n_g$
  - p Variablen.
- **Grundgesamtheit**
  - g Grundgesamtheiten
  - verschiedene Mittelwerte $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g)$
  - gleiche Covarianz (Σ)
  - multivariat normalverteilt.

Mögliche Anwendungen

1. Statistische Versuchsplanung mit multipler Response
2. Klassifikationsprobleme

Rückblick: univariate ANOVA

Zufallsstichproben $X_{l1}, X_{l2}, \ldots, X_{ln_l}$ aus $N(\mu_l, \sigma^2)$, $l = 1, \ldots, g$. Frage: Sind die Gruppenmittelwerte alle gleich, d.h. $\mu_1 = \mu_2 = \ldots = \mu_g$. Daraus folgt die Parametrisierung:

$$X_{lj} = \mu \quad + \quad \tau_l \quad + \quad e_{lj}$$

mit der Nebenbedingung $\sum_{l=1}^{g} n_l \tau_l = 0$, und der Nullhypothese:

$$H_o : \tau_1 = \tau_2 = \ldots = \tau_g = 0$$

Das Modell wird angepasst, indem jede Beobachtung in ihre Komponenten zerlegt wird:

$$X_{lj} = \bar{X} \ + \ (\bar{X}_l - \bar{X}) \ + \ (X_{lj} - \bar{X}_l), \ j = 1, \ldots, n_l; \ l = 1, \ldots, g$$

und die Quadratsummen für jede Komponente berechnet werden:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X})^2 = \sum_{l=1}^{g} n_l (\bar{X}_l - \bar{X})^2 + \sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2$$

...

ANOVA Tabelle

Quelle	QS	Fg
Modell (zwischen)	SS_Z	$g-1$
Residuen (innerhalb)	SS_I	$\sum_{l=1}^{g} n_l - g$
Gesamt	SS_G	$\sum_{l=1}^{g} n_l - 1$

Dann testen wir die Hypothese

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0 \;\text{ gegen }\; H_A : \text{nicht alle gleich } 0,$$

mittels

$$F = \frac{SS_Z/(g-1)}{SS_I/(\sum_{l=1}^{g} n_l - g)} \sim \mathcal{F}_{g-1, \sum n_l - g}(\alpha)$$

Multivariate ANOVA

Als Modellgleichung ergibt sich

$$\mathbf{X}_{lj} = \boldsymbol{\mu} \quad + \quad \boldsymbol{\tau}_l \quad + \quad \mathbf{e}_{lj}$$

wobei $\mathbf{X}_{lj}, \mathbf{e}_{lj}$ $n_l \times p$ Matrizen sind, und $\boldsymbol{\mu}, \boldsymbol{\tau}_l$ p-dim. Vektoren.
Die Datenzerlegung ist dann:

$$\mathbf{X}_{lj} = \bar{\mathbf{X}} \quad + \quad (\bar{\mathbf{X}}_l - \bar{\mathbf{X}}) \quad + \quad (\mathbf{X}_{lj} - \bar{\mathbf{X}}_l), \quad j = 1, \ldots, n_l;\ l = 1, \ldots, g$$

Die Quadratsummenzerlegung ist:

$$\sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{X}_{lj} - \bar{\mathbf{X}})(\mathbf{X}_{lj} - \bar{\mathbf{X}})' \quad = \quad \overbrace{\sum_{l=1}^{g} n_l (\bar{\mathbf{X}}_l - \bar{\mathbf{X}})(\bar{\mathbf{X}}_l - \bar{\mathbf{X}})'}^{\mathbf{B}_{p \times p}}$$

$$+ \quad \underbrace{\sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{X}_{lj} - \bar{\mathbf{X}}_l)(\mathbf{X}_{lj} - \bar{\mathbf{X}}_l)'}_{\mathbf{W}_{p \times p}}$$

MANOVA Tabelle und Test

- **MANOVA Tabelle**

Quelle	QS	Fg
Modell (zwischen)	**B**	$g-1$
Residuen (innerhalb)	**W**	$\sum_{l=1}^{g} n_l - g$
Gesamt	$\mathbf{B}+\mathbf{W}$	$\sum_{l=1}^{g} n_l - 1$

- **Hypothesentest**

$$H_o : \boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_l = \mathbf{0}$$

Problem: Hier werden nun Matrizen und keine einzelnen Werte verglichen. Diese Matizen entsprechen p–d Varianz Ellipsen. Als mögliche Teststatistik wird meist Wilks' lambda benutzt:

$$\Lambda^* = \frac{|W|}{|B+W|}$$

Wenn Λ^* zu klein ist, wird H_0 verworfen.

Verteilung von Wilks Λ

# Vars	# Gruppen	Verteilung von Λ^*
1	≥ 2	$\left(\frac{\sum n_l - g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim \mathcal{F}_{g-1,\sum n_l - g}$
2	≥ 2	$\left(\frac{\sum n_l - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim \mathcal{F}_{2(g-1),2(\sum n_l - g - 1)}$
≥ 1	2	$\left(\frac{\sum n_l - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim \mathcal{F}_{p,\sum n_l - p - 1}$
≥ 1	3	$\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim \mathcal{F}_{2p,2(\sum n_l - p - 2)}$

Wenn $\sum n_l = n$ groß ist, kann H_0 auf einem Signifikanz Niveau von α verworfen werden, falls gilt:

$$-(n-1-(p+g)/2)\ln\left(\frac{|W|}{|B+W|}\right) > \chi^2_{p(g-1)}(\alpha)$$

Wilks Λ über Eigenwerte

Λ^* lässt sich auch alternativ darstellen über

$$\Lambda^* = \prod_{i=1}^{\min(p,g-1)} \frac{1}{1+\lambda_i}$$

wobei $\lambda_i;\;\; i = 1, \ldots, \min(p, g-1)$ die Eigenwerte von $W^{-1}B$ sind.

Es gilt:

- Da Wilks' lambda auf Determinaten der Matrizen **W** und **B** basiert, werden im wesentlichen Volumina von Ellipsen verglichen.
- Eine andere Methode diese Volumina zu vergleichen ist über die Längen der Hauptachsen, d.h. Eigenwerte.

MANOVA: Beispiel

Daten $\mathbf{X}$, mit $p = 2$ Dimensionen, $g = 3$ Gruppen der Größe $n_1 = 3, n_2 = 2, n_3 = 3$.

$$\mathbf{X} = \begin{pmatrix} \begin{bmatrix} 9 \\ 3 \end{bmatrix} & \begin{bmatrix} 6 \\ 2 \end{bmatrix} & \begin{bmatrix} 9 \\ 7 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 4 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \\ \begin{bmatrix} 3 \\ 8 \end{bmatrix} & \begin{bmatrix} 1 \\ 9 \end{bmatrix} & \begin{bmatrix} 2 \\ 7 \end{bmatrix} \end{pmatrix}$$

und $\bar{\mathbf{X}} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}, \bar{\mathbf{X}}_1 = \begin{bmatrix} 8 \\ 4 \end{bmatrix}, \bar{\mathbf{X}}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \bar{\mathbf{X}}_3 = \begin{bmatrix} 2 \\ 8 \end{bmatrix}$.

$$\begin{aligned} \mathbf{B} &= \sum_{l=1}^{g} n_l (\bar{\mathbf{X}}_l - \bar{\mathbf{X}})(\bar{\mathbf{X}}_l - \bar{\mathbf{X}})' \\ &= 3 \cdot \left(\begin{bmatrix} 8 \\ 4 \end{bmatrix} - \begin{bmatrix} 4 \\ 5 \end{bmatrix} \right) \left(\begin{bmatrix} 8 \\ 4 \end{bmatrix} - \begin{bmatrix} 4 \\ 5 \end{bmatrix} \right)' + \ldots \\ &= 3 \begin{bmatrix} 4 \\ -1 \end{bmatrix} \begin{bmatrix} 4 & -1 \end{bmatrix} + \ldots \\ &= 3 \begin{pmatrix} 16 & -4 \\ -4 & 1 \end{pmatrix} + \ldots + 3 \begin{pmatrix} 4 & -6 \\ -6 & 9 \end{pmatrix} = \begin{pmatrix} 78 & -12 \\ -12 & 48 \end{pmatrix} \end{aligned}$$

Beispiel (cont.)

$$
\begin{aligned}
\mathbf{W} &= \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{X}_{lj} - \bar{\mathbf{X}}_l)(\mathbf{X}_{lj} - \bar{\mathbf{X}}_l)' \\
&= \left(\left[\begin{smallmatrix}9\\3\end{smallmatrix}\right] - \left[\begin{smallmatrix}8\\4\end{smallmatrix}\right]\right)\left(\left[\begin{smallmatrix}9\\3\end{smallmatrix}\right] - \left[\begin{smallmatrix}8\\4\end{smallmatrix}\right]\right)' + \ldots + \left(\left[\begin{smallmatrix}3\\8\end{smallmatrix}\right] - \left[\begin{smallmatrix}2\\8\end{smallmatrix}\right]\right)\left(\left[\begin{smallmatrix}3\\8\end{smallmatrix}\right] - \left[\begin{smallmatrix}2\\8\end{smallmatrix}\right]\right)' + \ldots \\
&= \left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]\left[1 \quad -1\right] + \ldots \qquad + \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]\left[1 \quad 0\right] + \ldots \\
&= \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \ldots \qquad + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \ldots \\
&= \begin{pmatrix} 6 & 6 \\ 6 & 14 \end{pmatrix} + \ldots + \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 10 & 1 \\ 1 & 24 \end{pmatrix}
\end{aligned}
$$

$$
\Lambda^* = \frac{|W|}{|B+W|} = \frac{10 \cdot 24 - 1 \cdot 1}{88 \cdot 72 - (-11 \cdot -11)} = 0.0385
$$

verwerfe H_0 bei einem Niveau von $\alpha = 0.01$ wegen

$$
\left(\frac{\sum n_l - g - 1}{g - 1}\right)\left(\frac{1 - \sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) = 8.19 > 7.01 = \mathcal{F}_{4,8}(0.01)
$$