Varianzvergleiche bei normalverteilten Zufallsvariablen

- *Nächste Anwendung:* Vergleich der Varianzen σ_A^2 und σ_B^2 zweier *normalverteilter* Zufallsvariablen $Y^A \sim N(\mu_A, \sigma_A^2)$ und $Y^B \sim N(\mu_B, \sigma_B^2)$ auf Grundlage zweier unabhängiger einfacher Stichproben $X_1^A, \ldots, X_{n_A}^A$ vom Umfang n_A zu Y^A und $X_1^B, \ldots, X_{n_B}^B$ vom Umfang n_B zu Y^B.
- *Idee:* Vergleich auf Grundlage der erwartungstreuen Schätzfunktionen

$$S_{Y^A}^2 = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} (X_i^A - \overline{X^A})^2 = \frac{1}{n_A - 1} \left(\left(\sum_{i=1}^{n_A} (X_i^A)^2 \right) - n_A \overline{X^A}^2 \right)$$

$$\text{bzw. } S_{Y^B}^2 = \frac{1}{n_B - 1} \sum_{i=1}^{n_B} (X_i^B - \overline{X^B})^2 = \frac{1}{n_B - 1} \left(\left(\sum_{i=1}^{n_B} (X_i^B)^2 \right) - n_B \overline{X^B}^2 \right)$$

  für die Varianz von Y^A bzw. die Varianz von Y^B.
- Es gilt $\frac{(n_A-1)\cdot S_{Y^A}^2}{\sigma_A^2} \sim \chi^2(n_A - 1)$ unabhängig von $\frac{(n_B-1)\cdot S_{Y^B}^2}{\sigma_B^2} \sim \chi^2(n_B - 1)$.
- Geeignete Testgröße lässt sich aus (standardisiertem) Verhältnis von $\frac{(n_A-1)\cdot S_{Y^A}^2}{\sigma_A^2}$ und $\frac{(n_B-1)\cdot S_{Y^B}^2}{\sigma_B^2}$ herleiten.

Die Familie der $F(m, n)$-Verteilungen

- Sind χ_m^2 und χ_n^2 stochastisch unabhängige, mit m bzw. n Freiheitsgraden χ^2-verteilte Zufallsvariablen, so heißt die Verteilung der Zufallsvariablen

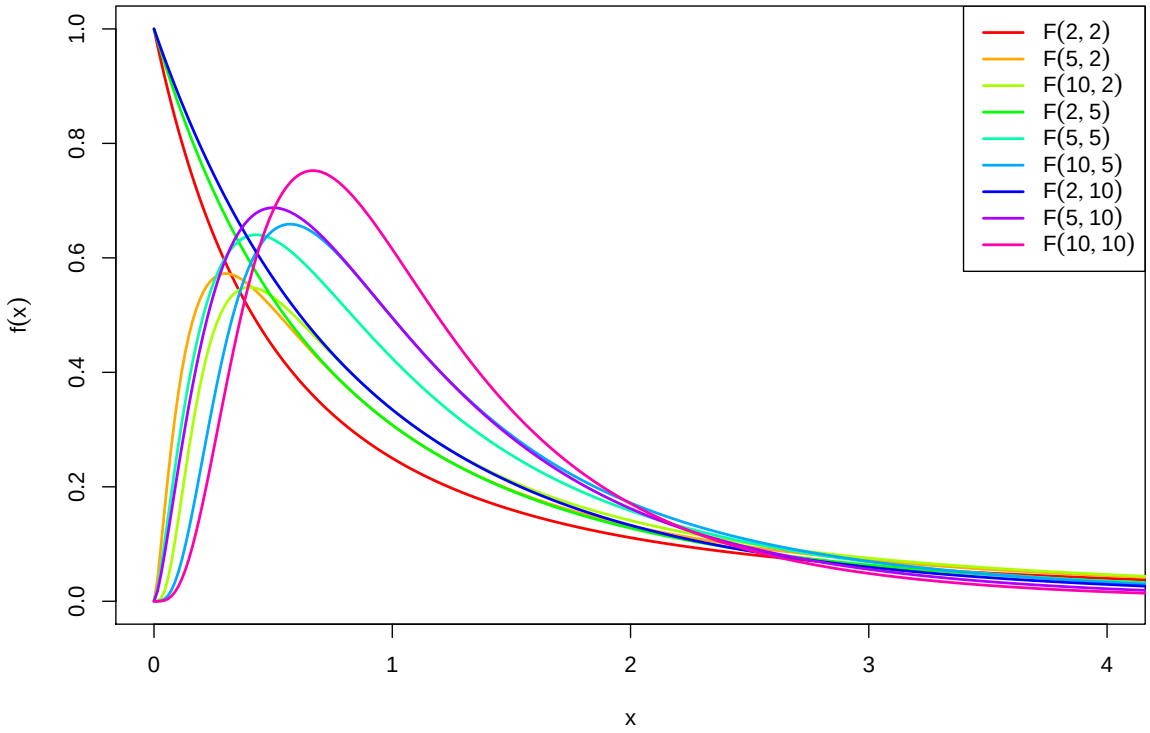
$$F_n^m := \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}} = \frac{\chi_m^2}{\chi_n^2} \cdot \frac{n}{m}$$

  F-**Verteilung mit** m **Zähler- und** n **Nennerfreiheitsgraden**, in Zeichen $F_n^m \sim F(m, n)$.
- Offensichtlich können $F(m, n)$-verteilte Zufallsvariablen nur nichtnegative Werte annehmen, der Träger ist also $[0, \infty)$.
- Für $n > 2$ gilt $\mathrm{E}(F_n^m) = \frac{n}{n-2}$.
- Als Abkürzung für α-Quantile der $F(m, n)$-Verteilung verwenden wir (wie üblich) $F_{m,n;\alpha}$.
- Für die Quantile der $F(m, n)$-Verteilungen gilt der folgende Zusammenhang:

$$F_{m,n;\alpha} = \frac{1}{F_{n,m;1-\alpha}}$$

Grafische Darstellung einiger $F(m, n)$-Verteilungen

für $m, n \in \{2, 5, 10\}$



Varianzvergleiche (Fortsetzung)

- Eine $F(n_A - 1, n_B - 1)$-verteilte Zufallsvariable erhält man also in der Anwendungssituation der Varianzvergleiche durch das Verhältnis

$$\frac{\frac{(n_A-1)\cdot S_{YA}^2}{\sigma_A^2}}{\frac{(n_B-1)\cdot S_{YB}^2}{\sigma_B^2}} \cdot \frac{n_B - 1}{n_A - 1} = \frac{\frac{S_{YA}^2}{\sigma_A^2}}{\frac{S_{YB}^2}{\sigma_B^2}},$$

  das allerdings von den (unbekannten!) Varianzen σ_A^2 und σ_B^2 abhängt.
- Gilt jedoch $\sigma_A^2 = \sigma_B^2$, so hat auch das Verhältnis

$$F := \frac{S_{YA}^2}{S_{YB}^2}$$

  eine $F(n_A - 1, n_B - 1)$-Verteilung und ist somit als Testgröße geeignet, wenn unter H_0 (eventuell im Grenzfall) $\sigma_A^2 = \sigma_B^2$ angenommen wird.
- Offensichtlich sprechen große Werte von F eher für $\sigma_A^2 > \sigma_B^2$, kleine eher für $\sigma_A^2 < \sigma_B^2$, Verhältnisse in der Nähe von 1 für $\sigma_A^2 = \sigma_B^2$.

- Da die Klasse der F-Verteilungen von 2 Verteilungsparametern abhängt, ist es nicht mehr möglich, α-Quantile für verschiedene Freiheitsgradkombinationen und verschiedene α darzustellen.
- In Formelsammlung: Tabellen (nur) mit 0.95-Quantilen für verschiedene Kombinationen von m und n für $F(m, n)$-Verteilungen verfügbar.
- Bei linksseitigen Tests (zum Niveau $\alpha = 0.05$) und zweiseitigen Tests (zum Niveau $\alpha = 0.10$) muss also regelmäßig die „Symmetrieeigenschaft“

$$F_{m,n;\alpha} = \frac{1}{F_{n,m;1-\alpha}}$$

verwendet werden, um auch 0.05-Quantile bestimmen zu können.
- Der resultierende Test ist insbesondere zur Überprüfung der Anwendungsvoraussetzungen für den 2-Stichproben-t-Test hilfreich.

Wichtig!

Die Normalverteilungsannahme für Y^A und Y^B ist wesentlich. Ist diese (deutlich) verletzt, ist auch eine näherungsweise Verwendung des Tests nicht mehr angebracht.

0.95-Quantile der $F(m, n)$-Verteilungen $F_{m,n;0.95}$

$n \backslash m$	1	2	3	4	5	6	7	8
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032
150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001

Zusammenfassung: F-Test zum Vergleich der Varianzen

zweier normalverteilter Zufallsvariablen

Anwendungs-voraussetzungen	exakt: $Y^A \sim N(\mu_A, \sigma_A^2)$, $Y^B \sim N(\mu_B, \sigma_B^2)$, $\mu_A, \mu_B, \sigma_A^2, \sigma_B^2$ unbek. $X_1^A, \ldots, X_{n_A}^A$ einfache Stichprobe zu Y^A, unabhängig von einfacher Stichprobe $X_1^B, \ldots, X_{n_B}^B$ zu Y^B.		
Nullhypothese Gegenhypothese	$H_0 : \sigma_A^2 = \sigma_B^2$ $H_1 : \sigma_A^2 \neq \sigma_B^2$	$H_0 : \sigma_A^2 \leq \sigma_B^2$ $H_1 : \sigma_A^2 > \sigma_B^2$	$H_0 : \sigma_A^2 \geq \sigma_B^2$ $H_1 : \sigma_A^2 < \sigma_B^2$
Teststatistik Verteilung (H_0)	$F = \dfrac{S_{Y^A}^2}{S_{Y^B}^2}$ F unter H_0 für $\sigma_A^2 = \sigma_B^2$ $F(n_A - 1, n_B - 1)$-verteilt		
Benötigte Größen	$\overline{X^A} = \frac{1}{n_A} \sum_{i=1}^{n_A} X_i^A$, $\overline{X^B} = \frac{1}{n_B} \sum_{i=1}^{n_B} X_i^B$, $S_{Y^A}^2 = \frac{1}{n_A - 1} \sum_{i=1}^{n_A} (X_i^A - \overline{X^A})^2 = \frac{1}{n_A - 1} \left(\left(\sum_{i=1}^{n_A} (X_i^A)^2 \right) - n_A \overline{X^A}^2 \right)$ $S_{Y^B}^2 = \frac{1}{n_B - 1} \sum_{i=1}^{n_B} (X_i^B - \overline{X^B})^2 = \frac{1}{n_B - 1} \left(\left(\sum_{i=1}^{n_B} (X_i^B)^2 \right) - n_B \overline{X^B}^2 \right)$		
Kritischer Bereich zum Niveau α	$[0, F_{n_A-1, n_B-1; \frac{\alpha}{2}})$ $\cup (F_{n_A-1, n_B-1; 1-\frac{\alpha}{2}}, \infty)$	$(F_{n_A-1, n_B-1; 1-\alpha}, \infty)$	$[0, F_{n_A-1, n_B-1; \alpha})$
p-Wert	$2 \cdot \min \{ F_{F(n_A-1, n_B-1)}(F),$ $1 - F_{F(n_A-1, n_B-1)}(F) \}$	$1 - F_{F(n_A-1, n_B-1)}(F)$	$F_{F(n_A-1, n_B-1)}(F)$

Beispiel: Präzision von 2 Abfüllanlagen

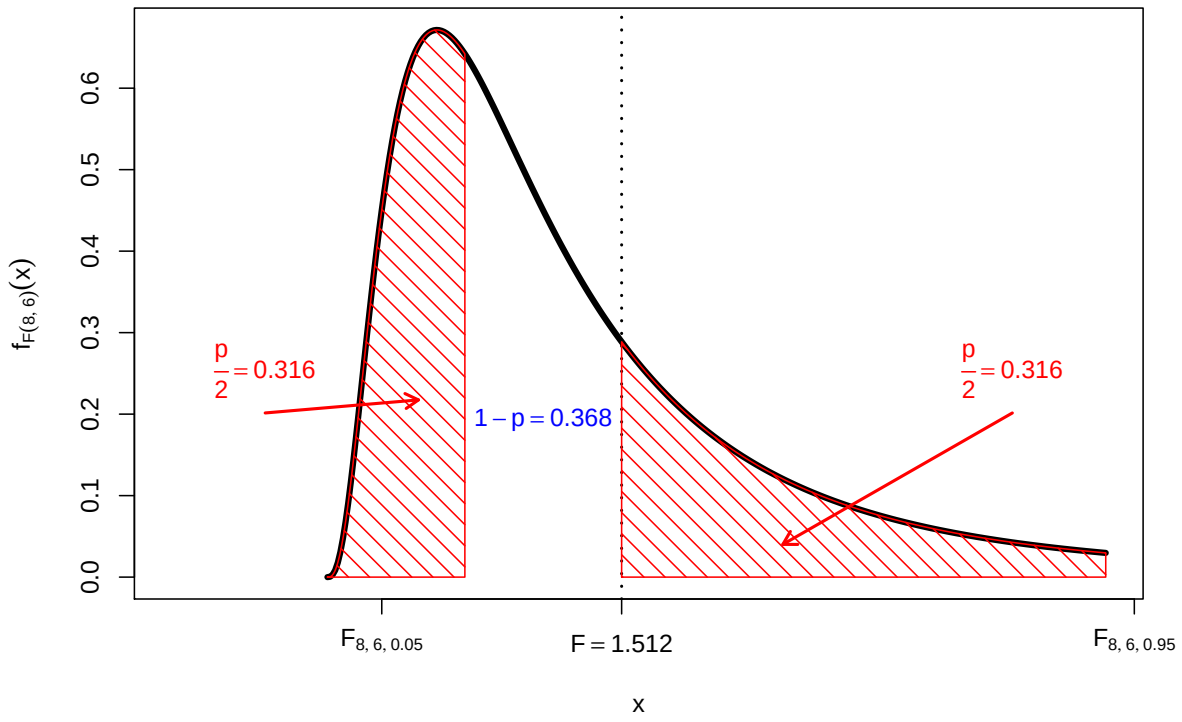
- Untersuchungsgegenstand: Entscheidung, ob Varianz der Abfüllmenge von zwei Abfüllanlagen übereinstimmt oder nicht.
- Annahmen: Abfüllmengen Y^A und Y^B jeweils normalverteilt.
- Unabhängige einfache Stichproben vom Umfang $n_A = 9$ zu Y^A und vom Umfang $n_B = 7$ zu Y^B liefern realisierte Varianzschätzungen $s_{Y^A}^2 = 16.22$ sowie $s_{Y^B}^2 = 10.724$.
- Gewünschtes Signifikanzniveau $\alpha = 0.10$.

Geeigneter Test: F-**Test für die Varianzen normalverteilter Zufallsvariablen**

1. **Hypothesen:** $H_0 : \sigma_A^2 = \sigma_B^2$ gegen $H_1 : \sigma_A^2 \neq \sigma_B^2$
2. **Teststatistik:** $F = \dfrac{S_{Y^A}^2}{S_{Y^B}^2}$ ist unter H_0 $F(n_A - 1, n_B - 1)$-verteilt.
3. **Kritischer Bereich zum Niveau** $\alpha = 0.10$**:** Mit $F_{8,6;0.05} = 1/F_{6,8;0.95} = 1/3.581 = 0.279$:
   $K = [0, F_{n_A-1, n_B-1; \frac{\alpha}{2}}) \cup (F_{n_A-1, n_B-1; 1-\frac{\alpha}{2}}, +\infty) =$
   $[0, F_{8,6;0.05}) \cup (F_{8,6;0.95}, +\infty) = [0, 0.279) \cup (4.147, +\infty)$
4. **Berechnung der realisierten Teststatistik:** $F = \dfrac{s_{Y^A}^2}{s_{Y^B}^2} = \dfrac{16.22}{10.724} = 1.512$
5. **Entscheidung:** $F \notin K \quad \Rightarrow \quad H_0$ wird nicht abgelehnt!

Beispiel: p-Wert bei F-Test für Varianzen (Grafik)

Abfüllanlagenbeispiel, realisierte Teststatistik $F = 1.512$, p-Wert: 0.632



Mittelwertvergleiche bei $k > 2$ unabhängigen Stichproben

- *Nächste Anwendung:* Vergleich der Mittelwerte von $k > 2$ normalverteilten Zufallsvariablen $Y_1 \sim N(\mu_1, \sigma^2), \ldots, Y_k \sim N(\mu_k, \sigma^2)$ mit *übereinstimmender* Varianz σ^2.
- Es soll eine Entscheidung getroffen werden zwischen

  $$H_0 : \mu_1 = \mu_j \text{ für alle } j \qquad \text{und} \qquad H_1 : \mu_1 \neq \mu_j \text{ für (mindestens) ein } j$$

  auf Basis von k unabhängigen einfachen Stichproben

  $$X_{1,1}, \ldots, X_{1,n_1}, \qquad \ldots, \qquad X_{k,1}, \ldots, X_{k,n_k}$$

  mit Stichprobenumfängen $n_1, \ldots, n_k$ (Gesamtumfang: $n := \sum_{j=1}^{k} n_j$).
- Häufiger Anwendungsfall: Untersuchung des Einflusses *einer* nominalskalierten Variablen (mit mehr als 2 Ausprägungen) auf eine (kardinalskalierte) Zufallsvariable, z.B.
  - Einfluss verschiedener Düngemittel auf Ernteertrag,
  - Einfluss verschiedener Behandlungsmethoden auf Behandlungserfolg,
  - Einfluss der Zugehörigkeit zu bestimmten Gruppen (z.B. Schulklassen).
- Beteiligte nominalskalierte Einflussvariable wird dann meist **Faktor** genannt, die einzelnen Ausprägungen **Faktorstufen**.
- Geeignetes statistisches Untersuchungswerkzeug: **Einfache Varianzanalyse**

Einfache Varianzanalyse

- Idee der einfachen („einfaktoriellen“) Varianzanalyse:
  Vergleich der Streuung der **Stufenmittel** (auch „Gruppenmittel“)

$$\overline{X}_1 := \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}, \qquad \ldots, \qquad \overline{X}_k := \frac{1}{n_k}\sum_{i=1}^{n_k} X_{k,i}$$

  um das Gesamtmittel

$$\overline{X} := \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{j,i} = \frac{1}{n}\sum_{j=1}^{k} n_j \cdot \overline{X}_j$$

  mit den Streuungen der Beobachtungswerte $X_{j,i}$ um die jeweiligen Stufenmittel $\overline{X}_j$ innerhalb der j-ten Stufe.
- Sind die Erwartungswerte in allen Stufen gleich (gilt also H_0), so ist die Streuung der Stufenmittel vom Gesamtmittel im Vergleich zur Streuung der Beobachtungswerte um die jeweiligen Stufenmittel *tendenziell* nicht so groß wie es bei Abweichungen der Erwartungswerte für die einzelnen Faktorstufen der Fall wäre.

- Messung der Streuung der Stufenmittel vom Gesamtmittel durch Größe SB (**„Squares Between“**) als (gew.) Summe der quadrierten Abweichungen:

$$SB = \sum_{j=1}^{k} n_j \cdot (\overline{X}_j - \overline{X})^2 = n_1 \cdot (\overline{X}_1 - \overline{X})^2 + \ldots + n_k \cdot (\overline{X}_k - \overline{X})^2$$

- Messung der (Summe der) Streuung(en) der Beobachtungswerte um die Stufenmittel durch Größe SW (**„Squares Within“**) als (Summe der) Summe der quadrierten Abweichungen:

$$SW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{j,i} - \overline{X}_j)^2 = \sum_{i=1}^{n_1}(X_{1,i} - \overline{X}_1)^2 + \ldots + \sum_{i=1}^{n_k}(X_{k,i} - \overline{X}_k)^2$$

- Man kann zeigen:
  - Für die Gesamtsumme SS (**„Sum of Squares“**) der quadrierten Abweichungen der Beobachtungswerte vom Gesamtmittelwert mit

$$SS = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{j,i} - \overline{X})^2 = \sum_{i=1}^{n_1}(X_{1,i} - \overline{X})^2 + \ldots + \sum_{i=1}^{n_k}(X_{k,i} - \overline{X})^2$$

    gilt die **Streuungszerlegung** $SS = SB + SW$.
  - Mit den getroffenen Annahmen sind $\frac{SB}{\sigma^2}$ bzw. $\frac{SW}{\sigma^2}$ unter H_0 unabhängig $\chi^2(k-1)$- bzw. $\chi^2(n-k)$-verteilt $\rightsquigarrow$ Konstruktion geeigneter Teststatistik.

- Da $\frac{SB}{\sigma^2}$ bzw. $\frac{SW}{\sigma^2}$ unter H_0 unabhängig $\chi^2(k-1)$- bzw. $\chi^2(n-k)$-verteilt sind, ist der Quotient

$$F := \frac{\frac{SB}{\sigma^2}}{\frac{SW}{\sigma^2}} \cdot \frac{n-k}{k-1} = \frac{SB}{SW} \cdot \frac{n-k}{k-1} = \frac{\frac{SB}{k-1}}{\frac{SW}{n-k}} = \frac{SB/(k-1)}{SW/(n-k)}$$

  unter H_0 also $F(k-1, n-k)$-verteilt.
- Zur Konstruktion des kritischen Bereichs ist zu beachten, dass **große** Quotienten F **gegen** die Nullhypothese sprechen, da in diesem Fall die Abweichung der Stufenmittel vom Gesamtmittel SB verhältnismäßig groß ist.
- Als kritischer Bereich zum Signifikanzniveau α ergibt sich $K = (F_{k-1,n-k;1-\alpha}, \infty)$
- Die Bezeichnung „Varianzanalyse“ erklärt sich dadurch, dass (zur Entscheidungsfindung über die Gleichheit der Erwartungswerte!) die Stichprobenvarianzen $SB/(k-1)$ und $SW/(n-k)$ untersucht werden.
- Die Varianzanalyse kann als näherungsweiser Test auch angewendet werden, wenn die Normalverteilungsannahme verletzt ist.
- Das Vorliegen gleicher Varianzen in allen Faktorstufen („Varianzhomogenität“) muss jedoch (auch für vernünftige näherungsweise Verwendung) gewährleistet sein! Überprüfung z.B. mit „Levene-Test“ oder „Bartlett-Test“ (hier nicht besprochen).

Zusammenfassung: Einfache Varianzanalyse

Anwendungs-voraussetzungen	exakt: $Y_j \sim N(\mu_j, \sigma^2)$ für $j \in \{1, \ldots, k\}$ approximativ: Y_j beliebig verteilt mit $\mathrm{E}(Y_j) = \mu_j$, $\mathrm{Var}(Y_j) = \sigma^2$ k unabhängige einfache Stichproben $X_{j,1}, \ldots, X_{j,n_j}$ vom Umfang n_j zu Y_j für $j \in \{1, \ldots, k\}$, $n = \sum_{j=1}^{k} n_j$
Nullhypothese Gegenhypothese	$H_0 : \mu_1 = \mu_j$ für alle $j \in \{2, \ldots, k\}$ $H_1 : \mu_1 \neq \mu_j$ für (mindestens) ein $j \in \{2, \ldots, k\}$
Teststatistik Verteilung (H_0)	$F = \frac{SB/(k-1)}{SW/(n-k)}$ F ist (approx.) $F(k-1, n-k)$-verteilt, falls $\mu_1 = \ldots = \mu_k$
Benötigte Größen	$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{j,i}$ für $j \in \{1, \ldots, k\}$, $\overline{x} = \frac{1}{n} \sum_{j=1}^{k} n_j \cdot \overline{x}_j$, $SB = \sum_{j=1}^{k} n_j \cdot (\overline{x}_j - \overline{x})^2$, $SW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{j,i} - \overline{x}_j)^2$
Kritischer Bereich zum Niveau α	$(F_{k-1,n-k;1-\alpha}, \infty)$
p-Wert	$1 - F_{F(k-1,n-k)}(F)$

- Alternative Berechnungsmöglichkeiten mit „Verschiebungssatz“
  - für Realisation von SB:
$$SB = \sum_{j=1}^{k} n_j \cdot (\overline{x}_j - \overline{x})^2 = \left(\sum_{j=1}^{k} n_j \overline{x}_j^2 \right) - n\overline{x}^2$$
  - für Realisation von SW:
$$SW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{j,i} - \overline{x}_j)^2 = \sum_{j=1}^{k} \left(\left(\sum_{i=1}^{n_j} x_{j,i}^2 \right) - n_j \overline{x}_j^2 \right)$$
- Liegen für $j \in \{1, \ldots, k\}$ die Stichprobenvarianzen
$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{j,i} - \overline{X}_j)^2$$
bzw. deren Realisationen s_j^2 für die k (Einzel-)Stichproben
$$X_{1,1}, \ldots, X_{1,n_1}, \qquad \ldots \qquad X_{k,1}, \ldots, X_{k,n_k}$$
vor, so erhält man die Realisation von SW offensichtlich auch durch
$$SW = \sum_{j=1}^{k} (n_j - 1) \cdot s_j^2 \ .$$

Beispiel: Bedienungszeiten an $k = 3$ Servicepunkten

- Untersuchungsgegenstand: Stimmen die mittleren Bedienungszeiten μ_1, μ_2, μ_3 an 3 verschiedenen Servicepunkten überein oder nicht?
- Annahme: Bedienungszeiten Y_1, Y_2, Y_3 an den 3 Servicestationen sind jeweils normalverteilt mit $\mathrm{E}(Y_j) = \mu_j$ und **identischer** (unbekannter) Varianz $\mathrm{Var}(Y_j) = \sigma^2$.
- Es liegen Realisationen von 3 unabhängigen einfache Stichproben zu den Zufallsvariablen Y_1, Y_2, Y_3 mit den Stichprobenumfängen $n_1 = 40, n_2 = 33, n_3 = 30$ wie folgt vor:

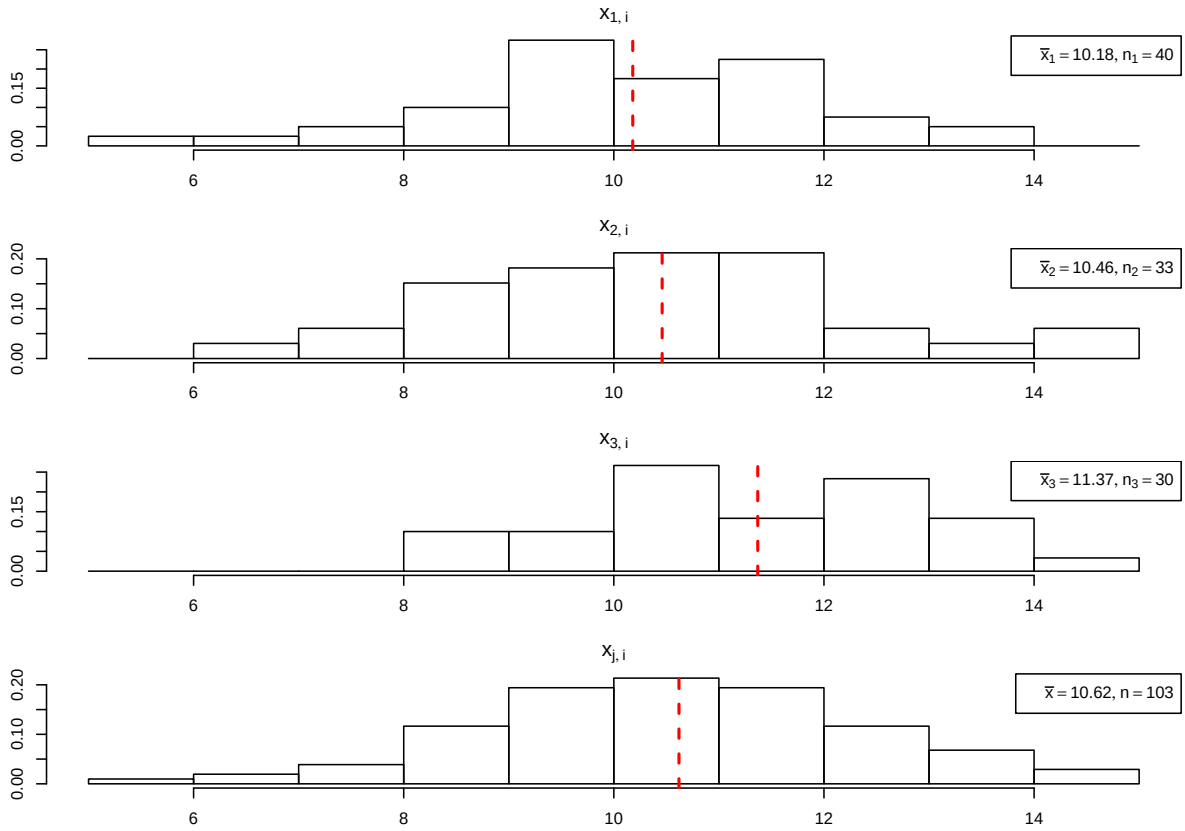
j (Servicepunkt)	n_j	$\overline{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{j,i}$	$\sum_{i=1}^{n_j} x_{j,i}^2$
1	40	10.18	4271.59
2	33	10.46	3730.53
3	30	11.37	3959.03

(Daten simuliert mit $\mu_1 = 10, \mu_2 = 10, \mu_3 = 11.5$, $\sigma^2 = 2^2$)

- Gewünschtes Signifikanzniveau: $\alpha = 0.05$

Geeignetes Verfahren: **Varianzanalyse**

Grafische Darstellung der Stichprobeninformation



1. **Hypothesen:**

   $H_0 : \mu_1 = \mu_2 = \mu_3 \qquad H_1 : \mu_1 \neq \mu_j$ für mindestens ein j

2. **Teststatistik:**

   $F = \dfrac{SB/(k-1)}{SW/(n-k)}$ ist unter H_0 $F(k-1, n-k)$-verteilt.

3. **Kritischer Bereich zum Niveau $\alpha = 0.05$:**

   $K = (F_{k-1;n-k;1-\alpha}, +\infty) = (F_{2;100;0.95}, +\infty) = (3.087, +\infty)$

4. **Berechnung der realisierten Teststatistik:**

   Mit $\overline{x}_1 = 10.18, \overline{x}_2 = 10.46, \overline{x}_3 = 11.37$ erhält man

$$\overline{x} = \frac{1}{103} \sum_{j=1}^{3} n_j \cdot \overline{x}_j = \frac{1}{103} \left(40 \cdot 10.18 + 33 \cdot 10.46 + 30 \cdot 11.37\right) = 10.62$$

   und damit

$$\begin{aligned} SB &= \sum_{j=1}^{3} n_j \left(\overline{x}_j - \overline{x}\right)^2 = n_1 \left(\overline{x}_1 - \overline{x}\right)^2 + n_2 \left(\overline{x}_2 - \overline{x}\right)^2 + n_3 \left(\overline{x}_3 - \overline{x}\right)^2 \\ &= 40(10.18 - 10.62)^2 + 33(10.46 - 10.62)^2 + 30(11.37 - 10.62)^2 \\ &= 25.46 \ . \end{aligned}$$

4. (Fortsetzung)
   Außerdem errechnet man

$$\begin{aligned} SW &= \sum_{j=1}^{3}\sum_{i=1}^{n_j}(x_{j,i}-\overline{x}_j)^2 = \sum_{j=1}^{3}\left(\left(\sum_{i=1}^{n_j}x_{j,i}^2\right) - n_j\cdot\overline{x}_j^2\right) \\ &= \left(\sum_{i=1}^{n_1}x_{j,i}^2\right) - n_1\cdot\overline{x}_1^2 + \left(\sum_{i=1}^{n_2}x_{j,i}^2\right) - n_2\cdot\overline{x}_2^2 + \left(\sum_{i=1}^{n_3}x_{j,i}^2\right) - n_3\cdot\overline{x}_3^2 \\ &= 4271.59 - 40\cdot 10.18^2 + 3730.53 - 33\cdot 10.46^2 + 3959.03 - 30\cdot 11.37^2 \\ &= 326.96\ . \end{aligned}$$

   Insgesamt erhält man

$$F = \frac{SB/(k-1)}{SW/(n-k)} = \frac{25.46/(3-1)}{326.96/(103-3)} = \frac{12.73}{3.27} = 3.89\ .$$

5. **Entscheidung:**
   $F = 3.89 \in (3.087, +\infty) = K \quad \Rightarrow \quad H_0$ wird abgelehnt!
   (p-Wert: $1 - F_{F(2,100)}(F) = 1 - F_{F(2,100)}(3.89) = 1 - 0.98 = 0.02$)

ANOVA-Tabelle

- Zusammenfassung der (Zwischen-)Ergebnisse einer Varianzanalyse oft in Form einer sog. ANOVA(ANalysis Of VAriance) - Tabelle wie folgt:

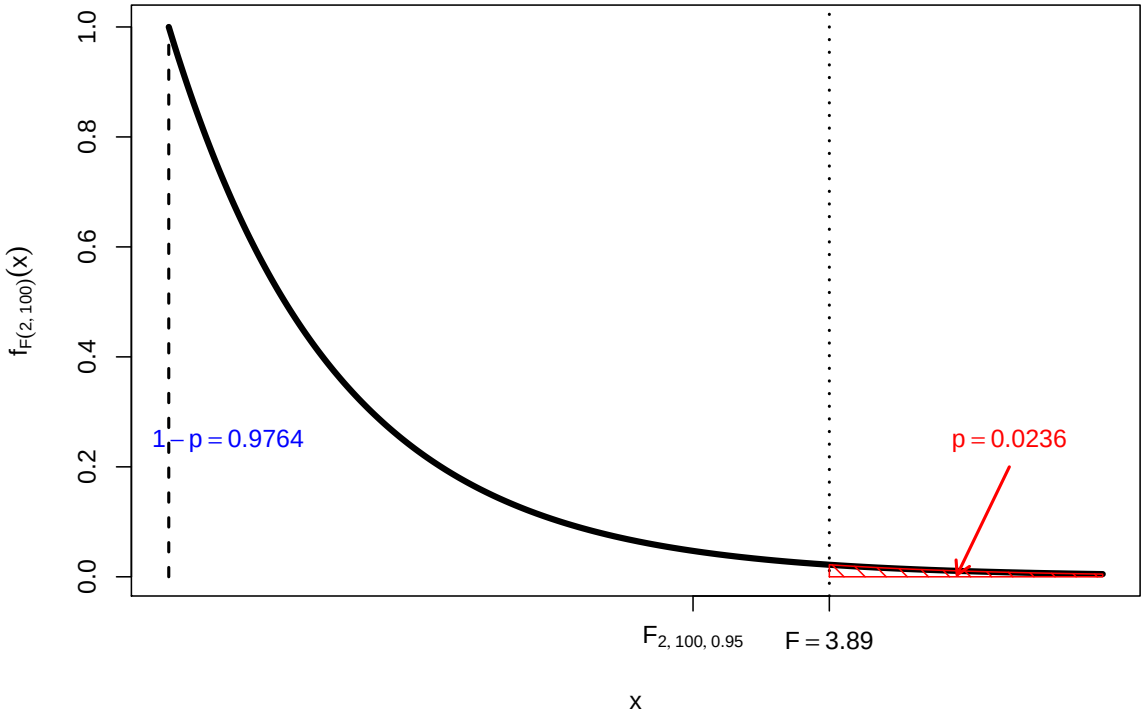
Streuungs-ursache	Freiheits-grade	Quadrat-summe	Mittleres Quadrat
Faktor	$k-1$	SB	$\frac{SB}{k-1}$
Zufallsfehler	$n-k$	SW	$\frac{SW}{n-k}$
Summe	$n-1$	SS	

- Im Bedienungszeiten-Beispiel erhält man so:

Streuungs-ursache	Freiheits-grade	Quadrat-summe	Mittleres Quadrat
Faktor	2	25.46	12.73
Zufallsfehler	100	326.96	3.27
Summe	102	352.42	

Beispiel: p-Wert bei Varianzanalyse (Grafik)

Bedienungszeiten-Beispiel, realisierte Teststatistik $F = 3.89$, p-Wert: 0.0236



Varianzanalyse und 2-Stichproben-t-Test

- Varianzanalyse zwar für $k > 2$ unabhängige Stichproben eingeführt, Anwendung aber auch für $k = 2$ möglich.
- Nach Zuordnung der beteiligten Größen in den unterschiedlichen Notationen ($\mu_A \equiv \mu_1$, $\mu_B \equiv \mu_2$, $X_i^A \equiv X_{1,i}$, $X_i^B \equiv X_{2,i}$, $n_A \equiv n_1$, $n_B \equiv n_2$, $n = n_A + n_B$) enger Zusammenhang zum 2-Stichproben-t-Test erkennbar:
  - Fragestellungen (Hypothesenpaare) und Anwendungsvoraussetzungen identisch mit denen des zweiseitigen 2-Stichproben-t-Tests für den Mittelwertvergleich bei unbekannten, aber übereinstimmenden Varianzen.
  - Man kann zeigen: Für Teststatistik F der Varianzanalyse im Fall $k = 2$ und Teststatistik t des 2-Stichproben-t-Tests gilt $F = t^2$.
  - Es gilt außerdem zwischen Quantilen der $F(1, n)$ und der $t(n)$-Verteilung der Zusammenhang $F_{1,n;1-\alpha} = t^2_{n;1-\frac{\alpha}{2}}$. Damit:

$$x \in (-\infty, -t_{n;1-\frac{\alpha}{2}}) \cup (t_{n;1-\frac{\alpha}{2}}, \infty) \iff x^2 \in (F_{1,n;1-\alpha}, \infty)$$

- Insgesamt sind damit die Varianzanalyse mit $k = 2$ Faktorstufen und der zweiseitige 2-Stichproben-t-Test für den Mittelwertvergleich bei unbekannten, aber übereinstimmenden Varianzen also äquivalent in dem Sinn, dass Sie stets übereinstimmende Testentscheidungen liefern!

Deskriptive Beschreibung linearer Zusammenhänge

- Aus deskriptiver Statistik bekannt: Pearsonscher Korrelationskoeffizient als Maß der Stärke des *linearen* Zusammenhangs zwischen zwei (kardinalskalierten) Merkmalen X und Y.
- *Nun:* Ausführlichere Betrachtung linearer Zusammenhänge zwischen Merkmalen (zunächst rein deskriptiv!):
  Liegt ein linearer Zusammenhang zwischen zwei Merkmalen X und Y nahe, ist nicht nur die Stärke dieses Zusammenhangs interessant, sondern auch die genauere „Form“ des Zusammenhangs.
- „Form“ linearer Zusammenhänge kann durch Geraden(gleichungen) spezifiziert werden.
- *Problemstellung:* Wie kann zu einer Urliste $(x_1, y_1), \ldots, (x_n, y_n)$ der Länge n zu (X, Y) eine sog. **Regressiongerade** (auch: Ausgleichsgerade) gefunden werden, die den linearen Zusammenhang zwischen X und Y „möglichst gut“ widerspiegelt?
- *Wichtig:* Was soll „möglichst gut“ überhaupt bedeuten?
  Hier: Summe der quadrierten Abstände von der Geraden zu den Datenpunkten (x_i, y_i) in **vertikaler** Richtung soll möglichst gering sein.
  (Begründung für Verwendung dieses „Qualitätskriteriums“ wird nachgeliefert!)

- Geraden (eindeutig) bestimmt (zum Beispiel) durch Absolutglied a und Steigung b in der bekannten Darstellung

$$y = f_{a,b}(x) := a + b \cdot x \ .$$

- Für den i-ten Datenpunkt (x_i, y_i) erhält man damit den vertikalen Abstand

$$u_i(a, b) := y_i - f_{a,b}(x_i) = y_i - (a + b \cdot x_i)$$

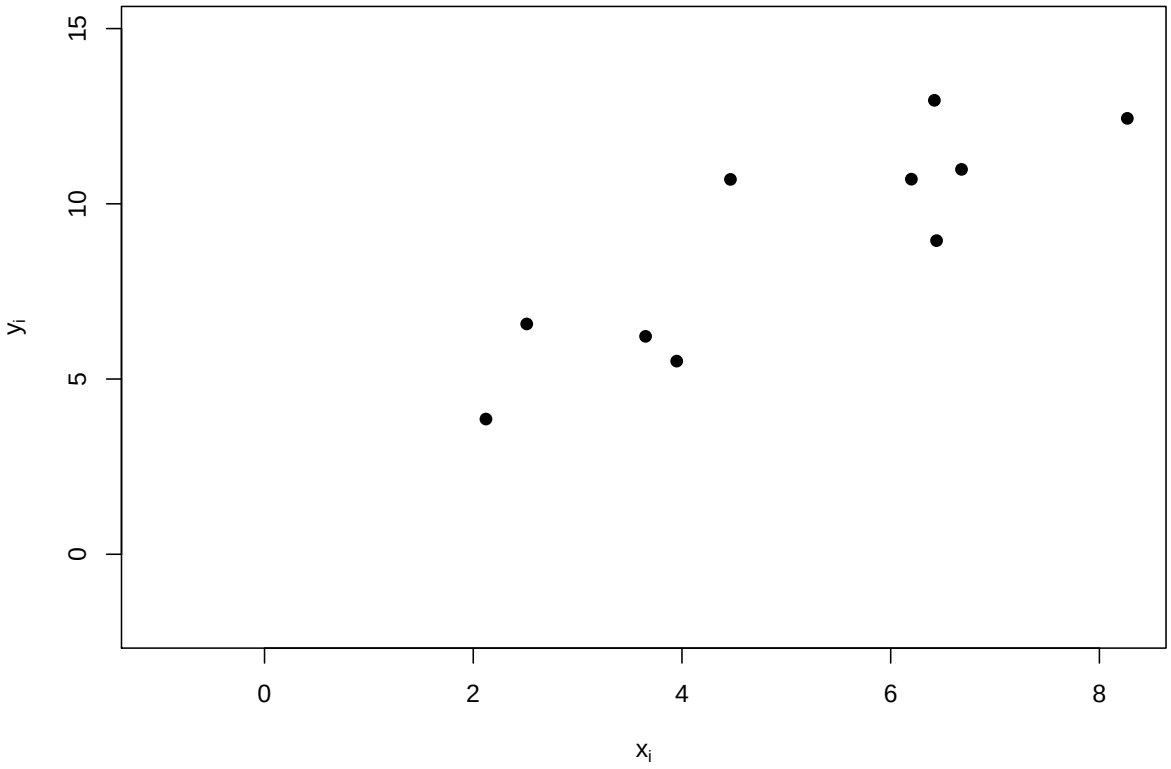
  von der Geraden mit Absolutglied a und Steigung b.
- Gesucht werden a und b so, dass die Summe der quadrierten vertikalen Abstände der „Punktwolke“ (x_i, y_i) von der durch a und b festgelegten Geraden,

$$\sum_{i=1}^{n} (u_i(a, b))^2 = \sum_{i=1}^{n} (y_i - f_{a,b}(x_i))^2 = \sum_{i=1}^{n} (y_i - (a + b \cdot x_i))^2 \ ,$$

  möglichst klein wird.
- Verwendung dieses Kriteriums heißt auch **Methode der kleinsten Quadrate (KQ-Methode)** oder **Least-Squares-Methode (LS-Methode)**.

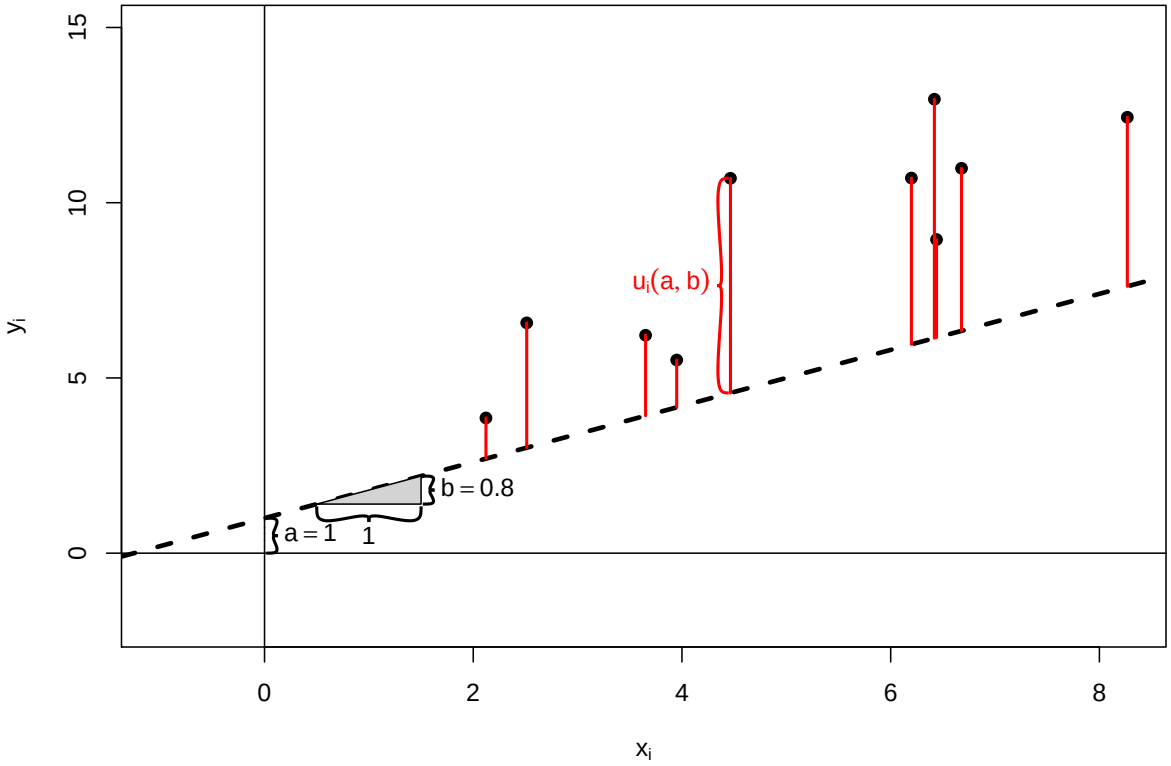
Beispiel: „Punktwolke“

aus $n = 10$ Paaren (x_i, y_i)



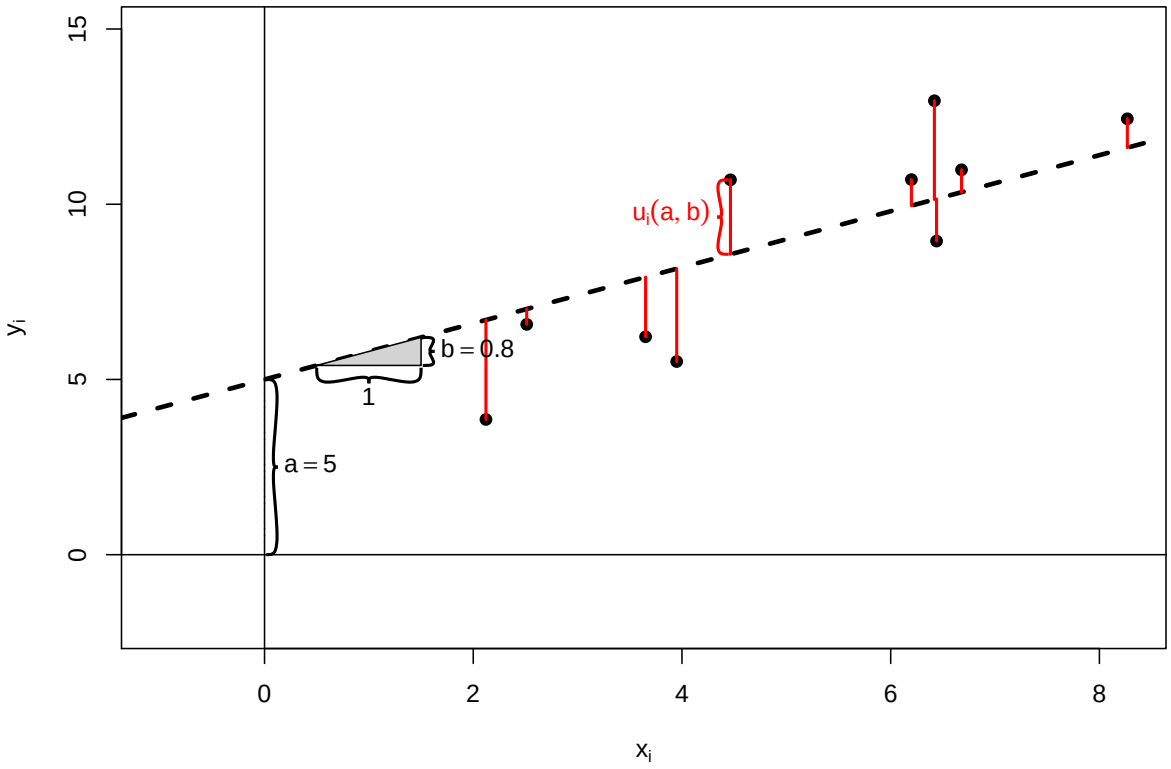
Beispiel: „Punktwolke“ und verschiedene Geraden (I)

$a = 1$, $b = 0.8$, $\sum_{i=1}^{n}(u_i(a,b))^2 = 180.32$



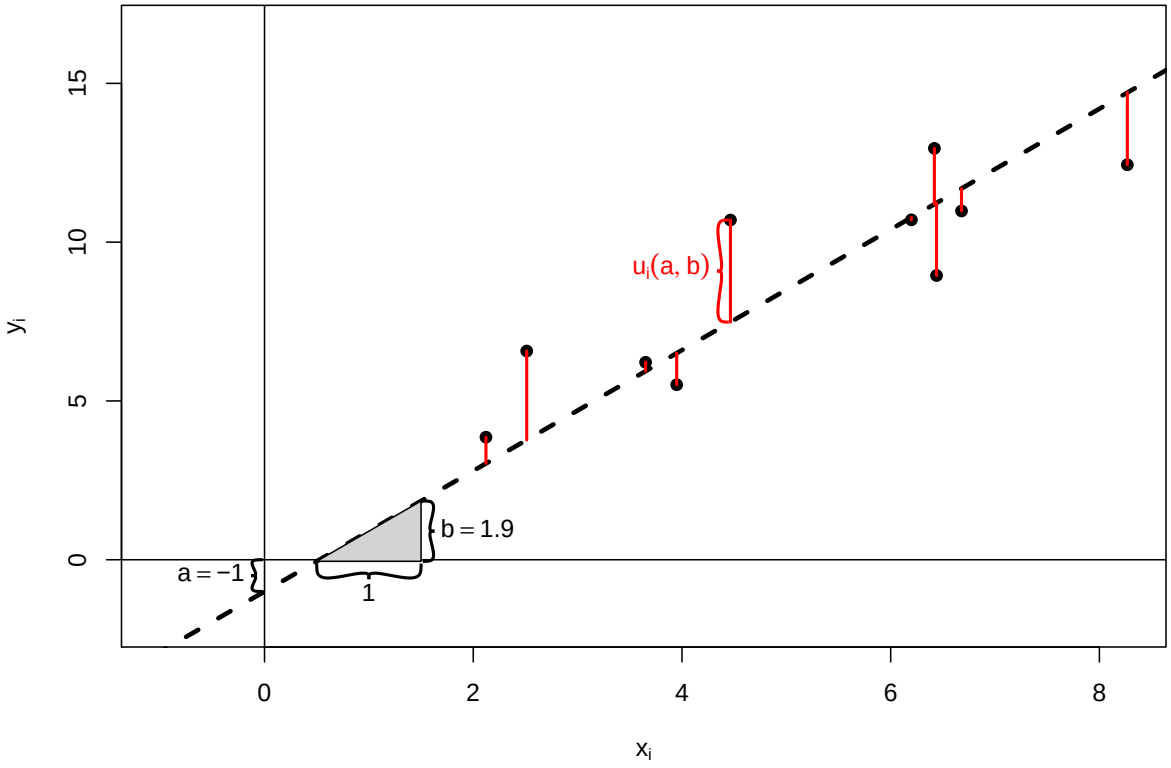
Beispiel: „Punktwolke“ und verschiedene Geraden (II)

$a = 5,\ b = 0.8,\ \sum_{i=1}^{n}(u_i(a,b))^2 = 33.71$



Beispiel: „Punktwolke“ und verschiedene Geraden (III)

$a = -1,\ b = 1.9,\ \sum_{i=1}^{n}(u_i(a,b))^2 = 33.89$



Rechnerische Bestimmung der Regressionsgeraden (I)

- Gesucht sind also $\widehat{a}, \widehat{b} \in \mathbb{R}$ mit

$$\sum_{i=1}^{n}(y_i - (\widehat{a} + \widehat{b}x_i))^2 = \min_{a,b\in\mathbb{R}} \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

- Lösung dieses Optimierungsproblems durch Nullsetzen des Gradienten, also

$$\frac{\partial \sum_{i=1}^{n}(y_i - (a + bx_i))^2}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) \stackrel{!}{=} 0$$

$$\frac{\partial \sum_{i=1}^{n}(y_i - (a + bx_i))^2}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i \stackrel{!}{=} 0 \ ,$$

führt zu sogenannten **Normalgleichungen**:

$$na + \left(\sum_{i=1}^{n} x_i\right) b \stackrel{!}{=} \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b \stackrel{!}{=} \sum_{i=1}^{n} x_i y_i$$

Rechnerische Bestimmung der Regressionsgeraden (II)

- Aufgelöst nach a und b erhält man die Lösungen

$$\widehat{b} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\widehat{a} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \cdot \widehat{b}$$

oder kürzer mit den aus der deskr. Statistik bekannten Bezeichnungen

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{und} \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

bzw. den empirischen Momenten $s_{X,Y} = \overline{xy} - \overline{x} \cdot \overline{y}$ und $s_X^2 = \overline{x^2} - \overline{x}^2$:

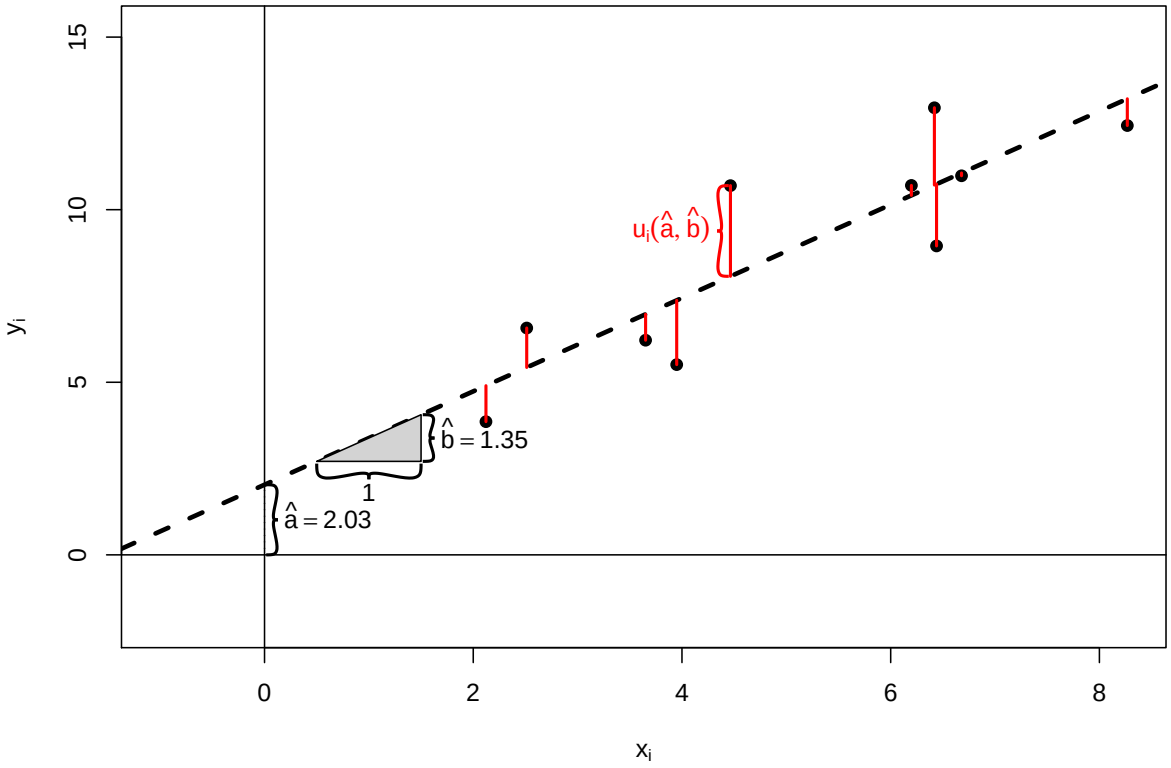
$$\widehat{b} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{s_{X,Y}}{s_X^2}$$

$$\widehat{a} = \overline{y} - \overline{x}\widehat{b}$$

- Die erhaltenen Werte $\widehat{a}$ und $\widehat{b}$ minimieren tatsächlich die Summe der quadrierten vertikalen Abstände, da die Hesse-Matrix positiv definit ist.

Beispiel: „Punktwolke“ und Regressionsgerade

$\widehat{a} = 2.03$, $\widehat{b} = 1.35$, $\sum_{i=1}^{n}(u_i(\widehat{a},\widehat{b}))^2 = 22.25$



- Zu $\widehat{a}$ und $\widehat{b}$ kann man offensichtlich die folgende, *durch die Regressionsgerade erzeugte* Zerlegung der Merkmalswerte y_i betrachten:

$$y_i = \underbrace{\widehat{a} + \widehat{b} \cdot x_i}_{=:\widehat{y}_i} + \underbrace{y_i - (\widehat{a} + \widehat{b} \cdot x_i)}_{=u_i(\widehat{a},\widehat{b})=:\widehat{u}_i}$$

- Aus den Normalgleichungen lassen sich leicht einige wichtige Eigenschaften für die so definierten $\widehat{u}_i$ und $\widehat{y}_i$ herleiten, insbesondere:
  - $\sum_{i=1}^{n} \widehat{u}_i = 0$ und damit $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$ bzw. $\overline{y} = \overline{\widehat{y}} := \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i$.
  - $\sum_{i=1}^{n} x_i \widehat{u}_i = 0$.
  - Mit $\sum_{i=1}^{n} \widehat{u}_i = 0$ und $\sum_{i=1}^{n} x_i \widehat{u}_i = 0$ folgt auch $\sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i = 0$.

  Mit diesen Eigenschaften erhält man die folgende Varianzzerlegung:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{Gesamtvarianz der } y_i} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2}_{\text{erklärte Varianz}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2}_{\text{unerklärte Varianz}}$$

- Die als Anteil der erklärten Varianz an der Gesamtvarianz gemessene Stärke des linearen Zusammenhangs steht in engem Zusammenhang mit $r_{X,Y}$; es gilt:

$$r_{X,Y}^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$