

# Mittelwertvergleiche bei $k > 2$ unabhängigen Stichproben

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

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

auf Basis von  $k$  unabhängigen einfachen Stichproben

$$X_{1,1}, \dots, X_{1,n_1}, \quad \dots, \quad X_{k,1}, \dots, X_{k,n_k}$$

mit Stichprobenumfängen  $n_1, \dots, n_k$  (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“)

$$\bar{X}_1 := \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1,i}, \quad \dots, \quad \bar{X}_k := \frac{1}{n_k} \sum_{i=1}^{n_k} X_{k,i}$$

um das Gesamtmittel

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

mit den Streuungen der Beobachtungswerte  $X_{j,i}$  um die jeweiligen Stufenmittel  $\bar{X}_j$  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:

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

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

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

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

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

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  $\alpha$  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 näherungsweise 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

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

- Alternative Berechnungsmöglichkeiten mit „Verschiebungssatz“
  - ▶ für Realisation von  $SB$ :

$$SB = \sum_{j=1}^k n_j \cdot (\bar{x}_j - \bar{x})^2 = \left( \sum_{j=1}^k n_j \bar{x}_j^2 \right) - n\bar{x}^2$$

- ▶ für Realisation von  $SW$ :

$$SW = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{j,i} - \bar{x}_j)^2 = \sum_{j=1}^k \left( \left( \sum_{i=1}^{n_j} x_{j,i}^2 \right) - n_j \bar{x}_j^2 \right)$$

- Liegen für  $j \in \{1, \dots, k\}$  die Stichprobenvarianzen

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (X_{j,i} - \bar{X}_j)^2$$

bzw. deren Realisationen  $s_j^2$  für die  $k$  (Einzel-)Stichproben

$$X_{1,1}, \dots, X_{1,n_1}, \quad \dots \quad X_{k,1}, \dots, X_{k,n_k}$$

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  $\mu_1, \mu_2, \mu_3$  an 3 verschiedenen Servicepunkten überein oder nicht?
- Annahme: Bedienungszeiten  $Y_1, Y_2, Y_3$  an den 3 Servicestationen sind jeweils normalverteilt mit  $E(Y_j) = \mu_j$  und **identischer** (unbekannter) Varianz  $\text{Var}(Y_j) = \sigma^2$ .
- 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:

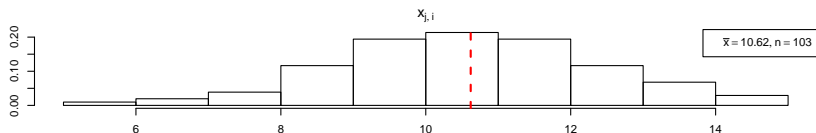
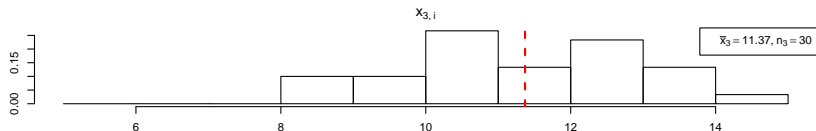
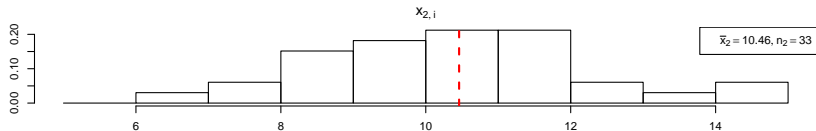
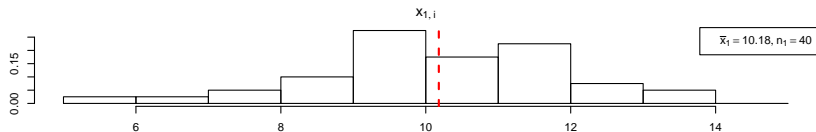
$j$ (Servicepunkt)	$n_j$	$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{j,i}$	$\sum_{i=1}^{n_j} x_{j,i}^2$
1	40	10.18	4271.59
2	33	10.46	3730.53
3	30	11.37	3959.03

(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:**

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad H_1 : \mu_1 \neq \mu_j \text{ für mindestens ein } j$$

② **Teststatistik:**

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

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

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

und damit

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

## 4 (Fortsetzung)

Außerdem errechnet man

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

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:**

$$F = 3.89 \in (3.087, +\infty) = K \quad \Rightarrow \quad H_0 \text{ wird abgelehnt!}$$

$$(\textit{p-Wert: } 1 - F_{F(2,100)}(F) = 1 - F_{F(2,100)}(3.89) = 1 - 0.98 = 0.02)$$

# ANOVA-Tabelle

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

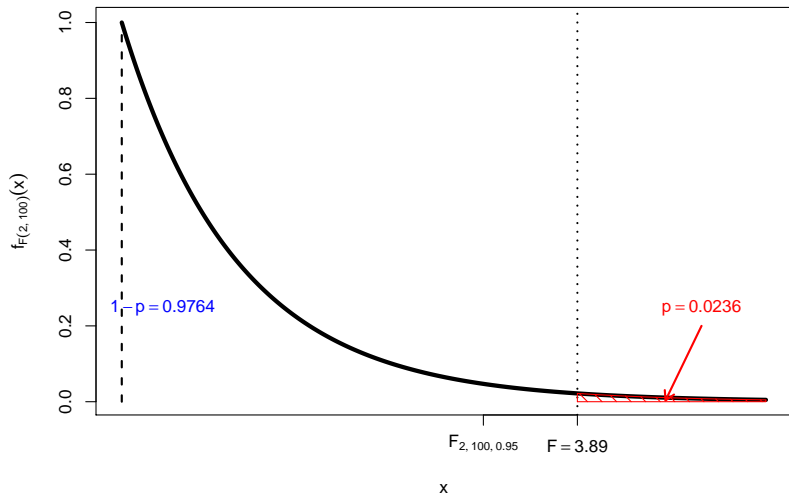
Streuungs- ursache	Freiheits- grade	Quadrat- summe	Mittleres Quadrat
Faktor	$k - 1$	$SB$	$\frac{SB}{k - 1}$
Zufallsfehler	$n - k$	$SW$	$\frac{SW}{n - k}$
Summe	$n - 1$	$SS$	

- 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 unbekanntem, 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  $F_{1, n; 1-\alpha} = t_{n, 1-\frac{\alpha}{2}}^2$ . 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 unbekanntem, 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  $(x_1, y_1), \dots, (x_n, y_n)$  der Länge  $n$  zu  $(X, Y)$  eine sog. **Regressionsgerade** (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,

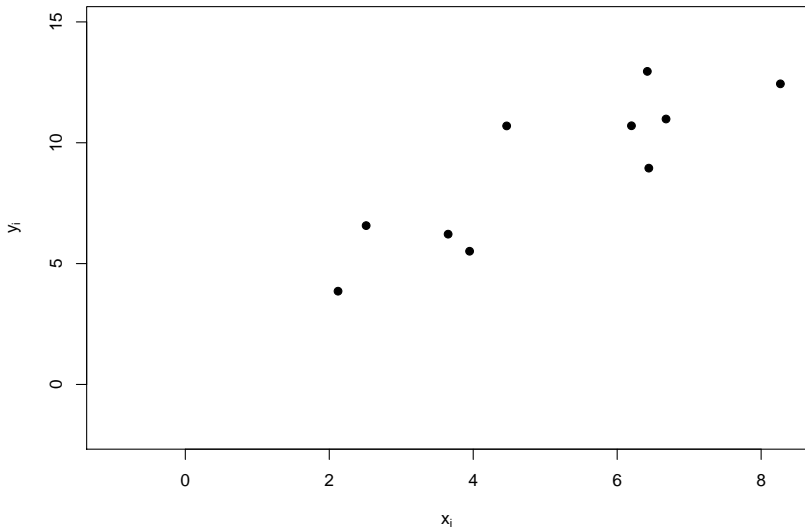
$$\sum_{i=1}^n (u_i(a, b))^2 = \sum_{i=1}^n (y_i - f_{a,b}(x_i))^2 = \sum_{i=1}^n (y_i - (a + b \cdot x_i))^2 ,$$

möglichst klein wird.

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

# Beispiel: „Punktwolke“

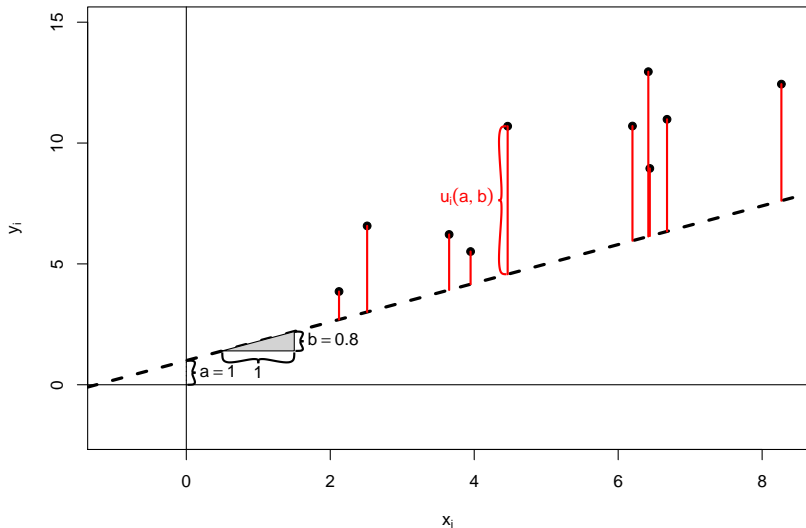
aus  $n = 10$  Paaren  $(x_i, y_i)$





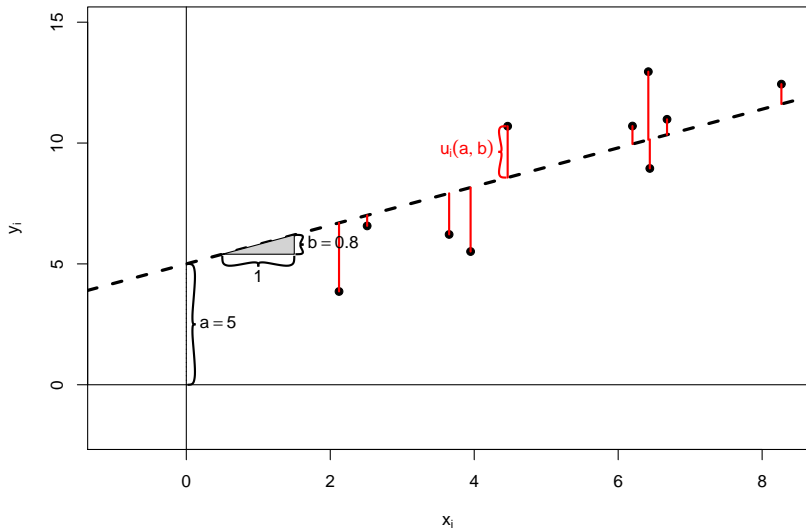
# Beispiel: „Punktvolke“ und verschiedene Geraden (I)

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



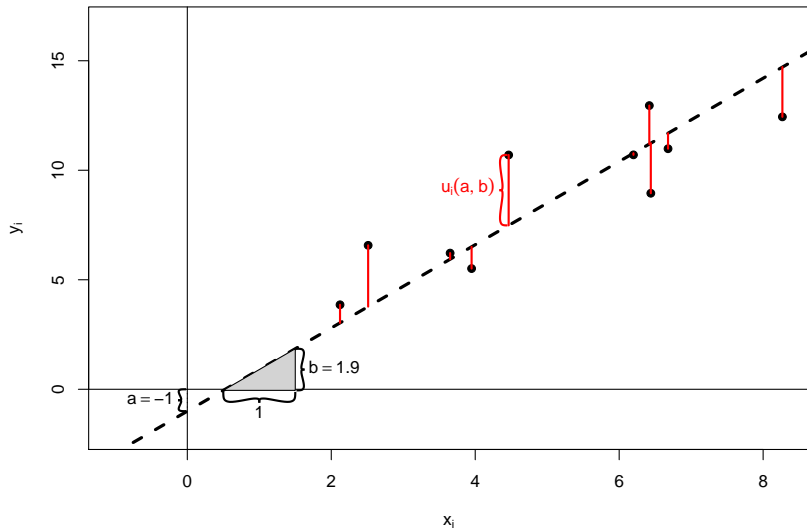
# Beispiel: „Punktvolke“ und verschiedene Geraden (II)

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



# 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  $\hat{a}, \hat{b} \in \mathbb{R}$  mit

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

- 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

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

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

oder kürzer mit den aus der deskr. Statistik bekannten Bezeichnungen

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

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

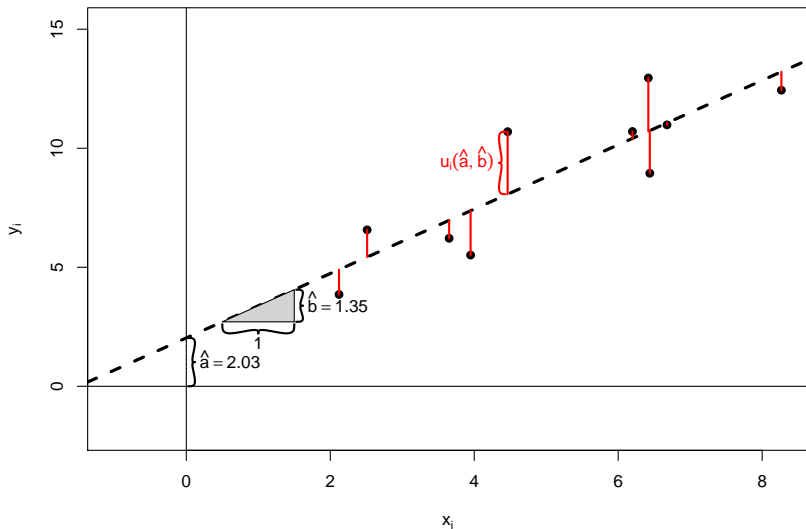
$$\hat{b} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{X,Y}}{s_X^2}$$

$$\hat{a} = \bar{y} - \bar{x} \hat{b}$$

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

# Beispiel: „Punktwolke“ und Regressionsgerade

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



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

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

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

Mit diesen Eigenschaften erhält man die folgende Varianzzerlegung:

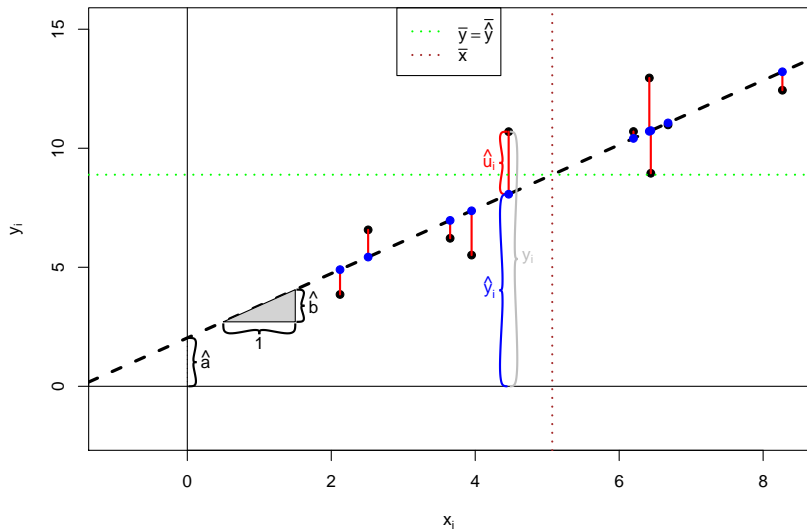
$$\underbrace{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}_{\text{Gesamtvarianz der } y_i} = \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}_{\text{erklärte Varianz}} + \underbrace{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2}_{\text{unerklärte Varianz}}$$

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

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

# Beispiel: Regressionsgerade mit Zerlegung $y_i = \hat{y}_i + \hat{u}_i$

$$\hat{a} = 2.03, \hat{b} = 1.35, \sum_{i=1}^n (\hat{u}_i)^2 = 22.25$$





## Beispiel: Berechnung von $\hat{a}$ und $\hat{b}$

- Daten im Beispiel:

$i$	1	2	3	4	5	6	7	8	9	10
$x_i$	2.51	8.27	4.46	3.95	6.42	6.44	2.12	3.65	6.2	6.68
$y_i$	6.57	12.44	10.7	5.51	12.95	8.95	3.86	6.22	10.7	10.98

- Berechnete (deskriptive/empirische) Größen:

$$\begin{aligned} \bar{x} &= 5.0703 & \bar{y} &= 8.8889 & \overline{x^2} &= 29.3729 & \overline{y^2} &= 87.9398 \\ s_X^2 &= 3.665 & s_Y^2 &= 8.927 & \overline{xy} &= 50.0257 & s_{X,Y} &= 4.956 \end{aligned}$$

- Damit erhält man Absolutglied  $\hat{a}$  und Steigung  $\hat{b}$  als

$$\hat{b} = \frac{s_{X,Y}}{s_X^2} = \frac{4.956}{3.665} = 1.352$$

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = 8.8889 - 1.352 \cdot 5.0703 = 2.03$$

und damit die Regressionsgerade

$$y = f(x) = 2.03 + 1.352 \cdot x .$$