# 10. Varianzanalyse

Mit der einfache Varianzanalyse (ANOVA = Analysis of Variance) wird die Hypothese geprüft, ob die Mittelwerte zweier oder mehrerer Stichproben identisch sind, die aus normalverteilten Grundgesamtheiten gezogen werden, die denselben Mittelwert besitzen. Die Varianzanalyse ist somit eine Erweiterung des *t*-Tests, mit dem ja nur zwei Gruppen untersucht werden können.
Im Gegensatz zur *einfachen* Varianzanalyse, bei der eine Stichprobe bzgl. einer *einzelnen* Variationsursache in *r* Gruppen unterteilt wird, lassen sich die Gruppen bei der *doppelten* Varianzanalyse noch zusätzlich nach weiteren Gesichtspunkten untergliedern. Somit kann gleichzeitig der Einfluss mehrerer Variationsursachen untersucht werden (z.B. der Einfluss von Magnesiumgehalt *und* Wärmebehandlung auf die Festigkeit von AlSi10Mg).

Bei der *einfachen* Varianzanalyse wird eine Stichprobe vom Umfang *n* in *r* Gruppen unterteilt. Dabei werden die Elemente der Gruppen wie folgt bezeichnet:

$x_{11}, x_{12}, \ldots, x_{1n_1}$ 1. Gruppe
$x_{21}, x_{22}, \ldots, x_{2n_2}$ 2. Gruppe
.
.
.
$x_{r1}, x_{r2}, \ldots, x_{rn_r}$ *r*. Gruppe

Dabei gilt: $n_1 + n_2 + \ldots + n_r = n$ .

Wir setzen weiter voraus, dass die *r* Gruppen von Zahlen aus *r* normalverteilten Grundgesamtheiten entstammen, die alle dieselbe Varianz $\sigma^2$ besitzen, die allerdings nicht bekannt zu sein braucht. Es soll geprüft werden, ob die Mittelwerte $\mu_1, \ldots, \mu_r$ der *r* Grundgesamtheiten gleich sind. Das Problem lösen wir mit folgender Arbeitsanleitung oder wie wir später sehen werden mit einem entsprechenden Computerprogramm.

**Einfache Varianzanalyse**

Test der Hypothese **H**, die *r* normalverteilten Grundgesamtheiten *gleicher* Varianz haben alle denselben Mittelwert.

(i) Wählen einer Signifikanzzahl $\alpha$ (5% oder 1%)

(ii) Berechnen der *r* Mittelwerte $\bar{x}_1, \ldots, \bar{x}_r$ der Gruppen. Hierbei ist

$$\bar{x}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ik} \quad .$$

Als Mittelwert der gesamten Stichprobe bekommen wir dann:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} n_i \bar{x}_i$$

(iii) Wir berechnen die *Quadratsummen zwischen den Mittelwerten der Gruppen*

$$q_1 = \sum_{i=1}^{r} (\bar{x}_i - \bar{x})^2$$

und die *Quadratsumme innerhalb der Gruppen*

$$q_2 = \sum_{i=1}^{r} \sum_{k=1}^{n_i} (x_{ik} - \bar{x}_i)^2 \ .$$

Damit bilden wir die Prüfgröße

$$v_0 = \frac{q_1/(r-1)}{q_2/(n-r)}.$$

(iv) Bestimmen der Lösung $c$ der Gleichung

$$P(V \le c) = 1 - \alpha$$

aus einer geeigneten Tafel für die $F$-Verteilung mit ($r$ - 1, $n$ - $r$) Freiheitsgraden. Ist $v_0 \le c$, so nehmen wir die Hypothese $\mu_1 = \mu_2 = ... = \mu_r$ an, ansonsten wird sie verworfen, d.h. wir nehmen an, dass die Mittelwerte nicht alle gleich sind.

Arbeitsanleitung 10.1: Einfache Varianzanalyse

***Beispiel 10.1:***

Bei Folien, die aus einer Titanlegierung hergestellt werden, soll geprüft werden, ob die Zugfestigkeit an allen Stellen dieselbe ist. Es wurden vier Folien untersucht. Die Ergebnisse sind in der folgenden Tabelle dargestellt.

| **Messstelle** | **Messwerte** | | | |
|---|---|---|---|---|
| 1. Gruppe (Ecke) | 137 | 142 | 128 | 137 |
| 2. Gruppe (Mitte) | 140 | 139 | 117 | 137 |
| 3. Gruppe (Kante) | 142 | 140 | 133 | 141 |

Gruppenmittelwerte: $\bar{x}_1 = 136,\ \bar{x}_2 = 133.25,\ \bar{x}_3 = 139$

Mittelwert der gesamten Stichprobe: $\bar{x} = 136{,}083$ .

Die Quadratsumme zwischen den Gruppen lautet:

$$q_1 = 4[(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + (\bar{x}_3 - \bar{x})^2] = 66{,}167 .$$

Für die Quadratsumme innerhalb der Gruppen erhalten wir:

$$q_2 = \sum_{i=1}^{3}\sum_{k=1}^{4}(x_{ik} - \bar{x}_i) = (137-136)^2 + (142-136)^2 + ... + (141-139)^2 = 508{,}75 .$$

Somit folgt für die Prüfgröße:

$$v_0 = \frac{q_1/2}{q_2/9} = \frac{33{,}083}{56{,}528} = 0{,}585 .$$

Wegen $r$ = 3 und $n$ = 12 folgt $df$ = (2,9). Aus einer entsprechenden Tafel für die $F$-Verteilung erhalten wir demnach als Lösung der Gleichung

$$P(V \le c) = 0{,}95$$

$c$ = 4,26. Da $v_0 < c$, nehmen wir die Hypothese **H**: $\mu_1 = \mu_2 = \mu_3$ an. Mit der Varianzanalyse folgt aus der vorgelegten Stichprobe, dass die Zugfestigkeit der Folie an den verschiedenen Messstellen nur zufallsbedingt schwankt; d.h. der Unterschied in den Messwerten ist nicht signifikant.

Dieses Ergebnis erhalten wir auch ohne großen Rechenaufwand, wenn wir diese Analyse mit MS-Excel durchführen. Wir wählen dazu als Analyse-Funktion die *Einfaktorielle Varianzanalyse* und erhalten folgende Ausgabe:

Anova: Einfaktorielle Varianzanalyse

ZUSAMMENFASSUNG

| *Gruppen* | *Anzahl* | *Summe* | *Mittelwert* | *Varianz* |
|---|---|---|---|---|
| Zeile 1 | 4 | 544 | 136 | 34 |
| Zeile 2 | 4 | 533 | 133,25 | 118,916667 |
| Zeile 3 | 4 | 556 | 139 | 16,6666667 |

ANOVA

| *Streuungsursache* | *Quadratsummen* | *Freiheitsgrade (df)* | *Mittlere Quadratsumme* | *Prüfgröße* | *P-Wert* | *kritischer F-Wert* |
|---|---|---|---|---|---|---|
| Zwischen den Gruppen | 66,16666667 | 2 | 33,08333333 | 0,58525799 | 0,576830021 | 4,256492048 |
| Innerhalb der Gruppen | 508,75 | 9 | 56,52777778 | | | |
| | | | | | | |
| Gesamt | 574,9166667 | 11 | | | | |

Da der $p$-Wert größer als 0,05 ist, nehmen wir die Hypothese **H**: $\mu_1 = \mu_2 = \mu_3$ an.

Wie wir bei der *doppelten* Varianzanalyse, bei der die *r* Gruppen noch in *p* Klassen unterteilt werden, vorgehen müssen, kann der folgenden Arbeitsanleitung entnommen werden.

**Doppelte Varianzanalyse**

Test der Hypothese **H**, die $n = rp$ normalverteilten Grundgesamtheiten gleicher Varianz haben alle denselben Mittelwert.

(i) Wählen einer Signifikanzzahl $\alpha$ (5% oder 1%)

(ii) Wir berechnen

$$q_1 = p\sum_{i=1}^{r}\left(\left(\frac{1}{p}\sum_{k=1}^{p}x_{ik}\right)-\bar{x}\right)^2,\ q_2 = r\sum_{k=1}^{p}\left(\left(\frac{1}{r}\sum_{i=1}^{r}x_{ik}\right)-\bar{x}\right)^2,\ q=\sum_{i=1}^{r}\sum_{k=1}^{p}(x_{ik}-\bar{x})^2$$

und daraus $q_3 = q - q_1 - q_2$.

(iii) Wir setzen $s_1^2 = \frac{q_1}{r-1}$, $s_2^2 = \frac{q_2}{p-1}$ und $s_3^2 = \frac{q_3}{(r-1)(p-1)}$. Damit bekommen wir die zwei Prüfgrößen

$$v_1 = \frac{s_1^2}{s_3^2} \quad \text{und} \quad v_2 = \frac{s_2^2}{s_3^2}$$

(iv) Bestimmen der Lösung $c_1$ der Gleichung

$$P(V \le c_1) = 1-\alpha$$

aus einer geeigneten Tafel für die *F*-Verteilung mit [$r$ – 1, ($r$ – 1)($p$ – 1)] Freiheitsgraden. Ist $v_1 \le c_1$, so wird angenommen, dass zwischen den Zeilen kein signifikanter Unterschied besteht. Ist $v_1 > c_1$, wird angenommen, dass ein solcher Unterschied besteht.

(v) Bestimmen der Lösung $c_2$ der Gleichung

$$P(V \le c_2) = 1-\alpha$$

aus einer geeigneten Tafel für die *F*-Verteilung mit [$p$ – 1, ($r$ – 1)($p$ – 1)] Freiheitsgraden. Ist $v_2 \le c_2$, so wird angenommen, dass zwischen den Spalten kein signifikanter Unterschied besteht. Ist $v_2 > c_2$, wird angenommen, dass ein solcher Unterschied besteht. Ist $v_1 \le c_1$ und $v_2 \le c_2$, so wird die Hypothese **H:** alle $n$ Mittelwerte $\mu_{11},...,\mu_{rp}$ sind gleich, angenommen. Ansonsten wir die Hypothese **H** verworfen.

Arbeitsanleitung 10.2: Doppelte Varianzanalyse

***Beispiel 10.2:***

In der folgenden Tabelle sind die Werte des Stroms [mA], der auf dem Schirm eines bestimmten Typs von Fernsehröhren eine gegebene Helligkeit hervorruft, aufgeführt. Wir untersuchen, ob der Einfluss der Glasart oder der Phosphorart signifikant ist.

| **Glasart** | **Phosphorart** | | | |
|---|---|---|---|---|
| | **A** | **B** | **C** | **D** |
| I | 285 | 302 | 282 | 290 |
| II | 235 | 245 | 225 | 300 |
| III | 240 | 260 | 255 | 295 |

Wir benutzen MS-Excel und wählen als Analyse-Funktion die Zweifaktorielle Varianzanalyse ohne Messwiederholung. Excel liefert folgende Ausgabe:

Anova: Zweifaktorielle Varianzanalyse ohne Messwiederholung

| *ZUSAMMENFASSUNG* | *Anzahl* | *Summe* | *Mittelwert* | *Varianz* |
|---|---|---|---|---|
| Glasart I | 4 | 1159 | 289,75 | 77,5833333 |
| Glasart II | 4 | 1005 | 251,25 | 1122,91667 |
| Glasart III | 4 | 1050 | 262,5 | 541,666667 |
| | | | | |
| Phosphorart A | 3 | 760 | 253,333333 | 758,333333 |
| Phosphorart B | 3 | 807 | 269 | 873 |
| Phosphorart C | 3 | 762 | 254 | 813 |
| Phosphorart D | 3 | 885 | 295 | 25 |

ANOVA

| *Streuungsursache* | *Quadratsummen (SS)* | *Freiheitsgrade (df)* | *Mittlere Quadratsumme (MS)* |
|---|---|---|---|
| Zeilen | 3135,166667 | 2 | 1567,583333 |
| Spalten | 3423 | 3 | 1141 |
| Zufallsfehler | 1803,5 | 6 | 300,5833333 |
| | | | |
| Gesamt | 8361,666667 | 11 | |

| *Streuungsursache* | *Prüfgröße (F)* | *P-Wert* | *kritischer F-Wert* |
|---|---|---|---|
| Zeilen | 5,215137233 | 0,048698916 | 5,143249382 |
| Spalten | 3,795952315 | 0,077304279 | 4,757055194 |

*Analyse der Zeilen:*

Da die *Prüfgröße (F)* größer als der *kritische F-Wert* ist ($v_1 > c_1$), nehmen wir an, dass zwischen den Zeilen ein signifikanter Unterschied besteht. Wir hätten das auch damit begründen können, dass der *p-Wert* kleiner als 0,05 ist.

*Analyse der Spalten:*
Da die *Prüfgröße (F)* kleiner als der *kritische F-Wert* ist ($v_2 \le c_2$), nehmen wir an, dass zwischen den Spalten kein signifikanter Unterschied besteht. Wir hätten das auch damit begründen können, □ass der *p-Wert* größer als 0,05 ist.

*Gesamtanalyse:*
Da $v_1 > c_1$ nehmen wir insgesamt an, dass ein signifikanter Unterschied besteht; und zwar bezüglich der Glasart.

Wie wir bereits erwähnt haben, kann die Varianzanalyse auch bei Regressionsproblemen angewendet werden. Wir beziehen uns auf die Notation im *Kapitel 9* und setzen

(10.1) $$q = (n-1)s_y^2,\ q_1 = (n-1)\frac{s_{xy}^2}{s_x^2} \text{ und } q_2 = q - q_1.$$

In der folgenden Arbeitsanleitung wird für eine *lineare* Regression die Hypothese getestet, der Regressionskoeffizient $\alpha$ der Grundgesamtheit ist gleich Null:

**Varianzanalyse und lineare Regression**
Test der Hypothese **H**: $\alpha = 0$ gegen die Alternative **A**: $\alpha \neq 0$

(i) Wählen einer Signifikanzzahl $\alpha$ (5% oder 1%)

(ii) Aus der gegebenen Stichprobe $\{(x_1, y_1),...,(x_n, y_n)\}$ berechnen wir mit (10.1)

$$q,\ q_1 \text{ und } q_2$$

(iii) Als Prüfgröße setzen wir:

$$v_0 = \frac{q_1}{q_2(n-2)}$$

(iv) Bestimmen der Lösung *c* der Gleichung

$$P(V \le c) = 1 - \alpha$$

aus einer geeigneten Tafel für die *F*-Verteilung mit [1, (*n* - 1)] Freiheitsgraden.
Ist $v_0 \le c$, so wird die Hypothese **H**: $\alpha = 0$ angenommen, ansonsten wird **H** verworfen und die Alternative **A**: $\alpha \neq 0$ angenommen.

Arbeitsanleitung 10.3: Varianzanalyse und Regression

***Beispiel 10.3:***
Eine betriebliche Messung lieferte die folgende Ergebnisse:

| $x_i$ [cm] | $y_i$ [cm] | | | |
|---|---|---|---|---|
| 159 | 24 | 24 | | |
| 160 | 23 | 25 | 27 | |
| 161 | 24 | 24 | 25 | 26 |
| 162 | 23 | 24 | 24 | 29 |
| 166 | 23 | 23 | 25 | |
| 168 | 24 | 29 | 30 | 31 |
| 172 | 24 | 25 | | |

Wir führen eine Regressionsanalyse mit MS-Excel durch, bei der auch die sogenannte ANOVA-Table ausgegeben wird.

| *Regressions-Statistik* | |
|---|---|
| Multipler Korrelationskoeffizient | 0,33427185 |
| Bestimmtheitsmaß | 0,11173767 |
| Adjustiertes Bestimmtheitsmaß | 0,06732456 |
| Standardfehler | 2,31208222 |
| Beobachtungen | 22 |

| ANOVA | | | | | |
|---|---|---|---|---|---|
| | *(df)* | *Quadratsummen* | *Mittlere Quadratsumme* | *Prüfgröße (F)* | *F krit* |
| Regression | 1 | 13,44915254 | 13,44915254 | 2,51587101 | 0,1283902 |
| Residue | 20 | 106,9144838 | 5,345724191 | | |
| Gesamt | 21 | 120,3636364 | | | |

| | *Koeffizienten* | *Standardfehler* | *t-Statistik* | *P-Wert* | *Untere 95%* | *Obere 95%* |
|---|---|---|---|---|---|---|
| Schnittpunkt | -6,69337442 | 20,15929609 | -0,332024213 | 0,74332539 | -48,7449096 | 35,3581608 |
| X Variable 1 | 0,194915254 | 0,122885784 | 1,586149745 | 0,1283902 | -0,06141988 | 0,45125039 |

Excel gibt in der ANOVA-Tabelle folgendes Schema der Varianzanalyse zur Zerlegung an:

| Variation | (df) | Quadratsumme | Durchschnittsquadrate | Prüfgröße | kritischer Wert |
|---|---|---|---|---|---|
| Auf der Regression | 1 | $q_1$ | $q_1$ | $v_0 = \frac{q_1}{q_2(n-2)}$ | $c$ |
| Abweichung von der Regression (Residue) | $n$ - 2 | $q_2$ | $q_2/(n$ - 2) | | |
| Gesamt | $n$ - 1 | $q$ | | | |

Da 2,51587101 > 0,1283902, lehnen wir die Hypothese **H**: $\alpha = 0$ ab.

Mit Hilfe der Varianzanalyse können wir auch die Annahme testen, ob die Regression, die wir betrachten, linear ist. Dazu fassen wir in der Stichprobe

$S = \{(x_1, y_1), ..., (x_n, y_n)\}$ diejenigen Paare gruppenweise zusammen, die denselben *x*-Wert haben. Um das zu verdeutlichen, betrachten wir folgendes Beispiel:

Stichprobe *S* = {(1,5), (2,6), (1,4), (4,11), (1,6), (2,10), (4,7)}

Gruppe 1 - (1,4) (1,5) (1,6)

Gruppe 2 - (2,6) (2,10)

Gruppe 3 - (4,7) (4,11).

Für jede dieser Gruppe berechnen wir den *y*-Mittelwert $\bar{y}_i$, wobei *i* die Gruppennummer bezeichnet. Ist die Regression linear, so müssen die berechneten Mittelwerte $\bar{y}_i$ "ziemlich genau" auf einer Geraden liegen. Um das mathematisch genau zu formulieren, setzen wir:

*n* Stichprobenumfang

*r* Anzahl der Gruppen, also Anzahl der zahlenmäßig verschiedenen *x*-Werte

$n_i$ Anzahl der *y*-Werte in der *i*-ten Gruppe

$y_{ij}$ *j*-ter *y*-Wert in der *i*-ten Gruppe

$\bar{y}_i$ Mittelwert der *y*-Werte in der *i*-ten Gruppe, also $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n_i} y_{ij}$

$\bar{y}$ Mittelwert aller *y*-Werte, also $\bar{y} = \frac{1}{n}\sum_{i=1}^{r} n_i \bar{y}_i$ .

Klar ist, $\sum_{i=1}^{r} n_i = n$. Ferner setzen wir:

$$q = \sum_{i=1}^{r}\sum_{j=1}^{n_i} \left(y_{ij} - f(x_i)\right)^2 ,$$

dabei ist $f(x_i)$ die Gleichung der linearen Regressionsgeraden, die zur Stichprobe *S* ermittelt wurde. Wir zerlegen *q* in zwei Bestandteile

$$q = q_1 + q_2 ,$$

wobei $q_1 = \sum_{i=1}^{r} n_i \left(\bar{y}_i - f(x_i)\right)^2$ und $q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$ . Damit können wir nun folgende Arbeitsanleitung angeben, wobei wir wie in *Kapitel 9* voraussetzen, dass für jedes feste **X** = *x* die Zufallsvariable **Y** normalverteilt ist und die Varianz nicht von *x* abhängt.

**Test auf Linearität der Regression mit Hilfe der Varianzanalyse**

Test der Hypothese **H**: *Die Regression ist linear*
unter der oben genannten Voraussetzung

(i) Wählen einer Signifikanzzahl $\alpha$ (5% oder 1%)

(ii) Aus der gegebenen Stichprobe $\{(x_1, y_1), ..., (x_n, y_n)\}$ berechnen wir

$$q\ ,\ q_1 \text{ und } q_2 = q - q_1$$

(iii) Als Prüfgröße erhalten wir dann

$$v_0 = \frac{q_1 / (r-2)}{q_2 / (n-r)}$$

(iv) Bestimmen der Lösung *c* der Gleichung

$$P(V \le c) = 1 - \alpha$$

aus einer geeigneten Tafel für die *F*-Verteilung mit [(*r* - 2), (*n* - *r*)] Freiheitsgraden. Ist $v_0 \le c$, so wird die Hypothese **H** angenommen, ansonsten wird **H** verworfen.

Arbeitsanleitung 10.4: Test auf Linearität der Regression

Wir erhalten damit folgendes Schema der Varianzanalyse:

| Variation | (df) | Quadrat-summe | Durch-schnitts-quadrate | Prüfgröße | kritischer Wert |
|---|---|---|---|---|---|
| Auf der Regression | *r* - 2 | $q_1$ | $q_1/(r-2)$ | $v_0 = \frac{q_1/(r-2)}{q_2/(n-r)}$ | *c* |
| Abweichung von der Regression (Residue) | *n* - *r* | $q_2$ | $q_2/(n-2)$ | | |
| Gesamt | *n* - 2 | *q* | | | |

Leider ist dieser Test in Excel nicht voreingestellt. Es ist aber kein Problem diesen Test zu programmieren (kleine Übungsaufgabe).

Das oben berschriebenen Testverfahren kann auch auf die nichtlineare Regression übertragen werden. Wir können z.B. testen, ob die Regressionskurve ein Polynom *m*-ten Grades ist. Dazu müssen in der *Arbeitsanleitung 10.4* die Freiheitsgerade *df* = [(*r* - 2), (*n* - *r*)] durch *df* = [(*r* - *m* - 1), (*n* - *r*)] ersetzt werden (eine Gerade ist übrigens ein Polynom 1-ten Grades). Ferner muss bei der Berechnung von *q* die entsprechende Funktionsgleichung für *f* eingesetzt werden.