

Varianzanalyse - ANOVA

Daten: kiga_57_T.sav und kiga_57.sav in PSPP.

Im obigen Beitrag zum 2-Stichproben t-Test (groß) wurde der Einfluss von 2 Gruppen auf den Mittelwert des dmft untersucht. Bei den Gruppen handelte es sich um Kindergärten eines Landkreises aus jeweils unterschiedlichen Regionen, z.B. Gr.1 = Nordkreis und Gr.2 = Südkreis. Getestet wurde, ob sich die dmft-MW beider Gruppen unterscheiden. Mit $p = 0,0385$ ergab sich für ungleiche Varianzen (Welch-Test) ein (schwach) signifikanter Unterschied und ohne Berücksichtigung der Varianzhomogenität mit $p = 0,0373$ ein ähnlicher Wert.

Betrachtet man die Gruppen als Einfluss - **Faktor** und die 2 Ausprägungen in Gruppe 1 und Gruppe 2 als **Faktorstufen**, so ist in der Nomenklatur der Varianzanalyse (ANOVA) der obige t-Test eine **Einfaktorielle Varianzanalyse mit 2 Faktorstufen**. Modifikationen bezüglich der Varianzhomogenität (z.B. Welch-Test) sind in der ANOVA nicht vorgesehen, da hier die Homogenität vorausgesetzt wird. Zur Demonstration laden Sie die Datei [kiga_57_T.sav](#) in PSPP.

2-Stichproben t-Test (gleiche Varianzen angenommen)

Test bei unabhängigen Stichproben	T-Test für die Gleichheit der Mittelwerte						
		t	df	Sig. (2-seitig)	Mittlere Differenz	Stdfehler der Differenz	95% Konfidenzintervall der Differenz
							Untere Obere
dmf	Varianzen sind gleich	-2,083	2705,000	,037	-,266	,128	-,516 -,016

Einfaktorielle Varianzanalyse mit 2 Faktorstufen

ANOVA		Quadrat summe	df	Mittel der Quadrate	F	Sig.
dmf	Zwischen Gruppen	47,334	1	47,334	4,341	,037
	Innerhalb Gruppen	29497,57	2705	10,905		
	Gesamt	29544,91	2706			

Einfaktorielle Varianzanalyse mit 3 Faktorstufen

Häufig möchte man mehr als zwei Mittelwerte vergleichen. Beispielsweise soll untersucht werden, ob sich die dmft-MW von 3-, 4- und 5-jährigen Kindern einer Region unterscheiden. Der Faktor ist hier das Alter mit drei Faktorstufen (3,4,5). Bezeichnet man mit $dmft_i^*$ den Erwartungswert der i-ten Altersgruppe (dmft-MW der i-ten Grundgesamtheit), so lautet die Nullhypothese $H_0 : dmft_3^* = dmft_4^* = dmft_5^*$, d.h., die Erwartungswerte sind in allen drei Altersgruppen gleich. Für die Alternativhypothese gilt $H_1 : dmft_i^* \neq dmft_j^*$, d.h., mindestens zwei Erwartungswerte unterscheiden sich. Für diese Fragestellung bietet sich die ANOVA an mit folgenden Voraussetzungen:

- normalverteilte Meßwerte in den Grundgesamtheiten der Faktorstufen
- gleiche Varianzen der Grundgesamtheiten
- unabhängige Einzelbeobachtungen

Diese strengen Voraussetzungen sind häufig nicht erfüllt, doch ist die ANOVA robust bei annähernd gleichen Verteilungsformen der Subgruppen (Altersgruppen) und großen Stichprobenumfängen (Fallzahlen), was hier zutrifft.

Zur Demonstration laden Sie die Datei kiga_57.sav in PSPP.

ANOVA		Quadrat summe	df	Mittel der Quadrate	F	Sig.
dmf	Zwischen Gruppen	324,345	2	162,173	15,007	,000
	Innerhalb Gruppen	29220,56	2704	10,806		
	Gesamt	29544,91	2706			

Auf die Theorie der Varianzanalyse (Zerlegung der Streuung) soll hier nicht eingegangen werden. Im Ergebnis unterscheiden sich die dmft-MW von mindestens zwei Altersgruppen, doch man kann noch nicht erkennen, welche das sind. Hier stehen in Statistikprogrammen sog. **Post-Hoc-Tests** bereit, die gleichzeitig adjustieren, um das Signifikanzniveau von z.B. $\alpha = 5\%$ einzuhalten. Der Scheffe-Test z.B. liefert folgendes Ergebnis:

Mehrfach vergleiche (dmf)	(I) Familie	(J) Familie	Differenz Mittelwerte (I - J)	Standard fehler	Sig.	95% Konfidenzintervall	
						Untere Grenze	Obere Grenze
Scheffé	3	4	-,617	,181	,003	-1,061	-,173
		5	-,968	,178	,000	-1,402	-,533
	4	3	,617	,181	,003	,173	1,061
		5	-,351	,140	,043	-,693	-,009
	5	3	,968	,178	,000	,533	1,402
		4	,351	,140	,043	,009	,693

Alle paarweisen Vergleiche zeigen signifikante Unterschiede zwischen den dmft-Mittelwerten der drei Altersgruppen.

Alter	N	dmf- Mittelwert	Standardab- weichung
3	482	1.15	2.792
4	1037	1.77	3.341
5	1188	2.12	3.423
Gesamt	2707	1.81	3.304

In der Regel beeinflusst nicht nur ein Faktor (z.B. Alter) die abhängige Variable (z.B. dmft-MW). Mögliche weitere Einflußfaktoren wären hier z.B. Zahnpflege, Sozialstatus, Wohnort ua. Je mehr Einflußfaktoren berücksichtigt werden, desto schwieriger wird möglicherweise die Interpretation der Ergebnisse. Die Handrechnung eines Beispiels einer Zweifaktor-ANOVA findet man im Buch von Kim und Dailey, Seiten 246 - 253 (siehe Literatur & Software).

Zweifaktorielle Varianzanalyse mit unterschiedlichen Faktorstufen

Zwei Medizinprodukte (Faktor: **Produkt** mit 2 Faktorstufen) werden in drei Bundesländern (Faktor: **Land** mit 3 Faktorstufen) vertrieben. Anhand von Umsatzzahlen eines bestimmten Zeitraumes (in T€) möchte man den Einfluß der beiden Faktoren "Produkt" und "Land" untersuchen (**ZweifaktorANOVA_S250.dta**).

Für die mittleren Umsätze (Umsatz-MW) in den 3 Ländern erhält man folgende Tabelle:

Umsatz-MW	PRODUKT	
	1	2
LAND		
a	42.4	38.8
b	35.6	37.4
c	32.4	34.6

Quelle:

Fahrmeir L. et al.: Arbeitsbuch Statistik. S.250, Springer 2003

Im Land a werden die höchsten Umsätze erreicht, hier speziell mit Produkt 1. Den niedrigsten Umsatz-MW findet man in Land c, ebenfalls mit Produkt 1.

Die Zweifaktorielle Varianzanalyse mit STATA liefert folgende Tabellen, in der erkennbar das Land einen signifikanten Einfluss auf den Umsatz hat, der Faktor "Produkt" dagegen nicht. Die deutliche Wechselwirkung von "Produkt" und "Land" ist aus der obigen Tabelle und den untenstehenden Grafiken erkennbar.

```
anova umsatz produkt##land
```

```
Number of obs =      30      R-squared      = 0.8775
Root MSE      = 1.33542    Adj R-squared = 0.8520
```

Source	Partial SS	df	MS	F	Prob>F
Model	306.66667	5	61.333333	34.39	0.0000
produkt	.13333333	1	.13333333	0.07	0.7869
land	254.06667	2	127.03333	71.23	0.0000
produkt#land	52.466667	2	26.233333	14.71	0.0001
Residual	42.8	24	1.7833333		
Total	349.46667	29	12.050575		

Das partielle Eta² ist eine Effektgröße, die die Bedeutung der einzelnen Faktoren im Modell beschreibt. Auch hier erkennt man die vergleichsweise geringe Bedeutung des nichtsignifikanten Faktors "Produkt" für den mittleren Umsatz.

Eta²-Werte von 0,01 ; 0,06 ; 0,14 gelten als schwache, mittlere oder starke Effekte.

```
. estat esize
```

Effect sizes for linear models

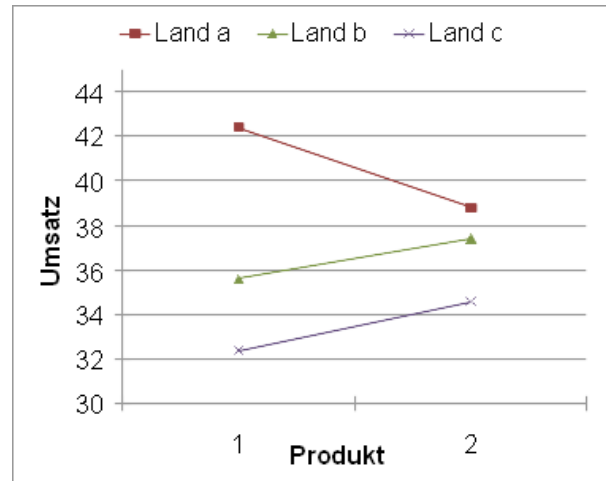
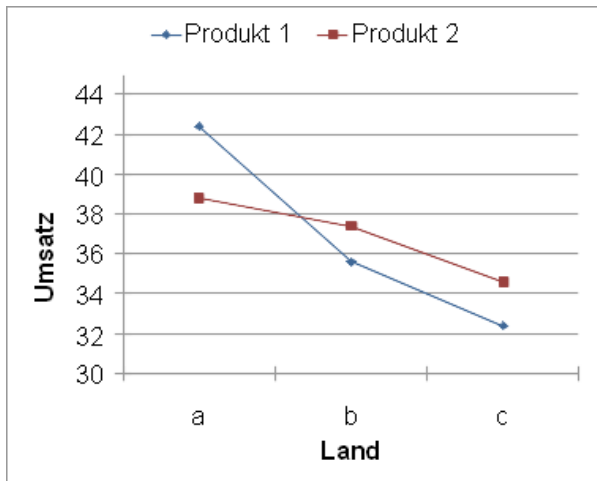
Source	Eta-Squared	df	[95% Conf. Interval]	
Model	.8775277	5	.7156497	.9073096
produkt	.0031056	1	.	.1446765
land	.8558275	2	.7010465	.9019042
produkt#land	.5507348	2	.2201578	.6926828

Note: Eta-Squared values for individual model terms are partial.

Auf der linken unteren Grafik ist zu sehen, dass Produkt 1 im Land a höhere Umsätze erzielt, als Produkt 2. In den anderen beiden Ländern ist es umgekehrt (Interaktion). Für beide Produkte besteht aber ein deutliches Umsatzgefälle von Land a zu Land c.

Das Diagramm auf der rechten Seite zeigt den gleichen Sachverhalt. Die Linien liegen weit auseinander (signifikanter Einfluss von Faktor Land auf den mittleren Umsatz).

Demgegenüber unterscheiden sich die Umsätze beider Produkte 1 und 2 in den Ländern eher gering.



Zweifaktorielle Varianzanalyse - zahnmedizinisches Beispiel (kiga_57.sav)

Der Datensatz ist bekannt aus anderen Beispielen und aus der Rubrik "Milchzahn aktuell".
Frage: Welchen Einfluss haben "Alter" und "Zahnpflege" auf den dmft-Mittelwert?

Faktor Alter (alter) mit 3 Faktorstufen (3,4,5) und

Faktor Zahnpflege (zpfl) mit 3 Faktorstufen (1,2,3).

Zahnpflege 1 - sehr gute Zahnpflege

Zahnpflege 2 - mäßige Zahnpflege

Zahnpflege 3 - schlechte Zahnpflege

Für die mittleren dmft-Werte in Abhängigkeit von Zahnpflege und Alter bekommt man folgende Tabelle:

ZPFL	ALTER		
	3	4	5
1	0.03	0.08	0.03
2	0.66	1.18	1.81
3	7.23	7.84	7.57

Je schlechter die Zahnpflege, desto größer ist der dmft-MW in jeder Altersgruppe.
Innerhalb einer Zahnpflegekategorien sind die Veränderungen weniger ausgeprägt.

Im Ergebnis der Zweifaktoriellen Varianzanalyse hat die Zahnpflege einen deutlich signifikanten Einfluss auf den dmft-MW. Der Faktor Alter ist ebenfalls, aber schwach signifikant. Es gibt eine Wechselwirkung zwischen beiden Faktoren, die aus der Varianz-Tabelle und den untenstehenden Grafiken erkennbar ist.

```
. anova dmft zpfl##alter
```

Number of obs =	2,707	R-squared =	0.4123
Root MSE =	2.53693	Adj R-squared =	0.4105

Source	Partial SS	df	MS	F	Prob>F
Model	12180.538	8	1522.5673	236.57	0.0000
zpfl	9398.7167	2	4699.3584	730.17	0.0000
alter	43.586351	2	21.793176	3.39	0.0340
zpfl#alter	104.87373	4	26.218433	4.07	0.0027
Residual	17364.368	2,698	6.436015		
Total	29544.907	2,706	10.918295		

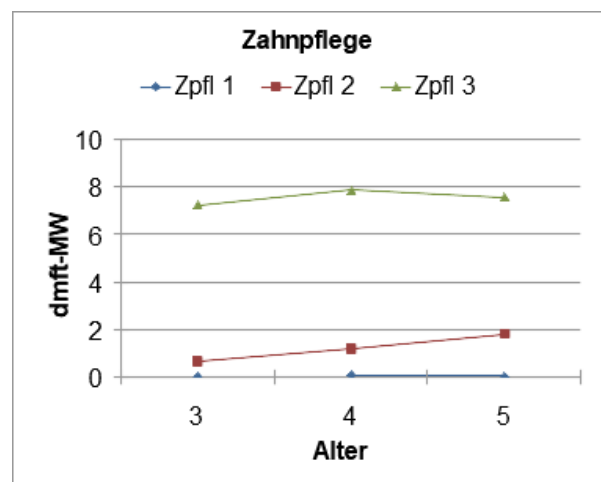
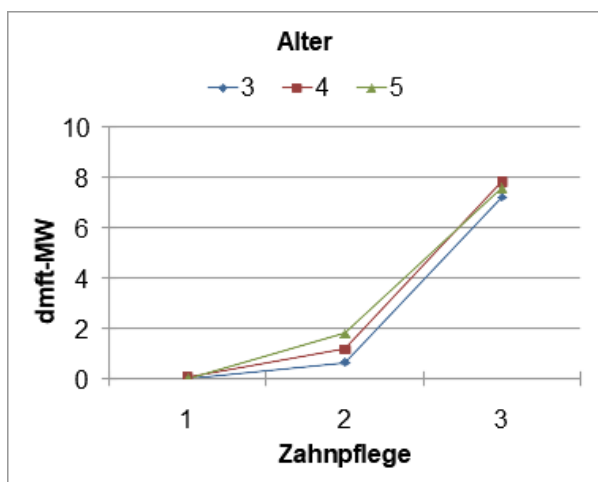
Das partielle Eta² beschreibt die Bedeutung der einzelnen Faktoren im Modell. Man erkennt für den Faktor "alter" und auch für die Wechselwirkung deren vergleichsweise geringen Effekt auf den dmft-MW.

```
. estat esize
Effect sizes for linear models
```

Source	Eta-Squared	df	[95% Conf. Interval]	
Model	.412272	8	.3846952	.4355156
zpfl	.3511821	2	.3241044	.3767616
alter	.0025038	2	.	.0071656
zpfl#alter	.0060033	4	.0008029	.0118167

Auf der linken unteren Grafik ist zu sehen, dass die dmft-Mittelwerte für jedes Alter mit schlechter werdender Zahnpflege ansteigen, insbesondere von Kategorie 2 nach 3. Die Linien liegen eng beieinander und verlaufen praktisch parallel bis auf eine geringe Abweichung bei 4- und 5-Jährigen (Interaktion).

Das Diagramm auf der rechten Seite zeigt den gleichen Sachverhalt. Die Linien liegen weit auseinander (signifikanter Einfluss von Faktor "zpfl" auf den dmft-MW). Demgegenüber unterscheiden sich die dmft-MW innerhalb der Zahnpflegekategorien eher gering.



Eine entsprechende dreidimensionale Darstellung für den Anteil Kinder mit naturgesunden Zähnen findet man in der Rubrik "Milchzahn aktuell".