Grundlagen der Mathematischen Statistik

Testen von Hypothesen
Teil III: ANOVA

Uwe Menzel, 2017
uwe.menzel@matstat.org
www.matstat.org

ANOVA

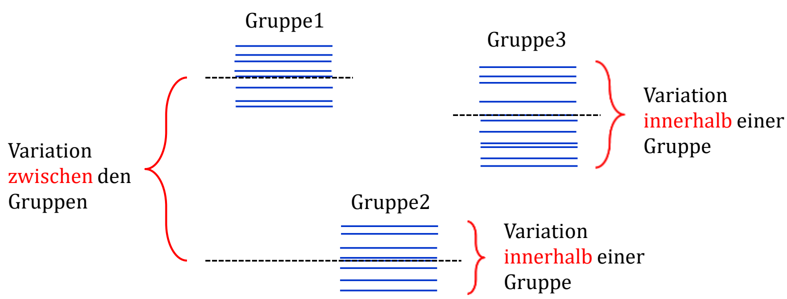
- testet ob mehrere (> 2) normalverteilte Populationen (="Gruppen") den gleichen Erwartungswert haben ("Erweiterung" des t-Tests auf mehr als 2 Gruppen)
- dafür werden die empirischen Varianzen (!) der Populationen ausgewertet (ANOVA = ANalysis Of VAriance)
- Nullhypothese: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (k Stichproben)
- alternative Hypothese: mindestens ein Gleichheitszeichen gilt **nicht**, d.h. mindestens eine Population hat einen abweichenden Erwartungswert
- (ANOVA ist also um eine parametrischer Test.)
- Der Test sagt nichts darüber aus, welcher Erwartungswert abweicht, falls H_0 verworfen wird → dafür kann ein post-hoc-Test benutzt werden.

$X_{n,i} \sim N(\mu_n, \sigma)$ mehr als 2 Populationen

ANOVA

Der Test wird ausgeführt, indem die Variation zwischen den Gruppen mit der Variation innerhalb der Gruppen ”verglichen” wird.

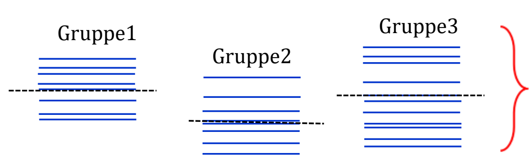
Observationen für drei Gruppen:



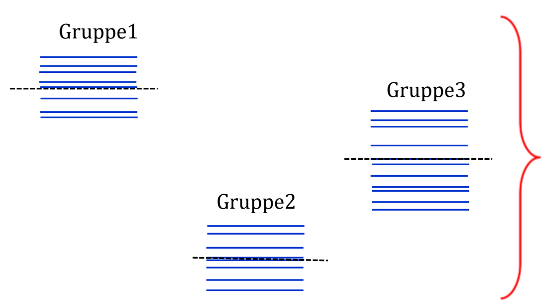
Intuition: Die Gruppen sind verschieden, wenn die Variation zwischen den Gruppen bedeutend größer ist als die Variation innerhalb den Gruppen.

ANOVA

Intuition: Die Gruppen sind verschieden, wenn die Variation zwischen den Gruppen bedeutend größer ist als die Variation innerhalb den Gruppen.



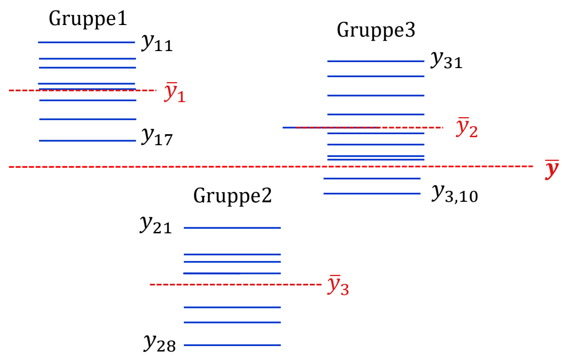
Wahrscheinlich kein signifikanter Unterschied zwischen den Erwartungs-werten (H_0 kann **nicht** verworfen werden), kleine Unterschiede können durch puren Zufall bestehen.



Die Populationen haben wahrscheinlich **nicht** den gleichen Erwartungswert (H_0 wird verworfen), die Unterschiede könnten signifikant sein.

ANOVA: Bezeichnungen

- y_{ij} : Observation j in Gruppe i
  - $i = 1 \ldots k$ (k Gruppen)
  - $j = 1 \ldots n_i$ (n_i Observationen in Gruppe i)
- $\bar{y}_i$: Mittelwert von Gruppe i
- $\bar{y}$: Mittelwert über alle Gruppen (”grand mean”)



ANOVA: ”Total Sum of Squares”

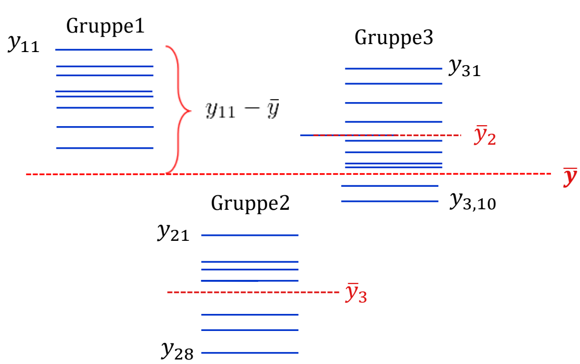
Um die Variation zu quantifizieren, werden Quadratsummen (”Sum of Squares”) angewendet :

$$SS_{tot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$$

y_{ij} : Gruppe i ; Observation j

$\bar{y}$: ”grand mean”

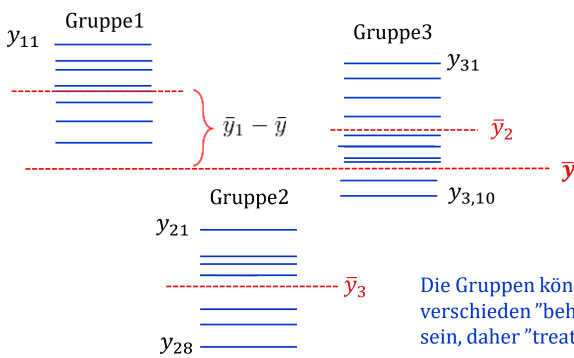
n_i : Anzahl der Observationen in Gruppe i



ANOVA: "Sum of Squares for Treatments"

$$SS_{tr} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2$$

$\bar{y}_i$: Mittelwert der Gruppe i
$\bar{y}$: "grand mean"
n_i : Anzahl der Observationen in Gruppe i



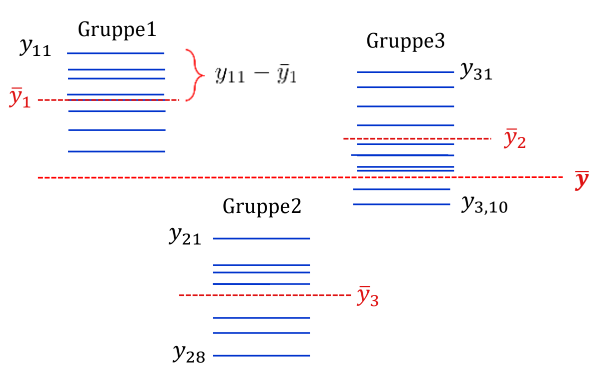
Die Gruppen können verschieden "behandelt" worden sein, daher "treatment".

www.matstat.org

ANOVA: " Sum of Squares for Error"

$$SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$$

y_{ij} : Gruppe i ; Observation j
$\bar{y}_i$: Mittelwert der Gruppe i
n_i : Anzahl der Observationen in Gruppe i



www.matstat.org

ANOVA

Es lässt sich allgemein zeigen, dass folgende Beziehung gilt:

$$SS_{tot} = SS_{tr} + SS_{err}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

SS_{tot} : "Total Sum of Squares"

SS_{tr} : "Sum of Squares for Treatments"

SS_{err} : Sum of Squares for Error

y_{ij} : Gruppe i ; Observation j

$\bar{y}$: "grand mean"

$\bar{y}_i$: Mittelwert der Gruppe i

n_i : Anzahl der Observationen in Gruppe i

$$\sum_{i=1}^{k} n_i = n$$

Totale Anzahl der Observationen

ANOVA: Teststatistik

$$SS_{tr} = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 \qquad \frac{1}{\sigma^2}\cdot SS_{tr} \sim \chi^2(k-1)$$

Chi-Quadrat-Verteilung mit $k-1$ Freiheitsgraden

$$SS_{err} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 \qquad \frac{1}{\sigma^2}\cdot SS_{err} \sim \chi^2(n-k)$$

Chi-Quadrat-Verteilung mit $n-k$ Freiheitsgraden

Allgemein gilt: $$\frac{\frac{\chi^2(n)}{n}}{\frac{\chi^2(m)}{m}} \sim F(n,m)$$

F-Verteilung mit n Zähler-Freiheitsgraden und m Nenner-Freiheitsgraden

Unter Annahme der Nullhypothese gilt daher:

$$F = \frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} \sim F(k-1, n-k)$$

F-Verteilung mit $k-1$ Zähler-Freiheitsgraden und $n-k$ Nenner- Freiheitsgraden

(Unter Annahme der Nullhypothese kommen die $\bar{y}_i$ von gleichen Verteilungen)

ANOVA: Teststatistik

$$F = \frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} \sim F(k-1, n-k)$$

F-Verteilung mit $k-1$ Zähler - Freiheitsgraden und $n-k$ Nenner -Freiheitsgraden

$$SS_{tr} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 \qquad SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

- Die F-Verteilung nimmt als Quotient von Quadratsummen nur positive Werte an.
- Die Testvariable F wird groß, wenn die Abweichungen der Gruppenmittelwerte vom globalen Mittelwert (=Zähler) groß sind im Vergleich zu den Abweichungen innerhalb der Gruppen (=Nenner).
- Für große Observationen von F sollte also die Nullhypothese verworfen werden.
- Wie zuvor für den Z-Test oder t-Test (Vorlesung **F12**) verwerfen wir daher die Nullhypothese auf dem Signifikanzniveau α wenn eine Observation von F das entsprechende α-Quantil von F überschreitet.

ANOVA: Kritische Region

Für ein zuvor festgelegtes Signifikanzniveau α kann der kritische Wert ω_α für die Teststatistik F aus der Forderung berechnet werden, dass die Wahrscheinlichkeit für den Fehler vom Typ I gleich α sein soll:

$$\boxed{P\Bigg(\underbrace{\frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} > \omega_\alpha}_{H_0 \text{ wird verworfen}} \mid H_0 \ wahr\Bigg) = \alpha}$$

Fehler Typ I

Bestimmungsgleichung für ω_α für vorgegebenes α (= Wahrscheinlichkeit für Fehler Typ I)

Unter H_0 ist der linke Term in der Klammer F-verteilt, wir können also schreiben (womit die Bedingung "| H_0 *wahr*""verarbeitet" ist):

$$P(F > \omega_\alpha) = \alpha \quad \text{mit} \quad F \sim F(k-1, n-k)$$

Allgemein gilt: $P(F > F_\alpha(k-1, n-k)) = \alpha$

$F_\alpha(k-1, n-k)$: Quantil für F-Verteilung mit $k-1$ bzw. $n-k$ Freiheitsgraden

Ein Vergleich der beiden letzten Beziehungen ergibt $\omega_\alpha = F_\alpha(k-1, n-k)$.
Die Nullhypothese wird also verworfen, wenn eine Observation von F $(= f_{obs})$ größer als $F_\alpha(k-1, n-k)$ wird.

ANOVA: Kritische Region

H_0 wird verworfen, wenn

$$f_{obs} > F_\alpha(k-1, n-k)$$

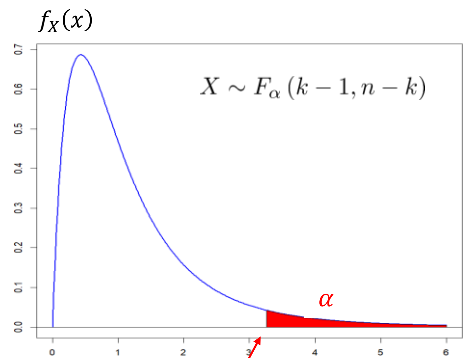
Kritische Region, Signifikanzniveau α

$$\Omega_\alpha = \{f_{obs} > F_\alpha(k-1, n-k)\}$$

Wenn die Observation von F in der kritischen Region (rot) liegt, wird die Nullhypothese verworfen.

Voraussetzungen:

- $X_i \sim N$ (oder ungefähr so)
- $\sigma_i = \sigma$ (oder ungefähr so)
- unabhängige Stichprobenwerte



$F_\alpha(k-1, n-k)$:
Quantil für F-Verteilung, mit $k-1$ bzw. $n-k$ Freiheitsgraden

ANOVA

Manchmal sind nicht alle Observationen y_{ij} verfügbar, sondern nur die Stichprobengrößen, die empirischen Mittelwerte und die Standardabweichungen der jeweiligen Gruppen. In diesem Fall werden die Quadratsummen mit den folgenden Formeln berechnet:

$$\bar{y} = \frac{\sum n_i \cdot \bar{y}_i}{\sum n_i}$$

$$SS_{tr} = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

n_i: Anzahl der Observationen in Gruppe i

$\bar{y}_i$: Mittelwert der Gruppe i

s_i: Standardabweichung der Gruppe i

$$SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2$$

siehe Anhang für Herleitung

ANOVA: Annahmen

- Alle Observationen müssen unabhängig voneinander sein.
- Die Populationen müssen (ungefähr) normalverteilt sein.
  - Der Kolmogorow-Smirnow Test, der Shapiro-Wilk-Test, oder der Anderson-Darling-Test können benutzt werden um zu testen, ob die Verteilungen signifikant von der Normalverteilung abweichen.
- Die Varianzen müssen in allen Gruppen gleich sein (Homoskedastizität, Varianzhomogenität). Hat man die gleiche Anzahl von Observationen in jeder Gruppe, muss dies nur ungefähr erfüllt sein.
  - Der Levene Test oder Bartlett's Test können benutzt werden, um zu testen, ob die Varianzen signifkant verschieden sind.

Welcher Mittelwert weicht ab?

- Ein post-hoc Test kann dies feststellen (wenn H_0 durch ANOVA verworfen wurde), z. B. Tukey's Test.
- Tukey's Test macht paarweise Vergleiche, korrigiert gleichzeitig für multiples Testen.

ANOVA: Beispiel, 4 Gruppen

$\alpha = 0.05$

A	B	C	D
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

$$n_1 = 6\ ; \quad n_2 = 7\ ; \quad n_3 = 6\ ; \quad n_4 = 4$$

$$\bar{y}_1 = 75.67\ ; \quad \bar{y}_2 = 78.43\ ; \quad \bar{y}_3 = 70.83\ ; \quad \bar{y}_4 = 87.75$$

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + n_3 \cdot \bar{y}_3 + n_4 \cdot \bar{y}_4}{n_1 + n_2 + n_3 + n_4} = \frac{1179}{23} = 77.35$$

$$SS_{tr} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = 712.6 \qquad M_{tr} = \frac{SS_{tr}}{k-1} = 237.5$$

$$SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2 = 1196.6 \qquad M_{err} = \frac{SS_{err}}{n-k} = 63.0$$

$$f_{obs} = \frac{M_{tr}}{M_{err}} = \frac{237.5}{63.0} = \underline{\underline{3.77}}$$

$$\Omega_\alpha = \{F > F_\alpha(k-1, n-k)\} = \{F > F_{0.05}(3, 19)\} = \underline{\underline{\{F > 3.13\}}}$$

Die Nullhypothese wird verworfen. Mindestens ein Erwartungswert unterscheidet sich signifikant von den anderen.

uwe.menzel@matstat.org

ANOVA: Beispiel, alternative Berechnung

A	B	C	D
65	75	59	94
87	69	78	89
73	83	67	80
79	81	62	88
81	72	83	
69	79	76	
	90		

$\Rightarrow$

	A	B	C	D
n_i	6	7	6	4
$\bar{y}_i$	75.67	78.43	70.83	87.75
s_i^2	66.67	50.62	91.77	33.58

$$\bar{y} = \frac{n_1 \cdot \bar{y}_1 + n_2 \cdot \bar{y}_2 + n_3 \cdot \bar{y}_3 + n_4 \cdot \bar{y}_4}{n_1 + n_2 + n_3 + n_4} = \frac{1179}{23} = 77.35$$

$$SS_{tr} = \sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = 712.8 \qquad M_{tr} = \frac{SS_{tr}}{k-1} = \frac{712.8}{3} = 237.5$$

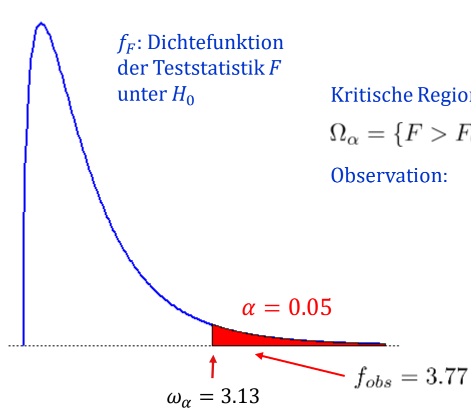
$$SS_{err} = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2 = 1196.66 \qquad M_{err} = \frac{SS_{err}}{n-k} = \frac{1196.66}{19} = 63$$

alternative Formel

$$f_{obs} = \frac{M_{tr}}{M_{err}} = \frac{237.5}{63.0} = \underline{\underline{3.77}}$$

$$\Omega_\alpha = \{F > F_\alpha(k-1, n-k)\} = \{F > F_{0.05}(3,19)\} = \underline{\underline{\{F > 3.13\}}}$$

ANOVA: Beispiel 4 Gruppen



Kritische Region:

$$\Omega_\alpha = \{F > F_{0.05}(3,19)\} = \{F > 3.13\}$$

Observation: $f_{obs} = 3.77$

Die Observation der Teststatistik F überschreitet den kritischen Wert. Die Nullhypothese wird daher verworfen. Mindestens ein Erwartungswert weicht signifikant von den anderen ab. (Signifikansniveau $\alpha = 0.05$)

www.matstat.org

ANOVA mit R

```
data(InsectSprays)
levels(InsectSprays$spray)
summary(InsectSprays$count)
boxplot(count ~ spray, data = InsectSprays, col="green")

# 1. Funktion "oneway.test"
oneway.test(count ~ spray, data = InsectSprays)

# Verschiedene Varianz in den Gruppen?
bartlett.test(count ~ spray, data = InsectSprays)  #  ungleich – Problem!

# Nicht-parametrischer Test:
kruskal.test(count ~ spray, data = InsectSprays)

# 2. alternative Funktion für ANOVA: aov
aov.out = aov(count ~ spray, data = InsectSprays)
summary(aov.out)

TukeyHSD(aov.out)  # post-hoc Test
plot(TukeyHSD(aov.out))  # paarweise Differenzen – signifikanter Unterschied wenn das KI nicht
die Null einschließt.
```


Anhang

Testen von Hypothesen
Teil III: ANOVA

Uwe Menzel, 2018
uwe.menzel@matstat.org
www.matstat.org

Alternative Formel für SS_{err}

$$SS_{err} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

ziehe Summe über i auseinander

$$SS_{err} = \underbrace{\sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)^2}_{(n_1 - 1)\cdot s_1^2} + \underbrace{\sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)^2}_{(n_2 - 1)\cdot s_2^2} + \ldots + \underbrace{\sum_{j=1}^{n_k} (y_{kj} - \bar{y}_k)^2}_{(n_k - 1)\cdot s_k^2}$$

denn $\quad s_i^2 = \frac{1}{n_i - 1} \cdot \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)$

Varianz in Gruppe i. Es wird über alle Observationen der Gruppe summiert (über Index j)

z. B. $\quad s_1^2 = \frac{1}{n_1 - 1} \cdot \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)$

y_{ij} : Gruppe i ; Observation j

$\bar{y}_i$: Mittelwert der Gruppe i

n_i : Anzahl der Observationen in Gruppe i

Wird der letzte Ausdruck für SS_{err} wieder in eine Summe umgewandelt, folgt:

$$SS_{err} = \sum_{i=1}^{k} (n_i - 1) \cdot s_i^2$$

www.matstat.org

F-Test, Verteilung für die Teststatistik

$$\frac{SS_{err}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

ziehe Summe über i auseinander

$$\frac{SS_{err}}{\sigma^2} = \underbrace{\frac{1}{\sigma^2} \cdot \sum_{j=1}^{n_1} (y_{1j} - \bar{y}_1)}_{\sim \chi^2 (n_1 - 1)} + \underbrace{\frac{1}{\sigma^2} \cdot \sum_{j=1}^{n_2} (y_{2j} - \bar{y}_2)}_{\sim \chi^2 (n_2 - 1)} + \ldots + \underbrace{\frac{1}{\sigma^2} \cdot \sum_{j=1}^{n_k} (y_{kj} - \bar{y}_k)}_{\sim \chi^2 (n_k - 1)}$$

$$\Rightarrow \quad \frac{SS_{err}}{\sigma^2} \sim \chi^2 (n - k) \quad \text{denn} \quad \sum_{i=1}^{k} (n_i - 1) = n - k$$

mit $\quad \sum_{i=1}^{k} n_i = n$

$Y_{n,i} \sim N(\mu, \sigma) \quad$ unter H_0

y_{ij} : Gruppe i ; Observation j

$\bar{y}_i$: Mittelwert der Gruppe i

n_i : Anzahl der Observationen in Gruppe i

www.matstat.org

F-Test, Verteilung für die Teststatistik

$$SS_{tot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$$

$$\frac{SS_{tot}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 \sim \chi^2(n-1) \qquad \text{mit} \quad \sum_{i=1}^{k} n_i = n$$

$$\underbrace{\frac{SS_{tot}}{\sigma^2}}_{\sim \chi^2(n-1)} = \frac{SS_{tr}}{\sigma^2} + \underbrace{\frac{SS_{err}}{\sigma^2}}_{\sim \chi^2(n-k)}$$

$$\Rightarrow \quad \frac{SS_{tr}}{\sigma^2} \sim \chi^(k-1)$$

denn werden χ^2-verteilte Zufallsvariablen addiert, so addieren sich deren Freiheitsgrade (χ^2 ist ”reproduktiv”)

www.matstat.org

F-Test, Verteilung für die Teststatistik

$$\frac{SS_{tr}}{\sigma^2} \sim \chi^(k-1)$$

$$F = \frac{SS_{tr}/(k-1)}{SS_{err}/(n-k)} = \frac{\dfrac{SS_{tr}}{\sigma^2\cdot(k-1)}}{\dfrac{SS_{err}}{\sigma^2\cdot(n-k)}} \sim \frac{\dfrac{\chi^2(k-1)}{(k-1)}}{\dfrac{\chi^2(n-k)}{(n-k)}} \sim F\,(k-1, n-k)$$

$$\frac{SS_{err}}{\sigma^2} \sim \chi^2(n-k)$$

www.matstat.org