# Überprüfen von Modellannahmen

von Peter Pfaffelhuber

Version: 27. November 2015

Sowohl bei der Regression, als auch bei der Varianzanalyse, haben wir angenommen, dass verschiedene Stichproben dieselbe Varianz aufweisen, oder sogar alle normalverteilt mit den gleichen Varianzen sind. Um Fehlinterpretationen der statistischen Verfahren auszuschließen, sollte man diese Annahmen überprüfen. Einige Tests, die hierfür zur Verfügung stehen, wollen wir hier vorstellen.

## 1 Gleichheit von Varianzen...

### 1.1 ...bei zwei Stichproben

Seien $X_1, ..., X_m$ unabhängig und nach $N(\mu_X, \sigma_X^2)$ verteilt, sowie $Y_1, ..., Y_n$ unabhängig und nach $N(\mu_Y, \sigma_Y^2)$ verteilt. Wir wollen die Hypothese $H_0 : \sigma_X^2 = \sigma_Y^2$ testen. Glücklicherweise ist dies einfach zu bewerkstelligen, da die empirischen Varianzen unabhängig sind und verteilt sind nach $(m-1)s^2(X)/\sigma_X^2 \sim \chi_{m-1}^2$ und $(n-1)s^2(Y)/\sigma_Y^2 \sim \chi_{n-1}^2$. Daraus ergibt sich bereits der $F$-Test auf ungleiche Varianzen

**$F$-Test auf gleiche Varianzen**

| | |
|---|---|
| Annahme | $X_1, ..., X_m \sim N(\mu_X, \sigma_X^2), Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$ |
| Hypothese | $H_0 : \sigma_X^2 = \sigma_Y^2 \quad$ gegen $\quad H_1 : \sigma_X^2 \neq \sigma_Y^2$ |
| Teststatistik | $F = \frac{s^2(X)}{s^2(Y)} \sim F(m-1, n-1)$ |
| Ablehnungsbereich | $F \in (-\infty, F_{m-1,n-1,\alpha/2}) \cup (F_{m-1,n-1,1-\alpha/2}, \infty)$ |
| $p$-Wert | $2(1 - P_{F(m-1,n-1)}(F)) \wedge 2(1 - P_{F(n-1,m-1)}(1/F))$ |

**Beispiel 1 (Verletzung der Modellannahmen).** Der $F$-Test testet auf Gleichheit zweier Varianzen von normalverteilten Stichproben. Damit lautet

$$H_0 : X \text{ und } Y \text{ sind normalverteilt mit } \sigma_X^2 = \sigma_Y^2.$$

Insbesondere steckt bereits in $H_0$ die Annahme der Normalverteilung der Daten. Wird also die Nullhypothese verworfen werden, so kann dies bedeuten, dass die Normalverteilungsannahme nicht stimmt. Als Veranschaulichung nehmen wir exponentialverteilte Daten und vergleichen deren Varianz mit normalverteilten Daten:

```
> x<-rexp(100)
> y<-rnorm(100)
> var.test(x,y)

F test to compare two variances

data:  x and y
F = 1.5575, num df = 99, denom df = 99, p-value = 0.02854
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.047926 2.314755
sample estimates:
ratio of variances
          1.557463
```

Obwohl die Varianzen der beiden Stichproben gleich sind, wird also $H_0$ aufgrund der unterschiedlichen Verteilung von $X$ und $Y$ auf dem Niveau von 5% abgelehnt.

**Levene- und Brown-Forsythe-Test**

| | |
|---|---|
| Annahme | $(X_{ki})_{k=1,...,p,i=1,...,n_k}$ unabhängig,<br>$(X_{ki})_{i=1,...,n_k}$ identisch verteilt, $k = 1, ..., p$ |
| Hypothese | $H_0 : \mathbb{V}[X_{k1}] = \mathbb{V}[X_{\ell 1}], k, \ell = 1, ..., p$ gegen<br>$H_1 : \mathbb{V}[X_{k1}] \neq \mathbb{V}[X_{\ell 1}]$ für ein Paar $k, \ell$ |
| Teststatistik | $W = \frac{\sum_{k=1}^{p} n_k(\overline{Z}_{k\bullet} - \overline{Z})^2/(k-1)}{\sum_{k=1}^{p}\sum_{i=1}^{n_k}(Z_{ki} - \overline{Z}_{k\bullet})^2/(N-p)} \overset{\text{approx}}{\sim} F(p-1, n-p)$ |
| | $Z_{ki} = \lvert X_{ki} - \overline{X}_{k\bullet}\rvert$, wobei |
| | $\overline{X}_{k\bullet} = \begin{cases} \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}, & \text{Levene-Test} \\ \text{Median von } (X_{ki})_{i=1,...,n_k}, & \text{Brown-Forsythe-Test} \end{cases}$ |
| | $\overline{Z}_{k\bullet} := \frac{1}{n_k}\sum_{i=1}^{n_k} Z_{ki}, \overline{Z} := \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_k} Z_{ki}$ |
| Ablehnungsbereich | $F \in (F_{p-1,n-p,1-\alpha}, \infty)$ |
| $p$-Wert | $1 - P_{F(p-1,n-p)}(F)$ |

### 1.2 ...bei $k$ Stichproben

Liegen nicht zwei, sonder $k$ Stichproben vor (etwa bei einer Varianzanalyse), könnten paarweise $F$-Tests Aufschluss über die Gleichheit der Varianzen geben, aber es gibt auch Alternativen. Oft verwendet werden hier der Levene-Test und der Brown–Forsythe-Test. Diese beschreiben wir lediglich, ohne auf genaue Eigenschaften einzugehen.

**Beispiel 2.** Wir verwenden dieselben simulierten Daten wie in Beispiel 1. Hier wird nun die Hypothese der gleichen Varianzen nicht verworfen. Beim Levene-Test handelt es sich um einen Test, der robuster ist gegen die Verletzung der Modellannahme der Normalverteilung.

```
> library(lawstat)
> x<-rexp(100)
> y<-rnorm(100)
> data = c(x,y)
> group = c(rep(1,100), rep(2, 100))
> levene.test(data, group)

modified robust Brown-Forsythe Levene-type test based on the absolute
deviations from the median

data:  data
Test Statistic = 0.0306, p-value = 0.8613
```

## 2 Testen der Normalverteilungsannahme

Sowohl beim $t$-Test, $\chi^2$-Test, als auch bei der Regression und der Varianzanalyse haben wir die Annahme gemacht, dass die Daten normalverteilt sind. Diese Annahme lässt sich auch testen. Verfahren hierzu werden wir nun besprechen.

### 2.1 QQ-Plots

Eine einfache grafische Möglichkeit, sich einen Eindruck zu verschaffen, ob ein Datensatz von reellwertigen Beobachtungen einer bestimmten Verteilung folgt, sind Plots der Quantile oder QQ-Plots. Hier werden die Quantile der empirischen Verteilung gegen Quantile der zu überprüfenden Verteilung geplottet. Etwa ist das 5%-Quantil der empirischen Verteilung der (oder ein) $y \in \mathbb{R}$, so dass unterhalb von $y$ genau 5% aller Datenpunkte zu finden sind.

In R sind solche QQ-Plots einfach zu bekommen. Hierzu verwenden wir den Datensatz `precip`, der die Niederschlagsmenge (in Zoll) für 70 Städte der USA (und Puerto Rico) angibt.

```
> head(precip)
    Mobile      Juneau     Phoenix Little Rock Los Angeles  Sacramento
      67.0        54.7         7.0        48.5        14.0        17.2
```

Für den QQ-Plot gibt es den Befehl

```
> qqnorm(precip),
```

der die empirischen Quantile gegen die einer Standardnormalverteilung plottet; siehe Abbildung 1.

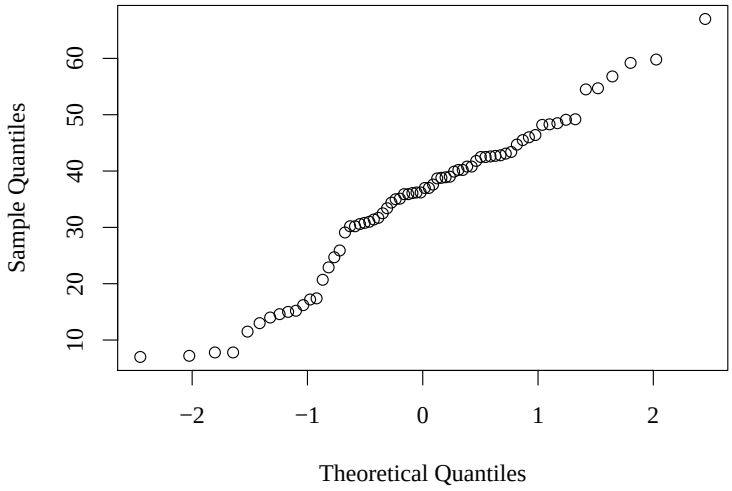


**Abbildung 1:** QQ-Plot der `precip`-Daten.

## 2.2 Der Kolmogorov-Smirnov-Test

Natürlich ist es gut, nicht nur einen grafischen Eindruck der möglichen Abweichung der Normalverteilungsannahme zu haben, sondern auch einen statistischen Test. Mit dem hier vorgestellten Kolmogorov-Smirnov-Test kann man testen, ob Daten einer beliebigen, vorgegebenen, stetigen Verteilung folgen. Er basiert auf der empirischen Verteilung der Stichprobe.

**Definition 3 (Empirische Verteilung).** *Sei $X = (X_1, ..., X_n)$ ein Vektor von Zufallsgrößen. Die empirische Verteilung von $X$ ist gegeben als*

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}.$$

*Sind die Zufallsvariablen reellwertig, dann ist die empirische Verteilnugsfunktion die Verteilungsfunktion der empirischen Verteilung und gegeben als*

$$t \mapsto S_n(t) := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(-\infty; t] = \frac{1}{n}\sum_{i=1}^{n}1_{X_i \leq t}.$$

**Bemerkung 4 (Satz von Glivenko-Cantelli).** Aus der Vorlesung Wahrscheinlichkeitstheorie ist bekannt: Sind $X, X_1, X_2, ...$ unabhängige und identisch verteilte Zufallsgrößen mit Verteilungsfunktion $F_X$. Dann gilt

$$D_n := \sup_{t\in\mathbb{R}} |S_n(t) - F_X(t)| \xrightarrow{n\to\infty}_{fs} 0.$$

Um dies einzusehen, sei bemerkt, dass $1_{X_1\leq t}, 1_{X_2\leq t}, ...$ unabhängig und identisch verteilt sind mit $\mathbb{E}[1_{X_1\leq t}] = \mathbb{P}(X_1 \leq t) = F_X(t)$. Damit ist mit dem Gesetz der großen Zahlen zumindest erklärt, warum $S_n(t) - F_X(t) \xrightarrow{n\to\infty}_{fs} 0$ für jedes feste $t$ gilt.

**Bemerkung 5 (Verteilung von $F_X(X_{(i)})$).** Sei $X$ eine Zufallsvariable mit Dichte und habe Verteilungsfunktion $F_X$.

1. Es ist $F_X(X) \sim U[0,1]$.
   Denn: Fast sicher ist $X$ so, dass $F_X^{-1}(X)$ existiert. Daraus folgt
   $$\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t)) = F_X(F_X^{-1}(t) = t.$$

2. Seien $X, X_1, ..., X_n$ unabhängig und identisch verteilt und $U_1, ..., U_n \sim U([0,1])$ unabhängig. Dann gilt $F_X(X_{(i)}) \sim U_{(i)}$.
   Denn: Genau wie oben ist $X_{(i)}$ fast sicher so, dass $F_X^{-1}(X_{(i)})$ existiert. Nun ist
   $$\begin{aligned}\mathbb{P}(F_X(X_{(i)}) \leq t) &= \mathbb{P}(X_{(i)} \leq F_X^{-1}(t)) = \mathbb{P}(X_j \leq F_X^{-1}(t) \text{ für } i \text{ verschiedene } j) \\ &= \mathbb{P}(U_j \leq t \text{ für } i \text{ verschiedene } j) = \mathbb{P}(U_{(i)} \leq t).\end{aligned}$$

**Proposition 6 (Verteilungsfreiheit von $D_n$).** *Sei $(X, \{\mathbb{P}_\theta : \theta \in \mathcal{P}\})$ ein reguläres, stetiges statistisches Modell. Dann ist für jedes $t \in \mathbb{R}$ die Statistik $D_n(t)$ verteilungsfrei.*

*Beweis.* Seien $X_{(1)}, ..., X_{(n)}$ die Ordnungsstatistiken von $X_1, ..., X_n$ sowie $X_{(0)} := -\infty$ und $X_{(n+1)} := \infty$. Dann ist
$$S_n(t) = \frac{i}{n} \text{ für } X_{(i)} \leq t < X_{(i+1)}.$$
Wir schreiben nun
$$\begin{aligned}D_n &= \sup_{t\in\mathbb{R}} |S_n(t) - F_X(t)| = \max_{1\leq i\leq n} \sup_{X_{(i)}\leq t<X_{(i+1)}} |S_n(t) - F_X(t)| \\ &= \max_{1\leq i\leq n} \sup_{X_{(i)}\leq t<X_{(i+1)}} \left|\frac{i}{n} - F_X(t)\right| \\ &= \max_{1\leq i\leq n} \max\left(\left|\frac{i}{n} - F_X(X_{(i)})\right|, \left|\frac{i}{n} - F_X(X_{(i+1)})\right|\right).\end{aligned}$$
Damit ist gezeigt, das $D_n$ nur von $F_X(X_{(0)}), ..., F_X(X_{(n+1)})$ abhängt. Diese Größen haben nach Bemerkung 5 dieselbe Verteilung wie die Ordnungsstatistiken eines $U(0,1)$-verteilten Vektors von Zufallsvariablen, und zwar unabhängig von $F_X$. Daraus folgt die Behauptung. □

**Kolmogorov-Smirnov-Test**

| | |
|---|---|
| Annahme | $X_1, ..., X_n$ reellwertig, unabhängig und stetig identisch verteilt |
| Hypothese | $H_0 : X_i$ hat Verteilungsfunktion $F_X$ gegen<br>$H_1 : X_i$ hat eine andere Verteilungsfunktion |
| Teststatistik | $D_n := \sup_{t\in\mathbb{R}} \lvert S_n(t) - F_X(t)\rvert$<br>$S_n(t) := \frac{1}{n}\sum_{i=1}^n 1_{X_i\leq t}$<br>Verteilung $(D_n)_*\mathbb{P}$ von $D_n$ ist in Theorem 7 angegeben |
| Ablehnungsbereich | $D_n >$ $(1-\alpha)$-Quantil von $(D_n)_*\mathbb{P}$ |
| $p$-Wert | $(D_n)_*\mathbb{P}((D_n, \infty))$ |

**Theorem 7 (Verteilung von $D_n$).** *Sei $X, X_1, .., X_n$ unabhängig und identisch verteilt mit Dichte sowie $F_X$ die Verteilungsfunktion von $X$. Dann gilt für $0 < s < (2n-1)/(2n)$*

$$\mathbb{P}\Big(D_n < \frac{1}{2n} + s\Big) = n! \int_{1/(2n)-s}^{1/(2n)+s} \int_{2/(2n)-s}^{2/(2n)+s} \cdots \int_{(2n-1)/(2n)-s}^{(2n-1)/(2n)+s} 1_{0<u_1\cdots<u_n<1} du_n \cdots du_1.$$

*Beweis.* Zunächst bemerken wir, dass immer $D_n \geq 1/2n$ gilt, da $F_X$ stetig ist, $S_n$ aber Sprünge der Größe $1/n$ macht. OBdA nehmen wir wegen der Verteilungsfreiheit von $D_n$ an, dass $F_X(x) = x$, d.h. $X \sim U([0,1])$. Wir schreiben mit $s' := \frac{1}{2n} + s$

$$\begin{aligned}
\mathbb{P}(D_n < s') &= \mathbb{P}(\sup_{t\in[0,1]} |S_n(t) - t| < s') \\
&= \mathbb{P}\Big(\Big|\frac{i}{n} - t\Big| < s' \text{ für alle } X_{(i)} \leq t < X_{(i+1)}, \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < t < \frac{i}{n} + s' \text{ für alle } X_{(i)} \leq t < X_{(i+1)}, \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i}{n} + s', \frac{i}{n} - s' < X_{(i+1)} < \frac{i}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i}{n} + s', \frac{i-1}{n} - s' < X_{(i)} < \frac{i-1}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{i}{n} - s' < X_{(i)} < \frac{i-1}{n} + s' \text{ für alle } i = 1, ..., n\Big) \\
&= \mathbb{P}\Big(\frac{2i-1}{2n} - s < X_{(i)} < \frac{2i-1}{2n} + s \text{ für alle } i = 1, ..., n\Big).
\end{aligned}$$

Daraus folgt die Behauptung, da die gemeinsame Verteilung von $X_{(1)}, ..., X_{(n)}$ die Dichte $n!1_{0\leq u_1<\cdots<u_n}$ hat. □

**Beispiel 8 (Der Kolmogorov-Smirnov-Test für $t$-verteilte Daten).** Es ist bekannt, dass die $t$-Verteilung mit $k$ Freiheitsgraden für große $k$ gegen $N(0,1)$ konvergiert. Wir wollen nun testen, ob der Unterschied der $t$-Verteilung mit $k = 10$ Freiheitsgraden und der $N(0,1)$-Verteilung erkennbar ist. Wir verwenden hierzu verschiedene Stichprobengrößen. Es ergibt etwa

```
> data = rt(1000, df=10)
> ks.test(data, "pnorm")
One-sample Kolmogorov-Smirnov test

data:  data
D = 0.0348, p-value = 0.178
alternative hypothesis: two-sided
```

also kann in dieser Stichprobe der Größe 1000 die Normalverteilungsannahme nicht verworfen werden. In einer deutlich größeren Stichprobe allerdings schon, wie wir nun sehen.

```
> data = rt(10000, df=10)
> ks.test(data, "pnorm")
One-sample Kolmogorov-Smirnov test

data:  data
D = 0.0207, p-value = 0.0003925
alternative hypothesis: two-sided
```

## 2.3 Der Lilliefour-Test

Will man prüfen, ob ein Datensatz einer Normalverteilung folgt, so kennt man zunächst die Parameter $\mu$ und $\sigma^2$ nicht. Deshalb ist es nicht möglich, den Kolmogorov-Smirnov-Test direkt anzuwenden, da man nicht weiß, gegen welche Verteilung genau getestet werden soll. Es liegt nun nahe, zunächst $\mu$ und $\sigma^2$ etwa durch $\overline{x}$ und $s^2(x)$ aus den Daten zu testen und anschließend die Normalverteilungsannahme dadurch zu überprüfen, ob die Daten $x$ einer $N(\overline{x}, s^2(x))$-Verteilung folgen. Allerdings verändert sich durch das Schätzen der Modellparameter aus den Daten die Verteilung der Teststatistik $D_n$. Die neue Verteilung von $D_n$ kann man mittels Simulation ermitteln.