Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine-Universität Düsseldorf

12. Januar 2011

Vergleich zweier Mittelwerte

Zwei Versuchsreihen liefern Messergebnisse. Der Test soll entscheiden, ob sich diese Ergebnisse signifikant unterscheiden.

Unverbundene Stichproben: Die Messerergebnisse werden an verschiedenen Populationen gewonnen.

Beispiel: 9 Maisfelder werden mit einem Bodenbakterium behandelt, 10 weitere bleiben unbehandelt. Bei allen wird der Befall mit Maiszünsler bestimmt.

Verbundene Stichproben: Beide Messungen werden an derselben Population wird unter identischen Bedingungen durchgeführt.

Beispiel: Bei 10 Patienten mit Bluthochdruck wird der Blutdruck vor und nach einer Therapie bestimmt.

Vergleich zweier Mittelwerte bei verbundenen Stichproben

- Beispiel “Blutdrucksenker”: Bei 10 Patienten mit unbehandeltem Bluthochdruck wird der Blutdruck vor und nach einer Therapie gemessen.
- Senkt die Therapie den Blutdruck?
- Es handelt sich um verbundene Stichproben.

t-Test für verbundene Stichproben

- Die Zufallsvariablen $X_1, \ldots, X_n$ und $Y_1, \ldots, Y_n$ sind verbunden, d. h. X_1 und Y_1 bezeichnen Messwerte des ersten Individuums unter verschiedenen Versuchsbedingungen, entsprechend für die anderen Individuen
- Für jedes j setze $Z_j = Y_j - X_j$
- Verteilungsvoraussetzungen sind:
  - Alle X_j besitzen denselben Erwartungswert μ_1
  - Alle Y_j besitzen denselben Erwartungswert μ_2
  - Die Differenzen Z_j sind normalverteilt mit unbekanntem Erwartungswert $\mu_2 - \mu_1$ und unbekannter Varianz σ^2
- Ziel: μ_1 und μ_2 sollen verglichen werden

t-Test für verbundene Stichproben, Fortsetzung

- x_j, y_j und z_j seien Realisierungen.
- Bestimme arithmetischen Mittelwert und Stichprobenstreuung der z-Daten

$$\overline{z} = \frac{1}{n}\sum_{j=1}^{n} z_j \qquad \text{und} \qquad s_z = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(z_j - \overline{z})^2}$$

- Beim Test für verbundene Stichproben ist die Teststatistik

$$t = \frac{\overline{z}}{s_z}\sqrt{n}$$

t-Test für verbundene Stichproben, Fortsetzung

- Das Signifikanzniveau sei α
- Bestimme zugehörige Quantile der t-Verteilung

$t_{n-1,1-\alpha/2}$	beim zweiseitigen Test
$t_{n-1,1-\alpha}$	bei einem einseitigen Test

- Entscheidung

  $H_0 = \{\mu_1 = \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn
  $$|t| > t_{n-1,1-\alpha/2}$$
  $H_0 = \{\mu_1 \leq \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn
  $$t < -t_{n-1,1-\alpha}$$
  $H_0 = \{\mu_1 \geq \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn
  $$t > t_{n-1,1-\alpha}$$

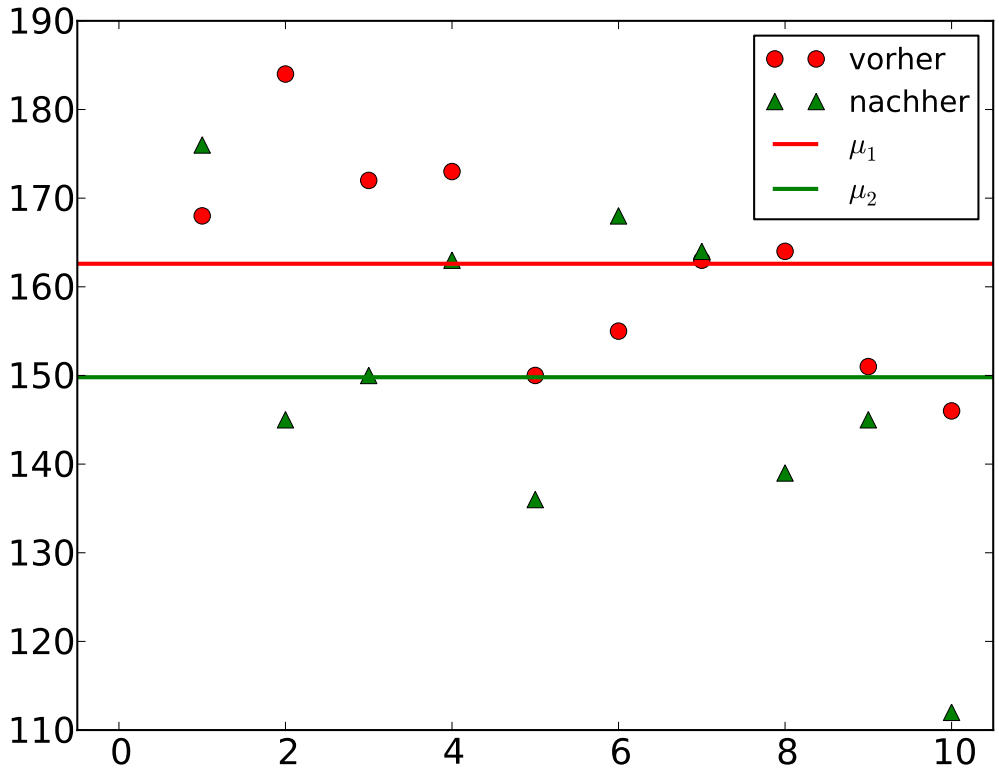
Verbundene Stichproben, Beispiel

- Bei 10 Patienten mit unbehandeltem Bluthochdruck wurde eine Therapie erprobt. Dazu wurde ihr Blutdruck vor und nach der Therapie gemessen

Blutdruck [mm hg]	1	2	3	4	5
vorher	168	184	172	173	150
nachher	176	145	150	163	136
Blutdruck [mm hg]	6	7	8	9	10
vorher	155	163	164	151	146
nachher	168	164	139	145	112

- Ist zum Signifikanzniveau $\alpha = 0.05$ sichergestellt, dass die Therapie den Blutdruck senkt?

Datensatz des Beispiels



Beispiel, Fortsetzung

- $X_1, \ldots, X_{10}$ sind die Daten vor und $Y_1, \ldots, Y_{10}$ die Daten nach der Therapie.
- Nullhypothese $H_0 = \{\mu_1 \leq \mu_2\}$
- Realisierungen der Z_j

1	2	3	4	5	6	7	8	9	10
8	-39	-22	-10	-14	13	1	-25	-6	-34

- Arithmetisches Mittel und Stichprobenstreuung

$$\bar{z} = -12.80 \qquad \text{und} \qquad s_z = 17.36$$

- Teststatistik

$$t = \frac{\bar{z}}{s_z} \cdot \sqrt{n} = \frac{-12.80}{17.36} \cdot \sqrt{10} = -2.332$$

- Quantil $t_{9,\,0.95} = 1.833$
- Also $t < -t_{9,\,0.95}$. Daher wird die Nullhypothese abgelehnt
- Die Therapie hat ihre Wirksamkeit gezeigt

Zusammenhang zum t-Test für Erwartungswerte

- Die Frage, ob $\mu_2 \geq \mu_1$ ist gleichwertig zur Frage, ob $E(Z_j) \geq 0$
- Daher ist der Vergleichstest für verbundene Stichproben eine Variante des t-Tests für Erwartungswerte

Unverbundene Stichproben

- Ein Versuch wird n_1-mal durchgeführt
- Ein Parameter wird geändert
- Der Versuch wird mit dem geänderten Parameter n_2-mal **mit einem anderen Kollektiv** wiederholt
- Die Messergebnisse werden verglichen
- Da die Stichproben unverbunden sind, ist $n_1 \neq n_2$ möglich

t-Test zum Vergleich zweier Mittelwerte bei unverbundenen Stichproben

- Gegeben sind Zufallsvariable $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$
- Verteilungsvoraussetzungen sind
  - Alle X_j sind normalverteilt mit Erwartungswert μ_1 und unbekannter Varianz σ^2
  - Alle Y_j sind normalverteilt mit Erwartungswert μ_2 und unbekannter Varianz σ^2
  - Die beiden Varianzen müssen also gleich sein
- Ziel: μ_1 und μ_2 sollen verglichen werden

t-Test für unverbundene Stichproben, Fortsetzung

- x_j und y_j seien Realisierungen
- Bestimme arithmetische Mittelwerte

$$\overline{x} = \frac{1}{n_1} \sum_{j=1}^{n_1} x_j \qquad \overline{y} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_j$$

- und Stichprobenstreuungen

$$s_x = \sqrt{\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_j - \overline{x})^2} \qquad s_y = \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \overline{y})^2}$$

- Bestimme die *Standardabweichung der gepoolten Stichproben*

$$s_p = \sqrt{\frac{(n_1 - 1) \cdot s_x^2 + (n_2 - 1) \cdot s_y^2}{n_1 + n_2 - 2}}$$

- Die Teststatistik ist

$$t = \frac{\overline{x} - \overline{y}}{s_p} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

Unverbundene Stichproben, Fortsetzung

- Das Signifikanzniveau sei α
- Bestimme zugehörige Quantile der t-Verteilung

$t_{n_1+n_2-2,\,1-\alpha/2}$	beim zweiseitigen Test
$t_{n_1+n_2-2,\,1-\alpha}$	bei einem einseitigen Test

- Entscheidung

$H_0 = \{\mu_1 = \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn

$$|t| > t_{n_1+n_2-2,\,1-\alpha/2}$$

$H_0 = \{\mu_1 \geq \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn

$$t < -t_{n_1+n_2-2,\,1-\alpha}$$

$H_0 = \{\mu_1 \leq \mu_2\}$: Die Nullhypothese H_0 wird abgelehnt, wenn

$$t > t_{n_1+n_2-2,\,1-\alpha}$$

Alternative Formel für die Standardabweichung der gepoolten Stichproben

$$
\begin{aligned}
s_p &= \sqrt{\frac{(n_1 - 1) \cdot s_x^2 + (n_2 - 1) \cdot s_y^2}{n_1 + n_2 - 2}} \\
&= \sqrt{\frac{1}{n_1 + n_2 - 2} \left(\sum_{j=1}^{n_1} (x_j - \overline{x})^2 + \sum_{j=1}^{n_2} (y_j - \overline{y})^2 \right)}
\end{aligned}
$$

Beispiel: Maiszünsler

- Der Maiszünsler soll mit einem Bodenbakterium bekämpft werden
- Die folgenden Befallraten (in Larven pro Quadratmeter) wurden beobachtet:

Unbehandelte Felder

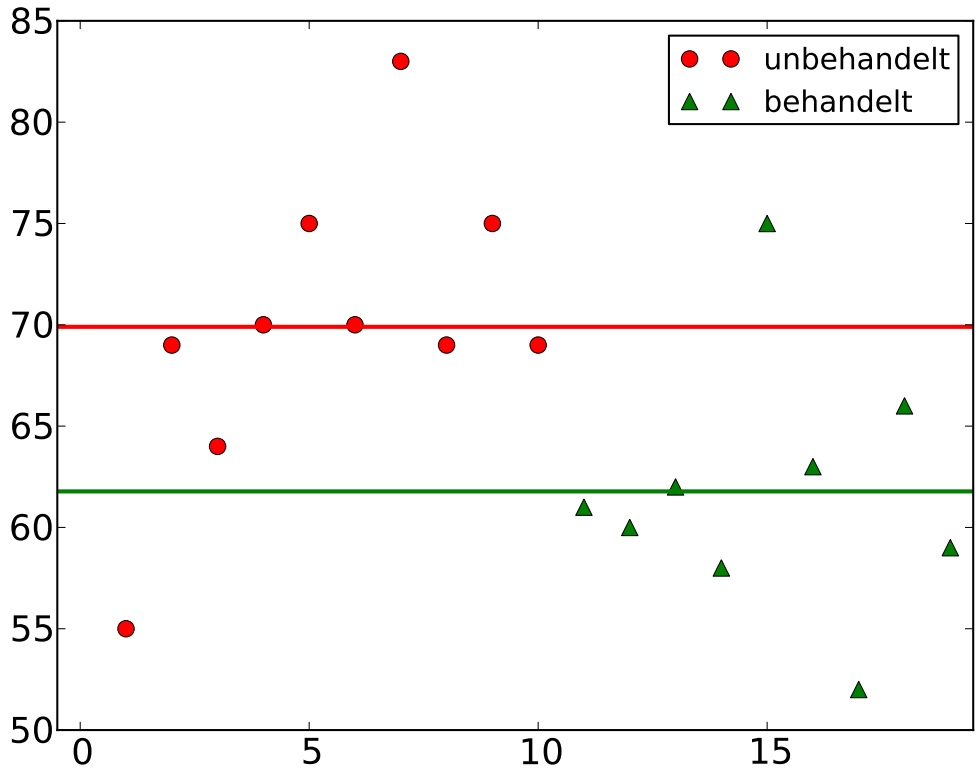
55	69	64	70	75	70	83	69	75	69

Behandelte Felder

61	60	62	58	75	63	52	66	59

- $n_1 = 10,\ n_2 = 9$
- Da die Stichproben unverbunden sind, sind unterschiedliche Umfänge der beiden Stichproben zulässig

Datensatz des Beispiels



Maiszünsler, Fortsetzung

- Zum Signifikanzniveau $\alpha = 0.05$ soll nachgewiesen werden, dass die Behandlung mit einem Bodenbakterium den Befall mit Maiszünsler verringert
- $X_1 \ldots, X_{10}$ sind die Werte für die unbehandelten Felder, $Y_1, \ldots, Y_9$ die Werte für die behandelten
- μ_1 ist der Erwartungswert der X_i, μ_2 der Erwartungswert der Y_i
- Die Nullhypothese ist, dass die Behandlung keinen Nutzen bringt, also

$$H_0 = \{\mu_1 \leq \mu_2\}$$

Maiszünsler, Fortsetzung

- Benötigtes Quantil der t-Verteilung

$$t_{17,\,0.95} = 1.740$$

- Die arithmetischen Mittel und die Stichprobenstreuungen betragen

$$\overline{x} = 69.00 \qquad s_x = 7.972$$
$$\overline{y} = 61.78 \qquad s_y = 6.280$$

- Die Streuungen der X_i und der Y_i müssen gleich sein. Wir gehen mal davon aus, dass diese Voraussetzung erfüllt ist. (Auch dafür gibt es einen Test.)
- Die Standardabweichung der gepoolten Stichproben beträgt

$$s_p = 7.226$$

Maiszünsler, Fortsetzung

- Damit kann man die Teststatistik ausrechnen

$$t = 2.175$$

- Das ist größer als $t_{17,\,0.95} = 1.740$, also wird die Nullhypothese abgelehnt
- Die Behandlung mit dem Bodenbakterium verringert den Befall mit Maiszünsler
- Der p-Wert ist 2.2%
- Dazu benötigt man allerdings eine Tabelle der t_{17}-Verteilung

Verteilungsannahmen

- Alle bisherigen Tests verwenden Verteilungsannahmen
- Entweder waren alle Zufallsvariablen normalverteilt oder binomialverteilt
- In der Praxis ist oft nicht klar, ob diese Voraussetzungen erfüllt sind
- Tests, die auch bei Verletzung der Verteilungsannahmen noch gute Ergebnisse liefern, heißen *konservativ*
- Der t-Test ist konservativ

Q-Q-Plot

- Mit dem Quantil-Quantil-Plot kann man auf graphischem Wege beurteilen, ob Messwerte Realisierungen einer normalverteilten Zufallsvariablen sind
- Man trägt dazu auf der x-Achse die Quantile der Standardnormalverteilung und auf der y-Achse die Quantile der Beobachtungsdaten auf
- Wenn diese Punkte annähernd auf einer Geraden liegen, sind die Daten näherungsweise normalverteilt, ansonsten nicht

Q-Q-Plot: Vorgehensweise

- Gegeben n verschiedene Messwerte
- Ordne sie der Reihe nach an

$$x_1 < x_2 < \cdots < x_n$$

- Interpretiere x_j als $\frac{j-\frac{1}{2}}{n}$-Quantil des Datensatzes
- j-ter Datenpunkt im Q-Q-Plot:

$$\begin{aligned} x\text{-Koordinate} &: \frac{j-\frac{1}{2}}{n}\text{-Quantil der Standardnormalverteilung} \\ y\text{-Koordinate} &: x_j \end{aligned}$$

- Liegen diese Punkte annähernd auf einer Geraden?
- Wenn ja, dann ist die Normalverteilungsannahme gerechtfertigt

Q-Q-Plot: Beispiel

- Wir legen Daten aus dem Beispiel “Blutdrucksenker” zu Grunde

  168 184 172 173 150 155 163 164 151 146

- Zur Bestimmung der Quantile ordnen wir sie der Größe nach an

  146 150 151 155 163 164 168 172 173 184

- Benötigt: Die Quantile $q_{0.05}, q_{0.15}, q_{0.25}, \ldots, q_{0.95}$ der Standardnormalverteilung
- Von diesen ist nur $q_{0.95} = 1.645$ explizit tabelliert, die anderen muss man in der Tabelle aufsuchen

$q_{0.05}$	$q_{0.15}$	$q_{0.25}$	$\ldots$	$q_{0.75}$	$q_{0.85}$	$q_{0.95}$
-1.645	-1.04	-0.675	$\ldots$	0.675	1.04	1.645

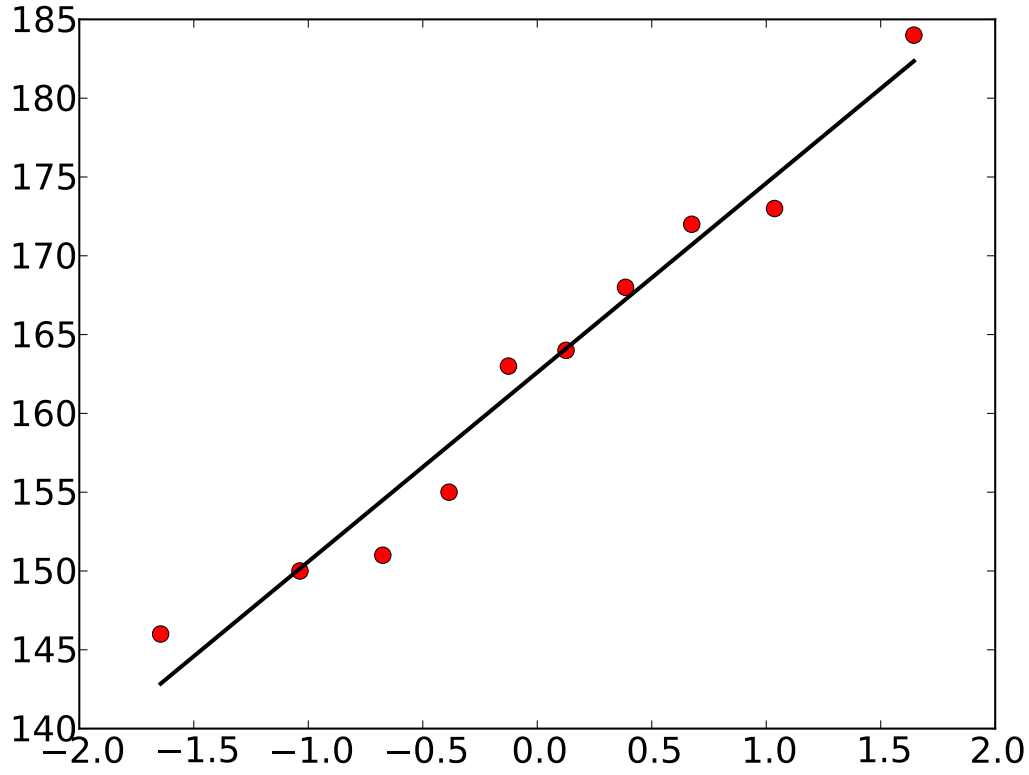
Tabelle der Standard-Normalverteilung, linke Seite

u	0.00	0.01	0.02	0.03	0.04
0.0	0,500000	0,503989	0,507978	0,511966	0,515953
0.1	,539828	,543795	,547758	,551717	,555670
0.2	,579260	,583166	,587064	,590954	,594835
0.3	,617911	,621720	,625516	,629300	,633072
0.4	,655422	,659097	,662757	,666402	,670031
0.5	,691462	,694974	,698468	,701944	,705401
0.6	,725747	,729069	,732371	,735653	,738914
0.7	,758036	,761148	,764238	,767305	,770350
0.8	,788145	,791030	,793892	,796731	,799546
0.9	,815940	,818589	,821214	,823814	,826391
1.0	,841345	,843752	,846136	,848495	,850830
1.1	,864334	,866500	,868643	,870762	,872857
1.2	,884930	,886861	,888768	,890651	,892512
1.3	,903200	,904902	,906582	,908241	,909877
1.4	0,919243	0,920730	0,922196	0,923641	0,925066

Tabelle der Standard-Normalverteilung, rechte Seite

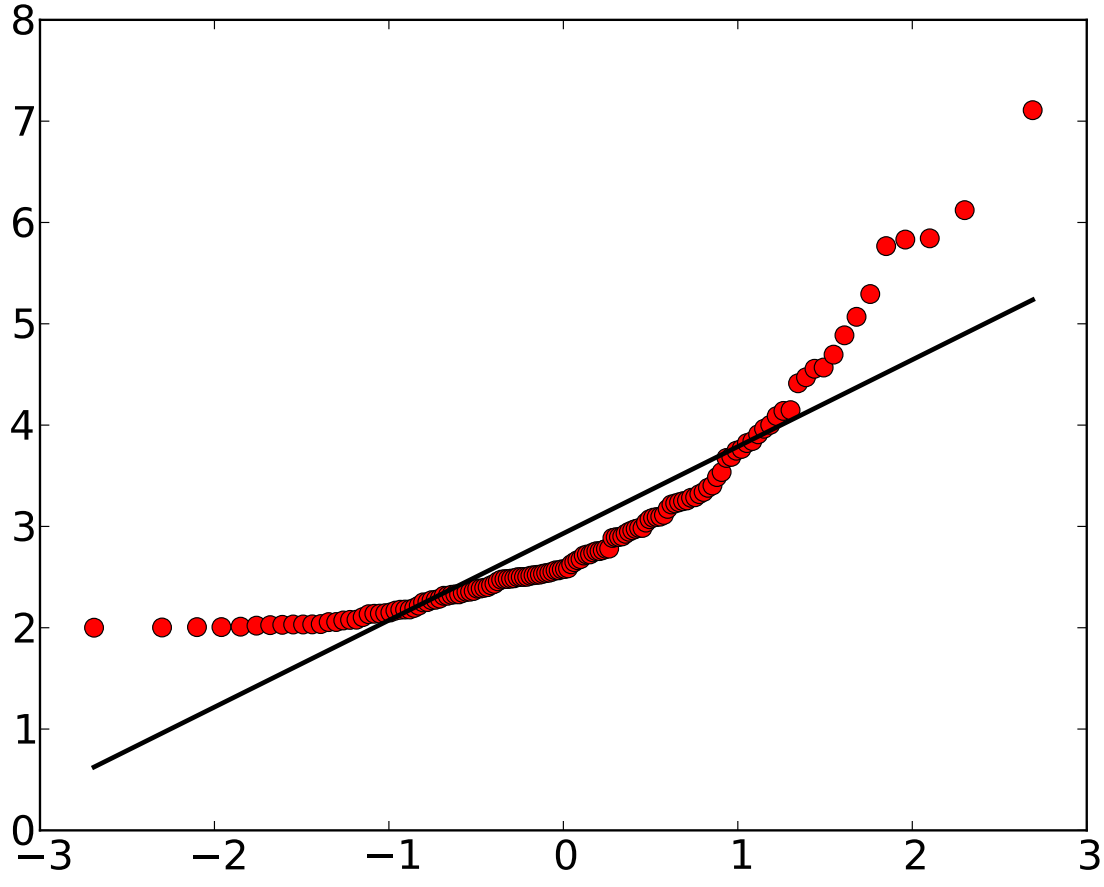
u	0.05	0.06	0.07	0.08	0.09
0.0	0,519939	0,523922	0,527903	0,531881	0,535856
0.1	,559618	,563559	,567495	,571424	,575345
0.2	,598706	,602568	,606420	,610261	,614092
0.3	,636831	,640576	,644309	,648027	,651732
0.4	,673645	,677242	,680822	,684386	,687933
0.5	,708840	,712260	,715661	,719043	,722405
0.6	,742154	,745373	,748571	,751748	,754903
0.7	,773373	,776373	,779350	,782305	,785236
0.8	,802337	,805105	,807850	,810570	,813267
0.9	,828944	,831472	,833977	,836457	,838913
1.0	,853141	,855428	,857690	,859929	,862143
1.1	,874928	,876976	,879000	,881000	,882977
1.2	,894350	,896165	,897958	,899727	,901475
1.3	,911492	,913085	,914657	,916207	,917736
1.4	0,926471	0,927855	0,929219	0,930563	0,931888

Q-Q-Plot

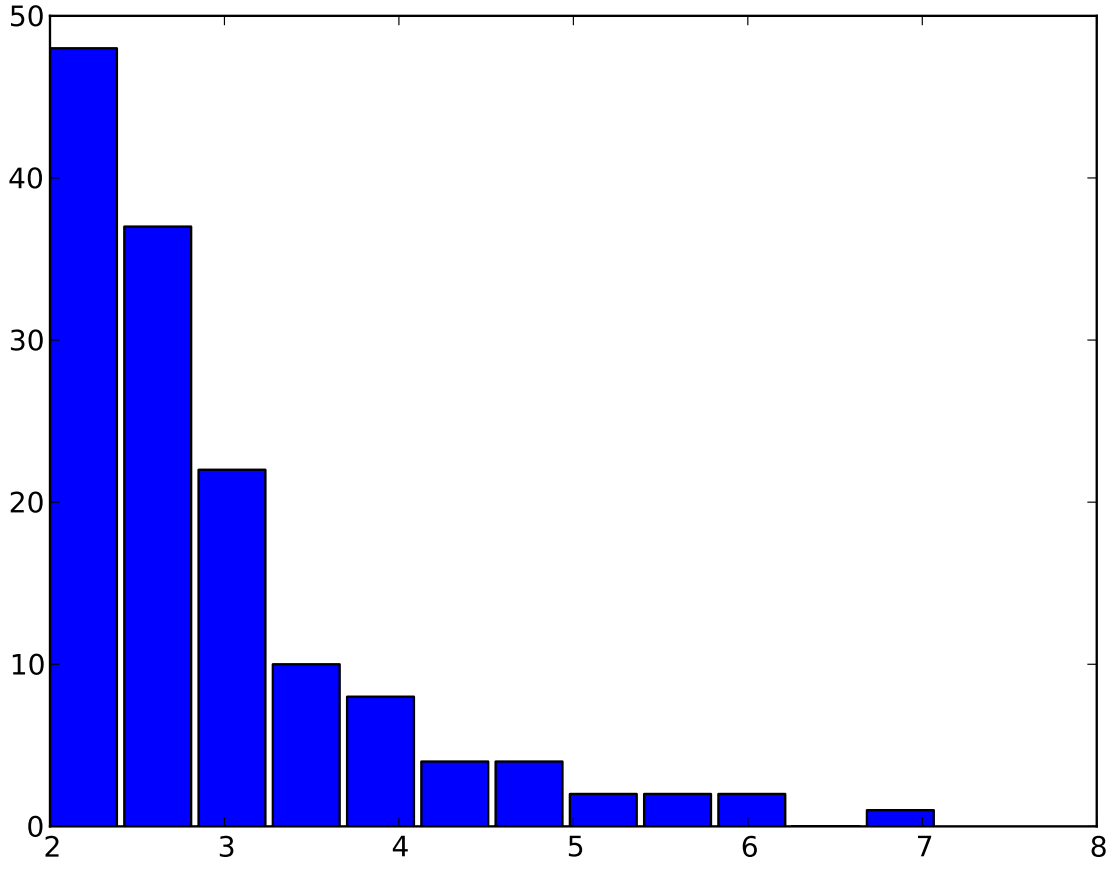


Der Q-Q-Plot der Blutdruckdaten zeigt, dass die Normalverteilungsannahme gerechtfertigt war

Q-Q-Plot von exponentialverteilten Daten



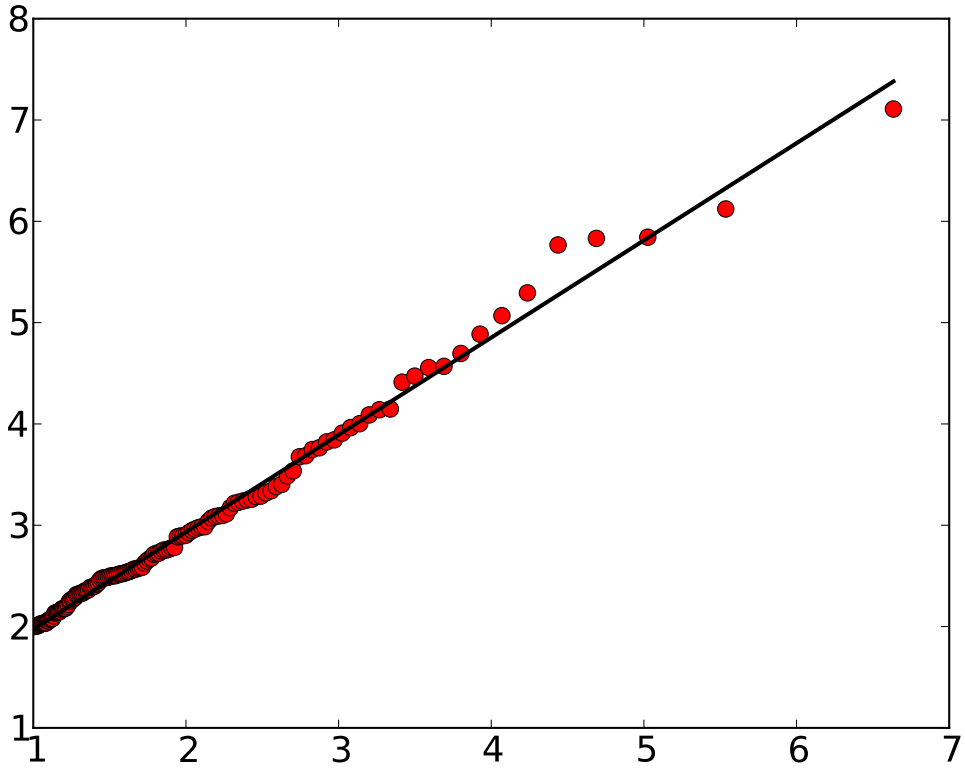
Histogramm der exponentialverteilten Daten



Q-Q-Plot ist Besonderheit der Normalverteilungen

- Man könnte daran denken, z. B. eine Exponentialverteilungsannahme zu überprüfen, indem man im Q-Q-Plot anstelle der Quantile der Standardnormalverteilung die Quantile der Exponentialverteilung benutzt
- **Das funktioniert nicht!**
- Grund: Die Normalverteilungen entstehen aus der Standardnormalverteilung durch eine einfache (affin) lineare Umrechnung, nämlich
- X ist $N(\mu, \sigma^2)$-verteilt, wenn $Y = \dfrac{X - \mu}{\sigma}$ standardnormalverteilt ist
- Bei den anderen Verteilungen besteht dieser einfache Zusammenhang nicht

Plot von Quantilen von exp(2)-verteilten Daten gegen die Quantile der exp(1)-Verteilung



Die Q-Q-Methode funktioniert nur zum Testen von Normalverteilungsannahmen

Rolle der Verteilungsannahmen

- Bisherige Tests hatten Verteilungsannahmen
- Im Prinzip wurden ein oder zwei Parameter geschätzt und daraus dann Schlüsse über die Fehlerwahrscheinlichkeiten gezogen
- Dieses Vorgehen liefert einen *parametrischen Test*
- Ohne Verteilungsannahmen könnte man versuchen, die ganze Verteilung zu schätzen
- Das geht aber nicht!

Parametrisch vs. Nichtparametrisch

- Die Verteilungsannahmen sind eine Zusatzinfo
- Deswegen ist ein passender parametrischer Test genauer als ein nichtparametrischer, vorausgesetzt, die Verteilungsannahmen sind erfüllt
- Ohne Verteilungsannahmen muss ein nichtparametrischer Test durchgeführt werden