QQ-Plots als Instrument zum Vergleich von Verteilungen

Holger Langkabel
inVentiv Health Germany GmbH
Große Hub 10d
65344 Eltville
holger.langkabel@inventivhealth.com

Zusammenfassung

In empirischen Studien ist es häufig notwendig sich zu vergewissern, ob eine angenommene Verteilung für die Daten angemessen ist. Eventuell muss man sogar erst eine Hypothese über die Datenverteilung aufstellen. Dieser Aufsatz möchte daher die Möglichkeit aufzeigen, solche Aufgaben mit Hilfe von Quantil-Quantil-Diagrammen (Quantile-Quantile-Plots/QQ-Plots) zu lösen. Dabei wird auch die Umsetzung dieser Methode in SAS besprochen. Darüber hinaus ist in bestimmten Analysen (bspw. in nicht-randomisierten Studien) zu überprüfen, ob die Verteilung bestimmter Variablen in Kontroll- und Versuchsgruppe identisch ist. Auch hierzu eignen sich Quantil-Quantil-Diagramme. Allerdings gibt es in SAS keine vorgefertigte Prozedur, mit der eine solche Graphik erstellt werden könnte (PROC UNIVARIATE ist für diese multivariate Methode nicht ausgelegt). Daher präsentiert dieser Aufsatz auch ein SAS-Makro des Autors, das diese Aufgabe übernimmt.

Schlüsselwörter: Quantil-Quantil-Diagramme, QQ-Plots, Datenvisualisierung, nichtparametrische Methoden

1 Einleitung

In empirischer Statistik tritt häufig das Problem auf, dass die einer metrischen Variablen zugrunde liegende wahre Verteilung unbekannt ist. Allerdings ist es auch häufig nötig eine Annahme über solche Verteilungen zu treffen, um die Anwendung inferenzstatistischer Methoden zu ermöglichen. Daher ist es wichtig, die Richtigkeit dieser Annahmen zu überprüfen. Eine Möglichkeit, dies zu tun, sind formale statistische Anpassungstests, die die angenommene Verteilung explizit mittels einer geeigneten Teststatistik testen. Eine zweite Methode, die ergänzend oder ersatzweise genutzt werden kann, sind Quantil-Quantil-Diagramme (englisch: Quantile-Quantile-Plot oder kurz: QQ-Plot). Dabei handelt es sich um eine grafische Methode, mit deren Hilfe eingeschätzt werden kann, welches Verteilungsmodell hinreichend gut auf die empirischen Daten passt.
Kurz zusammengefasst nutzt diese Methode die Tatsache, dass die empirisch beobachteten Quantile den Quantilen der angenommenen Verteilung recht nah kommen sollten. Daher sollte die Kurve der empirischen Quantile abgetragen gegen die theoretischen Quantile einem vordefinierten Muster folgen. Üblicherweise werden Quantil-Quantil-Diagramme so konstruiert, dass es sich bei diesem Muster um eine gerade Linie handelt. Auf Grund der Nähe oder Abweichung von dem vorgegebenen Muster kann dann beur-

teilt werden, ob die angenommene Verteilung auf die empirischen Daten passt oder nicht.
Ein zweites Feld, auf dem Quantil-Quantil-Diagramme benutzt werden können, ist der Vergleich einer empirischen Verteilung gegen eine andere empirische statt einer theoretischen Verteilung. Diese Aufgabe stellt sich bspw. beim Vergleich einer Verteilung über verschiedene Teilgruppen in den Daten (z. B. bei einem Vergleich von Kontroll- und Behandlungsgruppe). Dadurch kann ermittelt werden, ob die Verteilung einer interessierenden Variable über die Teilgruppen hinweg identisch ist, d.h., ob die Variable über die Gruppen hinweg „ausbalanciert" ist. In einer üblichen randomisierten kontrollierten Studie der medizinischen Forschung ist dies schon der Fall einfach auf Grund des Studiendesigns. In nicht-interventionellen Studien muss dies jedoch nicht zutreffen. Solche Studien sind anfällig für eine Selbstselektion der Patienten in Kontroll- und Behandlungsgruppe. Da eine solche Selektion auf Variablen beruhen kann, die auch den Behandlungserfolg beeinflussen, ist es wichtig zu überprüfen, ob solche Variablen dieselbe empirische Verteilung über die Teilgruppen hinweg aufweisen (z. B. dass in einer Studie über ein neues Blutdruckpräparat die Verteilung des BMI zwischen Kontroll- und Behandlungsgruppe dieselbe ist).
Im Weiteren ist dieser Artikel wie folgt gegliedert: Der nächste Abschnitt befasst sich kurz mit den theoretischen Grundlagen von Quantil-Quantil-Diagrammen; geht dabei aber nicht zu sehr ins (mathematische) Detail. Der dritte Abschnitt erläutert, wie man mit Hilfe von SAS Quantil-Quantil-Diagramme mittels geeigneter Prozeduren erstellen kann. Außerdem werden zahlreiche Beispiele zur Interpretation von Quantil-Quantil-Diagrammen vorgestellt. Der abschließende Abschnitt erläutert die Konstruktion von Quantil-Quantil-Diagrammen zum Vergleich zweier empirischer Verteilungen. Da SAS keine vorgefertigte derartige Funktionalität bietet, wird hierzu ein vom Autor entwickeltes Makro verwendet, das sich auch im Anhang findet. Im Literaturverzeichnis finden sich Referenzen zu weiterführender Literatur, auf der dieser Beitrag basiert.

2 Quantil-Quantil-Diagramme in der Theorie

Um ein Quantil-Quantil-Diagramm zu erstellen, müssen die Beobachtungen zunächst der Größe des Variablenwertes nach sortiert werden. Mathematisch kann das so notiert werden, dass die Stichprobe $x_1, x_2, \dots, x_n$ in die geordnete Reihenfolge $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ gebracht wird, wobei $x_{(1)}$ der kleinste Wert ist, $x_{(2)}$ der zweitkleinste usw. Definieren wir nun $F(x)$ als Wahrscheinlichkeitsfunktion einer nicht näher bestimmten Verteilung, dann sollte der theoretische Wert der Wahrscheinlichkeitsfunktion für $x_{(i)}$, $F\left(x_{(i)}\right)$ (d. h. die theoretische Wahrscheinlichkeit, dass ein Wert höchstens so groß wie $x_{(i)}$ auftritt), etwa dem tatsächlich beobachteten Gegenstück entsprechen:

$$F\left(x_{(i)}\right) \approx \frac{i}{n}$$

(falls n hinreichend groß). Hierbei stellt i/n die geschätzte Wahrscheinlichkeit, einen Wert höchstens so groß wie $x_{(i)}$ zu beobachten, dar (da genau i Beobachtungen der Stichprobe der Größe n einen Wert von x haben, der kleiner oder gleich $x_{(i)}$ ist).
Durch Umformung erhält man

$$x_{(i)} \approx F^{-1}\left(\frac{i}{n}\right).$$

Hierbei ist $F^{-1}(\cdot)$ die Inverse der Wahrscheinlichkeitsfunktion, die die Quantile der Verteilung liefert. Daher kann man ebenfalls erwarten, dass die geordneten Werte $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ in etwa den theoretischen Quantilen $F^{-1}(1/n)$, $F^{-1}(2/n), \ldots, F^{-1}(n/n)$ entsprechen.
Aus der zweiten Formel kann geschlossen werden, dass ein Diagramm, das die geordneten Werte gegen ihre entsprechenden theoretischen Werte abträgt, eine Kurve zeigt, die um die erste Winkelhalbierende streut. Folglich haben wir eine Diagrammanweisung, die es uns erlaubt, die Eignung eines bestimmten Verteilungsmodells zu beurteilen:

> Man trage die geordneten Werte der Stichprobe gegen die angenommenen theoretischen Quantile ab. Wenn die angenommene Verteilung korrekt ist, streuen die Diagrammpunkte zufällig um eine Gerade mit Steigung 1 und Achsenabschnitt 0.

Auf Grund der oben genutzten Umformung lässt sich erkennen, dass es genauso möglich wäre die Werte der empirischen Wahrscheinlichkeitsfunktion gegen diejenigen der theoretischen Wahrscheinlichkeitsfunktion abzutragen. Dabei handelt es sich um eine genauso berechtigte Methode die Passgenauigkeit einer angenommenen Verteilung zu prüfen. Solche Diagramme werden im Englischen als *probability-probability plot* oder *percent-percent plot* (kurz: PP-Plot) bezeichnet.

3 Quantil-Quantil-Diagramme in der Praxis

Schauen wir uns ein paar Beispiele an: Wenn wir einen Datensatz mit folgendem Programmcode erstellen

```
data normal;
     call streaminit(1234);
     do i=1 to 100;
          normal1 = rand('normal', 0,1);
          normal2 = rand('normal',10,4);
          exp = rand('weibull',1,0.5);
          output;
     end;
run;
```

erhalten wir einen Datensatz mit drei Variablen und jeweils 100 Beobachtungen. Die Variable NORMAL1 ist standardnormalverteilt, Variable NORMAL2 hat eine Normalverteilung mit Mittelwert 10 und Standardabweichung 4 und Variable EXP ist expo-

nentialverteilt mit $\lambda = 2$ (Dies entspricht in SAS einer Weibull-Verteilung mit Parameterwerten 1 und 0,5.).
SAS stellt einfache Quantil-Quantil-Diagramme in der UNIVARIATE-Prozedur bereit. Ein einfacher Aufruf der Prozedur sieht wie folgt aus:

```
proc univariate data=normal noprint;
     qqplot normal1 normal2 exp / normal(mu=0 sigma=1);
run;
```

Dies resultiert in den folgenden drei Graphiken. Bei der Option `normal(mu=0 sigma=1)` handelt es sich eigentlich um die Voreinstellung. Jedoch zeichnet SAS keine Referenzlinie ein, wenn keine Vergleichsverteilung angegeben wird. Abbildung 1 zeigt ein perfektes Ergebnis: Die abgetragenen Punkte bilden nicht nur eine gerade Linie, sondern sie streuen auch noch rein zufällig um die Referenzlinie. Dies liegt natürlich daran, dass die Variable tatsächlich standardnormalverteilt ist. Es ist auch nicht ungewöhnlich, dass die Punkte an den Enden der Verteilung etwas stärker vom Ideal abweichen. Abbildung 2 zeigt ein etwas anderes Bild: Obwohl die abgetragenen Punkte ebenfalls eine gerade Linie bilden, streuen sie nicht um die Referenzlinie. Die Form der Kurve zeigt an, dass die Normalverteilung angemessen ist, während das Abweichen von der Referenzlinie anzeigt, dass lediglich die Parameter falsch gewählt wurden. In Abbildung 3 bilden die Punkte überhaupt keine gerade Linie, was zu erwarten war, da die zugrundeliegende Verteilung die Exponential- und nicht die Normalverteilung ist.

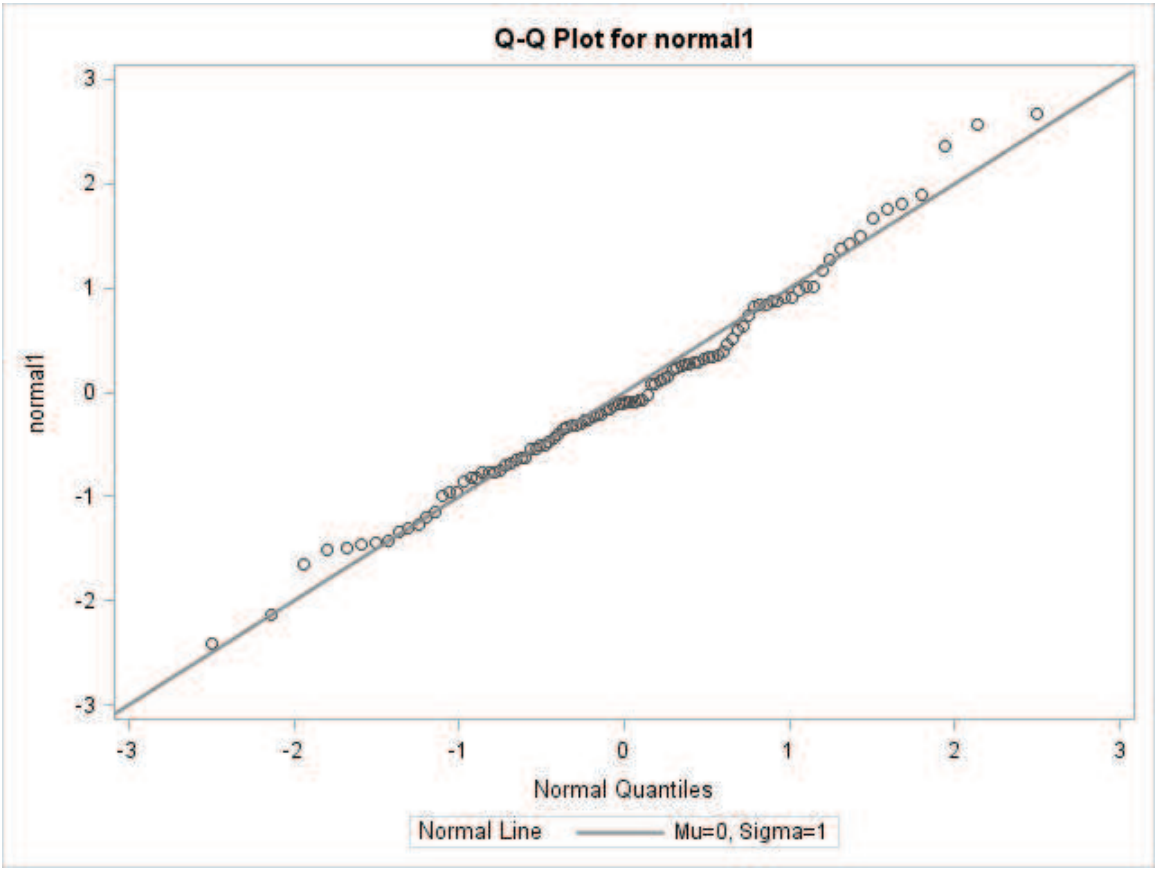

Abbildung 1: Quantil-Quantil-Diagramm für NORMAL1 mit angenommener Standardnormalverteilung.

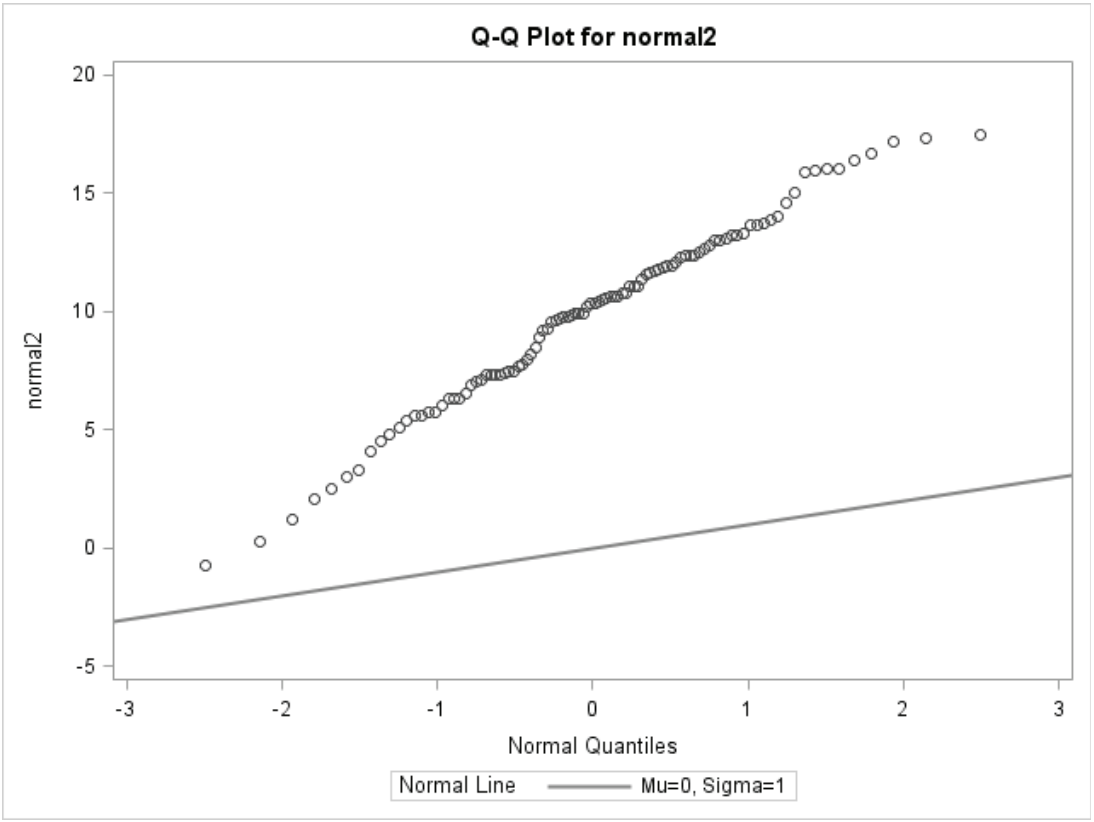

Abbildung 2: Quantil-Quantil-Diagramm für NORMAL2 mit angenommener Standardnormalverteilung.

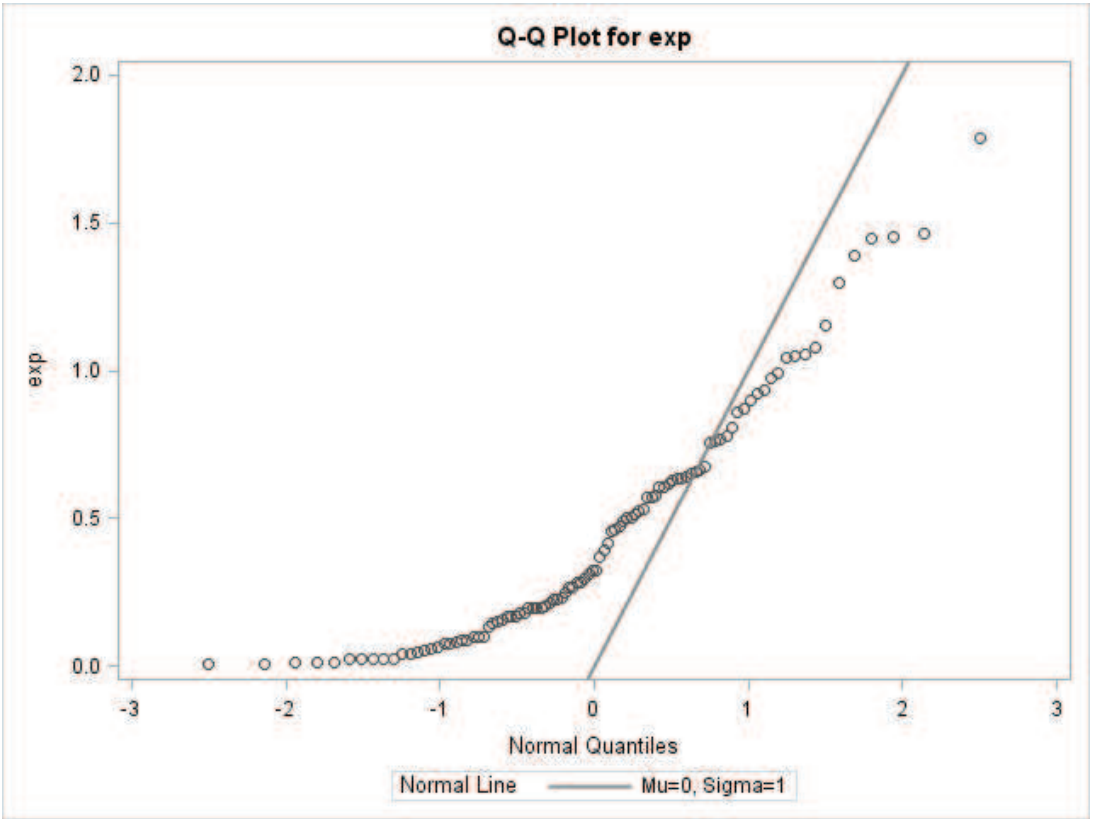

Abbildung 3: Quantil-Quantil-Diagramm für EXP mit angenommener Standardnormalverteilung.

Nutzt man diese Ergebnisse und passt das Diagramm für NORMAL2 entsprechend an, könnte sich Abbildung 4 ergeben, die die empirischen Quantile der Variable mit einer Normalverteilung mit Mittelwert 10 und Standardabweichung 4 vergleicht. Jetzt streuen die Punkte um die Referenzlinie. Analog könnte man die Variable EXP mit einer Exponentialverteilung vergleichen:

```
proc univariate data=normal noprint;
     qqplot exp / exponential(theta=0 sigma=0.5);
run;
```

Die resultierende Abbildung 5 zeigt eine deutlich bessere Anpassung. Nichtsdestotrotz mag die Richtigkeit der angenommen Verteilung durch Abbildung 5 in Frage gestellt werden, da die Punkte für die höheren Quantile teilweise stark von der Referenzlinie abweichen. Dies deutet auf die Subjektivität der Methode.

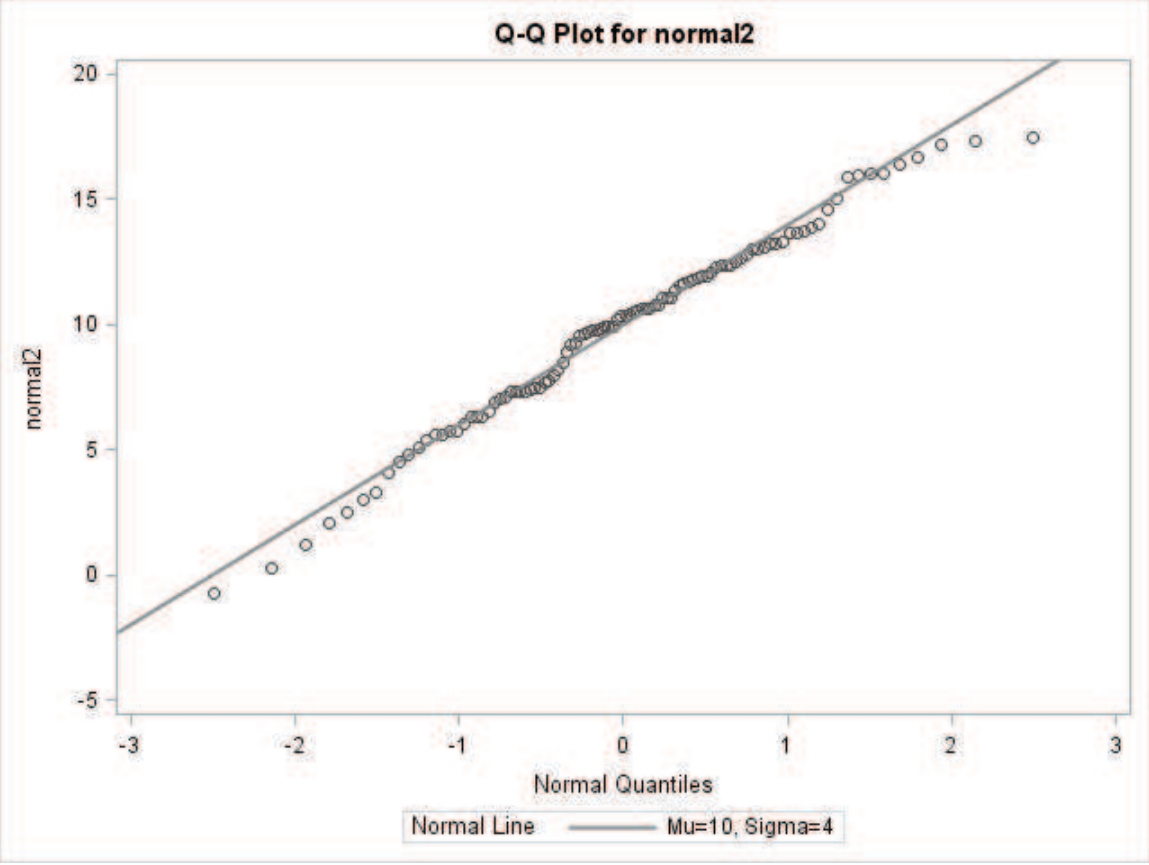

Abbildung 4: Quantil-Quantil-Diagramm für NORMAL2 mit angenommener $N(10;4)$-Verteilung.

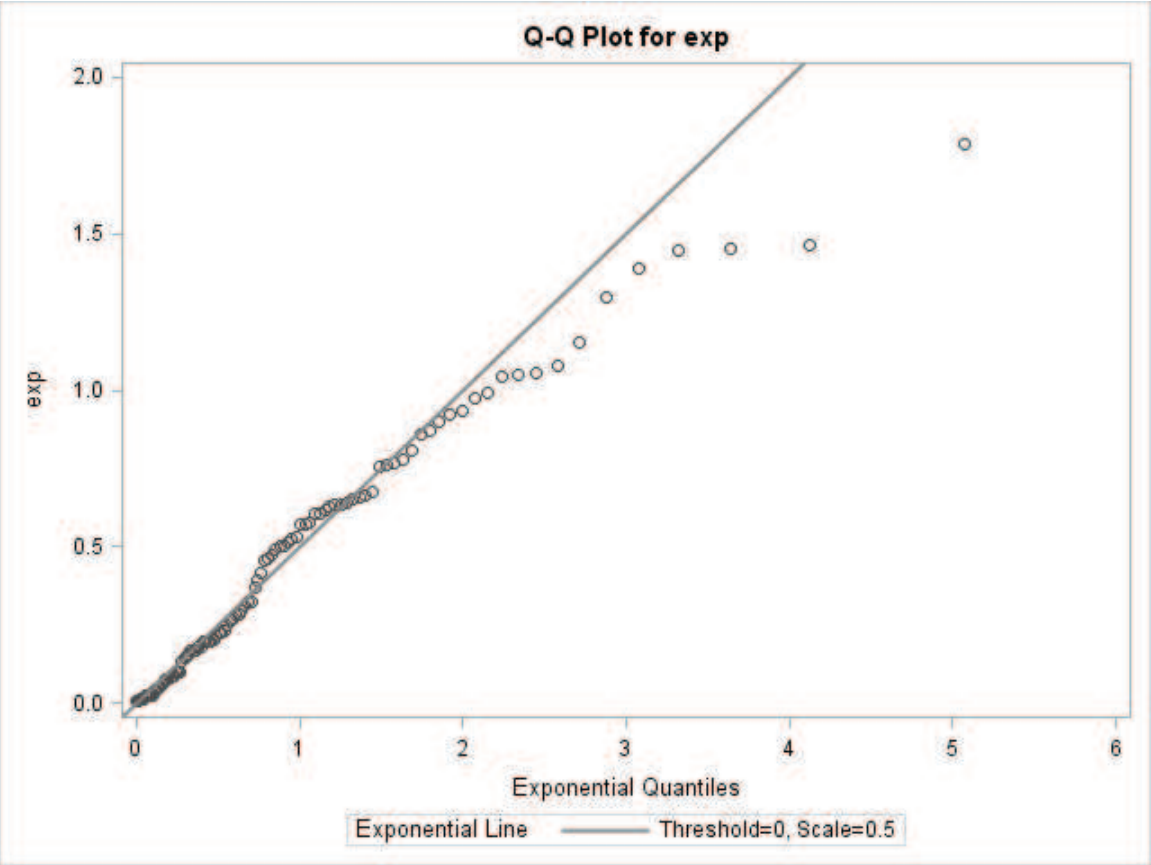

Abbildung 5: Quantil-Quantil-Diagramm für EXP mit angenommener Exponentialverteilung ($\lambda = 2$).

Zugegebenermaßen ist die Erstellung korrekter Graphiken recht einfach, wenn die wahre Verteilung bereits bekannt ist. Aber sobald eine Familie von Verteilungen als geeignet eingestuft wird, lassen sich die einzelnen Parameter der Verteilung aus dem Quantil-Quantil-Diagramm schätzen. So können bspw. die Parameter der Normalverteilung aus Abbildung 2 geschätzt werden. Dies wird in Abbildung 6 illustriert: Das Diagramm zeigt eine Referenzlinie für eine Normalverteilung mit den geschätzten Parametern $\mu = 9{,}8668$ und $\sigma = 3{,}9224$, die nahe bei den wahren Werten von 10 und 4 liegen. Das Diagramm wurde mit folgendem Programmcode erstellt:

```
proc univariate data=normal noprint;
     qqplot normal2 / normal(mu=est sigma=est);
run;
```

Für alle Verteilungen, die der UNIVARIATE-Prozedur bekannt sind[1], können die Verteilungsparameter durch das Stichwort `est` ersetzt werden. Dann wird SAS den (oder die) Parameter aus den Daten schätzen und eine Referenzlinie einzeichnen, die die entsprechende Schätzung nutzt.

[1] Dies sind: Beta-, Exponential-, Gamma-, Gumbel-, Lognormal-, Normal-, Pareto-, Power-Function-, Rayleigh-, dreiparametrische und zweiparametrische Weibull-Verteilung.

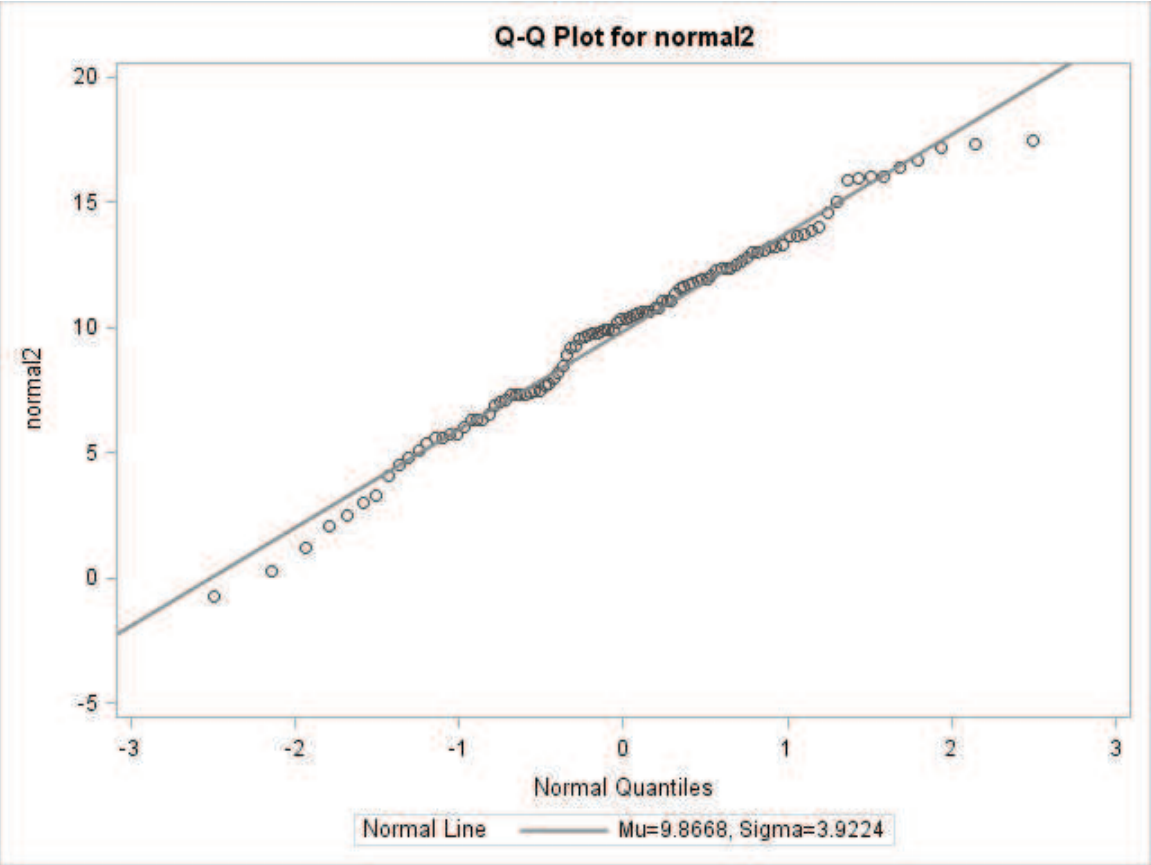

Abbildung 6: Quantil-Quantil-Diagramm für NORMAL2 mit angenommener Normalverteilung und geschätzten Parametern.

4 Quantil-Quantil-Diagramm zweier empirischer Verteilungen

Aber wie sieht es aus, wenn die Methode auf empirische und nicht auf simulierte Daten angewendet wird? Benutzen wir den SASHELP.BWEIGHT-Datensatz aus SAS 9.2. Dieser Datensatz enthält Daten des amerikanischen National Center for Health Statistics zum Geburtsgewicht von 50.000 Säuglingen aus dem Jahr 1997. Abbildung 7 zeigt ein Quantil-Quantil-Diagramm für die Variable WEIGHT unter der Annahme einer Normalverteilung. Abbildung 8 zeigt ein Quantil-Quantil-Diagramm für dieselbe Variable unter Annahme einer Weibull-Verteilung. Beide Diagramme zeigen eine recht gute Anpassung für die Mitte der Verteilung, jedoch nicht für das untere Ende.

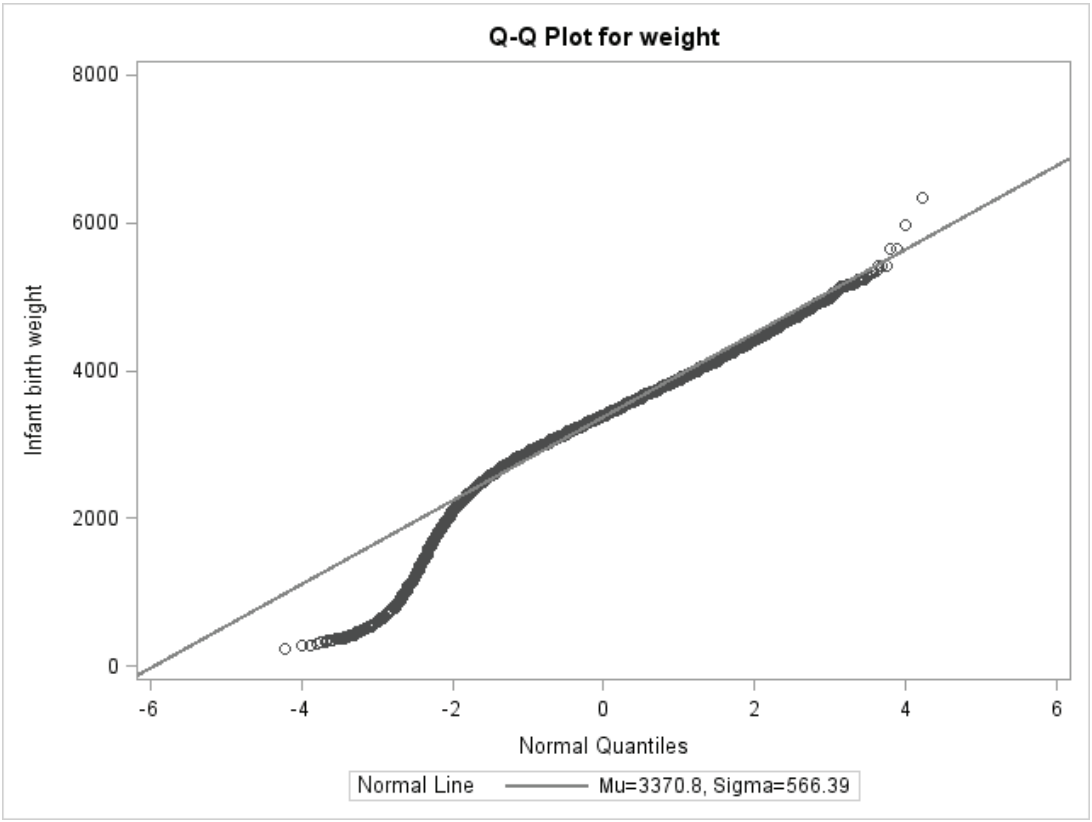

Abbildung 7: Quantil-Quantil-Diagramm für WEIGHT mit angenommener Normalverteilung und geschätzten Parametern.

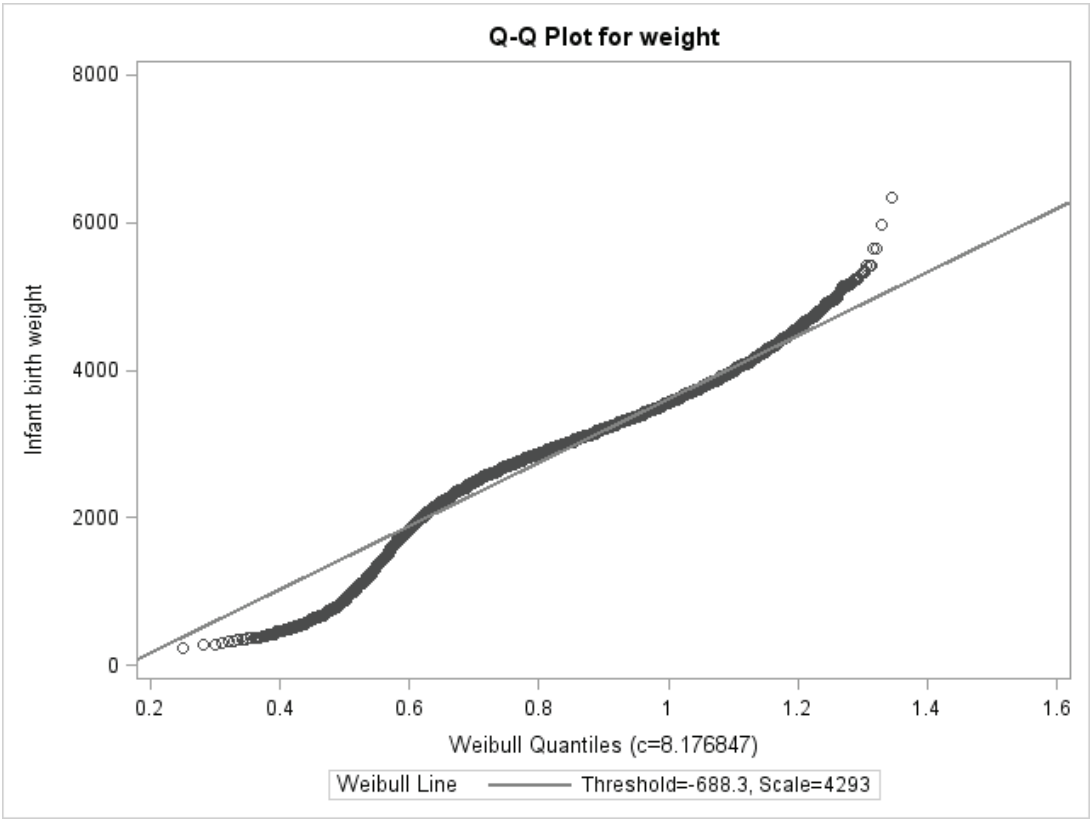

Abbildung 8: Quantil-Quantil-Diagramm mit angenommener Weibull-Verteilung und geschätzten Parametern.

Aber vielleicht ist man auch gar nicht an der zugrundeliegenden Verteilung des Geburtsgewichts selbst interessiert, sondern vielmehr an dem Unterschied in der Verteilung zwischen bestimmten Gruppen. Z. B. könnte man an dem Einfluss der Rauchgewohnheiten der Mutter auf das Geburtsgewicht eines Kindes interessiert sein. Dazu müsste man entweder die Verteilung je Teilgruppe untersuchen (d. h. man müsste zwei Quantil-Quantil-Diagramme erstellen: eines für rauchende Mütter und eines für nichtrauchende Mütter) oder – viel unkomplizierter – man vergleicht die Verteilungen in einem einzigen Diagramm, da man nicht zwangsläufig eine empirische mit einer theoretischen Verteilung vergleichen muss, sondern auch einfach die Quantile einer anderen empirischen Verteilung auf der Abszisse abtragen kann.
Bedauerlicherweise bietet SAS keine vorgefertigte Routine für den Vergleich zweier empirischer Verteilungen. Aus diesem Grund enthält der Anhang dieses Artikels ein Makro, das diese Aufgabe erledigt. Dieses Makro benötigt drei Parameter:

- `data`: der einzulesende Datensatz,
- `var`: die Variable, für die der Zwischengruppenvergleich durchgeführt werden soll, und
- `class`: eine Zahlen- oder Textvariable mit zwei Ausprägungen, die die Gruppenzugehörigkeit anzeigen.

Intern berechnet das Makro eigentlich nur die Quantile der zu untersuchenden Variable je Gruppe und trägt diese dann in einer Graphik gegeneinander ab.

Um unsere Analyseaufgabe zu lösen, benötigen wir folgenden Makroaufruf:

```
%qq(data=sashelp.bweight, var=weight, class=smoke);
```

Das Resultat ist Abbildung 9. Diese zeigt die Quantile des Geburtsgewichts für nichtrauchende Mütter auf der Abszisse und für rauchende Mütter auf der Ordinate. Die eingezeichnete Referenzlinie ist die erste Winkelhalbierende. Wäre die Verteilung identisch über die Gruppen hinweg, müssten die eingezeichneten Punkte zufällig um die Referenzlinie streuen. Dies ist hier offensichtlich nicht der Fall. Daraus kann geschlossen werden, dass die Verteilung des Geburtsgewichts nicht dieselbe für rauchende und nichtrauchende Mütter ist. Da die Punkte aber trotzdem eine gerade Linie bilden, kann geschlossen werden, dass die zugrundeliegende Verteilungsfamilie dieselbe ist. Die Parallelität von Referenz- und Datenlinie verrät, dass beiden Gruppen eine Verteilung mit derselben Streuung (Varianz) zu Grunde liegt, das Niveau des Mittelwerts für nichtrauchende Mütter jedoch höher ist (wobei es sich um ein erwartetes Ergebnis handelt).

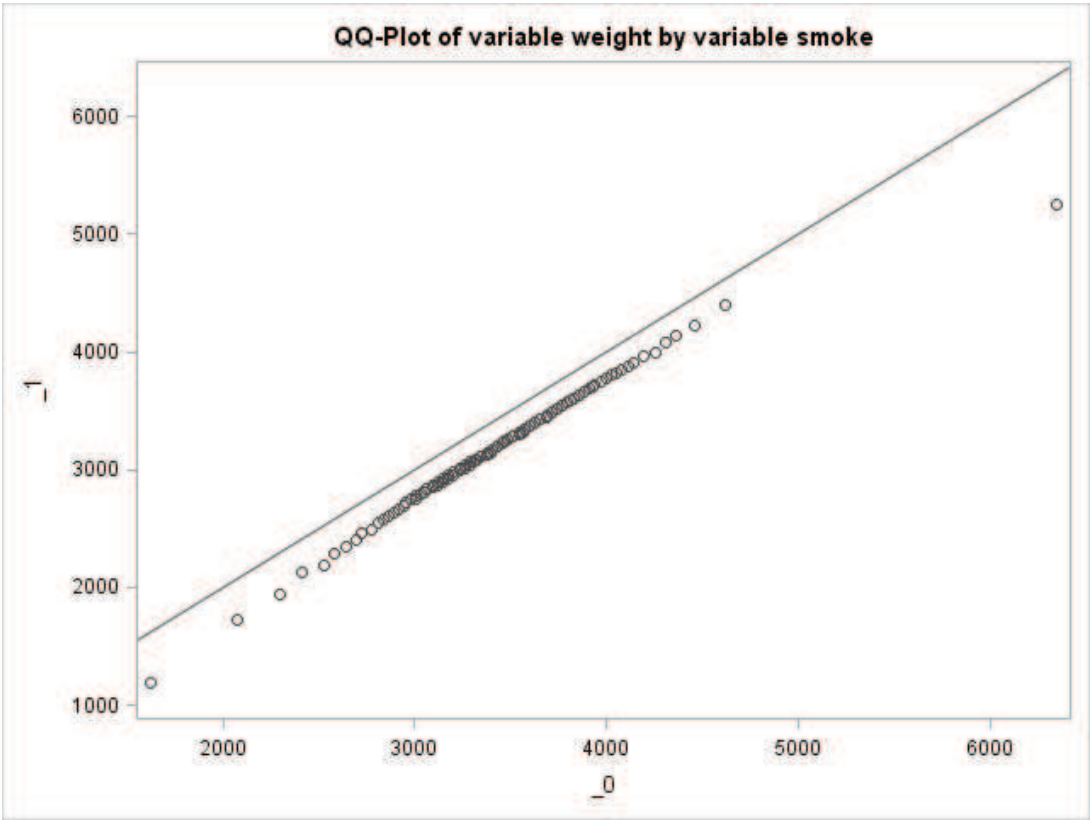

Abbildung 9: Quantil-Quantil-Diagramm für WEIGHT nach Gruppen in SMOKE.

5 Zusammenfassung

Quantil-Quantil-Diagramme sind ein theoretisch recht einfach herzuleitendes Instrument zum Vergleich von Verteilungen. Kurz zusammengefasst nutzen sie einfach die zu erwartende Ähnlichkeit der Quantile einer theoretischen Verteilung und einer empirischen Realisation dieser Verteilung. Stimmen theoretische Annahme und empirische Daten (teilweise) überein, lassen sich im Quantil-Quantil-Diagramm bestimmte Muster erkennen, mit deren Hilfe man die Korrektheit der angenommenen Verteilung überprüfen kann. Ein besonderer Vorzug dieser Methode ist, dass es sich um eine nicht-parametrische Methode handelt, da lediglich die Quantile zweier Verteilungen verglichen werden und daher gerade keine Verteilungsannahmen zur Anwendung der Methode erforderlich sind. Daraus ergibt sich auch die Möglichkeit zwei empirische Verteilungen direkt auf Übereinstimmung hin zu überprüfen, ohne dass man irgendeine Annahme über die zugrundeliegende Verteilung machen muss. Quantil-Quantil-Diagramme für den Test einer empirischen gegen eine theoretische Verteilung lassen sich in SAS ganz einfach mit Hilfe der UNIVARIATE-Prozedur erstellen. Für den Vergleich von zwei empirischen Verteilungen steht das im Anhang dargestellte Makro zur Verfügung.

Literatur

[1] J. M. Chambers, W. S. Cleveland, B. Kleiner, P. A. Tukey: Graphical Methods for Data Analysis. Wadsworth: Belmont 1983.

[2] D. C. Montgomery: Statistical Quality Control: A Modern Introduction. John Wiley & Sons, Singapur 2013.

Anhang: SAS-Makro zur Erstellung bivariater Quantil-Quantil-Diagramme

Bitte beachten Sie bei Benutzung dieses Makros, dass es unter SAS 9.2 in einer UNIX-Umgebung entwickelt wurde.

```
%macro qq(data, var, class);
    %local ngroups nobs value1 value2 increment;

    proc sort data=&data.;
        by &class.;
    proc means data=&data. noprint;
        var &var.;
        by &class.;
        output out=_n(where=(_stat_="N"));
    run;

    proc sql noprint;
        select count(*)
             , min(&var.)
             , &class.
        into :ngroups, :nobs, :value1-:value2
        from _n;
    quit;

    %if %sysfunc(anydigit(&value1.)) = 1 %then
        %let value1 = _&value1.;
    %if %sysfunc(anydigit(&value2.)) = 1 %then
        %let value2 = _&value2.;

    %if &ngroups ne 2 %then %do;
        %put ERROR: Macro qq: &class. does not have two distinct
values.;
        %goto abort_macro;
    %end;

    * Choose number of displayed percentiles dependent on number of
      observations;

          %if &nobs. >= 100 %then %let increment =   1;
    %else %if &nobs. >=  50 %then %let increment =   2;
    %else %if &nobs. >=  25 %then %let increment =   4;
```

```
    %else %if &nobs. >=  20 %then %let increment =   5;
    %else %if &nobs. >=  10 %then %let increment =  10;
    %else %if &nobs. >=   5 %then %let increment =  20;
    %else %if &nobs. >=   4 %then %let increment =  25;
    %else %if &nobs. >=   2 %then %let increment =  50;
    %else %if &nobs. >=   1 %then %let increment = 100;

    * Calculate percentiles;

    proc univariate data=&data. noprint;
        var &var.;
        class &class.;
        output out=_qq1 pctlpts=1 to 100 by &increment. pctlpre=p;
        format &class.;
    run;

    * Prepare data;

    proc transpose data=_qq1 out=_qq2;
        var p:;
        id &class.;
    run;

    * Plot;

    proc sgplot data=_qq2 noautolegend;
        title "QQ-Plot of variable &var. by variable &class.";
        scatter x=&value1. y=&value2.;
        lineparm x=0 y=0 slope=1;
    run;

    proc datasets library=work;
        delete _n _qq1 _qq2;
    run;

    %abort_macro:
%mend qq;
```