Bausteine der Datenanalyse

Methodenbausteine Statistik

Lukas Arnold Simone Arnold Florian Bagemihl
Matthias Baitsch Marc Fehr Franca Hollmann
Maik Poetzsch Sebastian Seipel

2026-03-19

Dichtekurven und Normalverteilung

Dichtekurven

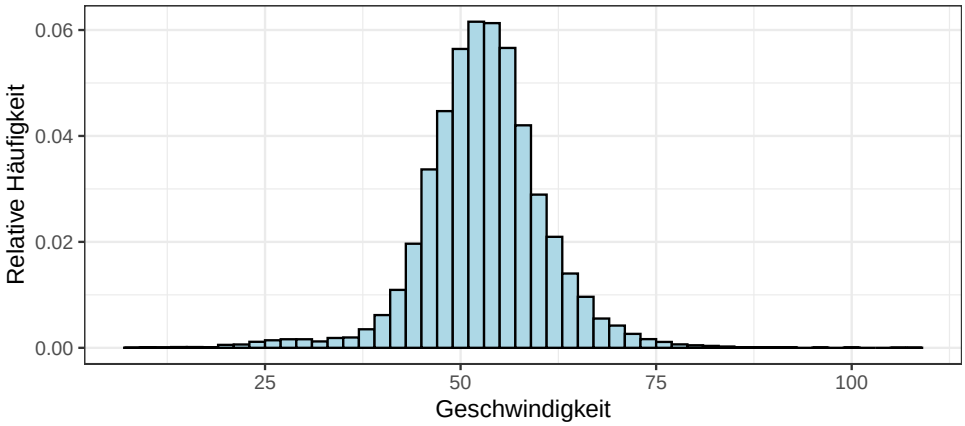
Dichtekurven sind eine Alternative zu Histogrammen für die Darstellung einzelner Merkmale. Die Grundidee soll am Beispiel einer Geschwindigkeitsmessung erläutert werden.

Beispiel Geschwindigkeitsmessung: In einer Geschwindigkeitsmessung mit dem Seitenradar an der Universitätsstraße in Bochum wurden innerhalb von zwei Tagen die Geschwindigkeiten von 20710 Fahrzeugen gemessen.

	Datum	Geschwindigkeit	Fahrzeug
1	13.12.2017 00:03:31	70	PKW
2	13.12.2017 00:03:37	59	PKW
3	13.12.2017 00:03:53	79	PKW
4	13.12.2017 00:05:42	58	PKW
5	13.12.2017 00:06:51	60	PKW
6..20709			
20710	14.12.2017 23:58:29	72	PKW

Daten bereitgestellt von Prof. Iris Mühlenbruch

Unten dargestellt ist das Histogramm der relativen Häufigkeiten der gemessenen Geschwindigkeiten, die Klassenbreite wurde dabei mit 2 km/h gewählt.



Man kann sich nun vorstellen, die Messung der Geschwindigkeiten mit einer höheren Genauigkeit durchzuführen und dabei über einen längeren Zeitraum zu messen. Dann wäre es durchaus möglich, das Histogramm mit einer kleineren Klassenbreite zu erstellen, zum Beispiel 0.1 km/h. Damit nähert sich der Verlauf des Histogramms der kontinuierlichen Kurve an, die sich im Grenzübergang zu einer unendlich langen und gleichzeitig exakten Messung mit einer gegen Null gehenden Klassenbreite ergeben würde.

Mit einer Dichtekurve wird dieser Grenzübergang vorweggenommen: Die Idee dabei ist es, das Histogramm durch eine kontinuierliche Funktion $f : \mathbb{R} \to \mathbb{R}$ zu approximieren. Die zugehörige Dichtekurve $y = f(x)$ approximiert und glättet das Histogramm. Dabei gehen gegebenenfalls Details verloren. Dafür ergibt sich eine optisch klare Darstellung, die sich manchmal sogar mit einer einfachen Formel für die Zuordnungsvorschrift von f angeben lässt.

Um eine solche Funktion charakterisieren zu können, rufen wir uns nochmal die Eigenschaften des Histogramms der relativen Häufigkeiten in Erinnerung:

1. Alle Rechtecke haben eine positive Höhe,
2. die Fläche eines Rechtecks ist der Anteil der Werte, die in der entsprechenden Klasse liegen,
3. die Summe der Rechtecksflächen ist demnach eins.

Diese Eigenschaften werden auf die Dichtekurve übertragen und es ergibt sich die folgende Definition:

Definition 0.1 (Dichtekurve)**.** Eine stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt Dichtekurve, wenn für alle $x \in \mathbb{R}$ gilt, dass $f(x) \geq 0$ ist und gleichzeitig die vom Graphen von f überdeckte Fläche gleich 1 ist. Die zweite Bedingung können wir mithilfe eines Integrals in der Form

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

aufschreiben.

In Abbildung 1 dargestellt ist ein Beispiel für eine solche Dichtekurve. Da die Dichtekurve eine Approximation des Histogramms ist, approximiert die Fläche unter dem Graphen von f zwischen a und b den Flächeninhalt des Rechtecks. Daher gilt

$$\int_a^b f(x)\,dx \approx \text{Anteil der Werte zwischen } a \text{ und } b.$$

Dieser Zusammenhang gilt sogar, wenn das Intervall von a bis b nicht mit Klassengrenzen des Histogramms übereinstimmt. Das Integral über die Dichtekurve spielt in verwandter Form eine besonders wichtige Rolle in der Wahrscheinlichkeitstheorie.

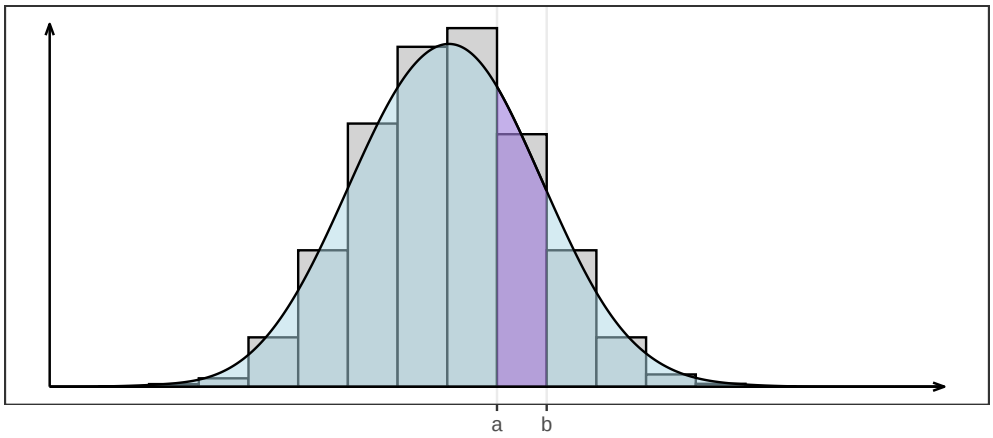


Abbildung 1: Dichtekurve

Quantile und Median

Zwei Größen, die sich unmittelbar von Häufigkeitsverteilungen auf Dichtekurven übertragen lassen, sind die Quantile und der Median (vergleiche Abbildung 2). Das p-Quantil x_p mit $0 < p < 1$ teilt die Fläche unter der Dichtekurve in zwei Teilflächen mit den Flächeninhalten p und $1 - p$. Der Median x_{med} teilt die Dichtekurve in zwei Teilflächen mit einem Flächeninhalt von jeweils $1/2$.

Normalverteilung

Eine besonders wichtige Klasse von Dichtekurven bilden die Normalverteilungen. Sie sind nach dem Mathematiker Carl Friedrich Gaus (1777 - 1855) benannt. Der Graph einer Normalverteilung wird häufig auch Gaußsche Glockenkurve genannt. Sie hat es sogar auf den 10-Mark-Schein geschafft.

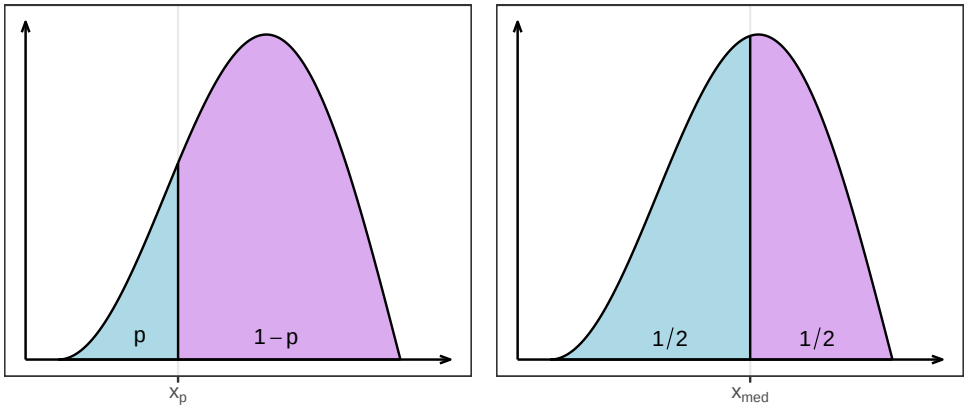


Abbildung 2: Quantil und Median einer Dichtekurve

Definiert ist die Dichtekurve der Normalverteilung mithilfe der Exponentialfunktion.

Definition 0.2 (Dichtekurven von Normalverteilungen)**.** Für $\mu, \sigma \in \mathbb{R}$ und $\sigma > 0$ ist die Dichte der Normalverteilung durch

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

gegeben. Dabei heißt μ Mittelwert und σ Standardabweichung der Verteilung.

Eigenschaften der Normalverteilung

Die Form der Normalverteilung ergibt sich aus dem Faktor $\exp(-1/2\,((x-\mu)/\sigma)^2)$ der Funktionsgleichung. Da das Argument der Exponentialfunktion negativ ist, nimmt die Funktion f an der Stelle $x = \mu$ ihr Maximum ein. Sie fällt von dort nach links und rechts ab und nähert sich für $|x|$ gegen ∞ der x-Achse an. Das geht umso schneller, je kleiner der Wert von σ ist. Weiterhin kann man sich überlegen, dass f in $x = \mu - \sigma$ und $x = \mu + \sigma$ jeweils eine Wendestelle besitzt, siehe Abbildung 3.

Der Vorfaktor $1/(\sigma\sqrt{2\pi})$ in der Funktionsgleichung von f sorgt dafür, dass die Fläche unter dem Graphen von f gleich 1 ist.

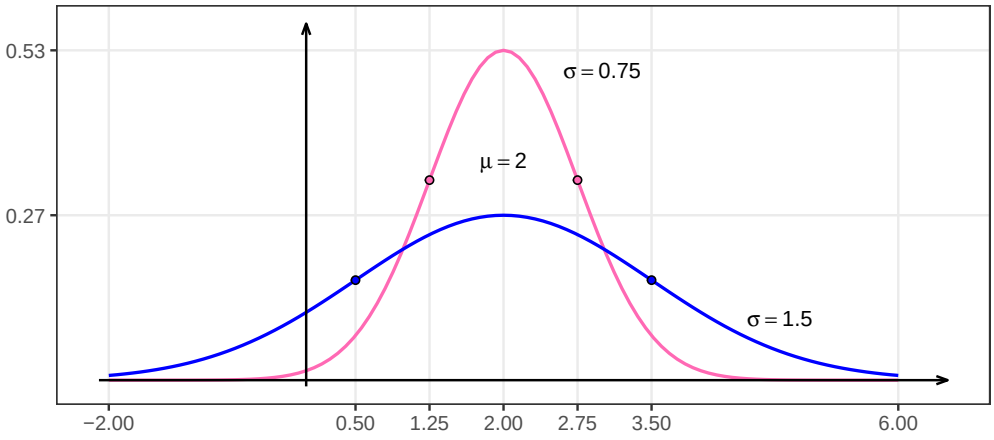


Abbildung 3: Zwei Normalverteilungen

Für die Anteile in bestimmten Intervallen um den Mittelwert μ gilt die **68-95-99.7** Regel: Von der Gesamtfläche unter der Dichtekurve liegen

$$68\% \quad \text{im Intervall} \quad [\mu - \sigma,\ \mu + \sigma],$$

$$95\% \quad \text{im Intervall} \quad [\mu - 2\sigma,\ \mu + 2\sigma],$$

$$99.7\% \quad \text{im Intervall} \quad [\mu - 3\sigma,\ \mu + 3\sigma].$$

Die Bereiche sind in Abbildung 4 farblich hervorgehoben.

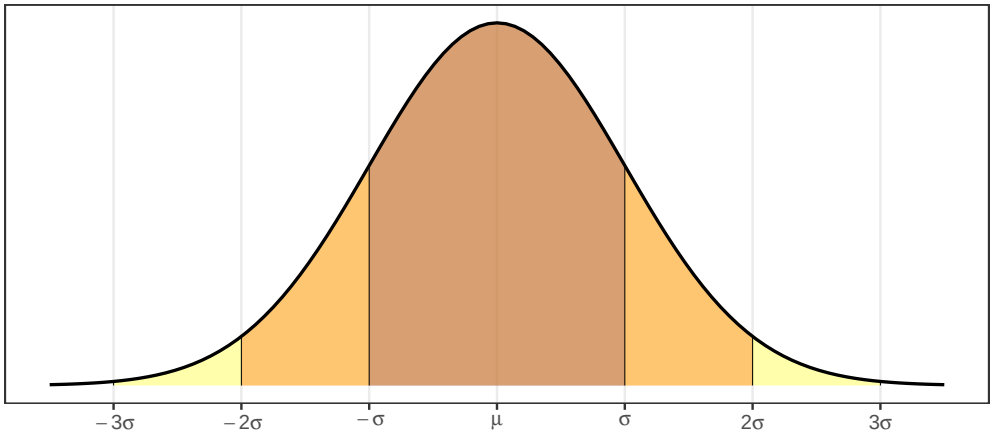


Abbildung 4: Flächen unter der Normalverteilung

Standardnormalverteilung

Eine Normalverteilung mit dem Mittelwert $\mu = 0$ und der Standardabweichung $\sigma = 1$ heißt Standardnormalverteilung, siehe Abbildung 5. Die zugehörige Dichtekurve ist

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

Das Quantil p zu einem Wert z_p ist, so haben wir das oben vereinbart, die Fläche unter der Dichtekurve $\phi(z)$ links von z_p. Natürlich kann man nun für beliebige Werte von z den zugehörigen Flächeninhalt unter der Dichtekurve links davon bestimmen. Dies führt auf die Verteilungsfunktion Φ, die unten rechts dargestellt ist. Aufgrund der Symmetrie von $\phi(z)$ und weil die gesamte Fläche unter $\phi(z)$ gleich 1 ist, muss $\Phi(0) = 0.5$ gelten. Für z gegen ∞ nähert sich der Funktionswert dem Wert 1 an. Erstaunlich ist, dass sich die Stammfunktion Φ nicht geschlossen angeben lässt, in der Regel wird sie numerisch ausgewertet.

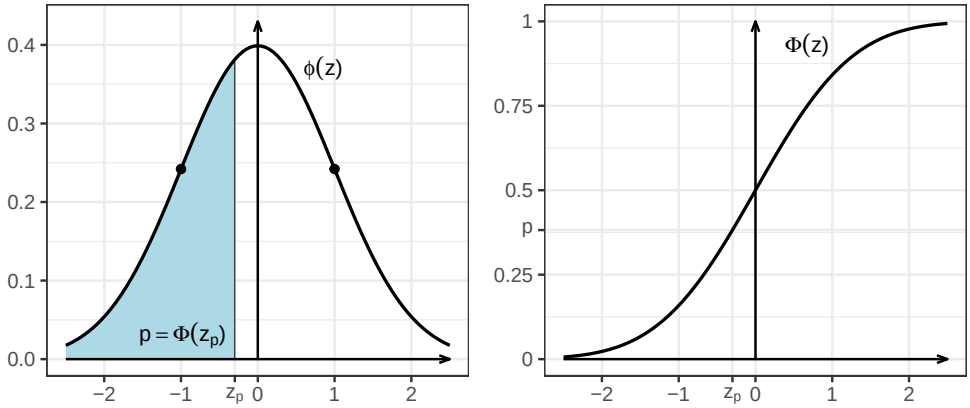


Abbildung 5: Standardnormalverteilung ϕ und zugehörige Verteilungsfunktion Φ

Normalverteilung anpassen

Um nun einen vorliegenden Datensatz durch eine Normalverteilung zu approximieren, liegt es nahe den Mittelwert $\bar{x}$ und die empirische Standardabweichung $\tilde{s}$ der Verteilung in die Normalverteilung einzusetzen. Mit $\mu = \bar{x}$ und $\sigma = \tilde{s}$ ergibt sich damit eine Approximation der Verteilung.

Beispiel Geschwindigkeitsmessung (Fortsetzung): Für die Geschwindigkeitsmessung auf Universitätsstraße erhalten wir die Kenngrößen

$$\bar{x} = 52.54\,\text{km/h} \quad \text{und} \quad \tilde{s} = 7.82\,\text{km/h}.$$

Damit ergibt sich für die Dichtekurve die Funktionsgleichung

$$f(x) = 0.05102 \cdot \exp\left(-0.008177 \cdot (x - 52.54)^2\right).$$

Für die Dichtekurve erhalten wir zusammen mit dem Histogramm das in Abbildung 6 dargestellte Bild. Es ist zu erkennen, dass die Normalverteilung den Verlauf des Histogramms zwar grob erfasst, insgesamt jedoch etwas breiter und dafür weniger hoch ausfällt. Offensichtlich lassen sich die gemessenen Geschwindigkeiten nur unzureichend mithilfe einer Normalverteilung abbilden. Um die Güte der Approximation besser beurteilen zu können, verwendet man so genannte Normal-Quantil-Plots.

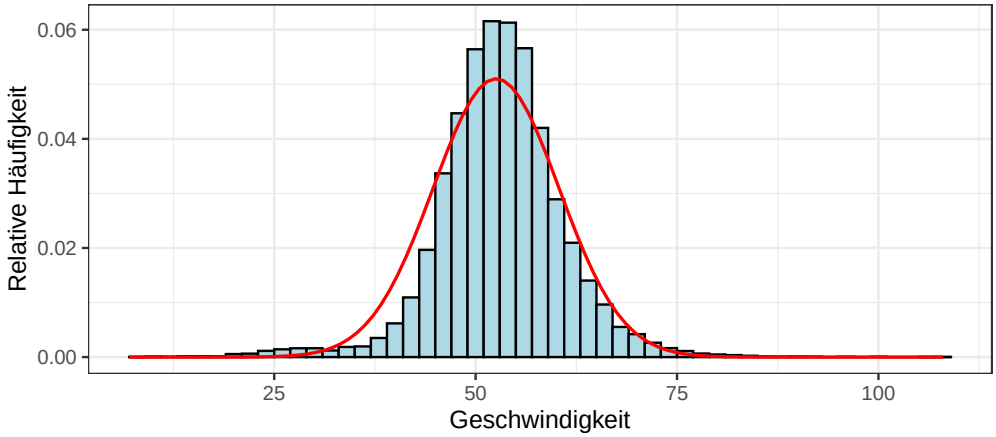


Abbildung 6: An Geschwindigkeitsmessung angepasste Normalverteilung

Normal-Quantil-Plots

Für den zu einem empirisch erhobenen Merkmal X gehen wir wieder von der geordneten Urliste $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ aus. Da die Liste geordnet ist, sind für den Wert $x_{(i)}$ genau i Werte kleiner oder gleich groß. Somit ist $x_{(i)}$ nichts anderes als das i/n-Quantil der Verteilung. Zu diesem i/n-Quantil bestimmt man nun den Wert z_i als zugehöriges Quantil der Standardnormalverteilung. Diese Wertepaare werden als Punkte in einem zx-Koordinatensystem aufgetragen. Dabei muss man ein bisschen aufpassen: Das 1-Quantil der Standardnormalverteilung liegt im Unendlichen, für den letzten Wert x_n würden wir daher keinen zugehörigen Werte z_n auftragen können. Daher werden die Quantile der Standardnormalverteilung nicht für i/n sondern für $(i - 0.5)/n$ berechnet (Stetigkeitskorrektur). Die gesamte Vorgehensweise ist in Abbildung 7 bildhaft erläutert.

Mit dem Normal-Quantil-Plot wird also veranschaulicht, inwieweit die Verteilung eines Merkmals mit der theoretischen Standardnormalverteilung übereinstimmt. Je besser die Verteilungen

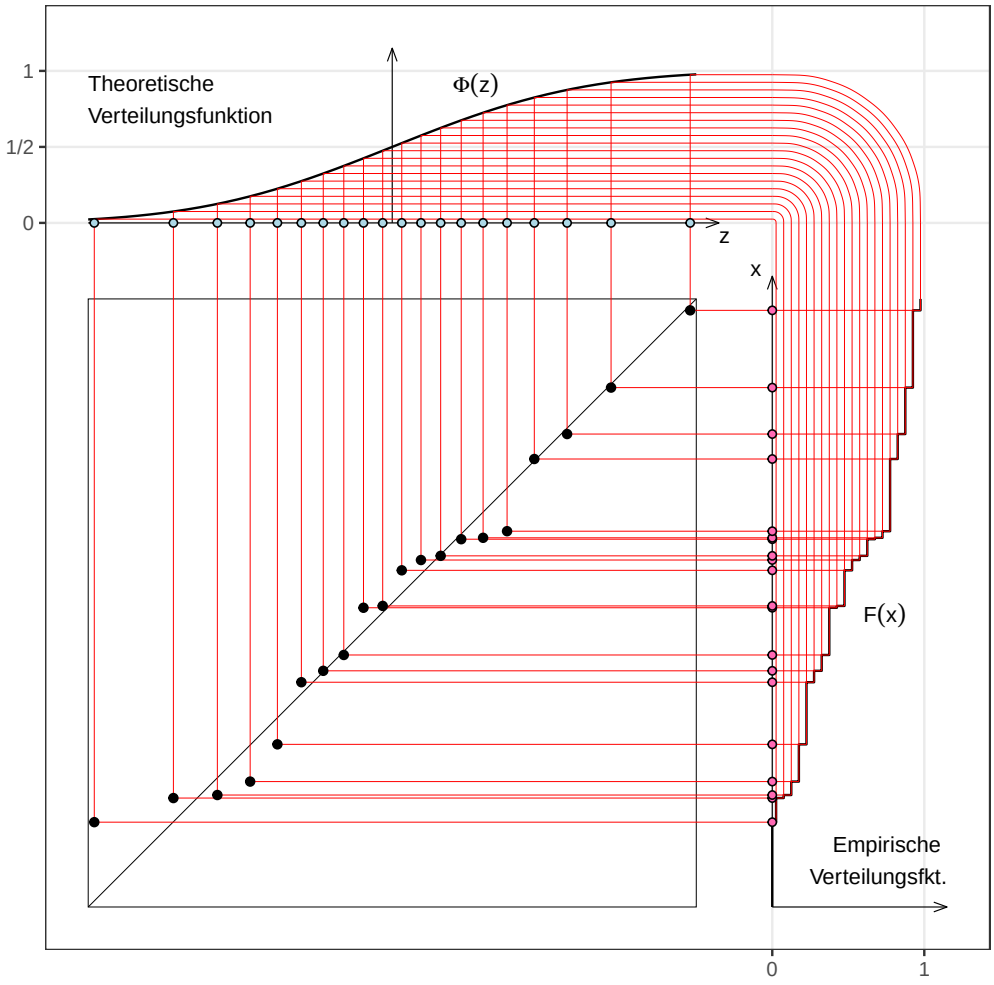


Abbildung 7: Entstehung des Normal-Quantil-Plots: Die roten Punkte sind die empirisch erhobenen Werte $x_{(i)}$. Zu jedem Wert kann man das zugehörige p_i (den Anteil der Werte, die kleiner gleich $x_{(i)}$ sind) aus der empirischen Verteilungsfunktion ablesen. Nun wird für jedes p_i das zugehörige Quantil z_{p_i} der Standardnormalverteilung bestimmt. Das sind die blauen Punkte. Die Lage der roten und blauen Punkte auf der x-Achse und der z-Achse bestimmt die Lage der Punkte im Normal-Quantil-Plot

übereinstimmen, desto näher liegen die Punkte $(z_{(i)}, x_{(i)})$ an der Geraden $x = \bar{x} + \tilde{s} \cdot z$. Für vier verschiedene Merkmale ist dieser Zusammenhang in Abbildung 8 dargestellt.

Beispiel Geschwindigkeitsmessung (Fortsetzung): Für die Geschwindigkeitsmessung an der Universitätsstraße stellt die Normalverteilung nur eine unzureichende Approximation der wirklichen Verteilung dar. Das sieht man auch an dem NQ-Plot in Abbildung 9.

Approximation von Dichtekurven

Häufig lässt sich die Verteilung eines Merkmals nicht sinnvoll durch eine Normalverteilung darstellen. Allerdings ist es für eine kontinuierliche Variable trotzdem häufig wünschenswert, die Verteilung durch eine Dichtekurve darzustellen. Hierfür sprechen folgende Gründe:

- Die willkürlich zu wählende Klassenbreite beeinflusst den optischen Eindruck
- Der Verlauf eines Histogramms kann sehr unregelmäßig sein
- Eine eigentlich stetige Funktion wird durch eine Treppenkurve dargestellt

Diese Nachteile lassen sich vermeiden, wenn man die Dichtekurve geeignet durch eine stetige Funktion approximiert.

Ausgangspunkt hierfür ist eine Kernfunktion $K : \mathbb{R} \to \mathbb{R}$, mit den Eigenschaften einer Dichtekurve, siehe Definition 0.2. Diese Kernfunktion wird über die Datenpunkte geschoben und dabei die Längen der Linien von den Beobachtungen x_i bis zum Graphen der Kernfunktion bestimmt. Die Summe der Längen geteilt durch die Anzahl der Werte n ist der Funktionswert der approximierten Dichtekurve (siehe Abbildung 10).

Gebräuchlich sind die Kerne mit den Funktionsgleichungen

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{für} \quad -1 \le u \le 1 \\ 0 & \text{sonst} \end{cases} \qquad \text{Epachenikov-Kern,}$$

$$K(u) = \begin{cases} \frac{15}{16}(1-u^2)^2 & \text{für} \quad -1 \le u \le 1 \\ 0 & \text{sonst} \end{cases} \qquad \text{Bisquare-Kern,}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \qquad \text{Gauß-Kern.}$$

Eine Darstellung der Graphen der Kernfunktionen ist in Abbildung 11 zu sehen.

Damit können wir den sogenannten Kerndichteschätzer definieren.

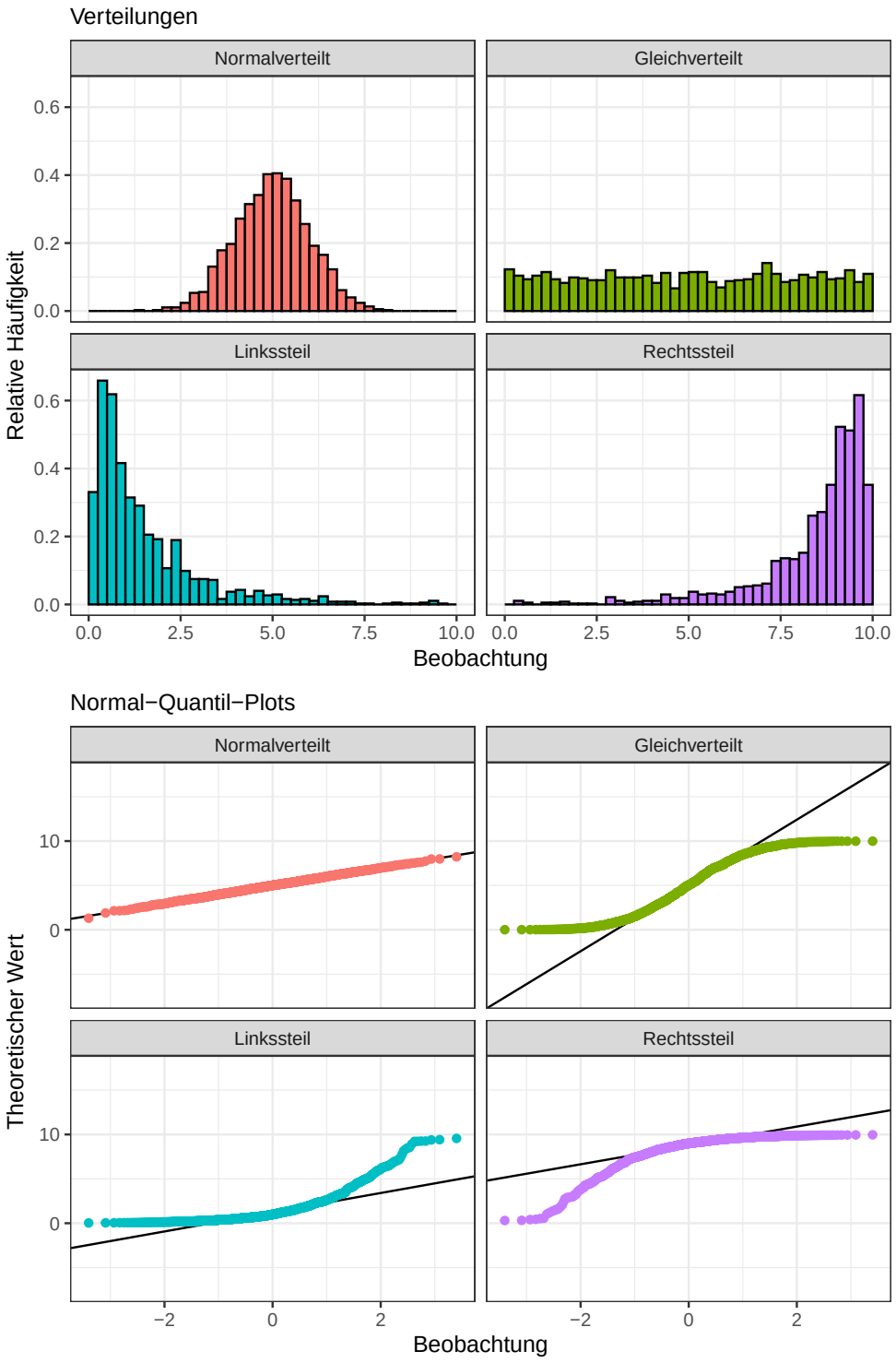


Abbildung 8: Verteilungen von Beobachtungen (oben) und zugehörige NQ-Plots (unten). Für das näherungsweise normalverteilte Merkmal liegen die Punkte fast auf einer Geraden, für die Gleichverteilung ober- und unterhalb der Geraden. Die asymetrischen Verteilungen ergeben eine konkave (linkssteile Verteilung) beziehungsweise eine konvexe (rechtssteile Verteilung) Anordnung der Punkte im NQ-Plot

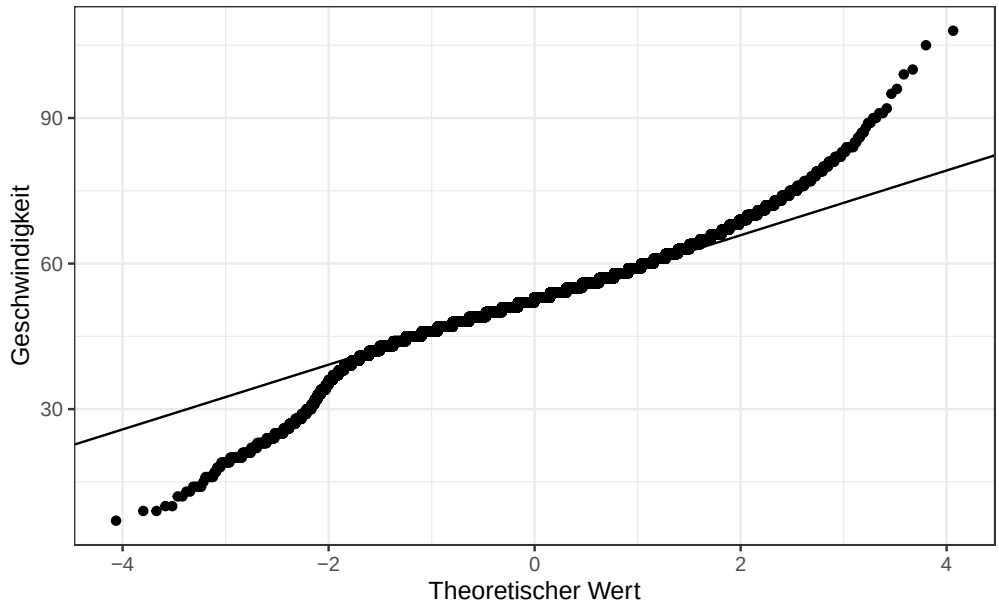


Abbildung 9: NQ-Plot zur Geschwindigkeitsmesung. Da die wirkliche Verteilung steilere Flanken besitzt als die angepasste Normalverteilung (Abbildung 6) ergibt sich der hierfür charakteristische Verlauf der Punkte: Kleine Werte liegen unterhalb der Geraden, große Werte oberhalb

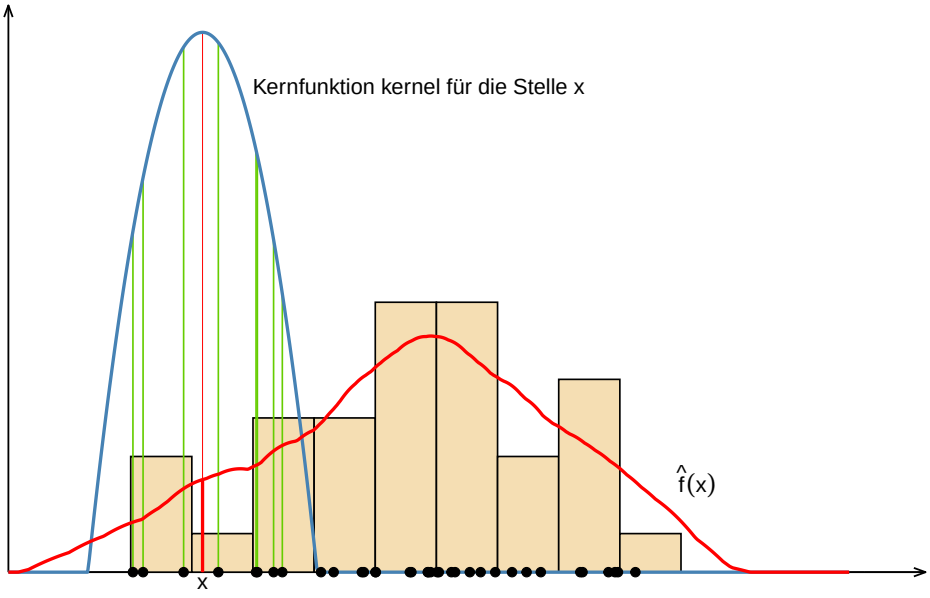


Abbildung 10: Approximation einer Dichtekurve

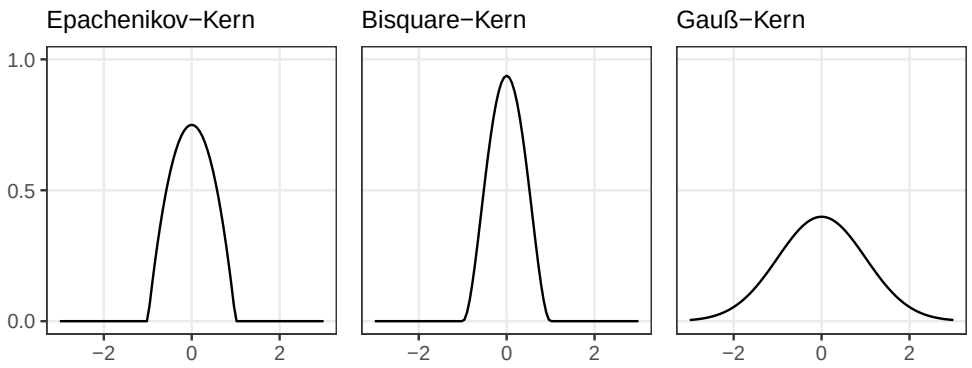


Abbildung 11: Gebräuchliche Kernfunktionen

Definition 0.3 (Kerndichteschätzer)**.** Zu einer Kernfunktion $K : \mathbb{R} \to \mathbb{R}$ und gegebenen Daten $x_1, x_2, \ldots, x_n$ ist die Funktion $\hat{f} : \mathbb{R} \to \mathbb{R}$ mit

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

ein Kerndichteschätzer für die unbekannte Dichtekurve f. Dabei legt der Parameter h die Breite des Kerns fest. Da K eine Dichtekurve ist lässt sich leicht zeigen, dass $\hat{f}$ ebenfalls eine Dichtekurve sein muss.

Index