Demokurs

Modul 31101 Grundlagen der Wirtschaftsmathematik und Statistik

Kurs 40601 Grundlagen der Statistik

13. Juli 2010

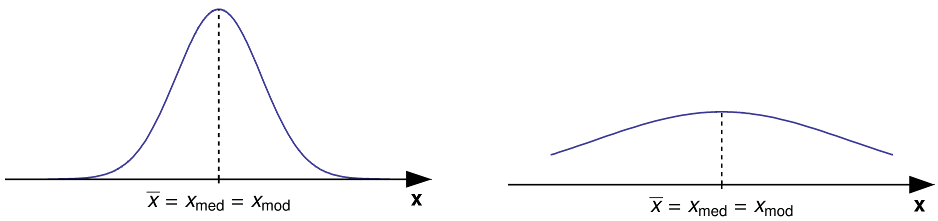
2.4 Schiefe und Wölbung einer Verteilung

Zur Beschreibung von Verteilungen können auch die Begriffe **Symmetrie, Schiefe**, **Steilheit** und **Wölbung** herangezogen werden.

Symmetrie, Schiefe, Steilheit, Wölbung

Die Schiefe gibt an, wie stark eine eingipflige Verteilung nach rechts bzw. nach links geneigt ist. Eine Verteilung ist symmetrisch, wenn sie in Bezug auf das arithmetische Mittel symmetrisch ist. Merkmalsausprägungen, die um den gleichen Betrag nach unten bzw. oben vom arithmetischen Mittel abweichen, haben dann die gleiche absolute bzw. relative Häufigkeit. Des Weiteren kann zwischen flachen und steilen Verteilungen unterschieden werden.

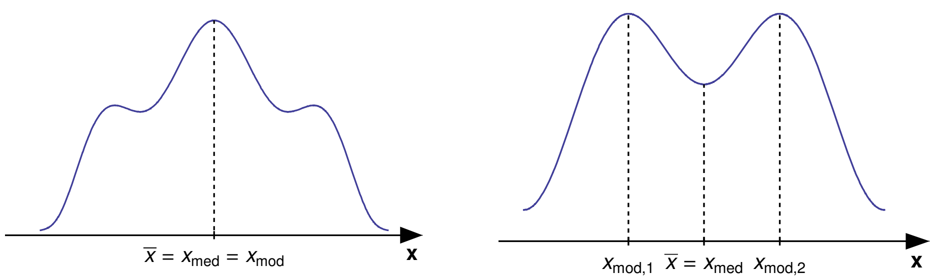
Arithmetisches Mittel, Median und Modalwert stimmen bei einer **eingipfligen, symmetrischen Verteilung** überein.



steile Verteilung flache Verteilung

Abbildung 2.4.1: *Eingipflige, symmetrische Verteilungen*

Auch bei **mehrgipfligen, symmetrischen Verteilung** stimmen das arithmetische Mittel und der Median stets überein. Aufgrund der Mehrgipfligkeit können jedoch mehrere Modalwerte auftreten.

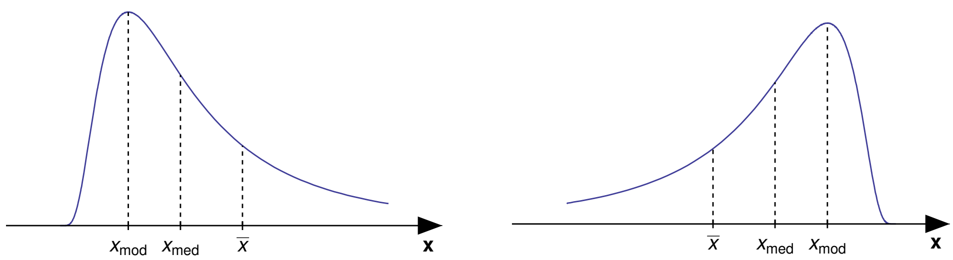


unimodale Verteilung bimodale Verteilung

Abbildung 2.4.2: *Mehrgipflige, symmetrische Verteilungen*

rechtsschief (linkssteil)
linksschief (rechtssteil)

Eingipflige Verteilungen können nach ihrer Schiefe beurteilt werden, wobei zwischen rechtsschief und linksschief unterschieden werden kann. **Rechtsschiefe Verteilungen** steigen von links nach rechts steil an und fallen dann nach rechts flach ab. **Linksschiefe Verteilungen** steigen dagegen von rechts nach links steil an und fallen dann nach links flach ab. Daher werden diese auch als **linkssteile** bzw. **rechtssteile Verteilungen** bezeichnet.



linkssteile bzw. rechtsschiefe Verteilung

linksschiefe bzw. rechtssteile Verteilung

Abbildung 2.4.3: *Schiefe Verteilungen*

Fechnersche Lageregel

Mit Hilfe der **Fechnerschen Lageregel** können Verteilungen hinsichtlich der Schiefe betrachtet werden, ohne eine Grafik hinzuzuziehen.

Fechnersche Lageregel:

$$x_{mod} = \overline{x} = x_{med} \quad \text{für symmetrische Verteilungen,}$$
$$x_{mod} < x_{med} < \overline{x} \quad \text{für rechtsschiefe Verteilungen,}$$
$$\overline{x} < x_{med} < x_{mod} \quad \text{für linksschiefe Verteilungen.}$$

Als Maßzahl zur Beurteilung der Schiefe einer eingipfligen Verteilung metrischer Merkmale kann der Momentenkoeffizient berechnet werden.

Momentenkoeffizient der Schiefe

Der **Momentenkoeffizient der Schiefe g_3** lautet:

$$g_3 = \frac{m_3}{\tilde{s}^3} = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})^3}{\left[\frac{1}{n}\sum_i^n (x_i - \overline{x})^2\right]^{\frac{3}{2}}}.$$

3.3 Arten von Zusammenhängen

Die Aufstellung und Analyse zweidimensionaler Verteilungen hat insbesondere den Zweck folgende Fragestellungen zu untersuchen:

- Besteht eine Abhängigkeit zwischen den betrachteten Merkmalen? Wenn ja, in welcher Ausprägung und von welcher Art liegt diese vor?
- Wie kann eine vorhandene Abhängigkeit quantitativ beschrieben werden?

Die Frage, ob überhaupt Abhängigkeit vorliegt, kann bei quantitativen Merkmalen in vielen Fällen durch Streudiagramme geklärt werden.

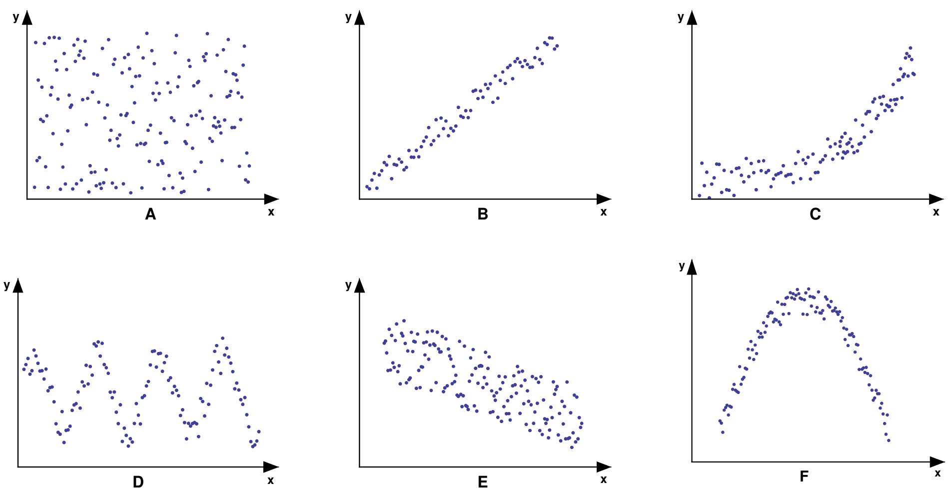


Abbildung 3.3.1: *Streudiagramme mit unterschiedlichen Abhängigkeiten der Merkmale*

In der Darstellung A sind keine Anzeichen von Abhängigkeit zwischen den Merkmalen X und Y zu erkennen. Die anderen Darstellungen weisen auf eine Abhängigkeit hin, welche jedoch nicht exakt im Sinne der Eindeutigkeit ist, wie sie bei Funktionen vorliegt. Die Abhängigkeit besteht tendenziell.

In B hat das Merkmal Y im Durchschnitt um so größere Werte, je größer die Werte von X sind, während in der Darstellung E die Werte von Y im Durchschnitt um so kleiner sind, je größer die Werte von X sind.

Beide Grafiken veranschaulichen einen linearen Zusammenhang. Grafik D deutet auf einen periodischen Zusammenhang hin, während in C die Werte von Y für zunehmende Werte des Merkmals X exponentiell steigen. Die Darstellung F verdeutlicht einen quadratischen Zusammenhang.

Allgemein ist es anhand einer grafischen Darstellung i.d.R. nicht möglich, Aussagen folgender Form zu treffen: „Wenn X den Wert x annimmt, dann nimmt Y immer genau den Wert y an.“

Die Überlegungen zu Bild 3.3.1 zeigen, dass bei quantitativen Merkmalen die grafische Darstellung der zweidimensionalen Verteilung in einem Streudiagramm bereits ausreichen kann, um grundlegende Aussagen über eventuelle Abhängigkeiten zu machen. Dabei ist oft auch bereits eine Aussage über den Grad (die Ausgeprägtheit) eines Zusammenhangs möglich.

Für Merkmale, die nach einer Nominalskala oder einer Ordinalskala klassifiziert werden können, ist eine grafische Darstellung einer zweidimensionalen Verteilung in einem Streudiagramm nicht möglich. Hier können nicht so schnell Urteile über eventuelle Abhängigkeiten gemacht werden.

Als Aufgabenstellung für die Untersuchung von Zusammenhängen zwischen Merkmalen kann festgehalten werden:

- Bestimmung von Maßzahlen, die angeben, ob überhaupt ein Zusammenhang vorliegt und wie ausgeprägt dieser ist. Die Maßzahlen werden im Allgemeinen so definiert, dass sie bei Unabhängigkeit den Wert 0 annehmen. Bei einem vorhandenen Zusammenhang soll aus dem Wert der Maßzahl deutlich werden, wie ausgeprägt der Zusammenhang ist.

- Bestimmung von Funktionen, die die durchschnittliche Tendenz eines Zusammenhangs wiedergeben. Das ist nur für metrisch messbare Merkmale möglich. Diese Funktionen sind so zu bestimmen, dass sie die in einem Zusammenhang steckende Tendenz möglichst gut beschreiben.

Zur Beschreibung des Zusammenhangs können für die Streudiagramme in Abbildung 3.3.1 folgende Funktionen verwendet werden: Für C bietet sich eine Exponentialfunktion, für D eine Sinusfunktion und für F eine Parabel an, während die Tendenz des Zusammenhangs in B und E am besten durch eine steigende bzw. fallende Gerade beschrieben wird.

1.4.4 Statistische Definition der Wahrscheinlichkeit

Die in diesem Abschnitt behandelte statistische Definition der Wahrscheinlichkeit beruht auf einem Zusammenhang zwischen relativen Häufigkeiten und Wahrscheinlichkeiten.

Betrachtet wird ein Zufallsexperiment, das unter völlig gleichen Bedingungen nacheinander n-mal durchgeführt wird. Nach jeder Durchführung wird die relative Häufigkeit für das Auftreten des Ereignisses A bestimmt. In den ersten Versuchen schwanken die berechneten relativen Häufigkeiten für das Auftreten des Ereignisses A sehr stark. Je größer die Anzahl der Versuche des Zufallsexperimentes ist, desto enger schwanken die relativen Häufigkeiten um einen festen Wert.

Beispiel 1.4.6:

Ein Würfel wurde 200-mal hintereinander geworfen Nach jedem Durchgang wurde die relative Häufigkeit für das Ereignis A=„Auftreten der Augenzahl 6“ registriert. Für jeden Durchgang ist die Anzahl n der Würfe (x-Achse) und die zugehörige relative Häufigkeit $f_n(A)$ (y-Achse) in Bild 1.4.1 grafisch dargestellt. Dieser Vorgang wurde 9 mal wiederholt.

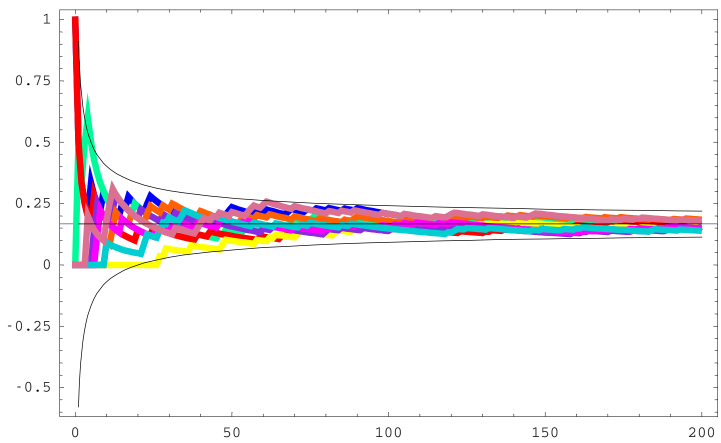


Abbildung 1.4.1: *Relative Häufigkeit für das Auftreten von „Augenzahl 6“ in Abhängigkeit von der Anzahl der Würfelwürfe*[4]

[4]siehe interaktive Mathematica-Applet „frequenzneu.nbp“ auf http://www.fernuni-hagen.de/ls_statistik/forschung/multimedia/eigene.shtml

In Beispiel 1.4.6 schwanken die relativen Häufigkeiten immer weniger um den Wert $\frac{1}{6}$. Je häufiger das Zufallsexperiment durchgeführt wird, desto besser stabilisieren sich die relativen Häufigkeiten. Offensichtlich streben die relativen Häufigkeiten einem „Grenzwert“ zu. Dieser Grenzwert entspricht der Wahrscheinlichkeit für das Ereignis A. Diese Eigenschaft der relativen Häufigkeit führt zu der statistischen Definition der Wahrscheinlichkeit.

statistische Definition der Wahrscheinlichkeit

Nach der statistischen Definition ist die Wahrscheinlichkeit für das Auftreten des Ereignisses A gleich dem Grenzwert der relativen Häufigkeiten, der sich ergibt, wenn das Zufallsexperiment unendlich oft durchgeführt wird:

$$\boldsymbol{P}(A) = \lim_{n \to \infty} f_n(A).$$

Da es in der Wirklichkeit unmöglich ist, ein Zufallsexperiment unendlich oft durchzuführen, ist es ebenso unmöglich, auf die angegebene Art eine Wahrscheinlichkeit zu bestimmen.

Die statistische Definition der Wahrscheinlichkeit bedeutet, dass über die Berechnung von relativen Häufigkeiten zumindest eine Näherung oder Schätzung der dem Zufallsexperiment zugrunde liegenden Wahrscheinlichkeiten bestimmt werden kann. In dem Fall wird von den sogenannten **empirischen Wahrscheinlichkeiten** gesprochen.

empirische Wahrscheinlichkeit

1.4.5 Das Gesetz der großen Zahlen

Im Zusammenhang mit Anwendungen der Statistik wird sehr häufig vom sogenannten Gesetz der großen Zahlen gesprochen, wobei zu beachten ist, dass es verschiedene „Gesetze der großen Zahlen“ gibt.

Die relative Häufigkeit für ein bestimmtes Ereignis eines Zufallexperimentes nähert sich mit zunehmender Anzahl der Wiederholungen des Zufallexperimentes einem bestimmten Wert immer mehr an. Dieser Wert ist im allgemeinen unbekannt, da die relative Häufigkeit mit ihm erst dann übereinstimmt, wenn unendlich viele Versuche durchgeführt werden. Bild 1.4.1 veranschaulicht, dass die relativen Häufigkeiten sich

3.3 Normalverteilung

3.3.1 Definition der Normalverteilung

Die **Normalverteilung** ist die wichtigste stetige Verteilung. Sie spielt bei nahezu allen Anwendungen der Statistik eine große Rolle.

Normalverteilung

Dichtefunktion

Die Dichtefunktion der **Normalverteilung** lautet:

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

Zu der Dichtefunktion $f_X(x)$ existiert keine elementare Stammfunktion, so dass die **Verteilungsfunktion** der Normalverteilung nicht mehr mit Hilfe elementarer Funktionen darstellbar ist. Die Werte der Verteilungsfunktion können mittels numerischer Integrationsverfahren oder spezieller Tabellen (s. Gesamtglossar) angegeben werden. Die Parameter der Normalverteilung lauten:

Erwartungswert und Varianz

$$E(X) = \mu \quad \text{und} \quad \mathrm{Var}(X) = \sigma^2.$$

Eine normalverteilte Zufallsvariable X wird als $N(\mu, \sigma^2)$-verteilt bezeichnet. Die Schreibweise lautet $\boldsymbol{X \sim N(\mu, \sigma^2)}$.

Erwartungswert und Varianz bzw. Standardabweichung der Normalverteilung lassen sich also unmittelbar aus der Dichtefunktion ablesen.

Aus der Dichtefunktion ergibt sich, dass die Normalverteilung in einem konkreten Fall durch die Angabe von μ und σ^2 jeweils spezifiziert werden muss. Es gibt also nicht nur eine Normalverteilung, sondern eine ganze Klasse von Normalverteilungen. Die Dichtefunktion der Normalverteilung hat folgende typische Gestalt:[5]

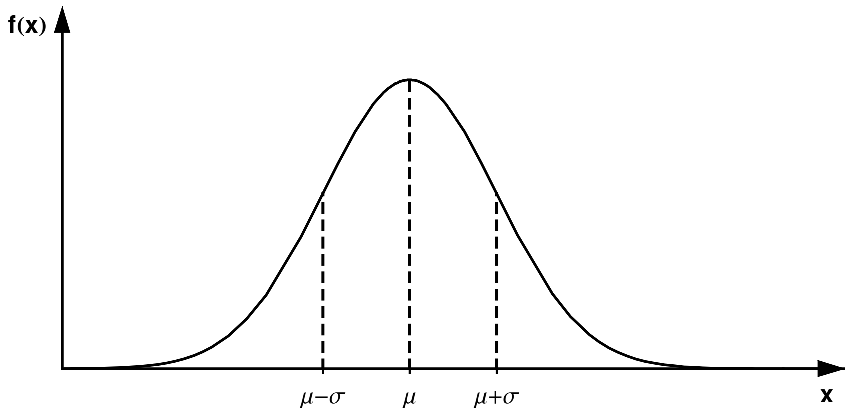


Abbildung 3.3.1: *Dichtefunktion der Normalverteilung*

[5]siehe interaktive Mathematica-Applets auf der Homepage des Lehrstuhls http://www.fernuni-hagen.de/ls_statistik/forschung/multimedia/

Die Dichtefunktion ist symmetrisch und hat ihren Gipfel bei $x = \mu$. An den Stellen $x = \mu - \sigma$ und $x = \mu + \sigma$ befinden sich Wendepunkte. Aufgrund der Symmetrie gilt, dass Erwartungswert, Median und Modalwert übereinstimmen. In der Abbildung 3.3.2 sind Normalverteilungen für verschiedene Werte von μ und σ^2 dargestellt.

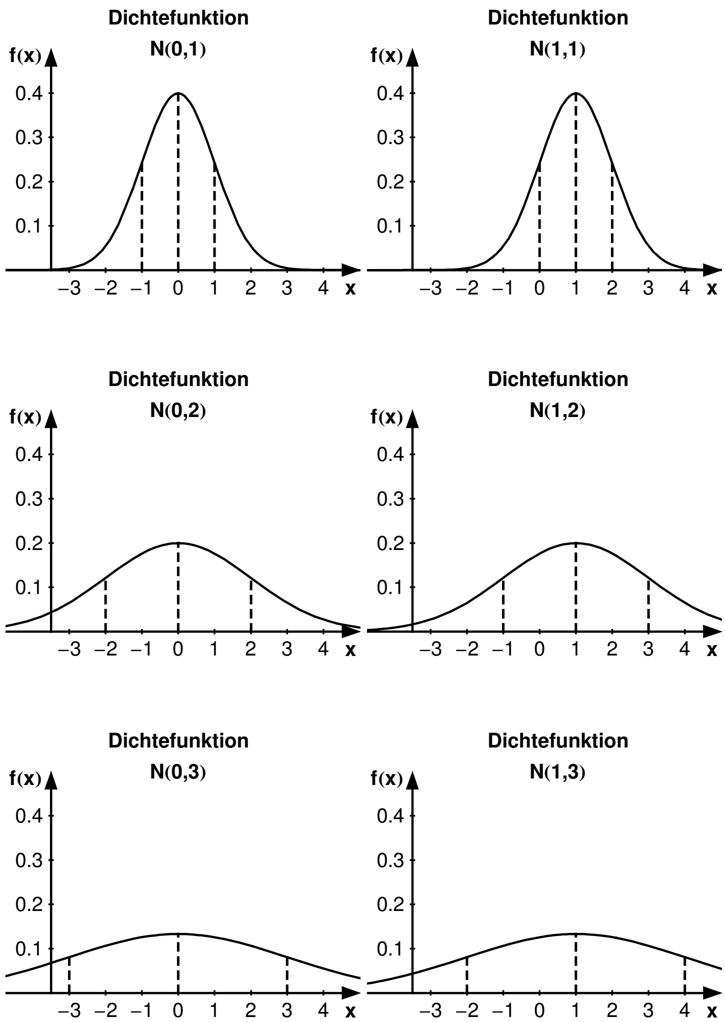


Abbildung 3.3.2: *Verschiedene Normalverteilungen*

> Ist eine Zufallsvariable normalverteilt mit dem Erwartungswert 0 und der Varianz 1, so liegt eine **Standardnormalverteilung** vor. In diesem Zusammenhang wir die Zufallsvariable mit Z bezeichnet, d.h. $\boldsymbol{Z \sim N(0,1)}$.

Standard-normalverteilung

3.4 Zusammenhang zwischen Konfidenzintervall und Testverfahren

Für eine normalverteilte Zufallsvariable X mit unbekanntem Parameter μ und **bekannter** Varianz σ^2, d.h. $X \sim N(\mu, \sigma^2)$ lautet das zweiseitige $(1-\alpha)$-Konfidenzintervall für μ

$$\text{KI}: \qquad \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,;\,\overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right].$$

Hierbei sind die Grenzen Zufallsvariablen, d.h. die Grenzen sind zufällige Größen. Der Parameter μ ist dagegen eine unbekannte, aber feste Größe.

Wird ein Parametertest für μ durchgeführt, so geben die kritischen Grenzen c_u und c_o einen Bereich für den Stichprobenmittelwert $\overline{X}$ an, bei dem die Nullhypothese $H_0: \ \mu = \mu_0$ nicht abgelehnt werden kann.

$$\text{Test}: \qquad \left[\mu_0 - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,;\,\mu_0 + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$$

Die Grenzen des Bereiches $[c_u; c_o]$ sind feste Größen, während die Prüfgröße $\overline{X}$ zufällig ist.

Die Grenzen des einen Bereiches können aus den Grenzen des anderen Bereiches hergeleitet werden ($\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$).

$$\begin{aligned}
& \boldsymbol{P}\left(\mu - z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}} \le \overline{X} \le \mu + z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}\right) \\
= \ & \boldsymbol{P}\left(-z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}} \le \overline{X} - \mu \le z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}\right) \\
= \ & \boldsymbol{P}\left(-\overline{X} - z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}} \le -\mu \le -\overline{X} + z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}\right) \\
= \ & \boldsymbol{P}\left(\overline{X} - z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}} \le \mu \le \overline{X} + z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}\right) = 1 - \alpha
\end{aligned}$$

Die Nullhypothese $H_0: \ \mu = \mu_0$ kann auch mittels des Konfidenzintervalls überprüft werden. Überdeckt das Konfidenzintervall nicht den Wert μ_0, so kann die Nullhypothese abgelehnt werden. Die Testentscheidung mittels der Prüfgröße $\overline{X}$ und die Testentscheidung mittels des Konfidenzintervalls liefern dasselbe Ergebnis. Das Konfidenzniveau „$1 - \alpha$“ entspricht „1−Signifikanzniveau“ beim Testen.

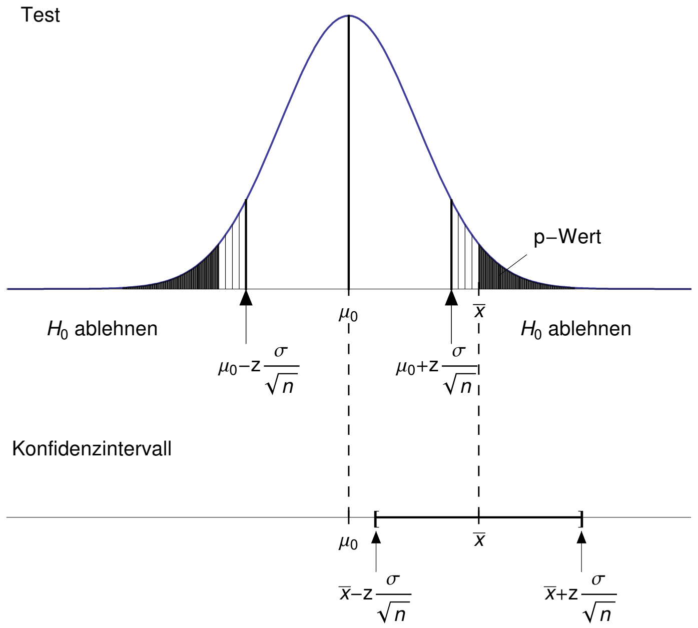


Abbildung 3.4.1: *Zusammenhang zwischen Konfidenzintervall und Testverfahren*

Zusammenfassend kann formuliert werden:

Das Konfidenzintervall mit der Wahrscheinlichkeitsaussage

$$\boldsymbol{P}\left(\overline{X} - z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}} \leq \mu \leq \overline{X} + z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}\right) = 1 - \alpha$$

ist ein Intervall, dessen Grenzen Zufallsvariablen sind und welches den unbekannten (aber festen) Parameter μ mit der Wahrscheinlichkeit $1 - \alpha$ überdeckt.

Das Intervall mit der Wahrscheinlichkeitsaussage

$$\boldsymbol{P}\left(\mu - z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}} \leq \overline{X} \leq \mu + z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}\right) = 1 - \alpha$$

hat feste Grenzen und gibt die Wahrscheinlichkeit $1 - \alpha$ an, mit der die Zufallsvariable $\overline{X}$ in den angegebenen Bereich fällt.