|  |
| --- |
| IUBH |
| Angewandte Statistik |
| MMET01-01\_E |

# Übergeordnete Lernziele

Die Welt, in der wir leben, enthält riesige Mengen von Daten, die in jeder Sekunde, Minute und Stunde gesammelt werden. Daten von unseren Telefonen, Autos, Computern oder sogar von unseren Körperfunktionen werden ununterbrochen auf Server geladen. In diesen Daten sind Erkenntnisse enthalten, die ohne die Methoden aus dem Studium der Statistik kaum durch Menschen zu entschlüsseln sind. Die Statistik ist ein Spezialgebiet der Mathematik, das sich mit der Sammlung, Analyse und Interpretation numerischer Daten befasst. Im Studienfach Angewandte Statistik werden wir uns mit diesem Spezialgebiet beschäftigen und seine spezielle Rolle und Bedeutung für praktische Entscheidungsprozesse im Geschäftsleben ergründen. Insbesondere werden wir lernen, die Relevanz der Daten für die Beantwortung empirischer Fragen zu verstehen, statistische Methoden auf den allgemeinen Kontext konkreter Probleme anzuwenden und statistische Probleme mit Spezialsoftware für Statistik zu lösen.

Die Statistische Analyse (auch schlicht Statistik genannt) ist ein wichtiges Werkzeug, das Führungskräfte benutzen, um trotz unsicherer Bedingungen Entscheidungen zu treffen. Eine statistische Kenngröße ist eine einzelne Zahl, die einen Datensatz beschreibt. In diesem Kurs lernen die Studierenden, statistische Werkzeuge anzuwenden, die bessere geschäftliche Entscheidungen ermöglichen.

# Lektion 1 – Grundlagen

### Lernziele

Nach der Bearbeitung dieser Lektion werden Sie in der Lage sein, ...

... Kenngrößen der deskriptiven Statistik zu definieren und zu berechnen.

... mit Hilfe der induktiven Statistik Erkenntnisse aus einer Stichprobe zu gewinnen.

... Wahrscheinlichkeiten zu berechnen und zu interpretieren.

# 1. Grundlagen

## Aus der Praxis

Das Finanzdienstleistungsunternehmen TechnoBank war in den vergangenden Jahren mit zunehmendem Wettbewerb konfrontiert und möchte nun Daten sammeln, um mehr über die Stimmung und Loyalität seiner Kundschaft zu erfahren. Das Marketing-Team unter der Leitung von Marketingchefin (CMO) Kristin Stark entscheidet, einen Fragebogen zu erstellen und eine Umfrage unter aktuellen und ehemaligen Kund:innen durchzuführen. Kristin beauftragt den Leiter der Marktforschungsabteilung, Stefan Rider, das Projekt zu übernehmen. Stefan umreißt den Prozess der Fragebogenerstellung und der Ermittlung der Befragungsteilnehmer aus dem Online-Banking-Programm der TechnoBank. Er plant, die Kund:innen systematisch zu befragen (jeweils eine:n von fünf), eine deskriptive Statistik zu erstellen, um Kenngrößen für Kundenloyalität und -zufriedenheit zu ermitteln, und dann diese Daten zu benutzen, um die Wahrscheinlichkeit für Kundenabwanderung zu berechnen. Kristin möchte in der nächsten Woche über den neuesten Stand informiert werden, und Stefan denkt über die folgenden vier Fragen nach:

1. Wie groß sollte die Stichprobe sein?
2. Wird diese Art von Stichprobenermittlung brauchbare Rückschlüsse über die Gesamtbevölkerung erlauben?
3. Welche Kenngrößen für zentrale Tendenzen und Streubereiche, und welche Diagramme sind geeignet, um aus den Daten eine Aussage zu entnehmen?
4. Welches sind die richtigen Fragen, die gestellt werden müssen, um die richtigen Daten zur Loyalität zu erhalten, und wie soll die Wahrscheinlichkeit der Kundenabwanderung dann berechnet werden?

Dies sind nur einige der statistischen Überlegungen, mit denen wir uns in dieser Lektion befassen werden, um so wichtige Unternehmensentscheidungen, wie die von TechnoBank, zu ermöglichen.

## 1.1 Deskriptive Statistik

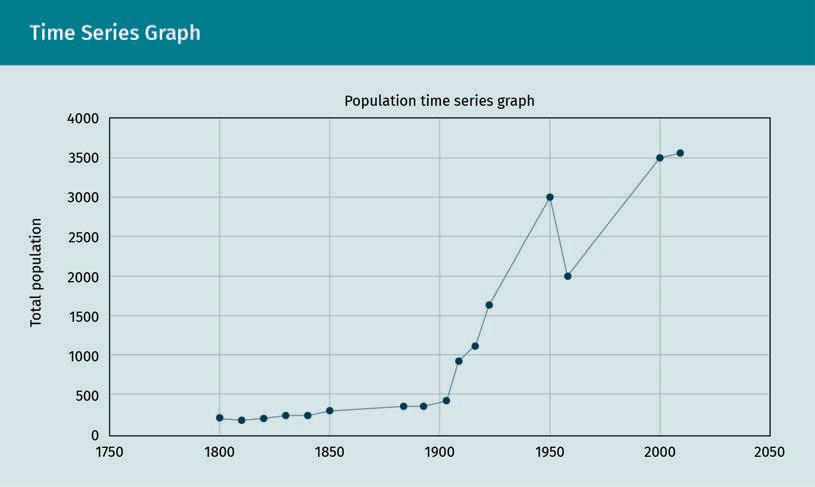
Die deskriptive Statistik bildet die Grundlage jeder statistischen Analyse. Wenn wir die Daten aus einem Datensatz verstehen wollen, ist der erste Schritt, sie zu beschreiben. Die deskriptive Statistik ist das wichtigste Werkzeug, um dieses Ziel zu erreichen. Ohne die Hilfe der deskriptiven Statistik können Menschen nicht mehr als einige wenige Datenreihen in einer Tabelle sinnvoll interpretieren. Deskriptive Kenngrößen ermöglichen es Forschenden, alle Einzelbeobachtungen einer Variablen zu verwenden und eine Zusammenfassung über die wahrscheinlichsten Werte und deren Unterschiede zueinander zu erstellen.

Kenngrößen der deskriptiven Statistik werden häufig in Zusammenhang mit anderen Kenngrößen angegeben, da eine einzige Zahl in der Regel nicht ausreicht, um die Merkmale der Variablen in einem Datensatz vollständig zu erklären. So ist es beispielsweise irreführend, den arithmetischen Mittelwert als Schätzwert für die Grundgesamtheit anzugeben, wenn die Daten stark gestreut sind. Darüber hinaus gibt es einige Phänomene, die bimodal sind und daher als zwei verschiedene Verteilungen angegeben werden sollten. Schließlich können einige Variablen, die in Verbindung miteinander erfasst werden, den Forschenden eine Vorstellung davon vermitteln, was in der Verteilung vor sich geht (z. B. könnten signifikant voneinander arithmetischer Mittelwert und Median auf eine Nicht-Normalverteilung hinweisen). Die deskriptive Statistik wird oft als eine einfache Form der Datenanalyse auf einer Entwicklungskurve betrachtet, die die Wettbewerbsvorteile der Analytik darstellt. Dies ist ein Irrtum, denn Kenngrößen aus der deskriptiven Statistik werden zur Unterstützung vieler übergeordneter Analysen benötigt.

### Statistische Fachbegriffe

Was wir uns zu Beginn eines Statistikstudiums als erstes verschaffen müssen, ist eine Übersicht über die benötigten Fachbegriffe. Wenn wir Daten erörtern, müssen wir eine Variable definieren, die ein veränderliches Merkmal eines Forschungsgegenstands ist. Beispiele für Variablen sind etwa Alter, Geschlecht und Einkommen. Diese Datenelemente variieren von einem erfassten Individuum zum anderen. Der Begriff „Daten“ ist ein Plural und bezieht sich auf die Gesamtheit oder auf einen Teil der gesammelten Beobachtungen zu einer Variablen, während ein Datensatz alle zu Forschungszwecken gesammelten Werte zu einer oder mehreren Variablen umfasst. Für viele erfolgt der erste Kontakt mit Variablen, Daten und Datensätzen bei der Verwendung von Excel-Tabellen.

Als nächstes werden wir uns mit den Datentypen befassen. Kategoriale Daten, die auch als qualitative Daten bezeichnet werden, sind Daten, die mit Worten und nicht mit Zahlen beschrieben werden. Ein Beispiel für kategoriale Merkmale sind Farben. Numerische Daten sind das Ergebnis zahlenmäßiger Messungen oder mathematischer Operationen. Beispiele für numerische Variablen sind Einkommen oder Tageszeit. Die Hauptarten numerischer Daten sind diskrete und kontinuierliche Variablen. Diskrete Variablen sind zählbar, wie z. B. die Anzahl von Kund:innen, die einen Laden betreten. Kontinuierliche Variablen sind nicht zählbar, beispielsweise Gewicht und Zeit, die beide unendlich viele verschiedene Messwerte haben können. Ein weiterer Datentyp, der insbesondere im Zusammenhang mit der Unternehmensanalyse zu erwähnen ist, sind Zeitreihen. Zeitreihen werden über einen bestimmten Zeitraum hinweg aufeinanderfolgend gemessen. Die Abbildung unten stellt eine Zeitreihe als Diagramm dar..



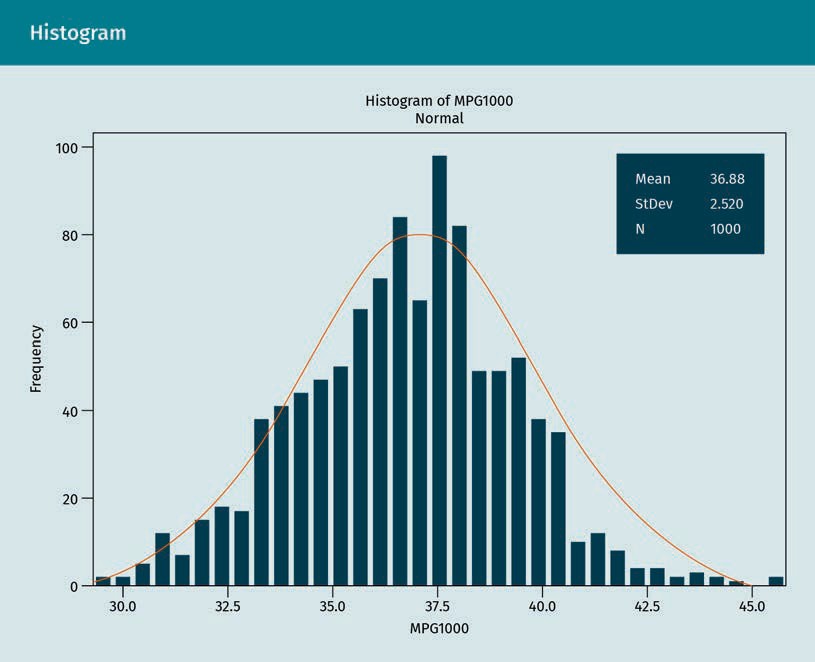
### Skalenniveaus

Die Art der Skala, in der Daten erfasst werden, ist wichtig für die Auswahl der richtigen statistischen Techniken. Die vier möglichen Skalenarten von Variablen sind Nominalskala, Ordinalskala, Intervallskala und Verhältnisskala. Nominale Daten, abgeleitet aus dem lateinischen „nomen“ oder Name, identifizieren eine Kategorie (Doane, 2016). Kategoriale Daten, wie z. B. Haarfarbe, werden auf einer Nominalskala gemessen. Als nächstes folgt die Ordinalskala. Ordinale Daten, wie z. B. die Kreditwürdigkeit von Unternehmen, haben eine Rangfolge, aber die Differenz zwischen den Beobachtungen ist außerhalb der Rangstufen nicht aussagekräftig. Dann kommen wir zur Intervallskala, wo der Unterschied zwischen den einzelnen Messpunkten eine Bedeutung hat. Es gibt jedoch keinen natürlichen Nullpunkt, und die Differenz zwischen den Skalenpunkten kann nicht für eine Verhältnisberechnung verwendet werden, da diese nicht vollständig bekannt sind. Grad Fahrenheit und Grad Celsius sind Beispiele für Messwerte oder Variablen auf einer Intervallskala.

Die letzte Skala, mit der wir uns befassen, ist die Verhältnisskala. Variablen, die auf einer Verhältnisskala gemessen werden, haben einen natürlichen Nullpunkt, so dass sie zueinander ins Verhältnis gesetzt werden können. Die Temperatureinheit Kelvin ist ein Beispiel für eine Variable auf einer Verhältnisskala (im Gegensatz zu Grad Fahrenheit oder Grad Celsius), da Kelvin durch die kinetische Energie kleiner Partikel definiert wird, die gemäß den Gesetzen der Thermodynamik einen natürlichen Nullpunkt hat. Verhältnis- und Intervallskalen sind für statistische Techniken am besten geeignet. Das Einkommen ist ein Beispiel für eine Variable auf einer Verhältnisskala, obwohl einige immer noch argumentieren würden, dass die Möglichkeit eines Kredits eine Person oder ein Unternehmen unter Null bringen könnte. Dennoch gibt es einen Punkt, an dem kein Kredit mehr erhältlich ist, so dass ein natürlicher Nullpunkt entsteht (Ahmad, 2016; Doane, 2016).

### Mittelwerte

Mittelwerte beruhen auf der Tatsache, dass viele Verteilungen in unserer Welt eine Art zentrale Tendenz haben, d. h. die meisten Beobachtungen liegen wahrscheinlich nahe am Zentrum einer Verteilung und sind keine Ausreißer. Üblicherweise wird die Form einer Verteilung in einem Histogramm dargestellt. Das Beispiel unten zeigt die Tendenz für einen typischen Messwert.



Bevor wir uns näher mit Mittelwerten befassen, benötigen wir einige statistische Fachbegriffe. Bitte schauen Sie sich dazu die folgende Liste an (Byjus Classes, 2021). Bitte beachten Sie, dass die Liste nicht alle existierenden Bezeichnungen aus der Statistik enthält, aber sie liefert alle für diesen Kurs relevanten Fachbegriffe.

|  |  |  |  |
| --- | --- | --- | --- |
| *P*(*A*) = | | Wahrscheinlichkeit, dass Ereignis A eintritt | |
| P(A | B) = | | Bedingte Wahrscheinlichkeit | |
| P(A ∪ B) = | | Wahrscheinlichkeit, das Ereignis A oder B eintritt | |
| µ*=* | | Arithmetischer Mittelwert der Grundgesamtheit | |
| var(X) = | | Varianz | |
| E(X | Y) = | | Bedingter Erwartungswert | |
| std(X) = | | Standardabweichung | |
| σ*2 =* | | Varianz der Grundgesamtheit | |
| x˜ = | | Median | |
| σ*X =* | | Standardabweichung der Grundgesamtheit | |
| corr(X,Y) = | | Korrelation | |
| cov(X,Y) = | | Kovarianz | |
| ρX,Y *=* | | Korrelation | |
| Mo = | | Modus | |
| Md = | | Median der Stichprobe | |
| MR *=* | | Quartilsabstand | |
| Q2 = | | Zweites Quartil | |
| Q1 = | | Erstes Quartil | |
| x = | | Arithmetischer Mittelwert der Stichprobe | |
| Q3 = | | Drittes Quartil | |
| s = | | Standardabweichung der Stichprobe | |
| s2 = | | Varianz der Stichprobe | |
| X ~ = | | Verteilung von X | |
| zx = | | Z-Wert | |

### Arithmetischer Mittelwert

Der **arithmetische Mittelwert** ist der bekannteste und vielseitigste Mittelwert. Es gibt Formeln sowohl für den arithmetischen Mittelwert der Grundgesamtheit, als auch für die Stichprobe. Da wir uns hauptsächlich mit der Berechnung des Arithmetischen Mittelwerts einer Stichprobe beschäftigen, verwenden wir Gleichung (1.1)(Ahmad, 2016; Doane, 2016):

**Arithmetischer Mittelwert**

Wenn Daten normalverteilt sind, sollte immer der arithmetische Mittelwert als Kenngröße für die zentrale Tendenz ausgewählt werden, weil er später in leistungsfähigeren statistischen Verfahren verwendet werden kann.

(1.1)

Wir können den arithmetischen Mittelwert auch mit der Excel-Funktion =MITTELWERT(Datenvariable) berechnen, wobei sich „Datenvariable“ auf die Spalte bezieht, die die tatsächlichen Werte enthält. Der arithmetische Mittelwert lässt sich nur für Variablen bestimmen, die auf einer Intervall- oder Verhältnisskala gemessen werden (Microsoft, n.d.). Er wird auch als "Durchschnitt" bezeichnet und wäre der Drehpunkt, wenn man alle Variablen auf einer Balkenwaage abwägen würde. Dies ist deshalb so, weil die Summe der Abstände zu den einzelnen Datenpunkten in einer Verteilung immer gleich Null ist. Es ist zu beachten, dass der arithmetische Mittelwert im Allgemeinen weniger genau einem typischen Wert in einem Datensatz entspricht, wenn es Ausreißer oder extreme Werte in den Daten gibt. Daher sollte zusätzlich zum arithmetischen Mittelwert auch die Standardabweichung angegeben werden.

**Standardabweichung**

Die Standardabweichung misst die Streuung der Beobachtungen um den arithmetischen Mittelwert.

### Median

Der Median entspricht dem fünfzigsten Perzentil, dem Mittelpunkt oder dem mittleren Wert der betrachteten Datenvariablen. Wenn die Daten eine ungerade Anzahl von Beobachtungen aufweisen, wird der Median berechnet, indem der Durchschnitt der beiden mittleren Beobachtungen genommen wird. Bei Extremwerten im Datensatz sollte anstelle des arithmetischen Mittelwerts der Median verwendet werden. Es ist jedoch anzumerken, dass dem Median einige der nützlichen mathematischen Eigenschaften des arithmetischen Mittelwerts fehlen, was ihn weniger vielseitig macht (Doane, 2016).

### Modus

Der Modus ist der am häufigsten auftretende Wert für eine Variable. Wenn die Daten zentral verteilt sind, können arithmetischer Mittelwert, Median und Modus gleich sein. Es kann jedoch auch sein, das es in einem Datensatz gar keinen Modus gibt. Stattdessen könnte der Datensatz mehrere häufige Variablen haben, wodurch er bimodal wird. Um den Modus zu ermitteln, muss man die Häufigkeit der einzelnen Werte einer Variablen ermitteln (z. B. mit der Excel-Funktion = MODE.EINF(Datenvariable)) (Microsoft, n.d.). Der Modus wird häufig zur Beschreibung der zentralen Tendenz bei kategorialen oder nominalen Variablen verwendet (Doane, 2016). Es gibt weitere Mittelwerte, doch arithmetischer Mittelwert, Median und Modus werden am häufigsten benutzt. Bitte beachten Sie, dass sich arithmetischer Mittelwert, Median und Modus sehr ähnlich sein können, wenn die Beobachtungen sich symmetrisch um einen zentralen Wert gruppieren (Normalverteilung).

### Streuungsmaße

Die Form einer Verteilung kann mit Hilfe eines Histogramms dargestellt werden. Eine Veranschaulichung ist aber auch möglich durch den Vergleich von arithmetischem Mittelwert und Median. Wenn die Daten rechtsschief sind, bedeutet dies, dass der arithmetische Mittelwert den Median übersteigt. Im Umkehrfall, wenn die Daten linksschief sind, bedeutet dies, dass der arithmetische Mittelwert geringer ist, als der Median. Wenn der arithmetische Mittelwert, der Median und der Modus ähnlich sind, wird die Verteilung als symmetrisch angesehen. Wenn eine Verteilung nicht symmetrisch ist, wird sie als schief bezeichnet.

Streuungsmaße werden verwendet, um zu messen, wie sich die Beobachtungen um den arithmetischen Mittelwert gruppieren. Dieses Maß wird als Abweichung der Daten bezeichnet. Ein einfaches Maß für die Streuung ist die Spannweite, die sich aus dem höchsten Wert der Daten abzüglich des niedrigsten Wertes ergibt und mit Gleichung (1.2) berechnet wird (Doane, 2016):

Spannweite = größter Datenwert - kleinster Datenwert

(1.2)

In Excel kann die Spannweite mit der Funktion = MAX(Datenvariable)- MIN(Datenvariable) berechnet werden (Microsoft, n.d.). Das beliebteste und vielseitigste Maß für die Streuung ist die Standardabweichung (Ahmad, 2016). Die Standardabweichung ist ein Maß für den durchschnittlichen Abstand der Datenvariablen vom arithmetischen Mittelwert und wird nach Gleichung (1.3) berechnet:

(1.3)

In Excel kann die Standardabweichung mit der Funktion = STABW.S(Datenvariable) berechnet werden (Microsoft, n.d.). In der Regel wird neben dem Mittelwert auch die Standardabweichung angegeben, um die Aussagekraft des Mittelwerts bei der Beschreibung der Daten zu dokumentieren. Der arithmetische Mittelwert ist als Kenngröße nicht besonders nützlich, wenn die Variable eine große Standardabweichung aufweist (Doane, 2016).

### Weitere Kenngrößen der deskriptiven Statistik

**Boxplot**

Die aus einer Box und Linien bestehende Darstellung repräsentiert die Quartile einer Verteilung. Ausreißer werden üblicherweise als Sternchen in der äußeren Region dargestellt.

Während Zentrumsmaße und Streuungsmaße die wichtigsten allgemeinen deskriptiven Kenngrößen sind, die von Forschern verwendet werden, umfassen andere relevante Kenngrößen die Häufigkeit (Prozent, Anzahl, Häufigkeit) und Positionsmaße (Quartile und Perzentile). Ein **Boxplot** wird häufig verwendet, um den Median, die Quartile und sogar Ausreißer eines Datensatzes darzustellen.

## 1.2 Induktive Statistik

**Induktive Statistik**

Der Prozess der Anwendung induktiver Statistik wird manchmal auch als Praxis der statistischen Schlussfolgerung bezeichnet. Statistische Schlussfolgerungen hängen stark von der Stichprobentheorie und der zentralen Tendenz ab.

Bei der **induktiven Statistik** geht es in erster Linie darum, von einer gegebenen Stichprobe auf Informationen über die Grundgesamtheit zu schließen. In vielen Fällen ist es entweder unpraktisch oder unmöglich, die gewünschten Informationen von der Grundgesamtheit zu ermitteln. Ein praktisches Beispiel sind politische Umfragen, bei denen nicht die Gesamtbevölkerung befragt wird, sondern lediglich eine Stichprobe von mehreren tausend Menschen. Diese Umfrage liefert eine bis auf wenige Prozentpunkte genaue Einschätzung – vorausgesetzt, die Teilnehmer:innen wurden unter Berücksichtigung geeigneter Verfahren aus der Wahrscheinlichkeitsrechnung ausgewählt. Im Gegensatz zu einer Befragung der Gesamtbevölkerung bietet dies eine erhebliche Kosteneinsparung. Wenn es darum ginge, zu zählen, wie viele Lachse derzeit im Columbia River leben, wäre dies eine unmögliche Aufgabe. Mit der richtigen Planung und Durchführung könnte dies jedoch recht einfach mit Hilfe von Stichproben durchgeführt werden. In einigen Fällen, wie z. B. bei einer Volkszählung der gesamten Bevölkerung eines Landes, die oft gesetzlich vorgeschrieben ist, kann eine Umfrage sogar noch genauer sein. Der Versuch, die Informationen von allen in einem Land lebenden Personen zu erhalten, ist eine Herausforderung, auch wenn die entsprechenden Ressourcen für die Durchführung der Volkszählung vorhanden sind. Zu den Arten von Analysen, die mit Hilfe der induktiven Statistik durchgeführt werden können, gehören die Schätzung von Kenngrößen, das Testen von Hypothesen sowie Regressions- und Trendanalysen. Es gibt auch viele Aspekte der Qualitätskontrolle, bei denen die induktive Statistik zum Einsatz kommt.

Für die Anwendung der induktiven Statistik ist es wichtig, die Daten für Stichprobe aus einer Grundgesamtheit in geeigneter Weise zu entnehmen. Dadurch wird sichergestellt, dass die geschlussfolgerten Kenngrößen verzerrungsfreie Näherungswerte für die Grundgesamtheit darstellen. Zu den Zufallsstichprobenverfahren gehören die einfache Zufallsauswahl, bei der jeder in einer Grundgesamtheit die gleiche Chance hat, ausgewählt zu werden. Die einfache Zufallsstichprobe ist eine Form der Wahrscheinlichkeitsstichprobe. Weitere Arten der Wahrscheinlichkeitsstichprobe sind die systematische und die geschichtete Zufallsstichprobe sowie die Zufallsziehung nach dem Klumpverfahren. Da alle diese Methoden allen Elementen der Grundgesamtheit eine Chance geben, ausgewählt zu werden, gelten sie als die wünschenswertesten Stichprobenmethoden zur Schätzung einer Grundgesamtheit mit Hilfe der induktiven Statistik (Knapp, 1980). Andere Stichprobenverfahren, bei denen nicht alle Mitglieder oder Elemente einer Grundgesamtheit die Chance haben, ausgewählt zu werden, sind z. B. willkürliche, Freiwilligen- und Schneeballstichproben. Methoden wie die willkürliche Stichprobe, deren Elemente auf der Grundlage zweckmäßig ausgewählter Teilbereiche einer Grundgesamtheit ausgewählt werden, sind in der Praxis zwar sehr beliebt, können aber zu erheblichen Abweichungen führen (Doane, 2016; Knapp, 1980).

**Wahrscheinlichkeitsstichprobe**

Es ist erwähnenswert, dass Forschende bei Wahrscheinlichkeitsstichproben Kalibrierungstechniken anwenden müssen, um sicherzustellen, dass die gewonnenen Erkenntnisse die Proportionen der Grundgesamtheit widerspiegeln.

Nach der Wahl eines Stichprobenverfahrens besteht der nächste Schritt darin, die Daten der Stichprobe zu erheben. Eine der wichtigsten Methoden zur Erfassung von Daten aus einer Stichprobe ist die Verwendung eines Fragebogens. Fragebögen können individuell entwickelt werden, um Daten zu praktisch jedem Forschungsthema zu erheben. Die Forschenden müssen jedoch zusätzliche Vorsichtsmaßnahmen ergreifen, um sicherzustellen, dass die aus dem Fragebogen gewonnenen Erkenntnisse sowohl gültig als auch zuverlässig in Bezug auf das interessierende Thema sind.

### Schätzung eines Parameters der Grundgesamtheit aus einer statistischen Kenngröße

Sobald wir unsere Stichprobe haben, geht es um die Schätzung von Parametern der Grundgesamtheit anhand der abgeleiteten Kenngrößen. Eine statistische Kenngröße ist ein Maß, das aus einer Stichprobe abgeleitet wird, und ein Parameter ist ein Maß, das die Grundgesamtheit beschreibt, aus der die Stichprobe ausgewählt wurde. Alle Stichproben sind mit Zufallsfehlern behaftet, daher wird die Differenz zwischen einer statistischen Kenngröße und einem Parameter als Stichprobenfehler bezeichnet. Da wir in der Regel keine Möglichkeit haben, die Parameter der Grundgesamtheit zu kennen, können wir mit Hilfe der induktiven Statistik die Parameter anhand einer hypothetischen Verteilung schätzen, die den Stichprobenfehler berücksichtigt.

Eine Punktschätzung ist eine Einzelwertschätzung eines Parameters (ein arithmetischer Mittelwert ist ein Beispiel für eine Punktschätzung). Eine Intervallschätzung ist ein Bereich von Werten, in dem sich der Parameter höchstwahrscheinlich befindet. Punktschätzungen sind zwar leicht zu verstehen, geben aber kaum Aufschluss über die Streuung des Parameters. Ein Konfidenzintervall ist eine der beliebtesten Intervallschätzungen, die eine Lösung für dieses Problem bietet. Jedes Konfidenzintervall enthält ein Konfidenzniveau, das aus einer theoretischen Verteilung abgeleitet ist, welche die Wahrscheinlichkeit beschreibt, dass das Intervall den geschätzten Parameter unter gleichen Bedingungen wiederholten Stichproben enthält.

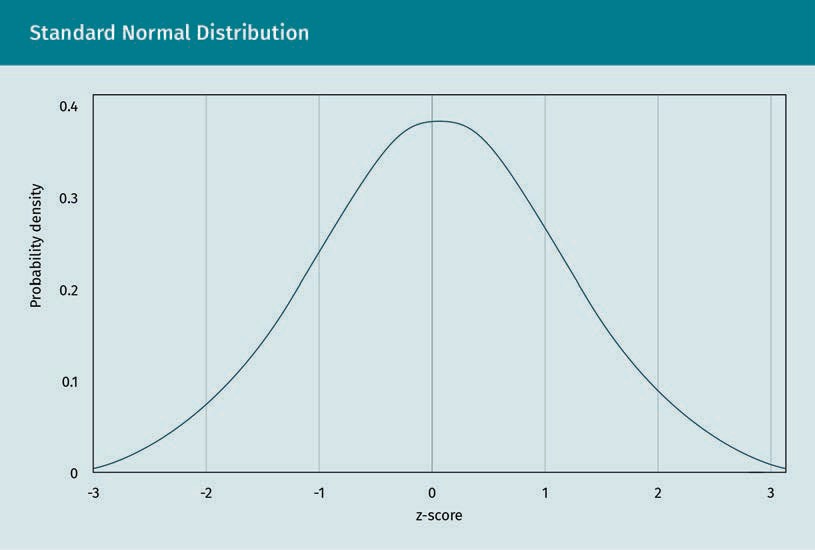
Ein 95-prozentiges Konfidenzintervall bedeutet, dass das Intervall bei einer unbestimmten Anzahl von Stichproben den wahren Parameter der Grundgesamtheit in 95 von 100 Fällen enthält. Wie ein Konfidenzintervall berechnet wird, zeigt Gleichung (1.4.). Weitere Bereiche der induktiven Statistik sind Hypothesentests, Korrelationsanalysen und Regressionsanalysen.

(1.4)

|  |  |
| --- | --- |
| CI = | Konfidenzintervall |
| x = | Arithmetischer Mittelwert der Stichprobe |
| z = | z-Wert |
| s = | Standardabweichung der Stichprobe |
| n = | Stichprobengröße |

wobei:

Ein z-Wert oder z-Score beschreibt die Position eines Probenwertes in Bezug auf seinen Abstand vom arithmetischen Mittelwert, gemessen als Faktor der Standardabweichung in einer Normalverteilung (Doane, 2016; Knapp, 1980). Die Abbildung unten stellt die Standard- oder Normalverteilung dar.



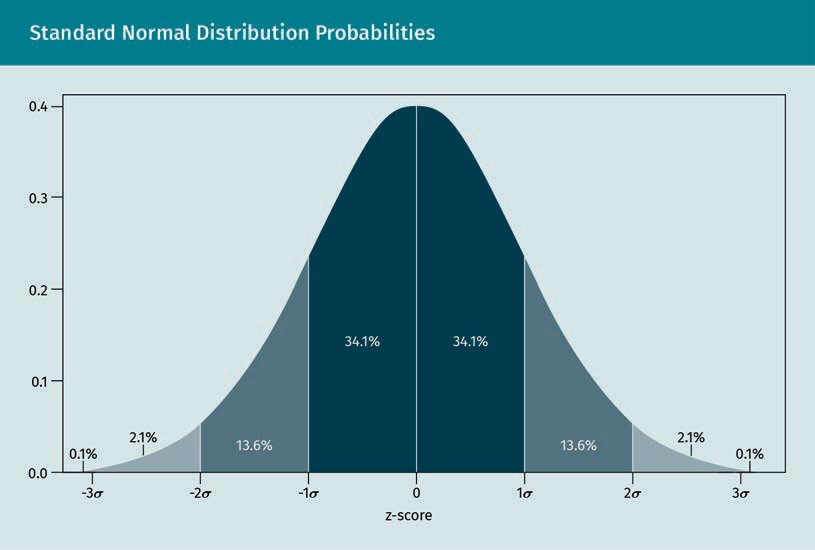
Der z-Wert ist positiv für alle Werte, die größer sind als der arithmetische Mittelwert und negativ für alle Werte, die kleiner sind als der arithmetische Mittelwert. Die Standardnormalverteilung in der Abbildung ist eine glockenförmige Verteilung, die einen arithmetischen Mittelwert von 0 und eine Standardabweichung von 1 hat. Die Formel zur Berechnung eines z-Wertes lautet z = (x-μ)/σ, wobei x ein Ergebniswert, μ der arithmetische Mittelwert der Grundgesamtheit, und σ die Standardabweichung der Grundgesamtheit ist. Da wir den Mittelwert und die Standardabweichung der Grundgesamtheit nur selten kennen, können wir natürlich den Mittelwert (x̄) und die Standardabweichung der Stichprobe (s) als unverzerrte Schätzwerte für die Parameter der Grundgesamtheit in die Formel einsetzen (Doane, 2016).

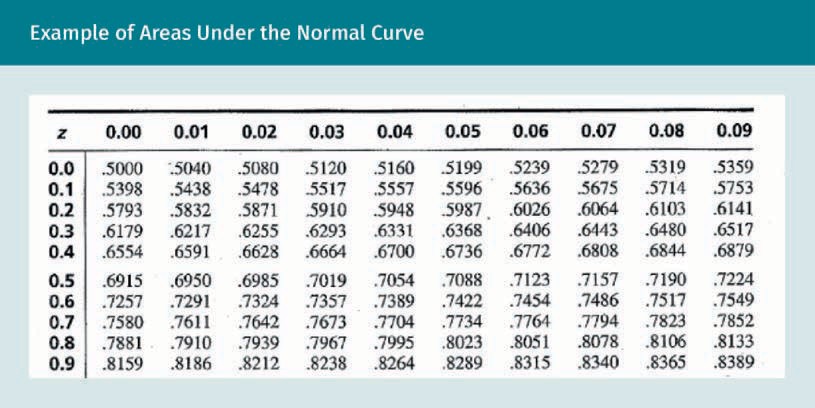
### Z-Standardisierung

Die Formel zur Berechnung des z-Wertes wird im Folgenden dargestellt. Dabei ist x ein Ergebniswert, μ der arithmetische Mittelwert und σ die Standardabweichung der Grundgesamtheit.

(1.5)

Sobald Sie einen z-Score berechnet haben, kann dieser in einer z-Tabelle nachgeschlagen werden, von der ein Teil in Anlehnung an die nachstehende Abbildung der Standardnormalverteilung dargestellt ist. Die vollständige Tabelle ist in der Literatur von Farber und Larson (2017) zu finden. Die z-Tabelle enthält die Flächeninhalte unter der Standardnormalverteilung, die es den Forschenden ermöglichen, die Wahrscheinlichkeit eines zufälligen Ergebniswertes aus der Stichprobenverteilung zu berechnen. Beispielsweise ist aus der folgenden Abbildung zu erkennen, dass eine 95,4-prozentige Wahrscheinlichkeit besteht, zufällig eine Punktzahl auszuwählen, die zwischen -2 und +2 Standardabweichungen vom Mittelwert liegt (Doane, 2016; Knapp, 1980).





## 1.3 Wahrscheinlichkeitsrechnung

Archäologische Ausgrabungen haben ergeben, dass antike Gesellschaften eine Art von Wahrscheinlichkeitsspielen mit primitiven würfelähnlichen Objekten aus Tierknochen spielten (Britannica, n.d.). Die ernsthafte Beschäftigung mit der Wahrscheinlichkeitsrechnung begann jedoch erst, als die Menschen das organisierte Glücksspiel für sich entdeckt hatten. Es sind noch viele Fragen offen, was Zufall wirklich bedeutet. In der klassischen Physik würde man sagen, dass echter Zufall existiert, weil wir unbekannte Eingaben haben. Die Quantenphysik, die unser heutiges Leben in Bezug auf Kommunikation und Datenverarbeitung beherrscht, zeichnet jedoch das Bild, dass es tatsächlich zufällige Phänomene gibt, solange wir die Welt aus einer subatomaren Perspektive betrachten (Fogarty, 2017). Was also denken wir im Alltag über Zufall, bzw. Wahrscheinlichkeit? Lassen Sie uns zunächst zwei Methoden zur Berechnung von Wahrscheinlichkeiten untersuchen. Die erste ist die klassische Wahrscheinlichkeitsrechnung. Bei klassischen Wahrscheinlichkeiten hat jedes Ereignis die gleiche Chance, einzutreten. Zum Beispiel sollte mit einem idealen Würfel die gleiche Wahrscheinlichkeit bestehen, eine 1, 2, 3, 4, 5 oder 6 zu würfeln. Das Gleiche gilt für das Werfen einer Münze, bei dem die Wahrscheinlichkeit, Kopf oder Zahl zu erhalten gleich groß ist, und auch die Wahrscheinlichkeit aus einem Standard-Skatspiel eine bestimmte Karte zu ziehen, beträgt für alle Karten 1/52 (Di Paola et al., 2018; Doane, 2016).

### Formeln der klassischen Wahrscheinlichkeitsrechnung

Die klassische Wahrscheinlichkeit, auch als theoretische Wahrscheinlichkeit bezeichnet, ist die erwartete Wahrscheinlichkeit, dass ein einfaches Ereignis eintritt (z. B. beim Würfeln). Dies ist die Anzahl der Ereignisse (eins, bei einem Würfelwurf) geteilt durch die Anzahl der möglichen Ergebnisse (sechs, für jede Seite eins). P(A) bedeutet „Wahrscheinlichkeit von Ereignis A“, wobei A das einzuschätzende Ereignis ist, z. B. ein Würfelwurf mit einer 6. f ist die Häufigkeit, d. h. die Anzahl der Ereignisse, und N ist die Anzahl der möglichen Ergebnisse.

Die Wahrscheinlichkeit, bei einem Wurf mit einem idealen Würfel eine Drei zu erhalten, ist beispielsweise eins zu sechs, also 1/6. Das ist ein mögliches Ergebnis geteilt durch die Anzahl der möglichen Ergebnisse (in diesem Fall sechs). Die Gewinnchancen bei der Powerball-Lotterie liegen bei 1/292,2 Millionen (Amadeo, 2021). Die "1" ist die Anzahl der möglichen Ereignisse (d.h. Sie gewinnen im Lotto), geteilt durch die Anzahl der möglichen Zahlenkombinationen (etwa 292.000.000) auf den verkauften Scheinen. Wenn Sie die Mega-Millions berechnen wollen, wäre dies 1/302,5 Millionen (Amadeo, 2021).

### Empirische Wahrscheinlichkeit

Die **empirische oder experimentelle Wahrscheinlichkeit** unterscheidet sich von der klassischen Wahrscheinlichkeit dadurch, dass sie nicht von einer hypothetischen Wahrscheinlichkeit ausgeht, die auf der Anzahl der Versuche und möglichen Ergebnisse beruht, sondern versucht, die tatsächliche Wahrscheinlichkeit des Eintretens von Ereignissen auf Basis eines Experiment zu berechnen (Doane, 2016). Ein interessantes Beispiel hierfür ist ein US-amerikanischer Statistiker, der während des Vietnamkonﬂikts in Kriegsgefangenschaft war und die Ergebnisse von Hunderttausenden von wiederholten Münzauszählungen während seiner Zeit in der Gefängniszelle aufzeichnete. Seine Ergebnisse zeigten, dass die Ergebnisse leicht in Richtung Kopf tendierten. Nach der klassischen Wahrscheinlichkeitsrechnung stünden die Chancen 50:50, aber empirische Wahrscheinlichkeitsberechnungen zeigen, dass es besser wäre, immer Kopf zu wählen (Di Paola et al., 2018; Doane, 2016). Mathematisch gesehen lautet die Formel für die empirische Wahrscheinlichkeit:

**Empirische Wahrscheinlichkeit**

Manchmal können sogar die am meisten akzeptierten klassischen Wahrscheinlichkeiten (z. B. die Wahrscheinlichkeit eines Münzwurfs) durch empirische Wahrscheinlichkeiten in Frage gestellt werden.

Empirische Wahrscheinlichkeit = (Häufigkeit des Auftretens eines Ereignisses / Gesamtzahl der Versuche)

(1.6)

Die experimentelle oder empirische Wahrscheinlichkeit eines Ereignisses basiert auf dem, was in einem bestimmten Experiment tatsächlich passiert ist, während die klassische oder theoretische Wahrscheinlichkeitsrechnung versucht, auf der Grundlage der Gesamtzahl der möglichen Ergebnisse vorherzusagen, was passieren wird. Wir erwarten, dass die experimentellen und theoretischen Wahrscheinlichkeiten konvergieren, wenn die Anzahl der Versuche in einem Experiment steigt (Di Paola et al., 2018; Doane, 2016).

### Subjektive Wahrscheinlichkeit

Die **subjektive Wahrscheinlichkeit** wird angewendet, wenn es kein wiederholbares Zufallsexperiment gibt. Bei der subjektiven Wahrscheinlichkeit wird ein begründetes Urteil zur Vorhersage der Wahrscheinlichkeit eines Ereignisses verwendet. In unserem täglichen Leben verwenden wir subjektive Wahrscheinlichkeiten. Beispiele aus dem wirklichen Leben, in denen Menschen subjektive Wahrscheinlichkeiten verwenden können:

**Subjektive Wahrscheinlichkeit**

Der menschliche Verstand hat sich so entwickelt, dass er uns hilft, subjektive Wahrscheinlichkeiten zu berechnen. Darüber hinaus sollten viele Geschäftsentscheidungen, die auf der Grundlage subjektiver Wahrscheinlichkeiten getroffen werden, anhand empirischer Wahrscheinlichkeiten validiert werden..

* das Ergebnis eines Vorstellungsgesprächs voraussagen
* ein Kaufangebot für ein Haus abgeben
* das Risiko bei Nichtanlegen des Sicherheitsgurts einschätzen

### Die sieben Grundregeln der Wahrscheinlichkeit

Die Grundregeln der Wahrscheinlichkeit sind sowohl logisch als auch intuitiv. Manchmal werden Wahrscheinlichkeitsprobleme kompliziert, aber auch dann können sie gelöst werden, wenn diese Regeln konsequent angewendet werden. Für die Anwendung der meisten dieser Regeln sind sowohl logische als auch mathematische Fähigkeiten erforderlich (Di Paola et al., 2018; Doane, 2016).

### Einfache Wahrscheinlichkeit (Regeln 1 – 4)

#### Wahrscheinlichkeitsregel 1

Die erste Grundregel der Wahrscheinlichkeitsrechnung besagt, dass alle Wahrscheinlichkeiten durch eine Zahl zwischen Null und Eins dargestellt werden. Die Wahrscheinlichkeitsberechnung ist dann falsch, wenn das Ergebnis kleiner als null oder größer als eins ist (Di Paola et al., 2018; Doane, 2016):

Für ein Ereignis A gilt: 0 ≤ P(A) ≤ 1.

(1.7)

#### Wahrscheinlichkeitsregel 2

Die zweite Grundregel der Wahrscheinlichkeit besagt, dass die Summe der Ergebnisse aller Wahrscheinlichkeiten gleich eins sein muss (Di Paola et al., 2018; Doane, 2016):

Summe aller Wahrscheinlichkeiten = 1

(1.8)

#### Wahrscheinlichkeitsregel 3

Die dritte Regel, auch Komplementärregel genannt, besagt, dass die Wahrscheinlichkeit, dass ein Ereignis nicht eintritt, durch eins minus der Wahrscheinlichkeit, dass das Ereignis eintritt, gegeben ist (Di Paola et al., 2018; Doane, 2016):

Summe aller Wahrscheinlichkeiten = 1

*P*(*A*¯¯¯)=1−*P*(*A*)??

(1.9)

#### Wahrscheinlichkeitsregel 4

Die vierte Grundregel der Wahrscheinlichkeitsrechnung ist auch als Additionsregel bekannt. Diese Regel besagt, dass die Wahrscheinlichkeit von zwei unabhängigen Ereignissen die Summe der Wahrscheinlichkeiten für jedes der beiden Ereignisse ist (Di Paola et al., 2018; Doane, 2016). Wenn A und B voneinander unabhängige Ereignisse sind, dann gilt:

P(A or B) = P(A) + P(B)

(1.10)

#### Wahrscheinlichkeitsregel 5

Die fünfte Grundregel der Wahrscheinlichkeitsrechnung ist auch als allgemeine Additionsregel bekannt. Diese Regel besagt, dass die Wahrscheinlichkeit, dass eines von zwei Ereignissen eintritt, berechnet werden kann, indem von der Summe der Wahrscheinlichkeiten, dass jedes Ereignis eintritt, die Wahrscheinlichkeit abgezogen wird dass beide eintreten. (Di Paola et al., 2018; Doane, 2016):

P(A or B) = P(A) + P(B) – P(A and B)

(1.11)

Beim Lösen von Wahrscheinlichkeitsproblemen sollte immer logisch und unter systematischer Anwendung der obigen Regeln vorgegangen werden.

### Weiterführende bedingte Wahrscheinlichkeiten und unabhängige Ereignisse (Regeln 6 – 7)

Jetzt werden wir zwei fortgeschrittenere Regeln für die Wahrscheinlichkeit behandeln, die als Multiplikationsregeln für die Bestimmung von P(A und B) bekannt sind. Bevor diese Regeln dargelegt werden, müssen wir uns zunächst mit zwei wichtigen Begriffen befassen: mit unabhängigen Ereignisse und bedingten Wahrscheinlichkeiten. Zwei Ereignisse, A und B, werden als unabhängig betrachtet, wenn die Tatsache, dass ein Ereignis eingetreten ist, keinen Einfluss auf die Wahrscheinlichkeit hat, dass das andere Ereignis eintritt. Wenn jedoch das Eintreten eines Ereignisses die Wahrscheinlichkeit des Eintretens des anderen Ereignisses beeinflusst, dann sind die beiden Ereignisse abhängig (Di Paola et al., 2018; Doane, 2016).

#### Wahrscheinlichkeitsregel 6

Die sechste Grundregel der Wahrscheinlichkeitsrechnung ist auch als Multiplikationsregel für unabhängige Ereignisse bekannt. Diese Regel besagt, dass, wenn A und B beide unabhängige Ereignisse sind, die Wahrscheinlichkeit, dass sowohl A als auch B eintreten, das Produkt aus der Wahrscheinlichkeit für A und der Wahrscheinlichkeit für B ist (Di Paola et al., 2018; Doane, 2016). Wenn A und B voneinander unabhängige Ereignisse sind, dann gilt:

P(A and B) = P(A) \* P(B)

(1.12)

#### Wahrscheinlichkeitsregel 7

Die siebte und letzte Grundregel der Wahrscheinlichkeit beruht auf der Tatsache, dass die Wahrscheinlichkeiten bestimmter Ereignisse davon abhängen, ob andere Ereignisse eingetreten sind oder nicht. Diese Regel ist als bedingte Wahrscheinlichkeitsregel bekannt. Sie besagt, dass die bedingte Wahrscheinlichkeit von Ereignis B, falls Ereignis A eingetreten ist, die Wahrscheinlichkeit des Eintretens beider Ereignisse A und B geteilt durch die Wahrscheinlichkeit des Eintretens von Ereignis A ist (Di Paola et al., 2018; Doane, 2016). Die bedingte Wahrscheinlichkeit für Ereignis B, vorausgesetzt A ist eingetreten, lautet:

P(B | A) = P(A and B) / P(A)

(1.13)

### Zusammenfassung

In dieser Einheit haben Sie einige der grundlegenden statistischen Prinzipien kennengelernt, darunter deskriptive Statistik, induktive Statistik und grundlegende Wahrscheinlichkeitsanalyse. Kenngrößen der deskriptiven Statistik sind wichtige Instrumente, die in jeder Datenanalyse enthalten sein sollten. Sie ermöglichen den Forschenden, einen Datensatz zu bewerten und helfen ihnen bei der Auswahl der am besten geeignetsten, fortschrittlicheren statistischen Technik für zukünftige Analysen. Die induktive Statistik umfasst leistungsfähige Instrumente, die es ermöglichen, anhand einer kleinen Stichprobe auf die Parameter einer Grundgesamtheit zu schließen. Dadurch erhalten wir die bedeutende Fähigkeit, die Welt auf effiziente und nachhaltige Weise zu verstehen, und ein wichtiges Handwerkszeug für das Studium und die kontinuierliche Anwendung in der Geschäftswelt.

Es gibt sieben Grundregeln der Wahrscheinlichkeit. Beim Lösen von Wahrscheinlichkeitsproblemen sollte immer logisch und unter systematischer Anwendung dieser Grundregeln vorgegangen werden, denn manchmal sucht unser Verstand nach einer einfachen Lösung und kann leicht einige der Einzelheiten eines Wahrscheinlichkeitsproblems übersehen. Bei den Regeln handelt es sich um grundlegende Prinzipien, die als Baustein für weiterführende Studien zu multivariaten statistischen Instrumenten und Konzepten verwendet werden können.

# Lektion 2 - Bivariate Analyse

### Lernziele

Nach der Bearbeitung dieser Lektion werden Sie in der Lage sein, ...

... mit zwei Methoden auf statistische Signifikanz zu testen und auszuwerten.

... Kreuztabellen zu erstellen und zu interpretieren.

... Korrelationskoeffizienten zu berechnen und zu interpretieren.

# 2. Bivariate Analyse

## Aus der Praxis

Die Lebensmittelkette Lesko möchte wissen, ob ihre Gutscheinkarten, die einen Rabatt von fünf Prozent auf ausgewählte Artikel im Geschäft bieten, die Verkaufszahlen steigern. Die Wachstumschefin von Lesko, Catherine Vauxhall, fragt sich, ob das Unternehmen durch das Angebot dieser Gutscheine an die Kunden zu viel von seinem Gewinn aufgibt, da das Geschäft in letzter Zeit einen Umsatzanstieg verzeichnet hat. Diese Gutscheine wurden nur den Abonnent:innen des Lesko-Treueprogramms angeboten. Catherine befürchtet, dass diese die falsche Zielgruppe sind, da sie wahrscheinlich ohnehin zum Einkaufen in den Laden kommen. Da alle derzeitigen Abonnenten von Lesko diese Gutscheine erhalten, überlegt Catherine, wie sie diesen Test am besten durchführen kann. Sollte eine Kontrollgruppe eingerichtet werden, die keine Gutscheinpost erhält? Welche Größe sollte die Stichprobe haben? Welchen statistischen Test und welche Software könnte sie benutzen, um ihre Hypothese nachzuweisen? Das sind die Fragen, von denen Catherine weiß, dass sie sie beantworten muss, bevor sie die gewünschten Informationen erhält.

Diese Lektion ist eine Einführung in die bivariate Analyse, d. h. in eine Reihe von statistischen Analyseverfahren für zwei Variablen. Dies steht im Gegensatz zur univariaten (für eine einzige Variable) und multivariaten (für drei oder mehr Variablen) Statistik. Die bivariate Analyse konzentriert sich auf Situationen mit einer abhängigen und einer unabhängigen Variable. Die Bezeichnungen dieser Variablen sind X und Y. Mit bivariaten Analysen untersuchen wir, ob und in welchem Ausmaß eine der Variablen die andere beeinflusst.

## 2.1 Kreuztabellen

**Kreuztabellen** sind für Unternehmen, die mit verpackten Waren handeln, unerlässlich. Diese Unternehmen beauftragen **Marktforschung**sunternehmen mit der Durchführung von Feldforschung und umfangreichen Umfragen unter Verbraucher:innen und aktuellen Kund:innen, wobei sie Kreuztabellen verwenden, um ihre Kundschaft zu gruppieren und Erkenntnisse über die Verwendung ihrer Produkte zu gewinnen. Ein Beispiel ist die Frage nach der Verwendung des Produkts in den letzten 30 Tagen, aufgeteilt nach verschiedenen Altersgruppen. Diese beiden Variablen können in einer Tabelle angeordnet werden, damit die Forschenden feststellen können, ob bestimmte Altersgruppen das Produkt häufiger nutzen (Doane, 2016).

**Marktforschung**

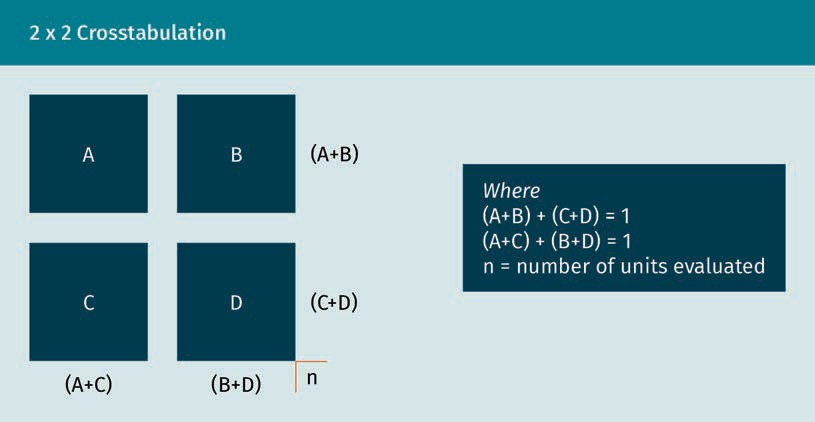
Einige schnell agierende Unternehmen für Verbrauchsgüter beauftragen Marktforschungsunternehmen, um ausgefeilte Techniken für Kreuztabellen zu entwickeln, die abhängig vom Stichprobenumfang viele Ebenen umfassen können.

**Kreuztabellen**

Die statistische Signifikanz einer Kreuztabelle wird in der Regel mit nicht-paramentrischen statistischen Verfahren gemessen, darunter die Chi-Quadrat-Analyse..

Eine Kreuztabelle für zwei Variablen wird auch als zweidimensionale Kontingenztabelle bezeichnet und hat ein rechteckiges Format. Die Zeilen (R) enthalten die Werte der X-Variablen, die Spalten (C) die der Y-Variablen. Für jede mögliche Kombination von X- und Y-Werten gibt es eine Zelle. In diese wird die absolute Häufigkeit, relative Häufigkeit oder die Wahrscheinlichkeit des gemeinsamen Auftretens der jeweiligen Kombination eingetragen. Eine R x C-Tabelle bezieht sich auf eine Kontingenztabelle mit R Zeilen und C Spalten (Doane, 2016).

Eine Variable mit nur zwei Kategorien wird als dichotome oder binäre Variable bezeichnet. Wenn beide Variablen binär sind, führt dies zu einer 2 x 2-Tabelle (Doane, 2016). Die Abbildung unten zeigt ein Beispiel dafür.



Eine Kreuztabelle enthält die drei wichtigsten Wahrscheinlichkeitsverteilungen, mit denen wir uns befassen: gemeinsame Verteilung, Randverteilung und bedingte Verteilung (Doane, 2016). Die gemeinsame Wahrscheinlichkeitsverteilung in einer Kreuztabelle beschreibt den Anteil einer bestimmten Kombination aus X- und Y-Werten. Dividiert man die Zelleninhalte der Kreuztabelle durch die Gesamtzahl, erhält man die gemeinsame Verteilung. Die Summe einer gemeinsamen Verteilung ist = 1, wie in der obigen Abbildung angegeben (Doane, 2016).

Die Randverteilung wird als die Verteilung der Variablen X (Zeile) oder Y (Spalte) allein beschrieben. Die Randverteilungen ergeben sich aus den Zeilen- und Spaltensummen der Kreuztabelle. Die Summe einer Randverteilung ist = 1, wie in der obigen Abbildung angegeben (Doane, 2016).

Die Verteilung der einen Variablen in Abhängigkeit von den Werten der anderen Variablen wird durch bedingte Verteilungen beschrieben. Die Zellinhalte der Kreuztabelle geteilt durch die Zeilen- oder Spaltensummen ergeben die bedingten Wahrscheinlichkeitsverteilungen. Die Summe einer bedingten Verteilung ist ebenfalls = 1, wie in der obigen Abbildung dargestellt (Doane, 2016).

Es gibt noch viele weitere Regeln für Kreuztabellen, die in der Fachliteratur weiter erforscht werden können. Dazu gehören Regeln für den Fall, dass beide Variablen zufällig sind, beide Variablen unterschiedliche Skalenniveaus haben oder dass Y- oder X-Variablen Kombinationen aus zufälligen und festgelegten Variablen sind (Doane, 2016).

## 2.2 Mittelwerttest

Woher wissen wir, ob der Unterschied zwischen zwei arithmetischen Mittelwerten oder Medianen auf einen wesentlichen Verarbeitungseffekt oder nur auf eine zufällige Variation zurückzuführen ist? Aus dem Studium der induktiven Statistik wissen wir bereits, dass der Stichprobenmittelwert aufgrund eines Stichprobenfehlers vom Mittelwert der Grundgesamtheit abweichen kann. Wenn wir also einen Unterschied zwischen zwei Stichprobenmittelwerten oder -medianen feststellen, woher wissen wir dann, ob dieser Unterschied auf einen Zufallsfehler oder auf einen Verarbeitungsprozess oder eine Geschäftstätigkeit zurückzuführen ist? Die Betrachtung von Mittelwert- und Median-Vergleichstests wird es uns ermöglichen, diesen Unterschied zu erkennen. In diesem Fall kann ein **nicht parametrischer** Test verwendet werden, um Mediane zu vergleichen. Dieser Test ist als Wilcoxon-Vorzeichen-Rang-Test bekannt. Nicht parametrische Tests sind einfacher durchzuführen als parametrische Tests und beruhen nicht auf so vielen Annahmen über die Verteilung der Daten. Der Begriff „nicht parametrisch" bezieht sich ausdrücklich auf Daten, die nicht normalverteilt sind. Der Wilcoxon-Vorzeichen-Rang-Test wird bei abhängigen Stichproben verwendet, die für die Untersuchung der bivariaten Analyse, mit der wir uns hier befassen, am interessantesten sind. Der Wilcoxon-Vorzeichen-Rang-Test wertet die Differenz zwischen dem Median zweier abhängiger Stichproben aus und nicht den arithmetischen Mittelwert. Dies trägt dazu bei, dass er bei anspruchsvolleren Stichproben eingesetzt werden kann, da der Median weniger anfällig für Ausreißer in den Daten ist (Fogarty, 2017). Ein Beispiel für zwei abhängige Mittelwerte sind die Ergebnisse zweier Tests zum Wissen über Unternehmensethik, von denen einer vor, der andere nach einer Online-Ethikschulung der Angestellten durchgeführt wurde. Wir können davon ausgehen, dass die beiden Stichproben voneinander abhängig sind, da dieselben Personen den Test sowohl vor als auch nach der Schulung absolvieren.

**Nicht parametrisch**

Statistische Verfahren, die nicht parametrisch sind, erfordern nicht die gleichen Annahmen über die Datenverteilungen wie parametrische Verfahren. Allerdings müssen im Gegenzug auch Abstriche bei der Leistungsfähigkeit der statistischen Methoden gemacht werden.

Der Wilcoxon-Vorzeichen-Rang-Test wurde von Frank Wilcoxon entwickelt, als er in den American Cyanamid Labs in Stamford, Connecticut, chemische Verbindungen für Produkte von Old Spice, Combat Insektizide und Pine-Sol Reiniger erforschte (Fogarty, 2017). Es gibt zwei leicht unterschiedliche Versionen des Tests, die sich jedoch sehr ähneln und in der Regel als "Wilcoxon"-Tests bezeichnet werden. Die Grundidee der Tests besteht darin, den Median der beiden Stichproben zu vergleichen und zu sehen, ob die Differenz außerhalb dessen liegt, was man aufgrund des Fehlers bei der Ziehung einer Stichprobe erwarten würde. In dem obigen Beispiel möchten wir wissen, ob das Online-Ethik-Training einen Einfluss auf die Ergebnisse hatte (Doane, 2016).

Erinnern Sie sich, dass bei normalverteilten Daten (unter der Annahme einer ausreichend großen Stichprobe) arithmetischer Mittelwert und Median gleich sind. Für die Anwendung des Wilcoxon-Tests lautet die **Nullhypothese**, dass die Mediane der beiden Stichproben gleich sind. Dies bedeutet, dass es keinen Verarbeitungseffekt gibt (Doane, 2016). Die Forschungsfrage, die wir anhand des obigen Beispiels untersuchen, lautet, ob der Unterschied zwischen dem Mittelwert vor und nach dem Test aufgrund der Ethikschulung statistisch signifikant ist.

**Hypothese**

Hypothesentests sind entscheidend, um festzustellen, ob die Unterschiede zwischen den Medianen auf eine Behandlung oder ein Programm zurückzuführen sind oder ob es sich um einen zufälligen Stichprobenfehler handelt.

Damit ein Wilcoxon-Vorzeichen-Rang-Test durchgeführt werden kann, müssen die folgenden Voraussetzungen müssen erfüllt sein:

* Die Datenerfassungsweise vor dem Test muss mit der Datenerfassungsweise nach dem Test übereinstimmen.
* Es sollte keine verbundenen Ränge geben. Wenn es sie doch gibt, muss eine Abhilfemaßnahme ergriffen werden.
* Die abhängige Variable muss kontinuierlich sein (für kategoriale abhängige Variablen stehen andere nicht-parametrische Tests zur Verfügung, z. B. der unten vorgeschlagene alternative Test), was bei diesen Testergebnissen der Fall ist (Doane, 2016).

Die sieben Schritte des Wilcoxon-Vorzeichen-Rang-Tests sind wie folgt:

1. Paarweises Berechnen der Differenzen zwischen den Ergebnissen vor und nach der Schulung.
2. Differenzen nach ihrem Betrag aufsteigend sortieren und mit Rangwerten versehen.
3. Rangwerte wieder mit Vorzeichen versehen.
4. Positive und negative Rangwerte jeweils getrennt aufsummieren.
5. Bezeichnen Sie den vorzeichenlosen Wert der kleineren Summe der Ränge mit 6 ; n = Stichprobengröße (Anzahl der Ränge), + = Mittelwert und 0,2 = Standardabweichung.
6. Berechnen Sie *μ* = *n(n +* 1)/4, σ2 = (2*n* + 1)*μ*/6, und dann *Z*= (*T* – μ)/σ.
7. Ermitteln Sie den p-Wert (als ob es sich um α handeln würde) aus der z-Tabelle für einen einseitigen Test, falls erforderlich.

Wenn der p-Wert sehr niedrig ist (beispielsweise weniger als 0,05), bedeutet dies, dass es sehr unwahrscheinlich ist, dass die Differenz der Mittelwerte rein zufällig ist (Doane, 2016). Hypothesentests sind vielleicht eines der wichtigsten Instrumente, die in der heutigen Zeit entwickelt wurden, um den Fortschritt der Wissenschaft zu unterstützen. Viele bedeutende wissenschaftliche Entdeckungen, darunter Impfstoffe und andere lebensrettende Arzneimittel, wurden durch Hypothesentests entdeckt.

## 2.3 Korrelationen

Um die Richtung der linearen Beziehung zwischen den beiden Variablen zu bestimmen, kann die Kovarianz verwendet werden. Die Kovarianz kann beliebige positive oder negative Werte von null bis unendlich annehmen. Die Kovarianz zweier Variablen kann durch die Summe des Produkts der Differenzen aus den Variablenwerten und ihren Mittelwerten bestimmt werden. Die Gleichung dazu lautet:

(2.1)

Die Varianzen der beteiligten Variablen bestimmen die minimalen und maximalen Werte der Kovarianz. Die Skalierung der Variablen und der Maßeinheiten kann sich direkt auf diese Varianzen auswirken. Daher ist die Kovarianz nicht sehr brauchbar, um das Ausmaß des Zusammenhangs zu bestimmen.

Eine andere Kenngröße hingegen, die Korrelation, die sehr eng mit der Kovarianz verwandt und ist, hat einige nützliche Eigenschaften, wie z. B. die Tatsache, dass sie von der Änderung des Maßstabs, der Abmessungen und des Ortes unbeeinflusst bleibt. Sie kann auch verwendet werden, um zwei Variablenpaare über verschiedene Bereiche hinweg zu vergleichen (Doane, 2016). Die Korrelation wird durch Normalisierung der Korrelation ermittelt. Dazu wird sie durch das Produkt der Standardabweichungen der beiden Variablen dividiert (Doane, 2016):

(2.2)

Dabei ist Cov die Kovarianz, σx ist die Standardabweichung der Grundgesamtheit von X, and σy ist die Standardabweichung der Grundgesamtheit von Y.

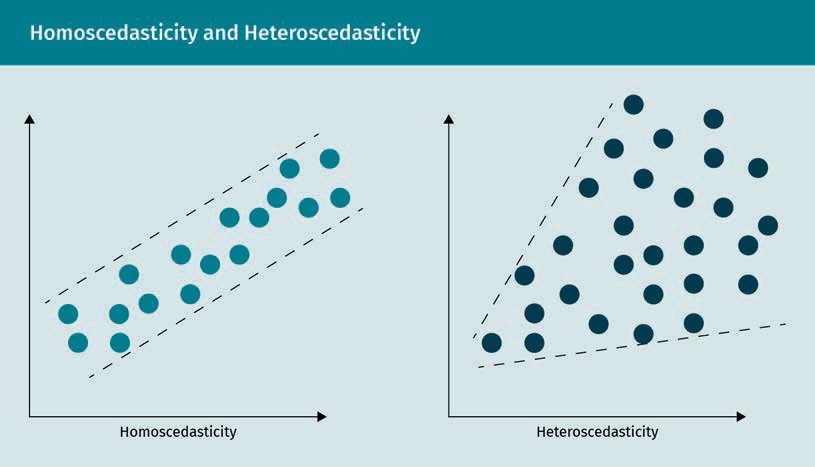
Der Korrelationskoeffﬁzient misst das Ausmaß und die Richtung der Beziehung zwischen zwei Variablen. Er ist ein Maß für den Grad der "linearen" Beziehung zwischen zwei Variablen. Der gebräuchlichste Korrelationskoeffﬁzient ist der Korrelationskoeffizient nach Pearson, in der Literatur als "r" bezeichnet (Doane, 2016).

Die Korrelation wurde ursprünglich von Sir Francis Galton, einem Cousin von Charles Darwin, im Rahmen seiner Forschungen zur Evolutionstheorie entdeckt. Karl Pearson hat später beim gleichen Forschungsthema im Zusammenhang mit Darwins Theorie die Berechnungen verfeinert – was zur Pearson-Produkt-Moment-Korrelation führte. Eine Kurzfassung der Herleitung erschien in einem 1896 von Pearson veröffentlichten Aufsatz (Fogarty, 2017).

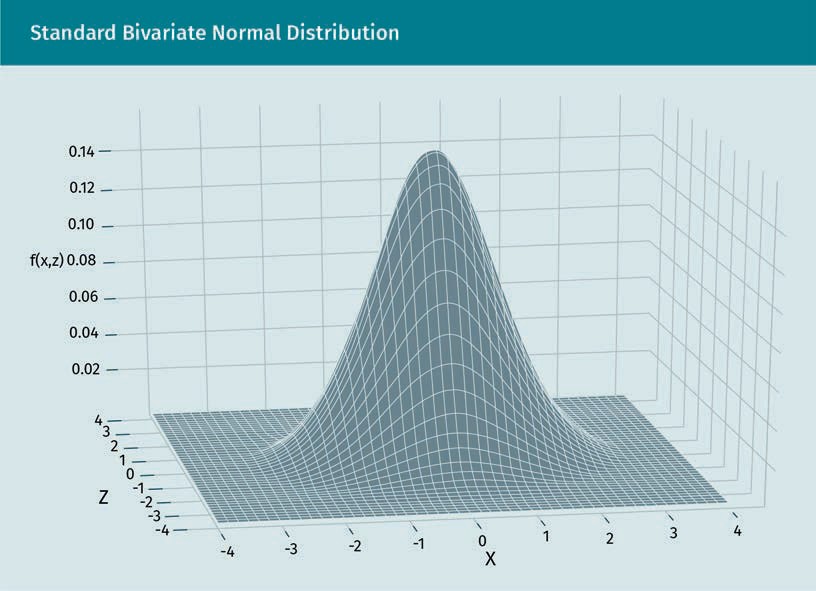
Die Korrelationskoeffizienten können zwischen -1 (perfekte inverse/negative Korrelation) und +1 (perfekte direkte/positive Korrelation) liegen. Ein Korrelationskoeffizient von (ungefähr) Null zeigt an, dass es keinen erkennbaren linearen Zusammenhang zwischen den Variablen gibt. Der quadrierte Korrelationskoeffﬁzient gibt den prozentualen Anteil der Variation in einer der Variablen an, der auf die Kenntnis der anderen Variablen zurückzuführen ist.

Wir müssen uns darüber im Klaren sein, dass der Korrelationskoeffﬁzient nur den geradlinigen oder „linearen" Zusammenhang zwischen Variablen misst; bei nichtlinearen Beziehungen wird er sehr ungenau. Eine Annahme bei der Korrelation ist die Homoskedastie. Die Homoskedastizitätsannahme besagt, dass die Varianz für alle Werte der Variablen auf der y-Achse eines Streudiagramms gleich ist. Homoskedastizität anzunehmen bedeutet, zu akzeptieren, dass die Varianz in einer Verteilung gleichmäßig ist. Bei schwerwiegenden Verstößen gegen die Homoskedastizitätsannahme wird der Pearson-Produkt-Moment-Korrelationskoeffﬁzient unterschätzt (Doane, 2016).

Die Heteroskedastizitätsannahme bei einer Korrelation ergibt sich aus einer Nichtnormalität einer der Variablen (wie in der Abbildung unten dargestellt), einer indirekten Beziehung zwischen den Variablen oder einer Datentransformation. Es ist wichtig zu beachten, dass Heteroskedastizität eine Korrelationsanalyse nicht ungültig macht, sie stellt vielmehr die Fähigkeit des Instruments in Frage, die Beziehung zwischen zwei Variablen zu messen. Homoskedastizität lässt sich mit Hilfe von Streudiagrammen feststellen und wird durch Transformationen der Variablen korrigiert (Doane, 2016).



Bei der Korrelation wird außerdem davon ausgegangen, dass jedes Paar von Variablen eine bivariate Normalverteilung aufweist. Bei einer bivariaten Normalverteilung ist jede der beiden Variablen eines Paares normalverteilt, und sie behalten eine Normalverteilung bei, wenn sie zusammen addiert werden (Doane, 2016). Die bivariate Normalverteilung wird als Glockenfläche dargestellt, wie in der folgenden Abbildung zu sehen ist.



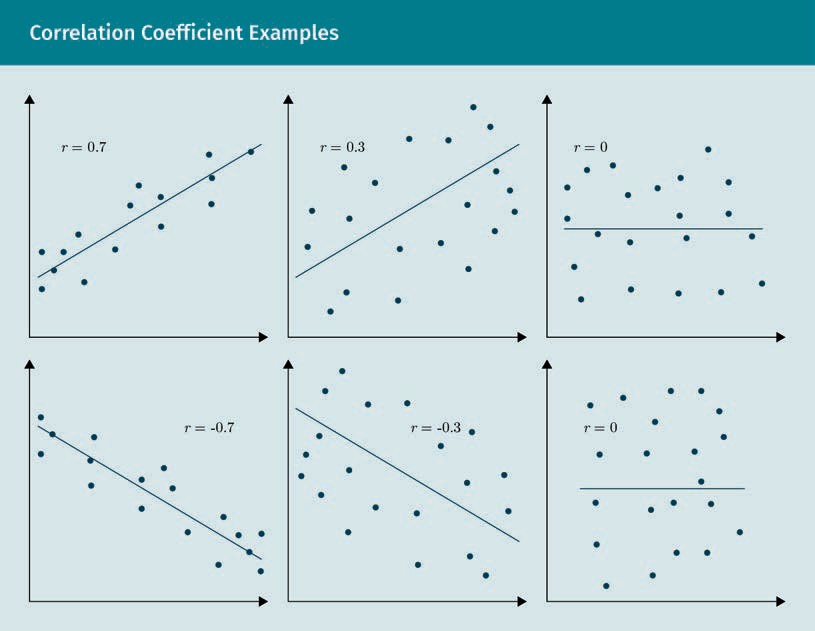
Weitere Probleme, die Forschende bei der Korrelationsanalyse beachten müssen, sind nichtlineare Beziehungen zwischen zwei Variablen, da Korrelationskoeffizienten nur für die Ermittlung linearer Zusammenhänge verwendbar sind. Ebenfalls zu ungültigen Korrelationskoeffizienten führen Ausreißer in den Daten. Daher sollten Forschende immer überprüfen, ob solche vorliegen und vor der Korrelationsanalyse Maßnahmen ergreifen, sie zu löschen oder anderweitig zu berücksichtigen. Zusätzlich zu Ausreißern können auch besonders auffällige Personengruppen die Interpretation von Korrelationskoeffizienten beeinflussen. Manchmal fassen Forschende Gruppen von Werten zusammen, um den Korrelationskoeffizienten zu berichtigen, aber dies muss mit Vorsicht geschehen, um keine falsche Interpretation zu erhalten. Des Weiteren können die in einen Korrelationskoeffizienten einbezogenen Wertebereiche gekürzt sein, was die Interpretation des Koeffizienten erschwert (Doane, 2016). Es gibt drei Arten von Korrelation:

1. Einfache Korrelation zwischen einer „abhängigen“ Variablen Y und einer „unabhängigen“ Variablen X (Doane, 2016).
2. Multiple Korrelation zwischen einer „abhängigen“ Variablen Y und mehreren „unabhängigen“ Variablen (Doane, 2016).
3. Kanonische Korrelation zwischen mehreren „abhängigen“ Variablen und mehreren „unabhängigen“ Variablen (Doane, 2016).

Die einfache Korrelation wird, wie bereits erwähnt, zwischen zwei Variablen berechnet. Die multiple Korrelation wird zwischen einer Variablen und zwei oder mehr Variablen berechnet (z. B. die Beziehung zwischen dem Body-Mass-Index einer Person als Funktion der Größe, des Alters und der durchschnittlichen täglichen Kalorienzufuhr). Das bedeutet, dass sie die Verwendung mehrerer Variablen beinhaltet (Doane, 2016).

Es ist wichtig zu beachten, dass der Korrelationskoeffizient nicht nur das Ausmaß einer linearen Beziehung angibt (nahe bei -1 oder +1), sondern auch ihre Richtung. Mit anderen Worten: Eine positive einfache Korrelation bedeutet, dass ein Anstieg der Variablen X mit einem Anstieg der Variablen Y verbunden ist (und umgekehrt). Eine negative einfache Korrelation besagt, dass ein Anstieg der X-Variablen mit einem Abfallen der Y-Variablen verbunden ist (und umgekehrt) (Akoglu, 2018).

Die nachstehenden Streudiagramme sind grafische Darstellungen der verschiedenen Ausmaße und Richtungen der linearen Zusammenhänge, die zwischen zwei Variablen bestehen können. Die gerade Linie zwischen den Punkten stellt den linearen Zusammenhang dar.



Wenn Forschende mehrere einfache Korrelationen für mehrere Variablen anzeigen möchten, verwenden sie eine Korrelationsmatrix. Die Korrelationen werden alle in eine Matrix eingetragen, in der die Zahl (1) immer die Hauptdiagonale belegt. Bei vier Variablen würden zum Beispiel vier einfache Korrelationskoeffizienten berechnet. Die Matrix enthielte dann diese vier Koeffizienten in den Elementen außerhalb der Diagonalen und der Zahl (1) in der Hauptdiagonalen Scheinkorrelationen treten auf, wenn zwei Variablen von ihren Zahlenwerten her zu korrelieren scheinen, es aber keinen wirklichen Zusammenhang gibt. Oftmals wird ihre Korrelation durch eine dritte, verborgene Variable erzeugt.

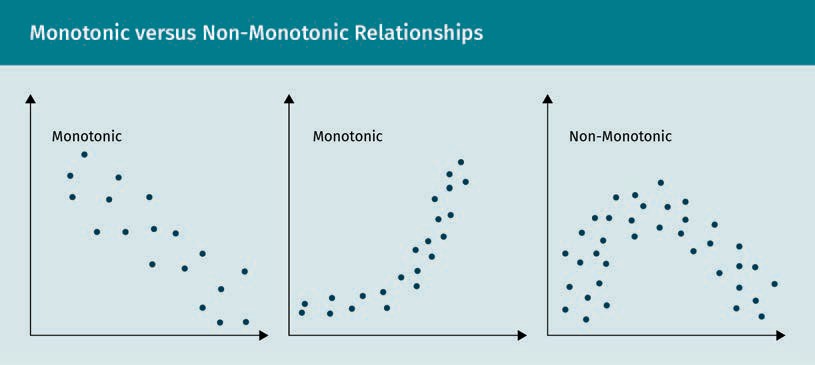
**Scheinkorrelationen**

Diese können durch eine starke Hypothese für den Grund des Zusammenhangs verhindert werden. Ohne eine Hypothese kann es leicht vorkommen, dass Korrelationen in den Daten gefunden werden, die tatsächlich gar nicht vorhanden sind.

Korrelations- und Regressionsanalyse gehen Hand in Hand, und je nach Anwendung kann die eine durch die andere ersetzt werden. Obwohl der Korrelationskoeffizient das Ausmaß und die Richtung eines Zusammenhangs beschreibt, gibt er nicht die tatsächliche lineare Beziehung an. Die Regressionsanalyse hingegen liefert die Gleichung der Geraden (der blauen Linie in den obigen Plots), die durch die Streudiagramme gelegt wird. Dabei ist b0 der Y-Achsenabschnitt und b1 die Steigung der Geraden. Der Korrelationskoeffizient und die Steigung haben stets das gleiche mathematische Vorzeichen (Doane, 2016).

Die Pearson-Produkt-Moment-Korrelation ist bei der Messung von zwei numerischen Variablen die effizienteste Gleichung. Es gibt andere Korrelationsarten, die vorteilhaft sein können, wenn die beiden Variablen unterschiedliche Skalenniveaus haben. In der Statistik betrachten wir zusätzlich zur Pearson-Korrelation drei weitere Arten von Korrelationen. Diese sind die Spearman-Korrelation, Kendall-Rangkorrelation und Punkt-biseriale Korrelation (Akoglu, 2018).

Denken Sie daran, dass bei der Pearson-Korrelation von einem linearen Zusammenhang zwischen den beiden Variablen ausgegangen wird. Die Spearman-Korrelation ist hilfreich, wenn Sie herausfinden, dass Ihre beiden numerischen Variablen nicht linear zusammenhängen (sondern z. B. exponentiell), oder eine der Variablen ordinal ist oder es Ausreißer in den Daten gibt. In jedem dieser Fälle können Sie immer noch das Ausmaß und die Richtung eines Zusammenhangs mit dem Spearman-Korrelationskoeffizienten bestimmen, der als nichtparametrischer Korrelationskennwert angesehen wird. Der Spearman-Korrelationskoeffizient ρ betrachtet die Rangwerte der zwei Variablen. Es ist wichtig zu beachten, dass ρ immer ein Wert zwischen -1 und 1 ist, da die Spearman-Korrelation in Bezug auf ihre Berechnung und Interpretation dem Pearson-Korrelationskoeffizienten entspricht. Schließlich ist zu beachten, dass die Spearman-Korrelation davon ausgeht, dass der Zusammenhang zwischen den beiden zu bewertenden Variablen monoton ist. Mit anderen Worten: Wenn eine Variable zunimmt, nimmt die andere tendenziell ab oder zu (aber nicht beides) (Doane, 2016). Die nachstehende Abbildung veranschaulicht monotone und nicht monotone Zusammenhänge.



Eine weitere nichtparametrische Version der Pearson-Korrelation ist als Kendall-Rangkorrelationskoeffizient bekannt. Dieser Koeffizient, der auch als Kendall-Tau-Korrelationskoeffizient bezeichnet wird, verwendet die Ränge der Daten, um statistische Zusammenhänge zu bewerten. Sowohl der Kendall-Tau-Korrelationskoeffizient als auch der Spearman-Korrelationskoeffizient sind gute Alternativen zum Pearson-Korrelationskoeffizenten, wenn eine von dessen Voraussetzungen nicht erfüllt ist. Der Kendall-Tau-Koeffizient wird häufig statt des Spearman-Korrelationskoeffizienten verwendet, wenn die Daten viele gleiche Ränge enthalten oder die Stichprobe sehr klein ist. In letzterem Fall sind die p-Werte mit dem Kendall-Tau-Korrelationskoeffizienten genauer (Akoglu, 2018).

Während andere Arten von Korrelationskoeffizienten (einschließlich der Pearson-Produkt-Moment-Korrelation) Beobachtungen als Grundlage für die Korrelation verwenden, benutzt der Kendall-Rang-Korrelationskoeffizient Paare von Beobachtungen und bewertet, ob es ein Muster von Konkordanz oder Diskordanz zwischen den Paaren gibt, und bestimmt die Stärke des Zusammenhangs auf der Grundlage des beobachteten Musters (Akoglu, 2018). Der Begriff "konkordant" bedeutet, dass die Paare in der gleichen Weise angeordnet sind (konsistent). Ein Paar von Beobachtungen gilt als übereinstimmend, wenn (x2 - x1) und (y2 - y1) das gleiche Vorzeichen haben. Dagegen werden diskordante Paare unterschiedlich (inkonsistent) angeordnet. Ein Paar von Beobachtungen gilt als diskordant, wenn (x2 - x1) und (y2 - y1) das gleiche Vorzeichen haben.

Die Bedingungen für die Verwendung der Kendall-Rang-Korrelation sind, dass die Variablen auf einer ordinalen oder kontinuierlichen Skala gemessen werden und dass eine monotone Beziehung zwischen den Variablen besteht (Akoglu, 2018). Die Korrelationsanalyse in ihren vielen Formen ist ein wichtiger Bestandteil der bivariaten statistischen Analyse und wird in vielen Studien eingesetzt. Außerdem bildet das menschliche Gehirn aufgrund von Erfahrungen ganz natürlich Assoziationen. Die Korrelation ist eine Möglichkeit, diese „Eingebungen“, die wir in unserem beruflichen und privaten Leben ständig haben, mathematisch zu quantifizieren.

### Zusammenfassung

Diese Lektion befasste sich mit bivariater Statistik und vermittelte einen ersten Einblick, wie Zusammenhänge zwischen zwei Variablen gemessen und beschrieben werden. Die Verwendung von Kreuztabellen ist eine der einfachsten Methoden, die Zusammenhänge zwischen zwei oder mehr diskreten Variablen zu zahlenmäßig zu erfassen. Wir betrachteten Korrelationen zwischen kategorialen Variablen und Unterschiede zwischen Mittelwerten. Wir befassten uns mit Korrelationskoeffizienten, die Richtung und Ausmaß von Zusammenhängen zwischen zwei numerischen Variablen bestimmen.

Der Wilcoxon-Vorzeichen-Rang-Test zum Vergleichen von Medianen wurde eingeführt. Der Pearson-Produkt-Moment-Korrelationskoeffizient misst ebenfalls das Ausmaß und die Richtung von Zusammenhängen zwischen zwei Variablen. In Situationen, in denen die Pearson-Korrelation nicht anwendbar ist (z. B. bei mehr als zwei Variablen), können wir die Spearman-Korrelation, die Kendall-Rangkorrelation oder die Punkt-biseriale Korrelation einsetzen. In der modernen Welt von „Big Data“ werden bivariate Statistiken benötigt, die Zusammenhänge zwischen Variablen quantifizieren. Sie ermöglichen das Verständnis von Faktoren, die mit Phänomenen zusammenhängen, die wir in der Wirtschaft und in unserem persönlichen Leben beobachten.

# Lektion 3 – Wahrscheinlichkeitsverteilungen

### Lernziele

Nach der Bearbeitung dieser Lektion werden Sie in der Lage sein, ...

... diskrete und numerische Zufallsvariablen zu identifizieren und zu verstehen.

... die Eigenschaften der Standardnormalverteilung und allgemeiner Normalverteilungen zu identifizieren.

... die Studentsche T-Verteilung zu interpretieren.

# 3. Wahrscheinlichkeitsverteilungen

## Aus der Praxis

Das Biotechnologie-Unternehmen Riedling hat sich auf die Entwicklung besserer Gesundheitslösungen für eine alternde Bevölkerung spezialisiert. Kürzlich hat die Firma Daten zu ihren Lösungen gesammelt, um die langzeitige Wirksamkeit ihrer Behandlungen besser einzuschätzen. Zuvor hatte das Unternehmen viele dieser Informationen in klinischen Tests erhoben. Nun, wo die Behandlungen auf dem Markt sind, möchte Riedling jedoch eine längerfristige Prognose gewinnen.

Die gesammelten Daten enthalten alle physischen Informationen über die getesteten Patienten. Dazu gehören unter anderem Größe, Gewicht, Alter, Geschlecht und Ehestand. Ebenfalls erfasst wurden medizinische Messergebnisse wie Blutdruck, Cholesterin- und Blutzuckerspiegel. Es wurden auch Daten über den Gesundheitszustand der Personen erhoben, einschließlich des Auftretens bestimmter schwerwiegender Krankheiten wie Diabetes, kardiopulmonaler Erkrankungen und vieler anderer klinischer Zustände.

Verantwortlich für diese Initiative war Harvey Kramer, der leitende Biostatistiker des Unternehmens. Er wollte einige ausgefeilte statistische Analysen mit den Daten durchführen, als erstes alle Variablen als diskret oder kontinuierlich klassifizieren und die Mittelwerte, Standardabweichungen und andere deskriptive Statistiken berechnen. Dies hatte er in seiner Ausbildung über beschreibende Statistik gelernt. Diesmal wollte er jedoch noch einen Schritt weiter gehen und Histogramme für alle Daten erstellen, da er wusste, dass er bei statistisch unauffälligen Daten Techniken wie Korrelationsanalysen und allgemeine lineare Modelle anwenden konnte. Ihm fiel sofort auf, dass Variablen wie Größe, Gewicht und Blutdruck normalverteilt waren, während andere Messwerte, wie das gesundheitliche Engagement, eine andere Verteilung hatten. Harvey erstellte zuerst eine neue Tabelle mit allen normalverteilten und nicht-normalverteilten Variablen, bevor er mit der eigentlichen Analyse begann.

## 3.1 Zufallsvariablen und ihre Verteilungen

In der Welt, in der wir leben, gibt es viele Zufallserscheinungen. Einige Wissenschaftler:innen, die in der klassischen Mechanik ausgebildet wurden, argumentieren, dass es im Universum keine echte Zufälligkeit gibt und dass der „Zufall“, wie wir ihn kennen, nur der Mangel an Informationen über das Universum ist. Wissenschaftler:innen, die mit dem Denkmodell der Quantenphysik vertraut sind, argumentieren hingegen, dass es im Universum echten Zufall gibt, und sie verfügen sogar über Quantenexperimente, die diese Vermutung untermauern. Unabhängig davon, wie man zu dieser Diskussion steht, werden wir als Menschen in unserem täglichen persönlichen und beruflichen Leben viel Zufälliges erleben.

**Zufallsvariablen**

Im Gegensatz zu algebraischen Variablen, die einen festen Wert haben, haben Zufallsvariablen eine ganze Reihe von Werten und können jeden davon annehmen.

In der Statistik wird die Zufälligkeit in Form von Zufallsvariablen erfasst und gemessen. In der Regel werden **Zufallsvariablen** aus Experimenten abgeleitet; sie können jedoch auch aus Prozessen stammen, die kontinuierlich um uns herum ablaufen. Zufallsvariablen sind entscheidende Komponenten der statistischen Analyse und ihre Kenntnis ist für ein umfassendes Verständnis von Wahrscheinlichkeitsverteilungen unbedingt erforderlich (McEvoy, 2018).

Zufallsvariablen erleichtern uns die zahlenmäßige Erfassung von Zufallsprozessen und die Durchführung statistischer Berechnungen. Eine der bekanntesten Zufallsvariablen entsteht durch den Münzwurf. Wenn wir eine Münze einhundertmal werfen und jedes Ergebnis erfassen, ist das aufgezeichnete Resultat eine Zufallsvariable. Zufallsvariablen werden entweder als diskret oder als kontinuierlich klassifiziert. Diskrete Zufallsvariablen, wie zum Beispiel die Anzahl von Studierenden in einem Klassenraum, können durch Zählen bestimmt werden. Kontinuierliche Zufallsvariablen hingegen können nur durch Messungen bestimmt werden. Ein Beispiel für eine kontinuierliche Zufallsvariable ist z. B. die Zeit, die in Jahrhunderte, Jahrzehnte, Jahre, Tage, Stunden, Minuten, Sekunden und Millisekunden unterteilt werden kann (McEvoy, 2018).

Kontinuierliche oder diskrete Zufallsvariablen aus Experimenten sind in der Statistik wichtig, weil sie Verteilungen mit Eigenschaften aufweisen, die bei der Durchführung statistischer Analysen hilfreich sind. Beispiele für diskrete Zufallsvariablen sind die Anzahl der Kätzchen in einem Wurf, die Anzahl der Besucher:innen eines Basketballspiels, die Anzahl der Schüler:innen in der Klasse und die Anzahl der Personen, die zu einer bestimmten Uhrzeit und an einem bestimmten Tag in einem Kaufhaus Schlange stehen. Kontinuierliche Zufallsvariablen hingegen können nicht gezählt werden und können unendlich viele Werte annehmen. Mit anderen Worten: Kontinuierliche Zufallsvariablen haben eine unbestimmte Genauigkeit und können nur durch Messungen ermittelt werden. Beispiele dafür sind Entfernung, Temperatur, Druck, Größe, Gewicht, Dichte und Volumen.

Es ist sinnvoll, darauf hinzuweisen, dass es keine absoluten Werte gibt. Diskrete Zufallsvariablen können teilweise kontinuierlich sein, und ebenso können kontinuierliche Zufallsvariablen teilweise diskret sein. Solche Variablen werden als Variablen mit gemischtem Typ bezeichnet.

## 3.2 Diskrete Wahrscheinlichkeitsverteilungen

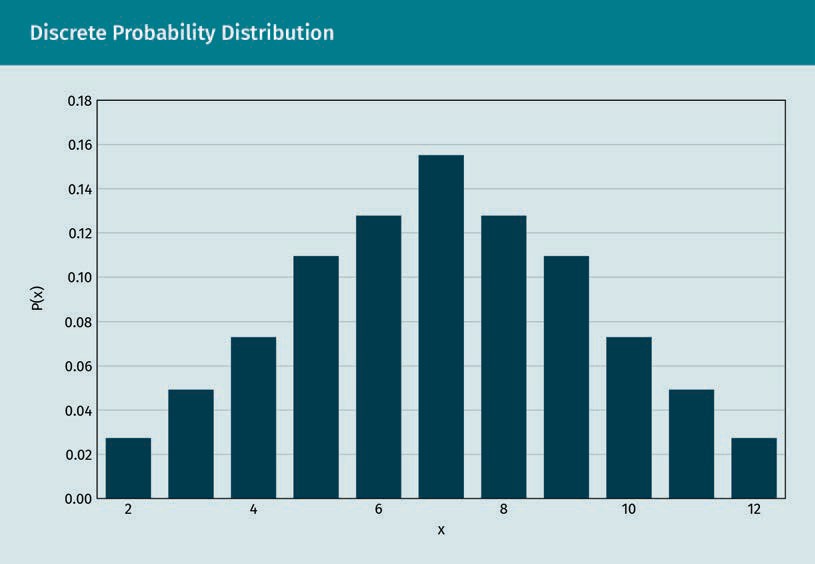
Diskrete Zufallsvariablen haben eine Verteilung, die als **diskrete Wahrscheinlichkeitsverteilung** bezeichnet wird (Viti et al., 2015; McEvoy, 2018). Die diskrete Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen ist ein Verzeichnis von Wahrscheinlichkeiten, die mit jedem ihrer möglichen Werte verbunden sind. Sie wird gelegentlich auch als Wahrscheinlichkeitsmassefunktion bezeichnet. Zu den bekanntesten diskreten Wahrscheinlichkeitsverteilungen gehören die binomiale Wahrscheinlichkeitsverteilung, die hypergeometrische Wahrscheinlichkeitsverteilung, die multinomiale Wahrscheinlichkeitsverteilung, die negative Binomialverteilung und die Poisson-Wahrscheinlichkeitsverteilung (Viti et al., 2015; McEvoy, 2018).

**Diskrete Wahrscheinlichkeitsverteilung**

Eine diskrete Wahrscheinlichkeitsverteilung stellt die Wahrscheinlichkeit des Auftretens jedes einzelnen Wertes einer diskreten Zufallsvariablen dar. Dabei kann jedem Wert eine Wahrscheinlichkeit zugeordnet werden, die größer als Null ist.

Bei einer diskreten Wahrscheinlichkeitsverteilung wird jeder mögliche Wert, den die Zufallsvariable annehmen kann, in Verbindung mit seiner Wahrscheinlichkeit aufgeführt. Die folgenden Bedingungen müssen erfüllt sein, damit eine Wahrscheinlichkeitsverteilung gültig ist. Dabei sei x eine diskrete Zufallsvariable mit den möglichen Ergebniswerten x1, x2, ..., xn. Erstens, jeder Wert der diskreten Zufallsvariablen hat eine Wahrscheinlichkeit zwischen 0 und 1, bzw. 0 ≤ P(xi) ≤ 1. Zweitens, die Summe aller Wahrscheinlichkeiten ∑P(xi) = 1 (Viti et al., 2015; McEvoy, 2018).

Diskrete Wahrscheinlichkeitsverteilungen können mit Hilfe eines Histogramms abgebildet werden, wie im folgenden Beispiel. Dieses Histogramm erhält man, weil X eine diskrete Zufallsvariable ist, die mehr als ein mögliches Ergebnis hat. Zunächst wird die Wahrscheinlichkeit auf der y-Achse und das Ergebnis der diskreten Variable auf der x-Achse aufgetragen. Wenn wir das Experiment über viele Versuche wiederholen und jedes Mal die Wahrscheinlichkeit jedes möglichen Ergebnisses aufzeichnen, erhalten wir ein Diagramm, das die Wahrscheinlichkeiten darstellt. Diese Darstellung wird als Wahrscheinlichkeitsverteilung bezeichnet. Die Höhe der einzelnen Balken im Diagramm für X gibt die Wahrscheinlichkeit des jeweiligen Ergebnisses an (Viti et al., 2015; McEvoy, 2018).



Wir haben nun gelernt, dass eine diskrete Wahrscheinlichkeitsverteilung die Verteilung einer diskreten Zufallsvariablen beschreibt. Eine diskrete Zufallsvariable ist eine Variable, die sich aus einem Experiment mit diskreten Zahlen oder abzählbaren Werten ergibt. So ist beispielsweise das Zählen der Ankünfte von Flugzeugen auf einem Flughafen eine diskrete Zufallsvariable. Jeder Wert einer diskreten Zufallsvariablen hat eine Wahrscheinlichkeit, die zwischen 0 und 1 einschließlich liegt. Wir haben nun gelernt, dass eine diskrete Wahrscheinlichkeitsverteilung die Verteilung einer diskreten Zufallsvariablen beschreibt. Eine diskrete Zufallsvariable ist eine Variable, die sich aus einem Experiment mit diskreten Zahlen oder abzählbaren Werten ergibt. So ist beispielsweise das Zählen der Ankünfte von Flugzeugen auf einem Flughafen eine diskrete Zufallsvariable. Jeder Wert einer diskreten Zufallsvariablen hat eine Wahrscheinlichkeit, die zwischen null und eins einschließlich liegt. Außerdem ergeben alle Wahrscheinlichkeiten in der Summe Irgendein Kram.

Diskrete Wahrscheinlichkeitsverteilungen können mit einem Histogramm oder einem Wahrscheinlichkeitsverteilungsdiagramm visualisiert werden. Zunächst wird die Wahrscheinlichkeit auf der y-Achse und das Ergebnis der diskreten Variable auf der x-Achse aufgetragen. Es gibt viele Arten von Verteilungen von diskreten Zufallsvariablen, darunter die binomische Wahrscheinlichkeitsverteilung und die hypergeometrische Wahrscheinlichkeitsverteilung (Viti et al., 2015; McEvoy, 2018).

3.3 Kontinuierliche Wahrscheinlichkeitsverteilungen

Wenn die Zufallsvariable kontinuierlich ist, heißt ihre Verteilung **kontinuierliche Wahrscheinlichkeitsverteilung**. Die Wahrscheinlichkeit, dass eine kontinuierliche Zufallsvariable einen bestimmten Wert in einer Verteilung annimmt, ist null, was ein wesentliches Unterscheidungsmerkmal zwischen einer kontinuierlichen und einer diskreten Wahrscheinlichkeitsverteilung ist (Crooks, 2019). Folglich muss eine kontinuierliche Wahrscheinlichkeitsverteilung mit einer Gleichung oder Formel beschrieben werden und nicht in Tabellenform wie eine diskrete Wahrscheinlichkeitsverteilung (Crooks, 2019).

**Kontinuierliche Wahrscheinlichkeitsverteilung**

Wird in der Regel in Form einer Kurve dargestellt; die darunterliegende Fläche gibt die kumulative Wahrscheinlichkeit an. Typischerweise beschreibt sie die Wahrscheinlichkeit durch eine spezifische Formel, Geometrie, Technologie oder Wahrscheinlichkeitstabelle.

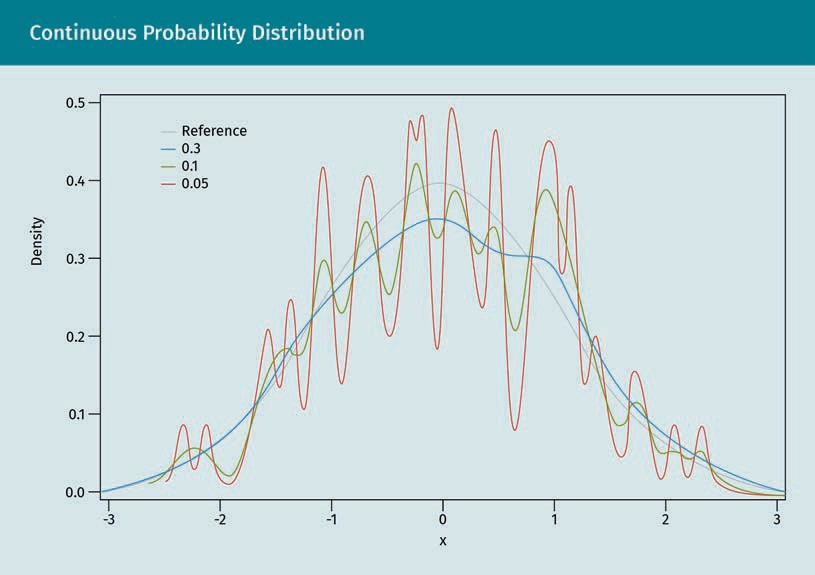
Eine diskrete Wahrscheinlichkeitsverteilung hat einen abzählbaren Wertebereich. Wenn eine Kartenfirma zum Beispiel Geburtstagskarten für alle möglichen Altersgruppen erstellen wollte, würden sie von null bis 122 reichen, wobei 122 für die älteste Person steht, die jemals gelebt hat. Eine kontinuierliche Verteilung von Zufallsvariablen hingegen hat einen Wertebereich, der unendlich ist, und daher nicht abzählbar ist. Eine kontinuierliche Verteilung einer Zufallsvariablen kann beliebig viele Zwischenwerte innerhalb des möglichen Bereichs der Zufallsvariablen annehmen.

Nehmen wir zum Beispiel an, dass die Temperatur in New York City im Monat Juli in den vergangenen zehn Jahren immer zwischen 34 und 44 °C gelegen hat. Die Temperatur kann jeden Wert zwischen 34 und 44 Grad Celsius annehmen. Die Temperatur kann an einem beliebigen Tag 39,15°C oder 40,15°C betragen, oder sie kann jeden Wert zwischen 39,15°C und 40,15°C annehmen. Wenn wir sagen, dass die Temperatur 39°C beträgt, bedeutet das, dass die Temperatur irgendwo zwischen 38,5°C und 39,5°C liegt. Jede Beobachtung, die gemacht wird, fällt in das Intervall (Crooks, 2019).

Zusätzlich zu den Unterschieden zwischen diskreten und kontinuierlichen Wahrscheinlichkeitsverteilungen wird eine Formel verwendet, um eine kontinuierliche Wahrscheinlichkeitsverteilung auszudrücken (Crooks, 2019). Während die Beschreibung diskreter Wahrscheinlichkeitsverteilungen oft mit Häufigkeitsdiagrammen oder Tabellen beschrieben erfolgt, werden kontinuierliche Wahrscheinlichkeitsverteilungen mit einer Formel ausgedrückt, die als **Wahrscheinlichkeitsdichtefunktion** bekannt ist (Crooks, 2019). Die nachstehende Abbildung ist ein Beispiel für eine typische kontinuierliche Wahrscheinlichkeitsverteilung mit der kontinuierlichen Zufallsvariablen auf der x-Achse und der Wahrscheinlichkeitsdichte auf der y-Achse.

**Wahrscheinlichkeitsdichtefunktion**

Da es sich bei vielen dieser Verteilungen nicht um geschlossene Lösungen handelt, wird die Integralrechnung verwendet, um die Fläche unter der Kurve zu ermitteln, die durch die Wahrscheinlichkeitsdichtefunktion repräsentiert wird.



Bei den kontinuierlichen Wahrscheinlichkeitsverteilungen ist die Normalverteilung nach wie vor die am weitesten verbreitete Variante. Sie kann nicht nur die Verteilung vieler kontinuierlicher Zufallsvariablen in der Biologie, den Sozialwissenschaften und der Wirtschaft darstellen, sondern auch die Poisson-Verteilung sowie die Binomial- und hypergeometrischen Verteilungen approximieren (Shakil et al., 2010). Weitere kontinuierliche Verteilungen, die in diesem Abschnitt nicht behandelt werden, aber dennoch für bestimmte Anwendungen wichtig sind, sind die Beta-, Cauchy-, Exponential-, Gamma-, logistische und Weibull-Verteilung. In der Literatur gibt es auch eine Reihe von weniger verbreiteten kontinuierlichen Verteilungen, z. B. die Shakil-Singh-Kibria-Verteilung, die auf den Whittaker-Funktionen basiert (Shakil et al., 2010).

Die Wahrscheinlichkeit, dass eine kontinuierliche Zufallsvariable einen bestimmten Wert annimmt, ist null und stellt somit ein wesentliches Unterscheidungsmerkmal zwischen einer diskreten und einer kontinuierlichen Wahrscheinlichkeitsverteilung dar. Folglich wird eine kontinuierliche Wahrscheinlichkeitsverteilung mit einer Gleichung oder Formel beschrieben und nicht in Tabellenform wie eine diskrete Wahrscheinlichkeitsverteilung (Crooks, 2019). Diese Formel wird häufig als Wahrscheinlichkeitsdichtefunktion bezeichnet. Kontinuierliche Zufallsvariablen gehören zu den wichtigsten Messgrößen menschlicher und wirtschaftlicher Aktivitäten auf unserem Planeten, und kontinuierliche Wahrscheinlichkeitsverteilungen sind für das Verständnis und die Vorhersage dieser Variablen von entscheidender Bedeutung. Es ist jedoch wichtig, sich daran zu erinnern, dass angesichts der Natur kontinuierlicher Wahrscheinlichkeitsverteilungen die Arbeit mit diskreten Verteilungen viel einfacher ist, deshalb beginnen viele Forschenden ihre Analyse mit einer diskretisierten Version einer kontinuierlichen Variablen (Crooks, 2019).

## 3.4 Normalverteilung

Bei frühen landwirtschaftlichen Studien wurde festgestellt, dass sich Pflanzen bei kontinuierlicher Züchtung mit überlegenen Sorten nicht unbegrenzt verbessern. Stattdessen verändern sie sich und bilden eine durchschnittliche Sorte, die zwar Ausreißer hat, aber die meisten Exemplare befinden sich um das Zentrum der Verteilung herum. Dieses Prinzip ist als Regression zum Mittelwert bekannt und wurde erstmals von Francis Galton beschrieben, der dies zunächst bei landwirtschaftlichen Experimenten feststellte, bevor er die Studie auf die Körpergröße von Eltern und ihren Kindern ausweitete. Ein russischer Mathematiker des 19. Jahrhunderts, P.L. Tschebyscheff, stellte fest, dass der Prozentsatz der Beobachtungen, die zwischen zwei verschiedene Werte fallen, deren Abweichungen vom Mittelwert denselben absoluten Wert haben, mit der Varianz der Grundgesamtheit zusammenhängt. Daraufhin entwickelte das Tschebyscheff-Theorem, das eine konservative Schätzung des oben genannten Prozentsatzes liefert (Fogarty, 2017). Für jede Grundgesamtheit oder Stichprobe, liegt mindestens

(3.1)

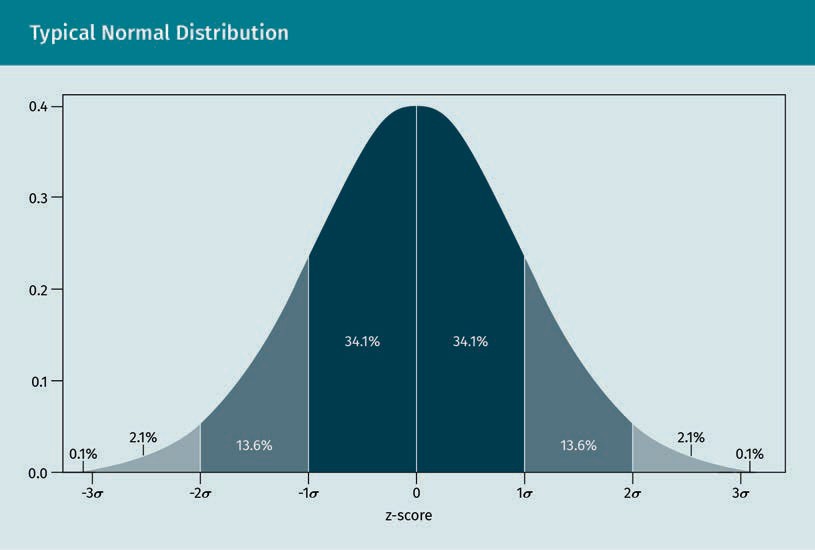
der Beobachtungen des Datensatz innerhalb von k Standardabweichungen zum arithmetischen Mittelwert, wobei k ≥ 1.

Schließlich leiteten die Mathematiker Gauß und Adrian unabhängig voneinander die Formeln für die **Normalverteilung** her. Es war jedoch Gauß, der dafür die Auszeichnung erhielt, indem die Normalverteilung nach ihm als Gauß-Verteilung benannt wurde(Fogarty, 2017). Da die Normalverteilung keine geschlossene Lösung ist, muss man die tatsächlichen Wahrscheinlichkeiten mit Hilfe der Integralrechnung ermitteln. Um die Berechnungen zu vereinfachen werden daher Wahrscheinlichkeitstabellen für die Standardnormalverteilung verwendet. Diese Wahrscheinlichkeiten entsprechen der Fläche unter der Normalverteilung, wobei die gesamte Fläche den Flächeninhalt 1 hat. Die einzelnen Wahrscheinlichkeiten hängen davon ab, wo ein Wert in der Verteilung im Verhältnis zum arithmetischen Mittelwert liegt (Fogarty, 2017).

**3.4 Normalverteilung**

Die Kenngrößen der Grundgesamtheit bestimmen die Form und die Wahrscheinlichkeiten jeder Wahrscheinlichkeitsverteilung, einschließlich der Normalverteilung. Die Form der Normalverteilung wird durch zwei Kenngrößen definiert: durch den arithmetischen Mittelwert und die Standardabweichung.

Die Normalverteilung weist eine Glockenform auf und ist symmetrisch um den Mittelwert. Dies bedeutet, das die Seite links vom Mittelpunkt das Spiegelbild der Seite rechts vom Mittelpunkt ist. Die Standardabweichung bestimmt die Breite der Verteilung, wobei höhere Werte eine wesentlich flachere Kurve beschreiben. Die Kenngrößen arithmetischen Mittelwert und Standardabweichung der Normalverteilung verändern die Form der Verteilung, was zu einer Vielzahl von Varianten führt (Attenborough, 2003). Die Abbildung unten zeigt ein Beispiel für eine typische Normalverteilung.



In Bezug auf ihre Bedeutung ist die Normalverteilung vielleicht die Nummer eins im Studium der Statistik, da viele kontinuierliche Zufallsvariablen im Geschäfts- und Alltagsleben diese glockenförmige Kurve aufweisen, wenn sie erfasst und grafisch dargestellt werden. Dies ist der Grund, warum die Normalverteilung oft als Glockenkurve bezeichnet wird (Doane, 2016).

Die Parametrische Statistik setzt voraus, dass die Daten normalverteilt sind. Sie umfasst einige der leistungsfähigsten Techniken, die in der Wirtschaft und in den Sozialwissenschaften eingesetzt werden. Wenn die Daten nicht normal verteilt sind, müssen die Forschenden sie entweder transformieren oder ein weniger leistungsfähiges statistisches Verfahren verwenden, das als nichtparametrische Statistik bekannt ist (Doane, 2016; McEvoy, 2018).

Die Mehrheit der kontinuierlichen Datenwerte in einer Normalverteilung tendiert dazu, sich um den arithmetischen Mittelwert zu gruppieren. Je weiter die Beobachtung vom Mittelwert abweicht, desto geringer ist die Wahrscheinlichkeit des Eintretens eines Ereignisses. Die Enden der Normalverteilung gehen in die Unendlichkeit über und sind asymptotisch, d. h. sie nähern sich der x-Achse bzw. dem Horizont, ohne diesen jedoch ganz zu erreichen (Doane, 2016).

Since the deviations from the normal curve are predictable in terms of the probability of deviating from the mean value, the area under the normal distribution represents the probability, and the total area under the normal curve represents all of the proba- bilities, therefore summing to one (Doane, 2016). Interestingly, in a perfect normal dis- tribution, the mean, median, and mode will be the same, or very similar, values and lie at the very peak of the curve. There are speciﬁc tests for normality of variables that take this principle into account by calculating and comparing the mean, median, and mode (Doane, 2016).

Oftentimes, students will come across the term “standard normal distribution,” but what is the difference between a normal distribution and a standard normal distribu- tion? A normal distribution is determined by two population parameters: the mean and the variance. These can take on any values, but share the same probabilities regardless. Taking this into account, a normal distribution can be standardized to have a mean of zero and a standard deviation of one. This standardization creates what is referred to as the standard normal distribution (Doane, 2016). The values, or raw scores, are typi- cally standardized by turning them into z-scores. This standardization procedure ena- bles researchers to ascertain the proportion of value that falls within a speciﬁed stand- ard deviation from the mean. This way, they are able to use the empirical rule (Doane, 2016; McEvoy, 2018).

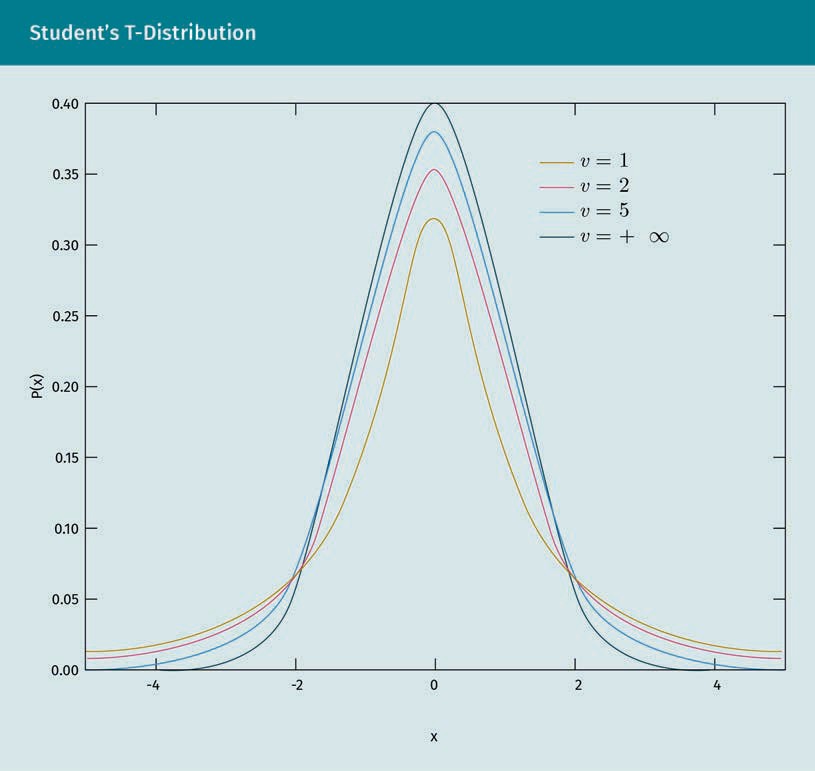
The empirical rule is a unique property of the normal distribution. It’s actually also referred to as the “95% rule” since this is the most frequently used interval in hypothe- sis tests. The 95% rule states that 95 percent (actually 95.44 percent) of observations fall within two standard deviations of the mean on a normal distribution (Doane, 2016; McEvoy, 2018).

The empirical rule expands this deﬁnition by stating that, in a normal distribution, around 68 percent of data will be within one standard deviation of the mean, around 95 percent will be within two standard deviations of the mean, and around 99.7 percent will be within three standard deviations of the mean. Therefore, in statistics, we often use the 68–95–99.7 rule as shorthand to remember the percentages of observations that fall into the three standard deviations from the mean in a normal distribution (Doane, 2016).

3.5 T-Distribution

The popular Pearson product moment correlation was ﬁrst developed and published by Karl Pearson, the British biostatistician and mathematician who was a protégé of Fran- cis Galton and developed statistical techniques to validate Darwin’s Theory of Evolu- tion. One of Pearson’s own protégés was William Sealy Gossett, who was a gifted statis- tician, hired by the famous Guinness Brewing Company to improve the quality processes at the brewery. Gossett found that the present statistical techniques that assumed large samples did not work very well in the applications he was working on at Guinness, which had a very small sample size. He therefore engaged in research around testing a statistical hypothesis with small samples, and developed knowledge around the t-distribution (Fogarty, 2017). Gossett corresponded with Karl Pearson, most famous for developing the Pearson product moment correlation.

Karl Pearson saw the value of the work that Gossett was doing to the broader scholarly community, as opposed it to being proprietary only to Guinness Brewery. The communi- cation of this type of information to scholars is primarily done through scholarly jour- nals or at academic conferences. Since Gossett did not want Guinness Brewery to know that he was publishing some of the statistical work he was doing at the brewery, he published his work under the pseudonym “Student.” Hence, we now know why the t- test is referred to as Student’s t-test. We are very lucky that Gossett published his work under a pseudonym despite facing the risk of becoming employed, as hypothesis test- ing, which includes the use of t-tests is regarded as one of the most important discov- eries of the nineteenth century (Fogarty, 2017). The illustration below depicts a typical t- distribution.



When compared with the shape of the standard normal distribution one can see that the t-distribution is still bell-shaped but has heavier tails. The t-distribution is deﬁned by the degrees of freedom. This is a correction related to sample size (Doane, 2016; Tat- tar, Ramaiah, & Manjunath, 2016; McEvoy, 2018).

### Degrees of Freedom

In applied statistics, the df, or **degrees of freedom**, is the number of independent val- ues that can vary in your analysis without violating any constraints. This concept is an important one that is used in many techniques, including the t-distribution and t-tests, and can impact the precision and power of one’s analyses (Fogarty, 2017; Frost, 2019).

**Degrees of freedom**

This concept was ﬁrst mentioned by Carl Frederick Gauss in the 1880s, and was brought into modern statistics by William Sealy Gos- sett in the early1900s

Degrees of freedom (typically a positive whole number) are usually calculated by taking the sample size and subtracting the number of parameters needed (Frost, 2019).

The idea of degrees of freedom is that the the quantity of independent information you have constrains the number of parameters that you can estimate (Frost, 2019). Using this logic researchers always desire more information to go into parameter esti- mates and have more powerful statistical analyses. Therefore, the more degrees of freedom, the better (Frost, 2019).

### Comparison of the T-Distribution with the Normal Distribution

Researchers are limited in their ability to choose the normal distribution (which requires the population standard deviation) and will default to using a t-distribution (which only requires using the sample standard deviation) for their hypothesis testing (Virginia Tech, 1999). However, since t-tests are recommended with sample sizes of less than 30 observations, researchers must take sample sizes into account when thinking about their critical values. However, the t-distribution exhibits asymptotic properties and will exhibit almost the same critical values as the normal distribution when sam- ple sizes exceed 30 observations (Virginia Tech, 1999).

However, the t-distribution is still very similar to the normal distribution in that it has a precise mathematical deﬁnition. Other similar features include the smooth shape of both the normal distribution and the t-distribution, and that the normal distribution and the t-distribution are both symmetrical (each side of the mean will mirror the other). Both distributions also have a mean of zero. The ultimate similarity is that, as the sample size approaches inﬁnity (i.e., becomes larger), the t-distribution very closely resembles the normal distribution. (Doane, 2016; Tattar et al., 2016; McEvoy, 2018).

## Zusammenfassung

Random variables make it easy for us to quantify random processes and perform statistical calculations. Random variables emerging from experiments, which can be either continuous or discrete, are important in statistics, because these variables all follow distributions that have properties that are useful when conducting statis- tical analysis. Discrete probability distribution describes the distribution of a dis- crete random variable, which is a variable resulting from an experiment with dis- crete number or countable values. Discrete probability distributions can be depicted using a histogram or probability distribution plot. If the random variable is continuous, then its distribution is classiﬁed as a continuous probability distri- bution.

The normal distribution recorded by Gauss is bell-shaped with probabilities repre- senting the area under the curve; the entire area and all of the individual values is equal to one, depending on where the value lies on the distribution in relation to the mean. The empirical rule states that, in a normal distribution, around 68 per- cent of data will be within one standard deviation of the mean, around 95 percent will be within two standard deviations, and around 99.7 percent will be within three standard deviations.

A t-distribution (i.e., Student’s t-distribution) describes the standardized distances of sample means (in terms of standard deviations) to the population mean when the population standard deviation is not known (and only the sample standard deviation is available for use), and the observations come from a normally distrib- uted population.

# Unit 4 – Statistical Estimation Methods

## Lernziele

Nach der Bearbeitung dieser Lektion werden Sie in der Lage sein, ...

… compute and analyze a point estimate.

… compute and analyze an interval estimate.

… decide when to use a conﬁdence interval versus a single point estimate.

# 4. Statistical Estimation Methods

## Aus der Praxis

The Swiss company, Accurate Timepieces, is a maker of luxury watches, and had recently done an initial public offering (IPO) and had shares traded on the Swiss stock exchange (SIX). The chief executive ofﬁcer (CEO), Hermann Marx, began working with Marie Schultz, the chief ﬁnancial ofﬁcer (CFO), to create an annual shareholder report. Hermann and Marie calculated many key performance indicators (KPIs) for the organi- zation from the business operations and ﬁnance, and wanted to report some of these ﬁgures in the annual report.

However, Hermann noticed from the analysis that quite a few of these measures con- tained a high degree of variability and, therefore, a single point estimate (i.e., average) may not be suitable for reporting. Doing this would be somewhat deceptive to share- holders. Marie had a great suggestion that any time some of the KPIs displayed a high degree of variance, as measured by their standard deviation, they would create conﬁ- dence intervals and report a range of values, within which the population parameter could reside. Hermann thought this was a great idea. Now all they needed to do was decide on the conﬁdence level for the calculations. “This is an easy decision,” exclaimed Hermann. “We are a Swiss watch company and precision matters. Therefore, it must be at least a 95 percent conﬁdence level,” Marie said, agreeing.

4.1 Point Estimation

**Point estimators** are metrics or functions that are utilized in inferential statistics to ascertain an approximate value of a population parameter from a random sample selected from a population. Point estimators extract the data from a sample of the population to calculate a point estimate or a statistic that serves as the most likely estimate of an unknown population parameter. Point estimators can be a variety of sta- tistics, including means, proportions, medians, and standard deviations (Doane, 2016; McEvoy, 2018).

**Point estimators**

A single value or point is used to draw inferences about a population by esti- mating the value of an unknown param- eter in a point esti-mator.

A point estimate can be useful if a researcher wants a single number to be representa- tive of the entire population. This is only possible where central tendency is present, such as in a normal distribution where the most likely value will be the mean. The probability of assuming any particular value is zero in a continuous probability distri- bution, so researchers attempt to generate a value that is as close as possible to the population parameter. If this objective is met, then the point estimator is deemed an unbiased estimator. Other characteristics used to evaluate an estimator include consis- tency and efﬁciency. Managers measuring performance favor point estimates because they can assign people, processes, and businesses a goal based on the measure. How- ever, managers are cautioned that the results may just be due to random ﬂuctuation without a hypothesis test or using an interval estimate (Doane, 2016; McEvoy, 2018).

There are diverse types of point estimators, each of which contains different properties in relation to being reliable estimators of the population. Some of the more popular ones include the Bayes estimator, minimum-variance mean-unbiased estimator (MVUE), best linear unbiased estimator (BLUE), minimum mean squared error (MMSE), maximum likelihood estimator (MLE), median-unbiased estimator, method of moments, and generalized method of moments (Lehmann & Casella, 1998).

A point estimate is used when an analyst desires a single point of the most likely value of a population parameter. This can be any univariate statistic, including the mean or even a proportion. However, it should be noted that the point estimate calculated from a sample of the population will always vary based on random sampling error. It is always recommended that a point estimate is presented along with some measure of dispersion (i.e., standard deviation) so that the consumer of the statistic will have a gauge as to how reliable that estimate is, given the amount of variation in the popula- tion or sample.

Take, for instance, a human resource manager who is asked by a candidate about the average salary of the ﬁrm. If the CEO and CFO of the ﬁrm, who are the founders, both make $500,000 per year, and a random sample of the 20 other employees of the com- pany who make less than $50,000 per year is calculated, then the average salary of the ﬁrm (point estimate) would be around $175,000. This is a deceiving ﬁgure. Now, the human resource ofﬁcer can present the median salary, which is not subjected to outli- ers, or they can present the mean using an interval estimate.

4.2 Interval Estimation

An **interval estimate**, as opposed to point estimate, can be useful if a researcher decides that a point estimate is not representative as a population inference due to a skewed distribution or from being uncomfortable (from a risk perspective) reporting a point estimate. Examples from the latter include business managers who fear they may be judged on their performance when making a point estimate, when a range of esti- mates would be more appropriate from a statistical perspective.

**Interval estimate**

An interval estimator draws inferences from a sample about a population by esti- mating the value of an unknown popula- tion parameter using an interval rather than a single point. With an interval esti- mator, we attempt to construct an interval that contains within it the actual popula- tion parameter with a speciﬁed probabil- ity

In this case, the interval estimate would be used. The interval estimate is a range of values, within which the actual population parameter has an estimated probability of falling. A conﬁdence interval is one of the most popular interval estimation techniques. Conﬁdence intervals can be calculated for a variety of statistics, including means and proportions. Interestingly, conﬁdence intervals can even be used to conduct hypothesis testing. In addition to being a range of values rather than a single point estimate, inter- val estimators are also different from point estimators in that they reﬂect the effects of sample sizes, with larger sample sizes tightening the width of the interval by reducing the standard error. For example, suppose we are trying to estimate the income of work- ers in a factory. An interval estimate would state that the (unknown population) mean income per hour is between $31 and $48 with a probability of 0.95. (Greenﬁeld et al., 1998; Altman & Bland, 2011). The formula for conﬁdence intervals (CIs) for the mean is depicted below:

(4.1)

where

CI = confidence interval

{"mathml":"<math xmlns=\"http://www.w3.org/1998/Math/MathML\"><menclose notation=\"top\"><mi>x</mi></menclose></math>"} top enclose x = sample mean

z = confidence level value

s = sample standard deviation

n = sample size

Oftentimes, the characteristic being measured is not continuous, but is a categorical variable. Examples of this include political afﬁliation, gender, and marital status. In this instance, the objective would be to estimate a population proportion, *p*, using a sample proportion, , plus or minus a predetermined margin of error. This activity will result in a conﬁdence interval for the population proportion, *p* (Greenﬁeld et al., 1998; Altman & Bland, 2011). The formula for the CI for the proportions is

(4.2)

wobei:

= sample proportion

z\* = confidence level value from the standard normal distribution table

n = sample size

Once you calculate the z-score, it can be looked up in a z-table, which is partially depic- ted below. For access to the full table, see the literature by Farber and Larson (2017).

|  |  |
| --- | --- |
| z\*–Values for Various Conﬁdence Levels | |
| Conﬁdence level | z\*-value |
| 50% | 0 |
| 70% | .5 |
| 80% | 1.28 |
| 90% | 1.645 (by convention) |
| 95% | 1.96 |
| 98% | 2.33 |
| 99% | 2.58 |

## Zusammenfassung

A point estimate is used when searching for a single point with the most likely value of a population parameter. This can be any univariate statistic, including the mean or a proportion. However, it should be noted that the point estimate calcula- ted from a sample of the population will always vary based on random sampling error. This variation is taken into account when calculating an interval estimate, as the interval estimate (i.e., conﬁdence interval) will return a range of values, into which the population parameter is likely to fall, taking into account random sam- pling error.

It is always recommended that a point estimate is presented along with some measure of dispersion (i.e., standard deviation) so that the consumer of the statis- tic will have a gauge as to how reliable that estimate is, given the amount of varia- tion in the population or sample. An interval estimate can address this all in a sin- gle statistic, although it’s not really a single statistic, but a range of values. That said, it is certainly easier to interpret than a point estimate.

## 5. Hypothesis Testing

## **Study Goals**

Nach der Bearbeitung dieser Lektion werden Sie in der Lage sein, ...

… identify and understand the need for hypothesis testing.

… conduct and construct a hypothesis test.

… interpret a hypothesis test.

## Aus der Praxis

The auto manufacturer the Black Forest Motor (BFM) is anticipating the implementation of a new onboard computer interface system in response to customer complaints. The new system is purported to be more user friendly when drivers want to obtain informa- tion from the vehicle or interact with and adjust certain systems, including climate con- trol. The challenge is that the cost of modifying the system across all vehicles is sub- stantial.

Hannah Brekker, the chief engineer of BFM, is discussing with Gerald Meier, the chief statistician, about how to conduct a test without making a big investment in retooling the plants. The decision is to custom engineer a small subset of cars and then provide these new cars to a random sample of customers who ordered the vehicle. BFM will then monitor the customer satisfaction scores over the next 24 months to see if the new onboard computer interface yields signiﬁcantly better results. In order to deter- mine statistical signiﬁcance, Gerald Meir recommends a t-test be performed on the data. However, he needs to discuss a few things with Hannah Brekker, including the sample size; level of signiﬁcance; 24-month observation window; and logic of using cus- tomer satisfaction scores over other options, such as net promoter scores. Hannah Brekker feels a little overwhelmed learning about all these new concepts, but, at the same time, is glad that this new interface can be successfully tested without making big investments. She reﬂects on how important hypothesis testing is as a tool available to organizations like BFM.

## 5.1 Hypothesis Testing when Population Standard Deviation is Known

Whether the population standard deviation is known or unknown makes a big differ- ence when hypothesis testing. This knowledge determines the probability distribution, which is used to conduct the test. Practically speaking, we almost never know the pop- ulation standard deviation, but it makes sense to study this since it facilitates a better understanding of the concept of hypothesis testing. Therefore, we will address the his- tory, overview, and some examples of hypothesis testing.

## Origins of Hypothesis Testing

Salsburg (2001) describes a tea party at the University of Cambridge, England, at which a lady made a claim that the taste of a cup of tea poured into milk can be distin- guished from milk poured into tea. Most of the members of the party disagreed with the lady on the grounds that her conjecture lacked a scientiﬁc basis, but Ronald Fisher, another party attendee who would later become knighted and known as Sir Ronald Fisher, proposed a type of test that could satisfy both the lady and her challengers. His test included subsequent randomized presentations of cups of tea to the lady in which the milk was introduced ﬁrst and vice versa, while recording the lady’s responses for each cup. This test he proposed was a hypothesis test, and the work of Fisher and his contemporaries, including Karl Pearson and William Sealy Gossett, created a testing strategy that has been responsible for helping to usher in some of the greatest discov- eries known to humankind (Salsburg, 2001).

So why do we need this type of testing? Well, one application is the simple security that when we make a conclusion about a phenomenon, it is not occurring by chance, and is instead related to the variable in question. In the case of the tea, the variable is the timing of adding the milk. We needed such a test in this case because the lady could have simply been lucky in guessing which tea was which. Fisher’s proposed test, however, was able to measure whether the lady’s chances of getting the number of cor- rect responses was beyond normal chance occurrence (Salsburg, 2001). The key to hypothesis testing is understanding that associations between two variables may appear to be correlated, but, in reality, are sometimes the result of randomness. In the following section, you will be exposed to the various types of hypothesis tests used in business and the social sciences today. This will allow you to test your own hypotheses about the world around you to see if your conjectures are real or simply due to random variation.

## Generalized Approach to Statistical Hypothesis Testing

Statistical hypothesis testing is a form of statistical inference in which conclusions about a population are drawn on the basis of a sample obtained from that population. Hypothesis testing provides a convenient and straightforward framework and method- ology for ascertaining the reliability or strength of evidence based on insights extracted from the sample. These results are then considered in relation to the population from which the sample was drawn. However, since it is just a sample of the population, any ﬁnding derived from that sample could be biased due to random sampling error. Hypothesis tests closely follow the scientiﬁc method, in which the researcher formu- lates a speciﬁc hypothesis based on an assumption or educated guess. The researcher then draws a sample from a population, evaluates data from the sample, and uses these data to decide whether they support the speciﬁc hypothesis.

## Hypothesis Testing Fundamentals

Hypothesis testing always starts with a research question or hypothesis, which is con- verted into a hypothesis test. This research question should ideally be based on an educated guess or theory in order to prevent a false positive. Hypothesis testing is about creating and testing two hypothesis statements about a probability distribution based on data observed from a sample distribution. A step-by-step process is deployed, enabling inferences about a population parameter using a sample and tak- ing into account the expected results if the underlying research hypothesis were cor- rect (Friedman, 2018). The ﬁve core steps to conducting a hypothesis test are as follows:

1. State the null hypothesis, i.e., H0 = 0.
2. Choose the appropriate distribution.
3. Ascertain the rejection and non-rejection regions.
4. Compute the value of the test statistic.
5. Make a decision based on the data.

**Step 1: State the null hypothesis**

In this step, you will create two hypothesis statements to determine the validity of a statistical claim; these statements will contain a null hypothesis and an alternative hypothesis. The null hypothesis makes a statement claiming a zero difference, i.e., a null. The null hypothesis undergoes the testing procedure, regardless of whether it is the original claim. H0 is the statistical notation for the null hypothesis. This represents what is assumed to be true, that which is the status quo, and always contains the equal sign. H0 = 0 is the statistical language used to represent the null hypothesis (Friedman, 2018).

It now stands to reason that alternative null statement must be true if the null hypoth- esis is false. The statistical notation for the alternative hypothesis is H1. It is the direct opposite of the null and is what the researcher hopes to be true, as it represents the research hypothesis. Unlike its opposite, the null hypothesis, it never contains the equal sign. H1 ≠ 0 is the statistical language used to represent the altervative hypothe- sis (Friedman, 2018). One question which is often asked is why we test for the null rather than the alternate hypothesis, which is the one we hope to prove. One of the reasons is that, in business and the social sciences, there are few things that are able to be determined 100 percent through statistical hypothesis testing. There are simply too many factors to consider establishing true causality. This is true with even tightly controlled experiments. Therefore, we assume there are other factors causing the out- come and we ﬁnd enough evidence against this to conclude that the alternate hypoth- esis must be a probable inﬂuence. Take the negative health effects on the use of tobacco products, for instance. It’s very difﬁcult to actually prove a link between the use of these products and ill health effects. However, evidence from tests and studies that have been conducted over the previous ﬁve decades would indicate otherwise. Enough evidence has been gathered to actually change the behavior of individuals, despite the fact that causality can never be established. This is the power of hypothesis testing (Friedman, 2018).

**Step 2: Choose the appropriate distribution**

For this step, the appropriate probability distribution is determined by whether you know the population standard deviation or need to rely on the sample standard devia- tion. Where the population standard deviation is known, one would use the z-value, which represents the standard normal distribution. For hypothesis tests where only the sample standard deviation is known, one would choose the t-distribution. It should be noted that, when sample sizes are large (i.e., 30 or more observations), both distribu- tions look the same and students are able to choose either (Friedman, 2018).

**Step 3: Ascertain the rejection and non-rejection regions**

In this step the signiﬁcance level, denoted as alpha, or α, is to be calculated. The signif- icance level represents the probability of rejecting the null hypothesis when it is true. For example, a signiﬁcance level of 0.01 indicates a one percent risk of concluding that a difference exists when there is no actual difference (Friedman, 2018).

**Step 4: Compute the value of the test statistics**

The calculation for the z-test statistic when the population standard deviation is known is

(5.1)

Where:

= Sample mean

sigma = Population standard deviation

n = Sample size

The rejection and non-rejection regions are separated by the values of the test statis- tic. The rejection region is the set of values for the test statistic that leads to rejection of H0. The non-rejection region is the set of values not in the rejection region, which lead to a non-rejection of H0 (Friedman, 2018).

The p-value is another quantitative measure for reporting the result of a hypothesis test. When the p-value is low, there is a greater likelihood of obtaining the same result. Therefore, a low p-value provides statistical evidence that the results of the test are not due to random sampling error alone. A p-value is equal to the chance of obtaining a test statistic equal to or more extreme, than the observed value of H0 (Friedman, 2018).

We then compare the p-value with α, there are three possible outcomes:

1. If the p-value < α, reject H0.
2. If the p-value >= α, do not reject H0.
3. If the p-value is low, then H0 must go.

The idea of hypothesis testing is to reject the null hypothesis if the sample data do not agree with the null hypothesis. Thus, if the observed test statistic is more weighted in the direction of the alternative hypothesis than is preferred from a decision risk per- spective, then one will be justiﬁed in rejecting the null hypothesis (Friedman, 2018).

**Step 5: Make a decision**

According to the results of your hypothesis test, you can ascertain whether your study fails to reject, accepts, or rejects the null hypothesis. It is important to note that the researcher should not accept the alternate hypothesis. This, as mentioned, is due to the fact that the hypothesis testing process is meant to gather evidence for a research hypothesis, as opposed to making a deﬁnitive conclusion. We also often discover that, when the results of a hypothesis test are reported in academic journal, it is common to ﬁnd that the researcher provides only the test statistic and its p-value, leaving it up to the reader to draw their own conclusions. This is also not a preferred practice; however, since the accuracy of precise p-values may be inﬂated (Friedman, 2018).

### One-Tailed versus Two-Tailed Hypothesis Tests

There are two considerations researchers need to make when conducting a hypothesis test: whether they are conducting a one-tailed or a two-tailed test. A two-tailed test involves rejecting the null hypothesis if the t-score for the sample is of a low probabil- ity in either direction. This test is justiﬁable when it is believed that the sample mean might differ from the hypothetical population mean, but we do not have good reason to expect the difference to go in any particular direction (McEvoy, 2018). An example of this would be wanting to compare the mean strength of auto parts from an original equipment manufacturer (OEM) parts supplier to a predetermined target value. In this case, we would want to use a two-tailed test because we are trying to ascertain whether the mean exceeds or is lower than the target value (McEvoy, 2018).

Conversely, we have the one-tailed test, with which we would only reject the null hypothesis if the t-score for the sample were a low probability in one direction that is speciﬁed before collecting the data. This test makes sense when we have good reason to expect that the sample mean will differ from the hypothetical population mean in a particular direction, which could be greater than or less than the hypothetical popula- tion mean. An example of this is that employee engagement scores would increase as a result of providing new incentives (McEvoy, 2018). When you have a one-tailed test you need to determine whether it is a right-tailed test or a left-tailed test, which are some- times also referred to as upper and lower tests. When one’s hypothesis has an inequal- ity that points to the right and contains a greater than (>) symbol, then this is a right- tailed test. For example, you might be comparing the life of refrigerator compressors before and after a manufacturing change (Glen, n.d.). If you want to know if the com- pressor life is greater than the original (e.g., 10 years), your hypothesis statements might be as follows:

* null hypothesis: No change or decrease observed (H0 ≤ 10).
* alternate hypothesis: Compressor life has increased (H1 > 10).

When one’s hypothesis has an inequality that points to the left and contains a less than (<) symbol, then this is a left-tailed test. For example, you might be comparing the length of time it takes to execute digital marketing campaigns before and after an implementation of customer relationship management software (Glen, n.d.). If you want to know if the cycle time for campaign execution is less than the original (let’s say three days), your hypothesis statements might be as follows:

* null hypothesis: No change or greater than (H0 ≥ 3).
* alternate hypothesis: Campaign execution time has decreased (H1 < 3).

It is important to note that whether a test is left- or right-tailed is determined by the alternate hypothesis (H1) and not the null hypothesis (Glen, n.d.).

Hypothesis tests are one of the greatest tools developed in the twentieth century. These tests are indirectly responsible for the development of vaccines, life-saving drugs (like the cure for Hepatitis-C), and doubling the life span of individuals in less than a century. Hypothesis tests are designed to determine whether two treatments or a single treatment has a signiﬁcant effect on the population, or whether any differen- ces are due to sampling error. All hypothesis tests start with a research hypothesis for the desired test element, followed by the creation of the null hypothesis, which assumes there is no statistically signiﬁcant difference between the population mean and the sample mean. If, after calculating the difference, the probability of the differ- ence happening by chance proves to be very low, then we have enough evidence to reject the null hypothesis, therefore supporting our alternate hypothesis related to our research hypothesis. The choice of which test to use depends on whether we know the sample standard deviation (t-test) or the population standard deviation (z-test).

## 5.2 Types of Hypothesis Testing When the Population Standard Deviation is Unknown

The t-test is used when the population standard deviation is unknown. The most important type of hypothesis testing in business and the social sciences, if we are focused on the differences between two means, is the t-test. In this section, we exam- ine three types of t-tests that are associated with speciﬁc research designs: the one- sample, dependent samples, and independent samples t-tests.

### One-Sample T-Test

The **one-sample t-test** is used when researchers want to compare a mean derived from a sample (M) with a hypothetical population mean (μ0). The null hypothesis is that the mean for the population (µ) is equal to the hypothetical population mean: μ = μ0. The alternative hypothesis is that the mean for the population is different from the hypothetical population mean: μ ≠ μ0 (McEvoy, 2018). To conduct this hypothesis test, we need to ascertain the probability of obtaining the sample mean if the null hypothesis were true. To do this we must first find the *p* value by computing a test statistic called *t* (McEvoy, 2018). The formula we use for *t* is as follows:

(5.2)

In this equation, top enclose x is the sample mean and (µ0 is the hypothetical population mean of interest), S2 is the sample standard deviation, and n is the sample size. If we calculate our t- score using a computer, we will receive both the t-score and the *p*-value as outputs. At this point, we would implement the ﬁnal step in the hypothesis testing process, i.e., make a decision based on the data. Assuming we decide to test the hypothesis at a 95 percent conﬁdence level, if *p* is less than .05, we reject the null hypothesis and con- clude that there is evidence that the population mean differs from the hypothetical mean. If *p* is greater than .05, we fail to reject the null hypothesis and conclude that there is not enough evidence to say that the population mean differs from the hypo- thetical mean (McEvoy, 2018).

If we want to calculate the t-score manually, we could use a table of **critical values** of t when alpha equals .05 to make our decision. Interestingly, these tables do not provide actual *p* values; instead, they provide the critical values of t for different degrees of freedom (df) when a is .05. There are also tables for different alpha levels, which vary depending on whether the test is one- or two-tailed. When implementing the decision making step, we consider any t-score beyond the critical value in either direction (for two-tailed tests only) to be in the most extreme ﬁve percent of t-scores when the null hypothesis is true and has a < value less than .05. Thus, if the t-score we compute is beyond the critical value in either direction, then we have enough evidence to reject the null hypothesis. However, if the t-score we compute is between the upper and lower critical values, then we fail to reject the null hypothesis (McEvoy, 2018).

**Critical values**

These are derived from a graph of a distribution that divides the graph into a reject region. If the t exceeds that value and enters the region, then the null hypothesis should be rejected.

### The Dependent Sample T-Test

The dependent sample t-test is often referred to as the paired t-test. This test is used to compare two means for the same sample tested at two different times or under two different conditions. An example of this in business is the comparison between a pre- test and posttest measuring knowledge of corporate ethics after employees take a new corporate ethics training module. The null hypothesis is that the means calculated from sample data at the two points in time (or under the two conditions) are the same in the population. The alternative hypothesis is that they are different. If there is good reason to expect that the difference will go in a particular direction (like in the previous example), then this test can be one-tailed. This test is appropriate for related samples since the same variation in both observations can be taken into account to make the test more powerful (McEvoy, 2018).

One can frame their interpretation of the dependent sample t-test as a particular type of one-sample t-test after the preparation process is completed. This process speciﬁ- cally involves subtracting the two scores for each participant to create a single, differ- ence score. With this ﬁrst step completed, the dependent sample t-test becomes a one- sample t-test using these difference scores. The hypothetical population mean ((0) is 0 because this is what the mean difference score would be if there were no difference (on average) between the two observations or two conditions of the same population. The null hypothesis can then be stated as the mean difference score in the population is 0 ((0 = 0), and the alternative hypothesis is stated as the mean difference score in the population is not 0 ((0 ≠ 0) (McEvoy, 2018).

### The Independent Sample T-Test

The third hypothesis testing method is the independent sample t-test, which is used to compare the means of two separate samples (Ma and Mb). The two samples might have been acquired via testing under different conditions in a between-subjects experiment, or they could be preexisting groups in a correlational study (e.g., married and single, or male and female). An example application of an independent sample t- test in business is if a random sample of customers were either handled through a live agent or a chatbot and we were to test whether there was a statistically signiﬁcant dif- ference in their mean satisfaction scores (McEvoy, 2018). The important thing to note about these samples is that they are chosen independently from one another. The null hypothesis is that the means of the two populations are the same ((1 = (2). The alter- native hypothesis is that they are not the same ((1 ≠ (2). This test should be designa- ted as one-tailed if the researcher has good reason to expect the difference in means to be either higher or lower (McEvoy, 2018). The 1 statistic in this test is considerably more complex due to the fact that were are now taking into account two sample sizes, two sample means, and two standard deviations (McEvoy, 2018). The formula is as fol- lows:

(5.3)

where

* mA and mB represent the mean value of the group A and B, respectively.
* nA and nB represent the sizes of the group A and B, respectively.
* s2 is an estimator of the pooled variance of the two groups.

It can be calculated as follows:

(5.4)

with degrees of freedom df=nA+nB−2.

Please note that the formula above contains the variances, which are the squared standard deviations contained inside the square root symbol. Also, lowercase n/ and n3 refer to the sample sizes in the two groups, which is the opposite of the symbol *N*, which, in statistical notation, generally refers to the total sample size. One ﬁnal thing to note is that there are N − 2 degrees of freedom for the independent sample t-test.

### Confidence Intervals for Hypothesis Testing

Another method of conducting hypothesis testing is through interval estimation, specif- ically conﬁdence intervals. Curran-Everett (2009) discusses the fact that, unlike hypoth- esis tests, conﬁdence intervals are a fairly recent phenomenon, developed by Jerzy Neyman in the 1930s. Both the hypothesis test and conﬁdence intervals can accomplish the same thing, but are often used for different purposes. For example, many research- ers will use hypothesis testing when they have a pre-speciﬁed hypothesis and signiﬁ- cance level and want to do a strict comparison. Conversely, when they want to describe the magnitude of an effect (e.g., mean difference or odds ratio), or a single sample, a conﬁdence interval may be more useful (Curran-Everett, 2009).

Which test to use depends on whether we know the sample standard deviation (t-test) or the population standard deviation (z-test). There are also different types of t-tests, which are chosen based on whether there is a single sample or two samples; for the latter, we must also determine whether the samples are related or independent. One ﬁnal note is that hypothesis testing is not just for means, but can also be used to test other statistics, including proportions.

### Zusammenfassung

Hypothesis tests are indirectly responsible for the development of vaccines, life- saving drugs, and doubling the life span of individuals in less than a century.Hypothesis tests are one of the greatest tools developed in the twentieth century. The premise of the hypothesis test is simple; they determine whether two treat- ments or a single treatment has a signiﬁcant effect over the population, or whether any differences are due to sampling error. All hypothesis tests start with a research hypothesis, followed by the creation of the null hypothesis, which assumes there is no statistically signiﬁcant difference between the population mean and the sample mean. If, after calculating the difference, the probability of the difference happen- ing by chance proves to be very low, then we then have enough evidence to reject the null hypothesis, which then supports our alternate hypothesis related to our research hypothesis.

Which test to use depends on whether we know the sample standard deviation (t- test) or the population standard deviation (z-test). This distinction becomes less important with larger sample sizes. There are also different types of t-tests, which are chosen based on whether there is a single sample or two samples. Many researchers will use traditional hypothesis testing when they have a pre-speciﬁed hypothesis and signiﬁcance level and want to do a strict comparison. Hypothesis testing is not just for means, but can also be used to test other statistics, including proportions when the data are normal, and statistics, such as medians with non- parametric tests in which the data are not normally distributed.

## 6. Simple Linear Regression

## **Study Goals**

Nach der Bearbeitung dieser Lektion werden Sie in der Lage sein, ...

… identify and understand the applications for practicing simple linear regression analysis.

… design and compute a simple linear regression analysis.

… interpret and validate a simple linear regression model.

## Aus der Praxis

The supermarket chain, Anderson’s, wants to ascertain whether advertising expendi- tures for targeted campaigns over the previous three years have had an impact on sales. The chief marketing ofﬁcer Alice Taylor elicits the assistance of Helen Weber, the chief digital ofﬁcer, to determine whether Anderson’s is able to attribute their ad spend to increased sales. Helen explains to Alice that it is indeed possible, and actually quite common for ﬁrms who spend a good portion of their budgets on advertising, to be able to measure the results of the campaigns. She goes on to explain that a popular method of measure in the efﬁcacy of the advertising effort of ﬁrms is to build attribution mod- els using simple linear regression models.

As a next step, Helen agrees to collect data related to the previous 36 months of adver- tising spend and company sales. The advertising spend serves as the independent vari- able and the gross sales serves as the dependent variable in this study. Helen plans to run the regression analysis in Excel and then share the results with Alice and the branding and advertising team. Alice anxiously awaits the results and hopes that there will be a strong positive relationship between their advertising efforts and sales so she can present the results to the executive team to gain additional funds to help grow the business.

## 6.1 Simple Linear Regression—Concepts, Approach, and Quality Assessment

Karl Pearson developed a mathematical rigor around the Pearson product moment cor- relation, and was inspired by Sir Francis Galton, the accomplished nineteenth century British scientist, who was originally thought to have conceived both correlation and regression analysis (Stanton, 2001). In an era of science focused on proving or disprov- ing the theory of evolution, Galton’s study of genetics sparked the idea that led to the development of linear regression.

In a famous experiment with sweet peas, Galton gained the idea for regression by plot- ting a scatter diagram of the sizes of daughter peas in relation to the sizes of their mother (sweet peas are self-fertilizing). While providing evidence for heredity, it also gave Galton the idea that two variables which were related (i.e., the size of mother and daughter sweet peas) could also be predicted. The generalized form of regression anal- ysis was developed later by numerous data scientists (Stanton, 2001).

Other contributors to regression analysis include Carl Friedrich Gauss and Adrien-Marie Legendre, scientists who both independently discovered an essential feature of regres- sion analysis: the method of least squares. Least squares is a statistical optimization technique in which the sum of the squared errors are minimized. Both of these scien- tists used this method to understand the orbits of celestial bodies (Stigler, 1981). While the applications of linear regression increased dramatically over the next 150 years, very little advancement to the technique was made, primarily due to the lack of high- speed computing (Hocking, 1983). More recent developments in the techniques of regression analysis, aided by increased computing power, include contributions from statisticians John Tukey, George Box, David Cox, Sanford Weisberg, and Maurice Kendall, and, even more recently, David Freedman and Andrew Gelman (Hocking, 1983).

## Generalized Approach to Regression

Almost all statistical programming languages from, R to SAS and even spreadsheet pro- grams such as Excel, make the calculation of regression analysis seem relatively easy while having linear algebra operating as their hidden engine powering the regression equations. We will brieﬂy discuss linear algebra, as it is necessary for an understanding of the linear algebraic origins of **linear regression**.

In its simplest form, appropriately referred to as simple linear regression, our goal is to ﬁnd the best ﬁt line (from which we can then predict with new observations) through a set of data points: (x1, y1), (x2, y2), … (xn, yn) (Sundaram, 2020). For example, this could be used to try to predict employee satisfaction scores as a function of average employee tenure. But the real question to ask is what does the best ﬁt mean? Keeping in mind that the equation for a straight line is y = mx + b, where y is our dependent variable, x is our independent variable, m is the slope of the line, and b (c in the UK) is the constant and also referred to as the y-intercept. Remember, the constant is y when x = 0. If we can ﬁnd a slope and an intercept for a single line that passes through all the possible data points, then it would certainly be considered the best ﬁt line, simply because it is a perfect ﬁt. In terms of regression, there would be no errors (Sundaram, 2020).

**Linear regression**

This describes a ser- ies of modeling techniques, which can have different error distributions, but all relate in some way to simple linear regression. They are known as general linear mod- els (GLM).

However, in the majority of cases, such a perfect line does not exist. Therefore, we default to searching for another line that measures the error of each data point when a connecting line is drawn parallel to the y-axis, from the data points to the regression line. This should ensure that the sum of all such errors should be minimized. Though it sounds quite simple, we deploy matrix and linear algebra behind the scenes to make this happen (Sundaram, 2020).

Regression, paired with hypothesis testing, are the key value drivers of statistical analy- sis. However, despite all of its popular applications, from credit scoring to sales fore- casting, regression should not be considered a single entity. Rather, regression is a vari- ety of statistical methods, all with one fundamental idea in common:

*Dependent Variable (y) = Constant (b) + Slope (m)\*Independent Variable (x) + Error (e)*

The dependent variable in the simple linear regression equation, also known as the outcome variable, is something the researcher will want to predict or explain. For example, in a customer management context we may want to ﬁnd a customer satisfac- tion score measured on a zero to ten rating scale. The independent variable in a simple linear regression equation, which is also known as the predictor variable, is what is used to explain or predict the dependent variable. Continuing with examples of cus- tomer management, we could assess the length of a customer service call in minutes (Gray & Gray, 2017). The constant term in the equation above as previously stated is also referred to as the Y intercept. The y represents the dependent or the outcome var- iable, and x is the independent variable. The slope shows amount Y changes when X changes by a speciﬁc amount (Gray & Gray, 2017).

The error term is a critical one and constantly reminds us that it's virtually impossible to predict y from x with 100 percent certainty. Sometimes no errors occur in a model, however, this should raise concerns that there may be something wrong with the model. Analysts should try to continually reduce the errors and improve the model (Gray & Gray, 2017).

An introduction of simple linear regression is necessary to be able to conceptualize the process. However, in many business applications, we have more than one independent variable and, in this case, we can utilize a more advanced from of regression analysis known as **multiple regression** (Gray & Gray, 2017).

**Multiple regression**

Solving multiple regression problems by hand requires one to think of the points on an n- dimensional surface in space, with n rep- resenting the num- ber of independent variables. The shapeof this surface depends on the model structure.

One of the key assumptions of simple linear regression is a linear relationship between the two variables. However, some business applications deal with nonlinear relation- ships. In these cases, similar to multiple regression, the regression class of algorithms also covers nonlinear relationships with more advanced models like polynomial regres- sion and Gompertz regression. Even taking these special algorithms into account, many linear regression models are, in fact, linear only after transformations of the variables have been made (Gray & Gray, 2017).

Some of the other considerations related to regression analysis are related to the type of dependent variable we are working with. Simple linear regression assumes a contin- uous dependent variable at the interval or ratio level of measurement. Sometimes, researchers want to predict an outcome which is dichotomous or binary. Loan default on a credit score is an example of such a variable, since it is measuring whether a cus- tomer repays their loan or does not. Because a loan default variable is a binary varia- ble and not a continuous numeric, using a simple linear regression that assumes a numeric dependent variable would not be an acceptable statistic. Logistic or probit regression in this case would be the correct choice (Gray & Gray, 2017). Moreover, if the dependent variable is categorical but has three or more categories, multinomial logistic regression (based on a multinomial distribution) is used. An application of regression where there are often three or more categories to be modeled is in conjoint analysis, in which respondents typically choose from three or more products in each choice task (Gray & Gray, 2017).

Sometimes, we have a categorical dependent variable in which the categories are ordered, e.g., light, medium, and dark colors. In this case one would choose ordinal logistic or probit regression (Gray & Gray, 2017). Finally, when researchers have a dependent variable, which is count data (i.e., how often a customer visited the store in one year), Poisson and negative binomial regression are two of the best choices for modeling said data (Gray & Gray, 2017). These examples are just some of the most pop- ular regression techniques. Others include quantile regression, box-cox regression, truncated and censored regression, hurdle regression, and nonparametric regression, as well as regression methods for time series data (Gray & Gray, 2017).

Developing a Linear Regression Model

As mentioned, linear regression is a statistical tool used to measure and predict the relationship between two variables: one independent (the explanatory variable) and one dependent (the predictor variable). This is done by ﬁtting a linear equation to observed data. For example, a business analyst might want to model and relate the annual purchases of customers to their income levels using a simple linear regression model (Yale University, 1997). Before even attempting to ﬁt a simple linear regression model to the data that have been collected, an analyst should ﬁrst look for some type of evidence to determine whether or not there is a relationship between the variables of interest. They should always start with a strong hunch or research hypothesis that it would be logical for a relationship to exist between two variables (Yale University, 1997).

When trying to ascertain the relationship between variables in advance of conducting a regression analysis, it’s critically important to recognize that a correlation between two variables does not necessarily assume or imply that one variable causes the other. For example, higher scores on a pre-employment test do not necessarily cause better employee performance. However, this does not take away from the fact that there is some signiﬁcant association between the two variables.

Dr. William Edwards Deming, the famous statistician and engineer, and one of the fathers of modern quality, often shared the idea that graphs are friends that help us to understand data. On that same note, a scatterplot of scatter diagram should be one of the ﬁrst tools used to ascertain the strength of the relationship between two variables. If there appears to be a lack of association (randomness) between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing linear trends by plotting the x and y variables), then ﬁtting a linear regression model to the data will probably not be a productive use of time.

In addition to creating and analyzing a scatter diagram, a researcher at this stage of the modeling process can also calculate a Pearson product moment correlation coefﬁcient, which is a numerical modeling of the association between two variables. The output of this model yields what is known as the correlation coefﬁcient, which is a value between-1 and 1, indicating the strength and direction of the association of the observed data for the two variables. A positive coefﬁcient indicates that, as the x variable increases, so does the y variable. A negative correlation indicates that as the x variable in increa- ses, the y variable, in turn, decreases (Yale University, 1997). The form of the simple lin- ear regression equation is B = 4 + CD, where D represents the explanatory variable and B represents the dependent variable. The slope of the line is represented by C, and the intercept is represented by a (remember a represents the value of y when x = 0) (Yale University, 1997).

Quality Assessment for Simple Linear Regression

There are several ways to assess the quality of a regression equation. The ﬁrst is a numeric method, and the second is a graphical method, which involves plotting the actual versus predicted values from the regression analysis. The numeric method involves the Pearson’s correlation coefﬁcient, which is a statistical tool that ranges from 0 to +1 and -1 and describes a linear relationship between two variables. In the case of simple linear regression, we use the coefﬁcient of determination r2, which is a measure of how well the regression model describes the observed data. In simple lin- ear regression analysis, r2 is simply the Pearson’s correlation coefﬁcient squared. The interpretation of this statistic is that it represents the percentage of variance in the dependent variable by knowing the independent variable. Of course, the higher the percentage, the better, as long as the model does not overﬁt. Overﬁtting is when a model is only performing on the sample from which it was developed, and cannot extrapolate consistently and effectively when new samples are introduced (Schneider et al., 2010).

One can also test the regression coefﬁcients for statistical signiﬁcance. The null hypothesis for a simple linear regression is b = 0, meaning there is no relationship between variables, and the regression coefﬁcient is therefore zero. This can then be tested with a t-test. A 95 percent conﬁdence interval for the regression coefﬁcient can also be computed (Schneider et al., 2010).

Regression analysis in both the simple linear regression form and its more sophistica- ted nonlinear and multiple regression counterparts are considered one of the most important statistical procedures in business and our personal lives today. Not only can this tool help researchers to ascertain whether or not on or more variables have a stat- istical relationship, but it can also be used to predict the future when new observations are made in the independent variable. This module introduced the concept of simple linear regression so that students can conceptualize and easily learn the process. The equation for simple linear regression follows the form B = 4 + CD, where D represents the explanatory variable and B represents the dependent variable. The slope of the line is represented by C, and the intercept is represented by 4 (Yale University, 1997). The idea is to used the least-squares algorithm to create a line across all points between the two variables which minimizes the sum of the squared errors. In the case if simple linear regression we use the coefﬁcient of determination r2, which is a meas- ure of how well the regression model describes the observed data to assess the quality of our regression analysis. In simple linear regression analysis, r2 is simply the Pear- son’s correlation coefﬁcient squared. The interpretation of this statistic is that it repre- sents the percentage of variance in the dependent variable there is by knowing the independent variable

6.2 Applications of Simple Least-Squares Regression

The least-squares is the most common method for fitting a regression. This method minimizes the sum of the squared errors. What this means is that a best-fitting line is created for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (the vertical deviation is zero if a point lies on the fitted line exactly). The squaring operation means the deviations are first squared, then summed, the positive and negative values do not cancel each other out (Yale University, 1997).

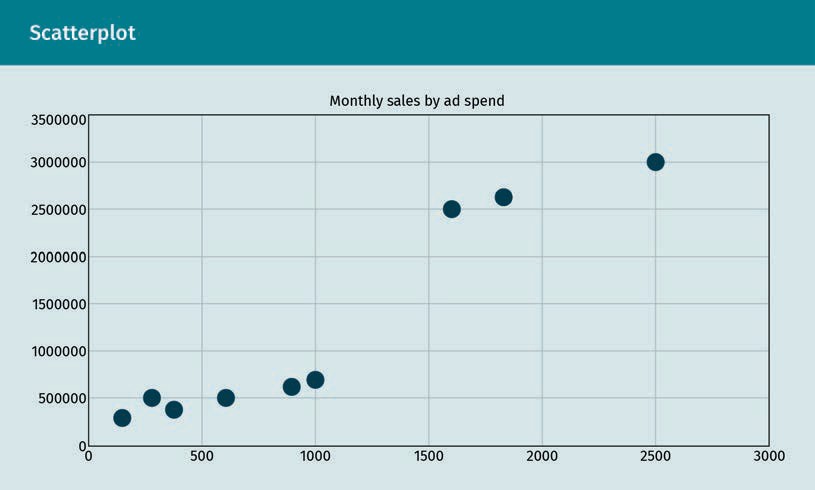
Example

This example will teach students how to run a simple linear regression analy- sis in Excel and how to interpret the summary output. Below are the sample data. The research question is as follows: Is there a relation between monthly sales (dependent variable) and monthly advertising spend (independent variable). Moreover, can we pre- dict sales if we know the advertising spend?

|  |  |
| --- | --- |
| Sample Advertising Data | |
| Monthly sales | Monthly ad spend |
| 3,000,000 | 2,500 |
| 500,000 | 300 |
| 300,000 | 200 |
| 400,000 | 400 |
| 500,000 | 600 |
| 2,000,000 | 1,600 |
| 2,100,000 | 1,800 |
| 600,000 | 900 |
| 700,000 | 1,000 |

### Scatterplot

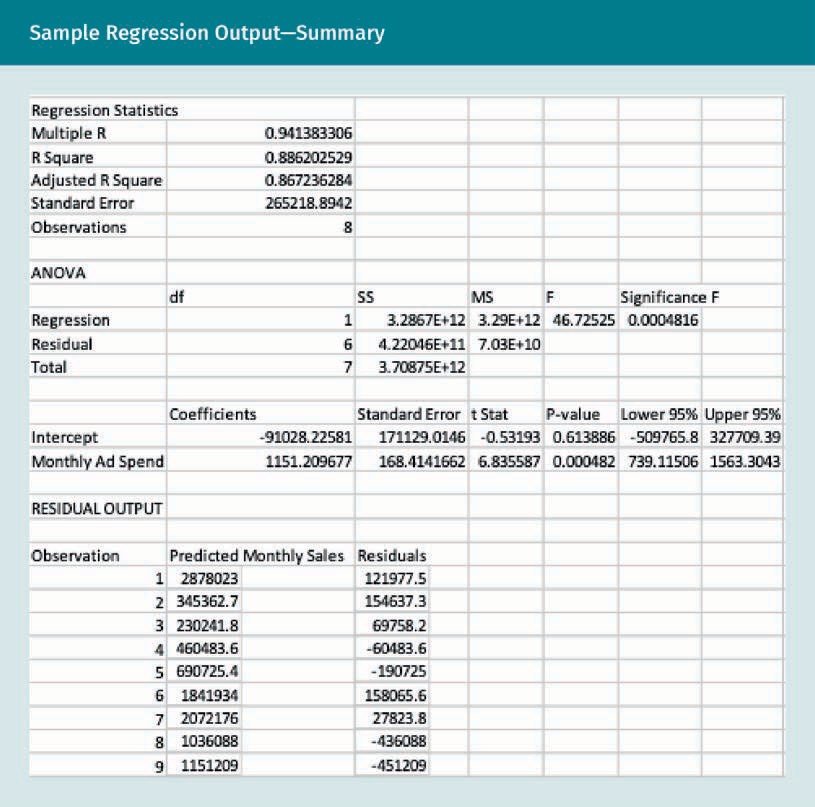
Before we begin the process of building a simple linear regression model, it makes sense to ﬁrst create a scatterplot to assess the extent of the linear relationship between the two variables (Excel Easy, n.d.). An example of such a scatterplot is shown below.



From the scatterplot above, we can see a fairly strong positive correlation between the two variables, which infers that, as advertising spend is increased, we will also observe an increase in monthly sales (Excel Easy, n.d.).

### Simple Linear Regression

Next, we run a regression analysis on this dataset in Excel using the Analysis ToolPak add-in to the Data Analysis Tab. An example of how the total output would be repre- sented in Excel is given in the ﬁgure below.



R-Square

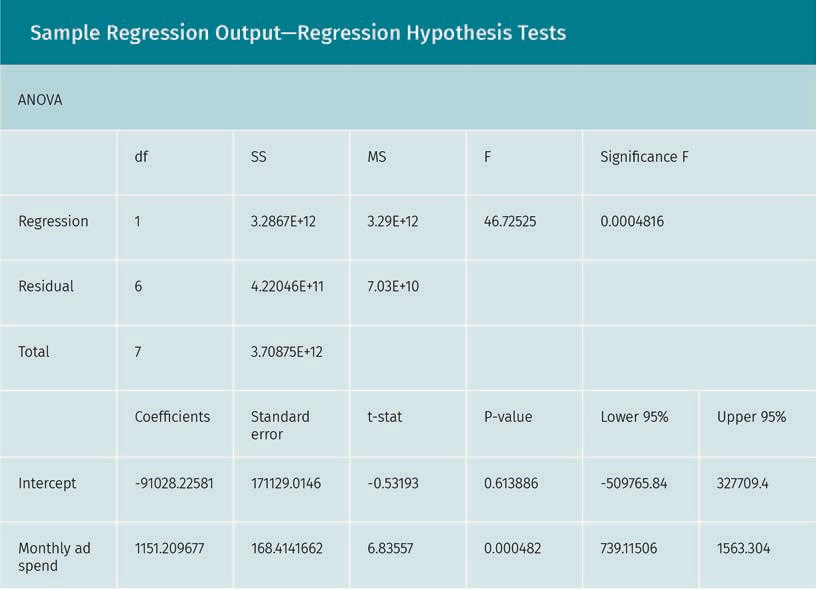
The ﬁrst thing we check on this analysis is the r-square in the summary output. R- square is the correlation coefﬁcient squared. The r-square, or 2², value in this regres- sion analysis is 0.886 (the square of the correlation coefﬁcient 0.941), indicating that 88.6 percent of the variation in one variable (i.e., monthly sales) may be explained by the other (i.e., advertising spend) (Excel Easy, n.d.).

|  |  |
| --- | --- |
| Sample Regression Output: Regression Statistics | |
| Summary output | |
| Multiple R | 0.941383306 |
| R-square | 0.886202529 |
| Adjusted R-square | 0.867236284 |
| Standard error | 265218.8942 |
| Observations | 8 |

### Significance F and P-values

Next, w need to ascertain whether the results are reliable and statistically signiﬁcant. Take a look at signiﬁcance F (0.000481). If this value is less than 0.05, then you can be generally conﬁdent in your analysis. If you discover that signiﬁcance F is greater than 0.05, you probably need to search for a new independent variable and rerun your regression analysis. This should be repeated until signiﬁcance F drops below 0.05.

Most, or all, p-values should be below 0.05. In our example, this is the case with monthly ad spend (0.000482). Our intercept is not statistically signiﬁcant because there isn’t sufﬁcient statistical evidence that it is different from zero (0.61388). However, it seems reasonable that, without advertising, there would be no sales, so we will keep this regression analysis (Excel Easy, n.d.).



### Coefficients

The regression line is y = monthly sales = -91028.225 + 1151.209 · monthly ad spend. In other words, for each unit increase in ad spend, sales increases by 1,151.209. This is val- uable information. You can now also create forecasts with this simple linear regression model. For example, if monthly ad spend equals 2657, you might be able to achieve a quantity sold of 0 + 1151.209 · 2657 = 3,058,762.31 (Excel Easy, n.d.).

### Residuals

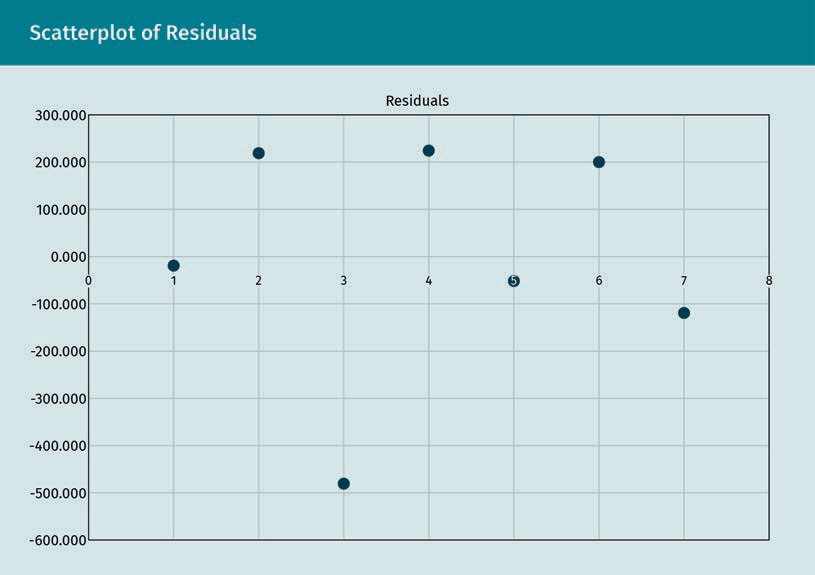
An analysis of the residuals can help us ascertain how far away the predicted data points are from the actual data points by using the given regression equation. For illus- trative purposes, the ﬁrst data point in our advertising spend example equals 2500. Using the equation, the predicted data point equals 0 + 1151.209 · 2500 = 2,878,023, giving a residual of 3,000,000 - 2,878,023 = 121,977 (Excel Easy, n.d.).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observation | Predicted monthly sales | | Residuals | |
| 1 | 2878023 |  | 121977.5 |  |
| 2 | 345362.7 |  | 154637.3 |  |
| 3 | 230241.8 |  | 69758.2 |  |
| 4 | 460483.6 |  | -60483.6 |  |
| 5 | 690725.4 |  | -190725 |  |
| 6 | 1841934 |  | 158065.6 |  |
| 7 | 2072176 |  | 27823.8 |  |
| 8 | 1036088 |  | -436088 |  |
| 9 | 1151209 |  | -451209 |  |

You can also create a scatter plot of these **residuals**.

**Residual**

When a residual plot shows a random pat- tern, it suggests a good ﬁt for a linear model. If the pat- terns in the residual plot are non-random (i.e., U-shaped and inverted U-shaped), this suggests a bet- ter ﬁt for a nonlinear model.



Outliers and Influential Observations

A point that lies far from a regression line that has been computed for a group of data, such as in the group shown above, is often referred to as an outlier or extreme value. It is important that any outlier be addressed, since these may result in a ill-ﬁtting regres- sion. The plot of the regression line for the advertising sample (created by highlighting the data and creating a scatterplot with a linear trendline option in Excel) is shown below.



Please note that points that lie far from the other data in the horizontal direction are referred to as inﬂuential observations. We make the distinction between inﬂuential observations and outliers because the former may have a signiﬁcant impact on the slope of the regression line. In the case of the data in this application, most of the observations tend to be close to the prediction, so we do not see any evidence of outli- ers or inﬂuential observations being present (Yale University, 1997).

Lurking or Confounding Variables

Whenever the relationship between two variables is signiﬁcantly affected by the pres- ence of a third variable that has not been included in the modeling effort, a lurking or confounding variable is present. The presence of lurking variables may manifest in the creation of nonlinear trends, which are visible in the relationship between an explana- tory and dependent variable (Yale University, 1997).

Extrapolation

The actual range of the data should always be carefully considered whenever a linear regression model is ﬁt to a dataset. It is inappropriate to attempt to use a regression equation to predict values outside of this range, as it may yield unreliable answers. This practice is known as extrapolation. One example of this is a linear model that relates employee health risk scores to age for younger employees in their twenties and thir- ties. Applying such a model to middle-aged employees in their forties and sixties would not be appropriate, since the relationship between age and employee health risk scores is not consistent for all age groups (Yale University, 1997).

### Simple Linear Regression Assumptions

When conducting a linear regression, certain assumptions must be upheld. Each of these assumptions may have higher or lower importance depending on the data and application. There are three key assumptions (Yale University, 1997):

1. Homogeneity of variance (homoscedasticity). With homoscedasticity, we assume that the size of the error in our prediction doesn’t vary signiﬁcantly across the values of the independent variable.
2. Independence of observations. With independent observations, we assume that there are no hidden relationships among the observations and that all of the obser- vations in the dataset were gathered using statistically valid sampling methods.
3. Normality. Normality assumes that the data are following a pattern that closely approximates the normal distribution.

Linear regression makes the additional assumption that the relationship between the dependent and independent variable is linear. The line of best ﬁt through the data points is a straight line, rather than a curve, which makes the relationship nonlinear or curvilinear. Please note that if your data do not meet the assumptions one through three above, you may be able to use a nonparametric regression, such as kernel analy- sis or regression trees. For violation of the assumption of linear regression, the nonlin- ear regression methods, including polynomial and Gompertz regression, may be poten- tial solutions (Yale University, 1997).

The unique property created by regression analysis lends it to useful applications including, but not limited to, marketing and operations forecasting, credit scoring, investment portfolio allocation, insurance underwriting, and pricing optimization. For simple linear regression, it is recommended to use the least squares algorithm, which is the most common method for ﬁtting a regression equation. This method minimizes the sum of the squared errors. This means that a best-ﬁtting line is created for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (the vertical deviation is zero if a point lies directly on the ﬁtted line). The least squares algorithm was used to create a line across all points between the two variables, which minimizes the sum of the squared errors for the mar- keting application we developed. Researchers must remember to always follow several key assumptions when conducting a regression analysis.

### Zusammenfassung

Regression analysis is helpful to researchers seeking to ascertain whether one or more variables have a statistical relationship, as well as to predict the outcomes when new observations are made about the independent variable. For these rea- sons, this tool is considered one of the most important statistical procedures in business and our personal lives today. The equation for simple linear regression follows the form y = a + bX, where x represents the explanatory variable and y represents the dependent variable. The slope of the line is represented by b, and the intercept is represented by a (Yale University, 1997).

The least squares algorithm is an optimization procedure utilized to create a line across all points between the two variables, which minimizes the sum of the squared errors. We then use the coefﬁcient of determination r2 to assess the qual- ity of our regression analysis. We interpret r2 as the percentage of variance in the dependent variable, knowing the independent variable.

This unit demonstrated how simple linear regression works in a marketing applica- tion, but additional applications of regression analysis exist, including other mar- keting applications, sales and operations forecasting, pricing optimization, credit scoring, investment portfolio allocation, and insurance underwriting. An analysis of the residuals should always be considered when conducting a regression analysis in order to help ascertain, by using the regression equation above, how far away the predicted data points are from the actual data points. Researchers must be aware of the presence of inﬂuential observations and outliers, because these can both have an impact on the regression analysis. The presence of lurking variables may also create unwanted nonlinear trends when conducting a regression analysis. Finally, there are four key assumptions that researchers must remember to follow when conducting a regression analysis. If these assumptions are violated, it could lead to biased or unreliable results when interpreting the regression analysis.

# Appendix 1 – References

Adhikari, A., & Pitman, J. (2020). *Probability for data science*. Prob 140. <http://prob140.org/fa18/textbook/chapters/Chapter_24/01_Bivariate_Normal_Distribution>

Ahmad, F. (2016). *Basic statistical analysis for research: Step by step analysis*. Penerbit Universiti Tun Hussein Onn Malaysia (UTHM).

Akoglu, H. (2018). User’s guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91—93. <https://doi-org.pxz.iubh.de:8443/10.1016/j.tjem.2018.08.001>

Altman D., G, & Bland J., M. (2011). Statistics notes: How to obtain the confidence interval from a P value. *BMJ: British Medical Journal*, *343*(7825), 681. <https://doi.org/10.1136/bmj.d2090>

Amadeo, K. (2021). What are the odds of winning the lottery? How to improve your luck. *The Balance*. https://[www.thebalance.com/what-are-the-odds-of-winning-the-lot-](http://www.thebalance.com/what-are-the-odds-of-winning-the-lot-) tery-3306232

Attenborough, M. (2003), *Mathematics for electrical engineering and computing.* Elsevier.

Bandhari, P. (2020, November 5). The standard normal distribution. *Scribbr*. <https://www.scribbr.com/statistics/standard-normal-distribution/>

Britannica. (n.d.). *Dice.* https://[www.britannica.com/topic/dice](http://www.britannica.com/topic/dice)

Byjus Classes. (2021). *Probability and statistics symbols*. [<https://byjus.com/maths/probability-and-statistics-symbols/>](https://byjus.com/maths/probability-and-statistics-symbols/)

Center for Innovation in Teaching and Learning (CITL). (n.d.). *Contents of nomenclature for the fourfold (2x2) contingency table.* Central Illinois University. <https://cdn.citl.illinois.edu/courses/KINES401/Lesson4_lectures/Lecture4_P4/web_data/file2contents.htm>

Crooks, G. (2019). *Field guide to continuous probability distributions*. Berkeley Institute for Theoretical Science.

Curran-Everett, D. (2009). Explorations in statistics: Confidence intervals. *Advances in Physiology Education*, 33(2), 87—90.

Di Paola, G., Bertani, A., De Monte, L., & Tuzzolino, F. (2018, February). A brief introduction to probability. *Journal of Thoracic Disease*, *10*(2), 1129—1132. <https://jtd.amegroups.com/article/view/18920/15055>

Doane, D. P. (2016). *Applied statistics in business and economics* (5th ed.). McGraw-Hill.

Excel Easy (n.d.) *Regression*. <https://www.excel-easy.com/examples/regression.html>

Farber, B., & Larson, R. (2017, November 8). Z-0 normal distribution table. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:TABELA_DE_DISTRIBUI%C3%87%C3%83O_NORMAL_DE_Z-0.png>

Fogarty, D. J. (2017). *From beer to derivatives: The history of analytics in business*.Sara Book.

Friedman, A. (2018). *Statistics for LIS with open-source R*. Rowman & Littlefield.

Frost, J. (2019). *Introduction to Statistics: An intuitive guide for analyzing data and unlocking discoveries*. Statistics by Jim Publishing.

Glen, S. (n.d.). One tailed test or two in hypothesis testing; One tailed distribution area. *Statistics How To*. <https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/one-tailed-test-or-two/>

Gray, K., & Gray, C. (2017). *Regression analysis: A primer*. KDnuggets. <https://www.kdnuggets.com/2017/02/regression-analysis-primer.html>

Greenfield, M. L., Kuhn, J. E., & Wojtys, E. M. (1998). A statistics primer: Confidence intervals. *The American Journal of Sports Medicine*, *26*(1), 145—149.

Hayes, D. (2011, December 8). Lab2 histogram of MPG1000. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Lab2_Histogram_of_MPG1000.gif>

Hocking, R. (1983). Developments in linear regression methodology: 1959—1982. *Technometrics*, *25*(3), 219—230. <https://doi.org/10.2307/1268603>

Knapp, T. R. (1980). *Regression toward the mean. Statistics. [and] Basic descriptive statistics. Descriptive statistics. [and] Approximations in probability calculations. Applications of statistics. Modules and monographs in undergraduate mathematics and its applications project. UMAP Units 406, 426, 443*. Education Development Center. <https://files.eric.ed.gov/fulltext/ED218128.pdf>

Laerd Statistics. (2019, March 3). Pearson correlation coefficient and associated scatterplots. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Pearson_Correlation_Coefficient_and_associated_scatterplots.png>

[Lehmann, E. L.](https://en.wikipedia.org/wiki/Erich_Leo_Lehmann), & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer.

Macfarlane, G. (2016, February 4). Population time series graph. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Population_time_series_graph.jpg>

Magiya, J. (2019). Kendall rank correlation explained. Towards Data Science. *Medium*. <https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535>

McEvoy, D. M. (2018). *A Guide to Business Statistics*. Wiley & Sons.

Microsoft. (n.d.). *Statistical functions (reference).* <https://support.microsoft.com/en-us/office/statistical-functions-reference-624dac86-a375-4435-bc25-76d659719ffd>

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Freeman and Co.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: Part 14 of a series on evaluation of scientific publications. *Deutsches Arzteblatt International, 107*(4), 776—782. <https://www.aerzteblatt.de/int/archive/article/79009>

Shakil, M., Singh, J. N., & Golam Kibria, B. M. (2010). On a family of product distributions based on the Whittaker functions and generalized Pearson differential equation. *Pakistan Journal of Statistics* 26(1).

Skbkekas. (2010). Student t pdf. *Wikimedia Commons.* <https://commons.wikimedia.org/wiki/File:Student_t_pdf.svg>

Stanton, J. M. (2001). Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, 9(3). <https://doi.org/10.1080/10691898.2001.11910537>

Stigler, S. (1981). Gauss and the invention of least squares. *Annals of Statistics*, 9(3), 465—474.

Sundaram, G. (2020). *Building Linear Regression (Least Squares) with Linear Algebra* Towards Data Science. <https://towardsdatascience.com/building-linear-regression-least-squares-with-linear-algebra-2adf071dd5dd>

Tattar, N. P., Ramaiah, S., &Manjunath, B. G. (2016). *A Course in Statistics with R*. Wiley & Sons.

Toews, M. W. (2007, April 7). Standard deviation diagram. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg>

Toews, M. W. (2007, July 9). Kernel density. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Kernel_density.svg>

Viti, A., Terzi, A., & Bertolaccini, L. (2015). A practical overview on probability distributions. *Journal of Thoracic Disease*, *7*(3), E7–E10. <https://doi.org/10.3978/j.issn.2072-1439.2015.01.37>

Viniciuslima94. (2017, June 29). Fig2 tipaper. *Wikimedia Commons*. <https://commons.wikimedia.org/wiki/File:Fig2_tipaper.png>

Virginia Tech. (1999). *The t-distribution and its use in hypothesis testing*. https://simon.cs.vt.edu/SoSci/converted/T-Dist/

Yale University (1997) *Statistics department: Linear regression course*. <http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm>

Yemelyanov, D. (2020). Learn heteroskedasticity in 2 minutes. *Medium*. <https://medium.com/riga-data-science-club/learn-heteroskedasticity-in-1-minute-42f678911389>

# Appendix 2 – List of Tables and Figures

**Time Series Graph**

Source: Macfarlane (2016). [CC BY-SA 4.0]

--------------------------------------------------------------------------------------

**Histogram**

Source: Hayes (2011). [CC BY-SA 3.0]

--------------------------------------------------------------------------------------

**Standard Normal Distribution**

Source: Bhandari (2020).

--------------------------------------------------------------------------------------

**Standard Normal Distribution Probabilities**

Source: Toews (2007, April). [CC BY 2.5]

--------------------------------------------------------------------------------------

**Areas Under the Normal Curve Example**

Source: Farber & Larson (2017). [CC BY-SA 4.0]

--------------------------------------------------------------------------------------

**2 x 2 Crosstabulation**

Source: CITL (n.d.).

--------------------------------------------------------------------------------------

**Homoscedasticity and Heteroscedasticity**

Source: Yemelyanov (2020).

--------------------------------------------------------------------------------------

**Standard Bivariate Normal Distribution**

Source: Adhikari & Pitman (2020).

--------------------------------------------------------------------------------------

**Correlation Coefficient Examples**

Source: Laerd Statistics (2019). [CC BY-SA 4.0]

--------------------------------------------------------------------------------------

**Monotonic vs. Non-Monotonic Relationships**

Source: Magiya (2019).

--------------------------------------------------------------------------

**Discrete Probability Distribution**

Source: Viniciuslima94 (2017). [CC BY-SA 4.0]

--------------------------------------------------------------------------

**Continuous Probability Distribution**

Source: Toews (2007, July). [CC BY-SA 4.0]

--------------------------------------------------------------------------

**Typical Normal Distribution**

Source: Toews (2007, April). [CC BY 2.5]

--------------------------------------------------------------------------

**Student’s t-Distribution**

Source: Skbkekas (2010). [CC BY 3.0]

--------------------------------------------------------------------------

**All other tables and figures**

Source: David Fogarty (2021).