COMPANION WEBSITE
WWW.PEARSON-STUDIUM.DE

st
scientific tools

Reinhold Hatzinger
Herbert Nagel

PASW Statistics

Statistische Methoden und Fallbeispiele

PEARSON
Studium

Eine kategoriale Variable

4

ÜBERBLICK

Kategoriale Daten entstehen durch die Klassifikation von Beobachtungen in Kategorien. Meistens wird die interessierende Information dadurch gewonnen, dass ausgezählt wird, wie oft bestimmte Kategorien vorkommen. Dadurch können Häufigkeitsverteilungen und Prozentsätze berechnet werden. Dieses Kapitel beschäftigt sich damit, wie man prüfen kann, ob beobachtete Häufigkeiten mit bestimmten Annahmen übereinstimmen. Es werden die Grundlagen statistischen Testens besprochen. Weiters wird behandelt, wie man für einen Prozentsatz, den man aus einer Stichprobe gewonnen hat, Bereiche emittelt, von denen man annehmen kann, dass sie mit einer bestimmten Sicherheit den wirklichen Prozentsatz, wie er in der Population vorkommt, enthält. Dieses Kapitel legt die Basis zum Verständnis der Inferenz- oder schließenden Statistik.

4.1 Einleitung

Kategoriale Information erhält man, wenn etwas (ein Merkmal oder Charakteristikum), das man an verschiedenen Personen, Unternehmen, Pflanzen etc. (also Beobachtungseinheiten) registriert, in eine von mehreren Kategorien fällt. Dieser Prozess kann unterschiedlich verlaufen. Es können die Kategorien schon von vorneherein feststehen und man braucht seine Beobachtungen nur mehr entsprechend zuzuordnen. Oder aber man sammelt die Information zunächst in freier Form (z. B. offene Fragen in einem Fragebogen), um anschließend nach einem Kategorisierungsschema (das man eventuell erst entwickeln muss) die Beobachtungen zuzuteilen. Diese beiden Vorgänge könnte man auch als KLASSIFIKATION bezeichnen.

Etwas anders verläuft der Prozess der AGGREGATION. Der Begriff kommt aus dem Lateinischen und bedeutet Anhäufung oder Vereinigung. Hier werden schon vorhandene Daten in einfachere, zusammenfassendere Strukturen transformiert. Zum Beispiel kann man die Körpergröße von Personen, die man in cm gemessen hat, zu drei Kategorien aggregieren, nämlich klein, mittel und groß. Oder, bei zugrunde liegender kategorialer Information, kann man z. B. die Berufe Tischler, Maurer etc. in die Berufskategorie Handwerker zusammenfassen. Das Resultat einer Aggregation ist also oft eine Reihe von Kategorien einer Variable. Im Extremfall, besonders bei metrischen Daten, kann aber auch eine einzelne Maßzahl das Ergebnis sein, wenn z. B. die durchschnittliche Geburtenrate für ein Land bestimmt wird. Der Exkurs 4.1 illustriert am Beispiel von Berufen in Deutschland solche Vorgänge.

Wenn man nun kategoriale Daten erhoben oder mittels Klassifikation bzw. Aggregation gewonnen hat, geht es darum, diese Information in geeigneter Weise aufzubereiten, um die enthaltene Information erfassen, untersuchen und weitervermitteln zu können.

Der Vorgang hierbei ist das Auszählen, d. h., man zählt ab, wie oft Kategorie 1, Kategorie 2 etc. ingesamt vorkommt. Das Resultat dieser Auszählung nennt man *Häufigkeit* oder *absolute Häufigkeit*.

(absolute) Häufigkeit:
die Zahl, die zustande kommt, wenn man abzählt, wie oft eine bestimmte Kategorie in Daten vorkommt

Eng verbunden mit dem Begriff absolute Häufigkeit ist die

(absolute) Häufigkeitsverteilung:
eine Zusammenstellung (tabellarisch oder grafisch), wie oft einzelne Kategorien einer Variable in einer Stichprobe (oder auch der Population) vorkommen

Bezieht man die absolute Häufigkeit auf die Gesamtanzahl von Beobachtungen, dann erhält man relative Häufigkeiten oder Anteile.

relative Häufigkeit oder **Anteil:**

$$\textit{relative Häufigkeit einer Kategorie} = \frac{\textit{absolute Häufigkeit}}{\textit{Gesamtanzahl von Beobachtungen}}$$

Anteile werden oft als Prozentsätze (also als Anteile von 100) angegeben, d. h.

Prozent:

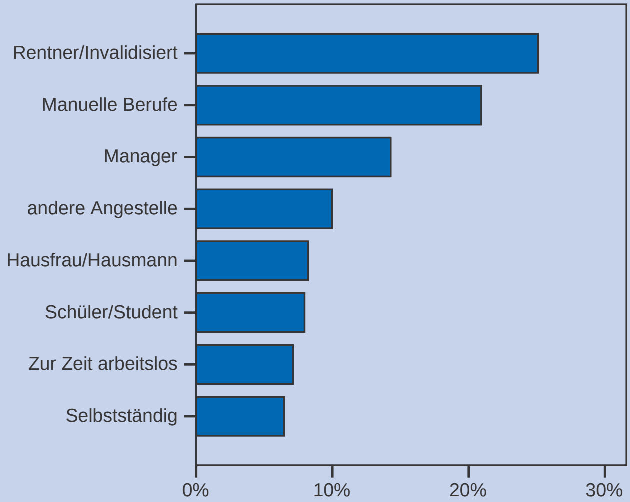
$$\textit{Prozentsatz einer Kategorie} = \textit{relative Häufigkeit} \times 100$$

Während Angaben von Anteilen in Prozentsätzen leichter zu lesen sind und auch allgemein üblich verwendet werden, hat die Darstellung mittels relativer Häufigkeiten den Vorteil, dass dadurch deren enge Verwandtschaft mit Wahrscheinlichkeiten ausgedrückt wird. Relative Häufigkeiten liefern (wenn zusätzlich die Gesamtzahl von Beobachtungen angegeben wird) die bedeutsamste Information, die man aus Daten gewinnen kann. Dies gilt nicht nur für kategoriale Daten. Auf diese Punkte werden wir später noch eingehen.

Exkurs 4.1 Wie kommt kategoriale Information zustande?

Beispiel: Berufe in Deutschland

Die simple Frage *Was arbeiten die Menschen eigentlich so?* ist gar nicht so einfach zu beantworten. Mögliche Quellen, wo man die entsprechende Information herbekommen könnte, sind die nationalen statistischen Behörden, für Deutschland das Statistische Bundesamt (www.destatis.de), für Österreich Statistik Austria (www.statistik.at), für die Schweiz das Bundesamt für Statistik (www.bfs.admin.ch) oder Eurostat als europäische Institution (ec.europa.eu/eurostat). Man könnte von einer dieser Quellen für Deutschland (2005) etwa folgende Grafik finden:



Die Grafik gibt einen guten Überblick, wie viel Prozent der Deutschen welcher Art von Tätigkeit nachgehen. Dieser Überblick ist allerdings aus zwei Gründen recht ungenau. Wenn wir wissen wollten, wie viele Personen tatsächlich selbstständig sind, müssten wir zunächst wissen, wie viele Personen insgesamt zur Erstellung der Grafik herangezogen worden waren (Kinder scheinen jedenfalls nicht dabei gewesen zu sein). Und zweitens kann man die angegebenen Prozentsätze nur ungefähr ablesen. Versuchen wir es dennoch. An anderer Stelle findet man auf der Webseite des deutschen statistischen Bundesamts, dass es im Jahr 2005 ca. 82 438 000 Einwohner in Deutschland gab, davon 11 649 800 Kinder und Jugendliche unter 16 Jahren. Für unsere Berechnung müssten wir als Gesamtzahl 70 788 200 verwenden. Laut Grafik sind etwa 7% selbstständig, insgesamt gab es also 2005 knapp 5 Millionen Selbstständige.

Wir sehen hier zwei verschiedene Arten der Angabe von Zahlen im Zusammenhang mit kategorialer Information, nämlich (absolute) Häufigkeiten, d. h., wie oft ist insgesamt eine Kategorie vorgekommen, und Anteilen (hier Prozentsätze), d. h., in welchem (relativen) Ausmaß kam eine bestimmte Kategorie im Verhältnis zu allen anderen Kategorien vor. Das eine Mal bezieht sich die Information nur auf eine bestimmte Kategorie (ohne die anderen zu berücksichtigen), das andere Mal erhält man indirekt auch Information über weitere Kategorien. Auf diese Unterscheidung werden wir noch näher eingehen.

Der vorherigen Grafik kann man entnehmen, dass ca. 10% als *andere Angestellte* bezeichnet werden. Was ist darunter zu verstehen? Aus den bisherigen Angaben ist dies nicht ohne Weiteres ersichtlich. Wir können annehmen, dass unter diesem Begriff eine größere Zahl von weniger häufig auftretenden Kategorien zusammengefasst wird. Solche Zusammenfassungen dienen der Übersichtlichkeit und der einfacheren Kommunizierbarkeit von wichtigen Informationsbestandteilen. Allerdings geht dabei natürlich ein Teil der ursprünglich vorhandenen Information verloren.

Tatsächlich wurde bei der Datenerhebung genauer gefragt.

1. Hausfrau/Hausmann und verantwortlich für den Haushaltseinkauf und den Haushalt (ohne anderweitige Beschäftigung)
2. Schüler/Student
3. Zur Zeit arbeitslos
4. Rentner/Pensionär/Frührentner/Invalidisiert

5. Landwirt
6. Fischer
7. Freie Berufe (z. B. Rechtsanwalt, Arzt, Steuerberater, Architekt usw.)
8. Ladenbesitzer, Handwerker usw.
9. Selbstständige Unternehmer, Fabrikbesitzer (Alleininhaber, Teilhaber)
10. Freie Berufe im Angestelltenverhältnis (z. B. angestellte Ärzte, Anwälte, Steuerberater, Architekten usw.)
11. Leitende Angestellte/Beamte, Direktor oder Vorstandsmitglied
12. Mittlere Angestellte/Beamte (Bereichsleiter, Abteilungsleiter, Gruppenleiter, Lehrer, Technischer Leiter)
13. Sonstige Büroangestellte/Beamte
14. Angestellte/Beamte ohne Bürotätigkeit mit Schwerpunkt Reisetätigkeit (Vertreter, Fahrer)
15. Angestellte/Beamte ohne Bürotätigkeit z. B. im Dienstleistungsbetrieb (Krankenschwester, Bedienung in Restaurant, Polizist, Feuerwehrmann)
16. Meister, Vorarbeiter, Aufsichtstätigkeit
17. Facharbeiter
18. Sonstige Arbeiter

Unter *andere Angestellte* wurden die Kategorien 13 und 14 zusammengefasst. Die mit etwa 15% relativ stark vertretene Gruppe der *Manager* umfasste die Kategorien 10 bis 12. Doch selbst diese feinere Gruppierung (wie sie in der Eurobarometerumfrage verwendet wird) kann in noch detailliertere Kategorien aufgeschlüsselt werden. So verwendet die deutsche Bundesagentur für Arbeit die vom deutschen statistischen Bundesamt entwickelte Berufssystematik, die auf der untersten Ebene 29 500 Berufsbezeichnungen definiert.

Anhand dieses Beispiels können wir einige Grundüberlegungen anstellen. Die Beobachtungseinheiten sind hier einzelne Personen. Die Variable, die beobachtet (bzw. aufgezeichnet) wurde, könnte man mit *berufliche Tätigkeit* bezeichnen. Die Ausprägung dieser Variable ist kategorial, d. h., es wurde jede einzelne Person gefragt, in welche der vorher festgelegten Kategorien sie fällt. Die Daten könnten folgendermaßen aussehen.

Person	Berufstätigkeit
1	Mittlere Angestellte/Beamte
2	Rentner/Invalidisiert
3	Hausfrau/Hausmann
4	Leitende Angestellte/Beamte
5	Mittlere Angestellte/Beamte
6	Sonstige Arbeiter
7	Rentner/Invalidisiert
⋮	⋮

Diese fiktive Liste enthält Einzelinformationen über die Art der Beschäftigung. Solch eine Liste nennt man *Rohdatenliste* und so ähnlich könnte auch eine Computerdatei aussehen, mittels derer die vorangegangene Grafik erstellt wurde. In dieser Art der Datendarstellung ist natürlich die meiste Information zu finden, zumal ja der Beruf jeder einzelnen Person verzeichnet ist. Allerdings kann man sich leicht vorstellen, wie lang diese Liste ist. Man muss daher die Daten reduzieren bzw. zusammenfassen, um die darin enthaltene Information untersuchen und weitervermitteln zu können.

4.2 Kommen alle Kategorien gleich häufig vor?

Fallbeispiel 1: Der gefährlichste Wochentag in New Jersey

Datenfile: `murder03.sav`

Die Kriminalstatistik für den US-Bundesstaat New Jersey im Jahr 2003 gab unter anderem die Zahl begangener Morde an einzelnen Wochentagen an. Die folgende Häufigkeitstabelle zeigt, wie viele Morde an den einzelnen Wochentagen verübt wurden.

So	Mo	Di	Mi	Do	Fr	Sa
53	42	51	45	36	37	65

Man sieht, dass die meisten Morde an Samstagen begangen wurden, gefolgt von Sonntag und Dienstag. Donnerstag und Freitag sind die Häufigkeiten geringer. Es stellt sich die Frage, ob tatsächlich eine größere Gefahr an Wochenenden bzw. am Dienstag besteht oder ob diese Häufungen nur zufällig sind.

Gibt es gefährlichere Wochentage oder ist es an allen Wochentagen gleich gefährlich, ermordet zu werden?

Bevor wir uns überlegen, wie man diese Frage beantworten kann, wollen wir einige Grundsätze besprechen, die bei der Analyse von Daten bzw. Fragestellungen beachtet werden sollten.

Im Beispiel der Morde an einzelnen Wochentagen ist die Grundstruktur der Daten relativ einfach. Trotzdem sind die Rohdaten sehr unübersichtlich. Nach Öffnen der Datei `murder03.sav` sieht man im **PASW-Daten-Editor** und dort in der **Datenansicht** auszugsweise die ersten zehn von 329 Fällen.

Hier haben wir zwar die maximal zur Verfügung stehende Information, allerdings ist diese so nicht kommunizierbar. Wir benötigen also Methoden, wie wir die Inhalte kompakt darstellen können, ohne zu viel Information zu verlieren. Dies kann anhand von Grafiken oder numerischen Zusammenfassungen (z. B. Tabellen) geschehen. Die entsprechenden Methoden werden unter dem Begriff Datenbeschreibung oder deskriptive Statistik zusammengefasst und bestehen, je nach Datentyp, aus

	Wochentag
1	Dienstag
2	Dienstag
3	Sonntag
4	Dienstag
5	Sonntag
6	Mittwoch
7	Freitag
8	Freitag
9	Dienstag
10	Mittwoch

Abbildung 4.1: Die ersten zehn Fälle aus der Datei `murder03.sav` im Daten-Editor

spezifischen Verfahren. Numerische und grafische Beschreibung der Daten sollten immer der Ausgangspunkt bei der Analyse von Daten sein.

4.2.1 Numerische Beschreibung

Bei einfacher kategorialer Information zählen wir aus, wie häufig jede Kategorie auftritt und erhalten absolute Häufigkeiten. Dividieren wir diese noch zusätzlich durch die Gesamtanzahl an Beobachtungen, ergibt das die relativen Häufigkeiten. Wir können beide in Tabellenform darstellen und erhalten eine übersichtliche numerische Beschreibung des Datenmaterials, aus der schon einige Aspekte zur Beantwortung der Fragestellung ablesbar sind.

In PASW würden wir folgendermaßen vorgehen:

PASW

Analysieren
  Deskriptive Statistiken
    Häufigkeiten...
      markieren der Variable (`Wochentag`) [➡] Variable(n):
      [✓] Häufigkeitstabellen anzeigen
      [OK]

Diese Darstellung bedeutet:

Zunächst wählen wir aus dem Menü Analysieren den Untermenüpunkt Deskriptive Statistiken und dort Häufigkeiten (▶ Abbildung 4.2).

In der folgenden Dialogbox (▶ Abbildung 4.3) markiert man die darzustellende Variable (hier Wochentag) und bewegt sie mit einem Klick auf den mittleren Pfeil [➡] ins rechte Fenster. Nachdem man sich vergewissert hat, dass die Checkbox Häufigkeitstabellen anzeigen gewählt ist [✔], klickt man auf [OK].

Abbildung 4.2: Auswahlmenü Häufigkeiten

Abbildung 4.3: Dialogbox Häufigkeiten zur Variablenauswahl

Im PASW-Viewer erscheint (nach einer kleinen Tabelle mit Angaben über gültige und fehlende Werte) das Resultat (► Abbildung 4.4)

Die ersten beiden Spalten zeigen die (absoluten) Häufigkeiten und Prozentwerte für die Anzahl an Morden an den einzelnen Wochentagen. Außerdem wird in der letzten Zeile die Gesamtanzahl 379 an Morden in New Jersey 2003 ausgegeben. Man sieht, dass an Samstagen und Sonntagen (eventuell auch noch Dienstag) die meisten Morde verübt werden, während die Zahl unter der Woche geringer ist.

Die Spalte gültige Prozent stimmt hier mit der Spalte Prozent überein, da in den Daten keine fehlenden Werte vorkommen. Die Spalte kumulierte Prozent ist bei nominalen Variablen sinnlos und kann hier ignoriert werden. Die Tabelle in Abbildung 4.4 ist nicht sehr übersichtlich und würde in dieser Form auch nicht in einer Publikation oder Präsentation verwendet werden, besonders nicht die beiden letzten Spalten (Exkurs 4.2). Glücklicherweise kann man in PASW das Aussehen einer Tabelle modifizieren. Dies wird in Kapitel 3.9.1 beschrieben.

Wochentag

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Sonntag	53	16,1	16,1	16,1
	Montag	42	12,8	12,8	28,9
	Dienstag	51	15,5	15,5	44,4
	Mittwoch	45	13,7	13,7	58,1
	Donnerstag	36	10,9	10,9	69,0
	Freitag	37	11,2	11,2	80,2
	Samstag	65	19,8	19,8	100,0
	Gesamt	329	100,0	100,0	

Abbildung 4.4: Häufigkeitstabelle zu Wochentagen

4.2.2 Grafische Beschreibung

Der Spruch *Ein Bild sagt mehr als tausend Worte* klingt sehr abgedroschen, trotzdem stimmt er uneingeschränkt. Gerade dann, wenn Information klar gemacht und vermittelt werden soll, ist ein gute grafische Darstellung besonders wichtig. Bei kategorialen Daten haben sich einige Methoden bewährt, die hier beschrieben werden sollen.

Exkurs 4.2 Einige Prinzipien zur Erstellung guter Tabellen

Will man kategoriale Information in Form von Zahlenmaterial geeignet präsentieren, so wird man wohl kaum eine Darstellungsform wie in Abbildung 4.4 wählen, wie sie direkt in PASW erzeugt wird. Folgende Punkte sollte man beachten:

- Eine Tabelle sollte für sich allein stehen können, d. h., alles Wichtige sollte ohne weitere Erklärung verständlich sein.
- Angabe eines geeigneten Titels und der Datenquelle
- Benennung der Kategorien (in PASW Wertelabels)
- Bei ungeordneten Kategorien: diese nach ihrer Häufigkeit ordnen
- Angabe der Gesamtanzahl von Beobachtungen
- Zahlen runden (besonders Nachkommastellen nur angeben, wenn sie für das Verständnis der Größenordnung wichtig sind)
- Sehr große Tabelle eventuell in Teiltabellen zerlegen
- Keine vertikalen Linien zeichnen

Die Tabelle aus Abbildung 4.4 könnte dann so aussehen:

Verteilung begangener Morde nach Wochentagen in New Jersey, 2005[a]

	Sonntag	Montag	Dienstag	Mittwoch	Donnerstag	Freitag	Samstag	Gesamt
Häufigkeit	53	42	51	45	36	37	65	329
Prozent	16	13	16	14	11	11	20	

a. Quelle: http://www.njsp.com

Balkendiagramm (Bar Chart)

Diese Darstellungsweise ist wohl die wichtigste grafische Methode bei kategorialen Daten. Bei einem Balkendiagramm werden die Kategorien auf der x-Achse und die absoluten oder relativen Häufigkeiten auf der y-Achse aufgetragen, wobei die Höhe der einzelnen Balken den Häufigkeiten in den einzelnen Kategorien entspricht. Um zu veranschaulichen, dass die darzustellende Variable kategorial ist, lässt man einen kleinen Leerraum zwischen den Balken (man macht dies im Gegensatz zu einem sogenannten Histogramm, das wir später besprechen werden). Manchmal findet man auch die beiden Achsen vertauscht vor (Exkurs 4.1), an der Information, die man aus solch einer Grafik ablesen kann, ändert sich dadurch aber nichts.

Für unser Beispiel der Morde nach Wochentagen erzeugt man ein Balkendiagramm in PASW folgendermaßen:

PASW

Diagramme
Diagrammerstellung... OK
in Galerie *auswählen von* Balken
Doppelklick auf linkes oberes Galeriediagramm einfache Balken
(*die Vorschau erscheint in der Zeichenfläche*)
Variable `Wochentag` *in das Feld* X-Achse? *ziehen*
OK

Das Resultat findet man wieder im PASW-Viewer (► Abbildung 4.5).

Am Balkendiagramm in ► Abbildung 4.5 sieht man (vielleicht noch besser als in der Häufigkeitstabelle), dass Samstag der gefährlichste Tag ist, gefolgt von Sonntag und Dienstag.

HINWEIS

Mit Balkendiagrammen (aber ganz allgemein mit Grafiken) kann man gut lügen. Indem man die Skala ändert, könnte man die Daten aus unserem Beispiel leicht viel dramatischer erscheinen lassen (► Abbildung 4.6). Plötzlich sieht der Samstag viel gefährlicher aus als vorher. Solche (bewusst oder unbewusst) manipulativen Veränderungen der Information findet man häufig in den Medien oder in Präsentationen. Achten Sie einmal darauf. Jedenfalls sollte bei einfachen Balkendiagrammen die Skala immer bei null beginnen. Tritt der Fall ein, dass aus irgendwelchen Gründen die reskalierte Form verwendet wird, dann sollte man zumindest darauf hinweisen.

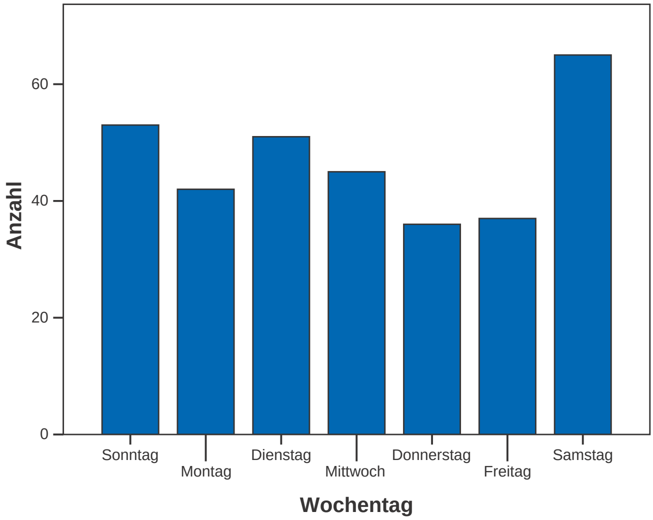


Abbildung 4.5: Balkendiagramm für die Anzahl von Morden nach Wochentagen

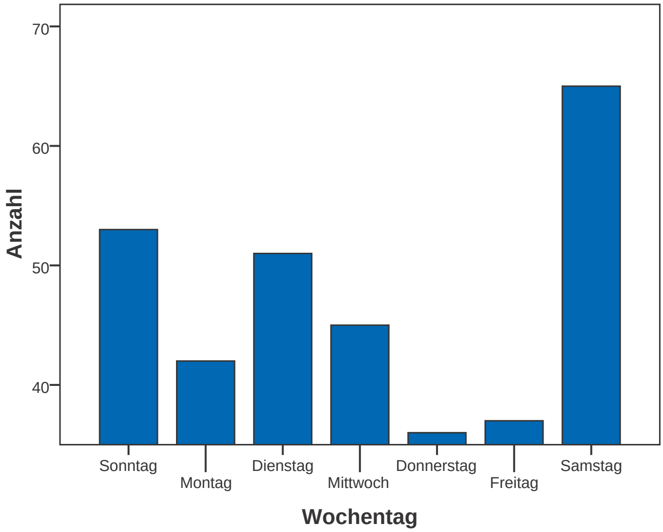


Abbildung 4.6: Verfälschtes Balkendiagramm für die Anzahl von Morden nach Wochentagen

Kreisdiagramm (Pie Chart)

Häufig begegnet man auch einer zweiten Form der grafischen Beschreibung von einfacher kategorialer Information, den Kreisdiagrammen (manchmal auch Torten- oder Kuchendiagramme genannt). Hierbei werden die Daten in Kreisform dargestellt. Die Kategorien bilden Kreissegmente, wobei deren Fläche proportional zu den relativen Häufigkeiten bzw. Prozenten des Auftretens der Kategorien dargestellt werden.

Obwohl sie relativ beliebt sind, haben Kreisdiagramme doch eine Reihe von Nachteilen. Standardmäßig gibt PASW nur das Kreisdiagramm und die Bezeichnungen der Kategorien aus.

Für die Daten der Morde nach Wochentagen erzeugt man in PASW ein Kreisdiagramm mit

PASW

Diagramme
Diagrammerstellung... [OK]
in Galerie *auswählen von* Kreis/Polar
Doppelklick auf Galerieelement [Kreisdiagramm]
(*die Vorschau erscheint in der Zeichenfläche*)
Variable `Wochentag` *in das Feld* Aufteilen nach? *ziehen*
[OK]

Im PASW-Viewer erhält man die Grafik in ▶ Abbildung 4.7.

Man kann nicht (wie beim Balkendiagramm) die tatsächlichen Häufigkeiten ablesen und vergleichsweise sind Unterschiede nur schwer zu erkennen, außer sie sind

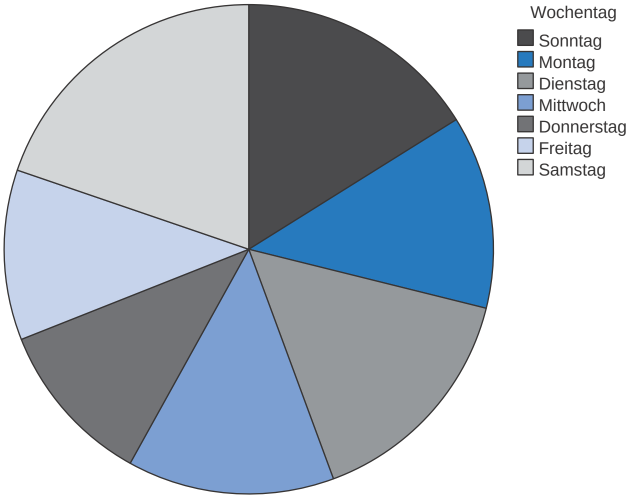


Abbildung 4.7: Kreisdiagramm für die Anzahl von Morden nach Wochentagen

sehr groß. Man muss einige Modifikationen anbringen, um zu einer geeigneteren Darstellung zu kommen.

Wenn man auf das Kreisdiagramm im PASW-Viewer doppelklickt, gelangt man in den Diagramm-Editor. Unter dem Menüpunkt Elemente wählt man Datenbeschriftungen einblenden. Es öffnet sich das Dialogfenster Eigenschaften (▶ Abbildung 4.8).

Hier bringen wir folgende Modifikationen an:

- Im Feld Nicht angezeigt markieren wir zunächst die Variable Wochentag und klicken mit dem rechten Pfeil nach oben in das Feld Angezeigt.
- Das Gleiche machen wir mit Prozent.
- Im oberen Feld Angezeigt können wir dann die Reihenfolge für die Darstellung mit den rechten Pfeilen nach oben/unten ändern.
- Im Feld Beschriftungsposition wählen wir die unterste Option Benutzerdefiniert und klicken auf das mittlere Dialogbild, auf dem das Kästchen innerhalb des Kreises erscheint.
- Die Änderungen werden mit Zuweisen abgeschlossen.
- Zur besseren Lesbarkeit kann man noch die neu erzeugten Beschriftungen adaptieren, indem man in Text-Layout den Abstand horizontal und vertikal auf z. B. 3 setzt, danach Zuweisen.

Nach Schließen der Fenster Eigenschaften und Diagramm-Editor erhalten wir das Kreisdiagramm wie in ▶ Abbildung 4.9.

Es ist ein letztlich eine Geschmacksfrage, ob man ein Balkendiagramm oder ein Kreisdiagramm zur Beschreibung kategorialer Information verwendet. Lesbarer und informativer dürften aber wohl Balkendiagramme (Abbildung 4.5) sein. Zu erwähnen ist noch, dass die Beschriftungen, wie wir sie für das Kreisdiagramm in Abbildung 4.9 angebracht haben, in analoger Weise auch für Balkendiagramme möglich sind. Notwendig sind sie dort aber nicht, weil durch die Achsenbeschriftungen die relevante Information schon mitgeliefert wird.

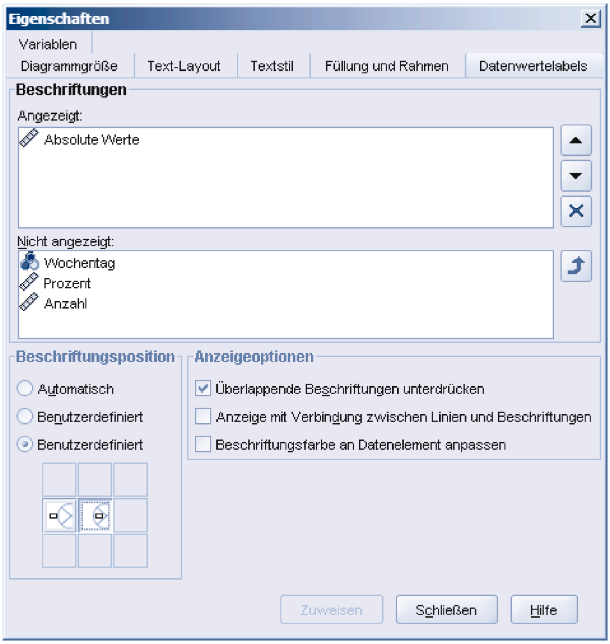


Abbildung 4.8: Dialogbox Eigenschaften im Diagramm-Editor

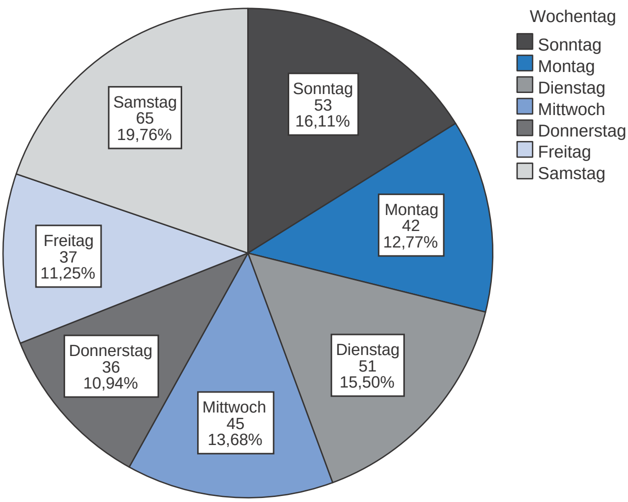


Abbildung 4.9: Modifiziertes Kreisdiagramm für die Anzahl von Morden nach Wochentagen

4.2.3 Statistische Analyse der Problemstellung

Nachdem wir nun besprochen haben, wie die Information, die in den Daten steckt, numerisch und grafisch beschrieben werden kann, wollen wir uns der Beantwortung der eigentlichen Problemstellung zuwenden. Die Frage ist, ob es bestimmte Wochentage gibt, an denen Morde häufiger auftreten als an anderen, oder ob sich die Morde über die Wochentage gleich verteilen.

Um diese Frage mittels statistischer Methoden beantworten zu können, benötigen wir das Konzept der ERWARTETEN RELATIVEN HÄUFIGKEITEN und der ERWARTETEN (ABSOLUTEN) HÄUFIGKEITEN. *Beobachtete Häufigkeiten* (absolut oder relativ) entstehen durch Abzählen, *erwartete* Häufigkeiten sind solche, die man unter bestimmten Annahmen erwarten würde. In Exkurs 4.3 wird dieses Konzept anhand der Idee des *fairen Würfels* erläutert.

Exkurs 4.3 Der faire Würfel

Ein Würfel wird dann als „fair" bezeichnet, wenn alle sechs Seiten gleich häufig auftreten, oder mit anderen Worten, wenn die Wahrscheinlichkeit für jede der möglichen Seiten 1 bis 6 gleich groß ist. Die Erfahrung zeigt uns, dass wir nicht damit rechnen können, dass bei sechsmaligem Würfeln als Resultat alle sechs Augenzahlen je einmal vorkommen. Allerdings, wenn wir nicht nur sechs Mal, sondern viel öfter würfeln, wie das auch bei Gesellschaftsspielen der Fall ist, dann erwarten wir schon, dass alle Seiten am Ende ungefähr gleich oft aufgetreten sind. Würden wir immer längere Serien würfeln, also 10.000 Mal, 100.000 Mal etc., dann sollten die Prozentsätze des Auftretens der einzelnen Augenzahlen immer ähnlicher werden und wir könnten sagen, dass der Würfel fair ist. Das setzt aber voraus, dass der Würfel verschiedene Voraussetzungen erfüllt, nämlich vollkommen gleichmäßige Beschaffenheit des Materials, aus dem er hergestellt ist, exakt gleiche Kantenlängen (nicht nur auf den hundertstel Millimeter genau) etc. Solch einen Würfel gibt es natürlich nicht, aber man kann einen herstellen, bei dem diese Abweichungen nur sehr klein und daher vernachlässigbar sind. Bei manchen Würfeln kann einem aber schon der Verdacht kommen, dass er nicht fair ist. Wie könnten wir das nun prüfen? Die Antwort ist Ausprobieren. Nehmen wir einmal an, wir hätten 30 Mal gewürfelt und uns die Häufigkeiten für die einzelnen Seiten wie folgt notiert:

gewürfelte Augenzahl	1	2	3	4	5	6
beobachtete Häufigkeit	3	8	4	5	6	4

Unter der Annahme, dass der Würfel fair ist, hätten wir eigentlich erwartet, dass jede Seite gleich häufig, nämlich zu 1/6 bzw. zu $16,7\%$ aufgetreten wäre. Diese erwarteten Anteile nennt man auch ERWARTETE RELATIVE HÄUFIGKEITEN. Multipliziert man sie mit der Gesamtanzahl der Würfe, hier mit 30, so erhält man die ERWARTETEN (ABSOLUTEN) HÄUFIGKEITEN. Demnach hätten wir je 5 Mal die Augenzahlen 1 bis 6 erwartet und nicht 3, 8, 4, ... Diese Abweichungen schreiben wir dem Zufall zu. Stutzig hätte uns womöglich gemacht, wenn z. B. die Seite mit 6 dreißig Mal und die anderen Seiten kein einziges Mal vorgekommen wären. Aber ab wann sollte uns die Sache verdächtig vorkommen? Diese Frage wollen wir versuchen zu beantworten.

Für unsere Fragestellung wollen wir die Annahme prüfen, ob die Morde sich gleichmäßig über die Wochentage verteilen. Unter dieser Annahme müssten von den insgesamt 329 Morden ungefähr $329/7 = 47$ an jedem Wochentag auftreten, wir würden also 47 Morde an jedem Tag *erwarten*.

In PASW können wir eine entsprechende Tabelle erzeugen.

PASW

Analysieren
Nichtparametrische Tests
Chi-Quadrat...
Variable (`Wochentag`) ➡ Testvariablen
OK

Im PASW-Viewer erhalten wir drei Tabellen, wovon uns vorläufig nur die mittlere interessiert.

Wochentag

	Beobachtetes N	Erwartete Anzahl	Residuum
Sonntag	53	47,0	6,0
Montag	42	47,0	-5,0
Dienstag	51	47,0	4,0
Mittwoch	45	47,0	-2,0
Donnerstag	36	47,0	-11,0
Freitag	37	47,0	-10,0
Samstag	65	47,0	18,0
Gesamt	329		

Abbildung 4.10: Tabelle der beobachteten und erwarteten Häufigkeiten

Die Tabelle in ▶ Abbildung 4.10 hat drei Spalten. In der ersten (Beobachtetes N) finden wir die tatsächlich beobachteten Häufigkeiten für die einzelnen Wochentage, in der nächsten (Erwartete Anzahl) die erwarteten Häufigkeiten und in der letzten (Residuum) die Differenz der ersten beiden Spalten.

Wir sehen, dass die Differenzen zwischen −2 und 18 liegen. Wie beim fairen Würfel können wir uns nun fragen, ob diese Abweichungen zufällig zustande gekommen sind oder ob mehr dahinter steckt. Je größer die Abweichungen insgesamt sind, umso weniger werden wir an Zufall glauben. Sind sie groß genug, dann werden wir die Frage verneinen, dass die Morde gleichmäßig verteilt sind. Aber wo ist die Grenze zwischen *rein zufällig* und nicht mehr *nur durch Zufall* erklärbar? Hier kann uns die Statistik weiterhelfen.

Exkurs 4.4 Statistische Tests – Hypothesen – p-Wert Was ist noch Zufall und was nicht mehr?

Am Beginn steht immer eine Fragestellung, z. B. ob ein Würfel fair ist. Die Methode, die man verwendet, um solch eine Frage zu beantworten, nennt man STATISTISCHEN TEST. Es gibt zwei mögliche Ergebnisse: Wir glauben entweder

- der Würfel ist fair oder
- der Würfel ist nicht fair.

Diese beiden Statements nennt man STATISTISCHE HYPOTHESEN. Die sogenannte NULLHYPOTHESE (auch mit H_0 bezeichnet) lautet:

- der Würfel ist fair oder
  H_0: alle Augenzahlen beim Würfeln sind gleich wahrscheinlich

Das logische Gegenteil ist die ALTERNATIVHYPOTHESE (oder H_A), die uns helfen soll, solche Annahmen zu überprüfen. Sie lautet:

- der Würfel ist nicht fair oder
  H_A: zumindest eine der Augenzahlen ist wahrscheinlicher als die anderen

Ein STATISTISCHER TEST hilft uns, eine Entscheidung zugunsten der Null- oder der Alternativhypothese zu treffen. Man kann dabei eine Wahrscheinlichkeit angeben, ob beobachetete Daten für die Nullhypothese sprechen. Diese Wahrscheinlichkeit bezeichnet man als P-WERT (oder wie in PASW SIGNIFIKANZ). Die Entscheidungsregel lautet:

- Ist der p-Wert klein, glaubt man nicht an die Nullhypothese, man glaubt eher, dass die Alternativhypothese zutrifft – der Würfel ist nicht fair.
- Ist der p-Wert groß, glaubt man an die Nullhypothese – der Würfel ist fair.

Im ersten Fall sagt man, *die Nullhypothese wird* (zugunsten der Alternativhypothese) *verworfen*, im zweiten Fall sagt man, *die Nullhypothese wird beibehalten*.

Wann ist ein p-Wert groß bzw. klein?
Dies ist keine statistische Frage, sondern ein Frage der subjektiven Sicherheit. Üblicherweise verwendet man einen Vergleichswert von 0,05 oder 0,01. Diese Festlegung wird SIGNIFIKANZNIVEAU (auch α) genannt. Ein Wert von 0,05 besagt, dass man sich in 1 von 20 Fällen irrt und dass man die Nullhypothese verwirft, obwohl sie eigentlich zutreffend ist. Das heißt, wenn man sehr viele Versuche mit einem fairen Würfel durchführen würde, dann käme man in 5% zufällig zu dem Ergebnis, dass der Würfel nicht fair ist, obwohl das eigentlich nicht stimmt (Anhang 4.8).

Wir registrieren Differenzen zwischen den beobachteten Häufigkeiten und den Häufigkeiten, die wir erwarten würden, wenn die Nullhypothese gültig wäre. Aus diesen Differenzen können wir uns eine einzelne Maßzahl überlegen, die solche Abweichungen beschreibt. Sie soll groß sein, wenn die Abweichungen groß sind und umgekehrt. Es gäbe natürlich verschiedene Möglichkeiten solch eine Maßzahl zu konstruieren, die wichtigste stammt aber von Karl Pearson. Wir wollen sie einem allgemeinen Sprachgebrauch nach Pearson X^2 nennen.

Berechnung von X^2 bei eindimensionalen Häufigkeitsverteilungen

$$X^2 = \sum_{j=1}^{J} \frac{(o_j - e_j)^2}{e_j} \tag{4.1}$$

- $o_j \ldots$ beobachtete Häufigkeit für die Kategorie j (o steht für *observed*)
- $e_j \ldots$ erwartete Häufigkeit für die Kategorie j
- $J \ldots$ Gesamtanzahl der Kategorien

Der X^2-Wert gibt uns Auskunft über die Größe der Abweichungen zwischen erwarteten und beobachteten Häufigkeiten. Anhand der Formel 4.1, in die ja die Differenz $o_j - e_j$ eingeht, sieht man, dass das X^2 mit der Größe der Abweichungen steigt. Für unser Beispiel (wie wir gleich sehen werden) ist dieser Wert $X^2 = 13{,}3$. Nun sollte uns dieser Wert einen Anhaltspunkt darüber geben, ob die Abweichungen zwischen den erwarteten und den beobachteten Häufigkeiten auffällig sind (man sagt auch *statistisch signifikant* bzw. *bedeutsam* in dem Sinn, dass man eher nicht daran glaubt, dass die Morde gleichmäßig über die Wochentage verteilt sind).

Die dritte Tabelle, die wir vorher erzeugt haben, gibt uns Aufschluss.

In ▶ Abbildung 4.11 finden wir drei Zahlen. Die erste, Chi-Quadrat, gibt uns den Wert 13,319. Das ist das vorher besprochene Pearson X^2 aus Formel 4.1. Der zweite Wert, df, bedeutet FREIHEITSGRADE (aus dem Englischen *degrees of freedom*) und ist die Anzahl der Kategorien minus 1 ($df = J - 1$). Die dritte Zahl, Asymptotische Signifikanz, ist hier die wichtigste. Sie ist eine Wahrscheinlichkeit und liegt daher zwischen 0 und 1. Sie gibt uns (sehr vereinfacht gesagt) an, wie plausibel unsere Annahme ist, dass die Morde gleichmäßig auf alle Wochentage verteilt sind. Diese Zahl ist sehr klein, d. h., die Nullhypothese ist unplausibel. Eine Interpretation des Ergebnisses könnte so aussehen.

Fallbeispiel 1: Interpretation

Der Chi-Quadrat-Test zur Überprüfung der Nullhypothese, dass die Morde an allen Wochentagen gleich häufig auftreten, musste verworfen werden ($X^2 = 13{,}3$, $df = 5$, $p < 0{,}04$). Die Daten weisen darauf hin, dass an verschiedenen Wochentagen unterschiedlich viele Morde auftreten. Die meisten Morde werden an Samstagen begangen, gefolgt von Sonntag und Dienstag. An den anderen Tagen sind die Häufigkeiten geringer.

Die Tabelle in Abbildung 4.11 und die Schlussfolgerungen, die wir daraus gezogen haben, bedürfen einiger zusätzlicher Erläuterungen. In Appendix 4.8 zu diesem Kapitel werden die theoretischen Grundlagen ausführlicher beschrieben.

Zunächst ist Ihnen vielleicht der Unterschied aufgefallen, dass wir X^2 verwendet haben, in PASW aber der Begriff Chi-Quadrat für das Gleiche aufgetaucht ist. PASW ist hier ein bisschen schlampig, weil man eigentlich zwischen X^2, einer Zahl die aus beobachteten Daten errechnet wurde, und Chi-Quadrat, einer theoretischen Größe, unterscheiden sollte (eine ausführlichere Erklärung gibt der erwähnte Appendix 4.8).

Statistik für Test

	Wochentag
Chi-Quadrat	13,319
df	6
Asymptotische Signifikanz	,038

Abbildung 4.11: Chi-Quadrat-Test für Morde nach Wochentagen

Der Begriff Freiheitsgrade wird uns noch öfter begegnen. Vereinfacht gesagt dienen Freiheitsgrade dazu, den dritten Wert, die Signifikanz, zu bestimmen. Freiheitsgrade werden auch im Appendix 4.8 näher erklärt.

Der dritte Begriff, Signifikanz, ist wie erwähnt der wichtigste Wert, wenn es darum geht, eine (statistische) Fragestellung zu beantworten bzw. die Gültigkeit einer Annahme zu evaluieren.

Wie interpretiert man ein Ergebnis?

Zweck einer Interpretation ist das Zusammenführen von technischen Ergebnissen eines statistischen Verfahrens mit den inhaltlichen Aspekten der Fragestellung, für die man die statistische Methode verwendet hat. Es gibt keine exakten, allgemeingültigen Regeln, wie man ein Ergebnis interpretiert. Die genaue Länge und Form hängt davon´ab, für welchen Zweck man eine Interpretation erstellt. Manche wissenschaftliche Zeitschriften oder Institutionen geben gewissen Formvorlagen oder Richtlinien. Im allgemeinen sollte eine Interpretation aus zwei Teilen bestehen, aus einem technischen und einem inhaltlichen.

- Technische Interpretation: Hier wird oft die Nullhypothese (eventuell auch Alternativhypothese) formuliert, welches statistische Verfahren (eventuell auch warum) angewendet wurde, Kennzahlen des speziellen statistischen Verfahrens und der p-Wert. Wird die Nullhypothese verworfen oder beibehalten.
- Inhaltliche Interpretation: Was bedeutet das technische Ergebnis. Die Fragestellung wird beantwortet und es werden wichtige beschreibende Fakten (wie z. B. Prozentsätze etc.) angegeben.

Je nach Fragstellung und Methode wird eine Interpretation anders aussehen. Beispiele können Sie den Kästen Interpretation entnehmen, die immer am Ende der Analyseabschnitte in diesem Buch angeführt werden.

4.3 Entsprechen Häufigkeiten bestimmten Vorgaben?

Im Abschnitt 4.2 haben wir uns mit der Frage beschäftigt, ob beobachtete Häufigkeiten für Kategorien einer Variable mit der Hypothese in Einklang stehen, dass in Wirklichkeit alle Kategorien gleich wahrscheinlich sind. Wir wollen diese Problemstellung nun insofern erweitern, als die Anteile für Kategorien unterschiedlich spezifiziert sein können. Nehmen wir an, eine Variable hätte drei Kategorien. Wir könnten prüfen, ob alle Kategorien zu je 1/3 vorkommen. Nun wollen wir den Fall untersuchen, ob zum Beispiel in der ersten Kategorie 50%, in der zweiten 35% und in der dritten 15% vorkommen. Die dabei gestellte statistische Frage ist, ob die Anteile von einzelnen Kategorien in einer Stichprobe den tatsächlichen Anteilen in der Population entsprechen. Wir wir sehen werden ist die statistische Methode zur Beantwortung solcher Fragen sehr ähnlich zum ersten Abschnitt. Eine typische Anwendung, wie sie in der Praxis häufig vorkommt, wird in Fallbeispiel 2 dargestellt.

Fallbeispiel 2: Repräsentativität einer Stichprobe

Eine Meinungsforscherin hat eine Telefonumfrage an 200 zufällig ausgewählten Telefonteilnehmern in Österreich zu Thema einer erwünschten Gesetzesmaßnahme zur speziellen Förderung von Familien mit Kindern durchgeführt. Aus Angaben der Statistik Austria für 2007 weiß sie, dass die insgesamt etwa 2,31 Millionen Familien sich folgendermaßen aufteilten: 31,2% Ehepaare ohne Kinder, 42,4% Ehepaare mit Kindern, 7,3% Lebensgemeinschaften ohne Kinder, 6,1% Lebensgemeinschaften mit Kindern und 13% alleinerziehende Elternteile. Die Meinungsforscherin interessierte, ob ihre Stichprobe repräsentativ bezüglich der Familienstruktur war, d. h., ob die Anteile der verschiedenen Arten von Familien in ihrer Telefon-Stichprobe mit den Anteilen aller österreichischen Familien übereinstimmen. In ihrer Stichprobe konnte sie folgende Häufigkeiten feststellen:

Familietyp	Häufigkeit	Prozent
Ehepaare ohne Kinder	42	21
Ehepaare mit Kindern	98	49
Lebensgemeinschaft ohne Kinder	6	03
Lebensgemeinschaft mit Kindern	20	10
Alleinerziehende Elternteile	34	17
Gesamt	200	

Ist die Stichprobe repräsentativ für die Population bezüglich der Familienstruktur?

Wie schon zuvor (Abschnitt 4.2) beginnen wir die Analyse mit einer Darstellung der Daten.

4.3.1 Numerische und grafische Beschreibung

Wir werden in PASW wieder eine Häufigkeitstabelle erstellen. Da hierzu aber kein Datenfile zur Verfügung steht, müssen wir ein solches selbst erzeugen. Wir öffnen also eine neue Datei über die Menüauswahl Datei, Neu und dann Daten. Es öffnet sich eine neue, noch leere Datenansicht im PASW-Daten-Editor. Zunächst wollen wir die Variable `Familientyp` definieren. Dazu tragen wir, nachdem wir in die Variablenansicht gewechselt haben, in der ersten Zeile unter Name `Familientyp` ein und änden noch das Messniveau auf Nominal. Als Nächstes sollten wir Wertelabels vergeben. Dies könnten wir unter der Spalte Wertelabels aber auch, etwas bequemer, über die Menüauswahl Daten folgendermaßen bewerkstelligen.

PASW

Daten
Variableneigenschaften definieren ... OK
Variablen (`Familientyp`) ➡ Zu durchsuchende Variablen ... Weiter
Eingabe der Wertelabels (▶ Abbildung 4.12)
OK

Im Dialogfenster **Variableneigenschaften definieren** tragen wir unter **Wert** die Kategoriennummern 1 bis 5 und unter **Variablenlabel** die entsprechenden Bezeichnungen ein (▶ Abbildung 4.12). (Achtung: **PASW** verwendet irrtümlicherweise den Begriff **Variablenlabel**, obwohl es eigentlich **Wertelabel** heißen müsste)

Prizipiell haben wir nun zwei Möglichkeiten, die Daten einzugeben. Wir könnten für alle 200 Personen die entsprechenden Kategorien, also den jeweils zutreffenden Familientyp eingeben. Dies entspräche der Datenstruktur wie in Abbildung 4.1. Allerdings wäre das relativ mühsam. Einfacher ist Folgendes.

Eingabe und Verwendung von Daten in Tabellenform

Wir tragen im **Daten-Editor** in der schon definierten Spalte **Familientyp** der Reihe nach die Zahlen 1 bis 5 ein. Daneben, in eine neue Spalte, schreiben wir zugehörigen Häufigkeiten, also 42, 98 etc., analog zu Fallbeispiel 2. In der **Variablenansicht** definieren wir in der zweiten Zeile den Namen dieser neuen Variable als `counts` und verwenden als Variablenlabel **Anzahl im Sample**. Die Daten sollten jetzt so aussehen (▶ Abbildung 4.13).

Wir müssen **PASW** nun mitteilen, dass die Kategorien 1 bis 5 nicht nur einmal vorkommen (wir haben ja nur 5 Zeilen in den Daten), sondern so oft, wie in der Spalte `counts` steht. Dies erreichen wir, indem wir die Fälle GEWICHTEN.

	Geändert	Fehlende Werte	Anzahl	Wert	Variablenlabel
1	☐	☐	1	1,00	Ehepaare ohne Kinder
2	☐	☐	1	2,00	Ehepaare mit Kindern
3	☐	☐	1	3,00	Lebensgemeinschaft ohne Kinder
4	☐	☐	1	4,00	Lebensgemeinschaft mit Kindern
5	☐	☐	1	5,00	Alleinerziehende Elternteile
6	☐	☐			

Abbildung 4.12: Ausschnitt des Dialogfensters zum Definieren von Wertelabels

	Familientyp	counts
1	Ehepaare ohne Kinder	42,00
2	Ehepaare mit Kindern	98,00
3	Lebensgemeinschaft ohne Kinder	6,00
4	Lebensgemeinschaft mit Kindern	20,00
5	Alleinerziehende Elternteile	34,00

Abbildung 4.13: PASW Statistics Daten-Editor: Ausschnitt der Familientypdaten

PASW

Daten
Fälle gewichten
Gewichtungsvariable (`counts`) ➡ Häufigkeitsvariable
OK

In der rechten unteren Ecke des **Daten-Editors** steht jetzt **Gewichtung aktiv**. Man sollte nie vergessen, die Gewichtung wieder zu deaktivieren, wenn man diese nicht mehr benötigt, da sonst alle weiteren Berechnungen sehr falsch werden können.

HINWEIS

Wir können nun die Häufigkeitstabelle für das Meinungsumfragenbeispiel (genauso wie im Abschnitt 4.2.1) erzeugen.

PASW

Analysieren
Deskriptive Statistiken
Häufigkeiten
Variable (`Familientyp`) ➡ Variablen
✓ Häufigkeitstabellen anzeigen
OK

► Abbildung 4.14 zeigt die Häufigkeitstabelle. Standardmäßig gibt es wieder die beiden hier unnötigen Spalten mit gültigen Prozent und kumulierten Prozent.

Familientyp

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Ehepaare ohne Kinder	42	21,0	21,0	21,0
	Ehepaare mit Kindern	98	49,0	49,0	70,0
	Lebensgemeinschaft ohne Kinder	6	3,0	3,0	73,0
	Lebensgemeinschaft mit Kindern	20	10,0	10,0	83,0
	Alleinerziehende Elternteile	34	17,0	17,0	100,0
	Gesamt	200	100,0	100,0	

Abbildung 4.14: Häufigkeitstabelle der Familientypen

Darstellung verschiedener Variablen in einer Tabelle

Die Tabelle in Abbildung 4.14 entspricht jener aus Fallbeispiel 2. Sie enthält zwar die gesamte Information zur Stichprobe, aber hinsichtlich der Fragestellung wäre es günstiger, auch noch die Zahlen aus der gesamten österreichischen Bevölkerung hinzuzufügen. Man bekäme dann gleich einen Eindruck, wie ähnlich oder unähnlich die Prozente aus der Population und der Stichprobe sind.

Hierzu müssen wir i) die Prozentwerte für die Stichprobe berechnen und ii) die Prozentwerte der Population zu den Daten hinzufügen.

Zum Berechnen der Prozentwerte für die Stichprobe wählen wir das Menü Transformieren und dann Variable berechnen. Wir erhalten ein Dialogfenster, in dem wir Folgendes eintragen (▶ Abbildung 4.15):

- In das Feld Zielvariable schreiben wir `s.perc` (das ist die neue Variable, in der die Prozentwerte aus der Stichprobe stehen sollen).
- In das Feld Numerischer Ausdruck setzen wir die Formel für die zu berechnenden Prozentwerte ein (`counts/200` gibt die relativen Häufigkeiten, 200 war ja die Stichprobengröße, und die Multiplikation mit 100 ergibt die Prozente).

Die berechneten Prozentwerte finden sich jetzt in der Datenansicht. Nun müssen wir noch die Prozentsätze aus der Population hinzufügen. Wir definieren eine Variable `p.perc` und tragen die Werte ein, wie sie im Fallbeispiel 2 angegeben sind. Danach vergeben wir in der Variablenansicht noch Variablenlabels für `s.perc`, nämlich Prozent Stichprobe, und für `p.perc` den Label Prozent Population. Die Daten sehen nun so aus (▶ Abbildung 4.16).

Die Tabelle, in der die Prozentsätze aus der Stichprobe und der Population einander gegenübergestellt werden, kann man folgendermaßen erzeugen.

PASW

Analysieren
  Tabellen
    Benutzerdefinierte Tabellen...
      erstes Fenster Benutzerdefinierte Tabellen *mit* [OK] *schließen*
      Einträge wie unten beschrieben (siehe auch Abbildung~4.17)
    [OK]

Im großen (zweiten) Dialogfenster Benutzerdefinierte Tabellen geht man so vor:

- Aus dem Feld Variablen (links oben) zieht man die kategoriale Variable (hier `Familientyp`) auf die Fläche Zeilen (in der mittleren Fläche).

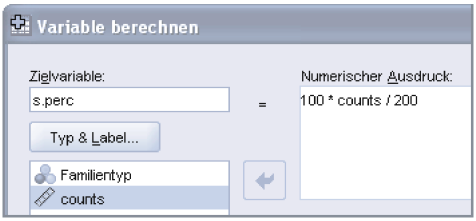


Abbildung 4.15: Ausschnitt aus dem Dialogfenster Variable berechnen. Es werden die Prozentsätze für die Familientypen aus der Stichprobe berechnet

	Familientyp	counts	s.perc	p.perc
1	Ehepaare ohne Kinder	42,00	21,00	31,20
2	Ehepaare mit Kindern	98,00	49,00	42,40
3	Lebensgemeinschaft ohne Kinder	6,00	3,00	7,30
4	Lebensgemeinschaft mit Kindern	20,00	10,00	6,10
5	Alleinerziehende Elternteile	34,00	17,00	13,00

Abbildung 4.16: PASW Statistics Daten-Editor: Ausschnitt der Familientypdaten nach Hinzufügen der Prozentwerte aus der Population

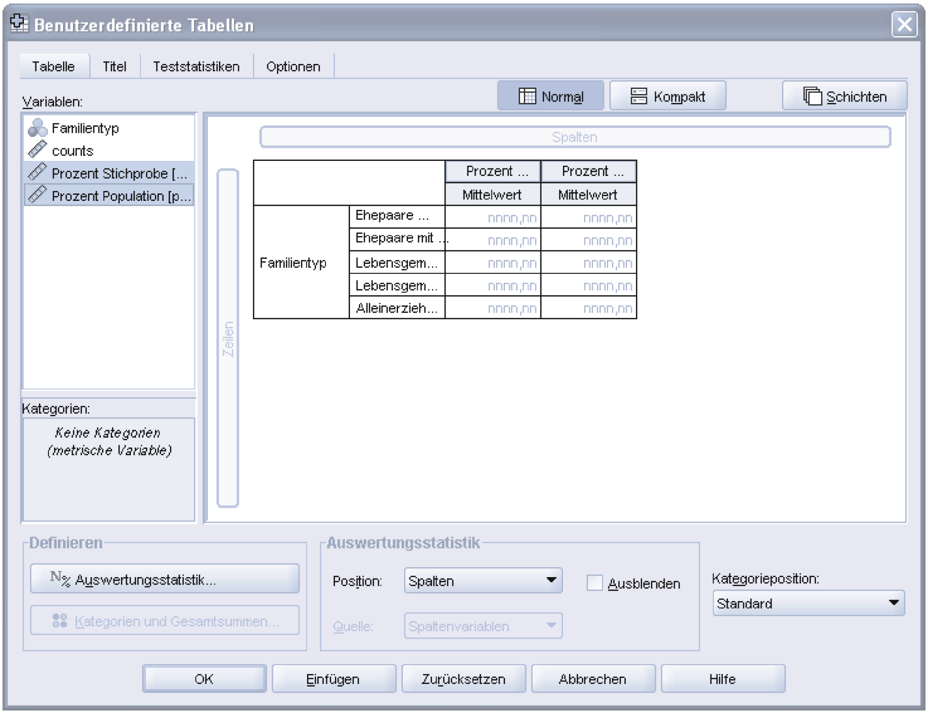


Abbildung 4.17: Dialogfenster Benutzerdefinierte Tabellen zur Erzeugung einer Tabelle aus zwei verschiedenen Variablen

- Man markiert die beiden Variablen, für die man die Prozentsätze darstellen will (hier `s.perc` und `p.perc`), wobei man beim Anklicken die Strg-Taste gedrückt hält.
- Diese zieht man dann auf die Fläche Spalten.

Das Resultat ist die Tabelle in ► Abbildung 4.18, in der man die Prozentsätze für die Familienarten aus der Stichprobe der Meinungsforscherin und der österreichischen Bevölkerung ablesen kann.

Man erkennt, dass die Prozentsätze differieren. Es fällt auf, dass generell Familien mit Kindern stärker in der Stichprobe vertreten sind als jene ohne Kinder. Da die Umfrage zum Thema einer Gesetzesmaßnahme zur speziellen Förderung von Familien mit Kindern stattfand, könnte eventuell die Bereitschaft zur Beantwortung bei

Tabelle 1

		Prozent Stichprobe	Prozent Population
		Mittelwert	Mittelwert
Familientyp	Ehepaare ohne Kinder	21,00	31,20
	Ehepaare mit Kindern	49,00	42,40
	Lebensgemeinschaft ohne Kinder	3,00	7,30
	Lebensgemeinschaft mit Kindern	10,00	6,10
	Alleinerziehende Elternteile	17,00	13,00

Abbildung 4.18: Tabelle mit Prozentsätzen der Familienarten aus einer Stichprobe und der Population

solchen Personen größer gewesen sein, die in einer Familie mit Kindern leben. Auch die Wahrscheinlichkeit, solche Personen eher zu Hause am Festnetz zu erreichen, könnte eine Rolle gespielt haben.

Die Tabelle (Abbildung 4.18) hat einen kleinen Schönheitfehler, nämlich die Zeile, in der Mittelwert steht. Diese lässt sich leicht entfernen (Abschnitt 3.9.1), allerdings sollten wir damit warten, bis wir die Grafiken erzeugt haben, da PASW die Mittelwertsfunktion dafür benötigt.

Darstellung verschiedener Variablen in einer Grafik

Zur grafischen Beschreibung der Daten aus Fallbeispiel 2 bieten sich GRUPPIERTE BALKENDIAGRAMME an. Sie sind eine Erweiterung der Balkendiagramme aus Abschnitt 4.2.2. Bei gruppierten Balkendiagrammen werden mehrere kategoriale Variablen gleichzeitig oder eine kategoriale Variable aufgeschlüsselt nach verschiedenen Gruppen dargestellt (im Detail gehen wir darauf in Kapitel 7 ein).

PASW

Diagramme
Diagrammerstellung... OK
in Galerie *auswählen von* Balken
Doppelklick auf Galeriediagramm Gruppierte Balken
(die Vorschau erscheint in der Zeichenfläche)
Variable `Familientyp` *in das Feld* X-Achse? *ziehen*
Variablen `s.perc` *und* `p.perc`*gemeinsam nach* Anzahl *ziehen*
OK *(Dialogfeld Zusammenfassung erstellen ignorieren)*
Elementeigenschaften
unter Eigenschaften bearbeiten von *wählen von* X-Achse1
unter Kleine/leere Kategorien *wählen von*
Nur in den Daten vorhandene Kategorien anzeigen
Zuweisen, dann Schließen
Grundelemente
Transponieren
OK

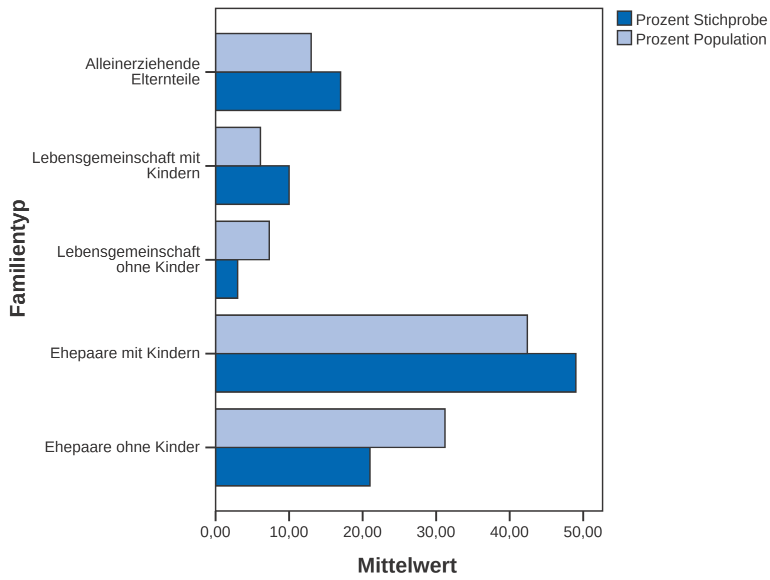


Abbildung 4.19: Transponiertes und umgruppiertes Balkendiagramm mit Prozentsätzen der Familienarten aus einer Stichprobe und der Population

Das so hergestellte gruppierte Balkendiagramm findet sich in ► Abbildung 4.19. Es ist natürlich in dieser Form nicht geeignet für Publikationen, Berichte oder Präsentationen. Durch Verwendung des Diagramm-Editors kann man aber alle notwendig erscheinenden Modifikationen leicht durchführen (siehe Kapitel 1).

Abbildung 4.19 ist zum Vergleich der Prozentsätze aus der Stichprobe und der Population gut geeignet. Man kann sehr schön die Unterschiede und deren Größenordnung erkennen. Es bekräftigt sich der Eindruck, den wir schon aus der Tabelle in Abbildung 4.18 gewonnen haben. In der Stichprobe sind jene Befragten überproportional vertreten, die in Familien mit Kindern leben. Gegenteiliges gilt für Ehepaare ohne Kinder, die in der Stichprobe unterrepräsentiert sind. Die Frage ist jetzt, sind diese Unterschiede bedeutsam oder nur auf Zufall zurückzuführen.

4.3.2 Statistische Analyse der Problemstellung

Ganz ähnliche Überlegungen, wie wir sie schon im Abschnitt 4.2.3 angestellt haben, treffen auch hier zu. Allerdings ist die Annahme jetzt nicht, dass alle Kategorien gleich häufig vorkommen (Nullhypothese), sondern dass sie in bestimmter Weise festgelegt sind. In unserem Beispiel, der Frage nach Repräsentativität der Stichprobe, sind die festgelegten Anteile jene der Familientypen in der Population. Wir können zwar die Formel (4.1) verwenden, aber wir müssen die erwarteten Häufigkeiten anders bestimmen.

Wir kennen die Anteile in der Population, nämlich 0,312 für Ehepaare ohne Kinder, 0,424 für Ehepaare mit Kindern etc. Wir können aus diesen jene Häufigkeiten bestimmen, die wir erwarten würden, wenn in der Stichprobe der 200 Personen die

Stichprobenhäufigkeiten mit den Populationswerten übereinstimmen würden. Wir müssen dazu nur die Populationsanteile mit der Stichprobengröße 200 multiplizieren. Wenn wir eine bezüglich der Familienstruktur repräsentative Stichprobe hätten, dann müssten $200 \times 0,312 = 62,4$, also 62 Befragte der Kategorie Ehepaar kinderlos angehören. Allgemein gilt

Berechnung von erwarteten Häufigkeiten

$$e_j = n\pi_j \tag{4.2}$$

- $e_j \ldots$ erwartete Häufigkeit für die Kategorie j
- $n \ldots$ Stichprobengröße
- $\pi_j \ldots$ relative Häufigkeit in der Population (Wahrscheinlichkeit für Kategorie j)

Die weiteren Schritte sind analog zu Abschnitt 4.2.3. Wenn die Abweichungen zwischen beobachteten und erwarteten Häufigkeiten zu groß werden, verwerfen wir die Nullhypothese, dass die beobachteten Werte mit den erwarteten übereinstimmen. Die Abweichungen werden dann nicht mehr dem Zufall zugeschrieben, sondern man glaubt an einen systematischen Unterschied.

Die Berechnung des Tests in PASW erfolgt über

PASW

Analysieren
  Nichtparametrische Tests
    Chi-Quadrat...
    Variable (`Familientyp`) [➡] Testvariablen
      Erwartete Werte: Werte *markieren*
      in das Feld neben Werte *der Reihe nach:*
        die erwarteten relativen Häufigkeiten einfügen
        durch [Hinzufügen] *in das große weiße Feld bringen*
          (► Abbildung 4.20)
    [OK]

Das Ergebnis des Chi-Quadrat-Tests ist in ► Abbildung 4.21. In der oberen Tabelle werden die beobachteten und erwarteten Häufigkeiten dargestellt.

Die untere Tabelle gibt uns das Resultat des Tests. Wie wir am Signifikanzwert sehen, ist die Plausibilität der Nullhypothese sehr klein (der p-Wert ist zumindest kleiner als 0,001, ganz null kann er nie werden). Daher verwerfen wir die Annahme, dass die Stichprobenanteile der einzelnen Familientyp-Kategorien jenen in der Population entsprechen.

Abbildung 4.20: Chi-Quadrat-Test: Variableauswahl und Spezifikation der erwarteten Anteile

Familientyp

	Beobachtetes N	Erwartete Anzahl	Residuum
Ehepaare ohne Kinder	42	62,4	-20,4
Ehepaare mit Kindern	98	84,8	13,2
Lebensgemeinschaft ohne Kinder	6	14,6	-8,6
Lebensgemeinschaft mit Kindern	20	12,2	7,8
Alleinerziehende Elternteile	34	26,0	8,0
Gesamt	200		

Statistik für Test

	Variablen
Statistik	Familientyp
Chi-Quadrat	21,238
df	4
Asymptotische Signifikanz	,000

Abbildung 4.21: Ergebnis des Chi-Quadrat-Tests zur Prüfung der Repräsentativität einer Stichprobe

Fallbeispiel 2: Interpretation

Der Chi-Quadrat-Test zur Überprüfung der Nullhypothese, dass die Häufigkeiten für die Familienarten in der Stichprobe jenen in der Population entsprechen, zeigt ein signifikantes Ergebnis ($X^2 = 21{,}2$, $df = 4$, $p < 0{,}001$). Demnach ist nicht davon auszugehen, dass die Stichprobe bezüglich der Familienarten repräsentativ ist. Familien mit Kindern sind stärker in der Stichprobe vertreten als in der Population.

4.4 Hat ein Prozentsatz (Anteil) einen bestimmten Wert?

Bisher haben wir untersucht, wie man analysieren kann, ob Häufigkeiten bzw. Anteile mehrerer Kategorien bestimmten Vorgaben entsprechen. Jetzt wollen wir uns auf einzelne Kategorien konzentrieren.

Fallbeispiel 3: Fehlerrate bei Lügendetektoren

Datenfile: `detektor.sav`

Nach wie vor gibt es in den USA Bemühungen, den perfekten Lügendetektor zu entwickeln. Neuere Ansätze stammen von Pavlidis et al. (2002), die versuchten, mit einer hochauflösenden, temperatursensiblen Kamera aus Gesichtsaufnahmen Lügen zu entdecken. Rosenfeld (2002) verwendete sogenannte ERPs (ereigniskorrelierte Potentiale), bestimmte Gehirnaktivitätssignale, die mittels an der Kopfhaut angebrachten Elektroden gemessen werden. Er untersuchte Studierende, die unter anderem Sätze, wie *Verwenden Sie einen gefälschten Ausweis?* vorlesen mussten. Es wurde erwartet, dass bei Studierenden, die tatsächlich einen gefälschten Ausweis verwenden, ein entsprechendes Hirnsignal auftritt. Von insgesamt $N = 17$ „Schuldigen“ wurden 13 (77%) richtig erkannt. Bei einer vergleichbaren Studie des amerikanischen Verteidigungsministeriums wurden 75% der „Schuldigen“ mithilfe eines traditionellen klassischen Lügendetektors (Polygraphen) richtig erkannt.

Liefert der neue Lügendetektor bessere Ergebnisse als der traditionelle?

Wenn man die Daten einer einzelnen Kategorie analysieren möchte, dann kann man im Wesentlichen die gleichen Methoden verwenden wie in den vorhergehenden Abschnitt 4.2.3 und Abschnitt 4.3.2. Man muss sich nur vor Augen halten, dass es für diese einzelne Kategorie eigentlich zwei Möglichkeiten gibt, nämlich *trifft zu* bzw. *trifft nicht zu*. Wenn man also eine einzelne Kategorie untersucht, dann verhält sich diese wie eine „neue“ Variable mit zwei Kategorien.

Numerische und grafische Beschreibung

Zur numerischen Beschreibung wird es wohl genügen, eine Häufigkeitstabelle über die Menüpunkte Analysieren ▷ Deskriptive Statistiken ▷ Häufigkeiten wie in Abschnitt 4.2.1 zu erstellen.

Als grafische Darstellung bietet sich wieder ein Balkendiagramm an. In Analogie zu Abschnitt 4.2.2 kann es folgendermaßen erstellt werden.

PASW

Diagramme
Diagrammerstellung...
in Galerie *auswählen von* Balken
Doppelklick auf linkes oberes Galeriediagramm [einfache Balken]
(die Vorschau erscheint in der Zeichenfläche)
Variable `ueberfuehrt` *in das Feld* X-Achse? *ziehen*
[Elementeigenschaften]
unter Eigenschaften bearbeiten von *auswählen von* Balken
unter Statistiken *auswählen von* Prozentsatz
unter Eigenschaften bearbeiten von *auswählen von* X-Achse1
unter Kleine/leere Kategorien *wählen von*
Nur in den Daten vorhandene Kategorien anzeigen
[Zuweisen], dann [Schließen]
[OK]

Das Resultat findet sich in ► Abbildung 4.22.

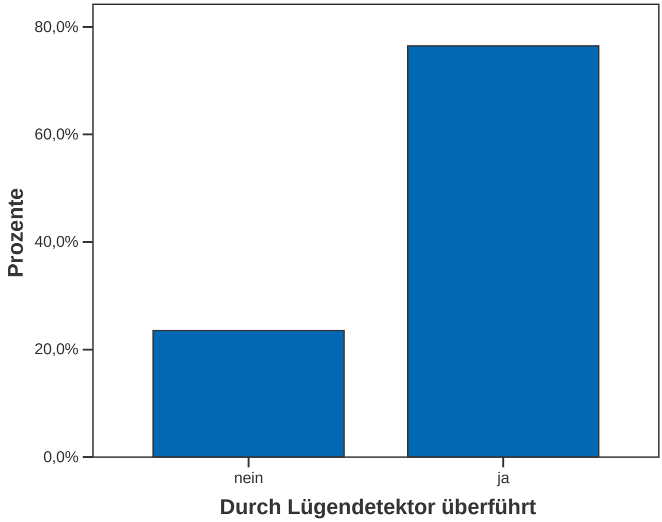


Abbildung 4.22: Balkendiagramm für Erfolg mit neuem Lügendetektor

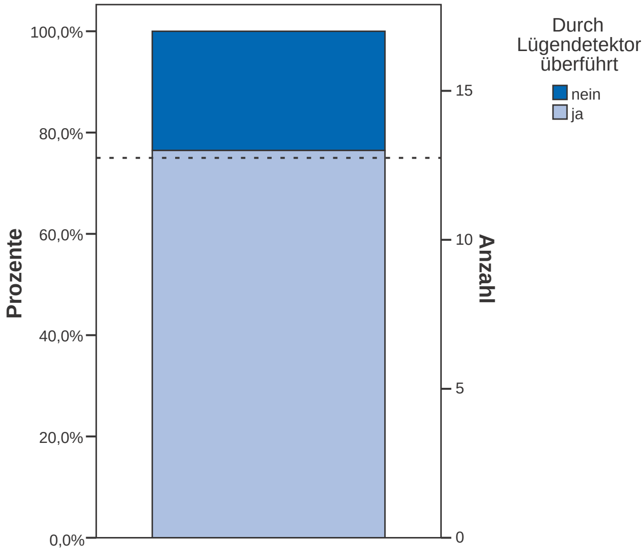


Abbildung 4.23: Gestapeltes Balkendiagramm mit zweiter Achse und Referenzlinie

Allerdings interessiert für die Fragestellung eigentlich nur die Kategorie `ja` (überführt). Dazu könnte man ein gestapeltes Balkendiagramm verwenden (diese werden in Abschnitt 5.1.2 genauer erklärt). Und man möchte vielleicht auch noch die 75%-Erfolgsrate der traditionellen Methode darstellen. Das Resultat könnte so aussehen (▶ Abbildung 4.23).

Man sieht, dass die Erfolgsrate des neuen Lügendetektors nur unwesentlich höher als jene traditioneller Detektoren ist. Die entsprechende statistische Methode, im nächsten Abschnitt, wird uns darüber Aufschluss geben, ob die Erfolgsrate tatsächlich höher ist oder nicht.

Die Grafik in Abbildung 4.23 ist keine Standard-PASW-Grafik und ihre Erstellung ist ein wenig komplizierter als das bisher Beschriebene. (Versuchen Sie, diese Abbildung durch Anwenden und Ausprobieren der entsprechenden Bedienelemente im Grafik-Editor zu erstellen. Der Aufwand lohnt sich.)

4.4.1 Statistische Analyse der Problemstellung

Die Frage, die wir untersuchen wollen, lautet: Liefert der neue Lügendetektor bessere Ergebnisse als der traditionelle? Übersetzt in eine statistische Fragestellung, könnte man auch so formulieren: Ist die relative Häufigkeit des Erfolgs des neuen Lügendetektors (in Wirklichkeit) höher als die relative Häufigkeit des Erfolgs nach der traditionellen Methode.

Als statistische Hypothesen formuliert:

- Nullhypothese H_0: $\pi = 0,75$
  Die relative Häufigkeit des Erfolgs nach der neuen Methode, π, entspricht der relativen Häufigkeit des Erfolgs nach der traditionellen Methode ($\pi_0 = 0,75$).
- Alternativhypothese H_A: $\pi > 0,75$

Die relative Häufigkeit des Erfolgs nach der neuen Methode, π, ist größer als die relative Häufigkeit des Erfolgs nach der traditionellen Methode ($\pi_0 = 0,75$).

Allgemein würde man schreiben:

Hypothesen beim Test eines Anteils

H_0: $\pi = \pi_0$ (Nullhypothese)

H_A: $\pi > \pi_0$ (Alternativhypothese)

- π ... die unbekannte wirkliche relative Häufigkeit des Erfolgs mit der neuen Methode
  Da wir sie nicht kennen, aber etwas über sie wissen wollen, verwenden wir für sie r, die beobachtete relative Häufigkeit aus der Stichprobe.
- π_0 ... der Wert, den wir kennen (oder den wir festlegen) und gegen den wir prüfen wollen, in unserem Beispiel ist er 0,75.

Die statistische Problemstellung ist von der Idee her gleich, wie wir sie schon in Abschnitt 4.2.3 kennengelernt haben:

- Wir gehen davon aus, dass in Wirklichkeit die relative Häufigkeit des Erfolgs nach der neuen Methode jener der traditionellen Methode entspricht, dass also tatsächlich H_0 gilt.
- Wir sammeln Daten aus einer Stichprobe (nach der neuen Methode), berechnen die relative Häufigkeit r (sie ist die beste Information, die wir für das unbekannte π haben) und vergleichen sie mit jener Zahl, die wir kennen, nämlich 0,75. Allgemein können wir π_0 statt 0,75 einsetzen.
- Da wir ja eine Zufallsstichprobe verwenden, wird r nicht genauso groß wie π sein, sondern ein wenig abweichen. Wenn die Abweichung aber zu groß wird, dann werden wir nicht glauben, dass H_0, sondern eher dass H_A gilt.
- Die Prüfmethode zur Entscheidung, welche der beiden Hypothesen zutrifft, nennen wir statistischen Test.

Im Unterschied zum Chi-Quadrat-Test aus Abschnitt 4.2.3 und Abschnitt 4.3.2, wo wir absolute Häufigkeiten (beobachtete und erwartete) verglichen haben, beschäftigen wir uns jetzt mit Anteilen oder relativen Häufigkeiten (auch beobachtete, nämlich r, und erwartete, nämlich π_0). Der Test, den wir jetzt verwenden werden, heißt EIN-STICHPROBEN-TEST FÜR ANTEILE bzw. BINOMIAL-TEST.

In PASW führen wir diesen Test folgendermaßen durch:

PASW

Analysieren
  Nichtparametrische Tests
    Binomial
      Variable (`ueberfuehrt`) ➡ Testvariablen
      unter Testanteil *eintragen von* `0,75` *(Nullhypothese)*
      [OK]

Das Ergebnis ist in ▶ Abbildung 4.24 dargestellt. Neben den absoluten (N) finden wir die relativen Häufigkeiten (Beobachteter Anteil), den Wert, den wir für die H_0 spezifiziert haben (Testanteil) sowie den *p*-Wert (Exakte Signifikanz 1-seitig). Dieser ist 0,574 und damit wesentlich größer als 0,05. Wir behalten daher die Nullhypothese bei. Die Erfolgsrate beim neuen Lügendetektor ist demnach nicht höher als beim traditionellen Lügendetektor.

Fallbeispiel 3: Interpretation

Der Binomial-Test zur Überprüfung der Nullhypothese, dass der Erfolgsanteil beim neuen größer ist als 75%, ergab kein signifikantes Ergebnis ($p = 0,54$). Demnach könnte nicht gezeigt werden, dass mit dem neuen Detektor Lügner in höherem Ausmaß überführt werden können als nach der traditionellen Methode.

Was bedeutet Exakte Signifikanz (1-seitig)? Der Begriff *einseitig* wird in Exkurs 4.5 besprochen. Exakt bedeutet vereinfacht gesagt, dass PASW statistische Methoden verwendet, bei denen der *p*-Wert (Signifikanz) für einen Test genau berechnet wird. Das ist aber nicht immer möglich, meist wenn eine Stichprobe groß ist. In diesem Fall werden Methoden verwendet, die den *p*-Wert möglichst genau approximieren. Je größer die Stichprobe ist, umso besser ist die Approximation. Die Approximationen beruhen auf der Grundidee, die zugrunde liegenden Wahrscheinlichkeiten so zu berechnen, als ob man unendlich viele Stichproben gezogen hätte. PASW nennt dann das Ergebnis Asymptotische Signifikanz, wie wir es schon beim Chi-Quadrat-Test gesehen haben (z. B. Abbildung 4.20).

Im Prinzip genügt es, in den PASW-Output-Tabellen nach Signifikanz zu suchen und dort den *p*-Wert abzulesen. Dadurch bekommt man die notwendige Information über die Plausibilität der Nullhypothese (und der Alternativhypothese). Natürlich setzt das voraus, dass man die Nullhypothese und die Alternativhypothese kennt. Aber ohne diese Kenntnis wüsste man ja die zugrunde liegende Fragestellung nicht und man hätte ja auch nicht die entsprechende statistische Methode gewählt, für die der Output erzeugt wurde.

Test auf Binomialverteilung

		Kategorie	N	Beobachteter Anteil	Testanteil	Exakte Signifikanz (1-seitig)
Durch Lügendetektor überführt	Gruppe 1	ja	13	,76	,75	,574
	Gruppe 2	nein	4	,24		
	Gesamt		17	1,00		

Abbildung 4.24: Ergebnis des Binomial-Tests für die Erfolgsquote des neuen Lügendetektors

Exkurs 4.5 Einseitige und zweiseitige Alternativhypothesen

Die Nullhypothese ist der Ausgangspunkt eines statistischen Tests. Sie legt die Annahme fest, gegen die wir erhobene Daten prüfen wollen. Meistens ist das Ziel einer Analyse, die Nullhypothese zu verwerfen. Die eigentlich interessierende Fragestellung wird als Alternativhypothese formuliert. Es werden zwei Arten von Alternativhypothesen unterschieden:

Einseitige Alternativhypothesen

Bei einseitigen Alternativhypothesen geht die Frage immer in eine bestimmte Richtung:

- $H_A: \pi > \pi_0$
  Die Erfolgsrate (π) der neuen Methode ist größer als erwartet (π_0).

Diese Frage würde z. B. jemand stellen, der Argumente für die Überlegenheit der neuen Methode sucht.

- $H_A: \pi < \pi_0$
  Die Erfolgsrate der neuen Methode ist kleiner als erwartet.

Diese Festlegung könnte von einem Zweifler an der neuen Methode stammen. In beiden Fällen setzt die einseitige Formulierung der Alternativhypothese ein gewisses Vorwissen oder eine bestimmte Absicht voraus. Oft weiß man aber nicht, in welche Richtung die Sache läuft. Es ist dann besser, die Alternativhypothese zweiseitig zu formulieren:

Zweiseitige Alternativhypothesen

- $H_A: \pi \neq \pi_0$
  Die Erfolgsrate der neuen Methode ist anders als man erwarten würde, sie kann größer, aber auch kleiner sein.

Es hängt von der Fragestellung ab, welche der drei Möglichkeiten man wählt. Man muss sich aber für eine entscheiden. Keinesfalls sollte man alle drei prüfen.

Was macht man, wenn PASW nur die einseitige Signifikanz (p-Wert) ausgibt, man aber zweiseitig testen möchte? In diesem Fall muss man den p-Wert verdoppeln, bevor man ihn mit dem Signifikanzniveau, üblicherweise 0,05, vergleicht.

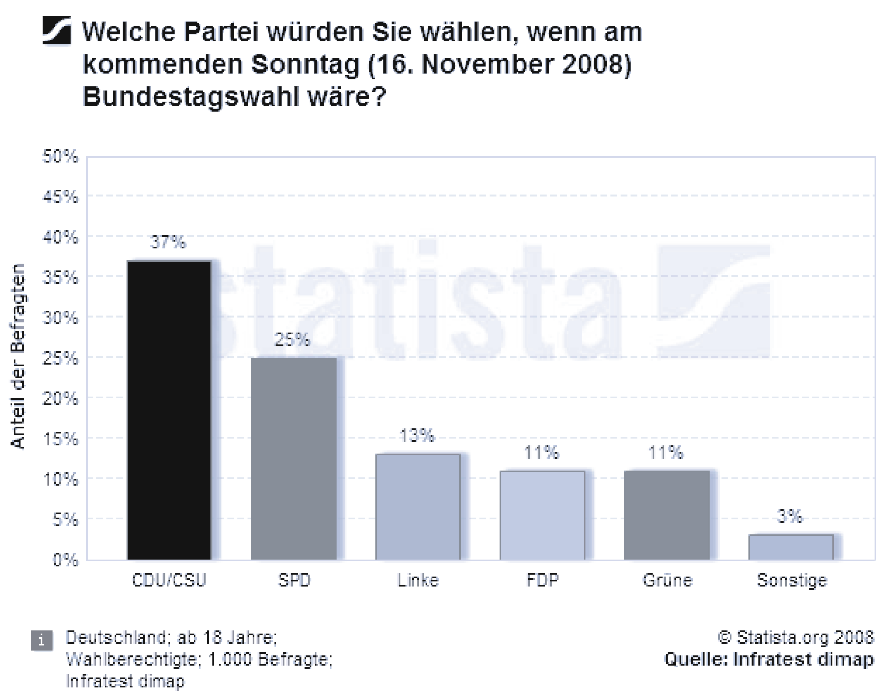
Was macht man wenn PASW nur die zweiseitige Signifikanz (p-Wert) ausgibt, man aber einseitig testen möchte? Dann muss man den p-Wert halbieren. Aber Achtung, das gilt nur dann, wenn die Daten der Alternativhypothese entsprechen, also wenn z. B. $H_A: \pi > \pi_0$ und das r in den Daten tatsächlich größer als π_0 ist.

Was macht man, wenn r kleiner als π_0 ist, man aber als Alternativhypothese $H_A: \pi > \pi_0$ prüfen wollte? In diesem Fall wird die Nullhypothese immer beibehalten. Auf keinen Fall sollte man nachträglich die Hypothesen umformulieren.

4.5 In welchem Bereich kann man einen Prozentsatz (Anteil) erwarten?

Fallbeispiel 4: Die Sonntagsfrage

Auf der Webseite http://www.statista.org fand sich im Herbst 2008 das Ergebnis einer Meinungsumfrage, in der die Stimmungslage der deutschen Bevölkerung zu den politischen Parteien erhoben wurde.



Wie wird die nächste Bundestagswahl ausgehen?

Oft findet man in den Medien Grafiken oder Angaben im Text, wo Prozentwerte zu irgendeinem Sachververhalt berichtet werden. Zum Beispiel: „Eine Studie der IASO (International Association for the Study of Obesity) für 2007 ergab, dass dreiviertel der Männer und 59% der Frauen in Deutschland übergewichig sind.“ Diese Zahlen werden so dargestellt, als ob sie Fakten wären. Nicht berichtet wird aber, ob die Daten aus einer Stichprobe stammen und wie groß diese war. Es ist nicht anzunehmen, dass man das Gewicht aller Deutschen gewogen hat. Aber wie verlässlich sind solche Zahlen? Wir wollen diese Problematik anhand eines prominenten Beispiels, der „Sontagsfrage“ im Fallbeispiel 4, diskutieren.

Das Ergebnis der Umfrage aus Fallbeispiel 4 ergab, dass CDU/CSU mit 37% gegenüber der SPD mit 25% vorne liegt und Linke mit 13%, FDP und Grüne mit jeweils 11% nahezu gleich viel Zustimmung erhielten. Alle anderen Gruppierungen kommen zusammen auf 3%. Was bedeuten diese Zahlen? Es gibt einige Gründe, warum man diese Zahlen nicht als bare Münze nehmen sollte. Erstens sind sie nicht das Ergebnis einer Wahl, sondern einer Umfrage, die Monate vor der nächsten Wahl stattgefunden hat. Es kann sich also noch einges ändern. Zweitens beinhalten Umfragen immer gewisse Unschärfen. Wichtige Ursachen für Ungenauigkeiten liegen unter anderem im Anteil an Unentschlossenen, Nicht- und Ungültig-Wählern und solchen, die nicht ehrlich antworten. Der wichtigste Grund aber, warum wir nicht erwarten dürfen, dass diese Zahlen genau stimmen, liegt darin, dass nur eine Stichprobe befragt wurde. Selbst unter der Annahme, dass die Stichprobe repräsentativ ist (dass sie also ein wirkliches Abbild der deutschen Wähler darstellt), alle ehrlich geantwortet haben und niemand mehr seine Meinung ändert, muss man mit gewissen Ungenauigkeiten rechnen, wenn man das Wahlergebnis vorhersagen wollte.

Der Grund hierfür sind Stichprobenschwankungen. Hätten die Meinungsforscher andere 1000 Personen ausgewählt und befragt, wäre das Ergebnis sicherlich ein anderes gewesen. Die gleichen Überlegungen haben wir schon angestellt, als wir uns fragten, ob alle Seiten eines Würfels gleich häufig auftreten. Bei wenigen Versuchen können wir nicht erwarten, dass alle Seiten gleich häufig gewürfelt werden, auch wenn der verwendete Würfel fair ist. Genauso verhält es sich mit Stichproben bei Meinungsumfragen. Es werden nicht alle Parteipräferenzen mit dem gleichen Prozentsatz wie in der Bevölkerung vorkommen. Allerdings, wie auch beim Würfeln, je mehr Personen man befragt, je größer also die Stichprobe ist, umso genauere Ergebnisse wird man bekommen. Wir benötigen also eine Methode, um bestimmen zu können, wie genau das Ergebnis einer Meinungsumfrage ist. Diese Methode nennt man Bestimmen von KONFIDENZINTERVALLEN oder VERTRAUENSBEREICHEN. In den Medien findet man auch oft den Begriff SCHWANKUNGSBREITE. Hinter dieser Methode stecken einige nicht ganz einfache mathematisch-statistische Überlegungen (ähnlich wie in Appendix 4.8), allerdings kann man Schwankungsbreiten relativ leicht ausrechnen.

Berechnung von Schwankungsbreiten für Anteile

$$c = 1,96 \cdot \sqrt{\frac{r_j(1-r_j)}{n}} \qquad (4.3)$$

- $c \ldots$ Schwankungsbreite
- $r_j \ldots$ relative Häufigkeit
- $n \ldots$ Stichprobengröße (Gesamtanzahl von Befragten)

Eigentlich müsste man in der obigen Formel c_j statt c schreiben, da die Schwankungsbreite von der Größe der relativen Häufigkeit für die jeweilige Kategorie j (hier Partei) abhängt. Am größten ist die Schwankungsbreite, wenn eine Kategorie einen Anteil von 50% hat. Ein Konfidenzintervall (Vertrauensbereich) besteht aus zwei Zahlen, die einen Bereich festlegen. Die obere Zahl (Grenze) erhält man, wenn man die Schwankungsbreite c zu der relativen Häufigkeit für die jeweilige Kategorie dazuzählt, die untere Zahl, wenn man sie abzieht.

Konfidenzintervall (Vertrauensbereich) für Anteile

$$\text{Konfidenzintervall: } [\, r_j - c\,;\; r_j + c\,] \tag{4.4}$$

Welche Bedeutung haben nun Schwankungsbreiten? Sie geben an, wie stark die aus einer Stichprobe errechneten Prozentsätze aufgrund von Zufallseinflüssen schwanken können. Und Konfidenzintervalle? Sie geben mit einer bestimmten Plausibilität (oder Sicherheit) an, zwischen welchen Grenzen der tatsächliche Anteil in der Population zu erwarten ist. Ein Wert von 1,96 (wie in Formel 4.3) würde 95%-ige Sicherheit bedeuten, würde man stattdessen 2,58 verwenden, wäre die Sicherheit 99%, das Intervall wäre dann aber auch größer.

Wie würden die Konfidenzintervalle für die Parteienzustimmung aussehen?

In PASW gibt es leider keine direkte Möglichkeit, Konfidenzintervalle für Prozentsätze auszurechnen, aber man kann das ganz leicht „händisch" bewerkstelligen. Wir erzeugen zunächst die notwendigen Daten und gehen dabei genauso vor wie in Abschnitt 4.3.1. Im Daten-Editor sollte dann Folgendes zu sehen sein (▶ Abbildung 4.25).

Nun müssen wir die oberen und unteren Grenzen berechnen. Wir benutzen die Menüs Transformieren ▷ Variable berechnen und machen die Einträge wie in ▶ Abbildung 4.26, wobei wir die Formeln 4.3 und 4.4 verwenden. Die Berechnung wird dann mit OK abgeschlossen.

Die untere Grenze nennen wir `unt.Grenze`, SQRT(...) ist die Funktion (*square root*), mit der man eine Wurzel berechnen kann. Nun benötigen wir noch die obere Grenze. Dazu

- ersetzen wir im Dialogfenster Variable berechnen `unt.Grenze` durch `ob.Grenze` (PASW merkt sich den letzten Eintrag, wir müssen nicht alles neu eintippen)
- und schreiben + statt − im rechten Feld.

Im Daten-Editor können wir das Resultat schon sehen. Zur Erzeugung eines Tabellen-Outputs gehen wir genauso vor wie in Abschnitt 4.3 auf Seite 118. Man erhält die Tabelle mit den Konfidenzintervallen (▶ Abbildung 4.27), aus der wir noch die Zeile mit Mittelwert entfernt haben.

Zur grafischen Darstellung stellt PASW den Typ der sogenannten Hoch-Tief-Diagramme zur Verfügung, der alle Diagramme umfasst, für die ein Datenbereich zwischen zwei Werten angezeigt werden soll. Wir wollen ein Hoch-Tief-Schluss-Diagramm verwenden.

	Partei	Anteil
1	CDU/CSU	0,37
2	SPD	0,25
3	Linke	0,13
4	FDP	0,11
5	Grüne	0,11

Abbildung 4.25: Sonntagsfrage: Anteile der Parteienpräferenz im Daten-Editor

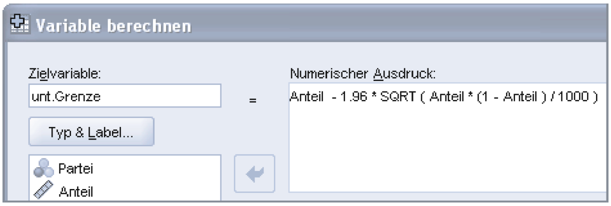


Abbildung 4.26: Berechnung der unteren Grenze von Konfidenzintervallen für Anteile der Parteienpräferenz

Tabelle 1

		unt.Grenze	ob.Grenze
Partei	CDU/CSU	,34	,40
	SPD	,22	,28
	Linke	,11	,15
	FDP	,09	,13
	Grüne	,09	,13

Abbildung 4.27: Konfidenzintervallen für Anteile der Parteienpräferenz

PASW

Diagramme
Diagrammerstellung... [OK]
in Galerie *auswählen von* Hoch-Tief
Doppelklick auf linkes oberes Galeriediagramm [Hoch-Tief-Schluss]
(die Vorschau erscheint in der Zeichenfläche)
Variable `Partei` *in das Feld* X-Achse? *ziehen*
Variable `ob.Grenze` *in das Feld* Hoch-Variable? *ziehen*
Variable `unt.Grenze` *in das Feld* Tief-Variable? *ziehen*
Variable `Anteil` *in das Feld* Schluss-Variable *ziehen*
[Elementeigenschaften]
unter Eigenschaften bearbeiten von *wählen von* X-Achse1
unter Kleine/leere Kategorien *wählen von*
Nur in den Daten vorhandene Kategorien anzeigen
[Zuweisen]
unter Eigenschaften bearbeiten von *wählen von*
Hoch-Tief-Schluss1
unter Statistiken *auswählen von* Wert ... [OK]
[Zuweisen], dann [Schließen]
[OK]

Wir erhalten das Diagramm, das uns die Konfidenzintervalle für die fünf Parteien darstellt (► Abbildung 4.28).

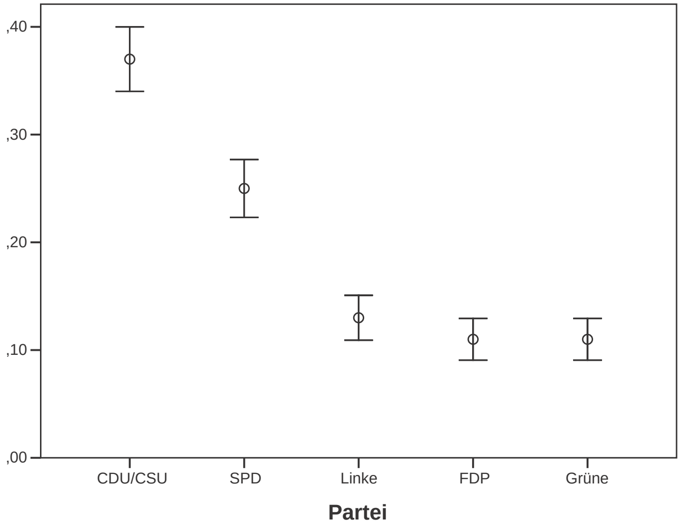


Abbildung 4.28: Hoch-Tief-Schluss-Diagramm: Konfidenzintervalle für Anteile der Parteienpräferenz

Bei Betrachten der Konfidenzintervalle in Tabelle 4.27 und Abbildung 4.28 fallen einige Dinge auf. Die untere Grenze die für CDU/CSU liegt höher als die obere Grenze für die SPD. Das bedeutet, dass die CDU/CSU eindeutig vor der SPD liegt, die Intervalle überlappen sich nicht. Genauso eindeutig liegt die SPD vor den anderen Parteien. Für diese kann man aber nicht sagen, welche bevorzugter sind. Die Linken liegen zwar um zwei Prozentpunkte vor den Grünen und der FPD, allerdings überlappen sich die Konfidenzintervalle hier und dieser scheinbare Vorzug in der Wählergunst kann sich auf Zufall (durch die spezifische Stichprobe bedingt) zurückführen lassen.

Fallbeispiel 4: Interpretation

Die Konfidenzintervalle für die Parteienzustimmung ergaben, dass die CDU/CSU klar vorne liegt. An zweiter Stelle folgt die SPD, die wiederum klar vor den anderen Parteien positioniert ist. Zwischen Linken, Grünen und FDP gibt es keine Unterschiede.

4.6 Zusammenfassung der Konzepte

Ein kategoriale Variable entsteht, indem man Beobachtungen nach bestimmten Kategorien klassifiziert.

- Häufigkeiten erhält man durch Auszählen, wie oft eine bestimmte Kategorie vorkommt. Daraus lassen sich relative Häufigkeiten oder Anteile bzw. Prozentsätze berechnen.
- Eine Häufigkeitsverteilung ist die tabellarische oder grafische Zusammenstellung der Häufigkeiten aller Kategorien einer Variable.
- Balken- bzw. Kreisdiagramme dienen zur grafischen Repräsentation der Häufigkeiten bzw. Prozentsätze, mit denen einzelne Kategorien vorkommen.
- Null- und Alternativhypothese: Annahmen, die man beim statischen Testen treffen muss. Das Resultat des Tests legt mit einer bestimmten Wahrscheinlichkeit (p-Wert) nahe, welche der beiden Hypothesen zutrifft.
- Signifikanzniveau: Wahrscheinlichkeit, mit der man eine Nullhypothese verwirft, obwohl sie in Wirklichkeit zutrifft.
- Chi-Quadrat-Test auf Gleichverteilung: prüft, ob alle Kategorien gleich wahrscheinlich sind.
- Chi-Quadrat-Test auf spezifizierte (vorgegebene) Verteilung: prüft, ob Häufigkeiten bzw. Prozentsätze bestimmten Vorgaben entsprechen.
- Test eines Anteilswerts: damit kann ermittelt werden, ob ein Prozentsatz kleiner, größer oder anders als ein bestimmter angenommener Wert ist.
- Konfidenzintervall für Anteile: ist der Bereich, in dem der wirkliche Wert (der Wert in der Population) eines Prozentsatzes mit einer bestimmten Sicherheit erwartet werden kann.

4.7 Übungen

1. Lottozahlen in Österreich 2007 und 2008

In Österreich ist Lotto sehr beliebt. Sechs Zahlen plus Zusatzzahl werden von einer Maschine aus der Menge der Zahlen von 1 bis 45 gezogen. Natürlich erwartet man, dass jede Zahl die gleiche Wahrscheinlichkeit hat, gezogen zu werden. Im Datenfile sind die Ziehungshäufigkeiten für alle 45 Lottozahlen des österreichischen Lottos für die Jahre 2007 und 2008 gegeben. Kann man aus den Daten schließen, dass der Mechanismus, mit dem die Lottozahlen ermittelt werden, fair ist, d. h., dass für alle Zahlen die Auswahlwahrscheinlichkeit gleich ist?

Daten: `lotto0708.sav`
Variablen: `lottozahl`, `h2007`, `h2008`

2. Multiple-Choice-Prüfung

Georg steht vor einer Prüfung, die als Multiple-Choice-Prüfung durchgeführt wird. Wie üblich hat Georg keine Ahnung vom Prüfungsstoff und er hofft, die Prüfung durch reines Raten zu bestehen. Allerdings hat Georg eine frühere Prüfung desselben Prüfers mit markierten richtigen Antworten erhalten. Diese richtigen Antworten sind im Datenfile enthalten. Verteilt der

Prüfer die richtigen Antworten zufällig über alle fünf Auswahlmöglichkeiten?

Daten: `mchoice.sav`
Variable: `richtig`

3. Notenverteilung

Die Noten, die von einem BW-Professor vergeben werden, folgten bisher einer symmetrischen Verteilung: 5% Sehr gut, 25% Gut, 40% Befriedigend, 25% Genügend und 5% Nicht genügend. Dieses Jahr wird eine Stichprobe von 150 Noten gezogen. Kann man daraus schließen (5% Signifikanzniveau), dass sich die Notenverteilung dieses Jahr von der früherer Jahre unterscheidet?

Daten: `noten.sav`
Variable: `note`

4. Autoklasse und Unfallhäufigkeit

Zulassungsdaten aus einem Land zeigen, dass 15% der Autos Kleinwagen, 25% Kompaktmodelle, 40% Mittelklassemodelle und der Rest größere oder Sondermodelle sind. Eine Zufallsstichprobe von Verkehrsunfällen mit Autos wird gezogen. Kann man schließen, dass bestimmte Größenklassen von Autos häufiger in Verkehrsunfälle verwickelt sind, als es die Zulassungszahlen vermuten lassen?

Daten: `unfaelle.sav`
Variable: auto

5. Konzentrationsleistung von Studierenden

Bei einem Konzentrationstest kann man 0 bis 50 Punkte erzielen. Es ist bekannt, dass 15% der Personen mehr als 40 Punkte erzielen. Der Test wurde an 200 zufällig ausgewählten Studierenden durchgeführt. Kann man aus den Ergebnissen schließen, dass Studierende besser abschneiden als die Gesamtbevölkerung (1% Signifikanzniveau)?

Daten: `ktest.sav`
Variable: `punkte`

WWW.PEARSON-STUDIUM.DE CWS COMPANION WEBSITE

Datendateien sowie Lösungen finden Sie auf der Webseite des Verlags.

4.8 Anhang: Die Chi-Quadrat-Verteilung oder wie entsteht ein *p*-Wert?

Im Exkurs 4.3 in Abschnitt 4.2.3 haben wir zwischen beobachteten und erwarteten Häufigkeiten unterschieden und einen Wert X^2 berechnet, der die Größe der Abweichungen zwischen den beiden beschreibt. Auf Basis dieses X^2 hat PASW dann eine Wahrscheinlichkeit, den p-Wert bzw. die Signifikanz, ausgegeben, die besagt, wie sehr die Abweichung für oder gegen die Nullhypothese spricht.

Wie kommt nun so ein p-Wert zustande. Stellen wir uns Folgendes vor: Der „faire“ Würfel aus Exkurs 4.3 wird 30 Mal geworfen. Aufgrund der beobachteten Häufigkeiten berechnen wir einen X^2-Wert und notieren ihn. Nun werfen wir den Würfel nochmals 30 Mal (Durchgang 2), berechnen wieder den X^2-Wert und notieren auch ihn. Da der Zufall im Spiel ist, werden die beiden X^2-Werte kaum gleich groß sein. Was uns aber bei diesem (zugegebenermaßen seltsamen) Spiel interessiert, ist, welche Werte X^2 überhaupt annehmen kann. Wir setzen also das Spiel fort und würfeln ein drittes Mal 30 Mal und berechnen und notieren wieder den neu berechneten X^2-Wert. Den ganzen Vorgang führen wir sehr lange fort, sagen wir 10 000 Mal. Nun bekommen wir natürlich einen ganz guten Eindruck davon, welche X^2-Werte überhaupt vorkommen und wie oft sie vorkommen (die Tabelle zeigt einen kleinen Ausschnitt aus den 10 000 Versuchen). Wir erhalten eine Verteilung der X^2-Werte (solch eine Verteilung nennt man in der Statistik *sampling distribution*).

Man kann jetzt bestimmte Aussagen machen, wie z. B. 75% aller X^2-Werte waren kleiner als sagen wir $6{,}6$ oder 5% waren größer als 11,1. Diese letzte Aussage ist der springende Punkt. Wir haben einen „fairen“ Würfel geworfen und in 5% der Fälle ergab das Resultat des Würfelns eine X^2-Wert, der größer als 11,1 war. Wie könnte das Ergebnis so eines Würfeldurchgangs aussehen, für das der X^2-Wert größer als 11,1 war. In Tabelle 4.1 finden wir ein solches bei Durchgang 343. Besonders auffällig ist, dass die Augenzahl 5 elfmal vorkam. Das ist doppelt so viel, wie man erwartet hätte. Ein so extremes Ergebnis ist also möglich, aber sehr unwahrscheinlich, es kommt in weniger als 5% der Durchgänge vor.

Jeder Durchgang entspricht dem Ziehen einer Stichprobe vom Umfang $n = 30$. Wir haben also 10 000 Stichproben gezogen. In der Praxis haben wir (meistens) nur eine Stichprobe, mehr Information steht nicht zur Verfügung. Sie muss uns aber helfen, eine Fragestellung zu testen. Hätten wir ein Ergebnis wie in Durchgang 343 bekommen, hätten wir die Nullhypothese (der Würfel ist fair) verworfen, weil der X^2-Wert in weniger als 5% auftritt, also sehr unwahrscheinlich ist. Wäre hingegen ein Ergebnis wie in Durchgang 344 eingetreten, dann hätten wir die Nullhypothese nicht verworfen und weiterhin an die „Fairness“ des Würfels geglaubt.

		Augenzahl					
Durchgang	X^2	1	2	3	4	5	6
⋮							
342	6.0	3	7	1	6	7	6
343	12.4	2	3	2	5	11	7
344	2.8	7	5	3	7	4	4
345	6.4	4	6	9	3	6	2
346	7.6	1	9	3	5	6	6
⋮							

Tabelle 4.1: Ausschnitt aus 10 000 Durchgängen, in denen 30 Mal mit einem „fairen“ Würfel gewürfelt wurde. Angegeben werden für jeden Durchgang der X^2-Wert und die beobachteten Häufigkeiten für die einzelnen Augenzahlen.

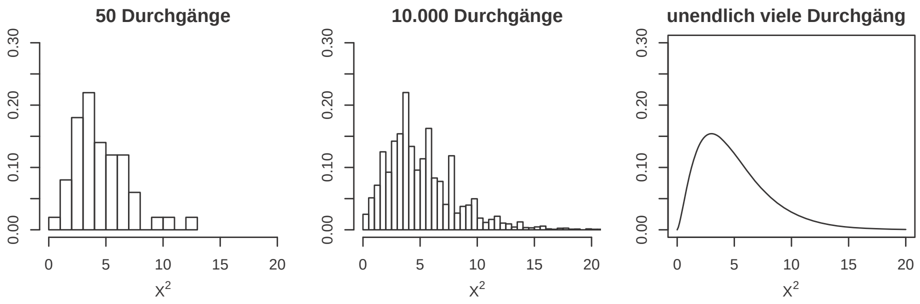


Abbildung 4.29: Sampling

Als Grenze haben wir 11,1 festgelegt, das war jener Wert, über dem nur mehr 5% der gesammelten X^2-Werte auftraten. Diese Grenze basierte auf 10 000 Durchgängen. Falls wir nur 20 Durchgänge gemacht hätten, dann wäre diese Grenze der größte aufgetretene X^2-Wert gewesen. Es ist intuitiv klar, dass je mehr Durchgänge zur Bestimmung dieser Grenze gemacht werden, diese umso genauer wird. Am besten, es wären unendlich viele Durchgänge. Für 50, 10 000 und unendlich viele Durchgänge ist die Verteilung der X^2-Werte in ► Abbildung 4.29 dargestellt.

Das ist natürlich in der Praxis nicht möglich, aber die mathematische Statistik kann dieses Problem lösen (die Idee dabei ist, den Grenzwert zu bestimmen, wenn die Anzahl der Durchgänge gegen unendlich und auch die Stichprobengröße, also aus wieviel Mal würfeln besteht ein Durchgang, gegen unendlich geht).

Das Resultat ist eine theoretische Verteilung, die sogenannte χ^2-Verteilung (χ ist der griechische Buchstabe *chi*, das „ch“ wird wie in „Sprache“ ausgesprochen). In Abbildung 4.29 rechts ist eine spezielle χ^2-Verteilung dargestellt, nämlich eine für 5 Freiheitsgrade. Für einen bestimmten X^2-Wert kann man damit ausrechnen, wie groß die Wahrscheinlichkeit ist, einen solchen oder noch größeren zu bekommen. Diese Wahrscheinlichkeit ist der p-Wert oder Signifikanz. Sie entspricht der Fläche unter der χ^2-Kurve, die ab dem X^2-Wert rechts noch übrig bleibt.

Je nach Anzahl der Kategorien, die man beobachtet, hat die χ^2-Verteilung unterschiedliche Form. Dies wird durch die Freiheitsgrade (df) ausgedrückt. Je höher die Anzahl der Freiheitsgrade, umso symmetrischer wird die χ^2-Verteilung und umso mehr rückt ihr Gipfel nach rechts. Die mathematisch-statistische Bedeutung von Freiheitsgraden ist nicht ganz einfach zu erklären, aber man kann sich Folgendes vorstellen: Man weiß z. B., dass die Summe aus drei Zahlen 10 ergibt, d. h. $x_1 + x_2 + x_3 = 10$. Wie viele der drei x-Werte kann man frei wählen? Man könnte z. B. für x_1 die Zahl 3 wählen. Dann blieben noch zwei weitere Zahlen, die in der Summe 7 ergeben müssen, damit sich die Gesamtsumme 10 ergibt. Wählen wir nun noch für x_2 die Zahl 5, dann haben wir damit auch automatisch $x_3 = 2$ festgelegt. Wir können also zwei Zahlen frei wählen, die dritte ist aufgrund der vorgegebenen Gesamtsumme festgelegt. Wir haben in diesem Beispiel 2 Freiheitsgrade oder $df = 2$. Bei einfachen kategorialen Daten ist die Anzahl der Freiheitsgrade immer die Anzahl der Kategorien minus 1. Beim Würfelbeispiel ist also $df = 5$.