GRUNDLAGEN DER STATISTIK UND EPIDEMIOLOGIE

Teil I
Einführung in die medizinische Statistik

Patrick Messner

1. Auflage

Folgendes Werk ist ausschließlich für den persönlichen Gebrauch bestimmt. Eine Weitergabe in veränderter Form, besonders die Unkenntlichmachung des Verfassers oder eine Verwertung unter finanziellen Aspekten ist verboten. Alle Bilder unterliegen geltenden Copyright-Bestimmungen und sind für den privaten Gebrauch bestimmt, eine darüberhinausgehende Verbreitung und Verwendung ist nicht gestattet.

Grundlagen der Statistik und Epidemiologie

Skriptum für das Studium
der einführenden Statistik
für Veterinärmediziner

Patrick Messner

Veterinärmedizinische Univ. Wien

Wien, Sommersemester 2017

1. Auflage

- Wien

Vorwort

Das Skriptum *Grundlagen der Statistik und Epidemiologie Teil I* soll einen Überblick in die Einführung der medizinischen Statistik und Epidemiologie geben. Dabei behandelt der erste Teil dieser zweiteiligen Skriptenreihe das Basiswissen und die Grundlagenkenntnisse der Statistik, um im zweiten Teil die epidemiologischen Berechnungen durchführen zu können.

Dabei wird vor allem versucht, das Verständnis für den Umgang mit wissenschaftlichen Daten zu vermittelt und gleichzeitig der Umgang mit dem Tabellekalkulationsprogramm Microsoft EXCEL anhand von Beispielen erklärt.

Wien im Mai 2016 *Patrick Messner*

Inhaltsverzeichnis

1 Kapitel 1: Begriffserklärungen

1.1 Einführung

Die Statistik ist die Wissenschaft der Entwicklung und Anwendung formaler Methoden zur Gewinnung, Beschreibung und letztendlicher Beurteilung von Daten. Sie kann aber auch als Wissenschaft zur quantitativen Erfassung und übersichtlichen Darstellung von massenhaft auftretenden Einzelerscheinungen angesehen werden.

1.2 Grundgesamtheit und Stichprobe

- Grundgesamtheit (Population)
  Ist die Menge aller potentieller Untersuchungseinheiten für eine bestimmte Fragestellung.

- Stichprobe
  Ist eine Teilmenge aus der Grundgesamtheit, die entweder zufällig oder nach bestimmten festgelegten Kriterien ausgewählt wurde.

- Repräsentative Stichprobe
  Eine Stichprobe, welche die Verhältnisse in der Grundgesamtheit möglichst genau widerspiegelt.

- Stichprobenumfang n
  Ist die Anzahl der Untersuchungseinheiten einer Stichprobe.

- Deskriptiven (beschreibende, empirische) Statistik
  Mit der deskriptiven Statistik werden die Daten einer Stichprobe in geeigneter Weise beschrieben, aufbereitet und auch zusammengefasst, wobei man mit ihren Methoden quantitative Daten zu Tabellen, grafischen Darstellungen und Kenngrößen (z.B. Mittelwert) verdichtet.

- Induktive (schließende) Statistik
  Mit der induktiven Statistikleitet man aus den Daten einer Stichprobe die Eigenschaften der zugehörigen Grundgesamtheit ab, worauf hier auf der Wahrscheinlichkeitstheorie basierende Schätz- und Testverfahren zurückgegriffen wird.

1.3 Merkmale und Skalenniveaus

- Beobachtungs- oder Merkmalsträger
  Sind die Untersuchungseinheiten wie etwa Tiere, Patienten, Menschen, wobei ein solches Merkmal z.B. das Geschlecht oder der Gesundheitszustand ist und die dazugehörigen Merkmalsausprägungen wären dann männlich/weiblich bzw. gesund/krank.

- Variable oder Zufallsvariabel
  Synonyme für den Begriff „Merkmale“, die oftmals in der Statistik verwendet werden.

- Skalenniveau
  Ein Skalenniveau (Messniveau, Skalendignität, Skalenqualität) ist eine in der Empirie wichtige Eigenschaft von Merkmalen bzw. von Variablen. Demnach kann man jedes Merkmal einem bestimmten Skalenniveau zuordnen, das entweder niedrig oder hoch sein kann und daher bestimmt, welche statistischen Verfahren angewendet werden. Hierbei haben nominale Daten das niedrigste und metrische Daten das höchste Skalenniveau, wobei man vom höheren auf ein niedrigeres Skalenniveau konvertieren kann, aber niemals umgekehrt.

- Nominale Daten (niedriges Skalenniveau)
  Unter nominalen Daten versteht man klassifizierte Daten, wie z.B. der Gesundheitszustand eines Patienten, der als gesund oder krank klassifiziert werden kann. Durch die Berechnung der absoluten oder relativen Häufigkeit kann eine statistische Auswertung erfolgen, wobei Kennzahlen wie ein Mittelwert nicht berechnet werden können.

- Ordinale Daten (mittleres Skalenniveau)
  Ist es möglich, bei kranken Patienten zusätzliche Merkmalsausprägungen wie etwa größer oder kleiner, besser oder schlechter anzugeben, so kann man von ordinalen Daten sprechen. Demzufolge ist der Schweregrad von Bauchschmerzen ein Beispiel für eine solche subjektive Bewertung einer Merkmalsausprägung, die in einer Reihenfolge aufgelistet werden kann.

- Metrische Daten (hohes Skalenniveau)
  Wenn Messwerte im eigentlichen Sinne vorliegen, spricht man von metrischen Daten, wobei man hier zwischen diskrete Messwerte (z.B. Anzahl von Parasiten pro Maus) oder kontinuierlichen Messwerten (z.B. Körpertemperatur) unterscheidet. Einzig allein metrische Daten können quantitativ ausgewertet werden, da sie die höchste Aussagekraft und somit das höchste Skalenniveau besitzen.

2 Kapitel 2: Häufigkeitsverteilung

2.1 Einführung

Mit der Berechnung und gleichzeitigen Darstellung von Häufigkeitsverteilungen wird oftmals der erste Schritt bei der Aufbereitung von Daten durchgeführt. Dabei behält man stets das Ziel im Auge, diese Daten durchschaubar zu machen.

Mit Aufbereiten meint man Ordnen, Zusammenfassen und Darstellen, wobei Häufigkeitsverteilungen auf drei Arten dargestellt werden können.

- tabellarisch
- grafisch durch Diagramme (Säulendiagramme, Histogramme, usw.)
- numerisch durch statistischen Kenngrößen (Mittelwert, Standardabweichung, usw.)

2.2 Häufigkeitsverteilungen für nominale Daten

- Klassifikation
  Als Klassifikation bezeichnet man die Zuordnung von Merkmalsausprägungen zu verschiedenen Klassen, wobei man zwischen absoluter Häufigkeit (Anzahl) und relativer Häufigkeit (Anteil) unterscheidet.

- Absolute Häufigkeit (Anzahl)
  Als absolute Häufigkeit bezeichnet man die gesamte Anzahl der Merkmalsausprägungen

- Relative Häufigkeit (Anteil)
  Als relative Häufigkeit bezeichnet man einen Anteil aus der absoluten Häufigkeit, die man mittels Division der absoluten Häufigkeit durch den Stichprobenumfang n ermittelt. Eine prozentuale Darstellung der relativen Häufigkeit erfolgt durch Multiplikation mit 100.

2.3 Häufigkeitsverteilungen für ordinale Daten

- Klasse nicht frei wählbar
  Bei ordinalen Daten ist eine Anordnung der Klassen nicht frei wählbar, da die Häufigkeiten der Merkmalsausprägungen eine natürliche Reihenfolge definieren.

- Häufigkeitssummen
  Man kann sowohl aus den absoluten als auch aus den relativen Häufigkeiten h(x) sogenannte Häufigkeitssummen H(x) berechnen, die auch als kumulierte Häufigkeitsverteilungen bezeichnet werden. Dabei muss man beachten, dass die Häufigkeitssummen der letzten Klasse immer n (absolut) oder 1 (relativ) ist, wobei hierbei eine Möglichkeit der Kontrolle erfolgen kann.

- Darstellungsmöglichkeiten von Häufigkeiten
  Man kann Häufigkeiten entweder absolut oder relativ bzw. kumuliert oder nicht kumuliert, darstellen. Hierbei muss man berücksichtigen, dass sich absolute und relative Häufigkeiten nur durch die Skalierung der Ordinate unterscheiden lassen.

2.4 Häufigkeitsverteilungen für metrische Daten

- Zusammenfassung von Merkmalsausprägungen
  Eine jede Häufigkeitsverteilung basiert auf einer Zusammenfassung von Merkmalsausprägungen zu Klassen, wobei sich bei nominalen und ordinalen Daten die Klasseneinteilung zu Folge natürlicher Abgrenzungen von selbst ergibt (z.B. Rinderrassen, Schulnoten, usw.). Im Gegensatz dazu muss man die Klassen bei metrischen Daten selbst definieren, wobei die Werte zu Klassen zusammengefasst werden müssen. Hierbei spricht man von Klassierung oder auch Klassenbildung.

- Klassierung
  Bevor eine Klassenbildung durchführt wird, muss überlegt werden, wie viele Klassen überhaupt definiert werden sollen. Damit eine Abschätzung einer sinnvollen Klassenzahl k aus dem Stichprobenumfang n erfolgen kann, kann folgende Formel (für n < 100) zur Hand genommen werden:

$$k \approx \sqrt{n}$$

- Klassenbildung
  Um die Klassenbildung einfach erklären zu können, wird folgendes Beispiel angenommen: es liegt eine Stichprobe mit n = 16 Merkmalsausprägungen (kurz: x-Werte) vor; im ersten Schritt sucht man den kleinsten und größten x-Wert, wobei diese im hiesigen Beispiel $x_{min} = 2$ und $x_{max} = 10$ sind. Man berechnet daraus die Spannweite S (engl. range), die mit folgender Formel ermittelt werden kann:

$$S = x_{max} - x_{min}$$

  In unserem Beispiel ergibt sich folglich eine Spannweite von S = 8. Folglich kann man eine Klassenzahl von k = 4 schätzen, sodass man 4 Klassen mit der Breite 2 erhält. Diese errechneten Klassen können wie folgt angegeben werden: [2 – 4], (4 – 6], (6 – 8], (8 – 10]. Die erste Klasse fasst demnach die Werte 2 – 4 zusammen, die zweite Klasse die Werte 4 – 6. Eine eckige Klammer bedeutet, dass der Grenzwert in der Klasse liegt, die runden Klammern besagen, dass er nicht mehr zur Klasse gehört.

- Klassenbildung in der Praxis
  Oftmals kommt es vor, dass keine ganzzahlige Klassenzahl k errechnet wird, wobei auch die Spannweite eine Dezimalzahl darstellen kann. Demzufolge empfiehlt es sich, für Klassengrenzen ganze Zahlen zu verwenden, so dass stets aufgerundet wird.

- Histogramm metrischer Daten
  Möchte man metrische Daten grafisch darstellen, so empfiehlt es sich statt einem Stabdiagramm ein Histogramm zu verwenden. Bei einer solchen Diagrammwahl ist nämlich auch neben der Anordnung der Höhe der Stäbe auch deren Breite von Bedeutung, sodass die Klassenhäufigkeit durch den Flächeninhalt repräsentiert wird.

2.5 Beobachtete und theoretische Häufigkeitsverteilungen

- Theoretische Häufigkeitsverteilung
  Bis zum jetzigen Zeitpunkt wurde implizit angenommen, dass Häufigkeitsverteilungen immer aus Beobachtungen errechnet werden. Aus diesem Grund spricht man auch von empirischen Verteilungen, die man mit h(x) bzw. H(x) abgekürzt hat. Im später folgenden Kapitel 6 wird über die theoretische Häufigkeitsverteilung f(x) bzw. F(x) diskutiert. Man bezeichnet diese auch als Wahrscheinlichkeitsverteilungen. Sie sind der Übergang von der Verteilung der Stichprobe h(x) auf die Verteilung der Grundgesamtheit f(x).

- Formen von Häufigkeitsverteilungen
  Es gibt verschiedene Formen von Häufigkeitsverteilungen: Schiefe (symmetrisch, linksschief, rechtsschief), Wölbung (normalverteilt, flacher als Normalverteilung, steiler als Normalverteilung), Gipfel (unimodal, bimodal, multimodal).

- EXCEL-Funktion zur Berechnung von Häufigkeiten
  EXCEL ermöglicht die Berechnungen von Häufigkeitsverteilungen, wobei diese im Anschluss auch gezeichnet werden können. Die wichtigsten Funktionen hierfür sind:

$$HÄUFIGKEIT(x_1 : x_n; Klassenobergrenze; kumuliert)$$

  Der Parameter kumuliert = 1 ermöglicht eine Berechnung der Summenhäufigkeit, wobei durch Differenzbildung eine nicht-kumulierte Häufigkeit errechnet wird. Wird die Häufigkeit durch n dividiert, so erhält man relative Häufigkeiten. Weitere Befehle sind:

$$MAX(x_1 : x_n)$$
$$MIN(x_1 : x_n)$$
$$SPANNWEITE(x_1 : x_n)$$
$$WURZEL(n)$$

3 Kapitel 3: Statistische Kenngrößen

3.1 Einführung

Folgende Kenngrößen oder Maße zur Beschreibung von Stichproben (z.B. zu der zugehörigen Grundgesamtheit) können zur Hand genommen werden:

- Lagemaß (Modalwert, Mittelwert, Quantile)
- Streumaß (Spannweite, Standardabweichung, Varianz, Variationskoeffizient)
- Formmaße (Schiefe, Wölbung)

Solche statistische Kenngrößen stehen in einem direkten Zusammenhang mit den Häufigkeitsverteilungen, indem sie die Lage und Streuung der x-Werte sowie die Form der Verteilung charakterisieren. Im Gegensatz dazu unterscheidet man noch Assoziationsmaße für den Zusammenhang mehrerer Merkmale (Kapitel 4).

3.2 Lagemaß

- Nominale Daten
  Bearbeitet man nominale Daten, so kann man den Modalwert (oder die Modalklasse) berechnen.

- Modalklasse
  Eine Modalklasse beschreibt die Klasse mit der größten Besetzungszahl.

- Median
  Bei ordinalen Daten kann auch der Median (Zentralwert) angegeben werden. Er ist jener Wert, der in der Mitte liegt, nachdem man die Zahlen in eine Rangordnung gebracht hat.

- Arithmetischer Mittelwert
  Als bekanntestes Lagemaß gilt der arithmetische Mittelwert, der allerdings nur aus metrischen Daten berechnet werden kann. Er wird aus der Summe der x-Werte geteilt durch den Stichprobenumfang n berechnet:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\,(x_1 + x_2 + \ldots + x_n)$$

- Mittelwert für klassierte Daten
  Kann man nicht auf die ursprünglichen x-Werte zurückgreifen, sondern nur auf die Häufigkeiten $h(x_i) = n_i$ mit den Klassen $i = 1, 2, \ldots, k$, kann folgende Formel für den Mittelwert herangezogen werden:

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i\, n_i = \frac{1}{n}\,[x_1 n_1 + x_2 n_2 + \ldots + x_k n_k]$$

- Lagemaß
  Man muss jedoch beachten, dass Lagemaße nur für 1-gipfelige Verteilungen sinnvoll interpretiert werden können (z.B. Modalwert = Modus = Gipfel).

Skalenniveau	**Zulässige Lage-Kenngröße**
Nominal	Modalklasse
Ordinal	Modalklasse, Quantil, Median
Metrisch	Modalklasse, Quantil, Median

- Quantile
  Quantilen spielen in der schließenden Statistik eine wichtige Rolle, wobei die gebräuchlichste Quantile der Median, die Quartile oder die Zentile ist. Dabei teilt der Median $x_{0,5}$ die geordnete Stichprobe in 2 Hälften und besitzt zugleich den Charakter eines Durchschnittswertes. Wird hingegen die Stichprobe in Viertel geteilt, so nennt man die 3 Trennpunkte $x_{0,25}$, $x_{0,5}$ und $x_{0,75}$ Quartile. In diesem Beispiel entspricht das 2. Quartil folglich dem Median ($x_{0,5}$). Teilt man aber die Stichprobe in 10 Teile $x_{0,1}$, $x_{0,2}$, ... $x_{0,9}$, dann sind Zentilen gemeint. Bei einer beliebigen Einteilung von Quantilen oder p-Quantilen (p = probability), ist zum Beispiel $x_{0,99}$ das p = 0,99 Quantil oder das 99 % Quantil.

- Lagemaß und Häufigkeitsverteilungen
  Lagemaße können in Häufigkeitsverteilungen eingetragen werden, wobei für normalverteilte Daten (Symmetrische Verteilung, Kapitel 6) folgender Spezialfall gilt:

$$Mittelwert = Modalwert = Median$$

3.3 Streumaß

- Streuung der Merkmalsausprägungen
  Zu beachten gilt, dass Häufigkeitsverteilungen sich nicht nur in ihrer Lage unterscheiden, sondern auch hinsichtlich der Streuung der Merkmalsausprägungen (x-Werte). Demzufolge gilt: liegen die x-Werte nahe beim Mittelwert, ist die Streuung klein. Daraus folgt, dass die Streuung umso größer ist, je stärker die Einzelwerte voneinander abweichen. Quantitative Streumaße sind auf metrische Daten beschränkt.

- Spannweite
  Unter der Spannweite versteht man die Differenz aus dem größten und kleinsten x-Wert, der auch als Variationsbreite (range) bezeichnet wird.

$$S = x_{max} - x_{min}$$

- Quartilsabstand (Inter-Quartilsabstand)
  Als Quartilsabstand beschreibt man die Differenz zwischen dem 3. und 1. Quartil von geordneten x-Werten. Demnach schließt der Quartilsabstand die zentralen 50 % der x-Werte ein und wird aus diesem Grund auch als Hälftespielraum bezeichnet.

$$Q = x_{0,75} - x_{0,25}$$

- Quantilsabstände (Inter-Quantilsabstände)
  Quantilsabstände sind eine Verallgemeinerung des Quartilsabstandes. So liegt z.B. der 80 % Spielraum zwischen dem 90 % und 10 % Quantil.

- Standardabweichung
  Als Basis gilt die Summe der Abweichungsquadrate, da für jeden x-Wert die Abweichung vom Mittelwert m gebildet und quadriert wird.

$$s = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n} (x_i - m)^2}$$

- Standardabweichung für klassierte Daten
  Rechnet man mit einer Häufigkeitsverteilung mit $h(x_i) = n_i$, dann wird über die Klassen i = 1, 2, ..., k summiert.

$$s = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{k} (x_i - m)^2 * n_i}$$

- Varianz
  Die Varianz ist ein wichtiges Streuungsmaß der Wahrscheinlichkeitsverteilung einer reellen Zuvallsvariablen, indem sie die erwartete quadratische Abweichung der Zufallsvariablen von ihrem Erwartungswert beschreibt. Wird die Varianz quadriert, bezeichnet man das Ergebnis als Standardabweichung der Zufallsvariablen.

$$var = s^2$$

- Box-and-Whisker-Diargramm (Boxplot)
  Das Box-and-Whisker-Diagramm (auch Boxplot genannt) ist eine häufig in Biowissenschaften verwendete Diagrammart um Median, Spannweite und Quartilsabstand darzustellen. Dabei grenzt die Box den Hälftespielraum ein, das Whisker (englisch: Schnurr- oder Barthaar) die Spannweite, sodass verschieden große Abstände zum Median die Schiefe der Verteilung anzeigen.

3.4 Formmaß

- Metrische Daten
  Formmaße können nur für metrische Daten berechnet werden.

- Schiefe (englisch: skew)
  Mit der Schiefe v wird ein Maß für die Asymmetrie einer Häufigkeitsverteilung angegeben.

$$v = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s}\right)^3$$

- Wölbung (griechisch: kurtosis) und Exzess
  Wölbung und Exzess sind Maße für die Steilheit einer Häufigkeitsverteilung. Wenn man wissen will, ob eine Verteilung steiler oder flacher als eine vergleichbare Normalverteilung ist, so wird anstelle der Wölbung w der Exzess e verwendet. Aufgrund dessen, dass die Wölbung einer Normalverteilung $w = 3$ ist, kann $e = w - 3$ verwendet werden, wobei man bei $e = 0$ von normalgipflig, bei $e > 0$ von steilgipflig und bei $e < 0$ von flachgipflig spricht.

$$w = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - m}{s}\right)^4$$
$$e = w - 3$$

- EXCEL-Funktionen zur Berechnung von statistischen Kenngrößen
  Wichtigste Funktionen zu den Lagemaßen:

$$MITTELWERT(x_1 : x_n)$$
$$QUANTILE(x_1 : x_n; \alpha)$$
$$QUARTILE(x_1 : x_n; f)$$
$$MEDIAN(x_1 : x_n)$$

  Wichtigste Funktionen zu den Streußmaßen:

$$STABW(x_1 : x_n)$$
$$VARIANZ(x_1 : x_n)$$

  Wichtigste Funktionen zu den Formmaßen:

$$SCHIEFE(x_1 : x_n)$$
$$KURT(x_1 : x_n)$$

4 Kapitel 4: Korrelation und Regression

4.1 Einführung

In den ersten Kapiteln hat man sich immer nur darauf beschränkt, ein Merkmal zu beschreiben. Liegen jedoch mehrere Merkmale vor, so kann man sich die Frage stellen, ob diese auch zusammenhängen, d.h., ob zwischen ihnen eine Assoziation besteht. Mit der Korrelationsrechnung beschreibt man die Stärke, mit der Regressionsrechnung die Art des Zusammenhangs.

- Korrelation und Kausalbeziehung
  Auch wenn zwei Merkmale, X und Y, korrelieren, kann man noch keinen sicheren Schluss auf eine Kausalbeziehung schließen. Demzufolge kann eine sinnvolle medizinische Deutung nur durch fachliche Überlegung geklärt werden, wobei man unterscheidet:

  Einseitige Abhängigkeit: X ist die Ursache von Y (z.B. die Körpertemperatur von wechselwarmen Tieren hängt von der Lufttemperatur ab).

  Wechselseitige Abhängigkeit: X kann die Ursache von Y oder umgekehrt sein (z.B. symbiotische Beziehung zwischen Organismen).

  Abhängigkeit von gemeinsamen Ursachen: Hierbei gibt es einen oder mehrere Faktoren, die auf beide Merkmale einwirken, sodass man dann von einer Scheinkorrelation spricht (z.B. Zusammenhang zwischen dem Rückgang der Störche im Burgenland mit dem Rückgang der Anzahl Neugeborener. Dabei reduziert eine dritte Größe, die Verstädterung, die Lebensräume der Störche und fördert gleichzeitig die Bildung von Kleinstfamilien).

4.2 Korrelationskoeffizienten

- Korrelationskoeffizienten:
  Ist der Überbegriff für Assoziations-, Kontingenz- und Korrelationsmaße.

Art der Daten	**Bezeichnung des statistischen Zusammenhangs**	**der Kenngröße**	
Vierfeldertafel	Assoziation	Assoziationskoeffizient	*ϕ-Koeffizient*
Kontingenztafel (nominal/ordinal)	Kontingenz	Kontingenzkoeffizient	*Cramérs V*
Ordinal	Rangkorrelation	Rangkorrelationskoeffizient	*Spearman*
Metrisch	Maßkorrelation (Korrelation)	Maßkorrelationskoeffizient (K, Korrelationskoeffizient)	*Pearson*

- Vierfelder-Korrelationskoeffizient
  Ist der ϕ-Koeffizient (für nominale Daten) und somit ein Korrelationsmaß für nominale Daten (z.B. vorhanden, nicht vorhanden), wie sie in Vierfeldertafeln dargestellt werden können. Dabei erreicht derϕ-Koeffizient für seinen maximalen Wert $r_\Phi = 1$, wenn $n_{11} = n_{22} = 0$ oder $n_{12} = n_{21} = 0$, ist. Außerdem gilt: bei $r_\Phi = 0$ gibt es keinen Zusammenhang. In der Regel liegen die Werte dazwischen, sodass $r_\Phi = 0{,}7$ (als Maß für die Stärke des gemeinsamen Auftretens von Krankheit X und Y) typisch wäre.

		Krankheit		
		vorhanden	nicht vorhanden	$\sum$
Krankheit Y	vorhanden	n_{11}	n_{21}	$n_{11} + n_{21} = n_{.1}$
	nicht vorhanden	n_{12}	n_{22}	$n_{12} + n_{22} = n_{.2}$
	$\sum$	$n_{11} + n_{12} = n_{1.}$	$n_{21} + n_{22} = n_{2.}$	$n_{1.} + n_{2.} = n_{.1} + n_{.2} = n$

Anm.: Besetzungszahlen (n_{11}, n_{21}, n_{12} und n_{22}) bzw. Randsummen ($n_{1.}$, $n_{2.}$, $n_{.1}$, $n_{.2}$)

- Berechnung des ϕ-Koeffizienten

$$r_\phi = \frac{|\, n_{11} * n_{22} - n_{12} * n_{21}|}{\sqrt{n_{.1} * n_{.2} * n_{1.} * n_{2.}}}$$

Die obenstehende Formel wurde aus der allgemeineren Gleichung gewonnen, die wie folgt lautet:

$$r_\phi = \sqrt{\frac{\chi^2}{n}} \qquad \text{oder} \qquad \chi^2 = n * n_\phi^2$$

Wie χ^2 (Chi-Quadrat) berechnet wird, wird in Kapitel 7 besprochen. Hier wird χ^2 als Testgröße für die Prüfung auf Unabhängigkeit verwendet. Folgende Einschränkungen müssen berücksichtigt werden: De ϕ-Koeffizient r_Φ gibt keinen Aufschluss darüber, ob die vorliegende Korrelation positiv oder gar negativ ist. Um diese Erkenntnis zu gewinnen, muss der Berechnende diese Fakten selbst aus der Vierfeldertafel ablesen. Hierbei ist der ϕ-Koeffizient mit dem Maßkorrelationskoeffizienten nach Pearson konsistent (r_Φ kann aus r abgeleitet werden).

- Kontingenzkoeffizient = Cramérs V (für nominale und ordinale Daten):
  Cramérs V ist ein Korrelationsmaß für nominale und ordinale Daten, wie sie beispielsweise in Kontingenztabellen beliebiger Dimension dargstellt werden können. Dazu wird Cramérs V wie folgt berechnet:

$$r_V = \sqrt{\frac{\chi^2}{n * (MIN[I, m] - 1}}$$

Hierbei ist MIN[I, m] der kleinere der Werte von I (Dimension der Spalten) und m (Dimension der Reihen). Für I = m = 2 erhält man so den Spezialfall einer 2 * 2 Kontigenztabelle. Demzufolge gilt für die Vierfeldertafel nach wie vor:

$$r_V = r_\phi$$

- Rangkorrelationskoeffizient nach SPEARMAN (für ordinale Daten)
  Der Rangkorrelationskoeffizient nach SPEARMAN ist ein Maß für die Stärke des Zusammenhanges zweier Merkmale (eines Wertepaares von x- und y-Werten) auf einer ordinalen Skala. So kann der Korrelationskoeffizient nach Spearman Werte zwischen r_S = +1 (vollständige positive Korrelation) und r_S = -1 (vollständige negative Korrelation) annehmen. $r_S \approx 0$ heißt, dass kein Zusammenhang besteht.

$$r_S = 1 - \frac{6}{n * (n^2 - 1)} * \sum_{i=1}^{n} d_i^2$$

d_i = Differenzen zwischen den Rängen der gepaarten x- und y-Werte

Zu beachten ist, dass es für den Korrelationskoeffizienten nach SPEARMAN keine EXCEL-Funktion gibt. Zur Abhilfe erhält man jedoch eine gute Näherung, wenn man die Daten mit *RANG()* reiht und die Ränge in *KORREL()* oder *PEARSON()* einsetzt. Treten dann gleiche Ränge auf, dann führt die Funktion *RANG.MITTELW()* zu besseren Ergebnissen.

- Varianz und Kovarianz
  Man kann für Werte metrischer Daten sowohl die Varianzen der x- und y-Werte, als auch die zugehörigen Kovarianzen (engl. covariance) berechnen. Dabei wird die Kovarianz (abs. Maß) dem Korrelationskoeffizienten (rel. Maß) vorgezogen, wenn keine allgemein vergleichbare Kenngröße benötigt wird.

$$var_x = s_x^2 = \frac{1}{n-1} * \sum(x - m_x)^2 \quad \text{bzw.} \quad var_y = s_x^2 = \frac{1}{n-1} * \sum(x - m_y)^2$$

$$cov_{xy} = \frac{1}{n-1} * \sum(x - m_x) * (y - m_y)$$

mx, sx = Mittelwert und Standardabweichung der x-Werte
my, sy = Mittelwert und Standardabweichung der y-Werte
Σ = Kurzschreibweise der Summe $\sum_{i=1}^{n}$

- Zweidimensional normalverteilte Daten
  Zweidimensional normalverteilte Daten sind die Voraussetzung, dass der Maßkorrelationskoeffizient nach PEARSON angwendet werden darf.
  Dabei wird die statistische Prüfung auf Normalverteilung in Kapitel 7 besprochen.

- Typische Werte des Maßkorrelationskoeffizienten r
  Der Korrelationskoeffizient kann hierbei Werte zwischen r = -1 (negativ korreliert) und r = +1 (positiv korreliert) annehmen. Bei r = nahe 0, so besteht kein statistischer Zusammenhang. Der Korrelationskoeffizient r ist nur für lineare Zusammenhänge definiert.

4.3 Regressionsanalyse

- Lineare Regression
  Zu beginn wird der Fall einer linearen Regression betrachtet. Darunter versteht man die Art des Zusammenhangs zweier Merkmale durch eine Gerade mit folgender Gleichung:

$$y = k * x + d$$

  y = Wert auf der y-Koordinate
  x = Wert auf der x-Koordinate
  d = Achsenabschnitt
  k = Steigung

  Die Gerade wird dabei so in die Punktewolke gesetzt, dass die Summe der quadratischen Abstände (in y-Richtung) minimal ist.

- Bestimmtheitsmaß (erklärte Varianz) r^2
  Unter dem Bestimmtheitsmaß versteht man das Quadrat des Korrelationskoeffizienten und gibt somit den erklärten Anteil der Variation (der Varianz) der Zielgröße durch die Einflussgröße an.

  Dabei wird bei der Regressionsanalyse r^2 auch als Maß für die Güte des gewählten Regressionsmodells verwendet. Bieten sich also mehrere Funktionen (linear oder nichtlinear) als statistisches Regressionsmodell an, so kann r^2 dazu verwendet werden, das beste davon auszuwählen. Es ist hierbei jedoch zu beachten, dass es oft sinnvoll ist, ein Regressionsmodell (eine Funktion) zu wählen, die biologisch interpretierbar ist.

  Weitere wichtige EXCEL-Funktionen zur Korrelations- und Regressionsanalyse:

$$VARIANZ(x_1 : x_n) \quad und \quad KOVAR(x_1 : x_n; y_1 : y_n)$$
$$ACHSENABSCHNITT(y_1 : y_n; x_1 : x_n)$$
$$STEIGUNG(y_1 : y_n; x_1 : x_n)$$

5 Kapitel 5: Wahrscheinlichkeitsfunktionen

5.1 Einführung

Unter Wahrscheinlichkeitsfunktionen versteht man theoretische Häufigkeitsverteilungen. Während die sogenannten empirischen Häufigkeitsverteilungen h(x) eine Stichprobe repräsentieren, beschreiben die theoretischen Häufigkeitsverteilungen f(x) die Grundgesamtheit (Population).

Man verfolgt in der Statistik folgendes Konzept: Aus der Stichprobe wirst zuerst eine Häufigkeitsverteilung berechnet. Im Anschluss sucht man dann eine passende Wahrscheinlichkeitsfunktion dazu und vollzieht damit den Übergang von der Häufigkeit (Stichprobe) zur Wahrscheinlichkeit (Grundgesamtheit).

Eine objektive Beurteilung, ob die ausgewählte Funktion f(x) die empirische Häufigkeitsverteilung h(x) auch wirklich hinreichen beschreibt, erfolgt mittels sogenannter Anpassungstests, die im später folgenden Kapitel 7 behandelt werden.

5.2 Dichte- und Verteilungsfunktion

- Häufigkeiten h(x) und Summenhäufigkeiten H(x)
  Man unterscheidet bei den empirischen Häufigkeitsverteilungen zwischen den Häufigkeiten h(x) und Summenhäufigkeiten H(x), wobei man letztere auch als kumulierte Häufigkeit bezeichnet.

- Theoretische Häufigkeitsverteilungen
  Gleichzeitig unterscheidet man bei den theoretischen Häufigkeitsverteilungen zwischen Dichtefunktion f(x) und Verteilungsfunktion F(x), wobei man erstere auch als probability density function (pfd) bezeichnet.

- EXCEL-Funktionen
  Manche der vorgegebenen EXCEL-Funktionen erlauben eine Berechnung der beiden Wahrscheinlichkeiten durch Angabe des Parameters „kumuliert“. Dabei muss man jedoch beachten, dass sowohl die im Folgenden angegebenen Formel als auch die EXCEL-Funktionen immer relative Wahrscheinlichkeitsfunktionen berechnen. Damit diese mit den aus einer Stichprobe berechneten absoluten Häufigkeiten vergliechen werden können, müssen sie mit dem Stichprobenumfang n multipliziert werden.

5.3 Theoretische Häufigkeitsverteilungen

Gleichverteilung Normalverteilung (Gauß-Verteilung) Binomialverteilung Negative Binomialverteilung	Die ersten 4 Verteilungen werden immer zur Beschreibung von Daten verwendet.
χ^2-Verteilung t-Verteilung (Student-Verteilung)	Die letzten2 Verteilungen sind sogenannte Testverteilungen.

- Diskrete und stetige Verteilungen
  Es wird zwischen diskreten und stetigen Verteilungen unterschieden. Dabei sind die Binomialverteilungen diskret, wohingegen die Gleichverteilung sowohl diskrete als auch kontinuierliche Daten beschreiben kann.

5.4 Gleichverteilung

- Diskret oder stetig
  Eine Gleichverteilung kann entweder diskret oder stetig sein. Ist die Wahrscheinlichkeit f(x) für jeden Wert x_i (i = 1, 2, ..., n) gleich groß, so liegt eine diskrete Gleichverteilung vor. In diesem Falle ist die Dichtefunktion einfach:

$$f(x) = \frac{1}{n}$$

  Dabei muss für eine stetige Gleichverteilung ein Intervall [a, b] definiert werden, sodass für alle x-Werte innerhalb dieses Intervalls folgendes gilt:

$$f(x) = \frac{1}{b-a} \quad und \quad F(x) = \frac{x-b}{b-a}$$

  bzw. für das normierte Intervall [0, 1]:

$$f(x) = 1 \quad und \quad F(x) = x$$

5.5 Normalverteilung

- Wichtigste stetige Wahrscheinlichkeitsfunktion
  Die Normalverteilung wird auch als Gauß-Verteilung bezeichnet und stellt die wichtigste stetige Wahrscheinlichkeitsfunktion dar. Hierbei ist die Dichtefunktion (Gauß'sche Glockenkurve) symmetrisch und wird wie folgt berechnet:

$$f(x) = \frac{1}{\sigma * \sqrt{2 * \pi}}^{[-\frac{(x-\mu)^2}{2*\sigma^2}]}$$

Hierbei heißt sie für, wenn für den Mittelwert μ = 0 und für die Standardabweichung σ = 1 (die statistische Maßzahlen der Grundgesamtheit werden mit griechischen Buchstaben bezeichnet) genommen wird, Standardnormalverteilung und ist in Statistikbüchern tabelliert. Dabei setzen viele statistische Methoden/Testverfahren normalverteilte Daten voraus. Gleichzeitig wird bei der Dichtefunktion f(x) für x die Klassenmitte, bei der Verteilungsfunktion F(x) die Klassenobergrenze, angegeben.

- Mittelwert
  Die freien Parameter werden vom Mittelwert μ und von der Standardabweichung σ definiert, wobei der Mittelwert ein Lagemaß (definiert die Lage der Kurve auf der x-Achse) ist.

- Standardabweichung
  Die freien Parameter werden vom Mittelwert μ und von der Standardabweichung σ definiert, wobei die Standardabweichung ein Streumaß (Streuung der x-Werte um den Mittelwert) angibt. Abschätzungen beider Parameter erfolgt mit der Momentenmethode:

$$\mu = m \quad bzw. \quad \sigma = s$$

5.6 Standardnormal- und Studentenverteilung

- z-Transformation
  Die sogenannte z-Transformation bringt normalverteilte x-Werte auf Standardnormalverteilung (z-Verteilung):

$$z = \frac{x - m}{s}$$

- Studentenverteilung (t-Verteilung)
  Die Studentenverteilung, auch t-Verteilung genannt, kann mit einer EXCEL-Funktion berechnet werden. Sie gilt als Prüfverteilung (wird in Kapitel 6 und 7 besprochen), wobei als Näherung oftmals die z-Verteilung verwendet wird (für Freiheitsgrade $\nu = \infty$ sind beide ident). Die EXCEL-Funktion der Studentenverteilung:

$$T.VERT(x;\ \nu;\ kumuliert)$$

5.7 Sigma-Regel (für normalverteilte Daten)

- Intervalle:

$$68{,}3\ \%\ der\ Werte\ liegen\ im\ Intervall\ \mu\ \pm 1 *\ \sigma$$

$$95{,}5\ \%\ der\ Werte\ liegen\ im\ Intervall\ \mu\ \pm 2 *\ \sigma$$

$$99{,}7\ \%\ der\ Werte\ liegen\ im\ Intervall\ \mu\ \pm 3 *\ \sigma$$

- $2 * \sigma$-Regel bei z-Verteilung

$$ca.\ 95\ \%\ der\ Werte\ liegen\ im\ Intervall\ [-2, +2]$$

5.8 Binomialverteilung

- Diskrete Wahrscheinlichkeitsverteilung
  Die Binomialverteilung ist eine diskrete Wahrscheinlichkeitsverteilung. Ihre Dichtefunktion f(x) gibt dementsprechend die Ereigniswahrscheinlichkeit bei n unabhängigen Versuchen konstanter Versuchswahrscheinlichkeit p an, um genau x Ergebnisse zu erzielen.

- Berechnung der Binomialverteilung

$$f(x) = \binom{n}{x} *\ p^x * (1-p)^{n-x}$$

  Die Binomialkoeffizienten „n über x“ werden über die Fakultäten berechnet:

$$\binom{n}{x} = \frac{n!}{x! * (n-x)!}$$

  Folgende EXCEL-Funktion dient dazu, um die Fakultäten zu berechnen:

$$FAKULTÄT(x)$$

  Weitere Funktionen:

$$BINOMVERT(x; n; p; kumuliert)$$

- Kumuliert
  Die Funktion/Verteilung besitzt die 2 freien Parameter n und p. Mit der EXCEL-Funktion kann über den Parameter „kumuliert“ gewählt werden, ob die Dichtefunktion f(x) oder die Verteilungsfunktion F(x) berechnet werden soll.

5.9 Negative Binomialverteilung

- Diskrete Wahrscheinlichkeitsverteilung
  Die negative Binomialverteilung ist eine diskrete Wahrscheinlichkeitsverteilung, die oft zur Charakterisierung der Häufigkeit von „Makro-Parasiten“ (z.B. Zecken) verwendet wird.

- Berechnung der negativen Binomialverteilung (analytisch)

$$f(x) = \binom{\kappa + x - 1}{x} * p^{\kappa} * (1-p)^{x}$$

  Der erste Term wird dabei oft alternativ angeschrieben und kann über Fakultäten berechnet werden:

$$\binom{\kappa + x - 1}{x} = \binom{\kappa + x - 1}{\kappa - 1} = \frac{(\kappa + x - 1)!}{x! * (\kappa - 1)!}$$

  Die dazugehörende EXCEL-Funktion lautet:

$$NEGBINOMVERT(x;\ \kappa; p)$$

- Berechnung der negativen Binomialverteilung (rekursiv)
  Es muss beachtet werden, dass die negative Binomialverteilung eine diskrete Verteilung ist und die x-Werte daher ganze Zahlen darstellen müssen. Gleichzeitig muss auch $\kappa > 0$ und ganzzahlig sein, da ansonsten die Binomialkoeffizienten nicht berechent werden können. Daher ist oftmals die EXCEL-Funktion unbrauchbar. Im Gegensatz dazu kann die rekursive Berechnung immer verwendet werden:

$$f(0) = p^{\kappa}$$

$$f(x) = \frac{x + \kappa - 1}{x} * (1-p) * f(x-1)$$

  Hierbei können die 2 freien Parameter κ und p aus Daten über die Momentenmethode geschätzt werden:

$$\kappa = \frac{m^2}{s^2 - m} \qquad bzw. \qquad p = \frac{m}{s^2}$$

6 Kapitel 6: Schätzen unbekannter Größen

6.1 Einführung

Es muss berücksichtigt werden, dass aus einer Stichprobe berechnete (empirisch bestimmte) statistische Kenngrößen immer Schätzungen für die wahren (unbekannten) Kenngrößen der Grundgesamtheit sind. Demnach beantworten Konfidenzintervalle die Frage, wie weit ein Schätzwert vom wahren Wert entfernt ist. Diese Intervalle gehören zu den häufig verwendeten Methoden der schließenden Statistik.

Um dies durchzuführen, werden ein unterer und ein oberer Grenzwert angegeben, die den wahren Wert mit einer bestimmten Sicherzeit (z.B. 90 %, 95 % oder 99 %) einschließen. Man bezeichnet diese Grenzwerte als Konfidenz-, Vertrauens- oder Mutungsgrenzen, wobei der von ihnen eingeschlossene Wertebereich als Konfidenzintervall, Vertrauens- oder Mutungsbereich bezeichnet wird.

6.2 Irrtumswahrscheinlichkeit α

- Intervallschätzung
  Einer solchen Intervallschätzung liegt eine Wahrscheinlichkeitsaussage (entsprechend einer bestimmten statistischen Sicherheit) zugrunde. Demzufolge spricht man von einem 90 %, 95 % oder 99 % Konfidenzintervall (KI).

- Irrtumswahrscheinlichkeit α
  Das Restrisiko und somit jener Schätzwert, der nicht im Konfidenzintervall liegt, wird auch als Irrtumswahrscheinlichkeit α bezeichnet. Demzufolge entsprechen Irrtumswahrscheinlichkeiten von 10 %, 5 % oder 1 % dann α-Werten von 0,1, 0,05 oder 0,01.

- Signifikanzniveau α
  In den meisten Fällen wird eine Irrtumswahrscheinlichkeit von α = 0,05 angenommen. Dabei muss beachtet werden, dass die Irrtumswahrscheinlichkeit auch als Signifikanzniveau α bezeichnet wird (Kapitel 7).

- Zufällige oder systematische Phänomene
  Die schließende Statistik beschäftigt sich damit, herauszufinden, ob Phänomenen (Unterschiede, Zusammenhänge, usw.) zufällig oder systematisch (überzufällig, signifikant) sind.

6.3 Standardfehler und Variationskoeffizient

- Standardfehler (Standardabweichung des Mittelwertes)
  Der Standardfehler gibt die Genauigkeit des Mittelwerts m der Stichprobe als Schätzwert für den Erwartungswert μ (Mittelwert $\pm$ Standardfehler) an. So wird mit zunehmendem Stichprobenumfang n „se“ immer kleiner und der Mittelwert wird eine immer bessere Schätzung für den Erwartungswert (für n = ∞ folgt demnach m = μ).

$$se = \frac{s}{\sqrt{n}}$$

- Variationskoeffizient (des Mittelwertes)
  Möchte man herausfinden, ob die Streuung stark oder gering anzusehen ist, ist dies leichter zu beurteilen, wenn man die Standardabweichung im Verhältnis zum Mittelwert betrachtet. Wird der Wert mit 100 multipliziert, so wird V auch als prozentualer Fehler des Mittelwerts bezeichnet.

$$V = \frac{s}{m}$$

6.4 Konfidenzintervall für Mittelwerte

- Konfidenzintervall KI_u und KI_o
  Die Grenzen des Konfidenzintervalls, die man auch als KIu und KIo abkürzt, liegen für den wahren Mittelwert = Erwartungswert μ (Grundgesamtheit) jeweilt unterhalb (KI_u) bzw. oberhalb (KI_o) des empirisch bestimmten Mittelwertes m (Stichprobe). Demzufolge gilt, dass bei gleicher Irrtumswahrscheinlichkeit α das Konfidenzintervall umso breiter ist, je größer die Streuung der Daten s und je kleiner der Stichprobenumfang n ist.

  Deshalb gilt für normalverteilte Daten:

$$m \pm t * \frac{s}{\sqrt{n}} = \begin{cases} KI_u \\ KI_o \end{cases} \quad oder \quad \mu = m \pm t * \frac{s}{\sqrt{n}}$$

  n, m, s = Umfang, Mittelwert und Standardabweichung der Stichprobe
  t = kritischer Wert der t-Verteilung, EXCEL-Funktion $TINV(\alpha, \nu)$
  ν = Freiheitsgrade $\nu = n - 1$

  So folgt mit dem Standardfehler se und der Präzision L aus der oben stehenden Gleichung:

$$\mu = m \pm t * se = m \pm L \quad mit: \quad L = t * \frac{s}{\sqrt{n}}$$

Hierbei muss beachtet werden, dass die EXCEL-Funktion

$$KONFIDENZ.T(\alpha, s, n)$$

nicht das Konfidenzintervall, sondern die Präzision L berechnet. Außerdem wird manchmal anstelle der t- die z-Verteilung verwendet, mit der dazugehörenden EXCEL-Funktion:

$$KONFIDENZ.NORM(\alpha, s, n)$$

Dieser Befehl ist eine Näherung:

$$\mu = m \pm z * se \quad mit \quad \alpha = 0{,}05 \quad folgt: \quad \mu = m \pm 1{,}96\, se$$

So liegt in erster Näherung demnach der Erwartungswert im Intervall:

$$\mu = m \pm 2 * se$$

6.5 Konfidenzintervall für die Differenz zweier Mittelwerte

- Konfidenzintervallgrenzen KI_u und KI_o
  Es werden die Grenzen des Konfidenzintervalls, KIu und KIo, für die Differenz der wahren Mittelwert μ_1 - μ_2 (Grundgesamtheit). Dabei werden jedoch normalverteilte Stichproben und Varianzhomogenität ($s_1 = s_2 = s$) vorausgesetzt.

$$t * s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \pm (m_1 - m_2) = \begin{cases} KI_u \\ KI_o \end{cases}$$

t = kritischer Wert der t-Verteilung, EXCEL-Funktion $TINV(\alpha, \nu)$
ν = Freiheitsgrade $\nu = n_1 + n_2 - 2$

Als alternative Schreibweise gilt:

$$\Delta\mu = \Delta m \pm L \quad mit: \quad L = t * s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- Anmerkung
  Mit der Berechnung des Konfidenzintervalls KI verwendet man eine Alternative zum t-Test für zwei unverbundene Stichproben (Kapitel 7). Man kann daher mit dem KI auf die statistische Signifikanz schließen.

6.6 Konfidenz- und Vorhersageintervall für Regressionsgeraden

- Konfidenzintervalle für Regressionsgeraden
  Liegt eine gegebene Irrtumswahrscheinlichkeit von beispielsweise α = 0,05 vor, so befindet sich der Erwartungswert (hier der mittels Regression berechnete Mittelwert y_{regr}) aller y-Werte an der Stelle x mit einer Wahrscheinlichkeit von 95 % innerhalb des berechneten Konfidenzintervalls (KI). Berechnet man somit die KI für alle x-Werte, so erhält man ein Konfidenzband.

- Vorhersageintervalle für Regressionsgeraden
  Das so genannte Vorhersageintervall (VI) ist so konzipiert, dass es zukünftige Beobachtungen y an einer bestimmten Stelle x mit der Wahrscheinlichkeit $1 - \alpha$ enthält. Berechnet man also die VI für alle x-Werte, so erhält man ein Vorhersageband. Dabei ist das Vorhersageband immer breiter als das Konfidenzband, da das KI sich auf Mittelwerte bezieht und das VI für Einzelwerte gilt.

- Berechnung des Konfidenzintervalls (confidence interval)

$$y_{regr} \pm \sqrt{2 * F} * s_{y1} = \begin{cases} KI_o \\ KI_u \end{cases}$$

s_{y1} = Standardabweichung für geschätzten y-Mittelwert
F = kritischer Wert der F-Verteilung, EXCEL-Funktion $FINV(0{,}025, 2, n-2)$

$$s_{y1} = s_{yx} * \sqrt{\frac{1}{n} + \frac{(x - m_x)^2}{\sum(x - m_x)^2}} \qquad mit: \qquad s_{yx} = s_y * \sqrt{1 - r^2}$$

s_y = Standardabweichung der y-Werte
m_x = Mittelwert der x-Werte
r = Korrelationskoeffizient

- Berechnung des Vorhersageintervalls (prediction interval)

$$y_{regr} \pm t * s_{y2} = \begin{cases} VI_o \\ VI_u \end{cases}$$

s_{y2} = Standardabweichung für geschätzte y-Mittelwerte
t = kritischer Wert der t-Verteilung, EXCEL-Funktion $TINV(0{,}025, n-2)$

$$s_{y2} = s_{yx} * \sqrt{1 + \frac{1}{n} + \frac{(x - m_x)^2}{\sum(x - m_x)^2}} \qquad mit: \qquad s_{yx} = s_y * \sqrt{1 - r^2}$$

s_y = Standardabweichung der y-Werte
m_x = Mittelwert der x-Werte
r = Korrelationskoeffizient

7 Kapitel 7: Statistische Testverfahren

7.1 Einführung

Oftmals wird die statistische Bearbeitung von Versuchsergebnissen mit der Anwendung von statistischen Tests gleichgesetzt. Dadurch wird die Bedeutung von statistischen Tests unterstrichen. So bieten beispielsweise Signifikanztests Schutz gegen täuschende Einflüsse des Zufalls und können empirische Wissenschaftler davon bewahren, Zufallseffekte (fälschlich) als Gesetzmäßigkeiten zu interpretieren.

Es muss beachtet werden, dass ein mit einem Signifikanztest „statistisch abgesichertes" Ergebnis von vielen Wissenschaftlern, Herausgebern von Zeitschriften, Gutachtern oder auch Betreuern von Diplom- und Doktorarbeiten gefordert wird. Sie sind jedoch nicht gleichbedeutend, dass das Ergebnis in jeder Hinsicht unantastbar ist. Demzufolge ist eine fehlerfreie rechnerische Ausführung noch lange keine Garantie dafür, dass nicht andere methodische Fehler oder Fehler bei der Planung und Durchführung von Experimenten (z.B. fehlende Randomisierung) gemacht wurden.

7.2 Nullhypothese H_0 und Alternativhypothese H_1

- Hypothese/Nullhypothese/Alternativhypothese
  Für gewöhnlich wird in der Statistik eine Hypothese (Aussage) formuliert, die man widerlegen (falsifizieren) möchte. Man bezeichnet eine solche Hypothese auch als Nullhypothese H_0. Eine solche Nullhypothese kann wie folgt lauten: Die Mittelwerte zweier Stichproben sind gleich - „ein Versuch zeigt einen Effekt":

  $$H_0: \mu_1 = \gamma_2$$

  Unter der Einbeziehung der Daten wird folglich mit einem statistischen Test die so genannte Teststatistik berechnet. Spricht diese Teststatistik gegen die Nullhypothese, so wird sie verworfen. Dies ist gleichbedeutend damit, dass man die Alternativhypothese annimmt. Hier – „ein Versuch zeigt einen Effekt":

  $$H_1: \mu_1 \neq \mu_2$$

  Dabei unterscheiden sich die Mittelwerte signifikant für eine definierte Irrtumswahrscheinlichkeit α. Als Alternative wird der p-Wert berechnet, wobei $p < \alpha$ angestrebht wird.

7.3 Signifikanzniveau

- Testgröße (Teststatistik)
  Mit statistischen Tests überprüft man, ob Unterschiede signifikant (d.h. nicht zufällig) sind. Um dies zu ermitteln, wird eine Testgröße (Teststatistik) ermittelt. Überschreitet diese Testgröße jedoch einen kritischen Wert, so wird bei einem gewählten Signifikanzniveau α die Nullhypothese abgelehnt. Der Unterschied lautet dann signifikant und somit nicht zufällig.

- Irrtumswahrscheinlichkeit α (Signifikanzniveau)
  Die Irrtumswahrscheinlichkeit α (Signifikanzniveau) gibt das Risiko an, die Nullhypothese irrtümlich abzulehnen, wobei folgende Werte gebräuchlich sind:

Wert:	**Signifikanz:**	**Abkürzung:**
$\alpha = 0{,}1$	tendenziell	
$\alpha = 0{,}05$	signifikant	*
$\alpha = 0{,}01$	hoch signifikant	**
$\alpha = 0{,}001$	höchst signifikant	***

In den meisten Fällen will man signifikante Unterschied ($\alpha = 0{,}05$) finden. Als Ausnahme gelten hier die Anpassungstests. Bei diesen Tests will man, dass h(x) durch f(x) beschrieben wird, wobei sie sich also nicht signifikant unterscheiden.

7.4 p-Wert und Signifikanzniveau α

- Kennzahl von Signifikanztests
  Der p-Wert (probability value) ist, wie die Teststatistik (Testgröße), eine Kennzahl von Signifikanztests. Überschreitet den so genannten kritischen Wert als Funktion von α, kann man von einem signifikanten Resultat sprechen. Heutzutage ist die direkte Angabe der Signifikanz durch den p-Wert üblich. Früher war es üblich, eine einfache Testgröße zu berechnen und den zugehörigen kritischen Wert aus einer Tabelle abzulesen. Seit der Einführung von Computern und Statistikprogrammen kann der p-Wert direkt berechnet werden.

- Referenzbereich des p-Werts
  Der p-Wert liegt stets zwischen 0 und 1, wobei folgende Regel gilt: je kleiner der p-Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese. Außerdem wird für Werte die kleiner als das im voraus festgesetzten Signifikanzniveau α = 0,05 (5 %), α = 0,01 (1 %) bzw. α = 0,001 (0,1 %) die Nullhypothese abgelehnt, gleichbedeutend mit einem signifikanten Resultat.

- Probe des p-Werts
  Wird der p-Wert anstelle von α zur Berechnung des kritischen Wertes eingesetzt, dann gilt: Testgröße = kritischer Wert.

7.5 Anpassungstests (für metrische Daten)

- Anpassungsprüfung
  Wird an eine empirische Häufigkeitsverteilung h(x) eine theoretische Verteilung f(x) angepasst, so werden dazu statistische Testverfahren verwendet um zu prüfen, ob die Anpassung auch erfolgreich war. Eine der hier gebräuchlichsten Anpassungstests ist der sogenannte χ^2-Anpassungstest.

- χ^2-Anpassungstest
  Der χ^2-Anpassungstest dient dem Nachweis, dass Daten einer theoretischen Verteilung genügen. Demnach folgt den Häufigkeiten $h(x_i) = n_i$ und den Erwartungswerten $f(x_i) = e_i$ die Testgröße χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{[h(x_i) - f(x_i)^2]}{f(x_i)} = \sum_{i=1}^{k} \frac{(n_i - e_i)}{e_i}$$

  Dafür kann der kritische Wert χ^{2*} aus der inversen Verteilung mit der entsprechenden EXCEL-Funktion bestimmt werden. Üblicherweise wird hierbei α = 0,05 gewählt. Außerdem müssen die Freiheitsgrade ν als Funktion der Klassenzahl k und der freien Parameter a der gewählten Wahrscheinlichkeitsfunktion angegeben werden.

$$\chi^{2*} = CHIINV(\alpha;\ \nu) \qquad mit: \qquad \nu = k - a - 1$$

- Null- und Alternativhypothese bei Anpassungstests
  Mithilfe von Anpassungstests möchte man nachweisen, dass es keine signifikanten Unterschiede zwischen der empirischen Verteilung von medizinischen Daten und einer gewählten theoretischen Verteilung gibt. Sie stehen daher im Gegensatz zu allen anderen statistischen Tests, die bekanntlich die Unterschiede nachweisen wollen.

- Nullhypothese H_0
  Die Nullhypothese ist eine Hypothese, dass sich zwei Verteilungen nicht unterscheiden – aus der gleichen Grundgesamtheit stammen. Hierbei werden beobachtete Abweichungen als zufällig erachtet. Möchte man zeigen, dass Daten normalverteilt sind (eine Voraussetzung für viele Testverfahren), will man die Nullhypothese beibehalten.

- Alternativhypothese H_1
  Hier unterscheiden sich die empirische und die theoretische Verteilung signifikant.

  Man strebt dabei an, dass die Nullhypothese H_0 beibehalten wird, gleichbedeutend mit:

$$Testgr\ddot{o}\ss e < kritischer\ Wert$$

Im Klartext heißt das, dass die Anpassung erfolgreich war. Das heißt, dass die Daten durch die gewählte theoretische Verteilung beschrieben werden. Als Alternative kann der p-Wert für den χ^2-Anpassungstest mit einer der beiden folgenden EXCEL-Funktion berechnet werden:

$$p = CHITEST(n_i;\ e_i) \quad Cave: setzt\ \nu = n - 1\ voraus$$

$$p = CHIVERT(x^2;\ \nu)$$

Wird also angestrebt, dass H0 beibehalten wird (im Gegensatz zu den folgenden Signifikanztests), dann muss p > α (dem gewählten Signifikanzniveau) sein.

7.6 Prüfung auf Unabhängigkeit (für nominale und ordinale Daten)

- Unabhängigkeitsprüfung
  Bei der Prüfung auf Unabhängigkeit für nominale und ordinale Daten bedient man sich dem χ^2-Unabhängigkeitstest (wird auch χ^2-Vierfeldertest genannt). Dieser ist äquivalent zum χ^2-Anpassungstest aber für Kontingenztabellen und somit für nominale und ordinale Daten formuliert.

beobachtete Häufigkeiten

		X		
		x_1	x_2	Σ
Y	y_1	n_{11}	n_{21}	$n_{.1}$
	y_2	n_{12}	n_{22}	$n_{.2}$
	Σ	$n_{1.}$	$n_{2.}$	n

beobachtete Häufigkeiten

		X		
		x_1	x_2	Σ
Y	y_1	e_{11}	e_{21}	$e_{.1}$
	y_2	e_{12}	e_{22}	$e_{.2}$
	Σ	$e_{1.}$	$e_{2.}$	n

- χ^2-Unabhängigkeitstest (für Vierfeldertafel = $2 * 2$ Kontingenztafel)
  Er ist der in der Medizin am häufigsten angewendete statistische Test zum Nachweis, dass zwei Merkmale unabhängig sind. So wird für den Spezialfall einer $2 * 2$ Kontingenztabelle die Testgröße χ^2 wie folgt berechnet:

$$\chi^2 = n * \frac{(n_{11} * n_{22} - n_{12} * n_{21})^2}{n_{1.} * n_{2.} * n_{.1} * n_{.2}}$$

Sie wird mit dem kritischen Wert verglichen:

$$\chi^{2*} = CHIINV(\alpha;\ \nu) \quad mit: \quad \nu = 1$$

So strebt man für eine vorgegebene Irrtumswahrscheinlichkeit von beispielsweise α = 0,05 an: Testgröße > kritischer Wert ($\chi^2 > \chi^{2*}$) und belegt damit einen signifikanten Zusammenhang der beiden Merkmale X und Y.

- χ^2-Unabhängigkeitstest (für $l * m$ Kontingenztabellen)
  Für Kontingenztabellen beliebiger Dimension wird die Testgröße χ^2 analog zum χ^2-Anpassungstest berechnet, und lautet wie folgt:

$$\chi^2 = \sum_{i=1}^{l} \sum_{i=j}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

  Gleich dem χ^2-Anpassungstest, wird hier die Summe der Abweichungsquadrate (der Häufigkeit von den Erwartungswerten) berechnet und mit einem kritischen Wert verglichen. Hierfür werden zunächst die Erwartungswerte e_{ij} berechnet:

$$e_{ij} = \frac{n_{i.} * n_{.j}}{n} \qquad mit\ den\ Randsummen\ n_{i.}\ und\ n_{.j}$$

  Die Werte werden dann in eine Kontingenztabelle für die erwarteten Häufigkeiten eingetragen.

- Nullhypothese H_0
  Die Nullhypothese H_0 ist die Hypothese, dass zwei Merkmale X und Y unabhängig sind. Demzufolge ist bei einem vorgegebenem Signifikanzniveau α die Nullhypothese H_0 abzulehnen (d.h. die Merkmale sind abhängig). Dies gilt dann, wenn die Testgröße χ^2 größer als der kritische Wert χ^{2*} ist χ^{2*} und wird wie folgt berechnet:

$$\chi^2 = CHIINV(\alpha;\ \nu) \quad mit: \quad \nu = (l-1) * (m-1)$$

  Alternativ dazu wird H_0 abgelehnt, wenn $p < \alpha$ ist:

$$p = CHITEST(n_{ij};\ e_{ij})$$

  Werden demzufolge zwei Merkmale als abhängig eingestuft, so stellt sich die Frage nach der Stärke dieser gegenseitigen Abhängigkeit. Diese kann mit r_ϕ und r_v berechnet werden. Aufgrund dessen ist der χ^2- Unabhängigkeitstest auch ein Signifikanztest für die Korrelationskoeffizienten r_ϕ und r_v.

7.7 Signifikanztest für die Korrelationskoeffizienten r und r_s

- Mit dem Signifikanztest für die Korrelationskoeffizienten r und r_s prüft man ob der Korrelationseffizient nach PEARSON signifikant ist (Nullhypothese H_0: r = 0) versus Alternativhypothese H_1: $r \neq 0$). Berechnung der Testgröße t:

$$t = \frac{r \quad \sqrt{n-2}}{\sqrt{1-r^2}}$$

So kann der kritische Wert t^* und der p-Wert wie zuvor mit den entsprechenden EXCEL-Funktionen berechnet werden:

$$t^* = TINV(\alpha;\ \nu) \quad mit: \quad \nu = n - 2$$

$$p = TVERT(t;\ \nu; 2)$$

Daraus folgt, dass r für $|\ t\ | \geq\ t^*$ oder $p\ \leq\ \alpha$ signifikant von Null abweicht.

- r^* Tabelle
  Korrelationskoeffizienten können sehr einfach mit Tabellen der kritischen Korrelation r* auf Signifikanz überprüft werden. Dabei ist r dann signifikant, wenn für ein vorgegebenes Signifikanzniveau α bei gegebenen n (bzw. ν) gilt: $|\ r\ | \geq r^*$.

$$r^* = \frac{1}{\sqrt{\frac{\nu}{t^{*2}} + 1}}$$

- Signifikanztest für den Rangkorrelationskoeffizienten r_S
  Oftmals wird für den Korrelationskoeffizienten nach SPEARMAN (Kapitel 4) der gleiche Signifikanztest wie für die Korrelation nach PEARSON vorgeschlagen. Einige Statistikbücher enthalten jedoch auch leicht alternierende Tabellen für die kritische Korrelation r_s^*, wie dir r_s^*-Tabelle nach Lorenz (1988).

7.8 Signifikanztests für zwei Stichproben (t-Test, WELCH-Test)

- Einführung
  In weiterer Folge werden verschiedene Signifikanztests für metrische Daten aus zwei Stichproben besprochen. Man unterscheidet dabei prinzipiell zwischen verbundenen (abhängigen, gepaarten) und unverbundenen (unabhängigen) Stichproben.

- Verbundene Stichproben
  Die Merkmale einer Stichprobe werden vor und nach einem Ereignis verglichen, z.B. Patienten werden vor und nach einer definierten medizinischen Behandlung untersucht oder Pflanzen werden zu einem definierten Zeitpunkt abgemessen und nach weiterer Wachsdauer erneut abgemessen. Daraufhin werden die Ergebnisse verglichen.

- Unverbundene Stichprobe
  Bei einer unverbundenen Stichprobe wird eine Stichprobe in zwei experimentelle Gruppen geteilt, die verglichen werden. Hierbei können z.B. Patienten einer Versuchsgruppe (behandelt) und einer Kontrollgruppe (unbehandelt) verglichen werden.

- t-Test für verbundene Stichproben
  Mit dem statistischen Test t-Test prüft man auf den signifikanten Unterschied der mittleren Differenz der gepaarten x-Werte. Demnach wird die Testgröße t wie folgt berechnet:

$$t = \frac{m(d_i)}{s(d_i)} * \sqrt{n}$$

  d_i = Differenz der gepaarten x-Werte
  $m(d_i)$ = Mittelwert der Differenz der gepaarten x-Werte
  $s(d_i)$ = Standardabweichung der Differenz der gepaarten x-Werte

  Folgende EXCEL-Funktion ermöglicht die Berechnung des kritischen Wertes t^*:

$$t^* = TINV(\alpha;\ \nu) \quad mit: \quad \nu = n_1 + n_2 - 2$$

$$p = TTEST(Stichprobe_1;\ Stichprobe_2; 2; 2)$$

- WELCH-Test für zwei unverbundene Stichproben
  Mit dem WELCH-Test prüft man auf den signifikanten Unterschied zweier Stichproben in Bezug auf ihre Mittelwerte. Die Testgröße t wird wie folgt berechnet:

$$t = \frac{m_1 - m_2}{\sqrt{se_1^2 + se_2^2}} \quad mit: \quad se = \frac{s}{\sqrt{n}}$$

  Dabei wird der kritische Wert t^* mit der EXCEL-Funktion berechnet, kann jedoch auch aus Tabellen, wie sie in den meisten Statistikbüchern am Anfang zu finden sind, entnommen werden:

$$t^* = TINV(\alpha;\ \nu)$$

  Für $|t| \geq t^*$ unterscheiden sich die Stichproben signifikant.

  Dabei ist die Berechnung der Freiheitsgrade ν deutlich aufwendiger und erfolt mit:

$$\nu = \frac{1}{\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}} \quad mit: \quad c = \frac{se_1^2}{se_1^2 + se_2^2}$$

  Für den WELCH-Test gelten dabei jedoch die Vorrausetzungen (annähernd), dass es sich um normalverteilte Stichproben handelt.

  Dabei kann der p-Wert wieder mit einer EXCEL-Funktion ermittelt werden, die zum Ausdruck bringt, dass dieser Test alternativ auch als t-Test bezeichnet wird. Die Berechnung des p-Wertes erfolgt mittels:

$$p = TTEST\ (Stichprobe_1;\ Stichprobe_2; 2; 3)$$

Literaturangabe

Das Skriptum basiert auf dem Wissen, das in der Lehrveranstaltung 804603 (Grundlagen der Statistik und Epidemiologie) vermittelt wird.

Sämtliche Fakten entstammen dem Lehrskript dieser Lehrveranstaltung:
„Grundlagen der Statistik und Epidemiologie Teil 1, Franz Rubel (Lehrveranstaltungsleiter), Katharina Brugger, Günther Schauberger, Melanie Walter, SS 2017“