Wirtschaftsstatistik

Universität Ulm
Abteilung Stochastik

Vorlesungsskript
Prof. Dr. Volker Schmidt
Stand: Sommersemester 2004

Ulm, im Juli 2004

Contents

1 Einleitung

- In dieser Vorlesung werden Begriffe und Methoden der Wahrscheinlichkeitsrechnung und Statistik weiter vertieft, die teilweise bereits im Grundkurs *"Stochastik für Wirtschaftswissenschaftler"* eingeführt worden sind.
- Das Ziel der *Vorlesung Wirtschaftsstatistik* besteht vor allem darin, solche Begriffe und Methoden der
  - *beschreibenden Statistik*

  sowie der
  - *beurteilenden Statistik*

  auf anschauliche Weise zu behandeln, die bei der *Gewinnung, Darstellung und Beschreibung von Wirtschaftsdaten* sowie bei deren *Analyse, Bewertung und Interpretation* nützlich sind.
- Die *beschreibende Statistik* (auch deskriptive Statistik bzw. explorative Datenanalyse genannt) befasst sich hauptsächlich damit,
  - Daten, die aus der Beobachtung von interessierenden Sachverhalten, Objekten bzw. Vorgängen gewonnen werden, in geeigneter Form darzustellen und zu beschreiben.
- Bei der Untersuchung von großen Datenmengen geht es dabei oft um eine geeignete Strukturierung der Daten, beispielsweise um eine geeignete Zusammenfassung/Komprimierung von Teildatensätzen bzw. deren graphische Darstellung, um somit wesentliche Eingenschaften der Daten stärker hervorzuheben (und Nebensächliches in den Hintergrund zu rücken).
- Die *beurteilende Statistik* (auch schließende, induktive bzw. inferentielle Statistik genannt) nutzt
  - *Begriffe und Methoden der mathematischen Stochastik.*
- Hierdurch wird eine wesentlich tiefergehende *Analyse, Bewertung und Interpretation der Daten* ermöglicht.

Die Vorlesung *Wirtschaftsstatistik* besteht aus zwei Teilen:

1. *Grundideen der beschreibenden Statistik*, insbesondere
   - Methoden der Datengewinnung (Datenerhebung, experimentelle Datengewinnung, Monte-Carlo-Simulation),
   - Kenngrößen zur Beschreibung von univariaten Daten (Maßzahlen für Lage und Streuung der Daten, Konzentrationsmaße, relative Häufigkeiten, empirische Verteilungsfunktion),
   - Kenngrößen zur Beschreibung von bivariaten Daten (Kontingenztafeln, Zusammenhangsmaße)
2. *Regressions- und Varianzanalyse*, insbesondere
   - einfache bzw. multiple lineare Regression (Schätzen und Testen von Modellparametern; Prognose von Zielgrößen)
   - einfaktorielle bzw. mehrfaktorielle Varianzanalyse (mit festen Effekten)

Die folgende Liste von einführenden Lehrbüchern umfasst lediglich eine kleine Auswahl von Texten, die neben dem Vorlesungsskript für ein ergänzendes und vertiefendes Studium empfohlen werden können.

- Bamberg, G. und Baur, F., *Statistik.* Oldenburg-Verlag, München, 2001.
- Bosch, K., *Elementare Einführung in die angewandte Statistik.* Vieweg-Verlag, Braunschweig, 2000.
- Bosch, K., *Statistik-Taschenbuch (3. Aufl.).* Oldenbourg-Verlag, München, 1998
- Fahrmeir, L., Künstler, R., Pigeot, I. und Tutz, G., *Statistik: Der Weg zur Datenanalyse (4. Aufl.).* Springer-Verlag, Berlin, 2002.
- Hartung, J., Elpelt, B. und Klösener, K.-H., *Statistik: Lehr- und Handbuch der angewandten Statistik.* Oldenbourg-Verlag, München, 2002.
- Kazmier, L.J., *Wirtschaftsstatistik.* McGraw-Hill, London, 1996.
- Mosler, K. und Schmid, F., *Beschreibende Statistik und Wirtschaftsstatistik.* Springer-Lehrbuch, Berlin 2003.
- Schlittgen, R. und Streitberg, B., *Zeitreihenanalyse.* Oldenbourg-Verlag, München, 1999.
- Storm, R., *Wahrscheinlichkeitsrechnung, mathematische Statistik und Qualitätskontrolle.* Hanser-Fachbuchverlag, Leipzig, 2001.

2 Grundideen der beschreibenden Statistik

2.1 Methoden der Datengewinnung

Die Gewinnung von Daten erfolgt durch die Beobachtung/Registrierung interessierender *Merkmale* von

- realen Sachverhalten, Objekten bzw. Vorgängen in Wirtschaft, Industrie, Natur und Gesellschaft (durch direkte Messung, Inventur, Befragung, etc.),
- Ergebnissen aus
  - Labor- bzw. Feldversuchen,
  - virtuellen Experimenten per Computer-Simulation.

Die beobachteten Merkmale werden dabei durch *Kenngrößen* bzw. *Variablen* ausgedrückt.

2.1.1 Klassifikation von Merkmalen/Kenngrößen/Variablen

Es gibt mehrere unterschiedliche Möglichkeiten zur *Klassifikation* von Merkmalen bzw. der zugehörigen Kenngrößen/Variablen:

1. *Diskrete und stetige Variablen*
   - Eine Möglichkeit, Merkmale bzw. die zugehörigen Kenngrößen/Variablen zu klassifizieren, besteht darin, die *Anzahl der Ausprägungen/Werte* zu betrachten, die die Merkmale/Kenngrößen/Variablen annehmen können.
     - Wenn die Menge der angenommenen Werte *endlich* oder *abzählbar unendlich* ist – zum Beispiel die (endliche) Menge $\{1, \ldots, n\}$ der ersten n natürlichen Zahlen oder die (abzählbar unendliche) Menge $\mathbb{N} = \{1, 2, \ldots\}$ *aller* natürlichen Zahlen – dann heißt das Merkmal bzw. die zugehörige Kenngröße/Variable *diskret*.
     - Ansonsten, d.h., wenn die Menge der angenommenen Werte *nicht abzählbar* ist – zum Beispiel die Menge $\mathbb{R} = (-\infty, \infty)$ aller reellen Zahlen oder das Intervall $[0, 1]$ –, dann spricht man von *stetigen* Variablen.
   - Beispiele
     - Diskrete Variablen treten oft bei der Registrierung von *Anzahlen* auf; beispielsweise bei der Beobachtung von Personen- bzw. Objektgruppen mit bestimmten Eigenschaften.
     - Aber auch bei der *Gruppierung* bzw. *Kategorisierung* von Ausprägungen, die eigentlich eher von stetiger Natur sind, treten diskrete Variablen auf.
     - So kann bei großen Datenmengen die Gruppierung der *Rohdaten* zur Erhöhung der Übersichtlichkeit beitragen.
     - Darüber hinaus kann die Darstellung von Variablen, die eigentlich stetig sind, als diskrete Variablen beispielsweise auch auf dem begrenzten Auflösungsvermögen von Messgeräten beruhen.
     - Solche Zwischenformen von Merkmalen/Kenngrößen/Variablen werden manchmal *quasi-stetige Variablen* genannt.
2. *Klassifikation gemäß dem Skalenniveau*
   - Eine andere Art, Merkmalen/Kenngrößen/Variablen zu klassifizieren, beruht auf dem *Skalenniveau*, auf dem ein Merkmal „gemessen" wird.
     - Ein Merkmal bzw. die zugehörige Kenngröße/Variable heißt *nominalskaliert*, wenn die Ausprägungen *Namen* oder *Kategorien* sind, die keine lineare Ordnung aufweisen. Den Ausprägungen werden dennoch meistens (natürliche) Zahlen zugeordnet, die jedoch lediglich der Kodierung dienen und keine numerischen Werte im üblichen Sinne sind.

  - Im Gegensatz hierzu heißt ein Merkmal bzw. die zugehörige Kenngröße/Variable *ordinalskaliert*, wenn die Ausprägungen (beispielsweise der Größe nach) geordnet werden können.
  - Wenn die Ausprägungen linear geordnet werden können und wenn die *Differenzen zwischen den Ausprägungen* eine einheitliche Interpretation besitzen, dann spricht man von *intervallskalierten* Merkmalen/Kenngrößen/Variablen (bzw. von der *Intervallskala* der Ausprägungen).
  - Wenn darüber hinaus die *Quotienten der Ausprägungen* von intervallskalierten Merkmalen/Kenngrößen/Variablen ebenfalls eine inhaltliche Interpretation besitzen, dann spricht man von *verhältnisskalierten* Merkmalen/Kenngrößen/Variablen.
  - Merkmale/Kenngrößen/Variablen, die sowohl intervall- als auch verhältnisskaliert sind, nennt man auch *kardinalskaliert* bzw. *metrisch skaliert*.
- Besipiele
  - Typischerweise wird die Zugehörigkeit von Personen bzw. Objekten zu bestimmten Gruppen durch nominalskalierte Variablen beschrieben. Beispiele hierfür sind die Staatsbürgerschaft, die Religionszugehörigkeit, das Geschlecht von Personen bzw. der Verwendungszweck oder das Herkunftsland von Produkten etc.
  - Als typische Beipiele für ordinalskalierte Variablen, die *nicht* intervallskaliert sind, gelten Schulnoten bzw. Warnstufen (etwa bei Sturm- oder Lawinengefahr). Denn der Abstand zwischen den Noten 1 und 2 hat beispielsweise eine völlig andere Bedeutung als der Abstand zwischen den Noten 4 und 5.
  - Ein Beispiel für intervallskalierte Variablen, die jedoch *nicht* verhältnisskaliert sind, ist die in Grad Celsius gemessene Temperatur, bei der man keinen sinnvollen Nullpunkt angeben kann. Somit besitzen Temperaturquotienten keine sinnvolle quantitative Interpretion.
  - Typische Beispiele für verhältnisskalierte Variablen sind der aktuelle Wert einer Währung (beispielsweise gegenüber dem Euro), der Luftdruck (bezogen auf den Normaldruck von 1 at).

3. *Qualitative und quantitative Variablen*

- Außerdem werden Merkmale/Kenngrößen/Variablen danach unterschieden, ob sie eher eine Qualitätsstufe oder ein Ausmaß beschreiben.
  - Von *qualitativen* Variablen spricht man, wenn es nur *endlich viele* Ausprägungen gibt und wenn diese Ausprägungen nominalskaliert sind.
  - Darüber hinaus sieht man auch ordinalskalierte Merkmale als qualitativ an, wenn die Ausprägungen eher eine Qualitätsstufe als ein Ausmaß darstellen.
  - Bei Ausprägungen, die eher ein Ausmaß bzw. eine Intensität darstellen, spricht man dagegen von *quantitativen* Variablen.
- Beispiele
  - Die obengenannten Zugehörigkeitsrelationen werden ausnahmslos durch qualitative (nominalskalierte) Variablen erfasst.
  - Ordinalskalierte Variablen können entweder qualitative, aber auch quantitative Variablen sein. Die obengenannten Beispiele der Schulnoten bzw. Warnstufen sind eher qualitativer Natur.
  - Dagegen führt das Messen von betriebsswirtschaftlichen Ergebnissen (wie Umsatz, Gewinn, Verlust etc.) zu quantitativen Variablen. Dies gilt auch für physikalische Kenngrößen wie Temperatur oder Druck.

4. *Univariate und multivariate Variablen*

- Falls die Bewertung der interessierenden Merkmale/Kenngrößen/Variablen durch reelle Zahlen erfolgt, dann spricht man von *univariaten Variablen*.
- Dieser Fall tritt dann auf, wenn nur ein einzelnes Merkmal/Kenngröße/Variable interessiert.
- Falls jedoch die Bewertung der interessierenden Merkmale/Kenngrößen/Variablen durch mehrdimensionale Vektoren von reelle Zahlen erfolgt, dann spricht man von *multivariaten Variablen*.

- Ein typisches Beispiel für diese Situtation liegt dann vor, wenn man sich gleichzeitig für *mehrere* (uni- bzw. multivariate) Merkmale/Kenngrößen/Variablen interessiert.

Zusammenfasung: *Typen von Merkmalen/Kenngrößen/Variablen*

diskret	endlich oder abzählbar unendlich viele Ausprägungen/Werte
stetig	alle (reellen) Zahlen eines Intervalls können mögliche Ausprägungen/Werte sein
nominalskaliert	Ausprägungen sind Namen, Ordnung nicht sinnvoll
ordinalskaliert	Ausprägungen/Werte können geordnet, aber Abstände nicht interpretiert werden
intervallskaliert	Ausprägungen/Werte sind Zahlen, deren Abstände interpretiert werden können
verhältnisskaliert	Ausprägungen/Werte besitzen absoluten Nullpunkt, der sinnvoll interpretiert werden kann
kardinalskaliert bzw. *metrisch skaliert*	Merkmale/Kenngrößen/Variablen, die sowohl intervall- als auch verhältnisskaliert sind,
qualitativ	endlich viele Ausprägungen/Werte, typischerweise nominalskaliert, gegebenenfalls auch ordinalskaliert
quantitativ	Ausprägungen/Werte stellen Ausmaß bzw. Intensität
univariat	Ausprägungen/Werte sind reelle Zahlen
multivariat	Ausprägungen/Werte sind Vektoren

2.1.2 Datenerhebung

Wenn die Gewinnung von Daten durch die Beobachtung/Registrierung interessierender Merkmale/Kenngrößen/Variablen von realen Sachverhalten, Objekten bzw. Vorgängen (durch direkte Messung, Inventur, Befragung, etc.) erfolgt, dann spricht man von *Datenerhebung*. Dabei gibt es verschiedene Arten der Datenerhebung.

1. *Primär- und sekundärstatistische Erhebung*
   - Man spricht von *primärstatistischer Erhebung*, wenn die Daten speziell im Zusammenhang mit der interessierenden Fragestellung erhoben werden, bzw. von
   - *sekundärstatistischer Erhebung*, wenn die Daten bereits vorhanden sind und beispielsweise lediglich aus größeren Datenbeständen/Datenbanken extrahiert werden müssen.
   - Beachte:
     - Wenn nicht die Rohdaten selbst, sondern nur vorverarbeitete (beispielsweise aggregierte/komprimierte) Daten vorliegen, dann spricht man von *tertiärstatistischer Datenerhebung*.
     - In der Regel sind Rohdaten besser als stark aggregierte/komprimierte Daten für statistische Analysen geeignet.
2. *Fehlende Daten*
   - Durch das Zusammenfassen, d.h. das Aggregieren bzw. Komprimieren von Daten entsteht stets ein Informationsverlust.
   - Ein ähnliches Problem sind *fehlende Daten*. Dieser Effekt tritt häufig bei der Befragung von Personen auf und kann die Ergebnisse statistischer Analysen negativ beeinflussen.

- Ein anderes Beispiel, bei dem die Problematik fehlender Daten typischerweise vorkommt, ist die statistische Bildanalyse. Denn die Daten, die den Rand von digitalen Bildern beschreiben, sind oft weniger informativ als die Daten, die den inneren Bildbereich beschreiben.
- Wenn der Entstehungsmechanismus bekannt ist, der zum Fehlen von Daten führt, dann kann dies allerdings durch eine entsprechende Wahl der statistischen Analyse-Methoden korrigiert werden.
- Zum Beispiel werden bei der Analyse von Bilddaten sogenannte *randkorrigierte Schätzer* betrachtet.

3. *Teil- und Vollerhebung*

- Häufig wird die Beobachtung interessierender Merkmale/Kenngrößen/Variablen lediglich für einen Teil der (insgesamt verfügbaren) realen Sachverhalte, Objekte bzw. Vorgänge durchgeführt.
- Man spricht dann vom Ziehen einer *Stichprobe* aus der insgesamt (zumindest prinzipiell) verfügbaren Grundgesamtheit von Sachverhalten/Objekten/Vorgängen bzw. von einer *Teilerhebung*.
- Manchmal (beispielsweise bei Volkszählungen) werden die interessierenden Merkmale/Kenngrößen/Variablen jedoch für sämtliche Einheiten der Grundgesamtheit beobachtet/registriert. In diesem Fall spricht man von einer *Vollerhebung*.
- Ein typisches Beispiel für Teilerhebungen ist die *Qualitätskontrolle* von Produkten, bei der eine Vollerhebung häufig aus Kostengründen bzw. wegen des hohen zeitlichen Auswandes nicht sinnvoll ist.

2.1.3 Gewinnung von experimentellen Daten

- Eine andere Methode der Datengewinnung beruht auf der Durchführung von *Experimenten* im Rahmen von Labor- bzw. Feldversuchen:
  - Beobachtet, d.h. registriert werden dabei die interessierenden Merkmale/Kenngrößen/Variablen der Versuchsergebnisse.
  - Typische Beispiele sind biologische, physikalische bzw. chemische Experimente, die in Forschungslabors durchgeführt werden.
  - Bei wirtschaftswissenschaftlichen Studien werden die Experimente vorwiegend als Feldversuche durchgeführt, wobei wesentlich umfangreichere Stichproben (beispielsweise von potentiellen Kunden) als bei Laborversuchen betrachtet werden.
  - Das Ziel solcher Studien können zum Beispiel *Marktanalysen* sein, bei denen die Nachfrage nach (neu entwickelten) Produkten untersucht wird.
  - Experimentelle medizinische Studien von (zufällig ausgewählten) Patientengruppen können sowohl als klinische Studie (d.h. als Laborversuch) bzw. als Feldstudie durchgeführt werden.
  - Diese Studien können beispielsweise das Ziel haben, die Wirkung von neu entwickelten Medikamenten bzw. Therapien zu beurteilen.
- Naturwissenschaftliche Experimente weisen *wesentliche Unterschiede* im Vergleich zu wirtschaftswissenschaftlichen bzw. medizinischen Experimenten auf:
  - Während bei naturwissenschaftlichen Experimenten unter *gleichbleibenden Bedingungen* kontrolliert Daten gesammelt werden können, ist dies bei wirtschaftswissenschaftlichen und medizinischen Untersuchungen oftmals nicht möglich.
  - Ein Grund hierfür ist, dass wirtschaftswissenschaftliche und medizinische Untersuchungen teilweise auf *Personenbefragungen* beruhen, wodurch die Kontrollierbarkeit der Versuchsbedingungen nur in einem geringerem Maße als bei naturwissenschaftlichen Experimenten gewährleistet ist.

Die Erzeugung von Daten im Zusammenhang mit der Durchführung von Experimenten muss sorgfältig vorbereitet werden. Man spricht in diesem Zusammenhang auch von *Versuchsplanung*. Dabei sind insbesondere die folgenden Punkte zu beachten.

1. *Ziele der Untersuchungen*

   - Zunächst müssen die Ziele der Untersuchungen abgesteckt werden.
   - Wesentlich für die Effizienz und den Erfolg der Untersuchungen ist, dass sämtliche Arbeiten bereits in diesem frühen Stadium in *enger Abstimmung* zwischen dem Statistiker und seinen Kooperationspartnern erfolgen.
   - Insbesondere müssen zunächst die relevanten Merkmale/Kenngrößen/Variablen der ínteressierenden Sachverhalte/Objekte/Vorgänge spezifiziert werden.
   - Eine anschließende *Literatur-Recherche* (beispielsweise per Internet) kann Informationen darüber liefern, ob ähnliche Fragestellungen bereits in anderen Projekten untersucht worden sind.
   - Im Ergebnis hiervon können schon *vorhandene Modellierungs- und Lösungsanätze* mit in die eigenen Untersuchungen einbezogen werden.
   - Die Literatur-Recherche kann jedoch auch dazu führen, dass die ursprünglich festgelegten Ziele der Untersuchungen präzisiert bzw. korrigiert werden müssen.

2. *Planung der Experimente*

   - Nachdem die Ziele der Untersuchungen festgelegt worden sind, sollte eine sorgfältige Planung der Experimente folgen.
   - Dabei hängt die Planung der Rahmenbedingungen, des Umfanges bzw. der Zeitdauer, die für die Durchführung der Experimente veranschlagt werden, natürlich von den vorhandenen finanziellen und technischen Ressourcen ab.
   - Bei wirtschaftswissenschaftlichen und medizinischen Untersuchungen, die auf Personenbefragungen beruhen, ist zunächst das Vorgehen bei der Auswahl der Probanden festzulegen.
   - In diesem Zusammenhang ist die *Randomisierung* der Stichprobe ein wichtiges Prinzip, um Verfälschungen (beispielsweise durch die bewusste Auswahl von besonders geeigneten Personen) zu vermeiden und somit eine *repräsentative Stichprobe* von Probanden zu erhalten.
   - Bei medizinischen Studien wird, zusätzlich zu der eigentlichen Gruppe der behandelten Probanden, oftmals noch eine *Placebo- bzw. Kontrollgruppe* betrachtet, um eine möglichst objektive Beurteilung der Versuchsergebnisse zu ermöglichen.
   - Bei der Planung der Experimente müssen auch bereits erste Vorstellungen darüber entwickelt werden,
     - welche relevanten Merkmale/Kenngrößen/Variablen der ínteressierenden Sachverhalte/Objekte/Vorgänge auf welche Weise untersucht werden sollen,
     - welche Merkmale/Kenngrößen/Variablen als (einstellbare) Ausgangsgrößen und welche Merkmale/Kenngrößen/Variablen als (beobachtbare) Zielgrößen aufgefasst werden,
     - wie die gewonnenen Daten dargestellt und statistisch verarbeitet werden sollen.
   - Um mehr Planungssicherheit für umfangreiche Feldversuche zu erlangen, ist es manchmal sinnvoll, zunächst eine *Pilotstudie* durchzuführen. Die Ergebnisse solcher Pilotstudien können dann zur Planung der eigentlichen Experimente verwendet werden.

3. *Erfassung/Protokollierung der Daten*

   - Die Protokollierung der Daten beginnt mit der Darlegung der Ziele der Untersuchungen und der Dokumentation der Versuchsplanung.
   - Das Protokoll sollte außerdem Angaben über die verwendeten Methoden zur Darstellung und statistischen Verarbeitung der Daten enthalten.
   - Die gewonnenen Daten, die aus den Einstellungen bzw. Messungen sämtlicher Einfluss- bzw. Zielgrößen resultieren, sollten möglichst detailliert erfasst und protokolliert werden.
   - Das Aggregieren bzw. Komprimieren von Rohdaten, das bei umfangreichen Feldversuchen aus Kapazitätsgründen geboten sein kann, führt stets zu Informationsverlusten und sollte deshalb nur bei begründeter Notwendigkeit erfolgen.

- Bei wirtschaftswissenschaftlichen und medizinischen Untersuchungen, die auf Personenbefragungen beruhen, sollten sämtliche relevanten Daten über die Probanden im Protokoll erfasst werden (auch über diejenigen Probanden, die im Verlauf der Experimente aus der Studie austreten).

2.1.4 Datengewinnung durch Computer-Simulation

- Neben der Gewinnung von experimentellen Daten im Rahmen von realen Labor- bzw. Feldversuchen erlangte die Erzeugung sogenannter *synthetischer Daten* durch Computer-Simulation in den letzten Jahren eine immer größere Bedeutung.
- Die Gründe für die zunehmende Nützlichkeit von Computer-Simulationen bei der Untersuchung von interessierenden Sachverhalten/Objekten/Vorgängen sind vielfältig:
  - An erster Stelle ist hier natürlich das rasant wachsende Leistungsvermögen moderner Computer-Systeme zu nennen, das sich in den letzten Jahren in einem ungeahnten Ausmaß weiterentwickelt hat und dabei Möglichkeiten eröffnet, die noch vor kurzem völlig unvorstellbar waren.
  - Im Zusammenhang damit ist die Datengewinnung durch Computer-Simulation oft viel *kostengünstiger* und mit *weniger Zeitaufwand* verbunden als die herkömmliche Gewinnung von experimentellen Daten im Rahmen von realen Labor- bzw. Feldversuchen.
  - Außerdem lassen sich (virtuelle) Computer-Experimente unter gleichbleibenden Versuchsbedingungen beliebig oft wiederholen, wogegen beispielsweise bei naturwissenschaftlichen Experimenten das untersuchte Objekt während der Versuche oft beschädigt bzw. zerstört wird.
- Ein weiterer Grund für die Nützlichkeit von Computer-Simulationen besteht darin, dass
  - der Umfang und die Struktur der zu analysierenden Datensätze oft sehr komplex ist,
  - die Verarbeitung und Bewertung der Daten dann typischerweise auf mathematischen Modellen beruht, deren charakteristische Eigenschaften nicht (oder nur teilweise) mit geschlossenen analytischen Formeln beschrieben werden können,
  - die Computer-Simulation der betrachteten Modelle in diesem Fall ein nützliches (alternatives) Analyse-Tool ist.

- Computer-Experimente zur Untersuchung von interessierenden Sachverhalten/Objekten/Vorgängen beruhen auf *stochastischen Simulationsalgorithmen*. Man spricht deshalb auch von *Monte-Carlo-Simulation*. Dabei gibt es unterschiedliche Arten von Simulationsalgorithmen.

  1. Die Grundlage zur Monte-Carlo-Simulation von (einzelnen) Merkmalen/Kenngrößen/Variablen bilden *Zufallszahlen-Generatoren*.
     - Dies sind Algorithmen, durch die Realisierungen von Zufallsvariablen per Computer erzeugt werden können, die *Pseudozufallszahlen* genannt werden.
     - Den Ausgangspunkt bilden dabei sogenannte *Standard-Zufallszahlen-Generatoren*, durch die Realisierungen von Zufallsvariablen erzeugt werden können, die auf dem Einheitsintervall $[0, 1]$ gleichverteilt sind, sogenannte *Standard-Pseudozufallszahlen*.
     - Hiervon ausgehend lassen sich dann durch gewisse *Transformations- bzw. Verwerfungsmethoden* auch Pseudozufallszahlen für Zufallsvariablen mit anderen Verteilungen erzeugen, zum Beispiel für binomialverteilte, Poisson-verteilte oder normalverteilte Zufallsvariablen.

  2. Computer-Experimente zur *Untersuchung der zeitlichen Entwicklung* von Sachverhalten/Objekten/Vorgängen beruhen auf anspruchsvolleren Algorithmen der *dynamischen* Monte-Carlo-Simulation.
     - Eine zentrale Rolle spielen dabei die Algorithmen der *Markov-Chain-Monte-Carlo-Simulation* (MCMC-Simulation), durch die *zeitstationäre Gleichgewichtszustände* von Sachverhalten/Objekten/Vorgängen näherungsweise simuliert werden können.

- Ein anderes Beispiel, bei denen Algorithmen der MCMC-Simulation angewendet werden, ist die *statistische Analyse von Bilddaten*.
- Eine aktuelle Forschungsthematik, zu der in den letzten Jahren zahlreiche Publikationen veröffentlicht wurden, sind sogenannte *Kopplungsalgorithmen der perfekten MCMC-Simulation*.
- Durch solche Kopplungsalgorithmen können beispielsweise zeitstationäre Gleichgewichtszustände von Sachverhalten/Objekten/Vorgängen nicht nur näherungsweise, sondern in einem gewissen Sinne „perfekt" simuliert werden können.

2.2 Kenngrößen zur Beschreibung von univariaten Daten

- Die beschreibende Statistik stellt eine Reihe von Kenngrößen bereit, durch die wesentliche Eigenschaften bzw. Gesetzmäßigkeiten von (großen) Datensätzen auf übersichtliche Weise dargestellt werden können. Der Vektor $(x_1, \ldots, x_n)$ der vorliegenden Daten $x_1, \ldots, x_n$ kann dabei im allgemeinen eine komplizierte Struktur aufweisen.
- Der „Wert" x_i muss nämlich nicht unbedingt eine Zahl sein, sondern x_i kann für jedes $i = 1, \ldots, n$ ein Vektor oder eine Matrix sein, die beispielsweise die Ausprägungen mehrerer Merkmale gleichzeitig beschreiben können.
- In diesem Abschnitt setzen wir jedoch voraus, dass $x_i \in \mathbb{R}$ für jedes $i = 1, \ldots, n$, d.h., wir betrachten sogenannte *univariate Daten*.
- Wir beginnen zunächst mit der Einführung einiger allgemeiner *Grundbegriffe* der beschreibenden Statistik.
  - Der Datenvektor $(x_1, \ldots, x_n)$ wird (konkrete) *Stichprobe* genannt.
  - Die Menge $C \subset \mathbb{R}^n$ aller (potentiell möglichen) Stichproben $(x_1, \ldots, x_n)$ heißt *Stichprobenraum*.
  - Für jedes $i = 1, \ldots, n$ nennt man x_i den i-ten *Stichprobenwert* von $(x_1, \ldots, x_n)$.
  - Die Anzahl n der Komponenten von $(x_1, \ldots, x_n)$ heißt *Stichprobenumfang*.
  - Unter einer *Kenngröße der Stichprobe* $(x_1, \ldots, x_n)$ verstehen wir dabei eine Abbildung

$$(x_1, \ldots, x_n) \longmapsto \varphi(x_1, \ldots, x_n)\,, \tag{1}$$

    die jeder Stichprobe $(x_1, \ldots, x_n)$ einen „Kennwert" $\varphi(x_1, \ldots, x_n)$ zuordnet.
  - Die in (1) betrachtete Abbildung wird auch *Stichprobenfunktion* genannt.

2.2.1 Lagemaßzahlen

In diesem Abschnitt betrachten wir eine spezielle Klasse von Kenngrößen der Stichprobe $(x_1, \ldots, x_n)$, sogenannte *Maßzahlen*, die die Lage der Daten beschreiben und die kurz *Lagemaßzahlen* genannt werden.

1. *Stichprobenmittel*
   - Wir betrachten zunächst die Stichprobenfunktion $\varphi : \mathbb{R}^n \to \mathbb{R}$ mit

$$\varphi(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i\,, \tag{2}$$

     d.h., wir betrachten das *arithmetische Mittel* $\overline{x}_n = (x_1 + \ldots + x_n)/n$ der Stichprobenwerte $x_1, \ldots, x_n$.
   - Die Zahl $\overline{x}_n$ wird *Stichprobenmittel* der (konkreten) Stichprobe $(x_1, \ldots, x_n)$ genannt.

- Beispiel
  - Während eines bestimmten Zeitraumes verkauften die 8 Angestellten der Abteilung „Lebensversicherungen" eines Versicherungsunternehmens jeweils die folgenden Anzahlen von Lebensversicherungsverträgen: 9, 12, 5, 13, 7, 11, 24, 11.
  - Wir fassen diese Daten als Stichprobenwerte $x_1, \ldots, x_8$ einer Stichprobe vom Umfang $n = 8$ auf.
  - Das Stichprobenmittel $\overline{x}_8$ dieser Stichprobe beträgt

$$\overline{x}_8 = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{9+12+5+13+7+11+24+11}{8} = 11.5\,,$$

    d.h., im Mittel wurden von den Angestellten jeweils 11.5 Verträge während des betrachteten Zeitraumes verkauft.
- Beachte
  - In der obenbetrachteten Beispiel-Stichprobe ist die maximale Anzahl der abgeschlossenen Verträge um mehr als 10 Verträge größer als die zweitgrößte Anzahl.
  - Streicht man den Maximalwert 24 aus dieser Stichprobe, so verändert sich das Stichprobenmittel $\overline{x}_8 = 11.5$ zu $\overline{x}_7 = 9.7$.
  - Das Stichprobenmittel reagiert also offensichtlich „empfindlich" auf extreme Werte, sogenannte *Ausreißer*, in den Daten.
- Weitere Eigenschaften des Stichprobenmittels
  - Es gilt stets

$$\sum_{i=1}^{n}(x_i - \overline{x}_n) = 0\,, \tag{3}$$

    d.h., das Stichprobenmittel $\overline{x}_n$ lässt sich als *Schwerpunkt* der Daten $x_1, \ldots, x_n$ interpretieren.
  - Außerdem kann man zeigen, dass das Stichprobenmittel $\overline{x}_n$ die *Summe der quadratischen Abweichungen*

$$e(z; x_1, \ldots, x_n) = \sum_{i=1}^{n}(x_i - z)^2$$

    minimiert, d.h., es gilt

$$e(\overline{x}_n; x_1, \ldots, x_n) = \min_{z\in\mathbb{R}} e(z; x_1, \ldots, x_n)\,.$$

2. *Stichprobenmedian*

- Lagemaßzahlen, die den Einfluss von Extremwerten begrenzen, heißen *resistent* oder *robust*. Eine derartige robuste Lagemaßzahl ist der Stichprobenmedian.
- Hierfür ordnet man die Stichprobenwerte $x_1, \ldots, x_n$ der Größe nach. Dies ergibt die *geordnete Stichprobe* $(x_{(1)}, \ldots, x_{(n)})$ mit $x_{(1)} \le \ldots \le x_{(n)}$.
- Insbesondere gilt

$$x_{(1)} = \min_{1\le i\le n} x_i \qquad \text{und} \qquad x_{(n)} = \max_{1\le i\le n} x_i\,, \tag{4}$$

  d.h., $x_{(1)}$ bzw. $x_{(n)}$ sind das *Minimum* bzw. das *Maximum* der Stichprobenwerte $x_1, \ldots, x_n$, die auch mit $x_{\min}$ bzw. $x_{\max}$ bezeichnet werden.
- In diesem Zusammenhang wird auch die *Stichprobenspannweite* $x_{\max} - x_{\min} = x_{(n)} - x_{(1)}$ betrachtet, die jedoch eine Maßzahl für die Streuung der Daten ist; vgl. Abschnitt 2.2.2.

- Manchmal ist es zweckmäßiger, anstelle des Stichprobenmittels $\overline{x}_n$ den *Stichprobenmedian* x_{med} zu betrachten, wobei

$$x_{\text{med}} = \begin{cases} x_{((n+1)/2)}\,, & \text{falls } n \text{ ungerade,} \\ \big(x_{(n/2)} + x_{((n/2)+1)}\big)/2\,, & \text{falls } n \text{ gerade,} \end{cases} \tag{5}$$

- Beachte
  - Der Stichprobenmedian ist also ebenfalls ein Mittelwert: Jeweils die Hälfte der Stichprobenwerte $x_1, \ldots, x_n$ ist kleiner bzw. größer als der Stichprobenmedian x_{med} .
  - Ein Vorteil des Stichprobenmedians x_{med} besteht darin, dass x_{med} wesentlich weniger als $\overline{x}_n$ von den extremalen Stichprobenwerten $x_{(1)}$ und $x_{(n)}$ abhängt.
  - Für das obenbetrachtete Zahlenbeispiel der Anzahlen von jeweils verkauften Lebensversicherungsverträgen gilt $x_{\text{med}} = 11$, und zwar sowohl für die gesamte Stichprobe aller $n = 8$ Stichprobenwerte als auch für die Teilstichprobe von $n - 1 = 7$ Stichprobenwerten, für die der „Ausreißerwert" 24 gestrichen wurde.
- Eine weitere allgemeine Eigenschaft des Medians x_{med} ist die Minimierung der *Summe der absoluten Abweichungen*

$$e'(z; x_1, \ldots, x_n) = \sum_{i=1}^{n} |x_i - z|$$

minimiert, d.h., es gilt

$$e'(x_{\text{med}}; x_1, \ldots, x_n) = \min_{z \in \mathbb{R}} e'(z; x_1, \ldots, x_n)\,.$$

3. *Empirische Quantile*

- In Verallgemeinerung des Medians x_{med} betrachtet man für jedes $p \in (0, 1)$ der Begriff des p–Quantils z_p der Stichprobe $(x_1, \ldots, x_n)$.

- Dabei geht man erneut von der *geordneten Stichprobe* $(x_{(1)}, \ldots, x_{(n)})$ mit $x_{(1)} \leq \ldots \leq x_{(n)}$ aus und definiert das *p–Quantil* z_p wie folgt:

$$z_p = \begin{cases} x_{([np]+1)}\,, & \text{falls } np \text{ nicht ganzzahlig,} \\ \big(x_{(np)} + x_{(np+1)}\big)/2\,, & \text{falls } np \text{ ganzzahlig,} \end{cases} \tag{6}$$

wobei $[np]$ die größte ganze Zahl bezeichnet, die kleiner oder gleich np ist.

- Beachte
  - Durch das p–Quantil z_p werden die Stichprobenwerte $x_1, \ldots, x_n$ in zwei Teilmengen zerlegt, so dass mindestens $p \cdot 100\%$ der Stichprobenwerte kleiner oder gleich z_p und mindestens $(1-p) \cdot 100\%$ der Stichprobenwerte größer oder gleich z_p sind.
  - Mit anderen Worten: Für das p–Quantil z_p gilt:

$$\frac{\text{Anzahl}\,\{i:\ 1 \leq i \leq n,\ x_i \leq z_p\}}{n} \geq p \qquad \text{und} \qquad \frac{\text{Anzahl}\,\{i:\ 1 \leq i \leq n,\ x_i \geq z_p\}}{n} \geq 1 - p\,.$$

  - Insbesondere ist das 0.5–Quantil $z_{0.5}$ gleich dem Median x_{med} der Stichprobe $(x_1, \ldots, x_n)$.

  - Außerdem ergibt sich aus der Definitionsgleichung (6), dass $x_{(1)} = z_p$, falls $np < 1$, und $x_{(n)} = z_p$, falls $np > n - 1$, d.h., das Minimum $x_{(1)}$ und das Maximum $x_{(n)}$ der Stichprobenwerte $x_1, \ldots, x_n$ können auch als Quantile aufgefasst werden.
  - Weitere wichtige Quantile sind die sogenannten *Quartile* $z_{0.25}$ und $z_{0.75}$.
- Für das obenbetrachtete Zahlenbeispiel der Anzahlen von jeweils verkauften Lebensversicherungsverträgen gilt $n = 8$ und $z_{0.25} = (7 + 9)/2 = 8$ und $z_{0.75} = (12 + 13)/2 = 12.5$.

4. *Modus*

- Derjenige Wert der Stichprobenwerte $x_1, \ldots, x_n$, der am häufigsten auftritt, wird *Modus* der Stichprobe genannt und mit x_{mod} bezeichnet.
- Beachte
  - Der Modus x_{mod} ist die wichtigste Lagemaßzahl für nominalskalierte Merkmale.
  - Für das obenbetrachtete Zahlenbeispiel ist der Modus x_{mod} gleich 11.
- Für intervallskalierte Merkmale können das Stichprobenmittel $\overline{x}_n$, der Median x_{med} und der Modus x_{mod} auch zur Beschreibung der *Symmetrie* bzw. *Schiefe* der Stichprobe benutzt werden. Man spricht von einer
  - *symmetrischen Verteilung* der Stichprobenwerte, falls $\overline{x}_n \approx x_{\text{med}} \approx x_{\text{mod}}$,
  - *linkssteilen Verteilung* der Stichprobenwerte, falls $\overline{x}_n > x_{\text{med}} > x_{\text{mod}}$,
  - *rechtssteilen Verteilung* der Stichprobenwerte, falls $\overline{x}_n < x_{\text{med}} < x_{\text{mod}}$.

5. *Geometrisches und harmonisches Mittel*

- Neben dem eigentlichen (arithmetischen) Stichprobenmittel und dem Median werden in der Literatur noch weitere Ansätze zur Mittelung der Stichprobenwerte $x_1, \ldots, x_n$ betrachtet:
  - Das *geometrische Mittel* der Stichprobenwerte $x_1, \ldots, x_n$ ist gegeben durch
$$\overline{x}_{\text{geo}} = (x_1 \cdot \ldots \cdot x_n)^{1/n}$$
  - und das *harmonische Mittel* von $x_1, \ldots, x_n$ ist gegeben durch
$$\overline{x}_{\text{har}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}} \, .$$
- Beispiele

  i) Das geometrische Mittel $\overline{x}_{\text{geo}}$ wird im Zusammenhang mit Wachstumsfaktoren von Beständen betrachtet (beispielsweise in der Finanz- und Versicherungswirtschaft, aber auch bei biologischen Wachstumsmodellen):
  - Ausgehend von einem Anfangsbestand b_0 sei $b_0, b_1, \ldots, b_n$ eine Folge von Bestandsdaten, die zu $n + 1$ aufeinanderfolgenden (äquidistanten) Zeitpunkten beobachtet wurden.
  - Dabei wird vorausgesetzt, dass $b_i > 0$ für jedes $i = 0, 1, \ldots, n$ gilt.
  - Man nennt $x_i = b_i/b_{i-1}$ den i-ten *Wachstumsfaktor* und $(b_i - b_{i-1})/b_{i-1}$ die i-te *Wachstumsrate* für $i = 1, \ldots, n$.
  - Offenbar gilt
$$b_n = b_0 \, x_1 \cdot \ldots \cdot x_n \qquad \text{bzw.} \qquad b_n = b_0 \, (\overline{x}_{\text{geo}})^n \, ,$$
    d.h., das geometrische Mittel $\overline{x}_{\text{geo}}$ kann man als Mittelung der Wachstumsfaktoren $x_1, \ldots, x_n$ auffassen.

– Außerdem ergibt sich durch Logarithmieren, dass

$$\ln \overline{x}_{\text{geo}} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \qquad \text{und somit} \qquad \ln \overline{x}_{\text{geo}} \leq \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \,,$$

wobei $\overline{x}_{\text{geo}} = \overline{x}_n$ genau dann, wenn $x_1 = \ldots = x_n$.

– Das (arithmetische) Stichprobenmittel $\overline{x}_n$ würde also einen (gegenüber $\overline{x}_{\text{geo}}$) überhöhten mittleren Wachstumsfaktor liefern.

ii) Das harmonische Mittel $\overline{x}_{\text{har}}$ wird beispielsweise bei der Berechnung von mittleren Geschwindigkeiten verwendet.

– Seien $x_1, \ldots, x_n$ die Datenübertragungsraten, mit denen n Nachrichten der Länge ℓ übertragen werden.

– Dann ist $(\ell/x_1) + \ldots + (\ell/x_n)$ die Gesamtdauer, die für die Übertragung der n Nachrichten benötigt wird.

– Die mittlere Datenübertragungsrate ist dann gegeben durch

$$\overline{x}_{\text{har}} = \frac{\ell + \ldots + \ell}{\dfrac{\ell}{x_1} + \ldots + \dfrac{\ell}{x_n}} \,.$$

2.2.2 Maßzahlen für die Streuung der Daten

Wir betrachten nun Kenngrößen der Stichprobe $(x_1, \ldots, x_n)$, die die Streuung der Daten $x_1, \ldots, x_n$ beschreiben.

1. *Stichprobenvarianz und Stichproben-Standardabweichung*

- Zunächst betrachten wir die Stichprobenfunktion $\widetilde{\varphi} : \mathbb{R}^n \to \mathbb{R}$ mit

$$\widetilde{\varphi}(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \big(x_i - \overline{x}_n\big)^2 \,, \tag{7}$$

die die *mittlere quadratische Abweichung* der Stichprobenwerte $x_1, \ldots, x_n$ vom (arithmetischen) Stichprobenmittel $\overline{x}_n$ beschreibt.

- Anstelle der in (7) gegebenen Stichprobenfunktion wird in der beschreibenden Statistik jedoch häufig die (modifizierte) mittlere quadratische Abweichung $\varphi : \mathbb{R}^n \to \mathbb{R}$ mit

$$\varphi(x_1, \ldots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n} \big(x_i - \overline{x}_n\big)^2 \tag{8}$$

betrachtet, die *Stichprobenvarianz* bzw. *empirische Varianz* heißt (und mit s_n^2 bezeichnet wird).

- Manchmal betrachtet man die Wurzel $s_n = \sqrt{s_n^2}$ von s_n^2, die *Stichproben-Standardabweichung* bzw. *empirische Standardabweichung* genannt wird.

- Beachte

– Wegen der Schwerpunkteigenschaft (3) des Stichprobenmittels $\overline{x}_n$, d.h. $\sum_{i=1}^{n}(x_i - \overline{x}_n) = 0$, ist die n-te Abweichung $x_n - \overline{x}_n$ bereits durch die Abweichungen $x_1 - \overline{x}_n, \ldots, x_{n-1} - \overline{x}_n$ der ersten $n-1$ Stichprobenwerte $x_1, \ldots, x_{n-1}$ vom Stichprobenmittel $\overline{x}_n$ eindeutig bestimmt.

– Somit sind nur $n-1$ Abweichungen frei wählbar, weshalb man in der Definitionsgleichung (8) von s^2 nicht durch n, sondern durch die Anzahl $n-1$ der sogenannten *Freiheitsgrade* dividiert.

- Weitere Eigenschaften der Stichprobenvarianz

i) Alternative Darstellungsformel für s^2
   - Aus der Definitionsgleichung (8) von s_n^2 folgt, dass

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-2x_i\overline{x}_n+\overline{x}_n^2) \\ &= \frac{1}{n-1}\Big(\sum_{i=1}^{n}x_i^2-n\overline{x}_n^2\Big)\,. \end{aligned}$$

   - Es gilt also die folgende (alternative) *Darstellungsformel* für die Stichprobenvarianz s^2:

$$s_n^2=\frac{1}{n-1}\Big(\sum_{i=1}^{n}x_i^2-n\overline{x}_n^2\Big)\,. \tag{9}$$

ii) Stichprobenvarianz von linear transformierten Stichproben
   - Für beliebige, jedoch fest vorgegebene Zahlen $\alpha,\beta\in\mathbb{R}$ sei die Stichprobe $(y_1,\ldots,y_n)$ gegeben durch $y_i=\alpha+\beta x_i$ für jedes $i=1,\ldots,n$, d.h., die Stichprobenwerte $y_1,\ldots,y_n$ ergeben sich durch eine *lineare Transformation* der ursprünglichen Stichprobenwerte $x_1,\ldots,x_n$.
   - Mit der Schreibweise $\overline{y}_n=(y_1+\ldots+y_n)/n$ ergibt sich dann, dass

$$\begin{aligned} s_n^2(y_1,\ldots,y_n) &= \frac{1}{n-1}\sum_{i=1}^{n}\big(y_i-\overline{y}_n\big)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\big(\alpha+\beta x_i-(\alpha-\beta\overline{x}_n)\big)^2 \\ &= \beta^2\frac{1}{n-1}\sum_{i=1}^{n}\big(x_i-\overline{x}_n\big)^2=\beta^2 s_n^2(x_1,\ldots,x_n)\,, \end{aligned}$$

   - Es gilt also

$$s_n^2(y_1,\ldots,y_n)=\beta^2 s_n^2(x_1,\ldots,x_n)\,. \tag{10}$$

- Beispiel
   - Gegeben seien die Stückzahlen

$$42,46,41,39,42,40,43,40,41,44,42,42,41,40,42,42,41,43,39,40$$

   eines bestimmten Erzeugnisses, die an $n=20$ aufeinanderfolgenden Tagen von der Produktionsabteilung eines Unternehmens hergestellt worden sind.
   - Für diese Stichprobe gilt offenbar $x_{\min}=39$ bzw. $x_{\max}=46$, woraus sich die Stichprobenspannweite $x_{\max}-x_{\min}=46-39=7$ ergibt.
   - Außerdem gilt $\overline{x}_{20}=41.5$ und $\sum_{i=1}^{20}x_i^2=34500$, so dass sich aus der Darstellungsformel (9) die folgenden Werte für die Stichprobenvarianz s_{20}^2 bzw. die Stichproben-Standardabweichung s_{20} ergeben:

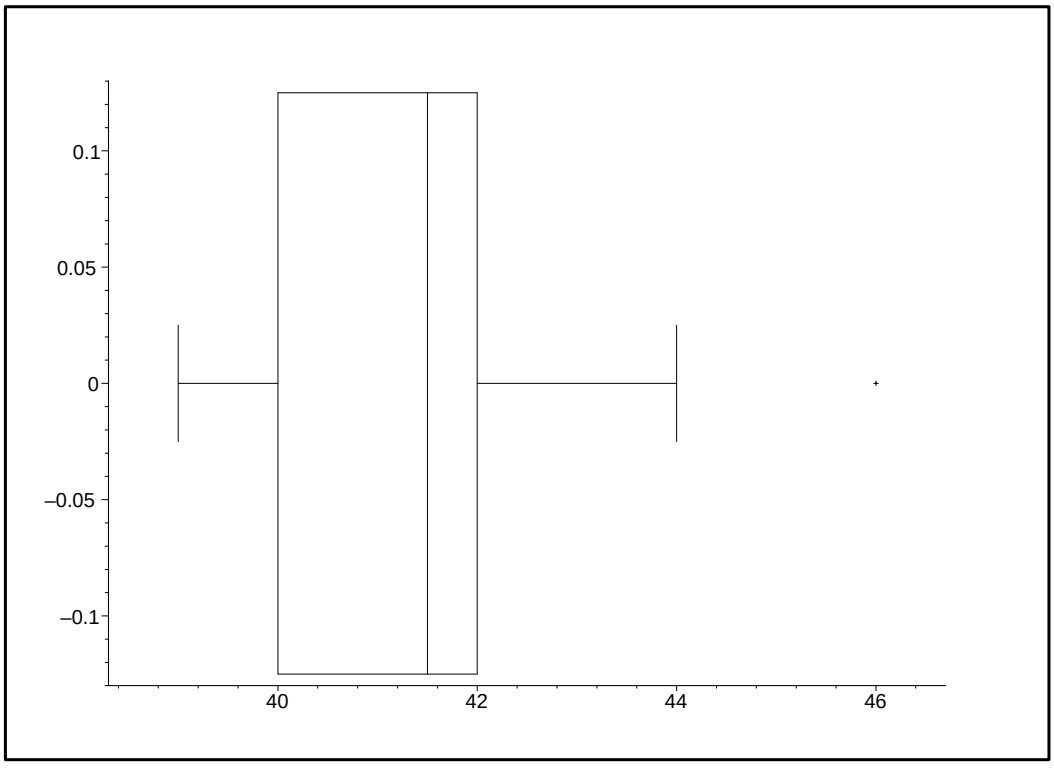
$$s_{20}^2=\frac{1}{19}\left(34500-20\cdot(41.5)^2\right)=2.89\,,\qquad s_{20}=1.70\,.$$

2. *Empirischer Variationskoeffizient*

- Für die Stichprobe $(x_1, \ldots, x_n)$ mit dem Stichprobenmittelwert $\overline{x}_n$ und der Stichproben-Standardabweichung s_n wird der *empirische Variationskoeffizient* durch den Quotienten $s_n/\overline{x}_n$ definiert.
- Der Quotient $s_n/\overline{x}_n$ wird manchmal auch *Variabilitätskoeffizient* genannt.

3. *Empirischer Quantilabstand*

- In Verallgemeinerung der bereits erwähnten Stichprobenspannweite $x_{\min} - x_{\max} = x_{(n)} - x_{(1)}$ wird manchmal auch für beliebiges $p \in (0, 1/2)$ der *empirische Quantilabstand* $z_{1-p} - z_p$ betrachtet, der ebenfalls eine Kenngröße zur Beschreibung der Streuung der Daten $x_1, \ldots, x_n$ ist.
- Insbesondere wird häufig der sogenannte *Interquartilsabstand* $z_{0.75} - z_{0.25}$ betrachtet.
- Beachte
  - Die Quartile $z_{0.25}$ und $z_{0.75}$, das Minimum $x_{\min}$, das Maximum $x_{\max}$ sowie den Median x_{med} nennt man die *Fünf-Maßzahlen-Charakteristik* (bzw. kurz *Fünfer-Charakteristik*) der Stichprobe $(x_1, \ldots, x_n)$.
  - Durch die Fünfer-Charakteristik $(x_{\min}, z_{0.25}, x_{\text{med}}, z_{0.75}, x_{\max})$ lässt sich die Stichprobe $(x_1, \ldots, x_n)$ in vier Teilstichproben zerlegen, wobei diese Teilstichproben jeweils etwa ein Viertel der Stichprobenwerte $x_1, \ldots, x_n$ enthalten.
  - Wenn zusätzlich noch die Quantile $z_{0.05}$ und $z_{0.95}$ betrachtet werden, dann ergibt sich die *Siebener-Charakteristik* $(x_{\min}, z_{0.05}, z_{0.25}, x_{\text{med}}, z_{0.75}, z_{0.95}, x_{\max})$ der Stichprobe $(x_1, \ldots, x_n)$.
  - Es ist üblich, solche Zerlegungen der Stichprobe $(x_1, \ldots, x_n)$ durch einen sogenannten *Box-Plot* graphisch darzustellen.
  - Für das obenbetrachtete Beispiel von Stückzahlen eines bestimmten Ereignisses ergibt sich der folgende Box-Plot der Siebener-Charakteristik, wobei in diesem Fall $x_{\min} = z_{0.05}$ gilt:



2.2.3 Konzentrationsmaße

- In diesem Abschnitt setzen wir voraus, dass
  - das beobachtete Merkmal/Kenngröße/Variable kardinalskaliert ist,
  - sämtliche Stichprobenwerte $x_1, \ldots, x_n$ nichtnegativ sind und
  - die sogenannte *Gesamtmerkmalssumme* $x_1 + \ldots + x_n$ positiv ist, d.h. $\sum_{i=1}^n x_i > 0$.
- Typische Beispiele solcher Merkmale/Kenngrößen/Variablen sind der Umsatz bzw. der Gewinn von Unternehmen, das Einkommen bzw. die Anzahl von Beschäftigten.
- Um die Bezeichnungsweise möglichst einfach zu halten, nehmen wir darüber hinaus (o.B.d.A.) an, dass die Stichprobenwerte $x_1, \ldots, x_n$ der Größe nach geordnet sind, d.h., es gelte

$$(x_1, \ldots, x_n) = (x_{(1)}, \ldots, x_{(n)}) .$$

Wir betrachten nun die folgenden beiden Charakteristiken zur Beschreibung von univariaten Daten, die in der Literatur *Konzentrationsmaße* genannt werden.

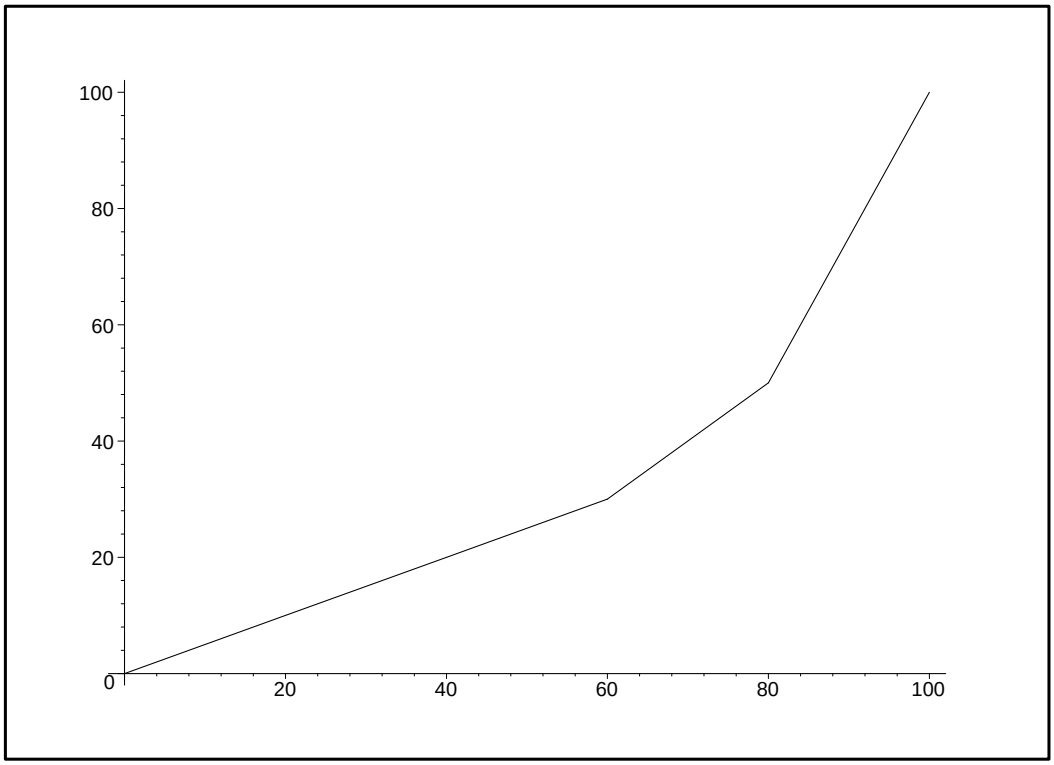
1. *Lorenzkurve*
   - Für jedes $j = 1, \ldots, n$ betrachtet man die sogenannte *relative Merkmalssumme*

$$v_j = \sum_{i=1}^{j} x_i \Big/ \sum_{i=1}^{n} x_i ,$$

     die die j kleinsten Stichprobenwerte $x_1, \ldots, x_j$ auf sich „konzentrieren", welche den Anteil $u_j = j/n$ von sämtlichen Stichprobenwerten ausmachen.
   - Den Polygonzug, der durch die $n+1$ Punkte $(0,0) = (u_0, v_0), (u_1, v_1), \ldots, (u_n, v_n) = (1,1)$ verläuft, nennt man *Lorenzkurve* der (geordneten) Stichprobe $(x_1, \ldots, x_n)$.
   - Beachte
     - Anstelle der (auf die Zahl Eins bezogenen) Anteile u_j und der relativen Merkmalssummen v_j betrachtet man manchmal die *prozentualen Anteile* $100u_j, 100v_j$.
     - Man kann sich leicht überlegen, dass jede Lorenzkurve *monoton wachsend* und *konvex* (d.h. nach unten gewölbt) ist.
     - Die „Stärke der Konzentration" der relativen Merkmalsummen in einem bestimmten Teilbereich der (geordneten) Stichprobe ist proportional zu der vertikalen Abweichung der Lorenzkurve von der (geradlinigen) Diagonalen, die die beiden Punkte $(0,0)$ und $(1,1)$ direkt miteinander verbindet.
   - Beispiel
     - Wir betrachten 5 Unternehmen aus einundderselben Branche, von denen 3 Unternehmen einen Marktanteil von jeweils 10%, ein Unternehmen einen Marktanteil von 20% und ein Unternehmen einen Marktanteil von 50% haben möge.
     - Als (geordnete) Stichprobe ergibt sich dann $(x_1, \ldots, x_5) = (10, 10, 10, 20, 50)$, und als Lorenzkurve ergibt sich der Polygonzug durch die (prozentualen Anteil-) Punkte

$$(0,0), (20,10), (40,20), (60,30), (80,50), (100,100)$$

       mit der graphischen Darstellung



2. *Gini-Koeffizient*

- Weil sich die Stärke der Konzentration durch die Entfernung der Lorenzkurve von der NO–Diagonalen ausdrücken lässt, ist es naheliegend, in die Definition eines weiteren Konzentrationsmaßes die Fläche zwischen der Lorenzkurve und der NO–Diagonalen einzubeziehen.
- Dabei betrachtet man den Quotienten dieses Flächeninhaltes zum „Gesamtflächeninhalt" des rechtwinkligen Dreiecks, das durch die Punkte $(0,0)$, $(1,0)$ und $(1,1)$ gebildet wird, und nennt diesen Quotienten *Gini-Koeffizient* der (geordneten) Stichprobe $(x_1, \ldots, x_n)$.
- Mit anderen Worten: Der Gini-Koeffizient γ ist gegeben durch

$$\begin{aligned} \gamma &= \frac{\text{Flächeninhalt zwischen Diagonale und Lorenzkurve}}{\text{Flächeninhalt zwischen Diagonale und } u\text{-Achse}} \\ &= 2 \cdot \text{ Flächeninhalt zwischen Diagonale und Lorenzkurve}. \end{aligned}$$

- Hieraus ergibt sich durch eine einfache Rechnung, dass

$$\gamma = \frac{2\sum_{i=1}^{n} i p_i - (n+1)}{n}, \tag{11}$$

wobei $p_i = x_i \Big/ \left(\sum_{i=1}^{n} x_i\right)$.

- Beachte
  - Falls sämtliche Stichprobenwerte $x_1, \ldots, x_n$ gleich sind, d.h. bei sogenannter *Nullkonzentration* mit $x_1 = \ldots = x_n$, gilt $\gamma = 0$.
  - Andererseits gilt $\gamma = (n-1)/n$ bei *maximaler Konzentration*, d.h., falls $x_1 = \ldots = x_{n-1} = 0$ und $x_n > 0$.

- Der maximal mögliche Wert des Gini-Koeffizienten γ hängt somit vom Stichprobenumfang ab.
- Aus diesem Grund wird manchmal der *normierte Gini-Koeffizient* γ^* betrachtet, wobei

$$\gamma^* = \frac{n}{n-1}\,\gamma\,.$$

- Beispiel
  - Für das obenbetrachtete Beispiel der Marktanteile der 5 Unternehmen gilt $n = 5$ und

$$p_i = \begin{cases} 0.1 & \text{für } i = 1, 2, 3, \\ 0.2 & \text{für } i = 4, \\ 0.5 & \text{für } i = 5. \end{cases}$$

  - Hieraus folgt, dass

$$\begin{aligned} \gamma &= \frac{1}{5}\left(2(0.1 + 0.2 + 0.3 + 0.8 + 2.5) - 6\right) \\ &= \frac{1}{5}\,1.8 = 0.36 \end{aligned}$$

    bzw. $\gamma^* = 0.45$.

2.2.4 Absolute und relative Häufigkeiten; Histogramm

1. *Absolute und relative Häufigkeiten*

- Die Stichprobenwerte $x_1, \ldots, x_n$ werden manchmal auch als *Urliste* bzw. als *Roh- oder Primärdaten* bezeichnet.
- Weil die direkte Auflistung der Rohdaten $x_1, \ldots, x_n$ mit wachsendem Stichprobenumfang n schnell unübersichtlich wird, ist es manchmal zweckmäßig bzw. notwendig, die Rohdaten in einer anderen Form darzustellen.
- So kann es beispielsweise bei diskreten Merkmalen/Kenngrößen/Variablen sinnvoll sein, anstelle der Rohdaten $x_1, \ldots, x_n$ zunächst
  - die Folge der vorhandenen bzw. potentiell möglichen Ausprägungen/Werte $c_1, \ldots, c_k$ (ohne Berücksichtigung eventuell vorkommender Wiederholungen) zu betrachten und der Größe nach zu ordnen, d.h. $c_1 < c_2 < \ldots < c_k$, und dann
  - für jedes $j = 1, \ldots, k$ die *absolute Häufigkeit* $h_j = h(c_j)$ bzw. die *relative Häufigkeit* $f_j = h_j/n$ der Ausprägung c_j zu bestimmen.

- Beispiel
  - Wir betrachten die Anzahlen $x_1, \ldots, x_{10}$ kariöser Zähne von 10 Schülern, wobei

i	1	2	3	4	5	6	7	8	9	10
x_i	5	1	1	0	5	1	0	2	0	0

  - Dann ergeben sich die folgenden absoluten bzw. relativen Häufigkeiten der Ausprägungen c_j:

c_j	0	1	2	3	4	5
h_j	4	3	1	0	0	2
f_j	0.4	0.3	0.1	0	0	0.2

- Beachte
  - Wenn der Stichprobenumfang n groß ist, dann ist die Menge $\{c_1, \ldots, c_k\}$ der überhaupt vorhandenen bzw. potentiell möglichen Ausprägungen/Werte typischerweise deutlich kleiner (und damit wesentlich übersichtlicher) als die Menge der Rohdaten $\{x_1, \ldots, x_n\}$.
  - Bei stetigen bzw. quasi-stetigen Merkmalen/Kenngrößen/Variablen können ebenfalls absolute bzw. relative Häufigkeiten betrachtet werden, wenn die Rohdaten vorher auf geeignete Weise zu Klassen zusammengefasst und die Häufigkeiten dann für die *gruppierten Daten* bestimmt werden.
  - Dabei ist jedoch zu beachten, dass die Gruppierung/Aggregation von Daten stets mit einem *Informationsverlust* verbunden ist.

2. *Histogramm*

- Wir betrachten nun Merkmale/Kenngrößen/Variablen, die zumindest ordinalskaliert sind, und zerlegen die (reelle) Zahlengerade $\mathbb{R}$ in k Intervalle $[a_j, b_j)$, die sich unmittelbar aneinander anschließen, d.h.

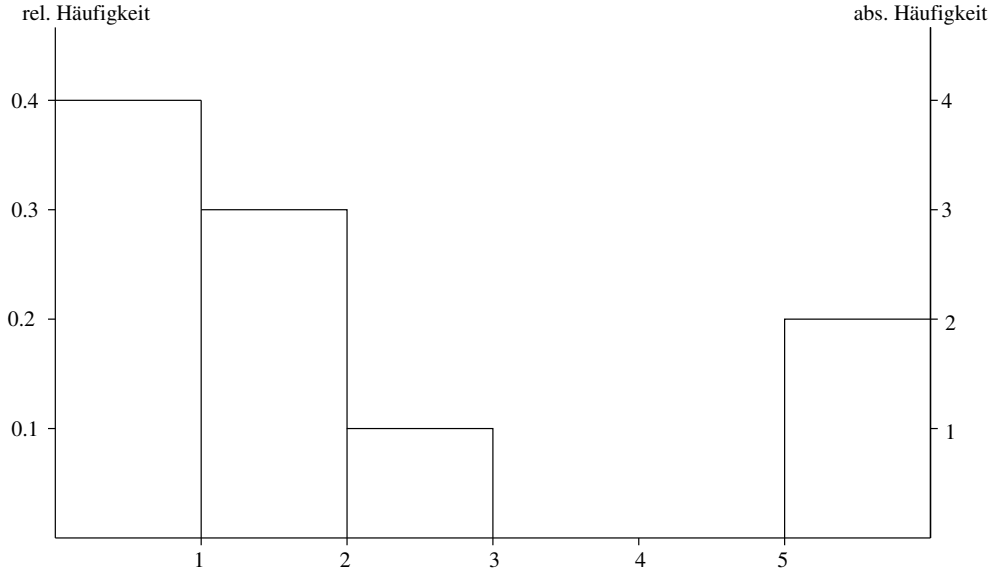
$$-\infty = a_1 < b_1 = a_2 < b_2 = \ldots = a_k < b_k = +\infty\,.$$

- Für jedes $j = 1, \ldots, k$ betrachten wir die absolute Häufigkeit h_j bzw. die relative Häufigkeit $f_j = h_j/n$ derjenigen Stichprobenwerte $x_1, \ldots, x_n$, die in das Intervall $[a_j, b_j)$ fallen.
- Ein *Histogramm* ist ein Säulendiagramm, wobei den *Klassen* $[a_1, b_1), \ldots, [a_k, b_k)$ Säulen zugeordnet werden, deren Flächeninhalte jeweils gleich oder proportional zu den absoluten bzw. relativen Häufigkeiten $h_1, \ldots, h_k$ bzw. $f_1, \ldots, f_k$ sind.

- Beispiel
  - Für den obenbetrachteten Datensatz der Anzahlen kariöser Zähne bei einer Gruppe von 10 Schülern besteht keine Notwendigkeit, die 6 beobachteten Werte $0, 1, 2, 3, 4, 5$ zu einer kleineren Anzahl von Klassen zusammenzufassen.
  - Um ein Säulendiagramm zu erhalten, werden dennoch die „Intervall-Klassen"

$$[-\infty, 1), [1, 2), \ldots, [4, 5), [5, , \infty]$$

    betrachtet, wobei diese Zerlegung der Zahlengerade $\mathbb{R}$ zu den (bereits obenerwähnten) absoluten Häufigkeiten $4, 3, 1, 0, 0, 2$ führt.
  - Hieraus ergibt sich das folgende Histogramm:



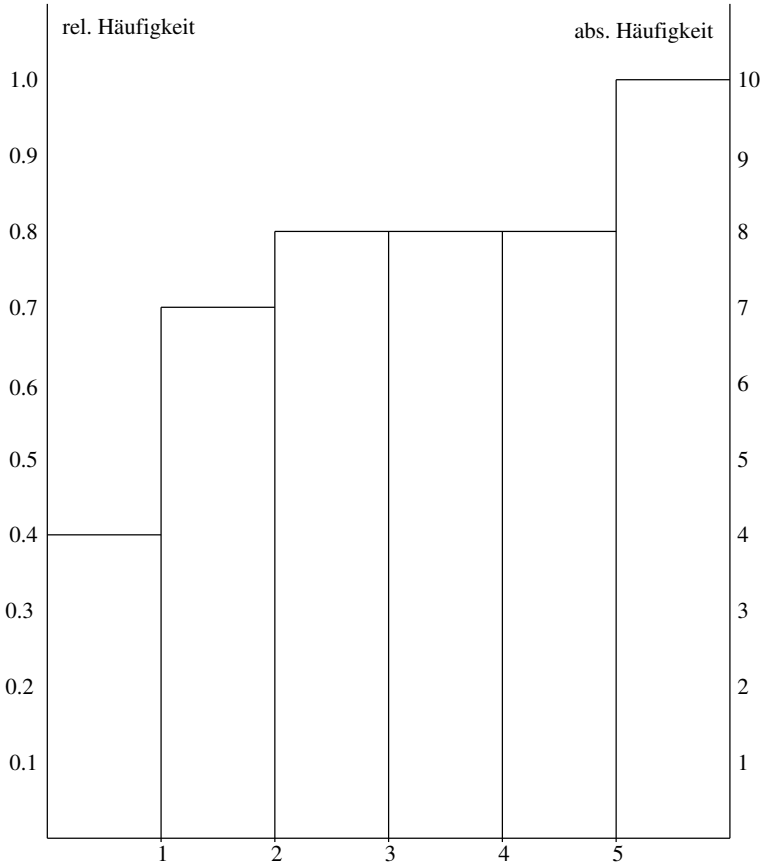
2.2.5 Kumulierte Häufigkeiten; empirische Verteilungsfunktion

- Anstelle der absoluten Häufigkeiten h_j bzw. der relativen Häufigkeiten $f_j = h_j/n$ werden manchmal die *kumulierten Häufigkeiten*

$$H_j = \sum_{i=1}^{j} h_i \qquad \text{bzw.} \qquad F_j = \sum_{i=1}^{j} f_i\,,$$

  derjenigen Stichprobenwerte $x_1, \ldots, x_n$ betrachtet, die in die ersten j Intervalle $[a_1, b_1), \ldots, [a_j, b_j)$ fallen.

- Beachte
  - Die kumlierten Häufigkeiten $H_1, H_2, \ldots, H_k$ bzw. $F_1, F_2, \ldots, F_k$ bilden *monoton wachsende* Zahlenfolgen.
  - Für den obenbetrachteten Datensatz der Anzahlen kariöser Zähne ergibt sich das folgende Histogramm (der kumulierten Häufigkeiten):



- Meistens werden die kumlierten Häufigkeiten $H_1, H_2, \ldots, H_k$ bzw. $F_1, F_2, \ldots, F_k$ als monoton wachsende *Treppenfunktionen* $H : \mathbb{R} \to [0, n]$ bzw. $F : \mathbb{R} \to [0, 1]$ dargestellt.

- Die Darstellung der relativen kumulierten Häufigkeiten $F_1, F_2, \ldots, F_k$ durch eine Treppenfunktion führt dann zu der sogenannten *empirischen Verteilungsfunktion* $F : \mathbb{R} \to [0, 1]$ mit

$$F(x) = F_j \quad \Big(= f_1 + \ldots + f_j\Big)\,, \qquad \text{falls } a_j \le x < b_j.$$

- Beachte

- Der Funktionswert $F(x)$ der empirischen Verteilungsfunktion $F : \mathbb{R} \to [0,1]$ an der Stelle $x \in [a_j, b_j)$ ist der Anteil derjenigen Stichprobenwerte $x_1, \ldots, x_n$, die in die ersten j Klassen $[a_1, b_1), \ldots, [a_j, b_j)$ fallen.
- Das heißt insbesondere, dass die empirische Verteilungsfunktion eine *monoton wachsende* Funktion ist.
- Für jedes $j = 1, \ldots, k-1$ springt die empirische Verteilungsfunktion F an der Intervallgrenze b_j $(= a_{j+1})$ um die relative Häufigkeit f_{j+1} nach oben.
- Dabei ist an den Sprungstellen der obere Wert der zugehörige Funktionswert, d.h., die empirische Verteilungsfunktion ist eine *rechtsseitig stetige* Funktion.
- Falls die Intervall-Klassen $[a_1, b_1), \ldots, [a_k, b_k)$ so gewählt sind, dass $b_1 \leq \min\{x_1, \ldots, x_n\}$ und $a_k > \max\{x_1, \ldots, x_n\}$, dann gilt außerdem $F(x) = 0$ für jedes $x < b_1$ bzw. $F(x) = 1$ für jedes $x \geq a_k$.

2.3 Kenngrößen zur Beschreibung von bivariaten Daten

- In diesem Abschnitt betrachten wir gleichzeitig *zwei* Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ von zwei diskreten Merkmalen/Kenngrößen/Variablen, die wir mit X bzw. Y bezeichnen.
- Neben Stichproben von diskreten Merkmalen/Kenngrößen/Variablen werden wir auch (gruppierte/ aggregierte) Stichproben von stetigen sowie quasi-stetigen Merkmalen/Kenngrößen/Variablen betrachten, die durch Einteilung der Daten in endlich viele Klassen „diskretisiert" worden sind.
- Dabei nehmen wir an, dass beide Stichproben den gleichen Stichprobenumfang n haben.
- Die Ausprägungen/Werte von X bezeichnen wir (so wie bisher) mit $c_1, \ldots, c_{k_1}$, und die Ausprägungen/Werte von Y bezeichnen wir mit $d_1, \ldots, d_{k_2}$.
- Von besonderem Interesse ist die Untersuchung der Frage, ob ein *Zusammenhang* (also eine Kontingenz) zwischen den Ausprägungen/Werten von X bzw. Y besteht.

2.3.1 Kontingenztafel der absoluten Häufigkeiten

- Analog zur Notation, die in Abschnitt 2.2.4 für den univariaten Fall eingeführt wurde, bezeichnen wir mit $h_{ij} = h(c_i, d_j)$ für jedes $i = 1, \ldots, k_1$ und für jedes $j = 1, \ldots, k_2$ die *absolute Häufigkeit*, mit der die Kombination (c_i, d_j) der Ausprägungen c_i und d_j in den Paaren (x_r, y_r) der Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ auftritt.
- Bei der tabellarischen Darstellung der absoluten Häufigkeiten $h_{ij} = h(c_i, d_j)$ werden auch die sogenannten *Randhäufigkeiten*

$$h_{i\cdot} = h_{i1} + \ldots + h_{ik_2} \qquad \forall\, i = 1, \ldots, k_1 \tag{12}$$

bzw.

$$h_{\cdot j} = h_{1j} + \ldots + h_{k_1 j} \qquad \forall\, j = 1, \ldots, k_2 \tag{13}$$

der Ausprägungen/Werte von X bzw. Y betrachtet, die sich ergeben, wenn lediglich die Ausprägungen/Werte von X (ohne Berücksichtigung der Ausprägungen/Werte von Y) bzw. umgekehrt lediglich die Ausprägungen/Werte von Y (ohne Berücksichtigung der Ausprägungen/Werte von X) betrachtet werden.
- Die Punktnotation $h_{i\cdot}$ bzw. $h_{\cdot j}$ in (12) bzw. (13) macht dabei deutlich, ob über j bzw. i summiert wird.
- Für die *Gesamtsumme* n sämtlicher Häufigkeiten wird manchmal die Notation $h_{\cdot\cdot}$ verwendet, wobei

$$h_{\cdot\cdot} = \sum_{i=1}^{k_1} h_{i\cdot} = \sum_{j=1}^{k_2} h_{\cdot j} \qquad \Big(= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} h_{ij} = n\Big).$$

- Unter der $k_1 \times k_2$–*Kontingenztafel* der absoluten Häufigkeiten von X und Y versteht man die folgende Tabelle:

	d_1	$\ldots$	d_{k_2}	
c_1	h_{11}	$\ldots$	h_{1k_2}	$h_{1\cdot}$
c_2	h_{21}	$\ldots$	h_{2k_2}	$h_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
c_{k_1}	$h_{k_1 1}$	$\ldots$	$h_{k_1 k_2}$	$h_{k_1\cdot}$
	$h_{\cdot 1}$	$\ldots$	$h_{\cdot k_2}$	$h_{\cdot\cdot} (= n)$

- Beispiel (vgl. L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz (2000) *Statistik*. Springer, Berlin, S. 111 ff.)
  - Bei einer Befragung von 447 männlichen deutschen Arbeitslosen, die vom Deutschen Institut für Wirtschaftsforschung (DIW) durchgeführt wurde, wurden u.a. die folgenden beiden Merkmale erfasst:
    1. *Ausbildungsniveau* mit den vier Ausprägungen „keine Ausbildung" (K), „Lehre" (L), „fachspezifische Ausbildung" (F), „Hochschulabschluss" (H) sowie
    2. *Dauer der Arbeitslosigkeit* mit den Kategorien „Kurzzeitarbeitslosigkeit" ($\leq$ 6 Monate), „mittelfristige Arbeitslosigkeit" (7–12 Monate), „Langzeitarbeitslosigkeit" ($>$ 12 Monate)
  - Dabei ergaben sich die folgenden Häufigkeiten:

	$\leq$ 6 Monate	7–12 Monate	$>$ 12 Monate	
keine Ausbildung	86	19	18	123
Lehre	170	43	20	233
fachspez. Ausbildung	40	11	5	56
Hochschulabschluss	28	4	3	35
	324	77	46	447

2.3.2 Kontingenztafeln für relative bzw. bedingte Häufigkeiten

- Wenn anstelle der absoluten Häufigkeiten $h_{ij} = h(c_i, d_j)$ die *relativen Häufigkeiten* $f_{ij} = h(c_i, d_j)/n$ betrachtet werden, dann ergibt sich die folgende $k_1 \times k_2$–*Kontingenztafel* der relativen Häufigkeiten der Kombinationen der Ausprägungen/Werte c_i, d_j) von X bzw. Y:

	d_1	$\ldots$	d_{k_2}	
c_1	f_{11}	$\ldots$	f_{1k_2}	$f_{1\cdot}$
c_2	f_{21}	$\ldots$	f_{2k_2}	$f_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
c_{k_1}	$f_{k_1 1}$	$\ldots$	$f_{k_1 k_2}$	$f_{k_1\cdot}$
	$f_{\cdot 1}$	$\ldots$	$f_{\cdot k_2}$	1

- Dabei sind

$$f_{i\cdot} = f_{i1} + \ldots + f_{ik_2} \qquad \forall\, i = 1, \ldots, k_1$$

bzw.

$$f_{\cdot j} = f_{1j} + \ldots + f_{k_1 j} \qquad \forall\, j = 1, \ldots, k_2$$

die relativen *Randhäufigkeiten* der Ausprägungen/Werte von X bzw. Y.

- Bei der Untersuchung der Frage, ob *Zusammenhänge* zwischen den Ausprägungen/Werten von X bzw. Y bestehen, interessieren darüber hinaus noch die *bedingten relativen Häufigkeiten*

$$f_Y(j \mid i) = \frac{h_{ij}}{h_{i\cdot}} \qquad \text{bzw.} \qquad f_X(i \mid j) = \frac{h_{ij}}{h_{\cdot j}} \qquad \forall\, i = 1, \ldots, k_1,\ j = 1, \ldots, k_2\,, \tag{14}$$

wobei $\frac{0}{0} = 0$ gesetzt wird.

- Dabei sind $f_Y(1 \mid i), \ldots, f_Y(k_2 \mid i)$ die relativen Häufigkeiten derjenigen Ausprägungen/Werte von Y, die zusammen mit der (fest vorgegebenen) Ausprägung c_i von X auftreten.
- Man spricht deshalb auch von den bedingten relativen Häufigkeiten $f_Y(1 \mid i), \ldots, f_Y(k_2 \mid i)$ der Ausprägungen/Werte von Y unter der Bedingung, dass $X = c_i$.
- Umgekehrt heißen $f_X(1 \mid j), \ldots, f_X(k_1 \mid j)$ die bedingten relativen Häufigkeiten der Ausprägungen/Werte von X unter der Bedingung, dass $Y = d_j$.
- Für das in Abschnitt 2.3.1 betrachtete Beispiel, bei dem die Merkmale „Ausbildungsniveau" und „Dauer der Arbeitslosigkeit" betrachtet wurden, ergibt sich dann die folgende 4×3–*Kontingenztafel* der bedingten relativen Häufigkeiten:

	≤ 6 Monate	7–12 Monate	> 12 Monate	
keine Ausbildung	0.699	0.154	0.147	1
Lehre	0.730	0.184	0.086	1
fachspez. Ausbildung	0.714	0.197	0.089	1
Hochschulabschluss	0.800	0.114	0.086	1

2.3.3 Zusammenhangsmaße

1. *Bedingte und relative Chancen*

   - Wir betrachten zunächst den Spezialfall $k_1 = k_2 = 2$, d.h., die Merkmale/Kenngrößen/Variablen X und Y besitzen jeweils nur zwei verschiedene Ausprägungen/Werte.
   - Die entsprechende 2×2–Kontingenztafel hat somit die Form

	d_1	d_2	
c_1	h_{11}	h_{12}	$h_{1\cdot}$
c_2	h_{21}	h_{22}	$h_{2\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	n

   - Für jedes $i = 1, 2$ heißt der Quotient der bedingten relativen Häufigkeiten

$$\gamma(d_1, d_2 \mid c_i) = \frac{f_Y(1 \mid i)}{f_Y(2 \mid i)} = \frac{h_{i1}/h_{i\cdot}}{h_{i2}/h_{i\cdot}} = \frac{h_{i1}}{h_{i2}} \tag{15}$$

   die *bedingte Chance* für $X = c_i$, wobei $h_{i2} > 0$ vorausgesetzt wird.

- Hieraus ergibt sich ein einfaches Zusammenhangsmaß zwischen den Chancen der ersten bzw. zweiten Zeile der 2×2–Kontingenztafel, das *relative Chance* genannt wird und gegeben ist durch den Quotienten

$$\gamma(d_1, d_2 \mid c_1, c_2) = \frac{\gamma(d_1, d_2 \mid c_1)}{\gamma(d_1, d_2 \mid c_2)} = \frac{h_{11}h_{22}}{h_{12}h_{21}}\,, \tag{16}$$

  wobei $h_{12}, h_{21} > 0$ vorausgesetzt wird.

- Beispiel
  - Für das in Abschnitt 2.3.1 diskutierte Beispiel betrachten wir jetzt nur die Ausprägungen „fachspezifische Ausbildung" (F) bzw. „Hochschulabschluss" (H) für das Merkmal „Ausbildungsniveau" sowie die Ausprägungen „Kurzzeitarbeitslosigkeit" ($\leq$ 6 Monate) bzw. „mittel- und langfristige Arbeitslosigkeit" ($\geq$ 7 Monate) für das Merkmal „Dauer der Arbeitslosigkeit".
  - Dann ergibt sich die folgende 2×2–*Kontingenztafel* der absoluten Häufigkeiten:

	d_1	d_2
c_1	40	16
c_2	28	7

  - Die „bedingte Chance" $\gamma(d_1, d_2 \mid c_1)$ von Personen mit fachspezifischer Ausbildung, kurzfristig arbeitslos zu sein (gegenüber einer mittel- bzw. langfristigen Arbeitslosigkeit), ist also gegeben durch

$$\gamma(d_1, d_2 \mid c_1) = \frac{40}{16} = 2.5\,.$$

  - Für Personen mit Hochschulabschluss ergibt sich dagegen der Wert

$$\gamma(d_1, d_2 \mid c_2) = \frac{28}{7} = 4.0\,.$$

  - Für die „relative Chance" ergibt sich

$$\gamma(d_1, d_2 \mid c_1, c_2) = \frac{280}{448} = \frac{70}{112} = 0.625\,,$$

    d.h., für Personen mit Hochschulabschluss stehen somit die „Chancen" deutlich besser.

- Beachte
  - Die Begriffe der bedingten bzw. relativen Chance lassen sich völlig analog auch für den Fall definieren, dass Merkmale mit mehr als 2 Ausprägungen betrachtet werden.
  - Die relative Chance zwischen $X = c_i$ und $X = c_{i'}$ bezüglich der bedingten Chancen von $Y = d_j$ und $Y = d_{j'}$ ist dann gegeben durch

$$\gamma(d_j, d_{j'} \mid c_i, c_{i'}) = \frac{\gamma(d_j, d_{j'} \mid c_i)}{\gamma(d_j, d_{j'} \mid c_{i'})} = \frac{h_{ij}\, h_{i'j'}}{h_{ij'}\, h_{i'j}}\,.$$

2. χ^2-*Koeffizient*

- Wir definieren nun den χ^2-Koeffizienten der beiden Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$,
  - der eine weitere Maßzahl zur Beschreibung des (eventuell vorhandenen) Zusammenhanges zwischen den Werten der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ der beiden Markmale X und Y ist,
  - wobei wir das Nichtvorhandensein eines solchen Zusammenhanges mit Hilfe der bedingten relativen Häufigkeiten $f_Y(j \mid i)$ beschreiben, die in (14) eingeführt worden sind.
- Man erwartet, dass die bedingten relativen Häufigkeiten $\{f_Y(j \mid i),\, j = 1, \ldots, k_2\}$ in diesem Fall

  - *nicht* von i abhängen,
  - was gleichbedeutend damit ist, dass

$$f_Y(j \mid i) = \frac{h_{\cdot j}}{n} \qquad \forall\, i = 1, \ldots, k_1,\ j = 1, \ldots, k_2\,.$$

- Mit anderen Worten: Falls die Ausprägungen/Werte der Merkmale/Kenngrößen/Variablen X und Y keinen Zusammmenhang aufweisen (d.h. unabhängig sind), dann sollte die (in diesem Fall) *erwartete Häufigkeit* $\widetilde{h}_{ij} = \widetilde{h}(c_i, d_j)$, mit der die Kombination (c_i, d_j) der Ausprägungen c_i und d_j auftritt, für jedes $i = 1, \ldots, k_1$ unf für jedes $j = 1, \ldots, k_2$
  - der folgenden Gleichung genügen:

$$\frac{\widetilde{h}_{ij}}{h_{i\cdot}} = \frac{h_{\cdot j}}{n}\,,$$

  - d.h., gegeben sein durch den *Produkt-Ansatz*

$$\widetilde{h}_{ij} = \frac{h_{i\cdot}\, h_{\cdot j}}{n}\,.$$

- Beachte
  - Falls die Ausprägungen/Werte der Merkmale/Kenngrößen/Variablen X und Y keinen Zusammenhang aufweisen, dann sollten sich die (tatsächlich beobachteten) Häufigkeiten h_{ij} und die (zu erwartenden) Häufigkeiten $\widetilde{h}_{ij}$ nicht zu sehr voneinander unterscheiden.
  - Als Zusammenhangsmaß betrachtet man deshalb den χ^2-*Koeffizienten* T, der eine sogenannte *Testgröße* ist und gegeben ist durch

$$T = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{\left(h_{ij} - \dfrac{h_{i\cdot}\, h_{\cdot j}}{n}\right)^2}{\dfrac{h_{i\cdot}\, h_{\cdot j}}{n}}\,, \tag{17}$$

    wobei vorausgesetzt wird, dass sämtliche Randhäufigkeiten $h_{1\cdot}, \ldots, h_{k_1\cdot}$ sowie $h_{\cdot 1}, \ldots, h_{\cdot k_2}$ positiv sind, und die Division durch $\widetilde{h}_{ij} = (h_{i\cdot}\, h_{\cdot j})/n$ lediglich der Normierung dient.
  - Im Spezialfall einer 2×2–Kontingenztafel

	d_1	d_2	
c_1	h_{11}	h_{12}	$h_{11} + h_{12}$
c_2	h_{21}	h_{22}	$h_{21} + h_{22}$
	$h_{11} + h_{21}$	$h_{12} + h_{22}$	n

    lässt sich der in (17) definierte χ^2-Koeffizient T leicht berechnen, denn in diesem Fall gilt

$$T = \frac{n\,(h_{11}\, h_{22} - h_{12}\, h_{21})^2}{(h_{11} + h_{12})(h_{11} + h_{21})(h_{12} + h_{22})(h_{21} + h_{22})}\,, \tag{18}$$

    wobei vorausgesetzt wird, dass die Randhäufigkeiten $h_{11} + h_{12}$, $h_{11} + h_{21}$, $h_{12} + h_{22}$ und $h_{21} + h_{22}$ positiv sind.

- Für den in (17) definierten χ^2-Koeffizienten gilt stets $T \geq 0$, wobei
  - T groß ist, wenn ein Zusammenhang zwischen X und Y besteht,
  - T klein ist, wenn X und Y voneinander unabhängig sind.

- Um genauer sagen zu können, wann T als klein bzw. groß anzusehen ist, sind tieferliegende mathematische Modelle der *beurteilenden Statistik* erforderlich, insbesondere sogenannte *Signifikanztests* zum Überprüfen von Modellannahmen; vgl. beispielsweise das Kapitel 3 des jetzigen Vorlesungsskriptes.

3. *Kontingenzkoeffizient*

- Der in (17) definierte χ^2-Koeffizient T hat den Nachteil, dass der Wertebereich von T vom Umfang n der Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ abhängt.
- Dieser Nachteil wird eliminiert, wenn anstelle des χ^2-Koeffizienten T der *Kontingenzkoeffizient* T' betrachtet wird,
  - der gegeben ist durch

$$T' = \sqrt{\frac{T}{n+T}}\,, \tag{19}$$

  - wobei T' nur Werte zwischen 0 und $T'_{\max} = \sqrt{(k_{\min}-1)/k_{\min}}$ annehmen kann; $k_{\min} = \min\{k_1, k_2\}$.

4. *Korrigierter Kontingenzkoeffizient*

- Ein gewisser Nachteil des Kontingenzkoeffizienten T' besteht noch darin, dass der Wertebereich der Testgröße T' von den Anzahlen k_1, k_2 der Ausprägungen von X bzw. Y abhängt.
- Durch einen weiteren Normierungsschritt wird deshalb der sogenannte *korrigierte Kontingenzkoeffizient* T^* eingeführt, der gegeben ist durch

$$T^* = \frac{T'}{T'_{\max}}$$

  und der nur Werte im Einheitsintervall $[0, 1]$ annehmen kann.
- Beispiel
  - Für das in Abschnitt 2.3.1 eingeführte Beispiel betrachten wir nun die Ausprägungen „keine Ausbildung" bzw. „Lehre" für das Merkmal „Ausbildungsniveau" sowie die Ausprägungen „mittelfristige Arbeitslosigkeit" (7–12 Monate) bzw. „langfristige Arbeitslosigkeit (> 12 Monate) für das Merkmal „Dauer der Arbeitslosigkeit".
  - Dann ergibt sich die folgende 2×2–*Kontingenztafel* der absoluten Häufigkeiten:

	d_1	d_2	
c_1	19	18	37
c_2	43	20	63
	62	38	100

(20)

  - Hieraus und aus (18) ergibt sich, dass

$$T = \frac{100\,(19 \cdot 20 - 18 \cdot 43)^2}{37 \cdot 63 \cdot 62 \cdot 38} = 2.826 \tag{21}$$

  sowie $T' = 0.165$ bzw. $T^* = 0.234$.

2.4 Beschreibung von metrischen bivariaten Daten

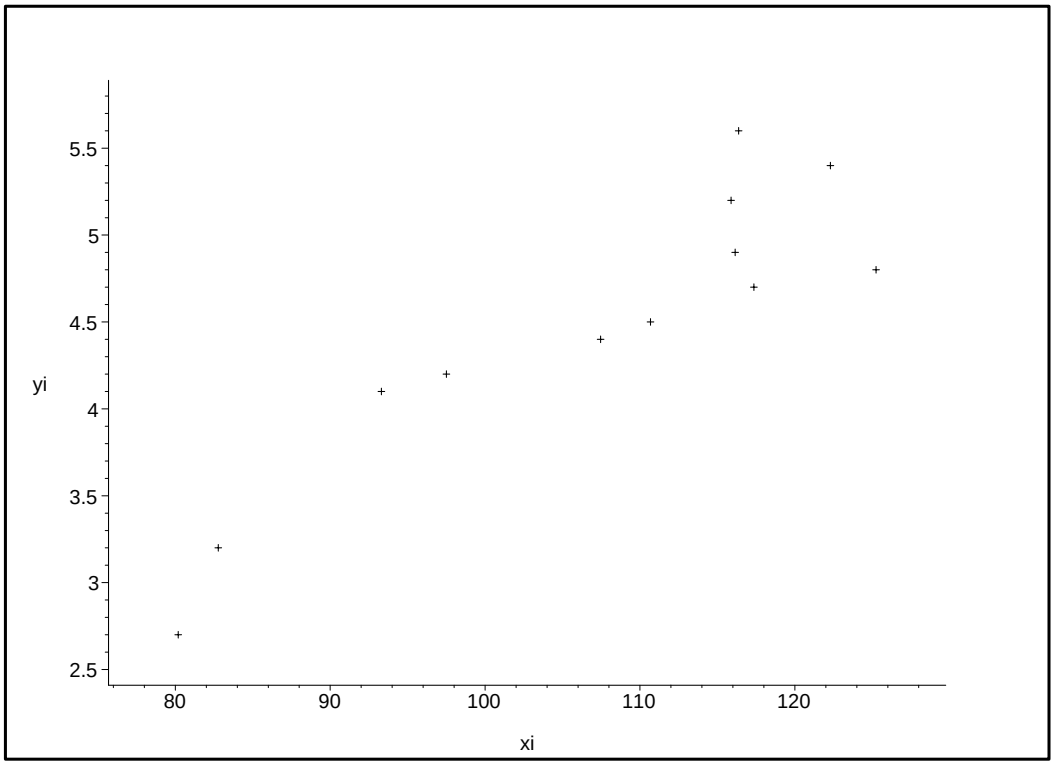
- Wir betrachten nun den Fall, dass
  - der Umfang n der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ groß ist und dass sie
  - die Ausprägungen/Werte von metrisch skalierten Merkmalen/Kenngrößen/Variablen X bzw. Y enthalten.
- Dann ist es sinnvoll, neben den in Abschnitt 2.3 diskutierten Methoden, noch weitere Verfahren zur Darstellung und Beschreibung des bivariaten Datensatzes $(x_1, y_1), \ldots, (x_n, y_n)$ heranzuziehen.

2.4.1 Streudiagramm (Scatterplot)

- Eine einfache graphische Darstellung des bivariaten Datensatzes $(x_1, y_1), \ldots, (x_n, y_n)$ ist durch ein sogenanntes *Streudiagramm* gegeben, wobei
  - die Daten $(x_1, y_1), \ldots, (x_n, y_n)$ als Punkte in der euklidischen Ebene $\mathbb{R}^2$ aufgefasst und in ein (orthogonales) kartesisches Koordinatensystem eingezeichnet werden, so dass
  - die x-Werte auf der Abszissenachse und die y-Werte auf der Ordinatenachse aufgetragen werden.
- Beispiel (vgl. Casella/Berger (2002) *Statistical Inference*, Duxbury, S. 540 ff.)
  - Im Weinanbau werden die jeweils im Herbst geernteten Erträge in Tonnen je 100 m^2 (t/ar) gemessen.
  - Es ist bekannt, dass der Jahresertrag bereits im Juli ziemlich gut prognostiziert werden kann, und zwar durch die Bestimmung der mittleren Anzahl von Beeren, die je Traube gebildet worden sind.
  - Dabei fassen wir den Jahresertrag als Zielvariable (Y) auf, und die mittlere Clusterzahl je Traube (X) als Ausgangsvariable.
  - Beobachtet wurden die folgenden Ausprägungen/Werte:

Jahr	Ertrag (y_i)	Clusterzahl (x_i)
1971	5.6	116.37
1973	3.2	82.77
1974	4.5	110.68
1975	4.2	97.50
1976	5.2	115.88
1977	2.7	80.19
1978	4.8	125.24
1979	4.9	116.15
1980	4.7	117.36
1981	4.1	93.31
1982	4.4	107.46
1983	5.4	122.30

  Die Daten des Jahres 1972 fehlen, weil in diesem Jahr das untersuchte Weinanbaugebiet von einem Wirbelsturm verwüstet worden war.
  - Für diesen Datensatz ergibt sich das folgende Streudiagramm:



2.4.2 Empirische Kovarianz; empirischer Korrelationskoeffizient

1. *Empirische Kovarianz*

   - Aus dem Streudiagramm des Beispiels, das in Abschnitt 2.4.1 betrachtet wurde, ergibt sich die Vermutung, dass
     - ein Zusammenhang zwischen den Merkmalen „Clusterzahl je Traube" (X) und „Jahresertrag"(Y) besteht, denn
     - für wachsende Werte des Merkmals X weist auch das Merkmal Y tendenzmäßig größere Werte auf.

   - Eine Maßzahl zur Beschreibung eines solchen Zusammenhanges ist die *empirische Kovarianz*

$$s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n) \tag{22}$$

   der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$, wobei

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

   die Stichprobenmittel von $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ bezeichnen.

   - Beachte

- Ein Nachteil des in (22) definierten Zuhammenhangsmaßes besteht darin, dass s_{xy}^2 skalenabhängig ist, d.h., von der Größe der Stichprobenwerte $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$ abhängt.
- Dieser Nachteil wird eliminiert, wenn anstelle der empirischen Kovarianz s_{xy}^2 der *empirische Korrelationskoeffizient* betrachtet wird.

2. *Empirischer Korrelationskoeffizient*

- Die Größe

$$\rho_{xy} = \frac{\sum_{i=1}^{n}\Big((x_i - \overline{x}_n)(y_i - \overline{y}_n)\Big)}{\sqrt{\sum_{i=1}^{n}\big(x_i - \overline{x}_n\big)^2 \sum_{i=1}^{n}\big(y_i - \overline{y}_n\big)^2}} \qquad \left(= \frac{s_{xy}^2}{\sqrt{s_{xx}^2 s_{yy}^2}}\right) \tag{23}$$

heißt *empirischer Korrelationskoeffizient* der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$, wobei die Stichprobenvarianzen s_{xx}^2 und s_{yy}^2 gegeben sind durch

$$s_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(x_i - \overline{x}_n\big)^2, \qquad s_{yy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(y_i - \overline{y}_n\big)^2.$$

- Man kann zeigen, dass für den in (23) definierten empirischen Korrelationskoeffizienten ρ_{xy} stets

$$-1 \le \rho_{xy} \le 1 \tag{24}$$

gilt, wobei
  - $|\rho_{xy}|$ groß ist, wenn ein Zusammenhang zwischen X und Y besteht, und
  - $|\rho_{xy}|$ klein ist, wenn X und Y voneinander unabhängig sind.

- Insbesondere kann man zeigen, dass
  - $\rho_{xy} = 1$, falls sämtliche Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ auf einer Geraden mit positivem Anstieg liegen,

  bzw.
  - $\rho_{xy} = -1$, falls sämtliche Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ auf einer Geraden mit negativem Anstieg liegen.

- Der empirische Korrelationskoeffizient ρ_{xy} misst darüber hinaus in dem folgenden Sinne die *Stärke des linearen Zusammenhanges* zwischen den Ausprägungen/Werten der Merkmale X und Y:
  - Je näher die Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ an einer Geraden mit positivem Anstieg liegen, um so näher liegt der empirische Korrelationskoeffizient ρ_{xy} bei 1, und
  - je näher die Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ an einer Geraden mit negativem Anstieg liegen, um so näher liegt der empirische Korrelationskoeffizient ρ_{xy} bei -1.

- Eine (grobe) Klassifikation des Zusammenhanges der Merkmale X und Y kann somit wie folgt beschrieben werden:
  - „schwacher Zusammenhang", falls $|\rho_{xy}| < 0.5$,
  - „mittlerer Zusammenhang", falls $0.5 \le |\rho_{xy}| < 0.8$,
  - „starker Zusammenhang", falls $|\rho_{xy}| \ge 0.8$.

3. *Alternative Darstellung des empirischen Korrelationskoeffizienten ρ_{xy}*

- Man kann zeigen, dass sich der in (23) definierte empirische Korrelationskoeffizient ρ_{xy} darstellen lässt in der Form

$$\rho_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}_n \overline{y}_n}{\sqrt{\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}_n^2\Big)\Big(\sum_{i=1}^{n} y_i^2 - n\overline{y}_n^2\Big)}}\,, \tag{25}$$

  wobei diese alternative Darstellung des empirischen Korrelationskoeffizienten ρ_{xy} günstiger für das praktische Rechnen ist.
- Übungsaufgabe. Bestimmen Sie für die in Abschnitt 2.4.1 betrachteten Daten über den Jahresertrag bzw. die mittlere Clusterzahl je Traube
  - die Stichprobenmittel $\overline{x}_{12}$ und $\overline{y}_{12}$ sowie
  - den empirischen Korrelationskoeffizienten ρ_{xy}.

4. *Empirischer Korrelationskoeffizient bei binären Daten*

- Außerdem lässt sich für *binäre Daten*, d.h., falls die Stichprobenwerte $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ nur 0 oder 1 sein können, noch eine weitere nützliche Darstellungsformel für den empirischen Korrelationskoeffizienten ρ_{xy} angeben.
- Mit der in Abschnitt 2.3.1 eingeführten Notation gilt dann

$$\rho_{xy} = \frac{h_{11}\,h_{22} - h_{12}\,h_{21}}{\sqrt{(h_{11}+h_{12})(h_{11}+h_{21})(h_{12}+h_{22})(h_{21}+h_{22})}}\,, \tag{26}$$

  wobei $h_{ij} = h(c_i, d_j)$ für jedes $i = 1, \ldots, k_1$ unf für jedes $j = 1, \ldots, k_2$ die absolute Häufigkeit bezeichnet, mit der die Kombination (c_i, d_j) der Ausprägungen $c_i \in \{0,1\}$ und $d_j \in \{0,1\}$ in den Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ auftritt.
- Beachte
  - Wenn man die Formeln (18) und (26) miteinander vergleicht, dann erkennt man, dass der χ^2-Koeffizient T und der empirische Korrelationskoeffizient ρ_{xy} bei binären Daten wie folgt zusammenhängen: Es gilt

$$\rho_{xy}^2 = \frac{T}{n}\,. \tag{27}$$

  - Wir betrachten nun erneut das in Abschnitt 2.3.1 eingeführte Beispiel mit den Ausprägungen „keine Ausbildung" bzw. „Lehre" für das Merkmal „Ausbildungsniveau" sowie den Ausprägungen „mittelfristige Arbeitslosigkeit" (7–12 Monate) bzw. „langfristige Arbeitslosigkeit (> 12 Monate) für das Merkmal „Dauer der Arbeitslosigkeit".
  - Wenn wir dabei die Eintragungen der 2×2–Kontingenztafel (20) in die Darstellungsformel (26) einsetzen, dann ergibt sich, dass

$$\rho_{xy} = -0.168\,.$$

  - Hieraus und aus (27) ergibt sich darüber hinaus, dass

$$T = n\rho_{xy}^2 = 100 \cdot 0.02826 = 2.826\,,$$

    was mit dem Ergebnis (21) übereinstimmt, das bereits am Ende von Abschnitt 2.3.3 ermittelt wurde.

5. *Invarianzeigenschaft bei linearer Daten-Transformation*

   - Die Deutung des empirischen Korrelationskoeffizienten ρ_{xy} als Maßzahl zur Quantifizierung des (linearen) Zusammenhanges zwischen den Ausprägungen zweier Merkmale X und Y wird auch durch die folgende *Invarianzeigenschaft* von $|\rho_{xy}|$ bei linearer Daten-Transformation gestützt.
   - Außer den ursprünglichen Stichprobenwerten $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$ betrachten wir noch die *linear transformierten Stichprobenwerte* $x'_1, \ldots, x'_n$ bzw. $y'_1, \ldots, y'_n$, wobei $x'_i = \alpha_x + \beta_x x_i$ bzw. $y'_i = \alpha_y + \beta_y y_i$ für jedes $i = 1, \ldots, n$ und für gewisse Konstanten $\alpha_x, \alpha_y \in \mathbb{R}$ und $\beta_x, \beta_y \neq 0$.
   - Es gilt dann

$$\rho_{x'y'} = \begin{cases} \rho_{xy} & \text{falls } \beta_x,\, \beta_y > 0 \text{ bzw. } \beta_x,\, \beta_y < 0, \\ -\rho_{xy} & \text{falls } \beta_x > 0,\, \beta_y < 0 \text{ bzw. } \beta_x < 0,\, \beta_y > 0. \end{cases} \tag{28}$$

2.4.3 Herleitung der Formeln für ρ_{xy}

1. *Herleitung der Formel* (24)

   - Mit den abkürzenden Bezeichnungen

$$u_i = \frac{x_i - \overline{x}_n}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}}, \qquad v_i = \frac{y_i - \overline{y}_n}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y}_n)^2}}$$

   gilt offenbar

$$\sum_{i=1}^{n} u_i = 0\,, \quad \sum_{i=1}^{n} u_i^2 = 1 \qquad \text{und} \qquad \sum_{i=1}^{n} v_i = 0\,, \quad \sum_{i=1}^{n} v_i^2 = 1$$

   sowie

$$\rho_{xy} = \sum_{i=1}^{n} u_i v_i\,.$$

   - Somit gilt für jedes $c \in \mathbb{R}$

$$0 \leq \sum_{i=1}^{n} (u_i - c v_i)^2 = \underbrace{\sum_{i=1}^{n} u_i^2}_{=1} - 2c \sum_{i=1}^{n} u_i v_i + c^2 \underbrace{\sum_{i=1}^{n} v_i^2}_{=1}\,.$$

   - Hieraus ergibt sich insbesondere für $c = \sum_{i=1}^{n} u_i v_i$, dass

$$0 \leq 1 - \Bigg(\underbrace{\sum_{i=1}^{n} u_i v_i}_{=\rho_{xy}} \Bigg)^2\,,$$

   - Dies impliziert, dass $\rho_{xy}^2 \leq 1$ bzw. $-1 \leq \rho_{xy} \leq 1$, womit die Gültigkeit von (24) bewiesen ist.

2. *Herleitung der Formel* (25)

- Durch Ausmultiplizieren des Zählers in (23) ergibt sich, dass

$$\begin{aligned}\sum_{i=1}^{n}\Big((x_i-\overline{x}_n)(y_i-\overline{y}_n)\Big) &= \sum_{i=1}^{n}\Big(x_iy_i-\overline{x}_ny_i-\overline{y}_nx_i+\overline{x}_n\overline{y}_n\Big)\\ &= \sum_{i=1}^{n}x_iy_i-n\overline{x}_n\overline{y}_n\,.\end{aligned}$$

- Auf ähnliche Weise erhalten wir die Identitäten

$$\sum_{i=1}^{n}(x_i-\overline{x}_n)^2=\sum_{i=1}^{n}(x_i^2-2x_i\overline{x}_n+\overline{x}_n^2)=\sum_{i=1}^{n}x_i^2-n\overline{x}_n^2$$

und

$$\sum_{i=1}^{n}(y_i-\overline{y}_n)^2=\sum_{i=1}^{n}(y_i^2-2y_i\overline{y}_n+\overline{y}_n^2)=\sum_{i=1}^{n}y_i^2-n\overline{y}_n^2\,.$$

- Wenn diese Ausdrücke in (23) eingesetzt werden, ergibt sich nun mühelos die Formel (25).

3. *Herleitung der Formel* (26)

- Für binäre Daten gilt offenbar

$$\sum_{i=1}^{n}x_iy_i=h_{22}\,,\qquad n\overline{x}_n=h_{21}+h_{22}\,,\qquad n\overline{y}_n=h_{12}+h_{22}$$

sowie $n=h_{11}+h_{12}+h_{21}+h_{22}$, wobei h_{ij} die in Abschnitt 2.3.1 eingeführte absolute Häufigkeit ist.

- Durch Einsetzen in (25) ergibt sich nun, dass

$$\begin{aligned}\rho_{xy} &= \frac{\displaystyle\sum_{i=1}^{n}x_iy_i-n\overline{x}_n\overline{y}_n}{\sqrt{\Big(\displaystyle\sum_{i=1}^{n}x_i^2-n\overline{x}_n^2\Big)\Big(\displaystyle\sum_{i=1}^{n}y_i^2-n\overline{y}_n^2\Big)}}\\ &= \frac{h_{22}-\dfrac{(h_{21}+h_{22})(h_{12}+h_{22})}{h_{11}+h_{12}+h_{21}+h_{22}}}{\sqrt{\Big((h_{21}+h_{22})-\dfrac{(h_{21}+h_{22})^2}{h_{11}+h_{12}+h_{21}+h_{22}}\Big)\Big((h_{12}+h_{22})-\dfrac{(h_{12}+h_{22})^2}{h_{11}+h_{12}+h_{21}+h_{22}}\Big)}}\\ &= \frac{(h_{11}+h_{12}+h_{21}+h_{22})h_{22}-(h_{21}+h_{22})(h_{12}+h_{22})}{\sqrt{(h_{21}+h_{22})(h_{11}+h_{12})(h_{12}+h_{22})(h_{11}+h_{12})}}\\ &= \frac{h_{11}h_{22}-h_{12}h_{21}}{\sqrt{(h_{21}+h_{22})(h_{11}+h_{12})(h_{12}+h_{22})(h_{11}+h_{12})}}\,.\end{aligned}$$

- Damit ist (26) bewiesen.

4. *Herleitung der Formel* (28)

- Die Invarianzeigenschaft (28) des empirischen Korrelationskoeffizienten ergibt sich unmittelbar aus der Definitionsgleichung (23).

- Falls $x'_1, \ldots, x'_n$ bzw. $y'_1, \ldots, y'_n$, wobei $x'_i = \alpha_x + \beta_x x_i$ bzw. $y'_i = \alpha_y + \beta_y y_i$ für jedes $i = 1, \ldots, n$ und für gewisse Konstanten $\alpha_x, \alpha_y \in \mathbb{R}$ und $\beta_x, \beta_y \neq 0$, dann gilt nämlich, dass

$$\begin{aligned}
\rho_{x'y'} &= \frac{\sum_{i=1}^{n}\Big((x'_i - \overline{x}'_n)(y'_i - \overline{y}'_n)\Big)}{\sqrt{\sum_{i=1}^{n}\big(x'_i - \overline{x}'_n\big)^2 \sum_{i=1}^{n}\big(y'_i - \overline{y}'_n\big)^2}} \\
&= \frac{\sum_{i=1}^{n}\Big((\alpha_x + \beta_x x_i - (\alpha_x + \beta_x \overline{x}_n))(\alpha_y + \beta_y y_i - (\alpha_y + \beta_y \overline{y}_n))\Big)}{\sqrt{\sum_{i=1}^{n}\big(\alpha_x + \beta_x x_i - (\alpha_x + \beta_x \overline{x}_n)\big)^2 \sum_{i=1}^{n}\big(\alpha_y + \beta_y y_i - (\alpha_y + \beta_y \overline{y}_n)\big)^2}} \\
&= \frac{\beta_x \beta_y}{|\beta_x||\beta_y|}\, \rho_{xy}\,.
\end{aligned}$$

- Damit ist (28) bewiesen.

2.4.4 Ränge von Stichprobenwerten; Rang-Korrelationskoeffizient

1. *Ränge von Stichprobenwerten*

   Eine weitere Maßzahl zur Beschreibung des (gebenenfalls vorhandenen) Zusammenhanges zwischen den Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ der Merkmale/Kenngrößen/Variablen X bzw. Y erhalten wir,

   - wenn wir die sogenannten Ränge der Stichprobenwerte $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$ betrachten.
   - Hierfür gehen wir von den geordneten Stichproben $(x_{(1)}, \ldots, x_{(n)})$ bzw. $(y_{(1)}, \ldots, y_{(n)})$ mit $x_{(1)} \leq \ldots \leq x_{(n)}$ bzw. $y_{(1)} \leq \ldots \leq y_{(n)}$ aus, die bereits in Abschnitt 2.2.1 betrachtet wurden.
   - Dabei nehmen wir (der Einfachheit wegen) an, dass

     $$x_{(1)} < \ldots < x_{(n)} \qquad \text{und} \qquad y_{(1)} < \ldots < y_{(n)}\,,$$

     d.h., sämtliche Werte in den Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ seien voneinander verschieden.
   - Für jedes $i \in \{1, \ldots, n\}$ gibt es dann eine (eindeutig bestimmte) Zahl $j \in \{1, \ldots, n\}$, so dass $x_i = x_{(j)}$ gilt.
   - Die auf diese Weise (für jedes $i \in \{1, \ldots, n\}$) gegebene Zahl $j = j(i)$ heißt der *Rang* des Stichprobenwertes x_i, wobei der Rang von x_i mit $\mathrm{rg}\,(x_i)$ bezeichnet wird.
   - Der Rang $\mathrm{rg}\,(y_i)$ des Stichprobenwertes y_i wird auf analoge Weise definiert.

2. *Rang-Korrelationskoeffizient*

   - Der Rang-Korrelationskoeffizient ρ'_{xy} (auch *Spearmans Korrelationskoeffizient* genannt) ist der in (23) bzw. (25) gegebene empirische Korrelationskoeffizient für die *Rangstichproben* $\big(\mathrm{rg}\,(x_1), \ldots, \mathrm{rg}\,(x_n)\big)$ und $\big(\mathrm{rg}\,(y_1), \ldots, \mathrm{rg}\,(y_n)\big)$, d.h.,
     - es gilt

$$\rho'_{xy} = \frac{\sum_{i=1}^{n} \mathrm{rg}\,(x_i)\mathrm{rg}\,(y_i) - n\overline{\mathrm{rg}}_x \overline{\mathrm{rg}}_y}{\sqrt{\Big(\sum_{i=1}^{n} \mathrm{rg}^{\,2}(x_i) - n(\overline{\mathrm{rg}}_x)^2\Big)\Big(\sum_{i=1}^{n} \mathrm{rg}^{\,2}(y_i) - n\overline{\mathrm{rg}}_y)^2\Big)}}\,, \tag{29}$$

  - wobei die Mittelwerte $\overline{\mathrm{rg}}_x$ bzw. $\overline{\mathrm{rg}}_y$ der Ränge der Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ gegeben sind durch

$$\overline{\mathrm{rg}}_x = \frac{1}{n} \sum_{i=1}^{n} \mathrm{rg}\,(x_i) = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n+1}{2}\,, \qquad \overline{\mathrm{rg}}_y = \frac{1}{n} \sum_{i=1}^{n} \mathrm{rg}\,(y_i) = \frac{1}{n} \sum_{i=1}^{n} i = \frac{n+1}{2}\,.$$

- Aus der Definitionsgleichung (29) folgt, dass
  - der Rang-Korrelationskoeffizient ρ'_{xy} *robuster* ist als der in Abschnitt 2.4.2 eingeführte empirische Korrelationskoeffizient ρ_{xy}, d.h.,
  - das Zusammenhangsmaß ρ'_{xy} reagiert weniger stark als ρ_{xy} auf extreme Wertepaare (x_i, y_i).
- der Rang-Korrelationskoeffizient ρ'_{xy} nimmt Werte im Intervall $[-1, 1]$ an, wobei
  - ρ'_{xy} eine Maßzahl für den *monotonen Zusammenhang* zwischen den Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ ist,
  - im Unterschied zum empirischen Korrelationskoeffizienten ρ_{xy}, der eine Maßzahl für den *linearen Zusammenhang* zwischen $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ ist.
- Dabei ist $\rho'_{xy} = 1$,
  - falls für beliebige Paare $i, j \in \{1, \ldots, n\}$ mit $i \neq j$ die Ungleichung $x_i < x_j$ genau dann gilt, wenn $y_i < y_j$.
  - In diesem Fall liegen sämtliche Rangpaare $(\mathrm{rg}\,(x_i), \mathrm{rg}\,(y_i))$ für $i = 1, \ldots, n$ auf einer Geraden mit positivem Anstieg.
- Der umgekehrte Extremfall $\rho'_{xy} = -1$ ist gegeben,
  - falls für beliebige Paare $i, j \in \{1, \ldots, n\}$ mit $i \neq j$ die Ungleichung $x_i < x_j$ genau dann gilt, wenn $y_i > y_j$.
  - In diesem Fall liegen sämtliche Rangpaare $(\mathrm{rg}\,(x_i), \mathrm{rg}\,(y_i))$ für $i = 1, \ldots, n$ auf einer Geraden mit negativem Anstieg.

3. *Alternative Darstellungsformel*

Falls sämtliche Werte in den Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ voneinander verschieden sind, dann lässt sich der Rang-Korrelationskoeffizient ρ'_{xy} auch in der folgenden Form darstellen:

$$\rho'_{xy} = 1 - \frac{6 \sum_{i=1}^{n} \big(\mathrm{rg}\,(x_i) - \mathrm{rg}\,(y_i)\big)^2}{(n^2 - 1)n}\,, \tag{30}$$

wobei diese alternative Darstellung des Korrelationskoeffizienten ρ'_{xy} günstiger für das praktische Rechnen ist.

4. *Invarianzeigenschaft bei monotoner Daten-Transformation*

- Außer den ursprünglichen Stichprobenwerten $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_n$ betrachten wir nun noch die *monoton transformierten Stichprobenwerte* $x'_1, \ldots, x'_n$ bzw. $y'_1, \ldots, y'_n$, wobei $x'_i = g(x_i)$ bzw. $y'_i = h(y_i)$ für jedes $i = 1, \ldots, n$ und für Funktionen $g, h : \mathbb{R} \to \mathbb{R}$, die entweder (streng) monoton wachsend oder fallend sind.
- Bei solchen monotonen Daten-Transformationen
  - ändern sich die Ränge der Stichprobenwerte nicht, falls g bzw. h monoton wachsend ist,
  - oder sie kehren sich um, falls g bzw. h monoton fallend ist.

- Unmittelbar aus der Definitionsgleichung (29) des Rang-Korrelationskoeffizienten ρ'_{xy} ergibt sich somit, dass

$$\rho'_{xy} = \begin{cases} \rho'_{x'y'} & \text{falls } g, h \text{ wachsend oder } g, h \text{ fallend,} \\ -\rho'_{x'y'} & \text{falls } g \text{ wachsend und } h \text{ fallend (oder umgekehrt).} \end{cases} \tag{31}$$

5. *Stichproben mit Bindungen*

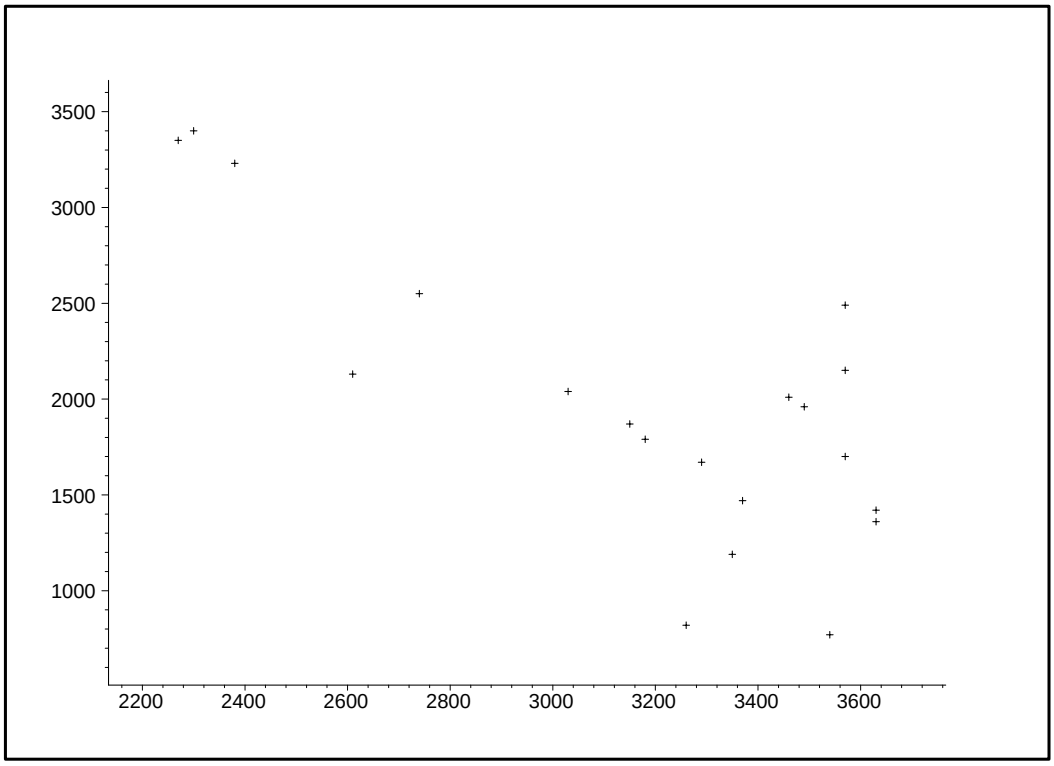
- Wenn nicht sämtliche Werte in den Stichproben $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ voneinander verschieden sind, d.h.,
  - wenn es Werte gibt, die mehrfach in $(x_1, \ldots, x_n)$ bzw. $(y_1, \ldots, y_n)$ vorkommen,
  - dann spricht man von Stichproben mit *Bindungen*.
- In diesem Fall bildet man sogenannte *Durchschnittsränge*, die jedoch bewirken, dass die Zuordnung $i \mapsto j(i)$ der Ränge nicht mehr eineindeutig ist.
- Falls beispielsweise
  - der Stichprobenwert 6.90 zweimal in der Stichprobe vorkommt und falls es nur einen Wert gibt, der kleiner als 6.90 ist,
  - so wird den beiden Stichprobenwerten, die gleich 6.90 sind, jeweils der Rang $\mathrm{rg}\,(6.90) = (2+3)/2 = 2.5$ zugeordnet (wobei die Ränge 2 und 3 dann entfallen).
- Mit Hilfe dieses allgemeineren Rang-Begriffes lässt sich der Rang-Korrelationskoeffizient ρ'_{xy} auch für Stichproben mit Bindungen durch die Definitionsgleichung (29) bestimmen.

6. *Beispiel* (vgl. J. Hüsler, H. Zimmermann (2001) *Statistische Prinzipien für medizinische Projekte*. Huber-Verlag, Bern, S. 183 ff.)

- Im Rahmen einer medizinischen Studie wurde das Geburtsgewicht (in g) von 20 Säuglingen sowie die Gewichtszunahme (in g) der Säuglinge zwischen dem 70. und 100. Tag untersucht.
- Dabei wurden die folgenden Daten beobachtet:

Säugling	Geburts-gewicht x_i	Gewichts-zunahme y_i	Säugling	Geburts-gewicht x_i	Gewichts-zunahme y_i
1	2740	2550	11	3260	820
2	3180	1790	12	3350	1190
3	3150	1870	13	3630	1360
4	3030	2040	14	3630	1420
5	3370	1470	15	3490	1960
6	2610	2130	16	3290	1670
7	3570	2150	17	3540	770
8	2270	3350	18	3570	1700
9	2300	3400	19	3460	2010
10	2380	3230	20	3570	2490

- Für diesen Datensatz ergibt sich das folgende Streudiagramm:



- Außerdem ergeben sich die folgenden Werte für
  - Stichprobenmittel bzw. Standardabweichung der Geburtsgewichte $x_1, \ldots, x_{20}$:

$$\overline{x}_{20} = 3169.5\,, \qquad s_x = 460.6\,,$$

  - Stichprobenmittel bzw. Standardabweichung der Gewichtszunahmen $y_1, \ldots, y_{20}$:

$$\overline{y}_{20} = 1968.5\,, \qquad s_y = 750.8\,,$$

  - sowie für den empirischen Korrelationskoeffizienten:

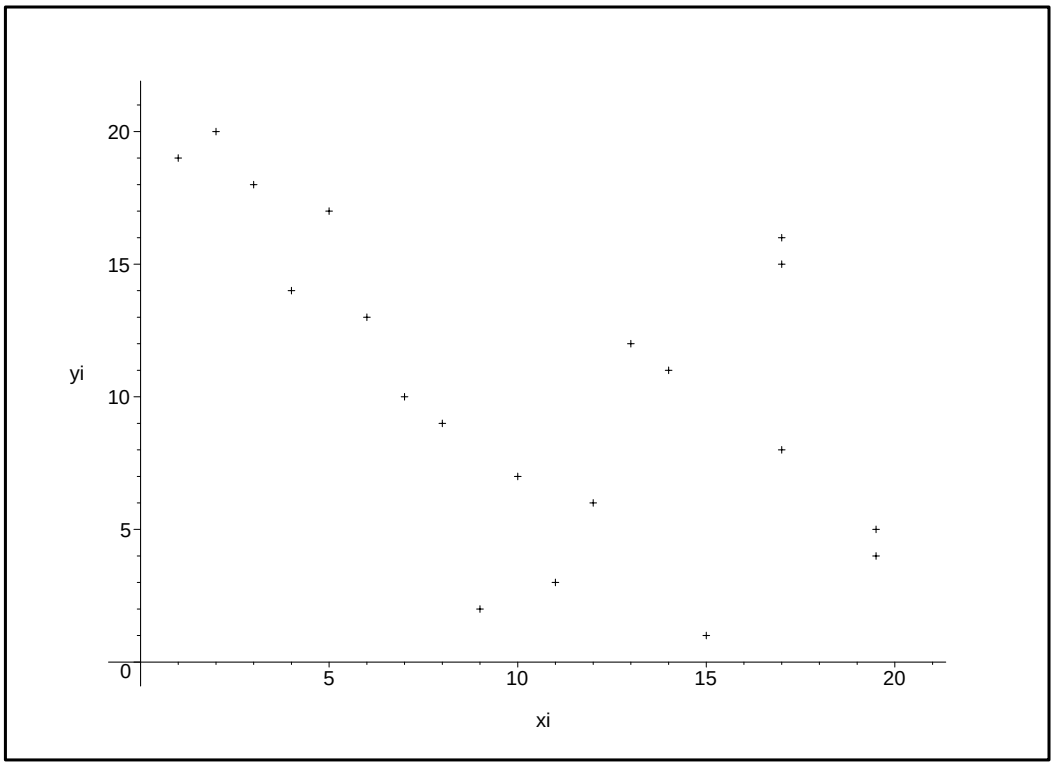
$$\rho_{xy} = -0.762\,. \tag{32}$$

- Beachte
  - Die beiden Merkmale „Geburtsgewicht" und „Gewichtszunahme" sind *negativ korreliert*, d.h.,
  - kleine Geburtsgewichte sind mit großen Gewichtszunahmen (und umgekehrt große Geburtsgewichte mit kleinen Gewichtszunahmen) zwischen dem 70. und 100. Tag verbunden.
  - Der in (32) gegebene empirische Korrelationskoeffizient ρ_{xy} liegt jedoch nicht in unmittelbarer Nähe von -1, d.h. die Wertepaare (x_i, y_i) liegen nicht sehr nahe an einer Geraden (mit negativem Anstieg), vgl. auch das Streudiagramm.
  - Aus dem Streudiagramm ist außerdem ersichtlich, dass der (näherungsweise) lineare Zusammenhang zwischen den Merkmalen „Geburtsgewicht" und „Gewichtszunahme" überwiegend durch eine kleine Anzahl von Wertepaaren (x_i, y_i) mit (extrem) geringem Geburtsgewicht x_i getragen wird.
  - Dies lässt vermuten, dass der Rang-Korrelationskoeffizient ρ'_{xy} der beiden Stichproben $(x_1, \ldots, x_{20})$ und $(y_1, \ldots, y_{20})$ näher bei 0 liegt als der in (32) gegebene empirische Korrelationskoeffizient ρ_{xy}.

- In der folgenden Tabelle sind die Ränge der Stichprobenwerte $x_1, \ldots, x_{20}$ bzw. $y_1, \ldots, y_{20}$ gegeben:

Säugling	Rang Geb.-gewicht $\mathrm{rg}(x_i)$	Rang Gew.-zunahme $\mathrm{rg}(y_i)$	Säugling	Rang Geb.-gewicht $\mathrm{rg}(x_i)$	Rang Gew.-zunahme $\mathrm{rg}(y_i)$
1	5	17	11	9	2
2	8	9	12	11	3
3	7	10	13	19.5	4
4	6	13	14	19.5	5
5	12	6	15	14	11
6	4	14	16	10	7
7	17	15	17	15	1
8	1	19	18	17	8
9	2	20	19	13	12
10	3	18	20	17	16

- Für diesen Datensatz ergibt sich das folgende Streudiagramm:



- In dem Streudiagramm der Ränge ist der (negativ korrelierte) lineare Zusammenhang der beiden Merkmale „Geburtsgewicht" und „Gewichtszunahme" weniger deutlich (als vorher in dem Streudiagramm der Rohdaten) ausgeprägt.
- Das wird auch durch den Rang-Korrelationskoeffizient $\rho'_{xy} = -0.56$ unterstrichen, den man durch Einsetzen in die Definitonsgleichung (29) erhält und der deutlich näher bei 0 liegt als $\rho_{xy} = -0.762$.

2.5 Lineare Regression

2.5.1 Modellbeschreibung

- Wir betrachten die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ von Ausprägungen zweier metrisch skalierter Merkmale/Kenngrößen/Variablen X bzw. Y, wobei wir Y als *Zielvariable* und X als *Ausgangsvariable* deuten.
- Falls der Betrag $|\rho_{xy}|$ des empirischen Korrelationskoeffizienten der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ groß ist, dann ist es sinnvoll anzunehmen, dass ein Zusammenhang zwischen den Ausprägungen $y_1, \ldots, y_n$ der Zielvariablen Y und den Ausprägungen $x_1, \ldots, x_n$ der Ausgangsvariablen X besteht.
- Diesen Zusammenhang modellieren wir dann wie folgt:
  - Wir nehmen an, dass es eine Funktion $\varphi : \mathbb{R} \to \mathbb{R}$ und reelle Zahlen $\varepsilon_1, \ldots, \varepsilon_n$ gibt mit

$$y_i = \varphi(x_i) + \varepsilon_i\,, \qquad \forall\, i = 1, \ldots, n\,. \tag{33}$$

  - Dabei gehen wir davon aus, dass sowohl die sogenannte *Regressionsfunktion* $\varphi : \mathbb{R} \to \mathbb{R}$ als auch die *Störgrößen* $\varepsilon_1, \ldots, \varepsilon_n$, durch die beispielsweise Messfehler modelliert werden können, nicht direkt beobachtbar (und somit unbekannt) sind.

- Beachte
  - Ein wichtiger Spezialfall liegt vor, wenn die Regressionsfunktion $\varphi : \mathbb{R} \to \mathbb{R}$ eine lineare Funktion ist, die sogenannte *Regressionsgerade*, d.h., wenn es reelle Zahlen α, β gibt mit

$$\varphi(x) = \alpha + \beta x\,, \qquad \forall\, x \in \mathbb{R}\,, \tag{34}$$

    wobei α die *Regressionskonstante* und β der *Regressionskoeffizient* genannt wird.
  - Das in (33) betrachtete Modell für den Zusammenhang zwischen den Ausprägungen $y_1, \ldots, y_n$ der Zielvariablen Y und den Ausprägungen $x_1, \ldots, x_n$ der Ausgangsvariablen X wird dann (einfaches) *lineares Regressionsmodell* genannt.

2.5.2 Methode der kleinsten Quadrate

- Wir nehmen nun an,
  - dass die Regressionsfunktion $\varphi : \mathbb{R} \to \mathbb{R}$ die in (34) gegebene Form besitzt, d.h., wir betrachten das lineare Regressionsmodell,
  - und passen die Gerade $\varphi(x) = \alpha + \beta x$, die auch *Ausgleichsgerade* genannt wird, mit der Methode der kleinsten Quadrate an die Datenpunkte $(x_1, y_1), \ldots, (x_n, y_n)$ an.
- Mit anderen Worten: Wir zeigen,

– wie die unbekannten Modellparameter α und β zu wählen sind, um den *mittleren quadratischen Fehler*

$$e(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - (\alpha+\beta x_i)\big)^2 \tag{35}$$

zu minimieren, und setzen dabei voraus,

– dass $n \geq 2$ und dass nicht alle Ausprägungen $x_1, \ldots, x_n$ von X einander gleich sind.

- Und zwar minimiert dann der Vektor $(\widehat{\alpha}, \widehat{\beta})$ mit

$$\widehat{\alpha} = \overline{y}_n - \widehat{\beta}\overline{x}_n \qquad \text{und} \qquad \widehat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} \tag{36}$$

den mittleren quadratischen Fehler $e(\alpha,\beta)$,

– wobei $\overline{x}_n$, $\overline{y}_n$ so wie bisher die Stichprobenmittel von $(x_1,\ldots,x_n)$ bzw. $(y_1,\ldots,y_n)$ bezeichnen, d.h.

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \overline{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$

– und die Stichprobenvarianzen s_{xx}^2, s_{yy}^2 bzw. die Stichprobenkovarianz s_{xy}^2 gegeben sind durch

$$s_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(x_i-\overline{x}_n\big)^2\,, \quad s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)(y_i-\overline{y}_n)\,, \quad s_{yy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(y_i-\overline{y}_n\big)^2\,.$$

- Beachte

– Die in (36) angegebene Lösung $(\widehat{\alpha}, \widehat{\beta})$ des Minimierungsproblems lässt sich wie folgt herleiten.

– Wenn man die Funktion $e(\alpha,\beta)$ nach α differenziert, so erkennt man, dass für jedes $\beta \in \mathbb{R}$ die Zahl

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i) = \overline{y}_n - \beta\overline{x}_n$$

den Wert des folgenden Ausdruckes minimiert:

$$\sum_{i=1}^{n}\big(y_i-(\alpha+\beta x_i)\big)^2 = \sum_{i=1}^{n}\big((y_i-\beta x_i)-\alpha\big)^2\,.$$

– Mit anderen Worten: Für jedes $\beta \in \mathbb{R}$ ist

$$\begin{aligned}\sum_{i=1}^{n}\big((y_i-\beta x_i)-(\overline{y}_n-\beta\overline{x}_n)\big)^2 &= \sum_{i=1}^{n}\big((y_i-\overline{y}_n)-\beta(x_i-\overline{x}_n)\big)^2\\ &= (n-1)\big(s_{yy}^2 - 2\beta s_{xy}^2 + \beta^2 s_{xx}^2\big)\end{aligned}$$

der kleinste Wert des mittleren quadratischen Fehlers.

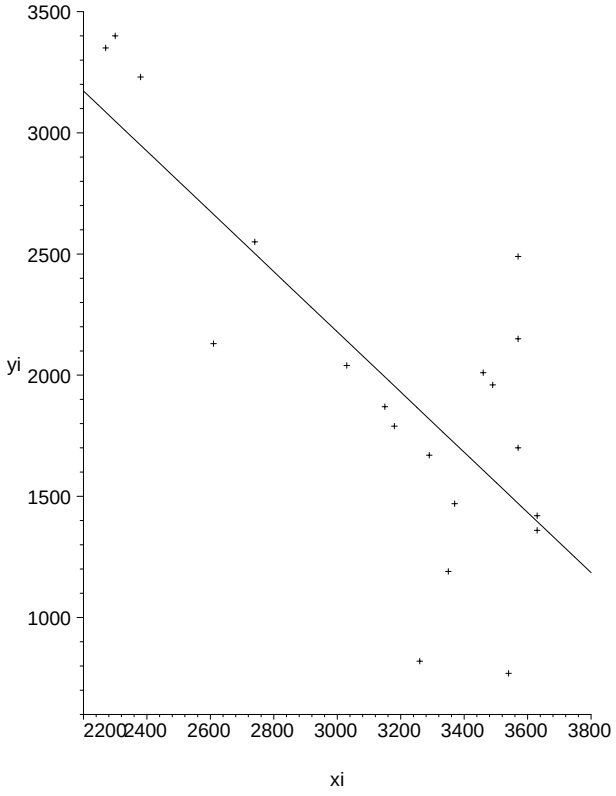
– Durch Differenzieren dieses Ausdruckes nach β ergibt sich nun, dass das globale Minimum an der folgenden Stelle angenommen wird;

$$\beta = \frac{s_{xy}^2}{s_{xx}^2}\,.$$

- Beispiel

– Für das in Abschnitt 2.4.4 betrachtete Beispiel der beiden Merkmale „Geburtsgewicht" und „Gewichtszunahme" ergibt sich, dass $\widehat{\alpha} = 5905$ und $\widehat{\beta} = -1.242$.

– An das Streudiagramm dieses Datensatzes lässt sich somit die folgende Regressionsgerade anpassen:



- Beachte
  - Der in (35) definierte mittlere quadratische Fehler $e(\alpha, \beta)$ ist der mittlere quadratische *vertikale* Abstand zwischen den „beobachteten" Punkten (x_i, y_i) und den entsprechenden Punkten $(x_i, \varphi(x_i))$ auf der Regressionsgeraden $\varphi(x) = \alpha + \beta x$ an den Stellen $x_1, \ldots, x_n$.
  - Anstelle der vertikalen Abstände kann man beispielsweise auch die horizontalen Abstände betrachten. Durch Vertauschen der Rollen von x und y ergibt sich dann, dass

$$\widehat{\alpha}' = \overline{x}_n - \widehat{\beta}' \overline{y}_n \qquad \text{und} \qquad \widehat{\beta}' = \frac{s_{xy}^2}{s_{yy}^2}$$

    die optimalen Werte der Parameter $\alpha', \beta' \in \mathbb{R}$ der (inversen) Regressionsgeraden $\varphi'(y) = x = \alpha' + \beta' y$ sind.
  - Wenn wir diese Geradengleichung nach y auflösen, dann ergibt sich die Gleichung

$$y = -\frac{\alpha'}{\beta'} + \frac{1}{\beta'} x \,,$$

    wobei allerdings die Werte $-\widehat{\alpha}'/\widehat{\beta}'$ und $\widehat{\beta}'^{-1}$ für Regressionskonstante bzw. Regressionskoeffizient im allgemeinen verschieden von den optimalen Werten $\widehat{\alpha}$ bzw. $\widehat{\beta}$ sind, die in (36) hergeleitet worden sind.

- Übungsaufgabe
  - Bestimmen Sie für das in Abschnitt 2.4.1 betrachtete Beispiel der Merkmale „Clusterzahl je Traube" unsd „Jahresertrag" die Werte $\widehat{\alpha}$ und $\widehat{\beta}$ sowie $\widehat{\alpha}'$ und $\widehat{\beta}'$ und zeichnen Sie die beiden Regressionsgeraden in das Streudiagramm dieses Datensatzes ein.
  - Prognostizieren Sie mit Hilfe der Regressionsgerade $\widehat{y} = \widehat{\alpha} + \widehat{\beta} x$ den Jahresertrag, der einer mittleren Clusterzahl von 100 Beeren je Traube entsprechen würde.

2.5.3 Güte der Modellanpassung; Quadratsummen-Zerlegung

- In diesem Abschnitt diskutieren wir eine Maßzahl, die
  - die *Anpassungsgüte des Regressionsmodells* $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$ an die beobachteten Werte $y_1, \ldots, y_n$ der Zielvariablen Y beschreibt,
  - falls die Regressionskonstante $\widehat{\alpha}$ und der Regressionskoeffizient $\widehat{\beta}$ durch (36) gegeben sind.
- In diesem Zusammenhang betrachten wir
  - für jedes $i = 1, \ldots, n$ die *Abweichung* $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ des beobachteten Wertes y_i von dem entsprechenden Wert $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ der Regressionsfunktion $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$,
  - wobei die Abweichungen $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$ auch *Residuen* genannt werden und
  - der in (35) eingeführte mittlere quadratische Fehler $e(\widehat{\alpha}, \widehat{\beta})$ das arithmetische Mittel

$$e(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

    der *Abweichungsquadrate* $\widehat{\varepsilon}_1^2, \ldots, \widehat{\varepsilon}_n^2$ ist.
- Es ist klar, dass die Anpassungsgüte des Regressionsmodells $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x$ an die beobachteten Werte $y_1, \ldots, y_n$ der Zielvariablen Y um so schlechter ist,
  - je größer die sogenannte *Residualstreuung* $e(\widehat{\alpha}, \widehat{\beta})$ ist,
  - wobei diese (absolute) Abweichungsmaßzahl noch mit der *Gesamtstreuung* $1/n \sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ der beobachteten Werte $y_1, \ldots, y_n$ normiert wird.
- Dies führt dann zu der *Bestimmtheitsmaßzahl*

$$R^2 = 1 - \frac{\displaystyle\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2}\,, \tag{37}$$

  die auch *Determinationskoeffizient* genannt wird.
- Beachte
  1. Das Bestimmtheitsmaß R^2 nimmt nur Werte zwischen 0 und 1 an, d.h., es gilt stets

$$0 \leq R^2 \leq 1\,. \tag{38}$$

     - Dies ergibt sich aus der folgenden *Quadratsummen-Zerlegung*

$$\sum_{i=1}^{n} (y_i - \overline{y}_n)^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2 + \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2\,. \tag{39}$$

     - Denn offenbar gilt $R^2 \leq 1$, und aus (39) folgt, dass

$$R^2 = 1 - \frac{\displaystyle\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2} = \frac{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2 - \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2} \overset{(39)}{=} \frac{\displaystyle\sum_{i=1}^{n} (\widehat{y}_i - \overline{y}_n)^2}{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y}_n)^2} \geq 0\,.$$

– Die hierbei verwendete Quadratsummen-Zerlegung (39) lässt sich wie folgt (durch Einfügen der „nahrhaften" Null $\widehat{y}_i - \widehat{y}_i$) herleiten, denn es gilt

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y}_n)^2 &= \sum_{i=1}^{n}\Big((y_i-\widehat{y}_i)+(\widehat{y}_i-\overline{y}_n)\Big)^2\\ &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+2\underbrace{\sum_{i=1}^{n}(y_i-\widehat{y}_i)(\widehat{y}_i-\overline{y}_n)}_{\overset{(36)}{=}0}+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y}_n)^2\\ &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y}_n)^2\,.\end{aligned}$$

2. Außerdem gilt

$$R^2=\rho_{xy}^2\,, \tag{40}$$

– d.h., das Bestimmtheitsmaß R^2 stimmt mit dem Quadrat des empirischen Korrelationskoeffizienten ρ_{xy} überein, der in Abschnitt 2.4.2 eingeführt wurde, wobei sich die Gültigkeit von (40) aus den folgenden Überlegungen ergibt.

– Aus (36) folgt zunächst, dass das arithmetische Mittel $\overline{\widehat{y}}$ von $\widehat{y}_1,\ldots,\widehat{y}_n$ mit dem Stichprobenmittel $\overline{y}_n$ übereinstimmt, denn es gilt

$$\overline{\widehat{y}}=\frac{1}{n}\sum_{i=1}^{n}\widehat{y}_i=\frac{1}{n}\sum_{i=1}^{n}(\widehat{\alpha}+\widehat{\beta}x_i)=\widehat{\alpha}+\widehat{\beta}\overline{x}_n\overset{(36)}{=}(\overline{y}_n-\widehat{\beta}\overline{x}_n)+\widehat{\beta}\overline{x}_n=\overline{y}_n\,.$$

– Hieraus ergibt sich, dass

$$\sum_{i=1}^{n}\big(\widehat{y}_i-\overline{y}_n\big)^2=\sum_{i=1}^{n}\big(\widehat{y}_i-\overline{\widehat{y}}\big)^2=\sum_{i=1}^{n}\Big(\widehat{\alpha}+\widehat{\beta}x_i-(\widehat{\alpha}+\widehat{\beta}\overline{x}_n)\Big)^2=\widehat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x}_n)^2$$

und somit

$$R^2=\frac{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y}_n)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_n)^2}=\frac{\widehat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x}_n)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_n)^2}\overset{(36)}{=}\frac{(s_{xy}^2)^2s_{xx}^2}{(s_{xx}^2)^2s_{yy}^2}=\left(\frac{s_{xy}^2}{\sqrt{s_{xx}^2s_{yy}^2}}\right)^2=\rho_{xy}^2\,.$$

3 Regressions- und Varianzanalyse

- Wir zeigen nun, wie die in Abschnitt 2 diskutierten Begriffe und Techniken der beschreibenden Statistik weiterentwickelt werden können,
  - um (neben der Darstellung bzw. Beschreibung von Datensätzen) eine tiefergehende *Analyse, Bewertung und Interpretation der Daten* zu ermöglichen,
  - wobei Begriffe und Techniken der *beurteilenden Statistik* (auch schließende, induktive bzw. inferentielle Statistik genannt) genutzt werden,
  - die auf *Methoden der mathematischen Stochastik* beruhen.
- Dabei werden Begriffe und Methoden der Wahrscheinlichkeitsrechnung und Statistik vertieft, die teilweise bereits im Grundkurs *"Stochastik für Wirtschaftswissenschaftler"* eingeführt worden sind.
- Insbesondere benötigen wir solche Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik wie
  - Zufallsvariable, Verteilung, Erwartungswert, Varianz bzw.
  - Zufallsstichprobe, Parameterschätzer, Konfidenzintervall, Signifikanztest,

  wobei die Kenntnis dieser Begriffe und ihrer grundlegenden Eigenschaften vorausgesetzt wird (und sie deshalb hier nur gelegentlich kurz wiederholt werden).

3.1 Einfache lineare Regression

- So wie bisher gehen wir bei der einfachen linearen Regression von zwei Datensätzen $(x_1, \dots, x_n) \in \mathbb{R}^n$ und $(y_1, \dots, y_n) \in \mathbb{R}^n$ aus, deren Eigenschaften jedoch nun mit Hilfe eines *stochastischen Modells* untersucht werden sollen.
- Dabei fassen wir die (nicht beobachtbaren!) Störgrößen $\varepsilon_1, \dots, \varepsilon_n$ in der Regressionsgleichung (2.33)
  - als Realisierungen von n (stochastisch) unabhängigen und identisch verteilten Zufallsvariablen auf,
  - durch die beispielsweise zufällige Messfehler modelliert werden können und
  - die wir (der Einfachheit der Schreibweise wegen) ebenfalls mit $\varepsilon_1, \dots, \varepsilon_n$ bezeichnen.
- Wir nehmen also insbesondere an, dass
  - die zufälligen Störgrößen $\varepsilon_1, \dots, \varepsilon_n : \Omega \to \mathbb{R}$ Abbildungen sind, die über einem (nicht näher spezifizierten) Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$ definiert sind.
  - Dabei gelte für jedes $i = 1, \dots, n$

    $$\mathbb{E}\,\varepsilon_i = 0\,, \qquad \operatorname{Var}\varepsilon_i = \sigma^2\,, \tag{1}$$

    wobei $\sigma^2 > 0$ ein gewisser (im allgemeinen unbekannter) Modellparameter ist.
- Die beobachteten Zielwerte $y_1, \dots, y_n$ fassen wir als Realisierungen von n Zufallsvariablen $Y_1, \dots, Y_n$ auf, die auf die folgende Weise von den (deterministischen) Ausprägungen $x_1, \dots, x_n$ der *Ausgangsvariablen* und von den (zufälligen) Störgrößen $\varepsilon_1, \dots, \varepsilon_n$ abhängen:

  $$Y_i = \alpha + \beta x_i + \varepsilon_i\,, \qquad \forall\, i = 1, \dots, n\,. \tag{2}$$

- Beachte
  - Die *Zielvariablen* $Y_1, \dots, Y_n$ sind zwar (stochastisch) unabhängige, jedoch typischerweise *nicht* identisch verteilte Zufallsvariablen.
  - Aus den allgemeinen Rechenregeln für Erwartungswert bzw. Varianz ergibt sich für jedes $i = 1, \dots, n$

    $$\mathbb{E}\,Y_i = \alpha + \beta x_i\,, \qquad \operatorname{Var} Y_i = \sigma^2\,. \tag{3}$$

  - Die Regressionskonstante α und der Regressionskoeffizient β sowie die Varianz σ^2 der Störgrößen sind *unbekannte Modellparameter*, die aus den beobachteten Daten $x_1, \dots, x_n$ und $y_1, \dots, y_n$ geschätzt werden sollen.

3.1.1 Kleinste-Quadrate-Schätzer

- Zur Erinnerung: Bei der Konstruktion von *Schätzern* für (reellwertige) Modellparameter geht man wie folgt vor.
  - Man betrachtet eine Abbildung $\varphi : \mathbb{R}^n \to \mathbb{R}$, die den beobachteten Daten $y_1, \ldots, y_n$, d.h. jeder Realisierung $(y_1, \ldots, y_n)$ der Zufallsstichprobe $(Y_1, \ldots, Y_n)$, den *Schätzwert* $\varphi(y_1, \ldots, y_n)$ zuordnet.
  - Der zugehörige *Schätzer* ist dann die (reellwertige) Zufallsvariable $\varphi(Y_1, \ldots, Y_n) : \Omega \to \mathbb{R}$, die sich ergibt, wenn die Abbildungen $Y_1, \ldots, Y_n : \Omega \to \mathbb{R}$ und $\varphi : \mathbb{R}^n \to \mathbb{R}$ nacheinander ausgeführt werden.
  - Zur Vereinfachung der Schreibweise werden wir gelegentlich sowohl den Schätzer als auch den (aus den jeweils beobachteten Daten bestimmten) Schätzwert mit $\varphi(Y_1, \ldots, Y_n)$ bezeichnen.
- Mit der bereits in Abschnitt 2.5.2 diskutierten Methode der kleinsten Quadrate erhalten wir die folgenden Schätzer $\widehat{\alpha}$, $\widehat{\beta}$ bzw. S^2 für die Modellparameter α, β und σ^2:

$$\widehat{\alpha} = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\overline{x}_n (x_i - \overline{x}_n)}{\sum_{j=1}^{n} (x_j - \overline{x}_n)^2} \right) Y_i \,, \qquad \widehat{\beta} = \sum_{i=1}^{n} \frac{x_i - \overline{x}_n}{\sum_{j=1}^{n} (x_j - \overline{x}_n)^2} \, Y_i \tag{4}$$

  und

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \widehat{\alpha} - \widehat{\beta} x_i \right)^2 . \tag{5}$$

- Hieraus ergibt sich für die Erwartungswerte dieser Schätzer, dass

$$\mathbb{E}\,\widehat{\alpha} = \alpha \,, \qquad \mathbb{E}\,\widehat{\beta} = \beta \,, \qquad \mathbb{E}\,S^2 = \sigma^2 \,, \tag{6}$$

  - d.h., die Modellparameter α, β bzw. σ^2 werden im Mittel durch $\widehat{\alpha}$, $\widehat{\beta}$ bzw. S^2 „richtig" geschätzt.
  - Mit anderen Worten: $\widehat{\alpha}$, $\widehat{\beta}$ bzw. S^2 sind sogenannte *erwartungstreue Schätzer* für α, β bzw. σ^2.
- Für die Varianzen der Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$ gilt:

$$\operatorname{Var} \widehat{\alpha} = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \,, \qquad \operatorname{Var} \widehat{\beta} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \,. \tag{7}$$

- Beachte
  - Die in (4) betrachteten Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$ für α bzw. β sind sogenannte *lineare Schätzer*, d.h., sie sind Linearkombinationen der Stichprobenvariablen $Y_1, \ldots, Y_n$.
  - Die linearen Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$ sind im allgemeinen *nicht* unabhängig, denn für die Kovarianz $\operatorname{Cov}(\widehat{\alpha}, \widehat{\beta})$ von $\widehat{\alpha}$ und $\widehat{\beta}$ gilt:

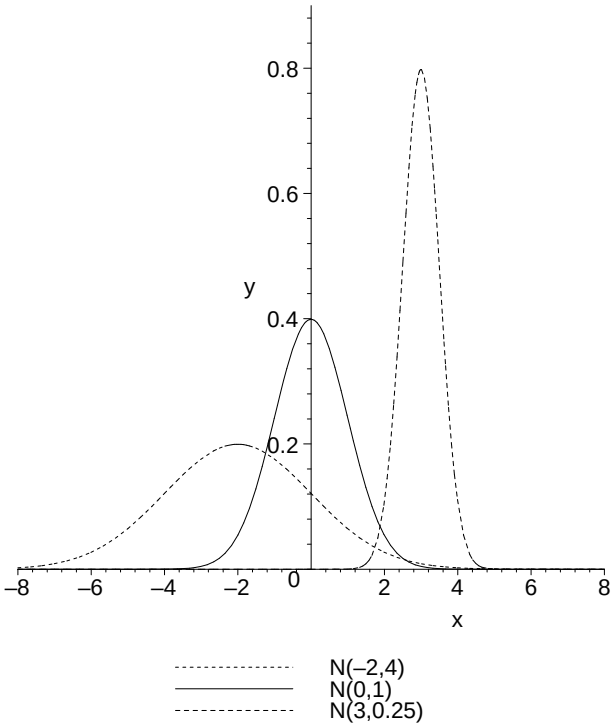
$$\operatorname{Cov}(\widehat{\alpha}, \widehat{\beta}) = - \frac{\sigma^2 \overline{x}_n}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2} \,. \tag{8}$$

3.1.2 Normalverteilte Störgrößen

- Um Aussagen über die Verteilungen der in (4) bzw. (5) betrachteten Kleinste-Quadrate-Schätzer $\widehat{\alpha}$, $\widehat{\beta}$ und S^2 machen zu können, benötigen wir Verteilungsannahmen über die (zufälligen) Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$.
  - Zusätzlich zu den Modellannahmen, die bisher in Abschnitt 3.1 gemacht wurden, setzen wir deshalb von nun an voraus, dass $n > 2$ und dass die (unabhängigen und identisch verteilten) Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ normalverteilt sind.
  - Zur Erinnerung: Seien $\mu \in \mathbb{R}$ und $\sigma^2 > 0$ beliebige, jedoch fest vorgegebene Zahlen. Man sagt, dass die Zufallsvariable $Z : \Omega \to \mathbb{R}$ *normalverteilt* ist mit dem Erwartungswert $\mathbb{E}\, Z = \mu$ und der Varianz $\operatorname{Var} Z = \sigma^2$ (und verwendet dann die Schreibweise $Z \sim \mathrm{N}(\mu, \sigma^2)$), falls die Dichte von Z gegeben ist durch

$$f_Z(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad \forall\, x \in \mathbb{R},$$

    mit der graphischen Darstellung:



  - Wegen (1) gilt dann $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, d.h., der Erwartungswert $\mathbb{E}\, \varepsilon_i$ der normalverteilten Störgröße ε_i ist 0 und die Varianz $\operatorname{Var} \varepsilon_i$ von ε_i ist σ^2 für jedes $i = 1, \ldots, n$.
  - Mit anderen Worten: Die *Wahrscheinlichkeitsdichte* $f_\varepsilon(x)$ der Zufallsvariablen $\varepsilon_1, \ldots, \varepsilon_n$ ist gegeben durch

$$f_\varepsilon(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad \forall\, x \in \mathbb{R}.$$

  - Wegen der Invarianzeigenschaften der Normalverteilung unter Linear-Transformation folgt hieraus außerdem, dass die (unabhängigen, jedoch im allgemeinen nicht identisch verteilten) Zielvariablen $Y_1, \ldots, Y_n$ ebenfalls normalverteilt sind, wobei

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2), \qquad \forall\, i = 1, \ldots, n. \tag{9}$$

- Weil die Kleinste-Quadrate-Schätzer $\widehat{\alpha}$ und $\widehat{\beta}$ Linearkombinationen der unabhängigen und normalverteilten Zielvariablen $Y_1, \ldots, Y_n$ sind, ergibt sich nun aus (4) und (9) (wegen der sogenannten *Faltungsstabilität* der Normalverteilung),
  - dass auch die Zufallsvariablen $\widehat{\alpha}$ und $\widehat{\beta}$ normalverteilt sind, wobei

$$\widehat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x}_n)^2}\right), \qquad \widehat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}\right). \tag{10}$$

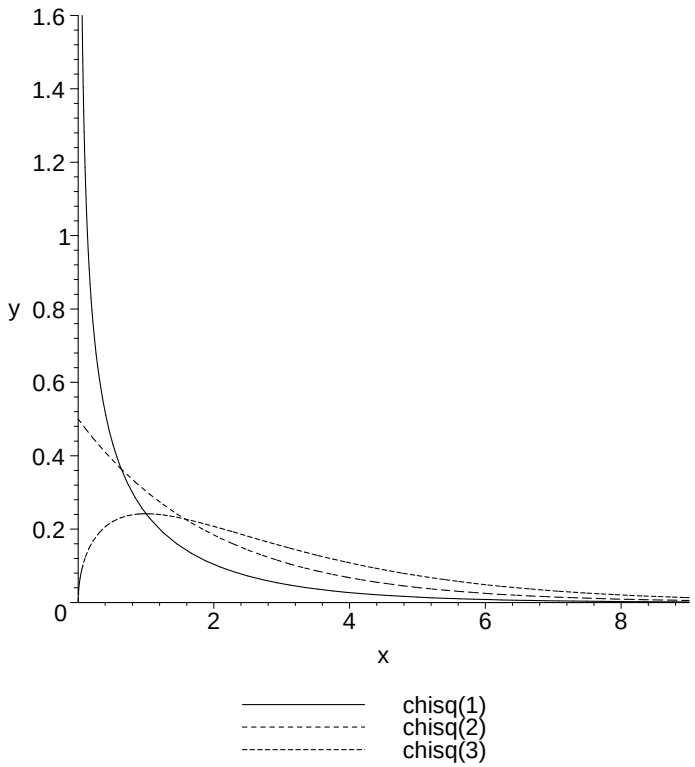
  - Außerdem kann man zeigen, dass der aus $\widehat{\alpha}$ und $\widehat{\beta}$ gebildete Zufallsvektor $(\widehat{\alpha}, \widehat{\beta})$ unabhängig ist von dem Kleinste-Quadrate-Schätzer S^2,
  - wobei die Zufallsvariable $(n-2)S^2/\sigma^2$ eine sogenannte χ^2-*Verteilung* mit $n-2$ *Freiheitsgraden* hat, d.h.,

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}. \tag{11}$$

- Zur Erinnerung: Sei $r \in \mathbb{N}$ eine beliebige natürliche Zahl, und seien $Z_1, \ldots, Z_r : \Omega \to \mathbb{R}$ unabhängige und $\mathrm{N}(0,1)$-verteilte Zufallsvariablen.
  - Dann sagt man, dass die Zufallsvariable $U_r = \sum_{i=1}^{r} Z_i^2$ eine χ^2-Verteilung mit r Freiheitsgraden hat. (Schreibweise: $U_r \sim \chi^2_r$).
  - Die Dichte von U_r ist gegeben durch

$$f_{U_r}(x) = \begin{cases} \dfrac{x^{(r-2)/2} e^{-x/2}}{2^{r/2}\,\Gamma(r/2)}, & \text{falls } x > 0, \\ 0 & \text{sonst}, \end{cases} \tag{12}$$

  wobei $\Gamma(1) = 1$, $\Gamma(1/2) = \sqrt{\pi}$ und $\Gamma(p+1) = p\Gamma(p)$, mit der graphischen Darstellung:



– Die χ^2-Verteilungen bilden eine Klasse von sogenannten *statistischen Prüfverteilungen*, die bei der Konstruktion von Signifikanztests bzw. Konfidenzintervallen nützlich sind, vgl. die nachfolgenden Abschnitte 3.1.3–3.1.6.

3.1.3 t-Tests für Regressionskonstante und Regressionskoeffizient

1. *Null-Hypothese und Alternativ-Hypothese*

- Für das Regressionsmodell mit normalverteilten Störgrößen kann man sogenannte t-*Tests* konstruieren,
  - um Hypothesen über die Regressionskonstante α bzw. den Regressionskoeffizienten β zu verifizieren,
  - wobei die in Abschnitt 3.1.2 betrachteten Verteilungseigenschaften der Kleinste-Quadrate-Schätzer $\widehat{\alpha}$, $\widehat{\beta}$ und S^2 nützlich sind.
- Für vorgegebene (hypothetische) Werte $\alpha_0, \beta_0 \in \mathbb{R}$ der Modellparameter α und β sind dabei hauptsächlich die folgenden Hypothesen-Paare von Interesse:

  i) *zweiseitige Hypothesen*

  $$H_0 : \alpha = \alpha_0 \text{ versus } H_1 : \alpha \neq \alpha_0 \qquad \text{bzw.} \qquad H_0 : \beta = \beta_0 \text{ versus } H_1 : \beta \neq \beta_0$$

  ii) *einseitige Hypothesen* (nach oben)

  $$H_0 : \alpha \geq \alpha_0 \text{ versus } H_1 : \alpha < \alpha_0 \qquad \text{bzw.} \qquad H_0 : \beta \geq \beta_0 \text{ versus } H_1 : \beta < \beta_0$$

  iii) *einseitige Hypothesen* (nach unten)

  $$H_0 : \alpha \leq \alpha_0 \text{ versus } H_1 : \alpha > \alpha_0 \qquad \text{bzw.} \qquad H_0 : \beta \leq \beta_0 \text{ versus } H_1 : \beta > \beta_0$$

- Das Grundprinzip, um zwischen der jeweiligen *Null-Hypothese* H_0 und der zugehörigen *Alternativ-Hypothese* H_1 abwägen zu können, lautet:
  - Falls die beobachteten Daten die Null-Hypothese H_0 nicht rechtfertigen, dann wird H_0 (zugunsten von H_1) abgelehnt.
  - Ansonsten wird die Null-Hypothese H_0 nicht abgelehnt.

2. *Konstruktion von Testgrößen*

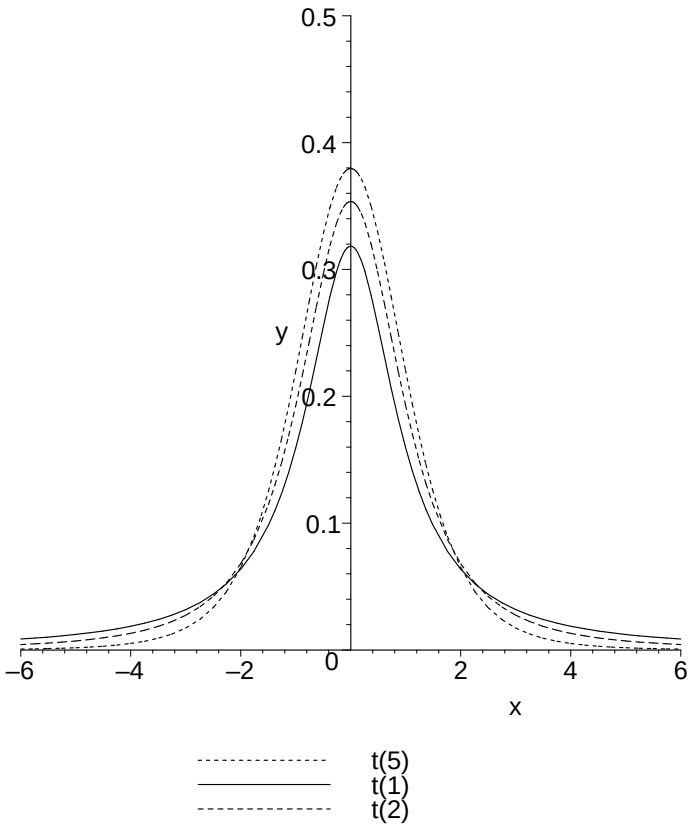
- Um eine solche Testentscheidung treffen zu können, werden Zufallsvariable konstruiert,
  - die *Testgröße* genannt werden und deren Verteilung *nicht* von den unbekannten Modellparametern abhängt.
  - Zur Verifizierung der obenbetrachteten Hypothesen-Paare werden insbesondere Testgrößen konstruiert, die t-verteilt sind, wobei die t-Verteilung eine weitere *statistische Prüfverteilung* ist.
- Zur Erinnerung: Sei $r \in \mathbb{N}$ eine beliebige natürliche Zahl, und seien Z und U_r unabhängige Zufallsvariablen mit $Z \sim \mathrm{N}(0,1)$ und $U_r \sim \chi^2_r$.
  - Dann sagt man, dass die Zufallsvariable

  $$V_r = Z \Big/ \sqrt{\frac{U_r}{r}}$$

  t-verteilt ist mit r Freiheitsgraden. (Schreibweise: $V_r \sim \mathrm{t}_r$)
  - Die Dichte von V_r ist gegeben durch

  $$f_{V_r}(v) = \frac{\Gamma((r+1)/2)}{\Gamma(r/2)} \, \frac{1}{\sqrt{r\pi}\,(1 + v^2/r)^{(r+1)/2}}\,, \qquad \forall\, v \in \mathbb{R}, \tag{13}$$

  mit der graphischen Darstellung:



- Aus den Verteilungs- und Unabhängigkeitseigenschaften der Kleinste-Quadrate-Schätzer $\widehat{\alpha}$, $\widehat{\beta}$ und S^2, die in Abschnitt 3.1.2 diskutiert worden sind, ergibt sich nun unmittelbar, dass

$$\frac{\widehat{\alpha}-\alpha}{S\sqrt{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big/\big(n(n-1)s_{xx}^2\big)}} \;\sim\; \mathrm{t}_{n-2} \tag{14}$$

und

$$\frac{\widehat{\beta}-\beta}{S/\sqrt{(n-1)s_{xx}^2}} \;\sim\; \mathrm{t}_{n-2}\,. \tag{15}$$

3. *Entscheidungsregel und Signifikanzniveau*

- Zur Erinnerung
  - Die Entscheidungsregel von statistischen Signifikanztests wird so konstruiert, dass die Wahrscheinlichkeit zu Fehlentscheidungen zu gelangen, unter einem vorgegebenem Schwellenwert liegt.
  - Insbesondere sollen richtige Null-Hypothesen nur extrem selten abgelehnt werden.
  - Die Wahrscheinlichkeit einer solchen Fehlentscheidung nennt man *Wahrscheinlichkeit des Fehlers erster Art.*
  - Üblicherweise werden für diese Wahrscheinlichkeit die Werte 0.01, 0.05 oder 0.1 vorgegeben und *Signifikanzniveau* des Tests genannt.
  - Dementsprechend wird die Null-Hypothese, falls sie richtig ist, mit der Wahrscheinlichkeit $\gamma = 0.99$, 0.95 bzw. 0.90 *nicht* verworfen.
- Wir erläutern nun, wie die obenbetrachten Hypothesen-Paare für das Regressionsmodell mit normalverteilten Störgrößen verifiziert werden können.

- Beim Test der Hypothese $H_0 : \alpha = \alpha_0$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \alpha \neq \alpha_0$) wird die Nullhypothese H_0 abgelehnt, falls

$$\frac{\left|\,\widehat{\alpha} - \alpha_0\,\right|}{S\sqrt{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big/\big(n(n-1)s_{xx}^2\big)}} \;>\; \mathrm{t}_{n-2,1-(1-\gamma)/2}\,, \tag{16}$$

  wobei $\mathrm{t}_{n-2,1-(1-\gamma)/2}$ das $(1-(1-\gamma)/2)$-Quantil der t-Verteilung mit $n-2$ Freiheitsgraden bezeichnet.
- Analog wird beim Test der Hypothese $H_0 : \beta = \beta_0$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \beta \neq \beta_0$) die Nullhypothese H_0 abgelehnt, falls

$$\frac{\left|\,\widehat{\beta} - \beta_0\,\right|}{S/\sqrt{(n-1)s_{xx}^2}} \;>\; \mathrm{t}_{n-2,1-(1-\gamma)/2}\,. \tag{17}$$

- Von besonderem Interesse ist der Test der Hypothese $H_0 : \beta = 0$ (gegen die Alternative $H_1 : \beta \neq 0$), wobei die Nullhypothese H_0 abgelehnt wird, falls

$$\frac{\left|\,\widehat{\beta}\,\right|}{S/\sqrt{(n-1)s_{xx}^2}} \;>\; \mathrm{t}_{n-2,1-(1-\gamma)/2}\,. \tag{18}$$

- Für einseitige Hypothesen werden ähnliche Entscheidungsregeln betrachtet:
  - So wird beispielsweise beim Test der Hypothese $H_0 : \alpha \geq \alpha_0$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \alpha < \alpha_0$) die Nullhypothese H_0 abgelehnt, falls

$$\frac{\widehat{\alpha} - \alpha_0}{S\sqrt{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big/\big(n(n-1)s_{xx}^2\big)}} \;<\; -\,\mathrm{t}_{n-2,\gamma}\,. \tag{19}$$

  - Umgekehrt wird beim Test der Hypothese $H_0 : \alpha \leq \alpha_0$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \alpha > \alpha_0$) die Nullhypothese H_0 abgelehnt, falls

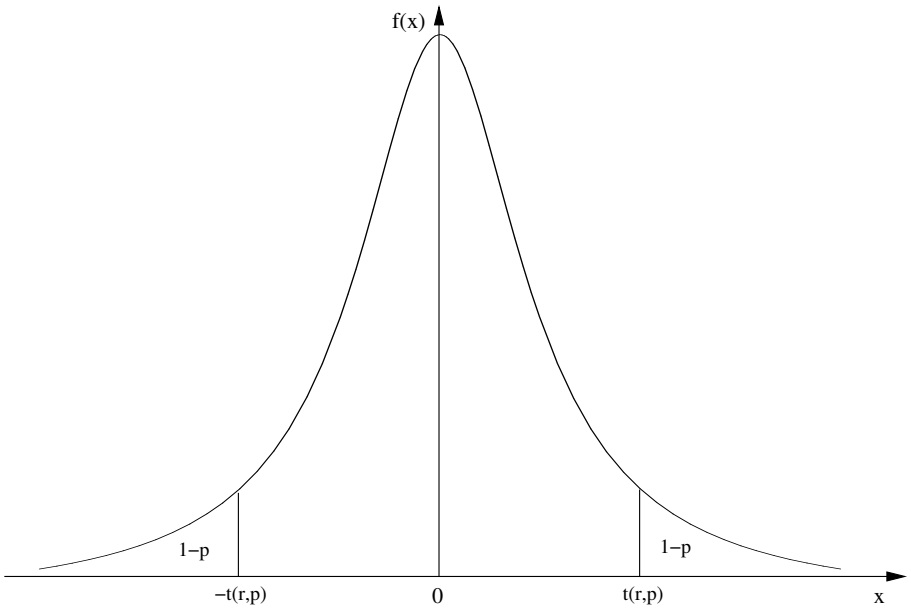
$$\frac{\widehat{\alpha} - \alpha_0}{S\sqrt{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big/\big(n(n-1)s_{xx}^2\big)}} \;>\; \mathrm{t}_{n-2,\gamma}\,. \tag{20}$$

  - Beim Testen von einseitigen Hypothesen über den Parameter β wird analog vorgegangen, wobei nun die Ungleichung (17) anstelle von (16) entsprechend modifiziert wird.

- Zur Erinnerung: Seien $r \in \mathbb{N}$ und $p \in (0,1)$ beliebige, jedoch fest vorgegebene Zahlen. Dann wird das *p-Quantil* $\mathrm{t}_{r,p}$ der t-Verteilung mit r Freiheitsgraden wie folgt definiert.
  - Sei $V_r \sim \mathrm{t}_r$ eine t-verteilte Zufallsvariable, und sei $f_{V_r} : \mathbb{R} \to \mathbb{R}$ die Dichte von V_r.
  - Dann ist $\mathrm{t}_{r,p}$ derjenige Schwellenwert, für den die Werte von V_r mit Wahrscheinlichkeit p kleiner (oder gleich) als $\mathrm{t}_{r,p}$ sind.
  - Mit anderen Worten: Für beliebige $r \in \mathbb{N}$ und $p \in (0.5,\, 1)$ genügt das *p-Quantil* $\mathrm{t}_{r,p}$ der folgenden *Quantilgleichung*:

$$\int_{-\infty}^{\mathrm{t}_{r,p}} f_{V_r}(v)\,dv = p\,, \tag{21}$$

  mit der graphischen Darstellung:



  - Das p-Quantil $\mathrm{t}_{r,p}$ kann durch die (numerische) Lösung der Integralgleichung (21) bestimmt werden, vgl. die Tabelle der Quantile der t-Verteilung in Abschnitt 4.

- Beachte
  - Die Quantile der t-Verteilung besitzen die folgende *Symmetrieeigenschaft*

$$\mathrm{t}_{r,p} = -\,\mathrm{t}_{r,1-p}\,, \qquad \forall\, p \in (0,1)\,. \tag{22}$$

  - Die in (16), (17) bzw. (18) betrachteten Ereignisse treten also jeweils mit der „Fehlerwahrscheinlichkeit" $1-\gamma$ ein, falls die betreffende Null-Hypothese richtig ist.

4. *Beispiel* (vgl. L.J. Kazmier (1999) *Wirtschaftsstatistik*. McGraw-Hill, S. 256 ff.)

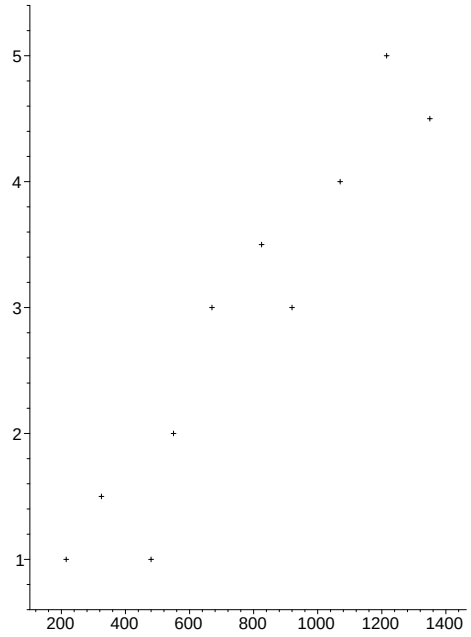
- Eine Speditionsfirma will anhand von 10 zufällig ausgewählten Lkw-Lieferungen untersuchen,
  - ob ein bzw. welcher Zusammenhang zwischen der Länge des Transportweges (in km) und der Lieferzeit (in Tagen) von der Abholbereitstellung bis zum Eintreffen der Lieferung beim Empfänger besteht.

  - Dabei wurden die folgenden Daten erhoben:

Nummer der Lieferung	1	2	3	4	5	6	7	8	9	10
Weglänge (in km)	825	215	1070	550	480	920	1350	325	670	1215
Lieferzeit (in Tagen)	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

    wobei die beobachteten Weglängen als Ausprägungen der Ausgangsvariablen und die zugehörigen Lieferzeiten als Zielwerte aufgefasst werden.

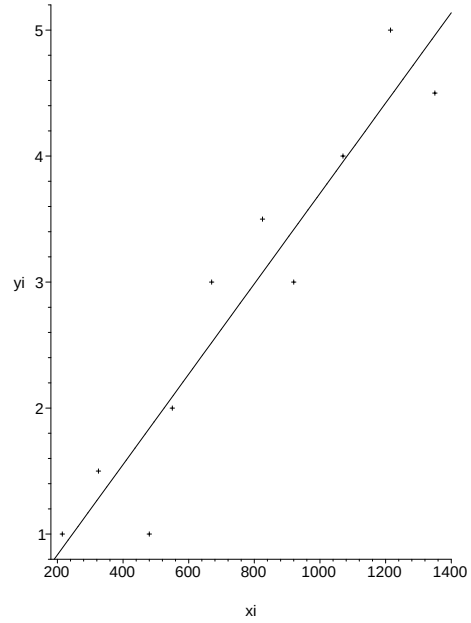
  - Aus diesen Daten ergibt sich das Streudiagramm:



- Außerdem ergeben sich gemäß (4) die folgenden Schätzwerte
  - für den Regressionskoeffizienten β bzw. die Regressionskonstante α:

$$\widehat{\beta} = \frac{s_{xy}^2}{s_{xx}^2} = 0.0036\,, \qquad \widehat{\alpha} = \overline{y}_{10} - \widehat{\beta}\overline{x}_{10} = 0.12\,.$$

  - Somit können wir die geschätzte Regressionsgerade $\widehat{y} = 0.12 + 0.0036x$ in das Streudiagramm einfügen:



- Für die geschätzten Lieferzeiten $\widehat{y}_i$ bzw. die Residuen $\widehat{\varepsilon}_i$ der beobachteten Weglängen x_i ergibt sich:

Nummer der Lieferung	1	2	3	4	5	6	7	8	9	10
beobachtete Lieferzeit	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0
geschätzte Lieferzeit	3.08	0.88	3.96	2.09	1.84	3.42	4.97	1.28	2.52	4.48
Residuum	0.42	0.12	0.04	-0.09	-0.84	-0.42	-0.47	0.22	0.48	0.52

- Als Schätzwert der Varianz σ^2 der Störgrößen $\varepsilon_1, \ldots, \varepsilon_{10}$ ergibt sich somit gemäß (5)

$$S^2 = \frac{1}{8} \sum_{i=1}^{10} \widehat{\varepsilon}_i^2 \approx 0.48^2 \,.$$

- Um zu prüfen, ob ein (signifikanter) Zusammenhang zwischen der Länge des Transportweges und der Lieferzeit besteht, wird nun
  - die Hypothese $H_0 : \beta = 0$ (gegen die Alternative $H_1 : \beta \neq 0$) zum Niveau $1 - \gamma = 0.05$ getestet werden.
  - Aus den beobachteten Daten ergibt sich, dass

$$\overline{x}_{10} = 762 \,, \quad \sum_{i=1}^{10} x_i^2 = 7104300 \,, \quad \sqrt{\sum_{i=1}^{10} x_i^2 - 10\,\overline{x}_{10}^2} = 1139.24$$

    und somit

$$\frac{|\,\widehat{\beta}\,|}{S/\sqrt{\sum_{i=1}^{10} x_i^2 - 10\,\overline{x}_{10}^2}} = \frac{0.0036}{0.48/1139.24} = \frac{0.0036}{0.0004} = 9.00 \,.$$

  - Andererseits gilt $t_{8,0.975} = 2.306$.
  - Gemäß (18) wird also die Hypothese $H_0 : \beta = 0$ abgelehnt, d.h., es besteht ein signifikanter Zusammenhang zwischen der Länge des Transportweges und der Lieferzeit.

5. *Beispiel* (Zusammenhang von Geburtsgewicht und Gewichtszunahme)

- Für die bereits in den Abschnitten 2.4.4 und 2.5.2 betrachteten Daten über das Geburtsgewicht von n=20 Säuglingen sowie deren Gewichtszunahme zwischen dem 70. und 100. Tag untersuchen wir nun noch die Frage,
  - ob auf der Basis dieses Datensatzes auf einen (statistisch signifikanten) linearen Zusammenhang von Geburtsgewicht und Gewichtszunahme geschlossen werden kann.
  - Mit anderen Worten: Wir prüfen die Hypothese $H_0 : \beta = 0$ (gegen die Alternative $H_1 : \beta \neq 0$) zum Niveau $1 - \gamma = 0.05$.
  - Für den Kleinste-Quadrate-Schätzer $\widehat{\beta}$ des Modellparameters β hatten wir in Abschnit 2.5.2 den Schätzwert $\widehat{\beta} = -1.242$ ermittelt.
  - Außerdem kann man zeigen, dass $S/\sqrt{19\,s_{xx}^2} = 0.249$. Für die in (18) betrachtete Testgröße ergibt sich somit der Wert

$$\frac{|\,\widehat{\beta}\,|}{S/\sqrt{19\,s_{xx}^2}} = 4.99 \,.$$

  - Andererseits gilt $t_{18,0.975} = 2.10 < 4.99$. Der beobachtete Wert der in (18) betrachteten Testgröße fällt somit in den Ablehnungsbereich der Null-Hypothese $H_0 : \beta = 0$.
  - Demzufolge ist der lineare Zusammenhang von Geburtsgewicht und Gewichtszunahme statistisch signifikant.

- Beachte
  - Die Hypothese, dass die Regressionsgerade durch den Nullpunkt verläuft (d.h. $H_0 : \alpha = 0$), wird bei diesem Beispiel ebenfalls verworfen,
  - denn für $1 - \gamma = 0.05$ gilt

$$\frac{\left|\,\widehat{\alpha}\,\right|}{S\sqrt{\Big(\sum_{i=1}^{20} x_i^2\Big)\Big/\big(20 \cdot 19\, s_{xx}^2\big)}} = \frac{5905}{796.2} = 7.42 > 2.10 = \mathrm{t}_{18,0.975}\,.$$

3.1.4 Konfidenzintervalle

1. *Konfidenzintervalle für die Modellparameter α und β*

- Zur Erinnerung: Bei der Konstruktion von *Konfidenzintervallen* für (reellwertige) Modellparameter, die auch *Vertrauensintervalle* genannt werden, geht man wie folgt vor.
  - Man betrachtet *zwei* Abbildungen φ_1, $\varphi_2 : \mathbb{R}^n \to \mathbb{R}$, die den beobachteten Daten $y_1, \ldots, y_n$, d.h. jeder Realisierung $(y_1, \ldots, y_n)$ der Zufallsstichprobe $(Y_1, \ldots, Y_n)$, die Schätzwerte $\varphi_1(y_1, \ldots, y_n)$ bzw. $\varphi_2(y_1, \ldots, y_n)$ zuordnen, und zwar so, dass

$$\varphi_1(y_1, \ldots, y_n) < \varphi_2(y_1, \ldots, y_n)\,.$$

  - Die zugehörigen Schätzer $\varphi_1(Y_1, \ldots, Y_n)$ und $\varphi_2(Y_1, \ldots, Y_n)$ ergeben dann ein *zufälliges Intervall* $\big(\varphi_1(Y_1, \ldots, Y_n),\, \varphi_2(Y_1, \ldots, Y_n)\big)$, das den unbekannten Modellparameter $\theta \in \mathbb{R}$ (zumindest) mit einer vorgegebenen *Überdeckungswahrscheinlichkeit* $\gamma \in (0, 1)$ enthalten soll.
  - Mit anderen Worten: Die Schätzer $\varphi_1(Y_1, \ldots, Y_n)$ und $\varphi_2(Y_1, \ldots, Y_n)$ liefern ein *Konfidenzintervall* für θ zum *Niveau* γ, falls

$$\mathbb{P}\big(\varphi_1(Y_1, \ldots, Y_n) < \theta < \varphi_2(Y_1, \ldots, Y_n)\big) \geq \gamma\,. \tag{23}$$

  - Die „Nichtüberdeckungswahrscheinlichkeit" $1 - \gamma$ wird *Irrtumswahrscheinlichkeit* genannt. Sie entspricht der in Abschnitt 3.1.3 betrachteten Fehlerwahrscheinlichkeit erster Art.
- Aus (16) und (17) ergeben sich ohne weiteres die folgenden Konfindenzintervalle zum Niveau $\gamma \in (0, 1)$ für die Modellparameter α und β. Und zwar gilt jeweils mit Wahrscheinlichkeit γ

$$\widehat{\alpha} - \mathrm{t}_{n-2,1-(1-\gamma)/2} S \sqrt{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big/\big(n(n-1)s_{xx}^2\big)} < \alpha < \widehat{\alpha} + \mathrm{t}_{n-2,1-(1-\gamma)/2} S \sqrt{\Big(\sum_{i=1}^{n} x_i^2\Big)\Big/\big(n(n-1)s_{xx}^2\big)} \tag{24}$$

und

$$\widehat{\beta} - \mathrm{t}_{n-2,1-(1-\gamma)/2} S\big/\sqrt{(n-1)s_{xx}^2} < \beta < \widehat{\beta} + \mathrm{t}_{n-2,1-(1-\gamma)/2} S\big/\sqrt{(n-1)s_{xx}^2}\,. \tag{25}$$

- Beispiele
  - i) Zusammenhang von Weglänge und Lieferzeit
    - Für die in Abschnitt 3.1.3 betrachteten Daten über Weglängen und Lieferzeiten von 10 zufällig ausgewählten Lkw-Lieferungen hatten wir gezeigt, dass der Zusammenhang von Weglänge und Lieferzeit statistisch signifikant ist.
    - Außerdem hatten wir gezeigt, dass

$$\frac{S}{\sqrt{\sum_{i=1}^{10} x_i^2 - 10\,\overline{x}_{10}^2}} = \frac{0.48}{1139.24} = 0.0004\,.$$

– Aus (25) ergibt sich somit das folgende Konfidenzintervall für β zum Niveau $\gamma = 0.95$:

$$(0.0036 \mp 2.306 \cdot 0.0004) = (0.0036 \mp 0.0009) = (0.0027,\, 0.0045)\,.$$

ii) Zusammenhang von Geburtsgewicht und Gewichtszunahme
– Wir betrachten erneut die Daten über die Merkmale „Geburtsgewicht" von Säuglingen sowie deren „Gewichtszunahme" zwischen dem 70. und 100. Tag, für die in Abschnitt 3.1.3 ein statistisch signifikanter Zusammenhang zwischen diesen beiden Merkmalen ermittelt wurde.
– Aus (25) ergibt sich das folgende Konfidenzintervall für β zum Niveau $\gamma = 0.95$:

$$(-1.262 \mp 2.10 \cdot 0.249) = (-1.765,\, -0.719)\,.$$

2. *Konfidenzintervall für den erwarteten Zielwert* $\alpha + \beta x_0$

- Auf ähnliche Weise kann man auch ein Konfidenzintervall zum Niveau $\gamma \in (0,1)$ für den erwarteten Zielwert $\alpha + \beta x_0$ herleiten, der einem vorgegebenen Ausgangswert $x_0 \in \mathbb{R}$ entspricht.
  – Von besonderem Interesse ist dabei natürlich der Fall, dass $x_0 \notin \{x_1, \ldots, x_n\}$, d.h., wenn an der Stelle x_0 *keine* Daten erhoben werden.
  – Sei also $x_0 \in \mathbb{R}$ eine (geeignet gewählte) reelle Zahl mit

$$\min\{x_1, \ldots, x_n\} < x_0 < \max\{x_1, \ldots, x_n\}\,. \tag{26}$$

  – Dann ist durch den Ansatz $\widehat{\alpha} + \widehat{\beta} x_0$ ein erwartungstreuer Schätzer für $\alpha + \beta x_0$ gegeben mit

$$\widehat{\alpha} + \widehat{\beta} x_0 \sim \mathrm{N}\left(\alpha + \beta x_0\,,\, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{(n-1)s_{xx}^2}\right)\right). \tag{27}$$

  – Mit Wahrscheinlichkeit γ gilt somit, dass

$$\widehat{\alpha} + \widehat{\beta} x_0 - \tau < \alpha + \beta x_0 < \widehat{\alpha} + \widehat{\beta} x_0 + \tau\,, \tag{28}$$

  wobei

$$\tau = \mathrm{t}_{n-2,1-(1-\gamma)/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{(n-1)s_{xx}^2}}\,.$$

- Beachte
  – Bei gegebenen Daten $(x_1, y_1), \ldots, (x_n, y_n)$ ist die Länge des in (28) hergeleiteten Konfidenzintervalls für $\alpha + \beta x_0$ eine monoton nichtfallende Funktion des Abstandes $|x_0 - \overline{x}_n|$, die ihr Minimum im Punkt $x_0 = \overline{x}_n$ annimmt.
  – Die in (28) hergeleitete Intervallschätzung für $\alpha + \beta x_0$ ist also dann am genauesten, wenn der Wert x_0 im „Zentrum" der (Ausgangs-) Werte $x_1, \ldots, x_n$ liegt.
  – Andererseits können Ergebnisse entstehen, die offenkundig falsch sind, falls x_0 eine der beiden Ungleichungen in (26) nicht erfüllt.
  – So würde sich aus den Daten über die Merkmale Geburtsgewicht von Säuglingen und deren Gewichtszunahme beispielsweise für das „Geburtsgewicht" $x_0 = 0$ eine geschätzte Gewichtszunahme von $\widehat{\alpha} + \widehat{\beta} x_0 = 5905$ ergeben.
- Beispiel
  – Aus den in Abschnitt 3.1.3 betrachteten Daten über Weglängen und Lieferzeiten von 10 zufällig ausgewählten Lkw-Lieferungen ergibt sich für die Weglänge $x_0 = 1000$, dass

$$\widehat{\alpha} + \widehat{\beta} x_0 = 0.11 + 0.0036 \cdot 1000 = 3.71\,.$$

– Außerdem gilt

$$S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{(n-1)s_{xx}^2}} = S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}}$$
$$= 0.46\sqrt{\frac{1}{10} + \frac{(1000-762)^2}{7104300 - (7620)^2/10}} \approx 0.17\,.$$

– Aus (28) ergibt sich somit das folgende Konfidenzintervall zum Niveau $\gamma = 0.95$ für die erwartete Lieferzeit $\alpha + \beta \cdot 1000$ bei der (angenommenen) Weglänge von $x_0 = 1000$:

$$(3.71 \mp 2.306 \cdot 0.17) = (3.71 \mp 0.39) = (3.32,\ 4.10)\,.$$

3.1.5 Prognose von Zielwerten

- In (28) ist ein zufälliges Intervall gegeben, in dem
  - der (unbekannte) Erwartungswert $\mathbb{E}\,Y_0 = \alpha + \beta x_0$ der Zufallsvariable $Y_0 = \alpha + \beta x_0 + \varepsilon_0$ mit der (vorgegebenen) Wahrscheinlichkeit γ liegt, wobei
  - die Störgröße ε_0 normalverteilt und unabhängig von den Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ ist; $\varepsilon_0 \sim \mathrm{N}(0, \sigma^2)$.
- Wir bestimmen nun ein zufälliges Intervall, in dem *nicht* der Erwartungswert $\mathbb{E}\,Y_0$, sondern die Zielgröße Y_0 selbst mit der Wahrscheinlichkeit γ liegt.
  - Dabei seien $L, U : \Omega \to \mathbb{R}$ zwei Zufallsvariablen, so dass $\mathbb{P}(L \le U) = 1$ und $\mathbb{P}(L < Y_0 < U) \ge \gamma$ gilt.
  - Das zufällige Intervall (L, U) wird dann *Prognoseintervall* für die Zielvariable Y_0 zum Niveau γ genannt.
- Man kann zeigen, dass mit Wahrscheinlichkeit γ

$$\widehat{\alpha} + \widehat{\beta} x_0 - \tau' < Y_0 < \widehat{\alpha} + \widehat{\beta} x_0 + \tau' \tag{29}$$

gilt, wobei

$$\tau' = \mathrm{t}_{n-2,1-(1-\gamma)/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{(n-1)s_{xx}^2}}\,.$$

- Beispiel
  - Aus den in Abschnitt 2.4.4 betrachteten Daten über die Merkmale „Geburtsgewicht" von Säuglingen sowie deren „Gewichtszunahme" zwischen dem 70. und 100. Tag ergibt sich, dass

$$S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x}_n)^2}{(n-1)s_{xx}^2}} = 590.7\,.$$

  – Aus (29) ergibt sich somit das folgende Prognoseintervall zum Niveau $\gamma = 0.95$ für die Gewichtszunahme $Y_0 = \alpha + \beta \cdot 2700 + \varepsilon_0$ bei dem (angenommenen) Geburtsgewicht von $x_0 = 2700$:

$$(2552 \mp 2.10 \cdot 590.7) = (1312,\ 3792)\,.$$

3.1.6 Simultane Konfidenzbereiche; Konfidenzbänder

1. *Simultane Konfidenzbereiche*

- Auf ähnliche Weise wie in Abschnitt 3.1.4 kann man sogenannte *simultane Konfidenzbereiche* zum Niveau $\gamma \in (0,1)$
  - *gleichzeitig für mehrere* erwartete Zielwerte $\alpha + \beta x_{0i}$, $i = 1, \ldots, m$ angeben, die vorgegebenen (Ausgangs-) Werten $x_{01}, \ldots, x_{0m} \in \mathbb{R}$ entsprechen.
  - Dabei ist erneut der Fall $x_{01}, \ldots, x_{0m} \notin \{x_1, \ldots, x_n\}$ von besonderem Interesse, d.h., wenn an den Stellen $x_{01}, \ldots, x_{0m}$ keine Daten erhoben werden.
  - Man kann nämlich zeigen, dass die Wahrscheinlichkeit mindestens gleich γ ist, dass

$$\widehat{\alpha} + \widehat{\beta} x_{0i} - \tau_i'' < \alpha + \beta x_{0i} < \widehat{\alpha} + \widehat{\beta} x_{0i} + \tau_i'' \tag{30}$$

    gleichzeitig für jedes $i = 1, \ldots, m$ gilt, wobei

$$\tau_i'' = \mathrm{t}_{n-2,1-(1-\gamma)/(2m)} S \sqrt{\frac{1}{n} + \frac{(x_{0i} - \overline{x}_n)^2}{(n-1)s_{xx}^2}}\,.$$

- Das kartesische Produkt $A_1 \times \ldots \times A_m$ der m Intervalle $A_i = \left(\widehat{\alpha} + \widehat{\beta} x_{0i} - \tau_i'', \widehat{\alpha} + \widehat{\beta} x_{0i} + \tau_i''\right)$ in (30) heißt *simultaner Konfidenzbereich* für den Vektor $(\alpha + \beta x_{01}, \ldots, \alpha + \beta x_{0m})$.

2. *Konfidenzbänder*

- Wir gehen nun noch einen Schritt weiter als in (30) und fragen,
  - ob es eine Zahl $a_\gamma > 0$ gibt, so dass die Wahrscheinlichkeit mindestens gleich γ ist, dass gleichzeitig für jedes $z \in \mathbb{R}$

$$\widehat{\alpha} + \widehat{\beta} z - a_\gamma S \sqrt{\frac{1}{n} + \frac{(z - \overline{x}_n)^2}{(n-1)s_{xx}^2}} < \alpha + \beta z < \widehat{\alpha} + \widehat{\beta} z + a_\gamma S \sqrt{\frac{1}{n} + \frac{(z - \overline{x}_n)^2}{(n-1)s_{xx}^2}}\,. \tag{31}$$

  - Die Menge $B_\gamma \subset \mathbb{R}^2$ mit

$$B_\gamma = \left\{ (z_1, z_2) \in \mathbb{R}^2 : \widehat{\alpha} + \widehat{\beta} z_1 - a_\gamma S \sqrt{\frac{1}{n} + \frac{(z_1 - \overline{x}_n)^2}{(n-1)s_{xx}^2}} < z_2 < \widehat{\alpha} + \widehat{\beta} z_1 + a_\gamma S \sqrt{\frac{1}{n} + \frac{(z_1 - \overline{x}_n)^2}{(n-1)s_{xx}^2}} \right\} \tag{32}$$

    heißt dann *Konfidenzband* zum Niveau γ für die Regressionsgerade $y = \alpha + \beta x$.

- Bei der Lösung dieser Fragestellung ist die F-Verteilung nützlich, die ebenfalls eine Klasse von *statistischen Prüfverteilungen* bildet.
  - Zur Erinnerung: Seien $r, s \geq 1$ beliebige natürliche Zahlen, und seien U_r, $U_s' : \Omega \to (0, \infty)$ unabhängige χ^2-verteilte Zufallsvariablen mit $U_r \sim \chi_r^2$ und $U_s' \sim \chi_s^2$.
  - Man sagt dann, dass die Zufallsvariable

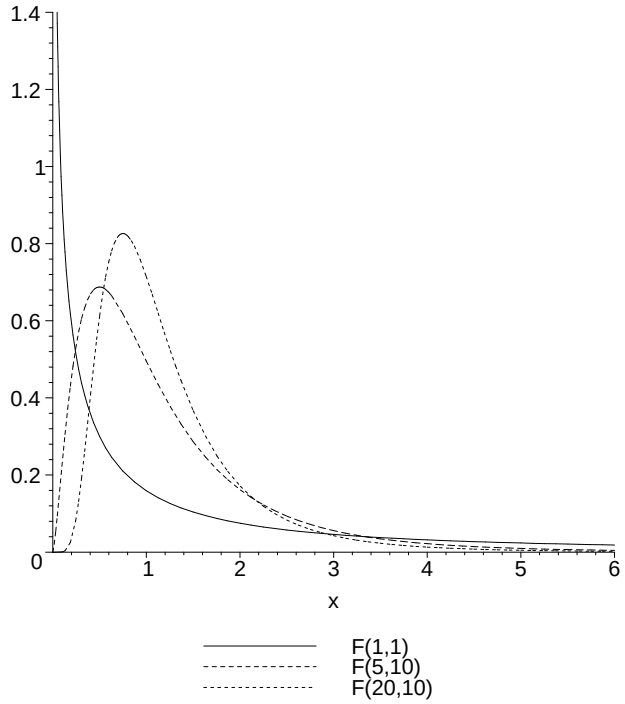
$$W_{r,s} = \frac{U_r / r}{U_s' / s}$$

    F-*verteilt* ist mit (r, s) Freiheitsgraden. (Schreibweise: $W_{r,s} \sim \mathrm{F}_{r,s}$)

– Die Dichte von $W_{r,s}$ ist gegeben durch

$$f_{W_{r,s}}(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{r+s}{2}\right)}{\Gamma\left(\dfrac{r}{2}\right)\Gamma\left(\dfrac{s}{2}\right)} \left(\dfrac{r}{s}\right)^{r/2} \dfrac{x^{(r/2)-1}}{\left(1+\dfrac{r}{s}\,x\right)^{(r+s)/2}}\,, & \text{falls } x>0, \\ 0 & \text{sonst,} \end{cases} \tag{33}$$

mit der graphischen Darstellung:



- Man kann zeigen, dass durch die in (32) definierte Menge $B_\gamma \subset \mathbb{R}^2$ ein *Konfidenzband* zum Niveau γ für die Regressionsgerade $y = \alpha + \beta x$ gegeben ist, wenn a_γ wie folgt gewählt wird:

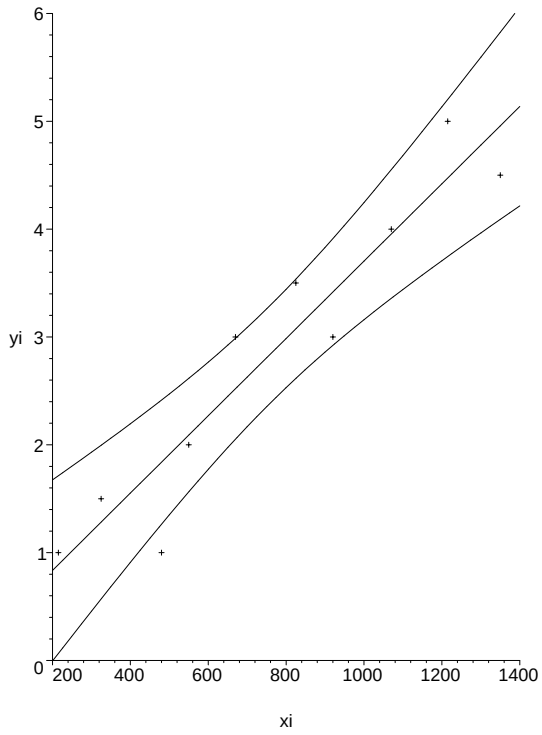
$$a_\gamma = \sqrt{2\,\mathrm{F}_{2,n-2,\gamma}}\,.$$

- Beachte
  - Es ist klar, dass $\mathrm{t}_{n-2,1-(1-\gamma)/(2m)} > \sqrt{2\,\mathrm{F}_{2,n-2,\gamma}}$ für jedes hinreichend große m gilt .
  - Hieraus folgt, dass der simultane Konfidenzbereich, der in (30) betrachtet wurde, für große m *größer* ist als der simultane Konfidenzbereich, der sich aus dem in (31) betrachteten Konfidenzband ergibt.
  - Auf den ersten Blick scheint dies ein Widerspruch zu sein, weil in (31) die Überdeckungseigenschaft für *alle* $x \in \mathbb{R}$ gefordert wird, während diese Eigenschaft in (30) nur für *endlich viele* Ausgangswerte $x_{01}, \ldots, x_{0m}$ betrachtet wird.
  - Der Grund, dass (31) für große m zu kleineren (d.h. besseren) simultanen Konfidenzbereichen führt, besteht darin, dass bei der Herleitung von (30) die sogenannte *Bonferroni-Ungleichung* der Wahrscheinlichkeitsrechnung verwendet wird, die für große m nur eine sehr ungenaue untere Schranke für die Wahrscheinlichkeit $\mathbb{P}\left(\bigcap_{i=1}^m A_i\right)$ liefert, wobei

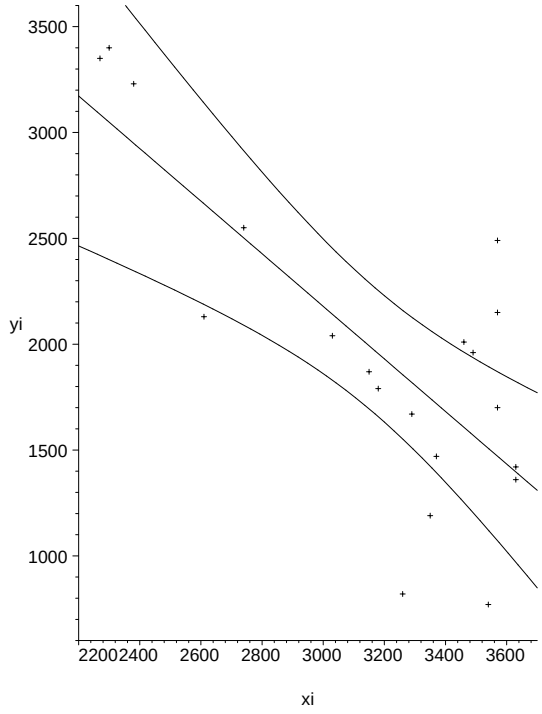
$$A_i = \left(\widehat{\alpha} + \widehat{\beta} x_{0i} - \tau'',\ \widehat{\alpha} + \widehat{\beta} x_{0i} + \tau''\right).$$

- Beispiele

– Aus den Daten über Weglängen und Lieferzeiten ergibt sich das folgende *Konfidenzband* zum Niveau $\gamma = 0.95$, das an der Stelle $\overline{x}_n = 762$ die kleinste Breite aufweist:



– Auf ähnliche Weise ergibt sich aus den Daten über die beiden Merkmale „Geburtsgewicht" und „Gewichtszunahme" das folgende *Konfidenzband* zum Niveau $\gamma = 0.95$, das an der Stelle $\overline{x}_n = 3170$ die kleinste Breite aufweist:



3.2 Einfaktorielle Varianzanalyse

3.2.1 Modellannahmen

- Wir nehmen nun an, dass die Zielvariablen $Y_1, \dots, Y_n$ durch das folgende stochastische Modell gegeben sind.
  - Und zwar zerlegen wir die Zufallsstichprobe $(Y_1, \dots, Y_n)$ in k Teilstichproben $(Y_{ij},\, j = 1, \dots, n_i)$, wobei angenommen wird, dass $n_i > 1$ für jedes $i = 1, \dots, k$ und $\sum_{i=1}^{k} n_i = n$.
  - Außerdem setzen wir voraus, dass die Stichprobenvariablen, die zu einundderselben Teilstichprobe gehören, jeweils den gleichen Erwartungswert haben.
- Mit anderen Worten: Wir nehmen an, dass

$$Y_{ij} = \theta_i + \varepsilon_{ij}\,, \qquad \forall\, i = 1, \dots, k,\ j = 1, \dots, n_i \tag{34}$$

  gilt, wobei
  - $\theta_1, \dots, \theta_k \in \mathbb{R}$ (unbekannte) Modellparameter sind,
  - die Störgrößen $\varepsilon_{ij} : \Omega \to \mathbb{R}$ unabhängig sind mit

$$\mathbb{E}\,\varepsilon_{ij} = 0\,, \qquad \operatorname{Var}\varepsilon_{ij} = \sigma_i^2\,, \qquad \forall\, i = 1, \dots, k,\ j = 1, \dots, n_i \tag{35}$$

  - und die Varianzen $\sigma_1^2, \dots, \sigma_k^2 > 0$ ebenfalls unbekannte Modellparameter sind.
- Beachte
  - Die Nummern $i = 1, \dots, k$ der Teilstichproben $(Y_{ij},\, j = 1, \dots, n_i)$ werden als *Stufen eines Einflussfaktors* gedeutet.
  - Von besonderem Interesse ist der Fall, dass $\sigma_1^2 = \dots = \sigma_k^2 = \sigma^2 > 0$. In diesem Fall spricht man von *homoskedastischen Störgrößen*, ansonsten von *heteroskedastischen Störgrößen*.
  - Die obengemachten Modellannahmen bedeuten insbesondere, dass die beobachteten Werte $y_1, \dots, y_n$ der Zielvariablen $Y_1, \dots, Y_n$ wie folgt tabellarisch strukturiert werden können:

Stufe	1	2	3	...	k
	y_{11}	y_{21}	y_{31}	$\cdots$	y_{k1}
	y_{12}	y_{22}	y_{32}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$	$\vdots$	$\cdots$	y_{k3}
			y_{3n_3}		$\vdots$
	y_{1n_1}				
		y_{2n_2}			y_{kn_k}

  - Der Begriff „Varianzanalyse" bedeutet *nicht*, dass die Varianzen der Stichprobenvariablen Y_{ij} untersucht werden, sondern es handelt sich um die Analyse der Variabilität der Erwartungswerte $\theta_1, \dots, \theta_k$.
  - In der englischsprachigen Literatur ist die Abkürzung ANOVA üblich (ANOVA = analysis of variance).

- Beispiel (vgl. L.J. Kazmier (1999) *Wirtschaftsstatistik*. McGraw-Hill, S. 235 ff.)
  - Eine Gruppe von 12 Personen führte das folgende Copmuter-Experiment durch.

- Dabei standen 3 verschiedene Tastaturen zur Verfügung, um einen vorgegebenen Text jeweils eine Minute lang in einen PC einzugeben.
- Es nutzen 5 Personen die erste Tastatur, 3 Personen die zweite Tastatur und 4 Personen die dritte Tastatur, wobei jeweils die folgenden Anzahlen von Wörtern je Minute in den PC eingegeben wurden:

Tastatur	1	2	3
	79	74	81
	83	85	65
	62	72	79
	51		55
	77		

- Der Einflussfaktor „Tastatur" hat also in diesem Fall $k = 3$ „Stufen" mit $n_1 = 5$, $n_2 = 3$ bzw. $n_3 = 4$.
- Die Größen θ_1, θ_2 bzw. θ_3 sind dabei jeweils die erwarteten Anzahlen von Wörtern, die mit der ersten, zweiten bzw. dritten Tastatur je Minute in den PC eingegeben werden.

3.2.2 Klassische ANOVA-Nullhypothese; Kontraste

- Die klassische ANOVA-Nullhypothese besteht darin zu prüfen,
  - ob die Erwartungswerte $\theta_1, \ldots, \theta_k$ der Stichprobenvariablen Y_{ij} gleich sind, also nicht von der Nummer i der betrachteten Teilstichprobe abhängen,
  - d.h., wir prüfen die Hypothese, ob die Stufen des Einflussfaktors *keine* statistische Signifikanz haben.
- Beachte
  - Beim Testen der ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ ist es nützlich zu beachten, dass diese Hypothese mit Hilfe von sogenannten Kontraste ausgedrückt werden kann.
  - Unter einem *Kontrast* versteht man dabei die folgenden Abbildung:

$$\mathbf{t} \to \sum_{i=1}^{k} a_i t_i \qquad \text{mit } \textstyle\sum_{i=1}^{k} a_i = 0,$$

    wobei $\mathbf{t} = (t_1, \ldots, t_k) \in \mathbb{R}^k$ ein beliebiger Vektor von Variablen und $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ ein Vektor von (bekannten) Konstanten ist.
- Man kann nämlich zeigen, dass die Gültigkeit von $\theta_1 = \ldots = \theta_k$ gleichbedeutend damit ist, dass $\sum_{i=1}^{k} a_i \theta_i = 0$ für jedes $\mathbf{a} \in \mathcal{A}$, wobei

$$\mathcal{A} = \Big\{ \mathbf{a} = (a_1, \ldots, a_k) : \ \mathbf{a} \neq \mathbf{o}, \ \sum_{i=1}^{k} a_i = 0 \Big\}$$

  und $\mathbf{o} = (0, \ldots, 0)$ den Nullvektor bezeichnet.
- Die ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ kann somit in der folgenden Form geschrieben werden:

$$H_0 : \ \sum_{i=1}^{k} a_i \theta_i = 0 \qquad \text{für jedes } \mathbf{a} \in \mathcal{A}, \tag{36}$$

  d.h., jeder Kontrast nimmt im Punkt $\mathbf{t} = (\theta_1, \ldots, \theta_k)$ den Wert 0 an.

- Um einen Test zur Verifizierung der ANOVA-Nullhypothese (36) zu konstruieren, wird nun zunächst
  - die Hypothese $H_0 : \sum_{i=1}^k a_i\theta_i = 0$ für *einen* vorgegebenen Kontrast $\mathbf{a} \in \mathcal{A}$ betrachtet,
  - d.h., es wird zunächst ein hypothetischer Wert für die *Linearkombination* $\sum_{i=1}^k a_i\theta_i$ der Erwartungswerte $\theta_1, \ldots, \theta_k$ getestet.

3.2.3 t-Test und Konfidenzintervall für Linearkombinationen von Erwartungswerten

- Sei $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ ein beliebiger Vektor.
  - Um einen Test zur Verifizierung der Hypothese $H_0 : \sum_{i=1}^k a_i\theta_i = 0$ bzw. ein Konfidenzintervall für die Linearkombination $\sum_{i=1}^k a_i\theta_i$ der Erwartungswerte $\theta_1, \ldots, \theta_k$ bzw. herzuleiten,
  - betrachten wir die folgenden Summen bzw. Mittelwerte der Stichprobenvariablen:

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}\,, \qquad Y_{\cdot j} = \sum_{i=1}^{k} Y_{ij} \tag{37}$$

  bzw.

$$\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}\,, \qquad \overline{Y}_{\cdot j} = \frac{1}{k}\sum_{i=1}^{k} Y_{ij}\,, \qquad \overline{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\,. \tag{38}$$

- Dabei werden wir von nun an ausschließlich homoskedastische Störgrößen betrachten, die normalverteilt sind,
  - d.h., wir setzen voraus, dass $\sigma_1^2 = \ldots = \sigma_k^2 = \sigma^2 > 0$
  - und dass $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ für alle $i = 1, \ldots, k$ und $j = 1, \ldots, n_i$.
- Man kann dann die Gültigkeit der folgenden *Verteilungs- und Unabhängigkeitseigenschaften* zeigen:
  1. Für jedes $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ gilt

$$\sum_{i=1}^{k} a_i\overline{Y}_{i\cdot} \sim \mathrm{N}\Big(\sum_{i=1}^{k} a_i\theta_i,\, \sigma^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}\Big)\,, \qquad \frac{(n-k)S_p^2}{\sigma^2} \sim \chi^2_{n-k}\,, \tag{39}$$

  2. die Zufallsvariablen $\sum_{i=1}^k a_i\overline{Y}_{i\cdot}$ und S_p^2 sind unabhängig, wobei

$$S_p^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij} - \overline{Y}_{i\cdot}\big)^2 \tag{40}$$

  die sogenannte *gepoolte Stichprobenvarianz* (pooled sample variance) ist.

- Hieraus und aus der Definition der t-Verteilung (vgl. Abschnitt 3.1.3) folgt, dass für jedes $\mathbf{a} \neq \mathbf{o}$

$$\frac{\sum\limits_{i=1}^{k} a_i\overline{Y}_{i\cdot} - \sum\limits_{i=1}^{k} a_i\theta_i}{\sqrt{S_p^2\sum\limits_{i=1}^{k}\frac{a_i^2}{n_i}}} \;\sim\; \mathrm{t}_{n-k}\,. \tag{41}$$

- Beachte

- Aus (41) ergibt sich ein Test der Hypothese $H_0 : \sum_{i=1}^k a_i\theta_i = 0$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \sum_{i=1}^k a_i\theta_i \neq 0$).
- Dabei wird die Nullhypothese H_0 abgelehnt, falls

$$\left| \frac{\sum\limits_{i=1}^k a_i \overline{Y}_{i\cdot}}{\sqrt{S_p^2 \sum\limits_{i=1}^k \frac{a_i^2}{n_i}}} \right| > t_{n-k,1-(1-\gamma)/2}\,. \tag{42}$$

- Außerdem ergibt sich aus (41) ohne weiteres das folgende Konfidenzintervall zum Niveau $\gamma \in (0,1)$ für $\sum_{i=1}^k a_i\theta_i$, denn mit Wahrscheinlichkeit γ gilt

$$\sum_{i=1}^k a_i\overline{Y}_{i\cdot} - t_{n-k,1-(1-\gamma)/2}\sqrt{S_p^2\sum_{i=1}^k\frac{a_i^2}{n_i}} < \sum_{i=1}^k a_i\theta_i < \sum_{i=1}^k a_i\overline{Y}_{i\cdot} + t_{n-k,1-(1-\gamma)/2}\sqrt{S_p^2\sum_{i=1}^k\frac{a_i^2}{n_i}} \tag{43}$$

- Durch eine geeignete Wahl des Vektors $\mathbf{a} = (a_1,\ldots,a_k) \in \mathbb{R}$ ergeben sich aus (42) *Tests spezifischer Eigenschaften* der Erwartungswerte $\theta_1,\ldots,\theta_k$.

  1. Für $\mathbf{a} = (1,-1,0,\ldots,0)$ ergibt sich aus (42) ein
     - Test der Hypothese $H_0 : \theta_1 = \theta_2$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \theta_1 \neq \theta_2$).
     - Dabei wird die Nullhypothese H_0 abgelehnt, falls

$$\left| \frac{\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \right| > t_{n-k,1-(1-\gamma)/2}\,. \tag{44}$$

  2. Für $\mathbf{a} = (1,-1/2,-1/2,0,\ldots,0)$ ergibt sich aus (42) ein
     - Test der Hypothese $H_0 : \theta_1 = (\theta_2+\theta_3)/2$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \theta_1 \neq (\theta_2+\theta_3)/2$).
     - Dabei wird die Nullhypothese H_0 abgelehnt, falls

$$\left| \frac{\overline{Y}_{1\cdot} - \frac{1}{2}\overline{Y}_{2\cdot} - \frac{1}{2}\overline{Y}_{3\cdot}}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{4n_2}+\frac{1}{4n_3}\right)}} \right| > t_{n-k,1-(1-\gamma)/2}\,. \tag{45}$$

- Beispiel
  - Für das in Abschnitt 3.2.1 betrachtete Beispiel der Eingabe von Wörtern über drei verschiedene Tastaturen wollen wir nun prüfen, ob sich die beiden ersten Tastaturen hinsichtlich der jeweils erwarteten Anzahlen der eingegebenen Wörter je Minute signifikant voneinander unterscheiden ($1-\gamma = 0.05$).
  - Mit anderen Worten: Wir testen die Hypothese $H_0 : \theta_1 = \theta_2$ zum Niveau $1-\gamma = 0.05$.
  - Hierfür berechnen wir zunächst den beobachteten Wert der Testgröße in (44), wobei sich für die Schätzer $\overline{Y}_{1\cdot}$, $\overline{Y}_{2\cdot}$, $\overline{Y}_{3\cdot}$ und S_p^2 die folgenden Werte ergeben:

$$\overline{Y}_{1\cdot} = 70.4\,, \qquad \overline{Y}_{2\cdot} = 77.0\,, \qquad \overline{Y}_{3\cdot} = 70.0$$

    und

$$S_p^2 = \frac{1}{9}\Big(8.6^2 + 12.6^2 + 8.4^2 + 19.4^2 + 6.6^2 + 3^2 + 8^2 + 5^2 + 11^2 + 5^2 + 9^2 + 15^2\Big) = 141.47\,.$$

– Hieraus folgt, dass

$$\left|\frac{\overline{Y}_{1\cdot}-\overline{Y}_{2\cdot}}{\sqrt{S_p^2\left(\frac{1}{5}+\frac{1}{3}\right)}}\right| = \left|\frac{70.4-77.0}{\sqrt{141.47\cdot 0.53}}\right| = 0.76 < 2.262 = t_{9,\,0.975}\,.$$

– Die Hypothese $H_0 : \theta_1 = \theta_2$, dass die Erwartungswerte θ_1 und θ_2 für die ersten beiden Tastaturen gleich sind, wird also *nicht* verworfen.

3.2.4 F-Test der ANOVA-Nullhypothese; Quadratsummenzerlegung

- Zur Erinnerung:
  - Die ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ ist äquivalent mit der Hypothese $H_0 : \boldsymbol{\theta} \in \bigcap_{\mathbf{a}\in\mathcal{A}} \Theta_{\mathbf{a}}$, wobei $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ und

$$\Theta_{\mathbf{a}} = \left\{\boldsymbol{\theta}' = (\theta_1', \ldots, \theta_k') : \sum_{i=1}^{k} a_i\theta_i' = 0\right\}.$$

  - Für jeden (einzelnen) Kontrast $\mathbf{a} \in \mathcal{A}$ hatten wir in Abschnitt 3.2.3 die Hypothese $H_{0,\mathbf{a}} : \boldsymbol{\theta} \in \Theta_{\mathbf{a}}$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_{1,\mathbf{a}} : \boldsymbol{\theta} \notin \Theta_{\mathbf{a}}$) getestet.
  - Dabei wurde die Nullhypothese $H_{0,\mathbf{a}} : \boldsymbol{\theta} \in \Theta_{\mathbf{a}}$ abgelehnt, falls die Testgröße

$$T_{\mathbf{a}} = \left|\frac{\sum_{i=1}^{k} a_i\overline{Y}_{i\cdot}}{\sqrt{S_p^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}}}\right| \tag{46}$$

    einen gewissen Schwellenwert c_γ überschreitet, der nur von γ (jedoch *nicht* von $\mathbf{a}$) abhängt.

- Dies führt zu folgendem *Ansatz*, um die klassische ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$, d.h. die Hypothese $H_0 : \boldsymbol{\theta} \in \bigcap_{\mathbf{a}\in\mathcal{A}} \Theta_{\mathbf{a}}$ (gegen die Alternative $H_1 : \boldsymbol{\theta} \notin \bigcap_{\mathbf{a}\in\mathcal{A}} \Theta_{\mathbf{a}}$) zu testen:
  - Weil H_0 genau dann abgelehnt wird, wenn die Hypothese $H_{0,\mathbf{a}} : \boldsymbol{\theta} \in \Theta_{\mathbf{a}}$ für ein $\mathbf{a} \in \mathcal{A}$ abgelehnt wird und
  - weil somit der Ablehnungsbereich der ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ die *Vereinigung* der Ablehnungsbereiche der Hypothesen $H_{0,\mathbf{a}}$ ist,
  - ist es naheliegend, die Hypothese H_0 genau dann abzulehnen, wenn

$$\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2 > c_\gamma^2\,, \tag{47}$$

    wobei der Schwellenwert c_γ^2 so gewählt wird, dass die Wahrscheinlichkeit des in (47) betrachteten Ereignisses unter H_0 nicht größer als $1-\gamma$ ist.

- Man kann nun zeigen, dass unter H_0 folgendes gilt:

$$\frac{1}{k-1}\sup_{\mathbf{a}\in\mathcal{A}} T_{\mathbf{a}}^2 = \frac{\sum_{i=1}^{k} n_i\left(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\right)^2}{(k-1)S_p^2} \sim \mathrm{F}_{k-1,n-k}\,, \tag{48}$$

wobei $\mathrm{F}_{k-1,n-k}$ die bereits in Abschnitt 3.1.6 erwähnte F-Verteilung mit $(k-1, n-k)$-Freiheitsgraden bezeichnet.

- Die ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ wird somit abgelehnt, falls

$$\frac{\sum_{i=1}^{k} n_i \big(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\big)^2}{(k-1)S_p^2} > \mathrm{F}_{k-1,n-k,\gamma}\,. \tag{49}$$

- Beachte
  - Durch die folgende *Quadratsummenzerlegung* ergibt sich eine anschauliche Deutung von Zähler und Nenner der in (49) betrachteten Testgröße.
  - Man kann nämlich zeigen, dass

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij} - \overline{Y}_{\cdot\cdot}\big)^2 = \sum_{i=1}^{k} n_i\big(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\big)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij} - \overline{Y}_{i\cdot}\big)^2\,. \tag{50}$$

  - Die Doppelsumme auf der linken Seite von (50) kann als eine Maßzahl für die (Gesamt-) *Variabilität der Stichprobenvariablen* $\{Y_{ij},\, i = 1,\ldots,k,\, j = 1,\ldots,n_i\}$ aufgefasst werden.
  - Die erste Summe auf der rechten Seite von (50) ist eine Maßzahl für die Variabilität *zwischen* den Stufen des Einflussfaktors, während die Doppelsumme auf der rechten Seite von (50) eine Maßzahl für die Variabilität *innerhalb* der Stufen des Einflussfaktors ist.
  - Wegen

$$S_p^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\big(Y_{ij} - \overline{Y}_{i\cdot}\big)^2$$

    ist die in (49) betrachtete Testgröße also proportional zu dem Quotienten, der aus der Variabilität zwischen den Stufen des Einflussfaktors und der Variabilität innerhalb der Stufen gebildet wird.
  - Die durch (49) gegebene Entscheidungsregel bedeutet somit, dass die ANOVA-Nullhypothese $H_0 : \theta_1 = \ldots = \theta_k$ abgelehnt wird, falls die Variabilität zwischen den Stufen signifikant größer als die Variabilität innerhalb der Stufen des Einflussfaktors ist.

- Beispiel
  - Für das bereits in den Abschnitten 3.2.1 bzw. 3.2.3 betrachtete Beispiel der Eingabe von Wörtern über drei verschiedene Tastaturen wollen wir nun prüfen, ob sich die drei Tastaturen hinsichtlich der jeweils erwarteten Anzahlen der eingegebenen Wörter je Minute signifikant voneinander unterscheiden $(1 - \gamma = 0.05)$.
  - Mit anderen Worten: Wir testen die ANOVA Nullhypothese $H_0 : \theta_1 = \theta_2 = \theta_3$ zum Niveau $1-\gamma = 0.05$.
  - Hierfür berechnen wir zunächst den beobachteten Wert der Testgröße in (49), wobei sich für die Schätzer $\overline{Y}_{1\cdot}$, $\overline{Y}_{2\cdot}$, $\overline{Y}_{3\cdot}$, $\overline{Y}_{\cdot\cdot}$ und S_p^2 die folgenden Werte ergeben:

$$\overline{Y}_{1\cdot} = 70.4\,, \qquad \overline{Y}_{2\cdot} = 77.0\,, \qquad \overline{Y}_{3\cdot} = 70.0\,, \qquad \overline{Y}_{\cdot\cdot} = 71.9\,, \qquad S_p^2 = 141.47$$

    und somit

$$\frac{\sum_{i=1}^{k} n_i\big(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\big)^2}{(k-1)S_p^2} = \frac{5\cdot(70.4-71.9)^2 + 3\cdot(77.0-71.9)^2 + 4\cdot(70.0-71.9)^2}{2\cdot 141.47} = 0.37\,.$$

  - Andererseits gilt $\mathrm{F}_{2,9,\,0.950} = 4.26 > 0.37$, weshalb $H_0 : \theta_1 = \theta_2 = \theta_3$ nicht abgelehnt wird.

3.3 Multiple lineare Regression

3.3.1 Modellbeschreibung

- Wir betrachten die folgende Verallgemeinerung des in Abschnitt 3.1 diskutierten (einfachen) linearen Regressionsmodells mit deterministischen Ausgangswerten $x_1, \dots, x_n$ eines *einzelnen* Einflussfaktors.

- Dabei lassen wir nun zu, dass die Zielvariablen $Y_1, \dots, Y_n$ nicht nur von den Werten $x_1, \dots, x_n$ eines einzelnen Einflussfaktors abhängen, d.h., wir betrachten $m-1$ Einflussfaktoren, wobei $m \geq 2$ eine beliebige, jedoch fest vorgegebene natürliche Zahl ist; $n \geq m$.

- Mit anderen Worten: Wir nehmen an, dass die Zielvariablen $Y_1, \dots, Y_n$ von $(m-1)$-dimensionalen vektoriellen Ausgangswerten $(x_{12}, \dots, x_{1m}), \dots, (x_{n2}, \dots, x_{nm})$ abhängen, d.h., es gelte

$$Y_i = \varphi(x_{i2}, \dots, x_{im}) + \varepsilon_i\,, \qquad \forall\, i = 1, \dots, n\,, \tag{51}$$

  wobei

  - die Regressionsfunktion $\varphi : \mathbb{R}^{m-1} \to \mathbb{R}$ gegeben ist durch

$$\varphi(x_2, \dots, x_m) = \beta_1 + \beta_2 x_2 + \dots + \beta_m x_m\,, \qquad \forall\,(x_2, \dots, x_m) \in \mathbb{R}^{m-1} \tag{52}$$

    mit der Regressionskonstante β_1 $(= \alpha)$ $\in \mathbb{R}$ und den Regressionskoeffizienten $\beta_2, \dots, \beta_m \in \mathbb{R}$,
  - die zufälligen Störgrößen $\varepsilon_1, \dots, \varepsilon_n : \Omega \to \mathbb{R}$ unabhängig sind mit

$$\mathbb{E}\,\varepsilon_i = 0\,, \qquad \operatorname{Var}\varepsilon_i = \sigma^2 \tag{53}$$

    für eine gewisse (unbekannte) Zahl $\sigma^2 > 0$.

- Die unbekannten Modellparameter $\beta_1, \dots, \beta_m$ und σ^2 sind aus den vorliegenden Daten $y_1, \dots, y_n$ bzw. $x_{12}, \dots, x_{1m}, \dots, x_{n2}, \dots, x_{nm}$ zu schätzen.

- Beachte

  - In Matrixschreibweise lässt sich das in (51) und (52) gegebene *multiple lineare Regressionsmodell* wie folgt formulieren:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,, \tag{54}$$

    wobei

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \dots & x_{nm} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{55}$$

  - Dabei wird $\mathbf{X}$ die *Designmatrix* des Regressionsmodells genannt.

3.3.2 Kleinste-Quadrate-Schätzer bei zwei Einflussfaktoren

- In diesem Abschnitt diskutieren wir zunächst den Spezialfall von zwei Einflussfaktoren, d.h. $m = 3$ und $n \geq 3$, wobei die Modellparameter $\beta_1, \beta_2, \beta_3$ und σ^2 erneut mit der Methode der kleinsten Quadrate geschätzt werden.
- Dabei ergeben sich Kleinste-Quadrate-Schätzer für $\beta_1, \beta_2, \beta_3$ durch Mnimierung des mittleren quadratischen Fehlers

$$e(\beta_1, \beta_2, \beta_3) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - (\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3})\big)^2 \tag{56}$$

  bzw. durch partielle Differentiation der Funktion $e(\beta_1, \beta_2, \beta_3)$ nach den Variablen β_1, β_2 bzw. β_3 und durch anschließendes Nullsetzen der Ableitungen.
- Hieraus folgt, dass die Kleinste-Quadrate-Schätzer $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ für $\beta_1, \beta_2, \beta_3$ den *Normalengleichungen*

$$\begin{aligned} n\widehat{\beta}_1 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \widehat{\beta}_3 \sum_{i=1}^{n} x_{i3} &= \sum_{i=1}^{n} Y_i \,, \\ \widehat{\beta}_1 \sum_{i=1}^{n} x_{i2} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 + \widehat{\beta}_3 \sum_{i=1}^{n} x_{i2} x_{i3} &= \sum_{i=1}^{n} x_{i2} Y_i \,, \\ \widehat{\beta}_1 \sum_{i=1}^{n} x_{i3} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} x_{i3} + \widehat{\beta}_3 \sum_{i=1}^{n} x_{i3}^2 &= \sum_{i=1}^{n} x_{i3} Y_i \end{aligned}$$

  genügen müssen, deren Lösung gegeben ist durch

$$\begin{aligned} \widehat{\beta}_1 &= \overline{Y} - \overline{x}_{\cdot 2}\widehat{\beta}_2 - \overline{x}_{\cdot 3}\widehat{\beta}_3 \,, \\ \widehat{\beta}_2 &= \sum_{i=1}^{n} \frac{c_{33}\big(x_{i2} - \overline{x}_{\cdot 2}\big) - c_{23}\big(x_{i3} - \overline{x}_{\cdot 3}\big)}{c_{22}c_{33} - c_{23}^2}\, Y_i \,, \\ \widehat{\beta}_3 &= \sum_{i=1}^{n} \frac{c_{22}\big(x_{i3} - \overline{x}_{\cdot 3}\big) - c_{23}\big(x_{i2} - \overline{x}_{\cdot 2}\big)}{c_{22}c_{33} - c_{23}^2}\, Y_i \,, \end{aligned} \tag{57}$$

  - wobei $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ bzw. $\overline{x}_{\cdot j} = \sum_{i=1}^{n} x_{ij}/n$ für $j = 2, 3$ und

$$c_{j_1 j_2} = \sum_{i=1}^{n} \big(x_{ij_1} - \overline{x}_{\cdot j_1}\big)\big(x_{ij_2} - \overline{x}_{\cdot j_2}\big) \qquad \text{für } j_1, j_2 = 2, 3 \tag{58}$$

  - und wobei vorausgesetzt wird, dass

$$c_{22}c_{33} - c_{23}^2 > 0 \,. \tag{59}$$

- Beachte
  - Die Bedingung (59) ist gleichbedeutend damit, dass die Designmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ 1 & x_{n2} & x_{n3} \end{pmatrix}$$

    vollen Spaltenrang $\operatorname{rg}(\mathbf{X}) = 3$ hat, d.h., dass $\det \mathbf{X} \neq 0$ bzw. dass $\mathbf{X}$ aus drei linear unabhängigen Spaltenvektoren besteht.

– Man kann zeigen, dass die Kleinste-Quadrate-Schätzer $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ für $\beta_1, \beta_2, \beta_3$ erwartungstreu sind, d.h., es gilt

$$\mathbb{E}\,\widehat{\beta}_1 = \beta_1\,, \qquad \mathbb{E}\,\widehat{\beta}_2 = \beta_2\,, \qquad \mathbb{E}\,\widehat{\beta}_3 = \beta_3\,. \tag{60}$$

– Die (zufällige) Abbildung $\widehat{Y} : \mathbb{R}^2 \to \mathbb{R}$ mit

$$\widehat{Y}(x_2, x_3) = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3\,, \tag{61}$$

die jedem Wertepaar (x_1, x_2) der beiden Einflussfaktoren die Zufallsvariable $\widehat{Y}(x_1, x_2)$ zuordnet, heißt *empirische Regressionsebene*.

– Außerdem kann man zeigen, dass die „Reststreueung" S^2 um die empirische Regressionsebene, die gegeben ist durch

$$S^2 = \frac{1}{n-3} \sum_{i=1}^{n} \Big(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_{i2} - \widehat{\beta}_3 x_{i3}\Big)^2\,, \tag{62}$$

ein erwartungstreuer Schätzer für σ^2 ist, d.h., es gilt $\mathbb{E}\,S^2 = \sigma^2$.

• Beispiel (vgl. R. Storm (2001) *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle*, Fachbuchverlag Leipzig, S. 260 ff.)

– In einem Unternehmen der Metallindustrie soll untersucht werden, inwiefern die Güte eines Erzeugnisses von zwei technologischen (Einfluss-) Faktoren des Produktionsprozesses abhängt.

– Dabei wird die Stahlproduktion eines Stahlwerkes betrachtet, wobei die Stahlausbeute (Zielvariable, gemessen in Prozent) in Abhängigkeit von der Anzahl der bisher erfolgten Abstiche (1.Einflussfaktor) und dem Schwefelgehalt (2. Einflussfaktor, gemessen in Prozent) untersucht wird.

– Die Qualitätskennzahl „Stahlausbeute" wurde für 26 Proben des Erzeugnisses bestimmt, wobei für die technologischen Einflussfaktoren „Anzahl der Abstiche" bzw. „Schwefelgehalt" jeweils unterschiedliche (Ausgangs-) Werte beobachtet wurden:

Nummer der Probe	1	2	3	4	5	6	7	8	9	10	11	12	13
Wert des 1. Faktors (x_{i2})	53	34	39	39	28	39	39	15	19	27	23	24	25
Wert des 2. Faktors (x_{i3})	8	8	7	9	9	8	9	12	12	8	8	8	8
Qualitätskennzahl (y_i)	19	70	0	77	85	70	0	100	78	78	98	59	87
Nummer der Probe	14	15	16	17	18	19	20	21	22	23	24	25	26
Wert des 1. Faktors (x_{i2})	27	9	37	20	23	11	10	13	45	6	7	15	22
Wert des 2. Faktors (x_{i3})	8	7	25	10	9	7	9	8	13	12	7	12	11
Qualitätskennzahl (y_i)	70	100	42	96	76	82	100	97	68	92	95	96	91

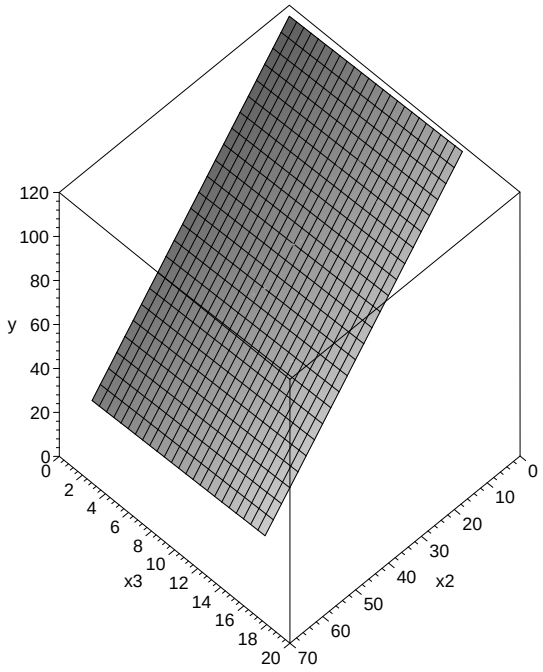
– Aus diesen Daten ergeben sich die folgenden Werte für die Größen, die in den Definitionsgleichungen (57) der Schätzer $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ vorkommen:

$$\overline{x}_{\cdot 2} = 24.96\,, \quad \overline{x}_{\cdot 3} = 9.69\,, \quad \overline{y} = 74.08\,, \quad c_{22} = 3995\,, \quad c_{33} = 326\,, \quad c_{23} = 155\,.$$

– Durch Einsetzen der ermittelten Werte in (57) ergibt sich nun, dass

$$\widehat{\beta}_1 = 114.972\,, \qquad \widehat{\beta}_2 = -1.695\,, \qquad \widehat{\beta}_3 = 0.146\,.$$

– Somit erhalten wir die geschätzte Regressionsebene $\widehat{y}(x_2, x_3) = 114.972 - 1.695 x_2 + 0.146 x_3$:



– Für die in (62) betrachtete Restreueung S^2 ergibt sich schließlich der Schätzwert $S^2 = 411.09$ bzw. $S = 20.28$.

3.3.3 Vektor- bzw. Matrixschreibweise

- Wir kehren nun zu dem (allgemeinen) multiplen linearen Regressionmodell mit einer *beliebigen* Anzahl $m-1$ von Einflussfaktoren zurück, das bereits in Abschnitt 3.3.1 betrachtet wurde.
  - So wie bisher schätzen wir die unbekannten Modellparameter $\beta_1, \ldots, \beta_m$ mit der Methode der kleinsten Quadrate aus den beobachteten Daten

$$(x_{12}, \ldots, x_{1m}), \ldots, (x_{n2}, \ldots, x_{nm}) \in \mathbb{R}^{m-1} \qquad \text{und} \qquad y_1 \ldots, y_n \in \mathbb{R}.$$

  - Mit anderen Worten: Es soll ein Zufallsvektor $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)$ bestimmt werden, so dass der *mittlere quadratische Fehler*

$$e(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - (\beta_1 + \beta_2 x_{i2} + \ldots + \beta_m x_{im})\big)^2 \tag{63}$$

    für $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ minimal wird, wobei wir voraussetzen, dass $n \geq m$ und dass der Rang der Matrix $\mathbf{X}$ gleich m ist.

- Ähnlich wie bei der Herleitung von (57) (d.h. bei der Lösung des entsprechenden Minimierungsproblems im Fall zweier Einflussfaktoren) kann man zeigen, dass für eine beliebige Anzahl von Einflussfaktoren

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \tag{64}$$

  - das (eindeutig bestimmte) Minimum des in (63) betrachteten Abweichungsmßes ist,
  - wobei $\mathbf{X}^\top$ die transponierte $m \times n$ Matrix bezeichnet, die sich durch Vertauschung der Zeilen und Spalten von $\mathbf{X}$ ergibt,
  - und $(\mathbf{X}^\top \mathbf{X})^{-1}$ die inverse Matrix von $\mathbf{X}^\top \mathbf{X}$ ist.

- Beachte
  - Der *Rang* $\operatorname{rg}(\mathbf{A})$ einer Matrix $\mathbf{A}$ ist die maximale Anzahl der linear unabhängigen Zeilenvektoren (bzw. Spaltenvektoren) von $\mathbf{A}$.
  - Zur Erinnerung: Die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_\ell \in \mathbb{R}^m$ heißen *linear abhängig*, falls es reelle Zahlen $c_1, \ldots, c_\ell \in \mathbb{R}$ gibt, die nicht alle gleich Null sind, so dass $c_1\mathbf{a}_1 + \ldots + c_\ell\mathbf{a}_\ell = \mathbf{o}$. Anderenfalls heißen die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_\ell \in \mathbb{R}^m$ *linear unabhängig*.
  - Weil wir voraussetzen, dass die Designmatrix $\mathbf{X}$ vollen (Spalten-) Rang $\operatorname{rg}(\mathbf{X}) = m$ hat, ist die symmetrische $m \times m$ Matrix $\mathbf{X}^\top\mathbf{X}$ *regulär*, d.h., es gilt $\det(\mathbf{X}^\top\mathbf{X}) \neq 0$.
  - Somit ist $\mathbf{X}^\top\mathbf{X}$ auch *invertierbar*, d.h., es gibt eine (eindeutig bestimmte) $m \times m$ Matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$, so dass
$$(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}) = (\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{I}\,,$$
wobei $\mathbf{I} = (\delta_{ij})$ die $m \times m$-dimensionale *Einheitsmatrix* bezeichnet mit
$$\delta_{ij} = \begin{cases} 1\,, & \text{falls } i = j, \\ 0\,, & \text{falls } i \neq j, \end{cases}$$

- Beachte
  - Der in (64) gegebene Kleinste-Quadrate-Schätzer $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_m)$ für den Parametervektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ ist ein sogenannter *linearer* Schätzer, d.h., $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ ist eine lineare Funktion der Zufallsstichprobe $\mathbf{Y} = (Y_1, \ldots, Y_n)$.
  - Der Schätzer $\widehat{\boldsymbol{\beta}}$ hat die folgenden *Güteeigenschaften*:
    1. Es gilt $\mathbb{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ für jedes $\boldsymbol{\beta} \in \mathbb{R}^m$, d.h., $\widehat{\boldsymbol{\beta}}$ ist *erwartungstreu*.
    2. Für jeden anderen linaren erwartungstreuen Schätzer $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_m)$ für $\boldsymbol{\beta}$ gilt
$$\operatorname{Var}\widehat{\beta}_i \leq \operatorname{Var}\widetilde{\beta}_i\,, \qquad \forall\, i = 1, \ldots, m\,, \tag{65}$$
wobei die Gleichheit in (65) genau dann für jedes $i = 1, \ldots, m$ gilt, wenn $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$, d.h., $\widehat{\boldsymbol{\beta}}$ ist *bester* linearer erwartungstreuer Schätzer für $\boldsymbol{\beta}$.
  - Die (zufällige) Abbildung $\widehat{Y} : \mathbb{R}^{m-1} \to \mathbb{R}$ mit
$$\widehat{Y}(\mathbf{x}) = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_m x_m\,, \tag{66}$$
die jedem Vektor $\mathbf{x} = (x_2, \ldots, x_m)$ von Werten der Einflussfaktoren die Zufallsvariable $\widehat{Y}(\mathbf{x})$ zuordnet, heißt *empirische Regressionshyperebene*.
  - Außerdem kann man in Verallgemeinerung von (62) zeigen, dass die „Reststreueung" S^2 um die empirische Regressionshyperebene, die gegeben ist durch
$$S^2 = \frac{1}{n-m}\,(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\,, \tag{67}$$
ein erwartungstreuer Schätzer für σ^2 ist, d.h., es gilt $\mathbb{E}\,S^2 = \sigma^2$.

3.3.4 t-Tests und Konfidenzintervalle für Regressionskonstante und Regressionskoeffizienten

- Zusätzlich zu den Modellannahmen, die bisher in Abschnit 3.3 gemacht wurden, setzen wir nun voraus, dass die zufälligen Störgrößen $\varepsilon_1, \ldots, \varepsilon_n$ normalverteilt sind. Wegen (53) gilt somit
$$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)\,, \qquad \forall\, i = 1, \ldots, n\,. \tag{68}$$

- Ähnlich wie bei den t-Tests, die in Abschnitt 3.1.3 für das einfache lineare Regressionsmodell betrachtet wurden, können wir dann *Hypothesen über einzelne Komponenten* des Parametervektors $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)$ testen:
  - Um einen hypothetischen Wert $\beta_{0,j}$ der j-ten Komponente β_j des Parametervektors $\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)$ zu testen, betrachten wir dabei die Testgröße
$$T_j = \frac{\widehat{\beta}_j - \beta_j}{S\sqrt{x^{jj}}} \sim \mathrm{t}_{n-m}\,, \tag{69}$$
wobei x^{jj} die (j,j)-te Eintragung der (inversen) Matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ bezeichnet; $j \in \{1, \dots, m\}$.
  - Beim Test der Hypothese $H_0 : \beta_j = \beta_{0,j}$ zum Niveau $1-\gamma \in (0,1)$ (gegen die Alternative $H_1 : \beta_j \neq \beta_{0,j}$) wird die Nullhypothese H_0 abgelehnt, falls
$$\frac{\left|\widehat{\beta}_j - \beta_{0,j}\right|}{S\sqrt{x^{jj}}} > \mathrm{t}_{n-m,1-(1-\gamma)/2}\,, \tag{70}$$
wobei $\mathrm{t}_{n-m,1-(1-\gamma)/2}$ das $(1-(1-\gamma)/2)$-Quantil der t-Verteilung mit $n-m$ Freiheitsgraden bezeichnet.
- Beachte
  - Für $j \in \{2, \dots, m\}$ ist der Test der Hypothese $H_0 : \beta_j = 0$ (gegen die Alternative $H_1 : \beta_j \neq 0$) von besonderem Interesse, weil damit verifiziert werden kann, inwieweit die Zielvariablen $Y_1, \dots, Y_n$ statistisch signifikant von dem $(j-1)$-ten Einflussfaktor abhängen.
  - Bei diesem Test auf Signifikanz des $(j-1)$-ten Einflussfaktors wird die Nullhypothese H_0 abgelehnt, falls
$$\frac{\left|\widehat{\beta}_j\right|}{S\sqrt{x^{jj}}} > \mathrm{t}_{n-m,1-(1-\gamma)/2}\,. \tag{71}$$
  - Außerdem ergibt sich aus (69) das folgende Konfidenzintervall für β_j zum Niveau $\gamma \in (0,1)$:
$$\widehat{\beta}_j - \mathrm{t}_{n-m,1-(1-\gamma)/2}\, S\sqrt{x^{jj}} < \beta_j < \widehat{\beta}_j + \mathrm{t}_{n-m,1-(1-\gamma)/2}\, S\sqrt{x^{jj}}\,. \tag{72}$$
- Beispiel
  - Für den bereits in Abschnitt 3.3.2 betrachteten Spezialfall von zwei Einflussfaktoren, d.h. $m = 3$, lassen sich die in (69) betrachteten Testgrößen T_2 und T_3 in der folgenden Form schreiben:
$$T_2 = \frac{\widehat{\beta}_2 - \beta_2}{S\sqrt{\dfrac{c_{33}}{c_{22}c_{33} - c_{23}^2}}} \sim \mathrm{t}_{n-3} \qquad \text{bzw.} \qquad T_3 = \frac{\widehat{\beta}_3 - \beta_3}{S\sqrt{\dfrac{c_{22}}{c_{22}c_{33} - c_{23}^2}}} \sim \mathrm{t}_{n-3}\,, \tag{73}$$
wobei die Größen c_{ij} und $S = \sqrt{S^2}$ in (58) bzw. (62) gegeben sind.
  - Wir betrachten nun erneut den in Abschnitt 3.3.2 gegebenen Beispiel-Datensatz und prüfen, ob die Zielvariable „Stahlausbeute" statistisch signifikant von den beiden technologischen Einflussfaktoren „Anzahl der bisherigen Abstiche" und „Schefelgehalt" des Produktionsprozesses abhängt.
  - Mit anderen Worten: Wir verifizieren die Hypothesen
$$H_0 : \beta_2 = 0 \text{ (versus } H_0 : \beta_2 \neq 0) \qquad \text{bzw.} \qquad H_0 : \beta_3 = 0 \text{ (versus } H_0 : \beta_3 \neq 0).$$
  - In Abschnitt 3.3.2 hatten wir gezeigt, dass
$$\beta_2 = -1.695\,, \quad \widehat{\beta}_3 = 0.146\,, \quad S = 20.28\,, \quad c_{22} = 3995\,, \quad c_{33} = 326\,, \quad c_{23} = 155\,.$$

- Durch Einsetzen in (73) ergeben sich nun unter $H_0 : \beta_2 = 0$ bzw. $H_0 : \beta_3 = 0$ die folgenden „Testwerte" zum Niveau $1 - \gamma = 0.05$:

$$T_2 = -5.23 \qquad \text{bzw.} \qquad T_3 = 0.13\,.$$

- Weil $|T_2| = 5.23 > 2.07 = t_{23,\,0.975}$ gilt, wird die Null-Hypothese $H_0 : \beta_2 = 0$ verworfen, d.h., die Qualitätskennzahl „Stahlausbeute" hängt statistisch signifikant von dem ersten technlogischen Einflussfaktor „Anzahl der bisherigen Abstiche" ab.
- Andererseits gilt $|T_3| = 0.13 < 2.07 = t_{23,\,0.975}$, d.h., die Abhängigkeit der Qualitätskennzahl „Stahlausbeute" von dem zweiten technlogischen Einflussfaktor „Schwefelgehalt" ist statistisch nicht gesichert.
- Aus (73) ergeben sich darüber hinaus die Konfidenzintervalle $(-2.366, -1.024)$ und$(-2.201, 2.493)$ für β_2 bzw. β_3, wenn dabei das Konfidenzniveau $\gamma = 0.950$ betrachtet wird.

3.3.5 Güte der Modellanpassung; Overall-F-Test

- Ähnlich wie bei der einfachen linearen Regression (vgl. Abschnitt 2.5.3) gilt auch bei der multiplen linearen Regression
  - die *Quadratsummen-Zerlegung*

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{74}$$

  der „Gesamtstreuung" $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ in die sogenannte „erklärte Streueung" $\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ und die „Reststreuung" $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$, wobei

$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_m x_{im}\,, \qquad \text{bzw.} \qquad \overline{Y} = \frac{Y_1 + \ldots + Y_n}{n}\,.$$

  - Für den Quotienten

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} \tag{75}$$

  von erklärter Streuung und Gesamtstreuung ergibt sich aus (74), dass

$$0 \le R^2 \le 1\,.$$

  - Das *sogenannte Bestimmtheitsmaß* R^2 ist eine Maßzahl für die Anpassungsgüte der geschätzten Regressionshyperebene $\widehat{Y}(\mathbf{x}) = \widehat{\beta}_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_m x_m$ an die beobachteten Werte $y_1, \ldots, y_n$ der Zielvariablen $Y_1, \ldots, Y_n$.

- Beachte
  - Der in (71) betrachtete t-Test dient lediglich zur Verifizierung der statistischen Signifikanz von *einzelnen* Einflussfaktoren.
  - Man kann jedoch einen (simultanen) F-Test konstruieren, um die statistische Signifikanz von *sämtlichen* Einflussfaktoren gleichzeitig zu prüfen.
  - Dabei wird die Hypothese $H_0 : \beta_2 = \beta_3 = \ldots = \beta_m = 0$ (gegen die Alternative $H_1 : \beta_j \neq 0$ für ein $j \in \{2, \ldots, m\}$) getestet.

– Unter H_0 gilt nämlich

$$T = \frac{n-m}{m-1}\,\frac{R^2}{1-R^2} \;\sim\; \mathrm{F}_{m-1,\,n-m}\,, \tag{76}$$

wobei R^2 die in (75) gegebene Maßzahl ist.

– Die Hypothese $H_0 : \beta_2 = \beta_3 = \ldots = \beta_m = 0$ wird abgelehnt, falls $T > \mathrm{F}_{m-1,\,n-m,\,\gamma}$.

4 Tabellen für Verteilungsfunktionen und Quantile

Tabelle 1 Verteilungsfunktion $\Phi(x)$ der Standardnormalverteilung

x	0	1	2	3	4	5	6	7	8	9
0,0	0,500000	0,503989	0,507978	0,511967	0,515953	0,519939	0,523922	0,527903	0,531881	0,535856
0,1	0,539828	0,543795	0,547758	0,551717	0,555670	0,559618	0,563559	0,567495	0,571424	0,575345
0,2	0,579260	0,583166	0,587064	0,590954	0,594835	0,598706	0,602568	0,606420	0,610261	0,614092
0,3	0,617911	0,621719	0,625516	0,629300	0,633072	0,636831	0,640576	0,644309	0,648027	0,651732
0,4	0,655422	0,659097	0,662757	0,666402	0,670031	0,673645	0,677242	0,680822	0,684386	0,687933
0,5	0,691462	0,694974	0,698468	0,701944	0,705402	0,708840	0,712260	0,715661	0,719043	0,722405
0,6	0,725747	0,729069	0,732371	0,735653	0,738914	0,742154	0,745373	0,748571	0,751748	0,754903
0,7	0,758036	0,761148	0,764238	0,767305	0,770350	0,773373	0,776373	0,779350	0,782305	0,785236
0,8	0,788145	0,791030	0,793892	0,796731	0,799546	0,802338	0,805106	0,807850	0,810570	0,813267
0,9	0,815940	0,818589	0,821214	0,823814	0,826391	0,828944	0,831472	0,833977	0,836457	0,838913
1,0	0,841345	0,843752	0,846136	0,848495	0,850830	0,853141	0,855428	0,857690	0,859929	0,862143
1,1	0,864334	0,866500	0,868643	0,870762	0,872857	0,874928	0,876976	0,878999	0,881000	0,882977
1,2	0,884930	0,886860	0,888767	0,890651	0,892512	0,894350	0,896165	0,897958	0,899727	0,901475
1,3	0,903199	0,904902	0,906582	0,908241	0,909877	0,911492	0,913085	0,914656	0,916207	0,917736
1,4	0,919243	0,920730	0,922196	0,923641	0,925066	0,926471	0,927855	0,929219	0,930563	0,931888
1,5	0,933193	0,934478	0,935744	0,936992	0,938220	0,939429	0,940620	0,941792	0,942947	0,944083
1,6	0,945201	0,946301	0,947384	0,948449	0,949497	0,950529	0,951543	0,952540	0,953521	0,954486
1,7	0,955435	0,956367	0,957284	0,958185	0,959071	0,959941	0,960796	0,961636	0,962462	0,963273
1,8	0,964070	0,964852	0,965621	0,966375	0,967116	0,967843	0,968557	0,969258	0,969946	0,970621
1,9	0,971284	0,971933	0,972571	0,973197	0,973810	0,974412	0,975002	0,975581	0,976148	0,976705
2,0	0,977250	0,977784	0,978308	0,978822	0,979325	0,979818	0,980301	0,980774	0,981237	0,981691
2,1	0,982136	0,982571	0,982997	0,983414	0,983823	0,984222	0,984614	0,984997	0,985371	0,985738
2,2	0,986097	0,986447	0,986791	0,987126	0,987455	0,987776	0,988089	0,988396	0,988696	0,988989
2,3	0,989276	0,989556	0,989830	0,990097	0,990358	0,990613	0,990863	0,991106	0,991344	0,991576
2,4	0,991802	0,992024	0,992240	0,992451	0,992656	0,992857	0,993053	0,993244	0,993431	0,993613
2,5	0,993790	0,993963	0,994132	0,994297	0,994457	0,994614	0,994766	0,994915	0,995060	0,995201
2,6	0,995339	0,995473	0,995603	0,995731	0,995855	0,995975	0,996093	0,996207	0,996319	0,996427
2,7	0,996533	0,996636	0,996736	0,996833	0,996928	0,997020	0,997110	0,997197	0,997282	0,997365
2,8	0,997445	0,997523	0,997599	0,997673	0,997744	0,997814	0,997882	0,997948	0,998012	0,998074
2,9	0,998134	0,998193	0,998250	0,998305	0,998359	0,998411	0,998462	0,998511	0,998559	0,998605
3,0	0,998650	0,998694	0,998736	0,998777	0,998817	0,998856	0,998893	0,998930	0,998965	0,998999
3,5	0,999767	0,999776	0,999784	0,999792	0,999800	0,999807	0,999815	0,999821	0,999828	0,999835
4,0	0,999968	0,999970	0,999971	0,999972	0,999973	0,999974	0,999975	0,999976	0,999977	0,999978

Tabelle 2a γ-Quantil $\chi^2_{r,\gamma}$ der χ^2-Verteilung mit r Freiheitsgraden

$r\backslash\gamma$	.005	.01	.025	.05	.10	.90	.95	.975	.99	.995
1	.00004	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

Tabelle 2b γ-Quantil $\chi^2_{r,\gamma}$ der χ^2-Verteilung mit r Freiheitsgraden

n \ γ	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84
4	4,88	5,39	5,99	6,74	7,78	9,49	11,14	13,28	14,86
5	6,06	6,63	7,29	8,12	9,24	11,07	12,83	15,09	16,75
6	7,23	7,84	8,56	9,45	10,64	12,59	14,45	16,81	18,55
7	8,38	9,04	9,80	10,75	12,02	14,07	16,01	18,48	20,28
8	9,52	10,22	11,03	12,03	13,36	15,51	17,53	20,09	21,95
9	10,66	11,39	12,24	13,29	14,68	16,92	19,02	21,67	23,59
10	11,78	12,55	13,44	14,53	15,99	18,31	20,48	23,21	25,19
11	12,90	13,70	14,63	15,77	17,28	19,68	21,92	24,73	26,76
12	14,01	14,85	15,81	16,99	18,55	21,03	23,34	26,22	28,30
13	15,12	15,98	16,98	18,20	19,81	22,36	24,74	27,69	29,82
14	16,22	17,12	18,15	19,41	21,06	23,68	26,12	29,14	31,32
15	17,32	18,25	19,31	20,60	22,31	25,00	27,49	30,58	32,80
16	18,42	19,37	20,47	21,79	23,54	26,30	28,85	32,00	34,27
17	19,51	20,49	21,61	22,98	24,77	27,59	30,19	33,41	35,72
18	20,60	21,60	22,76	24,16	25,99	28,87	31,53	34,81	37,16
19	21,69	22,72	23,90	25,33	27,20	30,14	32,85	36,19	38,58
20	22,77	23,83	25,04	26,50	28,41	31,41	34,17	37,57	40,00
21	23,86	24,93	26,17	27,66	29,62	32,67	35,48	38,93	41,40
22	24,94	26,04	27,30	28,82	30,81	33,92	36,78	40,29	42,80
23	26,02	27,14	28,43	29,98	32,01	35,17	38,08	41,64	44,18
24	27,10	28,24	29,55	31,13	33,20	36,42	39,36	42,98	45,56
25	28,17	29,34	30,68	32,28	34,38	37,65	40,65	44,31	46,93
30	33,53	34,80	36,25	37,99	40,26	43,77	46,98	50,89	53,67
40	44,16	45,62	47,27	49,24	51,81	55,76	59,34	63,69	66,77
50	54,72	56,33	58,16	60,35	63,17	67,50	71,42	76,15	79,49
60	65,23	66,98	68,97	71,34	74,40	79,08	83,30	88,38	91,95
70	75,69	77,58	79,71	82,26	85,53	90,53	95,02	100,43	104,21
80	86,12	88,13	90,41	93,11	96,58	101,88	106,63	112,33	116,32
90	96,52	98,65	101,05	103,90	107,57	113,15	118,14	124,12	128,30
100	106,91	109,14	111,67	114,66	118,50	124,34	129,56	135,81	140,17
150	158,58	161,29	164,35	167,96	172,58	179,58	185,80	193,21	198,36
200	209,99	213,10	216,61	220,74	226,02	233,99	241,06	249,45	255,26
500	516,09	520,95	526,40	532,80	540,93	553,13	563,85	576,49	585,21

Tabelle 3 γ-Quantil $t_{r,\gamma}$ der t-Verteilung mit r Freiheitsgraden

r \ γ	0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656
2	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
30	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
40	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704
50	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678
60	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660
70	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,994	2,381	2,648
80	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639
90	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,368	2,632
100	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626
150	0,386	0,526	0,676	0,844	1,040	1,287	1,655	1,976	2,351	2,609
200	0,386	0,525	0,676	0,843	1,039	1,286	1,653	1,972	2,345	2,601
500	0,386	0,525	0,675	0,842	1,038	1,283	1,648	1,965	2,334	2,586
1000	0,385	0,525	0,675	0,842	1,037	1,282	1,646	1,962	2,330	2,581

Tabelle 4a γ-Quantil $F_{r,s,\gamma}$ der F-Verteilung mit (r, s) Freiheitsgraden

$s \setminus r$	1	2	3	4	5	6	7	8	9	10	11	12	γ
1	161	200	216	225	230	234	237	239	241	242	243	244	0,95
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106	0,99
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41	0,95
	98,49	99,00	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40	99,41	99,42	0,99
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74	0,95
	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,92	0,99
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	0,95
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37	0,99
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68	0,95
	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89	0,99
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	0,95
	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	0,99
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57	0,95
	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47	0,99
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28	0,95
	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67	0,99
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07	0,95
	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11	0,99
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91	0,95
	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71	0,99
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79	0,95
	9,65	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40	0,99
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69	0,95
	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16	0,99

Tabelle 4b γ-Quantil $F_{r,s,\gamma}$ der F-Verteilung mit (r, s) Freiheitsgraden

$s \setminus r$	14	16	20	24	30	40	50	75	100	200	500	∞	γ
1	245	246	248	249	250	251	252	253	253	254	254	254	0,95
	6142	6169	6208	6234	6258	6286	6302	6323	6334	6352	6361	6366	0,99
2	19,42	19,43	19,44	19,45	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50	0,95
	99,43	99,44	99,45	99,46	99,47	99,48	99,48	99,49	99,49	99,49	99,50	99,50	0,99
3	8,71	8,69	8,66	8,64	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53	0,95
	26,83	26,69	26,60	26,50	26,41	26,35	26,27	26,23	26,18	26,14	26,12	34,12	0,99
4	5,87	5,84	5,80	5,77	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63	0,95
	14,24	14,15	14,02	13,93	13,83	13,74	13,69	13,61	13,57	13,52	13,48	13,46	0,99
5	4,64	4,60	4,56	4,53	4,50	4,46	4,44	4,42	4,40	4,38	4,37	4,36	0,95
	9,77	9,68	9,55	9,47	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02	0,99
6	3,96	3,92	3,87	3,84	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67	0,95
	7,60	7,52	7,39	7,31	7,23	7,14	7,09	7,02	6,99	6,94	6,90	6,88	0,99
7	3,52	3,49	3,44	3,41	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23	0,95
	6,35	6,27	6,15	6,07	5,98	5,90	5,85	5,78	5,75	5,70	5,67	5,65	0,99
8	3,23	3,20	3,15	3,12	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93	0,95
	5,56	5,48	5,36	5,28	5,20	5,11	5,06	5,00	4,96	4,91	4,88	4,86	0,99
9	3,02	2,98	2,93	2,90	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71	0,95
	5,00	4,92	4,80	4,73	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31	0,99
10	2,86	2,82	2,77	2,74	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54	0,95
	4,60	4,52	4,41	4,33	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91	0,99
11	2,74	2,70	2,65	2,61	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40	0,95
	4,29	4,21	4,10	4,02	3,94	3,86	3,80	3,74	3,70	3,66	3,62	3,60	0,99
12	2,64	2,60	2,54	2,50	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30	0,95
	4,05	3,98	3,86	3,78	3,70	3,61	3,56	3,49	3,46	3,41	3,38	3,36	0,99

Tabelle 4c γ-Quantil $F_{r,s,\gamma}$ der F-Verteilung mit (r, s) Freiheitsgraden

$s \setminus r$	1	2	3	4	5	6	7	8	9	10	11	12	γ
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60	0,95
	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	0,99
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53	0,95
	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	0,99
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48	0,95
	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	0,99
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	0,95
	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55	0,99
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38	0,95
	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45	0,99
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	0,95
	8,28	6,01	5,09	4,58	4,25	4,01	3,85	3,71	3,60	3,51	3,44	3,37	0,99
19	4,38	3,51	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,34	2,31	0,95
	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	0,99
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31	2,28	0,95
	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,30	3,23	0,99
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	0,95
	8,02	5,78	4,87	4,37	4,04	3,81	3,65	3,51	3,40	3,31	3,24	3,17	0,99
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,26	2,23	0,95
	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	0,99
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,24	2,20	0,95
	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	0,99
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18	0,95
	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	3,09	3,03	0,99
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24	2,20	2,16	0,95
	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,21	3,13	3,05	2,99	0,99

Tabelle 4d γ-Quantil $F_{r,s,\gamma}$ der F-Verteilung mit (r, s) Freiheitsgraden

$s \setminus r$	14	16	20	24	30	40	50	75	100	200	500	∞	γ
13	2,55	2,51	2,46	2,42	2,38	2,34	2,32	2,28	2,26	2,24	2,22	2,21	0,95
	3,85	3,78	3,67	3,59	3,51	3,42	3,37	3,30	3,27	3,21	3,18	3,16	0,99
14	2,48	2,44	2,39	2,35	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13	0,95
	3,70	3,62	3,51	3,43	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3,00	0,99
15	2,43	2,39	2,33	2,29	2,25	2,21	2,18	2,15	2,12	2,10	2,08	2,07	0,95
	3,56	3,48	3,36	3,29	3,20	3,12	3,07	3,00	2,97	2,92	2,89	2,87	0,99
16	2,37	2,33	2,28	2,24	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01	0,95
	3,45	3,37	3,25	3,18	3,10	3,01	2,96	2,89	2,86	2,80	2,77	2,75	0,99
17	2,33	2,29	2,23	2,19	2,15	2,11	2,08	2,04	2,02	1,99	1,97	1,96	0,95
	3,35	3,27	3,16	3,08	3,00	2,92	2,86	2,79	2,76	2,70	2,67	2,65	0,99
18	2,29	2,25	2,19	2,15	2,11	2,07	2,04	2,00	1,98	1,95	1,93	1,92	0,95
	3,27	3,19	3,07	3,00	2,91	2,83	2,78	2,71	2,68	2,62	2,59	2,57	0,99
19	2,26	2,21	2,15	2,11	2,07	2,02	2,00	1,96	1,94	1,91	1,90	1,88	0,95
	3,19	3,12	3,00	2,92	2,84	2,76	2,70	2,63	2,60	2,54	2,51	2,49	0,99
20	2,23	2,18	2,12	2,08	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84	0,95
	3,13	3,05	2,94	2,86	2,77	2,69	2,63	2,56	2,53	2,47	2,44	2,42	0,99
21	2,20	2,15	2,09	2,05	2,00	1,96	1,93	1,89	1,87	1,84	1,82	1,81	0,95
	3,07	2,99	2,88	2,80	2,72	2,63	2,58	2,51	2,47	2,42	2,38	2,36	0,99
22	2,18	2,13	2,07	2,03	1,98	1,93	1,91	1,87	1,84	1,81	1,80	1,78	0,95
	3,02	2,94	2,83	2,75	2,67	2,58	2,53	2,46	2,42	2,37	2,33	2,31	0,99
23	2,14	2,10	2,05	2,00	1,96	1,91	1,88	1,84	1,82	1,79	1,77	1,76	0,95
	2,97	2,89	2,78	2,70	2,62	2,53	2,48	2,41	2,37	2,32	2,28	2,26	0,99
24	2,13	2,09	2,02	1,98	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73	0,95
	2,93	2,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21	0,99
25	2,11	2,06	2,00	1,96	1,92	1,87	1,84	1,80	1,77	1,74	1,72	1,71	0,95
	2,89	2,81	2,70	2,62	2,54	2,45	2,40	2,32	2,29	2,23	2,19	2,17	0,99

Tabelle 4e γ-Quantil $F_{r,s,\gamma}$ der F-Verteilung mit (r, s) Freiheitsgraden

$s \setminus r$	1	2	3	4	5	6	7	8	9	10	11	12	γ
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	0,95
	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,17	3,09	3,02	2,96	0,99
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20	2,16	2,13	0,95
	7,68	5,49	4,60	4,11	3,79	3,56	3,39	3,26	3,14	3,06	2,98	2,93	0,99
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,15	2,12	0,95
	7,64	5,45	4,57	4,07	3,76	3,53	3,36	3,23	3,11	3,03	2,95	2,90	0,99
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18	2,14	2,10	0,95
	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,08	3,00	2,92	2,87	0,99
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12	2,09	0,95
	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,90	2,84	0,99
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04	2,00	0,95
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88	2,80	2,73	2,66	0,99
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98	1,95	0,95
	7,17	5,06	4,20	3,72	3,41	3,18	3,02	2,88	2,78	2,70	2,62	2,56	0,99
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	0,95
	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	0,99
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93	1,89	0,95
	7,01	4,92	4,08	3,60	3,29	3,07	2,91	2,77	2,67	2,59	2,51	2,45	0,99
80	3,96	3,11	2,72	2,48	2,33	2,21	2,12	2,05	1,99	1,95	1,91	1,88	0,95
	6,96	4,88	4,04	3,56	3,25	3,04	2,87	2,74	2,64	2,55	2,48	2,41	0,99
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88	1,85	0,95
	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,43	2,36	0,99
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87	1,83	1,80	0,95
	6,76	4,71	3,88	3,41	3,11	2,90	2,73	2,60	2,50	2,41	2,34	2,28	0,99
1000	3,85	3,00	2,61	2,38	2,22	2,10	2,02	1,95	1,89	1,84	1,80	1,76	0,95
	6,66	4,62	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,26	2,20	0,99

Tabelle 4f γ-Quantil $F_{r,s,\gamma}$ der F-Verteilung mit (r, s) Freiheitsgraden

$s \setminus r$	14	16	20	24	30	40	50	75	100	200	500	∞	γ
26	2,10	2,05	1,99	1,95	1,90	1,85	1,82	1,78	1,76	1,72	1,70	1,69	0,95
	2,86	2,77	2,66	2,58	2,50	2,41	2,36	2,28	2,25	2,19	2,15	2,13	0,99
27	2,08	2,03	1,97	1,93	1,88	1,84	1,80	1,76	1,74	1,71	1,68	1,67	0,95
	2,83	2,74	2,63	2,55	2,47	2,38	2,33	2,25	2,21	2,16	2,12	2,10	0,99
28	2,06	2,02	1,96	1,91	1,87	1,81	1,78	1,75	1,72	1,69	1,67	1,65	0,95
	2,80	2,71	2,60	2,52	2,44	2,35	2,30	2,22	2,18	2,13	2,09	2,06	0,99
29	2,05	2,00	1,94	1,90	1,85	1,80	1,77	1,73	1,71	1,68	1,65	1,64	0,95
	2,77	2,68	2,57	2,49	2,41	2,32	2,27	2,19	2,15	2,10	2,06	2,03	0,99
30	2,04	1,99	1,93	1,89	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62	0,95
	2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01	0,99
40	1,95	1,90	1,84	1,79	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51	0,95
	2,56	2,49	2,37	2,29	2,20	2,11	2,05	1,97	1,94	1,88	1,84	1,81	0,99
50	1,90	1,85	1,78	1,74	1,69	1,63	1,60	1,55	1,52	1,48	1,46	1,44	0,95
	2,46	2,39	2,26	2,18	2,10	2,00	1,94	1,86	1,82	1,76	1,71	1,68	0,99
60	1,86	1,81	1,75	1,70	1,65	1,59	1,56	1,50	1,48	1,44	1,41	1,39	0,95
	2,40	2,32	2,20	2,12	2,03	1,93	1,87	1,79	1,74	1,68	1,63	1,60	0,99
70	1,84	1,79	1,72	1,67	1,62	1,56	1,53	1,47	1,45	1,40	1,37	1,35	0,95
	2,35	2,28	2,15	2,07	1,98	1,88	1,82	1,74	1,69	1,62	1,56	1,53	0,99
80	1,82	1,77	1,70	1,65	1,60	1,54	1,51	1,45	1,42	1,38	1,35	1,32	0,95
	2,32	2,24	2,11	2,03	1,94	1,84	1,78	1,70	1,65	1,57	1,52	1,49	0,99
100	1,79	1,75	1,68	1,63	1,57	1,51	1,48	1,42	1,39	1,34	1,30	1,28	0,95
	2,26	2,19	2,06	1,98	1,89	1,79	1,73	1,64	1,59	1,51	1,46	1,43	0,99
200	1,74	1,69	1,62	1,57	1,52	1,45	1,42	1,35	1,32	1,26	1,22	1,19	0,95
	2,17	2,09	1,97	1,88	1,79	1,69	1,62	1,53	1,48	1,39	1,33	1,28	0,99
1000	1,70	1,65	1,58	1,53	1,47	1,41	1,36	1,30	1,26	1,19	1,13	1,08	0,95
	2,09	2,01	1,89	1,81	1,71	1,61	1,54	1,44	1,38	1,28	1,19	1,11	0,99