

Statistik Gurker WS 2014/2015

Zusammenfassung

Jakob Abfalter

11. Juli 2015

Inhaltsverzeichnis

1	Deskriptive und explorative Statistik	4
1.1	Diskrete univariate Merkmale	5
1.1.1	Kreisdiagramm	5
1.1.2	Balkendiagramm	5
1.1.3	Mosaikplot	6
1.1.4	Pareto-Diagramm	6
1.2	Stetige univariate Merkmale	7
1.2.1	Empirische Verteilungsfunktion	7
1.2.2	Stem- and Leaf-Plot	7
1.2.3	Klassierung	7
1.2.4	Histogramm	8
1.2.5	Kernschätzung	8
1.2.6	Quantile (Kenngröße der Lage)	9
1.2.7	QQ-Plot	9
1.2.8	Boxplot	10
1.3	Kennzahlen	10
1.3.1	Mittelwert:	11
1.3.2	Bruchpunkt:	11
1.3.3	Median:	11
1.3.4	Varianz:	11
1.3.5	MAD:	11
1.3.6	Modalwert:(Modus)	12
1.3.7	Momente:	12
1.3.8	Schiefte:	12
1.3.9	Kurtosis	12
1.4	Mehrdimensionale Daten	12
1.4.1	Korrelation	13
1.4.2	Kleinste Quadrate	13

2	Wahrscheinlichkeit	13
2.1	Gesetz der großen Zahlen	13
2.2	Merkmalraum:	14
2.3	Ereignisse:	14
2.4	Ereignissysteme:	14
2.5	Borelmengen	14
2.6	Wahrscheinlichkeitsmaße	14
2.7	Endliche W-Räume:	15
2.7.1	Geometrische Wahrscheinlichkeiten	15
2.8	Additionstheorem	15
2.9	Bedingte Wahrscheinlichkeiten	16
2.10	Multiplikationstheorem	16
2.11	Vollständige Wahrscheinlichkeit (totale Wahrscheinlichkeit)	16
2.12	Bayes'sche Formel	17
2.12.1	Odds-Form der Bay'schen Formel	17
2.13	Unabhängigkeit:	17
2.14	Mehrstufige Experimente	18
3	Stochastische Größen und Verteilungen	18
3.1	Verteilungsfunktion	18
3.1.1	Inverse	19
3.2	Diskrete Verteilung	19
3.3	Stetige Verteilung	19
3.4	Gemischte Verteilungen	20
3.5	Transformationen	20
3.5.1	Transformationen diskreter sGn	20
3.5.2	Transformationen stetiger sGn	20
3.5.3	Transformationssatz für Dichten	21
3.6	Erwartungswert	21
3.6.1	Erwartungswert einer diskreten sG	21
3.6.2	Erwartungswert einer stetigen sG	21
3.6.3	Erwartungswert einer Funktion von X	21
3.6.4	LotUS(Satz des unbewussten Statistiker)	22
3.7	Varianz	22
3.8	MAD	23
3.9	Simulation	23
4	Spezielle Verteilungen	23
4.1	Diskrete Verteilungen	23
4.1.1	Diskrete uniforme Verteilung (Gleichverteilung)	23
4.1.2	Bernoulli-Verteilung	24
4.1.3	Binomialverteilung	24
4.1.4	Negative Binomialverteilung	24
4.1.5	Geometrische Verteilung	24
4.1.6	Hypergeometrische Verteilung	25

4.1.7	Poisson- Verteilung	25
4.2	Stetige Verteilungen (Merkmalraum = Intervall)	26
4.2.1	Stetige uniforme Verteilung	26
4.2.2	Exponentialverteilung	26
4.2.3	Gamma- und Chiquadratverteilung	27
4.2.4	Normalverteilung	27
4.2.5	F-Verteilung	28
4.2.6	t-Verteilung	28
4.2.7	Betaverteilung	28
5	Multivariate Verteilungen	29
5.1	Bivariate Verteilungen	29
5.1.1	Diskrete stochastische Vektoren	29
5.1.2	Stetige stochastische Vektoren	30
5.2	Erwartungswert	30
5.3	Bedingte Verteilung	31
5.3.1	Stetiger Fall	31
5.4	Korrelation	32
5.5	Unabhängigkeit	32
5.6	Mehrdimensionale Erweiterungen	33
5.6.1	Varianz-Kovarianzmatrix	33
5.6.2	Transformationen	33
5.7	Spezielle multivariate Verteilungen	34
5.7.1	Multinomialverteilung	34
5.7.2	Polyhypergeometrische Verteilung	34
5.7.3	Multivariate Normalverteilung	34
6	Folgen von stochastischen Größen	35
6.1	Faltung	35
6.1.1	Diskrete Faltung	35
6.1.2	Stetige Faltung	35
6.2	Additionstheoreme	36
6.3	Ungleichungen	36
6.4	Gesetz der großen Zahlen	36
6.5	Schwaches Gesetz der großen Zahlen	36
6.6	Konvergenz in der Verteilung	36
6.7	Zentraler Grenzwertungssatz (ZGVS):	37
6.8	Normalapproximation	37
7	Schließende Statistik	37
7.1	Schätzer	38
7.1.1	Empirische Verteilungsfunktion	38
7.1.2	Satz von Gliwenko-Cantelli: (Fundamentalsatz der Statistik)	38
7.1.3	Dichteschätzer	38
7.1.4	Momentenschätzer	38

7.1.5	Maximum Likelihood	39
7.2	Gütekriterien für Schätzer	39
7.2.1	Erwartungstreue	39
7.2.2	Effizienz	40
7.2.3	Konsistenz	40
7.2.4	Asymptotische Normalverteilung	40
7.3	Konfidenzintervalle	40
7.3.1	Pivotmethode	40
7.3.2	Normalverteilung	41
7.3.3	Normalverteilung (zwei ua. Stichproben)	41
7.3.4	Normalverteilung (verbundene Stichproben)	41
7.4	Statistische Tests	41
7.4.1	Parametertests	41
7.5	p-Wert	42
7.6	Chiquadrat-Anpassungstests	43

1 Einleitung

Ich übernehme keine Garantie auf Richtigkeit und Vollständigkeit

2 Deskriptive und explorative Statistik

Die Deskriptive Statistik beschäftigt sich mit der tabellarischen und grafischen Aufbereitung von Daten sowie mit ihrer zahlenmäßigen Beschreibung. (z.B: Berechnung von Kenngrößen) Verwendet keine Modelle.

Explorative Datenanalyse (EDA) -> unbekannte Strukturen und Zusammenhänge in den Daten aufdecken.

Erster Schritt: Erhebung der Daten entweder durch

- Experimente (statistische Einheit: Versuchseinheit)
- Beobachtungsstudien (statistische Einheit: Beobachtungseinheit)

Experimentalstudien sind zu bevorzugen, weil kausale Zusammenhänge feststellbar sind. (Experimentalstudien sind aber nicht immer möglich)

Statistischen Einheiten, über die Aussagen getroffen werden sollen -> **Grundgesamtheit** (präzise Definition wichtig)

Repräsentative Teilauswahl aus der Grundgesamtheit -> **Stichprobe** (Auswahl zufällig)

N über n Auswahlmöglichkeiten für Stichprobe. Ziehen mit (einfacher) und Ziehen ohne Zurücklegen möglich. Die interessierenden Größen -> **Merkmale oder Variablen**. Werte die ein Merkmal annehmen kann -> **Merkmalausprägungen**. Merkmal ist eine Abbildung. Es gibt folgende Messskalen:

- Nominalskalen: reine Klassifikationen, keine weitere Relation bsp: Geschlecht

- Ordinalskalen: Rangmerkmale bsp: Prüfungsnoten
- Intervallskalen: reelle Zahlen (oder Vektoren) ohne Nullpunkt bsp: Zeiteinteilung
- Verhältnisskalen: Intervallskalen mit Nullpunkt bsp: Geschwindigkeit

Ausgangspunkt: Rohdaten -> in Datenmatrix

- univariaten Daten: 1 beobachtete Ausprägung ($p = 1$)
- multivariaten Daten: mehrere beobachtete Ausprägungen (z.B: $p = 6$)

Man unterscheidet außerdem zwischen diskreten und stetigen Merkmalen.

2.1 Diskrete univariate Merkmale

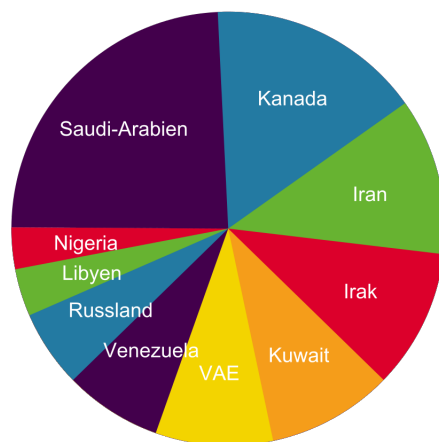
Erfolgt durch Bestimmung von Häufigkeiten und einer Visualisierung. Merkmal, das Werte $x_1 < x_2 < \dots$ annehmen kann, wird n Mal beobachtet. Absolute Häufigkeit mit der x_i beobachtet wird, wird mit n_i bezeichnet. Der größte beobachtete Merkmalswert ist x_k .

Relativen Häufigkeiten: $f_i = n_i/n$

2.1.1 Kreisdiagramm

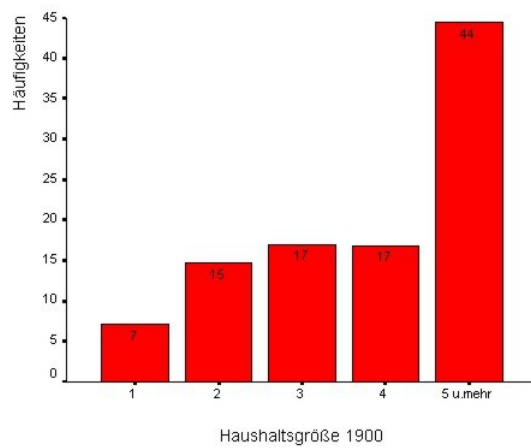
Gesamtwinkel wird entsprechend den absoluten oder relativen Häufigkeiten aufgeteilt.

**Verteilung der weltweiten Erdölreserven im Jahr 2004
(10 erdölreichste Länder)**



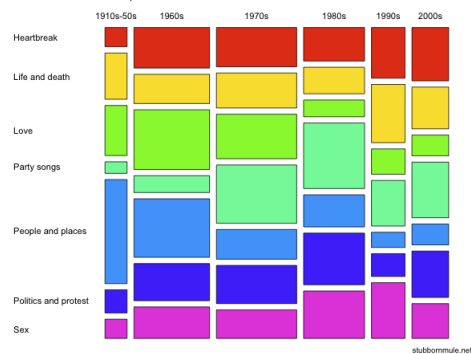
2.1.2 Balkendiagramm

Grafische Darstellung der absoluten (oder relativen) Häufigkeiten mit senkrechten Balken.



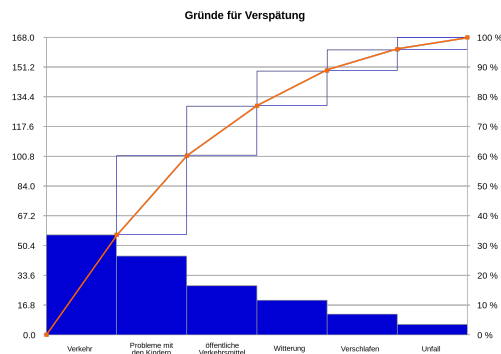
2.1.3 Mosaikplot

multivariat, schnell unübersichtlich



2.1.4 Pareto-Diagramm

wird auch 80/20 Regel genannt. Ähnlich Balkendiagramm. Wird häufig im Qualitätsmanagement verwendet.



2.2 Stetige univariate Merkmale

Ordnungsstatistiken: erster Schritt in Aufbereitung Sortierung nach Größe. Sind alle Werte verschieden -> Nummer i in der Anordnung = Rangzahl. Sind Beobachtungen identisch spricht man von einer **Bindung** und teilt allen Werten eine mittlere Rangzahl zu. Wird jede Beobachtung durch ihre Rangzahl ersetzt, spricht man von der Rangtransformation. (verzichtet auf einen Teil - metrische Informationen - der Informationen)

Rangtransformation spielen große Rolle in nichtparametrischen Statistik.

2.2.1 Empirische Verteilungsfunktion

Eine Funktion von grundlegender Bedeutung, die die Anzahl der Beobachtungen $\leq x$ zählt und durch die Gesamtzahl dividiert.

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(x_i)$$

Dabei bezeichnet I die Indikatorfunktion, welche 1 ist falls x in A und 0 wenn nicht.

Die empirische Verteilungsfunktion ist also eine Treppenfunktion mit Sprüngen an den Stellen $x(i)$ der Höhe $1/n$ Gültig sind bei Sprüngen jeweils nur die oberen Punkte.

2.2.2 Stem- and Leaf-Plot

bei kleineren Datensätzen. Vorderen Dezimalstellen = Stamm, hinteren die Blätter.

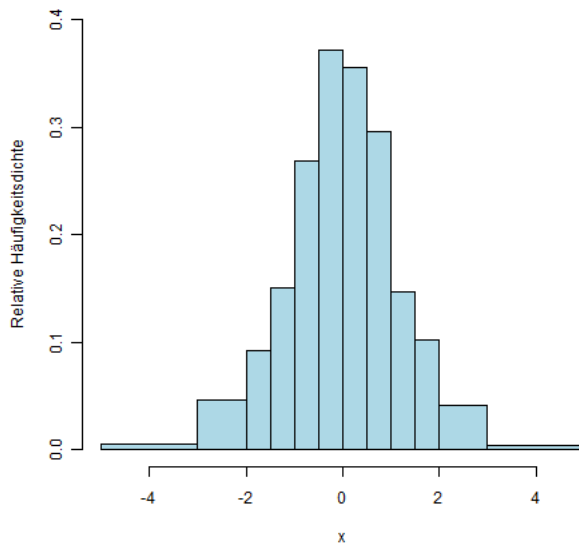
2.2.3 Klassierung

Bei größeren Stichproben Klassenbildung sinnvoll. Nicht eindeutig festgelegt, darauf zu achten, dass der gesamte Wertebereich überdeckt wird und jede Beobachtung eindeutig zuordbar.

Klassierung mit äquidistanten (abstandsgleich) Klassenbreiten zu bevorzugen. Spannweite dividiert durch \sqrt{n} oder 20 für größere Anzahl an Beobachtungen. Eine andere gängige Regel ist die Surges Rule.

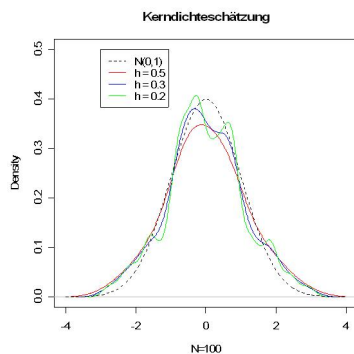
2.2.4 Histogramm

grafische Darstellung relativer Häufigkeitsverteilung, basierend auf einer Klassierung der Daten. Man sollte sich an das **Prinzip der Flächentreue** halten, bei welchem die Summe der Rechtecksflächen genau Eins beträgt. Höhe = relative Häufigkeit durch Breite der Klasse. Beachte: Nicht zu große, nicht zu kleine Klassen. $f(x)$ ist eine Dichte. (Dichte histogramm) Die Rechtecksfläche repräsentiert die relative Häufigkeit und die Höhe repräsentiert die Dichte.



2.2.5 Kernschätzung

Nehme keine feste Klasseneinteilung, sondern jedem x_i quasi ein eigenes Fenster, durch welches man auf einen Teil der Daten blickt. -> Konzept der Kerndichteschätzung. Durch Überlagerung aus den einzelnen Beobachtungen werden Kernfunktionen aufgebaut. (Jede Beobachtung = 1 Kern) **Kernfunktion** = symmetrische Funktion um Null, deren Fläche Eins ist. **Kerndichteschätzung** = stetige, glatte Funktion (Im Gegensatz dazu Histogramme stufig). Gebräuchliche Kernfunktionen (Rechteck, Dreieck, Normal). Vorteile: glatt, stetig, Brauchen keine Klasseneinteilung, realistischer. Eine **Bandbreite** muss gewählt werden, diese bestimmt Glattheit. Je größer, desto träger reagiert die Schätzung auf die Beobachtungen. Kritischer als die Wahl der Kernfunktion.



2.2.6 Quantile (Kenngröße der Lage)

Im wesentlichen Inverse zur emp. Verteilungsfunktion. Grob ein Wert der den Datensatz im Verhältnis $p : (1-p)$ teilt. Wichtige Quantile: Median und Quartile.

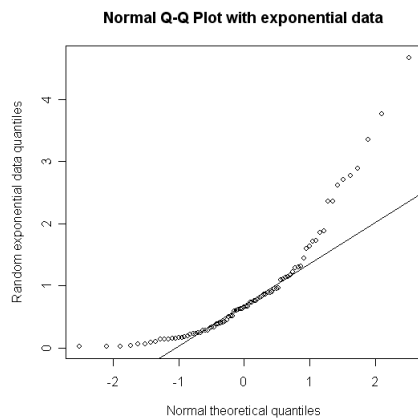
- Typ-1 Quantile: verallgemeinerten Inversen der emp. Verteilungsfkt.
- Typ-2 wie Typ1 allerdings bei Unstetigkeiten gemittelt, d.h auch Werte in der Mitte zw. Datenpunkten möglich
- Typ-4: lineare Interpolation der emp. Verteilungsfkt. Somit alle Werte zugelassen
- Typ-7: Ähnlich Typ 4, allerdings wird Intervall $[0,1]$ in $n-1$ Teilintervalle unterteilt

Wichtige Quantile:

- Median: das 50 Prozent Quantil. Teilt den Datensatz in zwei gleich große Hälften.
- Quartile: teilen den Datensatz in vier gleich große Stücke.
- Hinges: untere Hinge ist der Median der ersten Hälfte, der obere Hinge der Median der zweiten Hälfte. Bei ungerader Anzahl zählt der Median zu beiden Hälften.

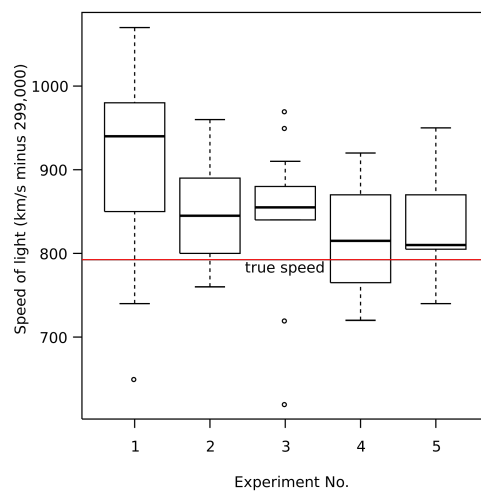
2.2.7 QQ-Plot

ist eine Art Streudiagramm zum Vergleich zweier Datensätze. Bei unterschiedlich großen Stichproben -> Reduzierung.



2.2.8 Boxplot

Graf. Darstellung eines Datensatzes auf Basis der Quartile. Somit mehrere Datensätze leicht vergleichbar. Box = Rechteck vom 1 zum 3 Quartil. Linie = Median. Fences = Einzeunungen Whiskers, Linien vom Rand der Box bis zu Fences. Außreiser extra eingezeichnet.



2.3 Kennzahlen

Neben graf. Aufbereitung, Berechnung von Stichprobenparametern wichtige Ergänzung. Wichtig ist die Robustheit der Kennzahlen. Es gibt Kennzahlen der Lage, der Streuung und der Verteilungsform

2.3.1 Mittelwert:

das Arithmetische Mittel. Bezeichnet als \bar{x} . **Minimumseigenschaft:** Mittelwert gibt immer eine minimalere Abweichung als eine beliebige Zahl c .

Gewichteter Mittelwert: Gewichtet nach Zahl der Beobachtungen. Bei Klassen berechnet man den Mittelwert als gewichtetes Mittel der Klassenmitten. Je nach Verteilung innerhalb der Klassen kann dieser vom tatsächlichen Mittelwert abweichen. Geometrisches Mittel: n -te Wurzel des Produkts.

Harmonischer Mittelwert. N durch Summe der Wurzeln. Bei Werten mit vielen Ausreißern -> getrimmtes Mittel, da Mittelwert sehr sensibel. Median am robustesten.

2.3.2 Bruchpunkt:

Robustheit in Bezug auf Ausreißer lässt sich durch seinen Bruchpunkt bemessen. = kleinsten Anteil der Datenwerte, den man ersetzen müsste, um den Schätzwert beliebig zu verändern.

2.3.3 Median:

Teilt Datensatz in zwei Hälften. Bruchpunkt ca. 50 Prozent. Erfüllt ebenfalls Minimumseigenschaft

2.3.4 Varianz:

Kennzahlen für die Charakterisierung des Streuverhaltens einer Verteilung. S^2 : Summe der quadratischen Abweichungen der Daten von ihrem Mittelwert.

Standardabweichung ist die Wurzel aus der Varianz.

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

Verwendet man keine Stichprobe, sondern eine Gesamtpopulation verwendet man $\frac{1}{n}$

statt $\frac{1}{n-1}$

Verschiebungssatz: Die Varianz lässt sich mittels des Verschiebungssatzes auch so berechnen:

$$s_n^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n(\bar{x}_n)^2 \right]$$

Diese Darstellung ist zu bevorzugen, da sich häufig auftretende Rundungsfehler hier weniger stark auswirken.

2.3.5 MAD:

Verwende Median um folgende Abstände zu bilden: $|x_1 - \text{Median}|, |x_2 - \text{Median}|, \dots$ Mittelwert dieser Abstände, genannt mittlere absolute Abweichung ist ein Streuungsmaß. Nicht sonderlich Robust, allerdings Robuster als Varianz. Alternativ MedMed (Statt Mittelwert, Median)

2.3.6 Modalwert:(Modus)

Nicht immer eindeutig. Bezugspunkt bei der Beurteilung der Form einer Verteilung. Merkmalsausprägung mit höchster Dichte. (höchster Beobachtungshäufigkeit). Betrachtet oft Histogramme. Manchmal auch multimodal.

2.3.7 Momente:

Für einen Datensatz x_1, x_2, \dots, x_n ist das (empirische) Moment der Ordnung r um den Nullpunkt definiert durch:

$$m'_r = \frac{1}{n} \sum_{i=1}^n x_i^r$$

Kurz nennt man m' einfach das r -te Moment der Daten. Bildet man die Momente um den Mittelwert, bekommt man die zentralen Momente. Varianz ein Moment zweiter Ordnung, Mittelwert ein Moment 1. Ordnung.

2.3.8 Schiefe:

Kann auch negativ sein. Um die Schiefe zu charakterisieren, kann man sich der Momente der Ordnung 3 und 2 bedienen. Oder berechne durch Quartile.

$$g_1 = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

$g_1 > 0$ = linksteil / rechtsschief

$g_1 = 0$ symmetrisch

$g_1 < 0$ rechtsteil/linksschief

2.3.9 Kurtosis

Immer positiv. Ziehe Momente der Ordnung 2 und 4 heran.

2.4 Mehrdimensionale Daten

Messung für mehrere Merkmale. Lassen sich in Form einer Datenmatrix zusammenfassen. (Datenframes)

Scatterplots: Im Fall von zwei Merkmalen kann man die Beobachtungspaare als Punkte interpretieren und als Scatterplot darstellen. Durch Überladen der Punkte (Farbe, Größe, ec.) können weitere Merkmale repräsentiert werden. Konzept der Kerndichteschätzung lässt sich auf den Fall mehrdimensionaler Beobachtungen erweitern. Der einfachste multivariate Kernschätzer basiert auf dem Produktkern.

2.4.1 Korrelation

Maß für Zusammenhang. \Rightarrow Korrelationskoeffizient misst den Grad der linearen Assoziation zwischen zwei Merkmalen. Betrachte Abstände vom Mittelwert der beiden Datensätze und bilde Produkte.

Gibt es eine **positive Assoziation**, werden diese Produkte großteils positive sein. (Da y-Werte die größer als ihr Durchschnitt sein, meist mit x Werten, die größer als ihr Durchschnitt sind, auftreten. Im Falle einer negativen Assoziation werden die Produkte großteils negativ sein. Der Korrelationskoeffizient ist nun der Durchschnitt dieser Produkte. Lässt sich auch so berechnen:

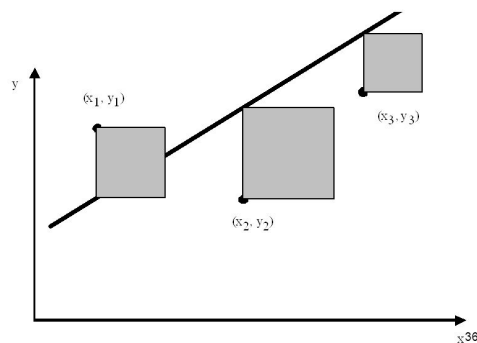
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Der Absolutwert von r sagt etwas über die Stärke der Assoziation. Für $r = 1$ liegen alle Punkte exakt auf einer Geraden. Spearmnsche Rangkorrelation. Robustes Korrelationsmaß, wegen Verzicht auf metrische Information und alleinige Verwendung der Ränge.

2.4.2 Kleinste Quadrate

Gerade die unter allen möglichen Geraden eine bestimmte Optimalitätseigenschaft aufweist. Es wird eine Kurve gesucht die möglichst nah an den Datenpunkten verläuft. Die Methode der kleinsten Quadrate besteht dann darin, die Kurvenparameter so zu bestimmen, dass die Summe der quadratischen Abweichungen der Kurve von den Punkten minimiert wird.

Prinzip: Kleinste Quadrate



3 Wahrscheinlichkeit

3.1 Gesetz der großen Zahlen

Man zählt, wie oft ein bestimmtes Ereignis A bei wiederholter Durchführung eines Zufallsexperiments eingetreten ist. N Wiederholungen eines Zufallsexperiments \rightarrow fragliches Ereignis tritt $H_n(A)$ -mal auf (Häufigkeit) so erhält man den Eindruck, dass sich

die relative Häufigkeit $h_n := H_n(A)/n$ von A einem Grenzwert nähert. Dieser Grenzwert $P(A)$ wird als Wahrscheinlichkeit betrachtet.

Die frequentistische Interpretation von Wahrscheinlichkeit beruht darauf, dass ein (statistisches) Experiment unter identischen Bedingungen beliebig oft wiederholbar ist.

3.2 Merkmalraum:

möglichen Ergebnisse von statistischen Experimenten lassen sich in einer Grundmenge zusammenfassen. Menge aller möglichen Versuchsausgänge = **Merkmalraum**.

$\Omega = \{ \omega \mid \omega \text{ möglicher Versuchsausgang} \}$

Dieser kann endlich, unendlich (abzählbar unendlich, überabzählbar), ein, mehrdimensional etc. sein.

3.3 Ereignisse:

Eine Teilmenge $A \subset \Omega$ nennt man Ereignis. Gilt für einen Versuchsausgang $\omega \in \Omega$ so sagt man, dass A eingetreten ist. Alle Ereignisse werden in einem Ereignissystem zusammengefasst.

3.4 Ereignissysteme:

haben algebraische Struktur (σ -Algebra) um auch komplexere Ereignisse betrachten zu können. Folgende Eigenschaften müssen erfüllt werden:

1. $\Omega \in A$ (Merkmalraum selbst ist ein Ereignis)
2. A ist abgeschlossen unter Komplementbildung
3. A ist abgeschlossen unter abzählbaren Vereinigungen

Die leere Menge ist ebenfalls in A und A ist auch abgeschlossen gegenüber abzählbaren Durchschnitten. Das Paar Ω und A nennt man auch Messraum. Ist der Merkmalraum endlich oder abzählbar unendlich, wählt man als Ereignissystem stets die **Potenzmenge**.

3.5 Borelmengen

Für überabzählbare Merkmalräume ist die Potenzmenge zu umfangreich. Um eine σ -Algebra zu erhalten erzeugt man die kleinste σ -Algebra, die alle Elemente umfasst. -> Borel σ -Algebra, welche eine echte Teilmenge der Potenzmenge ist. (Alle Teilmengen von \mathbb{R} , die man sich vorstellen kann -> Borelmengen. Analog lässt sich die Borel σ -Algebra auch für den mehrdimensionalen Fall definieren.

3.6 Wahrscheinlichkeitsmaße

Eine Abbildung $P : A \rightarrow \mathbb{R}$ heißt ein Wahrscheinlichkeitsmaß wenn sie folgende Eigenschaften erfüllt:

1. $P(A) \geq 0$ für alle A
2. $P(\Omega) = 1$
3. Wahrscheinlichkeit für die Summe der Ereignisse ist die selbe wie die Summe der Einzelwahrscheinlichkeiten (σ - Additivität)

Auf Basis dieser Definition lässt sich der Messraum zu einem Wahrscheinlichkeitsraum (Ω, \mathcal{A}, P) erweitern.

Die obigen Axiome haben eine Reihe von Konsequenzen.

- $P(\emptyset) = 0$ (leere Menge ist das unmögliche Ereignis)
- $P(\Omega) = 1$ (das sichere Ereignis)
- $P(A^c) = 1 - P(A)$
- $A \subset B \Rightarrow P(A) \leq P(B)$

Chancen(Odds): Quotient aus der Wahrscheinlichkeit und der Gegenwahrscheinlichkeit

3.7 Endliche W-Räume:

Ein W Raum heißt endlich, wenn der Merkmalraum Ω eine endliche Menge ist. Die Wahrscheinlichkeit eines beliebigen Ereignisses A ergibt sich dann durch Addition der Einzelwahrscheinlichkeiten.

Wenn die Elementarereignisse alle gleich wahrscheinlich sind nennt man das Experiment Laplace-Experiment. Die Wahrscheinlichkeit eines beliebigen Ereignisses A ist dann gegeben durch

$$P(A) = \frac{\text{Anzahl der Elemente von } A}{\text{Anzahl der Elemente von } \Omega}$$

(günstige durch mögliche Fälle)

3.7.1 Geometrische Wahrscheinlichkeiten

lässt sich anwenden, wenn der Merkmalraum als geometrisches Objekt und Ereignisse als Teilbereiche dieses Objekts interpretiert werden können, deren Wahrscheinlichkeit proportional zur Größe des Teilbereichs ist.

$$P(A) = \frac{\text{Grösse von } A}{\text{Grösse von } \Omega}$$

3.8 Additionstheorem

Addiere die Einzelwahrscheinlichkeiten, ziehe paarweise Durchschnitte ab und zähle dreifach Durchschnitte dazu usw. Betrachtet man die Teilsummen ergeben sich wechselweise **Abschätzungen** nach oben und unten. Die erste dieser Abschätzungen nennt

man **Boole'sche Ungleichung**.

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i)$$

3.9 Bedingte Wahrscheinlichkeiten

Die Kenntnis, dass ein Ereignis B eingetreten ist, kann informativ für das (mögliche) Eintreten von A sein und die Eintrittswahrscheinlichkeit ändern.

Man schreibt $P(A|B)$ (A wenn B eingetreten ist) und spricht von der bedingten Wahrscheinlichkeit. Wenn B eingetreten ist, sind nur noch diejenigen Versuchsausgänge relevant, die auch in B liegen.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bedingte Wahrscheinlichkeiten erfüllen alle Eigenschaften eines W-Maßes.

3.10 Multiplikationstheorem

Multipliziert man in der Definition der bedingten Wahrscheinlichkeit beide Seiten mit $P(B)$ so ergibt sich:

$$P(A \cap B) = P(A|B)P(B)$$

Vertauschen von A und B ergibt:

$$P(B \cap A) = P(B|A)P(A)$$

Eine Aussage dieser Art nennt man Multiplikationstheorem.

$$P\left(\bigcap_{i=1}^n A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\dots P(A_n|A_1A_2\dots A_{n-1})$$

Vorteil: komplizierte Probleme in mehrere einfachere Teilprobleme zerlegt.

3.11 Vollständige Wahrscheinlichkeit (totale Wahrscheinlichkeit)

Ermöglicht Wahrscheinlichkeiten von komplizierten Ereignissen aus Wahrscheinlichkeiten von einfacheren Ereignissen zusammenzusetzen. Grundlegend ist dabei die Zerlegung des Merkmalraums in paarweise disjunkte Ereignisse $C_1, C_2, \dots \in A$ (also eine höchstens abzählbare Partition von Ω)

$$\Omega = \bigcup_{i=1}^{\infty} C_i$$

Berechne Wsl. von A in allen Teilbereichen des Merkmalraums und summiere diese

auf.

$$P(A) = \sum_{i=1}^{\infty} P(A|C_i)P(C_i)$$

3.12 Bayes'sche Formel

Hat man eine Partition C_1, C_2, \dots des Merkmalraums und kennt für ein Ereignis A die bedingten Wahrscheinlichkeiten $P(A|C_i)$ -> wie berechne inverse bedingte Wsl. $P(C_i|A)$. -> Bay'sche Formel:

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{\sum_{j=1}^{\infty} P(A|C_j)P(C_j)}$$

Ergibt sich aus dem Multiplikationssatz

3.12.1 Odds-Form der Bay'schen Formel

einer Hypothese H ist das Verhältnis der Wahrscheinlichkeit von A bedingt durch H und bedingt durch H^c . Um die A posteriori-Odds zu erhalten, muss man also nur die A-priori-Odds mit den Likelihood Quotienten multiplizieren.

$$\frac{P(H|A)}{P(H^c|A)} = \frac{P(H)}{P(H^c)} * \frac{P(A|H)}{P(A|H^c)}$$

3.13 Unabhängigkeit:

Die bedingte Wahrscheinlichkeit $P(A|B)$ ist nicht gleich der unbedingten Wahrscheinlichkeit $P(A)$. Wenn bekannt ist, dass B eingetreten ist, verändert sich in der Regel die Wahrscheinlichkeit für den Eintritt von A . Gilt allerdings $P(A|B) = P(A)$, so hat der Eintritt von B keinen Einfluss, in diesem Fall sagt man, dass A und B unabhängig sind. Zwei Ereignisse sind unabhängig wenn

$$P(AB) = P(A)P(B)$$

→ Sind A und B ua., so sind auch die Kombinationen mit den Komplementen ua. Unabhängigkeit von drei Ereignissen ist gegeben wenn:

$$P(ABC) = P(A)P(B)P(C)$$

$$P(AB) = P(A)P(B)$$

$$P(AC) = P(A)P(C)$$

$$P(CB) = P(C)P(B)$$

Gelten nur die letzten drei Gleichungen, nennt man die Ereignisse paarweise unabhängig.

Unabhängigkeit \neq Disjunktheit

Nur Ereignisse, die etwas gemeinsam haben können auch unabhängig sein!

3.14 Mehrstufige Experimente

Bedingte Wahrscheinlichkeiten kommen auf natürliche Weise ins Spiel. Ablauf lässt sich durch einen **Wahrscheinlichkeitsbaum** repräsentieren. Die Wahrscheinlichkeiten berechnet man dann mit dem Satz von der vollständigen Wahrscheinlichkeit. Einen geeigneten Merkmalraum erhält man durch das kartesische Produkt aller Merkmalräume aller n Stufen. Um die oft komplexen W-Räume dieser Art zu vereinfachen, ist oft die Annahme gerechtfertigt, dass die bedingten Wahrscheinlichkeiten nur vom vorherigen Zustand abhängen.

4 Stochastische Größen und Verteilungen

Modellierung eines Zufallsexperiments mittels W-Raum ist nicht immer notwendig. Mathematisch betrachtet handelt es sich um eine Abbildung von Ω nach \mathbb{R} . Eine Abbildung dieser Art nennt man eine stochastische Größe (sG), die jedem $\omega \in \Omega$ eine reelle Zahl $X(\omega) = x$ zuordnet. Der Wertebereich (oder Merkmalraum) von X wird mit M_x bezeichnet.

$$M_x = \{\omega | x = X(\omega), \omega \in \Omega\}$$

X bezeichnet die sG (d.h. Die Abbildung) und x eine Realisation der sG (d.h. Ein konkreter Funktionswert). Eine Abbildung ist messbar, wenn das Urbild jeder Borelmenge $B \in \mathcal{B}$ ein Element von \mathcal{A} ist.

Verteilung von X : Ist X eine sG und A ein Ereignis in \mathcal{B} , so ist das Urbild $X^{-1}(A)$ ein Ereignis in \mathcal{A} .

P_x nennt man die Verteilung von X .

Neben punktförmigen Ereignissen spielen vor allem Ereignisse der Form $A = (a, b]$ eine große Rolle.

$$P_x(A) := P(X \in A) = P(X^{-1}(A))$$

Für **punktförmige** Ereignisse gilt:

$$P_x(x) = P(X = x) = P(X \leq x) - P(X < x)$$

Für **Intervalle** gilt:

$$P_X(a, b) = P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

4.1 Verteilungsfunktion

Die VF F_X einer sG X ist definiert durch: $F_X(x) := P(X \leq x), x \in \mathbb{R}$ und hat folgende Eigenschaften:

1. $0 \leq F(x) \leq 1$
2. F ist monoton wachsend d.h. aus $x < y$ folgt $F(x) < F(y)$
3. $\lim_{x \rightarrow -\infty} F(x) = 0$ und $\lim_{x \rightarrow \infty} F(x) = 1$

4. F ist rechtsstetig, d.h. bei Unstetigkeit gilt immer der obere Punkt

Sei X eine sG mit Verteilungsfunktion F_X , dann gilt für $a < b$:

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

$$P(X = x) = F(x) - F(x-) \text{ (linksseitiger Grenzwert)}$$

Dieser Wert entspricht der Höhe des Sprungs

4.1.1 Inverse

Vielfach benötigt man die Umkehrfunktion einer VF F . Da aber eine VF nicht notwendigerweise streng monoton wächst, muss man die Definition der Inversen modifizieren. Ist F streng monoton wachsend, entspricht F^{-1} der Umkehrfunktion von F .

Quantilenfunktion: $x_p = F^{-1}(p)$ für $p \in (0, 1)$ dabei ist F^{-1} die (verallgemeinerte) Inverse von F . Allgemein nennt man für ein festes p den Wert x_p das p -Quantil von F . Die Quantilenfunktion geht aus F_x durch Spiegelung an der 1. Mediane hervor.

4.2 Diskrete Verteilung

Eine sG hat eine diskrete Verteilung, wenn ihr Merkmalraum aus einer endlichen oder abzählbaren Menge von Punkten besteht. (Punktwahrscheinlichkeiten) Die Wahrscheinlichkeit für eine Teilmenge B wird wie folgt berechnet:

$$P(X \in B) = \sum_{x \in B} p_X(x)$$

Die Verteilungsfunktion einer diskreten sG ist eine Treppenfunktion mit Sprüngen der Höhe der Wahrscheinlichkeiten $p_X(x)$ an den Stellen x . Die Verteilungsfunktion berechnet sich aus allen Punktwahrscheinlichkeiten:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p_X(x_i)$$

4.3 Stetige Verteilung

Eine stochastische Größe X hat eine stetige Verteilung wenn die Verteilungsfunktion eine stetige Funktion auf \mathbb{R} ist. Punktwahrscheinlichkeiten sind bei stetigen Verteilungen gleich 0.

Die meisten stetigen Funktionen sind absolut stetig, d.h. Es gibt eine Funktion f_X sodass:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt$$

Eine Funktion mit dieser Eigenschaft nennt man **Dichtefunktion** von X .

1. $f_X(x) \geq 0$

2. $\int_{-\infty}^{\infty} f_X(t) dt = 1$

Die Menge S_X aller Punkte mit $f(x) > 0$ nennt man Träger von X . (Die Punkte wo die Dichte echt positiv ist) Um die Wahrscheinlichkeit eines Intervalls zu erhalten muss

man von Punkt a bis b integrieren.

$$P(a < X \leq b) = \int_a^b f_X(t) dt$$

4.4 Gemischte Verteilungen

In vielen Situationen sind Verteilung weder (rein) diskret noch (rein) stetig. (Keine Treppenfunktion, aber auch nicht überall stetig.) => gemischte Verteilung

$$F(x) = \alpha F_d(x) + (1 - \alpha) F_s(x)$$

Allgemeiner: Verteilungsfkt. F = Mischung von $m (m \geq 2)$ Verteilungen wobei mind. eine diskret und eine stetig ist.

Merkmalraum einer gemischten Verteilung:

$$M = \{x_1, x_2, \dots, x_k\} \cup \langle a, b \rangle$$

wobei $\langle a, b \rangle$ ein Intervall ist. Die diskreten Punkte x_i haben positive Wahrscheinlichkeiten und es gibt eine Dichte mit Träger $\langle a, b \rangle$ sodass:

$$\sum_{i=1}^m p(x_i) + \int_a^b f^*(x) dx = 1$$

(f^* keine vollständige Dichte) Wichtige Anwendung von gemischten Verteilungen: Berechnung von Lebensdauern, wenn Beobachtungen unvollständig sind.

4.5 Transformationen

Manchmal interessiert man sich für eine Transformation $Y = g(X)$, wobei g eine Funktion von \mathbb{R} nach \mathbb{R} ist. Da Y dann wieder eine sG ist, stellt sich die Frage nach ihrer Verteilung

4.5.1 Transformationen diskreter sGn

Ist g eine umkehrbar eindeutige Funktion lässt sich die W-Funktion von Y wie folgt bestimmen:

$$p_Y(g^{-1}(y))$$

4.5.2 Transformationen stetiger sGn

Die Verteilungsfunktion von $Y = g(x)$ lässt sich mittels der **Methode der Verteilungsfunktion** bestimmen.

Die Verteilungsfunktion von Y lässt sich durch die Verteilungsfunktion von X ausdrücken.

Falls die Funktion g umkehrbar eindeutig ist, kann die Dichte von $Y = g(X)$ mit Hilfe des folgenden Satzes auch direkt bestimmt werden.

4.5.3 Transformationssatz für Dichten

X sei eine stetige sG mit Dichte f_X und Träger S_X und g sei eine umkehrbar eindeutige differenzierbare Funktion. Dann ist die Dichte von Y gegeben durch:

$$f_Y(y) = f_X(g^{-1}(y)) \left(\text{Umkehrabbildung} \right) \left| \frac{dg^{-1}(y)}{dy} \right| \quad (\text{innere Ableitung})$$

Jacobian: Die Ableitung $dx/dy = dg^{-1}(y)/dy$ der Umkehrabbildung nennt man die Jacobian J .

Somit ist laut dem Transformationssatz für Dichten die Umkehrfunktion der Dichte mal dem Jacobian.

Wichtige Spezialfälle sind affine Transformationen der Form $Y = a + bX$ hier ist die Dichte:

$$f_Y(y) = 1/|b| f_X\left(\frac{y-a}{b}\right)$$

4.6 Erwartungswert

(Lageparameter einer Verteilung)

Verteilungsfunktion \Rightarrow gesamt verfügbare Wahrscheinlichkeitsinformation über eine sG X . In vielen Situationen genügen wenige Werte. Eine dieser Werte: Erwartungswert (auch Mittelwert) von X .

4.6.1 Erwartungswert einer diskreten sG

gewichteter Mittelwert

$\mathbb{E}(X) = \sum x p(x)$ wobei $p(\cdot)$.. W-Funktion und x die Ausprägung.

Erwartungswert als Schwerpunkt: Lässt sich als Schwerpunkt von Punktemassen interpretieren.

4.6.2 Erwartungswert einer stetigen sG

$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$... gewichtete Dichte. Der einer gemischten Verteilung ist die Kombination der beiden.

4.6.3 Erwartungswert einer Funktion von X

Die sG $Y = g(x)$ sei eine Funktion der sG X .

- Ist X diskret mit W-Funktion $p(x)$ ist der \mathbb{E} gegeben durch:

$$\mathbb{E}(Y) = \sum_{x \in S_x} g(x) p_X(x)$$

- Ist X stetig mit Dichte $f(x)$ dann ist der \mathbb{E} gegeben durch:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

→ LotUS

4.6.4 LotUS(Satz des unbewussten Statistiker)

Kernaussage: Zur Berechnung des Erwartungswertes einer Funktion $Y = g(X)$ muss man nicht zuerst die Verteilung von Y bestimmen, sondern kann auf die Verteilung von X zurückgreifen.

Alternativ kann man aber auch zuerst die Dichte von Y bestimmen (mittels Jacobian der Transformation und dem Transformationssatz)

Der Erwartungswert ist dann:

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} yf_Y(y)dy$$

Eigenschaften des Erwartungswerts:

- $\mathbb{E}(a) = a$ (a = Konstante)
- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
- $\mathbb{E}[k_1g(X) + k_2h(X)] = k_1\mathbb{E}[g(X)] + k_2\mathbb{E}[h(X)]$

4.7 Varianz

(Maßzahl für die Streuung) (immer positiv, da quadratisch) Mittelwert muss existieren.

$Var(X) = \mathbb{E}(X - \mu_X)^2$... mittlere quadratische Abweichung vom Mittelwert

Die positive Wurzel aus der Varianz nennt man die **Streuung** (oder Standardabweichung)

Für die einzelnen Fälle ist die Varianz gegeben durch:

diskret:

$$Var(X) = \sum_x [x - \mathbb{E}(X)]^2 p_X(x)$$

stetig:

$$Var(X) = \int_{-\infty}^{\infty} [x - \mathbb{E}]^2 f_x(x)dx$$

gemischt:

$$Var(X) = \sum_{i=1}^m [x_i - \mathbb{E}(X)]^2 p_X(x_i) + \int_{-\infty}^{\infty} [x - \mathbb{E}]^2 f^*(x)dx$$

Mittels **Verschiebungssatz**:

$$Var(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = \mathbb{E}(X^2) - \mu^2$$

Da Varianz positiv folgt daraus Ungleichung:
 $\mathbb{E}(X^2) \geq [\mathbb{E}(X)]^2$ Eigenschaften der Varianz:

1. $\text{Var}(a) = 0$
2. $\text{Var}(aX + b) = a^2 \text{Var}(X)$
3. $\sigma_{aX+b} = a\sigma_X$

4.8 MAD

mittlere absolute Abweichung vom Median

$$\text{MAD}(X) = \int_{-\infty}^{\infty} |x - \text{med}(X)| f_X(x) dx$$

4.9 Simulation

Erste und schwierigste Schritt: Erzeugung von Zufallszahlen uniform verteilt auf (0,1)
 Generatoren: Starte mit Anfangswert (Seed) und berechne rekursiv Werte. Haben Periode (ab der m-ten Zufallszahl wiederholt sich exakt die gleiche Folge).

Nächste Frage: Wie können Realisationen für eine beliebige sG X erzeugt werden?:

F ... beliebige Verteilungsfunktion $X = F^{-1}(U) \sim F$

F^{-1} ... verallgemeinerte Inverse von F .

Nennt man **Inversionsmethode**: Für die Erzeugung einer Realisation x einer sG $X \sim F$ genügen die folgenden Schritte:

1. Erzeuge eine Realisation u von U (uniform)
2. Bilde $x = F^{-1}(u)$

Inversionsmethode lässt sich grundsätzlich auch im diskreten (oder gemischten) Fall anwenden.

5 Spezielle Verteilungen

5.1 Diskrete Verteilungen

5.1.1 Diskrete uniforme Verteilung (Gleichverteilung)

Wenn jedes Element von M die gleiche Wahrscheinlichkeit hat.

$$\mathbb{E}(X) = \sum_{x \in M} xp(x) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \text{ (Arithmetisches Mittel der mögl. Ausprägungen)}$$

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Es existiert ein Spezialfall wenn der Merkmalraum M aus aufeinanderfolgenden ganzen Zahlen besteht. In diesem Fall ändert sich die Formel für den Mittelwert und der Varianz.

5.1.2 Bernoulli-Verteilung

Wenn man nur beobachtet, ob ein bestimmtes Ereignis A eintritt oder nicht. Die sG ist dann ein Indikator für den Eintritt von A. (1 = A tritt ein, 0 = A tritt nicht ein)

Wsl- Funktion ist gegeben durch

$$p(x) = p^x(1-p)^{1-x}$$

Erwartungswert: p (Erfolgswsl.)

Varianz: $p(1-p)$ (Verschiebungssatz Erfolgswsl * Misserfolgswsl.)

5.1.3 Binomialverteilung

(= Erweiterung der Bernoulli Verteilung)

Werden n unabhängige und identische Bernoulli Experimente durchgeführt, so ist das Ergebnis ein n -Tupel aus Nullen und Einsen. Häufig ist man aber nur an der **Anzahl der Erfolge** interessiert und nicht der Reihenfolge ihres Auftretens. Gibt es x Erfolge (und daher $n-x$ Misserfolge) so hat man $\binom{n}{x}$ verschiedene Möglichkeiten, die Position für die x Erfolge zu wählen. Die Wsl für jede dieser Möglichkeiten beträgt $p^x(1-p)^{n-x}$
=> die W-Funktion für X:

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

Man schreibt $X \sim B(n, p)$

$$\mathbb{E}(X) = np$$

$$\text{Var}(X) = np(1-p)$$

(Der Spezialfall $n = 1$ entspricht der Bernoulli Verteilung)

Es existieren 1 oder 2 Modalwerte (x-Werte mit $p(x) = \max$)

$$x_{\text{mod}} = (n+1)p \text{ oder } x_{\text{mod}} = (n+1)p - 1, (n+1)p$$

5.1.4 Negative Binomialverteilung

Betrachte Folge von unabhängigen Wiederholungen eines Bernoulli Experiments.

Ist x die Gesamtzahl der Versuche, die notwendig sind um r Erfolge zu bekommen.

$$p(x) = P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \text{ .. negative Binomialverteilung}$$

$$X \sim NB(r, p)$$

$$\mathbb{E}(X) = \frac{r}{p}$$

$$\text{Var}(X) = \frac{r(1-p)}{p^2}$$

5.1.5 Geometrische Verteilung

Die Negative Binomialverteilung für $r = 1$ nennt man Geometrische Verteilung und schreibt $G(p)$. Ist die X die Gesamtzahl der Versuche, die notwendig sind um exakt einen Erfolg zu bekommen:

$$p(x) = P(X = x) = p(1-p)^{x-1}$$

$$\mathbb{E}(X) = \frac{1}{p}$$

$$\text{Var}(X) = \frac{1-p}{p^2}$$

Gedächtnislosigkeit: Erinnert sich nicht an das was voraus gegangen ist. Wsl. für Erfolg auch nach X Verlusten gleich. Die $G(p)$ Verteilung ist die einzige diskrete Verteilung auf $1, 2, \dots$ ohne Gedächtnis.

5.1.6 Hypergeometrische Verteilung

Objekte ohne zurücklegen. Die sG X sei nun die Zahl der Erfolge in der Stichprobe. $X \sim H(N, A, n)$:

$$p(x) = P(X = x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}$$

$$\mathbb{E}(X) = n \frac{A}{N}, \text{Var}(X) = n \frac{A}{N} \left(1 - \frac{A}{N}\right) \frac{N-n}{N-1}$$

Letzteres ist der Korrekturfaktor an wessen die Ziehung ohne Zurücklegen schuld ist. Unter bestimmten Umständen lässt sich also die $H(N, A, n)$ - Verteilung durch die (einfachere) $B(n, p = A/N)$ - Verteilung approximieren.

FAUSTREGEL: Sind A und $N - n$ beide nicht zu klein und ist $n/N \leq 0.05$, so gilt in

guter Näherung: $P(X = x) \sim \binom{n}{x} \frac{A^x}{N^x} \left(1 - \frac{A}{N}\right)^{n-x}$

Anwendung z.B. In der Qualitätskontrolle

5.1.7 Poisson- Verteilung

Unendliche Verteilung basierend auf der Exponentialreihe:

$$1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots \text{ konvergiert gegen } e^\lambda$$

Die W-Fkt lässt sich also wie folgt definieren:

$$p(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Eine Verteilung mit der obigen W Funktion nennt man eine Poisson Verteilung und man schreibt $X \sim P(\lambda)$ Bsp. Die Anzahl X der von einer radioaktiven Substanz emittierten Teilchen in guter Näherung einer $P(\lambda)$ Verteilung. od. Zahl der Lackierungsfehler auf einem Autoblech. Zahl der Verkehrsunfälle in Zeitspanne. (Poisson-Prozesse)

Bedingungen für einen Poisson- Prozess:

1. Proportionalität im Kleinen
2. Nachwirkungsfreiheit (D.h für kleine Intervalle kann die Wsl. des Auftretens vernachlässigt werden)

3. Unabhängigkeit

$$\mathbb{E}(X) = \lambda, \text{Var}(X) = \lambda$$

Gibt es viele unabhängige und identische Bernoulli-Experimente mit kleiner Erfolgswahrscheinlichkeit, so folgt die Zahl der Erfolge in guter Näherung einer Poisson Verteilung (Verteilung der seltenen Ereignisse) Für $n \geq 50 \leq 1/10$ und $np \leq 10$ kann die Binomialverteilung als Poisson Verteilung approximiert werden.

5.2 Stetige Verteilungen (Merkmalraum = Intervall)

5.2.1 Stetige uniforme Verteilung

stetige Gleichverteilung auf dem Intervall (a, b) , wenn die Dichte von X

$$f(x) = \frac{1}{b-a} \text{ für } x \in (a, b) \text{ sonst } 0$$

Man schreibt $X \sim U(a, b)$

Die Verteilungsfunktion ist gegeben durch:

$$f(x) = \begin{cases} 0 & \text{für } x \leq a \\ \frac{x-a}{b-a} & \text{für } a < x < b \\ 1 & \text{für } x \geq b \end{cases}$$

$$\mathbb{E}(X) = \frac{a+b}{2}$$

$$\text{Var}(X) = (b-a)^2/12$$

5.2.2 Exponentialverteilung

Stetige Version der geometrischen Verteilung (Bis erster Erfolg) Eine sG hat eine Exponentialverteilung mit Skalierungsparameter $r > 0$ wenn ihre Dichte:

$$f(x) = \frac{1}{\tau} e^{-x/\tau}$$

man schreibt $Exp(\tau)$

Alternativ kann man sie auch mit $\lambda = \frac{1}{\tau}$ schreiben.

Verteilungsfunktion ist gegeben durch:

$$F(x) = 1 - e^{-\lambda x}$$

$$\mathbb{E}(X) = \frac{1}{\lambda} = \tau, \text{Var}(x) = \frac{1}{\lambda^2} = \tau^2$$

Wie die $G(p)$ Verteilung hat auch die $Exp(\lambda)$ Verteilung kein Gedächtnis.

5.2.3 Gamma- und Chiquadratverteilung

Verallgemeinerung der Exponentialverteilung.

Formparameter $\alpha > 0$ und den Skalierungsparameter $\beta > 0$

Man schreibt $X \sim \text{Gam}(\alpha, \beta)$ wenn ihre Dichte gegeben ist durch

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha}$$

Die Gammafunktion Γ ist eine wichtige Funktion in der Mathematik:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

Die Gammafunktion ist eine Verallgemeinerung der Fakultät auf positive reelle Zahlen
Zwei Spezialfälle:

1. $\alpha = 1$: Exponentialverteilung $\text{Exp}(\lambda)$
2. $\alpha = n/2$ und $\beta = 2$ ergibt sich die Chiquadratverteilung mit n Freiheitsgraden:

Man schreibt $X \sim \chi^2(n)$

$$\mathbb{E}(X) = \alpha\beta, \text{Var}(X) = \alpha\beta^2$$

Für Chiquadr.

$$\mathbb{E}(X) = n, \text{Var}(X) = 2n$$

5.2.4 Normalverteilung

Einer der wichtigsten Verteilungen.

Lageparameter μ und dem Skalierungsparameter σ .

Man schreibt $X \sim N(\mu, \sigma)$ Die Verteilung mit $\mu = 0$ und $\sigma = 1$ nennt man die **Standardnormalverteilung** $N(0, 1)$. Die Dichte dieser wird üblicherweise mit Φ bezeichnet. Für die Verteilungsfkt. der Normalverteilung gibt es keinen expliziten Ausdruck, sie lässt sich allerdings mittels Standardisierung auf die VF der $N(0,1)$ zurückführen. Diese ist ausführlich tabelliert.

Standardisierung: Ist X normalverteilt so ist $Z = \frac{X - \mu}{\sigma}$ die standardisierte Normalverteilung. Somit kann X als affine Transformation dargestellt werden.

$X = \mu + \sigma Z$ mit $Z \sim N(0, 1)$

$\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$

Beziehung zur Chiquadratverteilung:

$$X \sim N(\mu, \sigma^2) \rightarrow \left(\frac{X - \mu}{\sigma} \right)^2 \sim \chi^2(1)$$

5.2.5 F-Verteilung

Große Rolle in der klassischen Statistik. m und n Freiheitsgrade
Man schreibt $X \sim F(m, n)$

$$\mathbb{E}(X) = \frac{n}{n-2}, \text{Var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

Es gilt folgende Symmetriebeziehung:

$$X \sim F(m, n) \iff \frac{1}{X} \sim F(n, m)$$

Die Quantile sind für $p > 0.5$ ausführlich tabelliert. Spielt speziell in der Regressionsanalyse und der Varianzanalyse eine große Rolle.

5.2.6 t-Verteilung

n Freiheitsgrade Man schreibt $X \sim t(n)$ Dichte konvergiert für wachsende Freiheitsgrade gegen die Dichte der Standardnormalverteilung
t-Dichten haben schwerere Ausläufer als die Normaldichte. $t(1)$ -Verteilung nennt man auch Cauchy Verteilung $C(0, 1)$

$\mathbb{E}(X) = 0$ (wegen schwerer Ausläufer und $\text{Var}(X) = \frac{n}{n-2}$ Beziehung zur F-Verteilung:

$$X \sim t(n) \implies X^2 \sim F(1, n)$$

5.2.7 Betaverteilung

$Be(a, b)$ Formparametern a und b . Durch die beiden Formparameter zeigt die $Be(a, b)$ Verteilung eine große Formenvielfalt. Die Dichte der Betaverteilung lässt sich wie folgt schreiben:

$$f(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}$$

wobei die Betafunktion:

$$B(a, b) = \int_0^1 u^{a-1} (1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\mathbb{E}(X) = \frac{a}{a+b}, \text{Var}(X) = \frac{ab}{(a+b+1)(a+b)^2}$$

Symmetrie: $X \sim Be(a, b) \implies 1-X \sim Be(b, a)$

Beziehung zur F-Verteilung:

$$X \sim Be(m, n) \implies \frac{X/m}{(1-X)/n} \sim F(2m, 2n)$$

6 Multivariate Verteilungen

Häufig Beschreibung von Zufallsexperimenten mehrerer stochastische Größen. Müssen auf den Zusammenhang (oder das Fehlen desselben) zwischen den Größen beschreiben. D.h. Wir benötigen die gemeinsame Verteilung.

6.1 Bivariate Verteilungen

Merkmalraum Ω und zwei stochastische Größen X_1 und X_2 , die jedem Element $\omega \in \Omega$ eine reelle Zahl zuordnen.

$$X_1(\omega) = x_1 \text{ und } X_2(\omega) = x_2$$

Dann nennt man (X_1, X_2) einen stochastischen Vektor (sV) mit gemeinsamen Merkmalraum:

$$M = \{(x_1, x_2) | x_1 = X_1(\omega), x_2 = X_2(\omega), \omega \in \Omega\}$$

Wie im eindimensionalen Fall nennt man Teilmengen B aus M Ereignisse und die Wahrscheinlichkeit $P(X \in B)$ für den Eintritt von B lässt sich durch die Verteilungsfunktion charakterisieren.

Die gemeinsame Verteilungsfunktion ist definiert durch

$$P(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$$

(Wsl. das sG kleiner gleich eines gewissen Wertes) Die Wsl. von Ereignissen der Form $(a_1, b_1] \times (a_2, b_2]$ lässt sich wie folgt bestimmen:

$$P(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$$

6.1.1 Diskrete stochastische Vektoren

Ist der Merkmalraum endlich oder abzählbar \implies diskreter sV.

Gemeinsame W-Funktion:

$$P(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

Eigenschaften:

1. Immer zwischen 0 und 1
2. Summe ist 1

Ist die W-Funktion bekannt, lässt sich die Wsl. für ein beliebiges Ereignis wie folgt bestimmen:

$$P(X_1, X_2) \in B = \sum_{(x_1, x_2) \in B} P(x_1, x_2)$$

Träger: besteht aus allen Punkten (x_1, x_2) mit $p(x_1, x_2) > 0$

Um die WslFunktion der einzelnen sG zu erhalten:

$$X_1 : p(x_1) = \sum_{x_2} p(x_1, x_2) \quad X_2 : p(x_2) = \sum_{x_1} p(x_1, x_2)$$

Man hält also z.B. x_1 fest und summiert $p(x_1, x_2)$ über alle möglichen Werte von x_2 . Die auf diese Weise bestimmten Verteilungen nennt man die **Randverteilungen** von (X_1, X_2)

Aus der Kenntnis der beiden Randverteilungen von X_1 und X_2 allein kann die gemeinsame Verteilung von (X_1, X_2) nicht rekonstruiert werden, weil X_1 und X_2 nicht unabhängig sind.

6.1.2 Stetige stochastische Vektoren

Verteilungsfkt eines stetigen sV:

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(\omega_1, \omega_2) d\omega_1 d\omega_2$$

Den Integranden f nennt man die (gemeinsame) **Dichtefunktion** von (X_1, X_2) mit folgenden Eigenschaften:

1. nicht negativ aber nicht mit 1 beschränkt
2. Volumen unter der Dichte ist 1

Ist die Dichte bekannt, lässt sich die Wahrscheinlichkeit für ein Ereignis $\{(X_1, X_2) \in B\}$ wie folgt bestimmen:

$$P((X_1, X_2) \in B) = \int_B \int f(x_1, x_2) dx_1 dx_2$$

Entspricht dem Volumen unter der Fläche $z = f(x_1, x_2)$ über der Menge B . Der Träger besteht aus allen Punkten (x_1, x_2) mit $f(x_1, x_2) > 0$. Randdichten: Analog zum diskreten Fall bestimmt man die Randdichten von X_1 bzw. X_2 aus der gemeinsamen Dichte wie folgt:

$$X_1 : f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 \quad X_2 : f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$$

Integriere also über das überflüssige

6.2 Erwartungswert

Sei (X_1, X_2) ein stochastischer Vektor und g eine reellwertige Funktion, so ist $Y = g(X_1, X_2)$ eine sG und ihr Erwartungswert gegeben durch (LotUS)

$$\text{diskret} : \mathbb{E}(Y) = \sum_{x_1} \sum_{x_2} g(x_1, x_2) p(x_1, x_2)$$

$$stetig : \mathbb{E}(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f(x_1, x_2) dx_1 dx_2$$

Der Erwartungswert des stochastischen Vektors $X = (X_1, X_2)$ ist gegeben durch

$$\mathbb{E}(X) = [\mathbb{E}(X_1) \mathbb{E}(X_2)]$$

Ist also (X_1, X_2) ein sV kann man zuerst die Randdichte von X_1 bestimmen und dann nach Definition $\mathbb{E}(X_1)$ Die Andere Möglichkeit ist über den Lotus

6.3 Bedingte Verteilung

Häufig kennt man den Wert einer Variablen (dh. von X_1 oder X_2) und es stellt sich die Frage, welche Auswirkungen sich dadurch für die Verteilung der anderen Variablen ergeben.

$$P(X_2 = x_2 | X_1 = x_1) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_1 = x_1)} = \frac{p(x_1, x_2)}{p_1(x_1)}$$

Hat alle Eigenschaften einer W-Funktion. Der durch X_1 bedingte Erwartungswert von X_2 ist gegeben durch:

$$\mathbb{E}(X_2 | x_1) = \sum_{x_2} x_2 p(x_2 | x_1)$$

Ist $u(X_2)$ eine Funktion von X_2 , lässt sich der Lotus anwenden:

$$\mathbb{E}[u(X_2) | x_1] = \sum_{x_2} u(x_2) p(x_2 | x_1)$$

Bedingte Varianz:

$$Var(X_2 | x_1) = \mathbb{E}(X_2^2 | x_1) - [\mathbb{E}(X_2 | x_1)]^2$$

6.3.1 Stetiger Fall

bedingte Dichte:

$$f(x_2 | x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

Alle Eigenschaften einer Dichtefunktion.

Der bedingte Erwartungswert:

$$\mathbb{E}(X_2 | x_1) = \int_{-\infty}^{\infty} x_2 f(x_2 | x_1) dx_2$$

Ist $u(X_2)$ eine Funktion kann man wiederum den Lotus anwenden.

6.4 Korrelation

Die gemeinsame W-Funktion bzw. Dichte des stochastischen Vektors (X, Y) sei $p(x, y)$ bzw. $f(x, y)$

$$\text{Kovarianz} : \sigma_{12} = \text{Cov}(X, Y) = \mathbb{E}[(X - \mu_1)(Y - \mu_2)]$$

oder: (Verschiebungssatz für die Kovarianz)

$$\mathbb{E}(XY) - \mu_1\mu_2$$

Der Korrelationskoeffizient ist gegeben wenn σ_1 und σ_2 positiv ist:

$$p = p_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$$

Daraus ergibt sich mittels Verschiebungssatz der Erwartungswert für das Produkt XY

$$E(XY) = \mu_1\mu_2 + \text{Cov}(X, Y)$$

Eigenschaften des Korrelationskoeffizient:

1. Es gilt $-1 \leq p \leq 1$
2. Im Grenzfall $|p| = 1$ gilt Wsl. von Y und X sind auf einer Geraden.

Der Korrelationskoeffizient p lässt sich als Maß für den linearen Zusammenhang zwischen X und Y interpretieren. Dieser erfasst allerdings nur den linearen Zusammenhang, nicht beispielsweise den quadratischen.

6.5 Unabhängigkeit

Falls wir zwei Dichten f_1 und f_2 haben und falls die Dichte $f_2(x_2) = f(x_2|x_1)$ und die gemeinsame Dichte $f(x_1, x_2) = f_1(x_1)f_2(x_2)$ sind die Größen stochastisch **unabhängig**.

Im stetigen Fall muss dies nicht für alle Punkte gültig sein, sondern nur für eine Menge mit Wsl. 1, während im diskret Fall es für alle Punkte erfüllt sein muss. Die Unabhängigkeit lässt sich auch über die Verteilungsfunktion formulieren, hier muss man nicht zwischen stetig und diskret unterscheiden. unabhängig falls:

$$F(x_1, x_2)(\text{gemeinsame Verteilung}) = F_1(x_1)F_2(x_2)$$

Aus der Unabhängigkeit zweier stochastischen Größen folgt auch die lineare Unkorreliertheit. Die Umkehrung gilt allerdings nicht.

$$X, Y \text{ u.a.} \implies p_{XY} = 0$$

6.6 Mehrdimensionale Erweiterungen

Hier nennt man (X_1, X_2, X_n) einen (n-dimensionalen) stochastischen Vektor. Der Merkmalraum erfasst dann alle Elemente der Merkmalräume aller sGn.

Die gemeinsame Verteilungsfunktion ist definiert durch:

$$F(x) = F(x_1, x_2, x_3, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots)$$

Sind alle sG diskreter Form spricht man von einem diskreten stochastischen Vektor, sind alle stetig spricht man von einem stetigen stochastischen Vektor. Den Integranden nennt man die gemeinsame Dichtefunktion.

Ist $Y = u(X_1, X_2, X_n)$ eine Funktion des Vektors lässt sich der Erwartungswert wie folgt berechnen:

$$\begin{aligned} \text{diskret} : \mathbb{E}(Y) &= \sum_{x_1} \sum_{x_n} u(x_1, x_2, \dots, x_n) p(x_1, x_2, \dots, x_n) \\ \text{stetig} : \mathbb{E}(Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \end{aligned}$$

Das Konzept der Randverteilung lässt sich ebenfalls auf mehrere Dimensionen erweitern. Die bedingte Dichte und der bedingte Erwartungswert existieren ebenfalls. Die sG sind unabhängig wenn das Wissen eines Wertes nichts über die anderen Werte aussagt. Sind alle Größen unabhängig, so sind sie auch paarweise unabhängig. Die Umkehrung gilt allerdings nicht.

6.6.1 Varianz-Kovarianzmatrix

Der Erwartungswert $\mathbb{E}(X)$ ist der Vektor der Erwartungswerte.

Für einen stochastischen Vektor mit Mittelwert $\mu = \mathbb{E}(X)$ ist die Varianz-Kovarianzmatrix definiert durch

$$\text{Cov}(X) = \mathbb{E}(X - \mu)(X - \mu)'$$

Diagonal in der Matrix befinden sich die Varianzen der einzelnen sG und an den anderen Stellen die Kovarianzen der Paaren $\text{Cov}(X_i, X_j)$. Da $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ handelt es sich um eine symmetrische Matrix.

6.6.2 Transformationen

Der Transformationssatz für Dichten lässt sich auf (stetige) stochastische Vektoren verallgemeinern. Um die Verteilung einer Funktion von X zu bestimmen, kann man aber auch die **Methode der Verteilungsfunktion** anwenden. Die Bestimmung der Verteilung der Summe $Y = X_1 + X_2$ nennt man Faltung

6.7 Spezielle multivariate Verteilungen

6.7.1 Multinomialverteilung

Ergibt sich aus der Binomialverteilung. N identische und unabhängige Versuche, wobei jeder Versuch mit den Wahrscheinlichkeiten p_1, p_k auf eine von k Arten ausgehen kann. X = die Anzahl der Versuche, die auf die i -te Art ausgehen.

$M(n, p_1, p_2, p_k)$

Für $k = 2$ ergibt sich die Binomialverteilung $M(n, p_1, p_2) = B(n, p_1)$

Für $k = 3$ ergibt sich die Trinomialverteilung $M(n, p_1, p_2, p_3)$

Die Randverteilungen der Multinomialverteilung sind wieder Multinomialverteilungen.

$$\mathbb{E}(X_i) = np_i, \text{Var}(X_i) = np_i(1 - p_i) \text{Cov}(i, X_j) = -np_i p_j$$

6.7.2 Polyhypergeometrische Verteilung

Unter N Objekten gebe es A_i Objekte der i -ten Art. Es werden zufällig n Objekte ohne Zurücklegen gezogen und ist X_i die Zahl der dabei erhaltenen Objekte der i -ten Art, so hat der Vektor eine Polyhypergeometrische Verteilung. Die Randverteilungen der Polyhypergeometrischen Verteilung sind wieder Polyhypergeometrische Verteilungen.

$$\mathbb{E}(X_i) = n * \frac{A_i}{N}$$

$$\text{Var}(X_i) = n * \left(\frac{A_i}{N}\right) \left(1 - \frac{A_i}{N}\right) \left(\frac{N - n}{N - 1}\right)$$

(Korrekturfaktor) Man beachte dass die Größen X_i nicht unabhängig sind und negativ korrelieren.

6.7.3 Multivariate Normalverteilung

Auch die multivariate Verallgemeinerung ist von zentraler Bedeutung in Wahrscheinlichkeitstheorie und Statistik. Beispielsweise basiert die Regressionsanalyse auf der multivariaten Normalverteilung.

$X \sim N_n(\mu, \Sigma)$ wobei μ ein Vektor und Σ eine symmetrische ($n \times n$) Matrix.

Im Spezialfall $\mu = 0$ und einer n -dim Einheitsmatrix spricht man wiederum von einer Standardnormalverteilung.

Der Erwartungswert entspricht μ und die Cov der matrix. Solche sG sind abgeschlossen gegenüber affinen Transformationen. Randverteilungen einer multivariaten Normalverteilung sind wieder multivariate Normalverteilungen

Für $X \sim N_n(\mu, \Sigma)$ gilt, dass X_1 und X_2 genau dann unabhängig sind, wenn $\Sigma_{12} = 0$ (d.h. Wenn alle paarweisen Kovarianzen gleich Null sind).

Bedingte Verteilungen sind ebenfalls wieder multivariate Normalverteilungen. Die bedingten Erwartungswerte nennt man auch Regressionsgeraden von Y auf X . Die beiden Regressionsgeraden $\mathbb{E}(X|y)$ und $\mathbb{E}(Y|x)$ schneiden sich im Punkt (μ_1, μ_2)

7 Folgen von stochastischen Größen

speziell wichtig: lineare Funktionen:

$$T = \sum_{i=1}^n a_i X_i$$

Untersuchen nicht Eigenschaften für festes n sondern ihre Konvergenzverhalten. Für $n \rightarrow \infty$.

$$\mathbb{E}(T) = \sum_{i=1}^n a_i \mathbb{E}(X_i) \text{ (Erwartungswert der Linearkombinationen)}$$

$$\text{Var}(T) = \text{Cov}(T, T) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i < j} a_i a_j \text{Cov}(X_i, X_j)$$

Zähle also zu allen Varianzen die paarweise Kovarianzen dazu. Sind die sG unabhängig so sind die Kovarianzen 0 und es gilt:

$$\text{Var}(T) = \sum_{i=1}^n a_i^2 \text{Var}(X_i)$$

7.1 Faltung

Möchte man die Verteilung von $X+Y$ aus den Verteilungen von zwei unabhängigen sGn X und Y bestimmen, spricht man von Faltung. (Name bezieht sich auf geometrischen Aspekt)

7.1.1 Diskrete Faltung

$$p_{X+Y}(a) = \sum_{y \in M_y} p_X(a-y) p_Y(y)$$

für alle a aus Merkmalraum $X+Y$ Die obige Summe nennt man auch Faltprodukt.

7.1.2 Stetige Faltung

Zunächst bestimmen wir die Verteilungsfunktion von $X+Y$

$$F_{X+Y}(a) = \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy$$

Dichte:

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(x) f_Y(a-x) dx \text{.. Faltprodukt}$$

7.2 Additionstheoreme

Faltung lässt sich unschwer von zwei auf mehreren Größen erweitern. Hierbei gibt es für alle speziellen Verteilungen Theoreme.

7.3 Ungleichungen

Markow'sche Ungleichung:

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \text{ für } a > 0$$

Tschebyschew'sche Ungleichung:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \text{ für } k > 0$$

7.4 Gesetz der großen Zahlen

Stochastische Konvergenz: Die Folge stochastischer Größen X_n konvergiert gegen die sG X wenn:

$$\lim_{n \rightarrow \infty} P(|X_n - X| \geq c) = 0$$

Angenommen die Folge konvergiert gegen eine Konstante a . Dann gilt für eine an der Stelle a stetige Funktion g :

$$g(X_n) \rightarrow g(a)$$

7.5 Schwaches Gesetz der großen Zahlen

$\{X_n\}$ sei eine iid-Folge (ident und unabhängig) mit Mittelwert μ und Varianz σ^2 dann gilt:

$$\overline{X_n} \rightarrow \mu$$

Gibt auch ein starkes Gesetz der großen Zahlen:

$$P\left(\lim_{n \rightarrow \infty} \overline{X_n} = \mu\right) = 1 \text{ (fast sichere Konvergenz)}$$

7.6 Konvergenz in der Verteilung

$\{X_n\}$ sei eine Folge von stochastischen Größen und X eine andere stochastische Größe. Sind F_{X_n} und F_X die Verteilungsfunktionen von X_n bzw. X und ist $C(F_X)$ die Menge aller Stetigkeitspunkte von F_X , so konvergiert X_n in der Verteilung gegen X , wenn:

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \text{ für alle } x \in C(F_X)$$

7.7 Zentraler Grenzwertungssatz (ZGVS):

$\{X_n\}$ sei eine iid-Folge mit dem Mittelwert μ und der Varianz σ^2 . Dann konvergieren die Größen Y_n in der Verteilung gegen eine standardnormalverteilte stochastische Größe $Z \sim N(0, 1)$

7.8 Normalapproximation

Nach dem ZGVS Die Verteilung der Summe $\sum_{i=1}^n X_i$ von iid-Größen X_i (diskret oder stetig) mit Mittelwert μ und Varianz σ^2 für nicht zu kleines n in guter Näherung wie folgt durch eine Normalverteilung approximieren.

$$\sum_{i=1}^n \sim N(n\mu, n\sigma^2)$$

Stetigkeitskorrektur: Die Approximation einer diskreten Verteilung durch eine stetige Verteilung (insbesondere Normalverteilung) lässt sich häufig durch die Stetigkeitskorrektur verbessern: Am unteren Randpunkt wird $\frac{1}{2}$ abgezogen und am oberen Rand addiert.

$$P(a \leq X \leq b) \sim P\left(a - \frac{1}{2} \leq Y \leq b + \frac{1}{2}\right)$$

8 Schließende Statistik

die Grundaufgabe basierend auf Stichproben Rückschlüsse auf das zu Grunde liegende **statistische Modell** zu ziehen. Statistische Modelle durch Parameter charakterisiert und die Aufgabe besteht darin, diese Parameter zu schätzen. Aussagen über die Genauigkeit der Schätzungen treffen und Hypothesen über die Parameter testen.

Es gibt **parametrische** und **nicht parametrische** statistische Modelle. Die Menge aller möglichen Parameter nennt man den **Parameterraum**. Mann nennt die sG X_1, X_2, X_n eine Stichprobe einer sG X , wenn die Größen X_i unabhängig und so wie X verteilt sind.

Eine Funktion $T = T(X_1, X_2, \dots, X_n)$ einer Stichprobe X_1, X_2, \dots, X_n nennt man allgemein eine **Statistik**. Handelt es sich bei T um eine Abbildung in den Paramterraum Φ , nennt man die Statistik eine **Schätzfunktion** für den Parameter $\theta \in \Phi$.

Die wichtigsten Schätzfunktionen in der Statistik sind der **Stichprobenmittelwert** \bar{X}_n und die **Stichprobenvarianz**. (nicht robust, hängt stark von Werten ab)

Der Stichprobenmittelwert ist ein Schätzer für den Mittelwert und die Stichprobenvarianz ein Schätzer für die Varianz. Ein Schätzer für die Streuung ist die Stichprobenstreuung.

8.1 Schätzer

8.1.1 Empirische Verteilungsfunktion

Verteilung einer sG X ist durch ihre Verteilungsfunktion spezifiziert:

$$F(x) = P(X \leq x)$$

Hat man X mehrfach beobachtet, d.h. hat man eine Stichprobe X_1, X_2, \dots, X_n von X so stellt sich die Frage, wie F geschätzt werden kann. Dazu nehmen wir die empirische Verteilungsfunktion:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, x]}(X_i) \text{ für } x \in \mathbb{R}$$

Eigenschaften der empirischen VF:

1. $\mathbb{E}[F_n(x)] = F(x)$
2. $\text{Var}[F_n(x)] = \frac{F(x)[1 - F(x)]}{n}$
3. $F_n(x) \rightarrow F(x)$ für $n \rightarrow \infty$

Schätzung ist am präzisesten an den Rändern, am unpräzisesten am Median.

8.1.2 Satz von Glivenko-Cantelli: (Fundamentalsatz der Statistik)

$$P \left(\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0 \right) = 1$$

Die Wsl. das der größten Abstand der empirischen und der geometrischen Verteilung null ist, ist 1.

Dh. mit Wahrscheinlichkeit 1 konvergiert die empirische Verteilungsfunktion gegen die zugrunde liegende Verteilungsfunktion $F(x)$.

8.1.3 Dichteschätzer

Histogramm und die Kernschätzung. Die Schätzung der Dichte ist ein statistisch schwierigeres Problem als die Schätzung der Verteilungsfunktion.

8.1.4 Momentenschätzer

Die Idee hinter der Momentenmethode zur Schätzung von Parametern besteht darin, die theoretischen Momente der Verteilung den entsprechenden Stichprobenmomenten gleichzusetzen.

Das k -te Moment einer sG X ist definiert durch $\mathbb{E}(X^k)$ (Speziell ist etwa der Mittelwert $\mathbb{E}(X)$ das erste Moment.

Gibt es im Verteilungsmodell m unbekannte Parameter $\theta_1, \theta_2, \dots, \theta_m$, so lassen sich durch Auflösen des folgenden Gleichungssystems:

$$E(X^k) = \frac{1}{n} \sum_{i=1}^n X_i^k$$

Schätzer für die Parameter gewinnen

8.1.5 Maximum Likelihood

Eine konstruktive Methode zur Gewinnung von Schätzfunktionen. Man bekommt unter bestimmten Bedingungen optimale Schätzer (zumindest für große Stichproben)

diskreter Fall: sG X mit der W-Funktion $p(x; \theta)$, wobei $\theta \in \Phi$ ein einzelner unbekannter Parameter ist und sind x_1, x_2, \dots, x_n Beobachtungen einer Stichprobe, so ist die Likelihood Funktion:

$$L(\theta) = \prod_{i=1}^n p(x_i; \theta)$$

Also das Produkt der Wsfunktionen

Der Maximum-Likelihood-Schätzer (kurz ML-Schätzer) von θ ist nun jener Wert aus Φ , der $L(\theta)$ maximiert.

Im diskreten Fall entspricht $L(\theta)$ der Wahrscheinlichkeit, die Stichprobenwerte x_1, x_2, \dots, x_n zu beobachten. Der ML Schätzer ist jener θ -Wert, der die Beobachtung der (konkreten) Stichproben am wahrscheinlichsten macht.

Likelihood Prinzip: Entscheide dich für das plausibelste Verteilungsmodell, oder entscheide dich für jenes Modell, dass die Daten mit höchster Wsl. erzeugt hat. Es kann einfacher sein mit der Log-Likelihood (-Funktion) zu arbeiten.

stetiger Fall: Likelihoodfunktion ist gleich wie im diskreten Fall nur mit Dichten:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

der ML-Schätzer von θ ist jener Wert aus Φ , der $L(\theta)$ maximiert.

Invarianz der ML-Schätzung: Der ML-Schätzer macht die Transformation einer Funktion mit.

8.2 Gütekriterien für Schätzer

8.2.1 Erwartungstreue

Ein Schätzer für einen Parameter $\theta \in \Phi$ heißt erwartungstreu (oder unverzerrt), wenn:

$$\mathbb{E}_\theta(\theta_n) = \theta$$

Gilt dies für $n \rightarrow \infty$ nennt man den Schätzer **asymptotisch erwartungstreu**.

Anschaulich bedeutet die obige Definition, dass man bei Verwendung erwartungstreuer Schätzer im Mittel an der gewünschten Stelle ist.

8.2.2 Effizienz

Betrifft die Varianz. Man sagt, dass ein erwartungstreuer Schätzer effizienter als ein anderer ist, wenn:

$$\text{Var}(\theta_1) < \text{Var}(\theta_2)$$

Ist ein erwartungstreuer Schätzer des Parameters θ effizienter als jeder anderer erwartungstreue Schätzer desselben Parameters, nennt man ihn **effizient**.

8.2.3 Konsistenz

Ein Schätzer basierend auf einer Stichprobe der Größe n , heißt konsistent für θ wenn

$$\theta_n \rightarrow \theta \text{ f\"ur } n \rightarrow \infty$$

Anschaulich bedeutet Konsistenz, dass sich ein Schätzer mit dieser Eigenschaft für wachsendes n mit hoher Wahrscheinlichkeit in der Nähe des zu schätzenden Parameters aufhält. Eine sehr wünschenswerte Eigenschaft von "guten" Schätzern.

8.2.4 Asymptotische Normalverteilung

Ein Schätzer ist asymptotisch normalverteilt, wenn er in Verteilung gegen eine normalverteilte sG konvergiert.

8.3 Konfidenzintervalle

Ein wesentlicher Teil jeder Schätzprozedur sind Aussagen betreffend die Genauigkeit der Schätzer. Ist X eine Stichprobe der sG S , deren Verteilung von einem unbekannten Parameter abhängt und sind $T_1(X) < T_2(X)$ zwei Funktionen der Stichprobe, so nennt man das Zufallsintervall $(T_1(X), T_2(X))$ ein **Konfidenzintervall**. Es gibt mehrere Möglichkeiten zur Konstruktion von Konfidenzintervallen.

8.3.1 Pivotmethode

Unter einer Pivotgröße (Größe um die sich etwas dreht) versteht man eine sG T , die eine Funktion der Stichprobe X und des Parameters θ ist, deren Verteilung aber bekannt ist und nicht von θ abhängt.

Um nun ein KI mit Konfidenzkoeffizient $1 - \alpha$ zu konstruieren, nehmen wir das $(\alpha/2)$ und das $(1 - \alpha/2)$ Quantil der **Pivotverteilung**

$$P_\mu(z_{\alpha/2} < T < z_{1-\alpha/2}) = 1 - \alpha$$

Der Ausdruck in Klammern lässt sich äquivalent wie folgt schreiben

$$\overline{X}_n - z_{1-\alpha/2} \frac{1}{\sqrt{n}} < \mu < \overline{X}_n + z_{1-\alpha/2} \frac{1}{\sqrt{n}}$$

Wie lässt sich diese Intervall interpretieren? Zieht man wiederholt Stichproben der Größe n aus X und bestimmt jeweils das obige KI, so werden etwa $100(1 - \alpha)$ dieser Intervalle das wahre μ überdecken.

1. Formuliere ein statistische Modell für die Stichprobe X
2. Wähle eine geeignete Pivotgröße
3. Bestimme Verteilung des Privots
4. Bestimme zwei Quantile q_1 und q_2 .
5. Bringe das Ereignis in die Form $\{T_1(X) < \theta < T_2(X)\}$
6. $(T_1(X), T_2(X))$ ist ein $100(1 - \alpha)$ -Konfidenzintervall für θ .

8.3.2 Normalverteilung

Auf Basis einer Stichprobe X_1, X_2, \dots, X_n von $X \sim N(\mu, \sigma^2)$ können exakte Konfidenzintervalle für μ und σ^2 konstruiert werden, da $\overline{X_n}$ und S_n^2 unabhängig sind.

8.3.3 Normalverteilung (zwei ua. Stichproben)

Hat man zwei Stichproben die einer Normalverteilung folgen, lassen sich Konfidenzintervalle für die Differenz der Mittelwerte bzw für den Quotienten der Varianzen konstruieren.

8.3.4 Normalverteilung (verbundene Stichproben)

Entwicklung von Konfidenzintervallen für die Differenz der Mittelwerte, wenn die Stichproben abhängig sind. Betrifft in erster Linie Vorher/Nachher Situationen an denselben Untersuchungseinheiten. Die korrekte Vorgangsweise in diesem Fall ist die Bildung der Differenzen $D_i = X_i - Y_i$ der Beobachtungen.

8.4 Statistische Tests

Betrachten das Testen von statistischen Hypothesen. Unter einer **Parameterhypothese** versteht man eine Behauptung über (oder die) Parameter von einer (oder mehreren) Verteilungen) Bei dieser Art von Hypothesen wird angenommen, dass der Verteilungstyp bekannt ist. Ist der Verteilungstyp nicht bekannt und möchte man testen, ob eine bestimmte Verteilung ein zufriedenstellendes Modell für die vorliegenden Beobachtungen darstellt, spricht man von einer **Verteilungshypothese**.

8.4.1 Parametertests

Auf Grund einer Theorie gelte $\theta \in \Sigma_0$ oder $\theta \in \Sigma_1$, wobei Σ_0 und Σ_1 disjunkt sind. Die erste Behauptung nennt man **Nullhypothese**, die zweite **Alternativ- oder Gegenhypothese**.

Als Nullhypothese wählt man in der Regel diejenige Behauptung, die die bisherige Situation oder den "Normalfall" repräsentiert. Die Alternativhypothese ist häufig einfach das Komplement.

Testentscheidung: Eine auf einer Stichprobe X basierende Entscheidungsregel über Hypothesen nennt man einen Test. Ein Test wird durch seinen kritischen Bereich C charakterisiert. Dabei handelt es sich um eine Teilmenge des Stichprobenraumes M_X^n mit

Verwerfe H_0 (Akzeptiere H_1) falls X in C

Akzeptiere H_0 (Verwerfe H_1) falls X in C^c

Allgemein unterscheidet man **Typ I** und **Typ II-Fehler**. Der erste tritt auf, wenn H_0 verworfen wird, obwohl sie richtig ist, der zweite tritt auf, wenn H_0 nicht verworfen wird, obwohl sie falsch ist.

Die Wahrscheinlichkeit eines Typ I-Fehler bezeichnet (auch Signifikanzniveau genannt) man mit α :

$$\alpha = P(\text{Typ I} - \text{Fehler}) = P(X \in C)$$

Die Wahrscheinlichkeit eines Typ II Fehlers bezeichnet man mit β :

$$\beta = P(\text{Typ II} - \text{Fehler}) = P(X \in C^c)$$

In der Praxis verwendet man Tests, die eine vorgegebene Wahrscheinlichkeit für einen Typ-I-Fehler nicht überschreiten. Wird H_0 verworfen, spricht man daher von einer **starken** Schlussfolgerung. Wird H_0 nicht verworfen, hat man möglicherweise einen Typ II-Fehler begangen, und über die Größe weiß man meist nur wenig, da β vom wahren unbekannten Parameterwert abhängt. Nennt man dann **schwache** Schlussfolgerung.

Die **Schärfe** eines Tests ist die Wahrscheinlichkeit der Verwerfung der Nullhypothese H_0 , wenn die Alternativhypothese zutrifft. (d.h. die richtige Entscheidung zu treffen, wenn H_0 falsch ist).

8.5 p-Wert

Die meisten Statistikpakete verfolgen beim Testen von Hypothesen nicht die "klassische" Vorgehensweise, sondern berechnen statt dessen einen Wahrscheinlichkeitswert. Der p-Wert der H_0 entspricht der Wahrscheinlichkeit bei Zutreffen von H_0 - den beobachteten Wert der Teststatistik oder einen extremeren zu bekommen. Was unter extremer zu verstehen ist, hängt von der Gegenhypothese ab.

Bezug zum klassischen Testen: Der p-Wert der H_0 ist der größte Wert von α , für den die H_0 nicht verworfen wird.

Bei der Interpretation des p-Wertes hält man sich meist an das folgende Schema:

$< 0.01 \rightarrow$ sehr starke Einwände gegen H_0

$0.01 - 0.05$ starke Einwände gegen H_0

$0.05 - 0.10$ schwache Einwände gegen H_0

> 0.10 keine Einwände gegen H_0

Ein großer p-Wert bedeutet nicht automatisch eine Unterstützung für H_0 . Ein möglicher dafür könnte auch sein, dass die H_0 falsch ist, aber der Test eine zu geringe Schärfe hat um das zu erkennen.

Es gibt eine Beziehung zwischen Parametertests und Konfidenzintervallen, man kann die Konfidenzintervalle als Hypothesen verwenden.

8.6 Chiquadrat-Anpassungstests

Gehören zu den ältesten Methoden der schließenden Statistik:

$$\left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2(1)$$

Hat man n unabhängige sGn X_i , so gilt nach dem Additionstheorem für Chiquadrat-Verteilungen:

$$\sum_{i=1}^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n)$$

Einfacher Chiquadrat-Anpassungstest: Der Merkmalraum M zerfalle in k paarweise disjunkte Teilmengen A_1, \dots, A_i . Das Experiment werde n Mal wiederholt. Dann ist die Teststatistik gegeben durch:

$$Q_{k-1} = \sum_{i=1}^k \frac{(X_i - np_{i0})^2}{np_{i0}}$$

wobei:

$$H_0 : p_i = p_{i0} \text{ gegen } H_1 : \exists i \text{ mit } p_i \neq p_{i0}$$

H_0 wird verworfen, falls:

$$Q_{k-1} > \chi_{k-1; 1-\alpha}^2$$

Der Test hat nur approximativ das Niveau α . Verschiedene Regeln haben sich etabliert, um die Zulässigkeit der χ^2 -Approximation zu gewährleisten. Die übliche Regel besagt: nur wenn $np_{i0} \geq 5$.

Sind die Wsl. p_i nicht vollständig spezifiziert, ist der Anpassungstest etwas zu modifizieren. Man verwendet den sogenannten Minimum-Chiquadrat-Schätzer.