Deskriptive Statistik

Statistik umfasst die Entwicklung und Anwendung von Methoden zur Erhebung, Aufbereitung, Analyse und Interpretation von Daten.

Deskriptive Statistik: Beobachtete Merkmalsträger und ihre Verdichtung:

0. Vorbereiten
1. Relevante Daten zu Untersuchungsobjekten vollständig erheben
2. Daten aufbereiten (Tabelle, Grafik)
3. Daten analysieren + interpretieren → Kennzahlen (Mittelwert/Streuungsmass), Gesetzmässigkeiten, Berechnung Abhängigkeitsmass

Skalierungen:

Informationsniveau nimmt gegen unten zu!		
↓	Nominal	Skalenwerte = Namen; Gleichberechtigt, keine Reihenfolge Merkmale: qualitativ und/oder häufbar (ledig, verheiratet, geschieden, verwitwet)
	Ordinal	Skalenwerte = Klassenbezeichnungen; auf- oder absteigende Folge Merkmale: Vergleiche im Rang, intensitätsmässig abgestuft (++,+,0,-,--)
	Intervall (Metrisch)	Skalenwerte = Reelle Zahlen; besser/schlechter, früher/später, grösser/kleiner Merkmale: quantitativ → kein 0-Referenz → Kein Verhältnis zwischen Zahlen(°C)
	Verhältnis (Metrisch)	Skalenwerte = Reelle Zahlen; besser/schlechter, früher/später, grösser/kleiner Merkmale: quantitativ → 0-Referenz → Verhältnis zwischen Zahlen (Einkommen)

Eindimensionale Häufigkeitsverteilung → Ein Merkmal: Wie häufig ist Merkmalswert x aufgetreten
Mehrdimensionale Häufigkeitsverteilung → Mehrere Merkmale, Übersichtlich max. 2 Merkmale
Klassifizierte Häufigkeitsverteilungen → Viele Merkmalswerte: Klassenbildung (Merkmale von…bis unter)

Grundbegriffe:

Grundgesamtheit → Abgrenzung je nach Ziel: inhaltlich, sachlich, räumlich, Zeitpunkt/-raum (alle Bürger, alle PC)
↘Stichprobe = repräsentativer Teil der Grundgesamtheit.

Merkmale, welche Grundgesamtheit zeigt = interessante Eigenschaften bei Untersuchung (Kanton, Mängel PC)
↘Untergliederung: qualitativ, nicht messbar (FDP,SP,SVP,CVP,…) / quantitativ, messbar (++,+,0,-,--)
diskret, definierte Aufteilung (Zimmer:1 ½, 2) / stetig, ∞viele Werte→ messen (Körpergrösse)
häufbar, Merkmalsträger hat mehr als 1 Merkmalswert / nicht häufbar, nur 1 Merkmalswert

Bezeichnungen:

n	Gesamtzahl der Merkmalsträger	
v	Anzahl verschiedener Merkmalswerte	
x	Wert	
h_i	Absolute einfache Häufigkeit (oder Klassenhäufigkeit), Anzahl merkmalsträger mit Merkmal x_i (oder x_i, der in die j-te Klasse fällt) Zahl für entsprechende Klasse: in 0-100 sind es 30, in 100-150 sind es 34	$\sum_{i=1}^{v} h_i = n$
H_i	Absolute kumulierte Häufigkeit (oder Klassenhäufigkeit), Anzahl der Merkmalsträger mit Merkmalswert <= x_i (oder x_i <= die Obergrenze der j-ten Klasse); Σ absolute einfache Häufigkeit	$H_i = \sum_{k=1}^{i} h_k$ (h_1+h_2+h_3...)
f_i	Relative einfache Häufigkeit, Anteil Merkmalsträger mit Merkmal x_i Σ absolute einfache Häufigkeit = 100%; f_i = ?%	$f_i = \frac{h_i}{n}$
F_i	Relative kumulierte Häufigkeit, Anteil der Merkmalsträger mit Merkmalswert <= xi; Σ rel. Häufigkeit	$F_i = \sum_{k=1}^{i} f_k = \frac{H_i}{n}$ (f_1+f_2+f_3...)
HR_i	Resthäufigkeit → Wie viele Merkmalsträger mit Merkmalswert > x_i	$HR_i = n - H_i$
FR_i		$FR_i = 1 - F_i$
d_j	Häufigkeitsdichte: h_j / grösse von Klasse = 30 / (100-0)	$d_j = \frac{h_j}{x_j^o - x_j^u}$
x^l_j	Klassenmitte: von 0-100: (0+100)/2=50, von 100-150 (100+150)/2=125	$x_j^l = \frac{x_j^o + x_j^u}{2}$
j	Laufindex für die Klassen (Klassenindex) j=1,…,v	
x^u_j / x^o_j	Untergrenze der Klasse j / Obergrenze der Klasse j	

Tabellenbeispiel:

j	von….bis unter		h_i	H_i	f_i	F_i	d_j	x^l_i	$x^l_i * h_i$
	x^u_i	x^o_i							
1	0	20	10	10	8%	8%	0.5	10	100
2	20	40	20	30	16%	24%	1	30	600
3	40	60	60	90	48%	72%	3	50	3’000
4	60	100	35	125	28%	100%	.875	80	2’800
	Total (n) = 125			Total =100%			Total = 6’500		

Jahr	Einkom-men	Wachstums-	
		-faktor	-rate
98	56’000		
99	67’000	1.1964	+19.64%
00	74’600	1.1119	+11.19%
01	69’000	0.9261	-7.38%
02	78’500	1.1376	+13.76%

Mehrdimensionale Häufigkeitsverteilungen:

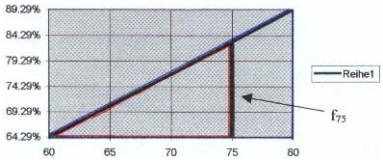
y_k / x_i	0		1		2		$\sum_{k=1}^{3} h_{ik}$
	Einf	Kum	Einf	Kum	Einf	Kum	
I	1 =	1 +	2 =	3 +	1 =	4	Σ Einfach
II	3 =	4 +	2 =	8 +	2 =	11	Σ Einfach
III	1 =	5 +	2 =	11 +	0 =	14	Σ Einfach
$\sum_{k=1}^{3} h_{ik}$	Σ Einfach		Σ Einfach		Σ Einfach		

Berechnung:
H_{33} = 1 + 2 + 1
+3 + 2 + 2
+1 + 2 + 0 = 14

oder $H_{ik} = \sum_{a=1}^{i}\sum_{b=1}^{k} h_{ab}$

$$H_{33} = \sum_{a=1}^{3}\sum_{b=1}^{3} h_{ab} = \sum_{b=1}^{3} h_{1b} + \sum_{b=1}^{3} h_{2b} + \sum_{b=1}^{3} h_{3b} = 1+2+1+3+2+2+1+2+0 = 14$$

Interpolation → Wert der sich mitten in einer Klasse befindet berechnen
Klassen: …, 40-60, 60-80, 80-100, … → Wo ist Wert von x=75? Kann man nicht herauslesen!
Lösung → Annahme linearer Verlauf, ähnliche Dreiecke:

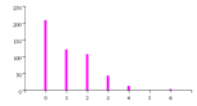
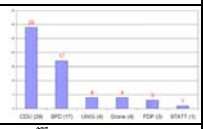
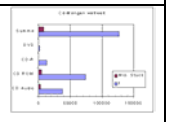
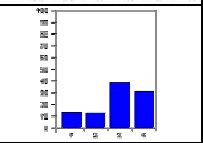
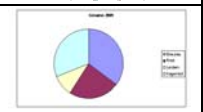
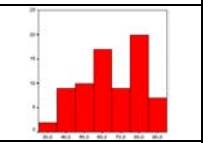
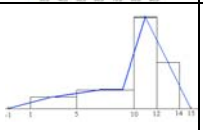
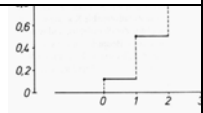


Herleitung: Seitenverhältnisse der Dreiecke sind gleich (gleiche Winkel)

$$\frac{f_{75}}{75-60} = \frac{89.29\% - 64.29\%}{80-60} \rightarrow f_{75} = \frac{89.29\% - 64.29\%}{80-60} * (75-60) \rightarrow F_{75} = F_{74} + f_{75}$$

also gilt $F_{(x)} = F_{j-1} + \left(\frac{x - x_j^u}{x_j^o - x_j^u}\right) * \left(F_j - F_{j-1}\right)$

Diagrammtypen:

Einfache Häufigkeitsverteilungen	**Stabdiagr.** ↓ Stäbe verbreitern ↓		Höhenproportional x=Tarife (x_i), y=Beschäftigte (h_i)
	Säulendiagr. ↓ ähnlich wie ↓		Höhenproportional x=Tarife (x_i), y=Beschäftigte (h_i) **Balkendiagramm** → gedrehtes Säulendiagramm 
	Rechteck-diagr.		Flächenproportional → Wenn Breite überall gleich, auch Höhenproportional
	Kreisdiagr.		Sektoren entsprechen flächenmässig den Häufigkeiten; U=2*r*π A=r^2* π U=360°
	Histogramm		Geeignet zur Darstellung klassifizierter Häufigkeitsverteilungen Flächenproportional mit Klassenbreite x=Klassenbreite, y=Häufigkeitsdichte (d_j) - Fläche über mehrere Klassen mit Mittelwert berechnen!
	Polygozug		Geeignet bei klassifizierten Häufigkeitsverteilungen → Vergleiche! x=Klassenbreite, y=Häufigkeitsdichte (d_j) 1. In Mitte d. Klassen Häufigkeitsdichten eintragen + linear verbinden 2. Fläche Polygonzug = Fläche Histogramm 3. Rand links → ½ ersten; Rand rechts → ½ letzten Häufigkeitsdichte 4. Anfang + Schluss Verlängerung, damit Fläche gleich Bleibt 5. Ungleichbreite Klassen:Flächen müsse gleich sein (abgeschnittene, wie dazugewonnene) → deshalb Korrektur (Mitte von Kastenhöhe)
Kumulierte Häufigkeitsverteilungen	**Treppen-funktion**		Geeignet bei ordinalskal. und diskreten nicht klassifiziert. Merkmalen x=Tarife (x_i), y=Anz Beschäftigte (H_i)
	Summen-polygon		Geeignet zur Darstellung klassifizierter Häufigkeitsverteilungen **Kumulierte Häufigkeiten:** x=Kumulierte Häufigkeiten (H_i), y=Klassenobergrenzen - Erster Punkt der Untergrenze der 1. Klasse mit 0 verbinden **Kumulierte relative Häufigkeiten (Ogive):** x=rel. Kumulierte Häufigkeiten (F_i), y=Klassenobergrenzen - Erster Punkt der Untergrenze der 1. Klasse mit 0 verbinden → auch empirische Verteilfunktion genannt!

Lageparameter → Wo liegt Verteilung, Mittelwert?

Mo=Modus: Derjenige Merkmalswert, der am häufigsten beobachtet wurde. *(Bei jeder Verteilung bestimmbar)*
Eignung: Für nominalsklaierte Grösse einziger Mittelwert; Braucht deutlichen Gipfel, dass er etwas bringt + nähere Umgebung höhere Konzentration, als Rest.
Nicht klassifiziert: Der Modus befindet sich in der Klasse mit der grössten Klassenhäufigkeit h_j.
Klassifiziert: Modus befindet sich in der Klasse mit der grössten Klassenhäufigkeit bei konstanter Klassenbreite, oder d_j bei unterschiedlicher Klassenbreite. Genaue Berechnung:
Konst. Kl.breite: $Mo = x_m^u + \left(x_m^o - x_m^u\right) * \frac{h_m - h_{m-1}}{(h_m - h_{m-1}) + (h_m - h_{m+1})}$ Untersch. Kl.breite: $Mo = x_m^u + \left(x_m^o - x_m^u\right) * \frac{d_m - d_{m-1}}{(d_m - d_{m-1}) + (d_m - d_{m+1})}$
Mo=Klassenuntergrenze + Klassenbreite * Prozentsatz (Näherung)
Falls Mo in unterster oder oberster Klasse → Für unbekannte Werte 0 einsetzen

Me=Median: Derjenige Merkmalswert, dessen Merkmalsträger in der Rangordnung aller Merkmalsträger genau die mittlere Position einnimmt. *(Merkmal mindestens ordinal skaliert → bei n ungerade geht's nicht mit Ordinals.)*
Eignung: Bei schiefen Verteilung, keine Gefahr der Verzerrung, da von Ausreissern unbeeinflusst.
Nicht klassifiziert: n ungerade: $Me = x_{\left[\frac{n+1}{2}\right]}$ n gerade: (n+1)/2 → Ø beider Merkmalsträger $Me = 0.5 * \left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right)$

$Me = x_{\left[\frac{22}{2}\right]} = 11$
Annahme n=21
11 haben 7Fehlt.

(20+1)/2=10.5
Ø 6 und 7 Fehltage
$Me = 0.5 * (6+7) = 6.5$
6.5 theoretische Mitte

Fehltage	0	2	5	6	7	11	12	14
hi	4	2	2	2	4	3	2	1
Hi	4	6	8	10	14	17	19	20

Klassifiziert:
1. Medianklasse bestimmen n/2 (Medianklasse = Klasse, in welcher dieser Wert liegt) → Ergebnis wo in H_j?
2. $Me = x_m^u + \frac{\frac{n}{2} - H_{m-1}}{h_m \text{ oder } H_m - H_{m-1}} * (x_m^o - x_m^u)$ 50% d.Merkmalsträger: Forderung>als Median, 50%: Forderung<Median

D=Quantile, Q=Quartile: Nicht klassifiziert: (1. Quartil) $Q1 = x_{\left[\frac{n+1}{4}\right]}$ **(3. Quartil)** $Q3 = x_{\left[\frac{3*(n+1)}{4}\right]}$

Klassifiziert:
1. Medianklasse bestimmen (dort, wo der Wert drin liegt), Bsp: 90% von n(=245) = 220.5 in welcher Klasse?
2. Formel anwenden:

Quantil: (Dezentil = Zehntel, Perzentil = Hundertstel) $D_9 = x_m^u + \frac{\frac{9}{10} * n - H_{m-1}}{H_m - H_{m-1}} * (x_m^o - x_m^u)$ 90% Forderung<Wert

Quartil: **(1. Quartil)** $Q_1 = x_m^u + \frac{\frac{1}{4} * n - H_{m-1}}{H_m - H_{m-1}} * (x_m^o - x_m^u)$ **(3.Quartil)** $Q_3 = x_m^u + \frac{\frac{3}{4} * n - H_{m-1}}{H_m - H_{m-1}} * (x_m^o - x_m^u)$

$\bar{x}$ **=Artithmetisches Mittel:** Ergibt sich bei gleichmässiger Verteilung der Σ aller beobachteten Merkmalswerte auf alle merkmalsträger. *(Abstände zwischen Merkmal messbar →Merkmal mind. Intervallskaliert)*
Eignung: Nur bei (nahezu) symmetrischen Häufigkeitsverteilungen, oder bei Verteilungen ohne Konzentrationen, nicht bei schiefen Verteilungen oder Ausreissern
Nicht klassifiziert: $\bar{x} = \frac{\sum_{i=1}^{v} x_i * h_i}{n} = \sum_{i=1}^{v} x_i * f_i$ bzw. $\frac{(h_1 * Wert1) + (h_2 * Wert2) + \ldots}{n}$
Klassifiziert: (Näherung) $\bar{x} = \frac{\sum_{j=1}^{v} x_j^l * h_j}{n} = \sum_{j=1}^{v} x_j^l * f_j$ bzw. $\frac{(Klassenmitte_1 * h_1) + (Klassenmitte_2 * h_2) + \ldots}{n}$
z.B. Für Ø Versicherungssumme bei verschiedenen Summen

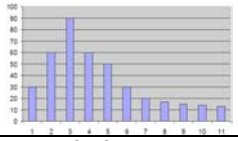
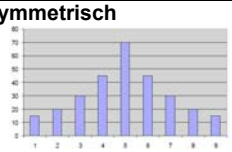
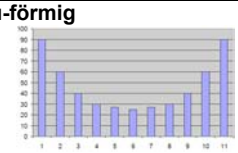
F_{GM}=Geometrisches Mittel: Basis = Merkmalswert, der zu verschiedenen Zeitpunkten erhoben wird. Werte Miteinander vergleichen → jährliche Veränderung = Faktor → Faktor – 1 = jährliche prozentuale Veränderung *(Grössen müssen verhältnismässig skaliert sein, nicht bei klassifizierten Häufigkeiten anwendbar!!!)*
Eignung: Ø prozentuale Entwicklung einer Grösse im Zeitablauf kann nur mit geom. Mittel ermittelt werden!
Wachstumsfaktor: Wert 2 / Wert 1 Wachstumsrate: Wert1 + Wachstumsrate [%] = Wert 2
Geom. Mittel (Variante 1): $F_{GM} = \sqrt[n]{\frac{Endwert}{Anfangswert}}$ → Wachstumsfaktor bei n Wachstumsfaktoren
Geom. Mittel (Variante 2): $MG = \sqrt[n]{F1 * F2 * \ldots * Fn} = \sqrt[n]{\prod_{i=1}^{n} F_i}$ → Geometrisches Mittel bei n Wachstumsfaktoren

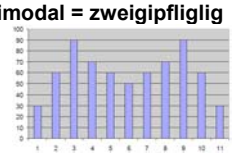
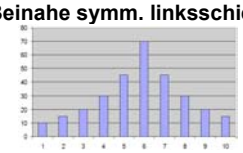
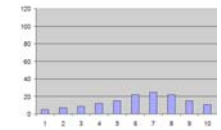
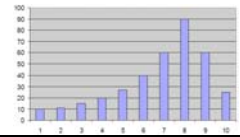
Auswertung Lageparameter:

- Modus uns Median sind unabhängig von Abweichungen (hohen Werten) links oder rechts (z.B. durch hohe Werte in grössere Klassen)
  - Modus: Häufigster Wert ist massgebend
  - Median: Gleich viele Werte links und rechts
- Arithm. Mittel wird von hohen Werten links oder rechts beeinflusst (z.B. durch hohe Werte in grössere Klassen)
- Geom. Mittel macht keinen Sinn, wenn Zahlen nicht miteinander in Abhängigkeit

→ bei ziemlich symmetrischer Verteilung liegen alle Parameter ziemlich nah beieinander. Ist die Verteilung unsymmetrisch, kann das arithm. Mittel abweichen.

Typen von Verteilungen → Diagrammarten und ihre Erscheinungsformen

Rechtsschief = linkssteil 	**Linksschief = rechtssteil** 
Symmetrisch 	**u-förmig** 

Bimodal = zweigipfligl 	**Beinahe symm. linksschief** 
Flach 	**Steil** 

Streuungsparameter → Treten Merkmalswerte in breitem, oder nur engem Bereich auf?

R=Spannweite (Variabilität): Differenz zwischen dem grössten und kleinsten beobachteten Wert *(Merkmal mind. Intervallskaliert, Praxis oft Ordinalskalierung genügend)* z.B. Börsenkurs mit Höchst- und Tiefstwerten
Eignung: Reagiert empfindlich auf Ausreisser. Ist aber anschauliches und einsichtiges Streuungsmass
Nicht klassifiziert: $R = x_n - x_1$ (Oberster Merkmalswert – Unterster Merkmalswert)
Beispiel: Überstunden von 0 – 12 Stunden → R = 12 – 0 = 12 →Überstunden streuen im Intervall von 12 Std.
Klassifiziert: $R = x^o_v - x^u_1$ (Grösster Wert: Obergrenze letzte Klasse - Kleinster Wert: Untergrenze 1. Klasse)

ZQA=Zentraler Quartilsabstand/Interquartilsabstand: Ist Entfernung zischen den beiden Merkmalswerten, welche in der Rangordnung zentral gelegen 50% der Merkmalsträger eingrenzen. Es wird also oben und unten je 25% ab. (**Zentraler 90%-Perzentilabstand** schneidet oben und unten je 5% ab. Berechnung analog) *(Merkmal mind. intervallskaliert; wenn nur Angabe der Quartilswerte (ohne Abstand), genügt auch Ord.skal.)*
Eignung: Anschaulich und leicht verständlich. Problem von Ausreissern ist hier nicht relevant. Gutes Mass, wenn z.B. nur der mittlere Teil der Werte von Interesse ist.
Nicht klassifiziert, Klassifiziert: ZQA = 3. Quartil – 1. Quartil (Berechnung Quartil → Siehe Quartil)

δ=Mittlere absolute Abweichung: Ø Entfernung aller beobachteten Merkmalswerte vom arithm. Mittel od. Median *(Merkmale mindestens intervallskaliert)*
Eignung: Problem von Ausreissern. DAS geeignete Mass in beschreib. Statistik, wird aber wegen besonderer Bedeutung von Varianz in der schlies. Statistik zunehmend durch Standartabweichung und Varianz verdrängt.
Nicht klassifiziert: $\delta = \frac{1}{n} * \sum_{i=1}^{v} |x_i - \bar{x}| * h_i$ **Klassifiziert:** $\delta = \frac{1}{n} * \sum_{j=1}^{v} |x_j^l - \bar{x}| * h_j$ (Unterstellung von Gleichverteilung in Klasse)

σ^2=Varianz, σ=Standartabweichung: Summe der quadrierten Entfernungen der Merkmalswerte vom arithmetischen Mittel dividiert durch die Anzahl Merkmalsträger *(Merkmale mind. Intervallskaliert)*
Eignung: Wenig Anschaulich, Merkmalswerte mit zunehmendem Abstand vom Mittelwert haben überproport. Einfl.
Nicht klassifiziert: Varianz

$\sigma^2 = \frac{1}{n} * \sum_{i=1}^{v} (x_i - \bar{x})^2 * h_i$ oder $\sigma^2 = \sum_{i=1}^{v} (x_i - \bar{x})^2 * f_i$ vereinfacht $\sigma^2 = \frac{1}{n} * \sum_{i=1}^{v} x_i^2 * h_i - \bar{x}^2$ oder $\sigma^2 = \sum_{i=1}^{v} x_i^2 * f_i - \bar{x}^2$

Standartabweichung:

$\sigma = \sqrt{\frac{1}{n} * \sum_{i=1}^{v} (x_i - \bar{x})^2 * h_i}$ oder $\sigma = \sqrt{\sum_{i=1}^{v} (x_i - \bar{x})^2 * f_i}$ vereinfacht $\sigma = \sqrt{\frac{1}{n} * \sum_{i=1}^{v} x_i^2 * h_i - \bar{x}^2}$ oder $\sigma = \sqrt{\sum_{i=1}^{v} x_i^2 * f_i - \bar{x}^2}$

Klassifiziert: Varianz

$\sigma^2 = \frac{1}{n} * \sum_{j=1}^{v} (x_j^l - \bar{x})^2 * h_j$ oder $\sigma^2 = \sum_{j=1}^{v} (x_j^l - \bar{x})^2 * f_j$ vereinfacht $\sigma^2 = \frac{1}{n} * \sum_{j=1}^{v} (x_i^l)^2 * h_j - \bar{x}^2$ oder $\sigma^2 = \sum_{j=1}^{v} (x_j^l)^2 * f_j - \bar{x}^2$

Standartabweichung:

$\sigma = \sqrt{\frac{1}{n} * \sum_{j=1}^{v} (x_j^l - \bar{x})^2 * h_j}$ oder $\sigma = \sqrt{\sum_{j=1}^{v} (x_j^l - \bar{x})^2 * f_j}$ vereinfacht $\sigma = \sqrt{\frac{1}{n} * \sum_{j=1}^{v} (x_j^l)^2 * h_j - \bar{x}^2}$ oder $\sigma = \sqrt{\sum_{j=1}^{v} (x_j^l)^2 * f_j - \bar{x}^2}$

VK=Variationskoeffizient: Misst nicht absolute Streuung, sondern setzt diese in Beziehung zur Lage der Häufigk.
(Standartabweichung wird als Prozentsatz des arithm. Mittels ausgedrückt, also Merkmal verhältnisskaliert)
Eignung: Ist geeignet beim Vergleich von Streuungen und Häufigkeitsverteilungen mit unterschiedlichen Lagen.
VK = $\sigma / \bar{x}$ bzw. falls $\bar{x}$ <0: VK = $\sigma / |\bar{x}|$

Tabellenbeispiel

von..	bis unter..	h_j	x_j^l	$x_j^l * h_j$	$\lvert x_j^l - \bar{x} \rvert$	$\lvert x_j^l - \bar{x} \rvert * h_j$	$(x_j^l - \bar{x})^2$	$(x_j^l - \bar{x})^2 * h_j$
50	100	15	75	1'125	245.92	3'688.78	60'476.65	907'149.70
100	200	50	150	7'500	170.92	8'545.92	29'231.65	1'460'682.32
200	300	80	250	20'000	70.92	5'673.47	5'029.65	402'371.71
300	400	40	350	14'000	29.08	1'163.27	845.65	33'825.86
400	600	40	500	20'000	179.08	7'163.27	32'069.65	1'282'785.86
600	1'000	20	800	16'000	479.08	9'581.63	229'517.65	4'590'352.93
	Total (n) = 245		Total = 78'625		Total = 35'816.33		Total = 8'677'168.37	

Um 320 herum streut es mit 188 bzw mit einer Varianz von 35517.

Variationskoeffizient:
VK = $\sigma / \bar{x}$
= 188.19 / 320.92
= 0.5864 = 85.64%

Arithm. Mittel: $\bar{x} = \frac{\sum_{j=1}^{v} x_j^l * h_j}{n} = \frac{78625}{245} = \underline{320.92}$

Mittlere abs. Abw: $\delta = \frac{1}{n} * \sum_{j=1}^{v} |x_j^l - \bar{x}| * h_j = \frac{1}{245} * 35816.33 = \underline{146.19}$

Varianz: $\sigma^2 = \frac{1}{n} * \sum_{j=1}^{v} (x_j^l - \bar{x})^2 * h_j = \frac{1}{245} * 8677168.37 = \underline{35417.01}$

Standartabweichung: $\sigma = \sqrt{\frac{1}{n} * \sum_{j=1}^{v} (x_j^l - \bar{x})^2 * h_j} = \sqrt{\sigma^2} = \sqrt{35417.01} = \underline{188.19}$

Zeitreihenanalyse → Zeitlich geordnete Folge von Merkmalswerten, deren Aufgabe es ist, Strukturen und Gesetzmässigkeiten einer Zeitreihe zu erkennen. *(Es wird ein Merkmal im Zeitablauf betrachtet)*

Trend: Grundrichtung der Entwicklung. I.d.R. ein glatte Kurve. Langfristig wirksame Einflüsse sollen sichtbar gemacht werden. Es gibt 2 Methoden:

1. gD=Methode der gleitenden Durchschnitte → Es erfolgt eine Glättung der Kurve dadurch, dass den einzelnen Perioden Durchschnitte zugeordnet werden.

Eignung: Idealerweise bei einer periodischen Schwankung über 4 Perioden sollte der gD4 verwendet werden. Bei 8- oder 12-Periodischer Schwankung → gD8 bzw. gD12 (kann eine weitere Glättung erzielen). Bei kurzen Betrachtungszeiträumen verschwinden mit höherer gleitenden Durchschnitten Werte am Anfang und Ende.

Ungerade Ordnung: Durchschnitt aus den Zeitreihenwerten (ungerader Anzahl) wird jeweils der mittleren Position Zugeordnet

Periode	1	2	3	4	5	6	7	8
Werte	23	18	19	25	27	22	23	29
gD3		20.0	20.7	23.7	24.7	24.0	24.7	
gD5			22.4	22.2	23.2	25.2		

$\hat{y}_2 gD3 = \frac{Zahl1 + Zahl2 + Zahl3}{3} = \frac{23+18+19}{3} = \underline{\underline{20}}$ analog bei anderen gD:

$\hat{y}_6 gD5 = \frac{Z1+Z2+Z3+Z4+Z5}{5} = \frac{25+27+22+23+29}{5} = \underline{\underline{25.2}}$ etc…

Gerade Ordnung: Man berücksichtigt immer einen Wert mehr, als man für die Ordnung braucht, wobei dann der erst und letzte Wert nur zur Hälfte berücksichtigt werden.

Periode	1	2	3	4	5	6	7
Werte	23	18	19	25	27	22	23
gD2		19.5	20.3	24.0	25.3	23.5	
gD4			21.8	22.8	23.8		

$\hat{y}_6 gD2 = \frac{\frac{Zahl1}{2} + Zahl2 + \frac{Zahl3}{2}}{2} = \frac{\frac{27}{2}+22+\frac{23}{2}}{2} = \underline{\underline{23.5}}$ analog bei anderen gD:

$\hat{y}_3 gD4 = \frac{\frac{Z1}{2}+Z2+Z3+Z4+\frac{Z5}{2}}{4} = \frac{\frac{23}{2}+18+19+25+\frac{27}{2}}{4} = \underline{\underline{21.8}}$ etc…

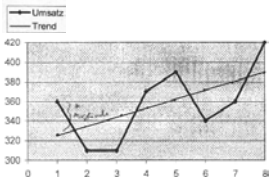
2. Methode der kleinsten Quadrate → Es wird eine Funktion gesucht, deren Werte als Trendwerte interpretiert werden. Es gibt lineare und nicht lineare Trendfunktionen (Exponential- und Potenz- und logische Funktion).

Linearer Trendverlauf: Gerade, dass die Summe der Abweichungen zwischen den effektiven Werten und den Trendwerten im Quadrat minimal sind.

Trendgerade $\hat{y} = a + b * x$ Arithm.Mittel

$b = \frac{\sum_{i=1}^{n} x_i * y_i - n * \bar{x} * \bar{y}}{\sum_{i=1}^{n} x_i^2 - n * \bar{x}^2}$ $a = \bar{y} - b * \bar{x}$ $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Quartal x_i	Umsatz y_i	$x_i * y_i$	x_i^2
1	360	360	1
2	310	620	4
3	310	930	9
4	370	1'480	16
5	390	1'950	25
6	340	2'040	36
7	360	2'520	49
8	420	3'360	64
36	2'860	13'260	204



Nicht linearer Trendverlauf: (wird hier nicht ausführlich behandelt) Angenommen, die erhobenen Daten haben einen Verlauf, der darauf hindeutet, dass er durch eine Exponentialfunktion beschrieben werden könnte, kann wie folgt gerechnet werden: $\hat{y} = a * b^x$ a>0, b>0 logarithmiert $\ln(\hat{y}) = \ln(a) + x * \ln(b)$

Periodische Schwankung: Periodisch wieder erkennbare Schwankungen um den Trend. Ermitteln wir die periodische bzw. saisonale Schwankung können wir darstellen, welche Werte aufgrund von Trend und saisonale Schwankung erwartet werden. Es werden periodisch wirksame Einflüsse dargestellt.

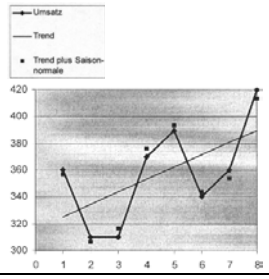
Restkomponente: Die Zahl (R_i), welche übrig bleibt, damit gilt:

Umsatz im Quartal i (2) = $Trend_i$ (3) + Saisonnorm. $Schwankung_i$ (8) + R_i (hier Unterstellung Additiver Zus'hang)

Additiver Zusammenhang → Komponenten wirken voneinander unabhängig.

Multiplikativer Zusammenhang → Komponenten voneinander abhängig: Umsatz = Trend * Period.Schwank. * R_i

Quar-tal	Um-satz	Trend	Abweichungen vom Trend im Qartal: 1.Q.	2.Q.	3.Q.	4.Q.	Saison-norm.	Saisonnor-male Sch-wankung	R_i
(1)	(2)	(3)	(4)=(2)-(3)	(5)=(2)-(3)	(6)=(2)-(3)	(7)=(2)-(3)	(8)	(9)=(3)+(8)	
1	360	325.0	35.00				31.43	356.43	3.57
2	310	334.3		-24.29			-27.86	306.43	3.57
3	310	343.6			-33.57		-27.14	316.43	-6.43
4	370	352.9				17.14	23.57	376.43	-6.43
5	390	362.1	27.86				31.43	393.57	-3.57
6	340	371.4		-31.43			-27.86	343.57	-3.57
7	360	380.7			-20.71		-27.14	353.57	6.43
8	420	390.0				30.00	23.57	413.57	6.43
		Ø der Abweichungen:	31.43	-27.86	-27.14	23.57			



Regressions- und Korrelationsrechung → Zusammenhang von zwei Merkmalen, die man kennt. *(Es werden zwei Merkmale zusammen betrachtet)*

Regression: Darstellung der Form bzw. Tendenz des Zusammenhangs zwischen den beiden Merkmalswerten. Zusammenhang kann linear, aber auch exponential sein. *(Merkmale mind. Intervall- oder verhältnisskaliert)*

- Die Methode der Kleinsten Quadrate können wir, wie bei der Zeitreihenanalyse auch hier anwenden:

Regressionsgerade $\hat{y} = a + b * x$ Arithmetisches Mittel

$b = \frac{\sum_{i=1}^{n} x_i * y_i - n * \bar{x} * \bar{y}}{\sum_{i=1}^{n} x_i^2 - n * \bar{x}^2}$ $a = \bar{y} - b * \bar{x}$ $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$

Anmerkung: Einziger Unterschied zur Zeitreihenanalyse: Hier wird nicht ein Merkmalswert im Zeitablauf, sondern zwei Merkmalswerte untersucht.

Streudiagramme sind nützlich um zu sehen, ob die Daten zueinander in Beziehung stehen oder nicht. Bei Beziehungen untereinander kann (linearer Anstieg im Streudiagramm) kann folgende Kennzahl ermittelt werden:

- Kovarianz: Ist analog zur Varianz definiert. (Varianz steht für ein Merkmal und die Kovarianz für zwei)

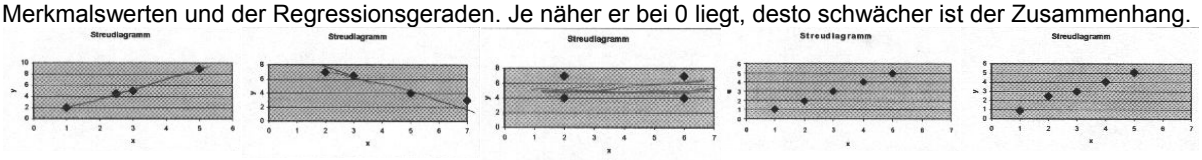
Varianz: x: $\sigma_x^2 = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x})^2$ y: $\sigma_y^2 = \frac{1}{n} * \sum_{i=1}^{n} (y_i - \bar{y})^2$ Kovarianz: $\sigma_{xy} = \frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})$ oder $\sigma_{xy} = \frac{1}{n} * \sum_{i=1}^{n} x_i * y_i - \bar{x} * \bar{y}$

Je nach „Steigung" die eine Trendgerade beschreiben würde, ist die Kovarianz positiv oder negativ. Die Kovarianz sagt uns, ob die beiden Merkmalswerte gleichläufig oder gegenläufig sind.

r=Korrelationskoeffizient, Masskorrelation: Aus Kovarianz, Standartabweichung x, Standartabweichung y. Es können nur noch Werte zwischen -1 und 1 entstehen.

$$r = \frac{\sigma_{xy}}{\sigma_x * \sigma_y} = \frac{\frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\frac{1}{n} * \sum_{i=1}^{n} (x_i - \bar{x})^2} * \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Das Vorzeichen des Korrelationskoeffizienten gibt an, ob die Merkmalswerte gleichläufig oder gegenläufig sind. Der Betrag variiert zwischen 0 und 1. Je näher er bei 1 liegt, desto stärker ist der Zusammenhang zwischen den Merkmalswerten und der Regressionsgeraden. Je näher er bei 0 liegt, desto schwächer ist der Zusammenhang.



r=0.997 r=-0.983 r=0 r=1 r=0.990

R^2=Bestimmtheitsmass: Ist normierte Kovarianz: $R^2 = r^2 = \left(\frac{\sigma_{xy}}{\sigma_x * \sigma_y}\right)^2$ falls $\sigma_{xy} > 0 \rightarrow +\sqrt{R^2}$ falls $\sigma_{xy} < 0 \rightarrow -\sqrt{R^2}$

Interpretation Bestimmtheitsmass:

1. R^2=1, wenn $\sum_{i=1}^{n} (y_i - b*(x_i - \bar{x}) - \bar{y})^2 = 0$ (wenn Abweichungen zwischen den beobachteten Werten=0)

   Je grösser die Abstände sind, desto kleiner wird R^2 $R^2 = 1 - \frac{\sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}}{\sigma_y^2}$

2. R^2 ist normierte Kovarianz
3. R^2 drückt aus, wie viel von der Gesamtvarianz durch die Regression erklärt werden kann $R^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$
4. U2=Unbestimmtheitsmass: U^2=1-R^2

ρ =Rangkorrelation: Untersucht den Zusammenhang zwischen zwei Merkmalen, wovon eines nur ordial- und das andere mindestens ordinalskaliert ist.

1. Rangordnung der beiden ordinalskalierten Merkmale erstellen. → Grad des Zusammenhangs zwischen den Merkmalen → die beiden Rangordnungen auf den Grad ihrer Übereinstimmung vergleichen
2. Sind zwei od. mehr Merkmale gleich → arithm. Mittel der Ränge zuordnen*1
3. Korrelationskoeffizient ermitteln: $\rho = r = \frac{\sigma_{xy}}{\sigma_x * \sigma_y}$ man erhält Rangkorrelation

   Sind die Merkmale vollständig, oder fast vollständig unterschiedlich (fast keine arithm. Mittel verwendet), kann folgende Formel angewandt werden: $\rho = 1 - \frac{6 * \sum_{i=1}^{n} D_i^2}{n^3 - n}$, wobei D_i=Rg(x_i)-Rg(y_i)

Weinmarke	Bewertung x_i	Preis y_i	Rang x_i	Rang y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) * (y_i - \bar{y})$	D_i	D_i^2
A	ausreichend	13.50	5	5	1.5	1.5	2.25	2.25	2.25	0	0
B	mangelhanft	15.20	6	4	2.5	0.5	6.25	0.25	1.25	2	4
C	gut	16.30	*1 2.5	3	-1	-0.5	1	0.25	0.5	-0.5	0.25
D	sehr gut	18.50	1	1	-2.5	-2.5	6.25	6.25	6.25	0	0
E	befriedigend	17.40	4	2	0.5	-1.5	0.25	2.25	-0.75	2	4
F	gut	12.90	*1 2.5	6	-1	2.5	1	6.25	-2.5	-3.5	12.25
		Mittelwert: 3.5		3.5		Summe: 17		17.5	7		20.5
						Vaianz: 2.83		2.92	Kovarianz: 1.17		
					Standartabweichung: 1.68			1.71	Rangkorrelation ρ: 0.406		

Berechnung ρ : $\rho = r = \frac{\sigma_{xy}}{\sigma_x * \sigma_y} = \frac{1.17}{1.68 * 1.71} = \underline{\underline{0.406}}$ oder $\rho = 1 - \frac{6 * \sum_{i=1}^{n} D_i^2}{n^3 - n} = 1 - \frac{6 * 20.5}{6^3 - 6} = \underline{\underline{0.414}}$

Rangkorrelation kann gleich interpretiert werden, wie der Korrelationskoeffizient. Es ist aber darauf zu achten, dass nur der Zusammenhang zwischen den Rängen untersucht wird, und nicht der Zusammenhang zwischen den Merkmalswerten selbst. Das Vorzeichen ergibt sich, je nach dem, wie man die Ränge wählt 1:sehr gut, 1: höchster Preis → positives Vorzeichen, oder 1:sehr gut, 1:tiefster Preis → negatives Vorzeichen.