KANTON
LUZERN

Bildungs- und Kulturdepartement
Kantonsschule Willisau

Auswertung von Umfragen und Experimenten

Umgang mit Statistiken in Maturaarbeiten
Realisierung der Auswertung mit Excel

Inhalt

Kenngrössen der beschreibenden Statistik

Für die Auswertung von Datenreihen werden verschiedene Kenngrössen (Lageparameter, Streuungsparameter) berechnet. In der Tabelle sind die wichtigsten Kenngrössen kurz beschrieben. Welche Grössen benutzt werden, hängt wesentlich von der Fragestellung und der gewünschten Antwort ab.

Lageparameter

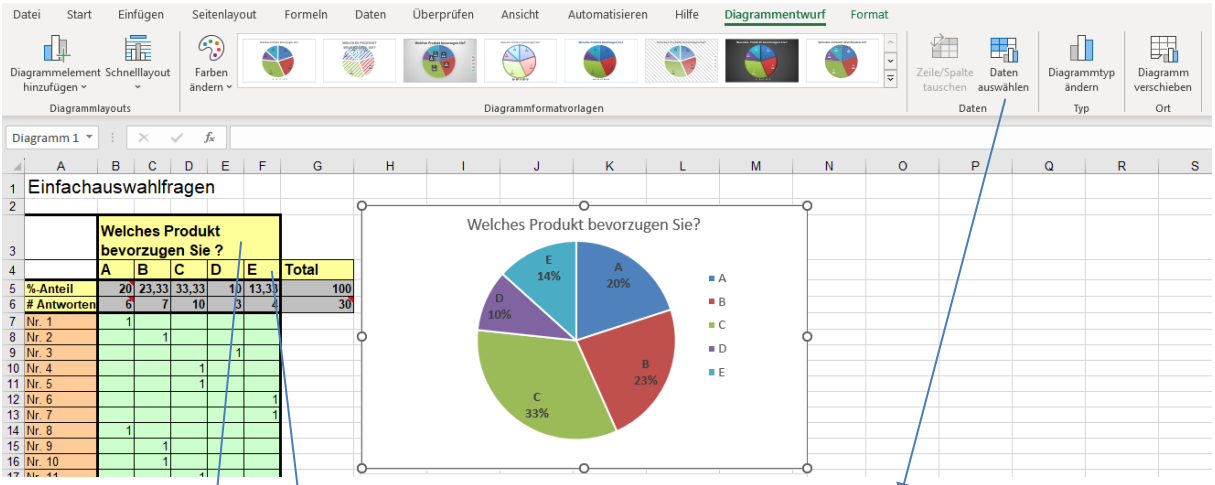
Grösse	Berechnung	Eigenschaften	Bemerkungen
Arithmetisches Mittel $\bar{x}$, AM	$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ ohne Klassen, Einzelmessungen x_i $\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \cdot h_i$ mit N Klassen der Häufigkeiten h_i **Excel**: =MITTELWERT(Bereich)	- Meist verwendeter Mittelwert - Künstliche Rechengrösse - Minimiert die Summe der Abstandsquadrate: $\sum_i (x_i - \bar{x})^2 = min$	- Starke Beeinflussung durch Ausreisser - Geeignet bei Binomial- und Normalverteilungen - Ungeeignet bei bimodalen Verteilungen
Median	Der Median oder Zentralwert ist jener Wert einer Verteilung, der die der Grösse nach geordneten Werte in 2 Hälften teilt. Links und rechts vom Median liegen je 50% der Werte. **Excel**: =MEDIAN(Bereich)	- Repräsentiert das Zentrum einer Verteilung - Liegt nie bei einem Extremalwert - Wird nicht durch die Grösse der Werte bestimmt - Wird nicht durch Ausreisser beeinflusst - Die Summe der Entfernungen zu den einzelnen Werten ist minimal: $\sum_i \lvert x_i - M \rvert = min$	Guter und häufig verwendeter Mittelwert. Eignet sich besonders für: - Durchschnittliches Einkommen - Konsumentenindex - Schulnoten - Bimodale Verteilungen - ...
Geometrisches Mittel GM	$GM = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$ **Excel**: =GEOMITTEL(Bereich)	Wird verwendet, wenn Daten multiplikativ miteinander verknüpft sind, z.B.: - Durchschnitte von Wachstumsraten - Durchschnittliche Zinsen bei Zinseszinsrechnungen	
Modus	Der mit der grössten Häufigkeit auftretende Wert einer Verteilung. **Excel**: =MODUS.EINF(Bereich)	- Einfach zu berechnen - Typischer Wert einer Verteilung	- Eignet sich, wenn ein Wert eindeutig dominiert - Bei starker Streuung wenig aussagekräftig - Es kann mehrere Modi geben

Streuungsparameter

Grösse	Berechnung	Eigenschaften	Bemerkungen
Spannweite	Die Spannweite berechnet sich aus der Differenz des grössten und des kleinsten Wertes einer Verteilung: $SW = Max(x_i) - Min(x_i)$ **Excel**: =MAX(Bereich)-MIN(Bereich)	Wird verwendet, wenn Extremwerte von Interesse sind: - Börsenkurse, Warenpreise, Schulnoten, ... - Qualitätskontrolle von Produkten	- Enthält keine Information über die Form der Verteilung - Einzelne Ausreisser bestimmen die SW
Standardabweichung	**Standardabweichung der Verteilung bzw. Grundgesamtheit:** $\sigma = \sqrt{\frac{1}{n} \cdot \sum_i (x_i - \mu)^2}$, $\mu =$ Erwartungswert der Verteilung **Excel**: =STABW.**N**(Bereich) **Stichprobenstandardabweichung**: $s = \sqrt{\frac{1}{n-1} \cdot \sum_i (x_i - \bar{x})^2}$ **Excel**: =STABW.**S**(Bereich)	- Meist verwendetes Mass für die Streuung - Bei einer den Messungen zu Grunde liegenden Normalverteilung (Binomial-verteilung) liegen 68% der Werte im Bereich $\bar{x} \pm s$ und 95% der Werte im Bereich $\bar{x} \pm 2s$ - Eignet sich gut für den Vergleich mehrerer Verteilungen	- Bei Experimenten, Umfragen, ... handelt es sich meist um **Stichproben**, $\bar{x}$ ist ein Schätzwert für den Erwartungswert μ der Grundgesamtheit bzw. der den Messungen zu Grunde liegenden Verteilung. Deshalb dividiert man zur Berechnung der Standardabweichung s der Stichprobe durch die Anzahl der Freiheitsgrade der Stichprobe: $n - 1$ und verwendet s als Schätzwert für σ. **Deshalb bei Excel =STABW.S verwenden!**
Standardfehler	**Standardabweichung des Mittelwertes** $s_{\bar{x}} = \frac{s}{\sqrt{n}}$	Sagt aus, wie gross die Standardabweichung eines gemessenen Mittelwertes ist, z.B.: 20 x würfeln: Wir erhalten einen Mittelwert $\bar{x} \cong 3.5$ mit Standardabweichung der einzelnen Würfe $s \cong 1.71$. Der Mittelwert selbst wird mit 95%-iger Sicherheit im Bereich $\bar{x} \pm 2 \cdot s_{\bar{x}}$ liegen: $3.5 \pm 2 \cdot \frac{1.71}{\sqrt{20}} \sim [2.74; 4.26]$	Vergleich von Mittelwerten: Ist der Unterschied von zwei gemessenen Mittelwerten (z.B. Klassendurchschnitte) signifikant? (---> T-Test)

Erstellen von Diagrammen

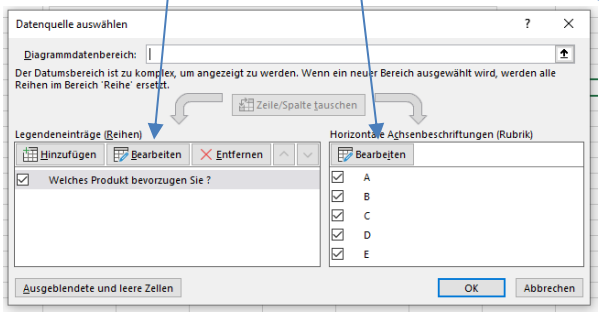
Datenreihen auswählen und bei **Einfügen** das geeignete Diagramm auswählen. Dann können bei **Diagrammentwurf** alle Einstellungen vorgenommen werden:



Legendeneinträge: Bei *Bearbeiten* Feld B3 auswählen (erscheint als Titel)

Bei *Daten auswählen* erscheint das Fenster *Datenquelle auswählen*

Horizontale Achsenbeschriftungen: Bei *Bearbeiten* den Bereich B4:F4 auswählen

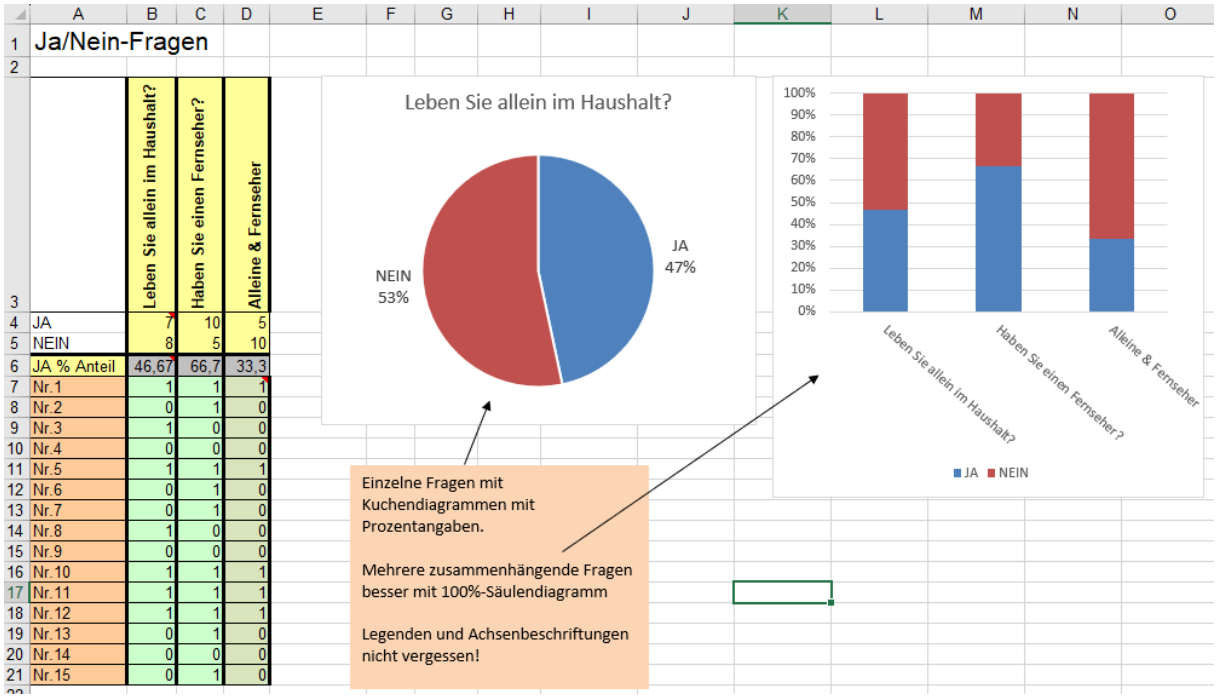


Ja/Nein-Fragen

In einer Tabelle werden die Antworten der Fragebögen im Bereich B7:D21 eingegeben.
Sinnvoll sind die Berechnung der prozentualen Ja/Nein-Anteile (Zeile 6) sowie der Kombination einzelner Fragen (Bsp. Wer lebt allein im Haushalt und hat einen Fernseher?) (Spalte D).

Diagramme: Bei einzelnen Fragen **Kuchendiagramme** mit Prozentangaben
Bei mehreren zusammenhängenden Fragen **Säulendiagramm** mit Prozentangaben

Völlig sinnlos: Mittelwerte, Standardabweichung, ...



Excelformeln:

- In den Zeilen 4 und 5 wird die Summe der Antworten (JA: 1, NEIN: 0) mit *ZÄHLENWENN* berechnet,
  z.B. B4: *=ZÄHLENWENN(B7:B21;1)*

- In Zeile 6 wird der prozentuale Anteil der "JA" berechnet, z.B. B6: *=B4/(B4+B5)*100*

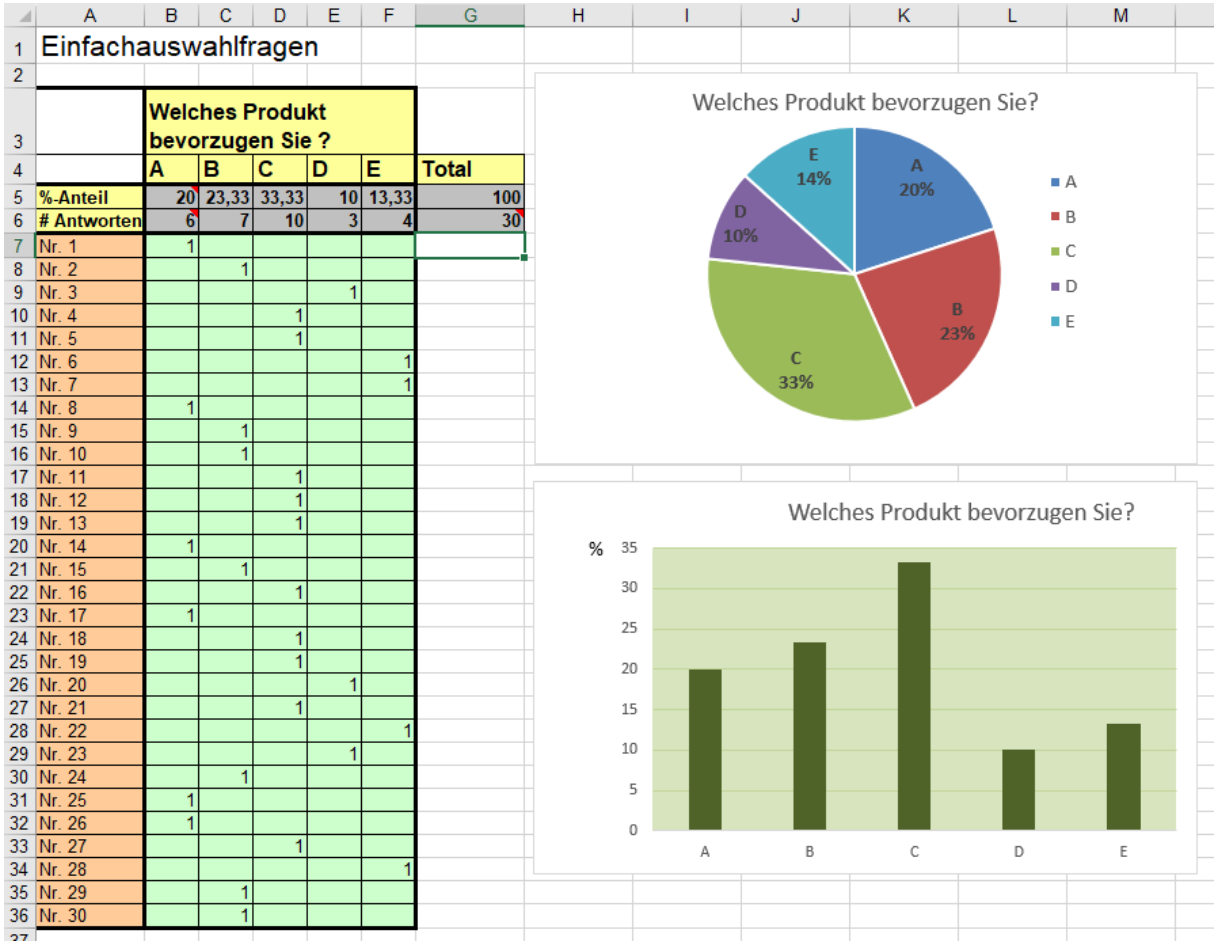
- Mehrere Fragen können auch kombiniert werden (Spalte D). Die Berechnung erfolgt mit einer bedingten Anweisung, z.B. D7: *=WENN(B7+C7=2;1;0)*.
  Falls beide Fragen mit JA beantwortet werden, wird der Wert 1 sonst der Wert 0 in das Feld D7 geschrieben.

Allgemein lautet die bedingte Anweisung:
=WENN(Wahrheitstest;Wert_wenn_wahr;Wert_wenn_falsch)

Einfachauswahlfragen

Von den möglichen Antworten muss genau eine angegeben werden. Die Antworten werden wie bei den Ja/Nein-Fragen in einer Tabelle zusammengefasst und ausgezählt (Zeile 6 mit *ZÄHLENWENN*)
Wichtig ist auch hier der prozentuale Anteil der Antworten untereinander (Zeile 5)

Diagramme: Kuchen- oder Säulendiagramme mit den prozentualen Anteilen.
Völlig sinnlos: Mittelwerte, Standardabweichung, Liniendiagramme, ...



	A	B	C	D	E	Total
1	Einfachauswahlfragen					
2						
3		Welches Produkt bevorzugen Sie ?				
4		A	B	C	D	E	Total
5	%-Anteil	20	23,33	33,33	10	13,33	100
6	# Antworten	6	7	10	3	4	30
7	Nr. 1	1					
8	Nr. 2		1				
9	Nr. 3				1		
10	Nr. 4			1			
11	Nr. 5			1			
12	Nr. 6					1	
13	Nr. 7					1	
14	Nr. 8	1					
15	Nr. 9		1				
16	Nr. 10		1				
17	Nr. 11			1			
18	Nr. 12			1			
19	Nr. 13			1			
20	Nr. 14	1					
21	Nr. 15		1				
22	Nr. 16			1			
23	Nr. 17	1					
24	Nr. 18			1			
25	Nr. 19			1			
26	Nr. 20				1		
27	Nr. 21			1			
28	Nr. 22					1	
29	Nr. 23				1		
30	Nr. 24		1				
31	Nr. 25	1					
32	Nr. 26	1					
33	Nr. 27			1			
34	Nr. 28					1	
35	Nr. 29		1				
36	Nr. 30		1				

Excelformeln:

Zeile 6: Summe der JA-Antworten, z.B. *=ZÄHLENWENN(B7:B36;1)*
Zeile 5: Prozentualer Anteil der JA-Antworten, z.B. *=B6/$G$6*100* . Das $-Zeichen bewirkt einen absoluten Bezug, d.h. beim Kopieren der Formel in die Felder C5-F5 wird der Divisor G6 beibehalten und nicht verschoben.
G5: *=SUMME(B5:F5)* Die Prozentsumme muss notwendigerweise 100% ergeben (Kontrolle).

Mehrfachauswahlfragen

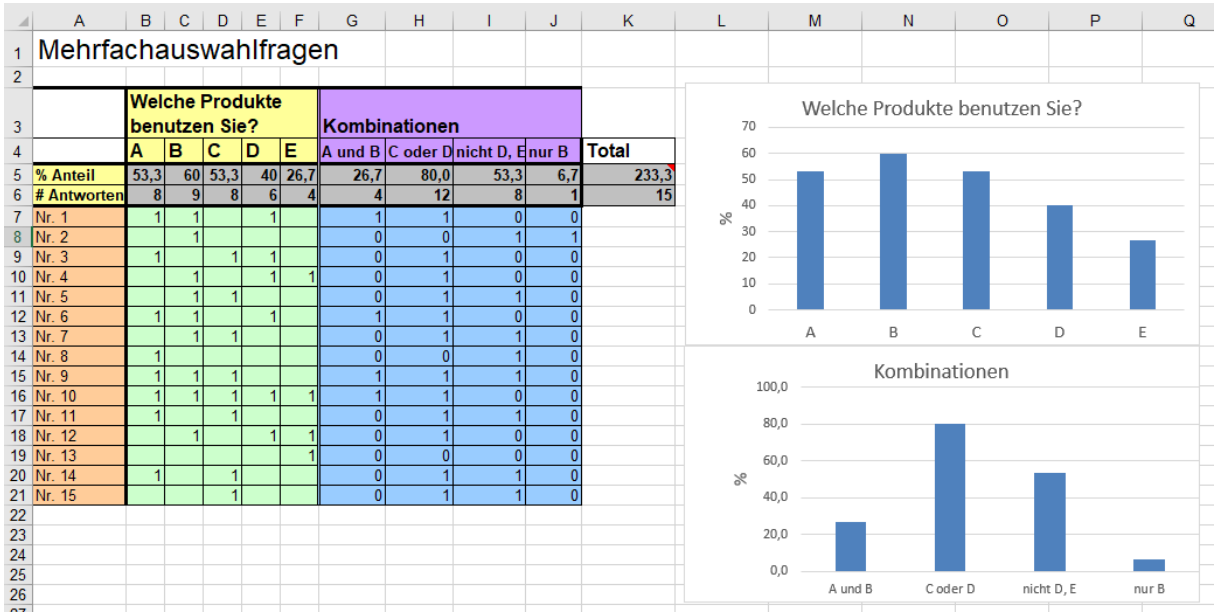
Bei diesem Fragetyp können mehrere Antworten angekreuzt werden. In der Tabelle werden die Antworten der 15 Fragebögen im Bereich B7:F21 festgehalten. In den Zeilen 6 und 5 werden die Anzahl der positiven Antworten sowie deren prozentualer Anteil berechnet. In den Spalten G bis J sind einige mögliche Kombinationen berechnet.

Prozentualer Anteil: Dieser berechnet sich aus der Anzahl positiver Antworten dividiert durch die Anzahl Fragebögen. Die Prozentsumme beträgt hier meist nicht 100%, da ja mehrere oder gar keine Produkte angekreuzt werden können. In K5 wird die Prozentsumme der Antworten A – E berechnet, ergibt hier 233.3%. Dies bedeutet, dass durchschnittlich 2.333 dieser Produkte A – E benutzt werden.

Diagramme: Säulendiagramme mit den prozentualen Anteilen.

Völlig falsch: Kuchendiagramme, da diese von der Prozentsumme 100 ausgehen.

Kombinationen: In den Spalten G bis J werden einige mögliche Kombinationen berechnet. Fragestellungen wie "Wer benutzt A und B" oder "Wer benutzt nur B und keine anderen" können bei diesem Fragetyp von Bedeutung sein. Unten werden die Excelformeln erklärt, die Werte sollen nicht von Hand berechnet werden.



Excelformeln:

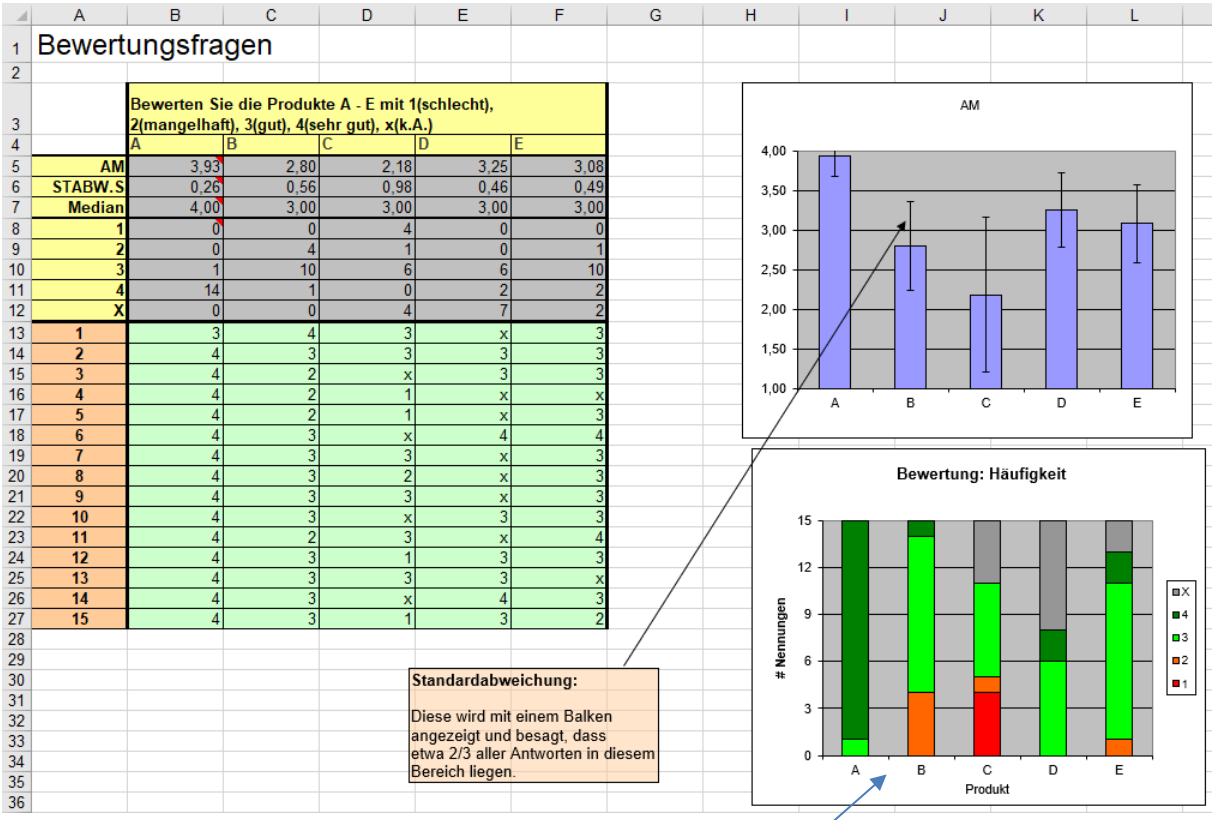
B6:J6: Anzahl Antworten *=ZÄHLENWENN(B7:B21;1)* Dieser Befehl zählt die Zahl der 1-er im Bereich B7:B21 und liefert hier das Resultat 8.
B5: Prozentualer Anteil: *=B6/$K$6*100* . Im absolut bezogenen Feld K6 steht die Zahl der Fragebögen.
G7: Kombination "A und B": *=WENN(SUMME(B7:C7)=2;1;0)*
H7: Kombination "C oder D": *=WENN(SUMME(C7:D7)>0;1;0)*
I7: Kombination "Weder D noch E": *=WENN(SUMME(E7:F7)=0;1;0)*
J7: Kombination "Nur B": *=WENN(UND(C7=1;B7+D7+E7+F7=0);1;0)*

Bewertungsfragen

Auswertung: Mittelwert, Standardabweichung und Median in den Zeilen 5 – 7.
Häufigkeitsverteilung der gegebenen Antworten in den Zeilen 8 – 12.
Die **Standardabweichung** ist hier aufschlussreich, da sie aussagt, wie stark die Meinungen der Testpersonen divergieren.
Bei der **Häufigkeitsverteilung** können Absolutwerte (im Bsp.) oder auch Prozentwerte verwendet werden (je nach Fragestellung).

Diagramme: Säulendiagramm der Mittelwerte und Standardabweichung
Anstelle des AM kann auch der Median verwendet werden.
Säulendiagramm der Häufigkeitsverteilung.
Die Häufigkeitsverteilung einer einzelnen Frage kann auch mit einem Kuchendiagramm dargestellt werden.

Völlig falsch: Liniendiagramme, Kurven



	A	B	C	D	E
1	Bewertungsfragen				
3		Bewerten Sie die Produkte A - E mit 1(schlecht), 2(mangelhaft), 3(gut), 4(sehr gut), x(k.A.)			
4		A	B	C	D	E
5	AM	3,93	2,80	2,18	3,25	3,08
6	STABW.S	0,26	0,56	0,98	0,46	0,49
7	Median	4,00	3,00	3,00	3,00	3,00
8	1	0	0	4	0	0
9	2	0	4	1	0	1
10	3	1	10	6	6	10
11	4	14	1	0	2	2
12	X	0	0	4	7	2
13	1	3	4	3	x	3
14	2	4	3	3	3	3
15	3	4	2	x	3	3
16	4	4	2	1	x	x
17	5	4	2	1	x	3
18	6	4	3	x	4	4
19	7	4	3	3	x	3
20	8	4	3	2	x	3
21	9	4	3	3	x	3
22	10	4	3	x	3	3
23	11	4	2	3	x	4
24	12	4	3	1	3	3
25	13	4	3	3	3	x
26	14	4	3	x	4	3
27	15	4	3	1	3	2

Excelformeln:
B5: Arithmetisches Mittel: *=MITTELWERT(B13:B27)*
B6: Standardabweichung: *=STABW.S(B13:B27)*
B7: Median: *=MEDIAN(B13:B27)*
B8: Häufigkeit: *=ZÄHLENWENN(B$13:B$27;$A8)*
Die notwendigen Absolutbezüge beachten!

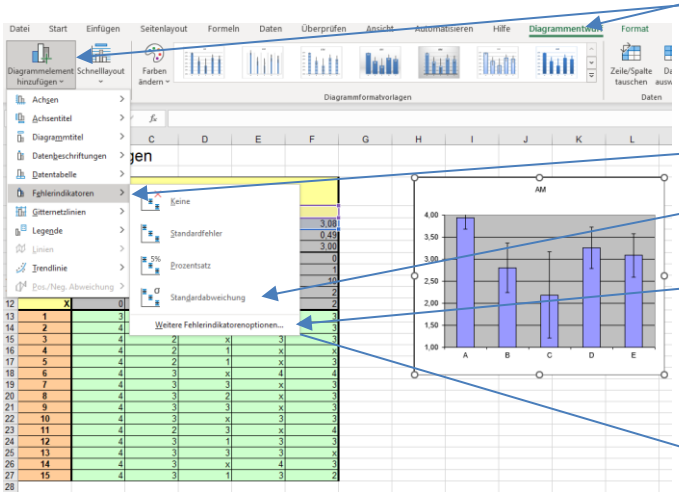
- positive Antworten oben, negative unten

- Klares Farbkonzept: z.B.
  Rot - negativ, grün (blau) - positiv, grau - unbestimmt
  Damit wird auf einen Blick klar, ob eine Frage positiv (A) oder eher negativ (C) beantwortet wird

- y-Achse in % oder absolut

Einfügen der Standardabweichung

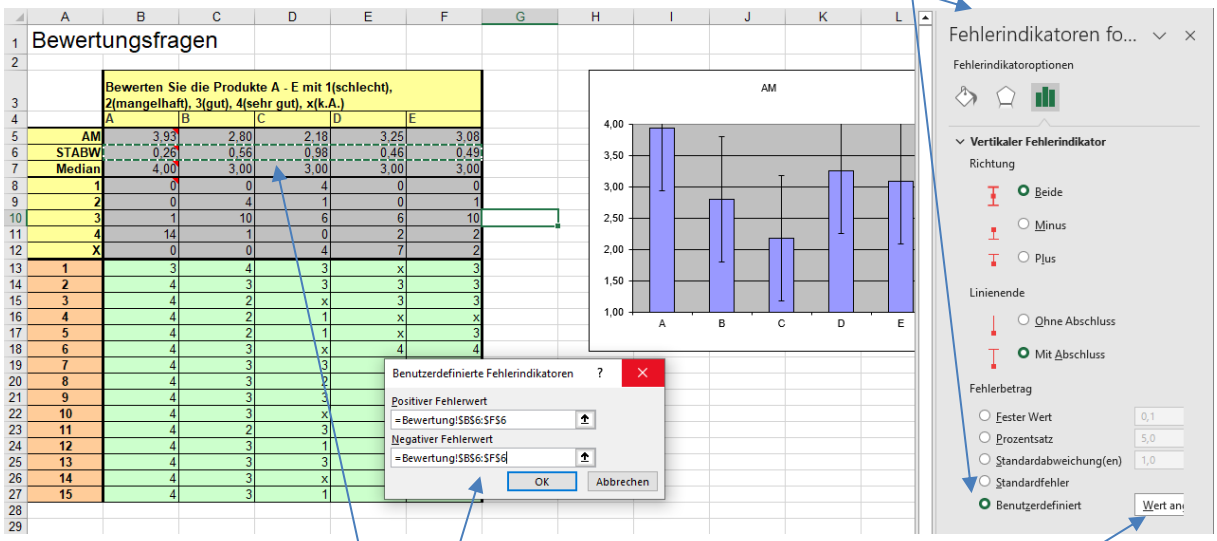
Im Menu *Diagrammentwurf* den Punkt *Diagrammelement hinzufügen* auswählen



Fehlerindikatoren ---> Standardabweichung

Weitere Fehlerindikatoroptionen:

Fehlerbetrag benutzerdefiniert



Wert angeben

Werte der Standardabweichung der Tabelle entnehmen (Bereich B6:F6)

Häufigkeitstabellen

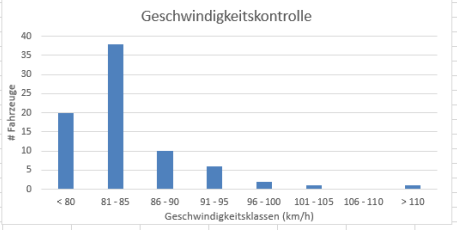
Viele Experimente liefern kontinuierlich verteilte Werte (Messungen in BI, PS, CH, ...).
Die Messwerte werden erfasst und in Häufigkeitsklassen zusammengefasst.
In folgendem Experiment wurde die Geschwindigkeit von Fahrzeugen ausserorts – also erlaubte 80 km/h – gemessen. Um herauszufinden, wie oft welche Art Busse verhängt wird, werden die gemessenen Geschwindigkeiten in Häufigkeitsklassen eingeteilt:

Häufigkeit, Klassenbildung

Messung der Geschwindigkeit in einer 80-er Zone:

Tempo (km/h)	# Fahrzeuge
64	3
66	1
72	2
73	1
78	8
79	5
80	7
81	12
82	15
83	9
84	2
86	6
87	4
92	3
94	3
96	2
101	1
136	1

Klassen	# Fahrzeuge	Bussen
< 80	20	
81 - 85	38	Toleranz
86 - 90	10	40
91 - 95	6	80
96 - 100	2	120
101 - 105	1	250
106 - 110	0	500
> 110	1	800



- 5 bis max. 10 Klassen
- Klassenbreite regelmässig
- Klassengrenzen praktisch wählen

Auswertung: Mittelwert, Standardabweichung, Median
Tabelle mit absoluten oder relativen Werten (%).
Relative Häufigkeit: Häufigkeit einer Klasse in %
Diagramme: Säulendiagramm, bei relativen Werten evtl. Kuchendiagramm
Völlig falsch: Liniendiagramme, Kurven

Lineare Regression, Korrelation

Häufig hängt eine Messgrösse y von einer Variablen x ab, z.B. Baumstammdicke (y) von der Zeit (x). Diese gemessenen Wertepaare (x/y) werden in einer Tabelle festgehalten und in einem Punktdiagramm als Punkte in einem Koordinatensystem dargestellt.
Durch diese Punkte soll eine Gerade (Regressionsgerade, linearer FIT) gelegt werden. Die Steigung m und der y-Achsenabschnitt q der Geraden g: y = mx + q werden so berechnet, dass die Summe der Abstandsquadrate der Messpunkte zur Geraden minimal wird (Ausgleichsrechnung):

$$\sum_i (y_i - g(x_i))^2 = min$$

Excel sowie die Taschenrechner sind in der Lage, diese Regressionsgerade auszurechnen.
Es gibt auch nichtlineare Regressionsmodelle, siehe nächste Seite.

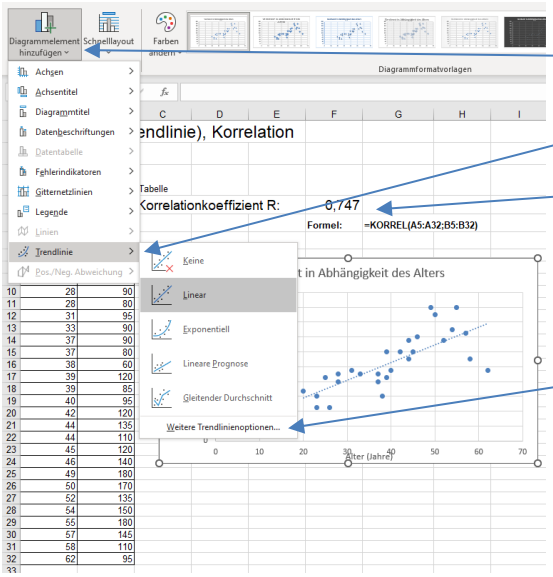
Der *Korrelationskoeffizient* R bzw. das *Bestimmtheitsmass* R^2 sagt aus, wie stark der Zusammenhang zwischen den Messgrössen x und y ist. **Es wird nichts darüber ausgesagt, ob dieser Zusammenhang *kausal* oder *zufällig* ist**.

Für eine vernünftige Korrelationsrechnung müssen genügend Datenpunkte vorhanden sein (N > 10).

$R^2 = 1$: Maximale Korrelation, bei linearer Regression liegen alle Messpunkte auf einer Geraden

$R^2 > 0.7$: Starke Korrelation

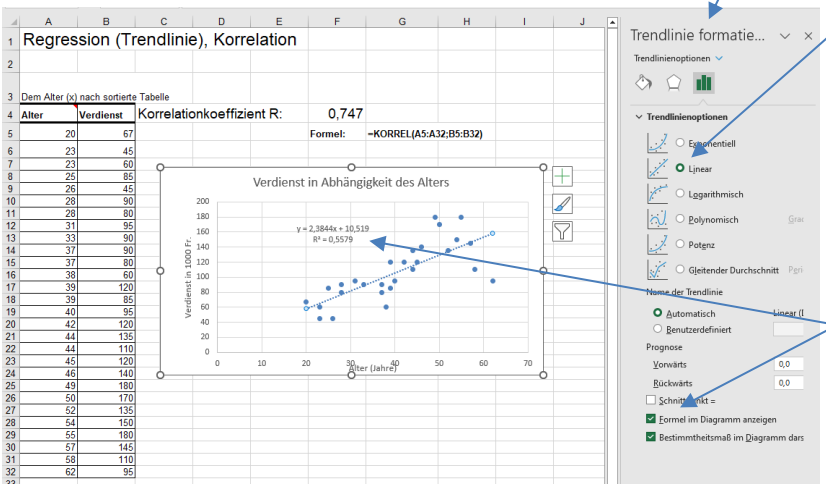
$R^2 < 0.3$: kein Hinweis auf eine Abhängigkeit zwischen x und y gemäss gewähltem Regressions-modell, bei linearer Regression: Gerade passt nicht zu den vorgegebenen Werten



Diagrammentwurf
--> Diagrammelement hinzufügen, Trendlinie

Der Korrelationskoeffizient kann auch aus den Tabellenwerten berechnet werden

Weitere Trendlinienoptionen: Die Trendline kann hier formatiert werden



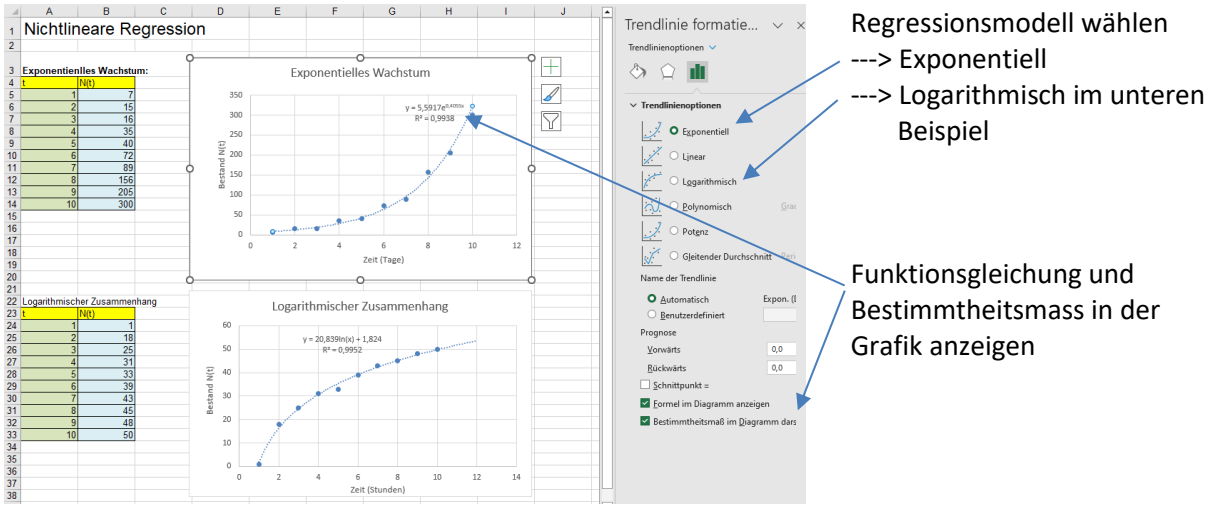
Regressionsmodell wählen

Geradengleichung und Bestimmtheitsmass in der Grafik anzeigen

Nichtlineare Regression

Nicht immer ist der Zusammenhang zwischen den Messgrössen linear. Bei Wachstumsproblemen suchen wir z.B. einen exponentiellen Zusammenhang, bei der Abhängigkeit des PH-Wertes von der Ionenkonzentration einen logarithmischen.

Da solche Modelle durch Linearisierung realisiert werden, wird ebenfalls der Korrelationskoeffizient R bzw. das Bestimmtheitsmass R^2 mit der gleichen Aussagekraft berechnet.



Regressionsmodell wählen
---> Exponentiell
---> Logarithmisch im unteren Beispiel

Funktionsgleichung und Bestimmtheitsmass in der Grafik anzeigen

Tipps:

Für die Bewertung nicht zu viele Stufen benutzen (Maximal 5: -- / - / 0 / + / ++),
im Beispiel wird mit 4 Stufen von schlecht (1) bis sehr gut (4) bewertet.

Die Fragen neutral stellen, die Fragestellung beeinflusst die Bewertung massiv.

Bei Diagrammen nicht vergessen: Achsenbeschriftung, korrekte Achsenskalierung.
Diagramme einfach gestalten, keine Farb- und Effektorgien.

Grosse Tabellen gehören in den Anhang.

Resultate vorsichtig interpretieren:
Beispiel: Produkt D wird mit 3.25, Produkt E mit 3.08 bewertet. Ob dieser Unterschied tatsächlich signifikant ist, muss mit einem geeigneten Test (hier: T-Test für unabhängige Stichproben) überprüft werden.

4. Auflage
Diese Broschüre hilft beim Verfassen und Betreuen von Maturaarbeiten.
Die 4. Auflage beinhaltet die Anpassungen mit Office 365. Insbesondere hat sich die Berechnung der Korrelation und der Regression stark geändert.
Die erwähnten Beispiele sind als Excel-Vorlage erhältlich und können von der Website heruntergeladen werden.
Für weitere Fragen und Auskünfte stehe ich gerne zur Verfügung,

Bernhard Scheel
KSW 2023