



Beate Sick, Oliver Dürr

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

beate.sick@zhaw.ch

Chur, 12. Mai 2015

1. Grundbegriffe der Datenanalyse

Grundgesamtheit, Urliste, Stichprobe, Stichprobenumfang, Rang

2. Univariate Daten charakterisieren, ordnen und klassieren

Rangliste, Klasseneinteilung, Häufigkeitstabelle

3. Univariate Daten visualisieren

Balkendiagramm, Kuchendiagramm, Histogramm, Boxplot

4. Diagramme charakterisieren und interpretieren

symmetrisch, schief, unimodal, multimodal

5. Bivariate Daten charakterisieren, visualisieren und interpretieren

6. Lagemasse berechnen und interpretieren

Mittelwert, Median, Modus

7. Streumasse berechnen und interpretieren

Standardabweichung, Quartilsdifferenz

8. Datengewinnung und Datenqualität beurteilen

9. Ausreisser und Extremwerte beurteilen

Lernziele

Statistische “Lesefähigkeit” erlangen

- Datendarstellung und Studienbeschreibungen in Medien lesen können
- Manipulativen Grafen und falschen Schlüssen nicht aufsitzen
- Argumentationen nachvollziehen und kritisch hinterfragen können

Statistische Fertigkeiten erlangen

- Daten organisieren,
- Erstellen von geeigneten Grafiken,
- Berechnung von geeigneten Kennzahlen,

Zweck und Logik einer statistischen Untersuchung verstehen

- **Daten variieren!**
- Arbeiten mit Stichproben und Rückschlüsse auf die Grundgesamtheit
- nur mit geplanten Experimenten kann Kausalität nachgewiesen werden

Zurich University
of Applied Sciences



Woher stammen Daten?

- **Geplante Experimente**
so kann Kausalität bewiesen werden
 - z.B. Wirksamkeit von Medikamenten
- **Beobachtungsstudien**
 - z.B. gezählte Anzahl Passagiere pro Tag
- **Umfragen**
 - z.B. wie viele % Kunden würden ein neues Produkt kaufen?
- **Datensammlungen**
 - z.B. lassen sich Kundengruppen mit charakteristischem Verhalten aus Geschäftsdaten erkennen?

Wir sammeln Daten

Erhebung persönliche Informationen

Klasse:

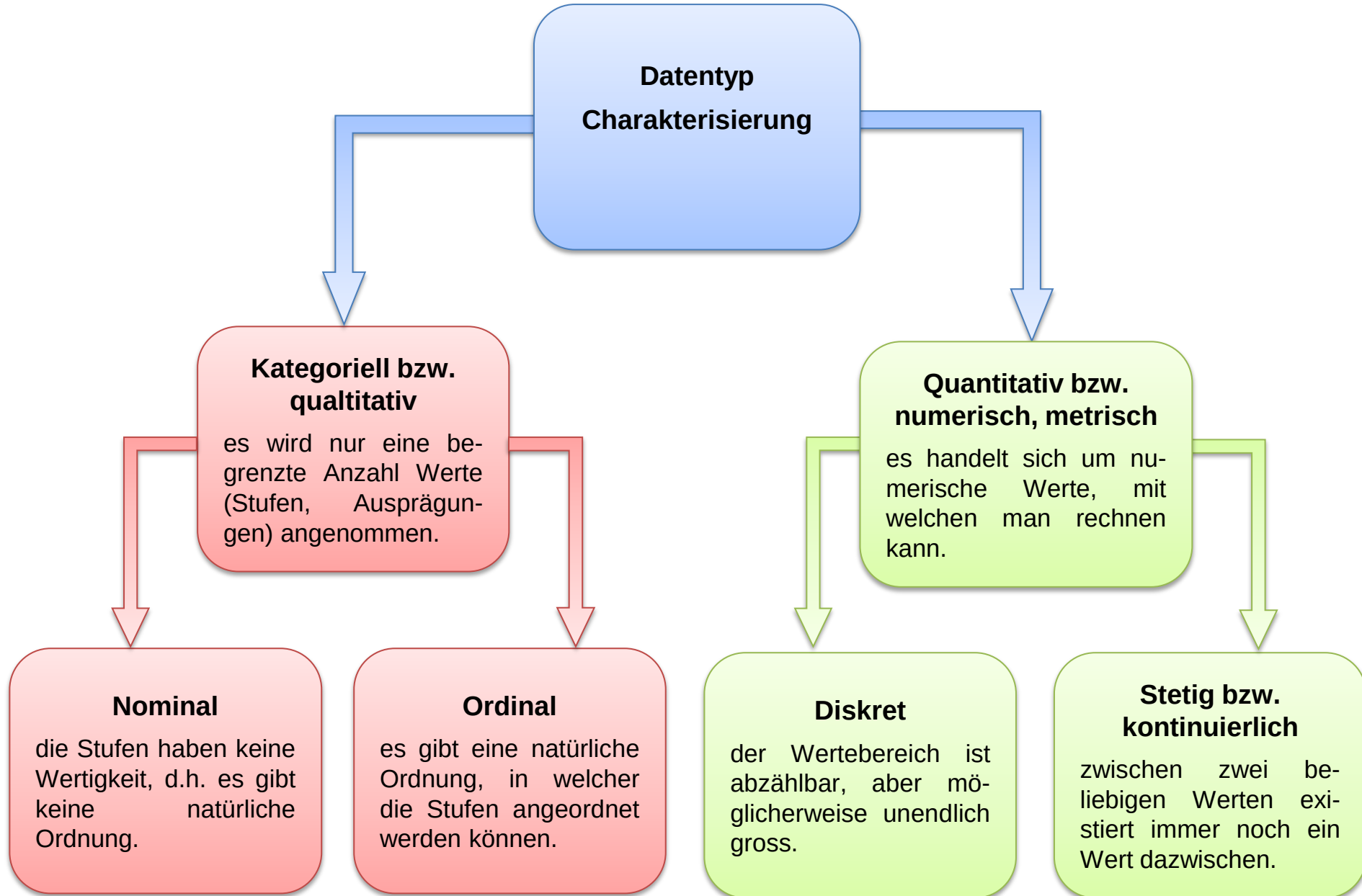
Semester:

Bitte beantworten Sie die nachfolgenden Fragen zu Ihrer Person. In Klammern ist jeweils das Format angegeben, in welchem die Antwort erfolgen soll. Verwenden Sie bitte das angegebene Format, bzw. bei einer Auswahl sind nur Antworten aus dieser Auswahl erlaubt.

Fragen	Antworten
Alter (in Jahren)	
Geschlecht (m/w)	
Augenfarbe (blau, braun, grün, haselnuss)	
Natürliche Haarfarbe (schwarz, braun, rot, blond)	
Körpergrösse (in cm)	
Armspanne (in cm)	
Handspanne* der grösseren Hand (in cm)	
Fusslänge des grösseren Fusses (in cm)	
Anzahl Geschwister	
Mitglied in einem Sportverein (ja/nein)	
Händigkeit (links, rechts)	

* Ausgehend von der gespreizten Hand, die Distanz in cm zwischen Fingerspitze des kleinen Fingers und des Daumens.

Welche Datensorten/Variablentypen gibt es?



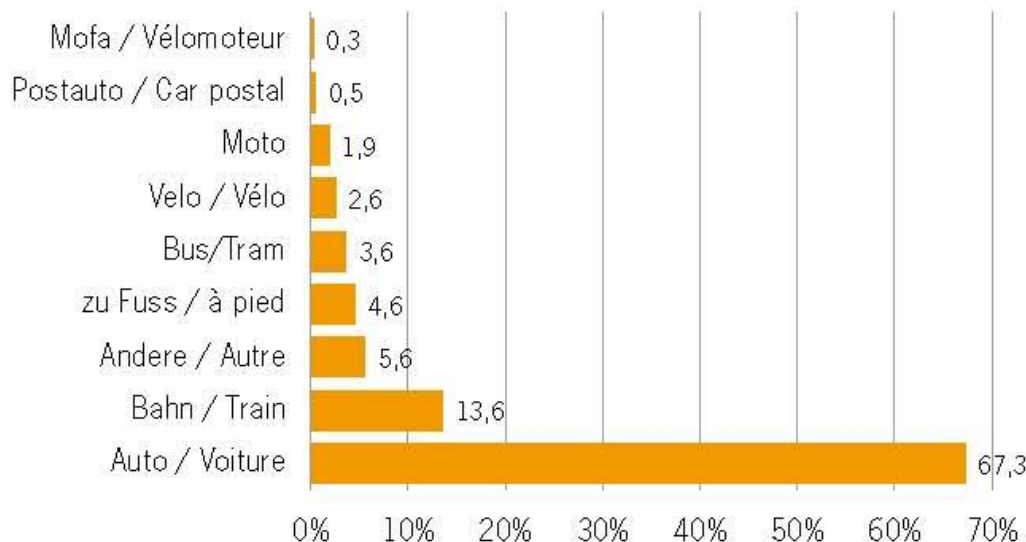
Visualisierung kategoriieller Daten

Balkendiagramm, Kuchendiagramm

Personenverkehr / Transport de personnes

Verkehrsmittelbenutzung¹ 2000, im Inland in %

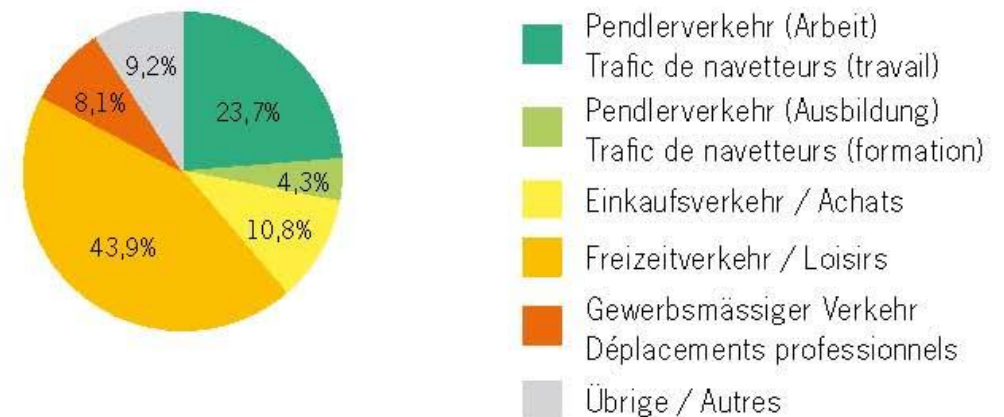
Utilisation des moyens de transports¹, en 2000, à l'intérieur du pays en %



PANORAMA © Bundesamt für Statistik, Februar 2007

Verkehrszwecke¹ 2000

Motifs de déplacement¹, en 2000

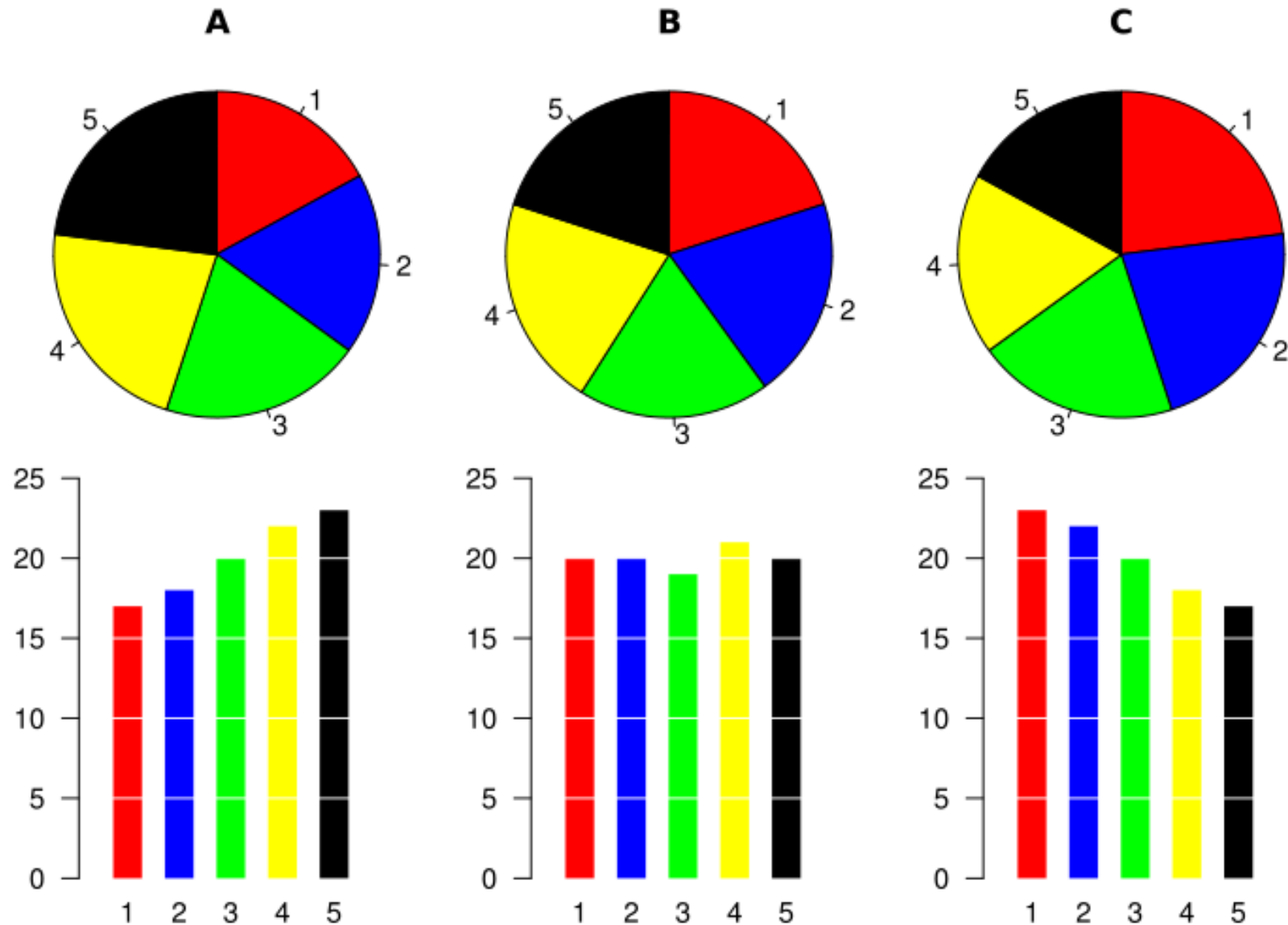


1 Anteile bezogen auf die täglich zurückgelegten km pro Person
Pourcentages des déplacements quotidiens en km par personne

PANORAMA © Office fédéral de la statistique, février 2007

Anmerkung: Auch numerische diskrete Daten können so dargestellt werden, falls es nur eine überschaubare Anzahl von möglichen diskreten Werten gibt.

Ein Experiment zur visuellen Wahrnehmung



Balkendiagramm oder Kuchendiagramm??

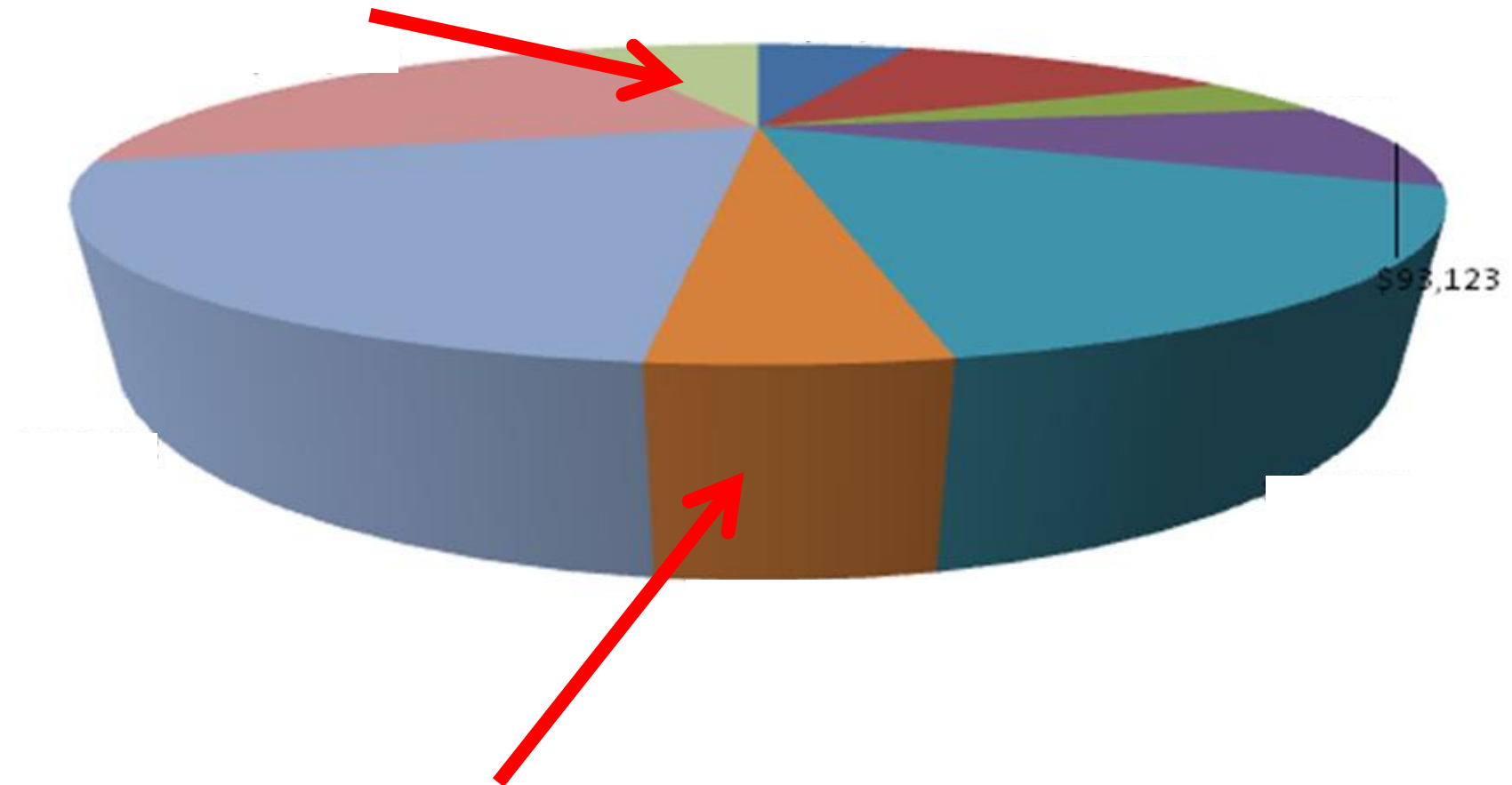
Kuchendiagramme sind weit verbreitet, bieten aber gegenüber Balkendiagrammen fast nur Nachteile. Der Grund:

das menschliche Auge kann Flächen nicht gut miteinander vergleichen

- => **es ist mit Kuchendiagrammen schwierig(er), die einzelnen Merkmale nach deren Häufigkeit zu ordnen (auch die Farbenwahl kann zu Schwierigkeiten beim Flächenvergleich führen)**
- => **ein Kuchendiagramm kann nur relative Häufigkeiten darstellen**
- => **es sollten bevorzugt Balkendiagramme eingesetzt werden**

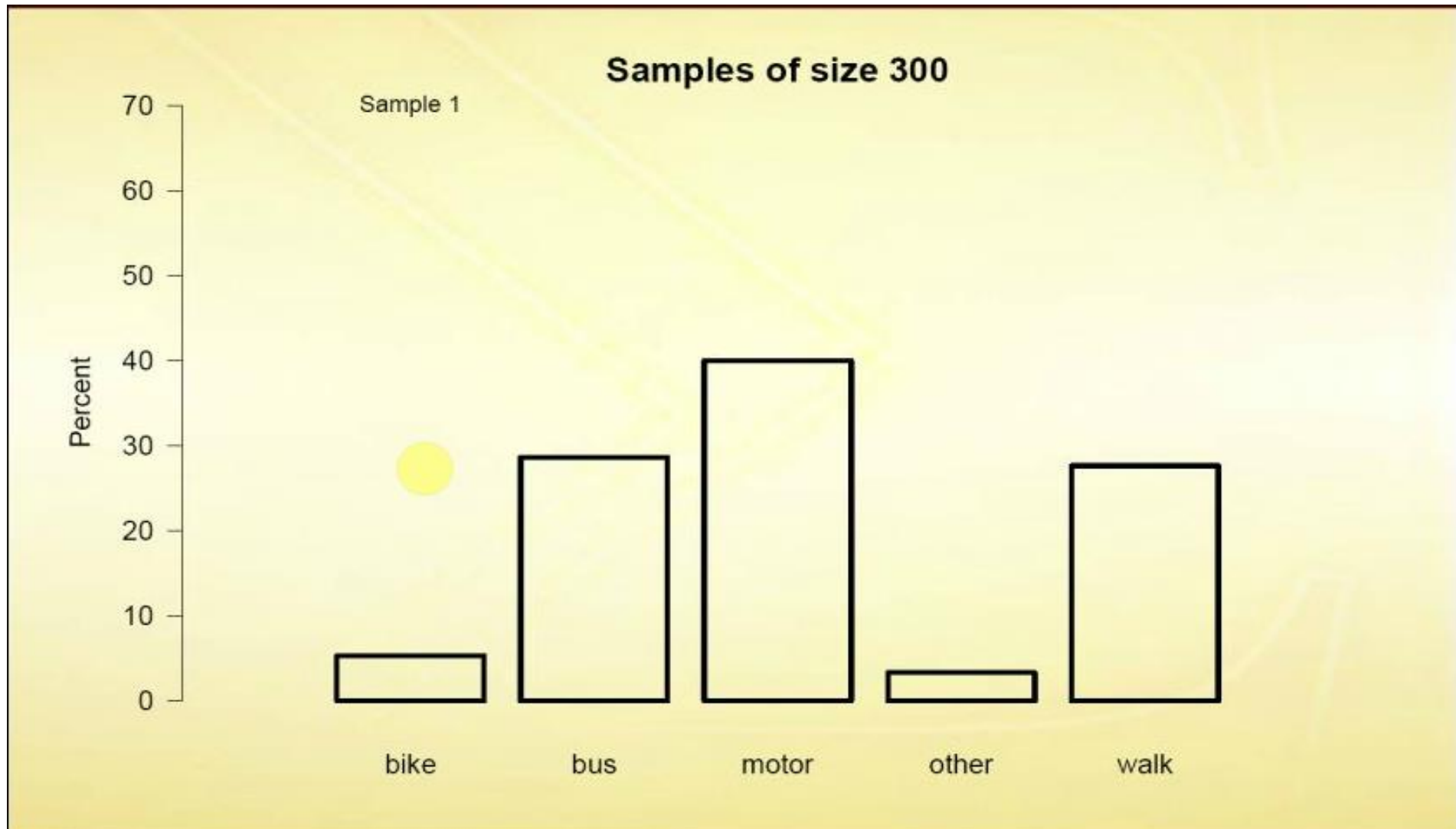
Welcher Anteil ist grösser?

Order Amounts by Salesperson



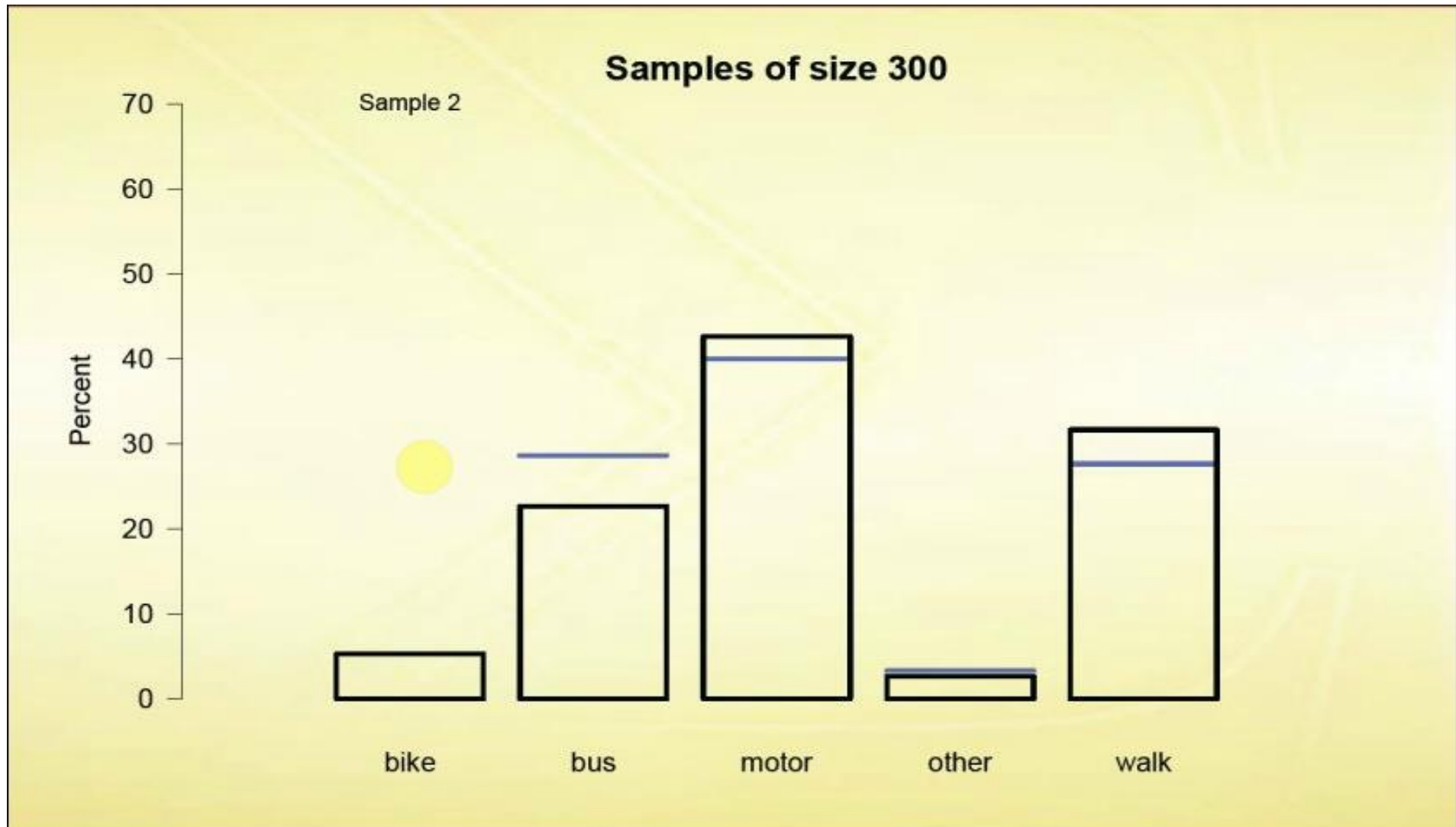
Wie aussagekräftig ist eine Stichprobe mit $n=300$?

Wie kommen neuseeländische Schüler in die Schule?



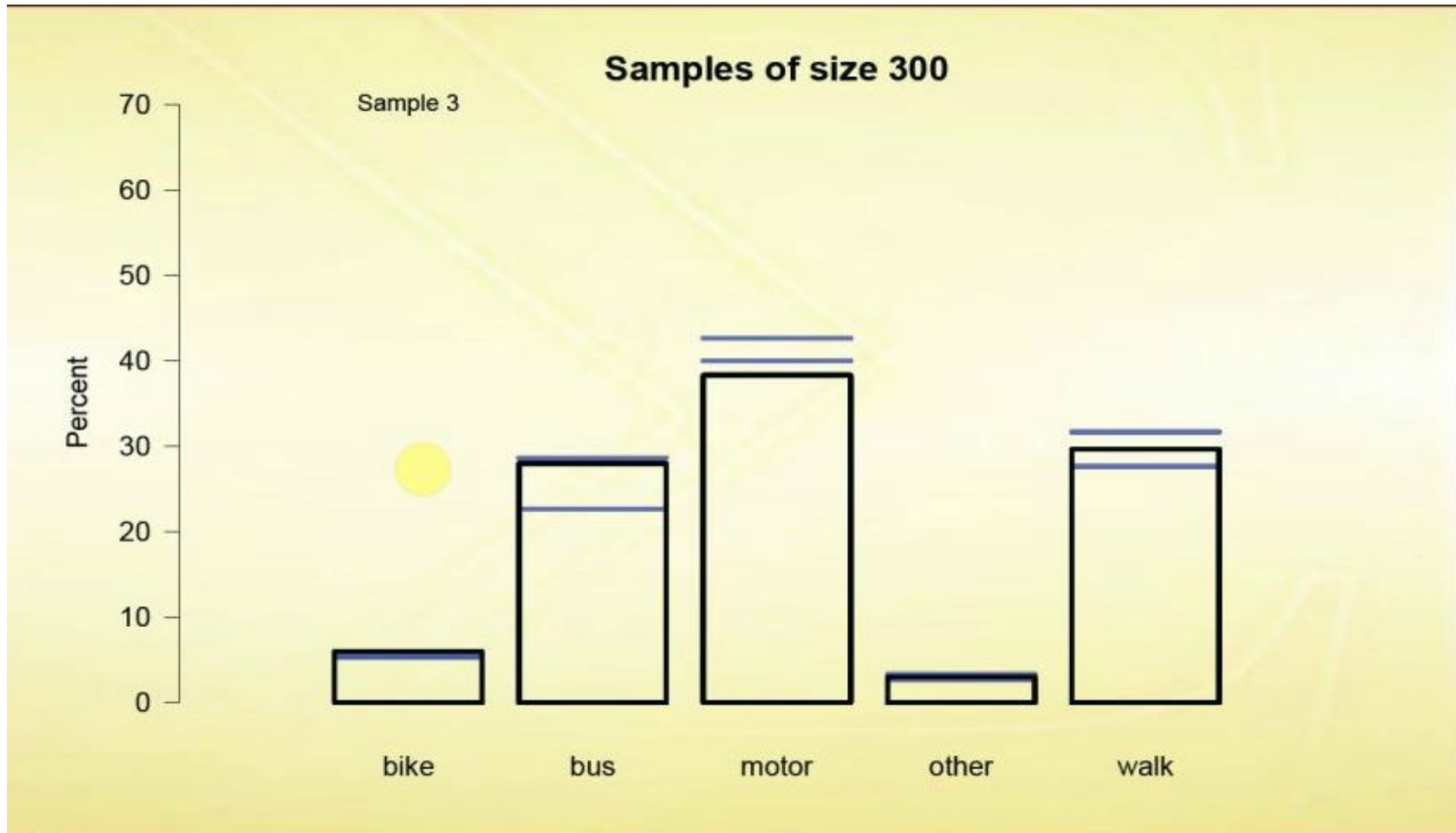
Wie aussagekräftig ist eine Stichprobe mit $n=300$?

Wie kommen neuseeländische Schüler in die Schule?



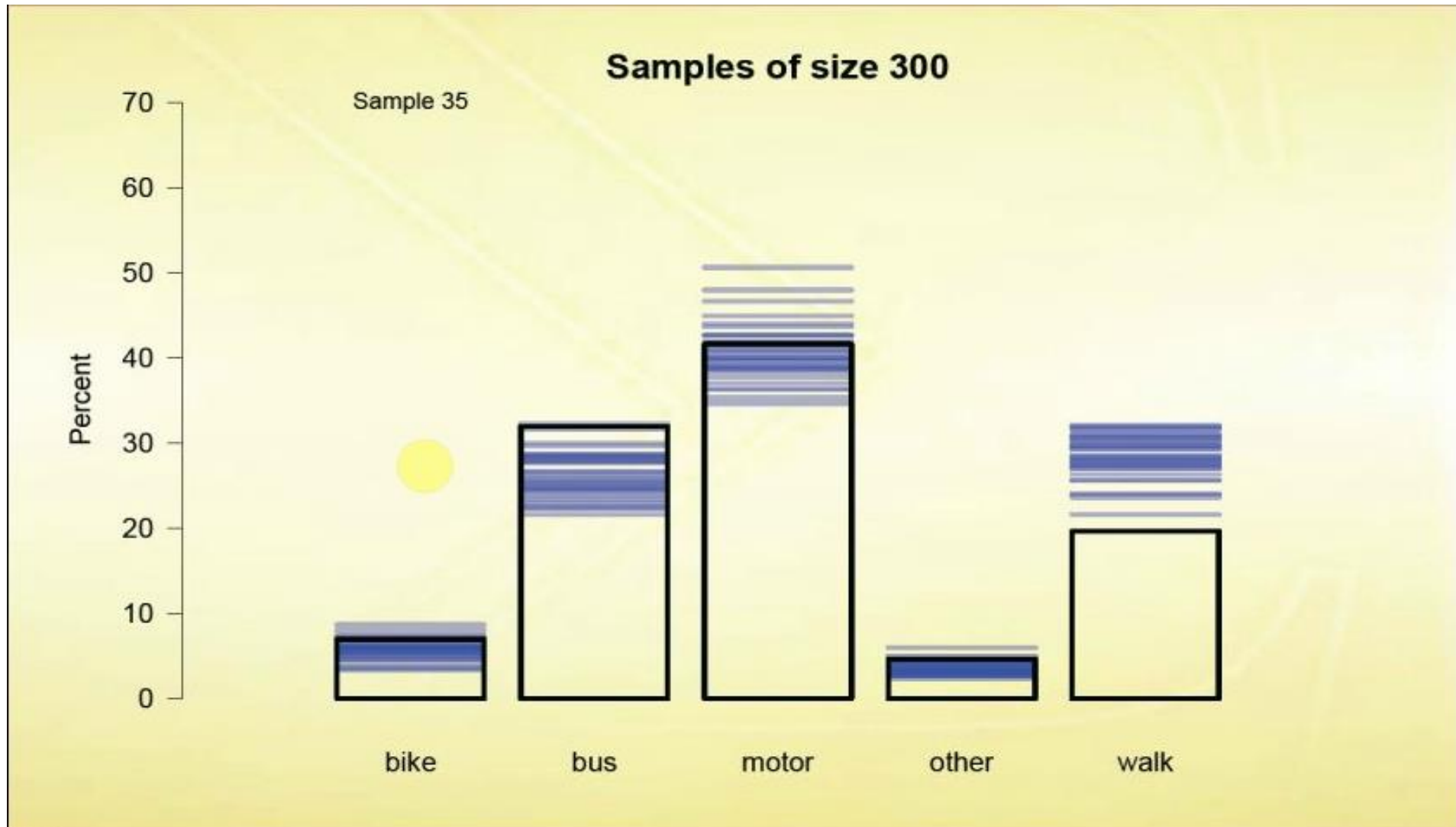
Wie aussagekräftig ist eine Stichprobe mit $n=300$?

Wie kommen neuseeländische Schüler in die Schule?



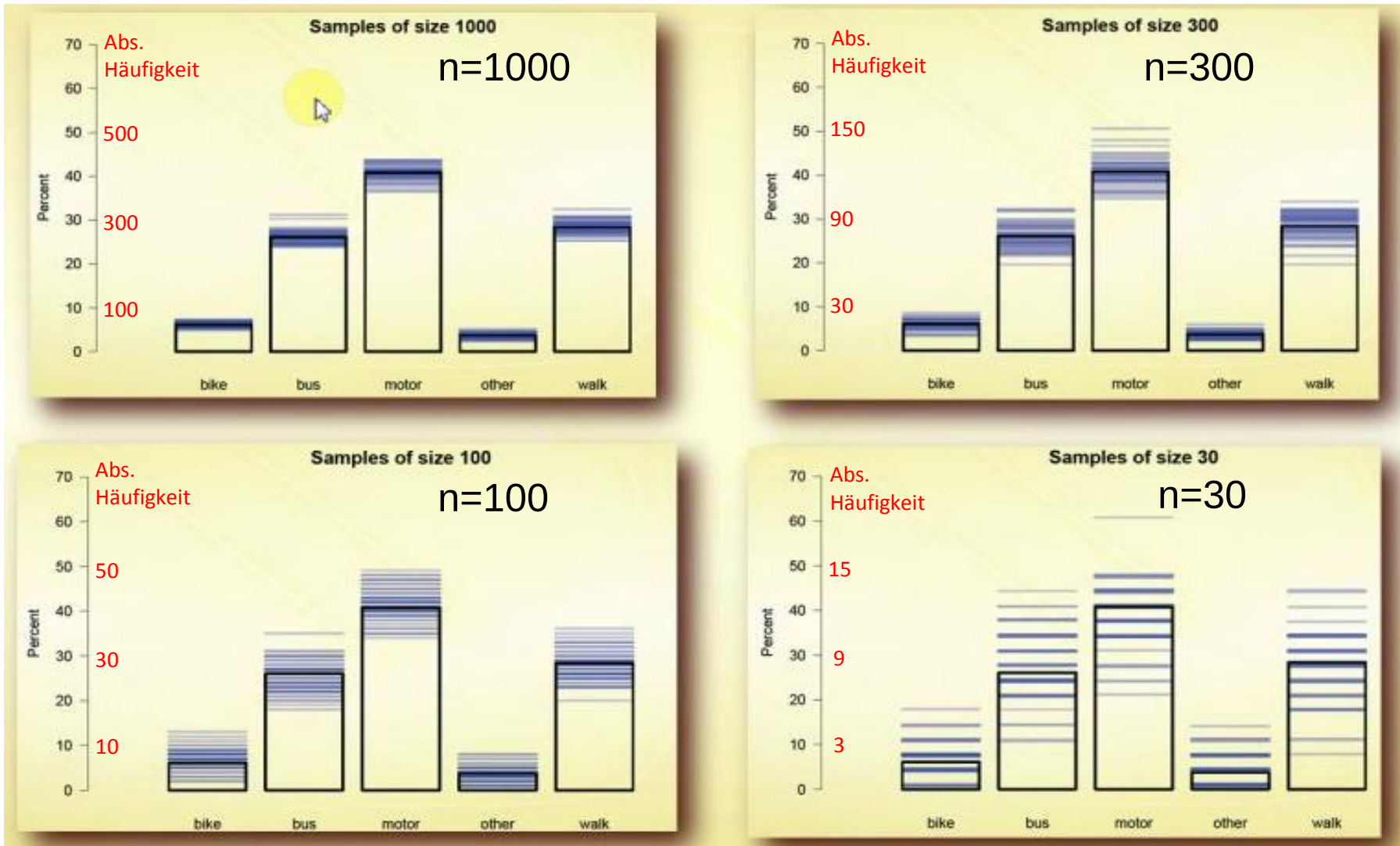
Wie aussagekräftig ist eine Stichprobe mit $n=300$?

Wie kommen neuseeländische Schüler in die Schule?



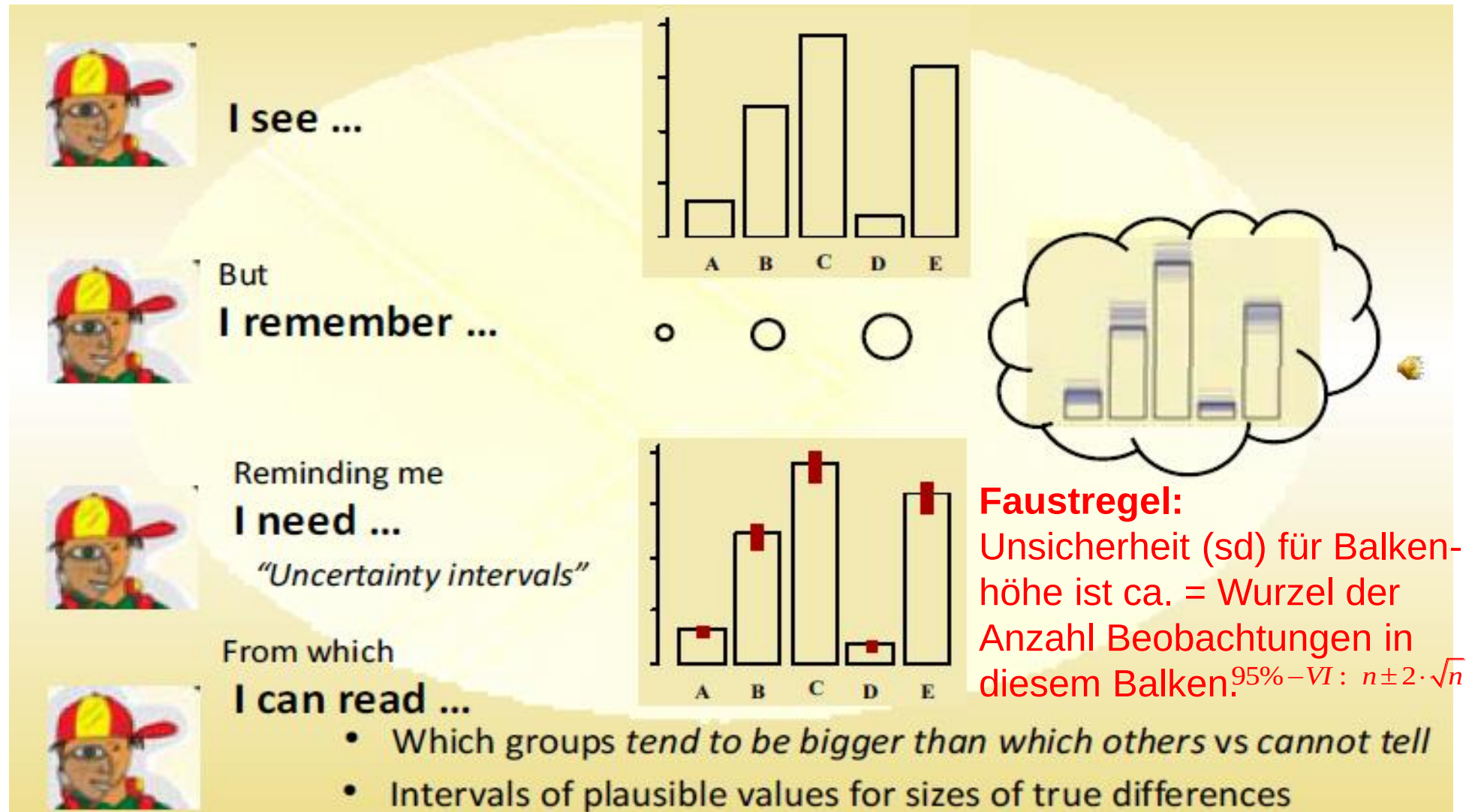
Wie aussagekräftig ist eine Stichprobe mit $n=300$?

Wie kommen neuseeländische Schüler in die Schule?

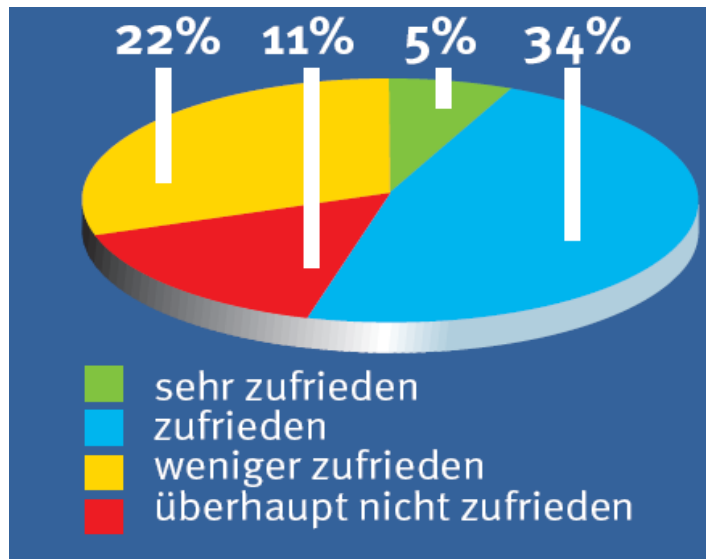


Wie aussagekräftig ist eine Stichprobe mit $n=300$?

Wie kommen neuseeländische Schüler in die Schule?

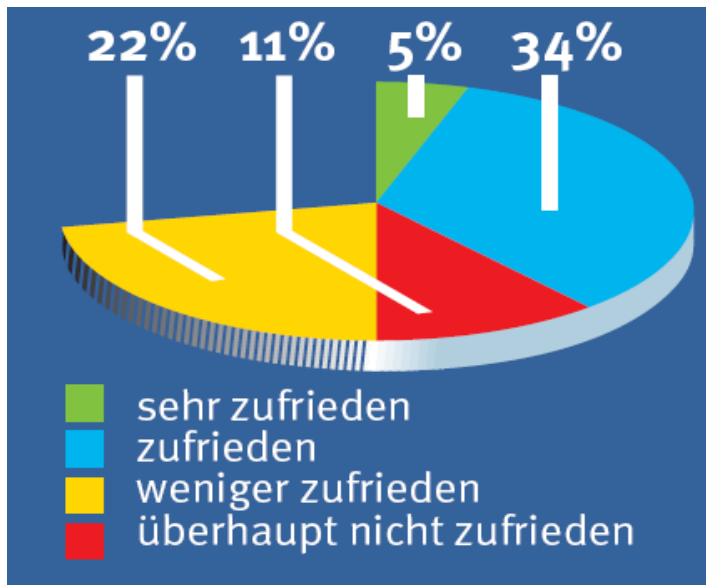


Ungefähr die Hälfte ist zufrieden - wirklich?



Grafik aus der „Bild“. Es wurde gefragt, ob die Leistung von Bundestrainer Klinsmann zufriedenstellend ist.

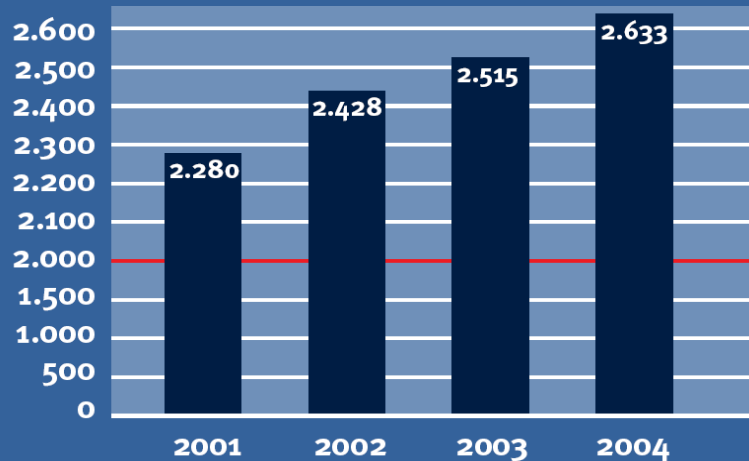
Manipulation: Fläche ist nicht proportional zum Anteil



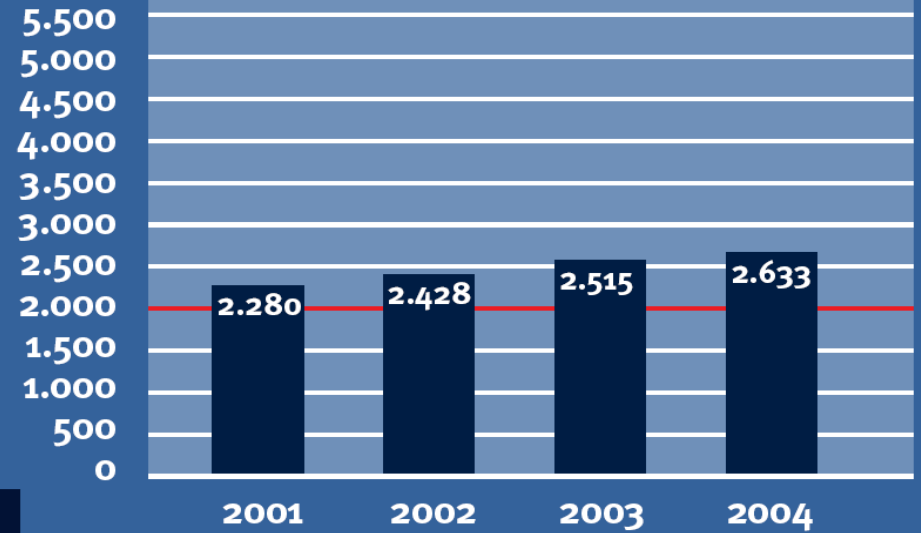
Grund: Prozentzahlen addieren sich nicht zu 100%

Guter Geschäftsverlauf - wirklich?

ENTWICKLUNG BILANZSUMME (in Mio. Euro)



G1



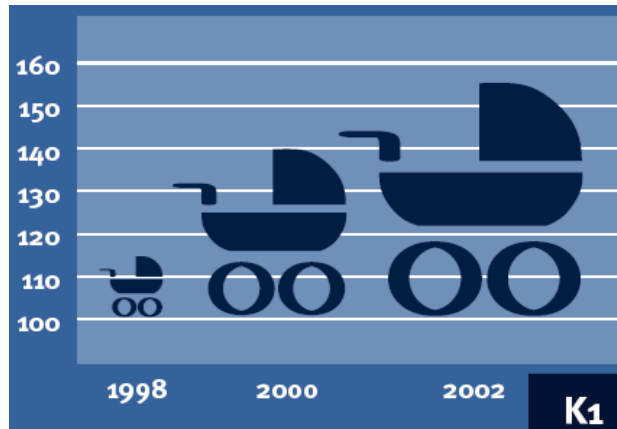
G2

Darstellung der psd-Bank Rhein-Ruhr in ihrem Geschäftsbericht 2004.

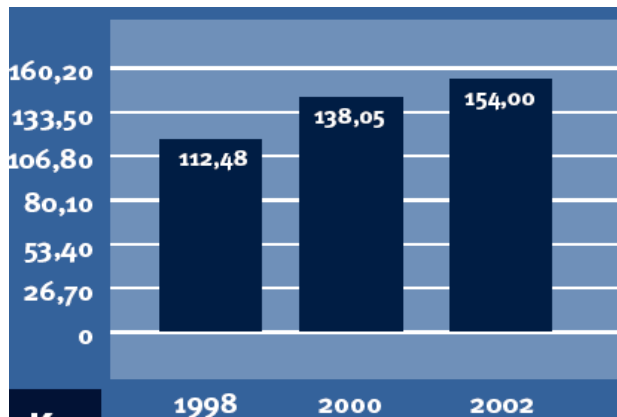
Manipulation: Balkenhöhe ist nicht proportional zur Bilanzsumme

Grund: Skala über 2000 wurde geändert.

Das Kindergeld ist gewaltig gestiegen - wirklich?

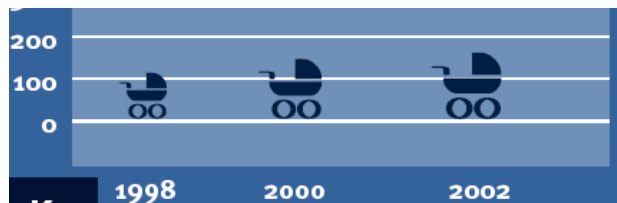


Grafik aus einer Regierungserklärung der deutschen rot-grünen Regierung in der agenda 2010.



Manipulation: Fläche ist nicht proportional zum Kindergeld

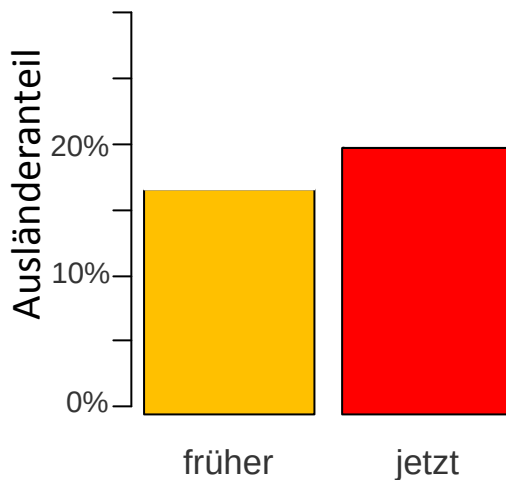
Grund: Skala beginnt nicht bei 0 und Fläche steigt stärker als proportional mit der Höhe des Kinderwagens



Der Ausländeranteil hat sich ver-10-facht- wirklich?



Internetauftritt der SVP



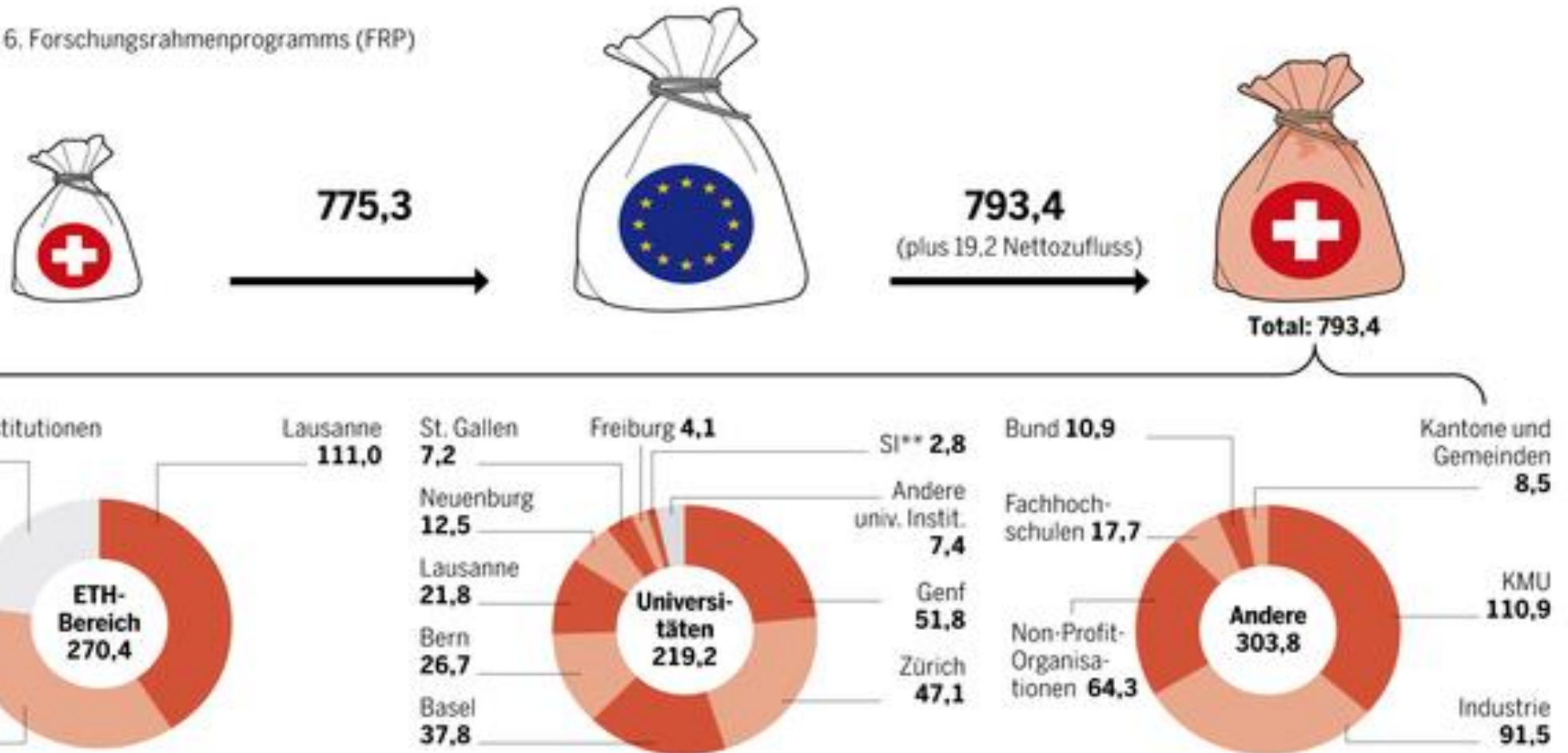
Manipulation: Balkenhöhe ist nicht proportional zum Ausländeranteil

Grund: Skala beginnt nicht bei 0 und Höhe steigt stärker als proportional mit der Höhe des Ausländeranteils

Wie stark profitiert die Schweizer Forschung von der EU?

Wie die Schweizer Forschung von der EU profitiert

Finanzfluss des 6. Forschungsrahmenprogramms (FRP)
in Mio. Fr.



Quantitative Daten

Armspanne der Klassenerhebung (in cm)

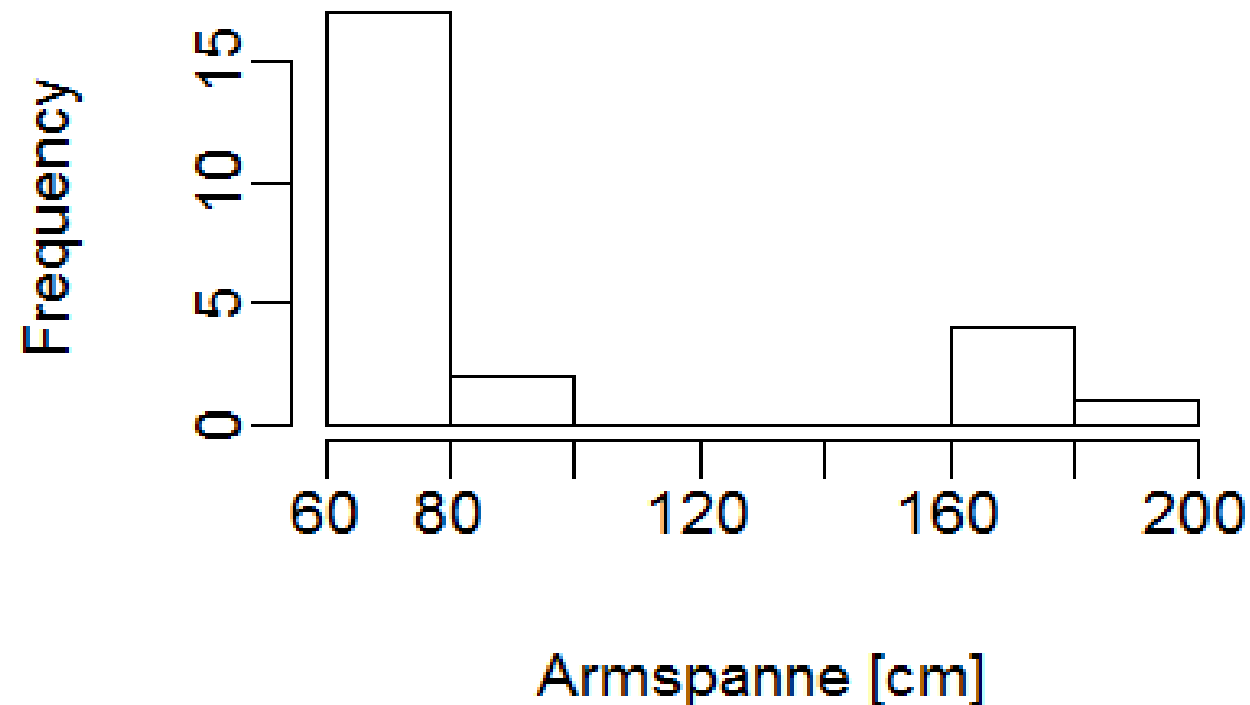
$x_1=82$ $x_2=72$ $x_3=77$ $x_4=77$ $x_5=60$ $x_6=80$ $x_7=65$ $x_8=84$ $x_9=180$
 $x_{10}=70$ $x_{11}=75$ $x_{12}=71$ $x_{13}=74$ $x_{14}=180$ $x_{15}=176$ $x_{16}=60$ $x_{17}=170$
 $x_{18}=73$ $x_{19}=72$ $x_{20}=75$ $x_{21}=70$ $x_{22}=77$ $x_{23}=183$ $x_{24}=72$

Wie können wir diese Daten übersichtlich darstellen?

Histogramm

Armspanne [cm]	Häufigkeit
[60, 80)	17
[80,100)	2
[100,120)	0
[120,140)	0
[140,160)	0
[160,180)	4
[180,200)	1

- Beobachtungen in (gleich breite) Klassen aufteilen
- Säule proportional zur absoluten Häufigkeit
- keine Lücken zwischen Säulen



Grössere Datenmengen

Messungen beim Old-Faithful-Geisirs im
Yellowstone-Nationalpark vom 1. August
1985 - 15. August 1985.

Wartezeiten zwischen zwei Eruptionen

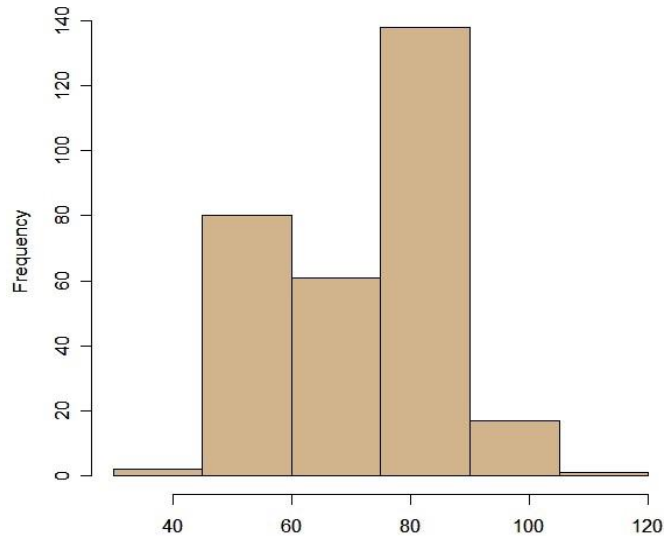
Urliste [in min]: 43-108



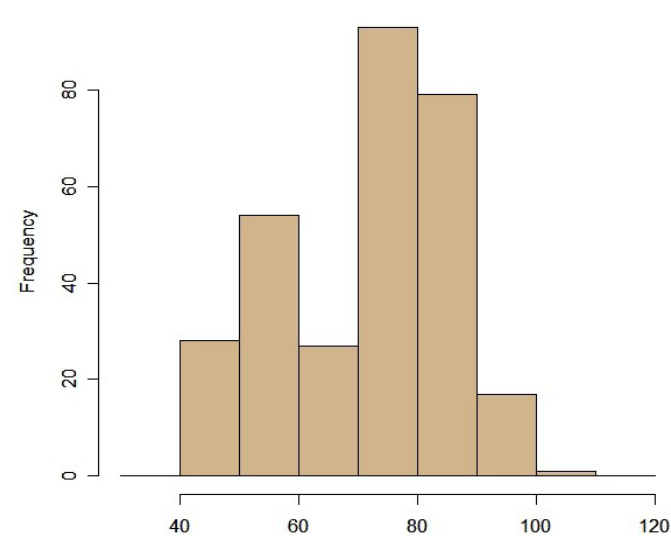
80	71	57	80	75	77	60	86	77	56	81	50	89	54	90	73	60	83	65	82	84	54	85	58	79	57	88	68	76
78	74	85	75	65	76	58	91	50	87	48	93	54	86	53	78	52	83	60	87	49	80	60	92	43	89	60	84	69
74	71	108	50	77	57	80	61	82	48	81	73	62	79	54	80	73	81	62	81	71	79	81	74	59	81	66	87	53
80	50	87	51	82	58	81	49	92	50	88	62	93	56	89	51	79	58	82	52	88	52	78	69	75	77	53	80	55
87	53	85	61	93	54	76	80	81	59	86	78	71	77	76	94	75	50	83	82	72	77	75	65	79	72	78	77	79
75	78	64	80	49	88	54	85	51	96	50	80	78	81	72	75	78	87	69	55	83	49	82	57	84	57	84	73	78
57	79	57	90	62	87	78	52	98	48	78	79	65	84	50	83	60	80	50	88	50	84	74	76	65	89	49	88	51
78	85	65	75	77	69	92	68	87	61	81	55	93	53	84	70	73	93	50	87	77	74	72	82	74	80	49	91	53
86	49	79	89	87	76	59	80	89	45	93	72	71	54	79	74	65	78	57	87	72	84	47	84	57	87	68	86	75
73	53	82	93	77	54	96	48	89	63	84	76	62	83	50	85	78	78	81	78	76	74	81	66	84	48	93	47	87
51	78	54	87	52	85	58	88	79																				

Wie viele Klassen sind hier sinnvoll?

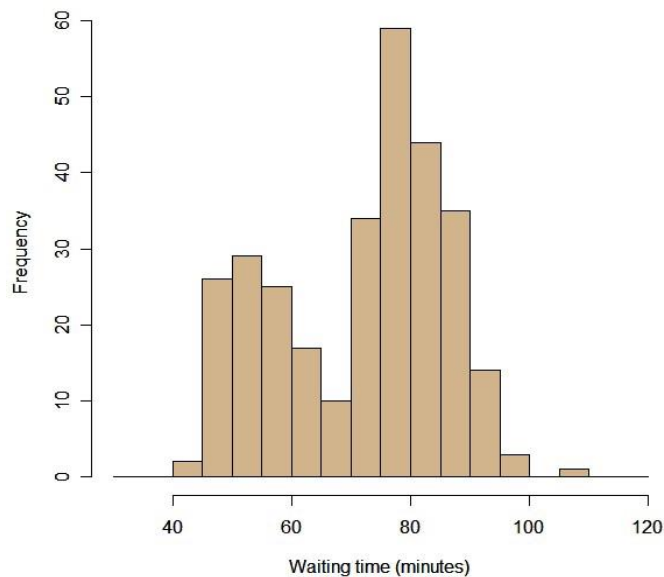
Times between Old Faithful eruptions



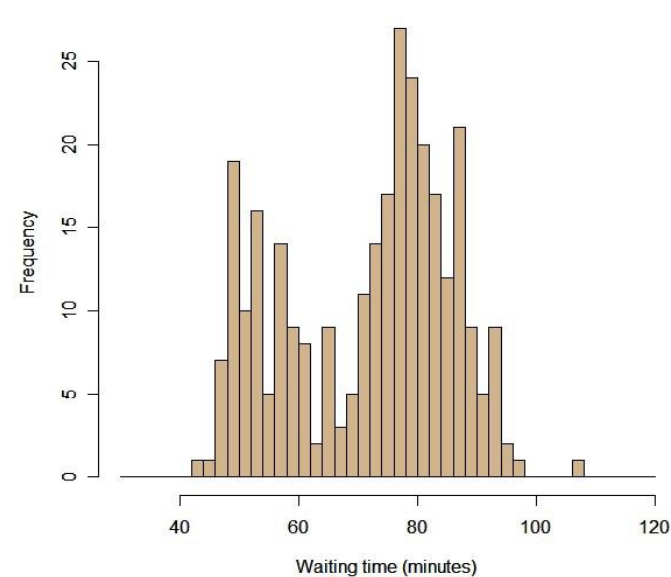
Times between Old Faithful eruptions



Times between Old Faithful eruptions



Times between Old Faithful eruptions



<http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html>



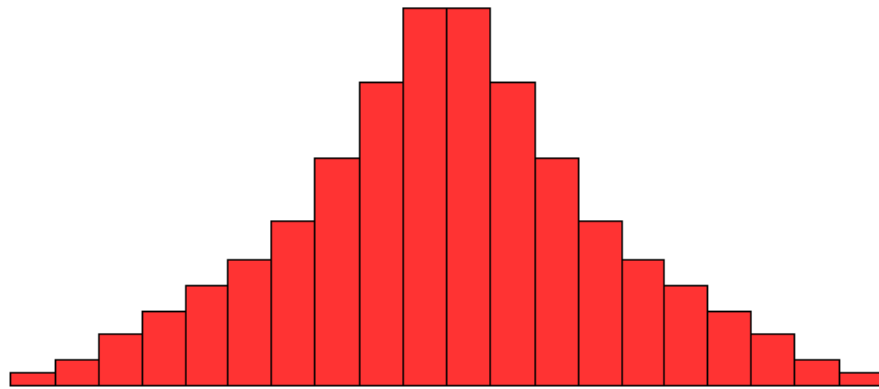
299 Eruptionsintervalle wurden beobachtet.

Achtung: Histogramms hängt von der Wahl der Klassen ab!

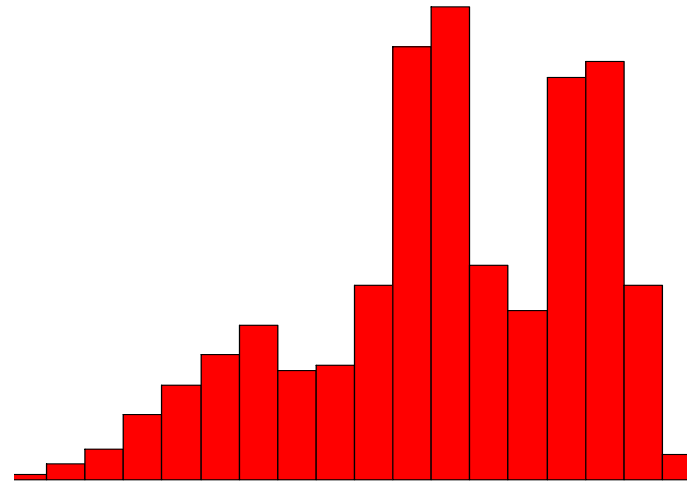
Merkmale zum Histogramm

- Das Histogramm ist eine der bekanntesten grafischen Darstellungen für quantitative meist stetige Daten.
- Für die Erstellung eines Histogramms **müssen stetige Daten in Klassen eingeteilt werden**. Die Klassen müssen nicht notwendig gleich breit sein, allerdings sind Histogramme mit gleichbreiten Klassen schwieriger zu interpretieren.
- Beim Histogramm der **absoluten Häufigkeiten** ist die **Balkenhöhe** gleich der Anzahl Beobachtungen der entsprechenden Klasse.
- Beim **skalierten Histogramm** ist die **Balkenfläche** gleich dem Anteil der Beobachtungen der entsprechenden Klasse an allen Beobachtungen.
- Ungleiche Klassenbreiten wenn möglich vermeiden!
- Wenn n Beobachtungen vorliegen, so sollte gilt es als Faustregel, die Daten in ca. $\sqrt{n}$ Klassen einzuteilen.

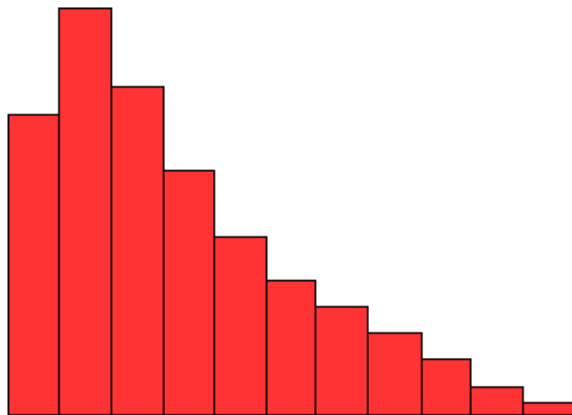
Form einer Verteilung



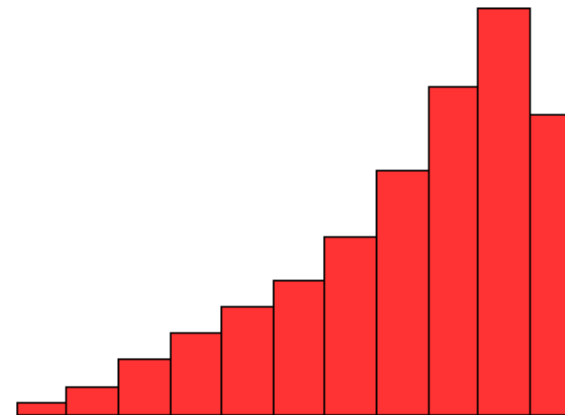
symmetrische Verteilung
uni-modale



multimodale Verteilung
leicht linksschief



rechtsschiefe Verteilung
uni-modale



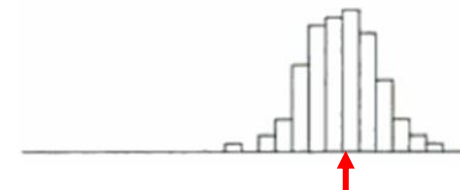
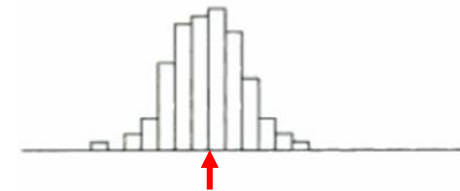
linksschiefe Verteilung
uni-modale

Daten, d.h. ein Zahlenhaufen, können nicht nur grafisch, sondern auch durch Kennzahlen (Masszahlen) charakterisiert werden. Wichtig sind vor allem jene Grössen, die **Lage** oder **Streuung** der Daten beschreiben.

Lagemasse

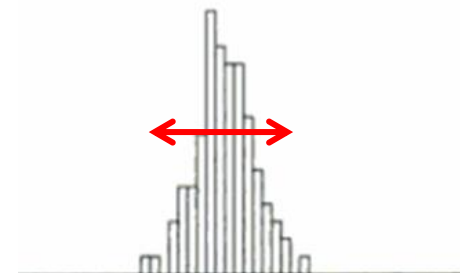
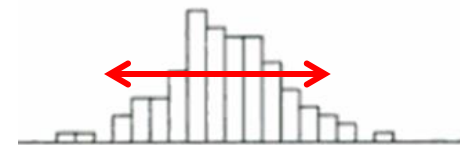
Eine Zahl, die den „mittleren“ Wert bzw. den Schwerpunkt einer Verteilung charakterisiert, d.h. die Lage der Verteilung.

(~> begrenzt sinnvoll wenn nicht unimodal)



Streumasse

Eine Zahl, die angibt, wie „breit“ die Verteilung ist, d.h. wie stark die Werte um den „mittleren“ Wert streuen.



Ist der mittlere Lohn ein «typischer Lohn»?

Bei Novartis lag der mittlere Lohn in der Schweiz im Jahr 2009 bei ca. 220'000



Schweizer Arbeitsplätze



 [Grafik vergrössern](#)

Novartis beschäftigt weltweit zurzeit rund 99.800 Mitarbeitende. Davon arbeiten rund 12.000 in der Schweiz – verteilt auf die acht Standorte in Basel BS/BL, Stein AG, Embrach ZH, Cham ZG, Bern BE, St-Aubin FR, Nyon VD und Locarno TI. Eine kürzlich veröffentlichte Studie hat ergeben, dass für jeden direkten Arbeitsplatz bei Novartis in der Schweiz indirekt 2,5 weitere Arbeitsplätze geschaffen werden.

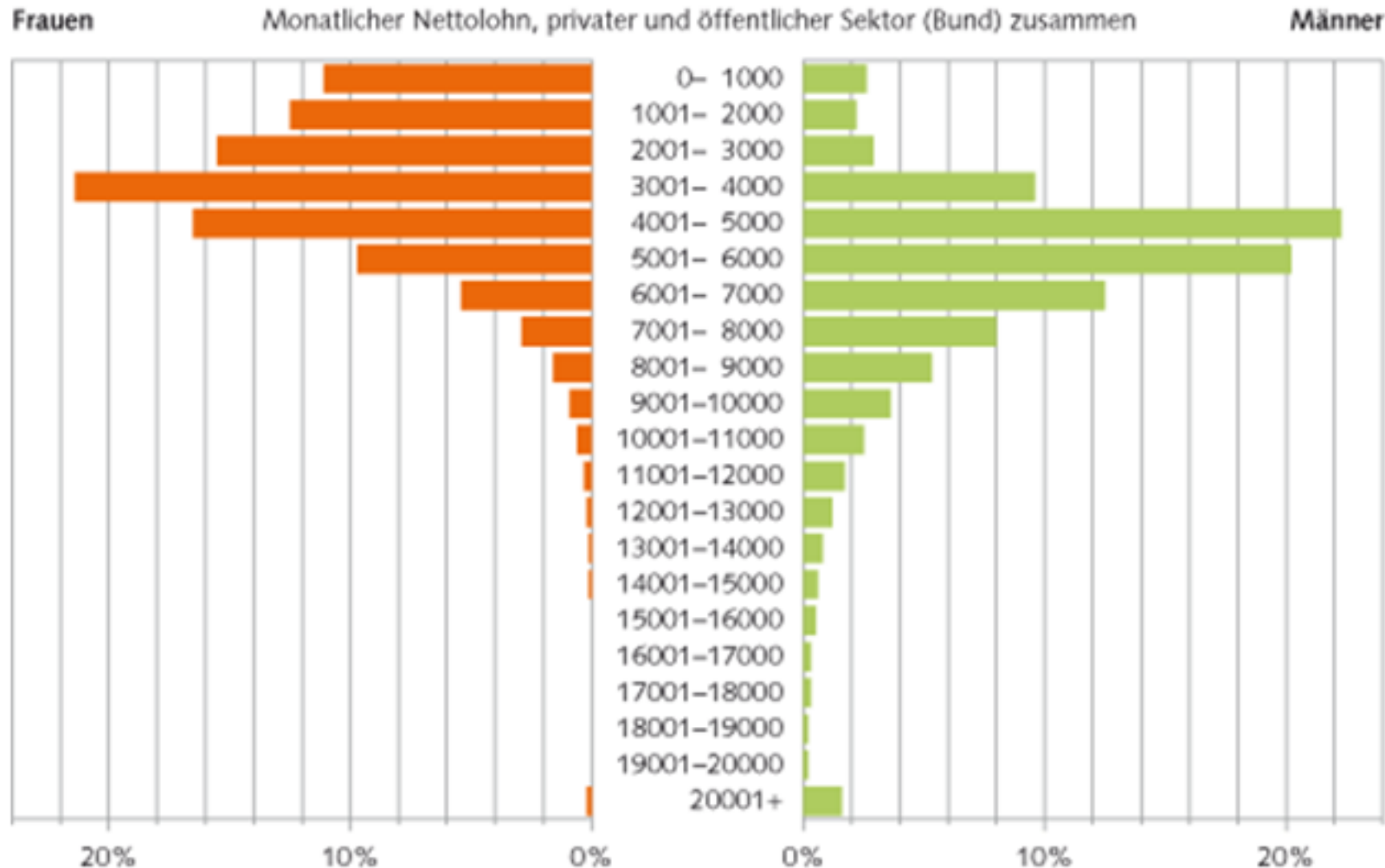
$$\text{mittl.Lohn} \approx \frac{2.6\text{Mrd.}}{12000} \\ \approx 220000\text{CHF}$$

Die Gesamtsumme der Lohn- und Sozialleistungen für Mitarbeitende von Novartis in der Schweiz betrug im Jahr 2009 rund 2,6 Milliarden Franken.

Visualisierung der Lohnverteilungen ist informativer

Lohnverteilungen sind typischerweise rechtsschief

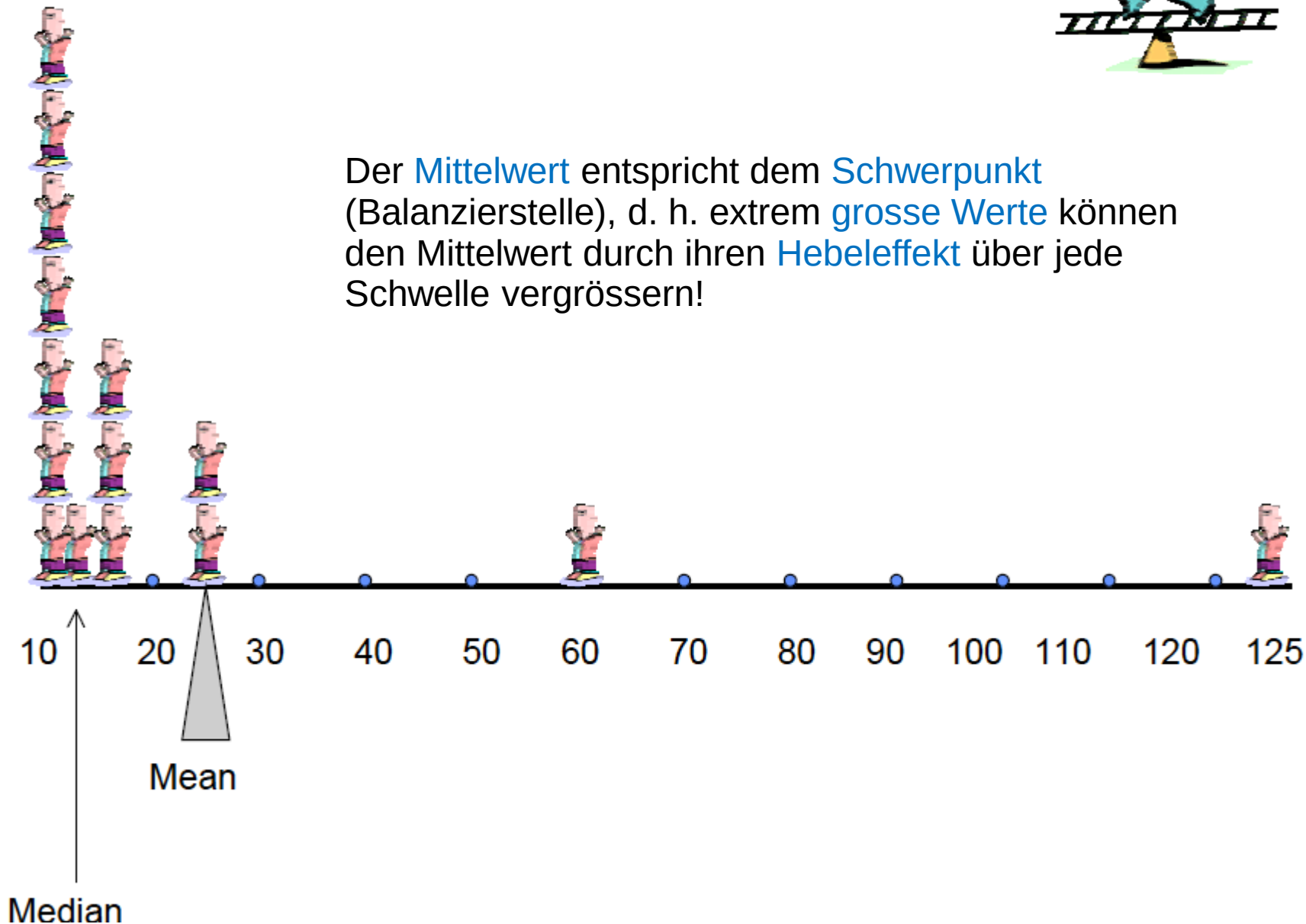
Häufigkeitsverteilung der Arbeitnehmenden nach Lohnhöhenklassen 2008



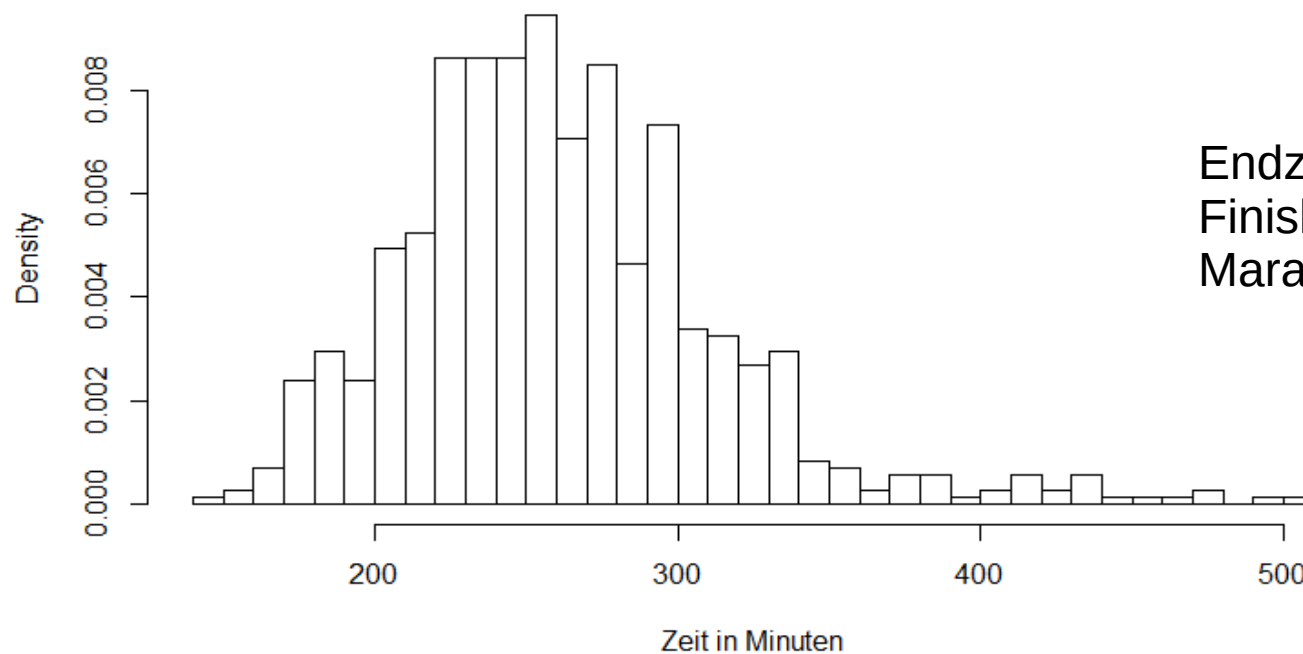
Bei rechtsschiefen Verteilungen ist der Mittelwert nicht „typisch“



Der **Mittelwert** entspricht dem **Schwerpunkt** (Balanzierstelle), d. h. extrem **grosse Werte** können den Mittelwert durch ihren **Hebeleffekt** über jede Schwelle vergrössern!



Motivation Quantile



Endzeiten von 1000 männlichen
Finishern der New York
Marathons 2002

Frage:

- Wie schnell muss ich laufen, so dass ich genau im Mittelfeld liege?

Median = 256.02 Minuten

- Wie schnell muss ich laufen, so dass ich zu den 10% schnellsten gehöre?

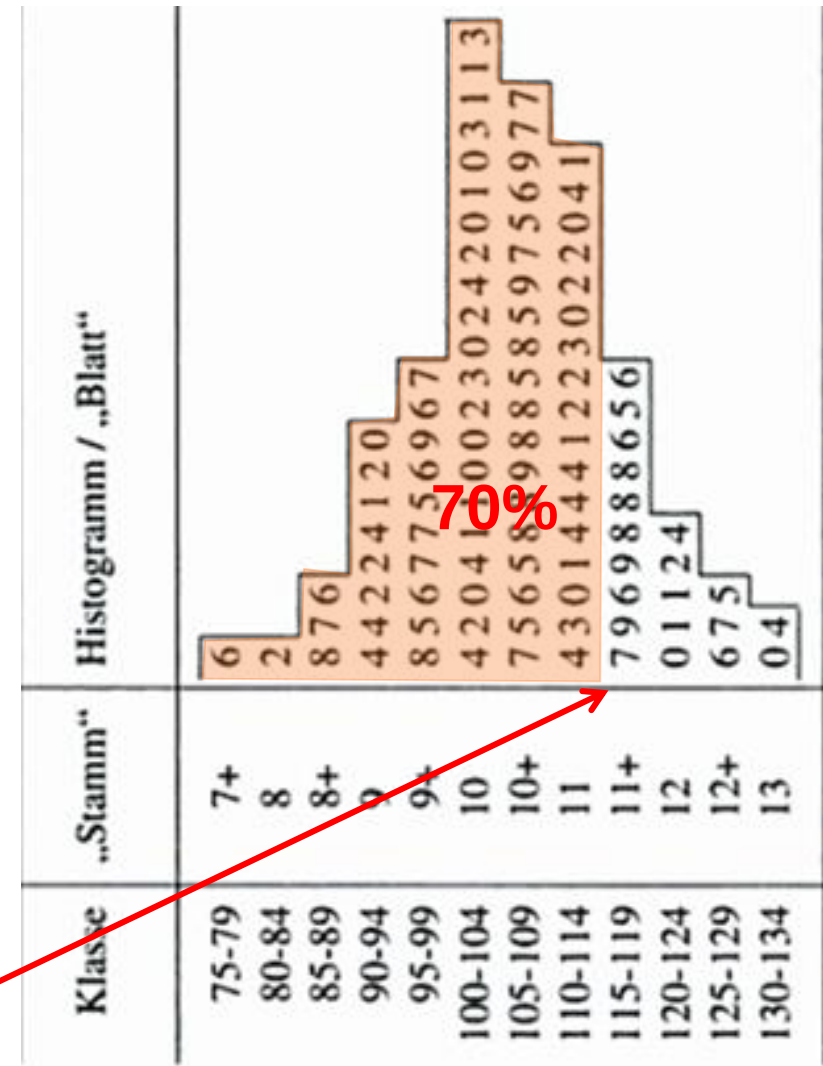
10%-Quantil=201.96

Flächen im skalierten Histogramm und Quantile

Das α -Quantil q_α ($0\% \leq \alpha \leq 100\%$) einer Stichprobe ist ein Wert im Wertebereich der Stichprobe, für den **gilt dass $\alpha \cdot 100\%$ aller Werte kleiner-gleich diesem Wert sind.**

- Die Fläche im Histogramm der relativen Häufigkeiten links von q_α ist gleich α .
- Zwischen dem 25%-Quantil $q_{25\%}$ und dem 75%-Quantil $q_{75\%}$ liegen die mittleren 50% aller Stichprobenwerte.

$$q_{70\%} = 117$$



Bemerkung: das Histogramm der abs. Häufigkeiten und das Histogramm der Dichte (rel. Häufigkeiten), haben bei gleicher Klassenbreite die gleiche Form => Flächenverlauf hat auch gleiche Form.

Berechnung von Quantilen aus geordneter Stichprobe

Das α -Quantil q_α (mit $0\% \leq \alpha \leq 100\%$) einer Stichprobe ist der Wert, für den $\alpha\%$ aller Werte kleiner-gleich diesem Wert sind.

Zuerst wird die Stichprobe geordnet - $x_{[k]}$ ist die Beobachtung die an der Ordnungsposition k steht, d.h. $k-1$ Beobachtungen sind kleinergleich als diese Beobachtung.

$$q_\alpha = \begin{cases} x_{[k]} & \text{falls } n \cdot \alpha \notin \mathbb{Z} , \quad k = \text{aufgerunden}(n \cdot \alpha) \\ \frac{1}{2} (x_{[n\alpha]} + x_{[n\alpha+1]}) & \text{falls } n \cdot \alpha \in \mathbb{Z} \end{cases}$$

Der Median ist das 50%-Quantil und ein robustes Lagemass

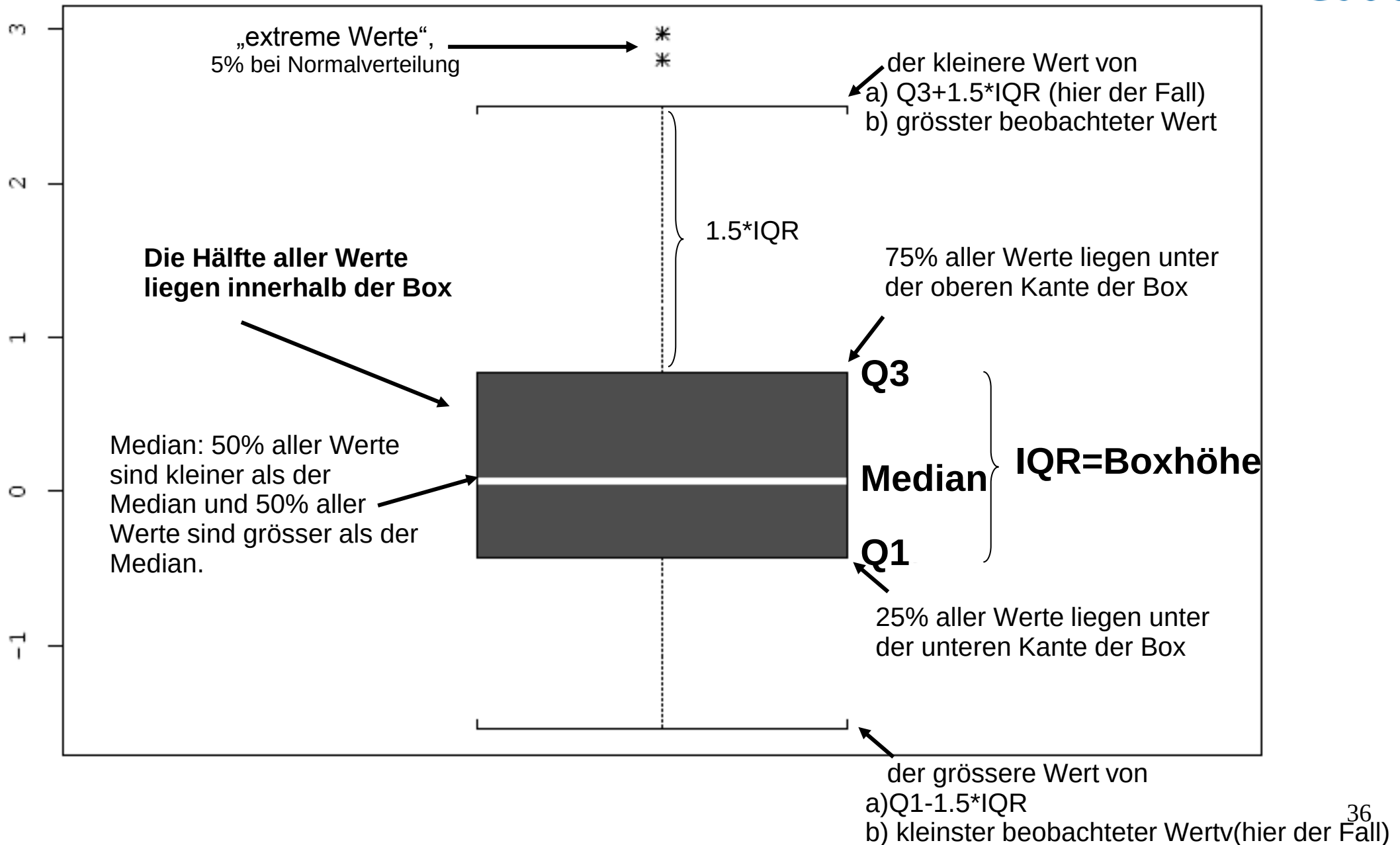
Das 1. Quartil Q_1 ist das 25%-Quantil

Der Interquartilsrange $IQR = Q_3 - Q_1 = q_{75\%} - q_{25\%}$ ist ein robustes Streumass

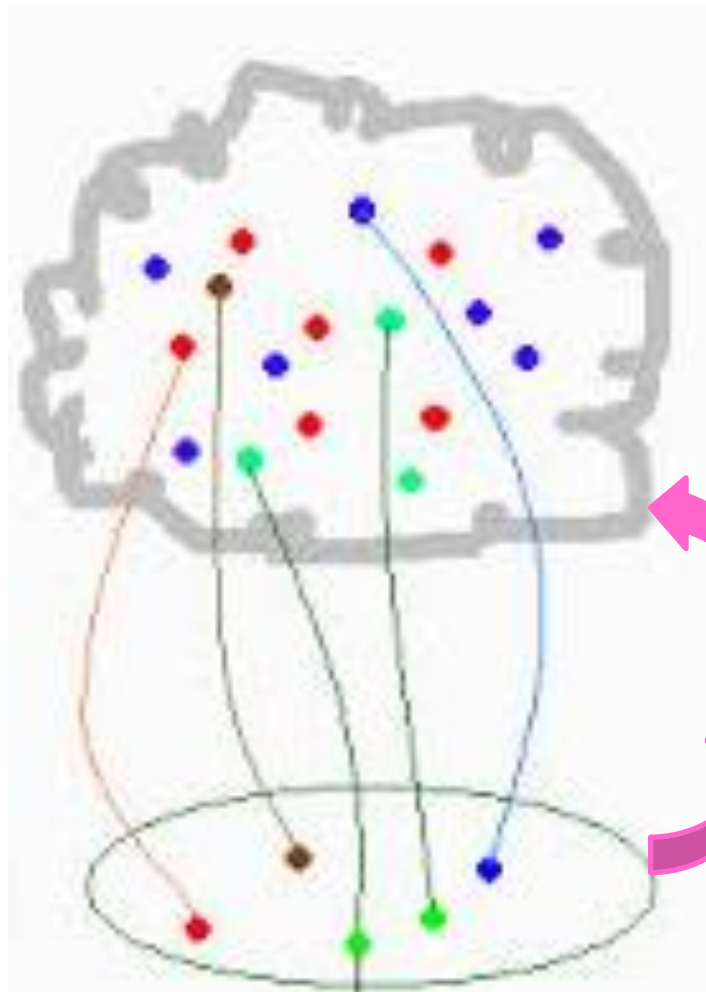
Das Kleingedruckte

Wenn $\alpha \cdot n$ keine ganze Zahl ergibt, so ist die nächste ganze Zahl gleich dem Rang des Stichprobenwerts, der dem α -Quantil entspricht. Falls $\alpha \cdot n$ eine ganze Zahl ergibt, so ist müssen die Stichprobenwerte mit dem Rang $\alpha \cdot n$ und dem Rang $\alpha \cdot n + 1$ gemittelt werden, um den Wert des α -Quantils zu erhalten.

Definition des Boxplots zur Visualisierung von quantitativen Daten



Warum sieht man 2 verschiedene Formeln für die Varianz?



Population

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

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

**Lernen,
Schliessen**

Stichprobe

$$S_{sample}^2 = \hat{\sigma}_{pop}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{sample} = \hat{\sigma}_{pop} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

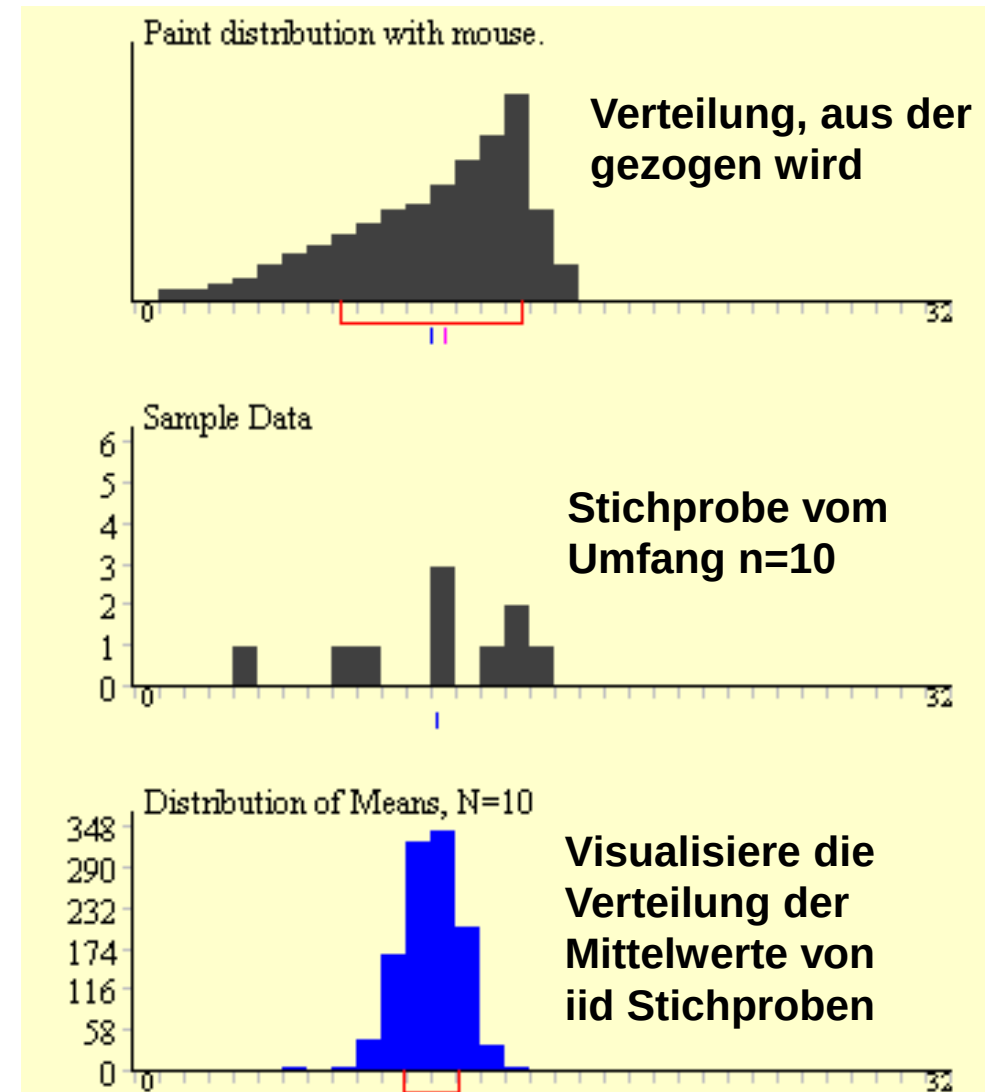
Um von einer Stichprobe eine unverzerzte (d.h. im Mittel korrekte) Schätzung der Populationsvarianz zu erhalten, muss die Formel der Stichprobenvarianz benutzt werden, die im Nenner (n-1) hat. Benutzt man die Formel mit n im Nenner, so schätzt man im Mittel eine zu kleine Varianz.

Ergebnisse einer Monte-Carlo-Studie führen auf den Zentralen Grenzwertsatz

- Die Mittelwerte streuen um den Erwartungswert der Verteilung, aus der die Beobachtungen gezogen werden

$$\bar{X} \xrightarrow{n \rightarrow \infty} E(X) = \mu_x$$

- Je grösser die Stichprobe, um so kleiner ist die Streuung der Mittelwerte.
- Egal wie die Verteilung der Einzelwerte aussieht, die Mittelwerte sind glockenförmig um den wahren Populationsmittelwert μ_x verteilt, wenn die Stichproben gross genug sind.



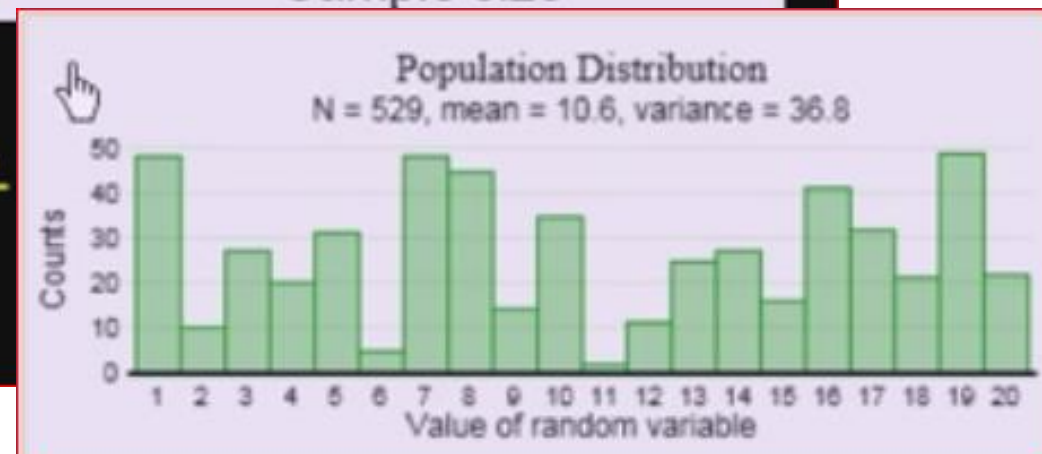
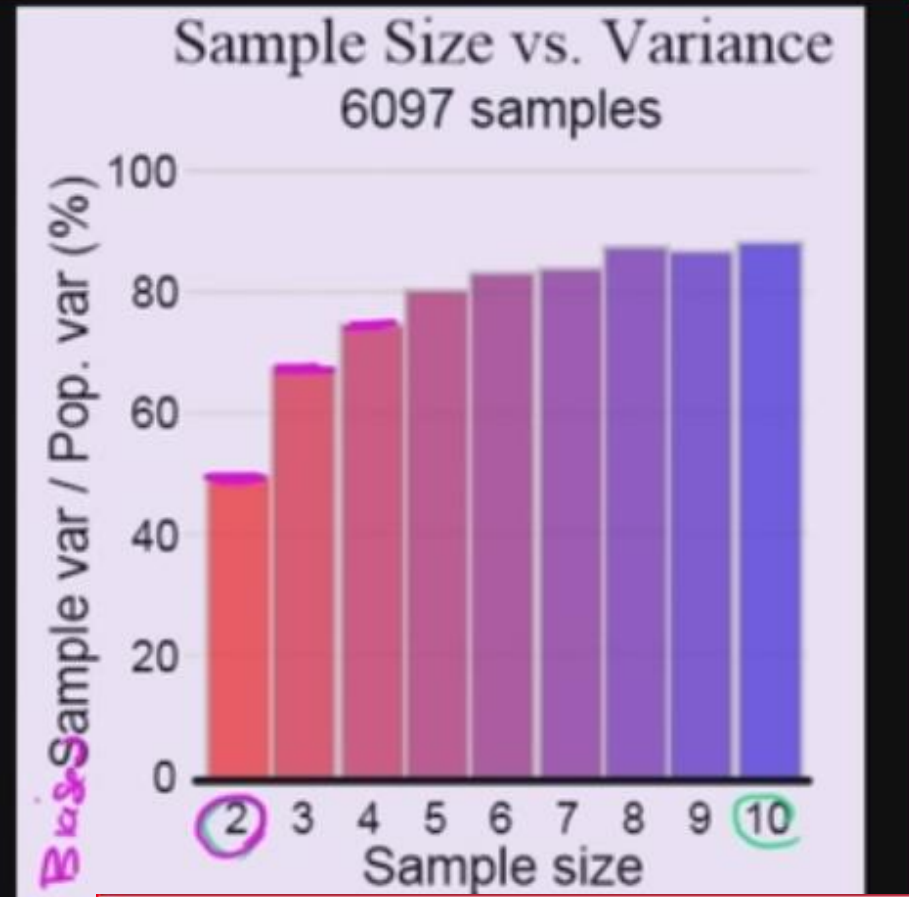
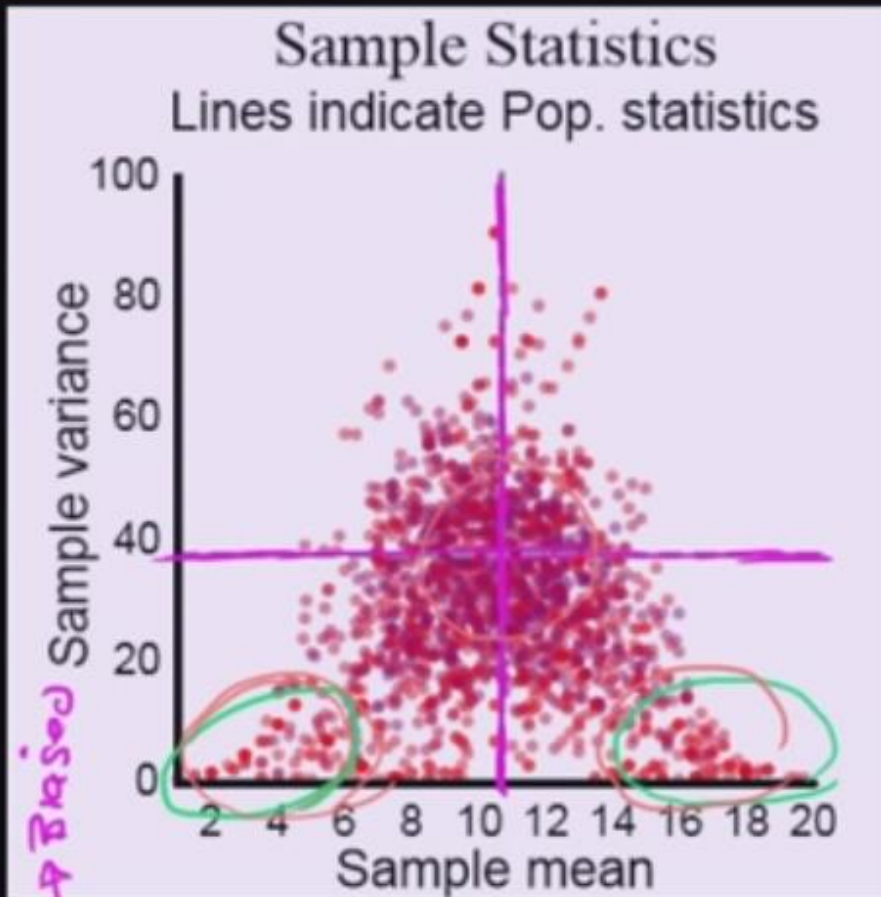
Der zentrale Grenzwertsatz

Der Mittelwert $T(X) = \bar{X}$ einer Stichprobe ist ein **erwartungstreuer Schätzer für den Populationsmittelwert**, weil:

Der Mittelwert $T(X) = \bar{X}$ einer Zufallsstichprobe $X_1, X_2, \dots, X_n$ vom Umfang n , mit X_i **i.i.d. nach einer beliebigen Verteilung V** also $X_i \sim V$ mit Erwartungswert $E(X) = \mu_x$ und Varianz $\text{Var}(X) = \sigma_x^2$, ist i.a. für (sehr) grosse Stichprobenumfänge näherungsweise normalverteilt, wobei gilt (**ZGS**):

$$\bar{X} \overset{a}{\sim} N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$$

Warum wird bei Stichproben-Varianz-Formel nicht durch n geteilt?



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

Vorschlag für einen Varianzschätzer und Vorbereitung für die Berechnung seines EW

Ein «natürlicher» Schätzer für die Varianz der Population wäre die mittlere quadratische Abweichung der Stichprobenwerte vom Stichprobenmittelwert:

$$\begin{aligned}\tilde{S}^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \\&= \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\&= \frac{1}{n} \left(\sum_{i=1}^n X_i^2 \right) - \frac{1}{n} \left(2\bar{X} \sum_{i=1}^n X_i \right) + \frac{1}{n} (n\bar{X}^2) \\&= E(X^2) - 2\bar{X}^2 + \bar{X}^2 \\&= E(X^2) - \bar{X}^2\end{aligned}$$

Berechnung des Erwartungswerts von $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

Benutze: $E(\tilde{S}^2) = E(X^2) - \bar{X}^2$ und $E(X^2) = \text{Var}(X) + E(X)^2$

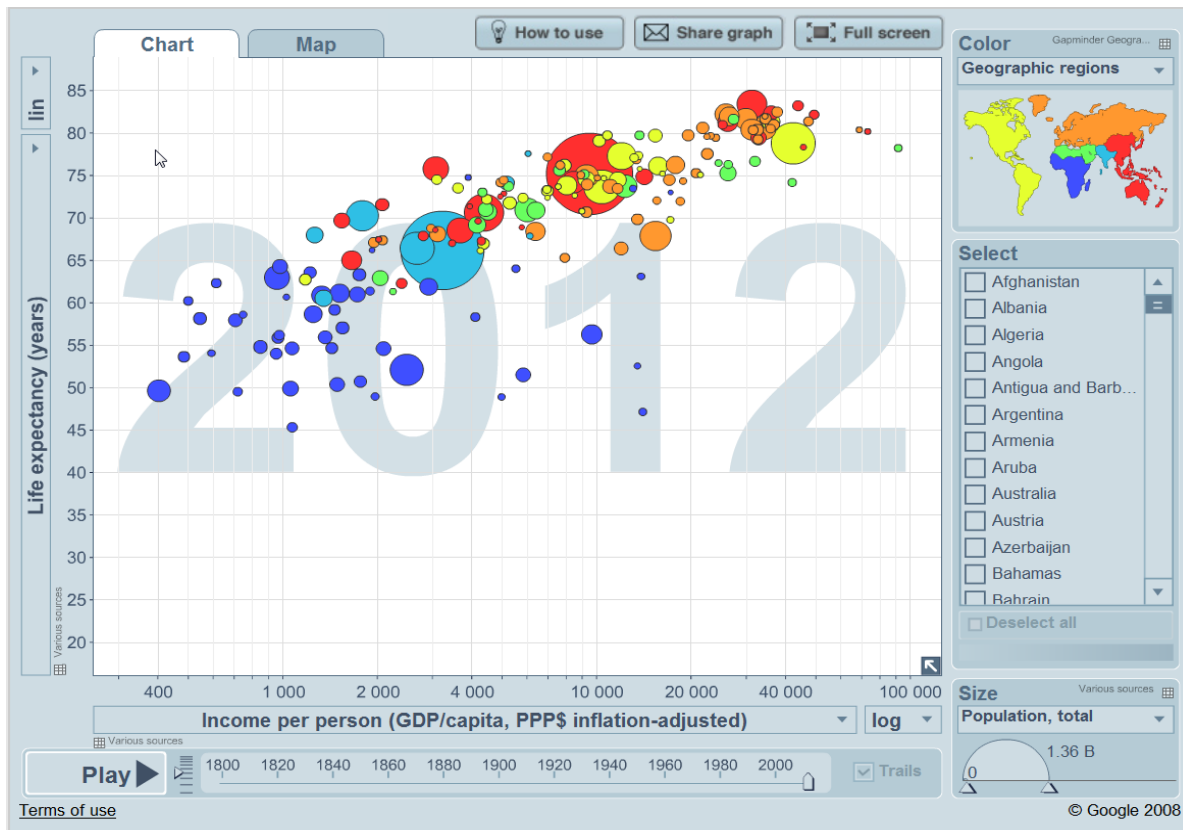
$$\begin{aligned}
 E(\tilde{S}^2) &= E(E(X^2) - \bar{X}^2) \\
 &= E(X^2) - E(\bar{X}^2) \\
 &= \text{Var}(X) + E(X)^2 - (\text{Var}(\bar{X}) + E(\bar{X})^2) \\
 &= \text{Var}(X) + E(X)^2 - \left(\frac{\text{Var}(X)}{n} + E(X)^2 \right) \\
 &= \frac{n \cdot \text{Var}(X)}{n} - \frac{\text{Var}(X)}{n} \\
 &= \frac{n \cdot \text{Var}(X) - \text{Var}(X)}{n} \\
 &= \frac{n-1}{n} \cdot \text{Var}(X) \\
 &\neq \text{Var}(X)
 \end{aligned}$$

Der Schätzer $\tilde{S}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
für die Varianz ist nicht erwartungstreu.

$$\begin{aligned}
 E(S^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\
 &= E\left(\frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\
 &= E\left(\frac{n}{n-1} \cdot \tilde{S}^2 \right) = \frac{n}{n-1} E(\tilde{S}^2) \\
 &= \frac{n}{n-1} \cdot \frac{n-1}{n} \cdot \text{Var}(X) \\
 &= \text{Var}(x)
 \end{aligned}$$

Der Schätzer $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
für die Varianz ist **erwartungstreu!**

Die Entwicklung der Weltbevölkerung...



[Gapminder zum selber benutzen:](http://www.gapminder.org/world/)
<http://www.gapminder.org/world/>

Interessanter TED Vortrag:



Hans Rosling

Debunking myths about the “third world”

http://www.ted.com/talks/hans_rosling_religions_and_babies?language=de