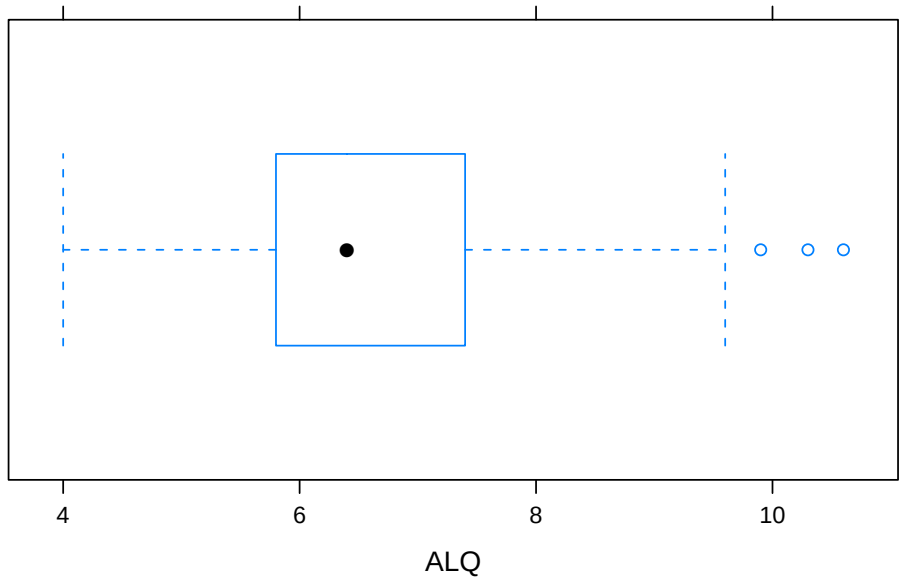
Lage- und Streuungsmaße

Statistik I

Sommersemester 2009



Tutorien

- Begleitend zur Vorlesung, inhaltlich identisch mit der Übung
- Mögliche Zeiten: Do 10-12, Do 16-18, Fr 8-10, Fr ab 16
- Wenn Sie ernsthaft bereit sind, an einem Tutorium teilnzunehmen (auch zu unattraktiven Zeiten) bitte eine leere Mail mit Betreff „Tutoriumsinteressent“ an `arzheimer@politik.uni-mainz.de`

Zum Nachlesen

- Agresti/Finlay: Kapitel 3.2 + 3.3
- Gehring/Weins: Kapitel 6
- Schumann: Kapitel 5

Zum Nachlesen

- Agresti/Finlay: Kapitel 3.2 + 3.3
- Gehring/Weins: Kapitel 6
- Schumann: Kapitel 5

- In der Vorlesung wird teilweise mehr Stoff präsentiert als in den Büchern
- Alles was in der Vorlesung gesagt wird ist Prüfungsstoff

Was ist eine Verteilung?

- Kategoriale und kontinuierliche Daten haben verschiedene Ausprägungen
- Häufigkeiten der Ausprägungen $\rightarrow$ Verteilung
- Graphische Darstellungen

Wie werden Verteilungen graphisch dargestellt?

- Kategoriale Daten
  - Nominal: Separate Balken, Reihenfolge egal
  - Ordinal: Separate Balken, Reihenfolge wichtig
- Kontinuierliche Daten: Keine Lücken zwischen Balken
  - Histogramm
  - Polygonzug
  - Dichteschätzung
- Univariate (eindimensionale) Darstellung

Was ist eine Verteilung? II

- Kategoriale und kontinuierliche Daten haben verschiedene Ausprägungen
- Häufigkeiten der Ausprägungen $\rightarrow$ Verteilung
- Graphische Darstellungen
- Verteilungen haben
  1. Eine oder mehrere „Gipfel“
  2. Eine „Mitte“ (zentrale Tendenz)
  3. Mehr oder weniger viel Variation um diese Mitte (nicht alle Werte identisch)
  4. Form (breit-/schmalgipflig, symmetrisch/asymmetrisch)

Was ist eine Verteilung?

- Kategoriale und kontinuierliche Daten haben verschiedene Ausprägungen
- Häufigkeiten der Ausprägungen $\rightarrow$ Verteilung
- Graphische Darstellungen
- Verteilungen haben
  1. Eine oder mehrere „Gipfel“
  2. Eine „Mitte“ (zentrale Tendenz)
  3. Mehr oder weniger viel Variation um diese Mitte (nicht alle Werte identisch)
  4. Form (breit-/schmalgipflig, symmetrisch/asymmetrisch)
- Verteilungen $\rightarrow$ graphische *und* numerische Darstellung (Häufigkeitstabelle, gleiche Information)

Was ist eine Verteilung?

- Kategoriale und kontinuierliche Daten haben verschiedene Ausprägungen
- Häufigkeiten der Ausprägungen $\rightarrow$ Verteilung
- Graphische Darstellungen
- Verteilungen haben
  1. Eine oder mehrere „Gipfel“
  2. Eine „Mitte“ (zentrale Tendenz)
  3. Mehr oder weniger viel Variation um diese Mitte (nicht alle Werte identisch)
  4. Form (breit-/schmalgipflig, symmetrisch/asymmetrisch)
- Verteilungen $\rightarrow$ graphische *und* numerische Darstellung (Häufigkeitstabelle, gleiche Information)
- Maßzahlen *verdichten* Information

Beispieldaten für heute

- (European Social Survey)
- ((fiktives) Alter von Kursteilnehmern)
- Französische Regionalwahl 2004
- Fälle: 94 Départements auf dem französischen Festland
- Variablen u. a. Stimmenanteil Front National, Arbeitslosenquote, Anteil Zuwanderer, Zugehörigkeit zu 21 Regionen

Beispieldaten für heute

- (European Social Survey)
- ((fiktives) Alter von Kursteilnehmern)
- Französische Regionalwahl 2004
- Fälle: 94 Départements auf dem französischen Festland
- Variablen u. a. Stimmenanteil Front National, Arbeitslosenquote, Anteil Zuwanderer, Zugehörigkeit zu 21 Regionen

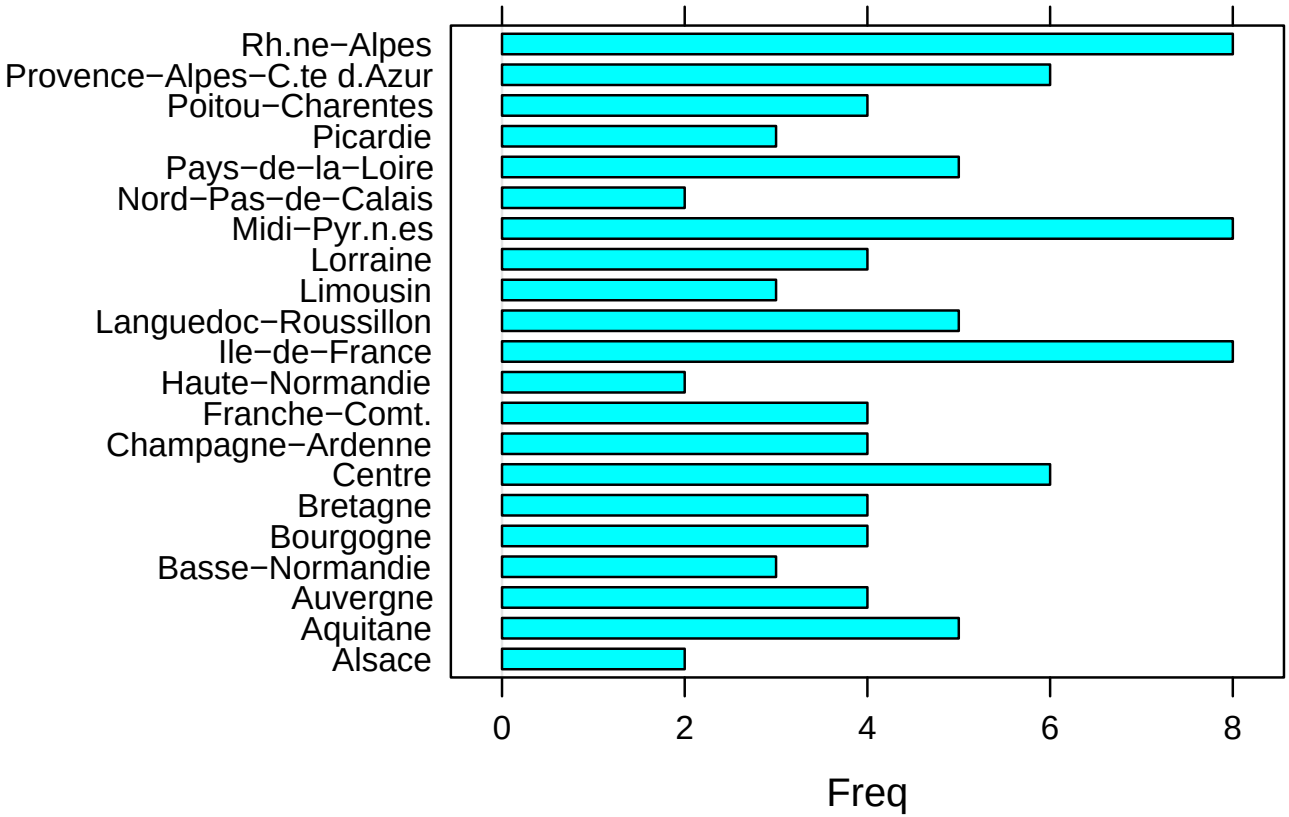
Beispieldaten für heute

- (European Social Survey)
- ((fiktives) Alter von Kursteilnehmern)
- Französische Regionalwahl 2004
- Fälle: 94 Départements auf dem französischen Festland
- Variablen u. a. Stimmenanteil Front National, Arbeitslosenquote, Anteil Zuwanderer, Zugehörigkeit zu 21 Regionen Skalenniveau?

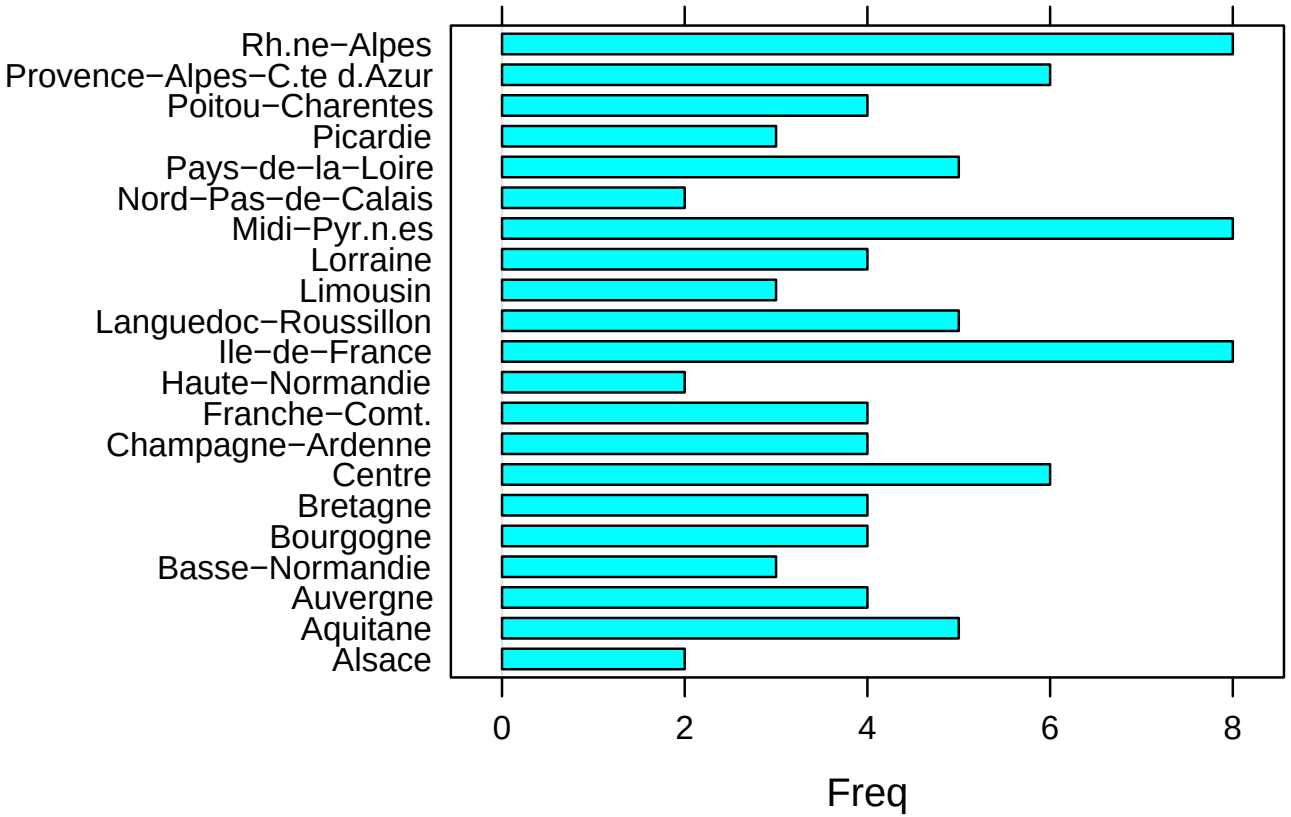
Beispieldaten für heute

- (European Social Survey)
- ((fiktives) Alter von Kursteilnehmern)
- Französische Regionalwahl 2004
- Fälle: 94 Départements auf dem französischen Festland
- Variablen u. a. Stimmenanteil Front National, Arbeitslosenquote, Anteil Zuwanderer, Zugehörigkeit zu 21 Regionen Skalenniveau?

Verteilung der Départements auf die Regionen

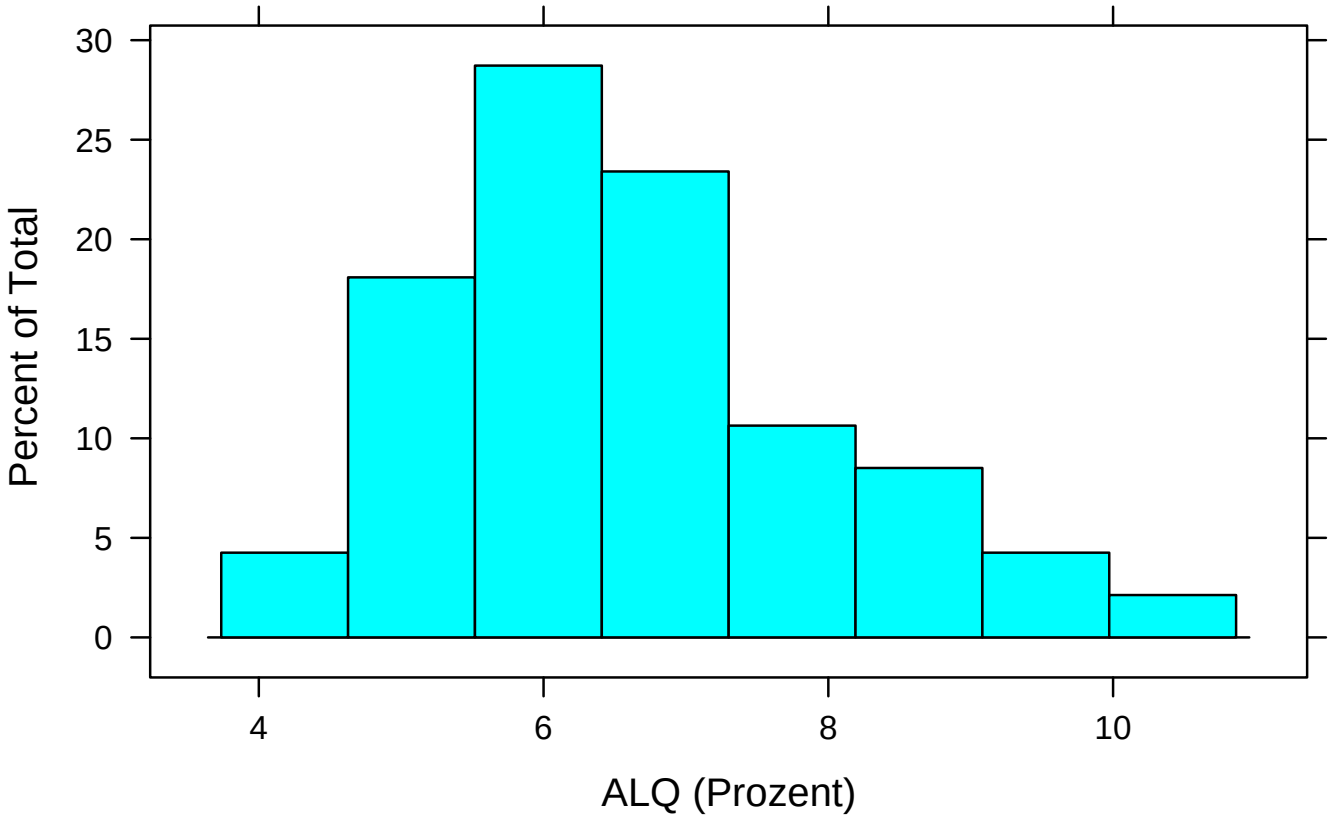


Verteilung der Départements auf die Regionen

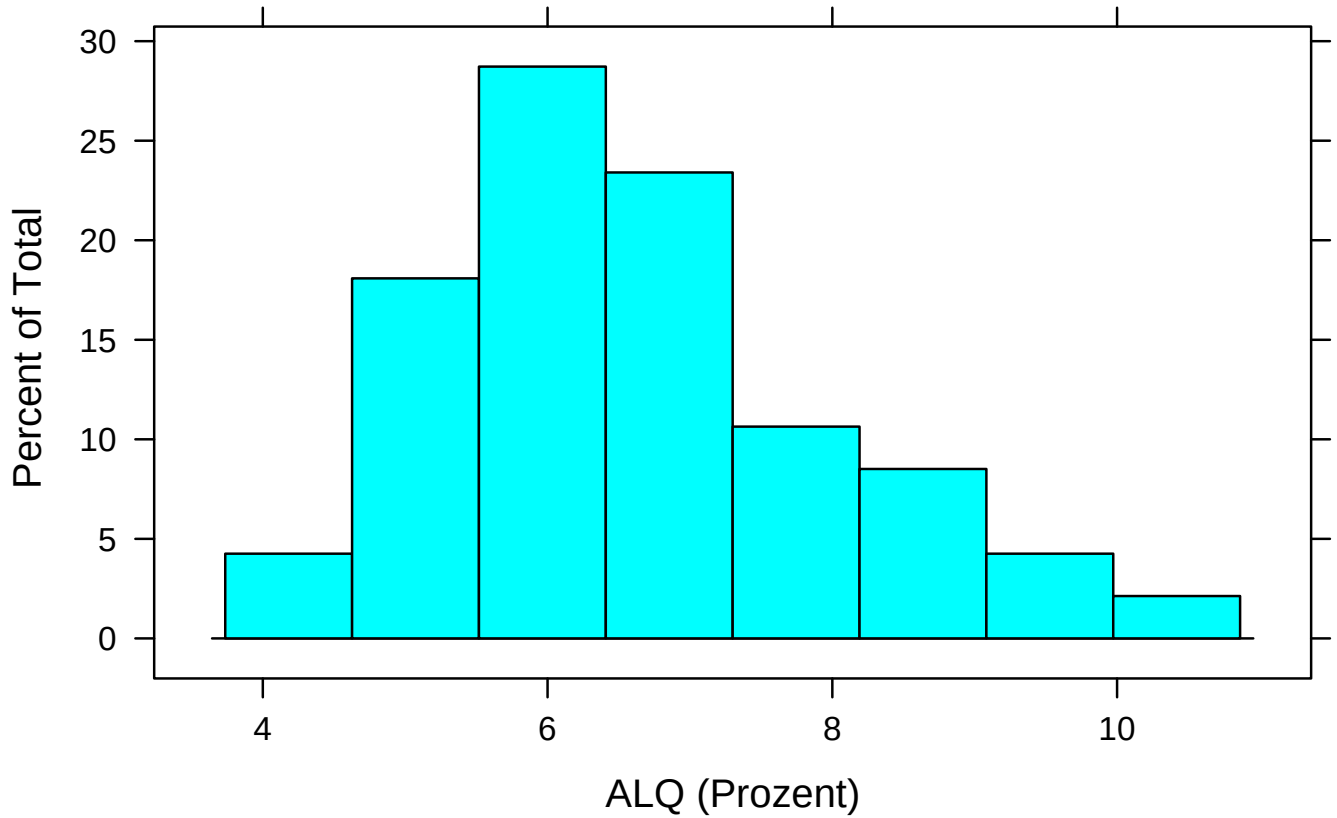


Gesamtlänge der Balken? Bedeutung?

Verteilung der ALQ (gemessen in den Départements)



Verteilung der ALQ (gemessen in den Départements)



Gesamtlänge der Balken? Bedeutung?

Rohdaten $\longrightarrow$ Häufigkeitstabelle

Region	Département
Alsace	Bas-Rhin
Alsace	Haut-Rhin
Aquitane	Dordogne
Aquitane	Gironde
Aquitane	Landes
...	...

Rohdaten $\longrightarrow$ Häufigkeitstabelle

Region	Département
Alsace	Bas-Rhin
Alsace	Haut-Rhin
Aquitane	Dordogne
Aquitane	Gironde
Aquitane	Landes
...	...

$\longrightarrow$

Region	Anzahl
Alsace	2
Aquitane	5
...	...

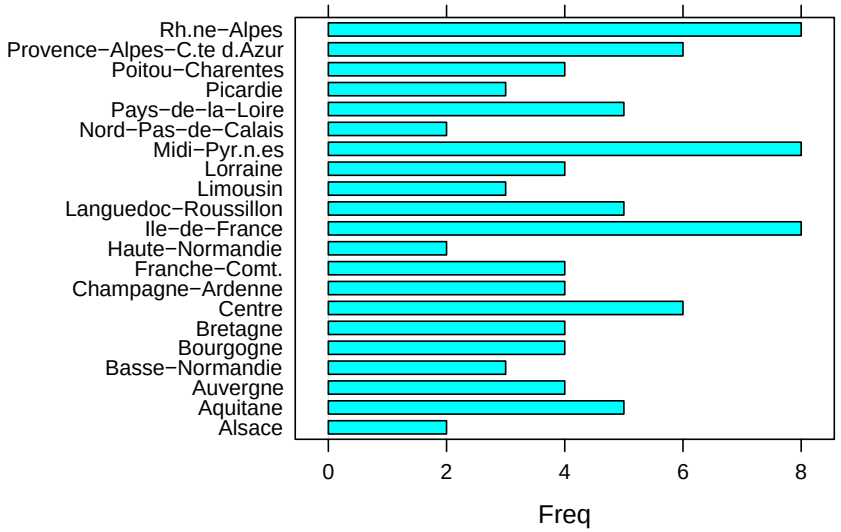
Verhältnis Häufigkeitstabelle - Graphische Darstellung

Region	Département
Alsace	Bas-Rhin
Alsace	Haut-Rhin
Aquitane	Dordogne
Aquitane	Gironde
Aquitane	Landes
...	...

Verhältnis Häufigkeitstabelle - Graphische Darstellung

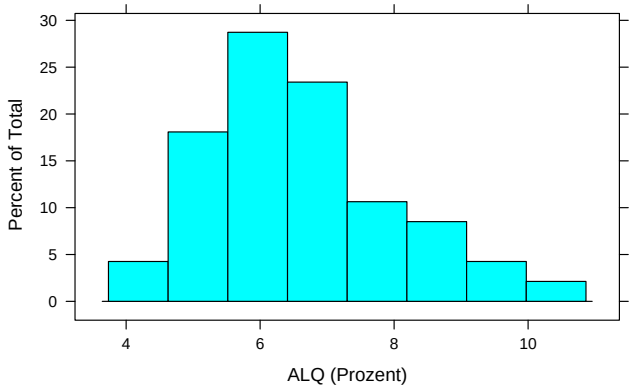
Region	Département
Alsace	Bas-Rhin
Alsace	Haut-Rhin
Aquitane	Dordogne
Aquitane	Gironde
Aquitane	Landes
...	...

=



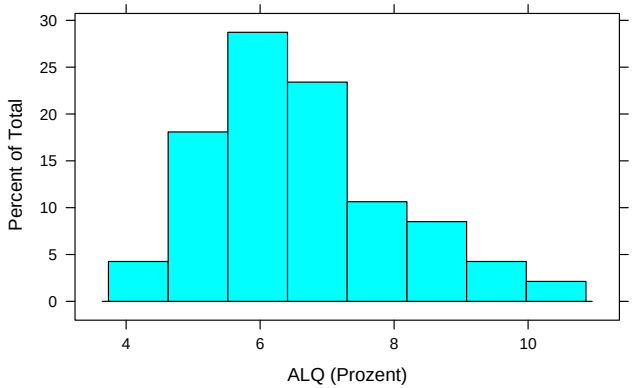
Wozu Maßzahlen?

- Graphische Darstellungen zeigen *vollständige* Verteilung
- Sind anschaulicher



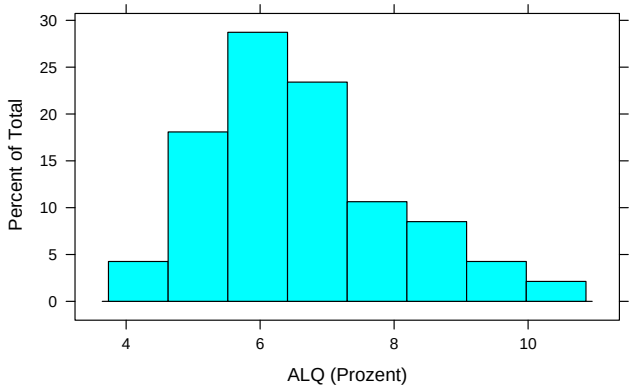
Wozu Maßzahlen?

- Graphische Darstellungen zeigen *vollständige* Verteilung
- Sind anschaulicher
- Aber . . .
  - Nehmen viel Platz ein
  - Aufwendig/unübersichtlich bei großen Datensätzen
  - Schlecht zu vergleichen



Wozu Maßzahlen?

- Graphische Darstellungen zeigen *vollständige* Verteilung
- Sind anschaulicher
- Aber . . .
  - Nehmen viel Platz ein
  - Aufwendig/unübersichtlich bei großen Datensätzen
  - Schlecht zu vergleichen
- Maßzahlen quantifizieren *Eigenschaften der Verteilung*



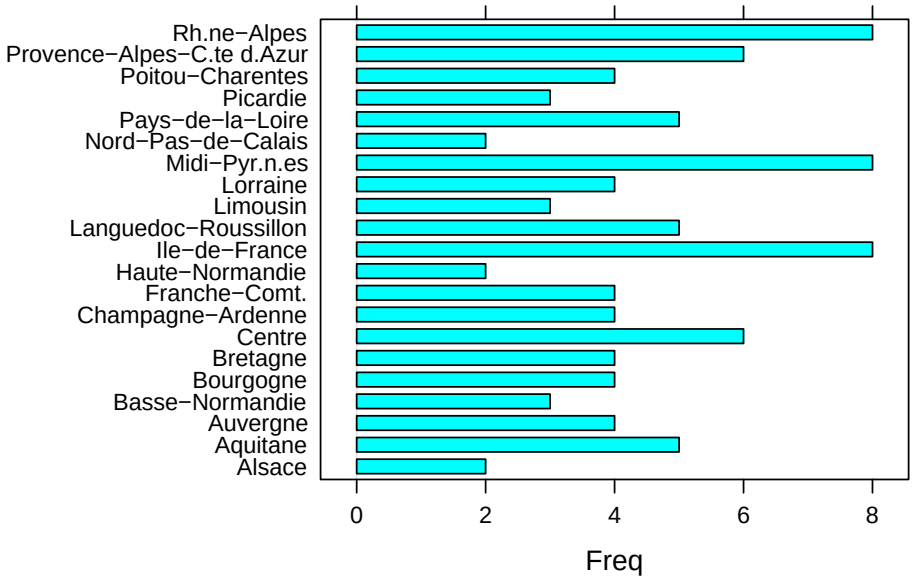
Was sind Mittelwerte?

- Beziehen sich auf „zentrale Tendenz“ (Mitte) der Verteilung
- Hat die Verteilung tatsächlich eine solche Mitte?
  - Ein, zwei, mehr Gipfel?
  - Berechnung sinnvoll?
  - Berechnung möglich!
- Vielzahl von Mittelwerten (Skalenniveaus)
  - Modus (=Modalwert, Mode)
  - Median
  - Arithmetisches Mittel (arithmetic mean)

Was ist der „Modus“

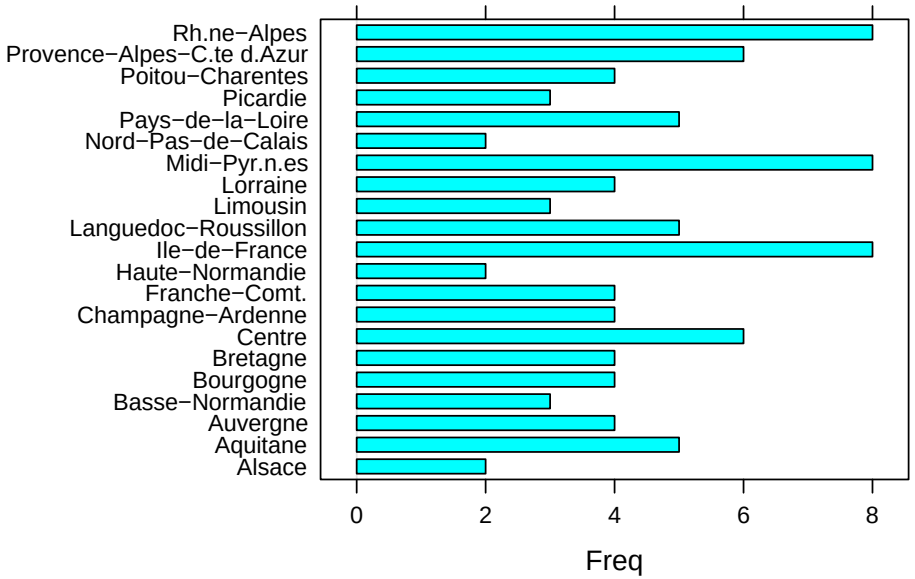
- Ausprägung, die am häufigsten vorkommt
- Abkürzung/Symbol: x_{Mo}
- Direkt aus Verteilung/Häufigkeitstabelle ablesbar
- Für alle Skalenniveaus berechenbar $\longrightarrow$ sinnvoll?
- Unempfindlich gegenüber Extremwerten („Ausreißern“)
- Wenn zwei oder mehr Ausprägungen gleich häufig
  - Arithmetisches Mittel bilden (wenn zulässig)
  - Auf Modus verzichten
- Niedriger Informationsgehalt

Beispiel: (Verteilung auf) Regionen



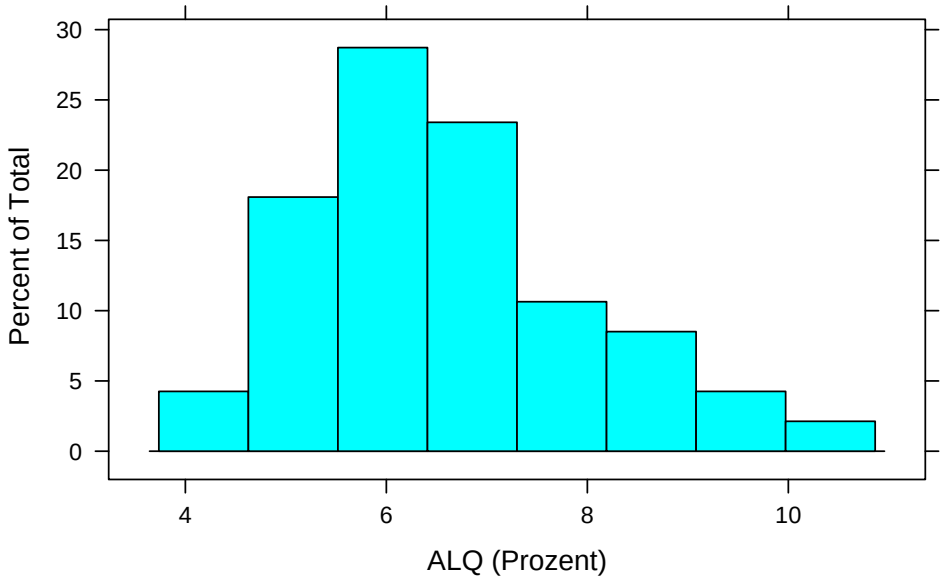
- Drei Regionen am häufigsten (jeweils 8 Départements):
  1. Île-de-France (Ordnungsnummer 11)
  2. Midi-Pyrénées (Ordnungsnummer 15)
  3. Rhône-Alpes (Ordnungsnummer 21)

Beispiel: (Verteilung auf) Regionen



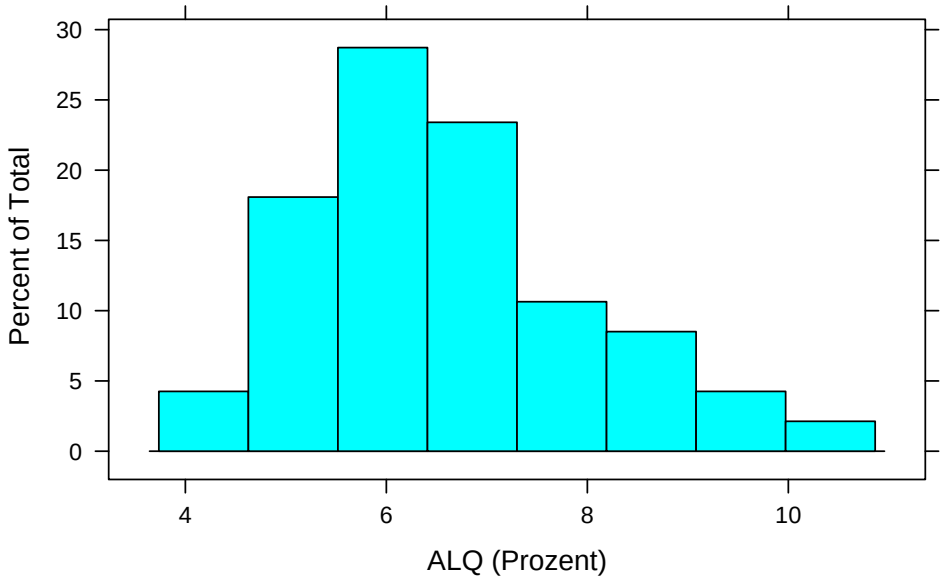
- Drei Regionen am häufigsten (jeweils 8 Départements):
  1. Île-de-France (Ordnungsnummer 11)
  2. Midi-Pyrénées (Ordnungsnummer 15)
  3. Rhône-Alpes (Ordnungsnummer 21)
- Mittelwertbildung *nicht* zulässig/sinnvoll!

Beispiel: Verteilung der ALQ



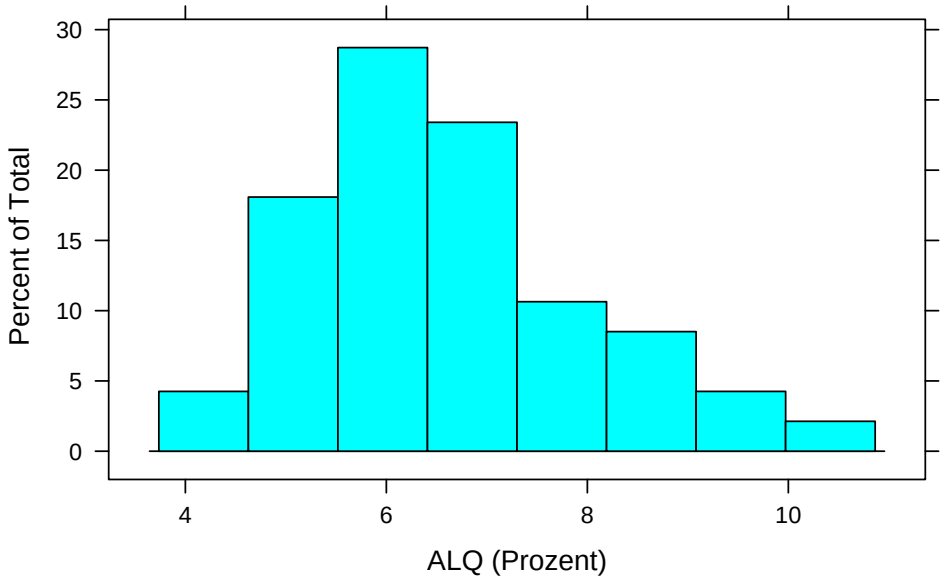
- Zwei Werte am häufigsten (jeweils 6 Départements):

Beispiel: Verteilung der ALQ



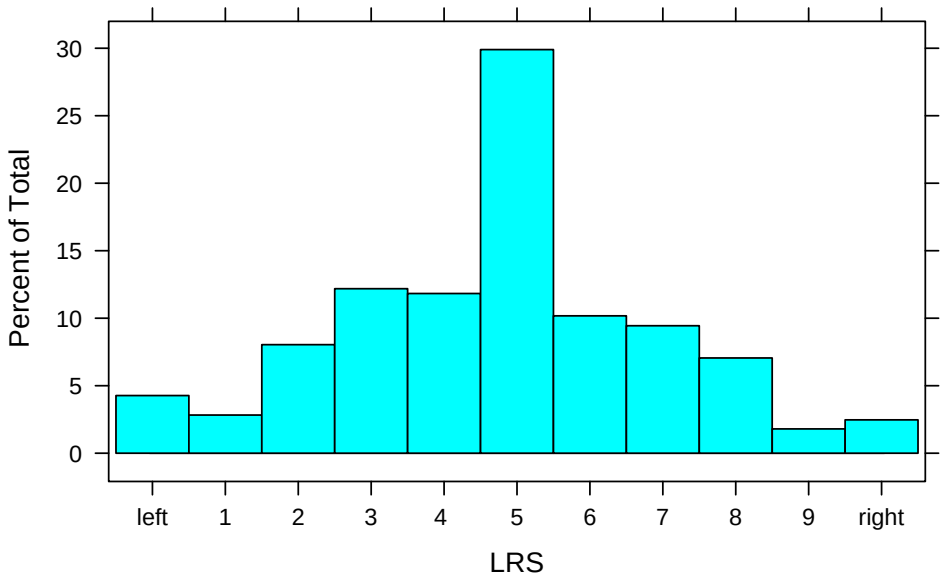
- Zwei Werte am häufigsten (jeweils 6 Départements):
  1. 7 Prozent
  2. 6.19999980926514 Prozent
  3. Arithmetisches Mittel ca. 6.6 Prozent

Beispiel: Verteilung der ALQ



- Zwei Werte am häufigsten (jeweils 6 Départements):
  1. 7 Prozent
  2. 6.19999980926514 Prozent
  3. Arithmetisches Mittel ca. 6.6 Prozent
- Für wirklich kontinuierliche Werte nur beschränkt sinnvoll

Besseres Beispiel: Ideologische Selbsteinstufung im ESS



- 5 463 Befragte (aus drei Ländern)
- Überschaubare Anzahl von (geordneten) Kategorien
- Selbsteinstufung von ganz links (0) bis ganz rechts (10)
- Wert 5 (neutral) ragt klar heraus

Was ist der „Median“

- Intuitivster von allen Mittelwerten – Mitte der Verteilung
- Werte werden aufsteigend sortiert
- Es ist kein Problem, wenn ein Wert mehrfach vorkommt
  - Bei ungerader Fallzahl: genau eine Beobachtung mit Wert, der Verteilung in der Mitte teilt
  - Bei gerader Fallzahl: Median zwischen zwei beobachteten Werten (evtl. Bildung arithmet. Mittel)
- Abkürzung/Symbol: $\tilde{x}$
- Berücksichtigt alle Meßwerte (informativ), zugleich robust gegen Ausreißer
- Entweder alle Fälle sortieren
- Oder Häufigkeitstabelle/gruppierte Daten betrachten

Fiktives Beispiel: Alter von Kursteilnehmern

- Rohdaten: 19,38,22,23,20

Fiktives Beispiel: Alter von Kursteilnehmern

- Rohdaten: 19,38,22,23,20
- Sortiert: 19,20,22,23,38

Fiktives Beispiel: Alter von Kursteilnehmern

- Rohdaten: 19,38,22,23,20
- Sortiert: 19,20,22,23,38
- Median?

Fiktives Beispiel: Alter von Kursteilnehmern

- Rohdaten: 19,38,22,23,20
- Sortiert: 19,20,22,23,38
- Median? **22**

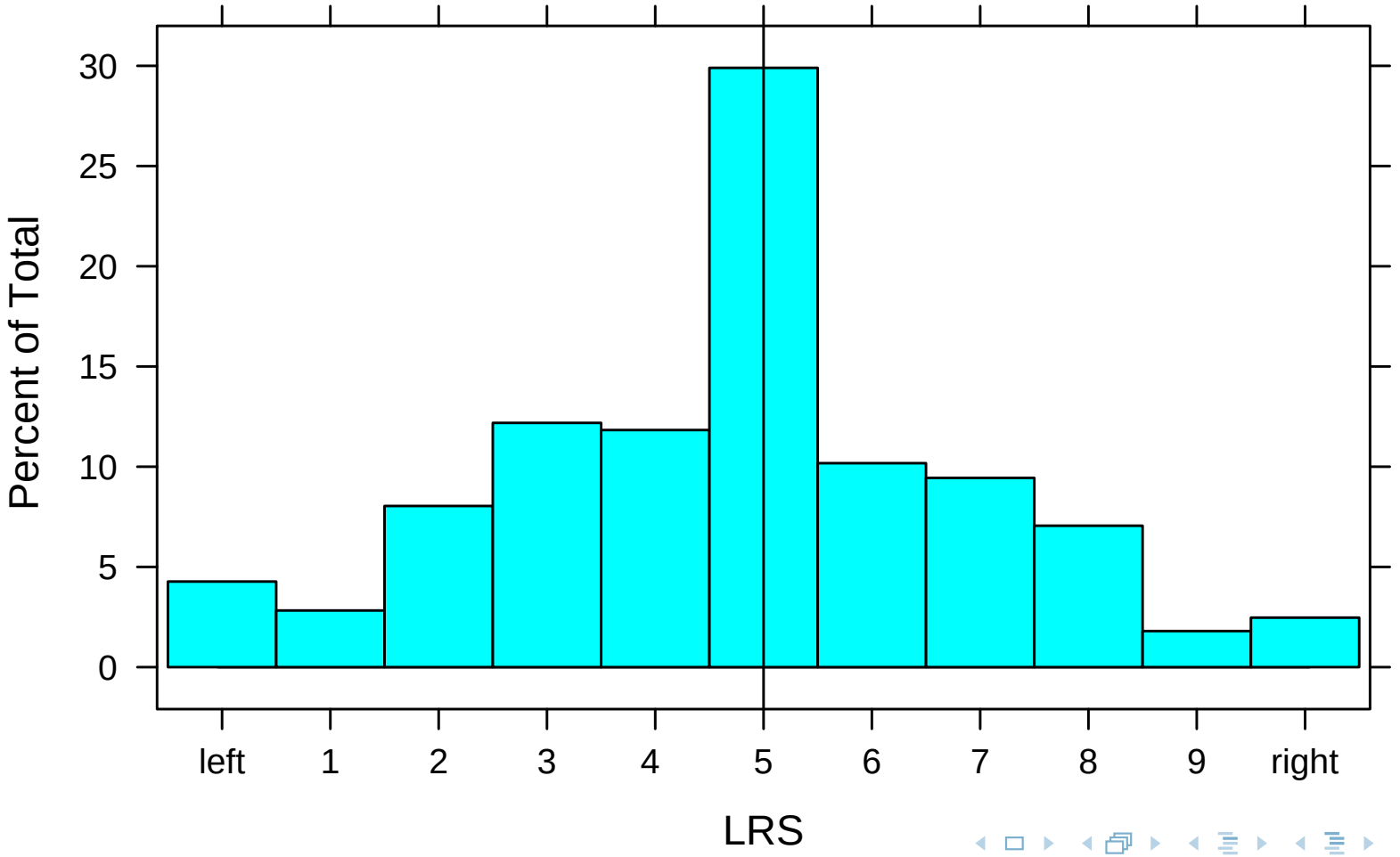
Beispiel: LRS

Wert	abs. Häufigkeit	Prozent	kumm. Prozent
0 (left)	204	4	4
1	135	3	7
2	384	8	15
3	582	12	27
4	565	12	39
5	1428	30	69
6	486	10	79
7	451	9	88
8	337	7	95
9	86	2	97
10 (right)	118	2	99

Beispiel: LRS

Wert	abs. Häufigkeit	Prozent	kumm. Prozent
0 (left)	204	4	4
1	135	3	7
2	384	8	15
3	582	12	27
4	565	12	39
5	1428	30	69
6	486	10	79
7	451	9	88
8	337	7	95
9	86	2	97
10 (right)	118	2	99

Median der LRS-Verteilung



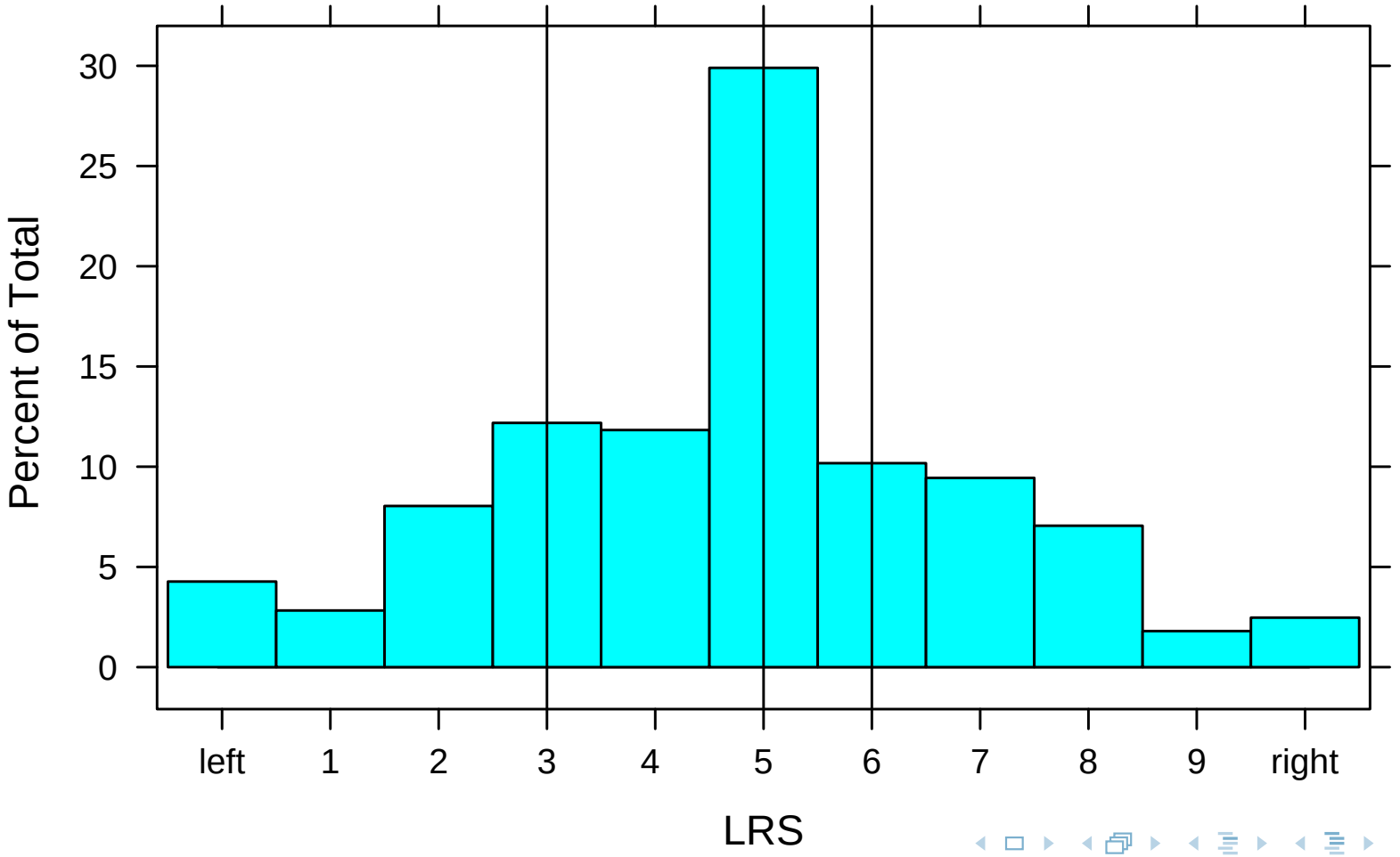
Median und extreme Werte

- Werte am Rand der Verteilung, weit weg von allen anderen: „Ausreißer“
  - Meßfehler?
  - Passen inhaltlich nicht zur Mehrheit der Fälle?
- Median (und Modus) reagieren nicht auf Ausreißer – „robust“
- Mehr dazu gleich

Was sind Perzentile?

- Idee des Medians läßt sich verallgemeinern
  - Median teilt Verteilung bei 50% $\longrightarrow$ zwei Teile
  - Drei Werte, Verteilung bei die 25%, die 50%, die 75% teilen $\longrightarrow$ vier Teile, Quartile
  - Vier Werte, die bei die 20%, die 40%, die 60%, ... teilen $\longrightarrow$ fünf Teile, Quintile
- allgemein: Perzentile
- Visualisiert durch Boxplot

Quartile der LRS-Verteilung



Beispiel: LRS

Wert	abs. Häufigkeit	Prozent	kumm. Prozent
0 (left)	204	4	4
1	135	3	7
2	384	8	15
3	582	12	27
4	565	12	39
5	1428	30	69
6	486	10	79
7	451	9	88
8	337	7	95
9	86	2	97
10 (right)	118	2	99

Was ist ein Boxplot?

- Hochformat/Querformat
- Boxplot (oder Box-Whisker-Plot) visualisiert
  - Unteres Quartil (Boden/linke Seite der Box)
  - Median (Linie oder Punkt in der Mitte der Box)
  - Oberes Quartil (Deckel/rechte Seite der Box)
  - Spannweite der Daten
  - „Ausreißer“ (unterschiedliche Definitionen)

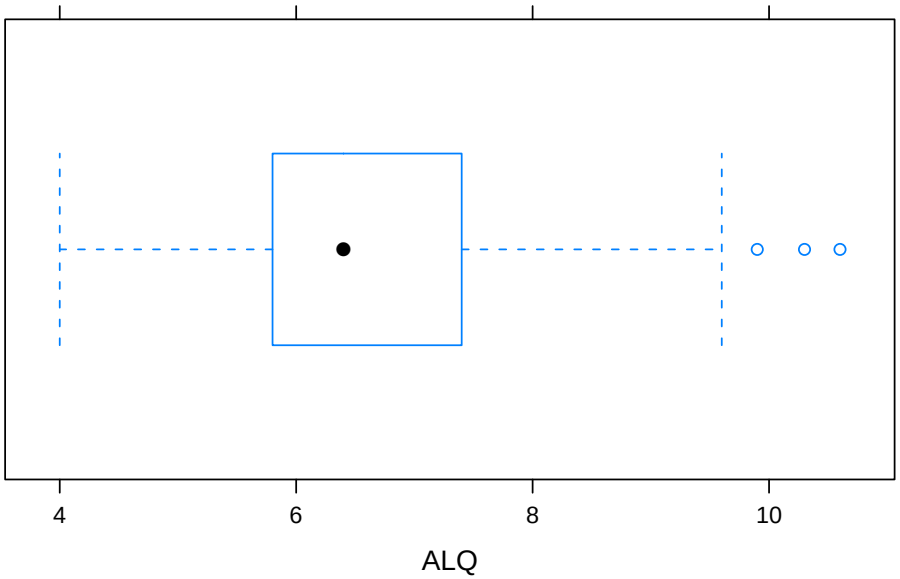
Was ist ein Boxplot?

- Hochformat/Querformat
- Boxplot (oder Box-Whisker-Plot) visualisiert
  - Unteres Quartil (Boden/linke Seite der Box)
  - Median (Linie oder Punkt in der Mitte der Box)
  - Oberes Quartil (Deckel/rechte Seite der Box)
  - Spannweite der Daten
  - „Ausreißer“ (unterschiedliche Definitionen)
- Box (IQR): mittlere 50% der Verteilung
- „Whiskers“: max(Extremwerte oder 1.5 $\times$ Länge der Box von der Box weg)
- Ausreißer: mehr als 1.5$\times$ IQR vom oberen/unteren Quartil entfernt

Was ist ein Boxplot?

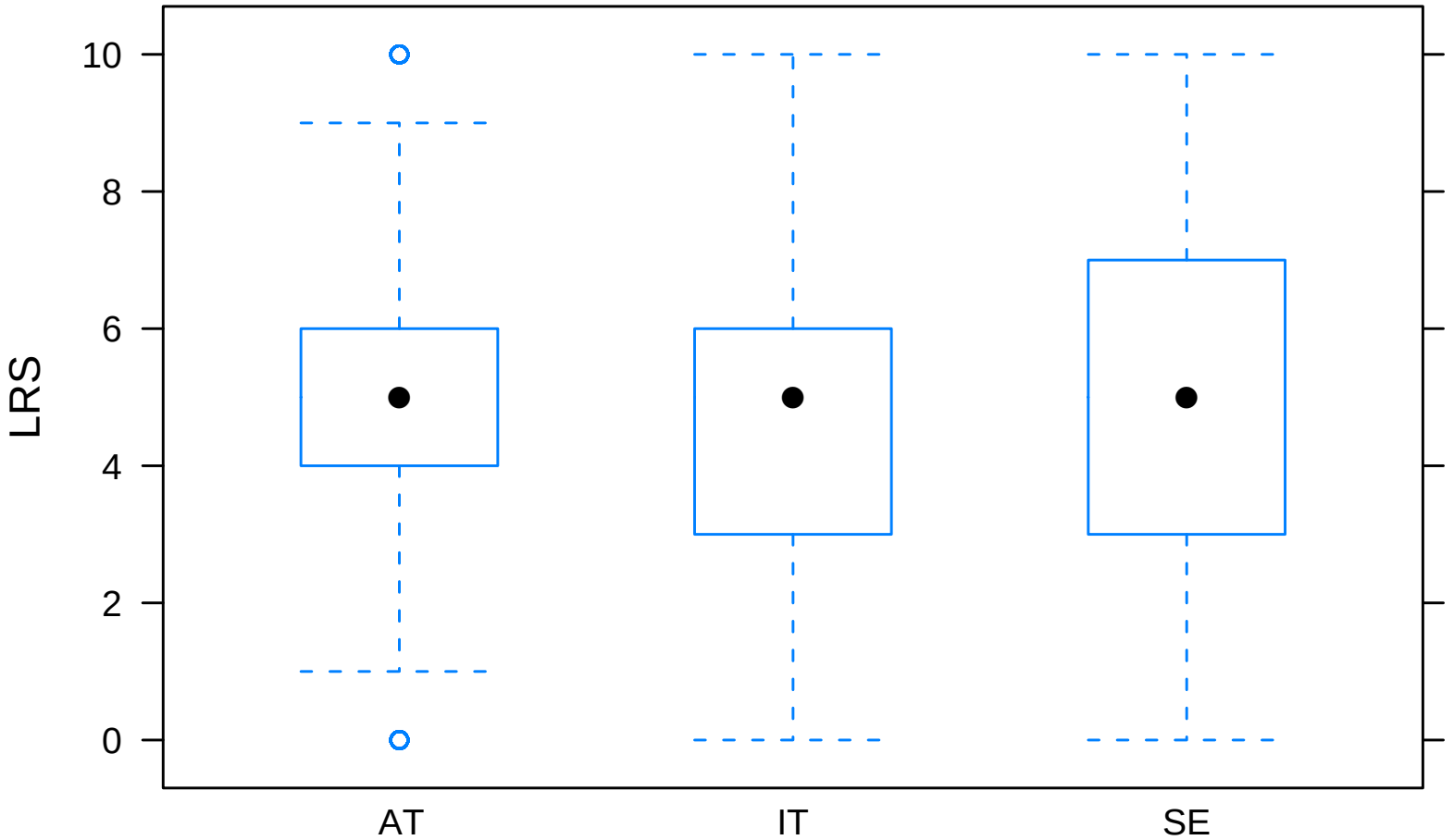
- Hochformat/Querformat
- Boxplot (oder Box-Whisker-Plot) visualisiert
  - Unteres Quartil (Boden/linke Seite der Box)
  - Median (Linie oder Punkt in der Mitte der Box)
  - Oberes Quartil (Deckel/rechte Seite der Box)
  - Spannweite der Daten
  - „Ausreißer“ (unterschiedliche Definitionen)
- Box (IQR): mittlere 50% der Verteilung
- „Whiskers“: max(Extremwerte oder 1.5 $\times$ Länge der Box von der Box weg)
- Ausreißer: mehr als 1.5$\times$ IQR vom oberen/unteren Quartil entfernt
- Kompakte Darstellung von Mittelwert/Median, Streuung und Form der Verteilung (mehr dazu gleich)
- Vergleich von zwei oder mehr Verteilungen

Beispiel: Verteilung ALQ

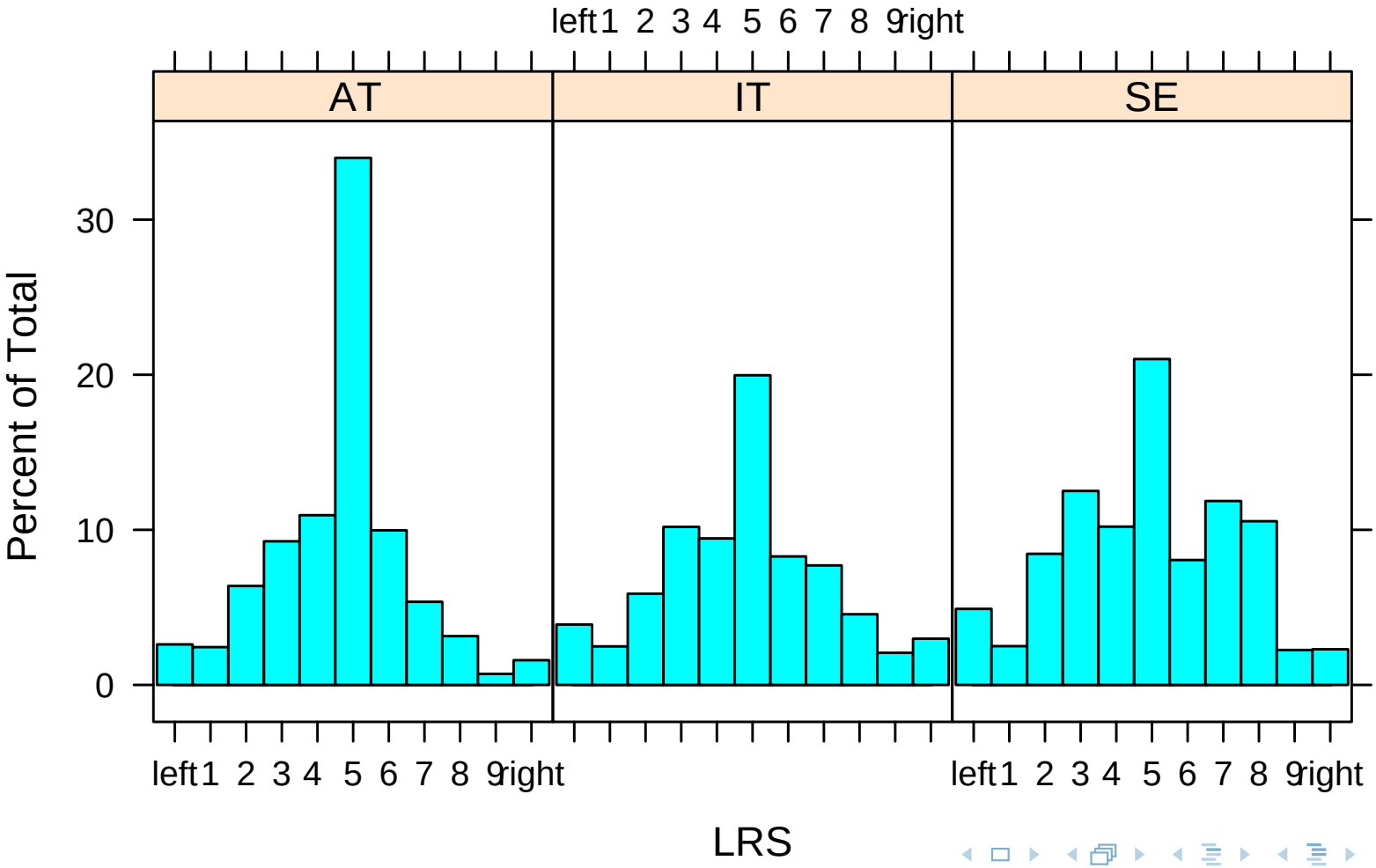


- 1., 2. (Median), 3. Quartil: 5.8, 6.4, 7.4 $\rightarrow$ Bedeutung?
- Whiskers 4.0, 9.6
- Ausreißer: 9.9, 10.3, 10.6

Beispiel: LRS in drei Ländern



Beispiel: LRS in drei Ländern II



Was ist das „arithmetische Mittel“?

- Der bekannteste aller Mittelwerte
- Grundprinzip: Alle Werte aufsummieren und durch Zahl der Fälle teilen
  - Z. B. Alter der fünf Kursteilnehmer:
    $(19 + 38 + 22 + 23 + 20)/5 = 24.4$
  - Mittlere ALQ der Départements =
    $(4 + 4.1 + \cdots + 10.3 + 10.6)/94 \approx 6.6$
- Nutzt Maximum an in den Daten enthaltenen Informationen
- Optimale Anpassung
  - Summe der einfachen Abweichungen $= 0$
  - Summe der quadrierten Abweichungen $=$ min.
- Reagiert empfindlich auf Ausreißer
  - Durchschnittsalter ohne ältesten Teilnehmer $= 21$
  - ALQ ohne Département mit niedrigstem Wert $= 6.7$

Trimmed Mean

- „Extreme“ Bereiche der Verteilung (z. B. obere/untere 10% der Werte weglassen)
- Beispiel ALQ
  - Sortieren
  - Neun Werte unten/neun Werte oben entfernen
  - $(5.1 + 5.2 \ldots + 8.6 + 8.8)/\mathbf{76} = 6.54$
- Arithmetisches Mittel für Bereich zwischen zwei Perzentilen (z. B. hier erstes/neuntes Dezentil) – Kompromiß zwischen Median und arithmetischem Mittel
- Informationsverlust

Rechnen mit dem Summenzeichen

- In Formeln wird das Summenzeichen Sigma (Σ) benutzt
- Aufsummierung über alle Fälle
- Index (Ordnungsnummer, z. B. $x_1, x_2, \cdots$
- $i = 1$: Indexwert startet mit 1
- n: Indexwert endet mit letztem Fall (n = Zahl der Fälle)
- Beide Zusätze oft weggelassen

Arithmet. Mittel

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Gruppierte Daten

- Vereinfachte Berechnung wenn Häufigkeitstabelle vorliegt
- Häufigkeit der Ausprägung mit Wert der Ausprägung multiplizieren
- Aufsummieren und durch Fallzahl teilen
- $0 \times 204 + 1 \times 135 \cdots + 10 \times 118 = 27548$
- $27548/4776 = \bar{x} = 5.8$

Was bedeutet Streuung (Varianz)?

- Alle Meßwerte untereinander (und mit dem Mittelwert) identisch – keine Streuung
- Heterogenere Werte – mehr Streuung
- Zweigipflige Verteilung, alle Werte an den Rändern – extreme Streuung
- Streuung = Variation um zentrale Tendenz der Verteilung (Mittelwert)
- Vergleich von Verteilungen
- Wie messen?

Range und Interquartilsabstand

- Range (Spannweite, V) einfachste Form der Streuungmessung
- Differenz zwischen höchstem und niedrigstem Wert
- Z. B. ALQ: 10.6-4=6.6
- Anschaulich
- Probleme
  - Per definitionem anfällig gegen Ausreißer
  - Berücksichtigt nur zwei Werte $\rightarrow$ nicht informativ
- Interquartilsabstand
  - Abstand zwischen 1. und 3. Quartil
  - ALQ: 7.375 - 5.825 = 1.55
  - Anschaulich
  - Betrachtet nur mittlere 50% der Daten
- $\rightarrow$ Varianz

Was ist die Varianz?

- Abkürzung s^2 (für Stichproben), σ^2 (für wahren Wert in Grundgesamtheit)
- Berücksichtigt alle Werte
- Entspricht mittlerer quadrierter Abweichung der Meßwerte von ihrem Mittelwert
- Quadrieren
  - Läßt Vorzeichen verschwinden (Summe der einfachen Abweichungen $= 0$)
  - Gibt größeren Abweichungen mehr Gewicht

Beispiel: fiktive Alterswerte

- Rohdaten: 19,38,22,23,20; $\bar{x} = 24.4$
- Einfache Abweichungen: -5.4 13.6 -2.4 -1.4 -4.4
- Quadrierte Abweichungen: 29.16 184.96 5.76 1.96 19.36
- Summe quadrierte Abweichungen (SAQ) = 241.2
- Mittlere quadrierte Abweichung = SAQ/n = $\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n} = s^2$

Beispiel: fiktive Alterswerte

- Rohdaten: 19,38,22,23,20; $\bar{x} = 24.4$
- Einfache Abweichungen: -5.4 13.6 -2.4 -1.4 -4.4
- Quadrierte Abweichungen: 29.16 184.96 5.76 1.96 19.36
- Summe quadrierte Abweichungen (SAQ) = 241.2
- Mittlere quadrierte Abweichung = SAQ/n = $\frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n} = s^2$
- Obacht: Varianz in der Stichprobe unterschätzt Varianz in der Grundgesamtheit
- Um einen Faktor von $\frac{n-1}{n}$ (geht für große Stichproben gegen 1)
- Schätzung für GG: Stichprobenvarianz mit Kehrwert $\frac{n}{n-1}$ multiplizieren
- $\frac{SAQ}{n} \times \frac{n}{n-1} = \frac{SAQ}{n-1}$

Standardabweichung

- Durch Quadrieren geht die Einheit verloren

Standardabweichung

- Durch Quadrieren geht die Einheit verloren
  - 48.2 Quadratjahre?
  - ALQ von 1.87 Quadratprozent?

Standardabweichung

- Durch Quadrieren geht die Einheit verloren
  - 48.2 Quadratjahre?
  - ALQ von 1.87 Quadratprozent?
- Standardabweichung (SD, s) = Quadratwurzel aus Varianz: $s_{ALQ} = \sqrt{1.87} = 1.37$, $s_{Alter} = \sqrt{48.2} = 6.95$

Standardabweichung

- Durch Quadrieren geht die Einheit verloren
  - 48.2 Quadratjahre?
  - ALQ von 1.87 Quadratprozent?
- Standardabweichung (SD, s) = Quadratwurzel aus Varianz: $s_{ALQ} = \sqrt{1.87} = 1.37$, $s_{Alter} = \sqrt{48.2} = 6.95$
- Ursprüngliche Einheit, aber keine einfache Interpretation

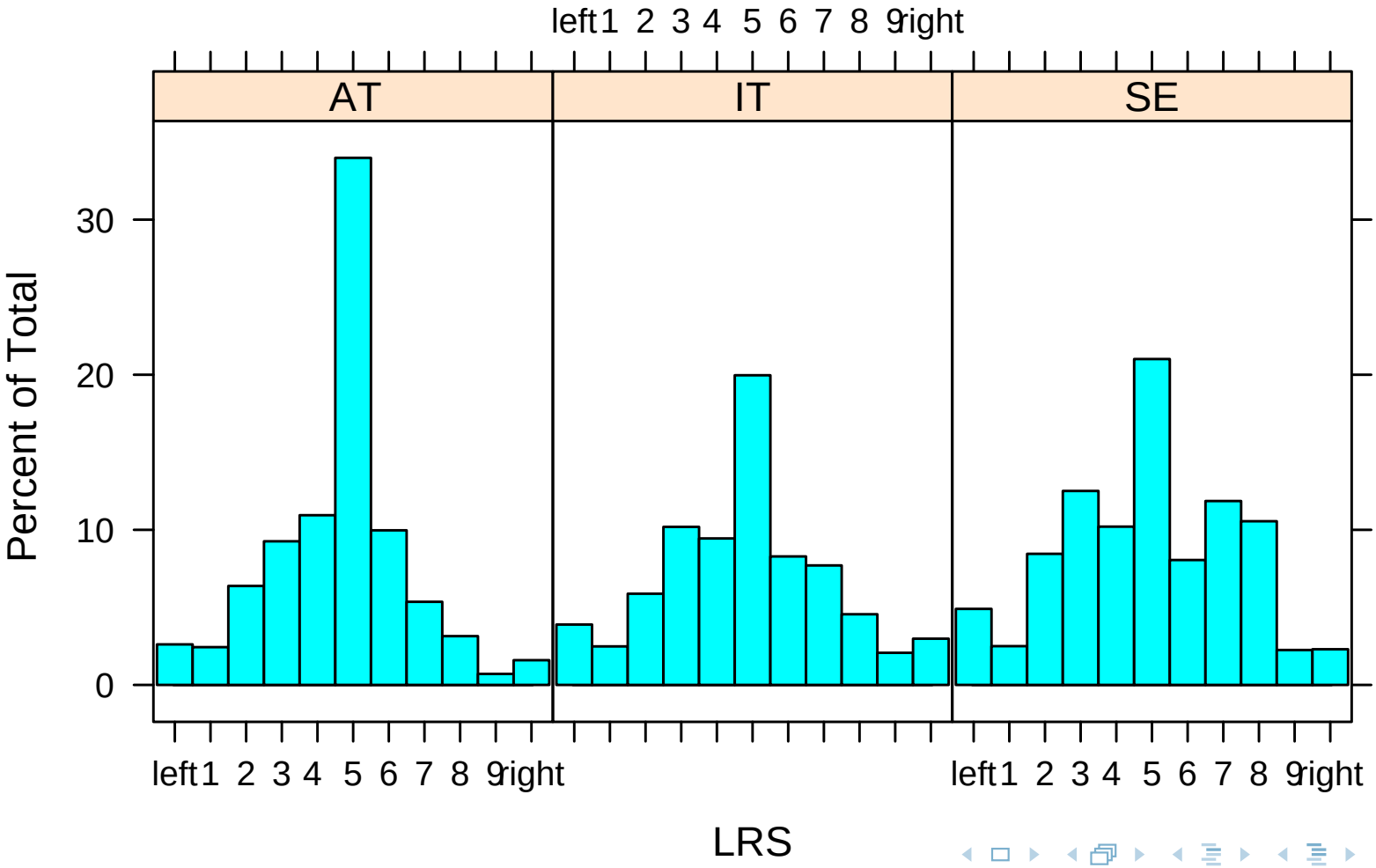
Standardabweichung

- Durch Quadrieren geht die Einheit verloren
  - 48.2 Quadratjahre?
  - ALQ von 1.87 Quadratprozent?
- Standardabweichung (SD, s) = Quadratwurzel aus Varianz: $s_{ALQ} = \sqrt{1.87} = 1.37$, $s_{Alter} = \sqrt{48.2} = 6.95$
- Ursprüngliche Einheit, aber keine einfache Interpretation
  - $s \neq$ mittlere Abweichung vom Durchschnittsalter
  - $s \neq$ Mittelwert der Beträge der einzelnen Abweichungen

Standardabweichung

- Durch Quadrieren geht die Einheit verloren
  - 48.2 Quadratjahre?
  - ALQ von 1.87 Quadratprozent?
- Standardabweichung (SD, s) = Quadratwurzel aus Varianz: $s_{ALQ} = \sqrt{1.87} = 1.37$, $s_{Alter} = \sqrt{48.2} = 6.95$
- Ursprüngliche Einheit, aber keine einfache Interpretation
  - $s \neq$ mittlere Abweichung vom Durchschnittsalter
  - $s \neq$ Mittelwert der Beträge der einzelnen Abweichungen
- Beispiel LRS
  - AT: $s = 1.9$
  - IT: $s = 2.3$
  - SE: $s = 2.4$

Beispiel: LRS in drei Ländern II



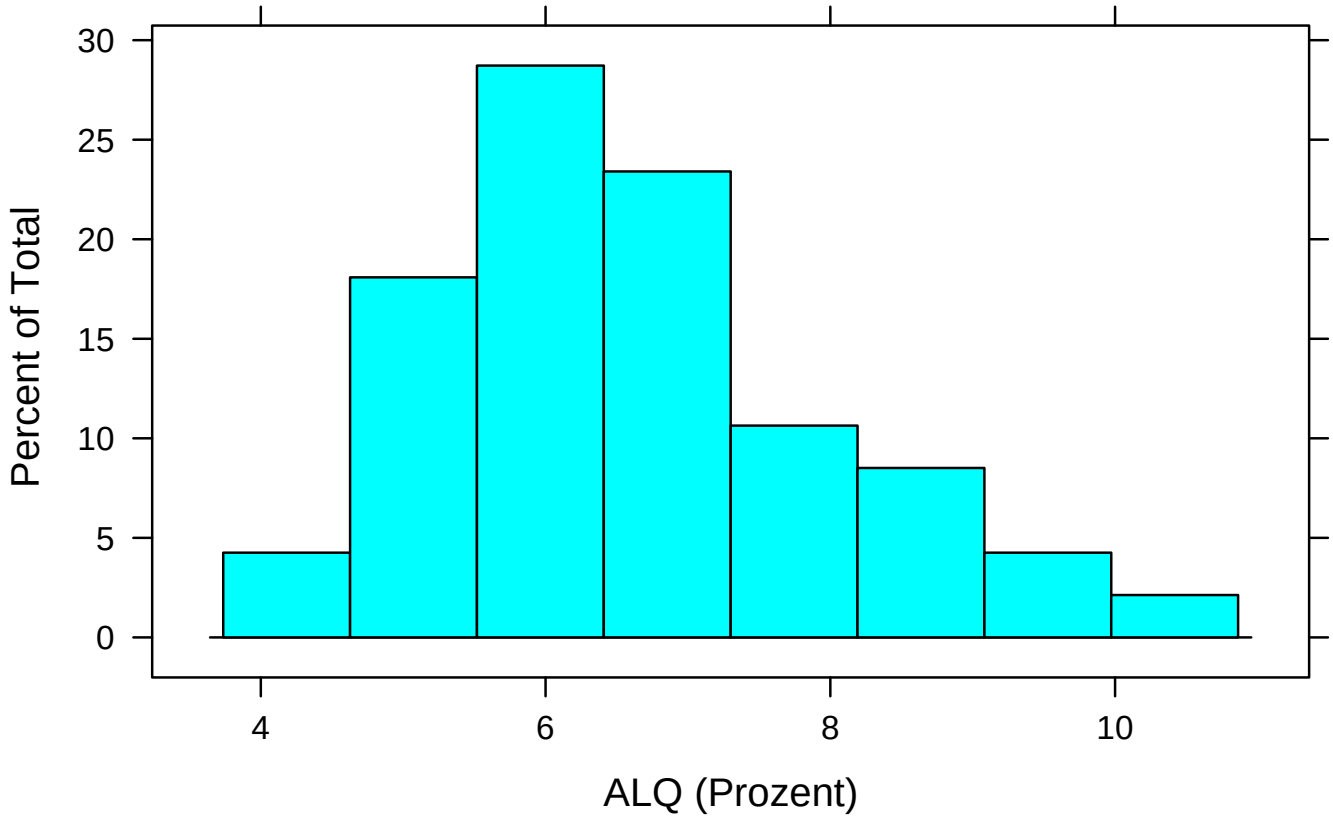
Was ist „Schiefe“ (Skewness, Skew)

- Verteilungen können symmetrisch oder schief sein
- Symmetrische Verteilungen: 0 (z. B. LRS)
- negative skewness: linksschief bzw. rechtssteil
- positive skewness: rechtsschief bzw. linkssteil
- Beispiel dafür: ALQ ($\gamma_1 = 0.7$)
- In den Texten angesprochene Faustregel Reihenfolge Mittelwerte $\longrightarrow$ Form der Verteilung funktioniert oft nicht

Skewness

$$\gamma_1 = \frac{\frac{1}{n}\sum(x_i - \bar{x})^3}{s^3}$$

ALQ: $\gamma_1 = 0.7$



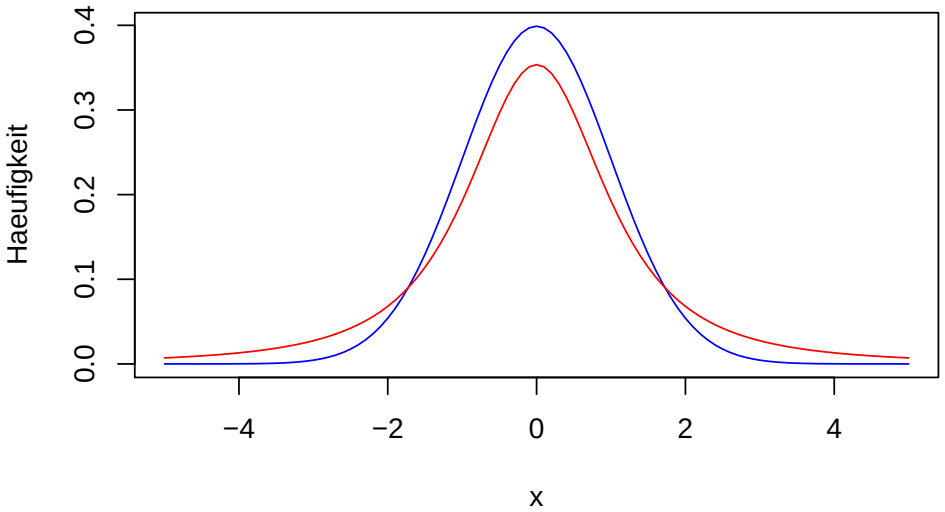
Was ist „Wölbung“, „Exzeß“, „Kurtosis“

- Schmaler/breiter Gipfel der Verteilung?
- $\gamma_2 > 0$: Schmaler Gipfel, relativ viele Fälle extremen Werten
- $\gamma_2 < 0$: Breiter Gipfel, relativ viele Fälle in der Nähe des Mittelwertes
- (Impliziter Vergleich mit Normalverteilung gleicher Varianz)
- ALQ: $\gamma_2 = 0.29$; LRS: $\gamma_2 = -0.06$

Kurtosis

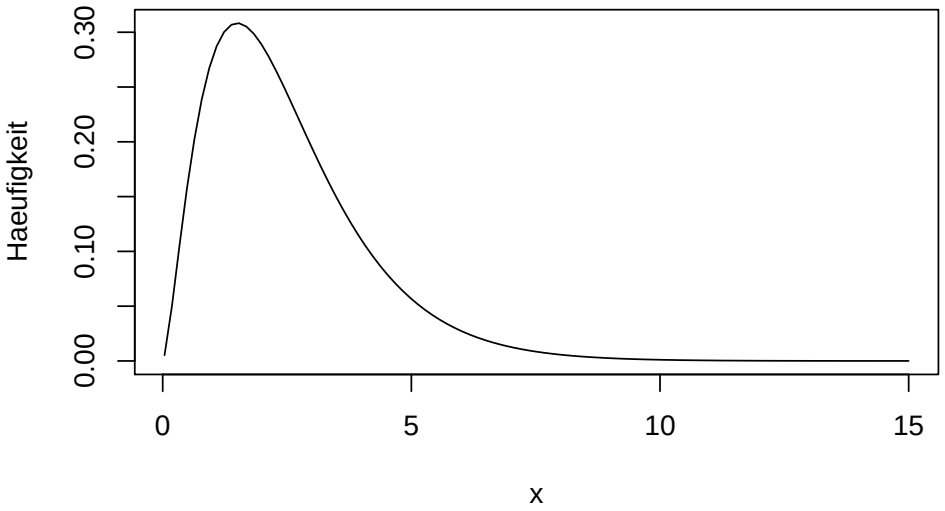
$$\gamma_2 = \frac{\frac{1}{n}\sum(x_i - \bar{x})^4}{s^4} - 3$$

Ein Beispiel



- Beide Verteilungen eingipflig und symmetrisch
- Rote Verteilung hat mehr extreme Werte (links und rechts)
- Weniger Werte in der Nähe des Mittelwertes → schmaler Gipfel, große Kurtosis

Ein finales Beispiel



- $\bar{x} = 2.48, \tilde{x} = 2.15$
- Quantile: 1.3, 3.3
- $s = 1.58$
- $\gamma_1 = 1.33, \gamma_2 = 2.68$

Warum z-Standardisierung?

- Werte aus unterschiedlichen Verteilungen (Mittelwert, Standardabweichung) vergleichen

1. Zentrierung: Mittelwert der Verteilung vom Meßwert abziehen
2. Standardisierung: Ergebnis durch Standardabweichung teilen

- Ergebnis: z-Werte
- Alter: 19 38 22 23 20
- zentriert: -5.4 13.6 -2.4 -1.4 -4.4
- standardisiert ($s = 6.95$): -0.78 1.96 -0.35 -0.20 -0.63

Warum z-Werte?

- Ursprungswerte werden als Abweichung (in Standardabweichungen) von ihrem Mittelwerte ausgedrückt
- Macht relative Position von Werten aus verschiedenen Verteilungen vergleichbar
- (wenn deren Form vergleichbar und am besten näherungsweise normal ist)
- Abweichungen von mehr als $\pm$ Standardabweichungen gelten als ungewöhnlich (Ausreißer)
- Z-standardisierte Werte haben ihrerseits einen Mittelwert von 0 und eine Standardabweichung von 1 (warum? $\rightarrow$ Hausaufgabe)
- Erleichtert den Umgang mit Normalverteilungen (mehr dazu bald)

Zusammenfassung

- Verteilungen von Variablen haben eine Mitte
- Eine Streuung um diese Mitte
- Sie sind symmetrisch oder schief
- Und haben eine mehr oder minder breiten Gipfel

Zusammenfassung

- Verteilungen von Variablen haben eine Mitte
- Eine Streuung um diese Mitte
- Sie sind symmetrisch oder schief
- Und haben eine mehr oder minder breiten Gipfel
- Für alle Eigenschaften gibt es Maßzahlen

Zusammenfassung

- Verteilungen von Variablen haben eine Mitte
- Eine Streuung um diese Mitte
- Sie sind symmetrisch oder schief
- Und haben eine mehr oder minder breiten Gipfel
- Für alle Eigenschaften gibt es Maßzahlen
- Z-Standardisierung macht Werte aus unterschiedlichen Verteilungen vergleichbar