Einige Grundbegriffe der Statistik

Friedrich Graef[1]

Dezember 2011

Inhaltsverzeichnis

Dies ist die Ausarbeitung einer zweistündigen Vorlesung für Studenten der Zahnmedizin, in der einige Methoden der beschreibenden und schließenden Statistik vorgestellt werden, die in wissenschaftlichen Arbeiten auf diesem Gebiet häufig Anwendung finden. Im wesentlichen wird erläutert, wie graphische Darstellungen gestaltet werden sollten und wie die Resultate statistischer Analysen zu interpretieren sind. Auf die Darstellung von Formeln und Rechenverfahren wird verzichtet. Dazu gibt es schließlich Computerprogramme.
Das erste Kapitel befasst sich mit der beschreibenden Statistik, speziell mit der Gestaltung von Histogrammen und den statistischen Kenngrößen Mittelwert und Standardabweichung.
Im zweiten Kapitel werden Konfidenzbereiche und Tests als wichtigste Hilfsmittel bei der Analyse von Messreihen vorgestellt.

Dezember 2011, Friedrich Graef

[1] Akademischer Direktor i.R., bis 2010 am Department Mathematik der FAU

1 Beschreibende Statistik

Die beschreibende Statistik befasst sich mit dem Aufspüren von Gesetzmäßigkeiten in *großen* Datenbeständen. Das wesentliche Hilfsmittel ist dabei das Komprimieren der Daten in Form von

- Zahlen: Tabellen
- Grafiken: Histogramme
- Kenngrößen: Mittelwert, Standardabweichung usw.

1.1 Die statistische Grundgesamtheit

Ausgangspunkt ist dabei die *Statistische Grundgesamtheit*. Das ist eine Menge von — in gewisser Weise gleichartigen— Objekten, an denen bestimmte Merkmale beobachtet werden.
Zur Veranschaulichung betrachten wir die Forbes2000-Liste der 2000 größten Unternehmen der Welt, wie sie alljährlich von der Wirtschaftszeitschrift Forbes zusammengestellt wird.

rank	name	country	category	sales	profits	assets	marketvalue
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
5	BP	United Kingdom	Oil & gas operations	232.57	10.27	177.57	173.54
6	Bank of America	United States	Banking	49.01	10.81	736.45	117.55
7	HSBC Group	United Kingdom	Banking	44.33	6.66	757.6	177.96
8	Toyota Motor	Japan	Consumer durables	135.82	7.99	171.71	115.4
9	Fannie Mae	United States	Diversified financials	53.13	6.48	1019.17	76.84
10	Wal-Mart Stores	United States	Retailing	256.33	9.05	104.91	243.74
11	UBS	Switzerland	Diversified financials	48.95	5.15	853.23	85.07
12	ING Group	Netherlands	Diversified financials	94.72	4.73	752.49	54.59
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Die Merkmale, die dabei an den Objekten „Unternehmen“ beobachtet werden, sind dabei Rang, Name, Land, Sparte, Umsatz, Gewinn, Kapital und Marktwert.

1.1.1 Merkmale und Ausprägungen

Die Werte, die bei der Beobachtung eines Merkmals gemessen werden, heißen *Ausprägungen*. Je nach Art der Ausprägungen unterscheidet man drei verschiedene Typen von Merkmalen.

Qualitative Merkmale: Name, Land, Sparte sind sog. qualitative Merkmale. Die Ausprägungen sind verschiedene Symbole oder Worte.

Quantitative Merkmale: Bei Zahlen als Ausprägungen gibt es zwei unterschiedliche Arten.

Ordinalzahlen: Der Rang eines Unternehmens ist eine natürliche Zahl: 1,2,3,. . . ,2000. Er ist eigentlich ein qualitatives Merkmal, aber mit einer Ordnungsstruktur: 1 ist besser als 2, 2 ist besser als 3, usw.

Kardinalzahlen: Das sind Umsatz, Gewinn u.ä., also Dezimalzahlen.

1.1.2 Häufigkeitsverteilungen

Statistische Gesetzmäßigkeiten sind im wesentlichen die *Häufigkeitsverteilungen*: Wie verteilen sich die Objekte einer Grundgesamtheit über die verschiedenen Ausprägungen eines Merkmals? Wie verteilen sich z.B. die Firmen auf Länder oder Sparten.

1.2 Histogramme

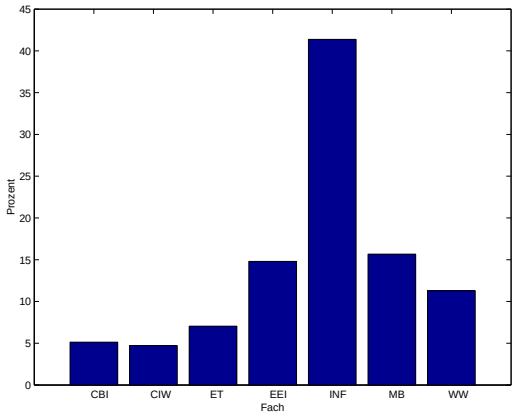
Häufigkeitsverteilungen werden zahlenmäßig als Tabelle, graphisch als Histogramm dargestellt. Wir befassen uns hier mit der Frage, was in diesem Zusammenhang bei qualitativen, ordinalen und kardinalen Merkmalen zu beachten ist.

1.2.1 Qualitative Merkmale

Die Häufigkeitsverteilung der Studenten einer Technischen Hochschule über die verschiedenen Fachrichtungen sei durch die folgende Tabelle gegeben:

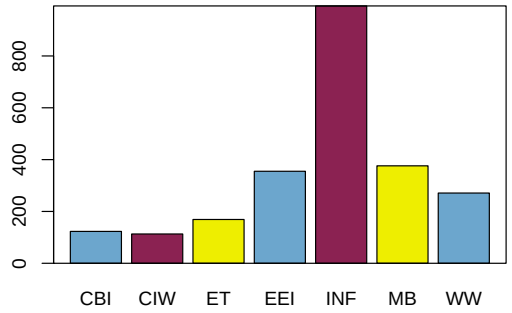
Fachrichtung	Anzahl	Prozent
CBI	123	5.13
CIW	113	4.71
ET	169	7.04
EEI	355	14.79
INF	993	41.38
MB	376	15.66
WW	271	11.29
	2400	100.00

Graphisch stellt man diese Verteilung durch ein *Histogramm* dar.

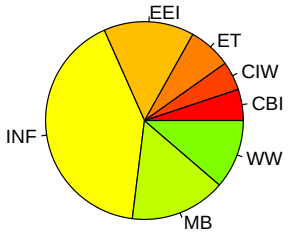


Für jede Fachrichtung wird eine Säule gezeichnet, deren Höhe der Anzahl bzw. dem Prozentsatz der Studenten in dieser Fachrichtung entspricht.
Zu beachten ist dabei:

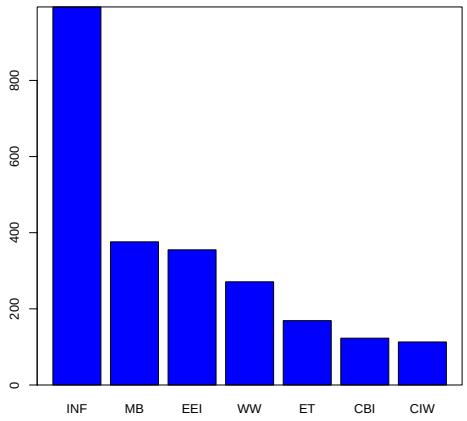
Gleiche Breite bei allen Säulen. Wie wir später noch sehen werden, wird optisch die Information nicht über die Höhe, sondern über die Fläche der Säule vermittelt. Um die Prozentsätze vergleichen zu können, dürfen die Säulen daher nicht unterschiedlich breit sein.
Gleiche Farbe für alle Säulen. Die einzige Information, die dieses Histogramm vermitteln soll, sind die Größenverhältnisse. Unterschiedliche Farben täuschen zusätzliche Informationen vor und stören beim Vergleich, wie man nachstehend sehen kann.



Tortendiagramme sind „out“: Tortendiagramme wie das nachstehende sind in wissenschaftlichen Arbeiten verpönt. Bei mehr als drei Ausprägungen sind die Sektoren bezüglich ihrer Größe kaum noch vergleichbar.



Schließlich sollte man, wenn nicht wie bei den folgenden Beispielen andere Gründe dagegen sprechen, die Säulen der Größe nach ordnen.



1.2.2 Ordinalzahlen

Ordinazahlen wie Noten, Klausurpunkte, Scores usw. sind eigentlich qualitative Merkmale mit Zahlen als Bezeichner. Sie sind aber zusätzlich mit einer Ordnungsstruktur behaftet:

- Je größer der Notenwert, desto schlechter das Ergebnis.
- Je größer die Punktezahl, desto besser das Ergebnis.

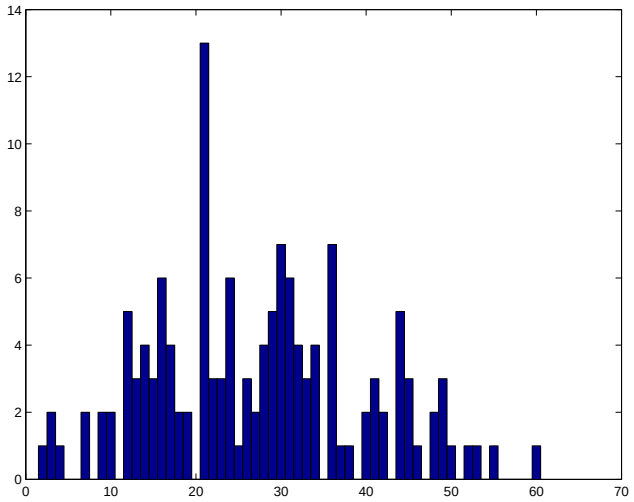
Als Beispiel zur graphischen Darstellung wählen wir die Ergebnisse einer Klausur, bei der maximal 60 Punkte erzielt werden konnten.

Teiln.	Punkte											
1–12	21	36	24	26	13	48	40	13	32	12	34	44
13–24	36	24	21	45	28	29	36	14	31	38	16	52
25–36	33	10	49	55	44	27	29	37	41	31	49	9
37–48	21	44	25	21	21	31	18	32	41	19	30	23
49–60	15	46	16	7	21	29	17	2	24	16	29	44
61–72	3	26	48	24	30	33	14	34	36	22	34	21
73–84	17	28	31	15	22	21	22	45	27	28	14	33
85–96	12	12	17	21	44	40	23	9	21	16	15	4
97–108	4	36	41	7	32	18	42	53	42	31	30	21
109–120	30	60	50	21	31	24	16	12	30	30	16	34
121–132	12	23	14	49	10	24	3	36	30	36	17	26
133–138	21	45	28	29	13	32						

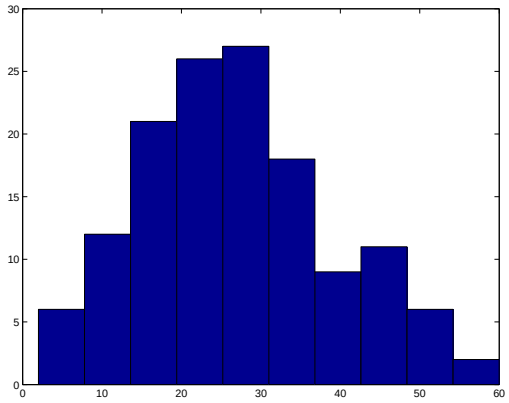
Zur graphischen Darstellung des Gesamtergebnisses stellen wir die Häufigkeitsverteilung der Klausurteilnehmer über die möglichen Punkte auf.

Pkte	Anz	Pkte	Anz	Pkte	Anz	Pkte	Anz	Pkte	Anz
1	0	13	3	25	1	37	1	49	3
2	1	14	4	26	3	38	1	50	1
3	2	15	3	27	2	39	0	51	0
4	1	16	6	28	4	40	2	52	1
5	0	17	4	29	5	41	3	53	1
6	0	18	2	30	7	42	2	54	0
7	2	19	2	31	6	43	0	55	1
8	0	20	0	32	4	44	5	56	0
9	2	21	13	33	3	45	3	57	0
10	2	22	3	34	4	46	1	58	0
11	0	23	3	35	0	47	0	59	0
12	5	24	6	36	7	48	2	60	1

Daraus ergibt sich das folgende Histogramm:



Ein „zerklüftetes“ Histogramm sieht nicht gut aus. Die Faustregel lautet, dass man möglichst nicht mehr als 8 Säulen verwenden sollte. Deshalb fasst man benachbarte Merkmalswerte zu Klassen gleicher Breite(!) zusammen. In unserem Fall ist die naheliegende Zusammenfassung die zu Notenstufen.



1.2.3 Kardinalzahlen

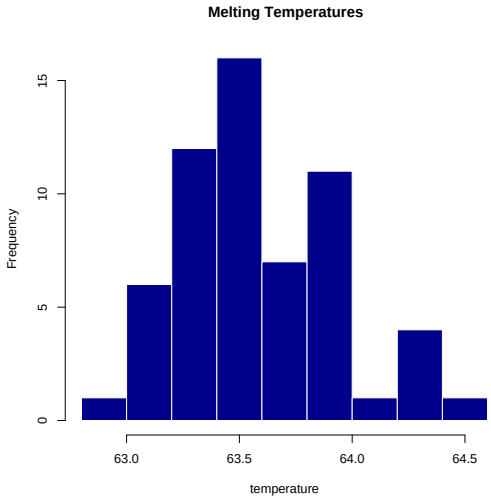
Unter Kardinalzahlen verstehen wir Messwerte in physikalischen Einheiten, wie z.B. Größe, Alter, Gewicht, und allgemein Werte, die *im Prinzip* beliebig viele Dezimalstellen haben können.
Als Beispiel betrachten wir die Schmelzpunkte von 59 Proben einer bestimmten Wachssorte, deren Schmelzpunkt theoretisch bei 63.5 Grad liegen sollte. Wie bei allen physikalischen Experimenten kann man auch hier nicht exakt messen, der zu messende Wert wird durch experimentell bedingte zufällig variierende Fehler überlagert.

63.78	63.45	63.58	63.08	63.40	64.42	63.27	63.10
63.34	63.50	63.83	63.63	63.27	63.30	63.83	63.50
63.36	63.86	63.34	63.92	63.88	63.36	63.36	63.51
63.51	63.84	64.27	63.50	63.56	63.39	63.78	63.92
63.92	63.56	63.43	64.21	64.24	64.12	63.92	63.53
63.50	63.30	63.86	63.93	63.43	64.40	63.61	63.03
63.68	63.13	63.41	63.60	63.13	63.69	63.05	62.85
63.31	63.66	63.60					

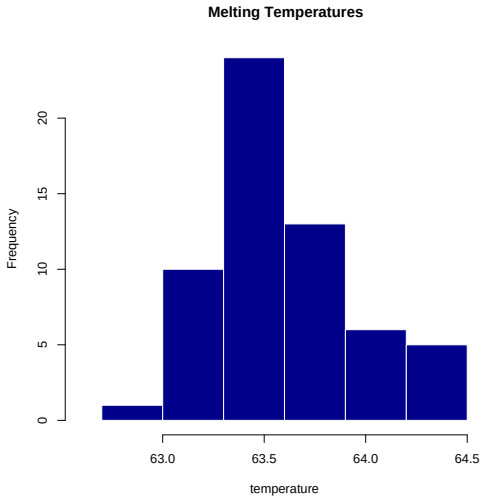
Zur Darstellung als Histogramm wird der Wertebereich in Intervalle (*Klassen*) eingeteilt und die Anzahl der Messwerte, die in den Intervallen liegen, bestimmt. Wenn man den Bereich von 62.0 bis 64.4 Grad in Intervalle der Länge 0.2 Grad einteilt, ergibt sich die folgende Häufigkeitsverteilung:

Klasse	Anzahl	Prozent
62.8–63.0	1	1.69
63.0–63.2	6	10.17
63.2–63.4	12	20.34
63.4–63.6	16	27.12
63.6–63.8	7	11.86
63.8–64.0	11	18.64
64.0–64.2	1	1.69
64.2–64.4	5	8.48
	59	100.00

und daraus das Histogramm



Hier ist durch eine ungünstige Wahl der Klassengrenzen ein falscher visueller Eindruck entstehen. Nach dem Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung sollte bei der Wiederholung immer des gleichen Experiments das Histogramm Glockengestalt besitzen, d.h. nach beiden Seiten einigermaßen gleichmäßig abfallen. Abweichungen wie oben sind entweder darauf zurückzuführen, das am experimentellen Aufbau zwischenzeitlich etwas geändert wurde oder häufiger dadurch, dass die Klassengrenzen ungünstig liegen.
Bei einer Einteilung des Wertebereichs in Klassen **gleicher** Breite 0.3 ergibt sich das Histogramm

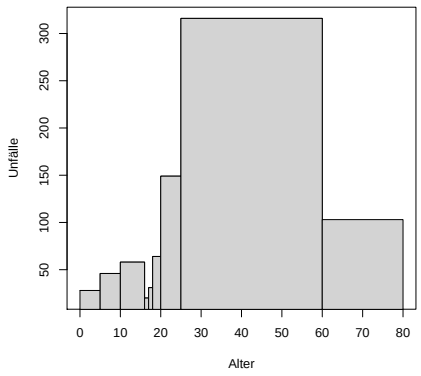


das den Erwartungen besser entspricht.

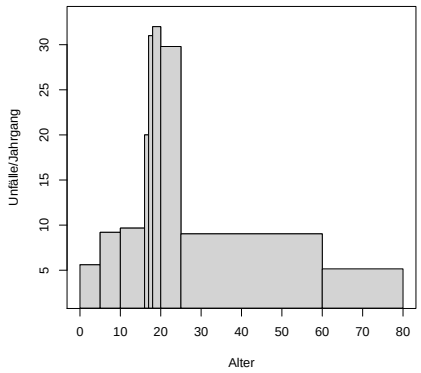
Unterschiedlichen Klassenbreiten
Als Beispiel betrachten wir eine Tabelle der Unfallhäufigkeiten an einer bestimmten Straßenkreuzung nach verschiedenen Altersklassen.

Altersgruppe	Häufigkeit
0 – 4	28
5 – 9	46
10 – 15	58
16	20
17	31
18 – 19	64
20 – 24	149
25 – 59	316
60 – 80	103

Die Klassen sind in diesem Fall vorgegeben. Wenn man jetzt auf der Altersskala über den Klassen einfach Säulen mit den Unfallhäufigkeiten zeichnet, erhält man das folgende Histogramm:



Dadurch wird optisch der Eindruck vermittelt, dass z.B. jeder einzelne Jahrgang zwischen 25 und 59 mit je 316 Unfällen vertreten ist. Um optisch die richtigen Verhältnisse aufzuzeigen, muss man Unfälle pro Jahrgang betrachten, d.h. die Unfallzahlen durch die Klassenbreite dividieren.



Wichtig bei unterschiedlichen Klassenbreiten!

- Nicht die Höhe sondern die *Fläche* einer Säule vermittelt optisch den richtigen Eindruck
- Höhe $= \dfrac{\text{Prozentsatz oder Häufigkeit}}{\text{Klassenbreite}}$
- Eine vertikale Skala ist dann u.U. sinnlos. Alternativ könnten die Säulen mit den Prozentsätzen beschriftet werden.

1.3 Die Momente

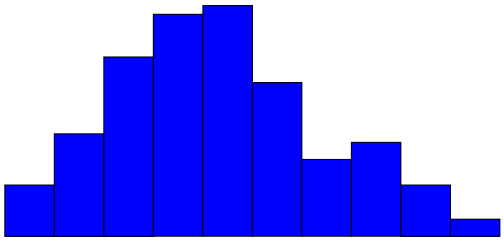
Zur mathematisch-statistischen Analyse von Messreihen, also Ausprägungen eines *quantitativen* Merkmals, sind Histogramme ungeeignet. Man benötigt dazu Kennzahlen, die eine solche Messreihe grob charakterisieren. Die wichtigsten Kennzahlen sind die *Momente* und darunter im wesentlichen nur die folgenden beiden:

- Das erste Moment oder der *Mittelwert*
- Das zweite zentrale Moment, die *Varianz*, bzw. ihre Quadratwurzel, die *Standardabweichung*

Zur Veranschaulichung des Mittelwerts und der Standardabweichung einer Messreihe

$$x_1, x_2, x_3, \ldots, x_{n-1}, x_n$$

verwenden wir aber doch ihr Histogramm

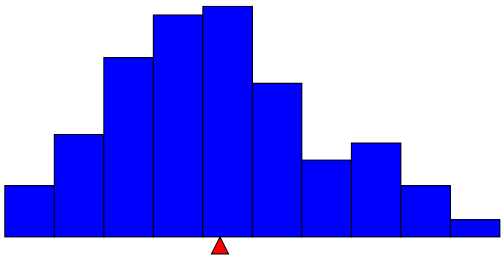


Als grobe Charakterisierung kann man sagen:

- Der *Mittelwert* entspricht dem Zentrum oder dem Schwerpunkt des Histogramms
- Die *Standardabweichung* ist ein Mass die Streubreite des Histogramms

1.3.1 Der Mittelwert

Stellt man sich das Histogramm als eine Reihe von Holzklötzen vor, die auf einem Brett stehen, so ist der Mittelwert der Schwerpunkt, d.h. die Stelle, an der man es unterstützen muss, damit es in der Schwebe bleibt.



Für eine Anzahl n von Messwerten

$$x_1, x_2, x_3, \ldots, x_{n-1}, x_n$$

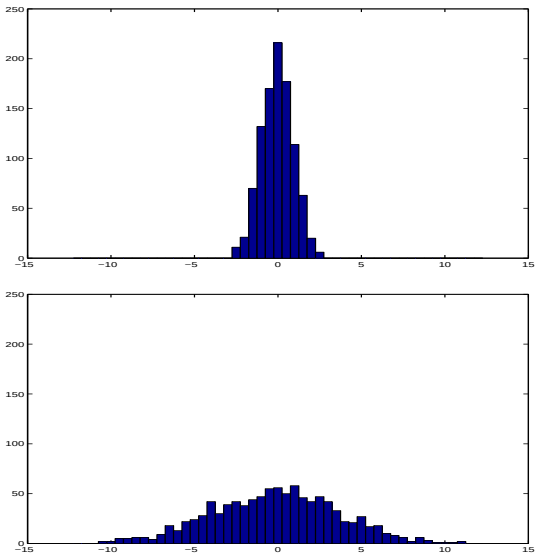
lautet die Formel für den Mittelwert (*mean value*)

$$\text{Mittelwert} = \frac{\text{Summe der Messwerte}}{\text{Anzahl der Messwerte}}$$

d.h.

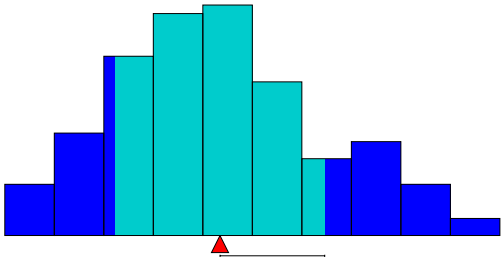
$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_{n-1} + x_n}{n}$$

Der Mittelwert alleine ist nicht ausreichend zur Charakterisierung. Die folgenden beiden Histogramme besitzen den gleichen Mittelwert 0.



1.3.2 Die Standardabweichung

Die Standardabweichung (*standard deviation*) ist ein Mass dafür, wie weit die Messwerte um den Mittelwert streuen.
In erster Näherung entspricht sie der Strecke vom Mittelwert bis zum Rand der hellblauen Fläche im nachstehenden Histogramm, wenn die hellblaue Fläche etwa 2/3 der Gesamtfläche des Histogramms ausmacht.



Die Formel für die Standardabweichung lautet

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

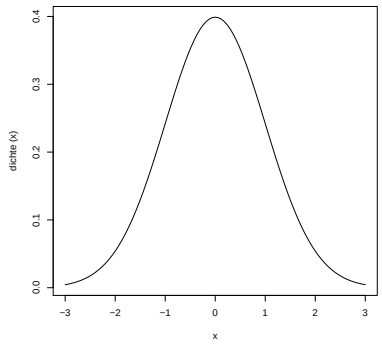
oder

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{k=1}^{n} x_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)^2\right)}$$

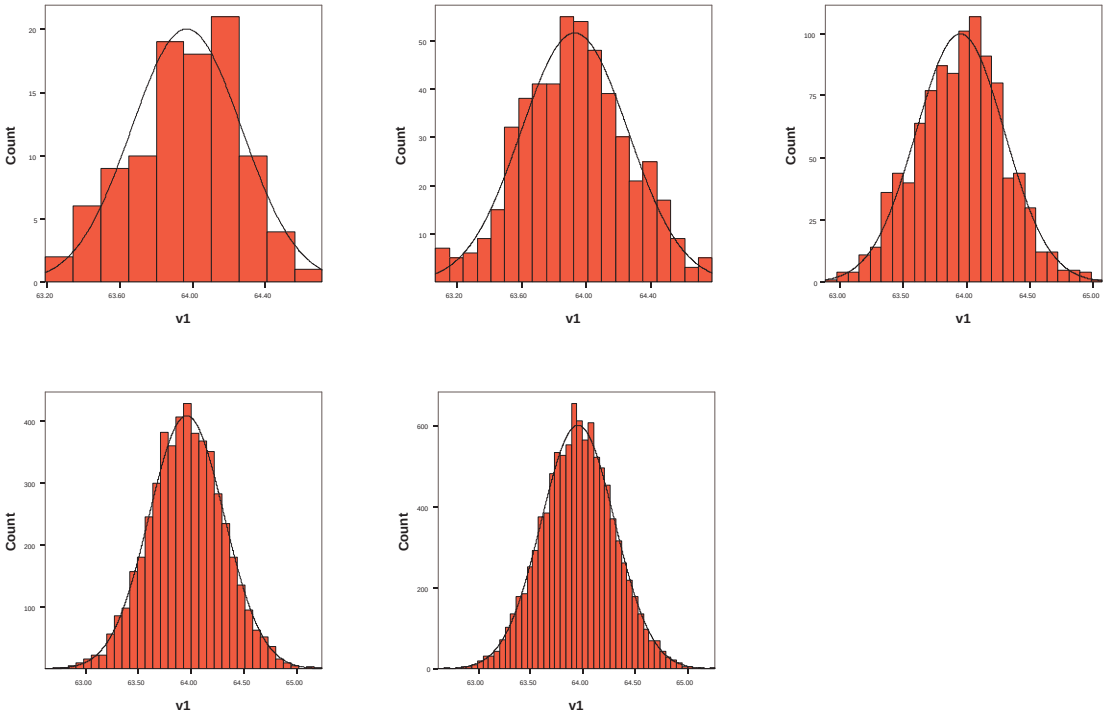
1.4 Die Gauss-Kurve

Die Bedeutung von Mittelwert und Standardabweichung folgt aus dem zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung.
Wiederholt man ein Experiment wie z.B. die Messung des Schmelzpunkts „bis ins Unendliche“, so nähern sich die auf Fläche 1 normierten Histogramme immer mehr der sog. Gauss-Kurve an.

$$f(x) = \frac{1}{\sqrt{2\pi}s}\mathrm{e}^{-\frac{(x-\overline{x})^2}{2s^2}}$$



Eine Simulation ergibt z.B. die folgende Entwicklung mit steigender Anzahl von Messungen



Die Gauss-Kurve

$$f(x) = \frac{1}{\sqrt{2\pi}s} \mathrm{e}^{-\frac{(x-\overline{x})^2}{2s^2}}$$

hängt aber nur vom Mittelwert $\overline{x}$ und von der Standardabweichung s der Messreihe ab. D.h. bei langen Messreihen benötigt man zur statistischen Analyse nur diese beiden Kenngrößen.

2 Schließende Statistik

Die schließende oder induktive Statistik beschäftigt sich mit dem „Schluss von der Stichprobe auf die Grundgesamtheit“.
Wir befassen uns hier mit dem kleinen Teilaspekt der Analyse von Messreihen, also quantitativen Merkmalen. Der Begriff Messreihe soll besagen, dass Messwerte aus der mehrfachen Wiederholung eines Experiments betrachtet werden, die von zufälligen „Störungen“ überlagert sind.
Die beiden wichtigsten statistischen Methoden, die bei wissenschaftlichen Arbeiten Anwendung finden, sind dabei *Konfidenzintervalle* und *Signifikanztests*, die wir im folgenden erläutern.

2.1 Beispiele

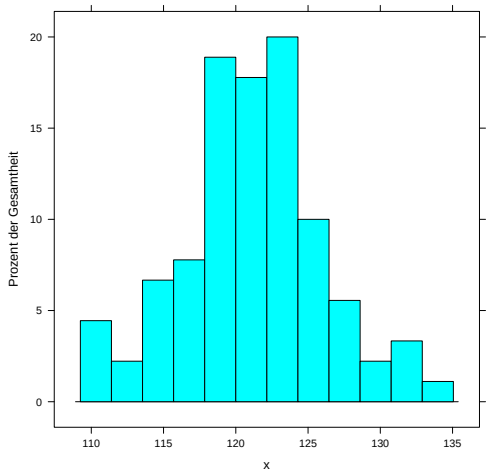
Zur Veranschaulichung verwenden wir die folgenden beiden Beispiele.
Beispiel 1: Messwerte der Ozonkonzentration in Mikrogramm pro Kubikmeter von verschiedenen Messstellen einer bestimmten Region:

127.1	120.8	122.8	110.3	113.5	124.0	119.3	122.2	119.0	110.2
122.9	125.4	123.2	124.0	124.6	125.2	114.9	124.8	128.9	122.2
123.4	119.8	125.8	111.1	123.5	117.4	121.9	120.4	119.7	119.0
118.0	120.8	114.1	118.8	116.1	120.1	117.3	124.3	121.0	117.1
111.1	119.1	121.6	123.4	117.7	122.8	121.7	114.9	114.6	124.2
121.7	118.1	118.1	118.8	126.2	123.5	121.5	126.8	127.1	121.1
116.2	131.4	118.9	124.7	118.4	121.0	117.1	131.0	114.8	121.0
121.4	132.0	124.2	120.3	121.9	122.4	119.6	124.4	112.2	122.9
134.1	115.5	119.2	118.7	125.4	129.0	127.3	119.7	127.5	122.8

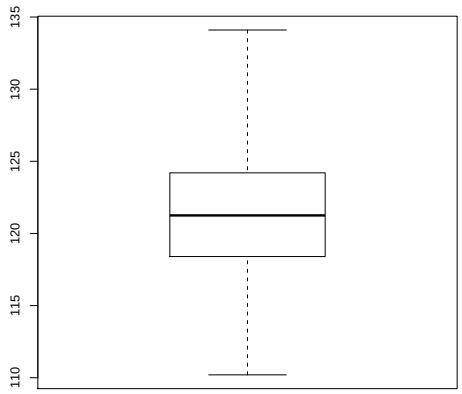
Als statistisches Arbeitsmodell gehen wir davon aus, dass es eine *mittlere Ozonkonzentration* μ der Region gibt und dass bei den Messwerten dieser Wert μ durch zufällig variierende „Messfehler“ überlagert ist.
Typische Fragen in diesem Zusammenhang sind etwa

- Kann man μ aus diesen Messwerten berechnen?
- Kann man aus diesen Messwerten schließen, dass der EU-Grenzwert von $120 \mu g/m^3$ überschritten ist?

Zur Beantwortung veranschaulicht man sich die Situation graphisch entweder anhand eines Histogramms



oder anhand eines sogenannten Boxplots

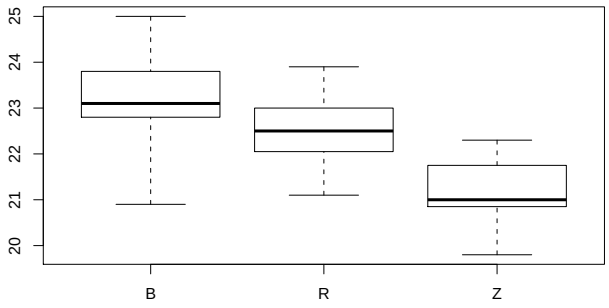


Erläuterung des Boxplots: Die Box enthält die mittleren 50% der Messwerte. Ober- und unterhalb liegen jeweils 25%. Der dicke Strich markiert den *Median*: Oberhalb und unterhalb liegen jeweils 50% der Messwerte.
Die Mehrzahl der gemessenen Werte liegt oberhalb von 120. Ob das allerdings für die Behauptung ausreicht, dass der Grenzwert überschritten ist, kann mit diesen Mitteln nicht entschieden werden.

Beispiel 2: Bei *O. H. Latter, Biometrika 1, 1901* findet sich die folgende Tabelle der Längen [mm] von Kuckuckseiern, die in den Nestern dreier Vogelarten gefunden wurden (B: Braune Grasmücke, R: Rotkehlchen, Z: Zaunkönig)

B	**R**	**Z**
22.0	21.8	19.8
23.9	23.0	22.1
20.9	23.3	21.5
23.8	22.4	20.9
25.0	22.4	22.0
24.0	23.0	21.0
21.7	23.0	22.3
23.8	23.0	21.0
22.8	23.9	20.3
23.1	22.3	20.9
23.1	22.0	22.0
23.5	22.6	20.0
23.0	22.0	20.8
23.0	22.1	21.2
	21.1	21.0
	23.0	

Im Vergleich der Boxplots



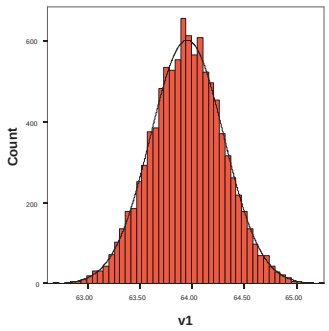
unterscheiden sich die Werte bei den verschiedenen Vogelarten optisch deutlich. Es stellt sich die Frage, ob diese Unterschiede statistisch nachweisbar sind.

2.2 Das statistische Arbeitsmodell

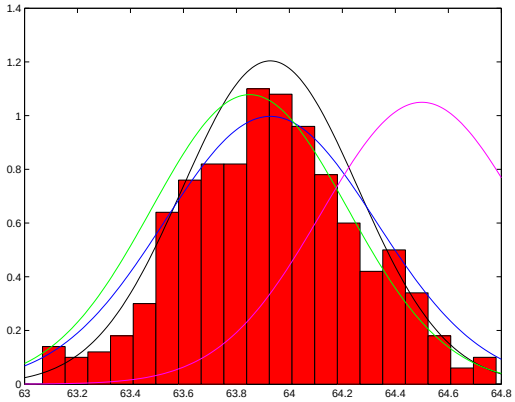
Am Beispiel der Ozonmesswerte erläutern wir kurz das theoretische Arbeitsmodell, das den statistischen Analysemethoden zugrundeliegt.
Gegeben ist eine *Statistische Grundgesamtheit* (*die Region*), die sich in einem *unbekannten Zustand* (*die Ozonkonzentration* μ) befindet.
Um Information über diesen Zustand zu erhalten, wird *eine Stichprobe vom Umfang* n *entnommen*, d.h. es werden n Messungen mit Ergebnissen $x_1, x_2, \ldots, x_n$ durchgeführt, wobei

$$x_k = \mu + e_k$$

mit zufälligen (unbekannten) Abweichungen e_k von μ.
(*Mit* $x = (x_1, x_2, \ldots, x_n)$ *bezeichnen wir die Liste aller dieser Messwerte.*)
Um Schlüsse ziehen zu können, benötigen wir noch einen Zusammenhang zwischen Messwerten und Zustand. In der Statistik ist das ein Wahrscheinlichkeitsgesetz, die sog. *Verteilungsannahme*.
Gemäß den Grenzwertbetrachtungen am Ende der letzten Vorlesung würde man mit gegen Unendlich wachsender Anzahl von Messungen als Histogramm eine Gausskurve erhalten, deren Mittelwert μ, die Ozonkonzentration, und deren Standardabweichung σ unbekannt sind.



Es steht aber nur eine *Stichprobe*, d.h. eine begrenzte Anzahl von zufällig gewonnenen Messwerten zur Verfügung. Ihr Histogramm weicht daher mehr oder weniger von der Gausskurve der Grundgesamtheit ab.



Man kann aus dem Histogramm der Messwerte also **keine sicheren Rückschlüsse** auf die Grundgesamtheit ziehen. Je weiter eine Gausskurve vom Histogramm der Messwerte entfernt liegt, desto unwahrscheinlicher ist es, dass diese Kurve die Grundgesamtheit beschreibt. Die Wahrscheinlichkeit für die Größe der Abweichung kann man formelmäßig beschreiben: Man kann dabei davon ausgehen, dass die Fehler e_k wenigstens näherungsweise normalverteilt sind.

2.3 Konfidenzintervalle

Das hat zur Folge, dass aus den Messwerten keine definitive Aussage über die Lage des Mittelwerts μ gewonnen werden kann. Eine Angabe wie

„μ liegt zwischen 119.5 und 121.0“

ist **nur mit einer bestimmten Wahrscheinlichkeit** richtig.

Man kann sich daher nur eine Sicherheitswahrscheinlichkeit γ vorgeben und unter dieser Voraussetzung einen Bereich ausrechnen, der die unbekannte Größe μ mit mindestens dieser Wahrscheinlichkeit enthält.
Die mathematische Definition dazu lautet:
Ein **Konfidenzintervall** oder **Konfidenzbereich** zur **Konfidenzzahl** γ ist ein aus den Messwerten berechneter Zahlenbereich $[T_1(x), T_2(x)]$ mit der Eigenschaft

$$P_{(\mu,\sigma)}(T_1(x) \leq \mu \leq T_2(x)) \geq \gamma$$

D.h. die Wahrscheinlichkeit, dass der berechnete Zahlenbereich die Größe μ enthält, ist mindestens gleich γ.

Für die Wahl der Konfidenzzahl bei konkreten Berechnungen haben sich bestimmte Konventionen eingebürgert. In der Medizin wählt man üblicherweise $\gamma = 0.95$.
Mit $\gamma = 0.95$ erhält man für unsere Ozon-Messwerte das Konfidenzintervall $[120.15, 122.16]$, also:

„Mit 95%-iger Wahrscheinlichkeit liegt μ zwischen 120.15 und 122.16“

Erhöht man die Sicherheitswahrscheinlichkeit, ohne zusätzliche Informationen wie zum Beispiel weitere Messungen zu berücksichtigen, wird die Aussage einfach etwas vorsichtiger:
Mit $\gamma = 0.9$ erhält man das Konfidenzintervall $[119.8185, 122.4904]$.

2.4 Tests

Die Frage aus Beispiel 1, ob der Grenzwert von $\mu_0 = 120\mu g/m^3$ für die Ozonkonzentration μ überschritten ist, lässt sich bereits durch die Berechnung des Konfidenzintervalls beantworten: Mit 95%-iger Wahrscheinlichkeit liegt μ zwischen 120.15 und 122.16, also oberhalb von 120, wenn auch nur knapp.
Üblicherweise wendet man aber zur ja/nein-Beantwortung von Fragen in der Statistik einen *Test* an. Zur Frage „Überschreitet die Ozonkonzentration μ den EU-Grenzwert von $\mu_0 = 120\mu g/m^3$?“ gibt es zwei mögliche Aussagen (Hypothesen) als Antwort:

$$H : \mu > \mu_0 \quad \text{und} \quad \overline{H} : \mu \leq \mu_0$$

Um diese Frage auf der Basis von Messwerten zu entscheiden, gibt man sich **vor** der Durchführung der Messungen eine Teilmenge $\mathcal{C}$ aus der Menge der möglichen Messergebnisse vor mit der Regel

Ist das tatsächliche $x \in \mathcal{C}$, so wird entschieden, dass H richtig ist,

andernfalls wird $\overline{H}$ als richtig angesehen.

Aufgrund des Zufallscharakters unserer Messungen sind Fehlentscheidungen unvermeidlich. Man kann nur versuchen, die Wahrscheinlichkeit für Fehlentscheidungen unterhalb vorgegebener Schranken zu halten.

Aber man kann die Wahrscheinlichkeit für **beide** möglichen Fehlentscheidungen **nicht gleichzeitig** nach unten drücken. Man muss auswählen, bei welcher der beiden Hypothesen die Entscheidung bis auf eine vorgegebene Irrtumswahrscheinlichkeit abgesichert sein soll.
Diese Hypothese nennt man die *Alternative* und bezeichnet sie mit A. Ihr Gegenteil heißt die *Nullhypothese*. Sie erhält das Symbol H_0.
Beispiel: Im Fall der Ozonkonzentration will die Stadtverwaltung sicher sein, dass der Grenzwert überschritten ist, bevor sie Verkehrsbeschränkungen erlässt, also: $A : \mu > \mu_0$.
Die möglichen Fehlentscheidungen haben dadurch unterschiedliche Gewichte und heißen

- Fehler 1. Art: Entscheidung auf A, obwohl H_0 zutrifft.
- Fehler 2. Art: Entscheidung auf H_0, obwohl A zutrifft.

Ein Test, bei dem die Wahrscheinlichkeit für den Fehler 1. Art unterhalb einer vorgegebenen Schranke liegt, heißt ein *Signifikanztest*. Die mathematische Definition lautet

> Ein **Signifikanztest** für die Hypothesen H_0 und A zur Signifikanzzahl α ist ein Entscheidungsverfahren $D(x)$ mit der Eigenschaft
>
> $$P_{H_0}(D(x) = A) \leq \alpha$$

Konventionen und Sprechweisen

- Standardwert für die Signifikanzzahl ist in der Medizin $\alpha = 0{,}05$.
- Ist $D(x) = A$, so heißt **die Alternative** A **signifikant nachgewiesen** oder **die Nullhypothese widerlegt**. (Die Wahrscheinlichkeit für einen Fehler 1. Art ist höchstens α)
- Ist $D(x) = H_0$, so sagt man, dass **die Nullhypothese** H_0 **angenommen wird**, bzw. die Alternative nicht nachgewiesen werden kann. (Die Wahrscheinlichkeit für einen Fehler 2. Art ist nicht bekannt)

Der p-Wert

Gibt man unsere Ozonmesswerte x und den Grenzwert 120 in das Statistik-Programm mit R ein (nach einer Signifikanzzahl wird man nicht gefragt):

```
> t.test(x, mu=120, alternative="greater")
```

so erhält man als Antwort

```
        One Sample t-test

data:  x
t = 2.2746, df = 89, p-value = 0.01267
alternative hypothesis: true mean is greater than 120
```

Die entscheidende Größe ist dabei der *p-value* oder *p-Wert*. Mathematisch nicht ganz korrekt formuliert ist der p-Wert (p-value) $p(x)$ die Irrtumswahrscheinlichkeit bei der Entscheidung auf A, die mit dem Ergebnis x des Experiments verbunden ist.

- Ist $p(x) \leq \alpha$, so gilt A als signifikant nachgewiesen.
- Ist $p(x) > \alpha$, so wird die Nullhypothese angenommen.

Bei der Veröffentlichung eines Testergebnisses wird grundsätzlich der p-Wert angegeben. Damit kann jeder Leser selbst entscheiden, wie er angesichts seiner bevorzugten Signifikanzzahl das Ergebnis bewerten soll.
Der p-Wert 0.01267 in unserem Beispiel besagt gerade: Bezüglich $\alpha = 0.05$ wäre die Alternative signifikant nachgewiesen, bezüglich $\alpha = 0.01$ nicht.

2.5 Zweistichprobentests

Eine der häufigsten Fragestellungen ist, ob sich zwei Messreihen signifikant unterscheiden oder ob die Unterschiede rein zufallsbedingt sind.
Gegeben seien zwei Stichproben

$$x = (x_1, x_2, \ldots, x_m) \quad \text{und} \quad y = (y_1, y_2, \ldots, y_m)$$

wobei x aus einer Grundgesamtheit mit Mittelwert μ_x und y aus einer mit Mittelwert μ_y stammt.
Zu entscheiden ist, ob die Mittelwerte gleich oder verschieden sind: $\mu_x = \mu_y$ oder $\mu_x \neq \mu_y$.
Man kann auf der Basis zufallsfehlerbehafteter Messwerte niemals „nachweisen“, dass die dahinterstehenden Parameter gleich sind, daher:

$$H_0 : \mu_x = \mu_y \quad \text{und} \quad A : \mu_x \neq \mu_y$$

Beispiel Wir vergleichen die Kuckuckseierlängen-Messreihen von Rotkehlchen und Zaunkönig aus Beispiel 2:

Rotkehlchen	21.8	23.0	23.3	22.4	22.4	23.0	23.0	23.0
	23.9	22.3	22.0	22.6	22.0	22.1	21.1	23.0
Zaunkönig	19.8	22.1	21.5	20.9	22.0	21.0	22.3	21.0
	20.3	20.9	22.0	20.0	20.8	21.2	21.0	

und fragen: „Gibt es hinsichtlich der Eigröße **signifikante Unterschiede**“ zwischen den zwei Vogelarten?
Die Visualisierung durch Boxplots zeigte, dass offensichtlich Unterschiede bestehen.

Zweistichprobentests

Zum statistischen Nachweis von Unterschieden gibt es zwei Testvarianten.

- Der Studentsche t-Test, Der klassische Test für den Vergleich von Mittelwerten. Er setzt aber voraus, dass die Standardabweichungen der beiden Grundgesamtheiten gleich sind: $\sigma_x = \sigma_y = \sigma$. (Sie müssen aber nicht bekannt sein).
- Der Zweistichprobentest von Welch. Dieser Test erfordert diese Voraussetzung nicht. Er ist aber ein sog. asympotischer Test, der eine größere Zahl von Messwerten erfordert und einen größeren p-Wert als der Studentsche t-Test liefert.

Mit dem Programm R ergibt der Studentsche t-Test angewandt auf unsere Beispielzahlen:

```
> t.test(x, y, mu = 0, alternative="two.sided",
         var.equal = TRUE)

        Two Sample t-test

data:  x and y
t = 5.5671, df = 29, p-value = 5.254e-06
alternative hypothesis: true difference in means
is not equal to 0
```

Und der Welch-Test

```
> t.test(x, y, mu = 0, alternative="two.sided",
         var.equal = FALSE)

        Welch Two Sample t-test

data:  x and y
t = 5.5485, df = 28.217, p-value = 6.052e-06
```

Der p-Wert ist also praktisch Null und der Unterschied signifikant nachgewiesen.

Beim Vergleich von Rotkehlchen und brauner Grasmücke

Rotkehlchen	21.8	23.0	23.3	22.4	22.4	23.0	23.0	23.0
	23.9	22.3	22.0	22.6	22.0	22.1	21.1	23.0
br. Grasmücke	22.0	23.9	20.9	23.8	25.0	24.0	21.7	
	23.8	22.8	23.1	23.1	23.5	23.0	23.0	

ergibt sich

```
> t.test(x, y, mu = 0, alternative = "two.sided")

        Welch Two Sample t-test

data:  x and y
t = -1.7, df = 21.806, p-value = 0.1033
```

Hier sind also bezüglich der üblichen Signifikanzzahlen keine Unterschiede nachweisbar.