W.11 allgemeine Erläuterungen

Stochastik ist der Oberbegriff für Statistik und Wahrscheinlichkeitsrechnung.
Die Statistik befasst sich eher mit dem Sammeln und der Auswertung von Daten. In der Statistik macht man Umfragen, überlegt natürlich wie man die Umfrage startet um repräsentative Ergebnisse zu erhalten. Dann berechnet man Durchschnitte, Standardabweichungen, etc.. um aus den Unmenge von Daten etwas herauslesen zu können. Die Häufigkeiten für die verschiedenen Fragestellungen verwendet man um Voraussagen für größere Bevölkerungsgruppen zu machen. Damit beschäftigt sich dann die Wahrscheinlichkeitsrechnung.
Die Wahrscheinlichkeitsrechnung verwendet also Daten [die die Statistik gesammelt hat] und erstellt daraus Prognosen für unbekannte bzw. zukünftige Ereignisse.

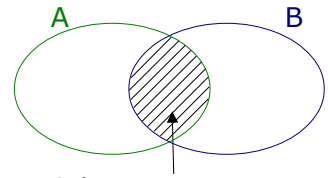
Beispiel: Sie möchten wissen, wie viele Frauen Hosen bzw. Röcke tragen.
Sie können nun alle Hundert Millionen Frauen anrufen und fragen, oder sie machen eine Umfrage [unter sagen wir mal: Tausend Frauen] und berechnen davon Durchschnitt und sonstige Daten.
Sie haben nun also eine statistische Erhebung durchgeführt.
Der Witz ist nun der, dass man automatisch davon ausgeht, dass die Durchschnittswerte, die man aus seiner Umfrage errechnet hat, für *alle* Frauen gelten, man überträgt die Werte aus der befragten Gruppe auf die ganze Bevölkerung (obwohl das ja eigentlich nicht sicher ist).
Verwendet man die Daten der Erhebung nun weiterhin für die *gesamte* Bevölkerung, ist der Sprung von der Statistik zur Wahrscheinlichkeitsrechnung vollzogen.

W.11.01 Begriffe der Stochastik (ꝕ)

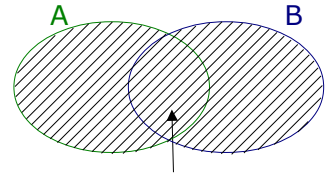
Symbole:

- Ω Grundraum, Ereignisraum, Ergebnismenge. Alle Elemente, die prinzipiell vorkommen können. [z.B. beim Würfel gilt: $\Omega=\{1;2;3;4;5;6\}$
- A Ein einzelnes gewünschtes Ereignis [=Elementarereignis] oder eine Menge von mehreren *gewünschten* Ereignissen.
- $\overline{A}$ Gegenereignis. Alle Elemente, die *nicht* in A vorkommen. Die Wahrscheinlichkeiten von A und $\overline{A}$ ergeben addiert 100%. Es gilt: $P(\overline{A})=1-P(A)$.
- $\overline{x}$ Erwartungswert [siehe →Kap.W.11.03]
- $\tilde{x}$ Median [siehe →Kap.W.11.03]
- $\cap$ Schnittmenge. Das sind die gemeinsamen Elemente beider Mengen.
- $\cup$ Vereinigung. Das alle Elemente, die in mindestens einer der beiden Mengen vorkommt.
- $\varnothing$ oder $\{\ \}$ Die leere Menge. Sie enthält natürlich kein Element und es gilt immer: $P(\varnothing) = P(\{\ \}) = 0$
- \ Differenz. A\B ist die Menge aller Elemente von A, die jedoch nicht in B enthalten sind.
- μ Erwartungswert [siehe →Kap.W.11.03]

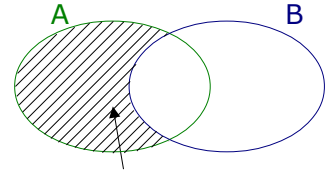
Venn-Diagramme
↓



Schnittmenge: A∩B



Vereinigung: A∪B



Differenz: A\B

σ Standardabweichung [siehe →Kap.W.11.05]

Begriffe:

„disjunkt" Zwei Mengen sind disjunkt, wenn sie keine gemeinsamen Elemente besitzen. Die Schnittmenge dieser beiden Mengen ist also leer.

„Erwartungswert" ist ein Mittelwert bzw. ein Durchschnitt. Man bezeichnet ihn mit E(x), mit μ oder mit $\bar{x}$.

„Klassen" haben in der Statistik nichts mit Schulklassen zu tun, sondern eher mit einer Gruppeneinteilung. Man kann die Klasseneinteilung einfach machen, man kann jedoch auch eine Wissenschaft daraus machen.

„Signifikanzniveau" ist das gleiche wie eine Irrtumswahrscheinlichkeit. Gehört beides zum Hypothesentest.

„Venn-Diagramme" sind Grafiken, mit denen man Zusammenhänge zwischen zwei Mengen beschreibt [siehe rechte obere Hälfte der ersten Kapitelseite]

„Zentralwert" ist ein Oberbegriff für die Begriffe: „Mittelwert=Erwartungswert" und „Median" und „Modus". [→Kap.W.11.03]

„Zufallsvariable" kann alles Mögliche sein. Eine Zufallsvariable ist in der Stochastik das, was im Alltag ein „Ding" ist. Eigentlich alles.

Schreibweisen:

Ein einzelnes Element steht in runden Klammern.	(A)
Eine Menge von mehreren Elementen wird in geschlungene Klammern geschrieben.	{(A);(B);(C);...}
Eine Wahrscheinlichkeit wird mit „P" bezeichnet. Dahinter steht das Ereignis in runder Klammer.	P(A)
Eine interessante Schreibweise entsteht z.B. wenn man die Wahrscheinlichkeit von einer Menge von mehreren Ereignissen hat.	P({(A);(B);(C);...})
Meist kürzt man diese Schreibweise jedoch „salopp" ab zu:	P(A; B; C; ...)

Eine Beziehung [die Sie vermutlich selten brauchen]: $\overline{A \cup B} = \overline{A} \cap \overline{B}$

bzw: $\overline{A \cap B} = \overline{A} \cup \overline{B}$

Bsp.1
Gegeben seinen die beiden Menge A und B mit:
A = { 1; 3; 5; 6; 7; 8 } und B = { 0; 1; 2; 4; 6; 8 }

Es gilt:
$A \cap B = \{ 1; 6; 8 \}$
$A \cup B = \{ 0; 1; 2; 3; 4; 5; 6; 7; 8 \}$
$\overline{A}$ kann man so nicht angeben. Man müsste erst wissen was Ω ist. Nehmen wir also an, Ω sei die Menge aller Zahlen von 0 bis 10. $\Rightarrow \overline{A} = \{ 0; 2; 4; 9; 10 \}$
$A \backslash B = \{ 3; 5; 7 \}$ $B \backslash A = \{ 0; 2; 4 \}$

Bsp.2
Knecht Ruprecht bereitet für den 6.Dez. zwei Säcke vor: Im ersten Sack befinden sich Äpfel, Walnüsse, Mandarinen und Haselnüsse. Im zweiten Sack befinden sich Orangen, Äpfel, getrocknete Bananen und ebenfalls Haselnüsse. Abgesehen von den genannten Früchten verfügt er noch über Pfirsiche und Kartoffeln.
Sein drittes Rentier heißt Mathilde und hat eine Vorliebe für Mengenlehre.
a) Bestimmen Sie für Mathilde die Vereinigungsmenge, die Schnittmenge und die Differenzmenge der beiden Sack-Mengen.
b) Sei S_1 der erste Sack und S_2 der zweite Sack.
Bestimmen Sie $\overline{S_1}$, $\overline{S_2}$, $\overline{S_1} \backslash S_2$, $S_1 \cup \overline{S_2}$ sowie $\overline{S_1 \cap S_2}$.

Lösung:
[Wir kürzen alle Früchte mit ihren Anfangsbuchstaben ab.]
$\Omega = \{Ä,B,H,K,M,O,P,W\}$ $S_1 = \{Ä,H,M,W\}$ $S_2 = \{Ä,B,H,O\}$
a) $S_1 \cup S_2 = \{Ä,B,H,M,O,W\}$ $S_1 \cap S_2 = \{Ä,H\}$ $S_1 \backslash S_2 = \{H,W\}$ $S_2 \backslash S_1 = \{B,O\}$
b) $\overline{S_1} = \{B,K,O,P\}$ $\overline{S_2} = \{K,M,P,W\}$ $\overline{S_1} \backslash S_2 = \{K,P\}$ $S_1 \cup \overline{S_2} = \{Ä,H,K,M,P,W\}$
$\overline{S_1 \cap S_2} = \{B,K,M,O,P,W\}$ [einfach auf $S_1 \cap S_2$ schauen]

W.11.02 absolute, relative und kumulierte Häufigkeit (卌)

Eine *absolute Häufigkeit* ist eine *Anzahl*. Es handelt sich daher immer um eine natürliche Zahl [0; 1; 2; …]. Als Bezeichnung verwendet man häufig den Kleinbuchstaben „h".

Eine *relative Häufigkeit* ist ein *prozentualer Anteil*, also eine *Wahrscheinlichkeit*. Es handelt sich daher immer um eine Zahl zwischen 0 und 1 [bzw. wenn sie in Prozent angegeben ist, dann zwischen 0% und 100%]. Als Bezeichnung verwendet man häufig den Kleinbuchstaben „f". Wahrscheinlichkeiten werden eigentlich immer mit „p" oder „P" bezeichnet. [Das Wort „Wahrscheinlichkeit" kürze ich in diesem Buch meist mit „***W.S.***" ab.]

Eine *kumulierte Häufigkeit* ist eine *Summe* von Häufigkeiten und heißt daher manchmal auch *Summenhäufigkeit* [sie kann eine absolute oder eine relative Häufigkeit sein]. Im Normalfall addiert man immer die Häufigkeiten von der betreffenden Zahl bis zu Null runter. Als Bezeichnung verwendet man häufig die Großbuchstaben „F" oder „H" [bei einer absoluten kumulierten Häufigkeit „H", bei einer relativen kumulierten Häufigkeit „F"].

Bsp.3

Gegeben sei eine Urne mit zwölf bunten und ganz tollen Kugeln. Fünf tragen die Nummer „1", vier tragen die Nummer „2" und drei tragen die Nummer „3".
Geben Sie die absoluten, die relativen und die zugehörigen kumulierten Häufigkeiten an.

Lösung:
Die absoluten Häufigkeiten sind eine Anzahl. Also: $h(1)=5,\ h(2)=4,\ h(3)=3$
Die relativen Häufigkeiten sind prozentuale Anteile: $f(1)=\frac{5}{12},\ f(2)=\frac{4}{12},\ f(3)=\frac{4}{12}$
Die absolute kumulierte Häufigkeiten sind:

$H(1)=h(1)=5$
$H(2)=h(1)+h(2)=5+4=9$
$H(3)=h(1)+h(2)+h(3)=5+4+3=12$

Die relative kumulierte Häufigkeiten sind:

$$F(1)=f(1)=\frac{5}{12}$$
$$F(2)=f(1)+f(2)=\frac{5}{12}+\frac{4}{12}=\frac{3}{4}$$
$$F(3)=f(1)+f(2)+f(3)=\frac{5}{12}+\frac{4}{12}+\frac{3}{12}=1$$

Bsp.4

Bei einer Umfrage unter 12400 Personen meinen 15%, maximal 25€ monatlich für Süßigkeiten auszugeben, 40% geben bis zu 50€ aus, 70% geben bis 75€ aus, der Rest gibt bis zu 100€ aus.
Geben Sie sinnvolle Klassen an, sowie deren absolute, relative und die zugehörigen kumulierten Häufigkeiten an.

Lösung:
Die Klasseneinteilung ist indirekt ja schon gegeben.
Klasse1: 0-25€; Klasse2: 25,01-50€; Klasse3: 50,01-75€; Klasse4: 75,01-100€

In der Aufgabenstellung sind die *relativen kumulierten* Häufigkeiten gegeben. [Bedenken Sie: „ 70% geben bis 75€ aus" bedeutet, dass 70% 0€-75€ ausgeben, da sind die Häufigkeiten der *drei* ersten Klassen drin].
Wir wissen also:

relative kumulierte Häufigkeiten der vier Klassen:
$F(1)=0,15$ $F(2)=0,40$ $F(3)=0,70$ $F(4)=1,00$

Die nicht kumulierten (relativen) Häufigkeiten kann man sich logisch herleiten.

f(1)=F(1)=0,15 f(2)=F(2)−F(1)=0,25
f(3)=F(3)−F(2)=0,30 f(4)=F(4)−F(3)=0,30

Die absoluten Häufigkeiten sind eine Anzahl. Da man Anzahlen erhält, in dem man die Gesamtanzahl mit den entsprechenden Häufigkeiten multipliziert, ergibt sich:

absolute Häufigkeiten:
h(1)=N·f(1)=12.400·0,15=1.860 h(3)=N·f(3)=12.400·0,30=3.720
h(2)=N·f(2)=12.400·0,25=3.100 h(4)=N·f(4)=12.400·0,30=3.720

absolute kumulierte Häufigkeiten:
H(1)=N·F(1)=12.400·0,15=1.860 H(3)=N·F(3)=12.400·0,70=8.680
H(2)=N·F(2)=12.400·0,40=4.960 H(4)=N·F(4)=12.400·1,0=12.400

W.11.03 Mittelwert, Median, Modus (∯)

Mittelwert, Median und Modus sind sogenannte Zentralwerte. Eine treffende [wenn auch mathematisch blöde] Formulierung wäre: sie liegen „in der Mitte" der Datenreihe. Je nachdem, wie man jedoch die „Mitte" definiert, verwendet man den ein oder anderen Wert.
Bei der Berechnung muss man unterscheiden, ob man:

- eine Liste von einzelnen Daten gegeben hat, oder
- ob Klassen [=Gruppen] mit deren prozentualer Häufigkeit gegeben sind

Zu beiden Fällen machen wir später Beispiele.

Der „**Mittelwert**" oder auch „Durchschnitt" oder auch „arithmetisches Mittel" ist das, was man klassisch tatsächlich unter Mittelwert versteht. Man zählt alle vorhandenen Werte zusammen und teilt durch deren Anzahl. Man bezeichnet den Mittelwert mit $\overline{x}$ oder μ. [Manchmal sieht man auch den englischen Begriff **„mean"**.] In der Wahrscheinlichkeitsrechnung verwendet man statt dessen eher den Begriff „Erwartungswert", der mit E(x) bezeichnet wird.
[Es gibt zwar minimale Unterschiede zwischen „Mittelwert" und „Erwartungswert", die sind aber für die Grundlagen der Stochastik nicht von Bedeutung].

Ist nicht die absolute Häufigkeit der einzelnen Daten gegeben, sondern nur deren prozentuale Häufigkeit, so multipliziert man jeden auftauchenden Wert mit seiner Häufigkeit und addiert diese Zwischenergebnisse.
Der Mittelwert ist der wichtigste Zentralwert.

Mittelwert:

bei Datenreihen:

$$\overline{x} = \mu = \frac{x_1+x_2+x_3+\ldots+x_n}{n}$$

bei Werten mit deren Häufigkeiten:

$$\overline{x}=\mu=x_1\cdot f(x_1)+x_2\cdot f(x_2)+\ldots$$

$x_1, x_2, x_3, \ldots$ auftauchende Werte
$f(x_1), f(x_2), \ldots$ prozentuale Häufigkeiten

bei Klassen:

$$\overline{x}=\mu=x_1\cdot f(x_1)+x_2\cdot f(x_2)+\ldots$$

$x_1, x_2, x_3, \ldots$ Klassenmitten
$f(x_1), f(x_2), \ldots$ Häufigkeiten der Klasse

Für den **„Median"** sortiert man alle gegebenen Daten der Größe nach [egal ob aufsteigend oder absteigend] und wählt nachher den Wert aus, der in der Mitte liegt.

Für den Median gibt keine gängige Bezeichnung. Eine, die man ab und zu sieht, ist: „$\tilde{x}$" oder „med".

Da der Median gleichzeitig auch das zweite Quartil ist, hat er auch die Bezeichnung: Q_2.

Hat man eine gerade Anzahl von Daten gegeben, so liegen zwei Werte in der Mitte. Von diesen beiden bildet man den Mittelwert.

Sind die Werte mit ihren prozentualen Häufigkeiten gegeben, ist die Vorgehensweise ein bisschen umständlicher. Zuerst sortiert man die Werte der Größe nach, danach berechnet man die kumulierten Häufigkeiten. Der erste Wert dessen kumulierte Häufigkeit „0,5" überschreitet, ist der Median.
[Sie werden in der Literatur hierzu leider unterschiedliche Definitionen finden].

Sind die Werte als Klassen gegeben, ist das Ganze noch ein bisschen hässlicher. Zuerst wählt man die Klasse aus, in welcher sich der Median befindet. Das geschieht indem man die Klasse der Größe nach sortiert, die kumulierten Häufigkeiten berechnet und dann diejenige Klasse auswählt, deren kum. Häuf. erstmals den Wert „0,5" erreicht. Nun setzt man die untere und obere Grenze dieser Klasse in nebenstehende Formel ein, um den Median zu erhalten.

Median:

bei Datenreihen:
alle Daten der Größe nach sortieren, danach den Wert aussuchen, der in der Mitte liegt. [Liegen zwei Werte in der Mitte, nimmt man deren Mittelwert.]

bei Werte mit deren Häufigkeiten:
alle Daten der Größe nach sortieren, danach den Wert aussuchen, dessen kumulierte W.S. erstmalig „0,5" erreicht.

bei Klassen:
Klassen der Größe nach sortieren, danach den Klasse aussuchen, dessen kumulierte W.S. erstmalig „0,5" erreicht.
Die Formel:

$$x_{un} + \frac{0{,}5 - F(x_{un})}{f(x_i)} \cdot (x_{ob} - x_{un})$$

liefert den Median

$F(x_{un})$: kumulierte Häufigkeit der unteren(!) [=vorhergehenden] Klasse.
$f(x_i)$: Häufigkeit der ausgesuchten (interessanten) Klasse.
x_{un}: untere Grenze der interessanten Klasse
x_{ob}: obere Grenze der interessanten Klasse

Der **„Modus"** oder „Modalwert" ist einfach nur der Wert, der am häufigsten auftaucht.

Bei Klassen oder bei gegebenen prozentualen Häufigkeiten ist es der Wert mit der größten prozentualen Häufigkeit.

Hat eine Datenreihe mehr als einen Modus, heißt sie „bimodal". [Wir gehen nicht weiter darauf ein.]

Modus:

bei Datenreihen:
derjenige Wert, der am häufigsten auftaucht.

bei Werte mit deren Häufigkeiten:
der Wert, der die größte Häufigkeit aufweist.

bei Klassen:
die Klassenmitte der Klasse, die die höchste Häufigkeit aufweist.

Bsp.5 (Datenreihe)
Die Notaufnahme einer Klinik untersucht 15 Nächte lang die Anzahl der eingehenden Notfälle.
Folgende Anzahl von Notfällen werden notiert:
3, 7, 5, 5, 1, 5, 2, 1, 3, 6, 5, 4, 2, 0, 2

a) Bestimmen Sie die durchschnittliche Anzahl von Notfällen je Nacht.
b) Was ist der Median der Datenreihe?
c) Geben Sie den Modus an.

Lösung:
a) Für den Durchschnitt zählt man alle Anzahlen zusammen und teilt durch die Anzahl. $\bar{x} = \mu = \frac{3+7+5+5+1+5+2+1+3+6+5+4+2+0+2}{15} = 3{,}4$
b) Für den Median sortieren wir erst die Daten der Größe nach:
0, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 5, 6, 7
Der mittlere Wert, hier der achte Wert.
Der Median ist der Wert „3".
c) Der Modus ist der Wert, der am häufigsten auftaucht. In diesem Fall taucht die „5" am häufigsten auf [vier Mal]. Der Modus ist „5".

Bsp.6 (Werte mit prozentualen Anteilen)
Familie Günther besitzt eine kleine Farm. 20% der Tiere sind Ziegen, 35% sind Schafe und 45% sind Hunde. Eine Ziege wiegt 40kg, ein Schaf 30kg und ein Hund wiegt 20kg.
a) Bestimmen Sie das durchschnittliche Gewicht aller Farmtiere.
b) Bestimmen Sie den Median an.
c) Geben Sie den Modus an.

Lösung:
a) Es geht um das durchschnittliche *Gewicht* der Tiere, daher sind die Werte, um die es geht, $x_1=20$, $x_2=30$, $x_3=40$ [ich habe die Werte gleich der Größe nach sortiert].
Die Häufigkeiten davon sind: $f(x_1)=0{,}45$ $f(x_2)=0{,}35$ $f(x_3)=0{,}20$.
$\bar{x} = \mu = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + x_3 \cdot f(x_3) = 20 \cdot 0{,}45 + 30 \cdot 0{,}35 + 40 \cdot 0{,}20 = 27{,}5\text{kg}$.
b) Für den Median benötigen wir die kumulierte Häufigkeit.
$x_1=20$: $f(20)=0{,}45$ ⇒ $F(20)=0{,}45$
$x_2=30$: $f(30)=0{,}35$ ⇒ $F(30)=0{,}45+0{,}35=0{,}8$
$x_3=40$: $f(40)=0{,}20$ ⇒ $F(40)=0{,}45+0{,}35+0{,}2=1$
Der erste Wert, dessen kumulierte Häufigkeit zum ersten Mal „0,5" erreicht [und überschreitet], ist der zweite Wert: $x_2=30$. ⇒ Der Median ist „30".
c) Der Modus ist der Wert mit größten Häufigkeit. Die größte Häufigkeit ist 0,45 und gehört zum Wert $x_1=20$. Der Modus ist „20".

Bsp.7 (Klassen)
Eine unglaublich wichtige biologische Untersuchung fördert zu Tage, dass 20% der Maikäfer maximal 0,6 Gramm wiegen, 25% der Maikäfer wiegen 0,6–0,8 Gramm, 40% wiegen 0,8–1,1 Gramm, der Rest wiegt 1,1–1,5 Gramm.
a) Bestimmen Sie den Durchschnitt des Gewichts.
b) Bestimmen Sie den Median der Maikäfer-Daten.
c) Geben Sie den Modalwert an.

Lösung:
Wir haben vier Klassen:

Nr.1:	0–0,6g	Klassenmitte: $x_1=0,3g$	Häufigkeit: $f(x_1)=0,20$	kumulierte Häuf. $F(x_1)=0,20$
Nr.2:	0,6–0,8g	Klassenmitte: $x_2=0,7g$	Häufigkeit: $f(x_2)=0,25$	kumulierte Häuf. $F(x_2)=0,45$
Nr.3:	0,8–1,1g	Klassenmitte: $x_3=0,95g$	Häufigkeit: $f(x_3)=0,40$	kumulierte Häuf. $F(x_3)=0,85$
Nr.4:	1,1–1,5g	Klassenmitte: $x_4=1,3g$	Häufigkeit: $f(x_4)=0,15$	kumulierte Häuf. $F(x_4)=1,00$

a) Die Maikäfer wiegen durchschnittlich:
$\bar{x} = \mu = x_1 \cdot f(x_1)+x_2 \cdot f(x_2)+x_3 \cdot f(x_3)+x_4 \cdot f(x_4) =$
$= 0,3 \cdot 0,20+0,7 \cdot 0,25+0,95 \cdot 0,40+1,3 \cdot 0,15 = 0,81$ Gramm.

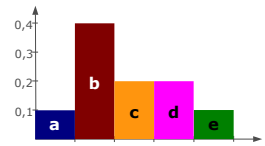
b) Um den Median zu erhalten, müssen wir die kumulierten Häufigkeiten betrachten. Die dritte Klasse ist die erste, deren Häufigkeit erstmalig über 0,5 liegt. Für die folgende Formel ist die dritte Klasse für uns interessant.
$$\text{Median} = x_{un} + \frac{0,5 - F(x_{un})}{f(x_i)} \cdot (x_{ob} - x_{un}) = 0,8 + \frac{0,5 - 0,45}{0,4} \cdot (1,1 - 0,8) = 0,8375$$
Der Median beträgt 0,8375.

c) Der Modus ist die Klassenmitte der Klasse mit der größten Häufigkeit.
Die dritte Klasse hat die größte Häufigkeit, nämlich „0,4". Daher ist die Klassenmitte der dritten Klasse der Modus. Der Modus ist 0,95 Gramm!

W.11.04 Diagramme (册)

Es gibt unzählige Typen von Diagrammen, ungefähr so viele, wie Graffiti in Madrid und die meisten Diagramme sind auch genau so wichtig.
An dieser Stelle werden wir nur drei Typen von Diagrammen betrachten, die die Ehre haben, unsere huldvolle Beachtung zu finden:
a) Histogramme b) Kreisdiagramme c) Boxplot-Diagramme

„Histogramme" sind vermutlich die wichtigsten [und häufigsten] Diagramme zur Veranschaulichung der Stochastik. [„Balkendiagramme" oder „Stabdiagramme" sind fast das Gleiche].
Es gibt schier unendlich viele Möglichkeiten und Regeln ein Histogramm zu zeichnen.
Eine einfache Möglichkeit ist, die Breite der Balken immer gleich breit zu zeichnen und auf der y-Achse die absolute oder relative Häufigkeit aufzutragen.



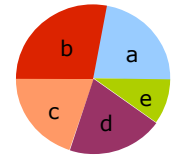
[siehe auch → Bsp.8a]

Eine andere Möglichkeit wäre, auf der x-Achse die Werte der Daten einzutragen, auf der y-Achse den Quotient zwischen der eingetragenen Balkenbreite und der relative Häufigkeit. Also y=f(x)/x. [siehe auch → Bsp.10]
[Bei dieser Variante würde die relative Häufigkeit als Fläche des Rechtecks auftauchen.]

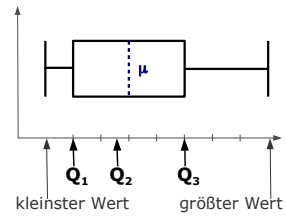
In →Bsp.9 wenden wir ein Zwischending zwischen beiden eben genannten Varianten an.

Ihnen bleibt also leider nichts anders übrig, als nachzuschlagen, welche Variante Ihr Lehrer / Professor / Mentor verwendet und dann ebenfalls diese Methode anzuwenden.

„Kreisdiagramme" sind sehr einfache Diagramme.
Jedem Ereignis wird ein Kreisausschnitt zugeordnet, wobei der Winkel durch die prozentuale Häufigkeit festgelegt ist [Winkel=Häufigkeit*360°].



Ein „Boxplotdiagramm" zeigt *keine* prozentuale Häufigkeit von Merkmalen auf [im Gegensatz zu den letzten beiden Diagrammarten], sondern zeigt in welchem Bereich die Daten liegen [also von welchem Wert bis zu welchem Wert] und zeigt auch noch auf in welchem Bereich die häufigsten Werte liegen.
Man benötigt für das Boxplott auf jeden Fall das Thema „Quartile", [siehe Kapitel →W.11.06].
Leider ist die Darstellung eines Boxplotdiagramms nicht standarisiert, d.h. es gibt kleine, aber viele Unterschiede [je nach Lehrer oder Buch].
Wir zeichnen das Boxplotdiagramm folgendermaßen:
Auf der y-Achse gibt es keine Einheiten.
In die Zeichnung wird ein Rechteck gezeichnet, welches links beim ersten Quartil, rechts beim dritten Quartil endet. Im Inneren des Rechtecks zeichnen wir den Mittelwert ein.
Am linken und rechten Ende des Rechtecks wird jeweils ein liegendes „T" drangesetzt, welche beim kleinsten bzw. beim größten Wert der Datenreihe endet.

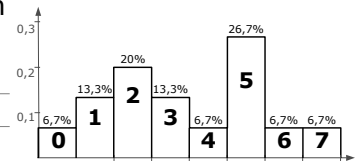


Bsp.8 (Datenreihen)
Die Notaufnahme einer Klinik untersucht 15 Nächte lang die Anzahl der eingehenden Notfälle. Folgende Anzahl von Notfällen werden notiert:
3, 7, 5, 5, 1, 5, 2, 1, 3, 6, 5, 4, 2, 0, 2

a) Zeichnen Sie ein Histogramm des Datensatzes.
b) Veranschaulichen Sie die Daten in einem Kreisdiagramm.
c) Zeichnen Sie ein Boxplotdiagramm.

a) Es tauchen die Werte „0" bis „7" auf, mit folgenden absoluten [h(x)] und relativen [f(x)] Häufigkeiten:

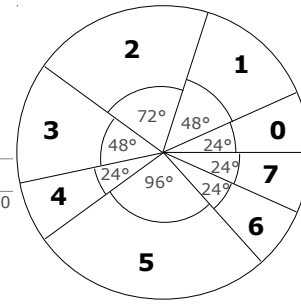
x	0	1	2	3	4	5	6	7
h(x)	1	2	3	2	1	4	1	1
f(x)	1/15= 0,067	2/15= 0,133	2/15= 0,2	2/15= 0,133	1/15= 0,067	4/15= 0,267	1/15= 0,067	1/15= 0,067



Die auftauchenden Werte tragen wir als Rechtecke ein. Die Breite der Rechtecke ist beliebig, als Höhe der Rechtecke wird die Häufigkeit [=Wahrscheinlichkeit] eingetragen. Andere Möglichkeiten für das Histogramm sind auch möglich.

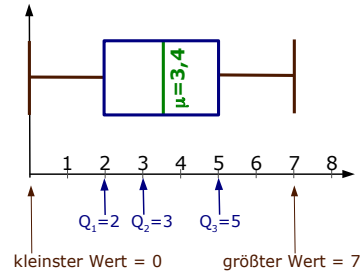
b) Die Häufigkeiten der Werte „0" bis „7" haben wir bereits berechnet. Daraus bestimmen wir die zugehörigen Winkel.

x	0	1	2	3	4	5	6	7
f(x)	0,067	0,133	0,2	0,133	0,067	0,267	0,067	0,067
α	0,067·360 =24°	0,133·360 =48°	0,2·360 =72°	0,133·360 =48°	0,067·360 =24°	0,267·360 =96°	0,067·360 =24°	0,067·360 =24°



Natürlich kann man die Kreissektoren auch einfärben oder sonstwie verschönern.

c) Für das Boxplotdiagramm benötigen wir die Quartile Q_1 und Q_3, und den Median $\tilde{x}$, die wir in →Kap.W.11.06, Bsp.14 berechnen werden.
Man erhält daraus: $Q_1=2$, $Q_2=\tilde{x}=3$, $Q_3=5$.
Den Mittelwert haben wir in Kap.W.11.03, Bsp.5a) berechnet. Wir erhielten $\bar{x}=3{,}4$. Der kleinste auftauchende Wert ist „0", der größte ist „7".

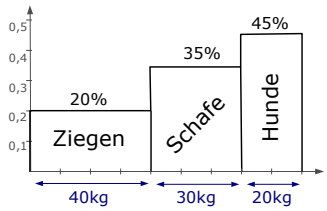


Bsp.9 (Werte mit prozentualen Anteilen)
Familie Günther besitzt eine kleine Farm. 20% der Tiere sind Ziegen, 35% sind Schafe und 45% sind Hunde. Eine Ziege wiegt 40kg, ein Schaf 30kg und ein Hund wiegt 20kg.
a) Zeichnen Sie ein Histogramm des Datensatzes.
b) Veranschaulichen Sie die Daten in einem Kreisdiagramm.
[Ein Boxplotdiagramm macht bei einer Menge von 3 Daten keinen Sinn, daher ist nicht danach gefragt]

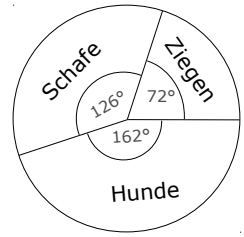
Lösung:
Zum Verständnis: Was sind überhaupt unsere Daten? Sind es die Namen der Tiere, also „Ziege", „Schaf" und „Hund" oder sind es deren Gewichte 40, 30 und 20?
Theoretisch ist beim Histogramm und beim Kreisdiagramm beides möglich. Beim Boxplotdiagramm wären Namen natürlich nicht möglich [oder wie möchten einen Durchschnitt oder ein Quartil von „Ziege", „Schaf" und „Hund" berechnen? ☺]

Name	Gewicht (Balkenbreite)	Häufigkeit (Balkenhöhe)	Winkel (für Kreisdiagramm)
Ziege	40	0,20	0,20·360 = 72°
Schaf	30	0,35	0,35·360 = 126°
Hund	20	0,45	0,45·360 = 162°

a)



b)



Bsp.10 (Klassen)

Eine unglaublich wichtige biologische Untersuchung fördert zu Tage, dass 20% der Maikäfer maximal 0,6 Gramm wiegen, 25% der Maikäfer wiegen 0,6–0,8 Gramm, 40% wiegen 0,8–1,1 Gramm, der Rest wiegt 1,1–1,5 Gramm.

a) Zeichnen Sie ein Histogramm der Maikäfer-Daten.

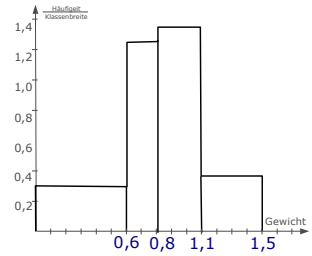
b) Veranschaulichen Sie die Daten in einem Kreisdiagramm.

c) Zeichnen Sie ein Boxplotdiagramm.

Lösung:

a) Wir notieren die wichtigen Daten der vier Klassen:

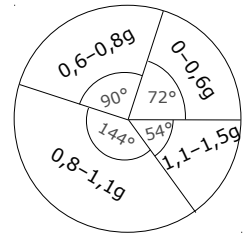
Gewicht	0–0,6g	0,6–0,8g	0,8–1,1g	1,1–1,5g
Klassenbreite	0,6	0,2	0,3	0,4
f(x) [Häufigkeit]	0,20	0,25	0,4	0,15
f(x)/Breite [Rechteckhöhe]	0,33	1,25	1,33	0,375



Wir zeichnen in dieser Aufgabe [willkürlich] ein Histogramm, in welchem auf der x-Achse die Klassenbreite aufgetragen wird und auf der y-Achse der Quotient aus der Häufigkeit und der Klassenbreite. [vierte Zeile der Tabelle]

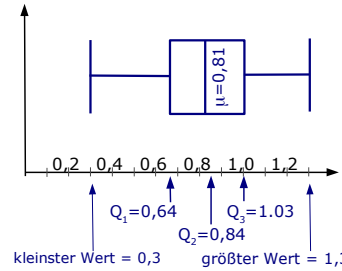
b) Die Mittelpunktswinkel der Klassen bestimmen...

Gewicht	0–0,6g	0,6–0,8g	0,8–1,1g	1,1–1,5g
f(x) [Häufigkeit]	0,20	0,25	0,4	0,15
Winkel α	0,20·360 =72°	0,25·360 =90°	0,4·360 =144°	0,15·360 =54°



Dann das Kreisdiagramm so oder so ähnlich wie dieses rechts eingezeichnete erstellen.

c) Für das Boxplotdiagramm benötigen wir die Quartile Q_1 und Q_3, und den Median $\tilde{x}$, die wir in →Kap.W.11.06, Bsp.15 berechnen werden.
Man erhält daraus: $Q_1=0{,}64$; $Q_2=\tilde{x}=0{,}84$; $Q_3=1{,}03$.
Den Mittelwert haben wir in Kap.W.11.03, Bsp.7a) berechnet. Wir erhielten $\bar{x}=0{,}81$. Der kleinste auftauchende Wert ist „0,3" [Mitte der kleinsten Klasse], der größte ist „1,3" [Mitte der größten Klasse].



W.11.05 Erwartungswert, Varianz, Streuung (ffff)

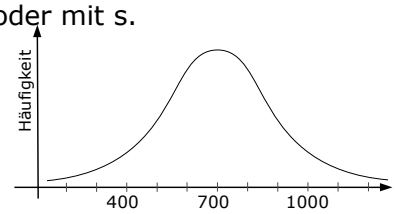
Der Erwartungswert ist ein Durchschnitt bzw. Mittelwert.
In Kap.W.11.03 haben wir das bereits ausführlich durchgekaut.

Die Varianz hat keine anschauliche Bedeutung, sie ist nur das Quadrat der Streuung. Streng genommen bräuchte man also den Begriff „Varianz" in der Mathematik nicht, man könnte alles über die Streuung berechnen.

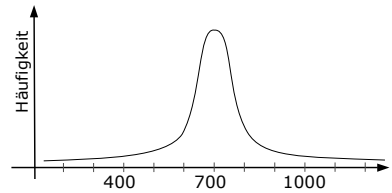
Die Streuung heißt auch Standardabweichung und gibt die Breite der Verteilung an.

Den Erwartungswert bezeichnet man meist mit E(x), μ oder $\bar{x}$.
Die Varianz bezeichnet man meist mit V(x) oder Var(x).
Die Standardabweichung bezeichnet man meist mit σ oder mit s.

Stellen Sie sich mal folgende Situation vor:
Sie verkaufen hauptberuflich in einer Imbissbude gebratene Hähnchen. Ihr Lieferant meint, die Hähnchen hätten ein durchschnittliches Gewicht von 700g. Das ist zwar schön und gut, aber theoretisch könnte es sein, dass die Hälfte der Hähnchen 300g wiegt und die andere Hälfte wiegt 1100g [der Durchschnitt von 300 und 1100 ist 700g]. Das wäre natürlich blöd, denn die Hälfte Ihrer Kunden würden ein unterernährtes Hähnchen bekommen und wären dann selber unterernährt. Die andere Hälfte der Kundschaft würde an Fettsucht sterben.
Es könnte auch sein, dass ALLE Hähnchen 700g wiegen. Dann wären alle glücklich.
Daher brauchen wir außer dem Durchschnittswert einen zweiten Wert, der uns sagt, wie weit die Werte auseinander liegen.
Dieser Wert ist die „Streuung" oder auch „Standardabweichung".



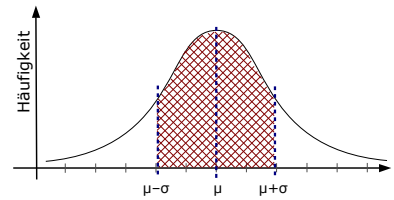
Das Gewicht aller Hähnchen ist recht weit um den Mittelwert gestreut. Viele Hähnchen sind deutlich leichter bzw deutlich schwerer.
Die Standardabweichung ist groß.



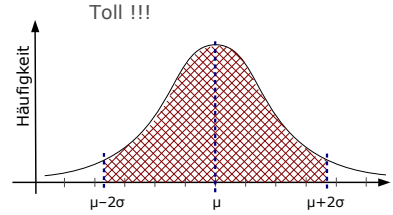
Das Gewicht aller Hähnchen liegt recht dicht um den Mittelwert. Fast alle Hähnchen wiegen ca 700 Gramm.
Die Standardabweichung ist klein.

Eine richtig gute anschauliche Bedeutung gibt es für die Standardabweichung nicht, aber für die grobe Orientierung reicht Folgendes:
Stellen Sie sich von einer Verteilung alle Werte vor, die oberhalb des Mittelwerts liegen und alle Werte, die unterhalb des Mittelwerts liegen. [In unserem Beispiel wäre das einerseits die Menge aller Hähnchen unterhalb 700g bzw. die Menge aller Hähnchen oberhalb 700g]. Die Hälfte der größeren Werte ist im Schnitt eine Standardabweichung größer als der Durchschnitt, die Hälfte der kleineren Werte ist eine Standardabweichung kleiner als der Durchschnitt.
Beispiel: Stellen Sie sich vor, der Hähnchenlieferant meint, die Hähnchen hätten ein Durchschnittgewicht von μ=700g bei einer Standardabweichung von 100g. Nun können Sie ganz grob überschlagen, dass die schwerere Hälfte der Hühnchen ca. 800g wiegen wird [$\mu+\sigma$=700+100], die leichtere Hälfte wiegt im Schnitt 600g [$\mu-\sigma$=700–100]. (Das stimmt zwar, wie gesagt nicht ganz, aber es reicht für eine erste, grobe Vorstellung der Verteilung).

Ebenfalls hilfreich ist die Eigenschaft, dass bei *jeder* beliebigen Verteilung ca. $^2/_3$ der Werte im Bereich liegen, der höchstens eine Standardabweichung vom Mittelwert entfernt sind, 95% aller Werte liegen in einem Bereich, der höchstens zwei Standard-abweichungen vom Mittelwert entfernt sind. [In unserem Hähnchenbeispiel mit μ=700 und σ=100 würde das bedeuten, dass ca. $^2/_3$ der Hähnchen ein Gewicht von 600g bis 800g haben. 95% aller Hähnchen hätten ein Gewicht von 500g bis 900g.]
Genau gesagt gilt:
Ca. 68% aller Werte liegen im Bereich [$\mu-\sigma$; $\mu+\sigma$].
Ca. 95% aller Werte liegen im Bereich [$\mu-2\sigma$;$\mu+2\sigma$].
Ca. 99% aller Werte liegen im Bereich [$\mu-3\sigma$;$\mu+3\sigma$].
[Siehe auch: →W.18.01 Bsp.2]



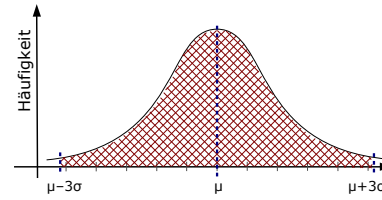
68% aller Werte liegen im Bereich: [$\mu-\sigma$;$\mu+\sigma$]. Toll !!!



95% aller Werte liegen im Bereich: [$\mu-2\sigma$;$\mu+2\sigma$]. Klasse!!!

Richtig genial ist, dass man jede Verteilung komplett mit den beiden Werten μ und σ vorhersagen kann, sprich: man kann die Wahrscheinlichkeitsfunktion nur über μ und σ aufstellen und somit jede gewünschte Wahrscheinlichkeit errechnen.
Wie das im Detail geht, ist nicht überlebenswichtig, wir sehen das im Kapitel: „W.18 Normalverteilung".

Genug herumgeredet. Jetzt die Formeln:



99% aller Werte liegen im Bereich: [$\mu-3\sigma$;$\mu+3\sigma$]. Suuper!!!

Nehmen wir mal an, es gibt mehrere Ereignisse, die wir x_1, x_2, x_3, … nennen.
[In unserem Fall mit den Brathähnchen könnte das das Gewicht der Hähnchen sein.]
Jedes der Ereignisse hat die Häufigkeit bzw. Wahrscheinlichkeit
$p(x_1)$, $p(x_2)$, $p(x_3)$, ...

Den Erwartungswert berechnet man wie folgt:

$$E(x) = \overline{X} = \mu = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + x_3 \cdot p(x_3) + \ldots$$

Die Varianz berechnet man wie folgt:

$$Var = \sigma^2 = p(x_1) \cdot (x_1 - \overline{x})^2 + p(x_2) \cdot (x_2 - \overline{x})^2 + p(x_3) \cdot (x_3 - \overline{x})^2 + \ldots$$

Die Standardabweichung ist die Wurzel aus der Varianz:

$$s = \sigma = \sqrt{Var}$$

Bsp.11 (Datenreihe)
Die Notaufnahme einer Klinik untersucht 15 Nächte lang die Anzahl der eingehenden Notfälle. Folgende Anzahl von Notfällen werden notiert:
3, 7, 5, 5, 1, 5, 2, 1, 3, 6, 5, 4, 2, 0, 2
a) Bestimmen Sie den Erwartungswert der Notfälle.
b) Bestimmen Sie die Varianz und die Standardabweichung.

Lösung:
Sowohl für den Erwartungswert als auch für die Standardabweichung brauchen wir die relativen Häufigkeiten. Also bestimmen wir zuerst die absoluten, dann die relativen Häufigkeiten (was die Wahrscheinlichkeiten sind).

x	0	1	2	3	4	5	6	7
h(x)	1	2	3	2	1	4	1	1
f(x)	1/15= 0,067	2/15= 0,133	2/15= 0,2	2/15= 0,133	1/15= 0,067	4/15= 0,267	1/15= 0,067	1/15= 0,067

a) Der Erwartungswert hat die Formel:

$$\mu = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \ldots =$$
$$= 0 \cdot 0{,}067 + 1 \cdot 0{,}133 + 2 \cdot 0{,}2 + 3 \cdot 0{,}133 + 4 \cdot 0{,}067 + 5 \cdot 0{,}267 + 6 \cdot 0{,}067 + 7 \cdot 0{,}067 \approx$$
$$\approx 3{,}41$$ [beachten Sie das (fast) gleiche Ergebnis wie der Durchschnitt aus Bsp.5]

b) Die Varianz berechnet man mit der Formel:

$$Var = \sigma^2 = p(x_1) \cdot (x_1 - \overline{x})^2 + p(x_2) \cdot (x_2 - \overline{x})^2 + p(x_3) \cdot (x_3 - \overline{x})^2 + \ldots$$
$$= 0{,}067 \cdot (0 - 3{,}41)^2 + 0{,}133 \cdot (1 - 3{,}41)^2 + 0{,}2 \cdot (2 - 3{,}41)^2 + 0{,}133 \cdot (3 - 3{,}41)^2 +$$
$$+ 0{,}067 \cdot (4 - 3{,}41)^2 + 0{,}267 \cdot (5 - 3{,}41)^2 + 0{,}067 \cdot (6 - 3{,}41)^2 + 0{,}067 \cdot (7 - 3{,}41)^2 \approx$$
$$\approx 3{,}983$$

Die Standardabweichung ist: $s = \sigma = \sqrt{3{,}983} = 1{,}996$

Bsp.12 (Werte mit prozentualen Anteilen)
Familie Günther besitzt eine kleine Farm. 20% der Tiere sind Ziegen, 35% sind Schafe und 45% sind Hunde. Eine Ziege wiegt 40kg, ein Schaf 30kg und ein Hund wiegt 20kg.

a) Bestimmen Sie den Erwartungswert des Gewichts eines zufällig ausgesuchten Tieres.
b) Bestimmen Sie die Varianz und die Standardabweichung für das Gewicht.

Lösung:
a) Wir brauchen den Erwartungswert [die Formulierung „zufällig ausgesucht" ist zweitrangig].
$\bar{x} = \mu = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + x_3 \cdot f(x_3) = 20 \cdot 0{,}45 + 30 \cdot 0{,}35 + 40 \cdot 0{,}20 = 27{,}5\text{kg}$.
b) Die Varianz berechnet man mit der Formel:
$\text{Var} = \sigma^2 = p(x_1) \cdot (x_1 - \bar{x})^2 + p(x_2) \cdot (x_2 - \bar{x})^2 + p(x_3) \cdot (x_3 - \bar{x})^2$
$= 0{,}20 \cdot (40 - 27{,}5)^2 + 0{,}35 \cdot (30 - 27{,}5)^2 + 0{,}45 \cdot (20 - 27{,}5)^2 \approx 58{,}75$
Die Standardabweichung ist daher: $s = \sigma = \sqrt{58{,}75} = 7{,}66$

Bsp.13 (Klassen)
Eine unglaublich wichtige biologische Untersuchung fördert zu Tage, dass 20% der Maikäfer maximal 0,6 Gramm wiegen, 25% der Maikäfer wiegen 0,6–0,8 Gramm, 40% wiegen 0,8–1,1 Gramm, der Rest wiegt 1,1–1,5 Gramm.

a) Bestimmen Sie den Erwartungswert des Maikäfergewichts.
b) Bestimmen Sie die Varianz und die Standardabweichung.

Lösung:
Mit den Klassen können wir in der Form, wie sie gegeben sind, nichts anfangen.
Wir brauche für jede Klasse nur *einen* Wert und das ist die Klassenmitte.

Nr.1:	0–0,6g	Klassenmitte: $x_1 = 0{,}3$g	Häufigkeit: $f(x_1) = 0{,}20$
Nr.2:	0,6–0,8g	Klassenmitte: $x_2 = 0{,}7$g	Häufigkeit: $f(x_2) = 0{,}25$
Nr.3:	0,8–1,1g	Klassenmitte: $x_3 = 0{,}95$g	Häufigkeit: $f(x_3) = 0{,}40$
Nr.4:	1,1–1,5g	Klassenmitte: $x_4 = 1{,}3$g	Häufigkeit: $f(x_4) = 0{,}15$

a) Den Erwartungswert berechnet man wie den Durchschnitt:
$\bar{x} = \mu = x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + x_3 \cdot f(x_3) + x_4 \cdot f(x_4) =$
$= 0{,}3 \cdot 0{,}20 + 0{,}7 \cdot 0{,}25 + 0{,}95 \cdot 0{,}40 + 1{,}3 \cdot 0{,}15 = 0{,}81$ Gramm.
b) Die Varianz berechnet man mit der Formel:
$\text{Var} = \sigma^2 = p(x_1) \cdot (x_1 - \bar{x})^2 + p(x_2) \cdot (x_2 - \bar{x})^2 + p(x_3) \cdot (x_3 - \bar{x})^2 + p(x_4) \cdot (x_4 - \bar{x})^2$
$= 0{,}2 \cdot (0{,}3 - 0{,}81)^2 + 0{,}25 \cdot (0{,}7 - 0{,}81)^2 + 0{,}4 \cdot (0{,}95 - 0{,}81)^2 + 0{,}15 \cdot (1{,}3 - 0{,}81)^2 \approx$
$\approx 0{,}0989$
Die Standardabweichung ist daher: $s = \sigma = \sqrt{0{,}0989} = 0{,}31$.

W.11.06 Quartile (⋕)

Ein Quartil ist ein Viertel.
Das erste Quartil ist also derjenige Wert, der beim ersten Viertel der Datenreihe liegt.
Das zweite Quartil ist derjenige Wert, der beim zweiten Viertel der Datenreihe liegt. Da das zweite Viertel die Hälfte ist, ist das zweite Quartil der Median. Kein normaler Mensch verwendet daher den Begriff „zweites Quartil", sondern spricht vom Median.[1]
Das dritte Quartil ist der Wert, der beim dritten Viertel der Datenreihe liegt.

Will man die Datenreihe nicht in Viertel, sondern Drittel, Fünftel, Zehntel, etc.. aufteilen, spricht man nicht von Quartilen, sondern von Quantilen.
Terzile sind Drittel, Quintile sind Fünftel, Dezile sind Zehntel, Perzentile sind Hundetstel.
Quartile und Quantile sind sogenannte Lagemaße.

Quartile / Quantile

Für Quartile und Quantile braucht man immer die kumulierten Häufigkeiten.

Das erste Quartil ist der Wert, dessen kumulierte Wahrscheinlichk. erstmals 0,25 erreicht oder überschreitet.

Das zweite Quartil [=Median] ist der Wert, dessen kumulierte Wahrscheinlichkeit erstmals 0,5 erreicht oder überschreitet.
Das dritte Quartil ist der Wert, dessen kumulierte Wahrscheinlichk. erstmals 0,75 erreicht oder überschreitet.

Bei Klassen ist das Ganze etwas umständlicher. Siehe weiter unten.

Es gibt leider, leider (wie bei Vielem in der Stochastik) auch zu Quartilen unterschiedliche Definitionen ☹

Bsp.14 (Datenreihen)
Die Notaufnahme einer Klinik untersucht 15 Nächte lang die Anzahl der eingehenden Notfälle. Folgende Anzahl von Notfällen werden notiert:
3, 7, 5, 5, 1, 5, 2, 1, 3, 6, 5, 4, 2, 0, 2
Bestimmen den Median, sowie das erste uns dritte Quartil.

Lösung:
[Zur Erinnerung: Für den Median und die Quartile braucht man die kumulierten Häufigkeiten. Das erste Quartil ist der erste Wert, dessen kumulierte Häufigkeit den Wert 0,25 erreicht oder überschreitet. Der Median ist der erste Wert, dessen kumulierte Häufigkeit den Wert 0,5 erreicht oder überschreitet. Das dritte Quartal ist der erste Wert, dessen kumulierte Häufigkeit den Wert 0,75 erreicht über überschreitet.]

x	0	1	2	3	4	5	6	7
f(x)	0,067	0,133	0,2	0,133	0,067	0,267	0,067	0,067
F(x)	0,067	0,2	0,4	0,533	0,6	0,867	0,933	1

Der Wert 0,25 wird zum ersten Mal von der „2" erreicht.
⇒ Das erste Quartil ist: $Q_1=2$
Der Wert 0,5 wird zum ersten Mal von der „3" erreicht.
⇒ Der Median ist: $\tilde{x}=3$.
[Natürlich kann man den Median auch wie in →Kap.W.11.03, Bsp.5 bestimmen, da man es mit einer Datenreihe zu tun hat.]
Der Wert 0,75 wird zum ersten Mal von der „5" erreicht.
⇒ Das dritte Quartil ist: $Q_3=5$

1 Natürlich vorausgesetzt, dass ein normaler Mensch *überhaupt* Begriffe der Stochastik verwendet.

Bsp.15 (Klassen)

Eine unglaublich wichtige biologische Untersuchung fördert zu Tage, dass 20% der Maikäfer maximal 0,6 Gramm wiegen, 25% der Maikäfer wiegen 0,6–0,8 Gramm, 40% wiegen 0,8–1,1 Gramm, der Rest wiegt 1,1–1,5 Gramm.
Bestimmen den Median, das erste und dritte Quartil.

Quartile bei Klassen

Formel für erstes Quartil:

1. Man sucht zuerst die Klasse aus, deren kumulierte W.S. erstmalig 0,25 erreicht oder überschreitet.
2. innerhalb dieser Klasse wird nun das erste Quartil nach der Formel bestimmt:

$$x_{un}+\frac{0{,}25-F(x_{un})}{f(x_i)}\cdot(x_{ob}-x_{un})$$

Hierbei ist x_{un} die untere Grenze der Klasse, x_{ob} ist die obere Grenze. F_{un} ist die kumulierte W.S. der vorhergehenden(!!) Klasse, $f(x_i)$ ist die Häufigkeit der Klasse.

3. Sucht man das **zweite** oder **dritte Quartil**, ersetzt man in der Formel die Zahl 0,25 durch 0,50 bzw. 0,75.

Lösung:

Die Häufigkeit der 4.Klasse beträgt natürlich:
1–0,20–0,25–0,40=0,15

Gewicht	0–0,6g	0,6–0,8g	0,8–1,1g	1,1–1,5g
Klassenmitte	0,3	0,7	0,95	1,3
f(x) [Häufigkeit]	0,20	0,25	0,4	0,15
F(x) [kum.Häuf.]	0,20	0,45	0,85	1

Q_1: Das erste Quartil liegt in jener Klasse, deren kumulierte Häufigkeit 0,25 erreicht oder überschreitet. Das in der zweiten Klasse der Fall.
Die untere Grenze dieser Klasse ist $x_{un}=0{,}6$; die obere Grenze dieser Klasse ist $x_{ob}=0{,}8$; die Häufigkeit der Klasse ist $f_i=0{,}25$; die kumulierte der nächstunteren(!) Klasse ist $F_{un}=0{,}20$.
Diese Werte setzen wir in die Formel für das erste Quartil ein:

$$Q_1 = x_{un}+\frac{0{,}25-F(x_{un})}{f(x_i)}\cdot(x_{ob}-x_{un}) = 0{,}6+\frac{0{,}25-0{,}2}{0{,}25}\cdot(0{,}8-0{,}6) = 0{,}64$$

$\tilde{x}$: Der Median liegt in jener Klasse, deren kumulierte Häufigkeit 0,5 erreicht oder überschreitet. Das in der dritten Klasse der Fall.
Die untere Grenze dieser Klasse ist $x_{un}=0{,}8$; die obere Grenze dieser Klasse ist $x_{ob}=1{,}1$; die Häufigkeit der Klasse ist $f_i=0{,}4$; die kumulierte der nächstunteren Klasse ist $F_{un}=0{,}45$.
Diese Werte setzen wir in die Formel für das erste Quartil ein:

$$\tilde{x} = x_{un}+\frac{0{,}5-F(x_{un})}{f(x_i)}\cdot(x_{ob}-x_{un}) = 0{,}8+\frac{0{,}5-0{,}45}{0{,}4}\cdot(1{,}1-0{,}8) \approx 0{,}84$$

Q_3: Das dritte Quartil liegt in jener Klasse, deren kumulierte Häufigkeit 0,75 erreicht oder überschreitet. Das in der dritten Klasse der Fall.
Die untere Grenze dieser Klasse ist $x_{un}=0{,}8$; die obere Grenze dieser Klasse ist $x_{ob}=1{,}1$; die Häufigkeit der Klasse ist $f_i=0{,}4$; die kumulierte der nächstunteren Klasse ist $F_{un}=0{,}45$.
Diese Werte setzen wir in die Formel für das erste Quartil ein:

$$Q_3 = x_{un}+\frac{0{,}75-F(x_{un})}{f(x_i)}\cdot(x_{ob}-x_{un}) = 0{,}8+\frac{0{,}75-0{,}45}{0{,}4}\cdot(1{,}1-0{,}8) \approx 1{,}03$$

W.11.07 Dichtefunktion (∮)

Eine Dichtefunktion ist eine Wahrscheinlichkeitsfunktion, also eine Funktion, mit deren Hilfe man Wahrscheinlichkeiten berechnen kann.
Die bekanntesten Wahrscheinlichkeitsfunktionen sind die Normalverteilung [→Kap.W.18] und die Binomialverteilung [→Kap.W.16]. Meistens verwendet man jedoch beliebige andere Funktionen.
Eine Funktion muss nur zwei Bedingungen erfüllen, um die große Ehre zu haben, zu einer Wahrscheinlichkeitsfunktion ernannt zu werden:

1. Sie darf nur positive Werte annehmen.
   [es gibt schließlich nur positive Wahrscheinlichkeiten]
2. Die Fläche zwischen der Funktion und der x-Achse im Bereich von -∞ bis +∞ muss 1 ergeben, also $\int_{-\infty}^{+\infty} f(x)\,dx = 1$.
   [Weil die gesamte Wahrscheinlichkeit 1 ergeben muss!]

Eine **Dichtefunktion** muss zwei Bedingungen erfüllen:

1. es werden nur positive Funktionswerte angenommen.
2. das Integral zwischen Funktion und x-Achse im Bereich von -∞ bis +∞ beträgt 1.

Die Wahrscheinlichkeit, dass ein Ereignis zwischen zwei Werten a und b liegt, berechnet man mit dem Integral:

$$P([a;b]) = \int_a^b f(x)\,dx$$

Bsp.16

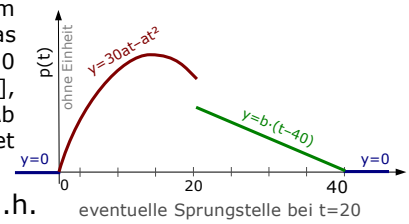
Ein Getränkehersteller testet ein neues Erfrischungsprodukt, welches den neuartigen Geschmackstoff HX71 enthalten soll. Leider verdunstet HX71 sehr leicht. Eine Forschungsgruppe untersucht, welcher prozentuale Anteil von HX71 im Laufe der Zeit bei Zimmertemperatur verdunstet und erhält folgendes Ergebnis:

$$p(t) = \begin{cases} 0 & \text{für } t<0 \\ 30a\cdot t - a\cdot t^2 & \text{für } 0\leqslant t<20 \\ b\cdot(t-40) & \text{für } 20\leqslant t\leqslant 40 \\ 0 & \text{für } t>40 \end{cases}$$

(t ist die Zeit in Minuten,
p(t) der Anteil des verdunsteten Geschmackstoffes)

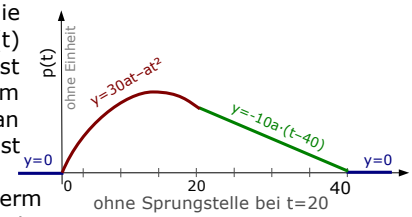
a) Bei welcher Temperatur verdunstet HX71 vollständig?
b) Begründen Sie, dass die Annahme b=-10a sinnvoll ist.
c) Bestimmen Sie a.
d) Welcher prozentuale Anteil verdunstet in den ersten zehn Minuten?
e) Welcher Anteil verdunstet von der 15. bis zur 30. Minute?
f) Welcher Anteil verdunstet zum Zeitpunkt 8^{30} ?

Lösung:
Interpretation von p(t): Wenn man sich die [zusammengesetzte] Funktion p(t) anschaut, stellt man fest, dass bis zu einem Zeitpunkt von t=0 nichts verdunstet [p(t)=0 für t<0], was irgendwie auch logisch ist. Im Zeitraum von t=0 bis t=20 verdunstet ein bisschen was [wird durch „30a·t−a·t²" beschrieben], ab t=20 wird die Verdunstung durch „b·(t−40)" beschrieben. Ab t=40 ist die Verdunstung wieder „0", d.h. ab t=40 verdunstet nichts mehr. Damit wäre Teilaufgabe a) beantwortet.



a) Ab dem Zeitpunkt t=40 verdunstet nichts mehr, d.h. zum Zeitpunkt t=40 ist HX71 vollständig verdunstet.

b) Die Funktion p(t) hat mehrere Sprungstellen. Bei t=0 springt p(t) vom Term „0" zu „30a·t–a·t²". Betrachtet man die y-Werte, bedeutet dass, das p(t) von „0" zu „30a·0–a·0²"=0 springt. Die y-Werte der Funktion springen bei t=0 also gar nicht, p(t) wechselt kontinuierlich von „0" zu „30a·t–a·t²". Das gleiche ist bei der letzten Sprungstelle, bei t=40. Setzt man in den Term „b·(t–40)" den Wert t=40 ein, erhält man „0", der Übergang an der Stelle t=40 vom Term „b·(t–40)" zum nächsten Term „0" ist also kontinuierlich.



Betrachten wir den Übergang bei t=20: p(t) wechselt vom Term „30a·t–a·t²" zum Term „b·(t–40)" Es wäre zu erwarten, dass die y-Werte hier auch kontinuierlich in einander übergehen, dass also zum Zeitpunkt t=20 gilt: $30a\cdot t-a\cdot t^2=b\cdot(t-40)$ [Es ist zwar nicht zwingend notwendig, dass das gilt, aber die Aufgabenstellung lautet auch nur zu begründen, dass es sinnvoll wäre]

Sinnvoll wäre, wenn die Übergänge an den drei Übergangsstellen t=0, t=20 und t=40 ohne Sprungstellen in der Funktion p(t) stattfinden würden.

An der Stelle t=20 würde daher gelten:

$30a\cdot t-a\cdot t^2=b\cdot(t-40)$ bzw $30a\cdot 20-a\cdot 20^2=b\cdot(20-40)$

$\Leftrightarrow\ 600a-400a=-20b\ \Leftrightarrow\ 200a=-20b\ \Leftrightarrow\ -10a=b.$

Dadurch erhält p(t) die Form:

$$p(t)=\begin{cases}0\\ 30a\cdot t-a\cdot t^2\\ -10a\cdot(t-40)\\ 0\end{cases}\quad\Leftrightarrow\quad p(t)=\begin{cases}0 & \text{für} & t<0\\ 30a\cdot t-a\cdot t^2 & \text{für} & 0\leqslant t<20\\ -10at+400a & \text{für} & 20\leqslant t\leqslant 40\\ 0 & \text{für} & t>40\end{cases}$$

c) Für eine Dichtefunktion muss gelten:

$\int_{-\infty}^{+\infty} p(t)dt = 1$, in unserem Fall also:

$$\int_{-\infty}^{0} 0\,dt + \int_{0}^{20} 30at-at^2\,dt + \int_{20}^{40} -10at+400a\,dt + \int_{40}^{+\infty} 0dt = 1$$

$$\Leftrightarrow\ 0 + \left[15at^2-\frac{1}{3}at^3\right]_0^{20} + \left[-5at^2+400at\right]_{20}^{40} + 0 = 1$$

$$\Leftrightarrow\ \left[15a\cdot 20^2-\frac{1}{3}a\cdot 20^3\right]-\left[15a\cdot 0^2-\frac{1}{3}a\cdot 0^3\right]+\left[-5a\cdot 40^2+400a\cdot 40\right]-\left[-5a\cdot 20^2+400a\cdot 20\right] = 1\ \Leftrightarrow$$

$$\left[6000a-\frac{8000}{3}\cdot a\right]-[0] + [-8000a+16000a]-[6000a] = 1$$

$$\Leftrightarrow\ \frac{16000}{3}\cdot a = 1\quad\Leftrightarrow\quad a=\frac{3}{16000}$$

Diesen Wert von a in p(t) einsetzen:

$$p(t)=\begin{cases}0 & \text{für} & t<0\\ \frac{9}{1600}\cdot t-\frac{3}{16000}\cdot t^2 & \text{für} & 0\leqslant t<20\\ -\frac{3}{1600}\cdot t+\frac{3}{40} & \text{für} & 20\leqslant t\leqslant 40\\ 0 & \text{für} & t>40\end{cases}$$

d) Den Anteil, der in den ersten 10 Minuten verdunstet, berechnet man über das Integral mit t=0 und t=10 als Grenzen.

$$\int_0^{10}\frac{9}{1600}\cdot t-\frac{3}{16000}\cdot t^2\,dt = \left[\frac{9}{3200}\cdot t^2-\frac{1}{16000}\cdot t^3\right]_0^{10} =$$

$$= \left[\frac{9}{3200}\cdot 10^2-\frac{1}{16000}\cdot 10^3\right]-\left[\frac{9}{3200}\cdot 0^2-\frac{1}{16000}\cdot 0^3\right] = \frac{7}{32}\approx 21{,}9\%$$

e) Der Anteil, der von der 15. bis zur 30. Minute verdunstet, ist natürlich wieder das Integral mit den Grenzen t=15 bis t=30. Der Unterschied zu Teilaufgabe c) ist nur der, dass wir das Integral aufteilen müssen, da p(t) bei t=20 vom einen Term zu einem anderen wechselt.

$$\int_{15}^{30} p(t)\,dt = \int_{15}^{20} \frac{9}{1600}\cdot t - \frac{3}{16000}\cdot t^2\,dt + \int_{20}^{30} -\frac{3}{1600}\cdot t + \frac{3}{40}\,dt =$$

$$= \left[\frac{9}{3200}\cdot t^2 - \frac{1}{16000}\cdot t^3\right]_{15}^{20} + \left[\frac{-3}{3200}\cdot t^2 + \frac{3}{40}\cdot t\right]_{20}^{30} =$$

$$= \left[\frac{9}{3200}\cdot 20^2 - \frac{1}{16000}\cdot 20^3\right] - \left[\frac{9}{3200}\cdot 15^2 - \frac{1}{16000}\cdot 15^3\right] + \left[\frac{-3}{3200}\cdot 30^2 + \frac{3}{40}\cdot 30\right] - \left[\frac{-3}{3200}\cdot 20^2 + \frac{3}{40}\cdot 20\right] =$$

$$= \left[\frac{5}{8}\right] - \left[\frac{27}{64}\right] + \left[\frac{45}{32}\right] - \left[\frac{9}{8}\right] = \frac{31}{64} \approx 48{,}4\%$$

f) Einen prozentualen Anteil [eine Wahrscheinlichkeit] berechnet man immer über ein Intervall mit *zwei* Grenzen. Wenn man, wie in dieser Aufgabe nur *eine* Grenze gegeben hat, geht das nicht. Das Ergebnis ist Null.
Die anschauliche Begründung wäre die, dass zum Zeitpunkt 8^{30} nichts verdunsten kann, weil es sich hier ja um einen unendlich kurzen Zeitpunkt handelt und nicht um eine Zeitspanne von .. bis ..