(A) Wachstum ohne Grenzen – exponentielles Wachstum

Populationen (Bestände) wachsen (ungehindert) meist so, dass in gewissen Zeitschritten immer ein bestimmter Anteil der Population (des Bestandes) dazukommt, also:

(1) *Neuer Bestand ist alter Bestand plus Anteil des alten Bestandes* oder

(2) *Die Änderung (Zuwachs: neu-alt) ist Anteil des alten Bestandes.*

In Formeln:

(1) $\underset{\text{neuer Bestand}}{x_{neu}} = \underset{\text{alter Bestand}}{x_{alt}} + \underset{\text{Anteil alter Bestand}}{w \cdot x_{alt}} = (1+w) \cdot x_{alt} = \underset{\text{Vielfaches von altem Bestand}}{b \cdot x_{alt}}$

(2) $x_{neu} - x_{alt} = w \cdot x_{alt}$

Es ist die Änderung („neu – alt"), die meist interessanter, manchmal auch einfacher zugänglich, als der Bestand ist oder auch: Die Anzahl der Infizierten ist nicht so ‚spannend', sondern die Änderung der Anzahl der Infizierten im Verlauf der Zeit. Dies modelliert man mathematisch so:

(a) Diskrete Zeitschritte ($n = 1, 2, 3, \ldots$):

$x_n - x_{n-1} = w \cdot x_{n-1}$; . x_0 ist ein Anfangswert bzw. $x_n = b \cdot x_{n-1}$ (vgl. oben (1)).

(b) Kontinuierlich: $f'(x) = k \cdot f(x)$. f' : Ableitung von f (x: reelle Zahl)

Kurz: **Änderung ist proportional zum Bestand**, Änderung ist Vielfaches des Bestandes.
Lösungen von (a) und (b) sind dann Exponentialfunktionen:

(a) $\boxed{x_n = a \cdot (1+w)^n = a \cdot b^n}$

(b) $\boxed{f(x) = a \cdot b^x = a \cdot e^{\ln(b) \cdot x} = a \cdot e^{k \cdot x}}$ (e: Eulerzahl: 2,71828…, ln: natürlicher Logarithmus)

Das exponentielle Wachstum hat eine ganz andere Qualität als das anschaulich naheliegende lineare Wachstum, wo pro Zeitschritt immer eine konstante Menge dazukommt. Ein Beispiel:

1 Zwei Jobangebote

Jonas erhält zwei verschiedene Jobangebote für einen Monat (20 Arbeitstage).

(A) Zu Beginn gibt es 50 €, am ersten Tag 55 € und dann jeden Tag 5 € mehr.

(B) Es gibt 1 Cent Startgeld, am ersten Tag 2 Cent, am zweiten 4 Cent, am dritten 8 Cent, am vierten 16 Cent usw.

Man entscheide zunächst „aus dem Bauch" und rechne dann.

Bekannt ist auch folgende Legende:

Das Märchen vom Reiskorn und dem Schachbrett

Im alten Persien erzählten sich die Menschen einst dieses Märchen: Es war einmal ein kluger Höfling, der seinem König ein kostbares Schachbrett schenkte. Der König war über den Zeitvertreib sehr dankbar und er sprach zu seinem Höfling: „Sage mir, wie ich dich zum Dank für dieses wunderschöne Geschenk belohnen kann. Ich werde dir jeden Wunsch erfüllen." „Nichts weiter will ich, als dass Ihr das Schachbrett mit Reis auffüllen möget. Legt ein Reiskorn auf das erste Feld, zwei Reiskörner auf das zweite Feld, vier Reiskörner auf das dritte, acht auf das vierte und so fort." Der König war erstaunt über so viel Bescheidenheit und ordnete sogleich die Erfüllung des Wunsches an…

- Wie viele Reiskörner liegen auf dem 64. Feld? Wie viele Körner sind es insgesamt?
- Wie viele Güterwagen (Länge: 15,8 m; Zuladung: 66,5 t) benötigt man, um die Reismenge zu transportieren? Wie lang wäre der Güterzug? Ein Kilogramm sind ungefähr 50000 Reiskörner.

Oder, weniger als Legende:

HOIMAR VON DITFURTH
(1921 – 1989)

Vergleich: Lineares und exponentielles Wachstum

Lineares und exponentielles Wachstum sind zwei grundlegende Wachstumsmodelle, die aber zu sehr unterschiedlichen Einschätzungen zukünftiger Entwicklungen führen können. Der Mensch neigt zu dem Denken, dass alle Prozesse linear sind, in der Natur verlaufen aber viele Prozesse häufig zunächst eher exponentiell. Hoimar von Ditfurth drückte das in den 80er Jahren des 20. Jahrhundert so aus: *Menschen zählen zeitlebens wie ABC-Schützen: eins, zwei… viele. Die Natur zählt, etwa bei Zellteilungen, anders: 400, 800, 1600, 3200… unendlich. Wenn der Mensch nicht bald lernt, naturgemäß zu zählen, sprich: mit der Natur zu rechnen, wird er ausgezählt.*

HOIMAR VON DITFURTH: „Unfähig zu zählen" (Natur 2/1985)

Übungen

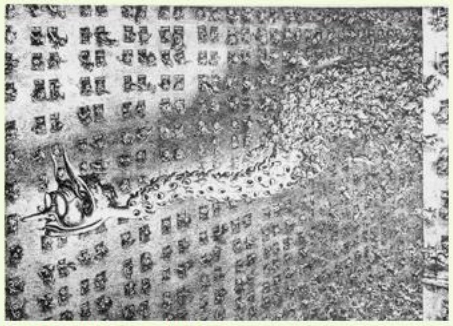
16 Seerosen

Auf einem 8 ha großen See befindet sich ein 100 m^2 großes Feld von Seerosen. Jährlich verdoppelt sich diese Fläche. Wenn ein See vollständig mit Seerosen bedeckt ist, stirbt er. Wann ist der See vollständig bedeckt? Wann ist er halbvoll bedeckt? Als er halbvoll bedeckt ist, bemerkt eine Spaziergängerin: *„Oh, wie schön, diese Seerosen!"* Was würden Sie ihr erzählen? Nehmen Sie Bezug zum Zitat von HOIMAR VON DITFURTH.

(B) Wachstum mit Grenze – logistisches Wachstum

Nach **(A)** wächst aber alles über alle Grenzen, das gibt es nicht. Wie kann man das jetzt modellieren? Zur Veranschaulichung:

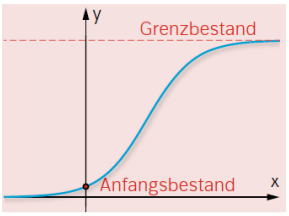
1 Ein Gerücht



P. Weber: Das Gerücht

Vier Schüler fangen an, um 8:00 Uhr in einer Schule mit 900 Schülerinnen und Schülern ein Gerücht zu verbreiten. Um 9:00 Uhr kennen schon 24 Schüler das Gerücht. Wie verbreitet sich das Gerücht?

Nun, zuerst wächst die Anzahl der ‚Gerüchtkenner' annähernd exponentiell, Schüler treffen zunächst ja meist nur auf Schüler, die es noch nicht kennen. Aber mit zunehmender Zeit treffen sie mehr auf ‚Kenner' als auf ‚Nichtkenner', die Ausbreitung verlangsamt sich. Wenn alle das Gerücht kennen, gibt es keine Zunahme der ‚Kenner' mehr, keine weitere Verbreitung (Schule hier als geschlossenes System). Die Verbreitung lässt sich qualitativ antizipieren wie es die Abbildung darstellt.



▪ Wie kann man das nun so modellieren, dass man auch quantitative Aussagen (Prognosen etc.) machen kann? Wann kennen alle Schüler das Gerücht? Wann die Hälfte?
Anfänglich wird sicherlich exponentielles Wachstum ein sinnvolles Modell sein, aber nur bis es zur Verlangsamung der Ausbreitungsgeschwindigkeit kommt („die Kurve oben steigt zunehmend weniger stark an."). Für diese Verlangsamung der Verbreitung ist die Annahme sinnvoll, dass die Zunahme der Verbreitung von der Anzahl der Schüler abhängt, die das Gerücht noch nicht kennen, das ist „die Anzahl aller Schüler (Grenze G) minus der Anzahl der Schüler, die das Gerücht kennen (Bestand f(x))".
Einfachste Möglichkeit dies zu modellieren ist wieder:
Die Änderung (Zuwachs) ist proportional zu „Grenze minus Bestand", also $G - f(x)$. Damit ergibt sich eine doppelte Proportionalität:
1. Änderung proportional zum Bestand: $f'(x) = k \cdot f(x)$
2. Änderung ist proportional zum Restbestand: $f'(x) = k \cdot (G - f(x))$
Nimmt der Bestand *f(x)* zu, dann wird $G - f(x)$ kleiner, bei einem konstanten *k* wird dann also $f'(x)$ kleiner, im Grenzfall, wenn $G = f(x)$, dann ist Änderung $f'(x) = 0$
Es gilt: Wenn eine Größe zu zwei Größen proportional ist, dann ist sie auch zum Produkt der Größen proportional, also:

$$\boxed{f'(x) = k \cdot f(x) \cdot (G - f(x))} \quad (*)$$

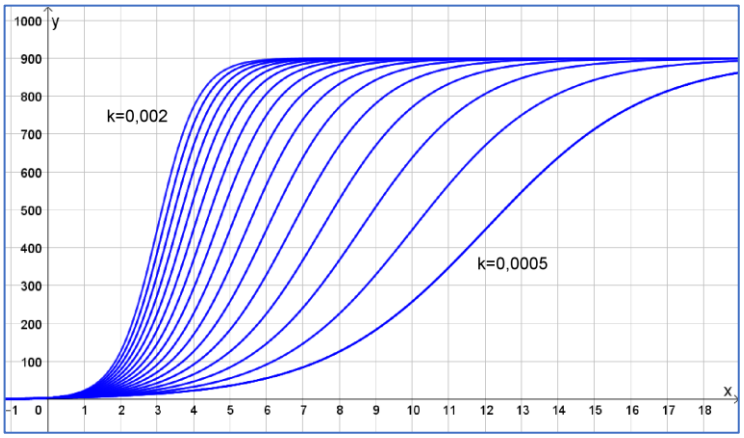
Das Besondere solcher Gleichungen ist, dass hier als Lösung nicht Zahlen x gesucht sind, sondern Funktionen *f(x)*, man nennt solche Gleichungen Differentialgleichungen (DGL). Sie sind das wesentliche Modellierungswerkzeug in den Anwendungen aus Naturwissenschaft und Technik, vor allem auch in den Umwelt- und Biowissenschaften. Durch Überlegungen (Theorie) kann man (Schüler auch) auf die Gleichung (*) kommen, auf die folgende Lösungsfunktion (**) aber nicht! Nachdenken über „Änderung" lohnt sich also (nicht nur in Zeiten ständiger Veränderung).
Lösungsfunktionen dieser Gleichung:

$$\boxed{f(x) = \frac{A \cdot G}{A + (G - A) \cdot e^{-k \cdot G \cdot x}}} \quad (**)$$ Anfangswert: $A = f(0)$; Grenze: G; *k*: Wachstumskonstante

Natürlich muss man wieder wissen, wie man solche Gleichungen löst. Das ist meist ‚höllisch' schwer, klappt in den meisten relevanten Fällen auch nur näherungsweise, aber dafür gibt es – vor allem auch mit digitalen Werkzeugen – sehr effektive Näherungsverfahren.

Hier die Grafik zu der Verbreitung des Gerüchts. $A = 4$; $G = 900$

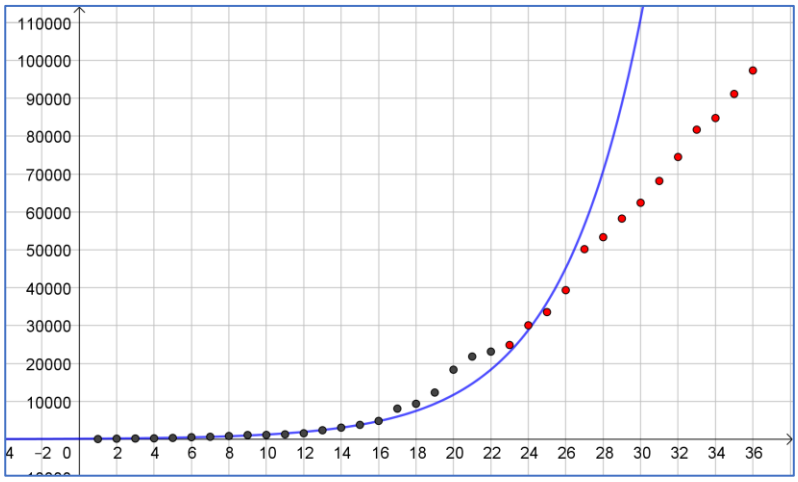
Die Wachstumskonstante wird variiert, sie beschreibt die unterschiedlich schnelle Verbreitung.
Anmerkung: Mit der Information, dass zu $x = 1$ (9:00 Uhr) 24 Schüler das Gerücht kennen, wird genau eine Wachstumskonstante festgelegt



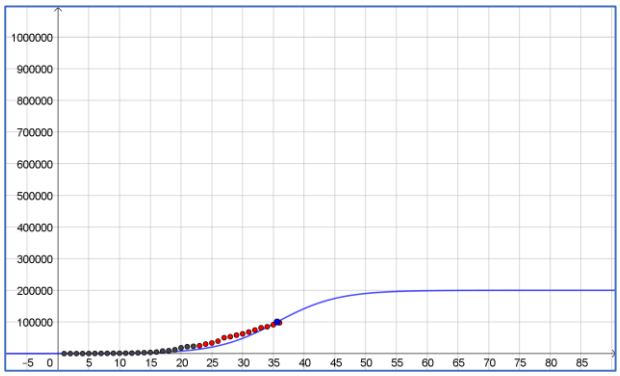
Man sieht deutlich, dass die Grenze invariant ist, nur die Annäherungsgeschwindigkeit variiert. K kann hier als Maß für die Ausbreitungsgeschwindigkeit angesehen werden:
Wie viele Schüler trifft jeder in welcher Zeit?
Wenn alle lange in ihren Klassenräumen sitzen und nur kurze Pausen stattfinden, wird k klein sein.
Wenn alle auf Schulfest in Aula und Umgebung zusammen sind, wird k groß sein.
Womit man schon (fast) bei Corona ist.

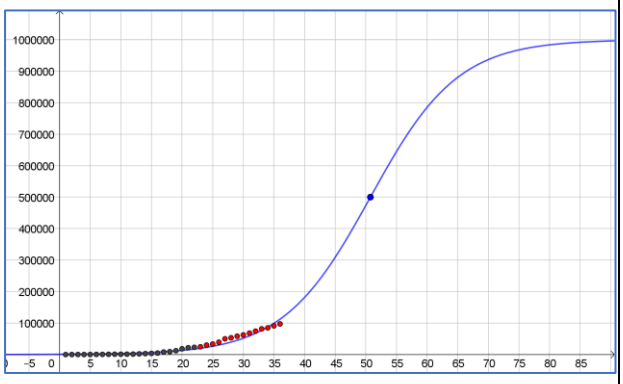
Henning Körner	Corona	17.04.2020	4

Zu Corona:
Zunächst eine Modellierung mit (A), in rot die Daten in Deutschland nach Beginn der Kontaktsperre. Man sieht deutlich, dass die ‚Explosivität' des Wachstums abgenommen hat. „Social Distancing" linearisiert fast!
https://de.statista.com/statistik/daten/studie/1102667/umfrage/erkrankungs-und-todesfaelle-aufgrund-des-coronavirus-in-deutschland/



Wenn man nun mit (B) modelliert, passiert Folgendes. Durch Variation von A, G und k wird Datenpassung erzeugt.





Es fällt auf, dass gänzlich unterschiedliche Prognosen möglich (und nach Datenlage gleichermaßen ‚rational') sind, aber gänzlich unterschiedliche Grenzzahlen an Infizierten prognostizieren. Erst wenn man den Punkt mit maximaler Steigung (maximale Änderungsrate, blauer Punkt) kennt, kann man sicherer prognostizieren. Man kann berechnen, dass nach Modell B zum Zeitpunkt der maximalen Wachstumsgeschwindigkeit (wo die Kurve am steilsten ist), der halbe Grenzbestand erreicht ist.
Wenn also der 05.04. (97351 Infizierte) diesen Zeitpunkt markiert, dann würde es nach dem Modell (B) ca. 200000 Infizierte geben. Aber genau das weiß man nicht, wird man auch erst etwas genauer wissen, wenn man einige Tage hinter diesem Zeitpunkt ist.

(C) Wachstum mit Gift oder Infektion mit Heilung (Immunisierung)

Modell (B) beschreibt den Verlauf einer Epidemie bis alle infiziert sind („Durchseuchung"). Wie kann man nun ‚Heilung' bzw. Immunisierung modellieren?

Wie beim Modellieren üblich, bildet Kritik am vorliegenden Modell Motor für Modifikation:

Es gilt: $f'(x) = k \cdot f(x) \cdot \big(G - f(x)\big) = \big(k \cdot G - k \cdot f(x)\big) \cdot f(x)$

Setzt man $g = k \cdot G$ und $s = k$ dann ergibt sich folgende Darstellung des logistischen Wachstums: $f'(x) = \big(g - s \cdot f(x)\big) \cdot f(x)$.

Der Faktor $g - s \cdot f(x)$ setzt sich aus der konstanten Geburtenrate g und einer zum Bestand proportionalen Sterberate $s \cdot f(x)$ zusammen.

Bei Bakterien hängt das Absterben, genau wie das Gesunden bei der Grippewelle, eher von der Zeit als vom Bestand ab. Es liegt daher folgende Modellierung nahe:

$f'(x) = (g - s \cdot x) \cdot f(x)$
$g > 0; \quad s > 0$

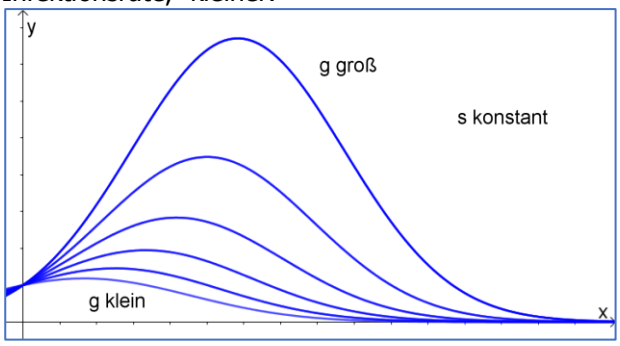
Dies passt auch zu Virenepidemien: Die Gesundungsrate s hängt eher von der Zeit ab ($s \cdot x$) als vom Bestand ($s \cdot f(x)$). Irgendwann („wenn nur genug Zeit vergeht", also x größer wird), wird $s \cdot x$ größer als das konstante g, dann wird $g - s \cdot t < 0$ und damit die Änderung negativ, der Bestand (Anzahl der Infizierten) nimmt ab…

Lösungsfunktionen der im weißen Kasten stehenden DGL sind die Funktionen

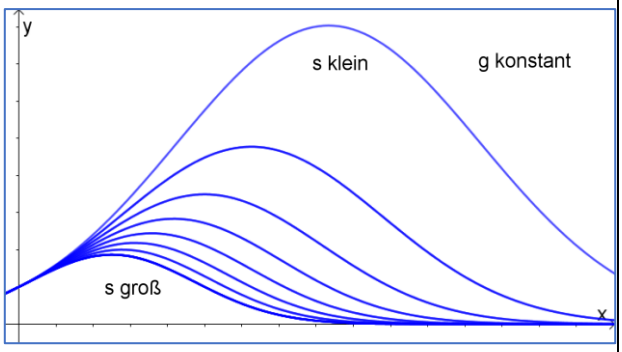
$$f(x) = e^{g \cdot x - \frac{s}{2} \cdot x^2}$$

Handlungen in Realität werden wieder durch Variationen der zugehörigen Parameter simuliert.

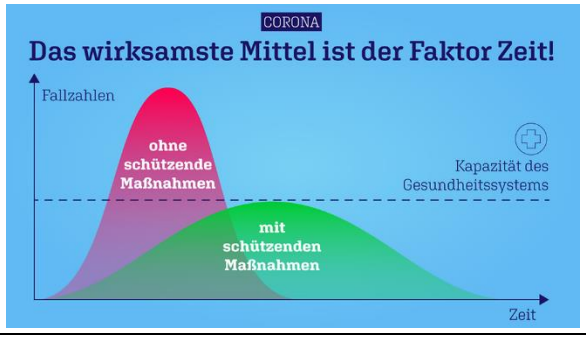
Schulschließungen, Kontaktsperren etc. machen g, die Ausbreitungsgeschwindigkeit und damit die Infektionsrate, kleiner.



Medikamentenentwicklung bis hin zu Impfungen machen s groß.



Die allseits bekannte Grafik aus den Medien.



(D) Wachstum mit Kranken, Heilung und Toten

Zwei Vorbemerkungen:
1. Die Modellierung mit mehreren Zustandsgrößen gelingt einfacher in diskreter Form (vgl. (A) (a))
2. Eine Population wird nur als ein geschlossenes System betrachtet (also z.B. die Bevölkerung Deutschlands).
3. Natürlich spielt bei den Zusammenhängen auch Zufälliges eine Rolle. Dies wird hier nicht berücksichtigt.
2. und 3. Sind natürlich starke Vereinfachungen (Reduktionen), aber: Hier wirkt die Dialektik des Modellierens. Das reduktive Vorgehen jeglicher Modellierung (Realitätsausschnitte, ausgewählte Parameter) einerseits und die Beherrschbarkeit von Realitätsbereichen andererseits durch gerade die Reduktion von Komplexität. Natürlich kann man auch Offenheit von Systemen mit stochastischen ‚Verschmutzungen' mathematisch modellieren, aber dann ist man in ganz anderen fachlichen Gefilden (der Autor ist da nicht mehr zuhause).

Modell:

Es werden Gesunde g_n , Kranke k_n und Tote T_n betrachtet.

g_5 sind also die Gesunden im 5. Zeitschritt (Tage, Wochen, Monate,…)

Modellannahmen:

(1) *Die Gesunden*
(i) Die Gesunden bleiben gesund,
(ii) Es gibt Kranke, die gesund werden
(iii) Es gibt Gesunde, die sich anstecken.
Mathematisierung:
Ein Anteil *a* der Kranken wird gesund. Die Begegnung von Gesunden mit Kranken wird mit dem Produkt modelliert, also $g_n \cdot k_n$. Beispiel: Es gibt 4 Gesunde und 3 Kranke. Dann kann jeder Gesunde jedem Kranken begegnen, also gibt es $4 \cdot 3$ Möglichkeiten der Begegnung. Natürlich finden nicht alle statt, deswegen „Anteilsfaktor" *b*. Man sieht. Bei Kontaktsperre ist b sehr klein, beim Rockfestival sehr groß. Und dann wird ja auch nicht jeder krank, wenn er einem Gesunden begegnet.

$$(1) \quad g_n = \underset{\text{(i)}}{g_{n-1}} + \underset{\text{(ii)}}{a \cdot k_{n-1}} - \underset{\text{(iii)}}{b \cdot g_{n-1} \cdot k_{n-1}}$$

(2) *Die Kranken*
(i) Kranke bleiben krank.
(ii) Kranke werden gesund.
(iii) Gesunde, die neu krank geworden sind.
(iv) Kranke, die gestorben sind.
Mathematisierung:
(ii) und (iii) gehen hier mit jeweils anderem Vorzeichen ein.

$$(2) \quad k_n = \underset{\text{(i)}}{k_{n-1}} - \underset{\text{(ii)}}{a \cdot k_{n-1}} + \underset{\text{(iii)}}{b \cdot g_{n-1} \cdot k_{n-1}} - \underset{\text{(iv)}}{c \cdot k_{n-1}}$$

(3) *Die Toten*
(i) Tote bleiben tot.
(ii) Gestorbene Kranke
Mathematisierung:

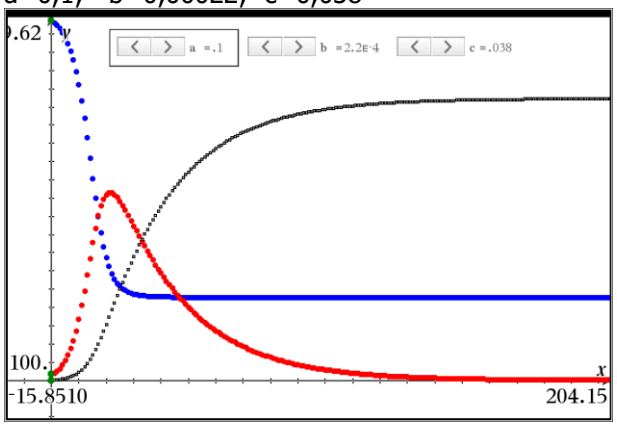
$$(3) \quad t_n = \underset{\text{(i)}}{t_{n-1}} + \underset{\text{(ii)}}{c \cdot k_{n-1}}$$

Mit (1), (2) und (3) hat man schon ein (kleines) vernetztes System, bei dem sich wohl kaum noch gedanklich langfristige Entwicklungen ‚scharf' prognostizieren lassen noch gar Zahlenwerte geschätzt werden können. Hier können nun digitale Werkzeuge ihre Kraft entfalten. Man löst keine Gleichung, man formt keine Terme mehr um, sondern man schaut sich tabellarische und grafische Entwicklung an. Mit Variation der Parameter (a: Gesundungsrate; b: Ansteckungsrate; c: Sterberate) können dann verschieden Szenarien entwickelt (simuliert) werden. Handlungen in der Realität entsprechen dann bestimmte Parameterwerte.

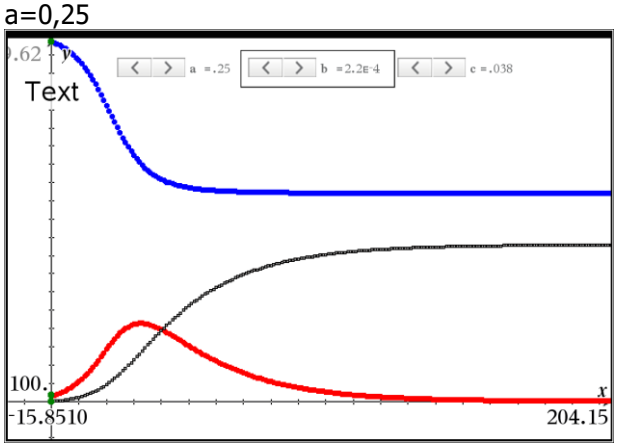
Simulationen: blau: Gesunde; rot: Kranke; schwarz: Tote
Es müssen (natürlich) noch Anfangswerte gesetzt werden.
Im Modell wird von einer Gesamtpopulation von 2000 ausgegangen.
Hier: $g_0 = 1970$; $k_0 = 30$; $t_0 = 0$
Sinnvolle Simulationen gehen so vor, dass man einen Parameter variiert und alle übrigen fest lässt.

1. Variation der Gesundungsrate a

a=0,1; b=0,00022; c=0,038

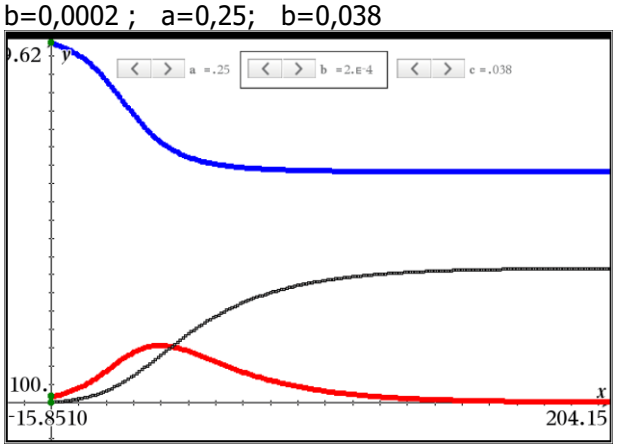


a=0,25

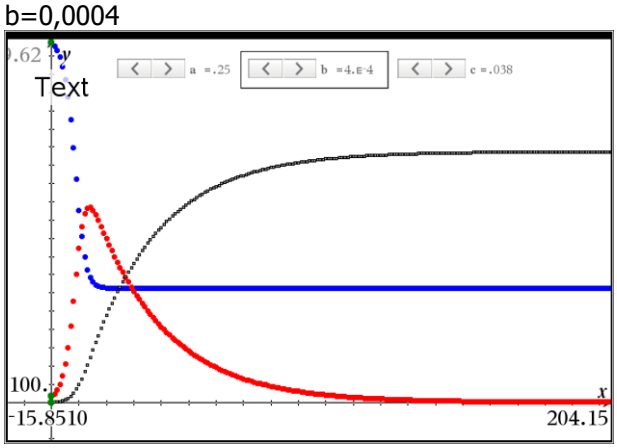


2. Variation der Ansteckungsrate b

b=0,0002 ; a=0,25; b=0,038

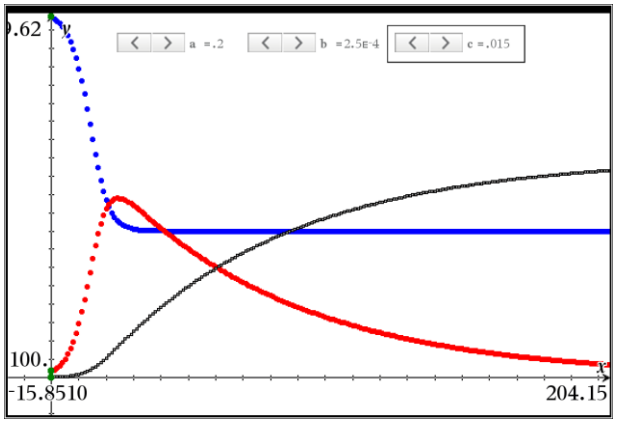


b=0,0004

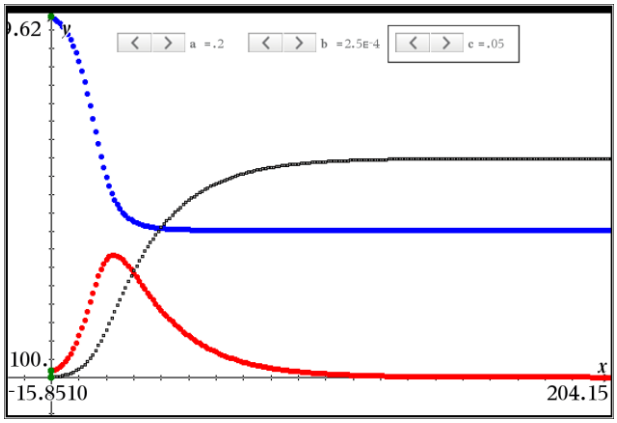


3. Variation der Sterberate

c=0,015 ; a=0,2; b=0,00025

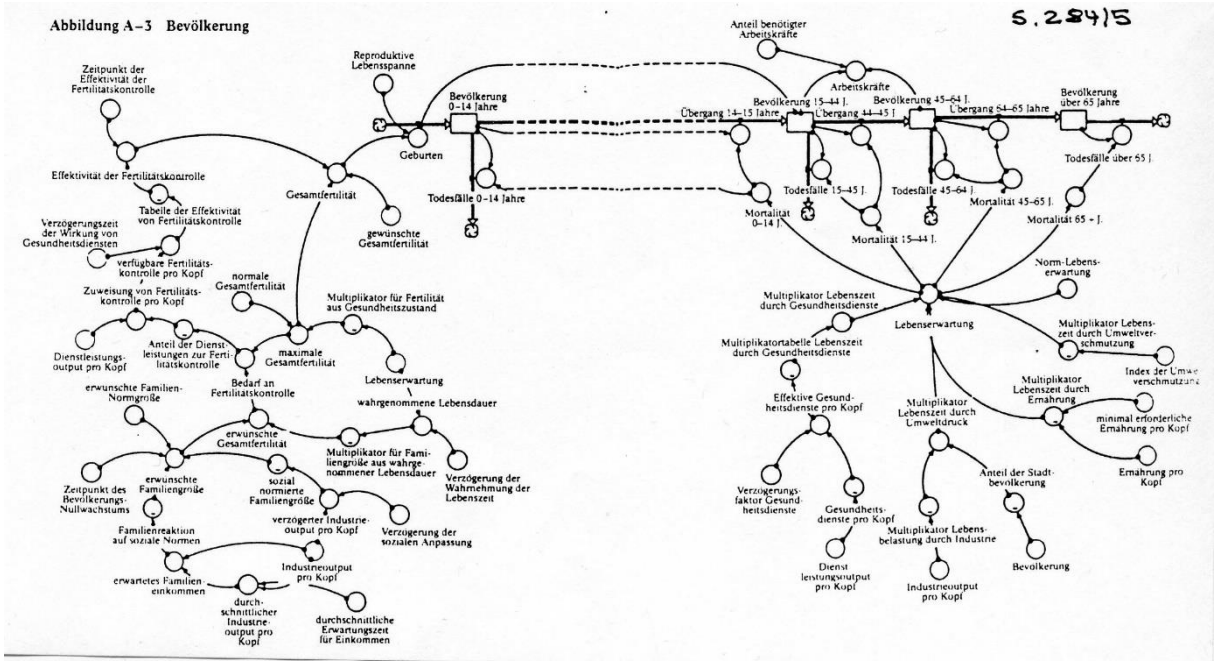


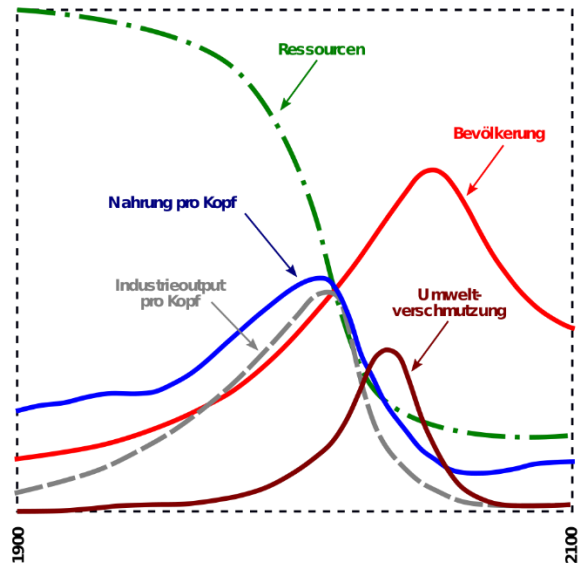
c=0,05

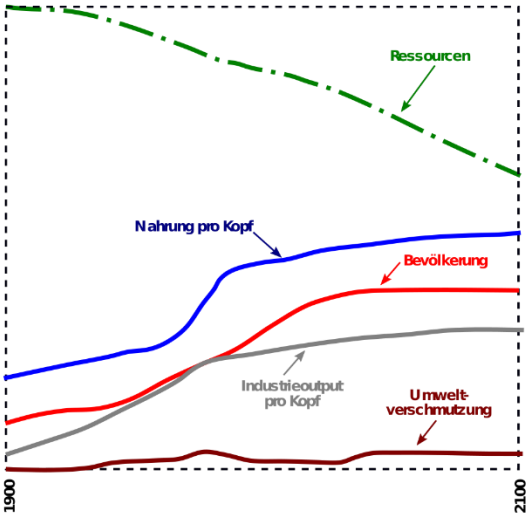


Man interpretiere die Szenarien selber.
Gibt es Fragen? Natürlich! Die Anzahl der letztendlich Gesunden (und damit Toten) scheint unabhängig von der Sterberate zu sein (3.). Wird das in der Realität so sein? Oder ist das nur in dem Modell so? …

Ausblick:
Die Modellierung in (D) ist archetypisch für Modellierungen von vernetzten Systemen in Natur- Bio- und Umweltwissenschaften. Beispiel: Die Modelle von Meadows/Meadows (Meadows/Meadows: Die neuen Grenzen des Wachstums, Hamburg 1993) modellieren mehrere tausend von Zustandsgrößen wo hier nur 3 Größen vernetzt wurden. Die Abbildung zeigt die Vernetzung von Größen bei der Modellierung von „Bevölkerung", die Graphen zwei Szenarien für unterschiedliche Parameter („gedachte Handlungen von Menschen) von verschiedenen Subsystemen.







(E) Die Dunkelziffer der Infizierten schätzen

Ein großes Problem liegt in der Dunkelziffer der Anzahl der Infizierten.
Wie lässt sich die Anzahl der Infizierten schätzen?
Einfachste Möglichkeit: Man testet alle Menschen einer Population...
Aber das ist zur Zeit aus verschiedenen Gründen nicht möglich.
Ein Beispiel:
Will man eine unbekannte Anzahl von Populationen schätzen, betrachtet man eine Stichprobe, ‚markiert' diese und gibt sie wieder in die Gesamtheit zurück.

Ein Beispiel:

New York Times Science March 16, 2012

Scientist at Work

Population estimate: a count of individual whales by photographic capture-recapture-study

From the Air

The scars and marks on the backs of bowhead whales (Grönlandwale) can be used to identify individuals. However, not all bowhead whales are well marked. An aerial survey team is flying over the open lead to assess the whales that are marked and to photograph („capture") marked animals. These photos will be compared with photos from previous years to „recapture" marked individuals. Although it is a gross oversimplification, knowing the proportion of recaptured animals from the population of marked whales and the overall proportion of marked (but not „captured") whales in the total population can lead to an estimate of the number of animals in the entire population.

Angenommen, den Wissenschaftlern sind 611 markierte Grönlandwale bekannt. Bei der beschriebenen Studie wurden auf den Fotos 856 Wale gezählt, darunter 38 markierte. Der Anteil der markierten Wale an allen Walen in der Studie beträgt damit $\frac{38}{856} = 0,044$. Also sind auf den Fotos ca. 4,4% markierte Wale. Dieser Anteil wird auf die Gesamtzahl hochgerechnet (Dreisatz): $x = \frac{856 \cdot 611}{38} \approx 13764$

Dieses Modell ist zwar einfach, aber auch nur sehr grob:
1. Ist die fotographierte Menge repräsentativ?
2. Jedes neue Foto führt zu anderen Anteilen und damit Schätzungen.

Solche Schätzungen heißen Punktschätzungen, weil sie einen ‚genauen Wert' angeben. Aber das ist natürlich extrem fehleranfällig und meist auch nicht das primäre Ziel. Man möchte eher ein gewisses Intervall der Bestandsmenge auf der Basis einer gewissen Sicherheit haben. Auf das Beispiel bezogen: Man möchte nicht wissen, ob es 13764 Grönlandwale oder 14128 gibt, sondern etwa so:
Mit (sehr) hoher Wahrscheinlichkeit sind es zwischen 13000 und 14000.
Oder so: Bei 200 Münzwürfen ist 100 sicher ein guter Schätzwert für „Zahl", aber 99 oder 102 wohl auch nicht viel schlechter. Allerdings: Wenn nur 70-mal Zahl fällt, dann...
Und: Die Wahrscheinlichkeit, bei 200 Münzwürfen genau 100-mal Zahl zu beträgt nur 5,6%!

In Österreich sind deutlich mehr Menschen mit dem Coronavirus infiziert als die offizielle Statistik ausweist. Laut einer Dunkelziffer-Studie sind es mehr als dreimal so viele Menschen.

Von Clemens Verenkotte, ARD-Studio Wien

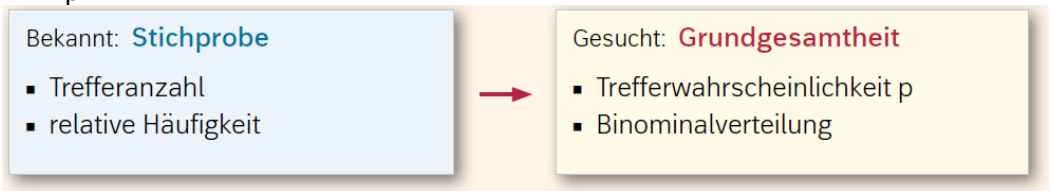
Es handele sich um eine echte Zufallsstudie, um die Dunkelziffer der Corona-Infektionen festzustellen, sagte Österreichs Forschungsminister Heinz Faßmann bei der Vorstellung der ersten repräsentativen Studie über die Anzahl der Infizierten im Land. Die sogenannte "Prävalenzstudie" sei vom Roten Kreuz, der Medizinischen Universität Wien und dem Meinungsforschungsinstitut SORA im Zeitraum vom 1. bis 6. April durchgeführt worden.

Bei 1544 Menschen aller Bevölkerungsgruppen und Altersgruppen sei von Rotkreuz-Mitarbeitern ein sogenannter "PCR-Test" - also ein Rachenabstrich - vorgenommen worden. Laut Studie seien 0,33 Prozent dieser Personen Corona-positiv gewesen. Aus den Ergebnissen sei dann - bezogen auf die Gesamtbevölkerung in Österreich - eine Schätzung von 28.500 Menschen ermittelt worden.

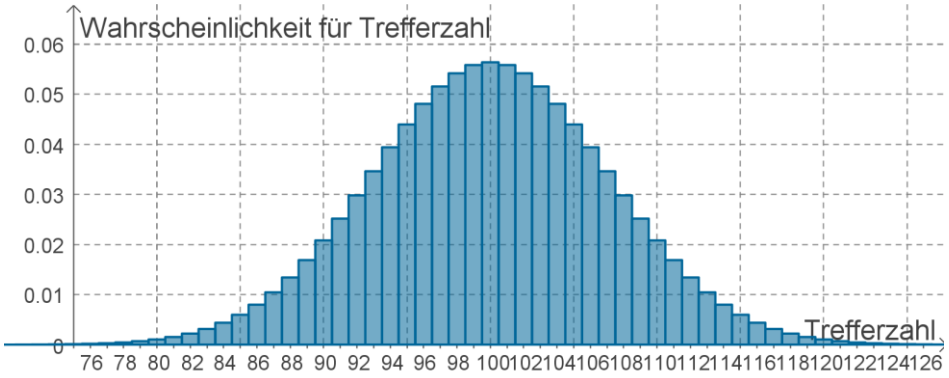
https://www.tagesschau.de/ausland/oesterreich-corona-101.html

Nimmt man diese Werte und rechnet nach „capture – recapture" hoch, erhält man ca. 8600000 als Gesamtbevölkerungszahl von Österreich (es sind wohl etwas mehr, 8,859 (Eurostat)). Der Artikel spricht richtig von „Schätzung", gibt aber nur „Punktschätzung" an (auf Hundert gerundet). Besser, allerdings etwas schwerer lesbar wäre eine Intervallschätzung wie oben angedeutet.

Prinzip:



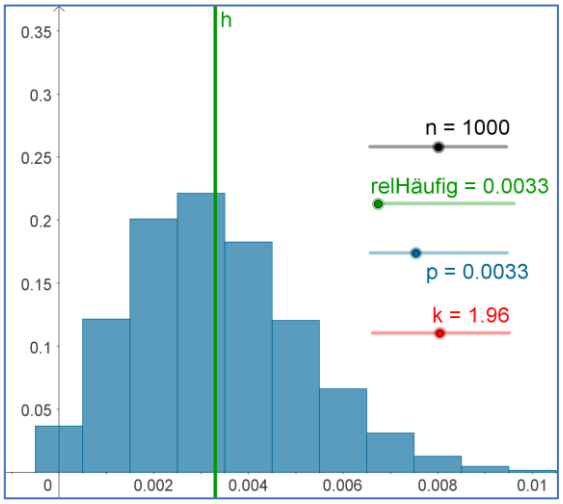
Die Frage hier: Wenn in einer Stichprobe $0,0033 = 0,33\%$ Infizierte sind (relative Häufugkeit), wie wahrscheinlich ist es dann, dass ein Österreicher infiziert ist, wenn ich einen beliebigen auswähle?
Nochmal zum Münzwurf: Das Diagramm zeigt die Wahrscheinlichkeiten beim 200-maligen Münzwurf für unterschiedliche Trefferzahlen (z.B. „Zahl").



Man sieht deutlich, dass die Wahrscheinlichkeit für 100 Treffer am größten ist (aber eben auch nur 5,6%) und die Wahrscheinlichkeit für 70 Treffer annähernd 0 ist, sie beträgt 0,0000063, d.h.: wenn ca. 160000 mal 200-mal eine Münze geworfen wir, gibt es ‚im Schnitt' einmal 70 Treffer.

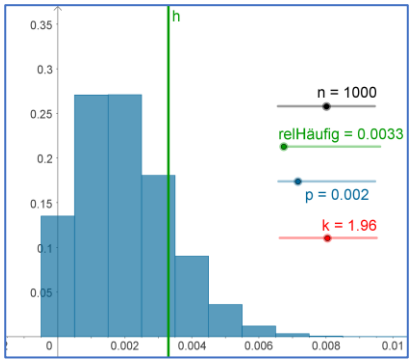
Zurück zu Corona:
Die Abbildungen beziehen sich auf eine Stichprobe von 1000 getesten Personen, weil die Software kein größeres n für eine Grafik zulässt. An den Abbildungen lässt sich aber alles Wesentliche erfassen (ohne jedes Rechnen usw.)
Grafische Untersuchung: Man markiert das Ergebnis der Stichprobe, die relative Häufigkeit (0,0033 grün) und skizziert Wahrscheinlichkeitsverteilungen (blau) für unterschiedliche vermutbare Wahrscheinlichkeiten in der Gesamtpopulation. Man schaut dann, ob das h irgendwie gut zu der Verteilung (blau) passt.

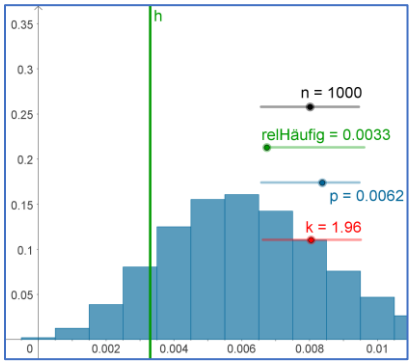
p=0,0033 passt natürlich ‚am besten', weil hier das Stichprobenergebnis genau der Wahrscheinlichkeit entspricht (h=p)

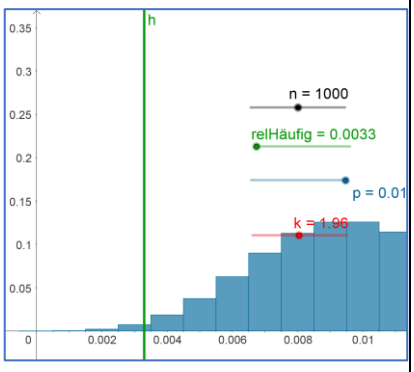


Wenn man nun p verändert, verschieben sich die Verteilungen.
Passt die relative Häufigkeit noch zu den Verteilungen?

p=0,002 passt wohl „mittelgut", p=0,0062 auch nicht richtig gut und p=0,01 passt ganz schlecht.

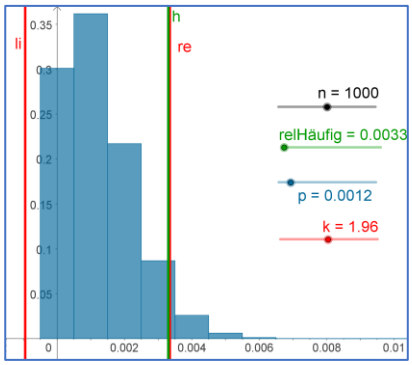


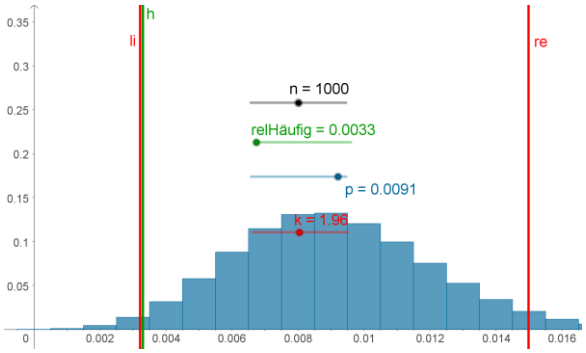




Es wird unmittelbar klar:
1. Ein eindeutiges Ergebnis kann es nicht geben. Vor allem: Man muss *vorher* festlegen, was man unter „gut passen" versteht.
2. Die relative Häufigkeit tritt in jeder Verteilung auf! Allerdings häufig nur mit verschwindend kleiner Wahrscheinlichkeit.
„Mittel gut", „ganz gut", „passt eher nicht" sind natürlich unbefriedigend. Das muss noch ‚messbar', weil dann vergleichbar, gemacht werden.
Die ‚Strenge' des Ergebnisses legt man vorgängig durch ein Intervall fest, in dem sich die relative Häufigkeit h befinden soll. Die Grenzen dieses Intervalls werden mit k („Signifikanzniveau", rot) gesteuert.
Zunächst gilt natürlich das nichtssagende Ergebnis:
Will man 100%-ige Sicherheit („absolute Strenge"), kann man nur sagen, dass jedes p passt (vgl. 2.).

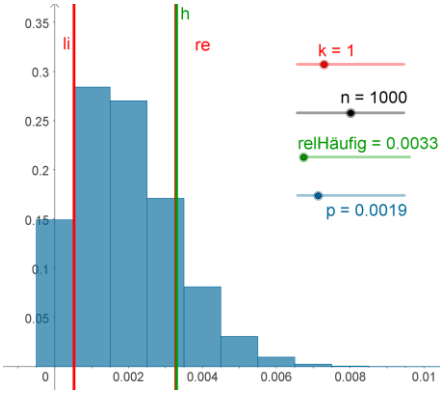
1. k=1,96: 5% der (blauen) Werte liegen außerhalb des Intervalls, also 95% im Intervall.
Man sagt dann: Mit einer Sicherheitswahrscheinlichkeit von 95% liegt die Wahrscheinlichkeit zwischen p=0,0012 und p=0,0091 (vgl. Angaben in Grafik).

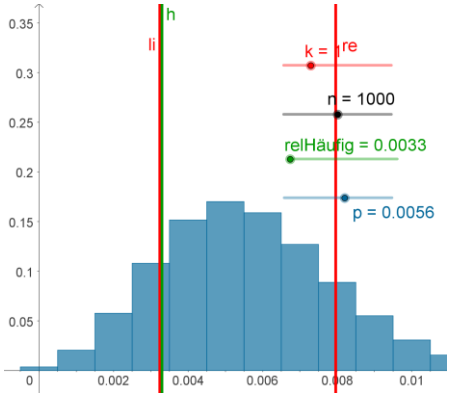




Wenn p kleiner als 0,0012 wird, liegt die grüne Gerade (h) nicht mehr zwischen li und re.

2. k=1: 32% der (blauen) Werte liegen außerhalb des Intervalls, also 68% im Imtervall.
Man sagt dann: Mit einer Sicherheitswahrscheinlichkeit von 68% liegt die Wahrscheinlichkeit zwischen p=0,0019 und p=0,0056.





Man erkennt:
Mit 1. und 2. erhält man unterschiedliche Ergebnisse. beide sind begründet, wenn man das Sicherheitsniveau explizit angibt. Umgekehrt gilt dann:

Ohne Angabe von solchen Sicherheitsniveaus sind alle Aussagen eigentlich leer. Wenn ich möchte, dass auch p=0,0001 oder p=0,2 eine ‚passende' Wahrscheinlichkeit ist, muss ich nur das Sicherkeitsniveau k entsprechend klein machen. Man kann also in der Tat eine Statistik so anlegen, dass man das Ergebnis erhält, das man wünscht. Ein Schelm, wer denkt, dass dies nicht geschieht. Aber eben auch umgekehrt: Wenn keiner nach den fehlenden Werten (Informationen) fragt, muss man sich nicht wundern. Aber dann müssten ja Zahlen etc. angegeben werden, das halbiert die Leserschaft.

Man kann die Intervalle natürlich auch berechnen:
Hier mit den Angaben aus der Nachricht (h=1544).
68%-Sicherheit: 0,21%<p<0,51%
95%-Sicherheit: 0,14%<p<0,76%
99%-Sicherheit: 0,11%<p<0,97%
Mit 8900000 Einwohner Österreichs erhält man die Intervalle:
68%-Sicherheit: [18690;45390]
95%-Sicherheit: [12460;67640]
99%-Sicherheit: [9790;86330]

$progli(n,p,k):=p-k\cdot\sqrt{\frac{p\cdot(1-p)}{n}}$ *Fertig*

$progre(n,p,k):=p+k\cdot\sqrt{\frac{p\cdot(1-p)}{n}}$ *Fertig*

$\text{solve}(progli(1544,p,1.96)<0.0033 \text{ and } progre(1544,p,1.96)>0.0$
$0.001421<p<0.007645$

$\text{solve}(progli(1544,p,1)<0.0033 \text{ and } progre(1544,p,1)>0.0033,p)$
$0.002127<p<0.005116$

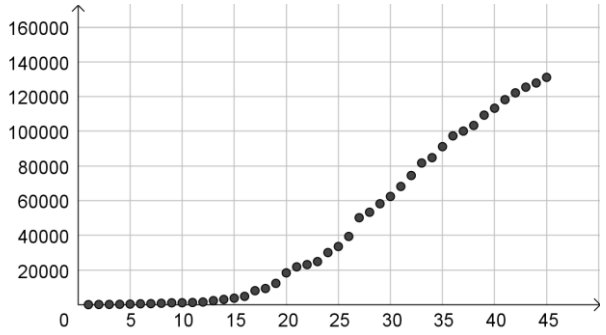
$\text{solve}(progli(1544,p,2.58)<0.0033 \text{ and } progre(1544,p,2.58)>0.0$
$0.001112<p<0.009752$

Klar: Die Schätzung aus der Nachricht liegt in allen Intervallen, aber...

Und: Es bleibt das Riesenproblem: Ist die Stichprobe repräsentativ? Aber das ist eine andere Geschichte.

Ergänzung 1:

Die Daten bis zum 14.04. Deutlich ist das ‚Verlassen' der exponentiellen Phase zu erkennen. Wenn man nun mit logistischem Wachstum modelliert und (35|90000) als Wendepunkt (maximale Steigung) annimmt, kann man eine Maximalzahl Infizierter von ca. 180000 prognostizieren. Aber: Dies setzt Beibehaltung der Kontaktsperre etc. voraus! Wenn die zu schnell zu stark gelockert wird, könnte es (wird es) neue exponentielle Phase geben (Wellenförmige Bewegung).



Ergänzung 2:

Warum bei den Statistiken Vorsicht geboten ist

CORONA-KRISE Verfügbare Zahlen bilden nur einen kleinen Teil der Realität ab – Zu wenig Tests

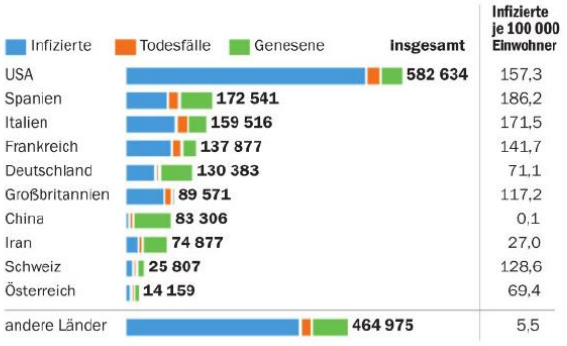
VON MARCO KREFTING

MÜNCHEN – Den größten Wert in Corona-Zeiten haben – neben Nudeln und Klopapier – wohl Zahlen. Auf einmal rufen Menschen, die mit Mathematik wenig und mit Statistik überhaupt nichts am Hut haben, mehrmals täglich Daten zu Coronavirus-Fällen ab. Statistiker warnen aber davor, sich allzu sehr auf die Zahlen zu verlassen.

Die verfügbaren Zahlen enthielten zu wenige Informationen, erklärt Katharina Schüller, Gründerin des Münchner Unternehmens Stat-Up und Leiterin der Arbeitsgruppe „Statistical Literacy" der Deutschen Statistischen Gesellschaft. „Sie bilden nur einen kleinen Teil der Realität ab, nämlich die schwer Erkrankten, einen Teil der leichter Erkrankten mit Symptomen und einen ganz kleinen Teil von Menschen ohne Krankheitszeichen, die getestet wurden, weil sie Verdachtsfälle waren."

Ob auch viele andere infiziert sind oder nicht, „das wissen wir nicht und können es auch nur mehr oder weniger begründet erraten", schreibt Schüller in einem Beitrag für das Hochschulforum Digitalisierung: „Wir wissen, dass jede unserer Modellrechnungen falsch sein muss." Trotzdem könnten die Schlussfolgerungen daraus richtig sein.

Covid-19: Die am stärksten betroffenen Länder



Die Weltgesundheitsorganisation (WHO) weist unter anderem auf „Unterschiede bei den Berichtsmethoden, rückwirkende Datenkonsolidierung und Verzögerungen bei der Berichterstattung" hin. Wegen der Inkubationszeit, der Zeit für den Test und der Meldeverzögerungen zeigen in Deutschland zum Beispiel Maßnahmen wie Kontaktverbote oft erst etwa 14 Tage später Folgen bei den Zahlen.

Es gibt viele solche Stolperfallen bei den Corona-Daten. Die Tücke liegt wie so oft im Detail. Besonders heikel sind Ländervergleiche. „Insbesondere hängen die erfassten Fallzahlen in jedem Land zentral davon ab, wie systematisch und umfangreich dort auf das Virus getestet wird", erklären die Macher der „Unstatistik des Monats", einem Angebot mehrerer Statistik-Experten, das auf mögliche Fehler bei der Interpretation von Statistiken hinweist.

Etliche Faktoren beeinflussen Stand und Schweregrad der Infektionen und können sich von Land zu Land immens unterscheiden: Einwohnerzahl, Altersstruktur, spezielle Erkrankungen in der Bevölkerung wie Tuberkulose, das Stadium der Ausbruchswelle, der Wille oder das Vermögen zu testen, die Richtlinien dafür, wer überhaupt getestet wird. In Altenheimen gestorbene Menschen etwa werden in einigen Ländern nachträglich getestet und fließen in die Statistik ein – in anderen nicht. Da vorwiegend Ältere mit Covid-19 sterben, kann das enorme Unterschiede zur Folge haben.

Die statistische Erfassung der Todesursachen variiere von Land zu Land erheblich, betonen auch die Macher der „Unstatistik", zu denen Katharina Schüller gehört. Dennoch werden immer wieder Vergleiche von Sterberaten diskutiert. Generell sei es falsch, einfach die Toten ins Verhältnis zu den bekannten Infizierten zu setzen. Werde die Dunkelziffer nicht berücksichtigt, werde die Letalität systematisch überschätzt.

Knifflig wird es auch bei Aussagen zur Zahl der Genesenen, die hier und da bis auf die letzte Stelle angegeben werden und damit ziemlich exakt aussehen. Doch wo nicht einmal alle Infizierten getestet und erhoben werden, kann natürlich noch viel weniger über die Zahl der Genesenen bekannt sein. Daher sind all diese Angaben immer nur Schätzungen – sehr grobe Schätzungen in vielen Fällen.

NWZ 15.04.2020 S.4

Die Ausführungen zu (E) mit den dortigen Ergebnissen sind eine Konkretion des markierten Textstücks.