Institut für Angewandte Statistik
www.jku.at/ifas

Studien-Info-Messe SIM 023

Andreas Quatember

Mustervorlesung zum JKU-Bachelorprogramm

STATISTIK UND DATA SCIENCE

Mit der Analyse von Daten am Puls der Zeit

JKU - PLATZ FÜR …

- etwa 23.000 Studierende aus 100 verschiedenen Ländern
- rund 100 Studienrichtungen an vier Fakultäten
- mehr als 170 Professor:innen an 140 Instituten und Universitäts-kliniken
- ca. 3900 Mitarbeiter:innen

DER JKU-CAMPUS - PLATZ FÜR MITEINANDER UND FÜR NEUES

Die Aufgabenstellung von Statistik und Data Science

Womit beschäftigen sich Statistiker:innen und Data Scientists ?

Gewinnung, Aufbereitung, Organisation, Verwaltung, Methodenentwicklung, Visualisierung und Analyse von Daten sowie Kommunikation der Ergebnisse (Data Literacy)

Was sind Daten ?

Mit Inhalt gefüllte Zahlen zu einem bestimmten Sachverhalt, die extra erhoben werden oder nebenbei in einem Prozess anfallen

Was ist der Zweck der Data Literacy ?

Das Aufdecken der in Daten vorhandenen Informationen, um Zusammenhänge besser verstehen, faktenbasiert Entscheidungen treffen und Entwicklungen vorhersagen zu können

Globale Bevölkerungsentwicklung

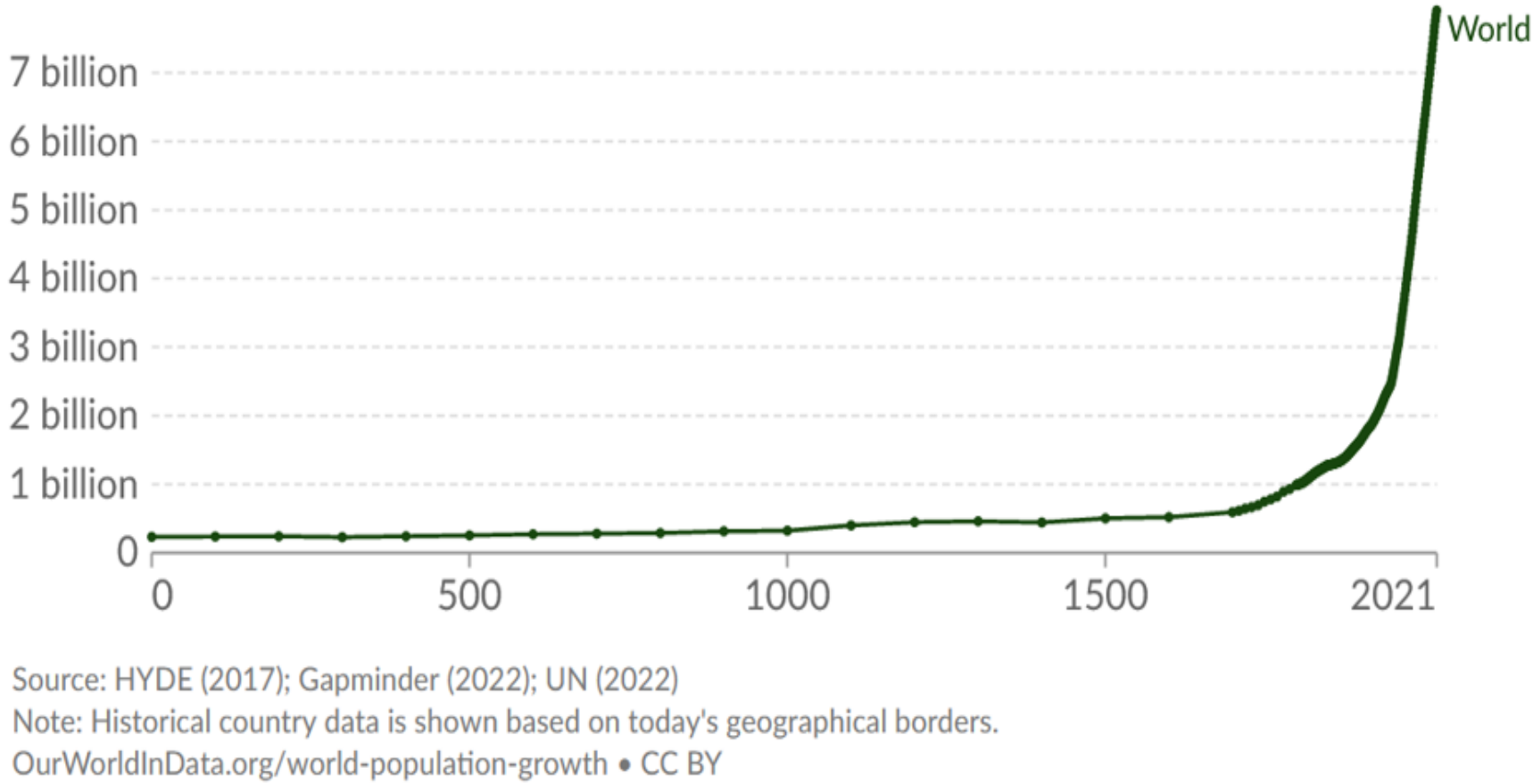
Updated on July 16, 2023 with the latest
July 2023-July 2024 estimates from the 2022 U.N. Revision

Current World Population

8,057,900,582

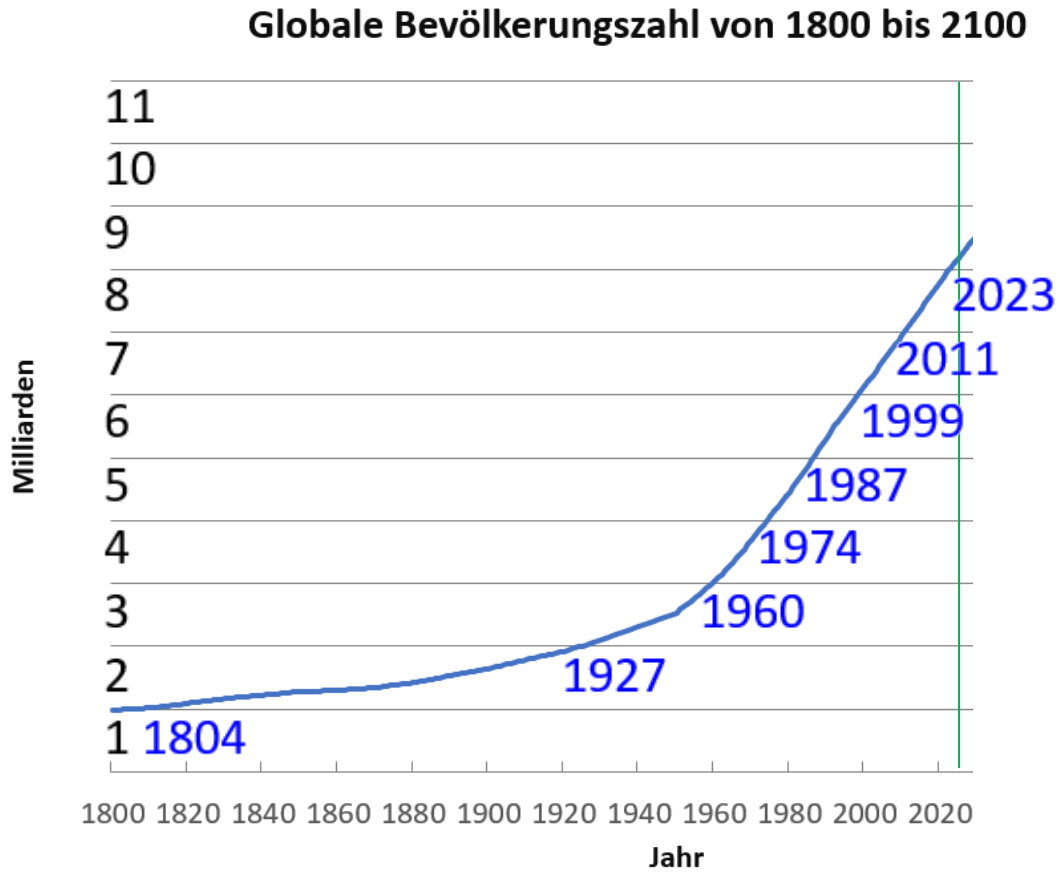
(https://countrymeters.info/de/World; Zugriff: 01.09.2023, 14:00)

Entwicklung der Weltbevölkerungszahl bis heute:



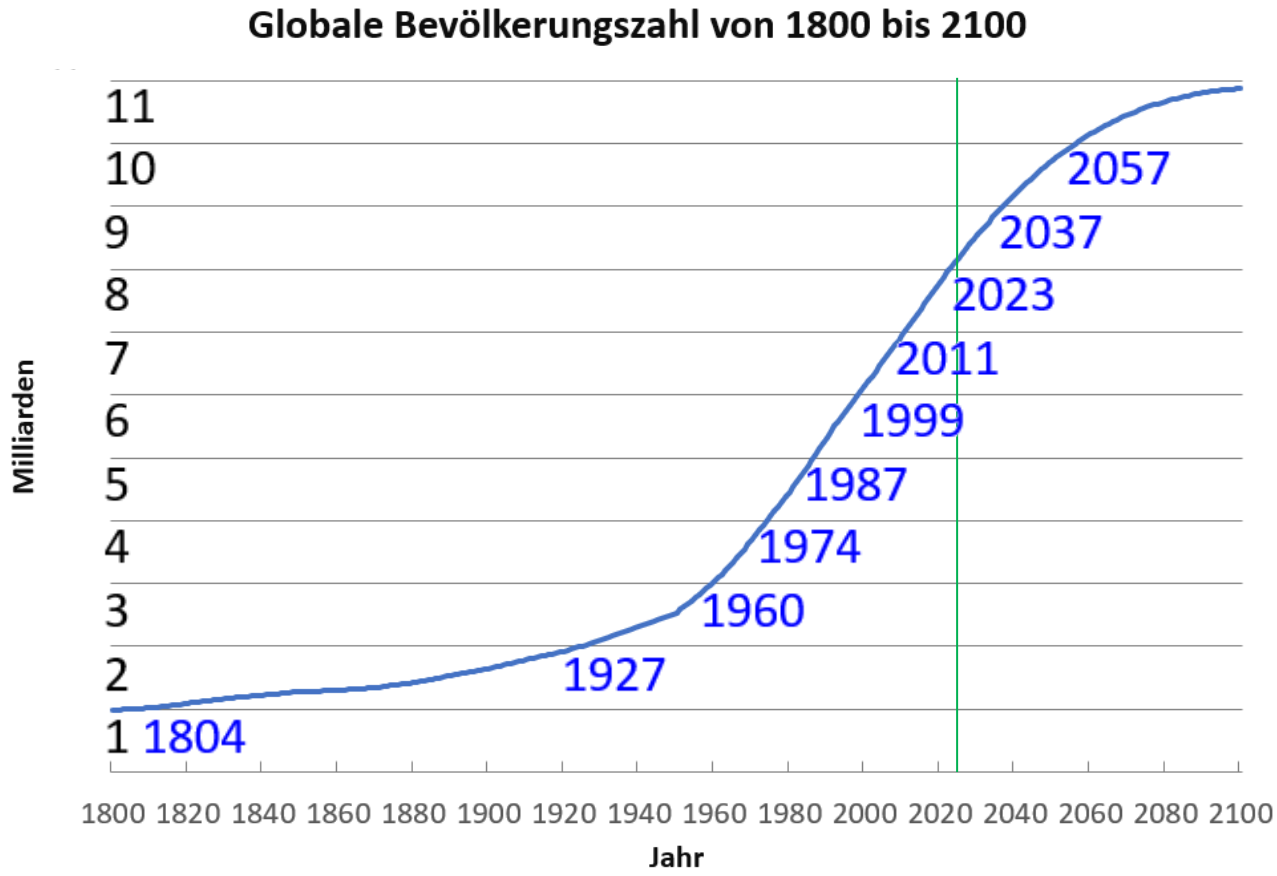
(https://ourworldindata.org/grapher/population; Zugriff: 05.09.2023)

UN-Prognosen bis 2100:



Warum wächst die Weltbevölkerung in naher Zukunft immer langsamer und stabilisert sich um das Jahr 2100 bei 11 Milliarden?

UN-Prognosen bis 2100:



Warum wächst die Weltbevölkerung in naher Zukunft immer langsamer und stabilisert sich um das Jahr 2100 bei 11 Milliarden?

Wesentliche Einflussfaktoren auf das Bevölkerungswachstum:

- Wohlstand (BIP/Kopf)
- Gesundheit (Lebenserwartung)
- Kindersterblichkeit (in den ersten fünf Lebensjahren)
- Geburtenrate (Mittelwert der Kinderzahl pro Frau)

Wohlstand und Gesundheit:

(https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1; Zugriff: 05.09.2023)

Globale Lebenserwartung:

1800: 30,5 Jahre
1900: 32,0 Jahre
1960: 49,9 Jahre
1990: 65,1 Jahre
2022: 74,0 Jahre

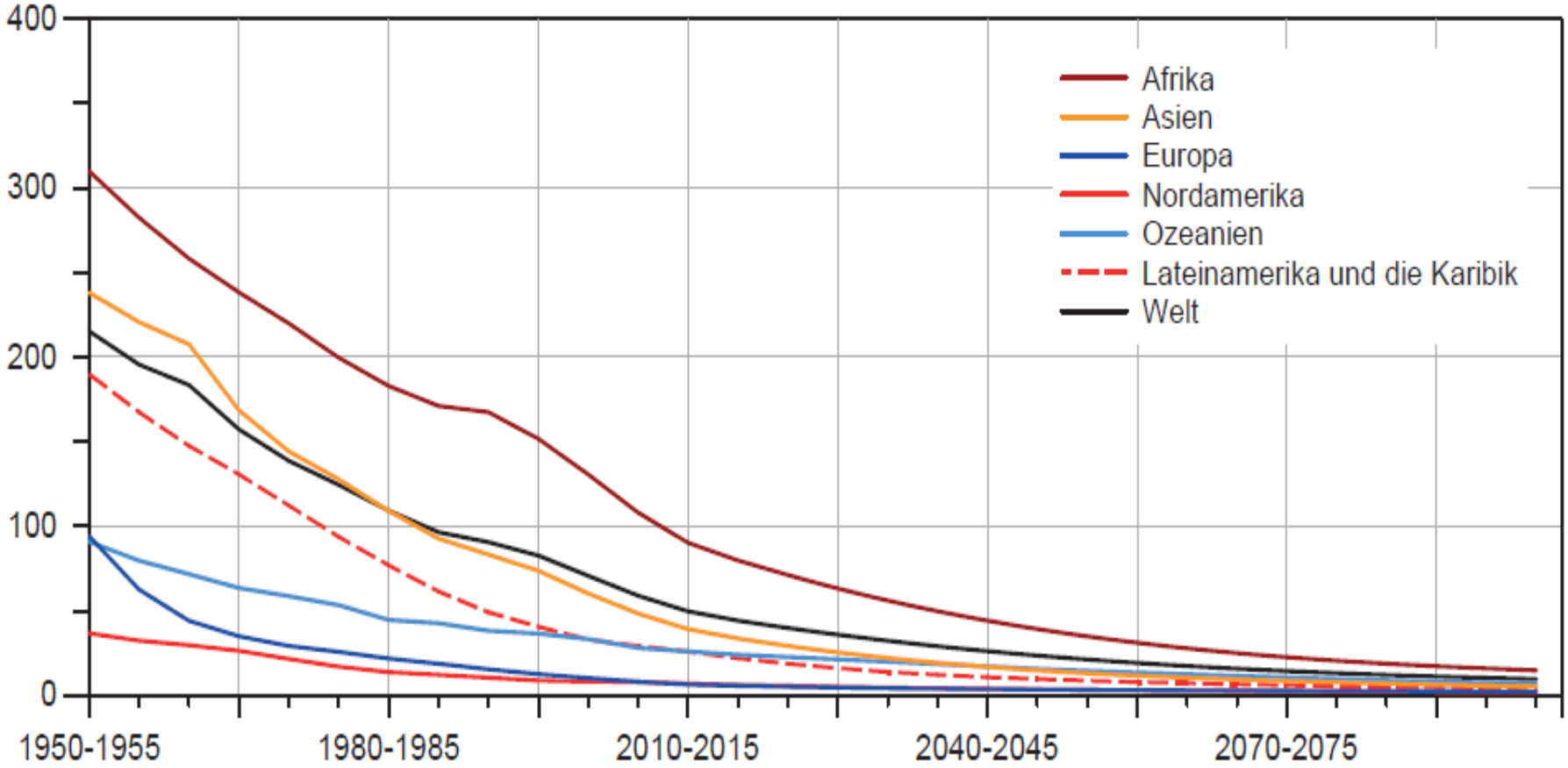
(https://www.gapminder.org/data/documentation/gd004/; Zugriff: 01.09.2023)

Weltweite Kindersterblichkeitsrate (Todesfälle unter 5 Jahre):

1800: 43,3 Prozent
1900: 36,2 Prozent
1960: 18,4 Prozent
1990: 9,3 Prozent
2021: 3,8 Prozent

(https://de.statista.com/statistik/daten/studie/915317/umfrage/weltweite-kindersterblichkeitsrate/; Zugriff: 01.09.2023)

Kindersterblichkeitsrate nach Kontinenten:



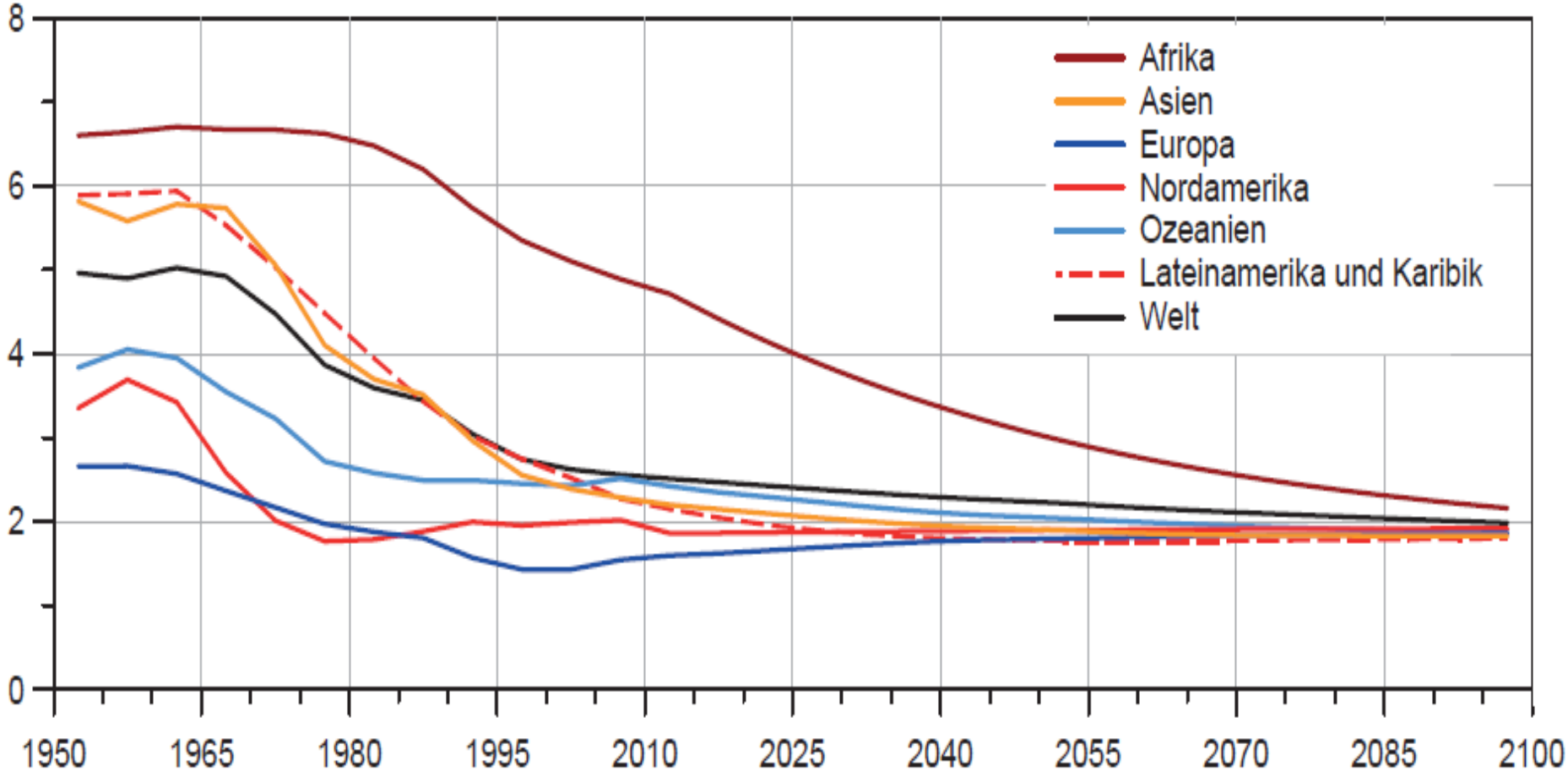
(https://www.bib.bund.de/Publikation/2018/Atlas-zur-Weltbevoelkerung.html?nn=9751912; S. 60, Zugriff: 01.09.2023)

Weltweite Geburtenrate:

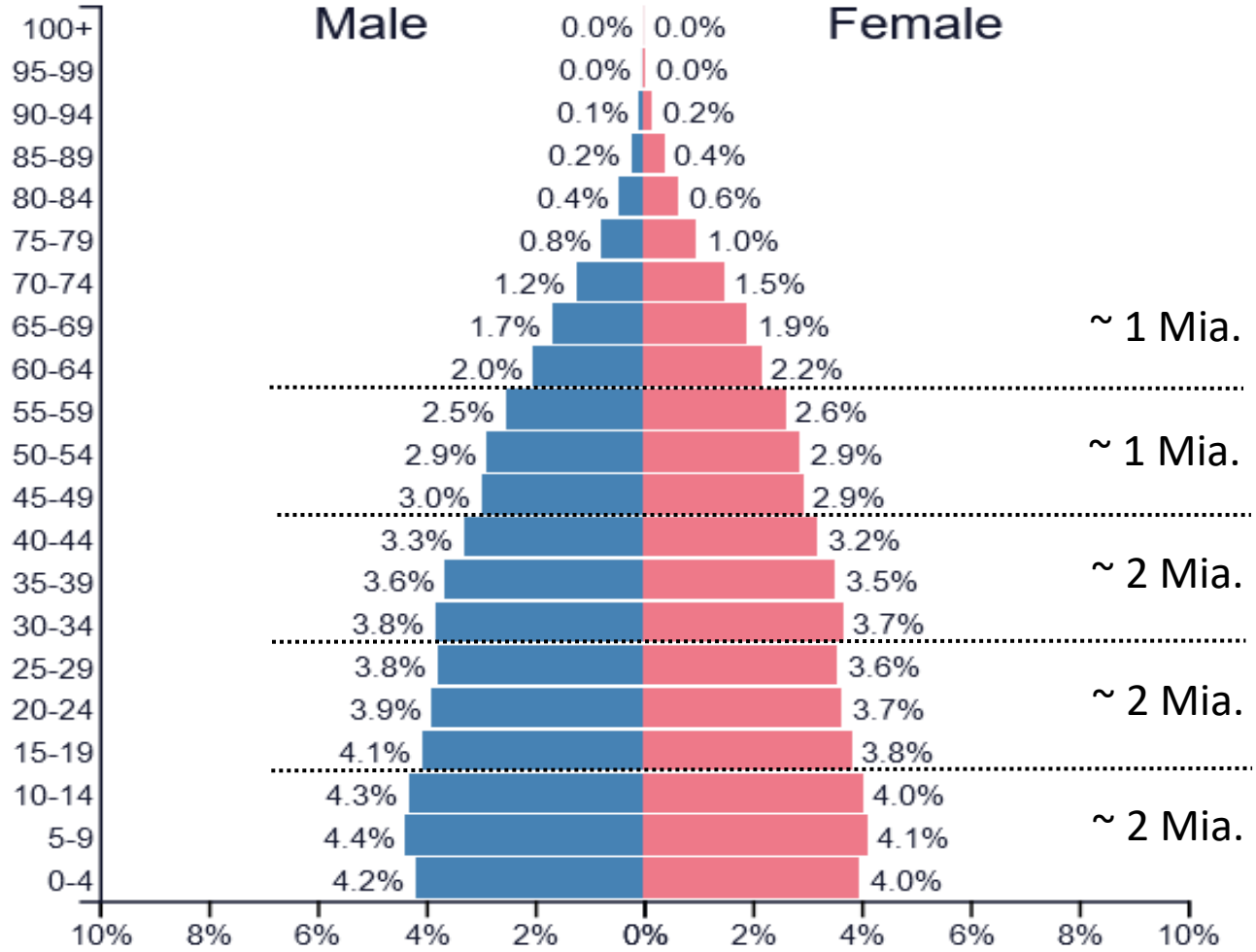
1800: 5,8
1900: 5,5
1960: 5,0
1990: 3,3
2022: 2,3 Kinder

(https://de.statista.com/statistik/daten/studie/1724/umfrage/weltweite-fertilitaetsrate-nach-kontinenten/; Zugriff: 01.09.2023)

Geburtenrate nach Kontinenten:

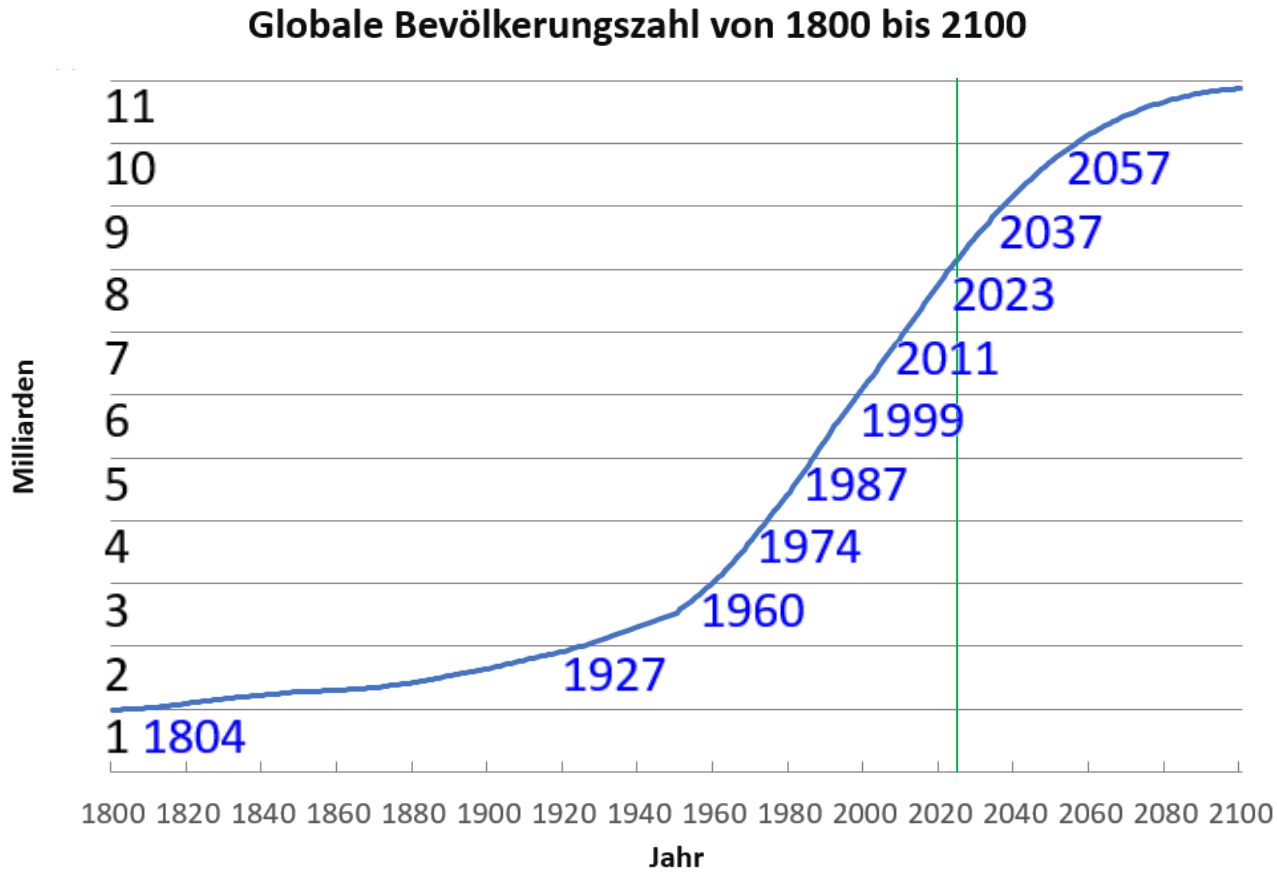


(https://www.bib.bund.de/Publikation/2018/Atlas-zur-Weltbevoelkerung.html?nn=9751912; S. 42, Zugriff: 01.09.2023)



(https://www.populationpyramid.net/; Zugriff: 01.09.2023)

UN-Prognosen bis 2100



Darum wächst die Weltbevölkerung in naher Zukunft immer langsamer und stabilisert sich um das Jahr 2100 bei 11 Milliarden!

Datenanalyse zum Klimawandel

Mittels hochkomplexer statistischer Methodik wurde unter Berücksichtigung

- von Messfehlern (früher ungenauere Messungen) und
- fehlenden Messungen (geografisch ungleiche Geräteaufteilung)

durch Statistikexpert:innen festgestellt, dass die globale Durchschnittstemperatur seit dem Vorliegen verwendbarer Messungen (ca. 150 Jahre) um mehr als 1° angestiegen ist

Laut Klimaforscher:innen liegt die kritische Grenze der Erwärmung bei einem Anstieg um 2° (mit Folgen wie steigende Meeresspiegel, anhaltende Dürre- bzw. Regenperioden)

In Österreich wurde der wärmste Silvestertag der 256-jährigen Messgeschichte registriert, der vergangene Oktober war der wärmste, der Sommer 2023 war so heiß wie nie, 18 der wärmsten Jahre seit Beginn der Aufzeichnungen gab es ab der Jahrtausendwende, die letzten sieben Jahre waren global die wärmsten seit Beginn der Messungen, das letzte Jahrzehnt war das wärmste Jahrzehnt

… ein Rekord jagt den anderen

Rekorde können selbst bei konstantem Klima rein zufällig auftreten

Wie können solche Fragestellungen untersucht werden ?

Gesetzmäßigkeiten von Rekorden am Beispiel des wiederholten Werfens einer riesigen Anzahl an Würfeln:

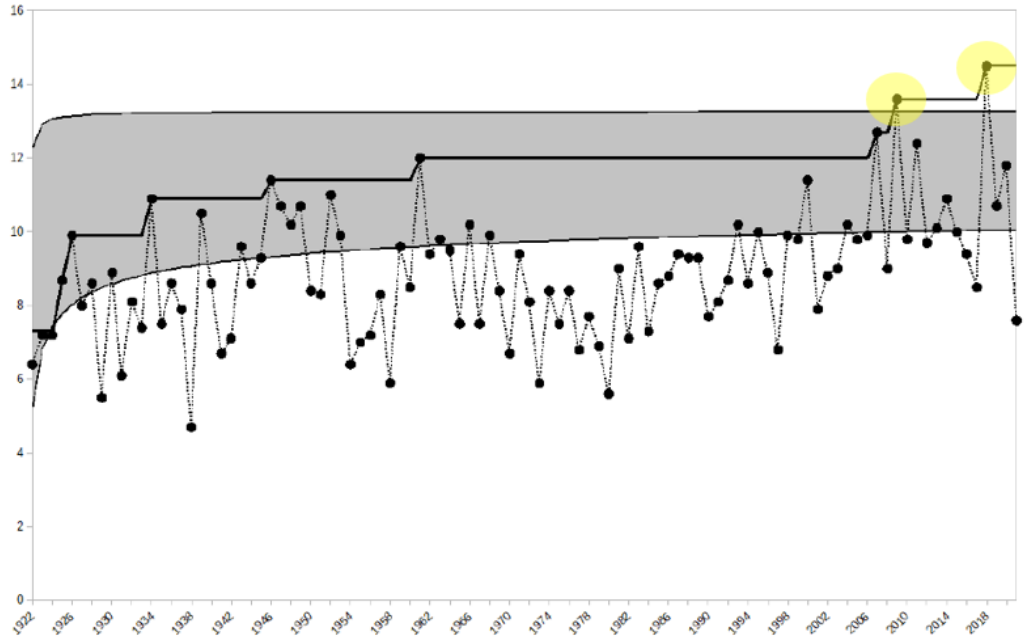
Neuer Rekord der Augensumme ist beim dritten Versuch weniger wahrscheinlich als beim Zweiten, beim Vierten weniger wahrscheinlich als beim Dritten usf.

Rekorde werden unter gleichen Bedingungen immer un*wahrscheinlicher*, wenngleich nicht un*möglich*, ihre Häufigkeit muss aber abnehmen

▶ Klassische Fragestellung der schließenden Statistik: Sind die Beobachtungen unter einer bestimmten Annahme (gleiche Würfel, konstantes Klima) so unwahrscheinlich, dass sie als starkes Indiz gegen diese Annahme zu interpretieren sind?

Vergleich des zu erwartenden Auftretens von Rekorden bei konstantem Klima mit ihrem tatsächlichen Auftreten

Entwicklung der Temperaturrekorde z. B. der Aprildurchschnittstemperaturen in den letzten 100 Jahren in Kremsmünster:



(Müller, Quatember, Waldl (2023). *When things get extreme: Records and crashes*)

Weitere Anwendungen der Extremwerttheorie: Aufdecken von dopingbedingten Sportrekorden, Bau von ausreichend hohen Überschwemmungsschutzdämmen oder Wellenbrechern, in der Finanz- oder Versicherungsmathematik (Risikomanagement), …

Know your status, but also know your probability

„Know your status" ist eine Kampagne zur Förderung des Bewusstseins für HIV durch ein Bevölkerungsscreening: „Einen HIV-Test zu machen und damit den eigenen Immunstatus zu kennen ist der erste Schritt zur Eindämmung von HIV/AIDS"

(https://lifeplus.org/know-your-status-uequalsu/; Zugegriffen: 06.09.2023)

Für den HIV-Schnelltest „rapid point-of-care" sind die bedingten Wahrscheinlichkeiten korrekter Testentscheidungen

- für Infizierte (Sensitivität): 99,6 % [= P(+|inf.)]
- für Nicht-Infizierte (Spezifität): 99,3 % [=P(–|noninf.)]

▶ *"All HIV Tests are very accurate."*

(https://www.catie.ca/client-publication/i-know-my-hiv-status-facts-about-hiv-testing; Zugriffe: 06.06.2023)

Angegeben ist also die Wahrscheinlichkeit für ein positives Testresultat, wenn jemand infiziert ist: P(+|inf.) = 0,996

Für Schnell-Getestete ist die entscheidende Frage aber: *Wie groß ist die Wahrscheinlichkeit für eine Infektion, wenn das Testresultat positiv ist: P(inf.|+) = ?*

▶ *„What does the result mean? If you test HIV positive, it means that you have HIV."*

(https://www.catie.ca/en/hivtesting; Zugegriffen: 23.04.2021)

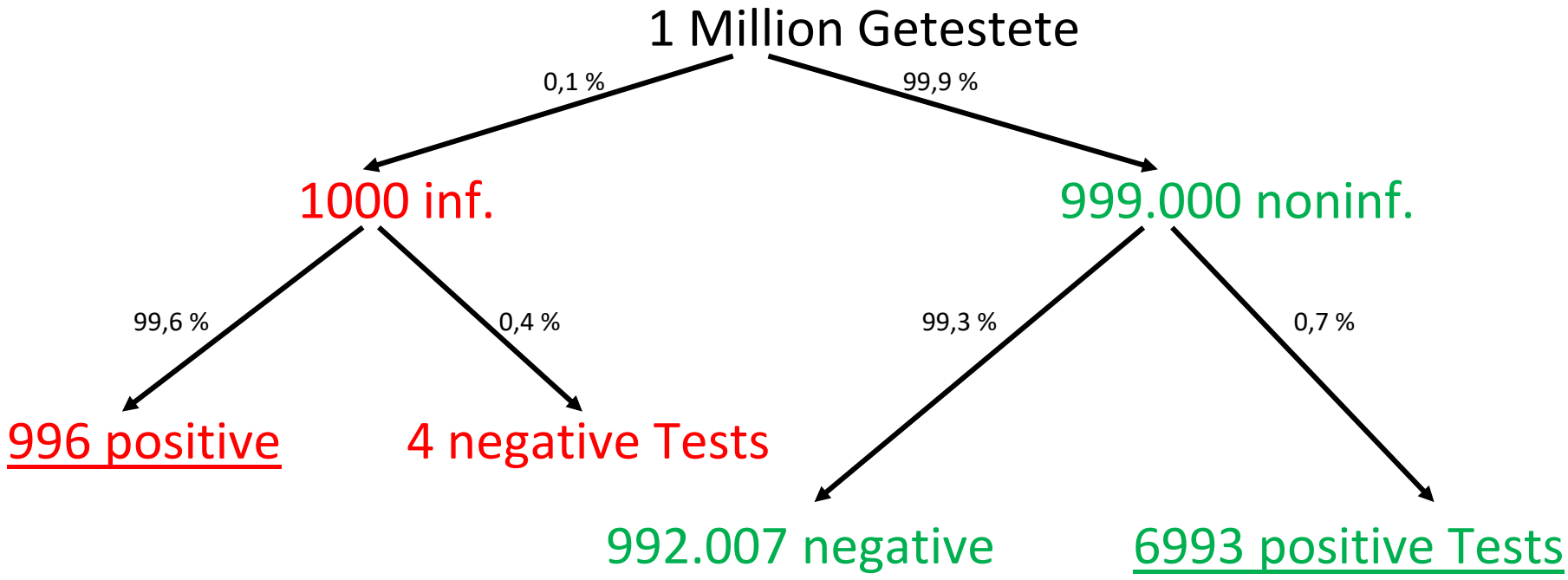
In einem Bevölkerungsscreening mit P(+|inf.) = 0,996 und P(+|non-inf.) = 0,007 und einer Prävalenz von 0,1 % [P(inf.) = 0,001], kann man P(inf.|+) nach der Regel von Bayes berechnen:

$$P(\text{inf.}|+) = \frac{P(+|\text{inf.}) \cdot P(\text{inf.})}{P(+|\text{inf.}) \cdot P(\text{inf.}) + P(+|\text{non-inf.}) \cdot P(\text{non-inf.})}$$

$$= \frac{0.996 \cdot 0.001}{0.996 \cdot 0.001 + 0.007 \cdot 0.999}$$

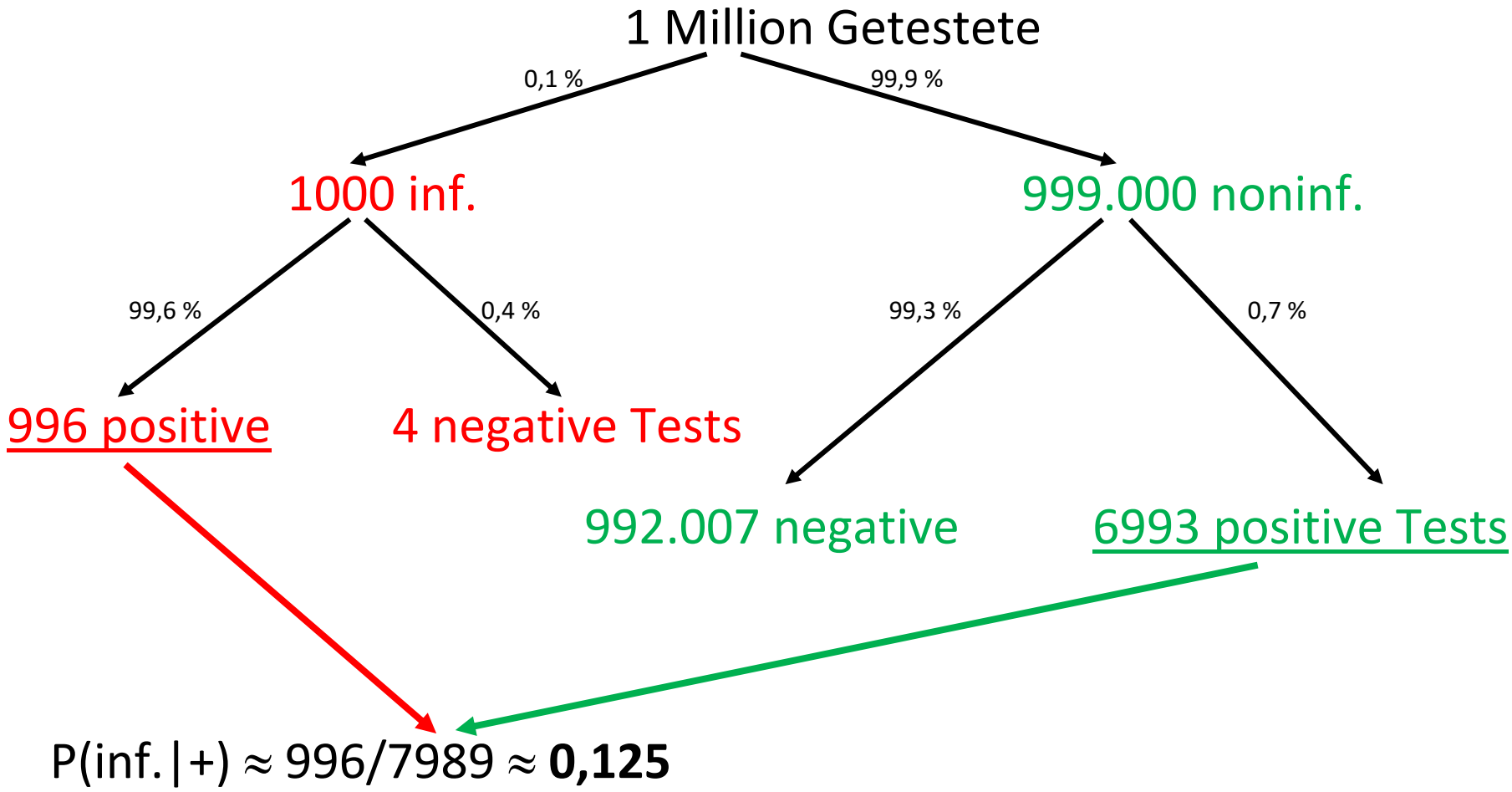
$$\approx 0.125$$

Mit P(+|inf.) = 0,996 und P(–|noninf.) = 0,993 gilt in einem Screening, wenn 0,1 % der Population infiziert sind [=P(inf.)]:



P(inf.|+) ≈

Mit P(+|inf.) = 0,996 und P(–|noninf.) = 0,993 gilt in einem Screening, wenn 0,1 % der Population infiziert sind [=P(inf.)]:



Im Screening steigert sich die Wahrscheinlichkeit für eine Infektion von 0,1 Prozent *ohne* auf 12,5 Prozent *mit* einem positiven Schnelltest

▶ *„If you test HIV positive, it means that you have HIV"* ?

Diese Botschaft, P(inf.|+) ≈ **0,125** (trotz hoher Testsensitivität und -spezifität), muss Schnelltestenden *vor* dem Test mitgeteilt werden, um z. B. Verzweiflungstaten zu vermeiden

Solche Berechnungen sind bei beinahe allen diagnostischen Tests durchzuführen:

Test method	Prevalence	Sensitivity	Specifity	P(inf.|+)
SARS-CoV-2 Rapid Antigen Test (Roche)	0.002	0.955	0.992	0.193
Mammography	0.008	0.9	0.93	0.094
Colon cancer blood test	0.003	0.5	0.97	0.048
PSA-Prostata-Test	0.08	0.808	0.384	0.102
Blood test on Down-syndrome	0.002	1	0.999	0.667

(https://diagnostics.roche.com/global/en/products/params/sars-cov-2-rapid-antigen-test.html; Retrieved on 24.05.2023; Gigerenzer, G. (2016). *Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken*. 2. Auflage. München: Piper Verlag; Paul, R., Breul, J., und Hartung, R. (1995). Sensitivität, Spezifität und positiver Vorhersagewert von PSA, PSA-Density, digital rektaler Untersuchung und transrektalem Ultraschall zur Früherkennung des Prostatakarzinoms. *Aktuelle Urologie* 26, 164-169)

Zweifelhafte Repräsentativität

Das Schlussfolgern auf Basis eines bewussten Auswählens des Teils, der für das Ganze stehen soll, wird als *Stichprobenmethode* bezeichnet

Alltägliche Anwendungen: Speisen abschmecken, Weine verkosten, Parfüms testen, Blut untersuchen, Prüfungen ablegen, …

Repräsentativität … Ähnlichkeit der Probe zum Ganzen

Der Geschmack aller möglichen „gut verrührten“ Kostproben schwankt dennoch leicht

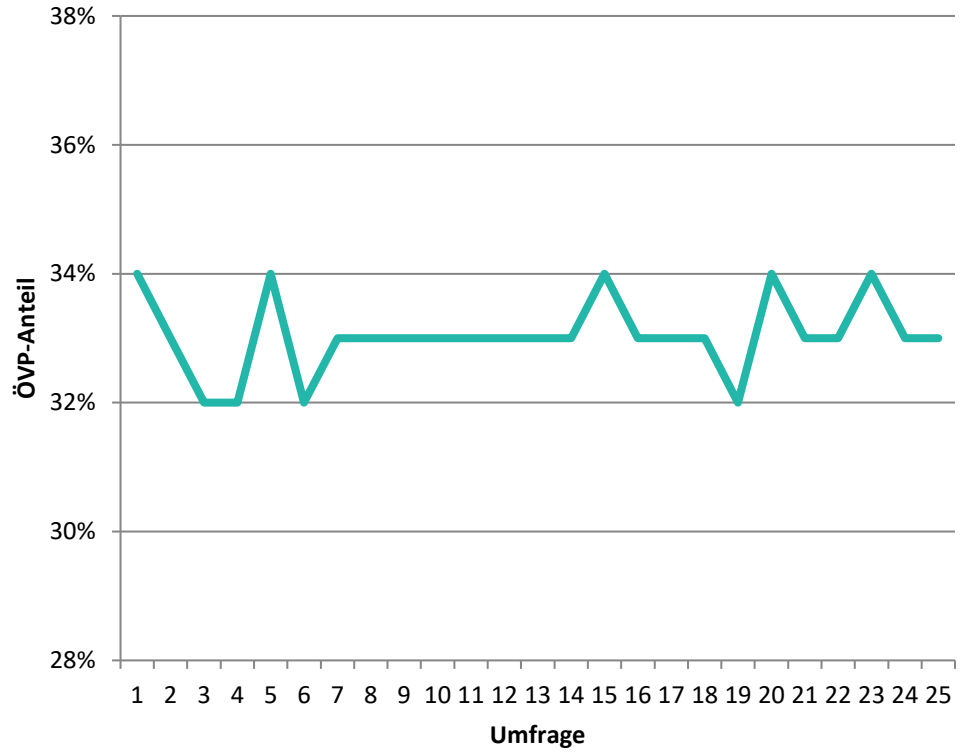
Auch die Ergebnisse von Zufallsstichproben unterliegen einer natürlichen, berechenbaren Stichprobenschwankung.

Bei n = 400 und einem Populationsanteil π = 0,33 (= 33 Prozent) muss der Stichprobenanteil in einer *einfachen Zufallsauswahl* z. B. mit einer Wahrscheinlichkeit von $1-\alpha$ = 0,95 zwischen 28,4 und 37,6 Prozent liegen:

$$\pi \pm u_{1-\alpha/2} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = 0{,}33 \pm 0{,}046$$

So stark schwanken Stichprobenergebnisse eben!

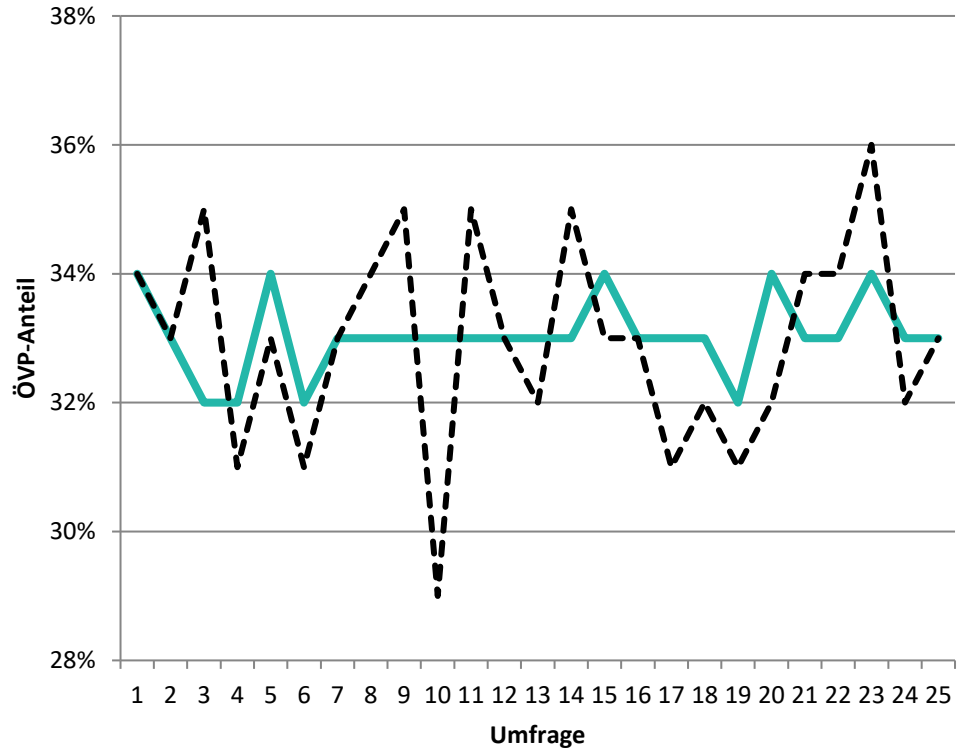
Zeitreihe der Stichprobenanteile für die ÖVP in den letzten 25 „unabhängigen“ Umfragen (Wahl 2017):



(https://neuwal.com/wahlumfragen; Zugegriffen: 27.04.2021)

Möglicherweise ist der *Populations*anteil konstant geblieben, *Stichproben*ergebnisse müssen dennoch schwanken!

Zeitreihe der Stichprobenanteile für die ÖVP in den letzten 25 „unabhängigen“ Umfragen (Wahl 2017) und eine korrekt schwankende Zeitreihe:



(https://neuwal.com/wahlumfragen; Zugegriffen: 27.04.2021)

Möglicherweise ist der *Populations*anteil konstant geblieben, *Stichproben*ergebnisse müssen dennoch schwanken!

Das JKU-Bachelorstudium *Statistik und Data Science*

Daten werden als der neue Treibstoff unserer Informationsgesellschaft bezeichnet

Expert:innen in Statistik und Data Science werden am Arbeitsmarkt stark nachgefragt

Ausbildungsziel ist eine umfassende **Data Literacy** als Schlüsselkompetenz im 21. Jahrhundert

Ausbildungsschwerpunkte:

- Datengewinnung (*Wie komme ich zu relevanten Daten?*)
- Spannungsfeld Datenqualität/Datenquantität (*Was habe ich bei Big Data-Analysen zu beachten?*)
- Datenverwaltung (*Wie organisiere ich die Daten auf effiziente Weise?*)
- Datenanalyse (*Welche statistischen Methoden sind für die jeweilige Fragestellung passend?*)
- Datenverarbeitung (*Kann ich vorhandene Software verwenden oder soll ich den Prozess selbst programmieren?*)
- Ergebnisvermittlung (*Wie vermittle ich Nichtexpert:innen anschaulich die Analyseergebnisse?*)

STATISTIK UND DATA SCIENCE

Vorstellung des Studiums um 11:30 Uhr im Hörsaal 3