Was ist „applied statistics"?

Applied Statistics for Data Science

Peter Büchel

HSLU I

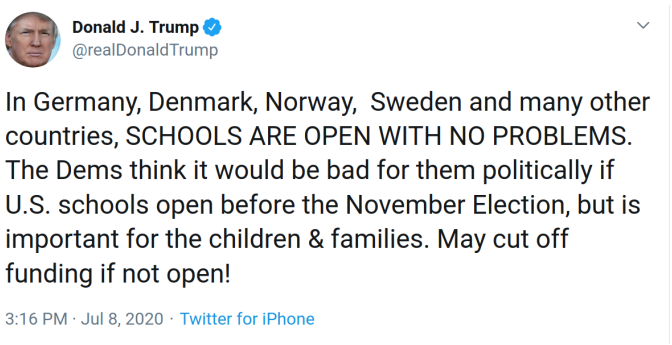
9. Februar 2021

Einführung

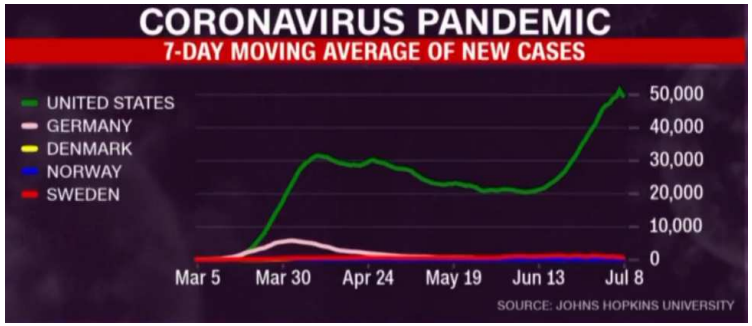
- Angewandte Statistik: Anwendung der Statistik auf reale Alltagsprobleme
- Illustration mit einem Beispiel: Thematisch (*kritisches Denken*) etwas ausserhalb dieses Moduls
- Vorteil dieses Beispiels: Es werden keine statistischen Vorkenntnisse vorausgesetzt

Beispiel

- Folgende Daten: Vom 8. und 9. Juli 2020
- Es begann, wie so vieles andere auch, mit einem Tweet von Donald Trump:

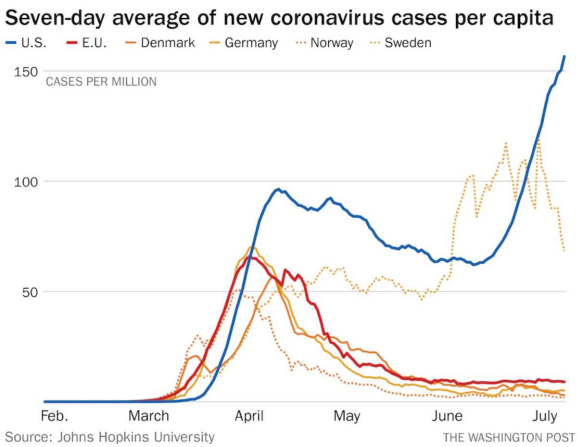


- Trumpf wollte wegen Covid 19 geschlossene Schulen wieder öffnen
- Begründete Öffnung mit: Deutschland, Dänemark, Norwegen und Schweden hatten dies ebenfalls getan
- CNN: Wollte zeigen, dass der Vergleich irreführend ist (ist er)
- Aber: Die Argumentation von CNN war eher unsinnig
- Begründung mit Grafik:



- Argumentation von CNN:
  - In Europa sind Fallzahlen viel niedriger als in den USA
  - Macht Sinn, dass europäische Länder Schulen wieder öffnen, da Fallzahlen niedrig
  - Macht für USA keinen Sinn, weil Fallzahlen sehr hoch
- Plausibel, aber absurd: Vergleich von Äpfeln mit Birnen
- Zahlen in obiger Abbildung: *Absolute* Zahlen
- Aber: Norwegens Bevölkerung etwa 5.5 Millionen; USA 330 Mio
- Kurve von Norwegen *muss* viel niedriger und flacher als USA sein
- Auch wenn Kurve Norwegens *relativ* (prozentual) ähnlich wie Kurve der USA aussieht, sieht sie in *absoluten* Zahlen nicht viel anders aus als die in der obigen Abbildung

- Basierend auf denselben Daten: Grafik in Zeitung *Washington Post*:



- Fallzahlen: Pro Million Einwohner
- Macht mehr Sinn: Kurven können miteinander verglichen werden

- *Jetzt*: Deutschland, Norwegen und Dänemark sind in Bezug auf *relative* Fallzahlen in viel besserer Verfassung als USA
- Schweden nicht: Ähnliche Kurve wie USA
- Letzte Beobachtung überhaupt nicht offensichtlich in Abb. Folie 4
- Wiedereröffnung von Schulen: Sinnvoll in Deutschland, Norwegen und Dänemark
- Vielleicht nicht so in den USA und Schweden

- Aber: Auch relative Fallzahlen oder Todeszahlen können problematisch sein
- Tabelle: Absolute Todeszahlen vom 9. September:

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
	World	27,523,031	+42,072	897,625	+1,069	19,624,327	7,001,079	60,296	3,531	115.2			
1	USA	6,486,426	+851	193,586	+52	3,758,629	2,534,211	14,589	19,575	584	88,067,850	265,771	331,367,517
2	Brazil	4,147,794		127,001		3,355,564	665,229	8,318	19,488	597	14,408,116	67,694	212,842,596
3	India	4,284,103	+6,519	72,843	+27	3,324,060	887,200	8,944	3,099	53	50,650,128	36,636	1,382,530,286
4	Mexico	637,509	+3,486	67,781	+223	446,715	123,013	2,836	4,935	525	1,435,703	11,114	129,184,522
5	UK	350,100		41,554		N/A	N/A	69	5,152	612	17,619,897	259,295	67,953,144
6	Italy	278,784		35,553		210,238	32,993	142	4,612	588	9,271,810	153,393	60,444,825
7	France	328,980		30,726		87,836	210,418	537	5,038	471	8,500,000	130,167	65,300,897
8	Peru	691,575		29,976		522,251	139,348	1,488	20,921	907	3,386,625	102,450	33,056,487
9	Spain	525,549		29,516		N/A	N/A	1,034	11,240	631	9,987,326	213,595	46,758,226
10	Iran	391,112	+2,302	22,542	+132	337,414	31,156	3,713	4,645	268	3,431,646	40,760	84,191,615
11	Colombia	671,848		21,615		529,279	120,954	863	13,178	424	2,964,722	58,150	50,983,652
12	Russia	1,035,789	+5,099	17,993	+122	850,049	167,747	2,300	7,097	123	38,758,184	265,565	145,946,376
13	South Africa	639,362		15,004		566,555	57,803	539	10,755	252	3,808,949	64,073	59,446,940
14	Chile	424,274		11,652		395,717	16,905	930	22,159	609	2,641,589	137,965	19,146,844
15	Ecuador	110,092		10,576		91,242	8,274	424	6,223	598	330,998	18,709	17,692,261
16	Argentina	488,007		10,179	+50	366,590	111,238	2,698	10,779	225	1,412,149	31,192	45,273,109
17	Belgium	88,769	+402	9,909	+2	18,576	60,284	52	7,653	854	2,449,055	211,141	11,599,139
18	Germany	254,168	+543	9,407	+2	227,000	17,761	223	3,032	112	12,383,035	147,708	83,834,622
19	Canada	132,142		9,146		116,459	6,537	54	3,495	242	5,841,880	154,531	37,803,923
20	Indonesia	200,035	+3,046	8,230	+100	142,958	48,847		730	30	2,484,807	9,067	274,061,093

- Länder erste Spalte: Geordnet nach absoluten Zahlen in zweiter Spalte
- Sollte mit letzter Kolumne Population verglichen werden

- Tabelle: Nach *relativen* Todeszahlen (4. Spalte von rechts) geordnet

#	Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop	Population
1	San Marino	716		42		660	14	1	21,093	1,237	6,865	202,239	33,945
2	Peru	691,575		29,976		522,251	139,348	1,488	20,921	907	3,386,625	102,450	33,056,487
3	Belgium	88,769	+402	9,909	+2	18,576	60,284	52	7,653	854	2,449,055	211,141	11,599,139
4	Andorra	1,261		53		934	274	3	16,316	686	137,457	1,778,504	77,288
5	Spain	525,549		29,516		N/A	N/A	1,034	11,240	631	9,987,326	213,595	46,758,226
6	UK	350,100		41,554		N/A	N/A	69	5,152	612	17,619,897	259,295	67,953,144
7	Chile	424,274		11,652		395,717	16,905	930	22,159	609	2,641,589	137,965	19,146,844
8	Bolivia	121,604	+835	7,054	+46	73,150	41,400	71	10,391	603	253,647	21,675	11,702,383
9	Ecuador	110,092		10,576		91,242	8,274	424	6,223	598	330,998	18,709	17,692,261
10	Brazil	4,147,794		127,001		3,355,564	665,229	8,318	19,488	597	14,408,116	67,694	212,842,596
11	Italy	278,784		35,553		210,238	32,993	142	4,612	588	9,271,810	153,393	60,444,825
12	USA	6,486,426	+851	193,586	+52	3,758,629	2,534,211	14,589	19,575	584	88,067,850	265,771	331,367,517
13	Sweden	85,707		5,838	+4	N/A	N/A	13	8,477	577	1,124,269	111,192	10,111,092
14	Mexico	637,509	+3,486	67,781	+223	446,715	123,013	2,836	4,935	525	1,435,703	11,114	129,184,522
15	Panama	97,578		2,099		70,247	25,232	149	22,550	485	369,420	85,371	4,327,225
16	France	328,980		30,726		87,836	210,418	537	5,038	471	8,500,000	130,167	65,300,897
17	Sint Maarten	516		19		321	176	7	12,009	442	2,450	57,022	42,966
18	Colombia	671,848		21,615		529,279	120,954	863	13,178	424	2,964,722	58,150	50,983,652
19	Netherlands	76,548	+964	6,244	+1	N/A	N/A	45	4,466	364	1,648,103	96,144	17,142,061
20	Ireland	29,774		1,777		23,364	4,633	7	6,017	359	906,432	183,191	4,948,020

- Beide Ranglisten sehen völlig unterschiedlich aus
- Nie gehört: San Marino am schlimmsten von Covid 19 betroffen
- Wie kann das sein?

- Letzte Spalte: Länder mit sehr kleinen Ländern in Bezug auf die Bevölkerung
- Wenige absolute Fälle haben grossen Einfluss auf relative Todeszahl
- Erste Tabelle: Grosse oder sehr grosse Länder in Bezug auf Bevölkerung

- Gerade gemachte Beobachtung kommt sehr oft bei angewandten Problemen vor:
  - Relative Fallzahlen klären Argumentation von CNN
  - Aber: Relative Fallzahlen sind nicht immer besser oder aussagekräftiger als absolute
- Für angewandte Probleme: *Es gibt kein Rezept, wie das Problem gelöst werden soll*

Problemlösung in der angewandten Statistik

- Beispiel oben sehr einfach
- Enthält aber viele Aspekte, die für die Lösung von Problemen in der angewandten Statistik relevant sind
- Erstens: Unklar, was die Frage oder das Problem ist
  - Begonnen mit dem Trump-Tweet
  - Und jetzt?
  - CNN, im Allgemeinen Anti-Trump, hielt es für erwähnenswert
- Zweitens: Es ist nicht klar, wie die Lösung aussehen soll
  - *Wie* reagiert man auf einen solchen Tweet?
  - CNN: Argumentation auf der Basis von Fallzahlen

- Drittens: Es ist nicht klar, welche Elemente für die Lösung verwendet werden sollen
  - CNN beschloss, absolute Fallzahlen anstelle der relativen Fallzahlen zu verwenden
- Viertens: Es ist nicht klar, wie das Ergebnis zu interpretieren ist
  - Häufig der schwierigste Punkt zur Lösung von Problemen in der angewandten Statistik
  - CNN ging in die falsche Richtung: Verglich Dinge, die nicht vergleichbar sind

Beispiel aus der Schulmathematik

- Löse folgende Gleichung nach x auf:

$$2x + 1 = 5$$

  - Problem klar
  - Lösung ist klar (oft geübt)
  - Nichts zu diskutieren oder zu interpretieren über Lösung $x = 2$
- *Diese Art von Problemen gibt es in der angewandten Statistik oder Wissenschaft nicht*

Alltagsprobleme (aus Eric Mazur; Principle & Practice in Physics)

- Sie haben es eilig, irgendwo hinzukommen, aber Sie können Ihre Autoschlüssel nicht finden
- Ihnen geht das Mehl beim Backen eines Geburtstagskuchens aus und der Supermarkt ist geschlossen
- Ihr Flug wurde auf dem Weg zu einem Vorstellungsgespräch gestrichen
- Sie wollen diese tollen neuen Schuhe kaufen, aber es ist kein Geld auf der Bank

- Für all diese Probleme gibt es keine Lösungsanweisung
- Kann auch nicht durch Formeln gelöst werden
- Vier-Schritte-Problemlösungsstrategie aus Mazur
- Die meisten statistischen Probleme sind in Worten formuliert

Erstens: Erste Schritte

- Es ist nicht klar, welches der effizienteste Weg ist, um auf ein bestimmtes Problem zu reagieren
- Der erste Schritt unserer Problemlösungsstrategie, der Anfang, ist oft der schwierigste
- Es ist daher sinnvoll, mit etwas zu beginnen, das Sie tun *können*:
  - Organisieren Sie die gegebenen Informationen und vergewissern Sie sich, dass Ihnen klar ist, was genau in dem Problem erforderlich ist
  - Stellen Sie sicher, dass Ihnen klar ist, welche Informationen in dem Problem enthalten sind
  - Formulieren Sie das Problem mit Ihren eigenen Worten
  - Schliesslich stellen Sie fest, ob Sie alle Informationen haben oder nicht, die zur Lösung des Problems notwendig

Zweitens: Plan erstellen

- Der nächste Schritt ist die Ausarbeitung eines Plans zur Lösung Ihres Problems, d.h. herauszufinden, was Sie tun müssen, um das Problem zu lösen
- Ein guter Plan ist es, die Schritte aufzuzeigen, die Sie unternehmen müssen, um eine Lösung zu finden

Drittens: Plan ausführen

- Sie führen Ihren Plan aus, indem Sie die von Ihnen umrissenen Schritte befolgen

Viertens: Resultat interpretieren

- Sie denken vielleicht, dass Sie fertig sind
- Aber es gibt einen letzten - und sehr wichtigen - Schritt: Interpretieren Sie Ihre Antwort
  - Prüfen Sie, ob Ihr Ergebnis überhaupt möglich ist
  - Wenn zum Beispiel eine Wahrscheinlichkeit negativ wird, dann muss etwas schief gelaufen sein
  - Interpretieren Sie das Ergebnis in den Worten des Problems

- Schulmathematik: Der 3. Punkt ist oft der wichtigste
- In der angewandten Statistik: Nicht so wichtig, Berechnungen werden mit R gemacht
- Punkt 4 ist der wichtigste Punkt
- Beispiel R-Output (was der Code bewirkt, ist jetzt nicht wichtig)

```
x <- c(1, 4, 6, 8)

t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 3.1814, df = 3, p-value = 0.05004
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.00151774  9.50151774
## sample estimates:
## mean of x
##      4.75
```


- Diese Ausgabe muss interpretiert werden
- Alle 4 Punkte müssen korrekt ausgeführt werden
- Selbst wenn der 4. Punkt der wichtigste ist, nützt es nichts, ihn richtig zu machen, wenn in den ersten 3 Schritten ein Fehler gemacht wurde
- Die Anwendung dieser vier Punkte erscheint bei sehr einfachen Aufgaben überflüssig
- Aber im weiteren Verlauf dieses Moduls, wenn die Probleme komplexer werden, bieten sie sehr gute Anhaltspunkte für die Lösung dieser Probleme
- Wir wissen aus Erfahrung, dass es für die Studierenden schwierig ist, aus der Problemdefinition herauszulesen, was tatsächlich gefordert wird

Was ist Statistik, die nicht angewandt ist?

- Angewandte Statistik: Verfahren und Methoden werden verwendet und beschrieben
- Kann oft auf einfache Weise erklärt werden
- Aber was nicht getan wird, warum diese Verfahren und Methoden genau das tun, was sie tun sollten
- Obwohl das Prinzip oft einfach ist, sind die Details schwierig

- Details werden in *mathematischer Statistik* nachgewiesen und das sieht aus wie

$$S^2 := \frac{1}{m+n-2}\left\{\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{j=1}^{m}(Y_j-\bar{Y})^2\right\}.$$

Note that $\bar{X}$ has expectation μ and variance σ^2/n, and $\bar{Y}$ has expectation $\mu+\gamma$ and variance σ^2/m. So $\bar{Y}-\bar{X}$ has expectation γ and variance

$$\frac{\sigma^2}{n}+\frac{\sigma^2}{m}=\sigma^2\left(\frac{n+m}{nm}\right).$$

The normality assumption implies that

$$\bar{Y}-\bar{X} \text{ is } \mathcal{N}\left(\gamma,\sigma^2\left(\frac{n+m}{nm}\right)\right)\text{–distributed.}$$

Hence

$$\sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y}-\bar{X}-\gamma}{\sigma}\right) \text{ is } \mathcal{N}(0,1)\text{–distributed.}$$

To arrive at a pivot, we now plug in the estimate S for the unknown σ:

$$Z(\mathbf{X},\mathbf{Y},\gamma) := \sqrt{\frac{nm}{n+m}}\left(\frac{\bar{Y}-\bar{X}-\gamma}{S}\right).$$

- Keine Angst, so etwas werden Sie nicht sehen