Statistical Analysis of COVID-19 and Development of a Prediction Formula

Robert M. Kuhnhenn, DO, MS
August 27, 2020

1

Topics

- Data
- Statistical analysis
  - Basic statistics terminology
- Analysis of large dataset
  - Risk vs Odds
  - Prediction model

2

Data

- Data set: collection of data objects
- Data object: an instance, data point
- Represented as a set of **attributes**
  - a characteristic or feature of data object
  - dimension, feature, variable, or field
- Types of data attributes
  - Qualitative
  - Quantitative

3

Statistical Software

- Microsoft Excel
- R
- Python
- WEKA
- JMP, SAS, SPSS, many others

4

Statistical Description of Data

Data Summaries

Qualitative Data

- Class frequency
- Class relative frequency
- Graphical visualization
  - Bar plots
  - Pie charts

Quantitative Data

- Numerical summaries
  - Center
    - mean, median
  - Spread
    - range, quartiles, variance, standard deviation
- Graphical visualization
  - Boxplots
  - Histograms

5

Numerical Summaries

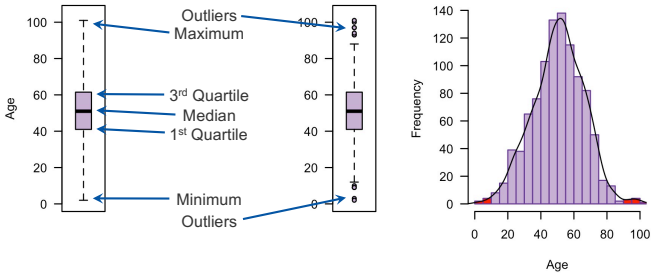
- Mean $x = \frac{x_1 + x_2 + \cdots + x_n}{n}$
- Median
- Range $= x_{max} - x_{min}$
- Quartiles: 1st, 2nd, and 3rd
  - represent 25%, 50%, and 75% of the data
  - IQR = 3rd Quartile – 1st Quartile
- Variance $s^2 = \frac{(x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2}{n - 1}$

$$= \frac{1}{n-1}\sum_{i=1}^{n} x_i - x)^2$$

- Standard deviation $s = \sqrt{s^2}$

6

Boxplots and Histograms

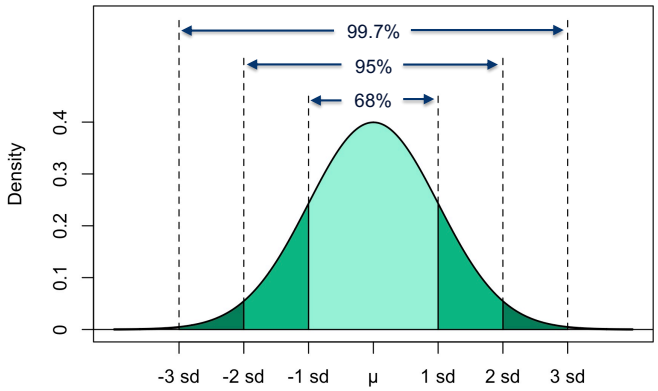


- Outliers:
  - < 1st quartile − 1.5 × IQR
  - > 3rd quartile + 1.5 × IQR

7

Normal Distribution

68-95-99.7 Rule



8

Statistical Analysis

- Statistics tests
  - *z*-statistic, *t*-statistic,
  - chi square,
  - regression models,
  - many, many others
- State hypothesis
  - $H_0: x_1 = x_2$ $H_0: x_1 \geq x_2$ $H_0: x_1 \leq x_2$
  - $H_1: x_1 \neq x_2$ $H_1: x_1 < x_2$ $H_1: x_1 > x_2$
- Choose a significance level (α) for determining statistical significance

9

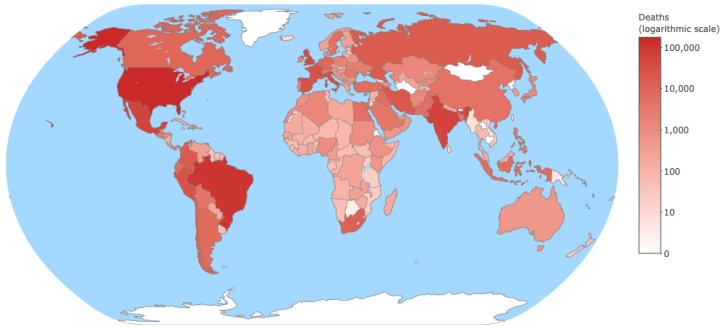
COVID-19 Deaths by Country

Robert M. Kuhnhenn, DO, MS

Data updated: August 24, 2020 4:27 AM GMT

Source: Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, Baltimore, MD
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports

Total world cases: 23,419,697
Total world deaths: 808,754



10

Mexican Government Dataset

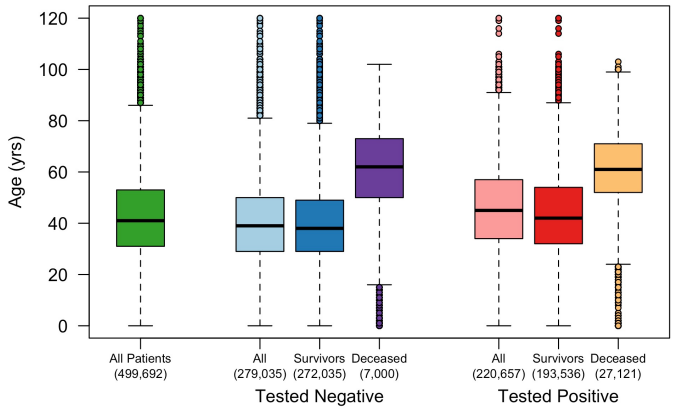
- Total rows: 566,602
- Patient type:
  - Ambulatory: 444,689 Hospitalized: 121,913
- COVID test results:
  - Negative: 279,035 Positive: 220,657 Results pending: 66,910
- Attributes:

id	sex	patient_type	entry_date
date_symptoms	date_died	intubated	pneumonia
age	pregnancy	diabetes	copd
asthma	immusupr	hypertension	other_disease
cardiovascular	obesity	renal_chronic	tobacco
contact_other_covid	covid_res	icu	

Mukherjee T. COVID-19 patient pre-condition dataset. Kaggle. Accessed August 14, 2020.
https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset

11

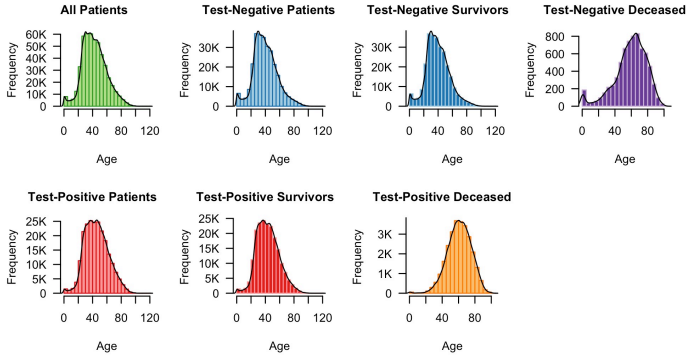
Age Distribution Boxplots



Mukherjee T. COVID-19 patient pre-condition dataset. Kaggle. Accessed August 14, 2020.
https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset

12

Histograms



Mukherjee T. COVID-19 patient pre-condition dataset. Kaggle. Accessed August 14, 2020. https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset

13

Preprocessing

- Removed several attributes:
  - pregnancy
  - patient type (ambulatory vs hospitalized)
  - other disease
  - contact with other COVID patient
  - ICU

14

Chi square test

- Test of independence of nominal variables
- Null hypothesis:
  - the variables are independent of each other
- Create contingency table
- Calculate expected values
  - $\frac{\text{sum of row} \times \text{sum of column}}{\text{Total}}$
- Compute the χ^2 test statistic
  - $\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$

Where O is the observed value,
E is the expected value,
i is the "i^{th}" position in the contingency table, and
k is the total number of pairs of expected and observed counts.

15

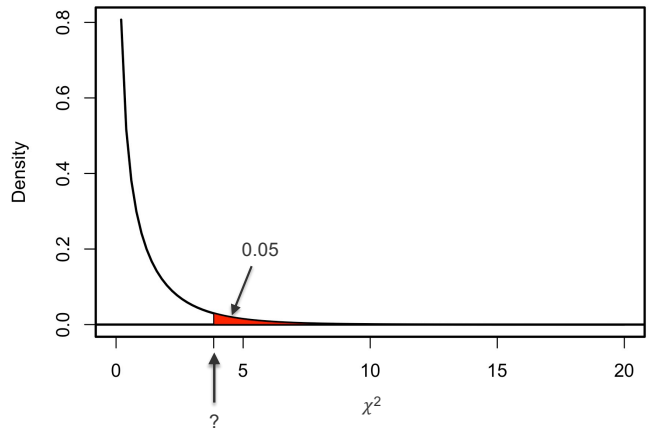
Chi square test

	Male	Female	Sum (row)
Survived	102,224 (105,221.46)	90,192 (87,194.54)	192,416
Deceased	17,631 (14,633.54)	9,129 (12,126.46)	26,760
Sum (column)	119,855	99,321	219,176

- Expected value $= \frac{192{,}416 \times 119{,}855}{219{,}176} = 105{,}221.46$
- $\chi^2 = \frac{(102224-105221.46)^2}{105221.46} + \frac{(90192-87194.54)^2}{87194.54} + \frac{(17631-14633.54)^2}{14633.54} + \frac{(9129-12126.46)^2}{12126.46}$
  $= 1{,}543.33$
- Degrees of freedom
  $= (\text{number of rows} - 1) \times (\text{number of columns} - 1) = 1$
- Look up chi square value for the specified df and α

16

Chi Square Distribution Curve



17

TABLE IV

Chi-Square (χ^2) Distribution

Area to the Right of Critical Value

Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Source: https://faculty.elgin.edu/dkernler/statistics/ch09/chi-square-table.pdf

18

Risk vs Odds

	Male	Female	Sum (row)
Survived	102,224	90,192	192,416
Deceased	17,631	9,129	26,760
Sum (column)	119,855	99,321	219,176

- Risk = $\frac{\text{event of interest}}{\text{all possible events}}$ Odds = $\frac{\text{probability of an event}}{\text{probability of event not occurring}}$
- Risk of death: Males: 17,631 ÷ 119,855 = 0.1471 = 14.71%
  Females: 9,129 ÷ 99,321 = 0.0919 = 9.19%
- Risk difference: 0.1471 − 0.0919 = 0.0552 = 5.52%
- Relative risk: 0.1471 ÷ 0.0919 = 1.60
- Similar results to Chinese study: relative risk = 2.4 (p = 0.016)[1]
- Odds of death: Males: 17,631 ÷ 102,224 = 0.172
  Females: 9,129 ÷ 90,192 = 0.101
- Odds Ratio: 0.172 ÷ 0.101 = 1.70

1. Jin J-M, Bai P, He W, et al. Gender Differences in Patients with COVID-19: Focus on Severity and Mortality. *Frontiers in Public Health*. 2020;8:152. doi:10.3389/fpubh.2020.00152

19

Logistic Regression

- Used when the dependent variable is dichotomous and there are one or more independent variables
- Two uses:
  - To predict the probability of the dependent variable based on the independent variables
  - To determine if there is a relationship between the independent variable and the dependent variable
- Similar to linear regression but uses the natural logarithm (ln) of the odds of an event and a linear combination of the explanatory variable(s)
- The base of the natural logarithm is e

  e = Euler's number, approximately equal to 2.718281828459

20

Logistic Regression

Patients that tested positive for SARS-COV-2; *n* = 219,176

	β Estimate	Pr(>|z|)	
(Intercept)	-6.2474882	< 2e-16	***
sex	0.4486531	< 2e-16	***
age	0.0524516	< 2e-16	***
pneumonia	2.2566323	< 2e-16	***
diabetes	0.3268673	< 2e-16	***
copd	0.1591509	0.000181	***
asthma	-0.1086288	0.042220	*
immsupr	0.3851192	2.7e-13	***
hypertension	0.1891916	< 2e-16	***
cardiovascular	-0.0647844	0.104911	
obesity	0.2443608	< 2e-16	***
chronic renal	0.7568419	< 2e-16	***
tobacco	-0.0688820	0.016662	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

21

Calculating Odds Ratio and Risk of Death from Logistic Regression Results

- $\mathrm{OR} = e^{\beta x}$
- $p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}$
- Example: 55 yo male, obese, + COPD, + HTN

$$p = \frac{e^{-6.2474882+(0.4486531\times 1)+(0.0524516\times 55)+\ldots(-0.0688820\times 0)}}{1+e^{-6.2474882+(0.4486531\times 1)+(0.0524516\times 55)+\ldots(-0.0688820\times 0)}}$$

$$= \frac{e^{0.09814634}}{1+e^{0.09814634}} = 0.08937456 = 8.94\%$$

22

Odds Ratios

	OR	95% Conf. Interval 2.5 %	97.5 %
sex	1.5662	1.5162	1.6179
age	1.0539	1.0527	1.0550
pneumonia	9.5509	9.2472	9.8645
diabetes	1.3866	1.3385	1.4365
copd	1.1725	1.0788	1.2744
asthma	0.8971	0.8078	0.9962
immsupr	1.4698	1.3256	1.6297
hypertension	1.2083	1.1660	1.2521
cardiovascular	0.9373	0.8667	1.0136
obesity	1.2768	1.2311	1.3242
chronic renal	2.1315	1.9760	2.2993
tobacco	0.9334	0.8823	0.9876

23

References

1. Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. *CSSEGISandData/COVID-19*.; 2020. Accessed August 26, 2020. https://github.com/CSSEGISandData/COVID-19
2. Mukherjee T. COVID-19 patient pre-condition dataset | Kaggle. Kaggle. Accessed August 14, 2020. https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset
3. Jin J-M, Bai P, He W, et al. Gender Differences in Patients with COVID-19: Focus on Severity and Mortality. *Frontiers in Public Health*. 2020;8:152. doi:10.3389/fpubh.2020.00152

Other recommended references:

1. Krousel-Wood MA, Chambers RB, Muntner P. Clinicians' guide to statistics for medical practice and research: Part i. *Ochsner J*. 2006;6(2):68-83. https://pubmed.ncbi.nlm.nih.gov/21765796
2. Krousel-Wood MA, Chambers RB, Muntner P. Clinicians' guide to statistics for medical practice and research: Part ii. *Ochsner J*. 2007;7(1):3-7. https://pubmed.ncbi.nlm.nih.gov/21603472
3. Anderson RP, Jin R, Grunkemeier GL. Understanding logistic regression analysis in clinical reports: an introduction. *The Annals of Thoracic Surgery*. 2003;75(3):753-757. doi:10.1016/S0003-4975(02)04683-0

24