Biostat 100A Lab #3

Section 1D: Wednesday 9:00 - 9:50 pm

Power Transformation, Assessing Normality

Slides Courtesy of Qingyuan Liu

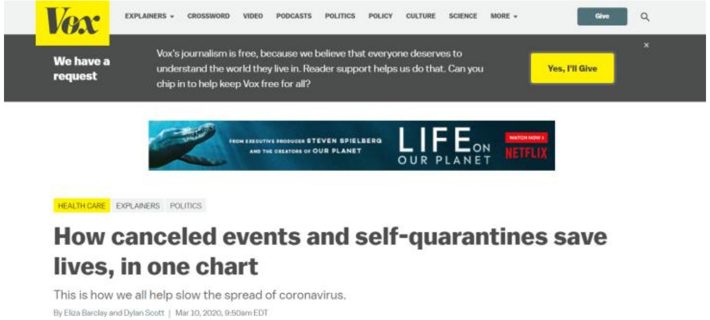
LAB 2, C3: Apply Descriptive Techniques Commonly used to Summarize Public Health Data

- Why do we care about having a GOOD write-up?
  - Not everyone knows how to INTERPRET statistics and especially in an age where media runs rampant and data is everywhere, it is EXTREMELY important that you can CORRECTLY INTERPRET DATA for others in a way that remains consistent with SCIENTIFIC TRUTH
  - Numbers mean NOTHING unless you assign some sort of meaning to them:
    - "Age is just a number"
      - In the context of **public health**, this is (usually) not true. Why?

      Example: During the COVID-19 pandemic, it was important to vaccinate high-risk populations, namely, **the elderly** because they are the most likely to become the most ill, have higher symptom severity, and more likely to be immunocompromised, compared to younger age groups

LAB 2, C3: Apply Descriptive Techniques Commonly used to Summarize Public Health Data

- When would you need to do a write-up?
  - Policy Briefs → People in Health Policy and Management are likely very familiar with this
  - Scientific Papers/Writing
  - If you are presenting in front of people who have no public health knowledge

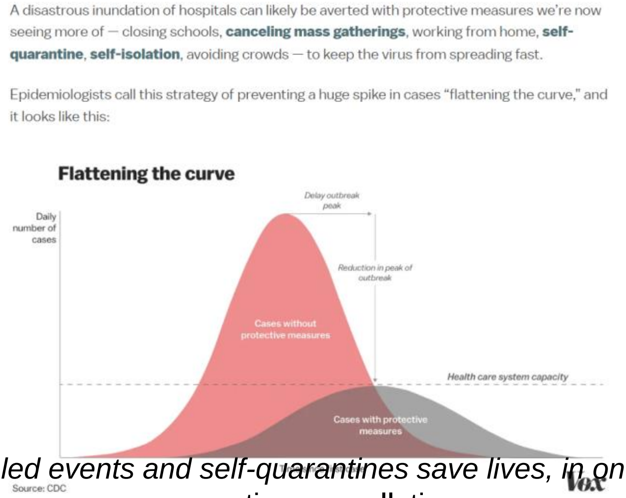


HEALTH CARE EXPLAINERS POLITICS

How canceled events and self-quarantines save lives, in one chart

This is how we all help slow the spread of coronavirus.

By Eliza Barclay and Dylan Scott | Mar 10, 2020, 9:50am EDT


A disastrous inundation of hospitals can likely be averted with protective measures we're now seeing more of — closing schools, **canceling mass gatherings**, working from home, **self-quarantine**, **self-isolation**, avoiding crowds — to keep the virus from spreading fast.

Epidemiologists call this strategy of preventing a huge spike in cases "flattening the curve," and it looks like this:

Flattening the curve

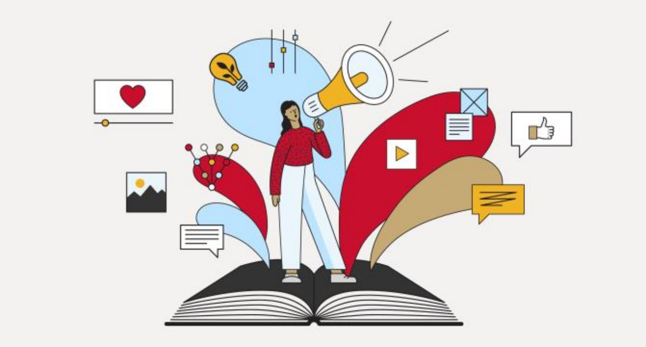
Daily number of cases
Delay outbreak peak
Reduction in peak of outbreak
Cases without protective measures
Health care system capacity
Cases with protective measures
Vox

Source: CDC

Barclay, E., & Scott, D. (2020, March 10). *How canceled events and self-quarantines save lives, in one chart*. Vox. https://www.vox.com/2020/3/10/21171481/coronavirus-us-cases-quarantine-cancellation

LAB 2, C3: Apply Descriptive Techniques Commonly used to Summarize Public Health Data

- What is YOUR ROLE?
  - You act as a bridge between the scientific and non-scientific community
  - You are a data **STORYTELLER**
    - State what you **found/evidence**
    - State the **Significance of your study**
    - If you **can, explain why the data is the way it is (or, data generating process/the WHY)**
    - Any **novel or surprising/not surprising finds**



Example of an Excellent Write-Up

In acyanotic congenital heart disease, blood has normal oxygen content but is pumped throughout the body abnormally (1). Cyanotic congenital heart disease leads to less oxygen-rich blood to be delivered to the body because oxygen-rich and -poor blood mix in the heart (1, 2). Thus, children with cyanotic congenital disease tend to have a rise in hemoglobin levels to compensate for the decreased oxygen content (3), while children with acyanotic disease would not have this. We see this in the data provided: acyanotic children had a lower median hemoglobin level (13.1) than cyanotic children (14.8) [see C1 #3-4 for explanation on why medians are compared]. **This is bolstered by** the shapes of the frequency distributions for acyanotic children, which is centered around 11-15 g/cc, compared to 12.4-19.4 g/cc for cyanotic children. Both distributions are skewed by outliers; for acyanotic children, there is one abnormally large outlier (20.5) leading to right-skew, and for cyanotic children, there is one small outlier (5.4) leading to left-skew. The spread, as indicated by interquartile range, is also smaller for acyanotic children (1.55) than cyanotic children (3.3); **this increased variability can potentially explained by differences in how low blood oxygen content is in each cyanotic child due to different disease etiologies—the lower the blood level, the greater the compensation needed. In conclusion. cyanotic children tend to have larger and more variable hemoglobin levels than acyanotic children, reflecting poor blood oxygen content.**

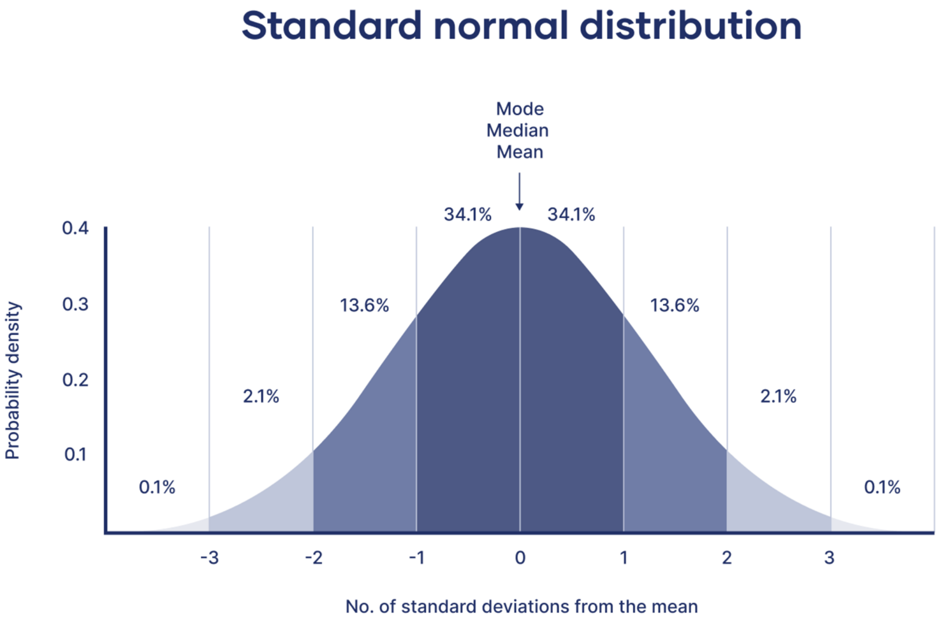
Green – Context, Claim/Hypothesis

Black – Supporting Evidence, how does the data support or undermine the claim/hypothesis?

Blue – Explanation behind the data, the WHY

Purple – Conclusion/Takeaway

Normality



Usually, normally distributed data is the most ideal, It is well-understood and has many benefits:

- Easy to understand: mean as a central tendency and sd as a variance parameter
- Can be directly compared with other data
- Meet the assumptions of many statistical analysis.
- ….

But in many cases, our data is far from a normal distribution…

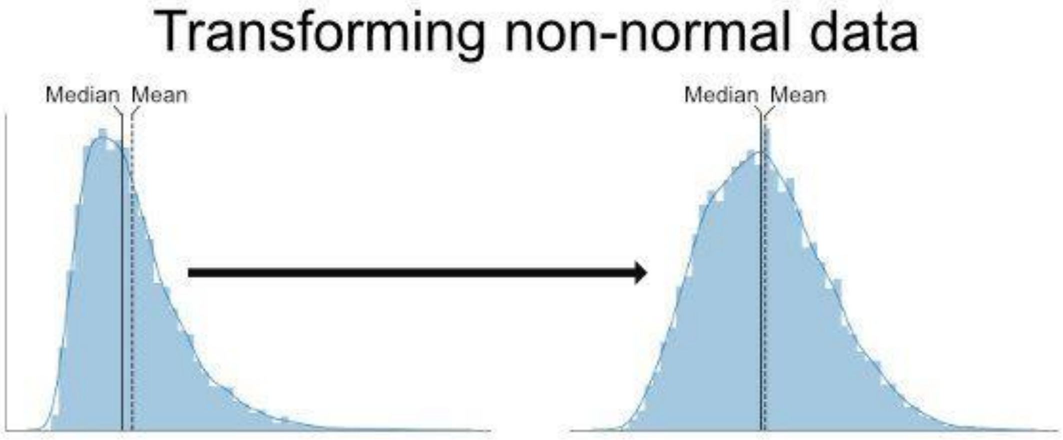
Slides Courtesy of Qingyuan Liu

Power Transformation

Data transformation can transform non-normal data to approximately normal distribution. Common transformations include the **logarithm, square root, and square**

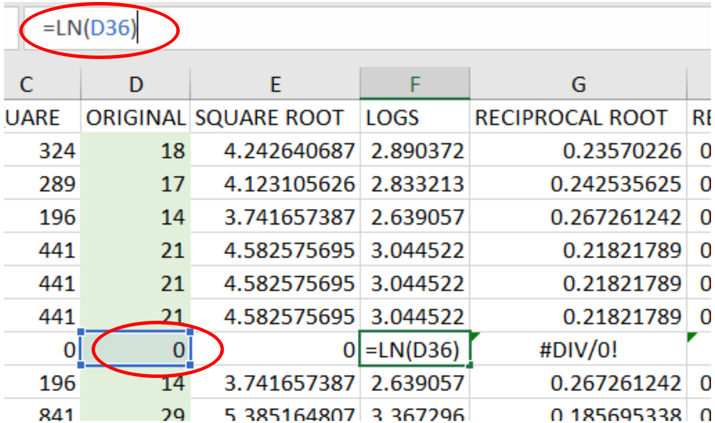
Log-transformation Example:

- Log Transformation is usually used when the data is **positively skewed**, where data in the long tail is far away from the center
- Applying a logarithmic transformation compresses the range by **pulling in the extreme values more than the lower values,** which makes it more symmetric.



Slides Courtesy of Qingyuan Liu

Caution: Zero Values in Power Transformation



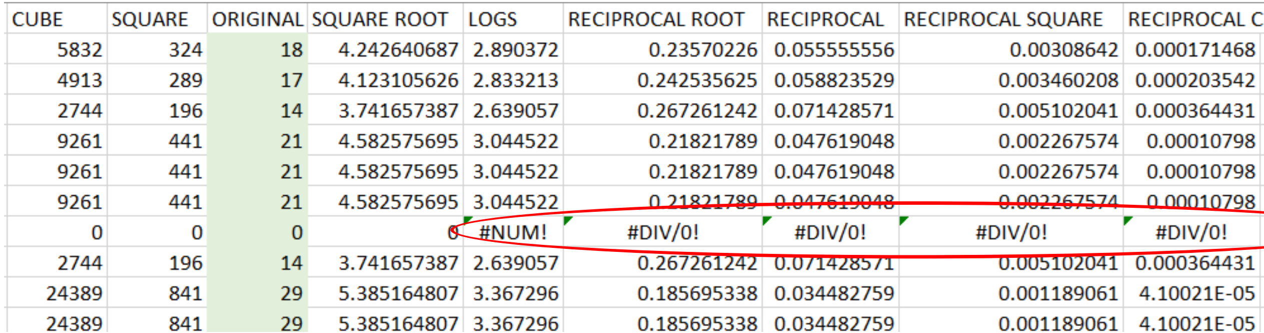
=LN(D36)

C	D	E	F	G	
UARE	ORIGINAL	SQUARE ROOT	LOGS	RECIPROCAL ROOT	RE
324	18	4.242640687	2.890372	0.23570226	0
289	17	4.123105626	2.833213	0.242535625	0
196	14	3.741657387	2.639057	0.267261242	0
441	21	4.582575695	3.044522	0.21821789	0
441	21	4.582575695	3.044522	0.21821789	0
441	21	4.582575695	3.044522	0.21821789	0
0	0	0	=LN(D36)	#DIV/0!	
196	14	3.741657387	2.639057	0.267261242	0
841	29	5.385164807	3.367296	0.185695338	0

- Many power transformation, such as log, is not applicable if the data contains 0's!
- It is important to check your original and transformed data to make sure there is no error produced

Logarithm: Log(0) = UNDEFINED

Reciprocal: 1/0 = UNDEFINED



CUBE	SQUARE	ORIGINAL	SQUARE ROOT	LOGS	RECIPROCAL ROOT	RECIPROCAL	RECIPROCAL SQUARE	RECIPROCAL CUBE
5832	324	18	4.242640687	2.890372	0.23570226	0.055555556	0.00308642	0.000171468
4913	289	17	4.123105626	2.833213	0.242535625	0.058823529	0.003460208	0.000203542
2744	196	14	3.741657387	2.639057	0.267261242	0.071428571	0.005102041	0.000364431
9261	441	21	4.582575695	3.044522	0.21821789	0.047619048	0.002267574	0.00010798
9261	441	21	4.582575695	3.044522	0.21821789	0.047619048	0.002267574	0.00010798
9261	441	21	4.582575695	3.044522	0.21821789	0.047619048	0.002267574	0.00010798
0	0	0	0	#NUM!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
2744	196	14	3.741657387	2.639057	0.267261242	0.071428571	0.005102041	0.000364431
24389	841	29	5.385164807	3.367296	0.185695338	0.034482759	0.001189061	4.10021E-05
24389	841	29	5.385164807	3.367296	0.185695338	0.034482759	0.001189061	4.10021E-05

Slides Courtesy of Qingyuan Liu

Solutions for error messages in transformed data

- One way to omit the error: **IF()** combined with **ISNUMBER()** inside summary functions

  =SUM(IF(ISNUMBER(A1:A10), A1:A10))

  =PERCENTILE(IF(ISNUMBER(A1:A10), A1:A10), 0.25)

-In this case, IF() **evaluates whether the value is a valid number** for each value one by one, return True or False

-If True, this data is included, if false, it is not. The error message “#NUM!” is not a valid number, thus it will not be included in the summary functions

- For average(), you can also use **AVERAGEIFS()** instead:

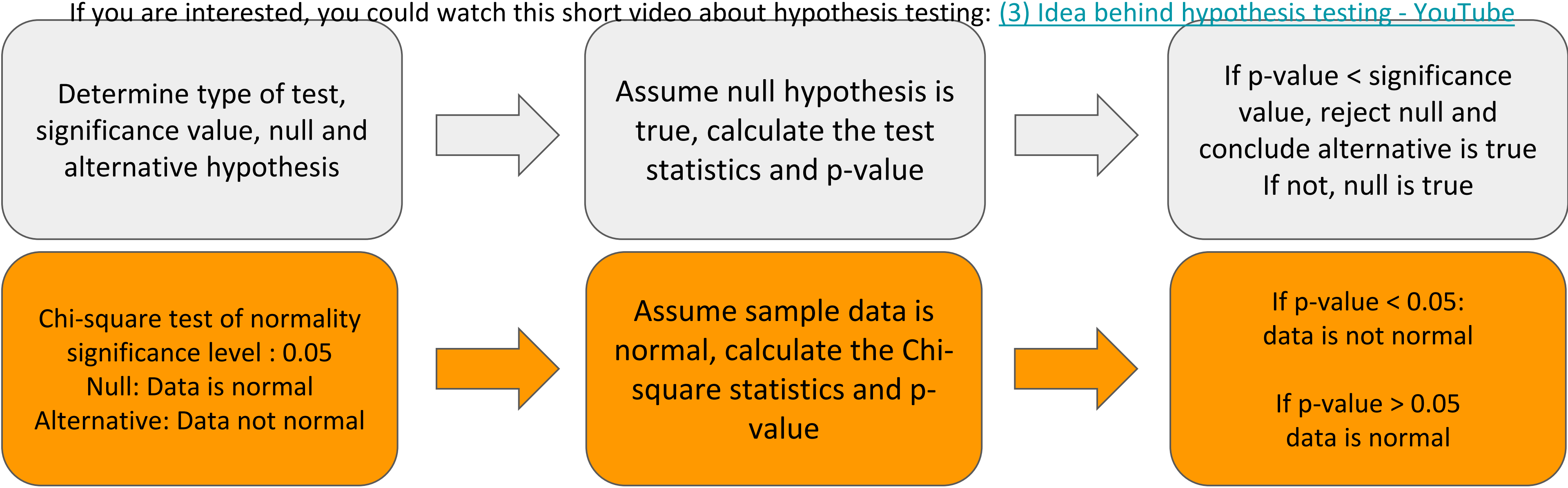
  =AVERAGEIFS(A1:A10, "<>#DIV/0!")

Slides Courtesy of Qingyuan Liu

Normality Test - Hypothesis Testing

The mechanism of normality test is **beyond the scope**, this diagram is a simple illustration of general hypothesis testings, you will learn more in the lectures. For this lab, you only need to know **how to interpret p-values.**

If you are interested, you could watch this short video about hypothesis testing: [(3) Idea behind hypothesis testing - YouTube]



Determine type of test, significance value, null and alternative hypothesis

Assume null hypothesis is true, calculate the test statistics and p-value

If p-value < significance value, reject null and conclude alternative is true
If not, null is true

Chi-square test of normality
significance level : 0.05
Null: Data is normal
Alternative: Data not normal

Assume sample data is normal, calculate the Chi-square statistics and p-value

If p-value < 0.05:
data is not normal

If p-value > 0.05
data is normal

Note: p-value is not a golden standard, the result could be misleading

Slides Courtesy of Qingyuan Liu

Venn Diagram

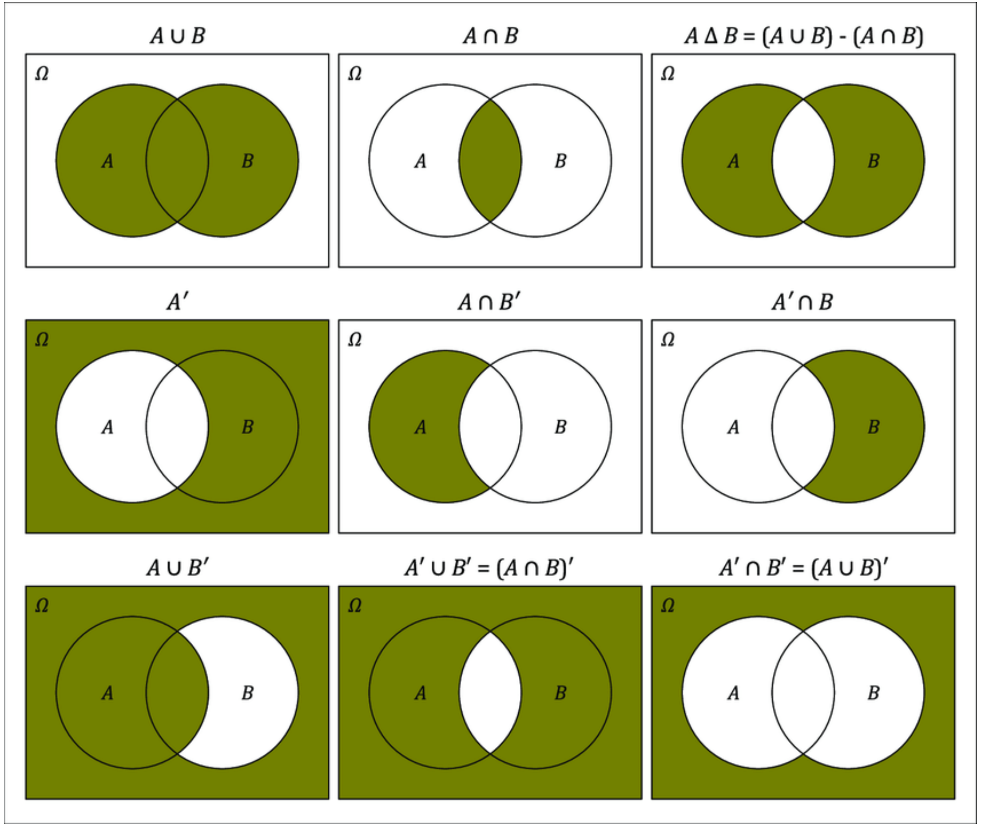
Each circle in a diagram indicates an event, and the area of their overlap means the intersection of events.

Notations:

A ∪ B: Union of A and B

A ∩ B: Intersection of A and B

A^c or A’ : Complement of A


$A \cup B$
$A \cap B$
$A \Delta B = (A \cup B) - (A \cap B)$
A'
$A \cap B'$
$A' \cap B$
$A \cup B'$
$A' \cup B' = (A \cap B)'$
$A' \cap B' = (A \cup B)'$
Ω
A
B

Slides Courtesy of Qingyuan Liu