Four general ways to address violations:

- **Ignore**
  - Sometimes you can use a method even if assumptions are violated
  - Thank the Central Limit Theorem for robustness
    - *means of large samples are normally distributed*
    - **tests that are based on** comparing sample **means** will be robust when they have sufficient sample size
    - **this doesn't help with F-test etc.**
  - Especially true if sample sizes are large ($n_i >> 50$) and violations are not extreme
    - **sample size must increase to accommodate how extreme violations are between groups**
      - especially if two samples both differ in opposite directions
  - Even with CLT we can't always ignore:
    - **outliers**
    - **frequency distribution between groups is very different**

Four general ways to address violations:

- Ignore
  - sometimes you can use a method even if assumptions are violated
  - especially true if sample sizes are large an violations are not extreme
- Transform
  - attempt to force normality and other assumptions onto data
  - We will investigate various tools
    - Usually boils down to: ***take the log of the data***
    - *Changes each measurement in the same way (1 to 1 correspondence) so that you can* ***transform back to get original data without ambiguity***
      - **monotonic relationship with original values**
      - remember to transform back the upper and lower limits for a Confidence Interval
  - work that does not always pay off but you at least maintain power of the test that you are using (since nonparameteric tests usually reduce power)

Four general ways to address violations:

- Ignore
    - sometimes you can use a method even if assumptions are violated
    - especially true if sample sizes are large an violations are not extreme
- Transform
    - attempt to force normality and other assumptions onto data
    - We will investigate various tools
    - work that does not always pay off
- Use Non-parametric method
    - classes of methods that do not require assumption of normality
    - not cost free! Often lose power etc

Power = 1-P[FTR Ho|Ho is incorrect] = P[reject|Ho is incorrect]

Four general ways to address violations:

- Ignore
  - sometimes you can use a method even if assumptions are violated
  - especially true if sample sizes are large an violations are not extreme
- Transform
  - attempt to force normality and other assumptions onto data
  - We will investigate various tools
  - work that does not always pay off
- Use Non-parameter method
  - classes of methods that do not require assumption of normality
  - not cost free! Often lose power etc
- **Computationally Intensive Methods**
  - Simulation
  - Bootstrap
  - randomization/permutation test

Data Transformation:

- a data transformation changes each data point by some simple mathematical formula
  - Used to improve fit of the normal distribution to the data to make standard deviation more similar between groups
  - 1 to 1 correspondence between transformed data and original scale
    - requires the same transformation be applied to each individual
    - Monotonic relationship with original values
    - e.g. larger values stay larger

Data Transformation:

- you can try out different transformations until you find one that makes he data fit assumptions

- you cannot keep trying until you manage to get a P-value < 0.05

Remember: Repeated testing leads to inflated Type I error

Just a reminder of an earlier question:

Two independent studies are performed to test the same null hypothesis.

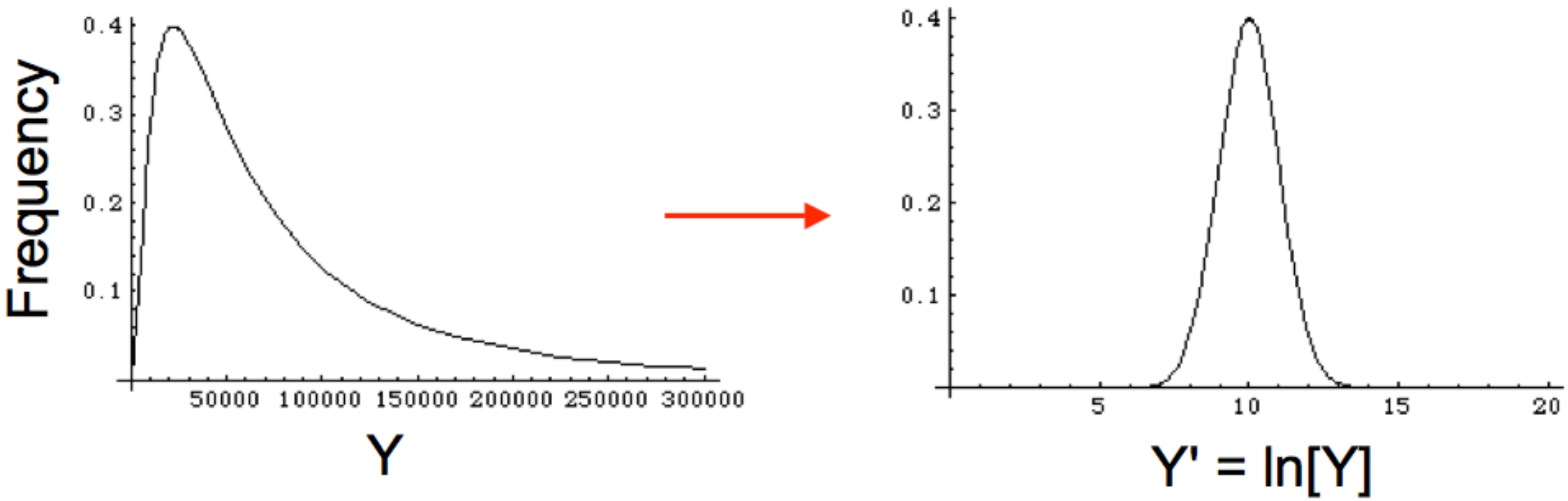
What is the probability that one or both of the studies obtains a significant result and rejects the null hypothesis ***even if the null hypothesis is true***? Assume that in each study there is a 0.05 probability of rejecting the null hypothesis.

---

a. **0.10**
b. **0.075**
c. **0.05**
d. **0.0975**

Log Transformation

$$Y' = \ln[Y]$$



Log Transformation

$$Y' = \ln[Y]$$

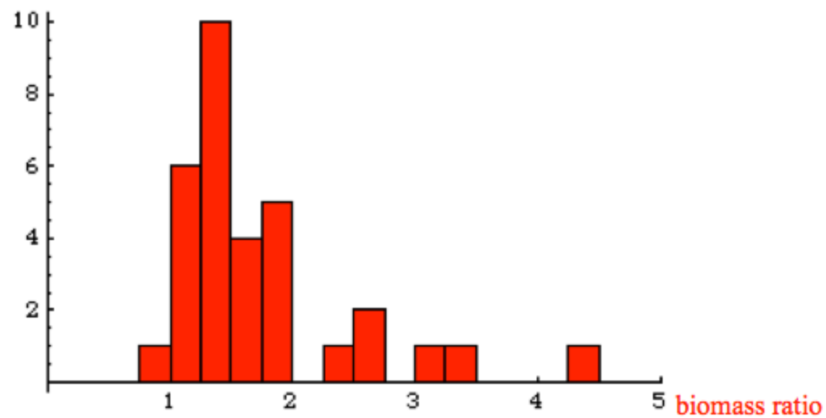
- most common
- It is most helpful when:
  - data spans several orders of magnitude
  - variables are > 0 (Ln(0) is undefined)
  - the variable is likely to be the result of the multiplication of several factors
    - log(ab) = log(a)+log(b)
  - the frequency distribution of the data is skewed to the right
  - the variance seems to increase as the mean gets larger (in comparison across groups)

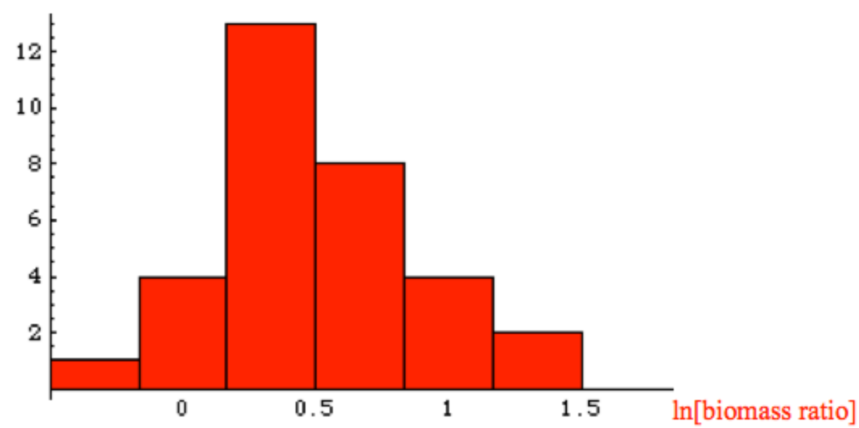
Log Transformation

$$Y' = \ln[Y]$$

Right skewed

Biomass ratio	ln[Biomass Ratio]
1.34	0.30
1.96	0.67
2.49	0.91
1.27	0.24
1.19	0.18
1.15	0.14
1.29	0.26

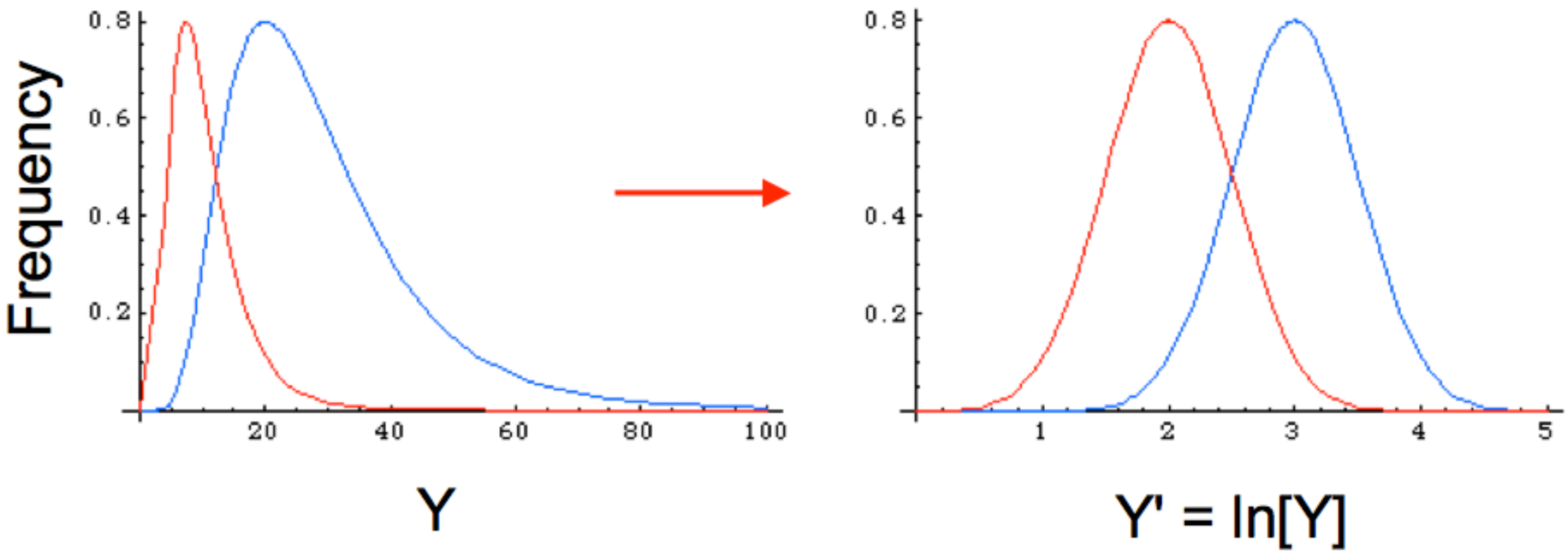




Log Transformation

$$Y' = \ln[Y]$$

Variance increases



Log Transformation

$$Y' = \ln[Y]$$

Use the same transformation on your hypothesis tests as you do on your data

H_0: The mean biomass ratio is unaffected by reserve protection(μ=1)

H_A: The mean biomass ratio is affected by reserve protection ($\mu \neq 1$)

H_0: The mean biomass ratio is unaffected by reserve protection(μ' =0)

H_A: The mean biomass ratio is affected by reserve protection ($\mu' \neq 0$)

Log Transformation

Y' = ln[Y]

Example: Confidence Interval with log-transformed data

Data:	5	12	1024	12398
Log Data:	1.61	2.48	6.93	9.43

$$\bar{Y}' = 5.11$$

$$s_{\ln[Y]} = 3.70$$

$$\bar{Y}' \pm t_{0.05(2),3} \frac{s_{\ln[Y]}}{\sqrt{n}} = 5.11 \pm 3.18 \frac{3.70}{\sqrt{4}} = 5.11 \pm 5.88$$

$$-0.773 < \mu_{\ln[Y]} < 10.99$$

$$e^{-0.773} < e^{\mu_{\ln[Y]}} < e^{10.99}$$

$$0.46 < \mu < 59278$$

Which one of these transformation helps to equalize standard deviations between groups when the group with the higher mean also has the higher standard deviation.

A- Arcsine transformation

B- Log transformation

C- Square root transformation

D- A and C

E- All the choices above

Other Common transformations:

Arcsine	• Exclusively on proportion data since they don't have equal standard deviations	$p' = \arcsin[\sqrt{p}]$
Square-root	• data counts	$Y' = \sqrt{Y + 1/2}$
Square	•Frequency dist^{n} skewed left	$Y' = Y^2$
Antilog	•Data skewed right	$Y' = e^Y$
Reciprocal	• When square transformation doesn't work	$Y' = 1/Y$