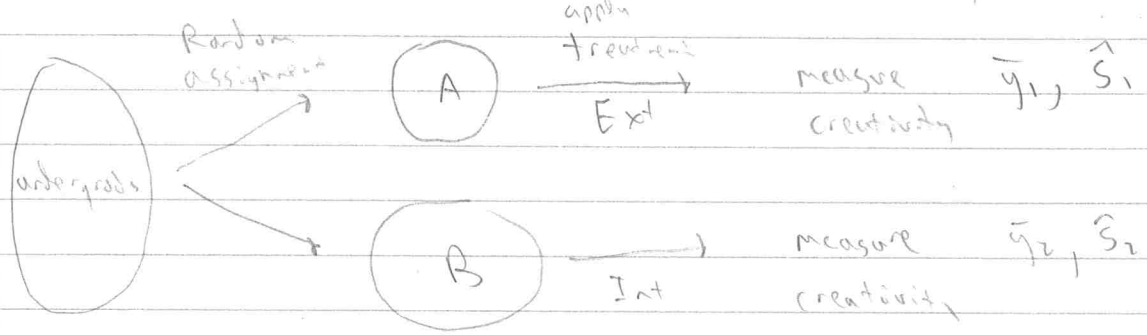
Learning objectives: Causality, generalizability, sampling distributions, randomization/permutation tests.

Stat 514, Sections 1.1 to 1.4 in Statistical Sleuth

- Case Study 1.1.1:

- Design:
  - Conscript 47 undergrads to participate
  - Randomly assign 24 to one group and 23 to another
  - Apply different treatments to each group
  - Measure creativity score within each group
  - Compare creativity scores between groups



↑ Can differences in average creativity be attributed to random assignment?

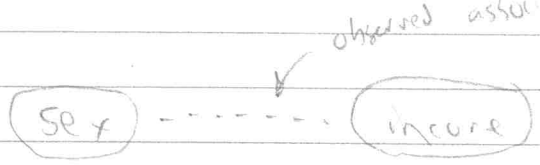
- Case Study 1.1.2:

- Design:
  - Get all clerks at bank and measure their sex and income
  - Compare mean income between the sexes
  - Can differences be attributed to random chance "mental model used randomly assign salaries"

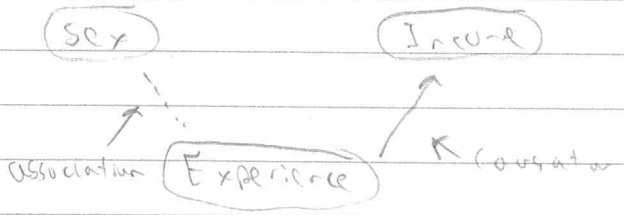
- Two types of Treatment allocations

Stage 1

- Observational Study
  - No treatment imposed on the observational units
  - Goal - describe (infer) properties of a population
  - Case 1.1.2
    - Sex not determined by investigators
    - want to make statements about clerks at the bank
  - Cannot claim causation
    - a third variable might cause an association
      - confounder

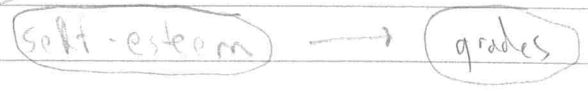


possible

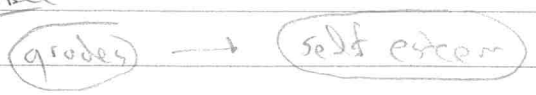


- Side Note: other ways that causation can be incorrect

Hypothesized

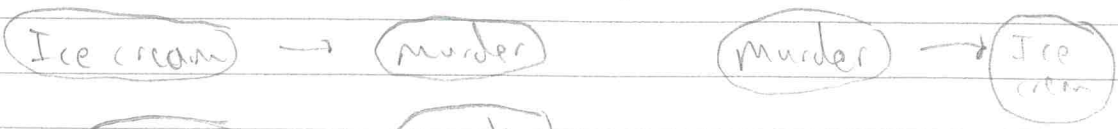


possible

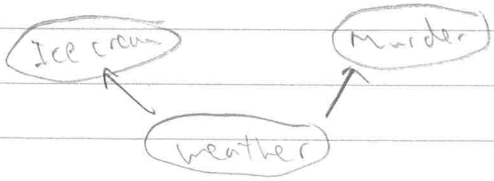


- Observed Ice cream ↑ and murder rate ↑

Hypothesis



possible



- Causation can never be fully proven by an observational study

- Design 2: Randomized experiment
  - treatment randomly assigned to observational units
  - case 1.1.1
  - can infer causation

- Why can we claim causation?
  - all other variables are on avg same on average
  - eg. possible to place really good writers in INT group and really bad in EXT, but randomly because we made groups by random

- Do observational studies b/c
  1. Goal is prediction, so causation doesn't matter
     - predict if rain, correlated variables are ok
  2. Combine w/ science to establish causation
     - smoking, animal trials

- How are observational units selected?
  - Random sampling
    - results are generalizable to population
  - Non-random sampling
    - scope of inference is limited.

- Random sampling makes sample look like population — some percent women, race, experience level etc.

Eg. Case 1.1.1: only good writers selected, so results may only hold for good writers.

	Random Experiment	Observational study
Random Selection	Causation generalizable	observation (possibly) non-generalizable
Non-random selection	Causation non-generalizable	observation (possibly) non-generalizable

- Inference:

- General setup:

1.) Model with population parameters
↳ parameters describe aspects of population we are interested in

Additive treatment effect model

Y_i = creativity score if subject i in Ext group

Y_i^* = " " " " " " " " Int "

$Y_i^* = Y_i + \delta$ ← Additive treatment Effect Model

↑ parameter, same for all individuals.

$\delta > 0 \Rightarrow$ Int improves creativity
$\delta < 0 \Rightarrow$ Ext improves creativity
$\delta = 0 \Rightarrow$ No effect

2.) Set a hypothesis based on parameters

$H_0: \delta = 0$
$H_A: \delta \neq 0$

Null hypothesis = "simpler state of affairs"

3.) Test statistic

test statistic = function of sample used to measure plausibility of a hypothesis

Expect $\bar{Y}_1 \approx \bar{Y}_2$

$\bar{Y}_1$: Ext creativity average

$\bar{Y}_2$: Int creativity average

So test statistic = $\bar{Y}_2 - \bar{Y}_1$

4.) Evaluate how rare our test statistic is under the Null hypothesis

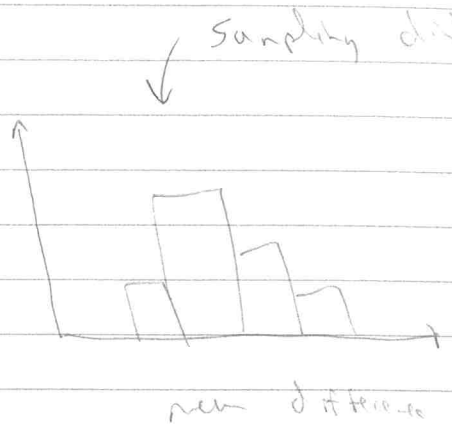
- RANDOMIZATION DISTRIBUTION SLIDES

- P-value = Probability we would have observed a test statistic as extreme or more extreme than what we observed if H_0 were true.

Go over Twin Study, Case Study 2.1.2

- Paired t-test §2.2
- Mean Difference in twin study = 0.1987. This difference is random.
- Recall sampling distribution

Take a sample of 15 twins
↳ average difference
Take a sample
↳ average difference
Take a sample
↳ average difference



- Slides on Sampling Dist

- Only one sample actually taken
  - How can we know anything about it?
  - Statistical theory!

- Thm: If population mean = μ
  population SD = σ
  then mean of sampling distribution of average = μ
  SD of sampling distribution of average = $\sigma/\sqrt{n}$

  n = sample size

- Central Limit Theorem: Shape of sampling distribution of sample average is more nearly normal for larger n.

- Recap: Even though we don't know the distribution of X, we know the distribution of $\bar{X}$ (for large n).

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

$\bar{X}$: sample average; μ: $E[X]$; σ^2: $Var(X)$; n: sample size

- This is useful because we want to know how extreme $\bar{X}$ is

- General setup of statistics (again)

1.) Posit. Model

$E[X] = \mu$ (expected difference is μ)

2.) Set interpretable hypotheses in terms of parameters

H_0: $\mu = 0$ (No difference)
H_A: $\mu \neq 0$ (some difference)

3.) Come up with test statistic

$\bar{X}$? $\quad \frac{\bar{X}}{\sigma/\sqrt{n}}$?

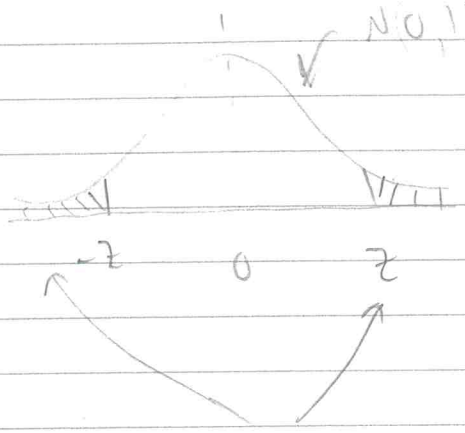
↑ larger values of each provide evidence against H_0

4.) Determine how rare test stat is under H_0

under H_0, $\bar{x} \sim N(0, \sigma^2/n)$

$$z = \frac{\bar{x}}{\sigma/\sqrt{n}} \sim N(0,1)$$

So compare z to a $N(0,1)$



shaded regions provide evidence against H_0

- Problem: Don't know σ^2

Solution: we can estimate σ^2 w/ s^2

σ^2: pop. variance; s^2: sample variance

But then it is better to use

$$\frac{\bar{x}}{s/\sqrt{n}} \sim t_\nu \quad \text{where } \nu = n-1$$

$= $ "degrees of freedom"

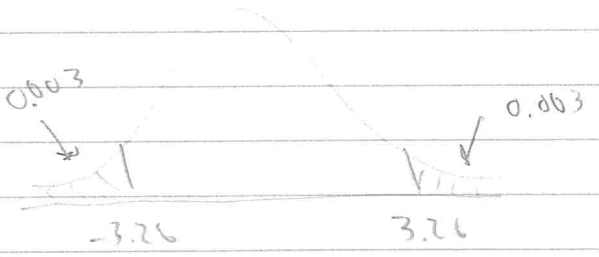
- Why t?
  - Estimating σ^2, so need to account for added uncertainty
  - t is more variable than $N(0,1)$ (longer tails)
  - You will be over confident in weirdness of $\bar{x}$ if you use $N(0,1)$

- R code: `t.test()`

- Twin study: $\bar{x} = 0.1846$
  $s/\sqrt{n} = 0.0615$
  $\nu = n - 1 = 14$

$$T = \frac{\bar{x}}{s/\sqrt{n}} = 3.236$$

compare to t_{14}



p-value = 0.006

↕ very sure if H_0 is true!

- So reject H_0 and conclude difference $\neq 0$

- Confidence Intervals for Effect

- Question: What are some plausible values for μ?

- Use general result

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

- Then we know that in 95% of samples, the following is true

$$t_{n-1}(0.025) = qt(0.025, df = n-1) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq qt(0.975, df = n-1) = t_{n-1}(0.975)$$

I solve for μ

$$\Rightarrow t_{n-1}(0.025)\, S/\sqrt{n} \leq \bar{X} - \mu \leq t_{n-1}(0.975)\, S/\sqrt{n}$$

$$\Rightarrow \bar{X} - t_{n-1}(0.025)\, S/\sqrt{n} \geq \mu \geq \bar{X} - t_{n-1}(0.975)\, S/\sqrt{n}$$

- Note: $-t_{n-1}(0.025) = t_{n-1}(0.975)$

- 95% Confidence Interval: $\bar{Y} \pm t_{n-1}(0.975)\, S/\sqrt{n}$

- Interpretation: See slides

Two Sample Inference:

Go over Beach Study, Case Study 2.1.1

- Goal: Still to explore difference in population means, but now have <u>two independent samples.</u>

- <u>General Setup of Statistics (again)</u>

1.) Posit a model

Let $X_{1,1}, X_{n_1}$ = Beach Depths in 1976
$Y_{1,1}, Y_{n_2}$ = Beach Depths in 1978

$X_i = \mu_1 + \varepsilon_i$

μ_1: mean Beach depth in 1976
ε_i: some noise, centered at 0, variance = σ_1^2

$Y_j = \mu_2 + \delta_j$

μ_2: mean Beach depth in 1978
δ_j: some more noise, centered at 0, variance = σ_2^2

2.) Set interpretable hypotheses

H_0: $\mu_1 = \mu_2$ (same mean beach-depth)

H_A: $\mu_1 \neq \mu_2$ (different mean beach-depth)

3.) Come up with a test statistic

$$\bar{X} - \bar{Y}$$

↑ values further from 0 support bigger difference

4.) Determine how rare test stat is under H_0

Under H_0: $\bar{X} - \bar{Y} \sim N\left(0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

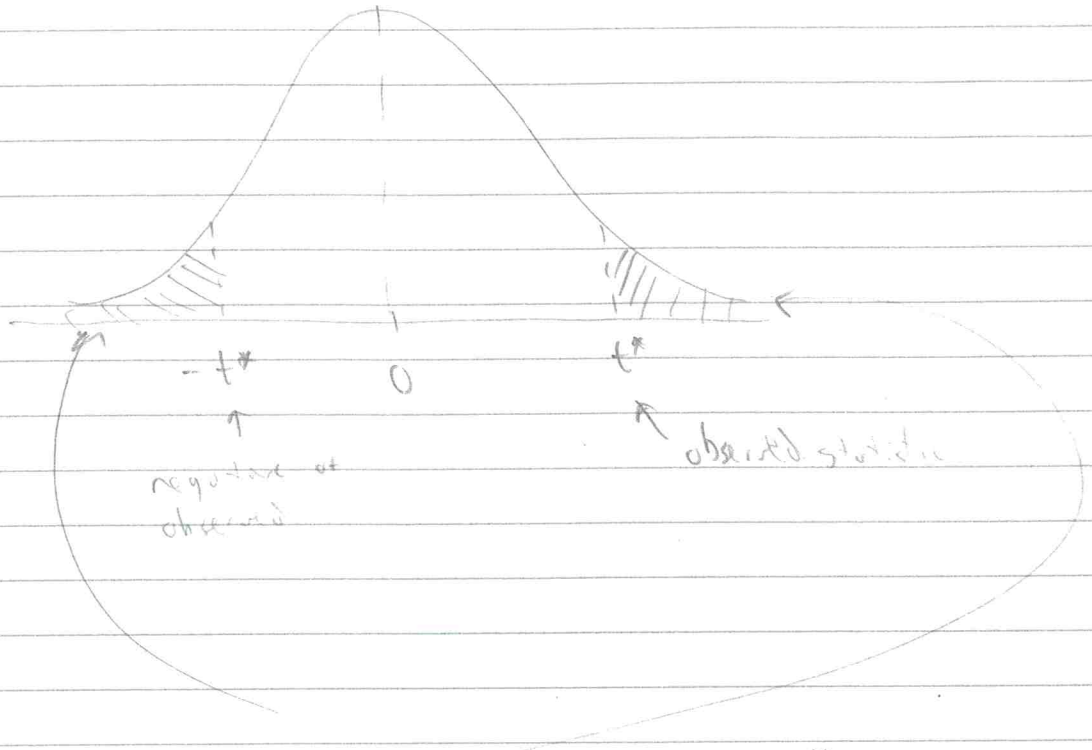
$$\Rightarrow \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

- Problem: don't know σ_1^2, σ_2^2
- Solution: replace with s_1^2 and s_2^2

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

ν: something weird "Satterthwaite's Approximation"

- We compare our observed test-statistic to the theoretical "null distribution"



proportion of t-stats that would provide as much or more evidence against the null than our observed t-stat.

p-value = $P_r(|T| \geq |t^*|)$ where $T \sim t_\nu$.

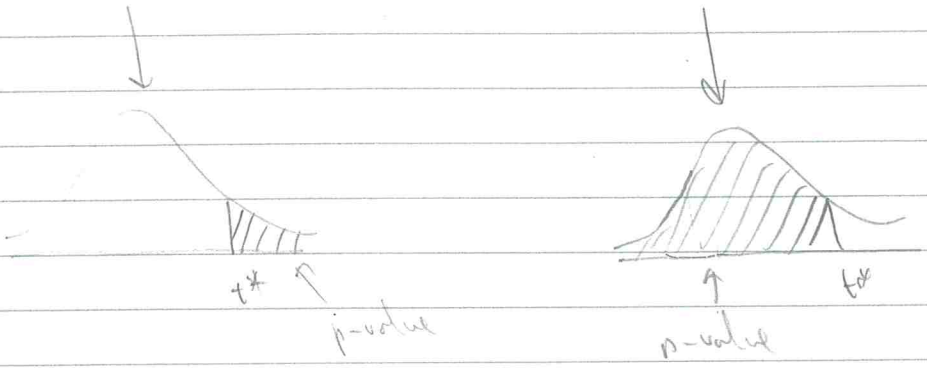
- Small p-values provide evidence against the null.
- Can also calculate confidence intervals, hypothesis tests, etc.

$$\bar{x} - \bar{y} \pm t_\nu(0.975)\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$$

Some Things to Think About

1.) One-sided tests
↑ set alternatives to be one sided

$H_0: \mu = 0$ or $H_0: \mu = 0$
$H_A: \mu > 0$ $H_A: \mu < 0$



Always report if p-value is one or 2 sided

2.) Non-zero Nulls

$H_0: \mu = a$, $H_A: \mu \neq a$

$$t^* = \frac{\bar{x} - a}{s/\sqrt{n}} \sim t_{n-1}$$

3.) "Significance" XKCD

p-value ≤ 0.05 is not magical, it's stupid.

4.) When reporting p-values, always either report the sample size or the confidence intervals, or both.

Chapter 3: Assumptions of 2-sample t-tools

- Assumptions in order of importance:

1.) Independence between observational units

2.) Equal Variance Assumption (if assuming equal variances)

3.) Normality (no skew / outliers)

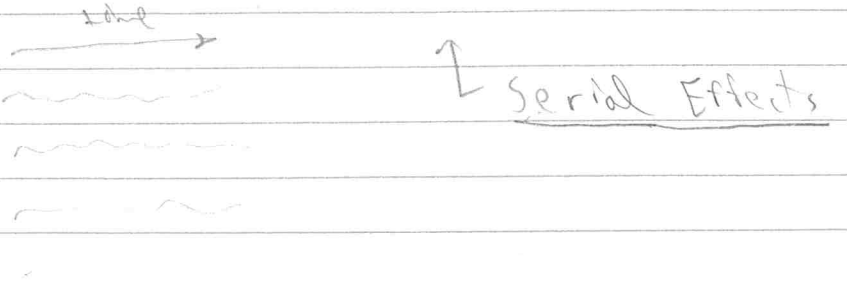
- Each of these assumptions can be checked
- Solutions exist for each problem.

1.) Independence:

↕ Knowing the value of one observation does not tell you any information on value of second

- Example of violation:
  - measure expression levels of a gene in two groups (disease / control)
  - Individual samples were collected by two technicians
  - technician 1 tended to produce samples with higher levels
  - cluster effects

- Example of violation:
  - measure expression levels on the same individuals over time



↕ Serial Effects

- To detect:
  1.) think carefully about how data were collected
    - were different responses measured on same subject?
    - were data collected in groups?
    - were groups treated differently unrelated to treatment?
  2.) Residual plots (chapter 8)

- Issues:

Recall $\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

- If samples are not independent, the variance formula is too small
  - You have less info than you think
  - So p-values are too small
  - Confidence intervals are too narrow

- Extreme Example:
  You have 4 people, measure each one 50 times

Data:

group 1	Ind 1	0	0	0 ...	0
	Ind 2	1	1	1 ...	1
group 2	Ind 3	2	2	2 ...	2
	Ind 4	3	3	3 ...	3

Sample size 4, but you think it's 200!

Correct variance estimate: $\frac{\sigma_1^2}{2} = \frac{\sigma_2^2}{2} = \frac{0.25}{2} = 0.125$

Assumed variance estimates $\frac{\sigma_1^2}{100} = \frac{\sigma_2^2}{100} = \frac{0.25}{100} = 0.0025$

- Solutions:
  - use more sophisticated stats
  - ANOVA (ch 12, 13) for cluster effects
  - Longitudinal analysis (ch 15) for serial effects

2.) Unequal Variances:

- Sometimes people assume $\sigma_1^2 = \sigma_2^2$
  ↑ easier to generalize to more complicated methods.

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Estimate σ^2 with $\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

- To detect:
  - make side-by-side boxplots
  - only worry if extreme

- Issues:
  - Not really estimating σ, so variance estimate of $\bar{X} - \bar{Y}$ is also wrong.
  - only really a problem when n_1 very different from n_2 and variances are very different

- Solution:
  - assume unequal variances
  - Could transform data.

3.) Normality (skew):

- t-test assume $X_1, \ldots, X_{n_1}$ or $Y_1, \ldots, Y_{n_2}$ are normal, or that the sample size is "large enough"

- More skew => larger sample size needed

- Issues:
  - Confidence intervals do not have 95% coverage
  - Hard to predict if conservative or liberal
  - P-values do not have correct interpretation

- To detect:
  - Make quantile-quantile plots (slides)

- Solutions:
  - transformation
  - Do permutation test

4.) Normality (outliers)

- To detect
  - histograms, boxplots, qq-plots

- Issues:
  - Results are unstable (remove point, change results)

- Solution:
  - Run <u>some</u> analysis with and without outliers
  - If results are same, do not remove
  - If results differ, either
    1.) Run a robust analysis (ch 4)
    2.) Report both analyses.

- Transformations:
  - many transformations of data are possible
    $\sqrt{y}$, $\arcsin(y)$, $\log(y)$
  - log is usually the only one used

- When to use log:
  1.) values are positive
  2.) Larger mean ⇒ larger variance
  3.) data are skewed

↕ makes such data symmetric w/ equal variance

Rainfall slides

I Interpreting transformations in original units

• Interpreting log-transformations in observational study model,

$w_i = \log(Y_i) \Rightarrow Y_i = \exp\{w_i\}$

$w_i = \mu + \varepsilon_i$

$z_i = \log(x_i) \Rightarrow x_i = \exp\{z_i\}$

$z_j = \mu + \delta + \xi_j$

$\delta = \text{Average}(z) - \text{Average}(w) = \text{Average}(\log(x)) - \text{Average}(\log(y))$

↳ want to interpret on X, Y scale

- Note $\log(\text{Average}(x)) \neq \text{Average}[\log(x)]$

E.g.) $x = 2, 4$

$\log_2(x) = 1, 2, \quad \text{Average}(\log_2 x) = 1.5$

$\text{Average}(x) = 3, \quad \log_2(\text{Ave } x) = \log_2(3) \approx 1.6$

- So not true that $E[X] = e^{\delta} E[Y]$

- But is true that $\text{Median}[X] = e^{\delta}\, \text{median}[Y]$ if $\log(X)$ & $\log(Y)$ or both symmetric.

e^{δ} = multiplicative difference in medians

for ideal: $e^{\delta} = e^{\text{Ave}\log(x)} \, e^{-\text{Ave}\log(y)} \Rightarrow e^{\delta} = \text{Ave}(x) / \text{Ave}(y)$

true $e^{\delta} = e^{\text{med}(\log x)} \, e^{-\text{med}\log y} \Rightarrow e^{\delta} = \text{med}(x) / \text{med}(y)$

δ : Average($\log(x)$) − Average($\log(y)$)

: median($\log(x)$) − median($\log(y)$) (if distributions of $\log x$ and $\log y$ are symmetric)

: $\log(\text{median}(x)) - \log(\text{median}(y))$

: $\log\left(\frac{\text{median}(x)}{\text{median}(y)}\right)$

$\therefore e^{\delta} = \frac{\text{Median}(x)}{\text{Median}(y)}$

$\therefore$ Median(x) = e^{δ} Median(y)