From Bigger to Smaller

CHIDATA

Recap: Simple Random Sampling

Population

Sample

“Sample Mean”

“Population Mean”

Error in an estimate is normally distributed

$$\epsilon \approx N(0, \frac{\sigma^2_{pop}}{K})$$

Variance of the population

Size of the sample

CHIDATA

Recap: Rule of Thumb

Draw a sample of size K from a population with a range of values in [a,b].

The difference between the sample mean and the population average is:

Error within $\boxed{\pm \frac{(b-a)}{\sqrt{K}}}$ with 95% probability

"With 95% probability, the sample mean is within that error of the true value"

CHIDATA

Aside…

Lot's of questions about this on Slack. Rule of thumb is a simple analytical formula to understand error rates.

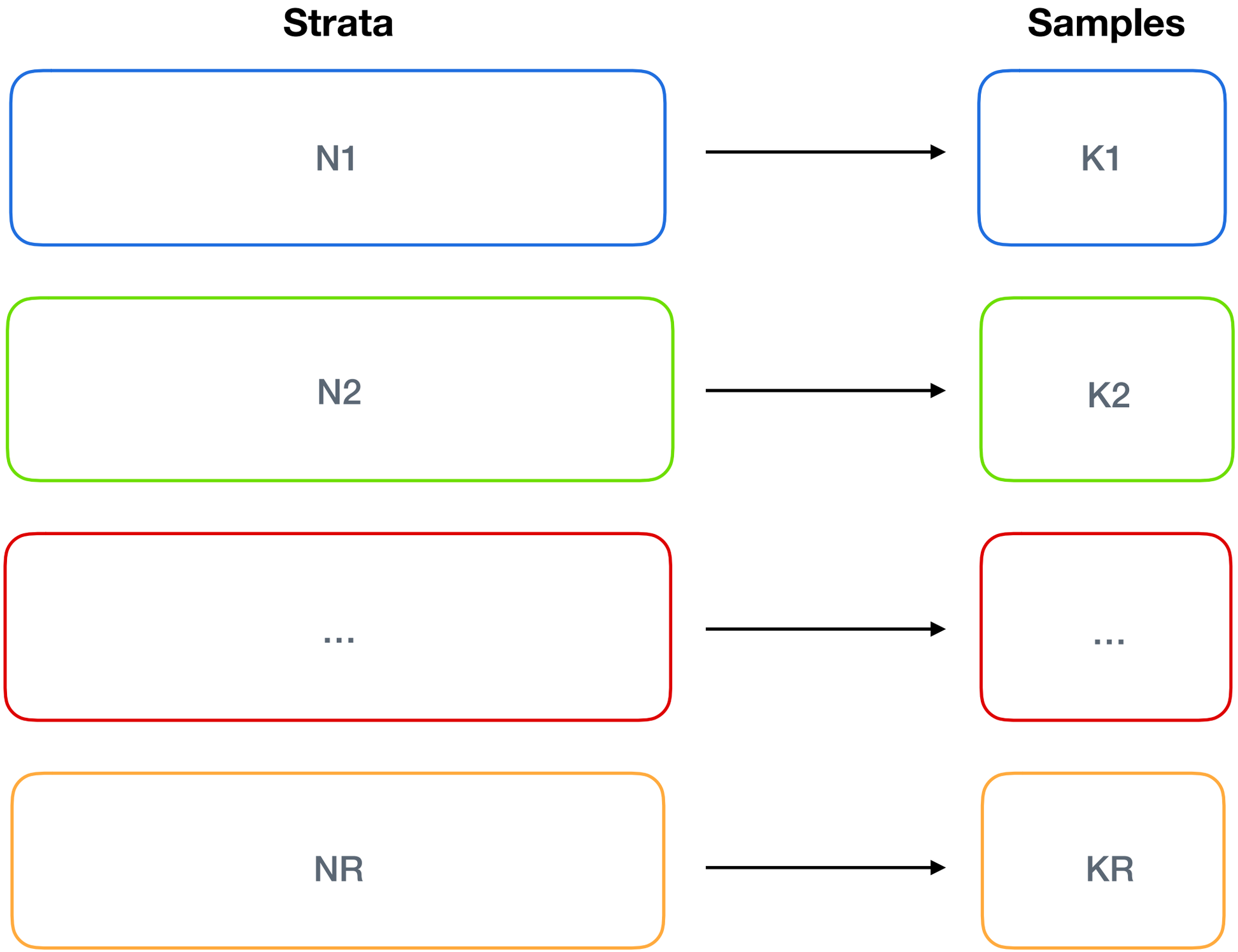
$$\pm \gamma \cdot \boxed{\sqrt{\frac{\sigma^2_{pop}}{K}}} \longleftarrow \textbf{Standard "Error"}$$

$$\gamma = 2 \approx 95\%$$

$$\gamma = 2.5 \approx 99\%$$

CHIDATA

Stratified Samples



CHIDATA

Assumptions!!!

Samples

K1

K2

...

KR

Calculate local sample mean

$\bar{\mu}_1$

$\bar{\mu}_2$

$\bar{\mu}_R$

All samples return at least 1 item

Weighted Combination

$$N = \sum_{i=1}^{R} N_i$$

$$\bar{\mu} = \sum_{i=1}^{R} \bar{\mu}_i \cdot \frac{N_i}{N}$$

You know Ni/N

CHIDATA

Stratified Sampling Error

Add up the error from each local SRS estimate:

$$\epsilon = \sum_{i=1}^{R} \epsilon_i \cdot \frac{N_i}{N} = \sum_{i=1}^{R} \epsilon_i \cdot w_i$$

$$\epsilon \approx \sum_{i=1}^{R} w_i \cdot \mathcal{N}(0, \frac{\sigma_i^2}{K_i})$$

Using rules for normal distributions:

$$\epsilon \approx \mathcal{N}(0, \sum_{i=1}^{R} \frac{\sigma_i^2}{K_i} w_i^2)$$

Rule of Thumb

Assume all strata are roughly the same size

Rule of thumb Confidence Interval

$$\pm \frac{RMS(local)}{\sqrt{R}}$$

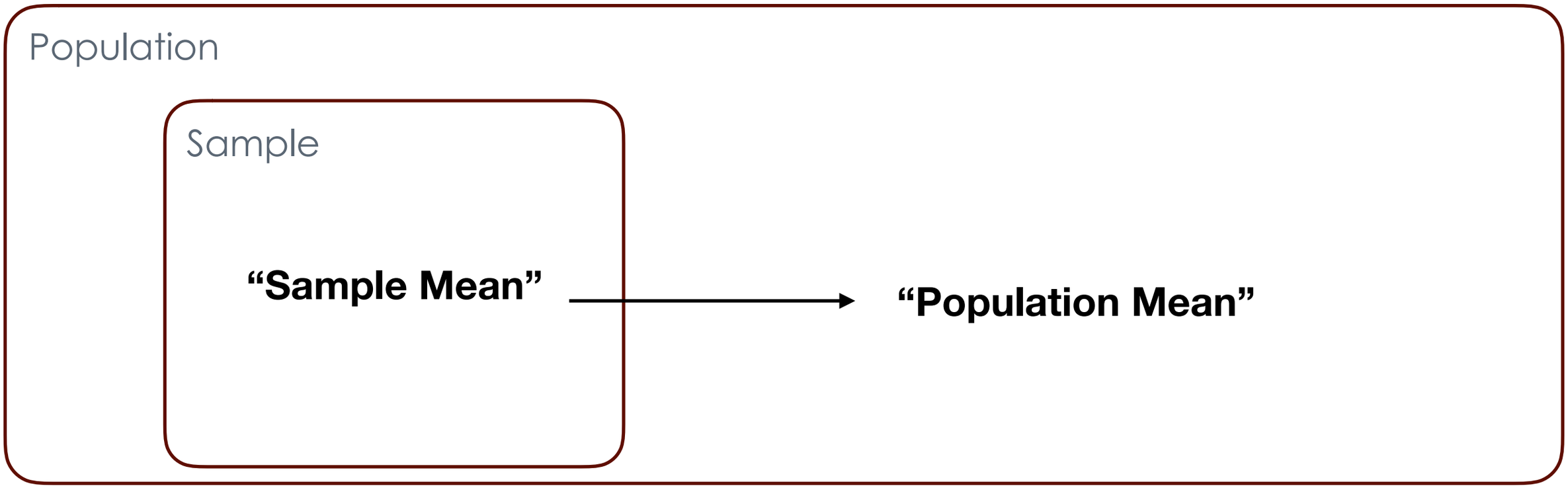
Calculate local CIs Root-Mean-Square

Number of strata

Can be more accurate than the sum of the parts!

Robust to Response Biases.

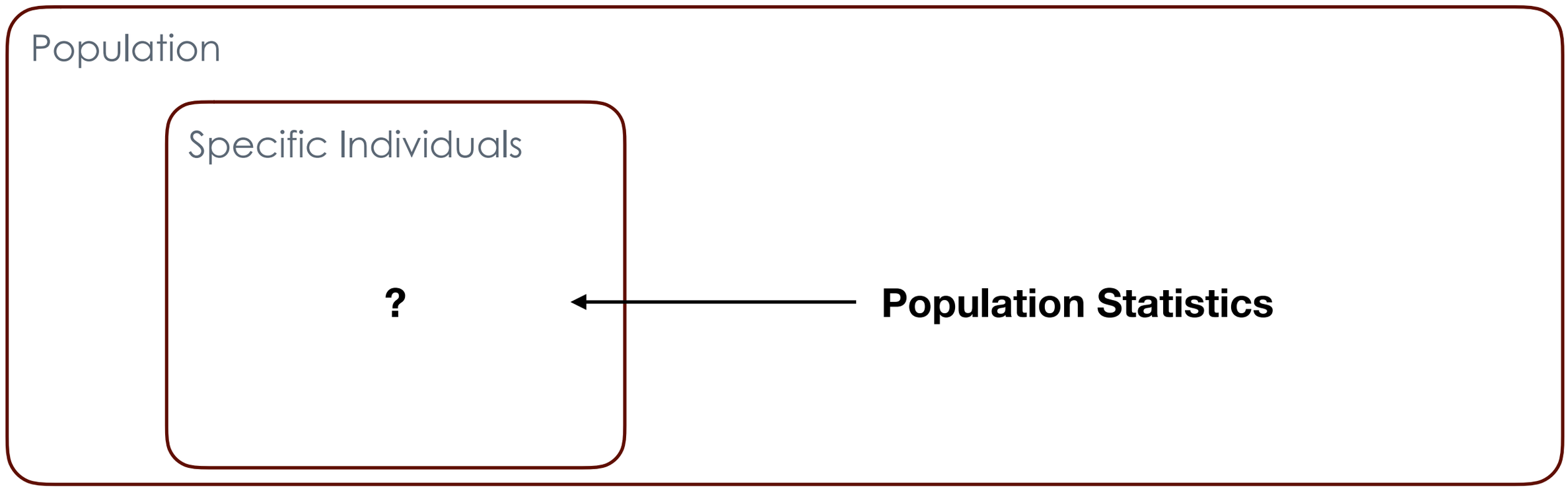
A Minore Ad Maius:
From the Smaller to the Bigger



Infer something about the population from the sample

Small to big….

A Maiori Ad Minus:
From the Bigger to the Smaller



Suppose you know something about the full population, what can you say about specific individuals

Big to small

CHIDATA

Comparisons Between Populations

Population 1

Average income

Admission Rate into College

Median age

Most frequent name

Population 2

Average income

Admission Rate into College

Median age

Most frequent name

stat1

stat2

Meaningful v.s. Significant

How meaningful: does the measured difference imply the desired differences in individuals.

How significant: could the measured difference be attributed to random chance.

CHIDATA

Start with a simple problem: Means

Population 1

Mean Age = 45.6

Population 2

Mean Age = 32.1

What does this tell us about both populations?

CHIDATA

Start with a simple problem: Means

Population 1

Mean Age = 45.6

[6.4, 97.6]

Population 2

Mean Age = 32.1

[32.0,32.2,32.1,32.5,32.1]

Actually not a lot….

At least one individual in Population 1 that has a value greater?

When is the “mean” meaningful?

Population 1

Mean Age = 45.6

Population 2

Mean Age = 32.1

The populations have a similar “spread” (or variance)

$$var = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

CHIDATA

When is the "mean" meaningful?

Chebyshev inequality. For any r.v X

$$\mathbf{P}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

"k" stds

If two populations X,Y have different means but the same "spread" and assume Mean(Y) > Mean(X)

Can calculate: $$\mathbf{P}[Y - \mu_x \leq 0] \leq \frac{\sigma^2}{\sigma^2 + \Delta^2}$$

Fraction of individuals in population Y less than the mean

CHIDATA

Rules of thumb

Can calculate: $$\mathbf{P}[Y - \mu_x \leq 0] \leq \frac{\sigma^2}{\sigma^2 + \Delta^2}$$

- Obviously symmetric P[X - Mean(y) > 0]….

- When the variances are the same and delta is roughly a standard deviation, a majority of Y is greater than X

CHIDATA

Start with a simple problem: Means

Population 1

Mean Age = 45.6

Population 2

Mean Age = 32.1

Reasonable comparison between two populations when the spread is similar and the expected difference is at least as big as the spread.

CHIDATA

How could you directly get this property?

Can calculate: $$\mathbf{P}[Y - \mu_x \leq 0] \leq \frac{\sigma^2}{\sigma^2 + \Delta^2}$$

- Obviously symmetric P[X - Mean(y) > 0]….

- When the variances are the same and delta is roughly a standard deviation, a majority of Y is greater than X

CHIDATA

Medians are Robust

Population 1

Median Age

Population 2

Median Age

Always guarantees that a majority of the greater population is larger without distributional assumptions.

CHIDATA

Sanjay's $0.02

Population 1

Median Age

Population 2

Median Age

Always use medians unless you are interested in measuring a "rate" or an "expected effect".

If someone uses "average" or mean when they are comparing populations, they are probably trying to mislead you.

CHIDATA

Sanjay's $0.02

Tax cuts saved the "average" tax payer $2000…

“Meaningful”

Population 1

stat1

Population 2

stat2

Is the quantity that we calculate meaningful

Is the population division meaningful?

CHIDATA

Example…

	PhD Admission Rate
West Coast Students	5%
East Cost Students	3%

East Coast students are less likely to be admitted, what is the implication of this result?

CHIDATA

Simpson's Paradox

	Major 1	Major 2
West Coast Students	7% (N=100)	2% (N=200)
East Cost Students	11% (N=12)	3% (N=1000)

East Coast students are actually admitted more frequently per subdivision.

Aggregates of heterogenous populations can be misleading

CHIDATA

Principle of Similar Confidence

	Major 1	Major 2
West Coast Students	7% (N=100)	2% (N=200)
East Cost Students	11% (N=12)	3% (N=1000)

Should only compare results of similar “confidence”

Use rule of thumbs!!

CHIDATA

How to lie with statistics

Population 1

stat1

Population 2

stat2

Is the quantity that we calculate meaningful

Is the population division meaningful?