Beyond a Reasonable Doubt

in statistics…

CHIDATA

Recap: 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

Rule of thumb

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 on the order of magnitude of the spread**.

CHIDATA

Means v.s. Medians

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.

“Meaningful”

Population 1

stat1

Population 2

stat2

Is the quantity that we calculate meaningful

Is the population division meaningful?

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)

Aggregates of heterogenous populations can be misleading

Principle of similar confidence! Compare and aggregate things with similar spread and similar size.

Recap: 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

“Significance”: Example

Treatment

Avg. Weight Loss

[-1.3,-4.5,10.5]

[1.7,5.2,10.5]

Control

Avg. Weight Loss

[1.7,5.2,10.5]

?

Uncertainty about who got the treatment

“Significance”: Example

Interface 1

Click-Through-Rate

Interface 2

Click-Through-Rate

?

Randomly assign users to interfaces

CHIDATA

Randomized Controlled Trial

Randomly assign individuals to treatment and control

Why do you need randomness?

Same as opinion polling, you want to determine the full population (or expected) effect of an

Same as opinion polling, randomness must be independent.

Picture looks like this

Sample 1

stat1

Sample 2

stat2

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

CHIDATA

Picture looks like this

Sample 1

stat1

Sample 2

stat2

Can think of it as a generalization of SRS….

A “Null Hypothesis”

Control (Sample 1)

Stat 1

Control (Sample 2)

Stat 2

Start with the assumption that there is **no meaningful** difference between the populations.

CHIDATA

Null Hypothesis: Example

Treatment (No Effect)

Avg. Weight Loss

Control

Avg. Weight Loss

Avg Weight Loss for Treatment and Control are the same

Null Hypothesis: Example

Interface 1

Click-Through-Rate

Interface 2

Click-Through-Rate

Click-Through-Rate on both interfaces is the same

Null Hypotheses are Subjective

Control (Sample 1)

Stat 1

Placebo (Sample 2)

Stat 2

Think of it as an intellectual baseline…

No single “correct” choice (we’ll see this on Friday!)

CHIDATA

Significance (or P-Value)

Stat 1

Stat 2

Probability that the difference between two statistics is at least as big assuming the null hypothesis is true.

CHIDATA

Two-Sample Z-Test: Comparing Means

A simple p-value calculation that gives us intuition

Sample 1

Mean

$$\bar{\mu}_1 \approx \mu_1$$

Sample 2

Mean

$$\bar{\mu}_2 \approx \mu_2$$

Apply ideas from sampling stats!

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

Variance of the population

Size of the sample

Step 1. Assume the Null Hypothesis

Assume: $\mu_1 = \mu_2$

Sample 1

$$\bar{\mu}_1 \approx \mu_1$$

Sample 2

$$\bar{\mu}_2 \approx \mu_2$$

Define: $Z = \bar{\mu}_1 - \bar{\mu}_2$

Variances add up

No expected difference

$$Z \sim \mathcal{N}(0, var(\bar{\mu}_1) + var(\bar{\mu}_2))$$

CHIDATA

Step 1. Assume the Null Hypothesis

Assume: $\mu_1 = \mu_2$

Define: $Z = \bar{\mu}_1 - \bar{\mu}_2$

Variances add up

No expected difference

$$Z \sim \mathcal{N}(0, var(\bar{\mu}_1) + var(\bar{\mu}_2))$$

Variance of the sample

$$Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2})$$

Number of samples

CHIDATA

Step 1. Assume the Null Hypothesis

Assume: $\mu_1 = \mu_2$

$$Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2})$$

"Null Hypothesis Model": A model for the world assuming the null hypothesis is true

CHIDATA

Step 2. Observe the Actual Difference

Sample 1

$$\bar{\mu}_1 \approx \mu_1$$

Sample 2

$$\bar{\mu}_2 \approx \mu_2$$

$$\Delta = \bar{\mu}_1 - \bar{\mu}_2$$

CHIDATA

Step 3. Calculate P-Value

Assume:

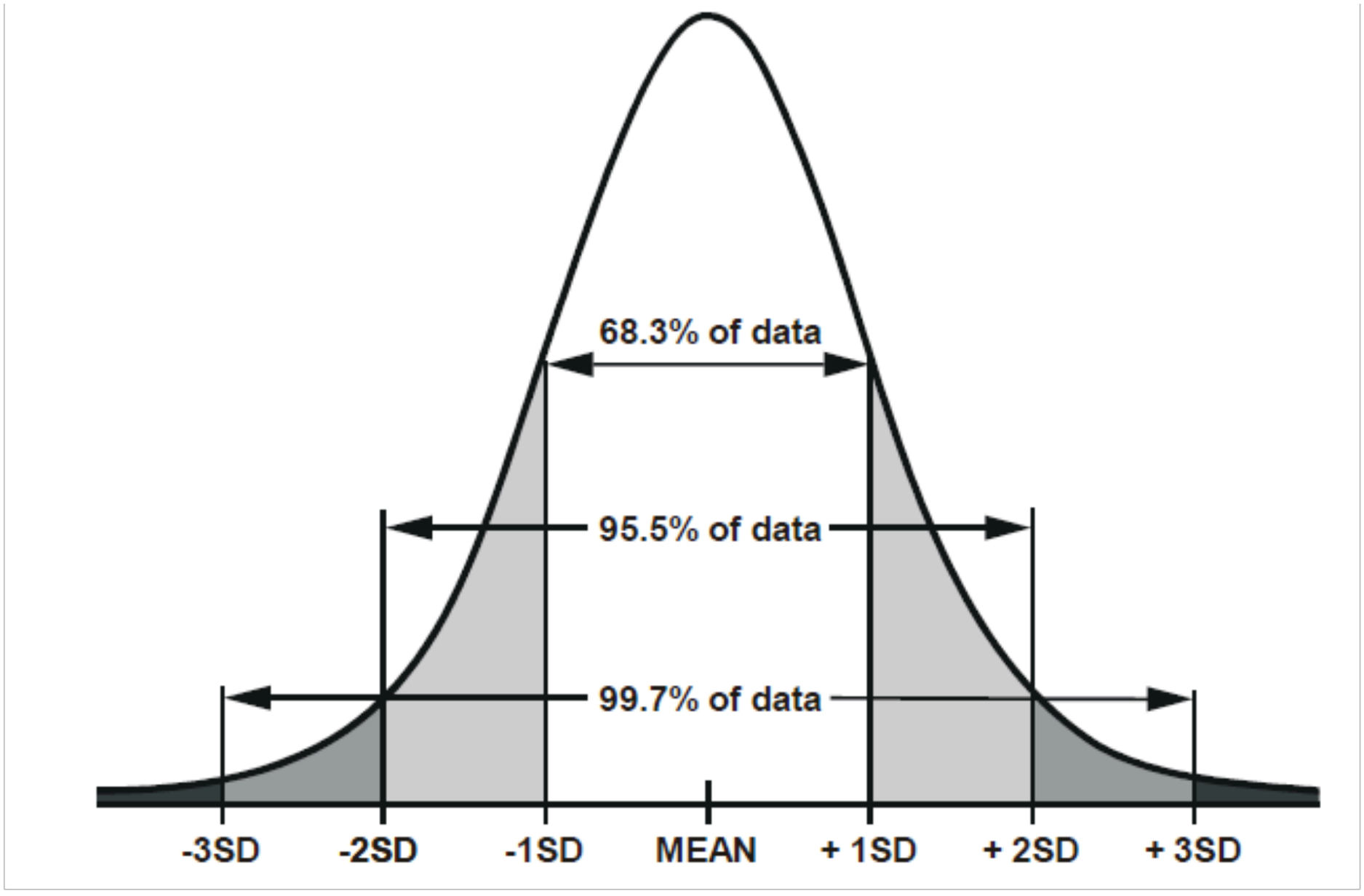
$$Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2})$$

Calculate:

$$p = \mathbf{Pr}[Z > \mu]$$

CHIDATA

Rules of Thumb Again



CHIDATA

Step 3. Calculate P-Value

Assume:

$$Z \sim \mathcal{N}(0, \frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2})$$

Calculate:

$$p = \mathbf{Pr}[Z > \Delta]$$

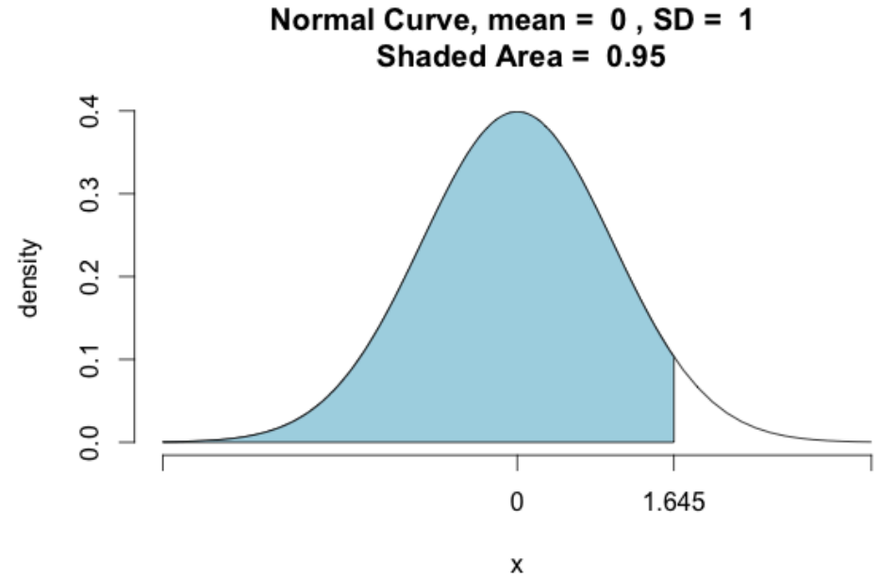
$$z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}}$$

→ # standard deviations of difference
Can be turned into a p-value using a table

CHIDATA

Step 3. Calculate P-Value

$$z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}}$$




```
>>> import scipy.stats as st
>>> st.norm.ppf(.95)
1.6448536269514722
>>> st.norm.cdf(1.64)
0.94949741652589625
```

(1- p-value)!

CHIDATA

Two-Sample Z-Test: Comparing Means

A simple p-value calculation that gives us intuition

Sample 1

$$\bar{\mu}_1, \bar{\sigma}_1^2, K_1$$

Calculate mean, variance, and size of sample

Sample 2

$$\bar{\mu}_2, \bar{\sigma}_2^2, K_2$$

Calculate mean, variance, and size of sample

$$z = \frac{\Delta}{\sqrt{\frac{\sigma_1^2}{K_1} + \frac{\sigma_2^2}{K_2}}}$$

Z-statistic

CHIDATA

Randomized Controlled Trial

Randomly assign individuals to treatment and control

Apply the desired intervention

Observe results and calculate a p-value

Accept conclusion when p-value is below some confidence threshold (e.g., 0.05 or 0.01)

CHIDATA

What are p-values?

Control the probability of "accidentally" accepting a null-hypothesis

$$p < \alpha$$

Calculated p-value **Acceptance thresh**

False Discovery Rate: The chance of accepting a null hypothesis in a RCT procedure

CHIDATA

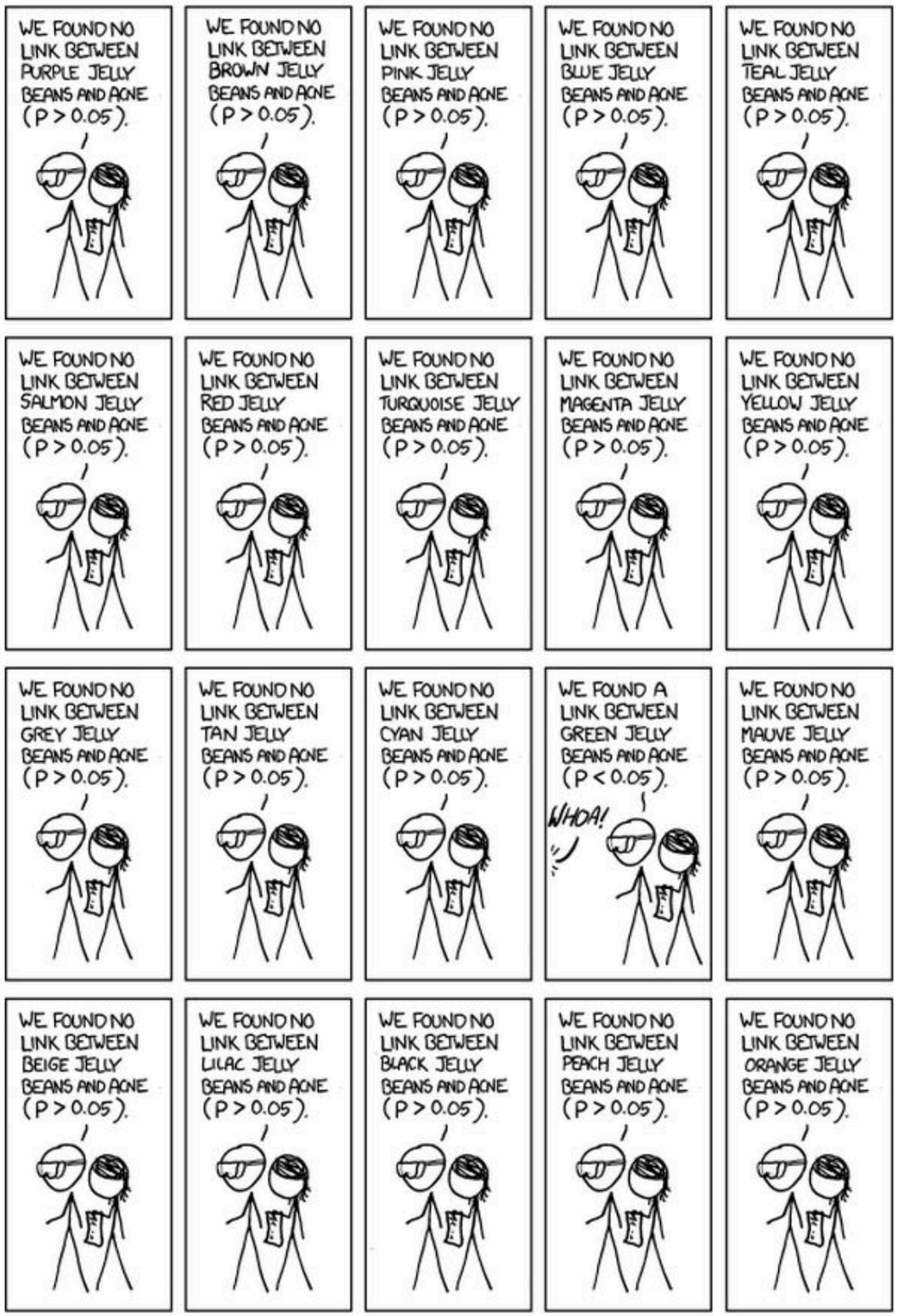
Multiple Hypothesis Testing

Scenario: Test 20 different drugs report results for the most significant treatment.

Why is this bad? 20 trials each with a small FDR means the overall FDR is much higher.

Test more hypotheses, you need to lower the acceptance threshold.

Multiple Hypothesis Testing



$$p \leq \alpha / R$$

Calculated p-value **Acceptance thresh**

R is the number of tests you run, divide acceptance threshold by number of tests

CHIDATA

RCT Assumptions?

Randomly assign individuals to treatment and control

Samples are drawn uniformly and randomly

Apply the desired intervention

Treatment effects are independent

Observe results and calculate a p-value

There is a reasonable null hypothesis

Accept conclusion when p-value is below some confidence threshold (e.g., 0.05 or 0.01)

The threshold is statistically meaningful

CHIDATA

Comparing Estimated Quantities

Sample 1

stat1

Sample 2

stat2