



Hypothesis Testing With Python

True Difference or Noise?

169.61

169.88

Which is better?

Noise?

That's a question.

Mosky



- Python Charmer at Pinkoi.
- Has spoken at: PyCons in TW, MY, KR, JP, SG, HK, COSCUPs, and TEDx, etc.
- Countless hours on teaching Python.
- Own the Python packages: ZIPCodeTW, MoSQL, Clime, etc.
- <http://mosky.tw/>

Outline

1. Tests with datasets
2. How tests work
3. Common tests
4. Complete a test

The Packages

- \$ pip3 install jupyter numpy scipy sympy matplotlib ipython pandas seaborn statsmodels scikit-learn
- Or
- > conda install jupyter numpy scipy sympy matplotlib ipython pandas seaborn statsmodels scikit-learn
- Or
- \$ curl -fLo Pipfile.lock <https://raw.githubusercontent.com/moskytw/data-science-with-python/master/Pipfile.lock>
- \$ pipenv sync

Tests with datasets

Hypothesis testing

- “The means of two populations are equal.”
- $p\text{-value} = 0.90$
- $p\text{-value} = 0.80$
- ...
- $p\text{-value} = 0.05$
- $p\text{-value} = 0.01$

Hypothesis testing

- “The means of two populations are equal.”
- Case 1: $p\text{-value} \geq 0.05$
 - “Hmmm ... can't tell.”
- Case 2: $p\text{-value} < 0.05$
 - “Not equal! So extreme!”

Hypothesis testing in a “null” taste

- $\langle \text{null hypothesis} \rangle$
- Case 1: $p\text{-value} \geq 0.05$
 - Can't reject $\langle \text{null hypothesis} \rangle$.
- Case 2: $p\text{-value} < 0.05$
 - Reject $\langle \text{null hypothesis} \rangle$.

Hypothesis testing in an “alternative” taste

- <*alternative hypothesis*>, \equiv not <*null hypothesis*> usually.
- Case 1: $p\text{-value} \geq 0.05$
 - Can't accept <*alternative hypothesis*>.
- Case 2: $p\text{-value} < 0.05$
 - Accept <*alternative hypothesis*>.

Hypothesis testing in “-+” taste

- “The case is negative.”
- Case 1: $p\text{-value} \geq \alpha$
 - “Hmmm ... somehow negative.”
- Case 2: $p\text{-value} < \alpha$
 - “Positive! So extreme!”

P-value, α , wording, and summary

p-value & α	Wording	Summary
$p\text{-value} < 0.001$	Very significant	***
$p\text{-value} < 0.01$	Very significant	**
$p\text{-value} < 0.05$	Significant	*
$p\text{-value} \geq 0.05$	Not significant	ns

Go with the notebooks

- *01_tests_with_simulated_datasets.ipynb*
- *02_tests_with_actual_datasets.ipynb*
- The notebooks are available on <https://github.com/moskytw/hypothesis-testing-with-python>.

How tests work

Seeing is believing

- $p\text{-value} = 0.0027 (< 0.01)$
 - 
- $p\text{-value} = 0.0271 (0.01\text{--}0.05)$
 -  ?  ? ? ? ?
- $p\text{-value} = 0.2718 (\geq 0.05)$
 - ? ? ? ? ? ?
- *03_how_tests_work.ipynb*

Common tests

The cheat sheet

- If testing independence:
 - If total sample size < 1000 , or more than 20% of cells have expected frequencies < 5 , **Fisher's exact test**.
 - Else, **Chi-squared test**.
- If testing difference:
 - If median is better, don't want to trim outliers, variable is ordinal, or any group size ≤ 20 :
 - If groups are paired, **Wilcoxon signed-rank test**.
 - If groups are independent, **Mann–Whitney U test**.
 - Else:
 - If groups are paired, **Paired Student's t-test**.
 - If groups are independent, **Welch's t-test**, not Student's.

More cheat sheets & references

- More cheat sheets:
 - Selecting Commonly Used Statistical Tests – Bates College
 - Which statistical test should I use? – University of Sheffield
 - Choosing a statistical test – HBS
- References:
 - Fisher's exact test of independence – HBS
 - Statistical notes for clinical researchers – Restor Dent Endod
 - Nonparametric Test and Parametric Test – Minitab
 - Advantages and limitations – Welch's t-test – Wikipedia
 - Dependent t-test for paired samples – Student's t-test – Wikipedia

The tests in Python

- *04_common_tests.ipynb*

Complete a test

Are p-value & α enough?

- sample size?
- power?
- β ?
- raw effect size?
- ? ? ? ? ?

Raw effect size, α , and β

- DSM5: The case for double standards – James Coplan, M.D.
 - The figures explain *raw effect size*, α , and β and perfectly, but due to the copyright, only put the link here.
 - “The distance between the means”: *raw effect size*
 - “FP”: α
 - “FN”: β

- Given α , β , raw effect size, get the sample size.
- Given α , raw effect size, sample size, get the β .
- Increase sample size to decrease α , β , or raw effect size.

Confusion matrix, where $A = 00_2 = C[0, 0]$

		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
actual positive CD	false negative C	true positive D	

False positive rate = B/AB = 4/(96+4) = 4/100

.....

		predicted negative AC	predicted positive BD
actual negative AB	96 A	4 B	
	9 C	41 D	

Predefined acceptable confusion matrix

		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
	false negative C	true positive D	
actual positive CD			

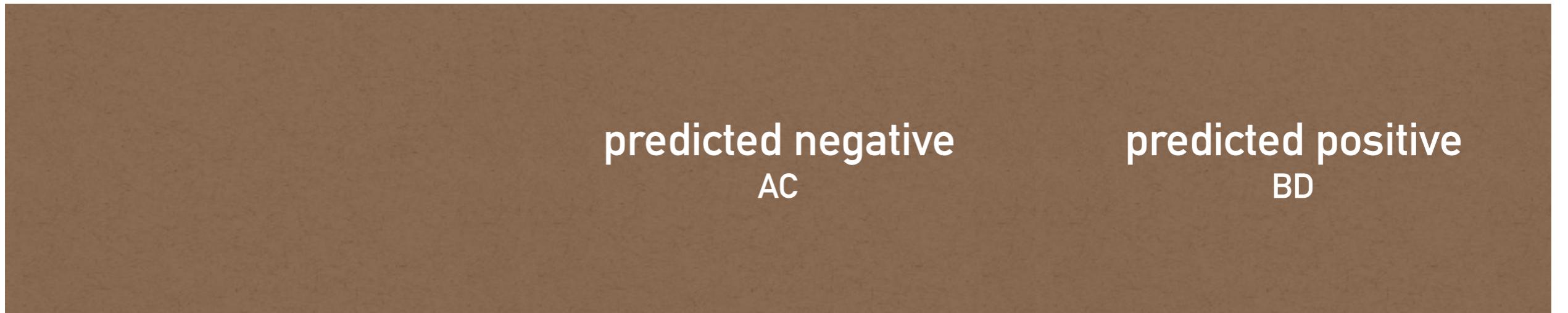
$$a = B/AB = P(BD|AB)$$

.....

		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
	false negative C	true positive D	
actual positive CD			

$$\beta = C/CD = P(AC|CD)$$

.....



actual negative
AB

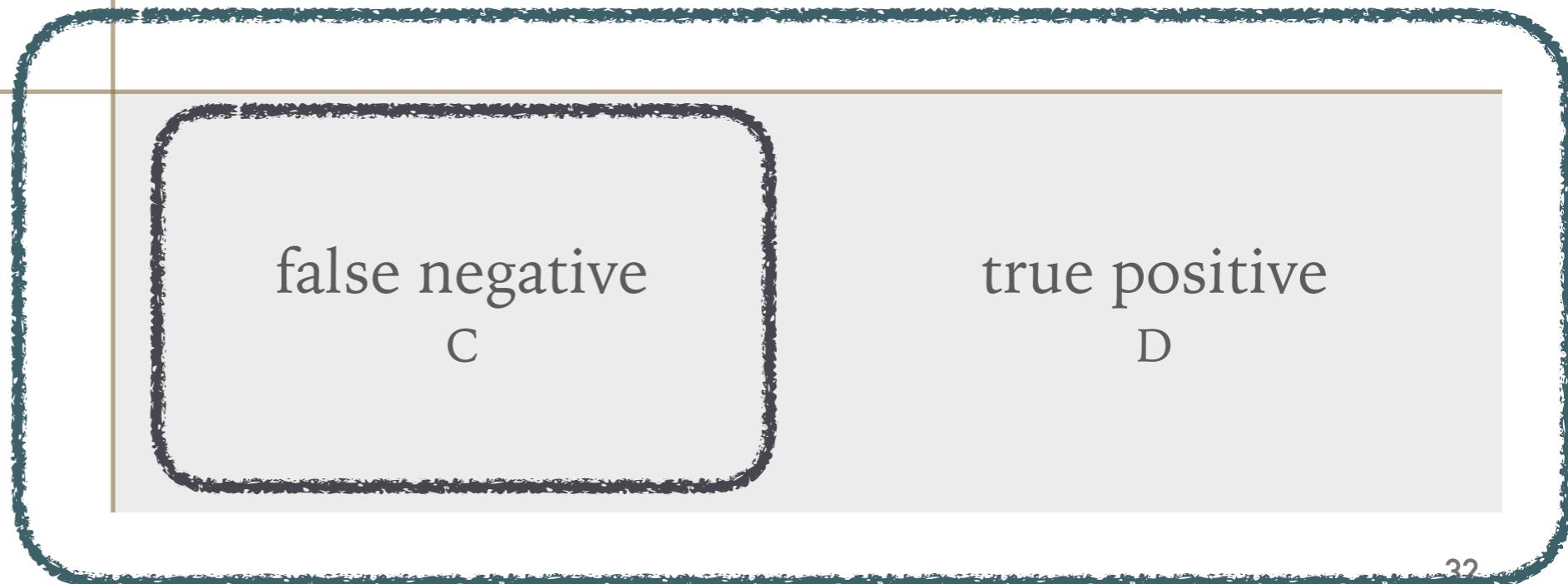
true negative
A

false positive
B

actual positive
CD

false negative
C

true positive
D



Given α , raw effect size, sample size, get the β

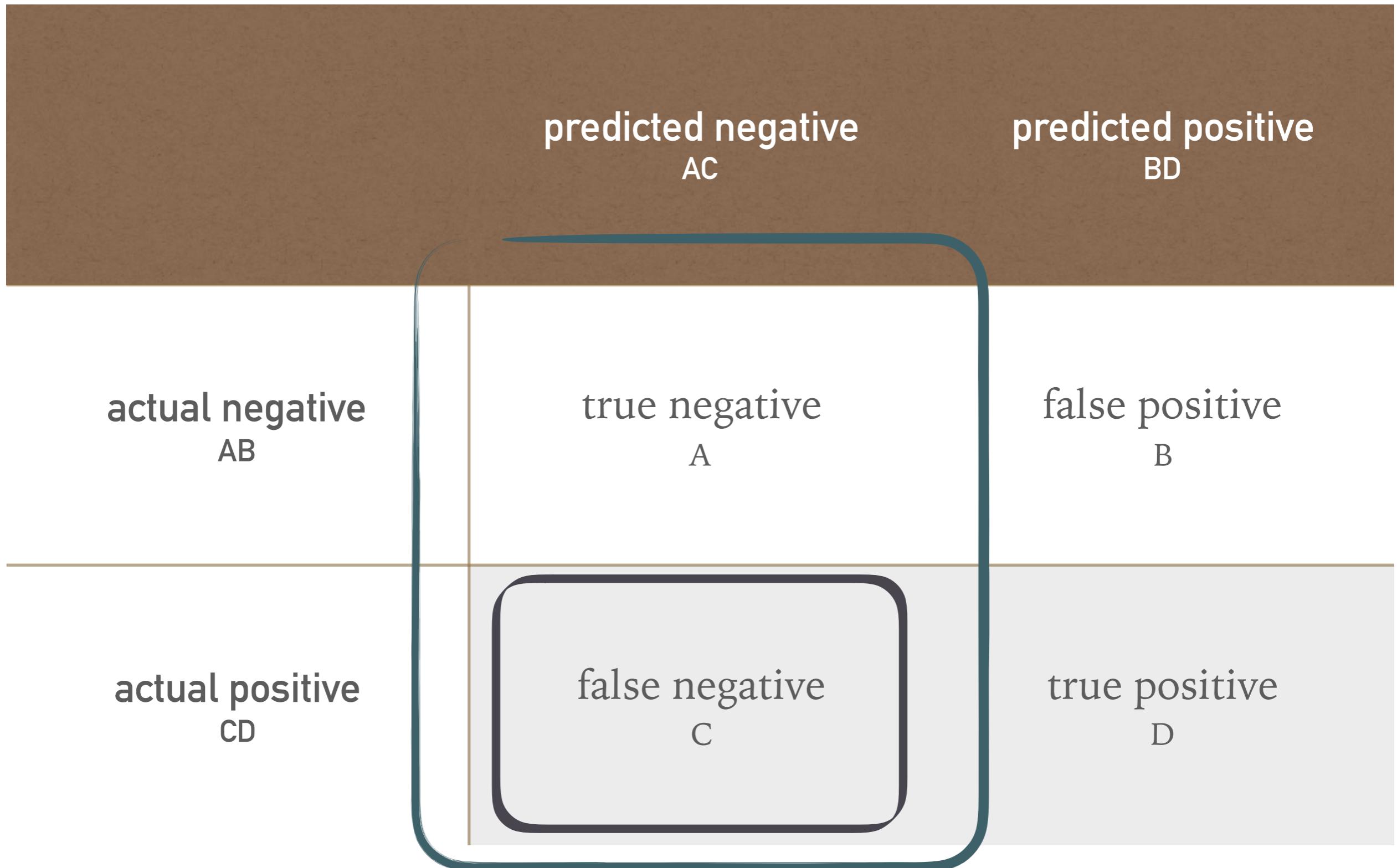
- “The conversion rates are the same.”
- $\alpha = 0.05$
- $\text{raw effect size} = 4\% - 3\% = 1\%$
- With two-proportion z-test, given:
 - $\text{sample size} = 1,000$, get $\beta = 0.86$ 🙄
 - $\text{sample size} = 10,000$, get $\beta = 0.22$
 - $\text{sample size} = 17,550$, get $\beta = 0.05$ 🎉

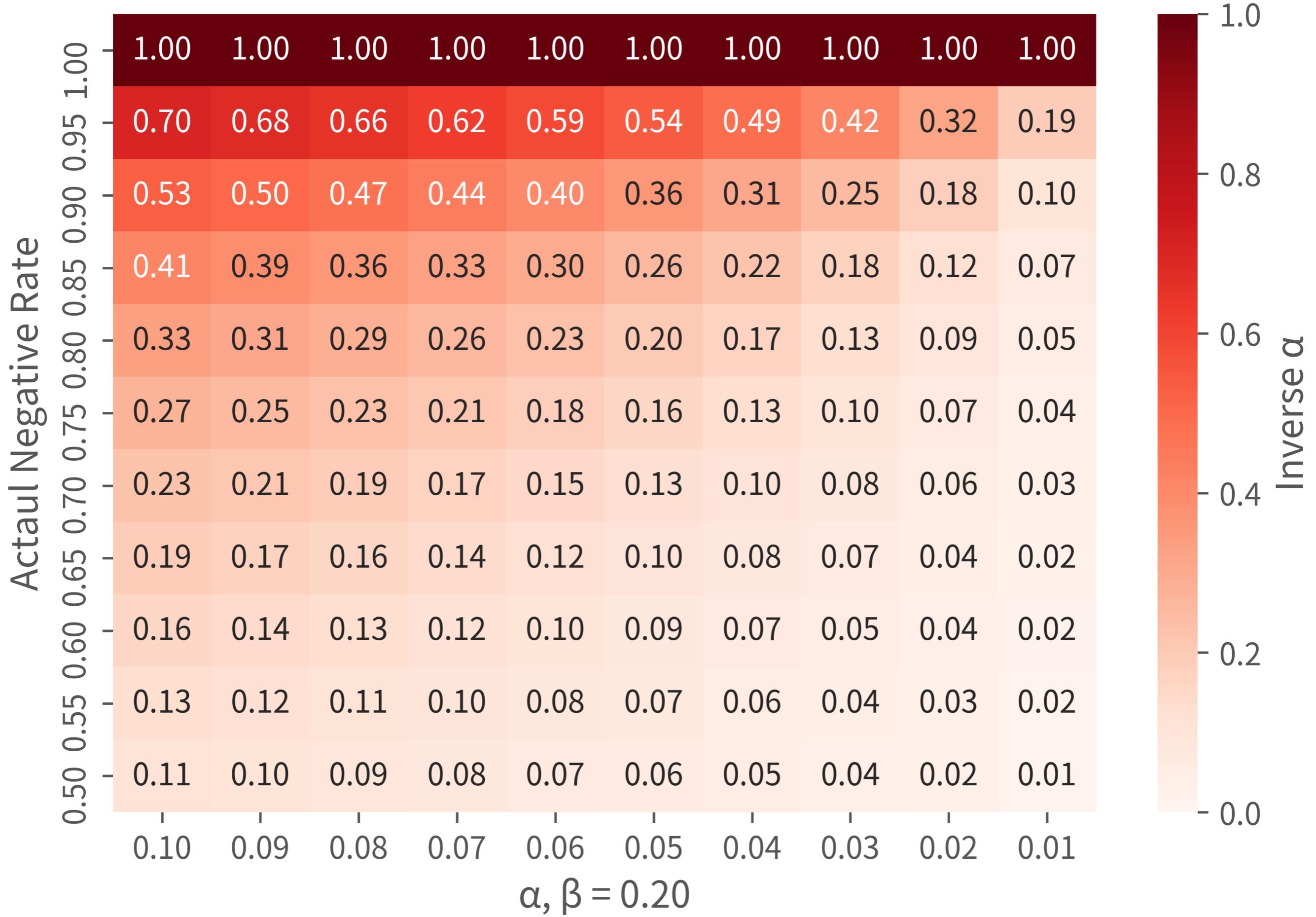
$$\text{Inverse } a = B/BD = P(AB|BD)$$

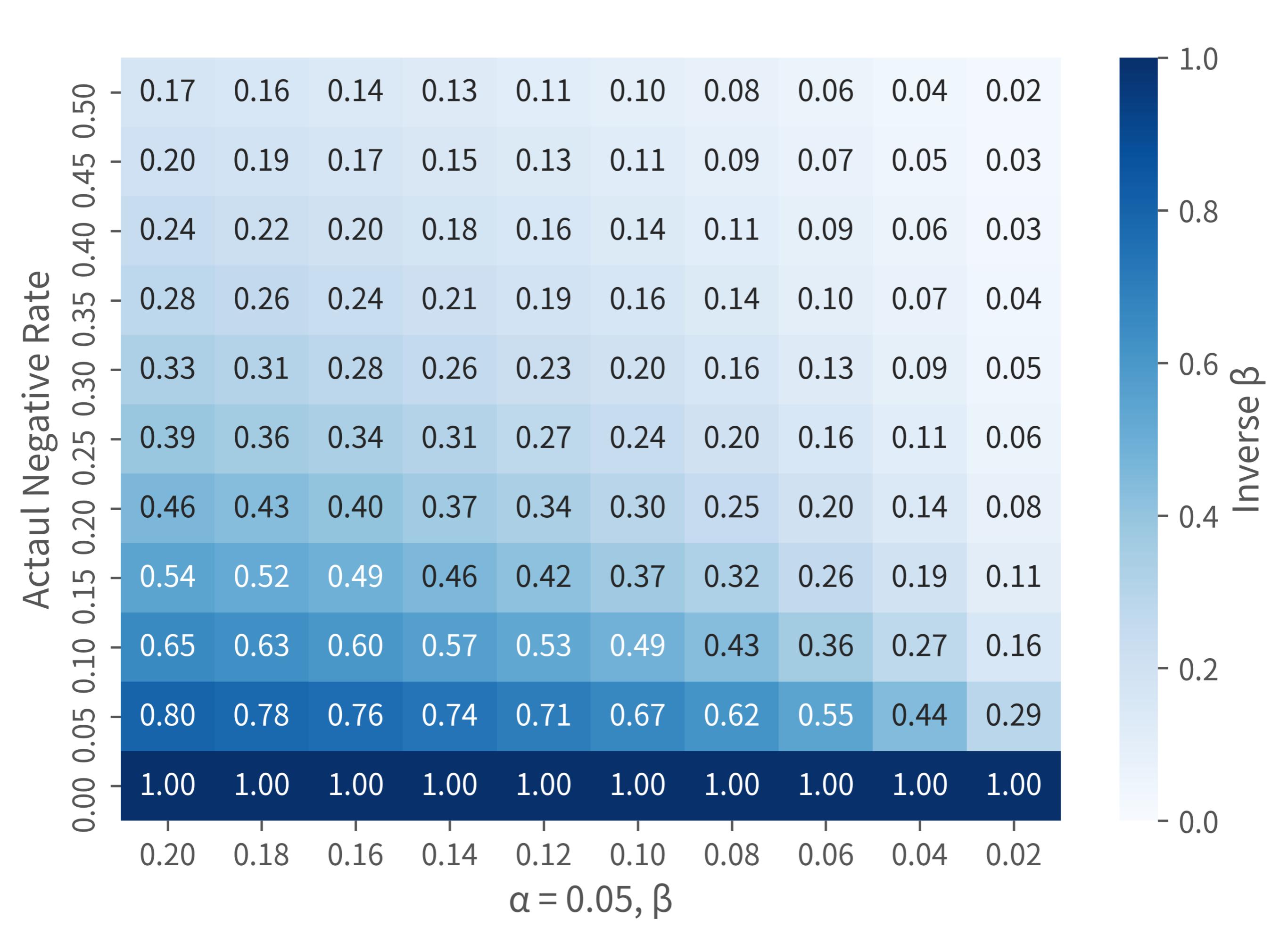
	predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B
actual positive CD	false negative C	true positive D

$$\text{Inverse } \beta = C/AC = P(CD|AC)$$

.....







Rates in predefined acceptable confusion matrix

= = = predefined

α	B/AB	significance level type I error rate	false positive rate
β	C/CD	type II error rate	false negative rate
inverse α	B/BD		false discovery rate
inverse β	C/AC		false omission rate
confidence level	A/AB	1- α	specificity
power	D/CD	1- β	sensitivity recall

Rates in confusion matrix

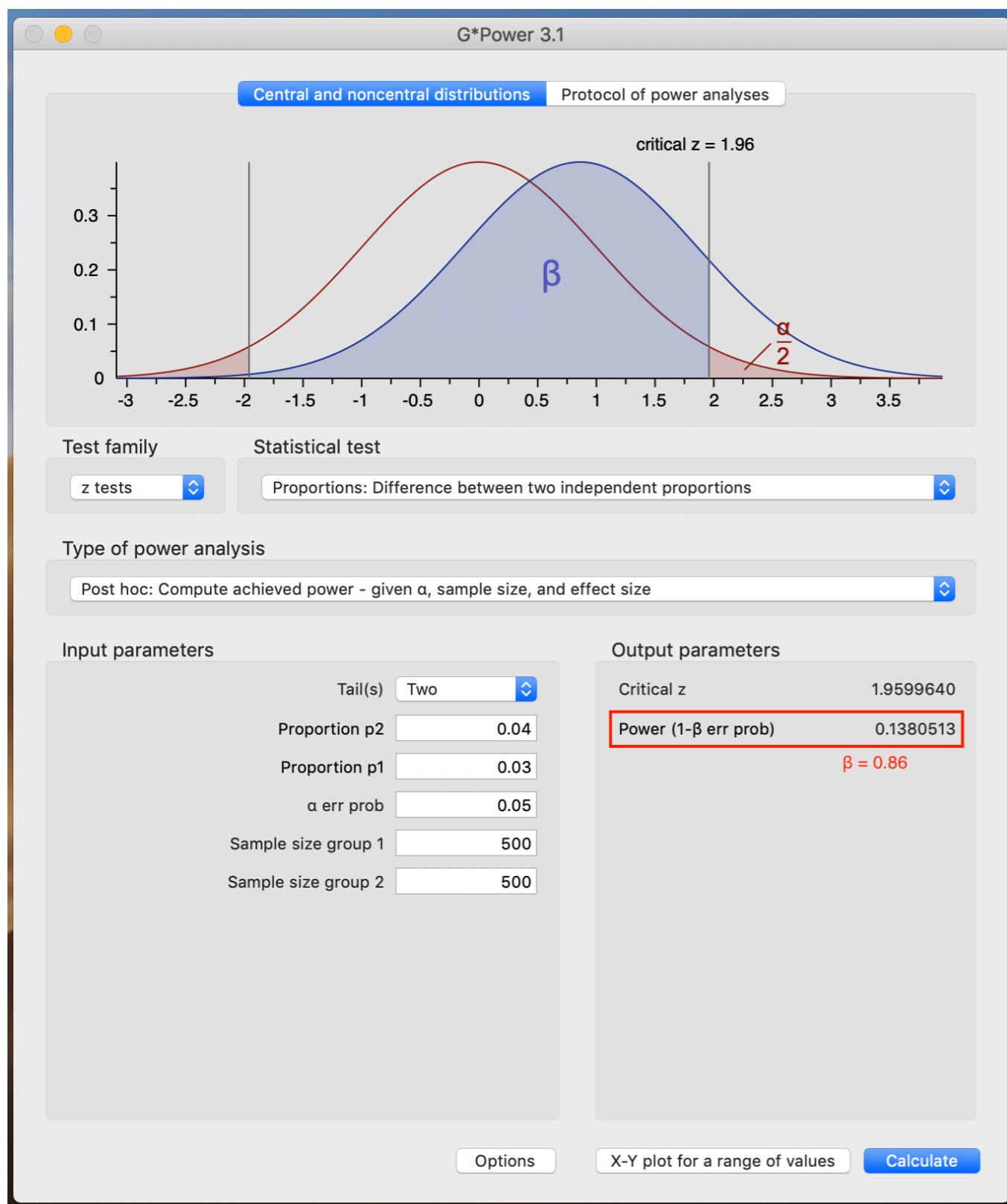
	=	=	= observed
false positive rate	B/AB		α
false negative rate	C/CD		β
false discovery rate	B/BD		inverse α
false omission rate	C/AC		inverse β
actual negative rate	AB/ABCD		
sensitivity	D/CD	recall	power
specificity	A/AB		confidence level
precision	D/BD		inverse power
recall	D/CD	sensitivity	power

Most formal steps

- State the hypothesis → what *test*.
- Estimate the *actual negative rate*.
- The *actual negative rate* → how much α, β are required.
- The $\alpha, \beta, \text{ raw effect size}$ → how large *sample size* is required.
- Still collect a sample as large as possible.
- Understand and preprocess the sample.
 - Missing data, outliers, Q–Q plot, transform, etc.
- Test and report fully.

Calculate by yourself

.....



- G*Power is awesome.
- <http://www.gpower.hhu.de/>
- *05_complete_a_test.ipynb*

Keep learning

- Statistics
 - [Seeing Theory](#)
 - [Statistics – SciPy Tutorial](#)
 - [StatsModels](#)
 - [Biological Statistics](#)
 - [Research Methods](#)
- Machine Learning
 - [Scikit-learn Tutorials](#)
 - [Standford CS229](#)
 - [Hsuan-Tien Lin](#)

Recap

- *p-value*: “tail” probability given “actual negative” hypothesis.
- *actual negative rate* does matter.
- β does matter.
- $\alpha, \beta, \text{raw effect size} \downarrow; \text{sample size} \uparrow$.
- Visualization and simulation do help.
- Bonus: `a1_figures.ipynb` .
- Let's identify noise efficiently!