



# 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, model, and probability

---

1. “The means of two populations are equal.”
2. Build a mathematical model based on the hypothesis.
3. Put the data into the model, we get:
  - $p\text{-value} = 0.90$
  - $p\text{-value} = 0.80$
  - ...
  - $p\text{-value} = 0.05$
  - $p\text{-value} = 0.01$

# 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

---

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

# 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>.

# P-value, $\alpha$ , wording, and summary

---

p-value & $\alpha$	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

---

- The notebooks are available on <https://github.com/moskytw/hypothesis-testing-with-python>.
- *01\_tests\_with\_simulated\_datasets.ipynb*
- *02\_tests\_with\_actual\_datasets.ipynb*

# 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 homogeneity:
  - 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 equality:
  - If median is better, don't want to trim outliers, variable is ordinal, or any group size  $\leq 20$ :
    - 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

# Complete a test

# Are p-value & $\alpha$ enough?

---

- raw effect size?
- $\beta$ ?
- power?
- sample size?
- ? ? ? ? ?

**False positive rate =  $P(BD|AB) = B/AB = 4/(96+4) = 4/100$**

.....

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

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

---

		predicted negative	predicted positive
actual negative	AB	AC	BD
	true negative	A	false positive
actual positive	CD	CF	DE
	false negative	C	true positive

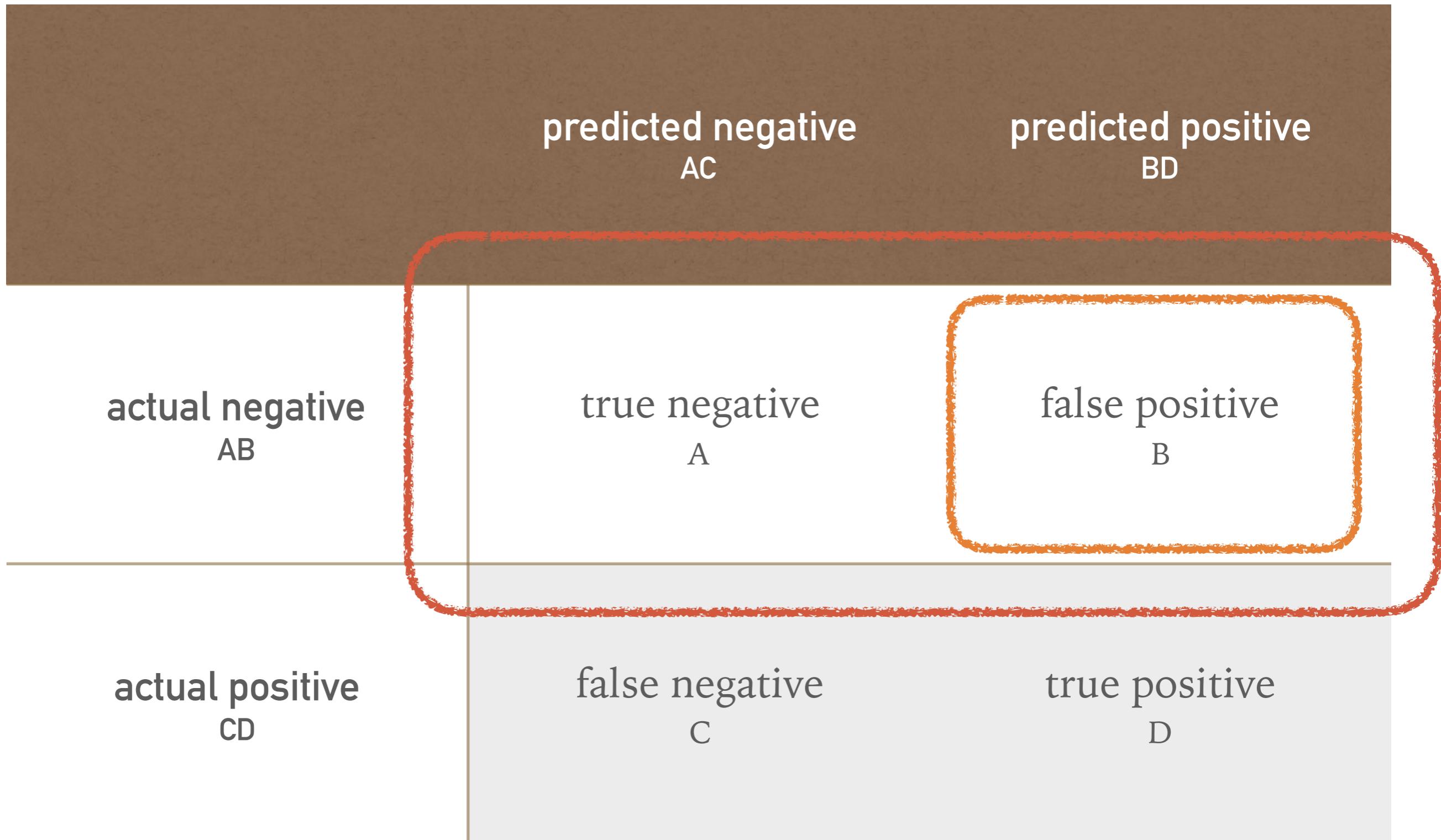
# Hypothesis testing in negative-positive taste

---

- “The case is negative.”
- Case 1:  $p\text{-value} \geq \alpha$ 
  - Can't accept positive.
- Case 2:  $p\text{-value} < \alpha$ 
  - Accept positive.

$p\text{-value} = P(BD|AB) = P(\text{predicted positive}|\text{actual negative})$

.....



$$a = P(BD|AB) = P(\text{predicted positive}|\text{actual negative})$$

.....

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

# False positive, p-value, and $\alpha$

---

false positive rate

Calculated with the actual answer.

p-value

Calculated false positive rate  
by a reasonable hypothesis.

$\alpha$

Predefined acceptable  
false positive rate.

# Predefined acceptable confusion matrix

---

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

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

.....

		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, raw effect size, and $\beta$

---

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

- Given  $\alpha$ ,  $\beta$ , raw effect size, get the sample size.
- Given  $\alpha$ , raw effect size, sample size, get the  $\beta$ .
- Increase sample size to decrease  $\alpha$ ,  $\beta$ , or raw effect size.

# Given $\alpha$ , raw effect size, sample size, get the $\beta$

---

- “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  $\beta = 0.05$  🎉

# Complete a test

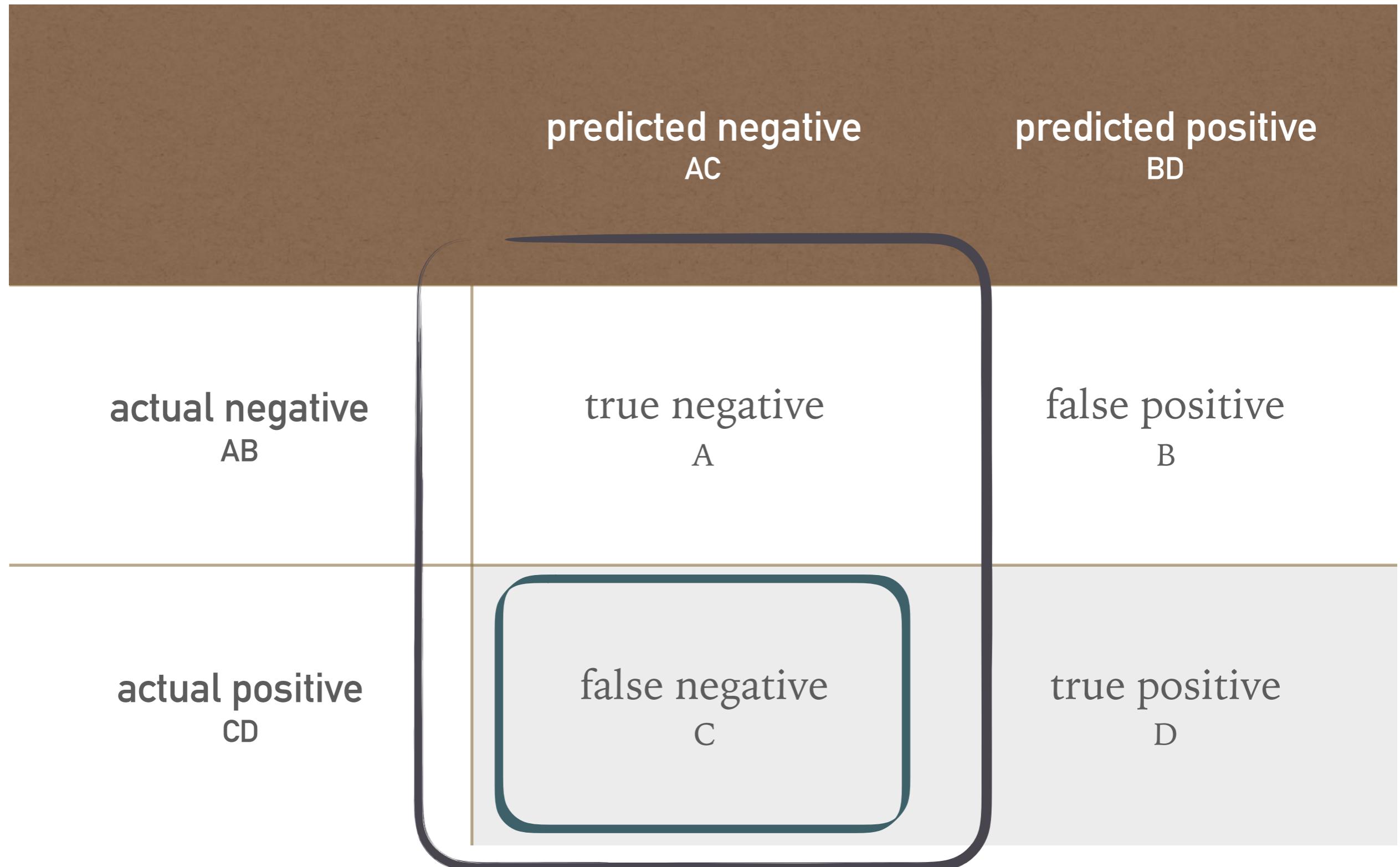
$$\text{Inverse } a = P(AB|BD) = B/BD$$

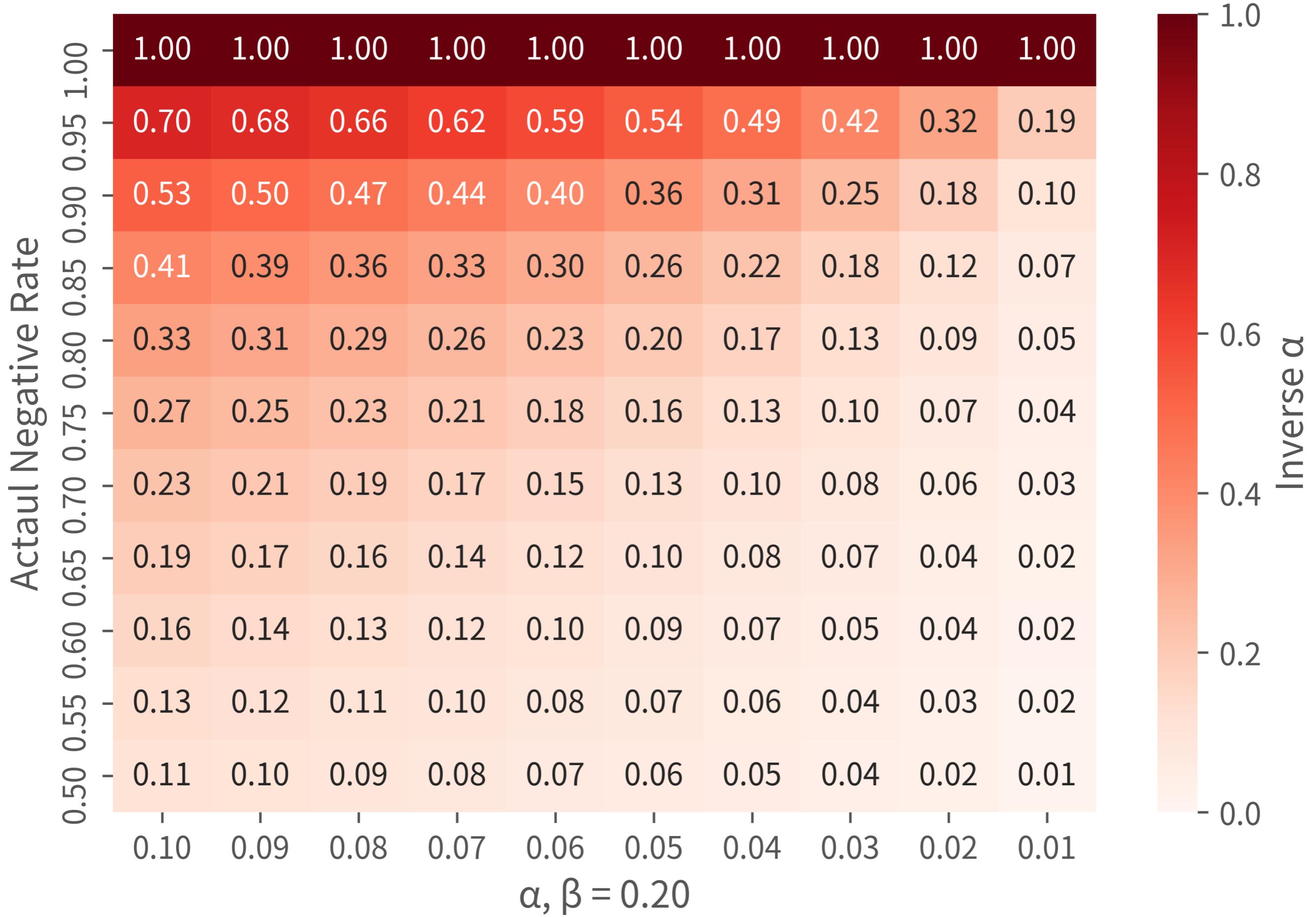
.....

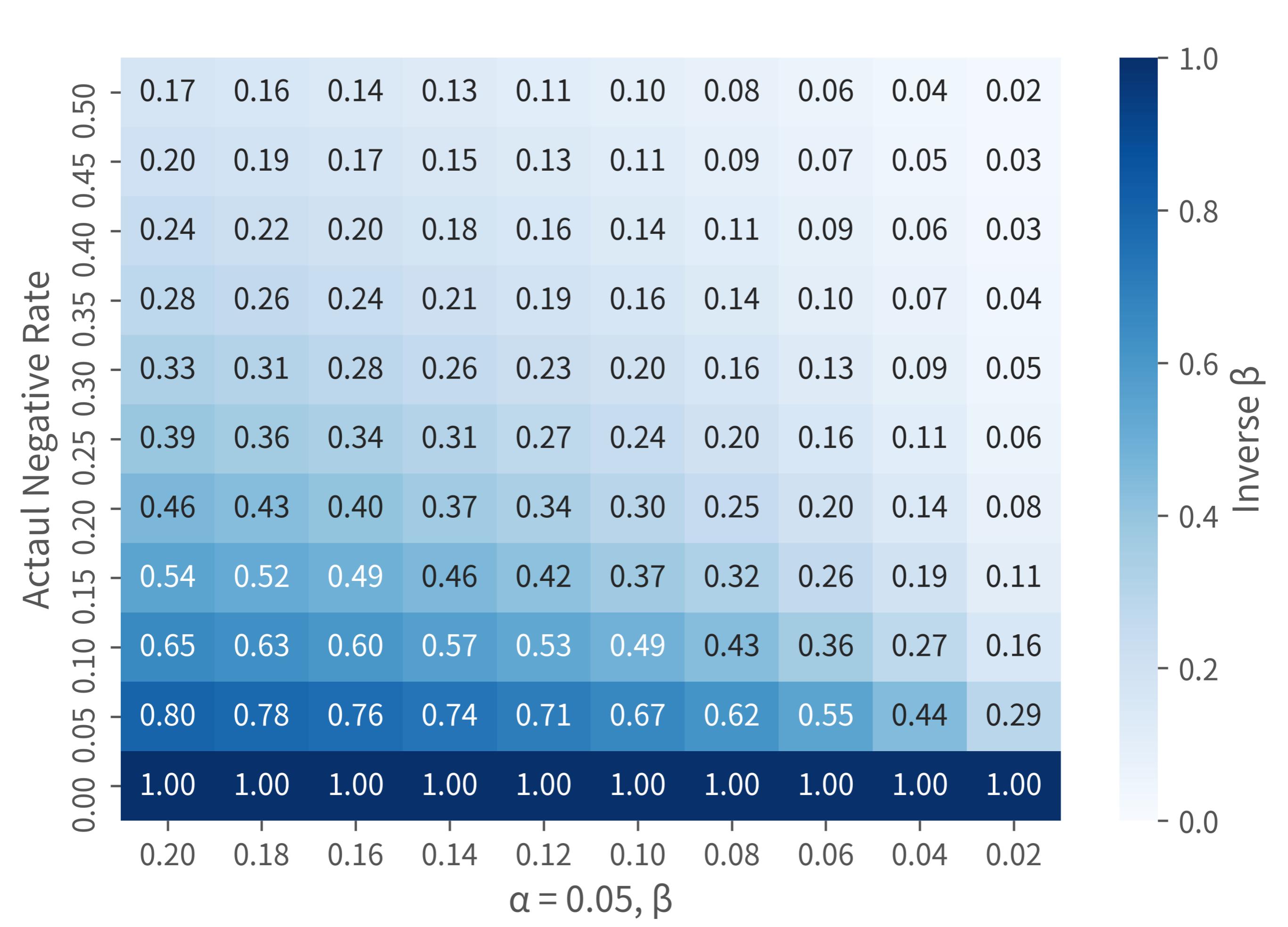
		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
	false negative C	true positive D	
actual positive CD			

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

.....







# Rates in predefined acceptable confusion matrix

= = = predefined

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

# Rates in confusion matrix

---

	=	=	= observed
false positive rate	B/AB		$\alpha$
false negative rate	C/CD		$\beta$
false discovery rate	B/BD		inverse $\alpha$
false omission rate	C/AC		inverse $\beta$
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

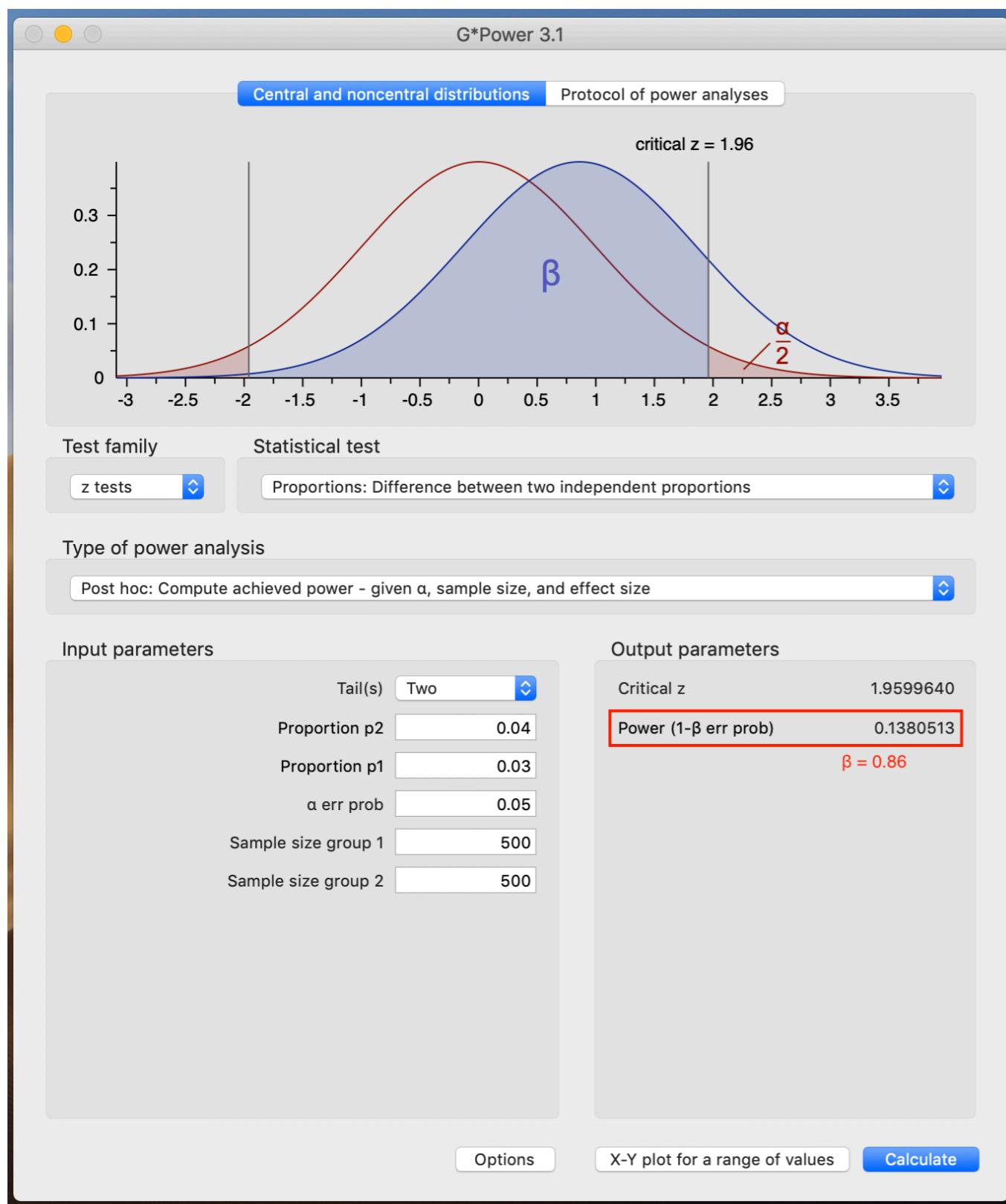
# Most formal steps

---

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

# Calculate by yourself

.....



➤ G\*Power is awesome.

➤ <http://www.gpower.hhu.de/>

# Keep learning

---

1. Seeing Theory
2. Statistics – SciPy Tutorial
3. StatsModels
4. Biological Statistics
5. “Research Method”

# Recap

---

1. p-value:
  - The tail probability given a hypothesis.
  - The false positive rate by a reasonable hypothesis.
2. Besides  $\alpha$ , we should consider  $\beta$  and actual negative rate.
3. The  $\alpha$ ,  $\beta$ , raw effect size decides the sample size.
4. Visualization and simulation do help.
5. More: 04, 05, and a1 notebooks in handouts.
6. Let's identify true difference efficiently!