



Hypothesis Testing With Python

True Difference or Noise?

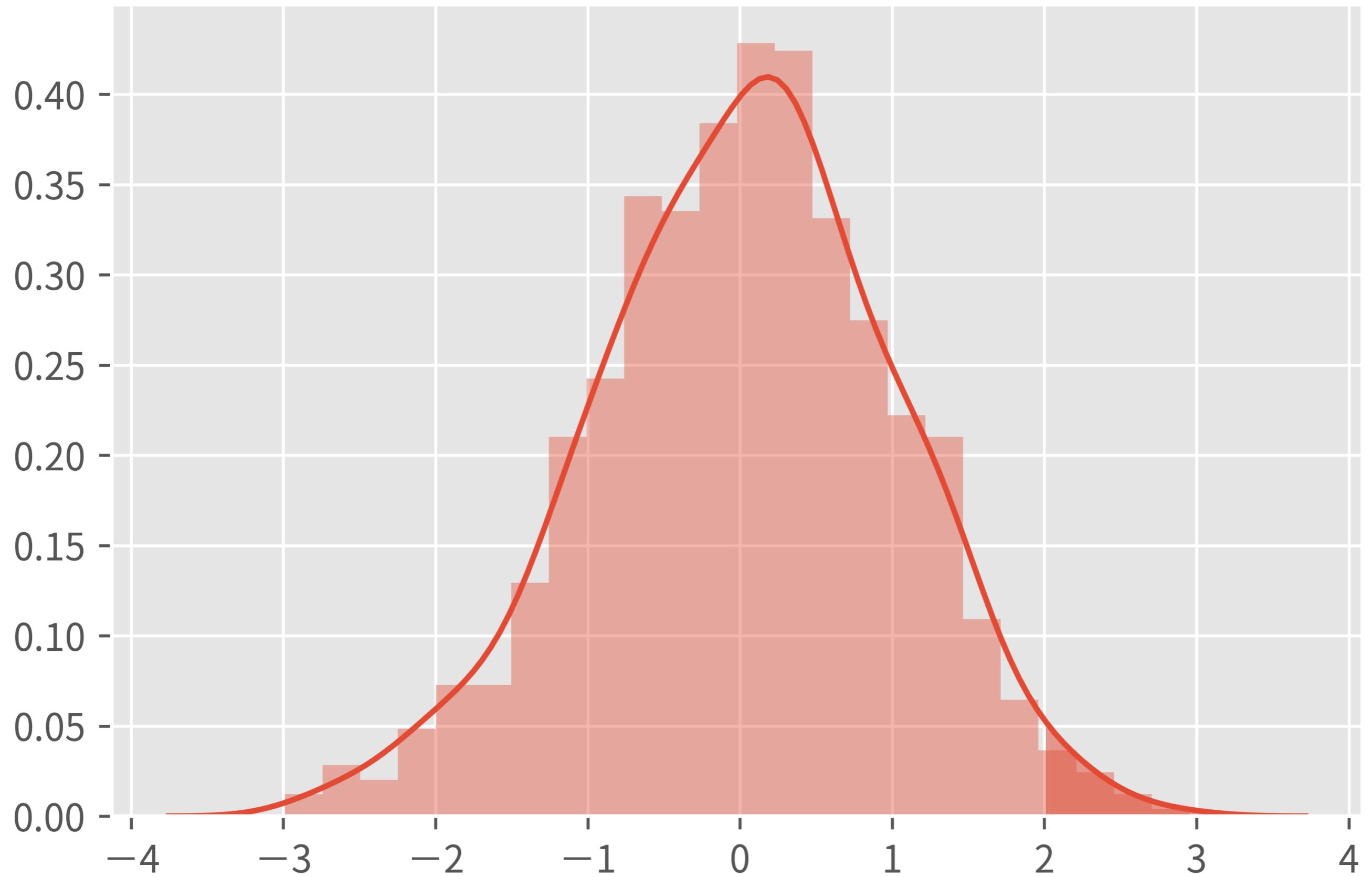
169.61

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

- Tests with simulated datasets
- Tests with actual datasets
- How tests work
- Common tests
- Complete a test

Tests with datasets

Go with the notebooks

- *01_tests_with_simulated_datasets.ipynb*
- *02_tests_with_actual_datasets.ipynb*

P-value & α

.....

P-value & α

Wording

p-value < 0.01

Very significant

p-value < 0.05

Significant

p-value ≥ 0.05

Not significant

How tests work

Seeing is believing

► $p\text{-value} = 0.0123$



► $p\text{-value} = 0.0314$



► $p\text{-value} = 0.2718$



► `03_how_tests_work.ipynb`

► At least, parametric tests.

Fair coin testing

- “The coin is fair.”
- Case 1: Toss the coin 100 times, comes up 53 heads.
 - “Hmmm ...”
- Case 2: Toss the coin 100 times, comes up 87 heads.
 - “Not fair!”

Hypothesis testing

- “The means of two populations are equal.”
- Case 1: $p\text{-value} > 0.05$.
 - “Hmmm ...”
- Case 2: $p\text{-value} \leq 0.05$.
 - “Not equal!”

Hypothesis testing in a “null” taste

- $\langle \text{null hypothesis} \rangle$
- Case 1: $p\text{-value} > a.$
 - Can't reject $\langle \text{null hypothesis} \rangle$.
- Case 2: $p\text{-value} \leq a.$
 - Reject $\langle \text{null hypothesis} \rangle$.

Hypothesis testing in an “alternative” taste

- $\langle \text{alternative hypothesis} \rangle \equiv \text{not } \langle \text{null hypothesis} \rangle$
- Case 1: $p\text{-value} > a$.
 - Can't accept $\langle \text{alternative hypothesis} \rangle$.
- Case 2: $p\text{-value} \leq a$.
 - Accept $\langle \text{alternative hypothesis} \rangle$.

Hypothesis testing in “+-” taste

- Negative.
- Case 1: $p\text{-value} > \alpha$.
 - Negative.
- Case 2: $p\text{-value} \leq \alpha$.
 - Not negative = positive.

Common tests

The cheat sheet

- If testing difference:
 - Use Welch's t-test, not Student's.
 - If groups are paired, Paired Student's t-test.
 - If median is better, don't want to trim outliers, variable is ordinal, or sample size < 20:
 - If groups are independent, Mann–Whitney U test.
 - If groups are paired, Wilcoxon signed-rank test.
- If testing independence:
 - Use Fisher's exact test.
 - If sum > 1000, Chi-squared test.

More cheat sheets & references

- More cheat sheets:
 - http://abacus.bates.edu/~ganderso/biology/resources/stats_flow_chart_v2014.pdf
 - <http://www.biostathandbook.com/testchoice.html>
 - https://www.sheffield.ac.uk/mash/what_test
- References:
 - https://en.wikipedia.org/wiki/Welch%27s_t-test#Advantages_and_limitations
 - https://en.wikipedia.org/wiki/Student%27s_t-test#Independent_two-sample_t-test
 - <http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test>
 - <http://www.biostathandbook.com/fishers.html>
 - <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426219/>

The tests in Python

- *04_common_tests.ipynb*

Complete a test

Is p-value enough?

- sample size?
- alpha?
- beta?
- effect size?
- ? ? ? ? ?

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

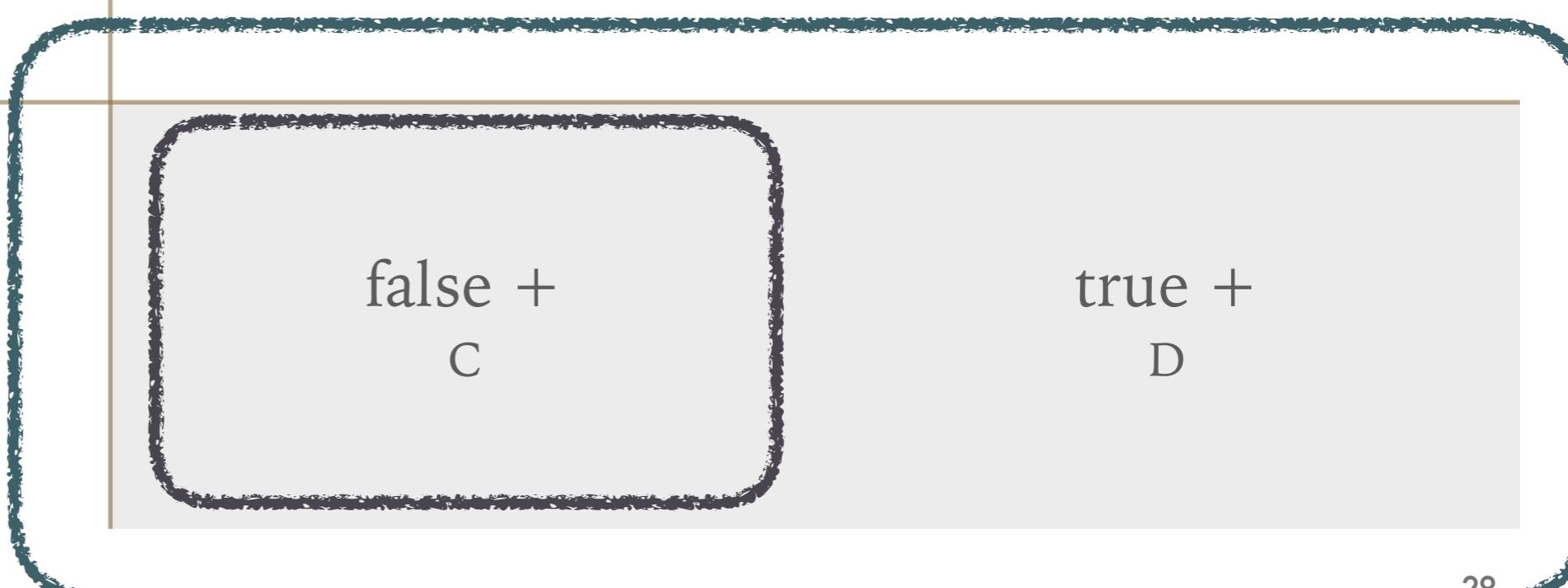
		predicted - AC	predicted + BD
actual - AB	true - A	true - A	false - B
	false + C	false + C	true + D
actual + CD			

False positive rate = B / AB = observed α

		predicted - AC	predicted + BD
actual - AB	true - A	false - B	
actual + CD	false + C	true + D	

False negative rate = C / CD = observed β

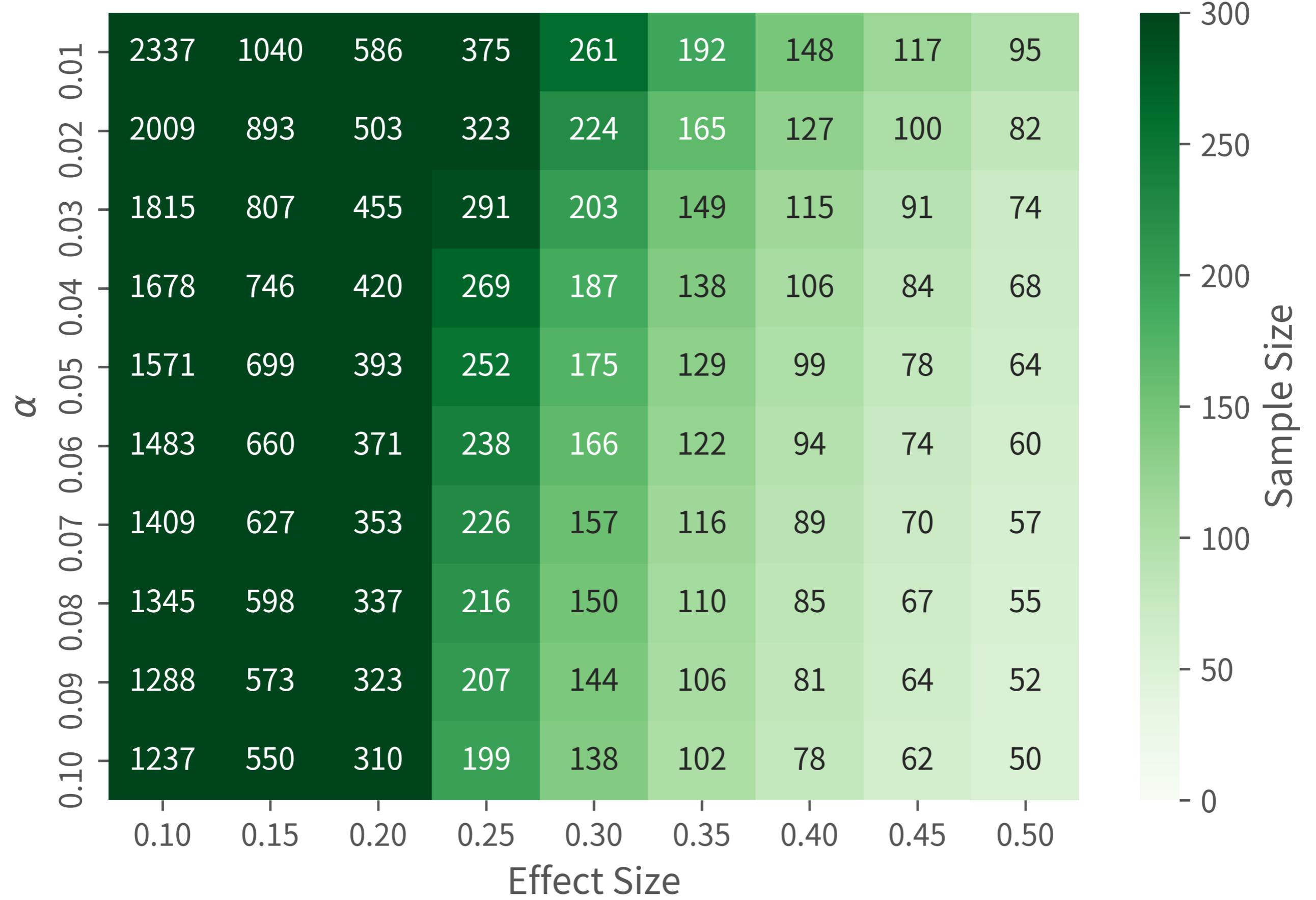
		predicted - AC	predicted + BD
actual - AB	true -	false - B	
	A		
actual + CD	false + C		true + D



The diagram illustrates a 2x2 confusion matrix. The top row is labeled "predicted -" and "predicted +". The left column is labeled "actual -" and "actual +". The cell "false +" is highlighted with a thick black border.

When sample size \uparrow ; α, β , effect size \downarrow

- Increase *sample size* to decrease $\alpha, \beta, \text{effect size}.$
 - The *effect size* is the distance between groups.
$$\text{➤ } = \frac{\mu_1 - \mu_2}{\sigma}$$
 - <http://www.drcoplan.com/dsm5-the-case-for-double-standards>
 - Explains α, β perfectly,
but due to the copyright, we only put the link here.
 - Stop increasing when low enough.

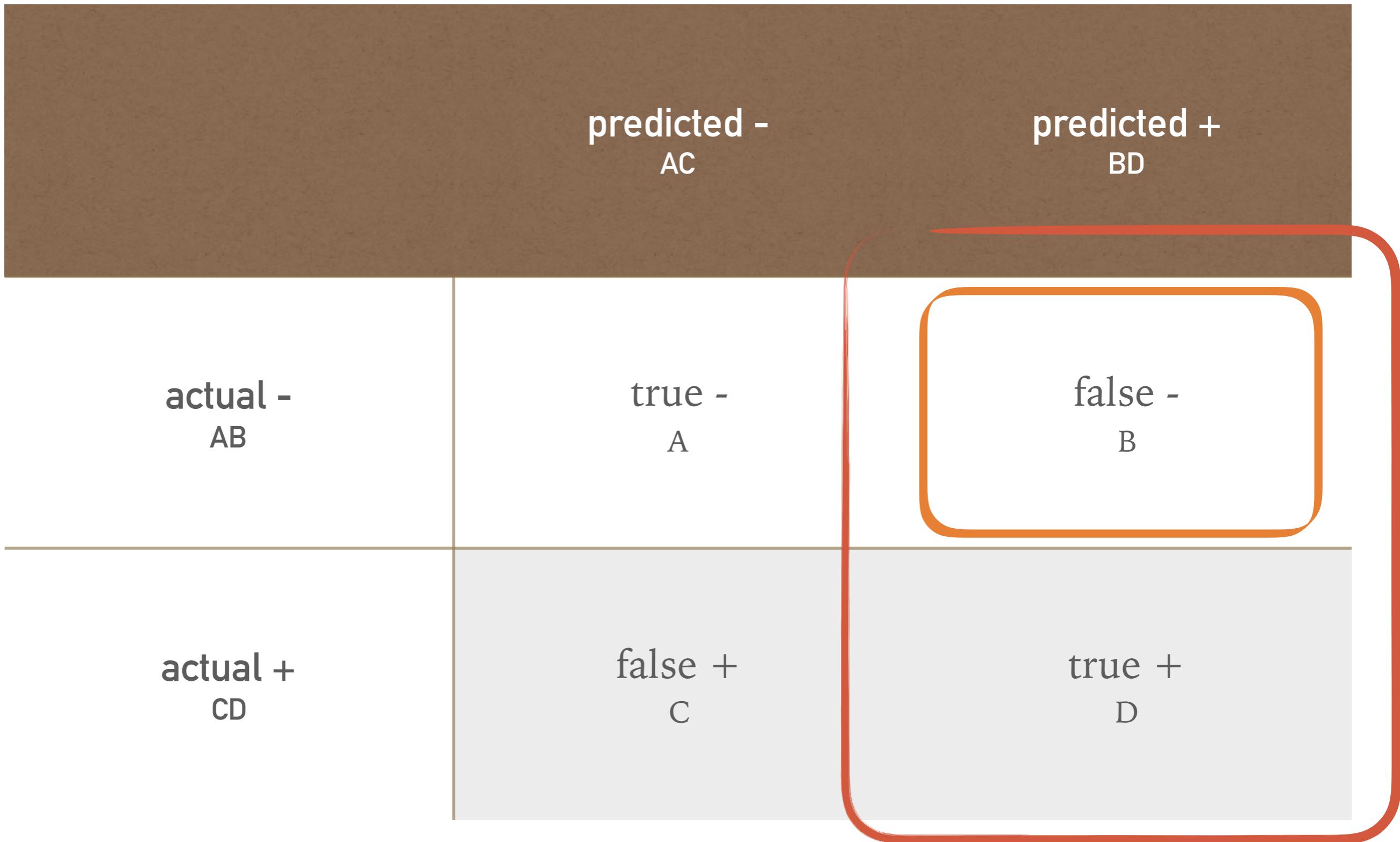


False discovery rate

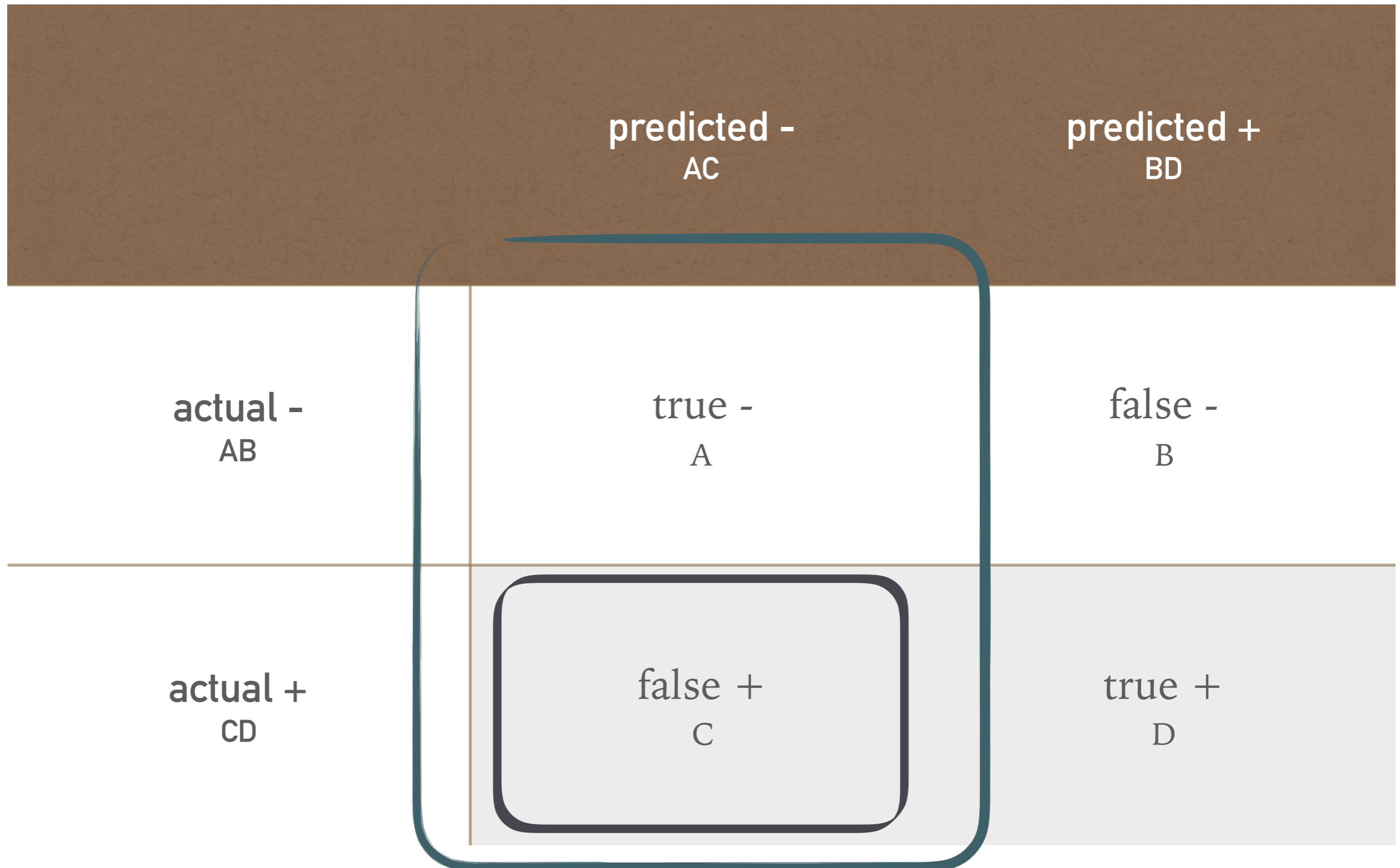
- Suppose:
 - A drug test has observed $\alpha = 1\%$ and observed $\beta = 1\%$
 - Only 0.5% of people are drug users.
- What is the probability that a person with a positive test is a drug user?

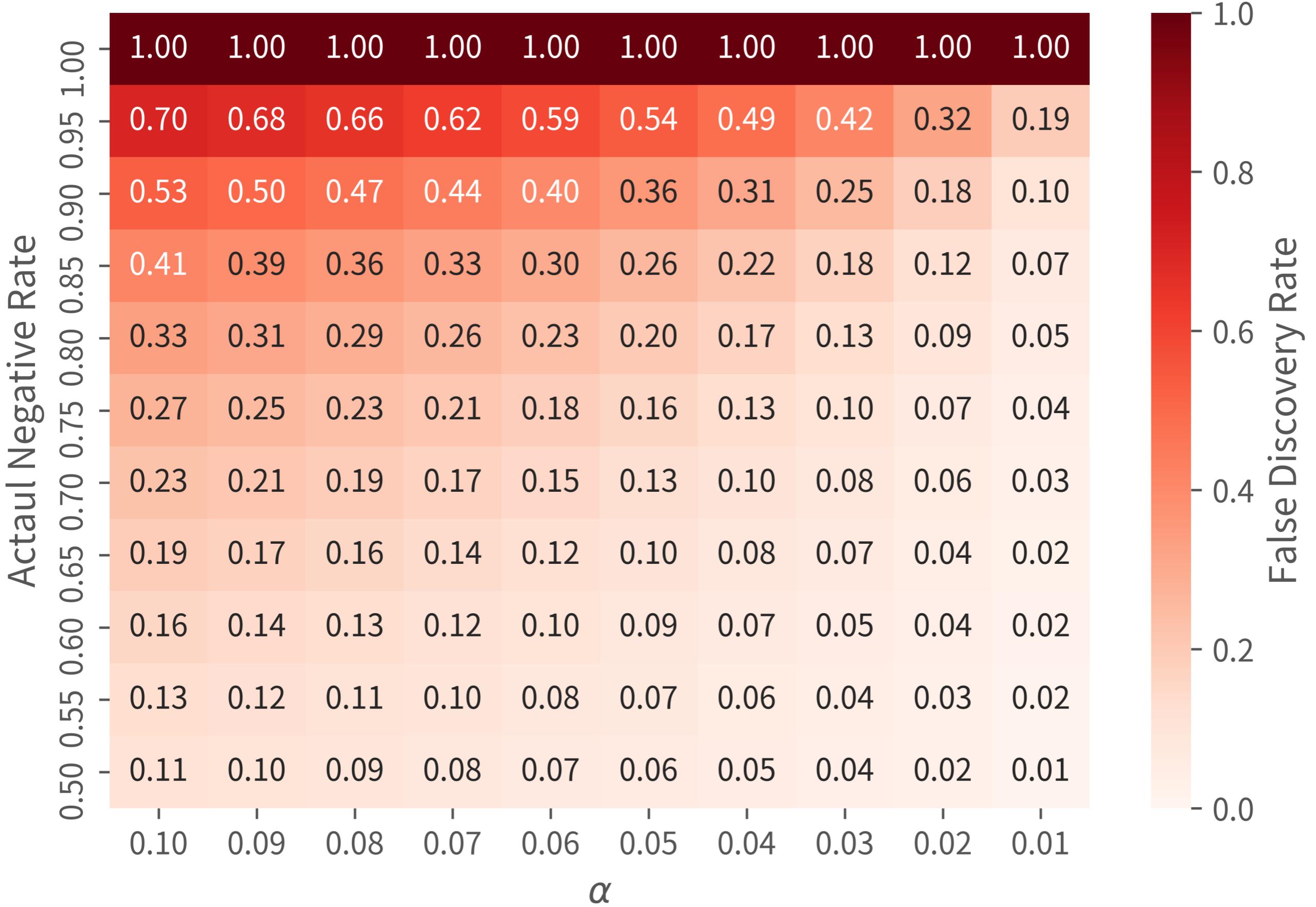
$$\begin{aligned} P(\text{User} \mid +) &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+)} \\ &= \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})} \\ &= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \\ &\approx 33.2\% \end{aligned}$$

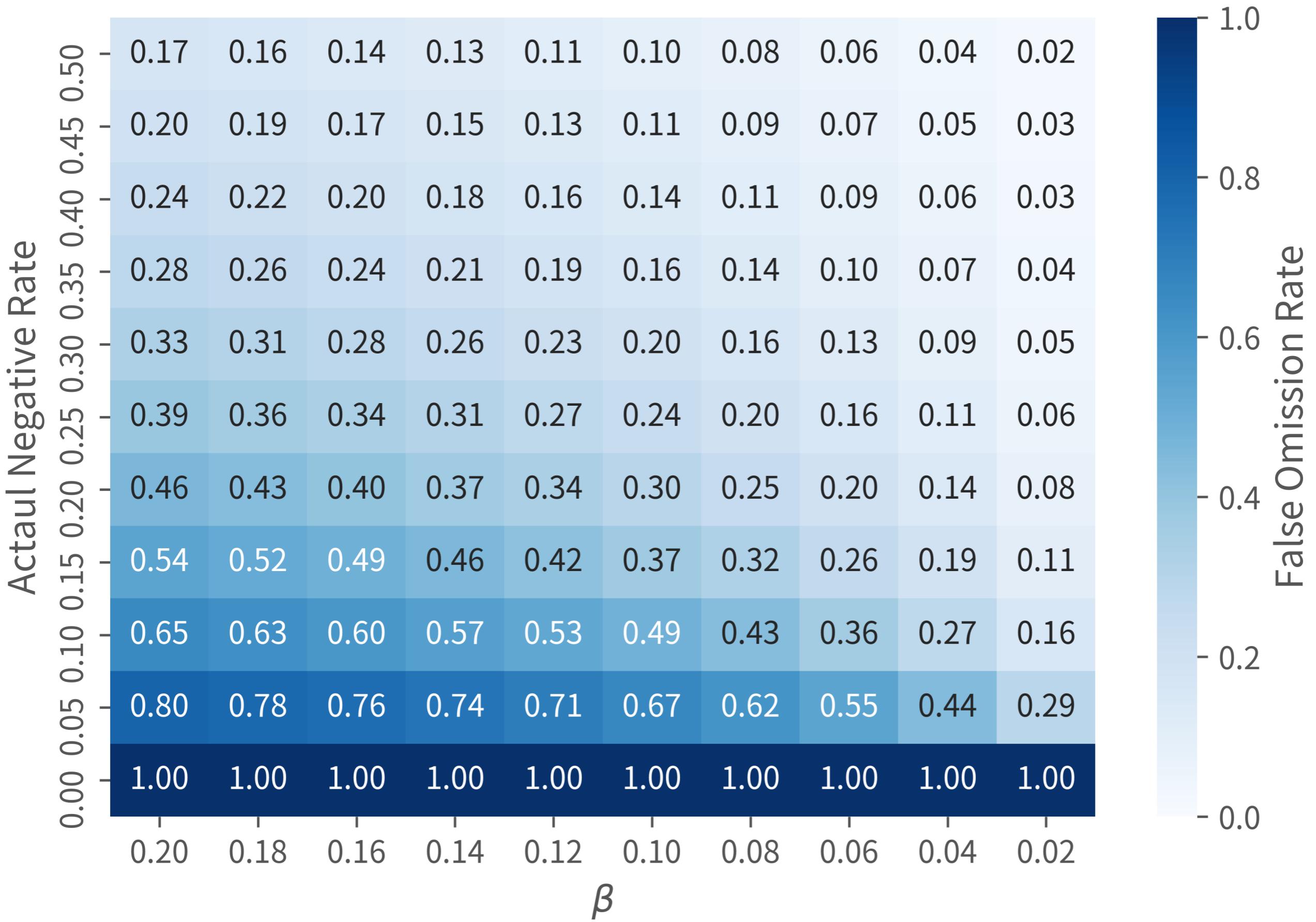
False discovery rate = B / BD



False omission rate = C / AC







Common “rates” in confusion matrix

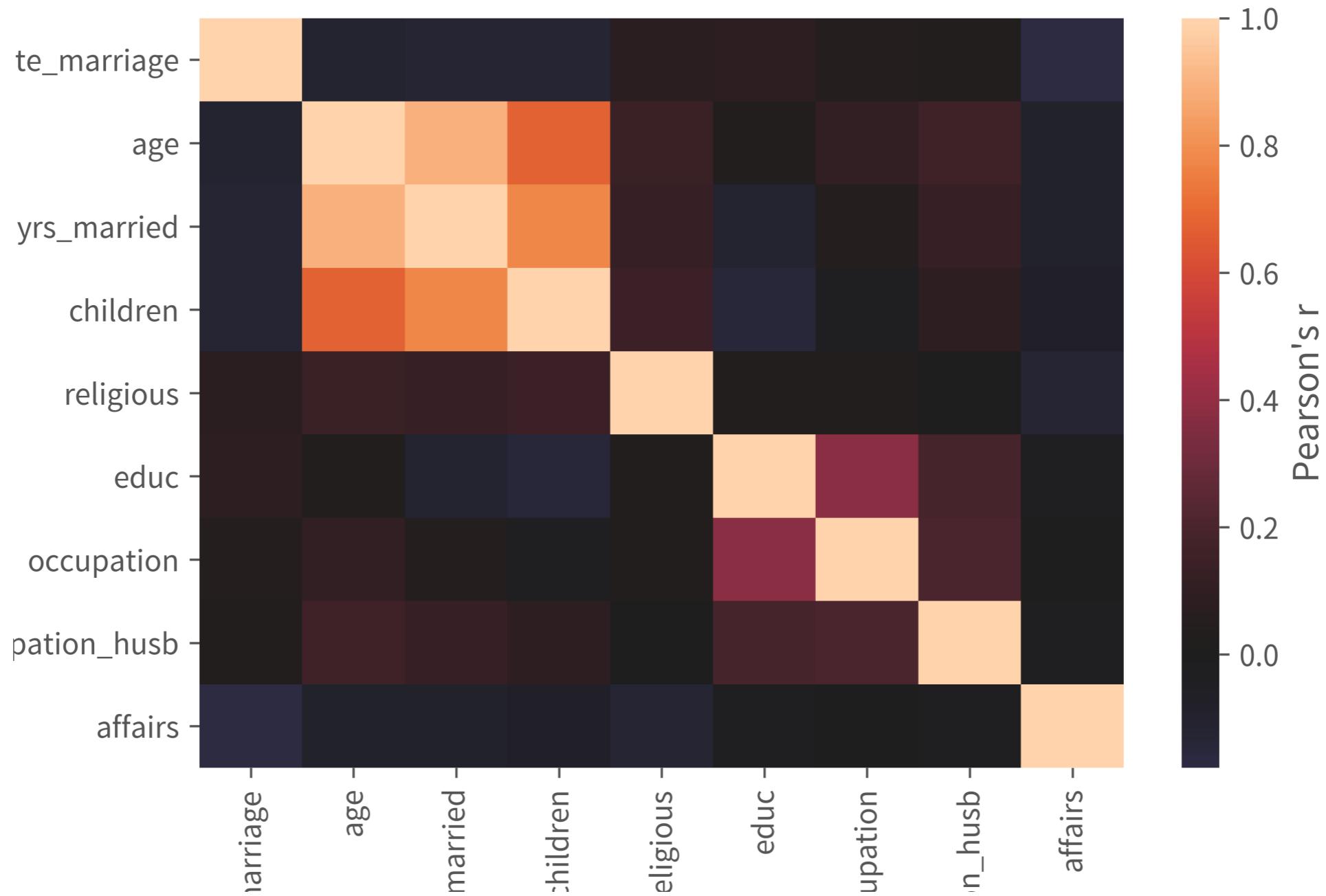
- false positive rate = B / AB = observed α
- false negative rate = C / CD = observed β
- false discovery rate = B / BD
- false omission rate = C / AC
- actual negative rate = AB / N
- sensitivity = D / CD = observed power
- **specificity** = A / AB = observed confidence level
- precision = positive predictive value = $1 - FDR$
- recall = sensitivity

Most formal steps

- State the hypothesis → what *test*.
- Estimate the *actual negative rate*.
- The *actual negative rate* → what α, β is low enough.
- The $\alpha, \beta, \text{effect size}$ → what *sample size* is required.
- Collect a sample as big as possible yet.
- Understand the sample.
 - Missing data, outliers, Q–Q plot, transform, etc.
- Report fully with *confidence interval*.
- *05_complete_a_test.ipynb*

Other statistical tools

Correlation analysis



Regression analysis

OLS Regression Results

Dep. Variable:	children	R-squared:	0.597			
Model:	OLS	Adj. R-squared:	0.597			
Method:	Least Squares	F-statistic:	9437.			
Date:	Thu, 05 Jul 2018	Prob (F-statistic):	0.00			
Time:	02:15:43	Log-Likelihood:	-8430.3			
No. Observations:	6366	AIC:	1.686e+04			
Df Residuals:	6364	BIC:	1.688e+04			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	0.0259	0.018	1.429	0.153	-0.010	0.062
yrs_married	0.1522	0.002	97.142	0.000	0.149	0.155
Omnibus:	449.258	Durbin-Watson:		1.972		
Prob(Omnibus):	0.000	Jarque-Bera (JB):		709.624		
Skew:	0.559	Prob(JB):		8.07e-155		
Kurtosis:	4.193	Cond. No.		18.5		

Keep learning

- Statistics
 - Biological Statistics
 - scipy.stats + StatsModels
 - Research Methods
- Machine Learning
 - Scikit-learn Tutorials
 - Standford CS229
 - Hsuan-Tien Lin

Recap

- *p-value* – the “tail” probability given “actual -”.
- *confidence interval* – the values the middle probability maps to.
- *actual negative rate* = AB / N
- *false discovery rate* & *false omission rate* do matter.
- α , β , *effect size*, and *sample size*
- Simulation and visualization do help.
- Bonus: *a1_figures.ipynb*
- Let's identify noise efficiently!