



Hypothesis Testing With Python

True Difference or Noise?

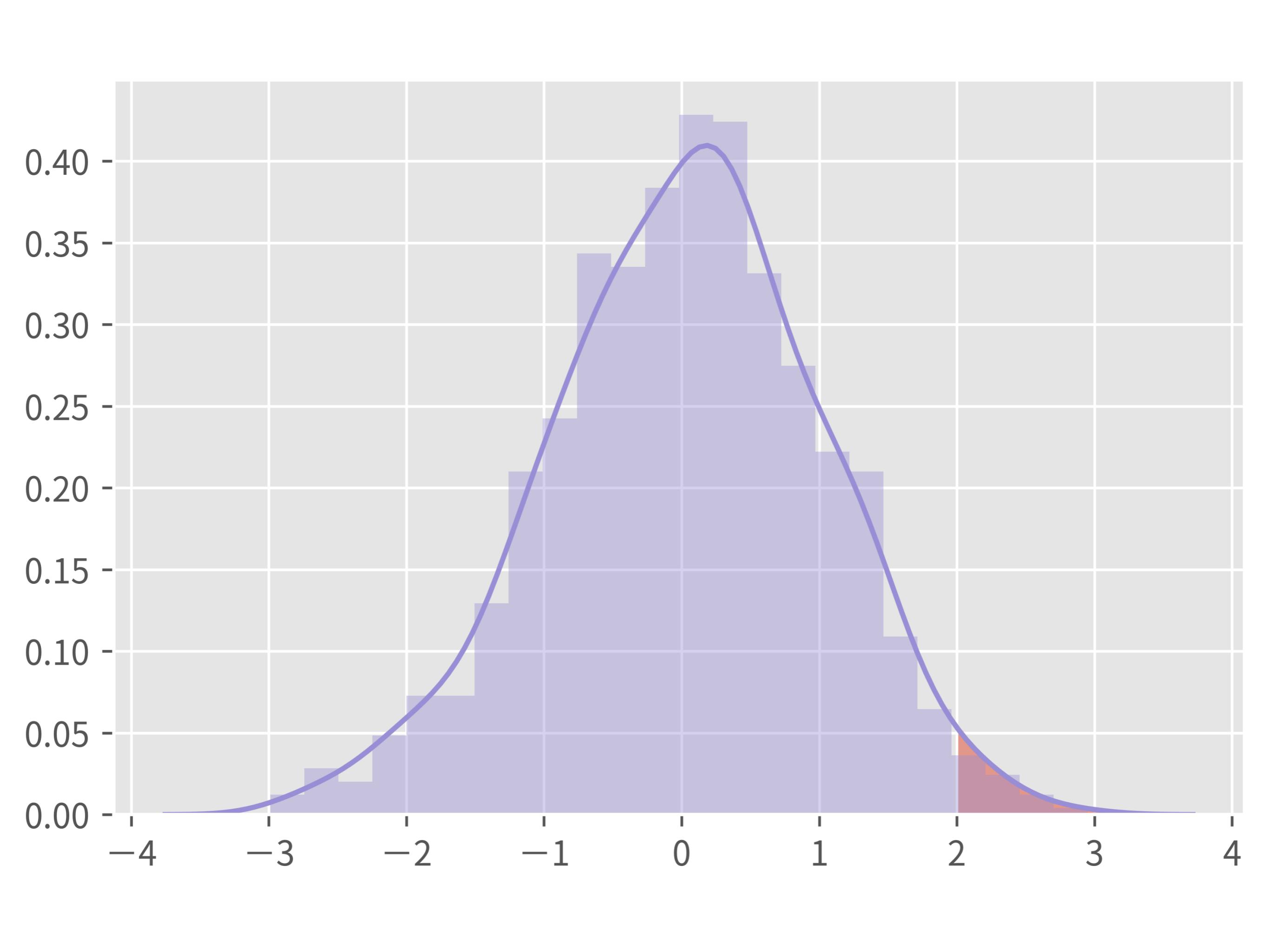
169.61

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

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

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

Fair coin testing

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

Hypothesis testing

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

Hypothesis testing in a “null” taste

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

Hypothesis testing in an “alternative” taste

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

Common tests

The cheat sheet

- If testing independence:
 - If total size < 1000, or more than 20% of cells have expected frequencies < 5, **Fisher's exact test**.
 - Use **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**.
 - 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:
 - 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:
 - <http://www.biostathandbook.com/fishers.html>
 - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426219/#_sec5title
 - <http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test>
 - https://www.sheffield.ac.uk/mash/what_test
 - https://en.wikipedia.org/wiki/Welch%27s_t-test#Advantages_and_limitations
 - https://en.wikipedia.org/wiki/Student%27s_t-test#Dependent_t-test_for_paired_samples

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 | false + B |
| | 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 matrix for a diagnostic test. The columns represent the predicted outcome (predicted - and predicted +) and the rows represent the actual condition (actual - and actual +). The matrix elements are labeled as follows:

- Top row: predicted - (AC), predicted + (BD)
- Middle row: true - (A), false + (B)
- Left column: actual - (AB)
- Bottom-left cell: false - (C)
- Bottom-right cell: true + (D)
- Bottom row: actual + (CD)

A dark blue rounded rectangle highlights the bottom-left cell (false -). A brown rectangular box surrounds the entire matrix.

When sample size ↑ ; α , β , effect size ↓

- Increase *sample size* to decrease α , β , *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>
 - The figures explain α , β perfectly,
but due to the copyright, only put the link here.
- When α , β , *effect size* are defined, get the *sample size*.
- When α , β , *sample size* are defined, get the *effect size*.

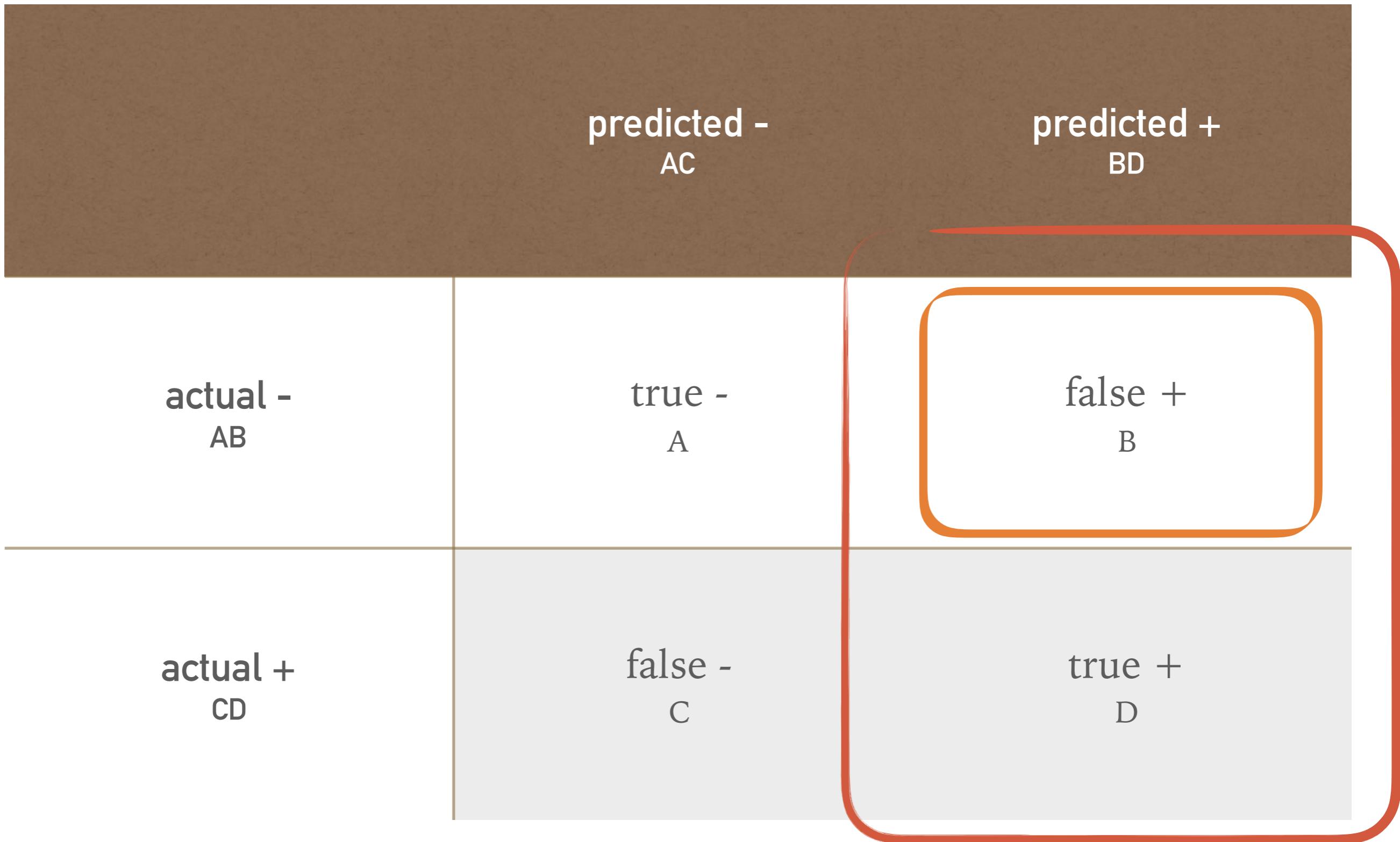


False discovery rate

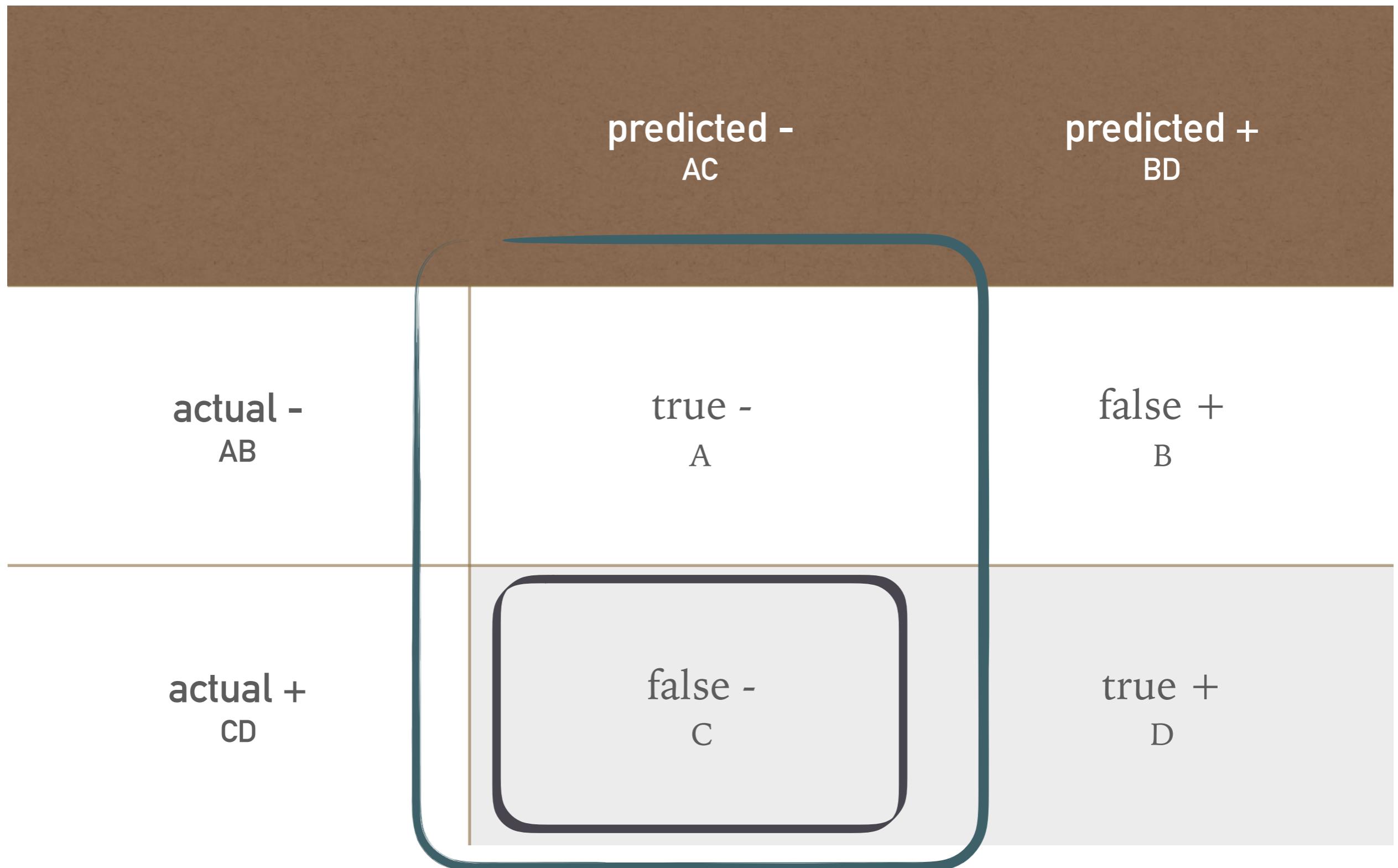
- Suppose:
 - A drug test has observed $\alpha = 1\%$ and observed $\beta = 1\%$
 - 99.5% of people are *not* drug users.
- What is the probability that a person with a positive test is a drug user?

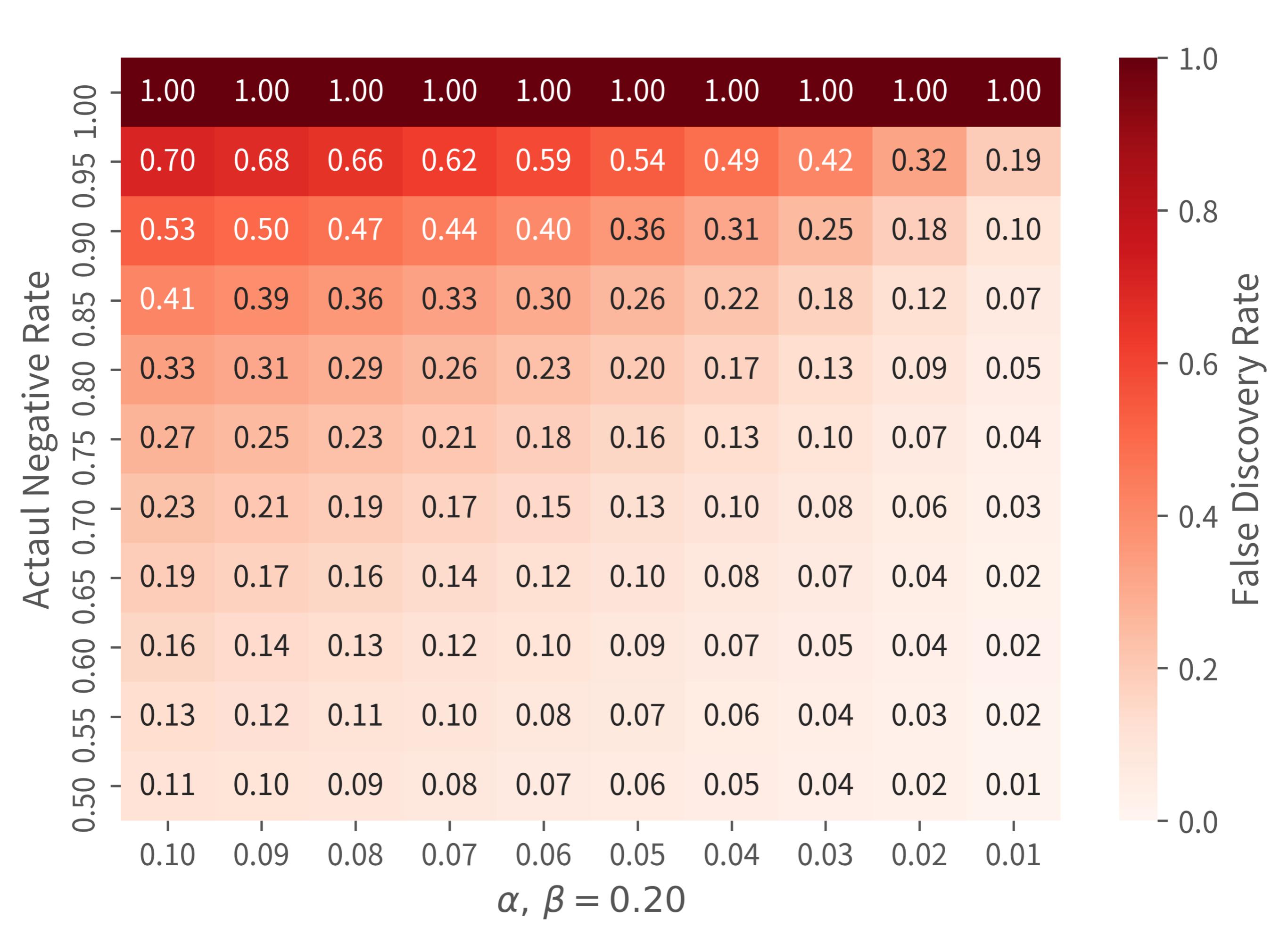
$$\begin{aligned}P(\text{non-user} \mid +) &= \frac{P(+ \mid \text{non-user})P(\text{non-user})}{P(+)} \\&= \frac{P(+ \mid \text{non-user})P(\text{non-user})}{P(+ \mid \text{non-user})P(\text{non-user}) + P(+ \mid \text{user})P(\text{user})} \\&= \frac{0.01 \times 0.995}{0.01 \times 0.995 + 0.99 \times 0.005} \\&\approx 66.8\%\end{aligned}$$

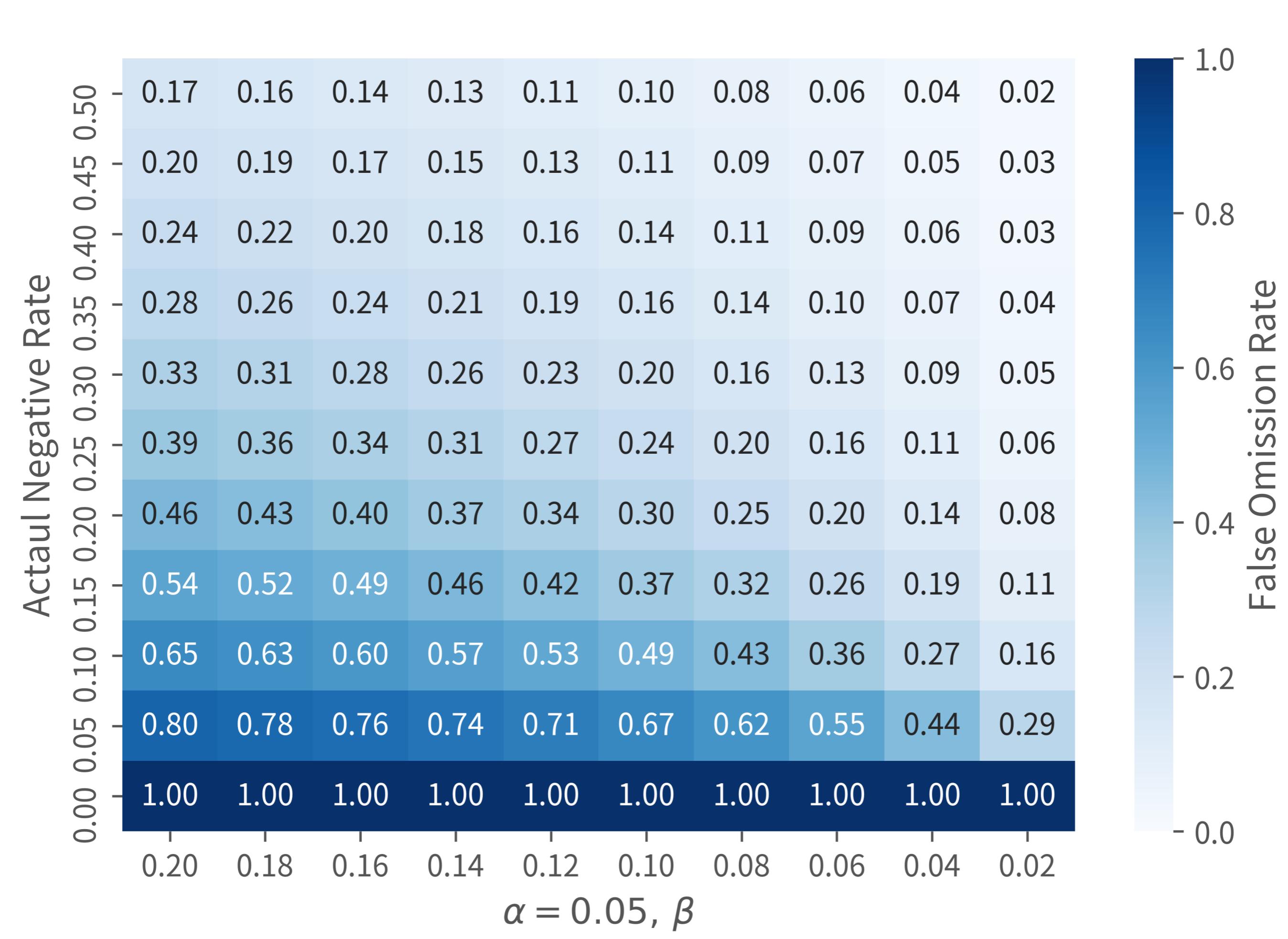
False discovery rate = B / BD



False omission rate = C / AC







Common “rates” in confusion matrix

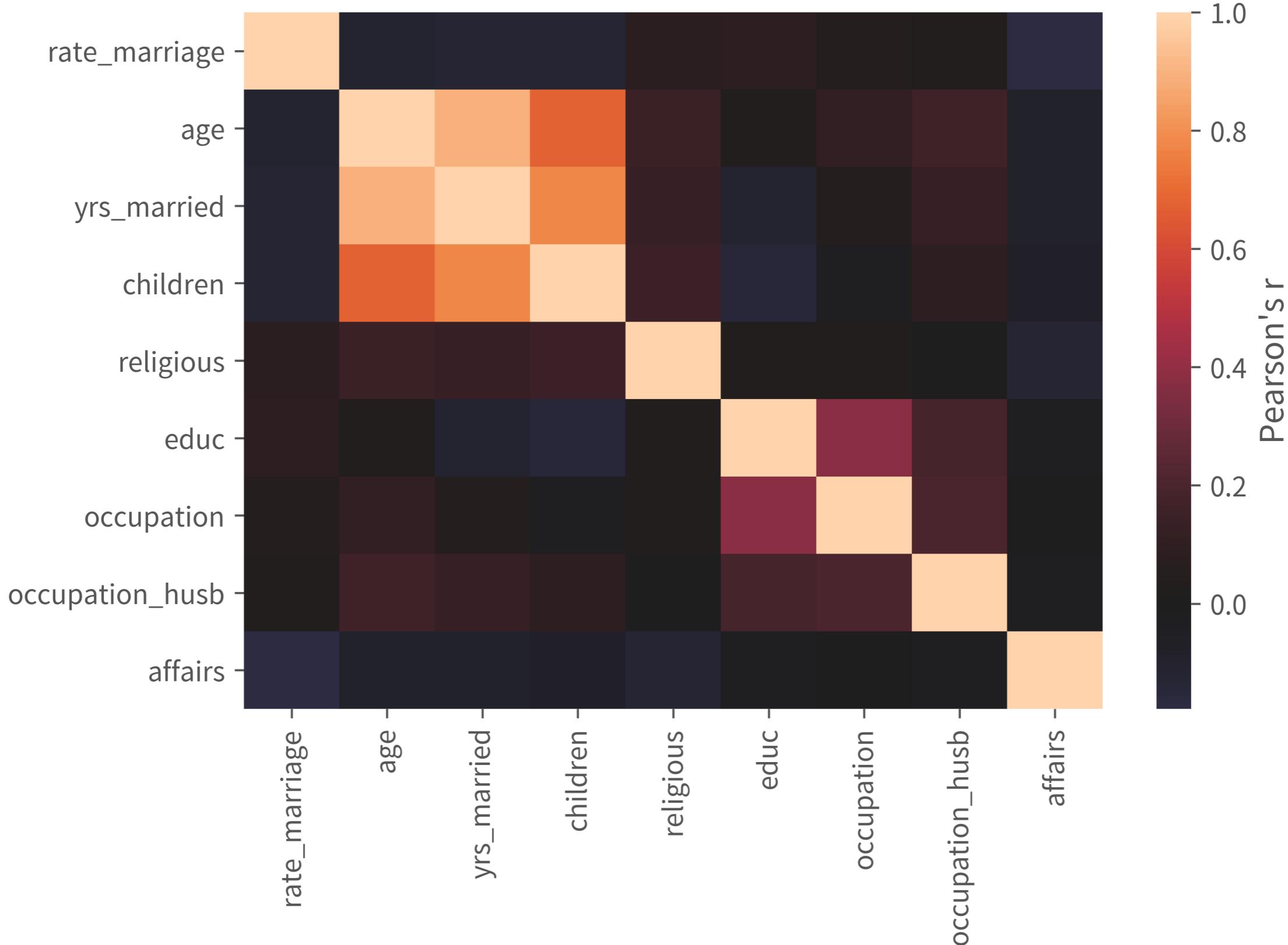
- false positive rate = B / AB = observed α
- false negative rate = C / CD = observed β
- false discovery rate = B / BD
- false omission rate = C / AC
- actual negative rate = $AB / ABCD$
- sensitivity = D / CD = observed power
- **specificity** = A / AB = observed confidence level
- precision = positive 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 required.
- The $\alpha, \beta, \text{effect size}$ → what *sample size* required.
- Still collect a sample as large as possible.
- Understand the sample.
 - Missing data, outliers, Q–Q plot, transform, etc.
- Test and report fully.
- *05_complete_a_test.ipynb*

Other statistical tools

Correlation analysis



Regression analysis

```
In [7]: fair_df = sm.datasets.fair.load_pandas().data  
ols_res = smf.ols('children ~ yrs_married', fair_df).fit()  
ols_res.summary()
```

Out[7]: OLS Regression Results

| Dep. Variable: | children | R-squared: | 0.597 | | | |
|-------------------|------------------|---------------------|-----------|-------|--------|--------|
| Model: | OLS | Adj. R-squared: | 0.597 | | | |
| Method: | Least Squares | F-statistic: | 9437. | | | |
| Date: | Fri, 06 Jul 2018 | Prob (F-statistic): | 0.00 | | | |
| Time: | 00:40:17 | 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
 - Seeing Theory
 - 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* $> 0.5 \uparrow \alpha \downarrow$
- *actual negative rate* $< 0.5 \downarrow \beta \downarrow$
- $\alpha, \beta, \text{effect size} \downarrow \text{sample size} \uparrow$
- Visualization and simulation do help.
- Bonus: *a1_figures.ipynb* .
- Let's identify noise efficiently!