Hypothesis Testing With Python

True Difference or Noise?

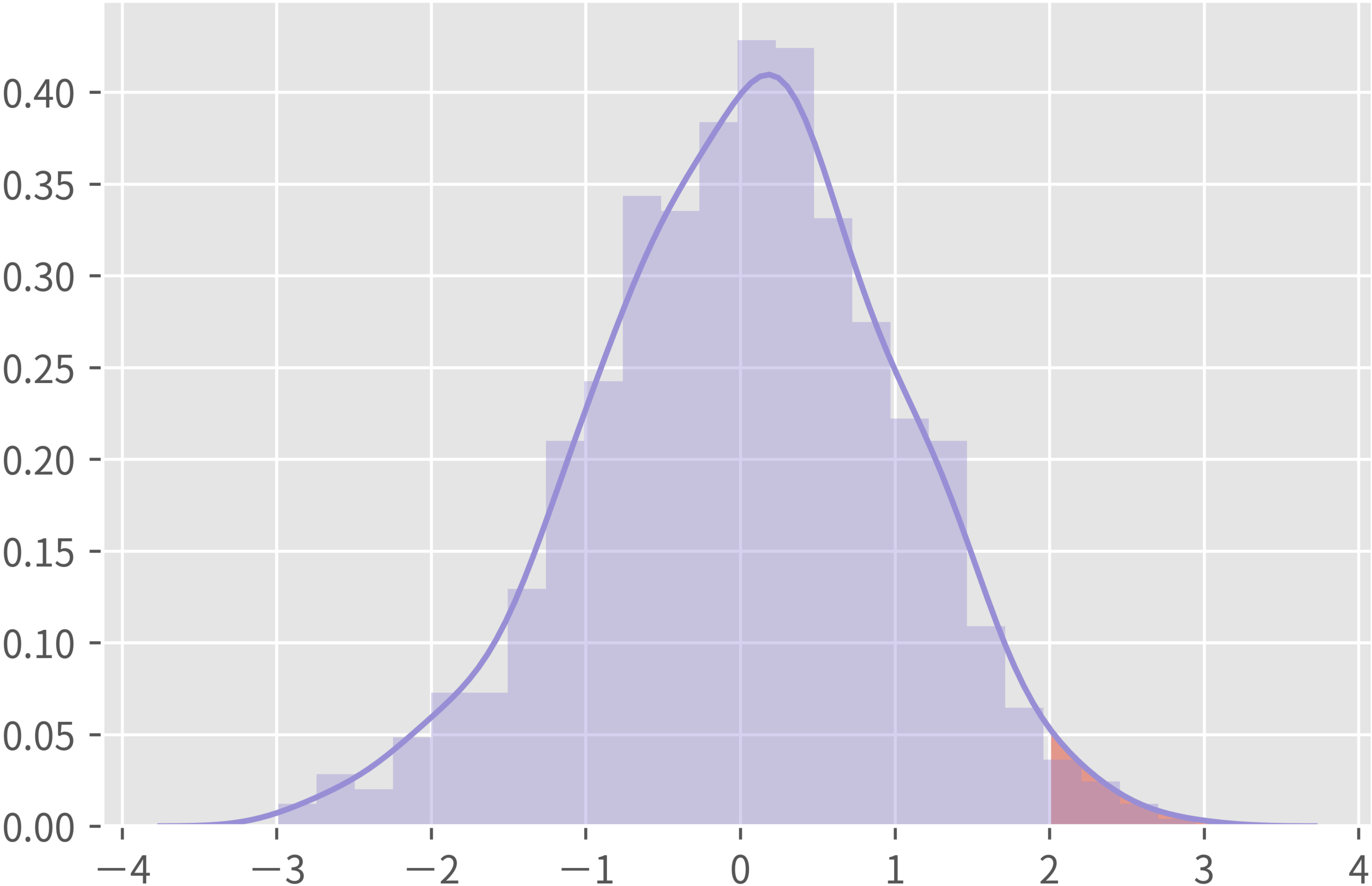
169.61

169.88

Which is better?

Noise?

That's a question.


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4

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-value* = 0.0027 (< 0.01)
  - 👏🏻👏🏻👏🏻
- *p-value* = 0.0271 (0.01–0.05)
  - 👏🏻❓👏🏻❓❓❓
- *p-value* = 0.2718 (≥ 0.05)
  - ❓❓❓❓❓❓
- *03_how_tests_work.ipynb*

Fair coin testing

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

Hypothesis testing

- "The means of two populations are equal."
- Case 1: *p-value* ≥ 0.05.
  - "Hmmm ... somehow equal."
- Case 2: *p-value* < 0.05.
  - "Not equal! So extreme!"

Hypothesis testing in a “null” taste

- *<null hypothesis>*
- Case 1: *p-value* ≥ *a*.
  - Can't reject *<null hypothesis>*.
- Case 2: *p-value* < *a*.
  - Reject *<null hypothesis>*.

Hypothesis testing in an "alternative" taste

- <*alternative hypothesis*> ≡ not <*null hypothesis*>
- Case 1: *p-value* $\geq a$.
  - Can't accept <*alternative hypothesis*>.
- Case 2: *p-value* $< a$.
  - Accept <*alternative hypothesis*>.

Hypothesis testing in “-+” taste

- “The case is negative.”
- Case 1: $p\text{-}value \geq a$.
  - “Hmmm … somehow negative.”
- Case 2: $p\text{-}value < a$.
  - “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
actual + CD	false + C	true + D

False positive rate = B / AB = observed α

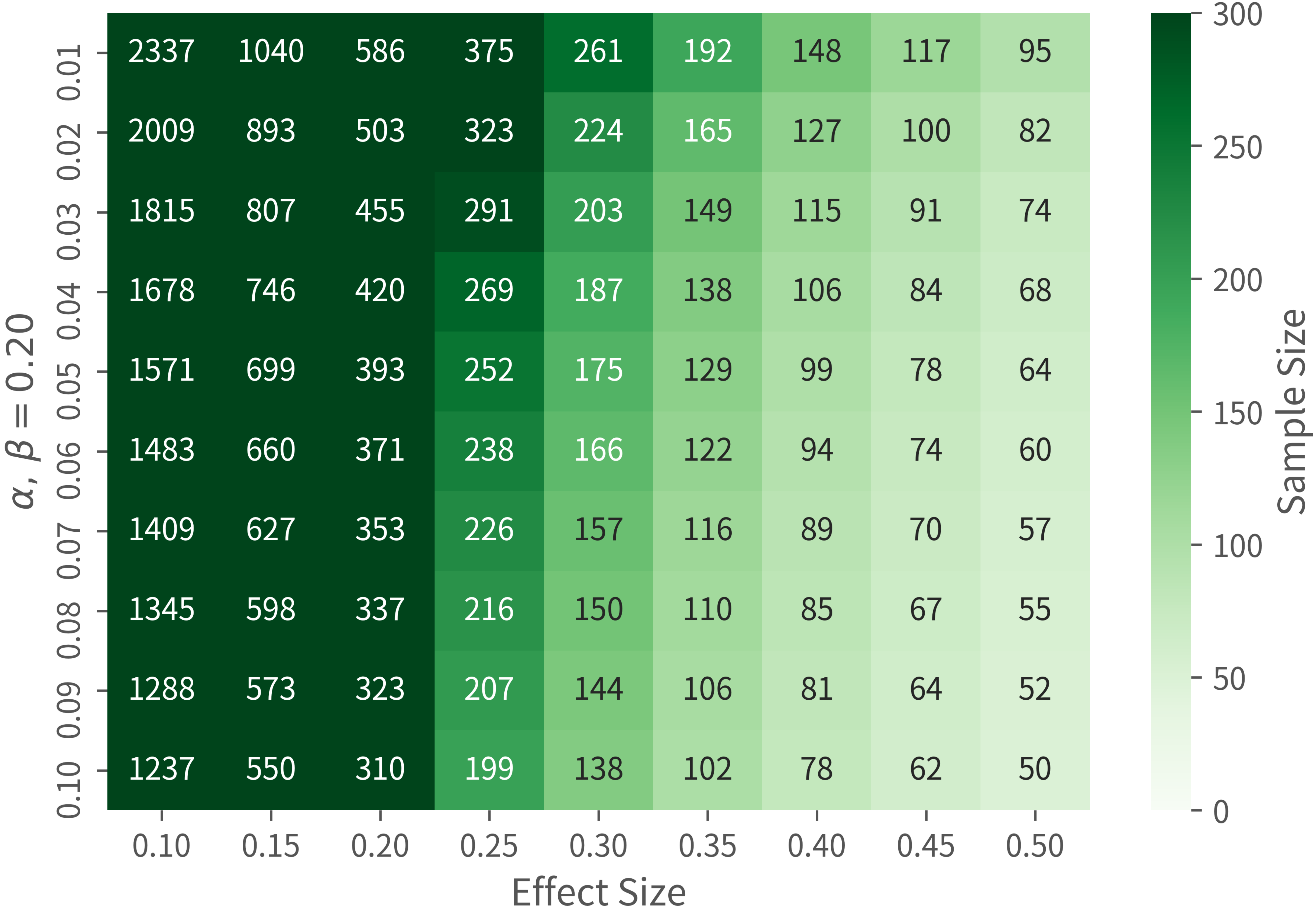
	predicted - AC	predicted + BD
actual - AB	true - A	false - B
actual + CD	false + C	true + D

False negative rate = C / CD = observed β

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

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

- Increase *sample size* to decrease *a*, *β*, *effect size*.
  - The *effect size* is the distance between groups.
    - $= \dfrac{\mu_1 - \mu_2}{\sigma}$
  - http://www.drcoplan.com/dsm5-the-case-for-double-standards
    - The figures explain *a*, *β* perfectly,
      but due to the copyright, only put the link here.
- When *a*, *β*, *effect size* are defined, get the *sample size*.
- When *a*, *β*, *sample size* are defined, get the *effect size*.


α, β = 0.20
0.01 2337 1040 586 375 261 192 148 117 95
0.02 2009 893 503 323 224 165 127 100 82
0.03 1815 807 455 291 203 149 115 91 74
0.04 1678 746 420 269 187 138 106 84 68
0.05 1571 699 393 252 175 129 99 78 64
0.06 1483 660 371 238 166 122 94 74 60
0.07 1409 627 353 226 157 116 89 70 57
0.08 1345 598 337 216 150 110 85 67 55
0.09 1288 573 323 207 144 106 81 64 52
0.10 1237 550 310 199 138 102 78 62 50
0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
Effect Size
Sample Size
300
250
200
150
100
50
0

False discovery rate

- Suppose:
  - A drug test has observed $a = 1\%$ 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?

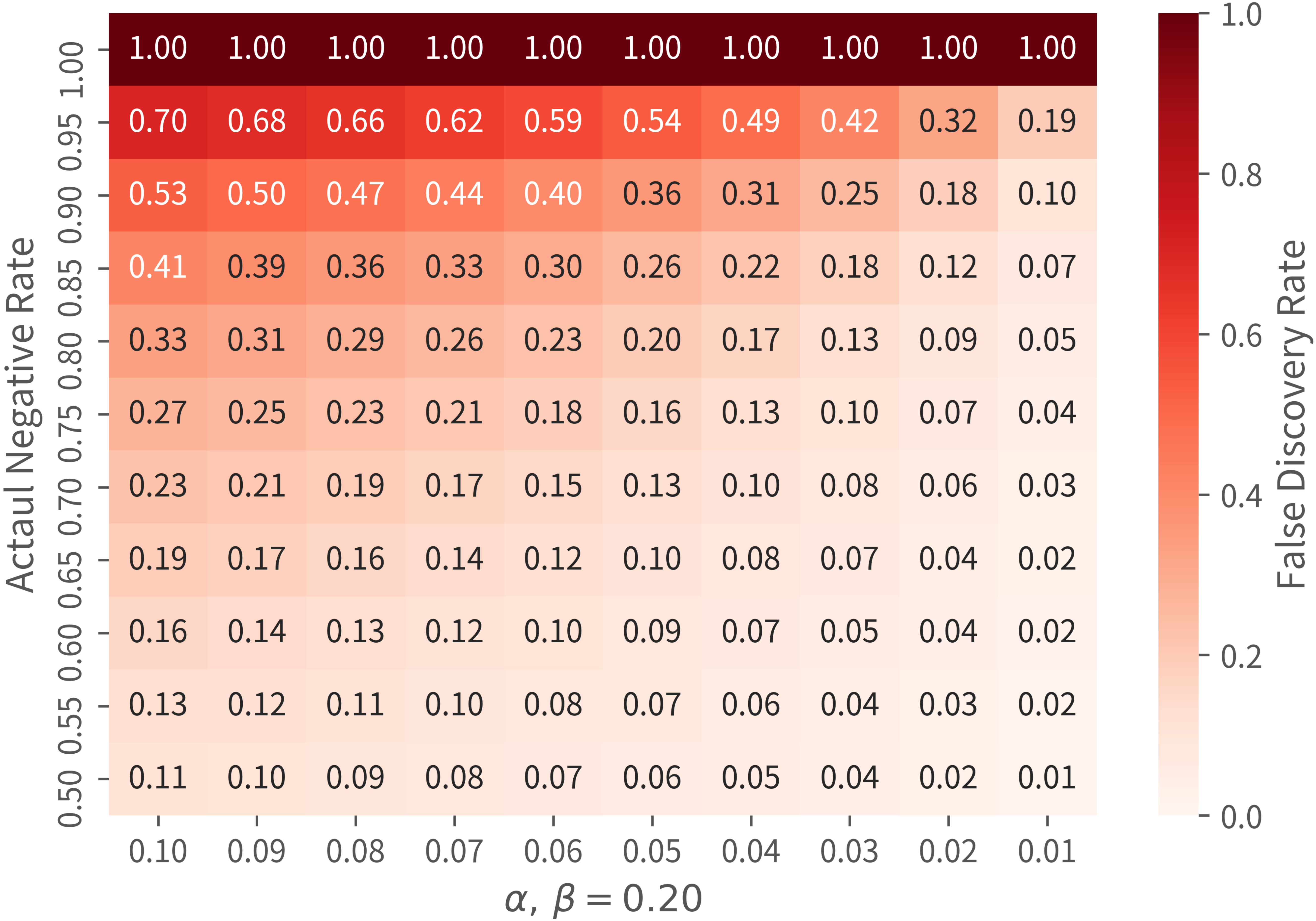
$$
\begin{aligned}
P(\textit{non-user} \mid +) &= \frac{P(+ \mid \textit{non-user})P(\textit{non-user})}{P(+)} \\
&= \frac{P(+ \mid \textit{non-user})P(\textit{non-user})}{P(+ \mid \textit{non-user})P(\textit{non-user}) + P(+ \mid \textit{user})P(\textit{user})} \\
&= \frac{0.01 \times 0.995}{0.01 \times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

False discovery rate = B / BD

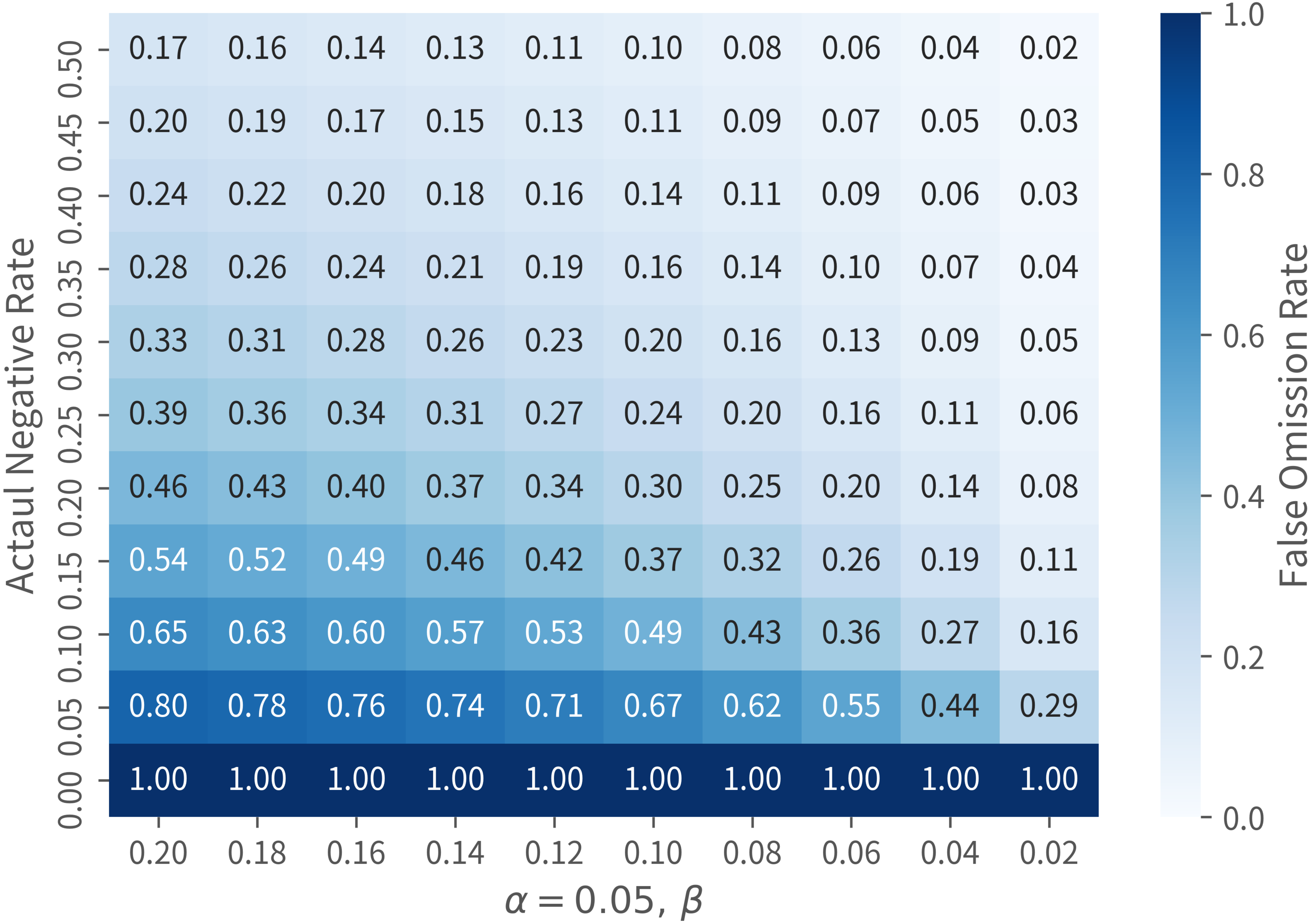
	predicted - AC	predicted + BD
actual - AB	true - A	false - B
actual + CD	false + C	true + D

False omission rate = C / AC

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



Actaul Negative Rate
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00
0.70 0.68 0.66 0.62 0.59 0.54 0.49 0.42 0.32 0.19
0.53 0.50 0.47 0.44 0.40 0.36 0.31 0.25 0.18 0.10
0.41 0.39 0.36 0.33 0.30 0.26 0.22 0.18 0.12 0.07
0.33 0.31 0.29 0.26 0.23 0.20 0.17 0.13 0.09 0.05
0.27 0.25 0.23 0.21 0.18 0.16 0.13 0.10 0.07 0.04
0.23 0.21 0.19 0.17 0.15 0.13 0.10 0.08 0.06 0.03
0.19 0.17 0.16 0.14 0.12 0.10 0.08 0.07 0.04 0.02
0.16 0.14 0.13 0.12 0.10 0.09 0.07 0.05 0.04 0.02
0.13 0.12 0.11 0.10 0.08 0.07 0.06 0.04 0.03 0.02
0.11 0.10 0.09 0.08 0.07 0.06 0.05 0.04 0.02 0.01
0.10 0.09 0.08 0.07 0.06 0.05 0.04 0.03 0.02 0.01
α, $\beta = 0.20$
False Discovery Rate
1.0
0.8
0.6
0.4
0.2
0.0


Actaul Negative Rate
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.17 0.16 0.14 0.13 0.11 0.10 0.08 0.06 0.04 0.02
0.20 0.19 0.17 0.15 0.13 0.11 0.09 0.07 0.05 0.03
0.24 0.22 0.20 0.18 0.16 0.14 0.11 0.09 0.06 0.03
0.28 0.26 0.24 0.21 0.19 0.16 0.14 0.10 0.07 0.04
0.33 0.31 0.28 0.26 0.23 0.20 0.16 0.13 0.09 0.05
0.39 0.36 0.34 0.31 0.27 0.24 0.20 0.16 0.11 0.06
0.46 0.43 0.40 0.37 0.34 0.30 0.25 0.20 0.14 0.08
0.54 0.52 0.49 0.46 0.42 0.37 0.32 0.26 0.19 0.11
0.65 0.63 0.60 0.57 0.53 0.49 0.43 0.36 0.27 0.16
0.80 0.78 0.76 0.74 0.71 0.67 0.62 0.55 0.44 0.29
1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00
0.20 0.18 0.16 0.14 0.12 0.10 0.08 0.06 0.04 0.02
$\alpha = 0.05,\ \beta$
1.0
0.8
0.6
0.4
0.2
0.0
False Omission Rate

Common “rates” in confusion matrix

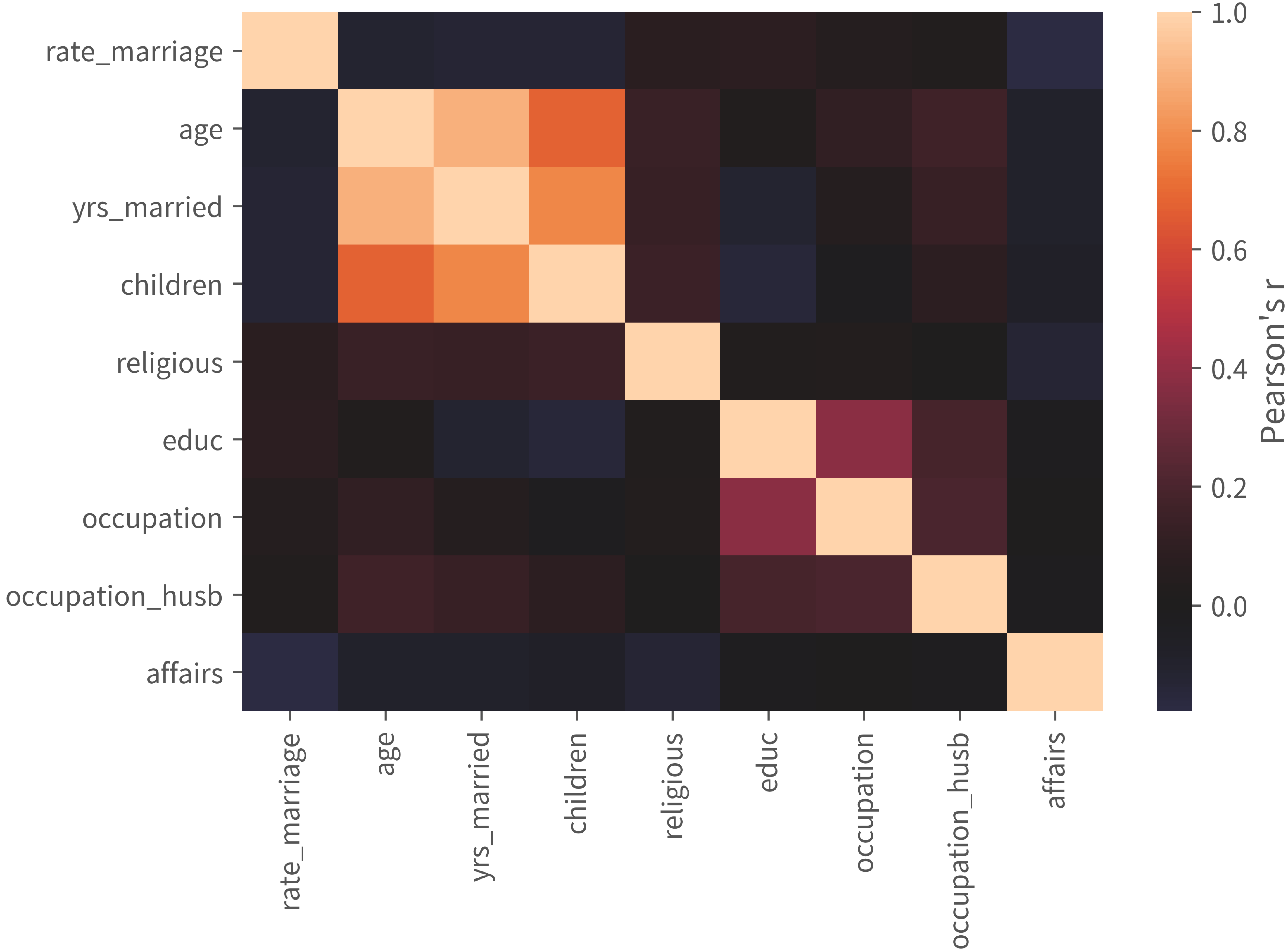
- false positive rate = B / AB = observed α
- false negative rate= C / CD = observed β
- false discovery rate = B / BD
- false omission rate = C / AC
- actual negative rate = AB / N
- sensitivity = D / CD = observed power
- specificity = A / AB = observed confidence level
- precision = positive predictive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

Most formal steps

- State the hypothesis → what *test*.
- Estimate the *actual negative rate*.
- The a*ctual negative rate* → what α, β required.
- The α, β, *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 ↑ *a* ↓
- *actual negative rate* < 0.5 ↓ *β* ↓
- *a, β, effect size* ↓ *sample size* ↑
- Visualization and simulation do help.
- Bonus: *a1_figures.ipynb* .
- Let's identify noise efficiently!