



# Hypothesis Testing With Python

---

*True Difference or Noise?*

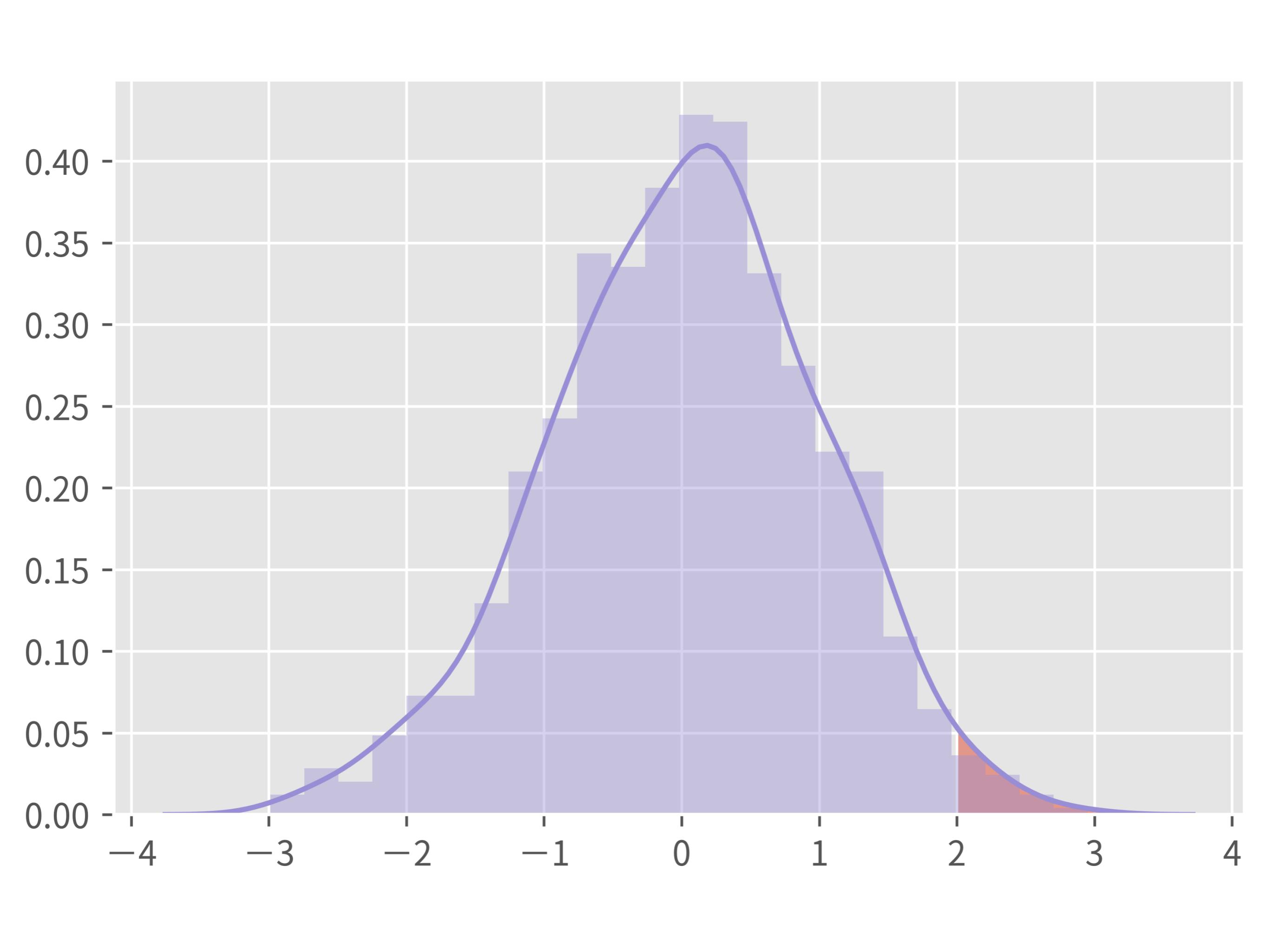
**169.61**

**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

# 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
- Visit the Pipfile for the exact version numbers.

# 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 & $\alpha$

.....

P-value &  $\alpha$

Wording

p-value < 0.01

Very significant

p-value < 0.05

Significant

p-value  $\geq 0.05$

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

# Hypothesis testing in an “alternative” taste

---

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

# 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 $\alpha$

---

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

# False negative rate = C / CD = observed $\beta$

---

		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 Left (Actual -): true - A
- Top Right (Actual +): false + B
- Bottom Left (Actual +): false - C
- Bottom Right (Actual -): true + D

The bottom-right quadrant, which contains element D, is highlighted with a thick dark blue border.

## When sample size ↑; $\alpha$ , $\beta$ , effect size ↓

---

- Increase *sample size* to decrease  $\alpha$ ,  $\beta$ , *effect size*.
  - The *effect size* is the distance between groups.
$$\text{➤ } = \frac{\mu_1 - \mu_2}{\sigma}$$
  - DSM5: The case for double standards – James Coplan, M.D.
    - The figures explain  $\alpha$ ,  $\beta$  perfectly,  
but due to the copyright, only put the link here.
  - When  $\alpha$ ,  $\beta$ , *effect size* are defined, get the *sample size*.
  - When  $\alpha$ ,  $\beta$ , *sample size* are defined, get the *effect size*.





- If the  $n$  is large enough, a reasonable approximation to a binomial distribution is given by the normal distribution:
  - $X \sim B(n, p)$
  - When  $n \rightarrow \infty$ :
    - $\mu = np$
    - $\sigma = \sqrt{np(1-p)}$
    - $X \sim N(\mu, \sigma)$
- See also: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3775042/>.

# False discovery rate

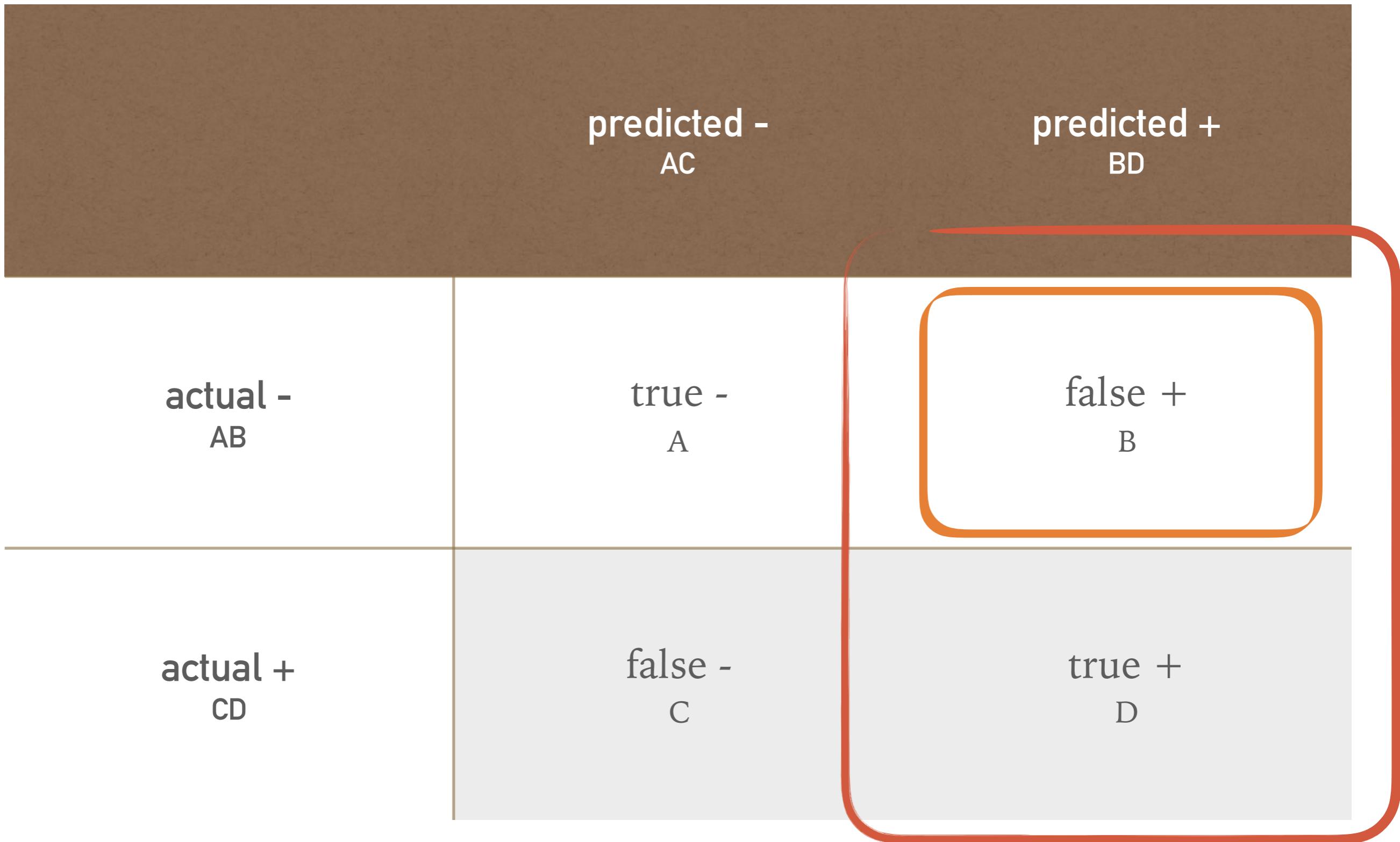
---

- 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 not a drug user?

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

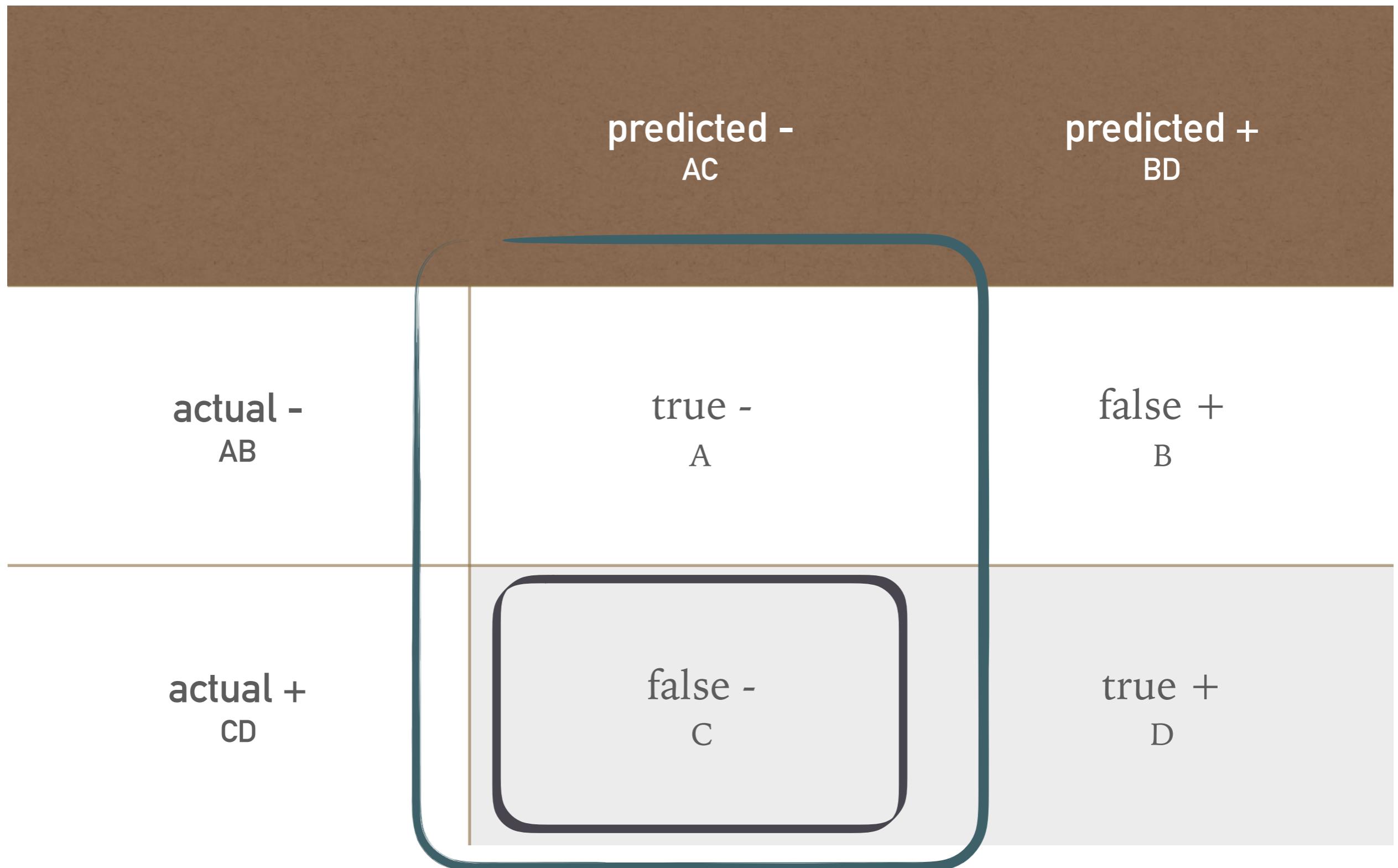
# False discovery rate = B / BD

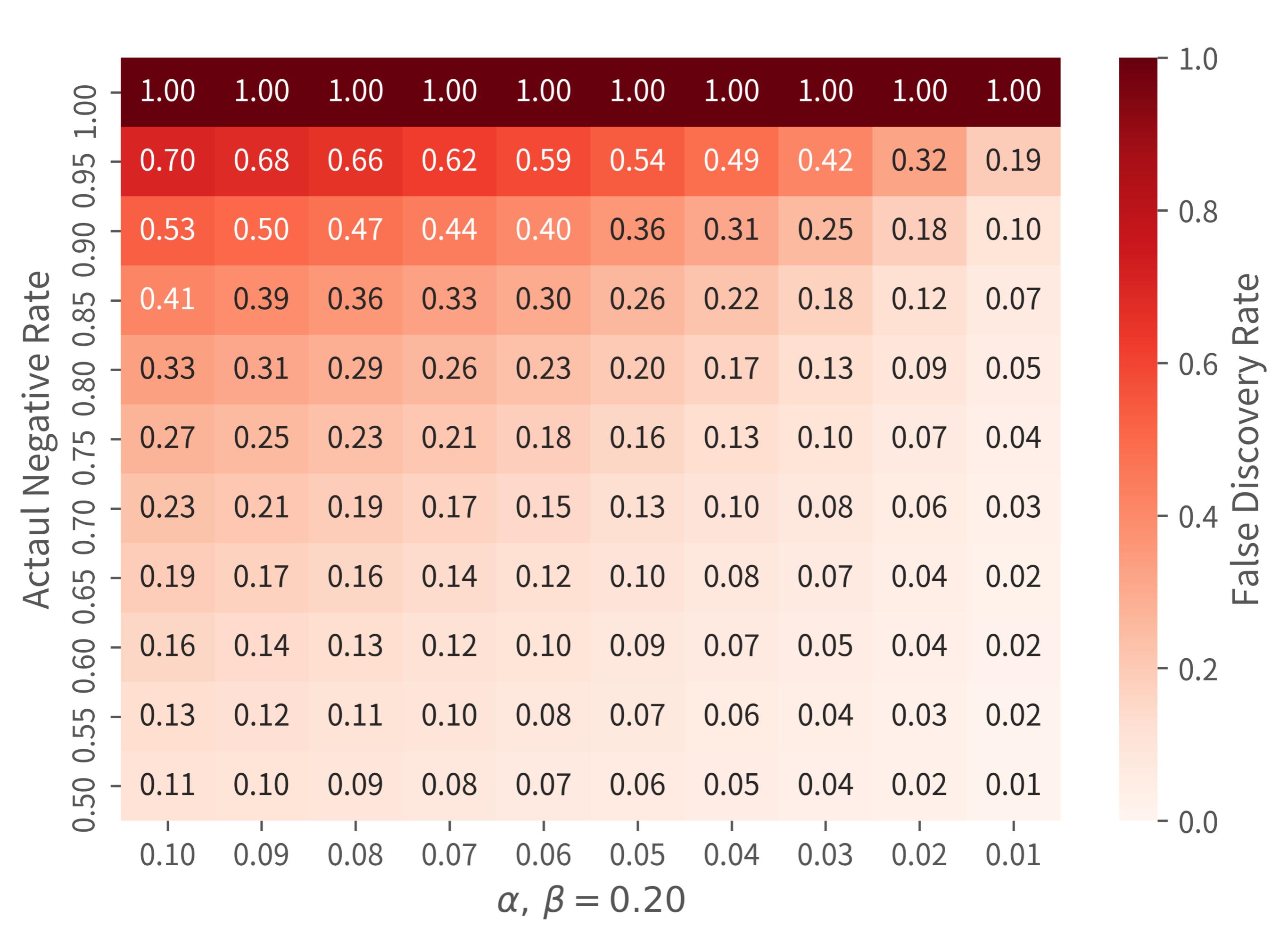
---

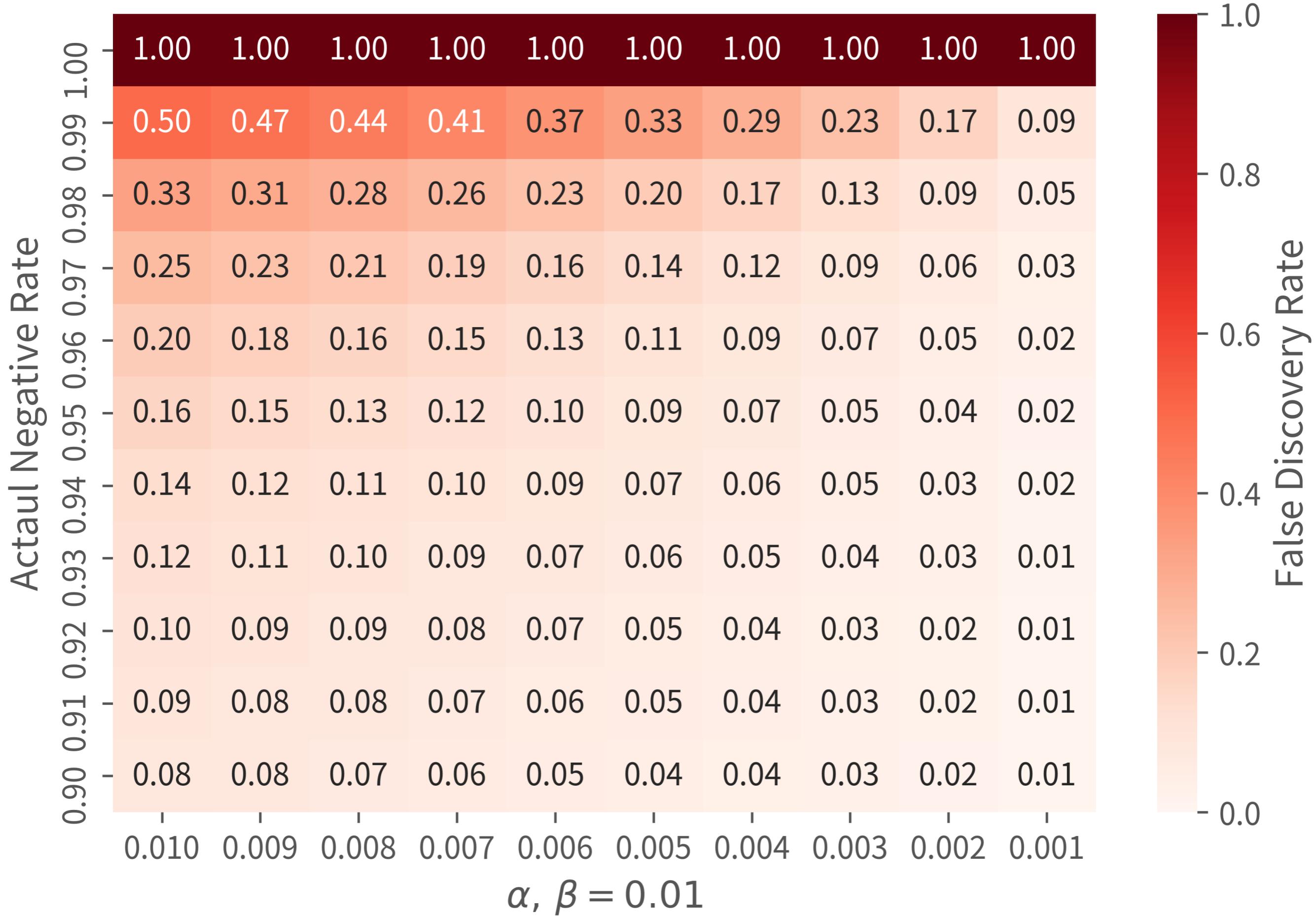


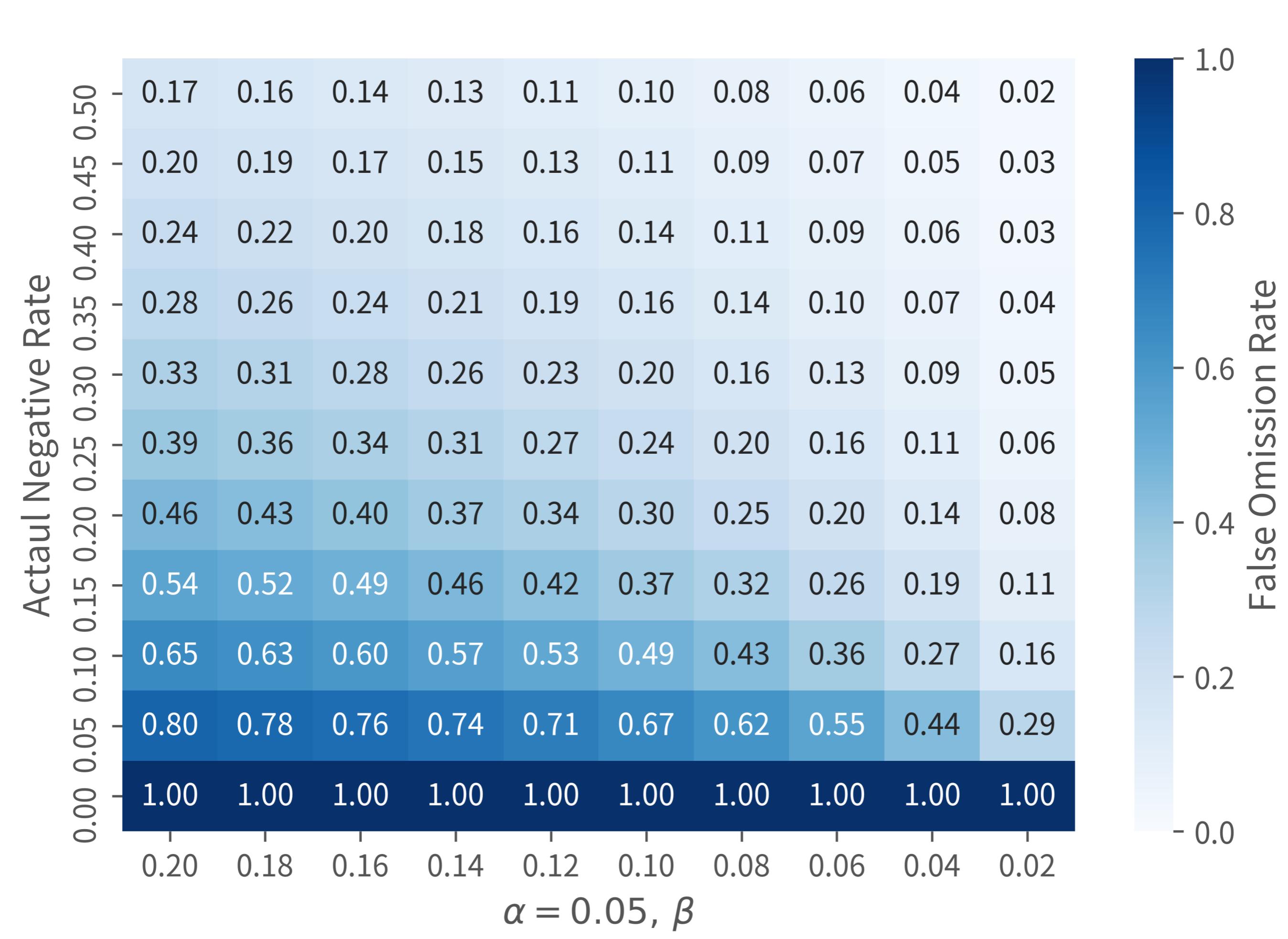
# False omission rate = C / AC

---









# Common “rates” in confusion matrix

---

- false positive rate =  $B / AB$  = observed  $\alpha$
- false negative rate =  $C / CD$  = observed  $\beta$
- actual negative rate =  $AB / ABCD$
- false discovery rate =  $B / BD$  = observed inverse  $\alpha$
- false omission rate =  $C / AC$  = observed inverse  $\beta$
- sensitivity =  $D / CD$  = observed power = recall
- specificity =  $A / AB$  = observed confidence level
- precision =  $D / BD$  = observed inverse power
- recall =  $D / CD$  = observed power = sensitivity

# Most formal steps

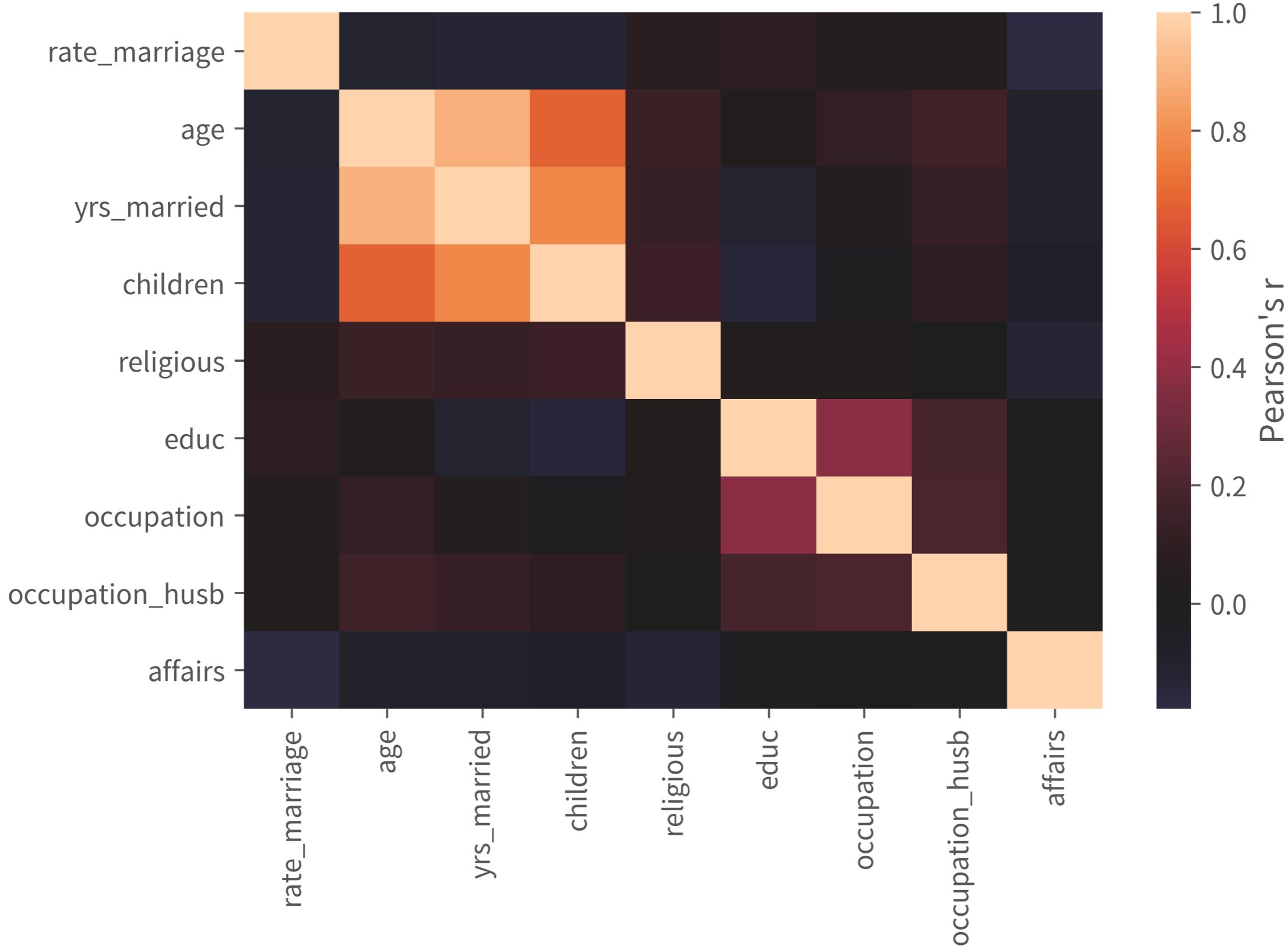
---

- State the hypothesis → what *test*.
- Estimate the *actual negative rate*.
- The *actual negative rate* → what  $\alpha, \beta$  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](#)
  - [Statistics – SciPy Tutorial](#)
  - [StatsModels](#)
  - [Biological Statistics](#)
  - [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!