



# 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
  - What methods work not well?
  - What methods work well?
  - Simple mean, etc.
- 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 versions.

# Tests with datasets

# Go with the notebooks

---

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

# False positive rate, 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.

# 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:  $p\text{-value} \geq a$ 
  - 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!”

# 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

- *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 independence:
  - 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 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**.
  - 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

# The tests in Python

---

- *04\_common\_tests.ipynb*

# Complete a test

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

---

- power?
- sample size?
- $\beta$ ?
- effect size?
- false discovery rate?
- ? ? ? ? ?

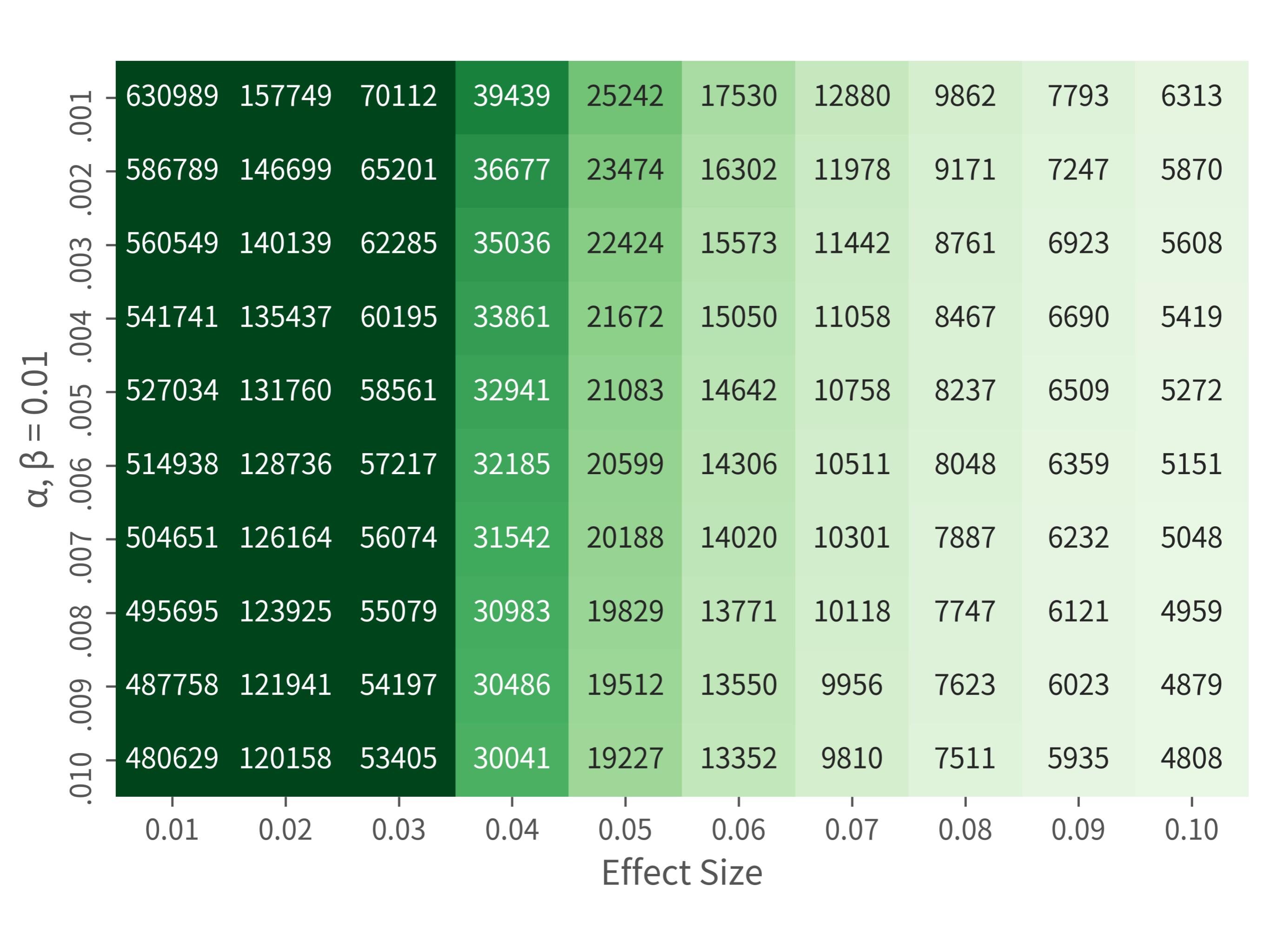
# $\alpha$ , $\beta$ , and effect size

---

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

- When  $\alpha$ ,  $\beta$ , *effect size* are defined, get the *sample size*.
- When  $\alpha$ ,  $\beta$ , *sample size* are defined, get the *effect size*.
- Increase *sample size* to decrease  $\alpha$ ,  $\beta$ , or *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
	false -	C	true + D
actual + CD			

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

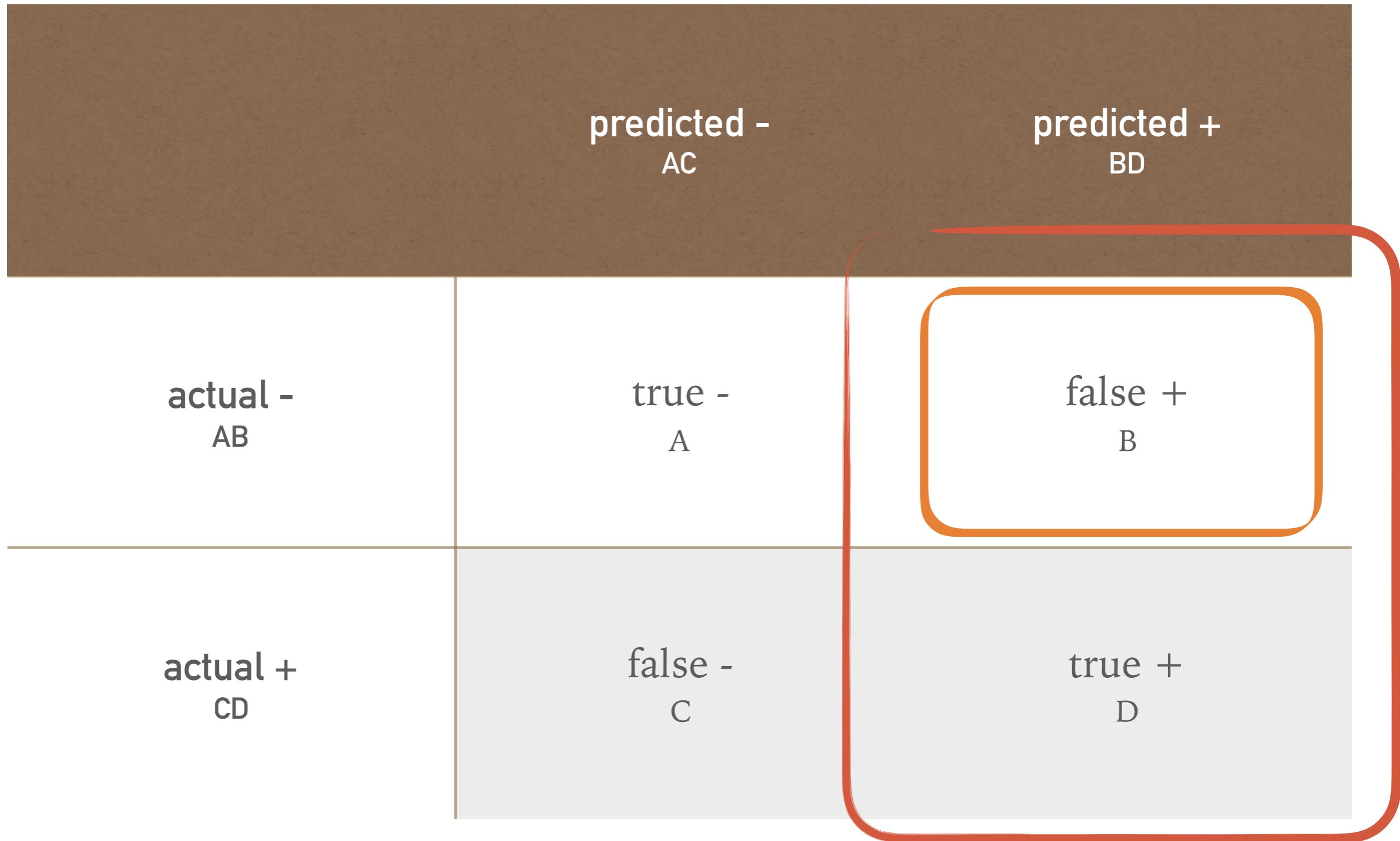
---

		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, showing the relationship between predicted and actual outcomes. The columns represent predicted outcomes: 'predicted -' (top) and 'predicted +' (bottom). The rows represent actual outcomes: 'actual -' (left) and 'actual +' (right). The four cells are labeled A, B, C, and D. Cell A is 'true -' (predicted - and actual -). Cell B is 'false +' (predicted + and actual -). Cell C is 'false -' (predicted - and actual +). Cell D is 'true +' (predicted + and actual +). A large rounded rectangle highlights cell C, and a smaller rounded rectangle highlights the entire row for 'actual +'.

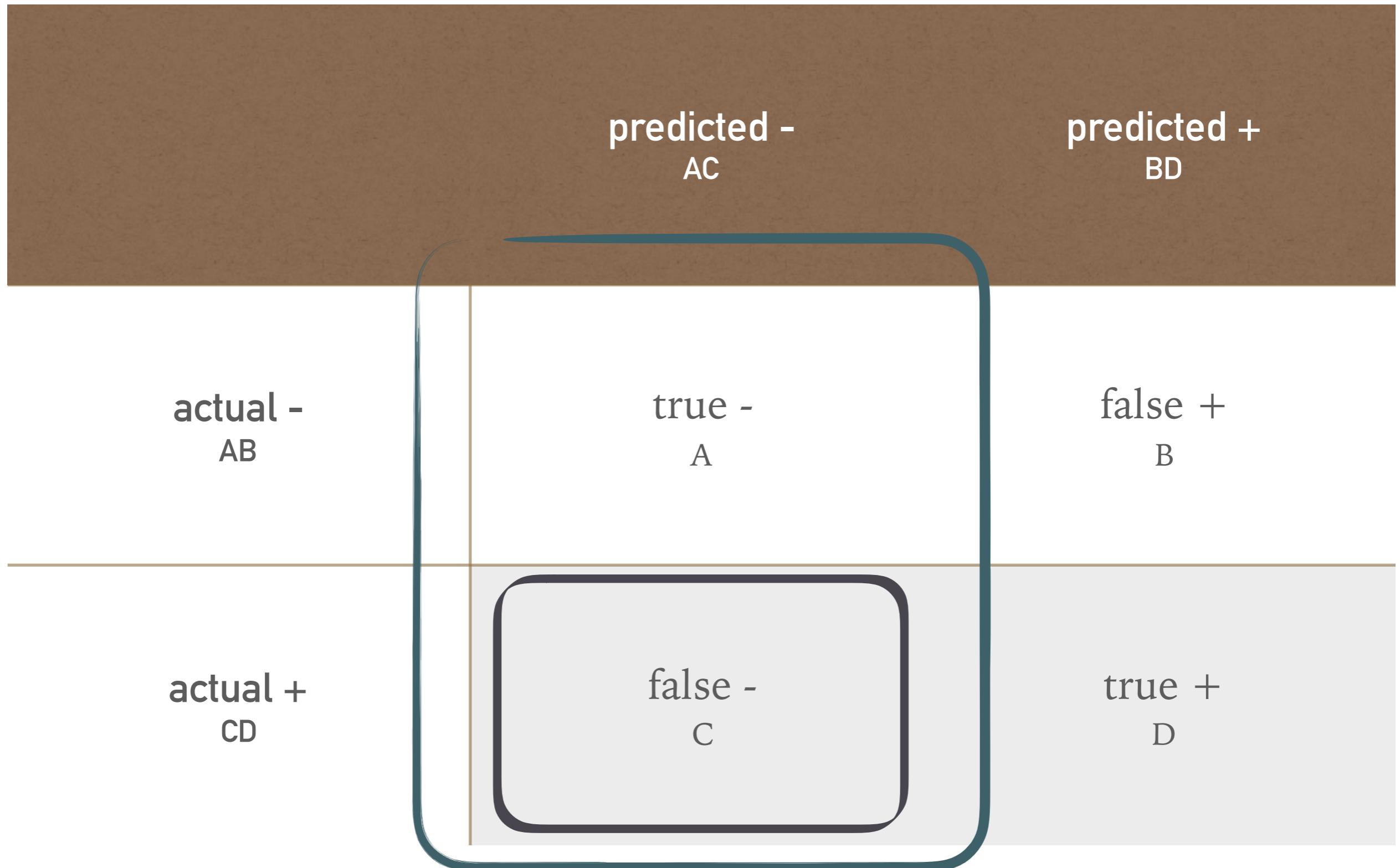
# False discovery rate = B / BD = observed inverse a

---



# False omission rate = C / AC = observed inverse $\beta$

.....

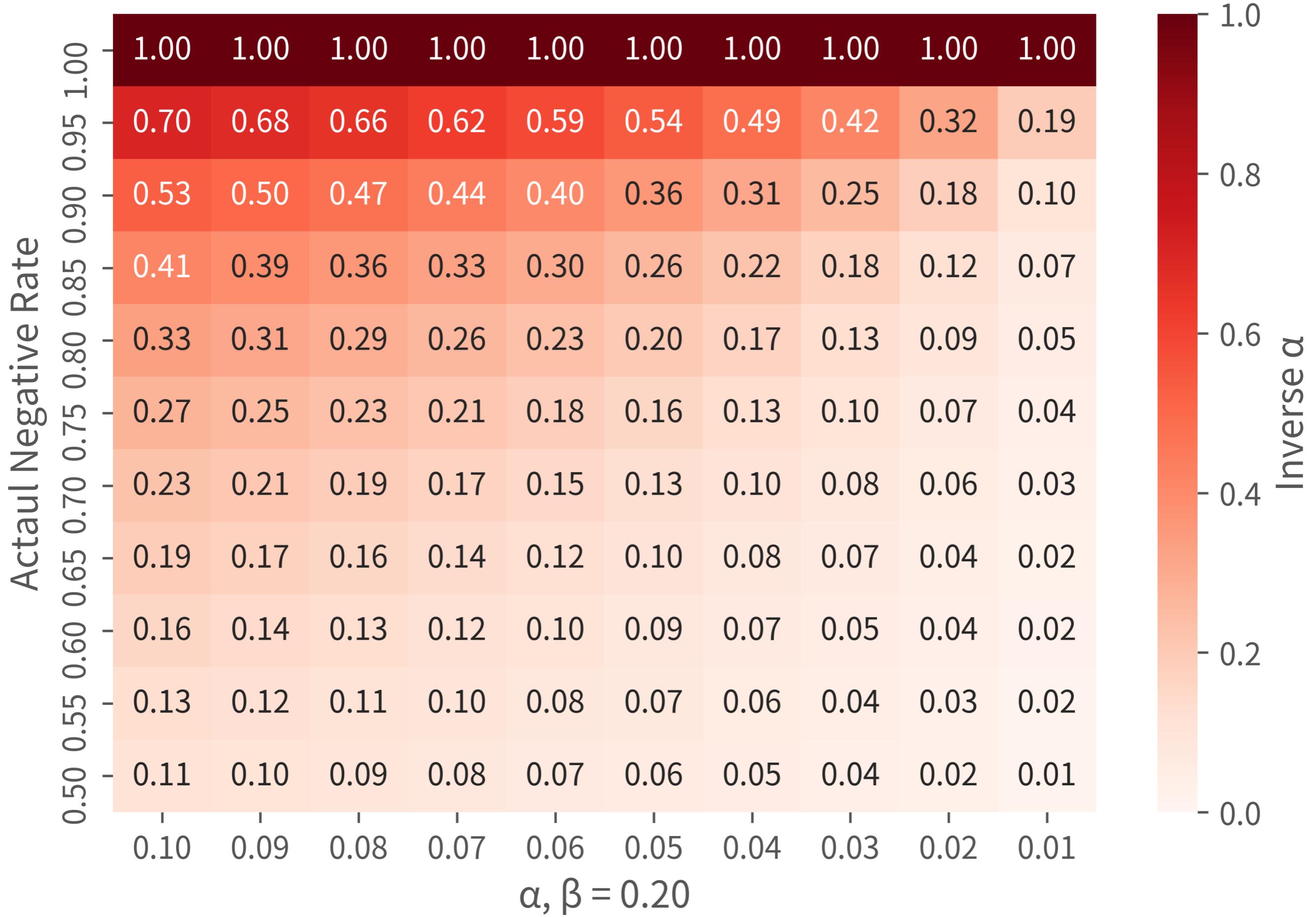


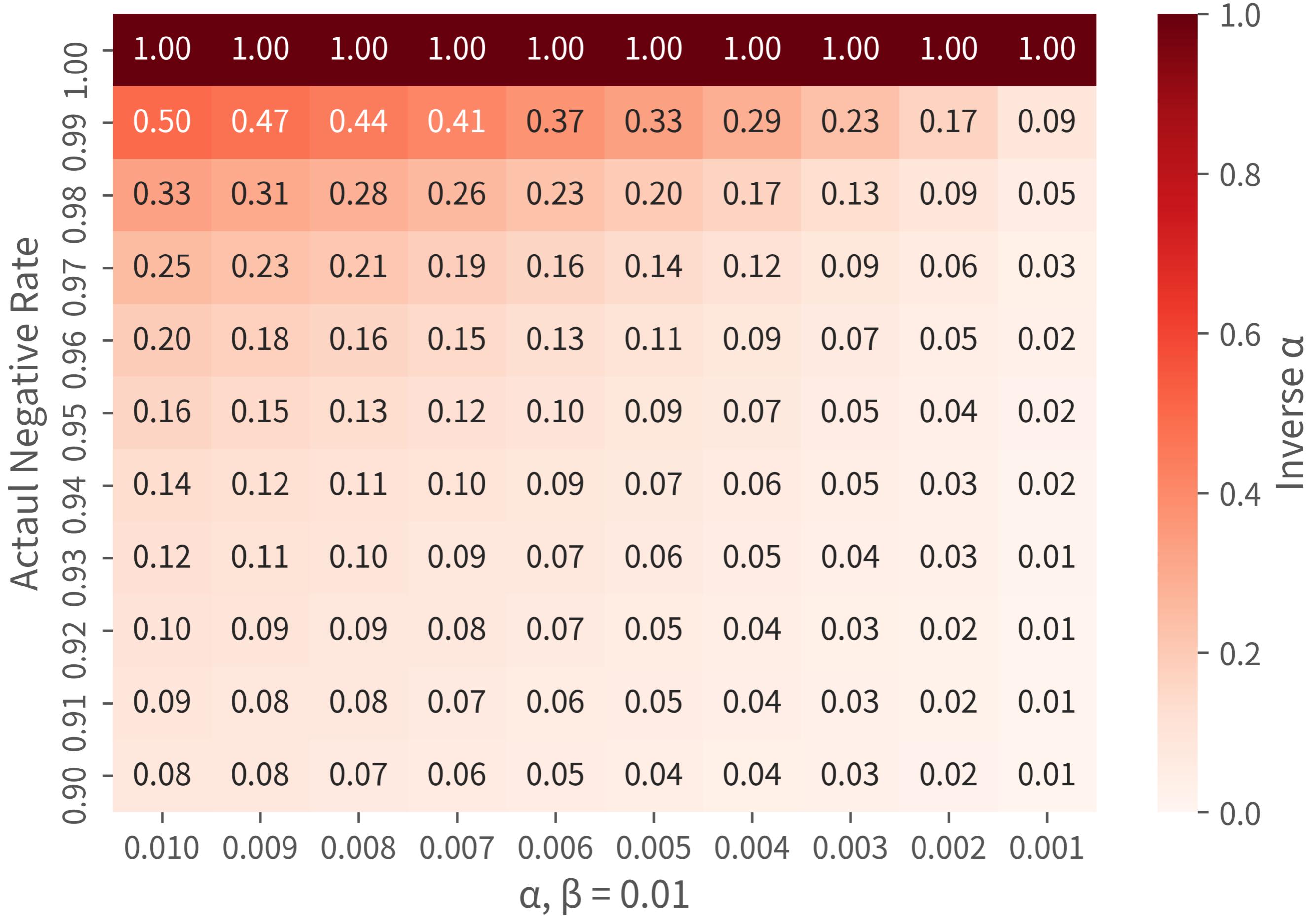
## Inverse a

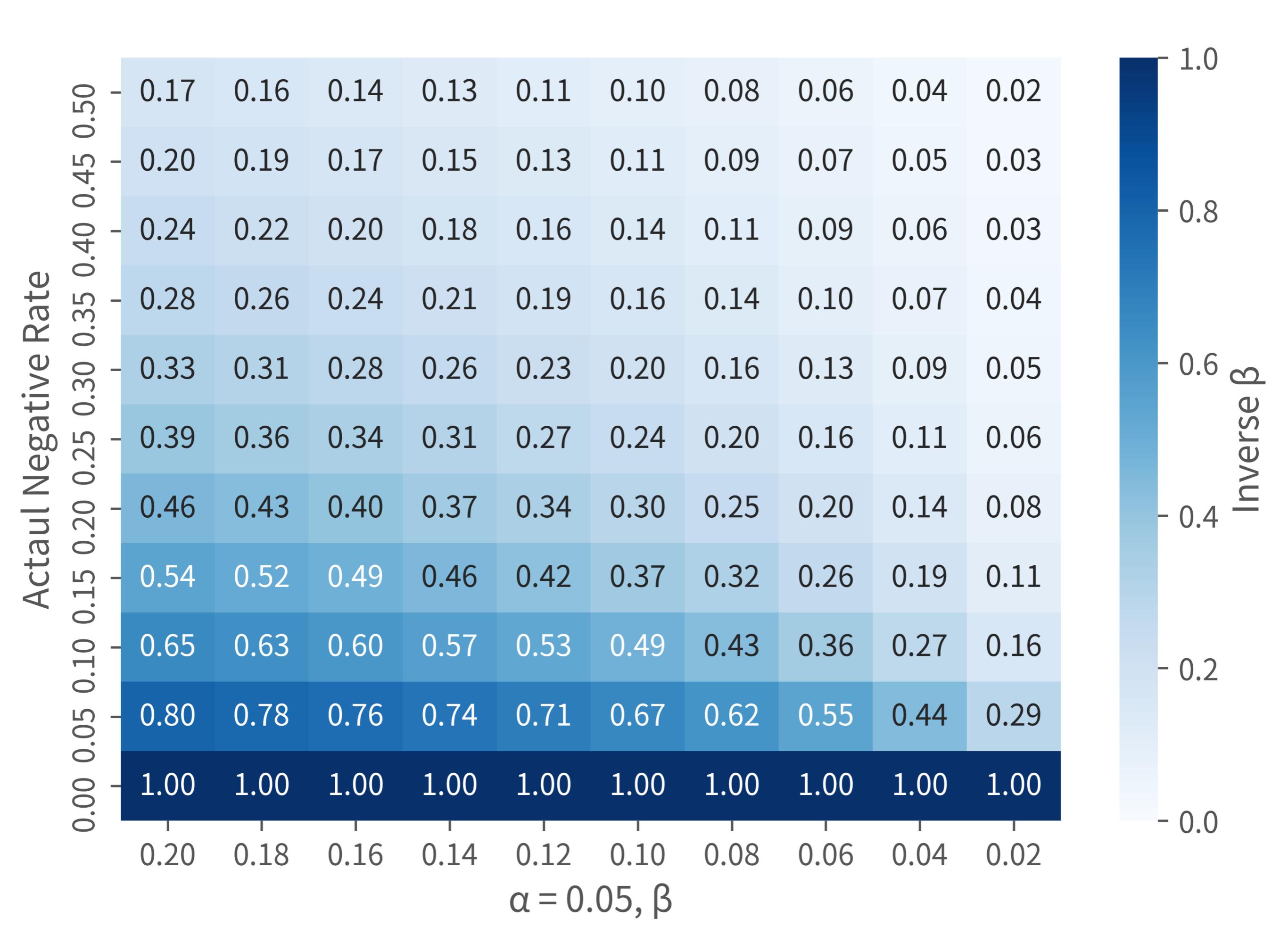
---

- 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?

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







# Common “rates” in hypothesis testing

---

- $\alpha$  = predefined acceptable false positive rate
  - $\alpha$  = significance level = type I error rate
- $\beta$  = predefined acceptable false negative rate
  - $\beta$  = type II error rate
- inverse  $\alpha$  = predefined acceptable false discovery rate
- inverse  $\beta$  = predefined acceptable false omission rate
- power =  $1 - \beta$  = predefined acceptable sensitivity or recall
- confidence level =  $1 - \alpha$  = predefined acceptable specificity

# 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

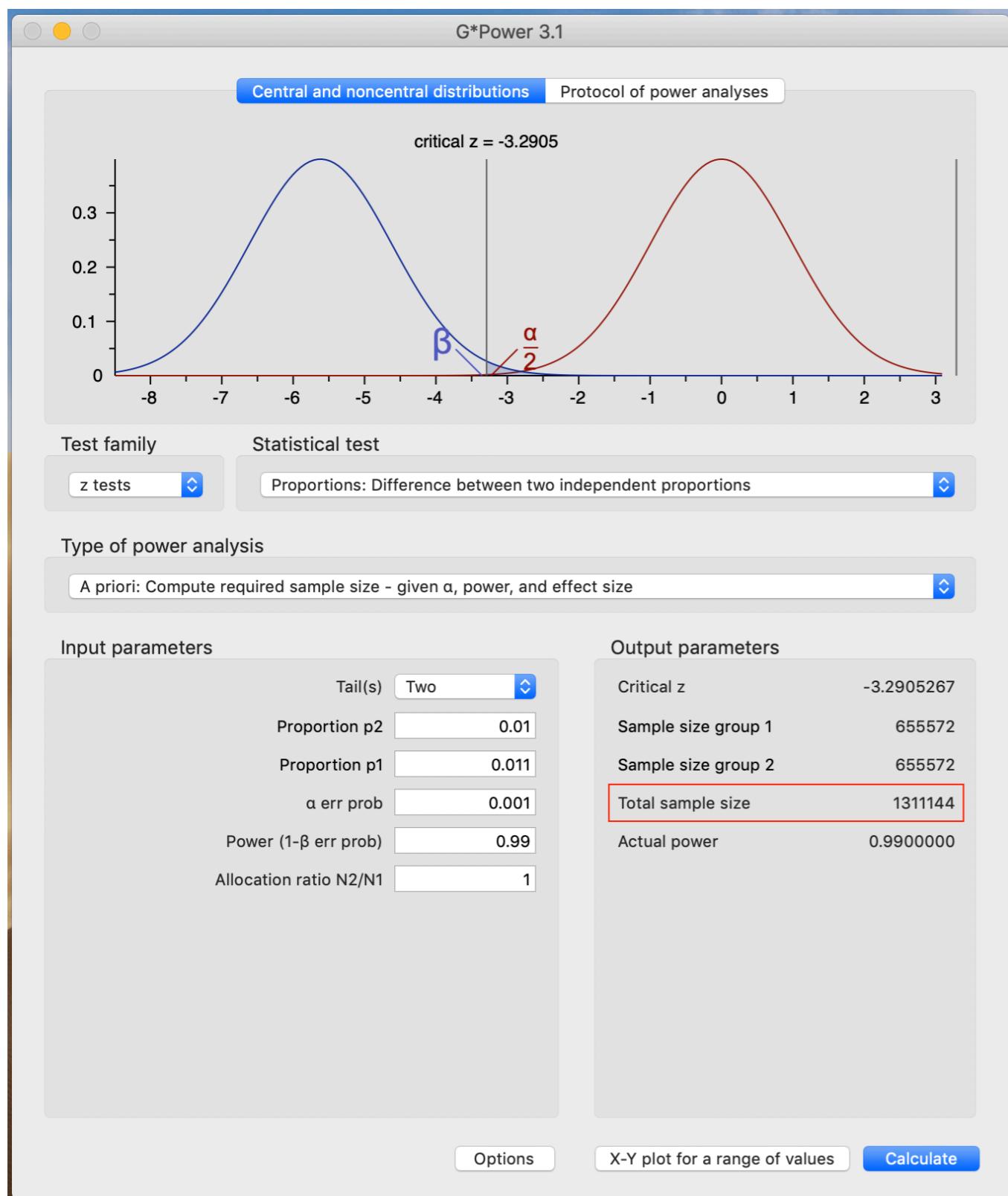
# Most formal steps

---

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

# Calculate by yourself

.....



➤ G\*Power can calculate:

➤ *sample size*

➤  $\alpha$

➤  $power = 1 - \beta$

➤ *effect size*

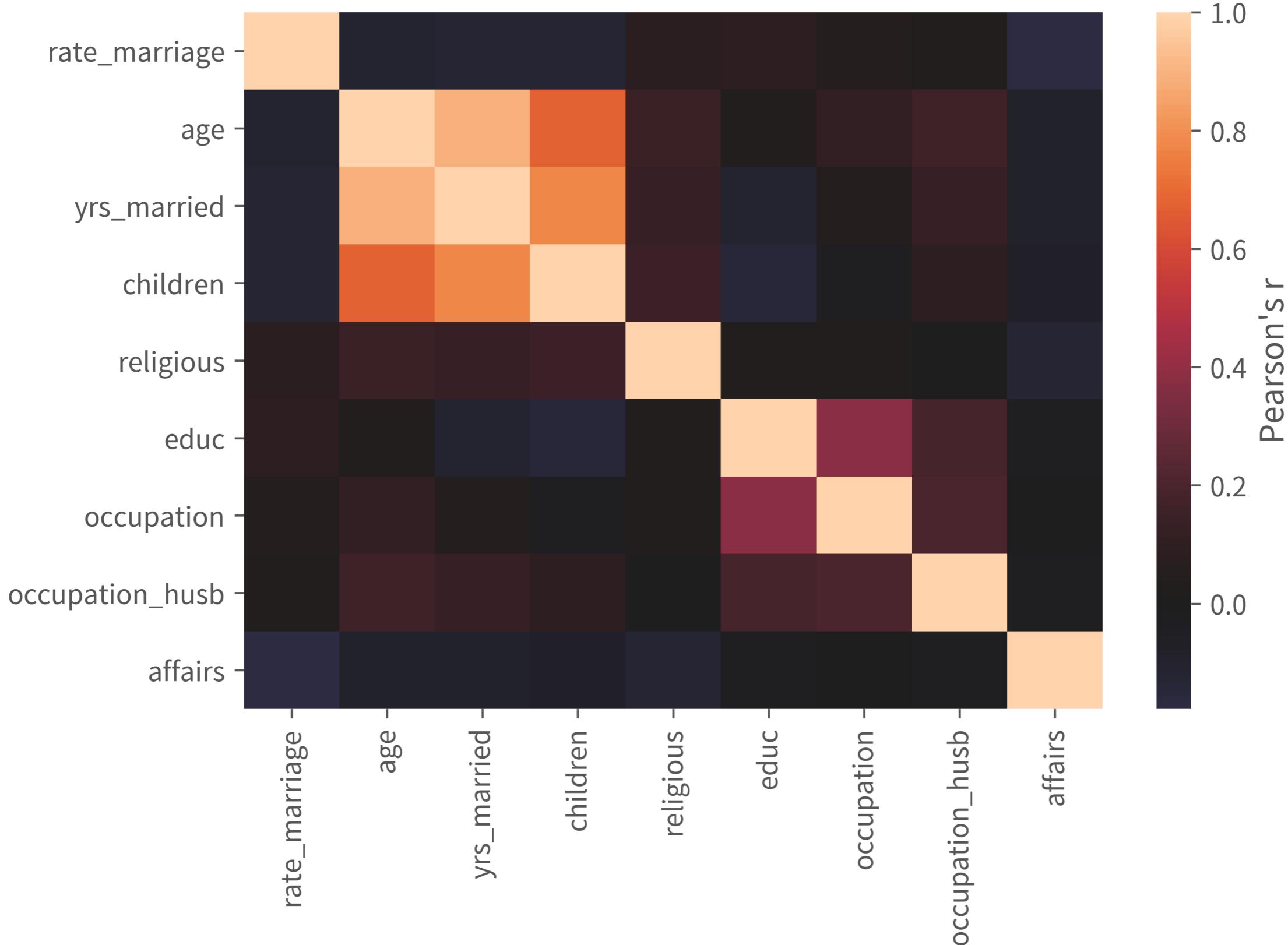
➤ <http://www.gpower.hhu.de/>

➤ *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*: “tail” probability given “actual negative” hypothesis.
- *actual negative rate* does matter.
- $\alpha, \beta, \text{effect size} \downarrow; \text{sample size} \uparrow.$
- Visualization and simulation do help.
- Bonus: `a1_figures.ipynb` .
- Let's identify noise efficiently!