



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

1. Tests with datasets
2. How tests work
3. Common tests
4. 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
- Or:
- \$ curl -fLo Pipfile.lock <https://raw.githubusercontent.com/moskytw/data-science-with-python/master/Pipfile.lock>
- \$ pipenv sync

Tests with datasets

To buy, or not to buy

- Going to buy a **bulb** on an online store.
- If see *10/100* bad reviews? Hmm ...
- If see *5/100* bad reviews? Good to buy.
- If see *1/100* bad reviews? Good to buy.

To buy, or not to buy (cont.)

- Going to buy a **notebook computer** on an online store.
- If see *10/100* bad reviews? Hmm ...
- If see *5/100* bad reviews? Hmm ...
- If see *1/100* bad reviews? Maybe good enough.

Build our “bad reviews” in statistics

- Build a statistical model.
 - “The means of two populations are equal.”
- Put the data into the model, get a probability, *p-value*.
 - “How compatible the data and the model are.”
 - If see *p-value* = 0.10?
 - If see *p-value* = 0.05?
 - If see *p-value* = 0.01?
 - Depends on your research question.
- It is just the “Hypothesis Testing”.

Null, alternative, and p-value

- Null hypothesis: can build a model directly:
 - “The means of two populations are equal.”
 - ≡ “The expected value of difference is zero.”
 - E.g., GMV didn't change.
- Alternative hypothesis: can build a model by negating it.
 - “The means of two populations are different.”
 - ≡ Not “The expected value of difference is zero.”
 - E.g., GMV changed.
- P-value: how compatible data and a null hypothesis are.

Misunderstandings of p-values

- “Buy or not” is a decision based on the study context.
 - In other words, “accept” or “reject” a hypothesis.
 - Not “prove a hypothesis is true or false.”
- Misunderstandings of p-values – Wikipedia

Suggested formatting

p-value & α	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

- Many researchers also suggest to report **without** formatting.
 - Since the largely misunderstandings, e.g.,
 - $p\text{-value} < 0.05 \equiv$ the hypothesis is false (**wrong**)
- Scientists rise up against statistical significance – Natural
 - “We are not calling for a ban on P values. Nor are we saying they cannot be used as a decision criterion in certain specialized applications.”
 - “We are calling for a stop to the use of P values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis.”

Go with the notebooks

- The notebooks are available on <https://github.com/moskytw/hypothesis-testing-with-python>.
- *01_tests_with_simulated_datasets.ipynb*
- *02_tests_with_actual_datasets.ipynb* [skip]

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 homogeneity:
 - If total sample size < 1000, or more than 20% of cells have expected frequencies < 5, **Fisher's exact test**.
 - Else, **chi-squared test**, or **two-proportion z-test** ($\equiv 2 \times 2$ chi-squared test).
- If testing equality:
 - 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:
 - Common statistical tests are linear models – Lindeloev
 - 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
 - Chi-squared test of proportions is identical to z-test squared – Rinterested
 - 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

Complete a test

The elements of a complete test

1. The null hypothesis, data, p-value, α .
 2. The raw effect size, β , sample size.
 3. The false negative rate, inverse α , inverse β .
- Will introduce them by the confusion matrix.

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

		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
actual positive CD	false negative C	true positive D	

False positive rate = $P(BD|AB) = B/AB = 4/(96+4) = 4/100$

.....

		predicted negative AC	predicted positive BD
actual negative AB	96 A	4 B	
	9 C	41 D	

How similar! Let's try to rename:

.....

null hypothesis is true

actual negative

alternative hypothesis is true

actual positive

accept null hypothesis

predict negative

accept alternative hypothesis

predict positive

$$\alpha = P(\text{accept alt} | \text{null}) = P(\text{predicted positive} | \text{actual negative})$$

.....

		predicted negative	predicted positive
actual negative	AB	AC	BD
	true negative	A	false positive
actual positive	CD	CB	DA
	false negative	C	true positive

Predefined acceptable confusion matrix

		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
	false negative C	true positive D	
actual positive CD			

False positive, p-value, and α

false positive rate

Calculated with the actual answer.

p-value

Calculated false positive rate
by a null hypothesis.

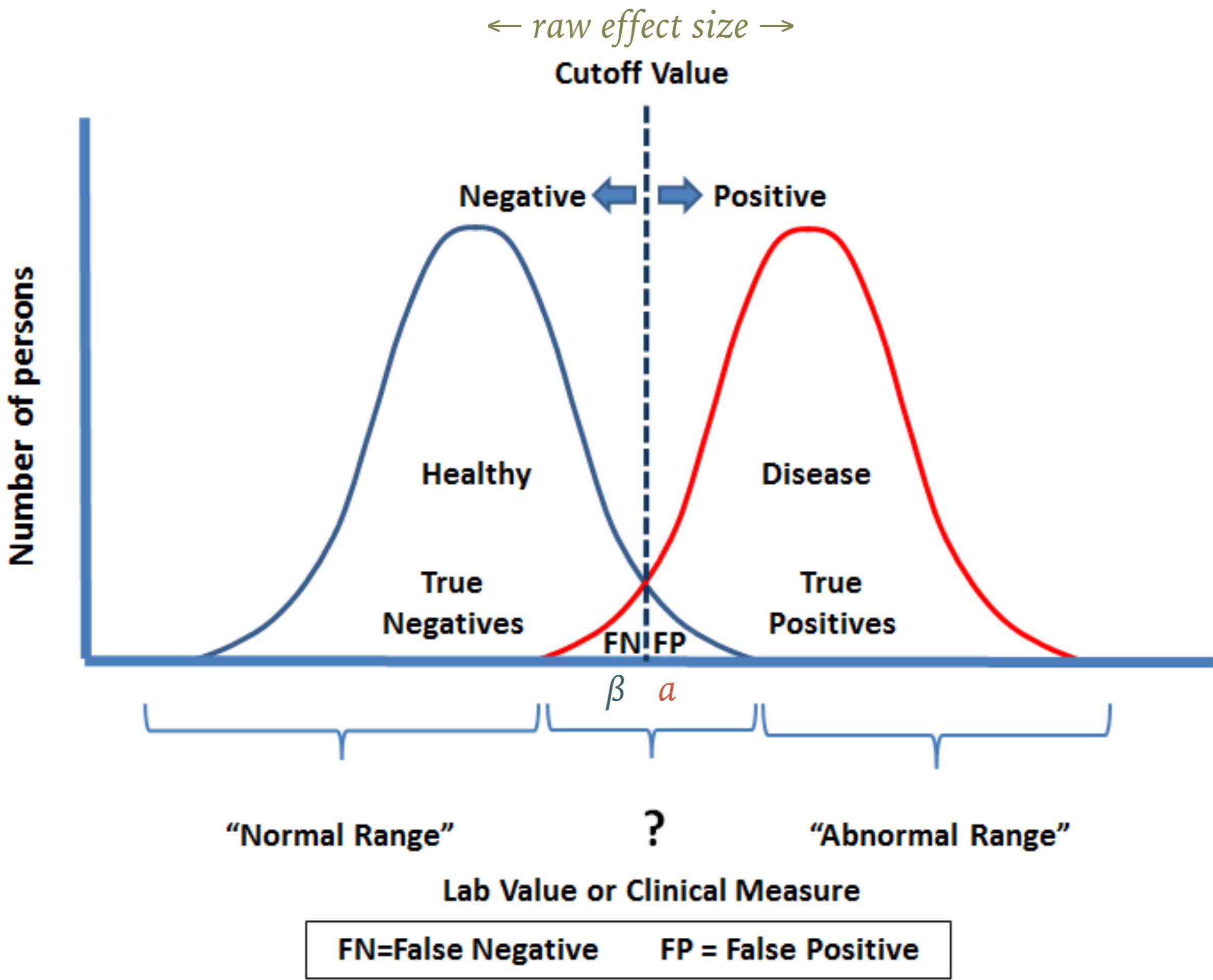
α

Predefined acceptable
false positive rate.

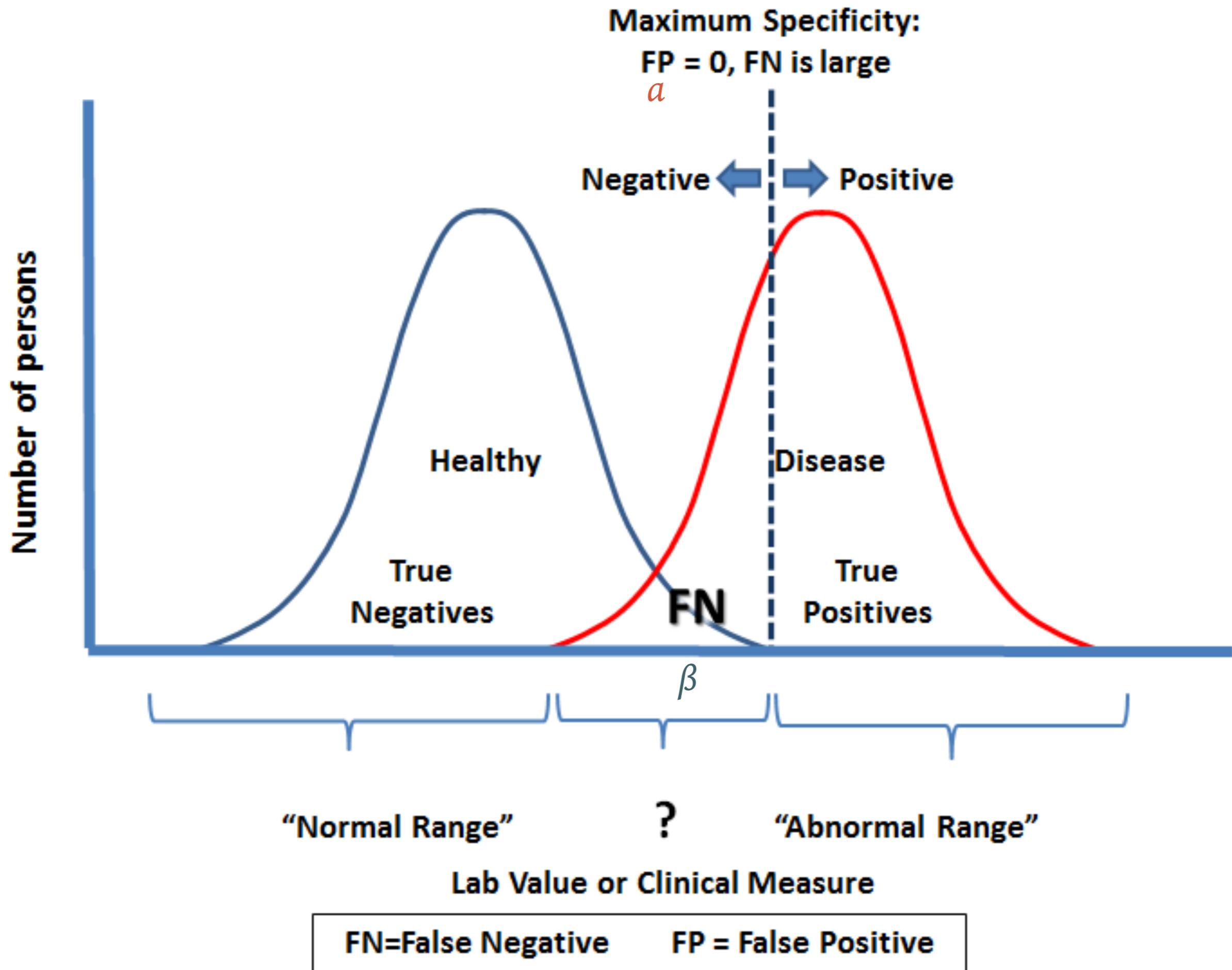
Raw effect size, and β

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

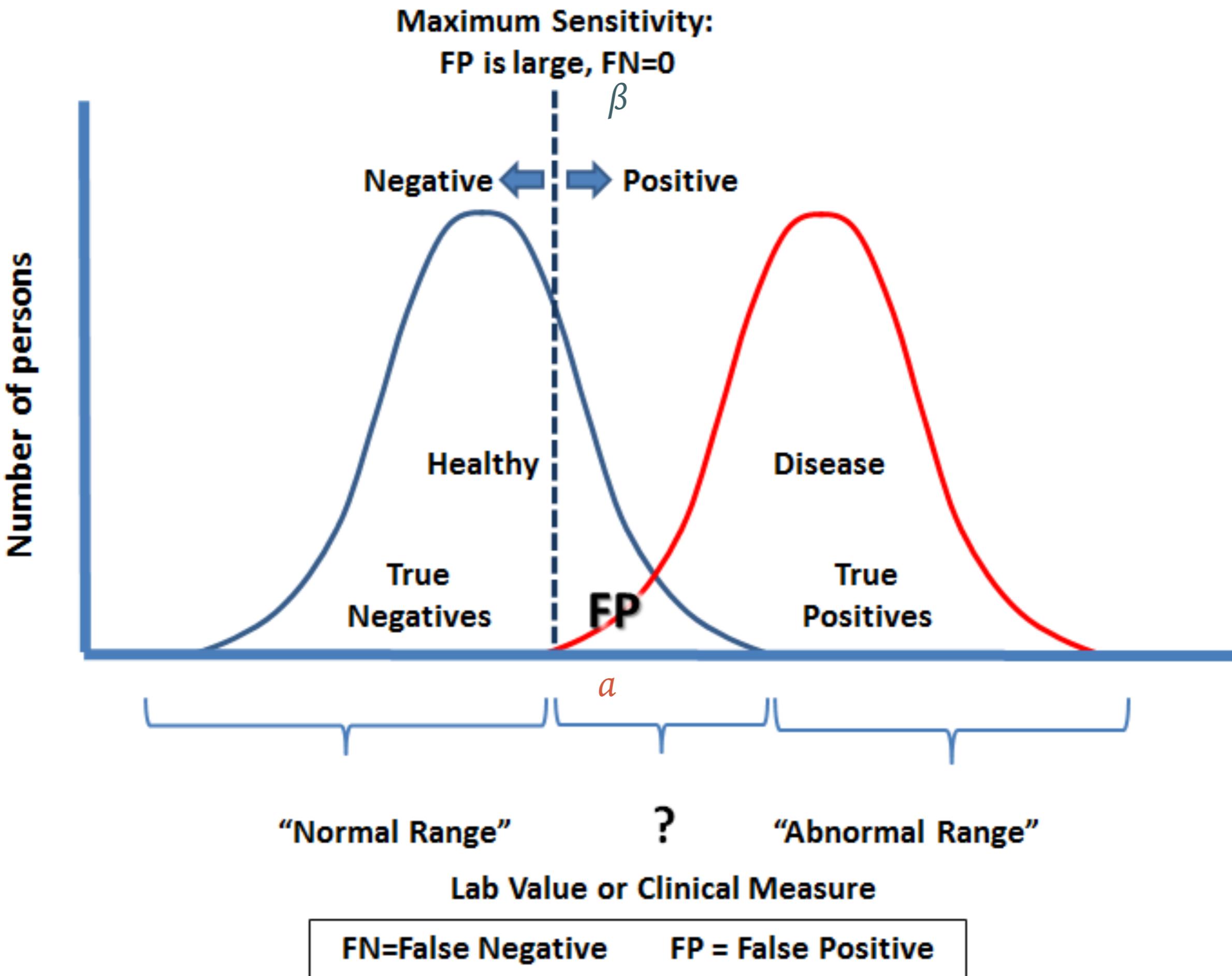
B.



C.



D.



sample size ↑

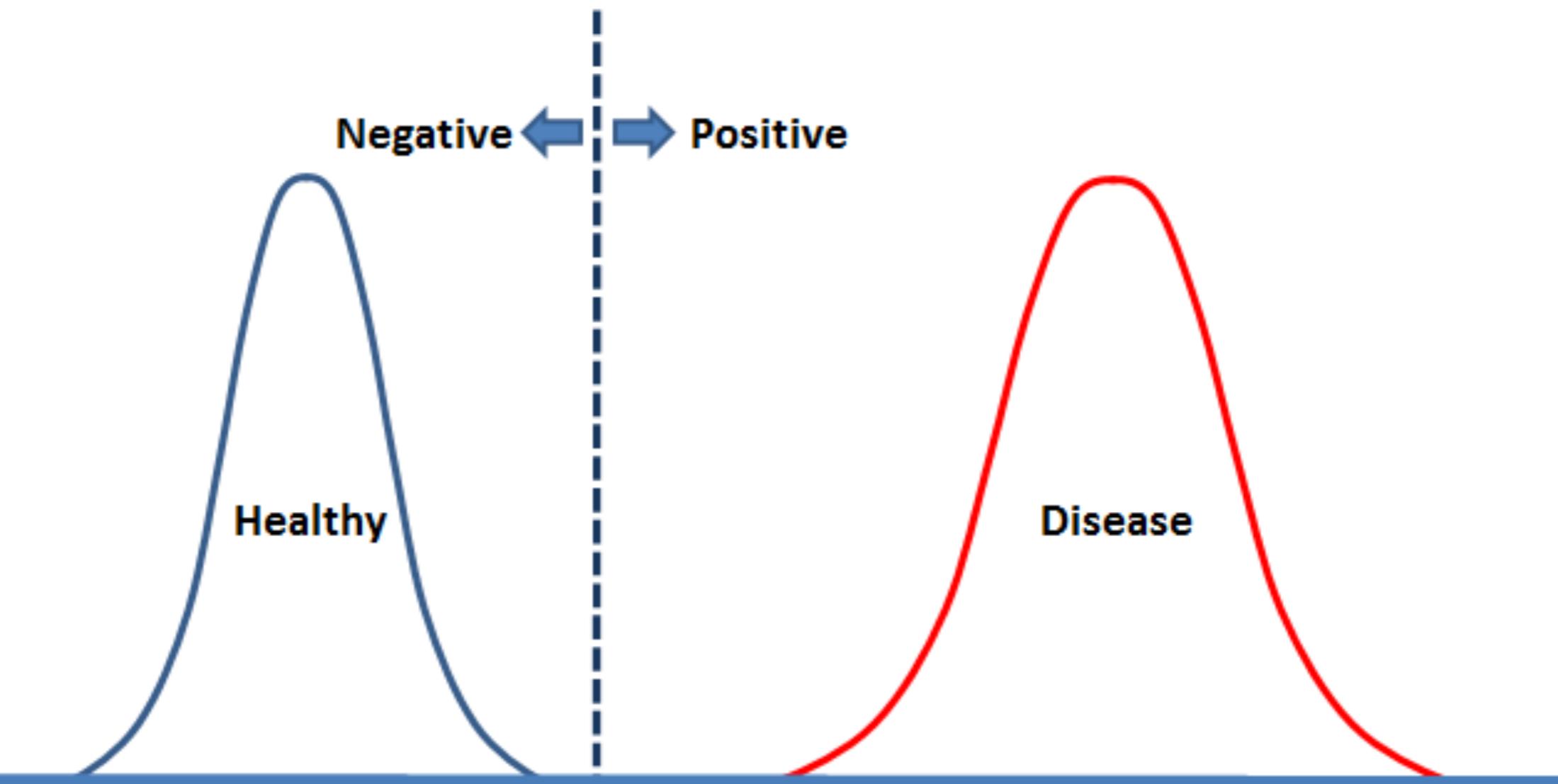
Cutoff Value

Negative ← → **Positive**

Number of persons

Healthy

Disease



“Normal Range”

“Abnormal Range”

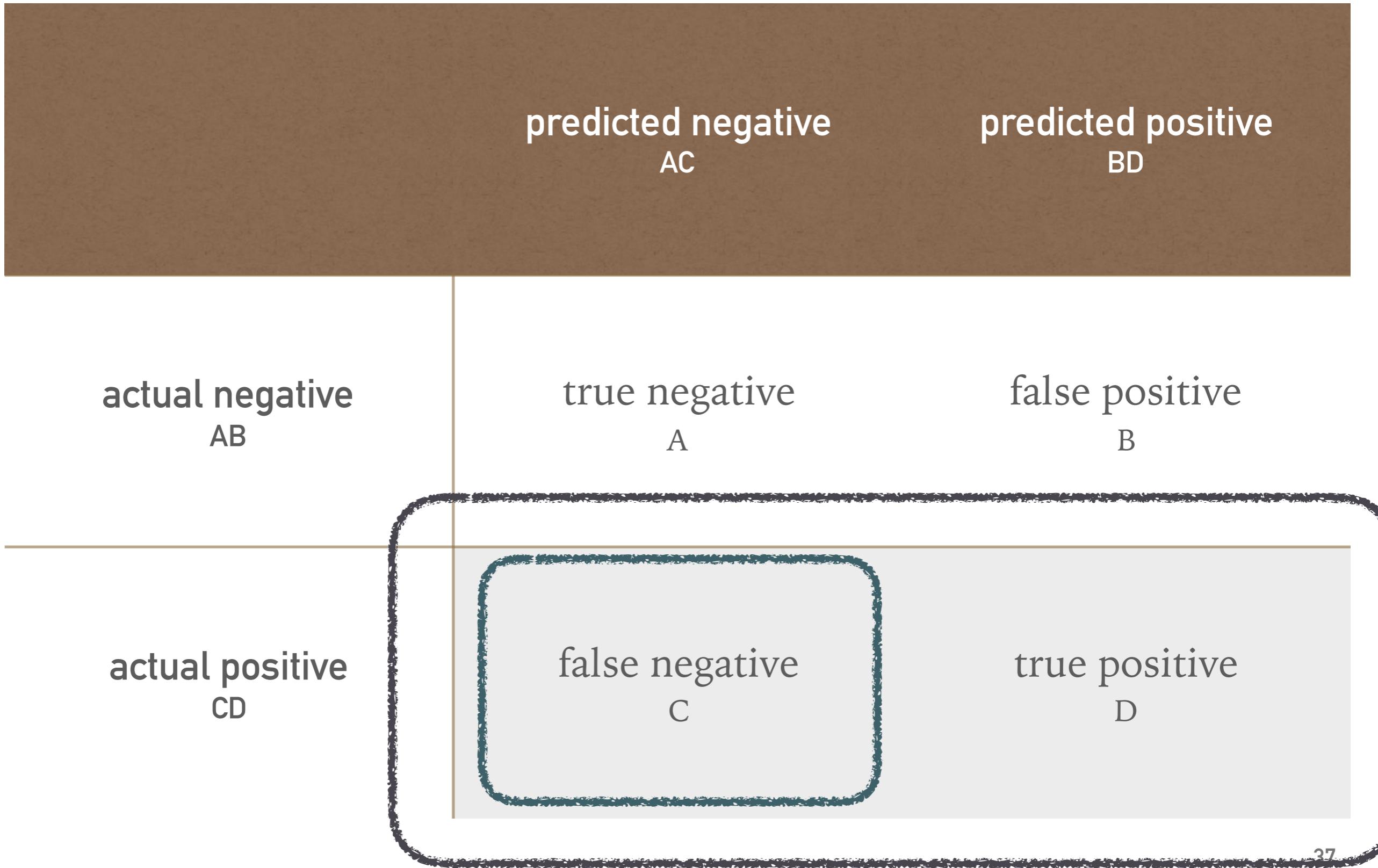
Lab Value or Clinical Measure

A.

$$\beta = P(AC|CD) = C/CD$$

.....

		predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B	
actual positive CD	false negative C	true positive D	



- Given α , β , raw effect size, get the sample size.
- Given α , raw effect size, sample size, get the β .
- Increase sample size to decrease α , β , or raw effect size.

Given α , raw effect size, sample size, get the β

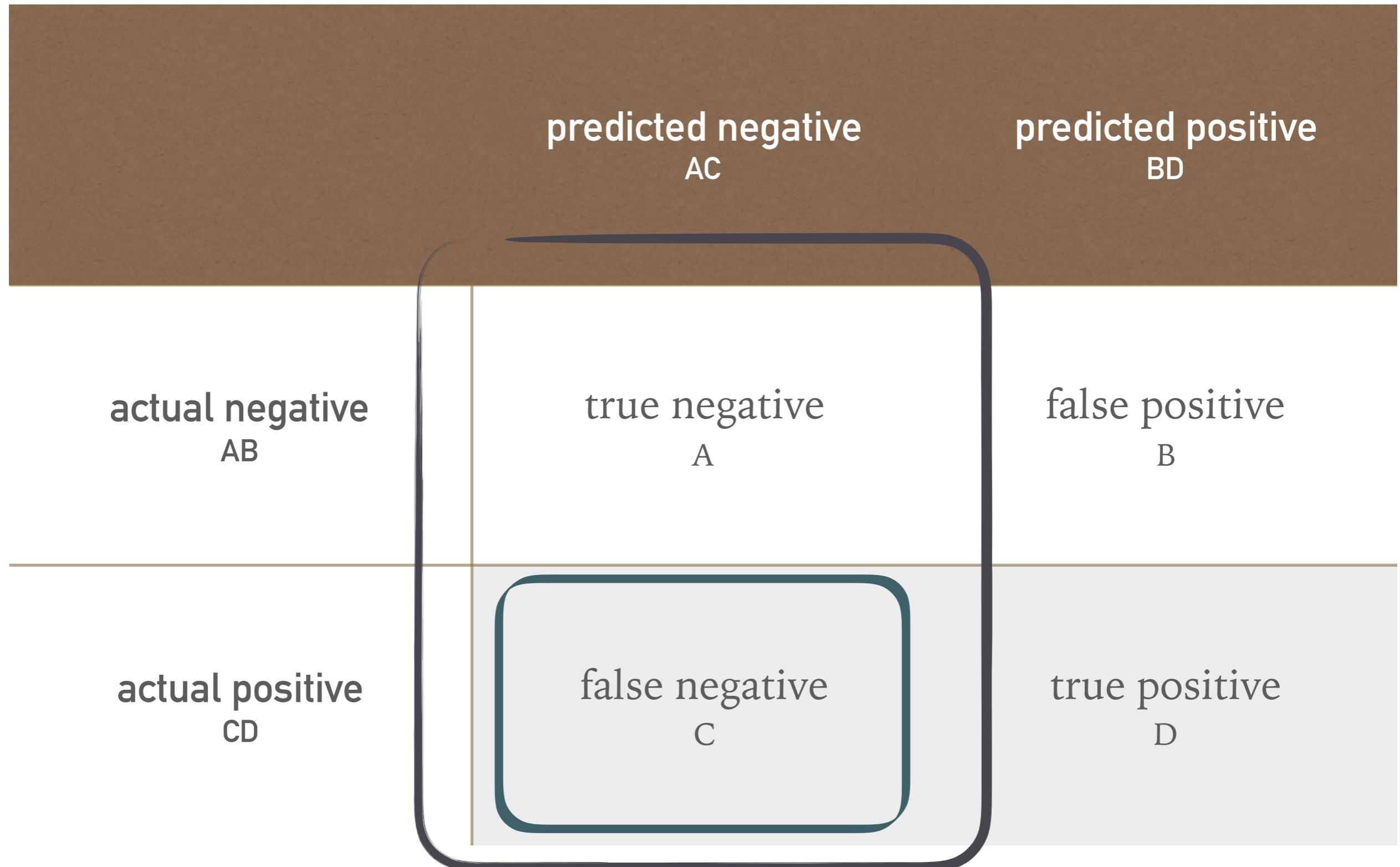
- “The conversion rates are the same.”
- $\alpha = 0.05$
- $\text{raw effect size} = 4\% - 3\% = 1\%$
- With two-proportion z-test, given:
 - $\text{sample size} = 1,000$, get $\beta = 0.86$
 - $\text{sample size} = 10,000$, get $\beta = 0.22$
 - $\text{sample size} = 17,550$, get $\beta = 0.05$

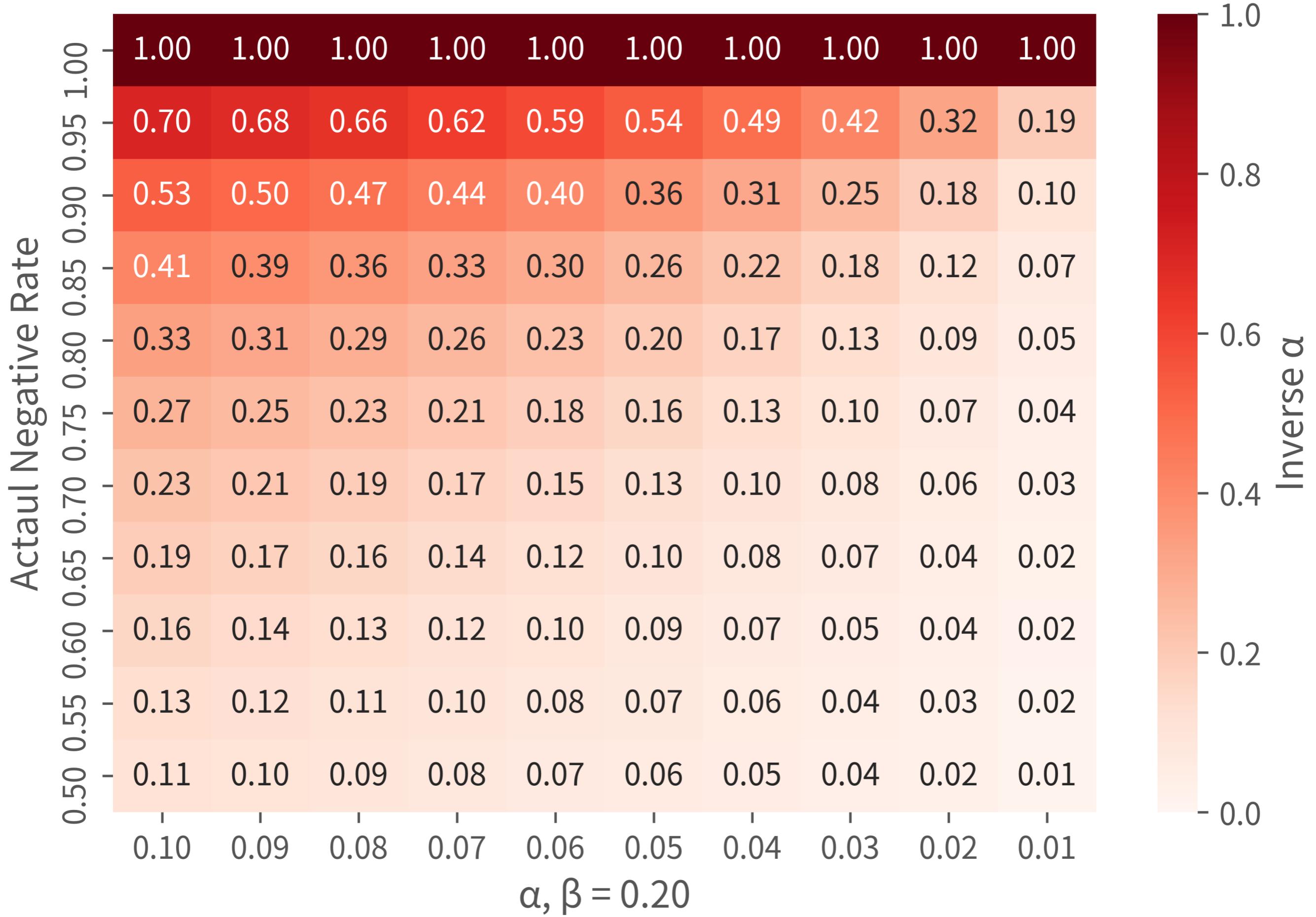
$$\text{Inverse } a = P(AB|BD) = B/BD$$

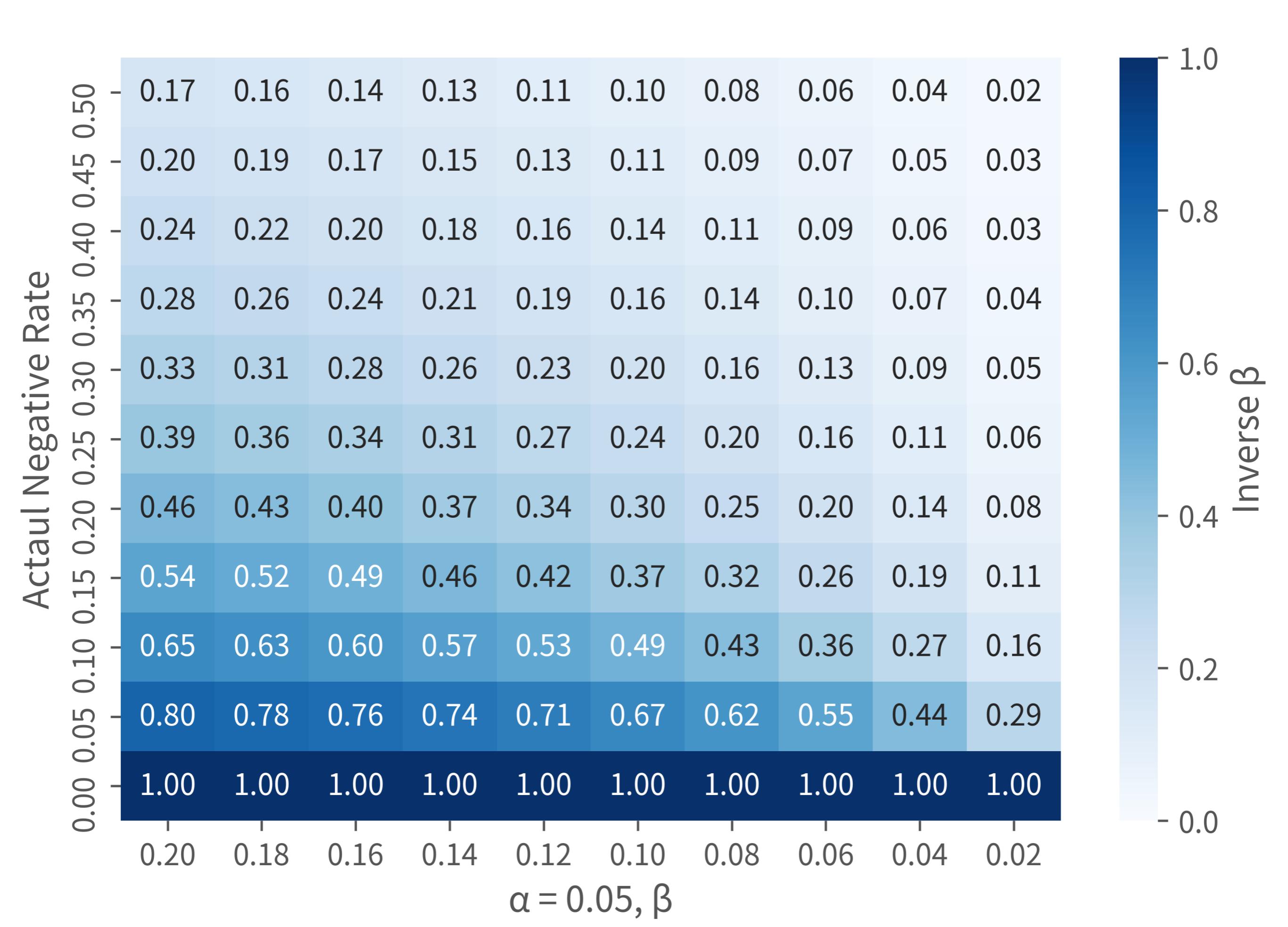
	predicted negative AC	predicted positive BD
actual negative AB	true negative A	false positive B
actual positive CD	false negative C	true positive D

$$\text{Inverse } \beta = P(CD|AC) = C/AC$$

.....







Rates in predefined acceptable confusion matrix

= = = predefined

a	B/AB	significance level type I error rate	false positive rate
β	C/CD	type II error rate	false negative rate
inverse a	B/BD		false discovery rate
inverse β	C/AC		false omission rate
confidence level	A/AB	1-a	specificity
power	D/CD	1- β	sensitivity recall

Rates in confusion matrix

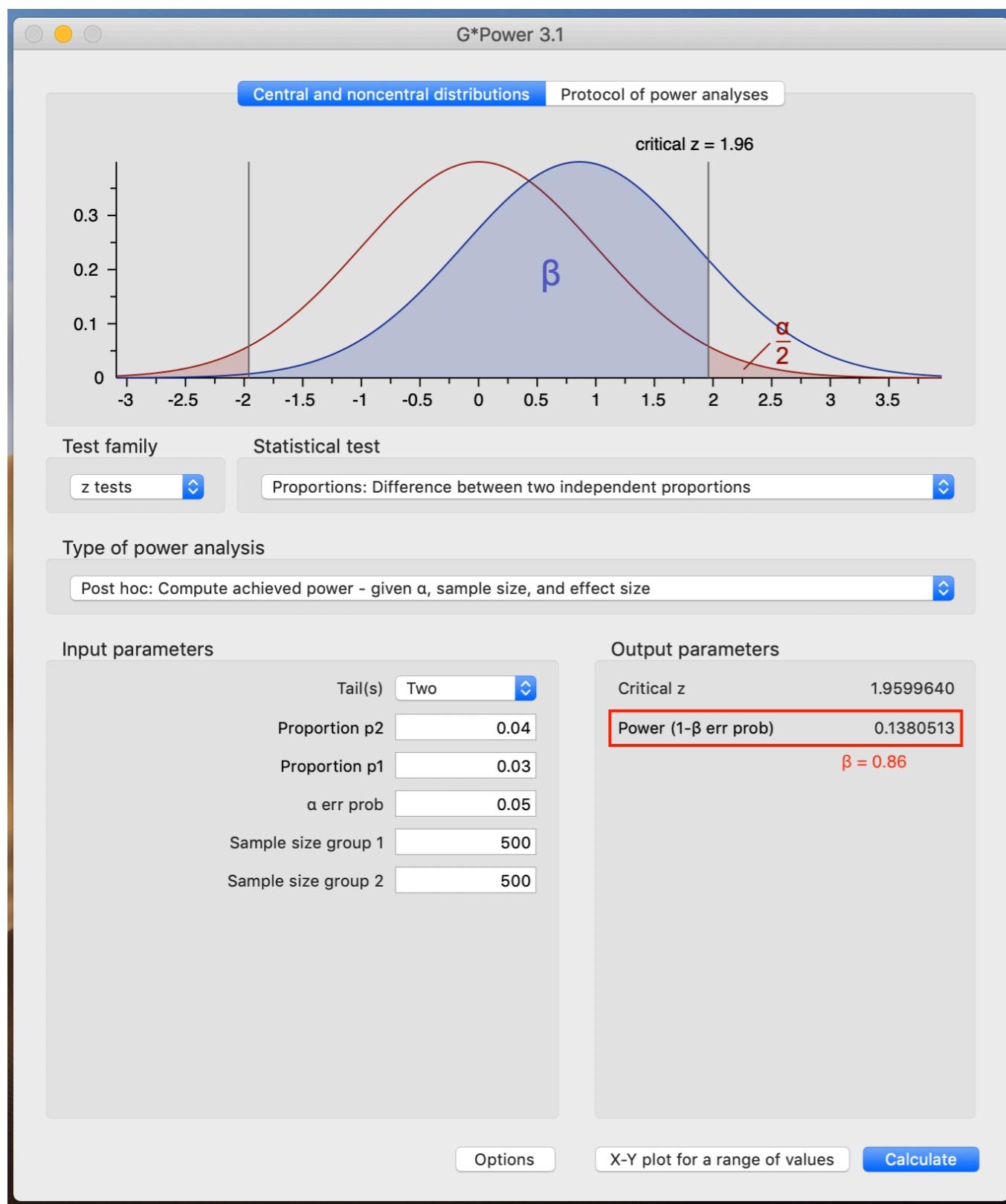
	=	=	= observed
false positive rate	B/AB		α
false negative rate	C/CD		β
false discovery rate	B/BD		inverse α
false omission rate	C/AC		inverse β
actual negative rate	AB/ABCD		
sensitivity	D/CD	recall	power
specificity	A/AB		confidence level
precision	D/BD		inverse power
recall	D/CD	sensitivity	power

Most formal steps

1. The hypothesis → what test.
2. The *actual negative rate* → how much α, β are required.
3. The $\alpha, \beta, \text{ raw effect size}$ → how large *sample size* is required.
4. Still collect a sample as large as possible.
5. Understand and preprocess the sample.
 - Plotting, missing data, outliers, transform, etc.
6. Test.
7. Report fully.
 - The *mean, confidence interval, p-value, study context*, etc.

Calculate by yourself

.....



➤ G*Power is awesome.

➤ <http://www.gpower.hhu.de/>

Keep learning

1. Seeing Theory
2. Statistics – SciPy Tutorial
3. StatsModels
4. Biological Statistics
5. Research method
 - Study design, experimental design, survey design etc.

Recap

1. *p-value*:
 - How compatible data and a null hypothesis are.
 - The false positive rate by a null hypothesis.
2. β does matter.
3. *actual negative rate* does matter.
4. The study context does matter.
5. Visualization and simulation do help.
 - More: 04, 05, and a1 notebooks in handouts.
6. Let's accept or reject a hypothesis efficiently!