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

Hypothesis testing

- "The means of two populations are equal."
- $p\text{-}value = 0.90$
- $p\text{-}value = 0.80$
- ...
- $p\text{-}value = 0.05$
- $p\text{-}value = 0.01$

Hypothesis testing

- "The means of two populations are equal."
- Case 1: *p-value* $\geq$ *0.05*
  - "Hmmm ... can't tell."
- Case 2: *p-value* $<$ *0.05*
  - "Not equal! So extreme!"

Hypothesis testing in a "null" taste

- *<null hypothesis>*
- Case 1: *p-value* $\geq$ *0.05*
  - Can't reject *<null hypothesis>*.
- Case 2: *p-value* $<$ *0.05*
  - Reject *<null hypothesis>*.

Hypothesis testing in an "alternative" taste

- <*alternative hypothesis*>, ≡ not <*null hypothesis*> usually.
- Case 1: *p-value* ≥ *0.05*
  - Can't accept <*alternative hypothesis*>.
- Case 2: *p-value* < *0.05*
  - Accept <*alternative hypothesis*>.

Hypothesis testing in "−+" taste

- "The case is negative."
- Case 1: *p-value* ≥ *a*
  - "Hmmm ... somehow negative."
- Case 2: *p-value* < *a*
  - "Positive! So extreme!"

P-value, α, wording, and summary

p-value & α	Wording	Summary
p-value < 0.001	Very significant	***
p-value < 0.01	Very significant	**
p-value < 0.05	Significant	*
p-value ≥ 0.05	Not significant	ns

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.

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*

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 & α enough?

- sample size?
- power?
- β?
- raw effect size?
- ? ? ? ? ?

Raw effect size, α, 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.
  - "The distance between the means": *raw effect size*
  - "FP": a
  - "FN": β

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

Confusion matrix, where $A = OO_2 = C[0, 0]$

	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 = B/AB = 4/(96+4) = 4/100

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

Predefined acceptable confusion matrix

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

$\alpha = B/AB = P(BD|AB)$

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

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

	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, raw effect size, sample size, get the β

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

Inverse α = B/BD = P(AB|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

Inverse β = C/AC = P(CD|AC)

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

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

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

Inverse β

1.0
0.8
0.6
0.4
0.2
0.0

Rates in predefined acceptable confusion matrix

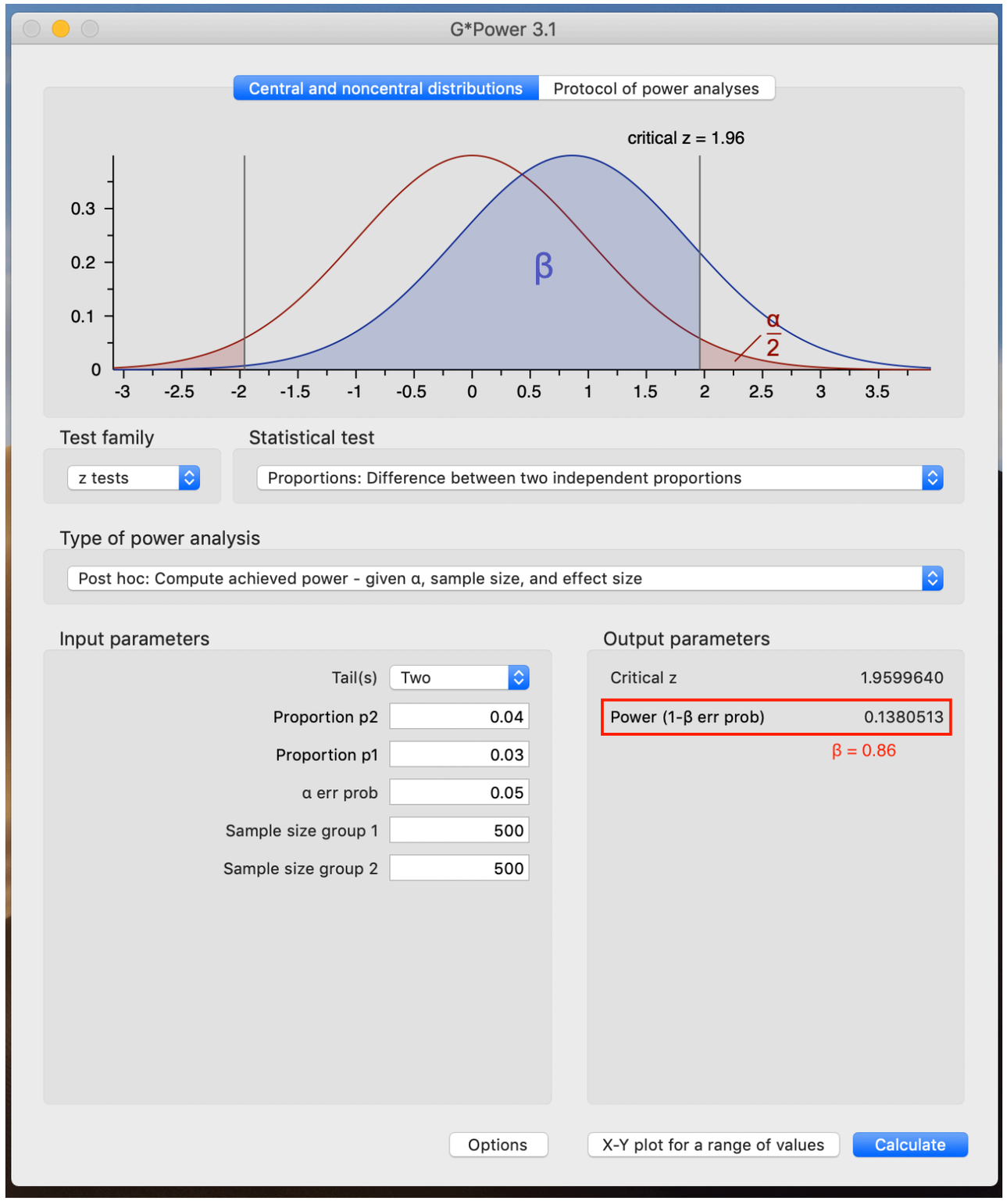
	=	=	= predefined
α	B/AB	significance level type I error rate	false positive rate
β	C/CD	type II error rate	false negative rate
inverse α	B/BD		false discovery rate
inverse β	C/AC		false omission rate
confidence level	A/AB	1-α	specificity
power	D/CD	1-β	sensitivity recall

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

- State the hypothesis → what *test*.
- Estimate the *actual negative rate*.
- The a*ctual negative rate* → how much a, β are required.
- The a, β, *raw 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 is awesome.
  - http://www.gpower.hhu.de/
- *05_complete_a_test.ipynb*

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.
- β does matter.
- *a, β, raw effect size ↓; sample size ↑.*
- Visualization and simulation do help.
- Bonus: a1_figures.ipynb .
- Let's identify noise efficiently!