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 versions.

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, 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

False positive rate, p-value, and α

false positive rate	Calculated with the actual answer.
p-value	Calculated false positive rate by a reasonable hypothesis.
α	Predefined acceptable false positive rate.

How tests work

Seeing is believing

- $p\text{-}value = 0.0027\ (< 0.01)$
  - 👏👏👏
- $p\text{-}value = 0.0271\ (0.01–0.05)$
  - 👏❓👏❓❓❓
- $p\text{-}value = 0.2718\ (\geq 0.05)$
  - ❓❓❓❓❓❓
- *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-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* $\geq a$
  - Can't reject *<null hypothesis>*.
- Case 2: *p-value* $< a$
  - Reject *<null hypothesis>*.

Hypothesis testing in an "alternative" taste

- <*alternative hypothesis*>, ≡ not <*null hypothesis*> usually.
- Case 1: *p-value* ≥ *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-value* ≥ *a*
  - "Hmmm ... somehow negative."
- Case 2: *p-value* < *a*
  - "Positive! So extreme!"

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?

- β?
- sample size?
- effect size?
- **? ? ? ? ?**

Confusion matrix, where A = OO_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

Sample size ↑ ; α, β, effect size ↓

- Increase *sample size* to decrease *a*, *β*, or *effect size*.
  - The *effect size* is the detectable distance between groups.
    - $= \frac{\mu_1 - \mu_2}{\sigma}$
  - DSM5: The case for double standards – James Coplan, M.D.
    - 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	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
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

Effect Size

Sample Size

300
250
200
150
100
50
0

α, β = 0.01	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
.001	630989	157749	70112	39439	25242	17530	12880	9862	7793	6313
.002	586789	146699	65201	36677	23474	16302	11978	9171	7247	5870
.003	560549	140139	62285	35036	22424	15573	11442	8761	6923	5608
.004	541741	135437	60195	33861	21672	15050	11058	8467	6690	5419
.005	527034	131760	58561	32941	21083	14642	10758	8237	6509	5272
.006	514938	128736	57217	32185	20599	14306	10511	8048	6359	5151
.007	504651	126164	56074	31542	20188	14020	10301	7887	6232	5048
.008	495695	123925	55079	30983	19829	13771	10118	7747	6121	4959
.009	487758	121941	54197	30486	19512	13550	9956	7623	6023	4879
.010	480629	120158	53405	30041	19227	13352	9810	7511	5935	4808

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 $n \to \infty$:
    - $\mu = np$
    - $\sigma = \sqrt{np(1\text{-}p)}$
    - $X \sim N(\mu, \sigma)$
- See also: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3775042/ .

Inverse α

- 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 not a drug user?

$$P(non\text{-}user \mid +) = \frac{P(+ \mid non\text{-}user)P(non\text{-}user)}{P(+)}$$

$$= \frac{P(+ \mid non\text{-}user)P(non\text{-}user)}{P(+ \mid non\text{-}user)P(non\text{-}user) + P(+ \mid user)P(user)}$$

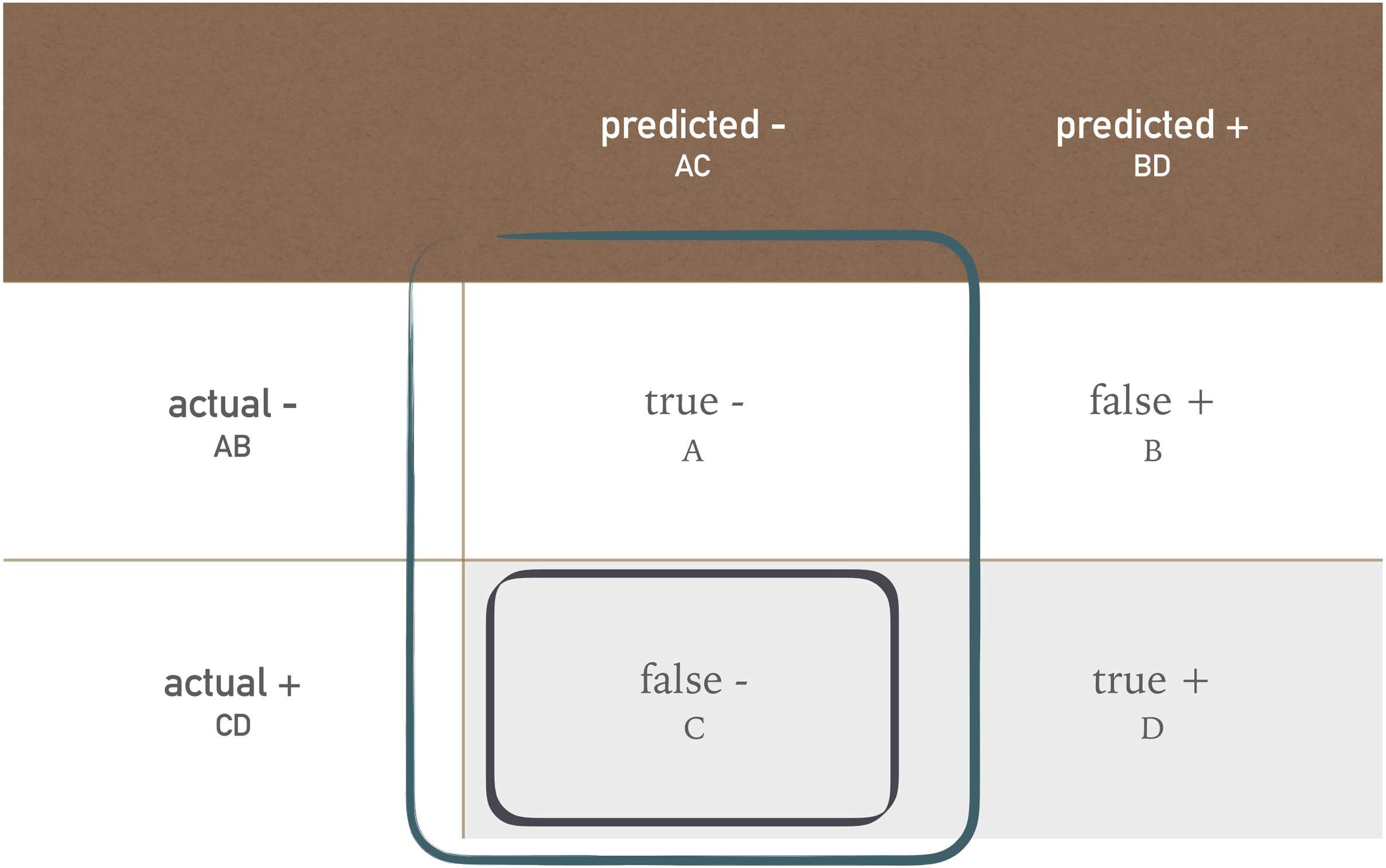
$$= \frac{0.01 \times 0.995}{0.01 \times 0.995 + 0.99 \times 0.005}$$

$$\approx 66.8\,\%$$

False discovery rate = B / BD = observed inverse α

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

False omission rate = C / AC = observed inverse β



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 \ α, β = 0.01	0.010	0.009	0.008	0.007	0.006	0.005	0.004	0.003	0.002	0.001
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.99	0.50	0.47	0.44	0.41	0.37	0.33	0.29	0.23	0.17	0.09
0.98	0.33	0.31	0.28	0.26	0.23	0.20	0.17	0.13	0.09	0.05
0.97	0.25	0.23	0.21	0.19	0.16	0.14	0.12	0.09	0.06	0.03
0.96	0.20	0.18	0.16	0.15	0.13	0.11	0.09	0.07	0.05	0.02
0.95	0.16	0.15	0.13	0.12	0.10	0.09	0.07	0.05	0.04	0.02
0.94	0.14	0.12	0.11	0.10	0.09	0.07	0.06	0.05	0.03	0.02
0.93	0.12	0.11	0.10	0.09	0.07	0.06	0.05	0.04	0.03	0.01
0.92	0.10	0.09	0.09	0.08	0.07	0.05	0.04	0.03	0.02	0.01
0.91	0.09	0.08	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01
0.90	0.08	0.08	0.07	0.06	0.05	0.04	0.04	0.03	0.02	0.01

Inverse α: 1.0, 0.8, 0.6, 0.4, 0.2, 0.0

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

Common "rates" in confusion matrix

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

Common "rates" in hypothesis testing

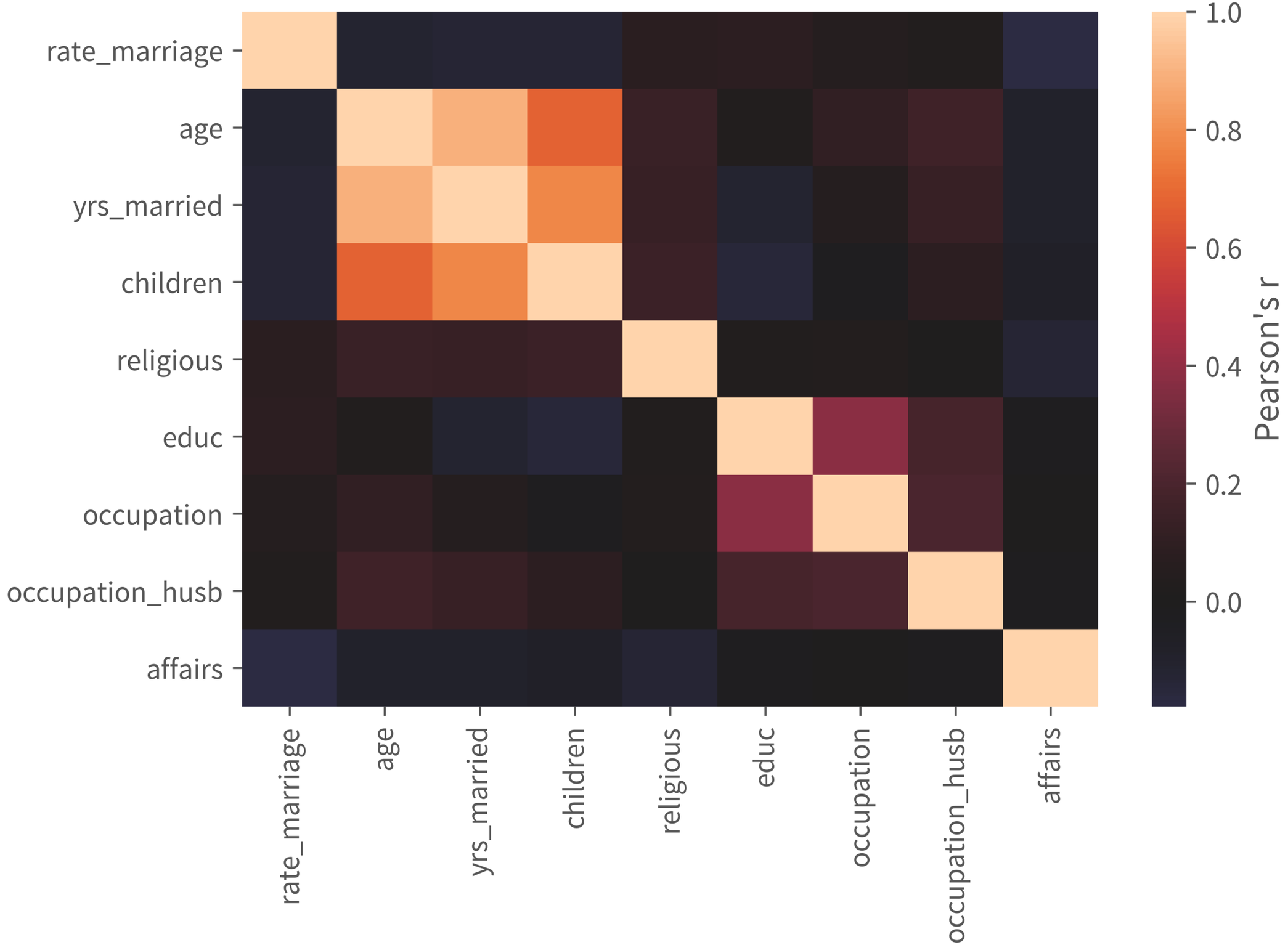
- α = predefined acceptable false positive rate
  - = significance level = type I error rate
- β = predefined acceptable false negative rate
  - = type II error rate
- inverse α = predefined acceptable false discovery rate
- inverse β = predefined acceptable false omission rate
- power = 1 - β = predefined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

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, *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.
- *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 negative".
- *confidence interval*: the values the middle probability maps to.
- *actual negative rate* does matter.
- *a, β, effect size ↓; sample size ↑.*
- Visualization and simulation do help.
- Bonus: *a1_figures.ipynb* .
- Let's identify noise efficiently!