Hypothesis Testing With Python

True Difference or Noise?

0.7196

0.7552

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

- Introduction
- Welch's t-test
- Chi-squared test
- Power analysis
- More tests
- Complete steps

- Theory
  - P-value & α
  - Raw effect size, β, sample Size
  - Actual negative rate, inverse α, inverse β

The PDF, Notebooks, and Packages

- The PDF and notebooks are available on https://github.com/moskytw/hypothesis-testing-with-python .
- 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`

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.

- 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.
- Context matters.

Build our "bad reviews" in statistics

- Build a statistical model by a hypothesis.
  - "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 context.

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."
  - ≡ ¬ "The expected value of difference is zero."
  - E.g., GMV changed.
- **P-value**: the probability to observe the data, given **null**.

Misunderstandings of p-values

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

Suggested formatting

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

- Many researchers also suggest to report **without** formatting.
  - Since the largely misunderstandings, e.g.,
    - *p-value* $< 0.05 \equiv$ the null 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."

Define assumptions

- The hypothesis testing:
  - Suitable to answer a yes–no question:
    - "Means or medians of two populations are equal?"
      - E.g., "The order counts of A and B are equal?"
    - "Proportions of two populations are equal?"
      - E.g., "The conversion rates of A and B are equal?"

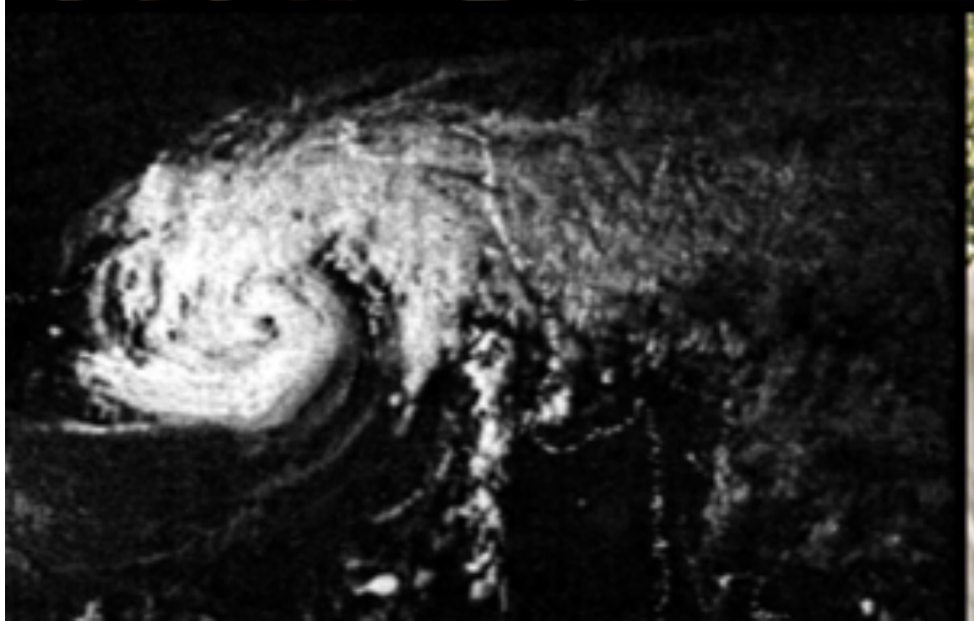
- “Poor or non-poor marriage has different affair times?”
- “Occupations have different affair times?”
- “Poor or non-poor marriage has different affair proportion?”
- “Occupations have different affair proportion?”

Validate assumptions

- Collect data ...
  - The "Fair" dataset:
    - Fair, Ray. 1978. "A Theory of Extramarital Affairs," Journal of Political Economy, February, 45-61.
      - A dataset from 1970s.
      - Rows: 6,366
      - Columns: (next slide)
- The full version of the analysis steps: http://bit.ly/analysis-steps .

1. *rate_marriage*: 1∼5; very poor, poor, fair, good, very good.
2. *age*
3. *yrs_married*
4. *children*: number of children.
5. *religious*: 1∼4; not, mildly, fairly, strongly.
6. *educ*: 9, 12, 14, 16, 17, 20; grade school, some college, college graduate, some graduate school, advanced degree.
7. *occupation*: 1, 2, 3, 4, 5, 6; student, farming-like, white-colloar, teacher-like, business-like, professional with advanced degree.
8. *occupation_husb*
9. *affairs*: n times of extramarital affairs per year since marriage.


JESUS SAV
All up
with the
Holy
Ghost
& Fire

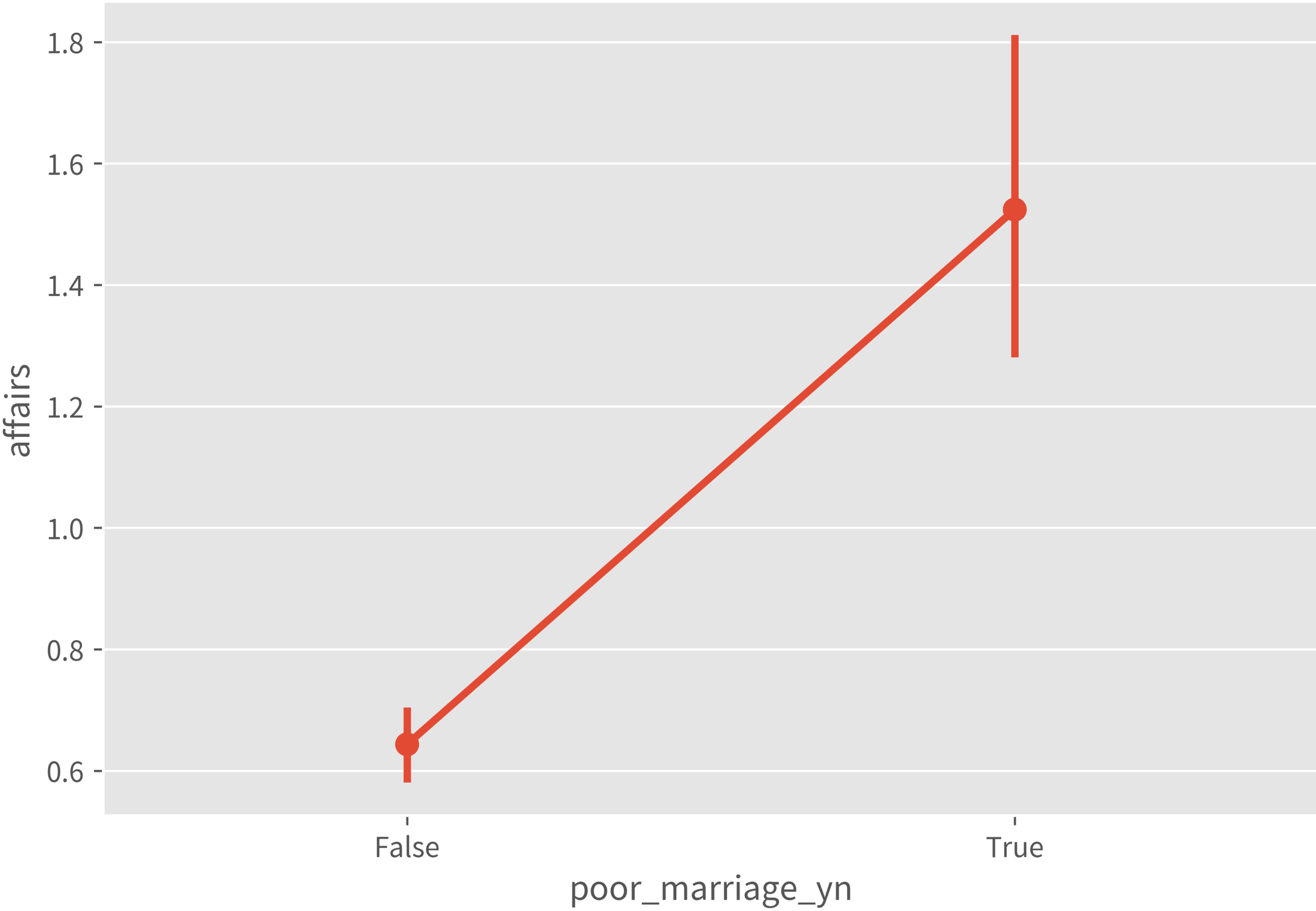


poor_marriage_yn	count	mean	std
False	5919.0	0.643549	2.116982
True	447.0	1.524038	3.015937

```
p-value: 2.7446844166802127e-09
```

Welch's t-test #1

- Preprocess:
  - Group into poor or not.
- Describe.
- Test:
  - Assume no difference, the probability to observe it: super low.
- So, we accept the means are different at 99.9% confidence level (= 1-a):
  - Non-poor: 0.6435
  - Poor: 1.5240


affairs
1.8
1.6
1.4
1.2
1.0
0.8
0.6
False
True
poor_marriage_yn

```
import scipy as sp
import statsmodels.api as sm


print(sm.datasets.fair.SOURCE,
      sm.datasets.fair.NOTE)


# -> Pandas's Dataframe
df_fair = sm.datasets.fair.load_pandas().data


df = df_fair
# 2: poor
# 3: fair
df = df.assign(poor_marriage_yn
               =(df.rate_marriage <= 2))
df_fair_11 = df
```

```
df = df_fair_11

display(df
        .groupby('poor_marriage_yn')
        .affairs
        .describe())

a = df[df.poor_marriage_yn].affairs
b = df[~df.poor_marriage_yn].affairs

# ttest_ind(...) === Student's t-test
# ttest_ind(..., equal_var=False) === Welch's t-test
print('p-value:',
      sp.stats.ttest_ind(a, b, equal_var=False)[1])
```

```
df = df_fair_11
sns.pointplot(x=df.poor_marriage_yn,
              y=df.affairs)
```

Why Welch's t-test, not Student's t-test?

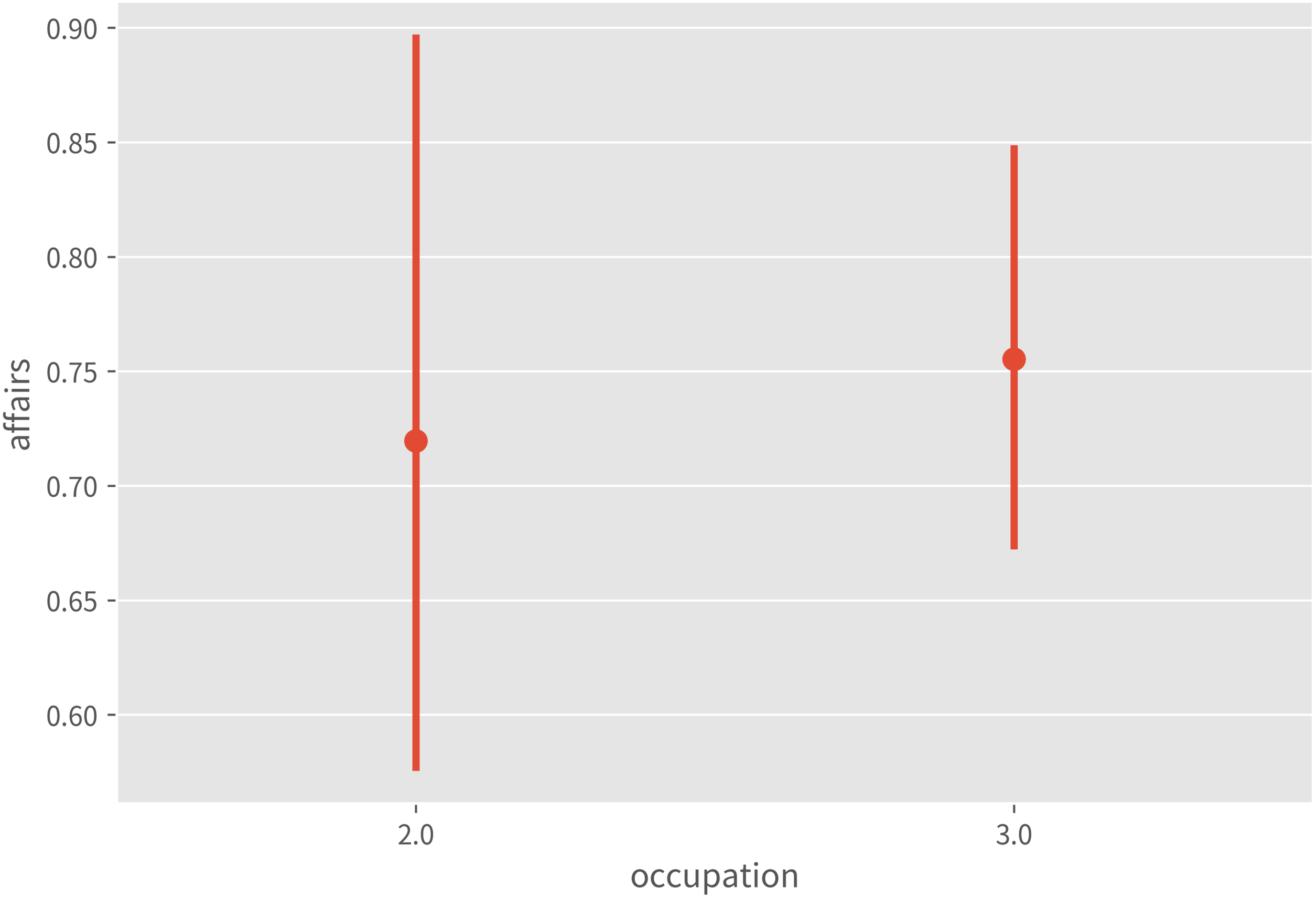
- Student's t-test assumed the two populations have the same variance, which may not be true in most cases.
- Welch's t-test relaxed this assumption without side effects.
- So, just use Welch's t-test directly. [ref]

Welch's t-test #2

occupation	count	mean	std
2.0	859.0	0.719556	2.375644
3.0	2783.0	0.755248	2.305594

```
p-value: 0.698381462473247
```

- Preprocess:
  - Select the two occupations.
  - Group by occupations.
- Describe.
- Test:
  - Assume no difference, the probability to observe it: 70%.
- So, we can't accept the means are different at 99% confidence level.

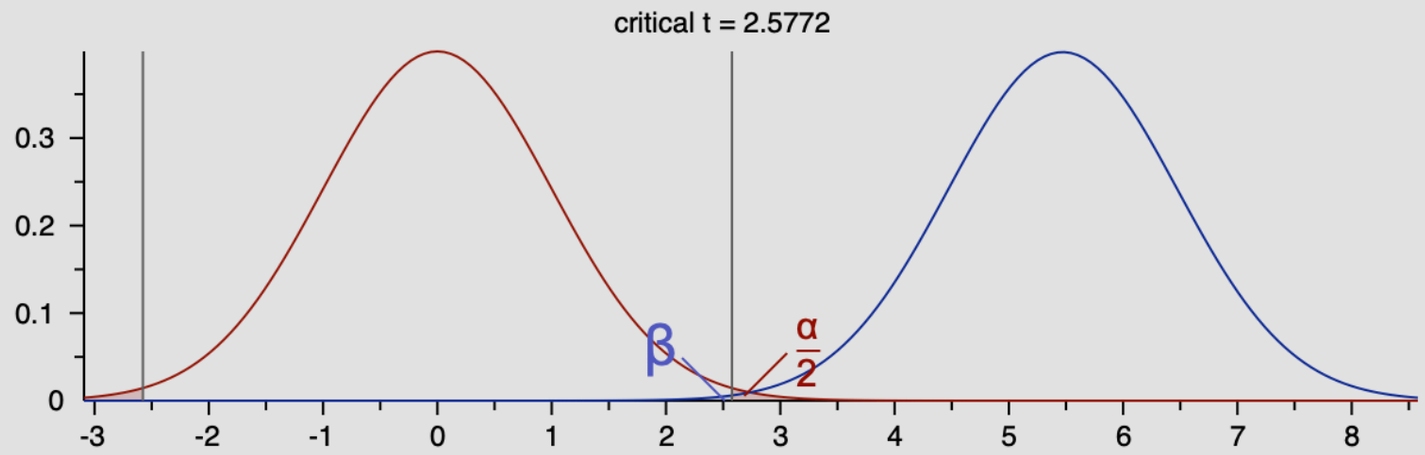

0.90
0.85
0.80
0.75
0.70
0.65
0.60
affairs
2.0
3.0
occupation

Power analysis for Welch's t-test #2

- We can't accept the alter, but if the alter is true?
- The α is "given null, accept alter", = 1-*confidence level*.
- The β is "given alter, don't accept alter", = 1-*power*.
- Hope to detect ≥ 0.5 difference at 99% confidence level:
  - *raw effect size* = 0.5
  - $a = 0.01$
- Use G*Power or StatsModels:
  - *power* = 0.9981
  - $\beta = 0.0019$
- Since the *p-value* ≥ a and β is 0.2%, we accept the difference < 0.5.
- If β is high, collect more sample or relax the α or effect size.

G*Power 3.1

Central and noncentral distributions | Protocol of power analyses



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: Post hoc: Compute achieved power - given α, sample size, and effect size

Input parameters

Tail(s)	Two
Effect size d	0.2135952
α err prob	0.01
Sample size group 1	859
Sample size group 2	2783

Determine

Output parameters

Noncentrality parameter δ	5.4723605
Critical t	2.5771807
Df	3640
Power (1-β err prob)	0.9980984

X-Y plot for a range of values | Calculate

n1 ≠ n2

Mean group 1	0
Mean group 2	1
SD σ within each group	0.5

n1 = n2

Mean group 1	0.719556
Mean group 2	1.219557
SD σ group 1	2.375644
SD σ group 2	2.305594

Calculate | Effect 0.2135952

Calculate and transfer to main window

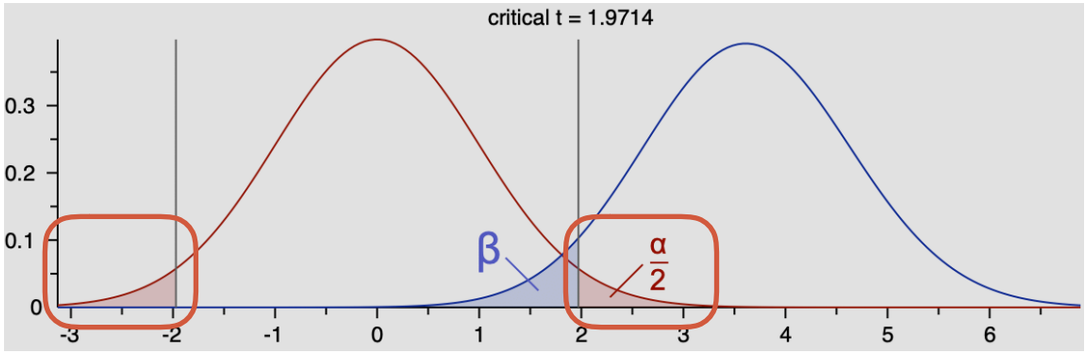
Close effect size drawer

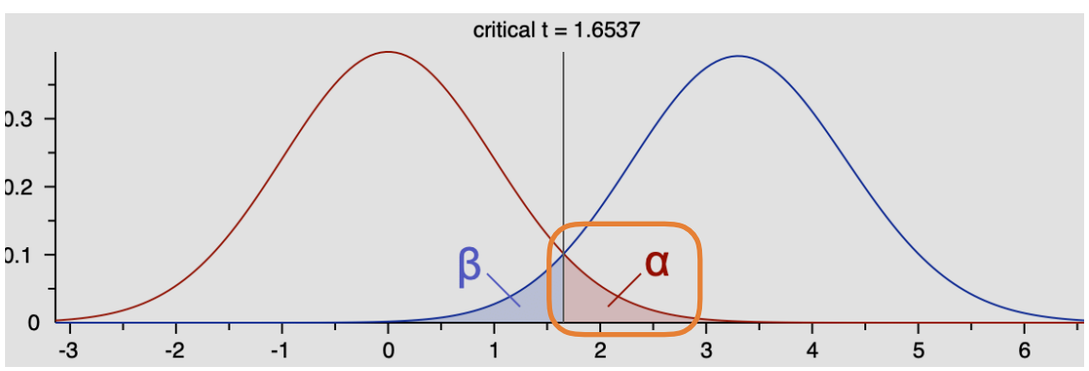
```
df = df_fair
# 2: farming-like
# 3: white-colloar
df = df[df.occupation.isin([2, 3])]
df_fair_12 = df


df = df_fair_12

display(df
        .groupby('occupation')
        .affairs
        .describe())

a = df[df.occupation == 2].affairs
b = df[df.occupation == 3].affairs

print('p-value:',
      sp.stats.ttest_ind(a, b, equal_var=False)[1])
```

```
df = df_fair_12
sns.pointplot(x=df.occupation,
              y=df.affairs,
              join=False)


print('p-value:',
      sp.stats.ttest_ind([1, 2, 3, 4, 5, 6],
[1, 2, 3, 4, 5, 60],
equal_var=False)[1])
```

Two tails or one tail

Two tails





One tail

- "B is different from A."
  - Regardless of directions.
  - Use two tails.
- "B is greater than A."
  "B is lesser than A."
  - One of directions.
  - Use one tail.
  - If a p-value function doesn't support and the distribution is symmetrical, just /2 to obtain one-tailed p-value.

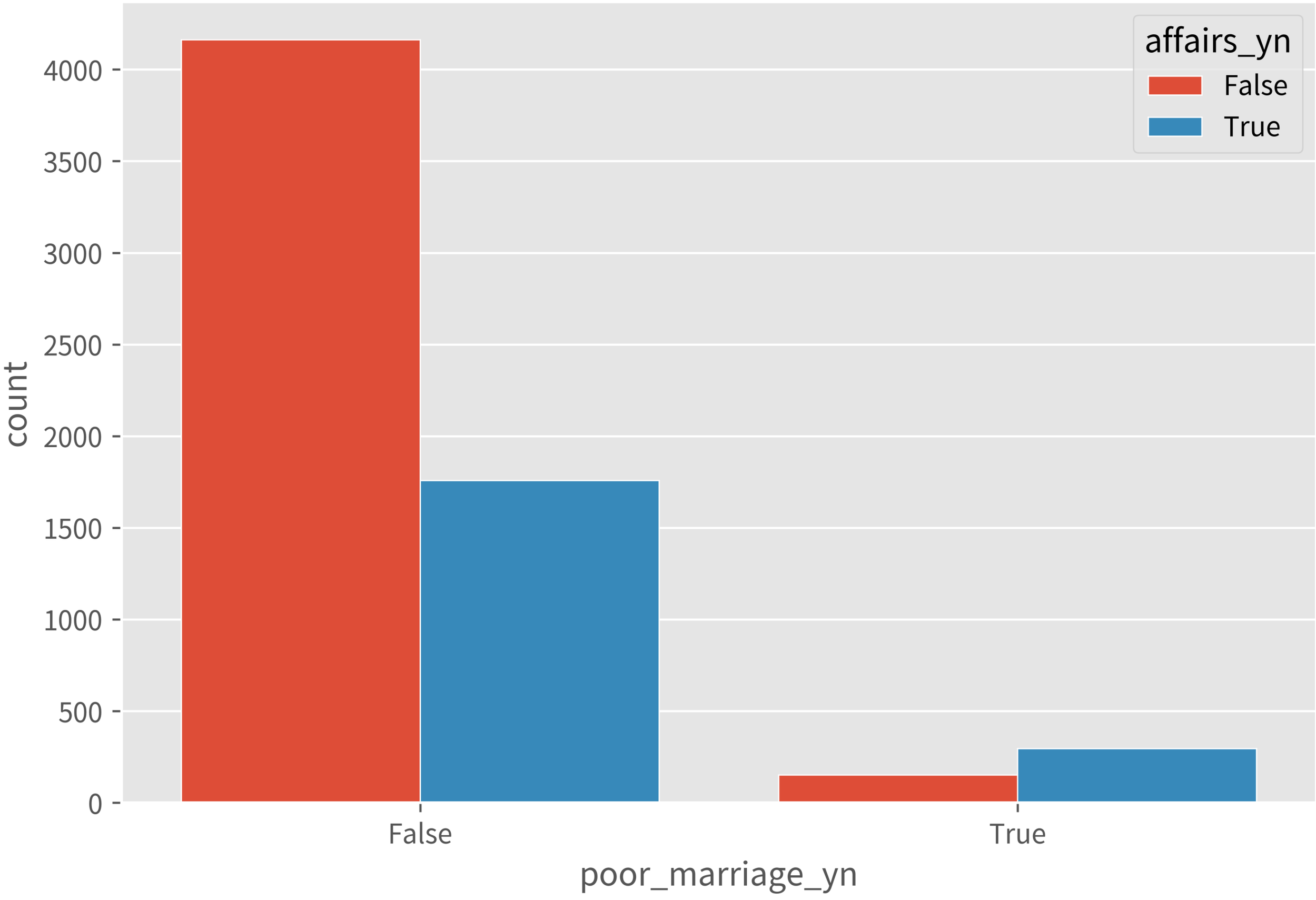
Chi-squared test #1

affairs_yn poor_marriage_yn	False	True
False	4161	1758
True	152	295

affairs_yn poor_marriage_yn	False	True
False	0.702990	0.297010
True	0.340045	0.659955

p-value: 1.9460298519537103e-56

- Preprocess:
  - Transform affairs > 0 as true.
  - Select the two occupations.
  - Group by occupations.
- Describe.
- Test:
  - Assume no difference, the probability to observe it: super low.
- So, we accept the proportions are different at 99.9% confidence level:
  - Non-poor: 30%; Poor: 66%


affairs_yn
False
True
count
4000
3500
3000
2500
2000
1500
1000
500
0
False
True
poor_marriage_yn

```
df = df_fair
# 2: poor
# 3: fair
df = df.assign(poor_marriage_yn
                  =(df.rate_marriage <= 2),
                  affairs_yn=(df.affairs > 0))
df_fair_21 = df
```

```
df = df_fair_21

df = (df
      .groupby(['poor_marriage_yn', 'affairs_yn'])
      [['affairs']]
      .count()
      .unstack()
      .droplevel(axis=1, level=0))
```

```
df_pct = df.apply(axis=1, func=lambda r: r/r.sum())

display(df, df_pct)

print('p-value:',
      sp.stats.chi2_contingency(
          df,
          correction=False
      )[1])
```

```
df = df_fair_21
sns.countplot(data=df,
                  x='poor_marriage_yn', hue='affairs_yn',
                  saturation=0.95, edgecolor='white')
```

Chi-squared test #2

affairs_yn	False	True
occupation		
2.0	607	252
3.0	1818	965

affairs_yn	False	True
occupation		
2.0	0.706636	0.293364
3.0	0.653252	0.346748

```
p-value: 0.0037369587127306517
```

- Preprocess:
  - Transform affairs > 0 as true.
  - Group by occupation.
- Describe.
- Test:
  - Assume no difference, the probability to observe it: 0.4%.
- So, we accept the proportions are different at 99.9% confidence level:
  - Farming-like: 29%
  - White-colloar: 35%


affairs_yn
False
True
count
1750
1500
1250
1000
750
500
250
0
2.0
3.0
occupation

```
df = df_fair_21
# 2: farming-like
# 3: white-colloar
df = df[df.occupation.isin([2, 3])]
df_fair_22 = df
```

```
df = df_fair_22

df = (df
      .groupby(['occupation', 'affairs_yn'])
      [['affairs']]
      .count()
      .unstack()
      .droplevel(axis=1, level=0))
```

```
df_pct = df.apply(axis=1, func=lambda r: r/r.sum())

display(df, df_pct)

print('p-value:',
      sp.stats.chi2_contingency(
          df,
          correction=False
      )[1])
```

```
df = df_fair_22
sns.countplot(data=df,
                x='occupation', hue='affairs_yn',
                saturation=0.95, edgecolor='white')


print('p-value:',
      sp.stats.chi2_contingency(
          [[607, 252],
           [1818, 965]],
          correction=False
)[1])
```

Power analysis

- $f(a, \textit{effect size}, \beta) = \textit{sample size}$
- Before collecting data,
  - Define a, *effect size*, β to calculate required *sample size*.
- After test,
  - If *p-value* $< a$, good to say there is a difference.
  - If *p-value* $> a$, or closes to a, may investigate the β.
- The a, *effect size*, β here are "to-achieve", not "observed".
- Two-proportion z-test ≡ 2×2 chi-squared test. [ref]
  - The power analysis of two-proportion z-test is much easier.

Confidence level, α, β, and power

	don't accept alter	accept alter
null	confidence level $= P(\textit{don't accept alter} \mid \textbf{\textit{null}})$	α $= P(\textit{accept alter} \mid \textbf{\textit{null}})$
alter = null + effect size	β $= P(\textit{don't accept alter} \mid \textbf{\textit{alter}})$	power $= P(\textit{accept alter} \mid \textbf{\textit{alter}})$

The mini cheat sheet

- If testing means, Welch's t-test.
- If testing medians, Mann–Whitney U test.
- If testing proportions, chi-squared test.

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** ($\equiv 2\times2$ 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:
  - Selecting Commonly Used Statistical Tests – Bates College
  - 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
  - Dependent t-test for paired samples – Student's t-test – Wikipedia

Complete steps

1. Decide what test.
2. Decide α, effect size, β to achieve.
3. Calculate sample size.
4. Still collect a sample as large as possible.
5. Test.
6. Investigate β if need.
7. Report fully, not only significant or not.

Keep learning

- Seeing Theory
- Statistics – SciPy Tutorial
- StatsModels
- Biological Statistics
- Research design

Recap

- The null hypothesis is for building a model.
- The p-value is:
  - Given null hypothesis, the probability to observe the data.
    - Not the probability of null hypothesis.
    - "How compatible the data and the model are."
- Power analysis to understand sample size or β.
- Welch's t-test, Mann–Whitney U test, and chi-squared test.
- Report fully, not only significant or not.
- Let's identify the true difference or noise efficiently! 📊

P-value & α

Theory

Seeing is believing

- $p\text{-}value = 0.0027\ (< 0.01)$
  - 👏👏👏
- $p\text{-}value = 0.0271\ (0.01–0.05)$
  - 👏❓👏❓❓❓
- $p\text{-}value = 0.2718\ (\geq 0.05)$
  - ❓❓❓❓❓❓
- *appendixes/theory_01_how_tests_work.ipynb*

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 = P(BD|AB) = 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

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

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

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, β, sample size

Theory

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.

Raw effect size, and β

- DSM5: The case for double standards – James Coplan, M.D.
  - The figures explain *a*, *raw effect size*, 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*

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

	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, raw effect size, β, 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.

Actual negative rate, inverse α, inverse β

Theory

Inverse α = 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

Inverse β = P(CD|AC) = C/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

- *appendixes/theory_02_complete_a_test.ipynb*
- *appendixes/theory_03_figures.ipynb*
- That's all. 📚