T-tests

Data Science Immersive

FLATIRON SCHOOL

Overview

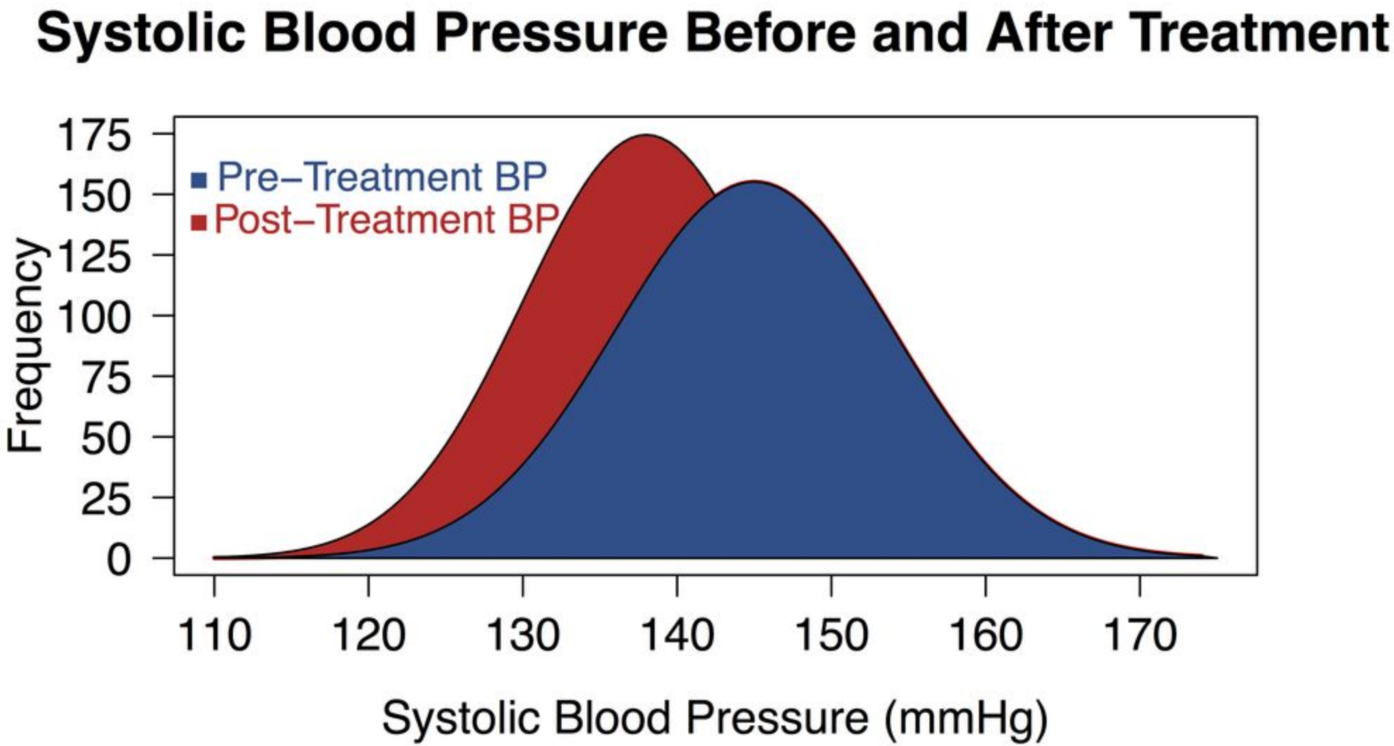
- Probability distribution versus sample distribution
- Touch on some other misconceptions
- Go over some examples and then code what we’ve learned


GUINNESS

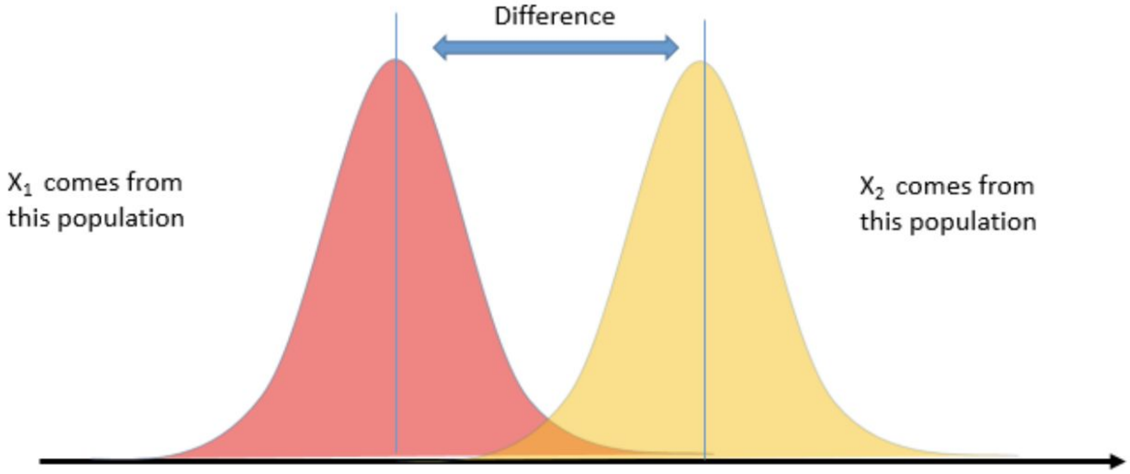
Examples

- Chemistry - do inputs from two different barley fields produce different yields?
- Astrophysics - do star systems with near-orbiting gas giants have hotter stars?
- Economics and social science - demography, surveys, etc.
- Medicine - BMI vs. Hypertension, etc.
- Business - which ad is more effective given engagement?

Problem


Systolic Blood Pressure Before and After Treatment
Pre–Treatment BP
Post–Treatment BP
Frequency
175
150
125
100
75
50
25
0
110
120
130
140
150
160
170
Systolic Blood Pressure (mmHg)

Problem



Difference
X_1 comes from this population
X_2 comes from this population

Hypothesis

- Are the means the same or different?

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

- Calculate test statistic which corresponds to a p-value
- Compare p to alpha

What’s a t-test

What’s different?

- Sometimes the population standard deviation is irrelevant, and sometimes it’s unknown. (we’ll get to the different types of t-test later)
- Sometimes a sample is too small to be confident that it’s an accurate representation of reality

A t-test is like a modified z-test:

- Penalize for small sample size - “degrees of freedom”
- Use sample std. dev. $\boldsymbol{s}$ to estimate population $\boldsymbol{\sigma}$

Assumptions

- underlying population is normally distributed
- Normally distributed (and similarly distributed) samples
- Controlled for selection bias

What’s a t-test

The z-distribution is skewed to fit the level of uncertainty of a t-test.

t curve approaches z curve as df -> ∞

t curve and *z curve*

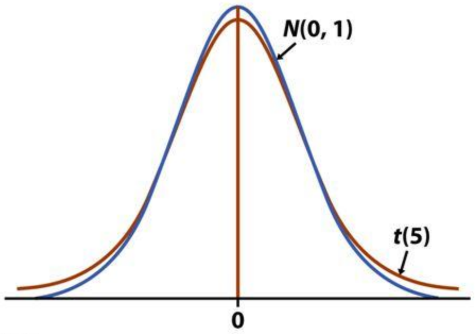


Figure 7-1
Introduction to the Practice of Statistics, Fifth Edition

Both the standard normal curve *N(0,1)* (the *z* distribution), and all *t(k)* distributions are density curves, symmetric about a mean of 0, but *t* distributions have more probability in the tails.

As the sample size increases, this decreases and the *t* distribution more closely approximates the *z* distribution. ***By n = 1000 they are virtually indistinguishable from one another.***

What's going on?

Z statistic

$$\boxed{z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}}$$

Unknown σ

t statistic

$$\boxed{t = \frac{\bar{X} - \mu}{s / \sqrt{n}}}$$

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

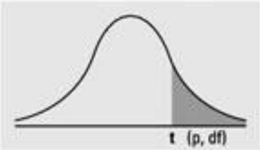
Low *v*

v approaches infinity

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Tables

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150

3 types of t-test

- Do we have a good sample?
  - **Single sample**
  - The One Sample *t* Test determines whether the sample mean is statistically different from a known or hypothesized population mean. (sample vs. population)
  - Note this is the only one where we need the population mean

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Has our sample changed?
  - **Paired data**
  - Two samples of related data points (same person, same wheat field, etc.)

$$t = \frac{m}{s/\sqrt{n}}$$

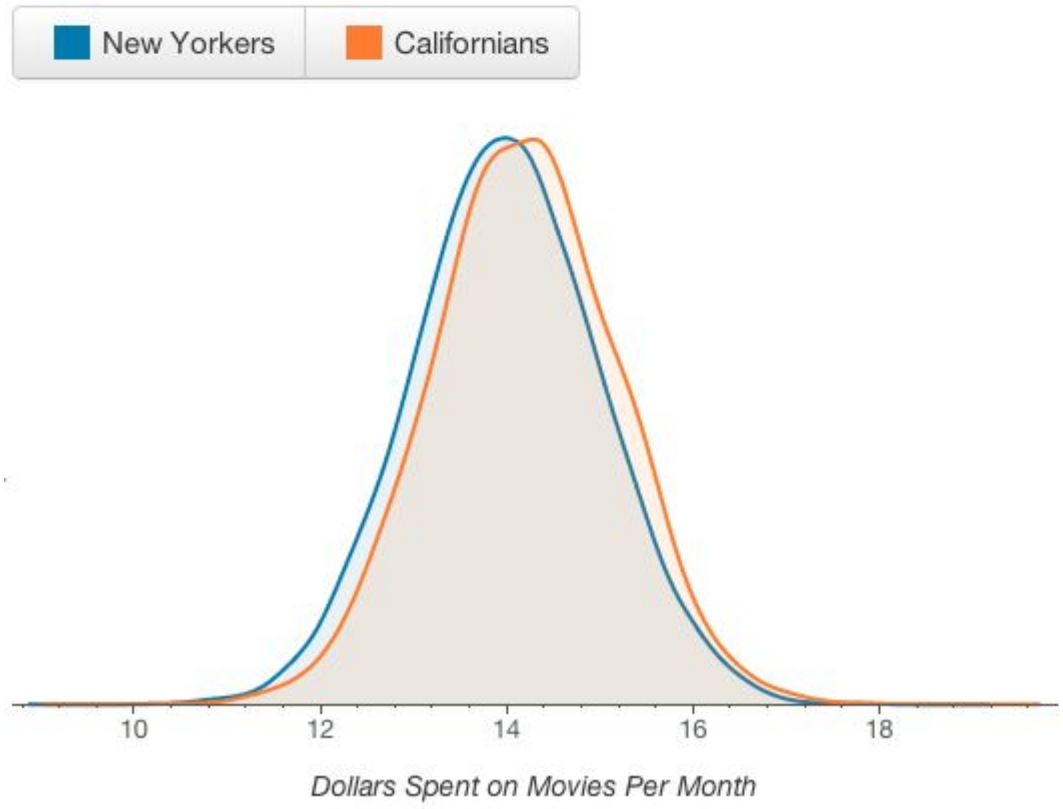
- Are two different samples different or alike?
  - **Independent samples**
  - Two samples of unrelated data points (not before/after, etc.) (different people, different wheat fields)

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

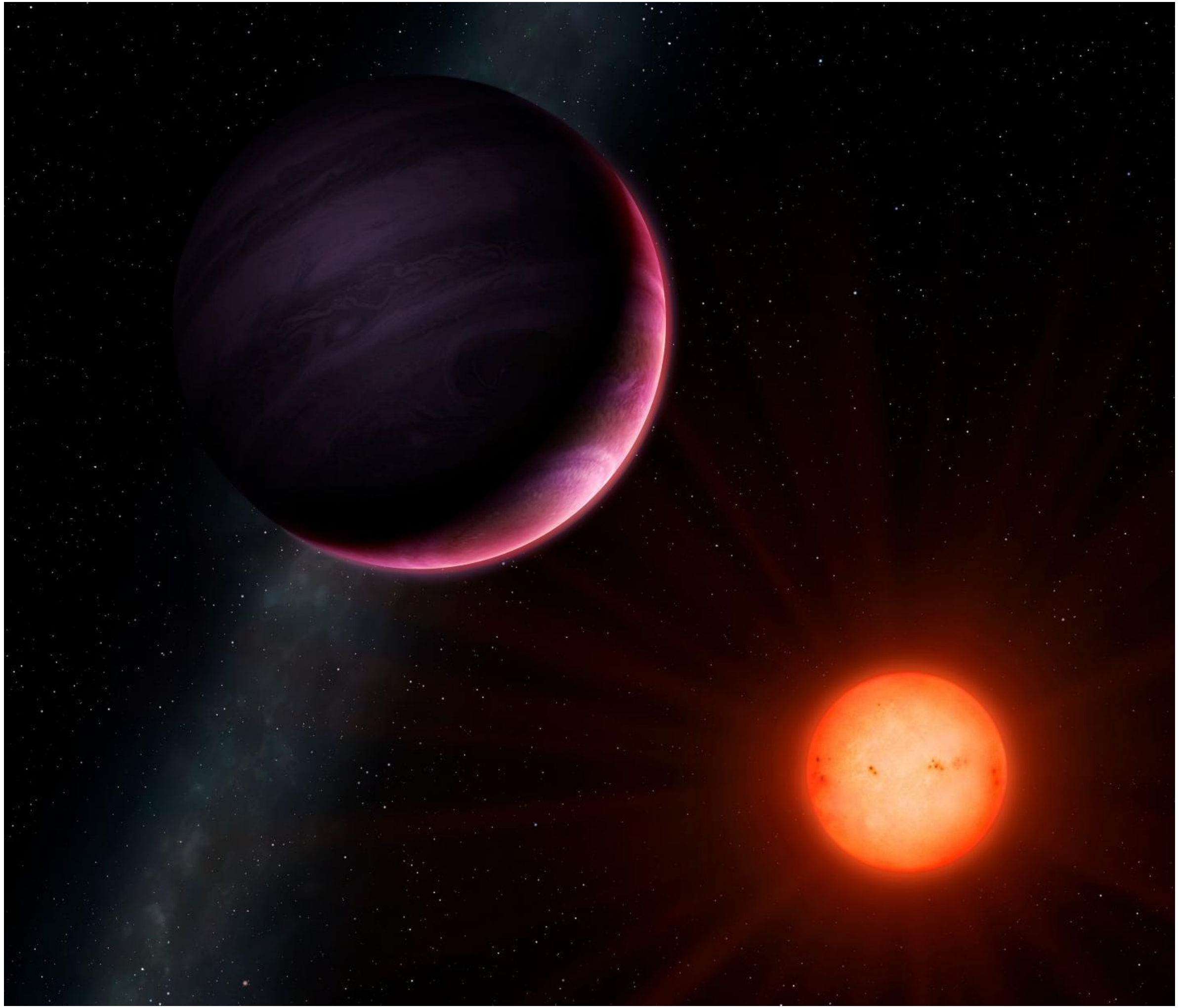
Practice

Effect Size


New Yorkers
Californians
10
12
14
16
18
Dollars Spent on Movies Per Month

Interpreting Results

- Reject or fail to reject null hypothesis
- Language is important!
- Don’t throw out failed experiments
  - This methodology, with this data, does not produce significant results
  - More data
  - More time
  - More details



Full Proof of t-distribution