15-388/688 - Practical Data Science: Hypothesis testing and experimental design

J. Zico Kolter
Carnegie Mellon University

Spring 2018

Outline

Motivation

Background: sample statistics and central limit theorem

Basic hypothesis testing

Experimental design

Outline

Motivation

Background: sample statistics and central limit theorem

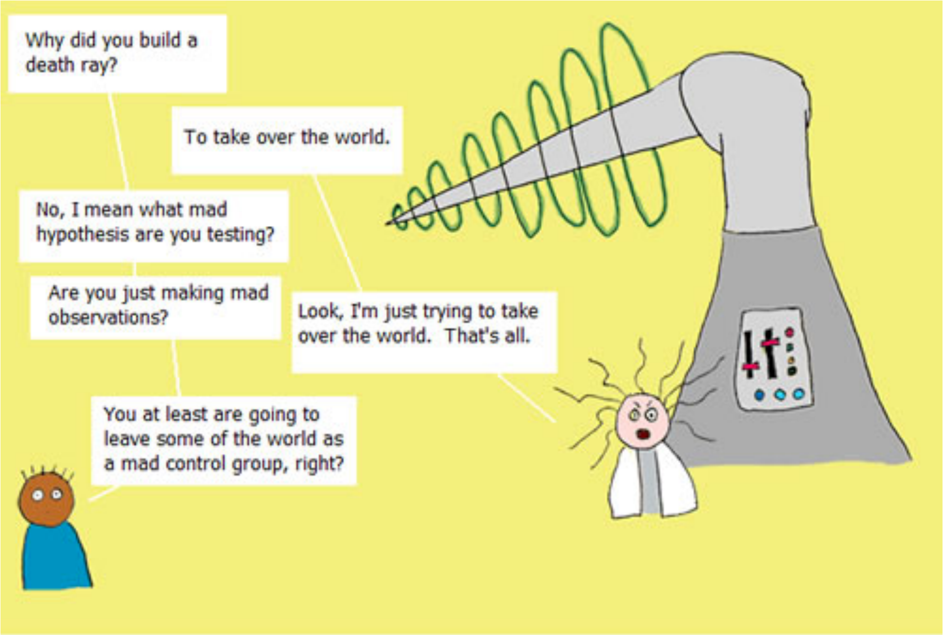
Basic hypothesis testing

Experimental design

Motivating setting

For a data science course, there has been very little “science” thus far…

“Science” as I’m using it roughly refers to “determining truth about the real world”


Why did you build a death ray?
To take over the world.
No, I mean what mad hypothesis are you testing?
Are you just making mad observations?
Look, I'm just trying to take over the world. That's all.
You at least are going to leave some of the world as a mad control group, right?

Sad truth: Most "mad scientists" are actually just mad engineers

Asking scientific questions

Suppose you work for a company that is considering a redesign of their website; does their new design (design B) offer any statistical advantage to their current design (design A)?

In linear regression, does a certain variable impact the response? (E.g. does energy consumption depend on whether or not a day is a weekday or weekend?)

In both settings, we are concerned with making actual statements about the nature of the world

Outline

Motivation

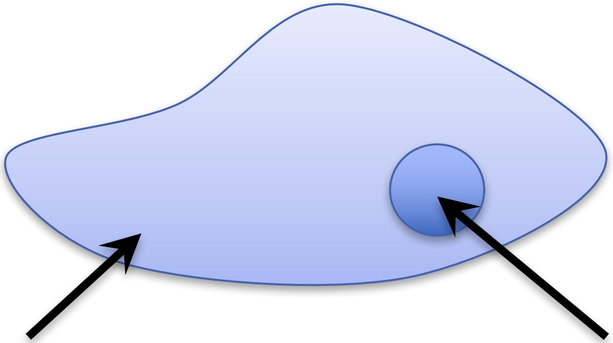
Background: sample statistics and central limit theorem

Basic hypothesis testing

Experimental design

Sample statistics

To be a bit more consistent with standard statistics notation, we'll introduce the notion of a *population* and a *sample*



	Population	**Sample**
Mean	$\mu = \mathbf{E}[X]$	$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
Variance	$\sigma = \mathbf{E}[(X-\mu)^2]$	$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x^{(i)} - \bar{x})^2$

Sample mean as random variable

The same mean is an empirical average over m independent samples from the distribution; it can also be considered as a random variable

This new random variable has the mean and variance

$$\mathbf{E}[\bar{x}] = \mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m} \mathbf{E}[X] = \mathbf{E}[X] = \mu$$

$$\mathbf{Var}[\bar{x}] = \mathbf{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] = \frac{1}{m^2}\sum_{i=1}^{m} \mathbf{Var}[X] = \frac{\sigma^2}{m}$$

where we used the fact that for *independent* random variables X_1, X_2

$$\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2]$$

When estimating variance of sample, we use s^2/m (the square root of this term is called the **standard error**)

Central limit theorem

Central limit theorem states further that $\bar{x}$ (for “reasonably sized” samples, in practice $m \geq 30$) actually has a Gaussian distribution *regardless of the distribution of* X

$$\bar{x} \rightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right) \text{ (or equivalently) } \frac{\bar{x} - \mu}{\sigma / m^{1/2}} \rightarrow \mathcal{N}(0,1)$$

In practice, for $m < 30$ and for estimating σ^2 using sample variance, we use a Student’s t-distribution with $m - 1$ degrees of freedom

$$\frac{\bar{x} - \mu}{s / m^{1/2}} \rightarrow T_{m-1}, \qquad p(x;\nu) \propto \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Aside: why the $m-1$ scaling?

We scale the sample variance by $m-1$ so that it is an *unbiased* estimate of the population variance

$$
\begin{aligned}
\mathbf{E}\left[\sum_{i=1}^{m}(x^{(i)}-\bar{x})^2\right] &= \mathbf{E}\left[\sum_{i=1}^{m}\left((x^{(i)}-\mu)-(\bar{x}-\mu)\right)^2\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{m}(x^{(i)}-\mu)^2 - 2(\bar{x}-\mu)\sum_{i=1}^{m}(x^{(i)}-\mu) + m(\bar{x}-\mu)^2\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{m}(x^{(i)}-\mu)^2\right] - m\mathbf{E}\left[\sum_{i=1}^{m}(\bar{x}-\mu)^2\right] \\
&= m\mathbf{Var}[X] - \frac{m\mathbf{Var}[X]}{m} = (m-1)\sigma^2
\end{aligned}
$$

Outline

Motivation

Background: sample statistics and central limit theorem

Basic hypothesis testing

Experimental design

Hypothesis testing

Using these basic statistical techniques, we can devise some tests to determine whether certain data gives evidence that some effect “really” occurs in the real world

Fundamentally, this is evaluating whether things are (likely to be) true about the population (all the data) given a sample

Lots of caveats about the precise meaning of these terms, to the point that many people debate the usefulness of hypothesis testing at all

But, still incredibly common in practice, and important to understand

Hypothesis testing basics

Posit a *null hypothesis* H_0 and an alternative hypothesis H_1 (usually just that "H_0 is not true"

Given some data x, we want to accept or reject the null hypothesis in favor of the alternative hypothesis

	H_0 **true**	H_1 **true**
Accept H_0	Correct	Type II error (false negative)
Reject H_0	Type I error (false positive)	Correct

$p(\text{reject } H_0 | H_0 \text{ true})$ = "significance of test"

$p(\text{reject } H_0 | H_1 \text{ true})$ = "power of test"

Basic approach to hypothesis testing

Basic approach: compute the probability of observing the data *under the null hypothesis* (this is the p-value of the statistical test)

$$p = p(\mathrm{data}|H_0 \text{ is true})$$

Reject the null hypothesis if the p-value is below the desired significance level (alternatively, just report the p-value itself, which is the lowest significance level we could use to reject hypothesis)

Important: p-value is $p(\mathrm{data}|H_0 \text{ is true})$ not $p(H_0 \text{ is true } |\mathrm{data})$

Poll: p-value hacking

Suppose you adopt the following procedure. You test 100 patients to see if a drug has a statistically significant effect. If so, you stop the test and publish your current p-value. If not, you collect 100 additional patients, test the drug again, and publish that p-value (statistically significant or not). Is this a valid experimental design?

1. Yes

2. No

3. Depends on what p value you achieve

Canonical example: t-test

Given a sample $x^{(1)}, \dots, x^{(m)} \in \mathbb{R}$

$$H_0\colon \mu = 0 \ \text{(for population)}$$
$$H_1\colon \mu \neq 0$$

By central limit theorem, we know that $(\bar{x} - \mu)/(s/m^{\frac{1}{2}}) \sim T_{m-1}$
(Student's t-distribution with $m - 1$ degrees of freedom)

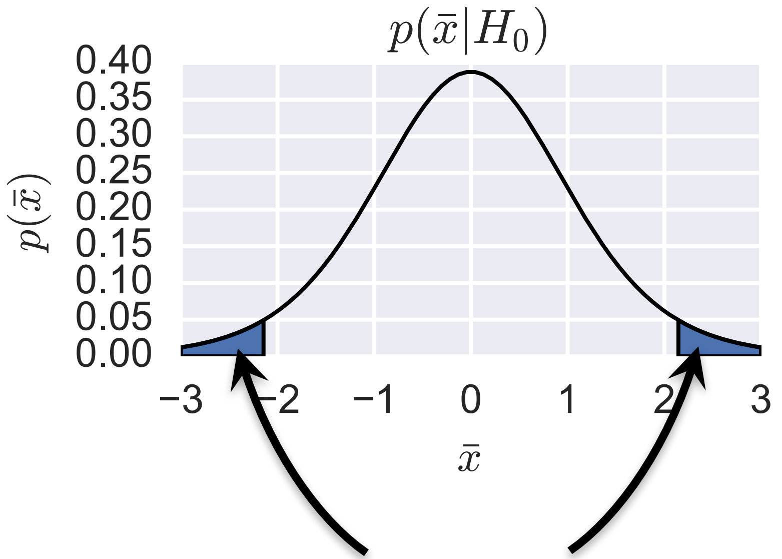
So we just compute $t = \bar{x}/\left(s/m^{\frac{1}{2}}\right)$ (called *test statistic*), then compute

$$p = p(x > |t|) + p(x < -|t|) = F(-|t|) + 1 - F(|t|) = 2F(-|t|)$$

(where F is cumulative distribution function of Student's t-distribution)

Visual example

What we are doing fundamentally is modeling the distribution $p(\bar{x}|H_0)$ and then determining the probability of the observed $\bar{x}$ or a more extreme value



$p = \text{Area}$

Code in Python

Compute t statistic and p value from data

```
import numpy as np
import scipy.stats as st
x = np.random.randn(m)

# compute t statistic and p value
xbar = np.mean(x)
s2 = np.sum((x - xbar)**2)/(m-1)
std_err = np.sqrt(s2/m)
t = xbar/std_err

t_dist = st.t(m-1)
p = 2*td.cdf(-np.abs(t))

# with scipy alone
t,p = st.ttest_1samp(x, 0)
```

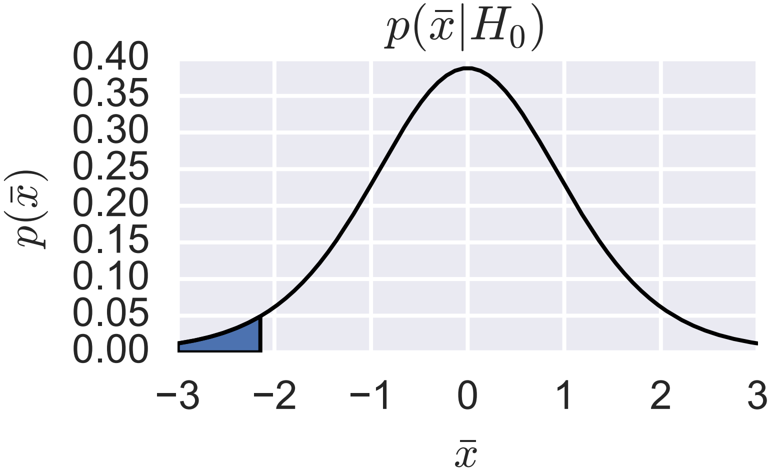
Two-sided vs. one-sided tests

The previous test considered deviation from the null hypothesis in both directions (two-sided test), also possible to consider a one-sided test

$$H_0\colon \mu \geq 0 \text{ (for population)}$$
$$H_1\colon \mu < 0$$

Same t statistic as before, but we only compute the area under the left side of the curve

$$p = p(x < t) = F(t)$$



Confidence intervals

We can also use the t statistic to create *confidence intervals* for the mean

Because $\bar{x}$ has mean μ and variance s^2/m, we know that $1-\alpha$ of its probability mass must lie within the range

$$\bar{x} = \mu \pm \frac{s}{m^{1/2}} \cdot F^{-1}\left(1 - \frac{\alpha}{2}\right) \equiv \mu + CI(s,m,\alpha)$$
$$\Longleftrightarrow \mu = \bar{x} \pm CI(s,m,\alpha)$$

where F^{-1} denotes the inverse CDF function of t-distribution with $m-1$ degrees of freedom

```
# simple confidence interval compuation
CI = lambda s,m,a : s / np.sqrt(m) * st.t(m-1).ppf(1-a/2)
```

Outline

Motivation

Background: sample statistics and central limit theorem

Basic hypothesis testing

Experimental design

Experimental design: A/B testing

Up until now, we have assumed that the null hypothesis is given by some *known* mean, but in reality, we may not know the mean that we want to compare to

Example: we want to tell if some additional feature on our website makes user stay longer, so we need to estimate both how long users stay on the current site and how long they stay on redesigned site

Standard approach is A/B testing: create a *control group* (mean μ_1) and a *treatment group* (mean μ_2)

$$H_0 : \mu_1 = \mu_2 \text{ (or e.g. } \mu_1 \geq \mu_2)$$
$$H_1 : \mu_1 \neq \mu_2 \text{ (or e.g. } \mu_1 < \mu_2)$$

Independent t-test (Welch's t-test)

Collect samples (possibly different numbers) from *both* populations

$$x_1^{(1)}, \ldots, x_1^{(m_1)}, \qquad x_2^{(1)}, \ldots, x_2^{(m_2)}$$

compute sample mean $\bar{x}_1, \bar{x}_2$ and sample variance s_1^2, s_2^2 for each group

Compute test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{(s_1^2/m_1 + s_2^2/m_2)^{1/2}}$$

And evaluate using a t distribution with degrees of freedom given by

$$\frac{(s_1^2/m_1 + s_2^2/m_2)^2}{\frac{(s_1^2/m_1)^2}{m_1 - 1} + \frac{(s_2^2/m_2)^2}{m_2 - 1}}$$

Starting seem a bit ad-hoc?

There are a huge number of different tests for different situations

You probably won't need to remember these, and can just look up whatever test is most appropriate for your given situation

But the basic idea in call cases is the same: you're trying to find the distribution of your test statistic under the hull hypothesis, and then you are computing the probability of the observed test statistic or something more extreme

All the different tests are really just about different distributions based upon your problem setup

Hypothesis testing in linear regression

One last example (because it's useful in practice): consider the linear regression $y \approx \theta^T x$, and suppose we want to perform a hypothesis test on the coefficients of θ

Example: suppose that instead of just two website, you have a website with multiple features that can be turned on/off, and your sample data includes a wide variety of different samples

We would like to ask the question: is the ith variable relevant for predicting the output?

We've already seen ways we can do this (i.e., evaluate cross-validation error, but it's a bit difficult to understand what this means)

Formula for sample variance in linear regression

There is an analogous formula for sample variance on the *errors* that a linear regression model makes

$$s^2 = \frac{1}{m-n}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2$$

Use this to determine sample covariance of *coefficients*

$$\mathbf{Cov}[\theta] = s^2(X^T X)^{-1}$$

Can then evaluate null hypothesis $H_0 \colon \theta_i = 0$, using t statistic

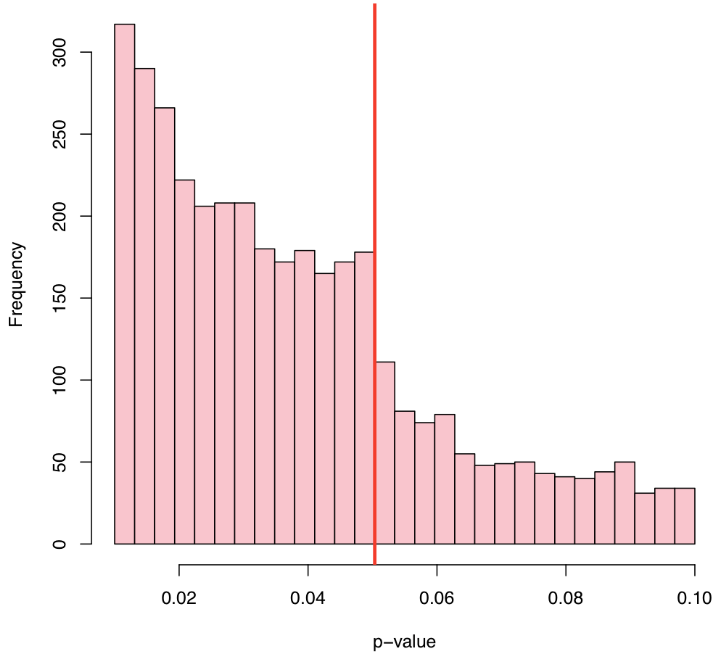
$$t = \theta_i / \mathbf{Cov}[\theta]_{i,i}^{1/2}$$

Similar procedure to get confidence intervals of coefficients

P-values considered harmful

A basic problem is that $p(\mathrm{data}|H_0) \neq p(H_0|\mathrm{data})$ (despite being frequently interpreted as such)

People treat $p < 0.05$ with *way* too much importance



Histogram of p values from ~3,500 published journal papers
(from E. J. Masicampo and Daniel Lalande, *A peculiar prevalence of p values just below .05*, 2012)