Data Science Methods for Clean Energy Research

Week 4, Lecture 1: Hypothesis testing and p-values

January 23, 2017

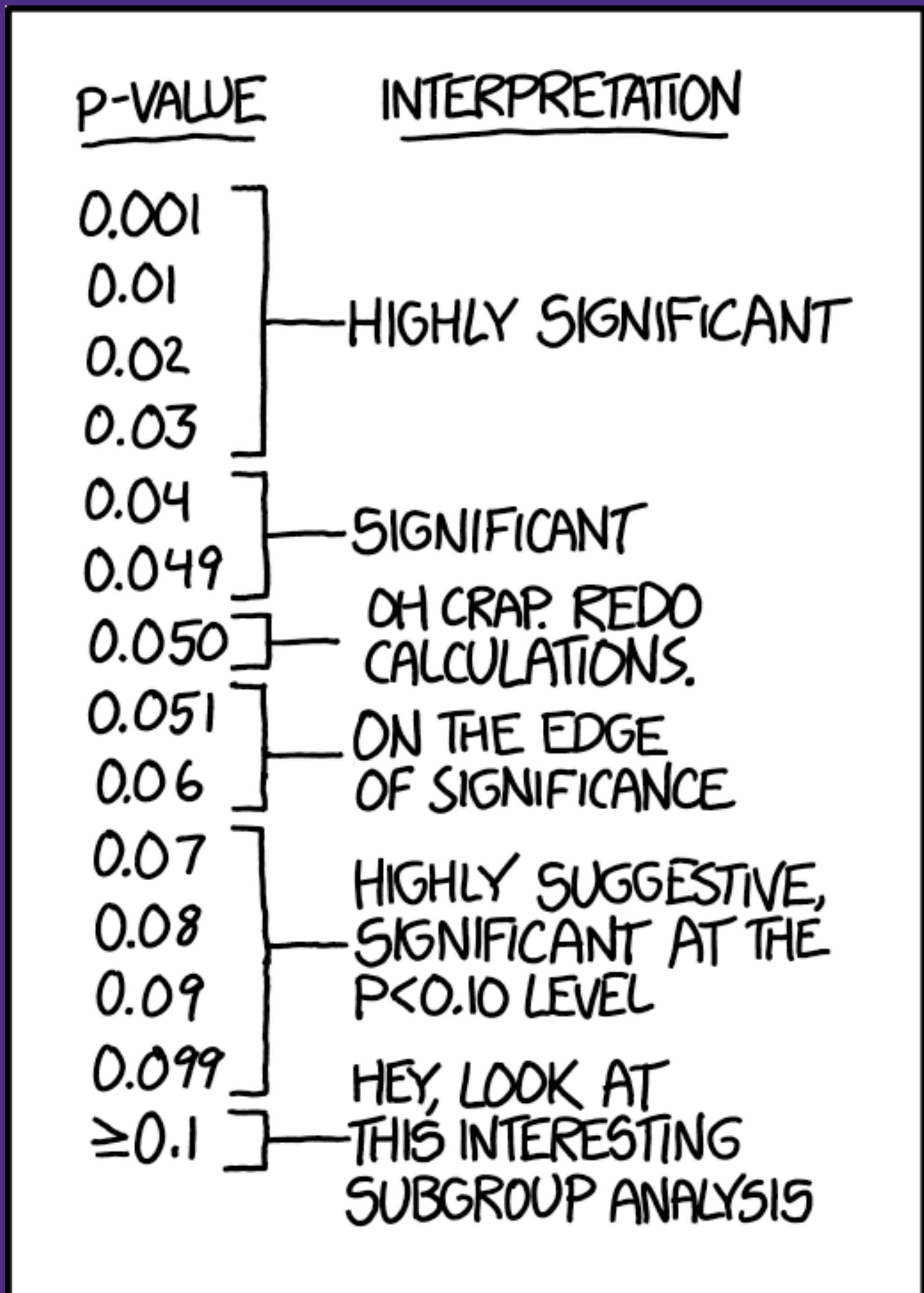
Please start a new Jupyter notebook and load the same software stack we have been using.

We will use later in the lecture

No file download

UNIVERSITY of WASHINGTON

P values via xkcd



Outline

- Reading assignment for W4
  - Please do this
- Quick review
- Warmup
- One sample t-test
  - Theory
  - Practice
- Nuzzo, the ASA, and the onslaught against p-values

Figures today mostly drawn from Krzywinski & Altman nature papers I gave you last week… or xkcd

Key concepts from last time

- Histogram
- PDFs
- Central limit theorem

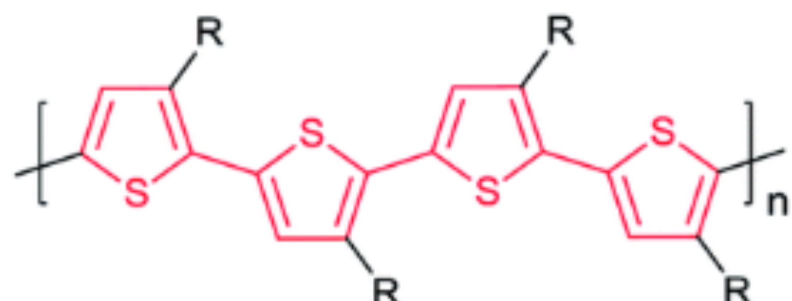
Warm up – 5 minutes discussion

> **Choose someone at the table who is not wearing jeans – they are the** facilitator**.**
  - **Tie break: whoever has birthday closest to today**

> **Each person**
  1. **Give an example of data you have collected (either in research or class)**
  2. **What did you do with the data? (e.g., error analysis, graphing, p-value calculation)**
  3. **Did you make any scientific conclusions about the data?**
  4. **What statistical methods** (if any) **were the conclusions based on?**

The concept of a hypothesis test in statistics

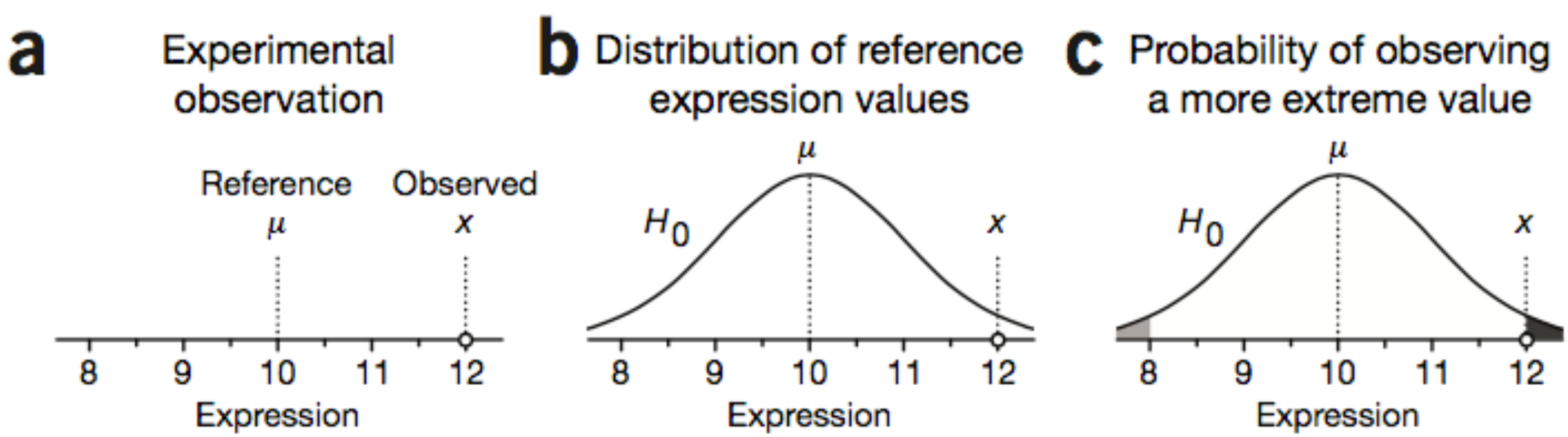
- A scientific hypothesis: **Adding additional hydrophobic groups to conjugated polymers will accelerate their crystallization rate in polar solvents**
- A scientific experiment to test the hypothesis: **We will synthesize P3AT of various –R length, and measure crystallinity as a function of time in different solvents and compare to consensus values from literature**
- A statistical hypothesis: **There is a 95% probability that the measured crystallization rates (with varying –R) are drawn from different distributions than the consensus literature values**



Polythiophene

The concept of a hypothesis test in statistics

> **Panel a illustrates the situation: we want to know if a reference value is difference from the observation**
> **Panel b proposes there is some underlying distribution around the reference (**assume normal for now**)**
> **Panel c illustrates the hypothesis test:** what is the probability the data at least as extreme as measured?
  – ***<u>The p-value is the shaded area under the curve, e.g., p=.05</u>***



The concept of a hypothesis test in statistics

> **Panel a illustrates the situation: we want to know if a reference value is difference from the observation**

> **Panel b proposes there is some underlying distribution around the reference (**assume normal for now**)**

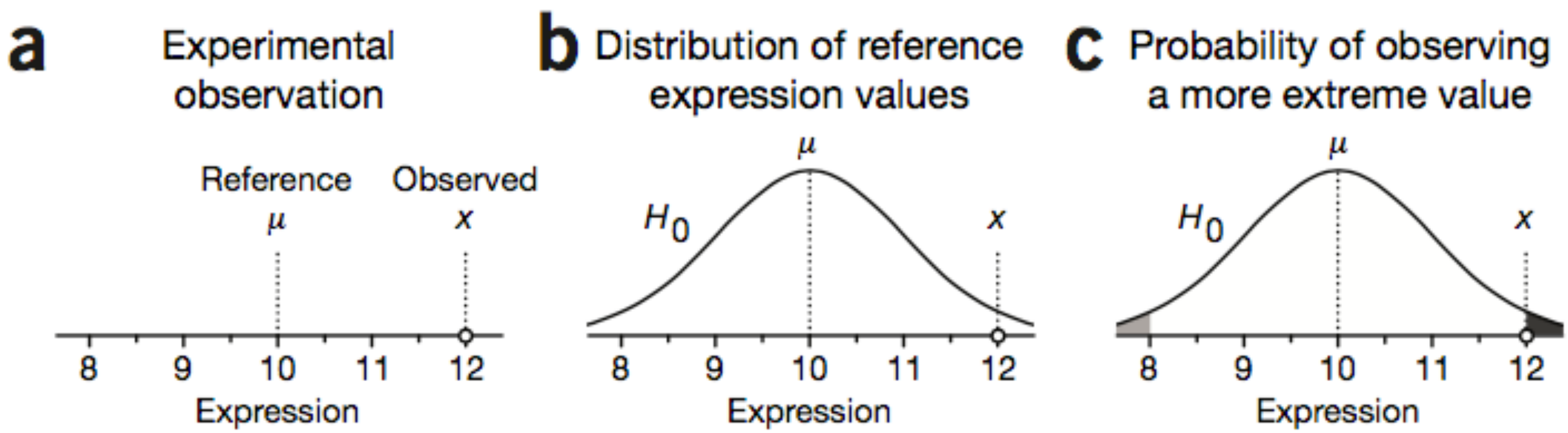
> **Panel c illustrates the hypothesis test:** what is the probability the data at least as extreme as measured?

   – ***<u>The p-value is the shaded area under the curve, e.g., p=.05</u>***

"Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." – From the American Statistical Association statement on p-values

The null hypothesis

- **H_0 is the** null **hypothesis, it states that the observations (measurements) are drawn from the same distribution as the the reference. It is characterized by a null distribution (also H_0)**
- **You will find many statistics texts that contrast the null with the so-called** alternate hypothesis **(H_A) which may state something about the likelihood that the data are "non random". These are falling out of favor.**
- **Presumably because of the way the this shapes your thinking about what the data are showing you (**more on this soon**)**



Short interlude – the CLT and SEM

- **Before we explain how to characterize H_0 (**the distribution**) and test its significance (**the hypothesis**) we have add one more piece of information about what happens when we sample means and how to characterize the error of the means.**

Sampling from distributions and the central limit theorem

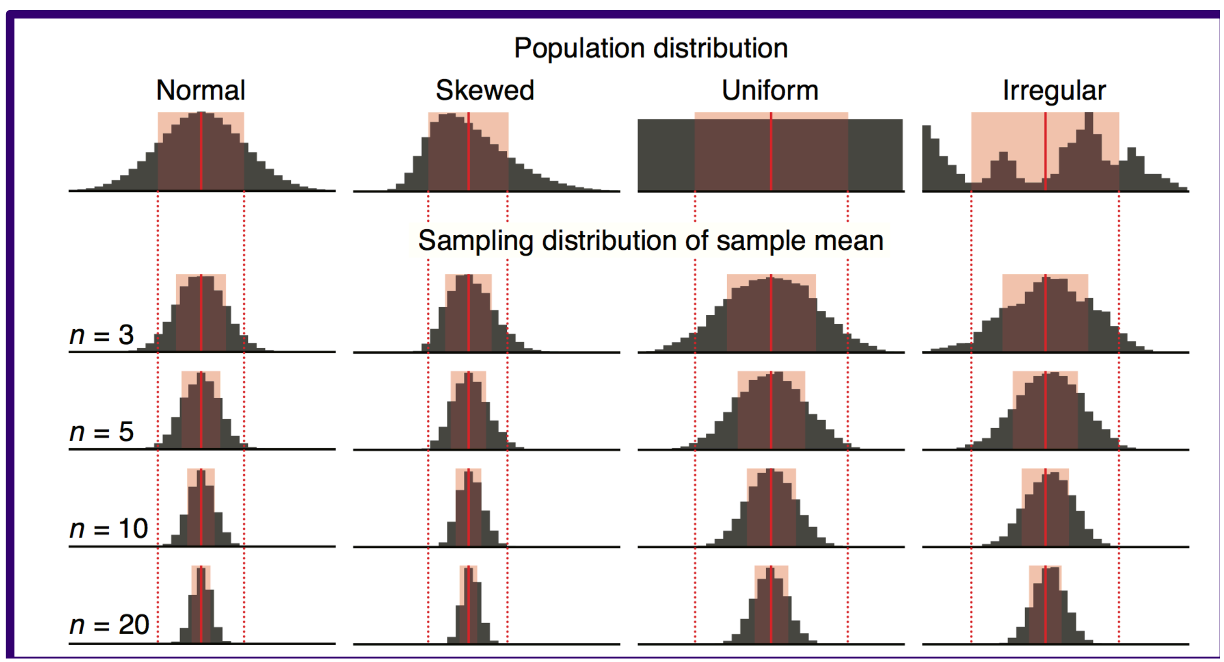
> **Recall from the CLT that the sampling distribution of the** sample mean **will always drift towards a normal distribution**

> **We can define the standard error of the mean** (SEM) **as a descriptor of this distribution:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$SEM = \frac{s_x}{\sqrt{n}}$$



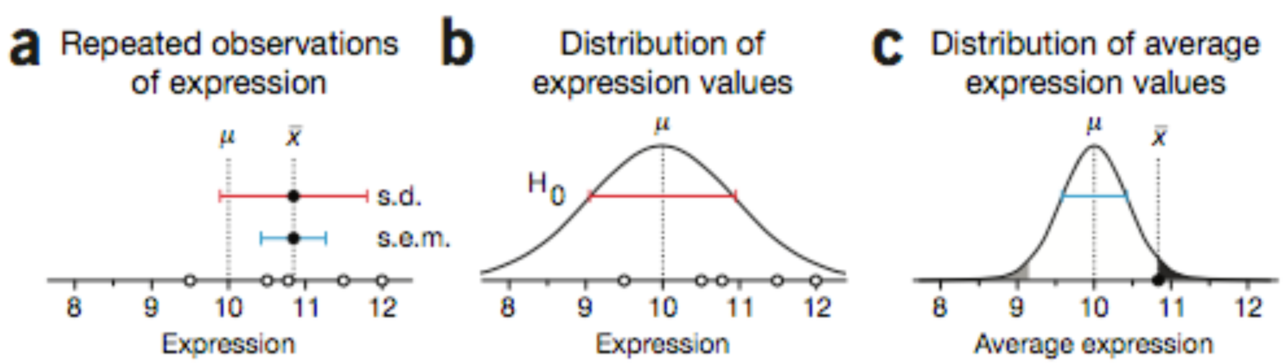
Sampling from distributions and the central limit theorem

- **Panel a shows the calculation of the sample mean, standard deviation, and standard error**
- **Panel b highlights a major assumption:** the variance of your sample is the same as the variance of the null distribution H_0
- **Panel c shows how the p-value is actually estimated:** Assuming the s.e.m. of your sample is distributed about your reference value

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

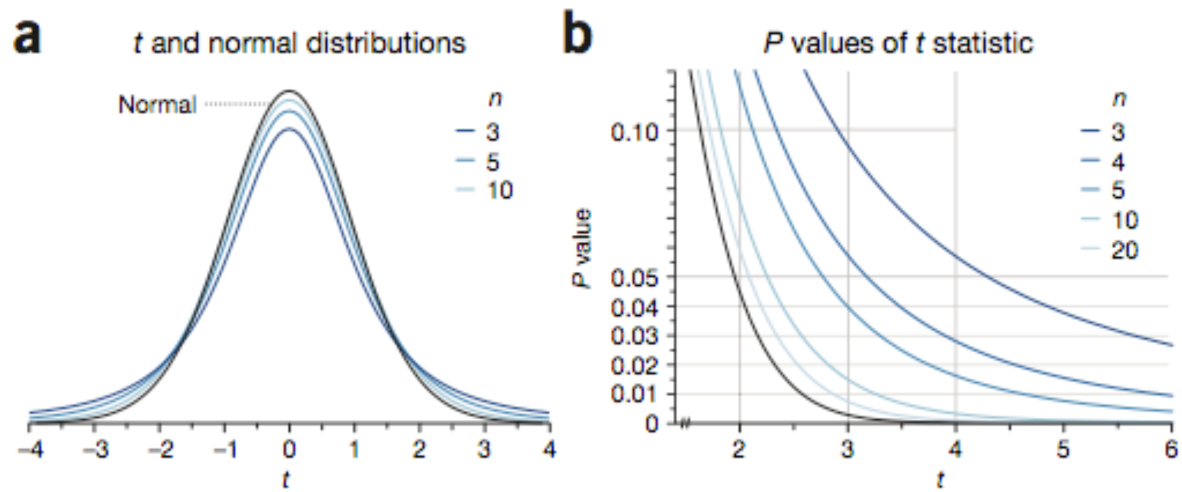
$$s.d. = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$s.e.m = \frac{s_x}{\sqrt{n}}$$



Just one more detail: The sampling distribution

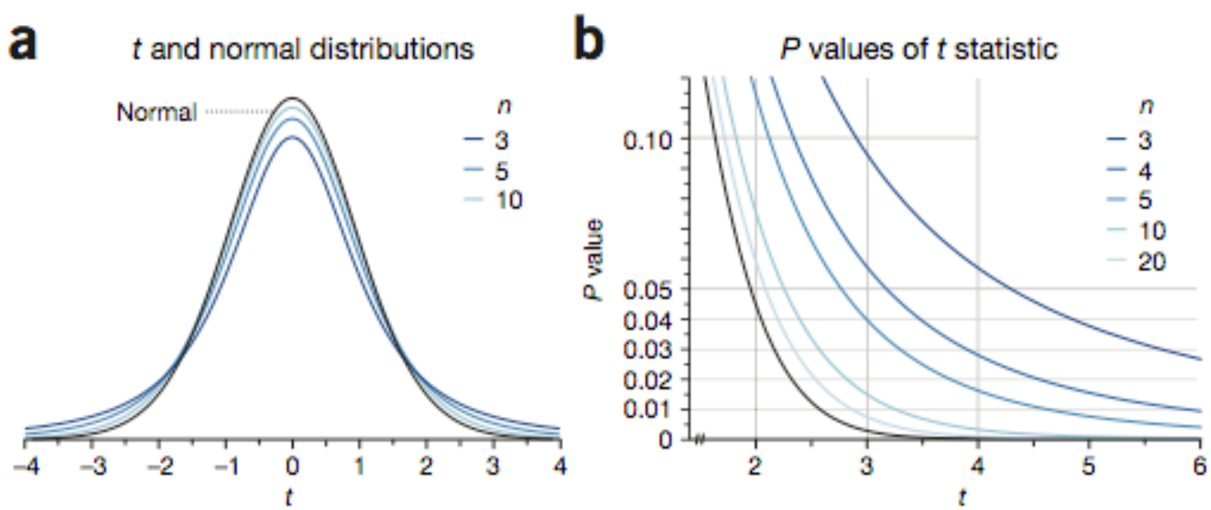
- **In general, it is known that the sample variance is an** underestimate **of the true distribution variance.**
- **There is a distribution closely related to the normal called** "the t-distribution" **or** "student's t-distribution"
- **The t-distribution gives you an easy recipe for calculating p-values with small sample sizes**



The recipe for a 1 sample t–test

- **State your hypothesis (H_0) and collect data**
- **Determine your desired** significance level **(usually called α). We usually use P=0.05**
- **Calculate so-called t-statistic**
- **Use a lookup table (or function) to determine P**
- **If $P < P_\alpha$, you may reject H_0**

$$\boxed{t = \frac{\bar{x} - \mu}{s.e.m} = \frac{\bar{x} - \mu}{s_x / \sqrt{n}}}$$



Practice

- Please open a python notebook – lets "capacity build!"
- Our goal will be to reproduce experiments drawn from this distribution and test their significance level