Data Science Methods for Clean Energy Research

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

January 23, 2017

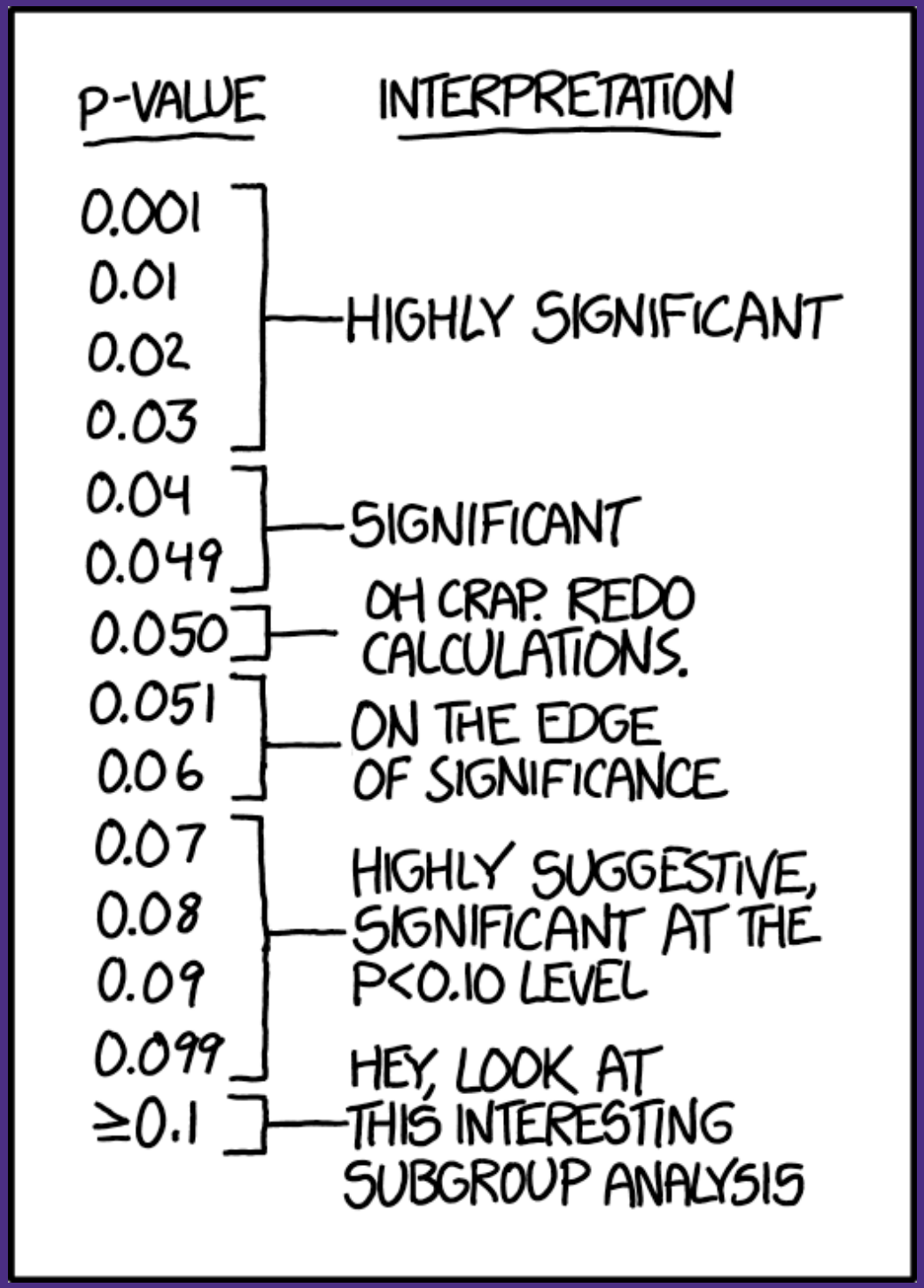
Please start a new Jupyter notebook and load the same software stack we have been using.

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

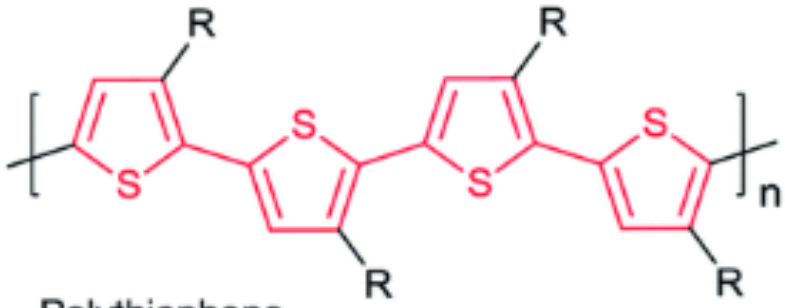
- Histogram
- PDFs
- Central limit theorem

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

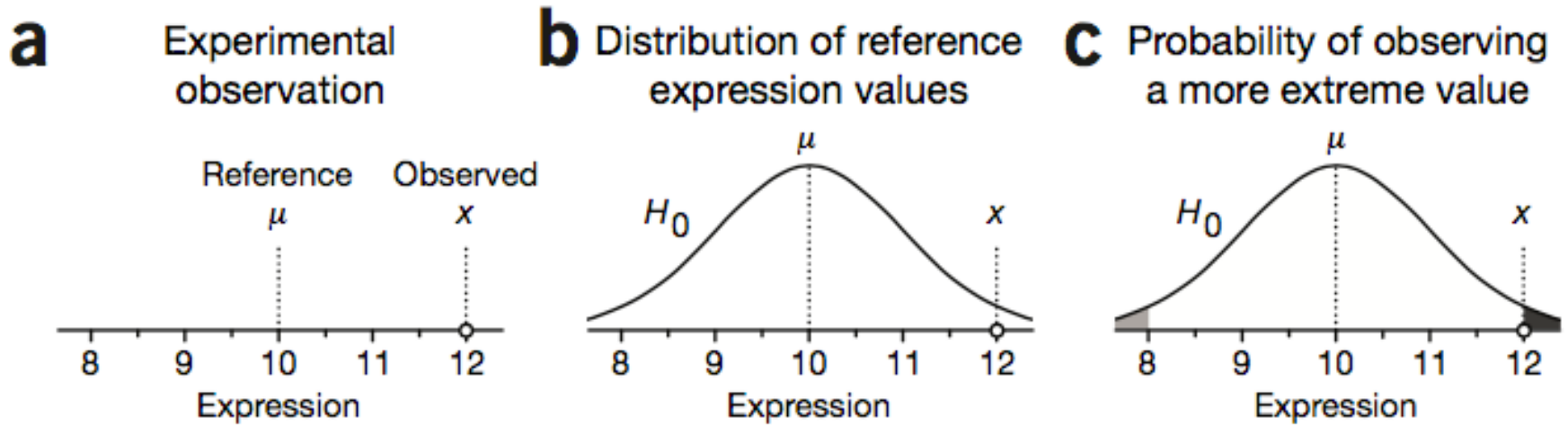
- A scientific hypothesis: **Adding additional hydrophobic groups to conjugated polymers will accelerate their crystallization rate in polar solvents**
- A scientific experiment to test the hypothesis: **We will synthesize P3AT of various –R length, and measure crystallinity as a function of time in 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?
  - ***The p-value is the shaded area under the curve, e.g., p=.05***



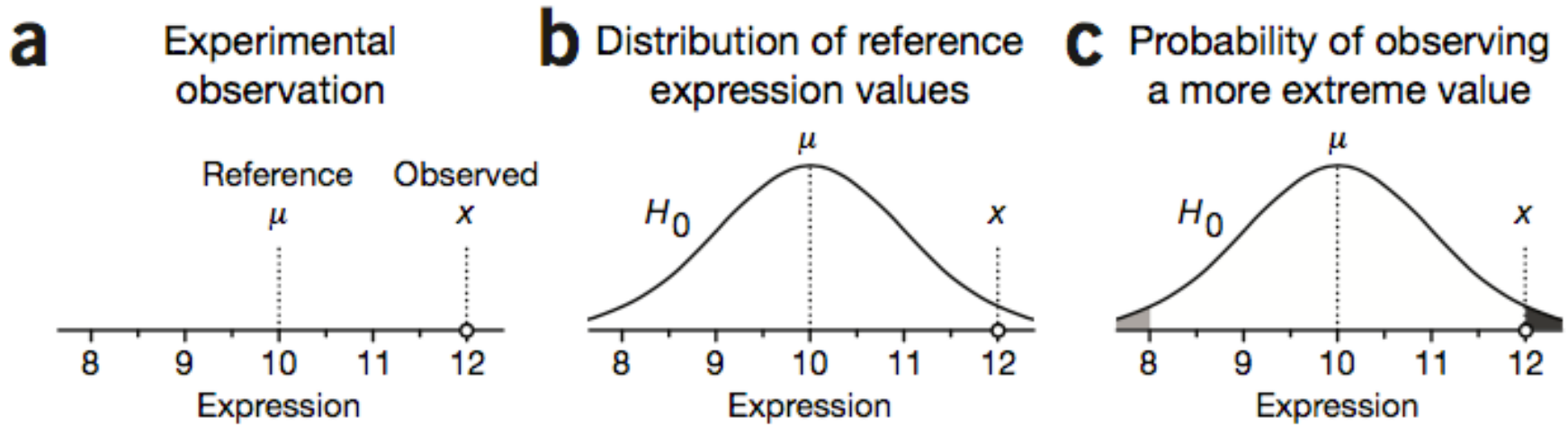
The concept of a hypothesis test in statistics

- **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?
  - ***The p-value is the shaded area under the curve, e.g., p=.05***

"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.**

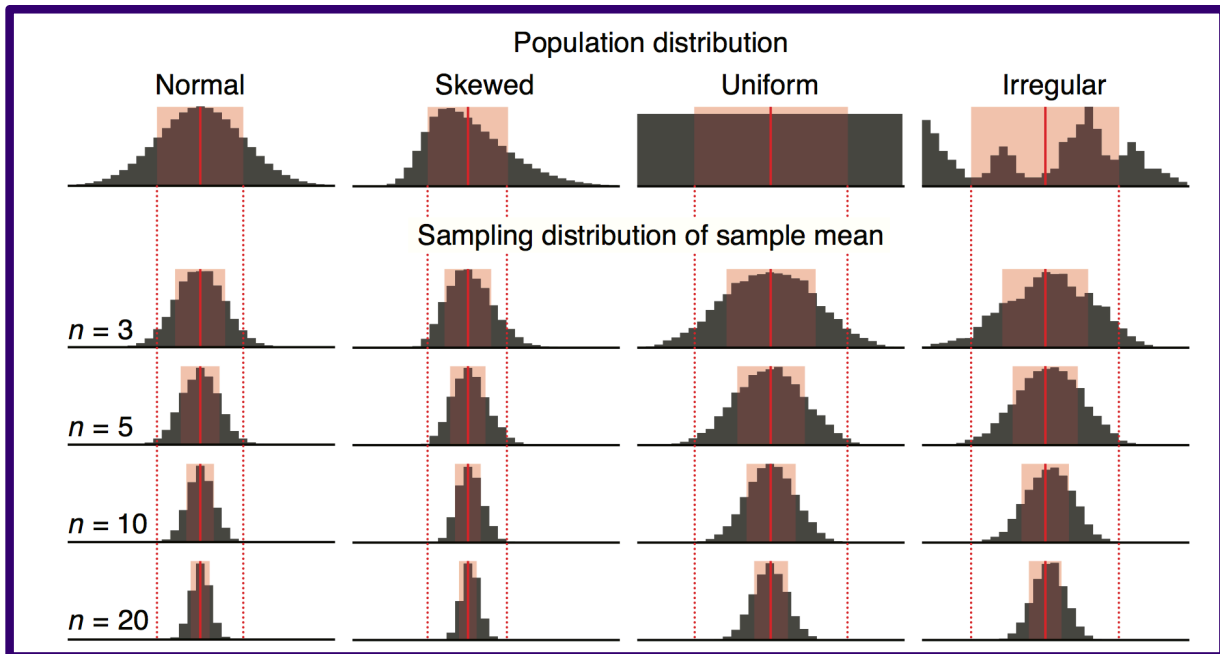
Sampling from distributions and the central limit theorem

- 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:

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

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

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



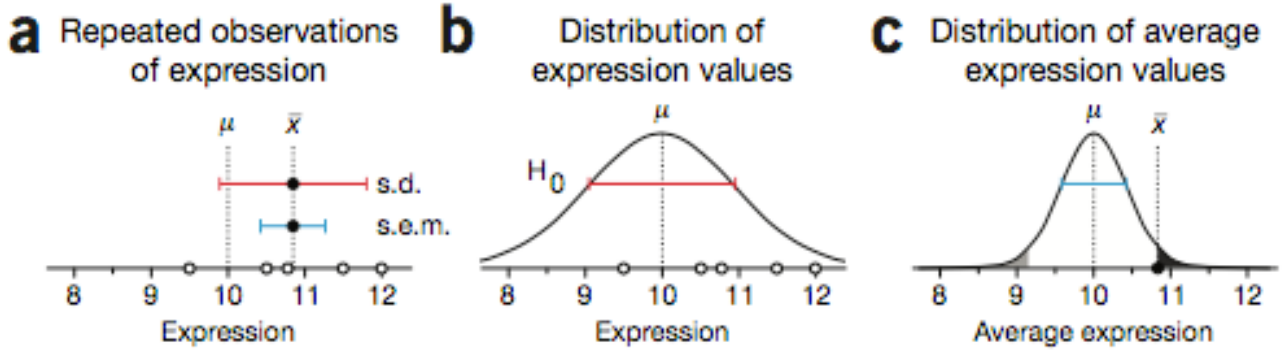
Sampling from distributions and the central limit theorem

- **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

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

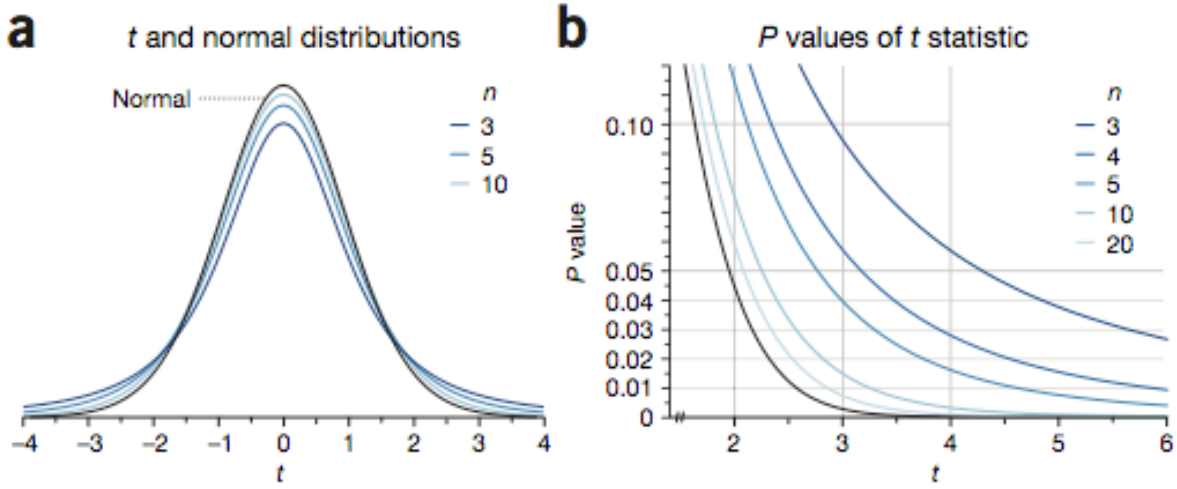
$$s.d. = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

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



Just one more detail: The sampling distribution

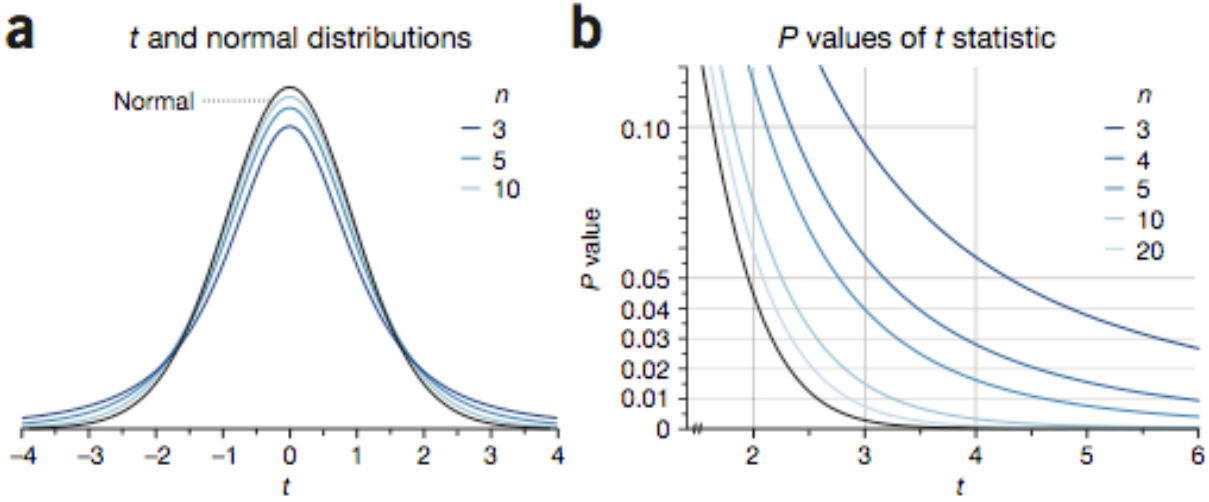
- **In general, it is known that the sample variance is an** underestimate **of the true distribution variance.**
- **There is a distribution closely related to the normal called "**the t-distribution**" or "**student's t-distribution**"**
- **The t-distribution gives you an easy recipe for calculating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



The recipe for a 1 sample t-test

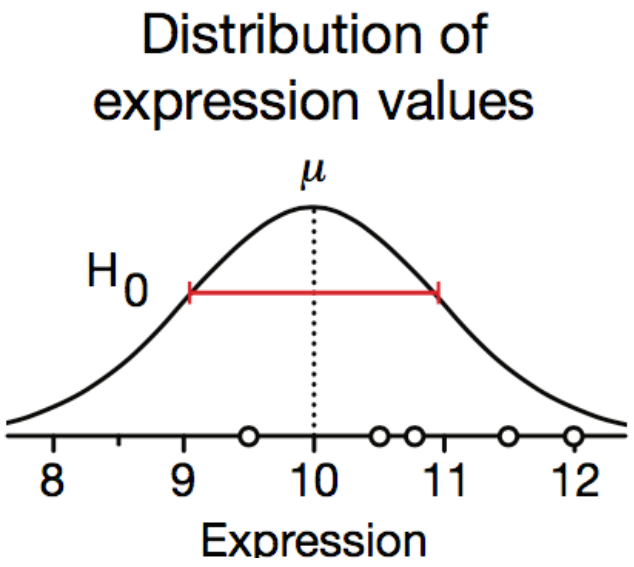
- State your hypothesis (H_0) and collect data
- Determine your desired significance level (usually called α). We usually use P=0.05
- Calculate so-called t-statistic
- Use a lookup table (or function) to determine P
- If $P < P_\alpha$, you may reject H_0

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



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



Closing the loop from the SGID (ELT)

- **Positives**
- **Negatives**
- **Changes I’m going to make**
- **A change I will** try **to make**

Quick review from last time

- Open up your notebook from Monday (hopefully you still have it!)
- Hypotheses in statistics
- Defining the P-value
- One sample t-test
- Practicing with some data

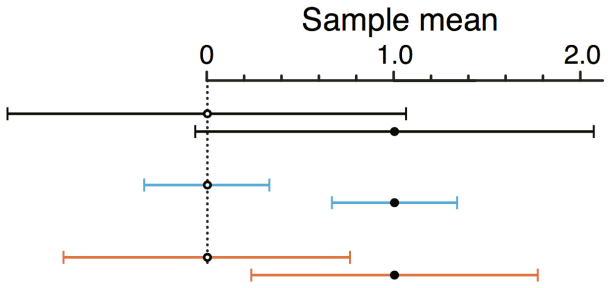
Boltzmann, my dog, chewing on a “bully stick”. Go ahead and google what that is real quick…

P-values, t-tests and confidence intervals

- **Three types of error bars are commonly used**
  - **The s.d** $\bar{x} \pm s.d.$
  - **The s.e.m** $\bar{x} \pm s.e.m.$
  - **The 95% confidence interval** $\bar{x} \pm t^* * s.e.m$
- **The value t* is the** critical t-value **at some desired significance (e.g., P=.05 and d.o.f.:** usually n-1**)**
- **Confidence intervals, error bars, and hypotheses testing often get all jumbled up in our quest to assign meaning to data we've collected!**
- If time: **back to python for more practice!**

Using error bars and confidence intervals

- Do not get in the habit of drawing scientific conclusions based on error bars



$\bar{x} \pm s.d.$

$\bar{x} \pm s.e.m.$

$\bar{x} \pm t^* * s.e.m$

Comparison of two sample means (0 and 1.0), which are different with a P-value of 0.05

- Always define your error bar and, to maximum extent possible, provide your data
- Guide the reader and their expectations

Nuzzo 2014 – 3 fallacies of p-values

TOP DEFINITION

p-hacking

Exploiting –perhaps unconsciously - researcher degrees of freedom until p<.05.

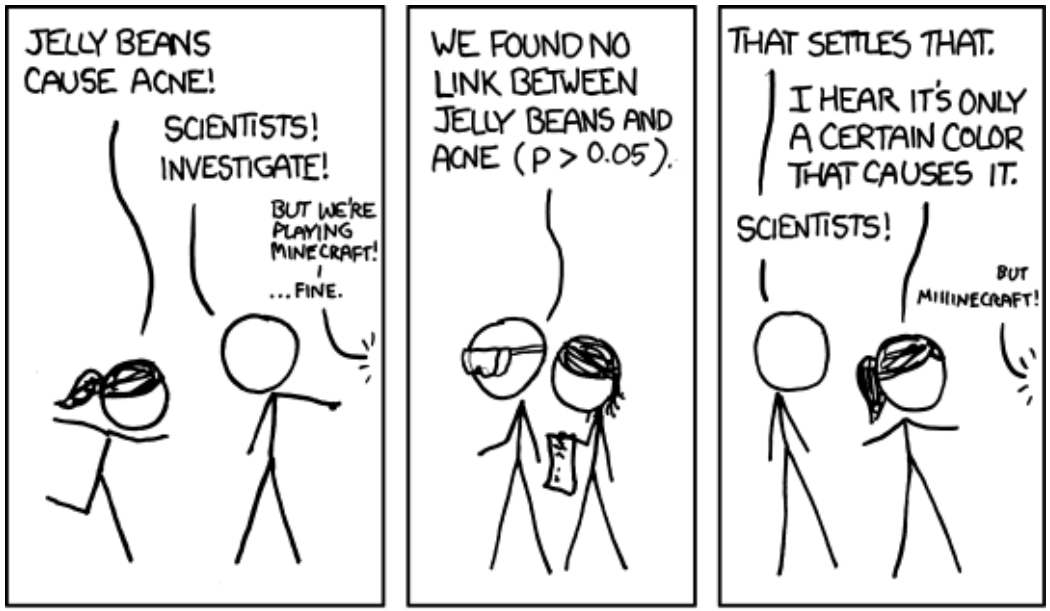
That finding seems to have been obtained through p-hacking, the authors dropped one of the conditions so that the overall p-value would be less than .05.

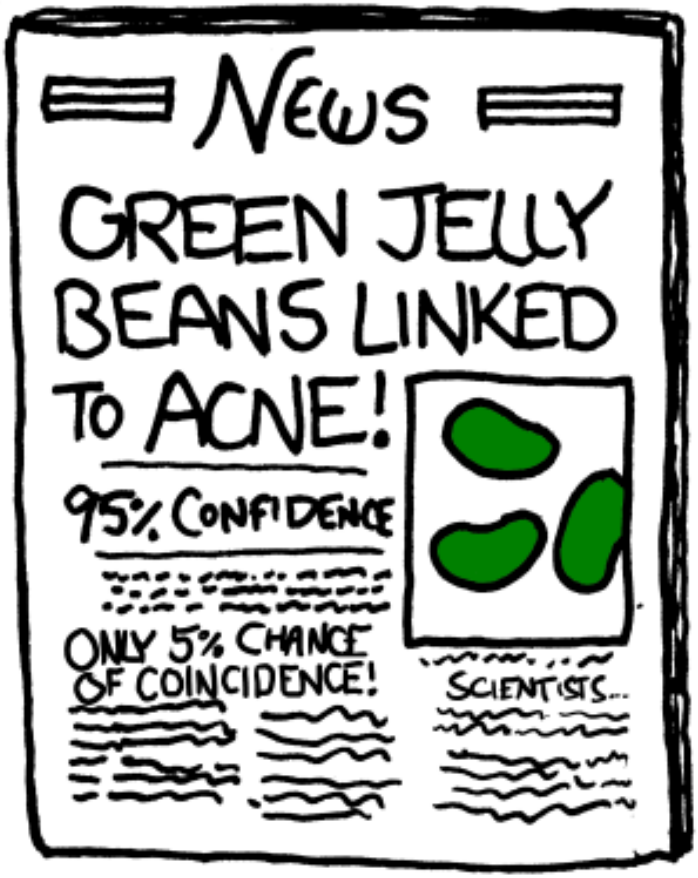
She is a p-hacker, she always monitors data while it is being collected.

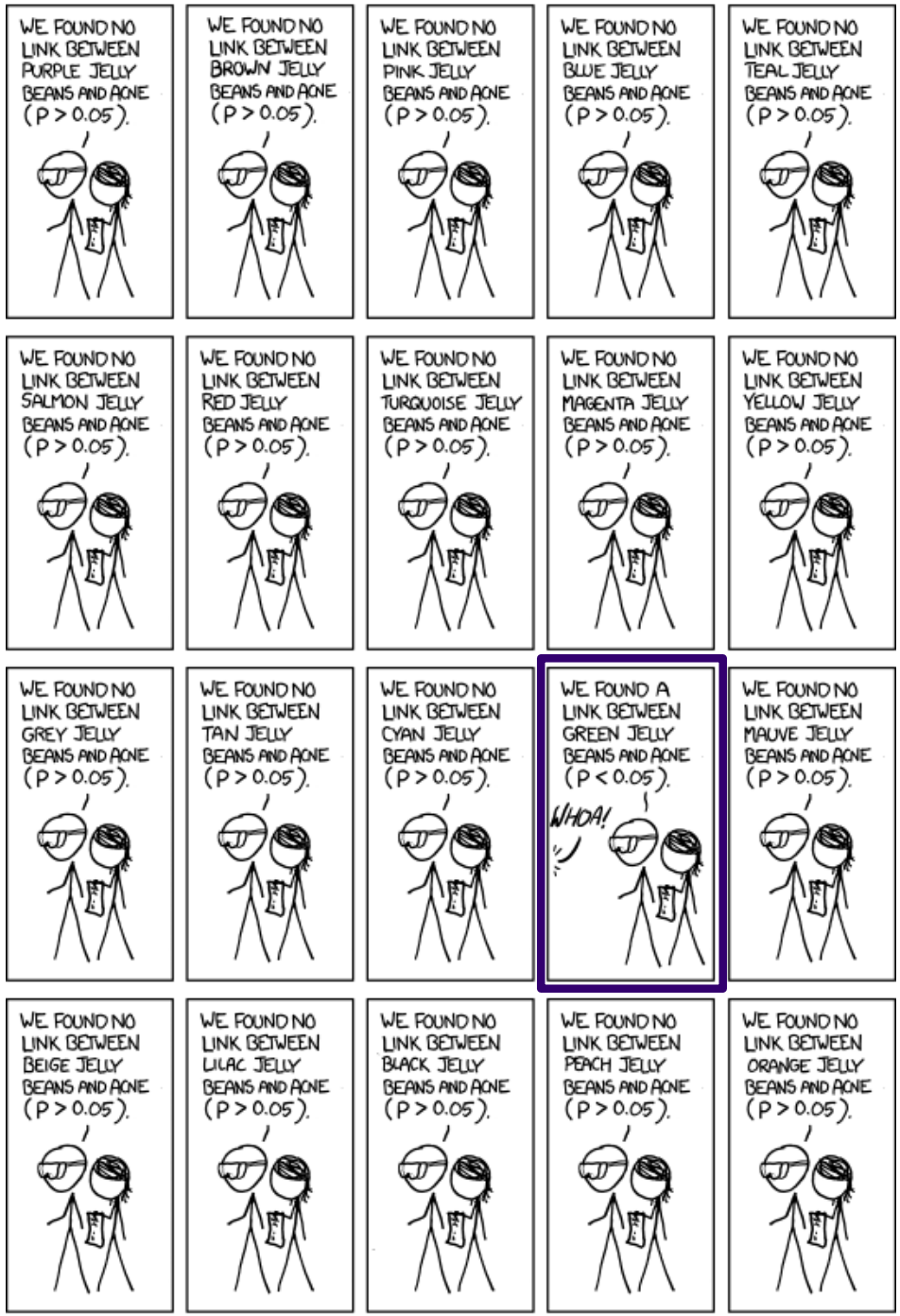
#psychology #false-positive #data monitoring #statistics #researcher degrees of freedom

by PProf January 30, 2012

Dictionary







Hypothesis testing via xkcd

The ASA statement on P-values: high notes

1. **P-values do indicate how** incompatible **data are with a specified** statistical model
2. **P-values do not measure: 1) the probability a hypothesis is true or 2) the probability your data were produced by random chance**
3. **You should never base scientific conclusions on whether or not a P-value test passes a threshold**
4. **Proper interpretation requires full reporting and transparency**
5. **A P-value does not measure the** size of an effect **or the** importance of a result
6. **By itself, a P-value does not** measure the evidence **regarding a model or a hypothesis**

Other types of one sample tests – regression

- One sample tests can have great use in understanding the outcome of regression analysis

$$y = \beta_0 + \beta_1 x$$

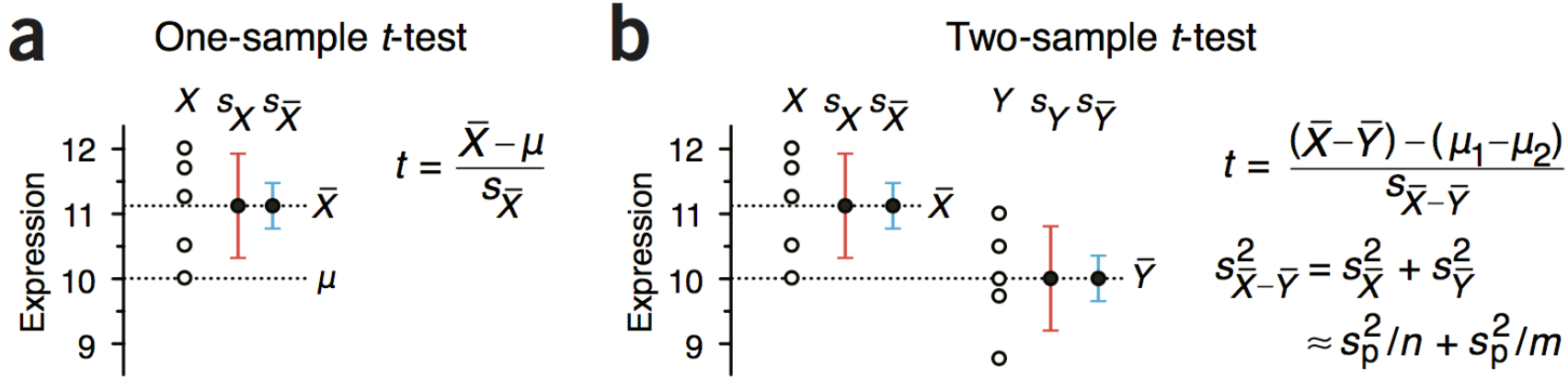
- What null hypothesis might we test to evaluate the fit coefficients?
  - H_0: $\beta_1=0$ or sometimes H_0: $\beta_0=0$

What about two samples?

- In many cases you do not have a single reference value.
  - Consider the case of experiments with a control
    - The control has a sample mean and variance as do each of the experiments testing your effects
    - There are many other examples
  - Uncertainty in your estimation of the significance level of the results comes from the variance in the reference as well as variance in the measurement!
- Fortunately, the one-sample test can be easily extended to a two-sample test

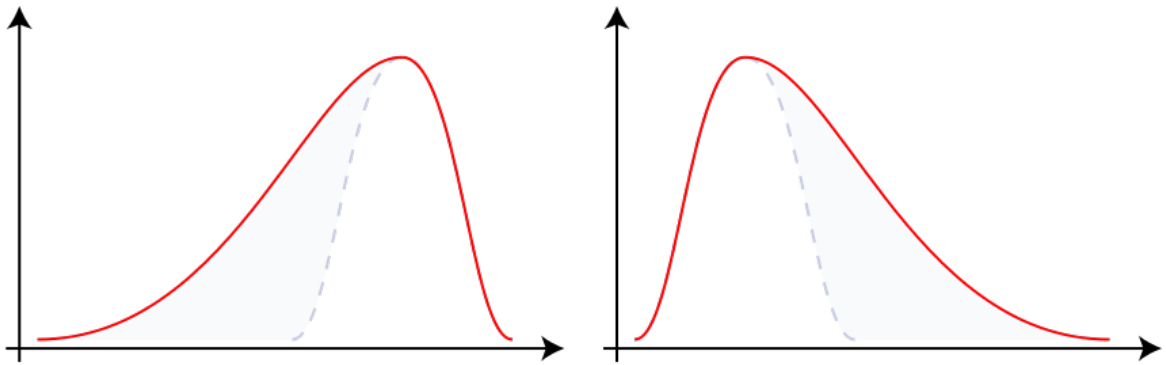
What about two samples?

- **Panel a shows the one-sample test we have already worked on**
- **Panel b shows the concept of the two sample test**
  - **Key takeaway: a modified t-statistic is calculated that includes the difference between the sample means (X/Y) as well as the ability to test for hypothesized differences in means ($\mu_1 - \mu_2$) (**n.b., this is zero when testing for equal means**)**
  - **Use the pooled sample variance, estimated from the two** s.e.m.
  - **Use** d.o.f. **= n+m-2 for P-value calculation**



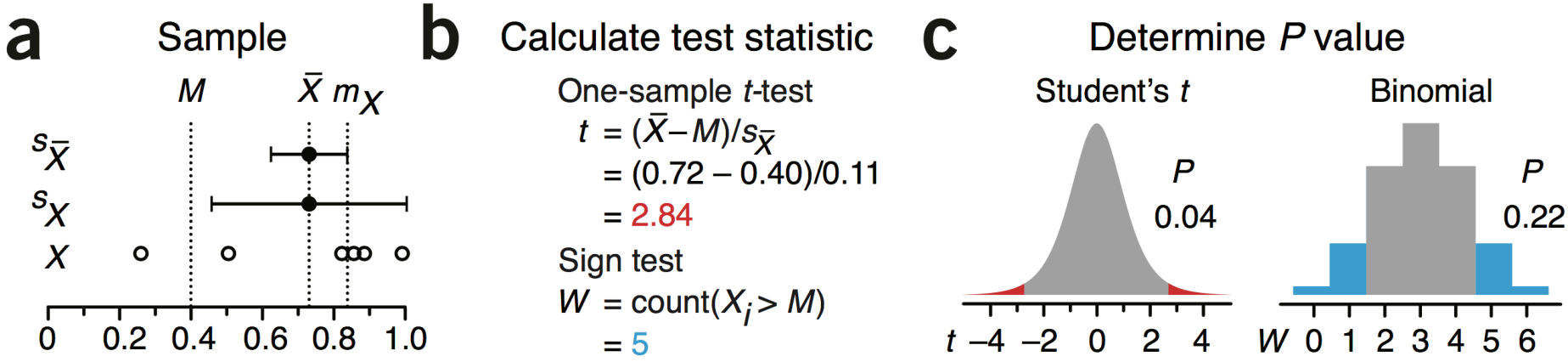
Two other issues…

- What about discrete data?
- What about heavily skewed data sets (i.e., not normally distributed)?
  - Skewness is known as the ‘third moment’ of the distribution – measures the deviation from a symmetric distribution



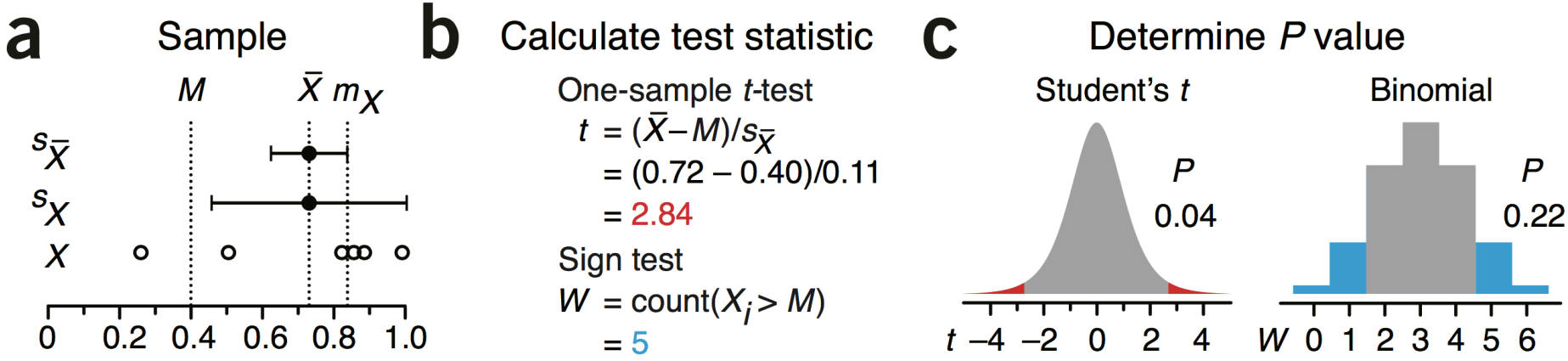
Two approaches are commonly used (1)

- We can use a count based approach that uses the sample medians
- The test statistic is named after "Wilcox" and is called the "W" value
- Panel A shows the data, population median (*M*) and sample median m_x
- Panel B compares the two test statistics we have seen
- Panel C illustrates the P values tested for the t and W scores



Two approaches are commonly used (1)

- **The W score is the number of points that are greater than the proposed population median**
- **The one sample "W-test" (called the** sign test**) looks at the likelihood of drawing a more extreme sample from the binomial distribution (**using ranked data**)**
- **Panel C (right) tabulates all probabilities of values of W and shows the cumulative probability of 0/1 + 5/6 (the cases of more extreme values): this is the two-sided** sign test



Two approaches are commonly used (1)

- **In practice you are usually comparing two sets of discrete data, so a similar but slightly more complicated variant: The "**Wilcoxon rank sum test**" is used**
- **Not sufficient time to cover this quarter, but hopefully we covered enough to get the concept**
- **More information here:**
- **See doi:10.1038/nmeth.2937 (Krzywinski & Altman, 2014) for more information**

Other topics we probably won't get to

- Two sample discrete testing and additional pproaches to hypothesis testing with discrete and/or heavily skewed data
  - See doi:10.1038/nmeth.2937 (Krzywinski & Altman, 2014) for more information

What's up next

> Regression and machine learning