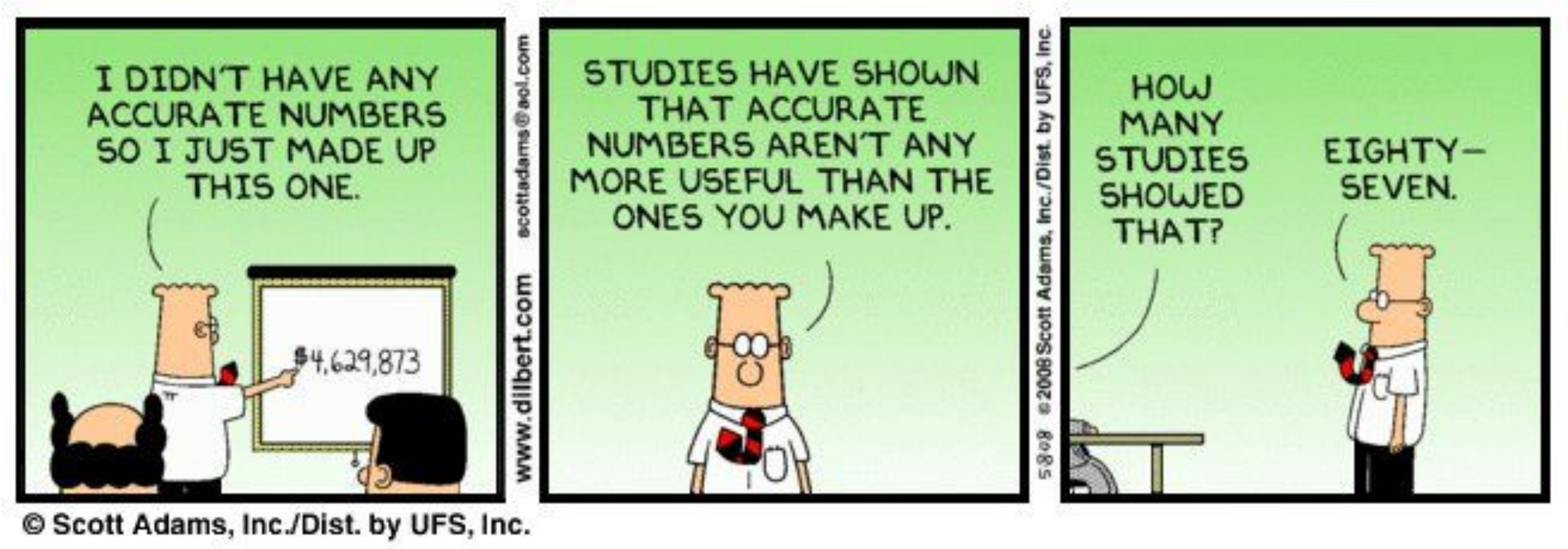
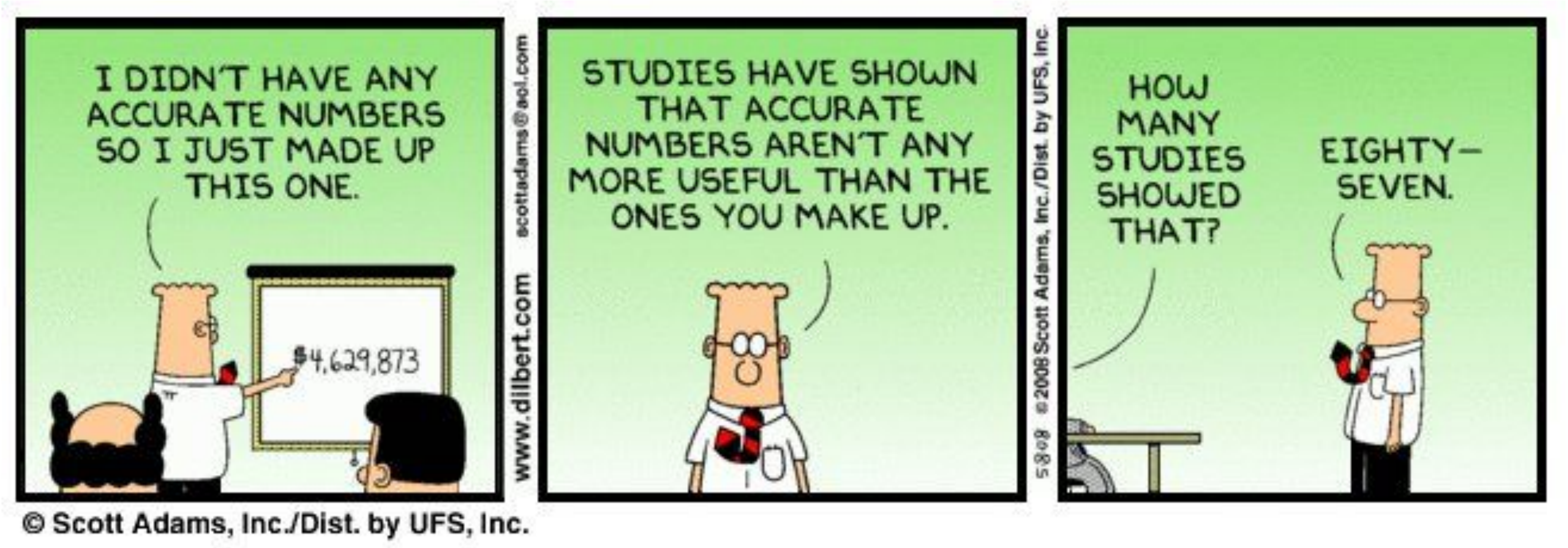
Lecture 19: The Bootstrap



Announcements and reminders

- HW 4 posted!
  Due Friday 5 April at 5 PM

- ~~Quizlet 10 due Wednesday 3 April at 10 AM~~



Previously, on CSCI 3022…

t-dist: small sample from $\boxed{\text{Normal}}$ pop. & don't know σ

A 100·(1-α)% confidence interval for the mean μ when the value of σ is known is given by

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

GPA

A 100·(1-α)% confidence interval for the difference between means:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

A 100·(1-α)% confidence interval for the difference between proportions:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}$$

What about other statistics?

We’ve seen methods for computing CIs for means, proportions and variances…

But what about the median? The skew?

Rather than develop separate theory for each statistic, wouldn’t it be nice if we had a method to compute CIs that would work for almost any statistic we could care about?

What if we don't have enough data?

In real scenarios, data comes at a cost:

- **Money** -- eg, data collected by aircraft (samples from clouds)
- **Time** -- polling people in surveys is time consuming
- **Privacy trade-offs** -- storing a person's genome in the database incurs ethical risk/cost, even if it is cheap time-/money-wise

What if we don't have enough data?

In real scenarios, data comes at a cost:

- **Money** -- eg, data collected by aircraft (samples from clouds)
- **Time** -- polling people in surveys is time consuming
- **Privacy trade-offs** -- storing a person's genome in the database incurs ethical risk/cost, even if it is cheap time-/money-wise

Today: a technique that enables us to tackle the not-enough-data problem, and the I-want-other-statistics problem!

… Today, we tackle the **bootstrap!**

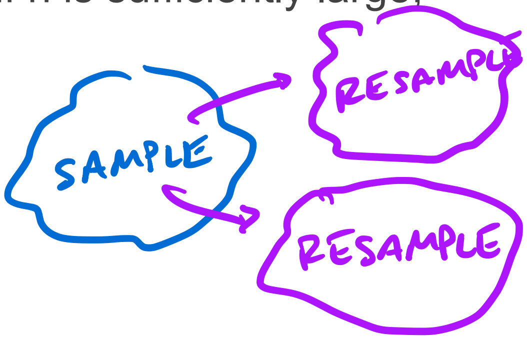
What are bootstraps?

- Bootstraps are the straps you use to pull your boots on
- To “pull yourself up by your bootstraps” is to somehow life yourself upward by pulling on your own shoes… obviously physically impossible.
- In statistics, however, bootstrapping means to accomplish what you need with what you’ve got
- The statistical bootstrap is to **make the most** of a smaller data set without sacrificing statistical rigor or collecting more samples

Confidence intervals for the mean

Recall: If we have n samples from a distribution, the CLT tells us that if n is sufficiently large, the CI for the mean is given by

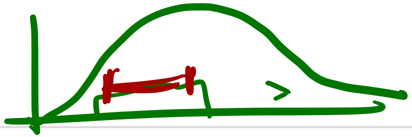
$$\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \quad \text{or} \quad \bar{X} \pm z_{\alpha/2}\sqrt{\frac{s^2}{n}}$$



The **bootstrap** is a different approach. Consider the same sample $X_1, X_2, \ldots, X_n$ as above, but instead of computing a CI analytically from the sample, we instead **re-sample** the sample many times and examine those.

Definition: A **<u>bootstrapped resample</u>** is a set of n draws from the original sample set **with replacement**.

Confidence intervals for the mean



Definition: A **bootstrapped resample** is a set of n draws from the original sample set **with replacement**.

Example: S'pose we have the data X = [2, 2, 4, 7, 9]

estimate the median

$\tilde{x} = 4$

- Resample 1 might be: [2, 4, 4, 4, 7] → $\tilde{x}_1 = 4$
- Resample 2 might be: [4, 4, 4, 4, 4] → $\tilde{x}_2 = 4$
- Resample ~~2~~ 3 might be: [4, 2, 7, 9, 9] → $\tilde{x}_3 = 7$

2, 4, 7, 9, 9

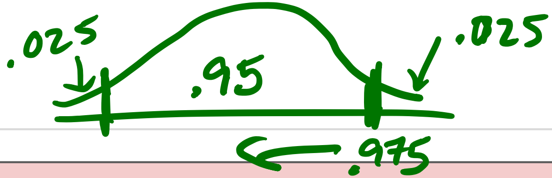
percentile($\tilde{x}$'s, [2.5, 97.5])

= 95% bootstrap CI

Given the example above, what does ***sample with replacement*** mean?

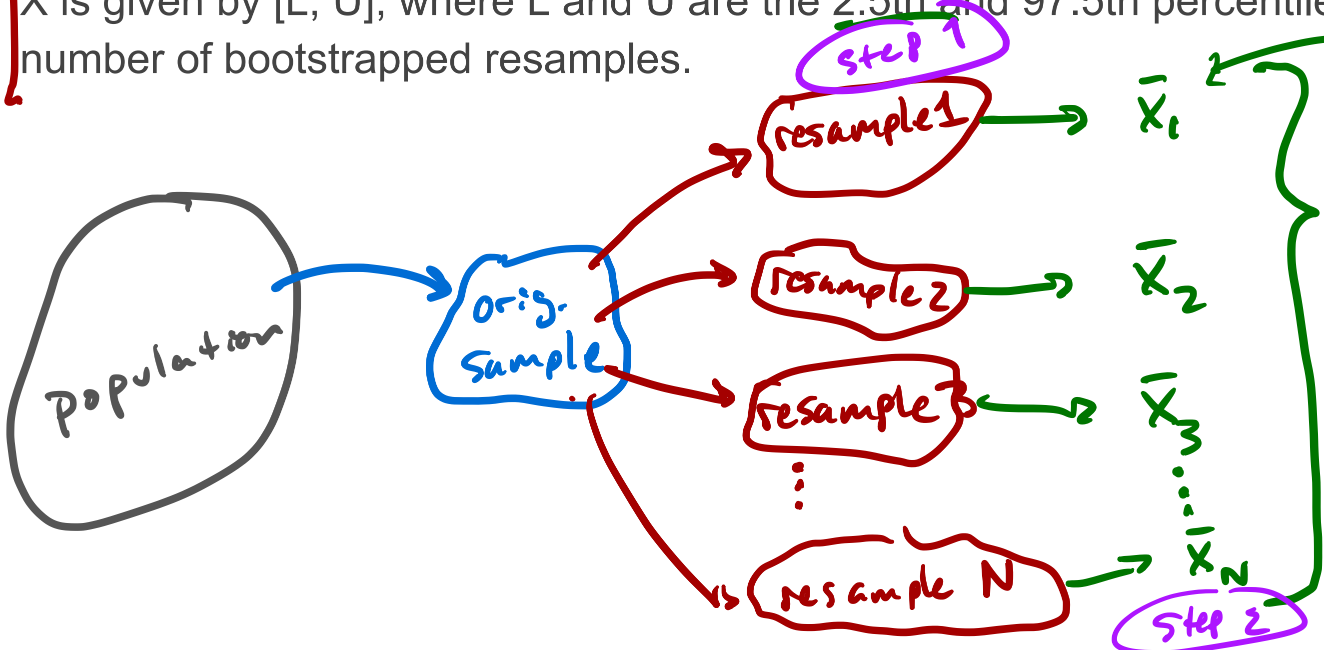
General rule: The bootstrapped resample should contain the same number of observations as the original sample.

Confidence intervals for the mean



Definition: A **bootstrapped resample** is a set of n draws from the original sample set **with replacement**.

Proposition: A suitable estimate of the 95% confidence interval for the mean of the population X is given by [L, U], where L and U are the 2.5th and 97.5th percentiles of the means of a large number of bootstrapped resamples.



all info "xbars", a list

95% CI for pop. mean:

np.percentile(xbars, [2.5, 97.5])

step 3

We <3 bootstrap

- The bootstrap for a CI around the mean is convenient, particularly when there are **not enough samples** to use the CLT
- Of course, if we can use the CLT, we should.   So why is bootstrap so great?

We <3 bootstrap

- The bootstrap for a CI around the mean is convenient, particularly when there are **not enough samples** to use the CLT
- Of course, if we can use the CLT, we should.   So why is bootstrap so great?

We can use bootstrap CIs for things besides the mean!

- Median
- SD
- Other statistical measures that we don’t even have theory for!

Bootstrap for the median

Example: Let's write pseudocode for how we would bootstrap a 90% CI for the median.

meds = []

Given: orig_sample

def not "sample"

for a large # of resample:

resample = random.CHOICE (orig_sample, w/ rep.)

med_resample = np.median(resample)

meds.append(med_resample)

CI = np.percentile(meds, [5, 95])

Bootstrap for the variance

Example: Let's write pseudocode for how we would bootstrap a 90% CI for the variance.

NOPE just find + replace median $\longrightarrow$ variance

Non-parametric bootstrap

In the literature -- your book, Wikipedia, etc. -- you may see the previous methodology referred to as the "**non-parametric bootstrap**". So… what does that mean?

Non-parametric bootstrap

In the literature -- your book, Wikipedia, etc. -- you may see the previous methodology referred to as the "**non-parametric bootstrap**". So… what does that mean?

Definition: **parametric statistics** assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters.

→ Can you name some **examples** of distributions with parameters?

$N(0, 1)$

μ σ^2

→ Can you give an example of a ***non***-parametric distribution we have talked about?

$P(X=0) = 1/2, \quad P(X=1) = 1/3, \quad P(X=2) = 1/6$

Parametric bootstrap

We call the bootstrap discussed in class today the **non-parametric bootstrap** because it doesn't assume any particular parametric distribution. What you resample is what you get.

Definition: The **parametric bootstrap** estimates a CI for a desired property in ~~two~~ three steps:

1) Estimate the parameter(s) of the known distribution from your sample $\mu \approx \bar{x}$
2) Draw bootstrap resamples from the distribution, ***assuming*** $\underbrace{\text{the estimated parameter}}_{}$
3) Compute a CI for the desired property from your resamples. assume is true: draw from $N(\bar{x}, \_)$

Parametric bootstrap

We call the bootstrap discussed in class today the **non-parametric bootstrap** because it doesn't assume any particular parametric distribution. What you resample is what you get.

Definition: The **parametric bootstrap** estimates a CI for a desired property in two steps:

1) Estimate the parameter(s) of the known distribution from your sample
2) Draw bootstrap resamples from the distribution, ***assuming*** the estimated parameter
3) Compute a CI for the desired property from your resamples.

est. CI for median of our data

Example: a) Assume data ~ Pois(λ), and estimate λ from your data set $\underbrace{(\text{mean}(X))}_{\hat{\lambda}}$

b) Generate N bootstrap resample data sets, sampling from Pois($\hat{\lambda}$)

c) Use each of those N resampled data sets to get N estimates of the median of the actual Poisson dist.

d) Compute the CI for the median from that pool of N medians (using percentiles)

Parametric bootstrap

HERE

We call the bootstrap discussed in class today the **non-parametric bootstrap** because it doesn't assume any particular parametric distribution. What you resample is what you get.

Definition: The **<u>parametric bootstrap</u>** estimates a CI for a desired property in two steps:

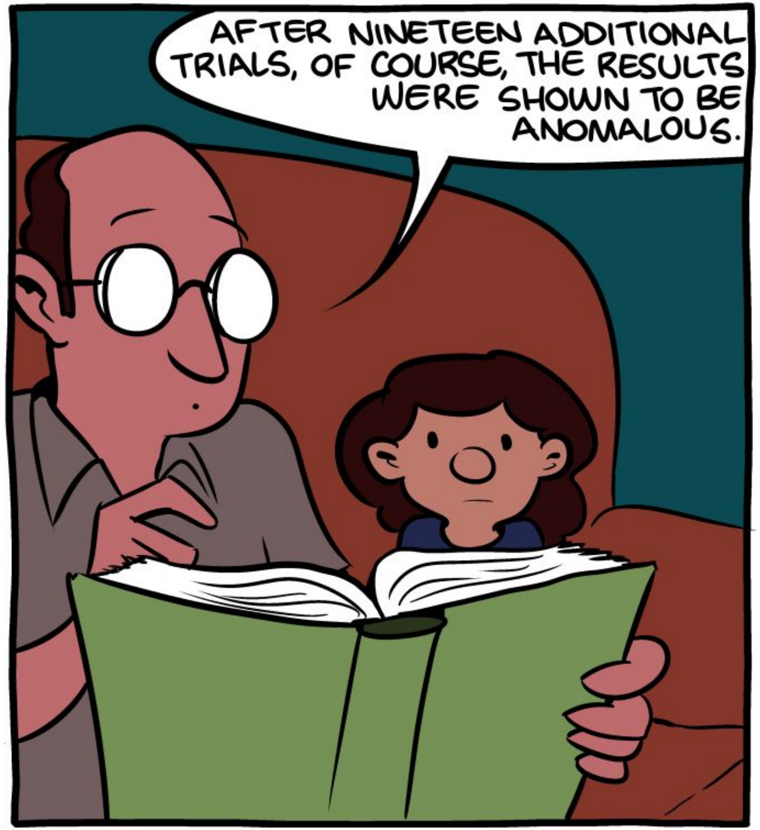
1) Estimate the parameter(s) of the known distribution from your sample
2) Draw bootstrap resamples from the distribution, ***assuming*** the estimated parameter
3) Compute a CI for the desired property from your resamples.

Pro: the parametric bootstrap can be shown to do a better job than the non-parametric bootstrap in particular scenarios

Con: works great if the population has the distribution you assumed. Not so great otherwise.

What just happened?

- **The bootstrap** happened!
- A way to compute confidence intervals even when…
  - We don't have theory for statistics
  - We don't have enough samples for CLT



"The Tortoise And The Hare" is actually a fable about small sample sizes.

Okay! Let's get to work!

Get in groups, get out laptops, and open **nb 16** notebook

Lets…

- Write a function that takes in samples, and computes a 95% confidence interval for the mean by bootstrapping the sample
- Compare the bootstrapped CI with the traditional 95% CI
- Come up with a way to test empirically whether this is working or not…
- Generate some bootstrapped CIs for the median and standard deviation
- Explore the parametric bootstrap