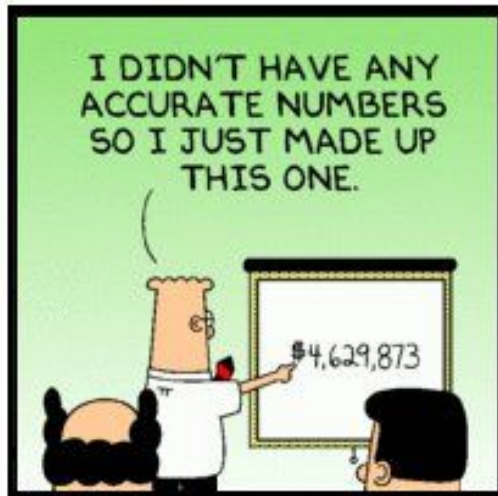
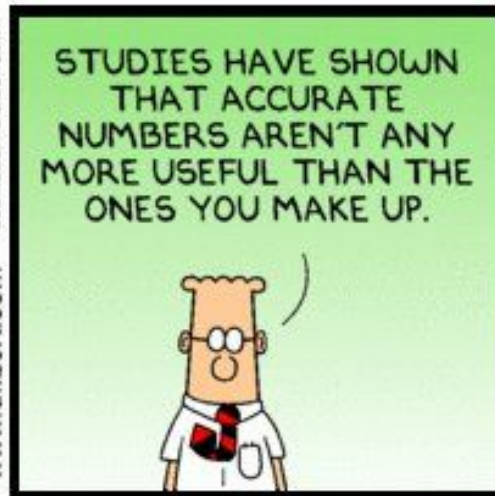




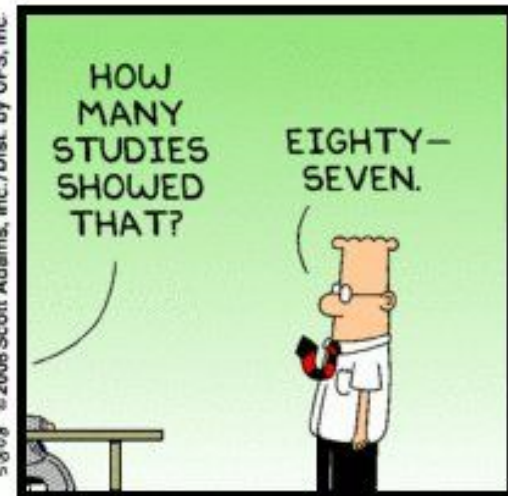
Lecture 19: The Bootstrap



www.dilbert.com
scottadams@aol.com



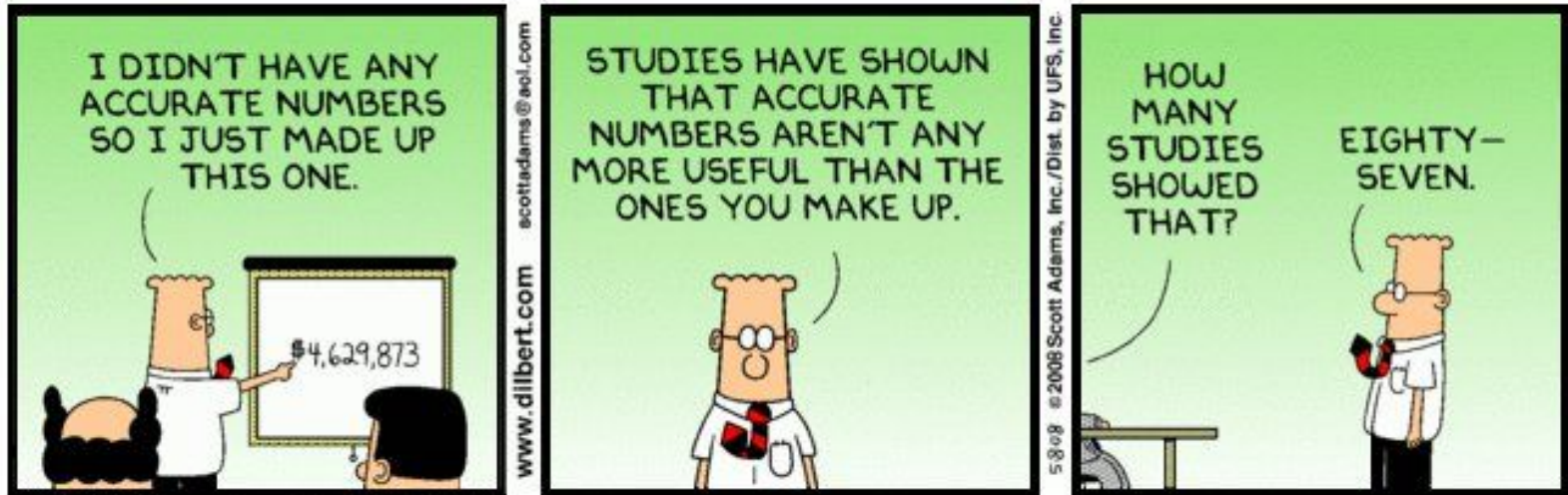
580g ©2008 Scott Adams, Inc./Dist. by UFS, Inc.



© Scott Adams, Inc./Dist. by UFS, Inc.

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...

A $100 \cdot (1-\alpha)\%$ confidence interval for the mean μ when the value of σ is known is given by

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

A $100 \cdot (1-\alpha)\%$ 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 $100 \cdot (1-\alpha)\%$ 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 lift 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

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

sample 

The **bootstrap** is a different approach. Consider the same sample X_1, X_2, \dots, 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 bootstrapped resample 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**.

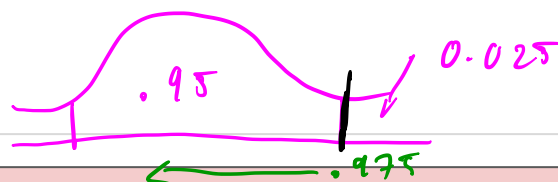
Example: S'pose we have the data $X = [2, 2, 4, 7, 9]$ *estimate the median $\tilde{x} = 4$*

- Resample 1 might be: $[2, 4, 4, 4, 7] \rightarrow \tilde{x}_1 = 4$
 - Resample 2 might be: $[4, 4, 4, 4, 4] \rightarrow \tilde{x}_2 = 4$
 - Resample 3 might be: $[4, 2, 7, 9, 9] \rightarrow \tilde{x}_3 = 7$
2, 4, 7, 9, 9
- per centile $(\tilde{x}', [2.5, 97.5])$
= 95% bootstrapped CI*

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

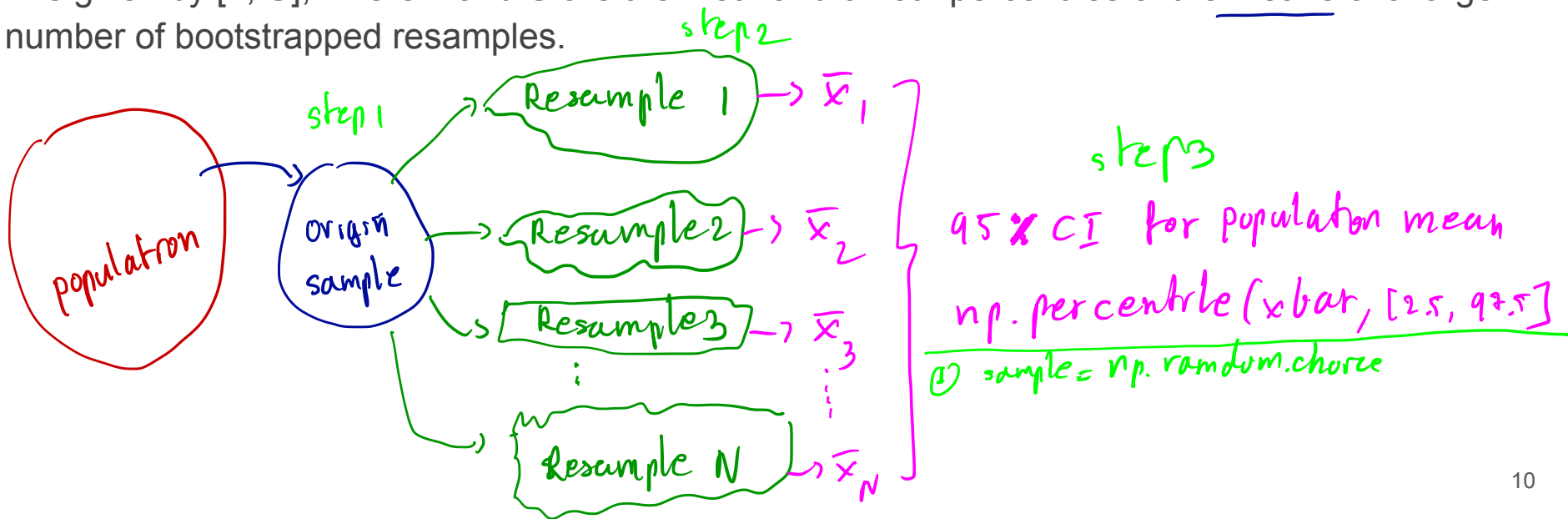
General rule: The bootstrapped resample should contain the same number of observations as the original sample.

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.



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`

for a large of resample:

`resample = np.random.choice(orig.sample, with replace)`

`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.

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?

$$N(0, 1)$$

$\uparrow \quad \uparrow$
 $\mu \quad \sigma^2$

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

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

flexible

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:

$$\mu = \bar{x}$$

- 1) Repeatedly estimate the parameter(s) of the known distribution via bootstrap
- 2) Compute a CI for the desired property by sampling from the known distribution using the parameters that you inferred

② draw bootstrap resample from distribution, assuming the estimated parameter

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) Repeatedly estimate the parameter(s) of the known distribution via bootstrap
- 2) Compute a CI for the desired property by sampling from the known distribution using the parameters that you inferred

Example: a) Create N bootstrap resample data sets

b) Assume data $\sim \text{Pois}(\lambda)$, and estimate λ for each of the N data sets

c) Use each of those N values of λ , and compute the median of that actual Poisson dist.

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

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:

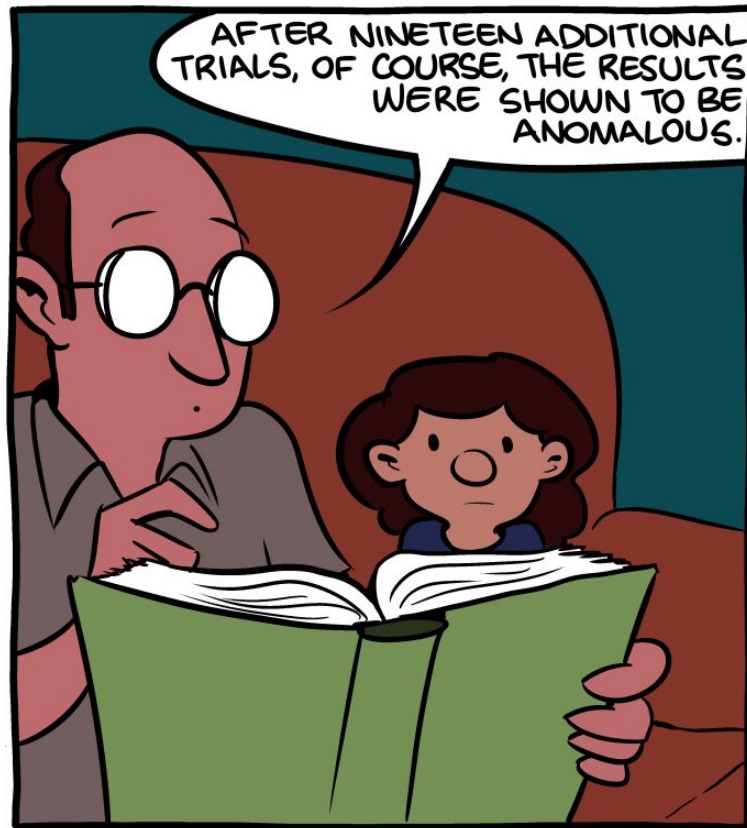
- 1) Repeatedly estimate the parameter(s) of the known distribution via bootstrap
- 2) Compute a CI for the desired property by sampling from the known distribution using the parameters that you inferred

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.
