STA130 TUT0110 Week6

TA: Gloria

shiyi.hou@mail.utoronto.ca

Midterm Reminder

- No class or tutorials next week (reading week). No OH but Piazza will be monitored.
- Midterm: Friday after reading week (Feb 28) during regular tutorial time in EX100.
  - You MUST attend the correct section's term test.
  - No calculators are allowed.
  - See Quercus for other details and practice tests.
  - Take advantage of our TA office hours in HS381/390!

Agenda

Material and vocabulary review

Group discussion

planning your poster project

Writing activity: poster project plan

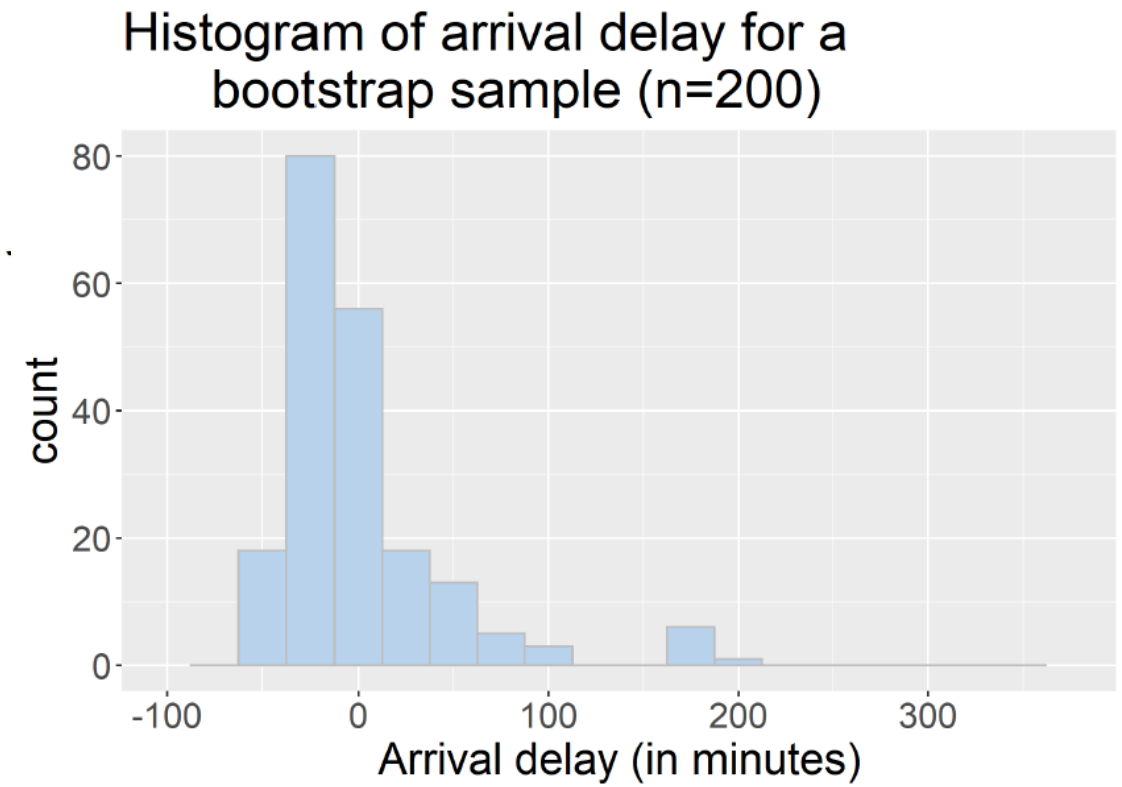
Vocabularies

- Parameter
- Statistic
- Population
- Sample
- Sampling distribution
- Random sampling
- Resampling

Vocabularies

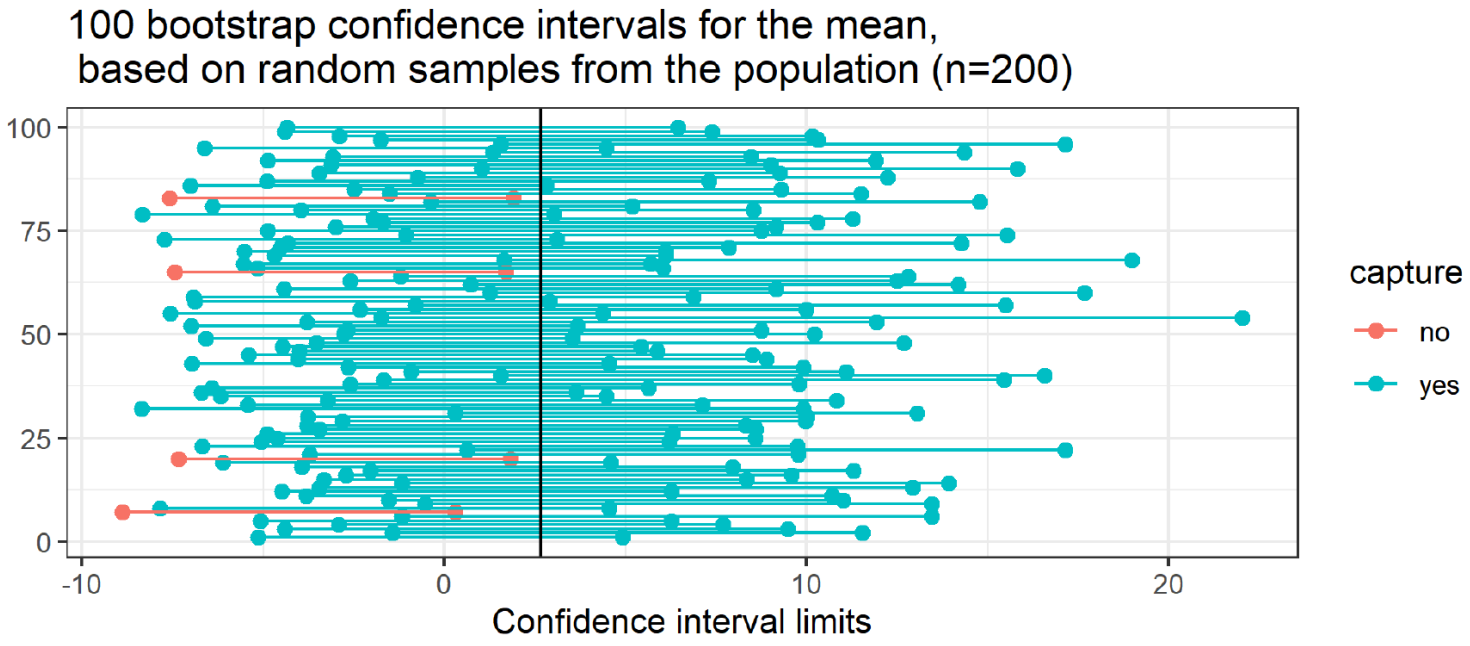
- Bootstrap
- Percentile (Quantile)
- Confidence interval
- Confidence level
- Testing
- Estimation
- Representative

```
boot_samp <- observed_data %>%
  sample_n(size=200, replace=TRUE)
```



Vocabularies

- Bootstrap
- Percentile (Quantile)
- **Confidence interval**
- Confidence level
- Testing
- Estimation
- Representative



Vocabularies

- Bootstrap
- Percentile (Quantile)
- **Confidence interval**
- Confidence level
- Testing
- Estimation
- Representative

→ Always check that your CI range makes sense.

- E.g. if reported the CI for a proportion, it needs to be bounded by zero and one. You can't have a probability less than zero or greater than 1.

Vocabularies

- Bootstrap
- Percentile (Quantile)
- Confidence interval
- Confidence level
- Testing → testing for null hypothesis given an observed statistics
- Estimation → estimate the range of true parameter given a observed sample
- Representative

Why do we do bootstrap?

To estimate the sampling distribution of a statistics!

- E.g. CI.

Why do we construct CI?

To obtain an estimate the parameter that reflects sampling variability.

- E.g. estimate the proportion of people in Toronto who use the TTC;
- E.g. number of coffees people in this class drink each week, etc.

Wide CI: if you had taken a different sample from the population, you could arrive at a very different estimate.

Narrow CI: if you had taken a different sample from the population, you could expect to get a similar estimate.

Does percentile bootstrap method always work?

The percentile bootstrap method (that we are using this week) works best for large samples and when the bootstrap distribution is **approximately symmetric** and **continuous**.

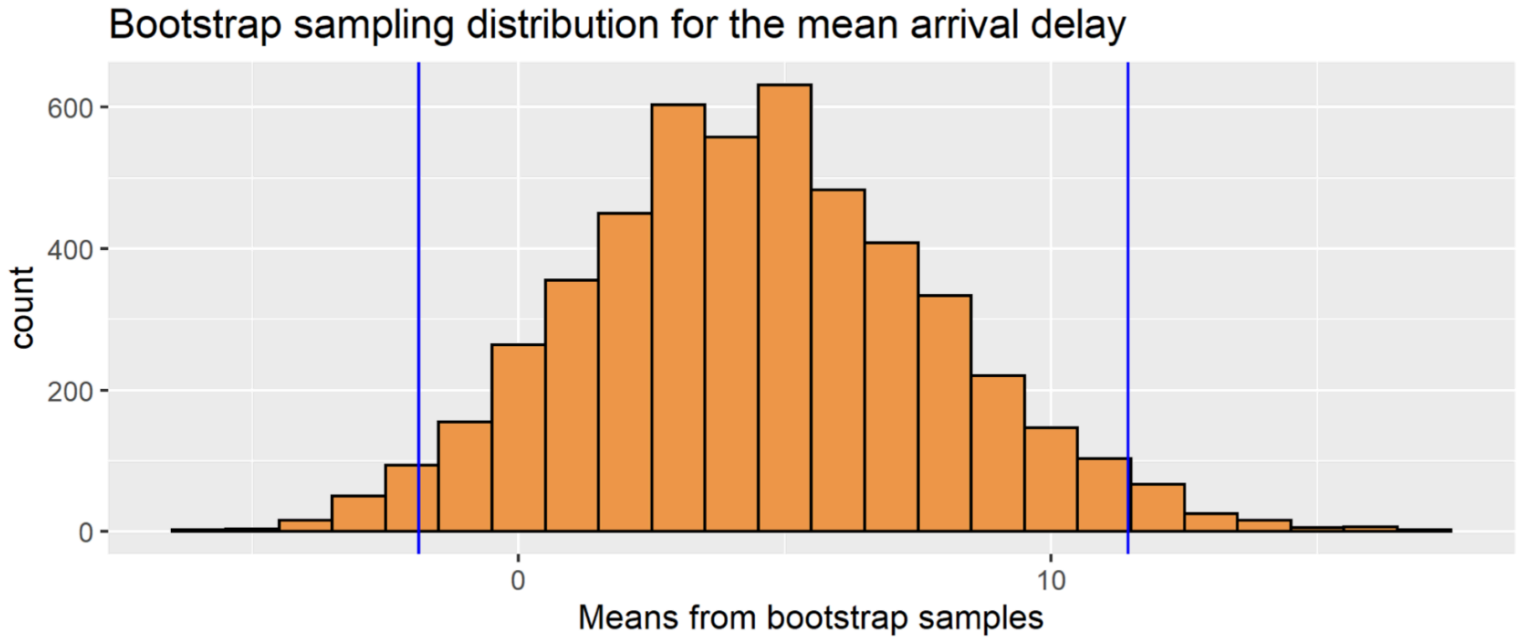
- Therefore, your CIs should be roughly symmetric around the point estimate.
- You will see other versions of the bootstrap method in future statistics courses.

Paired Discussion

- Are the use of p-values and confidence intervals mutually exclusive?
- What do the two have in common?
- How do they differ?
- Think about under which circumstances you may want to use each of these

Paired Discussion

- Are the use of p-values and confidence intervals mutually exclusive?
- What do the two have in common?
- How do they differ?
- Think about under which circumstances you may want to use each of these



Paired Discussion

- If we want to be more confident in correctly capturing the correct proportion of our outcome, such as the percent of couples who tilt their heads right when kissing, should we use a larger or smaller confidence interval?
- How do you think this relates to type I and/or type II errors?

Paired Discussion

- If you and your partner both applied the same bootstrap sampling method to the same data, do you expect that you both arrive at the same estimate and CI?
- What are some factors that you would need to consider (and hold constant) to ensure that you both arrived at the same answer?

Paired Discussion

It's Valentine's Day! You are interested in whether there is a difference in the proportion of couples who tilt their heads to the right or left when they kiss! You survey several students on campus and find that 35.5% tilt their heads to the left when kissing; 95% CI: (27%, 44%).

Which of the below (Question 1d) is the correct interpretation of the 95% CI?

- (i) We are 95% confident that between 27% and 44% of kissing couples in this sample tilt their head to the left when they kiss
- (ii) There is a 95% chance that between 27% and 44% of all kissing couples in the population tilt their head to the left when they kiss.
- (iii) We are 95% confident that between 27% and 44% of all kissing couples in the population tilt their head to the left when they kiss.
- (iv) If we considered many random samples of 124 couples, and we calculated 95% confidence intervals for each sample, 95% of these confidence intervals will include the true proportion of kissing couples in the population who tilt their heads to the left when kissing.

Poster Project!

- Using Toronto Police Service (TPS) data to investigate break & enter robberies (B&Es) over the past few years.
- Somebody from TPS will be coming to class next class to talk more about this project.
- You will receive the data after the midterm.
- We'd like you to start thinking of a research question *before you* see the data.
  - We have provided a handout containing some high-level information in the methods you will learn in later weeks.

Poster Project!

For the poster project, you'll be exploring data on "break and enter" robberies (B&Es) in the city of Toronto over the past several years. Our partners at the Toronto Police Service (TPS) would like *your* help to visualize and analyze their data to identify any patterns and anomalies in B&Es.

Based on these data, you are being asked to provide appropriate analysis(es) that could inform recommendations to members of the community and/or TPS to reduce B&Es. For example, are there any temporal or spatial trends in B&Es; i.e., are there any times or places that are more prone to B&Es?

You should make sure that your final recommendations are specific and tied to the data; i.e. not general recommendations like, "lock your doors and have an alarm", which are not tied to these data. You may consider searching for additional sources of data about Toronto neighborhoods to better understand spatial trends. These data can be found online.

The above objective is very broad, so you and your group will need to formulate 2 or 3 specific questions to help TPS better deploy their resources to prevent/combat B&Es in the future.

Poster Project!

So far in the course, we have discussed many types of plots which you may find useful to visualize the distribution of interesting variables, including histograms, bar plots, scatterplots, and boxplots. We have also discussed two methods of statistical analysis so far:

- ✓ Hypothesis test for one proportion
- ✓ Randomization test to compare two groups; i.e., comparing means, medians, proportions, etc.

Later in the term, we will discuss more statistical methods, which you may want to use in your poster project.

Methods

Week	Method	Type of variable(s) or hypotheses	Example questions
Week 4	Hypothesis test for one proportion	H_0: $p = p_0$ H_A: $p \neq p_0$ (be sure to define the parameters)	- Is the proportion of male students in this class 50%?
Week 5	Hypothesis test (Randomization test) comparing the means or proportions between two groups	Examples: H_0: $p_1 = p_2$; H_A: $p_1 \neq p_2$ *or* H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$ (be sure to define the parameters)	- Is the proportion of male students in this class the same as Canada? (In case you're interested, the current national sex ratio is 49.63%!) - Is the mean time spent commuting to class similar between students who live with their family or those who live elsewhere (e.g. in a shared house, on campus, etc.)?
Week 6 & 7	Bootstrap confidence intervals	Focuses on a population parameter; e.g. population mean, median, proportion, etc.	- What is a range of plausible (i.e., reasonable) values for the average post-graduation salary of somebody with a UofT undergrad degree in statistics? - What is a range of plausible values for the average treatment effect of a new weight loss drug?
Week 8	Classification trees	- Response: categorical - Predictors: categorical and/or numerical	- Based on information that we know about somebody (e.g. based on their age, gender, etc.), can we predict their favorite type of music? (e.g. pop, rock, rap, etc.)
Weeks 9 & 10	Linear regression	- Response: numerical - Predictor(s): numerical and/or categorical	- Based on some information that we know about a person (e.g. past grades, time spent studying, etc.), can we predict their grade on the STA130 midterm? - Is there an association between the number of tutorials attended and final STA130 grade?

Data

Variable	Description
Index	Record Unique Identifier
event_unique_id	Event Unique Identifier
occurrencedate	Date of occurrence
reporteddate	Date occurrence was reported
premisetype	Premise where occurrence took place. This variable takes the following values: • Apartment • Commercial • House • Outside: Outside of a building (e.g. shed or garage) • Other
offence	Offence related to the occurrence, and takes the following values: • B&E: Break & Enter • B&E - M/Veh To Steal a Firearm: B&E to motor vehicle to steal a firearm • B&E - To Steal Firearm: Break & Enter to steal a firearm • B&E Out: Break & Enter outside (ex: a shed or garage) • B&E W'Intent: Break & Enter with intent to commit offence (usually theft) • Unlawfully In Dwelling-House: Usually means the person has access to the house (e.g. a key) but has been asked not to enter or has been asked to leave and refuses. For example, if landlords evict a tenant but they retain their key and continue to be in the house - technically they are not "breaking in".
reportedyear	Year occurrence was reported
reportedmonth	Month occurrence was reported
reportedday	Day occurrence was reported
reporteddayofyear	Day of week occurrence was reported
reporteddayofweek	Day of year Occurrence was reported
reportedhour	Hour occurrence was reported
occurrenceyear	Occurrence year
occurrencemonth	Occurrence month
occurrenceday	Occurrence day
occurrencedayofyear	Occurrence day of year
occurrencedayofweek	Occurrence day of week
occurrencehour	Occurrence hour
MCI	Major Crime Indicator related to the offence
Division	Division where event occurred
Hood_ID	Neighbourhood ID assigned to occurrence after offsetting longitude and latitude Coordinates to nearest intersection node
Neighbourhood	Neighbourhood name assigned to occurrence after offsetting longitude and latitude Coordinates to nearest intersection node
Lat	Latitude of point extracted after offsetting longitude and latitude Coordinates to nearest intersection node
Lon	Longitude of point extracted after offsetting longitude and latitude Coordinates to nearest intersection node

For each question you formulate, go through the following checklist:

- Would the Toronto Police Service be interested in knowing the answer to this question and / or would it help them do their job better?
- Do we have the data required to answer this question? (e.g. do we currently have the necessary variables, or can we find them somewhere else?)
- Have we learned an appropriate statistical method which will allow us to answer this question? (For today's writing activity, focus only on questions which can be answered using the statistical methods covered so far in the course).

Form your poster group

- Find your own group mates
- This will be your tentative final group
- Each group will have 3-4 people to collaborate

Group Discussion
for potential research proposal for the poster project

- The research question(s) you are interested in investigating.
  - What is something that the Toronto Police might be interested in knowing about B&Es patterns?
- Any interesting visualizations to consider? Remember, these should be interesting and useful for the audience.
- Think about whether you will need to join datasets to answer their research question.
  - You can use additional sources of data
  - **NOTE:** We will discuss joins after reading week, there is a link to some examples on the project page.
- What will each group member be responsible for?
- How often will you meet?

Aspects to consider in your research plan:

- Aims/objectives
- 2-3 research question(s)
- The hypotheses
- Research design: data/ variables, methods*, visualizations, etc.

Writing Activity

- Write-up your research plan individually, based on what you discussed with your group.
- Consider formatting it to include the following sections
  - Research question;
  - Hypothesis;
  - Objective;
  - Research design: data/ variables, methods, visualization