STA130H1S TUT0109
W6: Sampling, Bootstrap Samples, and Confidence Intervals

Feb 15, 2019

Email: sonia.chhay@mail.utoronto.ca
Website: soniachhay.github.io

Announcements

- No class and tutorial next week (reading week!)
  - No OH but Piazza will be checked regularly
  - New College Stats Aid Centre will be open during reading week (Wetmore Hall, Room 68A)
- Example midterms have been posted to Quercus
  - Midterm details are on the next slide

About the Midterm

- When: During your usual tutorial time (**Fri March 1st**)
- Where: You MUST attend the correct section's midterm
  - **AM section**: EX 200
  - **PM section**:
    - MS 3154: Last names from A - Lo
    - WB 116: Last names from Lu - Z
- Includes: All material up to & including Feb 25th (mostly a review class)
- Format:
  - Multiple choice
  - Fill in the blanks
  - Written answers (**make sure to write in complete sentences**)

Overview

- Material and vocabulary review
- Group discussion
- Oral presentations

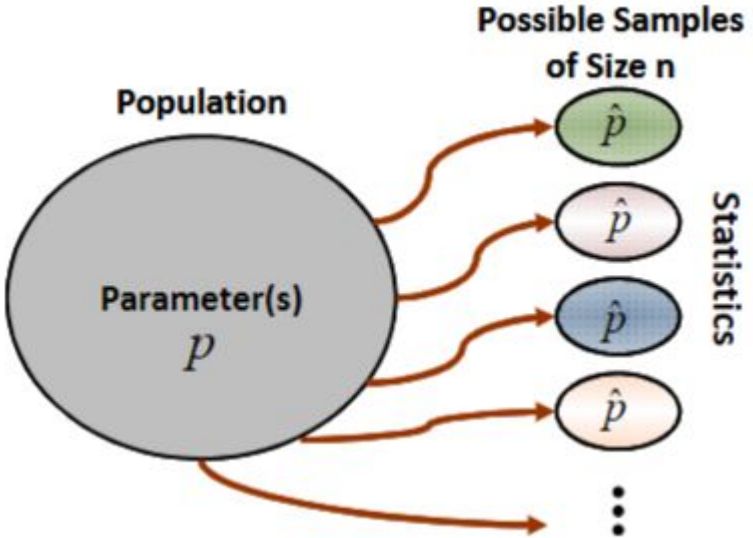
Vocabulary for this Week (1/2)

- Percentile (Quantile)
- Sampling distribution
- Population
- Sample
- Parameter
- Statistic

Percentile (Quantile)

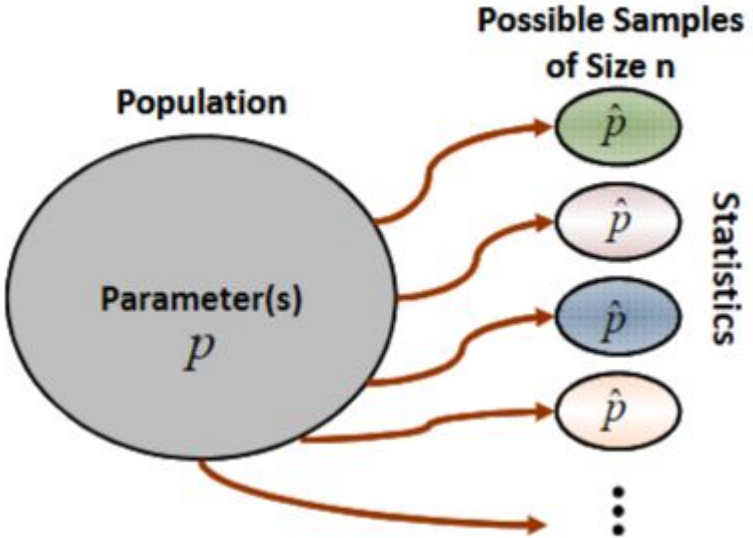
- Percentile: A measure *p* between 0 and 100, in which the p^{th} percentile is the smallest value that is larger or equal to p% of all the values
  - 25^{th} percentile = Quarter of points are less than it
- Quantile: Divides the data into equally sized groups
  - 0.25 or 25% quantile = Quarter of the points are less than it

Sampling Distribution



- Sampling distribution of a statistic: Distribution of statistic values taken for all possible samples of the same size (n) from the same population
- Sampling distribution:
  - Take random samples from a certain population
  - Calculate the statistic of each sample
  - Plot and observe distribution

Sampling Distribution (break-down)



- Population: A whole - Every member/element of a group
- Sample: A fraction or percentage of that group
- Parameter (p): Value that tells you something about a population
- Statistic (p̂): Tells you something about a small *part* of the population

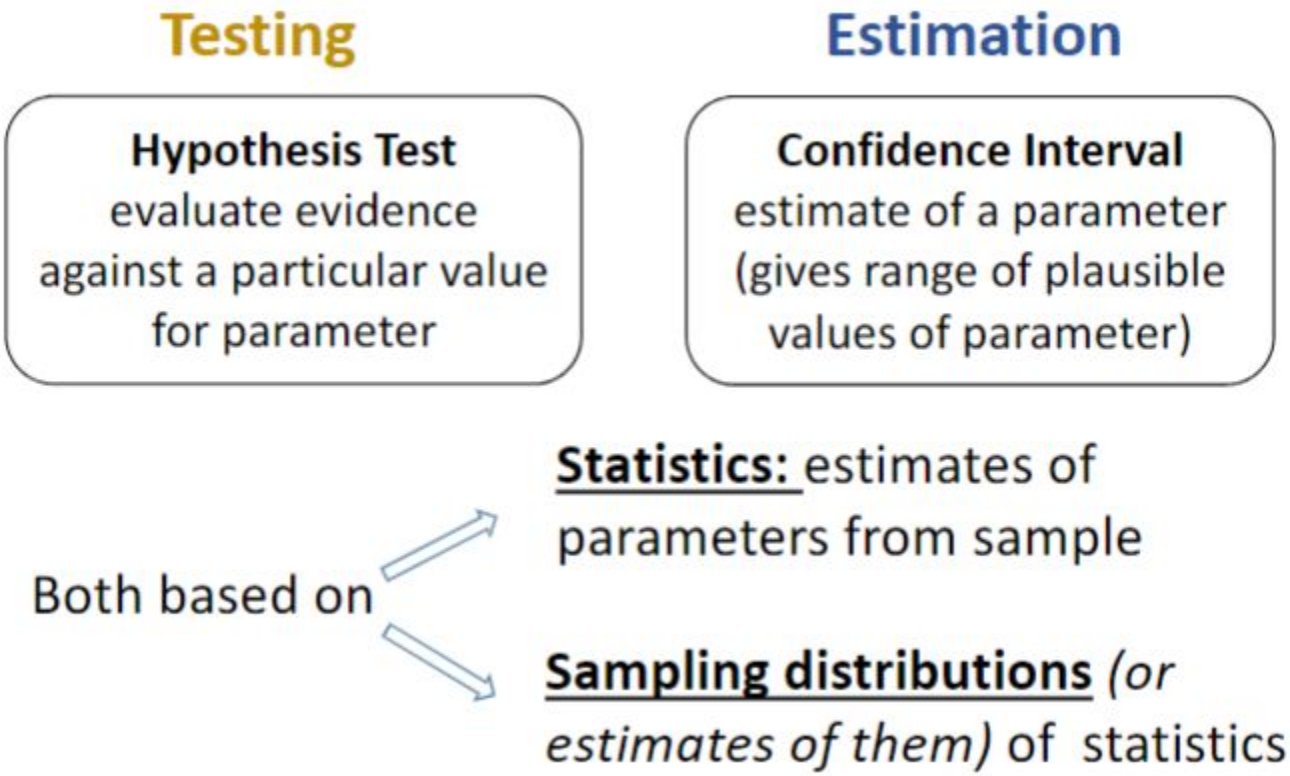
https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/what-is-sampling-distribution/v/introduction-to-sampling-distributions

Vocabulary for this Week (2/2)

- Testing
- Estimation
- Representative

- Resampling
- Bootstrap
- Confidence interval
- Confidence level

Testing and Estimation



Resampling and Bootstrap

- Bootstrap: An approach that helps us describe the variability of statistics based on one sample
  - Does not create new data, only reuses sample data!
  - Resampling method used to estimate statistics on a population by sampling a dataset with replacement
  - I.e. Resampling from the sample -> Take the sample that you have and resample from it
- Purpose of bootstrap: Estimate the sampling distribution of a statistic
  - Allows us to get a confidence interval (CI)

Confidence Interval & Confidence Level

Recall: Using only our sample data, we are trying to come up with a range of values that would be plausible for the true parameter value

- Confidence interval (CI): Captures the parameter value a certain percentage of the time
  - E.g. 95% CI includes the parameter for 95% of possible samples
  - Here, 95% is called the confidence level
- Purpose of CI: Obtain an estimate for the parameter that reflects sampling variability
- Examples of when you could use CI:
  - Wish to estimate proportion of people living in Toronto who use the TTC
  - Number of coffees people in this class drink each week

Confidence Interval (Cont'd)

- 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

Note: Always check that your CI range makes sense!

- E.g. If you reported CI for a proportion, it must be bounded by [0;1]

Confidence Interval (Cont'd)

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

Group Discussion

(10 min)

Group Discussion (10 min)

1. 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.
2. How do you expect the **width (precision)** of a 95% CI to compare to a 90% CI? Compared to a 99% CI? How do you think this relates to **type I** and/ or **type II errors**?
3. 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?

Group Discussion (#1)

1. 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.

Response:

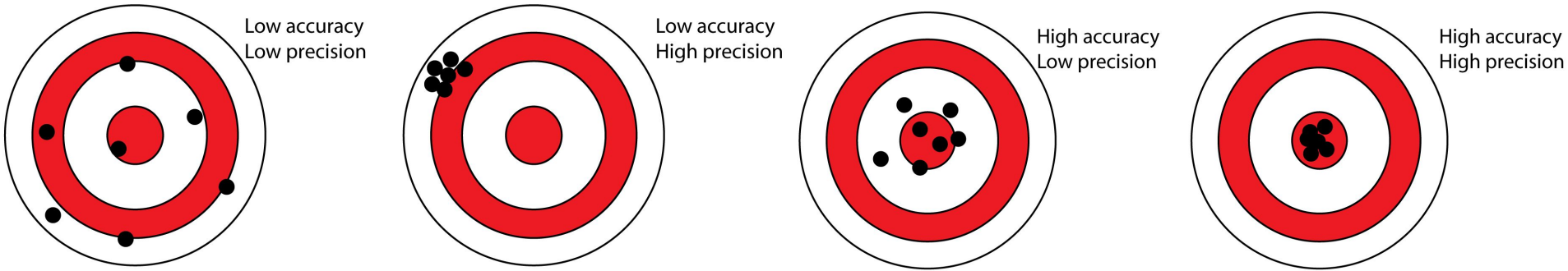
- Both provide insight into statistical significance
- CI also provides information regarding precision of an estimate
- Both are based on statistics and sampling distributions

Group Discussion (#2)

2. How do you expect the **width (precision)** of a 95% CI to compare to a 90% CI? Compared to a 99% CI? How do you think this relates to **type I** and/ or **type II errors**?

Group Discussion (#2 Cont'd)

- Accuracy: Defined in terms of whether or not the CI contains the true population parameter
- Precision: Refers to width of a CI

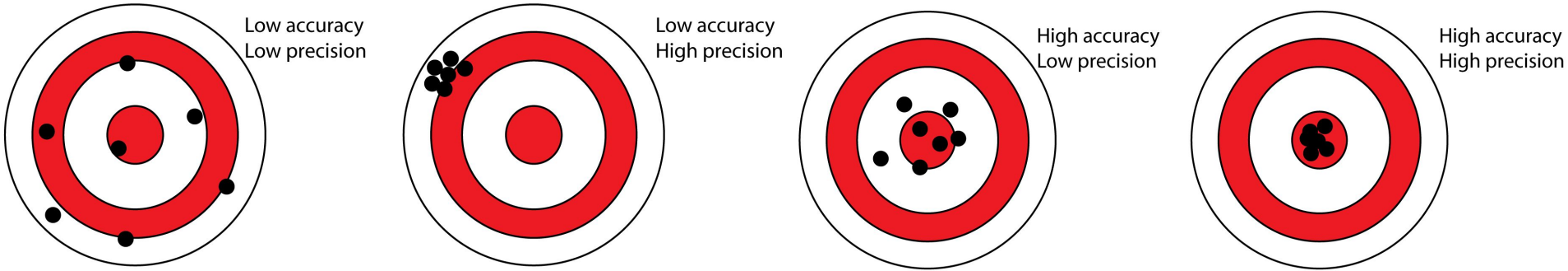


http://researchhubs.com/post/ai/data-analysis-and-statistical-inference/accuracy-vs-precision.html

Group Discussion (#2 Cont'd)

Example: Suppose the weather forecast tells you that the next day will have a low of -30°C and high of 40°C

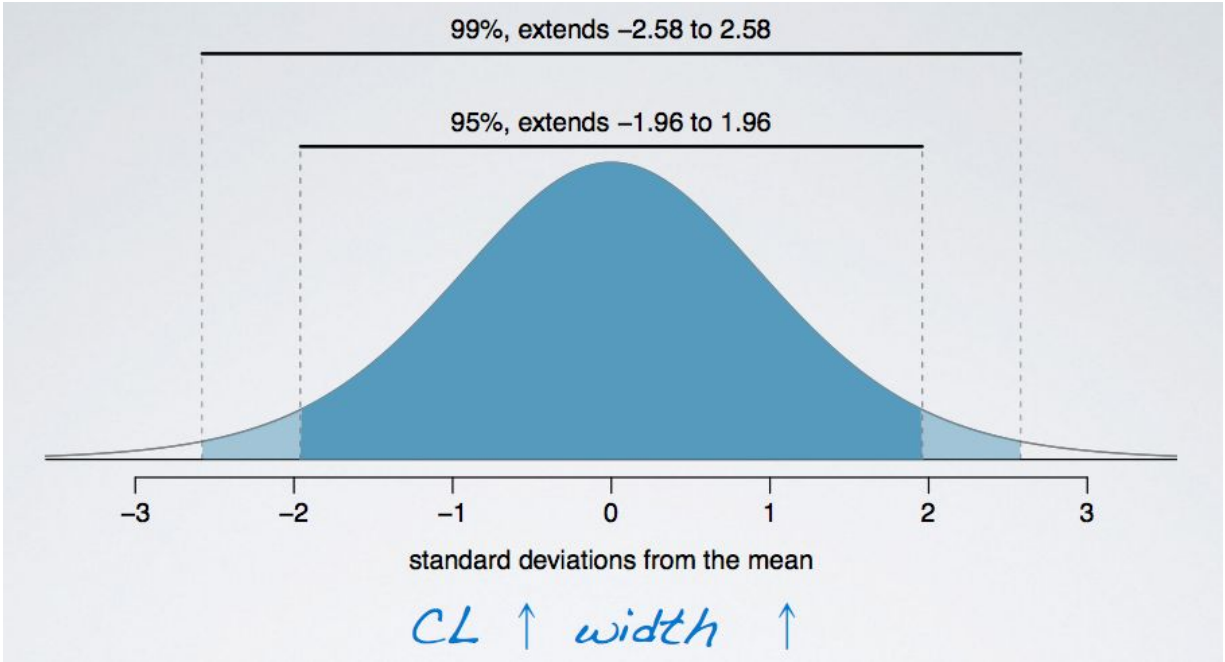
- Accurate? Technically, yes
- But is this informative (i.e. precise)? Not really



CL↑ Width↑ Accuracy↑, but Precision↓

Group Discussion (#2 Cont'd)

- As confidence level increases, so does the width of the CI



Group Discussion (#2)

2. How do you expect the **width (precision)** of a 95% CI to compare to a 90% CI? Compared to a 99% CI? How do you think this relates to **type I** and/ or **type II errors**?

Response:

- 90% CI narrower than 95% CI
  - As precision of CI increases (i.e. CI width decreasing), the reliability of an interval containing the actual mean decreases (less of a range to possibly cover the mean
  - E.g. 100% confidence; in order to ensure mean is captured with 100% certainty, the interval must contain every possible value
- Probability of committing type 1 error is called alpha (i.e. level of statistical significance)
  - For 99% CI, alpha=0.01 (confidence level)
  - Means that there's a 1% probability of making a type 1 error

Group Discussion (#3)

3. 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?

Response:

- With any simulation, results will be different every time
  - To get same results, should set the same seed (starting point for simulations), have same number of repetitions, etc.
- Need a large N to ensure estimates are stable

Group Presentation

Prepare an oral presentation based on the following topics regarding Question #3

(40 min to prepare, 5 min presentation)