“In statistics we apply probability to draw conclusions from data.”

---Prof. J. Orloff

Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 3.16.2021
Last time

- Sample mean
- Confidence interval
- t-distribution (I)
Objectives

- Review Sample mean, CI
- t-distribution (II)
- Bootstrap simulation
1) Why is **sample mean** a random variable?

**ans:** **No.**

2) Is $E[X^{(N)}] = \text{mean}\{x\}$? \(\downarrow\) \(\{x\}\) is some realized data of size \(N\), drawn from the population \(\{X\}\) with replacement.

3) What is the distribution of $X^{(N)}$?

4) What are $E[X^{(N)}]$, $\text{var}[X^{(N)}]$? 

$= \text{popmean}$ 

$= \frac{\text{popvar}}{N}$
About the distribution of $X^{(N)}$

If $N \to \infty$, $X^{(N)} \sim \text{Normal}(\mu, \sigma)$

\[
\mu = \mathbb{E}[X^{(N)}] = \text{pop mean} \\
\sigma = \text{std}[X^{(N)}] = \frac{\text{pop std}}{\sqrt{N}}
\]

If $X^{(n)}$ is from a Normal like population,

\[T = \frac{\text{mean}\{x\} - \text{pop mean}}{\text{stderr}\{x\}} \sim t \text{ distribution with DOF } N-1\]

\[X^{(n)} = \text{mean}\{x\}\]

iid $X^{(n)}$
A tale of two statisticians

\{X\} = \{1, 2, 3, \ldots, 12\} \quad N_p = 12

The task: use only a subset of \{X\}: \{x\} with \(N=5\) to estimate the popmean(\{X\}) with some confidence report.
A tale of two statisticians

\[ \{ X \} = \{ 1, 2, 3, \ldots, 12 \} \quad N_p = 12 \]

\[ \{ X^b \} = \{ 1, 4, 5, 7, 11 \} \]
\[ \{ X^b_1 \} = \{ 1, 1, 4, 5, 7 \} \]
\[ \{ X^b_2 \} = \{ 4, 5, 7, 7, 13 \} \]
\[ \{ X^b_n \} = \{ 5, 5, 5, 5, 5 \} \]

\[ \text{if } N \to \infty \]
\[ X^{(N)} \sim N(\mu, \sigma) \]
\[ \mu = E[X^{(N)}] = \text{mean}\{x\} \]
\[ \sigma = \text{stddev}\{x^{(N)}\} \]

Histogram of \( X^{(N)} \)

mean(\( X^{(b)} \)) = 18/5
Motivation of sampling: the poll example

This senate election poll tells us:
- The sample has 1211 likely voters
- Ms. Hyde-Smith has realized sample mean equal to 51%

What is the estimate of the percentage of votes for Hyde-Smith?

How confident is that estimate?
Expected value of one random sample is the population mean

🌟 Since each sample is drawn uniformly from the population

\[ E[X^{(1)}] = \text{popmean}(\{X\}) \]

therefore

\[ E[X^{(N)}] = \text{popmean}(\{X\}) \]

🌟 We say that \( X^{(N)} \) is an unbiased estimator of the population mean.

\[ E[X^{(N)}] \approx \text{mean}(\{x_i\}) \]
Standard deviation of the sample mean

- We can also rewrite another result from the lecture on the weak law of large numbers

\[ \text{var}[X^{(N)}] = \frac{\text{popvar}\{X\}}{N} \]

- The standard deviation of the sample mean

\[ \text{std}[X^{(N)}] = \frac{\text{popsd}\{X\}}{\sqrt{N}} \]

- But we need the population standard deviation in order to calculate the \( \text{std}[X^{(N)}] \)!
Unbiased estimate of population standard deviation & Stderr

The unbiased estimate of $\text{popsd}(\{X\})$ is defined as

$$\text{stdunbiased}(\{x\}) = \sqrt{\frac{1}{N - 1} \sum_{x_i \in \text{sample}} (x_i - \text{mean}(\{x_i\}))^2}$$

So the **standard error** is an estimate of

$$\text{std}[X^{(N)}] = \frac{\text{popsd}(\{X\})}{\sqrt{N}} \approx \text{stderr}(\{x\})$$

$$\frac{\text{popsd}(\{X\})}{\sqrt{N}} = \frac{\text{stdunbiased}(\{x\})}{\sqrt{N}} = \text{stderr}(\{x\})$$
**Standard error: election poll**

What is the estimate of the percentage of votes for Hyde-Smith? 51%

Number of sampled voters who selected Ms. Smith is: 1211(0.51) ≈ 618

Number of sampled voters who didn’t select Ms. Smith was 1211(0.49) ≈ 593
Standard error: election poll

\( \text{stdunbiased}(\{x\}) \)

\[
= \sqrt{\frac{1}{1211 - 1} \left( 618(1 - 0.51)^2 + 593(0 - 0.51)^2 \right) } = 0.5001001
\]

\( \text{stderr}(\{x\}) \)

\[
\approx \frac{0.5}{\sqrt{1211}} \approx 0.0144
\]

\[
= \frac{\text{stdunbiased}(\{x\})}{\sqrt{N}} \quad N = 1211
\]
Interpreting the standard error

- **Sample mean** is a random variable and has its own probability distribution, stderr is an estimate of sample mean’s standard deviation.

- When \( N \) is very large, according to the **Central Limit Theorem**, sample mean is approaching a normal distribution with

\[
\mu = \text{popmean}\left(\{X\}\right) ; \quad \sigma \approx \frac{\text{popsd}\left(\{X\}\right)}{\sqrt{N}} \approx \frac{\text{stderr}\left(\{x\}\right)}{\sqrt{N}}
\]

\[
\text{stderr}\left(\{x\}\right) = \frac{\text{stdunbiased}\left(\{x\}\right)}{\sqrt{N}}
\]
Interpreting the standard error

Probability distribution of sample mean tends normal when $N$ is large

99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation

Credit: wikipedia

Population mean

$\mu$ + Standard error
Confidence intervals

Confidence interval for a population mean is defined by fraction

Given a percentage, find how many units of stderr it covers.

For 95% of the realized sample means, the population mean lies in [sample mean - 2 stderr, sample mean + 2 stderr]

\[ \frac{\text{sample mean} - 2 \times \text{stderr}}{\text{sample mean} + 2 \times \text{stderr}} \]
Confidence intervals when N is large

- For about 68% of realized sample means
  \[ \text{mean}\{x\} - \text{stderr}\{x\} \leq \text{popmean}\{X\} \leq \text{mean}\{x\} + \text{stderr}\{x\} \]

- For about 95% of realized sample means
  \[ \text{mean}\{x\} - 2\text{stderr}\{x\} \leq \text{popmean}\{X\} \leq \text{mean}\{x\} + 2\text{stderr}\{x\} \]

- For about 99.7% of realized sample means
  \[ \text{mean}\{x\} - 3\text{stderr}\{x\} \leq \text{popmean}\{X\} \leq \text{mean}\{x\} + 3\text{stderr}\{x\} \]
Q. Confidence intervals

What is the 68% confidence interval for a population mean?

A. [sample mean - 2*stderr, sample mean + 2*stderr]
B. [sample mean - stderr, sample mean + stderr]
C. [sample mean - std, sample mean + std]

[highlighted answer: B]
We estimate the population mean as 51% with stderr 1.44%.

The 95% confidence interval is

\[ [51\%-2\times1.44\%, \ 51\%+2\times1.44\%] = [48.12\%, \ 53.88\%] \]
Q.

A store staff mixed their fuji and gala apples and they were individually wrapped, so they are indistinguishable. If I pick 30 apples and found 21 fuji, what is my 95% confidence interval to estimate the popmean is 70% for fuji? (hint: stderr > 0.05)

[A. [0.7-0.17, 0.7+0.17]
B. [0.7-0.056, 0.7+0.056]
What if N is small? When is N large enough?

If samples are taken from normal distributed population, the following variable is a random variable whose distribution is Student’s t-distribution with $N-1$ degree of freedom.

$$T = \frac{\text{mean}\{x\} - \text{popmean}\{X\}}{\text{stderr}\{x\}} \approx \frac{\text{mean}\{x\} - \text{popmean}}{\text{stderr}\{x\}}$$

Degree of freedom is $N-1$ due to this constraint:

$$\sum_{i}(x_i - \text{mean}\{x\}) = 0$$
t-distribution is a family of distributions with different degrees of freedom.

- t-distribution with $N=5$
- t-distribution with $N=30$

![Graph showing pdf of t-distribution](image)

Credit: wikipedia

William Sealy Gosset 1876-1937
When $N=30$, t-distribution is almost Normal

t-distribution looks very similar to normal when $N=30$.

So $N=30$ is a rule of thumb to decide $N$ is large or not
Confidence intervals when \( N < 30 \)

- If the sample size \( N < 30 \), we should use t-distribution with its parameter (the degrees of freedom) set to \( N-1 \)

\[ t\text{-distri. is also symmetric.} \]
Centered Confidence intervals

- Centered Confidence interval for a population mean by \( \alpha \) value, where

\[
P(T \geq b) = \alpha
\]

\[
P\left(\frac{\text{mean} \times \frac{1-p}{s\text{tdm}}} {\text{stderr}} \geq b\right) = \alpha
\]

For \( 1-2\alpha \) of the realized sample means, the population mean lies in

\([\text{sample mean} - b \times \text{stderr}, \text{sample mean} + b \times \text{stderr}]\)
2d Confidence Interval

\[
P\left( \frac{\text{mean}(\{x\}) - \text{popmean}}{\text{stderr}(\{x\})} \right) \geq b \]

\[= P\left( \text{popmean} \leq \text{mean}(\{x\}) + b \cdot \text{stderr}(\{x\}) \right)\]

\[
P\left( \frac{\text{mean}(\{x\}) - \text{popmean}}{\text{stderr}(\{x\})} \right) \leq -b \]

\[= P\left( \text{popmean} \geq \text{mean}(\{x\}) - b \cdot \text{stderr}(\{x\}) \right)\]

\[\alpha = 5\% \Rightarrow 1 - 2\alpha = 90\%\]
The 95% confidence interval for a population mean is equivalent to what $1-2\alpha$ interval?

A. $\alpha = 0.05$

B. $\alpha = 0.025$

C. $\alpha = 0.1$
Sample statistic

- A **statistic** is a function of a dataset
  - For example, the mean or median of a dataset is a statistic

- **Sample statistic**
  - Is a statistic of the data set that is formed by the realized sample
  - For example, the realized sample mean
Q. Is this a sample statistic?

The largest integer that is smaller than or equal to the mean of a sample

A. Yes  
B. No.
Q. Is this a sample statistic?

The interquartile range of a sample

A. Yes

B. No.
Confidence intervals for other sample statistics

- **Sample statistic** such as *median* and others are also interesting for drawing conclusion about the population.

- It’s often difficult to derive the analytical expression in terms of stderr for the corresponding random variable.

- So we can use simulation...
Bootstrap is a method to construct confidence interval for *any* sample statistics using resampling of the sample data set.

Bootstrapping is essentially uniform random sampling with replacement on the sample of size $N$. 
Bootstrap for confidence interval of other sample statistics

Credit: E S. Banjanovic and J. W. Osborne, 2016, PAREonline

Figure 1. Summary of Bootstrapping Process
The realized sample of student attendance
{12, 10, 9, 8, 10, 11, 12, 7, 5, 10}, \( N=10 \), median=10

Generate a random index uniformly from [1,10] that correspond to the 10 numbers in the sample, ie. if index=6, the bootstrap sample’s number will be 11.

Repeat the process 10 times to get one bootstrap sample

<table>
<thead>
<tr>
<th>Bootstrap replicate</th>
<th>Sample median</th>
</tr>
</thead>
<tbody>
<tr>
<td>{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}</td>
<td>10</td>
</tr>
</tbody>
</table>
The realized sample of student attendance \{12,10,9,8,10,11,12,7,5,10\}, \(N=10\), median=10

<table>
<thead>
<tr>
<th>Bootstrap replicate</th>
<th>Sample median</th>
</tr>
</thead>
<tbody>
<tr>
<td>{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}</td>
<td>10</td>
</tr>
<tr>
<td>{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}</td>
<td>10</td>
</tr>
<tr>
<td>{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}</td>
<td>8.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Q. How many possible bootstrap replicates?

A. $10^{10}$  B. $10!$  C. $e^{10}$

\[ \frac{3!}{2!} \]

<table>
<thead>
<tr>
<th>Bootstrap replicate</th>
<th>Sample median</th>
</tr>
</thead>
<tbody>
<tr>
<td>${11, 11, 12, 10, 10, 10, 12, 10, 7, 10}$</td>
<td>10</td>
</tr>
<tr>
<td>${7, 10, 10, 10, 9, 7, 9, 10, 12, 10}$</td>
<td>10</td>
</tr>
<tr>
<td>${9, 7, 10, 8, 5, 10, 7, 10, 12, 8}$</td>
<td>8.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example of Bootstrap for confidence interval of sample median

- Do the bootstrapping for $r = 10000$ times, then draw the histogram and also find the stderr of sample median.

<table>
<thead>
<tr>
<th>Bootstrap replicate</th>
<th>Sample median</th>
</tr>
</thead>
<tbody>
<tr>
<td>{11, 11, 12, 10, 10, 10, 12, 10, 7, 10}</td>
<td>10</td>
</tr>
<tr>
<td>{7, 10, 10, 10, 9, 7, 9, 10, 12, 10}</td>
<td>10</td>
</tr>
<tr>
<td>{9, 7, 10, 8, 5, 10, 7, 10, 12, 8}</td>
<td>8.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example of Bootstrap for confidence interval of sample median

- Bootstrapping for \( r = 10000 \) times, then draw the histogram and also find the stderr of sample median.

\[
\text{stderr}([S]) = \sqrt{\frac{\sum_i [S(\{x\}_i) - \bar{S}]^2}{r - 1}}
\]

mean(Sample Median) = 9.73625
stderr(Sample Median) = 0.7724446
Errors in Bootstrapping

- The distribution simulated from bootstrapping is called empirical distribution. It is not the true population distribution. **There is a statistical error.**

- The number of bootstrapping replicates may not be enough. **There is a numerical error.**

- When the statistic is not a well behaving one, such as maximum or minimum of a data set, the bootstrap method may fail to simulate the true distribution.
The realized sample of CEO salary $N=59$, median=350 K

$r = 10000$

$\text{mean(Sample Median)} = 348.0378$

$\text{stderr(Sample Median)} = 27.30539$
The realized sample of CEO salary $N=59$, median=350 K

$r = 10000$

mean(Sample Median) = 348.0378
stderr(Sample Median) = 27.30539
Checking whether it’s normal by Normal Q-Q plot

- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph.
- Linear means similar to normal!

Read Pg 64, 3.2.3, “Introductory statistics with R”
Checking whether it’s normal by Normal Q-Q plot

- Q-Q compares a distribution with normal by matching the kth smallest quantile value pairs and plot as a point in the graph.

- Linear means similar to normal!

Read Pg 64, 3.2.3, “Introductory statistics with R”
CEO salary sample median’s Q-Q plot

- Q-Q plot of CEO salary’s bootstrap sample medians
- It’s roughly linear so it’s close to normal.
- We can use the normal distribution to construct the confidence intervals
95% confidence interval for the median CEO salary from the bootstrap simulation

348.0378 ± 2 × 27.30539

= [293.427, 402.6486]
Assignments

- Read Chapter 7 of the textbook
- Week 8 module on Compass
- Next time: hypothesis testing
Additional References

- Charles M. Grinstead and J. Laurie Snell
  "Introduction to Probability"

- Morris H. Degroot and Mark J. Schervish
  "Probability and Statistics"
See you next time

See you!