“In statistics we apply probability to draw conclusions from data.”
---Prof. J. Orloff

Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 3.09.2021
Last time

- Exponential distribution
- Normal (Gaussian) distribution
Objectives

- Sample mean
- Confidence interval
- t-distribution
Motivation for drawing conclusion from samples

* In a study of new-born babies’ health, random samples from different times, places and different groups of people will be collected to see how the overall health of the babies is like.
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?
Population

What is a population?

- It’s the entire possible data set \( \{ X \} \)
- It has a countable size \( N_p \)
- The population mean \( \text{popmean}(\{ X \}) \) is a number
- The population standard deviation is \( \text{popsd}(\{ X \}) \) and is also a number

The population mean and standard deviation are the same as defined previously in chapter 1
\[ \{X\} = \{1, 2, 3, \ldots , 12\} \quad N_p = 12 \]

\[
\text{popmean}(\{X\}) = ? \\
\text{popstd}(\{X\}) = ? \\
\sqrt{\frac{\sum (X_i - 6.5)^2}{12}} \\
3.4\ldots
\]
Sample

- The sample is a random subset of the population and is denoted as \( \{x\} \), where sampling is done with replacement.

- The sample size \( N \) is assumed to be much less than population size \( N_p \).

- The sample mean of a population is \( X^{(N)} \) and is a random variable.
Sample mean example

Money box

* shake and take one and put back.

\[ X_1 \text{ takes } x_1 = 10 \]
\[ X_2 \text{ takes } x_2 = 10 \]
\[ X_3 \text{ takes } x_3 = 25 \]
\[ \vdots \]
\[ X_N \text{ takes } x_N = 10 \]

\[ E[X] = ? \]

\[ \bar{X}^{(N)} = \frac{\sum X_i}{N} \]
Sample \( \{ X \} \) and Sample Mean \( X^{(N)} \)

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

One random sample:

\[
\{ x \} = \{ 1, 1, 2, 3, 3 \} \quad N = 5
\]

\( X^{(N)} \) RV takes a value?

\[
X^{(N)} = \frac{X_1 + X_2 + \cdots + X_N}{N}
\]

Another random sample:

\[
\{ 1, 1, 1, 1, 1 \} \Rightarrow X^{(N)} = 1
\]

\( N < N_p \)
Sample mean of a population

- The sample mean is the average of IID samples
  \[ X^{(N)} = \frac{1}{N} (x_1 + x_2 + \ldots + x_N) = \text{mean} \{ x \} \]

- By linearity of the expectation and the fact the sample items are identically drawn from the same population with replacement
  \[ E[X^{(N)}] = \frac{1}{N} (E[X^{(1)}] + E[X^{(1)}] + \ldots + E[X^{(1)}]) = E[X^{(1)}] \]
Expected value of one random sample is the population mean

Since each sample is drawn uniformly from the population

\[ E[X^{(1)}] = \text{popmean}(\{X\}) = \frac{1}{N_p} \sum X_i \frac{1}{N_p} = \text{popmean} \]

therefore

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

We say that \( X^{(N)} \) is an unbiased estimator of the population 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 \( \text{std}[X^{(N)}] \)!

\[ \frac{1}{N} \cdot \text{var}(\sum x_i) = \sum \text{var}(\text{RV}_i) \]

if \( \text{RV}_i \) are indpt.
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 $\text{approximation}$

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

$$\frac{\text{popsd}(\{X\})}{\sqrt{N}} \approx \frac{\text{stdunbiased}(\{x\})}{\sqrt{N}} = \text{stderr}(\{x\})$$
What is the estimate of the percentage of votes for Hyde-Smith?

Number of sampled voters who selected Ms. Smith is:

\[ 1211(0.51) \approx 618 \]

Number of sampled voters who didn’t select Ms. Smith was

\[ 1211(0.49) \approx 593 \]
Standard error: election poll

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

- \( \text{stderr}(\{x\}) \)
  \[
  = \frac{0.5001001}{\sqrt{1211}} \approx 0.0144
  \]

\( N = 1211 \)

\( X = \frac{x_1 + \ldots + x_N}{N} \)

\( 618 \) "i" for Smith

\( 593 \) "o" not for her
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

\[
\mu = \text{popmean}\{X\} \ ; \ \sigma = \frac{\text{popsd}\{X\}}{\sqrt{N}} = \text{stderr}\{x\}
\]

\[
\text{stderr}\{x\} = \frac{\text{std unbiased}\{x\}}{\sqrt{N}}
\]

\[
pdf: f_{X}\left(x\right) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}
\]
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% of the data are within 2 standard deviations.

68% of the data are within 1 standard deviation.

\( \mu \) = mean of \( x \)

\( \sigma \) = std.err (mean)

Credit: wikipedia

if \( N \) is large

\( X \sim N(\mu, \sigma^2) \)
Confidence intervals

- Confidence interval for a population mean is defined by fraction $c \% = 95\%$
- 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]
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
  \[ \text{mean}(\{x\}) - 3\text{stderr}(\{x\}) \leq \text{popmean}(\{X\}) \leq \text{mean}(\{x\}) + 3\text{stderr}(\{x\}) \]
What is the 68% confidence interval for a population mean?

A. [sample mean-2stderr, sample mean+2stderr]
B. [sample mean-stderr, sample mean+stderr]
C. [sample mean-std, sample mean+std]
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: strerr > 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.

$$T = \frac{\text{mean}\left(\{x\}\right) - \text{popmean}\left(\{X\}\right)}{\text{stderr}\left(\{x\}\right)}$$

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

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

The t-distribution with $N=5$ and $N=30$ is shown in the graph.

威廉·西利·高赛特 (William Sealy Gosset, 1876-1937)
When \( N=30 \), \( t \)-distribution is almost Normal

distribution looks very similar to normal when \( N=30 \).

So \( N=30 \) is a rule of thumb to decide \( N \) is large or not
Assignments

- Read Chapter 7 of the textbook
- Next time: Bootstrap, Hypothesis tests
- Prepare for Midterm1
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!