Lecture A3. Statistics Review

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

SEOULTECH 서울과학기술대학교 데이터사이언스학과

Population and Sample

- A population set (모집단) is the entire group that you want to draw conclusions about.
- A sample set (표본집단) is the subset of population that you have an access to collect data from.
- The size of the sample is always less than the total size of the population.
- It is a researcher's primary concern to draw conclusion on the population set, by studying the behavior from the sample set.

Population statistics

- Population
  - Suppose that you are interested in Korean male's hand length. Let X be a distribution of population set (entire Korean male's hand length).
  - Let μ be the mean of X and σ^2 be the variance of X.
  - That is, $\mu = \mathbb{E}X$ and $\sigma^2 = \mathbb{E}[(X - \mathbb{E}X)^2]$.
  - These *population statistics* are what we are after, specifically, *population mean* and *population variance*.
  - Since these are what we aim to estimate, we often call them as *true values*, specifically *true mean* and *true variance*.
- Sample
  - In order to estimate μ and σ^2, you collect n samples of Korean male's hand length.
  - Typically, these collected samples are denoted as $X_1, X_2, ..., X_n$, or $\{X_i, 1 \leq i \leq n\}$.

Sample statistics

- Estimation
  - You want to draw conclusions on the *population mean* (μ) and *population variance* (σ^2) by studying the sample $\{X_i, 1 \leq i \leq n\}$.
  - From the sample, we compute some value that should be similar to population statistics.
- Sample Mean
  - It is known that $\sum_{i=1}^{n} X_i/n$ is similar value to the population mean.
  - This quantity is typically notated as $\overline{X}$, i.e., $\overline{X} = \sum_{i=1}^{n} X_i/n$.
  - This quantity is called as *sample mean* for obvious reason.
  - Sample mean is obtained by taking an arithmetic average of all samples.
- Sample Variance
  - It is known that $\frac{\sum(X_i - \overline{X})^2}{n-1}$ is similar value to the population variance.
  - This quantity is typically notated as s^2, i.e. $s^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$.
  - This quantity is called as *sample variance* for obvious reason.
  - Sample variance is obtained by 1) summing up squared deviations of all samples and 2) divide it by $n - 1$.

- Summary

	Mean	Variance
Population	$\mu = \mathbb{E}X$	$\sigma^2 = \mathbb{E}[(X - \mathbb{E}X)^2]$
Sample	$\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$	$s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$

Estimation

- Remind that it is mentioned that 'Sample mean is *believed to be a similar value* to the population mean'.
- Like such, we call the process of 'Finding sample statistics that is *believed to be a similar value* to the population statistics.' as *estimation*.

- For true mean μ, there may be various estimation efforts that aims to find similar value to the μ. We call these *similar value to the true value*, as an *estimator*.
- Again, *estimator* is not a true value, but an estimation effort. To distinguish between the *true value* and *estimator*. Notation of 'hat' is typically used. For example, $\hat{\mu}$ indicates an estimator for μ, and $\hat{\sigma}^2$ indicates an estimator for σ^2.

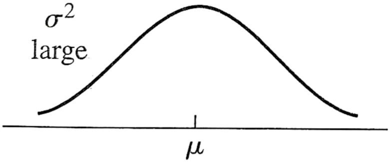
- Sample mean serves as an estimator for the true mean.
- Sample variance serves as an estimator for the true variance.

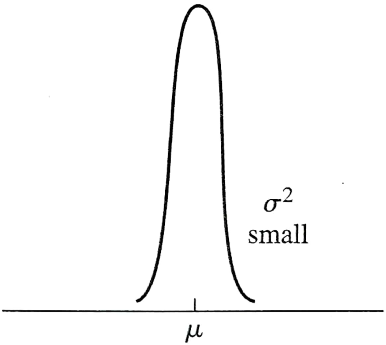
Desired properties of estimators

- Is $\frac{\sum_{i=1}^{n} X_i}{n}$ a good estimator for the true mean? What it means by *good*?
- There are many criteria for *good* estimator such as
  - *unbiased* estimator - Expected value of estimator must be same as true value.
  - *consistent* estimator - As the number of sample increases, the estimator converges to the true value.
  - *maximum-likelihood (ML) estimator* - The probability that the estimator is exactly equal to true value is maximal.
- For mathematical expression, let's notate the true statistics we are after as θ, and the estimator as $\hat{\theta}$. Then,
  - $\hat{\theta}$ is an *unbiased* estimator if $\mathbb{E}\hat{\theta} = \theta$.
  - $\hat{\theta}$ is a *consistent* estimator if $\hat{\theta} \to \theta$ as $n \to \infty$.
  - $\hat{\theta}$ is a *maximum-likelihood (ML)* estimator if $\hat{\theta} = argmax_x \mathbb{P}(\theta = x)$.

- It is known that
  - $\frac{\sum_{i=1}^{n} X_i}{n}$ is an *unbiased*, *consistent*, and *maximum-likelihood* estimator for the true mean.
  - $\frac{\sum (X_i - \overline{X})^2}{n-1}$ is an *unbiased* and *consistent* estimator for the true variance, but it is not a *maximum-likelihood* estimator.
  - $\frac{\sum (X_i - \overline{X})^2}{n}$ is a *consistent* and *maximum-likelihood* estimator for the true variance, but it is not an *unbiased* estimator. In other words, it is *biased* estimator.

- Normal variable $X \sim N(\mu, \sigma^2)$





Central limit theorem (CLT)

Theorem 1

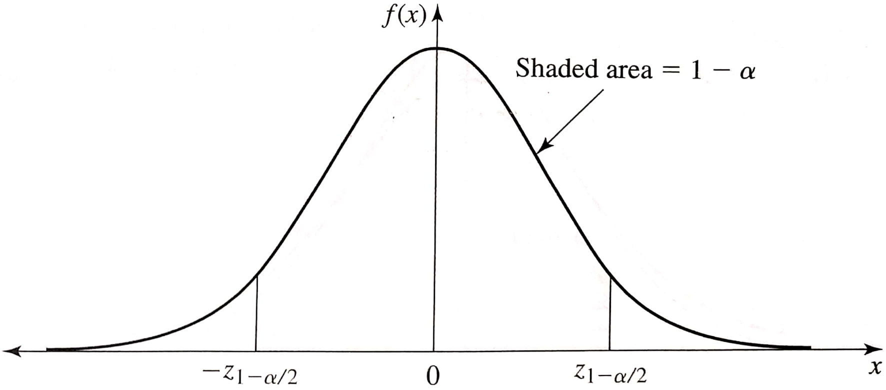
For a random variable X, whatever the distribution of X is, its sample mean $\overline{X}$ follows a normal distribution as long as the number of samples n is larger than 30. That is

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

- It is intriguing that the population distribution may not be a normal distribution, but the sample mean from the population will always follow a normal distribution as long as the number of sample is larger than 30.
- It is also intriguing that the uncertainty of closeness between the estimator and true value is nicely quantified with the variance σ^2/n.

- Questions
  1. Is $\overline{X}$ an unbiased estimator for μ?
  2. Is $\overline{X}$ a consistent estimator for μ?
  3. Is $\overline{X}$ a ML estimator for μ?

Normal variable's quantile



Confidence interval

- From $\overline{X} \sim N(\mu, \sigma^2/n)$, we can use normal distribution's property to say:

$$\mathbb{P}[\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}] = 0.95$$

- Two issues with the above confidence interval.
  1. The above expression is a confidence interval for the estimator ($\overline{X}$), not for the true value (μ).
  2. We do not know the true value σ.

- To tackle the first issue, the following effort is made.

$$\begin{aligned}
\overline{X} \sim N(\mu, \sigma^2/n) \quad &\Rightarrow \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) = Z \\
&\Rightarrow \quad \frac{\mu - \overline{X}}{\sigma/\sqrt{n}} \sim Z \\
&\Rightarrow \quad \mu \sim N(\overline{X}, \sigma^2/n)
\end{aligned}$$

- From the last expression, $\mu \sim N(\overline{X}, \sigma^2/n)$, we still have the second issue of not knowing σ. We must replace σ with s.
- In replacing σ with s, it is known that $\frac{\mu-\overline{X}}{\sigma/\sqrt{n}} \sim Z$ becomes

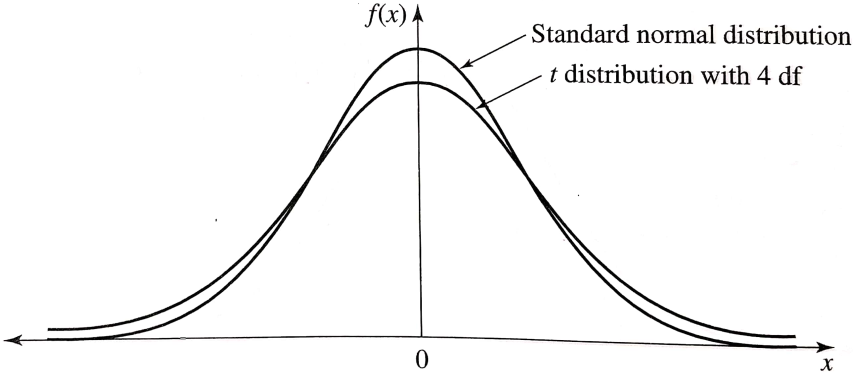
$$\frac{\mu - \overline{X}}{s/\sqrt{n}} \sim t_{n-1}$$

- Now we are ready to state the confidence interval for μ as following.

$$\mathbb{P}[\overline{X} - t_{0.975,n-1} \cdot s/\sqrt{n} \leq \mu \leq \overline{X} + t_{0.975,n-1} \cdot s/\sqrt{n}] = 0.95$$

- To get the some sence of what $t_{0.975,n-1}$ might be depending on n,
  - If $n = 30$, $\mathbb{P}[\overline{X} - 2.045 \cdot s/\sqrt{30} \leq \mu \leq \overline{X} + 2.045 \cdot s/\sqrt{30}] = 0.95$
  - If $n = 60$, $\mathbb{P}[\overline{X} - 2.000 \cdot s/\sqrt{60} \leq \mu \leq \overline{X} + 2.000 \cdot s/\sqrt{60}] = 0.95$
  - If $n = 120$, $\mathbb{P}[\overline{X} - 1.980 \cdot s/\sqrt{120} \leq \mu \leq \overline{X} + 1.980 \cdot s/\sqrt{120}] = 0.95$
  - If n is bigger, $\mathbb{P}[\overline{X} - 1.960 \cdot s/\sqrt{n} \leq \mu \leq \overline{X} + 1.960 \cdot s/\sqrt{n}] = 0.95$
- For the most applications in this course, n is so big enough that we are generally fine using 1.96.

Normal dist. vs t dist.



Exercise 1

You randomly sample 1,600 Korean male and measured their hand length. The sample mean is 20cm and the sample standard deviation is 2cm. What is the 95% confidence interval for Korean male's hand length?

```
"Man can learn nothing unless he proceeds from the known to the unknown. - Claude Bernard"
```