Section 7.1: Basic Properties of Confidence Interval
Preliminary

We have the follow three important distributions. They are all defined based on normal distributions.
Chi-square distribution ($\chi^2$-distribution)

- *(Definition).* Let $X_1, \cdots, X_n$ be independent $N(0, 1)$. Then, the distribution of

$$Y = \sum_{i=1}^{n} X_i^2$$

is called the $\chi^2$ distribution with $n$-degrees of freedom, denoted by $Y \sim \chi^2_n$.

- If $Y \sim \chi^2_n$, then $Y$ cannot be negative.
- If $Y \sim \chi^2_n$, then $E(Y) = n$ and $V(Y) = 2n$.
- If $X_1, \cdots, X_n$ are iid $N(0, \sigma^2)$, then

$$Y = \sum_{i=1}^{n} X_i^2 \sim \sigma^2 \chi^2_n.$$  

- We can use Table A.7 to find the value of $c$ for

$$P(\chi^2_n \geq c) = \alpha$$

for a given $\alpha$. The value of $c$ denoted by $\chi^2_{\alpha, n}$. For example, if we need

$$P(\chi^2_{10} \geq c) = 0.025,$$

then we have $c = \chi^2_{0.025, 10} = 20.483$. 
**t-distribution**

- **(Definition)** If $X \sim N(0,1)$ and $Y \sim \chi^2_n$ independently, then the distribution of

$$T = \frac{X}{\sqrt{Y/n}}$$

is called the $t$-distribution with $n$-degrees of freedom, denoted by $T \sim t_n$.

- If $n$ is large (e.g., $n \geq 40$), then $t_n \approx N(0,1)$.

- If $X \sim N(0,\sigma^2)$ and $Y \sim \sigma^2 \chi^2_n$ independently, then

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n.$$

- $t_n$ is symmetric about 0, i.e.

$$P(t_n \leq -c) = P(t_n \geq c).$$
• We can use Table A.5 to find the value of $c$ for

$$P(t_n \geq c) = \alpha$$

for a given $\alpha$. It is denoted by $t_{\alpha,n}$. For example, if $c$ satisfies

$$P(t_{10} \geq c) = 0.025,$$

then we have $c = t_{0.025,10} = 2.28$.

• By symmetry, we have

$$P(t_{10} \leq -2.28) = P(t_{10} \geq 2.28) = 0.025$$

and

$$P(|t_{10}| \geq 2.28) = 0.05.$$ 

• For large $n$, you can approximately use the normal table.
(Definition) If \( X \sim \chi^2_m \) and \( Y \sim \chi^2_n \) independently, then the distribution of
\[
F = \frac{X/m}{Y/n}
\]
is called the \( F \)-distribution with \( m \) degrees of freedom on the numerator (1st DF) and \( n \) degrees of freedom on the denominator (2nd DF), denoted by \( F \sim F_{m,n} \).

We can use Table A.9 to find \( c \) for \( P(F_{m,n} \geq c) = \alpha \) for a given \( \alpha \). It is denoted by \( F_{\alpha,m,n} \). For example, if \( P(F_{8,6} \geq c) = 0.05 \), then \( c = F_{0.05,8,6} = 4.15 \).

By the property, we have
\[
P(F_{m,n} \geq c) = P(F_{n,m} \leq 1/c).
\]
To find \( c \) such that \( P(F_{8,6} \geq c) = 0.95 \), we can use
\[
P(F_{8,6} \leq c) = 0.05 \Rightarrow P(F_{6,8} \geq 1/c) = 0.05.
\]
We obtain \( 1/c = 3.58 \). We have \( c = 1/3.58 = 0.279 \). Thus,
\[
P(0.279 \leq F_{8,6} \leq 4.15) = 0.9.
\]
• Let \( x_1, \cdots, x_n \) be observations.

• Suppose they are iid normal distributed, i.e., \( x_1, \cdots, x_n \) are observed values of iid random variables \( X_1, \cdots, X_n \) as we can write

\[
X_1, \cdots, X_n \sim iid \ N(\mu, \sigma^2).
\]

• We define the sample mean

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

and the sample variance

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]

We call \( s \) as the sample standard deviation.

• Then,

\[
\tilde{X} \sim N(\mu, \frac{\sigma^2}{n}).
\]

• Thus, we have

\[
P \left( -\frac{z_\alpha}{2} \leq \frac{\tilde{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{z_\alpha}{2} \right) = 1 - \alpha.
\]
Therefore, with probability $1 - \alpha$ there is

$$-z_\alpha^2 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha^2$$

which is equivalent to say that with probability $1 - \alpha$ there is

$$\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

We may take the above as the confidence interval, i.e, when $\sigma$ is known.

Therefore, we have the following formula: suppose $x_1, \cdots, x_n$ are iid observations of a normal population and assume the standard deviation $\sigma$ is known, then the $1 - \alpha$ level confidence interval for $\mu$ is

$$\bar{x} \pm z_\alpha \frac{\sigma}{\sqrt{n}} = [\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}].$$

The interpretation of the confidence interval is that if we repeat the procedure many many times, with probability $1 - \alpha$ the above confidence interval contains the true value of $\mu$. 
• An often asked question is about the length of confidence interval. How large is the sample size $n$ so that the $1-\alpha$ level confidence interval is less than $w$.

• Note that the length of the $1-\alpha$ level confidence interval is

$$2z_\alpha \frac{\sigma}{\sqrt{n}}.$$ 

Thus, we have

$$2z_\alpha \frac{\sigma}{\sqrt{n}} \leq w \Rightarrow n \geq \left(2z_\alpha \frac{\sigma}{w}\right)^2 = \frac{4z_\alpha^2 \sigma^2}{w^2}.$$
First example of Section 7.1: examples 7.1 and 7.2 on textbook.

- Study the preferred height for an experimental keyboard with large forearm-wrist support.

- Take sample size $n = 31$. Observed sample mean $\bar{x} = 80.0$ cm. We known the standard deviation $\sigma = 2.0$. Assume height is normally distributed.

- The 95% confidence interval is

$$\bar{x} \pm z_{0.025} \frac{\sigma}{\sqrt{n}} = 80.0 \pm 1.96 \times \frac{2.0}{\sqrt{31}} = [79.3, 80.7].$$
Second example of Section 7.1: example 7.3 on textbook.

- Study the hole diameters. Assume it is normally distributed.

- Historical data had suggested \( \sigma = 0.1 \text{mm} \). Assume it does not change.

- A sample of \( n = 40 \) units reported the sample mean \( \bar{x} = 5.426 \).

- The 90\% confidence interval is

\[
5.426 \pm 1.645 \times \frac{0.1}{\sqrt{40}} = [5.400, 5.452].
\]
Third example of Section 7.1: example 7.4 on textbook. Suppose the response time is normally distributed with $\sigma = 25$ms. How large the sample $n$ we need so that the 95% confidence interval is not greater than 10.

Answer:

$$n \geq \left(2 \times 1.96 \times \frac{25}{10}\right)^2 = 96.04$$

Therefore, we should choose $n = 97$ since the least integer $\geq 96.04$ is 97.