Sampling Distribution

Let $Y_1, \ldots, Y_n$ be independent identically distributed (i.i.d.) random variables. If we want to estimate the population mean $\mu$, we take sample $y_1, \ldots, y_n$, and use $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ to estimate $\mu$.

**Definition**: A statistic is a function of the observable random variables in a sample and known constants.

**Example**: $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

All statistics have probability distributions, which we call their sampling distributions.

**Example**: Toss a die three times.

$Y_1, Y_2, Y_3$ are 1's on the upper face of the die

$\bar{Y} = \frac{Y_1 + Y_2 + Y_3}{3}$

Find $E(\bar{Y}), \text{Var}(\bar{Y}), \text{dist} \text{h} \text{n} \text{ of} \bar{Y}$

$E(Y_i) = \sum y_i P(y) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6}$

$= 3.5$

$E(\bar{Y}) = \frac{1}{3} E(Y_1 + Y_2 + Y_3) = \frac{1}{3} \left[ E(Y_1) + E(Y_2) + E(Y_3) \right]$

$= \frac{1}{3} \cdot 3 \cdot 3.5 = 3.5$
\[ \text{Var}(\overline{Y}) = \frac{1}{g} \text{Var}(Y_1 + Y_2 + Y_3) \]
\[ = \frac{1}{g} \left[ \text{Var}(Y_1) + \text{Var}(Y_2) + \text{Var}(Y_3) \right] \]

\[ \text{Var}(Y_i) = \text{E}(Y_i^2) - \left( \text{E}(Y_i) \right)^2 \]

\[ \text{E}(Y_i^2) = \sum y^2 p(y) = 1^2 \frac{1}{6} + \ldots + 6^2 \frac{1}{6} \]

\[ \Rightarrow \frac{1}{6} \cdot 91 - 3.5^2 = 2.9167 \]

\[ \text{Var}(\overline{Y}) = \frac{1}{g} \cdot 3 \cdot 2.9167 = 0.9722 \]

\[ W = Y_1 + Y_2 + Y_3, \quad \overline{Y} = \frac{W}{3} \]

\[ W = 3, \ldots, 18 \]

\[ \text{P}(W = 3) = \text{P}(\overline{Y} = 1) = \text{P}(1, 1, 1) = \frac{1}{6^3} \]

\[ \text{P}(W = 4) = \text{P}(\overline{Y} = \frac{Y}{3}) = \text{P}(1, 1, 2) + \text{P}(1, 2, 1) \]
\[ + \text{P}(2, 1, 1) = \frac{3}{6^3} \]

\[ \text{P}(W = 5) = \text{P}(\overline{Y} = \frac{5}{3}) = \text{P}(1, 1, 3) + \]
\[ + \text{P}(1, 3, 1) + \text{P}(3, 1, 1) + \text{P}(1, 2, 2) \]
\[ + \text{P}(2, 1, 2) + \text{P}(2, 2, 1) = \frac{6}{6^3} \ldots \]
**Theorem:** Let \( Y_1, \ldots, Y_n \sim iid N(\mu, \sigma^2) \). Then \( \bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \).

**Note:** \( Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \).

**Proof:**

\[
\bar{Y} = \frac{1}{n} \left( Y_1 + \ldots + Y_n \right)
\]

\[
E(\bar{Y}) = \frac{1}{n} E(Y_1 + \ldots + Y_n)
= \frac{1}{n} \left[ E(Y_1) + \ldots + E(Y_n) \right], E(Y_i) = \mu
= \frac{1}{n} \cdot n \mu = \mu
\]

\[
Var(\bar{Y}) = \frac{1}{n^2} Var(Y_1 + \ldots + Y_n)
= \frac{1}{n^2} \left[ Var(Y_1) + \ldots + Var(Y_n) \right]
= \frac{1}{n^2} \cdot n \sigma^2 = \frac{\sigma^2}{n}
\]

\[
\Rightarrow \bar{Y} \sim N\left( \mu, \frac{\sigma^2}{n} \right)
\]
Example: (#7.11) A forester studying the effects of fertilization on certain pine forests is interested in estimating the average basal area of pine trees. In studying basal areas, he has discovered that these measurements are normally distributed with standard deviation appr. 4 sq. inches. If the forester samples $n = 9$ trees, find the probability that the sample mean will be within 2 sq. in. of the population mean.

Solution:

$$\Pr \left( \left| \bar{Y} - \mu \right| \leq 2 \right)$$

$\bar{Y} = \bar{Y}$, $\text{Var} (\bar{Y}) = \frac{\sigma^2}{n}$, $s_{\bar{Y}} = \text{St. dev.}$

$$s_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{9}} = \frac{4}{3}$$

$$\Pr \left( \left| \bar{Y} - \mu \right| \leq 2 \right) = \Pr \left( -2 \leq \frac{\bar{Y} - \mu}{s_{\bar{Y}}} \leq 2 \right)$$

$$= \Pr \left( -\frac{2}{\frac{4}{3}} \leq Z \leq \frac{2}{\frac{4}{3}} \right) \quad Z \sim \mathcal{N}(0,1)$$

$$= \Pr \left( -1.5 \leq Z \leq 1.5 \right)$$

$$= 1 - 2 \Pr (Z > 1.5) = 0.8664$$
Theorem: Let $Y_1, ..., Y_n \sim iid \ N(\mu, \sigma^2)$. Then

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \text{ and } \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$ 

Proof: done.

Example: $Z_1, ..., Z_6 \sim iid \ N(0,1)$. Find a number $b$ such that $P\left(\sum_{i=1}^{6} Z_i^2 \leq b\right) = 0.95$.

Solution: \[ X = \sum_{i=1}^{6} Z_i^2 \sim \chi_6^2 \]

$$P\left( X \leq b \right) = 0.95$$

from the table

\[ b = 12.59 \]
Suppose we want to make an inference about the population variance $\sigma^2$ based on $Y_1, \ldots, Y_n$ from normal population. We estimate $\sigma^2$ by the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

**Theorem:** Let $Y_1, \ldots, Y_n \sim iid \ N(\mu, \sigma^2)$. Then $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$. Also $\bar{Y}$ and $s^2$ are independent random variables.

**Proof:**

$$\bar{Y} = \frac{Y_1 + Y_2}{2}, \quad s^2 = \frac{1}{2-1} \sum_{i=1}^{2} (Y_i - \bar{Y})^2$$

$$= (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 = (Y_1 - \frac{Y_1 + Y_2}{2})^2 + (Y_2 - \frac{Y_1 + Y_2}{2})^2$$

$$= 2 \left( \frac{Y_1}{2} - \frac{Y_2}{2} \right)^2 = \frac{(Y_1 - Y_2)^2}{2}$$

$$\frac{(n-1)s^2}{\sigma^2} = \frac{(Y_1 - Y_2)^2}{2 \sigma^2} = \frac{\sqrt{Y_1 - Y_2}}{\sqrt{2 \sigma^2}}$$

$Y_1 \sim N(\mu, \sigma^2)$

$Y_2 \sim N(\mu, \sigma^2)$

$\Rightarrow Y_1 - Y_2 \sim N(0, 2\sigma^2)$
So, \[\frac{Y_1 - Y_2 - 0}{\sqrt{26^2}} \sim N(0,1)\]

\[\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_1, \quad n = 2\]

---

**Fact.** \(Y_1, Y_2 \sim N(\mu, \sigma^2)\), independent

Then \(Y_1 + Y_2, Y_1 - Y_2\) are independent

\[\overline{Y} = \frac{Y_1 + Y_2}{2}\quad \text{is a function of } Y_1 + Y_2\]

\[s^2 = \frac{(Y_1 - Y_2)^2}{2}\quad \text{is a function of } Y_1 - Y_2\]

\[\therefore \overline{Y} \text{ is independent of } s^2.\]
Definition: Let $Z \sim N(0,1)$ and $W \sim \chi^2$. Then, if $Z$ and $W$ are independent, $T = \frac{Z}{\sqrt{W/\nu}}$ is said to have a $t$ distribution with $\nu$ degrees of freedom.

\[
Y_1, \ldots, Y_n \sim \text{i.i.d. } N(\mu, \sigma^2)
\]

\[
Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)
\]

\[
W = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}
\]

\[
T = \frac{Z}{\sqrt{W/\nu}} = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{Y} - \mu}{\frac{s}{\sqrt{n}}} \sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)} \sim t_{n-1}
\]
Example: (#7.30) Let $Z \sim N(0,1)$ and $W \sim \chi^2_v$. $Z$ and $W$ are independent, then $T = \frac{z}{\sqrt{W/v}} \sim t_v$.

(a) Find $E(Z^2)$;

(b) Given $E(W^\alpha) = \frac{\Gamma(\frac{\nu}{2} + a)}{\Gamma(\frac{\nu}{2})} \cdot 2^\alpha$, $\nu > -2a$.

Show that $E(T) = 0$, $\nu > 1$ and $Var(T) = \frac{\nu}{\nu - 2}$, $\nu > 2$;

Solution:

(a) $\text{Var}(Z) = 1 = E(Z^2) - E(Z)^2$

$\Rightarrow E(Z^2) = 1$

(b) $E(T) = E\left(\frac{Z}{\sqrt{W/\nu}}\right)$

$= E(2 \cdot W^{-\frac{1}{2}}) \cdot \sqrt{\nu} = E(2) \cdot E(W^{-\frac{1}{2}})$

$= 0$, $\nu > -2a$, $a = -\frac{1}{2} \Rightarrow \nu > 1$

$\text{Var}(T) = E(T^2) - E(T)^2$

$= E(T^2) = E(2^2 \cdot W^{-\frac{1}{2}}$)
\[ E(T^2) = 1 \cdot E(\frac{v^2}{\sigma^2}) E(W^{-1}) \]

\[ = 1 \cdot 1 \cdot \frac{\Gamma\left(\frac{\nu}{2} - 1\right)}{\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{\nu^{-\frac{\nu}{2}-1}}{\frac{\nu}{2}-1}, \nu > 2 \]

\[ \Gamma(d) = (d-1) \Gamma(d-1) \]

\[ \Gamma(d-1) = \frac{\Gamma(d)}{d-1} \]

\[ = 1 \cdot \frac{\Gamma\left(\frac{\nu}{2}\right)}{\frac{\nu}{2}-1} \cdot \frac{1}{\frac{\nu}{2}} \]

\[ = \frac{\nu}{\nu - 2}, \nu > 2 \]
Definition: Let $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$ be independent. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2) \text{ or } F_{\nu_1,\nu_2} \text{ (F distribution)}$$

Suppose we want to compare the variances of two normal populations. We take a sample of size $n_1$ from one population, and a sample of size $n_2$ from another.

We look at $\frac{s_1^2}{s_2^2}$.

Theorem: $\frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$, if $\sigma_1^2 = \sigma_2^2$.

Proof:

$$W_1 = \frac{(n_1 - 1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1 - 1}$$

$$W_2 = \frac{(n_2 - 1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2 - 1}$$

$$F = \frac{W_1/n_1 - 1}{W_2/n_2 - 1} = \frac{(n_1 - 1)s_1^2}{\sigma_1^2/n_1 - 1} \sim \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$
Example: If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal population variances, find $b$ such that $P\left(\frac{s_1^2}{s_2^2} \leq b\right) = 0.95$.

Solution: 
\[ \sigma_1^2 = \sigma_2^2 \implies F = \frac{S_1^2}{S_2^2} \sim F(5, 9) \]

\[ P\left(F \leq b\right) = 0.95 \quad P\left(F \geq b\right) = 0.05 \]

From the table, $b = 3.48$

Example: (#7.34) Show $E(F) = \frac{\nu_2}{\nu_2 - 2}, \nu_2 > 2$

\[ \sqrt{\text{Var}(F)} = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \nu_2 > 4 \]

Solution:
\[ E(\chi^2) = \frac{\Gamma\left(\frac{1}{2} + \alpha\right)}{\Gamma\left(\frac{1}{2}\right)} \cdot 2^\alpha, \sqrt{\alpha} > -2 \alpha \]

\[ E(F) = E\left(\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}\right) = \frac{\nu_2}{\nu_1} E\left(\chi_1^2\right) E\left(\chi_2^{-1}\right) = \frac{\nu_2}{\nu_1} \cdot \frac{\Gamma\left(\frac{\nu_1}{2} + 1\right)}{\Gamma\left(\frac{\nu_1}{2}\right)} \cdot 2^{\frac{\nu_1}{2}} \cdot \frac{\Gamma\left(\frac{\nu_2}{2} - 1\right)}{\Gamma\left(\frac{\nu_2}{2}\right)} \cdot 2 = \]
\[
\begin{align*}
\Gamma \left( \frac{1}{2} \right) &= \frac{\Gamma \left( \frac{1}{2} \right) \cdot \sqrt{2}}{\Gamma \left( \frac{1}{2} \right)} \\
&= \frac{\sqrt{2}}{\frac{\sqrt{2}}{2} - 1} = \frac{\sqrt{2}}{\sqrt{2} - 2} \\
\text{Var}(F) &= E(F^2) - E(F)^2 \\
E(F^2) &= E \left( \frac{W_1^2 / V_1}{W_2^2 / V_2} \right) = \frac{V_2^2}{V_1} E(W_1^2) E(W_2^{-2}) \\
&= \frac{V_2^2}{V_1^2} \frac{\Gamma \left( \frac{1}{2} + 2 \right)}{\Gamma \left( \frac{1}{2} \right)} \frac{\Gamma \left( \frac{V_1}{2} - 2 \right)}{\Gamma \left( \frac{V_2}{2} - 2 \right)} \\
\Gamma(2 + 2) &= \Gamma(2) \cdot 2(2 + 1) \\
&= \frac{V_2^2}{V_1^2} \frac{\frac{1}{2} \left( \frac{1}{2} + 1 \right)}{\left( \frac{1}{2} - 1 \right) \left( \frac{1}{2} - 2 \right)} \\
\text{Var}(F) &= \frac{\sqrt{2}}{V_1} \frac{(V_1 + 2)}{V_1 (V_2 - 2)(V_2 - 4)} - \frac{\sqrt{2}}{(V_2 - 2)^2}
\end{align*}
\]
Central Limit Theorem

We already showed that if $Y_1, \ldots, Y_n$ is a random sample from any distribution with mean $\mu$ and variance $\sigma^2$, then $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \frac{\sigma^2}{n}$.

Now we want to develop an approximation for the sampling distribution of $\bar{Y}$ (regardless of the distribution of the population).

Consider random sample of size $n$ from exponential distribution with mean 10:

$$f(y) = \begin{cases} 
\frac{1}{10} e^{-y/10}, & y > 0 \\
0, & \text{elsewhere}
\end{cases}$$
\begin{align*}
\mathbb{E}(\overline{Y}) &= \mu = 10 \\
\text{Var}(\overline{Y}) &= \frac{\sigma^2}{n} = \frac{100}{n}
\end{align*}

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Average of 1000 sample means</th>
<th>( \mu_{\overline{Y}} = \mu )</th>
<th>Variance of 1000 sample means</th>
<th>( \sigma_{\overline{Y}}^2 = \frac{\sigma^2}{n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 5 )</td>
<td>9.86</td>
<td>10</td>
<td>19.63</td>
<td>20</td>
</tr>
<tr>
<td>( n = 25 )</td>
<td>9.95</td>
<td>10</td>
<td>3.93</td>
<td>4</td>
</tr>
</tbody>
</table>

**Theorem (CLT):** Let \( Y_1, \ldots, Y_n \) be i.i.d. with \( E(Y_i) = \mu \) and \( \text{Var}(Y_i) = \sigma^2 < \infty \).

Define \( U_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma/\sqrt{n}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \) \( \xrightarrow{d} \) \( \mathcal{N}(0,1) \)

Then \( \lim_{n \to \infty} P(U_n \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \) for all \( u \).

In other words, \( \overline{Y} \) is asymptotically normally distributed with mean \( \mu \) and variance \( \sigma^2/n \).

CLT can be applied to \( Y_1, \ldots, Y_n \) from any distribution provided that \( E(Y_i) = \mu \) and \( \text{Var}(Y_i) = \sigma^2 \) are both finite and \( n \) is large.
Example: (#7.42) The fracture strength of tempered glass averages 14 and has standard deviation 2.

(a) What is the probability that the average fracture strength of 100 randomly selected pieces of this glass exceeds 14.5?

(b) Find the interval that includes, with probability 0.95, the average fracture strength of 100 randomly selected pieces of this glass.

Solution:

(a) $\overline{Y} = \text{sample mean of 100 pieces}$

$\overline{Y} \sim \text{appr. } N(14, \frac{1}{100})$

$\frac{\overline{Y} - 14}{\frac{1}{\sqrt{100}}} \sim N(0,1)$

$p(\overline{Y} > 14.5) = p\left( z > \frac{14.5 - 14}{\frac{1}{\sqrt{100}}} \right)$

$= p(z > 2.5) = 0.0062$

(b) $p(-1.96 \leq z \leq 1.96) = 0.95$

$p(\overline{a} \leq \overline{Y} \leq 6) = 0.95$

$p\left( \frac{a - 14}{0.2} \leq z \leq \frac{6 - 14}{0.2} \right) = 0.95$

$\Rightarrow \frac{a - 14}{0.2} = -1.96 \Rightarrow a = 13.6$; $\frac{6 - 14}{0.2} = 1.96 \Rightarrow b = 14.4$
Normal Approximation to Binomial Distribution

\[ Y \sim \text{Bin}(n, p) \]
\[ Y = \sum X_i \]
\[ \frac{Y}{n} = \frac{1}{n} \sum X_i = \bar{X} \]
\[ \frac{Y}{n} \sim N \left( p, \frac{p(1-p)}{n} \right) \]

\[ X_i = \begin{cases} 1, & \text{if } \text{ith trial is success} \\ 0, & \text{otherwise} \end{cases} \]
\[ = \begin{cases} 1 \text{ with prob } p \\ 0 \text{ with prob } 1-p \end{cases} \]
\[ X_i \sim \text{Bernoulli}(p) \]
\[ E(X_i) = 1 \cdot p + 0 \cdot (1-p) = p \]
\[ E(X_i^2) = 1^2 \cdot p + 0^2 (1-p) = p \]
\[ \text{Var}(X_i) = p - p^2 = p(1-p) \]

\[ Y \overset{\text{approx}}{\sim} N \left( np, np(1-p) \right) \]
\[ Y = n \bar{X} \Rightarrow B(Y) = nE(\bar{X}) = np \]
\[ \text{Var}(Y) = n^2 \text{Var}(\bar{X}) = n^2 \frac{p(1-p)}{n} \]
\[ = np(1-p) \]
Example: (#7.80) The median age of residents of the US is 31 years. If a survey of 100 randomly selected US residents is to be taken, what is the appr. probability that at least 60 will be under 31 years of age?

Solution:

\[ Y = \# \text{ of residents younger than 31} \]

\[ Y \sim \text{Bin}(100, \frac{1}{2}) \approx \text{appr } N(50, 25) \]

\[ P(Y \geq 60) = P\left( \frac{Y - 50}{\sqrt{25}} \geq \frac{60 - 50}{\sqrt{25}} \right) \]

\[ = P(Y \geq 59.5) = P(z \geq \frac{59.5 - 50}{\sqrt{5}}) \]

(continuity correction)

\[ = P(z \geq 1.9) \]

\[ = 0.0287 \]