Section 4.12: Sampling Distributions

High-dose short-term folate administration modifies ambulatory blood pressure in postmenopausal women. A placebo-controlled study

http://www.nature.com/ejcn/journal/vaop/ncurrent/fig_tab/ejcn200958t1.html

Table 1. Mean (±s.e.) baseline and post-treatment parameters of postmenopausal women randomized to receive for 3 weeks a pill containing 15 mg of 5-MTHF (n=15) or placebo (n=15)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Baseline</th>
<th>5-MTHF</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI (kg/m²)</td>
<td>26.3 ± 1.3</td>
<td>26.3 ± 1.4</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>65.9 ± 4.7</td>
<td>65.9 ± 4.8</td>
</tr>
<tr>
<td>Cholesterol (mmol/l)</td>
<td>5.90 ± 0.3</td>
<td>5.92 ± 0.29</td>
</tr>
<tr>
<td>HDL (mmol/l)</td>
<td>1.61 ± 0.1</td>
<td>1.61 ± 0.1</td>
</tr>
<tr>
<td>Triglycerides (mmol/l)</td>
<td>1.14 ± 0.12</td>
<td>1.29 ± 0.13</td>
</tr>
<tr>
<td>Glucose (mmol/l)</td>
<td>5.04 ± 0.21</td>
<td>4.95 ± 0.28</td>
</tr>
<tr>
<td>Insulin (pmol/l)</td>
<td>79.0 ± 11.1</td>
<td>63.1 ± 0.71²</td>
</tr>
<tr>
<td>HOMA-IR</td>
<td>2.58 ± 0.04</td>
<td>2.03 ± 0.04b</td>
</tr>
<tr>
<td>Homocysteine (µmol/l)</td>
<td>11.77 ± 1.15</td>
<td>8.71 ± 0.50²</td>
</tr>
</tbody>
</table>

Abbreviations: BMI, body mass index; HDL, high-density lipoprotein; 5-MTHF, 5-methyltetrahydrofolate.

No difference was found in baseline parameters between the two groups.

² P<0.03; b P<0.01 vs baseline, by paired t-test.

A parameter is an attribute of an entire population.

- µₐ = population mean baseline triglycerides
- µₘ = population mean triglyceride level with drug M
- µ₅ = population mean difference in triglycerides after taking drug 5-MTHF
- σ₅ = population standard deviation of differences in triglycerides after taking drug 5-MTHF

Suppose Y = blood pressure in mm.

- µ = E(Y) = population mean (mm)
  - For a population of size N
    - \[ \mu = \frac{\sum_{i=1}^{N} y_i}{N} \]
    - E(Y) = Expected value of Y.
      - On average y is E(Y)
\( \sigma^2 = \text{population variance (mm}^2) \)

\[
\sum_{i=1}^{N} (y_i - \mu)^2
\]

\[
E(Y-\mu)^2 = \frac{1}{N}
\]

Average squared deviation from mean

\( \sigma = \sqrt{\sigma^2} = \text{population standard deviation (mm)} \)

- Back in original scale of mm

- For examples in the class the populations will be considered very large, essentially infinite.

An estimator is a value computed from data as an estimate of (guess at) the population parameter.

- \( \bar{Y}_b = 1.14 = \text{sample mean, average, triglycerides for baseline of n=15 subjects} \)
  - Estimator of \( \mu_b \)

- \( \bar{Y}_m = 1.29 = \text{sample mean, average, triglycerides after drug M of n=15 subjects} \)
  - Estimator of \( \mu_m \)

- \( \bar{d} = 0.15 = \text{sample mean, average, triglycerides difference from n=15 differences} \)
  - Estimator of \( \mu_d \)

- \( s_d = \text{sample standard deviation for triglycerides from n=15 differences} \)
  - Estimator of \( \sigma_d \)

- The sample mean is just the average of the n sampled, data values

\[
\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}
\]

- The sample variance is computed as

\[
S^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}
\]

- Why divide the sample standard variance, \( s^2 \), by \( n-1 \)?

- If we knew \( \mu \), we estimate \( \sigma^2 \) with \( \frac{\Sigma(y_i-\mu)^2}{n} \)

- But if we use \( \frac{\Sigma_{i=1}^{n} (y_i-\bar{y})^2}{n} \), this will be too small on average.

  - This is because \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \) is smaller than \( \sum_{i=1}^{n} (y_i - c)^2 \) for any other value c.

  - \( \bar{y} \) is the “least squares” estimator of \( \mu \)
    - \( \bar{y} \) minimizes the sum of squared deviations between \( y_i \) and our estimate.

  - So \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \) will be smaller than \( \sum (y_i - \mu)^2 \) (unless \( \bar{y} \) just happens to be exactly \( \mu \), which is highly unlikely)
    - For example with n=1 \( \frac{\Sigma_{i=1}^{n} (y_i-\bar{y})^2}{n} = 0 \) which is obviously too small for estimating \( \sigma^2 \)
• So we need to divide \( \sum_{i=1}^{n} (y_i - \bar{y})^2 \) by something less than \( n \) to make the resulting estimate of \( \sigma^2 \) not come out too small.
  o Dividing by \( n-1 \) makes this come out exactly right on average.
  o \( S^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \)
  o Notice that \( s^2 \) is undefined if \( n=1 \); we can't divide by zero.
    ▪ This makes sense. If we have only \( n=1 \) value, this tells us nothing about variability in the population.

  o The variance of an unbiased estimator is its expected squared deviation from the true population parameter.
    o \( \text{Var}(\bar{Y}) = E((\bar{Y} - \mu)^2) \)
    o Average squared error from using \( \bar{Y} \) as a guess at, estimator of, \( \mu \)
  o The standard error, \( SE \), of an estimator is the square root of its variance.
    o \( SE_{\bar{Y}} = \sqrt{\text{Var}(\bar{Y})} = \sqrt{E((\bar{Y} - \mu)^2)} \)
    o \( \sqrt{\text{Average squared error}} \)
    o This is the "standard" size of an error in using the estimator as a guess at the corresponding population parameter.

  o Important fact
    o For a mean of \( n \) independent values
      ▪ \( \text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \)
      ▪ \( SE_{\bar{Y}} = \sqrt{\text{Var}(\bar{Y})} = \frac{\sigma}{\sqrt{n}} \)
      ▪ Means from larger samples, bigger \( n \), have smaller potential error, \( SE \).
    o When sampling without replacement, the chosen values are not independent.
      ▪ If choosing \( n \) of \( N \) values without replacement
        \[ \text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \]
        This follows from the variance of a hypergeometric distribution.
    • Example 4.23 sampling \( n=5 \) of \( N=500 \) pennies should use this formula.
Suppose blood pressures in a population are normal with
Population mean $\mu = 120$
Population standard deviation $\sigma = 15$

If we take a random sample of $n = 4$ BP’s, how far might the sample mean, $\bar{y}$, be from $\mu = 120$?

Imagine over and over again, finding the mean of $n = 4$ BP’s.
Each student selects $n=4$ random subjects and records their blood pressures.

<table>
<thead>
<tr>
<th>Student</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
<th>$\bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>122</td>
<td>117</td>
<td>129</td>
<td>95</td>
<td>115.75</td>
</tr>
<tr>
<td>2</td>
<td>112</td>
<td>93</td>
<td>100</td>
<td>130</td>
<td>108.75</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>125</td>
<td>118</td>
<td>105</td>
<td>123.25</td>
</tr>
<tr>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
<tr>
<td>$\infty$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>

Doing this infinitely long, we would get the following means and variances

<table>
<thead>
<tr>
<th></th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
<th>$\bar{y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
<td>120</td>
</tr>
<tr>
<td>Variance</td>
<td>$15^2$</td>
<td>$15^2$</td>
<td>$15^2$</td>
<td>$15^2$</td>
<td>$\frac{15^2}{4}$</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td>$\sigma^2$</td>
<td>$n$</td>
<td></td>
</tr>
<tr>
<td>$\sqrt{\text{Variance}}$</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>$\frac{\sqrt{15}}{4}$ = 7.5</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\sigma$</td>
<td>$\text{SE}<em>{\bar{y}} = \sigma</em>{\bar{y}}$</td>
</tr>
</tbody>
</table>
If we drew a histogram of all \( \bar{y} \)'s

- \( \bar{y} - \mu = \bar{y} - 120 \) is our “error” when using \( \bar{y} \) as a guess (estimate) for \( \mu \).
- A usual size of error is:
  \[
  \text{Standard Error} = SE_{\bar{y}} = \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = 7.5
  \]
- It would not be at all unusual for \( \bar{y} \) to deviate from \( \mu \) by 7.5.
- For a normal population, 95% of the time, \( \bar{y} \) differs from \( \mu \) by no more than \( 1.96(SE_{\bar{y}}) \) or about 2 SE’s.

\[
P(\mu - 1.96\sigma_{\bar{y}} < \bar{y} < \mu + 1.96\sigma_{\bar{y}}) = 0.95
\]

\[
P(\bar{y} - 1.96\sigma_{\bar{y}} < \mu < \bar{y} + 1.96\sigma_{\bar{y}}) = 0.95
\]

This tells us how to find confidence intervals for \( \mu \).

A 95% confidence interval for \( \mu \) would be \( \bar{y} \pm 1.96\sigma_{\bar{y}} = \bar{y} \pm 1.96 \times SE_{\bar{y}} \)
Section 5.2: Estimation of $\mu$

- 95% of the time, $\mu$ is within
  \[ \bar{y} \pm 1.96\sigma_y = \bar{y} \pm 1.96 \times SE_y \]
- This is a 95% confidence interval for $\mu$.
- Usual notation with "$\alpha$" Confidence $= 100(1 - \alpha)\%$
- From the table of normal Z probabilities or the bottom of the t table

<table>
<thead>
<tr>
<th>Confidence</th>
<th>$\alpha$</th>
<th>$\alpha/2$</th>
<th>$z_{\alpha/2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.10</td>
<td>0.05</td>
<td>1.645</td>
</tr>
<tr>
<td>95%</td>
<td>0.05</td>
<td>0.025</td>
<td>1.96</td>
</tr>
<tr>
<td>99%</td>
<td>0.01</td>
<td>0.005</td>
<td>2.58</td>
</tr>
</tbody>
</table>

1st student before

\[ \bar{y} = 115.75 \quad \sigma_y = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{4}} = 7.5 \]

\[ 115.75 \pm 1.96(7.5) \]

\[ 115.75 \pm 14.7 \]

\[ 101.05 \pm 130.45 \quad \text{95% confidence interval for } \mu \]

- We are “95% confident” that $\mu$ is between 101.05 and 130.45.
- This does not say that 95% of the individual BP’s are between 101.45 and 130.45.
- 5% of such confidence intervals fail to include the true mean $\mu$. 
Section 5.3: Choosing n

- To be $100(1 - \alpha)\%$ sure that $\bar{y}$ deviates from $\mu$ by no more than error $E$
- We know that with a $100(1 - \alpha)\%$ chance,
  - the error in $\bar{y}$ is no more than $z_{\alpha/2}SE_{\bar{y}} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
  - We need $n$ large enough so that
    - $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = E$
    - $n = \frac{(z_{\alpha/2})^2 \sigma^2}{E^2}$
- In planning the experiment, we make our best guess at $\sigma^2$.
- The required sample size $n$ increases as $\sigma$ goes up $E$ goes down $\alpha$ goes down

**Example**: Patients with a new blood pressure medicine have population mean $\mu$

We want to be 95% certain that $\bar{y}$, our estimate of $\mu$, is within $\pm 2$mm of $\mu$ if $\sigma = 15$

$$1.96 \times SE_{\bar{y}} = 1.96 \times \frac{15}{\sqrt{n}} = 2$$

$$n = \frac{1.96^2(15^2)}{2^2} = 216.06$$

Strictly speaking, $n \geq 217$ to achieve no more than the required error in our estimate of the population mean, $\mu$. Round up