1 the same variance

Assume that $X_1, ..., X_m$ is a sample drawn from $N(\mu_X, \sigma^2)$, and $Y_1, ..., Y_n$ is a sample drawn from $N(\mu_Y, \sigma^2)$. Also assume that the two samples are independent. We can summarize these assumptions using the following statistical model:

$$
X_i = \mu_X + \epsilon_{X_i} \\
Y_j = \mu_Y + \epsilon_{Y_j}
$$

where $\{\epsilon_{X1}, \ldots, \epsilon_{Xm}, \epsilon_{Y1}, \ldots, \epsilon_{Yn}\} \overset{iid}{\sim} N(0, \sigma^2)$.

In many cases, we are interested in whether their mean values are the same. For example, there is a treatment group and a control group and we want to know the effect of the treatment group. Consider the null hypothesis $H_0 : \mu_X = \mu_Y$ and three alternatives:

- $H_1 : \mu_X \neq \mu_Y$
- $H_2 : \mu_X > \mu_Y$
- $H_3 : \mu_X < \mu_Y$.

The difference between these two groups can be characterized by the difference in their mean values, i.e., $\mu_X - \mu_Y$. An estimate of it is $\bar{X} - \bar{Y}$.

Recall that

$$
\bar{X} \sim N(\mu_X, \frac{1}{m}\sigma^2)
$$

and

$$
\bar{Y} \sim N(\mu_Y, \frac{1}{n}\sigma^2)
$$

we have

$$
\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right))
$$

If $\sigma^2$ is known, we can make inference (e.g., calculating C.I. of $\mu_X - \mu_Y$ and doing hypothesis test) based on

$$
Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}},
$$

which follows the standard normal distribution.
In practice, $\sigma^2$ is unknown and has to be estimated from observed data by calculating the pooled sample variance

$$s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}$$

where

$$s^2_X = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2$$

and

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

It is easy to see that the pooled sample variance is a weighted sum of the two sample variances, with the weights proportional to the degrees of freedom. Does this make sense? Yes. Consider $n=5$, and $m=50$. Of course the sample variance from the larger sample is more reliable than that from the smaller sample; by giving a larger weight to the former, we believe it is more reliable than the latter.

1.1 Theorem

Under the statistical model (assumptions) for two-sample t-test,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Proof: Let

$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

Since $U$ is a linear function of independent normal random variables, it follows a normal distribution. It is not difficult to verify that $E[U] = 0$ and $\text{Var}[U] = 1$. Therefore, $U \sim N(0,1)$.

From STAT120B, $(m-1)s_X^2/\sigma^2 \sim \chi^2_{m-1}$ and $(n-1)s_Y^2/\sigma^2 \sim \chi^2_{n-1}$. Together with the fact that the two samples are independent,

$$V = (m+n-2) \frac{s_p^2}{\sigma^2} \sim \chi^2_{m+n-2}$$

Also, from STAT120B we know that $U$ and $V$ are independent. By the definition of t-distribution,

$$\frac{U}{\sqrt{V/(m+n-2)}} \sim t_{m+n-2}$$
Consider the following test statistic

\[ T = \frac{(\bar{X} - \bar{Y})}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2} \]

Under the null hypothesis of equal mean, the test statistic follows \( t_{m+n-2} \).

Let \( s_{\bar{X}-\bar{Y}} = s_p \sqrt{\frac{1}{m} + \frac{1}{n}} \), then a 100(1 - \( \alpha \))% confidence interval for \( \mu_X - \mu_Y \) is

\[ (\bar{X} - \bar{Y}) \pm t_{m+n-2,1-\alpha/2} s_{\bar{X}-\bar{Y}} \]

The rejection regions:

- For \( H_1 \), \[ \frac{X - Y}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} > t_{m+n-2,1-\alpha/2} \]
- For \( H_2 \), \[ \frac{X - Y}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} > t_{m+n-2,1-\alpha} \]
- For \( H_3 \), \[ \frac{X - Y}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} < -t_{m+n-2,1-\alpha} \]

### 1.2 Equivalence of the two-sample t test and the likelihood ratio test (LRT)

We will demonstrate that the t test for \( H_0 : \mu_X = \mu_Y \) v.s. \( H_1 : \mu_X \neq \mu_Y \) is equivalent to the LRT. Under the statistical model (normality, independence, and equal variance), the likelihood for \( \mu_X, \mu_Y, \sigma^2 \) is

\[
L(\mu_X, \mu_Y, \sigma^2) = f(X_1, \ldots, X_m, Y_1, \ldots, Y_n | \mu_X, \mu_Y, \sigma^2) = \prod_{i=1}^{m} f(X_i | \mu_X, \sigma^2) \prod_{j=1}^{n} f(Y_j | \mu_Y, \sigma^2) = \left( \frac{1}{2\pi} \right)^{m+n} \left( \frac{1}{\sigma^2} \right)^{m+n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{m}(X_i - \mu_X)^2 + \sum_{j=1}^{n}(Y_j - \mu_Y)^2 \right\}
\]

The log-likelihood is (in many situations it is easier to work on log-likelihood than on likelihood)

\[
l(\mu_X, \mu_Y, \sigma^2) = -\frac{m+n}{2} \log(2\pi) - \frac{m+n}{2} \log(\sigma^2) - \frac{\sum_{i=1}^{m}(X_i - \mu_X)^2 + \sum_{j=1}^{n}(Y_j - \mu_Y)^2}{2\sigma^2}
\]

Under the full parameter space, i.e.,

\[ \Omega = \{ -\infty < \mu_X < \infty, -\infty < \mu_Y < \infty, 0 < \sigma < \infty \} \]
The mle

\[ \hat{\mu}_X = \bar{X} \]
\[ \hat{\mu}_Y = \bar{Y} \]
\[ \hat{\sigma}^2 = \frac{1}{m+n} \left( \sum_{i=1}^{m} (X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{n} (Y_j - \hat{\mu}_Y)^2 \right) \]

Substituting them back to the log-likelihood leads to

\[ l(\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}^2) = -\frac{m+n}{2} \log 2\pi - \frac{m+n}{2} \log \hat{\sigma}^2 - \frac{m+n}{2} \]

Under the reduced parameter space, i.e.,

\[ \Omega_0 = \{ -\infty < \mu_X = \mu_Y < \infty, 0 < \sigma < \infty \} \]

The mle

\[ \hat{\mu}_0 = \frac{1}{m+n} \left( \sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j \right) = \frac{m}{m+n} \bar{X} + \frac{n}{m+n} \bar{Y} \]

(The mle for \( \mu_0 \) is a weighted sum of the two sample means)

\[ \hat{\sigma}_0^2 = \frac{1}{m+n} \left( \sum_{i=1}^{m} (X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{n} (Y_j - \hat{\mu}_0)^2 \right) \]

Substituting them back to the log-likelihood leads to

\[ l(\hat{\mu}_0, \hat{\sigma}_0^2) = -\frac{m+n}{2} \log 2\pi - \frac{m+n}{2} \log \hat{\sigma}_0^2 - \frac{m+n}{2} \]

The log of the likelihood ratio is

\[ \log(\Lambda) = l(\hat{\mu}_0, \hat{\sigma}_0^2) - l(\hat{\mu}_X, \hat{\mu}_Y, \hat{\sigma}^2) = \frac{m+n}{2} \log(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}) \]

The LRT rejects \( H_0 \) for small values of \( \Lambda \), i.e., it rejects \( H_0 \) for large values of

\[ \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{m} (X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{n} (Y_j - \hat{\mu}_0)^2}{\sum_{i=1}^{m} (X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{n} (Y_j - \hat{\mu}_Y)^2} \]
It is easy to verify that
\[
\sum_{i=1}^{m}(X_i - \mu_0)^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2 + m(\bar{X} - \mu_0)^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2 + \frac{mn}{(m+n)^2}(\bar{X} - \bar{Y})^2
\]
and
\[
\sum_{j=1}^{n}(Y_j - \mu_0)^2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2 + \frac{mn}{(m+n)^2}(\bar{X} - \bar{Y})^2
\]
Using the above identities, the numerator of the ratio is
\[
\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2 + \frac{mn}{m+n}(\bar{X} - \bar{Y})^2 = (m+n-2)s_p^2 + \frac{mn}{m+n}(\bar{X} - \bar{Y})^2
\]
So we reject the \(H_0\) for large values of
\[
1 + \frac{mn}{m+n}\frac{(\bar{X} - \bar{Y})^2}{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2} = 1 + \frac{mn}{(m+n)(m+n-2)}\frac{(\bar{X} - \bar{Y})^2}{S_p^2}
\]
or, equivalently, we reject \(H_0\) for large values of
\[
\frac{|\bar{X} - \bar{Y}|}{S_p}
\]
which is the same as the t statistic for \(H_0\), except for constants that do not depend on data.

2 Different variances

So far, we have assumed that the two samples have the same variance. What if this assumption is violated? Of course the test we deceived is not appropriate. Since the two samples are independent,
\[
v\text{ar}(\bar{X} - \bar{Y}) = v\text{ar}(\bar{X}) + v\text{ar}(\bar{Y}) = \frac{s_X^2}{m} + \frac{s_Y^2}{n}
\]
Under the null hypothesis,
\[
\frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}} \sim t_{df}
\]
where
\[
df = \frac{(s_X^2/m + s_Y^2/n)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}}
\]
3  An example

We want to compare the effects of two drugs for hypertension. In a clinical trial with 25 patients, 10 randomly selected patients took drug 1 and the remaining 15 subjects took drug 2. The table below shows measures of systolic blood pressure after each treatment:

<table>
<thead>
<tr>
<th>Drug 1</th>
<th>Drug 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>142</td>
<td>138</td>
</tr>
<tr>
<td>120</td>
<td>133</td>
</tr>
<tr>
<td>135</td>
<td>124</td>
</tr>
<tr>
<td>126</td>
<td>139</td>
</tr>
<tr>
<td>141</td>
<td>122</td>
</tr>
<tr>
<td>130</td>
<td>118</td>
</tr>
<tr>
<td>137</td>
<td>126</td>
</tr>
<tr>
<td>130</td>
<td>140</td>
</tr>
<tr>
<td>117</td>
<td>132</td>
</tr>
<tr>
<td>137</td>
<td>121</td>
</tr>
<tr>
<td>124</td>
<td></td>
</tr>
<tr>
<td>131</td>
<td></td>
</tr>
<tr>
<td>129</td>
<td></td>
</tr>
<tr>
<td>118</td>
<td></td>
</tr>
<tr>
<td>123</td>
<td></td>
</tr>
</tbody>
</table>

We can use two-sample t-test to test whether the two drugs have different effects, i.e., $H_0$: means are equal v.s. $H_1$: means are not equal. To use R to perform the two-sample t test, we first read the data:

```r
> drug1 = c(142, 120, 135, 126, 141, 130, 137, 130, 117, 137)
> drug2 = c(138, 133, 124, 139, 122, 118, 126, 140, 132, 121, 124, 131, 129, 118, 123)
> boxplot(drug1,drug2,names=c("drug1","drug2"))
> t.test(drug1, drug2, var.equal=T)
```

The R command to perform the test,

```r
> t.test(drug1, drug2, var.equal=T)
```

And the output is

```
Two Sample t-test
data:  drug1 and drug2
t = 1.1358, df = 23, p-value = 0.2677
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -2.984294  10.250961
sample estimates:
  mean of x  mean of y
  131.5  127.8667
```
If you only know the basic commands (e.g., mean and variance), you can still calculate the test statistic:

```r
> mean1 = mean(drug1) # the sample mean for drug 1
> mean2 = mean(drug2) # the sample mean for drug 2
> m = length(drug1)
> n = length(drug2)
> s1 = var(drug1) # the sample variance for drug 1
> s2 = var(drug2) # the sample variance for drug 2
> sp = ((m - 1) * s1 + (n - 1) * s2) / (m + n - 2) # pooled sample variance
> t = (mean1 - mean2) / (sqrt(sp) * sqrt(1/m + 1/n)) # the test statistic
> t
> qt(0.975, 23) # the 2.5th upper point of t_{18}, or the 97.5th percentile
```

The value of t test statistic is 1.14, which is less than the 97.5th percentile of $t_{23}$ (2.068658). We therefore fail to reject the $H_0$ at level 0.05. So there is not enough evidence of different effects of the two drugs.

The above analysis assumes equal variance. If we drop this assumption, we will use the two-sample t-test with unequal variance:

```r
> t.test(drug1, drug2) # the default of t.test is unequal variance.
```

To learn more about this function, use the help command:

```r
> help(t.test)
```

**question**, what if there are more than two drugs/treatments?