Concentration of benzene in a cigar

Suppose that we want to determine whether the concentration of benzene in a brand of cigars is the same as that of cigarettes. Suppose further that we know that the mean concentration of benzene in cigarettes is $\mu = 81 \mu g/g$ of tobacco, but are unsure of the variability of that measurement in cigars.

Had we known $\sigma$, the test would be based on the statistic $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$. Since we do not, we must estimate it, using the sample standard deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$
We can then “plug in” $s$ into the previous test statistic and use $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ in our testing.

Note however, that we are given less information than we were, when the population standard deviation was known. Thus, $t$ is not distributed according to a standard normal distribution. In fact we should expect $t$ to be more variable than $z$, and its distribution should reflect this.

The distribution of $t$ is called the *Student’s t* distribution (or *t* distribution).
The $t$ distribution

- The $t$ distribution is symmetric, and centered around zero.
- It has “fatter” tails compared to the standard normal distribution.
- The $t$ distribution is defined by $n-1$ degrees of freedom ($n$ is the sample size).
Degrees of freedom are essentially the number of independent pieces of information provided by the sample. Initially, every sample has $n$ independent pieces of information (as many as the number of observations). However, after we calculate the sample mean, there are only $n-1$ independent pieces. Recall that $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$. Thus, if we know the first $n-1$ observations, we can compute the $n^{th}$ one (that would be $x_n = \bar{x} - \sum_{i=1}^{n-1} (x_i - \bar{x})$), and thus there are $n-1$ independent pieces of information.
Concentration of benzene in cigars (continued)

A random sample of 7 cigars had mean benzene concentration $\bar{x}_7 = 15 \mu g/g$ and std. deviation $s = 9 \mu g/g$. Is it possible that the benzene concentration is the same as that of the cigarettes? To answer this question, we proceed as follows:

1. $H_0: \mu = \mu_o$
2. $H_a: \mu \neq \mu_o$
3. The alpha level of the test is 5%
What is the probability that the cigar population mean benzene concentration is 81 µg/g? Since \( t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{151 - 81}{9/\sqrt{7}} = 20.6 \), the probability that a sample mean of 151 or higher is less than 0.0001.

4. Since this is less than the alpha level of the test we reject the null hypothesis. Cigars have higher concentration of benzene than cigarettes.
STATA Output: benzene concentration example

```
.ttesti 7 151 9 81, level(95)
```

```
Number of obs = 7

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
x |       151     3.40168   44.3898   0.0000       142.6764    159.3236
------------------------------------------------------------------------------

Degrees of freedom: 6

Ho: mean(x) = 81

Ha: mean < 81       Ha: mean ~= 81       Ha: mean > 81

  t = 20.5781       t = 20.5781       t = 20.5781
  P < t = 1.0000       P > |t| = 0.0000       P > t = 0.0000
```

Since we are performing a two-sided test, we concentrate on the middle test. Since \( P > |t| = 0.0000 \), which is much smaller than 0.05, we *reject* the null hypothesis.
# Independent Samples

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$\mu_1$</td>
<td>$\mu_2$</td>
</tr>
<tr>
<td>Std. deviation</td>
<td>$\sigma_1$</td>
<td>$\sigma_2$</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>$\bar{x}_1$</td>
<td>$\bar{x}_2$</td>
</tr>
<tr>
<td>Std. deviation</td>
<td>$s_1$</td>
<td>$s_2$</td>
</tr>
<tr>
<td>Sample size</td>
<td>$n_1$</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>
Comparing two independent Samples

Example. Consider the comparison of the serum iron levels of healthy children versus children suffering from cystic fibrosis. The (population) mean of the serum iron level among healthy children is $\mu_1$ while the mean serum iron level among children suffering from cystic fibrosis is $\mu_2$. Comparison of these unknown means is performed by taking two samples of size $n_1 = 9$ and $n_2 = 13$ children from the two populations.
Comparing two independent samples (continued)

In the case of two independent samples consider the following issues:

1. The two sets of measurements are independent (because each comes from a different group (e.g., healthy children, children suffering from cystic fibrosis).

2. In contrast to the one-sample case, we are simultaneously estimating two population means instead of one. Thus, there are now two sources of variability instead of one (one from each sample) instead of just one as was the case in the one-sample tests. As a result, the standard deviation is going to be roughly double (!) compared to the one-sample case.
Comparing two independent samples (continued)

3. The following assumptions must hold:

   a. The two samples must be independent from each other
   
   b. The individual measurements must be roughly normally distributed
   
   c. The variances in the two populations must be roughly equal
   
   d. If a-c are satisfied, inference will be based on the statistic

\[
T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},
\]

is distributed according to a *t* distribution with \( n_1 + n_2 - 2 \) degrees of freedom.
Testing of two independent samples (assuming equal variances)

STEP 1. Based on two random samples of size $n_1$ and $n_2$ observations compute the sample means $\bar{x}_1$ and $\bar{x}_2$, and the std. deviations

$$s_1^2 = \left(\frac{1}{n_1-1}\right) \sum_{j=1}^{n_1} (x_{j1} - \bar{x}_1)^2 \quad \text{and} \quad s_2^2 = \left(\frac{1}{n_2-1}\right) \sum_{j=1}^{n_2} (x_{j2} - \bar{x}_2)^2.$$

STEP 2. Compute the pooled estimate of the population variance

$$s_p^2 = \frac{(n_1-1)}{(n_1+n_2-2)} s_1^2 + \frac{(n_2-1)}{(n_1+n_2-2)} s_2^2$$

$$= \frac{\sum_{j=1}^{n_1} (x_{j1} - \bar{x}_1)^2 + \sum_{j=1}^{n_2} (x_{j2} - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

STEP 3. The estimate of the standard deviation is $s_p = \sqrt{s_p^2}$.
Serum iron levels and cystic fibrosis (continued)

\[ n_1 = 9 \quad n_2 = 13 \]

\[ \bar{x}_1 = 18.9 \text{ µmol/l} \quad s_1 = 5.9 \text{ µmol/l} \quad \bar{x}_2 = 11.9 \text{ µmol/l} \quad s_2 = 6.3 \text{ µmol/l} \]

Based on two samples the *pooled* estimate of the population variance

\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}
\]

\[
s_p^2 = \frac{(8)(5.9)^2 + (12)(6.3)^2}{(9+13-2)}
\]

\[
s_p^2 = 37.74
\]

The estimate of the standard deviation is \( s_p = \sqrt{37.74} = 6.14 \).
Serum iron levels and cystic fibrosis (continued)

1. State the null hypothesis $H_0: \mu_1 - \mu_2 = 0$

2. Set up the alternative hypothesis: Two-sided alternative $H_a: \mu_1 \neq \mu_2$

3. The $\alpha$ level is 5% (the significance level of the test is 95%)

4. Rejection rule: Reject the $H_0$, if
   $$\frac{(18.9 - 11.9)}{\sqrt{\frac{1}{9} + \frac{1}{13}}} = 2.63 \geq t_{20;0.025} = 2.086$$

That is, we are sure at the 95% level that children suffering from cystic fibrosis have significantly different levels of iron in their serum compared to healthy children.

It appears that these children have an iron deficiency.
STATA output: Blood Iron Levels

```
. ttesti 9 18.9 5.9 13 11.9 6.3
          x: Number of obs =         9
          y: Number of obs =        13

------------------------------------------------------------------------------
  Variable |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------
          x |      18.9    1.966667   9.61017   0.0000       14.36486    23.43514
          y |      11.9    1.747306   6.81049   0.0000       8.092948    15.70705
          diff |         7    2.663838   2.62779   0.0161       1.443331    12.55667
------------------------------------------------------------------------------

Degrees of freedom: 20
Ho: mean(x) - mean(y) = diff = 0
Ha: diff < 0          Ha: diff ~= 0          Ha: diff > 0
    t =   2.6278          t =   2.6278          t =   2.6278
    P < t =  0.9919       P > |t| =  0.0161       P > t =  0.0081
```
The two-sided test corresponds to the middle alternative (Ha:diff~0). The p-value (P>|t|=0.0161) is less than the α level, so we reject H₀. Children with cystic fibrosis (group y) have different levels of iron in their blood from healthy children. The sample mean is less than that of the healthy children meaning that children with cystic fibrosis have lower blood iron levels.
Paired samples

Example. Sixty-three adult males suffering from coronary artery disease were tested. The test involved challenging the subjects’ cardiovascular system by riding a stationary bicycle until the onset of angina (chest pain). After rest, the subjects rode again until the onset of angina and the percent of time of earlier onset of pain was recorded. On the first visit subjects were breathing clean air, while on a subsequent visit, the subject repeated the same series of tests, but CO was mixed in the air.

The percent difference in the time to the onset of angina on the first series of tests (when breathing regular air) and the percent difference of time to onset of angina during the second series of tests (when breathing air mixed with CO) were compared.
Each patient accounts for a *pair* of observations, so there are two issues to consider:

1. The two sets of measurements are not independent (because each pair is measured on the same patient) and each patient serves as his own “control”. The advantage of this design is that we are able to account for individual (biological) patient variability. Someone that tends to experience angina faster on clean air will more likely experience angina faster when the air is mixed with CO. Similarly someone that experienced angina later when breathing clean air, will likely experience symptoms later when breathing CO as well.

2. It is not appropriate to think that we have $2n$ distinct (independent) data points (or units of information) available to us, since each data point on the same subject provides a great deal of information on the subsequent data points collected on the same subject.
Hypothesis testing of paired samples

In a random sample of size \( n \) paired observations, compute the sample mean \( \bar{d}_n \) of the differences between the pairs of observations \( d_i = x_{ci} - x_{ri}, \ i=1,\ldots,\ n. \)

Carry out the test like a usual single sample \( t \) test based on these differences.

1. State the null hypothesis \( H_o: \delta(=\mu_C-\mu_T)=0 \)

2. Set up the alternative hypothesis
   a. One-sided tests: \( H_a: \delta \geq 0 \) (equivalent to \( \mu_C \geq \mu_T \))
      or \( H_a: \delta \leq 0 \) (equivalent to \( \mu_C \leq \mu_T \))
   b. Two-sided tests: \( H_a: \delta \neq 0 \)

3. Choose the \( \alpha \) level (the significance level of the test is \( 1-\alpha \)).
Hypothesis testing of paired samples (continued)

4. Rejection rule: Reject the null hypothesis,

   a. One-sided tests. If \( \frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1;\alpha} \) or if \( \frac{\bar{d}}{s_d/\sqrt{n}} < -t_{n-1;\alpha} \)

   b. Two-sided tests. If \( \frac{\bar{d}}{s_d/\sqrt{n}} > t_{n-1;\alpha/2} \) or \( \frac{\bar{d}}{s_d/\sqrt{n}} \leq -t_{n-1;\alpha/2} \) (or
CO study (continued)

The sample size is \( n = 63 \). The mean time to occurrence of angina was \( \bar{x}_c = 3.35\% \) faster for control subjects (breathing clean air on both occasions) and \( \bar{x}_r = 9.63\% \) faster for subjects breathing air mixed with CO during the second study visit. The difference between the two means is \( \bar{d} = -6.63\% \) with standard deviation \( s_d = 20.29\% \).
CO study (continued)

The hypothesis “is breathing CO associated with faster onset of angina” is tested as follows:

1. State the null hypothesis $H_0: \delta(=\mu_C-\mu_T)=0$

2. Set up the alternative hypothesis: $H_a: \delta \leq 0$ (equivalent to $\mu_C \leq \mu_T$ that is, when breathing air mixed with CO angina occurs faster)

3. The $\alpha$ level is 5%

4. Rejection rule: Reject $H_0$, since $\frac{\bar{d}}{s_d/\sqrt{n}} = -2.59 \leq -t_{62;0.05} = -1.673$

   (or equivalently since $P(T \leq -2.59) \approx 0.005 \leq 0.05$).

Subjects when breathing air with CO experience angina faster than when breathing air without CO.
Computer implementation

To carry out the above test of hypothesis by STATA we use the one-sample $t$-test command as before, noting that our data are now comprised by differences of the paired observations and the mean under the null hypothesis is zero.

The output is as follows:

```
. ttesti 63 -6.63 20.29 0

Number of obs =       63

------------------------------------------------------------------
                  |      Mean    Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x  |     -6.63      2.5563  -2.59359   0.0118        1.52003    11.73997
------------------------------------------------------------------

Degrees of freedom: 62
Ho: mean(x) = 0
Ha: mean < 0             Ha: mean ~= 0              Ha: mean > 0
 t =  -2.5936           t =  -2.5936              t =  -2.5936
 P < t =  0.0059   P > |t| =   0.0118         P > t =   0.9941
```
Since $P < t = 0.0059$ is less than 0.05, we reject the null hypothesis. Patients experience angina faster (by about 6.63%) when breathing air mixed with CO than when breathing clean air.