Math 425

Student t distribution

William Gosset 1908
worked for Guinness

had to publish under a pen name: Student

We have a population, we measure some variable X which has the normal (= Gaussian) distribution. All we know is the average (= expected value): \( \mu \).

Now we have a sample of \( n \) individuals.
We measure $X$ for them to get readings $X_1, \ldots, X_m$.

The null hypothesis: their result are the same as the population. How can we test this?

The **Student t-test** is: let

\[
\bar{X} = \frac{1}{m} \left( X_1 + \cdots + X_m \right).
\]

\[
T = \frac{\bar{X} - \mu}{\sqrt{\frac{\sum_{i=1}^{m} (X_i - \bar{X})^2}{m(m-1)}}} \leq \text{really should consider the absolute value}
\]

Compare it with the Student t-table.

Degrees of freedom $k = m-1$. 
Example: The mean score on a 115 exam is 70. A group of 5 students got scores 75, 72, 80, 69, 75.  
(1) Are they significantly different from the population? 

(2) Are they significantly better than the population? 

\[ X = \frac{1}{5} (75 + 72 + 80 + 69 + 75) \]

\[ T = \frac{74.2 - 70}{\sqrt{\frac{0.8^2 + 2.2^2 + 5.8^2 + 5.2^2 + 0.8^2}{20}}} \]

\[ = 2.298 \]
4 degrees of freedom

One-tail  0.05

Two-tail  0.05

4 degrees of freedom  2.132  2.776

not significantly different

(1) Two-tail : statistically insignificant

(2) One-tail : statistically significant

Example: suppose an average grade in a class is a B. Suppose then student got an A. Are they significantly better?
\[ \frac{?}{0} = ? \]

Why is this wrong? Answer: Grades only (this is clearly wrong).

Takes on discrete values (very few values - we cannot approximate by the normal distribution here).

Comment: With more degrees of freedom, the values in the table converge:

<table>
<thead>
<tr>
<th>One-tailed 0.05:</th>
<th>1</th>
<th>2</th>
<th>10</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>degrees of freedom</td>
<td>t</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.314</td>
<td>2.920</td>
<td>1.812</td>
<td>1.660</td>
<td>1.653</td>
</tr>
</tbody>
</table>
In the $t$-test, the degrees of freedom are determined by the size of the sample.

In the $X^2$ test, the degrees of freedom were determined by the number of outcomes.

The $t$-test is a part of small sample theory.

The $X^2$-test is an example of large sample theory.

Let us determine the probability distribution of the random variable $T$. 
\[ T = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{m}} \]

\[ \bar{X} - \mu_0 : \text{A sum of independent centered Gaussian} \]

\[ \text{let } \text{var}(X_i) = \sigma^2 \]

\[ \text{var}(X_1 + \cdots + X_m) = m\sigma^2 \]

\[ \text{var}(\bar{X}) = \text{var}(\bar{X} - \mu_0) = \frac{m\sigma^2}{m^2} = \frac{\sigma^2}{m} \]

A standard Gaussian variable:

\[ Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{m}} \]
Now the denominator:

\[ X_n - \bar{X} \] are also independent Gaussian variables

\[ \bar{Y}_i = \frac{X_i - \bar{X}}{\sigma} \] a standard Gaussian variable

The denominator is

\[ \frac{\sum_{i=1}^{n} Y_i^2 \cdot \sigma^2}{n(n-1)} = \frac{\sigma}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} Y_i^2 / (n-1)} \]

Summarize:

\[ T = \frac{Z}{\sqrt{\sum_{i=1}^{n} Y_i^2 / (n-1)}} \]
But $s = \sum_{i=1}^{n} \epsilon_i^2 \sim \chi^2$ with $n-1$ degrees of freedom.

Conclusion: \[ T = \frac{Z}{\sqrt{s/k}} \]

where $T, s$ are independent, $Z$ is standard Gaussian and $s$ is $\chi^2$ with $n-1$ degrees of freedom.

From now on, let $k = n-1 = \# \text{ degrees of freedom}$.

\[ T = \frac{Z}{\sqrt{s/k}} \]

Why do we divide the denominator by $\sqrt{k}$?
The reason is that there is a limit

$$\lim_{h \to 0} \frac{5}{h}$$

We will discuss the limit when we compute the t distribution.

HW: Look up the t table on the internet.

Suppose an average height of a person in a population is 5'8". Suppose a sample of 10 people on an island have heights:

5'1, 5'0, 5'5, 6'0, 5'9, 5'0, 5'1, 4'9, 5'5, 5'6.
Are they significantly shorter than the standard population?

Are they significantly different from the standard population? [Use 95% certainty]