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): μ.

Now we have a sample of n individuals.

We measure X for them to get readings $X_1, \dots, X_n$.
The null hypothesis: their results are the same as the population. How can we test this?

The Student t-test: let

$$\bar{X} = \frac{1}{n}(X_1 + \dots + X_n).$$

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n(n-1)}}}$$

← really should consider the absolute value

Compare it with the Student t table.
Degrees of freedom $k = n-1$.

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?

$$\bar{X} = \frac{1}{5}(75+72+80+69+75) = 74.2$$

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

$$\doteq 2.298$$

4 degrees of freedom

	One-tail	Two-tail
	0.05	0.05
4 degrees of freedom	2.132	2.776

① Two-tail – statistically insignificant
not significantly different

② One-tail : statistically significant
significantly better

Example: Suppose an average grade in a class is a B. Suppose three student got an A. Are they significantly better?

$$T = \frac{?}{0} = \infty \qquad ??$$

Why is this wrong? ← (this is clearly wrong).

Answer: Grades only take 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:

One-tail 0.05:

degrees of freedom	1	2	10	100	200
t	6.314	2.920	1.812	1.660	1.653

In the t test, the degrees of freedom are determined by the size of the sample

In the Pearson χ^2 test, the degrees of freedom were determined by the number of outcomes.

The t-test is a part of small sample theory

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

Let us determine the probability distribution of the random variable T:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\dfrac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n(n-1)}}}$$

$\bar{X} - \mu_0$: A sum of independent centered Gaussians

Let $\operatorname{var}(X_i) = \sigma^2$

$$\operatorname{var}(X_1 + \dots + X_n) = n\sigma^2$$

$$\operatorname{var}(\bar{X}) = \operatorname{var}(\bar{X} - \mu_0) =$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

A standard Gaussian variable:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Now the denominator:

$X_i - \bar{X}$ are also independent Gaussian variables

$Y_i = \dfrac{X_i - \bar{X}}{\sigma}$ a standard Gaussian variable

The denominator is

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

will cancel out

Summarize:

$$T = \frac{Z}{\sqrt{\left(\sum_{i=1}^{n} Y_i^2\right)/(n-1)}}$$

But $s = \sum_{i=1}^{n} Y_i^2$ is χ^2 with $n-1$ degrees of freedom.

<u>Conclusion</u>: $$T = \frac{Z}{\sqrt{\frac{s}{n-1}}}$$ where Z, s are independent, Z is standard Gaussian and s is χ^2 with $n-1$ degrees of freedom.

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

$$T = \frac{Z}{\sqrt{\frac{s}{k}}}.$$

Why do we divide the denominator by $\sqrt{k}$?

The reason is that there is a limit

$$\lim_{k \to \infty} \frac{s}{k}$$

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

HW 1 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.

(a) Are they significantly shorter than the standard population?

(b) Are they significantly different from the standard population?

Use 95% certainty