

MthStat 465, Spring 2005, Lecture Number 24
Student's t-distribution

Suppose we wanted to construct a test of hypotheses where the null hypothesis was simply that the population mean was 6.8., and the alternative hypothesis was that it was not. How could we go about constructing a rejection region?

A natural reaction would be to try to use the Central Limit Theorem as follows. Let R_1, R_2, \dots, R_N be independent random variables having the population distribution function. If the null hypothesis is true, all we know is the population mean. So we will reject the null hypothesis if

$$\left| \frac{R_1 + \dots + R_N}{N} - 6.8 \right| > C$$

where C is to be chosen so that the level (probability of Type I error) of the test is α . That is, given α , we need

$$\Pr_{H_0} \left(\left\{ \left| \frac{R_1 + \dots + R_N}{N} - 6.8 \right| > C \right\} \right) \geq \alpha$$

with equality for some distribution in the null hypothesis. Notice that

$$\begin{aligned} \left\{ \left| \frac{R_1 + \dots + R_N}{N} - 6.8 \right| > C \right\} &= \{|R_1 + \dots + R_N - 6.8N| > CN\} \\ &= \left\{ \left| \frac{R_1 + \dots + R_N - 6.8N}{\sigma\sqrt{N}} \right| > \frac{C}{\sigma\sqrt{N}} \right\} \end{aligned}$$

for any positive number σ . If the population mean were 6.8 (which it is if H_0 is true) and σ were the population standard deviation then we could say that

$$\Pr_{H_0} \left(\left\{ \left| \frac{R_1 + \dots + R_N}{N} - 6.8 \right| > C \right\} \right) \approx \int_{-C/\sqrt{N}}^{C/\sqrt{N}} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

and get an estimate of C . However, the null hypothesis makes no mention of the value of the population variance at all! There are many distributions in the null hypothesis, and, therefore, many population variances. A way out was described by William Gosset in the 1800's. His idea now seems quite obvious: replace σ^2 with $(s(R_1, \dots, R_N))^2$, the unbiased estimate of the population variance described in the last lecture. Gosset was able to calculate a formula for the density of

$$\frac{R_1 + \dots + R_N - N\mu}{\sqrt{N}s(R_1, \dots, R_N)}$$

under the assumption that the random variables R_k are normally distributed with expected value μ . This formula depends only on N , and not on μ . These densities are proportional to the expressions

$$\left(\frac{1}{1 + (u^2/N)} \right)^{(N+1)/2}$$

so they have a "bell shape" and converge to the the normal density as $N \rightarrow \infty$ since $(1 + (x/N))^N$ converges to e^x as $N \rightarrow \infty$. These densities are called *t-densities* and can be used in place of the standard normal density to design hypothesis tests where only the mean is described in the null hypothesis.

0.1. **A note on normal distributions.** A random variable N is said to have the standard normal distribution if the density function of its distribution function is $e^{-u^2/2}/\sqrt{2\pi}$. It is straightforward calculus to show that if N has the standard normal distribution then $E[N] = 0$ and $\text{Var}[N] = 1$. A random variable R is said to have a normal distribution if we can write $R = aN + b$ where $a > 0$ and N has the standard normal distribution. It is then easy to show from properties of expected value that $E[R] = b$ and $\text{Var}[R] = a$. As for the density of R :

$$\begin{aligned} \Pr(R \leq x) &= \Pr(aN + b \leq x) \\ &= \Pr(N \leq (x - b)/a) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-b)/a} e^{-u^2/2} du. \end{aligned}$$

Since we can recover the density function from the distribution function by differentiating we have from the fundamental theorem of calculus and the chain rule that the density of R , denoted by $f_R(x)$ is

$$\begin{aligned} f_R(x) &= \frac{\partial}{\partial x} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-b)/a} e^{-u^2/2} du \\ &= \frac{1}{\sqrt{2\pi}} e^{-((x-b)/a)^2} \frac{\partial}{\partial x} \frac{x-b}{a} \\ &= \frac{1}{a\sqrt{2\pi}} e^{-((x-b)/a)^2} \end{aligned}$$