Welch's t-test

Page Discussion Read View source View history Tools

Welch's t-test is a **draft** programming task. It is not yet considered ready to be promoted as a complete task, for reasons that should be found in its talk page.

Given two lists of data, calculate the p-value used for Welch's t-test. This is meant to translate R's `t.test(vector1, vector2, alternative="two.sided", var.equal=FALSE)` for calculation of the p-value.

Task Description

Given two sets of data, calculate the p-value:

```
x = {3.0,4.0,1.0,2.1}
y = {490.2,340.0,433.9}
```

Your task is to discern whether or not the difference in means between the two sets is statistically significant and worth further investigation. P-values are significance tests to gauge the probability that the difference in means between two data sets is significant, or due to chance. A threshold level, alpha, is usually chosen, 0.01 or 0.05, where p-values below alpha are worth further investigation and p-values above alpha are considered not significant. The p-value is not considered a final test of significance, only whether the given variable should be given further consideration .

There is more than one way of calculating the t-statistic, and you must choose which method is appropriate for you. Here we use Welch's t-test, which assumes that the variances between the two sets `x` and `y` are not equal. Welch's t-test statistic can be computed:

$$t \quad = \quad \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

where

$\overline{X}_n$ is the mean of set n,

and

N_n is the number of observations in set n,

and

s_n is the square root of the unbiased sample variance of set n, i.e.

$$s_n = \sqrt{\frac{1}{N_n - 1}\sum_{i=1}^{N_n}\left(X_i - \overline{X}_n\right)^2}$$

and the degrees of freedom, ν can be approximated:

$$\nu \quad \approx \quad \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{s_1^4}{N_1^2 (N_1 - 1)} + \frac{s_2^4}{N_2^2 (N_2 - 1)}}$$

The two-tailed p-value, p, can be computed as a cumulative distribution function

$$p_{2-tail} = I_{\frac{\nu}{t^2+\nu}}\left(\frac{\nu}{2}, \frac{1}{2}\right)$$

where I is the regularized incomplete beta function. This is the same as:

$$p_{2-tail} = \frac{\mathrm{B}(\frac{\nu}{t^2+\nu}; \frac{\nu}{2}, \frac{1}{2})}{\mathrm{B}(\frac{\nu}{2}, \frac{1}{2})}$$

Keeping in mind that

$$\mathrm{B}(x; a, b) = \int_0^x r^{a-1}\,(1-r)^{b-1}\,\mathrm{d}r.$$

and

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} = \exp(\ln \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}) = \exp((\ln(\Gamma(x)) + \ln(\Gamma(y)) - \ln(\Gamma(x+y)))$$

p_{2-tail} can be calculated in terms of gamma functions and integrals more simply:

$$p_{2-tail} = \frac{\int_0^{\frac{\nu}{t^2+\nu}} r^{\frac{\nu}{2}-1}\,(1-r)^{-0.5}\,\mathrm{d}r}{\exp((\ln(\Gamma(\frac{\nu}{2})) + \ln(\Gamma(0.5)) - \ln(\Gamma(\frac{\nu}{2} + 0.5)))}$$

which simplifies to

$$p_{2-tail} = \frac{\int_0^{\frac{\nu}{t^2+\nu}} \frac{r^{\frac{\nu}{2}-1}}{\sqrt{1-r}}\,\mathrm{d}r}{\exp((\ln(\Gamma(\frac{\nu}{2})) + \ln(\Gamma(0.5)) - \ln(\Gamma(\frac{\nu}{2} + 0.5)))}$$

The definite integral can be approximated with Simpson's Rule but other methods are also acceptable.

The $\ln(\Gamma(x))$, or `lgammal(x)` function is necessary for the program to work with large `a` values, as Gamma functions can often return values larger than can be handled by `double` or `long double` data types. The `lgammal(x)` function is standard in `math.h` with C99 and C11 standards.

11l