Q8 Spring 2013
Seth Boren

Q8 Statistics

(Dr. Slice; Spring 2013)

Q8.) Consider the standard normal, Student's t, **χ**2, and the Fisher-Snedecor distributions.

a) Describe these distributions.

A standard normal distribution (sometimes known as a bell curve) is a distribution used to represent variables whose distributions are not known. According to the central limits theorem, the averages of samples of random variable observations independently drawn from independent distributions converge to become normally distributed when the observations are sufficiently large. A standard normal distribution will always appear as a symmetrical bell-shaped curve.

A Student's t distribution is a distribution that arises when estimating the mean of a normally distributed population in situations such that sample sizes are small and the population standard deviation is unknown. A Student's t distribution appears like a symmetrical bell-shaped curve (same as the normal standard distribution).

A χ2 (chi-squared) distribution is a distribution of a sum of the squares of k independent standard normal random variables with k-degrees of freedom. The line of the chi-squared distribution depends on the degrees of freedom available.

A Fisher-Snedecor (F) distribution is the ratio of two variance estimates. The distribution is used to in analyses of variance to compute probability values.

b) Discuss their use.

Standard Normal Distributions are widely used in physical, chemical, and biological sciences. Any phenomenon that has a quantitative mean and variance in distribution can be analyzed using a standard normal disturbution.

The Student's t test uses a Student's t distribution to determine if two sets of data are statistically significant from each other.

The Chi-square test uses the chi-square distribution to evaluate the likelihood of observed differences in categorical data arising by chance.

Fisher-Snedecor distributions are used for Analyses of Variance and Multivariate Analyses of Variance.

c) Discuss the relationship amongst them.

The F distribution can be defined as the distribution of the ratio of two independent chi-squared random variables, each divided by their respective degrees of freedom.

Two datasets analyzed for similarities t test can be analyzed in chi-square test so long as each dataset is divided into categories. The t distribution is used to analyze whether differences between samples is significant, whereas the chi-square is used to determine if a relation between samples exists.

d) Where does the binomial distribution fit in?

A binomial distribution is the probability distribution of binomial random variables, which are the number of successes [x] in [n] repeated trials. A binomial distribution requires multiple comparisons with repeated trials, whereas standard normal, Student's t, chi-squared, and F distributions require only a single comparison.

e) Why are we so obsessed with the standard normal?

The only two statistics needed to establish a standard normal distribution is a mean and a variance. More complex distribution can be understood as modifications of the standard normal. Any natural phenomenon for which observations follow the Central Limits Theorem can be studied using the standard normal distribution.

f) Outline a program to statistically test whether or not two samples are drawn from populations having the same mean *without* reference to any of the above distributions?

Assuming we do not know the distribution, mean, variance, or standard deviation of either population.

Assuming the two random samples are independent of one another.

X = mean of the random sample from 1st population

Y = mean of the random sample from 2nd population

μ_X = unknown expected mean of 1st population

μ_Y = unknown expected mean of 2nd population

μ_X - μ_Y = unknown expected difference in the means of the two populations

X – Y = difference between mean of two random samples

The expected value of X – Y is μ_X - μ_Y :

E(X – Y) = E(X) - E(Y) = μ_X - μ_Y

The variance of X – Y is the sum of the variance for μ_X and the variance for μ_Y .

V(X – Y) = V(X) + V(Y) , where V(X) is the variance for sample from 1st population

σ_X = standard deviation of sample one

σ_Y = standard deviation of sample two

V(X) $= \dfrac{(\sigma_X)^2}{m}$ where m = number of sample values in 1st sample

V(Y) $= \dfrac{(\sigma_Y)^2}{n}$ where n = number of sample values in 2nd sample

The standard deviation of X – Y is denoted by $\boldsymbol{\sigma_{X-Y}}$.

$$\boldsymbol{\sigma_{X-Y}} = \sqrt{\frac{(\sigma_X)^2}{m} + \frac{(\sigma_Y)^2}{n}}$$

The standard deviation of the difference between the two sample means is also called the standard error of the mean difference.

By the central limit theorem, the distribution of the "sample mean" values for each population should be normal.

<u>We conduct the following two sample independent "t " test as follows</u>:

Make the null hypothesis that the means of the two populations are equal.

H_0 : $\mu_X = \mu_Y$ versus H_1 : $\mu_X \neq \mu_Y$

Calculate the t statistic from the two random samples.

calculated t statistic = $\dfrac{(\mathbf{X - Y})}{\boldsymbol{\sigma_{X-Y}}}$

Assume a two tailed α level test of 0.05, where α/2 = 0.025 .

The degrees of freedom **ν** for the t test is the smaller of (m -1) or (n-1).

ν = smaller of (m -1) or (n-1).

Obtain the critical value $t_{\left(\frac{\alpha}{2}\right),\nu}$ from the t test table.

Calculate a confidence interval with the formula:

$$(X - Y) \pm \left(t_{\left(\frac{\alpha}{2}\right),\nu} \right)\left(\sigma_{X-Y} \right)$$

where X and Y represent the mean for each of the population samples.

If the calculated t statistic falls within the confidence interval, then the null hypothesis is not rejected.

If the calculated t statistic does not fall within the confidence interval, then the null hypothesis is rejected.

From Wikipedia Source:

The standard error (SE) of a statistic (most commonly the mean) is the standard deviation of its sampling distribution, or sometimes an estimate of that standard deviation.

The equation for the standard error of the mean depicts the relationship between the dispersion of individual observations around the population mean (the standard deviation), and the dispersion of sample means around the population mean (the standard error). Different samples drawn from that same population would in general have different values of the sample mean, so there is a distribution of sampled means (with its own mean and variance). The relationship with the standard deviation is such that, for a given sample size, the standard error equals the standard deviation divided by the square root of the sample size. As the sample size increases, the sample means cluster more closely around the population mean and the standard error decreases.