The Central Limit Theorem and Related Distributions

Gene Quinn

The Central Limit Theorem

We have developed various techniques for the two major problems in classical (parametric) statistics, estimation and hypothesis testing.

The Central Limit Theorem

We have developed various techniques for the two major problems in classical (parametric) statistics, estimation and hypothesis testing.

One might expect at this point that we would develop a catalog of statistical tests specific to certain distributions.

A suprising aspect of classical applied statistics is that the most commonly used techniques at some level assume that the test statistics are either normally distriuted, or follow one of the distributions such as the t and f that are closely associated with samples from a normal population.

The Central Limit Theorem

We have developed various techniques for the two major problems in classical (parametric) statistics, estimation and hypothesis testing.

One might expect at this point that we would develop a catalog of statistical tests specific to certain distributions.

A suprising aspect of classical applied statistics is that the most commonly used techniques at some level assume that the test statistics are either normally distriuted, or follow one of the distributions such as the t and f that are closely associated with samples from a normal population.

The justification for this is the famous *central limit theorem*

The Central Limit Theorem

Loosely speaking, the central limit theorem states that under a very mild set of assumptions, the probability distribution of a sum of independent random variables tends to be normal as the number of terms in the sum increases.

The Central Limit Theorem

Loosely speaking, the central limit theorem states that under a very mild set of assumptions, the probability distribution of a sum of independent random variables tends to be normal as the number of terms in the sum increases.

There are a number of statements of the central limit theorem involving different sets of *sufficient* conditions.

The Central Limit Theorem

Loosely speaking, the central limit theorem states that under a very mild set of assumptions, the probability distribution of a sum of independent random variables tends to be normal as the number of terms in the sum increases.

There are a number of statements of the central limit theorem involving different sets of *sufficient* conditions.

By *sufficient* conditions, we mean conditions which, if met, guarantee that the theorem will hold.

The Central Limit Theorem

Loosely speaking, the central limit theorem states that under a very mild set of assumptions, the probability distribution of a sum of independent random variables tends to be normal as the number of terms in the sum increases.

There are a number of statements of the central limit theorem involving different sets of *sufficient* conditions.

By *sufficient* conditions, we mean conditions which, if met, guarantee that the theorem will hold.

The various statements of the central limit theorem differ in the sufficient conditions they suppose are satisfied.

The Central Limit Theorem

Sufficient conditions that cover most applications are easy to state, but are not the most lenient. Less restrictive sufficient conditions are more difficult to state and usually do not appear in introductory texts.

The Central Limit Theorem

Sufficient conditions that cover most applications are easy to state, but are not the most lenient. Less restrictive sufficient conditions are more difficult to state and usually do not appear in introductory texts.

The version that appears in the text assumes that we have an infinite sequence

$$W_1, W_2, \ldots$$

of random variables with the W_i having the following properties:

- The W_i are independent
- The W_i each have the same distribution
- The mean μ and variance σ^2 of this distribution are finite

The Central Limit Theorem

The first two conditions, that the W_i are independent and identically distributed, is a setting that commonly arises in random sampling, so this is a natural choice.

The Central Limit Theorem

The first two conditions, that the W_i are independent and identically distributed, is a setting that commonly arises in random sampling, so this is a natural choice.

Theorem: Let $W_1, W_2, \ldots$ be an infinite sequence of independent, identically distributed random variables having finite mean μ and finite standard deviation σ^2.

Then for any $a, b \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(a \leq \frac{W_1 + \cdots + W_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz$$

The Central Limit Theorem

An equivalent form that reflects commonly used test statistics more closely is obtained by replacing each W_i by W_i/n and calling their sum $\overline{W}$:

Theorem: Let $W_1, W_2, \ldots$ be an infinite sequence of independent, identically distributed random variables having finite mean μ and finite standard deviation σ^2.

Then for any $a, b \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(a \leq \frac{\overline{W} - \mu}{\sigma/\sqrt{n}} \leq b\right) \; = \; \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz$$

The Central Limit Theorem

This statement of the theorem is adequate for most purposes.

The Central Limit Theorem

This statement of the theorem is adequate for most purposes.

Loosely speaking, the more lenient sets of sufficient conditions require:

- That the random variables W_i be independent
- That the means and variances μ_i and σ_i^2 are finite
- That no small subset of the W_i dominates the variance of their sum

Note that the W_i are not required to be identically distributed in this setting.

This level of generality is not usually required in practice.

The Chi Square Distribution

Definition: The pdf of the random variable

$$U = \sum_{i=1}^{m} Z_j^2$$

where the Z_i are independent and identically distributed as standard normal $Z_i \sim N(0,1)$ is called the **chi square distribution with m degrees of freedom**

The Chi Square Distribution

Theorem: Let

$$Y_1, Y_2, \ldots, Y_n$$

be a random sample from a normal distribution with mean μ and variance σ^2, that is, $Y_i \sim N(\mu, \sigma^2)$.

Then:

- The sample variance S^2 and sample mean $\overline{Y}$ are independent
- $$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

  has a chi square distribution with $n-1$ degrees of freedom.

The t Distribution

Definition: Let Z be a standard normal random variable and U a chi square random variable with n degrees of freedom distributed independently of Z. The distribution of the ratio

$$T_n = \frac{Z}{sqrt\frac{U}{n}}$$

is called the **Student t distribution with n degrees of freedom**

The F Distribution

Definition: Let V and U be independently distributed chi square random variables with m and n degrees of freedom, respectively. The distribution of the ratio

$$F_{m,n} = \frac{V/m}{U/n}$$

is called the **F distribution with** m **and** n **degrees of freedom**