Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Outline

- Make-up Classes
- χ^2 Distribution
- T-Distribution
- F-Distribution
- Hypergeometric Distribution
- R Lab
  - Problem Set 2
  - Chapter 3
  - Chapter 3 Supplemental

χ2 Distribution

- The chi-squared distribution (also chi-square or χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables.
- If Z_1, …, Z_k are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2$$

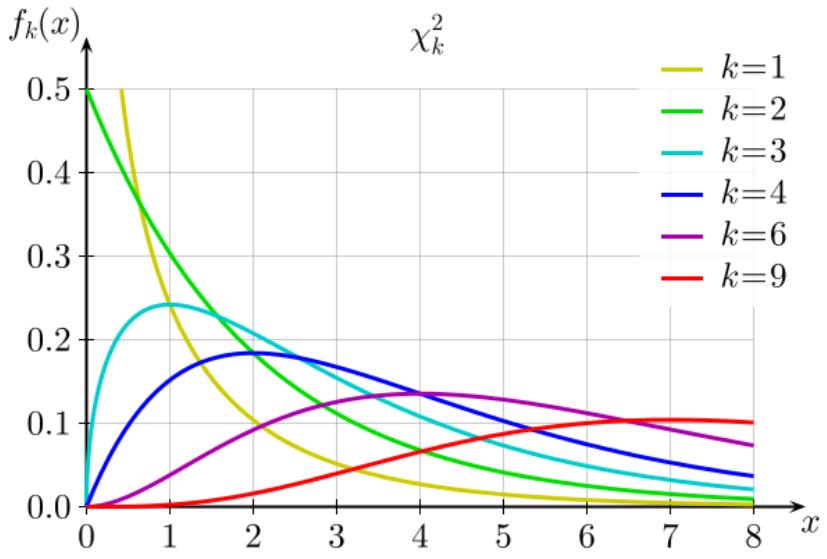
- is distributed according to the chi-squared distribution with k degrees of freedom. This is usually denoted as

$$Q \sim \chi^2(k) \quad or \quad Q \sim \chi_k^2$$

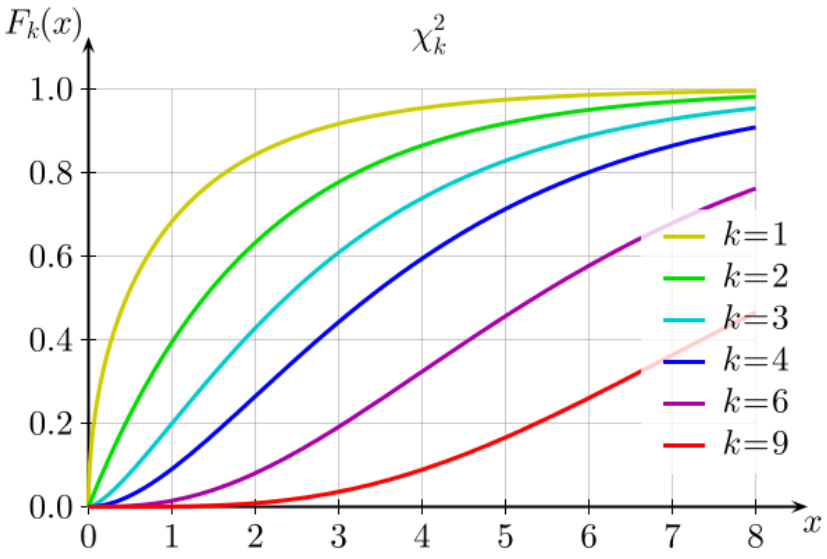
- The chi-squared distribution has one parameter: k — a positive integer that specifies the number of degrees of freedom (i.e. the number of Z_i's)

χ2 Distribution

Probability density function



Cumulative distribution function



χ2 Distribution

- The chi-squared distribution is used in the common chi-squared tests for:
  - goodness of fit of an observed distribution to a theoretical one
  - the independence of two criteria of classification of qualitative data
  - in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.
  - Many other statistical tests also use this distribution, like Friedman's analysis of variance by ranks.

- The chi-squared distribution is a special case of the gamma distribution.

(Student's) t-Distribution

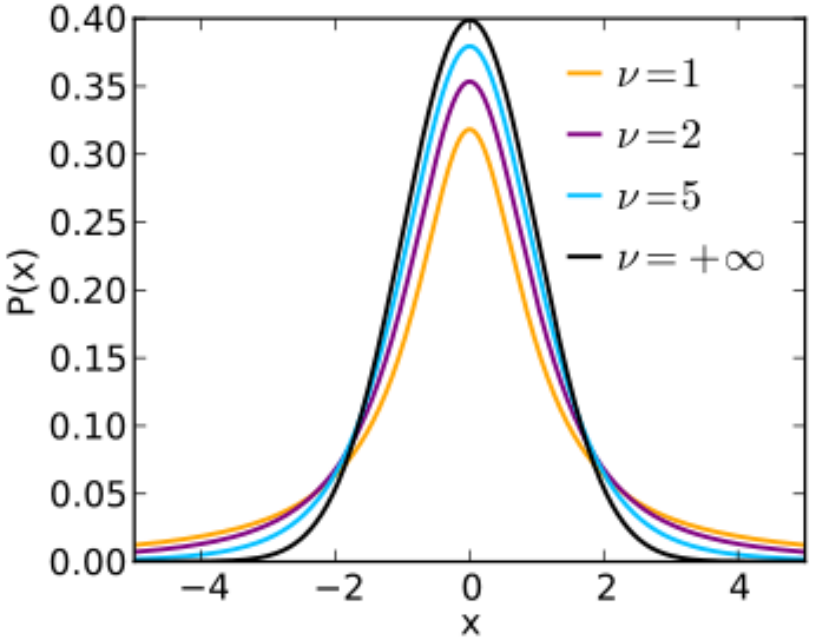
- Student's t-distribution (or simply the t-distribution) is a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small (< 30) and population standard deviation is unknown.
- It plays a role in a number of widely used statistical analyses including
  - the Student's t-test for assessing the statistical significance of the difference between two sample means
  - the construction of confidence intervals for the difference between two population means
  - linear regression analysis.

(Student's) t-Distribution

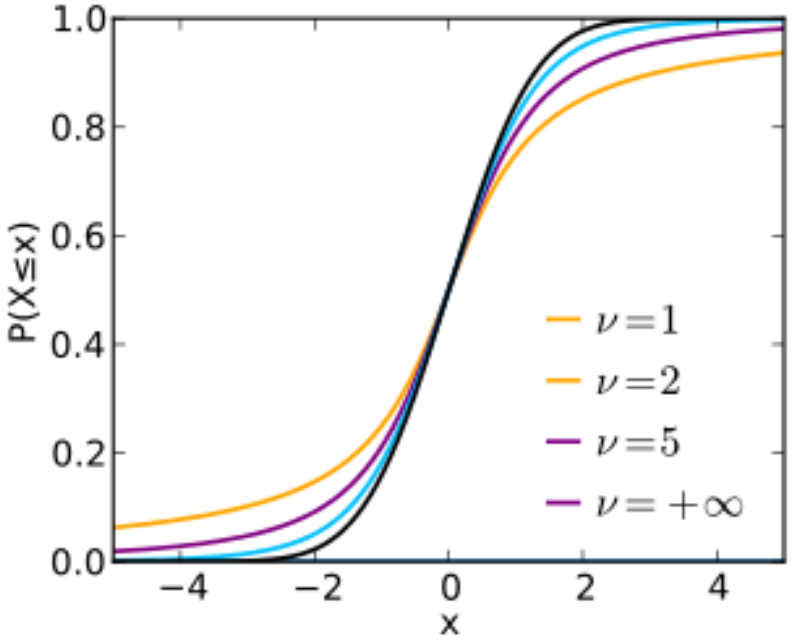
- If the data are normally distributed, then the values of $\frac{(\bar{x}-\mu)}{s/\sqrt{n}}$ follow a T-distribution with n-1 degrees of freedom, where s is the sample standard deviation and $\bar{x}$ is the sample mean.
- The T-distribution is approximately equal to the normal distribution when the sample size is thirty or greater.
- If we take a sample of n = v+1 observations from a normal distribution, compute the sample mean and plot it, and repeat this process infinitely many times (for the same n), we get the t-Distribution probability density function for that n.
- If we compute the sample variance for these n observations, then the t-distribution (for n-1) can be defined as the distribution of the location of the true mean relative to the sample mean, divided by the sample standard error. In this way, the t-distribution can be used to estimate how likely it is that the true mean lies in any given range.
- The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean

(Student's) t-Distribution

Probability density function



Cumulative distribution function

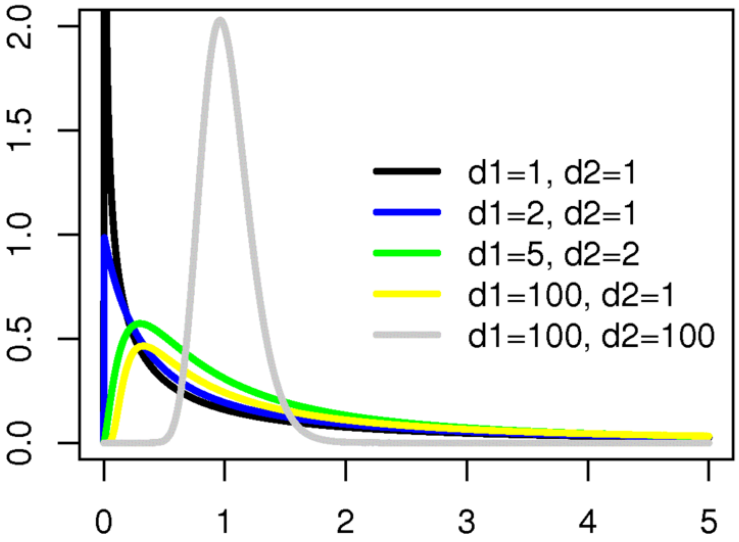


F-Distribution

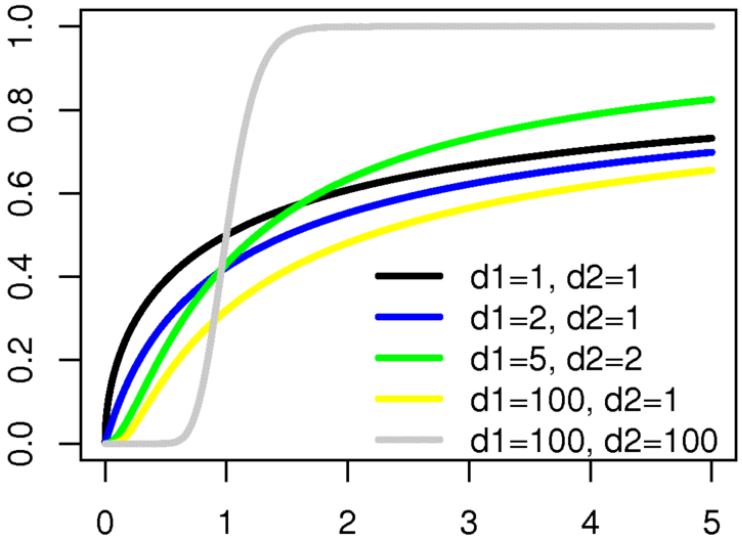
- Also known as Snedecor's F distribution or the Fisher-Snedecor distribution
- The F-distribution is important for testing the equality of two variances.
- It can be shown that the ratio of variances from two independent sets of normally distributed random variables follows an F-distribution.
- Specifically, if the two population variances are equal $\sigma_1^2 = \sigma_2^2$, then $\frac{s_1^2}{s_2^2}$ follows an F-distribution with $n_1 - 1$; $n_2 - 1$ degrees of freedom, where s_1^2 is the sample variance of the first sample set, s_2^2 that of the second sample, n_1 is the number of observations in the first sample and n_2 in the second.

F-Distribution

Probability density function



Cumulative distribution function



Hypergeometric Distribution

- The hypergeometric distribution is a discrete probability distribution that describes the probability of k successes in n draws without replacement from a finite population of size N containing exactly K successes.

- This is in contrast to the binomial distribution, which describes the probability of k successes in n draws with replacement.

Hypergeometric Distribution

- The hypergeometric distribution applies to sampling without replacement from a finite population whose elements can be classified into two mutually exclusive categories like Pass/Fail, Male/Female or Employed/Unemployed.

- As random selections are made from the population, each subsequent draw decreases the population causing the probability of success to change with each draw.

- The following conditions characterize the hypergeometric distribution:

- The result of each draw can be classified into one or two categories.
- The probability of a success changes on each draw.

Hypergeometric Distribution

- A random variable X follows the hypergeometric distribution if its probability mass function (pmf) is given by

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

N is the population size
K is the number of success states in the population
n is the number of draws
k is the number of successes