Other Common Continuous Distributions: Chi-Squared, F, and t

BIOS 6611

CU Anschutz

Week 3

Chi-Squared Distribution

The Normal and Chi Squared Relationship

The **chi-squared** distribution (also chi-square distribution or χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k IID standard normal random variables. In other words...

If $Z_1, Z_2, \ldots, Z_k \overset{iid}{\sim} N(0,1)$, then

$$\begin{aligned} X &\equiv Z_1^2 + Z_2^2 + \ldots + Z_k^2 \\ &\sim \chi_k^2 \end{aligned}$$

Degrees of Freedom

In general, the **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.

In the chi-squared case, there are k numbers that are free to vary, $Z_1, \ldots, Z_k$.

The shape of the chi-square distribution changes as its degrees of freedom change.

Why is the Chi-Squared distribution important?

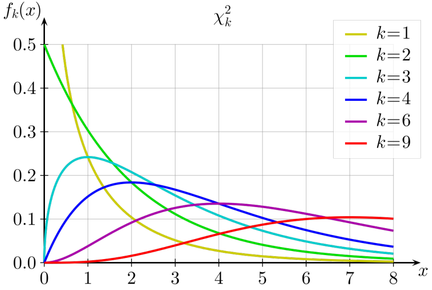
- One of most widely used distributions in inferential statistics
  - Hypothesis testing
  - Confidence Intervals
  - Tests for goodness of fit
  - Tests for independence
  - Tests for comparing models
- Distribution of sample variance
- Not as often applied in direct modeling of natural phenomena
- An important component of $t-$ and $F-$distributions

Chi-Squared Distribution Details

If $X \sim \chi^2_k$,

$$\begin{aligned} x &\in [0, \infty) \\ E[X] &= k \\ Var[X] &= 2k \end{aligned}$$

PDF of Chi-Square Distribution



Example Problem

Let $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Construct X from Y_i such that $X \sim \chi^2_n$, assuming you know μ and σ^2.

Solution:

Example Problem

Let $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \ldots, n$. Construct X from Y_i such that $X \sim \chi^2_n$.

Solution: Need n standard normal random variables, so standardize Y_i:

$$\begin{aligned} Z_i &\equiv \frac{Y_i - \mu}{\sigma} \\ \text{Then } E[Z_i] &= 0 \\ Var[Z_i] &= 1 \end{aligned}$$

Finally, construct X as

$$X \equiv \sum_{i=1}^{n} Z_i^2$$

It follows that $X \sim \chi^2_n$.

F-Distribution

Relationship between F- and Chi-Squared

The **F-distribution** is the ratio of two Chi-squared random variables scaled by their degrees of freedom. It has two parameters, k_1 and k_2, are those are both degrees of freedom. In other words...

If $U_1 \sim \chi^2_{k_1}$ and $U_2 \sim \chi^2_{k_2}$, then

$$\begin{aligned} X &\equiv \frac{U_1/k_1}{U_2/k_2} \\ &\sim \mathrm{F}(k_1, k_2) \end{aligned}$$

Why is the F-distribution important?

Widely used:

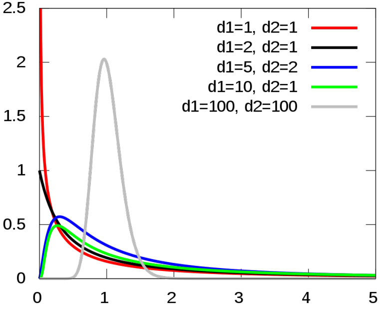
- Distribution of ratio of two independent estimators of population variances
- Tests equality of variances between two independent normal samples
- ANOVA
- Tests for comparing models

F-Distribution Details

If $X \sim F(k_1, k_2)$,

$$
\begin{aligned}
x &\in [0, \infty) \\
E[X] &= \frac{k_2}{k_2 - 2} \text{ for } k_2 > 2 \\
Var[X] &= \frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)} \\
\text{for} & \quad k_2 > 4
\end{aligned}
$$

PDF of F-Distribution



t-distribution (Student's t-distribution)

Relationship between sample mean, sample variance, and t-distribution

If $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and we define the sample mean and sample variance as usual:

$$\begin{aligned} \bar{Y} &= \frac{1}{n}\sum_{i=1}^{n} Y_i \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \end{aligned}$$

then

$$X \equiv \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

We say X has a t-distribution with $n-1$ degrees of freedom.

Why $n-1$ degrees of freedom?

If we know $\bar{Y}$, then it follows that

$$Y_1 + \ldots + Y_n - n\bar{Y} = 0$$

So, if you know $n-1$ of the values of Y_i, then you know what the last value is.

Why is t-distribution important?

Arises when. . .

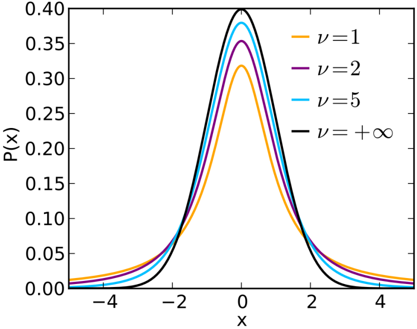
- Estimating mean of normally distributed population with small sample size and unknown true standard deviation
- Testing and confidence intervals for difference between two sample means (t-test)
- Testing significance in regression
- Bayesian analysis of data from a normal family

Details of t-distribution

If $X \sim t(\nu)$,

$$\begin{aligned} x &\in (-\infty, \infty) \\ E[X] &= 0 \text{ for } \nu > 1 \\ Var[X] &= \frac{\nu}{\nu - 2} \text{ for } \nu > 2 \end{aligned}$$

PDF of t-Distribution



Origin of name “Students t-distribution”

- Statistician named William Sealy Gosset published paper in *Biometrika* under pseudonym "Student"
- Gosset worked at Guinness Brewery in Dublin
- He was interested in small sample problems, such as chemical properties of barley with small sample size
- Some speculation as to why used Student:
  - Guinness preferred staff to use pen names when publishing scientific papers
  - Guinness did not want competitors to know they were using the t-test to determine quality of ingredients