Statistical Tests (2)

MAE301 Applied Experimental Statistics

Yi Ren, Yabin Liao

School for Engineering of Matter, Transport Energy
Arizona State University

September 15, 2015

Outline

distribution of sample variance

$E(S^2) = \sigma^2$

Similar to the sample mean, the sample variance

$$S^2 = \frac{1}{\mathbf{n-1}} \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{1}$$

is also a random variable.

The sample distribution of the variance S^2 will be used in learning about the population variance σ^2.

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \tag{2}$$

∞ 1

has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

chi-squared distribution

The chi-squared distribution random variable χ^2 has probability density function

$$f(\chi^2) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{\nu/2-1}e^{-\chi^2/2}, & \chi^2 > 0 \\ 0, & \text{elsewhere} \end{cases} \tag{3}$$

where ν is the degrees of freedom (a positive integer) and the gamma function

$$\Gamma(\nu/2) = \int_0^\infty u^{\nu/2-1}e^{-u}du \tag{4}$$

The mean and variance of chi-square distribution are $\mu = \nu$, $\sigma^2 = 2\nu$.

chi-squared distribution

Similar to T statistic, it is customary to let χ^2_α represent the χ^2 value above which we find an area (or upper-tail probability) of α.

If the sample mean and variance are obtained from 8 measurements. Calculate $\chi^2_{0.95}$ and $\chi^2_{0.05}$.

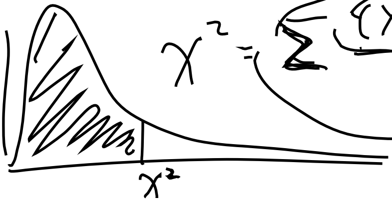
Inference on the population variance

95% of a chi-square distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$

A value falling to the right of $\chi^2_{0.025}$ is not likely to occur unless our assumed value of σ^2 is too small.

A value falling to the left of $\chi^2_{0.975}$ is not likely to occur unless our assumed value of σ^2 is too large.

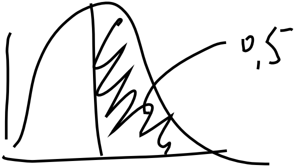
exercise



$$\chi^2 = \sum \frac{(X_i - \bar{x})^2}{\sigma^2}$$

A manufacturer of car batteries guarantees that this batteries will last, on the average, 3 years with a standard deviation of 1 year. Assume that the battery lifetime follows a normal distribution.

If five of these batteries have lifetime of 1.9, 2.4, 3.0, 3.5, and 4.2 years, is the manufacturer still convinced that his batteries have a standard deviation of 1 year?



F-distribution

Previously we have discussed situations where two means are compared (two-sample z-test or t-test). For many applications, variability is equally important and the F-distribution can be used to compare variances of two or more populations.

Consider the case where we take samples of size n_1 and n_2 respectively, from either a single population or two populations. We determine the variances from the samples and then find the ratio s_1^2/s_2^2. If we did this for a very large number of pairs of samples, the ratios would form a distribution known as the F distribution.

F-distribution definition

The statistic F is defined to be the ratio of two independent chi-square random variables, each divided by its number of degrees of freedom

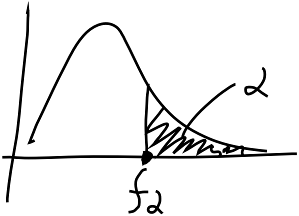
$$F = \frac{U/\nu_1}{V/\nu_2} \tag{5}$$

where U and V are independent random variables having chi-squared distribution with ν_1 and ν_2 degrees of freedom, respectively.

The probability density function of F statistic is

$$h(f) = \begin{cases} \frac{\Gamma((\nu_1+\nu_2)/2)(\nu_1/\nu_2)^{\nu_1/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{f^{\nu_1/2-1}}{(1+\nu_1 f/\nu_2)^{(\nu_1+\nu_2)/2}}, & f > 0 \\ 0, & f \leq 0 \end{cases} \tag{6}$$

F-distribution curve



f_α is the f-value above which we find an area equal to α:

$$P(F > f_\alpha) = \alpha \tag{7}$$

In addition, the f-value has property $f_{1-\alpha}(\nu_1, \nu_2) = 1/f_\alpha(\nu_2, \nu_1)$.

F-distribution with two sample variances

Suppose that random samples of size n_1 and n_2 are selected from two normal populations with variance σ_1^2 and σ_2^2, respectively. The following random variables in terms of the sample variances S_1^2 and S_2^2:

$$U = \frac{(n_1 - 1)S_1^2}{\sigma_1^2}, \qquad V = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \tag{8}$$

have chi-squared distributions with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

This results in the F-distribution for the sample variances

$$F = \frac{U/\nu_1}{V/\nu_2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \tag{9}$$

exercise

$$f = \frac{s_1^2}{s_2^2}$$

Consider the following measurements of the heat producing capacity of the coal produced by two mines (in millions of calories per ton):

Mine 1	8260	8130	8350	8070	8340	
Mine 2	7950	7890	7900	8140	7920	7840

Do the measurements support the statement that the two population variances are equal?

summary of the class

- χ^2 statistic (for one sample variance) and F statistic (for comparison between two sample variances)

Python code for demos in the class

```
# chi-square distribution
from scipy.stats import chi2
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1, 1)
df = 10
x = np.linspace(chi2.ppf(0.0001, df),chi2.ppf(0.9999, df), 100)
ax.plot(x, chi2.pdf(x, df),'r-', lw=5, alpha=0.6, label='chi2 pdf')

# chi-square test
from scipy.stats import chi2
import matplotlib.pyplot as plt
sigma = 1.0
x = [1.9, 2.4, 3.0, 3.5, 4.2]
s = np.std(x, ddof=1)
chisquare = (len(x)-1)*s**2/sigma**2
chi2.cdf(chisquare,len(x)-1)
```

Python code for demos in the class

```
# f distribution
from scipy.stats import f
import matplotlib.pyplot as plt
fig, ax = plt.subplots(1, 1)
df1 = 20
df2 = 10
x = np.linspace(f.ppf(0.0001, df1, df2),f.ppf(0.9999, df1, df2), 100)
ax.plot(x, f.pdf(x, df1, df2),'r-', lw=5, alpha=0.6, label='f pdf')

# f test
from scipy.stats import f
sample1 = [8260,8130,8350,8070,8340]
sample2 = [7950,7890,7900,8140,7920,7840]
xbar1 = np.mean(sample1)
s1 = np.std(sample1,ddof=1)
df1 = len(sample1)-1
xbar2 = np.mean(sample2)
s2 = np.std(sample2,ddof=1)
df2 = len(sample2)-1
F = s1**2/s2**2
f.cdf(F,df1,df2)
```