ACADGILD

Mastering Data Science

Statistics

Session 18 - Student t Distribution, Chi-Squared Distribution and F Distribution

Agenda

1. Student T Distribution
2. Determining Student t Values
3. Using the T table for all values
4. F Distribution
5. Determining Values of F
6. Check Requisite Conditions
7. Inference about Population Variance
8. Testing and Estimating Population Variance
9. Comparing Two Populations
10. Making Inference about $\mu_1 - \mu_2$
11. **Test Statistics for $\mu_1 - \mu_2$** (Equal Variances)
12. Test Statistics for $\mu_1 - \mu_2$ (Non-equal Variances)
13. Test Estimate for $\mu_1 - \mu_2$ (Equal Variances)
14. Inference about the ratio of the Two Variables

Student t Distribution

Here the letter t is used to represent the random variable, hence the name. The density function for the Student t distribution is as follows,

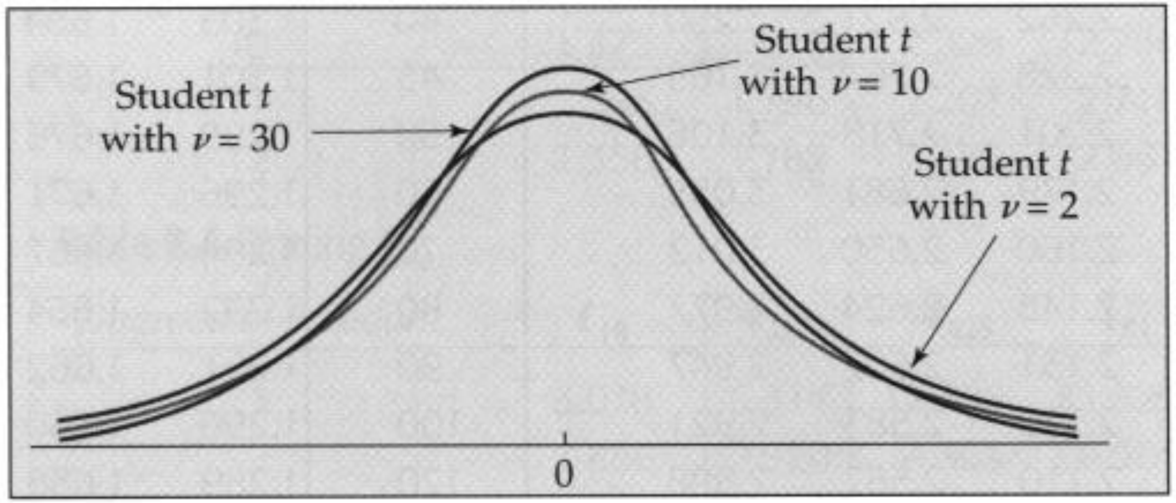
$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2}$$

ν (nu) is called the degrees of freedom, and

Γ (Gamma function) is Γ (k)=(k-1)(k-2)…(2)(1)

Student t Distribution

In much the same way that μ and σ define the normal distribution, ν, the degrees of freedom, defines the student t distribution:

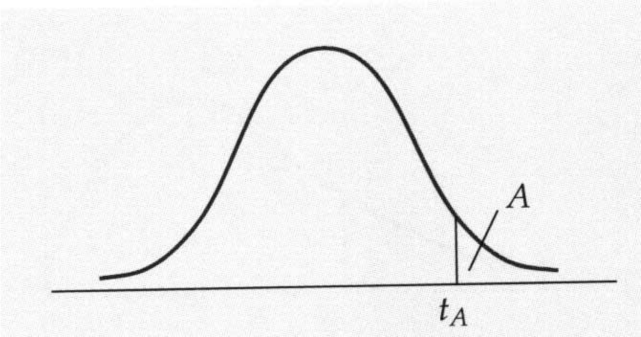


As the number of degrees of freedom increases, the t distribution approaches the standard normal distribution.

Determining Student t values

- The Student 't' distribution is used extensively in statistical inferences.
- That is, values of a Student 't' random variable with ν degrees of freedom such that:

$$P(t > t_{A,\nu}) = A$$



- The values for A are pre-determined "critical" values, typically in the 10%, 5%, 2.5%, 1% and 1/2% range.

Using the t table for all values

Example

➢ If we want the value of t with 10 degrees of freedom such that the area under the Student ‘t’ curve is 0.05:

Area under the curve value (t_A) : COLUMN

$$t_{.05,10}$$

$t_{.05,10}=1.812$

Degrees of Freedom : ROW

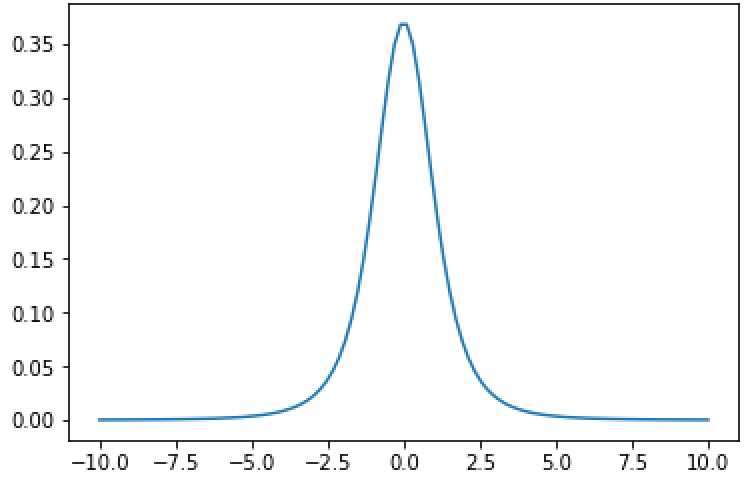
DEGREES OF FREEDOM	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106

Student T Distribution using Python

```
import scipy.stats
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(-10, 10, 100)
df = 3.34
mean, var = scipy.stats.t.stats(df, moments='mv')
print('mean: {:.2f}, variance: {:.2f}'.format(mean, var))
plt.plot(x, scipy.stats.t.pdf(x,df))
plt.show()
```

```
mean: 0.00, variance: 2.49
```



Practice Code

F Density Function

The F Density Function is given by:

$$f(F) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{F^{\frac{\nu_1 - 2}{2}}}{\left(1 + \frac{\nu_1 F}{\nu_2}\right)^{\frac{\nu_1 + \nu_2}{2}}}$$

F > 0. Two parameters define this distribution, and like we've already seen these are again **degrees of freedom.**

is ν_1 the "numerator" degrees of freedom and

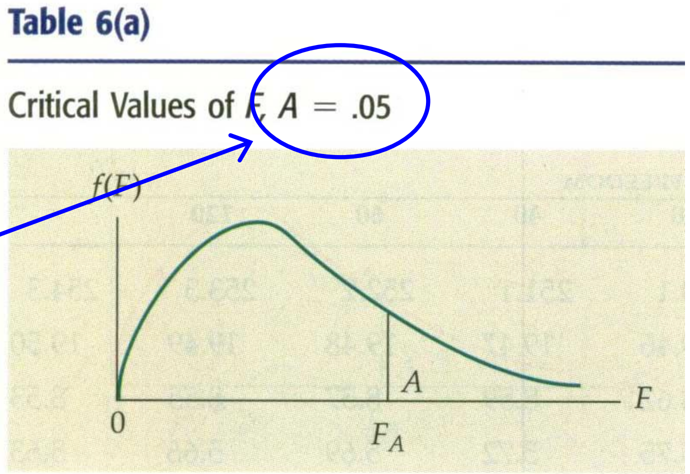
is ν_2 the "denominator" degrees of freedom.

Example

What is the value of F for 5% of the area under the right hand “tail” of the curve, with a numerator degree of freedom of 3 and a denominator degree of freedom of 7?

Solution: Use the F look-up

There are different tables for different values of A. Make sure you start with the correct table!!



$$F_{.05,3,7}$$

Denominator Degrees of Freedom : ROW

Numerator Degrees of Freedom : COLUMN

Determining Values of F

For areas under the curve on the left hand side of the curve, we can leverage the following relationship:

$$F_{1-A,v_1,v_2} = \frac{1}{F_{A,v_2,v_1}}$$

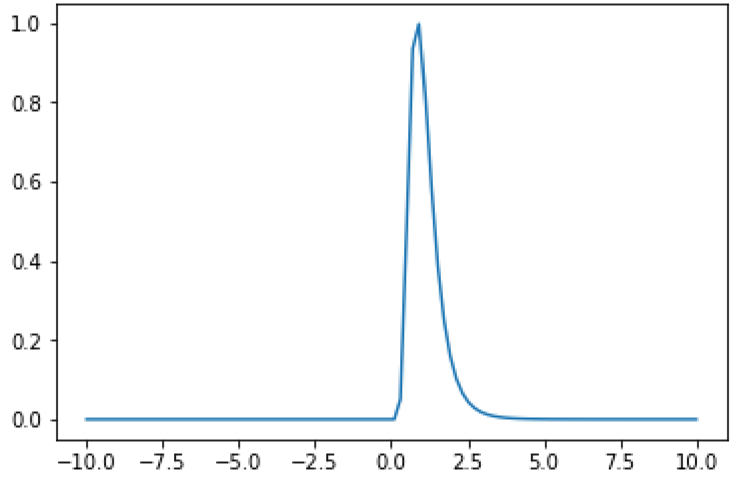
Pay close attention to the order of the terms!

F Distribution using Python

```
import scipy.stats
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(-10, 10, 100)
dfn = 29
dfd = 18
mean, var, skew, kurt = scipy.stats.f.stats(dfn, dfd, moments='mvsk')
print('mean: {:.2f}, variance: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))
plt.plot(x, scipy.stats.f.pdf(x,dfn, dfd))
plt.show()
```

```
mean: 1.12, variance: 0.28, skewness: 1.81, kurtosis: 7.07
```



[Practice Code]

Check Requisite Conditions

- The Student ‘t’ distribution is robust, which means that if the population is non-normal, the results of the t-test and confidence interval estimate are still valid provided that the population is not extremely non-normal.
- To check this requirement, draw a histogram of the data and see how bell shaped the resulting figure is. If a histogram is extremely skewed (say in that case of an exponential distribution), that could be considered “extremely non-normal” and hence t-statistics would not be valid in this case.

Inference About Population Variation

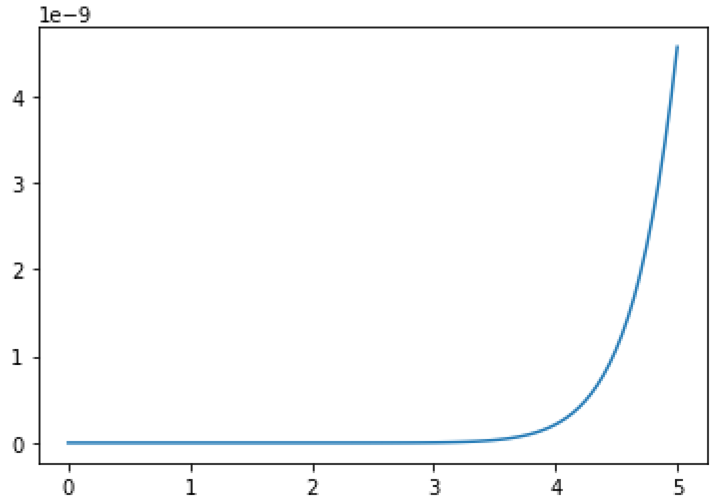
- If we are interested in drawing inferences about a population's variability, the parameter we need to investigate is the population variance: σ^2
- The sample variance (s2) is an unbiased, consistent and efficient point estimator for σ^2 .
- Moreover, the statistic, $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$, has a chi-squared distribution, with n–1 degrees of freedom.

Chi-Square Distribution using Python

```
import scipy.stats
import numpy as np
import matplotlib.pyplot as plt

x = np.linspace(0, 5, 100)
df = 34
mean, var, skew, kurt = scipy.stats.chi2.stats(df, moments='mvsk')
print('mean: {:.2f}, variance: {:.2f}, skewness: {:.2f}, kurtosis: {:.2f}'.format(mean, var, skew, kurt))
plt.plot(x, scipy.stats.chi2.pdf(x,df))
plt.show()
```

```
mean: 34.00, variance: 68.00, skewness: 0.49, kurtosis: 0.35
```



[Practice Code]

Testing and Estimating Population Variance

- Combining this statistic: $\chi^2 = \dfrac{(n-1)s^2}{\sigma^2}$

- With the probability statement: $P\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$

Yields the confidence interval estimator for σ^2

$$LCL = \frac{(n-1)s^2}{\chi^2_{\alpha/2}} \qquad UCL = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

lower confidence limit *upper confidence limit*

Comparing Two Populations

- Previously we looked at techniques to estimate and test parameters for one population:

Population Mean μ, Population Variance σ^2

We still consider these parameters when we are looking at two populations, however our interest will now be:

- The difference between two means
- The ratio of two variances

Difference between Two Means

- In order to test and estimate the difference between two population means, we draw random samples from each of two populations. Initially we will consider independent samples, that is samples that are completely unrelated to one another.

Because we compare two population means, we use the statistic: $\bar{x}_1 - \bar{x}_2$

Difference between Two Means

- In order to test and estimate the difference between two population means, we draw random samples from each of two populations. Initially we will consider independent samples, that is samples that are completely unrelated to one another.

Because we compare two population means, we use the statistic: $\bar{x}_1 - \bar{x}_2$

Sampling Distribution of Comparing two population means

- $\bar{x}_1 - \bar{x}_2$ is normally distributed if the original population is normal – or – approximately normal if the population are non-normal and the sample sizes are large (n_1, n_2 > 30)

- The expected value of $\bar{x}_1 - \bar{x}_2$ is $\mu_1 - \mu_2$

- The variance of $\bar{x}_1 - \bar{x}_2$ is $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

- The Standard error is $\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Making Inferences about $\mu_1 - \mu_2$

- Since $\bar{x}_1 - \bar{x}_2$ is normally distributed, If the original population is normal/approximately normal if the population are non-normal and the sample sizes are large (n_1, n_2 > 30), then:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Z is a standard normal (or approximately normal) random variables. We could use this to build test statistics or confidence interval estimators for $\mu_1 - \mu_2$

Making Inferences about $\mu_1 - \mu_2$

➢ Except that, in practice, the z statistics is rarely used since the population variances are unknown.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

??

Instead we use a t-statistic. We consider two cases for the unknown population variances: When we believe they are **equal** and conversely when they are **not equal.**

When are Variances Equal?

➢ How do we know when the population variances are equal?

Since the Population variances are unknown, we cant know for certain whether they are equal, but we can **examine the sample variances** and **informally judge** their relative values to determine whether we can assume that the population variance are equal or not.

Test Statistics for $\mu_1 - \mu_2$ (Equal Variances)

1. Calculate s_p^2 - The Pooled variance estimator as,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

2. And use it here as,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \nu = n_1 + n_2 - 2$$

Degrees of Freedom

CI Estimator for $\mu_1 - \mu_2$ (Equal Variances)

1. The confidence interval estimator for $\mu_1 - \mu_2$ when the population variances are equal is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad \nu = n_1 + n_2 - 2$$

Pooled Variance Estimator *Pooled Variance Estimator*

Test Statistics for $\mu_1 - \mu_2$ (Unequal Variances)

- The test statistic for $\mu_1 - \mu_2$ when the population variances are unequal is given by:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Degrees of Freedom

- Likewise the confidence interval estimator is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Inference about the ratio of two variances

- So far we have looked at comparing measures of central location, namely the mean of two populations.
- When looking at two population variances, we consider the ratio of the variances, i.e. the parameter of the interest to us is:

$$\sigma_1^2/\sigma_2^2 \quad or \quad \frac{\sigma_1^2}{\sigma_2^2}$$

- The sampling statistic: $\boxed{\dfrac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}}$ is F distributed with $\nu_1 = n_1 - 1 \;\; and \;\; \nu_2 = n_2 - 1$ egrees of freedom.

Inference about the ratio of two variances

- Our Null Hypothesis is always:

  H_0: $\sigma_1^2 / \sigma_2^2 = 1$

That is the variances of the two populations will be equal, hence their ratio will be one.

- Therefor, our statistic simplifies to: $F = s_1^2 / s_2^2$
  - df1 = n_1 - 1
  - df2 = n_2 - 1

Email us - support@acadgild.com