INTRODUCTION TO PROBABILITY AND STATISTICS FOURTEENTH EDITION

Chapter 10
Inference from Small Samples

INTRODUCTION

○ When the sample size is small, the estimation and testing procedures of Chapter 8 are not appropriate.

○ There are equivalent small sample test and estimation procedures for

   ✓ μ, the mean of a normal population

   ✓ $\mu_1 - \mu_2$, the difference between two population means

   ✓ σ^2, the variance of a normal population

   ✓ The ratio of two population variances, σ_1^2 / σ_2^2.

GENERAL PRINCIPLE

Once we observe data: (1) we specify a model that can describe the data; (2) we focus on model parameters.

For a dataset, we do the following:
1. Model specification
2. Point estimtor(s)
3. Standardize the point estimator and identify its distribution
   1. Find the Confidence Interval
   2. Find the test statistics in a Hypothesis Test

MODELS WE HAVE IN THIS COURSE

- No assumption on the model, but
  - n is large: ch8, 9 Central Limit Theorem
    - $X \sim \mathscr{G}(\mu, \sigma^2)$ for some distribution $\mathscr{G}$
    - $X \sim Bernoulli(p)$
  - n is small: ch15 Nonparametric statistics
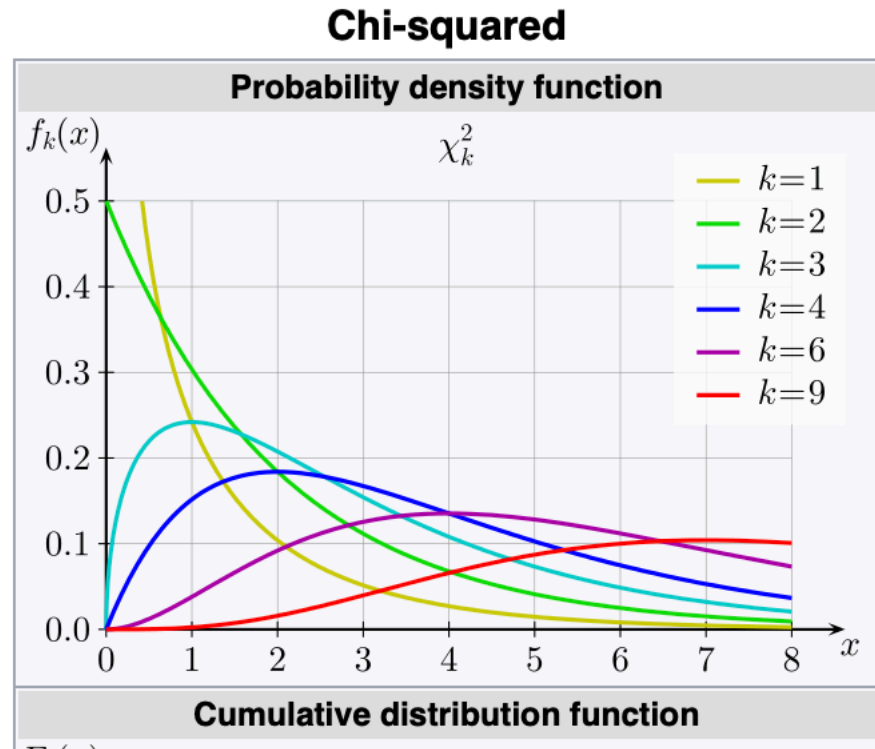- n is small but $X \sim N(\mu, \sigma^2)$: ch 10

10.1 Student's t distribution

CHI-SQUARED DISTRIBUTION

If $Z_i \overset{i.i.d.}{\sim} N(0,1)$ for $i = 1, \ldots, m$, then

$$\sum_{i=1}^{m} Z_i^2 \sim \chi_m^2.$$

χ_m^2 is called the chi-squared distribution with degrees of freedom m.

Chi-squared

Probability density function

$f_k(x)$

χ_k^2

—	$k=1$
—	$k=2$
—	$k=3$
—	$k=4$
—	$k=6$
—	$k=9$

0.5

0.4

0.3

0.2

0.1

0.0

0   1   2   3   4   5   6   7   8 x

Cumulative distribution function

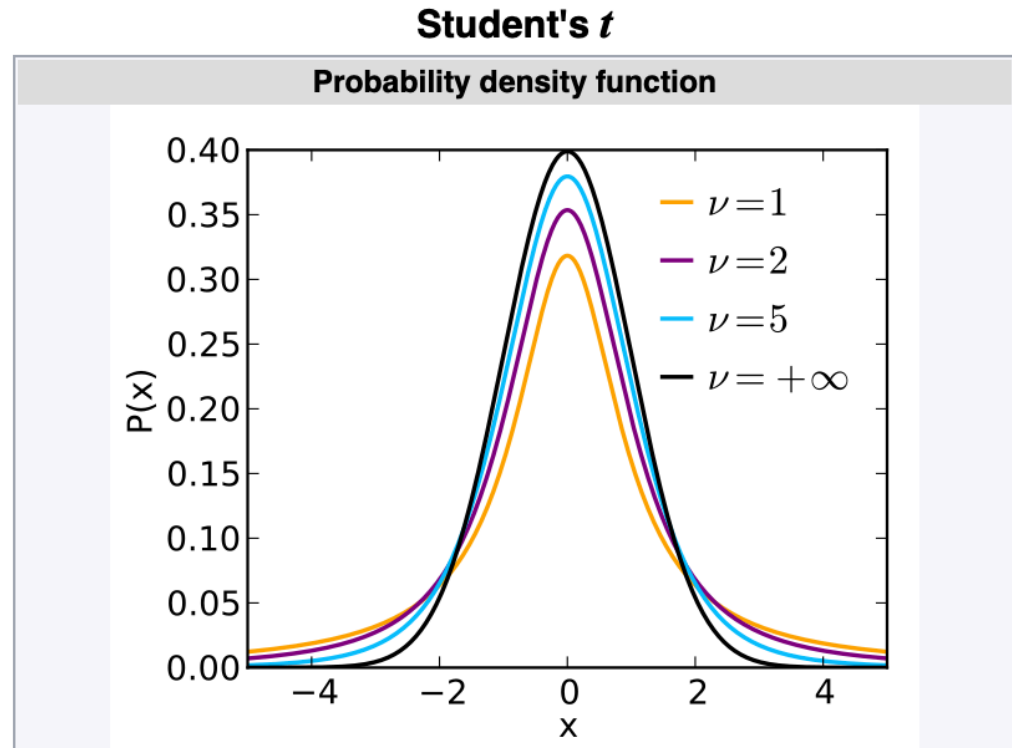# STUDENT'S t DISTRIBUTION

If $U \sim \chi_m^2$,
$Z \sim N(0,1)$, and U
and Z are
independent, then

$$\frac{Z}{\sqrt{U/m}} \sim t_m.$$

t_m is called the t
distribution with
degrees of
freedom m.

Student's t

Probability density function

Features of Student's t distribution

1. Symmetric and mounded shaped.

2. When degrees of freedom goes to infinity, t_m approaches the standard normal distribution.

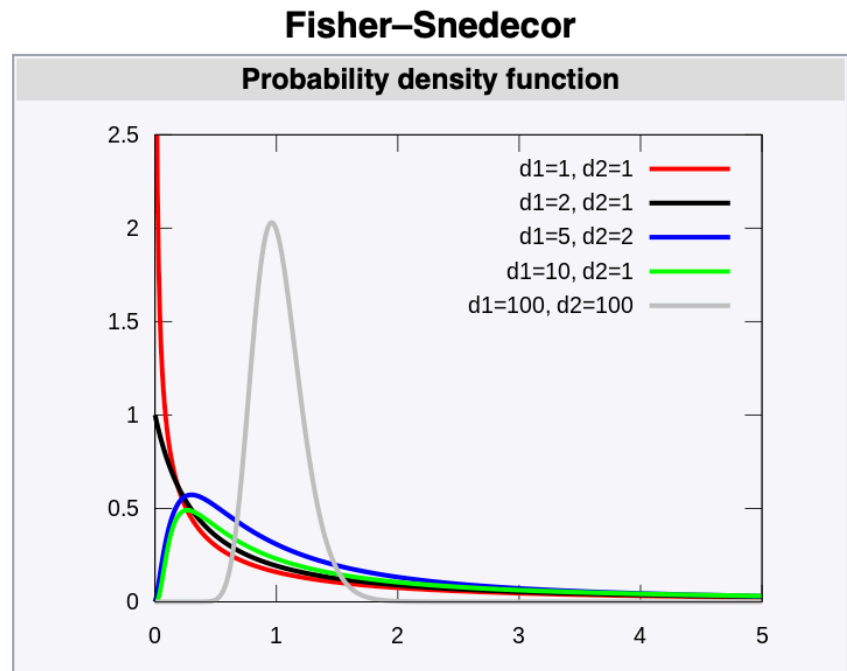Statistician William Sealy Gosset, known as "Student"

F DISTRIBUTION

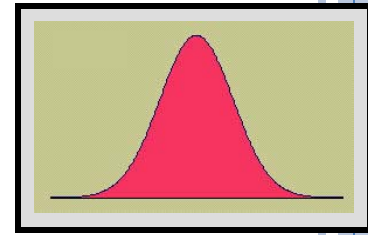If $U \sim \chi^2_m$, $V \sim \chi^2_n$, and U and V are independent, then

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

$F_{m,n}$ is called the F distribution with degrees of freedom m and n.

Fisher–Snedecor

Probability density function

THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

○ When we take a sample from a normal population, the sample mean $\bar{X}$ has a normal distribution for any sample size $n,$ and

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution.

○ But if σ is unknown, and we must use S to estimate it, the resulting statistic **is not normal**. $\dfrac{\bar{X} - \mu}{S/n}$ is not normal!
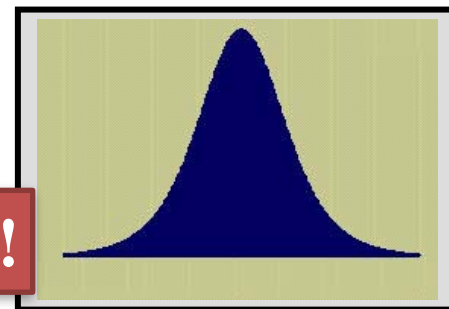
STUDENT'S T DISTRIBUTION

○ Fortunately, this statistic does have a sampling distribution that is well known to statisticians, called the **Student's t distribution**, with **_n_-1 degrees of freedom.** $t = \dfrac{\bar{X} - \mu}{S/n} \sim t_{(n-1)}$

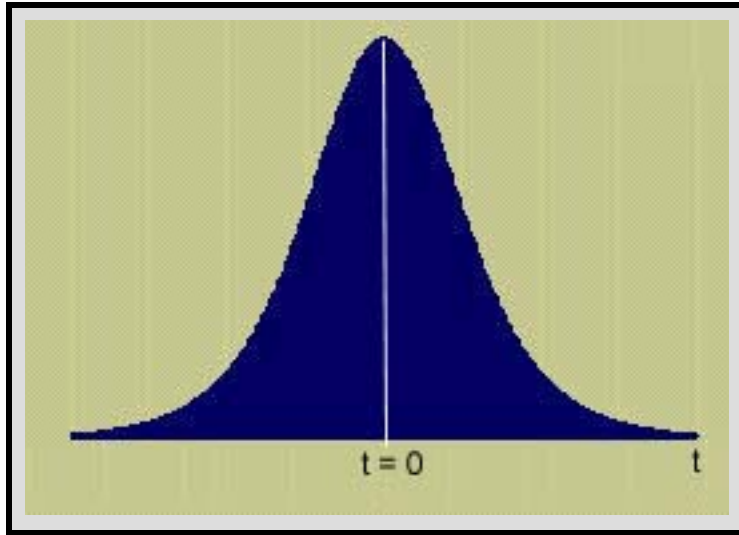**Why? Go to Math Stat!**

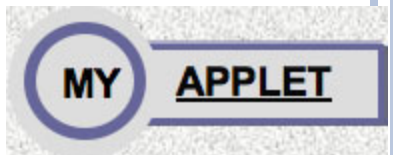- We can use this distribution to create estimation testing procedures for the population mean μ.
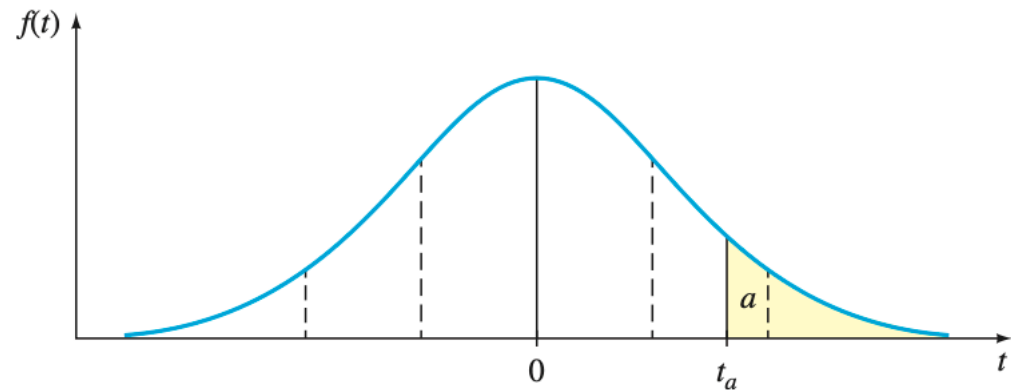
PROPERTIES OF STUDENT'S *T*

- **Mound-shaped** and symmetric about 0.
- **More variable than z**, with "heavier tails"

○ Shape depends on the sample size *n* or the **degrees of freedom, *n*-1.**

MY APPLET

○ As *n* increases the shapes of the *t* and *z* distributions become almost identical.

STUDENT'S t TABLE

- I use $t_{\nu;\alpha}$ to denote critical point of the ν distribution for the right-tailed probability of α.

- $P(t_\nu > t_{\nu;\alpha}) = \alpha$.

■ Table 10.1 Format of the Student's t Table from Table 4 in Appendix I

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
.	.	.	.	.	.	.
.	.	.	.	.	.	.
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
30	1.310	1.697	2.042	2.457	2.750	30
.	.	.	.	.	.	.
.	.	.	.	.	.	.
100	1.290	1.660	1.984	2.364	2.626	100
200	1.286	1.653	1.972	2.345	2.601	200
300	1.284	1.650	1.968	2.339	2.592	300
400	1.284	1.649	1.966	2.336	2.588	400
500	1.283	1.648	1.965	2.334	2.586	500
inf.	1.282	1.645	1.96	2.326	2.576	inf.

USING THE *T*-TABLE

- Table 4 gives the values of t that cut off certain critical values in the tail of the t distribution.
- Index **df** and the appropriate tail area **a** to find t_a, the value of t with area **a** to its right.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764
11	1.363	1.796	2.201	2.718
12	1.356	1.782	2.179	2.681
13	1.350	1.771	2.160	2.650
14	1.345	1.761	2.145	2.624
15	1.341	1.753	2.131	2.602

For a random sample of size $n = 10$, find a value of t that cuts off .025 in the right tail.

Row = $df = n - 1 = 9$

Column subscript = $a = .025$

$t_{.025} = 2.262$

Recall Ch 8 & 9

Suppose $X_i \overset{i.i.d.}{\sim} \mathcal{G}(\mu, \sigma^2)$ for $i = 1,\ldots,n$.

(1) When σ is known, the Central Limit Theorem:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d.} N(0,1) \text{ as } n \to \infty$$

(2) When σ is unknown, Slusky's theorem:

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \xrightarrow{d.} N(0,1) \text{ as } n \to \infty$$

$\xrightarrow{d.}$
- **Asymptotically distributed as** 漸進分配
- **Converges in distribution** 分配收斂

CHAPTER 10

Suppose $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

(1) When σ is known,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \, .$$

(2) When σ is unknown,

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1} \, .$$

10.2 SMALL-SAMPLE INFERENCES CONCERNING A POPULATION MEAN

KEY INGREDIENTS (WHEN σ IS UNKNOWN)

1. Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and we focus on the estimation of μ.

2. Point estimator of μ: $\bar{X}$

3. If σ^2 is unknown, we use the sample variance S^2 to estimate it. Then, the *standardized point estimator and its distribution*: $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$.

Confidence Interval

$$1 - \alpha = P(-t_{(n-1);\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{(n-1);\alpha/2})$$

$$= P(-t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}})$$

$$= P(-t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}} < \mu - \bar{X} < t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}})$$

$$= P(\bar{X} - t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}})$$

- With data, the $100(1 - \alpha)\%$ CI for the population mean μ as $\bar{X} \pm t_{(n-1);\alpha/2}\frac{S}{\sqrt{n}}$.

HYPOTHESIS TEST

1. $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ (or > or <)
2. Set the value of α.
3. The test statistic and its sampling distribution is

$$t_{STAT} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{(n-1)}.$$

4. Calculate the realized statistics, $t^* = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}.$

5. Find the rejection region or p-value.
6. Conclude.

    - Reject or do not reject H_0.

    - Conclude in plain language and get back to the scenario.

EXAMPLE

A sprinkler system is designed so that the average time for the sprinklers to activate after being turned on is no more than 15 seconds. A test of 6 systems gave the following times:

$$17, 31, 12, 17, 13, 25$$

Is the system working as specified? Test using $\alpha = .05$.

$$H_0 : \mu = 15 \text{ (working as specified)}$$

$$H_a : \mu > 15 \text{ (not working as specified)}$$

EXAMPLE

Data: 17, 31, 12, 17, 13, 25

First, calculate the sample mean and standard deviation, using your calculator or the formulas in Chapter 2.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{115}{6} = 19.167$$

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{2477 - \frac{115^2}{6}}{5}} = 7.387$$
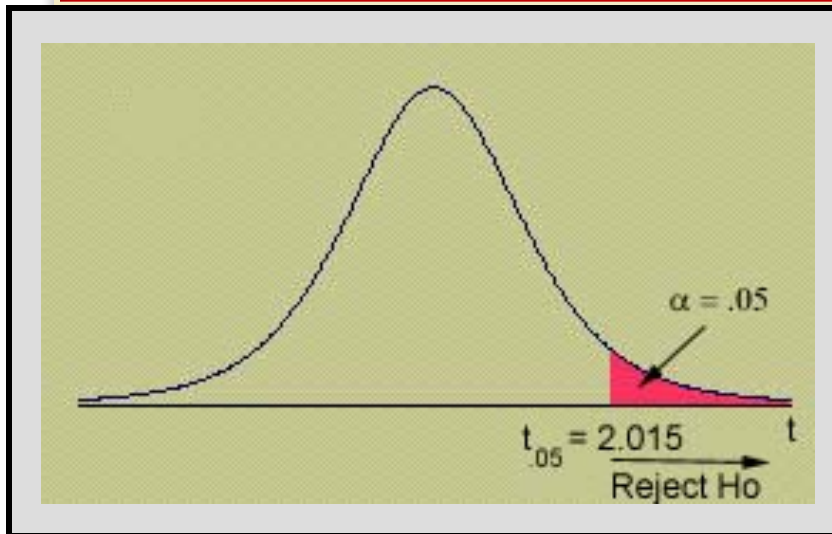
EXAMPLE

Data: 17, 31, 12, 17, 13, 25

Calculate the test statistic and find the rejection region for α =.05.

Test statistic :

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{19.167 - 15}{7.387 / \sqrt{6}} = 1.38$$

Degrees of freedom :

$$df = n - 1 = 6 - 1 = 5$$

α = .05

$t_{.05} = 2.015$ t

Reject Ho

Rejection Region: Reject H_0 if $t > 2.015$. If the test statistic falls in the rejection region, its p-value will be less than α = .05.

CONCLUSION

Data: 17, 31, 12, 17, 13, 25

Compare the observed test statistic to the rejection region, and draw conclusions.

$$H_0 : \mu = 15$$
$$H_a : \mu > 15$$

$$\text{Test statistic} : t = 1.38$$
$$\text{Rejection Region} :$$
$$\text{Reject } H_0 \text{ if } t > 2.015.$$

Conclusion: For our example, $t = 1.38$ does not fall in the rejection region and H_0 is not rejected. There is insufficient evidence to indicate that the average activation time is greater than 15.

STATISTICAL CONCLUSION (CAREFUL!)

○ Because realized statistic does not fall into the rejection region, so we do not reject H_0.

○ Do not reject H_0

○ Insignificance evidence to show H_0 is incorrect $\approx$ Insignificant evidence to show that H_a is correct

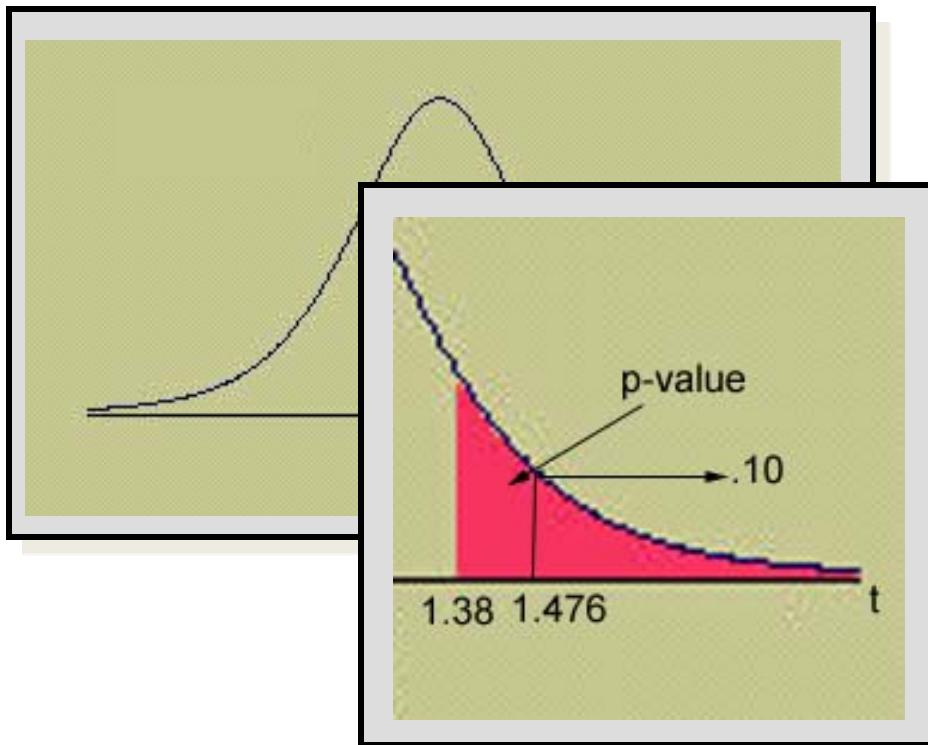# SOP with Rejection Region Approach

1. $H_0 : \mu = 15$ versus $H_a : \mu > 15$

2. Set $\alpha = 0.05$.

3. Define the test statistic $t_{STAST} = \dfrac{\bar{X} - 15}{S/\sqrt{n}} \sim t_{(6-1)}$.

4. Calculate the realized test statistic $t* = 1.38$.

5. Rejection region $\{t : t > 2.015\}$.

6. Because $t*$ does not fall into the rejection region, we do not reject H_0.

APPROXIMATING THE P-VALUE

○ You can only approximate the *p*-value for the test using Table 4.

df	$t_{.100}$	$t_{.050}$
1	3.078	6.314
2	1.886	2.920
3	1.638	2.353
4	1.533	2.132
5	1.476	2.015

p-value

.10

1.38 1.476   t

Since the observed value of *t* = 1.38 is smaller than $t_{.10}$ = 1.476,

p-value > .10.

p-VALUE APPROACH

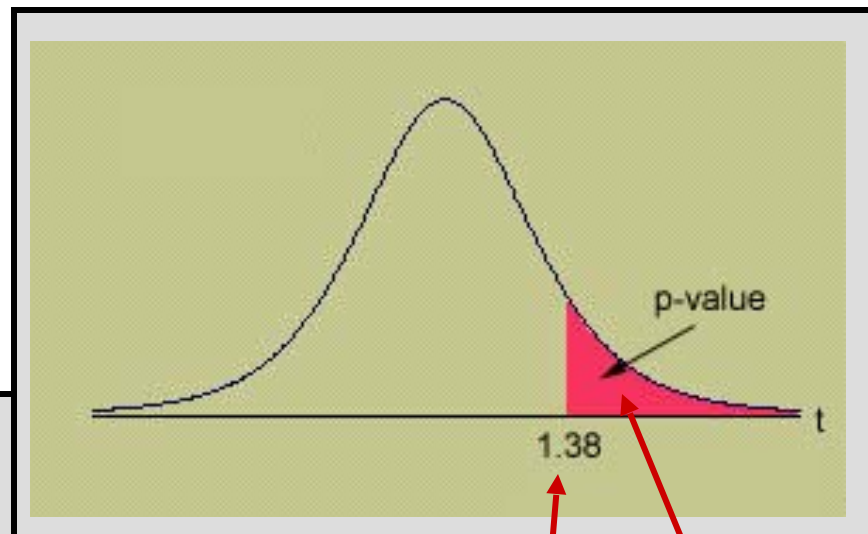5. The p-value $> 0.10 > \alpha = 0.05$

6. We do not rejection H_0.

THE EXACT *P*-VALUE

◦ You can get the exact *p*-value using some calculators or a computer.

p-value = .113 which is greater than .10 as we approximated using Table 4.

One-Sample T: Times

```
Test of mu = 15 vs > 15
```

Variable	N	Mean	StDev	SE Mean	95% Lower Bound	T	P
Times	6	19.1667	7.3869	3.0157	13.0899	1.38	0.113

10.3

Small sample inference for the difference between two population means: independent random sample (SKIPPED)

10.4

Small-sample inferences for the difference between two population means: A paired difference test (Skipped)

10.5 INFERENCES CONCERNING A POPULATION VARIANCE

KEY INGRENDIENTS

1. Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

2. Point estimator of σ^2: use sample variance S^2 to estimate σ^2.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

3. Standardize the point estimator and find its distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

An idea but not proof: $\dfrac{(n-1)S^2}{\sigma^2} = \sum \left(\dfrac{X_i - \bar{X}}{\sigma} \right)^2 \approx \sum Z_i^2$
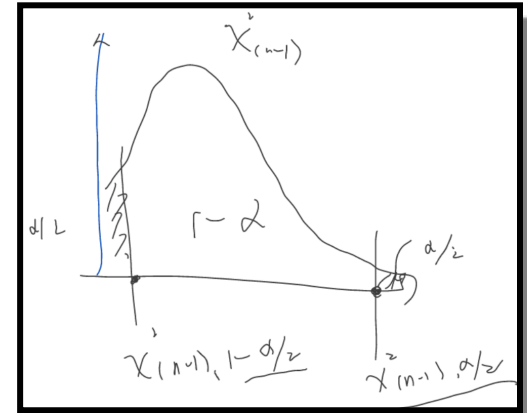
A $100(1-\alpha)\%$ CONFIDENCEN INTERVAL FOR σ^2

◦ In a similar manner,

$$1 - \alpha = P\left(\chi^2_{(n-1);1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{(n-1);\alpha/2}\right)$$

$$= P\left(\frac{1}{\chi^2_{(n-1);\alpha/2}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{(n-1);1-\alpha/2}}\right)$$

$$= P\left(\frac{(n-1)S^2}{\chi^2_{(n-1);\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1);1-\alpha/2}}\right).$$

◦ Thus, a $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-1)s^2}{\chi^2_{(n-1),\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{(n-1),1-\alpha/2}})\right].$$
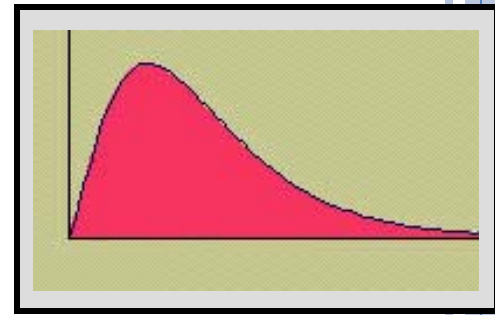
HYPOTHESIS TEST

1. $H_0 : \sigma^2 = \sigma_0^2$ v.s. $H_1 : \sigma^2 \neq (> , <)\sigma_0^2$.

2. Set α.

3. Test statistic and its sampling distribution

$$\chi^2_{STAT} = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{(n-1)}.$$

4. Calculate the realized statistic, $\chi^{2*} = \frac{(n-1)s^2}{\sigma_0^2}$.

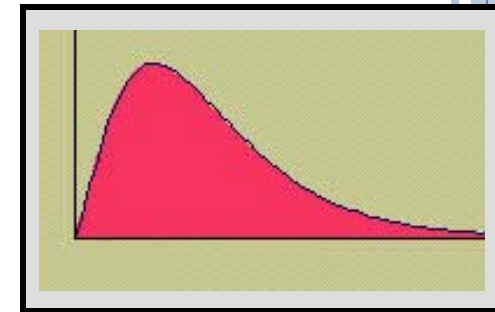5. Find the rejection region or p-value.

6. Conclude.

Inference Concerning a Population Variance

- Sometimes the primary parameter of interest is not the population mean μ but rather the population variance σ^2. We choose a random sample of size n from a normal distribution.

- The sample variance s^2 can be used in its standardized form:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

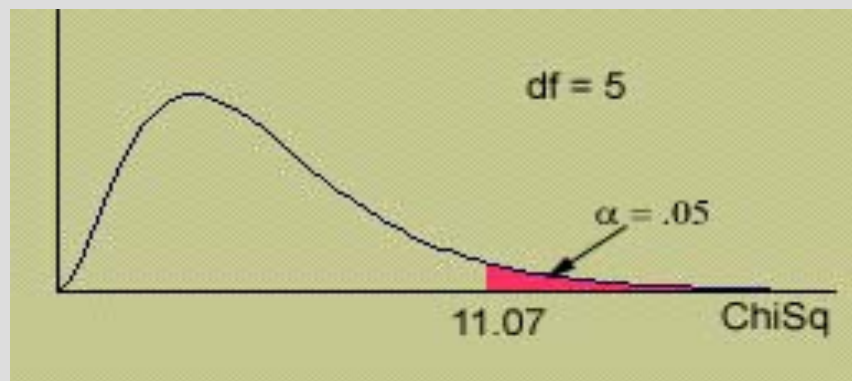- which has a Chi-Square distribution with $n - 1$ degrees of freedom.

INFERENCE CONCERNING A POPULATION VARIANCE

- Table 5 gives both upper and lower critical values of the chi-square statistic for a given *df*.

TABLE 5 (continued)	$X^2_{.100}$	$X^2_{.050}$	$X^2_{.025}$	$X^2_{.010}$	$X^2_{.005}$	df
	2.70554	3.84146	5.02389	6.63490	7.87944	1
	4.60517	5.99147	7.37776	9.21034	10.5966	2
	6.25139	7.81473	9.34840	11.3449	12.8381	3
	7.77944	9.48773	11.1433	13.2767	14.8602	4
	9.23635	11.0705	12.8325	15.0863	16.7496	5
	10.6446	12.5916	14.4494	16.8119	18.5476	6
	12.0170	14.0671	16.0128	18.4752	20.2777	7

For example, the value of chi-square that cuts off .05 in the upper tail of the distribution with *df* = 5 is χ^2 =11.07.

EXAMPLE

• A cement manufacturer claims that his cement has a compressive strength with a standard deviation of 10 kg/cm² or less. A sample of n = 10 measurements produced a mean and standard deviation of 312 and 13.96, respectively.

A test of hypothesis:

H_0: σ^2 = 10 (claim is correct)

H_a: σ^2 > 10 (claim is wrong)
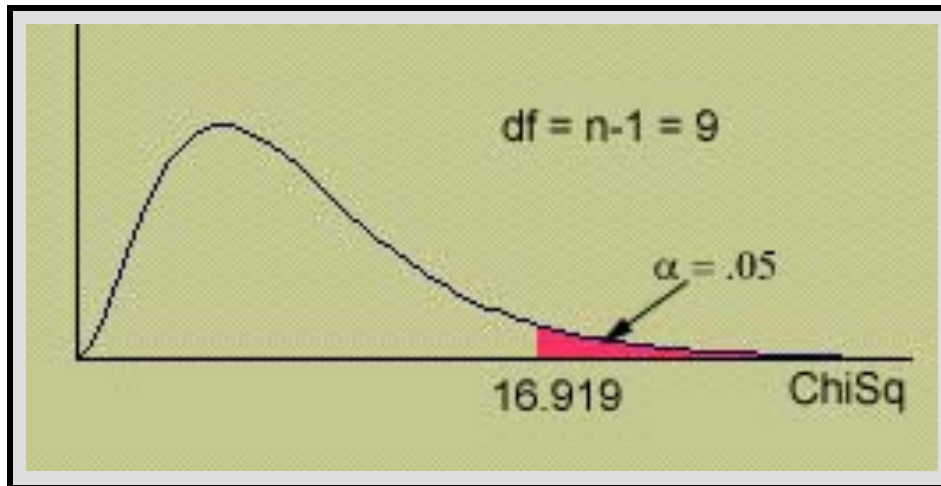
uses the test statistic:

$$\chi^2 = \frac{(n-1)s^2}{10^2} = \frac{9(13.96^2)}{100} = 17.5$$

EXAMPLE

- Do these data produce sufficient evidence to reject the manufacturer's claim? Use α = .05.

df = n-1 = 9
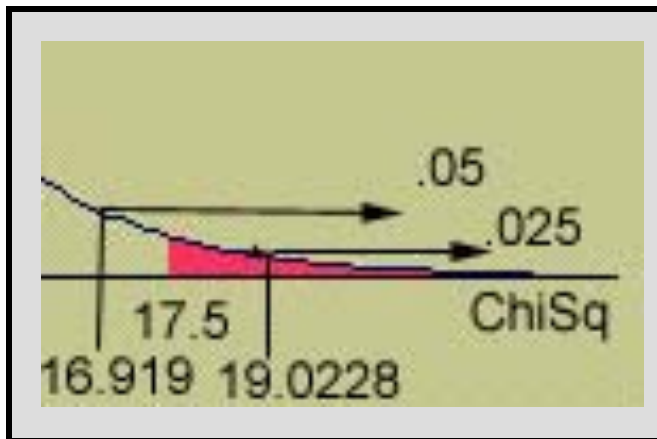
α = .05

16.919   ChiSq

Rejection region:
Reject H_0 if $\chi^2 > 16.919$ (α = .05).

Conclusion: Since χ^2 = 17.5, H_0 is rejected. The standard deviation of the cement strengths is more than 10.

APPROXIMATING THE *P*-VALUE

$$p\text{-value}: P(\chi^2 > 17.5) \text{ with } df = n - 1 = 9$$

.025 < *p*-value < .05

Since the *p*-value is less than $\alpha = .05$, H_0 is rejected. There is sufficient evidence to reject the manufacturer's claim.

$X^2_{.100}$	$X^2_{.050}$	$X^2_{.025}$	$X^2_{.010}$	$X^2_{.005}$	df
2.70554	3.84146	5.02389	6.63490	7.87944	1
4.60517	5.99147	7.37776	9.21034	10.5966	2
6.25139	7.81473	9.34840	11.3449	12.8381	3
7.77944	9.48773	11.1433	13.2767	14.8602	4
9.23635	11.0705	12.8325	15.0863	16.7496	5
10.6446	12.5916	14.4494	16.8119	18.5476	6
12.0170	14.0671	16.0128	18.4753	20.2777	7
13.3616	15.5073	17.5346	20.0902	21.9550	8
14.6837	16.9190	19.0228	21.6660	23.5893	9

STAT 10.6

Comparing two population variances (Skipped)