CENTRE DE RECHERCHE PUBLIC
crp SANTÉ
Microarray Center

APPLIED STATISTICS

Lecture 8

Inferences about Population Variance and Test of Goodness of Fit

Petr Nazarov
petr.nazarov@crp-sante.lu

25-11-2009

- **Interval estimation for population variance**
  - variance sampling distribution, χ^2 statistics
  - calculation of interval estimation
  - hypothesis tests for a population variance
- **Comparison of variances of two populations**
  - *F*-statistics
  - formulation of hypotheses and testing
- **Test of goodness of fit and independence**
  - goodness of fit for multinomial population
  - test for independence
  - goodness of fit for continuous distributions

INTERVAL ESTIMATION FOR VARIANCE

Variance Sampling Distribution

Variance
A measure of variability based on the squared deviations of the data values about the mean.

population $$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

sample $$s^2 = \frac{\sum (x_i - m)^2}{n-1}$$

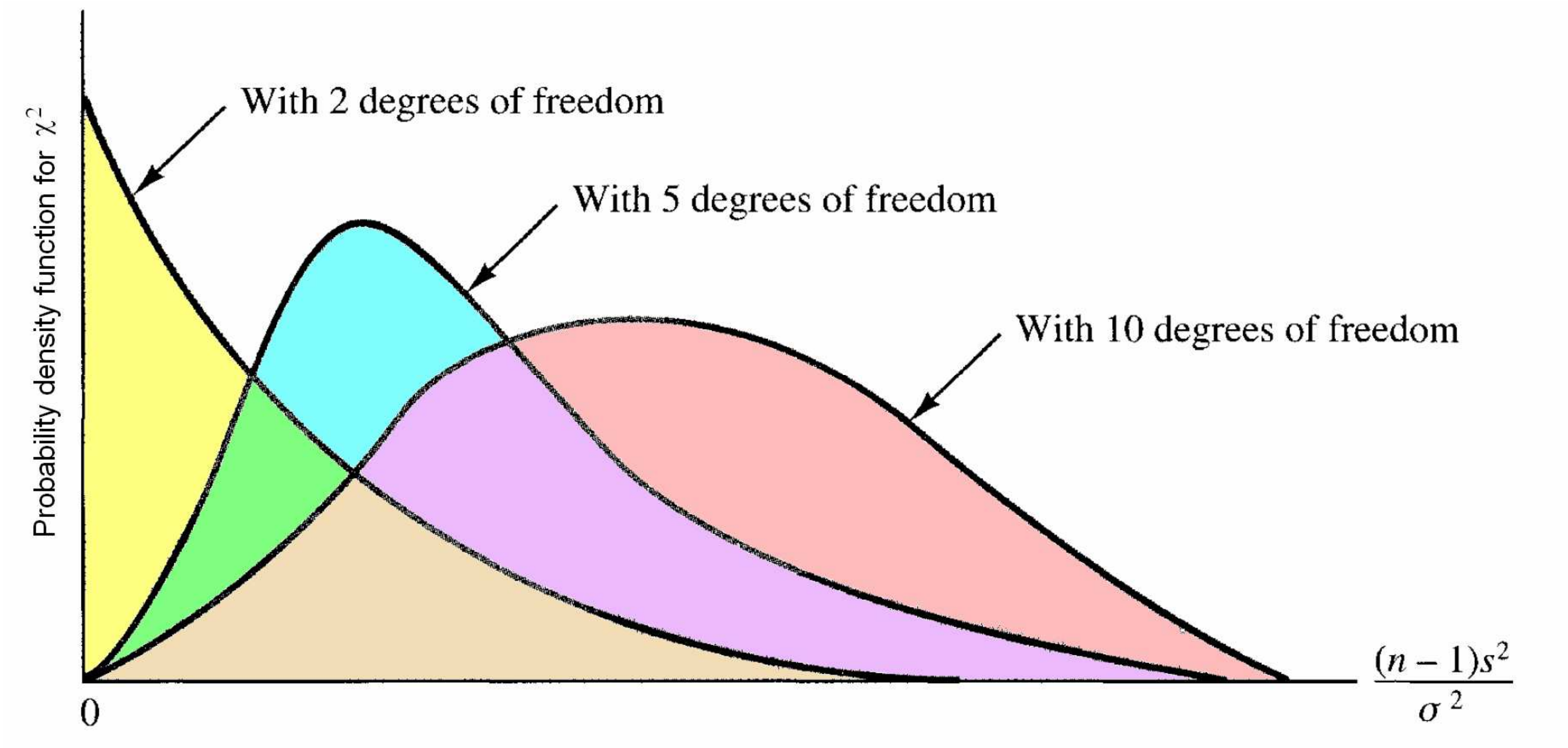
The interval estimation for variance is build using the following measure:

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

Sampling distribution of (n-1)s^2/σ^2
Whenever a simple random sample of size n is selected from a normal population, the sampling distribution of **(n-1)s^2/σ^2** has a **chi-square distribution** (χ^2) with n-1 degrees of freedom.

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

χ^2 Distribution

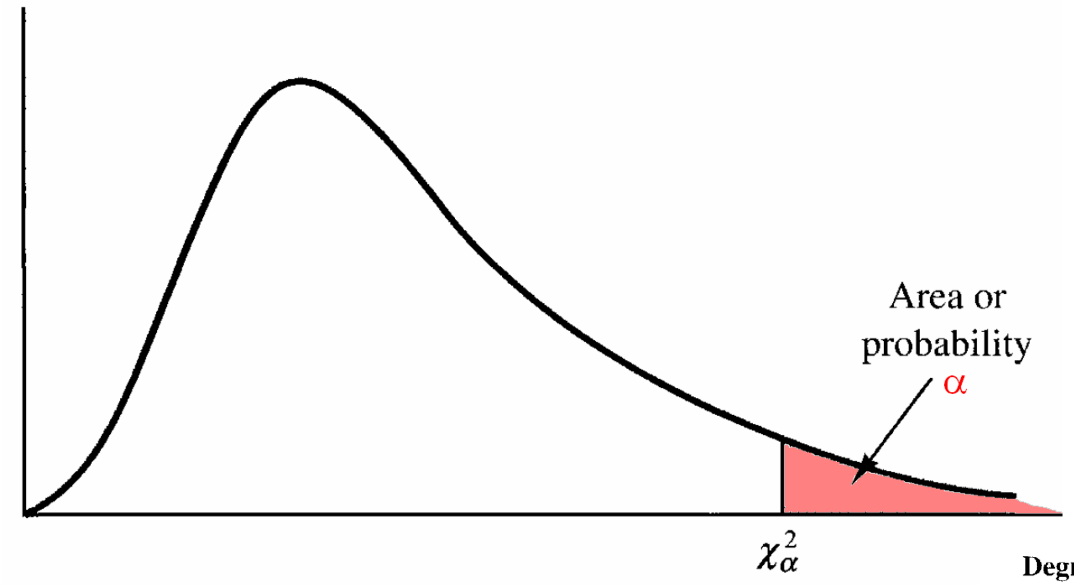


χ^2 distribution works only for sampling from normal population

$$\chi^2_{df=k} = \sum_{i=1}^{k} x_i^2 \quad where\ x_i - normal$$

INTERVAL ESTIMATION FOR VARIANCE

χ^2 Probabilities in Table and Excel



Degrees of Freedom	Area in Upper Tail .99	.975	.95	.90	.10	.05	.025	.01
1	.000	.001	.004	.016	2.706	3.841	5.024	6.635
2	.020	.051	.103	.211	4.605	5.991	7.378	9.210
3	.115	.216	.352	.584	6.251	7.815	9.348	11.345
4	.297	.484	.711	1.064	7.779	9.488	11.143	13.277
5	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086
6	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191

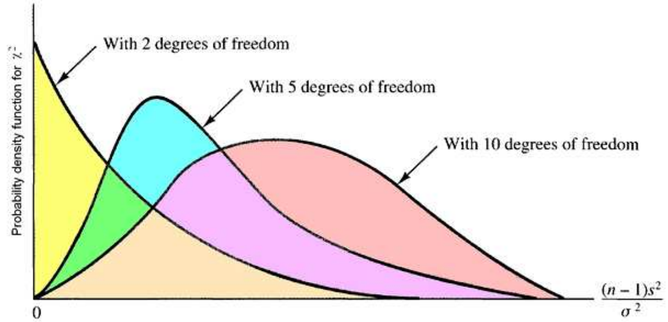
In Excel use functions:

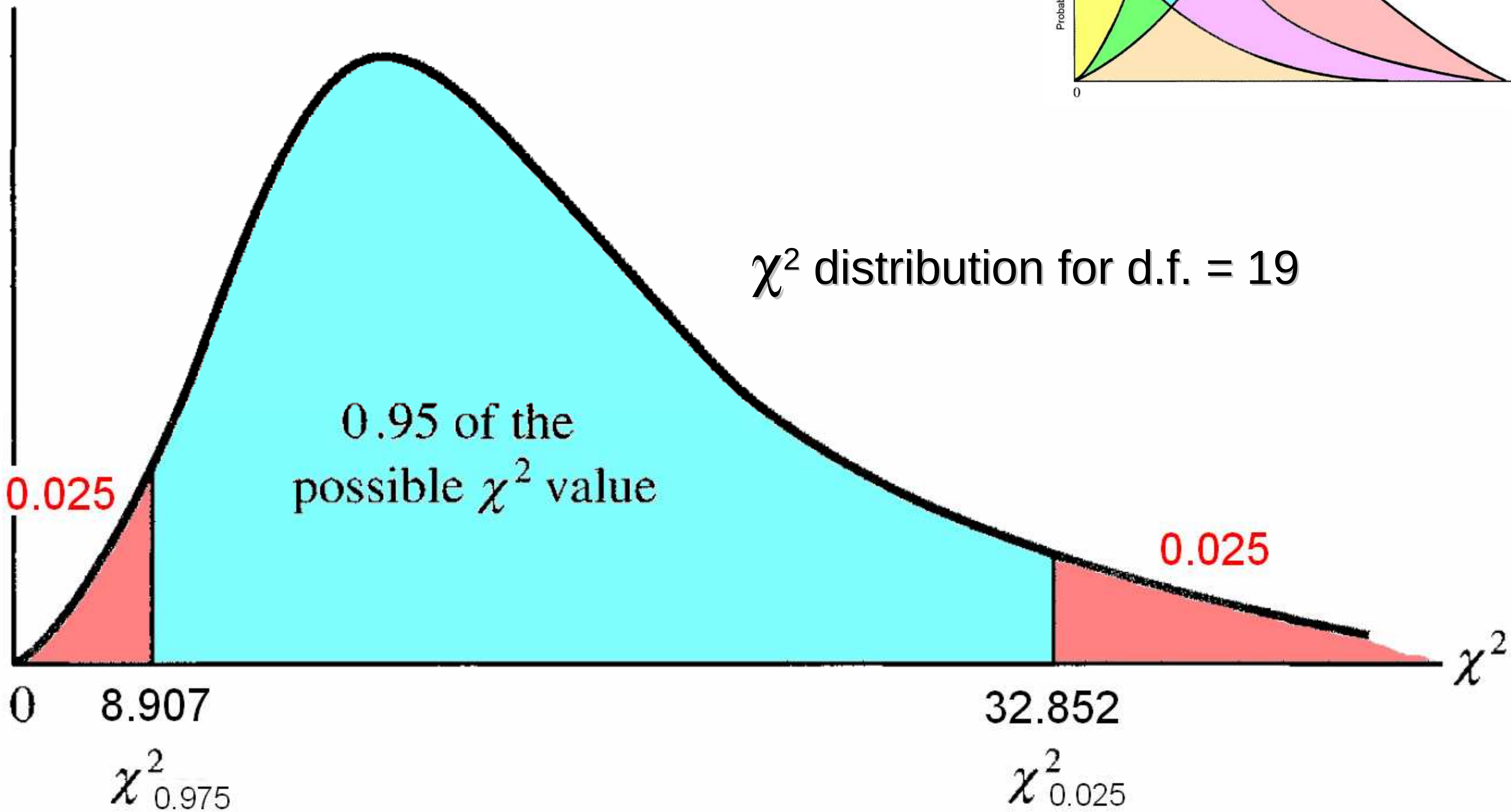
- **= CHIDIST(χ^2, n-1)**
- **= CHIINV(α, n-1)**

INTERVAL ESTIMATION FOR VARIANCE

χ^2 Distribution for Interval Estimation

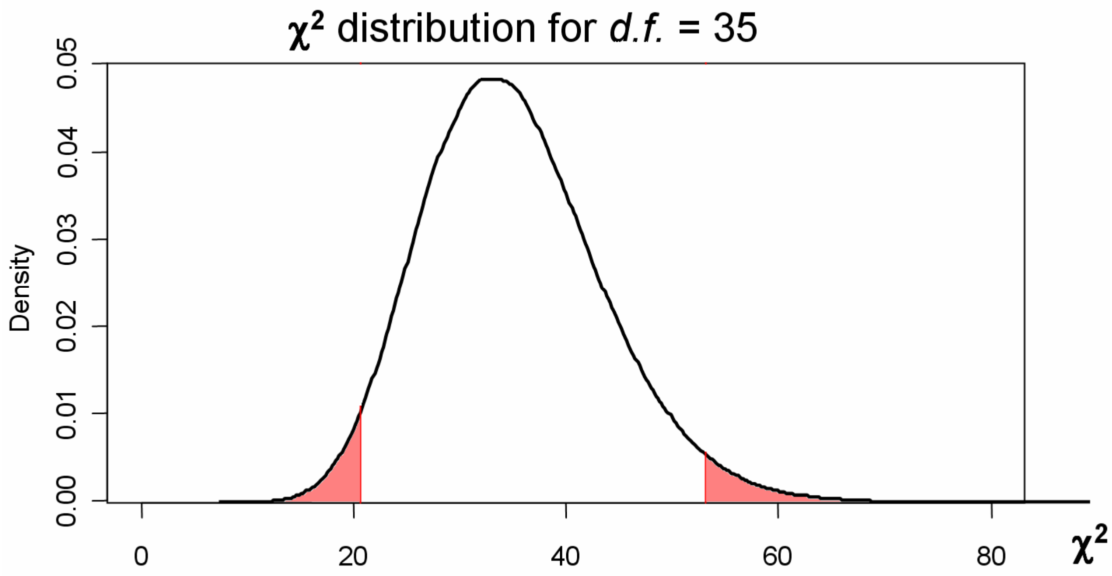
$$\chi^2 = (n-1)\frac{s^2}{\sigma^2}$$





$$\chi^2_{1-\alpha/2} \le (n-1)\frac{s^2}{\sigma^2} \le \chi^2_{\alpha/2}$$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$



Suppose sample of n = 36 coffee cans is selected and m = 2.92 and s = 0.18 lbm is observed. Provide 95% confidence interval for the standard deviation

$$\frac{(36-1)0.18^2}{53.203} \le \sigma^2 \le \frac{(36-1)0.18^2}{20.569}$$

◆ = CHIINV(α, n-1)

$$0.0213 \le \sigma^2 \le 0.0551$$

$$0.146 \le \sigma \le 0.235$$

INTERVAL ESTIMATION FOR VARIANCE

Hypotheses about Population Variance

H_0: $\sigma^2 \leq$ const
H_a: $\sigma^2 >$ const

H_0: $\sigma^2 \geq$ const
H_a: $\sigma^2 <$ const

H_0: $\sigma^2 =$ const
H_a: $\sigma^2 \neq$ const

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0 : \sigma^2 \geq \sigma_0^2$ $H_a : \sigma^2 < \sigma_0^2$	$H_0 : \sigma^2 \leq \sigma_0^2$ $H_a : \sigma^2 > \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_a : \sigma^2 \neq \sigma_0^2$
Test Statistic	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha)}$	Reject H_0 if $\chi^2 \geq \chi^2_{\alpha}$	Reject H_0 if $\chi^2 \leq \chi^2_{(1-\alpha/2)}$ or if $\chi^2 \geq \chi^2_{\alpha/2}$

In many statistical applications we need a comparison between variances of two populations. In fact well-known ANOVA-method is base on this comparison.

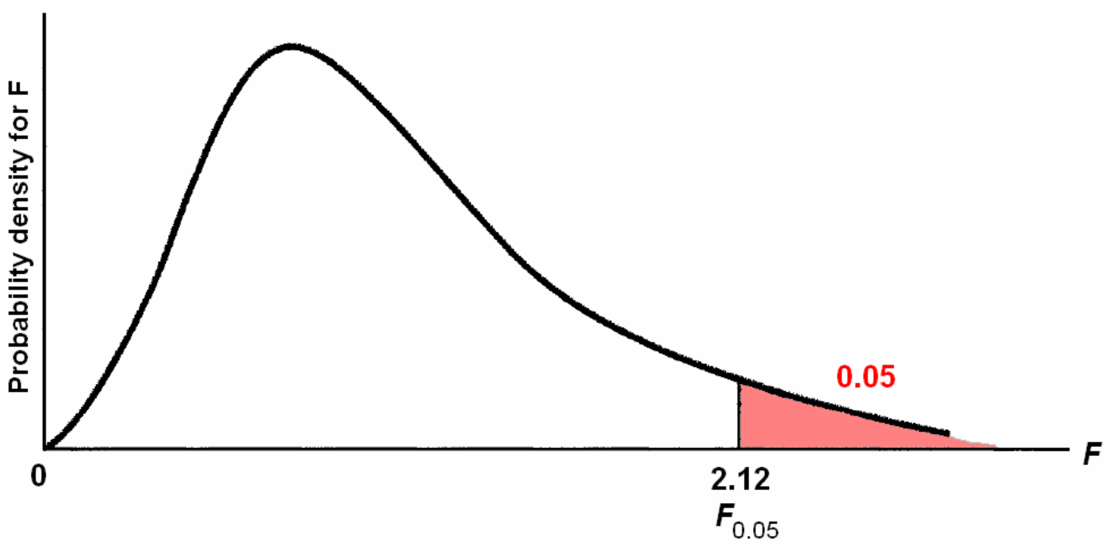
The statistics is build for the following measure:

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

Sampling distribution of s_1^2/s_2^2 when $\sigma_1^2 = \sigma_2^2$

Whenever a independent simple random samples of size n_1 and n_2 are selected from two normal populations with equal variances, the sampling of s_1^2/s_2^2 has **F-distribution** with n_1-1 degree of freedom for numerator and n_2-1 for denominator.

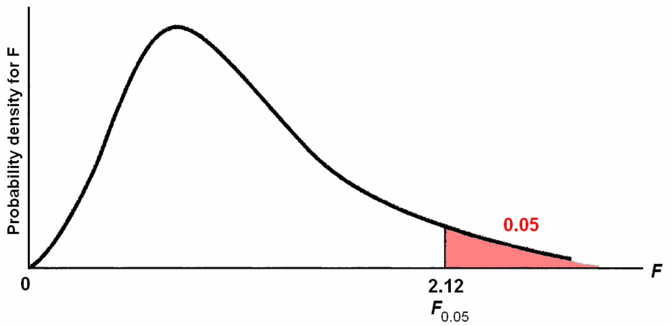
F-distribution for 20 d.f. in numerator and 20 d.f. in denominator



In Excel use functions:

- `= FDIST(F,n1-1,n2-1)` — $= \text{FDIST}(F, n_1-1, n_2-1)$
- $= \text{FINV}(\alpha, n_1-1, n_2-1)$
- = FTEST(data1,data2)

Hypotheses about Variances of Two Populations



$H_0: \sigma_1^2 \le \sigma_2^2$

$H_a: \sigma_1^2 > \sigma_2^2$

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

$H_a: \sigma_1^2 \ne \sigma_2^2$

	Upper Tail Test	**Two-Tailed Test**
Hypotheses	$H_0: \sigma_1^2 \le \sigma_2^2$ $H_a: \sigma_1^2 > \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \ne \sigma_2^2$ *Note: Population 1 has the lager sample variance*
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
Rejection Rule: p-Value Approach	Reject H_0 if p-value $\le \alpha$	Reject H_0 if p-value $\le \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $F \ge F_\alpha$	Reject H_0 if $F \ge F_\alpha$

VARIANCES OF TWO POPULATIONS

Example

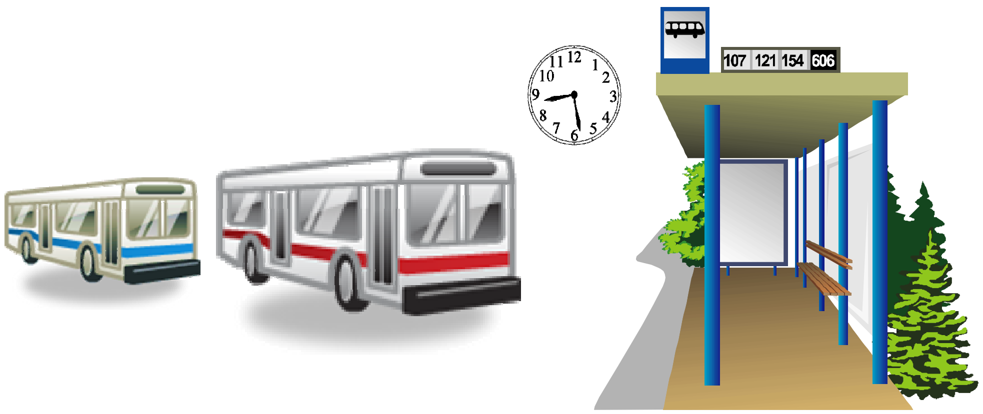
schoolbus.xls

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

Dullus County Schools is renewing its school bus service contract for the coming year and must select one of two bus companies, the Milbank Company or the Gulf Park Company. We will use the variance of the arrival or pickup/delivery times as a primary measure of the quality of the bus service. Low variance values indicate the more consistent and higher-quality service. If the variances of arrival times associated with the two services are equal Dullus School administrators will select the company offering the better financial terms However, if the sample data on bus arrival times for the two companies indicate a significant difference between the variances, the administrators may want to give special consideration to the company with the better or lower variance service. The appropriate hypotheses follow.

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

If H_0 can be rejected, the conclusion of unequal service quality is appropriate. We will use a level of significance of $\alpha = .10$ to conduct the hypothesis test.



Example

schoolbus.xls

#	Milbank	Gulf Park
1	35.9	21.6
2	29.9	20.5
3	31.2	23.3
4	16.2	18.8
5	19.0	17.2
6	15.9	7.7
7	18.8	18.6
8	22.2	18.7
9	19.9	20.4
10	16.4	22.4
11	5.0	23.1
12	25.4	19.8
13	14.7	26.0
14	22.7	17.1
15	18.0	27.9
16	28.1	20.8
17	12.1	
18	21.4	
19	13.4	
20	22.9	
21	21.0	
22	10.1	
23	23.0	
24	19.4	
25	15.2	
26	28.2	

1. Let us start from estimation of the variances for 2 data sets

Milbank: $s_1^2 = 48$

Gulf Park: $s_2^2 = 20$

interval estimation (optionally)

Milbank: $\sigma_1^2 \approx 48 \ (29.5 \div 91.5)$

Gulf Park: $\sigma_2^2 \approx 20 \ (10.9 \div 47.9)$

2. Let us calculate the *F*-statistics

$$F = \frac{s_1^2}{s_2^2} = \frac{48}{20} = 2.40$$

3. … and p-value = 0.08

p-value = 0.08 < α = 0.1

In Excel use one of the functions:

- `= 2*FDIST(F,n₁-1,n₂-1)`
- `= FTEST(data1,data2)`

Part II

Goodness of Fit and Independence

Multinomial population
A population in which each element is assigned to one and only one of several categories. The multinomial distribution extends the binomial distribution from two to three or more outcomes.

Contingency table = Crosstabulation
Contingency tables or crosstabulations are used to record, summarize and analyze the relationship between two or more categorical (usually) variables.

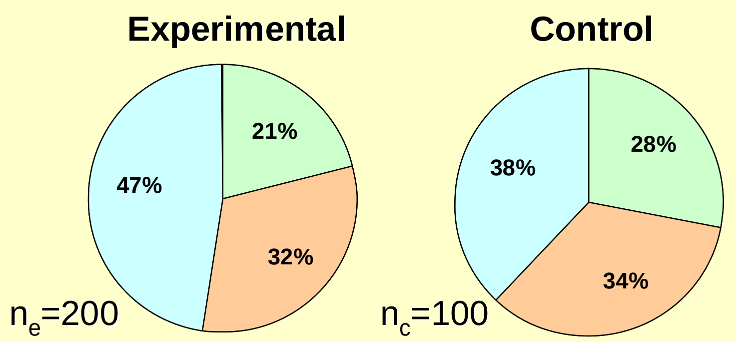
The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **A** – patient is **completely treated**
- **B** – disease transforms into a **chronic form**
- **C** – treatment is **unsuccessful** ☹

In parallel the 100 patients treated with standard methods are observed

Category	Experimental	Control
A	42	28
B	64	34
C	94	38
Sum	200	100

◆ The proportions for 3 “classes” of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

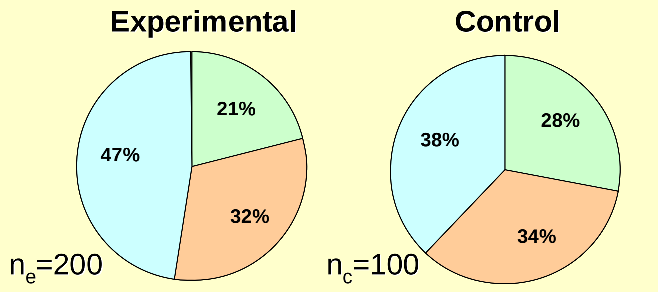
Goodness of fit test
A statistical test conducted to determine whether to reject a hypothesized probability distribution for a population.

Model – our assumption concerning the distribution, which we would like to test.

Observed frequency – frequency distribution for experimentally observed data, f_i

Expected frequency – frequency distribution, which we would expect from our **model**, e_i

◆ The proportions for 3 "classes" of patients with and without treatment are:



Are the proportions ***significantly different*** in control and experimental groups?

Hypotheses for the test:

H_0: the population follows a multinomial distribution with the probabilities, specified by **model**

H_a: the population does not follow … **model**

Test statistics for goodness of fit

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 has $k-1$ degree of freedom

At least 5 expected must be in each category!

TEST OF GOODNESS OF FIT

Example

The new treatment for a disease is tested on 200 patients. The outcomes are classified as:

- **A** – patient is **completely treated**
- **B** – disease transforms into a **chronic form**
- **C** – treatment is **unsuccessful** ☹

In parallel the 100 patients treated with standard methods are observed

Category	Experimental	Control
A	42	28
B	64	34
C	94	38
Sum	200	100

1. Select the model and calculate expected frequencies

Let's use control group (classical treatment) as a model, then:

2. Compare expected frequencies with the experimental ones and build χ^2

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Category	Control frequencies	Model for ctrl.	Expected freq., e	Experimental freq., f
A	28	0.28	56	42
B	34	0.34	68	64
C	38	0.38	76	94
Sum	**100**	**1**	**200**	**200**

Category	(f-e)2/e
A	3.500
B	0.235
C	4.263
Chi2	**7.998**
p-value	**0.01833**

3. Calculate p-value for χ^2 with d.f. = k–1

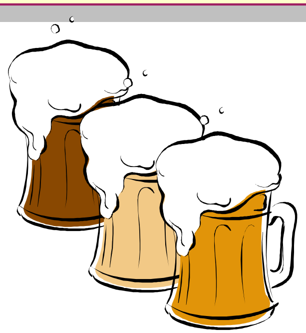
- `= CHIDIST(`χ^2`,d.f.)`
- `= CHITEST(`f, e`)`

p-value = 0.018, reject H_0

Goodness of Fit for Independence Test: Example

Alber's Brewery manufactures and distributes three types of beer: **white**, **regular**, and **dark**. In an analysis of the market segments for the three beers, the firm's market research group raised the question of whether preferences for the three beers differ among **male** and **female** beer drinkers. If beer preference is independent of the gender of the beer drinker, one advertising campaign will be initiated for all of Alber's beers. However, if beer preference depends on the gender of the beer drinker, the firm will tailor its promotions to different target markets.

beer.xls



H_0: Beer preference is **independent** of the gender of the beer drinker

H_a: Beer preference is **not independent** of the gender of the beer drinker

sex\beer	White	Regular	Dark	Total
Male	20	40	20	**80**
Female	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

Goodness of Fit for Independence Test: Example

1. Build model assuming independence

sex\beer	White	Regular	Dark	Total
Male	20	40	20	**80**
Female	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

	White	Regular	Dark	Total
Model	0.3333	0.4667	0.2000	1

2. Transfer the model into expected frequencies, multiplying model value by number in group

sex\beer	White	Regular	Dark	Total
Male	26.67	37.33	16.00	**80**
Female	23.33	32.67	14.00	**70**
Total	**50**	**70**	**30**	**150**

$$e_{ij} = \frac{(Row\ i\ Total)(Column\ \ j\ Total)}{Sample\ Size}$$

3. Build χ^2 statistics

$$\chi^2 = \sum_i^n \sum_j^m \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}}$$

$\chi^2 = 6.122$

χ^2 distribution with d.f.=$(n-1)(m-1)$, provided that the expected frequencies are 5 or more for all categories.

4. Calculate p-value

p-value = 0.047, reject H_0

Test for Normality: Example

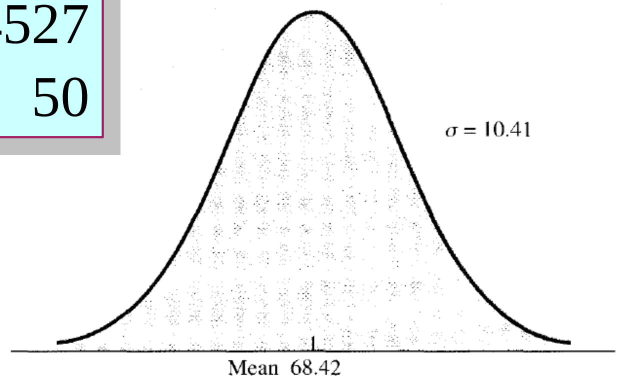
Chemline hires approximately 400 new employees annually for its four plants. The personnel director asks whether a normal distribution applies for the population of aptitude test scores. If such a distribution can be used, the distribution would be helpful in evaluating specific test scores; that is, scores in the upper 20%, lower 40%, and so on, could be identified quickly. Hence, we want to test the null hypothesis that the population of test scores has a normal distribution. The study will be based on 50 results.

chemline.xls

Aptitude test scores

71	86	56	61	65
60	63	76	69	56
55	79	56	74	93
82	80	90	80	73
85	62	64	54	54
65	54	63	73	58
77	56	65	76	64
61	84	70	53	79
79	61	62	61	65
66	70	68	76	71

Mean	68.42
Standard Deviation	10.4141
Sample Variance	108.4527
Count	50



H_0: The population of test scores **has a normal distribution** with mean **68.42** and standard deviation **10.41**

H_a: the population **does not have** a mentioned distribution

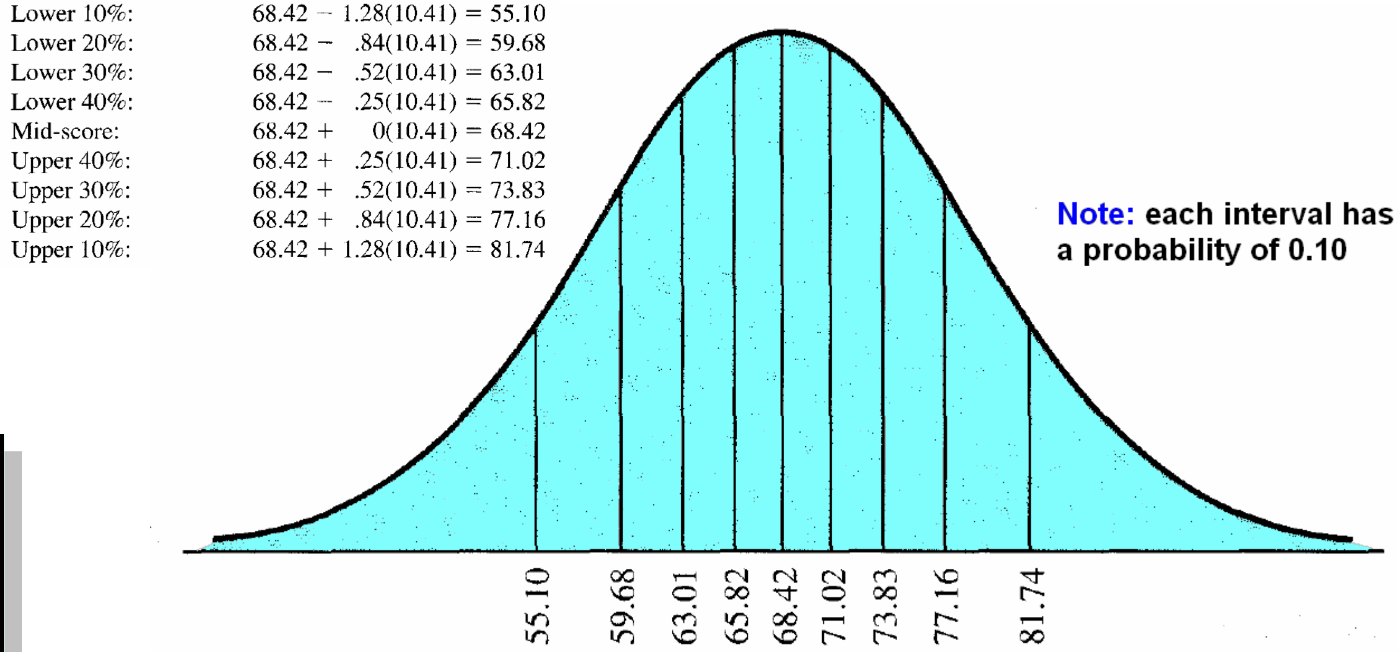
TEST FOR CONTINUOUS DISTRIBUTIONS

Test for Normality: Example

chemline.xls

Mean	68.42
Standard Deviation	10.4141
Sample Variance	108.4527
Count	50

Bin	*Observed frequency*	*Expected frequency*
55.1	5	5
59.68	5	5
63.01	9	5
65.82	6	5
68.42	2	5
71.02	5	5
73.83	2	5
77.16	5	5
81.74	5	5
More	6	5
Total	**50**	**50**



$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

χ^2 distribution with d.f.= $n - p - 1$,

where p – number of estimated parameters

$p = 2$ includes mean and variance

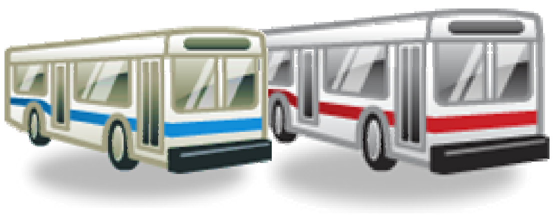
d.f. = 10 – 2 – 1

$\chi^2 = 7.2$

p-value = 0.41, cannot reject H_0

QUESTIONS ?

Thank you for your χ attention



to be continued…