Chapter 6: The χ^2 and *F* distributions

- The χ^2 distribution is used to:
  - construct confidence interval estimates of a variance
  - compare a set of actual frequencies with expected frequencies
  - test for association between variables in a contingency table

The χ^2 and *F* distributions (continued)

- The *F* distribution is used to
  - test the hypothesis of equality of two variances
  - conduct an analysis of variance (ANOVA), comparing means of several samples

Case 1: Estimating a variance

- A random sample of size $n = 20$ yields a standard deviation of $s = 25$. How do we estimate the population variance?

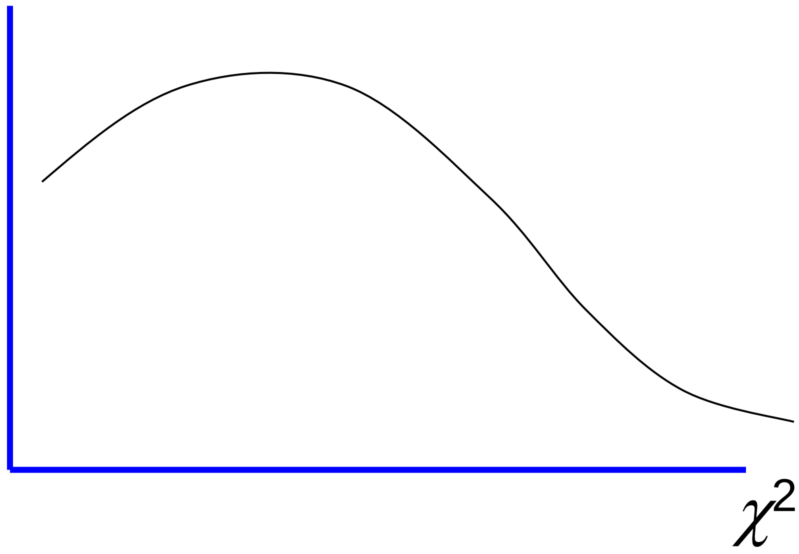
- Point estimate: use $s^2 = 25^2 = 625$ which is unbiased ($E(s^2) = \sigma^2$)

- Interval estimate: we need the sampling distribution of s^2...

The sampling distribution of s^2

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

- n-1 gives the degrees of freedom for the χ^2 distribution, 19 in this example.



Limits to the confidence interval

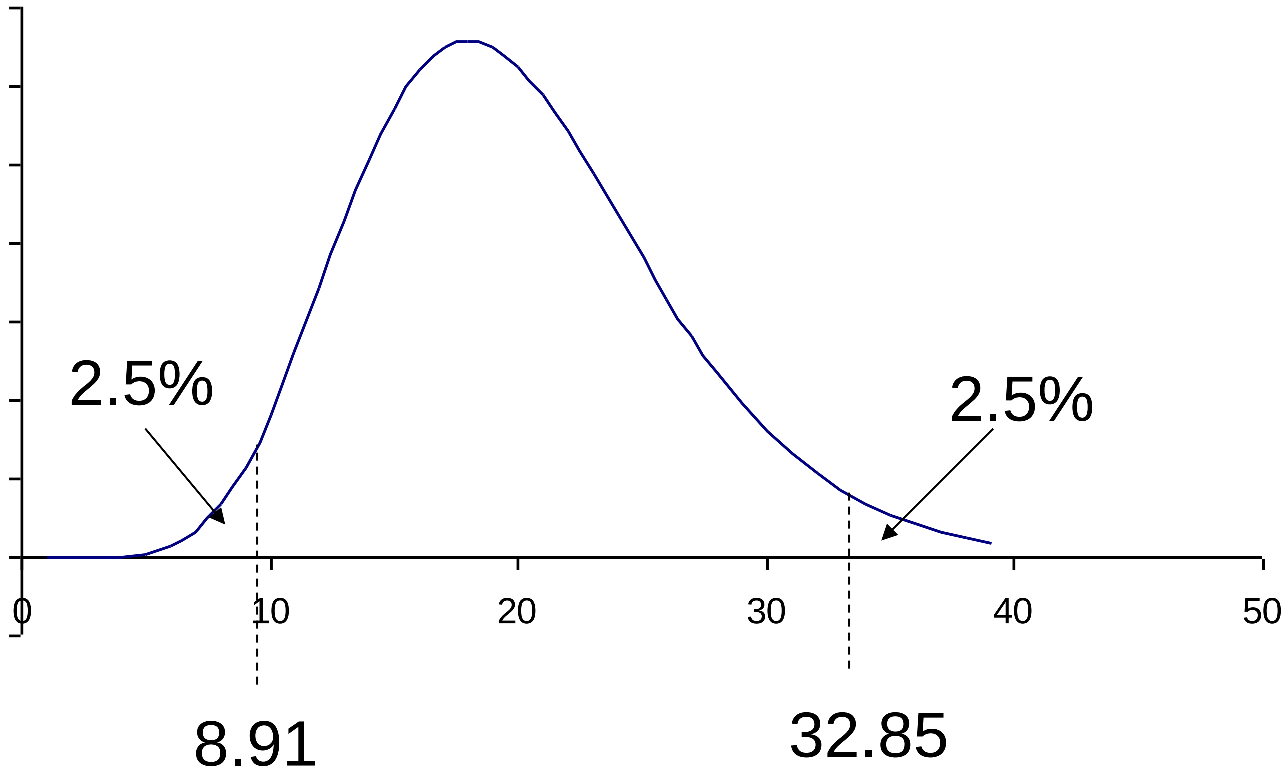
- For the 95% CI, we need the χ^2 values cutting off 2.5% in each tail of the distribution

Excerpt from Table A4:

ν	**0.990**	**0.975**	**…**	**0.050**	**0.025**	**0.010**
1	0.000	0.001	…	3.841	5.024	6.635
2	0.020	0.051	…	5.991	7.378	9.210
3	0.115	0.216	…	7.815	9.348	11.345
:	:	:	…	:	:	:
18	7.015	8.231	…	28.869	31.526	34.805
19	7.633	8.907	…	30.144	32.852	36.191
20	8.260	9.591	…	31.410	34.170	37.566

Slide 6.6

Tails of the χ^2_{19} distribution



Tails of the χ^2_{19} distribution (continued)

- We can be 95% confident that $(n\text{-}1)s^2/\sigma^2$ lies between 8.91 and 32.85 (for n = 20)

$$8.91 \le \frac{(n-1)s^2}{\sigma^2} \le 32.85$$

- Rearranging:

$$\frac{(n-1)s^2}{32.85} \le \sigma^2 \le \frac{(n-1)s^2}{8.91}$$

- Substituting s^2 = 625 and n = 20:

$$361.5 \le \sigma^2 \le 1{,}332.8$$

- gives the 95% CI estimate

Slide 6.8

Case 2: Comparing actual versus expected frequencies

- 72 rolls of a die yield:

Score on die	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	6	15	15	7	15	14

- From a fair die one would expect each number to come up 12 times.
- Is this evidence of a biased die?

The test statistic

- H_0: the die is fair
  H_1: the die is biased

- This can be tested using

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- which has a χ^2 distribution with k-1 degrees of freedom, k = 6 in this case.

Calculating the test statistic

Score	Observed frequency (O)	Expected frequency (E)	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
1	6	12	–6	36	3.00
2	15	12	3	9	0.75
3	15	12	3	9	0.75
4	7	12	–5	25	2.08
5	15	12	3	9	0.75
6	14	12	2	4	0.33
Totals	72	72	0		7.66

Calculating the test statistic (continued)

- The test statistic, 7.66, is less than the critical value of χ^2 with $\nu = 5$, 11.1
- Hence the null is not rejected, the variation is random
- Note the critical value cuts off 5% (not 2.5%) in the upper tail of the distribution. Only large values of the test statistic reject H_0

Case 3: Contingency tables

- The association between two variables can be analysed via the χ^2 distribution
  - Voting behaviour based on a sample of 200:

Social class	Labour	Conservative	Liberal Democrat	Total
A	10	15	15	40
B	40	35	25	100
C	30	20	10	60
Total	80	70	50	200

Slide 6.13

Are social class and voting behaviour related?

- H_0: no association between social class and voting behaviour
  H_1: some association

- Expected values are calculated, based on the null of no association

- E.g. if there is no association, 40% (80/200) of every social class should vote Labour, i.e. 16 from class A, 40 from B and 24 from C

Observed and (expected) values

Social class	Labour	Conservative	Liberal Democrat	Total
A	10(16)	15(14)	15(10)	40
B	40(40)	35(35)	25(25)	100
C	30(24)	20(21)	10(15)	60
Total	80	70	50	200

Calculating the test statistic

$$\frac{(10-16)^2}{16}+\frac{(15-14)^2}{14}+\frac{(15-10)^2}{10}+$$
$$\frac{(40-40)^2}{40}+\frac{(35-35)^2}{35}+\frac{(25-25)^2}{25}+$$
$$\frac{(30-24)^2}{24}+\frac{(20-21)^2}{21}+\frac{(10-15)^2}{15}=8.04$$

For ν = (*rows*-1) × (*columns*-1) = 4, the critical value of the χ^2 distribution is 9.50, so the null of no association is not rejected at the 5% significance level.

Testing two variances - the *F* distribution

- Do two samples have equal variances (i.e. come from populations with the same variance)?

- Data:

  $n_1 = 30 \qquad s_1 = 25$

  $n_2 = 30 \qquad s_2 = 20$

Slide 6.17

Testing two variances - the *F* distribution (continued)

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

  H_1: $\sigma_1^2 = \sigma_2^2$

  or, equivalently

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

  H_1: $\sigma_1^2/\sigma_2^2 \neq 1$

The test statistic

- The test statistic is

$$\frac{s_1^2}{s_2^2} \sim F_{n_1 - 1, n_2 - 1}$$

- Evaluating this: $F = \frac{25^2}{20^2} = 1.5625$
- $F^*_{29,29}$ = 2.09 >1.5625, so the null is not rejected. The variances may be considered equal.

Excerpt from Table A5(b): the *F* distribution

ν_1	**1**	**2**	**…**	**24**	**30**	**40**
ν_2						
1	647.79	799.48	…	997.27	1001.40	1005.60
2	38.51	39.00	…	39.46	39.46	39.47
:	:	:	…	:	:	:
28	5.61	4.22	…	2.17	2.11	2.05
29	5.59	4.20	…	2.15	2.09	2.03
30	5.57	4.18	…	2.14	2.07	2.01

(Using ν_1 = 30 (rather than 29) makes little practical difference.)

One or two tailed test?

- As long as the larger variance is made the numerator of the test statistic, only 'large' values of *F* reject the null.

- The smallest possible value of *F* is 1, which occurs if the sample variances are equal. H_0 should not be rejected in this case.

- So, despite the "≠" in H_1, this is a one tailed test.

Case 2: Analysis of variance (ANOVA)

- A test for the equality of several means, not just two as before.

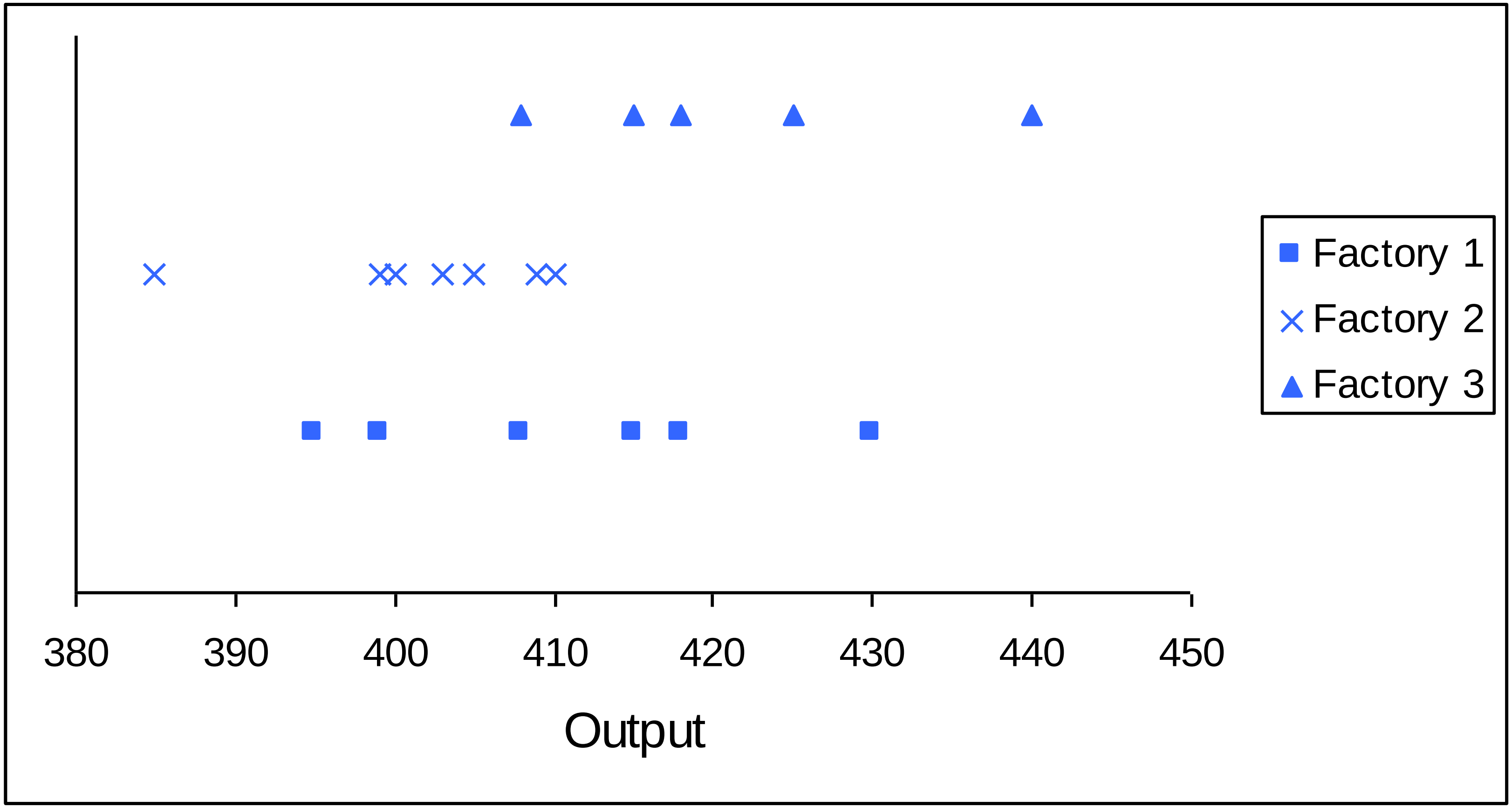
- In our example we test for the equality of output of three factories, i.e. are they equally productive, on average, or not?

Data - daily output of three factories

Observation	Factory 1	Factory 2	Factory 3
1	415	385	408
2	430	410	415
3	395	409	418
4	399	403	440
5	408	405	425
6	418	400	
7		399	

Slide 6.23

Chart of output



The hypothesis to test

- H_0: $\mu_1 = \mu_2 = \mu_3$

  H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

- Principle of the test: break down the total variance of all observations into the within factory variance and the between factory variance

- If the latter is large relative to the former, reject H_0

Sums of squares

- Rather than variances, work with sums of squares

Sum of squares

$$s^2 = \frac{\sum \left(x - \bar{x}\right)^2}{n-1}$$

Variance

Three sums of squares

- Total sum of squares (TSS)
  - Sum of squares of all deviations from the overall average
- Between sum of squares (BSS)
  - Sum of squares of deviations of factory means from overall average
- Within sum of squares (WSS)
  - Sum of squares of deviations within each factory, from factory average

Slide 6.27

Test statistic

$$F = \frac{BSS/(k-1)}{WSS/(n-k)}$$

- The *F* statistic is the ratio of BSS to WSS, each adjusted by their degrees of freedom (*k*-1 and *n*-*k*)
- Large values of *F* $\Rightarrow$ BSS large relative to WSS $\Rightarrow$ between factories deviations large $\Rightarrow$ reject H_0

The calculations

- TSS = $\sum_j \sum_i \left(x_{ij} - \bar{x}\right)^2$

  (*j* indexes factories, *i* indexes observations)

- = $(415 – 410.11)^2 + (430 – 410.11)^2 + \ldots + (440 – 410.11)^2 + (425 – 410.11)^2$ = 2,977.778

  (410.11 is the overall, or grand, average)

The calculations (continued)

- BSS = $\sum_j \sum_i \left(\bar{x}_i - \bar{x}\right)^2$

  where $\bar{x}_i$ is the average output of factory *i*

- $= 6 \times (410.83 - 410.11)^2 + 7 \times (401.57 - 410.11)^2 + 5 \times (421.2 - 410.11)^2 = 1{,}128.43$

- (410.83, 401.57, 421.11 are the three averages, respectively)

The calculations (continued)

- WSS = TSS – BSS = 2,977.778 – 1,128.430
  = 1,849.348

- Alternatively, WSS = $\sum_{j}\sum_{i}\left(x_{ij} - \bar{x}_i\right)^2$

  = $(415-410.83)^2 + \ldots + (418-410.83)^2 + (385-401.57)^2 + \ldots + (399-401.57)^2 + (408-421.2)^2 + \ldots + (425-421.2)^2$
  = 1,849.348

Result of the test

$$F = \frac{BSS/(k-1)}{WSS/(n-k)} = \frac{1128.43/(3-1)}{1849.348/(18-3)} = 4.576$$

- $F^*_{2,15}$ = 3.682 (5% significance level)
- $F > F^*$ hence we reject H_0. There are significant differences between the factories.

ANOVA table (Excel format)

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Factory 1	6	2465	410.833	166.967
Factory 2	7	2811	401.571	70.6191
Factory 3	5	2106	421.2	147.7

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1128.430	2	564.215	4.576	0.028	3.68
Within Groups	1849.348	15	123.290			
Total	2977.778	17				

Summary

- Use the χ^2 distribution to
  - Calculate the CI for a variance
  - Compare actual and expected values
  - Analyse a contingency table

- Use the *F* distribution to
  - Test for the equality of two variances
  - Test for the equality of several means