MATH273001

This question paper consists of 11 printed pages, each of which is identified by the reference **MATH2730**.

Statistical tables are provided at the end of this examination paper.

Only approved basic scientific calculators may be used.

Examination for the Module MATH2730
(January 2005)

ANALYSIS OF EXPERIMENTAL DATA

Time allowed: **2 hours**

Attempt not more than FOUR questions.
All questions carry equal marks.

Note that separate statistical tables are **not** provided.
Instead, the necessary statistical tables are included on pages 8 to 11 of this paper

1. Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad\qquad i = 1, \dots, t; j = 1, \dots, n,$$

where Y_{ij} is the jth response on treatment i, μ and α_i represent the overall mean and effect of the ith treatment respectively, and ε_{ij} represents random variation with each ε_{ij} being independently $N(0, \sigma^2)$ distributed. Let $N = nt$ be the total number of observations and assume that $\sum_{i=1}^{t} \alpha_i = 0$.

(a) Denote the mean of the observations in group i by $\overline{Y}_{i\bullet} = n^{-1} \sum_{j=1}^{n} Y_{ij}$ and the mean of all the observations by $\overline{Y}_{\bullet\bullet} = N^{-1} \sum_{i=1}^{t} \sum_{j=1}^{n} Y_{ij}$. Write down the distributions of Y_{ij}, $\overline{Y}_{i\bullet}$, and $\overline{Y}_{\bullet\bullet}$.

(b) Let the error sum of squares SS_E and the mean square error MS_E be defined by

$$SS_E = \sum_{i=1}^{t} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_{i\bullet})^2 \qquad \text{and} \qquad MS_E = \frac{1}{N-t} SS_E.$$

Show that $E(MS_E) = \sigma^2$ and write down, without proof, the distribution of SS_E/σ^2.

(c) We wish to test the null hypothesis H_0: $\alpha_i = 0$ for $i = 1, \dots, t$, against the alternative H_1: at least one $\alpha_i \neq 0$. Let the sum of squares and mean square for treatments be

$$SS_T = \sum_{i=1}^{t} n(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 \qquad \text{and} \qquad MS_T = \frac{1}{t-1} SS_T.$$

Given your answer to part (b) and the fact that, under H_0, $SS_T/\sigma^2 \sim \chi^2_{t-1}$, derive the distribution of the test statistic $F = MS_T/MS_E$.

(d) In an experiment to determine whether carbon tetrachloride is an effective anti-worm drug, 20 rats were infested with worm larvae. Eight days later, five rats were treated with each of three different doses of carbon tetrachloride (low, medium, or high doses), while five rats were left untreated. After two more days, the rats were killed and the number of worms in each was counted. The resulting data are given below.

Carbon tetrachloride data

	No carbon tetrachloride	Carbon tetrachloride dose		
		Low	Medium	High
	279	378	172	381
	338	275	335	346
	334	412	335	340
	198	265	282	471
	303	286	250	318
Totals	1452	1616	1374	1856

Given that $\sum_{i=1}^{t} \sum_{j=1}^{n} y_{ij}^2 = 2074428$ and $y_{\bullet\bullet} = 6298$, construct an ANOVA table for the carbon tetrachloride data. Is there significant evidence that carbon tetrachloride is an effective anti-worming agent?

2. Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad\qquad i = 1, \ldots, t; j = 1, \ldots, n,$$

where Y_{ij} is the jth response on treatment i, μ and α_i represent the overall mean and effect of the ith treatment respectively, and the independent $\varepsilon_{ij} \sim N(0, \sigma^2)$ represent random variation. Conventionally, the treatment effects α_i are constrained so that $\sum_i \alpha_i = 0$. Let $N = nt$ be the total number of observations.

(a) Show that the least-squares estimate of μ is given by

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{t} \sum_{j=1}^{n} Y_{ij}$$

and find the least squares estimates $\widehat{\alpha_i}$ of the α_i for $i = 1, \ldots, t$.

(b) The growth rings of 50 mature trees felled in each of four forests were recorded and analysed in R giving the following output, where some of the values have been omitted.

```
> growth.aov = aov(rings ~ forest)
> summary(growth.aov)
              Df Sum Sq Mean Sq F value    Pr(>F)
forest         i 131344      ii      iv   < 2.2e-16
Residuals    196   4270     iii
---
```

Give the values represented by `i` to `iv`. What conclusions can you draw from the ANOVA table?

(c) The average number of rings counted in the trees felled in each forest were as follows. Compute the parameter estimates $\widehat{\mu}$ and $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_4$.

Mean ring counts

Forest of Dean	Grisedale Forest	Kielder Forest	New Forest
96.0	85.6	151.0	125.3

(d) Use the method of least significant difference and Bonferroni's method to compare the average number of rings counted in trees in each of the four forests at the 5% level. What conclusions can you draw from your results?

In carrying out these comparisons, note that the exact values that you need are not available from the standard statistical tables included; use appropriate values close to those needed. In particular, use values with approximately the correct degrees of freedom and significance level.

3. In the two-way fixed effects ANOVA model, the kth response on level i of factor A and level j of factor B is typically expressed as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad (1)$$

for $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, n$.

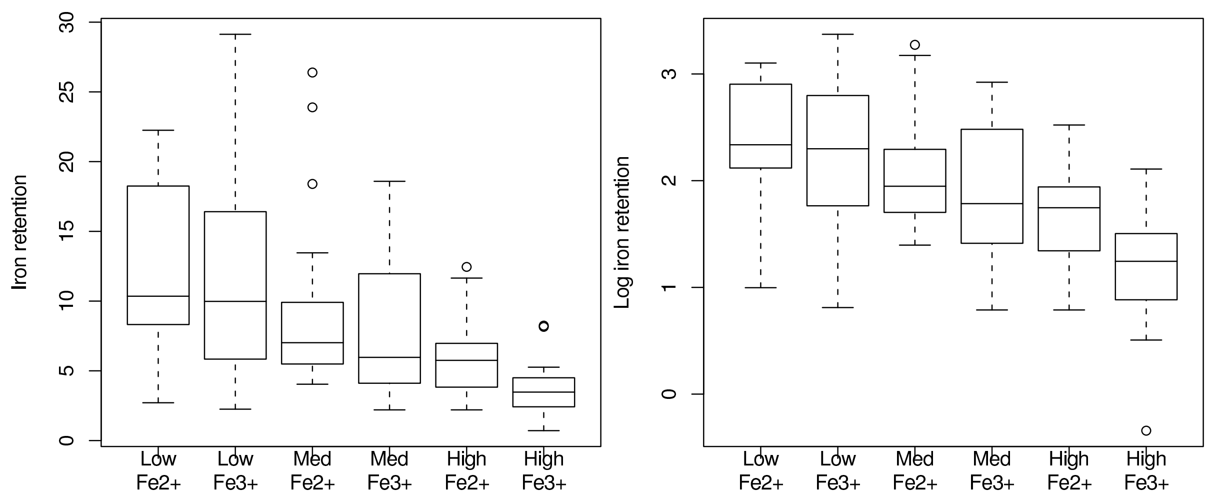
(a) Explain what each of the terms on the right hand side of (1) represent, and write down a set of constraints that will ensure that the terms are identifiable.

(b) For both the two-way random effects model and the two-way mixed effects model, write down equivalent equations to (1). For any terms which appear in your equations but are not in (1), explain what they represent.

(c) An experiment was conducted to determine the most effective type and dosage of iron to use as a dietary supplement. A total of 108 healthy volunteers took tablets containing low, medium, or high doses of either Fe^{2+} or Fe^{3+} and the percentage of iron retained after one day was recorded.

Boxplots of the retention and of the log retention are shown below. Explain why these indicate that it would be better to analyse the log retention data rather than the retention data.

Which of the fixed, random, or mixed ANOVA models is most appropriate to use here? Justify your choice.

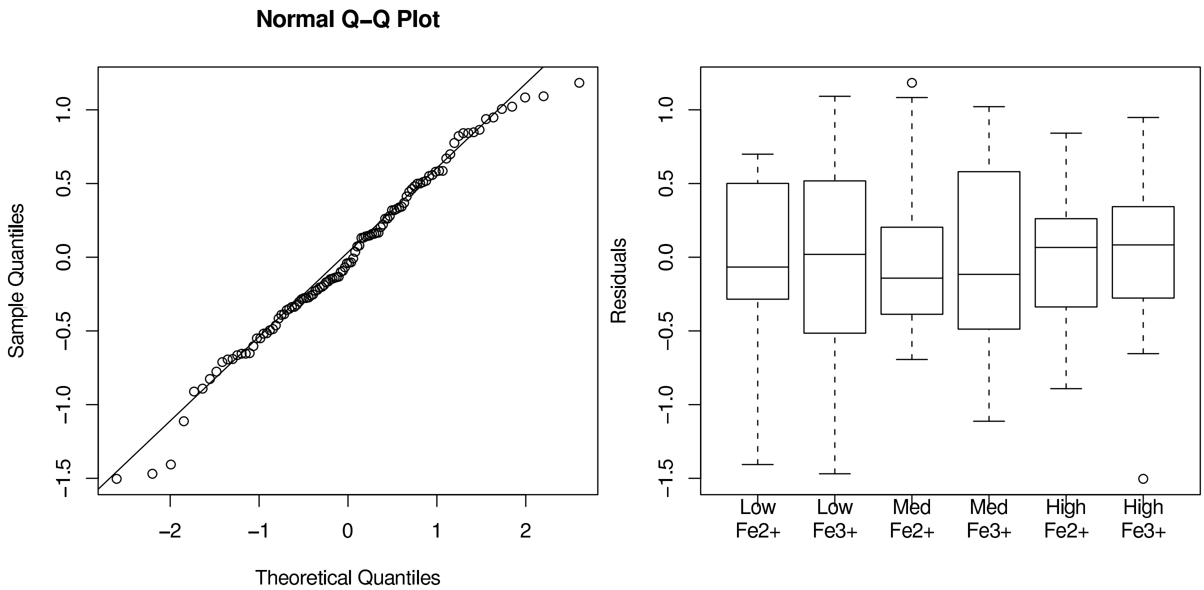


 QUESTION 3 CONTINUED...

(d) The logged data were analysed in R with the results reproduced below, where some of the values have been omitted. Give the values represented by `i` to `vii`. From your completed ANOVA table summarise the conclusions that you can draw from these data.

```
> iron.aov = aov(log(retention) ~ type * conc)
> summary(iron.aov)
             Df Sum Sq Mean Sq F value
type          1  2.074       i       v
conc          2 15.588      ii      vi
type:conc     2  0.810     iii     vii
Residuals   102 35.296      iv
```

(e) A QQ plot of the residuals for the ANOVA model fitted in part (d) and boxplots of the residuals in each iron type / dosage group are shown below. Comment on these plots and their implications, if any.



4. The following equation defines a fixed effects nested design model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk},$$

where Y_{ijk} denotes the kth response in the jth subgroup of the ith main group for $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, n$. Here, μ represents the overall mean, α_i represents the effects of the ith main group, $\beta_{j(i)}$ represents the effect of the jth subgroup within the ith main group, and the random error terms ε_{ijk} are independently $N(0, \sigma^2)$ distributed.

(a) Explain when a nested design is appropriate. Explain what makes the nested model above different to a one-way fixed effects model and say how you could analyse this type of data using a one-way model.

(b) Let the mean squares for main groups, subgroups, and error be defined by

$$
\begin{aligned}
MS_{MG} &= \frac{bn}{a-1}\sum_{i=1}^{a}(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet})^2, \\
MS_{SG} &= \frac{n}{a(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet})^2, \\
\text{and} \quad MS_E &= \frac{1}{ab(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1^n}(Y_{ijk} - \overline{Y}_{ij\bullet})^2,
\end{aligned}
$$

where, in the usual notation, a $\bullet$ indicates that a subscript has been summed over and a bar over a variable indicates the mean of that variable.

As usual, $E(MS_E) = \sigma^2$. Write down, without proof, the values of $E(MS_{MG})$ and $E(MS_{SG})$.

There are two hypotheses of potential interest: that there are no main group effects, and that there are no subgroup effects. Formulate both of these as statistical null hypotheses, and give the appropriate alternative hypotheses and test statistics.

Write down the distributions of the test statistics under their null hypotheses and use the expressions for $E(MS_E) = \sigma^2$, $E(MS_{MG})$ and $E(MS_{SG})$ to indicate how these distributions change if the null hypothesis is not true.

(c) In a mill, rope with a nominal breaking strain of 100kg is spun on four different machines, each machine filling bobbins on two independent spindles at once. Over a period of one week, five samples of rope are taken from each spindle and the breaking strain tested. The following data (breaking strain in 10 kg units) were gathered.

Rope breaking strain (10s of kgs)

Machine:	one		two		three		four	
Spindle:	A	B	A	B	A	B	A	B
	7.5	14.2	10.7	9.7	8.2	9.9	10.3	8.8
	5.8	15.1	13.1	8.5	10.4	11.1	10.1	9.9
	7.9	13.5	11.2	8.4	6.2	9.6	9.5	11.8
	5.7	14.2	9.5	11.5	10.1	9.0	11.5	10.7
	7.3	13.0	9.4	10.2	7.7	11.1	9.4	10.5
Totals:	34.2	70.0	53.9	48.3	42.6	50.7	50.8	51.7

Given the summary statistics $y_{\bullet\bullet\bullet} = 402.2$, $\sum_{ijk} y_{ijk}^2 = 4235.08$, $\sum_{ij} y_{ij\bullet}^2 = 20946.52$, and $\sum_i y_{i\bullet\bullet}^2 = 40513.62$, complete an ANOVA table for these data.

Are there significant differences in the breaking strain of the rope between machines or spindles? What action, if any, would you advise the mill owners to take?

5. Given n pairs of observed data $\{(X_i, Y_i); i = 1, \ldots, n\}$, where Y_i is an observed value of a response variable with corresponding covariate value X_i, we may use the centred simple linear regression model

$$Y_i = \alpha + \beta(X_i - \overline{X}) + \varepsilon_i. \qquad (1)$$

Assume that the ε_i are independently $N(0, \sigma^2)$ distributed. Recall that the least squares estimates of α and β are

$$\widehat{\alpha} = \overline{Y} \qquad \text{and} \qquad \widehat{\beta} = \frac{S_{XY}}{S_{XX}}, \qquad (2)$$

where $S_{XY} = \sum_i (X_i - \overline{X})(Y_i - \overline{Y})$ and $S_{XX} = \sum_i (X_i - \overline{X})^2$.

(a) Write (1) in vector form with the parameters α and β being incorporated into a parameter vector $\boldsymbol{\theta}$. Use this vector representation of the simple linear regression model to show that the least squares estimates of α and β are as given in (2).

Briefly explain the advantage of using the vector representation of the simple linear regression model.

(b) Raw material used in the production of synthetic fibre was stored in a room with no air conditioning. For 12 successive production batches, the relative humidity of the storage area and the moisture content of the resulting fibre were recorded (both as percentages), giving the following data.

Fibre humidity and moisture content

	% Humidity	% Moisture content
	46	12
	53	14
	37	11
	42	13
	34	10
	29	8
	60	17
	44	12
	41	10
	48	15
	33	9
	40	13
Totals:	507	144

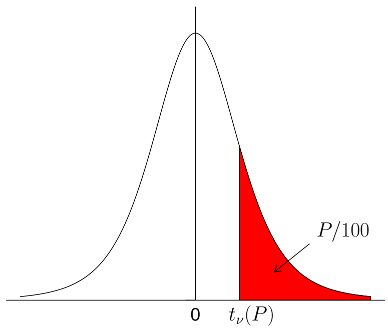
Regarding moisture content as the response, and given that $S_{xy} = 230$ and $S_{xx} = 844.25$, fit a linear regression model to these data and report the fitted regression line.

Given that the residual sum of squares about the regression line is $SS_{RES} = 11.34$, find a 95% confidence interval for the regression parameter β.

Comment on the implications of your regression line and confidence interval.

Percentage Points of the t-Distribution

This table gives the percentage points $t_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the fi gure to the right.



The lower percentage points are given by symmetry as $-t_u(P)$, and the probability that $|t| \geq t_u(P)$ is $2P/100$.

	Percentage points P						
ν	**10**	**5**	**2.5**	**1**	**0.5**	**0.1**	**0.05**
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

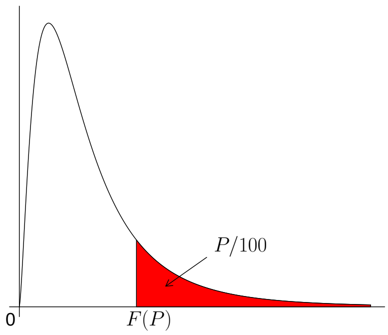
CONTINUED...

10 Percent Points of the F-Distribution

This table gives the percentage points $F_{\nu_1,\nu_2}(P)$ for $P = 0.10$ and degrees of freedom ν_1, ν_2, as indicated by the figure to the right.

The lower percentage points, that is the values $F'_{\nu_1,\nu_2}(P)$ such that the probability that $F \leq F'_{\nu_1,\nu_2}(P)$ is equal to $P/100$, may be found using the formula

$$F'_{\nu_1,\nu_2}(P) = 1/F_{\nu_1,\nu_2}(P)$$



ν_2 \ ν_1	**1**	**2**	**3**	**4**	**5**	**6**	**12**	**24**	**∞**
2	8.526	9.000	9.162	9.243	9.293	9.326	9.408	9.450	9.491
3	5.538	5.462	5.391	5.343	5.309	5.285	5.216	5.176	5.134
4	4.545	4.325	4.191	4.107	4.051	4.010	3.896	3.831	3.761
5	4.060	3.780	3.619	3.520	3.453	3.405	3.268	3.191	3.105
6	3.776	3.463	3.289	3.181	3.108	3.055	2.905	2.818	2.722
7	3.589	3.257	3.074	2.961	2.883	2.827	2.668	2.575	2.471
8	3.458	3.113	2.924	2.806	2.726	2.668	2.502	2.404	2.293
9	3.360	3.006	2.813	2.693	2.611	2.551	2.379	2.277	2.159
10	3.285	2.924	2.728	2.605	2.522	2.461	2.284	2.178	2.055
11	3.225	2.860	2.660	2.536	2.451	2.389	2.209	2.100	1.972
12	3.177	2.807	2.606	2.480	2.394	2.331	2.147	2.036	1.904
13	3.136	2.763	2.560	2.434	2.347	2.283	2.097	1.983	1.846
14	3.102	2.726	2.522	2.395	2.307	2.243	2.054	1.938	1.797
15	3.073	2.695	2.490	2.361	2.273	2.208	2.017	1.899	1.755
16	3.048	2.668	2.462	2.333	2.244	2.178	1.985	1.866	1.718
17	3.026	2.645	2.437	2.308	2.218	2.152	1.958	1.836	1.686
18	3.007	2.624	2.416	2.286	2.196	2.130	1.933	1.810	1.657
19	2.990	2.606	2.397	2.266	2.176	2.109	1.912	1.787	1.631
20	2.975	2.589	2.380	2.249	2.158	2.091	1.892	1.767	1.607
25	2.918	2.528	2.317	2.184	2.092	2.024	1.820	1.689	1.518
30	2.881	2.489	2.276	2.142	2.049	1.980	1.773	1.638	1.456
40	2.835	2.440	2.226	2.091	1.997	1.927	1.715	1.574	1.377
50	2.809	2.412	2.197	2.061	1.966	1.895	1.680	1.536	1.327
100	2.756	2.356	2.139	2.002	1.906	1.834	1.612	1.460	1.214
∞	2.706	2.303	2.084	1.945	1.847	1.774	1.546	1.383	1.002

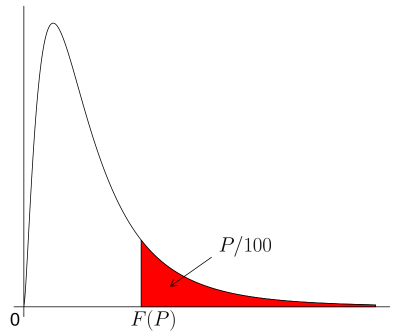
CONTINUED...

5 Percent Points of the F-Distribution

This table gives the percentage points $F_{\nu_1,\nu_2}(P)$ for $P = 0.05$ and degrees of freedom ν_1, ν_2, as indicated by the figure to the right.

The lower percentage points, that is the values $F'_{\nu_1,\nu_2}(P)$ such that the probability that $F \leq F'_{\nu_1,\nu_2}(P)$ is equal to $P/100$, may be found using the formula

$$F'_{\nu_1,\nu_2}(P) = 1/F_{\nu_1,\nu_2}(P)$$



	ν_1								
ν_2	**1**	**2**	**3**	**4**	**5**	**6**	**12**	**24**	∞
2	18.513	19.000	19.164	19.247	19.296	19.330	19.413	19.454	19.496
3	10.128	9.552	9.277	9.117	9.013	8.941	8.745	8.639	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	5.912	5.774	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.678	4.527	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.000	3.841	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.575	3.410	3.230
8	5.318	4.459	4.066	3.838	3.687	3.581	3.284	3.115	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.073	2.900	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	2.913	2.737	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	2.788	2.609	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.687	2.505	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.604	2.420	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.534	2.349	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.475	2.288	2.066
16	4.494	3.634	3.239	3.007	2.852	2.741	2.425	2.235	2.010
17	4.451	3.592	3.197	2.965	2.810	2.699	2.381	2.190	1.960
18	4.414	3.555	3.160	2.928	2.773	2.661	2.342	2.150	1.917
19	4.381	3.522	3.127	2.895	2.740	2.628	2.308	2.114	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.278	2.082	1.843
25	4.242	3.385	2.991	2.759	2.603	2.490	2.165	1.964	1.711
30	4.171	3.316	2.922	2.690	2.534	2.421	2.092	1.887	1.622
40	4.085	3.232	2.839	2.606	2.449	2.336	2.003	1.793	1.509
50	4.034	3.183	2.790	2.557	2.400	2.286	1.952	1.737	1.438
100	3.936	3.087	2.696	2.463	2.305	2.191	1.850	1.627	1.283
∞	3.841	2.996	2.605	2.372	2.214	2.099	1.752	1.517	1.002

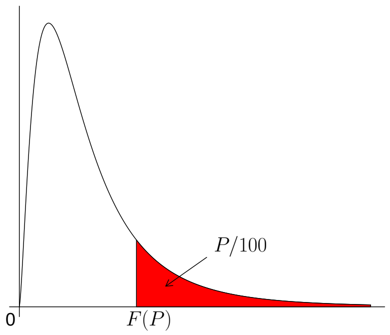
CONTINUED...

1 Percent Points of the F-Distribution

This table gives the percentage points $F_{\nu_1,\nu_2}(P)$ for $P = 0.01$ and degrees of freedom ν_1, ν_2, as indicated by the figure to the right.

The lower percentage points, that is the values $F'_{\nu_1,\nu_2}(P)$ such that the probability that $F \leq F'_{\nu_1,\nu_2}(P)$ is equal to $P/100$, may be found using the formula

$$F'_{\nu_1,\nu_2}(P) = 1/F_{\nu_1,\nu_2}(P)$$



	ν_1								
ν_2	**1**	**2**	**3**	**4**	**5**	**6**	**12**	**24**	∞
2	98.503	99.000	99.166	99.249	99.299	99.333	99.416	99.458	99.499
3	34.116	30.817	29.457	28.710	28.237	27.911	27.052	26.598	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.374	13.929	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	9.888	9.466	9.020
6	13.745	10.925	9.780	9.148	8.746	8.466	7.718	7.313	6.880
7	12.246	9.547	8.451	7.847	7.460	7.191	6.469	6.074	5.650
8	11.259	8.649	7.591	7.006	6.632	6.371	5.667	5.279	4.859
9	10.561	8.022	6.992	6.422	6.057	5.802	5.111	4.729	4.311
10	10.044	7.559	6.552	5.994	5.636	5.386	4.706	4.327	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.397	4.021	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.155	3.780	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	3.960	3.587	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	3.800	3.427	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	3.666	3.294	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	3.553	3.181	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.455	3.084	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.371	2.999	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.297	2.925	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.231	2.859	2.421
25	7.770	5.568	4.675	4.177	3.855	3.627	2.993	2.620	2.169
30	7.562	5.390	4.510	4.018	3.699	3.473	2.843	2.469	2.006
40	7.314	5.179	4.313	3.828	3.514	3.291	2.665	2.288	1.805
50	7.171	5.057	4.199	3.720	3.408	3.186	2.562	2.183	1.683
100	6.895	4.824	3.984	3.513	3.206	2.988	2.368	1.983	1.427
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.185	1.791	1.003

END