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 2007)

ANALYSIS OF EXPERIMENTAL DATA

Time allowed: **2 hours**

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

Throughout this examination paper, replacing a subscript by a $\bullet$ denotes that the subscript has been summed over. A bar over a quantity indicates that averaging has taken place. For example, given data Y_{ij} for $i = 1, \ldots, t$ and $j = 1, \ldots, n$, use the notation

$$Y_{i\bullet} = \sum_{j=1}^{n} Y_{ij}, \qquad Y_{\bullet\bullet} = \sum_{i=1}^{t} \sum_{j=1}^{n} Y_{ij},$$

$$\overline{Y}_{i\bullet} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}, \quad \text{and} \quad \overline{Y}_{\bullet\bullet} = \frac{1}{nt} \sum_{i=1}^{t} \sum_{j=1}^{n} Y_{ij}.$$

1. (a) Prove the one-way ANOVA identity

$$\sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2 = n\sum_{i=1}^{t}(\overline{Y}_{i\bullet}-\overline{Y}_{\bullet\bullet})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})^2. \qquad (1)$$

Each of the terms in (1) is a sum of squares term. Name the terms and say what they represent. How can the last term be used to estimate σ^2?

(b) An airline company wanted to find out whether people who rang their ticket sales line were more or less likely to remain on "hold" according to the type of recording they listened to. Callers were randomised to one of three recordings: advertising for other services, pop music, or classical music. (Only a small number of callers were randomised as these callers were not answered but left to see how long it would take them to hang up.) The time it took for callers to hang up in minutes were as follows.

Time on hold

Recording	Holding times y_{ij}	Mean $\overline{y}_{i\bullet}$
(1) Advert	5, 1, 11, 2, 8	5.4
(2) Pop music	0, 1, 4, 6, 3	2.8
(3) Classical	13, 9, 8, 15, 7	10.4

Analysing these data using a one-way layout in R gave the following edited output.

```
> anova(lm(hold.time ~ recording))
Analysis of Variance Table

Response: time.A
          Df Sum Sq Mean Sq F value
recording  i  149.2      ii     iii
Residuals 12  139.2    11.6
```

Fill in the missing values `i`, `ii`, and `iii`. Use your values to determine whether the type of recording has a significant effect on the time that a person will remain on hold. What type of music would you advise the airline company to use?

(c) Consider the following hypothetical data set.

Time on hold

Recording	Holding times y_{ij}	Mean $\overline{y}_{i\bullet}$
(1) Advert	3, 1, 13, 1, 9	5.4
(2) Pop music	0, 0, 5, 8, 1	2.8
(3) Classical	16, 7, 7, 16, 6	10.4

Assuming that $SS_E = 271.2$ for these data, estimate σ^2 and compare your estimate to that from part (b).

Analysing these data gives an observed F-value of $F_{obs} = 3.300$. Does this value lead to a different conclusion to that in part (b)? If so, why do these data lead to a different conclusion despite the group means being the same?

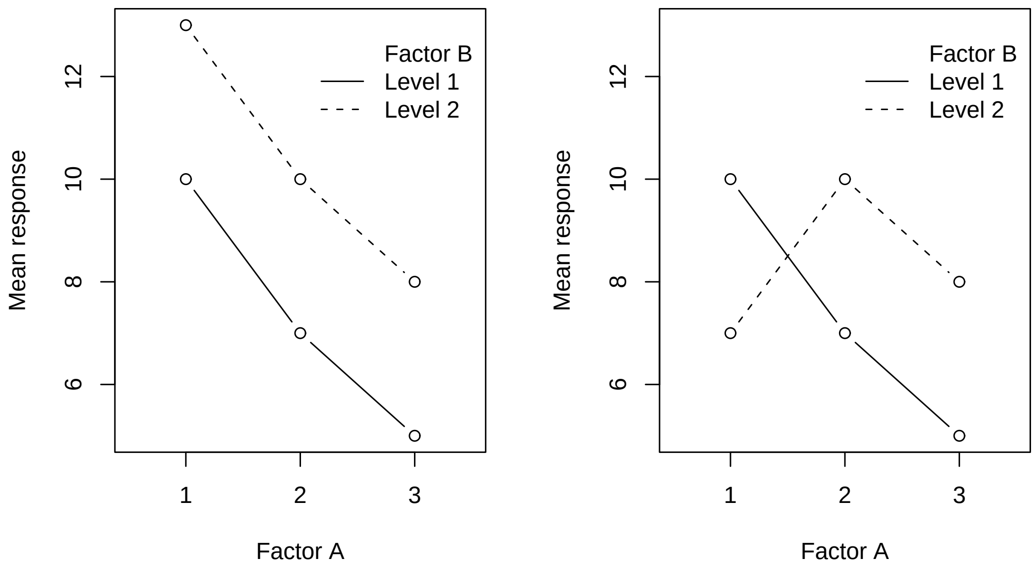
2. Consider a two-way fixed effects ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{1}$$

with $i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n$ and the ε_{ijk} being independently $N(0, \sigma^2)$ distributed.

(a) Explain what each of the terms on the right hand side of (1) represent and list constraints that we usually apply to the terms α_i, β_j, and γ_{ij}.

(b) Assuming that $a = 3$ and $b = 2$, the figures below shows the mean response $\overline{y}_{ij\bullet}$ for each combination of factors A and B for two particular data sets. What can you deduce about the parameters γ_{ij} for each panel of the plot? What does this tell you about the presence or absence of interaction effects?



(c) A randomised experiment studied the weight gain (in grams) of male rats given a diet varying by source of protein (beef, cereal, pork) and level of protein (low, high). Ten rats were randomised to each condition and their average weight gains are shown in the table below.

	Low protein	High protein
Beef	79.2	100.0
Cereal	83.9	85.9
Pork	78.7	99.5

Sketch a plot like those in part (b) for these data. From your plot, which of protein source, protein level, and the interaction between protein source and protein level would you expect to have a significant effect on weight gain?

3. In a one-way random effects model we assume that observation Y_{ij} is modelled by the equation $Y_{ij} = \mu + A_i + \varepsilon_{ij}$ for $i = 1, \ldots, t$ and $j = 1, \ldots, n$. Here, $A_i \sim N(0, \sigma_A^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$ with all A_i and ε_{ij} being mutually independent.

(a) Let $SS_A = n\sum_{i=1}^{t}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2$. Show that $SS_A = n\sum_{i=1}^{t}\overline{Y}_{i\bullet}^2 - N\overline{Y}_{\bullet\bullet}^2$ and use this result to show that $E(SS_A) = (t-1)\sigma^2 + n(t-1)\sigma_A^2$.

(b) Power generation companies are not permitted to burn coal with over 5% sulphur content. A power company was concerned over the sulphur content of coal from one supplier. To determine whether hoppers of coal were consistent in their sulphur content, six hoppers of coal were chosen at random. From each hopper, four samples were taken and their sulphur content measured giving the data below.

Percentage sulphur content of coal samples

	Hopper					
	I	II	III	IV	V	VI
	2	2	4	3	6	4
	3	2	4	3	7	6
	4	2	4	3	5	5
	4	4	5	5	6	7
$Y_{i\bullet}$	13	10	17	14	24	22

Why is a random effects model suitable to analyse these data? If instead of hoppers of coal, these batches represented different types of coal, would random effects still be appropriate? Justify your answer.

Assuming that $\sum_{i=1}^{t}\sum_{j=1}^{n} y_{ij}^2 = 470$ and $n^{-1}\sum_{i=1}^{t} y_{i\bullet}^2 = 453.5$, complete an ANOVA table and carry out a suitable hypothesis test to determine whether there are any significant differences in the sulphur content of the hoppers of coal.

Estimate the components of variance for these data and find a 95% confidence interval for σ^2. Comment on the proportion of variation in the data that is due to the hopper effects.

CONTINUED...

4. Consider a two-way fixed effects randomised complete block design (RCBD) which models observations Y_{ij} as $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ for $i = 1, \ldots, a$ and $j = 1, \ldots, b$. Here α_i represents the effect of the ith treatment and β_j the effect of the jth block, subject to the constraints $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$.

(a) By minimising the sum of squared errors $S = \sum_{i=1}^{a} \sum_{j=1}^{b} \varepsilon_{ij}^2$, find the least squares estimates of the parameters μ, α_i and β_j.

(b) By explaining the distinction between treatments and blocks, discuss how a RCBD helps to systematically reduce the variation in an experiment that is due to sources in which we are not interested. You should illustrate your answer by saying which are the treatments and which are the blocks in the experiment described in part (c) below and justifying your choice.

(c) A company was about to invest in new production equipment. Machines were obtained from seven different suppliers. Ten experienced machine operators used each machine for one week (the order in which they used the machines was randomised). Productivity for each machine/operator pair was measured and these data were used to determine which machine design would be ordered in future.

The data gathered in the above experiment were analysed as a RCBD in R with the following edited results.

```
Analysis of Variance Table

Response: productivity
          Df  Sum Sq Mean Sq F value
machine    6  5021.2   836.9  4.0743
operator   9  4360.8   484.5  2.3590
Residuals 54 11091.7   205.4
```

The mean productivity for each machine and each operator are given below (higher values indicate greater productivity). Machines are labelled I to VII while operators are labelled A to J.

Mean productivity by machine

I	II	III	IV	V	VI	VII
61.88	82.33	78.38	82.50	64.87	68.46	83.06

Mean productivity by operator

A	B	C	D	E	F	G	H	I	J
66.23	85.79	73.71	70.51	60.34	74.26	81.70	69.23	85.46	77.74

What conclusions can you draw from the ANOVA table? Was the use of a RCBD necessary in this case?

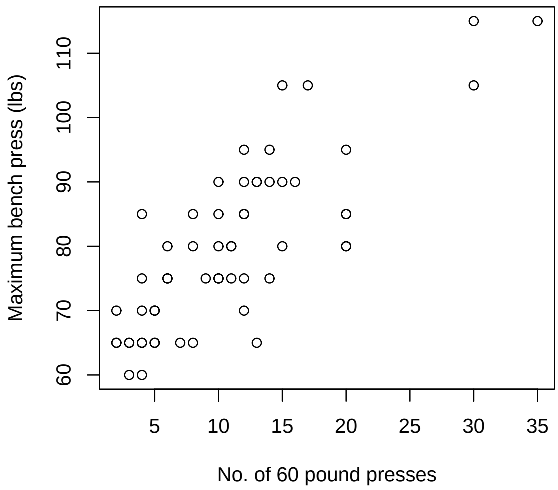
In light of your conclusions and the mean productivity figures, what recommendations would you give to the company?

5. Consider the centred simple linear regression model $Y_i = \alpha + \beta(X_i - \overline{X}) + \varepsilon_i$ with the ε_i being independent $N(0, \sigma^2)$ error terms.

An Australian study into the health of female high school students recorded the number of times a student could bench press a 60 pound weight (x) and the maximum weight the student could bench press (y, in pounds) for $n = 57$ students.

(a) The figure below shows the data gathered in the study. Is linear regression an appropriate tool to analyse these data?

Assuming that linear regression is appropriate, give one issue with the data that might contradict the assumptions of linear regression.



(b) Given that $\overline{x} = 10.982, \overline{y} = 79.912, S_{xx} = 2856.982, S_{yy} = 9874.561$, and $S_{xy} = 4259.912$, estimate the regression parameters α and β. Using the fact that $SS_{RES} = 3522.806$, carry out a two-sided test of the hypothesis H_0: $\beta = 0$.

(c) Construct an ANOVA table for these data and use your table to test the hypothesis H_0: $\beta = 0$.

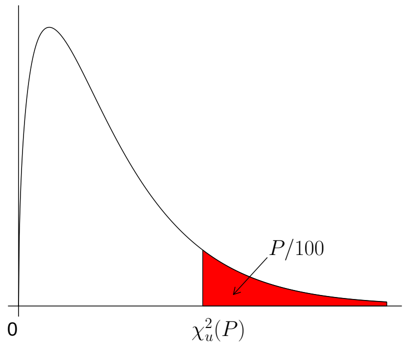
(d) Let $U \sim N(0, 1), V \sim \chi^2_n$, and $W \sim \chi^2_1$ be three independent random variables. Define the t_n and $F_{1,n}$ distributions in terms of U, V, and W. Hence deduce a relationship between the t_n and $F_{1,n}$ distributions. What relevance does this have to your answers to parts (b) and (c)?

Percentage Points of the χ^2-Distribution

This table gives the percentage points $\chi^2_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the figure to the right.

If X is a variable distributed as χ^2 with ν degrees of freedom, $P/100$ is the probability that $X \geq \chi^2_\nu(P)$.

For $\nu > 100$, $\sqrt{2X}$ is approximately normally distributed with mean $\sqrt{2\nu - 1}$ and unit variance.



	Percentage points P						
ν	**10**	**5**	**2.5**	**1**	**0.5**	**0.1**	**0.05**
1	2.706	3.841	5.024	6.635	7.879	10.828	12.116
2	4.605	5.991	7.378	9.210	10.597	13.816	15.202
3	6.251	7.815	9.348	11.345	12.838	16.266	17.730
4	7.779	9.488	11.143	13.277	14.860	18.467	19.997
5	9.236	11.070	12.833	15.086	16.750	20.515	22.105
6	10.645	12.592	14.449	16.812	18.548	22.458	24.103
7	12.017	14.067	16.013	18.475	20.278	24.322	26.018
8	13.362	15.507	17.535	20.090	21.955	26.124	27.868
9	14.684	16.919	19.023	21.666	23.589	27.877	29.666
10	15.987	18.307	20.483	23.209	25.188	29.588	31.420
11	17.275	19.675	21.920	24.725	26.757	31.264	33.137
12	18.549	21.026	23.337	26.217	28.300	32.909	34.821
13	19.812	22.362	24.736	27.688	29.819	34.528	36.478
14	21.064	23.685	26.119	29.141	31.319	36.123	38.109
15	22.307	24.996	27.488	30.578	32.801	37.697	39.719
16	23.542	26.296	28.845	32.000	34.267	39.252	41.308
17	24.769	27.587	30.191	33.409	35.718	40.790	42.879
18	25.989	28.869	31.526	34.805	37.156	42.312	44.434
19	27.204	30.144	32.852	36.191	38.582	43.820	45.973
20	28.412	31.410	34.170	37.566	39.997	45.315	47.498
25	34.382	37.652	40.646	44.314	46.928	52.620	54.947
30	40.256	43.773	46.979	50.892	53.672	59.703	62.162
40	51.805	55.758	59.342	63.691	66.766	73.402	76.095
50	63.167	67.505	71.420	76.154	79.490	86.661	89.561
80	96.578	101.879	106.629	112.329	116.321	124.839	128.261

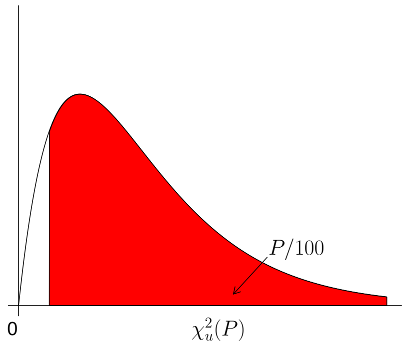
CONTINUED...

Percentage Points of the χ^2-Distribution

This table gives the percentage points $\chi^2_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the figure to the right.

If X is a variable distributed as χ^2 with ν degrees of freedom, $P/100$ is the probability that $X \geq \chi^2_\nu(P)$.

For $\nu > 100$, $\sqrt{2X}$ is approximately normally distributed with mean $\sqrt{2\nu - 1}$ and unit variance.

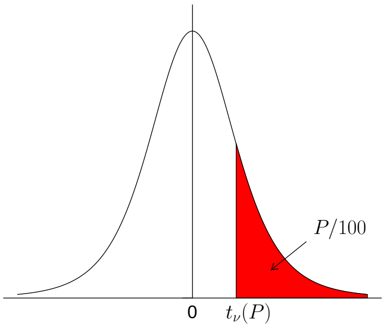


ν	Percentage points P							
	99.95	**99.9**	**99.5**	**99**	**97.5**	**95**	**90**	**80**
1	3.9e-07	1.6e-06	3.9e-05	1.6e-04	9.8e-04	3.9e-03	1.6e-02	6.4e-02
2	0.001	0.002	0.010	0.020	0.051	0.103	0.211	0.446
3	0.015	0.024	0.072	0.115	0.216	0.352	0.584	1.005
4	0.064	0.091	0.207	0.297	0.484	0.711	1.064	1.649
5	0.158	0.210	0.412	0.554	0.831	1.145	1.610	2.343
6	0.299	0.381	0.676	0.872	1.237	1.635	2.204	3.070
7	0.485	0.598	0.989	1.239	1.690	2.167	2.833	3.822
8	0.710	0.857	1.344	1.646	2.180	2.733	3.490	4.594
9	0.972	1.152	1.735	2.088	2.700	3.325	4.168	5.380
10	1.265	1.479	2.156	2.558	3.247	3.940	4.865	6.179
11	1.587	1.834	2.603	3.053	3.816	4.575	5.578	6.989
12	1.934	2.214	3.074	3.571	4.404	5.226	6.304	7.807
13	2.305	2.617	3.565	4.107	5.009	5.892	7.042	8.634
14	2.697	3.041	4.075	4.660	5.629	6.571	7.790	9.467
15	3.108	3.483	4.601	5.229	6.262	7.261	8.547	10.307
16	3.536	3.942	5.142	5.812	6.908	7.962	9.312	11.152
17	3.980	4.416	5.697	6.408	7.564	8.672	10.085	12.002
18	4.439	4.905	6.265	7.015	8.231	9.390	10.865	12.857
19	4.912	5.407	6.844	7.633	8.907	10.117	11.651	13.716
20	5.398	5.921	7.434	8.260	9.591	10.851	12.443	14.578
25	7.991	8.649	10.520	11.524	13.120	14.611	16.473	18.940
30	10.804	11.588	13.787	14.953	16.791	18.493	20.599	23.364
40	16.906	17.916	20.707	22.164	24.433	26.509	29.051	32.345
50	23.461	24.674	27.991	29.707	32.357	34.764	37.689	41.449
80	44.791	46.520	51.172	53.540	57.153	60.391	64.278	69.207

CONTINUED...

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 figure 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 $2P/100$.

ν	Percentage points P						
	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

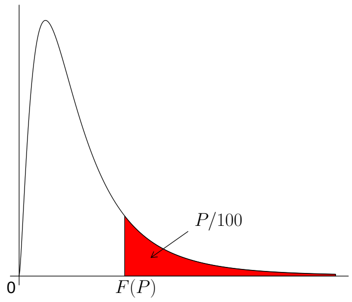
CONTINUED...

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 $F \le F'_{\nu_1,\nu_2}(P)$ 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

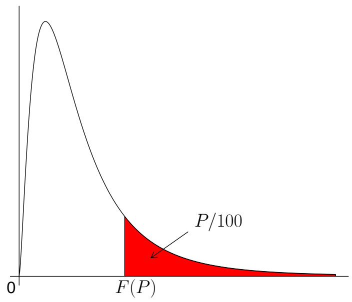
CONTINUED...

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