Published online in Wiley Online Library: 3 August 2015

(wileyonlinelibrary.com) DOI: 10.1002/cem.2734

The *F* distribution and its relationship to the chi squared and *t* distributions

by Richard G. Brereton

Journal of Chemometrics Columnist

We have seen that the *t* distribution is appropriate for estimating critical values or confidence limits when a population has an underlying normal distribution, but the sample size is small. This is primarily a consequence of the difficulty of determining a population standard deviation, and using a method that more often than not underestimates it, the apparent distribution from the mean is distorted.

Many fundamental multivariate methods use the F distribution and its associated tests and critical values: it is the basis of the many common statistical tests in chemometrics, for example, for detecting outliers or whether an observation belongs to a predefined class.

CHI SQUARED

When we discussed the chi squared distribution [1], we noted that this represented the distribution of squared Mahalanobis distances from the mean, and in particular that if more than one variable is measured, there is no specific positive or negative direction, and as such, using squared distances (which are independent of direction) was essential. Hence, the chi-square distribution naturally extends from univariate to multivariate data.

The *F* distribution can be regarded as the equivalent extension of the *t* distribution when there is more than one variable but small sample sizes. There are numerous ways of introducing this distribution in the literature, which is widely employed in many diverse areas. In this and the next article, we focus primarily on the distribution of data in multidimensional space: the *F* distribution is often introduced in the context of analysis of variance. We will come across this distribution and its associated statistic in other contexts in later articles.

R. A. FISHER

The *F* is named after R. A. Fisher, who was a pioneering statistician most active in the 1920s and 1930s, and who worked in agricultural science in the UK. Many of the fundamental multivariate methods such as several approaches for one class classification or class modelling, including SIMCA and multivariate statistical process control, use the *F* distribution and its associated tests and critical values. It is the basis of the many common statistical tests in chemometrics, for example, for detecting outliers or whether an observation belongs to a predefined class. However, in order to understand it, it is necessary to also understand its relationship to other distributions.

DEGREES OF FREEDOM

- The *F* distribution is characterized by two different types of degrees of freedom.
- It is often written $F(\nu_1, \nu_2)$. The horizontal axes of an *F* distribution cumulative distribution function (cdf) or probability density function represent the *F* statistic. We will see in the next article that if there is more than one variable, it is not equal to the squared Mahalanobis distance, unlike the chi-square statistic.
- In our context, if we consider a sample consisting of *n* observations and *k* variables, then ν_1 represents the number of variables, and ν_2 the number of observations minus the number of variables $(n - k)$.
- Note that $F(\nu_1, \nu_2) \neq F(\nu_2, \nu_1)$.
- If a dataset is represented in matrix format, then the number of rows equals $\nu_2 + \nu_1$, and the number of columns equals ν_1.
- Note that it is common to define the dimensions of a matrix using rows first and then columns, but for the purpose of the *F* distribution, we swap these round, with the first degree of freedom referring to the number of variables or columns in a data matrix. This is illustrated in Figure 1.
- Note that *n* cannot be less than *k*. This apparent limitation has been discussed in the context of the Mahalanobis distance [2]. However, it can be overcome by performing principal components analysis first to reduce the number of variables.
- It is important to remember that, rather like the chi squared distribution, an *F* distribution is only obtained if the variables are independent. If, however, we use the Mahalanobis distance measure, this will always be so as discussed previously [2] as this in practice is equivalent to performing a principal component transformation.

UNIVARIATE RELATIONSHIPS

If there is only 1 degree of freedom, the data are univariate, and several straightforward relationships can be derived.

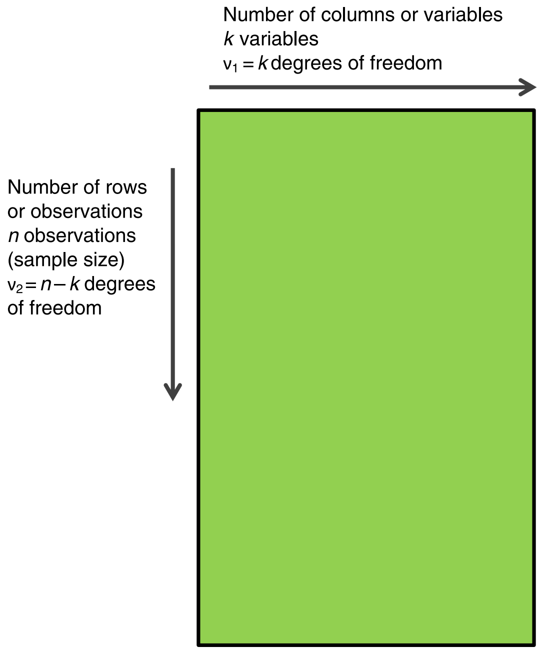


Figure 1. A data matrix.

- For the sake of illustration, we consider a measurement with an *F* value of 4.
- If there is a large sample size, then the *F* distribution, chi squared distribution, and the t^2 distributions all give the same results.
- In Excel, type F.DIST(4,1,10 000 − 1,TRUE), putting $n = 10\,000$: the 4 representing the value of *F*, the 1 equal to v_1, and the 10 000 − 1 equal to v_2. The logical value 'TRUE' represents a cumulative distribution. This should give the proportion of the data that is expected to have an *F* statistic less than 4.
- To check the chi squared distribution, type CHISQ.DIST(4,1,TRUE).
- For *t*, we should remember that we are dealing with squared distances for *F* and chi squared, but their square root when using *t*, although of course this example involves only one variable. Type (T.DIST(SQRT(4),10 000,TRUE) − 0.5)*2. This rather long expression is because we need to calculate the expected proportion of the data whose *t* statistic lies between +2 and −2, either side of the mean.
- The answer in all cases should be 0.954. Hence, 95.4% of the population have a chi squared or *F* statistic less than 4, or a *t*-statistic less than 2 (the square root of 4).

DISTRIBUTIONS

If there are a large number of observations (i.e. v_2 is large), then the shape of the *F* distribution is very similar to the chi squared distribution with v_1 degrees of freedom as illustrated in Figure 2, although there is a shift in position (in fact, chi squared equals v_1 *F*, and for 1 degree of freedom, they are both the same as $v_1 = 1$). Note that if both v_1 and v_2 are large, the *F* distribution also resembles the normal distribution, with a mean of 1.

Figure 3 illustrates several different *F* distributions. We can note several things.

- *F*(2, 5) and *F*(2, 50) are both very similar. In such situations, the sample sizes (7 and 52 respectively) are substantially greater than the number of variables (=2).

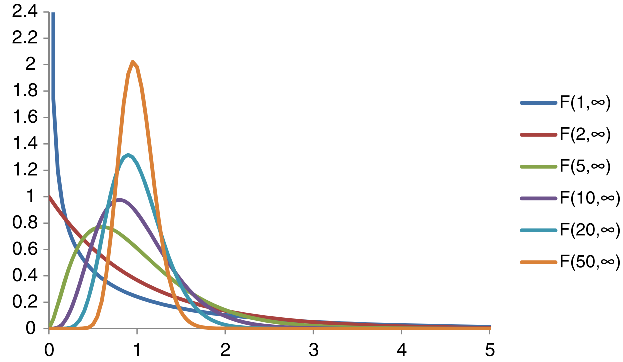


Figure 2. *F* distribution with a very large (effectively infinite) number of observations, as the number of variables (or v_1) increases. Horizontal axis, *F*-statistic; vertical axis, probability density function.

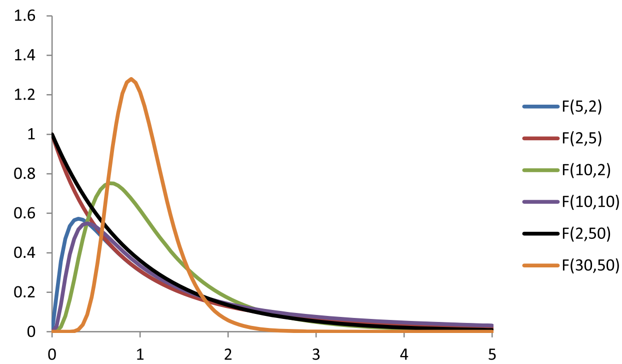


Figure 3. Several different *F* distributions. Horizontal axis, *F* statistic; vertical axis, probability density function.

- *F*(5, 2) and *F*(10, 10) are also similar in shape. The sample sizes (7 and 20) are not so much greater than the number of variables (5 and 10 respectively). Note that there is no problem if $v_1 > v_2$, only *k* must be less than *n*.
- *F*(2, 5) and *F*(5, 2) are very different. Never confuse the two different types of degrees of freedom.
- *F*(30, 50) resembles a normal distribution, as both degrees of freedom are large.

Note the following properties.

- The mean of the distribution is equal to $v_2/(v_2 - 2)$ for $v_2 > 2$.
- The variance is equal to $[2v_2^2\,(v_1 + v_1 - 2)]/[v_1(v_2 - 2)^2\,(v_2 - 4)]$ for $v_2 > 4$

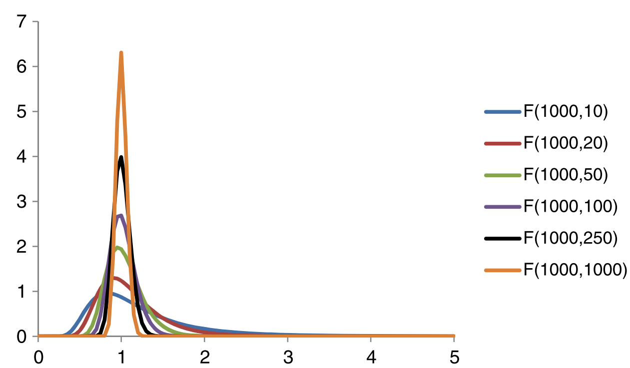
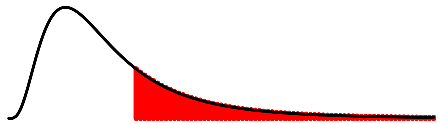


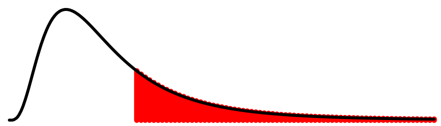
Figure 4. Several different *F* distributions with a very large number of variables (1000). Horizontal axis, *F* statistic; vertical axis, probability density function.

Table I. Critical values of F distribution, $p = 0.01$



ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	75	100	250	500	∞
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143	6157	6209	6260	6286	6302	6324	6334	6353	6360	6366
2	98.50	99.00	99.16	99.25	99.30	99.33	99.36	99.38	99.39	99.40	99.41	99.42	99.42	99.43	99.43	99.45	99.47	99.48	99.48	99.48	99.49	99.50	99.50	99.499
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92	26.87	26.69	26.50	26.41	26.35	26.28	26.24	26.17	26.15	26.125
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25	14.20	14.02	13.84	13.75	13.69	13.61	13.58	13.51	13.49	13.463
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.963	9.888	9.825	9.770	9.722	9.553	9.379	9.291	9.238	9.166	9.130	9.064	9.042	9.020
6	13.75	10.92	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.790	7.718	7.657	7.605	7.559	7.396	7.229	7.143	7.091	7.022	6.987	6.923	6.901	6.880
7	12.25	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.538	6.469	6.410	6.359	6.314	6.155	5.992	5.908	5.858	5.789	5.755	5.692	5.671	5.650
8	11.26	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.734	5.667	5.609	5.559	5.515	5.359	5.198	5.116	5.065	4.998	4.963	4.901	4.880	4.859
9	10.56	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.178	5.111	5.055	5.005	4.962	4.808	4.649	4.567	4.517	4.449	4.415	4.353	4.332	4.311
10	10.04	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.772	4.706	4.650	4.601	4.558	4.405	4.247	4.165	4.115	4.048	4.014	3.951	3.930	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.462	4.397	4.342	4.293	4.251	4.099	3.941	3.860	3.810	3.742	3.708	3.645	3.624	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.220	4.155	4.100	4.052	4.010	3.858	3.701	3.619	3.569	3.501	3.467	3.404	3.382	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	4.025	3.960	3.905	3.857	3.815	3.665	3.507	3.425	3.375	3.307	3.272	3.209	3.187	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.864	3.800	3.745	3.698	3.656	3.505	3.348	3.266	3.215	3.147	3.112	3.048	3.026	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.730	3.666	3.612	3.564	3.522	3.372	3.214	3.132	3.081	3.012	2.977	2.913	2.891	2.868
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.294	3.231	3.177	3.130	3.088	2.938	2.778	2.695	2.643	2.572	2.535	2.468	2.445	2.421
30	7.562	5.390	4.510	4.018	3.699	3.473	3.305	3.173	3.067	2.979	2.906	2.843	2.789	2.742	2.700	2.549	2.386	2.299	2.245	2.170	2.131	2.057	2.032	2.006
40	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.727	2.665	2.611	2.563	2.522	2.369	2.203	2.114	2.058	1.980	1.938	1.860	1.833	1.805
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698	2.625	2.563	2.508	2.461	2.419	2.265	2.098	2.007	1.949	1.868	1.825	1.742	1.713	1.683
75	6.985	4.900	4.054	3.580	3.272	3.052	2.887	2.758	2.653	2.567	2.494	2.431	2.377	2.329	2.287	2.132	1.960	1.866	1.806	1.720	1.674	1.583	1.551	1.516
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503	2.430	2.368	2.313	2.265	2.223	2.067	1.893	1.797	1.735	1.646	1.598	1.501	1.466	1.427
250	6.737	4.691	3.861	3.395	3.091	2.875	2.711	2.583	2.479	2.392	2.319	2.256	2.202	2.154	2.111	1.953	1.774	1.674	1.608	1.511	1.457	1.343	1.297	1.244
500	6.686	4.648	3.821	3.357	3.054	2.838	2.675	2.547	2.443	2.356	2.283	2.220	2.166	2.117	2.075	1.915	1.735	1.633	1.566	1.465	1.408	1.285	1.232	1.164
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.248	2.185	2.130	2.082	2.039	1.878	1.696	1.592	1.523	1.419	1.358	1.220	1.153	1.000

Table II. Critical values of F distribution, $p = 0.05$



ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	30	40	50	75	100	250	500	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	244.7	245.4	245.9	248.0	250.1	251.1	251.8	252.6	253.0	253.8	254.1	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42	19.43	19.45	19.46	19.47	19.48	19.48	19.49	19.49	19.49	19.50
3	10.13	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.763	8.745	8.729	8.715	8.703	8.660	8.617	8.594	8.581	8.563	8.554	8.537	8.532	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891	5.873	5.858	5.803	5.746	5.717	5.699	5.676	5.664	5.643	5.635	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655	4.636	4.619	4.558	4.496	4.464	4.444	4.418	4.405	4.381	4.373	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976	3.956	3.938	3.874	3.808	3.774	3.754	3.726	3.712	3.686	3.678	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550	3.529	3.511	3.445	3.376	3.340	3.319	3.290	3.275	3.248	3.239	3.230
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259	3.237	3.218	3.150	3.079	3.043	3.020	2.990	2.975	2.947	2.937	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048	3.025	3.006	2.936	2.864	2.826	2.803	2.771	2.756	2.726	2.717	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887	2.865	2.845	2.774	2.700	2.661	2.637	2.605	2.588	2.558	2.548	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761	2.739	2.719	2.646	2.570	2.531	2.507	2.473	2.457	2.426	2.415	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660	2.637	2.617	2.544	2.466	2.426	2.401	2.367	2.350	2.318	2.307	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577	2.554	2.533	2.459	2.380	2.339	2.314	2.279	2.261	2.229	2.218	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507	2.484	2.463	2.388	2.308	2.266	2.241	2.205	2.187	2.154	2.142	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.507	2.475	2.448	2.424	2.403	2.328	2.247	2.204	2.178	2.142	2.123	2.089	2.078	2.066
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.310	2.278	2.250	2.225	2.203	2.124	2.039	1.994	1.966	1.927	1.907	1.869	1.856	1.843
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.126	2.092	2.063	2.037	2.015	1.932	1.841	1.792	1.761	1.718	1.695	1.652	1.637	1.622
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.038	2.003	1.974	1.948	1.924	1.839	1.744	1.693	1.660	1.614	1.589	1.542	1.526	1.509
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.986	1.952	1.921	1.895	1.871	1.784	1.687	1.634	1.599	1.551	1.525	1.475	1.457	1.438
75	3.968	3.119	2.727	2.494	2.337	2.222	2.134	2.064	2.007	1.959	1.919	1.884	1.853	1.826	1.802	1.712	1.611	1.555	1.518	1.466	1.437	1.381	1.360	1.338
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927	1.886	1.850	1.819	1.792	1.768	1.676	1.573	1.515	1.477	1.422	1.392	1.331	1.308	1.283
250	3.879	3.032	2.641	2.408	2.250	2.135	2.046	1.976	1.917	1.869	1.827	1.791	1.759	1.732	1.707	1.613	1.505	1.443	1.402	1.341	1.306	1.232	1.202	1.166
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850	1.808	1.772	1.740	1.712	1.686	1.592	1.482	1.419	1.376	1.312	1.275	1.194	1.159	1.113
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831	1.789	1.752	1.720	1.692	1.666	1.571	1.459	1.394	1.350	1.283	1.243	1.152	1.106	1.000

If v_1 and v_2 are large, then we can see from the equations earlier that the mean is approximately equal to 1 and the variance to $4/v_2$, so the larger v_2, the sharper the Gaussian (and more symmetric). This is illustrated in Figure 4 in the case where $v_1 = 1000$. Such situations, whilst rare in traditional statistical applications, may often be encountered in chemometrics where there may be a large number of variables, although would still require large sample sizes. Note that the mode changes position as the distribution becomes more symmetric as v_2 increases.

PROBABILITY VALUES

The probability values corresponding to *F* distribution can easily be computed in Excel or most common environments such as MATLAB, and we have already introduced the basic syntax earlier. Note that earlier versions of Excel may have a simplified syntax; we refer to Office 2010 or above.

- What is the probability that an observation has an *F* value of greater than 2, if we measure 10 variables and our sample size is 30, that is, we have a data matrix of 30 rows and 10 columns?
- First, calculate v_1 that equals 10. Then calculate v_2 that equals 30–10 or 20.
- So, we are looking at *F*(10, 20).
- The syntax is F.DIST(2,10,20,TRUE) and should give an answer 0.910. The ‘TRUE’ implies that we are interested in the cdf of the *F* distribution. This implies that providing the underlying data are normally distributed, 9% of the data are expected to have an *F* value greater than 2.

TABLES

In traditional statistical texts, it is usual to present *F* distribution tables. Because there are rather many possible *F* distributions, these are usually presented as critical values. A critical value of $p = 0.01$ gives the value of the *F* statistic that is expected to be exceeded by only 1% of the data, or in some cases, this can be called the 99% confidence limit. These tables are self-evident and are given in Tables I and II for two critical values. Note that *F* tables can be presented for different critical values. There are several more comprehensive tables available on the web [3,4] although it is recommended that *p* values are calculated in Excel or any other common environment. Note that the tables later are presented for the one-tailed *F* cdf in this article. In some contexts, it is appropriate to look at two-tailed *F* tests, but we will not at this phase be concerned with this.

Richard G. Brereton

University of Bristol, School of Chemistry, Cantocks Close,
Bristol, United Kingdom, BS8 1TS
E-mail: r.g.brereton@bris.ac.uk

REFERENCES

1. Brereton RG. The chi squared and multinormal distributions. *J. Chemometr.* 2014; **29**: 9–12. DOI: 10.1002/cem.2680.
2. Brereton RG. The Mahalanobis distance and its relationship to principal component scores. *J. Chemometr.* 2015; **29**: 143–145. DOI: 10.1002/cem.2692.
3. NIST/SEMATECH e-Handbook of Statistical Methods, April 2012, 1.3.6.7.3. Upper Critical Values of the F Distribution, http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm
4. F distribution critical values, *Medcalc*, http://www.medcalc.org/manual/f-table.php