

## Lecture 14

# ANOVA and the F test

### 14.1 Example: Breastfeeding and intelligence

A study was carried out of the relationship between duration of breastfeeding and adult intelligence. The subjects were part of the Copenhagen Perinatal Cohort, 9125 individuals born at the Copenhagen University Hospital between October 1959 and December 1961. As reported in [MMSR02], a subset of the cohort was contacted for follow-up as adults (between ages 20 and 34). 983 subjects completed the Danish version of the Wechsler Adult Intelligence Scale (WAIS).

Table 14.1 shows the average scores for 3 tests, for 5 classes of breastfeeding duration. Look first at the rows marked “Unadjusted mean”. We notice immediately that 1) Longer breastfeeding was associated with higher mean intelligence scores, but 2) The longest breastfeeding (more than 9 months) was associated with lower mean scores. We ask whether either or both of these associations is reliably linked to the duration of breastfeeding, or whether they could be due to chance.

### 14.2 Digression: Confounding and the “adjusted means”

Before we can distinguish between these two possibilities — causation or chance association — we need to address another possibility: **confounding**. Suppose, for instance, that mothers who smoke are less likely to breastfeed. Since mother’s smoking is known to reduce the child’s IQ scores, this would produce higher IQ scores for the breastfed babies, irrespective of any causal influence of the milk. The gold standard for eliminating confounding is

Table 14.1: Intelligence scores (WAIS) by duration of breastfeeding.

Test		Duration of Breastfeeding (months)				
		$\leq 1$	2-3	4-6	7-9	$> 9$
	N	272	305	269	104	23
Verbal IQ	Unadjusted mean	98.2	101.7	104.0	108.2	102.3
	SD	16.0	14.9	15.7	13.3	15.2
	Adjusted Mean	99.7	102.3	102.7	105.7	103.0
Performance IQ	Unadjusted mean	98.5	100.5	101.8	106.3	102.6
	SD	15.8	15.2	15.6	13.9	14.9
	Adjusted Mean	99.1	100.6	101.3	105.1	104.4
Full Scale IQ	Unadjusted mean	98.1	101.3	103.3	108.2	102.8
	SD	15.9	15.2	15.7	13.1	14.4
	Adjusted Mean	99.4	101.7	102.3	106.0	104.0

the double-blind random controlled experiment. Subjects are assigned at random to receive the treatment or not, so that the only difference between the two groups is whether they received the treatment. (“Double blind” refers to the use of protocols that keep the subjects and the experimenters from knowing who has received the treatment and who is a control. Without blinding, the two groups would differ in their knowledge of having received the treatment or not. We then might be unable to distinguish between effects which are actually due to the treatment, and those that come from *believing* you have received the treatment — in particular, the so-called **placebo effects**.)

Of course, it is usually neither possible nor ethical to randomly assign babies to different feeding regimes. What we have here is an observational study. The next best solution is then to try to remove the confounding. In this case, the researchers looked at all the factors that they might expect to have an effect on adult intelligence — maternal smoking, maternal height, parents’ income, infant’s birthweight, and so on — and adjusted the scores for each category to compensate for a preponderance of characteristics that might be expected to raise or lower IQ in that category, regardless of infant nutrition. Thus, we see that the first and last categories both had their means adjusted substantially upward, which must mean that the infants

who were nursed more than 9 months and those nursed less than 1 month both had, on average, characteristics (whether their own or their mothers') that would seem to predispose them to lower IQ. For the rest of this chapter we will work with the adjusted means.

The statistical technique for doing this, called *multiple regression*, is outside the scope of this course, but it is fairly straightforward, and most textbooks on statistical methods that go beyond the most basic techniques will describe it. Modern statistical software makes it particularly easy to adjust data with multiple regression.

### 14.3 Multiple comparisons

Let us consider the adjusted Full Scale IQ scores. We wish to determine whether the scores of individuals with the same breastfeeding class might have come from the same distribution, with the differences being solely due to random variation.

#### 14.3.1 Discretisation and the $\chi^2$ test

One approach would be to group the IQ scores into groups — low, medium, and high, say. We would then have an incidence table. If these were categorical data — proportions of subjects in each breastfeeding class who scored “high” and “low”, for instance — we could produce an incidence table such as that in Table 14.2. (The data shown here are purely invented, for illustrative purposes.) You have learned how to analyse such a table to determine whether the vertical categories (IQ score) are independent of the horizontal categories (duration of breastfeeding), using the  $\chi^2$  test.

The problem with this approach is self-evident: We have thrown away some of the information that we had to begin with, by forcing the data into discrete categories. Thus, the power to reject the null hypothesis is less than it could have been. Furthermore, we have to draw arbitrary boundaries between categories, and we may question whether the result of our significance test would have come out differently if we had drawn the boundaries otherwise. (These are the same problems, you may recall, that led us to prefer the Kolmogorov-Smirnov test over  $\chi^2$ . The  $\chi^2$  test has the virtue of being wonderfully general, but it is often not quite the best choice.)

Table 14.2: Hypothetical incidence table, if IQ data were categorised into low, medium, and high

Full IQ score	Breastfeeding months				
	$\leq 1$	2-3	4-6	7-9	$> 9$
high	100	115	120	40	9
medium	72	85	69	35	9
low	100	115	80	29	5

### 14.3.2 Multiple t tests

Alternatively, we can compare the mean IQ scores between two different breastfeeding categories, using the t test — effectively, this is the z test, since the number of degrees of freedom is so large, but we still need to pool the variance, because one of the categories has a fairly small number of samples. (The large number of samples also allows us to be reasonably confident in treating the mean as normally distributed, as discussed in section 9.4.) For instance, suppose we wish to compare the children breastfed less than 1 month with those breastfed more than 9 months. We want to test for equality of means, at the 0.05 level.

We compute the pooled standard deviation by

$$s_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} = \sqrt{\frac{271 \cdot 15.9^2 + 22 \cdot 14.4^2}{293}} = 15.8,$$

and the standard error by

$$SE_{\text{diff}} = s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 3.43.$$

This yields a t statistic of

$$t = \frac{\bar{x} - \bar{y}}{SE_{\text{diff}}} = \frac{-4.6}{3.42} = -1.34.$$

Since the cutoff is 1.96, we do not reject the null hypothesis.

If we repeat this test for all 10 pairs of categories, we get the results shown in Table 14.3. We see that 4 out of the 10 pairwise comparisons show statistically significant differences. But what story are these telling together? Remember that if the null hypothesis were true — if the population

means were in fact all the same — 1 out of 20 comparisons should yield a statistically significant difference at the 0.05 level. How many statistically significant differences do we need before we can reject the overall null hypothesis of identical population means? And what if none of the differences were individually significant, but they all pointed in the same direction?

Table 14.3: Pairwise t statistics for comparing all 10 pairs of categories. Those that exceed the significance threshold for the 0.05 level are shown in red.

	2-3	4-6	7-9	> 9
$\leq 1$	-1.78	-2.14	-3.78	-1.34
2-3		-0.47	-2.58	-0.70
4-6			-2.14	-0.50
7-9				0.65

## 14.4 The F test

We will see in lecture 15 how to treat the covariate — the duration of breastfeeding — as a *quantitative* rather than *categorical* variable. That is, how to measure the effect (if any) per unit time breastfeeding. Here we concern ourselves only with the question: Is there a nonrandom difference between the mean intelligence in the different categories? As we discussed in section 14.3, we want to reduce the question to a single test.

What should the test statistic look like? Fundamentally, a test statistic should have two properties: 1) It measures significant deviation from the null hypothesis; that is, one can recognise from the test statistic whether the null hypothesis has been violated substantially. 2) We can compute the distribution of the statistic under the null hypothesis.

### 14.4.1 General approach

Suppose we have independent samples from  $K$  different normal distributions, with means  $\mu_1, \dots, \mu_K$  and variance  $\sigma^2$  (so the variances are all the same). We call these  $K$  groups **levels** (or sometimes **treatments**). We have  $n_i$  samples from distribution  $i$ , which we denote  $X_{k1}, X_{k2}, \dots, X_{kn_k}$ . The goal

is to determine from these samples whether the  $K$  **treatment effects**  $\mu_k$  could be all equal.

We let  $N = \sum_{k=1}^K n_k$  be the total number of observations. The average of all the observations is  $\bar{X}$ , while the average within level  $i$  is  $\bar{X}_i$ :

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

The idea of analysis of variance (ANOVA) is that under the null hypothesis, which says that the observations from different levels really all are coming from the same distribution, the observations should be about as far (on average) from their own level mean as they are from the overall mean of the whole sample; but if the means are different, observations should be closer to their level mean than they are to the overall mean.

We define the **Between Groups Sum of Squares**, or **BSS**, to be the total square difference of the group means from the overall mean; and the **Error Sum of Squares**, or **ESS**, to be the total squared difference of the samples from the means of their own groups. (The term “error” refers to a context in which the samples can all be thought of as measures of the same quantity, and the variation among the measurements represents random error; this piece is also called the Within-Group Sum of Squares.) And then there is the **Total Sum of Squares**, or **TSS**, which is simply the total square difference of the samples from the overall mean, if we treat them as one sample.

$BSS$	$= \sum_{i=1}^K n_i (\bar{X}_i - \bar{X})^2;$	
$ESS$	$= \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$	
	$= \sum_{i=1}^K (n_i - 1) s_i^2$ where $s_i$ is the SD of observations in level $i$ .	$TSS = \sum_{i,j} (X_{ij} - \bar{X})^2$
	$= (n - 1) s^2$ , where $s$ is the sample SD of all observations together.	

The initials **BMS** and **EMS** stand for **Between Groups Mean Squares** and **Error Mean Squares** respectively.

The **analysis of variance (ANOVA)** is based on two mathematical facts. The first is the identity  $TSS = ESS + BSS$ .

In other words, all the variability among the data can be divided into two pieces: The variability within groups, and the variability among the means of different groups.

Our goal is to evaluate the apportionment, to decide if there is “too much” between group variability to be purely due to chance.

Of course,  $BSS$  and  $ESS$  involve different numbers of observations in their sums, so we need to normalise them. We define

$$BMS = \frac{BSS}{K-1} \quad EMS = \frac{ESS}{N-K}.$$

This brings us to the second mathematical fact: if the null hypothesis is true, then EMS and BMS are both estimates for  $\sigma^2$ . On the other hand, interesting deviations from the null hypothesis — in particular, where the populations have different means — would be expected to increase BMS relative to EMS. This leads us to define the deviation from the null hypothesis as the ratio of these two quantities:

$$F = \frac{BMS}{EMS} = \frac{N-K}{K-1} \cdot \frac{BSS}{ESS}.$$

We reject the null hypothesis when  $F$  is too large: That is, if we obtain a value  $f$  such that  $P\{F \geq f\}$  is below the significance level of the test.

Table 14.4: Tabular representation of the computation of the F statistic.

	SS	d.f.	MS	F
Between Treatments	BSS (A)	$K-1$ (B)	BMS ( $X = A/B$ )	$X/Y$
Errors (Within Treatments)	ESS (C)	$N-K$ (D)	EMS ( $Y = C/D$ )	
Total	TSS	$N-1$		

Under the null hypothesis, the F statistic computed in this way has a known distribution, called the F distribution with  $(K-1, N-K)$  degrees of freedom. We show the density of  $F$  for  $K=5$  different treatments and different values of  $N$  in Figure 14.1.

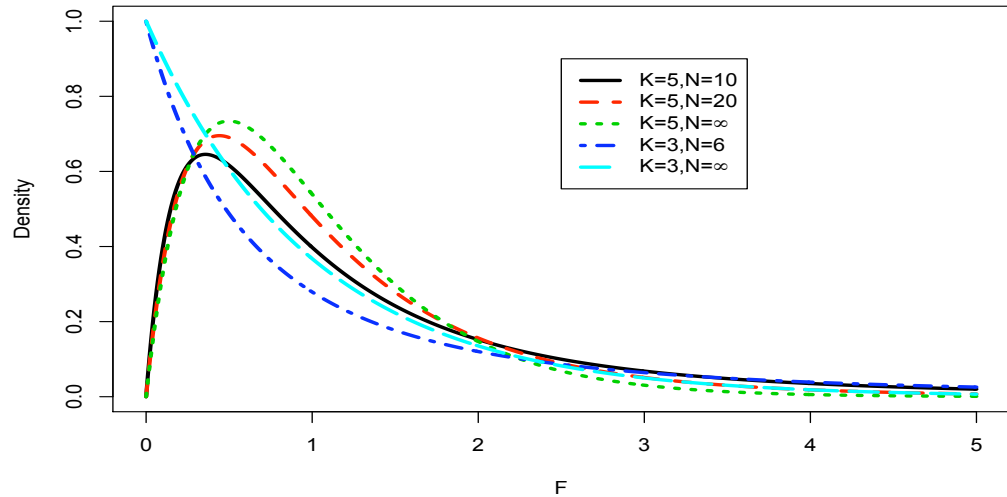


Figure 14.1: Density of F distribution for different values of  $K$  and  $N$ .

#### 14.4.2 The breastfeeding study: ANOVA analysis

In this case, since we do not know the individual observations, we cannot compute TSS directly. We compute

$$\begin{aligned}
 ESS &= \sum_{k=1}^5 (n_k - 1) s_k^2 \\
 &= 271 \cdot 15.9^2 + 304 \cdot 15.2^2 + 268 \cdot 15.7^2 + 104 \cdot 13.1^2 + 23 \cdot 14.4^2 \\
 &= 227000; \\
 BSS &= \sum_{k=1}^5 n_k (x_{k.} - \bar{x})^2 \\
 &= 272 \cdot (99.4 - 101.7)^2 + 305 \cdot (101.7 - 101.7)^2 + 269 \cdot (102.3 - 101.7)^2 \\
 &\quad + 104 \cdot (106.0 - 101.7)^2 + 23 \cdot (104.0 - 101.7)^2 \\
 &= 3597.
 \end{aligned}$$

We complete the computation in Table 14.5, obtaining  $F = 3.81$ . The numbers of degrees of freedom are (4, 968). The table in the official booklet is

quite small — after all, there is one distribution for each pair of integers. The table gives only the cutoff only for select values of  $(d_1, d_2)$  at the 0.05 level. For parameters in between one needs to interpolate, and for parameters above the maximum we go to the row or column marked  $\infty$ . Looking on the table in Figure 14.2, we see that the cutoff for  $F(4, \infty)$  is 2.37. Using a computer, we can compute that the cutoff for  $F(4, 968)$  at level 0.05 is actually 2.38; and the cutoff at level 0.01 would be 3.34.

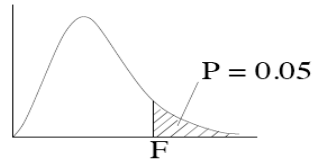
Table 14.5: ANOVA table for breastfeeding data: Full Scale IQ, Adjusted.

	SS	d.f.	MS	F
Between Samples	3597 (A)	4 (B)	894.8 ( $X = A/B$ )	3.81
Errors (Within Samples)	227000 (C)	968 (D)	234.6 ( $Y = C/D$ )	
Total	230600 (TSS=A+C)	972 ( $N - 1$ )		

### 14.4.3 Another Example: Exercising rats

We consider the following example, adapted from [MM98, Chapter 15]. A study was performed to study the effect of exercise on bone density in rats. 30 rats were divided into three groups of ten: The first group carried out ten “high jumps” (60 cm) a day for eight weeks; the second group carried out ten “low jumps” (30 cm) a day for eight weeks; the third group had no special exercise. At the end of the treatment period, each rat’s bone density was measured. The results are given in Table 14.6.

We wish to test the null hypothesis that the different groups have the same mean bone density, against the alternative hypothesis that they have different bone densities. We first carry out the ANOVA analysis. The total sum of squares is 20013.4. The error sum of squares (ESS) is computed as  $9s_1^2 + 9s_2^2 + 9s_3^2 = 12579.5$ . The between-groups sum of squares (BSS) is computed as  $10(638.7 - 617.4)^2 + 10(612.5 - 617.4)^2 + 10(601.1 - 617.4)^2 = 7433.9$  (here the overall mean is 617.4). Note that indeed  $TSS = ESS + BSS$ . We complete the computations in Table 14.7, obtaining  $F = 7.98$ . Looking in the column for 2, and the row for 30 (since there is no row on your



Variance ratio  $F = s_1^2/s_2^2$  with  $\nu_1$  and  $\nu_2$  degrees of freedom respectively.

$\nu_2$	$\nu_1$	1	2	3	4	5	6	8	12	24	$\infty$	$\nu_2$
6		5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67	6
8		5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93	8
10		4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54	10
12		4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30	12
14		4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13	14
16		4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01	16
18		4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92	18
20		4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84	20
30		4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62	30
40		4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51	40
60		4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.39	60
$\infty$		3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00	$\infty$

Figure 14.2: Table of F distribution; finding the cutoff at level 0.05 for the breastfeeding study.

table for 27), we see that the cutoff at level 0.05 is 3.32. Thus, we conclude that the difference in means between the groups is statistically significant.

### 14.5 Multifactor ANOVA

The procedure described in section 14.4 leads to obvious extensions. We have observations

$$x_{ki} = \mu_k + \epsilon_{ki}, \quad k = 1, \dots, K; \quad i = 1, \dots, n_k$$

where the  $\epsilon_{ki}$  are the normally distributed “errors”, and  $\mu_k$  is the true mean for group  $k$ . Thus, in the example of section 14.4.3, there were three groups, corresponding to three different exercise regimens, and ten different samples for each regimen. The obvious estimate for  $\mu_k$  is

$$\bar{x}_{k\cdot} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki},$$

and we use the F test to determine whether the differences among the means are genuine. We decompose the total variance of the observations into the

(a) Full data										
High	626	650	622	674	626	643	622	650	643	631
Low	594	599	635	605	632	588	596	631	607	638
Control	614	569	653	593	611	600	603	593	621	554

(b) Summary statistics		
Group	Mean	SD
High	638.7	16.6
Low	612.5	19.3
Control	601.1	27.4

Table 14.6: Bone density of rats after given exercise regime, in  $\text{mg} / \text{cm}^3$

portion that is between groups and the portion that is within groups. If the between-group variance is too big, we reject the hypothesis of equal means.

Many experiments naturally lend themselves to a two-way layout. For instance, there may be three different exercise regimens and two different diets. We represent the measurements as

$$x_{kji} = \mu_k + \nu_j + \epsilon_{kji}, \quad k = 1, 2, 3; \quad j = 1, 2; \quad i = 1, \dots, n_{kj}.$$

It is then slightly more complicated to isolate the exercise effect  $\mu_k$  and the diet effect  $\nu_j$ . We test for equality of these effects by again splitting the variance into pieces: the total sum of squares falls naturally into four pieces, corresponding to the variance over diets, variance over exercise regimens, variance over joint diet and exercise, and the remaining variance within each group. We then test for whether the ratios of these pieces are too far from the ratio of the degrees of freedom, as determined by the F distribution.

Multifactor ANOVA is quite common in experimental practice, but will not be covered in this course.

## 14.6 Kruskal-Wallis Test

Just as there is the non-parametric rank-sum test, similar to the t and z tests for equality of means, there is a non-parametric version of the F test, called the Kruskal-Wallis test. As with the rank-sum test, the basic idea is

Table 14.7: ANOVA table for rat exercise data.

	SS	d.f.	MS	F
Between Samples	7434 (A)	2 (B)	3717 ( $X = A/B$ )	7.98
Errors (Within Samples)	12580 (C)	27 (D)	466 ( $Y = C/D$ )	
Total	20014 (TSS=A+C)	29 ( $N - 1$ )		

simply to substitute ranks for the actual observed values. This avoids the assumption that the data were drawn from a normal distribution.

In Table 14.8 we duplicate the data from Table 14.6, replacing the measurements by the numbers 1 through 30, representing the ranks of the data: the lowest measurement is number 1, and the highest is number 30. In other words, suppose we have observed  $K$  different groups, with  $n_i$  observations in each group. We order all the observations in one large sequence of length  $N$ , from lowest to highest, and assign to each one its rank. (In case of ties, we assign the average rank.) We then sum the ranks in group  $i$ , obtaining numbers  $R_1, \dots, R_K$ . Then

The Kruskal-Wallis test statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^K \frac{R_i^2}{n_i} - 3(N+1).$$

Under the null hypothesis, that all the samples came from the same distribution,  $H$  has the  $\chi^2$  distribution with  $K - 1$  degrees of freedom.

In the rat exercise example, we have the values of  $R_i$  given in Table 14.7(b), yielding  $H = 10.7$ . If we are testing at the 0.05 significance level, the cutoff for  $\chi^2$  with 2 degrees of freedom is 5.99. Thus, we conclude again that there is a statistically significant difference among the distributions of bone density in the three groups.

(a) Full data

High	18.5	27.5	16.5	30	18.5	25.5	16.5	27.5	25.5	20.5
Low	6	8	23	11	22	3	7	20.5	12	24
Control	14	2	29	4.5	13	9	10	4.5	15	1

(b) Summary statistics

Group	Sum
High	226.5
Low	136.5
Control	102

Table 14.8: Ranks of data in Table 14.6.