SUM OF SQUARES, MEAN SQUARES, AND THE ANALYSIS OF VARIANCE

Overview

In this lecture, we will introduce the analysis of variance (ANOVA). This is the basic tool we will use to analyze many designed experiments.

This lecture is a bit long, because it connects many statistical concepts–old and new—in a coherent framework.

It is OK to take pauses when you study this lecture.

In this lecture, we will introduce the analysis of variance—ANOVA. This is the basic tool we will use to analyze data from many designed experiments.

This lecture is a bit long, because it connects many statistical concepts–old and new—in a coherent framework.

It is OK to take pauses when you study this lecture.

SUM OF SQUARED ERRORS (SSE) AND MEAN SQUARE ERROR (MSE)

Recall the separate-means model

$$y_{ij} = \mu_i + e_{ij} \quad (i = 1, \ldots, t; j = 1, \ldots, r),$$

where e_{ij}'s are assumed i.i.d. normal $N(0, \sigma^2)$.

Recall the separate-means model where each observation y i j equals a group mean, mu i, plus an error term e i j,

Here, i indexes the treatments and j indexes individuals receiving the same treatment.

The error terms, e i j's, are assumed i.i.d. normal with mean 0 and variance sigma squared.

Parameter estimation using observations from a single group

Within group i, the separate-means model simply says all the observations are i.i.d. normally distributed with mean μ_i and variance σ^2:

$$y_{ij} \quad i.i.d. \quad N(\mu_i, \sigma^2), \quad j = 1, \ldots, r.$$

Based only on observations in group i, it is reasonable to estimate the group population mean μ_i by the group sample mean,

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{r} \sum_{j=1}^{r} y_{ij}$$

and estimate the group population variance σ^2 with the sample variance of the r observations in group i

$$s_i^2 = \frac{1}{r-1} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{i\cdot})^2.$$

Within group i, the separate-means model simply says all the observations are i.i.d. normally distributed with mean mu i and variance sigma squared.

Based only on observations in group i, it is reasonable to estimate the group population mean, mu i, by the group sample mean.

and estimate the group population variance sigma squared with the sample variance of the r observations in group i.

Parameter estimation using the complete data set

With the complete data set (data from all groups), the group sample means are still the best estimates of the group population means:
$$\hat{\mu}_i = \overline{y}_{i\cdot}, \quad i = 1, \ldots, t.$$
Since observations from different groups share the same variance parameter σ^2, we can try improving the estimate of the common population variance σ^2 by **pooling information from all groups**.

For this, we will introduce the sum of squared errors. As we will see, the sum of squared errors also tells us how well a statistical model fit a data set, and therefore can be used for comparing two (nested) statistical models being fit to the same data set.

With the complete data set including data from all groups, the group sample means are still the best estimates of the group population means.

Since observations from different groups share the same population variance, sigma squared, we can try improving the estimate of sigma squared by pooling information from all groups.

For this, we will introduce the sum of squared errors. As we will see, the sum of squared errors also tells us how well a statistical model fit a data set, and therefore can be used for comparing two nested statistical models being fit to the same data set.

The sum of squared errors

If $\hat{y}_{ij}$ is the predicted (fitted) value for the ijth observation according to some statistical model, the **error** or the **residual** of unit ij from the fitted model is:

- $y_{ij} - \hat{y}_{ij}$.

The sum of squared errors (SSE) of the fitted model is thus,

- $SSE = \sum(\text{observed} - \text{predicted})^2 = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \hat{y}_{ij})^2$.

For the separate-means model, $\hat{y}_{ij} = \overline{y}_{i\cdot}$, so

- $SSE = \sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij} - \overline{y}_{i\cdot})^2$.

Note that

- $SSE = \sum_{i=1}^{t}\left[\sum_{j=1}^{r}(y_{ij} - \overline{y}_{i\cdot})^2\right] = \sum_{i=1}^{t}(r-1)\cdot s_i^2$,

where s_i^2 is the sample variance of group i observations.

If y hat i j is the predicted or fitted value for the (i, j)th observation according to some statistical model, the error or the residual of unit (i, j) from the fitted model is defined as the difference between the observed value and the fitted value.

The sum of squared errors (SSE) of the fitted model is thus, as the name suggests, the sum of squared errors, in other words, it is the sum of squared differences between the observed values and the corresponding fitted values.

For the separate-means model, the predicted value is simply the group mean, so the SSE equals the sum of squared differences between the observed values and corresponding group sample means.

Note that the SSE equals the sum of each group sample variance multiplied by the group size minus 1 over all treatment groups.

The mean square error (MSE)

If the true variance is the same in all groups (which is true if the experiment is completely randomized), it makes sense to estimate it with a pooled sample variance (a weighted average of variances of each group):

- $\hat{\sigma}^2 = \frac{(r-1)s_1^2 + (r-1)s_2^2 + \cdots + (r-1)s_t^2}{(r-1) + \cdots + (r-1)} = \frac{\sum_{i=1}^{t}(r-1)s_i^2}{\sum_{j=1}^{t}(r-1)} = \frac{SSE}{t(r-1)} = \frac{SSE}{N-t}$

where $N = rt$ is the total number of observations. This is the so-called **mean square error** (MSE, or the mean squared error).

The MSE estimates the population variance (σ^2) under the corresponding fitted model.

It is simply the SSE divided by the number of **degrees of freedom** of the corresponding model.

The population variance σ^2 is a population parameter and the MSE is a sample statistic that estimates σ^2.

If the true population variance is the same in all groups (which is true if the experiment is completely randomized), it makes sense to estimate it with a pooled sample variance, i.e., a weighted average of the sample variances of each group.

Here N = rt is the total number of observations.

This is the so-called mean square error or the mean squared error. In short, MSE.

The MSE estimates the population variance under the corresponding fitted model.

It is simply the SSE divided by the numbre of degrees of freedom of the corresponding model.

The population variance sigma squared is a population parameter and the MSE is a sample statistic that estimates sigma squared.

The number of degrees of freedom for the sample variance

The sample variance in group i is the sum of squared errors from group i divided by the group size minus 1:

- $s_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}(y_{ij} - \bar{y}_{i.})^2.$

$r - 1$ is called **the number of degrees of freedoms (d.f.)** — it counts how many **free-varying** components we have in a sum of squares:

- Here the r error terms or residuals, $y_{ij} - \bar{y}_{i.}, j = 1, \ldots, r$, in the sum are not all free: they add up to 0.
- There are only $r - 1$ free-varying components among the r residuals: once we know the values of any $r - 1$ of them, the remaining one is determined.

The sample variance in group i is the sum of squared errors from group i divided by the group size minus 1.

r-1 is called the number of degrees of freedoms (d.f.):
it counts how many free-varying components we have in a sum of squares.

Here the r error terms or residuals in the sum are not all free: they add up to 0.

There are only r-1 free-varying components among the r residuals: once we know the values of any r-1 of them, the remaining one is determined.

The number of degrees of freedom for the SSE

The number of d.f. for the SSE,

$$SSE = \sum_{i=1}^{t} \left[\sum_{j=1}^{r} (y_{ij} - \overline{y}_{i.})^2 \right] = \sum_{i=1}^{t} (r-1) \cdot s_i^2,$$

is

- the sum of the numbers of d.f. of each group variance and is thus $t(r-1)$.
- $rt - t$, the number of squared error terms in the sum minus the number of estimated parameters: here the t group means μ_i are estimated by the corresponding group sample means $\overline{y}_{i.}$.

The MSE is the SSE divided by the number of d.f. of the SSE.

The number of degrees of freedom for the SSE is the sum of the numbers of degrees of freedom of each group variance and is thus t times (r-1).

The number of degrees of freedom, rt - t, is also the number of squared error terms in the sum minus the number of estimated parameters: here the t group means are estimated by the corresponding group sample means.

The mean square error is the sum of squared errors divided by the number of degrees of freedom of the SSE.

In general, the number of d.f. is also the number of independent normal components we have in the sum after suitable transformation of the data.

SS TREATMENT AND THE F-TEST FOR COMPARING TWO NESTED LINEAR REGRESSION MODELS

Different versions of SSE and MSE

Different models can be fit to the same data set, giving rise to different versions of SSE and MSE with different degrees of freedom.

Different models can be fit to the same data set, giving rise to different versions of SSE and MSE with different degrees of freedom.

A equal-means model

For example, consider an equal-means model, which says that there is no effect of group membership on the response and all of the variation in the y's is 'noise':

- $y_{ij} \quad i.i.d. \quad N(\mu, \sigma^2), i = 1, \ldots, t, j = 1, \ldots, r.$ Or equivalently,
- $y_{ij} = \mu + e_{ij}, e_{ij} \sim N(0, \sigma^2), i = 1, \ldots, t, j = 1, \ldots, r.$

Under this model, there is only one unknown mean parameter μ and it can be estimated by the overall sample mean of all observations:

- $\bar{y}_{..} = \frac{y_{..}}{rt} = \frac{\sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}}{rt}.$

The fitted value is $\bar{y}_{..}$ for all observations and the SSE is

- $SSE_r = \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \hat{y}_{ij})^2 = \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{..})^2.$

The number of d.f. is $rt - 1$ for this SSE_r, and we can estimate σ^2 by

- $MSE_r = SSE_r/(rt - 1).$

For example, consider an equal-means model, which says that there is no effect of group membership on the response and all of the variation in the response variable y is 'noise'.

Under this model, the y i j's are iid normal with mean mu and variance sigma squared.

Or equivalently, each y i j can be written as the sum of the overall mean and an error term that is i i d normally distributed with mean 0 and variance sigma squared.

Under this model, there is only one unknown parameter, the population mean mu, and it can be estimated by the overall sample mean of all observations.

This overall sample mean is also the fitted value for all observations under this model. So the sum of squared errors is the sum of squared differences between each observation and the overall sample mean.

The number of d.f. is rt minus 1 for this SSE. Again, the d.f. is the number of squared error terms in the sum minus the number of estimated mean parameters: in this model, there is only one mean parameter to estimate.

Under this model, we can estimate the population variance sigma squared by the

MSE

Compare the separate-means model and the equal-means model

The separate-means model:

- y_{ij} $i.i.d.$ $N(\mu_i, \sigma^2), j = 1, \ldots, r$, for each group $i = 1, \ldots, t$.

The equal-means model:

- y_{ij} $i.i.d.$ $N(\mu, \sigma^2), i = 1, \ldots, t, j = 1, \ldots, r$, for all observations

What are the differences and connections between the two models?

If we compare the separate-means model to the equal-means model, what are the differences and connections between the two models?

Thank about it for a minute before moving on to the next slide.

Compare the separate-means model and the equal-means model

The separate-means model:

- y_{ij} i.i.d. $N(\mu_i, \sigma^2), j = 1, \ldots, r$, for each group $i = 1, \ldots, t$.
- The separate-means model has t mean parameters and one variance parameter.

The equal-means model:

- y_{ij} i.i.d. $N(\mu, \sigma^2), i = 1, \ldots, t, j = 1, \ldots, r$, for all observations
- The equal-means model has a single mean parameter and one variance parameter.

Therefore, the equal-means model is simpler or coarser. So relatively speaking, we will call the separate-means model the **full model** and the equal-means model the **reduced model**. The separate-means model will reduce to the equal-means model if $\mu_1 = \mu_2 = \ldots = \mu_t$ (which happens to be a hypothesis we often want to test).

We can see that the separate-means model has t mean parameters and one variance parameter, the equal-means model has a single mean parameter and one variance parameter.

Therefore, the equal-means model is simpler or coarser.

So relatively speaking, we will call the separate-means model the full model and the equal-means model the reduced model.

The equal-means model is reduced from the full model by requiring all group means to be equal.

This constraint that all group means are equal happens to be a hypothesis we often want to test.

Partition of the sums of squares

In the full (separate-means) model, the total variability in the data is partitioned as follows (Kuehl, p. 52):

- $\sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \overline{y}_{..})^2 = r \sum_{i=1}^{t} (\overline{y}_{i.} - \overline{y}_{..})^2 + \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \overline{y}_{i.})^2$
- SS Total = SS Treatment + SSE_f, (f for 'full').

The numbers of d.f. associated with these sums of squares are:

- $N - 1$ for SS Total (N observations, grand mean estimated);
- $N - t$ for SSE (N observations, t group means estimated); and
- $(N - 1) - (N - t) = t - 1$ for SS Treatment .

Note that here *SS Total* is the same as the SSE_r (r for "reduced") from the reduced model (the equal-means model). Therefore:

$$\text{SS Treatment} = \text{SS Total} - SSE_f = SSE_r - SSE_f$$

The *SS Treatment* is the amount of variability in the data that is explained by the treatment factor.

In the separate-means model, the total variability in the data is partitioned as follows:

In this equation, the left-hand side the sum of squared difference between each observed value and the overall sample mean over all observations. This term measures the total variation in the data set, so we call it the sum of squares total.

The first term on the right is the sum of squared difference between each group sample mean and the overall sample mean multiplied by the group size. This term measures the variation among group sample means. The group sample means reflect treatment effects. So this term is called sum of squares treatment.

The second term on the right is the sum of squared difference between each observation and the correspond group sample mean. Since the sample means are also the fitted values under the separate-means model. This term is also the sum of squared errors under the separate-means model.

The numbers of degrees of freedoms for SS Total, SS Treatment and SSE are N-1, N-t and t -1 respectively.

Note that sum of squares total is the same as the sum of squared errors from the equal-means model.

Therefore we can also write the partition of sum of squares as

SS Treatment equals SS Total minus SSE from the full model, which in turn equals SSE from the reduced model minus SSE from the full model.

SS Treatment is the amount of variability in the data that is explained by the treatment factor.

The figures on the next two pages illustrates the partitioning of variability in the separate-means and equal-means models.

The figures on the next two pages illustrates the partitioning of variability in the separate-means and equal-means models.

Errors in the equal-means model
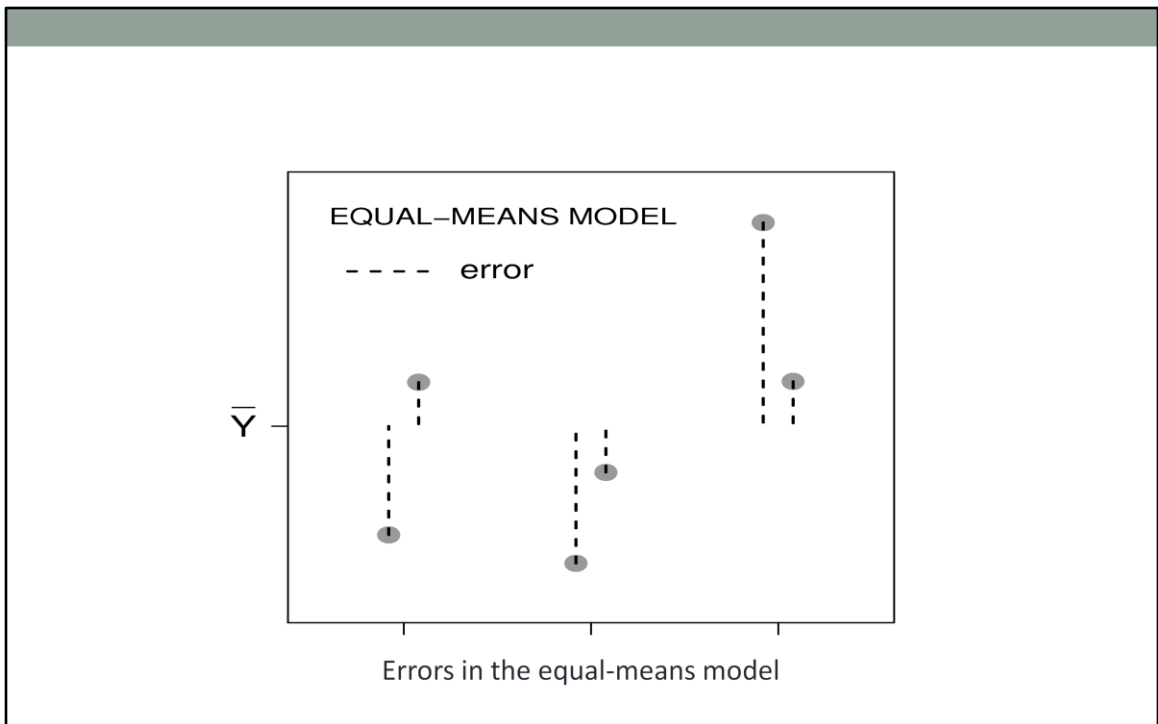
Under the equal-means model, the fitted value is the overall sample mean. All fitted values are the same.

The error from the fitted model is the difference betwen observed value and the overall sample mean.

Under the equal-means model, the sum of squared errors is thus the sum of squared differences between the each observed value and the overall sample mean.

SEPARATE–MEANS MODEL

---- error
—— model

$\overline{Y}_3$

$\overline{Y}$

$\overline{Y}_1$

$\overline{Y}_2$

1   2   3
Group
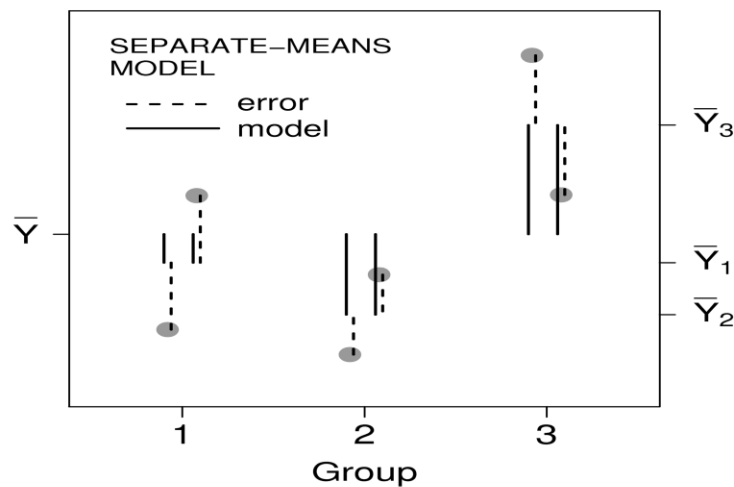
Errors and treatment/group effects in the separate-means model

Under the seperate-means model, the fitted value is the group sample mean. The fitted values are the same within each treatment group, but each group can have a different group sample mean.

The error from the fitted model is the difference betwen observed value and the group sample mean.

Under the separate-means model, the sum of squared errors is thus the sum of squared differences between the each observed value and the corresponding group sample mean.

Interpretation of the SS Treatment

- SSE_r and SSE_f are sums of squared errors from a reduced model and a full model respectively. Their difference, $SST = SSE_r - SSE_f$, measures how much the full model improves upon the simpler reduced model.

- The *SS Treatment* (or *SST*),

$$r \sum_{i=1}^{t} (\bar{y}_{i.} - \bar{y}_{..})^2,$$

- measures the variation in group sample means (more specifically, the deviation of the group sample means from the grand sample mean):
  - SST will be contributions from only noises if all group population means are actually the same.
  - Otherwise, the SST will reflect how much the groups population means differ.

- To test whether all group population means are equal, we can compare MST (SST divided by its number of d.f.) to MSE: if the ratio MST/MSE is large, it indicates the variation in group sample means is much greater than expected from a model where all group population means are the same.

SSE reduced and SSE full are sums of squared errors from a reduced model and from a full model respectively. Their difference, SST, measures how much the full model improves upon the simpler reduced model.

The SS Treatment or SST measures the variation in group sample means, more specifically, the deviation of the group sample means from the grand sample mean.

SST will be contributions from only noises if all group population means are actually the same. Otherwise, the SST will reflect how much the groups population means differ.

To test whether all group population means are equal, we can compare MST (SST divided by its number of d.f.) to MSE: if the ratio MST/MSE is large, it indicates the variation in group sample means is much greater than expected from a model where all group population means are the same.

Expected values of the mean squares and hypothesis testing

The mean squares (MS) are the sums of squares divided by their degrees of freedom. We'll use MSE and MST for the mean square error and mean square treatment, respectively. The expected values of MSE and MST are

- $E(MSE) = \sigma^2$ always; and

- $E(MST) = \sigma^2 + \frac{\sum_{i=1}^{t}(\mu_i - \mu)^2}{t-1}$. $E(MST)$ is equal σ^2 only if $\mu_1 = \cdots = \mu_t = \mu$.

So, if the treatment means differ, $E(MST) > E(MSE)$. MST/MSE can be used to test group mean differences. If $H_0: \mu_1 = \cdots = \mu_t$ is true, then

- $F = \frac{MST}{MSE} = \frac{(SSE_r - SSE_f)/(t-1)}{SSE_f/(N-t)}$

has the F-distribution with $t - 1$ numerator d.f. and $N - t$ denominator d.f. We reject H_0 at level α if our calculated statistic F^* exceeds $F_{\alpha, t-1, N-t}$, the $(1 - \alpha)$-th quantile of the F-distribution with $(t - 1, N - t)$ d.f..

The mean squares are the sums of squares divided by their numbers of degrees of freedom. We'll use MSE and MST for the mean square error and mean square treatment, respectively.

Recall that the expected value of a random variable is the long-run average of the results of many realization of that random variable.

The expected value of MSE is always sigma squared under the fitted model.

The expected value of MST is sigma squared plus the sum of squared differences between each group population mean, mu i, and the overall population mean mu, divided by t-1.

The expected value of MST equals sigma squared only when all group population means are the same.

So, if the treatment means mu i's differ, the expected value of MST will be greater than the expected value of MSE. Therefore, the ratio of MST over MSE can be used to test group mean differences.

If all group population means are the same, then the ratio of MST ove MSE will have an F-distribution with t-1 and N-t degrees of freedom in numerator and denominator

respectively.

We reject H_0 at level alpha if our calculated F-statistic value, denoted here by F asterisk, exceeds the (1 - alpha)-th quantile of the F-distribution with t - 1 and N - t degrees of freedom.

Often in a hypothesis test, we compare a test statistic value to its distribution under the null distribution to get the p-value, which measure the evidence against the null hypothesis.

More generally, the F-test can be thought as comparing a full model, here the separate-means model, to a reduced model, here the equal-means model.

The general form of such an F-test is given in this equation:

The denominator the the MSE under the full model.

The numerator is the difference between SSE from the reduced model and SSE from the full model divided by t-1, which is also the differences in the numbers of degrees freedom between the full model and the reduced model.

It measures the reduction in sum of squared errors achieved by fitting the more complex full model relative to the added number of degrees of freedom for the model.

So basically, this F test is asking "Is it worth it to fit the more complicated full model? Can then added model degrees of freedom in the full model be justified by the amount of reduction in the sum of squares?" Note that added model degrees of freedom means reduced error degrees of freedom.

The constraint put on the reduced model corresponds to the null hypothesis we want to test: in today's example, it is H naught, all group means are the same.

Under H naught, the F-test statistic above has an F-distribution, t-1 and N-t are the numbers of d.f. for the numerator sum of squares (SST) and denominator sum of squares (SSE), and are thus called the numerator and denominator d.f. of this F-distribution.

If the null hypothesis is rejected when the p-value is small, it says the data provide evidence that the simpler reduced model is not adequate: for a CRD experiment, it says the data provide evidence that the groups means are actually differ. The p-value gives the strength of the evidence.

P-VALUE (REVIEW)

p-value

The p-value is the probability of obtaining a test-statistic value at least as extreme as the observed value, given that H_0 is true. A smaller p-value gives stronger evidence **against** the null hypothesis.

If F represents the random variable, and F^* is the value calculated from our data, the p-value for the F-test is

$$p = Pr(F \geq F^* | \mu_1 = \ldots = \mu_t) = Pr(F_{t-1,N-t} \geq F^*),$$

because, if H_0 is true, then $F \sim F_{t-1,N-t}$.

In the blood-pressure example, $F^* = 4.85$ and $p = Pr(F_{2,12} > 4.85) = 0.029$. We have strong evidence ($p = 0.029$) that the mean change in blood pressure varies among dose groups.

The p-value is the probability of obtaining a test-statistic value at least as extreme as the observed value, given that the null hypothesis is true. A smaller p-value gives stronger evidence against the null hypothesis.

If F represents the random variable, and F asterisk is the value calculated from our data, the p-value for the F-test is the probability that a F distributed random variable exceeds the F test statistic value computed from the observed data set.

In the blood-pressure example, the F-statistic value is 4.85 and the corresponding p-value is 0.029. We have strong evidence that the mean change in blood pressure varies among dose groups.

Meaning of the p-value

The p-value, $p = 0.029$, means if the null is actually true—that is, the group means in blood pressure changes are actually the same for all three dose groups—the chance that we observe the data as or more extreme than the set of data we actually observed is 0.029: in other words, if the null hypothesis is actually true, and if we were to repeat our experiment many many times, we would only observe data set as or more extreme in 2.9% of the experiments.

When we see a small p-value, there are two possible scenarios:
- The null is actually true, we just observed an extreme data set by chance.
- The null hypothesis is false.

The smaller the p-value the less likely that the first scenario is true, and thus the more likely that the null hypothesis is false.

The p-value, 0.029, means if the null hypothesis is actually true—that is, the group means in blood pressure changes are actually the same for all three dose groups— then the chance that we observe the data as or more extreme than the set of data we actually observed is 0.029.

In other words, if the null hypothesis is actually true, and if we were to repeat our experiment many many times, only in about 2.9 percent of the experiments, we would expect see a F-test statistic value as or more extreme than the F-test statistic value for the actually observed data set.
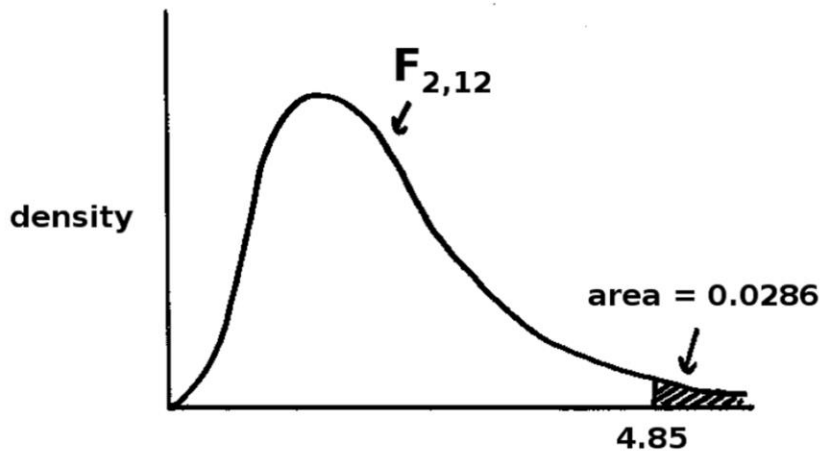
When we see a small p-value, there are two possible scenarios:

One, the null is actually true, we just observed an extreme data set by chance.

Two, the null hypothesis is false.

The smaller the p-value the less likely that the first scenario is true, and thus the more likely that the null hypothesis is false.

Figure of the F-distribution and p-value

This figure illustrates the calculation of F-test p-value.

4.85 is the F statistic computed from the observed data.

The density curve is the the distribution of the F-test statistic under the null.

The p-value is the area of the shaded tail region, which is the probability that the F-statistic will exceed the observed value, 4.85, under the null.

Recall that the F-test statistic measure the difference among treatment mean levels, so a large value of F statistic gives evidence that the group means are different, so F-test is typically one-sided.

Get probabilities and quantiles in R:

$pf(q, df1, df2)$ gives the area of the region under the $F_{df1,df2}$ density curve to the left of q. i.e.,

$$Pr(F_{df1,df2} \leq q).$$

In the previous example, the P-value corresponds to the area of the right tail, which can be computed as

```
1 - pf(4.85, 2, 12)
## [1] 0.02859757
```

$qf(p, df1, df2)$ computes the F-quantiles.

$qf(p, df1, df2)$ give the number such that the area under the density curve to the left of the number is p. (In Kuehl's notation, this would be $F_{1-p,df1,df2}$.)

In R, we can use the function pf to compute probabilties for F-distributed random vairables.

pf (q, df1, df2) gives the area under the F density curve of the region to the left of q, i.e., the probability that the F-distributed random variable is less than q.

In the previous example, the p-value corresponds the area under the density curve of the right tail region, so we compute it as

1 - pf(4.85, 2, 12)

which gives a p-value of 0.029.

We use the function qf to compute quantile of the F-distribution. In other words,

qf(p, df1, df2) give the number such that the area under the density curve to the left of the number is p.

F test is typically one-sided

The only "important" deviation of F^* is in the direction of large values.

An unusually small value of F^* implies that the variation among treatment means is less than you'd expect based on the within-treatment variability. This could happen by chance, but it could also mean that there is some problem with dependence in the data (caused, for example, by non-random sampling).

We explained earlier that the only "important" deviation of the F-statistic is in the direction of large values.

An unusually small value of F value implies that the variation among treatment means is less than you'd expect based on the within-treatment variability. This could happen by chance, but it could also mean that there is some problem with dependence in the data (caused, for example, by non-random sampling).

Rejecting the null hypothesis, type-I and type-II errors (review)

Summarizing a test result of a hypothesis test by a p-value is often enough.

But sometimes, researchers prefer to have a "yes or no" answer. For that, we can choose a cutoff for the p-value, *a significance level*, denoted by α, and **reject** H_0 if the p-value is less than the significance level.

A significance level of $\alpha = 0.05$ is often used, but other values (0.02, 0.01, and so on) can also be used.

In our example, $p = 0.029$, we can reject H_0 at $\alpha = 0.05$ level, but not at $\alpha = 0.02$ level.

If we decide to reject H_0, there is still a chance that H_0 is true. If that happens, we say we made a *type-I error* (false positive), the p-value is the probability of making a type-I error.

A more stringent significance level will lead to less false positives, but also less rejections even when H_0 is false (*type II error*, false negative).

Summarizing a test result of a hypothesis test by a p-value is often enough.

But sometimes, researchers prefer to have a "yes or no" answer. For that, we can choose a cutoff for the p-value, called a significance level, denoted by alpha, and reject the null hypothesis H naught if the p-value is less than the significance level.

A significance level of alpha equals 0.05 is often used, but other values such as 0.02, 0.01, and so on can also be used.

In our earlier example, p=0.029, we can reject the null hypothesis at alpha=0.05 level, but not at alpha = 0.02 level.

If we decide to reject null hypothesis, there is still a chance that H_0 is actually true. If that happens, we say we made a type-I error, or we get a false positive, the p-value is the probability of making a type-I error.

A more stringent significance level will lead to less false positives, but also less rejections even when the null hypothesis is actually false. Failing to reject H naught when it is actually false is a type II error or a false negative.

Understanding type-I and type-II errors are useful for power and sample-size analysis.

We will never accept the null hypothesis (review)

When concluding a hypothesis test, we will never accept the null hypothesis.

When the p-value is big (greater than the specified significance level), we will say we fail to reject the null hypothesis.

If a p-value is large, it says the model(s) specified by the null hypothesis is/are compatible with the data, but there could be many other models (which we did not explore) that are compatible with the data too.

When concluding a hypothesis test, we will never accept the null hypothesis.

When the p-value is big and greater than the specified significance level, we will say we fail to reject the null hypothesis.

If a p-value is large, it says the model(s) specified by the null hypothesis is/are compatible with the data, but there could be many other models that we did not explore that are also ompatible with the data too.
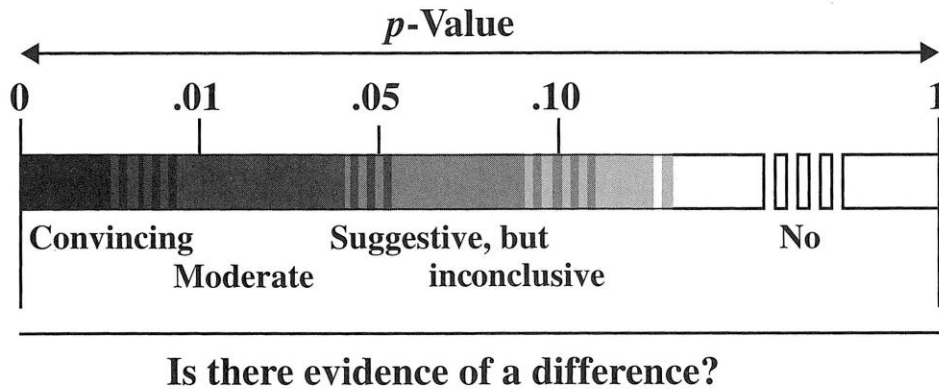
Describing the strength of a p-value

Get in the habit of interpreting the p-value as a continuous measure of the strength of evidence against the null hypothesis, rather than simply rejecting or failing to reject the null, depending on whether the P-value is less than or greater than 0.05. The figure on next page from the Statistical Sleuth gives the idea.

Get in the habit of interpreting the p-value as a continuous measure of the strength of evidence against the null hypothesis, rather than simply rejecting or failing to reject the null, depending on whether the p-value is less than or greater than 0.05.

The figure on next page from the Statistical Sleuth gives the idea.

Is there evidence of a difference?

p-Value

0   .01   .05   .10   1

Convincing   **Suggestive, but**   **No**
     Moderate   **inconclusive**

Is there evidence of a difference?

This figure gives a few adjectivese that can be used to describe the strength of a p-value.

For example, when the p-value is between 0.01 and 0.05, one can say there is moderate evidence against the null hypothesis.

When thinking about the strength a p-value, it is also helpful to relate it some random experiments: for example, a p-value of 0.0625 is the chance to toss a fair coin and get four heads in a row. How surprising is that outcome to you?

THE ANOVA TABLE

The ANOVA table

The partitioning of the sum of squares, the F-test statistic, and the p-value can be compactly summarized in a table called *the ANOVA table*.

The partitioning of the sum of squares, the F-test statistic, and the p-value can be compactly summarized in a table called the ANOVA table or the analysis of variance table.

	d.f.	Sum of squares	Mean square	F	P-value
Treatment	$t - 1$	SST	$MST = \frac{SST}{t-1}$	$F^* = \frac{MST}{MSE}$	$Pr(F_{t-1,N-t} > F^*)$
Error	$N - t$	SSE	$MSE = \frac{SSE}{N-t}$		
Total	$N - 1$	SS Total			

The partition of variability in the full (separate-means) model.

In the ANOVA table, the rows correspond to sources of errors.

The columns are sum of squares, mean squares, F test statistic and p-value, and the numbers of degrees of freedom

The column of sum of squares summarizes the partitioing of the total sum of squares.

Sometimes, the row of total sum of squares is omitted from the table.

The ANOVA table is important tool for analyzing designed experiments. In this class, for each different type of design, we will discuss the structure of its ANOVA table.

This ANOVA table corresponds to a completely randomized design with one treatment factor. For more complex designs, there will be more rows in the ANOVA table.

The ANOVA table for the bp1 data.

	df	SS	MS	F	P-value
Dose (Treatment)	2	411.73	205.87	4.8477	0.0286
Residuals (Error)	12	509.60	42.47		
Total	14	921.33			

Changes in blood pressure in the three dose groups (cont'd).

This is the ANOVA table for analyzing the bp1 data. In practice, R can help us compute the ANOVA table, but we have to let R know the design of the experiment.

If the model equation we specify in R does not match the design of the actual experiment, the ANOVA table we get from R will be wrong.

Note that R cannot tell the design of an experiment simply from the structure of input data. Neither can we: we have to know the design of the experiment to perform the correct analysis of the data.

See **script1** for instructions on fitting the separate-means model in R.

Unequal replication

An **unbalanced** design has different numbers of replicates in the different treatment groups.

In the expressions for the sums of squares, we replace r by $r_i (i = 1, \dots, t)$. For example:

$$\text{SS Treatment} = \sum_{i=1}^{t} \sum_{j=1}^{r_i} (\overline{y}_{i.} - \overline{y}_{i..})^2 = \sum_{i=1}^{t} r_i \, (\overline{y}_{i.} - \overline{y}_{i..})^2.$$

If $r_1 = r_2 = \cdots = r_t = r$, this becomes $r \sum_{i=1}^{t} (\overline{y}_{i.} - \overline{y}_{i..})^2$ as before.

It's probably best to aim for equal replication, but unequal sample sizes in the single-factor design do not affect inference or cause any particular technical difficulties—unlike the situation with some more complicated designs that we'll consider later.

An unbalanced design has different numbers of replicates in the different treatment groups.

In the expressions for the sums of squares, we replace r by r i. For example:

In SS Treatment, the coeffcient for each squared difference between group sample mean and overall sample mean will be the group size r_i.

If all r i's are the same, this becomes the same expression as before.

It's probably best to aim for equal replication, but unequal sample sizes in the single-factor design do not affect inference or cause any particular technical difficulties—unlike the situation with some more complicated designs that we'll consider later.

Summary

The ANOVA table sums it up all:
- The sum of squares, the mean squares, and corresponding degrees of freedom
- The partition of the sum of squares
- The F-test (comparing the full model to a reduced model)

Go over each term in the ANOVA table and explain its meaning. In particular, know how to correctly interpret a p-value.

You should be able to fill the entire ANOVA table given only two of the SS's and their corresponding numbers of d.f. (you will need some software to compute the p-value of the F-test).

To summarize this lecture, we can focus the ANOVA table.

The ANOVA table summarizes the sum of squares, the mean squares, and corresponding degrees of freedom.

It summarizes the partition of the sum of squares and the F-test for comparing the full model to a reduced model.

As a review, we can go over each term in the ANOVA table and explain its meaning. In particular, we should know how to correctly interpret a p-value.

You should be able to fill the entire ANOVA table given only two of the sums of squares and their corresponding numbers of degrees of freedom.

You will need some software to compute the p-value of the F-test.