Inference for Unstructured Multisample Studies (Ch. 7.1, 7.2, 7.4)

Yifan Zhu

Iowa State University

Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

Confidence interval for linear combination of means

The one-way ANOVA model

- Suppose we have:
    - Some response variable, Y
    - Some covariate factor, X, with levels $i = 1, 2, \dots, r$ and n_i observations at level i.
- The **one-way ANOVA model**, sometimes called the one-way normal model, is:

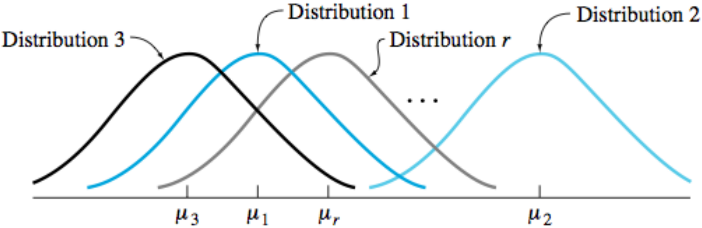
$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:

- The ε_{ij}'s are iid $N(0, \sigma^2)$
- μ_i is the true mean response at level i of the factor.
- $j = 1, 2, \dots, n_i$.

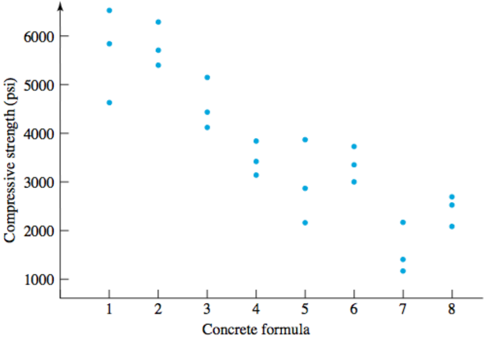
The one-way ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$



Example: concrete

- Compressive strengths of 8 different formulas of concrete:



- But the order of the numbers given to the formulas is meaningless. It wouldn't make sense to do a simple linear regression of strength on formula.

Example: concrete

- Instead of:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

with Y_i as strength and X_i as the formula index, we use:

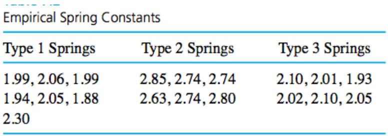
$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:

  - i is the formula index, $i = 1, 2, \ldots, 8$
  - j is the index of a specimen within the formula i group.

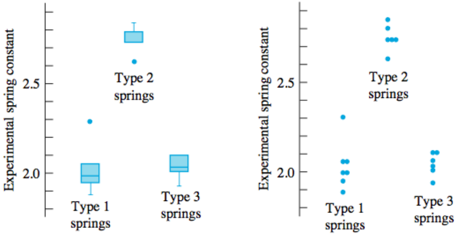
Example: springs

- Spring constants of three types of steel springs:



Empirical Spring Constants

Type 1 Springs	Type 2 Springs	Type 3 Springs
1.99, 2.06, 1.99 1.94, 2.05, 1.88 2.30	2.85, 2.74, 2.74 2.63, 2.74, 2.80	2.10, 2.01, 1.93 2.02, 2.10, 2.05



Example: springs

- Doesn't make sense to regress exponential spring constant on spring type.
- Instead, we apply:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:

    - Y_{ij} is the exponential spring constant of spring type i spring number j.
    - μ_i is the true mean exponential spring constant of type i.
    - i is the formula index, $i = 1, 2, \ldots, 8$
    - j is the index of a specimen within the formula i group.

Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

Confidence interval for linear combination of means

Fitted values

- Similarly to before, $\widehat{y}_{ij}$ is the fitted value corresponding to y_{ij}. It represents an estimate of the true mean response at factor level i and sample unit j.
- We treat all sample units equally, letting;

  $$\widehat{y}_{ij} = \hat{\mu}_i = \overline{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

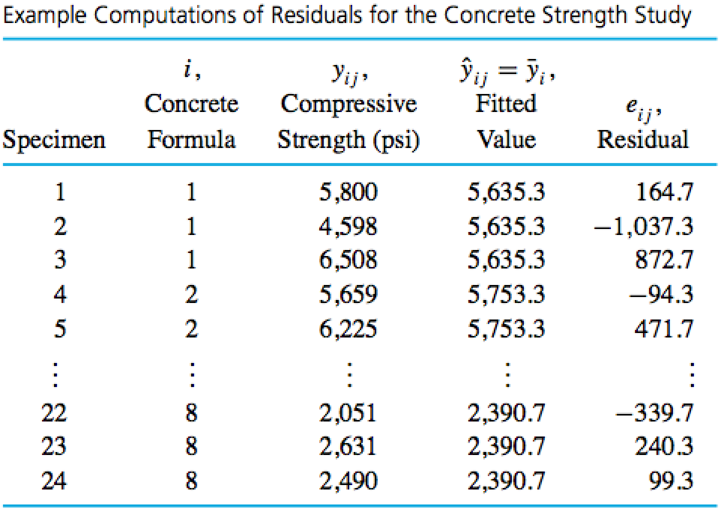
  the average of all the responses at factor level i.
- We get $\widehat{\mu}_i = \overline{y}_{i.}$ by minimizing the loss function:

  $$S(\mu_1, \mu_2, \ldots, \mu_r) = \sum_{ij} (y_{ij} - \mu_i)^2$$

  over all the choices of $\mu_1, \mu_2, \ldots, \mu_r$, selecting $\overline{y}_{i.}$ to estimate μ_i.
- The residuals e_{ij} are then:

  $$e_{ij} = y_{ij} - \overline{y}_{i.}$$

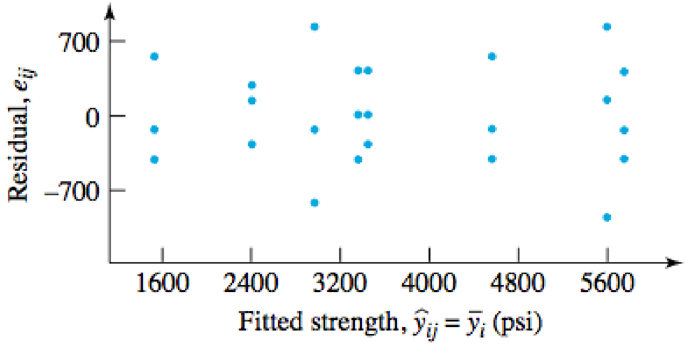
Example: concrete



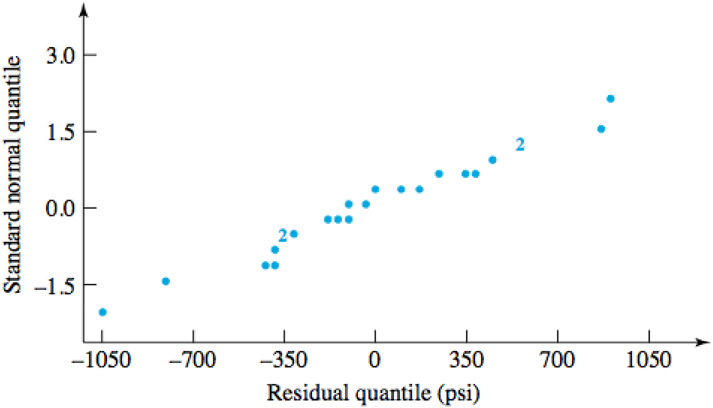
Example Computations of Residuals for the Concrete Strength Study

Specimen	i, Concrete Formula	y_{ij}, Compressive Strength (psi)	$\hat{y}_{ij} = \bar{y}_i$, Fitted Value	e_{ij}, Residual
1	1	5,800	5,635.3	164.7
2	1	4,598	5,635.3	−1,037.3
3	1	6,508	5,635.3	872.7
4	2	5,659	5,753.3	−94.3
5	2	6,225	5,753.3	471.7
⋮	⋮	⋮	⋮	⋮
22	8	2,051	2,390.7	−339.7
23	8	2,631	2,390.7	240.3
24	8	2,490	2,390.7	99.3

Example: concrete



Example: concrete



Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

Confidence interval for linear combination of means

Variance estimation

- We can compute a sample variance for each factor level:

$$s_i^2 = \frac{1}{n_i - 1} \sum_j (y_{ij} - \overline{y}_{ij})^2$$

- And we can compute a **pooled sample variance**:

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_r - 1)s_r^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_r - 1)}$$

- The pooled sample standard deviation is just $s_P = \sqrt{s_P^2}$

Variance estimation

- If $n = \sum_i n_i$, then:

$$
\begin{aligned}
s_P^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_r - 1)s_r^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_r - 1)} \\
&= \frac{(n_1 - 1)\left(\frac{1}{n_1 - 1}\right)\sum_j (y_{1j} - \overline{y}_1)^2 + \cdots + (n_r - 1)\left(\frac{1}{n_r - 1}\right)\sum_j (y_{lj} - \overline{y}_l)^2}{n - r} \\
&= \frac{1}{n - r}\sum_{ij} (y_{ij} - \overline{y}_i)^2 \\
&= \frac{1}{n - r}\sum_{ij} e_{ij}^2
\end{aligned}
$$

- As it turns out,

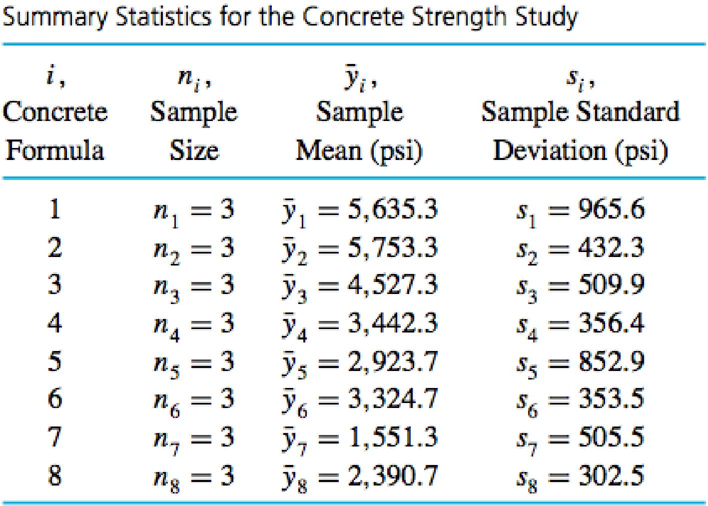
$$E(s_P^2) = \sigma^2$$

$$\frac{n - r}{\sigma^2} s_P^2 \sim \chi^2_{n-r}$$

- A $1 - \alpha$ confidence interval for σ^2 is of the form:

$$\left(\frac{n - r}{\chi^2_{n-r,\ 1-\alpha/2}} s_P^2,\ \frac{n - r}{\chi^2_{n-r,\ \alpha/2}} s_P^2 \right)$$

Example: concrete



Summary Statistics for the Concrete Strength Study

i, Concrete Formula	n_i, Sample Size	$\bar{y}_i$, Sample Mean (psi)	s_i, Sample Standard Deviation (psi)
1	$n_1 = 3$	$\bar{y}_1 = 5,635.3$	$s_1 = 965.6$
2	$n_2 = 3$	$\bar{y}_2 = 5,753.3$	$s_2 = 432.3$
3	$n_3 = 3$	$\bar{y}_3 = 4,527.3$	$s_3 = 509.9$
4	$n_4 = 3$	$\bar{y}_4 = 3,442.3$	$s_4 = 356.4$
5	$n_5 = 3$	$\bar{y}_5 = 2,923.7$	$s_5 = 852.9$
6	$n_6 = 3$	$\bar{y}_6 = 3,324.7$	$s_6 = 353.5$
7	$n_7 = 3$	$\bar{y}_7 = 1,551.3$	$s_7 = 505.5$
8	$n_8 = 3$	$\bar{y}_8 = 2,390.7$	$s_8 = 302.5$

Example: concrete

$$s_P^2 = \frac{(3-1)(965.6)^2 + (3-1)(432.3)^2 + \cdots + (3-1)(302.5)^2}{(3-1) + \cdots + (3-1)}$$

$$= 2\frac{965.6^2 + 432.3^2 + \cdots + 302.5^2}{16}$$

$$= 338213 \text{ psi}^2$$

$$s_P = \sqrt{338213} = 581.6psi$$

Example: concrete

- $n = 24$, $r = 8$, $n - r = 16$.
- $\chi^2_{16,\ 0.95} = 26.296$, $\chi^2_{16,\ 0.05} = 7.962$
- Hence, a 90% 2-sided confidence interval for σ^2 is:

$$\left(\frac{16 \cdot 581.6^2}{26.296},\ \frac{16 \cdot 581.6^2}{7.962}\right)$$
$$= (205816,\ 679745.9)$$

and you can make a 90% confidence interval for σ by transforming the endpoints of the confidence interval for σ^2:

$$(\sqrt{205816},\ \sqrt{679745.9}) = (453.7,\ 824.5)$$

- We're 90% confident that the true overall standard deviation of compressive strength of the concrete within factor levels is between 453.7 psi and 824.5 psi.

Outline

Standardized residuals

- Just as before, even though $\varepsilon_{ij} \sim$ iid $\mathrm{N}(0, \sigma^2)$, the e_{ij}'s don't have constant variance.
- The **standardized residuals** for the one-way ANOVA model are of the form:

$$e_{ij}^* = \frac{e_{ij}}{s_P\sqrt{\frac{n_i-1}{n_i}}}$$

which are approximately $N(0, 1)$.

Outline

Inference for the one-way ANOVA model

1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$, H_a : not all the μ_i's are equal.
2. α is some sensible value.
3. The test statistic is:

$$F = \frac{MSR}{MSE} = \frac{SSR/(r-1)}{SSE/(n-r)}$$

- Here,
  - n is the number of observations.
  - r is the number of levels of the covariate.
  - $SSR = \sum_{ij}(\widehat{y}_{ij} - \overline{y}_{..})^2 = \sum_{ij}(\overline{y}_{i.} - \overline{y}_{..})^2$
  - $SSE = \sum_{ij}(y_{ij} - \widehat{y}_{ij})^2 = \sum_{ij}(y_{ij} - \overline{y}_{i.})^2$
- Assume H_0 is true, the model is valid, and the ε_{ij}'s are iid $N(0, \sigma^2)$
- Then, $F \sim F_{r-1,\ n-r}$.
- Reject H_0 if $F > F_{r-1,\ n-r,\ 1-\alpha}$

Inference for the one-way ANOVA model

4. Compute the observed F using data. To do that, we can construct the ANOVA table:

Source	SS	df	MS	F
Covariate	SSR	$r-1$	$SSR/(r-1)$	MSR/MSE
Error	SSE	$n-l$	$SSE/(n-r)$	

5. If $observedF > F_{r-1,n-r,1-\alpha}$, reject H_0; or we can compute the p-value:

$$P(F_{r-1,n-r} > observedF)$$

   If the p-value is small, we reject H_0.

6. Conclusion in layman's term.

Example: concrete

1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_8$, H_a : not all the μ_i's are equal.
2. $\alpha = 0.05$
3. The test statistic is:

$$F = \frac{MSR}{MSE} = \frac{SSR/(r-1)}{SSE/(n-r)} = \frac{SSR/7}{SSE/16}$$

- Assume H_0 is true, the model is valid, and the ε_{ij}'s are iid $N(0, \sigma^2)$
- Then, $F \sim F_{r-1,\ n-r}$.
- Reject H_0 if $F > F_{r-1,\ n-r,\ 1-\alpha} = F_{7,16,0.95} = 2.66$

Example: concrete

4. We start by calculating SST, s_P^2, and SSE:

$$
\begin{aligned}
&(5{,}800 - 3{,}693.6)^2 + (4{,}598 - 3{,}693.6)^2 + (6{,}508 - 3{,}693.6)^2 \\
&\quad + \cdots + (2{,}631 - 3{,}693.6)^2 + (2{,}490 - 3{,}693.6)^2 \\
&= 52{,}772{,}190 \text{ (psi)}^2
\end{aligned}
$$

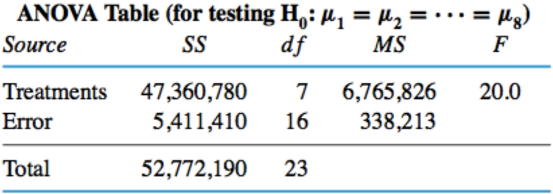
$$s_{\mathrm{P}}^2 = 338{,}213.1 \text{ (psi)}^2 \text{ and } n - r = 16, \text{ so}$$

$$SSE = (n - r)s_{\mathrm{P}}^2 = 5{,}411{,}410 \text{ (psi)}^2$$

Lastly, we calculate SSR:

$$\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 = 47{,}360{,}780$$

Example: concrete



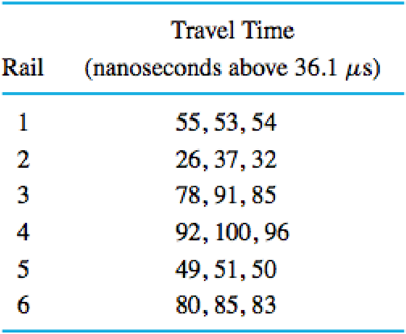
ANOVA Table (for testing $H_0: \mu_1 = \mu_2 = \cdots = \mu_8$)

Source	*SS*	*df*	*MS*	*F*
Treatments	47,360,780	7	6,765,826	20.0
Error	5,411,410	16	338,213	
Total	52,772,190	23		

5. With $observedF = 20.0 > 2.66$, we reject H_0 and conclude H_a.
6. There is enough evidence to conclude that the compressive strength of the concrete varies with formula.

Example: railroad rails

- The following data are taken from the paper *Zero- Force Travel-Time Parameters for Ultrasonic Head-Waves in Railroad Rail* by Bray and Leon- Salamanca (Materials Evaluation, 1985).
- Given are measurements in nanoseconds of the travel time (in excess of 36.1 μs) of a certain type of mechanical wave induced by mechanical stress in railroad rails.



Rail	Travel Time (nanoseconds above 36.1 μs)
1	55, 53, 54
2	26, 37, 32
3	78, 91, 85
4	92, 100, 96
5	49, 51, 50
6	80, 85, 83

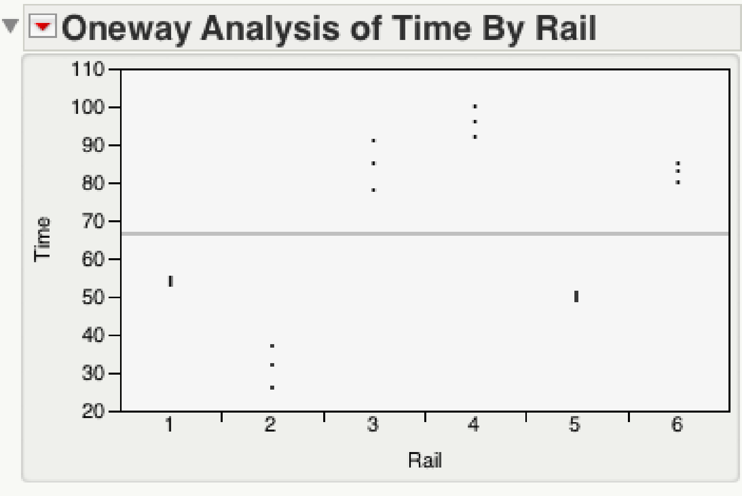
Example: railroad rails

- We apply the model:

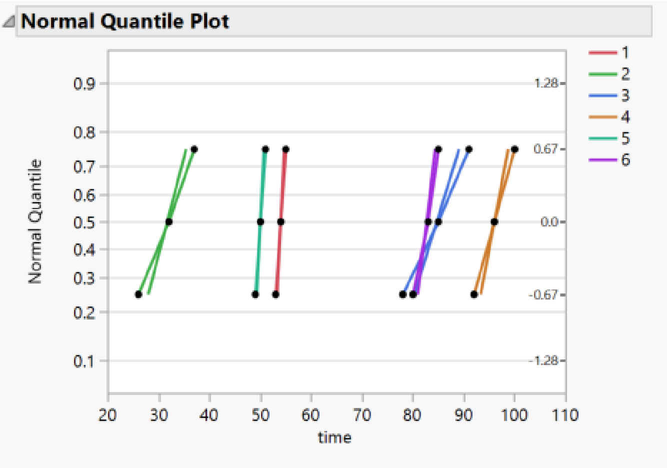
$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:
  - Y_{ij} is the observed travel time (ns) of the wave in excess of 26.1 μs for Rail i wave j.
  - μ_i is the true mean travel time (ns) in excess of 26.1 μs of waves through Rail i.

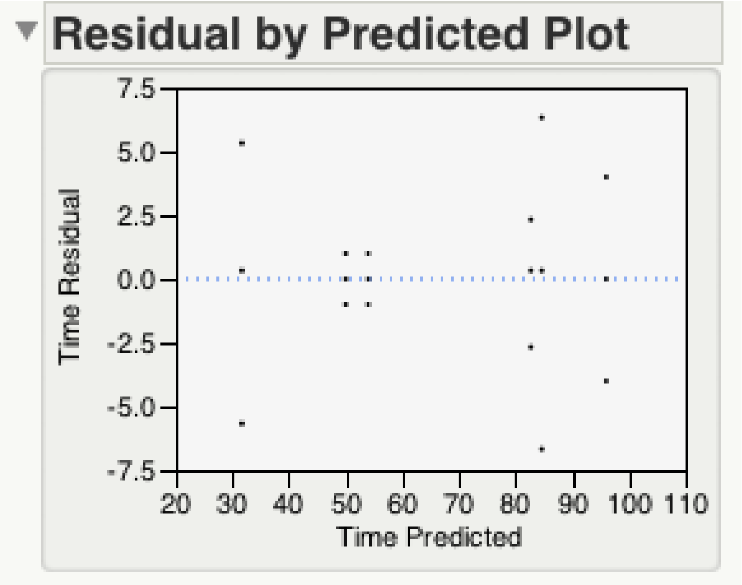
Example: railroad rails


Oneway Analysis of Time By Rail

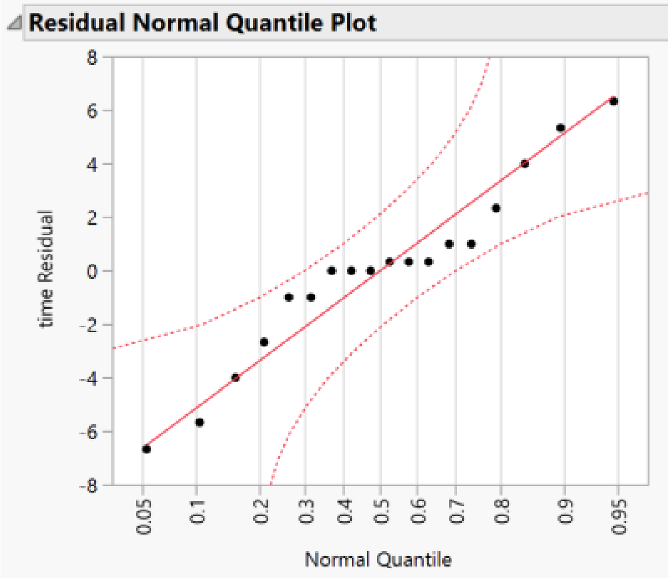
Example: railroad rails



Example: railroad rails



Example: railroad rails



Example: railroad rails

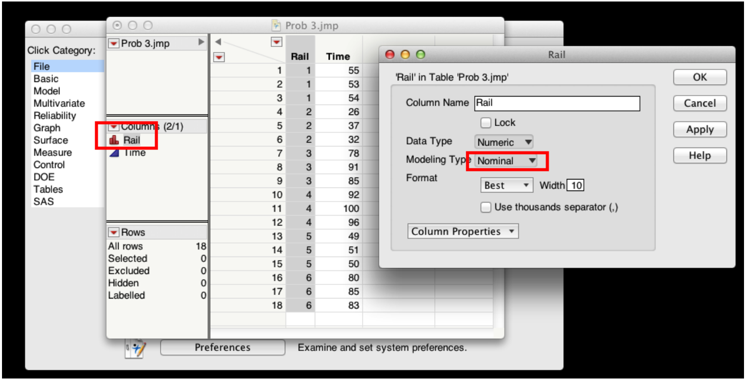
1. $H_0 : \mu_1 = \mu_2 = \cdots = \mu_6$, H_a : not all the μ_i's are equal.
2. $\alpha = 0.05$
3. The test statistic is:

$$F = \frac{MSR}{MSE} = \frac{SSR/(r-1)}{SSE/(n-r)} = \frac{SSR/(6-1)}{SSE/(18-6)} = \frac{SSR/5}{SSE/12}$$

- Assume H_0 is true, the model is valid, and the ε_{ij}'s are iid $N(0, \sigma^2)$
- Then, $F \sim F_{r-1,\ n-r}$.
- Reject H_0 if $F > F_{r-1,\ n-r,\ 1-\alpha} = F_{5,12,0.95} = 3.11$

Example: railroad rails

4. Load the data into JMP and fit travel time on rail, and *make sure the rail variable is a factor*.



	Rail	Time
1	1	55
2	1	53
3	1	54
4	2	26
5	2	37
6	2	32
7	3	78
8	3	91
9	3	85
10	4	92
11	4	100
12	4	96
13	5	49
14	5	51
15	5	50
16	6	80
17	6	85
18	6	83

All rows	18
Selected	0
Excluded	0
Hidden	0
Labelled	0

Rail

'Rail' in Table 'Prob 3.jmp'

Column Name Rail

☐ Lock

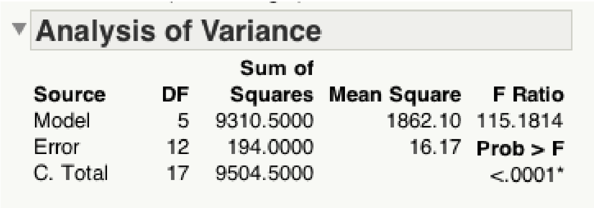
Data Type Numeric

Modeling Type Nominal

Format Best Width 10

☐ Use thousands separator (,)

Example: railroad rails



Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	5	9310.5000	1862.10	115.1814
Error	12	194.0000	16.17	**Prob > F**
C. Total	17	9504.5000		<.0001*

5. With $observedF = 115.18 > 3.11$, we reject H_0 and conclude H_a.
6. There is enough evidence to conclude that the true mean excess travel time of waves along the rails depends on the rail.

Outline

The one-way ANOVA model

Residuals and fitted values

Variance estimation

Standardized residuals

Inference

Confidence interval for linear combination of means

Confidence interval for linear combination of means

- When we have multiple samples with means $\mu_1, \mu_2, \dots, \mu_r$, we want to compare these means
- There are many possibilities: μ_1, $\mu_1 - \mu_2$, $\mu_1 - \mu_2$, $\frac{1}{2}(\mu_1 + \mu_2) - \mu_3 \dots$.
- We will construct a confidence interval for a linear combination of these means. We denote the linear combination as

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r$$

Confidence interval for linear combination of means

- Since the estimate for μ_i is $\bar{y}_{i\cdot}$, the estimate of L is

$$\hat{L} = c_1\bar{y}_{1\cdot} + c_2\bar{y}_{2\cdot} + \cdots + c_r\bar{y}_{r\cdot}$$

- Under the one-way normal model, $\bar{y}_{i\cdot} \sim N(\mu_i, \sigma^2/n_i)$. So

$$\hat{L} \sim N\Bigg(c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r, \\ \sigma^2\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}\right)\Bigg)$$

Confidence interval for linear combination of means

- Thus

$$\frac{\hat{L} - L}{\sigma\sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}} \sim N(0, 1)$$

- Replacing the unknown σ with the pooled sample standard deviation, then

$$\frac{\hat{L} - L}{s_P\sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}} \sim t_{n-r}$$

- Therefore a two-sided $1 - \alpha$ confidence interval is

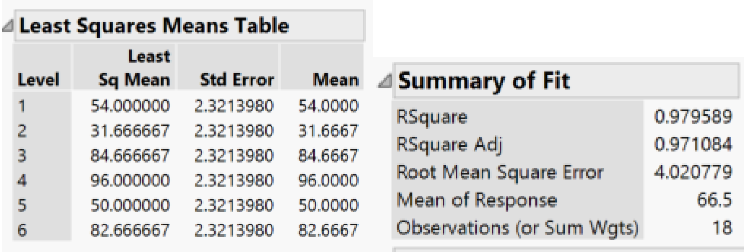
$$\hat{L} \pm t_{n-r,1-\alpha/2} s_P \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}$$

The one-sidede confidence intervals are analogous.

Example: railroad rails

Construct the following confidence intervals with the JMP output:

1. 95% two-sided confidence interval for μ_1
2. 95% two-sided confidence interval for $\mu_1 - \mu_2$
3. 95% lower confidence bound for $\mu_3 - \mu_5$
4. 90% two-sided confidence interval for $\frac{1}{2}(\mu_1 + \mu_2) - \mu_3$



Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
1	54.000000	2.3213980	54.0000
2	31.666667	2.3213980	31.6667
3	84.666667	2.3213980	84.6667
4	96.000000	2.3213980	96.0000
5	50.000000	2.3213980	50.0000
6	82.666667	2.3213980	82.6667

Summary of Fit

RSquare	0.979589
RSquare Adj	0.971084
Root Mean Square Error	4.020779
Mean of Response	66.5
Observations (or Sum Wgts)	18

Example: railroad rails

1. $t_{n-r,1-\alpha/2} = t_{18-6,1-0.05/2} = t_{12,0.975} = 2.179$. So

$$\begin{aligned} &\bar{y}_{1\cdot} \pm t_{12,0.975} s_P \sqrt{\frac{1^2}{3}} \\ =&54.000 \pm 2.179(4.021)\sqrt{1/3} \\ =&54 \pm 5.059 \\ =&(48.941, 59.059) \end{aligned}$$

We are 95% confident that the mean mechanical wave travel time for Rail 1 is any number between 48.941 and 59.059 nanoseconds.

Example: railroad rails

2. $t_{n-r,1-\alpha/2} = t_{18-6,1-0.05/2} = t_{12,0.975} = 2.179$. So

$$\begin{aligned}&\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm t_{12,0.975} s_P \sqrt{\frac{1^2}{3} + \frac{(-1)^2}{3}}\\ =&(54.000 - 31.667) \pm 2.179(4.021)\sqrt{2/3}\\ =&22.333 \pm 7.154\\ =&(15.179, 29.487)\end{aligned}$$

We are 95% confident that the mean mechanical wave travel time for Rail 1 is longer than that of Rail 2 by any number between 15.179 and 20.487 nanoseconds.

Example: railroad rails

3. $t_{n-r,1-\alpha} = t_{18-6,1-0.05} = t_{12,0.95} = 1.782$. So the lower 95% confidence bound is

$$\begin{aligned}
&\bar{y}_{3\cdot} - \bar{y}_{5\cdot} - t_{12,0.95} s_P \sqrt{\frac{1^2}{3} + \frac{(-1)^2}{3}} \\
=&(84.667 - 50.000) \pm 1.782(4.021)\sqrt{2/3} \\
=&34.667 - 5.851 \\
=&28.816
\end{aligned}$$

We are 95% confident that the mean mechanical wave travel time for Rail 3 is longer than that of Rail 5 by at least 28.816 nanoseconds.

Example: railroad rails

4. $t_{n-r,1-\alpha/2} = t_{18-6,1-0.1/2} = t_{12,0.95} = 1.782$. So

$$
\begin{aligned}
&\frac{1}{2}\bar{y}_{1\cdot} + \frac{1}{2}\bar{y}_{2\cdot} - \bar{y}_{3\cdot} \pm t_{12,0.95} s_P \sqrt{\frac{(1/2)^2}{3} + \frac{(1/2)^2}{3} + \frac{(-1)^2}{3}} \\
=&(\frac{1}{2}54.000 + \frac{1}{2}31.667 - 84.667) \pm 1.782(4.021)\sqrt{1/2} \\
=&-41.834 \pm 5.067 \\
=&(-46.901, -36.767)
\end{aligned}
$$

We are 90% confident that the average of the mean mechanical wave travel times for Rail 1 and Rail 2 is shorter than the mean mechanical wave travel time of Rail 3 by any number between 46.901 and 36.767 nanoseconds.