# Applied Statistics for Life Sciences

Dmitri D. Pervouchine

Centre de Regulació Genòmica

## Module 5: Statistical Modelling, Regression and ANOVA

# Contents

1. Regression and ANOVA
    - Least Squares
    - Residual plot
    - Pearson Correlation Coefficient
    - Coefficient of Determination
    - Confidence and Prediction Intervals

2. ANOVA
    - One-way ANOVA
    - Two-way ANOVA

3. Summary

## The Least Squares

- A simple data set consists of data pairs $(x_i, y_i)$, $i = 1, \ldots, n$, where $x_i$ is called *independent variable* and $y_i$ is called *dependent variable*

- We are looking for the model function of the form $y = a + bx$ such that it gives "best" fit to the data

- "best" in what sense?

## Residuals

- A **residual** $r_i$ is defined as the difference between the values of the dependent variable and the predicted values from the estimated model

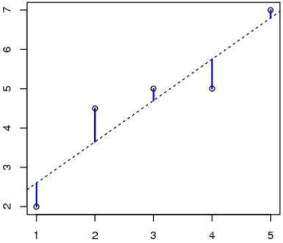
$$r_i = y_i - \hat{y}_i, \text{ where } \hat{y}_i = a + bx_i$$

- The least squares method defines "best" model as when

$$S = \sum_{i=1}^{n} r_i^2$$

is at minimum

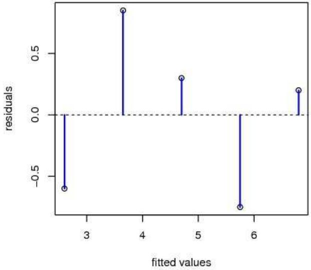
# Regression Line



- Residuals are shown in blue
- Residuals are positive for data points above the line
- Residuals are negative for data points below the line
- Sum of squares of residuals is at minimum

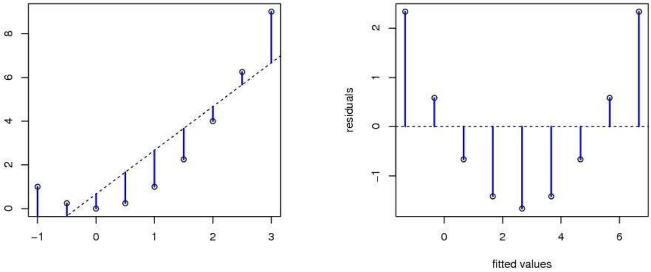
# Residual plot

The residual plot is the scatterplot of residuals vs. fitted values, i.e., $y_i - \hat{y}_i \sim \hat{y}_i$



- The sum of the residuals w.r.t least square line is equal to zero

# Residual plot



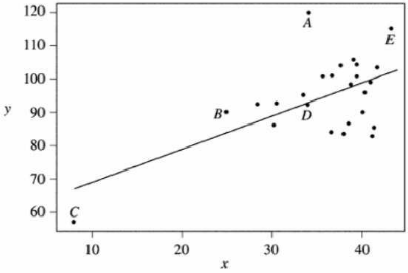
A pattern in the residual plot indicates that a non-linear model should be used

## Influential Scores and Outliers

- In regression, an **outlier** is a data point with large residual
- An **influential score** is the data point which significantly influences the regression line
- If an influential score is removed from the sample, the regression line will change significantly

### Problem 1.1

*Which of the five points is an outlier, and which is an influential score?*



### Solution

*Correct: (A) is an outlier; (C) is an influential score*

# Solving the Regression

$$S = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \to \min$$

$$\begin{cases} \frac{\partial S}{\partial a} = -2\sum\limits_{i=1}^{n}(y_i - (a + bx_i)) = 0 \\ \frac{\partial S}{\partial b} = -2\sum\limits_{i=1}^{n} x_i(y_i - (a + bx_i)) = 0 \end{cases}$$

$$\begin{cases} \sum\limits_{i=1}^{n} y_i = an + b\sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i y_i = a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 \end{cases}$$

If $\sum\limits_{i=1}^{n} x_i = 0$ then $b = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. In general, $b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$

# Regression Slope and Intercept

$$b = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{s_{xy}}{s_{xx}},$$

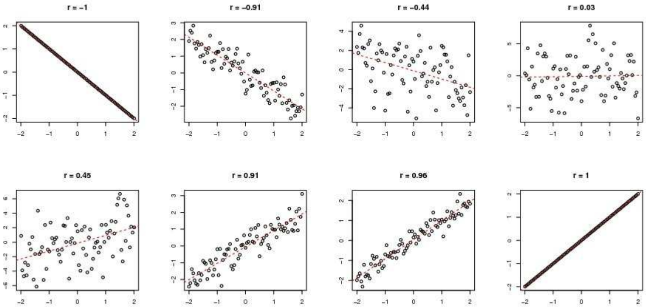
where

$$s_{xy} = \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y}),$$

$$s_{xx} = \frac{1}{n-1}\sum(x_i - \overline{x})(x_i - \overline{x}),$$

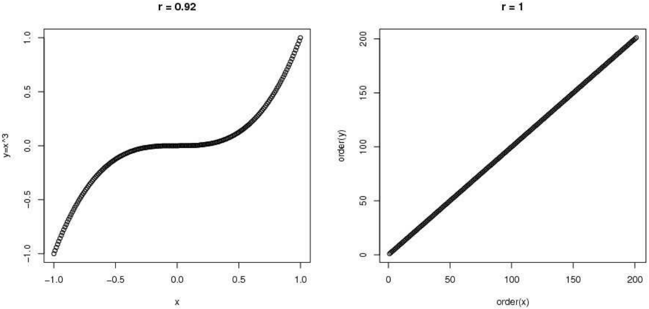
$$a = \overline{y} - b\overline{x}$$

# Pearson Correlation Coefficient



- The Pearson correlation coefficient $r$ indicates the degree of *linear dependence*
- $r \in [-1, 1]$
- $r$ and the regression slope have the same sign
- Regression slope is *not* determined by the value of $r$
- Variables with zero correlation are *uncorrelated* but not necessarily independent

# Spearman Correlation Coefficient



- Spearman correlation coefficient $r_s$ is equal to the Pearson correlation of ranks
- $r_s \in [-1, 1]$
- $r_s$ is sensitive to the order of observations, not their absolute value
- $r_s$ indicates the degree of *monotonous*, not necessarily linear dependence
- Unlike Pearson correlation coefficient, $r_s$ is not sensitive to outliers or influential scores

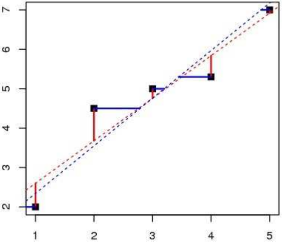
# Correlation and Regression Slope

- $r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{s_{xy}}{s_x s_y}$

- $s_{xy} = \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y})$

- $b = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$

# Regression $y$ vs. $x$ and $x$ vs. $y$



- Residuals in $y$ vs. $x$ and in $x$ vs. $y$ are different
- $y = a + bx \Leftrightarrow x = -\frac{a}{b} + \frac{1}{b}y$, the product of slopes of inverse lines is 1
- $b = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$, $b^* = \frac{s_{xy}}{s_y^2} = r\frac{s_x}{s_y}$
- $b \cdot b^* = r^2$, i.e., the product of slopes of inverse regression lines is $r^2 \leq 1$

## Coefficient of Determination

$$R^2 = r^2 = \frac{SSX}{SST}$$

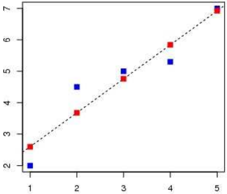
- SST = total sum of squares
- SSX = sum of squares explained by X
- SSE = sum of squares of residuals
- SST = SSX+SSE

- The square of the sample correlation coefficient, which is also known as the *coefficient of determination*, is the fraction of the variance in $y$ that is accounted for by a linear fit of $x$.

# Decomposition of Sums of Squares

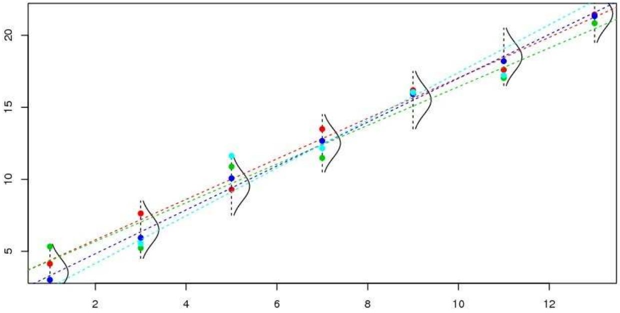
$$(n-1)s_Y^2 = \sum(y_i - \overline{y})^2 = \sum\left((y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\right)^2 =$$
$$\sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2 + 2 \cdot \text{cross-product} = SSE + SSX$$

$$\text{cross-product} = \sum(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum r_i(a + bx_i - a - b\overline{x}) = b\sum r_i(x_i - \overline{x}) = 0$$



$y_i$ vs. $\hat{y}_i$

# From Least Squares to Statistics



- The regression line is a result of random sampling
- Different samples produce different lines
- There is a family of lines for given population; you see just one

## Linear Regression Model

The model postulates that $y_i = \alpha + \beta x_i + e_i$, where

- $\alpha$ and $\beta$ are unknown parameters
- $x_i$ are non-random
- $e_i$ and, consequently, $y_i$ are random, where
  - $e_i \sim \mathcal{N}(0, \sigma_e^2)$
  - $\sigma_e$ is the same for all $i$ (**homoscedasticity**)
  - $e_i$ and $e_j$ are independent for $i \neq j$

# SE of the Regression Slope

- $\hat{\alpha} = a$ and $\hat{\beta} = b$ from LS are unbiased effective estimators of $\alpha$ and $\beta$

- $\mathrm{SE}(\hat{\beta}) = \frac{\sigma_e}{\sqrt{\sum(x_i - \overline{x})^2}}$

- $\hat{\sigma}_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$

## Problem 1.2

*Growth hormones are often used to increase the weight gain in chickens. In an experiment using 15 chickens, five different doses of growth hormone were injected into chickens (three for each dose) and the subsequent weight gain was recorded. An experimenter plots the data and finds that a linear relationship appears to hold. The output of the software is*

| Term | *Extimate* | *Std Error* | *t-ratio* | *P* |
|---|---|---|---|---|
| *Intercept* | *4.5458533* | *0.616518* | *7.37* | *0.0001* |
| *dose* | *4.83233426* | *1.016403* | *4.75* | *0.0004* |

- *What is the equation for the fitted line?*
- *Find an approximate 95% confidence interval for the regression slope?*
- *Test the hypothesis that the slope is non-zero?*

## Solution

- $gain = 4.55 + 4.83 * dose$
- $SE(\hat{\beta}) = 1.016$, $\beta = \hat{\beta} \pm t_{0.025}(13) SE(\hat{\beta}) = 4.83 \pm 2.16 * 1.016 = [2.64; 7.02]$
- $$H_0 : \beta = 0$$ $$H_a : \beta > 0$$

  $t = \frac{4.83-0}{1.016} = 4.75$, $\mathrm{P}(t(13) > 4.75) = 0.0002$, *i.e.,* $H_0$ *is rejected at the 1% significance level. That is, the weight gain in chicken significantly depends on the dose of the growth hormone*

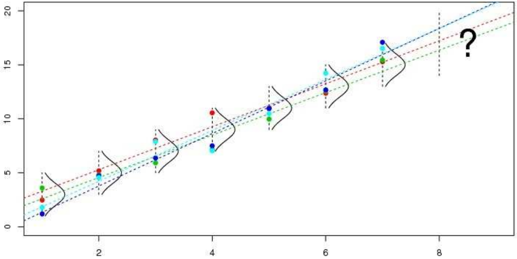
## Problem 1.3

*A marine biologist wants to test the effect of water temperature on the average dive duration for sea otters. Seven otters are available for the study. The biologist collects the data with the following summary statistics.* $\sum X = 80$, $\sum Y = 639$, $\sum X^2 = 1088$, $\sum Y^2 = 60457$, $\sum XY = 7888$. *Find the regression line and a 95% confidence interval for the regression slope.*

## Solution

- $\overline{X} = \frac{80}{7} = 11.4$, $\overline{Y} = \frac{639}{7} = 91.3$, $s_x^2 = \frac{1088 - 7 \cdot 11.4^2}{6} = 29.7$, $s_y^2 = \frac{60457 - 7 \cdot 91.3^2}{6} = 351.2$, $s_{xy} = \frac{7888 - 7 \cdot 11.4 \cdot 91.3}{6} = 100.4$
- $b = \frac{s_{xy}}{s_X^2} = \frac{100.4}{29.7} = 3.38$, $a = \overline{Y} - b\overline{X} = 91.3 - 3.38 * 11.4 = 52.77$
- $\texttt{duration} = 52.77 + 3.38\texttt{temp}$
- $r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{100.4}{\sqrt{29.7 * 351.2}} = 0.98$
- $SSY = (n-1)s_y^2 = 6 * 351.2$, $SSE = (1 - R^2)SSY = (1 - 0.98^2) * 6 * 351.2 = 83.44$, $SE(\hat{\beta}) = \frac{\sqrt{83.44/(7-2)}}{\sqrt{6 * 29.7}} = 0.30$
- $\beta = \hat{\beta} \pm t_{0.025}(5) SE(\hat{\beta}) = 3.38 \pm 2.57 * 0.30 = 3.38 \pm 0.77$, *i.e., we are 95% confident that the dive duration increases by on average* $3.38 \pm 0.77$ *minutes with each additional Celcius degree of water*

## Confidence vs. Prediction Interval



- Suppose I fuel my car 7 days a week, from Sunday to Sunday, each day at a randomly chosen gas station. I get a sample of gasoline prices for 7 days.
- *Confidence interval* is for the average gasoline price on Monday
- *Prediction interval* is for a gasoline price at a randomly chosen gas station on Monday

# Confidence vs. Prediction Interval

- Confidence Interval

$$\hat{y}_0 = \alpha + \beta x_0 = a + bx_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}_e\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2}}$$

- Prediction Interval

$$\hat{y}_0 = \alpha + \beta x_0 = a + bx_0 \pm t_{\alpha_2}(n-2)\hat{\sigma}_e\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2} + 1}$$

## Problem 1.4 (problem 1.3 revisited)

*A marine biologist wants to test the effect of water temperature on the average dive duration for sea otters. Seven otters are available for the study. The biologist collects the data with the following summary statistics.* $\sum X = 80$, $\sum Y = 639$, $\sum X^2 = 1088$, $\sum Y^2 = 60457$, $\sum XY = 7888$. *Find the 95% confidence and prediction intervals for the dive duration at 25 degrees Celsius.*

## Solution

- $\overline{X} = \frac{80}{7} = 11.4$, $\overline{Y} = \frac{639}{7} = 91.3$, $s_x^2 = \frac{1088-7\cdot 11.4^2}{6} = 29.7$, $s_y^2 = \frac{60457-7\cdot 91.3^2}{6} = 351.2$, $s_{xy} = \frac{7888-7\cdot 11.4\cdot 91.3}{6} = 100.4$
- $b = \frac{s_{xy}}{s_X^2} = \frac{100.4}{29.7} = 3.38$, $a = \overline{Y} - b\overline{X} = 91.3 - 3.38 * 11.4 = 52.77$
- $SSE = (1-R^2)SSY = (1-0.98^2) * 6 * 351.2 = 83.44$
- $\hat{\sigma} = \sqrt{MSE} = \sqrt{83.44/(7-2)} = 4.8$, $t_{0.025}(5) = 2.57$
- *Confidence* $\hat{y}_0 = 52.77 + 3.38 \cdot 25 \pm 2.57 \cdot 4.8\sqrt{\frac{1}{7} + \frac{(25-11.4)^2}{6*29.7^2}} = 137.3 \pm 5.2$ *min, i.e., we are 95% confident that the average dive duration at 25°C is* $137.3 \pm 5.2$ *minutes.*
- *Prediction* $\hat{y}_0 = 52.77 + 3.38 \cdot 25 \pm 2.57 \cdot 4.8\sqrt{\frac{1}{7} + \frac{(25-11.4)^2}{6*29.7^2} + 1} = 137.3 \pm 13.4$ *min, i.e., we are 95% confident that when a sea otter dives at 25°C next time, the duration will be* $137.3 \pm 13.4$ *minutes*

## ANOVA: Analysis of Variance

- A collection of models, in which the variance of the observed set is partitioned into components due to explanatory variables
- Assumptions:
  - Independence of observations
  - The distributions in each of the groups are normal
  - Variance homogeneity, called homoscedasticity: the variance of data in groups should be the same.

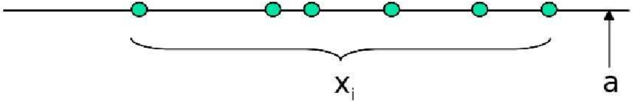
# ANOVA: Analysis of Variance

A manager wishes to determine whether the mean times required to complete a certain task differ for the three levels of employee training. He randomly selected 10 employees with each of the three levels of training.

| Level | $n$ | $\overline{x}$ | $s^2$ |
|---|---|---|---|
| Advanced | 10 | 24.2 | 21.54 |
| Intermediate | 10 | 27.1 | 18.64 |
| Beginner | 10 | 30.2 | 17.76 |

Do the data provide sufficient evidence to indicate that the mean times required to complete a certain task differ for at least two of the three levels of training?

# Steiner's Theorem



$$I(x_1, x_2, \ldots, x_n; a) = \sum_{i=1}^{n} (x_i - a)^2 = \text{Moment of inertia}$$

$$I(x_1, x_2, \ldots, x_n; a) = (x_1, x_2, \ldots, x_n; \overline{x}) + n(\overline{x} - a)^2$$

# One-way ANOVA example

Three different milling machines were being considered for purchase by a manufacturer. Potentially, the company would be purchasing hundreds of these machines, so it wanted to make sure it made the best decision. Initially, five of each machine were borrowed, and each was randomly assigned to one of 15 technicians (all technicians were similar in skill). Each machine was put through a series of tasks and rated using a standardized test. The higher the score on the test, the better the performance of the machine. The data are:

| Machine 1 | Machine 2 | Machine 3 |
|---|---|---|
| 24.50 | 28.40 | 26.10 |
| 23.50 | 34.20 | 28.30 |
| 26.40 | 29.50 | 24.30 |
| 27.10 | 32.20 | 26.20 |
| 29.90 | 30.10 | 27.80 |

## Decomposition of Sum of Squares

- SST = SSA + SSE
- SST = total sum of squares
- SSA = sum of squares for factor A
- SSE = sum of squares of errors

# Decomposition of Sum of Squares

### Observed

| | M1 | M2 | M3 | mean |
|---|---|---|---|---|
| | 24.50 | 28.40 | 26.10 | |
| | 23.50 | 34.20 | 28.30 | |
| | 26.40 | 29.50 | 24.30 | |
| | 27.10 | 32.20 | 26.20 | |
| | 29.90 | 30.10 | 27.80 | |
| Mean | 26.28 | 30.88 | 26.54 | 27.90 |

### Expected

| | M1 | M2 | M3 | mean |
|---|---|---|---|---|
| | 26.28 | 30.88 | 26.54 | |
| | 26.28 | 30.88 | 26.54 | |
| | 26.28 | 30.88 | 26.54 | |
| | 26.28 | 30.88 | 26.54 | |
| | 26.28 | 30.88 | 26.54 | |
| Mean | 26.28 | 30.88 | 26.54 | 27.90 |

$$(24.50 - 27.90)^2 + (23.50 - 27.90)^2 + \cdots + (29.90 - 27.90)^2 + \cdots =$$

$$(24.50 - 26.28)^2 + (23.50 - 26.28)^2 + \cdots + (29.90 - 26.28)^2 + 5 * (26.28 - 27.90)^2 \cdots = SSE + SSA$$

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \overline{x}_{\bullet j})^2$$

$$SSA = n \sum_{j=1}^{m} (\overline{x}_{\bullet j} - \overline{x}_{\bullet\bullet})^2$$

$$SSE = SST - SSA$$

# Decomposition of Sum of Squares

| | 1 | 2 | 3 | |
|---|---|---|---|---|
| | $x_{11}$ | $x_{12}$ | $x_{13}$ | |
| | $x_{21}$ | $x_{22}$ | $x_{23}$ | |
| | $x_{31}$ | $x_{32}$ | $x_{33}$ | |
| | $x_{41}$ | $x_{42}$ | $x_{43}$ | |
| | $x_{51}$ | $x_{52}$ | $x_{53}$ | |
| *mean* | $x_{\bullet 1}$ | $x_{\bullet 2}$ | $x_{\bullet 3}$ | $x_{\bullet\bullet}$ |

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{x}_{\bullet\bullet})^2 = \sum_{j=1}^{m}\left(\sum_{i=1}^{n}(x_{ij} - \overline{x}_{\bullet j})^2 + n(\overline{x}_{\bullet j} - \overline{x}_{\bullet\bullet})^2\right) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{x}_{\bullet j})^2 + n\sum_{j=1}^{m}(\overline{x}_{\bullet j} - \overline{x}_{\bullet\bullet})^2 = SSE + SSA$$

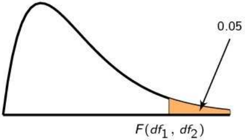
**Assumption:** $x_{ij} - \overline{x}_{\bullet j} \sim \mathcal{N}(0, \sigma^2)$ **are independent**$^*$

## One-way ANOVA table

- SST = SSA + SSE = Total sum of squares
- SSA = Sum of squares Factor A
- SSE = Sum of squares Error
- MSA = Mean sum of squares Factor
- MSE = Mean sum of squares Error

| | SS | df | MS | F | P-value |
|---|---|---|---|---|---|
| Factor | SSA | k-1 | SSA/(k-1) | MSA/MSE | $P(F > \dots)$ |
| Error | SSE | n-k | SSE/(N-k) | | |
| Total | SST | n-1 | | | |

# Fisher $F$-distribution



| $df_2$ | $df_1 = 1$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 161.4476 | 199.5000 | 215.7073 | 224.5832 | 230.1619 | 233.9860 | 236.7684 | 238.8827 | 240.5433 | 241.8817 |
| **2** | 18.5128 | 19.0000 | 19.1643 | 19.2468 | 19.2964 | 19.3295 | 19.3532 | 19.3710 | 19.3848 | 19.3959 |
| **3** | 10.1280 | 9.5521 | 9.2766 | 9.1172 | 9.0135 | 8.9406 | 8.8867 | 8.8452 | 8.8123 | 8.7855 |
| **4** | 7.7086 | 6.9443 | 6.5914 | 6.3882 | 6.2561 | 6.1631 | 6.0942 | 6.0410 | 5.9988 | 5.9644 |
| **5** | 6.6079 | 5.7861 | 5.4095 | 5.1922 | 5.0503 | 4.9503 | 4.8759 | 4.8183 | 4.7725 | 4.7351 |
| **6** | 5.9874 | 5.1433 | 4.7571 | 4.5337 | 4.3874 | 4.2839 | 4.2067 | 4.1468 | 4.0990 | 4.0600 |
| **7** | 5.5914 | 4.7374 | 4.3468 | 4.1203 | 3.9715 | 3.8660 | 3.7870 | 3.7257 | 3.6767 | 3.6365 |
| **8** | 5.3177 | 4.4590 | 4.0662 | 3.8379 | 3.6875 | 3.5806 | 3.5005 | 3.4381 | 3.3881 | 3.3472 |
| **9** | 5.1174 | 4.2565 | 3.8625 | 3.6331 | 3.4817 | 3.3738 | 3.2927 | 3.2296 | 3.1789 | 3.1373 |
| **10** | 4.9646 | 4.1028 | 3.7083 | 3.4780 | 3.3258 | 3.2172 | 3.1355 | 3.0717 | 3.0204 | 2.9782 |

$$\mathrm{P}\left(F(df_1, df_2) < x\right) = \mathrm{P}\left(\frac{1}{F(df_1, df_2)} > \frac{1}{x}\right) = \mathrm{P}\left(F(df_2, df_1) > \frac{1}{x}\right)$$

## Solution to the example

| Source of Variation | SS | df | MS | F | P-value |
|---|---|---|---|---|---|
| Between Groups | 66.77 | 2 | 33.39 | 7.14 | 0.009073 |
| Within Groups | 56.13 | 12 | 4.68 | | |
| Total | 122.9 | 14 | | | |

Conclusion: $H_0$ is rejected at 5% significance level, i.e., there is enough evidence to suspect that machines are different.

## Problem 2.1

*Some varieties of nematodes feed on the roots of lawn grasses and crops such as strawberries and tomatoes. Four brands of nematocides are to be compared. Twelve plots of land of comparable fertility that were suffering from nematodes were planted with a crop. The yields of each plot were recorded and part of the ANOVA table appears below:*

| Source of Variation | SS | df | MS | F | P-value |
|---|---|---|---|---|---|
| Nematocides | 3.456 | * | * | * | * |
| Error | 1.200 | 8 | * | | |
| Total | 4.656 | 11 | | | |

*Find the value of F and* P-value.

## Solution

| Source of Var | SS | df | MS | F | P-value |
|---|---|---|---|---|---|
| Nematocides | 3.456 | 11-8=3 | $\frac{3.4456}{3} = 1.152$ | $\frac{1.152}{0.15} = 7.68$ | $\mathrm{P}(F(3,8) > 7.68) = 0.009$ |
| Error | 1.200 | 8 | $\frac{1.2}{8} = 0.15$ | | |
| Total | 4.656 | 11 | | | |

# Two-way ANOVA

- **One-way ANOVA** Group A is given vodka, Group B is given gin, and Group C is given a placebo. Groups are tested with a memory task.
- **Two-way ANOVA** In an experiment testing the effects of expectations, subjects are randomly assigned to four groups:
  - expect vodka — receive vodka
  - expect vodka — receive placebo
  - expect placebo — receive vodka
  - expect placebo — receive placebo

  Each group is then tested on a memory task.

## Decomposition of Sum of Squares

- $\mathsf{SST} = \mathsf{SSA} + \mathsf{SSB} + \mathsf{SSE}$
- SST = total sum of squares
- SSA = sum of squares for factor A
- SSB = sum of squares for factor B
- SSE = sum of squares of errors

# Decomposition of Sum of Squares

|  |  |  |  | *mean* |
|---|---|---|---|---|
|  | $x_{11}$ | $x_{12}$ | $x_{13}$ | $x_{1\bullet}$ |
|  | $x_{21}$ | $x_{22}$ | $x_{23}$ | $x_{2\bullet}$ |
|  | $x_{31}$ | $x_{32}$ | $x_{33}$ | $x_{3\bullet}$ |
|  | $x_{41}$ | $x_{42}$ | $x_{43}$ | $x_{4\bullet}$ |
|  | $x_{51}$ | $x_{52}$ | $x_{53}$ | $x_{5\bullet}$ |
| *mean* | $x_{\bullet 1}$ | $x_{\bullet 2}$ | $x_{\bullet 3}$ | $x_{\bullet\bullet}$ |

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \overline{x}_{\bullet\bullet})^2 = \sum_i\sum_j(x_{ij} - \overline{x}_{i\bullet})^2 + m\sum_i(\overline{x}_{i\bullet} - \overline{x}_{\bullet\bullet})^2 =$$

$$\sum_i\sum_j\left((x_{ij} - \overline{x}_{i\bullet} - \overline{x}_{\bullet j} + \overline{x}_{\bullet\bullet}) + (\overline{x}_{\bullet j} - \overline{x}_{\bullet\bullet})\right)^2 + SSA = \sum_i\sum_j(x_{ij} - \overline{x}_{i\bullet} - \overline{x}_{\bullet j} + \overline{x}_{\bullet\bullet})^2 +$$

$$n\sum_j(\overline{x}_{\bullet j} - \overline{x}_{\bullet\bullet})^2 + SSA = \sum_i\sum_j(x_{ij} - \overline{x}_{i\bullet} - \overline{x}_{\bullet j} + \overline{x}_{\bullet\bullet})^2 + SSB + SSA = SSA + SSB + SSE$$

**Assumption:** $x_{ij} - \overline{x}_{i\bullet} - \overline{x}_{\bullet j} + \overline{x}_{\bullet\bullet} \sim \mathcal{N}(0, \sigma^2)$ **are independent**$^*$

# Two-way ANOVA example

Three different milling machines were being considered for purchase by a manufacturer. Potentially, the company would be purchasing hundreds of these machines, so it wanted to make sure it made the best decision. Initially, five of each machine were borrowed. *Machines are operated by 5 different crew technicians:*

| | Machine 1 | Machine 2 | Machine 3 |
|---|---|---|---|
| Crew 1 | 24.50 | 28.40 | 26.10 |
| Crew 2 | 23.50 | 34.20 | 28.30 |
| Crew 3 | 26.40 | 29.50 | 24.30 |
| Crew 4 | 27.10 | 32.20 | 26.20 |
| Crew 5 | 29.90 | 30.10 | 27.80 |

# What is the Error Term?

Observed

|  | M1 | M2 | M3 | mean |
|---|---|---|---|---|
| Crew 1 | 24.50 | 28.40 | 26.10 | 26.30 |
| Crew 2 | 23.50 | 34.20 | 28.30 | 28.70 |
| Crew 3 | 26.40 | 29.50 | 24.30 | 26.70 |
| Crew 4 | 27.10 | 32.20 | 26.20 | 28.50 |
| Crew 5 | 29.90 | 30.10 | 27.80 | 29.30 |
| mean | 26.28 | 30.88 | 26.54 | 27.90 |

Expected

|  | M1 | M2 | M3 | mean |
|---|---|---|---|---|
| Crew 1 | 24.70 | 29.30 | 25.00 | 26.30 |
| Crew 2 | 27.00 | 31.60 | 27.30 | 28.70 |
| Crew 3 | 25.10 | 29.70 | 25.40 | 26.70 |
| Crew 4 | 26.90 | 31.50 | 27.10 | 28.50 |
| Crew 5 | 27.60 | 32.20 | 27.90 | 29.30 |
| mean | 26.28 | 30.88 | 26.54 | 27.90 |

$$X_{ij} = 24.50, \quad \overline{x}_{i\bullet} + \overline{x}_{\bullet j} - \overline{x}_{\bullet\bullet} = 26.28 + 26.30 - 27.90 = 24.70$$

# Two-way ANOVA Table

| | SS | df | MS | F | P-value |
|---|---|---|---|---|---|
| Factor A | SSA | a-1 | SSA/(a-1) | MSA/MSE | $\mathsf{P}(F > \dots)$ |
| Factor B | SSB | b-1 | SSB/(b-1) | MSB/MSE | $\mathsf{P}(F > \dots)$ |
| Error | SSE | n-a-b+1 | SSE/(N-a-b+1) | | |
| Total | SST | n-1 | | | |

| Source of Variation | SS | df | MS | F | P-value |
|---|---|---|---|---|---|
| Rows | 19.89 | 4 | 4.97 | 1.098 | 0.4199 |
| Columns | 66.77 | 2 | 33.39 | 7.37 | 0.0153 |
| Error | 36.23 | 8 | 4.53 | | |
| Total | 122.9 | 14 | | | |

Conclusion: At 5% significance level there is enough evidence to suspect that machines are different, but not enough evidence to suspect that operators are different.

# Decomposition of Sum of Squares

| | | | | *mean* |
|---|---|---|---|---|
| | $\{x_{11},\ldots\}$ | $\{x_{12},\ldots\}$ | $\{x_{13},\ldots\}$ | $x_{1\bullet}$ |
| | $\{x_{21},\ldots\}$ | $\{x_{22},\ldots\}$ | $\{x_{23},\ldots\}$ | $x_{1\bullet}$ |
| | $\{x_{31},\ldots\}$ | $\{x_{32},\ldots\}$ | $\{x_{33},\ldots\}$ | $x_{1\bullet}$ |
| | $\{x_{41},\ldots\}$ | $\{x_{42},\ldots\}$ | $\{x_{43},\ldots\}$ | $x_{1\bullet}$ |
| | $\{x_{51},\ldots\}$ | $\{x_{52},\ldots\}$ | $\{x_{53},\ldots\}$ | $x_{1\bullet}$ |
| *mean* | $x_{\bullet 1}$ | $x_{\bullet 2}$ | $x_{\bullet 3}$ | $x_{\bullet\bullet}$ |

$$SST = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\alpha=1}^{k}(x_{ij,\alpha} - \overline{x}_{\bullet\bullet})^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\alpha=1}^{k}(x_{ij,\alpha} - \overline{x}_{ij,\bullet})^2 + k\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij,\bullet} - \overline{x}_{\bullet\bullet})^2 =$$

$$SSE + SSA + SSB + \sum_{i}\sum_{j}\sum_{\alpha}(x_{ij,\alpha} - \overline{x}_{ij,\alpha})^2 = SSE + SSA + SSB + SSAB$$

**SSAB = interaction of factors A and B**

**Assumption: SSE is the sum of squares of independent** $\mathcal{N}(0, \sigma^2)$

## Problem 2.2

*The following data on corn yields are obtained by planting three seed types using five fertilizers.*

| | *Fert I* | *Fert II* | *Fert III* | *Fert IV* | *Fert V* |
|---|---|---|---|---|---|
| *Seed A-402* | *106, 110* | *95, 100* | *94, 107* | *103, 104* | *100, 102* |
| *Seed B-894* | *110, 112* | *98, 99* | *100, 101* | *108, 112* | *105, 107* |
| *Seed C-952* | *94, 97* | *86, 87* | *98, 99* | *99, 101* | *94, 98* |

*Test at 5% significance level the hypothesis that corn yield depends on the seed type, fertilizer type, or the combination of the two.*

## Solution

*By using R statistics* `summary(aov(value seed+fert+seed*fert, data))`

| | *Df* | *Sum Sq* | *Mean Sq* | *F value* | *Pr(>F)* |
|---|---|---|---|---|---|
| *seed* | *2* | *512.87* | *256.43* | *28.28* | *0.0000* |
| *fert* | *4* | *449.47* | *112.37* | *12.39* | *0.0001* |
| *seed:fert* | *8* | *143.13* | *17.89* | *1.97* | *0.1221* |
| *Residuals* | *15* | *136.00* | *9.07* | | |

Given

| | Fert I | Fert II | Fert III | Fert IV | Fert V |
|---|---|---|---|---|---|
| Seed A-402 | 106, 110 | 95, 100 | 94, 107 | 103, 104 | 100, 102 |
| Seed B-894 | 110, 112 | 98, 99 | 100, 101 | 108, 112 | 105, 107 |
| Seed C-952 | 94, 97 | 86, 87 | 98, 99 | 99, 101 | 94, 98 |

$SS = 1241.467$

By cell

| | Fert I | Fert II | Fert III | Fert IV | Fert V | mean |
|---|---|---|---|---|---|---|
| Seed A-402 | 108.00 | 97.50 | 100.50 | 103.50 | 101.00 | 102.10 |
| Seed B-894 | 111.00 | 98.50 | 100.50 | 110.00 | 106.00 | 105.20 |
| Seed C-952 | 95.50 | 86.50 | 98.50 | 100.00 | 96.00 | 95.30 |
| mean | 104.83 | 94.17 | 99.83 | 104.50 | 101.00 | |

$SS = 552.7$

By row (seed)

| | Fert I | Fert II | Fert III | Fert IV | Fert V |
|---|---|---|---|---|---|
| Seed A-402 | 102.10 | 102.10 | 102.10 | 102.10 | 102.10 |
| Seed B-894 | 105.20 | 105.20 | 105.20 | 105.20 | 105.20 |
| Seed C-952 | 95.30 | 95.30 | 95.30 | 95.30 | 95.30 |

$SS = 256.4$

By column (fert)

| | Fert I | Fert II | Fert III | Fert IV | Fert V |
|---|---|---|---|---|---|
| Seed A-402 | 104.83 | 94.17 | 99.83 | 104.50 | 101.00 |
| Seed B-894 | 104.83 | 94.17 | 99.83 | 104.50 | 101.00 |
| Seed C-952 | 104.83 | 94.17 | 99.83 | 104.50 | 101.00 |

$SS = 224.7$

**ANOVA**

| | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|---|
| **seed** | **2** | **256.4*2=512.8** | **256.43** | **28.28** | **0.0000** |
| **fert** | **4** | **224.7*2=449.4** | **112.37** | **12.39** | **0.0001** |
| **seed:fert** | **8** | **(552.7-256.4-224.7)*2=143** | **17.89** | **1.97** | **0.1221** |
| **Residuals** | **15** | **1241.5-552.7*2=136.1** | **9.07** | | |
| **Toal** | **30** | | | | |

# Summary

- Residual is the difference between the observed and the fitted value
- Sum of the residuals w.r.t. LS line is equal to zero
- The Pearson correlation coefficient indicates the degree of linear dependence
- The Spearman correlation coefficient indicates the degree of monotonous dependence
- The coefficient of determination $R^2$, numerically equal to the square of the Pearson correlation coefficient, is the fraction of the variance in y that is explained by a linear fit of x
- Confidence interval is an estimate for the average model value
- Prediction interval is an estimate for a random deviation from the average value
- ANOVA assumes independent observations, normal populations, and variance homogeneity
- One-way ANOVA deals with one factor
- Two-way ANOVA deals with two factors and, possibly, their interactions