Topic 10

Analysis of variance (ANOVA)

Sergey Mastitsky ©
Klaipeda, 28-30 September 2011

10. Analysis of variance (ANOVA)

10.1. One-way ANOVA

Brushing up the theory

- With a t-test we can adequately compare only means of two groups
- One-way ANOVA is designed to compare mean values of two or more groups
- A little bit of notation:
  - x_{ij} – observation j from group i
  - $\bar{x}$ is the mean for group i
  - $\bar{X}$ is the grand mean (i.e. average of all observations)

Variance decomposition in ANOVA

- We can decompose the variation of observations as follows:

$$x_{ij} = \bar{X} + (\bar{x}_i - \bar{X}) + (x_{ij} - \bar{x}_i)$$

deviation of group mean from grand mean

deviation of observation from group mean

- This corresponds to the linear model:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

The null hypothesis of ANOVA

- The effect of the factor under study is insignificant, i.e. the observed differences between group means are accidental, and in reality all groups are just random samples from the same normal distribution with mean μ
- This implies that all a_i are zero:

$$X_{ij} = \mu + \varepsilon_{ij}$$

- How can we test this hypothesis?

The sums of squares

$$x_{ij} = \overline{X} + (\overline{x}_i - \overline{X}) + (x_{ij} - \overline{x}_i)$$

Consider the sums of squares of these deviations

$$SSD_B = \sum_i \sum_j (\overline{x} - \overline{X})^2 =$$

$$\sum_i n_i (\overline{x} - \overline{X})^2$$

Variation between groups

$$SSD_W = \sum_i \sum_j (x_{ij} - \overline{x}_i)^2$$

Variation within groups

Author: Sergey Mastitsky

Total variation is the sum of SSD_B and SSD_W

- It can be shown that

$$SSD_{total} = SSD_B + SSD_W = \sum_i \sum_j (x_{ij} - \bar{X})^2$$

- Thus, grouping of observations "explains" part of the total variation – the larger this part, the higher the influence of the grouping factor
- But what is "large"? We need to compare against something

Author: Sergey Mastitsky

Mean squares

- The best way is to compare SSD_B against SSD_W
- To do that, we first have to normalize both quantities by their corresponding d.f.:

$$MS_W = SSD_W/(N - k)$$

$$MS_B = SSD_B/(k - 1)$$

- The resulting quantities are called *mean squares*

Analysis of variance

- MS_W is the pooled variance obtained by combining individual group variances, and is the estimate of σ^2
- MS_B is also an estimate of σ^2, but if there *is* a group effect MS_B will tend to be larger than MS_W
- Thus, we simply need to compare MS_W with MS_B to test the effect of the grouping factor
- This is why the analysis is called ANOVA, even though we intend to compare mean values

Analysis of variance

- A formal test to compare the two estimates of variance is the *F*-test:

$$F = \frac{MS_B}{MS_W}$$

- Ideally, *F* should be ~1; but if there is a strong effect of the grouping factor, *F* will be much larger than 1
- As with the *t*-test, the null hypothesis of no effect is rejected if *F* falls outside the acceptance region of the *F*-distribution (at a certain α-level and d.f.)

10. Analysis of variance (ANOVA)

10.2. One-way ANOVA: implementation in R

Specifying the model in R

- As we've already seen, ANOVA is, in fact, a linear model ($X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$)
- Thus, the same R-function that we used for linear regression, can be used for ANOVA – `lm()`
- `lm()` is used to calculate the model object, and then the function `anova()` is used to extract the actual ANOVA table

An example: red cell folate data

```
> library(ISwR)
> data(red.cell.folate)
> summary(red.cell.folate)
```

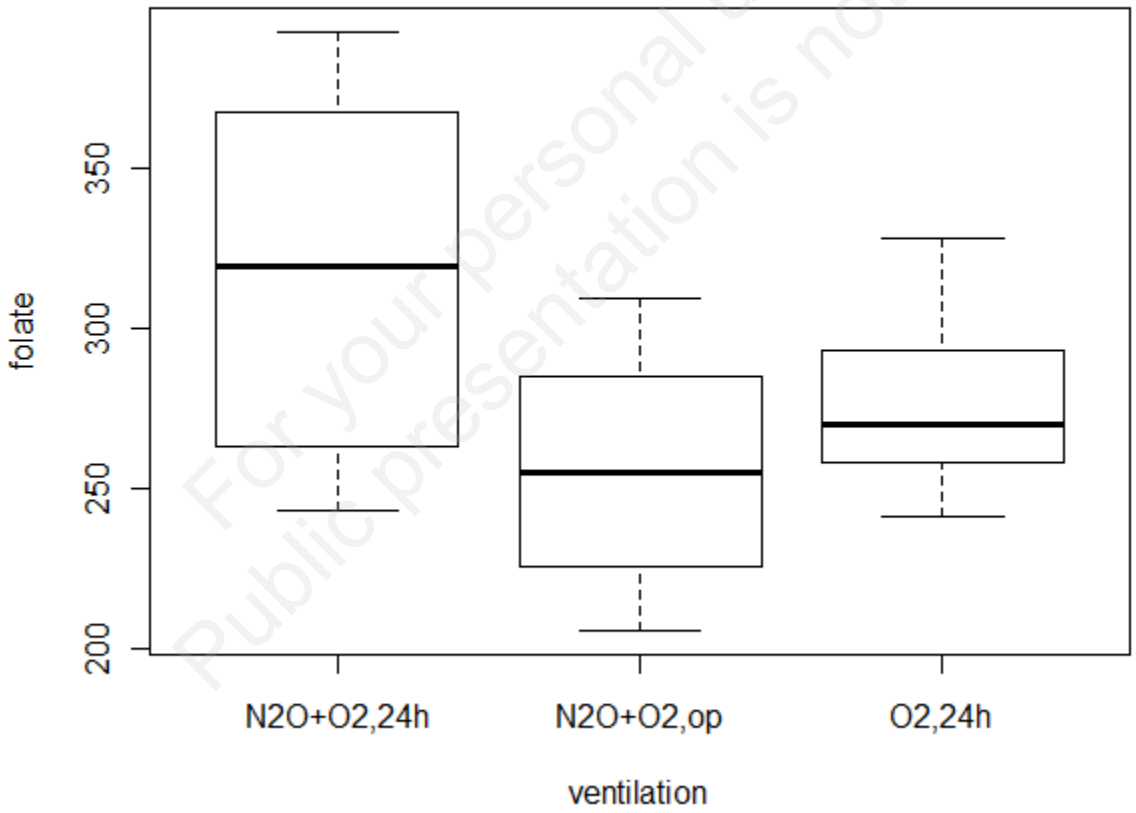
```
     folate           ventilation
 Min.   :206.0   N2O+O2,24h:8
 1st Qu.:249.5   N2O+O2,op :9
 Median :274.0   O2,24h    :5
 Mean   :283.2
 3rd Qu.:305.5
 Max.   :392.0
```

Concentrations of folate (natural source of B9) in red blood cells

Three groups of patients with different regimes of anesthesia

Graphical summary of the data: are the observed differences significant?

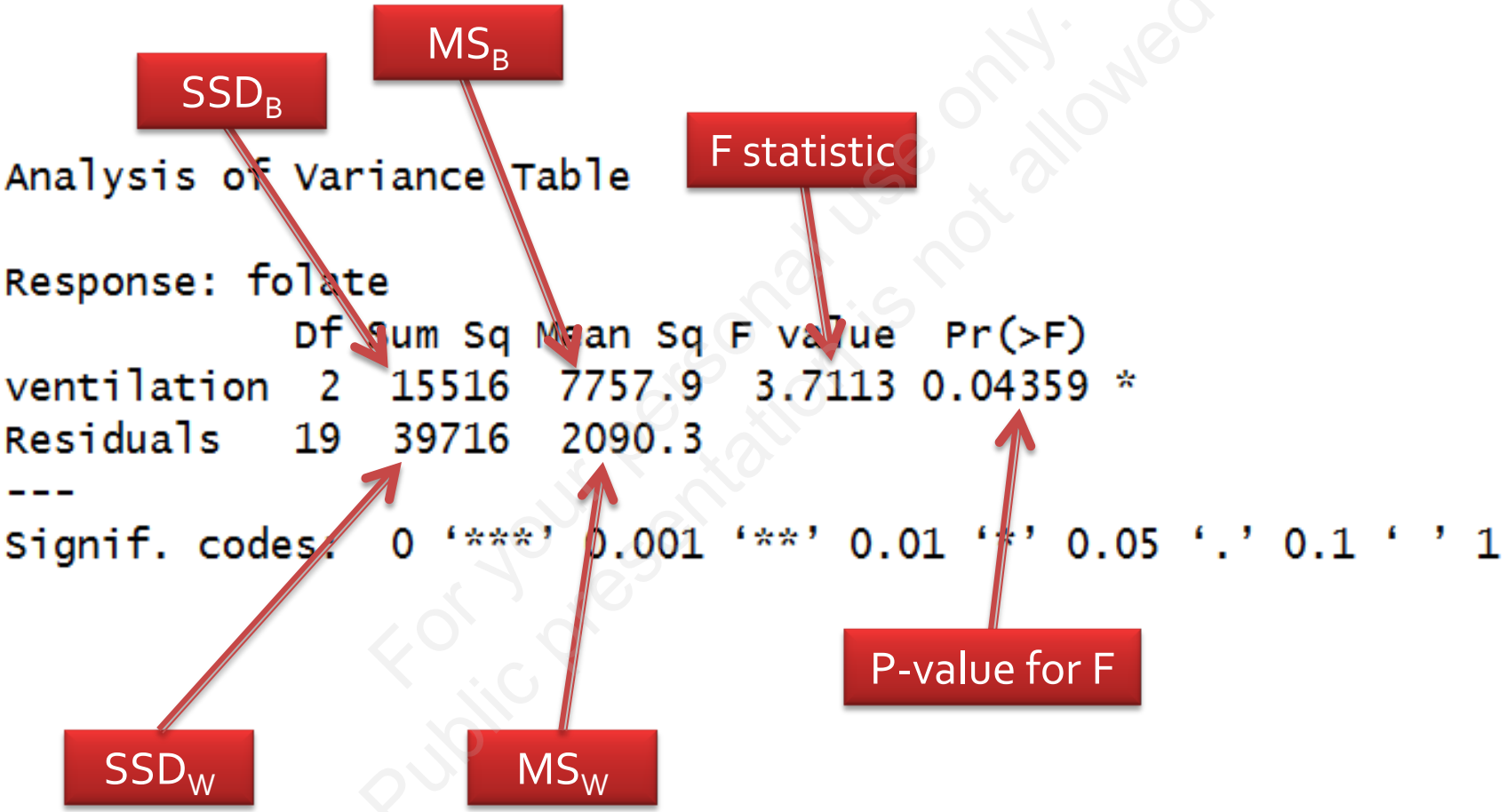
```
> plot(folate ~ ventilation,
       data = red.cell.folate)
```



One-way ANOVA in R

```
# Calculate model object:
> cell.mod <-
lm(folate ~ ventilation,
        data = red.cell.folate)

# Extract ANOVA table:
> anova(cell.mod)
```

Results of ANOVA



Results of ANOVA

- The *F*-test suggest that there is a (marginally) significant difference among the average concentrations of folate (P = 0.0436)
- This is the only conclusion that ANOVA allows us to make! It does not show where exactly the differences lie (i.e. which particular pairs of groups differ)

10. Analysis of variance (ANOVA)

10.3. *Post-hoc* analysis: pairwise comparisons and multiple testing

How to find which particular groups differ in ANOVA?

- Part of this information can be found in the regression coefficients stored in the model object:

```
# Extract regression coefficients from cell.mod object:
> summary(cell.mod)
...

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op  -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h     -38.62      26.06  -1.482   0.1548
```

Interpretation of the regression coefficients table in ANOVA

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op  -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h     -38.62      26.06  -1.482   0.1548
```

- `Intercept` – mean folate concentration in the first group, `N2O+O2,24h`
- Second coefficient – difference between `Intercept` and `N2O+O2,op`
- Third coefficient – difference between `Intercept` and `O2,24h`

Author: Sergey Mastitsky

Contrasts

- There are multiple ways of representing the effect of a factor variable in linear models
- The representation is made in terms of *contrasts*
- We don't go deep into this topic – just know that in our case we used the default *treatment contrasts*
- In treatment contrasts the first group is treated as a baseline and the other groups are given relative to that

Interpretation of the regression coefficients table in ANOVA

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           316.62      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op  -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h     -38.62      26.06  -1.482   0.1548
```

- Mean concentration of folate in group `N2O+O2,24h` significantly differs from that in `N2O+O2,op` (P = 0.014, t-test)
- Mean concentration of folate in group `N2O+O2,24h` doesn't differ from that in `O2,24h` (P = 0.155, t-test)
- (!) We can say nothing about the difference between `N2O+O2,op` and `O2,24h`

Multiple testing

- If we want to compare all the groups pair-wise, we have to correct the resulting P-values for multiple testing because the more tests we perform on the same data, the higher the chance of mistakenly rejecting at least one true null hypothesis
- A common adjustment method is the Bonferroni correction – the test's P-values are multiplied by the number of comparisons

Bonferroni correction for multiple t-tests using the `pairwise.t.test()`

```
> attach(red.cell.folate)
> pairwise.t.test(folate,
  ventilation, p.adj =
  "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  folate and ventilation

          N2O+O2,24h N2O+O2,op
N2O+O2,op 0.042      -
O2,24h    0.464      1.000

P value adjustment method: bonferroni
```


Author: Sergey Mastitsky

The Tukey HSD test

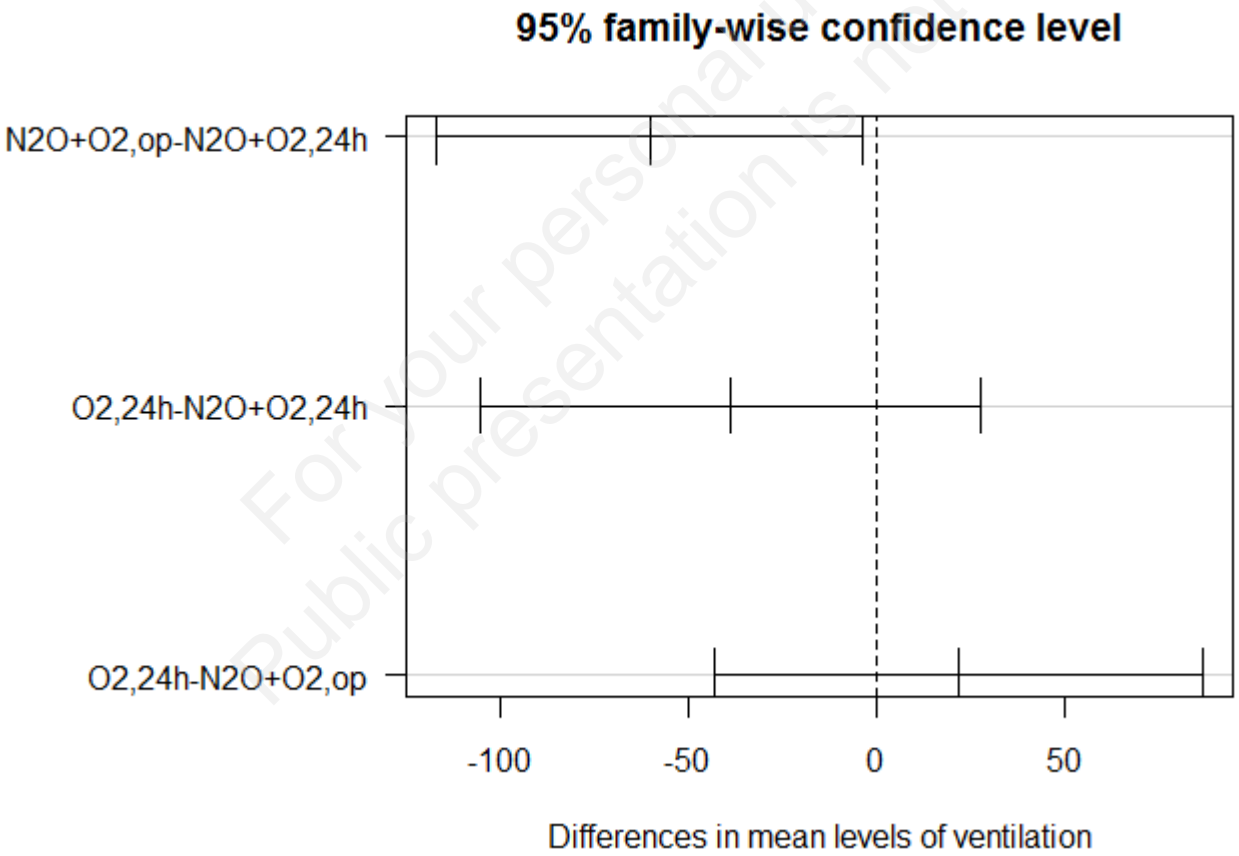
- With a large number of groups, Bonferroni correction becomes extremely *conservative*
- In such cases it makes more sense to use less conservative Tukey HSD test:

```
> TukeyHSD(aov(cell.mod))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = cell.mod)

$ventilation
                              diff        lwr      upr     p adj
N2O+O2,op-N2O+O2,24h -60.18056 -116.61904 -3.74207 0.0354792
O2,24h-N2O+O2,24h    -38.62500 -104.84037 27.59037 0.3214767
O2,24h-N2O+O2,op      21.55556  -43.22951 86.34062 0.6802018
```

Graphical representation of the Tukey HSD test

```
> par(mar = c(4.5, 12, 3.5, 1))
> plot(TukeyHSD(aov(cell.mod)))
```



10. Analysis of variance (ANOVA)

10.4. Testing for and relaxing the homogeneity of variance assumption

Homogeneity of variance assumption

- The traditional ANOVA assumes that the group variances are statistically equal

- This can be checked with the Bartlett's test (careful – sensitive to normality!):

```
> bartlett.test(folate~ventilation)

        Bartlett test of homogeneity of variances

data:  folate by ventilation
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

Relaxing the variance assumption

- As with the t-test, we can apply the Welch's approach and relax the variance assumption of ANOVA:

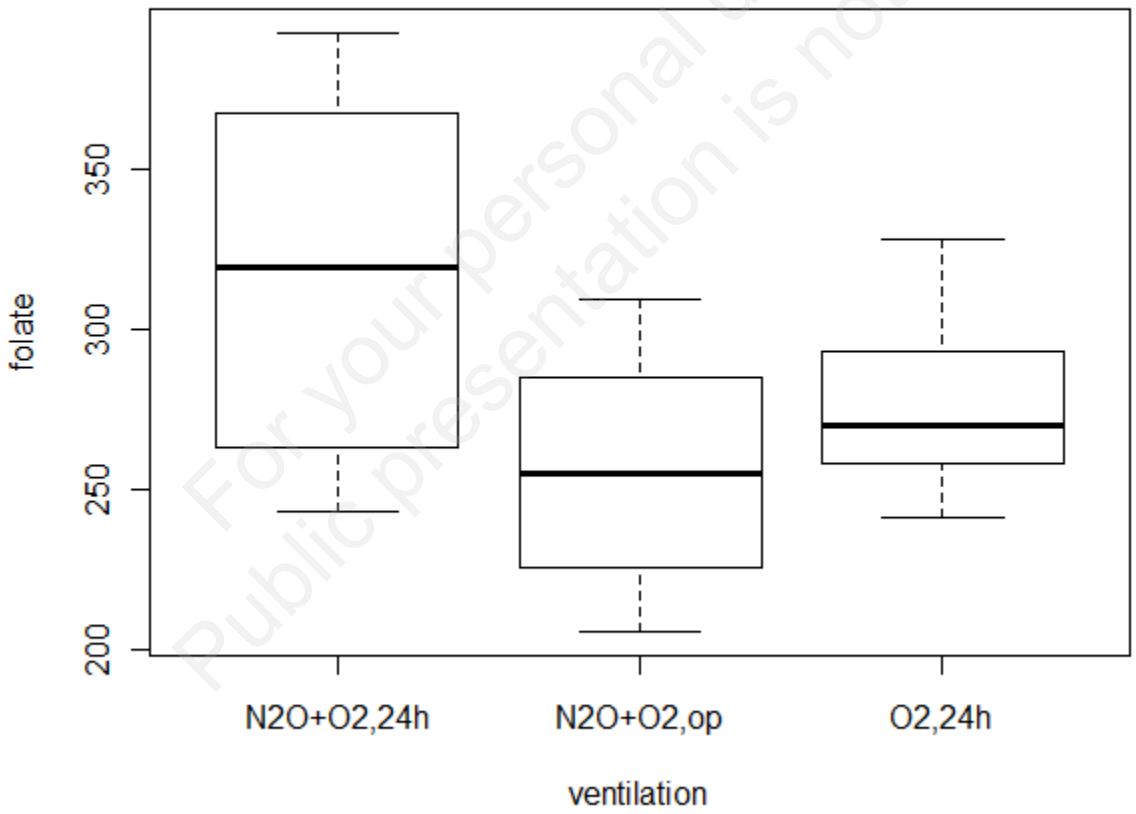
```
> oneway.test(folate~ventilation)

 One-way analysis of means (not assuming equal variances)

data:  folate and ventilation
F = 2.9704, num df = 2.000, denom df = 11.065, p-value = 0.09277
```

10. Analysis of variance (ANOVA)

10.5. Graphical representation of the ANOVA-type data (i.e. grouped data)

There are many ways, e.g. a boxplot:

```
> plot(folate ~ ventilation,
       data = red.cell.folate)
```



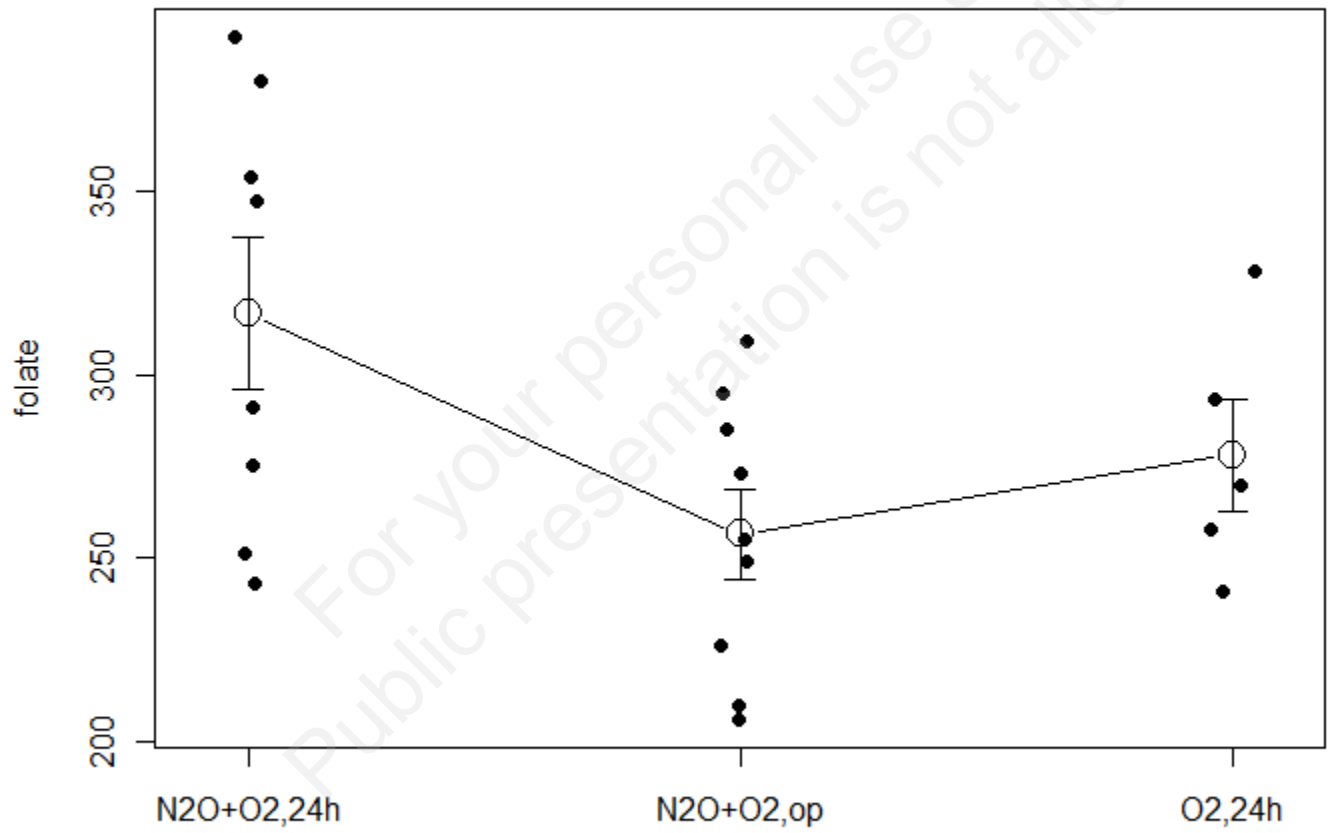
A plot of raw data with group means and their SEs indicated

```
# Calculate group means, SDs, sample sizes, and
  SEs:
> xbar <- tapply(folate,
    ventilation, mean)
> s <- tapply(folate,
    ventilation, sd)
> n <- tapply(folate,
    ventilation, length)
> sem <- s/sqrt(n)
```

A plot of raw data with group means and their SEs indicated

```
> stripchart(folate ~ ventilation,
    method = "jitter",
    jitter = 0.05,
    pch = 16, vertical = TRUE)
> arrows(1:3, xbar+sem,
    1:3, xbar-sem,
    angle = 90,code = 3,length = .1)
> lines(1:3, xbar, pch = 1,
     type = "b", cex = 2)
```

A plot of raw data with group means and their SEs indicated



10. Analysis of variance (ANOVA)

10.6. Kruskal-Wallis test

Kruskal-Wallis ANOVA by ranks

- When the group distributions substantially differ from normal, one cannot trust the P-value produced by the traditional one-way ANOVA
- To resolve this issue, one can use a non-parametric counterpart of the traditional ANOVA – Kruskal-Wallis ANOVA
- Similar to Wilcoxon test, K-W ANOVA is based on ranks of the raw data

Kruskal-Wallis ANOVA in R

```
> kruskal.test(folate~ventilation)

        Kruskal-Wallis rank sum test

data:  ventilation by folate
Kruskal-Wallis chi-squared = 21, df = 21, p-value = 0.4589
```

Conclusion: there is no statistically significant difference in folate concentration among the three experimental groups

This is not surprising:

1) The ANOVA results was marginally significant
2) K-W test is non-parametric = less powerful

10. Analysis of variance (ANOVA)

10.7. Two-way ANOVA

Two-way ANOVA

- We don't go into theory and just note that, similar to the one-way analysis of variance, the two-way ANOVA can be considered as a linear model:

$$X_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Two-way ANOVA in R

```
> load(pH_experiment.rda)
> names(LWdata)
```

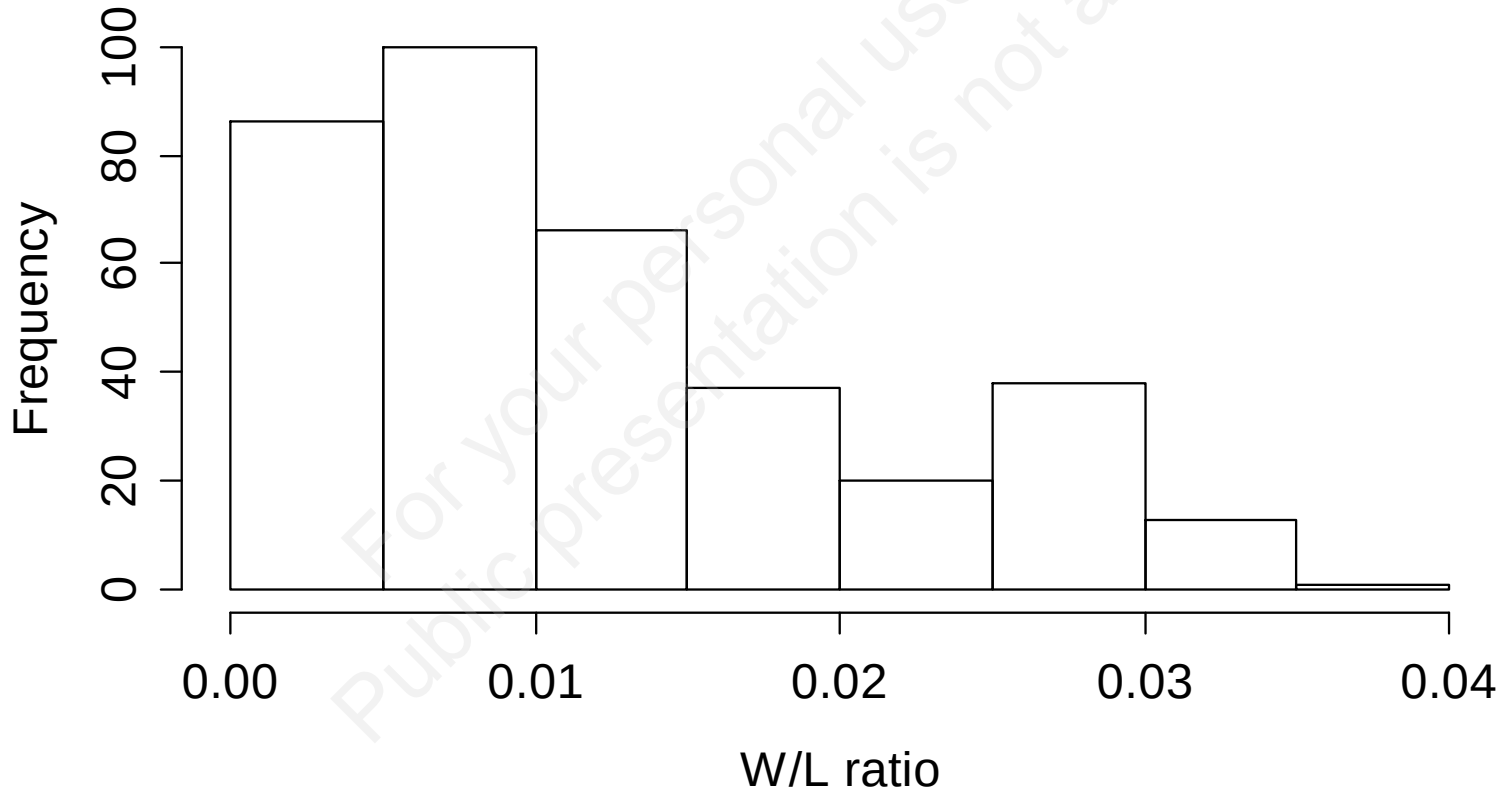
Suppose, we want to check the effect of Treatment and Barrel on the Weight/Length ratio

First, let's check the distribution of W/L:

```
> attach(PHdata)
> hist(Weigth/Length)
```

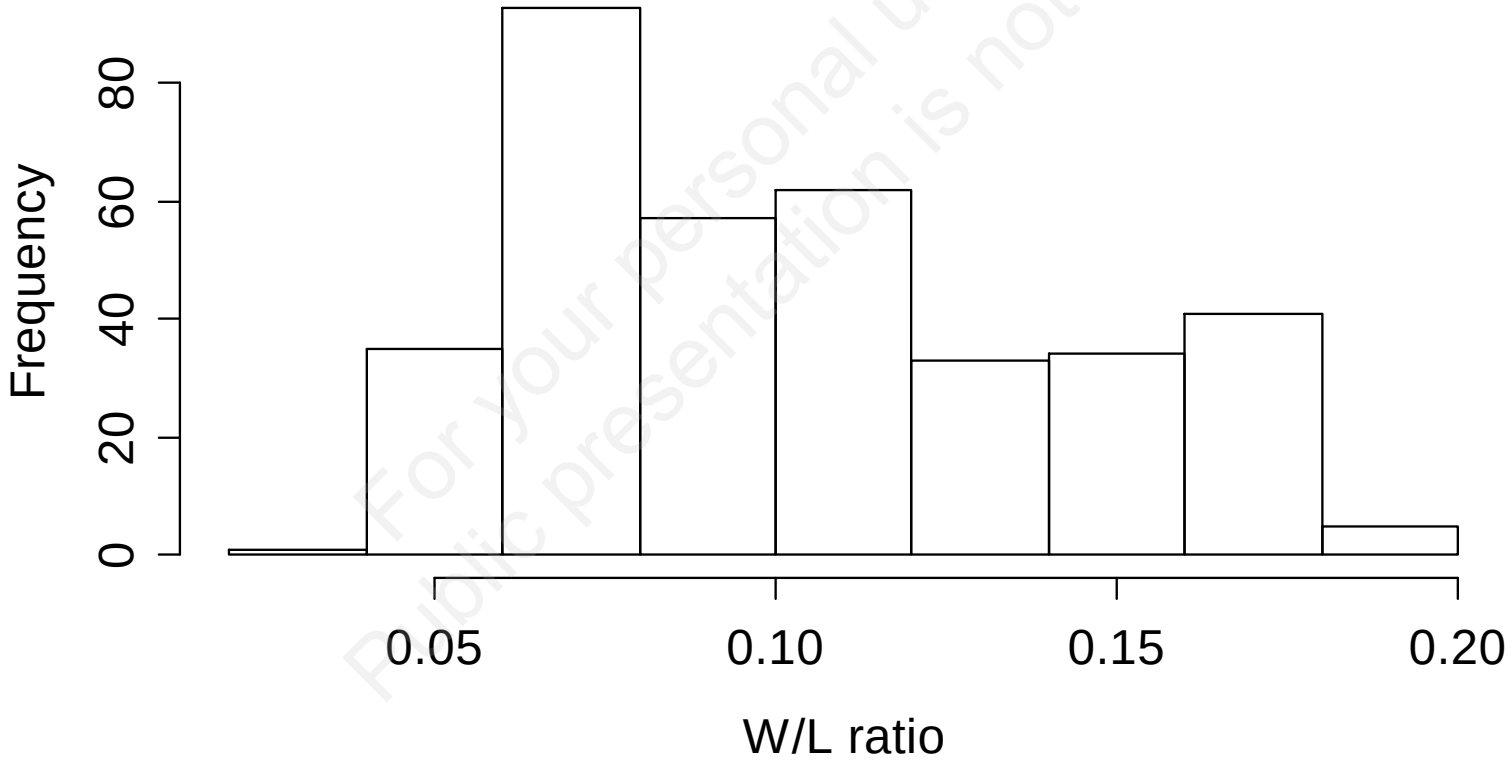
EDA of the pH experiment data

- The W/L ratio values are not normally distributed



EDA of the pH experiment data

```
> hist(sqrt(Weigth/Length))
```



Two-way ANOVA in R

```
> LWratio <- sqrt(Weight/Length)
> pH.mod <- lm(
    LWratio ~ Treatment + Barrel)
> anova(pH.mod)
```

```
Analysis of Variance Table

Response: LWratio
           Df  Sum Sq   Mean Sq F value  Pr(>F)
Treatment   3 0.01309 0.0043629  2.9675 0.03202 *
Barrel      8 0.00164 0.0002047  0.1392 0.99736
Residuals 349 0.51311 0.0014702
```


Author: Sergey Mastitsky