Statistics II
Single ANOVA

Çağrı Çöltekin

ideas/examples/slides from
John Nerbonne & Hartmut Fitz

University of Groningen, Dept of Information Science

May 8, 2013

Correlation

- The correlation coefficient (r) is a standardized symmetric measure of covariance between two variables.
- The correlation coefficient ranges between -1 and 1.
- Correlation and regression are strongly related.
- The most common correlation coefficient is Pearson's r, which assumes a linear relationship between two variables.
- When this assumption is not correct, non-parametric alternatives Spearman's ρ or Kendall's τ can be used.
- Correlation is not causation!

Simple regression

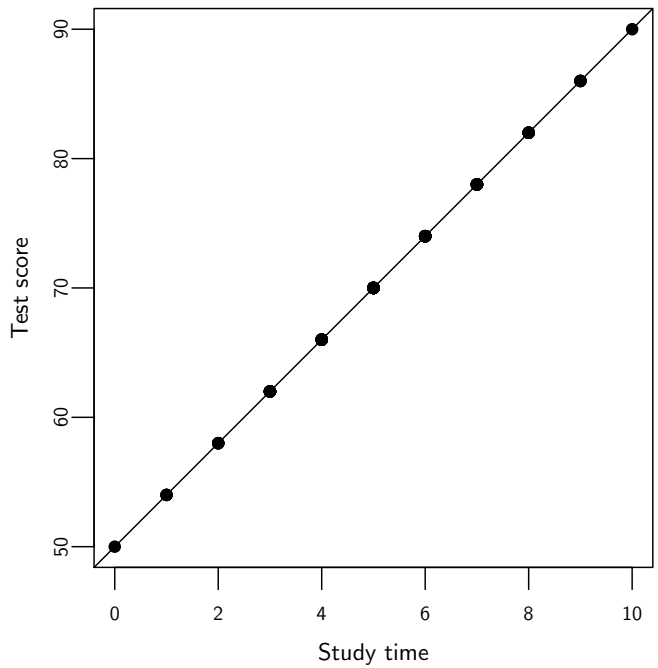
$$y_i = a + bx_i + e_i$$

- y is the *response* (or outcome, or dependent) variable. The index i represent each unit observation/measurement (sometimes called a 'case').
- x is the *predictor* (or explanatory, or independent) variable.
- a is the intercept.
- b is the slope of the regression line.
- $a + bx$ is the *deterministic* part of the model (we sometimes use $\hat{y}$).
- e is the residual, error, or the variation that is not accounted for by the model. Assumed to be (approximately) normally distributed with 0 mean (e_i are assumed to be i.i.d).

Regression by (another) example

Assume we want to see the relation between exam scores and study time.

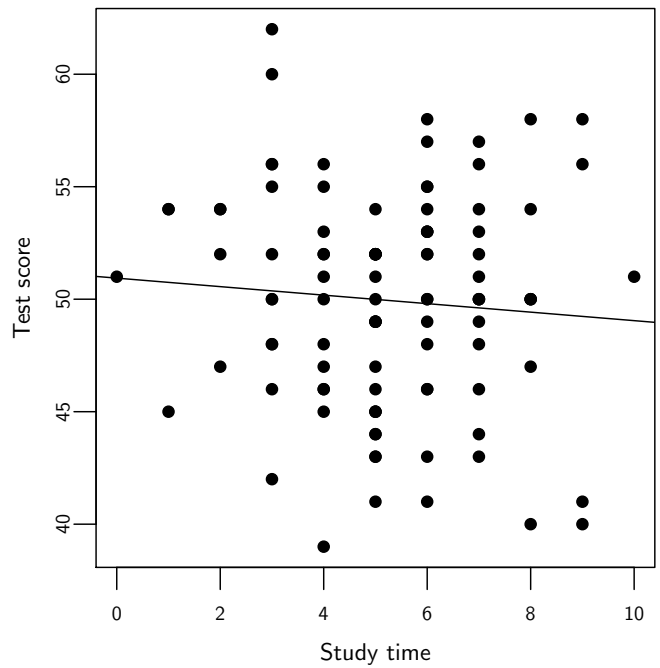
- If the study time was the prefect (linear) predictor of the score, we'd get a perfect correlation.



Regression by (another) example

Assume we want to see the relation between exam scores and study time.

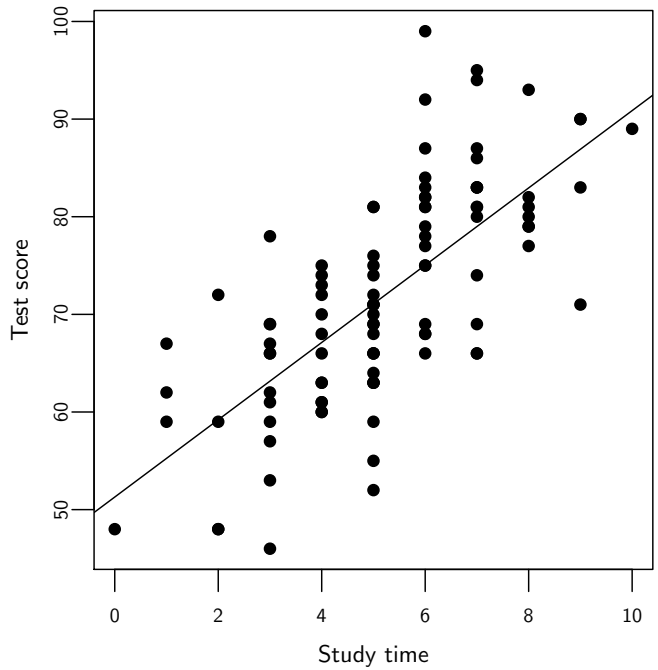
- If the study time was the prefect (linear) predictor of the score, we'd get a perfect correlation.
- If the test was irrelevant, we would expect no correlation.



Regression by (another) example

Assume we want to see the relation between exam scores and study time.

- If the study time was the prefect (linear) predictor of the score, we'd get a perfect correlation.
- If the test was irrelevant, we would expect no correlation.
- Real world is in between these two extremes.



Regression towards the mean

If we did a re-test (without additional study).

- If the study time was the perfect predictor, we expect everyone to get the exact same scores.

Regression towards the mean

If we did a re-test (without additional study).

- If the study time was the perfect predictor, we expect everyone to get the exact same scores.
- If we gave the test to monkeys, we would expect a complete re-shuffling of the scores: most successful monkey is highly likely to achieve worse, and least-successful monkey is highly likely to achieve better.

Regression towards the mean

If we did a re-test (without additional study).

- If the study time was the perfect predictor, we expect everyone to get the exact same scores.
- If we gave the test to monkeys, we would expect a complete re-shuffling of the scores: most successful monkey is highly likely to achieve worse, and least-successful monkey is highly likely to achieve better.
- Real world is in between these two extremes: success in both tests will be similar, however, extreme scores in the first test will tend to regress towards the mean.

Regression analysis step by step

1. Collect/check your data: cases should be independent.
2. Fit your model (let the computer do it).
3. Check assumptions or problem indications:

linearity	scatter plot of 'y vs. x' or 'residuals vs. fitted'.
normality	(of residuals!) histogram, Q-Q (or P-P) plot.
constant variance	(of residuals!) 'residuals vs. fitted' plot.
outliers	scatter plot of 'y vs. x' together with regression line, residual histogram or box plot.
influential cases	scatter plot of 'y vs. x', 'residuals vs. fitted', or more specialized statistics like *Cook's distance*.

4. Interpret your results:
   - Model parameters (coefficients): intercept and slope estimates.
   - Model fit: coefficient of determination (r^2).
   - Generalizability of the estimates: F-test for the model, and t-tests for the coefficients.
   - Prediction: confidence intervals for regression line (expected value of the response variable), and future observations.

Multiple regression

$$y_i = \underbrace{a + b_1 x_{i,1} + b_2 x_{2,i} + \ldots + b_k x_{k,i}}_{\hat{y}} + e_i$$

- Multiple regression is a generalization of the simple regression, where we predict the outcome using multiple predictors.
- Multicollinearity causes problems in estimation and interpretation of multiple-regression models.
- Model selection (finding a model that fits the data well, but not more complex than necessary) is important.

Hypothesis testing

Aim: make inferences about the population based on a sample regarding a research question.

Procedure:

- Formulate your question as two explicit hypotheses:

  alternative hypothesis (H_a) supports what you expect to find in the population.

  null hypothesis (H_0) is the formulation of the case where your expectations were wrong.

- Set a probability level (α-level) at which to reject the H_0. Typical values are 0.05, 0.01, 0.001.
- Calculate the probability, p, of obtaining the sample you have, if H_0 was true.
- If $p < \alpha$, we reject the H_0, otherwise, we fail to reject the H_0.

Hypothesis testing: example

We want to know whether the new design of a web page is easier to use based on responses to a questionnaire from two groups, one on old design, one on new design.

Procedure:

- Formulate your question as two explicit hypotheses:

  H_a the mean response score is different for each group ($\mu_1 \neq \mu_2$).

  H_0 the mean response score is the same for both groups ($\mu_1 = \mu_2$).

- We set $\alpha = 0.05$.
- The p-value for obtaining these samples given H_0 is true, can be calculated using the t-distribution (given the response scores for both groups are normally distributed, and the variances are similar).
- If $p < \alpha$, we reject the H_0, otherwise, we fail to reject the H_0.

Hypothesis testing: Type I and Type II errors

		Real world	
		H_0 is false	H_0 is true
Test decision	Reject H_0 ($p < \alpha$)	Correct decision True positive	Type I error False positive
	Fail to reject H_0 ($p \geq \alpha$)	Type II error False negative	Correct decision True negative

- Note that accepting H_0 means we will be wrong (committing a Type I error) with probability α.
- If we set $\alpha = 0.05$, and repeat an experiment 20 times, we expect to reject the null hypothesis once even it is true.

Independent samples t-test

Independent samples t-test is used when,

- we have a numeric variable (e.g., height, test score) measured for two groups (e.g., male/female, healthy/patient)
- the groups (samples) are independent.
  - If related: use paired t-test
- The samples are approximately normal, and variances are similar.
  - If violated: use non-parametric alternatives

Independent samples t-test

Independent samples t-test is used when,

- we have a numeric variable (e.g., height, test score) measured for two groups (e.g., male/female, healthy/patient)
- the groups (samples) are independent.
  - If related: use paired t-test
- The samples are approximately normal, and variances are similar.
  - If violated: use non-parametric alternatives

What if we have more than two groups?

Example problems for ANOVA

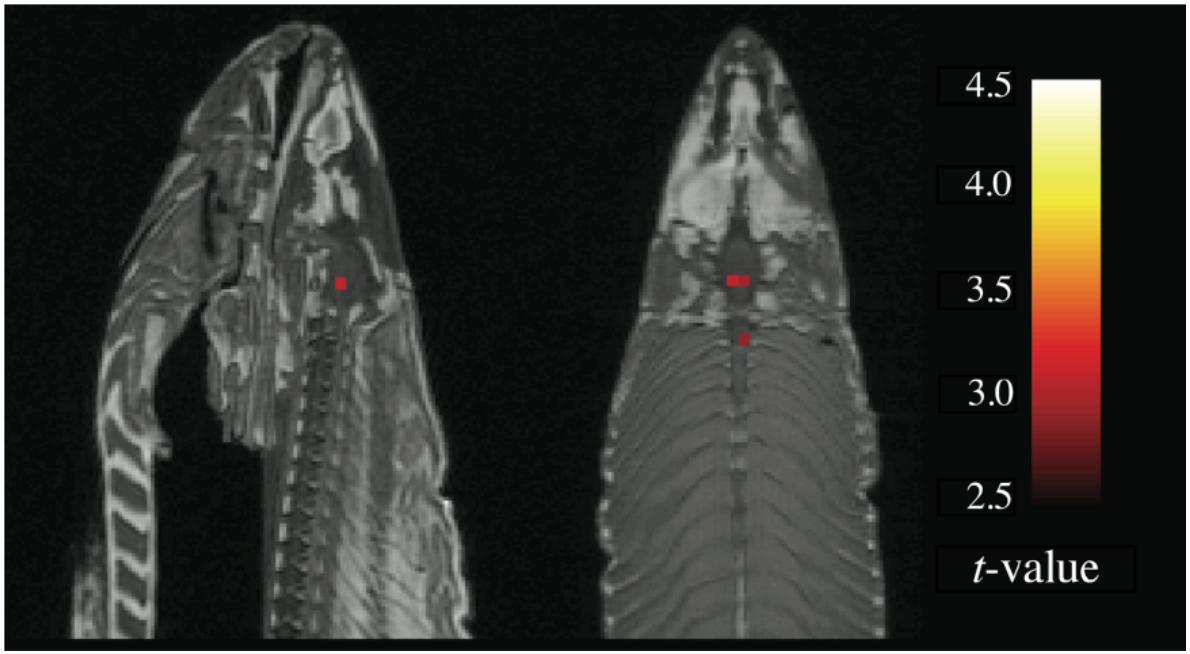
- Compare time needed for lexical recognition in
  1. healthy adults
  2. patients with Wernicke's aphasia
  3. patients with Broca's aphasia
- Effect of background color choice in a web site.
- Compare Dutch proficiency scores of second language learners based on their native language.

Why not multiple t-tests?

- Multiple comparisons over the same sample increases the chances of rejecting the null hypothesis (finding an effect where there is none).
- With $\alpha = 0.05$, if you do 20 different t-tests on the same sample, we expect to one of them being significant if the null hypothesis was true.
- We need
  3 comparisons 3 groups,
  6 comparisons for 4 groups,
  10 comparisons for 5 groups,
  45 comparisons for 10 groups,
  4950 comparisons for 100 groups.
- In general, for k groups, we need $\binom{k}{2}$ comparisons.

An extreme demonstration

finding emotional response in a dead salmon's brain activity



Subject One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at he time of scanning.

Task The task administered to the salmon involved completing an open-end mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Results Several active voxels were discovered in a cluster located within the salmon's brain cavity . . . with a cluster-level significance of $p = 0.001$. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

* From the poster by Bennett et al. (2009).

ANOVA

- ANOVA (analysis of variance) is a method to compare means of more than two groups.
- For two groups the result is equivalent to t-test.
- ANOVA indicate whether there is any difference at all. For k groups:

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_k$$

- A limited number and type of comparisons can be carried out by specifying contrasts.
- Otherwise, post-hoc pairwise comparisons can be carried out using corrected α-levels.
- ANOVA is strongly related to regression (later today).

Logic of ANOVA

We want to know whether there are any differences between the means of k groups.

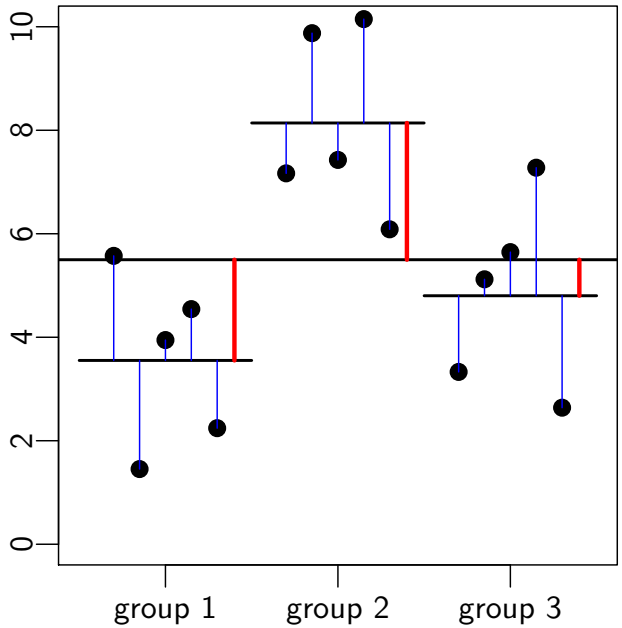
- If the variance between the groups is higher than the variance within the groups, there must be a significant group effect.
- Between group variance ($MS_{between}$, or MS_M or MS_G) is characterized by variance between the group means.
- Within group variance (MS_{within}, or MS_R or MS_E) is characterized by variance of data round the group means.

Then, the statistic of interest is

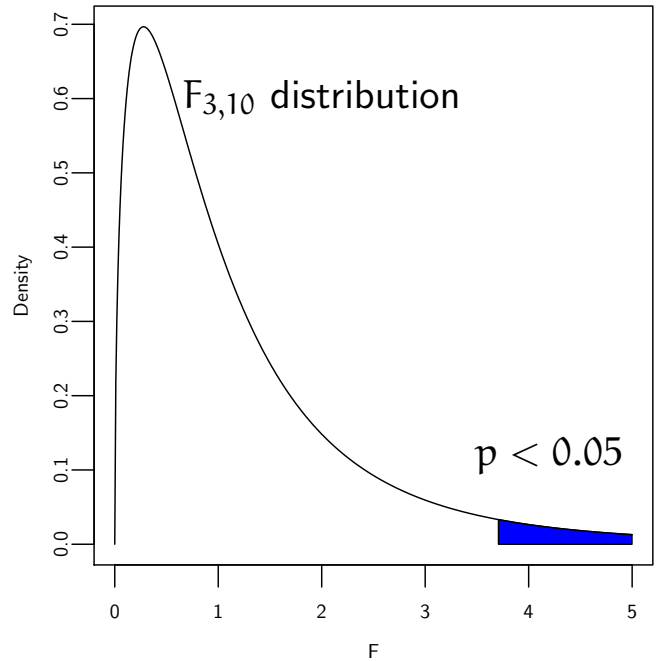
$$F = \frac{MS_{between}}{MS_{within}}$$

.

ANOVA: visualization



F-distribution



ANOVA: assumptions

- All observations are independent.
- The data for each group follow an approximately normal distribution.
- The variances for each group are approximately the same.

ANOVA: example

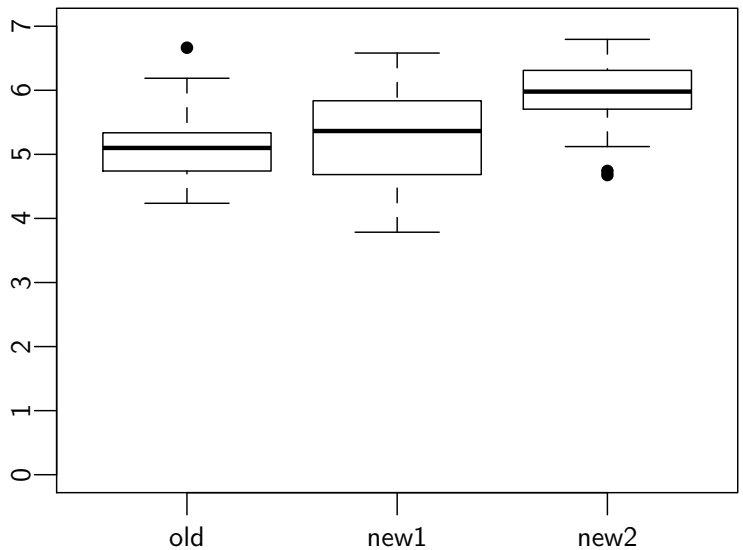
We have two new designs for RuG web site, want to know which one is easier to use. We test the new web site prototypes and the old one on three different group, and get their opinion via a questionnaire with 7-point scale. The data looks like:

	Old	New 1	New 2
	4.4	6.6	5.9
	5.8	6.2	4.9
	⋮	⋮	⋮
Mean	4.76	5.03	6.11
Variance	1.11	1.12	0.97

Note: rows in the table are not related!

Visualizing the data

Box-and-whisker plots (or box plots) are one of the best ways to visualize this type of data.



ANOVA results from software

```
Analysis of Variance Table
Response: ease
          Df Sum Sq Mean Sq F value    Pr(>F)
design     2 10.796  5.3978  13.955 5.541e-06 ***
Residuals 87 33.652  0.3868
```

- There is a significant effect (p-values is 0.0000055)
- but we do not know where the effect is.

Regression with categorical predictors: some terminology

- We take grouping variables (like design) as a categorical, or factor, variable.
- The values a grouping variable take are called levels.
- A categorical variable with k levels is converted to $k-1$ numeric variables, called 'indicator' or 'dummy' variables.

Regression with categorical predictors

We will use an example, where we measured speech rate of phrases within certain linguistic contexts.

- Consider 'context' variable with three levels ('A', 'B', 'C'), we can code it as two variables, 'contextB', 'contextC' :

level	contextB	contextC
A	0	0
B	1	0
C	0	1

- Other coding options (contrasts) are possible. With some constraints, the inferences will not change.

An example with only two levels

We want to check whether means of two of the contexts differ (labeled as 'A' and 'C').

An example with only two levels

We want to check whether means of two of the contexts differ (labeled as 'A' and 'C').
Normally we would do a t-test:

```
> t.test(rate2 ~ context2, var.equal=T)

        Two Sample t-test

data:  rate2 by context2
t = -1.4806, df = 98, p-value = 0.1419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.5596945  0.2267907
sample estimates:
mean in group A mean in group C
       6.428031        7.094483
```

Doing t-test with regression

- We have two levels of the predictor (A and C).
- We code 'A' as 0 and 'C' as 1.

$$y_i = a + b \times \text{contextC}_i + e_i$$

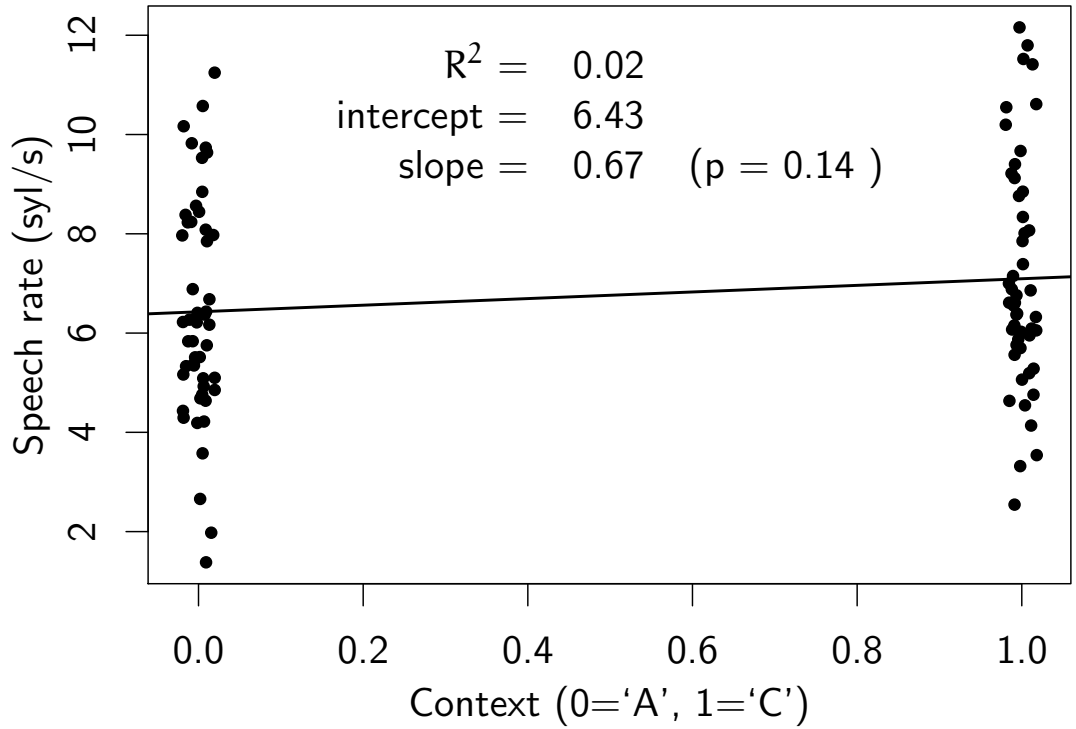
a (intercept) is the mean of level 'A'.

b (slope) is the mean difference between 'A' and 'B'.

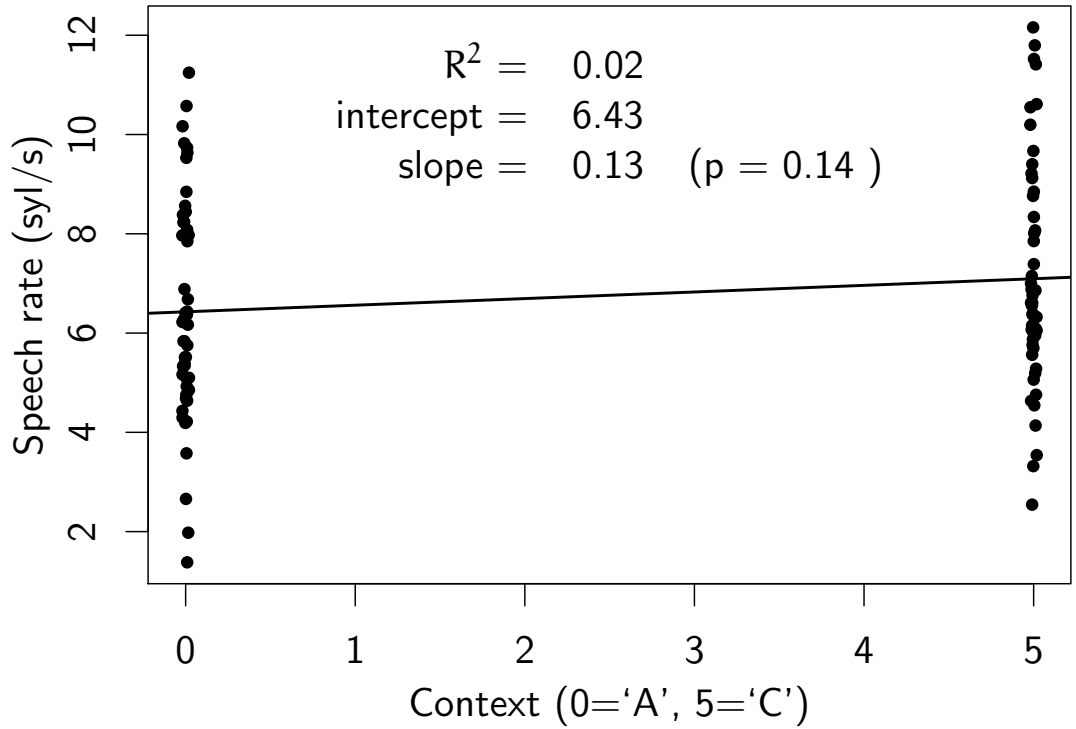
Doing t-test with regression: practice

```
> summary(lm(rate2 ~ context2))

Call:
lm(formula = rate2 ~ context2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0466 -1.3540 -0.4838  1.6895  5.0638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4280     0.3183  20.196   <2e-16 ***
context2C     0.6665     0.4501   1.481    0.142
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.251 on 98 degrees of freedom
Multiple R-squared: 0.02188,    Adjusted R-squared: 0.0119
F-statistic: 2.192 on 1 and 98 DF,  p-value: 0.1419
```

T-test as regression: the picture



T-test as regression: the picture



ANOVA as regression

Remembering that we code three levels as two indicator (dummy) variables:

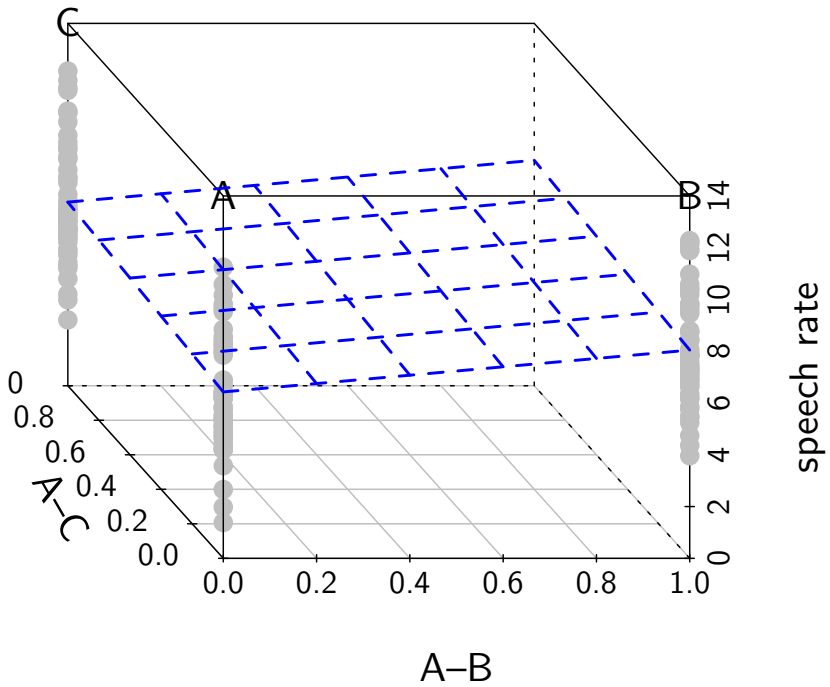
$$y_i = a + b_1 \times \text{contextB}_i + b_2 \times \text{contextC}_i + e_i$$

- a (intercept) is the mean of context 'A'.
- b_1 (slope of contextB) is the mean difference between 'A' and 'B'.
- b_2 (slope of contextC) is the mean difference between 'A' and 'C'.

ANOVA as regression: practice

```
> summary(lm(rate ~ context))

Call:
lm(formula = rate ~ context)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0466 -1.3719 -0.4616  1.6664  5.0638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4280     0.3128  20.548  < 2e-16 ***
contextB      1.6165     0.4424   3.654 0.000359 ***
contextC      0.6665     0.4424   1.506 0.134105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.212 on 147 degrees of freedom
Multiple R-squared: 0.08404,    Adjusted R-squared: 0.07158
F-statistic: 6.744 on 2 and 147 DF,  p-value: 0.001577
```

ANOVA as regression: the picture



ANOVA as regression: ANOVA table

```
> anova(lm(rate ~ context))
Analysis of Variance Table

Response: rate
           Df  Sum Sq Mean Sq F value    Pr(>F)
context     2  66.00  32.998  6.7437 0.001577 **
Residuals 147 719.29   4.893
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the fitted model is the same, we only summarize the results differently.

Contrast coding

- For k levels (or groups) we have $k - 1$ coefficients.
- We can code some comparisons (contrasts) into these coefficients to test for differences without running the risk of committing Type I errors.
- If the contrasts does not inflate the t-value (does not cause additional Type I error) it is called an orthogonal contrast.

More about contrasts next week.

Post-hoc comparisons

- In most cases, you will have a specific hypothesis and a (small) set of comparisons to make.
- You can do pairwise comparisons once you found a significant ANOVA result.
- Every comparison you make increases finding a significant difference where there isn't any (Type I error).
- If you do multiple comparisons you need to correct for it.
- Correction is applied such that your α-level is adjusted (called family-wise error rate).

Post-hoc comparisons (2)

Remember: finding a significant difference means that there is a chance (for example, $p = 0.05$) of finding a difference when there is no difference (null hypothesis is true).

- The simplest (and most conservative) correction is called 'Bonferroni' correction, which is obtained by dividing α to number of comparisons. If you have $\alpha = 0.05$ and n comparisons your family-wise α should be $\frac{0.05}{n}$.
- Bonferroni correction is safe in all cases, but increases the Type II error rate.
- There are other multiple-comparison methods that tare more powerful, but they typically apply only in specific cases.

Summary

- Single ANOVA is used when we have a single grouping variable with more than two groups.
- ANOVA tests whether there is a difference between means of the groups by comparing the variance within the groups, and variance of the means of the groups.
- The ratio of variances follow F distribution.
- ANOVA only tests for 'any difference', you can inspect specific differences through planned contrasts, or post-hoc comparisons.
- ANOVA is a specific case of regression.

Next week: more ANOVA. Reading: Ch. 12, 'factorial ANOVA'.