Comparing means with ANOVA

Prof. Maria Tackett

[Click here for PDF of slides]

Topics

Topics

- Compare groups using analysis of variance

Topics

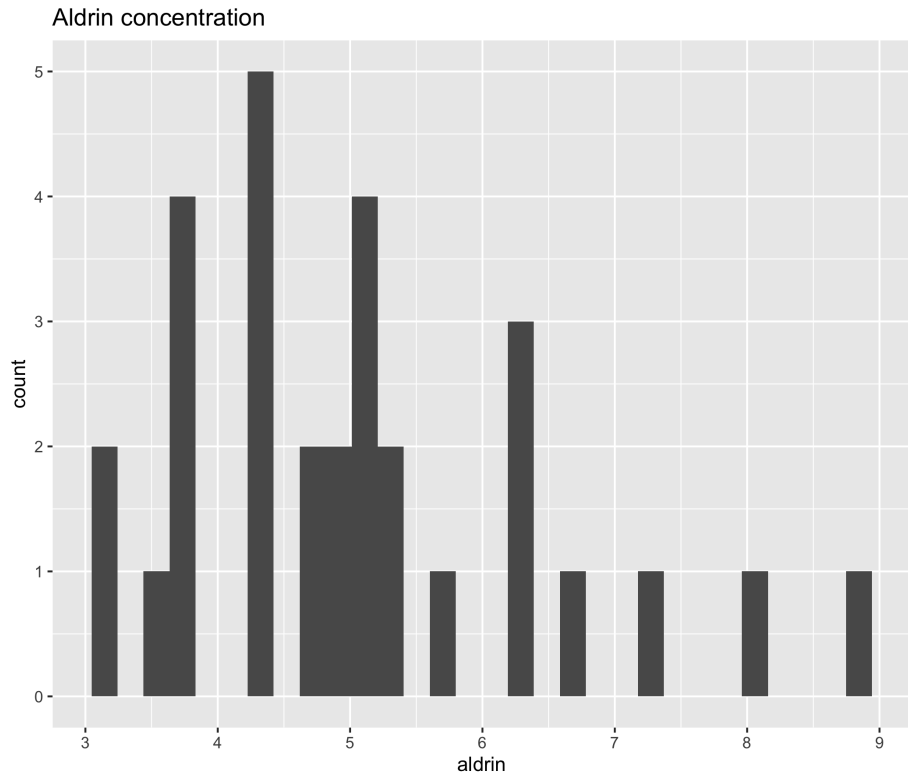
- Compare groups using analysis of variance

Aldrin in the Wolf River

- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides).
- These highly toxic organic compounds can cause various cancers and birth defects.

Aldrin in the Wolf River

```
## # A tibble: 30 x 2
##    aldrin depth
##     <dbl> <chr>
##  1    3.8 bottom
##  2    4.8 bottom
##  3    4.9 bottom
##  4    5.3 bottom
##  5    5.4 bottom
##  6    5.7 bottom
##  7    6.3 bottom
##  8    7.3 bottom
##  9    8.1 bottom
## 10    8.8 bottom
## # … with 20 more rows
```



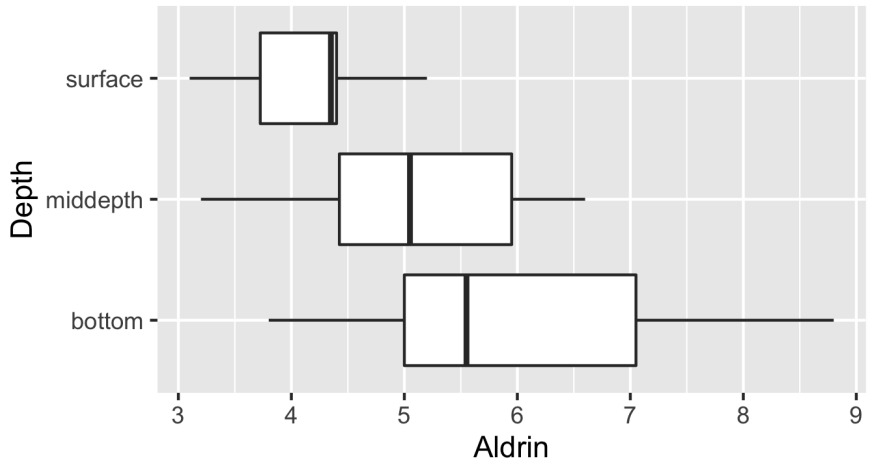
Aldrin in the Wolf River

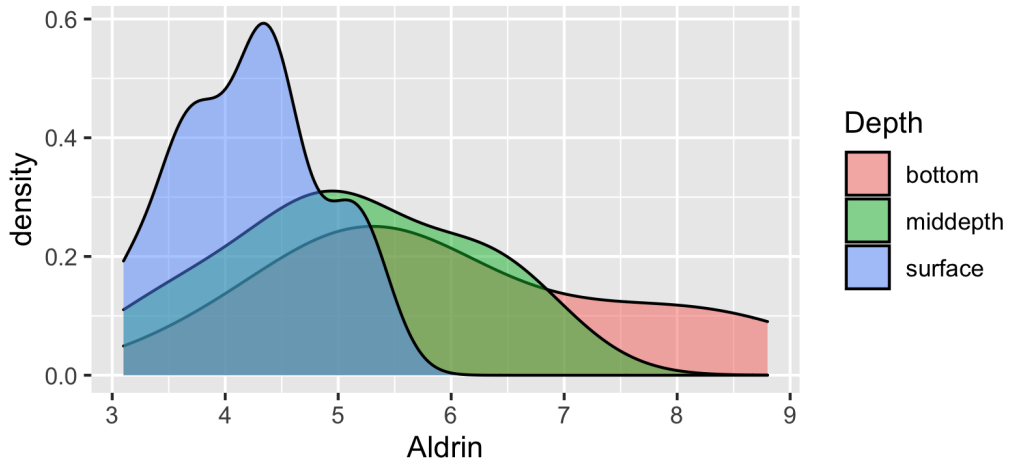
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth.

- These compounds are denser than water and their molecules tend to stick to particles of sediment, they are more likely to be found in higher concentrations near the bottom than near mid-depth.

Is there a difference between the mean aldrin concentrations among the three depth levels?

Aldrin by depth





depth	n	mean	sd
bottom	10	6.04	1.579
middepth	10	5.05	1.104
surface	10	4.20	0.660

So far, we have used a **quantitative** predictor variable to understand the variation in a quantitative response variable.

Now, we will use a **categorical (qualitative)** predictor variable to understand the variation in a quantitative response variable.

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \ldots, K$.

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \ldots, K$.
- n_i is number of observations in group i

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \ldots, K$.
- n_i is number of observations in group i
- $n = n_1 + n_2 + \cdots + n_K$ is the total number of observations in the data

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \ldots, K$.
- n_i is number of observations in group i
- $n = n_1 + n_2 + \cdots + n_K$ is the total number of observations in the data
- y_{ij} is the j^{th} observation in group i, for all i, j

Notation

- K is number of mutually exclusive groups. We index the groups as $i = 1, \ldots, K$.
- n_i is number of observations in group i
- $n = n_1 + n_2 + \cdots + n_K$ is the total number of observations in the data
- y_{ij} is the j^{th} observation in group i, for all i, j
- μ_i is the population mean for group i, for $i = 1, \ldots, K$

Using ANOVA to compare means

- **Question of interest** Is the mean value of the response y the same for all groups, or is there at least one group with a significantly different mean value?

- To answer this question, we will test the following hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$
$$H_a : \text{At least one } \mu_i \text{ is not equal to the others}$$

What's happening...

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$
$$H_a : \text{At least one } \mu_i \text{ is not equal to the others}$$

- If the sample means are "far apart", " there is evidence against H_0
- We will calculate a test statistic to quantify "far apart" in the context of the data

Analysis of Variance (ANOVA)

Main Idea: Decompose the total variation in the data into the variation between groups (model) and the variation within each group (residuals)

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{K} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

Analysis of Variance (ANOVA)

Main Idea: Decompose the total variation in the data into the variation between groups (model) and the variation within each group (residuals)

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{K} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

- If the variation between groups is significantly greater than the variation within each group, then there is evidence against the null hypothesis.

ANOVA table

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Total variation

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Total variation

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Total variation: variation between and within groups

$$SS_{Total} = 16.961 + 37.329 = 54.290$$

$$DF_{Total} = 2 + 37 = 29$$

$$s_y^2 = \frac{SS_{Total}}{DF_{Total}} = \frac{54.290}{29} = 1.872$$

Between variation

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Between variation

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Between variation: variation in the group means

$$SS_{Between} = 16.961$$

$$DF_{Between} = 2$$

$$MS_{Between} = \frac{SS_{Between}}{DF_{Between}} = \frac{15.961}{2} = 8.480$$

Within variation

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Within variation

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Within variation: variation within each group

$$SS_{Within} = 37.329$$

$$DF_{Within} = 27$$

$$MS_{Within} = \frac{SS_{Within}}{DF_{Within}} = \frac{37.329}{27} = 1.383$$

Using ANOVA table to test difference in means

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_a : At least one depth level has μ_i that is not equal to the others

Using ANOVA table to test difference in means

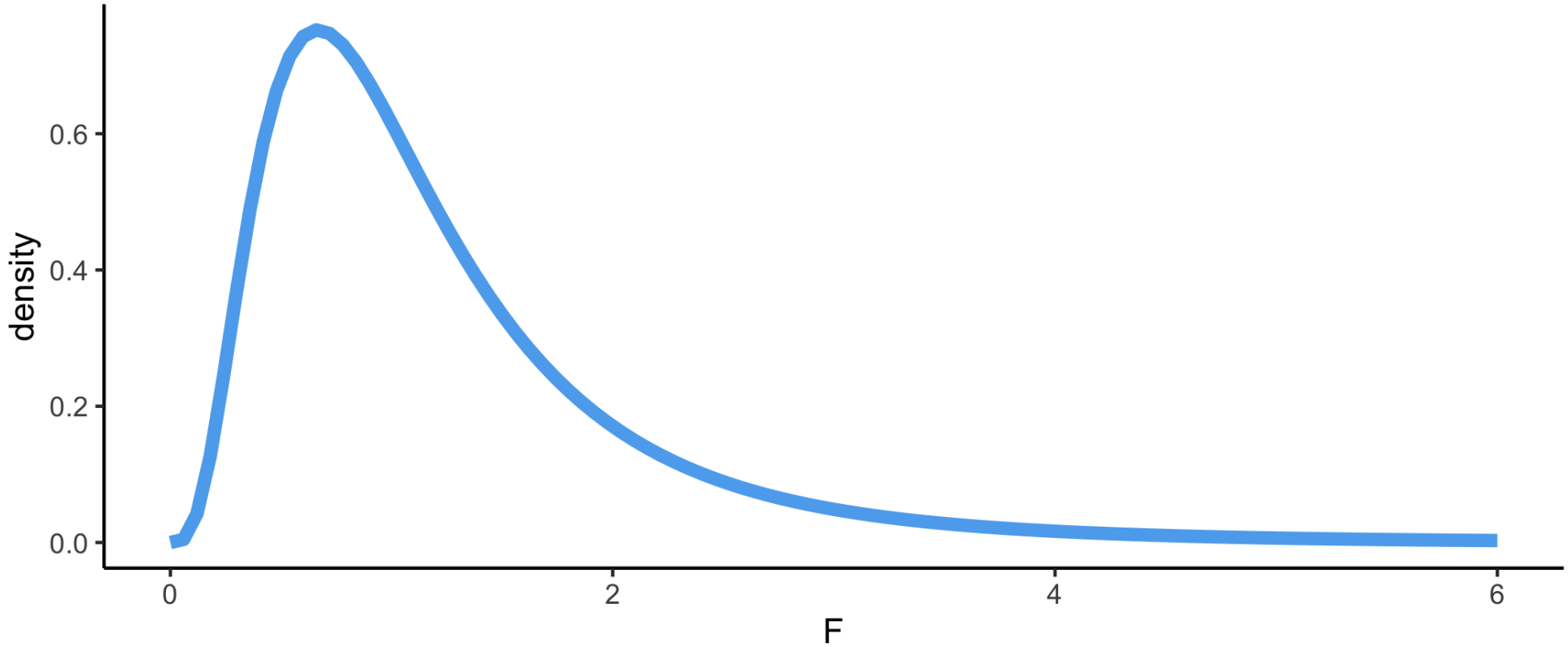
term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

Test statistic: Ratio of between group and within group variation

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{8.480}{1.383} = 6.134$$

Calculate p-value

Calculate the p-value using an F distribution with $K - 1$ and $n - K$ degrees of freedom



Using ANOVA table to test difference in means

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

P-value: Probability of observing a test statistic at least as extreme as *F Stat* given the group means are equal

Using ANOVA table to test difference in means

term	df	sumsq	meansq	statistic	p.value
depth	2	16.961	8.480	6.134	0.006
Residuals	27	37.329	1.383		

P-value: Probability of observing a test statistic at least as extreme as *F Stat* given the group means are equal

The p-value is very small (≈ 0), so we reject H_0. The data provide sufficient evidence that at least one depth level has a mean aldrin concentration that differs from the others.

Assumptions for ANOVA

Assumptions for ANOVA

Assumptions for ANOVA

1️⃣ **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$

Assumptions for ANOVA

1️⃣ **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$

2️⃣ **Constant variance:** The population distribution for each group has a common variance, σ^2

Assumptions for ANOVA

1️⃣ **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$

2️⃣ **Constant variance:** The population distribution for each group has a common variance, σ^2

3️⃣ **Independence:** The observations are independent from each other

- This applies to observations within and between groups

Assumptions for ANOVA

1 **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$

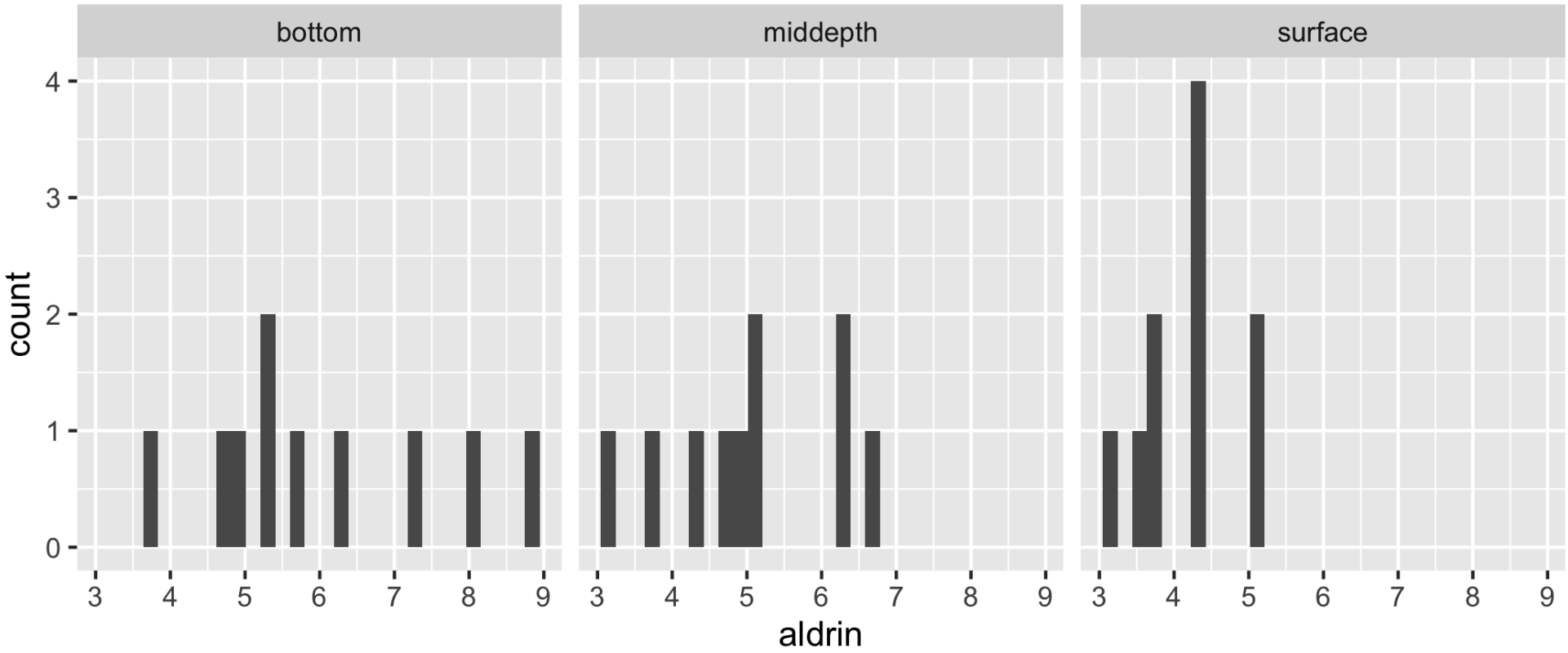
2 **Constant variance:** The population distribution for each group has a common variance, σ^2

3 **Independence:** The observations are independent from each other

- This applies to observations within and between groups

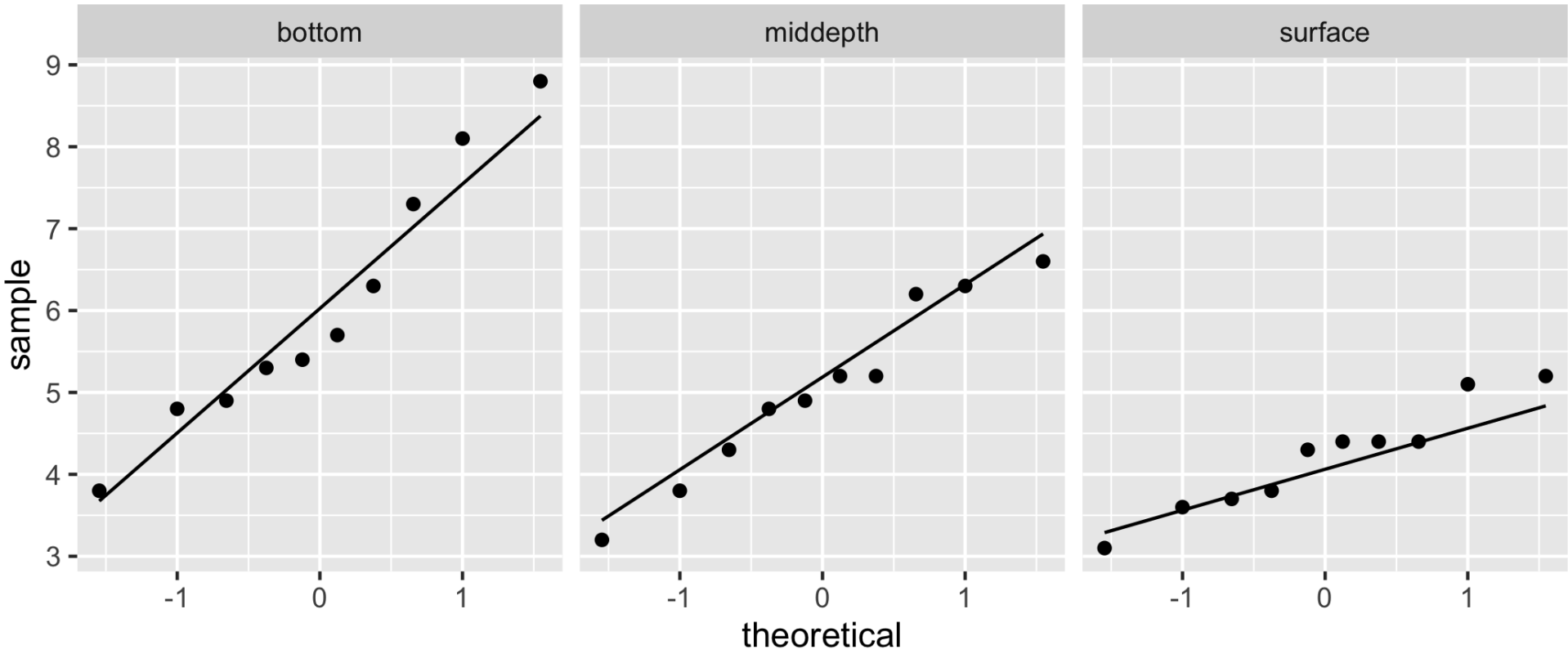
For ANOVA, we can typically check these assumptions in the exploratory data analysis

Checking Normality



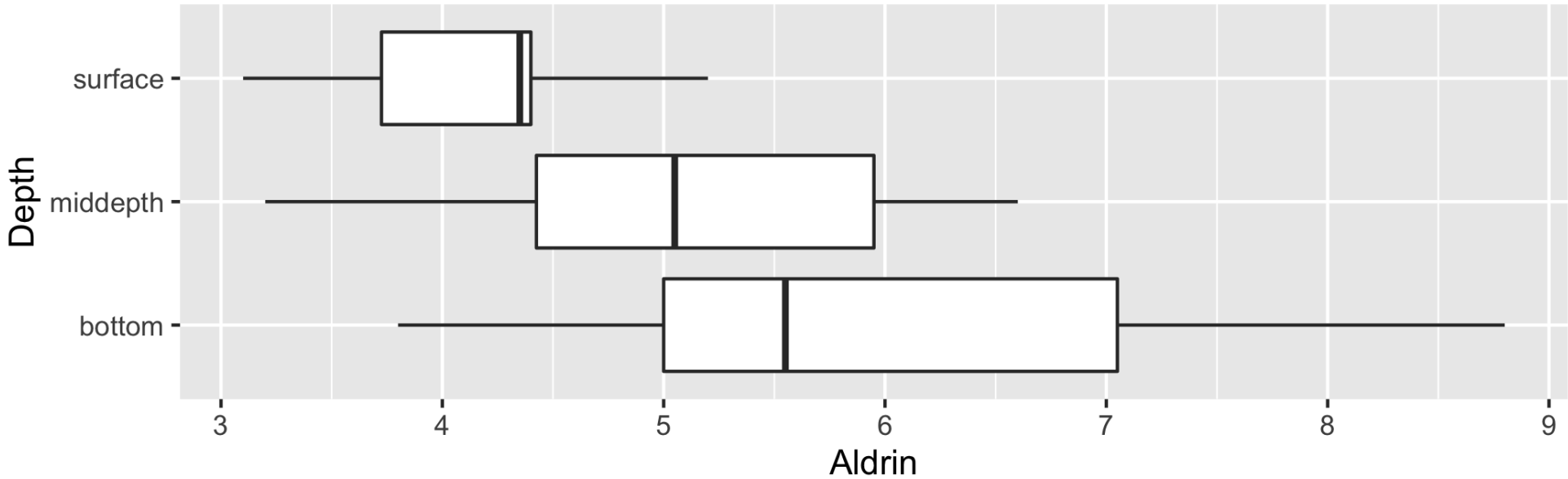
✅ No major skewness or outliers.

Checking Normality



✅ Points fall relatively along the diagonal line.

Checking constant variance




```
## # A tibble: 3 x 4
##   depth        n  mean    sd
##   <chr>    <int> <dbl> <dbl>
## 1 bottom      10  6.04 1.58
## 2 middepth    10  5.05 1.10
## 3 surface     10  4.2  0.660
```

✅ The maximum standard deviation is about 2.4 times the smallest one. This is OK given the small sample size.

Checking independence

✅ Based on what we know about the study, we have no reason to believe that the aldrin concentrations are not independent of each other.

Robustness to Assumptions

Robustness to Assumptions

- **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$
  - ANOVA relatively robust to departures from Normality.
  - Concern when there are strongly skewed distributions with different sample sizes (especially if sample sizes are small, < 10 in each group)

Robustness to Assumptions

- **Normality:** $y_{ij} \sim N(\mu_i, \sigma^2)$
  - ANOVA relatively robust to departures from Normality.
  - Concern when there are strongly skewed distributions with different sample sizes (especially if sample sizes are small, < 10 in each group)
- **Independence:** There is independence within and across groups
  - If this doesn't hold, should use methods that account for correlated errors

Robustness to Assumptions

- **Constant variance:** The population distribution for each group has a common variance, σ^2
  - Critical assumption, since the pooled (combined) variance is important for ANOVA
  - **General rule:** Satisfied if $SD_{max}/SD_{min} \leq 2$. OK if this is somewhat > 2 when sample sizes are small.

Recap

Recap

- Used ANOVA to compare means across groups

Acknowledgements

- Analysis example and map image from OpenIntro Statistics