Chapter 7: Inference for numerical data

OpenIntro Statistics, 4th Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Computing the power for a 2-sample test

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

Type 2 error rate

If the alternative hypothesis is actually true, what is the chance that we make a Type 2 Error, i.e. we fail to reject the null hypothesis even when we should reject it?

- The answer is not obvious.
- If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
- If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- Clearly, β depends on the *effect size* (δ)

Example - Blood Pressure (BP), hypotheses

Suppose a pharmaceutical company has developed a new drug for lowering blood pressure, and they are preparing a clinical trial to test the drug's effectiveness. They recruit people who are taking a particular standard blood pressure medication, and half of the subjects are given the new drug (treatment) and the other half continue to take their current medication through generic-looking pills to ensure blinding (control). What are the hypotheses for a two-sided hypothesis test in this context?

$$H_0 : \mu_{treatment} - \mu_{control} = 0$$

$$H_A : \mu_{treatment} - \mu_{control} \neq 0$$

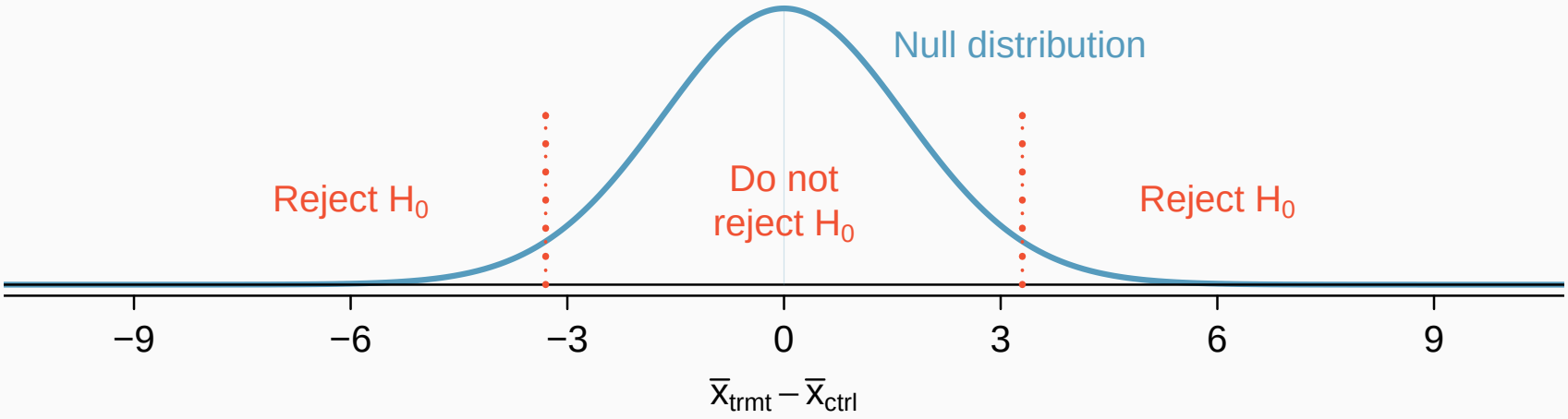
Example - BP, standard error

Suppose researchers would like to run the clinical trial on patients with systolic blood pressures between 140 and 180 mmHg. Suppose previously published studies suggest that the standard deviation of the patients' blood pressures will be about 12 mmHg and the distribution of patient blood pressures will be approximately symmetric. If we had 100 patients per group, what would be the approximate standard error for difference in sample means of the treatment and control groups?

$$SE = \sqrt{\frac{12^2}{100} + \frac{12^2}{100}} = 1.70$$

Example - BP, minimum effect size required to reject H_0

For what values of the difference between the observed averages of blood pressure in treatment and control groups (effect size) would we reject the null hypothesis at the 5% significance level?



The difference should be at least

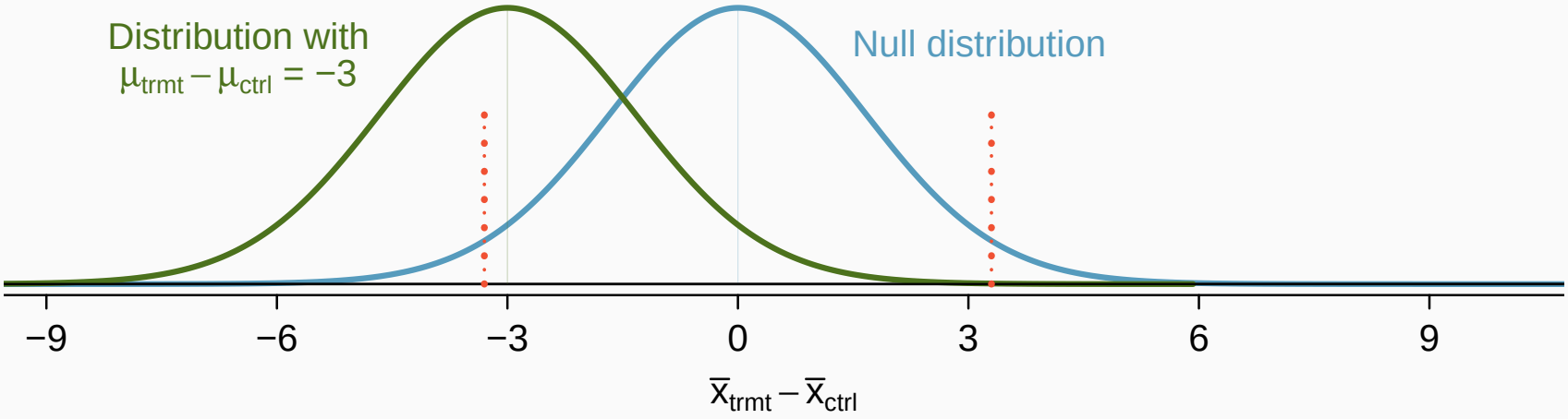
$$1.96 * 1.70 = 3.332$$

or at most

$$-1.96 * 1.70 = 3.332.$$

Example - BP, power

Suppose that the company researchers care about finding any effect on blood pressure that is 3 mmHg or larger vs the standard medication. What is the power of the test that can detect this effect?

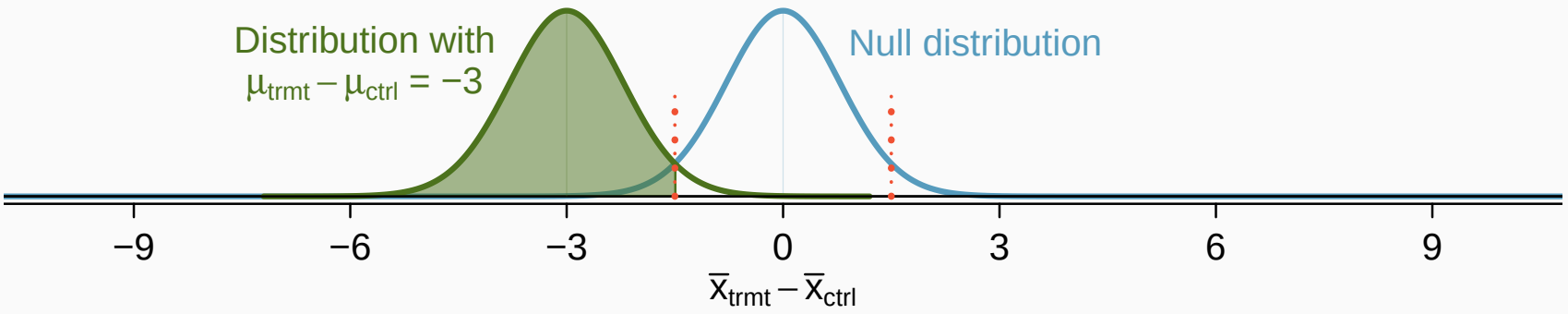


$$Z = \frac{-3.332 - (-3)}{1.70} = -0.20$$

$$P(Z < -0.20) = 0.4207$$

Example - BP, required sample size for 80% power

What sample size will lead to a power of 80% for this test?



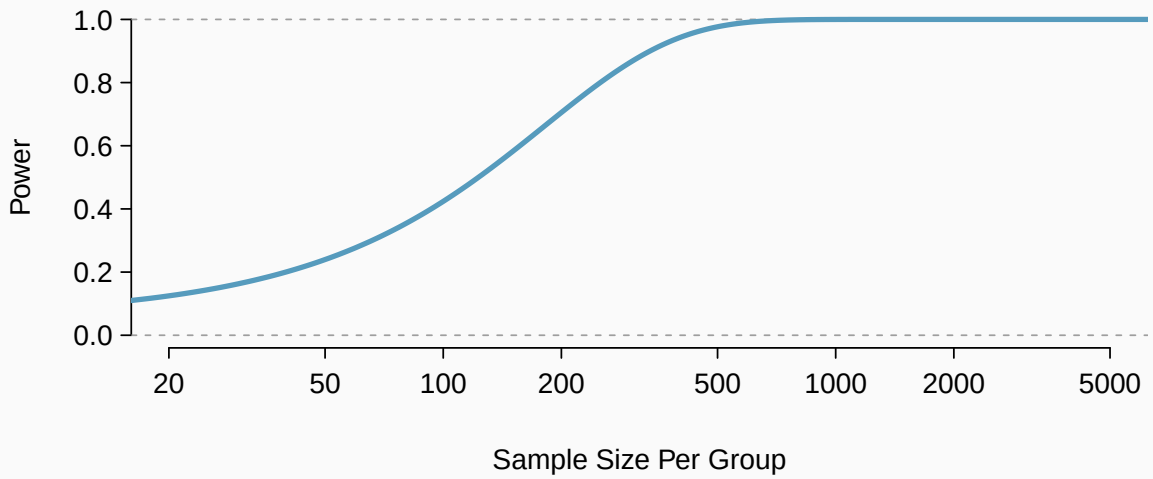
$$SE = \frac{3}{2.8} = 1.07142$$

$$1.07142 = \sqrt{\frac{12^2}{n} + \frac{12^2}{n}}$$

$$n = 250.88 \rightarrow n \geq 251$$

Recap

- Calculate required sample size for a desired level of power
- Calculate power for a range of sample sizes, then choose the sample size that yields the target power (usually 80% or 90%)



Achieving desired power

There are several ways to increase power (and hence decrease type 2 error rate):

1. Increase the sample size.
2. Decrease the standard deviation of the sample, which essentially has the same effect as increasing the sample size (it will decrease the standard error). With a smaller s we have a better chance of distinguishing the null value from the observed point estimate. This is difficult to ensure but cautious measurement process and limiting the population so that it is more homogenous may help.
3. Increase α, which will make it more likely to reject H_0 (but note that this has the side effect of increasing the Type 1 error rate).
4. Consider a larger effect size. If the true mean of the population is in the alternative hypothesis but close to the null value, it will be harder to detect a difference.

Comparing means with ANOVA

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

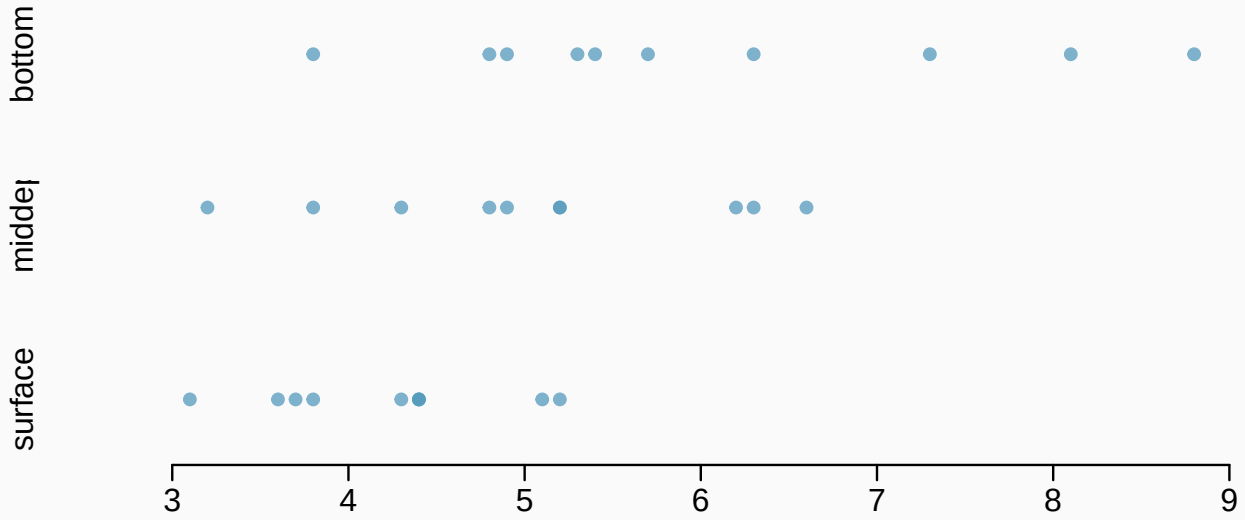
Data

Aldrin concentration (nanograms per liter) at three levels of depth.

	aldrin	depth
1	3.80	bottom
2	4.80	bottom
...		
10	8.80	bottom
11	3.20	middepth
12	3.80	middepth
...		
20	6.60	middepth
21	3.10	surface
22	3.60	surface
...		
30	5.20	surface

Exploratory analysis

Aldrin concentration (nanograms per liter) at three levels of depth.



	n	mean	sd
bottom	10	6.04	1.58
middepth	10	5.05	1.10
surface	10	4.20	0.66
overall	30	5.1 0	1.37

Research question

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

- To compare means of 2 groups we use a Z or a T statistic.
- To compare means of 3+ groups we use a new test called *ANOVA* and a new statistic called *F*.

ANOVA

ANOVA is used to assess whether the mean of the outcome variable is different for different levels of a categorical variable.

H_0 : The mean outcome is the same across all categories,

$$\mu_1 = \mu_2 = \cdots = \mu_k,$$

where μ_i represents the mean of the outcome for observations in category i.

H_A : At least one mean is different than others.

Conditions

1. The observations should be independent within and between groups
    - If the data are a simple random sample from less than 10% of the population, this condition is satisfied.
    - Carefully consider whether the data may be independent (e.g. no pairing).
    - Always important, but sometimes difficult to check.
2. The observations within each group should be nearly normal.
    - Especially important when the sample sizes are small.

How do we check for normality?

3. The variability across the groups should be about equal.
    - Especially important when the sample sizes differ between groups.

How can we check this condition?

z/t test vs. ANOVA - Purpose

z/t test

Compare means from *two* groups to see whether they are so far apart that the observed difference cannot reasonably be attributed to sampling variability.

$$H_0 : \mu_1 = \mu_2$$

ANOVA

Compare the means from *two or more* groups to see whether they are so far apart that the observed differences cannot all reasonably be attributed to sampling variability.

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

z/t test vs. ANOVA - Method

z/t test

Compute a test statistic (a ratio).

$$z/t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

Compute a test statistic (a ratio).

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$

- Large test statistics lead to small p-values.
- If the p-value is small enough H_0 is rejected, we conclude that the population means are not equal.

z/t test vs. ANOVA

- With only two groups t-test and ANOVA are equivalent, but only if we use a pooled standard variance in the denominator of the test statistic.
- With more than two groups, ANOVA compares the sample means to an overall *grand mean*.

Hypotheses

What are the correct hypotheses for testing for a difference between the mean aldrin concentrations among the three levels?

(a) $H_0 : \mu_B = \mu_M = \mu_S$

$H_A : \mu_B \neq \mu_M \neq \mu_S$

(b) $H_0 : \mu_B \neq \mu_M \neq \mu_S$

$H_A : \mu_B = \mu_M = \mu_S$

(c) $H_0 : \mu_B = \mu_M = \mu_S$

H_A : *At least one mean is different.*

(d) $H_0 : \mu_B = \mu_M = \mu_S = 0$

H_A : At least one mean is different.

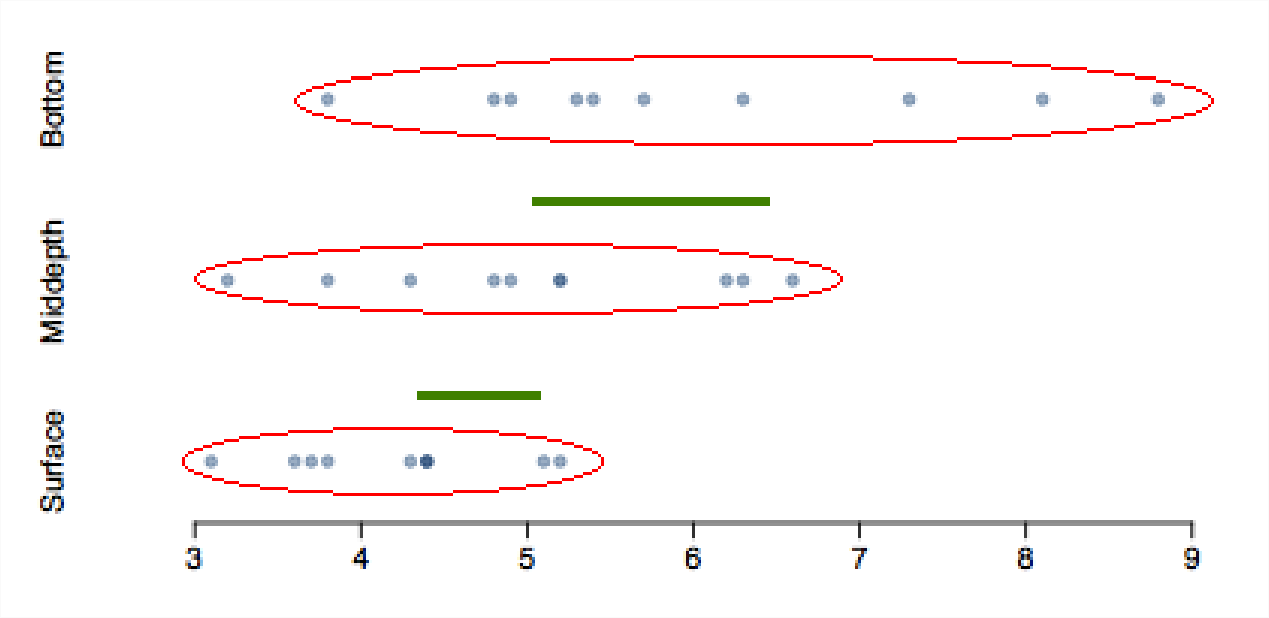
(e) $H_0 : \mu_B = \mu_M = \mu_S$

$H_A : \mu_B > \mu_M > \mu_S$

Test statistic

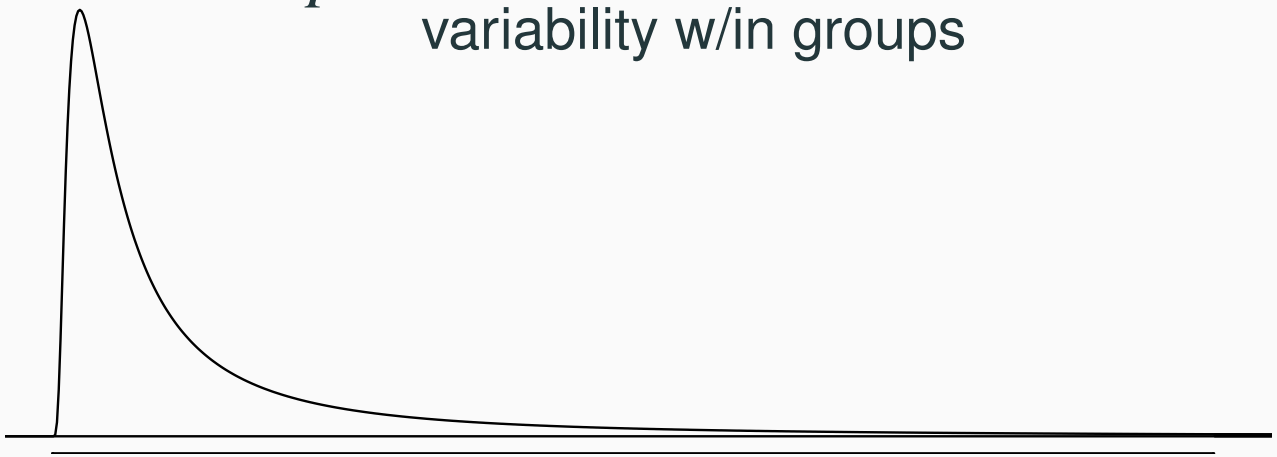
Does there appear to be a lot of variability within groups? How about between groups?

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



F distribution and p-value

$$F = \frac{\text{variability bet. groups}}{\text{variability w/in groups}}$$



- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- In order to obtain a large F statistic, variability between sample means needs to be greater than variability within sample means.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Degrees of freedom associated with ANOVA

- groups: $df_G = k - 1$, where k is the number of groups
- total: $df_T = n - 1$, where n is the total sample size
- error: $df_E = df_T - df_G$

- $df_G = k - 1 = 3 - 1 = 2$
- $df_T = n - 1 = 30 - 1 = 29$
- $df_E = 29 - 2 = 27$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	*16.96*	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	*16.96*	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	*16.96*	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$SSG = \left(10 \times (6.04 - 5.1)^2\right)$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	*16.96*	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Sum of squares between groups, SSG

Measures the variability between groups

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

where n_i is each group size, $\bar{x}_i$ is the average for each group, $\bar{x}$ is the overall (grand) mean.

	n	mean
bottom	10	6.04
middepth	10	5.05
surface	10	4.2
overall	30	5.1

$$\begin{aligned} SSG &= \left(10 \times (6.04 - 5.1)^2\right) \\ &+ \left(10 \times (5.05 - 5.1)^2\right) \\ &+ \left(10 \times (4.2 - 5.1)^2\right) \\ &= 16.96 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	*54.29*			

Sum of squares total, SST

Measures the variability between groups

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})$$

where x_i represent each observation in the dataset.

$$\begin{aligned} SST &= (3.8 - 5.1)^2 + (4.8 - 5.1)^2 + (4.9 - 5.1)^2 + \cdots + (5.2 - 5.1)^2 \\ &= (-1.3)^2 + (-0.3)^2 + (-0.2)^2 + \cdots + (0.1)^2 \\ &= 1.69 + 0.09 + 0.04 + \cdots + 0.01 \\ &= 54.29 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	6.13	0.0063
(*E*rror)	Residuals	27	*37.33*	1.38		
	*T*otal	29	54.29			

Sum of squares error, SSE

Measures the variability within groups:

$$SSE = SST - SSG$$

$$SSE = 54.29 - 16.96 = 37.33$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	*8.48*	6.13	0.0063
(*E*rror)	Residuals	27	37.33	*1.38*		
	*T*otal	29	54.29			

Mean square error

Mean square error is calculated as sum of squares divided by the degrees of freedom.

$$\begin{aligned} MSG &= 16.96/2 = 8.48 \\ MSE &= 37.33/27 = 1.38 \end{aligned}$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	*6.14*	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

Test statistic, F value

As we discussed before, the F statistic is the ratio of the between group and within group variability.

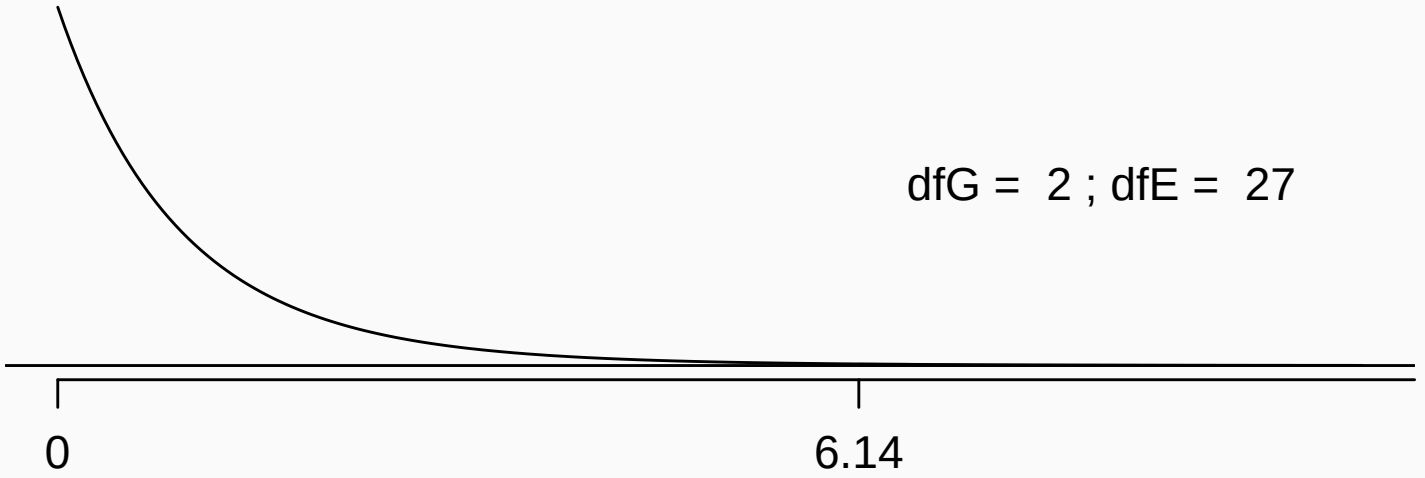
$$F = \frac{MSG}{MSE}$$

$$F = \frac{8.48}{1.38} = 6.14$$

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(*G*roup)	depth	2	16.96	8.48	*6.14*	0.0063
(*E*rror)	Residuals	27	37.33	1.38		
	*T*otal	29	54.29			

p-value

p-value is the probability of at least as large a ratio between the "between group" and "within group" variability, if in fact the means of all groups are equal. It's calculated as the area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



Conclusion - in context

What is the conclusion of the hypothesis test?

The data provide convincing evidence that the average aldrin concentration

(a) is different for all groups.

(b) on the surface is lower than the other levels.

(c) *is different for at least one group.*

(d) is the same for all groups.

Conclusion

- If p-value is small (less than α), reject H_0. The data provide convincing evidence that at least one mean is different from (but we can't tell which one).
- If p-value is large, fail to reject H_0. The data do not provide convincing evidence that at least one pair of means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

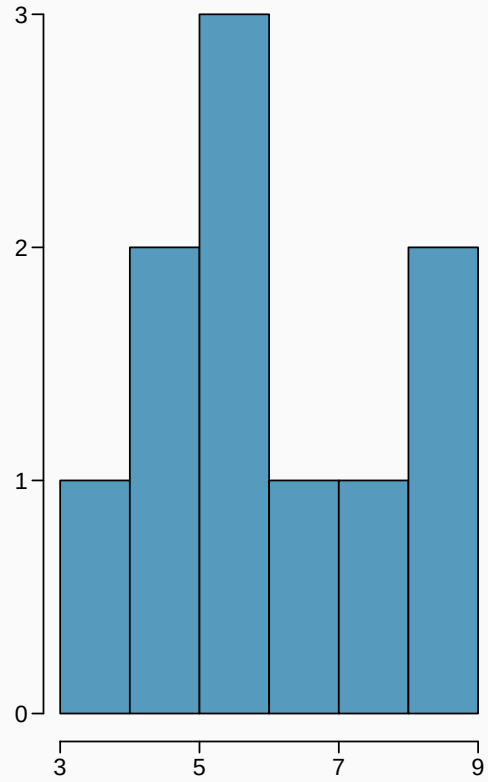
(1) independence

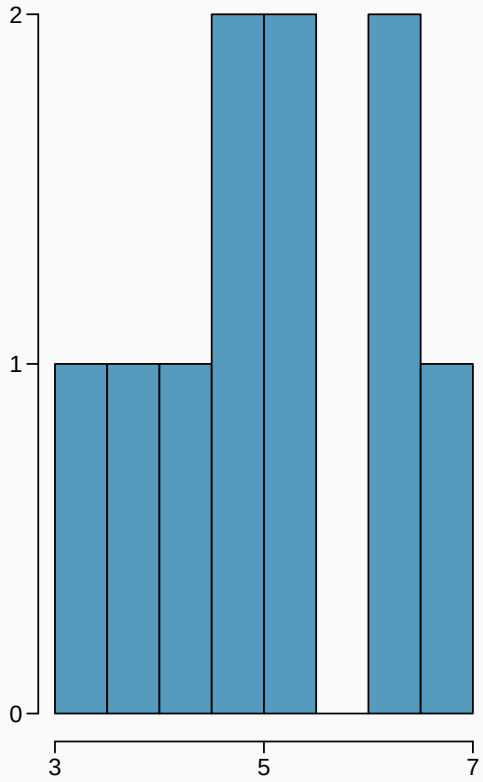
Does this condition appear to be satisfied?

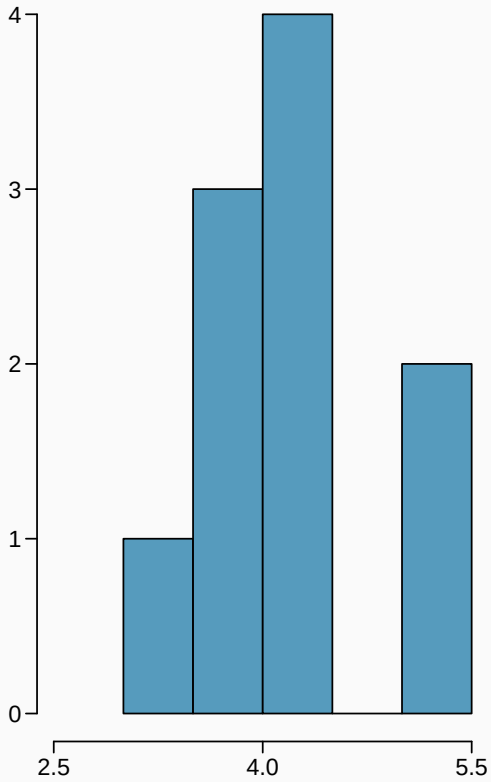
In this study the we have no reason to believe that the aldrin concentration won't be independent of each other.

(2) approximately normal

Does this condition appear to be satisfied?

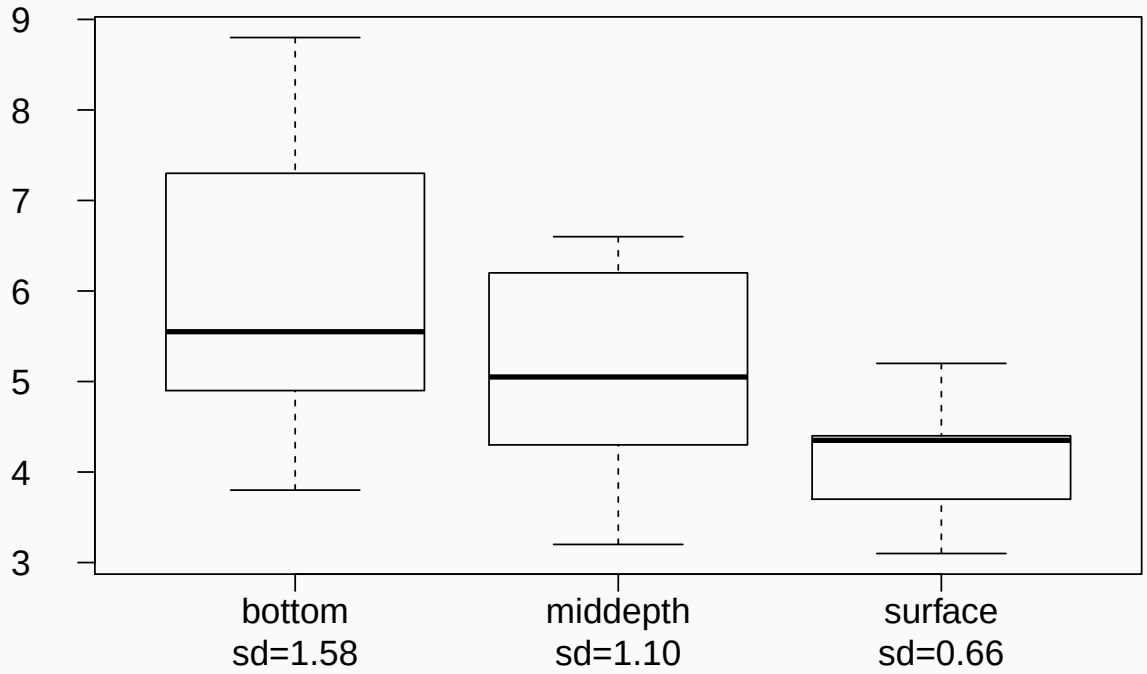






(3) constant variance

Does this condition appear to be satisfied?



Which means differ?

- Earlier we concluded that at least one pair of means differ. The natural question that follows is “which ones?”
- We can do two sample t tests for differences in each possible pair of groups.

Can you see any pitfalls with this approach?

- When we run too many tests, the Type 1 Error rate increases.
- This issue is resolved by using a modified significance level.

Multiple comparisons

- The scenario of testing many pairs of groups is called *multiple comparisons*.
- The *Bonferroni correction* suggests that a more *stringent* significance level is more appropriate for these tests:

$$\alpha^\star = \alpha / K$$

  where K is the number of comparisons being considered.
- If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Determining the modified α

In the aldrin data set depth has 3 levels: bottom, mid-depth, and surface. If $\alpha = 0.05$, what should be the modified significance level for two sample t tests for determining which pairs of groups have significantly different means?

(a) $\alpha^* = 0.05$

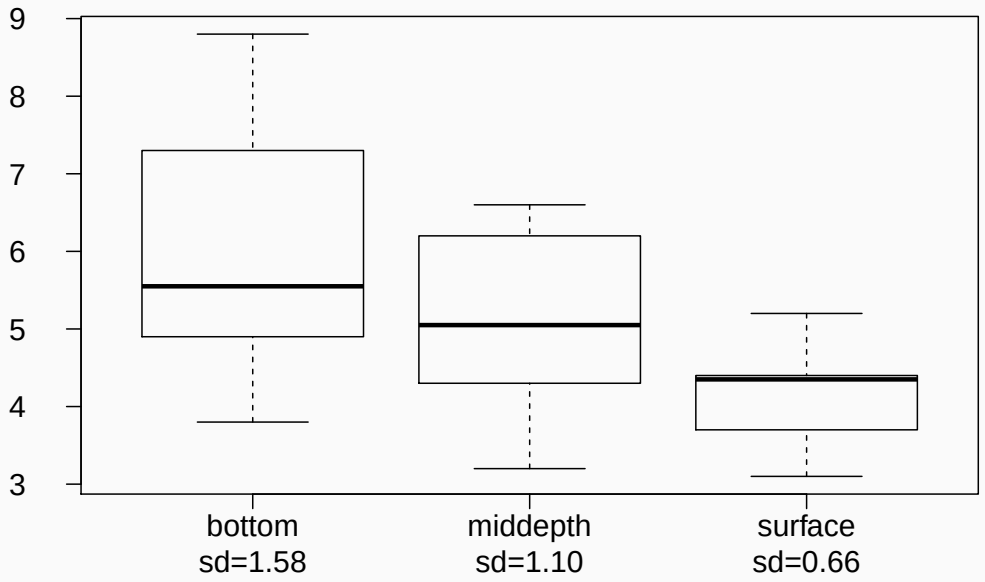
(b) $\alpha^* = 0.05/2 = 0.025$

(c) $\alpha^* = 0.05/3 = 0.0167$

(d) $\alpha^* = 0.05/6 = 0.0083$

Which means differ?

Based on the box plots below, which means would you expect to be significantly different?



(a) bottom & surface

(b) bottom & mid-depth

(c) mid-depth & surface

(d) bottom & mid-depth; mid-depth & surface

(e) bottom & mid-depth; bottom & surface; mid-depth & surface

Which means differ? (cont.)

If the ANOVA assumption of equal variability across groups is satisfied, we can use the data from all groups to estimate variability:

- Estimate any within-group standard deviation with $\sqrt{MSE}$, which is s_{pooled}
- Use the error degrees of freedom, $n - k$, for t-distributions

Difference in two means: after ANOVA

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

Is there a difference between the average aldrin concentration at the bottom and at mid depth?

	n	mean	sd
bottom	10	*6.04*	1.58
middepth	10	*5.05*	1.10
surface	10	4.2	0.66
overall	30	5.1	1.37

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
depth	2	16.96	8.48	6.13	0.0063
Residuals	*27*	37.33	*1.38*		
Total	29	54.29			

$$T_{df_E} = \frac{(\bar{x}_{bottom} - \bar{x}_{middepth})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{middepth}}}}$$

$$T_{27} = \frac{(6.04 - 5.05)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{0.99}{0.53} = 1.87$$

$$0.05 < p - value < 0.10 \quad \text{(two-sided)}$$

$$\alpha^\star = 0.05/3 = 0.0167$$

Fail to reject H_0, the data do not provide convincing evidence of a difference between the average aldrin concentrations at bottom and mid depth.

Pairwise comparisons

Is there a difference between the average aldrin concentration at the bottom and at surface?

$$
\begin{aligned}
T_{df_E} &= \frac{(\bar{x}_{bottom} - \bar{x}_{surface})}{\sqrt{\frac{MSE}{n_{bottom}} + \frac{MSE}{n_{surface}}}} \\
T_{27} &= \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81 \\
p - value &< 0.01 \qquad \textit{(two-sided)} \\
\alpha^\star &= 0.05/3 = 0.0167
\end{aligned}
$$

Reject H_0, the data provide convincing evidence of a difference between the average aldrin concentrations at bottom and surface.