Lecture 21: Small Sample Inference, One-way ANOVA

- Small sample inference for a proportion
- Small sample inference for proportions
- One-way ANOVA

Famous predictors

Before this guy...

There was this guy...

Paul the Octopus - psychic?

- Paul the Octopus (26 January 2008 –26 October 2010) predicted 8 World Cup games, and predicted them all correctly
- Does this provide convincing evidence that Paul actually has psychic powers?
- How unusual would this be if he was just randomly guessing (with a 50% chance of guessing correctly)?
- Hypotheses:

  $H_0: \ p = 0.5$
  $H_A: \ p > 0.5$

Conditions

1. *Independence:* We can assume that each guess is independent of another.
2. *Sample size:* The number of expected successes is *smaller than 10*.

$$8 \times 0.5 = 4$$

So what do we do?

Since the sample size isn't large enough to use CLT based methods, we use a simulation method instead.

Which of the following methods is best way to calculate the p-value of the hypothesis test evaluating if Paul the Octopus' predictions are unusually higher than random guessing?

(a) Flip a coin 8 times, record the proportion of times where all 8 tosses were heads. Repeat this many times, and calculate the proportion of simulations where all 8 tosses were heads.

(b) Roll a die 8 times, record the proportion of times where all 8 rolls were 6s. Repeat this many times, and calculate the proportion of simulations where all 8 rolls were 6s.

(c) Flip a coin 10,000 times, record the proportion of heads. Repeat this many times, and calculate the proportion of simulations where more than 50% of tosses are heads.

(d) Flip a coin 10,000 times, calculate the proportion of heads.

Which of the following methods is best way to calculate the p-value of the hypothesis test evaluating if Paul the Octopus' predictions are unusually higher than random guessing?

(a) *Flip a coin 8 times, record the proportion of times where all 8 tosses were heads. Repeat this many times, and calculate the proportion of simulations where all 8 tosses were heads.*

(b) Roll a die 8 times, record the proportion of times where all 8 rolls were 6s. Repeat this many times, and calculate the proportion of simulations where all 8 rolls were 6s.

(c) Flip a coin 10,000 times, record the proportion of heads. Repeat this many times, and calculate the proportion of simulations where more than 50% of tosses are heads.

(d) Flip a coin 10,000 times, calculate the proportion of heads.

Simulate

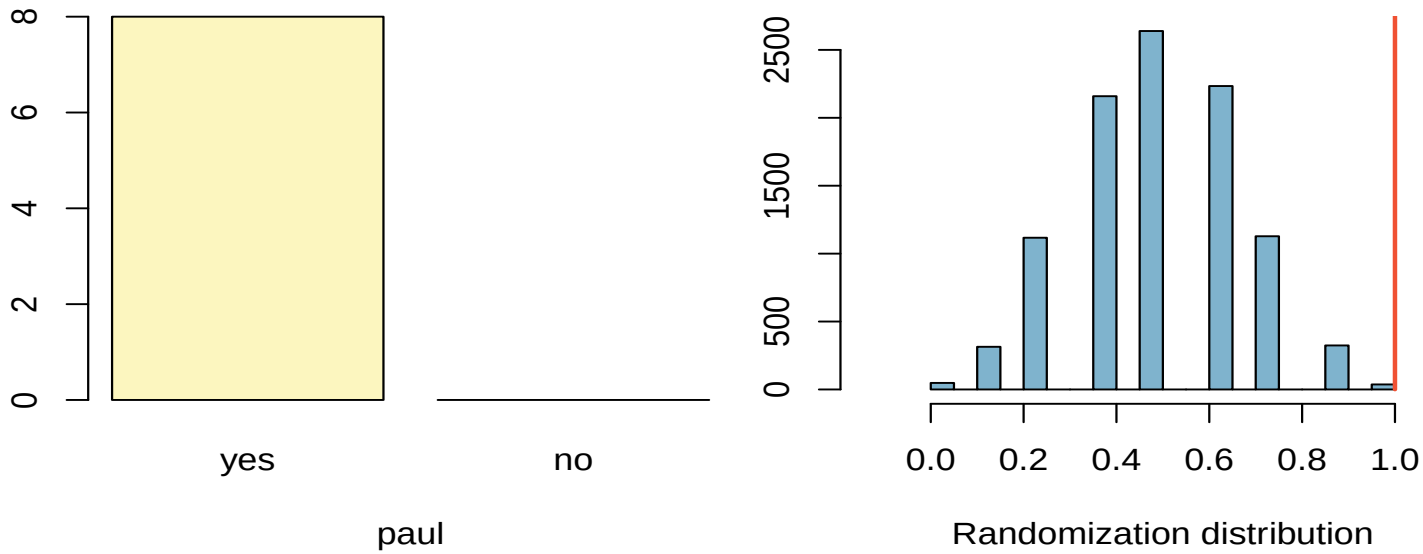
Flip a coin 8 times. Did you get all heads?

(a) Yes

(b) No

```
source("http://www.openintro.org/stat/slides/inference.R")
paul = factor(c(rep("yes", 8), rep("no", 0)), levels = c("yes","no"))
inference(paul, est = "proportion", type = "ht", method = "simulation",
          success = "yes", null = 0.5, alternative = "greater", seed = 290)
```

```
Single proportion -- success: yes
Summary statistics: p_hat = 1 ;  n = 8
H0: p = 0.5
HA: p > 0.5
p-value =  0.0037
```



Conclusions

Which of the following is false?

(a) If in fact Paul was randomly guessing, the probability that he would get the result of all 8 games correct is 0.0037.

(b) Reject H_0, the data provide convincing evidence that Paul did better than randomly guessing.

(c) We may have made a Type I error.

(d) The probability that Paul is psychic is 0.0037.

Conclusions

Which of the following is false?

(a) If in fact Paul was randomly guessing, the probability that he would get the result of all 8 games correct is 0.0037.

(b) Reject H_0, the data provide convincing evidence that Paul did better than randomly guessing.

(c) We may have made a Type I error.

(d) *The probability that Paul is psychic is 0.0037.*

Back of the hand

There is a saying “know something like the back of your hand". Describe an experiment to test if people really do know the backs of their hands.

In the MythBusters episode, 11 out of 12 people guesses the backs of their hands correctly.

Hypotheses

What are the hypotheses for evaluating if people are capable of recognizing the back of their hand at a rate that is better than random guessing. Remember, in the MythBusters experiment, there were 10 pictures to choose from, and only 1 was correct.

H_0 : $p = 0.10$ (random guessing)

H_A : $p > 0.10$ (better than random guessing)

Conditions

1. *Independence:* We can assume that each person guessing is independent of another.
2. *Sample size:* The number of expected successes is *smaller than 10*.

$$12 \times 0.1 = 1.2$$

So what do we do?

Since the sample size isn't large enough to use CLT based methods, we use a simulation method instead.

Simulation scheme

Describe how you test if results of this experiment to determine if people are capable of recognizing the back of their hand at a rate that is better than random guessing.

$$H_0 : p = 0.10 \qquad H_A : p > 0.10 \qquad \hat{p} = 11/12 = 0.9167$$

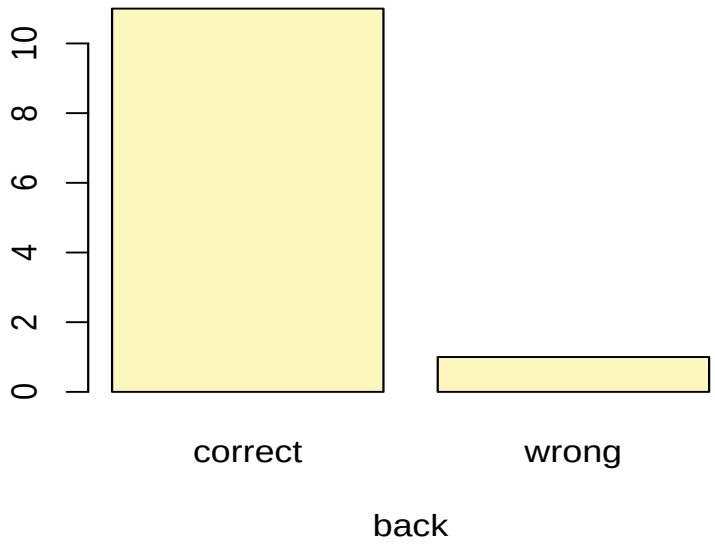
1. Use a 10-sided fair die to represent the sampling space, and call 1 a success (guessing correctly), and all other outcomes failures (guessing incorrectly).
2. Roll the die 12 times (representing 12 people in the experiment), count the number of 1s, and calculate the proportion of correct guesses in one simulation of 12 rolls.
3. Repeat step (2) many times, each time recording the proportion of successes in a series of 12 rolls of the die.
4. Create a dot plot of the simulated proportions from step (3) and count the number of simulations where the proportion was at least as high as 0.9167 (the observed proportion).

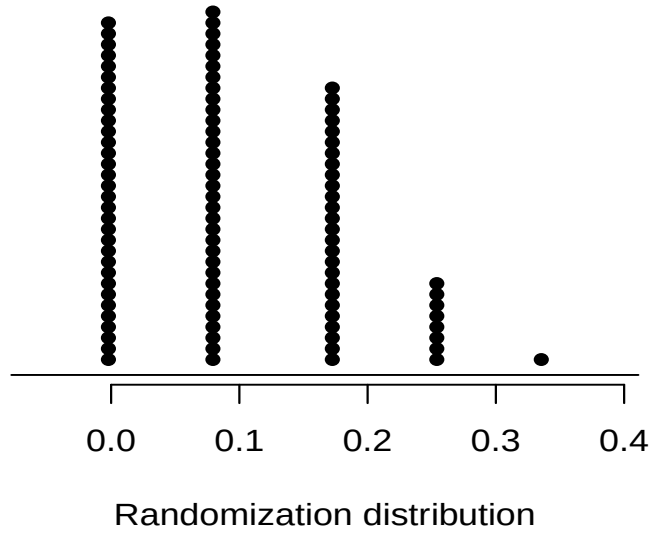
Simulation results

- In the next slide you can see the results of a hypothesis test (using only 100 simulations to keep things simple).
- Each dot represents a simulation proportion of success. There were 25-30 simulations where the success rate ($\hat{p}$) was 10%, 40-45 simulations where the success rate was slightly less than 10%, about 20 simulations where the success rate was slightly less than 20% and 1 simulation where the success rate was more than 30%.
- There are no simulations where the success rate is as high as the observed success rate of 91.67%.
- Therefore we conclude that the observed result is near impossible to have happened by chance (p-value = 0).
- And hence that these data suggest that people are capable of recognizing the back of their hand at a rate that is better than random guessing.

```
back = as.factor(c(rep("correct", 11), rep("wrong", 1)))
inference(back, est = "proportion", type = "ht", method = "simulation",
        success = "correct", null = 0.1, alternative = "greater", seed = 654, nsim = 100)
```

```
Single proportion -- success: correct
Summary statistics: p_hat = 0.9167 ;  n = 12
H0: p = 0.1
HA: p > 0.1
p-value =  0
```





Comparing back of the hand to palm of the hand

MythBusters also asked these people to guess the palms of their hands. This time 7 out of the 12 people guesses correctly. The data are summarized below.

	Palm	Back	Total
Correct	11	7	18
Wrong	1	5	6
Total	12	12	24

Proportion of correct guesses

	Palm	Back	Total
Correct	11	7	18
Wrong	1	5	6
Total	12	12	24

- Proportion of correct in the back group: $\frac{11}{12} = 0.916$
- Proportion of correct in the palm group: $\frac{7}{12} = 0.583$
- Difference: 33.3% more correct in the back of the hand group.

Based on the proportions we calculated, do you think the chance of guessing the back of the hand correctly is different than palm of the hand?

Hypotheses

What are the hypotheses for comparing if the proportion of people who can guess the backs of their hands correctly is different than the proportion of people who can guess the palm of their hands correctly?

H_0: $p_{back} = p_{palm}$

H_A: $p_{back} \neq p_{palm}$

Conditions?

- Independence - within groups, between groups?
  - Within each group we can assume that the guess of one subject is independent of another.
  - Between groups independence is not satisfied - we have the same people guessing.
- Sample size?
  - $\hat{p}_{pool} = \frac{11+7}{12+12} = \frac{18}{24} = 0.75$
  - Expected successes in back group: $12 \times 0.75 = 9$, failures = 3
  - Expected successes in palm group: $12 \times 0.75 = 9$, failures = 3

Since independence and S/F condition fails, we need to use simulation to compare the proportions.

Simulation scheme

1. Use 24 index cards, where each card represents a subject.
2. Mark 18 of the cards as “correct" and the remaining 6 as “wrong".
3. Shuffle the cards and split into two groups of size 12, for back and palm.
4. Calculate the difference between the proportions of “correct" in the back and palm decks, and record this number.
5. Repeat steps (3) and (4) many times to build a randomization distribution of differences in simulated proportions.

Interpreting the simulation results

When simulating the experiment under the assumption of independence, i.e. leaving things up to chance.

If results from the simulations based on the *null model* look like the data, then we can determine that the difference between the proportions correct guesses in the two groups was simply *due to chance*.

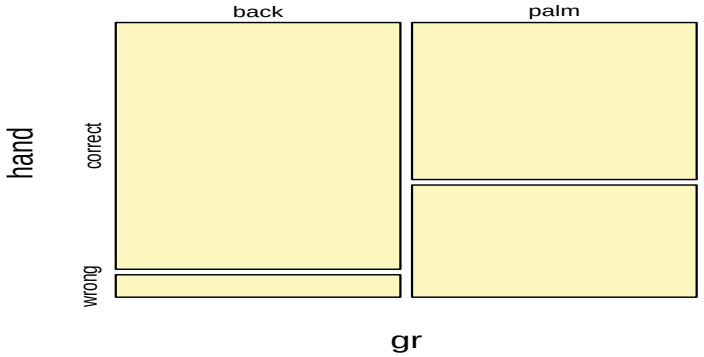
If the results from the simulations based on the *null model* do not look like the data, then we can determine that the difference between the proportions correct guesses in the two groups was not due to chance, but *because people actually know the backs of their hands better*.

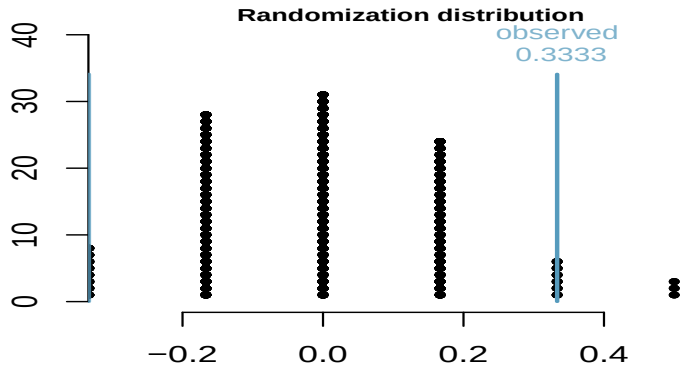
Simulation results

- In the next slide you can see the result of a hypothesis test (using only 100 simulations to keep the results simple).
- Each dot represents a difference in simulated proportion of successes. We can see that the distribution is centered at 0 (the null value).
- We can also see that 9 out of the 100 simulations yielded simulated differences at least as large as the observed difference (p-value = 0.09).

```
hand = as.factor(c(rep("correct", 7), rep("wrong", 5), c(rep("correct", 11), rep("wrong", 1))))
gr = c(rep("palm",12),rep("back",12))
inference(hand, gr, est = "proportion", type = "ht", null = 0, alternative = "twosided",
        order = c("back","palm"), success = "correct", method = "simulation", seed = 879,
        nsim = 100)
```

```
Response variable: categorical, Explanatory variable: categorical
Difference between two proportions -- success: correct
Summary statistics:
         x
y         back palm Sum
  correct   11    7  18
  wrong      1    5   6
  Sum       12   12  24
Observed difference between proportions (back-palm) = 0.3333
H0: p_back - p_palm = 0
HA: p_back - p_palm != 0
p-value =  0.18
```





Conclusion

Do the simulation results suggest that people know the backs of their hands significantly better?
(Remember: There were 33.3% more correct in the back group in the observed data.)

(a) Yes

(b) No

p-value = 0.09 > 0.05, fail to reject H_0. The data do not provide convincing evidence that people know the backs of their hands better than the palms of their hands.

Conclusion

Do the simulation results suggest that people know the backs of their hands significantly better?
(Remember: There were 33.3% more correct in the back group in the observed data.)

(a) Yes

(b) *No*

p-value = 0.09 > 0.05, fail to reject H_0. The data do not provide convincing evidence that people know the backs of their hands better than the palms of their hands.

Randomization for contingency tables

Simulation scheme

- Create a randomized contingency table under the *null hypothesis*, then compute a chi-square test statistic χ^2_{sim}
- Repeat this many times and examine the distribution of these simulated test statistics –*null distribution*
- As before, we can use the upper tail of this null distribution to calculate the p-value.

Remark

- This randomization approach is valid for any sized sample, especially for cases where one or more expected cell counts do not meet the minimum threshold of 5
- When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution

ANOVA

1. Comparing means of many different groups with ANOVA (analysis of variance)
2. ANOVA compares between group variation to within group variation
3. To identify which means are different, use t-tests and the Bonferroni correction
    1. Use a modified significance level in multiple comparisons
    2. The pooled standard deviation estimate from ANOVA

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

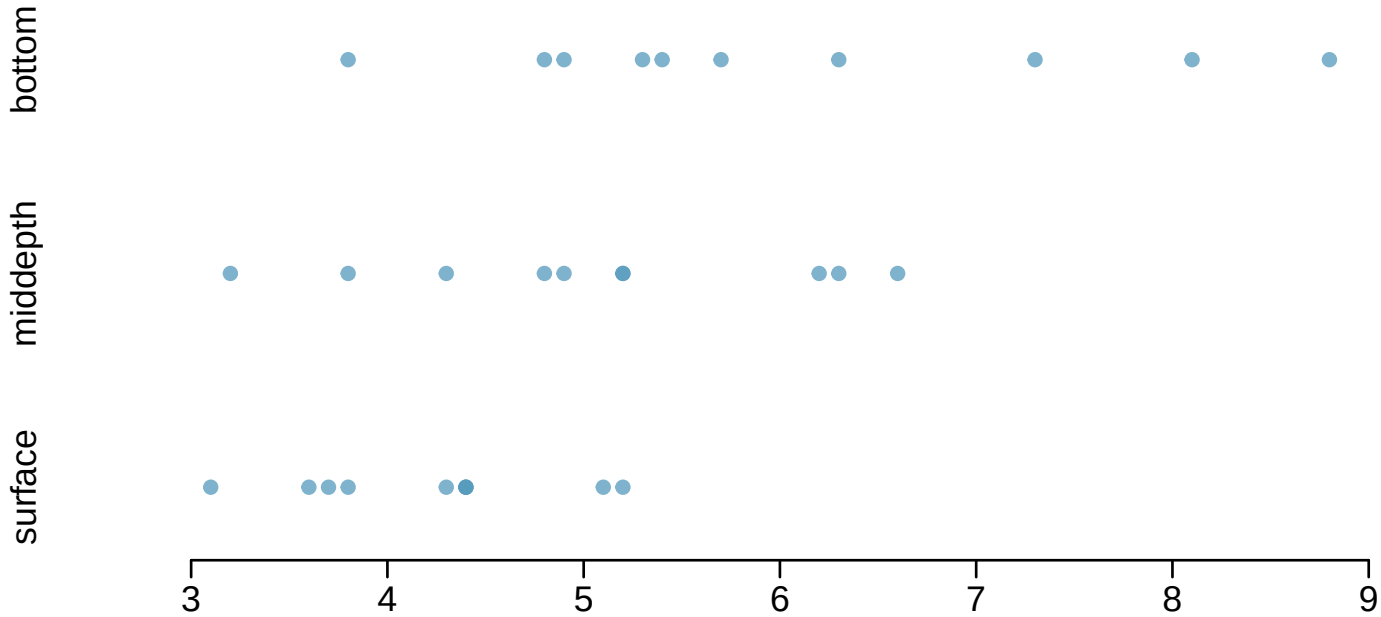
Data

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	5.10	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.

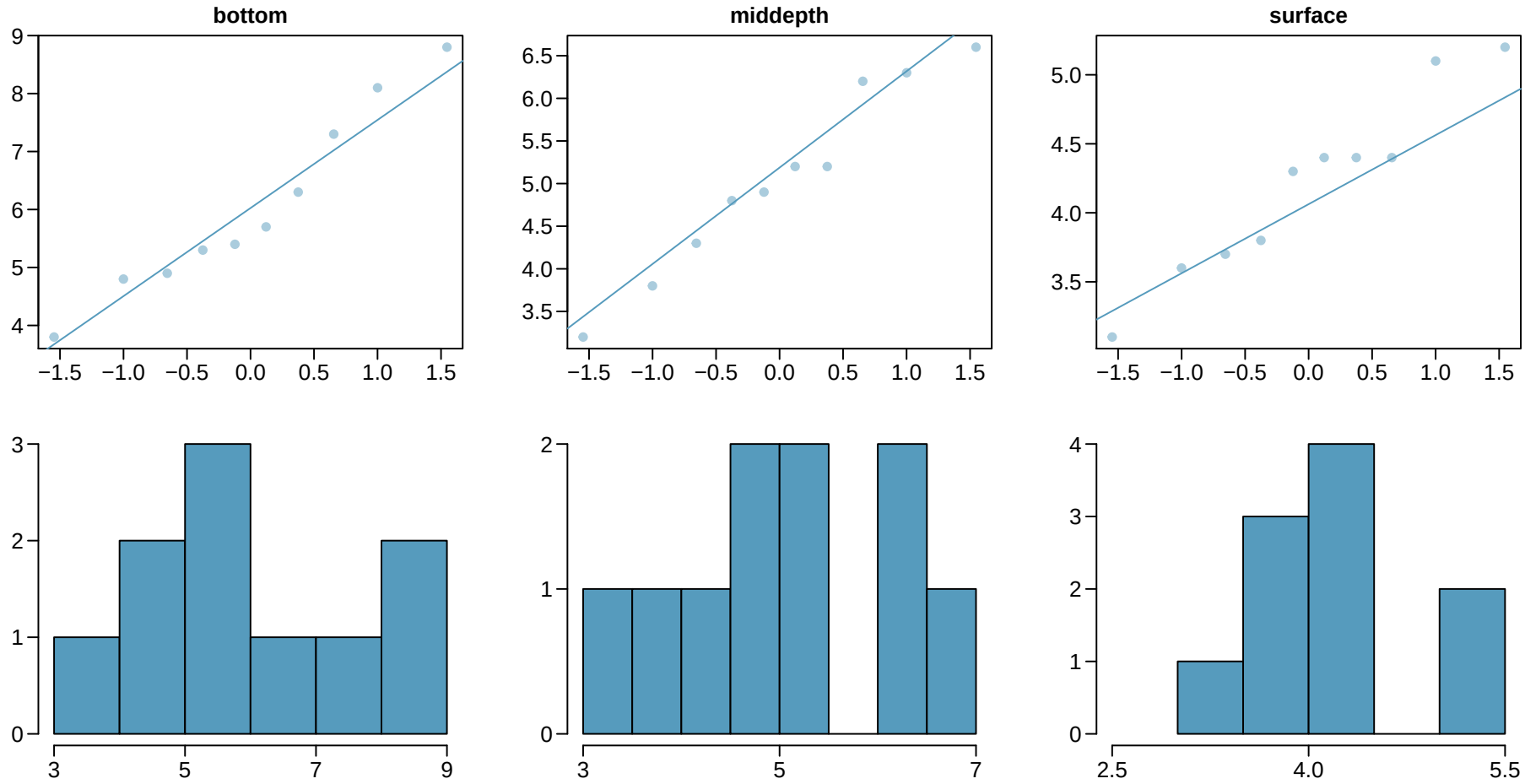
H_A : At least one mean is different than others.

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?

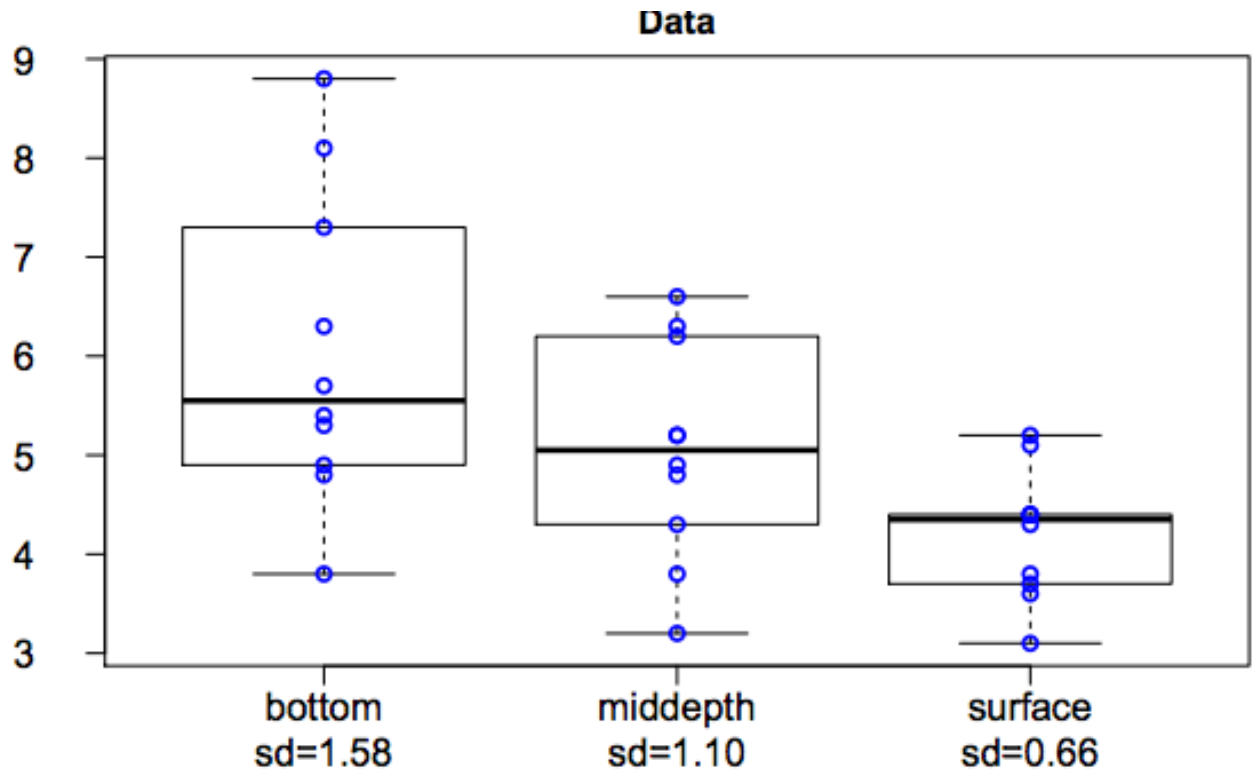
Checking conditions

Does the "approximately normal" condition appear to be satisfied?



Checking conditions

Does the "constant variance" condition appear to be satisfied?



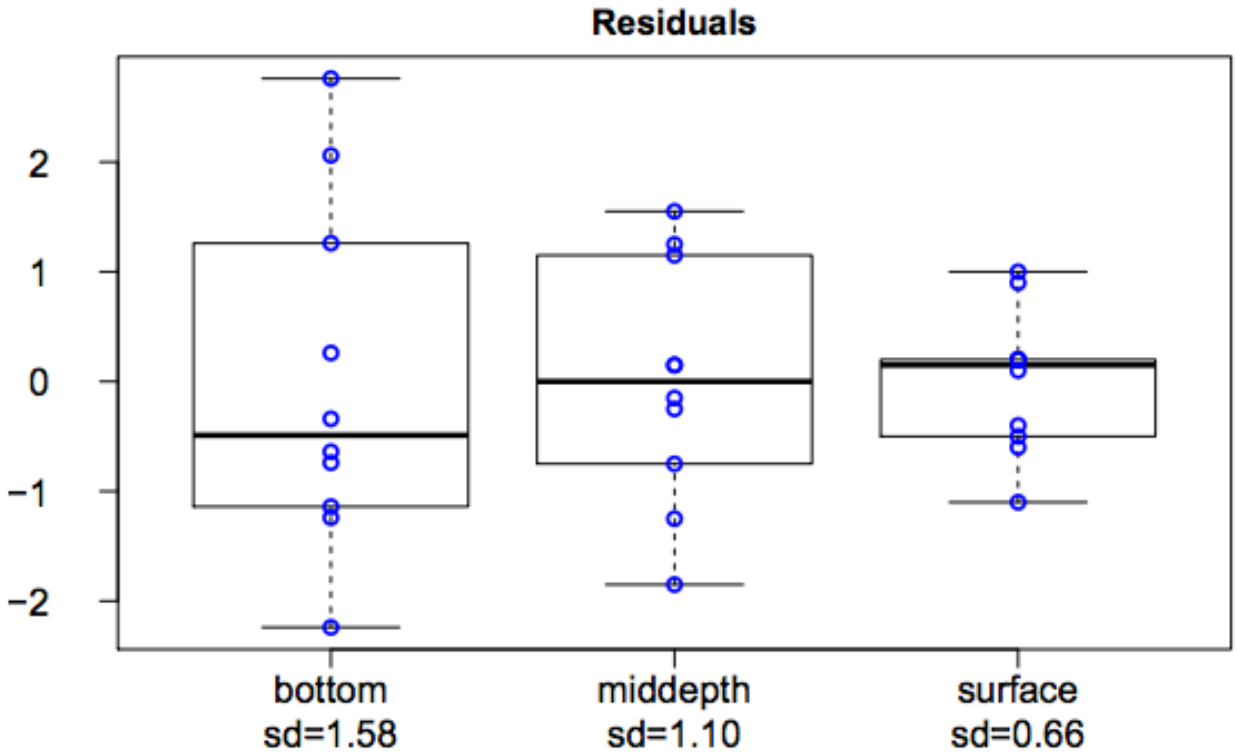
In this case it is somewhat hard to tell since the means are different.

Checking conditions

One of the ways to think about each data point is as follows:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where ϵ_{ij} is called the residual ($\epsilon_{ij} = y_{ij} - \mu_i$)



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) $H_0 : \mu_B = \mu_M = \mu_S$
$H_A : \mu_B > \mu_M > \mu_S$

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.*

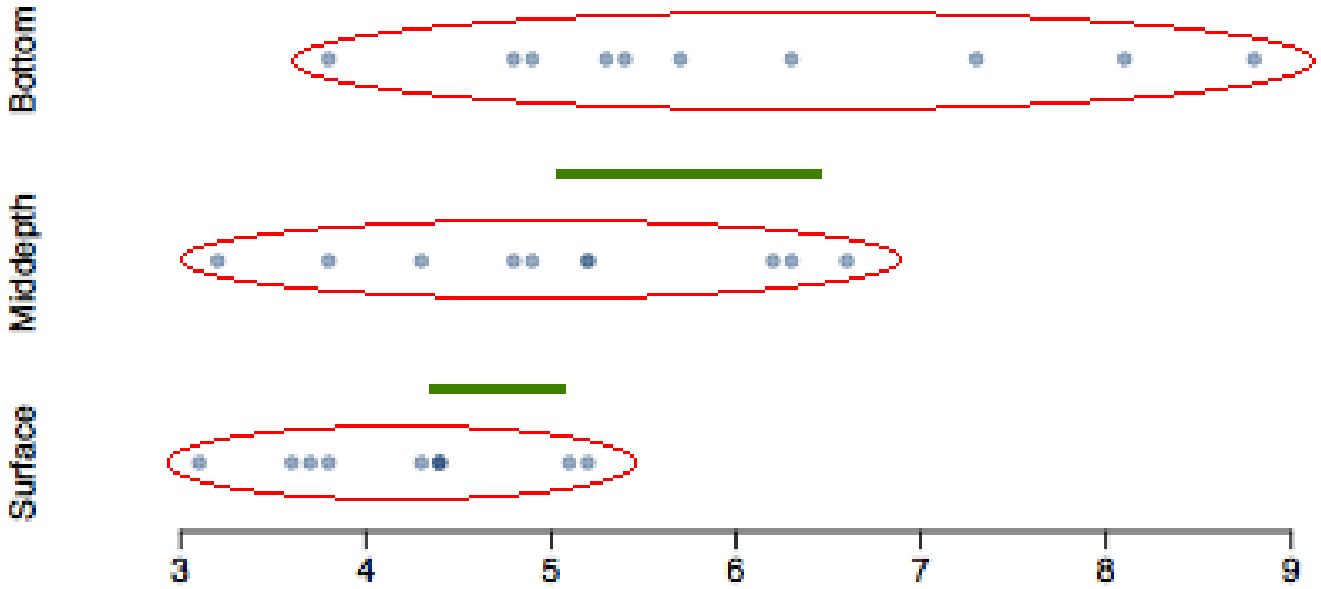
(d) $H_0 : \mu_B = \mu_M = \mu_S = 0$
H_A : At least one mean is different.

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

Test statistic

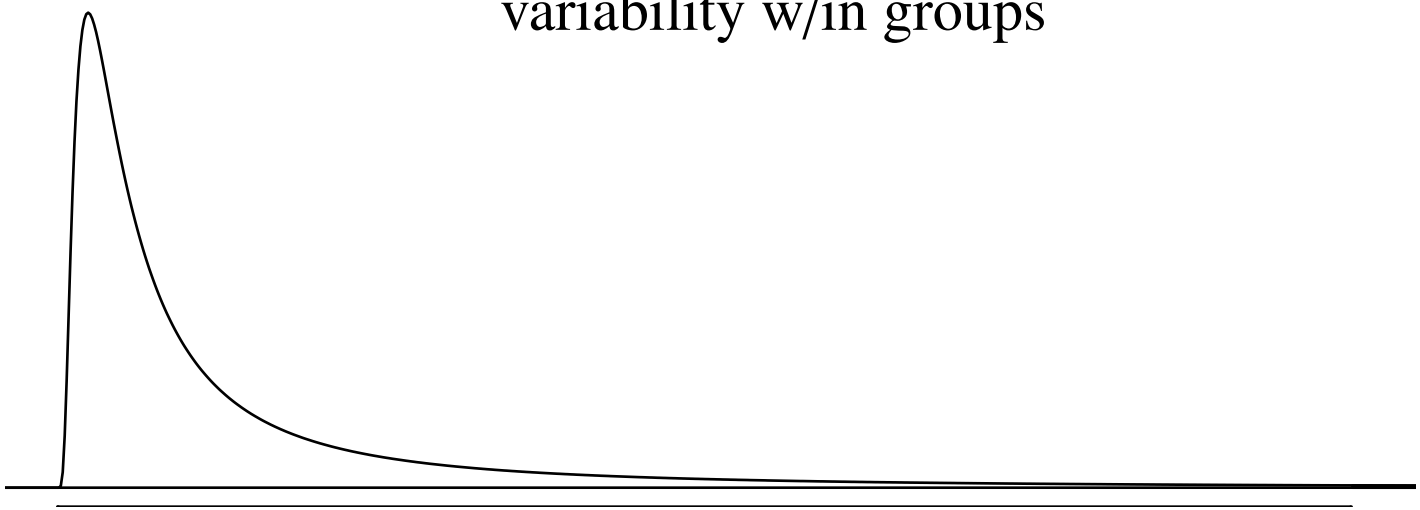
Does there appear to be a lot of variability within groups? How about between groups?

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



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.

Types of Variability

For ANOVA we think of our variability (uncertainty) in terms of three separate quantities:

- **Total variability** — all of the variability in the data, ignoring any explanatory variable(s). (You can think of this as being analogous to the sample variance of all the data)
- **Group variability** — variability between the group means and the grand mean
- **Error variability** — the sum of the variability within each group. (You can think of this as being analogous to the sum of sample variance for each group or the sum of variance of the residuals)

Sum of Squares and Variability

Mathematically, we can think of the unnormalized measure of variability as follows:

- **Total variability** — Sum of Squares Total

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_.)^2 = \text{Var}(Y_{ij})$$

- **Group variability** — Sums of Squares Group

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mu_i-\mu_.)^2 = \sum_{i=1}^{k} n_i(\mu_i-\mu_.)^2$$

- **Error variability** — Sum of Squares Error

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2 = \sum_{i=1}^{k} \text{Var}(Y_{i.})$$

Partitioning Sums of Squares

With a little bit of careful algebra we can show that:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\mu_{.})^2 = \sum_{i=1}^{k} n_i(\mu_i-\mu_{.})^2 + \sum_{i=1}^{k}\text{Var}(Y_{i.})$$

Total Variability $=$ Group Variability (between) $+$ Error Variability (between)

Sum of Squares Total $=$ Sum of Squares Group $+$ Sum of Squares Error

ANOVA Output

Includes these measures of uncertainty as well as the calculation of the F test statistic.

		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.

	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

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

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.

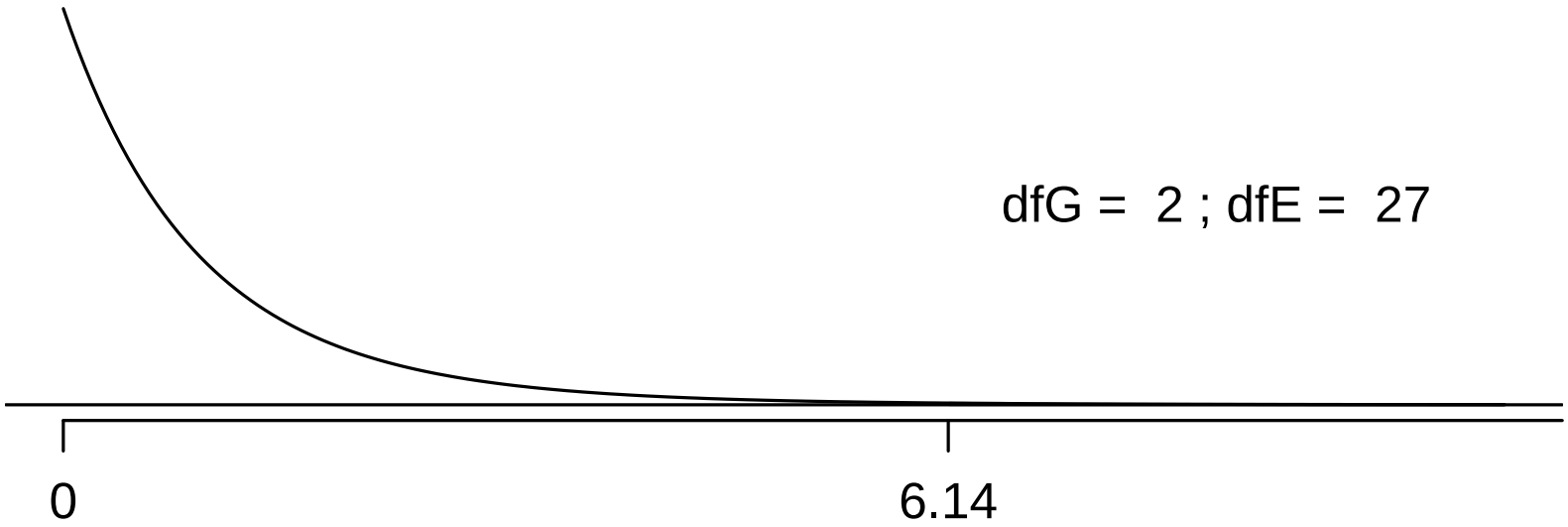
$$F = \frac{MSG}{MSE}$$

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

		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 - 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).

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

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) $\alpha^* = 0.05$

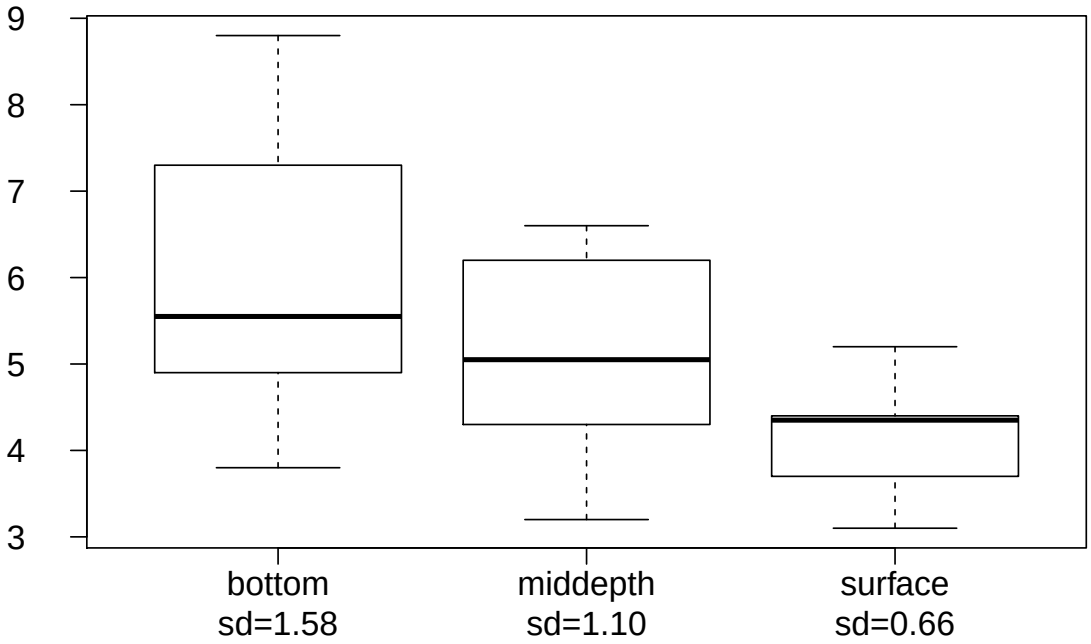
(b) $\alpha^* = 0.05/2 = 0.025$

(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 make the t-distribution approach slightly more precise by using a *pooled standard deviation*:

- The pooled standard deviation is a way to use data from all groups to better estimate the standard deviation from each group
- By pooling all the data, we can use a larger degree of freedom for the t-distribution
- Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ if the standard deviations of the groups are equal

Pooled standard deviation estimate from ANOVA

- The standard deviation of each group is estimated as $s_{pooled} = \sqrt{MSE}$
- Use the error degrees of freedom, $n - k$, for t-distributions
- The standard error of test statistic

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

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			

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

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			

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

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

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			

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

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

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

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			

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

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

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

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

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.

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

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

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

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

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

$$T_{27} = \frac{(6.04 - 4.02)}{\sqrt{\frac{1.38}{10} + \frac{1.38}{10}}} = \frac{2.02}{0.53} = 3.81$$

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

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

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

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

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

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

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

Recap

Today we learned about small sample inference based on simulation, as the epilogue for inference for categorical data, and one-way ANOVA, as a prelude of linear regression.

Suggested reading:

- D.S. Sec. 9.7, 11.6
- OpenIntro3: 5.5, 6.5, 6.6