Class 30

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-11-18

Analysis Tools (So Far)

- 1-sample proportion- or Z-test for a single proportion

Analysis Tools (So Far)

- 1-sample proportion- or Z-test for a single proportion
- 2-sample proportion- or Z-test for the difference between 2 proportions

Analysis Tools (So Far)

- 1-sample proportion- or Z-test for a single proportion
- 2-sample proportion- or Z-test for the difference between 2 proportions
- Chi-squared test for independence between 2 categorical variables

Analysis Tools (So Far)

- 1-sample proportion- or Z-test for a single proportion
- 2-sample proportion- or Z-test for the difference between 2 proportions
- Chi-squared test for independence between 2 categorical variables
- 1 sample t-test for a single mean

Analysis Tools (So Far)

- 1-sample proportion- or Z-test for a single proportion
- 2-sample proportion- or Z-test for the difference between 2 proportions
- Chi-squared test for independence between 2 categorical variables
- 1 sample t-test for a single mean
- Paired means t-test for the difference between 2 means from same unit

Analysis Tools (So Far)

- 1-sample proportion- or Z-test for a single proportion
- 2-sample proportion- or Z-test for the difference between 2 proportions
- Chi-squared test for independence between 2 categorical variables
- 1 sample t-test for a single mean
- Paired means t-test for the difference between 2 means from same unit
- 2-sample t-test for the difference between 2 unpaired means

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means
- Pearson correlation test for dependence between 2 numeric variables

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means
- Pearson correlation test for dependence between 2 numeric variables
- Linear regression for dependence between 1 or more explanatory variables (numeric or categorical) and a numerical or binary (0/1) response variable (if time)

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means
- Pearson correlation test for dependence between 2 numeric variables
- Linear regression for dependence between 1 or more explanatory variables (numeric or categorical) and a numerical or binary (0/1) response variable (if time)
- Developing a research question with a testable hypothesis

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means
- Pearson correlation test for dependence between 2 numeric variables
- Linear regression for dependence between 1 or more explanatory variables (numeric or categorical) and a numerical or binary (0/1) response variable (if time)
- Developing a research question with a testable hypothesis
- Communicating statistical methods and analysis results

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means
- Pearson correlation test for dependence between 2 numeric variables
- Linear regression for dependence between 1 or more explanatory variables (numeric or categorical) and a numerical or binary (0/1) response variable (if time)
- Developing a research question with a testable hypothesis
- Communicating statistical methods and analysis results
- Data visualization tips & tricks, do's & dont's

Remaining Tools

- ANOVA test for the difference between 3+ unpaired means
- Pearson correlation test for dependence between 2 numeric variables
- Linear regression for dependence between 1 or more explanatory variables (numeric or categorical) and a numerical or binary (0/1) response variable (if time)
- Developing a research question with a testable hypothesis
- Communicating statistical methods and analysis results
- Data visualization tips & tricks, do's & dont's
- Statistical analysis best practice

ANOVA and the F-Test for Comparing 3+ Means

- The ANOVA test (***An***alysis ***o***f ***Va***riance) tests for a difference between the means μ_i of k groups ($k \geq 3$).

ANOVA and the F-Test for Comparing 3+ Means

- The ANOVA test (***An**alysis **o**f **Va**riance*) tests for a difference between the means μ_i of k groups ($k \geq 3$).
- Compares the ratio of the ***between-group*** variability in means to what would be expected based on the ***within-group*** variability in the means.

$$\text{test statistic} = \frac{\text{Between-Groups Variability/Error}}{\text{Within-Groups Variability/Error}}$$

ANOVA and the F-Test for Comparing 3+ Means

- The ANOVA test (***An**alysis **o**f **Va**riance*) tests for a difference between the means μ_i of k groups ($k \geq 3$).
- Compares the ratio of the ***between-group*** variability in means to what would be expected based on the ***within-group*** variability in the means.

$$\text{test statistic} = \frac{\text{Between-Groups Variability/Error}}{\text{Within-Groups Variability/Error}}$$

- The probability of the observed difference in between-group variability vs within-group variability is calculated using the F distribution.

ANOVA and the F-Test for Comparing 3+ Means

- The ANOVA test (***An****alysis* ***o****f* ***Va****riance*) tests for a difference between the means μ_i of k groups ($k \geq 3$).
- Compares the ratio of the ***between-group*** variability in means to what would be expected based on the ***within-group*** variability in the means.

$$\text{test statistic} = \frac{\text{Between-Groups Variability/Error}}{\text{Within-Groups Variability/Error}}$$

- The probability of the observed difference in between-group variability vs within-group variability is calculated using the F distribution.
- Rarely calculated by hand. We will perform these exclusively with R.

ANOVA F-Test Hypotheses

- Null hypothesis: the mean outcome μ_i is the same across all k groups, such that each group has the same population mean μ.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$$

ANOVA F-Test Hypotheses

- Null hypothesis: the mean outcome μ_i is the same across all k groups, such that each group has the same population mean μ.

$$H_0: \mu_1 = \mu_2 = ... = \mu_k = \mu$$

- Alternate hypothesis: at least one mean μ_i is different from the other $k-1$ means, such that there is no single population mean μ.

$$H_A: \text{At least 1 } \mu_i \neq \mu$$

Assumptions

1. ***Independent & Identically Distributed Observations***: Observations are independent both within and between groups and identically distributed within groups.

Assumptions

1. ***Independent & Identically Distributed Observations***: Observations are independent both within and between groups and identically distributed within groups.
2. ***Sample size***: There are more than 30 observations in each of the k groups ($n_i \geq 30$).

Assumptions

1. ***Independent & Identically Distributed Observations***: Observations are independent both within and between groups and identically distributed within groups.
2. ***Sample size***: There are more than 30 observations in each of the k groups ($n_i \geq 30$).
3. ***Normality***: Especially when n_i are small, data within each of the k groups is normally distributed. This condition relaxes as n_i increases ($n_i \to \infty$).

Assumptions

1. ***Independent & Identically Distributed Observations***: Observations are independent both within and between groups and identically distributed within groups.
2. ***Sample size***: There are more than 30 observations in each of the k groups ($n_i \geq 30$).
3. ***Normality***: Especially when n_i are small, data within each of the k groups is normally distributed. This condition relaxes as n_i increases ($n_i \to \infty$).
4. ***Equal variance***: Within-group variance is approximately equal across the k groups. This condition relaxes as sample sizes n_i become more balanced between the k groups ($\frac{n}{k} \to n_i$).

The F-Distribution

- 2 degree of freedom parameters
  - The number of groups k minus 1: $\text{df}_1 = k - 1$
  - The number of observations n minus the number of groups k: $\text{df}_2 = n - k$

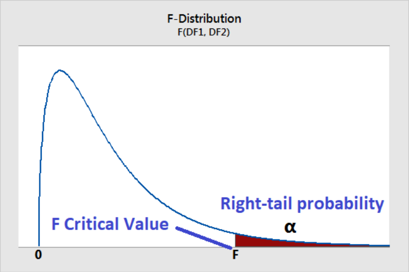
The F-Distribution

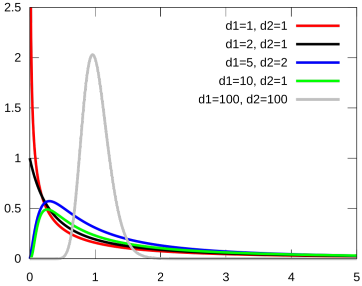
- 2 degree of freedom parameters
  - The number of groups k minus 1: $\mathrm{df}_1 = k - 1$
  - The number of observations n minus the number of groups k: $\mathrm{df}_2 = n - k$
- The larger the test statistic F, the less likely the observed data is under the null hypothesis

The F-Distribution

- 2 degree of freedom parameters
  - The number of groups k minus 1: $\text{df}_1 = k - 1$
  - The number of observations n minus the number of groups k: $\text{df}_2 = n - k$
- The larger the test statistic F, the less likely the observed data is under the null hypothesis
- Like the Chi-squared (χ^2) distribution, the upper tail probability is always used for hypothesis testing

The F-Distribution





ANOVA Results Table

Source	df	SS	MS	F	p
Between Groups (Factor)	$k-1$	$\sum_k n_k(\bar{x}_k - \bar{x}.)^2$	$\frac{SS_{Between}}{df_{Between}}$	$\frac{MS_{Between}}{MS_{Within}}$	Area to the right of $F_{k-1,\, n-k}$
Within Groups (Error)	$n-k$	$\sum_k \sum_i (x_{ik} - \bar{x}_k)^2$	$\frac{SS_{Within}}{df_{Within}}$		
Total	$n-1$	$\sum_k \sum_i (x_{ik} - \bar{x}.)^2$			

Legend

k	Number of groups
n	Total sample size (all groups combined)
n_k	Sample size of group k
$\bar{x}_k$	Sample mean of group k
$\bar{x}.$	Grand mean (i.e., mean for all groups combined)
SS	Sum of squares
MS	Mean square
df	Degrees of freedom
F	F-ratio (the test statistic)

Example: Wolf River Sediments

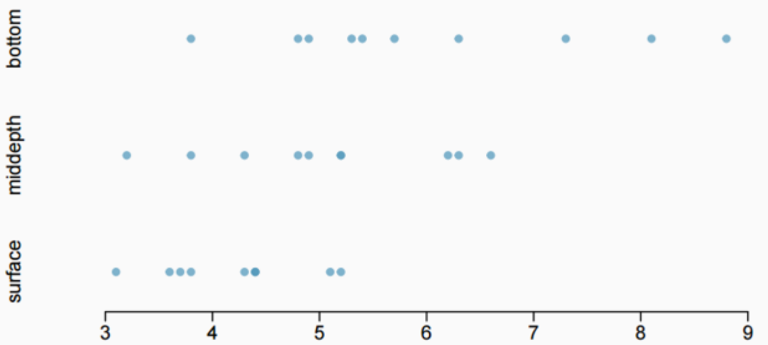
- The Wolf River in Tennessee flows past an abandoned site once used by the pesticide industry for dumping wastes, including chlordane (pesticide), aldrin, and dieldrin (both insecticides)
- These highly toxic organic compounds can cause various cancers and birth defects
- The standard methods to test whether these substances are present in a river is to take samples at six-tenths depth
- Since these compounds are

Figure 1: ***Research Question:*** Does the average aldrin concentration vary between the bottom, mid-depth, and surface?

The Data

	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
...		

Exploratory Analysis & Sample Statistics



	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

Equal Variance Assumption

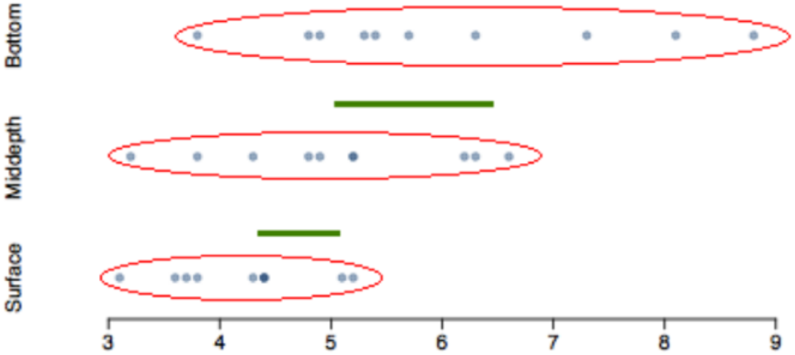


Figure 2: Do these variances look approximately equal?

ANOVA Results

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Group)	depth	2	16.96	8.48	*6.14*	0.0063
(Error)	Residuals	27	37.33	1.38		
	Total	29	54.29			

Interpreting The Results

- The test statistic F was 6.14, meaning the between-groups variability was 6.14 times as large as the within-group variability.

Interpreting The Results

- The test statistic F was 6.14, meaning the between-groups variability was 6.14 times as large as the within-group variability.
- The p-value was 0.0063, meaning it was very unlikely to have seen this much variability between groups in samples of these sizes, given the null hypothesis that each mean is the same ($\mu_i = \mu$).

Interpreting The Results

- The test statistic F was 6.14, meaning the between-groups variability was 6.14 times as large as the within-group variability.
- The p-value was 0.0063, meaning it was very unlikely to have seen this much variability between groups in samples of these sizes, given the null hypothesis that each mean is the same ($\mu_i = \mu$).
- The p-value is less than a significance threshold of $\alpha = 0.05$, so we would reject the null hypothesis that the means of the 3 groups are equal. The mean of at least one group differs from the other means.

Interpreting The Results

- The test statistic F was 6.14, meaning the between-groups variability was 6.14 times as large as the within-group variability.
- The p-value was 0.0063, meaning it was very unlikely to have seen this much variability between groups in samples of these sizes, given the null hypothesis that each mean is the same ($\mu_i = \mu$).
- The p-value is less than a significance threshold of $\alpha = 0.05$, so we would reject the null hypothesis that the means of the 3 groups are equal. The mean of at least one group differs from the other means.
- But we made some pretty strong assumptions.

Interpreting The Results

- The test statistic F was 6.14, meaning the between-groups variability was 6.14 times as large as the within-group variability.
- The p-value was 0.0063, meaning it was very unlikely to have seen this much variability between groups in samples of these sizes, given the null hypothesis that each mean is the same $(\mu_i = \mu)$.
- The p-value is less than a significance threshold of $\alpha = 0.05$, so we would reject the null hypothesis that the means of the 3 groups are equal. The mean of at least one group differs from the other means.
- But we made some pretty strong assumptions.
- If you want to know which means are different from each other, you will have to do additional pairwise tests between group means.