WEEK 6: STATISTICS (I)

Dr. Kai Li
School of Information Sciences
University of Tennessee, Knoxville
Spring 2025

Review of Week 4

- Data cleaning/wrangling.
  - *You have to play with the data before using it and get your hands (even arms) dirty!*
  - *Understand the distribution, spot error and outliers, work with missing values…*
- Any questions from last week?

Statistical methods

- Descriptive: focusing on summarizing and describing
- Inferential: focusing on understanding relationships between variables, and make predictions!
  - *So inferential methods are definitely more interesting.*
  - *But we will talk about regression and prediction, the more interesting part of this topic, in the next two classes.*

Inferential statistical models

- For most of the inferential methods, we generally have one dependent variable and one to more independent variables to construct a model.
  - *Dependent variable: or response, is the focus or outcome that we want to understand in the research.*
  - *Independent variable: or treatment, is the factor that could lead to the outcome.*
- The model is often expressed as:
  - *Dep ~ Ind1 + Ind2*

Data standardization/normalization

- I am not trying to distinguish these concepts, but focus on their similarity: to rescale the data, so that all variables will have a similar scale and come to a normal shape.
  - *But both concepts have a lot different definitions.*
  - *And there are a lot of debates about if these methods will be applicable to many contexts.*
- These methods are important because of (1) normality assumption and (2) data scale assumption in many statistical models.

Data standardization

- Square root method:
  - *Apply square root to all values in the case where there is not negative value.*
- Logarithm method:
  - *Apply natural logarithm or base-10 logarithm to all values (again, only to positive numbers).*
- Max-min method:
  - *(value - min) / (max - min) to get a decimal number between 0 and 1*
- Z-score method
  - *(value – mean value) / standard deviation to get values with a mean of 0 and sd of 1.*

STATISTICAL ANALYSIS DECISION MAKING

Two group comparison	Mean	Parametric		Independent 2 sample t test
		Nonparametric		Mann Witney U test
	Percentage	Chi-Square Test		
One group comparison	Mean	Single mean		One sample t test
		Mean difference	Parametric	Paired t test
			Non parametric	Wilcoxan Signed Scale test
More than 2 group comparison	Mean	Parametric		ANOVA
		Non parametric		Kruskal Walli's test
	Percentage	Chi square test		

https://www.slideshare.net/PrincyFrancisM/statistical-analysis-119122733

Inferential statistical methods

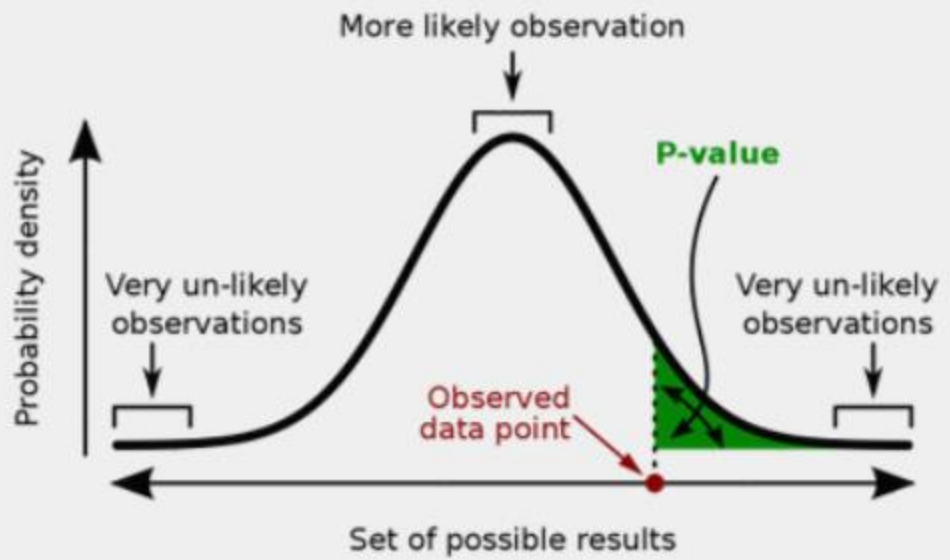
- This set of methods is characterized by the testing of hypothesis or assumption.
  - *We set up a null hypothesis (H_0), i.e, there is no significance, that we are trying to disprove and an alternative hypothesis (H_1) that we can prove if the null hypothesis is rejected.*
  - *From the test, we will get a p-value, that indicates the probability of the result occurring given the null hypothesis is true. → We generally reject the null hypothesis if p <= 0.05.*
  - *In some scientific fields, we rely on lower threshold (such as 0.01 or even 0.001).*

Important:

Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is committing an egregious logical error: **the transposed conditional fallacy.**



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

P-VALUE

An example

- In the regression model, the null hypothesis is that there is no correlation between the independent variable (biking) and the outcome (heart disease).
- The results show that we can reject the null hypothesis at 0.05-level and reach the conclusion that biking is correlated with heart disease.

```
Call:
lm(formula = heart.disease ~ biking + smoking, data = heart.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1789 -0.4463  0.0362  0.4422  1.9331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.984658   0.080137  186.99   <2e-16 ***
biking      -0.200133   0.001366 -146.53   <2e-16 ***
smoking      0.178334   0.003539   50.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.654 on 495 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9795
F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

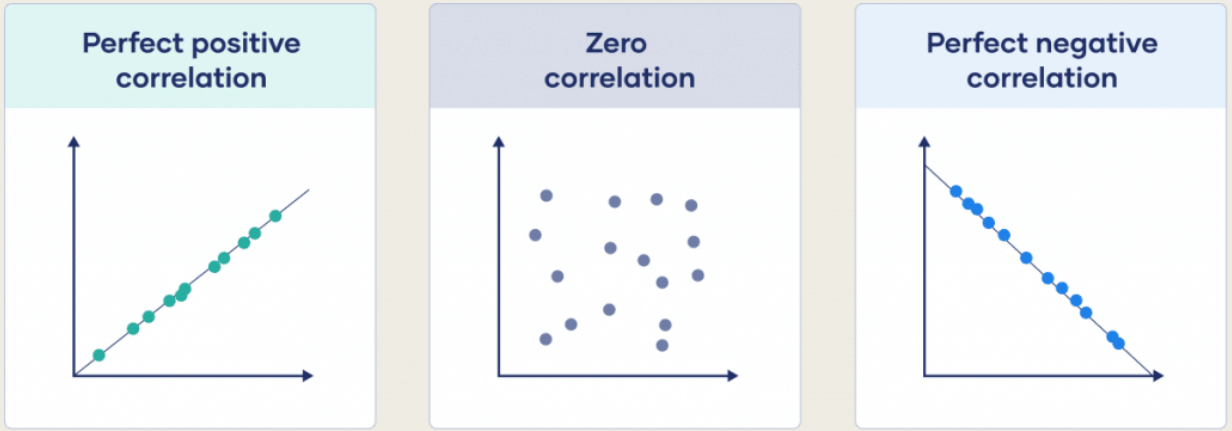
Issues with p-value

- There are many issues with overreliance on p-value to determine statistical significance, which can be very broadly called “p-hacking.”
  - *manipulate data (only use selected observations, or transform the data) to get significant results;*
  - *selectively report significant results.*
- The above practices are VERY bad ones that I hope you will not use in your career.

Inferential statistical methods

- Some examples of inferential methods:
  - *Correlation*
  - *Difference*
  - *Regression*

Correlation



- Correlation measures the statistical relationship between two variables.
- It is a value between -1 and 1 and can be negative or positive.
- Results from correlation analysis contains coefficient (showing the strength) and p-value (showing the statistical significance).
  - *Null hypothesis: there is no correlation between the two variables.*

"There was a positive correlation between the two variables, r = 0.985, n = 5, p = 0.002."

Correlations

		WATER	SKIN
WATER	Pearson Correlation	1.000	.985**
	Sig. (2-tailed)	.	.002
	N	5	5
SKIN	Pearson Correlation	.985**	1.000
	Sig. (2-tailed)	.002	.
	N	5	5

**. Correlation is significant at the 0.01 level

Correlation method

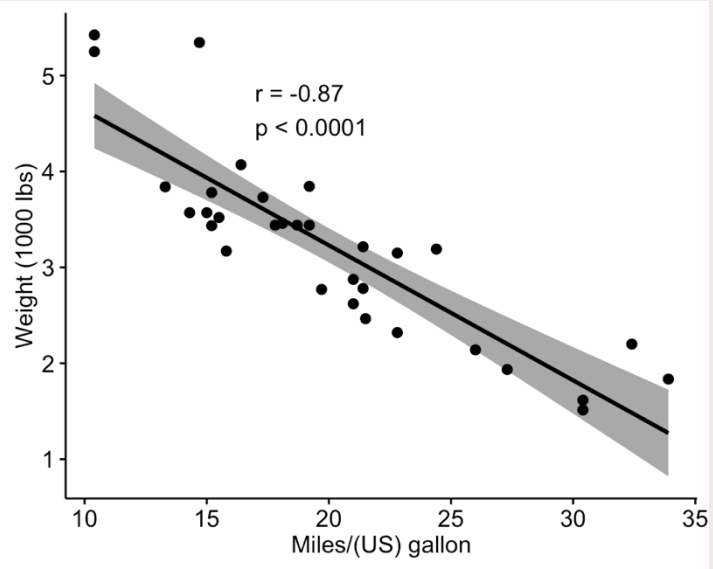
- Two very popular correlation methods:
  - *Pearson Correlation: relationship between two numeric variables*
  - *Spearman Correlation: relationship between two ranked variables*
- Correlation does not indicate causation!
  - *We cannot say "factor A causes factor B."*

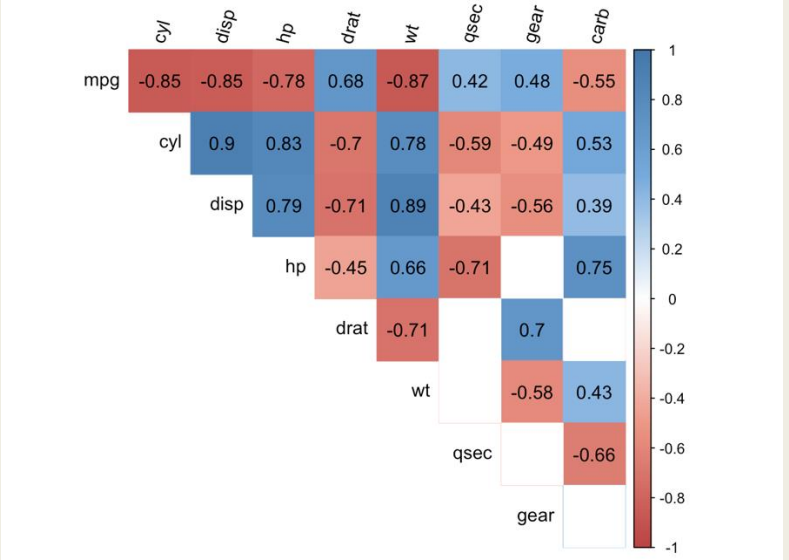
Assumptions of Pearson correlation

- Pearson correlation has the following requirements for the data:
  - *Both variables being continuous and paired*
    - Examine the class of variables and understand their meanings
  - *No outlier in the data*
    - Histogram or boxplot
  - *The two variables have linear relationship*
    - Scatterplot; but we will talk about one more approach in linear regression models
- Ideally, we want to meet these requirements so that our results are reliable.
- All the other models have requirements for data as well!

Visualization of correlation

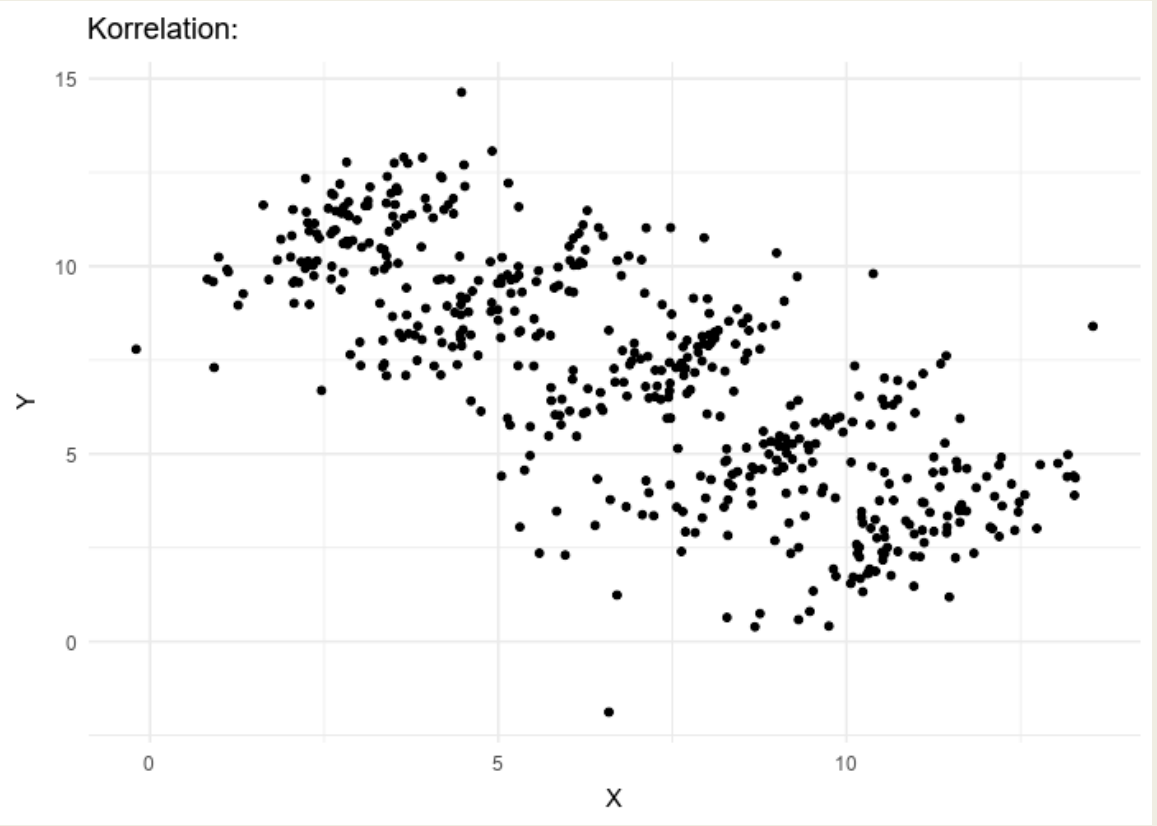
- Scatterplots
  - *Small multiples*
  - *Trendline*
- Heat map





Simpson's Paradox

- All subcategories in the data can have one type of correlation (positive) but the overall dataset shows the opposite relationship (negative).
- The overall trend may not reflect the actual pattern in the data!



What statistical methods can prove causality?

- Most of the statistical models can only prove correlation, including regressions models (which also shows the strength of the correlation relationship).
- We generally need experiments, beyond just statistical methods, to establish causality.
  - *We can use controlled experiment (selecting two very similar sample and give them different treatment) or quasi experiment (measure things before and after a spark) to collect data.*

Difference

- We can use various methods to compare the difference between two or multiple groups.
  - *T-test: a set of methods to compare the difference between the means of two groups.*
  - *ANOVA (analysis of variance): it is similar with t-test but can be used to compare more than two groups.*
  - *Chi-square test: it is used to compare the differences between categorical variables (for example, whether a new medicine can treat a disease better by curing more people).*

T-test

- There are a few different types of t-test.
  - One-sample t-test compare a group of value to a known mean (we can set the number).
  - Two-sample t-test compare the values from two independent groups.
  - Paired t-test compared two samples that are paired together.
    - *For example, A/B test or before/after test.*
- *We need to provide one dependent variable (numeric) and one independent variable (categorical).*
  - We will see how the two groups in the independent variable could results significant differences on the dependent variable.

T-test results

- The null hypothesis of t-test is that A and B have no significant difference.
- T-value and p-value are linked in t-test.
  - *T-value shows the size of difference and p-value shows if the difference is significant statistically.*
    - > 2 or < -2 are generally the threshold for significance.
  - *A larger t-value means a smaller p-value.*

```
> t.test(cholesterol ~ group, var.equal=TRUE, data = istt)

        Two Sample t-test

data:  cholesterol by group
t = 3.776, df = 19, p-value = 0.001278
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2274728 0.7932545
sample estimates:
 mean in group control mean in group exercise
              5.064000                4.553636
```

ANOVA

- Similarity between t-test and ANOVA:
  - *Both are focused on comparing differences across groups (unlike regression).*
  - *Both t-test and ANOVA can only use categorical independent variables and numeric dependent variable.*
- Differences:
  - *ANOVA can support the comparison between more than two groups.*
    - If we use ANOVA to compare just two groups, then we can get the same results with using t-test.
  - *ANOVA can support up to more than one independent variable.*
    - Two- or three-way ANOVA can show the interaction between the IVs for the outcome.

ANOVA results: one-way

- One-way ANOVA is just like t-test.
- In our model to use species to understand sepal length of irises, we see that there is statistical significance between species, meaning that at least one of the species group is significantly different from each other!
  - *We don't know how exactly they are different from each other, yet.*

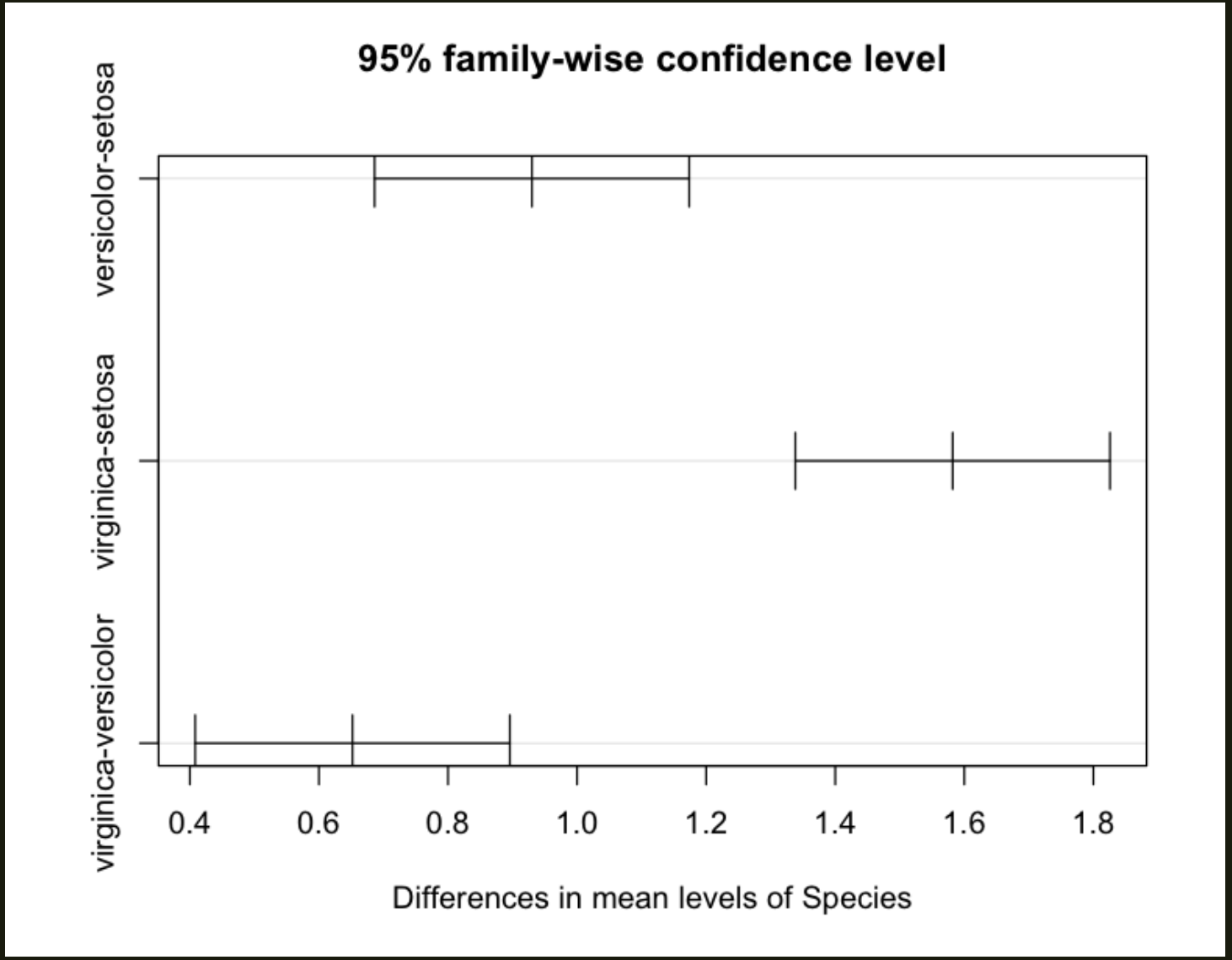
```
> summary(model_anova)
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16
Residuals   147  38.96   0.265

Species     ***
Residuals
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA results: two-way

- For two-way ANOVA, we have the option to show the interaction between the two categorical variables.
- In this case, besides the main effects of individual variables (Species and Sepal.Width.group), we also have their interaction effect: if each of the variable is affected by the other variable in terms of its relationship with the outcome?

	Df	Sum Sq	Mean Sq	
Species	2	63.21	31.606	
Sepal.Width.group	1	4.74	4.742	
Species:Sepal.Width.group	2	0.04	0.022	
Residuals	144	34.17	0.237	
	F value	Pr(>F)		
Species	133.196	< 2e-16	***	
Sepal.Width.group	19.985	1.57e-05	***	
Species:Sepal.Width.group	0.093	0.911		
Residuals				



ANOVA POST-HOC TEST

We can use Tukey HSD Test to examine the differences between groups in ANOVA test.

Reporting of the results

- I didn’t talk about reporting of results.
- But you can find some general guidelines about reporting results from each of the methods based on different writing styles (like APA) from Google.
- I hope you can do this homework when you start using the methods.

Demonstration

- 1. Standardization
- 2. Correlation
- 3. T-test
- 4. Anova

- Activity: Please see the end of the code file!