Midterm I Review

PH142 Fall 2025

Logistics

- **Date:** Friday, October 3rd
- **Time:** 8:10-9:00AM, arrive no later than 8:00AM
- **Location(s):** Wheeler, Stanley, Dwinelle
    - Room assignments will be emailed to you next week
- **Material Covered:** Lectures 1-10, Lab 1-3

What to Bring

- **Student ID**
- **Pencil/Pen**
- **Cheat Sheet** (single sided, handwritten, 8.5x11”)
- **Scientific Calculator** (non-graphing)

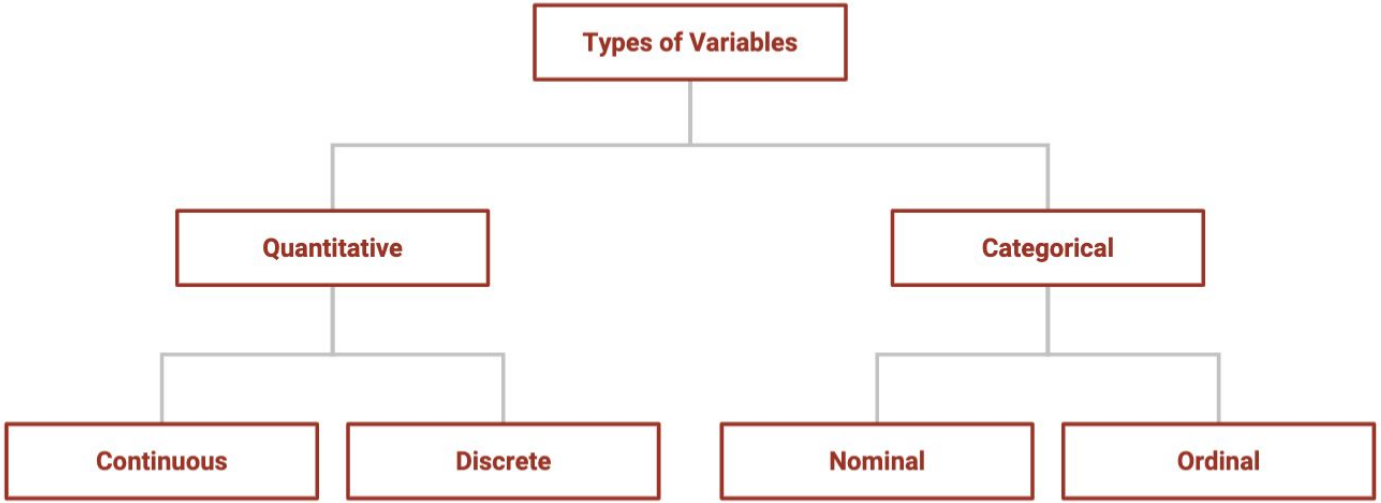
PPDAC

- **Problem:** A clear statement of what we are trying to achieve.
- **Plan:** The procedures we use to carry out the study.
- **Data:** The data which is collected according to the Plan.
- **Analysis:** The data is summarized and analyzed to answer the questions posed by the Problem.
- **Conclusion:** Conclusions are drawn about what has been learned about answering the Problem.

Problem Types

- **Descriptive:** *Learning about some particular attribute of a population*
  - Example: Trying to determine the mean amount that grad students pay for rent at Berkeley
- **Causative/Etiologic**: *Do changes in explanatory variables cause changes in response variables?*
  - Example: If we change the stipend amount for grad students, how does that change amount of money grad students at Berkeley choose to spend on rent?
- **Predictive:** *How can we best predict the value of the response variable for an individual?*
  - Example: Can we identify some explanatory variables that help predict the amount of money that grad students choose to spend on rent? It may simply be an association relationship rather than a causal relationship.

Variable Classification



Variable Classification

Quantitative Variables

- A **discrete variable** is a variable whose value is obtained by counting.
  - Number of children each mother has had
  - Number of cigarettes smoked each day
- A **continuous variable** is a variable whose value is obtained by measuring.
  - Hospital fees (in dollars)
  - Time it takes for a wound to heal

Variable Classification

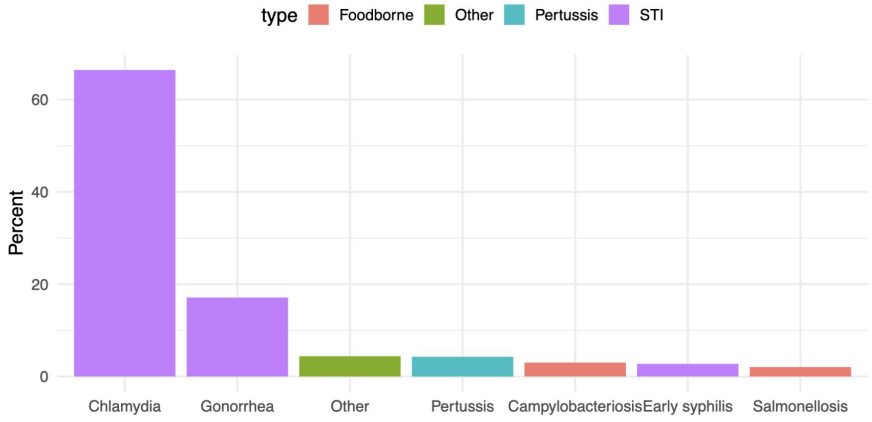
Categorical Variables

- A **nominal variable** is a categorical variable that has no particular order.
  - Eye color (several groups)
  - Disease status (2 groups: has disease vs. does not have)
- An **ordinal variable** is a categorical variable that has a particular order.
  - Income group
  - Pain chart at a doctor's office ordered from sad face to happy face

Plots: Bar Charts

Bar charts are used to illustrate the distribution of a categorical variable, with spaces between bars.

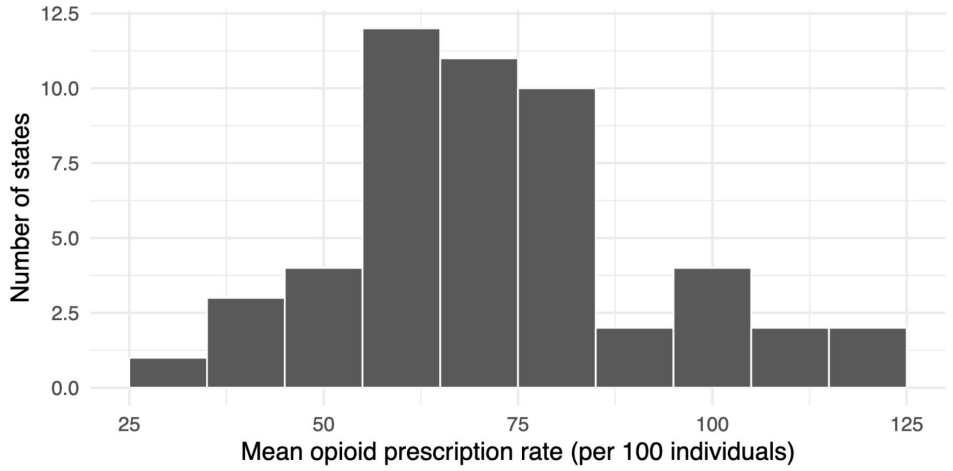
The “distribution” of a categorical variable is the count or percent of individuals in each category.



Plots: Histograms

Histograms are used to illustrate the distribution of a numeric (continuous or discrete) variable. There are no spaces between bars.

*Note that it does not make sense to rearrange the bars of a histogram



Describing Histograms

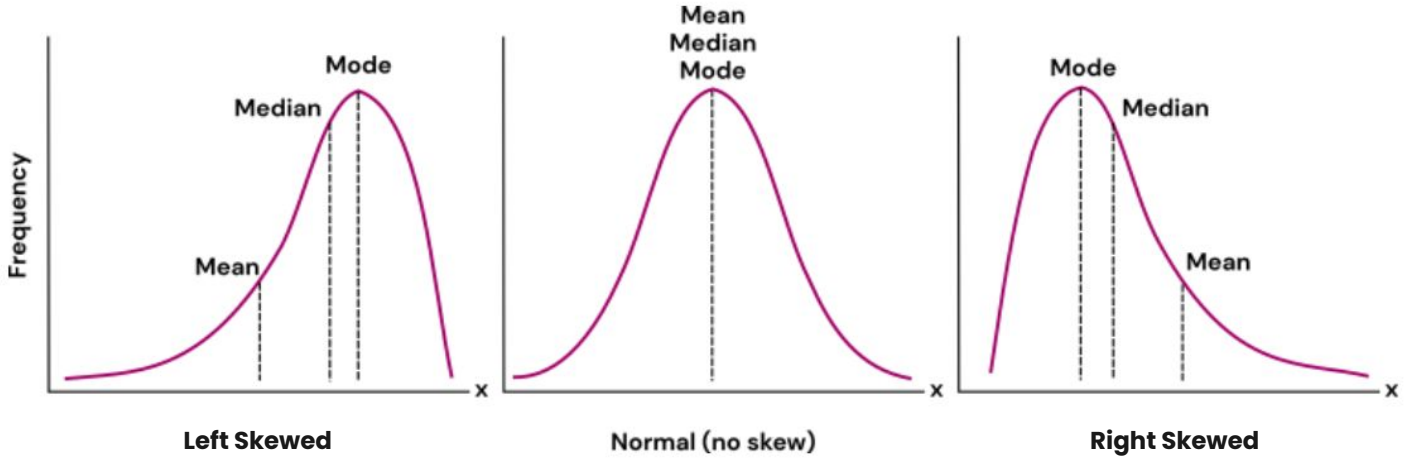
Shape: Is the distribution symmetric or skewed to the left or right?

Center: Does the histogram have one peak (unimodal), or two (bimodal) or more?

Spread: How spread out are the values? What is the range of the data?

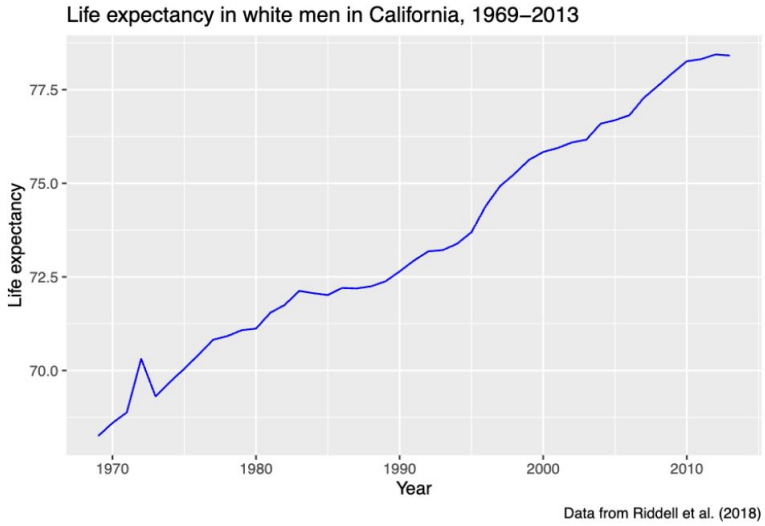
Outliers: Do any of the measurements fall outside of the range of most of the data points?

Describing Histograms



Plots: Line Plots

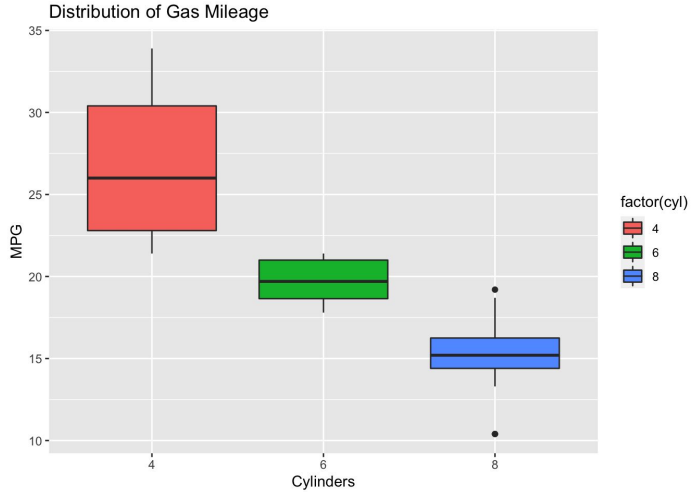
Line plots are used to visualize how continuous variables change over time.



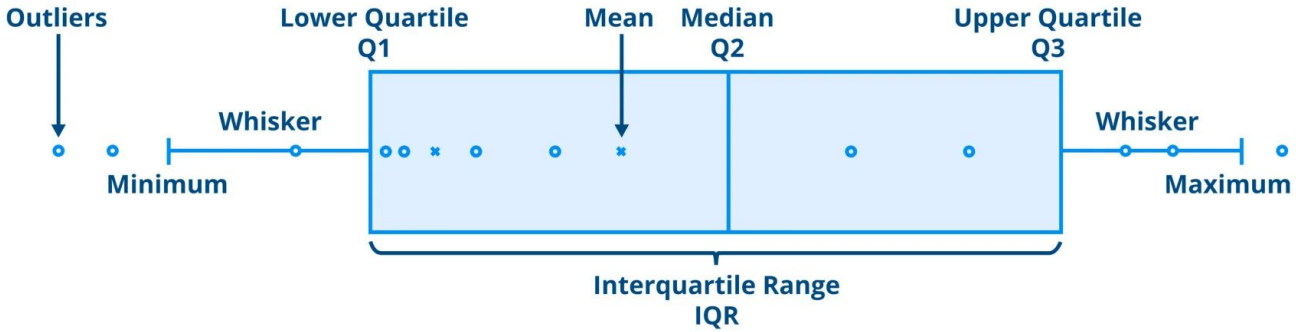
Plots: Box Plots

Box plots are used to visualize the distribution of continuous variables.

They give us a five-number summary: minimum, Q1, median, Q3, and maximum.

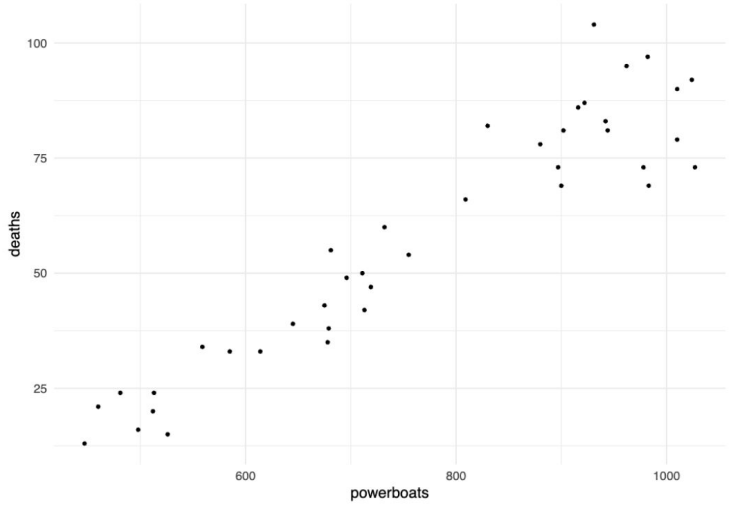


Plots: Box Plots



Plots: Scatter Plots

Scatter plots are used to visualize the relationship between two continuous variables.



Correlation

- Correlation measures the **strength** and **direction** of a linear relationship between two quantitative variables.
- Also written as ***r***
- Takes values between -1 to 1, inclusive.
- r^2 is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.

Linear Regression

- **Regression:** Straight line fitted to data to minimize distance between the data and fitted line
  - “Line of best fit” = **a + bx**
  - **a** = **intercept** (expected value of y when x=0)
  - **b** = **slope**
- Interpretation
  - Intercept : the value of the outcome when x = 0
  - Slope: For a one-unit change in x, the outcome changes by [number] [units]

Linear Regression

- **lm()** is the function for a linear model
  - lm(formula = y ~ x, data = your_dataset)
  - Add regression line to a scatterplot using: geom_abline(slope=, intercept=)
- **Interpretation of lm():** A one unit change in **X** is associated with a ____ increase/decrease of **Y**.
- **Interpretation of r-squared**: the fraction of the variation in the values of y that is explained by the line of best fit (the regression of y on x)

Two-Way Tables

Marginal Distribution: The distribution of a single categorical variable in the entire population

- Use the totals in the margins of the table to calculate proportions
- Percent of Exposed individuals= (A+B)/(A+B+C+D)
- Percent of Unexposed individuals = (C+D) / (A+B+C+D)

Exposure group	Disease	No disease	Row total
Exposed	A	B	A+B
Not Exposed	C	D	C+D
Column total	A+C	B+D	A+B+C+D

Two-Way Tables

Conditional Distribution: The distribution of one categorical variable within the other

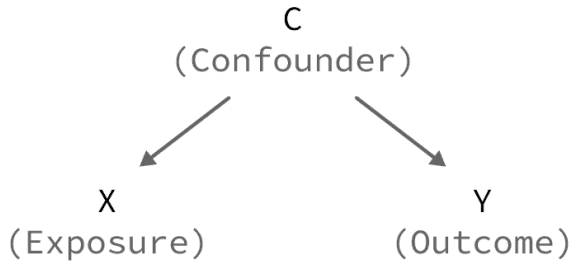
- Use a single row or column to calculate proportions
- Percent of Diseased given exposure = A / (A+B)
- Percent of Exposed given disease = A / (A+C)

Exposure group	Disease	No disease	Row total
Exposed	A	B	A+B
Not Exposed	C	D	C+D
Column total	A+C	B+D	A+B+C+D

Simpson's Paradox

Simpson's Paradox: When an association that seems to hold for all or several groups *reverses directions* when you look at data for the whole group.

The key to Simpson's paradox is that one of the categorical variables is a **confounder**. That is, a variable that is not part of the explanatory-response relationship but influences the interpretation of the relationship between these variables.



Samples and Study Design

Observation vs Experimentation

A study is observational if the researcher **observes** what happens and does not control who is treated or exposed.

- Does not control for confounding

A study is experimental if the investigator is **experimenting** (or intervening) by controlling who is treated or exposed.

- In an experimental study, the exposure is assigned by a randomization mechanism that is controlled by the investigator

Samples and Study Design

Population of Interest

Target Population: The target population is the entire group of individuals who we want information, and to whom we would like to apply the estimates and conclusions.

Study Population: The study population is the part of the population that you can draw a sample of individuals from.

Study Sample: The study sample is composed of individuals who have been sampled from the study population. This is the group on which you gather data. Samples are used to draw conclusions about the target population

Samples and Study Design

Experimental Validity

Internal Validity: confidence that the study's design and conduct allow for unbiased conclusions about causal effects (free from confounding, bias, etc.)

External Validity: generalizability of findings to the target population

Samples and Study Design

Observational Study Designs

Cross-Sectional: a “snapshot” of exposure and outcome at a specific time

Cohort: choosing participants based on exposure

Case-Control: choosing participants based on outcome

Samples and Study Design

Experimental Design Terms

- **Experimental units:** individuals or larger groups
- **Randomization:** randomly assigned different levels of treatment to individuals or groups
- **Factor:** An explanatory variable that is being manipulated
- **Treatment:** A specific experimental condition
- **Blinding:** the process used in experimental design by which individuals, clinicians and data collectors are kept unaware of the randomization.
- **Placebo:** An inactive treatment meant to mimic the look or feel of the treatment being tested in an RCT but that has no active ingredients
- **Placebo effects:** the measured effect on the outcome in the placebo “arm” of the RCT

Samples and Study Design

Simple Random Sampling

Simple Random Sample (SRS): A sample chosen by chance, where each individual in the dataset has an equal chance of being selected

Functions in R:

- slice_sample(n = 100), selects n rows at random
- slice_sample(prop = 0.05), selects a random proportion of rows
- set.seed(#), makes results reproducible by taking the same sample

Example: CS_100 <- CS_data %>% slice_sample(n = 100)

Coding Review

R Language Components

Packages

- A bunch of code that we want to use (for our purposes, a collection of functions)
- Loaded with library(package_name)
- E.g. dplyr, ggplot2

Functions

- Actions that do things (usually to objects)
- Need parentheses after the name, and any parameters go inside the parentheses (usually)
- E.g. mutate(), select(), ggplot(), etc.

Objects

- A “thing” that we can do things to, such as:
    - Name it (a.k.a. Assign it to a variable)
    - Change it (e.g. adding a column to a data frame)
    - Use it (e.g. use it to create a plot)

Functions

dplyr Function Review

Function	Purpose	Input	Output	Notes
rename()	Rename variables	rename(old_dataset, new_name = old_name)	Renamed dataset	Can do multiple variables at one!
select()	Select a subset of variables	select(dataset, column1, column2)	Dataset with specific columns	Remove variables with -
arrange()	Sort by a variable	arrange(dataset, column)	Dataset arranged by specific columns	Can sort by multiple variables, default is ascending order, descending order with -

Functions

dplyr Function Review

Function	Purpose	Input	Output	Notes
filter()	Select a subset of rows	filter(dataset, condition)	Dataset with rows based on conditions	==, <, >, <=, >=, != Or (|), and (&)
mutate()	Add new variables to a dataset	mutate(dataset, new_column = data)	Dataset with new variable	You can call existing variables in the dataset to create your new variable
group_by	Group data by a categorical variable	group_by(column)	Dataset grouped by variable	Use with summarize()!
summarize()	Summarize a statistic	summarize(statistic_name = calculation)	Tibble with grouped data and statistic for each group	Use with group_by()!

Functions

Describing Data with Functions

dim(): Returns the number of rows/individuals and number of columns/variables

names(): Returns the variable names

head(): Returns the first six rows

str(): Returns information about the types of variables found in the data set in terms of whether they are quantitative (int, num) or categorical (Factor) and provides some information about each one.

Visualizing Data

ggplot

Template: ggplot(data= dataset, aes(x=var1, y=var2)) + geom_point(

- **Bar chart:** geom_bar( )
- **Histogram:** geom_histogram( )
- **Line Plot:** geom_line( )
- **Scatterplot:** geom_point()
- **Box Plot:** geom_boxplot()

Add Title/Axes: + labs(title=" " , y= " ", x=" ")

Visualizing Data

ggplot Aesthetics

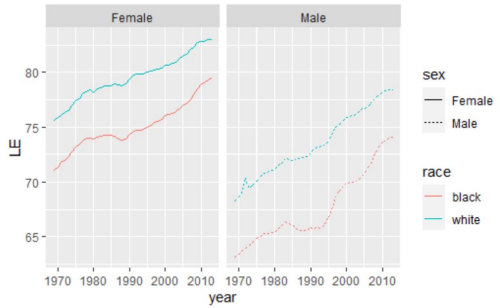
Inside of the aes() function, you can link plot features to variables using the following arguments:

- **`col`** : Colors points, lines, and borders
- **`fill`** : Fills the interior of bars and boxplots
- **`size`** : Adjusts point size and line thickness
- **`lty`** : Controls line type (solid, dashed, etc)

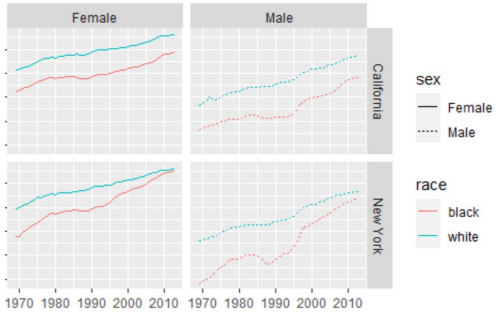
Visualizing Data

Faceting Plots

Use **facet_wrap(~ var1)** to make separate plots for different levels of ONE variable



Use **facet_grid(var1 ~ var2)** to make separate plots for combinations of TWO variables



Practice Problems

Variable Types

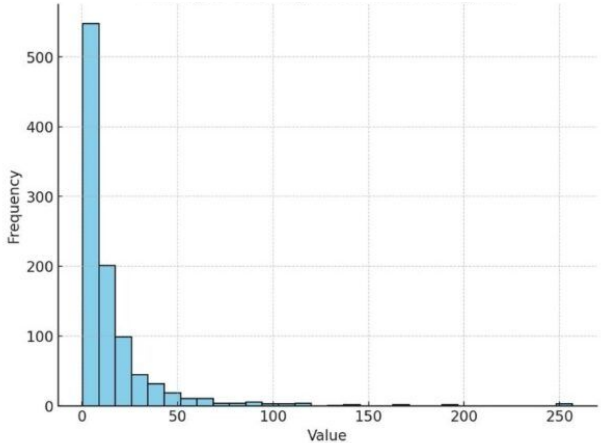
1. **For each of the following, determine the type of variable:**
    A. Number of hospital visits in the past year
    B. Blood pressure (mmHg)
    C. Type of health insurance (Medicare, Medicaid, Private, None)
    D. Pain severity scale (None, Mild, Moderate, Severe)
    E. Daily hours of sleep
    F. Favorite fruit (apple, banana, mango, etc.)

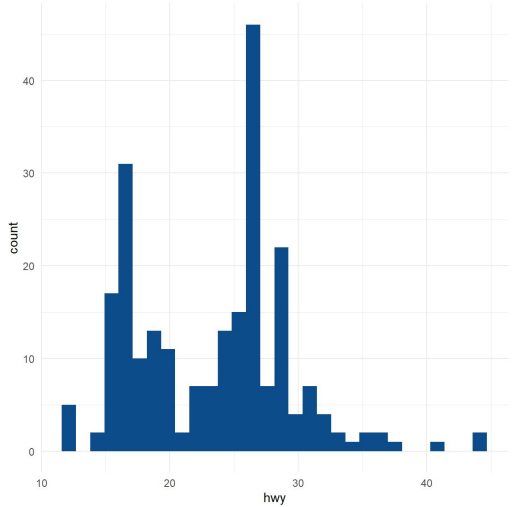
Variable Types (Key)

1. **For each of the following, determine the type of variable:**
   A. Number of hospital visits in the past year → **Discrete**
   B. Blood pressure (mmHg) → **Continuous**
   C. Type of health insurance (Medicare, Medicaid, Private, None) → **Nominal**
   D. Pain severity scale (None, Mild, Moderate, Severe) → **Ordinal**
   E. Daily hours of sleep → **Continuous**
   F. Favorite fruit (apple, banana, mango, etc.) → **Nominal**

Describing a Distribution

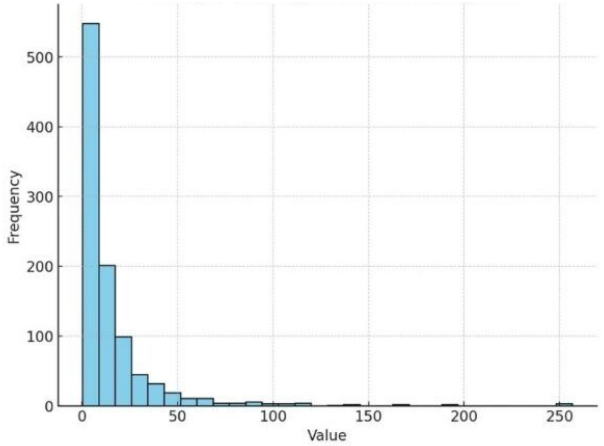
2. Using the shape/center/spread/outliers method, describe the each histogram:



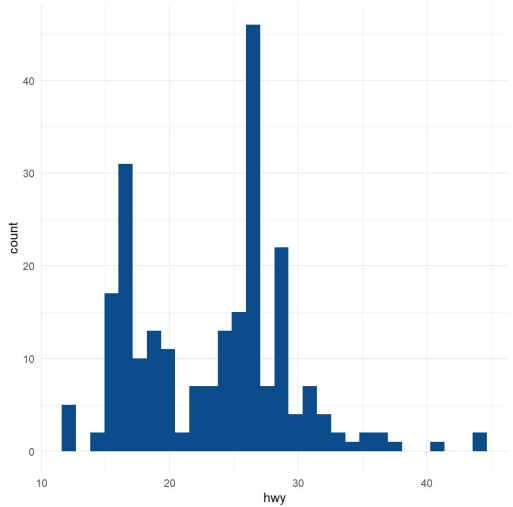


Describing a Distribution (Key)

2. Using the shape/center/spread/outliers method, describe the each histogram:



Unimodal, right skew, outliers

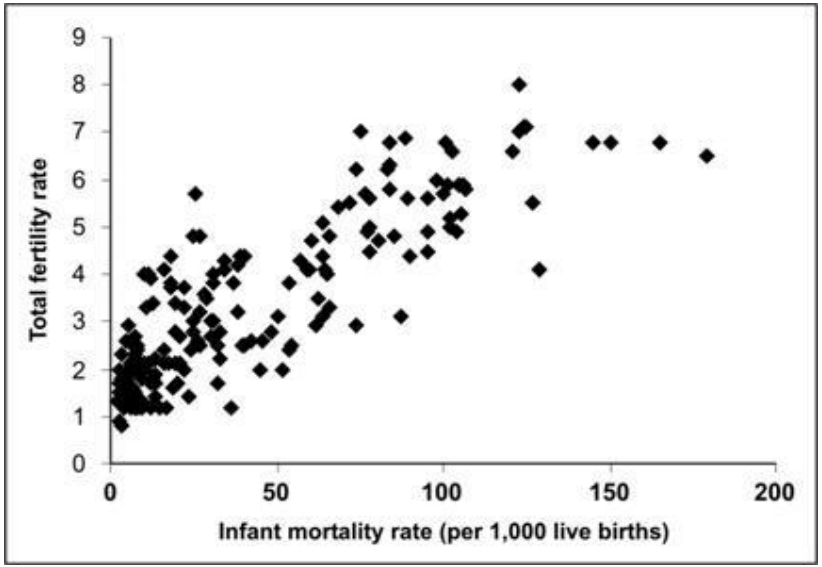


Bimodal, asymmetric

Correlation

3. Based on the scatterplot:

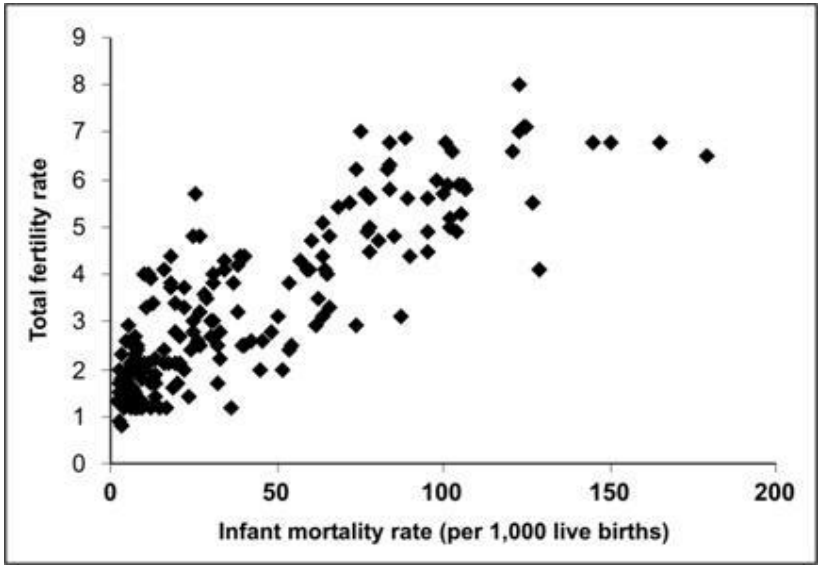
- What shape of association does this scatterplot show?
- How strong is this relationship?
- What is the direction of the relationship?
- Do you see any apparent outliers?



Correlation (Key)

3. Based on the scatterplot:

- **Shape: linear for the first 80%, then starts to flatten**
- **Strength: medium in terms of linearity**
- **Direction: Positive**
- **Outliers: None**



Linear Models

4. This is a data frame called OFCdata that is from a neuroscience research study that examines the relationship between two regions in the orbitofrontal cortex of the brain (OFC1 and OFC2). The researcher planted electrodes into both brain regions for several individuals and records the activity level during different independent stimuli. The results are shown here:

```
## # A tibble: 6 x 2
##     OFC1  OFC2
##    <dbl> <dbl>
## 1  0.759  6.31
## 2  4.10   7.94
## 3  2.01   9.46
## 4 12.6   11.0
## 5  4.98  11.1
## 6  1.92   8.90
```

Linear Models

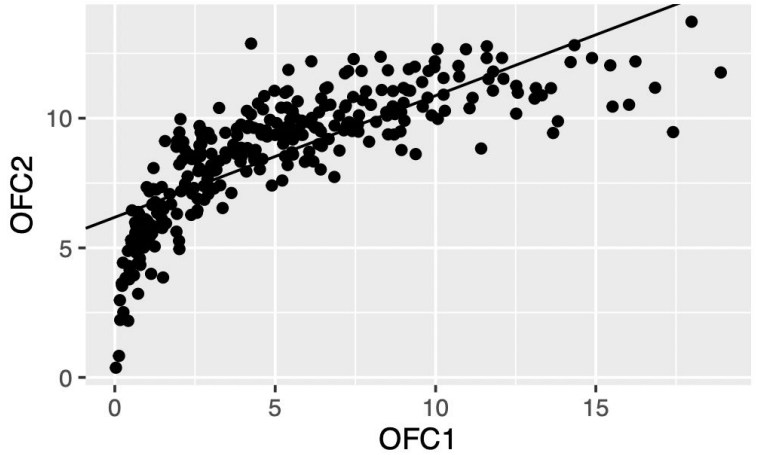
4A. You decide to fit a linear regression to the data. Fill in the blanks to generate the output shown below.

```
fit <- __[A]__( __[B]__ ~  __[C]__, OFCdata)

 __[D]__(fit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    6.18     0.145       42.6 7.26e-129
## 2 OFC1           0.469    0.0220      21.4 4.90e- 62
```

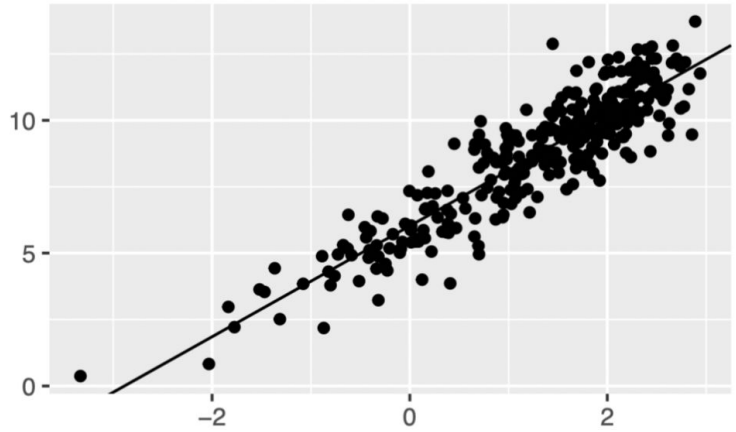
Linear Models

4B. Below is a plot of the OFCdata with the line of best fit that you generated in part a. Based on this plot, visually describe the relationship between OFC1 and OFC2.



Linear Models

4C. You then perform a transformation on the data points and generate a new model. Without being given any numbers, which variable(s) were transformed? Which plot (original vs. transformed) has a stronger correlation coefficient? Why?



Linear Models

4D. Which of the following is a plausible value for the correlation coefficient between log(OFC1) and OFC2?

A. 0.6

B. 0.9

C. 1.0

D. 0.3

Linear Models

4E. Interpret the slope parameter from the output below in the context of the problem.

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic   p.value
##   <chr>           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     6.03    0.0937      64.4 7.73e-177
## 2 log(OFC1)       2.09    0.0574      36.4 1.22e-111
```

Linear Models (Key 4A)

A. You decide to fit a linear regression to the data. Fill in the blanks to generate the output shown below.

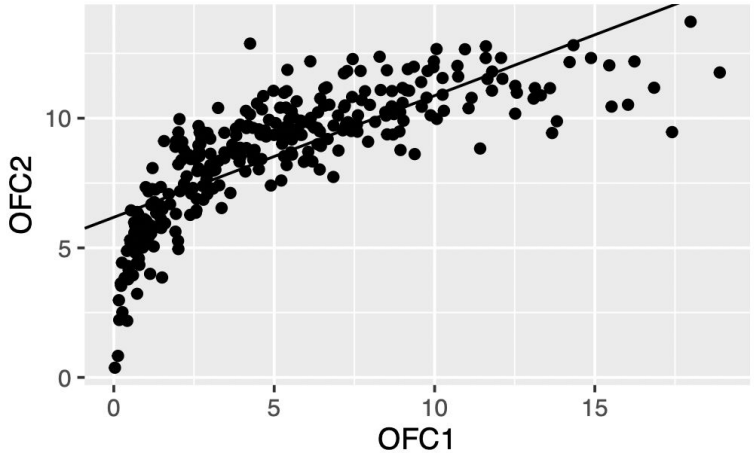
```
fit <- __[A]__( __[B]__ ~  __[C]__, OFCdata)

 __[D]__(fit)
```

A. lm
B. OFC2
C. OFC1
D. tidy

Linear Models (Key 4B)

B. Below is a plot of the OFCdata with the line of best fit that you generated in part a. Based on this plot, visually describe the relationship between OFC1 and OFC2.

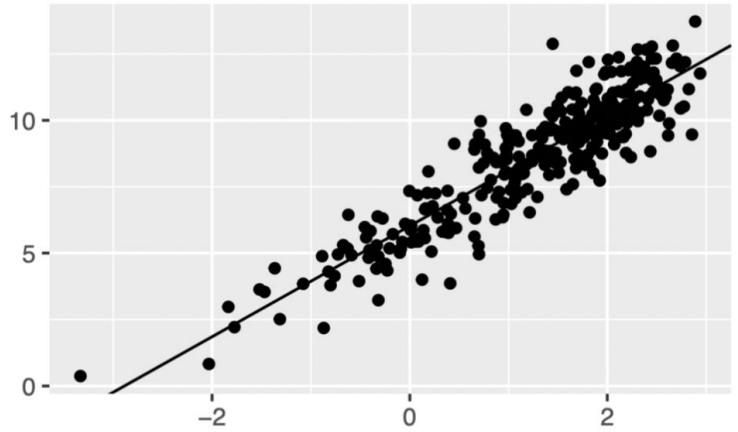
Strength: weak/moderate
Form: steep then flattens (not linear)
Direction: positive
Outliers: none



Linear Models (Key 4C)

C. You then perform a transformation on the data points and generate a new model. Without being given any numbers, which variable(s) were transformed? Which plot (original vs. transformed) has a stronger correlation coefficient? Why?

The OFC1 (x variable) was transformed to log(OFC1). Transformed will have stronger correlation coefficient (closer to 1) Points are more closely aligned with line of best fit in transformed plot; the data in the original plot is too steep and then flattens out (patterned, not linear)



Linear Models (Key 4D)

D. Which of the following is a plausible value for the correlation coefficient between log(OFC1) and OFC2?

A. 0.6

B. 0.9

C. 1.0

D. 0.3

Linear Models (Key 4E)

E. Interpret the slope parameter from the output below in the context of the problem.

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)     6.03    0.0937      64.4 7.73e-177
## 2 log(OFC1)       2.09    0.0574      36.4 1.22e-111
```

For every 1 unit increase in the log(OFC1), there is an increase of 2.09 units in OFC2.

Two Way Tables

5. Based on the 2x2 table:

A. What is the marginal distribution of breast cancer?
B. What is the marginal distribution of the BRCA1 mutation?
C. What is the conditional distribution of breast cancer among those with the mutation?
D. What is the conditional distribution of mutation among those with breast cancer?

	Breast Cancer	No Breast Cancer	
BRCA1 Mutation	152	688	840
No BRCA1 Mutation	66	773	839
	218	1461	1679

Two Way Tables (Key)

5. Based on the 2x2 table:

A. Distribution of breast cancer in the population: **Has breast cancer: 218/1679 = 12.98%**
No breast cancer = 87.02%

B. Distribution of BRCA1 mutation: **Has mutation: 840/1679 = 50.03% , No mutation: 839/1679 = 49.97%**

C. Conditional distribution of breast cancer among those with the mutation: **Has BC: 152/840 = 18.1%, No BC: 81.9%**

D. Conditional distribution of mutation among those with BC: **mutation: 152/218: 69.7%, no mutation: 66/218 = 30.3%**

	Breast Cancer	No Breast Cancer	
BRCA1 Mutation	152	688	840
No BRCA1 Mutation	66	773	839
	218	1461	1679

Simpson's Paradox

6. Based on the data below, answer the following:

A. Create a two-way table for the overall distribution of smoking and death. What percent of non-smokers survived? What about smokers? Is this surprising?

B. Show that within each age group, more non-smokers survived. Explain why this is an example of Simpson's paradox.

Ages 18 to 44

	Smoker	Not
Dead	19	13
Alive	269	327

Ages 45 to 64

	Smoker	Not
Dead	78	52
Alive	167	147

Ages 65+

	Smoker	Not
Dead	42	165
Alive	7	25

Simpson's Paradox (Key A)

6. Based on the data below, answer the following:

A. Create a two-way table for the overall distribution of smoking and death. What percent of non-smokers survived? What about smokers? Is this surprising?

	Smoker	Not	Total
Dead	139	230	369
Alive	443	499	942
Total	582	729	1311

Percent of surviving smokers: 76%

Percent of surviving non-smokers: 68%

Simpson's Paradox (Key B)

B. This is an example of Simpson's paradox because within most strata of age group, a higher percentage of non-smokers survive but overall a higher percentage of smokers survive.

	Smoker	Not
Dead	19	13
Alive	269	327
Total	288	340

Smoker survival: 269/288 = 93.4%
Non-smoker survival: 327/340=96.1%

	Smoker	Not
Dead	78	52
Alive	167	147
Total	245	199

Smoker survival: 167/245 = 68.2%
Non-smoker survival: 147/199 = 73.9%

	Smoker	Not
Dead	42	165
Alive	7	25
Total	49	190

Smoker survival: 7/49 = 14.3%
Non-smoker survival: 25/190 = 13.2%

Good Luck!

- The PH142 Teaching Team