Class 18

DATA1220-55, Fall 2024

Sarah E. Grabinski

2024-10-14

Chapter 1 Review

- Data basics
  - Data organization
  - Numerical and categorical variable types
  - Relationship between 2 numeric variables
- Research Methods
  - Study designs/types
  - Basic data collection strategies

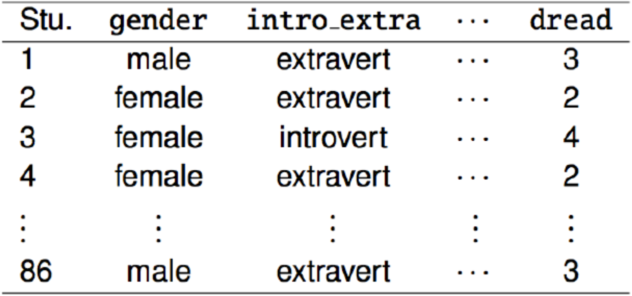
Data Organization

- Data is arranged in a ***matrix*** format such that...
    - Each row (ideally) contains 1 unique observation of the data for each of the measured variables
    - Each column (ideally) contains all the observations for 1 unique variable that was measured
- You may be familiar with this format from Excel or Google spreadsheets
- In R, we typically call these data tables "dataframes"

Data Matrix

variable

↓



Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

←

observation

Figure 1: Data is best organized in rows and columns where each row represents an observation and each column represents a variable.

Types of Data

- Numerical data is ***quantitative***
  - Numerical variables typically have units of measurement
  - Quantitative variables are recorded and used in calculations as numbers

Types of Data

- Numerical data is ***quantitative***
  - Numerical variables typically have units of measurement
  - Quantitative variables are recorded and used in calculations as numbers
- Categorical data is ***qualitative***
  - Variables consist of 2+ categories
  - Categories may be ordered or unordered

Continuous Numerical Data

Continuous numerical data takes on any possible value within a given range.

Continuous Numerical Data

Continuous numerical data takes on any possible value within a given range.

- When numbers have decimal places, it is often continuous

Continuous Numerical Data

Continuous numerical data takes on any possible value within a given range.

- When numbers have decimal places, it is often continuous
- When visualized, there are no meaningful “gaps” between the different values

Continuous Numerical Data

Continuous numerical data takes on any possible value within a given range.

- When numbers have decimal places, it is often continuous
- When visualized, there are no meaningful “gaps” between the different values
- If there are “gaps”, it would be possible to observe that value if you took another sample

Discrete Numerical Variables

Discrete numerical data has a limited set of potential values.

Discrete Numerical Variables

Discrete numerical data has a limited set of potential values.

- When numbers are integers (whole numbers) or rounded, it is often discrete

Discrete Numerical Variables

Discrete numerical data has a limited set of potential values.

- When numbers are integers (whole numbers) or rounded, it is often discrete
- When visualized, there are meaningful “gaps” between the different values

Discrete Numerical Variables

Discrete numerical data has a limited set of potential values.

- When numbers are integers (whole numbers) or rounded, it is often discrete
- When visualized, there are meaningful “gaps” between the different values
- If there are “gaps”, it’s because it was not possible for you to observe that value in your sample

Discrete Numerical Variables (cont.)

Why does it matter?

Discrete Numerical Variables (cont.)

Why does it matter?

- When you have very few distinct numbers (e.g. <10) in your discrete variable, you may have to treat it like a categorical variable

Discrete Numerical Variables (cont.)

Why does it matter?

- When you have very few distinct numbers (e.g. <10) in your discrete variable, you may have to treat it like a categorical variable
- When you have many distinct numbers (e.g. >10) in your discrete variable, you may be able to treat it like a continuous variable

Homework 1: Continuous Numerical Data

```
select(baby_df, bwt)

 1  Variables      1236  Observations
------------------------------------------------------
bwt
      n  missing distinct     Info     Mean      Gmd
   1236        0      107        1    119.6    20.33      8
    .25      .50      .75      .90      .95
  108.8    120.0    131.0    142.0    149.0

lowest :  55  58  62  63  65, highest: 169 170 173 174 176
------------------------------------------------------
```

Homework 1: Discrete Numerical Data

```
select(baby_df, age)

 1  Variables      1236  Observations
-----------------------------------------------------------
age
       n  missing distinct     Info     Mean      Gmd
    1234        2       30    0.997    27.26    6.506
     .25      .50      .75      .90      .95
      23       26       31       36       38

lowest : 15 17 18 19 20, highest: 41 42 43 44 45
-----------------------------------------------------------
```

Categorical Data

- ***Nominal*** categorical variables are *unordered* (i.e. shuffling categories doesn't change how you interpret data)

Categorical Data

- ***Nominal*** categorical variables are *unordered* (i.e. shuffling categories doesn't change how you interpret data)
- ***Ordinal*** categorical variables are *ordered* (i.e. shuffling categories changes how you interpret data)

Categorical Data

- ***Nominal*** categorical variables are *unordered* (i.e. shuffling categories doesn't change how you interpret data)
- ***Ordinal*** categorical variables are *ordered* (i.e. shuffling categories changes how you interpret data)
- ***Binary*** categorical variables have only 2 categories.

Categorical Data

- ***Nominal*** categorical variables are *unordered* (i.e. shuffling categories doesn't change how you interpret data)
- ***Ordinal*** categorical variables are *ordered* (i.e. shuffling categories changes how you interpret data)
- ***Binary*** categorical variables have only 2 categories.
- Any variable with 3 or more categories is called ***multi-categorical***.

Homework 1: Binary Categorical Data

```
select(baby_df, smoke)

 1  Variables      1236  Observations
-----------------------------------------------------------
smoke
       n  missing distinct     Info      Sum     Mean    0.4
    1226       10        2    0.717      484   0.3948   0.4

-----------------------------------------------------------
```

Terminology for Variable Relationships

- ***Independent*** or ***explanatory*** variable
    - Typically on the x-axis
    - The “cause” variable

Terminology for Variable Relationships

- ***Independent*** or ***explanatory*** variable
  - Typically on the x-axis
  - The “cause” variable
- ***Dependent*** or ***response*** variable
  - Typically on the y-axis
  - The “effect” variable

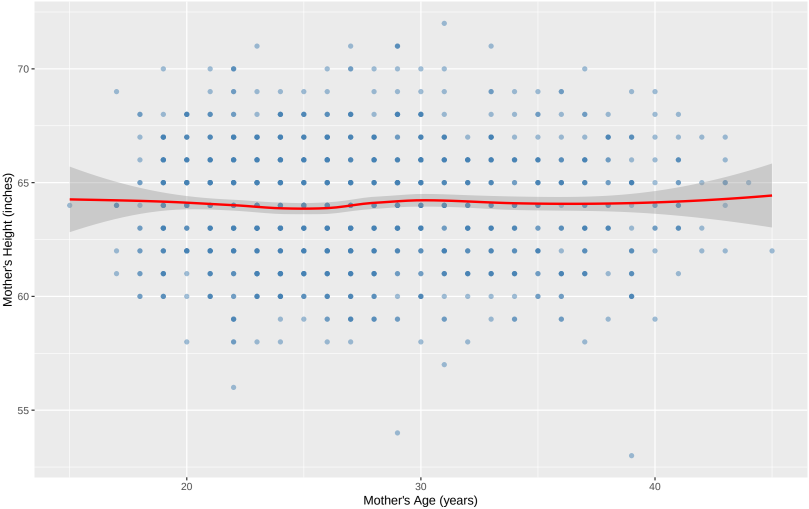
Describing Variable Relationships

- We say there's an association or relationship between 2 variables when a change in X or the ***independent/explanatory*** variable is associated with a change in Y or the ***dependent*** or ***response*** variable
  - ***Positive:*** as X increases, Y increases
  - ***Negative:*** as X increases, Y decreases

Describing Variable Relationships

- We say there's an association or relationship between 2 variables when a change in X or the ***independent/explanatory*** variable is associated with a change in Y or the ***dependent*** or ***response*** variable
  - ***Positive:*** as X increases, Y increases
  - ***Negative:*** as X increases, Y decreases
- We say that 2 variables are independent when a change in X or the ***independent/explanatory*** variable is *NOT* associated with a change in Y or the ***dependent*** or ***response*** variable

Homework 1: Independent Variables



Height in Inches by Age in Years of Mothers

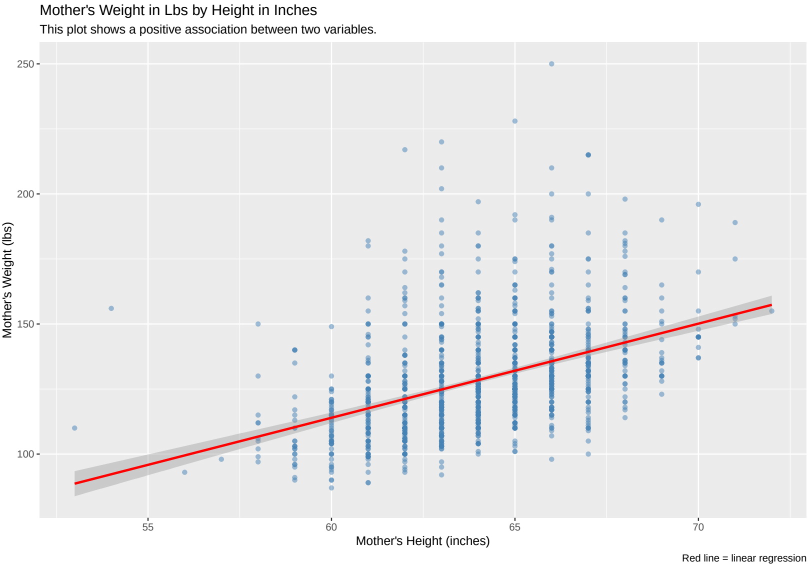
This plot shows the relationship between two independent variables.

Red line = LOESS regression

Homework 1: Independent Variables (Code)

```r
baby_df |>
  ggplot(aes(x = age,
             y = height)) +
  geom_point(col = 'steelblue', alpha = 0.5) +
  geom_smooth(method = 'loess', col = 'red') +
  labs(x = "Mother's Age (years)",
       y = "Mother's Height (inches)",
       title = 'Height in Inches by Age in Years of Mothers
       subtitle = 'This plot shows the relationship between
       caption = 'Red line = LOESS regression')
```

Homework 1: Positive Association



Homework 1: Positive Association (Code)

```r
baby_df |>
  ggplot(aes(x = height,
             y = weight)) +
  geom_point(col = 'steelblue', alpha = 0.5) +
  geom_smooth(method = 'lm', col = 'red') +
  labs(x = "Mother's Height (inches)",
       y = "Mother's Weight (lbs)",
       title = "Mother's Weight in Lbs by Height in Inches'
       subtitle = 'This plot shows a positive association b
       caption = 'Red line = linear regression')
```

Approaches to Collecting Data

- ***Case Study/Anecdotal Evidence***: very few observations, often $n = 1$

Approaches to Collecting Data

- ***Case Study/Anecdotal Evidence***: very few observations, often $n = 1$
- ***Sampling***: a subset of all possible observations

Approaches to Collecting Data

- ***Case Study/Anecdotal Evidence***: very few observations, often $n = 1$
- ***Sampling***: a subset of all possible observations
- ***Census***: all possible observations

Types of Sampling

- ***Simple random sample*** - randomly select cases from a study population with no regard to the individual characteristics of those cases

Types of Sampling

- ***Simple random sample*** - randomly select cases from a study population with no regard to the individual characteristics of those cases
- ***Voluntary response*** - cases self-select from study population by choosing to participate in a study and volunteer their data

Types of Sampling

- ***Simple random sample*** - randomly select cases from a study population with no regard to the individual characteristics of those cases
- ***Voluntary response*** - cases self-select from study population by choosing to participate in a study and volunteer their data
- ***Convenience sample*** - cases are subjects in the study population it is most convenient to get data from

Advanced Random Sampling

- ***Stratified***: simple random samples of cases are taken from each pre-defined cluster of similar cases in the study population

Advanced Random Sampling

- ***Stratified***: simple random samples of cases are taken from each pre-defined cluster of similar cases in the study population
- ***Cluster:*** a simple random sample of pre-defined clusters of non-similar cases in the study population which uses all observations from the sampled clusters

Advanced Random Sampling

- ***Stratified***: simple random samples of cases are taken from each pre-defined cluster of similar cases in the study population
- ***Cluster:*** a simple random sample of pre-defined clusters of non-similar cases in the study population which uses all observations from the sampled clusters
- ***Multistage:*** simple random samples of cases are taken from a simple random sample of pre-defined clusters of non-similar cases in the study population

Types of Study

- ***Observational:*** researchers do not affect the data being collected

Types of Study

- ***Observational:*** researchers do not affect the data being collected
- ***Interventional:*** researchers do *something* which affects the data collected from subjects

Types of Study

- ***Observational:*** researchers do not affect the data being collected
- ***Interventional:*** researchers do *something* which affects the data collected from subjects
- ***Prospective:*** subjects identified in advance and data collected over time

Types of Study

- ***Observational:*** researchers do not affect the data being collected
- ***Interventional:*** researchers do *something* which affects the data collected from subjects
- ***Prospective:*** subjects identified in advance and data collected over time
- ***Retrospective***: subjects identified in the present based on data available from the past

Population Definitions

- ***Study Population:*** all cases that could possibly have been in the data

Population Definitions

- ***Study Population:*** all cases that could possibly have been in the data
- ***Sample Population:*** all cases that were actually in the data

Population Definitions

- ***Study Population:*** all cases that could possibly have been in the data
- ***Sample Population:*** all cases that were actually in the data
- ***Target Population***: cases we'd like to apply the conclusions from our data to

Evaluating Study Data

- ***Reliable***: is our data a reasonably good representation of the sample population?

Evaluating Study Data

- ***Reliable***: is our data a reasonably good representation of the sample population?
- ***Valid***: is our sample population a reasonably good representation of the study population?

Evaluating Study Data

- ***Reliable***: is our data a reasonably good representation of the sample population?
- ***Valid***: is our sample population a reasonably good representation of the study population?
- ***Generalizable***: is our study population a reasonably good representation of the target population?

Reliability

Is it reasonable to think that the sample statistics from our sample would accurately describe the distribution of the sample population?

Reliability

Is it reasonable to think that the sample statistics from our sample would accurately describe the distribution of the sample population?

- Is data self-reported or otherwise subjectively measured?
- Are variables poorly defined and/or measured?
- Were cases non-randomly sampled or is data missing?
- When sample sizes are small, unusual or extreme data points have a large impact

Validity

Is it reasonable to think that a sampling distribution based on the sample statistics from our sample would accurately describe the expected distribution in the study population?

Validity

Is it reasonable to think that a sampling distribution based on the sample statistics from our sample would accurately describe the expected distribution in the study population?

- Were subjects non-randomly sampled?
- Is data missing or were subjects excluded from the sample population?
- Is the sample data unreliable?
- Sampling distributions based on small sample sizes will be less accurate and less precise estimators

Generalizability

Is it reasonable to think that a sampling distribution based on the sample statistics from our sample would accurately describe the expected distribution in the target population?

Generalizability

Is it reasonable to think that a sampling distribution based on the sample statistics from our sample would accurately describe the expected distribution in the target population?

- How different is the target population from the study population?
- Sampling distributions based on small sample sizes are less likely to be accurate for large, more diverse target populations
- When data is unreliable and/or invalid, conclusions may also not be generalizable

Homework 1: Problem 2

Researchers in the UK wanted to answer the question of how much crime there was in Britain and whether it was going up or down. They used 2 different approaches to gather data for their investigation, but they need help determining the validity of their approach.

Data Set 1

The Crime Survey for England and Wales is a survey in which approximately 38,000 people are questioned about their experiences with crime. People surveyed are 16 years of age or older and were not living in communal residences. Answers are self-reported.

Data Set 2

UK Police keep administrative records of crimes they have investigated. Police use internal definitions of crimes and their discretion when creating these records.

Homework 1: Study Population

Data Set 1

People from England and Wales who are age 16+ and not living in communal residences

Data Set 2

Crimes committed in the UK which were reported to and recorded by police and/or the victims of those crimes.

Homework 1: Sampling Strategy

Data Set 1

Data came from subjects who responded to a survey, so the data was collected via *voluntary response*.

Data Set 2

Data came from all the records that were available, so the data was collected via a *census*.

Never assume a random sample was taken unless it explicitly says so.

Homework 1: Sample Population

Data Set 1

38,000 people from England and Wales that were age 16+ and not living in communal residences who responded to the survey

Data Set 2

Crimes committed in the UK which were reported to and recorded by police and/or the victims of those crimes.

Homework 1: Target Population

From the instructions: “Researchers in the UK wanted to answer the question of how much crime there was in Britain and whether it was going up or down.”

The target population is crime in Britain and/or the potential victims of that crime.

Homework 1: Reliability and Validity

Both data sets were self-report

Homework 1: Reliability and Validity

Both data sets were self-report

- Subjects self-report survey responses, recall events in the past
- Both victims AND police self-report incidents of crime
- Police rely on internal definitions of crime

Homework 1: Reliability and Validity

Both data sets were missing data

Homework 1: Reliability and Validity

Both data sets were missing data

- People who choose not to respond to the survey aren't represented
- Crimes that are not often reported and their victims will be underrepresented
- Police may have an incentive to not record all crimes reported
- Subjects were non-randomly sampled for the survey

Homework 1: Generalizability

- Not many cases excluded from study populations, large sample sizes
  - Broad eligibility criteria for survey responses
  - All data available for police records

Homework 1: Generalizability

- Not many cases excluded from study populations, large sample sizes
  - Broad eligibility criteria for survey responses
  - All data available for police records
- May not be generalizable to crimes against children/businesses/property, crimes which are underreported or not well documented by police, victims of crime in the UK who are not residents