L10: Midterm 1 review

What have we learned so far?

What have we learned so far?

Key topics covered on midterm :

- 1 Course overview, PPDAC
- 2 Describing data structure, visualization(bar chart and histogram)
- 3 Describing data with numbers
- 4 Scatterplots and correlation
- 5 R basics working with data and visualizing data
- 6 Intro to regression
- 7 Two categorical variables
- 8 Samples and observational studies
- 9 Designing Experiments and Ethics

Lecture 1

Problem, Plan, Data, Analysis, Conclusion

You should know these terms.

You should also be familiar with the three types of problems we talked about and be able to identify them.

Lecture 1: Types of problems

Q1: We are interested in projecting the number of immunization doses that will be needed in a clinic during the month of November based on the previous year's data.

Q2: We are creating a visualization of the mean exam scores in PH142 by program and year of student.

Q3: We are planning an intervention to reduce e-cigarette use and assessing the role of exposure to advertisements in e-cigarette use among young people.

Lecture 2

R basics working with data and visualizing data

You should be able to recognize what a code chunk is doing, and be able to fill in blanks in a code chunk for syntax we have seen multiple times in lecture, lab and problem sets.

Lecture 2

What function would i use to do the following:

restrict my dataset to a smaller number of variables?

restrict my datset to a smaller set of observations?

check to see how many observations are in the dataset?

Create a new variable?

types of variables

Identifying the unit of analysis

Differentiating between the types of variables

Types of studies and types of variables

from Iverson et al. Abstract Exposure to industrial solvents has been associated with encephalopathy. Styrene is a neurotoxic industrial solvent, and we investigated the long-term risk of encephalopathy and unspecified dementia following styrene exposure. We followed 72,465 workers in the reinforced plastics industry in Denmark (1977–2011) and identified incident cases of encephalopathy ($n = 228$) and unspecified dementia ($n = 565$) in national registers. Individual styrene exposure levels were modeled from information on occupation, measurements of work place styrene levels, product, process, and years of employment.

check your knowledge Lectures 1 and 2 and 8/9

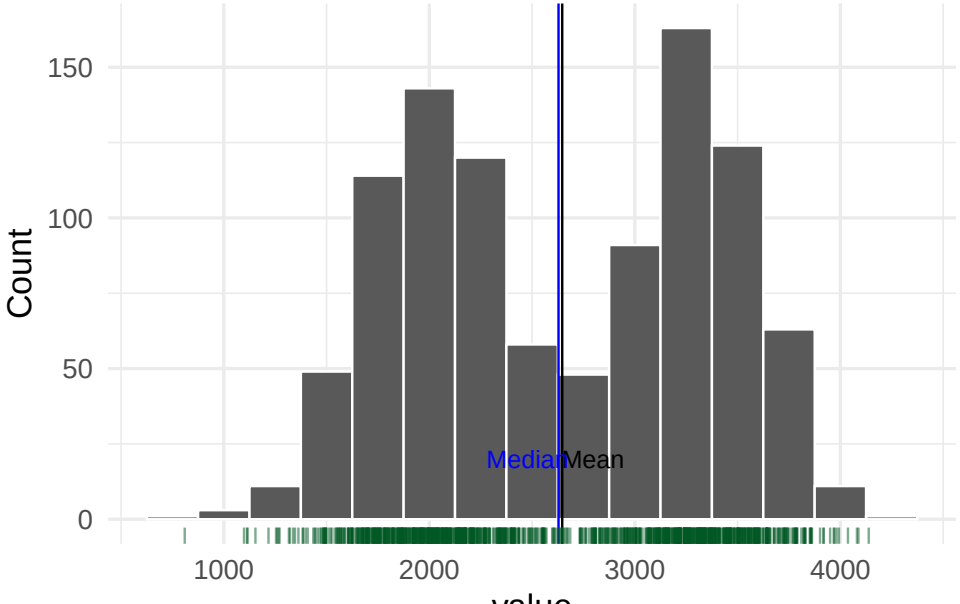
What type of problem is being addressed here (descriptive, causative, predictive?) What type of variable is the outcome? What kind of a study design is this?

L03 types of visualizations

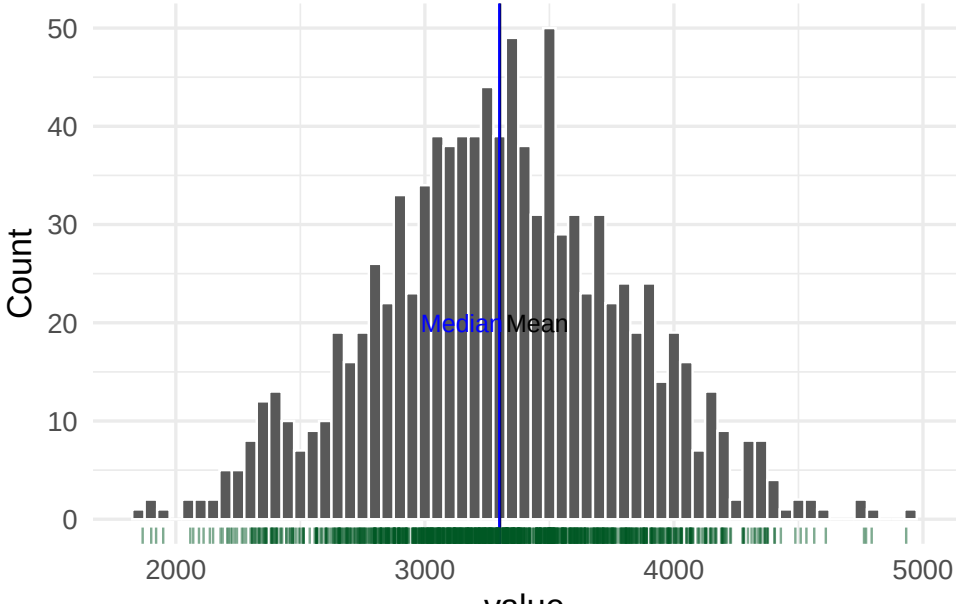
Creating visual summaries of variables

- Make bar charts using ggplot()'s geom_bar()
- Make histograms using ggplot()'s geom_histogram()
- Make scatterplots using ggplot()'s geom_point()

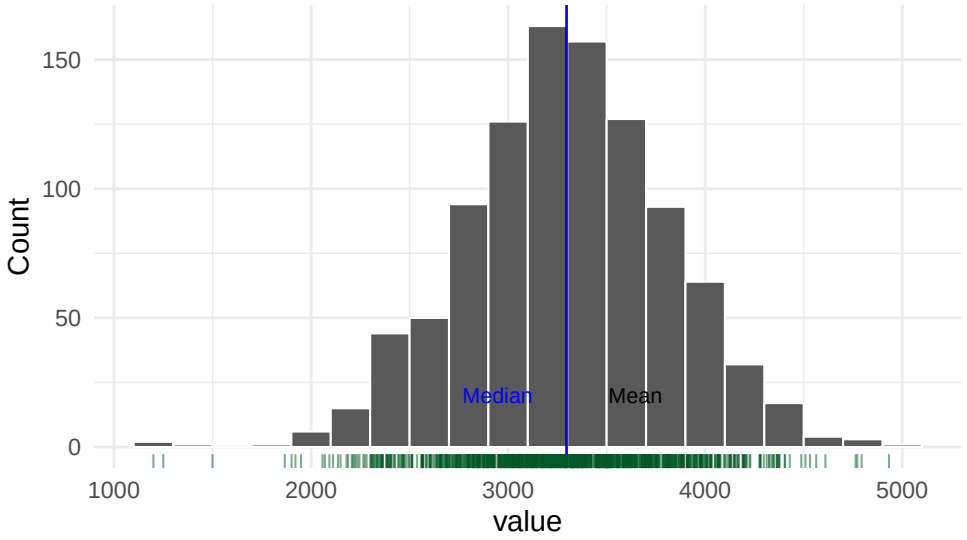
Examples



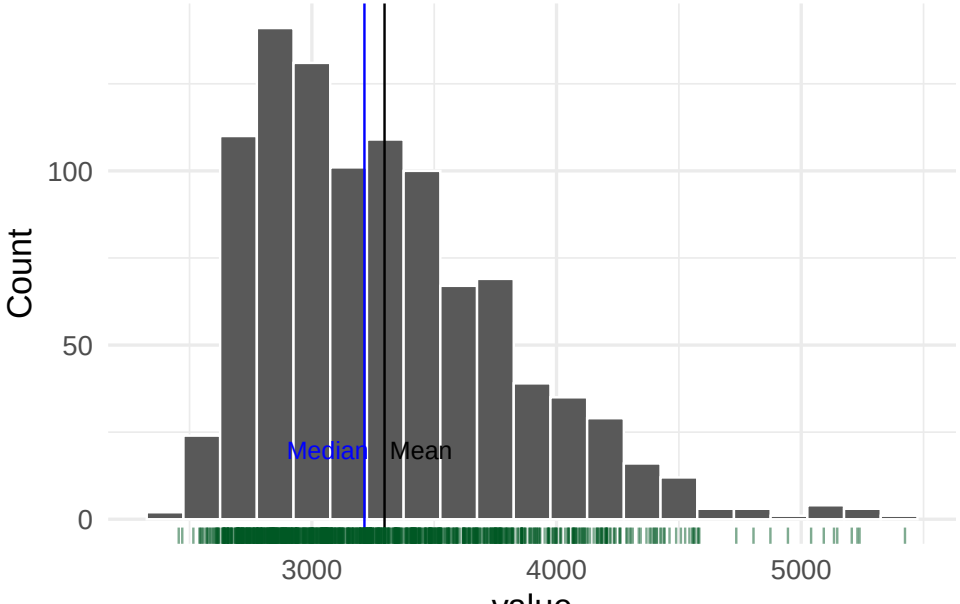
Examples



Examples



Examples



L04 Describing data with numbers

1. Investigate measures of centrality
   - mean and median, and when they're the same vs. different
2. Investigate measures of spread
   - IQR, standard deviation, and variance
3. Create a visualization of the "five number summary"
   - boxplots using `ggplot`
4. Calculate the variance and standard deviation

L05 Scatterplots and correlation

- Determine which variable is explanatory and which is response, or when it doesn't matter
- Visually describe the relationship between two variables (form, direction, strength, and outliers)
- Numerically describe the relationship with the correlation coefficient r

Lecture 6 Intro to regression

Be able to pull relevant pieces of information from r output and interpret them including - intercept - slope - r squared

Know how to use the equation to find predicted values of Y at a given X

Equation of the line of best fit

The line of best fit can be represented by the equation for a line:

$$y = a + bx$$

where a is the intercept and b is the slope.

Linear regression

Below are images of the output from code that will run a linear regression model on the relationship between age and `charges` for the subset population and produce the following outputs. Here the population is subset to smokers who are of normal BMI.
We will assign the regression model to the name `insure_model`. Write the missing commands to fill in blanks in the code

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   10609.    1325.       8.01 2.14e-10
## 2 age             244.      33.0      7.41 1.74e- 9
```

insure_model <- ___(formula = charges ~ age, data = insure_subset)

*answer is lm

Linear regression

Write the equation for the line of best fit for the subset data. Interpret the slope parameter, in one sentence what does the slope parameter tell you.

Using the model, predict the medical charges for a smoker of normal BMI who is 30 years old.

Example 2:

Below is a dataframe 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 plant electrodes into both brain regions for several individuals and records the activity level during different independent stimuli. The results are shown below.

```
## # 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
```

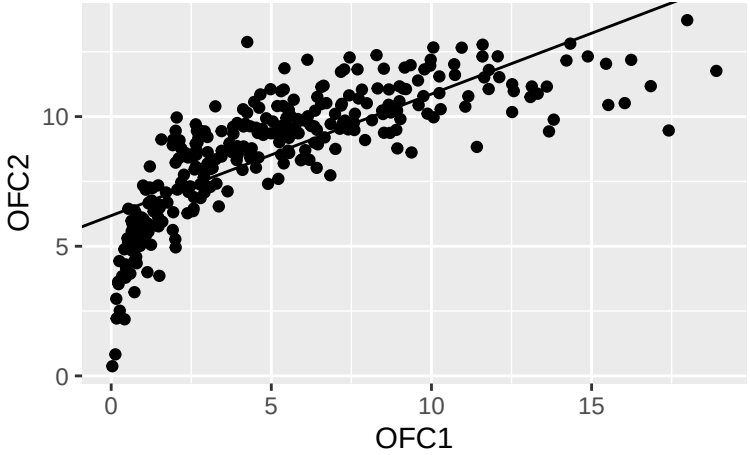
Q2.1 [2 pts] You decide to fit a linear regression to the data. Fill in the blanks to generate the output shown below.

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

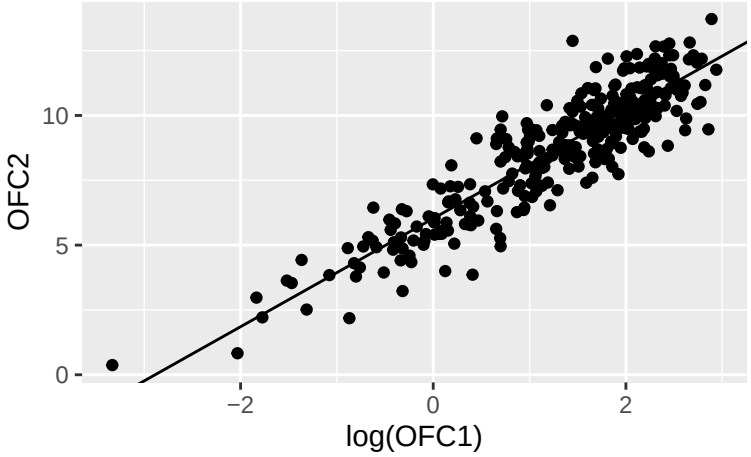
SOLUTION: lm OFC2 OFC1 tidy

Q2.2 [1 pt] 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 OCF1 and OCF2.



SOLUTION: strength: weak/moderate; form: steep then flattens (not linear);

Q2.3 [2 pts] You then perform a logarithmic transformation on the OFC1 data points and generate a new model. Without being given any numbers, which plot (original vs. transformed) has a stronger correlation coefficient? Why?



Q2.4 [1 pt] Which of the following is a plausible value for the correlation coefficient between log(OFC1) and OFC2?

- 0.6
- 0.9
- 1.0
- 0.3

Solution: b.

Q2.5 [1 pt] 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
```

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

Q2.6 [1 pt]

You are given one more observation of an individual with an OFC1 value of 40.14 and would like to use your transformed model to predict their OFC2 value. Calculate the predicted OFC2 value and round to two decimal places. Is this prediction appropriate given your data? Why or why not?

Solution: 6.03 + 2.09*log(OFC1) = 13.80865 Not necessarily - extrapolation

Confounding

A relationship between your variable of interest (exposure, treatment) and your outcome of interest (disease status, health condition etc) is confounded when there is a variable that is associated with both the exposure and outcome, and is not on the causal pathway between the two.

Variables that are on the causal pathway are those that represent a way in which the exposure acts on the outcome.

Lecture 7 Two categorical variables

- Two way tables
  - marginal vs conditional distributions
- Bar graphs
  - side by side
  - stacked
- Simpson's paradox

categorical variables

Suppose I am a teacher in elementary school and there is a head lice outbreak at my school. I know that some of the children went on a field trip recently to a fire station. While they were at the fire station they were allowed to pass around a fire helmet and try it on. I suspect that this caused the outbreak. I collect the following data:

Group	head lice	no head lice
Field trip	49	62
No Field trip	15	89

Calculate the conditional probability of head lice among students who attended the field trip, and among those who did not.

categorical variables

solution: among field trip $49/(49+62) = 0.44$ among non filed trip $15/(15+89) = 1.14$

Categorical vars

Why are we interested in the conditional rather than the marginal probabilities here?

What do these data suggest?

Solution:
We want to know if there is a potential association between the field trip and head lice These data suggest there is - the conditional probability is much higher in the field trip group.

Lecture 8 Samples and observational studies

- Whether the treatment or exposure is controlled by an investigator
  - Experimental vs observational designs
- The population of interest
  - Target population
  - Study population
- How the sample was drawn from the population
  - Complete sample (census)
  - Random sampling
  - Convenience sampling
  - Volunteer sampling
- Was selection conditional on exposure or outcome

Lecture 9 Desigining Experiments

- Understanding different types of experimental designs
- Thinking about sources of bias
  - bias from the design or conduct of sampling
  - bias from lack of adherence to protocol
  - bias in assessment
  - bias in analysis