STAT011 Statistical Methods I

Lecture 13 Midterm I Review

Lu Chen
Swarthmore College
3/5/2019

Structure of a data set

Data set structure

- **Cases/Observations**: usually by rows
  - The objects described by a set of data. They can be customers, companies, subjects in a study, units in an experiment, or other objects.
- **Variables**: usually by columns
  - Characteristics of cases/observations
    - **Quantitative variable**: numerical values
    - **Categorical variable**: several groups or categories
    - **Label variable**: A special variable used in some data sets to distinguish different cases/observations. For example, names, IDs. Each observation has a **unique** value.

STAT011 Lecture 2 | Lu Chen | 1/24/2019 | 3 / 30

Exploratory data analysis

Exploratory Data Analysis		No Explanatory	Explanatory	
			Categorical	Quantitative
Response	Categorical	• Table of counts and proportions • Bar plot • Pie chart *(Lecture 2)*	• Two-way tables - Joint distribution - Marginal distribution - Conditional distribution • Bar plot *(Lecture 6)*	____
	Quantitative	• Mean, SD • Median, IQR • Histogram, density curve • Boxplot *(Lecture 2~4)*	• Table of summary statistics • Histogram, density curve • Boxplot *(Lecture 7)*	• Correlation • Regression • Scatterplot (Lecture 5~6)

- What are the summary statistics and visualization methods for a single variable and the relationship between two variables?

Difference between mean and median

Quantitative variables - Mean and Median

Comparing Mean and Median

Statistic	Mean	Median
Pros	Taking all the values into account	Resistant to extreme values
Cons	Sensitive (not resistant) to extreme values	Losing information on data values

- There is no definite answer which one is better.
  - Mean is better for roughly symmetric distributions;
  - Median is better for skewed distributions with extreme values.
- In exploratory data analsyis, people usually look at both of them.

- For a variable with left/right-skewed distribution, would the mean greater or smaller than the median?

Effect of linear transformations

Linear transformations

A linear transformation changes the original variable X into the new variable X_{new} given by an equation of the form

$$X_{new} = a + bX$$

Adding the constant a shifts all values of X upward or downward by the same amount. In particular, such a shift changes the origin (zero point) of the variable. Multiplying by the positive constant b changes the size of the unit of measurement.

- **Question**: A variable X with values $1, 2, \cdots, 10, 15$ has
  mean $\bar{x} = 6.4$, standard deviation $s = 4.1$,
  median $M = 6$, quartiles $Q_1 = 3.5$ and $Q_3 = 8.5$
  and IQR $Q_3 - Q_1 = 5$.
  What are the mean, SD, median, quartiles and IQR for variable $Y = 3 + 2X$?

Effect of linear transformations

Effect of linear transformations

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (standard deviation and interquartile range) by b.
- Adding the same number a (either positive or negative) to each observation adds a to measures of center and to quartiles and other percentiles but does not change measures of spread.

- Therefore, For the variable $Y = 3 + 2X$,
  $\bar{y} = 3 + 2 \times 6.4 = 15.8, s = 2 \times 4.1 = 8.2,$
  $M = 3 + 2 \times 6 = 15, Q_1 = 3 + 2 \times 3.5 = 10, Q_3 = 3 + 2 \times 8.5 = 20$
  and $IQR = 2 \times 5 = 10$

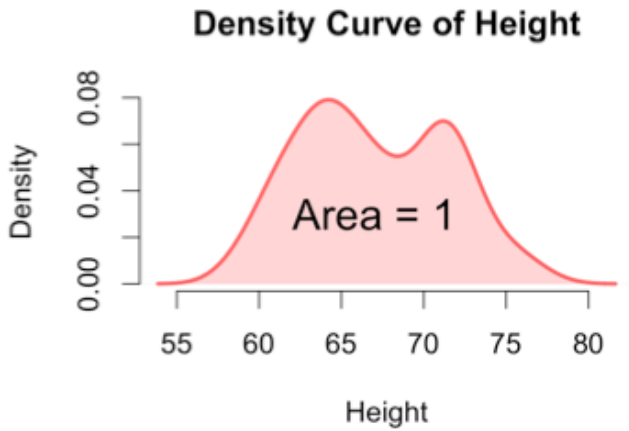
- This is then applied to standardization of a Normal variable. If
  $X \sim N(\mu, \sigma)$,
  $Z = \frac{X-\mu}{\sigma} \sim N(0, 1$

Normal distribution

Density curve

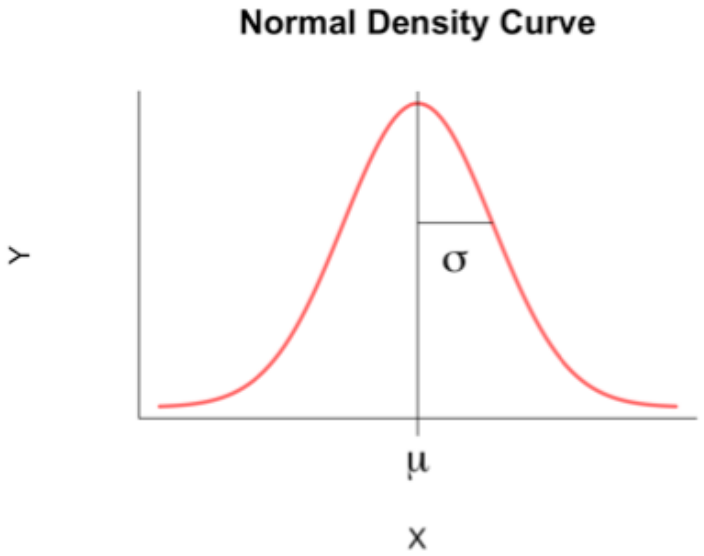
A **density curve** describes the overall pattern of a distribution. The **area** under the curve and above any range of values is the **proportion** of all observations that fall in that range.

- It is always on or above the horizontal axis.
- It has area exactly 1 underneath it.



Normal distribution

Normal distribution



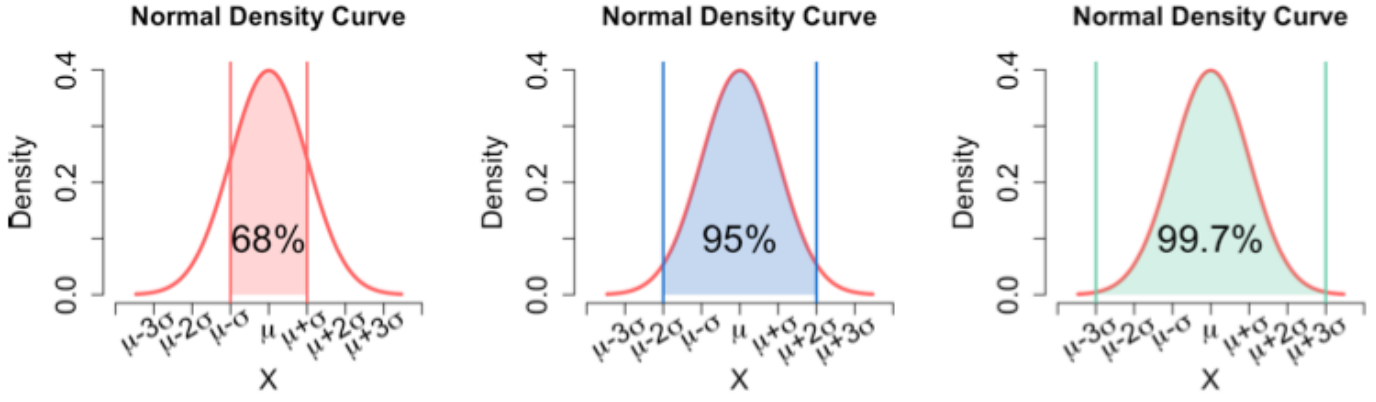
The Normal density curve is characterized by

- Center μ: mean of the distribution
- Spread σ: standard deviation of the distribution

and expressed as $X \sim N(\mu, \sigma)$, "X follows a Normal distribution with mean μ and standard deviation σ".

Normal distribution

The 68-95-99.7 rule



- The area under the curve within σ of μ is 0.68.
- The area under the curve within 2σ of μ is 0.95.
- The area under the curve within 3σ of μ is 0.997.

STAT011 Lecture 4 | Lu Chen | 1/31/2019 | 13 / 29

- Use the rule to calculate the area under the curve (proportion) within an interval.

Normal distribution

Standardization

If a variable X has **any** Normal distribution $N(\mu, \sigma)$ with mean μ and standard deviation σ, then the **standardized** variable

$$Z = \frac{X - \mu}{\sigma}$$

has the standard Normal distribution.

The tranformation from X to Z is called **standardization**.

- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.
- For example, $X \sim N(5, 2)$, then $Z = \frac{X-5}{2} \sim N(0, 1)$.

Normal distribution

Standardization

- Upper case X and Z usually denote variables.
- Lower case x and z usually denote specific values from the variables.

If x is an observation from $X \sim N(\mu, \sigma)$, the **standardized value** of x is

$$z = \frac{x - \mu}{\sigma}$$

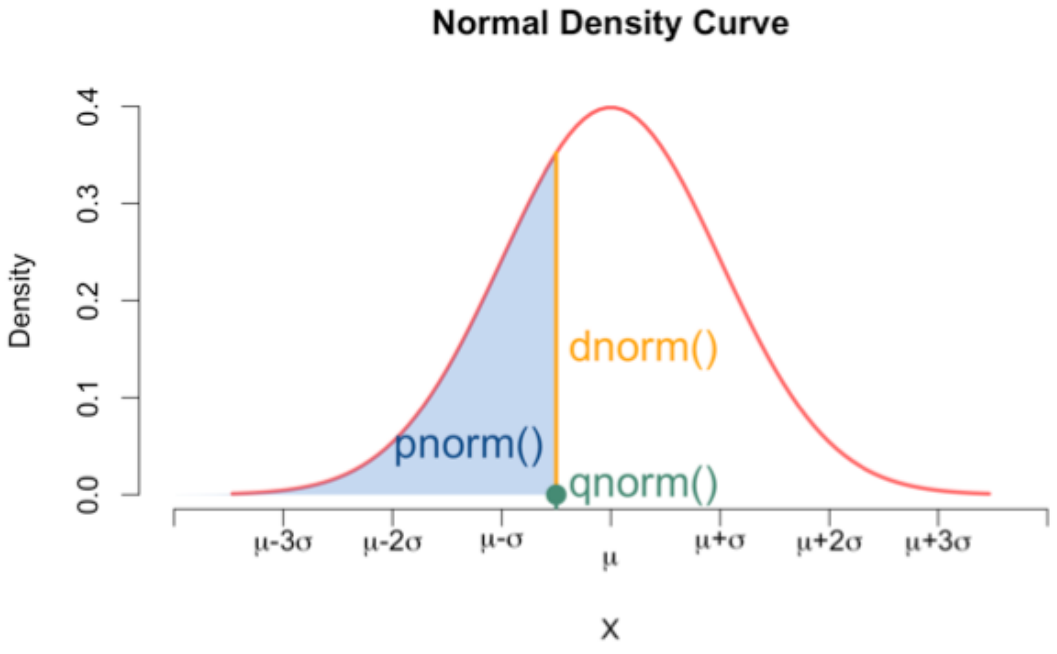
It is called the **standardized score of** x, or z**-score**.

- For $X \sim N(5, 2)$, suppose $x = 3$ is an observation from X. Then $z = (x - 5)/2 = (3 - 5)/2 = -1$is the z-score of x.
- What about $x = 10$? If $z = 0.5$, what's the corresponding x value?
- $x = 10 \Rightarrow z = (10 - 5)/2 = 2.5$

  For $z = 0.5, z = \frac{x-5}{2} = 0.5 \Rightarrow x = 2 \times 0.5 + 5 = 6$

Normal distribution

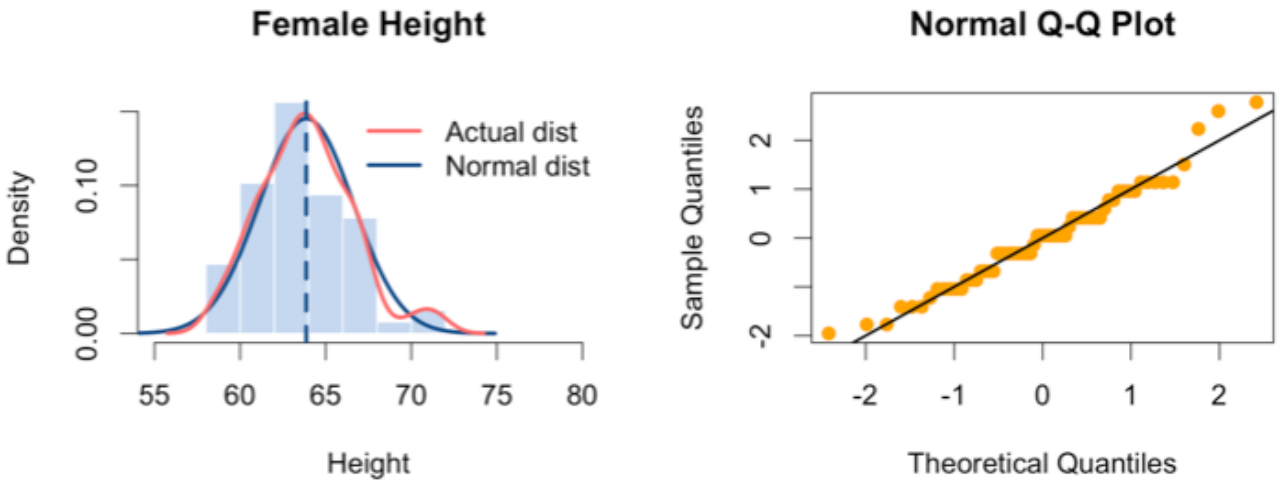
Normal distribution in R



Normal distribution

Assessing Normality

Normal Quantile-Quantile Plot



- Normal Q-Q plot compares the distribution of interest (usually after standardization) to the standard Normal distribution.
- If the distribution of interest is close to a Normal distribution, the points on the Q-Q plot should **lie close to the $y = x$ line**.

STAT011 Lecture 4 | Lu Chen | 1/31/2019 | 27 / 29

- Normal Q-Q plot is later used to assess the distribution of the residuals of a regression model and the sampling distribution of a sample mean.

Correlation coefficient

Correlation coefficient

The **correlation** measures the *direction* and *strength* of the **linear relationship** between two quantitative variables. Correlation is usually written as ***r***.

Suppose that we have data on variables X and Y for n individuals. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between X and Y is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

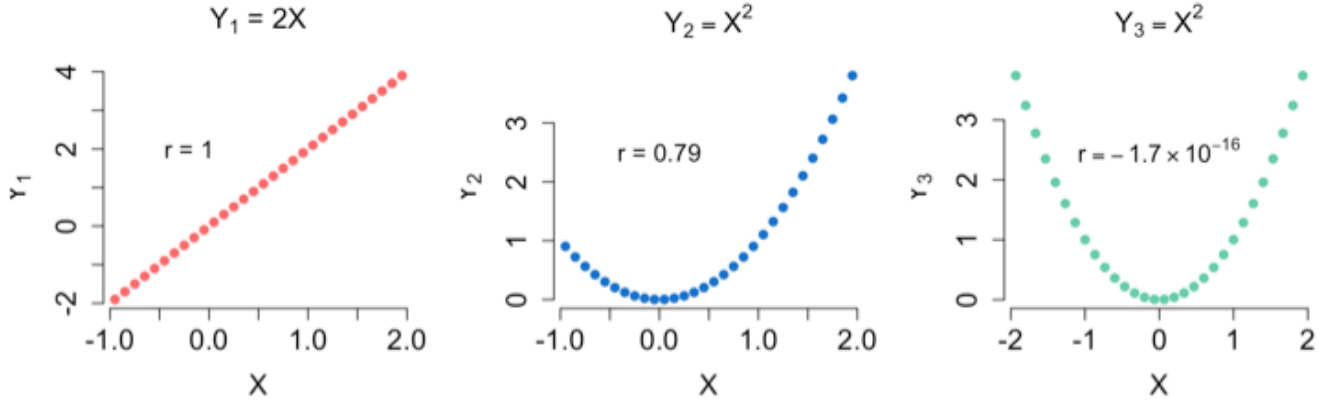
- $\frac{x_i-\bar{x}}{s_x}$ and $\frac{y_i-\bar{y}}{s_y}$: standardized x and y values - no units.
- $\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$ can be positive or negative.

- $-1 \leq r \leq 1$
- $r = \pm 1$: perfect relationship
- $r = 0$: no relationship

Correlation coefficient

Correlation coefficient

- When calculating correlation, there is no distinction in explanatory or response variable.
- Both varaibles must be quantitative.
- Linear transformation does not alter the value of correlation.
- Correlation only captures the linear relationship betweem two variables.



Least-squares regression

Least Square Regression (LSR)

> The **least-squares regression** line of y on x is the line that minimizes the sum of the squares of the vertical distances from the data points to the line.

- Observed data (x_i, y_i) for the i^{th} data point; the total number of data points is n.
- Predicted values $\hat{y_i} = b_0 + b_1 x_i$ for the i^{th} data point.
- Vertical distance from the data points to the line:

$$\text{Residual} = \text{Observed } y - \text{Predicted } y$$
$$e = y - \hat{y}$$
$$e_i = y_i - \hat{y_i}$$

- In least squares regression, we minimize

$$\sum(\text{residual})^2 = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Least-squares regression

Assessing LSR - Coefficient of determination

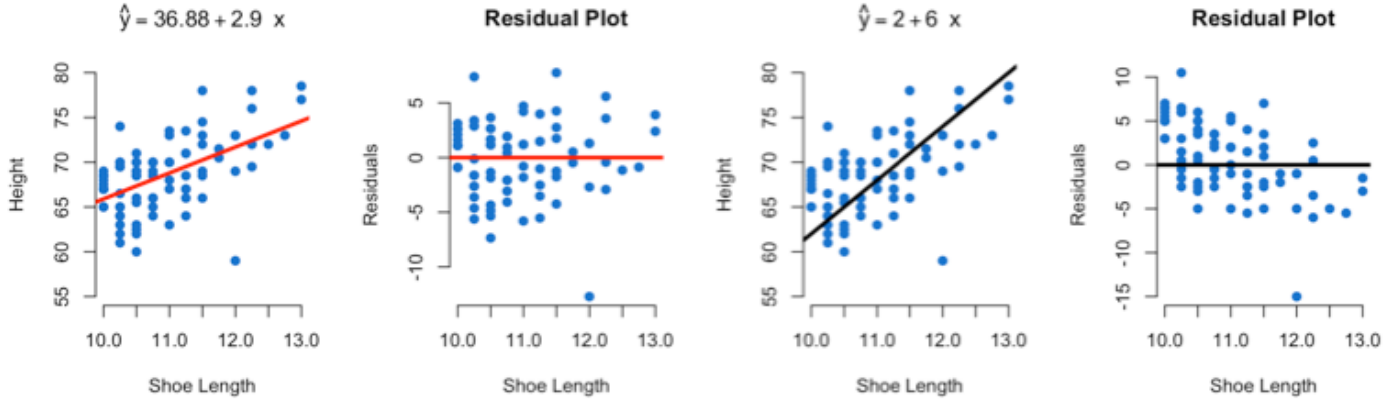
$$r^2 = \frac{\text{Variance}(\hat{y})}{\text{Variance}(y)}$$

- In terms of value, r^2 is the square of r; they are essentially the same.
- In terms of meaning/interpreation, they are **different**.
  - The correlation r measures the direction and strength of a linear relationship. *No regression is involved.*
  - The coefficient of determination r^2 measures the fraction of variation in y explained by the least squares regression line $\hat{y}$. *It directly assesses a regression line.*

- $r^2 = 0.7.70\%$ of the variability in the response variable is explained by the regression model.

Least-squares regression

Assessing LSR - Residual plot

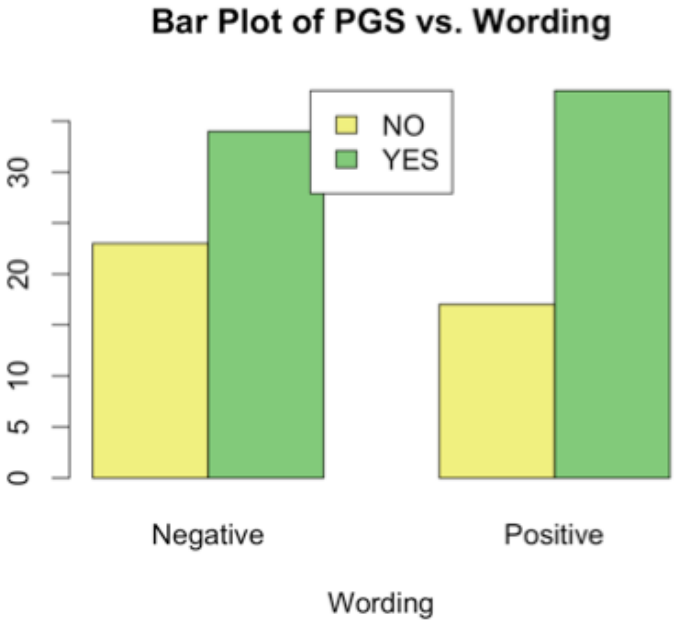


- If the regression line catches the overall pattern of the data, there should be *no pattern* in the residuals.
- If the residual plot shows any pattern, the regression line is NOT the best way to describing the data.

Relationship: cateogrical vs. categorical

Personal Genome Services (*PGS*) and *Wording*

PGS	Negative	Positive	Total
NO	23	17	40
YES	34	38	72
Total	57	55	112



- Two-way table: cell count, column total, row total, table total.
- Bar plot: the explanatory variable should be placed on x-axis.

Relationship: cateogrical vs. categorical

Personal Genome Services (*PGS*) and *Wording*

Joint and marginal distribution

PGS	Negative	Positive	Total
NO	21%	15%	36%
YES	30%	34%	64%
Total	51%	49%	100%

Conditional distribution

PGS	Negative	Positive
NO	40%	31%
YES	60%	69%
Total	100%	100%

- The conditional distribution suggests the relationship between the two categroical variables.

Relationship: Quantitative vs. categorical

Relationship: Quantitative vs. Categorical

Height	Female	Male	Other
Mean	65.0	70.0	63.0
SD	3.4	3.6	NA

Coffee	Fr	So	Jr	Sr
Median	0.0	0.0	1.0	2.0
IQR	3.0	5.0	3.0	2.5

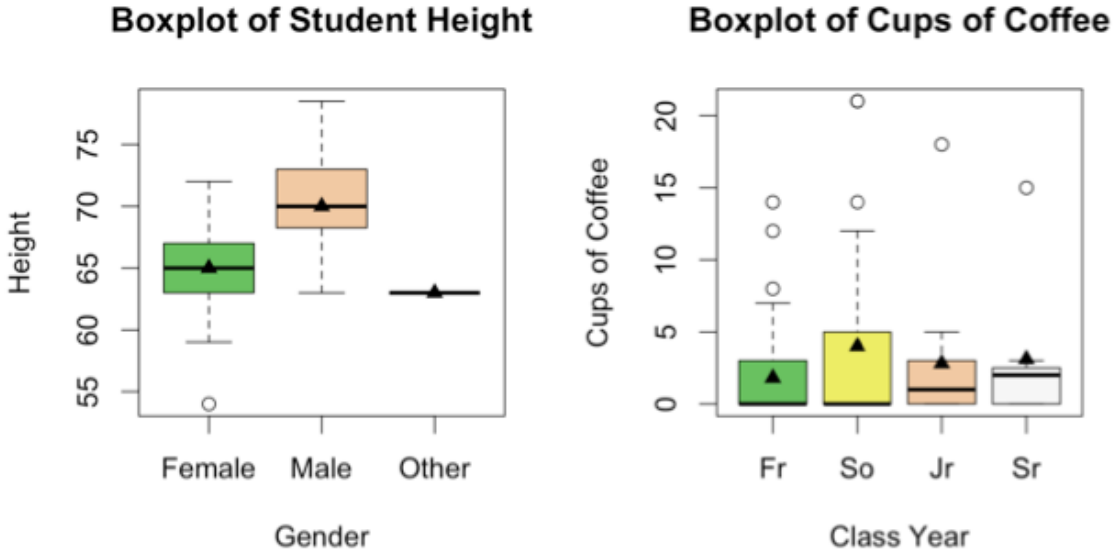
Why do we report mean and SD for *Height* but median and IQR for *Coffee*?

- Table of summary statistics by categories.

Relationship: Quantitative vs. categorical

Relationship: Quantitative vs. Categorical

```
boxplot(Response ~ Explanatory, data= , col= , main= , xlab= , ylab= )
points(c(mean1, mean2, mean3, ...), pch= , col= )
```

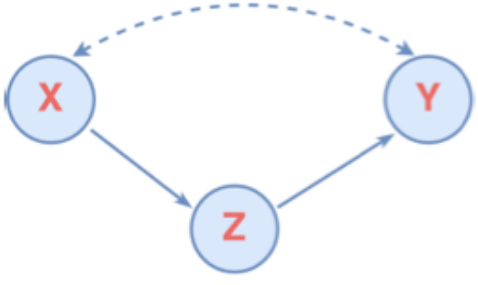


Association and causation

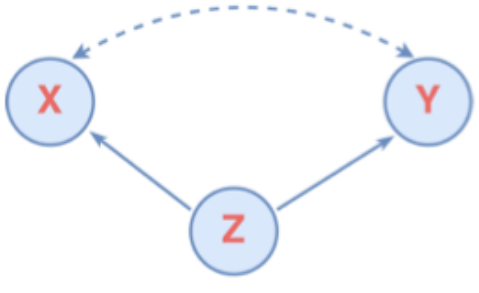
Types of associations

Direct Causation

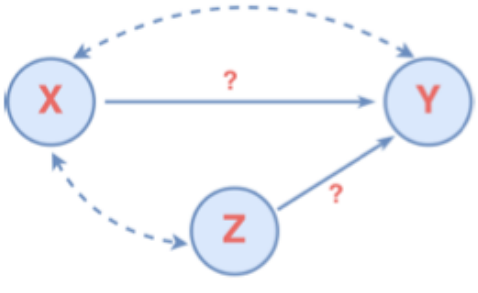
Mediation



Common Response



Confounding



- Association does not imply causation.

Design of experiments sampling design

Design of experiments

- Comparative experiment
- Matched pairs design
- Block design

Sampling design

- Simple random sample
- Stratified random sample
- Multistage random sample

Use of the `sample()` function

Bias and variability

Lecture 8

> The design of a study is **biased** if it systematically favors certain outcomes.

Lecture 9

> **Bias** concerns the center of the sampling distribution.
>
> $$\textbf{Bias} = \textbf{Mean of the statistic} - \textbf{Population parameter}$$
>
> A statistic used to estimate a parameter is an **unbiased estimator** if its mean is equal to the true value of the parameter being estimated.

> The **variability** of a statistic is described by the spread of its sampling distribution. This spread is determined by the sampling design and the sample size n. Statistics from larger samples have smaller spreads.

Bias and variability

Manage bias and variability

For SRS,

- The mean of the statistic is always close to the population parameter. We reasonally infer that SRS is an unbiased sampling procedure.
- Larger sample size will always result in smaller spread of the statistic.

Therefore,

> To **reduce bias**, use random sampling. When we start with the entire population, simple random sampling produces unbiased estimates – the values of a statistic computed from an SRS neither consistently overestimate nor consistently underestimate the value of the population parameter.
>
> To **reduce the variability** of a statistic from an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

Sampling distribution of a sample mean

Sampling distribution of a sample mean

Let $\bar{x}$ be the mean of an SRS of size n from a population having Normal distribution with mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

$$\mu_{\bar{x}} = \mu,$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

And $\bar{x}$ has the $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution.

This says that if $X \sim N(\mu, \sigma)$, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Central limit theorem (CLT)

Central Limit Theorem

Draw an SRS of size n from **any population** with mean μ and finite standard deviation σ. When ***n* is large**, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Central limit theorem holds for data with any population distribution.
- **Amazing Fact!**

Central limit theorem (CLT)

Sample size for CLT to hold

Population distribution	Sample size n	Sampling distribution of $\bar{x}$
Exactly Normal	Any	Exactly Normal
Close to Normal	$n > 20$	Approximately Normal
Highly skewed	$n > 60$	Approximately Normal

- These suggestions for samples size are only rules of thumb.
- The appropriate sample size always depends on each specific problem.
- Generally, as sample size n increases, the sampling distribution of mean $\bar{x}$ will become **less variable** and **more Normal**.

Central limit theorem (CLT)

Bernoulli Distribution

- Denote population proportion as p and sample proportion as $\hat{p}$.

	Mean	Standard deviation	Proportion
Population Parameter	μ	σ	p
Sample Statistic	$\bar{x}$	s	$\hat{p}$

- A dummy variable X with values 0 and 1 follows a **Bernoulli distribution**

$$X \sim Bernoulli(p)$$

- The proportion of $X = 1$ (usually called success) is p.
- The proportion of $X = 0$ (usually called failure) is $1 - p$.
- The mean of X is p.
- The SD of X is $\sqrt{p(1-p)}$.

Central limit theorem (CLT)

Sampling distribution of a proportion

Normal approximation for proportions

Draw an SRS of size n from a large population having population proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. When n is large, the sampling distribution of $\hat{p}$ is approximately Normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$:

$$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Central limit theorem (CLT)

Review

Central Limit Theorem

- Population distribution is Normal, $X \sim N(\mu, \sigma)$,

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Population distribution is not Normal, $\mu_X = \mu$, $\sigma_X = \sigma$,

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  - Population distribution is Bernoulli, $X \sim Bernoulli(p)$, $\mu_X = p$, $\sigma_X = \sqrt{p(1-p)}$,

$$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

STAT011 Lecture 11 | Lu Chen | 2/26/2019 | 2 / 25

Confidence interval for a population mean

Confidence interval for a population mean

Choose an SRS of size n from a population having unknown mean μ and known standard deviation σ. The margin of error for a level C confidence interval for μ is

$$m = z^* \frac{\sigma}{\sqrt{n}}.$$

Here, z^* is the value on the standard Normal curve with area C between the critical points $-z^*$ and z^*. The **level C confidence interval for μ** is

$$\bar{x} \pm m = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}.$$

The confidence level of this interval is exactly C when the population distribution is Normal and is approximately C when n is large in other cases.

Significance test for a population mean

z Test for a population mean

To test the hypothesis $H_0 : \mu = \mu_0$ based on an SRS of size n from a population with unknown mean μ and known standard deviation σ, compute the **test statistic**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of a standard Normal random variable Z, the P-value for a test of H_0 vs.

$$\begin{aligned} H_a &: \mu > \mu_0 \quad \text{is } P(Z \geq z) \\ H_a &: \mu < \mu_0 \quad \text{is } P(Z \leq z) \\ H_a &: \mu \neq \mu_0 \quad \text{is } 2P(Z \geq |z|) \end{aligned}$$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.