MATH 3339
Statistics for the Sciences
Live Lecture Help

James West
jdwest@uh.edu

University of Houston

Session 4

Office Hours: see schedule in the "Office Hours" channel on Teams
Course webpage: www.casa.uh.edu

Email policy

When you email me you **MUST** include the following

- MATH 3339 Section 20024 and a description of your issue in the **Subject Line**
- Your name and ID# in the **Body**
- Complete sentences, punctuation, and paragraph breaks
- Email messages to the class will be sent to your Exchange account (user@cougarnet.uh.edu)

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/

Updates

- Test 1 scheduling opens 9/23 at 12 AM.
- Homework 2 is available

Outline

1. Recap
2. Examples
3. Student submitted questions

Parameters and Statistics

- A **parameter** is a number that describes the **population**. A parameter is a fixed number, but in practice we usually do not know its value.
- A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample.
- The purpose of sampling or experimentation is usually to use statistics to make statements about unknown parameters, this is called **statistical inference**.

Notation of Parameters and Statistics

Name	Statistic	Parameter
mean	$\bar{x}$	μ (mu)
standard deviation	s	σ (sigma)
correlation	r	ρ (rho)
regression coefficient	$\hat{\beta}$ or b	β (beta)
proportion	$\hat{p}$	p

Measures of Variability

Population Variance

The population variance is defined as

$$\sigma^2 = \frac{1}{N}\left[(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_N-\mu)^2\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

and the population standard deviation is given by $\sigma = \sqrt{\sigma^2}$, the square root of the population variance.

In R: $\text{var}(x)\left(\frac{N-1}{N}\right)$

Measures of Variability

Sample Variance

The sample variance is defined as

$$s^2 = \frac{1}{n-1}\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2\right]$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

and the sample standard deviation is given by s, the square root of the sample variance. The reason for modifying the definition for the sample variance has to do with its properties as an estimate of the population variance.

in R: var(x)

Measures of Variability

The Coefficient of Variation

The coefficient of variation measures **relative variability**.

$$cv(x) = \frac{\sigma(x)}{\mu(x)} \text{ or } = \frac{s}{\bar{x}}$$

The advantage of the CV is that it is unitless. This allows CVs to be compared to each other in ways that other measures cannot be. It is used frequently in regression analysis.

Measures of Variability

Mean Absolute Deviation

The mean absolute deviation is defined as

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - median(x)|$$

Median Absolute Deviation

The median absolute deviation (mad) is defined as

$$mad(x) = median(|x - median(x)|)$$

Note:In R, the mad is adjusted by a constant factor of 1.4826.

In R: mad(x, constant = 1)

Jointly Distributed Variables

Covariance and Correlation

If x and y are jointly distributed numeric variables, we define their covariance as

$$cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu(x))(y_i - \mu(y))$$

for a population and

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

for a sample.

Covariance is a measure of how changes in one variable are associated with changes in a second variable.

Jointly Distributed Variables

Covariance and Correlation

If x and y are jointly distributed numeric variables, we define their correlation as

$$cor(x,y) = r = \frac{cov(x,y)}{sd(x) \cdot sd(y)}$$

The correlation is always such that $-1 \leq cor(x,y) \leq 1$.

The correlation indicates the strength and direction of a linear relationship.

If $\mid cor(x,y) \mid = 1$ then there is a perfect linear relationship between x and y.

R commands:

```
>cor(x,y)
```

Regression and Correlation

The Simple Linear Regression Model

A **response variable** (dependent) measures the outcome of a study.

An **explanatory variable** (independent) attempts to explain the observed outcomes.

The most common graphical display used to study the association between two variables is called a **scatter plot**.

To graphically analyze the data, we can display the data on a 2-dimensional graph. We need to identify which variable is the dependent (y) variable and which is the independent (x) variable.

Least-Squares Regression

- The **least-squares regression line (LSRL)** of Y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
- The linear regression model is: $Y = \beta_0 + \beta_1 x + \varepsilon$
  - Y is dependent variable (response).
  - x is the independent variable (explanatory).
  - β_0 is the population intercept of the line.
  - β_1 is the population slope of the line.
  - ε is the error term which is assumed to have mean value 0. This is a random variable that incorporates all variation in the dependent variable due to factors other than x.
  - The variability: σ of the response y about this line. More precisely, σ is the standard deviation of the deviations of the errors, ϵ_i in the regression model.
- We will gather information from a sample so we will have the least squares estimates model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. $= b_0 + b_1 x$

Least-Squares Regression

Formulas:

$$\hat{Y} = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = cor(x, y) \cdot \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Correlation Coefficient

Correlation Coefficient, r:

The correlation measures the strength and direction of the linear relationship between two quantitative variables.

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

The correlation, r is an average of the products of the standardized values of x and y for a sample of n data pieces.

Correlation Coefficient

Facts about Correlation:

1. Positive r indicates positive association and negative r indicates negative association between variables.
2. r is always between -1 and 1
3. Correlation is strongly influenced by outliers.

The R command for the correlation is cor(x,y)

Coefficient of Determination

The **coefficient of determination** is a measure that allows us to determine how certain one can be in making predictions with the line of best fit. It measures the proportion of the variability in the dependent variable that is explained by the regression model through the independent variable.

- The coefficient of determination is obtained by squaring the value of the correlation coefficient.
- The symbol used is r^2
- Note that $0 \leq r^2 \leq 1$
- r^2 values close to 1 would imply that the model is explaining most of the variation in the dependent variable and may be a very useful model.
- r^2 values close to 0 would imply that the model is explaining little of the variation in the dependent variable and may not be a useful model.

A party is held where everyone is offered and eats exactly one meal option. Of those in attendance 50% prefer the first meal (tacos), 30% prefer the second (pizza), and everyone else prefers the third (hot dog). Of the people who ate tacos, 1% got sick. Of the people who ate pizza, 2% got sick. 5% of the people who ate a hot dog got sick.

1. Draw a tree diagram for this problem.

A party is held where everyone is offered and eats exactly one meal option. Of those in attendance 50% prefer the first meal (tacos), 30% prefer the second (pizza), and everyone else prefers the third (hot dog). Of the people who ate tacos, 1% got sick. Of the people who ate pizza, 2% got sick. 5% of the people who ate a hot dog got sick.

2. If a guest is randomly selected, what is the probability that the guest ate pizza and did not get sick?

A party is held where everyone is offered and eats exactly one meal option. Of those in attendance 50% prefer the first meal (tacos), 30% prefer the second (pizza), and everyone else prefers the third (hot dog). Of the people who ate tacos, 1% got sick. Of the people who ate pizza, 2% got sick. 5% of the people who ate a hot dog got sick.

3. Given that a guest got sick, what is the probability that the guest ate hot dogs?

Using R and R-Studio

1. Download R from https://cran.r-project.org/
2. Download R-Studio from https://www.rstudio.com/