

Chapter 5 Review

Section 5.2 - 5.5: Scatter Plots and Correlation

Definitions

Bivariate Data

To understand the relationship between two variables, data on both variables need to be collected. This type of data is called bivariate data. With bivariate data, two observations are recorded from some entity.

Univariate Data

Measurements of one variable.

Scatterplot (or Scatter Diagram)

In the case of bivariate data, a scatterplot is the traditional graphical method used to display the relationship between two variables. In a scatterplot, measurements are plotted in pairs with one variable plotted on each axis.

Oct 26-9:53 AM

Linear Relationship

A **linear relationship** is graphically described as a line. Mathematically, a line is a set of points that satisfy the functional relationship

$$y = mx + b,$$

where m is the **slope** of the line and b is the point where the function crosses the y -axis, which is called the **y -intercept**.

Oct 26-10:15 AM

Formulas

Correlation Coefficient

The correlation coefficient is an index number used to summarize the strength of a linear relationship.

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

Computational Formula for the Correlation Coefficient

The computational formula for the correlation coefficient is as follows.

$$r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

Oct 26-10:15 AM

Properties of the Correlation Coefficient

- The value of r is always between -1 and $+1$.
- A value of r near -1 or $+1$ means the data are tightly bundled around a line.
- A value of r near -1 or $+1$ means that we have a very strong negative relationship or positive relationship, respectively, such that predictions within the scope of the data are very reliable.
- Positive association is indicated by $r > 0$ and an upward sloping relationship.
- Negative association is indicated by $r < 0$ and a downward sloping relationship.
- A value of r near zero means there is no linear relationship between x and y .
- It does not matter whether you correlate y with x or x with y ; you will still get the same value for r .

Oct 26-10:16 AM

Section 5.6 - 5.9: Fitting a Linear Model

Definitions

Independent Variable (or Explanatory Variable)

The variable in a relation that determines the value or values of other variables, usually denoted by x . (Also referred to as the input variable.)

Dependent Variable (or Response Variable)

The dependent variable, y , is the variable whose value is explained or determined by the value of the independent variable, x . (Also referred to as the output variable.)

Oct 26-10:16 AM

The difference between the observed value of y and the predicted value of y .

Least Squares Line (or Regression Line)

The line that best fits the data is the least squares line that is obtained by minimizing the sum of squared errors (SSE).

Intercept Coefficient

The intercept coefficient, b_0 , is the average value of the dependent variable, y , when the independent variable, x , is equal to zero.

Slope Coefficient

The slope coefficient, b_1 , is the average change in the dependent variable, y , for a one-unit change in the independent variable, x .

Coefficient of Determination

The proportion of the variation explained by the model is called the coefficient of determination and is denoted as R^2 . The coefficient of determination is the square of the correlation coefficient.

Oct 26-10:17 AM

Formulas

Least Squares Line (or Regression Line)

$$y = b_0 + b_1x$$

Sum of Squared Errors

The sum of squared errors (SSE) is given by

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (b_0 + b_1x_i))^2.$$

Mean Squared Error

The variance of the error terms is also known as the mean squared error and is given by

$$s_e^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} = \frac{SSE}{n - 2}.$$

Oct 26-10:17 AM

Total Sum of Squares (TSS)

The total variation in y is given by the total sum of squares (TSS).

$$TSS = \sum (y_i - \bar{y})^2$$

Total Sum of Squares (TSS)

The sum of squares of regression (SSR) is the explained variation in y and is equal to the total variation minus the unexplained variation.

$$SSR = TSS - SSE$$

Coefficient of Determination

The coefficient of determination, R^2 , is given by

$$R^2 = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS}.$$

The coefficient of determination is a value between 0 and 1, inclusive. That is, $0 \leq R^2 \leq 1$.

The coefficient of determination can also be calculated using the following computational formula.

Oct 26-10:19 AM

6.1 - 6.4 Classical Probability

Objectives

- ★ To learn the basic vocabulary used in counting techniques
- ★ To solve problems using the classical probability
- ★ To solve problems using the experimental probability

Oct 26-10:21 AM

Definition

A **random experiment** is defined as any activity or phenomenon that meets the following conditions.

1. There is one distinct outcome for each trial of the experiment.
2. The outcome of the experiment is uncertain.
3. The set of all distinct outcomes of the experiment can be specified and is called the **sample space**, denoted by S .

Oct 26-10:44 AM

Definition

A **simple event**, or outcome, is any member of the sample space.

Definition

An **event** is a set of simple events or outcomes.

Oct 26-10:50 AM

Experiment Number 1: Toss a coin and observe the outcome. Have we met the three conditions of a random experiment?

1. There is one distinct outcome for each trial of the experiment.
2. The outcome of the experiment is uncertain.
3. The set of all distinct outcomes of the experiment can be specified and is called the **sample space**, denoted by S .

Oct 26-10:52 AM

Experiment Number 2: Toss a coin three times and observe the number of heads. Have we met the three conditions of a random experiment?

1. There is one distinct outcome for each trial of the experiment.
2. The outcome of the experiment is uncertain.
3. The set of all distinct outcomes of the experiment can be specified and is called the **sample space**, denoted by S .

What kind of events could we describe?

Oct 26-10:54 AM

Experiment Number 3: Select a student from a class of size 100 and observe his or her grade point average (GPA). Have we met the conditions of a random experiment?

1. There is one distinct outcome for each trial of the experiment.
2. The outcome of the experiment is uncertain.
3. The set of all distinct outcomes of the experiment can be specified and is called the **sample space**, denoted by S .

1. Although we can only select one student at a time, we are measuring GPA, which is a continuous random variable between zero and four.
2. The outcome will be unknown before selecting the student.
3. The sample space cannot be specified with certainty.

Oct 26-10:56 AM

Experiment Number 4: Assume we have a deck of playing cards consisting of 13 hearts, 13 clubs, 13 spades, and 13 diamonds. Draw a card from a well-shuffled deck and observe the suit of the card. Have we met the three conditions of a random experiment?

1. There is one distinct outcome for each trial of the experiment.
2. The outcome of the experiment is uncertain.
3. The set of all distinct outcomes of the experiment can be specified and is called the **sample space**, denoted by S .

Experiment Number 4: Assume we have a deck of playing cards consisting of 13 hearts, 13 clubs, 13 spades, and 13 diamonds. Draw a card from a well-shuffled deck and observe the suit of the card. Have we met the three conditions of a random experiment?

1. There will be only one outcome.
2. The suit will be unknown since the card will be drawn at random.
3. The sample space consists of the following set of outcomes:

$$S = \{\text{heart, club, spade, diamond}\}.$$

This experiment meets the conditions of a random experiment.

If the random experiment involves drawing a card and observing a spade or a club, then the **event** would be given by the set of outcomes $\{\text{spade, club}\}$.

Oct 26-10:59 AM

Experiment Number 5: Inspect a transistor to determine if it meets quality control standards. Have we met the three conditions of a random experiment?

1. There is one distinct outcome for each trial of the experiment.
2. The outcome of the experiment is uncertain.
3. The set of all distinct outcomes of the experiment can be specified and is called the **sample space**, denoted by S .

1. There will only be one outcome.
2. The outcome of the experiment will be unknown if the transistor is selected from a manufacturing process that occasionally produces defective parts.
3. The sample space consists of the set of outcomes, $S = \{\text{meets standards, does not meet standards}\}$.

This experiment meets the conditions of a random experiment.

Oct 26-11:01 AM

Interpreting Probability: Relative Frequency

Someone who wanted to determine the probability of getting a head on the toss of a coin could toss a coin a large number of times and observe the number of times that a head appeared.

The probability could be computed as the number of times a head was observed divided by the number of times the coin was flipped.

This is the **relative frequency** interpretation of probability.

Oct 27-10:01 AM

Formula: Relative Frequency

If an experiment is performed n times, under identical conditions, and the event A happens k times, the **relative frequency** of A is given by the following expression.















$$\text{Relative Frequency of } A = \frac{k}{n}$$

If the relative frequency converges as n increases, then the relative frequency is said to be the **probability** of A .

Oct 27-10:03 AM

Let's flip a coin 4242 times and observe the relative frequency of a head during those tosses.











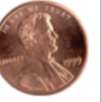


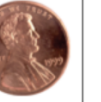
Table 6.1 – Relative Frequency of Heads for 42 Flips

Coin							
Flip Number	1	2	3	4	5	6	7
Relative Frequency	1.00	0.5	0.6667	0.75	0.6	0.5	0.4286
Coin							
Flip Number	8	9	10	11	12	13	14
Relative Frequency	0.375	0.4444	0.4	0.3636	0.4167	0.3846	0.3571

Oct 27-10:04 AM

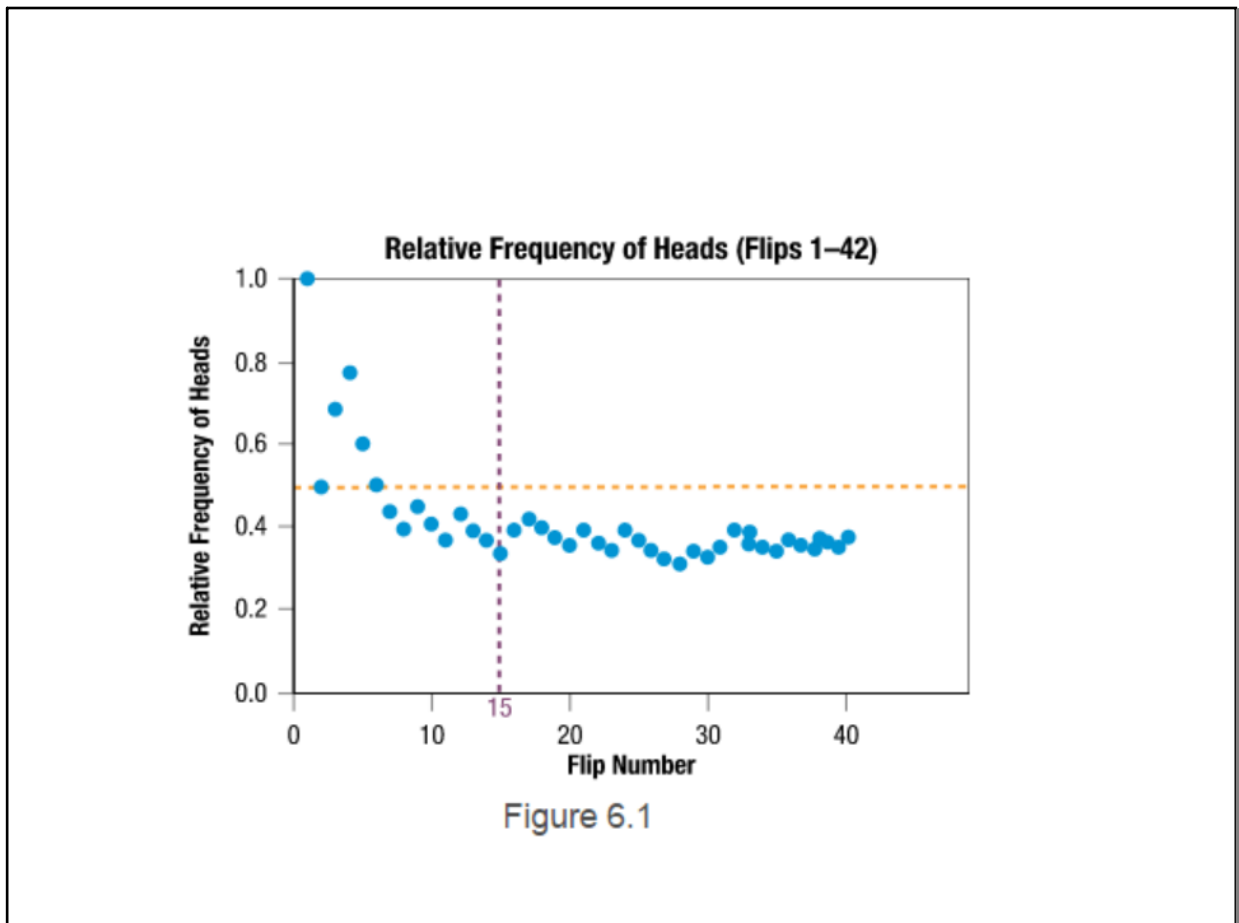
Coin							
Flip Number	15	16	17	18	19	20	21
Relative Frequency	0.3333	0.375	0.4118	0.3889	0.3684	0.35	0.3810
Coin							
Flip Number	22	23	24	25	26	27	28
Relative Frequency	0.3636	0.3478	0.375	0.36	0.3462	0.3333	0.3214

Oct 27-10:06 AM

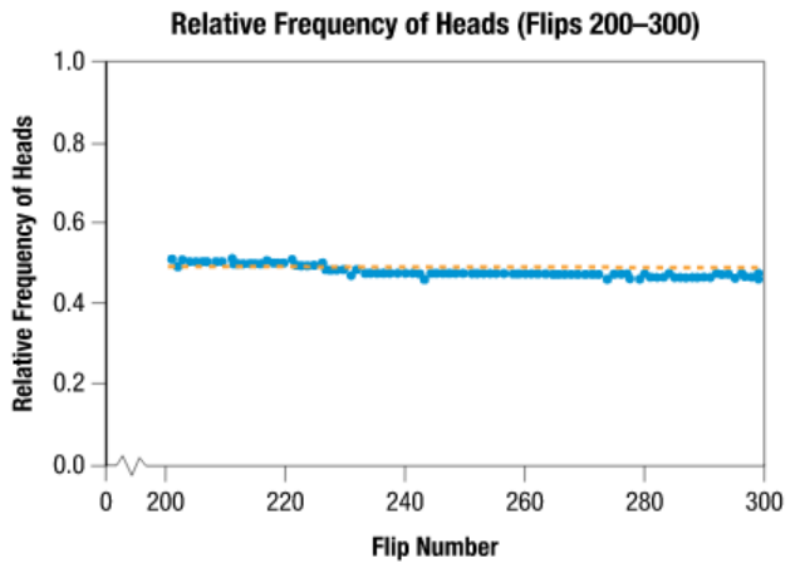
Coin							
Flip Number	29	30	31	32	33	34	35
Relative Frequency	0.3448	0.3333	0.3548	0.375	0.3636	0.3529	0.3429
Coin							
Flip Number	36	37	38	39	40	41	42
Relative Frequency	0.3611	0.3514	0.3421	0.3590	0.35	0.3415	0.3571

In Figure 6.1, you can see that the proportion of heads is very unstable during the first 15 flips. The proportion of heads begins to stabilize at around flip 20, although there is still some fluctuation in its value. Looking at the first 42 flips makes you wonder whether this is a "fair" coin, since heads is occurring only $0.3571(100) = 35.71\%$ of the time. In Figure 6.2, which starts at about 200 flips, the percentage of heads becomes very stable. By flip number 296, there are 141 heads and 155 tails which equate to (approximately) a 0.4764 probability (or relative frequency) of heads. Although this is slightly less than expected, such a percentage is reasonable considering the randomness of the coin toss.

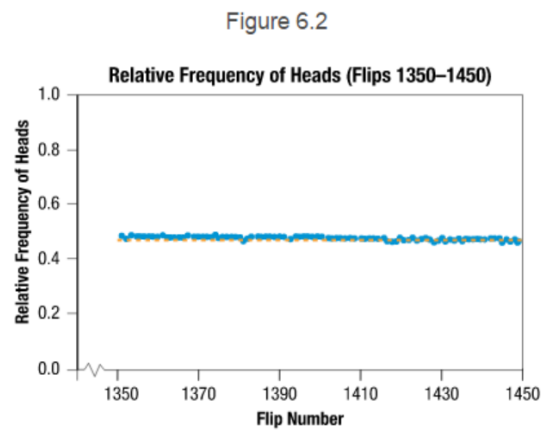
Oct 27-10:05 AM



Oct 27-10:10 AM



Oct 27-10:11 AM



As can be seen in Figure 6.2 and Figure 6.3, the percentage of heads converges on some point which is close to 50%. Figure 6.3 starts at 1350 flips. As you can see, the percentage of heads remains very stable. At flip number 1450 there are 718 heads and 732 tails which (approximately) equate to a 0.4952 probability (or relative frequency) of heads.

Oct 27-10:11 AM

What Is the Relative Frequency of a Head on Our Coin?

Our best available guess is 0.4952, since it is the observed relative frequency of heads using all 1450 tosses.

$$\text{Relative Frequency of } A = \frac{k}{n} = \frac{718}{1450} \approx 0.4952$$

How good is this guess for the relative frequency of heads? Since the coin has been tossed a large number of times and the observed frequency is very stable, the guess should be very good.

Oct 27-10:11 AM

Will the Observed Relative Frequency Ever Reach 0.5?

No mathematical or physical law requires the observed relative frequency to ever reach some predetermined level. But if the probability of observing a head was really 0.5 the observed relative frequency should closely approach this value after a large number of flips.

Oct 27-10:12 AM

A Summary

The experiment: Toss a coin and observe which side of the coin appears on top.

Duration of the experiment: Toss the coin 1450 times. $n = 1450$

Observe the event "getting a head": The event was observed 718 times. $k = 718$

Relative frequency of the event "getting a head": $\frac{k}{n} = \frac{718}{1450} \approx 0.4952$

The relative frequency of a head seems to converge to the expected relative frequency of 0.5. This kind of convergence is sometimes called statistical regularity. Although the outcomes of the experiment may vary, in the long run the relative frequency of an outcome tends to some value, its probability.

Oct 27-10:14 AM

Problems with the Relative Frequency Idea

- The problem with the relative frequency approach to defining probability is that probability only exists for events that can be repeated under the same conditions.
 - > Coin, dice, and card experiments can easily be repeated.
- However, because of the strict requirements of identical and repeatable experiments, many events in which it would be desirable to have relative frequency probabilities do not satisfy the requirement of repetition.
 - > the next launch of the rocket will be successful,
 - > whether you will make an A in your statistics course
 - are examples of experiments that are not repeatable under the exact same conditions.
 - they are not appropriate for the application of the relative frequency idea.
- This perspective greatly limits the application of the relative frequency interpretation of probability. Despite its limitations, the relative frequency approach is a widely held interpretation of probability.

Oct 27-10:14 AM

Example 6.1

Suppose we perform an experiment drawing three numbers, without replacement, from an urn containing 6464 numbers. Have we met the conditions of a random experiment?

Solution

No! This does not meet the conditions of a random experiment because it cannot be repeated under the same conditions. That is, as each number is selected, it is not returned to the urn prior to the subsequent selection, thus decreasing the amount of numbers remaining in the urn. For this reason, the experiment is different for each number drawn, and cannot be considered a random experiment.

Oct 27-10:20 AM

Interpreting Probability: Subjective Approach

- The subjective viewpoint regards the probability of an event as a measure of the degree of belief that the event has occurred or will occur.
 - > Someone's degree of belief in some event will depend on his or her life experiences.
 - > Different life experiences produce different degrees of belief. Hence, the subjective approach must allow for differences in the degree of belief among reasonable people examining the same evidence.
- One of the significant advantages of this view is the ability to discuss the probability of events that cannot be repeated.
 - > Thus, a subjectivist would be willing to assign the probability of making an A in your statistics course.
- Someone who adopts the subjective view could use the frequency interpretation to influence the determination of a subjective probability.
 - > For example, suppose that a coin had been tossed 20,000 times and had come up heads 63% of the time. It would certainly be reasonable for a subjectivist to use this information in the formulation of a statement of probability about the outcome of the next toss.

Oct 27-10:20 AM

Criticism of the Subjective View

If science is defined as finding out what is probably true, there should be a probability criterion on which all reasonable persons could agree. But if probability is subjective, how can it be used as a universally accepted criterion? Two reasonable persons might examine the same data and reach different conclusions about their degree of belief about some proposition.

Oct 27-10:28 AM

Interpreting Probability: Classical Approach

Classical probability can be measured as a simple proportion: the number of outcomes that compose the event divided by the number of outcomes in the sample space, when it can be assumed that all of the outcomes are equally likely.

Formula: Classical Probability

Using the classical approach to probability, the probability of an event A , denoted $P(A)$, is given by

$$P(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes in the sample space}}.$$

Oct 27-10:29 AM

Example 6.2

In experiment number 22, a coin was tossed three times and the number of heads was observed. The sample space consists of 8 outcomes {TTT, TTH, THT, THH, HTT, HTH, HHT, HHH}. Let A be the event of getting at least one head. What is $P(A)$?

Solution

Since the event A consists of 7 outcomes, {TTH, THT, THH, HTT, HTH, HHT, HHH}, and there are 88 equally likely outcomes in the sample space,

$$P(A)=7/8=0.875.$$

Oct 27-10:30 AM

Experiment Number 4: Assume we have a deck of playing cards consisting of 13 hearts, 13 clubs, 13 spades, and 13 diamonds. Draw a card from a well-shuffled deck and observe the suit of the card. Have we met the three conditions of a random experiment?

Example 6.3

In experiment number 4, let A be the event of drawing a heart. What is $P(A)$?

Solution

Since there are 13 outcomes in A (i.e., 13 hearts in a deck of cards) and 52 outcomes in the sample space (i.e., 52 cards in the deck) the probability of event A is as follows.

$$P(A)= 13/52 = 1/4 = 0.25$$

Oct 27-8:26 PM

If the sample space is composed of equally likely outcomes, then once the set of outcomes is determined, computing a probability is simply a matter of counting the members in each set and dividing by the total number of outcomes.

It is very important to remember that the classical approach rests on the assumption of equally likely outcomes. If the assumption is not reasonable, some other method of determining the probability must be used.

Oct 27-8:33 PM

Oct 27-8:34 PM