MAT 19.001 & 19.002

Introduction to Statistics Workbook

Arvind Borde

This workbook contains work done by: _____

Signature: _____

Introduction to Statistics

(1) Why are you here? _____

(2) What's ? _____

a) _____

b) _____

c) _____ and

d) _____

_____

_____

1

The "information" that is collected, organized, etc., in statistics is called _____

can be _____

(3) Write down three types of data about yourselves that are numerical, and three that are not:

Num: 1. _____ Non: 1. _____

2. _____ 2. _____

3. _____ 3. _____

2

Misuses of Statistics: Examples

A] A talk show host uses the number of people who call in with an opinion to measure how widespread the opinion is.

(4) Why is this misleading?

_____

(a) _____

(b) _____

3

B] A study says that gray cars are involved in fewer accidents than cars of other colors. The media announces that gray cars are safer.

(5) Correct conclusion? _____

_____

_____

4

C] Studies show that breast-fed children have higher IQs than those who were not breast-fed. They say on TV that breast-feeding increases a child's IQ

(6) Correct conclusion? _____

_____

_____

5

D] An athlete with a very successful later career offers a defense against the charge that steroids were involved by providing examples of other athletes with similarly successful later careers.

(7) Is this statistically valid?

_____

_____

6

ADDITIONAL NOTES

Examples of Statistical Statements

New York Times, January 8, 2016

Seattle Seahawks' Defense Stands Out, Even Across Eras

"Over the last four years, Seattle has allowed 15.73 points a game, relative to the league average of 22.89. The standard deviation in points allowed by the league's 32 teams over this period has been 2.62 points per game. This means Seattle, which has been over 7 points a game better than average, has been 2.73 standard deviations better than average, a statistic known as a Z-score."

7

New York Times, January 15, 2016

Signs of a Retail Rebound as Consumers Take On More Debt

"Statistics from the Federal Reserve Bank of New York show that auto loans increased by more than 12 percent from the third quarter of 2014 to the third quarter of 2015. That left the total of auto credit at $1.05 trillion"

8

New York Times, September 5, 2015

The Collateral Victims of Criminal Justice

"Between 1991 and 2007, the percentage of children with mothers in prison more than doubled, according to federal data – and that does not count the many more mothers who spent time in jail. It doesn't take statistics to grasp how damaging separation can be, but even so, the data shows these children have more depression, aggression, delinquency, absenteeism, asthma and migraines."

9

New York Times, September 5, 2015

Awash in Data, Thirsting for Truth

". . . in a numbers-soaked era. . . questions arise: How important is data to reporting? And does it get readers closer to the truth or obscure it?

"I'm thinking about these questions because of two recent pieces in The New York Times,. . .

"The first was a major exposé of brutal working conditions at Amazon, based on six months of reporting . . . some, including Amazon brass, contested it . . .

10

"The second piece was a magazine cover story, built on data, arguing that a feared digital-age 'creative apocalypse' never happened. . . . Many readers disputed the article's conclusions and charged that the numbers it relied upon had been 'cherry-picked' to make a flawed case. . . ."

– Margaret Sullivan, The Public Editor

11

Understanding statistical tests in the medical literature

"When writing or reading articles, one should be aware whether the statistical tests performed were appropriate for the type of data collected and used, thereby avoiding misleading conclusions. The goal of all statistical tests is to determine whether two (or more) variables are associated with one another or independent from each other. . . "

"One of the first things to keep in mind is the type of data and outcomes the author wants to measure and correlate. . .

12

ADDITIONAL NOTES

"The second important thing to keep in mind is how the results are distributed. Do they follow a 'bell curve'. . . , similar to biological phenomena and exam grading techniques, or do the results tend to cluster resulting in a skewed distribution?"

– National Institute of Health, September 2008

13

Health & Health Profession Statistics

"Nursing is the nation's largest health care profession, with more than 3.1 million registered nurses nationwide. Of all licensed RNs, 2.6 million or 84.8% are employed in nursing.

"Registered Nurses comprise one of the largest segments of the U.S. workforce as a whole and are among the highest paying large occupations. Nearly 58% of RNs worked in general medical and surgical hospitals, where RN salaries averaged $66,700 per year."

– American Association of Colleges of Nursing
September 2015

14

New York Times, August 19, 2017

The Stock Market Has Been Magical. It Can't Last.

"The bull market that started in March 2009 seems to have gone on forever, but that's only a small part of the story. What is astonishing is the way that stocks have risen in 2017. . . .

"Several statistical measurements demonstrate how unusual this market environment has been."

– Jeff Sommer

15

Here are some of the concepts mentioned in these examples that we'll study in this course:

- _____
- _____
- _____
  _____
- _____

16

Some Statistical Questions

(8) Here are some real baseball batting averages:

	1995	1996
Derek Jeter	.250	.314
David Justice	.253	.321

Who, would you guess, had the higher average over the two years? _____

17

(9) How is this possible?

Here are the details of the years:

	1995		1996	
Jeter	12/48	.250	183/582	.314
Justice	104/411	.253	45/140	.321

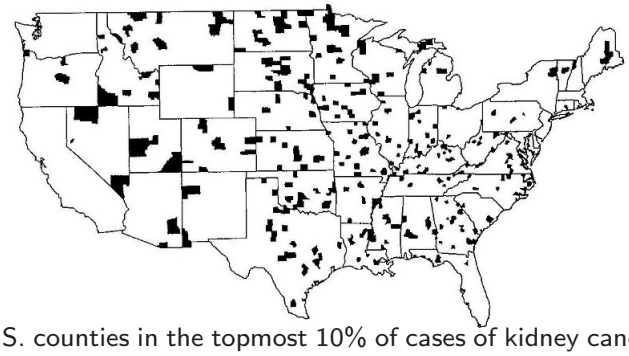
Now combine the two years:

Jeter: _____

18 Justice: _____

ADDITIONAL NOTES

(10) What pattern do you see here?



U.S. counties in the topmost 10% of cases of kidney cancer deaths among males, adjusted for age, 1980–89.

19

(11) _____

_____

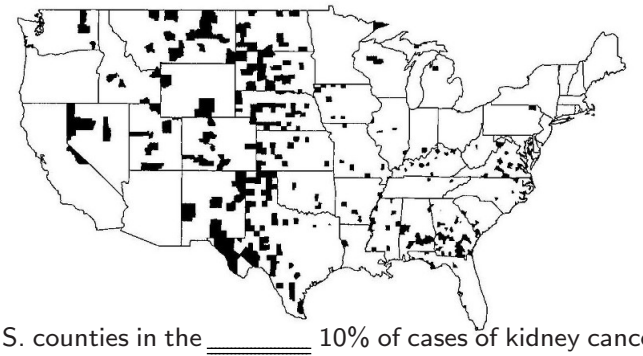
(12) _____

_____

_____

20

(13) What pattern do you see here?



U.S. counties in the _____ 10% of cases of kidney cancer deaths among males, adjusted for age, 1980–89.

21

_____

_____

(14) What gives?

_____

(15) _____

22

Some U.S. Counties, with population

1. LA, CA: _____
2. Cook IL: _____
5. SD, CA: _____
8. Kings, NY: _____
10. Queens, NY: _____

...

~3149. Loving, TX: _____

~3150. Kalawao, HI: _____

23

If, by pure chance, a county, with a population of 100 has one death in this category in the decade, as opposed to none, its rate will change from zero (low), to _____ (high).

A random fluctuation of a few deaths in LA, on the other hand, will not change its stats as much.

_____

_____

24

ADDITIONAL NOTES

(16) In WWII statistician Abraham Wald analyzed bullet damage on planes that returned from missions. He recommended that systematically damaged portions of aircraft be left unchanged, but that the *undamaged* parts be re-enforced. Why?

25

(17) I notice that when I get to a bus stop, I am more likely to see a bus going east than west. How might this be possible?

26

Some Statistical Terms

Definition

The entire group under study is the _____. An _____ is a person or object that is a member of the population. A _____ is a subset of the population.

27

Example A: I want to study the weekly spending habits of Post students. I survey this class.

(18) What is the population in this study?

(19) Who are the individuals?

(20) What is my sample?

28

Example B: I want to figure out who will win an election. I select 100 landline numbers at random and call them to ask who they plan to vote for.

(21) What is the population in this study?

(22) Who are the individuals?

(23) What is my sample? _____

29

Example C: I want to figure out who will win an election. I select 100 mobile numbers at random and call them to ask who they plan to vote for.

(24) What is the population in this study?

(25) Who are the individuals?

(26) What is my sample? _____

30

ADDITIONAL NOTES

Definition

_____ consist of organizing and summarizing data.

Typical ways to organize, summarize and display data are numerical summaries, tables, and graphs.

31

Definition

_____ uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

You want to describe and display information you've collected, and also to *infer* something from it. The process is never completely accurate, and you need a way to measure your confidence in the inference.

32

Definition

A _____ is a numerical summary of a sample.

If my survey of you (see slide 28) shows that you spend $102 per week on average, that number is a statistic.

33

Definition

A _____ is a numerical summary of a full population.

If *all* students at Post were asked, and the average weekly spending was found to be $93, that number is a parameter.

One goal of inferential statistics is to estimate parameters from statistics.

34

Variables in statistics are the characteristics of individuals within the population. They can be classified into two groups: _____ or _____

In my weekly spending example, students are the individuals. The amount of money they spend is the variable (varies from individual to individual).

35

Definition

_____ are descriptive characteristics.

These variables are often non-numerical, but don't have to be.

Example: _____

36

ADDITIONAL NOTES

Definition

_____ are numerical measures of individuals that can be manipulated mathematically (added, subtracted, etc.).

Example: _____

37

(27) If, instead of names, I identify you by your ID number, does that switch your identifier variable from qualitative to quantitative?

_____

_____

_____

38

(28) Identify these as quantitative or qualitative:

a) Your nationality. _____

b) Your temperature. _____

c) Your weight. _____

d) Your zip code. _____

39

Definition

A _____ is a quantitative variable that has either a finite number of possible values or a "countable" number of possible values.

The term countable means that the values result from counting, such as 0, 1, 2, 3, and so on. A discrete variable cannot take on every possible value between any two possible values.

40

Definition

A _____ is a quantitative variable that has an infinite number of possible values that are not countable.

A continuous variable may take on every possible value between any two values.

41

(29) One of the two quantitative variables we used as examples, height and weekly spending, is discrete and one is continuous. Which is which?

_____

_____

42

ADDITIONAL NOTES

These distinctions are not always clearcut.

(30) What type of variable is "handedness"?

Fill out the table on the next slide, as follows:

1 in R or L column if you *usually* use that hand.
2 if you *always* use that hand.
1 in *each* column if you often switch.

43

Handedness scoresheet

Task	Left	Right
Writing		
Drawing		
Throwing		
Scissors		
Toothbrush		
Cutting (knife)		
Spoon		
Hand on lid when opening jar		

44

R: Your total Right score:

L: Your total Left score:

Your handedness score: $\frac{\mathbf{R}-\mathbf{L}}{\mathbf{R}+\mathbf{L}} =$

45

Definition

An _____ measures the value of variables without attempting to influence them.

In an observational study, the researcher observes the behavior of the individuals without trying to influence the outcome.

46

Definition

A _____ changes the value of a variable, and then records the value of the response.

In a designed experiment, the researcher changes the situation in order to see if it influences the outcome.

47

A study of the possible relationship between cell phone use and brain tumors that looks at people with and without tumors, then studies if their cell phone usage differs is observational.

A study that bombards rats with cell phone frequency radiation to see if they develop tumors is a designed experiment.

48

ADDITIONAL NOTES

It may seem that an observational study does less harm. But it may also cause confusion.

One observational study has shown that seniors who get 'flu shots are far less likely to get the 'flu, or associated diseases such as pneumonia.

(31) Does this prove that the 'flu shot makes people healthier, or are there variables that have not been taken into account? _____

49

Definition

A _____ is a variable that was not considered in a study, but that affects the value of the response in the study.

In our 'flu example, those who got the shots were often more mobile than ones who did not, so mobility was a lurking variable. (Must not confuse _____ and _____.)

50

Sampling

How do we pick samples?

Definition

_____ is the process of using chance to select individuals from a population to be included in the sample.

An opposing approach is _____

51

(32) In the weekly spending example on slide 28, was my sampling random or convenience?

_____

(33) How might I randomize the sample?

_____

52

Definition

A sample is a _____ if each
(i) _____
(ii) _____
_____ and
(iii) _____
_____

53

For example, I have 27 students in class. I decide to give out 10 A grades at random. I write on pieces of paper all possible lists of 10 students, then pick one list out of a hat.

In practice, one picks one student at random, then the next, etc., but this is not *exactly* the same.
(34) Why not? _____

54

ADDITIONAL NOTES

We'll concentrate on simple random sampling in this course, but (for the record), here's another kind of random sampling: _____

_____

It's used in situations where the population has _____, or _____, and it's important to ensure that each is proportionately represented in the sample.

55

(35) Can you give an example where such situations might occur? _____

_____

(36) What would you do in situations where strata are important?

_____

56 _____

(37) If 5% of the population has an income above $\$1{,}000{,}000$, 60% between that and $\$100{,}000$, and the rest under $\$100{,}000$, what would a good stratified sample of 200 contain?

_____

_____

_____

57

Even with "random" sampling biases can occur. In our polling examples (slides 29 and 30) picking samples based on landlines vs. mobile, each introduce biases.

(38) Who is more likely to still have a landline?

_____

(39) Who is more likely to have a mobile?

58 _____

Estimation

(40) How many school busses in the U.S.?

59

ADDITIONAL NOTES

Math Review

Whole Numbers

(1) $2 + 2 = ?$

(2) $1 + 2 + 3 = ?$

(3) $12 - 3 + 4 = ?$

1

(4) $1 + 2 + 3 \ldots 100 = ?$

(5) How many such pairs are there? _____

(6) So, what is the sum? _____

2

Statistics often requires you to add a lot of numbers. Using "+" over and over again can get old.

Welcome Sigma,

$\sum$

our new bff.

3

What does

$$\sum_{i=1}^{100} i$$

mean?

It means

4

What does

(7) $\sum_{i=1}^{100} 3i$ mean?

(8) $\sum_{i=1}^{100} i^2$ mean?

5

What does

(9) $\sum_{i=1}^{50} \frac{i}{i+1}$ mean?

(10) $\sum_{i=1}^{n} \frac{1}{i}$ mean?

6

ADDITIONAL NOTES

(**11**) What does

$$\sum_{i=5}^{n} \frac{1}{i}$$

mean?

7

OK. Back to "smaller" calculations:

(**12**) $2 - 2 = ?$ _____

(**13**) $2 + (-2) = ?$ _____

(**14**) $2 - (-2) = ?$ _____

8

Fractions

(**15**) $\frac{1}{2} + \frac{2}{3} = ?$

9

(**16**) $\frac{1}{2} - \frac{2}{3} = ?$

10

(**17**) $\frac{3}{2} + \frac{3}{4} = ?$

11

(**18**) $\frac{3}{2} \times \frac{3}{4} = ?$

12

ADDITIONAL NOTES

(19) $\frac{3}{2} \times \frac{4}{5} = ?$

OR

13

(20) $\frac{3}{2} \div \frac{4}{5} = ?$

14

Division by any number is the same as multiplication by its reciprocal. Why?

Consider division by 2, e.g.: Dividing something by 2 is the same as halving it. Therefore,

$$N \div 2 = N \times \frac{1}{2}$$

This is true for division by any number.

15

Order of Operations

You will be expected to know the addition, subtraction, multiplication and division of basic fractions, as well as simplifying the answer.

You will also be expected to know in which order you carry out these operations.

16

(21) $\frac{1}{2} + \frac{2}{3} \times \frac{2}{5} = ?$

17

(22) $\left(\frac{1}{2} + \frac{2}{3}\right) \times \frac{2}{5} = ?$

18

ADDITIONAL NOTES

Equations

(23) Solve $3x - 15 = 0$.

19

(24) Solve $4x - 2 = 13 - 2x$.

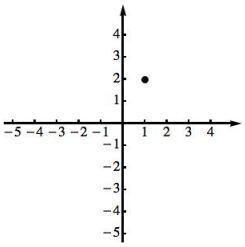
20

(25) Solve $-5x + 1 = 0$

21

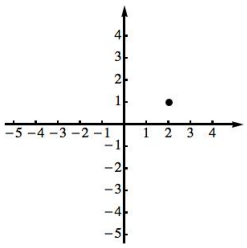
Coordinates

(26) What are the coordinates of



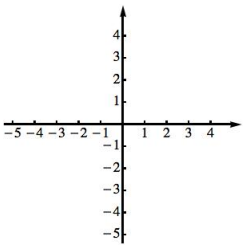
22

(27) What are the coordinates of



23

(28) Plot the points $(-1, 3)$ and $(2, -3)$.



24

ADDITIONAL NOTES

Functions and Graphs

Definition

A _____ expresses a relationship between two (or more) variables.

Examples:

25

We distinguish between the variable that you can change freely, _____, and the variable(s) that depend(s) on it, _____ variable(s). If two variables are separated, one on the left the other on the right, it's usually assumed that the right-hand variable is independent. For example, in $y = x^2 + 1$, _____

_____

26

In more complicated situations, for example

$$2^u + v^3 = u - \sqrt{v},$$

you'll be told which variable you should think of as independent.

The set of _____ _____ is called the _____ of the function, and the set of _____ _____ the _____.

27

For example, the domain of $y = 1/x^2$ is the set of all values of x _____, and the range is all values of y _____.

(**29**) Why?

Domain: _____

Range: _____

We can write the domain as $(-\infty, \infty) - \{0\}$, and the range as $(0, \infty)$.

28

(**30**) What are the domain and range of

$$y = \frac{1}{x-1}?$$

_____

_____

29

The dependent variable, say y, is called _____ _____.

We write this expression as _____, read as "_____." So,

$$y = x^2 + 1 \quad \text{and} \quad f(x) = x^2 + 1$$

have the same content. When we write $y = f(x)$ we're saying that y is a function of x.

30

ADDITIONAL NOTES

Graphs of Linear Functions

To plot the linear function $2x - 1$:

1) Write the equation $y = 2x - 1$.

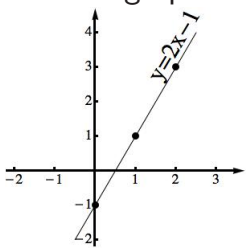
2) Make a table:

	x	$2x-1$	y	
Chosen	0	$2(0)-1$	-1	Calculated
	1	$2(1)-1$	1	
	2	$2(2)-1$	3	

3) Plot the (x, y) values and connect them.

31

Here's the graph:



Note:
1) y-intercept is -1. (Why?)
2) x-intercept is the solution of $0 = 2x - 1$: $x = 1/2$.
3) Graph points upward.

32

(**31**) Plot $2x + 1$ and get the intercepts.

33

(**32**) Plot $-2x + 3$ and get the intercepts.

34

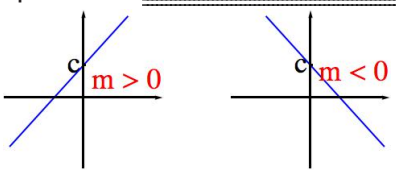
The General Linear Function: _____

1) Graph points upward (reading from left to right) when _____, and downward when _____.

m is called the _____.

2) y-intercept is _____.

3) x-intercept is the _____.



35

Slope

_____

_____

_____

_____

_____

_____

36

ADDITIONAL NOTES

Square Roots

(33) What is $\sqrt{4}$? _____.

(34) How do you know that ?

_____

(35) is there another number whose square is 4?

_____

_____

37 _____

(36) What's the difference between "positive" and "non-negative"?

_____

(37) $\sqrt{(2)^2} =$

(38) $\sqrt{(-2)^2} =$

Taking squares, then square roots, is a technique in statistics to eliminate negative numbers.

38

(39) Assuming a and b are positive, is it correct to say that

$$\sqrt{a^2 + b^2} = \sqrt{a^2} + \sqrt{b^2} = a + b?$$

39

Percentages

(40) Express these as percentages:

- $1/2$ _____
- $1/3$ _____
- $1/8$ _____
- 1 _____
- 2 _____

40

(41) Express these as factions and as decimals:

- 40% _____
- 55% _____
- 37% _____
- 250% _____
- 75% _____

41

(42) You get $8/20$ on test 1, and $16/20$ on test 2, by what percentage has your score *improved*?

_____

(43) What is your new score as a percentage of your old score? _____

(44) If the stock market goes down by 20% today, by what percentage must it go up tomorrow for you to recover your losses? _____

42

ADDITIONAL NOTES

Organizing Qualitative Data

First you collect data.

Example: injuries at a clinic.

The raw data needs to be _____ in order to be able to do statistics with it.

Back	Back	Hand
Wrist	Back	Groin
Elbow	Back	Back
Back	Shoulder	Shoulder
Hip	Knee	Hip
Neck	Knee	Knee
Shoulder	Shoulder	Back
Back	Back	Back
Knee	Knee	Back
Hand	Back	Wrist

Source: Krystal Catton, student at Joliet Junior College

43

A common organizational tool in statistics is a _____

Body Part	Tally	Frequency
Back		
Wrist		
Elbow		
Hip		
Shoulder		
Knee		
Hand		
Groin		
Neck		

44

(45) What might we mean by "frequency"?

_____

(46) For later, how many injuries are there in all in the clinic sample?

_____

(47) How many groin injuries?

_____

45

(48) Suppose a survey of another clinic shows 25 groin injuries there. Are groin injuries there more prevalent than at the first clinic?

_____

_____

_____

46

(49) Why "less prevalent" if sample size is 3,000?

_____

_____

In many cases the _____ is more useful. It is defined as

47

(50) What might be the advantages of relative frequencies?

_____

_____

_____

(51) Calculate the relative frequencies for the data from the clinic.

48

ADDITIONAL NOTES

Relative frequency table

Body Part	Frequency	Relative Frequency
Back	12	$\frac{12}{30} =$
Wrist	2	
Elbow	1	
Hip	2	
Shoulder	4	
Knee	5	
Hand	2	
Groin	1	
Neck	1	
Total	**30**	

49

Presenting Data Visually

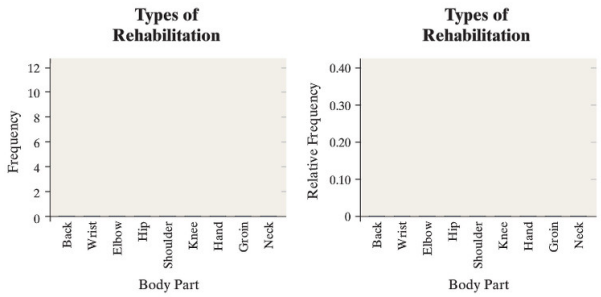
Tables are fine, but nothing beats a picture.

Graphs, bar graphs, pie charts, etc., are ways in which data can be presented pictorially.

50

Bar graphs

Frequencies and relative frequencies can be shown on bar graphs. Draw them for our injury example



51

(52) Do the freq. and rel. freq. bar graphs have similar profiles? _____

(53) Does that shock you? _____

(54) Cause even mild surprise? _____

(55) Why or why not?

_____

52

Side-by-side bar graphs

Consider these data:

Educational Attainment	1990	2009
Not a high school graduate	39,344	26,414
High school diploma	47,643	61,626
Some college, no degree	29,780	33,832
Associate's degree	9,792	17,838
Bachelor's degree	20,833	37,635
Graduate or professional degree	11,478	20,938
Totals	**158,870**	**198,283**

Source: U.S. Census Bureau

53

We want to compare the two years shown.

(56) In which sample were there more students with some college, but no degree? _____

(57) Can we conclude that there was a greater prevalence of such students then? _____

(58) Why/why not? _____

(59) What should we use? _____

54

ADDITIONAL NOTES

OK, since you think it's the right thing to do, make a relative frequency table:

Educational Attainment	1990	2009
Not a high school graduate		
High school diploma		
Some college, no degree		
Associate's degree		
Bachelor's degree		
Graduate or professional degree		

55

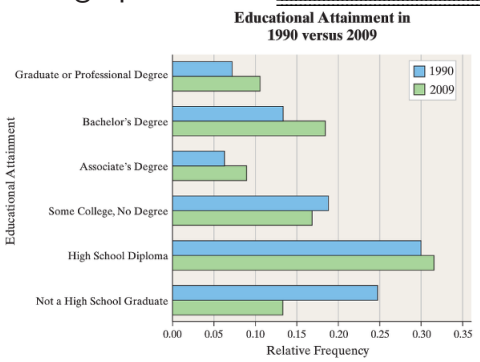
These numbers can be compared.

(60) When was there a greater prevalence of students with some college, but no degree? _____

The information is in the RF table, but we have to hunt for it. A better presentation is visual, such as by a side-by-side bar graph, where the two years are presented _____

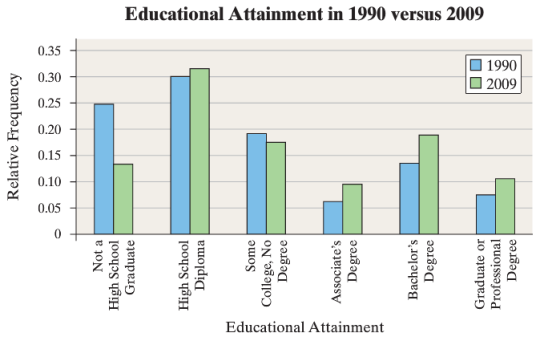
56

Side-by-side graphs can be _____



57

or _____



depending on preference, easy readability, etc.

58

Pie charts

These are another way to visually present one set of data (not two).

Each of the data categories is shown as a sector (a wedge or a slice) of the pie, where the angle in each sector (in degrees) is

_____

59

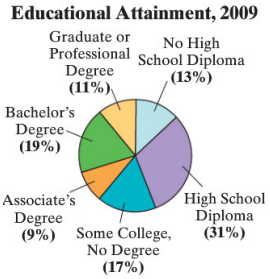
Calculate the degrees in each slice/sector (rounding off to whole numbers):

Educational Attainment	Frequency	Relative Frequency	Degree Measure of Each Sector
Not a high school graduate	26,414	0.1332	
High school diploma	61,626	0.3108	
Some college, no degree	33,832	0.1706	
Associate's degree	17,838	0.0900	
Bachelor's degree	37,635	0.1898	
Graduate or professional degree	20,938	0.1056	

60

ADDITIONAL NOTES

Then plot:



(61) How were the percentages obtained?

(Look at RFs.) _____

61

Bar graphs vs Pie charts: The smackdown

(62) Which is better and where?

Pie charts: _____

_____

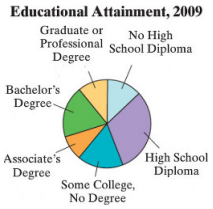
_____

_____

_____

62

For example, it is not easy to immediately see without numerical labels in this pie chart



how the percentages of associate's degrees, graduate degrees, and no high school diplomas compare.

63

Bar graphs: _____

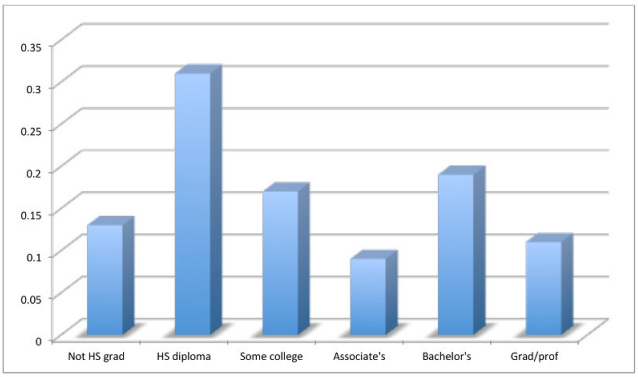
_____

_____

For example, for the "big picture" of educational attainment in 2009, a pie chart is a good visual summary. But, to compare bachelor's degrees to high school diplomas, a bar graph is better.

64

Like so:



65

Organizing Quantitative Data

Discrete Data

First you collect data:

Number of Arrivals at Wendy's							
7	6	6	6	4	6	2	6
5	6	6	11	4	5	7	6
2	7	1	2	4	8	2	6
6	5	5	3	7	5	4	6
2	2	9	7	5	9	8	5

(Number of customers who visit during 40 randomly selected 15-minute intervals at Wendy's.)

66

ADDITIONAL NOTES

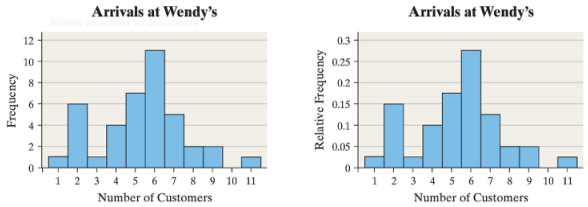
Then make frequency and rel. frequency tables:

Number of Customers	Tally	Frequency	Relative Frequency
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			

67

Presenting your results graphically

Common tool: histograms



68

Histo-what?

The origin of the term is unclear. It seems to first appear in a paper by William Pearce from 1894, and it *may* be an abbreviation for "historical digram."

69

X. *Contributions to the Mathematical Theory of Evolution.*—II. *Skew Variation in Homogeneous Material.*

By Karl Pearson, *University College, London.*

Communicated by Professor Henrici, *F.R.S.*

Received December 19, 1894,—Read January 24, 1895.

Plates 7–16.

The results are depicted, Plate 14, fig. 13.

It will be observed that so far as the observations can be plotted to the theoretical curve, it leaves little to be desired. The histogram* shows, however, the amount of deviation at the extremes of the curve.

* Introduced by the writer in his lectures on statistics as a term for a common form of graphical representation, *i.e.*, by columns marking as areas the frequency corresponding to the range of their base.

70

Bar graphs vs Histograms

Look similar, but:

A] Bar graphs plot numerical values as heights against categorical (or qualitative) variables, often several of them (different levels of educational attainment, for example). Histograms plot values against different values of a single quantitive variable (arrival times at Wendy's, for example).

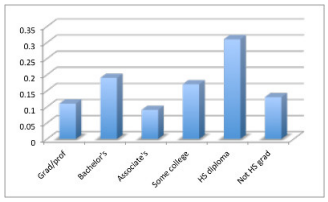
71

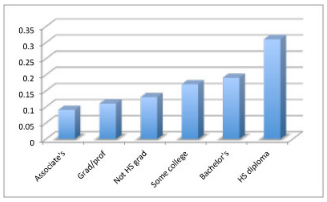
B] Since the categorical variables in a bar graph are distinct, they are usually drawn with gaps between the bars. Histograms do not have gaps because one set of values of the quantitative variable connect to the next set.

72

ADDITIONAL NOTES

C] You can rearrange the order of the bars on a bar graph for convenience. You can't on a histogram (see slide 68).





73

Continuous Data

Age	Number (in thousands)
25–34	12,967
35–44	13,904
45–54	13,005
55–64	10,357
65–74	4,584

Source: Current Population Survey, 2008

Bachelor's degrees

74

The _____ is the _____

_____

The _____ is the _____

The _____ is the _____

_____

75

(63) In the table on slide 74, what is the class width? _____

(64) Is there a potential problem with how the classes are defined? (Can they accommodate all ages?)

_____

_____

76

Here's the rate of return on your investment for a variety of mutual funds:

Five-Year Rate of Return of Mutual Funds (as of 10/7/10)							
3.27	3.53	3.45	5.98	4.55	3.54	4.91	4.75
3.30	10.87	3.25	3.98	5.78	4.43	4.44	10.90
5.38	4.37	4.27	3.33	8.56	11.70	3.54	5.93
4.04	3.22	4.86	3.28	11.74	6.64	3.25	3.57
4.19	4.91	12.03	3.24	4.18	4.10	3.28	3.23

Source: Morningstar.com

Using the classes 3–3.99, 4–4.99, etc., make a frequency and relative frequency table.

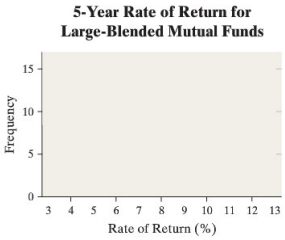
77

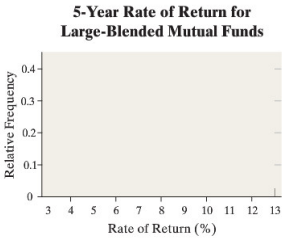
Class (5-year rate of return)	Tally	Frequency	Relative Frequency
3–3.99			
4–4.99			
5–5.99			
6–6.99			
7–7.99			
8–8.99			
9–9.99			
10–10.99			
11–11.99			
12–12.99			

78

ADDITIONAL NOTES

Draw histograms:





79

The Art of Choosing Classes

Too coarse?

Class (5-year rate of return)	Frequency
3–5.99	33
6–8.99	2
9–11.99	4
12–14.99	1

80

Too fine?

Class (5-year rate of return)	Frequency
3–3.49	11
3.5–3.99	5
4–4.49	8
4.5–4.99	5
5–5.49	1
5.5–5.99	3
6–6.49	0
6.5–6.99	1
7–7.49	0
7.5–7.99	0
8–8.49	0
8.5–8.99	1
9–9.49	0
9.5–9.99	0
10–10.49	0
10.5–10.99	2
11–11.49	0
11.5–11.99	2
12–12.49	1

81

Guidelines for Determining the Lower Class Limit of the First Class and Class Width:

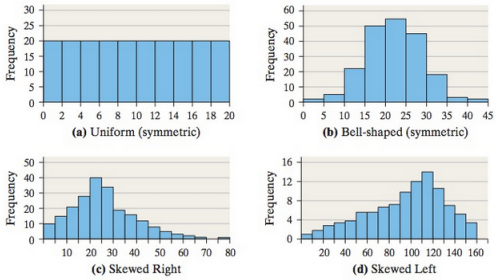
Choosing the Lower Class Limit of the first class:

82

For example, our smallest observation was 3.22. A convenient lower class limit of the first class is 3.

83

Shapes of Distributions



84

ADDITIONAL NOTES

Real-life data will rarely match theoretical shapes perfectly.

(65) Which theoretical shapes are the Wendy's and Mutual Fund histograms most like?

Wendy's: _____

Mutual funds: _____

85

Time-Series Data

If the value of a variable is measured at times, the data are called _____

You plot the time on the horizontal axis and the values of the variable on the vertical.

Since time-series data represent a progression, or an evolution, you usually connect the plotted points by lines in order to see trends.

86

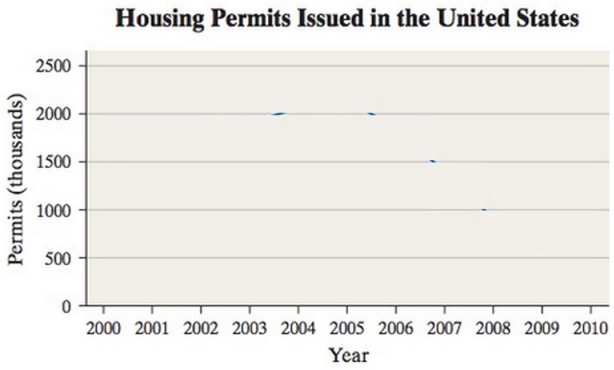
Plot the time-series data shown below:

TABLE 19

Year	Housing Permits (000s)
2000	1592.3
2001	1636.7
2002	1747.7
2003	1889.2
2004	2070.1
2005	2155.3
2006	1838.9
2007	1398.4
2008	905.4
2009	583.0
2010	592.9

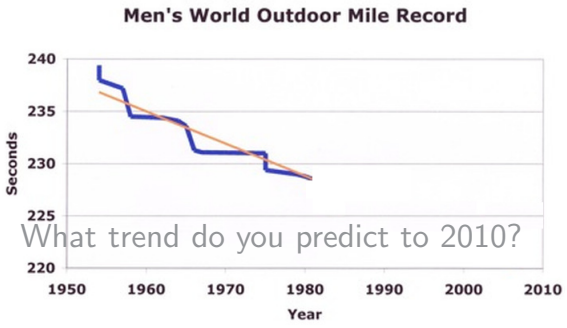
Source: U.S. Census Bureau

87



88

Here's a time-series plot of the record run-times for a mile, after the 4 min. barrier was broken:



What trend do you predict to 2010?

89

(66) Can the graph go down linearly, indefinitely in the future?

_____

(67) What trend do you predict for the future?

Where do you think will the graph level off?

90

ADDITIONAL NOTES

ADDITIONAL NOTES

Central Tendency

Central tendency describes the _____

_____

We try to see where the "middle of the data" is. "Average" is used in the sense of "middle."

Three common measures of central tendency are the _____, the _____, and the _____.

1

▷ The arithmetic **mean** of a variable is

The _____ ("mew"), is computed using all the individuals in a population. It's a _____

The _____ ("x-bar"), is computed using sample data. It's a _____

2

▷ The **median** of a variable is the value that's _____

_____

If there are two values in the middle (this happens when there's an even number of values) we take the mean of the two middle values.

We use ____ to represent the median.

3

▷ The **mode** of a variable is _____

_____

If no variable occurs more often than the others (for example if every value occurs just once), we say that there's no mode.

4

(1) I'm addressing a group of 10 people, and I ask them how much money they have in their pockets.

They say
$4, $6, $1, $4, $8, $2, $3, $1, $7, and $1.

What are the mean, median and mode of these values?

5

Mean:

6

ADDITIONAL NOTES

Median:
Arrange in ascending order:

Mode:

7

(2) In the previous example, which seems the least reliable measure of central tendency? _____

_____

(3) Someone now says "Wait a minute, I'm Mycroft Gates, Bill's better, older brother. By '$6' I meant $6 million." Which of the mean, median, and mode does this new information change?

8 _____

The mean becomes:

_____

The median is unchanged:

_____

The mode happens to stay unchanged.
(But could change in other circumstances.)

9

A numerical summary of data is called _____

_____

_____

_____

The word "substantially" is significant.
(4) If Mycroft had said, instead, "By '$1' I meant $1 million," which of the mean, median, and mode would *this* information change?

10

The mean becomes:

_____

The median also changes:

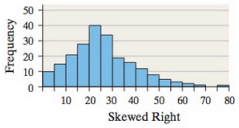
_____

_____

The mode changes to _____

11

There's a common (*not universal*) relationship between data-skew and mean, median and mode.
(5) When data are skewed right, how might you expect the mean, median and mode to be related?



_____

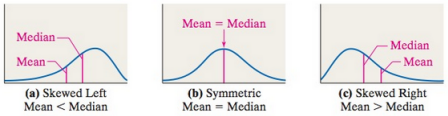
12 _____

ADDITIONAL NOTES

This is sometimes stated as a fact:

TABLE 4
Relation Between the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median



(a) Skewed Left Mean < Median (b) Symmetric Mean = Median (c) Skewed Right Mean > Median

but the "not always" of the previous slide is more accurate.

13

For example, consider the data

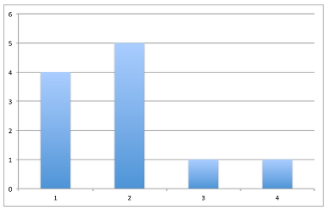
$$1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4$$

Make a frequency table:

x	f

14

The histogram looks like this



(6) Is this skewed right or left? _____

15

(7) Calculate the mean, median and mode of the preceding data (one decimal) and compare them.

Mean: _____

Median: _____

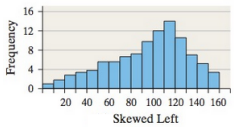
Mode: _____

So

_____

16

Similarly, when data are skewed left:

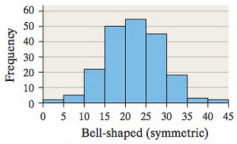


_____

_____

17

But when data are symmetric



_____

_____

18

ADDITIONAL NOTES

Dispersion

Only knowing the central tendency hides a lot about the data.

For example, suppose that the means on a test in two sections of a class, each with 10 students, are both 12 points out of 20.

(**8**) Does it follow that both sections did similarly?

19

(**9**) What are the means of the two sections?

§1: _____

§2: _____

20

Despite the means being the same, the two performances are quite different, and will need different approaches from the instructor.

21

Measures of Dispersion

We'll mainly study two measures: the _____ and the _____

The **range**, R, of a variable is defined as

_____

22

The **standard deviation** is an attempt to measure

_____

For a population, the deviation from the mean for the ith observation, x_i, is _____

To get the "average deviation" you might think you should _____

_____

23

But _____.

(**10**) Why?

_____

_____

_____

The cancelations can be avoided using the _____

_____

24

ADDITIONAL NOTES

We define the _____
(N observations) as:

(**11**) Why the squares? _____

_____

(**12**) Why the square root? _____

_____

25

X_i
82
77
90
71
62
68
74
84
94
88

Calculate the standard deviation of the data to the left.

Step 1: Get the mean.

_____

26

Then calculate $(x_i - \mu)^2$ for each row:

X_i	$(X_i-\mu)$	$(X_i-\mu)^2$
82		
77		
90		
71		
62		
68		
74		
84		
94		
88		

27

Add $(x_i - \mu)^2$: _____

Divide by number of observations: _____

Take square root (one decimal place):

$\sigma \approx$ _____

28

We define the _____
(sample size n) as:

(**13**) What is the main difference between this formula and the one for the population s.d.? _____

_____

29

The _____ of a variable is the square of the standard deviation.

The population variance is σ^2 and the sample variance is s^2.

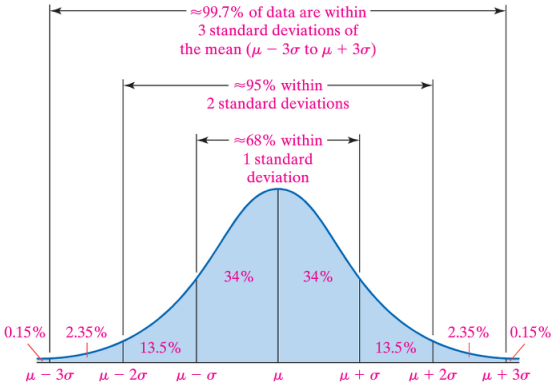
30

ADDITIONAL NOTES

The standard deviation is used with the mean to numerically describe distributions that are _____ _____ (and, so, symmetric).

The _____ of the distribution, and the _____

31



Empirical Rule

32

Chebyshev's Inequality

For any data set or distribution, symmetric or not, at least

of observations lie within k standard deviations of the mean, where k is any number greater than 1.

33

That is, at least $\left(1 - 1/k^2\right)100\%$ of the data lie between $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$.

(14) What percentage of observations lie within 2σ of the mean?

_____

34

ADDITIONAL NOTES

Grouped Data

Our calculations of mean and standard deviation work _____

Often, for various reasons, we only have a summary given by the frequency table.

1

The Mean from Grouped Data

When the data are quantitative and continuous, and all we have available is a frequency table, we need a "class representative" for each class.

We choose the _____ _____, defined as _____ _____

2

Look at this table from Week 2 (bachelor's degrees in the population, by age group):

TABLE 10

Age	Number (in thousands)
25–34	12,967
35–44	13,904
45–54	13,005
55–64	10,357
65–74	4,584

Source: Current Population Survey, 2008

(1) What is the class midpoint of each class?

3

Then the _____ is:

and the _____ is

where _____ _____

4

Consider this example, from week 2:

TABLE 12

Five-Year Rate of Return of Mutual Funds (as of 10/7/10)

3.27	3.53	3.45	5.98	4.55	3.54	4.91	4.75
3.30	10.87	3.25	3.98	5.78	4.43	4.44	10.90
5.38	4.37	4.27	3.33	8.56	11.70	3.54	5.93
4.04	3.22	4.86	3.28	11.74	6.64	3.25	3.57
4.19	4.91	12.03	3.24	4.18	4.10	3.28	3.23

Source: Morningstar.com

(2) What is the mean of these values (three decimals)? _____

5

Using the classes 3–3.99, 4–4.99, etc., we'd made a frequency table.

Fill in the class midpoints on the table (next slide).

6

ADDITIONAL NOTES

Class (%)	Freq. (f_i)	Mid (x_i)	$x_i f_i$
3–3.99	16		
4–4.99	13		
5–5.99	4		
6–6.99	1		
7–7.99	0		
8–8.99	1		
9–9.99	0		
10–10.99	2		
11–11.99	2		
12–12.99	1		

7

(3) What is the mean from the frequency distribution table?

8

Weighted Mean

This mean is an example of a _____

Each value, x_i is _____.

(If each $w_i = 1$, you have the standard mean.)

9

Example of weighting:

Every U.S. state has two senators. All their votes in the senate are equal; in other words they are weighted equally:

$$w_{CA} = w_{NY} = w_{IN} = w_{ND} = 1, \text{ etc.}$$

(4) What are the disadvantages of such a system?

10

(5) Why was such a system set up?

Now, suppose the value of each senators vote were weighted by half the population percentage of that state.

(6) Why half? _____

11

Selected population percentages

State	Rounded % of U.S. population
CA	%
TX	%
FL, NY	%
AZ, IN, MA, TN	%
ND	%
AK, VT, WY	%

12

ADDITIONAL NOTES

If the senate votes were weighted in a way that reflects population percentage, what would these weights be (remember two senators/state)?

- $w_{CA} =$ _____
- $w_{NY} =$ _____
- $w_{IN} =$ _____
- $w_{ND} =$ _____

13

Another example: GPA

Suppose that last term you took
a 5-credit chemistry class (got a C),
a 3-credit english class (B),
and a 6-credit dance class (A).

If an A is worth 4 points, a B worth 3 points and a C worth 2 points, what was your GPA?

14

(7) What are you averaging? _____

(8) Keeping in mind that you must weight the grade points from each class with the credits for that class, calculate your GPA.

15

The Standard Deviation from Grouped Data

where _____

_____

16

Let's return to our friend from week 2:

TABLE 12
Five-Year Rate of Return of Mutual Funds (as of 10/7/10)

3.27	3.53	3.45	5.98	4.55	3.54	4.91	4.75
3.30	10.87	3.25	3.98	5.78	4.43	4.44	10.90
5.38	4.37	4.27	3.33	8.56	11.70	3.54	5.93
4.04	3.22	4.86	3.28	11.74	6.64	3.25	3.57
4.19	4.91	12.03	3.24	4.18	4.10	3.28	3.23

Source: Morningstar.com

(9) What is the sample standard deviation (to one decimal place)? _____

(10) Now complete the table on the next slide (using the mean from Slide 8).

17

Class (%)	f_i	x_i	$(x_i - \bar{x})^2 f_i$
3–3.99	16	3.5	
4–4.99	13	4.5	
5–5.99	4	5.5	
6–6.99	1	6.5	
7–7.99	0	7.5	
8–8.99	1	8.5	
9–9.99	0	9.5	
10–10.99	2	10.5	
11–11.99	2	11.5	
12–12.99	1	12.5	

18

ADDITIONAL NOTES

(11) Get the sample standard deviation.

The s.d. from the raw data is _____: good agreement.

19

Comparisons

New York Times, January 8, 2016

Seattle Seahawks' Defense Stands Out, Even Across Eras

"Over the last four years, Seattle has allowed 15.73 points a game, relative to the league average of 22.89. The standard deviation in points allowed by the league's 32 teams over this period has been 2.62 points per game. This means Seattle, which has been over 7 points a game better than average, has been 2.73 standard deviations better than average, a statistic known as a Z-score."

20

Example: There are two sections of a course. The statistical results from them are:

- Section 1: Mean: 16.5; $\sigma = 2.8$.
- Section 2: Mean: 17; $\sigma = 2.9$.

Your friend is in section 1 and scored 18.2.
You're in section 2 and scored 18.5.
Who did better?

21

_____ to the rescue

In order to compare two values, we use their z-scores. For a variable x, its z-score is

A z-score tells you how many standard deviations you are from the mean.

22

(12) For the given z-scores, how many s.d.s are you above or slow the mean?

- $z = 1$: _____
- $z = -2.5$: _____

(13) Where did "above" and "below" come from?

_____

(14) For a value of the variable equal to the mean, what is the corresponding z-score? ____

23

(15) Calculate the test z-scores for you and your friend. Then say who did better.

$z_{\text{you}} =$

and

$z_{\text{friend}} =$

24

ADDITIONAL NOTES

(16) The statistics from two marathons are:

- Marathon 1: $\mu = 4.5$ hours; $\sigma = 0.75$ hours.
- Marathon 2: $\mu = 4.67$ hours; $\sigma = 0.5$ hours.

You're in marathon 1 and run it in 4.25 hours. Your friend is in marathon 2 and runs it in 4.6 hours. Who did better?

First get z-scores.

25

$z_{\text{you}} =$

and

$z_{\text{friend}} =$

(17) Which is bigger? _____

(18) Which is *better*? _____

_____

26

Percentiles

The kth percentile of a set of data, P_k, is the value such that $k\%$ of the values are less than or equal to it.

27

(19) You've seen another name for the 50th percentile, P_{50}. What is it ? _____

(20) If your SAT scores are in the 85th percentile what percentage of scores are better than yours?

_____

28

Two sections of a course, each with a hundred students, take a test.

(21) In one you get $16/20$; everybody else gets between 16.1 and 16.5. What percentile is your score in? _____

(22) In another your friend also gets $16/20$; everybody else gets between 14.8 and 15.9. What percentile is your friend's score in? _____

29

(23) What *percentage* is your score? _____

(24) What percentage is your friend's score?

_____

(25) What is the moral of the story?

_____

30

ADDITIONAL NOTES

Quartiles

Some percentiles have special names: _____

- Q_1: _____
- Q_2: _____
- Q_3: _____

For discrete data, Q_1 is the median of the "lower half" and Q_3 the median of the "upper half."

31

(**26**) What are the quartiles for

$$2, 5, 7, 9, 11, 1, 13, 4, 8?$$

First arrange in ascending order:

_____

Then get median: _____

Then get medians of lower and upper halves, *without the median*: _____

32

Quartiles and Dispersion

We've seen two measures of dispersion: the range and the standard deviation.

(**27**) Are they resistant? _____

The _____ is defined as

(**28**) Do you expect it to be resistant? _____

33

What are σ and R for

(**29**) $1, 2, 3, 4, 5, 6, 7$?

_____

(**30**) $1, 2, 3, 4, 5, 6, 7000$?

_____

As expected, these measures of dispersion are not resistant.

34

(**31**) What are the quartiles for the previous data sets?

Set 1: _____

Set 2: _____

(**32**) What are the IQRs?

_____

(**33**) Does the IQR seem resistant? _____

35

Outliers

A value in the data is considered an _____

_____

_____

For example,

_____, or

_____.

36

ADDITIONAL NOTES

The Five Number Summary of Data

Especially for skewed distributions (therefore, outliers), this is a good way to summarize data.

For symmetric distributions, μ and σ (or $\bar{x}$ and s) are good.

37

(34) Find the five number summary of these data:

$$25.55, 25.83, 26.28, 42.47, 49.17, 54.63, 28.58,$$

$$28.72, 30.95, 30.18, 30.35, 32.13, 33.23, 33.53,$$

$$36.68, 37.05, 19.95, 23.25, 23.32, 37.43, 41.42$$

38

ADDITIONAL NOTES

ADDITIONAL NOTES

ADDITIONAL NOTES

ADDITIONAL NOTES

Review of Standard Deviation

Population (N observations) Sample (sample size n)

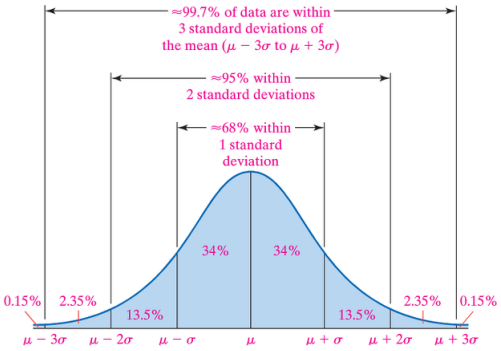
1

(1) Why $n-1$ in the sample formula, not n?

2 See the data on the right.

TEST 1	Scores	Sample
	16.5	*16.5*
	19.0	*20.0*
	14.0	*19.0*
	21.0	*20.0*
	20.0	*19.0*
	12.0	*16.0*
	16.0	*20.0*
	19.0	*16.0*
	18.0	*18.0*
	19.0	*16.0*
	20.0	
	21.0	
	19.0	
	16.0	
	20.0	
	21.0	
	20.0	
	20.0	
	18.0	
	21.0	
	16.0	
	17.0	
	20.0	
	16.5	
	15.0	
	18.0	
	17.0	
	21.0	
	16.0	
	18.0	
Mean	18.2	*18.1*
Pop. Std Dev.	2.3	*1.7*
Samp. Std Dev.		*1.8*

Where are most of the data?



Empirical Rule: Page 138, fig 13.

3

The standard deviation (sd) of data is a measure of how _____ the values are.

The _____, relative to the mean, the more _____ the data are from it.

The _____, relative to the mean, the more _____ the data are around it.

4

Examples

(2) What is the mean (μ) and the standard deviation (σ) of the data set below?

$\{1,1,1,1,1,1\}$ $\mu =$ ____ $\sigma =$ ____

(3) What is the mean (μ) and the standard deviation (σ) of the data set below?

$\{1,1,1,-1,-1,-1\}$ $\mu =$ ____ $\sigma =$ ____

5

(4) What is the mean (μ) and the standard deviation (σ) of the data set below?

$\{4,4,4,2,2,2\}$

$\mu =$ ____ $\sigma =$ ____

6

ADDITIONAL NOTES

(5) What is the mean (μ) and the standard deviation (σ) of the data set below?

$$\{20, 20, 20, 10, 10, 10\}$$

$\mu =$ _____ $\sigma =$ _____

7

(6) What is the mean (μ) and the standard deviation (σ) of the data set below?

$$\{-20, -20, -20, -10, -10, -10\}$$

$\mu =$ _____ $\sigma =$ _____

8

OK, Sherlocks, you've uncovered these properties:

- _____
  _____
- _____
  _____
- _____
  _____

Where c is a constant and X is the data set.

9

Relationships

So far we've discussed problems with a single variable: _____

We'll now discuss problems with two variables:

_____

We'll be interested in _____
_____

10

A _____ is a graph that shows the relationship between two quantitative variables on the same individual. Each individual is represented by a point in the diagram.

11

We view one variable as the _____ (dependent) variable and the other as _____ (independent).

The _____
_____, and the _____
_____

12

ADDITIONAL NOTES

It is _____

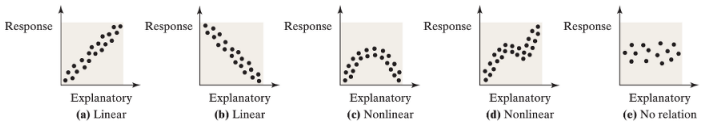
_____

For example, does high school GPA predict a student's SAT score or can SAT score predict GPA?

The researcher must determine which variable plays the role of explanatory variable based on the questions he or she wants answered.

13

Examples of scatter diagrams



We're interested in how variables that are linearly related might be _____.

14

Linearly related variables are _____ when above-average values of one are associated with above-average values of the other, and below-average values of one are associated with below-average values of the other.

That is, two variables are positively associated if,

_____

15 _____

Linearly related variables are _____ when above-average values of one are associated with below-average values of the other.

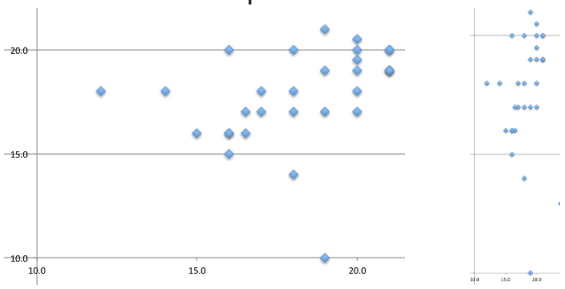
That is, two variables are negatively associated if,

_____

_____

16

Trying to spot an association visually is not reliable. Here's a scatter plot. Is there a correlation?



17

The _____ is a measure of the strength and direction of the linear relation between two quantitative variables. The letter r represents the sample correlation coefficient (and ρ for population).

18

ADDITIONAL NOTES

Sample Linear Correlation Coefficient

$\bar{x} =$ _____

$s_x =$ _____

$\bar{y} =$ _____

$s_y =$ _____

$n =$ _____

19

Properties of the Linear Correlation Coefficient, r

- _____
- If _____ _____ between the variables.
- If _____ _____ between the variables.

20

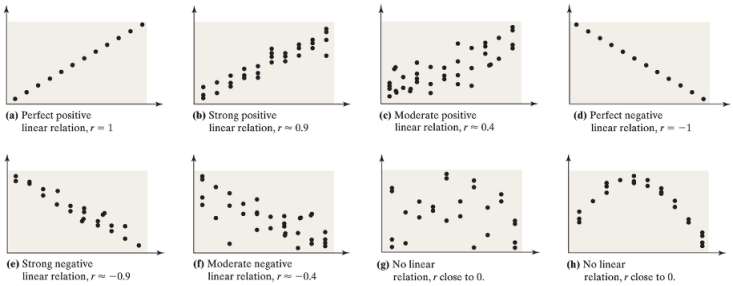
- The _____ _____ between the variables.
- The _____ _____ between the variables.
- If _____ _____ between the variables.

Note: _____ _____

21

- The linear correlation coefficient is _____. The unit of measure for x and y plays no role in the interpretation of r.
- The correlation coefficient is _____. An observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

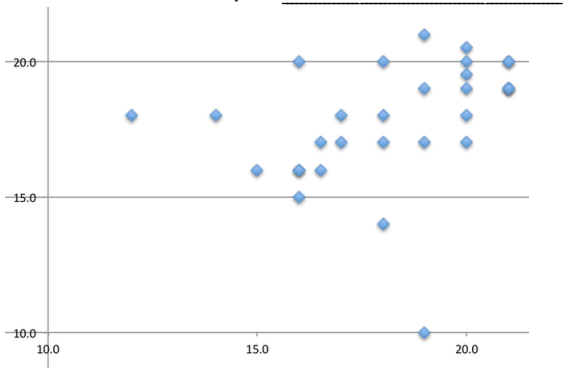
22



How close to zero does r have to be for you to know there's no linear relation?

23

Back to that example _____



24

ADDITIONAL NOTES

Is $r = 0.33$ close enough to zero to say there's no linear relation?

Depends on the sample size.
If the value of $|r|$ is lower than the _____ for _____ then there's low evidence for a linear relation.

25

Critical values for r					
n	r-crit	n	r-crit	n	r-crit
4	0.950	13	0.553	22	0.423
5	0.878	14	0.532	23	0.413
6	0.811	15	0.514	24	0.404
7	0.754	16	0.497	25	0.396
8	0.707	17	0.482	26	0.388
9	0.666	18	0.468	27	0.381
10	0.632	19	0.456	28	0.374
11	0.602	20	0.444	29	0.367
12	0.576	21	0.433	30	0.361

Appendix A, table 2

26

(7) OK, you statistical sluggers, is there a (positive) linear relation in our example?

27

Correlation vs Causation

The existence of an association does not prove causation.

The behavior of two variables may seem related, but neither may *cause* the behavior of the other.

For example, there's a positive association between air-conditioning bills and high crime.

28

(8) Does that show causation? Do people go crazy with high bills and go out and commit crimes?

A _____ is an initially hidden variable that's related to both explanatory and response variables.

29

(9) Dandelions and daisies tend to be found together on sports fields or in public parks in numbers that increase or decrease together (a positive association as they vary together in the same way). Causation or correlation?

30

ADDITIONAL NOTES

(10) Thyme tends to be found where the soil has low acidity. Is the relationship between the number of thyme plants and soil acidity a positive or negative association? _____

(11) Is the association between the number of thyme plants and soil acidity likely to be correlation or causation? _____

31

(12) Studies show a positive association between listening to loud music and active acne. Does listening to loud music cause acne? _____

(13) Does acne cause listening to loud music? _____

(14) If the positive association between listening to loud music and active acne is correlation, not causation, what might be a lurking variable? _____

32

Studies show a positive association between hand size and reading ability.

(15) Do bigger hands cause better reading ability? _____

(16) Does better reading give you bigger hands? _____

33

(17) If the positive association between hand size and reading ability is correlation, not causation, what might be a lurking variable? _____

34

(18) Studies show a negative association between cell phone use and sperm count. As cell phone use goes up, does sperm count seem (according to these studies) go up or down? _____

(19) Hard to know, for sure, but does lower sperm count cause higher cell phone use? _____

(20) Hard to know, for sure, but does higher cell phone use cause lower sperm count? _____

35

(21) If the negative association between cell phone use and sperm count is correlation, not causation, what might be a lurking variable? _____

36

ADDITIONAL NOTES

We are looking at _____ That is, we have two variables, say x and y, and are studying possible relationships between them. _____

_____

_____

We are looking at _____.

Unless otherwise stated, from now on _____

_____

1

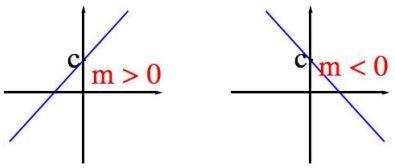
To further develop the idea of a linear association, we need to remind ourselves of linear functions.

2

The General Linear Function

$$mx + c$$

- Graph points upward (reading from left to right) when _____, and downward when _____. m is called the _____.
- The y-intercept is _____.
- The x-intercept is the _____ .



3

Slope

- Slope is _____ It tells you how much vertical increase or decrease you get as you increase the independent variable by 1.
- _____
  The greater the slope, the quicker the increase.
- _____
  The more negative the slope, the quicker the decrease.

4

(1) If $y = 3x - 4$, how much does y increase or decrease by when x increases by 1 unit?

_____

5

(2) If $2y = -5x + 2$, how much does y increase or decrease by when x increases by 1 unit?

_____

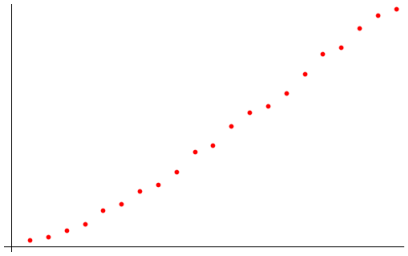
_____

_____

6

ADDITIONAL NOTES

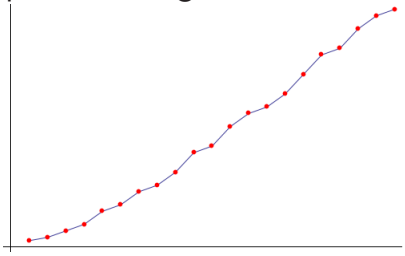
Linear Fitting



How do you fit a curve to these data?

7

"Curve fitting" is not simply drawing a curve by hand that passes through the data, like so:



The question is: _____

8

(3) Why do this?

_____

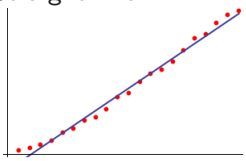
_____

_____

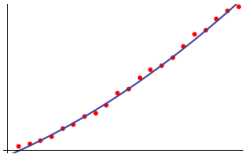
9

You can, in principle, try many different fits:

Straight line



Curve



We will confine ourselves to straight lines: _____

_____

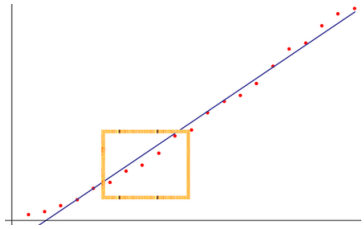
10

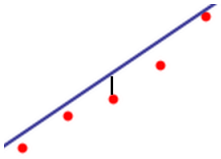
How do we get the best fit?

There's no one answer, but one method is to use the the _____.

To study this, let's "blow up" part of the example we have been looking at.

11





Observed value: _____

Predicted value: _____

Residual: _____

_____

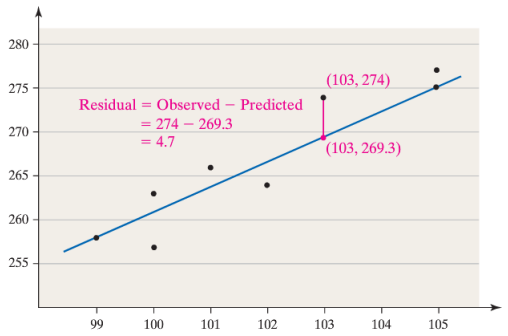
12

ADDITIONAL NOTES

(4) In the previous, for the indicated point, is the residual positive or negative? _____

13

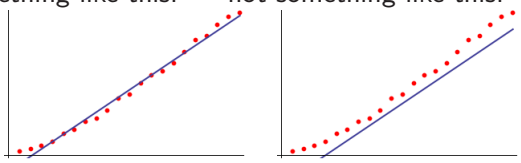
Another example:



14

Since the residual is a measure of the error, clearly we want to minimize it overall. We want

something like this: not something like this:



15

(5) Signs of the residuals in the previous?

_____

(6) If you add the residuals to get the total error, what answer would you expect for a line that offers a decent fit? _____

_____

(7) What to do when cancellations get in the way?

_____

16

_____

_____

$$\hat{y} = mx + c$$

where

$$m =$$

$$c =$$

_____

(Note: $\hat{y}$ distinguishes the predicted from actual value of y.)

17

(8) What are these?

- $\bar{x}$: _____
- $\bar{y}$: _____
- s_x: _____
- s_y: _____
- r: _____

18

ADDITIONAL NOTES

(**9**) Why must the l-s line pass through $(\bar{x}, \bar{y})$? (In other words, when $x = \bar{x}$, why *must* $\hat{y} = \bar{y}$?)

19

(**10**) What is the sign of the standard deviation?

_____

(**11**) Therefore, the sign of m, the slope of the least-squares regression line, is the same as what?

_____

When __ is positive, ___ is too. Ditto negative.

20

_____

_____

For example, if predicted value of the GPA when a student studies 15 hrs/wk is 3, then that's the mean of all students who study 15 hrs/wk, not necessarily the value for any actual student.

21

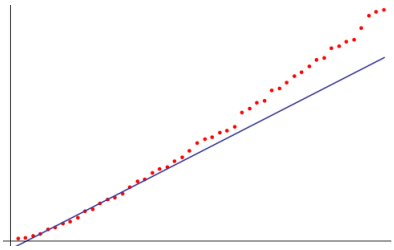
The slope of the regression line is the _____ _____ For every unit increase in x, we expect y to rise (or fall) by m on average.

The y-intercept, c, has meaning only if $x = 0$ makes sense for the particular scenario, and we have observations of x close to zero.

In general, we cannot push the regression line too far in any direction.

22

The data we have been looking at is the list of prime numbers. Our line fits the first 21 primes but fails if we look at the first 50:



23

Another example: there are (controversial) data that link income to height: The taller you are the _____ you seem to make.

One linear regression fit to data suggests that

$$\$I = 1560h - 84000$$

where $\$I$ is _____ and h is _____

24

ADDITIONAL NOTES

(12) What's the vertical ("y") intercept?

_____

(13) What is its interpretation?

_____

(14) Is that meaningful? _____

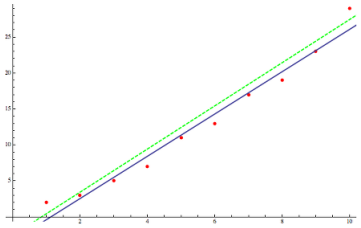
25

(15) Do you expect the least-squares linear regression line to be resistant, or not? _____

(16) Why? _____

26

Here are two different fits to the first 10 primes:



L-S line: $\hat{y}_{LS} = 2.9x - 3.3$

Other (dashed): $\hat{y}_O = 3x - 2.5$

Calculate squared residuals for each and add.

27

For example, the first prime ($x = 1$) is 2 ($y = 2$):

$\hat{y}_{LS} = 2.9(1) - 3.3 =$ _____

$\hat{y}_O = 3(1) - 2.5 =$ _____

So,

$(y - y_{LS})^2 = (2 - (-0.4))^2 = 2.4^2 =$ _____

$(y - y_0)^2 = (2 - (0.5))^2 = 1.5^2 =$ _____

Enter this on the table below, and fill in the rest.

28

x	y	$\hat{y}_{LS}$	$(y - \hat{y}_{LS})^2$	$\hat{y}_O$	$(y - \hat{y}_O)^2$
1	2				
2	3				
3	5				
4	7				
5	11				
6	13				
7	17				
8	19				
9	23				
10	29				

29

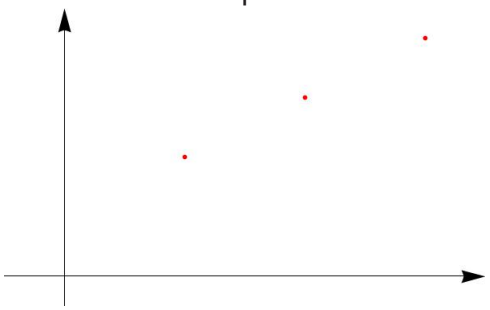
(17) On each of the next four graphs do this:

- Plot what you think a good straight-line fit would be.
- If there's a value of x that has two values of y plot the mean y with an open circle.
- Draw open circles around the other y values (that are each a single match for x).
- Draw a faint line that best fits the open circles.

30

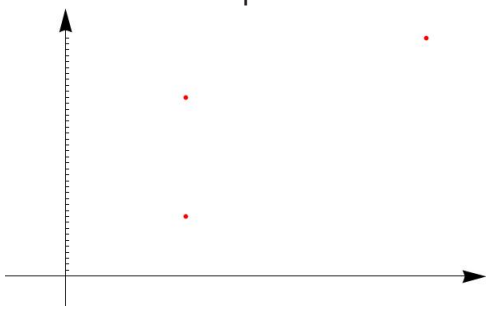
ADDITIONAL NOTES

Graph A



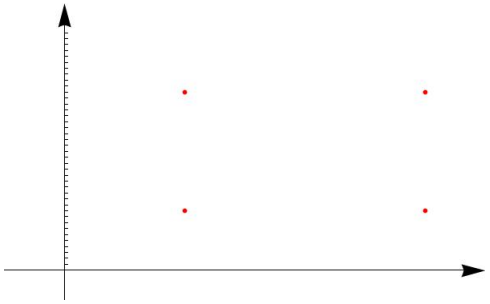
31

Graph B



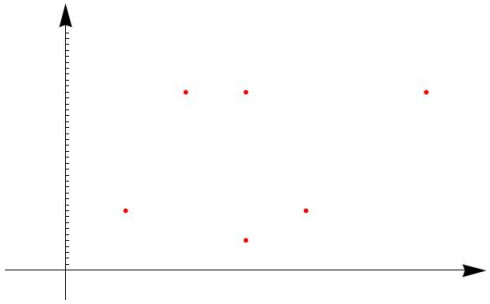
32

Graph C



33

Graph D



34

(**18**) For graph B, calculate the sum of the squares of the vertical distances of the given points from both your line and the shown l-s line, and compare:

Point	d^2 to your line	d^2 to l-s line
1		
2		
3		
Sum		

35

(**19**) Ditto for graph C:

Point	d^2 to your line	d^2 to l-s line
1		
2		
3		
4		
Sum		

36

ADDITIONAL NOTES

ADDITIONAL NOTES

ADDITIONAL NOTES

(1) What is the probability that a coin toss will come up tails?

_____

Does this mean that if you toss a coin

(2) 10 times, you'll get exactly 5 tails? _____

(3) 10,000 times, you'll get 5,000 tails? _____

(4) 10 billion times, you'll get 5 billion tails?

1 _____

Then, what does a probability of $1/2$ mean?

A clue is provided by asking this:

(5) In which of these cases (10 tosses, 10 thousand, or 10 billion) do you expect the *proportion* of tails to total tosses to be closest to $1/2$?

_____

2

The Law of Large Numbers

_____

_____

_____

_____

How is this "theoretical probability" calculated?

We'll need some terms...

3

An _____ is any process with uncertain results that can be repeated.

The result of any single trial of the experiment is not known ahead of time.

But, the results of the experiment over many trials produce regular patterns that enable us to predict results: _____

_____

4

The _____, $\mathcal{S}$, of a probability experiment is the collection of all possible outcomes.

An _____ is any collection of outcomes from a probability experiment. An event consists of one outcome or more than one outcome.

We'll denote events with one outcome, sometimes called simple events, by e_i. In general, events are denoted using capital letters such as E.

5

Probability attempts to capture _____

_____

If the event, E, is one possible outcome (or group of outcomes), the theoretical probability is

_____

6

ADDITIONAL NOTES

Suppose you're interested in the probability that a roll of one die will result in an even number.

Number of possible outcomes: _____

Number of "even" outcomes: _____

Therefore, the probability of getting an even number is _____.

7

(6) What is the probability that throwing a die will give a number greater than 4?

_____ _____

(7) What is the probability of rolling a prime?

_____ _____

(8) If you roll two dice, what is the probability that *at least* one of them will show a 3?

8

In order to answer this question you must first list all the possibilities:

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

9

There are _____ possible outcomes.

Which of these show at least one three?

10

(9) What is the probability that *exactly* one of the pair will show a three?

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

11

(10) What is the probability that the sum of the two dice will be seven?

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

12

ADDITIONAL NOTES

(**11**) What is the probability that the sum of the two dice will be less than or equal to five?

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

13

Of course, there's more in the world than two dice. There are also two coins. One way to list the possibilities is

(H,H)
(H,T)
(T,H)
(T,T)

There are _____ possible outcomes in all.

14

(**12**) What is the probability of both coins showing the same face?

(H,H)
(H,T)
(T,H)
(T,T)

15

(**13**) When you toss two coins what is the probability of at least one head?

(H,H)
(H,T)
(T,H)
(T,T)

16

(**14**) When you toss two coins what is the probability of exactly one head?

(H,H)
(H,T)
(T,H)
(T,T)

17

Here are all the outcomes if you toss 3 coins:

(H,H,H)
(H,H,T)
(H,T,H)
(H,T,T)
. . . . . .
(T,H,H)
(T,H,T)
(T,T,H)
(T,T,T)

18

ADDITIONAL NOTES

(15) Why are the possibilities on the previous slide grouped the way they are?

19

Based on how the 2-coin table (total of 4 possibilities) was used to build the 3-coin table (total of 8 possibilities), how may possibilities would there be totally if you tossed

(16) 4 coins? _____

(17) 5 coins? _____

20

(18) Can you work out a formula for the total number of possibilities if you toss n coins?

One coin: _____

Two coins: _____

Three coins: _____

Four coins: _____

Therefore, if you toss n coins, you'll have

21 _____

(19) In a 3-coin toss, what is the probability of exactly 2 heads?

(H,H,H)
(H,H,T)
(H,T,H)
(H,T,T)
(T,H,H)
(T,H,T)
(T,T,H)
(T,T,T)

22

(20) In a 3-coin toss, what is the probability of a single tail?

(H,H,H)
(H,H,T)
(H,T,H)
(H,T,T)
(T,H,H)
(T,H,T)
(T,T,H)
(T,T,T)

23

(21) In a 3-coin toss, what is the probability of at least one tail?

(H,H,H)
(H,H,T)
(H,T,H)
(H,T,T)
(T,H,H)
(T,H,T)
(T,T,H)
(T,T,T)

24

ADDITIONAL NOTES

General Properties of Probability

(22) Can an event have a probability of 5?

25

(23) What is the largest and smallest possible probability?

26

A probability of ___ is meant to capture the idea that an event will _____

For example, when you roll a die, none (zero) of the possible outcomes is -1.
Therefore, the probability of rolling -1 is

$$0/6 = 0.$$

27

A probability of ___ is meant to capture the idea that an event is _____

For example, the probability that you will roll *some* number between 1 and 6 is $6/6 = 1$.

28

The previous example can also be approached by adding the probabilities of getting each of the single outcomes ("simple events"): rolling 1, 2,...,6.

The probability for each is $1/6$. The sum is

$$\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{6}{6}=1.$$

29

That illustrates a general truth:

If a sample space, $\mathcal{S}$, consists of the outcomes $\{e_1, e_2, \ldots, e_n\}$, then

This statement, along with

must be satisfied by all "probability models" (lists of outcomes and their probabilities).

30

ADDITIONAL NOTES

(24) Is this a probability model?

Means of Travel	Probability
Drive alone	0.765
Carpool	0.11
Public transportation	0.05
Walk	0.025
Other means	0.015
Work at home	0.035

Table 3, p.237

31

In practice very few things of interest are either impossible or certain. What one has to look for is which events are likely, and how likely they are. It's useful to establish a "threshold" for events to be more likely to happen than not.

An event is considered more likely than not, if it has a probability _____.

32

Probabilities are sometimes expressed as percentages, simply by multiplying them by 100.

An event with a probability of 0.65 is said to have a _____ of occurring.

33

If there are only two distinct possible outcomes of a scenario, and the probability that one of them will occur is p, the probability of the other occurring is _____. (See Q 21.)

Such events are called _____.

For example, these are complementary events:

Rolling an odd number and rolling an even.

Picking a red card and picking a black card.

34

Now, the occurrence of an event and its *non-occurrence* are complementary. (Wither something will happen or it will not.)

So, another way to look at this is that if the probability that something will occur is p, then the

35

These are theoretical probabilities. How do they compare with what happens in the "real world"?

We define the _____ of an event E, denoted by $RF(E)$, in N trials as

$$\_\_\_\_\_$$

As _____ we expect _____

36

ADDITIONAL NOTES

This is called the _____ approach to probability.

Statisticians, such as yourselves, collect data and calculate empirical probabilities based on the data.

They then compare with the theoretical probability to see if there is a match, or – if there is not – why not.

37

For example, suppose a study of 500 3-child families reveals that 180 of these have two girls and one boy.

(**25**) Calculate the empirical property (i.e., RF) of such an event.

38

(**26**) What is the theoretical probability of such an event? (Hint: You know the answer.)

_____

_____

39

(**27**) The empirical probability is about 4%* smaller than the theoretical. What might be reasons?

40

Combining Probabilities

Let A and B be two probability events.

The probability of _____ happening is

_____

41

(**28**) Rolling two dice: What's the probability you'll get exactly one three, or that the sum will be five?

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

42

ADDITIONAL NOTES

In situations where two (or more) events have no overlap, the probability that one or the other will occur is very simple.

Such events are called _____.
For disjoint events, E_1, E_2,..., E_n, we have

43

Combining Probabilities: Independent Events

Two events are called _____ if the outcome of one doesn't affect the outcome of the other.

For example, getting an even number on the roll of a die and getting a heads on a coin toss are independent.

What's the probability that you will get both?

44

There are four groupings of possibilities:

$[(2 \text{ or } 4 \text{ or } 6), H]$ $[(2 \text{ or } 4 \text{ or } 6), T]$

$[(1 \text{ or } 3 \text{ or } 5), H]$ $[(1 \text{ or } 3 \text{ or } 5), T]$

In only one do you get both an even number on the die and a heads on the coin. So the probability is $1/4$.

45

The rule for calculating the probability that two independent events, A and B, will both happen is

$$P(A \text{ AND } B) = P(A) \cdot P(B).$$

(**29**) Apply this to the previous example.

46

The multiplication rule for the probabilities of independent events extends to more than two events.

For independent events, E_1, E_2,..., E_n, we have

47

Suppose that studies show that the probability that a randomly selected 24-year-old male will survive the year is 0.9986.

(**30**) What is the probability that three randomly selected males of this age will all survive the year?

48

ADDITIONAL NOTES

(31) What is the probability that at least one male in the age group above in a set of 20 such males will die this year?

To answer directly we would calculate:

$$P(\text{1 dies}) + P(\text{2 die}) + \ldots + P(\text{20 die})$$

This is difficult.

49

An indirect approach is to calculate the complementary probability.

(32) What is it?

_____

50

(33) From this what is the probability that at least one dies?

_____

_____

51

Conditional Probability

This is the probability of _____ _____ It is calculated by restricting the set of all possible outcomes to those that satisfy the condition.

The notion that is used is _____, and is read as the probability that the event F occurs, given that the event E has occurred.

52

For example, if you throw two dice what is the probability that at least one will show a 3 under the condition that the sum on the two is greater than or equal to 5?

53

Start by restricting the possibilities:

(1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1) (2,2) (2,3) (2,4) (2,5) (2,6)
(3,1) (3,2) (3,3) (3,4) (3,5) (3,6)
(4,1) (4,2) (4,3) (4,4) (4,5) (4,6)
(5,1) (5,2) (5,3) (5,4) (5,5) (5,6)
(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)

A total of 30 possibilities obey the condition.

54

ADDITIONAL NOTES

(34) How many of these show at least one three?

55

Another way: use the conditional probability rule:

$$\frac{\qquad\qquad}{\qquad\qquad\qquad\qquad\qquad\qquad}$$

(35) Use this rule in the previous example.

56

Extra Homework

A] See this Johnny Carson clip

http://www.cornell.edu/video/?videoid=2334.

About 12 seconds into it he makes a statement about probability, then attempts to check it by surveying the audience.

The experiment fails. What is his mistake?

57

B] You're on a game show. You are shown three closed doors. Behind one of them is a prize (a complete set of MP3s of lectures by famous mathematicians). You choose a door. The host opens *one of the others*, shows you there was no prize there, and asks whether you want to change your choice of door. Will changing your choice change your probability of winning (either go up or down) or will it stay the same?

58

Work out the previous problem yourself, then investigate to see if you were right. One discussion is at http://en.wikipedia.org/wiki/Monty_Hall_problem.

(The problem has been known by mathematicians for some time, but it rose to real fame after it was featured by the Parade Magazine columnist Marilyn vos Savant (supposedly the owner, for a while, of the world's highest IQ).(See http://en.wikipedia.org/wiki/Marilyn_vos_Savant.

59

ADDITIONAL NOTES

Counting

Calculating probabilities requires counting.

For example:

- What is the total number of outcomes?
- What is the number of outcomes of interest?

(1) So, do you know how to count?

1

(2) What's an efficient way to count these?

♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡♡♡
♡♡♡♡

2

Multiplication is a counting technique. On the right is the menu from a fancy New York restaurant.

(3) How many different meals are possible?

per se

CHEF'S TASTING MENU

March 17, 2017

"OYSTERS AND PEARLS"
"Salmon" of Pearl Tapioca with Island Creek Oysters and Sterling White Sturgeon Caviar

ROYAL KALUGA CAVIAR*
Cauliflower "Panna Cotta," Sorrell Salt Tuiles, Yuzu Gelee, and Three Avocado Purees
($60.00 supplement)

SALAD OF HOLLAND WHITE ASPARAGUS
Persian Cucumbers, Cherry Belle Radishes, Mustard Blossoms, and "Grilletes"

SLOW-POACHED HUDSON VALLEY MOULARD DUCK FOIE GRAS
Tokyo Turnips, Muscat Grapes, Toasted Pistachios, and Condensed Milk "Tuile"
($80.00 supplement)

FILLET OF SAUSH SEA HALIBUT
Pinkeye Crab, Melted Green Garlic, Cocktail Artichokes, and Nutty Pear Emulsion

BUTTER-POACHED NOVA SCOTIA LOBSTER
Golden Beets, Horseradish Créme Fraîche, Ruby Frost Apples, and "Brussels"

"BREAD AND BUTTER"
Bitter Cocoa Laminated Brioche and Diane St. Clair's Animal Farm Butter

LIBERTY FARMS PEKIN DUCK BREAST*
Wilted Arrowhead Spinach, Sarawan Mandarins, French Lentils, and "Sweet Perfection"

HAND-CUT "TAGLIATELLE"
"Parmigiano-Reggiano" and Shaved Black Winter Truffles
($25.00 supplement)

ELYSIAN FIELDS FARM "CARRÉ D'AGNEAU"*
Green Asparagus, Gem Lettuces, Spring Onion Confit, and Bonita's Bleu "Cheesique"

CHARCOAL-GRILLED WAGYU*
"Pioneers Acres," Morel Mushrooms, Sweet Carrots, and "Sweet Bordelaise"
($60.00 supplement)

"DOUGERE"
Amikuni Dairy "Brulé"

"ASSORTMENT OF DESSERTS"
Fruit, Ice Cream, Chocolate, and "Candies"

PRIX FIXE $295.00
SERVICE INCLUDED

3

Multiplication Rule: If a job has steps with p possibilities for the first step, q for the second, r for the third..., then the number of results is

_____

We saw this for coin tosses: 2 possibilities for the first coin, 2 for the second..., so tossing n coins has

$$\overbrace{2 \times 2 \times \cdots \times 2}^{n \text{ times}} = 2^n \text{ possibilities.}$$

4

The coin toss example, illustrates situations where repetitions are allowed: the second coin toss has the same two possibilities as the first. In other words, (H, H) is allowed, as is (T, T).

In some situations repetitions are not allowed.

You'll have to use that rare gift, common sense, to decide if repetitions are allowed or not.

5

(4) There are, say, 20 of you in the room. In how many ways can I distribute a red hat, a blue hat and a green hat among you? Nobody can wear two hats.

_____

_____

_____

6

ADDITIONAL NOTES

(5) Airport codes are three-capital-letter codes. The code for Boston's Logan airport is BOS, e.g. How many airports can the system represent?

7

In what follows, we'll need the _____

(6) What's $6!$? _____

(7) What's $1!$? ___

(8) What's $0!$? _____

8

(9) If I wanted to shake hand with all 30 of you, in how many different ways could I do it? ____

(10) What is $\frac{9!}{6!}$?

9

Permutations

We've been discussing "permutations."

_____

_____

_____

The symbol _____ represents the number of permutations of r objects selected from n objects.

10

The formula for ${}_nP_r$ is

(11) What's ${}_7P_3$?

11

You use the formula

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

to calculate the answer to questions _____

_____

_____

12

ADDITIONAL NOTES

(12) If I were to rank the top three students in a class of 30, how many ways could I do it?

(13) If I were to pick my best and second-best shirts from my wardrobe of five, how many ways could I do it?

13

Combinations

If the order isn't important, it's a combination.

The symbol _____ represents the number of combinations of r objects selected from n objects.

14

The formula for ${}_nC_r$ is

(14) What's ${}_7C_3$?

15

You can use the formula

$$ {}_nC_r = \frac{n!}{r!(n-r)!} $$

to calculate the answer to questions _____

16

(15) If I were to select three students in a class of 30 to be my little helpers, how many ways could I do it?

17

Permutations with non-distinct objects

where, clearly, $n =$ _____

18

ADDITIONAL NOTES

(16) How many different sequences can be formed using two A's, two C's, three G's, and one T?

19

(17) What formula do you use for the number of possibilities if you're selecting r objects out of n, with repetition allowed?

20

(18) What formula do you use for the number of possibilities if you're selecting r objects out of n distinct ones, where the order is important and repetition is not allowed?

21

(19) What formula do you use for the number of possibilities if you're selecting r objects out of n distinct ones, where the order is not important and repetition is not allowed?

22

(20) What formula do you use for the number of possibilities if you have n objects of which n_1 are of one kind, n_2 of another, $\ldots n_k$ of the last, and the objects of each type are indistinguishable from each other?

23

(21) What has this to do with probability?

24

ADDITIONAL NOTES

(22) Consider a lottery where there are balls numbered 1 to 52. Six balls are randomly chosen without replacement. If you choose six numbers, what is your probability of winning the lottery? The order does not matter.

25

Random Variables

26

For example, if you toss a coin twice, the total number of heads you get is a random variable.
(23) What are its possible values?

27

As another example, suppose I measure the time between arrivals of successive students in class. The time between arrivals is a random variable.
(24) What are its possible values?

28

(25) One of these is a discrete random variable, the other a continuous random variable. Which be which?

29

Discrete Probability Distributions

30

ADDITIONAL NOTES

Example: let X be the number of heads in a two-coin toss. You can tabulate the situation in the table on the right, where x represents the possible outcomes.

x	$P(x)$
0	
1	
2	

(26) Fill in the probabilities in col. 2.

What do they add to? ___

31

Rules for a Discrete Probability Distribution

Let $P(x)$ denote the probability that the random variable X has the value x; then

- _____ $\leqslant P(x) \leqslant$ _____
- $\sum P(x) =$ _____

32

Problem Which of the following is a discrete probability distribution?

(a)

x	$P(x)$
1	0.20
2	0.35
3	0.12
4	0.40
5	−0.07

(b)

x	$P(x)$
1	0.20
2	0.25
3	0.10
4	0.14
5	0.49

(c)

x	$P(x)$
1	0.20
2	0.25
3	0.10
4	0.14
5	0.31

33

Mean of a Discrete Random Variable

The mean of a discrete random variable, X is

where x represents the values of the variable, and $P(x)$ the probability of getting these values.

34

(27) What is the mean of X (the number of heads) in our two-coin toss example?

35

(28) What does it mean that the mean is 1 in our example?

_____

_____

_____

_____

36

ADDITIONAL NOTES

▷ Interpretation of the Mean of a Discrete Random Variable

Suppose an experiment is repeated n times and the value of the random variable X is recorded. As the number of repetitions of the experiment increases, the mean value of the n trials will approach μ_X, the mean of the random variable X.

37

If x_1 is the value of the random variable X after the first experiment, x_2 the value after the second, etc., then

$$\bar{x} =$$

The difference between $\bar{x}$ and μ_X gets closer to 0 as n _____

38

The mean of a discrete random variable, X, represents what we _____

_____

Therefore, it's often called the _____ of X, and denoted by _____

39

Example:
Suppose someone buys a \$250,000 life insurance policy with an annual premium of \$350. Suppose that the 1-year survival rate for that person's category (age, gender, etc.) is 0.998937.

40

(29) If the person dies during that year, what will he/she win or lose? What will the insurance company win or lose?
The person _____

_____

The company _____

_____

41

(30) If the person does not die during that year, what will he/she win or lose? What will the insurance company win or lose?
The person _____
The company _____

42

ADDITIONAL NOTES

(31) What is the probability that the company gains $350?

(32) What is the probability that the company loses $249,650?

43

(33) Let X be the amount of money the company makes. What is its expected value?

$E(X) =$

$=$

$=$

(34) What does that number represent?

44

Standard Deviation of a Discrete Random Variable

The standard deviation of a discrete random variable X is

45

Complete this table for our two-coin toss example and calculate σ_X (X is the number of heads). $\mu_X = 1$.

x	$P(x)$	$(x - \mu_X)^2 \cdot P(x)$
0	1/4	
1	1/2	
2	1/4	

46

Now $\sqrt{1/2} \approx$ _____

This says that if you toss two coins repeatedly, about 68% of the tosses will be between $0.3\,(= 1 - .7)$ and $1.7\,(= 1 + .7)$ heads.

Not meaningful here, but you get the idea.

47

Reminder: The variance is the square of the standard deviation (what you have before the square root in the formula for σ).

48

ADDITIONAL NOTES

ADDITIONAL NOTES

ADDITIONAL NOTES

▷ Reminder:
A random variable is a numerical measure of the outcome of a probability experiment. Its value is determined by chance.

Random variables are typically denoted using capital letters such as X.

Example: If you toss a coin 10 times, the number of heads you get is a random variable.

1

Keep in mind that if the random *variable* is X, then its *values* are conventionally denoted by x.

(**1**) If X is the number of heads in a ten-coin toss, what are its possible values, x?

_____

2

Binomial Distribution

_____

_____

_____

_____

The two outcomes are often referred to as success and failure .

3

▷ How to spot a binomial experiment:

1. _____

_____

(Each repetition is called a trial.)

2. _____
The outcome of one trial will not affect the outcome of the others.

4

3. _____

_____
They are called "success" and "failure." If p is the probability of success, then $1-p$ is the probability of failure.

4. _____

_____

5

The number of successes in a binomial experiment with n trials, X, is called a _____

_____

(**2**) What are the bounds on the values of X?

_____

6

ADDITIONAL NOTES

For the next two questions, decide if the given probability experiment is a binomial experiment.

If it is, identify the

(i) number of trials,

(ii) probabilities of success and failure, and

(iii) values of the random variable X.

7

(3) In random sample of 10 people, the number with blood type O-negative is recorded. (Studies show that 7% of people in the United States have blood type O-negative.)

8

1. _____
2. _____
3. _____
4. _____

9

Before we move on, are the trials in the previous problem truly independent?

The population of the U.S. today is $\sim 326{,}000{,}000$. If 7% are O-neg, that's $\sim 22{,}820000$. *If* the sample were too large, say the whole population, the probability of finding a subsequent O-neg would vary slightly with each previous O-neg find.

But the effect is negligible for small samples.

10

(4) A probability experiment in which three cards are drawn from a deck without replacement and the number of aces is recorded.

11

1. _____
2. _____

12

ADDITIONAL NOTES

We enumerate all possibilities in a binomial probability distribution exactly as we did for coin tosses ("heads" now is "success", and "tails" failure – or the other way around).

For example, on the right are the possibilities for a three-trial binomial probability experiment.

(S, S, S)
(S, S, F)
(S, F, S)
(S, F, F)
(F, S, S)
(F, S, F)
(F, F, S)
(F, F, F)

13

(5) There's one significant difference between calculating probabilities in a general binomial experiment and in a coin-toss. What is it?

14

If the probability of getting an O-neg is 0.07, what's the probability of getting

(6) three O-negs in a 3-trial experiment?

(7) no O-negs in a 3-trial experiment?

15

(8) What is the probability of getting exactly two O-negs in a 3-trial experiment?

16

There's a formula you can use that saves your having to correctly list all possibilities:

The probability of obtaining x successes in n independent trials of a binomial experiment is

17

(9) Apply this to the previous question.

18

ADDITIONAL NOTES

The next four questions are based on this scenario:

Suppose 25% of all U.S. households are wireless-only (no landline). We collect data from a random sample of 20 households.

(**10**) What's the probability that exactly 5 are wireless-only? _____

19

(**11**) What's the probability that fewer than 3 are wireless-only?

_____

20

(**12**) What is the probability that at least 2 are wireless-only?

_____

_____

_____

21

(**13**) What is the probability that the number of households that are wireless-only is between 5 and 7, inclusive?

_____

22

A binomial experiment with n independent trials and a probability of success p has a mean

and a standard deviation

23

(**14**) If 25% of households are wireless-only and 500 are sampled, find and interpret the mean and s.d. of the number of wireless-only households.

Mean:

Standard deviation:

Interpretation: _____

_____

_____

24

ADDITIONAL NOTES

A binomial probability distribution becomes symmetrical ("bell shaped") if either $p \approx 0.5$ or n becomes very large.

A rough rule of thumb is that we need

$$np(1-p) \geqslant 10$$

25

Suppose that our survey of 500 households found 90 to be wireless-only. Would this be statistically unusual (under 5% chance)?

First, is the distribution bell-shaped?

Yes, because we've seen that

$$np(1-p) = 93.75 > 10.$$

26

Therefore we expect 95% of surveys to give a result that's within $2\sigma = 2(9.7) = 19.4$ of the mean: i.e., between $125 - 19.4 = 106.6$ and $125 + 19.4 = 144.4$.

Our survey gives a result outside this range, and is therefore in the 5% of surveys that do this. It's statistically unusual, and needs explanation.

27

Continuous Probability Distributions

We've looked at discrete probability distributions.

Now look at continuous ones, concentrating on 2:

1) _____

and

2) _____

28

Calculating individual probabilities from continuous distributions is tricky.

(**15**) Is your arrival time to a 155-minute class a continuous or discrete variable? _____

(**16**) How many possible values does your arrival time have? _____

(**17**) What is the probability that you arrive at exactly 6:00 p.m.? _____

29

We use a _____ (pdf) to compute probabilities for continuous distributions.

This is a function with two properties:

1. _____

_____

2. _____

_____

30

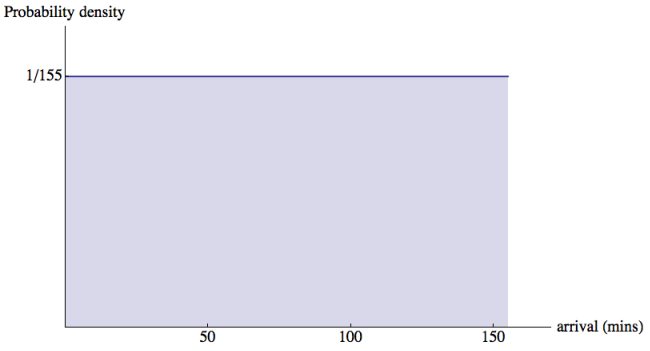
ADDITIONAL NOTES

(18) Do these remind you of anything?

1. _____
2. _____

31

Example: if all arrival times were equally likely, the pdf for this scenario would be graphed like this:



32

This is an _____

(19) Why must the height be $1/155$?

33

How do you use pdfs to calculate probabilities?

The area under the graph of a density function over an interval represents the probability of observing a value of the random variable *in that interval.*

34

(20) The probability that you arrive *at* 6:00 p.m. is zero. What is the probability that you arrive between 6:00 and 6:05 p.m?

_____

35

A uniform probability distribution may be thought of as one whose density function has a graph shaped like a rectangle.

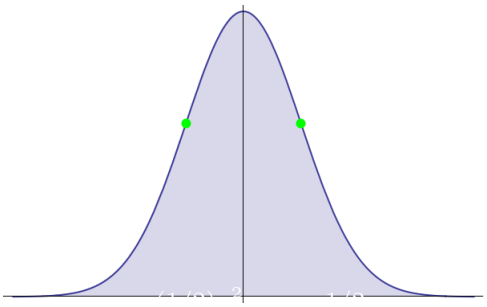
A _____ is one whose density function has a graph shaped like a normal curve.

Rectangles we can spot by their good looks. How do we spot a normal curve?

36

ADDITIONAL NOTES

Looks like this:



37

Properties of the Normal Density Curve

1. It is _____, μ.

2. Its _____; the curve has a single _____.

3. It has _____ (dots on previous graph).

38

4. The _____.

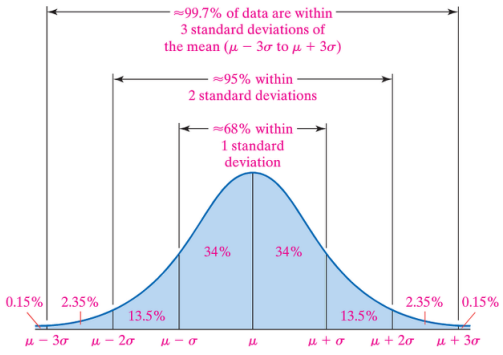
5. From the symmetry, _____ _____. Both are $1/2$.

6. As x gets larger and larger, the graph approaches, but never reaches, the horizontal axis. As x gets more and more negative, the graph approaches, but never reaches, the horizontal axis.

39

7. The Empirical Rule holds: Approximately 68% of the area under the normal curve is between $x = \mu - \sigma$ and $x = \mu + \sigma$; approximately 95% of the area is between $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$; approximately 99.7% of the area is between $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$.

40



41

ADDITIONAL NOTES

ADDITIONAL NOTES

The Normal Distribution (Continued)

▷ Area under a Normal Curve

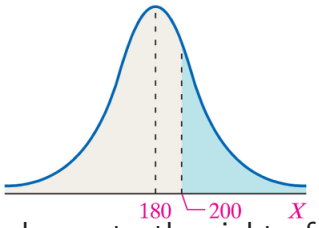
Suppose that a random variable X is normally distributed with mean μ and standard deviation σ.

1

This interval, equivalently, represents the proportion of the population that has those values.

2

Example: The cholesterol for males 20 to 29 years old is approximately normally distributed with mean $\mu = 180$ and $\sigma = 36.2$.



(1) If the size of the shaded area to the right of 200 is 0.2903, what does that mean?

3

Getting rectangular areas was easy. How do you get the area under part of a normal curve?

First, standardize the normal curve by switching to z-scores.

(2) What is the formula for the z-score of a random variable X?

4

(3) Looking at all the z-scores for a given random variable X, some will be positive and some negative, irrespective of the signs of the original values of X. Why?

5

(4) It follows that the sum of the z-scores always comes out to a fixed number. What is it?

(5) So, what's the mean of the z-scores?

(6) Any guesses on the standard deviation of the z-scores?

6

ADDITIONAL NOTES

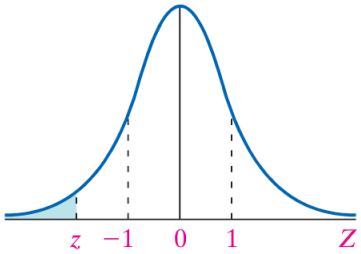
Using its z-scores instead of the original values of a normal random variable X gives a _____ _____.

For such a distribution, _____.

So, all standard distributions have the same graph, irrespective of the normal variable from which the z-scores arose.

7

This allows you to get areas from a table:



Given a particular value z, read the area to the left of it from the standard normal distribution area table.

8

How does getting the area under the standard normal distribution curve help us with the area under the curve of the actual data?

Suppose a data value x has a z-score of z.

It can be shown that _____ _____.

9

For example, IQ scores are normally distributed with an (adjusted) mean of 100, and a s.d. of 15.

(7) What's the probability that a randomly selected person will have an IQ under 135?

_____

_____

10

(8) What's the probability that a randomly selected person will have an IQ over 110?

_____.

_____.

11

Sometimes we want the reverse:

Example:
Suppose the heights of 3-year-old females are approximately normally distributed, with a mean of 38.72 inches and a standard deviation of 3.17 inches. Find the height of a 3-year-old female at the 20th percentile.

12

ADDITIONAL NOTES

(**9**) What does 20th percentile mean?

(**10**) Look up 0.2 in the *area part* of the table and read the z-score.

13

(**11**) Use that z-score and solve for x.

14

Notation: We use z_α to represent the z-score such that the area under the standard normal curve to the right of z_α is α.

For example $z_{0.2}$ is the z-score such that the area to the right of it is 0.2

15

(**12**) What is its value?

16

Testing Normality

How can we tell if the data we have is normally distributed?

We plot the observed values against the z-scores we expect.

How?

17

These are the steps:

1. Arrange data in ascending order.
2. Calculate $f_i = \dfrac{i - 0.375}{n + 0.25}$

where i is the position and n the total number of observations.

3. Find the z-scores that match the f_i.
4. Plot z vs. observed values.

18

ADDITIONAL NOTES

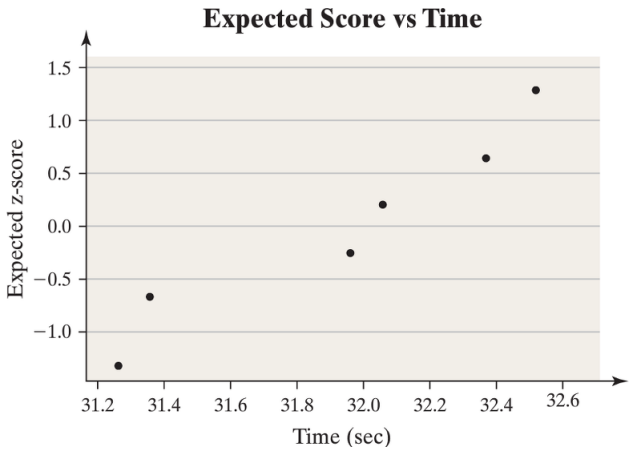
Example:

31.35	32.52
32.06	31.26
31.91	32.37

19

Index, i	Observed Value	f_i	Expected z-score
1	31.26	$\frac{1-0.375}{6+0.25} = 0.10$	−1.28
2	31.35	$\frac{2-0.375}{6+0.25} = 0.26$	−0.64
3	31.91	0.42	−0.20
4	32.06	0.58	0.20
5	32.37	0.74	0.64
6	32.52	0.90	1.28

20



21

Binomial and Normal Distributions: The Connection

▷ The Normal Approximation to the Binomial Probability Distribution

22

Sampling Distributions

When you want to get information on something such as household income, you typically survey a random sample of some size, n. From it you get quantities of interest such as the mean, $\bar{x}$.

What if you conduct a different sample of the same size?

23

You expect to get a _____ as you repeat the survey.

That leads to the idea of a sampling distribution.

24

ADDITIONAL NOTES

The _____ is a probability distribution for all possible values of the statistic computed from a sample of size n.

The sampling distribution of the sample mean, $\bar{x}$, is the _____ _____ _____.

25

Mean and Std. Dev. of the Sampling Distribution of x

Suppose that a simple random sample of size n is drawn from a population with mean μ and standard deviation σ. The sampling distribution of $\bar{x}$ has mean

and standard deviation

26

The standard deviation of the sampling distribution of $\bar{x}$, $\sigma_{\bar{x}}$, is called the _____ _____.

27

Example: The IQ, X, of humans is approximately normally distributed with mean $\mu = 100$ and standard deviation $\sigma = 15$. What is the probability that a simple random sample of size $n = 10$ results in a sample mean greater than 110. That is, compute $P(\bar{x} > 110)$.

28

(13) What are $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$?

$\mu_{\bar{x}} =$ _____.

$\sigma_{\bar{x}} =$.

29

(14) Convert $\bar{x} = 110$ to a z-score.

30

ADDITIONAL NOTES

(15) Look up the "area to the left" from the table, then convert to an "area to the right." That's the desired probability.

_____.

_____

_____

31

(16) If we take a hundred samples of size ten, how many do we roughly expect to yield a mean IQ greater than 110? _____

32

The Central Limit Theorem

_____

_____

_____

33

ADDITIONAL NOTES

ADDITIONAL NOTES

ADDITIONAL NOTES

Sampling Distributions

When you want to get information on something, such as household income or the proportion of households that might have a certain opinion, you typically survey a random sample of some size, n. From it you get quantities of interest such as the sample mean, $\bar{x}$, or the sample proportion, $\hat{p}$.

1

If you survey a different sample of the same size you expect different values of these quantities.

Repeating the surveys many times, you get a sampling distribution of these variables.

The sampling distribution of a statistic is a probability distribution for all possible values of the statistic computed from a sample of size n.

2

Distribution of the Sample Mean

The sampling distribution of the sample mean $\bar{x}$ is the probability distribution of all possible values of the random variable $\bar{x}$ computed from a sample of size n from a population with mean μ and standard deviation σ.

3

Mean and S.D. of Sampling Distribution of $\bar{x}$

Suppose that a simple random sample of size n is drawn from a population with mean μ and standard deviation σ. The sampling distribution of $\bar{x}$ has mean and standard deviation given by

and

4 _____, is called the standard error of the mean.

The Central Limit Theorem

Regardless of the shape of the underlying population, the sampling distribution of $\bar{x}$ becomes approximately normal as the sample size, n, increases.

5

Distribution of the Sample Proportion

Suppose that a random sample of size n is obtained from a population in which each individual either does or does not have a certain characteristic. The sample proportion, denoted $\hat{p}$ ("p-hat"), is given by

where x is the number of individuals in the sample with the specified characteristic.

6

ADDITIONAL NOTES

.

For a simple random sample of size n with a population proportion p, the shape of the sampling distribution of $\hat{p}$ is approximately normal provided

$$np(1-p) \geqslant 10.$$

7

Mean and S.D. of Sampling Distribution of $\hat{p}$

Mean:

Standard deviation:

8

A survey reveals that 76% of Americans believe that the state of moral values in the United States is getting worse.

Look at a simple random sample of $n = 60$ Americans and describe the sampling distribution of the sample proportion for Americans with this belief.

9

(**1**) Check that $np(1-p) \geqslant 10$.

(**2**) What are $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$?

10

Where are these concepts used? Here are statements one might see:

"24% of all voters believe the scientific debate about global warming is over. . . The survey of 1,000 Likely Voters was conducted on November 9-10, 2015 by Rasmussen Reports. The margin of sampling error is +/- 3 percentage points with a 95% level of confidence."

11

"Results for this Gallup poll are based on telephone interviews conducted July 8-21, 2015, on the Gallup U.S. Daily survey, with a random sample of 2,374 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is 2 percentage points at the 95% confidence level."

12

ADDITIONAL NOTES

We'll study both these concepts:

- _____ and
- _____

This is part of _____; i.e., extending information from a sample to a population. One area of inferential statistics is _____: sample data are used to estimate the value of parameters such as μ or p.

13

Making Estimates

A _____ is the value of a statistic that estimates the value of a parameter.

For example, the point estimate for the population proportion is $\hat{p} = x/n$, where x is the number of individuals in the sample with the specified characteristic and n is the sample size.

14

Confidence

A _____ for a parameter consists of an interval of numbers based on a point estimate.

The _____ represents the expected proportion of intervals that will contain the parameter if a large number of samples is obtained. The level of confidence is denoted

_____

15

(3) If the level of confidence is 95%, $\alpha = ?$

(4) What's $z_{0.025}$? _____

(5) What does it mean? _____

_____

_____

(6) What percentage of the data will lie between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$? _____

16

Let's apply this to a study of population proportions:

Let's say a sample reveals that a certain proportion of the population, $\hat{p}$, has some opinion (support for a political candidate, for example).

Different samples will yield different proportions, and given enough samples, we will have a distribution of proportions.

17

Assuming that the conditions are met for the distribution of sample proportions to be normal, we know that $\mu_{\hat{p}} = p$.

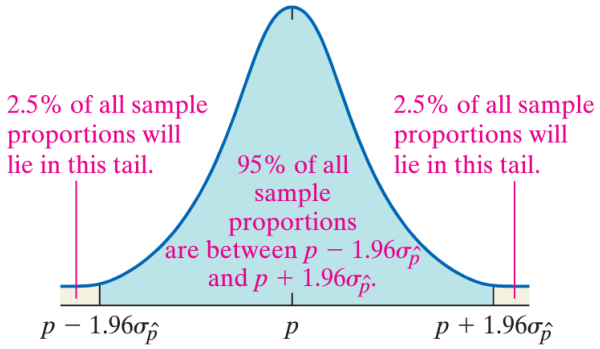
In words: the mean of the different proportions yielded by different samples will be the actual proportion of the population, p, that has that opinion.

How many of the sample proportions will lie a given "distance" from the mean, p?

18

ADDITIONAL NOTES

What we've seen, for example, is this:



19

A single sample gives you a point estimate $\hat{p}$.

What the previous diagram illustrates is that in 95% of the cases we have

$$p - 1.96\sigma_{\hat{p}} \ < \ \hat{p} \ < \ p + 1.96\sigma_{\hat{p}}.$$

(7) Negate that inequality:

20

(8) Add $p + \hat{p}$ to each side.

21

In words:

95% of the time we expect the (unknown) "actual" proportion p to lie within

$$\hat{p} \pm 1.96\sigma_{\hat{p}}$$

or, using the language of error, within

$$\hat{p} \pm 1.96(\text{standard error}).$$

The last term is called the _____.

22

This is generally expressed as follows:

We have a margin of error of $\pm 1.96\sigma_{\hat{p}}$ at the 95% confidence level.

(9) What would the margin of error be at the 99% confidence level? _____

23

Let's return to one of our examples:

"24% of all voters believe the scientific debate about global warming is over. . . The survey of 1,000 Likely Voters was conducted on November 9-10, 2015 by Rasmussen Reports. The margin of sampling error is +/- 3 percentage points with a 95% level of confidence."

24

ADDITIONAL NOTES

(**10**) What does this mean?

25

We've expressed the margin of error for a given confidence interval as a multiple of $\sigma_{\hat{p}}$.

We can estimate what that is, using

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Then, the $(1-\alpha)\cdot 100\%$ confidence interval for a point estimate $\hat{p}$ is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

26

Example:

If 272 of 800 teens surveyed say they text while driving, what is the 95% confidence interval for the proportion of teens who text while driving?

27

(**11**) Find the point estimate of the proportion of teens who text while driving.

(**12**) Is the underlying distribution likely to be normal? _____

(**13**) At the 95% confidence level, what is α and, therefore, $z_{\alpha/2}$? _____

28

(**14**) What is $\sigma_{\hat{p}}$?

(**15**) What is the margin of error, as a whole number percentage? _____.

29

(**16**) Put it all together, you little Rasmussens.

30

ADDITIONAL NOTES

We see that the margin of error is

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(**17**) If we increase the sample size, n, does it increase or decrease the margin of error? _____

(**18**) How much would it have to increase by to halve the margin of error? _____

31

ADDITIONAL NOTES

_____ is a procedure, based on sample evidence and probability, used to test hypotheses.

1

Steps in Hypothesis Testing:

1. _____

2. _____

3. _____

2

The _____, H_1 ("H-one"), is a statement that we are trying to find evidence to support.

3

Setting up null and alternative hypotheses:

1. Equal hypothesis versus not equal hypothesis (two-tailed test):
H_0: parameter has some value;
H_1: parameter does not have that value.

2. Equal versus less than (left-tailed test):
H_0: parameter has some value;
H_1: parameter has smaller value.

4

3. Equal versus greater than (right-tailed test):
H_0: parameter has some value;
H_1: parameter has greater value.

In each of the next three questions, determine the null and alternative hypotheses and what type of test is needed to try and establish the alternative.

5

(**1**) The Blue Book value of a certain used car is \$56,130. You wonder if your car is worth a different amount.

6

ADDITIONAL NOTES

(2) The standard deviation of the contents in a 64-ounce bottle of detergent using an old filling machine is 0.23 ounce. The manufacturer wants to know if a new filling machine has less variability.

7

(3) A company has a new antibiotic for children. Two percent of children taking competing antibiotics experience headaches. The FDA wants to know if the percentage of children who take the new antibiotic and experience headaches is $> 2\%$.

8

The Four Outcomes of Hypothesis Testing

		Reality	
		H_0 Is True	H_1 Is True
Conclusion	Do Not Reject H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

9

Example:

Suppose LIU wants to know if a majority of students favor a dress code on campus.

They conduct a survey of a sample of 100 students. The sample survey reveals that 53 students want a dress code ($\hat{p} = 0.53$).

10

(4) What are the null and alternative hypotheses?

11

(5) Get $\sigma_{\hat{p}}$.

12

ADDITIONAL NOTES

(6) Get the z-score for $\hat{p} = 53$.

13

The χ^2 Test

This is a commonly used method to test the validity of the null hypotheses.

Example: Suppose you roll three dice 100 times, and you get no sixes 48 times, one six 35 times, two sixes 15 times and three sixes 3 times.

Are these statistically what you'd expect from fair dice?

14

(7) What is the probability of obtaining x successes in n independent trials of a binomial experiment?

15

(8) What are the probabilities of getting no sixes, one six, two sixes and three sixes when you roll three dice?

16

(9) Convert to numbers you would expect after 100 throws of three dice.

17

Tabulate these against the observed results:

	Observed.	Calculated.
No One Two Three		

18

ADDITIONAL NOTES

The calculated values are the expected values. χ^2 is defined as

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

(**10**) Calculate χ^2 for our example.

19

We have thrown three dice. That's a system with "three degrees of freedom." Looking at the appropriate line in a χ^2 table, we see that $\chi^2 = 23.37$ lies well to the right of 0.005 and we can reject the null hypothesis that the dice are fair.

TABLE VII

Chi-Square (χ^2) Distribution
Area to the Right of Critical Value

Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

20

ADDITIONAL NOTES

ADDITIONAL NOTES

ADDITIONAL NOTES

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

HW ____ NAME: _____

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997