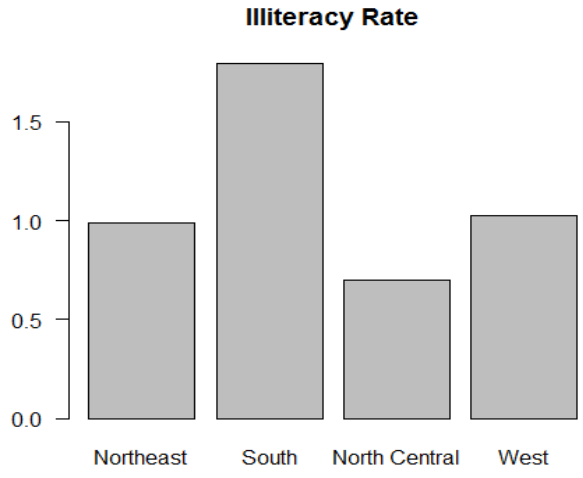
Review of Statistical Concepts

One basic application of statistics is to simply describe a dataset. Consider, for example, the data given in Table 1. These values represent characteristics of the fifty United States in the mid 1970's and were compiled from census information. Thus the values in Table 1 are based (at least theoretically) on all residents of the United States at the time the census was taken. We can summarize these values to obtain a "snapshot" of U.S. residents. Typically, we use means and standard deviations to summarize numeric data, although graphs (pie charts, bar charts, etc.) are also useful.

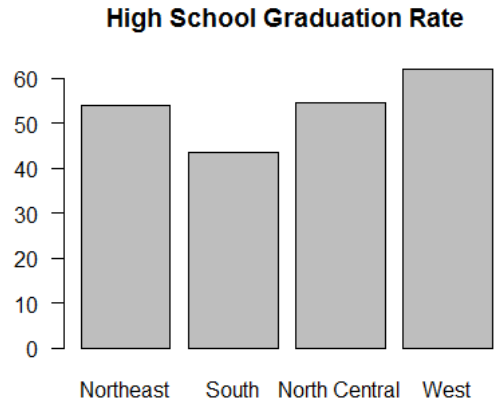
For the illiteracy rate, we calculate the average (across all 50 states) is 1.17 percent. This tells us that 1.17% of all U.S. residents are illiterate. If we incorporate the geographic region into our calculations, we find the following percentages: 0.99% of residents in the Northeast, 1.79% of residents in the South, 0.70% of residents in the North Central and 1.02% of residents in the West. These values could be presented in a bar graph or a small table.

Region	Illiteracy Rate
Northeast	0.99%
South	1.79%
North Central	0.70%
West	1.02%

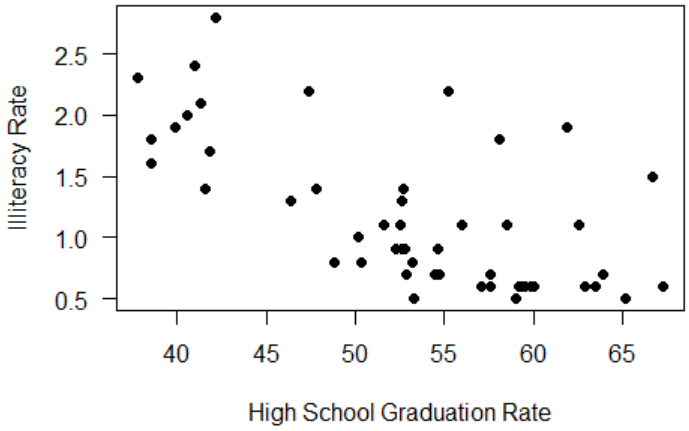


Clearly, states in the South have a higher illiteracy rate than the other regions. To explore this phenomenon, we can examine the high school graduation rates in the four regions.

As expected, the South has the lowest average high school graduation rate, in addition to the highest average illiteracy rate.



Since both illiteracy rate and high school graduation rate are numeric variables, we can directly compare them by generating a scatterplot. This is simply a plot of the (X, Y) pairs for each state.



(Note: We could use different symbols and/or colors to represent the four regions.)

The scatterplot shows that states with a small value for high school graduation rate tend to have a large value for illiteracy rate, and states with small graduation rates tend to have high illiteracy rates. This can be summarized by the correlation coefficient, which is –0.66. The negative part of the correlation coefficient tells us that large values for one variable are associated with small values for the other variable. Values for correlation are always between –1 and +1, and more extreme values (closer to –1 or +1) indicate stronger relationships between the two variables. In this case, a correlation of –0.66 indicates a moderately strong negative correlation between high school graduation rates and illiteracy rates. This is an intuitive result, since socioeconomic conditions that produce high illiteracy rates are also likely to produce low graduation rates.

Population vs. Sample

In the preceding analysis, one key point to recognize is that the data represent the entire U.S. population (in the mid-1970's), so we can describe this population simply by summarizing the data. In general, this is NOT the situation in which statistics are applied. We usually have a sample of data that has been taken from some larger population, and we want to describe the population based on the sample data.

Two key definitions:

- Population: all the items we are interested in studying
- Sample: the items from which we actually obtain data

In an ideal world, the items in the sample are randomly selected from the population. In practice, this is a very difficult thing to accomplish. For the data we will analyze in this course, we will assume that the items in the sample are randomly selected. More details about this process can be found in a course on experimental design.

When we are dealing with sample data, we are usually interested in making decisions regarding the population based on what we observe in the sample.

Example: Fat Free Yogurt

A particular brand of yogurt is advertised to be 98% fat free (i.e., only 2% fat), but a consumer watchdog group suspects that fat content is labeled incorrectly. The group will take action against the company if it can substantiate its suspicion with factual data. For this purpose, the group took a sample of 25 yogurt containers (each containing 170 grams) and measured the fat content of each container. If the company's claim is correct, then the mean fat content should not be different from 2%, i.e., 3.4 grams. For the 25 yogurt cups that were examined, the mean fat content was 3.7 grams with standard deviation 0.5 grams. Is there enough statistical evidence to support the consumer group's suspicion?

Population vs. Sample: The sample consists of the fat content of the 25 containers that were measured. The population is the fat content of all the containers of 98% fat free yogurt produced by this company. Presumably, the 25 containers were randomly selected, so they comprise a random sample of size n = 25 from the population.

The **population parameters** are μ and σ

μ = mean fat content of all yogurt from this company that is advertised as 98% fat free

σ = standard deviation of fat content of all yogurt from this company that is advertised as 98% fat free

These values are NOT known, but we can use the sample data to get estimates. In general, we use Greek letters to represent population parameters.

Sample statistics are $\bar{x}$ and s

$\bar{x}$ = mean fat content of the 25 containers

s = standard deviation of fat content of the 25 containers

These values ARE known. In this example, they are given in the statement of the problem ($\bar{x}$ = 3.7 and s = 0.5). If they are not given, you may need to use software to calculate them from the data that will be provided.

Point estimates: $\bar{x}$ is a point estimate for μ and s is a point estimate for σ. It is typical to use a "hat" to indicate an estimate, so that $\hat{\mu} = \bar{x}$ = 3.7 and $\hat{\sigma} = s$ = 0.5. Note that these are single values that estimate unknown population parameters. So we know that 3.7 grams is a reasonable value for the true mean fat content (μ) of all containers of 98% fat free yogurt produced by this company, but we do not know if 3.2 grams is a reasonable value, or if 3.69 grams is a reasonable value. At this point, all we know is that 3.7 grams is a reasonable value.

Interval estimate: A confidence interval for μ provides a range of plausible values for the unknown population mean. Every confidence interval has the form

(point estimate) $\pm$ (margin of error).

In addition, every confidence interval has a confidence coefficient, usually 95%, which indicates how sure we are that the confidence interval contains the true value of the parameter.

The point estimate was defined earlier ($\bar{x}$ is a point estimate for μ). The margin of error is the product (multiplication) of two parts: the standard error of the point estimate and the critical value.

For this example, the standard error of the mean is $\frac{s}{\sqrt{n}} = \frac{0.5}{\sqrt{25}} = 0.1$

Since this is a one-sample confidence interval for μ, the critical value comes from the table of probabilities for the t distribution. To use the t table, we need to know the degrees of freedom (which is n–1 = 24) and the confidence coefficient (95%). For this example, the critical value is 2.064.

When we put it all together, a 95% confidence interval for μ is

$$\text{(point estimate)} \pm \text{(margin of error)}$$

$$\bar{x} \pm (\text{critical value}) \times (\text{standard error})$$

$$3.7 \pm 2.064 \times 0.1$$

$$3.7 \pm 0.2064$$

$$3.4936 \text{ to } 3.9064$$

(To make this easier to follow, I will round these to 3.5 and 3.9 grams.)

We are 95% confident that the true value for μ is between 3.5 grams and 3.9 grams. Note that the confidence interval gives us a range of plausible values for the unknown population mean. Since the confidence interval does not contain 3.4 grams, we can conclude that 3.4 grams is not a reasonable value for μ. This leads us to believe that the company's claim is incorrect for this yogurt.

Hypothesis Testing

Constructing a confidence interval is one method for assessing the company's claim, but a more direct approach is to conduct a hypothesis test. The basic steps of a hypothesis test are outlined below, using the yogurt example.

Step 1: Define the null and alternative hypotheses.

null hypothesis is	H0: $\mu = 3.4$
alternative hypothesis is	Ha: $\mu \neq 3.4$

Note that the hypotheses involve the unknown population parameter μ, and not the sample statistic $\bar{x}$. The null hypothesis is that the manufacturer's label is accurate, that is, the mean fat content of the yogurt really is 3.4 grams. The alternative hypothesis that that manufacturer's label is not accurate, that is, the mean fat content of the yogurt is not 3.4 grams. Note that either H0 or Ha must be true, and that H0 and Ha cannot both be true.

A basic premise of hypothesis testing is that we choose to believe H0 is true unless the sample provides sufficient evidence to demonstrate that, beyond a reasonable doubt, H0 is false. The focus of the investigation is on the null hypothesis, so the final result of the hypothesis test will be to either "reject H0" or "fail to reject H0". We never "reject Ha" or "fail to reject Ha".

Step 2: Select a significance level for the test.

The significance level is the probability we reject the null hypothesis when, in fact, the null hypothesis is true. For the yogurt example, the consequence of this mistake is that we would conclude the yogurt is labeled incorrectly, when it is actually labeled correctly. The significance level controls the probability we make this kind of mistake. The significance level is sometimes called the alpha level of the test. Typical values for α are 0.01, 0.05 and 0.10. We will always use α=0.05 unless a different value is explicitly given.

Step 3: Identify the test statistic.

The test statistic is a single number that combines the sample information with the value specified in the hypotheses. The formula for calculating the test statistic depends on the precise nature of the hypotheses being tested and the type of information available from the sample. For the current example, we are working with a one-sample test for a population mean, so the test statistic is

$$TS = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3.7 - 3.4}{0.5/\sqrt{25}} = 3.0$$

This follows a t distribution with n – 1 = 24 degrees of freedom.

Note: There are many different hypothesis tests, and each test could have a different formula for calculating the test statistic. There are also many different distributions associated with the various test statistics. The key points I want you to know are

1) there is a test statistic for every hypothesis test;
2) the test statistic combines values from the sample with values from the hypothesis; and
3) the test statistic follows some probability distribution, called a reference distribution.

For the most part, we will use software to calculate the test statistic, but there are a small number of tests that we will have to perform "by hand". I will give you the formulas you need for these tests.

Step 4 (a): Identify the p-value and make a decision.

The p-value is a probability (so it is always between 0 and 1), and it indicates the strength of the evidence against H0. For almost all of the hypothesis tests we will perform in this course, we will use software to calculate the p-value. When the p-value is available, we compare the p-value to alpha to decide whether or not to reject H0.

- If the p-value is less than α, then reject H0.
- If the p-value is greater than α, do not reject H0.

For the yogurt example, the p-value is 0.0062, which is less than α=0.05. So we reject H0 and conclude the yogurt labels are incorrect.

Step 4 (b): Identify the critical value and make a decision.

When the p-value cannot be generated by software, we will use the critical value approach (instead of the p-value approach) to perform the final step of the hypothesis test. You will need to use a probability table (which I will provide) to identify the critical value. Specific details for doing this will given. To conduct the test, compare the critical value to the test statistic.

- If the test statistic is greater than the critical value, then reject H0.
- If the test statistic is less than the critical value, do not reject H0.

For the yogurt example, the critical value is 2.064. The test statistic is 3.0, which is greater than 2.064, so we reject H0. Note that this is the same conclusion we reached using the p-value approach. The critical value approach and the p-value approach should always produce the same result -- they are just two different ways of arriving at the same conclusion.

Connection between Confidence Intervals and Hypothesis Tests

There is a direct relationship between two-sided hypothesis tests and confidence intervals, when both are conducted at the same level of significance. Recall that, for the yogurt example, the 95% confidence interval for μ is 3.5 to 3.9 grams.

In the hypothesis test, we were deciding between two competing hypotheses:

H0: $\mu = 3.4$
Ha: $\mu \neq 3.4$

This is a two-sided test because the alternative is "not equal". We performed this test at significance level $\alpha = 0.05 = 5\%$. The confidence level of the confidence interval is 95% and the significance level of the test is 5%. These add up to 100%, so we can use the confidence interval to perform the test.

Since the confidence interval does not contain 3.4, we conclude that 3.4 is not a reasonable value for μ. We therefore reject H0 and conclude the true mean fat content of the yogurt is something other than 3.4 grams.

Statistical Significance and the Truth

When we conduct a test in which we reject H0, we say the "result is significant" or the "test is significant". This does not mean that H0 is false. It does mean that the data in the sample would be extremely unusual if H0 were true. The p-value gives us a measure of how unusual our sample data would be, assuming H0 is true.

When we conduct a test in which we do not reject H0, we say the result is "not significant". This does not mean that H0 is true. It does mean that the sample data is consistent with H0 being true.

Key Elements of a Hypothesis Test

1. The hypotheses, H0 (the null) and Ha (the alternative)
2. The significance level (α)
3. The test statistic
4. The reference distribution
5. Either the critical value or the p-value (sometimes both)

When you report the results of a hypothesis test, state both the test statistic and the p-value. If you don't have the p-value, report the critical value instead. For example, the results of the yogurt example should be reported as

> "We conclude the mean fat content of this yogurt is not 3.4 grams (t=3.0, p=0.0062)."

The test statistic is identified as "t=3.0", which also identifies the reference distribution as a t distribution. Both the test statistic and the p-value are given in parentheses at the end of the sentence that clearly states the results of the test.

The software we will be using automatically generates dozens of hypothesis tests for each data set we encounter. It will be very easy to get lost among all these tests, so you must be able to recognize what hypotheses are being tested so you can find the test(s) you need to answer

specific questions. You will also need to convince me that you are looking at the correct test(s), and the best way to do that is to provide both the test statistic and the p-value.

Sample vs. Population, re-visited

When the dataset contains information about every item in the population, it is appropriate to summarize the data, but it is not appropriate to perform hypothesis tests and/or generate confidence intervals. These techniques are designed specifically to make inferences about a population based on information contained in a sample from that population. If the dataset contains information about every item in a population, then statistical inference is not warranted.

The dataset itself does not contain any information that you can use to determine if the data represents an entire population or if it is a sample from the population. To make this determination, you must have additional information about the dataset. This additional information can be as simple as a paragraph explaining how the data was collected, or it could be a full report that provides complete details of an experimental design.

Most of the time, we will be dealing with sample data, not population data. Ideally, the items in the sample will be randomly selected from the population, but it is not always clear what the population is. In the yogurt example, the sample of 25 containers was (presumably) randomly selected from among all the containers of 98% fat-free yogurt produced by this manufacturer, so the population is all containers of 98% fat-free yogurt produced by this manufacturer. This sounds simple enough, but the real world is anything but simple. If you think about this for a minute, you should be able to think of some complications. One of the questions that came to my mind is "what flavor is the yogurt?" If all 25 containers in the sample were blueberry-flavored, then the results apply to the populations of all 98% fat-free *blueberry* yogurt produced by this manufacturer. Then our result (that the label is incorrect) would apply only to blueberry yogurt, and we would have no information about the other flavors. I will never purposely try to confuse you about this. If I say a dataset contains 1000 randomly selected U.S. residents, then the population is all U.S. residents. I mention this now, at the beginning of our course, because it is perhaps the most abused/overlooked/misunderstood aspect of statistics.

	Abbr	Area	Pop.	Income	Illiteracy	LifeExp	HSGrad
Region = North Central							
Illinois	IL	56,400	11,197	5107	0.9	70.14	52.6
Indiana	IN	36,291	5,313	4458	0.7	70.88	52.9
Iowa	IA	56,290	2,861	4628	0.5	72.56	59.0
Kansas	KS	82,264	2,280	4669	0.6	72.58	59.9
Michigan	MI	58,216	9,111	4751	0.9	70.63	52.8
Minnesota	MN	84,068	3,921	4675	0.6	72.96	57.6
Missouri	MO	69,686	4,767	4254	0.8	70.69	48.8
Nebraska	NE	77,227	1,544	4508	0.6	72.60	59.3
North Dakota	ND	70,665	637	5087	0.8	72.78	50.3
Ohio	OH	41,222	10,735	4561	0.8	70.82	53.2
South Dakota	SD	77,047	681	4167	0.5	72.08	53.3
Wisconsin	WI	56,154	4,589	4468	0.7	72.48	54.5
Region = West							
Alaska	AK	589,757	365	6315	1.5	69.31	66.7
Arizona	AZ	113,909	2,212	4530	1.8	70.55	58.1
California	CA	158,693	21,198	5114	1.1	71.71	62.6
Colorado	CO	104,247	2,541	4884	0.7	72.06	63.9
Hawaii	HI	6,450	868	4963	1.9	73.60	61.9
Idaho	ID	83,557	813	4119	0.6	71.87	59.5
Montana	MT	147,138	746	4347	0.6	70.56	59.2
Nevada	NV	110,540	590	5149	0.5	69.03	65.2
New Mexico	NM	121,666	1,144	3601	2.2	70.32	55.2
Oregon	OR	96,981	2,284	4660	0.6	72.13	60.0
Utah	UT	84,916	1,203	4022	0.6	72.90	67.3
Washington	WA	68,192	3,559	4864	0.6	71.72	63.5
Wyoming	WY	97,914	376	4566	0.6	70.29	62.9
Region = Northeast							
Connecticut	CT	5,009	3,100	5348	1.1	72.48	56.0
Delaware	DE	2,057	579	4809	0.9	70.06	54.6
Maine	ME	33,215	1,058	3694	0.7	70.39	54.7
Massachusetts	MA	8,257	5,814	4755	1.1	71.83	58.5
New Hampshire	NH	9,304	812	4281	0.7	71.23	57.6
New Jersey	NJ	7,836	7,333	5237	1.1	70.93	52.5
New York	NY	49,576	18,076	4903	1.4	70.55	52.7
Pennsylvania	PA	45,333	11,860	4449	1.0	70.43	50.2
Rhode Island	RI	1,214	931	4558	1.3	71.90	46.4
Vermont	VT	9,609	472	3907	0.6	71.64	57.1
Region = South							
Alabama	AL	51,609	3,615	3624	2.1	69.05	41.3
Arkansas	AR	53,104	2,110	3378	1.9	70.66	39.9
Florida	FL	58,560	8,277	4815	1.3	70.66	52.6
Georgia	GA	58,876	4,931	4091	2.0	68.54	40.6
Kentucky	KY	40,395	3,387	3712	1.6	70.10	38.5
Louisiana	LA	48,523	3,806	3545	2.8	68.76	42.2
Maryland	MD	10,577	4,122	5299	0.9	70.22	52.3
Mississippi	MS	47,716	2,341	3098	2.4	68.09	41.0
North Carolina	NC	52,586	5,441	3875	1.8	69.21	38.5
Oklahoma	OK	69,919	2,715	3983	1.1	71.42	51.6
South Carolina	SC	31,055	2,816	3635	2.3	67.96	37.8
Tennessee	TN	42,244	4,173	3821	1.7	70.11	41.8
Texas	TX	267,339	12,237	4188	2.2	70.90	47.4
Virginia	VA	40,815	4,981	4701	1.4	70.08	47.8
West Virginia	WV	24,181	1,799	3617	1.4	69.48	41.6

Table 1. Characteristics of the 50 United States, circa 1975.

Area = Land area of the state, in square miles
Pop. = Population of the state, in thousands
Income = Per capita income, in dollars
Illiteracy = Illiteracy rate (percent of population)
LifeExp = Life Expectancy, in years
HSGrad = Percent of population graduated from high school