Outline

PART II: Inferences about population values (Chapter 6)
- Review on confidence intervals
- Hypothesis testing
- Statistical tests on averages for large samples
- Statistical tests on percentages
Problem: What is the objective of statistics?

**ANSWER:** to make inferences about a population on the basis of an observed sample.

Decisions and predictions about the future are often made on the basis of observations - which are often confusing and hard to interpret!

For example:
1. Predictions for future investments based on financial data on stock returns
2. Analysis of time spent browsing a certain web page based on the access logs for a web site.
3. Assessment of software reliability based on the number of failures of a software or data on the memory load of a program.
Statistics provide formal procedures to make reliable decisions and predictions from observed data!

We will discuss methods for analyzing data on measurement variables.

First, we’ll focus on inferences about the population average (central value).

The statistical methods will deal with one of these two problems:

I. Estimate the value of the average/percentage of the population of interest

II. Test an hypothesis on the value of the population average / proportion.
Some confidence intervals for the population average:

\( \bar{x} \) is the sample average of \( n \) observations in a simple random sample of size \( n \), where \( n \) is large (\( >30 \))

\( s \) is the sample standard deviation of the \( n \) observations.

The 90% C.I. for the population mean:   \[ \bar{x} \pm 1.64* \frac{s}{\sqrt{n}} \]

The 95% C.I. for the population mean:   \[ \bar{x} \pm 1.96* \frac{s}{\sqrt{n}} \]

The 99% C.I. for the population mean:   \[ \bar{x} \pm 2.57* \frac{s}{\sqrt{n}} \]

See Table D for other values of \( z^* \) and the corresponding \( C \) (confidence levels).
Consider the example of the web application for the flight search. The investigator takes a sample of 100 flight searches. These are the summary statistics calculated by SAS:

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>14.811286</td>
<td>4.88768616</td>
<td>3.43076</td>
<td>26.88196</td>
</tr>
</tbody>
</table>

The estimated processing time is 14.811 seconds (= sample average)
The standard error is equal to s/sqrt(n) = 4.887/sqrt(100)=0.489.
What is the 95% confidence interval for the time employed by the web application to search for a flight schedule?

We can use the normal approximation, since the sample size is large enough. The 95% confidence interval is constructed as

\[ \text{estimate} \pm 1.96 \times \text{S.E.} \]

The estimate is the sample average = 14.811 and the S.E. is 0.489

The 95% confidence interval is

\[ (14.811 - 1.96 \times 0.489, 14.811 + 1.96 \times 0.489) = (13.852, 15.769) \]

The average processing time is a value between 13.8 seconds and 15.8 seconds
(on the basis of a method that fails 5 out of 100 times)
SAS procedure for C.I.

PROC MEANS DATA = data-name N MEAN STD CLM
ALPHA=value MAXDEC = number;
VAR measurement-variable;
RUN;

Where ALPHA=value is the (1-confidence level) value. So Thus ALPHA=0.05 for a 95% C.I., ALPHA =0.1. for a 90% CI, ALPHA=0.01 for a 99% C.I.
CLM is the option for C.I.’s
MAXDEC = defines how many decimal numbers (typically 1 to 4)
SAS example

```sas
proc means data=dist n mean std clm alpha=0.05 maxdec=4 ;
var x;
title "C.I. for population average";
run;
```

The MEANS Procedure
Analysis Variable : x waiting time

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>14.8113</td>
<td>4.8877</td>
<td>13.8415</td>
<td>15.7811</td>
</tr>
</tbody>
</table>
Confidence intervals are used when the goal of our analysis is to estimate an unknown parameter in the population.

A second goal of a statistical analysis is to verify some claim about the population on the basis of the data.

A test of significance is a procedure to assess the truth about a hypothesis using the observed data. The results of the test are expressed in terms of a probability that measures how well the data support the hypothesis.

Consider the following example
Example

To determine whether the mean nicotine content of a brand of cigarettes is greater than the advertised value of 1.4 milligrams, a health advocacy group takes a sample of 500 cigarettes and measures the amount of nicotine in the sample.

The sample average of nicotine is computed as 1.51 and the standard deviation of the observations is 1.016.

The estimated amount of nicotine is 1.51, based on the sample values. The standard error of the estimate is

\[ S.E. = \frac{s.d.}{\sqrt{n}} = \frac{1.016}{\sqrt{500}} = 0.045 \]

Is the difference between the sample value and the advertised value real? Or is it due to chance error?

To answer this question we need a test of significance:

If the observed value is “too many” standard errors away from the expected (conjectured) value then it is hard to explain it by chance!
Statistical language: stating hypotheses

The null hypothesis expresses the idea that the observed difference is due to chance. It represents the statement being tested in a significance test. The null hypothesis is abbreviated as $H_0$.

In the example, $\mu$ denotes the “true” average amount of nicotine. The null hypothesis states that the advertised amount is true: $H_0 : \mu = 14\text{mlg}$

The alternative hypothesis represents the idea that the difference is real. It is expressed as a statement we hope or suspect is true instead of the null hypothesis. The alternative hypothesis is abbreviated as $H_a$.

In the example, the alternative hypothesis states that the cigarettes contain a higher amount of nicotine, that is: $H_a : \mu > 14\text{mlg}$

Notice that the alternative hypothesis is often what we want to prove, and the null hypothesis is what we want to find evidence against!
Test statistics and significance levels

In the example of the amount of nicotine, the difference between the observed value and the conjectured value is

\[
\frac{1.51 - 1.40}{0.045} = 2.20
\]

This says that the observed value is 2.20 S.E.’s away from the advertised value. This is an example of a test statistic.

A test statistic is used to measure the difference between the data and what is expected under the null hypothesis.

The test statistic used above is called z-statistic and its general form is

\[
z = \frac{\text{observed} - \text{expected}}{\text{S.E.}}.
\]
Z-tests

The tests based on the z-statistic are called z-tests.

\[ z = \frac{\text{observed} - \text{expected}}{\text{S.E.}} \]

The z-statistic says how many S.E.’s away an observed value is from its expected value if the null hypothesis were true.

In the example the z-statistic is \( z = 2.20 \). In large samples, we can use the standard normal curve to check the area on the right of \( z = 2.20 \).

\[ p = 1.14\% \]
The chance of getting a sample average 2.20 S.E.’s or more above its expected values is extremely small, it is 0.014 or 1.4 %.

The test shows that the observed sample cannot be explained by the assumed null hypothesis, so it is unlikely that the null hypothesis is true.
P-values

The computed chance 1.4% in the example is called observed significance level. It is often denoted with P (for probability) and is called \( p \text{-value} \).

We looked at the area on the right of the z-value, because we compare the observed value with other possible values that would be even farther away from the value we want to test about.

A test of significance finds the probability of getting a test statistic as extreme or more extreme than the actually observed one. The chance is computed on the basis that the null hypothesis is true. The smaller this chance is, the stronger the evidence against the null hypothesis is.

The procedure for finding the p-value depends also on the alternative hypothesis.
The calculation of the p-value is based on the assumption that the null hypothesis is true.

Roughly speaking, a small p-value (such as 0.05 or smaller) indicates that the sample results are very unlikely under the assumption of the null hypothesis. Hence a small p-value is strong evidence against the null hypothesis, since the null hypothesis does not provide a “good explanation” for the observed sample.
If the p-value is small, then the null hypothesis should not be accepted (or should be **rejected** using the statistical terminology).

If the p-value is large (say larger than 0.05 or 0.01), the sample results do not provide evidence against the null hypothesis, the null hypothesis cannot be rejected!

**Remark:** The p-value is NOT the probability of the null hypothesis being right in the light of the data!!

The p-value “measures” the likelihood that a sample such as the one obtained, will occur when the null hypothesis is actually true.
Making a test of significance

Follow these steps:

1. **Set up the null hypothesis** $H_0$ – the hypothesis you want to test.
2. **Set up the alternative hypothesis** $H_a$ – what we accept if $H_0$ is rejected.
3. **Pick a test statistic**, to measure the difference between the data and what is expected under the null hypothesis.
4. **Compute the observed significance level** $P$. This is the probability, calculated assuming that $H_0$ is true, of getting a test statistic as extreme or more extreme than the observed one in the direction of the alternative hypothesis.
5. **State a conclusion**. You could choose a **significance level** $\alpha$. If the $P$-value is less or equal than $\alpha$, you conclude that the null hypothesis can be rejected at level $\alpha$, otherwise you conclude that the data do not provide enough evidence to reject $H_0$. 
Significance levels

In common statistical terminology:

• If P is less than $\alpha=5\%$, the null hypothesis is rejected at 5% significance level and the test result is called “statistically significant”.

• If P is less than $\alpha=1\%$, the null hypothesis is rejected at 1% significance level and the test result is called “highly significant”.

Notice that the significance levels are very popular for reporting the test results.

However, it is better practice to summarize the test results reporting what test was used, the P-value and whether the test was “statistically significant” or “highly significant”.
Example: Age of commercial jets

A report in USA today (July 7, 1995) stated that the average age of commercial jets in the U.S. is 14 years. An executive of a large airline company selects a sample of 40 airplanes and finds that the average age of the planes is 11.8 years. The standard deviation of the sample is 2.7 years. Is it true that the average age of the planes in his company is less than the national average?

We need to test the hypothesis

Null hypothesis $H_0$: $\mu = 14$

against the

Alternative hypothesis $H_a$: $\mu < 14$

The z-statistic is

$$z = \frac{11.8 - 14}{2.7 \sqrt{40}} = -5.11$$

The value of the z-statistic says that the observed value is more than 5 standard errors away from the age of 14 years that we assumed in the null hypothesis. Thus if the hypothesis $H_0$ were true, the observed sample would be very unlikely.
Assuming that the sample is random, we can compute the approximate p-value of the significance test, using the normal approximation to the z-statistic. The p-value is equal to the area under the standard normal curve to the left of z. We can use the table of areas under the standard normal table (Table A).

P-value is computed as \((100-99.9999)/2\), so the p-value is about 0.

**Conclusion**: since the p-value is so small, the data provides strong evidence against the hypothesis that the average age of the planes if 14 years. We can therefore accept the alternative hypothesis that the average age of the planes in the company is less than 14 years.
Z-tests for a population average in large samples

Consider a significance test for the hypothesis that a population average is equal to a certain value \( \mu^* \). The sample has size \( n \). The sample average is \( \bar{x} \) and the standard deviation in the sample is \( s \).

Compute the z-statistic:

\[
z = \frac{\text{observed} - \text{expected}}{\text{S.E.}} = \frac{\text{observed} - \text{expected}}{s / \sqrt{n}}
\]

Where the expected value is the value assumed under the null hypothesis and the S.E. of the sample estimate is S.E.=\( s/\sqrt{n} \).

If the alternative hypothesis is

1. Ha: \( \mu > \mu^* \) then the p-value is equal to the area on the right of \( z \)

2. Ha: \( \mu < \mu^* \) then the p-value is equal to the area on the left of \( z \)

These are called one-sided hypotheses, because they state that the true value is larger (1) or smaller (2) than the hypothesized value in \( H_0 \).
Two-sided Z-test for a population average

Consider a significance test for the hypothesis that a population average is equal to a certain value $\mu^*$. The sample has size $n$. The sample average is $\bar{x}$ and the standard deviation in the sample is $s$.

Compute the z-statistic:

$$z = \frac{\text{observed} - \text{expected}}{S.E.}.$$ 

where the S.E. of the sample average, that is $S.E. = s/\sqrt{n}$ and the expected value is the value assumed in the null hypothesis.

If the alternative hypothesis is $\text{Ha: } \mu \neq \mu^*$

then the p-value is equal to the sum of the areas on the left of $-z$ and on the right of $z$
Another example: Baseballs

In previous tests, baseballs were dropped 24ft onto a concrete surface and they bounced an average of 92.84 in. In a test of a sample of 40 new balls, the bounce heights had an average of 92.67 in. and standard deviation of 1.79in (based on data from Brookhaven National Lab and USA Today). Use a 0.05 significance level to determine whether there is sufficient evidence to support the claim that the new balls have bounce heights with a mean different from 92.84in. Does it appear that the new baseballs are different?

\textbf{Significance test:}

Null hypothesis: $\mu=92.84$ \quad Alternative hypothesis: $\mu \neq 92.84$

Test statistic:

$$z = \frac{92.67 - 92.84}{1.79 / \sqrt{40}} = -0.6$$

\textit{The value of the z-statistic says that the observed value is less than 1 standard error away from the expected value under the null hypothesis.}

Significance level $\alpha = 0.05$
The approximate p-value of the z-statistic is computed as the sum of the areas under the standard normal below:

![Normal Distribution](image)

The area between $-0.6$ and $0.6$ is $45.15\%$. The p-value is equal to $100-45.15=54.85\%$ or $0.55$. Thus, assuming that the null hypothesis is true, there is about a $55\%$ chance to observe a sample like the one we observed.

**Conclusion:** The p-value $=0.55$ is not smaller that $0.05$. So the data does not provide enough evidence to reject the null hypothesis that the new balls bounce at an average height of $92.84\text{in}$. The test results suggest that the new balls bounce at the same height of the old balls, on average.
Average speed reduced?

Suppose, over a long period of time in which the posted speed limit was 65 mph, the average speed along a certain stretch of the expressway was well established as 63 mph. After the speed limit dropped to 55 mph, the average speed of cars in a sample of size 100 was found to be 61.4 mph and the standard deviation was 4.6 mph. Does the lower speed limit cause a reduction in the average speed limit, or could it be only due to sampling variability?

1. Null hypothesis? Ho: \( \mu = 63 \text{mph} \)

2. Alternative hypothesis? Ha: \( \mu < 63 \text{mph} \)

3. Test statistic? \( z = \frac{61.4 - 63}{4.6 / \sqrt{100}} = -3.48 \)

*The observed value is more than 3 S.E.’s away from the value assumed under the null hypothesis. It is hard to explain it by chance variability.*
The area under the standard normal between \(-3.48\) and 3.48 is 99.95.
The p-value is \((100-99.95)/2=0.025\)% or 0.00025.
The small p-value indicates that the data provides very strong evidence against the null hypothesis. The test p-value is smaller than the level of significance \(\alpha=0.01\), hence the test result is highly significant.
The data suggests that after the change of speed limit, the average car speed is lower.
Use a 95% C.I. to get a range of plausible values for the average car speed with the reduced speed limit.
Example: Web server workloads

Fundamental to the goal of improving WWW performance is an understanding of WWW workloads. The number of requests that a server must process and the number of data bytes that the server must transfer were examined for various internet web servers from Dec 1992 to Dec 1994. The study stated that the NASA server (4 DEC Alpha 2100 servers with 128MB RAM) had an average daily number of 56,700 requests. Researchers thought that the number was not correct. They collected a sample of 1000 access logs of web requests in the same time period. The sample average was 58,534 with standard deviation equal to 7,436. Can they conclude that the previous study result was not correct?

Null hypothesis  Ho: $\mu = 56,700$

Alternative hypothesis  Ha: $\mu \neq 56,700$

Test statistic  $z = \frac{58,534 - 56,700}{7,436 / \sqrt{1000}} = 7.79$
The area under the standard normal curve between -7.79 and 7.79 is almost 100%. The test p-value is the sum of the areas under the s.n. curve for values on the left of -7.79 and on right of 7.79. Hence the test p-value is about zero.

The data provides very strong evidence that the null hypothesis should be rejected. The researchers are correct in thinking that the results from the previous study were not reliable.

We can use the sample values to get an estimate of the average number of requests for the NASA web server. The sample average is 58,534 requests. A 95% confidence interval for the average number of requests is computed as

$$58,534 \pm 1.96 \times \left(\frac{7,436}{\sqrt{1000}}\right) = 58,534 \pm 460.89$$

The average requests for the web server is between 58,073 and 58,994.
General remarks about tests

• The z-test is used for **reasonable large samples**, when the normal approximation can be used on the probability histograms for the sample average or the sample percentage. For small samples other techniques must be used.

• The z-test is used for **simple random samples**. If the sample is not random, test results are unreliable.

• A **significance test** answers the question of whether an observed difference is real (the *alternative hypothesis*) or just a chance variation (the *null hypothesis*).

• Set a level of significance (0.05 or 0.01) to make a decision about the null hypothesis in the light of the data. Choose a smaller level of significance (0.01) if the decision of rejecting $H_0$ in favor of $H_a$ will have strong practical consequences.

• If a difference is statistically significant, then it is hard to explain away as a chance variation. **Significant does not mean important!** The p-value depends on the sample size. If the sample is large, even a small difference can be “statistically significant”, that is hard to explain by the chance variability.
Using the z-test for percentages

Percentage of E-mail users: technology is dramatically changing the way we communicate. In 1997, an Internet provider claimed that less than 20% of U.S. households use e-mail. A survey of 880 U.S. households showed that 149 of them use email (based on data from the *World Almanac and Book of Facts*). Is the claim true?

We can use the data to test the claim on the population percentage. The significance test for percentages is similar to the test for averages.

The null hypothesis is $H_0: p = 0.20$

The alternative hypothesis is $H_a: p < 0.20$

What is the test statistic? It is the z-statistic $z = \frac{observed - expected}{S.E.}$
The observed percentage of households that use emails is

$$\hat{p} = \frac{149}{880} = 0.17.$$  

The expected percentage under the null hypothesis is $p = 0.20$.

Is the difference between the observed value and the expected value due to chance error or is it “real”?

We use the $z$-statistic to check whether that difference can be explained by sampling variability. The values in the statistic are now percentages!

$$z = \frac{\text{observed} - \text{expected}}{S.E.} = \frac{0.17 - 0.20}{0.013} = -2.30$$

The standard error is calculated using the expected value $p_0$ under the null hypothesis.

$$S.E.(\hat{p}) = \sqrt{p_0 (1 - p_0) / n} = \sqrt{0.2 * (1 - 0.2) / 880} = 0.013$$
The z-statistic is \( z = -2.30 \). So the observed value is \(-2.30\) S.E.’s away from the expected value … the difference cannot be explained by chance variation!

The p-value is the probability of observing \( z = -2.30 \) or a value smaller than \(-2.30\). The approximate p-value is

\[
P = \frac{(100\% - 97.86\%)}{2} = 1.07\% \text{ or } P = 0.0107
\]

The p-value is very small, we reject the null hypothesis with about 1% significance level. The test result is highly significant.

We can conclude that in 1997 the percentage of U.S. households using emails is less than 20%
Z-tests for percentages

Assume we have a large simple random sample of size n. We want to test the null hypothesis that the percentage of a certain characteristic in the population is $H_0: p=p_0$. The sample percentage is denoted by $\hat{p}$

Compute the z-statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 \ast (1 - p_0) / n}}$$

The approximate p-value for a test on $H_0$ against

- Ha: $p > p_0$ the p-value is equal to the area on the right of z
- Ha: $p < p_0$ the p-value is equal to the area on the left of z
- Ha: $p \neq p_0$ the p-value is equal to the sum of the area on the right of z and the area on the left of -z
Example: Tech news

Chicago land's technology professionals get local technology news from various newspapers and magazines. A marketing company claims that 25% of the IT professionals choose the Chicago Tribune as their primary source for local IT news. A survey was conducted last month to check this claim. Among a sample of 750 IT professionals in the Chicago land area, 23.43% of them prefer the Chicago Tribune. Can we conclude that the claim of the marketing company is true?

Null hypothesis: Ho: $p = 0.25$ & Alternative hypothesis: $p \neq 0.25$

$$Z\text{-statistic} \quad z = \frac{0.23 - 0.25}{\sqrt{0.25 \times (1 - 0.25) / 750}} = -0.981$$

The observed percentage is about 1 s.e. away from the expected percentage under the null hypothesis.
The p-value is the sum of the areas on the left of –0.981 and on the right of 0.981. The p-value is 100-67.03=32.97% or 0.33. The p-value is large, it indicates that there is about 33% chance to observe a sample like the one actually selected if the null hypothesis is true.

Hence the data provides evidence to support the hypothesis that 25% of IT professionals in the Chicago land prefer the Chicago Tribune.
Testing effectiveness of nicotine patches

In one study of 71 smokers who tried to quit smoking with nicotine patch therapy, 32 were not smoking one year after the treatment (from *Journal of the American Medical Association, Vol 274*). Use a 0.10 significance level to test the claim that among smokers who try to quit with nicotine patch therapy, the majority are smoking a year after the treatment. Do these results suggest that the nicotine patch therapy is not effective?

Let p be the probability that a smoker is still smoking a year after the treatment with nicotine patch therapy. We compute a significance test to study the hypothesis above.

Null hypothesis: $p=0.5$ & Alternative hypothesis: $p > 0.5$

From the sample the observed percentage is $\hat{p} = 39 / 71 = 0.55$

$Z$-statistic: $z = \frac{0.55 - 0.5}{\sqrt{0.5 \times 0.5 / 71}} = 0.833$
The p-value is equal to the area on the right of z. It is computed as $p-value=(100-59.05)/2=20.47\%$ or $p-value=0.204$. The p-value is larger than the chosen significance level $\alpha=0.10$. Hence the null hypothesis cannot be rejected.

On the basis of the data, we cannot reject the hypothesis than half of the smokers smoke a year after the treatment with the nicotine patch therapy.

Furthermore, we can compute a 95% confidence interval for the percentage of smokers that smoke a year after the treatment. It is given as

$0.55 \pm 1.96*0.06=0.55 \pm 0.118$ or $(0.432, 0.667)$

The percentage of smokers is between 43.2% and 66.7%. This is a large interval, it is hard to decide about the effectiveness of the nicotine patches.
Example: Geographic distribution of requests

A study to analyze the Web servers workload examined the requests received by six Web servers, including for instance the NASA Web server, the University of Calgary web server and the University of Waterloo web server. Part of the analysis examined the geographic distribution of the document requests, using the IP addresses of the requesting hosts in the access log. A conclusion of the analysis was that in all six Web servers, remote hosts send the most requests and receive the most data. Remote hosts account for over 75% in all web servers but Calgary. In fact at Calgary 53.6% of the requests were from remote hosts. If we know that the percentage was calculated from a sample of 1,000 requests, can we still conclude that the majority of the requests came from remote hosts?

Let $p$ be the percentage of requests sent by remote hosts.

Null hypothesis: $p = 0.5$ & Alternative hypothesis: $p > 0.5$

Z-statistic

$$z = \frac{0.536 - 0.5}{\sqrt{0.5 \times 0.5 / 1000}} = 2.25$$
The p-value is computed as \((100-97.56)/2=1.22\%\) or 0.0122. This indicates that there is 1.22\% chance to observe a sample such as the one observed if the null hypothesis is true. Ho is not a good explanation for the observations!!

Since the p-value is smaller than \(\alpha=0.05\), the test result is statistically significant. We can reject the null hypothesis at 5\% significance level. The data provides evidence to support the conclusion that the majority of web requests to the University of Calgary web server come from remote hosts.