CS555B1 Data Analysis and Visualization

Lecture 3

Kia Teymourian

Statistical Inference

Statistical inference provides methods for drawing **conclusions** about a **population** from **sample data**.

The goals of **statistical inference** are:

- **Draw conclusion** about a population based on sample data
- **Provide a statement**, expressed in the language of probability, of how much confidence to be placed in the conclusion

Two of the most common types of **statistical inference** include:

- **Confidence Intervals** for estimating the value of a population parameter
- **Tests of Significance** or **Hypothesis Tests** which assess the evidence for a scientific claim.

An example – calculating confidence interval

Suppose a company would like to estimate the average amount of time callers are on hold before they reach a customer service representative.

The **parameter of interest** is the **population mean**, μ.

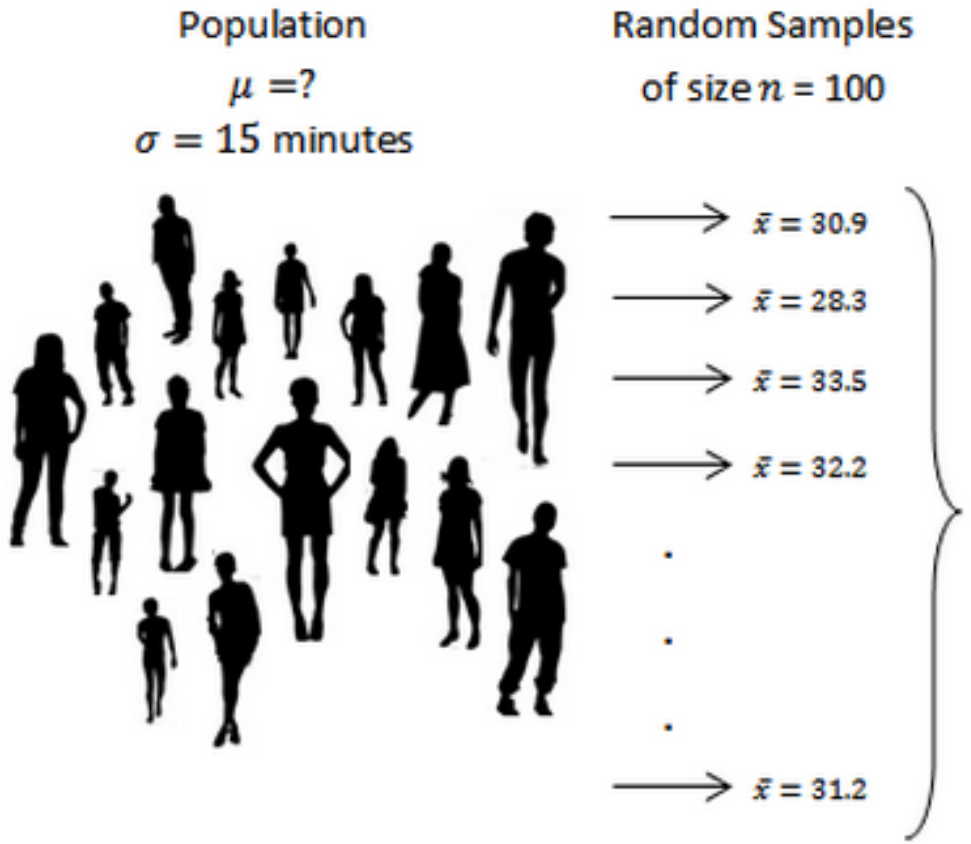
Executives from the company plan to **randomly sample 100** calls and use the sample mean, **x bar**, to estimate **population mean μ** (the average of all callers wait times).

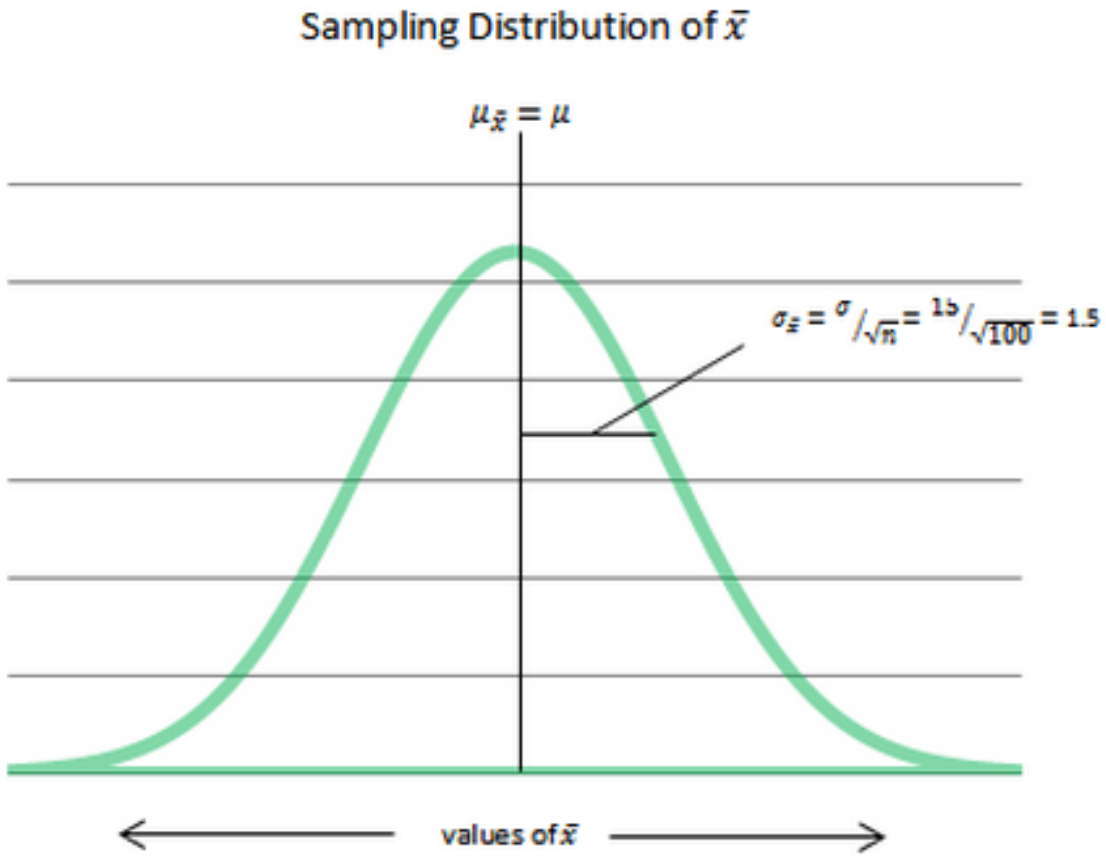
- Assume that the **population standard deviation** of wait times, **σ, is 15 minutes.**
- The sample mean of the wait times from the **100 random calls in the sample is 30.9 minutes**.

Executives also want to compute a confidence interval to assess the accuracy of the point estimate (the sample mean).

An example – calculating confidence interval

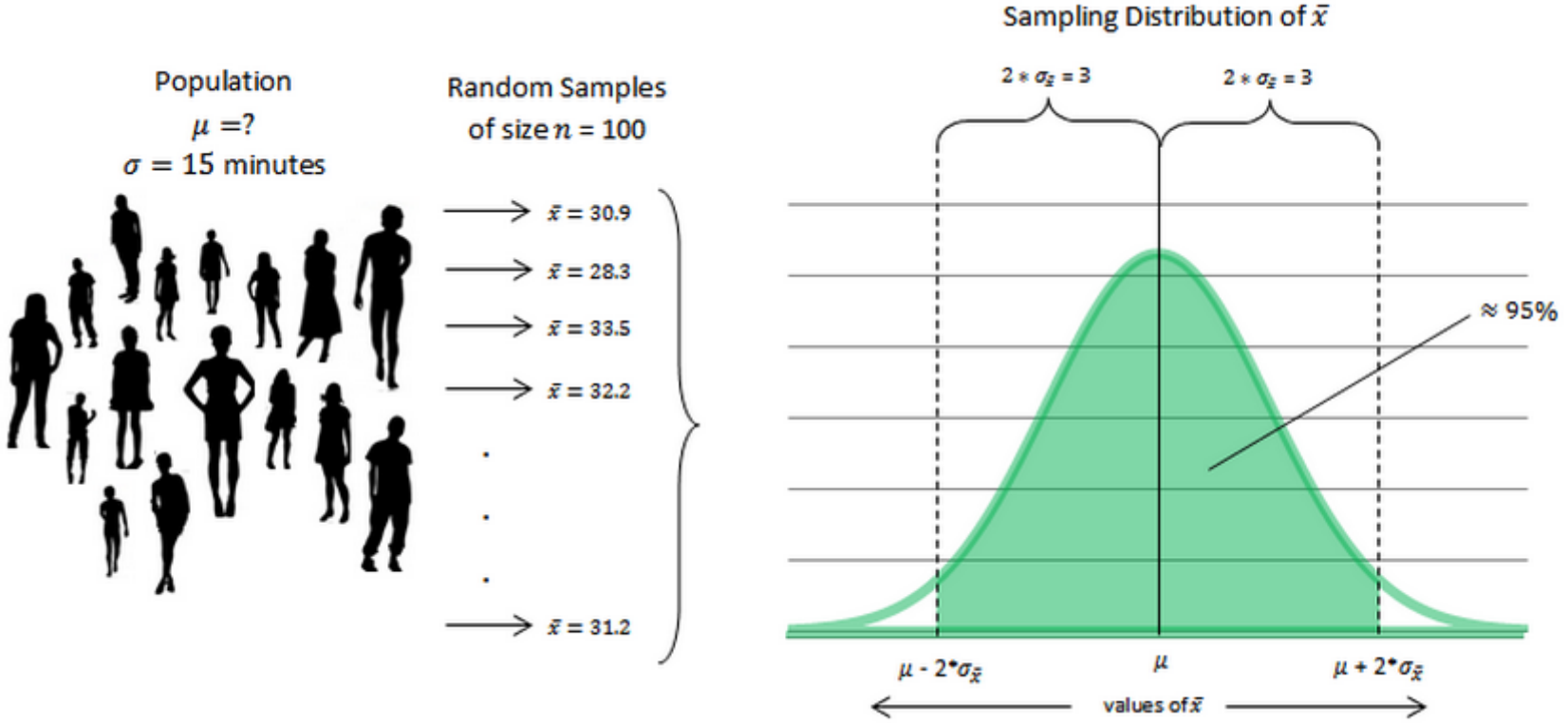
Since n≥30 (here, n=100), we know that the sample mean is approximately normally distributed with a mean of $\mu_{\bar{x}} = \mu$ and a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = 1.5$ minutes.





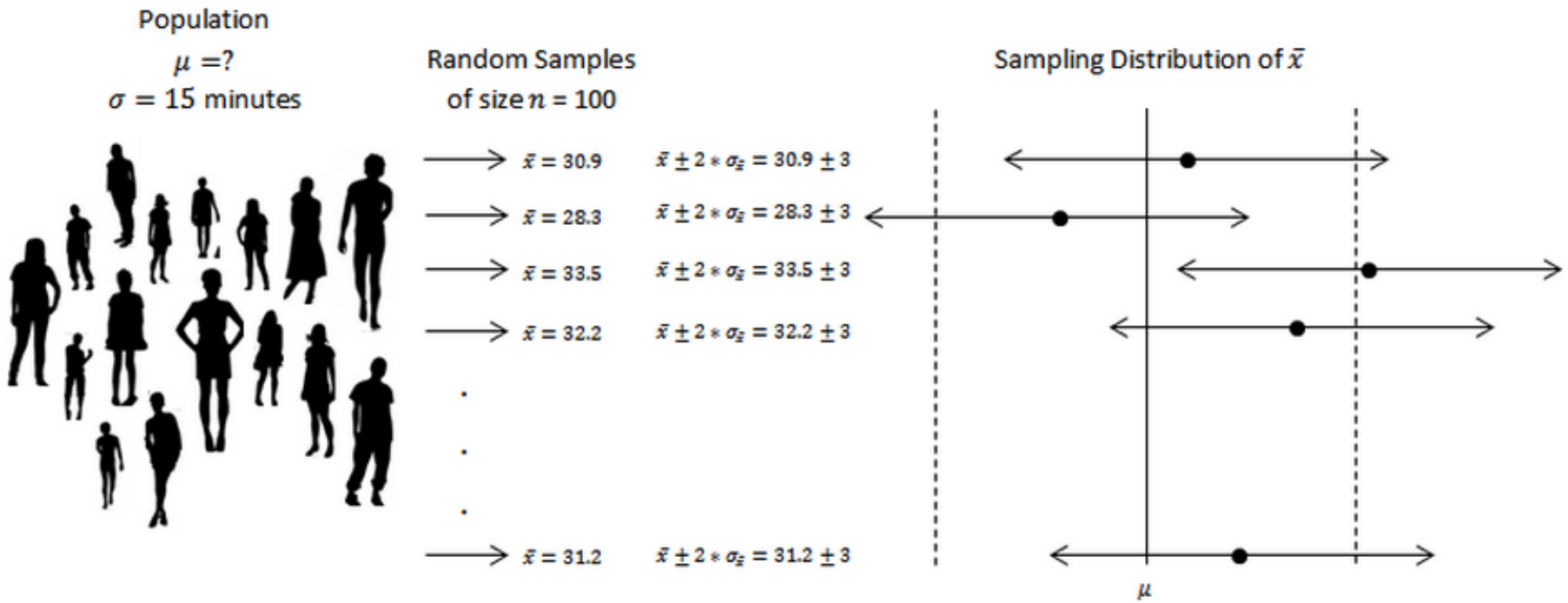
An example – calculating confidence interval

- In a normal distribution, 95% of the sample means are within 2 SDs of the population mean (z=1.96 corresponds to an area of 95% under the curve).
- The sample mean, **x bar**, and the population mean, **μ**, are within three minutes (**2 SDs**) of each other **95% of the time**.
- If our estimate of the population mean is between **xbar−3=30.9−3=27.9** and **xbar+3=30.9+3=33.9** then we will be right **95%** of the time.



An example – calculating confidence interval

- If we repeatedly take random samples of size n=100, then each separate sample will have its own unique sample mean.
- We know that 95% of the sample means will be within 2 SDs of the population mean.
- The sample mean ± 2 times the SD ($\bar{x} \pm 2 \cdot \sigma_{\bar{x}}$) is called the 95% confidence interval for the population mean.
- The two-way arrows in the below figure represent the confidence intervals from each random sample. The center point is the sample mean and the end of the arrows show the edges of each interval.



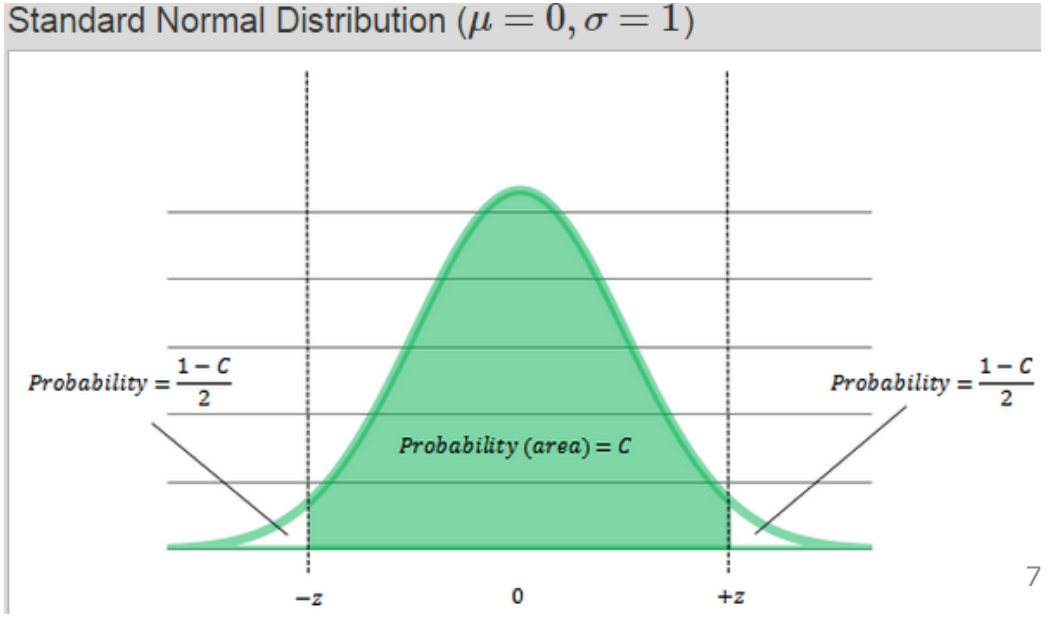
Confidence interval

- Most confidence intervals are of the form: **estimate ± margin of error**
- Generally, **95%** confidence intervals are standard. Other common intervals include **90%** confidence intervals or **99%** confidence intervals.

- To calculate a confidence interval with a **confidence level of C** for the population mean, μ, we use the following formula: $\overline{x} \mp z \cdot \frac{\sigma}{\sqrt{n}}$

  z is the appropriate critical value corresponding to the confidence level
- Typical values of C with their associated values of z are:

Confidence Level, C	90%	95%	99%
Critical Value, z	1.645	1.960	2.576



Exercise 1

Call Center Data – Waiting Time

Suppose the executives would like to estimate the average amount of time callers are on hold before they reach a customer service representative by computing a 99% confidence interval. The population standard deviation of wait times, σ, is 15 minutes. The sample mean of the wait times from the 100 random calls in the sample is 30.9 minutes.

To calculate a confidence interval with a confidence level of C for the population mean, μ, we use the formula: $\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$.

The 99% confidence interval is calculated as follows:

$$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}} = 30.9 \pm 2.576 * \frac{15}{\sqrt{100}} = 30.9 \pm 2.576 * 1.5 = (27.036, 34.764)$$

The Sample size and the margin of error

- In the case of the confidence interval for the population mean, the margin of error is equal to

$$z \cdot \frac{\sigma}{\sqrt{n}}$$

  - **z gets smaller.** **Smaller** values of z are associated **decreased** confidence levels. There is a **trade off between the confidence level and the margin of error**. To achieve a narrower confidence interval, one must be willing to accept a lower amount of confidence.
  - **sd gets smaller.** The standard deviation measures the variability of the population. When the variability of the individual observations is reduced, it becomes easier to estimate the population mean with higher precision.
  - **n gets larger.** Increasing the sample size n reduces the margin of error for a fixed confidence level C. In order to reduce the margin of error by half, we must take a sample that is four times as large.

- To obtain a specific margin of error, m, for a desired confidence level, C, you can use the following **formula** to determine the number of observations that you'd need:

$$n = \left(\frac{z \cdot \sigma}{m}\right)^2$$

Exercise 2

Call Center Data – Waiting Time

Suppose the executives would like to estimate the average amount of time callers are on hold before they reach a customer service representative by computing a 95% confidence interval. As before, the population standard deviation of wait times, σ, is 15 minutes. The executives would like the half width of the confidence interval to be 1.0 minutes (that is, they'd like the margin of error to be 1.0). How many callers must they sample to get their 95% confidence interval to have the desired width?

To obtain a specific margin of error, m, for a desired confidence level, C, we use the formula to determine the number of observations needed: $n = \left(\frac{z*\sigma}{m}\right)^2$

the number of observations needed is calculated as follows:

$$n = \left(\frac{z * \sigma}{m}\right)^2 = \left(\frac{1.96 * 15}{1}\right)^2 = 29.4^2 = 864.36$$

Confidence Level, C	90%	95%	99%
Critical Value, z	1.645	1.960	2.576

Note: In calculations involving sample size we always round up (instead of down). In our example 865 callers.

Example: Tests of Significance - On Time Flights

Example: Suppose an airline company claims that their flights arrive on time **90%** of the time. However, on your **past 10 flights** with the airline, only **4 of your flights** have been on time (= sample proportion of 40%).

You probably would be skeptical of their claim of 90% of flights being on time, wouldn't you?

- We took random samples of 10 flights with on time rate of 90%, it would be very rare that a random sample of this size would only give 4 flights that were on time.
- It gives strong evidence against the **claim of 90% of flights being on time**.
- If the airline's flights were really on time 90% of the time, the **probability of 4 or less flights on time** in a sample of **10 flights** is less than **0.02% (probability <0.0002)**.
- Such a small probability of this happening by chance if their claim of 90% of flights being on time was true gives **evidence against their claim**.

Set up the hypotheses

- We formally call the claim that we are hoping to **disprove the null hypothesis**. The null hypothesis is generally denoted as **H_0**.

- The **test of significance** is designed to assess the strength of the evidence **against** the **null hypothesis**.

- Usually the null hypothesis is a statement of **"no effect" or "no difference"**.

- The conclusion we'd like to make is captured in the **alternative hypothesis**. The **alternative hypothesis is generally denoted H_A or H_1**. This is written as the opposite of the null hypothesis and generally suggests that there is an **"effect" or "difference."**
- Hypotheses always refer to some population or distribution, not to a particular sample outcome. Thus we state H_0 and H_1 in terms of population parameters and not in terms of sample statistics.
- The **alternative hypothesis** states that a parameter differs from its null value in a specific direction **(an one-sided alternative)** or in either direction (**a two-sided alternative**).

Set up the hypotheses - one-sided alternative

Example: A gym is interested in whether or not a 6-week weight loss training program they launched has been successful in helping their clients lose weight. To assess this, they took a sample of **30 participants**. State the null and alternative hypotheses.

Let's denote the mean weight change, μ, as the population parameter of interest. The gym is specifically interested in whether program participants on average lost weight.

Thus they would be interested in an **one-sided alternative** where they seek to claim that program participants lose weight on average.

A shorthand notation to capture these in an easy to read form:

- **H_0:μ=0** (there is ***no effect on weight change*** of program participants)
- **H_1:μ<0** (program ***participants lose weight*** on average)

Set up the hypotheses - two-sided alternative

Example: County officials are interested in measuring a particular chemical in water sources in the county. High levels of this chemical are harmful as are low levels. Either indicate a need for intervention.

Normal levels of this chemical are **15 parts per million (ppm)**. County officials will start an intervention program if the chemical mean level in the water sources **is different than 15** ppm.

Samples from 50 water sources throughout the county are taken and the levels of this chemical are measured. State the null and alternative hypotheses.

Let's denote the mean level of the chemical, μ, as the population parameter of interest. The alternative hypothesis is that the levels are different than 15 ppm. It is a **two-sided alternative.**

A shorthand notation to capture these in an easy to read form:

- H_0:μ=15 (the mean level of the chemical is **normal**)
- H_1:$\mu \neq 15$ (the mean level of the chemical is **abnormal**)

Test statistics

Significance tests generally compare the value of the population parameter in the null hypothesis to the value of the estimate from the sample.

The test statistic is generally a measurement of how far the sample statistic is from the expected value under the null hypothesis.

Large values of the test statistic indicate that the estimate is far from the expected value (assuming that the null hypothesis is true) and give evidence against the null hypothesis.

The test statistic for hypotheses about the mean μ is the z statistic: $$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$\bar{x}$ is the sample mean, μ is the value of the population mean under the null hypothesis, σ is the standard deviation of the variable of interest in the population, and n is the number of observations in the sample

This measures how far $\bar{x}$ is from the value of μ (under the null hypothesis) in standard deviation units.

Z test

A Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution.

Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. Many statistical tests can be conveniently performed as approximate Z-tests if

- the sample size is large n≥30
- the population variance is known or unknown

Confidence interval is $\bar{x} \pm Z_{CL} * \frac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm Z_{CL} * \frac{S}{\sqrt{n}}$

If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large (n < 30), the Student's t-test may be more appropriate.

P-value of the test

One of the ways that we can measure how far the point estimate is from the expected value of the population parameter under the null hypothesis is to calculate the test **statistic's associated p-value**.
The P-value of the test is the **probability**, computed **assuming H_0 is true**, that the test statistic would take a value as extreme or more extreme than that actually observed.
The P-value is a measure of the strength of the evidence against H_0.

- **A small P-value** suggests that the observed result was **unlikely to occur if the null hypothesis is in fact true**.

- **Large P-values**, on the other hand, do not give evidence against the null hypothesis.

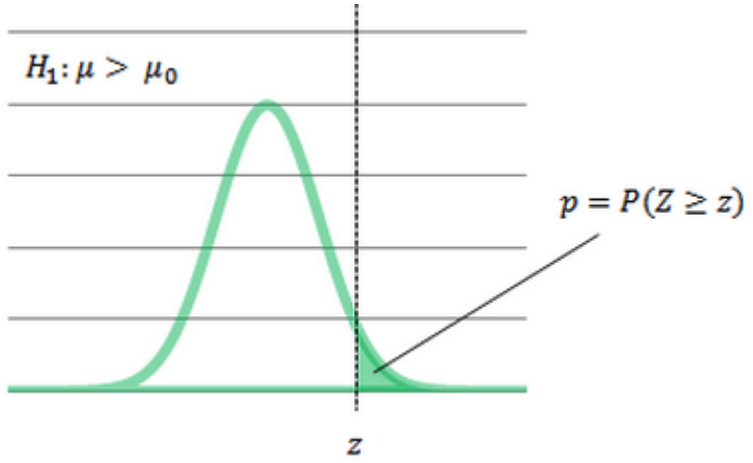
If the resulting p-value is large (meaning that the test statistic is small), then the sample did not give evidence against the null hypothesis.

A large P-value only means that the data are inconsistent with the null hypothesis, not that we have clear evidence that the null hypothesis is untrue. Statistical tests are set up to look for evidence against the null, not to prove that the null hypothesis is untrue.

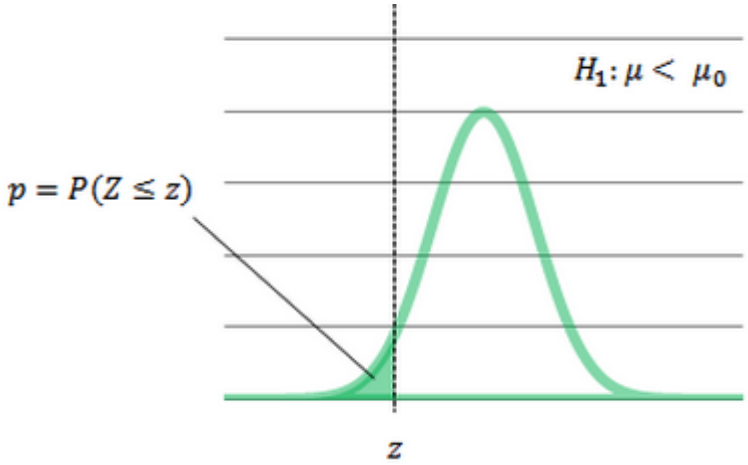
P-value of the test

The p-value for the z statistic is calculated using the standard normal distribution.

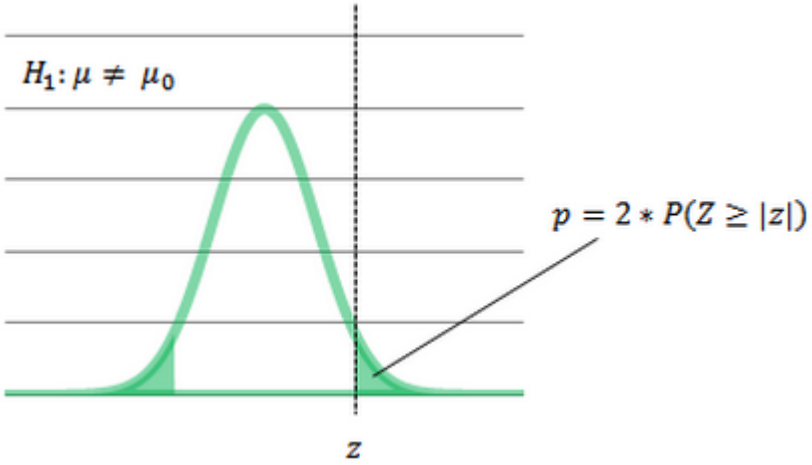
Depending on the alternative hypothesis, the p-value is calculated as follows:



An **one-sided alternative hypothesis** stating that the population mean is greater than a specified value



An **one-sided alternative hypothesis** stating that the population mean is less than a specified value



A two-sided alternative hypothesis stating that the population mean is different from a specified value

Example: Calculating P-value for an one-sided test

A gym is interested in whether a 6-week weight loss training program they launched has been successful in helping their clients lose weight. To assess this, they took a sample of 30 participants. They are interested in testing the following hypotheses:

- $\mathbf{H_0: \mu = 0}$ (there is **no effect on weight** change of program participants)
- $\mathbf{H_1: \mu < 0}$ (program participants **lose weight** on average)

Suppose we know that for the general population, the standard deviation of changes in weights over a six-week interval is 6 pounds.

The sample mean of the change in weight for the **30 participants** in the sample was **−2.98 pounds**.

Calculate the value of the test statistic and the associated **p-value.**

Example: Calculating the value of the Test statistic and P-value

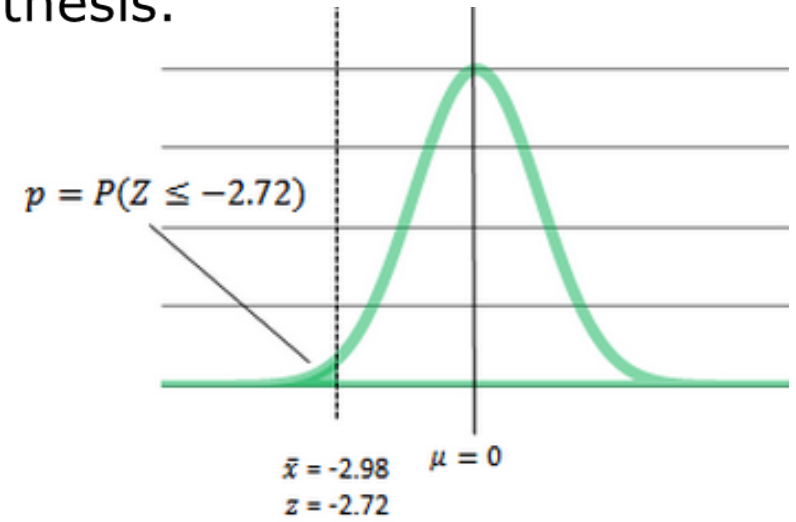
The standard deviation of changes in weights over a six-week interval is 6 pounds. The sample mean of the change in weight for the 30 participants was −2.98 pounds.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \\ &= \frac{-2.98 - 0}{\frac{6}{\sqrt{30}}} \\ &\approx \frac{-2.98}{1.0954} \\ &\approx -2.72 \end{aligned}$$

The p-value is the probability that the test statistic is −2.72 or more extreme, which is the probability that Z≤−2.72. Using the standard normal table, we can calculate: P=P(Z≤−2.72)=0.0033

This is a small p-value. It appears that the sample mean ($\bar{x}$=−2.98) is highly unlikely to have occurred if the true population mean μ=0. Thus we have strong evidence against the null hypothesis.



Calculating P-value for a two-sided test

Example: Normal levels of this chemical are 15 parts per million (ppm).

Samples from 50 water sources throughout the county are taken and the levels of this chemical are measured.

They are interested in testing the following hypotheses:

- **H0:μ=15 (the mean level of the chemical is normal)**
- **H1:μ≠15 (the mean level of the chemical is abnormal)**

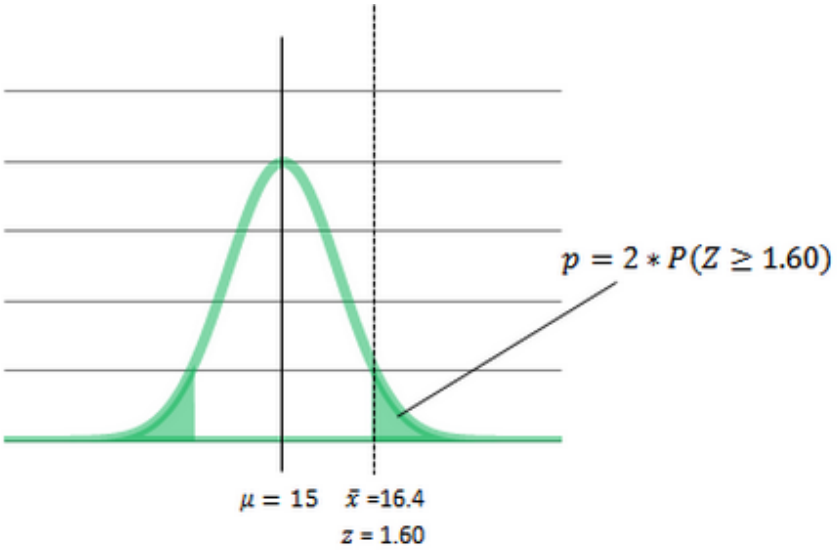
Suppose we know that the population standard deviation is 6.2. The sample mean from the 50 samples was 16.4 ppm.

Calculate the value of the test statistic and the associated p-value.

Calculating the value of the test statistic and P-value

The standard deviation of changes in weights over a six-week interval is **6** pounds. The sample mean of the change in weight for the 30 participants was **−2.98** pounds.

$$
\begin{aligned}
z &= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \\
&= \frac{16.4 - 15}{\frac{6.2}{\sqrt{50}}} \\
&\approx \frac{1.4}{0.8768} \\
&\approx 1.60
\end{aligned}
$$



The **p-value** is the probability that the test statistic is 1.60 or more extreme.

That is, the **p-value is the probability that Z≥1.60 or Z≤−1.60**. Using the standard normal table, we can calculate:

$$
\begin{aligned}
P &= P(Z \leq -1.60 \text{ or } Z \geq 1.60) \\
&= P(Z \geq 1.60) + P(Z \leq -1.60) \\
&= 2 \cdot P(Z \geq 1.60) \\
&= 2 \cdot 0.0548 \\
&= 0.1096
\end{aligned}
$$

It appears that the sample mean that we observed ($\bar{x}$=16.4) is moderately likely to have occurred if the true population mean was 15 ppm (if μ=15). **This means we don't have strong evidence against the null hypothesis.**

Evaluating Hypotheses using P-value

In order to decide whether to **reject the null hypothesis**, we can either compare:

- **the p-value to a pre-defined significance level**
- **the test statistic to a critical value**

For either method, the **significance level** needs to be pre-specified.
It is denoted with α **(the Greek letter "alpha")**.

The most common choice for **alpha is 0.05**. It means that the evidence from the data is so strong that the result obtained would **only appear 5%** of the time or less if the null hypothesis is in fact true.

If the **p-value ≤ α**, then we **reject the null hypothesis** in favor of the alternative hypothesis.

If the **p-value > α**, then we **fail to reject the null hypothesis**. We do not have sufficient evidence that the null hypothesis is not true.

Evaluating Hypotheses using critical values

For a significance test where the test statistic is normally distributed with a mean of 0 and a standard deviation of 1, the critical value is equal to the value of Z that corresponds with the predefined significance level.

The critical value depends on the **significance level** and whether the **test is one- or two-sided.**

α	10%	5%	2%	1%
1-sided critical value	1.282	1.645	2.054	2.326
2-sided critical value	1.645	1.960	2.326	2.576

If the absolute value of the test statistic ≥ the critical value, then **we reject** the null hypothesis in favor of the alternative hypothesis.

If the absolute value of the test statistic < the critical value, then **we fail to reject** the null hypothesis.

Significance Test Result

If the p-value ≤ α or (equivalently) the absolute value of the test statistic ≥ the critical value, then we say that the data are **"statistically significant at the α level."**

In statistics, **"Significant" means "not likely to have happened by chance."**

The **significance level** quantifies exactly how we are defining **"not likely."**

Generally, in reporting results we should report both the significance level as well as the p-value as others may be interested in a more or less conservative level of significance.

By provided the exact p-value, the **reader may compare the p-value with their own level of significance.**

Test of Significance (hypothesis test)

Tests of significance are used to assess the evidence provided by the data from a sample about some claim concerning the population.

There are **5 key steps in carrying out any significance test**:

1. Set up the **hypotheses** and **select the alpha level**
2. Select the appropriate **test statistic**
3. State the **decision rule**
4. Compute the **test statistic** and the associated **p-value**
5. State your **conclusion**

Example of a One-sided Test

A gym is interested in whether a 6-week weight loss training program they launched has been successful in helping their clients lose weight. To assess this, they took a sample of 30 participants.

Suppose we know that for the general population, the standard deviation of changes in weights over a six-week interval is 6 pounds. The sample mean of the change in weight for the 30 participants in the sample was −2.98 pounds.

Perform a significance test to determine whether the weight loss training program has been successful at the $\alpha=0.05$ level of significance.

Example of a One-sided Test

1. Set up the hypotheses and select the alpha level
- H_0:μ=0 (there is no effect on weight change of program participants)
- H_1:μ<0 (program participants lose weight on average)
- $\alpha = 0.0.5$

2. Select the appropriate test statistic

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

3. State the decision rule

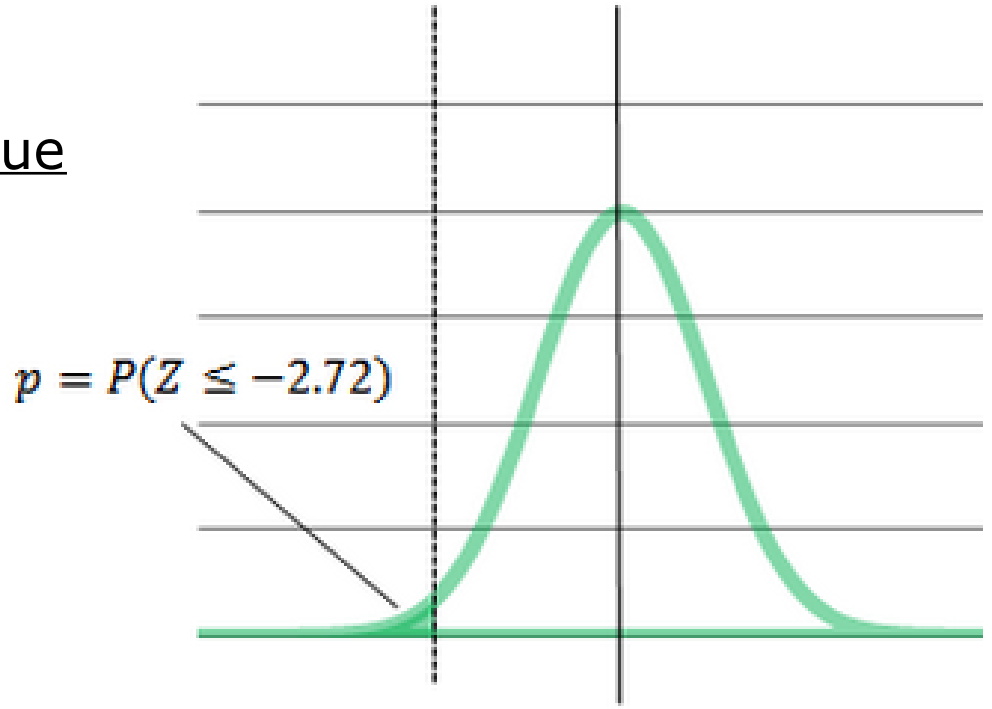
- Determine the appropriate critical value from the standard normal distribution table associated with a right hand tail probability of α=0.05. Using the table, the appropriate critical value is 1.645.
- Decision Rule: Reject H_0 if $|z| \geq 1.645$
- Otherwise, do not reject H_0

An one-sided test example

4. Compute the test statistic and the associated p-value

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{-2.98 - 0}{\frac{6}{\sqrt{30}}} \approx \frac{-2.98}{1.0954} \approx -2.72$$

$$p = P(Z \leq -2.72) = 0.0033$$



5. Conclusion

- Reject H_0 since $|-2.72| \geq 1.645$.
- We have significant evidence at the $\alpha = 0.05$ level that $\mu < 0$.
- We reject the null hypothesis that the weight loss program is no effect on weight change of program participants in favor of the alternative hypothesis that program participants lose weight on average (p = 0.0033).

A two-sided test example

Example: County officials were interested in measuring a particular chemical in water sources in the county. High levels of this chemical are as harmful as are low levels. Normal levels of this chemical are **15 parts per million** (ppm).

Samples from **50** water sources throughout the county are taken and the levels of this chemical are measured. Suppose we know that the **population standard deviation is 6.2**. The sample mean from the 50 samples was **16.4 ppm**.

A significance test was conducted to determine whether the mean levels are different than 15 ppm at the **α=0.05 level of significance**.

However, there was not enough evidence to suggest that the mean levels were not normal.

Given that the significance test did not reject the null hypothesis at the **α=0.05 level,** ***would you expect that the 95% confidence interval for the population mean to include 15 ppm or not?***

A two-sided test example

Let's confirm by calculating the confidence interval. To calculate a confidence interval with a confidence level of C for the population mean, μ, we use formula:

$$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, z is the appropriate critical value corresponding to the confidence level, σ is the population standard deviation, and n is the sample size.

The 95% confidence interval is calculated:

$$\begin{aligned} \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}} &= 16.4 \pm 1.960 \cdot \frac{6.2}{\sqrt{50}} \\ &= 16.4 \pm 1.960 \cdot 0.8768 \\ &= 16.4 \pm 1.7186 \\ &\approx (14.68, 18.12) \end{aligned}$$

We are 95% confident that the true mean level is between 14.69 ppm and 18.12 ppm.

The 95% confidence interval contained the null value of 15 ppm since the two-sided significance test at the α=0.05 level did not reject the null hypothesis that μ=15 (the mean levels are normal).

Z test R commands

- **Z-test** (population SD is known = popsd)

```
> z <- mean(data$variable)/(popsd/sqrt(nrow(data))
> pnorm(z) or 1- pnorm(z) or 2*(1- pnorm(abs(z)))
> lower <- mean(data$variable)-z*popsd/sqrt(nrow(data))
> upper <- mean(data$variable)+z*popsd/sqrt(nrow(data))
```

- **asbio package (A Collection of Statistical Tools for Biologists)**

```
> install.packages("asbio")
> library(asbio)
> one.sample.z(null.mu=[μ0], xbar=mean(data$variable), sigma=popsd, n=nrow(data),
alternative=[alternative], conf=[confidence level])
```

[alternative] = ‘less’, ‘greater’, or ‘two.sided’

```
# read more in manual page of one.samle.z
>?one.sample.z
```

Significance Tests and Confidence Intervals

There is a relationship between **two sided tests** conducted with a significance level of **α and $1-\alpha$ confidence intervals.**

A level α significance test rejects the **null hypothesis $H_0: \mu = \mu_0$** when the value of **μ_0 is not included in the $1-\alpha$ confidence interval for μ.**

On the other hand, a level α **significance test fails to reject the null hypothesis** $H_0: \mu = \mu_0$ when the value of **μ_0 is included in the $1-\alpha$ confidence interval for μ.**

The conclusion of a significance test (whether or not the null hypothesis is rejected) at the α level of significance can be determined by checking if the "**null**" value of the mean as specified by the null hypothesis (**μ_0**) is contained within the $1-\alpha$ confidence interval.

Further Reading Resources

- **Statistics Books:**
  - McClave, J. T., & Sincich, T. (2012). Statistics (12th ed.). Pearson.
  - Moore, D. S. (2003). The basic practice of statistics (3rd ed.). W. H. Freeman.

- **Online Book and Demos**
  - http://onlinestatbook.com