Inferences about population central values

- estimate the value of the population parameter
  (“What is the value of the population parameter?”)

- test a hypothesis about the value of the population parameter
  (“Is the parameter value equal to this specific value?”)
Estimation of $\mu$ (Ch. 5.2) – when $\sigma$ is known

When making inference about the population mean it is important to have some idea about the accuracy of the estimate.

When a random sample of size $n$ is drawn from a population, the sample mean $\bar{x}$ is approximately normally distributed with its mean equal to the population mean $\mu$ and the standard deviation $\sigma_x$ equal to the $\sigma / \sqrt{n}$ – Central Limit Theorem

Sampling distribution for $\bar{x}$

$\sim 95\%$ of the sample means lies here
From our knowledge of the Empirical rule we can say that the interval $\mu \pm 2\sigma_x$ includes approximately 95% of all the $\bar{x}$ in repeated sampling.

From our knowledge of areas under the standard normal curve we can say more accurately – the interval $\mu \pm 1.96\sigma_x$.

Sampling distribution of $\bar{x}$

Use Table 1 (Ott&Longnecker) to determine an appropriate $z$ value – for 95% confidence interval look for $z$ corresponding to 0.9750.
Unless an unlucky 5% chance, \( \bar{x} \) will lie between \( \mu - 1.96\sigma_x \) and \( \mu + 1.96\sigma_x \).

\[
\mu - 1.96\sigma_x \leq \bar{x} \leq \mu + 1.96\sigma_x
\]

**Conclusion:** unless an unlucky 5% chance occurred in drawing the sample

\[
\bar{x} - 1.96\sigma_x \leq \mu \leq \bar{x} + 1.96\sigma_x
\]

This interval is called a 95% confidence interval for the mean.
99% confidence interval for the mean?

Use Table 1 (Ott&Longnecker) to determine an appropriate \( z \) value
– for 99% confidence interval look for \( z \) corresponding to 0.9950

\[
\bar{x} - 2.58\sigma_x \leq \mu \leq \bar{x} + 2.58\sigma_x
\]
In general, the confidence interval for mean with confidence interval \((1-\alpha)\) is determined as \(\mu \pm z_{\alpha/2} \sigma_x\), where \(z_{\alpha/2}\) is a value of \(z\) having a tail area of \(\alpha/2\) to its right.

Table 5.2 (p.201) – a shortcut to the appropriate \(z\) values for some common confidence intervals.

The population standard deviation is known from past experiments or, if the sample size is large, a sample standard deviation is substituted for the population standard deviation.
Example 5.2:
An average number of “count” trees per acre on a 2,000 acre plantation?
A sample of randomly selected 50 1-acre plots ($n=50$).
Sample mean is 27.3,
sample standard deviation is 12.1.
99% confidence interval for the population mean?

From Table 1, $z_{\alpha/2}$ for 99% is equal to 2.58, hence, 99% confidence interval for mean is

$$27.3 \pm 2.58 \frac{12.1}{\sqrt{50}} = 27.3 \pm 4.41$$

We are 99% sure that the average number of count trees per acre is between 22.89 and 31.71.
Size of sample for estimating population mean from the sample data? (Ch. 5.3)

The researcher has to decide:

1) How accurate the estimate should be?

The estimate is desired to be within a certain limit \( \pm E \).

2) What is the probability that the mean is within this limit?

95%? 99%?
\[ \mu \pm E = \mu \pm z_{\alpha/2} \sigma / \sqrt{n} \]

\[ E = z_{\alpha/2} \sigma / \sqrt{n} \]

\( z_{\alpha/2} \) is determined from Table 1 based on the selected confidence level (95%..99%)

\[ n = (z_{\alpha/2})^2 \sigma^2 / E^2 \]
Example – Exercise 5.21

Biologist studies effect of an antibiotic on the growth of a particular bacterium. Of interest is the mean amount of bacteria present per plate of culture when a fixed amount of antibiotic is applied.

From previous studies – \( \sigma = 13 \text{ cm}^2 \)

If the desired \( E = 3 \text{ cm}^2 \), and the confidence interval is 90% - how many cultures need to be developed and tested?

\[
z_{\alpha/2} \quad (\text{for } 90\%) = 1.64
\]

\[
n = \left( z_{\alpha/2} \right)^2 \frac{\sigma^2}{E^2} = 1.64^2 \frac{13^2}{3^2} \approx 51
\]
Test a hypothesis about population mean (Ch. 5.4)

Difference from the parameter estimation – there is a preconceived idea about the value of the population parameter.

Example: the EPA standard for a pollutant concentration in surface water is 0.01 ppm. A sample mean from 100 samples taken in a particular lake is equal to 0.012 ppm – is the pollutant concentration in the lake significantly higher than the EPA standard?
There are two theories (two hypotheses) involved in a testing hypothesis process:

1) Research hypothesis – hypothesis proposed by the person conducting the study
   – e.g., my hypothesis is that the pollutant concentration in the lake exceeds the EPA standard;
   – common symbol for the mean: \( \mu_a \)
   – common symbol for the hypothesis: \( H_a \)

2) Null hypothesis – negation of the research hypothesis
   – e.g., the pollutant concentration in the lake does not exceed the EPA standard;
   – common symbol for the mean: \( \mu_0 \)
   – common symbol for the hypothesis: \( H_0 \).
What criterion should be used to judge which hypothesis is true?

The difference between the sample mean and the value of the mean stated in the null hypothesis $\mu_0 - \bar{x} - \mu_0$

When does the null hypothesis should be rejected?

Rejection region

Figure 5.7
The answer depends on how much risk we are willing to take = how high is the probability of an error.

We can make two types of error:

1) falsely reject the null hypothesis (Type I error) is denoted by the symbol $\alpha$.
2) falsely accept the null hypothesis (Type II error) is denoted by the symbol $\beta$.

Since probabilities of Type I and Type II errors are inversely related it is not possible to minimize the two errors simultaneously.

Power of the test – probability that the test will reject the false null hypothesis = $1 - \beta$
From Table 1 we determine the $z_\alpha$ value corresponding to the selected $\alpha$ – if the $z$-score calculated based on the sample mean is within the rejection region – reject the null hypothesis.
Example with the pollutant concentration:

\( \alpha = 0.05 \) (that is 5% chance of an Type I error – false rejection of the null hypothesis)

\( \bar{x} = 0.012, \ s=0.007, \ n=100 \)

\( \mu_0 = 0.01 \)

\[
z = \frac{\bar{x} - \mu_0}{\sigma_x} = \frac{0.012 - 0.01}{0.007/\sqrt{100}} = 2.86
\]

From Table 1,

\( z_\alpha = z_{0.05} \) (that is, \( z \) corresponding to the area under the curve of 0.95) = 1.64

\( z > z_{0.05} \) – we reject the null hypothesis in favor of the research hypothesis

– the pollutant concentration in the lake is greater than the EPA standard.
Steps of the statistical test for the mean:

1. Formulate the research hypothesis $H_a$ – e.g., pollutant concentration exceeds 0.01.
2. Formulate the null hypothesis $H_0$ – e.g., pollutant concentration does not exceed 0.01.
3. Determine the test statistics – $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$.
4. Determine the rejection region for the specified probability of Type I error $\alpha$.
5. Check assumptions and draw conclusions.
One-tailed test – the rejection region is located in only one tail of the $\bar{x}$ sampling distribution:

**Right-tailed test:**
Null hypothesis $H_0: \mu \leq \mu_0$ vs.
Research hypothesis $H_a: \mu > \mu_0$
Reject $H_0$ if $z \geq z_{\alpha}$
**Left-tailed test:**
Null hypothesis $H_0: \mu \geq \mu_0$ vs.
Research hypothesis $H_a: \mu < \mu_0$
Reject $H_0$ if $z \leq -z_\alpha$

Example – ex. 5.86 – A buyer wishes to determine whether the mean sugar content per orange shipped from a particular grove is less than 0.027 lb. A random sample of 50 oranges - $\bar{x} = 0.025$, $s = 0.003$ lb, $\alpha = 0.05$.

\[
z = \frac{\bar{x} - \mu_0}{\sigma_x} = \frac{0.025 - 0.027}{0.003/\sqrt{50}} = -4.71
\]

From Table 1, $-z_{0.05} = -1.64$ - reject the null hypothesis
Two-tailed test – the rejection region is located in both tails of the $\bar{x}$ sampling distribution:
Null hypothesis $H_0: \mu \neq \mu_0$ vs.
Research hypothesis $H_a: \mu = \mu_0$
Reject $H_0$ if $|z| \geq z_{\alpha/2}$

Example: Is yield of the particular variety of soybean in a given year different from a long-term average of 52 bu/acre. 36-one acre plots, $\bar{x} = 57.3, s = 12.4, \alpha = 0.05$.

$$z = \frac{\bar{x} - \mu_0}{\sigma_x} = \frac{57.3 - 52.0}{12.4 / \sqrt{36}} = 2.56$$

From Table 1, $z_{0.025} = 1.96$ - reject the null hypothesis
The level of significance of a statistical test (Ch. 5.6)

Example: we reporting that the mean was not significantly lower than a specified value at $\alpha=0.05$.
Of interest is: how far is the result short of being significant?

Reporting the actual probability associated with $z$ answers that question.

The probability is called a level of significance of a statistical test ($p$-value).

Level of significance – probability of occurrence for a deviation from the mean that is larger than the sample $z$, assuming that the null hypothesis is true.
Example - one-tailed test:

Null hypothesis $H_0: \mu \geq 0.027$ vs.
Research hypothesis $H_a: \mu < 0.027$

$$z = \frac{\bar{x} - \mu_0}{\sigma_x} = \frac{0.025 - 0.027}{0.003/\sqrt{16}} = -2.66$$

$p$-value$=P(z \leq -2.66) = 0.0039$ (from Table 1)

Example - two-tailed test:

Null hypothesis $H_0: \mu = 0.027$ vs.
Research hypothesis $H_a: \mu \neq 0.027$

$$|z| = \frac{\bar{x} - \mu_0}{\sigma_x} = \left| \frac{0.025 - 0.027}{0.003/\sqrt{16}} \right| = 2.66$$

$p$-value$=2P(z \geq |2.66|) = 2*0.0039 = 0.0078$ (from Table 1)