XII. Hypothesis testing

Instructor: Yanlin Qi

Institute of Transportation Studies

Department of Statistics

University of California, Davis

Sec 8.1 Hypotheses and test procedures

Sec 8.2: z Tests for hypotheses about a population

mean Sec 8.3: The one-sample t test

Introduction

Consider the brown egg problem again.

Suppose the weights of the eggs produced at the farm (population) are normally distributed with unknown mean μ but known standard deviation $\sigma = 2$ g.

It is claimed by the manufacturer that $\mu = 65$ g.

You bought a carton of 12 eggs, with an average weight of 61.5 g.

Question. Is such a discrepancy between sample mean and population mean purely due to randomness or significant evidence against the claim?

The formal procedure of hypothesis testing

First, we set up the following hypothesis test:

$$H_0 : \mu = 65 \quad \text{vs} \quad H_1 \, (or \; H_a) : \mu \neq 65$$

in which

- H_0 : **null hypothesis** (statement which we intend to reject)
- H_1 : **alternative hypothesis** (statement we suspect to be true)

The goal is to make a decision, based on a random sample $X_1, \ldots, X_n$ from the population, whether or not to reject H_0.

Hypothesis testing

There are two kinds of decisions:

- If the sample "strongly" contradicts H_0, then we reject H_0 and correspondingly accept H_1;
- If the sample "does not strongly" contradict H_0, then we fail to reject H_0, or equivalently we retain H_0.

Remark. This is essentially a proof by contradiction approach.

Hypothesis testing

Remark. There is a perfect analogy to courtroom trial. In this scenario, the following two hypotheses are tested:

- H_0: *Defendant is innocent*;
- H_a: *Defendant is guilty*.

The prosecutor presents evidence to the court, examined by the jury:

- If the jury thinks the evidence is strong enough (significant), the defendant will be convicted (H_0 is rejected and H_a is then accepted);
- Otherwise, the defendant is not found guilty and will be acquitted (the prosecutor has thus failed to convict the defendant due to insufficient evidence).

Hypothesis testing

Remark. It is also possible to use a one-sided alternative:

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu < 65.$$

In this case, the null is understood as "μ is *at least* 65 ($\mu \geq 65$)".

For example, the FDA's main interest is to know whether the eggs are lighter than 65 g (on average). It is not an issue if they are actually heavier (good for customers).

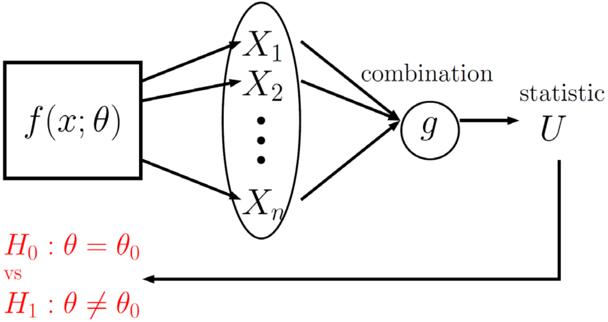
Similarly, for some other consideration, we might want to test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu > 65,$$

where the null is understood as "μ is *at most* 65 ($\mu \leq 65$)".

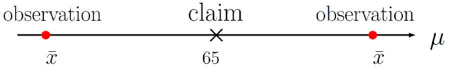
Test statistic

Typically, a test statistic needs to be specified to assist in making a decision. It is often a point estimator for the parameter being tested.

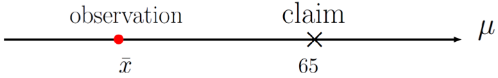


In the brown egg example, we can use $\bar{X}$ as a test statistic to test $H_0 : \mu = 65$ against

- $H_1 : \mu \neq 65$: "very small or large" values of $\bar{X}$ are evidence against the null and correspondingly in favor of the alternative hypothesis.



- $H_1 : \mu < 65$: only "very small" values of $\bar{X}$ are evidence against the null and correspondingly in favor of the alternative hypothesis.

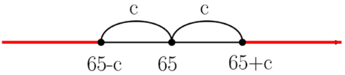


Decision rules

Clearly, a rule needs to be specified in order to decide when to reject the null $H_0 : \mu = 65$. This also defines a rejection region for the test.

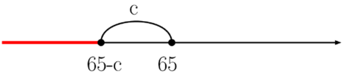
- For $H_1 : \mu \neq 65$:

  $|\bar{x} - 65| > c$



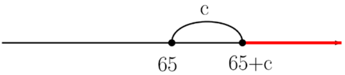
- For $H_1 : \mu < 65$:

  $\bar{x} < 65 - c$



- For $H_1 : \mu > 65$:

  $\bar{x} > 65 + c$



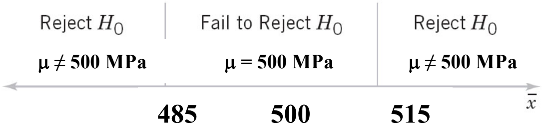
Tests of Statistical Hypotheses

- Consider the steel strength problem:
  - H_0: $\mu = 500$ MPa and H_1: $\mu \neq 500$ MPa
- Suppose a sample of $n = 10$ specimens is tested

- Decision Rule Based on $\bar{x}$:
  - If $\bar{x}$ falls close to the hypothesized value of 500 MPa:
    - This does not conflict with the null hypothesis H_0
    - We do not have enough evidence to reject H_0.
  - If $\bar{x}$ is considerably different from the hypothesized value of 500 MPa:
    - This provides evidence in support of the alternative hypothesis H_1.
    - We may reject the null hypothesis H_0.

Tests of Statistical Hypotheses

- The sample mean can take on many different values
- Suppose that if $485 \le \bar{x} \le 515$, we will not reject H_0, otherwise we will reject H_0 in favor of H_1



- The central region is called acceptance region
- The values of $\bar{x}$ that are < 485 and > 515 constitute the critical region for the test
- The boundaries are called critical values

Test errors

There are two kinds of test errors depending on whether H_0 is true or not.

		Decision	
		Retain H_0	Reject H_0
H_0	true	Correct decision	Type I error
	false	Type II error	Correct decision

Remark. In the courtroom trial scenario, a type I error is convicting an innocent person, while a type II error is acquitting a guilty person.

- Rejecting the null hypothesis H_0 when it is true is defined as a type I error

- Failing to reject the null hypothesis H_0 when it is false is defined as a type II error

- Because our decision is based on RVs, we can associate probabilities with the *type I* and *type II* errors
  - The probability of making a type I error is denoted by α
  - The probability of making a type II error is denoted by β

Type I Error

- It may happen that, even though the true mean strength is 500 MPa, we could select a RS that gives us a sample mean $\bar{x}$ that falls into the critical region

Type II Error

- It may happen that, even though the true mean strength is not 500 MPa, we could select a RS that gives us a sample mean $\bar{x}$ that falls into the acceptance region

Calculating the type-I error probability

Example 0.1. In the brown eggs problem, suppose the true population standard deviation is $\sigma = 2$ grams. A person decides to use the following decision rule (for a sample of size $n = 12$, i.e., a carton of eggs)

$$|\bar{x} - 65| > 1 \qquad \leftarrow \text{rejection region of the test}$$

to conduct the two-sided test

$$H_0 : \mu = 65 \qquad \text{vs} \qquad H_1 : \mu \neq 65.$$

What is the probability α of making a type-I error? (Answer: 0.0836)

THE UNIVERSITY OF CALIFORNIA

DAVIS

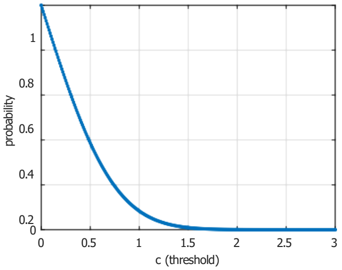
Example 0.2. (cont'd) Consider two different decision rules:

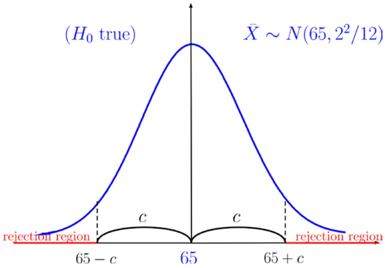
- $|\bar{x} - 65| > 0.5$
- $|\bar{x} - 65| > 2$

for conducting the same two-sided test. Verify that the corresponding probabilities of making a type-I error are 0.3844, 0.0006, respectively.

Hypothesis testing

Type-I error probabilities of tests with $|\bar{x} - 65| > c$ as rejection regions:





Observation: The larger the threshold (c), the smaller the rejection region (the less often we reject H_0), the smaller the type-I error probability.

Hypothesis testing

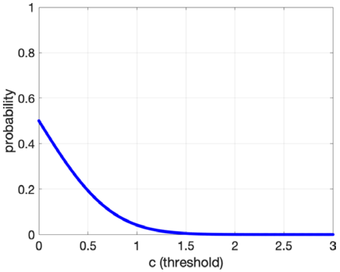
Example 0.3. Compute the probability of making a type-I error for the one-sided test $H_1 : \mu < 65$ with each of the following decision rules

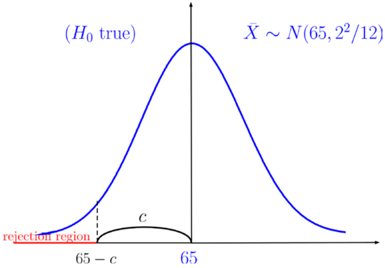
- $\bar{x} < 65 - 0.5 = 64.5$
- $\bar{x} < 65 - 1 = 64$
- $\bar{x} < 65 - 2 = 63$

(Answers: 0.1922, 0.0418, 0.0003)

Hypothesis testing

Type-I error probabilities of tests with $\bar{x} < 65 - c$ as rejection regions:





Similarly, the type-I error probability decreases as the threshold (c) is increased.

Hypothesis testing

- Too easy, too good?
- It seems that by increasing the threshold c (which would shrink the rejection region), we can make the type-I error probability arbitrarily small.
- This seems a bit too easy and too good to be true.
- This is indeed true, as far as only type-I error is concerned, but is it perhaps at the expense of something else?

Hypothesis testing

How is the type-II error affected?

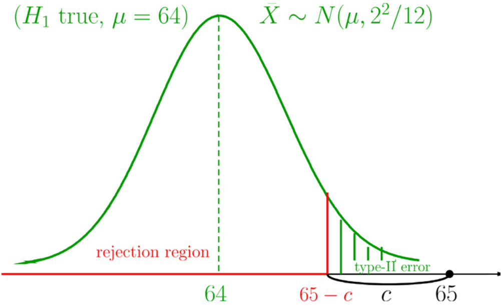
It turns out that reducing the rejection region will cause the probability of making a type-II error to increase:

- Making it hard to reject H_0 is good when H_0 is true (by using a small rejection region, this corresponds to type-I errors).
- But it would be bad when H_0 is false (we actually want to reject H_0 in this case).

The thing is that we don't know which hypothesis is true, so we have to choose a rejection region carefully such that both errors are small.

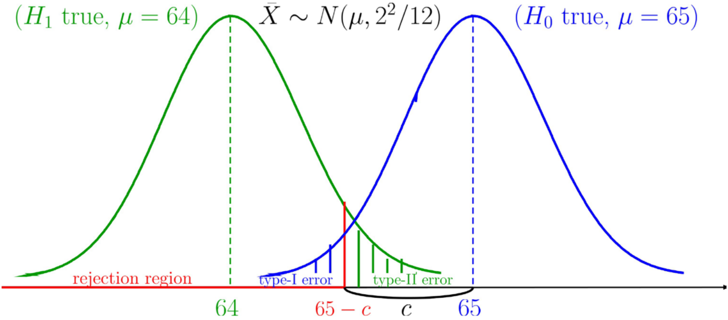
Illustration for a one-sided test when H_1 is true with $\mu = 64$

$$H_0 : \mu = 65 \text{ vs } H_1 : \mu < 65$$



Decision rule: $\bar{x} < 65 - c$

$$H_0 : \mu = 65 \quad \text{vs} \quad H_1 : \mu < 65$$



Decision rule: $\bar{x} < 65 - c$

Calculating the type-II error probabilities

Consider first the one-sided test

$$H_0 : \mu = 65 \qquad \text{vs} \quad H_1 : \mu < 65.$$

When $H_0 : \mu = 65$ is false (H_1 is correspondingly true), μ could be 64, or 63, or any other value contained by H_1.

For any fixed test with decision rule $\bar{x} < 65 - c$ (c given), the probability of making a type-II error depends on the true value of μ:

$$\beta(\mu) = P\,(\text{Fail to reject } H_0 \mid H_0 \text{ false}) = P\,(\bar{X} > 65 - c \mid H_1 \text{ true})$$

Thus, there is a separate type-II error probability at each μ in H_1.

Hypothesis testing

Remark.

- $1 - \beta(\mu)$ is the probability of making a correct decision by rejecting
  H_0 when it is false:

$$1 - \beta(\mu) = P\,(\text{Reject } H_0 \mid H_0 \text{ false}) = P\,(\bar{X} < 65 - c \mid H_1 \text{ true})$$

- It is called the power of the test (at μ).

- We would like

  – the type-II error probability $\beta(\mu)$ for a given μ to be small, and

  – the power of the test at the given μ to be large (80% or bigger).

Hypothesis testing

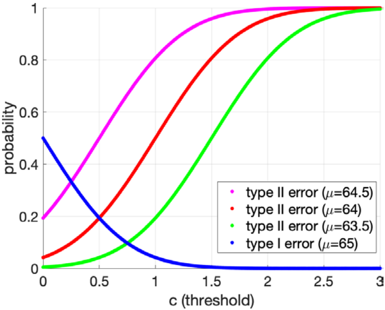
We demonstrate here how to find $\beta(64)$, the probability of making a type-II error when $\mu = 64$, by the following decision rules:

$$\bar{x} < 65 - c$$

By definition,

$$\begin{aligned}
\beta(64) &= P(\bar{X} > 65 - c \mid \mu = 64) \\
&= P\left(\frac{\bar{X} - 64}{2/\sqrt{12}} > \frac{(65 - c) - 64}{2/\sqrt{12}} \,\middle|\, \mu = 64\right) \\
&= P(Z > \sqrt{3}\,(1 - c)) = 1 - \Phi(\sqrt{3}(1 - c)) = \begin{cases} 0.1922, & c = 0.5 \\ 0.5, & c = 1 \\ 0.9582, & c = 2 \end{cases}
\end{aligned}$$

What about other values of c (and also other values of μ)?



Observations on the type-II errors (type-I error probability decreases as c increases):

- **For fixed value μ:** the larger c (the smaller the rejection region, and thus the harder to reject H_0), the larger the type-II error.

- **For fixed test (c):** the closer μ is to the value in H_0 (65), the larger the type II error.

Type-II error probabilities for two-sided tests can be computed similarly, but the process is a little harder.

Example 0.4. Consider the two-sided test:

$$H_0 : \mu = 65 \qquad \text{vs} \quad H_1 : \mu \neq 65$$

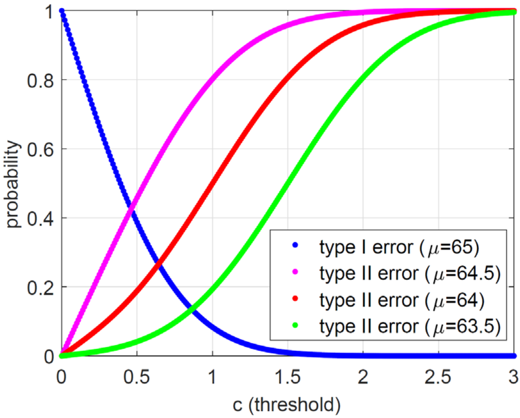
along with the following decision rule:

$$|\bar{x} - 65| > c.$$

Find the probability of making a type-II error when $\mu = 64$ for each value of $c = 0.5,\ 1,\ 2$.

(Answer: $\beta(64) = P\,(|\bar{X} - 65| < c \mid \mu = 64)$,0.9582 ,0.4997 ,$0.1875 =$ which has the same trend as c increases)

Hypothesis testing



How to control both errors together

Previously we assumed that both sample size n and test threshold c are fixed so as to evaluate the type-I and type-II errors of the test

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_a : \mu < \mu_0 \text{ (or } \mu \neq \mu_0\text{)}$$

Here we consider the inverse design problem by assuming the two types of error probabilities are given first:

- type-I error probability α (called level of the test) $\leftarrow$ typically 5%
- type-II error probability β (at a specified value μ^t) $\leftarrow$ typically 20%

and then trying to determine the required values of c and n as follows:

Hypothesis testing

1. For the given level of the test i.e., α, solve

$$
\begin{aligned}
\alpha &= P(\text{Reject } H_0 \mid H_0 \text{ true}) \\
&= P(\bar{X} < \mu_0 - c \mid \mu = \mu_0) \\
&= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -\frac{c}{\sigma/\sqrt{n}} \,\middle|\, \mu = \mu_0 \right) \\
&= P\left(Z < -\frac{c}{\sigma/\sqrt{n}} \right) \quad \longrightarrow \frac{c}{\sigma/\sqrt{n}} = z_\alpha
\end{aligned}
$$

This yields that $c = z_\alpha \frac{\sigma}{\sqrt{n}}$. That is, a level test for H_a: $\mu < \mu_0$ (for a fixed sample size n) is

$$
\bar{x} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}, \quad \text{or equivalently,} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha
$$

Hypothesis testing

2. For the choice of , choose sample size n to achieve type-II error probability β at an alternative value $\mu = \mu'$:

$$\begin{aligned}\beta &= P(\text{Fail to reject } H_0 \mid H_0 \text{ false}) \\ &= P(\bar{X} > \mu_0 - c \mid \mu = \mu') \\ &= P\left(\frac{\bar{X} - \mu'}{\sigma/\sqrt{n}} > \frac{\mu_0 - c - \mu'}{\sigma/\sqrt{n}} \,\middle|\, \mu = \mu'\right) \\ &= P\left(Z > -z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

This yields that

$$z_\beta = -z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}, \text{ and thus, } n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right)^2.$$

Hypothesis testing

Example 0.5. Assume the setting of the brown eggs example (with known $\sigma = 2$, but sample size n TBD). Consider the following one-sided test

$$H_0 : \mu = 65 \qquad \text{vs} \qquad H_a : \mu < 65$$

with corresponding decision rule

$$\bar{x} < 65 - c$$

Choose n, c so that the test has level 5% and power 80% (at $\mu = 64$).

Answer : $\quad c = z_\alpha \dfrac{\sigma}{\sqrt{n}} = 0.658, \quad n = \left(\dfrac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right)^2 = 25.$

Hypothesis testing

Remark. For a two-sided test such as

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_a : \mu \neq \mu_0$$

with corresponding decision rule

$$|\bar{x} - \mu_0| > c$$

the two equations (for determining n, c) become

$$\alpha = P\,(\text{Reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ true}) = P\,(|\bar{X} - \mu_0| > c \mid \mu = \mu_0)$$

$$\beta = P\,(\text{Fail to reject } \mathrm{H}_0 \mid \mathrm{H}_0 \text{ false}) = P\,(|\bar{X} - \mu_0| < c \mid \mu = \mu')$$

Hypothesis testing

The first equation has an exact solution

$$c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

but the second equation only has an approximation solution:

$$n \approx \left(\frac{\sigma(z_{\alpha/2}+z_{\beta})}{\mu_0-\mu'}\right)^2.$$

The corresponding level α test is

$$|\bar{x}-\mu_0| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \text{or equivalently,} \left|\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}$$

Hypothesis testing

Example 0.6. Redo the preceding example but instead for a two-sided test

$$H_0 : \mu = 65 \qquad \text{vs} \qquad H_a : \mu \neq 65$$

with decision rule

$$|\bar{x} - 65| > c$$

Answer :

$$c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.693, \quad n \approx \left(\frac{\sigma(z_{\alpha/2} + z_{\beta})}{\mu_0 - \mu'}\right)^2 = 32$$

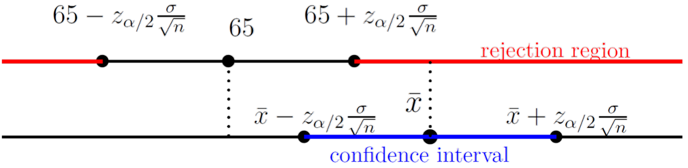
Connection to confidence intervals

In the last example, the rejection region of the two-sided test at level α is

$$|\bar{x} - 65| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

which is equivalent to

$$65 \notin \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{(CI)}$$



Hypothesis testing

That is, we reject the null at level α if and only if the $1 - \alpha$ confidence interval fails to capture the claimed value 65.

There is a similar connection between one-sided tests and one-sided confidence intervals: We reject the null at level α if and only if 65 is outside the one-sided confidence interval at level α:

$$\bar{x} < 65 - z_\alpha \frac{\sigma}{\sqrt{n}} \quad \Longleftrightarrow \quad 65 \notin \left(-\infty,\, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

Hypothesis testing

One can thus use a 1- or 2-sided $1 - \alpha$ confidence interval to conduct the corresponding hypothesis test at level α:

- Confidence interval captured $\mu = 65$: Do not reject H_0
- Confidence interval failed to capture $\mu = 65$: Reject H_0

Note the relationship between and interpretation of:

$1 - \alpha$ (confidence level) and α (level of the test).

Summary

A hypothesis test has the following components:

- Population: e.g., all brown eggs produced by the farm, whose weights have a normal distribution with unknown mean μ but known variance σ^2
- Null and alternative hypotheses: $H_0{:}\ \mu = \mu_0$ vs $H_a{:}\ \mu \neq \mu_0$;
- Random sample from the population: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$
- Test statistic: e.g., $\bar{X}$
- Decision rule (based on a specified rejection region):
- $|\bar{x} - \mu_0| > c$

Hypothesis testing

Evaluation of the test:

- **Type-I error:**

$$\alpha = P\,(\text{Reject H}_0 \mid \text{H}_0 \text{ true}) = P\,(|\bar{X} - \mu_0| > c \mid \mu = \mu_0)$$

If α is specified first as the level of the test, then set $c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

(or for $c = z_{\alpha}\frac{\sigma}{\sqrt{n}}$ a one-sided test)

- **Type-II errors (at a given $\mu = \mu'$)**

$$\beta = P\,(\text{Fail to reject H}_0 \mid \text{H}_0 \text{ false}) = P\,(|\bar{X} - \mu_0| < c \mid \mu = \mu')$$

To control both errors, we first choose c (dependent on n) to attain level α, then choose sample size n to achieve power $1 - \beta$ at μ':

Hypothesis testing

When σ^2 is *known*, a level α test for μ is

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$:

$$\text{Reject } H_0 \text{ if and only if } |\bar{x} - \mu_0| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$:

$$\text{Reject } H_0 \text{ if and only if } \bar{x} - \mu_0 < -z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

- $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$:

$$\text{Reject } H_0 \text{ if and only if } \bar{x} - \mu_0 > z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Hypothesis testing

To achieve a type-II error probability of β at an alternative value μ', the required sample size is

- for the two-sided test:

$$n \approx \left(\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'} \right)^2$$

- for both one-sided tests:

$$n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'} \right)^2$$