2.1.3 Significance of Parameters

Let say I want to buy a new home. However, before buying it, I want to have a general idea of how much the number of square meters affects the price of a house. To this end, I model the relationship between the price of a house and its ground surface as linear, i.e.:

2.1.11

$$y = wx + b$$

where x is the number of square meters and y is the price.
However, I do not have access to the actual coefficients w and b (the ground truth). However, I can visit several houses, and for the i-th house, measure the number of square meters x_i, and ask the owners for the price y_i. The relationship between x_i and y_i can be modeled as

$$y_i = wx_i + b + \epsilon_i, \quad i = 1, \ldots, N$$

where ϵ_i denotes measurement or model error, which in this case might be influenced by several factors such as:

- how old the house is
- how nice the neighborhood is
- how greedy the owner is
- with which amount of precision the ground surface was measured
- etc

I can then estimate the parameters w and b using linear regression. The corresponding least squares problem is

$$\min_{w,b \in \mathbb{R}} \frac{1}{2} \sum_{i=1}^{N} |wx_i + b - y_i|^2$$

and we know that the solution ($\hat{w}$, $\hat{b}$) to this problem is

2.1.13

$$\hat{w} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x}^2)}$$

$$\hat{b} = \overline{y} - \hat{w}\overline{x}$$

To save time, I visit 10 houses and ask a good friend of mine to visit 10 others. Then we each compute the linear regression corresponding to our own data, and draw the curve we each obtain

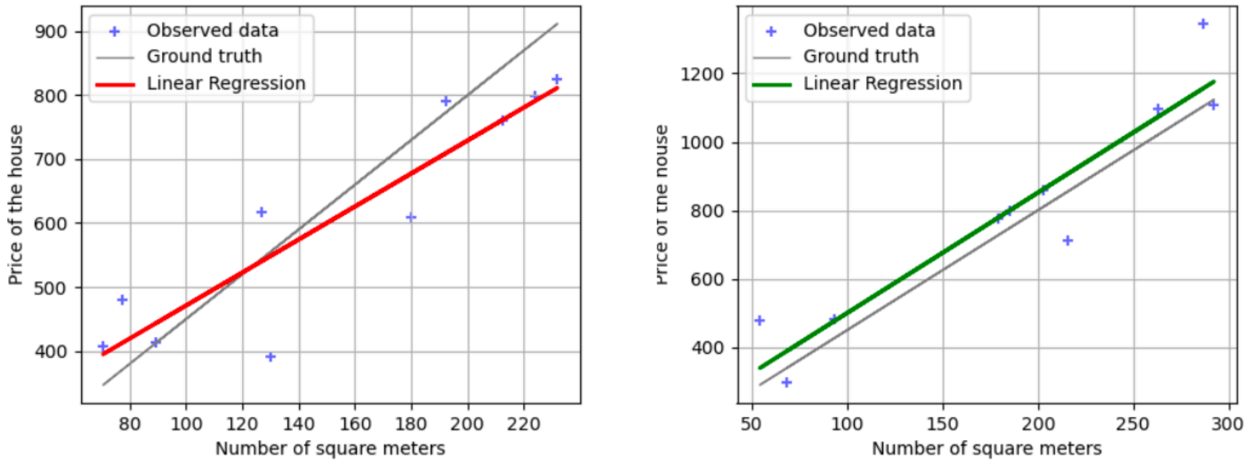


Figure 2.1: Linear regression curve I computed on the data I observed(left, in red) vs. the curve my friend obtains (right, in green). The ground truth (in grey) is unknown.

Surprisingly, the curves we obtain are pretty different. I think that this can be linked to different factors, like the few number of data points (x_i, y_i) my friend and I considered (called the sample size) or the amount of noise ϵ_i. In order to get another opinion, I ask a real estate agency for their data. They know the surface and price of over 100 different houses, and used this data to compute the linear regression showed below.

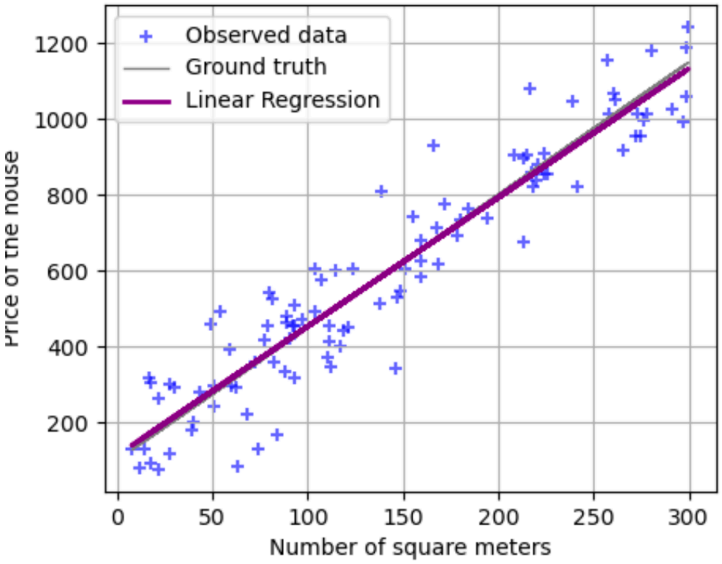


Figure 2.2: Linear regression curve the real estate agency computed on 100 samples

As we can see, this curve is different from the previous ones

Question

Since I do not know the actual values of w and b, which values of $\hat{w}$ and $\hat{b}$ should I trust more? Mine, my friend's or the one of the real estate agency?
This type of question is at the heart of *statistical tests* and forms the basis of modern scientific method. It is used in areas like pharmacology, epidemiology, sociology, psychology, finance, particle physics, etc.

Reminder of probability theory

Definition 2.1.2 (Random Variable)

Let $(\Omega, \mathbb{P})$ be a probability space. A random variable X is a function from Ω to $\mathbb{R}$. It assigns a real number to each outcome in probability space
Purpose: It translates abstract outcomes (e.g. "heads", "tails") into numerical values (e.g. 1 for heads, 0 for tails) so we can analyze probabilities mathematically

Example:

- Let $\Omega = \{\text{rain}, \text{snow}, \text{sunny}\}$
- Define $X(\text{rain}) = 0, X(\text{snow}) = 1, X(\text{sunny}) = 2$
- X is a random variable representing weather numerically

Definition 2.1.3 (Cumulative distributive function)

Let X be a random variable. Its cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ is defined as

$$F_X(x) := \mathbb{P}(X \leq x)$$

It gives the probability that X takes a value **less than or equal to** x

- $F_X(x)$ is non-decreasing
- $\lim_{x\to-\infty} F_X(x) = 0, \lim_{x\to+\infty} F_X(x) = 1$

Example
For a fair 6-sided die:

- $F_X(3) = \mathbb{P}(X \leq 3) = \frac{3}{6} = 0.5$
- $F_X(6) = \mathbb{P}(X \leq 6) = 1, F_X(0) = 0$

Definition 2.1.4 (Probability density function)

Let X be a random variable and suppose that F_X is differentiable. The probability density function is defined as $f_X = F_X'$
What it is: The derivative of the CDF:

$$f_X(x) = \frac{d}{dx}F_X(x)$$

Purpose: Describes the "density" of probability at a point x for **continuous** random variables

- Probabilities are calculated by integrating the PDF

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Example:
For a normal distribution, the PDF is the bell-shaped curve

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Remark 2.1.3 Suppose that X can only take values in a discrete set $\{x_i\}_{i\in\mathbb{N}} \subset \mathbb{R}$.
Then X is called a discrete random variable. A random variable that can take all values in $\mathbb{R}$ is called a *continuous random variable*
Discrete Random Variables
X takes a value in a countable set $\{x_i\}$
Instead of a PDF, discrete variables use a PMF: $\mathbb{P}(X = x_i) = p_i$
e.g. $\mathbb{P}(X = 1) = \frac{1}{6}$ for a fair die

Continuous Random Variables
X can take uncountably many values in $\mathbb{R}$
CDF has a smooth and differentiable function

Definition 2.1.5 (Expectation)

Let X be a continuous random variable with probability density function f_X. The expectation of X is defined as

$$\mathbb{E}(X) := \int_{\mathbb{R}} xf_X(x)dx$$

If X is a discrete random variable, its expectation is defined as

$$\mathbb{E}(X) := \sum_{i\in\mathbb{N}} x_i\mathbb{P}(X = x_i)$$

Exercise 2.1.1

Let X, Y be random variables and $a \in \mathbb{R}$. Show that

$$\mathbb{E}(aX + Y) = a\mathbb{E}(X) + \mathbb{E}(Y)$$

$$\mathbb{E}(aX + Y) = \mathbb{E}(aX) + \mathbb{E}(Y)$$

$$= a\mathbb{E}(X) + \mathbb{E}(Y)$$

Definition 2.1.6 (Variance and covariance)

The variance of a random variable X is defined as

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2]$$

or

$$\mathrm{Var}(X) := \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

The covariance of two random variables X, Y is defined as

$$cov(X, Y) := \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$$

Exercise 2.1.2

Show that

$$\mathbb{V}(X+Y) = V(X) + V(Y) + 2cov(X, Y)$$

Step 1:

$$V(X+Y) = \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2$$

Step 2: Look at $\mathbb{E}((X+Y)^2)$ and expand $(X+Y)^2$

$$(X+Y)^2 = X^2 + 2XY + Y^2$$

-->

$$\mathbb{E}((X+Y)^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2)$$

Step 3: Look at second part $(\mathbb{E}(X+Y))^2$ and expand $\mathbb{E}(X+Y)$
The expectation of a sum is the sum of expectations

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

So:

$$(\mathbb{E}(X+Y)^2) = (\mathbb{E}(X) + \mathbb{E}(Y))^2 = \mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2$$

Step 4: Substitute into variance formula

$$V(X+Y) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - (\mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2)$$

Rearranging gets us to

$$= (\mathbb{E}(X^2) - \mathbb{E}(X)^2) + (\mathbb{E}(Y^2) - \mathbb{E}(Y)^2) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y))$$

Recognize Variance and Covariance

$V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$,
$V(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$
$cov(x, y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

Substituting this gets us to

$$V(X + Y) = V(X) + V(Y) + 2cov(X, Y)$$

Definition 2.1.7 (Independence)

We say that two random variables X, Y are independent if for all $U, V \subset \mathbb{R}$

$$\mathbb{P}(X \in U, Y \in V) = \mathbb{P}(X \in U)\mathbb{P}(Y \in V)$$

Proposition 2.1.5 (Independance and covariance)

If X and Y are independent then

$$cov(X, Y) = 0$$

Remark 2.1.4 The converse is not true in general!!

Example 2.1.2 (Common distributions)

1. Let $p \in [0, 1]$. We say that X is a Bernoulli random variable (denoted $X \sim \mathcal{B}(p)$) if and only if

   $$\mathbb{P}(X = 1) = p$$

   and

   $$\mathbb{P}(X = 0) = 1 - p$$

   We have in this case:

   $$\mathbb{E}(X) = p$$

   and

   $$V(X) = p(1 - p)$$

   For $p = 1/2$, this can for instance model the behavior of a coin flip.
2. We say that a continuous random variable X is uniformly distributed on $[0, 1]$, and denote $X \sim Unif([0, 1])$, when the probability distribution function is $f_X(x) = 1$. In this case, we have

   $$\mathbb{E}(X) = \frac{1}{2}$$

   and

$$V(X) = \frac{1}{12}$$

3. Let $\mu, \sigma \in \mathbb{R}$. We say that X follows a Gaussian probability distribution (denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$) when the probability density function of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We have in this case:

$$\mathbb{E}(X) = \mu$$

and

$$V(X) = \sigma^2$$

Statistical significance

In this section we consider the even simpler model of only estimating the mean of the observed data, i.e., we use the model

2.1.14

$$y_i = b + \epsilon_i, \quad i = 1, \dots, N$$

A real-world application might be the following: suppose that you run an online store, and you want to change the design of your website from its original design A to a design B. You want that this change in design to have some impact on the amount your customers usually spend on your website. You already know that when presented with the original design A, a user usually spends in average b_A euros. You want to estimate the amount b_B an average user would spend when presented the design B. In order to do so, you present the new design B to a sample of N random users, and model the amount y_i a user spend by

2.1.15

$$y_i = b_B + \epsilon_i, \quad i = 1, \dots, N$$

We want to estimate the value of b_B and compare it to b_A. If $b_B \neq b_A$, we will definitely use the design B. If $b_B = b_A$, we keep the design A.
The hypothesis we are testing is $b_B = b_A$. This is what we call the **null hypothesis**, and is usually denoted by H_0. The alternative hypothesis (usually, the one we want to prove: here, $b_A \neq b_B$) is denoted by H_1. The idea is to reject the null hypothesis in such a way that the probability of wrongly rejecting it is very small.
However, we do not have access to the true value of b_B. The best we can do is to estimate it by its empirical mean obtained by least square approximation

$$\hat{b}_B = \frac{1}{N}\sum_{i=1}^{N} y_i$$

The way we will test if H_0 is likely to be true or false is to quantify how probable it is to observe the empirical value $\hat{b}_B$ under the assumption that H_0 is true. If this probability is small, then we will reject H_0. In order to do so, we need to use a test statistic, i.e. a random variable for which the distribution is known, provided that H_0 is true. To derive the test statistic, we make the strong assumptions that ϵ_i are independent and normally distributed random variables with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and that we know their variance σ^2. In this case, the y_i are i.i.d. (independent and identically distribitured) according to $\mathcal{N}(b_B, \sigma^2)$ and therefore $\hat{b}_B \sim \mathcal{N}(b_B, \frac{\sigma^2}{N})$
The test statistic in this case

$$T := \frac{\hat{b}_B - b_a}{\sqrt{V(\hat{b}_B)}} = \frac{\sqrt{N}}{\sigma}(\hat{b}_B - b_a)$$

Under the null hypothesis $H_0 := (b_B = b_A)$, we see that

2.1.16

$$T = \frac{\sqrt{N}}{\sigma}(\hat{b}_B - b_a)$$

and we can show that $T \sim \mathcal{N}(0, 1)$

Exercise 2.1.3

Show that we indeed have

$$\mathbb{E}(T) = 0$$

and

$$V(T) = 1$$

Step 1: Expected Value of T

We want to show $\mathbb{E}(T) = 0$.

1. **Expectation of $\hat{b}_B$:**

   Since $y_i = b_B + \epsilon_i$, the expected value of y_i is:

   $$\mathbb{E}(y_i) = b_B$$

   Therefore, the expectation of $\hat{b}_B$ (the sample mean) is:

   $$\mathbb{E}(\hat{b}_B) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(y_i) = b_B$$

2. **Under H_0:**

   If H_0 is true ($b_B = b_A$), then:

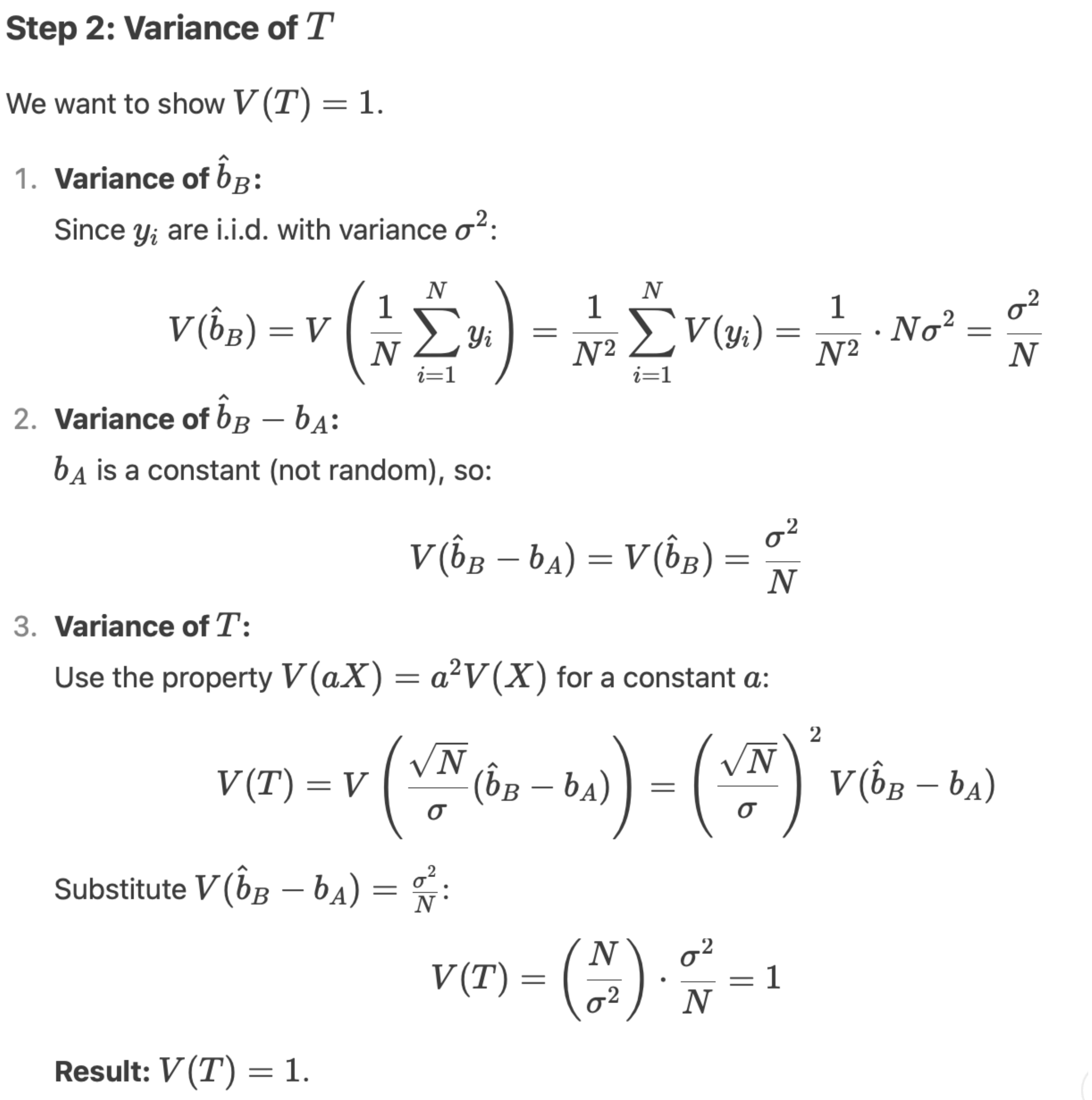
   $$\mathbb{E}(\hat{b}_B) = b_A$$

3. **Expectation of T:**

   Substitute $\mathbb{E}(\hat{b}_B) = b_A$ into T:

   $$\mathbb{E}(T) = \mathbb{E}\left(\frac{\sqrt{N}}{\sigma}(\hat{b}_B - b_A)\right) = \frac{\sqrt{N}}{\sigma}\left(\mathbb{E}(\hat{b}_B) - b_A\right) = \frac{\sqrt{N}}{\sigma}(b_A - b_A) = 0$$

   Result: $\mathbb{E}(T) = 0$.



Step 2: Variance of T

We want to show $V(T) = 1$.

1. **Variance of $\hat{b}_B$:**
   Since y_i are i.i.d. with variance σ^2:

$$V(\hat{b}_B) = V\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} V(y_i) = \frac{1}{N^2} \cdot N\sigma^2 = \frac{\sigma^2}{N}$$

2. **Variance of $\hat{b}_B - b_A$:**
   b_A is a constant (not random), so:

$$V(\hat{b}_B - b_A) = V(\hat{b}_B) = \frac{\sigma^2}{N}$$

3. **Variance of T:**
   Use the property $V(aX) = a^2V(X)$ for a constant a:

$$V(T) = V\left(\frac{\sqrt{N}}{\sigma}(\hat{b}_B - b_A)\right) = \left(\frac{\sqrt{N}}{\sigma}\right)^2 V(\hat{b}_B - b_A)$$

   Substitute $V(\hat{b}_B - b_A) = \frac{\sigma^2}{N}$:

$$V(T) = \left(\frac{N}{\sigma^2}\right) \cdot \frac{\sigma^2}{N} = 1$$

   Result: $V(T) = 1$.

Remark 2.1.5

If the null hypothesis is false, i.e. $b_B \neq b_A$, then we can show that

$$T \sim \mathcal{N}\left(\frac{\sqrt{N}}{\sigma}(b_B - b_a), 1\right)$$

Hence, with large N and small σ, we see that T "deviates" more and more from 0.
We then choose a number $\alpha \in (0,1)$, called the statistical significance, which has the property that

$$\mathbb{P}(H_0 \text{ is rejected } | H_0 \text{ is true}) \leq \alpha$$

i.e. the probability of wrongly rejecting H_0 is smaller than α. Usually, α is taken equal to 0.05, 0.01 or much lower depending on the field of study
The last thing we need is to actually define what it means to reject H_0. To this purpose, we need to define a *rejection rule*.
Using Equation 2.1.16, we see that T is very likely to be close to 0 when H_0 is true. Hence, we decide to reject H_0 when T is far enough from 0. However, T being a random variable, we decide to reject H_0 by choosing $M > 0$ such that

$$\mathbb{P}(|T| > M | H_0 \text{ is true}) \leq \alpha$$

When H_0 is true, we know that $T \sim \mathcal{N}(0,1)$. Hence,

$$\begin{aligned}\mathbb{P}[\text{H0 is rejected} \mid \text{H0 is true}] &:= \mathbb{P}[|T| \geq M \mid \text{H0 is true}] \\ &= \mathbb{P}_{t \sim \mathcal{N}(0,1)}[|t| \geq M]\end{aligned}$$

In the specific case for $\alpha = 0.02$, we can show that taking $M \approx 2.33$ guarantees that $\mathbb{P}_{t \sim \mathcal{N}(0,1)}[|t| \geq M]$
Consequently, we reject the null hypothesis if $|T| \geq 2.33$ (or equivalently $|\hat{\beta} - b_A| \geq 2.33 \frac{\sigma}{\sqrt{N}}$) and declare the value of $\hat{b}_B$ significant at the level α
Important It is incorrect to conclude that in this case the probability of the null hypothesis being true is 2%!
A quantity similar to the significance level is the so-called p-value, which in this specific case is defined as $p(T) = \mathbb{P}_{t \sim \mathcal{N}(0,1)}[|t| \geq |T|]$. In contrast to $\alpha, p(T)$ is a random variable and is not predetermined
The p-value $p(T)$ is the probability of observing a test statistic as extreme as T under H_0. For $\alpha = 0.02$, we reject H_0 if $|T| \geq 2.33$ corresponding to $p(T) \leq 0.02$

Skipped a few things here :)

Definition 2.1.8 (Hypothesis Test)

A hypothesis test for a null hypothesis H_0 consists of a test statistic T along with a decision rule against or in favor of H_0 based on T.
In the case of a real test statistic T and a decision rule of the form $T \geq M$ (or $T \leq M$) against H_0, we refer to it as a right tailed (or left-tailed) test. A decision rule of the form $T \geq M$ and $T \leq N$ for $M, N \in \mathbb{R}$ with $M \leq N$ refers to a two-tailed test.

Next, we define the p-value. For this purpose, we restrict ourselves to simple null hypotheses H_0 (e.g., $b = 0$, but not $b \leq 0$), which uniquely determine the distribution of $T | H_0$

Definition 2.1.19 (p-value)

In the case of a real test statistic T with an (unknown) distribution μ and a simple null hypothesis H_0, the p-value of T is defined as

- $p(T) := \mathbb{P}_{t\sim\mu}[t \geq T|H_0]$ for right-tailed tests
- $p(T) := \mathbb{P}_{t\sim\mu}[t \leq T|H_0]$ for left-tailed tests
- $p(T) :=. 2\min\{\mathbb{P}_{t\sim\mu}[t \leq T|H_0], \mathbb{P}_{t\sim\mu}[t \geq T|H_0]\}$ for a two-tailed test. In the case of a distribution of T that is symmetric around zero, it holds that $p(T) = \mathbb{P}_{t\sim\mu}[|t| \geq |T|\,|H_0]$

The p-value is the probability of observing data even more "extreme" than T under the assumption that H_0 is true.

Remark 2.1.6 It is important to note that by definition, the p-value $p(T)$ of a test statistic T is a random variable. Therefore, its values can vary significantly and are not suitable for a posteriori determination of a significance level!