PROBABILITY & STATISTICS FOR LIFE SCIENCES

Type I and Type II Errors

Alpha, beta, and power

Zan Ahmad

November 11, 2026 • EN.553.211

The question behind these notes

A test makes a decision from a random sample. How often does its decision go wrong, and how often can it detect a real effect? We will first distinguish the significance level from the p-value, then interpret and calculate β, and finally understand power as the complement of a Type II error.

1 A decision is not the same as the truth

Consider testing whether a laboratory protocol increases a population mean biomarker level above 100 units:

$$H_0 : \mu = 100, \qquad H_a : \mu > 100.$$

The hypotheses concern the *population*. The test uses a *sample*. Even a correctly designed test can reach a misleading decision because different random samples give different results.

Decision	H_0 **is true**	**A specified alternative is true**
Reject H_0	Type I error	Correct detection
Fail to reject H_0	Correct nonrejection	Type II error

Type I error: a false alarm. We reject H_0 even though it is true. In this example, we find evidence of an increase when the true mean is actually 100.

Type II error: a missed effect. We fail to reject H_0 even though an alternative is true. For example, the mean might really be 105, but our sample does not give sufficiently strong evidence of an increase.

Failing to reject does not prove the null

A test that fails to reject says the data did not provide enough evidence against H_0 under the chosen rule. It does not establish that there is no effect. We may simply have missed one.

2 Read the condition before computing the probability

Recall that a conditional probability has the form

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \qquad \mathbb{P}(B) > 0.$$

The condition B tells us which outcomes form the reference group. Among those outcomes, we measure the fraction for which A also occurs.

In hypothesis testing, we use this idea to imagine repeated samples generated under a specified population model. The words after the vertical bar tell us which model to use. We are *not* assigning a probability to the hypothesis itself.

Significance level and Type I error

For the continuous tests in these notes, the rejection rule is calibrated so that

$$\boxed{\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true}).}$$

Read this as: *If the null model generated many samples, what fraction would lead us to reject it?*

We choose α before inspecting the result. At $\alpha = 0.05$, about 50 out of 1,000 repeated studies under the null would reject it, on average. Each rejection in that null world is a Type I error. This is not a guarantee of exactly 50 errors.

2.1 Use the null distribution to locate the rejection region

Suppose the biomarker observations are independent and normally distributed, with **known population SD** $\sigma = 10$ and sample size $n = 25$. Then

$$\mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}} = 2, \qquad Z = \frac{\overline{X} - 100}{2} \sim N(0,1) \quad \text{under } H_0.$$

Here $N(0,1)$ is the standard normal distribution. The course notation z_α means the critical value with *right-tail* area α:

$$\mathbb{P}(Z > z_\alpha) = \alpha, \qquad z_{0.05} \approx 1.645.$$

Large positive sample means support H_a. Our level-0.05 rule is therefore

$$\boxed{\text{Reject } H_0 \text{ if } Z > 1.645 \iff \overline{X} > 103.29.}$$

Pause and interpret. The rule has been fixed, but no sample mean has yet been observed. The significance level describes the rule's false-alarm probability under the null, not the evidence from one particular sample.

3 Alpha and the p value use the same null world

Now collect the sample and suppose its mean is $\bar{x} = 104$. Its standardized observed value is

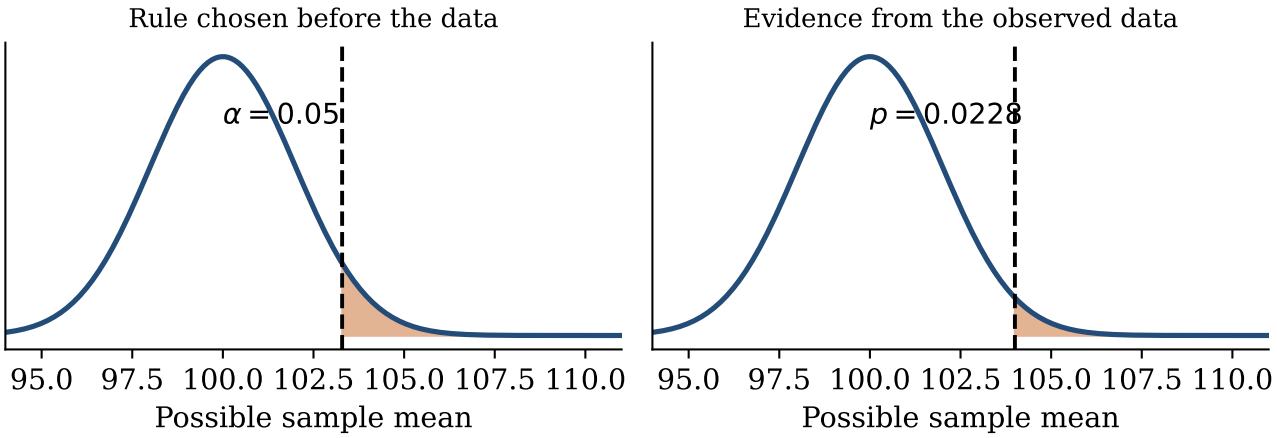
$$z_{\text{obs}} = \frac{104 - 100}{2} = 2.$$

For our right-tailed test, values at least as extreme in the direction of H_a are values at least as large as the observed one.

The p value for this right tailed test

$$\begin{aligned} p &= \mathbb{P}(\overline{X} \geq 104 \mid \mu = 100) \\ &= \mathbb{P}(Z \geq 2) = 1 - \Phi(2) \approx 0.0228. \end{aligned}$$

Here $\Phi(z) = \mathbb{P}(Z \leq z)$ is the standard-normal cumulative distribution function. We ask: *If the null were true, how often would a new sample produce a result this large or larger?*



Both pictures use the **null** sampling distribution. What changes is the boundary:

- For α, shade beyond the *prechosen rejection cutoff*, 103.29.
- For p, shade beyond the *observed sample mean*, 104.

Because $104 > 103.29$, the observed statistic is in the rejection region. Equivalently, $p = 0.0228 < 0.05 = \alpha$, so we reject H_0.

The p value is not the probability of a Type I error in this study

The p-value is not $\mathbb{P}(H_0 \text{ is true} \mid \text{data})$. Neither p nor α tells us whether this particular rejection is mistaken. A Type I error occurs only if H_0 is in fact true; the test does not reveal that truth with certainty.

For a two-sided alternative, "at least as extreme" includes both directions. For example, the two-sided normal p-value is $2[1 - \Phi(|z_{\text{obs}}|)]$.

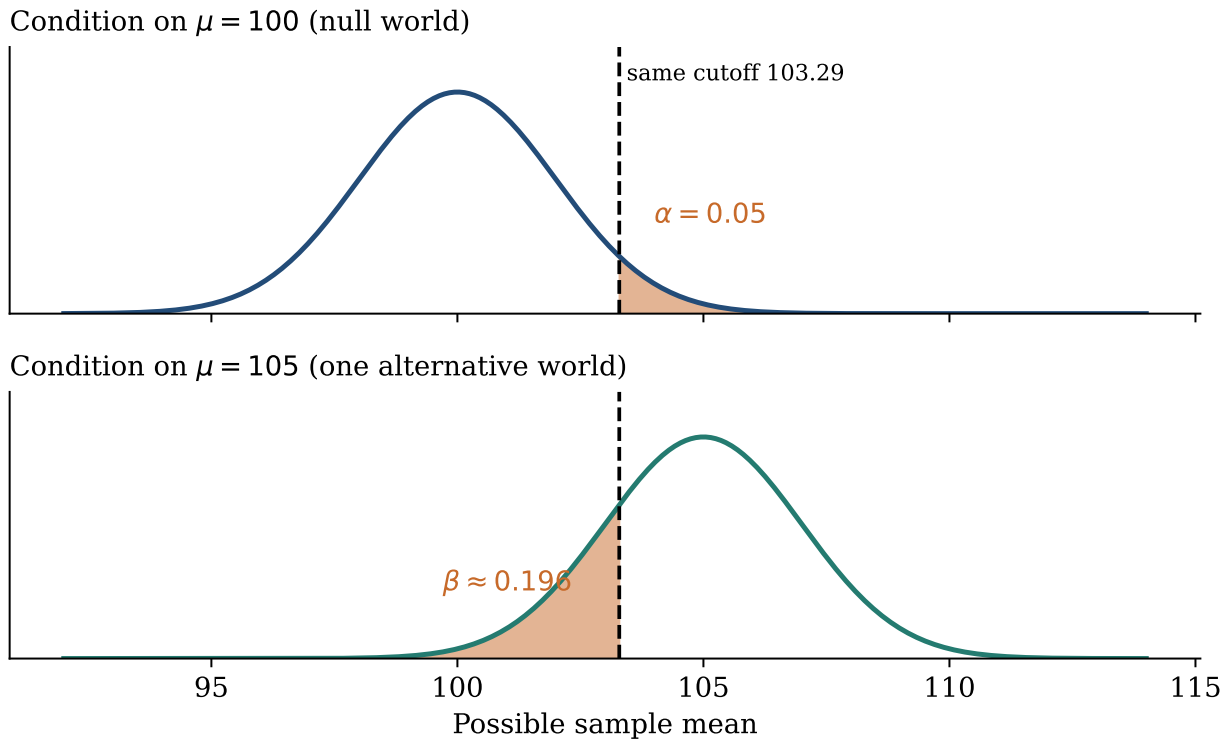
4 Beta changes the assumed truth but keeps the rule

A Type II error requires a false null. We must therefore specify an alternative value. Suppose the true mean is $\mu_1 = 105$. The decision rule is still "reject if $\overline{X} > 103.29$."

Type II error probability

$$\boxed{\beta(105, 0.05) = \mathbb{P}(\text{fail to reject } H_0 \mid \mu = 105).}$$

The course writes this more generally as $\beta(\mu, \alpha)$. It also depends on the sample size, population variability, and test being used, even when those inputs are not written in the notation.



Read the condition first. In the top picture, condition on $\mu = 100$; the shaded rejection tail is α. In the bottom picture, condition on $\mu = 105$; the shaded *nonrejection* region is β.

The vertical boundary does not move. We shift the sampling distribution because we have changed the assumed population mean. Both curves are distributions of *sample means*, not individual observations.

5 Calculate beta before defining power

Under $\mu = 105$, the sample mean has mean 105 and SD 2. Since the rule fails to reject when $\overline{X} \leq 103.29$,

$$\begin{aligned}\beta(105, 0.05) &= \mathbb{P}(\overline{X} \leq 103.29 \mid \mu = 105) \\ &= \mathbb{P}\left(\frac{\overline{X} - 105}{2} \leq \frac{103.29 - 105}{2} \,\middle|\, \mu = 105\right) \\ &= \Phi(-0.855) \approx 0.196.\end{aligned}$$

We standardize using **105**, not 100, because 105 is the true mean in this probability calculation. We still use the cutoff obtained from the null.

Among repeated studies in this alternative world, about 196 out of 1,000 would fail to reject H_0, on average. Each such failure misses the actual increase from 100 to 105.

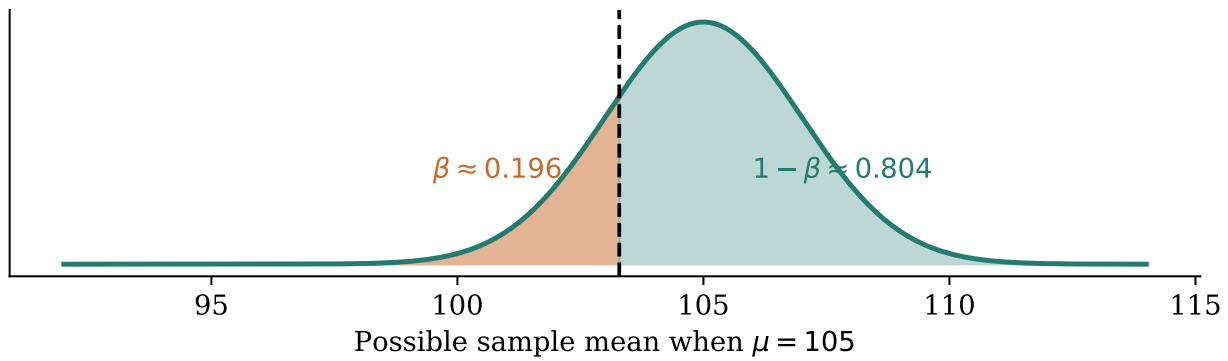
5.1 Power is the other decision under the same alternative

Once the alternative is fixed, rejection and nonrejection exhaust all possible decisions. Their probabilities sum to one.

Power to detect a specified effect

$$\begin{aligned}\text{Power at } \mu_1 &= \mathbb{P}(\text{reject } H_0 \mid \mu = \mu_1) \\ &= 1 - \beta(\mu_1, \alpha).\end{aligned}$$

For $\mu_1 = 105$, power is approximately $1 - 0.196 = 0.804$.



In this picture, both shaded regions use the *same alternative distribution*. Their areas add to one. About 804 out of 1,000 studies would detect this increase, on average.

Do not pair probabilities from different worlds

β + power = 1, but generally $\alpha + \beta \neq 1$. Alpha conditions on the null; beta conditions on a specified alternative. Also, $1 - p$ is not power.

6 A smaller effect is easier to miss

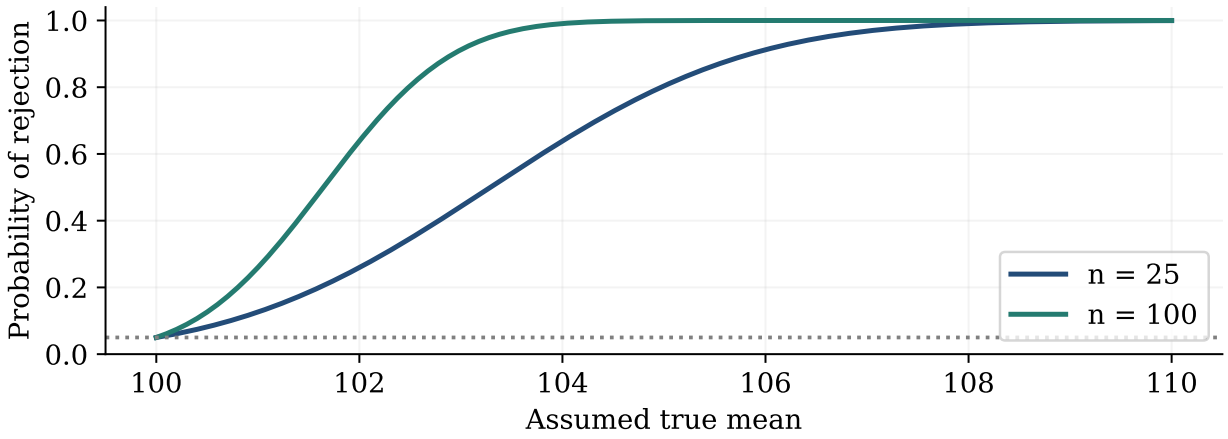
Keep $n = 25$, $\sigma = 10$, $\alpha = 0.05$, and the cutoff 103.29. Change only the assumed true mean from 105 to 103. Then

$$\beta(103, 0.05) = \Phi\left(\frac{103.29 - 103}{2}\right) = \Phi(0.145) \approx 0.558.$$

The power to detect this smaller increase is only about 0.442. More of the alternative distribution now lies inside the nonrejection region. **There is no single power for a test without specifying an effect.**

6.1 Compare designs using a power curve

Each point below answers: if this value on the horizontal axis were the true mean, how often would the test reject? Each sample size uses its own cutoff to keep $\alpha = 0.05$.



At $\mu = 100$, the rejection probability is α, because that point is the null. For means above 100, the curve gives power to detect a genuine increase.

Change, other inputs fixed	**Consequence for a true increase**
Increase n	Smaller standard error; lower β.
Increase $\mu_1 - \mu_0$	Effect is easier to distinguish; lower β.
Reduce σ	Less sampling variability; lower β.
Increase α	Easier to reject; lower β, but more false alarms.

Increasing sample size can improve power without increasing the Type I error probability. Changing α trades one kind of error against the other. We should not change a prespecified α merely because the observed p-value was disappointing.

7 Two sided tests have two rejection tails

Return to the carpenter example in Chapter 8. Test $H_0 : \mu = 19$ against $H_a : \mu \neq 19$, with $n = 100$, observed $\bar{x} = 20$, and $s = 6$. The slides use a large-sample normal approximation, with estimated standard error 0.6.

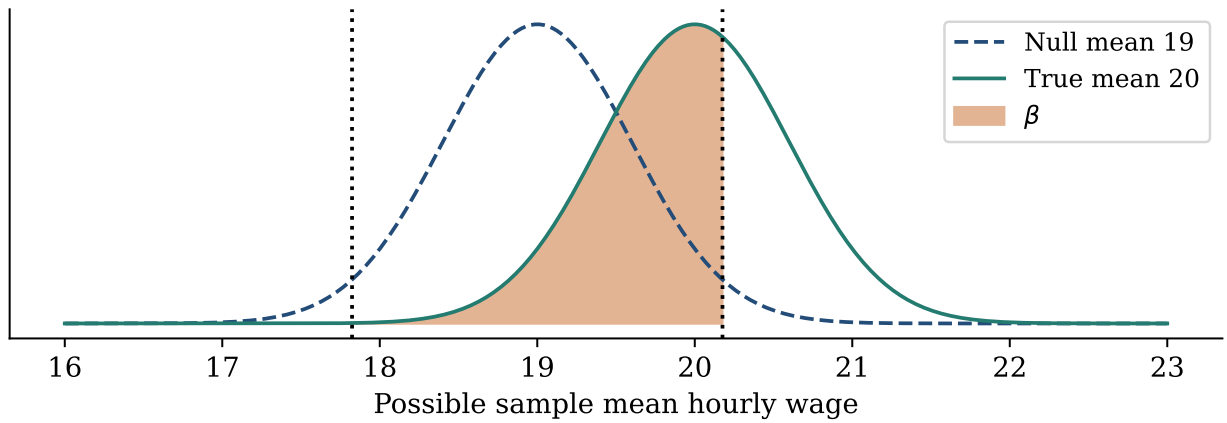
At $\alpha = 0.05$, the two critical values are $\pm z_{\alpha/2} = \pm 1.96$. We fail to reject when

$$-1.96 \leq \frac{\overline{X} - 19}{0.6} \leq 1.96 \quad \Longleftrightarrow \quad 17.824 \leq \overline{X} \leq 20.176.$$

For an approximate planning calculation, use 6 as a working population SD. If the true mean were 20, center the sampling distribution at 20 and compute the probability of falling *between both boundaries*:

$$\begin{aligned} \beta(20, 0.05) &\approx \mathbb{P}(17.824 \leq \overline{X} \leq 20.176 \mid \mu = 20) \\ &\approx \Phi\left(\frac{20.176 - 20}{0.6}\right) - \Phi\left(\frac{17.824 - 20}{0.6}\right) \\ &= \Phi(0.2933) - \Phi(-3.6267) \approx 0.6152. \end{aligned}$$

Thus power is approximately 0.3848. This is a **plug-in normal approximation**, not exact power for an unknown-σ test.



The orange area is β under the alternative. Power is the sum of the areas under that *same curve* beyond the two cutoffs. The dashed null curve is shown only for comparison.

For the observed data, $z_{\text{obs}} = 1.6667$ and the two-sided p-value is approximately 0.096 (0.095 with the slides' rounded $z = 1.67$). Failing to reject is therefore compatible with a design that frequently misses a true difference of \$1 per hour.

8 The same ideas apply to Chapter 9 tests

For a paired test, form the differences first. With n independent normal differences and unknown variance,

$$T = \frac{\overline{D} - \mu_{D,0}}{S_D/\sqrt{n}}, \qquad \text{df} = n - 1.$$

Under H_0, this statistic has a central t distribution. The course's critical-value notation is

$$\mathbb{P}(T_{\text{df}} > t_{\text{df},\alpha}) = \alpha.$$

For a two-sided test use $t_{\text{df},\alpha/2}$. For an independent pooled test with normal groups and a common unknown variance, $\text{df} = n_1 + n_2 - 2$.

The meanings of Type I error, Type II error, and power do not change. However, when the effect is nonzero, the relevant test statistic has a *noncentral* t distribution. Exact t-test power requires that alternative distribution, not just the usual central t critical-value table.

A general recipe for beta

1. Fix the hypotheses, significance level, and decision rule.
2. Specify an alternative parameter value and the other model inputs.
3. Find the sampling distribution under that alternative.
4. Calculate the probability of landing in the *nonrejection* region.
5. Subtract from one to obtain power for that alternative.

8.1 Why the normal example is useful

The biomarker example assumes normal observations and *known* σ. Its normal calculations are exact apart from rounding, even though $n = 25$. Small n alone does not require a t statistic; estimating an unknown population SD is the reason t enters the normal-population setting.

For nonnormal data, using a normal approximation requires suitable sampling conditions. A fixed rule such as "$n \geq 30$" is not a universal guarantee, particularly with heavy tails or strong dependence.

What a nonsignificant result does and does not say

A nonsignificant result is not evidence that the probability of an effect is small. Interpret it alongside the uncertainty interval and the effects the design could reasonably detect. Substituting the observed effect into a power calculation does not create new evidence beyond the observed test result.

9 Practice and interpretation

Use the biomarker setup unless a question states otherwise:

$$H_0 : \mu = 100, \quad H_a : \mu > 100, \quad \sigma = 10, \quad n = 25, \quad \alpha = 0.05.$$

The decision rule rejects when $\overline{X} > 103.29$. Useful values are $\Phi(-0.855) \approx 0.196$, $\Phi(0.145) \approx 0.558$, $\Phi(-3.355) \approx 0.000397$, and $\Phi(2) \approx 0.9772$.

1 Read the vertical bar

For each probability, state the assumed truth, the distribution to use, and the event whose probability is being calculated:

(a) $\mathbb{P}(\overline{X} > 103.29 \mid \mu = 100)$.

(b) $\mathbb{P}(\overline{X} \leq 103.29 \mid \mu = 105)$.

(c) $\mathbb{P}(\overline{X} \geq 104 \mid \mu = 100)$.

Which is α, which is β, and which is a p-value for an observed mean of 104?

2 Describe an error in context

Suppose the true mean is 103 but your sample does not cause rejection. Name the error. Does the experimenter generally know that this error occurred? Explain.

3 Calculate beta for a smaller effect

Assume $\mu = 103$. Write β as a conditional probability, standardize it using the correct center, and calculate power. Compare with the result at $\mu = 105$.

4 Decide from an observed result

You observe $\bar{x} = 104$. Calculate the p-value and state the decision. A classmate says: "The p-value is 0.0228, so the power is 0.9772." Explain why this is incorrect.

5 Change the sample size

Increase n to 100 while keeping $\alpha = 0.05$ and $\sigma = 10$.

(a) Find the new standard error and rejection cutoff.

(b) Calculate β and power when the true mean is 105.

(c) Did α change? Explain why the cutoff needed to change.

6 Draw a two sided test

Suppose a test fails to reject when $L \leq \overline{X} \leq U$. Sketch the null and a specified alternative distribution. Shade α, β, and power on the correct curves. Write the formula for β when $\overline{X}$ is normal with mean μ_1 and SD $\sigma/\sqrt{n}$.

10 Worked answers for self study

1(a). The condition is $\mu = 100$, so use the null sampling distribution with mean 100 and SD 2. The event is rejection. Its probability is $\alpha = 0.05$.

1(b). The condition is $\mu = 105$, so use the alternative distribution with mean 105 and SD 2. The event is failure to reject. Its probability is $\beta(105, 0.05) \approx 0.196$.

1(c). The condition is again the null. The boundary 104 comes from the observed result, not the prechosen cutoff. This right-tail area is the p-value, approximately 0.0228.

2. This is a Type II error: a real increase was missed. In a real study, the true population mean is usually unknown. The experimenter sees the nonrejection decision, but generally cannot know whether it is an error.

3. Keeping the rule fixed and changing the assumed truth gives

$$\beta(103, 0.05) = \mathbb{P}(\overline{X} \leq 103.29 \mid \mu = 103) = \Phi(0.145) \approx 0.558.$$

Power is approximately 0.442, lower than the power of about 0.804 at $\mu = 105$. A smaller increase is harder to distinguish from the null.

4. The observed standardized value is $z = 2$, so

$$p = 1 - \Phi(2) \approx 0.0228 < 0.05.$$

Reject H_0. The proposed power calculation is wrong because p uses the null distribution and the observed boundary. Power uses a specified alternative and the fixed rejection rule. Without specifying the alternative, the power is not determined.

5. The new standard error is $10/\sqrt{100} = 1$. The new cutoff is $100 + 1.645(1) = 101.645$. At $\mu = 105$,

$$\beta = \Phi\left(\frac{101.645 - 105}{1}\right) = \Phi(-3.355) \approx 0.000397, \qquad 1 - \beta \approx 0.999603.$$

Alpha remains 0.05: the cutoff was recalibrated to preserve that null tail area. The much larger power is specific to the assumed effect and known-σ model.

6. Shade the two rejection tails under the null for α. Under the alternative, shade the region between L and U for β, and the two outer tails for power:

$$\beta(\mu_1, \alpha) = \Phi\left(\frac{U - \mu_1}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{L - \mu_1}{\sigma/\sqrt{n}}\right).$$

These notes follow the notation and examples in the Chapter 8 slides (especially pp. 31–33, 41–53, and 79–82) and the Chapter 9 slides. The biomarker example is an illustrative model, not an empirical study.