Example:

New influenza vaccines are developed each year. The vaccines stimulate the production of antibodies that help fight the flu. But before the vaccines are put into production and widely distributed, they are tested to determine their efficacy.

Suppose that the vaccine is given to a group of 50 willing and eager participants. Twenty four hours later we take a sample of their blood and send it to a lab to determine if their antibody levels have increased enough to be protective.

Our statistical model is: $X_1,\ldots,X_{50} \sim \text{Ber}(\theta)$

where θ is the probability of a protective immune response from the vaccine.

Now some people, say 10%, will have an immune response just from being involved in a study (maybe they are allergic to the experimenters?).

And the vaccine has to be mildly effective (i.e., it has to work at least 25% of the time) to be worth the cost and effort.

So an appropriate null hypothesis here is that the probability of immune response is at least 35%.

So our null hypothesis is: H_0: $\theta \leq 0.35$

Selecting a rejection region

Now that we have the statistical model and null hypotheses, I only need to determine my rejection region before I can go out and collect some data to test this hypothesis.

Let's suppose that I decide to reject the null hypothesis if my sample estimate of this proportion is greater than 45%.

Why didn't I pick 35%? Because the sample proportion has some variability associated with it and I'm trying to account for that.

But admittedly this choice is ad-hoc at best. Hypothesis testing was developed to make this choice more precise (and we'll see how shortly), but for now let's entertain this choice.

So, to summarize, we have:

$X_1, \ldots, X_{50} \sim \text{Ber}(\theta)$ and H_0: $\theta \le 0.35$

And we will reject H_0 if the sample proportion is greater than 45%.

Now this hypothesis test is just as valid as any other. But it may not be the best in terms of how often it gives us the correct answer.

How often do we make a mistake? Let's calculate it.

What is the probability of rejecting the null hypothesis when it is true?

$$
\begin{aligned}
P(\text{Reject H}_0 \mid \text{H}_0 \text{ True}) &= P\left(\hat{\theta} > 0.45 \mid \theta_0 = 0.35\right) \\
&= P\left(\frac{\sqrt{n}\left(\hat{\theta} - \theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}} > \frac{\sqrt{n}\left(0.45 - \theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}} \mid \theta_0 = 0.35 \right) \\
&= P\left(Z > \frac{\sqrt{50}\left(0.45 - 0.35\right)}{\sqrt{0.35\left(1-0.35\right)}} \mid \theta_0 = 0.35 \right) \\
&= P\left(Z > 1.48 \mid \theta_0 = 0.35\right) \\
&= 0.069
\end{aligned}
$$

So when the null hypothesis is true, this test is pretty reliable. In that case, it only does the wrong thing (reject the null) 6.9% of the time. (This is called the Type I error.)

Now consider what happens when null hypothesis is false. In this case, we *want* to reject the null hypotheses. So let's calculate how often that happens.

But before we do, we have to be more precise about what it means for the null hypothesis to be false. That is, H_A: $\theta > 0.35$ is to vague to help me when I have to standardize; so I need a single number. Typically we pick a simple alternative, so let's say $\theta = 0.55$. Then,

$$
\begin{aligned}
P(\text{Reject } H_0 \mid H_0 \text{ false}) &= P\left(\hat{\theta} > 0.45 \mid \theta_A = 0.55\right) \\
&= P\left(\frac{\sqrt{n}\left(\hat{\theta}-\theta_A\right)}{\sqrt{\theta_A\left(1-\theta_A\right)}} > \frac{\sqrt{n}\left(0.45-\theta_A\right)}{\sqrt{\theta_A\left(1-\theta_A\right)}} \mid \theta_A = 0.55\right) \\
&= P\left(Z > \frac{\sqrt{50}\left(0.45-0.55\right)}{\sqrt{0.55\left(1-0.55\right)}} \mid \theta_A = 0.55\right) \\
&= P\left(Z > -1.42 \mid \theta_A = 0.55\right) \\
&= 0.92
\end{aligned}
$$

So when the true proportion is 55%, this test is very reliable. In that case, it correctly rejects the null hypothesis 92% of the time and only does the wrong thing (fails to reject the null) 8% of the time.

So we say this test has power of 92% to detect that the true portion is 55%. (And the Type II error of this test is therefore 8%.)

Notice that if I change the alternative, my power will also change, so the power is tied to a specific alternative hypothesis.

For example, the power to detect the alternative of 45% is

$$
\begin{aligned}
P(\text{Reject H}_0 \mid \text{H}_0 \text{ false}) &= P\left(\hat{\theta} > 0.45 \mid \theta_A = 0.45\right) \\
&= P\left(\frac{\sqrt{n}\left(\hat{\theta} - \theta_A\right)}{\sqrt{\theta_A\left(1-\theta_A\right)}} > \frac{\sqrt{n}\left(0.45 - \theta_A\right)}{\sqrt{\theta_A\left(1-\theta_A\right)}} \mid \theta_A = 0.45 \right) \\
&= P\left(Z > \frac{\sqrt{50}\left(0.45 - 0.45\right)}{\sqrt{0.45\left(1-0.45\right)}} \mid \theta_A = 0.45 \right) \\
&= P\left(Z > 0 \mid \theta_A = 0.45\right) \\
&= 0.5
\end{aligned}
$$

And this should make intuitive sense!

Today's version of "hypothesis testing," is only a slightly more complicated version of what we just did. The twist is that we first specify the type I error (because we want to control it) and then we work backwards to find the rejection region.

For example, suppose we want to choose the rejection point such that we only make errors under the null 5% of the time (i.e., we want to fix our type I error to be 5%). To do this we calculate:

$$\begin{aligned} 0.05 = P(\text{Reject H}_0 \mid \text{H}_0 \text{ True}) &= P\left(\hat{\theta} > ? \mid \theta_0 = 0.35\right) \\ &= P\left(Z > \frac{\sqrt{50}\left(? - 0.35\right)}{\sqrt{0.35\left(1-0.35\right)}} \mid \theta_0 = 0.35 \right) \\ &= P\left(Z > 1.645 \mid \theta_0 = 0.35\right) = 0.05 \end{aligned}$$

And we have to find the ?, so

$$\frac{\sqrt{50}\left(? - 0.35\right)}{\sqrt{0.35\left(1-0.35\right)}} = 1.645$$

$$\begin{aligned} ? &= 0.35 + 1.645\frac{\sqrt{0.35\left(1-0.35\right)}}{\sqrt{50}} \\ &= 0.46 \end{aligned}$$

So our rejection point is now 46%. By rejecting whenever the sample proportion is greater than 46% we can be sure that we won't make a Type I error more than 5% of the time.

And the power for this test (assuming that θ = 0.55):

$$
\begin{aligned}
P(\text{Reject H}_0 \mid \text{H}_0 \text{ false}) &= P\left(\hat{\theta} > 0.46 \mid \theta_A = 0.55\right) \\
&= P\left(\frac{\sqrt{n}\left(\hat{\theta} - \theta_A\right)}{\sqrt{\theta_A\left(1 - \theta_A\right)}} > \frac{\sqrt{n}\left(0.46 - \theta_A\right)}{\sqrt{\theta_A\left(1 - \theta_A\right)}} \mid \theta_A = 0.55 \right) \\
&= P\left(Z > \frac{\sqrt{50}\left(0.46 - 055\right)}{\sqrt{0.55\left(1 - 0.55\right)}} \mid \theta_A = 0.55 \right) \\
&= P\left(Z > -1.28 \mid \theta_A = 0.55\right) \\
&= 0.90
\end{aligned}
$$

Actually, to start this calculation, what we did was this:

$$
\begin{aligned}
P(\text{Reject H}_0 \mid \text{H}_0 \text{ false}) &= P\left(Z > 1.645 \mid \theta_A = 0.55\right) \\
&= P\left(\frac{\sqrt{n}\left(\hat{\theta} - \theta_0\right)}{\sqrt{\theta_0\left(1 - \theta_0\right)}} > 1.645 \mid \theta_A = 0.55 \right) \\
&= P\left(\hat{\theta} > \theta_0 + 1.645 \frac{\sqrt{\theta_0\left(1 - \theta_0\right)}}{\sqrt{n}} \mid \theta_A = 0.55 \right) \\
&= P\left(\hat{\theta} > 0.35 + 1.645 \frac{\sqrt{0.35\left(1 - 0.35\right)}}{\sqrt{50}} \mid \theta_A = 0.55 \right) \\
&= P\left(\hat{\theta} > 0.46 \mid \theta_A = 0.55\right)
\end{aligned}
$$

Notice here that we can switch back and forth between the "data" scale and the "standardized" scale. Specifying the tests on either scale is fine.

Rejecting when the same proportion is larger than some cutoff or rejecting when the test statistic is larger than the Z-value is the exact same thing!

That is,

$$\hat{\theta} > 0.46 \quad \text{is equivalent to} \quad \frac{\sqrt{50}\left(\hat{\theta} - 0.35\right)}{\sqrt{0.35(1-0.35)}} > 1.645$$

The first is on the scale of the data and the second is on the standardized scale, but they are mathematically equivalent.

The point is that the second expression is more generalizable and easier to deal with in statistical terms. So we use this more often.

To emphasize the point, suppose that when we collected our data we observed that the sample proportion was 50% (25 with the antibodies).

Since $\hat{\theta} = 0.50$ is greater than 46%, we would reject the null hypothesis at the 5% level.

The p-value is

$$\begin{aligned}
pvalue &= P\left(\hat{\theta} > 0.50 \mid \theta_0 = 0.35\right) \\
&= P\left(\frac{\sqrt{n}\left(\hat{\theta} - \theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}} > \frac{\sqrt{n}\left(0.50 - \theta_0\right)}{\sqrt{\theta_0\left(1-\theta_0\right)}} \mid \theta_0 = 0.35\right) \\
&= P\left(Z > \frac{\sqrt{50}\left(0.50 - 0.35\right)}{\sqrt{0.35\left(1-0.35\right)}} \mid \theta_0 = 0.35\right) \\
&= P\left(Z > 2.22 \mid \theta_0 = 0.35\right) \\
&= 0.013
\end{aligned}$$

Notice that (1) $pvalue = P\left(\hat{\theta} > 0.50 \mid \theta_0 = 0.35\right) = 0.013$

or (2) $pvalue = P\left(Z > 2.22 \mid \theta_0 = 0.35\right) = 0.013$ are equivalent.

Most computer packages will refer to (2), but (1) is much easier to think and talk about because it is on the scale of the data.