Building Statistical Tests

Jean-Baka Domelevo Entfellner

November 2013

Contents

1 Introduction

In many occasions you would like to make a statement concerning the underlying distribution of some random variable, knowing only a sample of values taken from that distribution. Such statements can have for instance one of the following forms:

- cancer patients treated with drug α live longer lives than if untreated;
- beads from one bag (containing tens of thousands) are bigger than their counterparts from another bag;
- temperature values across the world are normally distributed around $\mu = 20°\mathrm{C}$.

Statistical hypothesis testing is that process consisting in trying to answer questions about a probability distribution from a sample which data points have supposedly been generated by the above mentioned distribution. Statistical hypothesis testing is that process consisting in trying to answer questions about which probability distribution you suppose generated the data points you are interested in. To answer these questions, you only know one (or several) sample(s) of that random variable you measured. This entails four important remarks applying to all cases of statistical hypothesis testing:

1. you first have to state assumptions concerning the underlying probability distribution, which is unknown;

2. the data points within each of your samples should be i.i.d. (independent and identically distributed), otherwise the following will not be achievable;

3. you will derive a *statistic* (i.e. a quantity calculated from your sample), of which you know the distribution as soon as point (1) is correct;

4. finally, the answer to your test will necessarily be *stochastic*: no statistical test will lead to any form of certainty. To conclude your test, you will only be able to formulate things as: « After taking this statistical test, I am 99% sure that ... »

2 Describing the step-by-step framework for a test

Building a statistical test always has to be done according to a precise step-by-step framework that we are going to explain here. To illustrate the theory, we are going to use throughout a toy example. Let us suppose we are given daily measurements over one year of the temperature in Port Louis, Mauritius. All 365 measurements have been taken at noon from the same meteorological station. From these measurements, we want to answer the one question: is the average temperature in Port Louis at noon 25°C?

The steps to design the test are the following:

1. Acquire data.

2. State the hypotheses you want to test: what is the null hypothesis H_0, what is the alternate hypothesis H_a? Is it then a one-tailed or a two-tailed test?

3. Understand the two different types of errors involved (Type I and Type II errors). Define the rate of Type I error (α) you want to work with.

4. What is the test statistic t you are going to use ? Compute its realized value from your data points.

5. Find out what the probability distribution of t under H_0 is.

6. When possible, make explicit what is the probability distribution of t under H_a (necessary to compute the power of the test, but not compulsory for basic testing).

7. From the probability distribution $\Pr(t|H_0)$, work out the threshold value(s) t_{thresh}. Clearly mark the acception and rejection regions for H_0.

8. Comparing the calculated value of t and the threshold value(s), decide whether you reject or fail to reject H_0.

9. Optionally, compute the so-called *p-value*: the likelihood for t to get more extreme a value under H_0.

In the next subsections we are going to discuss all these points in detail, before we move on to several typical statistical tests.

2.1 Acquiring data

As we said earlier, it is important that the data points you work with be independent and identically distributed: in order to be able to draw accurate conclusions concerning the probability distribution (in a generative context, we rather say *stochastic process*) having generated your data points, you want these points to be:

1. all generated by that same stochastic process (*identically distributed*)
2. *independent* from each other: whatever the arbitrary ordering of your data points, the knowledge of the $n-1$ first points in your sample *does not* help you predict the nth point. In other terms, even after seeing as many of the data points drawn from the sample, one's prediction of the following data point is not further informed than from the mere knowledge of the generative stochastic process. If we denote $\mathcal{P}$ the generative stochastic process and $\{x_i\}_{i\in\{1,n\}}$ the data points, it means that for any ordering $\phi : [\![1,n]\!] \to [\![1,n]\!]$ we have $\Pr(x_{\phi(n)}|\mathcal{P}, x_{\phi(1)}, \ldots, x_{\phi(n-1)}) = \Pr(x_{\phi(n)}|\mathcal{P})$.

This double point shall not be overlooked, as a bias instrument or measuring protocol may well yield data points that are not only generated by the stochastic process but also influenced by the measurement process. For instance, when repeated measurements of weights are performed with a spring-loaded device, this device itself could be gradually damaged by the heavy weights and its spring would end up being overstretched. This way, the subsequent measures would be gradually skewed and overestimated: the resulting measured data points would not be i.i.d.

Here, let us name our 365 data points (temperatures collected throughout the year) x_1 to x_{365}. The whole sample is called $\mathbf{x}$.

Application on our toy example

We generate as some fake (but possible) data a random sample from a distribution centered on 26°C with a standard deviation of 3°C (remember the dimensionality of the standard deviation is that of the data itself). Just for our eyes (and to showcase some of the possibilities offered by R), we round the values to two decimal places and sort the resulting vector of temperatures in ascending order:

```
> x = sort(round(rnorm(n=365, mean=26, sd=3), digits=2))
> n = length(x)  # n = 365
```

We now forget how this vector has been built, but simply assume this is the data we have to deal with.

2.2 Stating hypotheses

Every statistical test is about testing one hypothesis, usually against another. First we have to state clearly an hypothesis that we call the *null* hypothesis and notate H_0. This so-called "null hypothesis" has to obey certain rules we describe below:

- H_0 has to be neatly utterable. No fuzzy statement here. "Apples are rather scarce on marketplaces this year" *cannot be* an H_0 hypothesis.

- H_0 has to be in the form of an *affirmative* statement. "The average temperature is not 25°C" *cannot be* an H_0 hypothesis.
- H_0 cannot be a *composite* hypothesis, it has to be a *simple* one. By this we mean that the distribution of the test statistic shall be fully specified by H_0. For instance, "the average temperature at noon in Port Louis is *above* 25°C" cannot be an H_0 hypothesis because it does not provide us with a unique distribution for the test statistic (being the mean measurement) under this hypothesis. This constraint says that the knowledge of H_0 should be enough to know with accuracy the probability distribution of the test statistic *given* H_0.
- if there is a "logical assumption" or an *a priori* belief in the matter, H_0 should be this *a priori* belief. For instance, "there are equal probabilities to toss a coin to any one of its two faces" should be the null hypotheses in simple cases related to tossing a coin. In a study concerned with birth ratios, "the number of baby boys born in one year is the same than that of baby girls" should be the null hypothesis, at least in most cases.

The **alternative hypothesis** is usually notated H_a or H_1. Here we will stick to the first notation. As suggested by its name, this hypothesis represents an alternative to H_0. It is designed so that *if* H_0 *does not hold, then* H_a *should*. In some cases, H_a is exactly the logical negation of H_0. In some other cases, H_a is not as broad as $\not(H_0)$, but the situations where neither H_0 nor H_a holds are disregarded, because they are obviously nonsensical or simply not interesting for the study.

Application on our toy example

In our example of the temperatures in Port Louis, H_0 will be rendered by the sentence "the average temperature at noon in Port Louis is 25°C". This is not necessarily an *a priori* belief, but because we want to test whether the average temperature is 25°C or not, this is the only suitable statement for an H_0 hypothesis.

In our toy test about the average temperature, H_a will be expressed by the sentence "the average temperature at noon in Port Louis is *not* equal to 25°C". If we write down this average temperature with μ_T, then it means that $H_a \equiv (\mu_T < 25 \text{ or } \mu_T > 25)$. In such situations we say that our test is *two-tailed*, because its rejection region (also known as "critical region") is composed of two separate parts at the two tails of the probability distribution for the test statistic under H_0. A preliminary sketch of such a two-tailed test is given in figure 1b. The one-tailed test with $H_a \equiv (\mu_T < 25)$ would be like in figure 1a. You will understand these two figures more clearly in a few moments, as we will keep on referring to them.

2.3 Type I and Type II errors: setting α

Whatever our statistical hypotheses and tests, we have to admit that there is a certain *reality* or a one and only *truth*. We don't know which is the truth, but we suppose it is either H_0 or H_a. The process of performing a statistical test will lead to a form of decision: we are finally going to reject or fail to reject (i.e. uphold in the absence of further contradicting evidence) our null hypothesis H_0. Doing so we could be wrong, in two different ways:

- the outcome of our test could be to reject H_0 while in reality it holds. We would then fail with a **Type I error**;

truth	test outcome	
	decide H_0 (acceptance)	decide H_a (rejection)
H_0	rightful acceptance	**Type I error**
H_a	**Type II error**	rightful rejection

Table 1: Different types of errors in a hypothesis test

- by our test we could decide symmetrically that H_0 is not to be rejected while in truth the alternative hypothesis H_a holds (and so H_0 should indeed be rejected). Here we are also wrong and commit a **Type II error**.

We sum this up in the small table 1.

There is no way we can avoid being possibly wrong in the conclusion of our test (otherwise we would be able to know the truth with certainty, which is not). Before starting a test we can decide, though, with what chance of committing a Type I error we want to work. The important point to consider is that we cannot control the risk level *both* for Type I *and* for Type II errors: if we decrease the chance of committing a Type I error, we will inevitably increase the risk of a Type II error, and inversely. One can consider for example the extreme situation of a test that would always decide H_0, whatever the data. With such a test it would be impossible to commit a Type I error, so the Type I error rate would simply be nil. The Type II error rate, though, would probably be huge: in many cases, we would decide to trust H_0 while it would be wrong. Similarly, a test deciding always in favour of H_a whatever the data would have a Type II error rate equal to 0, but possibly a very high Type I error rate.

The Type I error rate that we decide to work with is usually called α. The corresponding Type II error rate, which can be calculated only for simple (i.e. not composite) alternative hypotheses H_a, is called β.

Typical values for α are 0.01 (i.e. 1%) or 0.05 (which is 5%). But the designer of the test is really the one who has to chose. In some situations, making a Type I error can have severe consequences (for instance leading to the death of a patient by giving him/her inappropriate treatement). In such a case the designer of the test would like to have a very small α. On the contrary, if Type II errors are really to be avoided while Type I errors are less detrimental, we will set a somewhat larger value for α, in an attempt to get as low a β value as possible.

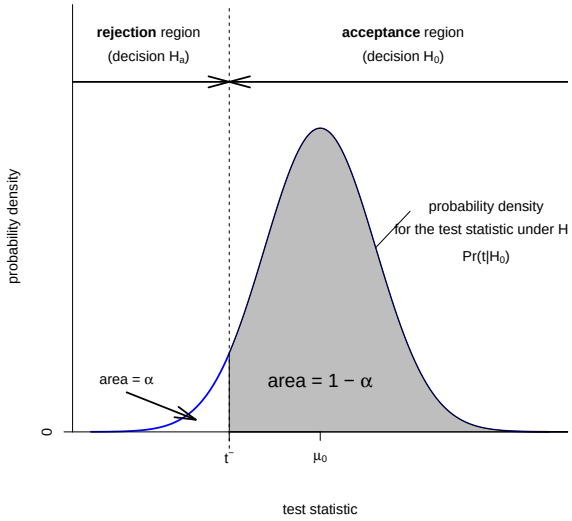
Application on our toy example

Here we do not have very severe worries about the results yielded by a wrong test outcome, either in one way (being too conservative on H_0) or on the other (being too prone to reject it), so we choose a rather standard value $\alpha = 5\%$.

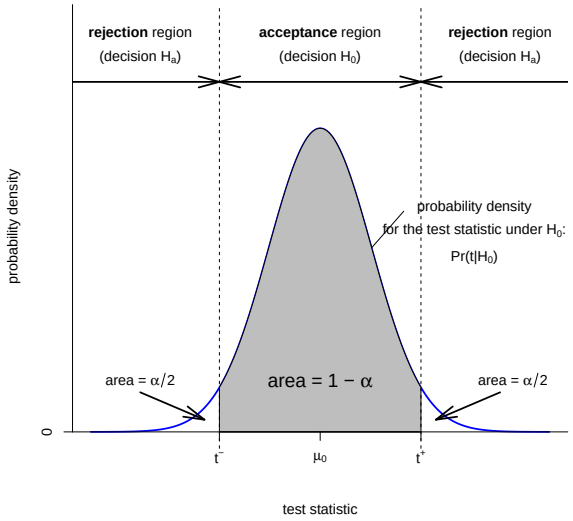
```
> alpha = 0.05
```

2.4 Defining a test statistic

Key to the test is the choice of a *test statistic*. This is a random variable t expressed as a function of the data sample. We choose this function so that the following two conditions are fulfilled:



(a) A one-tailed test with Type I error rate α. The rejection region is in one piece. On display here is an alternative hypothesis H_a of the form « $\mu < \mu_0$ ».



(b) The corresponding two-tailed test, with same Type I error rate α. The rejection region is in two pieces of equal area under the curve. This situation corresponds to a test where the alternative hypothesis H_a is of the form « $\mu \neq \mu_0$ ».

Figure 1: Sketches for one-tailed (left) and two-tailed (right) hypothesis tests. The curve displayed is the probability density function for the test statistic t when H_0 holds. Besides, the details we write in the subcaptions assume that t is an estimator for the unknown parameter μ, and that H_0 is the hypothesis « $\mu = \mu_0$ ».

1. we know the distribution $\Pr(t|H_0)$, possibly making some additional acceptable assumptions;
2. t is an estimator of a unknown parameter or quantity that is meaningful with regard to the question we want to answer through our testing procedure.

Point (2) can be illustrated with some examples:

- if our test is concerned with the true mean of the distribution having generated our data points, a sound test statistic will be the mean of our sample (t-tests, linked to the Student probability distribution);
- in the case our test is about determining whether two distributions have equal variances, the test statistic may well be the ratio of the two samples (homeoscedasticity test, linked to the Fisher probability distribution);
- if we want to see whether two groups are significantly different in terms of magnitude, without assuming normality, we could use as a test statistic some quantity involving the mean ranks of the groups (Kruskal-Wallis test) or the sum of the signed ranks between two paired samples (Wilcoxon signed-rank test), etc.

Application on our toy example

In our toy example we are concerned with the average temperature in Port Louis. We know from the theory of estimation that the **sample mean** $m_{\mathbf{x}}$ is the best estimator of that unknow average temperature μ_T. Logically enough, our test will be built around this quantity $m_{\mathbf{x}}$. It is not exactly our test statistic, though, because we know that without the knowledge of the true variance σ_T^2, we do not know the probability distribution followed by $m_{\mathbf{x}}$ under the null hypothesis H_0. And to proceed with our test, it is essential that we choose a test statistic enabling us with the expression thereof under the null hypothesis.

We know from the introductory course on probability theory that, when the n datapoints are independently generated by a unique normal distribution, $T(\mathbf{x}) = \frac{\mu_T - m_{\mathbf{x}}}{s_{\mathbf{x}}/\sqrt{n}}$ is distributed according to a Student probability distribution $t_{(n-1)}$. We notate this random variable $T(\mathbf{x})$ to stress the fact that it depends on the sample $\mathbf{x}$: $T(\mathbf{x})$ is that *sample statistic* we are going to use as our *test statistic*.

We can use the R functions `mean`, `var` and `sd` respectively for the sample mean, the sample variance and the sample standard deviation. We are going to demonstrate that R computes correctly these values by doing it "manually" also:

```
> mymean = sum(x) / n
> mymean - mean(x)
[1] 0
> myvar = sum((x - mymean)^2) / (n-1)
> myvar - var(x)
[1] 0
> mysd = sqrt(myvar)
> mysd - sd(x)
[1] 0
```

Now we have almost everything we want to proceed with our test (we know $m_{\mathbf{x}}$ and $s_{\mathbf{x}}$), but notice that we do not know the value of our test statistic $T(\mathbf{x})$ unless we specify a value for μ_T.

2.5 Probability distribution of the test statistic under H_0

The next step in our test procedure is to determine the distribution of our test statistic when H_0 holds. Here in our example we have:

$$H_0 \quad \equiv \quad \mu_T = \mu_0 = 25 \tag{1}$$

Under such a null hypothesis H_0, $T(\mathbf{x})$ is fully characterised: we know $\mu_T = \mu_0$, and from the sample we also know n (the sample size), $m_{\mathbf{x}}$ (the sample mean) and $s_{\mathbf{x}}$ (the sample variance, calculated with denominator $(n-1)$). We know $T(\mathbf{x})$ is distributed according to a $t_{(n-1)}$ probability distribution.

Application on our toy example

Under H_0 we can determine the value of our test statistic $T(\mathbf{x})$:

```
> test_statistic = (25 - mymean)/sqrt(myvar/n)
```

2.6 Threshold value(s) and the acceptance/rejection regions for H_0

As soon as we have made a decision on the Type I error rate α, we can draw the acceptance and rejection regions: the outcome of the test fully relies on the shape of the probability distribution of the test statistic when H_0 holds. The acceptance region always includes the majority of the weigth (i.e. area under curve, a.k.a AUC) of the probability density for $\Pr(t|H_0)$. The threshold value(s) is (are) determined in order to cut off an AUC equal to α: that AUC falls into the rejection domain.

Application on our toy example

In our toy test, H_0 is "the average temperature at noon in Port Louis is 25°C" while H_a is the logical negation of H_0, i.e. the sentence "the average temperature at noon in Port Louis is *not* equal to 25°C. We thus have a **two-tailed test**, with the rejection zone made of two separate intervals. As $\alpha = 0.05$, we want to have each of these two areas equal to 0.025 (symmetric setting). The two thresholds are given by the two corresponding quantiles of the Student probability distribution with $(n-1) = 364$ degrees of freedom:

```
> thresh1 = qt(alpha/2, df=n-1)
> thresh1
[1] -1.966503
> thresh2 = qt(1-alpha/2, df=n-1)
> thresh2
[1] 1.966503
```

The acceptance region corresponds to test statistics falling within the interval $[-1.967, 1.967]$. The rejection regions are the two intervals $]-\infty, -1.967]$ and $[1.967, \infty[$.

2.7 Test outcome

If the test statistic as calculated from the sample lies within the acceptance region, the outcome of the test will be to accept (or fail to reject) H_0. Otherwise (if the value of the test statistic is further from the peak than the threshold value), the test *rejects* H_0 and accepts H_a. A corresponding sketch is drawn for example on figure 1a, depicting the case of a single-tailed test where the test statistic is normally distributed.

Application on our toy example

With my data (remember that our data vector `x` has been randomly generated, so you can get a different value) I get the following:

```
> test_statistic = (25 - mymean)/sqrt(myvar/n)
> test_statistic
[1] -14.70914
```

The test statistic falls into the rejection region, so the outcome of the test here is to **reject** the null hypothesis H_0: with a Type I error rate set at $\alpha = 5\%$, we are confident that the average temperature at noon in Mauritius is *not* equal to 25°C. This statement is about the true mean of the underlying distribution, that we assume is normal.

truth	test outcome	
	decide H_0 (acceptance)	decide H_a (rejection)
H_0	$1-\alpha$	α (Type I err. rate)
H_a	β (Type II err. rate)	$1-\beta$ (power)

Table 2: Probabilities Pr(outcome|truth) in a hypothesis test

2.8 Probability distribution of the test statistic under H_a and power of the test

The **power** of a test is the probability, when one performs this test, to reject H_0 when indeed it is false. More precisely, it represents the ability of the test to detect that H_0 does not hold when H_a is true in reality. The power of the test is written $\pi(H_0, H_a) = 1-\beta$, where β is the Type II error rate, i.e. the probability that the test outcome be to accept H_0 when in fact H_a is true. Let us write these probabilities down into the 2x2 matrix in table 2.

If we know $\Pr(t|H_a)$, we can plot its density on the graph where we already have the density of $\Pr(t|H_0)$. See figure 2 for an example.

Application on our toy example

As we said earlier, we cannot compute the power of the test unless we give the alternative hypothesis a simple form, leading to a one-tailed test and enabling to calculate $\Pr(t|H_a)$. Let us say we want to test $H_0: \mu_T = \mu_0 = 25$ versus $H_a: \mu_T = \mu_a = 27$. Notice that this modifies the acceptance and rejection regions, and that we now have only one threshold value `t0`. On the left of that threshold we reject H_0, and on the right we accept H_0. This is consistent with the way we wrote our test statistic $T(\mathbf{x})$ under H_0, its numerator being $(\mu_0 - m_{\mathbf{x}})$: the bigger the observed temperature mean $m_{\mathbf{x}}$, the more negative $T(\mathbf{x})$ gets, and the more likely H_a becomes. The threshold value is still calculated only from H_0, but now with the rejection region all in one piece (AUC α on the lower tail of the density function):

```
> t0 = qt(alpha, df=n-1)
> t0
[1] -1.649051
```

From figure 2, you can see that the power of the test, $(1-\beta)$, is the AUC for the density of $\Pr(T(\mathbf{x})|H_a)$ to the left of our new threshold value `t0`. But a slight problem arises here: we should have for both curves to plot (the one corresponding to the density under H_0 and the other corresponding to the density under H_a the very same test statistic. But in our test statistic $T(\mathbf{x})$, so far we had to use the value of the mean under H_0, which is μ_0: $T(\mathbf{x}) = \frac{\mu_0 - m_{\mathbf{x}}}{s_{\mathbf{x}}/\sqrt{n}}$. This makes it necessary to be careful when we derive the value of the power. Let us go back to the definition of the power of the test: this is Pr(reject $H_0|H_a$). Here, because we know that

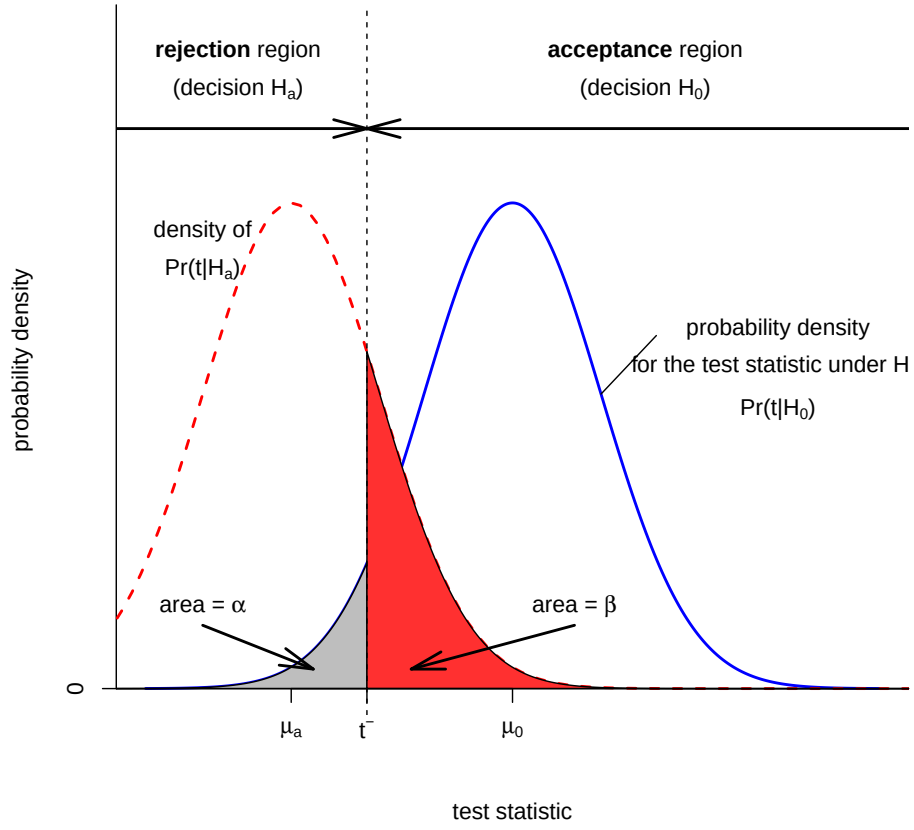


Figure 2: A one-tailed test with Type I error rate α and Type II error rate β. On display here is an alternative hypothesis H_a of the form « $\mu = \mu_a$ » with $\mu_a < \mu_0$.

the test outcome is to reject H_0 iff. $T(\mathbf{x}) < t_0$, we write:

$$\begin{aligned}
\pi(H_0, H_a) &= \Pr\left(\frac{\mu_0 - m_{\mathbf{x}}}{s/\sqrt{n}} < t_0 \;\middle|\; H_a\right) \\
&= \Pr\left(\frac{\mu_0 - m_{\mathbf{x}}}{s/\sqrt{n}} < t_0 \;\middle|\; \frac{\mu_a - m_{\mathbf{x}}}{s/\sqrt{n}} \sim t_{(n-1)}\right) \\
&= \Pr\left(\frac{\mu_0 - \mu_a}{s/\sqrt{n}} + \frac{\mu_a - m_{\mathbf{x}}}{s/\sqrt{n}} < t_0 \;\middle|\; \frac{\mu_a - m_{\mathbf{x}}}{s/\sqrt{n}} \sim t_{(n-1)}\right) \\
&= \Pr\left(\frac{\mu_a - m_{\mathbf{x}}}{s/\sqrt{n}} < t_0 + \frac{\mu_a - \mu_0}{s/\sqrt{n}} \;\middle|\; \frac{\mu_a - m_{\mathbf{x}}}{s/\sqrt{n}} \sim t_{(n-1)}\right) \\
&= \Pr\left(V < t_0 + \frac{\mu_a - \mu_0}{s/\sqrt{n}} \;\middle|\; V \sim t_{(n-1)}\right)
\end{aligned} \tag{2}$$

Calculating this probability with R makes use of a call to `pt`, the cumulative distribution function for the Student:

```
> new_threshold = t0 + (27 - 25) / sqrt(myvar/n)
> pt(new_threshold,df=n-1)
[1] 1
```

R approximates the power of this t-test as 1 (by definition, the power of a test can never be greater than 1). This means that having such a large number of data points ($n = 365$) is enough to detect a difference of 2 in the true mean, i.e. is enough to be sure to decide H_a whenever the data is indeed distributed normally around 27. A value μ_a closer to μ_0 and/or a smaller sample size would not give us such a good power.

2.9 Calculating that famous p-value

The p-value of the test, when H_0 is the conservative null hypothesis, says how extraordinary it would be to get such a remarkable value for your test statistic (i.e. a value departing so much from the null hypothesis) **if H_0 is in fact true**. That p-value is thus the weigh of the distribution tail beyond the measured value of the test statistic.

Application on our toy example

If we go back to the one-tailed test, the p-value is as follows:

```
> test_statistic
[1] -14.70914
> my_pval = pt(test_statistic, df=n-1)
> my_pval
[1] 4.571025e-39
```

This means that there would be virtually no chance to get as extreme a value for the test statistic if H_0 was indeed true: this is a measure of the high confidence we have in the outcome of the test (that was to reject H_0).

For the corresponding two-tailed test (with $H_a \equiv \mu_T \neq \mu_0$), we would get twice as large a p-value, because of the other tail that we would have to take into consideration:

```
> my_pval_2_tail = pt(test_statistic, df=n-1) + pt(-test_statistic, df=n-1, lower.tail=F)
> my_pval_2_tail
[1] 9.142049e-39
```

2.10 Using `t.test` and `power.t.test`

(see practical)