The student's t-distribution

So far we've been working with the normal distribution, which is that perfectly symmetrical bell-shaped distribution with a mean μ and a standard deviation σ.

The **standard normal distribution** (also called the z-distribution) is the normal distribution that specifically has mean $\mu = 0$ and standard deviation $\sigma = 1$, and we've been looking up z-values for the z-distribution in the z-table.

Of course, by the Empirical rule, about 68% of the area under the standard normal distribution is found between the z-scores $z = -1$ and $z = 1$, about 95% of the area is found between $z = -2$ and $z = 2$, and about 99.7% of the area is found between $z = -3$ and $z = 3$.

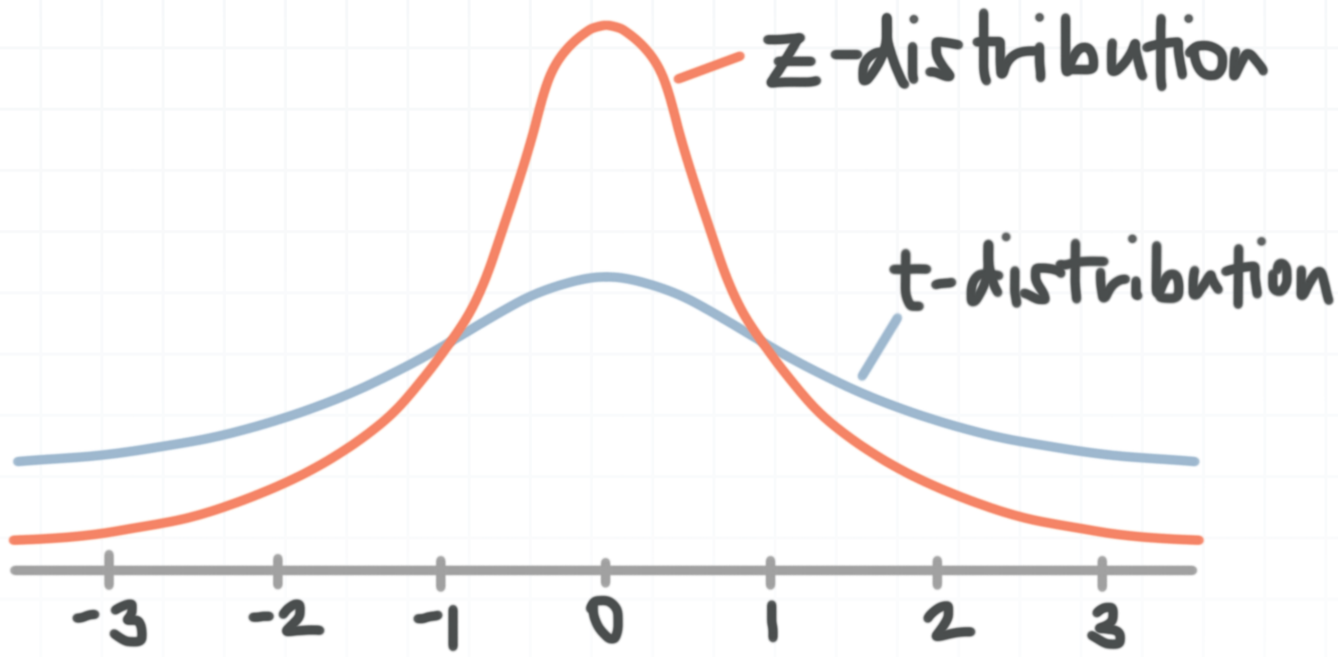
Now we want to turn our attention toward the t-distribution, and t-scores that we look up in the t-table.

The student's t-distribution

The **student's t-distribution** is similar to the standard normal distribution (z-distribution) in the sense that it's symmetrical, bell-shaped, and centered around the mean $\mu = 0$.

But the t-distribution is flatter and wider than the standard normal distribution, so more of the area under the t-distribution is pushed out toward the tails. That means that the standard deviation of the t

-distribution is larger than the standard deviation of the standard normal distribution.



Keep in mind that there isn't one generic t-distribution. The specific shape of a t-distribution, and therefore the t-scores we find in the t-table, will depend on the number of **degrees of freedom**, which is given by $n-1$, where n is our sample size.

So the t-distribution for a sample of size $n=10$ (with $n-1=10-1=9$ degrees of freedom) will look a little different than the t-distribution for a sample of size $n=20$ (with $n-1=20-1=19$ degrees of freedom).

In what way does the t-distribution for $n=10$ look different than the t-distribution for $n=20$. Well, regardless of the sample size (and therefore the number of degrees of freedom), the t-distribution is always normal. But the larger the sample size, the taller and narrower the t-distribution gets. The smaller the sample size, the shorter and wider the t-distribution gets.

In other words, as the sample size gets larger, the data becomes more tightly clustered around the mean, and the standard deviation gets smaller.

When the sample size reaches $n = 30$, the t-distribution gets just tall enough that its values very closely approximate the values of the z-distribution. So for sample sizes $n \geq 30$, the values from the t- and z-distributions will be almost identical. But for smaller samples $n < 30$, the t-distribution will do a better job estimating probability than the z-distribution.

Therefore, if we're using a sample size $n < 30$, we should calculate a t-score and look up that t-score in a t-table, instead of calculating a z-score and looking up that z-score in a z-table.

$n \geq 30$	Find a z-score, look it up in the z-table
$n < 30$	Find a t-score, look it up in the t-table

Because the z- and t-table return almost identical values when $n \geq 30$, such that either the z- or t-table could really be used for larger samples, and because we must use the t-table for smaller samples, some statisticians prefer to simplify things and always use the t-table, regardless of sample size. But many still choose to use a t-table for smaller samples and switch to the z-table for larger samples.

For our purposes, we'll follow this second approach, using a t-table for smaller samples and a z-table for larger samples.

There are also other conditions, besides sample size, that dictate when we should use the z-distribution vs. the t-distribution. For instance, if population standard deviation is unknown, we'll always use the t-table, regardless of sample size. For example, given a sample of size $n = 40$, if population standard deviation is unknown, we'd use a t-table.

Finding values in the t-table

Looking up t-values in a t-table is similar to looking up z-values in a z-table. To find values in the t-table, we need to know degrees of freedom ($n - 1$, when the sample size is n) and either the upper-tail probability or the confidence level.

We'll talk more later about upper-tail probability and confidence levels, but for now, we just want to understand that we can use either one (along with degrees of freedom) to locate values in the t-table.

In the t-table below, we see values for upper-tail probability along the top of the table, and confidence levels along the bottom of the table. So knowing either of these lets us locate the correct column of the table. And the degrees of freedom, which we see down the left side of the table, lets us locate the correct row.

	Upper-tail probability p									
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.765	0.987	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
	50%	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.8%**	**99.9%**
	Confidence level C									

For example, let's say we know that upper-tail probability is 0.01, and that our sample size is $n = 20$. Then we find $n - 1 = 20 - 1 = 19$ degrees of freedom on the left of the table, and upper-tail probability 0.01 along the top of the table. The intersection of those two values is the value we want to pull from the t-table.

	Upper-tail probability p									
df	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	**2.539**	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

Keep in mind that if we'd instead been given 19 degrees of freedom and a confidence level of 98%, we'd have found the exact same value from the t-table. That's because an upper-tail probability of 0.01 and a confidence level of 98% both put us in the same column of the t-table. The reason this is true is because "an upper-tail probability of 0.01" and "a confidence level of 98%" are just two different ways of saying exactly the same thing. Similarly,

- An upper-tail probability of 0.05 = a confidence level of 90%
- An upper-tail probability of 0.025 = a confidence level of 95%
- An upper-tail probability of 0.005 = a confidence level of 99%

Remember, the reason the full t-table above only extends to 30 degrees of freedom is because, when the sample size is $n \geq 30$, values in the z-table and t-table are approximately equivalent. Therefore, once we reach more than 30 degrees of freedom, using a z-table would give us a good enough approximation of the value we need.