Example: Joe and Mike have a large box with hundreds of thousands of tickets in it. Many years earlier, when they first gathered the tickets, they found that the average of the box was 15 with an SD of 5.3. But over the years tickets have been lost and other tickets have been added.

Joe believes that the average of the box is still 15 but Mike thinks that it must have changed. They don't want to look at all the tickets again, so they agree to draw a simple random sample from the box and settle their disagreement by looking at the sample average...

Data:

(*) Sample size: $n = 400$.

(*) Sample Average: $\overline{x} = 15.6$.

(*) Sample standard deviation: $s = 5.2$.

Mike: “$15.6 \neq 15$... I’m right!”

Joe: “No... $15.6 - 15 = 0.6$, that’s small: that difference can be explained by chance error. The sample SD is 5.2!”

$\Rightarrow$ *Why not use the original box SD of* 5.3*?*

Mike: “Hmmm... No, wait! The sample SD is not the right estimate for the chance error. We need to look at the *SE*!”

Joe: (grumbling) “Ok...”

$$SE = \frac{SD(box)}{\sqrt{n}} \approx \frac{SD(sample)}{\sqrt{n}} = \frac{5.2}{20} = 0.26.$$

Mike smiles and Joe frowns... *Why?*

Because $15.6 - 15 = 0.6 > 2 \times 0.26$, and it is very unlikely to see a sample average that is more than 2 SEs away from the expected value = box average.

Confidence interval approach:

$$\text{sample average} \pm 2SE = 15.6 \pm 0.52 = (15.08, 16.12).$$

There is a 95% chance that this interval contains the box average, but this interval does *not* contain the number 15, so it seems unlikely that the box average is 15.

In other words, if the box average is in fact 15, then there is only a 5% chance that we produce a 95%-confidence interval that *does not* contain the number 15.

There are 2 possible explanations for the observed results:

(1) *The average of the box hasn't changed:* The observed difference is due to chance error and Joe and Mike have just observed something *very unlikely*.

(2) *The average of the box has changed.*

Judgement call: If the 'very unlikely' is *too unlikely*, we choose the second option.

Tests of significance.

*A **test of significance** is a statistical procedure for determining the likelihood that the difference between an **observed value** and a hypothetical **'expected' value** is due to chance.*

- The expected value comes from the ***Null Hypothesis*** — a hypothesis about the composition of the box-model for the data. In many (but not all) applications, researchers expect the null hypothesis to be false, and are trying to collect statistical evidence to contradict it.
- The ***P-value*** of the test is the probability that the difference between the observed value and the expected value is ***due to chance***. The P-value is also called the ***observed significance level*** of the test.
- **Common terminology**: If the P-value is between 1% and 5%, the results are said to be ***significant***, and if P is 1% or less, the results are called ***highly significant***. If P is small enough, the null hypothesis is traditionally said to be ***rejected*** in favor of the ***alternative hypothesis*** (the opposite of the Null Hypothesis).

The steps:

1. Formulate the null and alternative hypotheses ***in terms of a box-model*** and a parameter associated with this box model (e.g., the average of the box or the percentage of $\boxed{1}$s in the box).

2. Choose an appropriate ***test statistic*** to measure the difference between the observed value(s) and null-hypothetical expected value(s).

3. Collect the data, and calculate the value of the test statistic.

4. Find the significance level (p-value) — this is the probability that the difference between the value of the sample statistic and the null-hypothetical expected value is due to ***chance error***. The nature of the box-model tells us how to do this.

Comment: In many scenarios, the test statistic follows the normal distribution. These types of significance tests are called z-tests.

5. Summarize the results and draw any appropriate conclusions.

Example 1: Joe and Mike's box.

(*) ***Box model:*** Their box; the parameter is the average of the box, μ.

(*) ***Hypotheses:***

$H_0 : \mu = 15$ (this is the null hypothesis)

$H_A : \mu \neq 15$ (this is the alternative hypothesis).

(*) ***Test statistic:*** $z_0 = \dfrac{\overline{x} - \mu_{H_0}}{SE}$

where $\overline{x}$ = sample average and $SE = \dfrac{\text{box } SD}{\sqrt{\text{sample size}}} \approx \dfrac{\text{sample } SD}{\sqrt{\text{sample size}}}$.

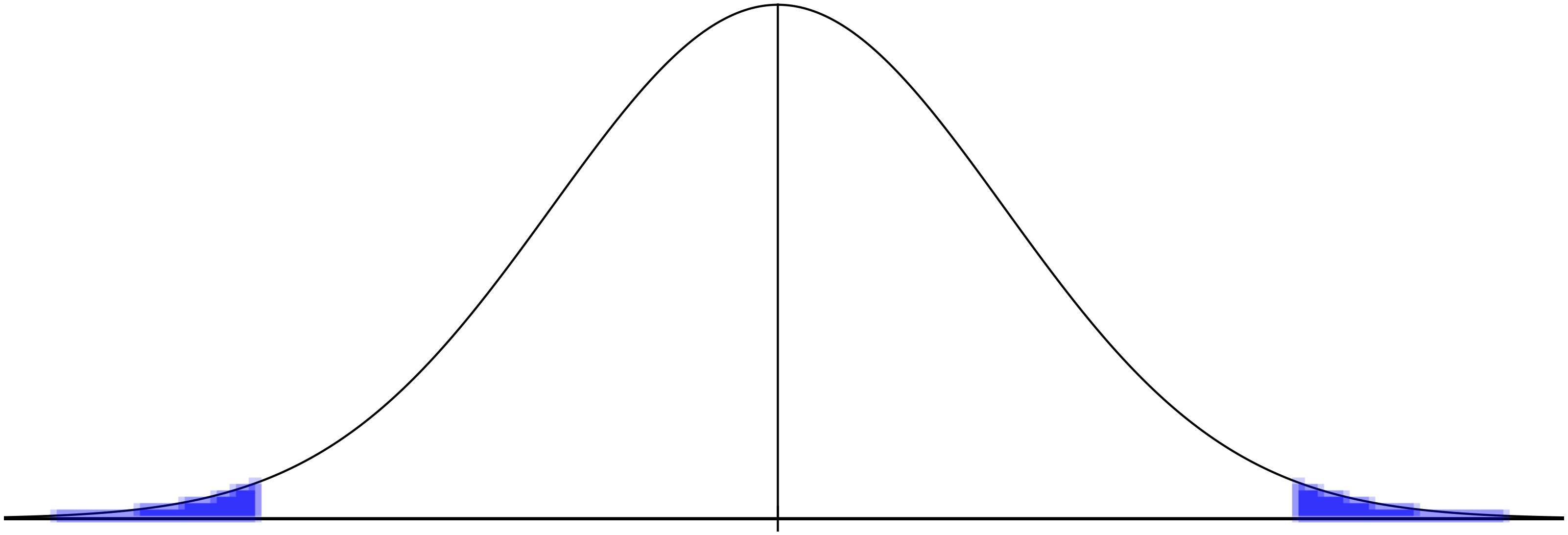
(*) ***Data*** $\Rightarrow$ observed value of test statistic: $z \approx \dfrac{15.6 - 15}{0.26} \approx 2.3$

(*) ***P-value:*** p = area under the normal curve *outside* the interval $(-2.3, 2.3)$, which is equal to $100\% - 97.86\% = 2.14\%$

Conclusion: The P-value is small enough (the results are *statistically significant*) that we reject H_0 and conclude that $\mu \neq 15$.

(*) The P-value is the probability of the test-statistic being *as extreme as* or *more extreme than* the observed value.

(*) The region of the normal curve used to calculate the P-value comes from the form of the alternate hypothesis. In this case $H_A : \mu \neq 15$, no direction is specified: it is the *size* of $|z| = |\overline{x} - \mu|/SE = 2.3$ that determines the value of p



$(\overline{x} - \mu)/SE < -2.3$ $(\overline{x} - \mu)/SE > 2.3$

This is an example of a *two-tailed test.*

Example 2:

- The story: an investigator (Kathy) from a marketing research company believes that the average number (μ) of connected devices per household in Metropolis is greater than 4. To test this belief she begins with...
- **Box model**: each household in Metropolis corresponds to a ticket in a box. The number on the ticket is the number of connected devices in the household. Then she formulates the
- **Null Hypothesis.** $H_0 : \mu = 4$ (some people have $H_0 : \mu \leq 4$)

  and the
- **Alternative hypothesis.** $H_A : \mu > 4$.

  Next she determines the...
- **Test statistic.** In this case, the test statistic is $z = \dfrac{\text{sample average} - \mu}{SE(avg)}$, where the value of μ is prescribed by H_0.

(*) This test statistic will follow the normal distribution if the sample is a ***simple random sample*** and the sample size is ***big enough***.

- ***Finally***... She collects data and does the calculations.

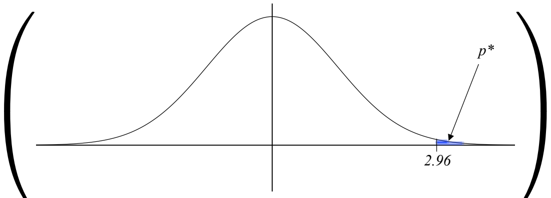
  Sample: A simple random sample of $n = 1600$ households.

  Sample average: $\bar{x} = 4.2$.

  Sample standard deviation: $SD = 2.7$.

- **Test statistic:** $z = \dfrac{4.2 - 4}{SE(avg)} \approx \dfrac{4.2 - 4}{2.7/\sqrt{1600}} \approx 2.96.$

- **P-value**:

  $p = \text{area}$ $\Bigg($  $\Bigg)$ $\approx 0.16\%$

  Comment: The calculation of p is based on (i) the nature of the test statistic (z follows the normal distribution) and (ii) the alternative hypothesis which specifies $\mu > 4$, so we only consider the area under the normal curve *to the right of* $z = 2.96$.

- **Conclusion:** The result is *highly significant*, $p < 1\%$. The likelihood is tiny that the difference between the sample average and the H_0-average is due to chance error, so we reject the null hypothesis.

Comments:

- The point of a test of significance is to see how strong the (statistical) evidence is ***in favor of the alternative hypothesis***.
- If the investigator believes that the parameter is specifically higher or lower than the null-hypothetical value, then a ***one-tailed*** test is appropriate. This means that the P value is computed by looking at the area either to the the right (or to the left) of z, as in the previous example.
- If an investigator is testing whether the 'truth' is different than the null hypothesis, but doesn't have specific expectations as to higher or lower, she should use a ***two-tailed*** test, as in the first example.

  This means that to compute the P value from the observed value of the test statistic z, we look at the area under the normal curve under ***both*** tails: less than $-|z|$ and greater than $|z|$.

⋆ One-tailed tests produce lower P values for the same value of z. Be sure that a one-tailed test is appropriate before citing one-tailed P values.

Example 3.

Five hundred readings are made of *span gas* with known CO concentration of 70 ppm, using a *spectrophotometer*.

(*) The average of the measurements is $\overline{x} = 70.1$ with standard deviation $SD = 2.07$.

Question: Does the machine need to be calibrated?

To answer, use the *Gauss Model* (for measurement error):

$$\boldsymbol{measured\ CO\ concentration} = \boldsymbol{70\ ppm} + \boldsymbol{bias} + \boldsymbol{chance\ error}$$

Box Model: The chance error behaves like random draws from a box with average 0, unknown SD (and a distribution that is approximately normal).

- ***Null hypothesis: bias*** $= 0$. I.e., the spectrophotometer is properly calibrated, the variation in the measured concentrations is due to chance error. ⇒ The ***expected average*** of the measurements is 70 ppm.
- ***Alternative hypothesis: bias*** $\neq 0$. There is bias, but *a priori* we don't know the direction of the bias. (The *spectrophotometer* needs to be calibrated.)

Test statistic:

$$z = \frac{\text{observed average} - H_0\text{-expected average}}{SE(\text{average})} = \frac{0.1}{2.07/\sqrt{500}} \approx 1.08$$

Recall: $SE \approx (\text{sample SD})/\sqrt{\text{sample size}}$.

P-value: The test statistic follows the normal curve (approximately) and the alternative hypothesis is *two-sided*, so $p =$ area under normal curve outside of $(-1.08, 1.08) \approx 28\%$.

(*) This is the probability of observing a test statistic as big as (or bigger than) the observed value, ***assuming that the null hypothesis is true.***

Conclusion: 'Fail to reject H_0' — the spectrophotometer is good to go.

Reality check: 500 measurements? Really?

In a practical, real-world setting, many fewer measurements are usually taken.

Example 4.

Five readings are made of *span gas* with known CO concentration of 70 ppm, using a *spectrophotometer*.

The measurements were: 74, 73, 69, 76, 70.

Question: Does the machine need to be calibrated?

Following the *Gauss Model* (again):

$$\textbf{\textit{measured CO concentration}} = \textbf{\textit{70 ppm}} + \textbf{\textit{bias}} + \textbf{\textit{chance error}}$$

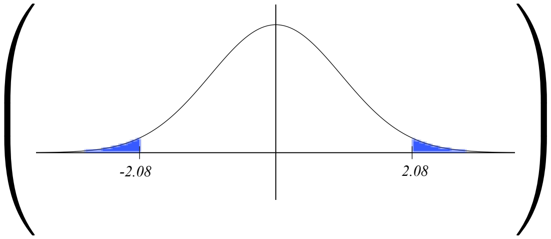
Box Model: The chance error behaves like random draws from a box with average 0, unknown SD (and a distribution that is approximately normal).

- ***Null hypothesis: bias*** = 0. I.e., the spectrophotometer is properly calibrated, the variation in the measured concentrations is due to chance error. ⇒ The ***expected average*** of the measurements is 70 ppm.

- ***Alternative hypothesis: bias*** ≠ 0. There is bias, but *a priori* we don't know the direction of the bias. (The *spectrophotometer* needs to be calibrated.)

The test:

- Average $= 72.4$; SD ≈ 2.58.
- Test statistic: $\dfrac{72.4 - 70}{2.57/\sqrt{5}} \approx 2.08$.
- P-value:

$p = \text{area}$  $\approx 4\%$

- Conclusion: The probability that the difference between observed and expected averages is due to chance error is low. *Recalibrate?*

Concern: *The sample size is **small!***

Problems:

- ***Problem 1.*** Sample size is small, so sample SD is ***likely to underestimate*** the SD of the 'error box'.

⇒ *This is true for all samples, but the difference is negligible when the sample size is large.*

- ***Solution 1.*** Use SD$^{+} = \sqrt{\frac{n}{n-1}}\times$SD...

$$SD^{+} = \sqrt{5/4} \times 2.57 \approx 2.87.$$

- ***Problem 2.*** The test statistic

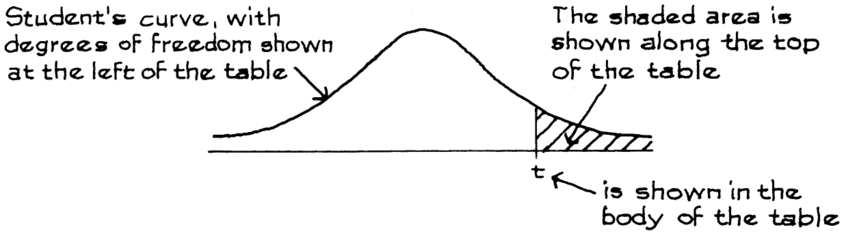
$$t = \frac{\text{observed} - \text{expected}}{SE} = \frac{\text{observed} - \text{expected}}{SD^{+}/\sqrt{n}}$$

does ***not*** follow the normal distribution...

- ***Solution 2.*** The test statistic ***does*** follow *Student's t-distribution* with $n-1$ degrees of freedom... ***as long as the box of errors has a normal distribution***.

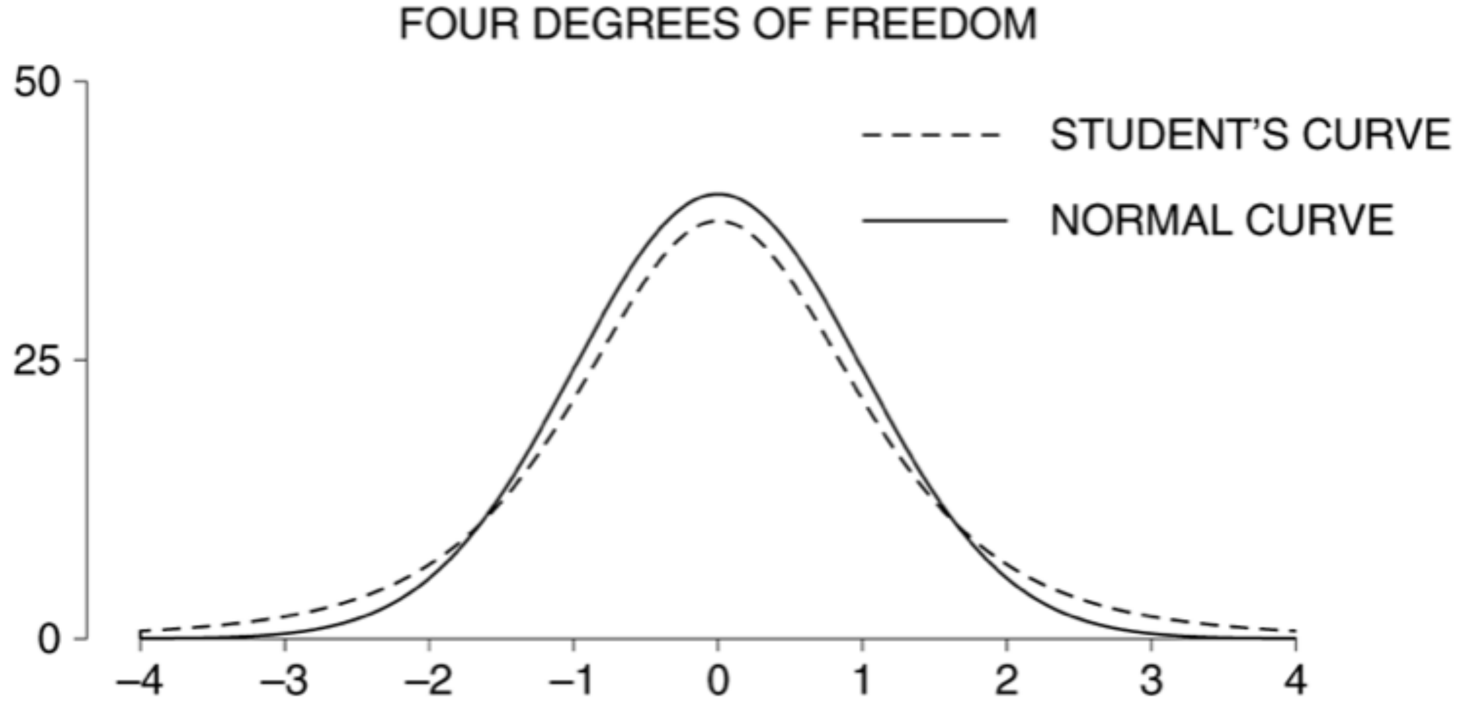
- We find P-values for the t-distributions from a '*t-table*'.
- The t-table is read differently than the normal table:
  - There is one row for every number of d.f.
  - The columns correspond to specific P-values — they give the t-value to the right of which the area under the t-curve is equal to the column header.

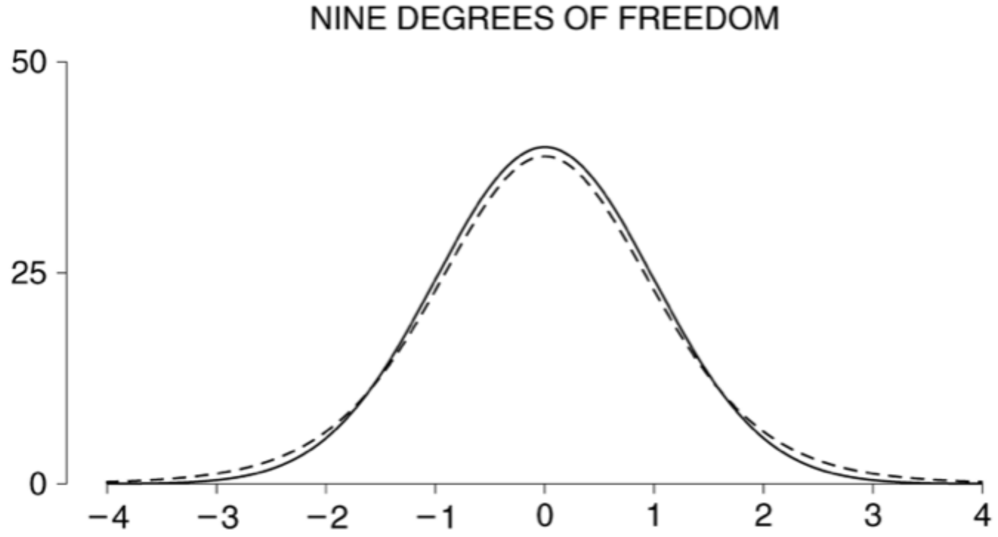
A t-TABLE

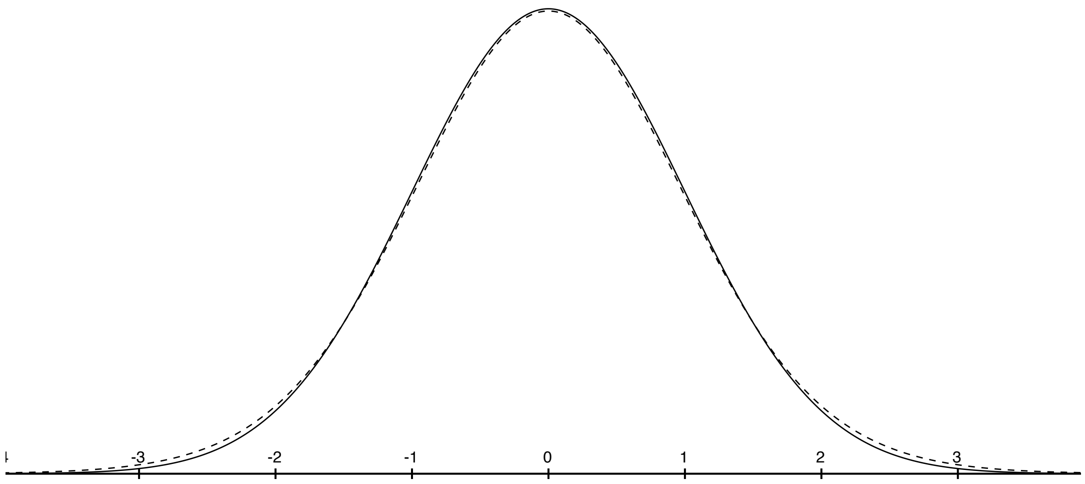

Student's curve, with degrees of freedom shown at the left of the table
The shaded area is shown along the top of the table
t
is shown in the body of the table

Degrees of freedom	*25%*	*10%*	*5%*	*2.5%*	*1%*	*0.5%*
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17

- The t-curves have the same general shape as the normal curve, but with *fatter tails* $\Rightarrow$ large values of $|t|$ are more likely (bigger P-values), than the same values of $|z|$.
- When the sample size (and so d.f.) is large, the difference between the t-curve and the normal curve becomes small (and perhaps negligible).



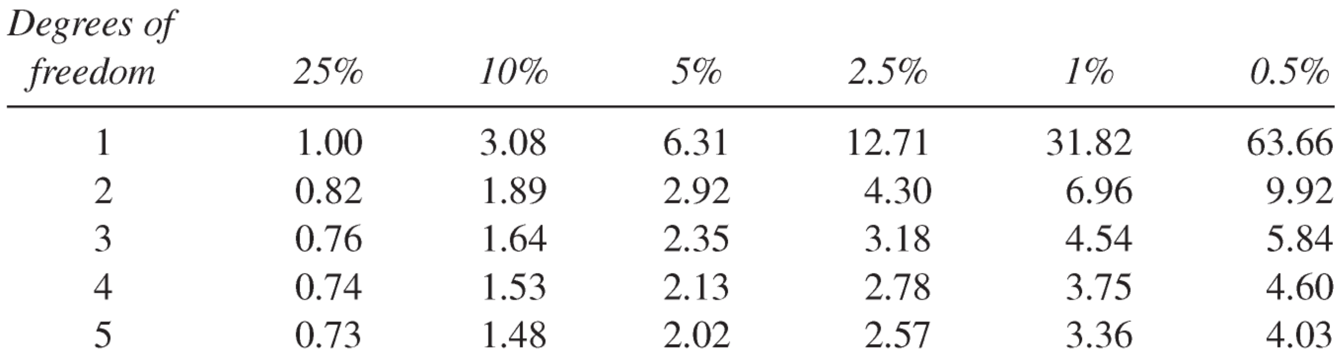




22 degrees of freedom

Back to the example...

- $t^* = \dfrac{72.4 - 70}{2.87/\sqrt{5}} \approx 1.87$
- The P-value is estimated from the row in the t-table corresponding to $5 - 1 = 4$ degrees of freedom:



Degrees of freedom	*25%*	*10%*	*5%*	*2.5%*	*1%*	*0.5%*
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03

- $t = 1.87$ falls between the columns corresponding to (the ***one-sided*** P-values) 10% and 5%.
- ***P-value:*** The probability that the difference between observed and expected is due to chance error is between 10% and 20% (actually $p \approx 13.5\%$), because this is a two-sided test.
- ***Conclusion:*** There is probably no need to recalibrate.

Summary:

- If the sample size n is large enough, then
$$z = \frac{\text{observed average} - \text{box average}}{\text{sample SD}/\sqrt{n}}$$
follows the normal curve reasonably well.

- If the *original* box follows the normal curve, then for *any* sample size $n > 1$,
$$t = \frac{\text{observed average} - \text{box average}}{SD^+/\sqrt{n}}$$
follows the Student t-distribution, with $n - 1$ degrees of freedom.