Example.

- In 2010 the median household income in Metropolis (1.5 million households) was \$33,000.
- Lex Luthor, running for reelection as mayor of Metropolis in 2016, claims that the policies of his administration have raised the median household income. As evidence, he cites a recent survey of 550 randomly selected households, where the median income was \$35,000.

  Is this a valid claim?
- To test this claim with a test of significance, we need a box model.

⇒ *The sampling distribution of the **median** is complicated, if the original population is not known to be normal.*

- Another look at the sample data: 290 of the sample households had incomes over \$33,000.

⇒ Use a zero-one box model:

$\boxed{1}$ ↔ household in Metropolis with income over \$33,000.

$\boxed{0}$ ↔ household in Metropolis with income under \$33,000.

- ***Null hypothesis:*** The median income in Metropolis did ***not*** change between 2010 and 2016.

$\Rightarrow$ H_0 : 50% of the tickets in the Metropolis box are $\boxed{1}$s.

- ***Alternative hypothesis:*** The median income in Metropolis increased between 2010 and 2016.

$\Rightarrow$ H_A : More than 50% of the tickets in the Metropolis box are $\boxed{1}$s.

- ***Test statistic:***

$$z = \frac{\text{observed percentage} - \text{expected percentage}}{SE(\%)}.$$

This test statistic follows the normal distribution.

- The standard error is computed based on the null hypothesis. In this case (a null hypothesis about percentages), then null hypothesis also tells us what the SD of the box is.

$$H_0 : \quad SD = \sqrt{0.5 \times 0.5} = 0.5.$$

- ***Data:*** Simple random sample of 550 households, observed percentage of $\boxed{1}$s $\dfrac{290}{550} \approx 52.7\%$.

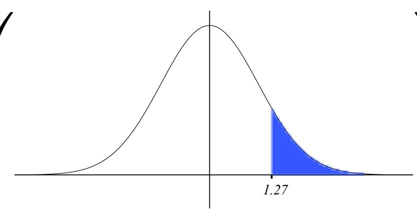
$\Rightarrow$ Observed value of z:

$$z^* = \frac{52.7\% - 50\%}{(0.5/\sqrt{550}) \times 100\%} \approx 1.27.$$

.

- ***P-value:*** p^* = area under the normal curve ***to the right of*** $z^* = 1.27$:

$$p^* = \text{area} \Bigg($$



$$\Bigg) \approx 10.5\%$$

(*) This is a ***one-tailed*** test because the alternative hypothesis specifies *new median > old median.*

- **Conclusion:** The results are *not* statistically significant (at the 5% significance level) — not super strong evidence in support of Lex Luthor's claim.

Comments:

- Typically, the point of the test is to provide statistical evidence ***in favor of the alternative hypothesis***.
- If an investigator expects the 'truth' to be different than the null hypothesis, but doesn't have specific expectations as to higher or lower, she should use a ***two-tailed*** test.

  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 $|z|$.
- If the true value (average, percentage, etc.) is thought to be specifically higher or lower than the null-hypothetical value, then a ***one-tailed*** test is more appropriate.

  This means that the P value is computed by looking at the area either to the left or to the right of z, as appropriate.
- 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.

The Gauss model for measurement error.

- Repeated measurements are made of the same quantity.
- Each measurement differs from the *truth* by ***chance error***.
- The chance error is like drawing a ticket at random from a box—the ***error box***. Independent measurements correspond to draws done with replacement.
- The error box has an average of 0, and usually an unknown SD. In practice the SD of the error box is inferred from the SD of the data.
- In many cases, it is assumed that the error box has a normal distribution.
- If there is no bias, then the average of independent measurements provides an accurate estimate of the *truth*.
- With bias: ***measured value*** = ***true value*** + ***bias*** + ***chance error***.

  This model can be used to test for ***bias***, assuming that the quantity being measured is known, or has a required value.

Example.

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

The measurements were: 75, 73, 70, 77, 70.

Question: Does the machine need to be calibrated?

Following the *Gauss Model* (for measurement error):

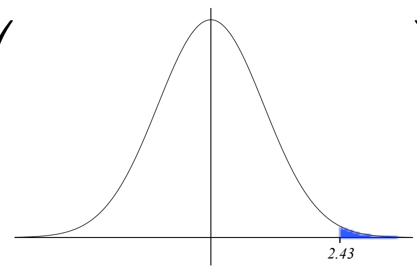
$$\textbf{\textit{measured CO concentration}} = \textbf{\textit{70 ppm}} + \textbf{\textit{bias}} + \textbf{\textit{chance error}}$$

The chance error behaves like random draws from a box with average 0, unknown SD and a probability distribution that is approximately normal. If the machine is properly calibrated, then the bias should be 0. If there is nonzero bias, then the machine needs to be calibrated.

- ***Null hypothesis: bias = 0***. I.e., the variation in the measured concentrations is due to chance error. (The *spectrophotometer* is properly calibrated)
- ***Alternative hypothesis: bias*** $\neq$ **0**. There is bias. (The *spectrophotometer* needs to be calibrated.)

The test:

- Average = 73; SD $\approx$ 2.76.
- Test statistic: $\dfrac{73-70}{2.76/\sqrt{5}} \approx 2.43.$
- P-value:

$p^* = \text{area}\Bigg($ 

$\Bigg) \approx 0.75\%$

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

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 sample sizes, but the difference is negligible when the sample size is large.*

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

$$SD^+ = \sqrt{5/4} \times 2.76 \approx 3.09.$$

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

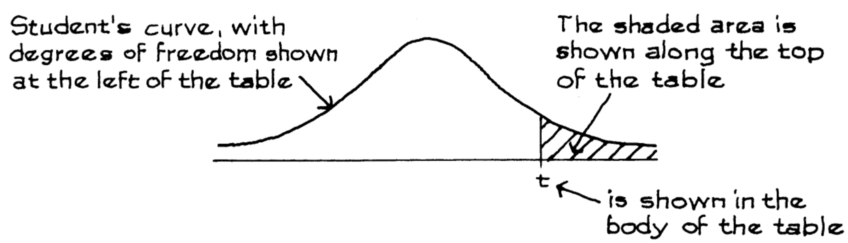
$$t = \frac{\text{observed} - \text{expected}}{SE} = \frac{\text{observed} - \text{expected}}{SD^+/\sqrt{n}}$$

doesn't follow the normal distribution...

- ***Solution 2.***... it ***does*** follow the ***t-distribution*** with $n-1$ degrees of freedom, ***as long as the box of errors has a normal distribution***.

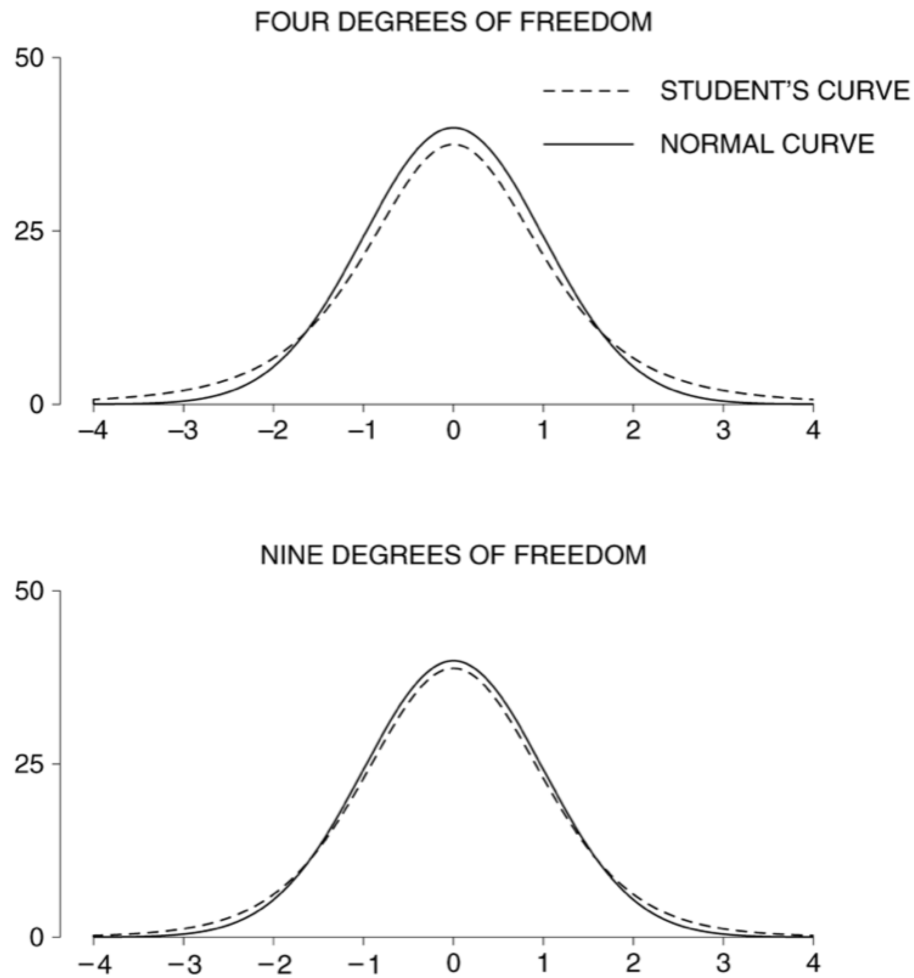
- We find P-values for the t-distributions from a '*t-table*'.
- The t-table is read differently than the normal table:
  - There is one row for every number of d.f.
  - The columns correspond to specific P-values — they give the t-value to the right of which the area under the t-curve is equal to the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

A t-TABLE



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*.
- When the sample size (and d.f.) is large, the difference between the t-curve and the normal curve becomes small (and perhaps negligible).



Back to the example...

- $t^* = \dfrac{73 - 70}{3.09/\sqrt{5}} \approx 2.17$
- 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

- $t = 2.17$ falls between the columns corresponding to 2.5% and 5%...
- Conclusion: if there is no bias, then the chance of observing an average of 73 ppm or more in 5 measurements on span gas is between 2.5% and 5% (actually, $p^* = 4.79\%$, using an online calculator).
- Recalibrate? That depends on the operational protocols of the Lab that uses the spectrophotometers.