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 is computed based on the **Null Hypothesis** — a hypothesis about the composition of the box-model for the data. In many, if not most examples, 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.

- 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 deemed **highly significant**. If P is small enough, the null hypothesis is said to be **rejected** in favor of the **alternative 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 and appropriate **test statistic** to measure the difference between the observed value(s) and null-hypothetical expected value(s). Then, collect the data, and calculate the value of the test statistic.

3. 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.

4. Summarize the results and draw any appropriate conclusions.
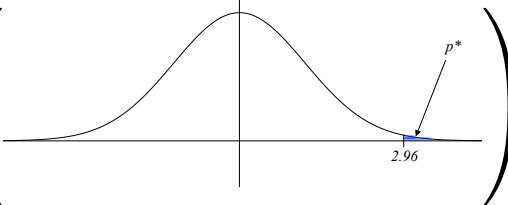
Example 1:

- The story: an investigator (Joe) 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 he 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 he formulates the ...

- **Null Hypothesis.** $H_0 : \mu = 4$ and the...

- **Alternative hypothesis.** $H_A : \mu > 4$. Next he 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...* He collects data and does the calculations.

Example 1 (cont):

- **Sample;** A simple random sample of $n = 1600$ households.

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

  Sample standard deviation: $SD = 2.7$.

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

- **P-value:**

$$p = \text{area} \left(\underset{\substack{\\ 2.96}}{\includegraphics{}} \,\,^{p*} \right) \approx 0.16\%$$

  Comment: The calculation of p is based on the nature of the test statistic (z follows the normal distribution) and 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 probability that the sample average is greater than 4 due to chance error is so small that we reject the null hypothesis in favor of the alternative.

Example 2.

- 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.

$\Rightarrow$ *The sampling distribution of the **median** is complicated if the original population is not known to be normal.*

Example 2.

- 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:**

$$z = \frac{\text{observed percentage} - H_0\text{-expected percentage}}{SE(\%)}.$$

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 should be (if H_0 is correct).

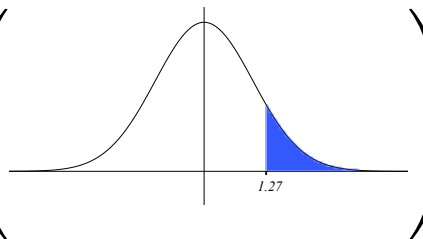$$H_0 : \quad SD = \sqrt{0.5 \times 0.5} = 0.5.$$

- **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** $z^* = 1.27$:

$$p^* = \text{area} \left(\begin{array}{c} \includegraphics \\ {\scriptstyle 1.27} \end{array} \right) \approx 10.5\%$$

(*) 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 strong evidence in support of Lex Luthor's claim. *We **fail to reject** the null hypothesis.* (But we don't accept it.)

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 examples.

- 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.

  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.

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: 74, 73, 69, 76, 70.

Question: Does the machine need to be calibrated?

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

measured CO concentration = 70 ppm + bias + 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.

- **Null hypothesis:** $\mu = 70$. I.e., the spectrophotometer is properly calibrated, the variation in the measured concentrations is due to chance error.

- **Alternative hypothesis:** $\mu \neq 70$. There is bias, but we don't know the direction of the bias. (The *spectrophotometer* needs to be calibrated.)

(*) μ = average of 'all' measurements.

The test:

- Average $= 72.4$; SD ≈ 2.58.

- Test statistic: $\dfrac{72.4 - 70}{2.57/\sqrt{5}} \approx 2.08$.

- P-value:

$$p = \text{area} \left(\vcenter{\hbox{}} \right) \approx 4\%$$

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

  Comment: This is a two-sided test. (The one-sided p-value is 2%).

Problems:

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

$\Rightarrow$ *This is true for all sample sizes, but the difference is negligible when the sample size is large.*

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

$$SD^+ = \sqrt{5/4} \times 2.57 \approx 2.87.$$
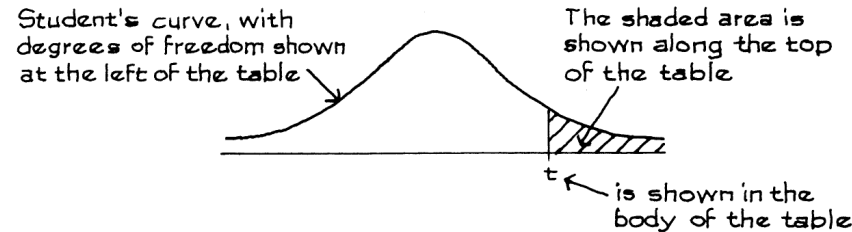
- ***Problem 2.*** The test statistic

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

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

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*.

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

FOUR DEGREES OF FREEDOM

----- STUDENT'S CURVE

——— NORMAL CURVE

NINE DEGREES OF FREEDOM
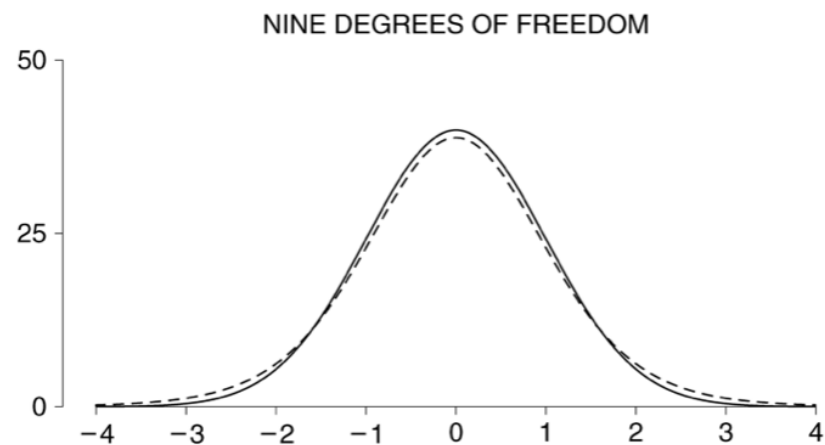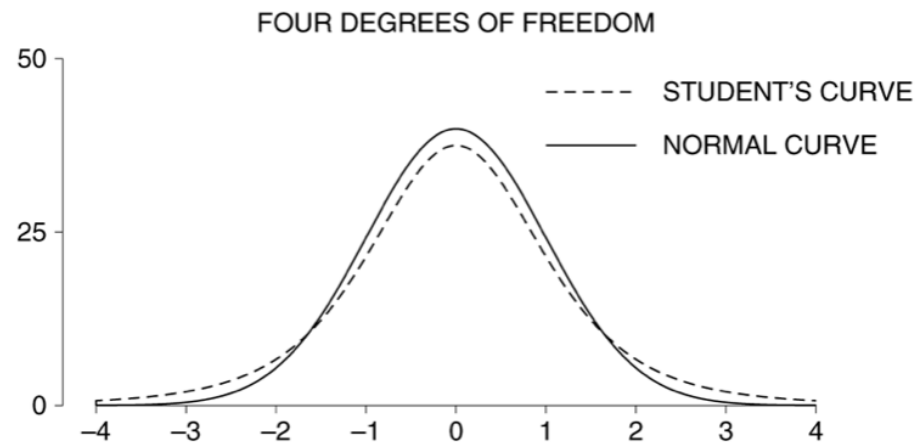
Back to the example...

- $t^* = \dfrac{72.4 - 70}{2.87/\sqrt{5}} \approx 1.87$

- 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 10% and 5%...

- ***Conclusion:*** The probability that the observed measurements are due to chance error is between 10% and 20% (actually $p \approx 13.5\%$)... because this is a two-sided test.

- There is probably no need to recalibrate — but this depends on the operational protocols of the Lab that uses the spectrophotometers.