

TESTS OF SIGNIFICANCE - MOTIVATING EXAMPLE

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures students' study habits and attitude toward school. Scores range from 0 to 200. The mean score for U.S. college students is about 115. A teacher suspects that older students have better attitudes toward school. She gives the SSHA to an SRS of 35 students who are at least 30 years old. The sample results are $\bar{X} = 125.7$ and $s = 30.1$. Is this good evidence that older students, on average, have better study habits and attitudes toward school than the typical college student?

THE MAIN CONCEPTS OF HYPOTHESIS TESTING

A statistical test begins by supposing for the sake of argument that the effect we seek is not present. We then look for evidence against this supposition and in favor of the effect we hope to find.

- For the null hypothesis, H_0 , state a claim that we will try to find evidence against. The null hypothesis is often a statement of "no effect" or "no difference". Nothing special has occurred, no change has taken place -- the "status quo" hypothesis.
- The statement we hope or suspect is true instead of H_0 is the alternative hypothesis, H_a .

A significance test looks for evidence against the null hypothesis and in favor of the alternative hypothesis. The evidence is strong if the outcome we observe would rarely come up when the null hypothesis is true.

That is, if the sample results can easily occur when H_0 is true, we attribute the relatively small discrepancy between the null hypothesis and the sample results to chance.

If the sample results cannot easily occur when H_0 is true, we explain the relatively large discrepancy between the null hypothesis and the sample results by concluding that H_0 is not true (and so we conclude that H_a is true).

- Hints:
- (1) The null hypothesis will always contain equality.
 - (2) It's often easier to write down the alternative hypothesis first.
 - (3) P-value helps us assess the amount of evidence the sample provides against H_0 and in favor of H_a . P-value tells us how unlikely the sample results are when H_0 is true. Very small p-values mean the sample results are very unlikely to occur when H_0 is true and therefore the evidence against H_0 is strong.
 - (4) Language: Based on the p-value, we either "reject H_0 in favor of H_1 " or we "fail to reject H_0 ." (Sometimes I say "retain H_0 ," instead of "fail to reject H_0 .)

Guidelines for p-value

P-value is defined as the probability of obtaining sample results as extreme (or more extreme) as those actually obtained, if H_0 were true. ("Extreme" means far from what we would expect if H_0 were true. The alternative hypothesis determines which directions count against H_0 .)

For example, $p\text{-value} = .02$ means sample results like those obtained only occur 2% of the time when H_0 is true.

P-value helps us assess the amount of evidence the sample provides against H_0 and in favor of H_a . P-value tells us how unlikely the sample results are when H_0 is true. Very small p-values mean the sample results are very unlikely to occur when H_0 is true and therefore the evidence against H_0 is strong.

Language: Based on the p-value, we either "reject H_0 in favor of H_a " or we "fail to reject H_0 ." (Sometimes I say "retain H_0 " instead of "fail to reject H_0 .)

The smaller the p-value, the stronger is the evidence against H_0 . The following can be used as guidelines when a significance level is not preset. They should not be viewed as p-value "cutoffs."

| | |
|----------------------------------|--------------------------------------|
| $p\text{-value} > .1$ | insufficient evidence against H_0 |
| $.05 < p\text{-value} \leq .10$ | some evidence against H_0 |
| $.01 < p\text{-value} \leq .05$ | fairly strong evidence against H_0 |
| $.001 < p\text{-value} \leq .01$ | strong evidence against H_0 |
| $p\text{-value} \leq .001$ | very strong evidence against H_0 |

Reporting a test of significance

1. Give the null and alternative hypotheses. Define the parameters involved in the study.
2. Summarize the sample data for your readers.
3. Give the test statistic and its distribution, the observed test statistic, and the p-value.
4. Use the p-value to draw a conclusion - reject the null hypothesis in favor of the alternative or retain the null hypothesis. State your conclusion in context of the problem.

Exercise: Each of the following situations calls for a significance test for μ , μ_d , or $\mu_1 - \mu_2$. For now, just state the null hypothesis H_0 and the alternative hypothesis H_a in each case. Also, define the parameters involved in the study (μ , μ_d , or μ_1 and μ_2) using the context of the problem. (We will finish working the problems with SAS.)

1. A fire insurance company felt that the mean distance from a home to the nearest fire department in a suburb of Chicago was at least 4.7 miles. It set its fire insurance rates accordingly. Members of the community set out to show that the mean distance was less than 4.7 miles. This, they felt, would convince the insurance company to lower its rates. They randomly identified 64 homes and measured the distance to the nearest fire department for each. The resulting sample mean was 4.4 miles and the sample standard deviation was 2.4 miles. Does the sample show sufficient evidence to support the community's claim? If yes, estimate the average distance from homes to the nearest fire department.

2. At Farmer's Dairy, a machine is set to fill 32-ounce milk cartons. Of course, the amount varies slightly from carton to carton but when the machine is working properly, the mean net weight of these cartons is 32 ounces. The quality control director at this dairy takes a sample of 35 such cartons each week to see if filling should be paused so the machine can be stopped and adjusted for overfilling or underfilling. (Both are undesirable since underfilling cheats the customers and overfilling costs the dairy money.) A recent sample of 35 cartons produced a mean net weight of 31.90 ounces and a standard deviation of .15 ounces. Based on this sample, would you conclude that the machine needs to be adjusted?

If you conclude that the machine needs to be adjusted, estimate the current fill weight for the machine so the quality control team can make the appropriate adjustments to get the machine in good working condition again. Use a 95% confidence interval and interpret your interval estimate. (For example, is the machine overfilling or underfilling, and by how much?)

3. National Paper Company must purchase a new machine for producing cardboard boxes. The company must choose between two machines. Since the machines produce boxes of equal quality, the company will choose the machine that produces the most boxes in a one-hour period. The company selected eight assembly workers to test the two types of machines. The table below gives the number of boxes produced in an hour on each type of machine for each of these eight workers. Based on the data, is there sufficient evidence that the mean number of boxes produced in an hour differs for the two machines?

If there is a significant difference, estimate the difference with a 98% confidence interval. Interpret your interval estimate.

| Machine Operator | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|------------------|----|----|----|----|----|----|----|----|
| Machine 1 | 53 | 60 | 58 | 48 | 46 | 54 | 62 | 49 |
| Machine 2 | 50 | 55 | 56 | 44 | 45 | 50 | 57 | 47 |

4. Ordinary corn doesn't have as much of the amino acid lysine as animals need in their feed. Plant scientists have developed varieties of corn that have increased amounts of lysine. In a test of the quality of high-lysine corn as animal feed, an experimental group of 20 one-day-old chicks at a ration containing the new corn. A control group of another 20 chicks received a ration that was identical except that it contained normal corn. The weight gains after 21 days are given below.

Perform a hypothesis test to decide if high lysine corn is effective in increasing weight gain.

If you find that the high lysine corn does yield higher weight gains, estimate the mean extra weight gain with a 95% confidence interval.

| Control (normal corn) | | | | Experimental (high lysine corn) | | | |
|--------------------------|-----|-----|-----|------------------------------------|-----|-----|-----|
| 380 | 321 | 366 | 356 | 363 | 449 | 403 | 377 |
| 283 | 349 | 402 | 462 | 436 | 405 | 395 | 428 |
| 356 | 410 | 329 | 399 | 408 | 320 | 469 | 409 |
| 350 | 384 | 316 | 272 | 429 | 422 | 479 | 394 |
| 345 | 455 | 360 | 431 | 432 | 341 | 412 | 328 |

5. A production engineer is investigating whether there is a difference in the washer diameters manufactured by two different methods. A random sample of washers from the production line that uses the first method yields the following diameters (in inches).

0.861 0.864 0.882 0.887 0.858 0.879 0.887 0.876 0.870
 0.894 0.884 0.882 0.869 0.859 0.887 0.875 0.863 0.887
 0.882 0.862 0.906 0.880 0.877 0.864 0.873 0.860 0.866
 0.869 0.877 0.863 0.875 0.883 0.872 0.879 0.861

The second method produces washers with the following diameters (in inches).

0.705 0.703 0.715 0.711 0.690 0.720 0.702 0.686 0.704
 0.712 0.718 0.695 0.708 0.695 0.699 0.715 0.691 0.696
 0.680 0.703 0.697 0.694 0.714 0.694 0.672 0.688 0.700
 0.715 0.709 0.698 0.696 0.700 0.706 0.695 0.715

Do the data show a significant difference in the average diameters for the two methods? If yes, estimate the difference with a 95% confidence interval. Interpret your interval estimate.

6. Advertisements for an instructional video claim that the techniques will improve the ability of Little League pitchers to throw strikes. To investigate this claim, we have 20 Little Leaguers throw 50 pitches each, and we record the number of strikes. After the players participate in the training program, we repeat the test. The table shows the number of strikes each player threw before and after the training.

Perform a hypothesis test to decide if the instructional video seems effective. If the instructional video is effective, use a 95% confidence interval to estimate the level of improvement in the ability of Little League pitchers to throw strikes.

| Number of strikes (out of 50) | | Number of Strikes (out of 50) | |
|----------------------------------|-------|----------------------------------|-------|
| Before | After | Before | After |
| 28 | 35 | 33 | 33 |
| 29 | 36 | 33 | 35 |
| 30 | 32 | 34 | 32 |
| 32 | 28 | 34 | 30 |
| 32 | 30 | 34 | 33 |
| 32 | 31 | 35 | 34 |
| 32 | 32 | 36 | 37 |
| 32 | 34 | 36 | 33 |
| 32 | 35 | 37 | 35 |
| 33 | 36 | 37 | 32 |

7. An experiment was conducted to evaluate the effectiveness of a treatment for tapeworm in the stomachs of sheep. A random sample of 24 worm-infected lambs of approximately the same age and health was randomly divided into two groups. Twelve of the lambs were injected with the drug and the remaining twelve were left untreated. After a 6-month period, the lambs were slaughtered and the following worm counts were recorded.

Perform a hypothesis test to decide if the treatment is effective in reducing the occurrence of tapeworm in sheep. If you conclude that the treatment is effective, estimate the average reduction in tapeworm count with a 95% confidence interval.

| Drug-Treated Sheep | Untreated Sheep |
|-------------------------------|----------------------------|
| 18 | 40 |
| 43 | 54 |
| 28 | 26 |
| 50 | 63 |
| 16 | 21 |
| 32 | 37 |
| 13 | 39 |
| 35 | 23 |
| 38 | 48 |
| 33 | 58 |
| 6 | 28 |
| 7 | 39 |