Tests of Significance

In Sweden, a controlled experiment was carried out to compare the mortality rate due to prostate cancer for those who have surgery and for those who choose watchful waiting.

For this study, from October 1989 to February 1999, 695 men with newly diagnosed prostate cancer were randomly assigned to watchful waiting or radical prostatectomy groups. 348 were assigned to the group to have the operation and 347 were assigned to watchful waiting. These subjects were followed through the year 2000.

During the time of the study 31 of 348 assigned to watchful waiting died of prostate cancer, while only 16 of the 347 assigned to radical prostatectomy died of prostate cancer, representing a 50% reduction in the death rate from prostate cancer.

Considering deaths from any cause, 62 of the 348 men in the watchful waiting group died and 53 in the radical prostatectomy group died.

Are these differences significant, or could they have been due to chance?

Emily Rosa, 11, conducted experiment two years ago

4th-Grader's Study Makes AMA Journal

It questions power of healing touch

By Terrence Mommonry and Louis Sahagun
Los Angeles Times

Loveland, Colo.

San Francisco Chronicle
April 1, 1998

Using little more than a towel and a piece of cardboard, a 9-year old girl conducted a "brilliant" study debunking therapeutic touch, an increasingly popular alternative treatment practiced by some 40,000 nurses and caregivers in the United States.

Along the way, Emily Rosa, now 11, apparently has become the youngest researcher to publish a scientific paper in the prestigious Journal of the American Medical Association. She co-authored the final report—which appears today—with her parents and a physician who specializes in uncovering medical fraud.

In a test that started out as her fourth-grade science project, Emily recruited 21 practitioners of therapeutic touch and found that they could not reliably detect another person's "energy field," contrary to one of the practice's central tenets.

She zeroed in on the idea that practitioners can sense another person's "energy field" with their own hands. Practitioners have described patients' energy as feeling cold, hot, sticky, tingling or throbbing, among other things.

In the study, each therapist sat across from Emily at a table, laying his or her arms out flat, palms up. A cardboard partition with cut-out armholes placed over their forearms blocked their view of their hands and of Emily. A towel draped over their arms also prevented peeking.

The test consisted of Emily placing one of her hands a few inches above a therapist's right or left hand, as determined by the flip of a coin. If the therapist could sense which hand better than 50 percent of the time, that would support the theory. Fourteen practitioners got 10 tries each, while seven got 20 tries. Overall, the average correct score was 44 percent, which is less than what would be expected by chance alone. "They were correct about half the time—about what you'd expect from guessing," Emily said. "Of course they came up with excuses. One said the room was too cold. Another complained that the air conditioning blew the force field away."

Taken together, the lack of supportive studies plus these new findings "suggest that (therapeutic) touch claims are groundless and that further use of the technique by health professionals is unjustified," Emily and her co-authors wrote.

"I think of me as a kid who did a simple science experiment," said Emily, an avid Spice Girls fan and budding flamenco dancer who lives with her mother, a registered nurse, and father, a mathematician and inventor, in this semirural town north of Denver.

"Age is irrelevant," the journal's editor, Dr. George D. Lundberg, said of the investigator's youth. "It's the quality of the science that matters."

Given the new findings, Lundberg urged in an editorial that patients should "save their money and refuse to pay for this procedure until or unless additional honest experimentation demonstrates an actual effect." .

Proponents of therapeutic touch disputed the study's importance, criticizing its premise and setup. They also said that the study is hardly dispassionate, because Emily's mother, Linda Rosa, a registered nurse, is an avowed critic who has spent years amassing evidence and lobbying against the procedure's acceptance.

Still, the study represents a strong challenge to a practice that has grown tremendously since it was first proposed in the 1970s as a modern version of the ancient laying on of hands. Practioners claim to promote healing by holding or moving their hands a few inches above a patient's body, which is said to realign "energy fields"' disrupted by illness.

Professional organizations such as the National League for Nursing and the American Nurses' Association have promoted therapeutic touch or energy healing, and some 80 hospitals in North America reportedly offer the treatment. The North America; Nursing Diagnostic Asioeiation recognizes "energy field disturbance" as a health problem.

Healing Touch International, a Colorado group, says the treatment can help with a range of illnesses and symptoms, including AIDS, multiple-sclerosis, cancer and arthritis. In a much publicized program at the Columbia University College of Physicians and Surgeons, therapeutic touch practitioners accompany patients during open-heart surgery.

Scientific evidence supporting the practice or the theories behind it has been elusive; despite many articles over the years reporting successful cases.

Emily and her co-authors say she may have been able to overcome practitioners' reluctance to subject themselves to testing "because the person conducting the test was a child who displayed no skepticism."

Summary

The test consisted of Emily placing one of her hands a few inches above a therapist's right or left hand, as determined by the flip of a coin. If the therapist could sense which hand better than 50 percent of the time, that would support the theory. Fourteen practitioners got 10 tries each, while seven got 20 tries, so there were 280 tries in all. Overall, the average correct score was 44 percent, which is less than what would be expected by chance alone. "They were correct about half the time—about what you'd expect from guessing," Emily said.

What would you expect from guessing? It's like drawing from what kind of box?

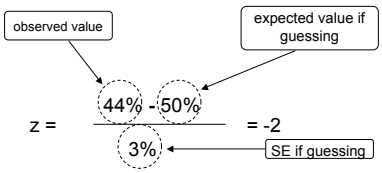
If guessing,

EV of % in sample =

SE of % =

So would you expect 44% if they are guessing?

The z-statistic



The **null hypothesis**: they are guessing

It says the difference is due to chance. There is nothing going on.

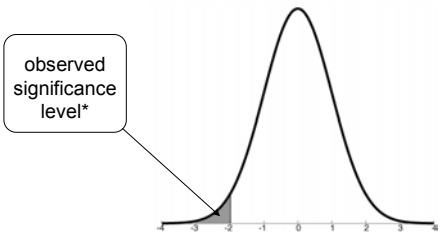
z is an example of a **test statistic**: it is used to measure the difference between the data and what is expected under the null hypothesis.

It is how many SEs the result is away from what it's expected value would be ***if*** the null hypothesis were true.

If z is very large there are two possibilities:

- The null hypothesis is true but a very rare event happened
- The null hypothesis is false

The larger the z-value the less plausible the former



P ~ .025

*also called "P-value"

Does the z-value of -2 and the p-value of 2.5% support the null hypothesis or the hypothesis of the therapists?

What if the z-value had been z=1 and p-value = 16%?

What if the z-value had been z=2 and the p-value = 2.5%?

What if the z-value had been z=3 and the p-value = .13%?

What if the z-value had been z=-3 and the p-value = .13%?

The Logic of Significance Testing

The **null hypothesis** says that there is no effect other than chance.

The **alternative hypothesis** says that there is an effect other than chance.

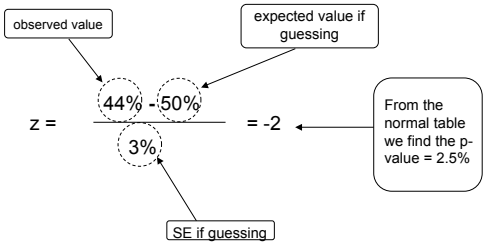
In order to prove that there is some kind of effect, you have to disprove the possibility that the results could be due to chance. Disprove the null hypothesis in order to establish the alternative hypothesis. You have to convince a skeptic.

The **P-value** is the chance of a z statistic as or more extreme than that observed occurring if the null hypothesis is true.

Small P-values cast doubt upon the null hypothesis.

The P-value is **not** the probability that the null hypothesis is true.

The P-value was calculated from the test statistic, z



The z-value is just how many SE's the observed value is from the expected value

Example

A newspaper carried a story reporting that a high school student got 9207 heads and 8743 tails in 17,950 coin tosses. Is this a significant discrepancy from the null hypothesis that the coin is fair?

What would we expect if the null hypothesis were true? The experiment would be like 17,950 draws with replacement from what box?

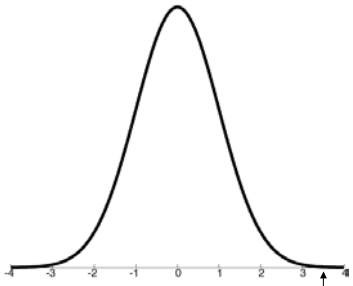
9207 heads in17,950 tosses

Null hypothesis: chance of a head is 1/2

Expected # heads =

SE of # heads =

$$z = \frac{\text{Observed value - expected value if null hypothesis true}}{\text{SE if null hypothesis is true}}$$



P-value = .0003

Example: Prostate Cancer Study

During the time of the study 31 of 348 assigned to watchful waiting died of prostate cancer, while only 16 of the 347 assigned to radical prostatectomy died of prostate cancer, representing a 50% reduction in the death rate from prostate cancer.

Considering deaths from any cause, 62 of the 348 men in the watchful waiting group died and 53 in the radical prostatectomy group died.

A method of testing significance

Assume that the surgery has no effect so that the chance of dying during the study is the same for the two groups. Then the 16 + 31 = 47 subjects in the watchful waiting group who died of prostate cancer would be equally likely to be in each group. Thus the subjects who died from prostate cancer that were in the watchful waiting group can be considered the result of tossing a coin 47 times and getting heads 31 times.

P-value: What is the chance of getting 31 or more heads when a coin is tossed 47 times?

EV =

Under the null hypothesis, SE =

$$Z =$$

P-value = .014

For testing significance of death from any cause, a similar computation gives P-value = .2, so the study does not establish that the surgery reduces the chance of death from any cause.

z-test for numerical data

Hypothetical example: items produced by a manufacturer have an average value with respect to some measurement of 100. A sample of 64 from one lot has an average value of 100.25 and an SD equal to 2.

Is there evidence of a problem with that lot?

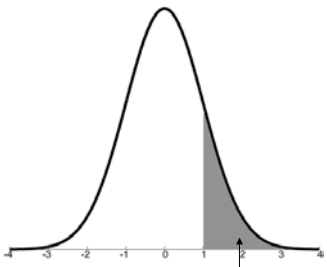
If the null hypothesis is true:

EV of sample average =

SE of sample average =

Test statistic:

$z =$.



P-value = 16%

P-Values and Tradition

A result with a P-value less than 5% is often called “statistically significant” or “significant at the .05 level.”

A result with a P-value less than 1% is often called “highly significant” or “significant at the .01 level.”

Some journals will not publish results unless they are significant at the .05 level or better.

Small Samples

The z-test is based on the central limit theorem: the probability histogram of the sum or average of a large number of draws can be approximated by the normal curve.

What if the sample size is small, like 10 or 15?

If you can assume that the histogram of the box (the population) is very close to normal, then there is an alternative - the **t-test**.

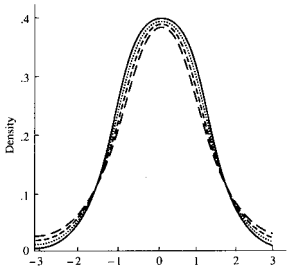
Rather than the z-statistic, you use the t-statistic:

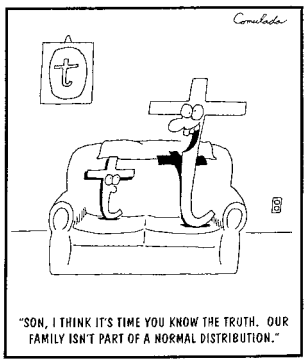
$$t = \frac{\text{observed - expected}}{\text{SE}}$$

The SE is found from a slightly different sample SD. (See book for details.)

Rather than a normal curve table a t-curve table is used. The “degrees of freedom” equal the sample size minus one.

Three t densities with 5 (long dashes), 10 (short dashes), and 30 (dots) degrees of freedom and the standard normal density (solid line).



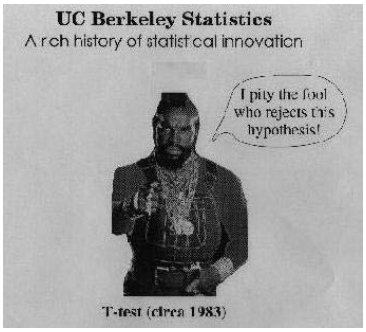


Despite its popularity in elementary statistics textbooks, the realm of applicability of the t test is quite limited.

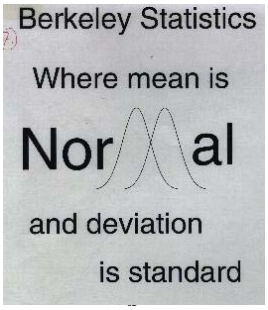
For sample sizes more than 25 or so, the results of the t-test and z-test are practically identical.

For small sample sizes, the assumption that the population is normal is very important. But how would you know?

The 1998 Statistics Department T-shirt



Another Classic Department T-Shirt



Which SE?
Some Examples

1. Last year 20% of the students at a large university had GPAs over 3.6. In a simple random sample of 400 students taken this year 25% had GPAs over 3.6. How strong is this evidence for grade inflation?

$$Z = \frac{\text{observed value} - \text{expected value under null}}{\text{SE of value under null}}$$

What value? Could use observed % (25%) or observed number (.25 x 400 =100).

How do we get SE under null? In this example the null specifies the box (20% 1's and 80% 0's) and from this we can get it's SD. The SD of the box is then used to find the SE of a number or a percent.

2. In 1994 the national average on verbal SAT was 423 with an SD equal to 110. A simple random sample of 300 scores were taken in a state, giving an average of 444 and an SD of 100. Did the state do better on average, or is this just chance variation?

observed value = 444 expected value = 423

Which SD to use to find SE of observed value – 110 or 100?

We want to use the SE of the observed value if the null is true. What exactly is the null? Two possibilities:

- The null says that the average is unchanged but says nothing about the SD. In this case use the SD of the sample – 100.
- The null says that the average and the SD are both unchanged. In this case use SD = 110.

The former seems more reasonable in this case.

Summary

A **significance test** is aimed at determining whether a result is real or could possibly be due to chance.

The **null hypothesis** says that the result is due to chance. A probability calculation is made under the null hypothesis via a box model.

A **test statistic** measures the difference between the data and what would be expected under the null hypothesis. We used the **z-statistic** and mentioned the **t-statistic.**

The **observed significance level,** or **P-value**, is the chance of getting a test statistic as or more extreme than the one observed if the null hypothesis were true.

Small P-values are evidence against the null hypothesis.

The z-statistic relies on the central limit theorem, so the sample size can't be too small.

If the sample size is small and the population is normally distributed, the t-statistic can be used.

Review Questions

1. Which of the following P-values is best for the null hypothesis: 1%, 5%, 25%

2. True or false:

   (a) The observed significance level depends on the data.

   (b) If the observed significance level is 5% there are 95 chances in 100 that the alternative hypothesis is correct.

3. True or false:

   (a) A result with a P-value of .001 cannot be due to chance.

   (b) In such a case the chance the null hypothesis is true is .001 or less.

   (c) In such a case the chance the alternative is true is .999 or better.

   (d) If there is no effect, the chance of such a P-value is .001.