Hypothesis Testing: Means

Roger B. Davis and Kenneth J. Mukamal

Circulation 2006;114;1078-1082
DOI: 10.1161/CIRCULATIONAHA.105.586461
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0009-7322. Online ISSN: 1524-4539

Hypothesis Testing
Means

Roger B. Davis, ScD; Kenneth J. Mukamal, MD, MPH

In most biomedical research, investigators hypothesize about the relationships of various factors, collect data to test those relationships, and try to draw conclusions about those relationships from the data collected. In many cases, investigators test relationships by comparing the average level of a factor between 2 groups or between 1 group and a standard reference. This framework is as true for understanding the basic role of cardiac myosin binding protein-C phosphorylation in cardiac physiology[1] as it is for evaluating non–high-density lipoprotein cholesterol (HDL-C) as a predictor of myocardial infarction in large groups of individuals.[2] In this article we describe hypothesis testing, which is the process of drawing conclusions on the basis of statistical testing of collected data, and the specific approach used to test means (or average levels of a collected data element). These concepts are covered in detail in many statistical textbooks at various levels, including Pagano and Gauvreau,[3] Zar,[4] and Kleinbaum et al.[5]

Hypothesis Testing

The purpose of statistical inference is to draw conclusions about a population on the basis of data obtained from a sample of that population. Hypothesis testing is the process used to evaluate the strength of evidence from the sample and provides a framework for making determinations related to the population, ie, it provides a method for understanding how reliably one can extrapolate observed findings in a sample under study to the larger population from which the sample was drawn. The investigator formulates a specific hypothesis, evaluates data from the sample, and uses these data to decide whether they support the specific hypothesis.

The first step in testing hypotheses is the transformation of the research question into a null hypothesis, H_0, and an alternative hypothesis, H_A.[6] The null and alternative hypotheses are concise statements, usually in mathematical form, of 2 possible versions of "truth" about the relationship between the predictor of interest and the outcome in the population. These 2 possible versions of truth must be exhaustive (ie, cover all possible truths) and mutually exclusive (ie, not overlapping). The null hypothesis is conventionally used to describe a lack of association between the predictor and the outcome; the alternative hypothesis describes the existence of an association and is typically what the investigator would like to show. The goal of statistical testing is to decide whether there is sufficient evidence from the sample under study to conclude that the alternative hypothesis should be believed.

Hypothesis testing has been likened to a criminal trial, in which a jury must use evidence to decide which of 2 possible truths, innocence (H_0) or guilt (H_A), is to be believed. Just as a jury is instructed to assume that the defendant is innocent unless proven otherwise, the investigator should assume there is no association unless there is strong evidence to the contrary. A jury's verdict must be either guilty or not guilty, in which case a not-guilty verdict does not equal innocence. Rather, it indicates that the burden of proof has not been met. Similarly, an investigator can only reject H_0 or fail to reject it; failure to reject does not prove that the null H_0 is true.

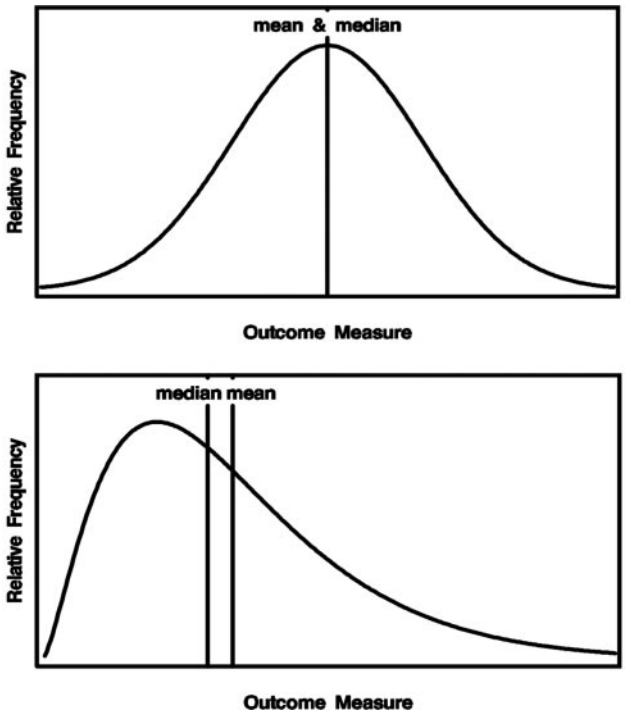
In a criminal trial in the United States, the required burden of proof is "beyond a reasonable doubt." For hypothesis testing, the investigator sets the burden by selecting the level of significance for the test, which is the probability of rejecting H_0 when H_0 is true. The standard value chosen for level of significance is 5% (ie, $P=0.05$), which is a much weaker standard than used in the criminal justice system. This standard means that even if no association between predictor and outcome exists in the population, the investigator is willing to accept a 1 in 20 chance of a false-positive conclusion that an association does exist.

Just as hypothesis testing can reject a true null hypothesis (referred to as a type I error), it can fail to reject H_0 when the predictor and outcome are associated (type II error). The probability of such a false-negative conclusion is called β. The quantity $(1-\beta)$ is called the power of the test and is simply the probability of drawing the correct conclusion (ie, rejecting H_0) when an association between predictor and outcome actually does exist.

In most cases, investigators are equally interested in whether a predictor leads to higher or lower levels of the outcome. In this situation, we specify a 2-sided statistical test, in which we accept a combined rate of false-positives (for both the higher and lower level of the outcome) of only 5%. If only 1 direction is of interest, a 1-sided test may be appropriate, but this choice requires strong justification. Because a 1-sided test is less stringent, many readers (and journal editors) appropriately view 1-sided tests with skepticism.[7] Two-sided tests should also be considered the default

From the Division of General Medicine and Primary Care, Beth Israel Deaconess Medical Center, Boston, Mass.
Correspondence to Roger B. Davis, ScD, Division of General Medicine and Primary Care, Beth Israel Deaconess Medical Center, 330 Brookline Ave, RO-108, Boston, MA 02215. E-mail rdavis@bidmc.harvard.edu
(*Circulation.* **2006;114:1078-1082.**)

Circulation **is available at http://www.circulationaha.org** **DOI: 10.1161/CIRCULATIONAHA.105.586461**



Hypothetical frequency distributions of variables with normal (top) and right-skewed (bottom) distributions.

option because an investigator's intuition about how a study will come out may be incorrect. If an investigator chooses a 1-sided test but observes results opposite to those expected, the strongest statement that can be made is that the null hypothesis was not rejected. For these reasons, the investigator should always specify the hypotheses, the methods of analysis, and the level of significance before initiating the research.

Means

In clinical practice and in biomedical research, we collect substantial amounts of numerical data. To analyze such data correctly, it is critical to recognize the different types of numerical data and the various methods specific to each type. Stevens[8] proposed 4 classes of measurement scales: nominal scales use numbers strictly as labels for categories with no natural ordering; ordinal scales represent categories with a natural ranking; interval scales use numbers in a truly quantitative sense in which differences between observations are meaningful (eg, temperature); and ratio scales are interval scales that also have a meaningful zero value (eg, height).

The mean of a measure for a population is simply its arithmetic average. It is usually denoted by μ. The mean from the sample that we actually observe, usually designated by $\bar{x}$, is the sum of the observed measurement for each individual in the sample, divided by n, the number in the sample. The mean is an appropriate measure for ordinal and ratio scales but not for nominal or ordinal scales.[4]

The Figure shows 2 theoretical distributions of data. The first pattern follows a normal distribution. The distribution is symmetrical (ie, the right-hand side is a mirror image of the left-hand side), and the mean and median occur at the same value. Many characteristics we observe approximate this pattern, such as height or HDL-C. The second distribution is skewed and asymmetrical; there are more observations far to the right of the mean than there are far to the left. The mean of this distribution is larger than its median, because the extreme values to the right increase the mean but do not affect the median. This general pattern is seen in the distributions of C-reactive protein, triglycerides, and coronary artery calcification, as well as medical costs and hospital length of stay. Analysts often perform logarithmic transformation of right-skewed variables like these to improve their fit to a normal distribution.

Although the mean can be skewed by extreme values, there are important reasons why it is the most commonly used measure of "center" in statistical testing. First, when the distribution of a measurement is reasonably symmetrical, statistical tests of the mean tend to have the most power (ie, when differences between groups exist, these tests are most likely to detect them). Second, for some measurements, we may want the center to reflect the pull of extreme values. For example, when measuring health care costs, we may want the "average" expenditure to reflect the almost inevitable presence of a few subjects with very high costs.[9] In such a case, the mean multiplied by the sample size recreates the total expenditure in the sample, but the median does not.

One-Sample Tests

In some research projects, the study design includes only a single sample, and the goal may be to determine whether the outcome measure for the population from which the sample was drawn has same mean as some standard population. Determining an appropriate standard for comparison for these designs is often an issue. Nonetheless, when well-established standards exist, investigators may wish to use these standards for maximal comparability. In this situation, we might perform a 1-sample (not 1-sided) t test.

To provide a concrete example, we examine data from a trial of black tea consumption in 28 adults (Table). As a preliminary step, we might be interested in testing whether the population from which these individuals derive tends to have baseline levels of HDL-C that differ from the overall US population as a test of their generalizability. The distribution of HDL-C in the US adult population is well characterized and has a mean of ≈50.7 mg/dL.[10] Therefore, we would want to determine whether the data from our 28-person sample support a conclusion that the population from which these older adults came has HDL-C levels that differ from 50.7 mg/dL. We would state the null and alternative hypotheses as follows:

$$H_0: \mu_{HDL\text{-}C} = 50.7$$

$$H_A: \mu_{HDL\text{-}C} \neq 50.7$$

To decide which of these hypotheses we believe, we first calculate the mean and standard deviation (SD; a measure of the "spread" or variability of the measurement) of baseline HDL-C in the sample. These are called $\bar{x}$ and s, respectively.

$$\bar{x} \text{ (sample mean)} = 63.2$$
$$s \text{ (sample standard deviation)} = 13.7$$
$$n \text{ (sample size)} = 28$$

Baseline and 6-Month HDL-C Levels Among 28 Participants in a Trial of Black Tea Consumption

ID	Sex	Baseline HDL-C, mg/dL	6-Month HDL-C, mg/dL
1	Female	64	74
2	Female	60	70
3	Female	59	65
4	Male	65	67
5	Female	64	62
6	Male	62	67
7	Male	54	51
8	Female	68	93
9	Female	67	56
10	Female	79	78
11	Female	45	58
12	Male	48	52
13	Female	59	60
14	Female	65	76
15	Female	87	74
16	Male	49	36
17	Male	46	42
18	Male	46	50
19	Female	97	79
20	Male	36	35
21	Male	67	60
22	Female	56	58
23	Male	62	57
24	Female	65	68
25	Female	65	60
26	Female	81	89
27	Female	83	58
28	Female	71	70

We then calculate the t statistic, as follows:

$$t = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})}$$

In this equation, the numerator is the difference between the observed sample mean HDL-C and the hypothesized mean if the null hypothesis is true (ie, 50.7). The denominator is the standard error, a measure of the variability of the sample mean. The farther the t statistic is from zero, the stronger the evidence that H_A is true. Put differently, we would conclude that the evidence against the null hypothesis is strong if the sample mean is far from the standard value compared with the inherent variability of the sample mean.

$$t = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})} = \frac{63.2 - 50.7}{(13.7/\sqrt{28})} = \frac{12.5}{2.6} = 4.83$$

To decide between H_0 and H_A, we compare the t statistic to the t distribution with $(n-1)$ *df*. Tables that provide critical values of the t distribution are available in introductory statistical texts and are published online.[11] For a 2-sided test at the 5% level of significance, the critical value of the t distribution with 27 *df* is 2.05. This value has an important interpretation; specifically, if H_0 is true (ie, the sample was truly drawn from a population with $\mu_{HDL\text{-}C} = 50.7$ mg/dL), 95% of samples of this size (n=28) will produce a t statistic between -2.05 and 2.05. Therefore, if the t statistic for our sample is >2.05 or <-2.05, we reject H_0 and conclude that the population from which these participants came has HDL-C levels that differ from the general population. If t is between -2.05 and 2.05, there is not enough evidence to refute the default assumption that this group's HDL-C is the same as in the general population. As seen in the calculation of the 1-sample t test, $t=4.83$, so we reject H_0 and conclude that our sample has a different HDL-C than does the general US population.

The mathematical derivation of the test statistic assumes that the mean HDL-C of the sample $\bar{x}$ is normally distributed. This assumption is satisfied if the outcome we are measuring (in this case HDL-C) is itself normally distributed. The t test performs reasonably well even if the underlying distribution of the measure deviates moderately from normality, a characteristic referred to as the test's robustness. Even if the underlying distribution of the measure itself deviates substantially from normality, the distribution of the mean typically approximates normality when the sample size is large, a result called the central limit theorem. How large is large enough is a complex question, but as a practical matter, statisticians seem reasonably comfortable with samples of 60 to 100 in most circumstances.

When the normality of the distribution is in question and the sample size is too small to invoke the central limit theorem, one relies on different, nonparametric tests such as the Wilcoxon signed rank test. Nonparametric tests (a topic that will be covered in a future article in this series) do not compute test statistics on the basis of the observed values of the outcome but rather on their rank ordering within the sample. Although these tests also examine the location of the distribution, they compare medians rather than means, and they tend to have less statistical power than t tests when the underlying distribution truly is normal.

Two-Sample Tests

Many studies obtain data from 2 samples and seek to test whether the means of the 2 populations represented by the samples are different. Typically, the statistical hypotheses are as follows:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Selection of the appropriate 2-sample statistical test depends on the study design, specifically whether the 2 samples are paired or independent of each other. In a paired design, each observation in 1 sample is linked in some way to 1 specific observation in the other sample. Examples include designs in which each individual is measured both before and after an intervention or studies of treated participants matched to individual untreated controls. Independent samples have no link between specific observations in the 2 samples.

Paired Tests

Whenever research is designed as a matched or paired study, the appropriate analysis takes the matching into account. The paired t test is the standard method for comparing means of paired samples. For each matched pair of observations, we

compute the difference between them, d_i. Note that if the 2 groups have the same mean (ie, if H_0 is true), we would expect the differences between pairs to center around zero. We next compute the mean and SD of the paired differences. The test statistic is

$$t = \frac{\bar{d}}{(s/\sqrt{n})}$$

in which the numerator is the mean of the paired differences, and the denominator is the standard error of $\bar{d}$. This test is identical to the 1-sample t test of H_0: $\mu_d = 0$.

The reason for designing a matched study is to eliminate a potential source of variability in the outcome being measured. This advantage is lost if the appropriate test is not performed. For matched designs, the paired t test will generally have greater statistical power than the equivalent test for independent samples if the matching is appropriate. However, if the matching criterion is not associated with the outcome measure, the matching is ineffective (ie, does not reduce a source of variability) and will not improve power.

In our tea study, we measured HDL-C levels in each participant 6 months apart. These measurements are obviously paired within each participant, and hence they comprise 28 pairs of data points. The Table shows the baseline and 6-month HDL-C levels for each of the participants. To test the hypothesis that HDL-C levels changed over time, we could test whether the baseline value of each pair differs substantially from the 6-month value. For each matched pair, we calculate the difference in levels of HDL-C. We calculate the mean and SD of the differences for the 28 pairs and use these to calculate the paired t statistic as follows:

$$H_0: \mu_{\text{baseline}} = \mu_{\text{6-month}}$$
$$H_A: \mu_{\text{baseline}} \neq \mu_{\text{6-month}}$$

$$d_1 = 74-64 = 10$$
$$d_2 = 70-60 = 10$$
$$\vdots$$
$$d_{28} = 70-71 = -1$$

$$\bar{d} = -\frac{5}{28} = -0.179$$

$$s = 10.234$$

$$t = \frac{\bar{d}}{(s/\sqrt{n})} = \frac{-0.179}{10.234/\sqrt{28}} = -0.09$$

We compare -0.09 to the t distribution with 27 degrees of freedom. From this, we determine $P=0.93$, so we fail to reject the null hypothesis and conclude HDL-C did not change from baseline to 6 months.

Independent Samples

When the data in the 2 samples are not matched, tests for independent samples are appropriate. Usually the assumption is made that the distributions in the 2 groups have the same variance, σ^2. Essentially, we assume that the predictor under investigation shifts the distribution of the outcome to the left or right but does not change its variability. This assumption can be tested by comparing the ratio of the estimated variances in the 2 groups to the F distribution (details are beyond the scope of this article).[4] Some statistical software packages automatically conduct this test when a 2-sample t test is requested. This test will reject the hypothesis that the variances are equal when the observed ratio is far from 1.0. As a general rule, ratios between 0.5 and 2.0 are acceptable for small samples (<30 per group), as are ratios between 0.67 and 1.5 for moderate samples (<100 per group).

Going back to our tea trial, suppose we want to test the hypothesis that HDL-C levels in men in the trial differ from levels in women, as we would expect. The first step is to compute the means $\bar{x}_1$ and $\bar{x}_2$ and SDs s_1 and s_2 in the 2 samples (ie, in the enrolled men and women).

$$H_0: \mu_{\text{male}} = \mu_{\text{female}}$$
$$H_A: \mu_{\text{male}} \neq \mu_{\text{female}}$$

$$\bar{x}_{male} = 53.5$$
$$s_{male} = 10.2$$
$$n_{male} = 10$$

$$\bar{x}_{female} = 68.6$$
$$s_{female} = 12.6$$
$$n_{female} = 18$$

On the basis of the observed SDs, the data do not provide evidence that the variances in the 2 groups are distinct

$$\left(\frac{s^2_{\text{female}}}{s^2_{\text{male}}}\right) = 1.53,$$

so it is reasonable to assume that HDL-C levels in men and women share a common variance.

Because we can now assume a single common variance, we compute the pooled estimate of the variance as follows,

$$s_p^2 \text{ (pooled variance)} = \frac{(n_m-1)\, s_m^2 + (n_f-1)\, s_f^2}{n_m + n_f - 2}$$
$$= \frac{(9)\,(10.2)^2 + (17)\,(12.6)^2}{26} = 139.8$$

which is a weighted average of the SD squared in the 2 groups of sample sizes n_1 and n_2. The t statistic is then calculated as

$$t = \frac{\bar{x}_m - \bar{x}_f}{s_p\sqrt{\left(\frac{1}{n_m}+\frac{1}{n_f}\right)}} = \frac{53.5-68.6}{11.8\sqrt{\left(\frac{1}{10}+\frac{1}{18}\right)}} = \frac{-15.1}{4.67} = -3.25$$

This test statistic is compared with the t distribution with (n_1+n_2-2) *df*. Just as in the 1-sample test, the numerator of this statistic is the difference between the means of the 2 samples, and the denominator is a measure of the variability of this difference between means. If the difference is large relative to the variability, then there is strong evidence against the null hypothesis of no difference.

We apply these methods to test whether HDL-C levels are the same for men and women in our trial. We compare -3.25

to the t distribution with 26 degrees of freedom. From this, we determine $P=0.003$.

In some circumstances, we should not assume that the 2 populations have equal variances. Because the 2 SDs are no longer assumed to be estimating the same parameter, the test statistic does not use a pooled estimate of the variance. The statistic, called Welch's t, is calculated as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

and is compared with a t distribution. To determine the number of df, we calculate

$$\nu \text{ (the degrees of freedom)} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

and round down to the nearest integer.

Just like the 1-sample t test, the 2-sample t tests assume that the sample means follow a normal distribution but are robust to moderate departures from that assumption. For data that deviate substantially from the normal distribution, there are nonparametric tests such as the Wilcoxon rank sum test. These tests compare the location of each sample's distribution but do not test their means per se.

t Tests and Confidence Intervals

Another concept related to hypothesis testing about means is the confidence interval (CI), which is closely linked to the probability value derived from a t test. A CI for a given mean estimates the range of values that, based on the sample mean and its variability, are likely to include the true population mean μ. In most cases, we are interested in the 95% CI, which corresponds directly to the 5% false-positive rate we accept in standard hypothesis testing.

Summary

In summary, we have described some of the standard methods for testing hypotheses about the means of observed measurements. These methods are appropriate for measures made on ratio or interval scales and include t tests to compare 1 sample to a reference group and to compare 2 paired or 2 independent samples. These methods tend to yield better power than nonparametric alternatives yet are typically robust to the distribution of the measurement being tested, especially when sample sizes are large. Methods for comparing means of >2 groups will be covered later in the series, as will methods for comparing means while adjusting for other factors.

Sources of Funding

The Tea's Effect on Atherosclerosis Pilot Study was funded by grants from the American Heart Association (0355638T) and the National Center for Complementary and Alternative Medicine (R21AT01899). This research was also supported in part by grant RR01032 to the Beth Israel Deaconess Medical Center General Clinical Research Center from the National Institutes of Health.

Disclosures

None.

References

1. Sadayappan S, Gulick J, Osinska H, Martin LA, Hahn HS, Dorn GW II, Klevitsky R, Seidman CE, Seidman JG, Robbins J. Cardiac myosin-binding protein-C phosphorylation and cardiac function. *Circ Res*. 2005; 97:1156–1163.
2. Pischon T, Girman CJ, Sacks FM, Rifai N, Stampfer MJ, Rimm EB. Non-high-density lipoprotein cholesterol and apolipoprotein B in the prediction of coronary heart disease in men. *Circulation*. 2005;112: 3375–3383.
3. Pagano M, Gauvreau K. *Principles of Biostatistics*. Belmont, Calif: Duxbury Press; 1993.
4. Zar JH. *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice Hall; 1999.
5. Kleinbaum DG, Kupper LL, Muller KE. *Applied Regression Analysis and Other Multivariable Methods*. Boston, Mass: PWS-KENT Publishing; 1988.
6. Browner WS, Newman TB, Hearst N. Getting ready to estimate sample size: hypotheses and underlying principles. In: Hulley SB, Cummings SR, Browner WS, Hearst N, eds. *Designing Clinical Research: An Epidemiological Approach*. 2d ed. Philadelphia, Pa: Lippincott Williams & Wilkins; 2001.
7. Ware JH, Mosteller F, Delgado F, Donnelly C, Ingelfinger JA. P values. In: Bailar JC, Mosteller F, eds. *Medical Uses of Statistics*. Boston, Mass: NEJM Books; 1992.
8. Stevens SS. On the theory of scales of measurement. *Science*. 1946;103: 677–680.
9. Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. *Annu Rev Public Health*. 1999;20: 125–144.
10. American Heart Association. *Heart Disease and Stroke Statistics—2005 Update*. Dallas, Tex: American Heart Association; 2005.
11. Upper critical values of the Student's-t distribution. In: *NIST/SEMATECH e-Handbook of Statistical Methods*. Available at: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm. Accessed August 7, 2006.