Testing Hypotheses for STAT112

Charles Fleming

January 23, 2015

If an association of a histogram with a probability distribution is convincingly made, then, in essence, the statistical problem of characterizing the population has been solved. But how can we tell analytically if the association is a valid one? The answer to that question tends to be controversial to say the least as each antagonist strives to make an association which will improve the advantage of his position over his opponent's position especially when a decision must be won in a political arena. The arena need not be a legislative forum or a court of law; it could be and undoubtedly will be the company conference room.

To set the stage for developing the process of testing whether a hypothesis should or should not be rejected, a simple experiment will illustrate that the origins of indecisiveness is a function of the probability of not wanting to commit the wrong decision. The less chance that is prescribed in allowing a commission of error, the less likely that a decision will be made. The level of chance that one is willing to take depends on the stakes at risk. There is no theoretical way to find the right balance between risks and benefits; the manager ultimately must make that decision.

Law 1 (Dow's Law). *In a hierarchical organization, the higher the level, the greater the confusion.*

Toss a coin 100 times and designate: $X_i = \begin{cases} 1 & \text{if a head appears} \\ 0 & \text{otherwise} \end{cases}$. The total number of heads that appear out of 100 tosses is determined by counting the number of 1's that appear in the data. Equivalently, the sum of the X_i's is the number of heads because the tails are represented by $X_i = 0$. Denoting T to be the number of heads that appear out of 100 tosses implies that $T = \sum_{i=1}^{100} X_i$. Each X_i is a Bernoulli random variable, hence $T \sim b(100, p)$.

Problem 1. *Suppose 35 heads were observed after tossing a coin 100 times. Is the coin a fair coin? Does* $P(X_i = 1) = \frac{1}{2}$*?*

If the coin were fair, then $T \sim b(100, .5)$ and that E[T]=np=100(.5)=50 and

$$var(T) = npq = 100(.5)(.5) = 25$$

Instead of 50 heads being observed, only 35 appeared. The probability of the event of getting at most 35 heads given that the coin is fair is: $P(T \leq 35) = .0017588$. This computation for a Binomial random variable was done on a computer because it would have been too arduous a task to do the computation by means of a hand calculator. In the absence of a computer or tables for a Binomial distribution, an approximation of that probability could have been obtained by using the Normal distribution. The usual procedure of finding the z-score leads to the statement: $P(\frac{T-50}{\sqrt{25}} \leq \frac{35-50}{\sqrt{25}}) = P(z \leq -3) = .0013499$. In either case, the probability of getting at most 35 heads is quite small when flipping a fair coin 100 times. The small probability suggests the conclusion that the coin is not fair, otherwise if the coin were in fact fair, the number of observed heads should have been much closer to the expected value of 50.

The construction of a confidence interval is certainly worth investigating for gaining an idea of the location of the true value of p. A more simple approach for our purposes will be to assume that the coin is fair and then find two numbers, a and b, which are symmetrically placed about 50 such that the probability that T is between them is 95 percent. By supposition, p=.5 and using the Normal distribution to approximate T, the lower limit, a=50-$\frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}$. There is only one T so that n=1, $z_{\frac{\alpha}{2}} = z_{.025} = 1.96$, and $\sigma^2 = 25 \to a = 40$. The upper limit, b=50+$\frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}} = 60$.

It is obvious at $\alpha = .05$ that $35 \notin (40, 60)$. This result and the previous one are two indications to substantiate the claim that the coin is not fair. One would be inclined to decide that the hypothesis that the coin is fair should be rejected in favor of the alternative hypothesis that it is not a fair coin.

The discussion thus far elicits the puzzling question concerning the feasibility of determining a definitive answer to a statistical hypothesis.

Law 2 (Walpole & Myers)**.** *The truth or falsity of a statistical hypothesis is never known with certainty, unless we examine the entire population.*

Because a feasible experiment will admit an examination of only a part of the entire population, a statistical conclusion can only be made at a certain level of confidence that it is correct. Our decision can only be based on a sample of a population, and, as a result, the validity of our decision will have some uncertainty in it. In a simple decision, we must either accept or reject a hypothesis. To reject a hypothesis means that based on the data, the hypothesis is false, while to accept the hypothesis means that there is insufficient data to believe otherwise. To accept a hypothesis does not mean that the hypothesis is true. Because of this terminology, we should always state the hypothesis which we hope to reject. The formulation of the hypothesis depends on the context of the problem.

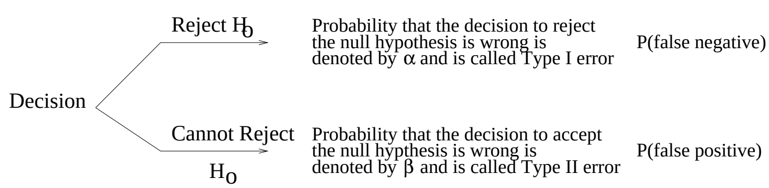
Example 1. *Suppose that you found a coin on the sidewalk. You identified it to be not a U.S. coin, but an ancient Scythian coin; you also know that one variety of that coin is extremely valuable. You know that such a coin is loaded so that p=.35 of getting heads, while the counterfeit variety is fair, p=.5 and is worthless. You are hoping that you have indeed the rare coin. The only expert in the U.S. who can appraise the coin lives in Boise, Idaho. It is now early January. You prefer not to travel to Boise, Idaho on a wild goose chase. An experiment of flipping the coin many times comes to your mind for demonstrating the falsity of the hypothesis that the coin is counterfeit. That is, you hope to reject, based on the experimental data, the hypothesis that p=.5 with the probability, α, that your decision is wrong.*

A translation of the problem into mathematical terminology would be the following:

$$H_0 : p = .5 \quad vs. \quad H_1 : p = .35 \quad at\ a\ level\ of\ significance\ \alpha = .05$$

H_0 denotes the *null hypothesis*, and H_1 denotes the *alternative hypothesis.*

In an experiment, the examination of the population is imperfect. The decision to reject the null hypothesis, based on experimental data might lead to a wrong decision. There are two possible outcomes to every decision as depicted in the following diagram.



Philosophically, it is better to say *cannot reject* than to say *accept* in the same sense that the innocence of an accused criminal is rejected to prove guilt. However, if due to insufficient evidence, the accused is not convicted, then the verdict does not mean that the innocence of the accused has been proven, since he may, in fact, be not innocent. We can only say that there is insufficient evidence to reject his innocence. The null hypothesis could very well be false, but the set of data might not be sufficient to reject the null hypothesis. Even if the null hypothesis cannot be rejected, we dare not accept the hypothesis as being true.

Example 2. *Suppose while walking home, a mysterious man wearing a trench coat whispered to you from out of a dark doorway as if not to be overheard by any one else but you. Suppose he advised you to sell all of your stock in Microsoft because next week it will declare bankruptcy. Then suddenly this mysterious man darted down a dark alley before you had a chance to respond. Of course, you would disbelieve him, because the government will never allow Microsoft to go out of business since too much of the tax base depends on it. Still you might be somewhat concerned. Should you sell or not sell your stock?*

You must make a decision. It can be framed as a statistical hypothesis: Microsoft will go bust next week, hence sell all stock. Shall you reject or not reject the hypothesis against the alternative hypothesis that Microsoft will continue to thrive. Which decision will be the worst one to make? If you reject the null hypothesis and keep your investment in Microsoft but next week you learned to your dismay that Microsoft indeed declared bankruptcy, then your investment is now worthless. You committed an error. Let us assume that you did sell your stock in Microsoft in accordance with the advice of that man whose character reminded you of Humphrey Bogart but next week you learned that Microsoft is as financially sound as ever. You have also committed an error. Which error is the worst error? The first one is the worst error because you lost your entire fortune whereas in the latter case you can still reinvest the capital given to you from the sale of the stock to buy it back again though at the cost of paying the commission. The first kind of error is the worst of the two. It is the Type-I Error; the second error is the Type-II Error.

Jerzy Neyman
1894-1981

The terminology which appears in the subject of testing hypotheses was coined by Jerzy Neyman who with Egon Pearson laid the foundations for the testing of hypotheses.

The probability that the decision to reject the null hypothesis is wrong is denoted by α; it is the same α that is used in calculating confidence intervals. The probability that the decision not to reject the null hypothesis is wrong is denoted by β. Not wanting to make the wrong decision dictates the desire for sufficient information. The more information that is made available for analysis the less likely that a wrong decision will be made. The discipline in statistics which is dedicated to this topic is called decision theory.

Suppose that a statistician presented a rule whereby if 40 or less heads are observed out of 100 tosses, then the decision to reject the null hypothesis that p=.5 against the alternative that p=.35 may be made. This rule is called a decision rule; the value, 40, is called the critical value. Decision rules are routinely employed in quality control in such matters as deciding whether to reject a lot of manufactured items or to approve it for shipment to a customer. The decision rule in that context is applied to a random sample. If more than 40 items are defective or non-conforming in the parlance of a legalist, then the lot is rejected and perhaps the assembly line will have to be shutdown until the cause of producing the defects is eliminated.

α denotes the probability of committing a type-I error;

$$\begin{aligned}\alpha &= P(\textit{decision to reject } H_0 \textit{ is wrong}) \\ &= P(T < 40|p = .5) = P(T \leq 39|p = .5) \\ &= \sum_{k=0}^{39} P(X = k|p = .5) = \sum_{k=0}^{39} \binom{n}{k} .5^k(1 - .5)^{n-k} \\ &= .0176001.\end{aligned}$$

By following the decision rule, the hypothesis is rejected or not rejected depending on whether we observe less than 40 heads or more than 40 heads. If the decision of rejecting the null hypothesis is made, then that decision will be wrong with a probability of .0176.

On the other hand, the alternative hypothesis is: $H_1 : p = .35$. If by following the decision rule, H_0 is not rejected but in reality that decision is the wrong one to have been made, then the decision maker committed a type-II error because in fact p=.35. The probability of this event occurring is denoted by β. $\beta = P(T \geq 40|p = .35) = 1 - P(T \leq 39|p = .35) = 1 - .827585 = .172415$. In the event that from following the decision rule, the null hypothesis is not rejected because more than 40 heads are observed, then the decision will be the wrong decision with probability β=.172.

From the preceding discussion, it should be clear that a decision rule depends on three things: the formulation of the hypothesis which includes the chosen parameter of a probability distribution like p for a Binomial distribution; the sampling size, n; and the critical value, C. It is informative to see how α and β change as either one of those three entities is varied. Keeping p and n fixed, Table 1 shows how α and β change as the critical value, C, varies.

The decision is based on the criterion that if the number of observed heads is less than C, then reject $H_0 : p = .5$ in favor of $H_1 : p = .35$, but if the number of observed heads is greater than C, then do not reject H_0. C is the critical value for making the decision. As the critical value changes, the probabilities of making the wrong decisions change. The probabilities α and β are in an inverse relationship with one another as a function of the critical value, C, and that leads to a problem of choosing an appropriate critical value. Choosing the right C is a decision of the manager who after weighing the the risks and benefits of a decision will stipulate values for α and β while recognizing that committing a type-I error is worse than committing a type-II error.

Given the same problem, but maintaining a constant ratio, $\frac{C}{n} = .4$, of the critical value, C, to the sampling size, n, the following table shows how the values of α and β change as

Table 1: Relationship of Type-I, Type-II, and Power of a Test

Critical Value C	$P(T < C \mid p = .5)$ α	$P(T \geq C \mid p = .35)$ β	Power of Test $1 - \beta$
35	.000895	.537563	.462431
36	.001758	.454164	.545836
37	.003318	.373075	.626946
38	.006016	.297551	.702449
39	.010400	.230192	.769870
40	.017600	.172415	.827585
41	.028144	.125022	.874977
42	.044313	.087678	.912321
43	.066605	.059430	.940569
$\vdots$	$\vdots$	$\vdots$	$\vdots$
48	.308649	.005019	.994998
49	.382176	.002748	.997252
50	.460205	.001450	.998549

the sampling size increases.

Table 2: Affect of Sampling Size on α, β, and Power of a Test

$\frac{C}{n}$	C	n	α	β	$1 - \beta$
.4	20	50	.059460	.273563	.726436
.4	40	100	.017600	.172415	.827585
.4	80	200	.001817	.080469	.919530
.4	400	1000	9.08×10^{-11}	.0005713	.9994287

The relationship of α and β according to the size of the sample agrees with our intuition that as more observations are made, the less likely a decision will be wrong. On the other hand, a large sample costs more money than small samples, consequently, the formulation of a decision rule must balance the sampling size and the critical value in such a way as to make α and β come as close to the stipulated values as promulgated by the management but with the least cost. The methodology which underpins this endeavor constitutes the design of experiments, an important branch of statistics upon which a practitioner of statistics constantly depends.

Upon completion of a well designed and successfully conducted experiment, the moment of anxious anticipation for substantiating the objective of the experiment arrives at the commencement of analyzing the data. From the analysis of the data, a parameter of the prospective probability distribution which is being proposed to describe a facet of the population is estimated and, based on that estimation, the decision whether or not to reject the claim that the estimate is the right one must be made. To appease the skepticism of an antagonist, the results of the analysis must be convincing; the claim must be tested against the facts. The analytical process of substantiating a claim is known as a test of hypothesis. There are many formulations of a hypothesis; we will be concerned with three of them.

Definition 1. *Let Θ be a parameter of a probability distribution.*

One-sided or One-tail Test	Two-sided or Two-tailed Test
$H_0 : \Theta = \Theta_0$ vs $H_1 : \Theta > \Theta_0$ $H_0 : \Theta = \Theta_0$ vs $H_1 : \Theta < \Theta_0$	$H_0 : \Theta = \Theta_0$ vs $H_1 : \Theta \neq \Theta_0$

1 Testing a Hypothesis between a Mean and a Constant

When X_i are i.i.d. $N(\mu, \sigma^2)$, the following table gives the test statistic and the criteria for rejecting the null hypothesis.

	H_0	Test Statistic	H_1	Reject When
When σ^2 is known	$\mu = \mu_0$	$Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	$\mu < \mu_0$	$Z < -z_\alpha$
			$\mu > \mu_0$	$Z > z_\alpha$
			$\mu \neq \mu_0$	$Z < -z_{\frac{\alpha}{2}}$ or $Z > z_{\frac{\alpha}{2}}$
When σ^2 is unknown	$\mu = \mu_0$	$T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}$	$\mu < \mu_0$	$T < -t_{n-1,\alpha}$
			$\mu > \mu_0$	$T > t_{n-1,\alpha}$
			$\mu \neq \mu_0$	$T < -t_{n-1,\frac{\alpha}{2}}$ or $T > t_{n-1,\frac{\alpha}{2}}$

Like the recipe for constructing a confidence interval, the recipe for testing a hypothesis is short and simple.

Step 1 Find α.

Step 2 Find the quantile that is appropriate for either a one or a two-sided test.

Step 3 Compute the test statistic.

Step 4 Does the test statistic satisfy the criterion for rejection?

Step 5 Make a decision.

Example 3. *Let $X_1\ X_2\ \ldots, X_{20}$ be i.i.d. Normal with $\sigma^2 = 1.0$, $\bar{x} = 14.5$. Test the hypothesis: $H_0 : \mu = 15$ vs $H_1 : \mu \neq 15$ at the level of significance, $\alpha = .01$.*

1. $\alpha = .01$ *(Two-sided test)* $\rightarrow \frac{\alpha}{2} = .005$
2. $z_{\frac{\alpha}{2}} = z_{.005} = 2.58$
3. $Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{14.5-15}{\frac{\sqrt{1}}{\sqrt{20}}} = -2.236$
4. *Is* $Z = -2.236 < -2.58$ *or* $Z = -2.236 > 2.58$*? No*
5. *Cannot reject null hypothesis at a level of significance of .01.*

Same problem but test at level of significance, $\alpha = .05$.

Let $X_1\ X_2\ \ldots, X_{20}$ be i.i.d. normal with $\sigma^2 = 1.0$, $\mu = 14.5$. Test the hypothesis: $H_0 : \mu = 15$ vs $H_1 : \mu \neq 15$ at the level of significance of .05.

1. $\alpha = .05$ *(Two-sided test)* $\rightarrow \frac{\alpha}{2} = .025$
2. $z_{\frac{\alpha}{2}} = z_{.025} = 1.96$
3. $Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{14.5-15}{\frac{\sqrt{1}}{\sqrt{20}}} = -2.236$
4. *Is* $Z = -2.236 < -1.96$ *or* $Z = -2.236 > 1.96$*? Yes*
5. *Reject null hypothesis at a level of significance of .05.*

Example 4. *Let $X_1\ X_2\ \ldots, X_{15}$ be i.i.d. $N(\mu, \sigma^2)$ where σ^2 is unknown but $\bar{x} = 40$ and $s^2 = 120$. Test the hypothesis: $H_0 : \mu = 45$ vs $H_1 : \mu < 45$ at the level of significance of $\alpha = .025$.*

1. $\alpha = .025$ *(One-sided test)*
2. $t_{n-1,\alpha} = t_{14,.025} = 2.145$
3. $T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{40-45}{\frac{\sqrt{120}}{\sqrt{15}}} = -1.7678$
4. *Is* $T = -1.7678 < -2.145$*? No*
5. *Cannot reject null hypothesis at a level of significance of .025.*

Same problem but test at a level of significance of .10.

Let $X_1\ X_2\ \ldots, X_{15}$ *be i.i.d.* $N(\mu, \sigma^2)$ *where* σ^2 *is unknown but* $\bar{x} = 40$ *and* $s^2 = 120$. *Test the hypothesis:* $H_0 : \mu = 45$ *vs* $H_1 : \mu < 45$ *at the level of significance,* $\alpha = .10$.

1. $\alpha = .10$ *(One-sided test)*

2. $t_{n-1,\alpha} = t_{14,.10} = 1.345$

3. $T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{40-45}{\frac{\sqrt{120}}{\sqrt{15}}} = -1.7678$

4. *Is* $T = -1.7678 < -1.345$*? Yes*

5. *Reject null hypothesis at a level of significance of .10.*

The preceding two examples illustrate the influence of α in determining the outcome of conducting a test. A smaller α betrays the manager's apprehension of committing an error in making a decision. Perhaps the risks are too high, as might be the case of a pharmaceutical company board of directors deciding to market a new drug or to forego an opportunity to make a profit. If the decision to market a new drug is wrong, then the company might have to bear the consequences of many expensive lawsuits for compensating the harmful effects of the drug on its users or if the decision not to market the new drug is wrong, then the company will lose an opportunity to make a profit. In light of the risk and benefits which will be incurred by the decision, the managers stipulate a level of tolerance for making a mistake. A small tolerance which a small α implies means that the null hypothesis that the drug is harmful can only be rejected based on overwhelming evidence. Not rejecting the null hypothesis does not mean that the drug is harmful; it means that there is insufficient evidence to conclude that the drug is beneficial. There is no theoretical method to ascertain the best α and β. Ultimately, an arbitrary decision must be made in practice by someone who is in a position of responsibility and who believes that the stipulated α and β properly address the concerns over risk and benefits which are incurred by making a decision. In response to the temptations of making big profits at the expense of compromising the health of consumers, Theodore Roosevelt signed into law in 1906 the Food and Drug Act which was at first administered by the United States Department of Agriculture. Through various amendments and changes in organization, the Food and Drug Administration was created in 1953 to guarantee that the tests of hypotheses for approving a drug are stringent enough to protect the welfare of society. Indeed, an understanding of the theory of statistics governs huge appropriations of public funds and lies at the core of running an efficient enterprise.

1.1 p-value

In the preceding example, two different conclusions were made even when the sets of data are identical. What changed was the specification of α. When $\alpha = .025$, the null hypothesis was not rejected, whereas when $\alpha = .10$, the null hypothesis was rejected. The conclusion not only depends on the hypothesis and on the data but also on α.

Table 3: Effect of Changing α on Making a Decision

α	$t_{n-1,\alpha}$	Decision
.025	2.145	Cannot Reject
.03	2.046	Cannot Reject
.04	1.887	Cannot Reject
.04943599	1.7678	????
.05	1.7613	Reject
.06	1.656	Reject
.07	1.565	Reject
.10	1.345	Reject

In that example, the test statistics, T=-1.7678, for the one sided hypothesis: $H_0 : \mu = 45$ vs $H_1 : \mu < 45$ does not change with the level of significance. Regardless of α, the test statistic remains fixed because it is derived from the data. It is then used to determine if the null hypothesis should be rejected according to the criterion: reject when $T < -t_{n-1,\alpha}$. It is the criterion which changes according to α. It is α which determines the decision whether or not to reject the null hypothesis. To illustrate this point, Table 3 shows the dependency of making a decision on α. As α changes, there comes a time when the decision reverses. At that point, α assumes a singular importance because it marks the boundary between accept and reject the null hypothesis. That special α occurs when $T = t_{n-1,\alpha}$ and it is marked in the table by ????. That α at which the quantile exactly equals the test statistic occurs exactly on the boundary of deciding to reject or not reject the null hypothesis. That particular value of α is called the p-value. An α which is larger than the p-value will cause the null hypothesis to be rejected, and when a value of α is less than the p-value the null hypothesis cannot be rejected.

Definition 2.

1. *That α such that $Z = z_\alpha$ is called the p-value for a one-sided test when σ^2 is known.*

2. *That α such that $Z = z_{\frac{\alpha}{2}}$ is called the p-value for a two-sided test when σ^2 is known.*

3. *That α such that $T = t_{n-1,\alpha}$ is called the p-value for a one-sided test when σ^2 is unknown.*

4. *That α such that $T = t_{n-1,\frac{\alpha}{2}}$ is called the p-value for a two-sided test when σ^2 is unknown.*

Because of animosity which erupted between Karl Pearson and Ronald Fisher who led the development of modern statistics, the p-value which Pearson favored was replaced by the use of testing hypotheses at regular values of α like .10, .05, and .025. Due to the feud which developed between them and Fisher's profound influence on the teaching of statistics, Fisher's use of regular values of α prevailed for decades. Fisher's custom of using regular values of α was a result of Fisher's lack of resources to reproduced the extensive tables of Pearson's which were protected by copyright. Today, with the ready access of computers, not only can tables which are protected by copyright be circumvented but the computation of p-values have become routine. In a certain sense, the computer has vindicated Pearson's opinion of using a p-value over the use of regular values of α.

A p-value offers a statistician an easy way to assess the importance of α. If a p-value is extremely small then in order to reject the null hypothesis, only a very slightly larger α is sufficient. In other words, a very small p-value like .0001 implies that an α of .00011 is sufficient to reject the null hypothesis. Such a small α indicates that it is very unlikely that a Type-I error will be committed in rejecting the null hypothesis. On the other hand, if the p-value is rather large like .30 then one can conclude that even as large an α as .29 is not sufficient to cause the null hypothesis to be rejected with much satisfaction since the odds of committing a Type-I error is about as bad as flipping a coin. The availability of a p-value provides a way for a statistician to make a quick decision about the null hypothesis.

Computing a p-value by hand requires access to a large volume of tables, and, on that account, it is not practical unless a computer is available. With the aid of a computer, p-values can be easily obtained, provided, of course, that a functional computer program exists.

Theorem 1. *Given that Z and T are the test statistics for the hypothesis $H_0 : \Theta = \Theta_0$ regarding the mean of a population, the p-values are given by:*

1	$p - value = P(z > Z)$	for one-sided test when σ^2 is known
2	$p - value = 2P(z > \|Z\|)$	for two-sided test when σ^2 is known
3	$p - value = P(t_{n-1} > T)$	for one-sided test when σ^2 is unknown
4	$p - value = 2P(t_{n-1} > \|T\|)$	for two-sided test when σ^2 is unknown

Proof. Proving two items will be sufficient to illustrate the proof of the others. Consider the case of a one-sided case when the variance is known. By definition of p-value, p is that α such that $Z = z_p$. By Theorem 2[1], $P(z \geq z_p) = p$, but $Z = z_p$ so that by replacing z_p

[1]

Theorem 2. $P(z \geq z_\alpha) = \alpha$.

Proof. $P(z \geq z_\alpha) = 1 - P(z \leq z_\alpha) = 1 - (1 - \alpha) = \alpha$. ■

with Z, we arrive at the formula, $P(z \geq Z) = p$.

For formula 4, the proof is similar. By definition of p-value, p is that α such that the T test statistic exactly equals the t quantile, that is, $|T| = t_{n-1,\frac{p}{2}}$. Since, $P(t_{n-1} \geq t_{n-1,\frac{p}{2}}) = \frac{p}{2}$ and by replacing $t_{n-1,\frac{p}{2}}$ with T, we arrive at the formula, $P(t_{n-1} > T) = \frac{p}{2}$ and the fourth formula is proved. ■

1.2 Difference between One-sided and Two-sided Tests

In the next example, a comparison is made between a one-sided test and a two-sided test at the same level of significance and using the same data.

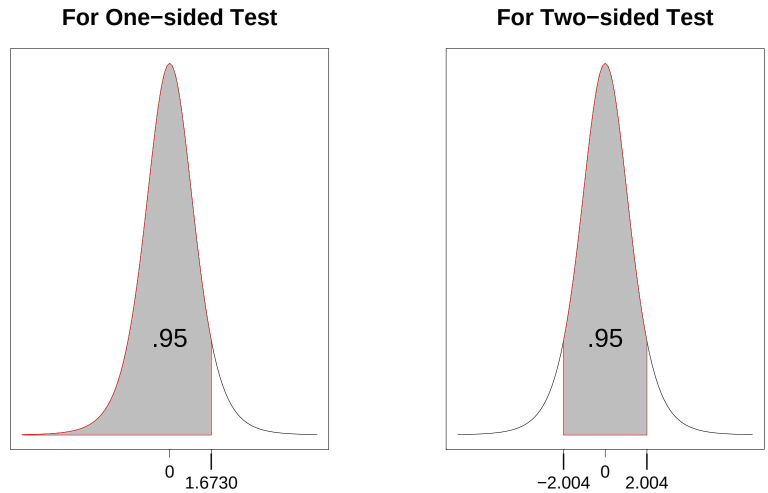
Example 5. *Let X_1 X_2 $\ldots, X_{56}$ be i.i.d. $N(\mu, \sigma^2)$ where σ^2 is unknown but $\bar{x} = 297.7$ and $s^2 = 4517.09$. Test the hypothesis: $H_0 : \mu = 315$ vs $H_1 : \mu \neq 315$ at the level of significance of .05.*

1. $\alpha = .05$ *(Two-sided test)* $\rightarrow \frac{\alpha}{2} = .025$
2. $t_{n-1,\frac{\alpha}{2}} = t_{55,.025} = 2.004045$
3. $T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{297.7-315}{\frac{\sqrt{4517.09}}{\sqrt{56}}} = -1.926242$
4. *Is* $T = -1.926242 < -2.004045$ *or* $T = -1.926242 > 2.004045$*? No*
5. *Cannot reject null hypothesis at a level of significance of .05.*

Let X_1 X_2 $\ldots, X_{56}$ be i.i.d. $N(\mu, \sigma^2)$ where σ^2 is unknown, but $\bar{x} = 297.7$ and $s^2 = 4517.09$. Test the hypothesis: $H_0 : \mu = 315$ vs $H_1 : \mu < 315$ at the level of significance of .05.

1. $\alpha = .05$ *(One-sided test)*
2. $t_{n-1,\alpha} = t_{55,.05} = 1.673034$
3. $T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{297.7-315}{\frac{\sqrt{4517.09}}{\sqrt{56}}} = -1.926242$
4. *Is* $T = -1.926242 < -1.673034$*? Yes*
5. *Reject null hypothesis at a level of significance of .05.*

It appears that the one-sided test is more stringent than the two-sided test because it leads to a definite rejection of the null hypothesis. Upon examining the alternative hypothesis, the one-sided hypothesis indicates that additional information is known, in that μ is presumed to be negative while on the other hand that knowledge is absent in the two-sided test. A picture of the probability distribution under each circumstance shows that the quantile in the two-sided test must be different than the quantile for the one-sided test, in order to keep the areas under the curves the same at .95. A larger quantile must be used in the two-side test in order to compensate for the lack of information about the sign of μ.



2 Equivalence of Testing Hypotheses and Confidence Intervals

The picture pertaining to the one-sided test on the left suggests an equivalent interpretation of testing a hypothesis. It would seem that one could assert that a one-sided test is the same as determining if $T = \frac{\mu_0 - \bar{x}}{\frac{s}{\sqrt{n}}} \in (-\infty,\ 1.6730)$ or with some algebraic rearranging, a test of hypothesis appears to be equivalent to ascertaining if $\mu_0 \in (-\infty, \bar{x} + \frac{s}{\sqrt{n}} t_{n-1,\alpha})$, that is, if $315 \in (-\infty,\ 312.7259)$ and, because 315 is not in that interval, the hypothesis that $\mu = 315$ is false at $\alpha = .05$. The picture on the right suggests, in a similar line of reasoning that in a two-sided test the null hypothesis is rejected if $\frac{\mu_0 - \bar{x}}{\frac{s}{\sqrt{n}}} \notin (-t_{n-1,\frac{\alpha}{2}}, t_{n-1,\frac{\alpha}{2}})$ or equivalently the null hypothesis is rejected when $\mu_0 \notin (\bar{x} - \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}})$. This last interval is nothing other than a confidence interval. Testing a hypothesis and looking at a confidence interval are equivalent approaches to answering the same question about whether or not an estimate is good.

Recall that $P(\mu \in (\bar{x} - \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}},\ \bar{x} + \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}})) = 1 - \alpha$. We do not know whether μ might be in the interval or outside it. A confidence interval gives an indication of where μ might be. It is asserted in the null hypothesis that $\mu = \mu_0$. Based on experimental data, an estimate of μ is given by $\bar{x}$ and, at the same time, σ^2 is estimated by s^2. The null hypothesis is rejected in a two-sided test when $T < -t_{n-1,\frac{\alpha}{2}}$ or $T > t_{n-1,\frac{\alpha}{2}}$. Given the test statistic, $T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ and applying a few simple algebraic manipulations, the criterion of rejecting H_0 is the same as $\bar{x} \leq \mu_0 - \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}$ or $\bar{x} \geq \mu_0 + \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}$. These inequalities are equivalent to the statement of rejecting H_0 when $\mu_0 \notin (\bar{x} - \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n}} t_{n-1,\frac{\alpha}{2}})$ at the $100(1-\alpha)\%$ level of significance.

The advantage of testing a hypothesis in terms of confidence intervals is that the use of confidence intervals provides an excuse for using pictures in one's presentation of the results. A judicious use of pictures which confidence intervals offers greatly improves the success of making a persuasive argument.

3 Paired Difference Test

A special kind of test, called the paired difference test, is useful when there is a need to determine if a treatment produces a noticeable difference on a subject. Commonly, the test is used to assess the difference in the effects of a treatment before and after it is applied to the same subject. There must be an exact pairing in order to make any sense out of a paired difference test.

Example 6. *An instructor at a college is curious to learn whether his students benefit from attending his lectures. He conducts an experiment in which he administers an examination to some students before a lecture and an examination to the same students after a lecture in such a way that the memory of the first examination does not affect a student's performance on the second examination. Over the years, the instructor has assiduously observed his students and having gained an insight into their attitudes, he applied his personal opinions in designing what would have to be well designed examinations. The results of the experiment are given below.*

Student's Name	Before Lecture	After Lecture	Improvement (Difference, d_i)
Abe	65	80	15
John	80	100	20
Tina	70	90	20
Jack	50	95	45
John	80	80	0
Total			100

The instructor, subsequently, tests the worst case, namely the hypothesis that attending lecture does not improve a quiz grade at a level of significance of .05, i.e. $H_0 : \mu_d = 0$ *vs* $H_1 : \mu_d > 0$ *at* $\alpha = .05$.

Of course, σ^2 is unknown; therefore, the instructor must resort to a T test statistic. From the data, the instructor calculates $\bar{d} = 20$ and the sample standard deviation of the differences to be $s = 16.2019$.

1. $\alpha = .05$ *(One-sided test)*

2. $t_{n-1,\alpha} = t_{4,.05} = 2.132$

3. $T = \frac{\bar{d}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{20-0}{\frac{16.2019}{\sqrt{5}}} = 2.7603$

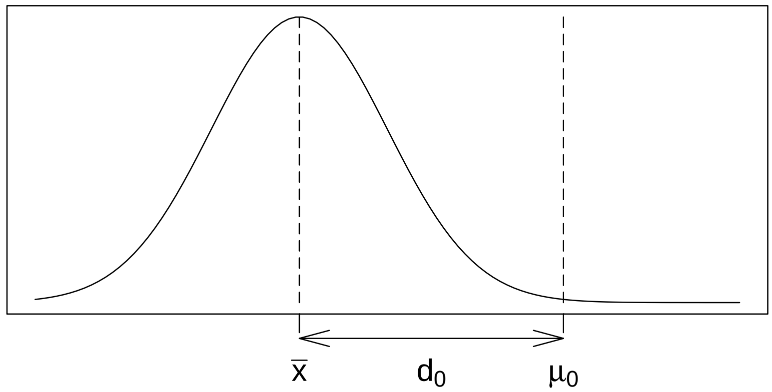
4. *Is* $T = 2.7603 > 2.132$*? Yes*

5. *Reject null hypothesis at a level of significance of .05 that attending lecture makes no difference in examination scores. In conclusion, the instructor's self-esteem is reassured to his immense satisfaction, and the students learn to appreciate, once again, the practical benefits of attending lecture.*

The decision to do a paired difference test depends on the design of the experiment. The key idea which must be kept in mind and which underscores the experiment is that a subject undergoes an examination twice; once before the treatment is applied and again afterwards. In the next section, the test of hypothesis between two means might easily be confused with the paired difference test. The design of the experiment dictates which test to use. If an element is drawn from a population and is examined twice, once before and again after the application of a treatment, then the paired difference test is appropriate. If two samples are drawn from different populations and the experiment is so designed as to examine the difference between the means of the two populations, then the test in the next section is appropriate.

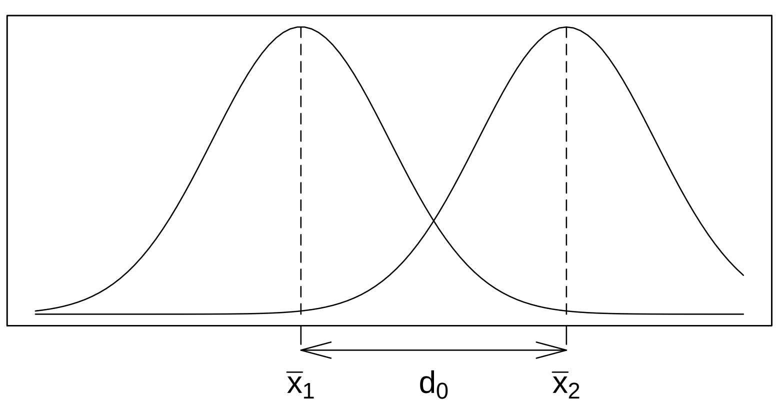
4 Testing a Hypothesis between Two Means

Thus far, the testing of a hypothesis involved a mean and a constant. The null hypothesis states that the distance, d_0, between the mean, μ, and a constant, μ_0, is zero; that is, $H_0 : \mu - \mu_0 = d_0 = 0$. A picture of this situation appears below where $\bar{x}$ is, as before, the estimate of μ.



If d_0 is very small, then one would surmise that for practical purposes $\mu = \mu_0$. If, on the other hand, the distance between $\bar{x}$ and μ_0 is very large, then it would seem very unlikely that μ and μ_0 would be the same. The null hypothesis is rejected, if d_0 is too big. It cannot be rejected when d_0 is small. The question of how big is big and how small is small is answered by referring to a measuring stick which we know as a quantile. The quantile is the statistical measuring stick by which d_0 is determined to be big or small. In order to use a quantile for that purpose, the distance between $\bar{x}$ and μ_0 must be transformed into a test statistic. If the test statistic is too big relative to the quantile, then the null hypothesis is rejected.

The origins of μ_0 might be obvious or theoretical. In any case, it is a parameter of a probability distribution which is being advanced to describe some characteristic of a population. It is certainly possible that μ_0 might not be known theoretically so that its value must be estimated from another experiment. In that case, it is no longer a constant but a random variable and must be associated with a probability distribution. One might wonder if this other distribution which now supersedes the constant, μ_0, is identical to the distribution of $\bar{x}$. In other words, could the distance between μ_1 which represents the mean of one distribution and μ_2 which represents the mean of the other distribution be so small, that, for practical purposes, the means are the same and thereby suggest that the distributions are actually identical.



The determination of whether d_0 is big or small is accomplished by comparing an appropriate test statistic with a quantile. If d_0 is very small, then for practical purposes, μ_1 and μ_2 can be deemed to be the same; otherwise, if d_0 is very large, then the assertion that the means are the same must be rejected. To begin that determination, the difference between $\bar{x}_1$ and $\bar{x}_2$ must be transformed into a test statistic which involves the process of weaving the information of two distributions together into one, in order to derive the right probability distribution for getting a quantile. Once the test statistic is computed, then it can be compared to a suitable quantile of that probability distribution. The criteria for rejecting the null hypothesis involving two Normal distributions is given below.

	H_0	Test Statistic	H_1	Reject When
When σ_1^2 and σ_2^2 are known	$\mu_2 - \mu_1 = d_0$	$Z = \frac{\bar{x}_2 - \bar{x}_1 - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\mu_2 - \mu_1 < d_0$ $\mu_2 - \mu_1 > d_0$ $\mu_2 - \mu_1 \neq d_0$	$Z < -z_\alpha$ $Z > z_\alpha$ $Z < -z_{\frac{\alpha}{2}}$ or $Z > z_{\frac{\alpha}{2}}$
When σ_1^2 and σ_2^2 are unknown but $\sigma_1^2 = \sigma_2^2$	$\mu_2 - \mu_1 = d_0$	$T = \frac{\bar{x}_2 - \bar{x}_1 - d_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\mu_2 - \mu_1 < d_0$ $\mu_2 - \mu_1 > d_0$ $\mu_2 - \mu_1 \neq d_0$	$T < -t_{\nu;\alpha}$ $T > t_{\nu;\alpha}$ $T < -t_{\nu;\frac{\alpha}{2}}$ or $T > t_{\nu;\frac{\alpha}{2}}$

Combining the essentials of two Normal distributions leads to the calculation of $S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ which is called the pooled variance and $\nu = n_1 + n_2 - 2$, the degrees of freedom.

Example 7. *An editor of an employment newspaper is curious if the wages paid by Wal-Mart and by K-Mart are different. Ten workers chosen at random from a local Wal-Mart store and 12 from a K-Mart store were interviewed. Based on the interviews the following information was obtained.*

	Store	$\bar{x}$	s
#1	Wal-Mart	\$23,600	\$3,200
#2	K-Mart	\$24,800	\$3,700

The question posed by the editor is: Does K-Mart pay better than does Wal-Mart at a level of significance of $\alpha = .05$*?*

Let μ_1 *be the salary from Wal-Mart and let* μ_2 *be the salary from K-Mart. Assume that* $\sigma_1 = \sigma_2$.

The question is equivalent to a test of hypothesis, namely: $H_0 : \mu_1 - \mu_2 = 0$ *vs* $H_1 : \mu_1 - \mu_2 < 0$ *at the level of significance of .05.*

1. $\alpha = .05$ *(One-sided test) where* ν*=10+12-2=20*

2. $t_{\nu,\alpha} = t_{20,.05} = 1.725$

3. $s_p^2 = \frac{(10-1)3200^2+(12-1)3700^2}{20} = 12137500 \rightarrow s_p = 3484$

4. $T = \frac{\bar{x}_1-\bar{x}_2-d_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{23600-24800-0}{3484\sqrt{\frac{1}{12}+\frac{1}{10}}} = -.8044$

5. *Is* $T = -.8044 < -1.725 = t_{20,.05}$*? No*

6. *Cannot reject null hypothesis at a level of significance of .05. There does not appear to be a difference in salaries paid by the stores.*

Example 8. *A random sample of size* $n_1 = 25$ *taken from a normal population with population standard deviation* $\sigma_1 = 5.2$ *has sample mean* $\bar{x}_1 = 81$*. A second random sample of size* $n_2 = 36$ *taken from a different normal population with a population standard deviation* $\sigma_2 = 3.4$ *has a mean* $\bar{x}_2 = 76$*. Test the hypothesis at the 0.06 level of significance that* $\mu_1 = \mu_2$ *against the alternative* $\mu_1 \neq \mu_2$*.*

$H_0 : \mu_1 = \mu_2$ *vs* $H_1 : \mu_1 \neq \mu_2$

or $H_0 : \mu_1 - \mu_2 = 0$ *vs* $H_1 : \mu_1 - \mu_2 \neq 0$ *where* $d_0 = 0$

What do we know?

$n_1 = 25 \quad n_2 = 36$

$\bar{x}_1 = 81 \quad \bar{x}_2 = 76$

$\sigma_1 = 5.2 \quad \sigma_2 = 3.4$

The alternative hypothesis indicates that the test is a two-sided test. We surmise that the variances are known since from the terminology, a population standard deviation implies that the variance is known and will be provided; therefore the Z test statistic is calculated and compared to a z-quantile. It should be noted that the order of $\bar{x}_1$ *and* $\bar{x}_2$ *is important and it must follow the order of* μ_1 *and* μ_2 *in the null hypothesis.*

1. $\alpha = .06$ *hence* $\frac{\alpha}{2} = .03$

2. $z_{.03} = 1.88$

3. $Z = \frac{\bar{x}_1-\bar{x}_2-d_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} = \frac{81-76-0}{\sqrt{\frac{5.2^2}{25}+\frac{3.4^2}{36}}} = \frac{5}{1.18436} = 4.22$

4. *Is* $4.22 < -1.88$ *or* $4.22 > 1.88$*? Yes*

5. *Reject null hypothesis.*

For the sake of curiosity, the 94% confidence interval for $\mu_1 - \mu_2$ *is constructed.*

1. $\alpha = .06$ *so* $\frac{\alpha}{2} = .03$

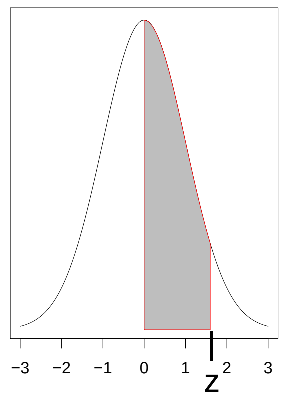
2. $z_{.03} = 1.88$

3. *Lower bound* $a = \bar{x}_1 - \bar{x}_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} z_{.03} = 5 - 1.18436(1.88) = 2.773403$

4. *Upper bound* $b = \bar{x}_1 - \bar{x}_2 + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} z_{.03} = 5 + 1.18436(1.88) = 7.226597$

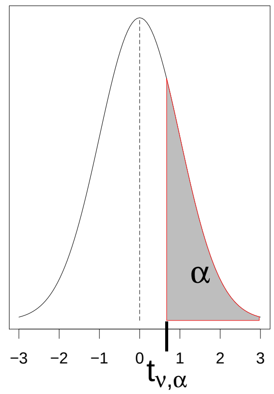
5. *94% CI for* $\mu_1 - \mu_2$ *= (2.77,7.23).*

6. *Question: Is* $0 \in (2.77, 7.22)$*? No. Therefore, we may reject the null hypothesis that* $\mu_1 = \mu_2$ *at a level of significance of .06.*



Cumulative Probabilities for a N(0,1) Distribution: $\Phi(z) - .5$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.0279	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47193	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49506	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49865	0.49869	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900



Quantiles for a Student's t Distribution

ν	$t_{\nu,.20}$	$t_{\nu,.15}$	$t_{\nu,.10}$	$t_{\nu,.05}$	$t_{\nu,.025}$	$t_{\nu,.01}$	$t_{\nu,005}$
1	1.37638	1.96261	3.07768	6.31375	12.7062	31.82052	63.65674
2	1.06066	1.38621	1.88562	2.91999	4.30265	6.964560	9.92484
3	0.97847	1.24978	1.63775	2.35338	3.18245	4.54070	5.84091
4	0.94096	1.18957	1.53321	2.13185	2.77645	3.74695	4.60410
5	0.91954	1.15577	1.47588	2.01505	2.57058	3.36493	4.03216
6	0.90570	1.13416	1.43976	1.94318	2.44691	3.14267	3.70743
7	0.89603	1.11916	1.41492	1.89458	2.36462	2.99795	3.49948
8	0.88889	1.10815	1.39682	1.85955	2.30600	2.89646	3.35539
9	0.88340	1.09972	1.38303	1.83311	2.26216	2.82144	3.24984
10	0.87906	1.09306	1.37218	1.81246	2.22814	2.76377	3.16927
11	0.87553	1.08767	1.36343	1.79588	2.20099	2.71808	3.10581
12	0.87261	1.08321	1.35622	1.78229	2.17881	2.68100	3.05454
13	0.87015	1.07947	1.35017	1.77093	2.16037	2.65031	3.01228
14	0.86805	1.07628	1.34503	1.76131	2.14479	2.62449	2.97684
15	0.86624	1.07353	1.34061	1.75305	2.13145	2.60248	2.94671
16	0.86467	1.07114	1.33676	1.74588	2.11991	2.58349	2.92078
17	0.86328	1.06903	1.33338	1.73961	2.10982	2.56693	2.89823
18	0.86205	1.06717	1.33039	1.73406	2.10092	2.55238	2.87844
19	0.86095	1.06551	1.32773	1.72913	2.09302	2.53948	2.86093
20	0.85996	1.06402	1.32534	1.72472	2.08596	2.52798	2.84534
21	0.85907	1.06267	1.32319	1.72074	2.07961	2.51765	2.83136
22	0.85827	1.06145	1.32124	1.71714	2.07387	2.50832	2.81876
23	0.85753	1.06034	1.31946	1.71387	2.06866	2.49987	2.80734
24	0.85686	1.05932	1.31784	1.71088	2.06390	2.49216	2.79694
25	0.85624	1.05838	1.31635	1.70814	2.05954	2.48511	2.78744
26	0.85567	1.05752	1.31497	1.70562	2.05553	2.47863	2.77871
27	0.85514	1.05673	1.31370	1.70329	2.05183	2.47266	2.77068
28	0.85465	1.05599	1.31253	1.70113	2.04841	2.46714	2.76326
29	0.85419	1.05530	1.31143	1.69913	2.04523	2.46202	2.75639
30	0.85377	1.05466	1.31042	1.69726	2.04227	2.45726	2.75000
40	0.85070	1.05005	1.30308	1.68385	2.02108	2.42326	2.70446
50	0.84887	1.04729	1.29871	1.67591	2.00856	2.40327	2.67779
75	0.84644	1.04365	1.29294	1.66543	1.99210	2.37710	2.64298
100	0.84523	1.04184	1.29007	1.66023	1.98397	2.36422	2.62589
150	0.84402	1.04003	1.28722	1.65508	1.97591	2.35146	2.60900
∞	0.84162	1.03643	1.28155	1.64485	1.95996	2.32635	2.57583