Inferences and Hypothesis Testing
Geography 450, Urban Research
Elvin Wyly

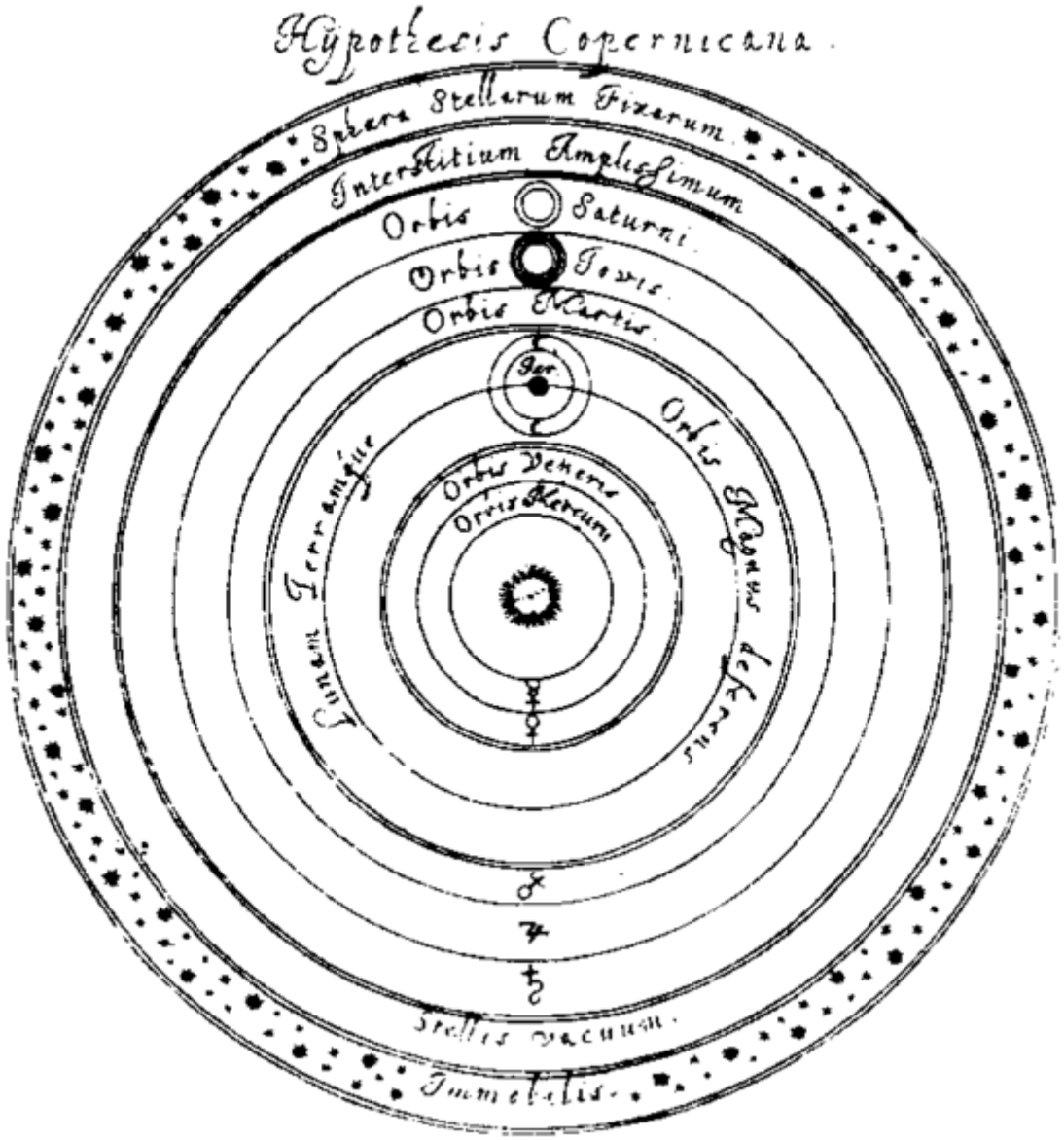


Figure 1. Hypothesis Copernicana. Ditaeva (2005). *Hypothesis Copernicana, Scanned from Dagfinn Døhl Dybvig & Magne Dybvig (2003). Det tenkende mennesket. Oslo: Tapir akademisk forlag*, p. 138. Released into the public domain with copyright expiry, via Wikimedia Commons.

> "...a fundamental use of hypothesis testing is to draw some inference about a larger body of unobserved data (the '*population*') from a *sample* of

> observations. ... If it were not possible to draw inferences about the population, any analysis would have very limited application and use."[1]
>
> "Statistics is the science of random processes, the standard alternative theory suggested by the phrase 'null hypothesis.' It has the basic form: 'there is nothing going on here but the generation of random motions in what the investigator thought was a causal space.' Because there is a great deal of random motion in social life, and because there is a great deal of random noise in social science techniques of observation, every social science finding has to show that it is not likely to be simply noise. Because that alternative theory is one of the few in social science that is well formulated mathematically, it is in general the hardest for ordinary social scientists to learn."[2]

In the most general terms, a 'hypothesis' is a testable proposition derived from theory, logic, or existing knowledge; if tested with proper, rigorous methods, the proposition will provide information that adds useful knowledge and contributes to theory. A hypothesis is not exactly the same as a hunch, an argument, or an assertion. The hunch that comes from intuition is a valuable element of creativity, but it must be clearly connected to existing knowledge -- and translated into a testable proposition -- before it can be considered a hypothesis. Assertions and arguments are ubiquitous, but they rarely involve propositions that can be clearly tested with legitimate methods or credible evidence.

'Hypothesis testing' refers to two distinct approaches in social research. One approach is usually understood in a broad, qualitative sense, while the other is usually described in narrow, quantitative terms.

Rhetorical Hypotheses

First, a hypothesis refers to a **rhetorical** approach: a technique of persuasion that involves an attempt to gain credibility by establishing common ground with a reader or listener. The word hypothesis comes from the Greek *hypothesis*, which in turn came from the Greek *hypotithenai*, "to place under," which came from *hypo* ('under') + *tithenai* (to put under). To advance a hypothesis is to put forward a postulate, an assumption, or a supposition. In many areas of scholarly research and policy discussion, participants quickly learn that they agree on many aspects of the issue at hand -- and indeed, there may be a consensus that is so broad that it approaches what John Kenneth Galbraith famously called the "conventional wisdom."

---

[1] A. Stewart Fotheringham (2008). "Analysing Numerical Spatial Data." In Robin Flowerdew and David Martin, eds., *Methods in Human Geography*. Harlow, England: Pearson Prentice Hall, 191-206, quote from p. 196.

[2] Arthur L. Stinchcombe (2005). *The Logic of Social Research*. Chicago: University of Chicago Press, p. 291.

But the points of disagreement are crucial, and narrow disputes can easily overshadow points of agreement. In order to promote a healthy, productive discussion, participants will often seek to formulate propositions that build, as far as possible, on points of agreement -- extending the implications to raise questions where there is disagreement, in ways that will help to adjudicate amongst alternatives. Proposing a hypothesis, then, is a way of identifying common ground in order to highlight an area of dispute in a productive way -- such that participants in a debate can either reach a consensus, or learn useful things from their disagreements. Consensus is rare, but productive disagreement can be quite healthy.

Both consensus and healthy disagreement, however, always require at least three preconditions. **First**, the rhetorical hypothesis must be presented in the logic and language of the critic or the opponent: to earn the trust of the audience you wish to engage, you may have to speak in their language. You need to appeal to things they find convincing. You need to find common ground. **Second**, the rhetorical hypothesis must strike a tone of fairness and impartiality. Again, this requires a careful consideration of the sensitivities of the audience. In some cases, the best move is an attempt to achieve rhetorical neutrality -- presenting a proposition in the most neutral terms possible, with a minimum of adjectives or any other words that signal particular interpretations. In other cases, the best approach is to present contrasting interpretations -- as clearly and as fairly as possible, with no caricatures or exaggerated implications. A rhetorical hypothesis is not an infallible guide to 'perfect' knowledge; but it *is* a legitimate attempt to build common ground in the search for better knowledge. Arguments and interpretations are usually least convincing when they are introduced too early -- before a foundation is built. The stronger the foundations of your hypothesis, the more your audience will feel compelled to follow the logic of your arguments and interpretations. **Third**, the hypothesis must be clearly testable. You do not have to specify all the detailed methods the first time you propose a hypothesis (indeed, it is common to have 'hypotheses' and 'methods' appear in different sections); but there should be sufficient information in the presentation of the hypothesis to make it clear that the proposition is testable, and that there is a reasonable body of evidence on which to base conclusions and inferences. An un-testable hypothesis is pretty much the same as an assertion or ideological statement; assertions and ideological manifestos certainly have their place, but for many purposes, for certain audiences, the diplomatic, fair, and testable hypothesis is absolutely crucial.

Statistical Hypotheses

The second meaning of hypothesis is more narrow, specific, and quantitative: *a testable proposition based on the characteristics of a sample, which is used to draw an inference about a broader population according to the mathematical principles of probability.* Peter Rogerson offers the example of a survey of commuting behaviors: we identify a simple random sample of fifty workers in a city, and we ask each of them how far they travel to their regular workplace.[3] The mean of their responses works out to 10.0 km, with a standard deviation of 9.0 km which we would typically denote in shorthand as

[3] Peter A. Rogerson (2006). *Statistical Methods for Geography*. Second Edition. Thousand Oaks, CA: Sage Publications, p. 93ff.

$\overline{X}$ =10.0, *s*=9.0. How much confidence can we place in this observed sample mean, $\overline{X}$, as an estimate of the true, unobserved population mean, μ? We know that if we were to choose a different sample -- fifty different workers, chosen randomly -- we would probably not obtain exactly the same estimate of the mean. But if we were to do this procedure repeatedly, drawing samples over and over again from the population, and calculating the mean $\overline{X}$ for each sample, probability theory gives us three valuable pieces of information.

1. The frequency distribution of the sample means $\overline{X}$ will resemble a normal distribution, *even if the underlying population does not conform to the perfect normal bell-curve*. Not all phenomena conform to the normal distribution, but the distribution of pure, random sampling error *does* conform to normality. This has been demonstrated repeatedly through experimental techniques (for instance, taking many repeated random samples from a known, easily-observed population, and analyzing the sample distributions).

2. With more samples, or with a larger number of observations in each sample, the sampling distribution becomes a more perfect normal distribution, with less variance in the scatter of sample means $\overline{X}$ above and below the true, unmeasured population mean, μ. In graphical terms, this means that the sampling distribution begins to look 'tighter,' with less 'spread' (Figure 1).

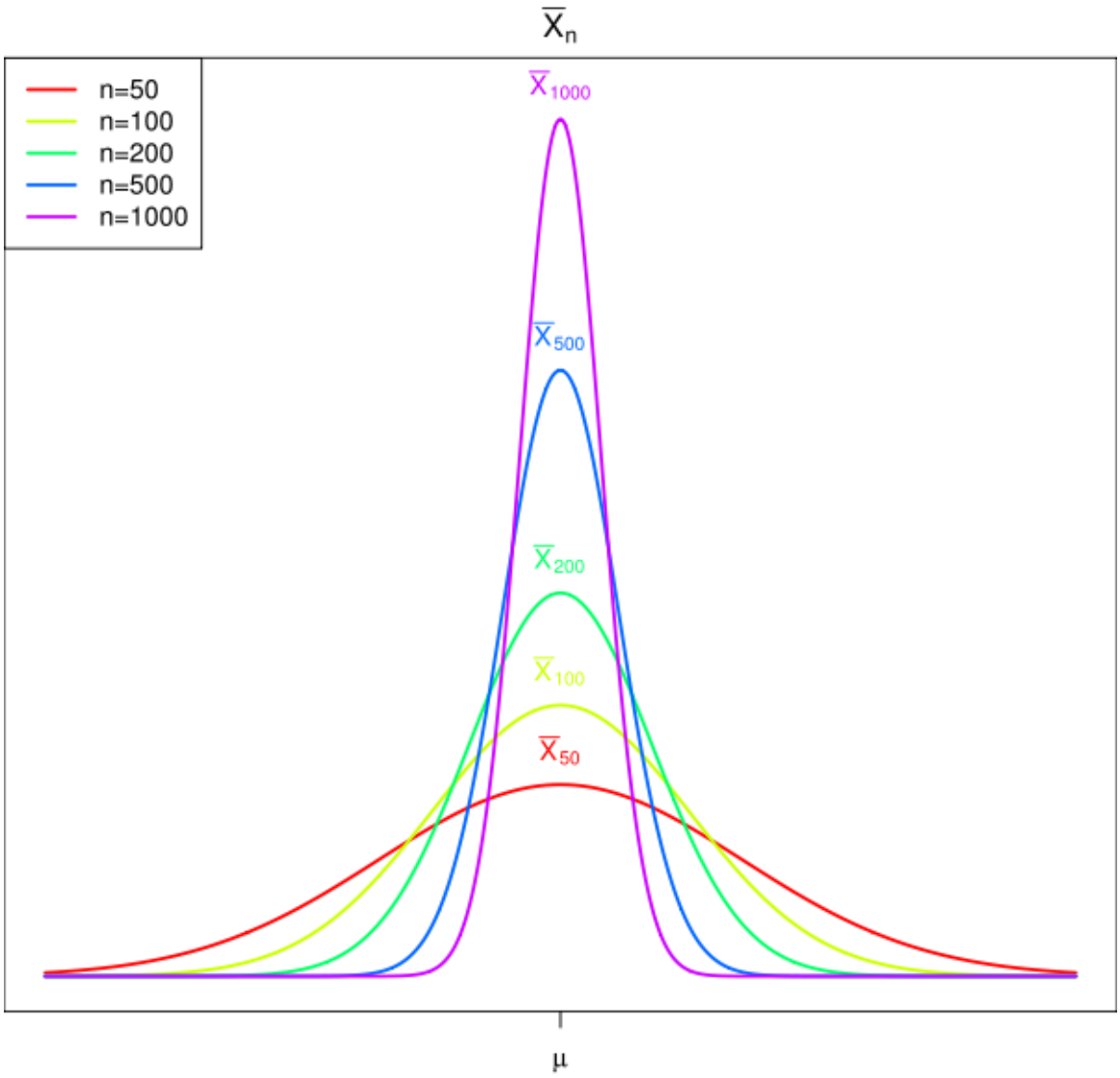


Figure 1. Comparison of Sampling Distributions with Different Sample Sizes. *Source:* Sigbert (2011). Distribution of Mean Estimator. Reproduced under Creative Commons Attribution-Share Alike 3.0 Unported license, via Wikimedia Commons.

More formally, the *variance* of the various sample estimates of the mean is

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

which means that if the population has a large variance σ^2, then we should expect a correspondingly high variability in the distribution of the sample means; conversely, with more samples or with more observations in each sample, the variability in these sample means decreases. If we take the square root of the variance of the sample estimates, we obtain *the standard error of the sample means*:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

When *n* approaches infinity, the variance of the sampling distribution approaches zero, and so does the standard error of the sample means. At that point the sample mean $\bar{X}$ becomes a perfect estimate of the population mean μ. This is intuitively logical: if we had the resources to do more than a simple survey of fifty workers -- if we could do a complete enumeration of all workers in a city with, say, one million workers -- then our 'sample' mean would be exactly the same as the 'population' mean.

3. When the sample size is sufficiently large, the standard deviation of the observations in a *sample* can be used to make statements about the likely value of the *population mean.* The sample means will cluster around the true population mean, following a perfect normal distribution. If everything is expressed in standardized units -- in terms of z-scores, standard deviations above and below the mean -- then this means we can use the **standard normal curve** to estimate the likelihood of observing various sample outcomes. In the standard normal curve, a normal distribution is transformed to z-scores, so that the mean is zero, the standard deviation is 1, and the total 'area under the curve' traced out by the histogram is also 1. In the standard normal curve, only a very small proportion of the values are below the curve in either of the 'tails.' Figure 2 shows this in the form of a histogram, while Figure 3 shows this as a table of numbers corresponding to the area under the curve at various z-scores to the right of the mean. Notice the row for z=1.9, and then look over several columns until the 'second decimal' is 0.06: the table indicates that 0.475 of the curve is included from the mean up to 1.96 standard deviations (z-scores) above the mean. The normal curve is symmetrical, and so we know that the area under the curve up to the mean is 0.500. So the total area under the curve all the way up to a z-score of +1.96 is 0.500+0.475=0.975. This is 0.025 shy of 1.000. This means that only 2.5 percent of the entire distribution is to be found under the curve in the upper tail -- above a z-score of 1.96; since the curve is perfectly symmetrical, another 2.5 percent of the entire distribution is to be found under the curve in the lower tail -- less than a z-score of -1.96. In turn, this means that the remainder of the entire distribution -- 95 percent -- is within the range of -1.96 to +1.96.

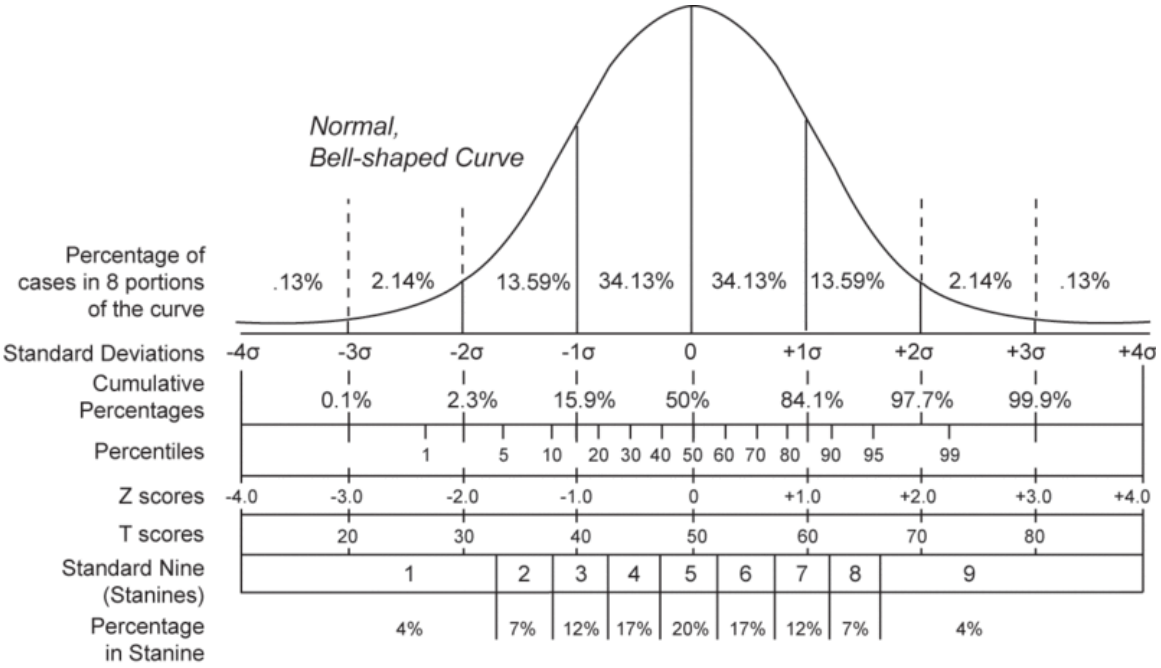


Figure 2. The Standard Normal Distribution. Source: Jeremy Kemp (2005). *Standard Normal Distribution with Scales, Adapted from Ward et al., Assessment in the Classroom.* Released into the public domain, via Wikimedia Commons.

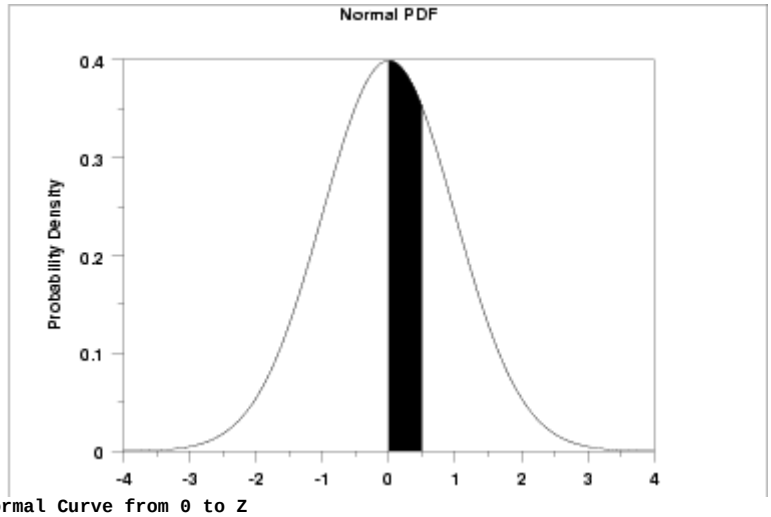
This means that if our sample is sufficiently large, we can use the characteristics of the sample to make statements about the likely location of the population mean. Specifically, if we have drawn a simple random sample to obtain a mean value $\overline{X}$, we can be sure that there is only a 2.5 percent chance that this sample is more than 1.96 standard deviations above the actual population mean μ.

Since the standard deviation of the sampling distribution was defined above as

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

this means that if we have a sufficiently large sample size, we can use the sample standard deviation (s) in place of the population standard deviation (σ). For our survey of workers' commute distances, then, this means that there is only a 2.5 percent chance that our observed sample mean $\overline{X}$ is more than 1.96 $\frac{s}{\sqrt{n}}$ above the true population mean; since our sample size is 50 and the sample standard deviation is 9, we know that there is only a 2.5 percent chance that $\overline{X}$ is more than $1.96\frac{9}{\sqrt{50}}$=2.49 units **above** the true mean. There is another 2.5 percent chance that $\overline{X}$ is more than 2.49 units **below** the true mean. The confidence interval for our sample mean of 10 km, then is 7.51 km to 12.49 km: we are 95 percent confident that if we were to draw an infinitely large number of other random samples of 50 workers from our city, and to calculate the average commute

distance of the workers in each sample, 95 percent of the means would lie within this range.



Area under the Normal Curve from 0 to Z

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	0.02790	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	0.41308	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
3.1	0.49903	0.49906	0.49910	0.49913	0.49916	0.49918	0.49921	0.49924	0.49926	0.49929
3.2	0.49931	0.49934	0.49936	0.49938	0.49940	0.49942	0.49944	0.49946	0.49948	0.49950
3.3	0.49952	0.49953	0.49955	0.49957	0.49958	0.49960	0.49961	0.49962	0.49964	0.49965
3.4	0.49966	0.49968	0.49969	0.49970	0.49971	0.49972	0.49973	0.49974	0.49975	0.49976
3.5	0.49977	0.49978	0.49978	0.49979	0.49980	0.49981	0.49981	0.49982	0.49983	0.49983
3.6	0.49984	0.49985	0.49985	0.49986	0.49986	0.49987	0.49987	0.49988	0.49988	0.49989
3.7	0.49989	0.49990	0.49990	0.49990	0.49991	0.49991	0.49992	0.49992	0.49992	0.49992
3.8	0.49993	0.49993	0.49993	0.49994	0.49994	0.49994	0.49994	0.49995	0.49995	0.49995
3.9	0.49995	0.49995	0.49996	0.49996	0.49996	0.49996	0.49996	0.49996	0.49997	0.49997
4.0	0.49997	0.49997	0.49997	0.49997	0.49997	0.49997	0.49998	0.49998	0.49998	0.49998

Figure 3. Cumulative Distribution Function of the Standard Normal Distribution. Source: National Institute of Standards and Technology (2010). *NIST/SEMATECH e-Handbook of Statistical Methods, Engineering Statistics Handbook*. Washington, DC: U.S. Department of Commerce, public domain.

One-Sample Z-Tests

This general approach is widely used to test whether a particular sample differs from some hypothesized value. Suppose we wish to compare our survey of 50 workers' commutes (mean 10.0 km, standard deviation of 9.0) to a known regional average of 13.1 km. Are the commutes of our city workers actually shorter than those of workers across the entire urban region? Or is it possible that our sample mean is just the product of random sampling variability?

If there is no systematic difference between the commutes of our 50 city workers and others across the region, then any differences in the means will result solely from random sampling variability. In that case, the difference between our observed mean and the true population mean would follow a standard normal distribution. The equations above can be easily rearranged to give a z-test statistic:

$$z = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

For our commuting example, the z-score is

$$z = \frac{10.0 - 13.1}{\frac{9}{\sqrt{50}}}$$

$$z = \frac{-3.1}{\frac{9}{\sqrt{7.071}}}$$

$$z = \frac{-3.1}{1.273}$$

$$z = -2.435$$

This means that if our sampled workers were no different from workers across the entire region, then our sample average is 2.43 standard deviations below population mean (the regional average). Consult the table for the standard normal curve, and you note that going 2.43 z-scores away from the mean takes you almost all the way to half of the entire area under the curve: 0.49245. This means that the rest of the way to 0.50000 is only 0.00755 (0.50000-0.49245). This applies symmetrically to negative z-scores, so this means that only 0.00755 of the area under the curve is below -2.43 z-scores. **We can**

thus conclude that there is less than a 1 percent chance that our sample of city commuters is no different from workers across the entire region.

In the language of hypothesis testing, our initial proposition -- that there is no systematic difference between the sampled workers and the broader regional population -- is referred to as the '*null hypothesis,*' or H_0. The z-score is our *test statistic*. And the thresholds that we look through on the table of the standard normal curve are often referred to as critical regions, or critical values: z-values that fall below -1.96 or above +1.96 are usually identified as "significant at P<0.05," because only five percent of all z-values will lie outside this range if indeed the null hypothesis is true. Z-values falling below -2.58 or above +2.58 are noted as "significant at P<0.01," because only one percent of all z-values will lie outside this range if the null hypothesis is true. In our example, we noted the probability 0.0075 for a z-score of -2.43; this is a one-tailed test of the hypothesis that the city workers' commutes are shorter than those of workers throughout the entire region. If we have no theoretical or logical reason to suspect that one group is higher or lower than the other, then we would use a more conservative, two-tailed test: the likelihood of a z-score below -2.43 or above +2.43 is (0.00755)*2, or 0.0151. This is significant at the 5 percent level, but not at the 1 percent level.

One-Sample t-Tests

At several points in the discussion above, we noted that "if the sample is sufficiently large," we could use the observed standard deviation of a sample (s) as an estimate of the population standard deviation (σ). In his classic work, *Social Statistics*, Blalock notes that "this was commonly done before the development of modern statistics. ... As it turns out, this procedure yields reasonably good results when n is large..." but "probabilities obtained in this manner can be quite misleading whenever n is relatively small."[4] This is intuitive if we recall Figure 1 above, which shows the effect of sample size on the shape of a sampling distribution: sampling distributions approach perfect normality only with large samples; smaller samples thus provide much less reliable information to allow us to infer the likely location of a true, unobserved mean. Sampling distributions quickly begin to resemble normal distributions even with fairly small n, but z-tests cannot be reliably used unless the sample size is at least 30. When sample sizes dip below 30, the normal distribution at the heart of sampling theory begins to show much greater variability -- the 'tails' become fatter, with larger proportions of deviations farther out from the true, unobserved mean. In the 1930s, a scientist by the name of W.C. Gossett wrote an anonymous article identifying the problems with the use of the sample standard deviation in small-sample tests, and demonstrated a distinctive sampling distribution that he called the "t" distribution. Gossett published his anonymous article under the name, "Student," and ever since, the test statistic has been known as the "Student's t distribution."

The t statistic is calculated almost exactly the same as a z-score:

[4]Blalock, *Social Statistics*, p. 145.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

but the reference points on the table for the area under the curve on the t distribution are different. The shape of the curve varies significantly with sample size, indicated by the "degrees of freedom," which is calculated as the sample size minus 1. Consider if our survey of commuters had involved a sample size of only 20, rather than 50. In that case, how confident could we be that our sample mean of 10.0 km was different from the regional figure of 13.1?

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{10.0 - 13.1}{\frac{9}{\sqrt{20}}}$$

$$t = \frac{-3.1}{\frac{9}{\sqrt{4.472}}}$$

$$t = \frac{-3.1}{2.01}$$

$$t = -1.54$$

With a sample size of 20, our degrees of freedom are 19, and the critical threshold for a confidence interval of 95 percent (2.5 percent in each tail) would be between -2.093 and +2.093. Since our t statistic lies within this range, we cannot reject the null hypothesis with 95 percent confidence. We have insufficient information with a sample size of 20 to back up the claim that our city commuters have shorter commutes than others in the broader region.

With larger sample sizes, the t- and z-statistics become perfect substitutes for one another.

Hypothesis testing is by no means infallible. A 95-percent confidence interval will support the wrong conclusion one time out of twenty (five percent). Whether this margin of sampling error -- typically referred to as alpha, or α=0.05 -- is acceptable is a judgment to be made on the basis of logic, theory, and policy. Alpha is also described as a Type I error -- the probability of falsely rejecting the null hypothesis. In our commuting example, with the larger sample of 50 workers, we found that the chance of observing a z-value as large as 2.43 above or below the true mean was 0.015; even so, with α=0.015, this means that fifteen times out of a thousand, a random sample will yield a z-score beyond this range, *even when there is no difference in the commutes of workers in the city compared to the entire region.* On the other hand, if there really is a systematic difference, and if we set alpha too low, then there is a greater chance that we will fail to reject the null hypothesis when we should have; this is known as Type II error.

One-Sample Tests for Proportions

Up to this point, our examples have focused on cases where we wish to evaluate the mean of a variable measured on an interval/ratio scale. A few adjustments are necessary if we wish to test hypotheses regarding proportions. Suppose in our sample of 50 city commuters, we ask each worker whether they work in the downtown central business district; ten (20 percent) say yes, while forty (80 percent) say no. If the true proportion of workers throughout the entire region who work downtown is 15.5 percent, do we have any evidence that our sample of workers who live in the city are more likely to work downtown?

For proportions, the observed proportion in our sample, *p*, can be understood as a random sample from a distribution that (if n is sufficiently large) approaches normality; the mean of this sampling distribution is the true, unobserved proportion, which we'll call ρ_0. The standard deviation of the sampling distribution -- the standard error of the mean -- is calculated with a formula that's a bit different from the one used for interval/ratio measures:

$$\sigma_{\rho_0} = \sqrt{\rho_0(1-\rho_0)/n}$$

This allows us to calculate a z-statistic to test against the standard normal curve:

$$z = \frac{p-\rho_0}{\sqrt{\rho_0(1-\rho_0)/n}}$$

if we do not know the true population proportion ρ_0, we can simply use the hypothesized proportion. For our example, the z-statistic is

$$z = \frac{0.20-0.155}{\sqrt{0.155(0.845)/50}}$$

$$z = \frac{0.045}{\sqrt{0.0026195}}$$

$$z = \frac{0.045}{0.051181}$$

$$z = 0.879$$

Consulting the reference table for the standard normal curve, we find that the probability of obtaining a z-statistic larger than 0.88 under the null hypothesis (i.e., when $p=\rho_0$) is .18943 (subtract the number you find in the table, .31057, from .50000). There is another .18943 probability 'in the other tail,' so the chance that we will obtain a z value more extreme than 0.88 is 0.37886. This figure is far larger than $\alpha=0.05$, and so we cannot reject the null hypothesis. We do not have sufficient evidence to conclude that the proportion of downtown workers is different for our city residents (20 percent) compared with workers throughout the region (15.5 percent).

Two-Sample Tests for Differences in Means

Suppose we compare our sample of 50 commuters, with their mean commute of 10.0 km and standard deviation of 9.0 km, to a survey using identical methods in another city; the other survey (n=50) yields a mean of 8.1 km, with a standard deviation of 8.8 km. Are the mean commute distances significantly different?

To test this kind of hypothesis, we first have to make assumptions about the 'spread' of the observations in each sample. The calculation of the t-statistic will be slightly different, depending on whether we have any reason to believe that the variances of the two samples should be equal. In general, it is more conservative -- meaning that we minimize Type I error, and we make it more difficult to reject the null hypothesis -- if we do not assume that the variances are equal. In this case the t-statistic is calculated as

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where the 1 and 2 subscripts refer to the different observed sample values. So for our comparison of commuters in two different cities, we have

$$t = \frac{10.0 - 8.1}{\sqrt{\frac{9.0^2}{50} + \frac{8.8^2}{50}}}$$

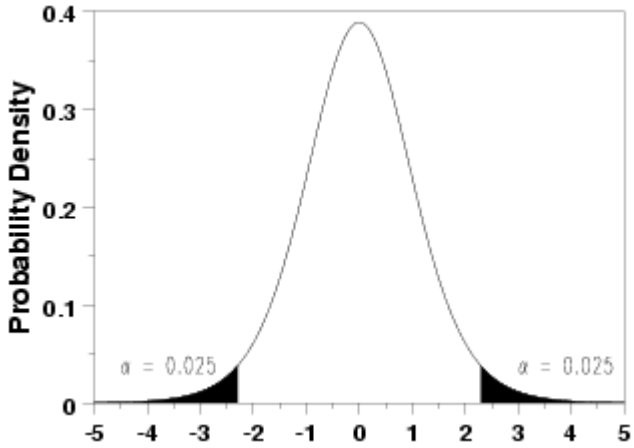
$$t = \frac{1.9}{\sqrt{\frac{81}{50} + \frac{77.44}{50}}}$$

$$t = \frac{1.9}{\sqrt{1.62 + 1.5488}}$$

$$t = \frac{1.9}{1.780}$$

$$t = 1.067$$

The most cautious, conservative calculation of the degrees of freedom for the two-sample t-test depends on the smallest sample size -- i.e., the minimum of (n1-1) and (n2-1); for our example, both of these terms yield a value of 49. Some tables of the t distribution do not provide all possible combinations of degrees of freedom, but we can see that for df=40, there is a 0.10 probability (a ten percent chance) of a t value larger than 1.303; for df=60, there is a 0.10 percent chance of a t value larger than 1.296. Our t-value falls far short of these thresholds, and so we cannot reject the null hypothesis: our evidence is insufficient to conclude that the mean commute distances between the two samples are significantly different.



Probability of exceeding the critical value (one-tailed)

df	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385
40.	1.303	1.684	2.021	2.423	2.704	3.307
60.	1.296	1.671	2.000	2.390	2.660	3.232
100.	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090

Figure 4. Distribution of Student's t. Source: National Institute of Standards and Technology (2010). *NIST/SEMATECH e-Handbook of Statistical Methods, Engineering Statistics Handbook*. Washington, DC: U.S. Department of Commerce, public domain.