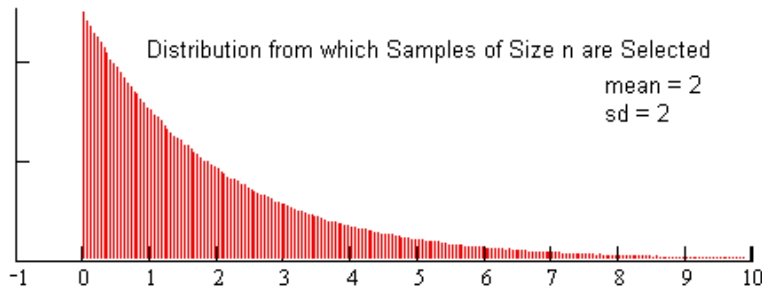


Understanding Tests of Hypothesis and Confidence Intervals

The two topics of confidence intervals and hypothesis test are very much related to each other. Students in an AP Statistics class will learn to construct confidence intervals for proportions and for means, and learn to perform hypothesis tests for both proportions and means. In this session, I will use the question of proportions to address the principles of confidence intervals and means to address the principles of hypothesis testing.

Confidence Intervals for Proportions

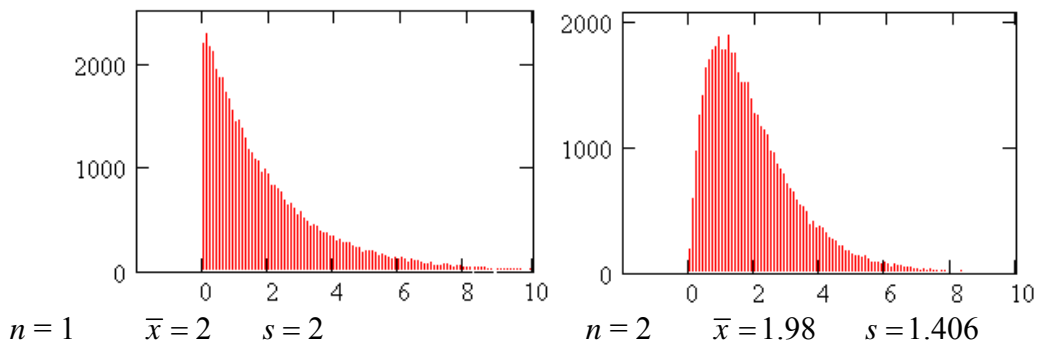
An understanding of Confidence Intervals for Proportions begins with the Central Limit Theorem, which is most easily approached through simulation. Suppose we repeatedly take a random sample of size n from a very asymmetric distribution with a mean $\mu = 2$ and a standard deviation $\sigma = 2$.

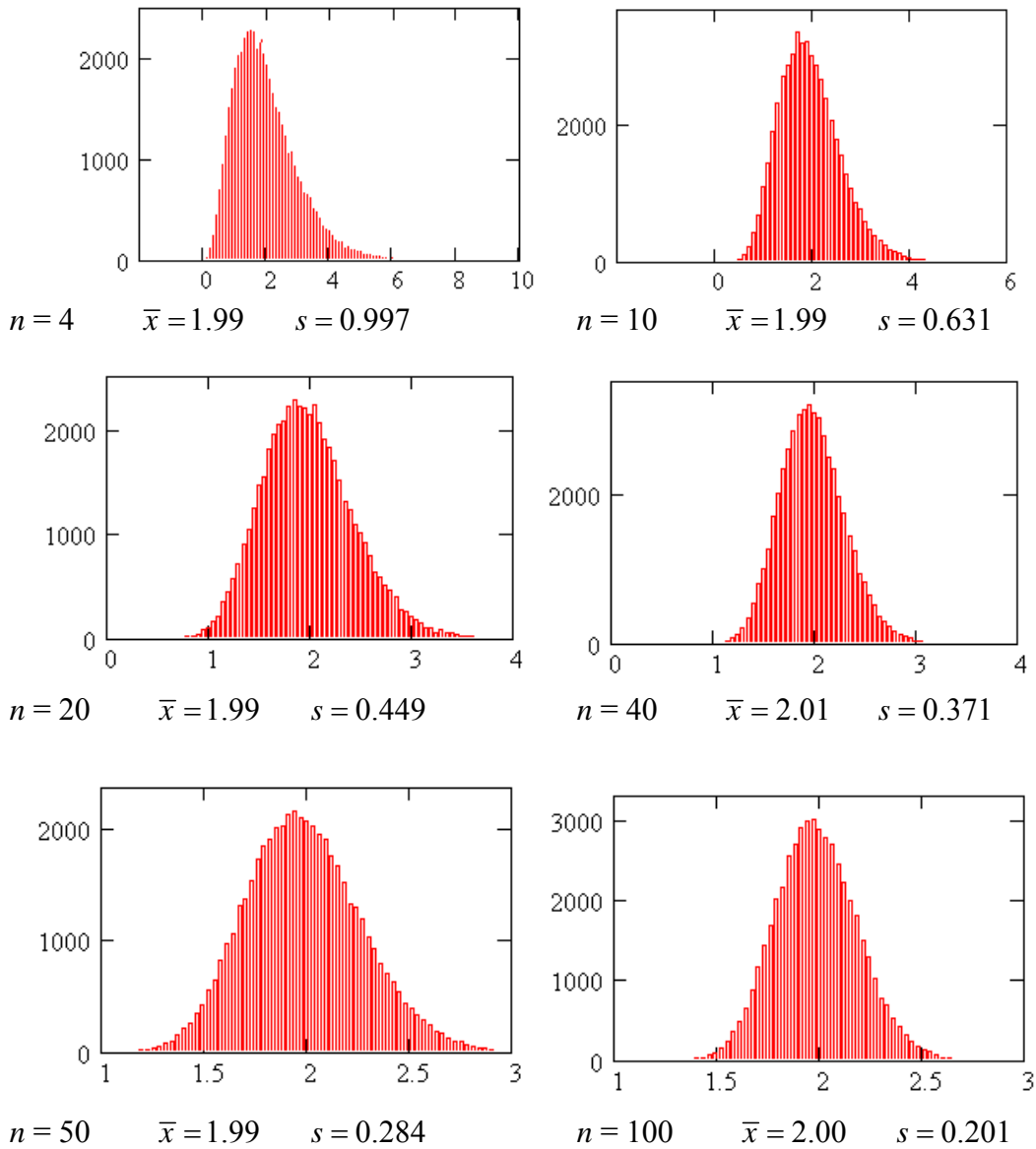


For each sample of size n , we compute the mean of the sample. We then plot a bar graph of those means. The distribution of these computed means of samples of size n is an approximation of the *Sampling Distribution of the Mean*. It is our knowledge of the sampling distribution of the mean that allows us to compute both confidence intervals and to perform tests of hypotheses.

Approximate Sampling Distribution of the Mean for Samples of Size n

In the plots below, 50,000 samples of size n are taken from the distribution shown above. The mean of each sample is computed and the distribution of those means plotted. Both the mean and standard deviation of the 50,000 sample means is shown.





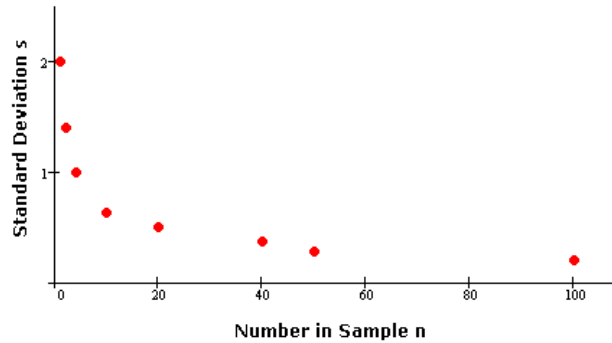
As we look at the plots, and the computed means and standard deviations, we notice several important things.

- The plots are getting more and more symmetric in shape.
- The mean of the sample means stays constant.
- The standard deviation of the sample means gets smaller and smaller.
-

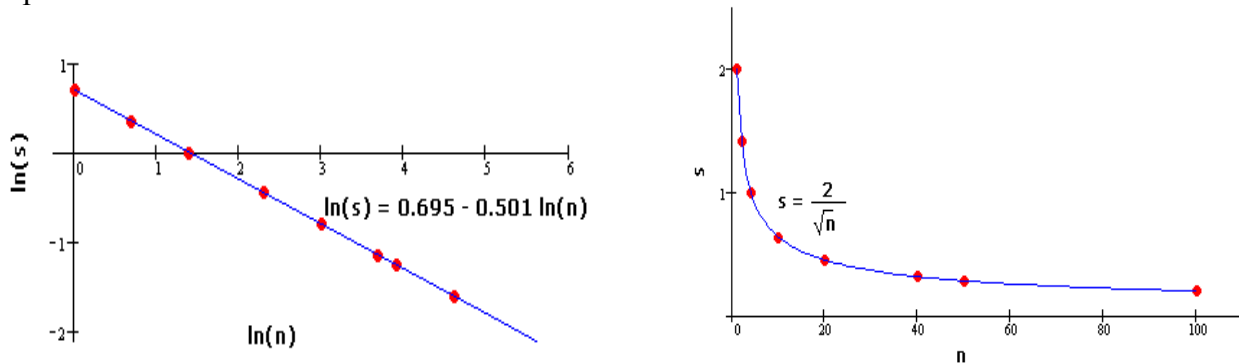
If we create a table of our values of n , \bar{x} (which approximates $\mu_{\bar{x}}$) and s (which approximates $\sigma_{\bar{x}}$), we can use data analysis to see what is happening to the standard deviations.

n	1	2	4	10	20	40	50	100
\bar{x}	2.00	1.98	1.99	1.99	1.99	2.01	1.99	2.00
s	2.00	1.406	0.997	0.631	0.499	0.371	0.284	0.201

The plot below shows the standard deviations plotted against the values of n . What function is being described by these data?



There appears to be both a vertical and horizontal asymptote in the plot, so we will try a log-log re-expression to linearize the data.



The log-log plot indeed is linear, and we find the line $\ln(s) = 0.695 - 0.501 \ln(n)$ models the re-expressed data. Solving for s , we find that $s = e^{0.695} n^{-0.501} = \frac{2}{\sqrt{n}}$.

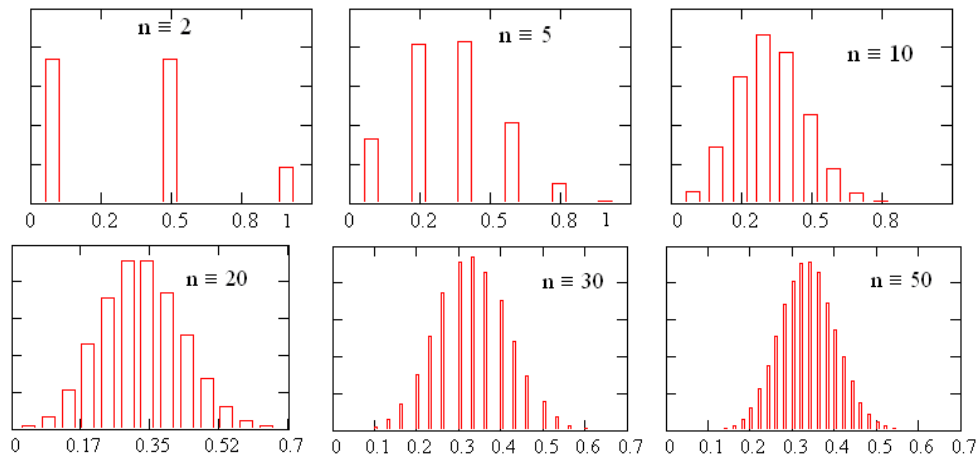
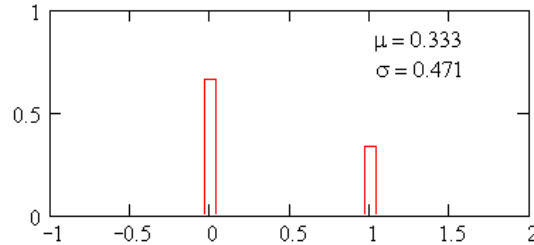
The Central Limit Theorem: If a random sample of size n is selected from a population X with mean μ and standard deviation σ , the *sampling distribution of the mean* approaches a normal distribution as n increases. Moreover, the sampling distribution of the mean has a mean of μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. So, \bar{x}_n has the approximate distribution $\bar{x}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ if n is large enough.

How large does n have to be? It depends on the shape of the distribution of x . In the example above, our population was very asymmetric. Once n was larger than 50, our approximation of the sampling distribution of the mean can be well approximated $\bar{x}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Why is this important? If we know a distribution is approximately normally distributed, then we know that 95% of the observations will fall within 2 standard deviations of the mean. If we know the mean and standard deviation of the parent distribution, we can predict the behavior of the mean.

Simulation from a 0-1 Box

The same simulation is repeated, only this time, we use a population of only zeros and ones. One-third of this population is a one, and two-thirds a zero. We can think of this a box with an unlimited number of balls, each labeled with a zero or a one. We draw n ball from the box and compute the mean of the numbers on the balls drawn.



Notice that the distribution becomes more and more “Normal” as n increases. The distribution is always discrete, but a normal approximation is reasonable once n is large enough. Notice that if $n = 30$, then $np = 10$, one of the common rules of thumb for using a normal approximation.

What’s Special about a 0-1 box?

Suppose a box has N balls, with n of them are marked with a one and $N - n$ marked with a zero. What is the mean and standard deviation of this population of numbers on the balls.

The mean is just $\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{n(1) + (N - n)(0)}{N} = \frac{n}{N}$. So $\mu = p$, the proportion of 1’s in the box.

Now compute $\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N (x_i - p)^2}{N} = \frac{n(1 - p)^2 + (N - n)(0 - p)^2}{N}$. This simplifies greatly to

$$\sigma^2 = \frac{n(1 - p)^2 + (N - n)(p)^2}{N} = \frac{n}{N}(1 - p)^2 + \left(1 - \frac{n}{N}\right)(p)^2 = p(1 - p)^2 + (1 - p)(p)^2$$

so

$\sigma^2 = p(1-p)^2 + (1-p)(p)^2 = p(1-p)(1-p+p) = p(1-p)$. Finally, $\sigma = \sqrt{p(1-p)}$. The zero-one box has the very special property that if we know the proportion of ones in the box, then we know that the mean is p , and we also know the standard deviation is $\sqrt{p(1-p)}$.

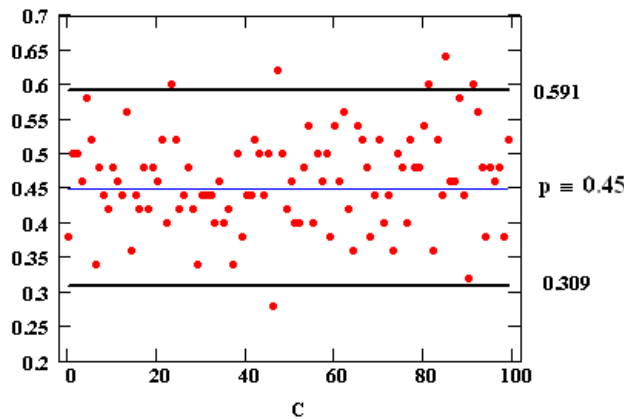
Combining this principle with the Central Limit Theorem, we have something very useful. If a box contains zeros and ones, with the proportion of 1's being p , we can predict very accurately what the mean of a random sample of n draws from the box will be. The mean will be the proportion of 1's in the box, so we can predict the proportion that will be in a sample of size n . We can do this because we know that the sampling distribution of the mean, \hat{p} , is approximately normal

with a mean of p and standard deviation of $\sigma = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$. So, 95% of our observations are

expected to be within $2\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ of p . We can try this and see. One hundred samples of size

$n = 50$ were taken from a population of 45% ones and 55% zeros. The 100 sample proportions are plotted below along with the line $2\frac{\sqrt{p(1-p)}}{\sqrt{n}} = 2\frac{\sqrt{0.45(0.55)}}{\sqrt{50}} = 0.141$ on either side. In this

simulation, 6 of the 100 sample proportions fell outside the designated region.



The principle behind Confidence Intervals and Hypothesis Testing:
If we know what is in the box we can predict what is likely to come out.

We can tell how likely it is to achieve any sample proportion, since I know the Sampling

Distribution of the Mean is $\bar{x}_n = \hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

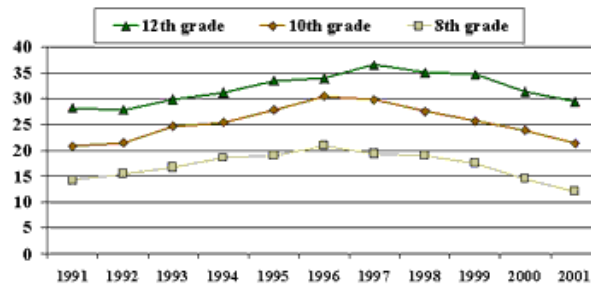
The Problem

OK. So life is good. If we know what is in the box, we can predict what comes out in a random sample from the box. Only in the world of Survey Sampling, *we know what came out of the box and what to predict what was in the box*. We know how to do the problem backwards! If we know the population proportion p , we can create an interval that will contain 95% of the sample

proportions \hat{p} . We expect 95% of the observed \hat{p} 's to be within $2\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ of p . But all we really have is one observed \hat{p} . We know how far we expect \hat{p} to be from p , but how far should p be from \hat{p} ? What boxes could that \hat{p} come out of? For which p 's would it be a typical result in a random draw. The confidence gives an answer to this question.

The following example comes from my colleague Floyd Bullard. The graphs and data below are from the Gallup organization via www.gallup.com.

Recent Trends in Teen Smoking:(% who have smoked a cigarette in the last 30 days.)



In the chart, we see that 29% of about 1000 12th graders polled in 2001 said they had smoked a cigarette in the last 30 days. So we have observed $\hat{p} = 0.29$, but of course it would have been different with a different sample. What are the potential population proportion p 's that could have given rise to this observation? If we had asked every 12th grader, what would the true proportion be?

The following table shows several different possible “models”, *i.e.*, different possible values of p . Complete the table by finding the mean and standard deviation of \hat{p} under each of those models, and then the range of “plausible” values of \hat{p} . (We will define a “plausible” value of \hat{p} to be any value of \hat{p} that falls within ± 2 standard deviations of its mean.

model	$\mu_{\hat{p}}$	$\sigma_{\hat{p}}$	$\mu_{\hat{p}} - 2\sigma_{\hat{p}}$	$\mu_{\hat{p}} + 2\sigma_{\hat{p}}$	Is $\hat{p} = 0.29$ plausible ?
p = 0.24	0.24	0.0135	0.2130	0.2670	No
p = 0.25	0.25	0.0137	0.2226	0.2774	No
p = 0.26	0.26	0.0139	0.2322	0.2878	No
p = 0.27	0.27	0.0140	0.2420	0.2980	Yes
p = 0.28	0.28	0.0142	0.2516	0.3084	Yes
p = 0.29	0.29	0.0143	0.2614	0.3186	Yes
p = 0.30	0.30	0.0145	0.2710	0.3290	Yes
p = 0.31	0.31	0.0146	0.2808	0.3392	Yes
p = 0.32	0.32	0.0148	0.2904	0.3496	No
p = 0.33	0.33	0.0149	0.3002	0.3598	No
p = 0.34	0.34	0.0150	0.3100	0.3700	No

For what proportions is our observation one of those that is likely to come out?

One interpretation of 95% confidence interval that is equivalent to the more popular one but which gets at THE particular confidence interval you constructed: “This is the set of all parameter values which are ‘consistent’ with the observed data, where ‘consistent’ means that the sample statistic observed falls in the middle (‘typical’) 95% of the sampling distribution for that parameter.” This is no less cumbersome than the other interpretation, but at least it refers to the particular confidence interval you created. The confidence interval for the sample above is (0.26, 0.32). Our observed value of 0.29 is a typical response from any proportion in that interval. It would be an unusual response from proportions outside that interval.

When we say we have computed a 95% confidence interval, how do we interpret this level of confidence? The classic incorrect interpretation that students always give is that the “true population proportion will be in the interval (0.26, 0.32) 95% of the time”, or “the probability that the interval (0.26, 0.32) captures the true proportion is 0.95”. Both of these interpretations are incorrect because they place confidence in the interval created, (0.26, 0.32). The confidence we have is not in the interval. The confidence is in the method for creating the interval. The method will “work” 95% of the time. We don’t know if it “worked” or not when we created the interval (0.26, 0.32).

Conditions: $np > 5$ (or 10)

One way to think about the np condition is to consider confidence intervals for proportions. If, when we compute a 95% confidence interval for proportions, we obtain an interval (–0.1, 0.3) or (0.85, 1.05), we know that our interval is based on too few data, since the endpoints exceed the limits of 0 and 1. To have confidence in our confidence interval, the endpoints of the interval must be between 0 and 1. How large must n be to achieve this?

We need $p - 2\sqrt{\frac{p(1-p)}{n}} > 0$ and $p + 2\sqrt{\frac{p(1-p)}{n}} < 1$ where p is the population parameter (approximated by \hat{p}). If $p - 2\sqrt{\frac{p(1-p)}{n}} > 0$, then $p > 2\sqrt{\frac{p(1-p)}{n}}$ or $p^2 > \frac{4p(1-p)}{n}$.

Simplifying, we find that $np > 4(1-p)$. This lower bound is only an issue when p is small and $1-p$ is close to 1. So if $np > 4$, we will get an interval whose lower bound is greater than 0. But we don’t know p , we only know \hat{p} . By requiring $n\hat{p} > 5$, we give ourselves some wiggle-room and confidence that, unless our \hat{p} is far from p , our confidence interval will be OK. Solving $p + 2\sqrt{\frac{p(1-p)}{n}} < 1$ in a similar manner and adding in the wiggle-room for \hat{p} will generate $n(1-\hat{p}) > 5$. If we want 99% confidence intervals we need $np > 9$ or $n\hat{p} > 10$.

Hypothesis Tests: t-tests

Hypothesis tests are related to Confidence Intervals, but we give an explicit statement of what we think is “in the box”. Each test of hypothesis has a null hypothesis. It is the null hypothesis, describing the situation when “nothing interesting is happening”, that tells us what is “in the box”. As before, if we know what is in the box we can predict what is likely to come out of the box. The basic principle of hypothesis testing is to compare our observation to the expectation

from a random draw from the box described by the null hypothesis. We will look at several examples.

One-Sample t-test

Is the normal human temperature 98.6? When this standard was determined, the mean temperature was measured in degrees Celsius, then rounded up to 37 degrees and converted to degrees Fahrenheit using $32 + 1.8(37) = 98.6$. Rounding up before conversion may have produced a result that is higher than the true average.

A random sample of $n = 18$ individual is selected and their mean temperature is $\bar{x}_{18} = 98.217$ with a standard deviation $s = 0.684$.

98.2	97.8	99.0	98.2	97.8	98.4	99.7	98.2	98.6
97.4	97.6	98.4	98.0	99.2	98.6	97.1	97.2	98.5

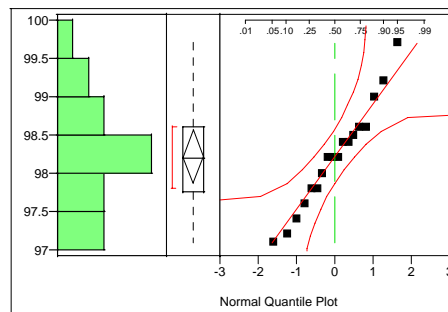
Do these data provide evidence that the true mean temperature is less than 98.6, or could the observed mean of 98.217 be easily explained as chance variation (is it one of those typical responses from the box)?

Our null hypothesis is that the true mean μ is 98.6 degrees. Our alternative is that the true mean is less than 98.6 degrees. We have a one-sided alternative, since we believe the rounding would create an error in only one direction. We state our hypotheses as $H_0 : \mu = 98.6$ and $H_a : \mu < 98.6$. If we take a

random sample of size n from a population whose distribution is normal, the distribution of the statistic $t = \frac{\bar{x}_n - \mu}{\left(\frac{s}{\sqrt{n}}\right)}$ is a t -distribution with $n - 1$ degrees of freedom. This tells us what is in the box,

and we can determine what is likely to come out of the box. So one condition that must be met is that the 18 observations could plausibly represent a random sample from a normal distribution. We use a normal probability plot to the plausibility of normality.

Body Temperature

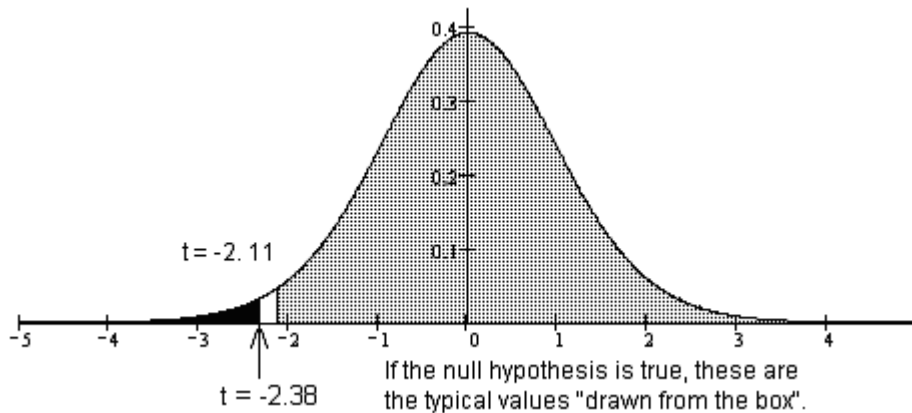


The data are not significantly non-normal, there is no indication that a normal distribution is implausible. There are no outliers in the data, so we can trust our computed statistics \bar{x} and s .

Suppose we set our α -level at $\alpha = 0.025$. This means that we will accept as implausible any mean that happens with a probability less than 0.025 in a random draw from the box described by the null hypothesis.

If the null hypothesis is true, we know what to expect to draw from the box. We would expect 95% of our observed means to generate a t -score between $t = -2.11$ and $t = 2.11$. So, we expect to find 97.5% of the observed means to generate t -scores larger than $t = -2.11$. If our observed t -score falls in this range, we consider it to be plausibly explained as chance variation. It is a value we would expect to get when the null hypothesis is true. If, however, our observed t -score falls outside this range, we think chance as an explanation is not plausible, so something must be going on. We will reject the null hypothesis in favor of the alternative.

$$\text{So, we compute our } t\text{-score of } t_{17} = \frac{\bar{x}_8 - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{98.217 - 98.6}{\left(\frac{0.684}{\sqrt{18}}\right)} = -2.38.$$



While chance can explain our observation, it is not a likely result if the null hypothesis is true. The p -value associated with a t -score of -2.38 with 17 degrees of freedom $p = 0.0146$. If the null hypothesis is true, we could have randomly drawn observations that would have given us a t -score of -2.38 or even further from zero, but this would have happened with a probability less than 0.015. We have a choice: either the null hypothesis is true and we have just seen a rare event or the null hypothesis is false and we have just seen a typical event. Believing in typical events, we therefore reject chance as an explanation. We reject the null hypothesis and say that our observations support the alternative hypothesis. We believe that the mean body temperature is less than 98.6 degrees.

z vs t

Older texts often suggest that the z -test should be used when $n > 50$ (or some other rule of thumb). This is an application of the Central Limit Theorem. Before computers and calculators made computing t -scores easy, this approach had many adherents. However, it is easy to compute p -values for t -scores with arbitrary degrees of freedom. If the standard deviation of the population is being estimated by s , then the sampling distribution of the mean is a t -distribution. It can be approximated by a Normal distribution. There is no reason to use the approximation when we can easily use the correct distribution. Moreover, the t -tests are robust. That means that they are fairly immune to non-compliance with the theoretical conditions. If the theoretical conditions are not met, the t -distribution is just as good (actually a little better) an approximation as the z . So, I never have my student use z instead of t .

Matched Pair Test: Often we are interested in comparing two different treatments. A matched pairs design compares the responses of two experimental units that are matched in some important

way. In this setting, we compute the differences in the responses of our matched units and are interested in whether the mean difference observed could plausibly be zero. The strongest matching is to use the same unit with both treatments, such as in a pre-test/post-test situation.

As an example, consider Problem 5 from the 1997 AP Statistics exam. A company bakes computer chips in two ovens, A and B, and wants to know if the percentage of defective chips produced by the two ovens are equal. The chips are randomly assigned to an oven and 100's of chips are baked each hour, so the percentage of defective chips are matched by the hour they were created. The experimenter has no option in the analysis. Since the experiment was conducted as a matched pair design (whether for good or ill) it must be analyzed as a matched pair design.

The data is given in the table below:

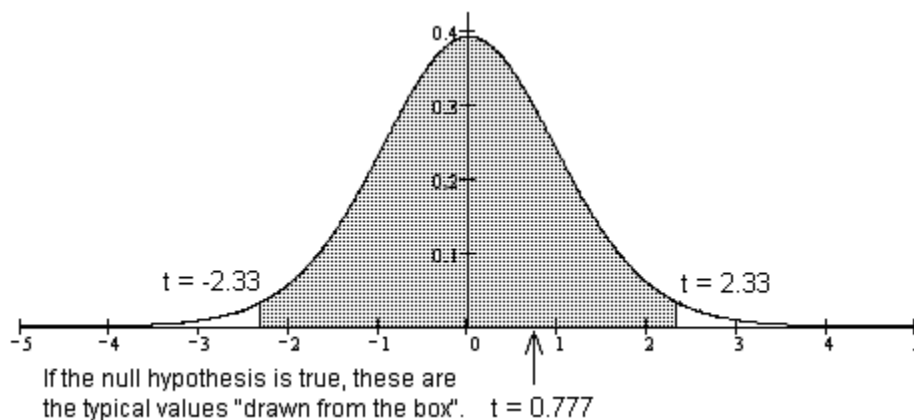
Hour	1	2	3	4	5	6	7	8	9
Oven A	45	32	34	31	35	37	31	30	27
Oven B	36	37	33	34	33	32	33	30	24
Difference	9	-5	1	-3	2	5	-2	0	3

Our null hypothesis is that the mean difference is zero. Do the observed differences 9, -5, 1, -3, 2, 5, -2, 0, 3 support this hypothesis?

Formally, our hypotheses are $H_0: \mu_D = 0$ and $H_a: \mu_D \neq 0$. There are no outliers and the normal probability

plot gives no evidence of non-normality, so we can proceed with our t -test on the differences. If the null hypothesis is true, then our 9 observations represent a random sample from a normal distribution with a mean of 0 and standard deviation estimated by $s = 4.285$. The sampling distribution of these means is a t -distribution with 8 degrees of freedom. From this distribution, we would expect 95% of the observed mean differences to generate t -scores between $t = -2.33$ and $t = 2.33$. If our t -score falls in this range, we consider it to be plausibly explained as chance variation. If our t -score falls outside this range, we think chance as an explanation is not plausible, so something must be going on. We will reject the null hypothesis in favor of the alternative.

Again, we compute our t -score of $t_8 = \frac{\bar{x}_9 - \mu}{\left(\frac{s}{\sqrt{n}}\right)} = \frac{1.11 - 0}{\left(\frac{4.285}{\sqrt{9}}\right)} = 0.777$.



Clearly, our observations produce a t -score that is consistent with the null hypothesis. It is among those t -scores that can plausibly be explained by natural variation or chance. The p -value

associated with a t -score of 0.777 with 8 degrees of freedom is $p = 0.4596$. There is no evidence that the mean difference is not zero. We “fail to reject” the null hypothesis. Notice that we do not accept the null hypothesis. We don’t know if the true mean difference is actually zero, but we don’t have any evidence that says it isn’t. We conclude that there is no evidence that the two ovens differ with respect to the percentage of defective chips they produce.

Two-Sample Tests: In the two sample setting, we again are interested in comparing mean the mean response from two treatments. In this case, there is not matching of experimental units. Our question concerns the observed difference in the mean responses. Can the observed difference be plausibly explained by chance?

As an example, consider problem #4 from the 2000 AP Statistics exam. A study compared the mental skills of two sets of babies, those who used walkers and those who never used walkers. The 54 babies in the study who used walkers had a mean mental skill score of 113 with a standard deviation of 12 while the 55 babies who didn’t use walkers had a mental skill score of 123 with a standard deviation of 15. The two mean scores are not the same. Does that mean that the two groups differ with respect to their mean mental scores? We will say they differ only if the observed difference can not be plausibly explained as chance variation.

In the two sample situation, the two sets of observations must be independent random samples from a normal distributions. If we are to believe our measures of center and spread, there must be no outliers in the data. Since we don’t have the data, we cannot test to see if these conditions are met. So we must assume that they are met.

Our null hypothesis is that the two means are equal, or that the difference in the two means is zero. Our alternative is that the difference is not zero. Formally, we say

$$\begin{aligned} H_0 : \mu_W - \mu_N &= 0 \\ H_a : \mu_W - \mu_N &\neq 0 \end{aligned}$$

There is rarely a good reason to use the pooled form of the two-sample test. We will discuss why shortly. If the null hypothesis is true, is our t -score one that would naturally result from random sampling from the distribution described by the null hypothesis?

$$\text{Our } t\text{-score is computed } t = \frac{(\bar{x}_W - \bar{x}_N) - (\mu_W - \mu_N)}{\sqrt{\frac{s_W^2}{n_W} + \frac{s_N^2}{n_N}}} = \frac{(113 - 123) - (0)}{\sqrt{\frac{12^2}{54} + \frac{15^2}{55}}} = \frac{-10}{2.6} = -3.84.$$

This test has 102.8 degrees of freedom and a p -value of 0.0002. While it is possible to achieve a t -score of -3.84 by a random sample under the null hypothesis, it is certainly not a likely event. Consequently, it doesn’t offer chance as a plausible explanation of our observed difference. We reject chance as the explanation, and therefore reject the null hypothesis of equal means in favor of the alternative of unequal means. The mean mental skills score for babies who used walkers is not the same as the mean mental skills score for babies who did not use walkers. We believe the babies who used walkers have a lower average mental skill score.

Pooled vs Unpooled

In the example above, we used the unpooled form of the 2-Sample Test. If two random samples are selected independently and randomly from normal populations with common standard deviations, and if the null hypothesis $\mu_1 = \mu_2$ is true, then T_p has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. However, the test based on this statistic is not robust with respect to departures from the assumption of common variances, especially if n_1 and n_2 are unequal. And we never can be sure that $\sigma_1 = \sigma_2$.

A natural test statistic for comparing μ_1 and μ_2 is the unpooled t , $t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. This

statistic does not have a t distribution even if $\sigma_1 = \sigma_2$ and $\mu_1 = \mu_2$. However, regardless of whether $\sigma_1 = \sigma_2$, the critical values of its null distribution (that is, when $\mu_1 = \mu_2$) can be approximated quite well by those of a t distribution whose degrees of freedom depend on the observed data. As John Tukey suggests, “an approximate solution to the right problem is better than an exact solution to the wrong problem.” The pooled t -test is the exact solution to the wrong problem and the unpooled t -test is the approximate solution to the right problem.

It seems advisable in an introductory statistics course just to teach and use the unpooled procedure for tests of $\mu_1 = \mu_2$ and for confidence intervals for $\mu_1 - \mu_2$. Teaching both methods in a first course would likely only confuse students. Once students move on to the techniques of ANOVA, they can learn about the pooled test, since ANOVA is a generalization of the pooled version of the 2-Sample Test.

Hypothesis Test vs Confidence Interval for Difference in Proportions

The standard error for a confidence interval for the difference in two proportions is $se = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ while the standard error for a hypothesis test is

$se = \sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where $\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$. Students are always baffled by the difference

in the computations used for hypothesis tests and confidence intervals for the difference in two proportions. Why are two different formulas used?

The distinction goes back to the null hypothesis. For a hypothesis test, the null hypothesis is that the two proportions are equal, so $p_1 = p_2$. As we have already seen, in a zero-one box for proportions, if the means are the same, then the standard deviations must also be the same. So, for the hypothesis test, we begin with the assumption of a common standard deviation. The value \hat{p}_c is our best pooled estimate of that common variance.

The confidence interval does not begin with an assumption that the two proportions are equal. In fact, we use a confidence interval to estimate the size of the difference between them. So we think we have two different proportions, which means we have two different standard deviations.