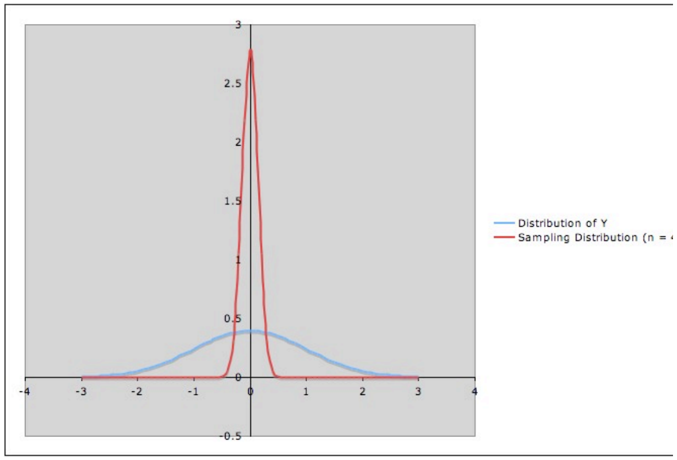


The following chart summarizes which model assumptions are necessary to prove which part of the theorem:

	Conclusions ***about Sampling Distribution*** (Distribution of $\overline{Y}_n$)		
	1: Normal	2: Mean μ	3: Standard deviation $\sigma/\sqrt{n}$
Assumption 1 (Y normal)	√		
Assumption 2 (simple random samples)	√		√

(Note: The conclusion that the sampling distribution $\overline{Y}_n$ has the same mean as Y does not involve either of the model assumption

- The conclusions of the theorem will allow us to do the following:
  - If we specify a probability (we'll use 0.95 to illustrate), we can find a number *a* so that

    (*) The probability that $\overline{Y}_n$ lies between μ - a and μ + a is approximately 0.95:

    $$P(\mu - a < \overline{Y}_n < \mu + a) \cong 0.95$$

    Caution: It is important to get the *reference category* straight here. This amounts to keeping in mind what is a random variable and what is a constant. Here, $\overline{Y}_n$ *is the random variable* (because the *sample* is varying), whereas μ *is constant*.

    Note: The z-procedure for confidence intervals is only an approximate procedure; that is why the "approximately" is in (*) and below. Many procedures are "exact"; we don't need the "approximately" for them.

- A little algebraic manipulation (which can be stated in words as, “If the estimate is within a units of the mean μ, then μ is within a units of the estimate”) allows us to restate (*) as

  (**) The probability that μ lies between $\overline{Y}_n$ - a and $\overline{Y}_n$ + a is approximately 0.95:
  $P(\overline{Y}_n - a < \mu < \overline{Y}_n + a) \cong 0.95$

  Caution: It is again important to get the reference category correct here. It hasn't changed: it is still the *sample* that is varying, *not* μ. So the probability refers to $\overline{Y}_n$, <u>*not*</u> to μ.

  Thinking that the probability in (**) refers to μ is a **common mistake** *in interpreting confidence intervals.*

  It may help to restate (**) as:

  (***) The probability that the interval from $\overline{Y}_n$ - a to $\overline{Y}_n$ + a contains μ is approximately 0.95.

  Note: The reference category is still the sample – the sample is varying, but μ is not varying, However, as the sample varies, so does $\overline{Y}_n$, and hence in this restatement, the *interval* is varying. This is helpful to remember.

- We are now faced with two possibilities (assuming the model assumptions are indeed all true):

  1) The sample we have taken is one of the approximately 95% for which the interval from $\overline{Y}_n$- a to $\overline{Y}_n$+ a *does* contain μ. ☺

  2) Our sample is one of the approximately 5% for which the interval from $\overline{Y}_n$ - a to $\overline{Y}_n$ + a does *not* contain μ. ☹

- Unfortunately, *we can't know which of these two possibilities is true for the sample we have*. ☹

- Since this is the best we can do, we calculate the values of $\overline{Y}_n$ - a and $\overline{Y}_n$ + a *for the sample we have*, and call the resulting interval an approximate *95% confidence interval for μ.*
    - We ***can*** say that *we have obtained the confidence interval by using a procedure that, for approximately 95% of all simple random samples from Y, of the given size n, produces an interval containing the parameter μ that we are estimating.*
    - Unfortunately, we ***can't know*** whether or not the sample we have used is one of the approximately 95% of "good" samples that yield a confidence interval containing the true mean μ, or whether the sample we have is one of the approximately 5% of "bad" samples that yield a confidence interval that does not contain the true mean μ.
    - We can just say that we have used a procedure that "works" about 95% of the time.
    - In other words, confidence is in *the degree of reliability of the method*, not in the result.
    - Various web demos can demonstrate.

In general: We can follow a similar procedure for many other situations to obtain confidence intervals for parameters.

- Each type of confidence interval procedure has its own model assumptions.
    - *If the model assumptions are not true, we can’t be sure that the procedure does what is claimed.*
    - However, some procedures are *robust* to some degree to some departures from models assumptions -- i.e., the procedure works pretty closely to what is intended if the model assumption is not too far from true.
    - Robustness depends on the particular procedure; there are no "one size fits all" rules.

- We can decide on the "level of confidence" we want;
  - E.g., we can choose 90%, 99%, etc. rather than 95%.
  - Just which level of confidence is appropriate depends on the circumstances. (More later)
- The *confidence level* is *the proportion (expressed as a percentage) of samples for which the procedure results in an interval containing the true parameter*. (Or approximate proportion, if the procedure is not exact.)
- However, *a higher level of confidence will produce a wider confidence interval*. (See demo)
  - i.e., less certainty in our estimate.
  - So *there is a trade-off between level of confidence and degree of certainty*.

- Sometimes the best we can do is a procedure that only gives *approximate* confidence intervals.
  - i.e., the sampling distribution can be described only approximately.
  - i.e., there is one more source of uncertainty.
  - This is the case for the large-sample z-procedure.

- If the sampling distribution is not symmetric, we can't expect the confidence interval to be symmetric around the estimate.
  - In this case, there might be more than one reasonable procedure for calculating the endpoints of the confidence interval.
  - This is typically the case for variances, odds ratios, and relative risks, which usually have skewed distributions such as F or chi-squared.

- There are variations such as "upper confidence limits" or "lower confidence limits" where we are only interested in estimating how large or how small the estimate might be.

V. MORE ON FREQUENTIST HYPOTHESIS TESTS

We'll now continue the discussion of hypothesis tests.

Recall: Most commonly used frequentist hypothesis tests involve the following elements:

1. Model assumptions
2. Null and alternative hypothesis
3. A test statistic (something calculated by a rule from a sample)
   - This needs to have the property that extreme values of the test statistic are rare, and hence cast doubt on the null hypothesis.
   - The test statistic will have a certain sampling distribution.
4. A mathematical theorem saying, "If the model assumptions and the null hypothesis are both true, then the sampling distribution of the test statistic has this particular form."

The exact details of these four elements will depend on the particular hypothesis test.

Illustration: One-sided t-test for a Sample Mean

In this situation, the four elements above are:

1. *Model assumptions*:
- The random variable Y is *normally distributed*.
- Samples are *simple random samples*.

2. *Null and alternate hypotheses:*
- *Null hypothesis*: The population mean μ of the random variable *Y is* μ_0. (i.e., $\mu = \mu_0$)
- *Alternative hypothesis*: The population mean μ of the random variable Y is greater than μ_0. (i.e., $\mu > \mu_0$)

3. *Test statistic:* For a simple random sample $y_1, y_2, \ldots, y_n$ of size n, we define the *t-statistic as*

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad ,$$

where

$$\bar{y} = (y_1 + y_2 + \ldots + y_n)/n \quad \text{(sample mean)},$$

and

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2} \quad \text{(sample standard deviation)}$$

The *sampling distribution* for this test is then the distribution of the random variable T_n defined by random process and calculation,

> “Randomly choose a simple random sample of size n and calculate the t-statistic for that sample.”

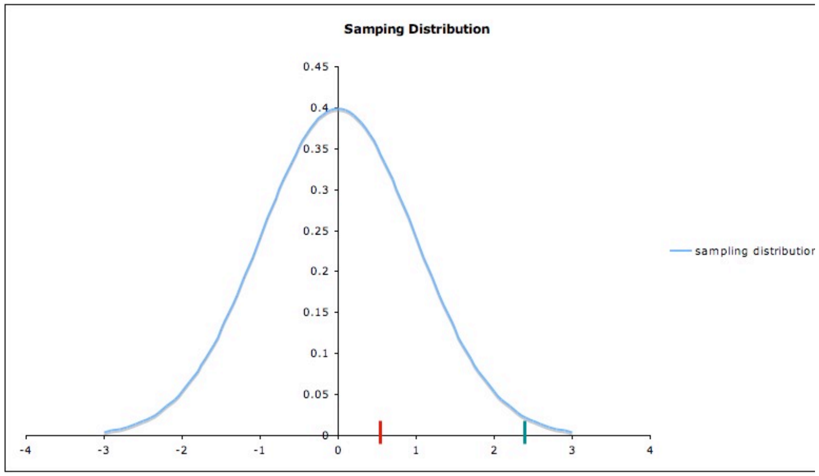
4. The mathematical theorem associated with this inference procedure (one-sided t-test for population mean) says:

> If *the model assumptions are true* and *the null hypothesis is true*, then the sampling distribution is the *t-distribution with n degrees of freedom*.

(For large values of n, the t-distribution looks very much like the standard normal distribution; but as n gets smaller, the peak gets slightly smaller and the tails go further out.)

The reasoning behind the hypothesis test uses the sampling distribution *and* the value of the test statistic for the sample that has actually been collected (the actual data).

1. First, *calculate the t-statistic for the data*
2. Then *consider where the t-statistic for the data at hand lies on the sampling distribution*. Two possible values are shown in red and green, respectively, in the diagram below.
    - *Remember that this picture depends on the validity of the model assumptions and on the assumption that the null hypothesis is true.*



Case 1: If the t-statistic lies at the *red* bar (around 0.5) in the picture, *nothing is unusual*; our data are consistent with the null hypothesis.

Case 2: If the t-statistic lies at the *green* bar (around 2.5), then the data would be fairly *unusual* -- assuming the null hypothesis is true.

So ***a t-statistic at the green bar would cast some reasonable doubt on the null hypothesis***.

A t-statistic even further to the right would cast even more doubt on the null hypothesis.

Note: A little algebra will show that if t = $\dfrac{\bar{y}-\mu_0}{s/\sqrt{n}}$ is unusually large, then so is $\bar{y}$, and vice-versa

p-Values

The idea: The *p-value* is a quantitative measure of how unusual a particular test statistic is, with lower p-values indicating more unusual data.

The general definition:

> *p-value* = the probability of obtaining a test statistic *at least as extreme as* the one from the data at hand, assuming the model assumptions and the null hypothesis are all true.

Elaboration: The interpretation of "*at least as extreme as*" depends on the *alternative* hypothesis.

- For the one-sided alternative hypothesis $\mu > \mu_0$ (as in our example), "at least as extreme as" means "at least as great as".
  - Recalling that the probability of a random variable lying in a certain region is the area under the probability distribution curve over that region, we conclude that for this alternative hypothesis, *the p-value is the area under the sampling distribution curve to the right of the test statistic calculated from the data.*
  - Note that, in the picture, the p-value for the t-statistic at the green bar is much less than that for the t-statistic at the red bar.

- Similarly, for the *other* one-sided alternative, $\mu < \mu_0$, the *p-value is the area under the sampling distribution curve to the left of the calculated test statistic*.
  - Note that for this alternative hypothesis, the p-value for the t-statistic at the green bar would be much greater than the t-statistic at the red bar, but both would be large as p-values go.
- For the two-sided alternative $\mu \neq \mu_0$, the *p-value would be the area under the curve to the right of the absolute value of the calculated t-statistic, plus the area under the curve to the left of the negative of the absolute value of the calculated t-statistic.*
  - Since the sampling distribution in the illustration is symmetric about zero, the two-sided p-value of, say the green value, would be twice the area under the curve to the right of the green bar.

Recall that in the sampling distribution, we are only considering samples

- from the *same random variable*,
- that *fit the model assumptions* and
- *of the same size as the one we have*.

So if we spelling everything out, the definition of p-value reads:

p-value = the probability of obtaining a test statistic *at least as extreme* as the one from the data at hand, *assuming*

- *the model assumptions are all true, and*
- *the null hypothesis is true, and*
- *the random variable is the same (including the same population), and*
- *the sample size is the same*.

Comment: The preceding discussion can be summarized as follows:

If we obtain an unusually small p-value, then (at least) one of the following must be true:

I. At least one of the model assumptions is not true (in which case the test may be inappropriate).
II. The null hypothesis is false.
III. The sample we have obtained happens to be one of the small percentage that result in an unusually small p-value.

Thus, if the p-value is small enough *and* all the model assumptions are met, then *rejecting the null hypothesis in favor of the alternate hypothesis* can be considered a rational decision, based on the evidence of the data used.

Comments:

1. How small is "small enough" is a judgment call.

2. "Rejecting the null hypothesis" does *not* mean the null hypothesis is false or that the alternate hypothesis is true. (Why?)

VI. MISINTERPRETATIONS AND MISUSES OF P-VALUES

Recall:

p-value = the probability of obtaining a test statistic at least as extreme as the one from the data at hand, *assuming*:

- the model assumptions for the inference procedure used are all true*, and*
- the null hypothesis is true*, and*
- the random variable is the same (including the same population)*, and*
- the sample size is the same.

Notice that this is a *conditional probability*: The probability that something happens, given that various other conditions hold. *One* **common mistake** *is to neglect some or all of the conditions*.

Example A: Researcher 1 conducts a clinical trial to test a drug for a certain medical condition on 30 patients all having that condition.

- The patients are randomly assigned to either the drug or a look-alike placebo (15 each).

- Neither the patients nor the medical personnel involved know which patient takes which drug.

- Treatment is exactly the same for both groups, except for whether the drug or placebo is used.

- The hypothesis test has null hypothesis "proportion improving on the drug is the same as proportion improving on the placebo" and alternate hypothesis "proportion improving on the drug is greater than proportion improving on the placebo."

- The resulting p-value is p = 0.15.

Researcher 2 does another clinical trial on the *same drug*, with the *same placebo*, and *everything else the same except* that 200 patients are randomized to the treatments, with 100 in each group. The same hypothesis test is conducted with the new data, and the resulting p-value is p = 0.03.

Are these results contradictory? No -- *since the sample sizes are different, the p-values are not comparable, even though everything else is the same.*

Indeed, *a larger sample size typically results in a smaller p-value*.

> The idea of why this is true is illustrated by the case of the z-test, since large n gives a smaller standard deviation of the sampling distribution, hence a narrower sampling distribution.

Comparing p-values for samples of different size is a **common mistake**.

Example B: Researcher 2 from Example A does everything as described above, but for convenience, his patients are all from the student health center of the prestigious university where he works.

- He *cannot* claim that his result applies to patients other than those of the age and socio-economic background, etc. of the ones he used in the study, because his sample was taken from a smaller *population*.

Example C: Researcher 2 proceeds as in Example A, with a sample carefully selected from the population to which he wishes to apply his results, but he is testing for equality of the means of an outcome variable for the two groups.

- The hypothesis test he uses requires that the variance of the outcome variable for each group compared is the same.
- He doesn't check this, and in fact the variance for the treatment group is twenty times as large as the variance for the placebo group.
- He is *not* justified in rejecting the null hypothesis of equal means, no matter how small his p-value (unless by some miracle the statistical test used is robust to such large departures from the model assumption of equality of variances.)

Another **common misunderstanding** *of p-values is the belief that the p-value is "the probability that the null hypothesis is true".*

- This is essentially a case of confusing a conditional probability with the reverse conditional probability: In the definition of p-value, “the null hypothesis is true” is the condition, not the event.
- The basic assumption of frequentist hypothesis testing is that the null hypothesis is either true (in which case the probability that it is true is 1) or false (in which case the probability that it is true is 0) – so unless p = 0 or 1, the p-value couldn't possibly be the probability that the null hypothesis is true.

Note: In the *Bayesian* perspective, it makes sense to consider "the probability that the null hypothesis is true" as having values other than 0 or 1.

- In that perspective, we consider "states of nature;" in different states of nature, the null hypothesis may have different probabilities of being true.
- The goal is then to determine the probability that the null hypothesis is true, given the data: P(H_0 true | data)
- This is essentially the *reverse conditional probability* from the one considered in frequentist inference (the probability of the data given that the null hypothesis is true – P(data | H_0 true).

VII: TYPE I ERROR AND SIGNIFICANCE LEVEL

Type I Error:

Rejecting the *null* hypothesis when it is in fact true is called a ***Type I error***.

Significance level:

Many people decide, before doing a hypothesis test, on a maximum p-value for which they will reject the null hypothesis. This value is often denoted α (alpha) and is also called the ***significance level***.

When a hypothesis test results in a p-value that is less than the significance level, the result of the hypothesis test is called *statistically significant*.

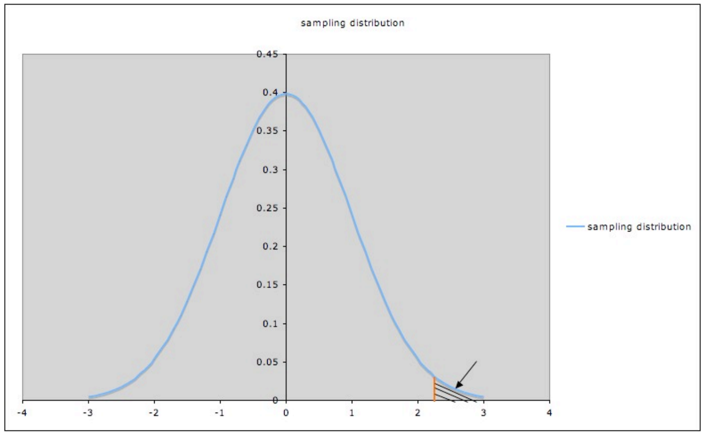
Confusing statistical significance and practical significance is a **common mistake**.

> *Example*: A large clinical trial is carried out to compare a new medical treatment with a standard one. The statistical analysis shows a statistically significant difference in lifespan when using the new treatment compared to the old one.
>
> - However, the increase in lifespan is at most three days, with average increase less than 24 hours, and with poor quality of life during the period of extended life.
> - Most people would not consider the improvement practically significant.

> *Caution*: The larger the sample size, the more likely a hypothesis test will detect a small difference. Thus *it is especially important to consider practical significance when sample size is large.*

Connection between Type I error and significance level:

A significance level α corresponds to a certain value of the test statistic, say t_α, represented by the orange line in the picture of a sampling distribution below (the picture illustrates a hypothesis test with alternate hypothesis "$\mu > 0$").



- Since the shaded area indicated by the arrow is the p-value corresponding to t_α, that p-value (shaded area) is α.
- To have p-value less than α, a t-value for this test must be to the right of t_α.
- So the probability of rejecting the null hypothesis when it is true is the probability that $t > t_\alpha$, which we have seen is α.
- In other words, *the probability of Type I error is α.*
- Rephrasing using the definition of Type I error:
  The significance level α is the probability of making the wrong decision when the null hypothesis is true.

Note:

- α is also called the *bound on Type I error*.
- Choosing a significance level α is sometimes called *setting a bound on Type I error*.

Common mistake: Claiming that an alternate hypothesis has been "proved" because it has been rejected in a hypothesis test.

- This is one instance of the mistake of "expecting too much certainty" discussed Monday.
- There is always a possibility of a Type I error; the sample in the study might have been one of the small percentage of samples giving an unusually extreme test statistic.
- This is why *replicating studies* (i.e., repeating the analysis with another sample) is important. The more (carefully done) studies that give the same result, the stronger the overall evidence.
- There is also the possibility that the sample is biased or the method of analysis was inappropriate; either of these could lead to a misleading result.

VIII: PROS AND CONS

OF SETTING A SIGNIFICANCE LEVEL

- Setting a significance level (*before* doing inference) has the *advantage* that the analyst is not tempted to chose a cut-off on the basis of what he or she hopes is true.
- It has the *disadvantage* that it neglects that some p-values might best be considered borderline.
    - *This is one reason why it is important to report p-values when reporting results of hypothesis tests.*
    - *It is also good practice to include confidence intervals corresponding to the hypothesis test.*
        - For example, if a hypothesis test for the difference of two means is performed, *also* give a confidence interval for the difference of those means.
        - If the significance level for the hypothesis test is .05, then use confidence level 95% for the confidence interval.