Tests of hypothesis

Hypothesis – a contention based on preliminary evidence of what appears to be fact (an educated guess), which may or may not be true.

- Formulating a hypothesis is the second step in the scientific method.
- A statement of the hypothesis is the first step in experimentation.

Test of hypothesis – a comparison of the contention with a set of newly gathered data.

Hypothesis testing procedure – we will consider 7 steps

I. Set up a meaningful hypothesis such as "The population mean is equal to some value" (call it μ_0)

H_0: $\mu = \mu_0$ or $\mu - \mu_0 = 0$

This is called the null hypothesis. It is a hypothesis of equality or of no difference (even if you believe there is a difference). Note that hypotheses are always stated in terms of the population parameters, not the sample statistics we actually measure, because we are drawing inference about the population.

II. Set up an alternative hypothesis

Alternative hypotheses are denoted H_1 or H_a. This hypothesis states what is correct if the null hypothesis in not correct. This is usually the case of actual interest.

Examples:

a) H_1: $\mu \neq \mu_0$ or H_1: $\mu - \mu_0 \neq 0$ (also called the non-directional alternative)

b) H_1: $\mu < \mu_0$ or H_1: $\mu - \mu_0 < 0$

c) H_1: $\mu > \mu_0$ or H_1: $\mu - \mu_0 > 0$

III. Consider the assumptions.

1) We will be using the Z distribution, so the distribution we are testing must be normal.

2) The observations should be independent. The best guarantee of independent observations is random sampling.

3) Strictly speaking, the variance should be known in order to use the Z distribution; however it is often used for very large samples. Later we will discuss the t-distribution that is used when the variance is not known and must be estimated from the sample.

There will be a few other assumptions for other test statistics. However, the tests of hypothesis we will be using are also "robust". Statistically speaking, robustness indicates that the test performs quite well even if the assumptions are not perfectly met.

IV. Select a probability of rejecting the null hypothesis (H_0) when it is true. This is called the alpha (α) value and the value chosen is somewhat arbitrary. By convention the values usually chosen is $\alpha = 0.05$ or sometimes $\alpha = 0.01$.

for $\alpha = 0.05$ then if H_0 is true we will reject it 5% of the time, or in one of 20 samples

for $\alpha = 0.01$ then if H_0 is true we will reject it 1% of the time, or in one of 100 samples

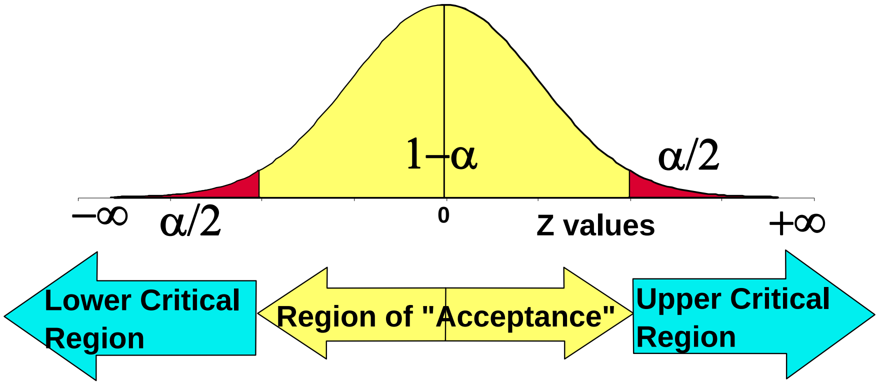
This value is sometimes called the significance level.

From this value, and the alternate hypothesis, we can determine the critical limits, those values of the test statistic that would cause us to reject the null hypothesis.

Determine a critical region, what is too large or too small by using the chosen probability or significance level.

Critical region – the area in the distribution which would lead to rejection of the null hypothesis (H_0:). When we reject we know that it is possible that the null hypothesis is true, but if it is we would only reject $\alpha*100\%$ of the time. So this type of error can be controlled.

Region of "acceptance" – the area under the distribution which would lead to "acceptance" of the null hypothesis (H_0:).



Notice that I have placed the word "acceptance" in quotes. We cannot really state that we "accept" the null hypothesis because it is also possible that we would be wrong in doing so. Unfortunately, in practice the probability of this type of error is unknown and therefore one cannot "accept" with a known probability of error (more later under Type II error and Power).

V. Draw a sample from the population of interest (as defined by the investigator), and

a) Compute an estimate of the parameter in the hypothesis; in our example the hypothesis was about μ so the statistic will be $\bar{Y}$, recall E($\bar{Y}$) = μ.

b) The value of $\bar{Y}$ from the sample now becomes one of many possible observations from the derived population of all sample means.

c) Recall that the derived population has,

$$\mu_{\bar{Y}} = \mu$$

$$\sigma^2_{\bar{Y}} = \sigma^2/n$$

$$\sigma_{\bar{Y}} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$$

d) Recall that the distribution of sample means ($\bar{Y}_k$) approaches a normal distribution as the value of n increases (according to the Central Limit Theorem). This helps meet our assumption of normality.

e) Recall the Z transformation $Z_i = \frac{\left(\overline{Y}_i - \mu_{\overline{Y}}\right)}{\sigma_{\overline{Y}}}$.

Our null hypothesis contends that the true value of $\mu_{\overline{Y}}$ is our hypothesized value, μ_0, so we will calculate a Z score using μ_0. This will follow a Z distribution **if the null hypothesis is correct**. If the null hypothesis is not correct we don't care what the distribution is, we just hope to reject the null hypothesis.

$$Z_i = \frac{\left(\overline{Y}_i - \mu_0\right)}{\sigma_{\overline{Y}}}$$

As a result, where μ_0 is the hypothesized value of the mean, if the null hypothesis is true and $\mu = \mu_0$ we would expect Z to be approximately zero (within reasonable limits, defined later). On the other hand, if $\mu \neq \mu_0$ we would expect Z to be different from zero by some amount. If Z is too much greater than zero (i.e. Z > 0), that suggests that $\mu_{\overline{Y}}$ is too large while if Z is much less than zero, then $\mu_{\overline{Y}}$ appears to be too small.

VI. Compare the test statistic from step V to the critical region determined in step IV.

VII. Draw conclusions and interpretations from the results of the test. The test statistic is not an end in itself.

Logic behind the test

A key aspect of a test of hypothesis is that we must have a test statistic with a known distribution.

We could sample from any one of numerous populations with many different distributions. The characteristics of these distributions are unknown, but if we can transform the sampled distribution to a known distribution, we can then make some probability statements.

Beyond this, we simply want to determine what is likely under the null hypothesis. If we hypothesize a mean of μ_0 and take a sample of mean that is actually close to μ_0, then the null hypothesis is probably true. If, on the other hand, the calculated sample mean is not close to μ_0, and if the difference big enough that it is not likely to have occurred due to sampling variation, then the alternate hypothesis is the more likely choice.

Reasonable limits – recall that we needed to define this

a set of limits of the critical region determined by the significance level (α) and by the alternative hypothesis (e.g. was it two tailed, or one tailed, and if one tailed, to which side). The value of α is what specifies what we feel would be unlikely under the null hypothesis.

SUMMARY OF THE 7 STEPS OF HYPOTHESIS TESTING

I. Establish a null hypothesis, H_0:

II. Determine an appropriate alternative hypothesis to the null, H_1:

III. Consider the assumptions

IV. Determine a value for α and find the critical limits and a critical region for the chosen statistic.

V. Obtain a sample of new data to test the hypothesis, and compute the appropriate test statistic from the sample.

VI. Using the critical region and the test statistic (e.g. Z), compare the values and make a decision to reject the H_0 or to fail to reject the H_0.

VII. Draw your conclusions from the test of hypothesis.

Example of a Test of Hypothesis

Extensive measurements done in eastern Tennessee have shown that the average 20-year-old White Oak produces an average of 12 Kg of acorns with a variance of 4 Kg^2. Five White Oaks in Georgia produced a mean of 14 Kg. Assuming that the variance is the same, test the hypothesis that the production is the same in Tennessee and Georgia.

1) $H_0\!: \mu = \mu_0$ or $H_0\!: \mu - \mu_0 = 0$ (where μ_0 = 12 Kg, the known value for Tennessee.

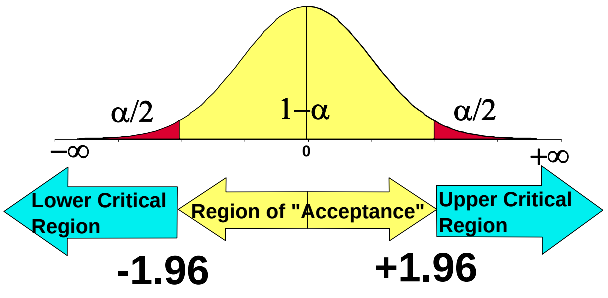
2) $H_1\!: \mu \neq \mu_0$ or $H_1\!: \mu - \mu_0 \neq 0$. We might be tempted to test the hypothesis, $H_1\!: \mu > \mu_0$, since the Georgia oaks had a mean of 14 Kg. However, remember that this is supposed to be a new data set to test the hypothesis and we would not have known this in advance.

3) Assume the sample of Oaks is random (independent) and normally distributed. We also have a known variance from Tennessee of 4 Kg^2.

4) Determine a value of α and obtain the critical limits for a critical region for the test statistic using our knowledge of H_1 and α.

We will somewhat arbitrarily choose a value of α = 0.05. This is a commonly used and accepted value.

The H_1 indicates that we are doing a 2 tailed test. To keep α at 0.05, place half the value of a in each tail (0.0250 per tail). This corresponds to critical Z values of ±1.96



The Critical Region: red areas in the tails are areas of rejection.

5) Obtain a new data set to test the hypothesis, and compute the appropriate test statistic from the sample ($\overline{Y}$ for testing differences in the means).

The results for our sample were $\overline{Y}$ = 14, and n = 5.

6) Calculate $Z = \frac{\left(\bar{Y}_i - \mu_0\right)}{\sigma_{\bar{Y}}} = \frac{\left(14-12\right)}{\sqrt{4/5}} = \frac{\left(2\right)}{2/\sqrt{5}} = \frac{\sqrt{5}\left(2\right)}{2} = \sqrt{5} = 2.236.$

This value of the test statistic is greater than the limit for the upper critical region (±1.96), so it falls in the region of rejection. This would be interpreted to indicate that it is unlikely that a value this large would arise by random chance alone if the null hypothesis were true.

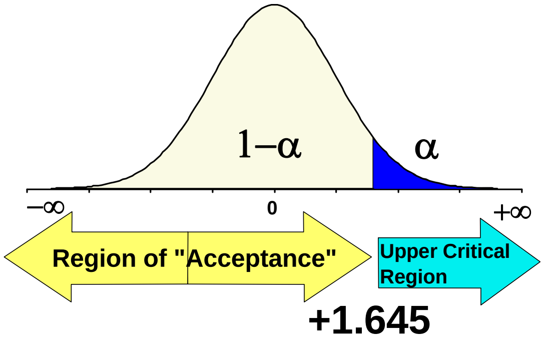
7) Conclusion: Reject the null hypothesis and conclude that the two areas differ in terms of acorn production.

We can also go one step further. Since the production levels are different we can also conclude that production is greater in Georgia since it had a greater value for the mean production.

One-tailed tests

Suppose our problem had been a little different, and that we had believed from the beginning that Georgia had a higher rate of production. Something we believed BEFORE we started the study (*a priori*). We might then want to test for only this alternative, i.e. H_1: $\mu > \mu_0$ or H_1: $\mu - \mu_0 > 0$, where μ_0 = 12 Kg (the Tennessee value).

Now, the test is altered because we will have a different critical value. We still want an α = 0.05, but we would put all 5% chance of error in the upper tail!



Note that this makes it "easier" to show significance, because we only need meet the 1.645 criteria instead of the 1.96 criteria.

However, it implies that we have some additional knowledge and have no interest in the lower tail. What if the calculated value was well into in the lower tail? Presumably this would be a spurious occurrence and not of interest, because we "know" it can't happen.

In fact, if our critical value was 1.645 in the upper tail, and we found the Georgia value to be less than Tennessee, no additional calculations would be needed because the calculated Z value would be negative and could not be in the upper tail. In other words, if our hypothesized value (μ_0) is greater than out observed value ($\bar{Y}$), then the calculation $Z = \frac{\left(\bar{Y}_i - \mu_0\right)}{\sigma_{\bar{Y}}}$ would be negative and could not be in the upper tail that was hypothesized.

Additional notes and terminology on hypothesis testing

Recall that a key aspect of a test of hypothesis is that we must have a test statistic with a known distribution. For our present discussion we are using the Z distribution.

Given $H_0: \mu = \mu_0$ and $H_1: \mu \neq \mu_0$ and $\alpha = 0.05$

1) If H_0 is true then $(1–\alpha)100\%$ of the samples will yield a Z test statistic that will fall in the region of "acceptance". That is, for $\alpha = 0.05$, then $(1–\alpha)100\% = 95\%$. This is sometimes referred to as the confidence level.

2) For a two-tailed test, half of the possible samples will have a Z test statistic score in the upper critical region $[(\alpha/2)100\%]$, and half of the samples will have a score in the lower critical region $[(\alpha/2)100\%]$. For $\alpha = 0.05$, then $(\alpha/2)100\% = 2.5\%$.

3) Since $H_1: \mu \neq \mu_0$ (implying we do not know "*a priori*" if the hypothesized value might be too large or too small), the probability statement then becomes.

$P(|Z| \geq Z_0) = \alpha = 0.05$ (the absolute value sign indicates Z may be positive or negative)

$2P(Z \geq Z_0) = \alpha = 0.05$

$P(Z \geq Z_0) = \alpha /2 = 0.025$

so $Z_0 = 1.96$ from the Z tables

4) If the calculated Z test statistic is between –1.96 and +1.96, we cannot reject the null hypothesis ($H_0: \mu = \mu_0$). This means that the observed statistic is consistent with the hypothesized value, BUT we can never actually PROVE that H_0 is true. It is relatively easy to prove that things are different, but almost impossible to prove that two things are identical.

So we resort to jargon; we say that ...

- there is no "statistically significant difference"
- there is no "significant difference"
- that "the data is consistent with the null hypothesis"
- that we "fail to reject the null hypothesis".

These statements are better (more correct) than stating that we actually "ACCEPT" the null hypothesis or that the null hypothesis is TRUE.

5) For a two tailed test, if the calculated |Z| is greater than, or equal to, the critical value of the test statistic (e.g. Z = 1.96), then reject the H_0, and conclude that the null hypothesis is correct. For one tailed tests, if Z > 1.96 reject for the hypothesis $H_1: \mu > \mu_0$ and if or Z < –1.96 conclude that $H_1: \mu < \mu_0$.

6) The size of the critical region is determined by α, the level of significance.

Note that when we reject the null hypothesis, there is a chance that we are in error, but that we know the probability of making that error. It is α. This is because we can set the level of fallibility in our conclusions for this type of error.

7) When we have a one tailed alternative, say $H_1: \mu > \mu_0$, versus the null hypothesis $H_0: \mu = \mu_0$, what happens to the cases that may be much less than the hypothesized

values? Since we have a one tailed test we must know that such cases are impossible, or are simply not of interest no matter how small they must be. In this case some investigators prefer notation where the other extreme is included in the null hypothesis.

$H_0: \mu \le \mu_0$ versus $H_1: \mu > \mu_0$

$H_0: \mu \ge \mu_0$ versus $H_1: \mu < \mu_0$

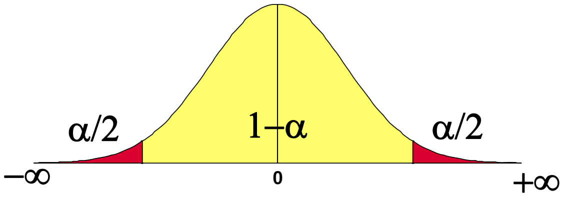
This is acceptable, but the statistical development of a test of hypothesis actually considers only the equality in the null hypothesis and doesn't really consider these cases.

Final notes on the one and two tailed alternatives

1) The two tailed test is called the "non directional alternative".

$H_0: \mu = \mu_0$ or $H_0: \mu - \mu_0 = 0$

$H_1: \mu \ne \mu_0$ or $H_1: \mu - \mu_0 \ne 0$



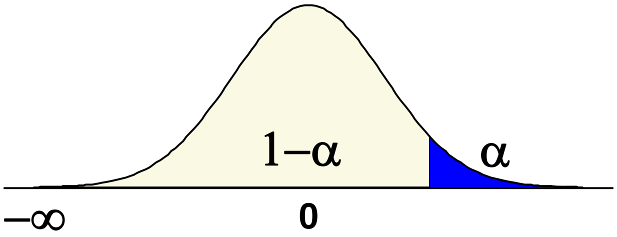
This means that we will accept either $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$ as fulfillment of the alternate hypothesis. Since either case is to be accepted we state our probability with an absolute value, $P(|Z| \ge Z_0) = \alpha = 0.05$ and for a 5% chance of error, we divide the 5% into equal parts (usually) and put half in each tail.

2) The one tailed test is called the "directional alternative".

$H_0: \mu = \mu_0$ or $H_0: \mu - \mu_0 = 0$ and

$H_1: \mu < \mu_0$ or $H_1: \mu - \mu_0 < 0$ or

$H_1: \mu > \mu_0$ or $H_1: \mu - \mu_0 > 0$



this indicates that we will accept only one of the two options, $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$ as fulfillment of the alternate hypothesis. Since only one case is to be accepted, we state our probability as either $P(Z \ge Z_0) = \alpha = 0.05$ or $P(Z \le Z_0) = \alpha = 0.05$, and for a 5% chance of error, we put all 5% into the tail of interest.

Why $\alpha = 0.05$, and not 0.04 or 0.09?

No particular reason. The value is not special, but has become something of a convention or traditional standard. This value represents a one chance in 20 of error. It is generally accepted as a reasonable chance of error, and is usually acceptable to referees, editors and journals. However, if you want to use another value, and have some **good** reason for doing so, this should be possible.

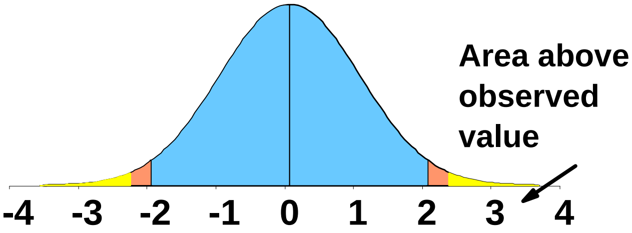
The value of 0.05 has traditionally been termed the level at which we have "statistically significant" results. A value of 0.01 is then considered a "highly significant" result.

P values in tests of hypothesis

Probability value or P values, like those we have discussed previously, just represent some area under a curve. However, in the context of hypothesis testing they indicate that area under the curve that represents a value equal or larger than some observed value of a test statistic.

Recent literature had tended to giving just the actual "P value", and letting the reader decide if it is "significant". The P-value is just the area in the tail above the calculated Z value. For example, with our Oak tree example, the calculated Z value was 2.236. This was larger than our critical value of 1.96. so the "tail" would be smaller than 0.025.

So, how unusual is a value of 2.236?
Actually, the probability of a randomly chosen value exceeding this value is 0.0127 in one tail. For a two tailed tests we would express this probability as 2(0.0127) = 0.0254 since we would reject for either – 2.236 or +2.236.



The P-value is then: P = 0.0254. For most tests that we do, SAS will give this value.

If smaller than the desired α, calculated test statistic value would be in the tail and would be rejected. If larger than the desired α, test statistic value would not be in the tail and would be not be rejected. Most tests in SAS are two–tailed, though a few are one-tailed.

Another Example

The mean for high school seniors on a nationally standardized reading test is 170 points with a variance of 400. The principal of a small rural high school hypothesizes that the 9 seniors in his school will score better than the national average. Test his hypothesis (data given later).

I. H_0: $\mu = \mu_0$ or H_0: $\mu - \mu_0 = 0$

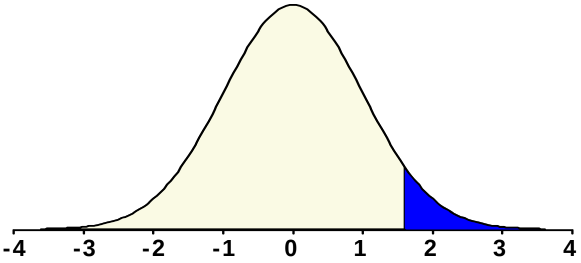
II. H_1: $\mu > \mu_0$ or H_1: $\mu - \mu_0 > 0$

III. Assume that the scores are (1) normally and (2) independently distributed with a (3) known variance of $\sigma^2 = 400$. (i.e. the distribution is NID(170, 400)).

IV. Let the probability of Type I error equal 5%. (i.e. α = 0.05)

V. Find the critical limits given that we want a one tailed test against the upper tail with α = 0.05. The Z value which will leave 5% in the upper tail is 1.645.

VI. Gather new data to test the hypothesis. The test results for the 9 students were: 164, 175, 186, 173, 191, 187, 189, 176 and 179. The summary statistics for this group are $\bar{Y} = 180$ and $S^2 = 634$. However, we know the true national variance ($\sigma^2 = 400$) for the test and can use this in a Z test.



The condition of "known variance" is really important to using a Z test, and should be added as a third assumption.

The test calculations are $Z = \dfrac{\bar{Y} - \mu_0}{\sqrt{\sigma^2 / n}} = \dfrac{180 - 170}{\sqrt{400/9}} = \dfrac{10}{6.6667} = 1.5$

VII. This value does not reach the critical value of 1.645, so we cannot conclude that these 9 seniors scored significantly higher than the national average. Apparently, it is not that unusual, at the 5% level, for any subgroup of 9 individuals to score 10 points above the national mean. However, the P value for the observed Z score is P = 0.0668, so it is not very common either.

Are we convinced that these 9 students are not above average? This would be our conclusion if the P value had reached 0.05, but it reached only 0.0668. Close! The principal may well claim that this was significant. As scientists we may decide it is just too close to call, and "reserve judgment" pending more data.

Summary

Logic: We need a known probability distribution and we need to determine what is likely for our known distribution under the null hypothesis.

Any conditions needed for this to work out are specified in the assumptions.

Both one and two-tailed alternative hypotheses are possible.

Review the 7 Steps of Hypothesis testing

I. Determine the H_0

II. Determine the H_1

III. Consider the assumptions

IV. Determine a value for α and obtain a critical region for a test statistic (e.g. Z), from your knowledge of alpha (α) and the H_1.

V. Obtain a sample of new data to test the Hypothesis. Compute the appropriate statistic from a sample (e.g. $\overline{Y}$) and calculate the value of the TEST STATISTIC (Z)

VI. Compare the calculated value of the test statistic to the CRITICAL VALUES. Make your decision to either reject the H_0 or to FAIL to Reject the H_0.

VII. Draw you conclusions from the test of Hypothesis and interpret your results.

The 5 steps of Hypothesis Testing according to Freund & Wilson.

1) Establish H_0 , H_1 and a value for α.

2) Determine the test statistic and a region for rejection

3) Draw a sample, calculate the test statistic

4) Compare the test statistic to the critical limits and make a decision to reject or fail to reject.

5) Interpret the results

Hypothesis testing Concepts

The logic of test of hypothesis is based on the chosen probability of error, α (or significance level) for the test statistic (Z) which determines the range of what would be expected due to chance alone assuming H_0 is true.

Significance level notation, commonly used levels and terminology

"Statistically significant"	$\alpha = 0.05$
"Highly significant"	$\alpha = 0.01$

Errors!

When we do a test of hypothesis is it possible that we are wrong?

Yes, unfortunately, it is always possible that we are wrong. Furthermore, there are two types of error that we could make!

Types of error

	Data indicates: H_0 is true	Data indicates: H_0 is false
True result: H_0 is true	NO ERROR	**Type I Error: Reject TRUE H_0**
True result: H_0 is false	**Type II Error: Fail to Reject FALSE H_0**	NO ERROR

Type I Error: Type I error is the rejection of a true null hypothesis.

This type of error is also called α (alpha) error. This is the value that we choose as the "level of significance", so we actually set the probability of making this type of error.

The probability of a type I error = α

Type II Error: Type II error is the failure to reject of a null hypothesis that is false. This type of error is also called β (beta) error.

We do not set this value, but we call the probability of a type II error = β.

Furthermore, in practice we will never know this value. This is another reason we cannot "accept" the null hypothesis, because it is possible that we are wrong and we cannot state the probability of this type of error.

The good news, it is only possible to make one error at a time.

If you reject H_0, then you may have made a type I error, but you cannot have made a type II error.

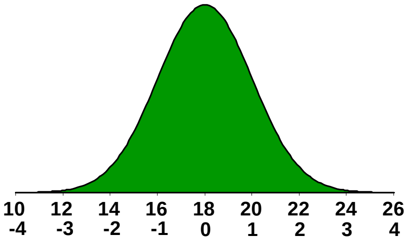
If you fail to reject H_0, then you may have made a type II error, but you cannot have made a type I error.

The probability of Type II Error

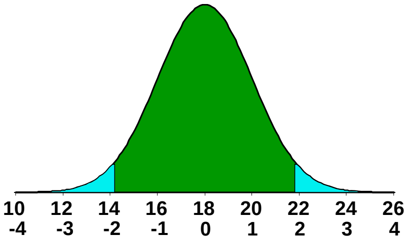
This is a probability that we will not know. This probability is called β. However, we can do several things to make the error smaller, so this will be our objective.

First, let's look at how these errors occur.

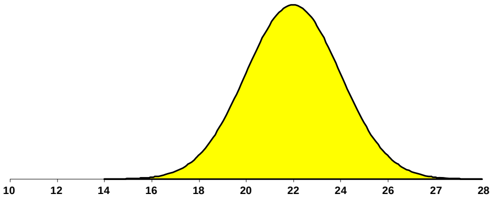
Examine the hypothesized distribution (below) that we believe to have a mean of 18.



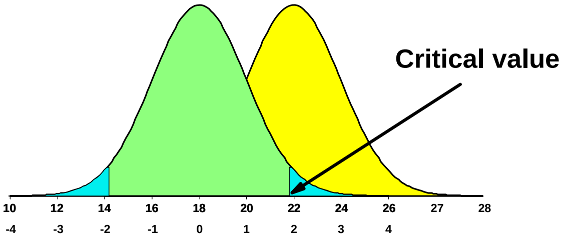
We are going to do a 2 tailed test with an α value of 0.05. Our critical limits will be ±1.96.



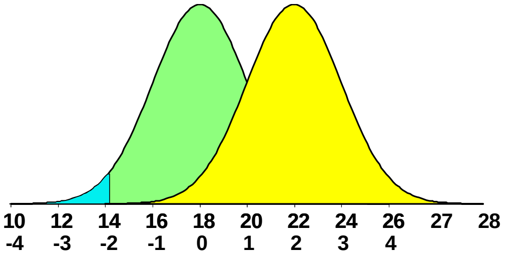
So we will reject any test statistic over 1.96 (or under –1.96). But let's suppose the null hypothesis is false!!! Let's suppose that the alternate hypothesis is true. Then the hypothesized distribution is not real, there is another "real" distribution that we are sampling from. What might it look like?



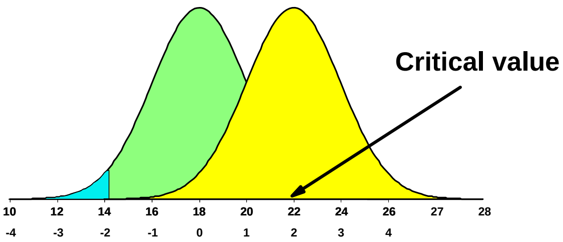
Here is the real distribution. It actually has a mean of 22, but we don't know that. If we did, we would not have hypothesized a mean of 18!



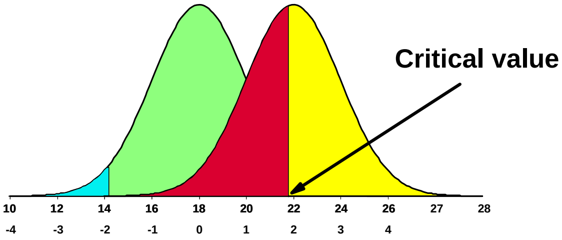
So where on the real distribution is our critical limit. This is the key question.



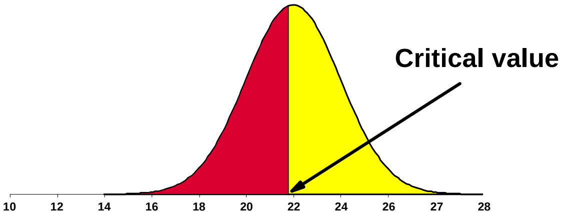
Note that with the Z transformations each change of 1 unit of Z corresponds to a change of 2 on the original *Y* scale. This means that on the original scale $\sigma = 2$.



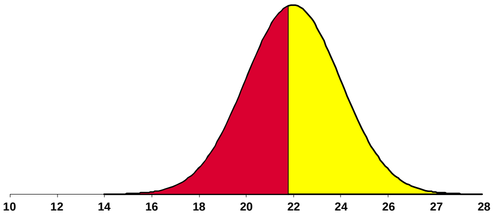
Now we draw our sample from the real distribution. If our result says reject the H_0, we make no error.



But if our result causes us to fail to reject, we err.



And in this case it appears we have pretty close to a 50-50 chance of going either way.



So we take our sample and do our test. Will we err? Maybe we will, and maybe we won't. Our sample could come from anywhere in this "real" distribution. If our sample happens to fall in the lower red area (below about 22), we would not reject H_0, and we would err. But if our sample happens to fall in the upper yellow area (above about 22), we will reject H_0. In this case there is no error, we draw the correct conclusion.

The Probability of Type II error or β error

For $\alpha = 0.05$, our critical limit, in terms of Z, would be 1.96.

The critical limit translates to a value on the original scale of $Y_i = \mu + Z_i\sigma = 18 \pm 1.96(2) = 18 \pm 3.92$. The lower bound is 14.08 and the upper bound is 21.92. The lower bound is so far down on the real distribution that the probability of getting a sample that falls there is near zero. The upper bound is the one that falls in the middle of the "real" distribution.

In this fictitious case we know that the true mean is 22. Normally we wouldn't know the true mean. Since we know the true mean in this case, we can calculate the probability of drawing a sample above and below the critical limit (21.92 on the *Y* scale, –0.04 on the Z scale of the real distribution). The probability of falling below this value, and of making a type II error, is 0.484, or about 48.4%. This is the probability we call beta (β).

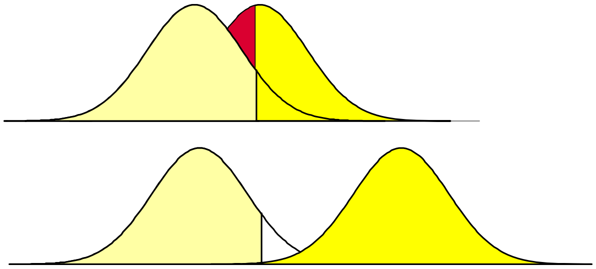
The probability of falling above this value, and of NOT making a type II error, is 0.516, or 51.6%. So in this case we can calculate β, the probability of a Type II error. In practice we cannot usually know these probabilities because we never know the real value of the mean.

We define a new term POWER, this is the probability of NOT making a type II error ($1 - \beta$). This was 0.516 in our example.

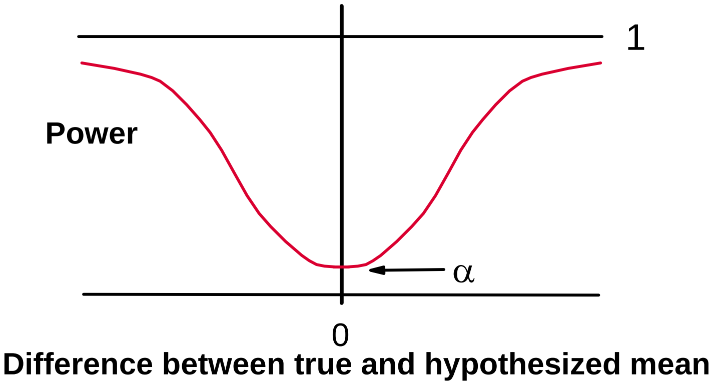
Power and Type II Error

Since we don't actually know the value of the true mean (or we wouldn't be hypothesizing something else), we cannot know in practice the type II error rate (β). However, it is affected by a number of things, and we can know about these.

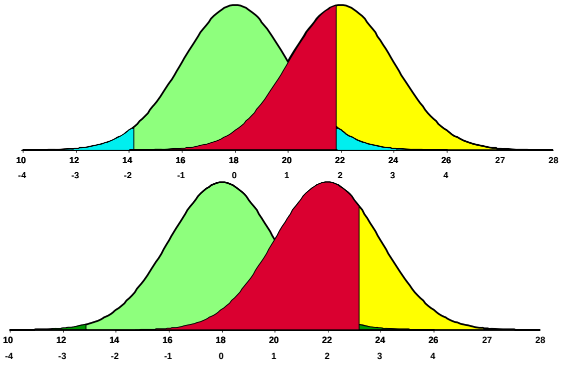
1) Power is affected by the distance between the hypothesized mean (μ_0) and true mean (μ).



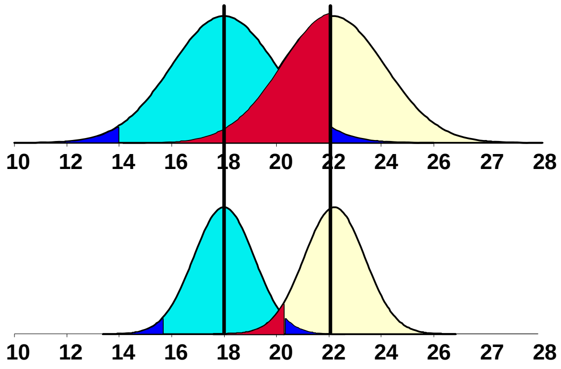
The Power Curve



2) Power is affected by the value chosen for Type I error (α).



3) Power is affected by the variability or spread of the distribution.



Influencing the power of a test of hypothesis

The capability of the test to reject H_0 when it is false is called Power = $1 - \beta$. Anything done to enhance this value will improve your ability to test for differences among populations. Which of the 3 factors influencing power can you control?

For testing means you may be able to control sample size (n). This reduces the variability and increases power.

You probably cannot influence the difference between μ and μ_0.

You can choose any value of α. However, this cannot be too small or Type II error becomes more likely. Too large and Type I error becomes likely.

Methods of increasing the power of a test

How would we use our knowledge of factors affecting power to increase the power of our tests of hypothesis?

Increase the significance level (e.g. from $\alpha = 0.01$ to $\alpha = 0.05$)

If H_0 is true we would increase α, the probability of a Type I error.

If H_0 is false then we decrease β, the probability of a Type II error, and by decreasing β, we are increasing the POWER of test.

For a given α, the POWER can be increased by ....

Increasing n, so $\sigma_{\bar{Y}} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$ decreases, and the amount of overlap between the real and hypothesized distributions decreases.

For example, let's suppose we are conducting a test of the hypothesis H_0: $\mu = \mu_0$ against the alternative H_1: $\mu \neq \mu_0$. We believe $\mu_0 = 50$ and we set $\alpha = 0.05$. We also know that $\sigma^2 = 100$ and that n = 25.

From this information we can calculate $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 10/5 = 2$. The critical region in terms of Z is then $P(|Z| \geq Z_0) = 0.05$ and $Z_0 = 1.96$, and the critical value on the original scale Y variable scale is $Y_i = \mu + Z_i\sigma = 50 + 1.96(2) = 53.92$.

If the REAL population mean is 54, calculate $P(Y \geq 53.92)$, given that the TRUE mean is 54 we calculate the Z value as $Z = (53.92 - 54)/2 = -0.08 / 2 = -0.04$.

The probability of a TYPE II error (β) is the probability of not drawing a sample that falls above this value and not rejecting the false null hypothesis. The value is β = P(Z ≤ –0.04) = 0.4840.

So for an experiment with n = 25, the power is 1 – β = 1 – 0.4840 = 0.516.

But suppose we had a larger sample, say n = 100. Now $\sigma_{\bar{Y}} = \sigma / \sqrt{n} = 10/10 = 1$. The critical region stays at Z_0 = 1.96, but on the original scale this is now $Y_i = \mu + Z_i\sigma$ = 50 + 1.96(1) = 51.96. For a true mean of 54 we now get Z = (51.96–54)/1 = –2.04/1 = – 2.04.

The value of β = is P(Z ≤ –2.04) = 0.0207, and the power for this test is 1 – β = 0.9793.

The bottom line,

With n = 25, the power is 0.5160.

With n = 100, the power is 0.9793.

This is why statisticians recommend larger sample sizes so strongly. We may never really know what power is, but we know how to increase it and reduce the probability of TYPE II error.

Summary

Hypothesis testing is prone to two types of errors, one we control (α) and one we do not (β).

Type I error is the REJECTION of a true null hypothesis.

Type II error is the FAILURE TO REJECT a null hypothesis that is false.

The "Power" of a test is 1 – β

Not only do we not control TYPE II error, we probably do not even know its value. However, we can hopefully reduce this error, and increase power, by

Controlling the distance between μ and μ_0 (not really likely)

Selecting a value of α that is not too small (0.05 and 0.01 are the usual values)

Getting a larger sample size (n), this is the factor that is usually under the most control of the investigator.

The t-test of hypotheses

The t distribution is used the same as Z distribution, except it is used where sigma (σ) ,is unknown (or where $\bar{Y}$ is used instead of μ to calculate deviations). The t distribution is a bell shaped curve, like the Z distribution, but not the same. The Z distribution is normal because it has a normal distribution in the numerator (Y_i) and all other terms in the transformation are constant. The t distribution has a normal distribution in the numerator but the sample variance in the denominator is another statistic with a chi square distribution.

$t_i = \frac{\left(Y_i - \bar{Y}\right)}{S}$; the t distribution applied to individual observations