STAT 515 fa 2023 Lec 14

Hypothesis testing

Karl Gregory

Hypothesis testing

A *statistical inference* is a conclusion about a population based on a random sample. More specifically, given a statement about the population, a statistical inference is a decision to reject or not to reject the statement in light of the sample data. In the language of statistical inference, the statement subject to our rejection or non-rejection is called the *null hypothesis*. The statement which conveys the opposite about the population is called the *alternate hypothesis*. We typically denote the null and alternate hypotheses as H_0 and H_1. We read H_0 as "H nought."

For example, suppose X is a random variable with unknown mean μ. It could represent a draw from a population with an unknown mean μ. A null hypothesis could be

$$H_0\text{: } \mu = 5,$$

to which the alternate hypothesis would be

$$H_0\text{: } \mu \neq 5.$$

To reach a decision regarding H_0 and H_1, we consider the evidence contained in a random sample $X_1, \ldots, X_n$ of X values. Our statistical inference, that is our decision regarding H_0 and H_1, is always one of the following:

1. "We reject H_0 and therefore conclude that H_1 is true."
2. "We do not reject H_0."

For the second case people also like to say, "We fail to reject H_0." It is important to note that if we do not reject H_0, we do *not* conclude that H_0 is true. We simply say, "We do not reject H_0," or, "Our data do not lead us to reject H_0."

Deciding from the sample data whether to reject or not to reject the null hypothesis is referred to as *testing* the null hypothesis. In order to test the null hypothesis, we will compute from our random sample a quantity called a *test statistic*. From the test statistic we will judge the plausibility of the null hypothesis in light of the data.

Forms of the null and alternate hypothesis for μ and p

The null and alternate hypotheses are mathematical expressions involving population parameters. As a convention, the null hypothesis H_0 always contains an equality. In this course, for statistical inference about the mean μ, the hypotheses H_0 and H_1 will always be one of the following:

$$\begin{array}{lll} H_0\colon\ \mu \geq \mu_0 & \text{or} & H_0\colon\ \mu = \mu_0 \quad \text{or} \quad H_0\colon\ \mu \leq \mu_0 \\ H_1\colon\ \mu < \mu_0 & & H_1\colon\ \mu \neq \mu_0 \qquad\quad H_1\colon\ \mu > \mu_0, \end{array}$$

where μ_0 is a specific value of the unknown parameter μ used to define the null hypothesis. Sometimes we call μ_0 the *null value* of μ. Note that the book will write

$$\begin{array}{lll} H_0\colon\ \mu = \mu_0 & \text{or} & H_0\colon\ \mu = \mu_0 \quad \text{or} \quad H_0\colon\ \mu = \mu_0 \\ H_1\colon\ \mu < \mu_0 & & H_1\colon\ \mu \neq \mu_0 \qquad\quad H_1\colon\ \mu > \mu_0. \end{array}$$

For statistical inference about a proportion p, the hypotheses H_0 and H_1 in this course will always be one of the following:

$$\begin{array}{lll} H_0\colon\ p \geq p_0 & \text{or} & H_0\colon\ p = p_0 \quad \text{or} \quad H_0\colon\ p \leq p_0 \\ H_1\colon\ p < p_0 & & H_1\colon\ p \neq p_0 \qquad\quad H_1\colon\ p > p_0, \end{array}$$

where p_0 is a specific value of the unknown parameter p used to define the null hypothesis. Sometimes we call p_0 the *null value* of p. Note that the book will write

$$\begin{array}{lll} H_0\colon\ p = p_0 & \text{or} & H_0\colon\ p = p_0 \quad \text{or} \quad H_0\colon\ p = p_0 \\ H_1\colon\ p < p_0 & & H_1\colon\ p \neq p_0 \qquad\quad H_1\colon\ p > p_0. \end{array}$$

More complex hypotheses could be constructed, but hypotheses of these forms are the most common in practice.

Example. Consider 6.45 from the text. What are the relevant hypotheses H_0 and H_1? What logic led everyone to reject H_0?

Type I and Type II errors

Since we are using data from a random sample to make conclusions about a population, there is always a chance that our conclusions will be false. If we were to repeat our experiment or gather our data many many times, we would sometimes draw a sample leading us to reject H_0 and would sometimes draw a sample leading us to not reject H_0. We would like to test hypotheses in such a way that we control the probability of reaching a false conclusion. There are two ways in which our conclusion can be false:

1. Reject H_0 when H_0 is true. This is called a *Type I error*.

2. Fail to reject H_0 when H_0 is false. This is called a *Type II error*.

We approach hypothesis testing with a view to controlling the probability of a Type I error. Suppose we denote by α the maximum Type I error rate (or probability of a Type I error) we are willing to allow. Then we would like to test hypotheses in such a way that we do not make Type I errors more than a proportion α of the time. The value α is also called the *significance level*.

We denote by β the rate at which we commit Type II errors, that is, we let β represent the probability with which we fail to reject H_0 when H_0 is false. It turns out that this probability is not directly under our control, so we approach hypothesis testing with the Type I error probability α in mind.

We can summarize the outcomes of a test of hypotheses in the following table:

	H_0 true	H_0 false
reject H_0	Type I error	correct decision
fail to reject H_0	correct decision	Type II error

We aim to define our tests of hypotheses so that for a choice of α made by the researcher, the probabilities corresponding to these outcomes satisfy

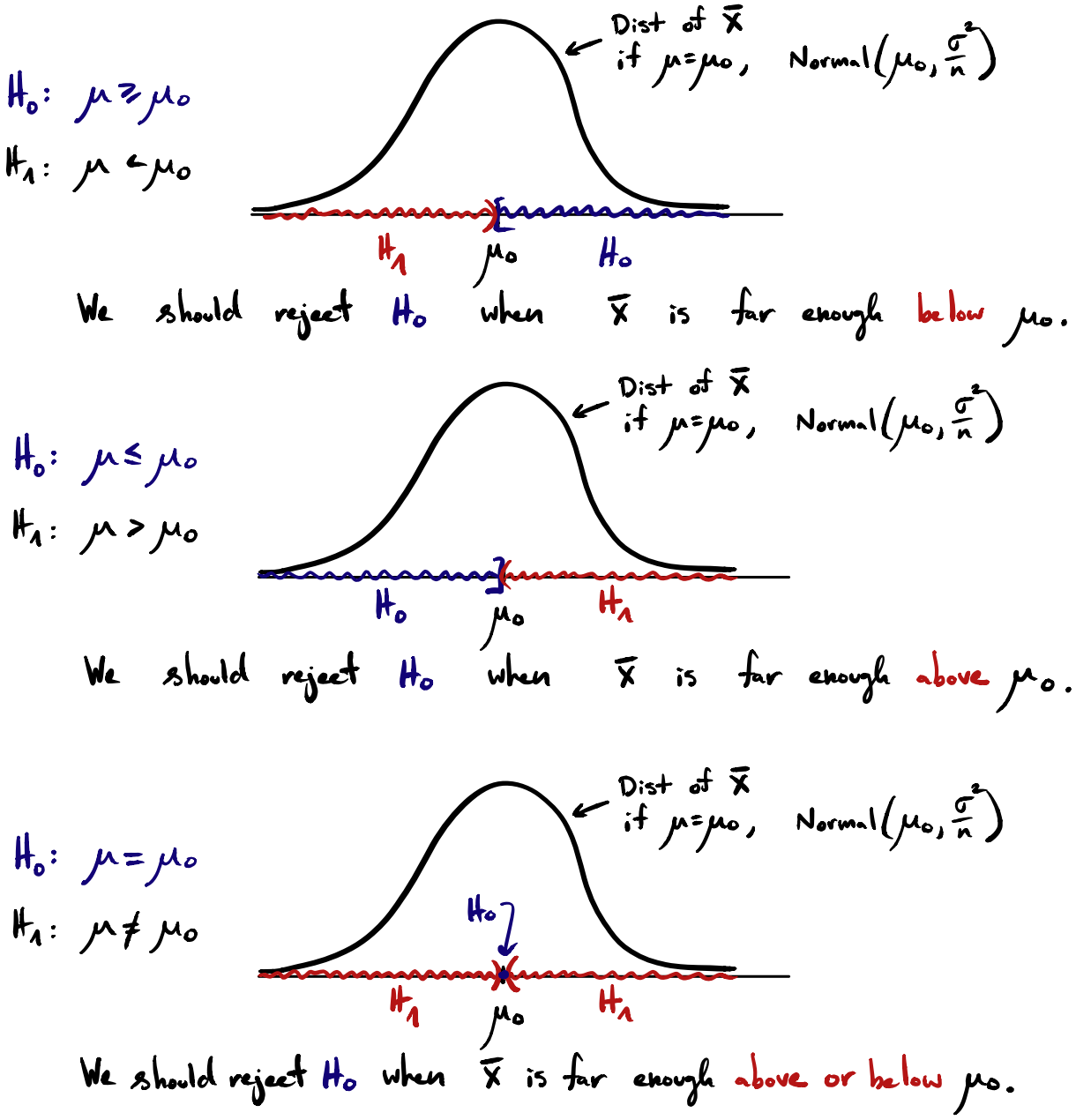
	H_0 true	H_0 false
P(reject H_0)	$\leq \alpha$	$= 1 - \beta$
P(fail to reject H_0)	$> 1 - \alpha$	$= \beta$

for β as small as possible.

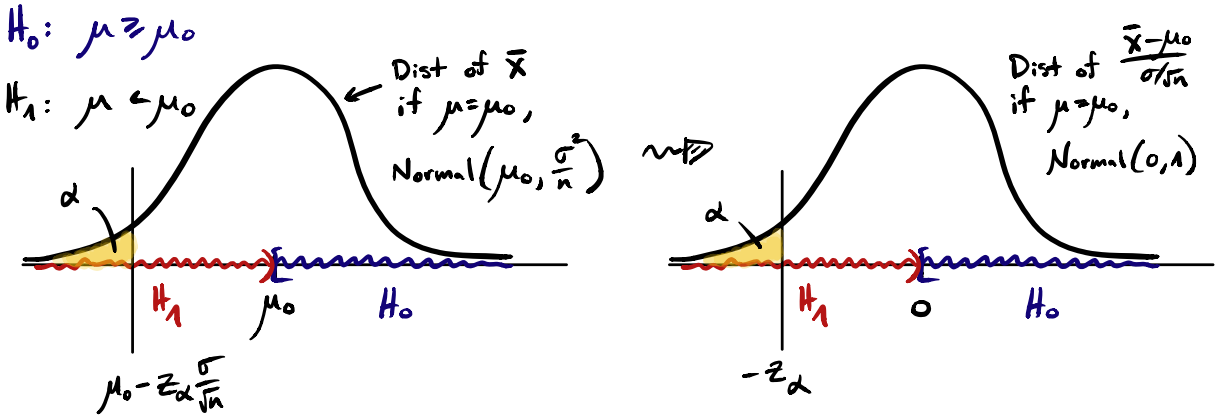
It is important to remark that β is not under our control: it depends on *the extent* to which the null hypothesis is false. We will discuss β and $1 - \beta$ in greater detail later on.

Testing hypotheses about μ (σ known)

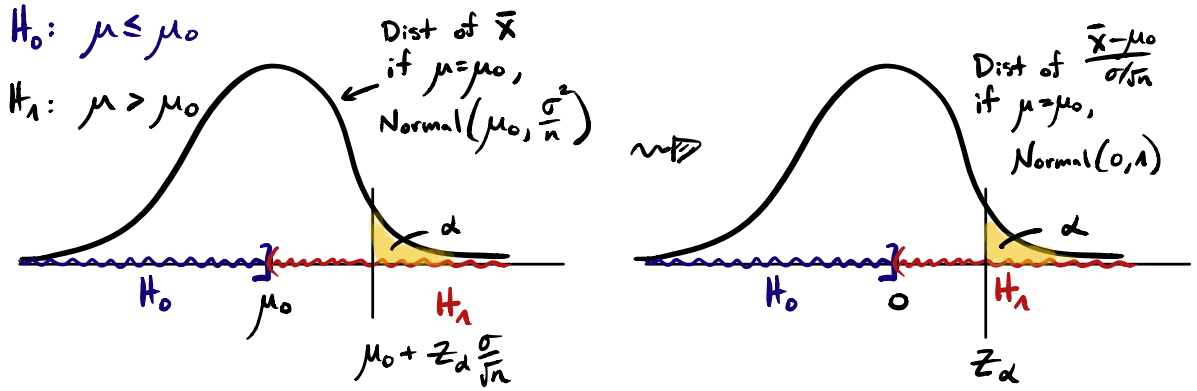
Suppose that $X_1, \ldots, X_n$ are a random sample of X values, where X has unknown mean μ and *known* variance σ^2. Our best guess of μ from the sample is $\bar{X}_n$, so in order to test a hypothesis concerning μ, we should start by looking at the value of $\bar{X}_n$. Let's consider each possible set of hypotheses about μ in turn:



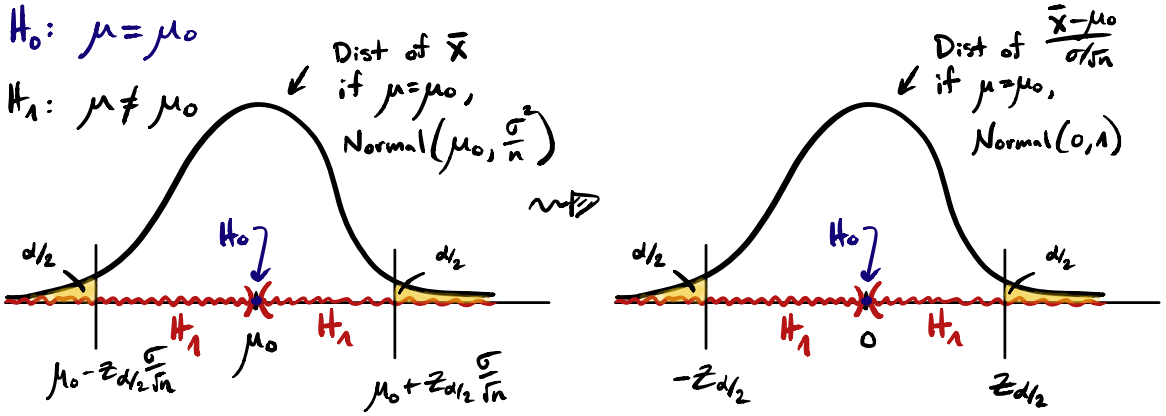
Values of $\bar{X}_n$ which are far away from μ_0 in the direction of the alternative hypothesis cast doubt on the null hypothesis. When there is enough doubt, we will reject H_0. To be precise about how much doubt is enough doubt we specify for each set of hypotheses a region such that if $\bar{X}_n$ falls within this region we reject H_0. Moreover, we choose this region such that the Type I error probability does not exceed α. We can depict our choices of rejection region as follows:



We should reject H_0 when $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < -z_\alpha \iff \bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$.



We should reject H_0 when $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha \iff \bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$.



We should reject H_0 when $|\bar{X}-\mu| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \iff \left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}$.

Let us define the quantity Z_{test} to be the quantity

$$Z_{\text{test}} = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}},$$

since our decisions about rejecting H_0 depend in all cases on this quantity. We shall call this quantity the *test statistic*. Now we can summarize our decision rules (how we decide to reject H_0 or not) in terms of the test statistic Z_{test}. At significance level α, when σ is known, we have the following decision rules:

H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$
Reject H_0 if $Z_{\text{test}} < -z_\alpha$	Reject H_0 if $\lvert Z_{\text{test}}\rvert > z_{\alpha/2}$	Reject H_0 if $Z_{\text{test}} > z_\alpha$

The middle test is called a *two-sided* test, because we reject H_0 if $\bar{X}_n$ is far enough above or below μ_0; the other two tests are called *one-sided* tests, because we reject H_0 only when $\bar{X}_n$ is far enough below μ_0 or only when $\bar{X}_n$ is far enough above μ_0.

The value to which we compare the test statistic in order to make our decision whether to reject H_0 is called the *critical value*. The critical values of the above tests are $-z_\alpha$, $z_{-\alpha/2}$, and z_α. For each test, the critical value defines what we call a *rejection region*; the rejection region of a test is the set of values such that when the test statistic lies in that set the null hypothesis is rejected.

Exercise. Refer to 6.92 of the textbook. Suppose a bottler of soft-drinks claims that its bottling process results in an internal pressure of 157 psi with a standard deviation $\sigma = 3$ psi. You are interested in contracting with the bottler, but you will not do so if the mean internal pressure is less than what the producer stated.

1. What are the relevant hypotheses?

   Answer: Look for the strict inequality. This goes in the alternate hypothesis. So H_1: $\mu < 157$ and thus H_0: $\mu \geq 157$. If we reject H_0, we conclude H_1 and we do not purchase from this bottler.

2. Suppose you collect a sample of size $n = 40$ and get a sample mean of $\bar{X}_n = 155.7$. Suppose $\sigma = 3$. Do you reject or not reject the null hypothesis at the $\alpha = 0.05$ significance level?

   Answer: We have a sample mean which is below the mean claimed by the bottler, but is it low enough for us to reject the bottler's claim? To find out, we must compute the test statistic:

   $$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{155.7 - 157}{3/\sqrt{40}} = -2.741.$$

This value lies in the rejection region, as it is less than the critical value $-z_{0.05} = -1.645$, so we reject the null hypothesis, concluding that the bottler's claim is false.

Exercise. Refer to 6.84 of the textbook. A machine should produce ball bearings such that the standard deviation of the diameters is $\sigma = 0.001$ inches. The mean diameter should be 0.5 inches. You would like to know whether the mean diameter of the ball bearings is different from the targeted diameter of 0.5 inches.

1. What are the relevant hypotheses?

   Answer: Put the strict inequality in the alternative: The key word is "different from," which is "$\neq$". So we have H_1: $\mu \neq 0.5$ and H_0: $\mu = 0.5$. If we reject H_0, we conclude that the machine is not producing ball bearings with the targeted mean diameter.

2. Suppose that the diameters are Normally distributed. You take a random sample of 5 ball bearings and compute a mean diameter of 0.499. Make a decision to reject or not to reject the H_0 at the $\alpha = 0.05$ significance level.

   Answer: We will reject H_0 if the sample mean is sufficiently far away from 0.5. In order to know if the $\bar{X}_n = 0.499$ is far enough away from 0.5 for us to reject H_0: $\mu = 5$, we must compute the test statistic:

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{0.499 - 0.5}{0.001/\sqrt{5}} = -2.24.$$

   We compare the absolute value of this with the critical value $z_{0.025} = 1.96$. Since $2.24 > 1.96$, the test statistic lies in the rejection region, so we reject H_0 and conclude at the 0.05 significance level that the machine is not producing ball bearings with mean diameter 0.5.

Exercise. Suppose you have a random sample of size $n = 35$ with sample mean $\bar{X}_n = 25$ from a right-skewed population with unknown mean μ and variance $\sigma = 10$.

1. Test the hypotheses

$$H_0\text{: } \mu \geq 27 \text{ versus } H_1\text{: } \mu < 27$$

   at significance level $\alpha = 0.05$.

   Answer: The null value μ_0 of μ is $\mu_0 = 27$. We will reject H_0: $\mu \geq 27$ if the sample mean $\bar{X}_n$ is far enough below $\mu_0 = 27$. To determine whether $\bar{X}_n$ is far enough below $\mu_0 = 27$, we compute the test statistic

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} = \frac{25 - 27}{10/\sqrt{35}} = -1.18.$$

We now compare the value of the test statistic to the critical value $-z_\alpha = -z_{.05} = -1.645$. Since $-1.18 > -1.645$, the test statistic does not lie in the rejection region, so we fail to reject H_0: $\mu \geq 27$ at the 0.05 significance level; $\bar{X}_n$ is not far enough below $\mu_0 = 27$ for us to reject the claim that $\mu \geq 27$.

2. Test the hypotheses

$$H_0\text{: } \mu = 27 \text{ versus } H_1\text{: } \mu \neq 27$$

at significance level $\alpha = 0.05$.

Answer: The value of the test statistic is the same, but now we compare its absolute value to the critical value $z_{\alpha/2} = z_{0.025} = 1.96$. Since $|-1.18| < 1.96$, the test statistic does not lie in the rejection region, so we fail to reject H_0: $\mu = 27$ at the 0.05 significance level; the sample mean $\bar{X}_n$ is not far enough away from $\mu_0 = 27$ in order for us to reject the claim that $\mu = 27$.

3. Test the hypotheses

$$H_0\text{: } \mu \leq 27 \text{ versus } H_1\text{: } \mu > 27$$

at significance level $\alpha = 0.05$.

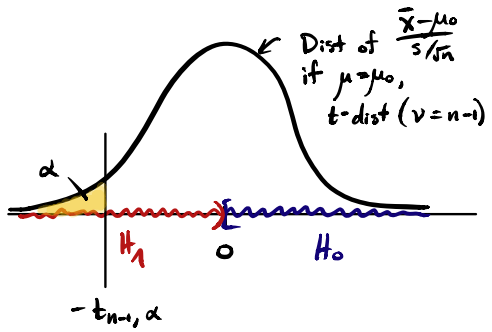
Answer: We note that the value of the sample mean $\bar{X}_n = 25$ is in support of the null hypothesis. The data cannot lead us to reject it. We fail to reject H_0.

Testing hypotheses about μ (σ unknown)

We just replace σ with s in the above as well as z_α with $t_{n-1,\alpha}$ and $z_{\alpha/2}$ with $t_{n-1,\alpha/2}$. We can draw pictures in this setting which are analogous to those in the previous section:

$H_0: \mu \geq \mu_0$

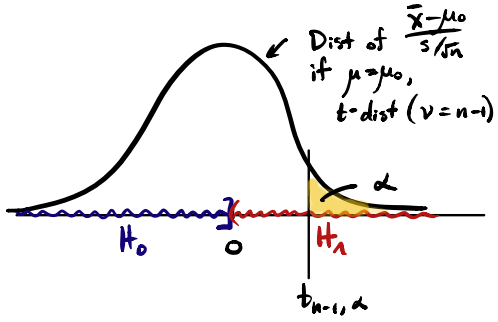
$H_1: \mu < \mu_0$



We should reject H_0 when $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha}$ $\Leftrightarrow$ $\bar{x} < \mu_0 + t_{n-1,\alpha}\frac{s}{\sqrt{n}}$

$H_0: \mu \leq \mu_0$

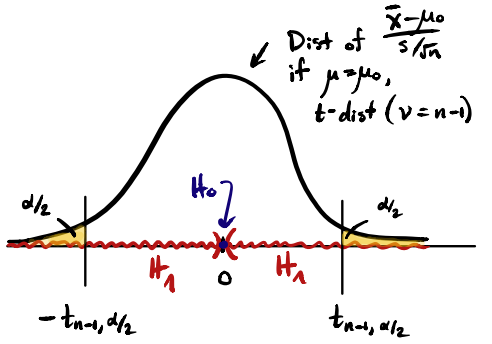
$H_1: \mu > \mu_0$



We should reject H_0 when $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} > t_{n-1,\alpha}$ $\Leftrightarrow$ $\bar{x} > \mu_0 + t_{n-1,\alpha}\frac{s}{\sqrt{n}}$

$H_0: \mu = \mu_0$

$H_1: \mu \neq \mu_0$



We should reject H_0 when $\left|\frac{\bar{x}-\mu}{s/\sqrt{n}}\right| > t_{n-1,\alpha/2}$ $\Leftrightarrow$ $|\bar{x}-\mu| > t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$

Let us define

$$T_{\text{test}} = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}},$$

noting that our decisions about rejecting H_0 depend in all cases on this quantity. Now T_{test} is our test statistic when σ is unknown, and we have the following decision rules for rejecting H_0 at significance level α:

H_0: $\mu \geq \mu_0$ H_1: $\mu < \mu_0$	H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	H_0: $\mu \leq \mu_0$ H_1: $\mu > \mu_0$
Reject H_0 if $T_{\text{test}} < -t_{n-1,\alpha}$	Reject H_0 if $\lvert T_{\text{test}}\rvert > t_{n-1,\alpha/2}$	Reject H_0 if $T_{\text{test}} > t_{n-1,\alpha}$

The critical values $-t_{n-1,\alpha}$, $t_{n-1,\alpha/2}$, and $t_{n-1,\alpha}$ which define the rejection regions come from the t-distribution with $n-1$ degrees of freedom.

Exercise. The average height of 14 randomly selected ten-yr-old Loblolly pine trees was $\bar{X}_n = 27.44$ and the sample standard deviation was $S_n = 1.54$. Assume that the heights of ten-yr-old Loblolly pine trees are Normally distributed.

1. Test the hypotheses

   $$H_0\text{: } \mu \leq 26 \text{ versus } H_1\text{: } \mu > 26$$

   at significance level $\alpha = 0.05$.

   Answer: The null value μ_0 of μ is $\mu_0 = 26$. We will reject H_0 if $\bar{X}_n$ is far enough above $\mu_0 = 26$. To see whether $\bar{X}_n$ is far enough above $\mu_0 = 26$, we compute the test statistic

   $$\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{27.44214 - 26}{1.537887/\sqrt{14}} = 3.5.$$

   The critical value to which we must compare the test statistic is $t_{n-1,\alpha} = t_{13,0.05} = 1.77$. Since the value of our test statistic exceeds the $\alpha = 0.05$ critical value, it lies in the rejection region, so we reject H_0: $\mu \leq 26$ at the 0.05 significance level.

2. Test the hypotheses

   $$H_0\text{: } \mu = 26 \text{ versus } H_1\text{: } \mu \neq 26$$

   at significance level $\alpha = 0.05$.

   Answer: We compute the same test statistic: Its value is 3.5. Now, however, we check whether it is greater than the critical value $t_{13,\alpha/2}$ in absolute value. We have $t_{13,0.025} = 2.16$. Since $3.5 > 2.16$, the test statistic lies in the rejection region, so we reject H_0: $\mu = 26$ at the 0.05 significance level.

3. Test the hypotheses

$$H_0\text{: } \mu \geq 26 \text{ versus } H_1\text{: } \mu < 26$$

at significance level $\alpha = 0.05$.

Answer: The value of $\bar{X}_n$ exceeds $\mu_0 = 26$, so it lies in region specified by H_0. The sample contains no evidence against H_0, so we fail to reject H_0: $\mu \geq 26$.

Connection between CIs and two-sided tests about μ

Whenever we have hypotheses of the form

$$H_0\text{: } \mu = \mu_0 \text{ versus } H_1\text{: } \mu \neq \mu_0$$

we can perform a test at significance level α by simply checking whether the $(1-\alpha)100\%$ confidence interval contains μ_0. If the confidence interval contains μ_0, we fail to reject H_0. If the confidence interval does not contain μ_0, we reject H_0.

Exercise. The average height of 14 randomly selected ten-yr-old Loblolly pine trees was $\bar{X}_n = 27.44$ and the sample standard deviation was $s = 1.54$. Assume that the heights of ten-yr-old Loblolly pine trees are Normally distributed. Test the hypotheses

$$H_0\text{: } \mu = 26 \text{ versus } H_1\text{: } \mu \neq 26$$

at significance level $\alpha = 0.05$.

Answer: We can build a 95% confidence interval and see if it contains 26. The 95% confidence interval is

$$\bar{X}_n \pm t_{13,\alpha/2}\frac{S_n}{\sqrt{n}} = 27.44 \pm 2.160\frac{1.54}{\sqrt{14}} = (26.55, 28.33)$$

The 95% confidence interval does not contain 26, so we reject H_0: $\mu = 26$ at the $\alpha = 0.05$ significance level.

Testing hypotheses about a proportion p

As in the case of the mean μ, we will consider the following possible sets of hypotheses about the population proportion p:

$$\begin{array}{lll} H_0\text{: } p \geq p_0 & \text{or} & H_0\text{: } p = p_0 \quad \text{or} \quad H_0\text{: } p \leq p_0 \\ H_1\text{: } p < p_0 & & H_1\text{: } p \neq p_0 \qquad\quad H_1\text{: } p > p_0, \end{array}$$

We note again that the textbook will write

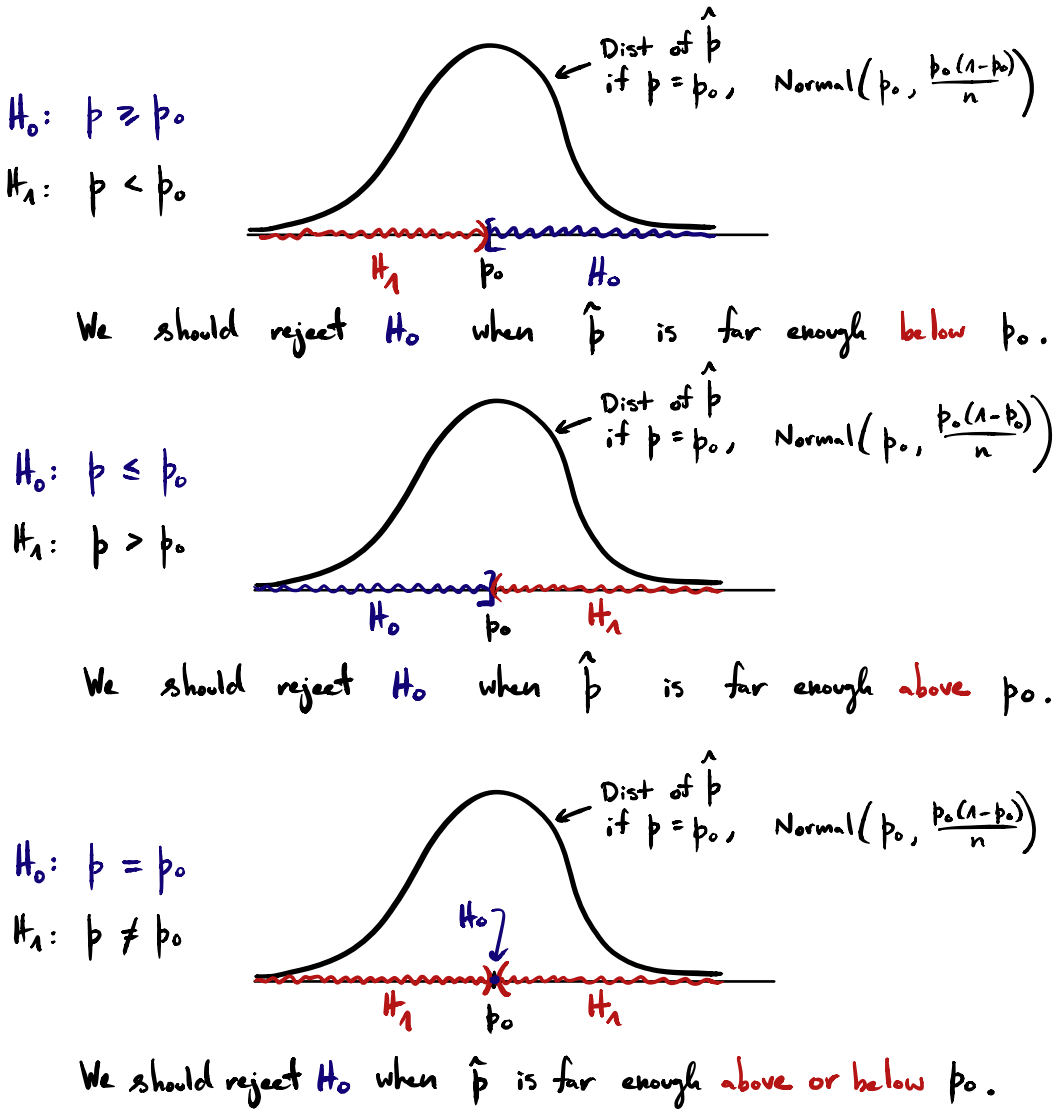
$$\begin{array}{lll} H_0\colon\ p = p_0 & \text{or} & H_0\colon\ p = p_0 \quad \text{or} \quad H_0\colon\ p = p_0 \\ H_1\colon\ p < p_0 & & H_1\colon\ p \neq p_0 \qquad\quad H_1\colon\ p > p_0. \end{array}$$

To test any of these hypotheses, we will consider how far away the sample proportion $\hat{p}$ is from p_0 in the direction of the alternative. We will reject H_0 if $\hat{p}$ is far enough away from the null value p_0 in the direction of the alternative. In order to determine how far is far enough, we will use the fact that if the true population proportion p is equal to p_0, the quantity

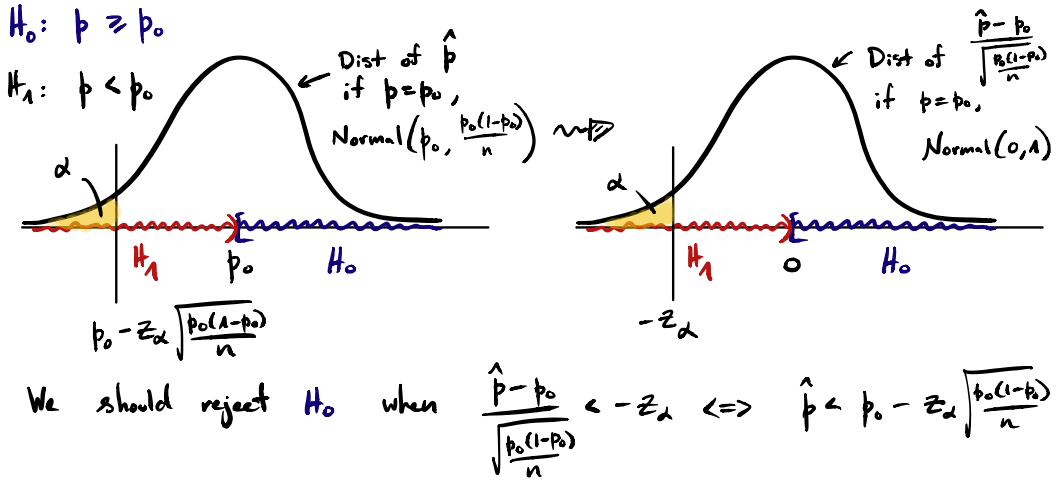
$$\frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

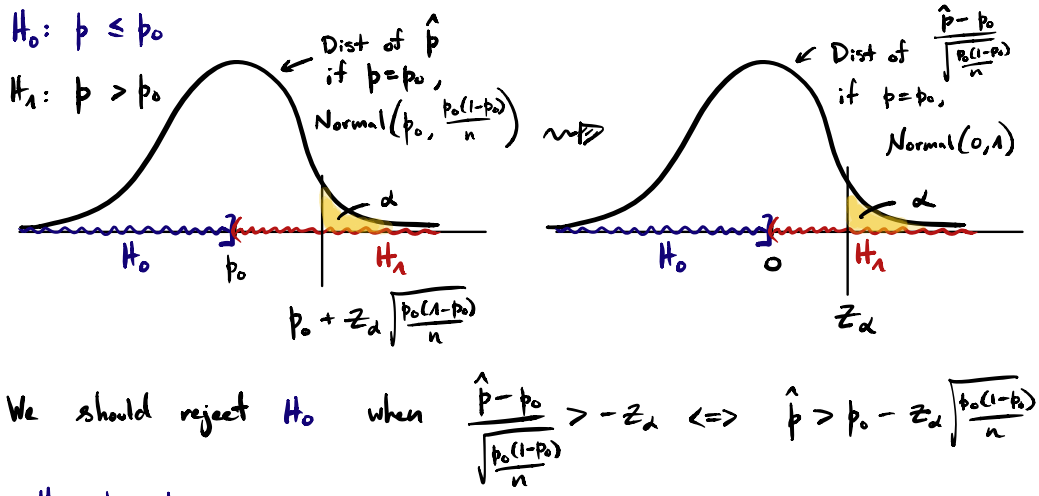
has approximately the Normal$(0, 1)$ distribution for large enough n. This quantity will be our *test statistic*.

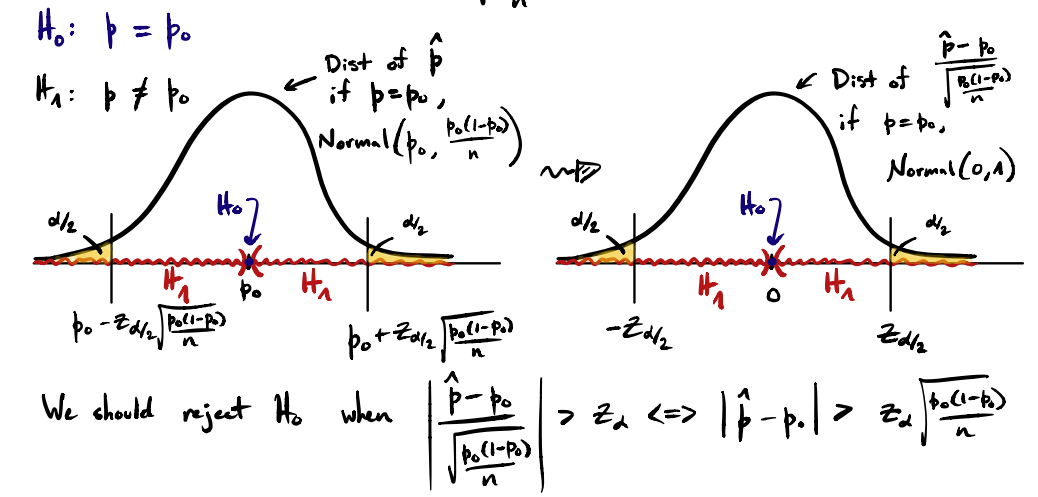
We can draw pictures analogous to those we drew when considering hypotheses about a mean μ. Firstly, our intuition tells us whether we should reject H_0 when $\hat{p}$ is far below, far above, or far away from p_0 in either direction:



We may define the critical regions of the tests by going to the Z-world; however, we must remember that $\hat{p}$ is only Normally distributed when we expect a sufficient number of successes and failures to be present in the sample. In the context of hypothesis testing, we require that $np_0 \geq 15$ and $n(1-p_0) \geq 15$. Then we may draw the following pictures:







Let us use reuse the notation Z_{test} for the test statistic

$$Z_{\text{test}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

upon which our decision to reject H_0 in all cases depends. Then we may summarize our decision rules for rejecting H_0 at significance level α by the following:

H_0: $p \geq p_0$ H_1: $p < p_0$	H_0: $p = p_0$ H_1: $p \neq p_0$	H_0: $p \leq p_0$ H_1: $p > p_0$
Reject H_0 if $Z_{\text{test}} < -z_\alpha$	Reject H_0 if $\lvert Z_{\text{test}} \rvert > z_{\alpha/2}$	Reject H_0 if $Z_{\text{test}} > z_\alpha$

Exercise. A scientist is interested in seeing whether the presence of a parasite tips the sex ratio of the hosts' offspring in favor of females (which would be advantageous to the parasite, as it inhabits only females). A sample of size $n = 500$ offspring from parasite-infected females is collected, among which there are 287 females.

1. What are the relevant hypotheses?

   Answer: Assuming that the proportion of females in the offspring of the host species is 1/2 in the absence of the parasite, the hypotheses of interest are

   $$H_0\text{: } p \leq 1/2 \text{ versus } H_1\text{: } p > 1/2,$$

   where p is the proportion of females in the offspring of parasite-infected hosts.

2. Carry out a test of the hypotheses at the $\alpha = 0.05$ significance level.

   Answer: We first compute the test statistic

   $$Z_{\text{test}} = \frac{287/500 - 1/2}{\sqrt{1/2(1-1/2)/500}} = 3.309.$$

   The critical value is $z_{0.05} = 1.645$. Since $3.309 > 1.645$, the test statistic lies in the rejection region, so we reject H_0 and conclude that the parasite indeed tips the sex ratio of offspring in favor of females.

Exercise. Refer to 8.82 of the textbook. In an tasting experiment, 121 students were blindfolded, and each student was fed either a red or green gummy bear, (red with probability 1/2 and green with probability 1/2) and asked to identify which color it was based on the taste. Among the 121 students, 97 correctly identified the color of the gummy bear.

1. If the students guessed "red" or "green" based on flipping a coin, with what probability would they guess the color correctly?

   Answer: The students would guess correctly with probability 1/2.

2. Suppose you wish to know if the students are doing better or worse than guessing. What are the relevant hypotheses?

   Answer: If p is the probability of guessing correctly based on the taste, we are interested in testing

   $$H_0\text{: } p = 1/2 \text{ versus } H_1\text{: } p \neq 1/2.$$

3. Test the hypotheses at the $\alpha = 0.01$ significance level.

   Answer: The test statistic is

   $$Z_{\text{test}} = \frac{97/121 - 1/2}{\sqrt{1/2(1-1/2)/121}} = 6.644.$$

   The $\alpha = 0.01$ critical value is $z_{0.005} = 2.576$. Since $6.644 > 2.576$, the test statistic lies in the rejection region, so we reject H_0 at the 0.01 significance level.