Lab 8: Hypothesis Testing

For this lab, it will be helpful to have a copy of the knitted version of this document to answer the questions as much of it is written using mathematical notation that may be difficult to read when the document is not knitted. For your convenience, a pdf of this document is in the lab folder on blackboard. Please ignore the code used to generate the figures in this lab.

Lab Goal

The purpose of this lab is to explore hypothesis testing for an unknown population mean, μ, when $\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$. There are three main topics in the lab:

(I) how to form a rejection region at a given significance level
(II) how to calculate a p-value
(III) how to calculate the power of a test (or choose n to get a desired power)

Recap of "psychic" example in lecture

In class, Dr. Bien used the binomial distribution to test whether or not he was psychic. What were the basic ideas behind this test?

1) The **null hypothesis** is that Dr. Bien is not psychic, meaning he is just random guessing. If he were random guessing, then he would only have a 1 in 4 chance of guessing the correct number for each person in the class, so $H_0 : p = 1/4$ (where p is the probability of his being right on any single guess).

2) We would like to test whether Dr. Bien is psychic. This implies that the actual probability of Dr. Bien guessing the correct number for a given student is greater than what we would expect by random guessing. This establishes the **alternative hypothesis** of our test, $H_a : p > 0.25$.

3) He repeated the experiment $n = 89$ times. Under the null hypothesis, we would expect him to get, on average, $0.25 \times 89 \approx 22$ correct. We will decide to reject H_0 if we have enough evidence in favor of H_a over H_0. What does "evidence in favor of H_a over H_0" look like? That depends on the form of H_a. In this case, $H_a : p > 0.25$, so if the number of correct predictions Dr. Bien made is a lot bigger than 22, that would lead us to reject H_0 in favor of H_a. How do we determine a specific cutoff for what "a lot bigger" means? We determine the distribution of the test statistic under H_0 and then choose the cutoff so that the probability of rejecting H_0 when H_0 is true is no more than α, the desired **significance level** of the test. The set of values (of the test statistic) that would lead to our test rejecting H_0 in favor of H_a is called the **rejection region**.

Overview of hypothesis testing using a rejection region

Step 1. Identify null and alternative hypotheses.

Step 2. Determine distribution of the test statistic under the null hypothesis.

Step 3. Determine rejection region based on (1) and (2) and the desired significance level α chosen for this test.

Step 4. Compute test statistic on the particular sample of data you collected. Does this fall in rejection region? If so, reject H_0 in favor of H_a. If not, fail to reject H_0 in favor of H_a.

Step 5. Check assumptions.

Part I of lab

We go through Step 1 to Step 4, with a different kind of alternative hypothesis and when test statistic is normal. We will skip Step 5 for now since we'll see how to check normality in a future lab.

Step 1: Choosing alternative hypothesis

Depending on the particular problem you are studying, there are three types of alternative hypothesis to consider about the population mean μ (in "psychic" example, p is the population mean of the Bernoulli distribution).

A1) H_a: $\mu \neq \mu_0$ ("two-sided alternative")

A2) H_a: $\mu > \mu_0$ (used in "psychic" example with $\mu_0 = 0.25$)

A3) H_a: $\mu < \mu_0$

If in the "psychic" example, we were interested in whether Dr. Bien was random guessing or not, we should have chosen $H_a : p \neq 0.25$. But since being psychic would mean specifically that $p > 0.25$, we chose this alternative instead.

Step 2: Distribution of test statistic under null hypothesis

In the "psychic" example, we based the test of whether to reject H_0 in favor of H_a on the distribution of a test statistic under the assumption that H_0 is true. This is the standard framework for hypothesis testing.

For the example given in class, H_0 states that $p = 0.25$. The null hypothesis indicates that for any single student there is a 0.25 chance that Dr. Bien will succeed in guessing his/her number correctly. This indicates that the number of correct guesses in 89 trials (89 students) has a Binomial(89,0.25) distribution under the assumption that Dr. Bien is not psychic (H_0). Thus, in the example given in class, this was the appropriate null distribution for the number of correct guesses.

In many situations, we wish to perform a hypothesis test about the population mean μ, and the most common choice of test statistic in this context is the sample mean, $\bar{X}_n$. For this lab, we will assume that $\bar{X}_n \sim N(\mu, \sigma/\sqrt{n})$. This is approximately true when (i) the Central Limit Theorem holds or when (ii) the individual X_i are themselves approximately normal and independent of each other. Under the null hypothesis that $H_0 : \mu = \mu_0$, we have that $\bar{X}_n \sim N(\mu_0, \sigma/\sqrt{n})$.

Example: Hard Alcohol Content of Mixed Drinks in NYC The industry standard for the amount of alcohol poured into many types of drinks is 1.5 oz. Regulators are interested in whether on average the amount of alcohol poured into drinks purchased in New York City is the industry standard. Forty bartenders are chosen at random from bars and restaurants in NYC. Each were asked to pour the amount of hard alcohol into a glass needed to make a mixed drink. The amount of hard alcohol poured by each bartender was recorded. The standard deviation of the amount of hard alcohol poured to make a drink in NYC is known to be about 0.39 oz. The sample mean of the amount of hard alcohol poured for the mixed drink by the 40 bartenders was 1.7 oz.

a) Define μ and determine the null and alternative hypotheses of this study.

μ = *Mean amount of hard alcohol in mixed drinks made in NYC*

$H_0 : \mu = 1.5$ *and* $H_a : \mu \neq 1.5$

b) What is the approximate distribution of the sample mean under the null hypothesis for this study?

$\bar{X}_{40} \sim N(1.5, 0.39/\sqrt{40})$

Step 3: Determine rejection region

In the "psychic" example, we decided that if the test statistic (number of correct guesses) was higher than a certain cutoff, we should reject H_0 in favor of H_a. The set of values above this cutoff are called the rejection region for the test.

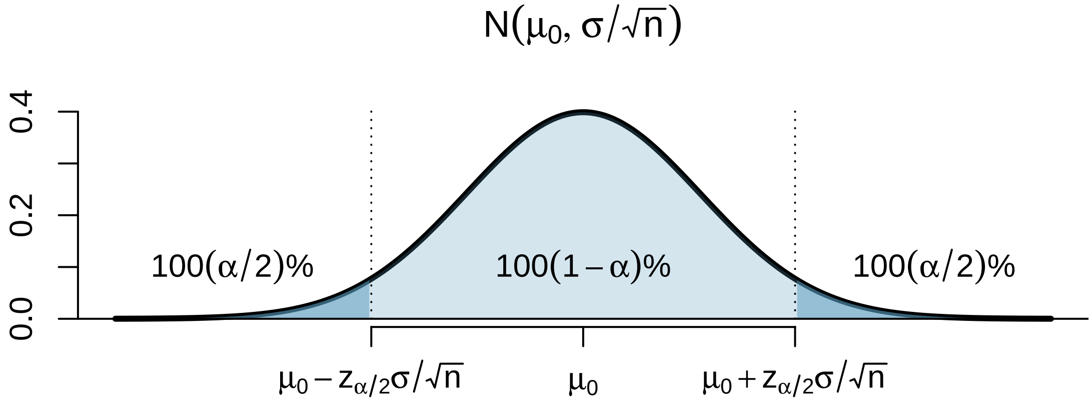
For any test, the **rejection region** is the range of values of the test statistic for which H_0 will be rejected in favor of H_a. The rejection region is based on three things:

1) The desired significance level, α, of the test. This is the largest that we will allow the probability of making a Type I error (rejecting H_0 when H_0 is true) to be.
2) The alternative hypothesis. Looking at the alternative tells us whether high values or low values (or both) of the test statistic would be considered evidence against the null and *in favor of the alternative.*
3) The distribution of the test statistic under the null (from Step 2) tells us what the cutoff values should be for the rejection region to make the test have significance level α.

In class, the rejection region for $H_a : p > 0.25$ was $X \geq 30$. This rejection region guarantees that if H_0 is true, there is at most a 5% chance that H_0 will be rejected.

What are the rejection regions based on the sample mean for the three alternatives above?

A1) If H_a: $\mu \neq \mu_0$, H_0 will be rejected if the observed sample mean from our data, $\bar{x}_n$, is either a lot bigger or a lot smaller than what we would expect under H_0. How large is "a lot bigger" and how small is "a lot smaller"? These cutoffs come from considering the null distribution of $\bar{X}_n$ and the significance level α. It's easiest looking at a picture:

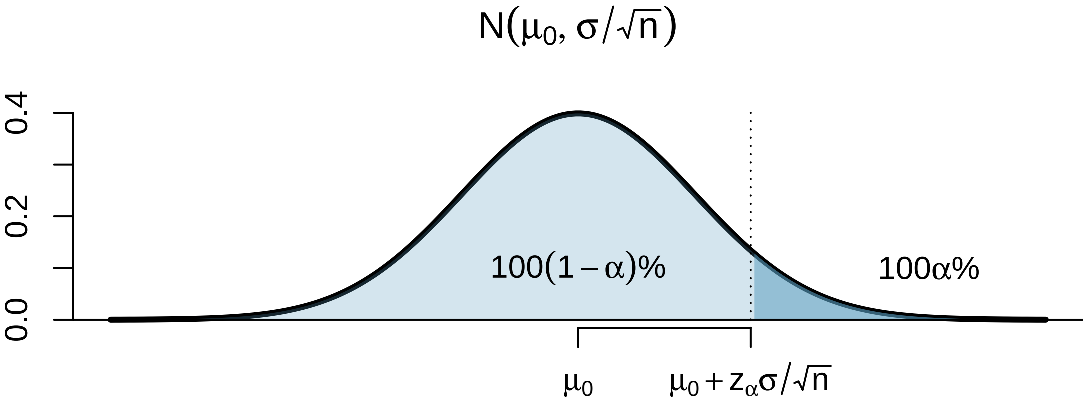


Reject H_0 if $\bar{x}_n$ is less than the $\alpha/2$ quantile (a.k.a., the $100(\alpha/2)$ percentile) of the distribution of $\bar{X}_n$ under H_0 *or* if $\bar{x}_n$ is greater than the $(1-\alpha/2)$ quantile (a.k.a., the $100(1-\alpha/2)$ percentile) of the distribution of $\bar{X}_n$ under H_0. We see from the picture that we reject if $\bar{x}_n$ is farther than $z_{\alpha/2}\sigma/\sqrt{n}$ from μ_0:

$$|\bar{x}_n - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}.$$

That is we reject if $\bar{x}_n > \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$ *or* if $\bar{x}_n < \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$.

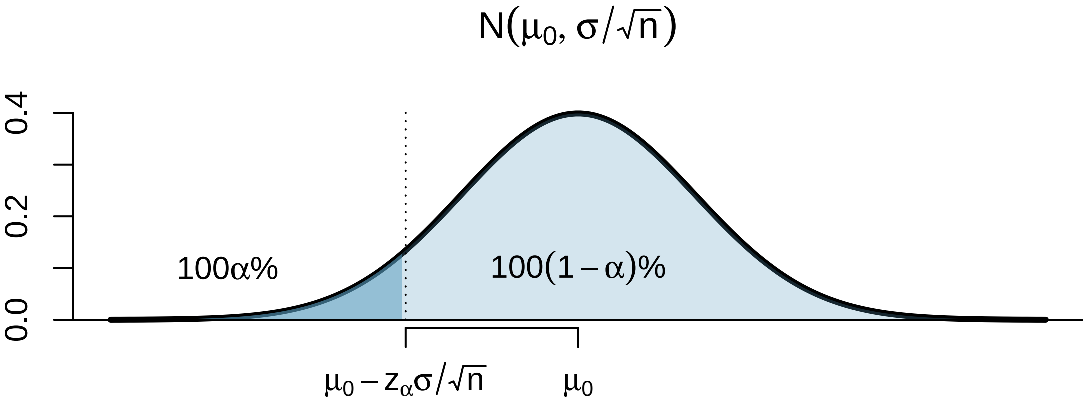
A2) If H_a: $\mu > \mu_0$, H_0 will be rejected if the observed sample mean from our data, $\bar{x}_n$, is a lot bigger than what we would expect under H_0. How large is "a lot bigger"? The cutoff comes from considering the null distribution of $\bar{X}_n$ and the significance level α. Again, it's always best to draw a picture:



Reject H_0 if $\bar{x}_n \geq$ the $100(1-\alpha)$ percentile of the distribution of $\bar{X}_n$ under H_0. We see from the picture that we reject if

$$\bar{x}_n > \mu_0 + z_\alpha \sigma/\sqrt{n}.$$

A3) If H_a: $\mu < \mu_0$, H_0 will be rejected if the observed sample mean from our data, $\bar{x}_n$, is a lot smaller than what we would expect under H_0. How small is "a lot smaller"? The cutoff comes from considering the null distribution of $\bar{X}_n$ and the significance level α. Again, we draw a picture:



Reject H_0 if $\bar{x}_n \leq 100\alpha$ percentile of the distribution of $\bar{X}_n$ under H_0. We see from the picture that we reject if

$$\bar{x}_n < \mu_0 - z_\alpha \sigma/\sqrt{n}.$$

Example: Hard Alcohol Content of Mixed Drinks in NYC We continue with the next part of our example problem. Suppose the regulators want to perform the hypothesis test determined above at a significance level of 0.01.

c) What is the rejection region of this test?

As always, we determine the rejection region based on (i) the type of alternative hypothesis, (ii) the distribution of the test statistic under H_0, and (iii) the significance level α. Then we check whether $\bar{x}_{40} = 1.7$ falls in this region (basic idea: Is 1.7 an unusual realization of $\bar{X}_{40}$ under H_0?). Above, it was determined under H_0 that $\bar{X}_{40} \sim N(1.5, 0.39/\sqrt{40})$. There are two ways we can calculate this in R.

- *Method 1: Our alternative hypothesis is $\mu \neq \mu_0$, so (referring to A1's picture), we should reject if $\bar{x}_n > \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}$ or if $\bar{x}_n < \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}$.*

```
mu0 = 1.5
sigma = 0.39
n = 40
alpha = 0.01
```

```
z = -qnorm(alpha / 2)
mu0 + z * sigma / sqrt(n) # reject above this
```

```
## [1] 1.658837
```

```
mu0 - z * sigma / sqrt(n) # or below this
```

```
## [1] 1.341163
```

- *Method 2: Looking at the A1 picture, we really are just trying to find the $\alpha/2$ and $1-\alpha/2$ quantiles of the $N(1.5, 0.39/\sqrt{40})$ distribution. So in R we could do this more directly:*

```
qnorm(0.995, mean = mu0, sd = sigma/sqrt(n)) # 99.5th Percentile
```

```
## [1] 1.658837
```

```
qnorm(0.005, mu0, sigma/sqrt(n)) # 0.5th Percentile
```

```
## [1] 1.341163
```

Both methods are doing the same thing: We reject H_0 when $\bar{x}_{40}$ is either larger than or equal to the 99.5th percentile of $N(1.5, 0.39/\sqrt{40})$ or smaller than or equal to the 0.5th percentile of $N(1.5, 0.39/\sqrt{40})$. This ensures that the probability that $\bar{X}_n$ is in the rejection region when H_0 is true is at most 0.01. Using either method we get the same answer: We reject H_0 in favor of H_a if $\bar{x}_{40} \geq 1.66$ or $\bar{x}_{40} \leq 1.34$

d) Based on the rejection region, should we reject H_0?

Since $1.7 > 1.66$, we reject H_0 in favor of H_a at the $\alpha = 0.01$ significance level. There is evidence that the mean amount of hard alcohol in mixed drinks purchased in NYC is not the industry standard of 1.5 oz.

Overview of hypothesis testing using P-values

Instead of the rejection region approach, we now consider the (highly related) p-value approach to hypothesis testing.

Step 1. Identify null and alternative hypotheses.

Step 2. Determine distribution of the test statistic under the null hypothesis.

Step 3. Compute the p-value based on the particular sample of data you collected.

Step 4. Reject H_0 in favor of H_a if p-value is smaller than desired significance level α.

Step 5. Check assumptions.

The **p-value** of a test is the probability under the null hypothesis of seeing something as extreme or more extreme than what was actually observed.

Part II of lab

We go through these steps (again, skipping step 5 for this lab) with alternative hypotheses A1 to A3 when the test statistic is normal.

Step 1 and Step 2

Identical to Steps 1 and 2 in rejection region approach.

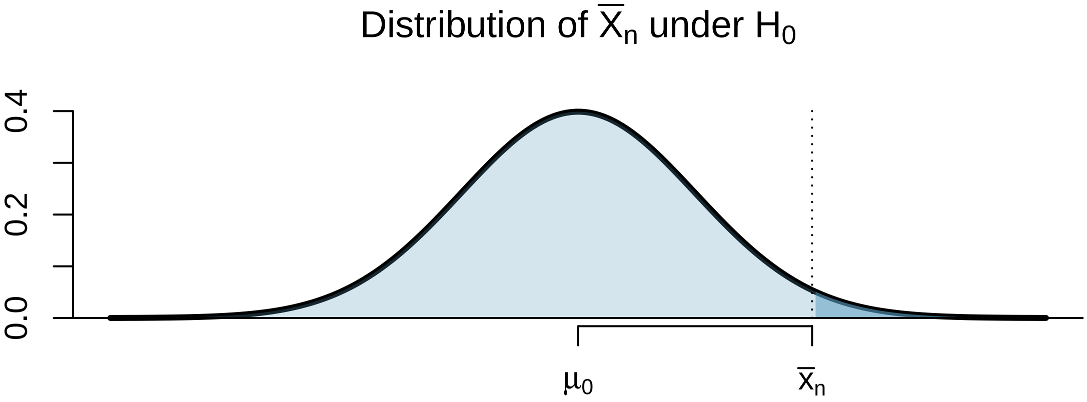
Step 3: Compute the p-value

We observed $\bar{x}_n$ and want to know how likely it is for $\bar{X}_n \sim N(\mu_0, \sigma/\sqrt{n})$ (this is the null distribution) for $\bar{X}_n$ to be as extreme or more extreme as $\bar{x}_n$. What does "extreme" mean? This is defined by the alternative hypothesis. Because A1 is the most complicated, we'll do this last.

A2) H_a: $\mu > \mu_0$. In this case, "extreme" means being much larger than μ_0. So we want to know the probability (calculated under H_0) that $\bar{X}_n$ would be as big as $\bar{x}_n$ (our observed value) or larger. That is, we want to know

$$P(\bar{X}_n > \bar{x}_n \mid H_0 \text{ true})$$

Recalling that H_0 says that $\bar{X}_n \sim N(\mu_0, \sigma/\sqrt{n})$ we can draw a picture to show the probability we want:



The p-value in this case is

$$P(\bar{X}_n > \bar{x}_n \mid H_0 \text{ true}).$$

In R this is given by

```
1 - pnorm(xbar, mu0, sigma / sqrt(n))
```

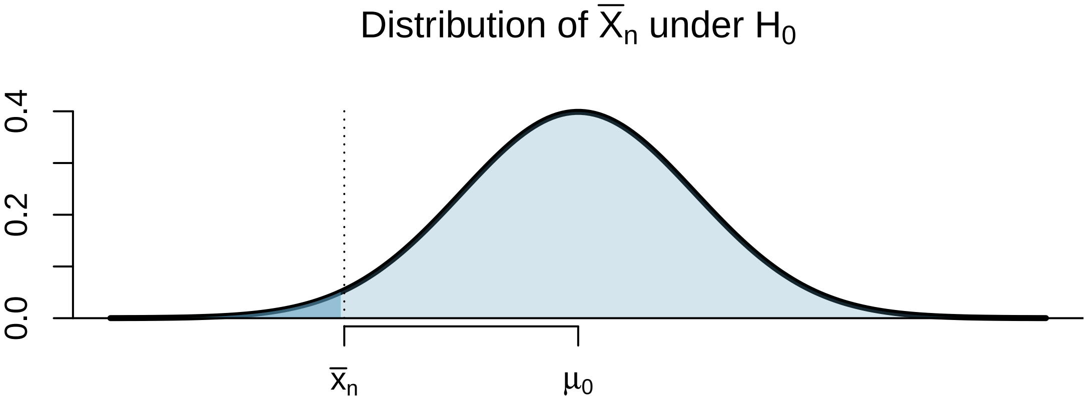
or we could standardize to z-values (to get same answer):

```
zvalue = (xbar - mu0) / (sigma / sqrt(n)) # standardize according to null
1-pnorm(zvalue) # under null, 'zvalue' is a realization from a N(0,1)
```

A3) H_a: $\mu < \mu_0$. In this case, "extreme" means being much smaller than μ_0. So we want to know the probability (calculated under H_0) that $\bar{X}_n$ would be as small as $\bar{x}_n$ (our observed value) or smaller. That is, we want to know

$$P(\bar{X}_n < \bar{x}_n \mid H_0 \text{ true})$$

It's always a good idea to draw a picture:



The p-value in this case is

$$P(\bar{X}_n < \bar{x}_n \mid H_0 \text{ true}).$$

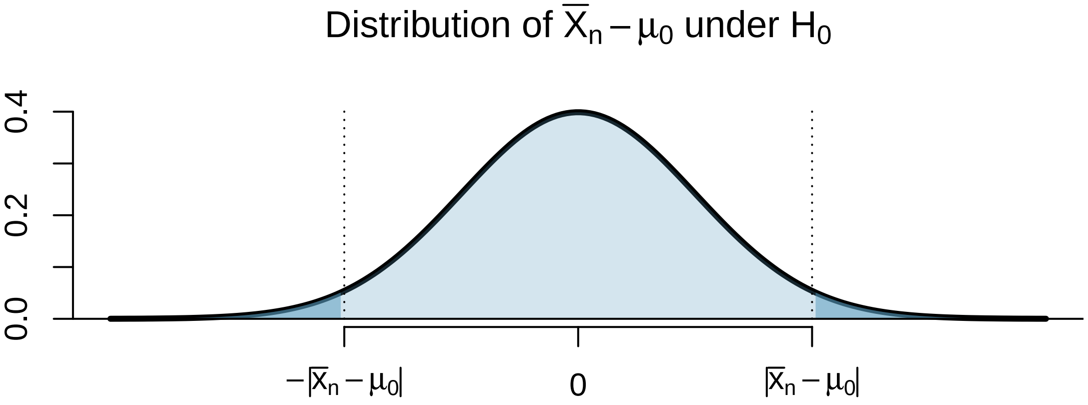
In R this is given by

```
pnorm(xbar, mu0, sigma / sqrt(n))
```

or we could standardize to z-values (to get same answer):

```
zvalue = (xbar - mu0) / (sigma / sqrt(n)) # standardize according to null
pnorm(zvalue) # under null, 'zvalue' is a realization from a N(0,1)
```

A1) H_a: $\mu \neq \mu_0$. In this case, "extreme" means either much larger or much smaller than μ_0. So we want to know the probability (calculated under H_0) that $\bar{X}_n$ would be as far or farther from μ_0 as we observed $\bar{x}_n$ to be. That is, we want to know if $|\bar{x}_n - \mu_0|$ is unusually high based on what we'd expect for $|\bar{X}_n - \mu_0|$ under H_0. Note that H_0 tells us that $\bar{X}_n - \mu_0 \sim N(0, \sigma/\sqrt{n})$. Let's draw a picture to show what we mean:



The p-value in this case is

$$P(\bar{X}_n - \mu_0 > |\bar{x}_n - \mu_0| \text{ or } \bar{X}_n - \mu_0 < -|\bar{x}_n - \mu_0| \mid H_0 \text{ true}) = 2P(\bar{X}_n - \mu_0 > |\bar{x}_n - \mu_0| \mid H_0 \text{ true}).$$

For calculating this probability, it's easier to deal with the standard normal. Under the null,

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

so we want to know if we calculate

$$z_n = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}},$$

we want to calculate

$$P(N(0,1) > |z_n| \text{ or } N(0,1) < -|z_n|) = 2P(N(0,1) > |z_n|).$$

In R, this is given by

```
2 * (1 - pnorm(abs(z)))
```

P-value: Hard Alcohol Content of Mixed Drinks in NYC

a) What is the p-value of this test?

Since this is a two-sided test (A1), the p-value is $2P(N(0,1) > |z_n|)$ *where* $z_n = (1.7 - 1.5)/(0.39/\sqrt{40})$ *since under the null hypothesis* $(\bar{X}_{40} - 1.5)/(0.39/\sqrt{40}) \sim N(0,1)$*. The p-value is therefore*

```
z = (1.7 - 1.5) / (0.39/sqrt(40))
2 * (1 - pnorm(abs(z)))
```

```
## [1] 0.001181281
```

b) Based on the p-value, should we reject H_0?

Since the p-value of $0.0012 < 0.01$*, we reject* H_0 *in favor of* H_a *at the 0.01 significance level.*

Power (Part III of lab)

The power of a test is the probability we will reject H_0 when it is false. If $H_0 : \mu = \mu_0$ is false, a value of μ contained in the range of the alternative hypothesis must be the correct mean of $\bar{X}_n$. The power of a test cannot be calculated without specifying an alternative value of μ contained in the range of H_a. We call this value μ_a. The power of a test is calculated under the assumption that $\bar{X}_n \sim N(\mu_a, \sigma/\sqrt{n})$ is the correct distribution of the sample mean.

Generally μ_a is chosen so that $|\mu_0 - \mu_a|$ represents the smallest difference between μ_0 and μ for which you would like to reject H_0 at the calculated power level.

Problem 2

In reference to problem 1, suppose that NYC officials don't mind if the population mean amount of hard alcohol in a mixed drink in NYC is a bit higher than 1.5 oz, but they do want to make sure that they can detect when it is in truth 1.65 oz (or higher). That is, they want to make sure that when the population mean is in fact 1.65oz, their testing will have a high probability of revealing this fact. Thus, to calculate the power of the test, we'll set $\mu_a = 1.65$ oz. and $\bar{X}_{40} \sim N(1.65, 0.39/\sqrt{40})$.

a) Approximate the power of the test performed in Problem 1 using the following steps:

```
i) Simulate 100,000 independent samples from the distribution of $\bar X_{40}$ under the alternative hyp
```

```r
set.seed(1)
n=40
nsim = 100000
sampmean.alt = rnorm(nsim, 1.65, 0.39/sqrt(n))
```

ii) Based on these 100,000 samples, estimate the probability that the sample mean of the study will fall into the rejection region determined in Problem 1(b). This is the power of the test when the sample size, n, is equal to 40.

Recall that the rejection region is RR: $\bar{x}_{40} \geq 1.66$ or $\bar{x}_{40} \leq 1.34$. So, we will count how many times our simulated values fall into the rejection region and divide by the total number of simulated values.

```r
n=40
lowercutoff = qnorm(.005, 1.5, 0.39 / sqrt(n))
uppercutoff = qnorm(.995, 1.5, 0.39 / sqrt(n))
# recall that | means 'or'...
num_reject = sum((sampmean.alt <= lowercutoff) | (sampmean.alt >= uppercutoff))
power = num_reject / nsim
power
```

```
## [1] 0.44329
```

b) Calculate the actual power of this test, $P(\bar{X}_n \leq 1.34 \text{ or } \bar{X}_n \geq 1.66 \mid \mu = 1.65)$ using the `pnorm()` function in R.

```r
n=40
lowercutoff = qnorm(.005, 1.5, 0.39 / sqrt(n))
uppercutoff = qnorm(.995, 1.5, 0.39 / sqrt(n))
pnorm(lowercutoff, 1.65, 0.39/sqrt(n)) + (1 - pnorm(uppercutoff, 1.65, 0.39/sqrt(n)))
```

```
## [1] 0.4430237
```

c) Re-run the code from part (b) using different values of n. By trial and error, find the smallest value of n such that the power of the test is greater than or equal to 0.9.

```r
n=101
lowercutoff = qnorm(.005, 1.5, 0.39 / sqrt(n))
uppercutoff = qnorm(.995, 1.5, 0.39 / sqrt(n))
pnorm(lowercutoff, 1.65, 0.39/sqrt(n)) + (1 - pnorm(uppercutoff, 1.65, 0.39/sqrt(n)))
```

```
## [1] 0.9013891
```

n=101 for the power of this test to be at least 0.90.