Strategies on how to approach questions for hypothesis testing

Abraham Arbelaez

Steps for hypothesis testing

1. Identify all the variables in the question.
2. Once you identify all the variables, find the alternative hypothesis (might be underlined or in bold, as a hint).
3. After writing down your alternative hypothesis, write down your null hypothesis by stating the opposite of what you have for the alternative.
4. Once you have the steps from above, you will follow one of three paths:
   a. For mean: Find the t-test statistic by following the code in R: `t = (xbar-mu0)/(s/sqrt(n))`. Once you have the number, you have three different options:
      i. If your alternative hypothesis is $\mu < \mu_0$ is a left-tailed, we calculate the p-value by using `pt(t, n-1)`.
      ii. If your alternative hypothesis is $\mu > \mu_0$ is a right-tailed, we calculate the p-value by using `1-pt(t, n-1)`.
      iii. If your alternative hypothesis is $\mu \neq \mu_0$ is a two-tailed, we calculate the p-value by finding the smaler between the left-tailed and right-tailed and multiply it by 2. Therefore, we could have `2*pt(t, n-1)` or `2*(1-pt(t, n-1))`.
   b. For proportion: Make sure all assumptions hold. That is, $np_0 \geq 5$ and $n(1 - p_0) \geq 5$. Once you verify both, you have to find the z-test statistic by following the code in R: `z = (phat-p0)/sqrt((p0*(1-p0))/n)`. Once you have the number, you have three different options:
      i. If your alternative hypothesis is $p < p_0$ is a left-tailed, we calculate the p-value by using `pnorm(z, 0, 1)`.
      ii. If your alternative hypothesis is $p > p_0$ is a right-tailed, we calculate the p-value by using `1-pnorm(z, 0, 1)`.
      iii. If your alternative hypothesis is $p \neq p_0$, we calculate the p-value by finding the smaler between the left-tailed and right-tailed and multiply it by 2. Therefore, we could have `2*pnorm(z, 0, 1)` or `2*(1-pnorm(z, 0, 1))`.
   c. For variance: Check assumptions of normal data by using `shapiro.test()`. If the p-value of the test is high, there is no evidence the data are not normal. Find the test statistic by following the code in R: `chi = ((n-1)*s^2)/sigma0^2`. Once you have the number, you have three different options:
      i. If your alternative hypothesis is $\sigma^2 < \sigma_0^2$ is a left-tailed, we calculate the p-value by using `pchisq(chi, n-1)`.
      ii. If your alternative hypothesis is $\sigma^2 > \sigma_0^2$ is a right-tailed, we calculate the p-value by using `1-pchisq(chi, n-1)`.

iii. If your alternative hypothesis is $\sigma^2 \neq \sigma_0^2$, we calculate the p-value by finding the smaler between the left-tailed and right-tailed and multiply it by 2. Therefore, we could have `2*pchisq(chi, n-1)` or `2*(1-pchisq(chi, n-1))`.

5. Once you have your p-value, contrast it to the significance level (α) to have a decision rule.

   a. If p-value $< \alpha$, we **reject the null hypothesis**. This means, that we have enough evidence to reject our null hypothesis, in other words, our alternative hypothesis is true.

   b. If p-value $> \alpha$, we **do not reject the null hypothesis/fail to reject the null hypothesis**. This means, that we don't have enough evidence to reject the null hypothesis. In other words, our alternative hypothesis is false.

6. While we may be confident in our statistical tests, there may be cases that even though our numbers say one thing, real life says something different. We can incur in false positives and false negatives.

- Type Error I: Happens when we reject the null hypothesis but the null hypothesis in real life is true.
- Type Error II: Happens when we do not reject the null hypothesis, but in real life the alternative hyppothesis is true.

Examples following the strategies

- *Suppose a beverage company claims that the mean caffeine content in its energy drink is 100 mg per 8 oz can. However, a consumer advocacy group believes that the actual mean caffeine content is higher. A random sample of 30 cans of the energy drink is selected, and the caffeine content of each can is measured obtaining a sample mean of 105 mg with a sample standard deviation of 10 mg. Suppose our significance level is* $\alpha = 0.05$

1. We first identify all the variables in the question. So:

- $\bar{x}$ is the sample mean: 105
- s is the sample standard deviation: 10
- n is the sample size: 30
- μ_0 is the specific value that we want to test: 100
- α is the significance level: 0.05

2. Once we identify all the variables, we find the alternative hypothesis. Our alternative hypothesis is the actual mean caffeine content is higher than 100.

3. After writing down the alternative hypothesis, we can write down our null hypothesis by stating the opposite of what we already have. Our null hypothesis is the actual mean caffeine content is less or equal to 100. We can also implement our correct notation:

- $H_o : \mu \leq 100$
- $H_a : \mu > 100$

4. Since we are calculating the mean. We will follow the steps for the mean with the test statistic. Given that our alternative hypothesis is a right-tailed (case b), we will follow the code `1-pt(t, n-1)`.

```
# saving our parameters
xbar = 105
s = 10
n = 30
mu0 = 100

# test statistic
t = (xbar-mu0)/(s/sqrt(n))

# p-value
1 - pt(t, n-1)
```

```
## [1] 0.005218695
```

5. Our p-value is 0.00521. Since our significance level is $\alpha = 0.05$, we contrast it to our p-value. Our p-value is lower than our α $(0.05 > 0.00521)$, which means that we **reject the null hypothesis**, concluding that at significance level 0.05, the true mean of caffeine in the beverages is greater than 100 mg.

- *The administration at Kansas State University claims that the proportion of undergraduate students who graduate within four years is at least 50%. However, a group of faculty members suspects that the actual proportion of students graduating within four years is lower. A random sample of 400 undergraduate students at K-State is selected, and their graduation status within four years are recorded. Out of the 400 students, 180 graduated within four years. Suppose the significance level is 0.10*

1. We first identify all the variables in the question. So:

- $\hat{p}$ is the sample proportion: $\frac{180}{400} = 0.45$
- n is the sample size: 400
- p_0 is the specific value that we want to test: 0.5
- α is the significance level: 0.10

2. Once we identify all the variables, we find the alternative hypothesis. Our alternative hypothesis is whether the actual proportion of students graduating within four years is lower than 50%.

3. After writing down the alternative hypothesis, we can write down our null hypothesis by stating the opposite of what we already have. Our null hypothesis is whether the actual proportion of students graduating within four years is greater or equal to 50%. We can also implement our correct notation:

- $H_o : p \geq 0.5$
- $H_a : p < 0.5$

4. Since we are calculating the proportion. We will follow the steps for the proportion with the test statistic. Given that our alternative hypothesis is a left-tailed (case a), we will follow the code `pnorm(z, 0, 1)`.

```r
# saving our parameters
phat = 0.45
n = 400
p0 = 0.5

# test statistic
z = (phat-p0)/sqrt((p0*(1-p0))/n)

# p-value
pnorm(z, 0, 1)
```

```
## [1] 0.02275013
```

5. Our p-value is 0.02275013. Since our significance level is $\alpha = 0.10$, we contrast it to our p-value. Our p-value is lower than our α ($0.10 > 0.02275013$), which means that we **reject the null hypothesis**, concluding that at significance level 0.10, the true proportion of students that graduate within four yours is lower than 50%. In other words, the administration was wrong (cue dramatic music).