Intro to Hypothesis Testing Exercises

Pages	Suggested Reading
371 – 376	Sections 9.3, 9.6, 9.7, and 9.8

Pages	Problems
396 – 406	(***Section 9.16***) 2, 3, 34, 41, 44

For all of the following, assume that the tests described are operating at an α = 5% level; that is, a P – value that comes in under 5%will indicate that we should believe the research hypothesis (i.e., "something changed"), whereas a P – value above 5% means we shouldn't believe the research hypothesis (i.e., "nothing changed").

E1: The National Center for Education Statistics monitors many aspects of elementary and secondary education nationwide. Their year 1996 survey numbers are often used as a baseline to assess changes. In 1996, 34% of students were present in school every day during the previous month (the "perfect attendance" rate). In the year 2000, many school administrators and teachers thought that rate had fallen. A study was done, and it was found that a significantly lower percentage of students had attended school every day ($P = 0.027$).

a. What is H_1, the research hypothesis?
b. What is H_0, the null hypothesis?
c. What's the P – value? That is, what's the chance that these researchers saw rates as least as extremely low as they did, assuming that the rate, indeed, hadn't changed significantly from 34%?
d. Based on the previous answers, does the "perfect attendance" rate appear to have significantly dropped?
e. What type of error could we have made (Type I or Type II)?
f. Define the error you identified in part e.

E2: In the article "Antithrombotic Potential of Grape Juice and Red Wine for Preventing Heart Attacks" (Folts, John D., *Pharmaceutical Biology*, Dec98 Supplement, Vol. 36, p21), a study involving red and white wine was conducted. In particular, the researchers wanted to see if the consumption of these wines could aid in the treatment of atherosclerosis (hardening of the arteries). The subjects in the study first had their blood tested for "blood platelet aggregation" (that is, the clumping of platelets in the blood). Then, they consumed 5 ml/kg of either red or white wine. Then, they had their blood tested again for aggregation. Platelet aggregation decreased ($P < 0.01$) after the red wine but there were no significant changes after the white wine ($P > 0.05$). If platelet aggregation is lessened, then the treatment effect shows potential for helping lessen the effects of atherosclerosis.

a. With both the red and white wine portions of the study, what is H_1?
b. With both the red and white wine portions of the study, what is H_0?
c. ***p***(the study got lowered blood platelet aggregation | the **red** wine had no lowering effect)
d. ***p***(the study got non – lowered blood platelet aggregation | the **white** wine had no lowering effect)
e. Based on answer **c**, does **red** wine appear to be significantly lowering blood platelet aggregation?
f. Based on answer **d**, does **white** wine appear to be significantly lowering blood platelet aggregation?
g. What type of error could you have made with respect to **red** wine?
h. Define that error, in context.
i. What type of error could you have made with respect to **white** wine?
j. Define ***that*** error.

E3: Atrial Fibrillation (AF) is the most common heart rhythmic disorder, affecting 2.3 million people in the US alone. The sufferers of AF are 500% times more likely to suffer a stroke (when compared to the general population), as AF makes it more likely for clots to form in the atria. A new experimental anticoagulant drug called Rivaroxaban was tested for efficacy in staving off stokes in AF sufferers. In the study ("Phase III ROCKET AF Study of Rivaroxaban meets its primary efficacy endpoint with Comparable Safety vs. Warfarin", 2010**)**, Rivaroxaban's efficacy was compared to that of Warfarin, the "status quo" anticoagulant drug, in many categories. Two were 1) efficacy in reducing strokes, and 2) tendency to decrease the blood's ability to coagulate.

During the double – blind study, more than 14,000 sufferers of AF were randomly assigned to one of two groups. One group received a course of Rivaroxaban, the other a course of Warfarin.

After the 4 – year study, Rivaroxaban was found to reduce the relative risk of stroke (RRR) better than Warfarin (P = 0.015). The rates of minor and major bleeding with subjects using Rivaroxaban were similar to those using Warfarin (P=0.442 and P=0.576, respectively).

a. This one's got three different sets of null and research hypotheses; see if you ID which is which:

(with respect to RRR)

Rivaroxaban has a lower RRR than Warfarin.

Rivaroxaban has the same RRR as Warfarin.

(with respect to rates of minor bleeding)

Rivaroxaban has the same rate of minor bleeding as Warfarin.

Rivaroxaban has a different rate of minor bleeding as Warfarin.

(with respect to rates of major bleeding)

Rivaroxaban has a different rate of major bleeding as Warfarin.

Rivaroxaban has the same rate of major bleeding as Warfarin.

b. ***p***(Rivaroxaban's RRR was lower than Warfarin's | Rivaroxaban has the same RRR as Warfarin)

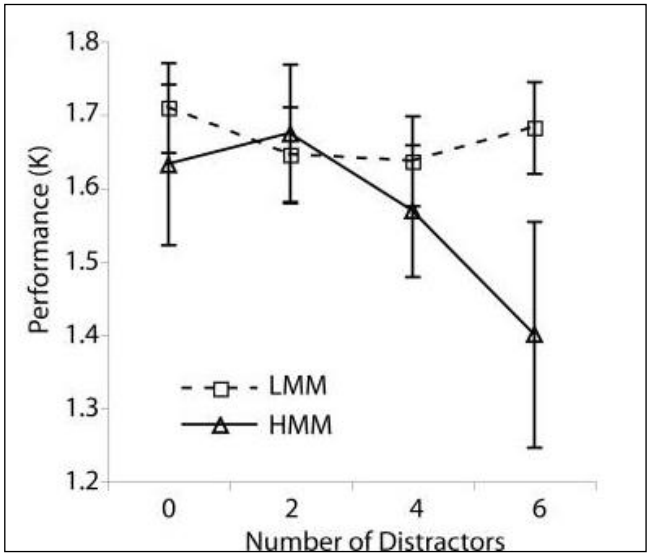
c. ***p***(Rivaroxaban's rate of minor bleeding was similar to Warfarin's | Rivaroxaban has the same rate of minor bleeding as Warfarin)

d. ***p***(Rivaroxaban's rate of major bleeding was similar to Warfarin's | Rivaroxaban has the same rate of major bleeding as Warfarin)

e. ID and define which errors could have been made in each case.

E4: In the article **"**Cognitive Control in Media Multitaskers" (Ophir, et. Al., Stanford University, 2010), subjects were broken into two groups; Light Media Multitaskers (LMM) and Heavy Media Multitaskers (HMM), and their ability to mentally process information while multitasking was analyzed. One component of the study was "filtering environmental distractions" (i.e., being able to recognize important information when confronted with unimportant "distractors").

HMM performance was inversely proportional to the number of distractors ($P < 0.01$), while LMM performance was essentially constant, even as distractors increased ($P > 0.68$). See diagram at right for data from study[1].



a. ID both sets of hypotheses (for HMM and LMM)
b. ***p***(HMM were negatively influenced by distractors | HMM are not negatively influenced by distractors.
c. ***p*** (LMM were not negatively influenced by distractors | LMM are not negatively influenced by distractors.
d. ID and define which errors could have been made in each case.

Let's do a little error analysis of our own:

- A bike light company is conducting a test to see if their new LED tail light results in an increased visibility of its users. Their previous tail light was seen at a distance of 100 meters approximately 50% of the time. They want to check to see if the new tail light results in a higher "percentage of visibility." Their null hypothesis is that the new light is no better than the old one, visibility – wise.

  E5: ID and define each type of error that ***could*** occur after this test.

  E6: In your opinion, which type of error is potentially worse?

- Two drugs are being compared for effectiveness in treating the same condition. Drug 1 is very affordable, but Drug 2 is extremely expensive. The null hypothesis is "both drugs are equally effective," and the alternate is "Drug 2 is more effective than Drug 1."

  E7: ID and define each type of error that could occur after this test.

  E8: In your opinion, which type of error is potentially worse?

[1]You might say, "Hey wait a minute! Those confidence intervals are overlapping! You can't make any conclusions based on overlapping CI's!" Well, if you remember, when CI's don't overlap, we claim to have found a significant difference. However, when they do overlap, we have to say we're not sure...but hypothesis tests (in this case, an ANOVA) are able to hone in on differences that CI's can't. More to come!

- Two drugs are known to be equally effective for a certain condition. They are also each equally affordable. However, there is some suspicion that Drug 2 causes a serious side-effect in some patients, whereas Drug 1 has been used for decades with no reports of the side effect. The null hypothesis is "the incidence of the side effect in both drugs is the same", and the alternate is "the incidence of the side effect in Drug 2 is greater than that in Drug 1."

  E9: ID and define each type of error that could occur after this test.

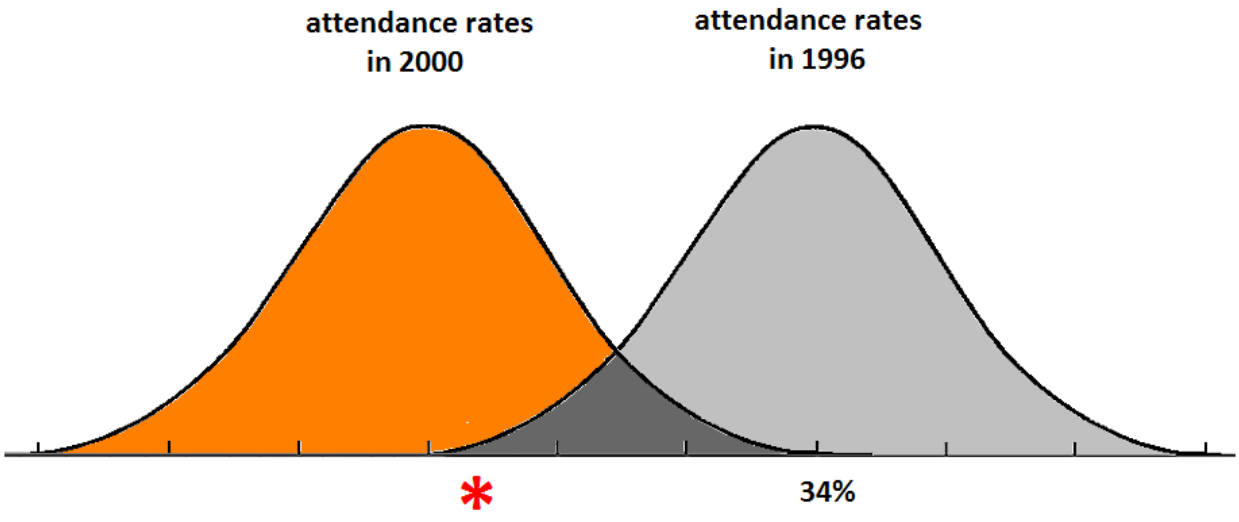
  E10: In your opinion, which type of error is potentially worse?

E11: Here's a ***great*** question that I hear every term: "Sean, what if the P – value comes in right ***at*** 5%? What do we do then?" I have an extension to that question: What if the P – value comes in ***close*** to 5% (say, 4% or 6%)? I mean, if it's below 5%, then you should believe H_1; otherwise, you don't…but what if it's so close that it seems to be inconclusive? How would you respond to these questions?

Answers.

E1:

a. Faculty and administrators are interested in seeing if the perfect attendance rate has fallen, H_1 would be that the perfect attendance rate is below 34%, the baseline from 1996.
b. H_0 would be that the perfect attendance rate is not below 34% (or, it's the same as 34%).
c. 2.7%. This P – value is a likelihood, based on the assumption that the perfect attendance rate is still at 34%, that we got so much of a lower rate. In other words, there's not much of a chance of us seeing data that low on the curve.
d. It appears to have gone down significantly; there's only a 2.7% chance that it fell due to random fluctuation. That is, if the rate were still 34% (as it was in 1996), there's only a 2.7% chance that we will have gotten such low data. Consider the following diagram, to see the logic:



Assuming our data point falls at the *, ask yourself, "what's more likely: that the data is in the grey curve, or the orange one?" Remember that likelihood in these curves is given by how high the curve is above your data point. So, since the orange curve is much higher above the asterisk than the gray curve, it must be that the data is more likely from the orange population than the gray one (in fact, the P – value tells us that there's only a 2.7% chance it comes from the gray curve). Since the orange population has a lower center than 34%, we can conclude that the center (that is, the enrollment percentage) has shifted downward in 2000, and is significantly lower than it was in 1996.

e. Since you acted on a small P – value (that is, you said something changed), you could have committed a false positive error, or Type I.

f. Saying that the attendance rates are lower, when, in fact, they are not. Remember, it's still *possible* (but not ***probable***) that you're in the gray curve above.

E2.

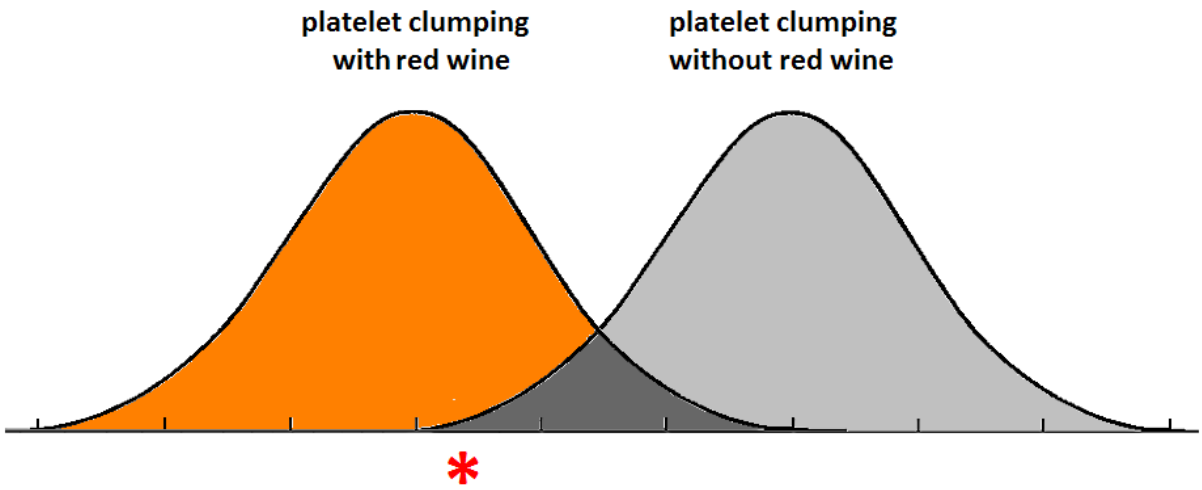
a. H_1 : the wine decreases platelet aggregation.

b. H_0? the wine does not decrease platelet aggregation (no reason to think it does unless data shows it).

c. ***p***(we got data as least as extreme as our data | the **red** wine had no lowering effect) < 1% (*are you OK with the conditional probability "|" I used there? Thanks.* ☺)

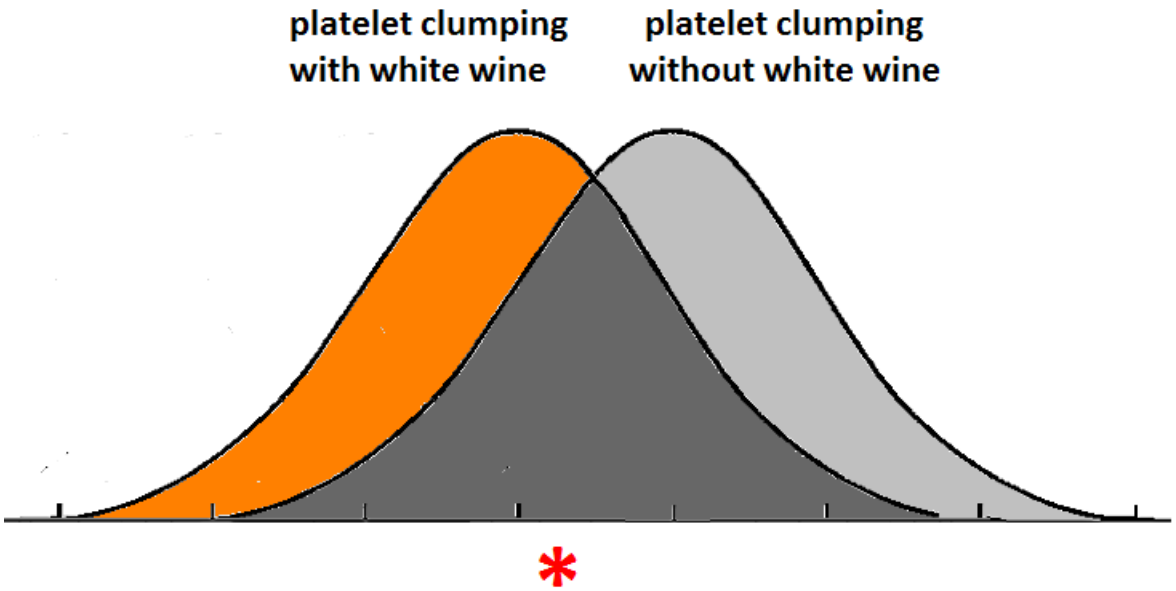
d. ***p***(we got data as least as extreme as our data | the **white** wine had no lowering effect) > 5%

e. It appears that red wine is significantly effective at reducing platelet clumping; from part a., there's only a 1% chance that our data would have come from a population where red wine was NOT being effective:



Like in the previous question, assume your data falls at the red asterisk. What's more likely: that the low platelet clumping we saw after drinking red wine was just "lucky" (that is, red wine had no effect on the clumping, and our data landed in the tail of the gray curve), or that the platelet clumping was lower due to the fact that the red wine ***caused*** it to be lower (in which case it would land near the center of the orange curve, which is centered lower than the gray curve)?

f. The white wine shows ***no*** promise is reducing clumping; the data fell within 2 standard deviations of the "had no effect" population, so there's no reason to believe it HAS an effect. We can also look at this one graphically:



What the high P – value implies is that it's impossible to tell which distribution the data comes from. In other words, any difference between those two curves isn't significantly different, and we must consider them to be statistically exactly the same curve. Any time you get a *P* – value higher than 5%, you conclude that the effect that you are studying has no effect at all on your data. So, in this case, drinking white wine has no measurable effect on platelet clumping (in fact, this study went even further to explain why. Check it out if you're interested)

g. Type I (since you got a small P – value)
h. Saying that red wine reduces platelet aggregation, when in fact, it doesn't.
i. Type II (since you got a large P – value)
j. Saying that white wine doesn't reduce platelet aggregation, when in fact, it does.

E3: a. Null hypotheses always assume nothing's different, and research hypotheses always say something changed, so:

(with respect to RRR)

H_1: Rivaroxaban has a lower RRR than Warfarin.
H_0: Rivaroxaban has the same RRR as Warfarin.

(with respect to rates of minor bleeding)

H_0: Rivaroxaban has the same rate of minor bleeding as Warfarin.
H_1: Rivaroxaban has a different rate of minor bleeding as Warfarin.

(with respect to rates of major bleeding)

H_1: Rivaroxaban has a different rate of major bleeding as Warfarin.
H_0: Rivaroxaban has the same rate of major bleeding as Warfarin.

a. ***p***(we got our data | Rivaroxaban has the same RRR as Warfarin) = 0.015…looks like the Rivaroxaban's working to lower RRR, eh?

b. ***p***(we got our data | Rivaroxaban has the same rate of minor bleeding as Warfarin) = 0.442

c. ***p***(we got our data | Rivaroxaban has the same rate of minor bleeding as Warfarin) = 0.576 … these last two P – values are basically coin flips, so it's impossible to determine whether or not one drug causes more bleeding than the other (this is what Rivaroxaban's producers want to show: that it's more effective than Warfarin, with no additional side effects).

d. For RRR, you could have made a Type I error (saying Rivaroxaban reduces RRR when, in fact, it doesn't). For both types of bleeding rates, you could have made a Type II error (saying that there is no difference in bleeding rates between the two drugs, when, in fact, there is). Which type of error, do you think, is worse in this case?

E4: a.

(HMM)

H_0: HMM's performance is not influenced by increased numbers of distractors.

H_1: HMM's performance is influenced by increased numbers of distractors.

(LMM)

H_0: LMM's performance is not influenced by increased numbers of distractors.

H_1: LMM's performance is influenced by increased numbers of distractors.

b. P(we got our data | HMM are not negatively influenced by distractors) < 0.01….agrees with the statement "*HMM performance was inversely proportional to the number of distractors*."

c. P(we got our data | LMM are not negatively influenced by distractors) > 0.68…which agrees with "*LMM performance was essentially constant, even as distractors increased*."

d. For the HMM's, we could have made a Type IU error (saying that their performance went decreased, when, in fact, it didn't). For the LMM's, Type II (saying their performance was unchanged, when, in fact, it wasn't).

E5: **Type I** – saying the light has increased visibility when, in fact, it doesn't. **Type II** - saying the light doesn't have increased visibility when, in fact, it does.

E6: My vote is for Type I; if you erroneously think that a light is more visible, you might put yourself in harm's way more – a type II error result would only make you ***no less cautious***.

E7: **Type I** – saying Drug 2 is more effective than Drug 1 when, in fact, it isn't. **Type II** - saying Drug 2 is not more effective than Drug 1 when, in fact, it is.

E8: I think Type 1 is much worse in this case – if you're going to pay a LOT more for a drug, it had most likely work! Thus, I'd actually change the significance of this test to less than 5% to make it harder to "prove" that it's working better (a Type 2 error has a smaller impact in this case, assuming Drug 1 actually ***does*** treat the condition it's designed to treat).

E9: **Type I** – saying Drug 2 has worse side effects when, in fact, it doesn't. **Type II** - saying Drug 1 and Drug 2 have the same side effects.

E10: All things being equal, a Type 2 error in this case would be worse. If you get a false positive (Type I) result, than more testing would be done, and, hopefully, it would be realized that the results was in error. However, a Type 2 would allow a drug with awful side effects out into the market. Yuck. So, here's one of those cases, where I'd want to minimize the chance of a Type II by increasing the chance of a Type I (i.e., I'd lower confidence from 95% to, say, 90%. We talked about why this worked in class, right? Ask me about it if we didn't).

E11: The quick and easy answer is "get more data!"

Why does this work? Well, suppose you're testing a research hypothesis that states that a new cross country ski wax (A) creates preferable glide in a double blind "taste test". You aim to test this new wax against another leading wax (B) and see how they compare:

H_0: A and B are equally preferred for "glide".
H_1: A is preferred more for glide than B.

You run a double blind test with a small number of skiers (say 20) and, of those 20, 14 (or 70%) prefer wax A. This gives your study a P – value of about 4%. So, protocol tells you that you should believe H_1, right? But, there's something nagging at you…I mean, 4% is ***so close*** to 5%. What if it was just luck?

Now, suppose that, in your heart of hearts, you truly ***do*** believe that wax A is superior, and you're confident that, if you run the test again, regardless of how many people you ask, people will still prefer wax A over wax B. So, this time, you survey 50 skiers, and guess what? Exactly 70% of them (or 35) prefer wax A to wax B…and now, your P – value is way less than 5% (it's about 0.002).

And, at the same time you're surveying 50 people, a friend of yours is testing the same hypotheses, but he only asks 10 skiers. Lo and behold, 7 of the 10 prefer wax A, but his P – value is over 10%, so he can't believe that wax A is preferable (by the $\alpha = 5\%$ guideline).

How can it be that the same result (70% prefer A in each study) can yield such wildly different results (in the form of P – values)? The only difference in those three cases was the sample size…and, by the central limit theorem, that's ***everything***!

Recently, we learned about confidence intervals. Awesome little tools, those. However, in this case, we need something a bit lopsided: we need 1- sided (not 2 – sided) CI's. Why? Because we want to show that Wax A is ***superior*** to wax B, not just *different*. 1 – sided CI's will do that for us[2].

Here's the data you and your friend collected, with P – values gotten in each case:

Sample Size (n)	Number preferring Wax A (x)	$\hat{p}$ (proportion in sample preferring Wax A)	P – value of result
20	**14**	**0.7**	**4%**
50	**35**	**0.7**	**0.2%**
10	**7**	**0.7**	**> 10%**

It appears (and it's true) that 70% does not always statistically equal 70%. Why? You know why! Because that 70% carries a ***margin of error***!

[2] Just so you know, the individual techniques of hypothesis testing take the place of 1 – sided CI's.

Here's the deal: if people really didn't care one way or the other between Wax A and Wax B, you'd expect them to (basically) flip a coin to decide which one they liked. In other words, the rate of Wax A's acceptance would be 50%. The question we have to address is this: of all of those 70%'s in the table above, which ones are within CI's that contain 50%? If any of their CI's contain 50%, then we can not safely say that a greater proportion prefer Wax A (remember the fishies swimming in their intervals).

Here's the same table, with a new column: the 1 – sided CI of the results. They'll look a little weird, as they're no longer symmetric (like 2 – sided CI's are), but the same logic applies: **we're 95% confident that this interval captures the true proportion of those who prefer Wax A**.

n	x	$\hat{p}_{\text{Wax A}}$	P – value of result	1 – sided CI for *p*, the proportion of those preferring Wax A[3]
20	**14**	**0.7**	**4%**	**(0.53, 1)**
50	**35**	**0.7**	**0.2%**	**(0.59, 1)**
10	**7**	**0.7**	**> 10%**	**(0.46,1)**

A – HA! You can now really ***see*** the tie – in between small P – values at their CI's! ***A small P – value means that you have escaped 2 standard deviations of the CI***. Notice how the CI's built around sample sizes of 20 and 50 don't contain 0.5? That means that we can be statistically certain that more than half of the skiers out there prefer Wax A (at least 53% and 59%, respectively). However, when your friend only sampled 10, he couldn't show that more than 50% of folks preferred Wax A; the proportion could be as low as 46%!

It appears as though, as the sample size gets larger, the margin of error gets smaller. Why is that?

Even better…if you pooled all of your results together (assuming that the samples were all independent and random), you'd get a P – value of about 0.0002 and a CI of (0.62, 1). Even more evidence why studies that can be replicated are looked upon more favorably.

[3] Feel free to ask me how you create this interval. It's not technically part of the class, but it's ***way*** important.

Intro to Hypothesis Testing Quizzes

Quiz 1.

In the article "Efficacy of multidisciplinary treatment in a tertiary referral headache centre." (Zeeberg P, Olesen J, Jensen R, *Cephalalgia*. 2005;25(12):1159.), a study was done to "set out to describe the procedures, characterize the patients and evaluate the treatment results" for a headache center. Patients in the study (n = 336) were divided into two groups: patients with medication overuse headaches (MOH) and those without (non – MOH). The study had two variables:

1. headache intensity for MOH vs. non – MOH for tension type headaches (TTH).
2. headache intensity for MOH vs. non – MOH for post traumatic headaches (PTH).

The study's null hypothesis, for each of these variables, was the same:

1. Headache intensity is the same for MOH vs. non – MOH with respect to TTH.
2. Headache intensity is the same for MOH vs. non – MOH with respect to PTH.

The *P* – value for the null about TTH was *P* < .05, while the P – value for the null about the PTH was P > 0.05.

1. (**2 points**) Let's find some P – Values. The first is "The chance that the researchers got headache intensity differences as extreme as they did between MOH and non – MOH with respect to TTH, when, in fact, there is no difference between these two with respect to TTH". Hint: it's up there in **bold**.

2. (**1 point**) Circle the correct choice: It **appears / does not appear** that there is a difference between headache intensity with respect to TTH, between MOH and non – MOH.

3. (**2 points**) Now – find "the chance that the researchers got headache intensity differences as extreme as they did between MOH and non – MOH with respect to PTH, when, in fact, there is no difference between these two with respect to PTH". Hint: it's up there in **bold**.

4. (**1 point**) Circle the correct choice: It **appears / does not appear** that there is a difference between headache intensity with respect to PTH, between MOH and non – MOH.

For both of these tests, the errors are the same: type 1 would be saying the headache intensities are different, when in fact they aren't, and type 2 would be saying the headaches intensities are the same, when in fact they're not.

5. (**1 point**) In your opinion, which type of error is worse?

6. (**3 points**) Why?

Quiz 2.

In the article "Learning and Attention Problems Among Children With Pediatric Primary Hypertension" (**Adams, Szilagyi, Gebhardt, Lande,** *PEDIATRICS* Vol. 126 No. 6 December 2010, pp. e1425-e1429), a study is outlined. The objective in the study was to "determine whether children with sustained primary hypertension are at increased risk for learning disabilities (LDs)."

The children involved in the study were classified as either hypertensive (HT) or not hypertensive, based on blood pressure testing at set intervals. In addition, the parents of the children told researchers whether or not their child had a documented learning disability (LD).

The study's null hypothesis was that HT children and non – HT children had the same rate of LD. The ***P*** – value for this null hypothesis was ***P*** = 0.0002.

1. (**2 points**) The chance that these researchers observed differences in LD rates data at least as variant as they did between the two groups (HY children and non – HT children), assuming that there was no difference between the LD rates of these two groups, is _____.

2. (**2 points**) Does it appear that children with HT have a different rate of LD than those children without HT?

3. (**2 point**) What type of error could we have made?

4. (**2 points**) Define that error, in the context of the problem.

5. (**2 point**) What is the probability of that error? Hint: Assuming you're using 95% confidence.

Quiz 3.

Do a little Googling and find a study (or some studies) that has (have) been done, and its (or their) corresponding P – value(s). You may not use ones that we already discussed in class! Select one P – value to use for this quiz!

1. (**1 point**) What is that P – value? By this, I mean what's the number (it's probably a percentage, but it might be a decimal).

2. (**2 points each**) What are the hypotheses that accompany this P – value? If it helps you, remember that the P – value assumes the null hypothesis is true!

3. (**2 points**) Based on the size of your P – value (greater than or less than 5%), which type of error could have been made?

4. (**2 points**) What would that error be, in context? By "in context", I mean don't just say "false positive" or "false negative" – actually tell me the context of this particular study!

5. (**1 point**) Provide the course for your study!

Quiz 4.

(**1 point for each error correctly identified**) If you'd like another HT quiz, start by looking at question 2 on page 396 of your text. Pick 5 of these claims ***other*** than a, c, d, or i, and then answer #2 for these 5 claims.