Testing Worksheet 4

Question 1 – Conceptual

Part A: What does a test-statistic quantify?

Part B: What does the 'significance level' refer to in hypothesis testing?

Part C: What is another name for a Type I error?

Part D: What is another name for a Type II error?

Part E: Explain why using $\alpha = 0.05$ as a strict cut-off for signifance is problematic.

Question 2 – Decision Error

Part A: What type of error occurs if H_0 is False but we fail to reject it?

Part B: What type of error occurs if H_A is True but we do reject it?

Part C: How does the sample size affect the test-statistic, and as a consequence the p-value, with everything else held constant?

Part D: How does the standard error affect the test-statistic, and as a consequence the p-value, with everything else held constant?

Conclusions for the 'Decision Making' Approach

Making conclusions with this approach will differ only slightly from what we did using 'Strength of evidence.' Use the significance level α you are given in a problem and compare the p-value to this. (If a significance level is not given use best judgement to pick one)

You will answer the following question:

Do we reject H_0 at significance level α?

If p-value $\leq \alpha \rightarrow$ **Yes**

If P-value $> \alpha \rightarrow$ **No**

Then we apply this answer to the context (what does H_0 and H_A actually say?) to answer the research question. Usually I like to see the following in the conclusion:

- a statement of the test-statistic value and the p-value
- mention of whether we reject H_0
- what that actually means in context of the study

Example: With a test-statistic value of Z = 1.62 and a p-value of .104, we fail reject the null hypothesis that the coin is biased. There is not sufficient data to indicate the coin is biased.

Question 3 – Piano Revisited

Georgianna claims that in a small city renowned for its music school, the average child takes less than 5 years of piano lessons. We have a random sample of 35 children from the city, with a sample mean of 4.6 years of piano lessons and a sample standard deviation of 2.2 years.

Evaluate Georgianna's claim using a hypothesis test at the 0.05 significance level.

The test-statistic and p-value from the previous worksheet is

```
T = (4.6 - 5) / (2.2 / sqrt(35))
T
```

```
## [1] -1.075651
```

```
pt(T, df=34)
```

```
## [1] 0.1448287
```

Do we reject the null hypothesis?

Write a conclusion to summarize our results.

Practical vs Statistical Significance

We have so far been talking about 'statistical significance' when it comes to test-statistics and p-values. What we are doing is basically going 'our results are so many standard errors away from the hypothesized value.'

In essence, we are saying there is (or there is not) a difference. But even if we reject a null hypothesis, it does not mean that an effect has any practical difference in the real world.

Examine the following scenarios related to coin flipping we've seen before. For both we have the same null and alternate hypotheses:

H_0: p = .5 (the coin is fair), H_A: p > .5 (coin is biased towards heads)

Scenario 1

We flip the coin 20 times and get 16 heads. (ignore conditions for now)

```
Z = (.8 - .5) / sqrt(.5 * .5 / 20)
Z
```

```
## [1] 2.683282
```

```
round(pnorm(Z, lower.tail=F), 4)
```

```
## [1] 0.0036
```

Scenario 2

We flip the coin 1 million times and get 501342 heads.

```
Z = (.501342 - .5) / sqrt(.5 * .5 / 1000000)
Z
```

```
## [1] 2.684
```

```
round(pnorm(Z, lower.tail=F), 4)
```

```
## [1] 0.0036
```

In both scenarios we get nearly the same test-statistic and the same p-value rounded to 4 decimal places. In the real world, are you going to notice the affect of the biased coin in scenario 1? Probably. What about scenario 2? Probably not. I am willing to bet most real world coins are actually probably more biased than the coin in scenario 2, and yet it is hard to tell when actually flipping them.

(continued on next page)

If we want an idea of practical significance – "HOW biased is the coin," we can pair hypothesis tests with confidence intervals.

Scenario 1

```
CI_lower = .8 - 1.96*sqrt(.8 * .2 / 20)
CI_upper = .8 + 1.96*sqrt(.8 * .2 / 20)
paste(round(CI_lower, 4), round(CI_upper, 4))
```

```
## [1] "0.6247 0.9753"
```

Scenario 2

```
CI_lower = .501342 - 1.96*sqrt(.501342 * (1-.501342) / 1000000)
CI_upper = .501342 + 1.96*sqrt(.501342 * (1-.501342) / 1000000)
paste(round(CI_lower, 4), round(CI_upper, 4))
```

```
## [1] "0.5004 0.5023"
```

Both scenarios have 'statistical significance' since there are low p-values. In scenario 1 we can see an estimated proportion of heads much larger than the .5 proportion of heads for a fair coin.

But in scenario 2 the actual difference between the CI estimates and a fair coin proportion of 0.5 is basically nothing. You will not notice the effects of the coin bias. A bias this small won't affect the # of heads flipped compared to a fair coin until you flip the coin 435 times.

```
1 / (.5023 - .5)
```

```
## [1] 434.7826
```

Effect Size

The actual difference between the statistic and the hypothesized value ($\mu - \mu_0$, $\mu_1 - \mu_2$, $p - p_0$, $p_1 - p_2$) is called the **effect size**. In lieu of having confidence intervals that give you a range of estimates of the actual parameter, seeing reports of the effect size can help us judge practical significance too.

Scenario 1: The effect size is $p - p_0 = 0.8 - 0.5 = 0.3$. 30% is usually meaninful in the real world regardless of the actual circumstances

Scenario 2: the effect size is $p - p_0 = 0.501342 - 0.5 = .001342$, which is so small that it won't really affect anything

Question 4 – ASA P-value Statement

Read the "ASA Statement on P-values" link on the course webpage and answer the following:

Part A: Explain in your own words what is meant by the following phrase: "Over time it appears the p-value has become a gatekeeper for whether work is publishable."

Part B: The articles says a common misconception is that 'P-values measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.' What *do* p-values measure?

Part C: Briefly explain some methods statisticians use to supplement (or even replace) using p-values. (at least one of these we've already seen)