

Homework 5

Kai Aragaki

Problem 1

A

Power can be calculated by:

$$Power = Pr\left(Z > C - \frac{\Delta}{\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}}\right) + Pr\left(Z < -C - \frac{\Delta}{\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}}\right)$$

So if we assume $\Delta = 1$ and both $\sigma = \Delta$, $n = m$ and set $Power = 0.8$ then

$$0.8 = Pr\left(Z > C - \frac{\Delta}{\sqrt{\frac{2\Delta^2}{n}}}\right) + Pr\left(Z < -C - \frac{\Delta}{\sqrt{\frac{2\Delta^2}{n}}}\right)$$

$$0.8 = Pr\left(Z > C - \frac{1}{\sqrt{\frac{2}{n}}}\right) + Pr\left(Z < -C - \frac{1}{\sqrt{\frac{2}{n}}}\right)$$

$$0.8 = Pr\left(Z > C - \sqrt{\frac{n}{2}}\right) + Pr\left(Z < -C - \sqrt{\frac{n}{2}}\right)$$

Though not given, it's reasonable to assume that our desired $\alpha = 0.05$, so our critical values C become

$$0.8 = Pr\left(Z > 1.96 - \sqrt{\frac{n}{2}}\right) + Pr\left(Z < -1.96 - \sqrt{\frac{n}{2}}\right)$$

We then set a threshold value that, when we draw from a normal distribution, we expect to get a value more extreme than it 80% of the time on average. To select such value, we assume that the probability of selecting a value less extreme than these 'thresholds' contributes very minimally to the over 80% power, so much so that we can approximate:

$$0.8 \approx Pr\left(Z > 1.96 - \sqrt{\frac{n}{2}}\right)$$

Finding this value is then simple:

```
z <- qnorm(0.2)
```

We use a normal distribution rather than a t-distribution as we are given the population standard deviation (σ) rather than the sample standard deviation (s)

The resultant Z is roughly -0.84. This is the value that when we sample from a $Normal(0, 1)$ distribution, we would expect to get values greater than it on average 80% of the time.

Calculating n , then:

$$Z > 1.96 - \sqrt{\frac{n}{2}}$$

$$2(-Z + 1.96)^2 < n$$

so,

```
n_min <- 2 * (-z + 1.96)^2
```

The minimum whole value of n is 16. That is the number of ELISAs we would need per group to obtain a power of 80%.

B

If we only ran 10 ELISAs per group, then

$$Power \approx Pr(Z > 1.96 - \sqrt{10/2})$$

So our threshold value is

```
z_min <- 1.96 - sqrt(10/2)
```

Giving a probability of

```
p <- pnorm(z_min, lower.tail = FALSE)
```

Approximately 61%

C

Our critical value then shifts to account for a single tailed hypothesis:

```
cv <- qnorm(0.05, lower.tail = FALSE)
```

So our threshold is now:

```
z_min <- cv - sqrt(10/2)
```

Giving a probability of

```
p <- pnorm(z_min, lower.tail = FALSE)
```

Approximately 72%

Problem 2

We know that to calculate a confidence interval with an α of 0.05, we use the following equation:

$$\bar{X} \pm 1.96\sigma/\sqrt{n}$$

We assume we know σ , as we know the standard deviation of the nation and have made the decision to assume that it is the standard deviation of the population of patients. We want each ‘arm’ of the confidence interval to be 1.5, but we don’t know how many patients need to be sampled (that is, n). We can solve for n in the above equation and plug in our population standard deviation. Additionally, we drop \bar{X} as it is not relevant (we are only interested in the width of the interval and not where it is centered).

$$1.5 = 1.96\sigma/\sqrt{n}$$

$$n = (1.96 * 4.5/1.5)^2$$

```
n_min <- (1.96*4.5/1.5)^2
```

The closest integer, rounding up, is 35 - so she should pull 35 patient records to produce a mean that she has 95% confidence captures the patient population mean, and additionally has a width of 3 BMI units.

Problem 3

A

```
a_dat <- c(55.2, 58.1, 41.7, 44.9, 44.8, 48.9, 47.5, 48.4, 51.6, 40.6, 48.0)
b_dat <- c(48.7, 52.6, 65.2, 70.4, 44.2, 54.7, 44.0, 66.5, 56.8)

tt <- t.test(a_dat, b_dat, alternative = "two.sided", var.equal = TRUE)
tt

##
## Two Sample t-test
##
## data: a_dat and b_dat
## t = -2.2729, df = 18, p-value = 0.03552
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -14.9046999 -0.5862092
## sample estimates:
## mean of x mean of y
## 48.15455 55.90000
```

By this method we reject the null hypothesis and gain confidence for the conclusion that these populations are truly different from one another.

B

```
wc <- wilcox.test(a_dat, b_dat, alternative = "two.sided")
wc

##
## Wilcoxon rank sum exact test
##
## data: a_dat and b_dat
## W = 27, p-value = 0.09518
## alternative hypothesis: true location shift is not equal to 0
```

By this method we fail to reject the null hypothesis.

C

If we assume that the response values drawn from each strain population are normally distributed, we reject the null hypothesis that these populations have the same population mean, and gain confidence that there is a differences of the populations means. If we use a signed rank test, which relaxes the assumption of normality of the populations, we no longer reach statistical significance, so in this instance we would fail to reject the null hypothesis. As we did not test for normality, we cannot say with these analyses alone which would be most ‘proper’.