Final Review Session Spring 2023

requested topics

- ANOVA
- Sign rank
- Power

Additional examples for review

ANOVA

The `melanoma` data set contains data on 205 patients from Denmark with malignant melanoma. You have joined a lab in which the principal investigator is interested in determining whether mean tumor thickness (mm) differs by patient status (1 = died from melanoma, 2 = alive, 3 = died from other causes).

```
head(melanoma)
```

```
##   time status sex age year thickness ulcer status2
## 1   10      3   1  76 1972      6.76     1       3
## 2   30      3   1  56 1968      0.65     0       3
## 3   35      2   1  41 1977      1.34     0       2
## 4   99      3   0  71 1968      2.90     0       3
## 5  185      1   1  52 1965     12.08     1       1
## 6  204      1   1  28 1971      4.84     1       1
```

ANOVA

```
anova <- aov(thickness ~ status2, data = melanoma)
tidy(anova)
```

```
## # A tibble: 2 x 6
##   term          df sumsq meansq statistic     p.value
##   <chr>         <dbl> <dbl>  <dbl>     <dbl>       <dbl>
## 1 status2         2  180.   90.2        11.3  0.0000216
## 2 Residuals     202 1606.    7.95      NA    NA
```

Your task is to help your PI analyze the results of this ANOVA run in R. For the ANOVA above, what are the null and alternative hypotheses?

ANOVA

```
tidy(anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic    p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>      <dbl>
## 1 status2       2  180.  90.2       11.3  0.0000216
## 2 Residuals   202 1606.   7.95      NA   NA
```

Based on the results shown, what would you conclude?

ANOVA

Based on these results what would be your next step in the analysis process? Justify your answer in 1-2 sentences. (would you continue your analysis with an additional test and if so, what test would you use)

ANOVA

What would you conclude from these results?

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = thickness ~ status2, data = melanoma)
##
## $status2
##           diff        lwr       upr     p adj
## 2-1 -2.0663511 -3.1192614 -1.013441 0.0000191
## 3-1 -0.5931955 -2.5792549  1.392864 0.7606857
## 3-2  1.4731557 -0.3970052  3.343316 0.1531399
```

Swimming

A group in the athletic department is working with the swim team. They implement a new training program and want to know if there has been an improvement in the 50m swim time (in seconds) at 12 weeks following the start of the program.

Swimming

They have collected the following data:

Pre program	Post Program
24.23	24.26
24.12	24.09
24.15	24.11
24.12	24.13
24.16	24.15
24.18	24.19
24.51	24.42
24.69	24.69
24.88	24.82
25.01	24.94
25.58	25.55
25.47	25.45
25.66	25.67

swimming

Calculate the appropriate test statistic. Show your work by writing the formula needed to calculate with values plugged in.

$$Z_T = \frac{T - \mu_T}{\sigma_T}$$

Where

$$\mu_T = \frac{n(n+1)}{4}$$

and

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

swimming

Write the line of code that would give you the appropriate p-value for this test statistic.:

swimming

```
pnorm(-1.922, mean = 0, sd = 1)
```

```
## [1] 0.02730288
```

swimming

Based on your findings would you recommend that the athletic department continue this training program? Why or why not?

Power

Data from the Framingham study allow us to compare the distribution of initial serum cholesterol levels for two populations of males: those who go on to develop coronary heart disease and those who do not. The mean serum cholesterol level in men has a standard deviation of σ=41 mg/100 ml.

The mean initial serum cholesterol level of men who eventually develop coronary heart disease is μ is 244 mg/100ml.

Since it is believed that the mean serum cholesterol for those who do not develop heart disease cannot be higher than the mean level for men who do, a once sided test conducted at the α=0.05 level of significance is appropriate. For this scenario, what is the probability of type I error?

Power

Type I error or α here is the probability of rejecting the null when in fact the null is true. Here we are setting alpha at 0.05 so the probability of making a type I error is 0.05 or 5%.

We presume a mean serum cholesterol of 219 among those who do not develop heart disease. If a sample size of 25 is selected from the population of men who do not go on to develop coronary heart disease, what is the probability of making a type II error?

Power

Remember that Beta (type II error) is evaluated under the condition that the alternative hypothesis is true. Here our alternative proposed mean is 219. We first need to find the value in actual measured units at which we would reject the null.

We can do this two ways:

Find the cuttoff in terms of Z score and then convert it to mg/100ml

Note that

$$Z = \frac{x - \mu}{(\sigma/\sqrt{n})}$$

Re-arrange to solve for the x

$$Z = \frac{x - \mu}{(\sigma/\sqrt{n})}$$

$$x = Z * (\sigma/\sqrt{n}) + \mu$$

Power

$$x = Z * (\sigma / \sqrt{n}) + \mu$$

```
Zalpha_cutpoint<-qnorm(0.05)
Zalpha_cutpoint
```

```
## [1] -1.644854
```

```
#convert
Zalpha_cutpoint_converted=Zalpha_cutpoint*(41/sqrt(25))+244

Zalpha_cutpoint_converted
```

```
## [1] 230.5122
```

Power

Or we can use R to give us the cutpoint in units of mg/100ml directly by modifying the qnorm statement

```
cutpoint_alpha<-qnorm(0.05, mean=244, sd=(41/sqrt(25)))

cutpoint_alpha
```

```
## [1] 230.5122
```

note also that we are looking at qnorm of 0.05 here because our hypothesized mean (219) is lower than the null mean (244) so we are interested in the lower tail (the default in R).

Power

Now that we have this cutpoint in terms of the value we are measuring, we can find where this is on the distribution under the scenario where the alternative hypothesis (219) is the truth.

```
pnorm(cutpoint_alpha, mean=219, sd=(41/sqrt(25)), lower.tail=FALSE)
```

```
## [1] 0.08017032
```

This represents the probability of failing to reject the null when we should reject the null.

Note that we are interested in the upper tail here, because the distribution of our alternative hypothesis (mu=219) is lower than (to the left of) the null hypothesis distribution so the cutpoint (230.5) at which we would reject the null is on the right side of our alternative hypothesis distribution.

Power

Note that here again we are using R to give us probability relative to the distribution of means in units of mg/100ml.

We could instead have converted the cutpoint to Z units (this time with respect to the alternative distribution):

$$Z = \frac{230.5122 - 219}{41/\sqrt{25}}$$

```r
(230.5122-219)/(41/sqrt(25))
```

```
## [1] 1.403927
```

```r
pnorm(1.403927, lower.tail=FALSE)
```

```
## [1] 0.08017029
```

Power

What is the power of the test?

Remember that Power is the probability that you reject the null when the hypothesized alternative is true, it is the complement of type II error.

Power is 1-β

so power here is 1-0.08 or 0.92

This means that if the alternative of 219 is true and we draw a 25 person sample, then when we test against a null hypothesis of 244, we would correctly reject the null 92% of the time at an α of 0.05.

Power

How could you increase the power?

Power

The easiest way to increase the power here is to increase the sample size.

Let's say we wish to test the null hypothesis mu $=$ 244 mg/100ml against the one sided alternative hypothesis that mu $<$ 244 mg /100ml at the alpha $=$ 0.05 level of significance. If the true population mean is as low as 219 mg/100ml, and you want to risk only a 5% chance of failing to reject the null when the null should be rejected. How large a sample would be required?

Power

Here we are looking for the n, so we need to start with finding the cutpoint (in terms of Z) at which we would reject the null with an alpha of 0.05 and a one sided test.

```
qnorm(0.05)
```

```
## [1] -1.644854
```

Note that we keep the negative here because we are interested in the lower tail.

Power

Now we re-arrange our Z equation to put x (the cutpoint) on one side of the equation:

$$-1.645 = \frac{x - 244}{41/\sqrt{n})}$$

$$x = -1.645 * (41/\sqrt{n}) + 244$$

Now we look for the cutpoint (in terms of Z) for Beta. In our problem we are now setting the Beta to 0.05 (this is fairly stringent - many studies default to a 0.2 for Beta)

```
qnorm(0.05, lower.tail=FALSE)
```

```
## [1] 1.644854
```

Keep note of which side of the distribution we are working with.

Power

We will re-arrange this Z equation to put x (the cutpoint) on one side of the equation:

$$1.645 = \frac{x - 219}{41/\sqrt{n})}$$

$$x = 1.645 * (41/\sqrt{n}) + 219$$

Since we know that the cutpoints must have the same value in real units, we can now set these two equations equal to each other:

$$-1.645 * (41/\sqrt{n}) + 244 = 1.645 * (41/\sqrt{n}) + 219$$

and now there is only one variable that is unknown (n) which we can solve for.

```r
(41/(-25/(-1.645-1.645)))^2
```

```
## [1] 29.1125
```

Power

How would the sample size change if you were willing to risk a 10% chance of failing to reject a false null hypothesis?

If we were less stringent, we would need a smaller sample size.

You can check this by finding the Z for this other Beta:

```r
qnorm(0.1, lower.tail=FALSE)
```

```
## [1] 1.281552
```

Power

Substituting this in to our previous calculations we would get:

$$-1.645 * (41/\sqrt{n}) + 244 = 1.282 * (41/\sqrt{n}) + 219$$

which we can solve for n.

```r
(41/(-25/(-1.645-1.282)))^2
```

```
## [1] 23.04269
```

So we would require a sample size of 24 under these criteria