2 Calculations Exercises

6/9/2020

General instructions.

There are 5 exercises below, and you will be required to provide five solutions, each worth ten points. You have three options for completing the exercises.

1. In the past, I've required that R and SAS be used for at least one solution each. If you wish to delevop skill in both languages, you can use this option. You can provide both R and SAS solutions for the same exercises (e.g. R and SAS code for exercise 1), and then provide three solutions by choosing among the remaining four exercises (e.g. 2,4,5) in the language of your choice.
2. We have effectively two cohorts in these classes (600, 601, 602). One cohort is expected to learn SAS, one will not be using SAS in their program. Thus, for this summer, I will not require solutions in both languages. You can use either language to complete all the exercises. You will be *allowed* to submit homework as in past summers (point 1), but not *required*.
3. Last summer I allowed that one solution could be implemented in Python. I will continue that practice this summer. You can embed Python code in R Markdown, using the syntax `{python}` instead of `{r}`. If you choose this, you will be expected to comment on the differences or similarities between R/SAS and Python, and I would prefer your Python solution to be included in R Markdown. I won't be teaching Python this summer, but if you're familiar with Python, this may help understand the inner workings of R or SAS.

Exercise 1

Cohen gives a formula for effect size, d, for the difference between two means m_1 and m_2, as

$$d = \frac{|m_1 - m_2|}{s_{pooled}}$$

where s_{pooled} is a pooled standard deviation. Use the formula

$$s_{pooled} = \sqrt{(s_1^2 + s_2^2)/2}$$

Calculate the effect size d for the differences among calories per serving,

- 1936 versus 2006

- 1936 versus 1997
- 1997 versus 2006

Use the values from Wansink, Table 1 as given in Homework 1 or in the course outline.

Answer

Enter the R code in the chunks below. If you choose SAS for this exercise, use the marked portion in the SAS homework template.

1936 versus 2006

```
m_1936 <- 268.1
s_1936 <- 124.8
m_2006 <- 384.4
s_2006 <- 168.3
s_pooled <- sqrt((s_1936^2 + s_2006^2)/2)
d <- abs(m_1936 - m_2006)/s_pooled
cat('d =', d)
```

```
## d = 0.784987603959
```

1936 versus 1997

```
m_1936 <- 268.1
s_1936 <- 124.8
m_1997 <- 288.6
s_1997 <- 122.0
s_pooled <- sqrt((s_1936^2 + s_1997^2)/2)
d <- abs(m_1936 - m_1997)/s_pooled
cat('d =', d)
```

```
## d = 0.166115727787
```

1997 versus 2006

```
m_1997 <- 288.6
s_1997 <- 122.0
m_2006 <- 384.4
s_2006 <- 168.3
s_pooled <- sqrt((s_1997^2 + s_2006^2)/2)
d <- abs(m_1997 - m_2006)/s_pooled
cat('d =', d)
```

```
## d = 0.651769377713
```

Cohen recommends that $d = 0.2$ be considered a small effect, $d = 0.5$ a medium effect and $d = 0.8$ a large effect. Should any of the differences be considered *large*?

Comment - The different between 1936 and 2006 is $\sim$0.78 which, if rounded is equal to 0.8. Therfore, this has a large effect size which means that it is large enough and consistent enough to be seen by the naked eye.

Exercise 2.

Suppose you are planning an experiment and you want to determine how many observations you should make for each experimental condition. One simple formula (see Kuehl, “Design of Experiments : Statistical Principles of Research Design and Analysis”) for the required replicates n is given by

$$n \geq 2 \times \left(\frac{CV}{\%Diff}\right)^2 \times \left(z_{\alpha/2} + z_{\beta}\right)^2$$

where

and z are quantiles from the normal distribution with $\mu = 0$ and $\sigma^2 = 1$.

Use this formula to calculate the number of replicates required to detect differences between calories per serving,

- 1936 versus 2006
- 1936 versus 1997
- 1997 versus 2006

You will need to research how to use the normal distribution functions (`*norm` in R,). Use $\alpha = 0.05$ and $\beta = 0.2$ for probabilities, and let `mean = 0` and `sd = 1` (both z should be positive).

Since n must be an integer, you will need to round *up*. Look up the built in functions for this.

Answer

Enter the R code in the chunks below. If you choose SAS for this exercise, use the marked portion in the SAS homework template.

1936 versus 2006

```r
m_1936 <- 268.1
s_1936 <- 124.8
m_2006 <- 384.4
s_2006 <- 168.3
s_pooled <- sqrt((s_1936^2 + s_2006^2)/2)

alpha <- 0.05
beta <- 0.2
z_alpha <- qnorm(1-alpha/2)
z_beta <- qnorm(1-beta)

cv <- s_pooled/((m_1936 + m_2006)/2)
dif <- (m_1936 - m_2006)/((m_1936 + m_2006)/2)
```

```
rrep <- 2*((cv/dif)^2)*((z_alpha + z_beta)^2)

delta <- (m_2006-m_1936)/s_pooled
check <- 16/(delta^2)
cat('n >= ',rrep,'    Rule of Thumb: n >=', check)

## n >=  25.4748756564     Rule of Thumb: n >= 25.9653622107
```

1936 versus 1997

```
m_1936 <- 268.1
s_1936 <- 124.8
m_1997 <- 288.6
s_1997 <- 122.0
s_pooled <- sqrt((s_1936^2 + s_1997^2)/2)

alpha <- 0.05
beta <- 0.2
z_alpha <- qnorm(1-alpha/2)
z_beta <- qnorm(1-beta)

cv <- s_pooled/((m_1936 + m_1997)/2)
dif <- (m_1936 - m_1997)/((m_1936 + m_1997)/2)
rrep <- 2*((cv/dif)^2)*((z_alpha + z_beta)^2)

delta <- (m_1997-m_1936)/s_pooled
check <- 16/(delta^2)
cat('n >= ',rrep,'    Rule of Thumb: n >=', check)

## n >=  568.874102995     Rule of Thumb: n >= 579.827055324
```

1997 versus 2006

```
m_1997 <- 288.6
s_1997 <- 122.0
m_2006 <- 384.4
s_2006 <- 168.3
s_pooled <- sqrt((s_1997^2 + s_2006^2)/2)

alpha <- 0.05
beta <- 0.2
z_alpha <- qnorm(1-alpha/2)
z_beta <- qnorm(1-beta)

cv <- s_pooled/((m_1997 + m_2006)/2)
dif <- (m_1997 - m_2006)/((m_1997 + m_2006)/2)
rrep <- 2*((cv/dif)^2)*((z_alpha + z_beta)^2)

delta <- (m_1997-m_2006)/s_pooled
check <- 16/(delta^2)
cat('n >= ',rrep,'    Rule of Thumb: n >=', check)
```

```
## n >=  36.9530054638     Rule of Thumb: n >= 37.66448891
```

To check your work, use the rule of thumb suggested by van Belle ("Statistical Rules of Thumb"), where

$$n = \frac{16}{\Delta^2}$$

with

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$$

.

How does this compare with your results? Why does this rule of thumb work? How good is this rule of thumb?

A comment about the notation. When planning for experiments, we can assume known parameters (i.e. σ^2), but when we plan for experiments using the results of past experiments, we can use the corresponding estimates (i.e. s^2).

Comment - Using s_pooled as the standard deviation in the delta, the results of the rule of thumb is very close to the actual calculation. However, you could substitute either σ to get a close estimate. The reason this works is because this formula is derived from the required replicates formula with estimates for alpha and beta.Depending on the type of test you would conduct, you would change the estimated numerator. In this case, since it is a 2 sample test with a beta of 0.2, the numberator is 16, if it was a one sample test, the numerator would be half or 8.

In normal experiment conditions, you would have to estimate the standard deviation you expect. Therefore, this rule of thumb is only as good as the estimates you put into calculation. When substituting the standard deviation from eitheryear rather than the s_pooled calculation, the number isn't as accurate, but close enough for an estimate.

Exercise 3

The probablity of an observation x, when taken from a normal population with mean μ and variance σ^2 is calculated by

$$L(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For values of $x = \{-0.1, 0.0, 0.1\}$, write code to calculate $L(x;\mu = 0, \sigma = 1)$.

Answer

Enter the R code in the chunks below. If you choose SAS for this exercise, use the marked portion in the SAS homework template.

$x = -0.1$

```r
x <- -0.1
mu <- 0
sig <- 1

prob_obs_x <-(1/(sig*sqrt(2*pi)))*(exp(1)^((-1*(x-mu)^2)/2*sig^2))

check <- dnorm(x,mu,sig)

cat('L(x;mu,sigma) = ',prob_obs_x, '  Check is',prob_obs_x == check)

## L(x;mu,sigma) =  0.396952547477   Check is TRUE
```

$x = 0.0$

```r
x <- -0.0
mu <- 0
sig <- 1

prob_obs_x <-(1/(sig*sqrt(2*pi)))*(exp(1)^((-1*(x-mu)^2)/2*sig^2))

check <- dnorm(x,mu,sig)

cat('L(x;mu,sigma) = ',prob_obs_x, '  Check is',prob_obs_x == check)

## L(x;mu,sigma) =  0.398942280401   Check is TRUE
```

$x = 0.1$

```r
x <- 0.1
mu <- 0
sig <- 1

prob_obs_x <-(1/(sig*sqrt(2*pi)))*(exp(1)^((-1*(x-mu)^2)/2*sig^2))

check <- dnorm(x,mu,sig)

cat('L(x;mu,sigma) = ',prob_obs_x, '  Check is',prob_obs_x == check)

## L(x;mu,sigma) =  0.396952547477   Check is TRUE
```

You can confirm your results using the built in normal distribution function. Look up `dnorm` in R help and use the same values for `x`, `mean` and `sigma` as above. You should get matching results to at least 12 decimal places.

Comment - The confirmation is located above in each block of exercise 3. The check confirms that my calculations match the dnorm calculation.

Exercise 4

Part a

Write code to compute

$$7 - 1 \times 0 + 3 \div 3$$

Type this in verbatim, using only numbers, +,-,* and /, with no parenthesis. Do you agree with the result? Explain why, one or two sentences.

Answer

```
7-1*0+3/3

## [1] 8
```

Comment - Yes, I agree with this result because it is following the order of operations, or PEMDAS (Parentheses, Exponents, Multiplication, Division, Addition, Subtraction). Therefore 1 * 0 = 0, 3/3 = 1, so 7-0+1 = 8.

Part b

According to “Why Did 74% of Facebook Users Get This Wrong?” (https://profpete.com/blog/2012/11/04/why-did-74-of-facebook-users-get-this-wrong/), most people would compute the result as 1. Use parenthesis `(  )` to produce this result.

Answer

```
(7-1)*0+3/3

## [1] 1
```

Comment - If you do not follow the order of operations and calculate the formula in order, you would result in an incorrect answer of 1.

Part c

Several respondents to the survey cited in Part 2 gave the answer 6. Add *one* set of parenthesis to produce this result.

Answer

```
7-1*(0+3/3)

## [1] 6
```

Exercise 5.

Part a

Quoting from Wansink and Payne

Because of changes in ingredients, the mean average calories in a recipe increased by 928.1 (from 2123.8 calories … to 3051.9 calories …), representing a 43.7% increase.

Show how 43.7% is calculated from 2123.8 and 3051.9, and confirm W&P result.

Answer

```
m1 <- 2123.8
m2 <- 3051.9

change1 <- (m2-m1)/m1
for_change1 <- paste(round(change1*100,1),"%",sep='')
cat('My calculated result is ', for_change1)

## My calculated result is  43.7%
```

Comment - Using a standard change formula I confirmed W&P's result of 43.7%

The resulting increase of 168.8 calories (from 268.1 calories … to 436.9 calories …) represents a 63.0% increase … in calories per serving.

Part b

Repeat the calculations from above and confirm the reported 63.0% increase in calories per serving. Why is there such a difference between the change in calories per recipe and in calories per serving?

Answer

```
m1 <- 268.1
m2 <- 436.9

change2 <- (m2-m1)/m1
for_change2 <- paste(round(change2*100,1),"%",sep='')
cat('My calculated result is', for_change2)

## My calculated result is 63%
```

Comment - Using a standard change formula I confirmed W&P's result of 63.0%. There is a large change in calories per serving as compared to calories per recipe due to the fact that the number of calories per serving is smaller and therefore has a larger percent change impact than the larger calorie per recipe calculation.

Part c

Calculate an `average_calories_per_serving` by dividing `average_calories_per_recipe` by `average_servings_per_recipe`, for years 1936 and 2006, then calculate a percent increase. Which of the two reported increases (a or b) are consistent with this result?

Answer

```
avg_cal_per_rec_2006 <- 3051.9
avg_cal_per_rec_1936 <- 2123.8
avg_serv_per_rec_2006 <- 12.7
```

```r
avg_serv_per_rec_1936 <- 12.9

avg_cal_per_serv_2006 <- avg_cal_per_rec_2006/avg_serv_per_rec_2006
avg_cal_per_serv_1936 <- avg_cal_per_rec_1936/avg_serv_per_rec_1936

change3 <- (avg_cal_per_serv_2006-
avg_cal_per_serv_1936)/avg_cal_per_serv_1936
for_change3 <-paste(round(change3*100,1),"%",sep='')
var_1 <- (change3-change2)*100

cat('My calculated result is', for_change3,'which has a variance
of',round(var_1,0),'pts from W&P\'s result.')
```

```
## My calculated result is 46% which has a variance of -17 pts from W&P's
result.
```

Comment - The result has a more consistent result with (a) rather than (b). However, the average calories per serving differ by a considerable amount in the article than from my results which puts into question how the studies calories per servings were calculated.