Bios 6301: Assignment 3 - HW3 for HW 4

Abdurrahman Abdulhamid

2023-01-06

Due Tuesday, 27 September, 1:00 PM

50 points total.

Add your name as `author` to the file's metadata section.

Submit a single knitr file (named `homework3.rmd`) by email to tianyi.sun@vanderbilt.edu. Place your R code in between the appropriate chunks for each question. Check your output by using the `Knit HTML` button in RStudio.

$5^{n=day}$ points taken off for each day late.

Question 1

15 points

Write a simulation to calculate the power for the following study design. The study has two variables, treatment group and outcome. There are two treatment groups (0, 1) and they should be assigned randomly with equal probability. The outcome should be a random normal variable with a mean of 60 and standard deviation of 20. If a patient is in the treatment group, add 5 to the outcome. 5 is the true treatment effect. Create a linear model for the outcome by the treatment group, and extract the p-value (hint: see assigment1). Test if the p-value is less than or equal to the alpha level, which should be set to 0.05.

Repeat this procedure 1000 times. The power is calculated by finding the percentage of times the p-value is less than or equal to the alpha level. Use the `set.seed` command so that the professor can reproduce your results.

1. Find the power when the sample size is 100 patients. (10 points)
2. Find the power when the sample size is 1000 patients. (5 points)

set.seed(1000) n_sims <- 1000 # we want 1000 simulations p_vals <- c() for(i in 1:n_sims){ group1 <- rnorm(100,60,20) # simulate group 1 group2 <- rnorm(100,65,25) # simulate group 2 p_vals[i] <- t.test(group1, group2)$p.value # run t-test and extract the p-value } p_vals mean(p_vals <= .05) # check power (i.e. proportion of p-values that are smaller than alpha-level of .05) lm(group1~group2)

[1] 0.592

Question 2

14 points

Obtain a copy of the football-values lecture. Save the `2021/proj_wr21.csv` file in your working directory. Read in the data set and remove the first two columns.

1. Show the correlation matrix of this data set. (4 points)
2. Generate a data set with 30 rows that has a similar correlation structure. Repeat the procedure 1,000 times and return the mean correlation matrix. (10 points)

Question 3

21 points

Here's some code:

```
nDist <- function(n = 100) {
    df <- 10
    prob <- 1/3
    shape <- 1
    size <- 16
    list(
        beta = rbeta(n, shape1 = 5, shape2 = 45),
        binomial = rbinom(n, size, prob),
        chisquared = rchisq(n, df),
        exponential = rexp(n),
        f = rf(n, df1 = 11, df2 = 17),
        gamma = rgamma(n, shape),
        geometric = rgeom(n, prob),
        hypergeometric = rhyper(n, m = 50, n = 100, k = 8),
        lognormal = rlnorm(n),
        negbinomial = rnbinom(n, size, prob),
        normal = rnorm(n),
        poisson = rpois(n, lambda = 25),
        t = rt(n, df),
        uniform = runif(n),
        weibull = rweibull(n, shape)
    )
}
```

1. What does this do? (3 points)

```
round(sapply(nDist(500), mean), 2)
```

```
##           beta       binomial     chisquared    exponential              f
##           0.10           5.17           9.94           0.94           1.16
##          gamma      geometric hypergeometric      lognormal    negbinomial
##           1.04           1.97           2.61           1.79          32.13
##         normal        poisson              t        uniform        weibull
##          -0.03          24.79           0.09           0.52           1.02
```

answer here

2. What about this? (3 points)

```
sort(apply(replicate(20, round(sapply(nDist(10000), mean), 2)), 1, sd))
```

```
##           beta        uniform              f         normal        weibull
##    0.000000000    0.002236068    0.006882472    0.007181848    0.009445132
##              t    exponential          gamma hypergeometric       binomial
##    0.010699238    0.010711528    0.010894228    0.012680279    0.016944181
##      lognormal      geometric     chisquared        poisson    negbinomial
##    0.020749128    0.026137289    0.040974960    0.054386531    0.102910283
```

```
answer here
```

In the output above, a small value would indicate that `N=10,000` would provide a sufficent sample size as to estimate the mean of the distribution. Let's say that a value *less than 0.02* is "close enough".

3. For each distribution, estimate the sample size required to simulate the distribution's mean. (15 points)

Don't worry about being exact. It should already be clear that N < 10,000 for many of the distributions. You don't have to show your work. Put your answer to the right of the vertical bars (|) below.

distribution	N
beta	?
binomial	?
chisquared	?
exponential	?
f	?
gamma	?
geometric	?
hypergeometric	?
lognormal	?
negbinomial	?
normal	?
poisson	?
t	?
uniform	?
weibull	?