QSCI 482 Story 5

Simon Hans Edasi

2023-11-13

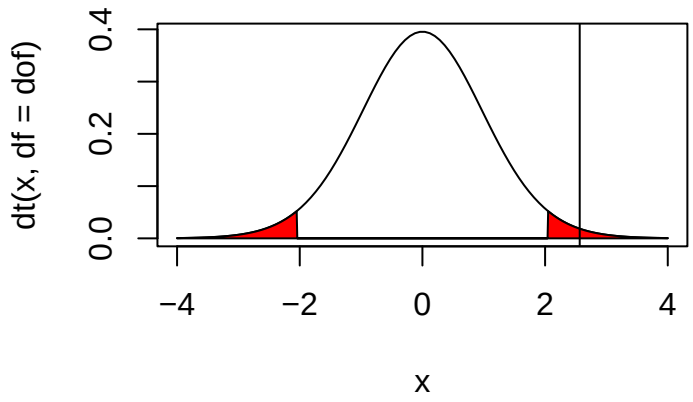
Question 1

a. Conduct a paired-sample parametric test the long way around to determine whether the highway time is significantly different from the back route time

$H_0 : \mu_h = \mu_b$

$H_0 : \mu_h \neq \mu_b$

```
alpha <- 0.05
n <- length(h)
d <- h-b
dbar <- mean(d)
s_d <- var(d)
s_dbar <- sqrt(s_d / n)
dof <- n - 1
t <- dbar / s_dbar
t_crit <- abs(qt(alpha/2, df = dof, lower.tail = F))
p_value <- 2*pt(t,df = dof, lower.tail = F)
t_plot(alpha, dof = dof, t_obs = t)
```




```
glue('degrees of freedom: {dof}\nt_obs: {t}\nt_crit: {t_crit}\np-value: {p_value}')
```

```
## degrees of freedom: 29
## t_obs: 2.56449856760162
```

```
## t_crit: 2.0452296421327
## p-value: 0.0157743290340233
```

We see significantly higher times on the highway than the backroads and conclude the backroute is faster.

b. Conduct the test the short way around. Your answers should match.

```
t.test(h,b,paired = T)
```

```
##
##  Paired t-test
##
## data:  h and b
## t = 2.5645, df = 29, p-value = 0.01577
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  10.39416 92.27251
## sample estimates:
## mean difference
##        51.33333
```

Yay, same values!

Question 2

a. Conduct a Wilcoxon paired-sample test (this is non-parametric), the long way around, to determine whether the highway time is significantly difference from the back route time. Report the test statistic, p-value etc.

$H_0 : \mu_h = \mu_b$

$H_0 : \mu_h \neq \mu_b$

```
diff <- h-b
# find index of nonzero differences
r <- which(diff != 0)
diff <- diff[r]
h <- h[r]
b <- b[r]
absd <- abs(diff)
ranks <- rank(absd)
n <- length(diff)

T_pos <- ranks[diff > 0]
T_p <- sum(T_pos)

T_neg <- ranks[diff < 0]
T_n <- sum(T_neg)

smaller_T <- min(abs(T_p),abs(T_n))
left <- psignrank(smaller_T, n)
p <- 2 * left
glue('T+: {T_p}\nT-: {T_n}\nleft: {left}\np-value: {p}')
```

```
## T+: 349
## T-: 116
```

```
## left: 0.00773000158369541
## p-value: 0.0154600031673908
```

b. Conduct the same test but obtain the p-value using the normal approximation (use twice the area for a two-tailed test).

```
df <- data.frame(h,b,diff)
df$absd <- abs(df$diff)
df$rank <- rank(abs(diff))
df$sign <- (df$diff / abs(df$diff))
# Get the df index of ranks in ascending rank order
rank_order <- order(df$rank)
# Identify df indexes where rank is not a whole number
repeated_ranks <- rank_order[ranks[ranks != as.integer(ranks)]]
repeats <- df[repeated_ranks, ]
# Repeats Of Different Sign
ros <- (repeats$diff + (repeats$diff * repeats$sign))
rods <- repeats[ros == 0,]
# w = number of columns in rods
w <- dim(rods)[1]
print(w)
```

```
## [1] 0
```

Okay, we have no repeats of different signs so we don't have to worry about correcting for it. I'm still gonna code it for future use!

```
rods_correction <- function(rods){
  t_c <- 0
  if (w == 0){
    return(t_c)
  }
  if (w != 0){
    t_vec <- vector()
    n <- 1
    for (i in (unique(rods$absd))){
      t_vec[n] <- nrow(rods[,rods$diff == i])
      n <- n+1
    }
    for (i in 1:length(t_vec)){
      t_i <- t_vec[i]
      t_c <- t_c + (t_i^3 - t_i)
    }
    t_c <- (t_c / 2)
    return(t_c)
  }
}
t_c <- rods_correction(rods)
```

```
mu <- (n * (n+1)) / 4
sig <- sqrt(  (n*(n+1)*(2*n + 1) - t_c) / 24  )
Z <- (abs(smaller_T - mu) - 0.5) / sig
p <- 2*(pnorm(Z,lower.tail = F))
print(p)
```

```
## [1] 0.01703612
```

c. Conduct a Wilcoxon paired-sample test the short way around in R. If this differs from 2a) or 2b), explain why.

```
wilcox.test(x = h,y = b,paired = T, alternative = 'two.sided')
```

```
## Warning in wilcox.test.default(x = h, y = b, paired = T, alternative =
## "two.sided"): cannot compute exact p-value with ties
```

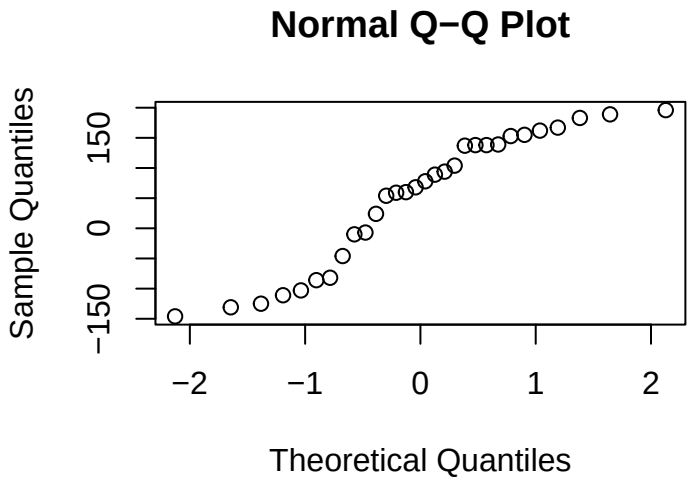
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  h and b
## V = 349, p-value = 0.01703
## alternative hypothesis: true location shift is not equal to 0
```

This differs from the answer in 2a because the built in test uses the normal approximation automatically and toggles correction on or off.

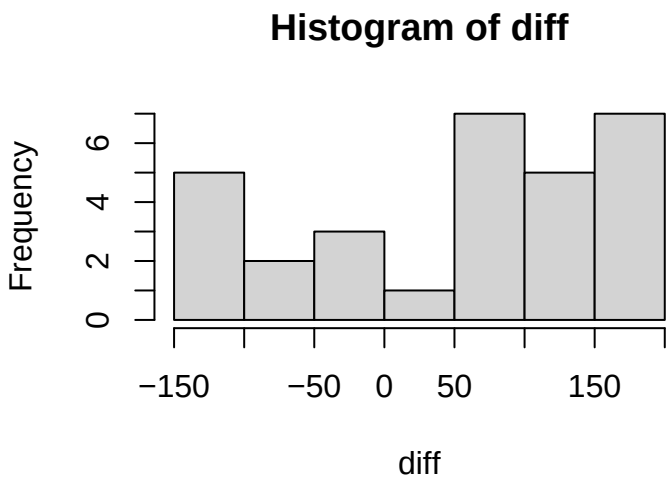
Question 3

a. Test whether the differences in time are normally distributed using qqplot, histogram, and the Shapiro-Wilk test. What do you conclude about normality?

```
qqnorm(diff)
```




```
hist(diff)
```




```
shapiro.test(diff)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diff
## W = 0.90831, p-value = 0.01349
```

All three of these distribution tests show that the differences are not normally distributed.

b. Given the answer in 3a), which of the tests used in Q1 or Q2 is appropriate, and what do you conclude about travel time on the two routes? Was your roommate correct or were you correct?

We show that our data is not normally distributed, which means we can not use the t-test as our assumption of normality is violated. We also can not use the normal approximation, as we have too few data points. Therefore, we must use the non-parametric Wilcoxon signed-rank test to determine if there is a difference in travel times and conclude that I am right to take the backroute home.

Question 4

a. Run a chi-square goodness-of-fit test the long way around to determine if the distribution of fruit flies among the four categories could be due to random luck, or are due to a preference of fruit flies for some substances over others.

H_0 : Fruit fly attraction is independent of substance.

H_A : Fruit fly attraction is dependent on substance.

```
k <- c('banana', 'vinegar','honey','sugarwater')
obs <- c(23, 18, 5, 10)
n <- sum(obs)
K <- length(k)
preds <- vector()
for (i in k)
  preds[i] <- n/K
```

```r
chisq.value <- sum((obs - preds)^2 / preds)
p <- pchisq(abs(chisq.value), df = K-1, lower.tail = F)
glue('\nChiSQ goodness of fit\nchisq: {chisq.value}\np_value: {p}')
```

```
## ChiSQ goodness of fit
## chisq: 13.8571428571429
## p_value: 0.00310619766757956
```

b. Run a direct R test.

```r
chisq.test(x = obs, p = preds / n)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 13.857, df = 3, p-value = 0.003106
```

Question 5

Long way around

a. Run a chi-square test with no corrections

H_0 : Offspring anadromy is independent of maternal anadromy

H_A : Offspring anadromy depends on maternal anadromy

```r
obs <- matrix(c(16,13,5,1),byrow = T,nrow = 2, ncol = 2)
obs
```

```
##      [,1] [,2]
## [1,]   16   13
## [2,]    5    1
```

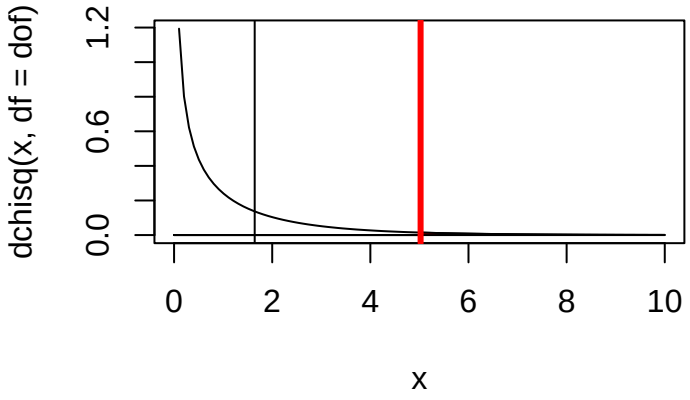
```r
alpha <- 0.05
n <- sum(obs)

Rvec <- rowSums(obs)
Cvec <- colSums(obs)

chi.stat <- (det(obs)^2 * n) / prod(Rvec * Cvec)
chi.crit <- qchisq(alpha/2,df = 1,lower.tail = F)
p_value <- pchisq(chi.stat,df =1,lower.tail = F)
glue('chi.stat: {chi.stat}\nchi.crit: {chi.crit}\np-value: {p_value}')
```

```
## chi.stat: 1.64272030651341
## chi.crit: 5.02388618731489
## p-value: 0.199952639507696
```

```r
chisq_plot(0.05,1,chi.stat)
```



b. Run a chi-square test with the Yates continuity correction

```
alpha <- 0.05
n <- sum(obs)

Rvec <- rowSums(obs)
Cvec <- colSums(obs)

chi.stat <- (  (abs(det(obs)) - (n/2))^2 * n  ) / prod(Rvec * Cvec)
chi.crit <- qchisq(alpha/2,df = 1,lower.tail = F)
p_value <- pchisq(chi.stat,df =1,lower.tail = F)
glue('chi.stat: {chi.stat}\nchi.crit: {chi.crit}\np-value: {p_value}')
```

```
## chi.stat: 0.678879310344827
## chi.crit: 5.02388618731489
## p-value: 0.409972896845415
```

c. Now compare the results of 5a-5b with the built-in chi-square test in R. Which form is used by default in R for a 2x2 contingency table?

```
chisq.test(obs)
```

```
## Warning in chisq.test(obs): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  obs
## X-squared = 0.67888, df = 1, p-value = 0.41
```

The built in r function applies the Yates correction by default.

d. What conclusions can be drawn from these data and the test?

We can conclude that offspring anadromy is independent from maternal anadromy

e. The data in Q5 can only have a certain number of possible outcomes, given the number of anadromous mothers, and the number of non-anadromous mothers. For example, there were six anadromous mothers, and thus the data for offspring could only be 0-6, 1-5, 2-4, 3-3, 4-2, 5-1, or 6-0 for the offspring. In this question, you will need to explore what it would take to change the result of the test (this is related to the power of the test). If the test was significant, what is the smallest change in the data that would lead to a non- significant result? If the test was not significant, what is the smallest change in the data that would lead to a significant result? What does this tell you about the available sample size compared to the required sample size to reject the null hypothesis?