140.61.HW7.Delahanty.Jeremy

Jeremy Delahanty

2024-04-10

1

1a Apply a Chi-square test to the deer-gland and tick data.

```
# Create the table of datapoints
dataset <- rbind(c(24, 5), c(18, 5), c(23,4), c(20,4), c(17,8),
                 c(25, 3), c(21,6), c(25,2))

# Perform the test but do NOT use the ocntinuity correction because
# our data is not continuous.
test <- chisq.test(x = dataset, correct = FALSE)
```

```
## Warning in chisq.test(x = dataset, correct = FALSE): Chi-squared approximation
## may be incorrect
```

```
cat("\nP-value: ", test$p.value, "\n")
```

```
##
## P-value:  0.403441
```

```
# Extract from this the generated table of expecte values where the null
# is that the data are independent
test$expected
```

```
##          [,1]     [,2]
## [1,] 23.89048 5.109524
## [2,] 18.94762 4.052381
## [3,] 22.24286 4.757143
## [4,] 19.77143 4.228571
## [5,] 20.59524 4.404762
## [6,] 23.06667 4.933333
## [7,] 22.24286 4.757143
## [8,] 22.24286 4.757143
```

We see htere that our p-value is substantially higher than an alpha of 0.05 and therefore we fail to reject the null hypothesis.

1b Apply the lilkelihood ratio test

```r
# Use the formula from lecture for this test statistic
lrt <- 2*sum(dataset*log(dataset/test$expected))

# Determine DoF from the dataset: #rows - 1  * #columns - 1
dof <- (8-1) * (2-1)

# Calculate the p-value from the test statistic
lrt_pvalue <- pchisq(q = lrt, df = dof, lower.tail = FALSE)

# Display the p-value
cat("LRT p-value is: ", lrt_pvalue)
```

```
## LRT p-value is:  0.4045542
```

Again, the p-value is substantially above 0.05 and we fail to reject the null hypothesis.

1c Apply Fischer’s Exact Test

```r
fisher.test(dataset)
```

```
##
##  Fisher’s Exact Test for Count Data
##
## data:  dataset
## p-value = 0.4265
## alternative hypothesis: two.sided
```

The p-value is again substantially higher than the alpha value of 0.05 and so we fail to reject the null hypotehsis.

1d

We can conlude that the dataset, regardless of the statistic used, does not provide evidence at a significance value of 0.05 that there is a difference in the mean probability of choosing the treated tube and the untreated tube. All tests yield a similar p-value of approximately 0.4.

2

We can use the built-in `dhyper` function for this test at a significance level of 0.05. The hypergeometric distribution should only use tables that are at least as extreme as the observed data. We will want to use a one sided t-test because we are only interested in whether the use of HGH will yield taller individuals.

```r
dataset <- rbind(c(26, 29), c(18, 39))

# Now, check which tables have non-zero probabilities with dhyper
n <- sum(dataset)
yes_tot <- apply(dataset, 2, sum)[1]
no_tot <- apply(dataset, 2, sum)[2]
hgh_tot <- apply(dataset, 1, sum)[1]

# Create distribution tables for what was actually observed
p_observed <- dhyper(dataset[1,1], yes_tot, no_tot, hgh_tot)

# Do this computation
p_all <- dhyper(yes_tot:0, yes_tot, no_tot, hgh_tot)

# Gather only the tables that are less than p_obs
usable_probs <- p_all <= p_observed
usable_probs
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [37]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```r
# We see that the first 19 values are all TRUE and will therefore use those.
# We will focus only those entries which are formulated around HGH height is
# greater than non HGH height
p_value <- sum(p_all[1:19])

# Display the p-value
cat("P-value is: ", p_value)
```

```
## P-value is:  0.06578953
```

```r
# Now use Fischer's Test to see if we get the same answer!
fisher.test(dataset, alternative = "greater")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dataset
## p-value = 0.06579
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.9488927       Inf
## sample estimates:
## odds ratio
##   1.930871
```

We can see that our 2 p-values agree with one another! We must have calculated this test correctly. They are both above the alpha value of 0.05 and therefore this data offers insufficient evidence for rjecting the null hypothesis. There is not sufficient evidence that HGH treatment leads to growth increase over the placebo.

3

3a Is there evidence that the infections are *not* independent of each other?

We can answer this question by generating contingency tables and then manually calculating both the chi-square test statistic and the likelihood ratio test. Finally, we can use the built in `R` functions for comparing our manual calculation against the built in functions' values.

```r
# Calculate expected values
infected <- rbind(c(10, 25), c(14, 101))
# Calculate the marginal values:
row_sum <- apply(infected, 1, sum)
col_sum <- apply(infected, 2, sum)
infected_tot <- sum(infected)
expected <- outer(row_sum, col_sum, "*")/infected_tot

# Calculate the DoF
df <- (2-1)*(2-1)

# Manually calculate the chi-square test statistic
chisquare <- sum((infected - expected)^2/expected)
cat("Chi-square test statistic (manually):" , chisquare, "\n")
```

```
## Chi-square test statistic (manually): 5.368234
```

```r
p_value_xsq <- pchisq(chisquare, df = df, lower.tail = FALSE)
cat("Chi-square p-value (manually):", p_value_xsq, "\n")
```

```
## Chi-square p-value (manually): 0.02050673
```

```r
# Now do this procedure for the LRT
lrt <- 2*sum(infected * log(infected/expected))
cat("LRT test statistic (manually):", lrt, "\n")
```

```
## LRT test statistic (manually): 4.835653
```

```r
p_value_lrt <- pchisq(lrt, df = df, lower.tail = FALSE)
cat("LRT p-value (manually):", p_value_lrt)
```

```
## LRT p-value (manually): 0.02787709
```

```r
# Finally, use the built in functions to verify our values; set correct to FALSE
# so we do not use Yates' continuity correction
chisq.test(infected, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  infected
## X-squared = 5.3682, df = 1, p-value = 0.02051
```

```
fisher.test(infected)
```

```
##
##  Fisher’s Exact Test for Count Data
##
## data:  infected
## p-value = 0.03283
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.010996 7.909161
## sample estimates:
## odds ratio
##   2.861389
```

We can see that our analytic solutions are nearly the same as the built iln values found through `R` functions! Using both tests, we have p-values less than 0.05 and can therefore reject the null hypotehsis. We can conclude that there is is dependence between the infections.

1b Are HIV/HepC prevalences the same? Calculate both the analytic solution and uses the built in functions.

We can use our contingency tables constructed above for this question and ignore any individuals who (unfortunately) have both HepC and HIV. We will again use a significance value of 0.05. The built in test we can use here is McNemar’s Chi-squared test

```
hepc <- infected[1, 2]
hiv <- infected[2, 1]

# Calculate chi-square
chisquare <- (hepc - hiv)^2 / (hepc + hiv)
cat("Chi-square test statistic (manually):", chisquare, "\n")
```

```
## Chi-square test statistic (manually): 3.102564
```

```
# Calculate the p-value
p_value <- pchisq(chisquare, df = df, lower.tail = FALSE)
cat("P-value from chi-square (manually):", p_value)
```

```
## P-value from chi-square (manually): 0.07816909
```

```
# Use McNemar's test with the original data without continuity correction
mcnemar.test(infected, correct = FALSE)
```

```
##
##  McNemar’s Chi-squared test
##
## data:  infected
## McNemar’s chi-squared = 3.1026, df = 1, p-value = 0.07817
```

We can see that our test statistic and p-value are the same! We performed our analytic solutions correctly. Our dataset does not have sufficient evidence to reject the null hypothesis. Therefore, we will conclude that the prevalence of the diseases in the population are the same.

4

4a

We can again use a contingency table for our dataset. From lecture, we also know that if $n \times \hat{p}$ *and* $n \times \hat{p} \times (1 - \hat{p})$ are greater than 5, we can use a Normal approximation for the confidence interval.

```
# Put together a table and make variables for useful sums
gaits <- rbind(c(198, 104), c(91, 218))
low_speed <- apply(gaits, 1, sum)[1]
normal_speed <- apply(gaits, 1, sum)[2]
less_68 <- apply(gaits, 2, sum)[1]
geq_68 <- apply(gaits, 2, sum)[2]

# Calculate phat for less than 68 within a low speed group
p_less_68 <- gaits[1,1] / less_68

# Now generate the confidence interval
z_score <- qnorm(0.975) # tails of 0.025
lower_bound <- p_less_68 - z_score * sqrt(p_less_68*(1 - p_less_68)/less_68)
upper_bound <- p_less_68 + z_score * sqrt(p_less_68*(1 - p_less_68)/less_68)

# Display the computed confidence interval
cat("95% CI for Low Gait Speed:", lower_bound, upper_bound, "\n")
```

```
## 95% CI for Low Gait Speed: 0.6315717 0.7386705
```

```
# Now do the same thing for individuals over 68
p_slow_geq68 <- gaits[1, 2] / geq_68
lower_bound <- p_slow_geq68 - z_score * sqrt(p_slow_geq68 * (1 - p_slow_geq68) / geq_68)
upper_bound <- p_slow_geq68 + z_score * sqrt(p_slow_geq68 * (1 - p_slow_geq68) / geq_68)

# Display the computed confidence interval
cat("95% CI for GEQ 68 Gait Speed:", lower_bound, upper_bound, "\n")
```

```
## 95% CI for GEQ 68 Gait Speed: 0.2719063 0.3740565
```

4b) Compute 95% confidence interval for difference in proportions

The formula we will use relies upon the $\hat{p}$s computed before as well as the standard errors of the differences in proportions to get our confidence interval.

```
# Compute the lower bound
lower_bound <- (p_less_68 - p_slow_geq68) - z_score * sqrt(
  ((p_less_68*(1 - p_less_68)/less_68) + (p_slow_geq68*(1 - p_slow_geq68)/geq_68))
  )
upper_bound <- (p_less_68 - p_slow_geq68) + z_score * sqrt(
  ((p_less_68*(1 - p_less_68)/less_68) + (p_slow_geq68*(1 - p_slow_geq68)/geq_68))
  )

# Displ ay the confidence interval
cat("95% CI for Difference in Proportions:", lower_bound, upper_bound)
```

```
## 95% CI for Difference in Proportions: 0.2881384 0.4361411
```