Categorical Data Hypothesis Testing Review Sheet

Advanced Statistics

1. One Proportion

You flip 60 heads in 100 coin tosses. What is the chance that you'd have this many heads or more with a fair coin?

With the Standard Normal Approximation

```
sd_heads = sqrt(100*0.5*0.5)
zscore = (59.5 - 50)/sd_heads
# using 59.5 instead of 60 is the continuity correction
1-pnorm(zscore)
```

```
## [1] 0.02871656
```

Using The Binomial Formula

```
sum(dbinom(60:100, 100, 0.5))
```

```
## [1] 0.02844397
```

Using R's built-in Proportion Test

```
prop.test(x=60, 100, p=0.5, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  60 out of 100, null probability 0.5
## X-squared = 3.61, df = 1, p-value = 0.02872
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5127842 1.0000000
## sample estimates:
##   p
## 0.6
```

2. Comparing Two Proportions

In a class experiment, Jonah made 60 out of 100 free throws while Owen made 48 out of 105 free throws. What is the chance that one of the two students would out-perform the other by this much or more if they were equally good free throw shooters?

Full Calculation (without a continuity correction)

```
pooled_success_rate = (60+48)/(100+105)

sd_jonah_rate = sqrt(pooled_success_rate*(1-pooled_success_rate)/100)

sd_owen_rate = sqrt(pooled_success_rate*(1-pooled_success_rate)/105)

sd_diff_rates = sqrt(sd_jonah_rate^2 + sd_owen_rate^2)

diff_in_rates = 60/100 - 48/105

zscore_diff = diff_in_rates/sd_diff_rates

2*(1-pnorm(zscore_diff))
```

```
## [1] 0.04058503
```

```
# We used "2*" because this is "two-sided"
# the alternative hypothesis isn't that
# Jonah is better it's that one of the
# two students is better than the other.
```

With a continuity correction:

```
diff_in_rates = 59.5/100 - 48.5/105
zscore_diff = diff_in_rates/sd_diff_rates

2*(1-pnorm(zscore_diff))
```

```
## [1] 0.05641508
```

Using R's built in function

```
prop.test(x = c(60, 48), n=c(100, 105), alternative = "two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(60, 48) out of c(100, 105)
## X-squared = 3.6398, df = 1, p-value = 0.05642
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.002177206  0.287891492
## sample estimates:
```

```
##    prop 1    prop 2
## 0.6000000 0.4571429
```

3. Chi Square Test for Goodness of Fit

Our example, results from possibly unfair dice:

side	num_times
1	80
2	103
3	93
4	105
5	96
6	123

How likely is it that we'd see results this uneven (or more uneven) from a fair die?

Full Calculations:

```
num_times <- c(80, 103, 93, 105, 96, 123)
expected <- rep(100, 6)
differences = num_times-expected
chi_square = sum((differences^2)/expected)
1 - pchisq(chi_square, df=5) # to find the p-value
```

```
## [1] 0.06767935
```

Easy Way:

```
num_times <- c(80, 103, 93, 105, 96, 123)
chisq.test(x = num_times, p=rep(1/6, 6))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  num_times
## X-squared = 10.28, df = 5, p-value = 0.06768
```

4. Chi Square Test for Independence

Our example, kids taking multiple science classes by grade:

	One_Science_Class	Mult_Science_Classes
9	57	26
10	35	45
11	61	21
12	38	41

What are the chances that the rates of doubling up by grade would be this different or more different if doubling up and grade were independent?

Full Calculations:

```
observed = c(57, 26, 35, 45, 61, 21, 38, 41)

row_tots = c(83, 83, 80, 80, 82, 82, 79, 79)

col_tots = rep(c(191, 133),4)
grand_total = sum(observed)
expecteds = row_tots*col_tots/grand_total
chi_square = sum(((observed-expecteds)^2)/expecteds)
1 - pchisq(chi_square, df=3) #for the p-value
```

```
## [1] 4.439895e-05
```

Easy Way:

```
observed = c(57, 26, 35, 45, 61, 21, 38, 41)

values = matrix(observed, ncol=2, byrow=TRUE)

chisq.test(values)
```

```
##
##  Pearson's Chi-squared test
##
## data:  values
## X-squared = 22.802, df = 3, p-value = 4.44e-05
```