Analyzing Categorical Data using the Pearson Statistic and R

Consider a multinomial experiment with $k = 4$ possible outcomes. After $n = 100$ trials are performed, we observe the following counts of each outcome.

Suppose that we wanted to determine whether the probabilities of each outcome differ; this analysis will echo that of Example 14.1. More precisely, we want to test the null hypothesis $H_0 : p_1 = p_2 = p_3 = p_4 = 0.25$. To do so, we use the Pearson test statistic X^2 with approximate χ^2 distribution; in this case, X^2 has $k-1$ degrees of freedom since the only assumption used about the outcome probabilities is that $p_1 + p_2 + p_3 + p_4 = 1$.

We can calculate X^2 using the following code.

```
# Set outcome probabilities under H0.
p <- rep(0.25, 4)

# Define vector of observed counts.
y <- c(20, 17, 31, 32)
n <- sum(y)

# Calculate X^2.
Xsq = sum((y - p*n)^2/(p*n))
```

This gives $X^2 = 6.96$.

We have an α-level test if we reject H_0 when $X^2 > \chi^2_{\alpha,k-1}$. For $\alpha = 0.05$ and $k - 1 = 3$, we have the rejection region

$$X^2 > 7.815.$$

In this case, we cannot reject H_0 since the observed value of X^2 is outside the rejection region.

Using the `chisq.test` function

We could have also performed the test using the `chisq.test` function, as in the following code.

```
chisq.test(y)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  y
## X-squared = 6.96, df = 3, p-value = 0.07318
```

When called, `chisq.test(y)` calculates the value of X^2 for the observed counts stored in the vector `y`, as well as the observed attained significance level. Here, we have $p = 0.07318$ which indicates that we cannot

Outcome	Observed Count
1	20
2	17
3	31
4	32

reject H_0 (using $\alpha = 0.05$).