Stat 344 – Test 2

Trey Tipton

June 30, 2022

```
wrld <- read.csv('https://sldr.netlify.app/data/sustainable-tanzania.csv')
```

Categories, Categories

```
tally( ~ Age_Group + Own_Land, data = wrld)
```

```
##           Own_Land
## Age_Group  no yes
##   30_49     7  40
##   50_plus   2  29
##   under_30  4   5
```

Hypotheses:

$H_0 : \Omega_0 = \{\pi | \pi_{ij} \geq 0 \text{ and } \sum_{i,j} \pi_{ij} = 1\}$,

Age group and Owning Land are independent. Knowing the age group does not help to know whether they own land.

$H_a : \Omega_0 = \{\pi | \pi_{ij} = \pi_{i.}\pi_{j.} \text{ and } \sum_i \pi_{i.} = \sum_j \pi_{.j} = 1\}$,

Age group and Owning Land relate to each other, knowing the age group helps to know whether they own land.

Test Statistic:

$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

```
chisq.test(wrld$Age_Group, wrld$Own_Land)
```

```
## Warning in chisq.test(wrld$Age_Group, wrld$Own_Land): Chi-squared approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  wrld$Age_Group and wrld$Own_Land
## X-squared = 7.9218, df = 2, p-value = 0.01905
```

Our χ^2 test statistic came out to be 7.9218.

At a low p-value of 0.01905, we reject the null hypothesis and can say that there is some sort of association between age group and owning land, and they are not independent.

The test we just carried out relies on the fact that $G = \chi^2$, which becomes more accurate as n gets larger. Because we have a few bins with low values, we do have a reason to suspect that our test may not hold.

```
set.seed(032601)
n_sim = 10000
sims <- do(4)*chisq.test(wrld$Age_Group, wrld$Own_Land, simulate.p.value = TRUE, B = n_sim)

sims$p.value
```

```
## [1] 0.01799820 0.01949805 0.01919808 0.01849815
```

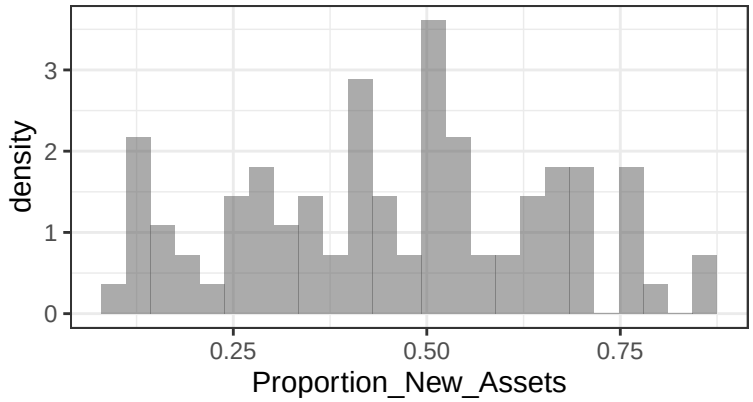
The p-values are 0.018, 0.0195, 0.192, and 0.0185.

Using simulations, the p-values we got surround the p-value we got from the original chi-squared test. Each p-value still leads to a rejection in the null hypothesis.

We can now conclude that although some bins have low values, our p-value is still accurate because our simulations gave us p-values similar to and above and below our original p-value.

Testing, Testing

```
gf_dhistogram(~Proportion_New_Assets, data = wrld)
```




```
LL <- function(theta, x){
  mu <- theta[1]
  sigma <- theta[2]
  if (mu < 0) return(NA)
  if (sigma < 0) return(NA)
  dnorm(x, mean = mu, sd = sigma, log = TRUE)
}

maxLik(LL, start = c(mean = .5, sd = .25), x = wrld$Proportion_New_Assets)
```

```
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: 18.96732 (2 free parameter(s))
## Estimate(s): 0.4614054 0.1945719
```

```
mu <- 0.4614054
sigma <- 0.1945719
```

Using maximum-likelihood estimation, I chose to fit a normal distribution to the data for Proportion of New Assets. The resulting parameter estimates are $\mu = 0.4614054$ and $\sigma = 0.1945719$.

We can now carry out a Goodness of Fit test for the model we just fitted:

Hypotheses:

$H_0 : \Omega_0 = \{\theta | \mu \geq 0, \sigma \geq 0\}$,

A normal distribution generates the data for the Proportion of New Assets.

H_a : A normal distribution does not generate the data for the Proportion of New Assets.

Test Statistic:

$G = -2 * log(\lambda) = 2 * \sum o_i * log(\frac{o_i}{e_i})$

```
bins <- c(-.1, .2, .3, .4, .5, .6, .7, .8, Inf)
wrld <- wrld %>%
  mutate(binned = cut(wrld$Proportion_New_Assets, breaks = bins))

tally(~binned, data = wrld)
```

```
## binned
## (-0.1,0.2]  (0.2,0.3]  (0.3,0.4]  (0.4,0.5]  (0.5,0.6]  (0.6,0.7]  (0.7,0.8]
##          12         10         13         20         10         11          9
##  (0.8,Inf]
##          2
```

```
count_dat <- data.frame(tally(~binned, data = wrld))
count_dat
```

```
##       binned Freq
## 1 (-0.1,0.2]   12
## 2  (0.2,0.3]   10
## 3  (0.3,0.4]   13
## 4  (0.4,0.5]   20
## 5  (0.5,0.6]   10
## 6  (0.6,0.7]   11
## 7  (0.7,0.8]    9
## 8  (0.8,Inf]    2
```

```
count_dat <- count_dat %>%
  mutate(probs = diff(pnorm(bins, mean = mu, sd = sigma)),
  e = sum(count_dat$Freq) * probs )

count_dat
```

```
##       binned Freq      probs         e
## 1 (-0.1,0.2]   12 0.08760170  7.621348
## 2  (0.2,0.3]   10 0.11384244  9.904293
## 3  (0.3,0.4]   13 0.17275685 15.029846
## 4  (0.4,0.5]   20 0.20246094 17.614102
## 5  (0.5,0.6]   10 0.18324521 15.942333
## 6  (0.6,0.7]   11 0.12808634 11.143512
## 7  (0.7,0.8]    9 0.06913983  6.015165
## 8  (0.8,Inf]    2 0.04091169  3.559317
```

```
G <- 2 * sum( count_dat$Freq * log( count_dat$Freq / count_dat$e))
G
```

```
## [1] 7.730431
```

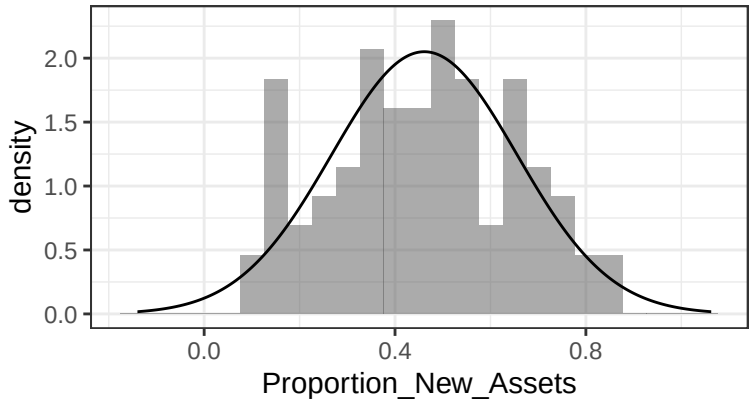
```
1 - pchisq(G, df = nrow(count_dat) -1 -2)
```

```
## [1] 0.1717312
```

Our sample statistic is G = 7.730431. Using k-1-m degrees of freedom, we got a p-value of 0.1717312.

With a p-value of 0.17 we fail to reject the null hypothesis that a normal distribution generated this data. Therefore, we conclude that it is plausible for the Proportion of New Assets to have been generated from a normal distribution.

```
gf_dhistogram(~Proportion_New_Assets, data = wrld)%>%
  gf_dist(dist = "norm", params = list(mean = 0.4614054, sd = 0.1945719))
```



The overlayed pdf (see above) on the histogram is consistent with our hypothesis test results. It is not perfect, but the data seems to be somewhat normal and it is plausible that our data was generated by that distribution, and our p-value of 0.17 communicates the same idea.

Just Testing

Since we just failed to reject that a normal distribution generated the data for the Proportion of New Assets, I have decided to use this variable. It seems that about half of the people's assets are "new", so I will be testing whether the mean is .5 with unknown standard deviation.

Hypotheses:

$\theta =< \mu, \sigma >$

$H_0 : \Omega_0 = \{\theta | \mu = .5, \sigma \geq 0\}$,

The proportion of new assets has a normal distribution with a mean of 0.5.

$H_a : \Omega_0 = \{\theta | 0 \leq \mu \geq 1, \sigma \geq 0\}$,

The proportion of new assets has a normal distribution, but not with a mean of 0.5.

```
ll_norm <- function(theta, x){
  mu <- theta[1]
  sigma <- theta[2]
  if (mu < 0) return(NA)
  if (sigma < 0) return(NA)
  sum(log(dnorm(x, mean = mu, sd = sigma)))
}


maxLik(logLik = ll_norm, start = c(mu = .5, sigma = 1), x = wrld$Proportion_New_Assets)
```

```
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 6 iterations
```

```
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: 18.96732 (2 free parameter(s))
## Estimate(s): 0.4614054 0.1945719
```

```
maxLik(logLik = ll_norm, start = c(mu = .5, sigma = 1), fixed = 1, x = wrld$Proportion_New_Assets)
```

```
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 7 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: 17.28861 (1 free parameter(s))
## Estimate(s): 0.5 0.1983627
```

```
W <- 2*(18.96732 - 17.28861)
W
```

```
## [1] 3.35742
```

```
1 - pchisq(W, 1)
```

```
## [1] 0.06690281
```

Test Statistic:

$\lambda = \frac{L(\Omega_0)}{L(\hat{\Omega})}$ and $W = -2 * log(\lambda) = 2 * (l(\hat{\Omega}) - l(\Omega_0)) = 2$ * (18.96732 - 17.28861) = 3.35742

p-value and conclusion:

From 1 - pchisq(W, 1), using df = 1 since $dim(\hat{\Omega}) - dim(\Omega_0) = 2 - 1 = 1$, we get a p-value of 0.06690281.

With a p-value of 0.067 and a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis that the mean proportion of new assets is 0.5. Although the p-value is still relatively low, there is not enough evidence to say that the true proportion of new assets data does not come from a normal distribution with a mean of 0.5.