SDA - Problem Sheet 8

Hans Reimann

8 1 2022

Exercise 3

In order to test customer satisfaction with a given service, we conduct a survey and define a random variable Y_i as follows:
$Y_i = 1$ if customer i is satisfied and $Y_i = 0$ if customer i is not satisfied.

Accordingly, we define a Bernoulli distributed sample $y_{1:n}$ with $Y_{1:n} \sim_{iid} \mathcal{B}(1-\theta)$. We wand to test the hypotheses H_0: $\theta = \theta_0 = 0.52$ and $H1$: $\theta = \theta_1 = 0.48$.

1.

Construct the likelihood of the observations $y_{1:n}$ and explain the rejection region of H_0 from the test of Neyman and Pearson. Assume $\alpha = 0.1$ for numerical application.

Solution:

We observe $\theta_1 = 1 - \theta_0$ and the other way around.

$\mathcal{L}(\theta) = \mathcal{L}(\theta; y_{1:n}) = p(y_{1:n}; \theta) = \prod_{i=1}^n p(y_i; \theta) = \prod_{i=1}^n (1-\theta)^{y_i} \cdot \theta^{1-y_i} = (1-\theta)^{\sum_{i=1}^n} \cdot \theta^{n-\sum_{i=1}^n}$

$\implies \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{(1-\theta_1)^{\sum_{i=1}^n} \cdot \theta_1^{n-\sum_{i=1}^n}}{(1-\theta_0)^{\sum_{i=1}^n} \cdot \theta_0^{n-\sum_{i=1}^n}} = \frac{\theta_0^{\sum_{i=1}^n} \cdot \theta_1^{n-\sum_{i=1}^n}}{\theta_1^{\sum_{i=1}^n} \cdot \theta_0^{n-\sum_{i=1}^n}} = \frac{\theta_1^{n-2\cdot\sum_{i=1}^n}}{\theta_0^{n-2\cdot\sum_{i=1}^n}} = (\frac{\theta_1}{\theta_0})^{n-2\cdot\sum_{i=1}^n y_i} = (\frac{\theta_1}{\theta_0})^n \cdot (\frac{\theta_0}{\theta_1})^{2\cdot\sum_{i=1}^n y_i}$.

$\theta_0 > \theta_1 \implies \frac{\theta_0}{\theta_1} > 1$ and thus $(\frac{\theta_0}{\theta_1})^{2\cdot\sum_{i=1}^n y_i}$ is increasing for $\sum_{i=1}^n y_i$ increasing. This has the following implication for our Likelihood-Ratio test:

$$\Lambda_{LR}(y_{1:n}) = \begin{cases} 1, & \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} > k \\ \gamma, & \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = k \\ 0, & \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} < k \end{cases}$$

$\iff$

$$\Lambda_{LR}(y_{1:n}) = \begin{cases} 1, & \sum_{i=1}^n y_i > c \\ \gamma, & \sum_{i=1}^n y_i = c \\ 0, & \sum_{i=1}^n y_i < c \end{cases}$$

.

From the lecture we know that:

$\mathbb{E}_{\theta_0}[\Lambda_{LR}(Y_{1:n})] = 1 \cdot \mathbb{P}_{\theta_0}[\sum_{i=1}^n Y_i > c] + \gamma \cdot \mathbb{P}_{\theta_0}[\sum_{i=1}^n Y_i = c] = \alpha = 0.1$

$\implies \mathbb{P}_{\theta_0}[\sum_{i=1}^n Y_i > c] \leq 0.1$ and $\mathbb{P}_{\theta_0}[\sum_{i=1}^n Y_i \geq c] > 0.1$ (1)

$\iff \mathbb{P}_{\theta_0}[\sum_{i=1}^n Y_i \leq c] \geq 0.9$ and $\mathbb{P}_{\theta_0}[\sum_{i=1}^n Y_i < c] < 0.9$.

As we don't know the sample size n, it is rather difficult to accurately compute the constant value c. However, from the context of the task we may assume that n is sufficiently large ($n >> 50$) as a survey would be rather pointless otherwise. This allows to utilize the theorem of Moivre and Laplace to approximate the given situation with the standard normal distribution:

$\implies \mathbb{P}_{\theta_0}[\frac{\sum_{i=1}^n Y_i - n\cdot(1-\theta_0)}{\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}} \leq \frac{\tilde{c}-n\cdot(1-\theta_0)}{\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}}] \geq 0.9$ and $\mathbb{P}_{\theta_0}[\frac{\sum_{i=1}^n Y_i - n\cdot(1-\theta_0)}{\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}} < \frac{\tilde{c}-n\cdot(1-\theta_0)}{\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}}] < 0.9$ with $\frac{\sum_{i=1}^n Y_i - n\cdot(1-\theta_0)}{\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}} \sim \mathcal{N}(0,1)$

So we are basically solving for the 0.9-quantile of the standard normal distribution with

$\implies z_{0.9} = \frac{\tilde{c}-n\cdot(1-\theta_0)}{\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}} \iff \tilde{c} = \sqrt{n\cdot\theta_0\cdot(1-\theta_0)}\cdot z_{0.9} + n\cdot(1-\theta_0)$

or in numbers:

```
qnorm(0.9)*sqrt(0.54*(1-0.52))
```

```
## [1] 0.6524595
```

$\iff \tilde{c} = 0.6525\cdot\sqrt{n} + 0.48\cdot n = n\cdot(\frac{0.6525}{\sqrt{n}} + 0.48)$. In short, we are treating $\tilde{c}$ as a function of n, our sample size. As we are actually dealing with a discrete distribution just approximated by a continuous distribution, $\tilde{x}$ needs to be rounded accordingly for (1) to be true. Likely this means $c = \lceil\tilde{c}\rceil$ in order to ensure everything holds true.

So what does all of this mean for the rejection region of the test? H_0 is rejected if $\Lambda_{LR}(y_{1:n}) = 1$. This is equivalent to $\sum_{i=1}^n y_i > c = \lceil\tilde{c}\rceil = \lceil n\cdot(\frac{0.6525}{\sqrt{n}} + 0.48)\rceil$. As by construction, $\alpha \leq 0.1$.

A whole different story would be to compute γ, however, in this we are not interested as of now.

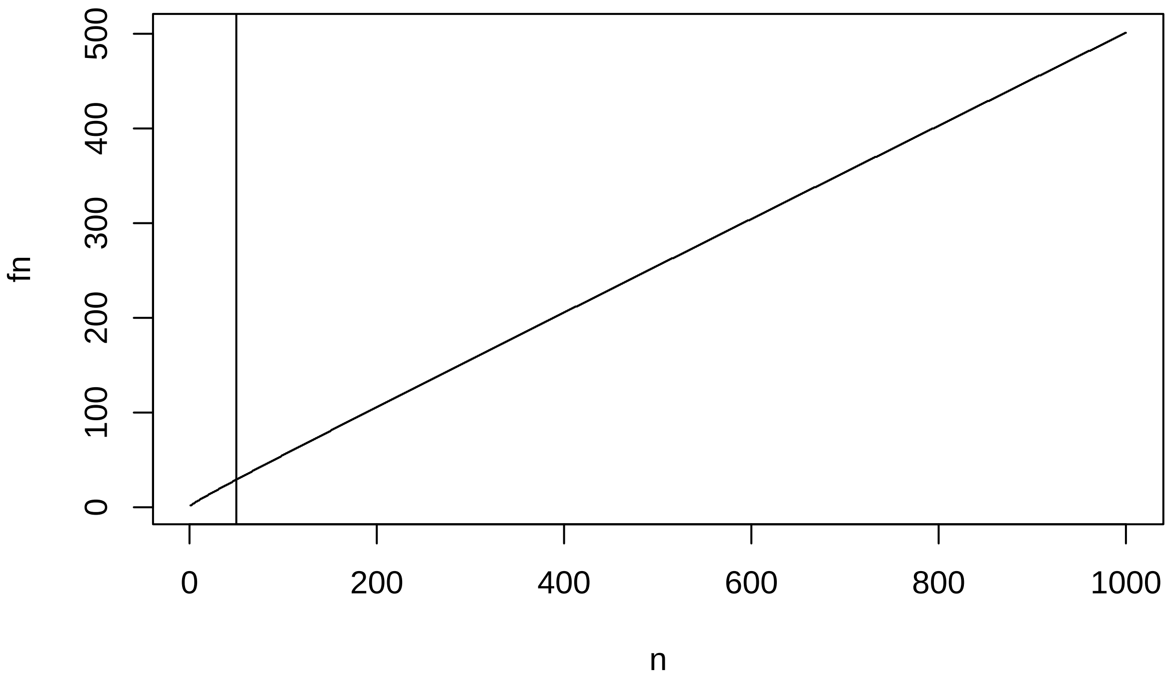
```
f <- function(x){
  return(ceiling(x*((0.6525/sqrt(x))+0.48)))
}

n <- 1:1000

fn <- f(n)
c <- round(fn/n, digits = 2)

summary(c[51:1000])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5000  0.5000  0.5100  0.5149  0.5200  0.5900
```

```
summary(c[901:1000])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.5     0.5     0.5     0.5     0.5     0.5
```

```
plot(x = n, y = fn, type = "l")
abline(v = 50)
```



2.

Determine $\mathbb{P}[H_0 \text{ rejected}|H_1 \text{ true}]$.

Solution:

we deal with the task similarly to before. We already know the rejection region for H_0 and thus just need to assume that H_1 is true: $\mathbb{P}_{\theta_1}[\sum_{i=1}^n Y_i > c]$.

From construction we know that $\sum_{i=1}^n Y_i > c > \tilde{c} > (\sum_{i=1}^n Y_i) - 1$.

Similar to before, to actually work with this without knowing the sample size n, we need to assume that n is sufficiently large in order to use Moivre-Laplace:

$\mathbb{P}_{\theta_1}[\sum_{i=1}^n Y_i > c] = \mathbb{P}_{\theta_1}[\sum_{i=1}^n Y_i > \tilde{c}] \approx \mathbb{P}_{\theta_1}[\frac{\sum_{i=1}^n Y_i - n\cdot(1-\theta_1)}{\sqrt{n\cdot\theta_1\cdot(1-\theta_1)}} > \frac{\tilde{c}-n\cdot(1-\theta_1)}{\sqrt{n\cdot\theta_1\cdot(1-\theta_1)}}]$ with $\frac{\sum_{i=1}^n Y_i - n\cdot(1-\theta_1)}{\sqrt{n\cdot\theta_1\cdot(1-\theta_1)}} \sim \mathcal{N}(0,1)$ for $Y_i \sim_{iid} \mathcal{B}(1-\theta_1)$.

Let's investigate the other side of the inequality:

$\frac{\tilde{c}-n\cdot(1-\theta_1)}{\sqrt{n\cdot\theta_1\cdot(1-\theta_1)}} = \frac{(\sqrt{n\cdot\theta_0\cdot(1-\theta_0)}\cdot z_{0.9}+n\cdot(1-\theta_0))-n\cdot(1-\theta_1)}{\sqrt{n\cdot\theta_1\cdot(1-\theta_1)}}$

$= z_{0,9} + \sqrt{n\cdot\frac{\theta_1}{\theta_0}} - \sqrt{n\cdot\frac{\theta_0}{\theta_1}} = 1,282 - 0.08\cdot\sqrt{n}$

```r
qnorm(0.9)
```

```
## [1] 1.281552
```

```
sqrt(0.48/0.52)-sqrt(0.52/0.48)
```

```
## [1] -0.08006408
```

So in conclusion, $\mathbb{P}_{\theta_1}[\sum_{i=1}^{n} Y_i > c] \approx \mathbb{P}[N > 1,282 - 0.08 \cdot \sqrt{n}]$ with $N \sim \mathcal{N}(0,1)$. What is interesting about this result? With increasing sample size n, the probability for H_0 to be rejected if H_1 is true also increases. This is something good and desirable for a test!

```
h <- function(x){
  return(1.282-0.08*sqrt(x))
}

hn <- h(n)

p <- 1 - pnorm(hn)

plot(x = n, y = p, type = "l")
abline(v = 50)
```