Hypothesis Testing

JamesQuacinella

April 19, 2017

Z Scores

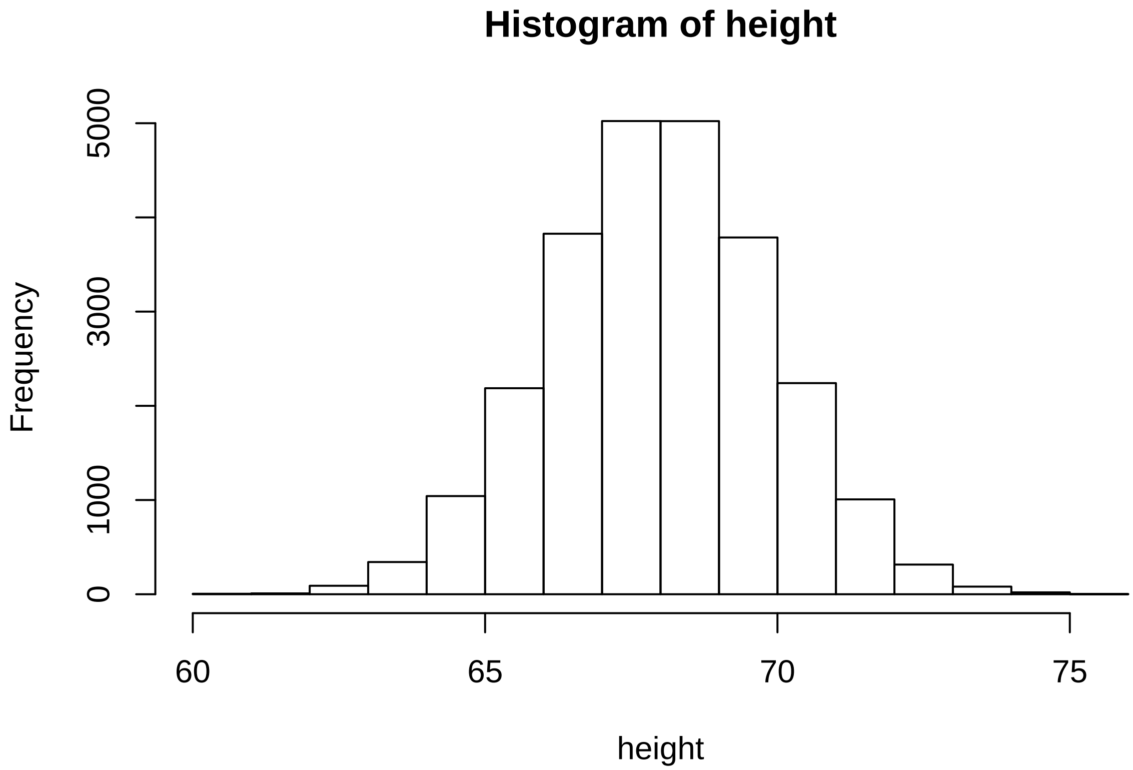
Here I will load a data set of heights and weights of people from Hong Kong:

```
data <- read.csv('~/Desktop/height_data.csv')
head(data)
```

```
##   index   height   weight
## 1     1 65.78331 112.9925
## 2     2 71.51521 136.4873
## 3     3 69.39874 153.0269
## 4     4 68.21660 142.3354
## 5     5 67.78781 144.2971
## 6     6 68.69784 123.3024
```

Lets inspect the height data:

```
height <- data$height
hist(height) #histogram
```



Its easy in R to find the parameters for the population distribution:

```
#population parameter calculations
N <- length(height)  # Population size
```

```r
pop_sd <- sd(height) * sqrt( (length(height) - 1) / (length(height)) )   # Why this correction? sd in R
                                                                         # a sample sd, which has n-1 i
                                                                         # Bessel's correction. Advance

pop_mean <- mean(height) # Mean of the data

writeLines(paste("Population mean:", pop_mean))
```

```
## Population mean: 67.9931135968
```

```r
writeLines(paste("Population   sd:", pop_sd))
```

```
## Population   sd: 1.90164073724984
```

Lets review z-scores. Z-scores allow us to talk about values in terms of how far that value is from the mean in units of standard deviation, What is the probability of getting a height of < 72inches? How do we convert that to a probability > 72in?

```r
data_point = 72

# z-score calculation
z <- (data_point - pop_mean) / pop_sd

p_less_than1 <- pnorm(data_point, pop_mean, pop_sd)    #using x, mu, and sigma
p_less_than2 <- pnorm(z)                               #using z-score of 2.107

# Convert to prob > the data point
p_greater1 <- 1 - p_less_than1
p_greater2 <- 1 - p_less_than2


# Probability of greater or equal to 72 is same no matter which way we calculate
writeLines(paste("Prob >= 72in:", p_greater1))
```

```
## Prob >= 72in: 0.0175558415018912
```

```r
writeLines(paste("Prob >= 72in:", p_greater2))
```

```
## Prob >= 72in: 0.0175558415018912
```

Hypothesis Testing

Here we are going to take a biased sample from the height data, focusing more on taller people than shorter. We do this to setup a hypothesis test to see if this sample is statistically different than the population.

```r
# Sample size
n <- 50

# tall-biased sample
cut <- 1:N
weights <- cut^.6
sorted_height <- sort(height)

set.seed(123)
```

```r
height_sample_biased <- sample(sorted_height, size=n, prob=weights)
head(height_sample_biased)
```

```
## [1] 69.66027 67.41253 69.09773 66.76883 66.19564 71.57334
```

The null hypothesis is that there is no difference in the means of the population and the sample. The alternative hypothesis, as opposed to the web page this was taken from, will be that the sample mean is greater than the population mean.

Now that we have the sample, lets find the sample mean and sample sd:

```r
# Calculate the sample mean and sd
sample_mean <- mean(height_sample_biased)
sample_sd <- sd(height_sample_biased)

writeLines(paste("Sample mean:", sample_mean))
```

```
## Sample mean: 68.5925406
```

```r
writeLines(paste("Sample   sd:", sample_sd))
```

```
## Sample   sd: 1.65874447545322
```

```r
#t-stat
t <- (sample_mean - pop_mean) / (sample_sd / sqrt(n))
t
```

```
## [1] 2.5553
```

From this t score we can calculate the probability of getting a sample mean greater than or equal to what we observed:

```r
#p-value for t-test
p_value <- 1 - pt(t, n-1)
p_value
```

```
## [1] 0.006882297
```

We can do this using t.test:

```r
# http://www.stat.columbia.edu/~martin/W2024/R2.pdf
t.test(height_sample_biased, alternative="greater", mu=pop_mean)
```

```
##
##  One Sample t-test
##
## data:  height_sample_biased
## t = 2.5553, df = 49, p-value = 0.006882
## alternative hypothesis: true mean is greater than 67.99311
## 95 percent confidence interval:
##  68.19925      Inf
## sample estimates:
## mean of x
##  68.59254
```