Confidence in the Sample Mean

Chase Baggett

September 21, 2017

2.1 An investigator found that a random sample of size 100 from a normal population resulted in a 95 percent confidence interval for the population mean with length 4. The investigator would like a confidence interval with length about 1. About how many observations should the investigator take to obtain a confidence interval with length about 1?

Because the standard error is divided by the square root of the standard deviation, we say that information is gained at the square rate. Because of this, to make our confidence interval 4 times better, we have to increase our sample size by 16.

I have included an example in code.

```
library(alr4)
library(ggpubr)
library(Hmisc)
perfect_normal <- function(n,mean,sd) { mean+sd*scale(rnorm(n)) }
original_sample_size <- 100
sample <- perfect_normal(100,mean=50,sd=40)
std.error <- function(x){sd(x)/sqrt(length(x))}
std.error(sample)
```

```
## [1] 4
```

```
new_sample_size <- 1600
new_sample <- perfect_normal(new_sample_size,mean=50,sd=40)
std.error(new_sample)
```

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

2.2 (b) The investigator would like to test simultaneously the hypotheses mean = 0 and variance = 1 about mean and variance. Suggest a test statistic and a test procedure, and discuss their rationale.

We could hypothesis test the mean with a one sample t test. After that, a Chi-square test would allow us to test the variance. If we create a t statistic of the sum of squared residuals around the hypothesis mean, divided by the hypothesis variance of 1, then t is a Chi-squared distribution. We can identify the degrees of freedom and test the t statistic to a specific confidence level.

2.3 It is reasonable to treat the data for problem 1.1 in your text as a random sample? If so, what is your reason for this conclusion and what is the population? If not, what is the point of studying the data?

No, it is not a random sample of the earth's population because they only collected data from UN countries. It might, however, be considered a random sample of the sub-population of UN member countries, as they don't explcitly tell us how they collected it.

Studying non-random samples can still be useful outside the realm of statistics. Without the desire to infer anything from the population, It still says something about the people who were sampled, even if we can't say anything about people not sampled.

2.5 Assume that mheight in the data discussed in Section 1.1 in your text is a random sample from a normal distribution with meanmean and variance variance.

(a) derive the variance of the estimator for the estimated variance of the variance.

The estimation variance is the variance divided by the sample size.

```
sd(Heights$mheight)^2 / nrow(Heights)
```

```
## [1] 0.004033827
```

(b) Give the standard error of ^variance.

The standard error is the square root of the estimation variance.

```
sqrt(sd(Heights$mheight)^2 / nrow(Heights))
```

```
## [1] 0.06351241
```

2.6 Assume that (mheight, dheight) in the data discussed in Section 1.1 in your text is a random sample from a bivariate normal population.

(a) Estimate the correlation coefficient and describe what it is estimating.

The correlation coefficient is the measure of linear correlation or dependence between mheight and dheight. Because our value is positive, they are positively correlated.

```
cor(Heights$mheight,Heights$dheight)
```

```
## [1] 0.4907094
```

```
#or
cov(Heights$mheight,Heights$dheight)/(sd(Heights$mheight) * sd(Heights$dheight))
```

```
## [1] 0.4907094
```

(c) Estimate the conditional mean E(dheight | mheight) and variance var(dheigh | mheight), and describe the connection with the regression of part (b).

Part b asked us for the expected value of mheight given dheight, whereas now we are looking for the expected height of a daughter conditional on knowing the mother's height.

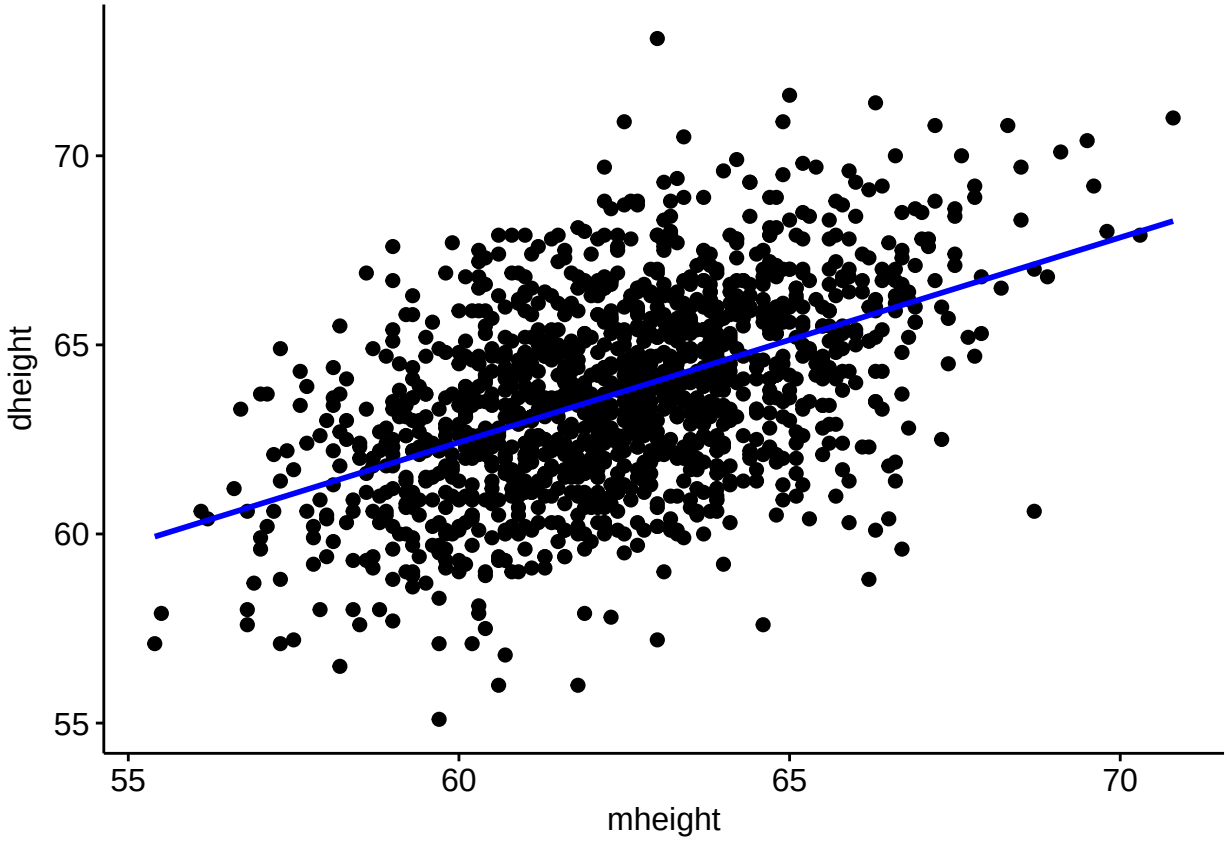
We first fit a regression to the mean to understand the conditional mean, and see that we can expect the conditional mean a daughter's height given her mother's height is 29.91744 + .54175 * mheight.

```
fit <- with(Heights,lm(dheight~mheight))
summary(fit)
```

```
##
## Call:
## lm(formula = dheight ~ mheight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```
ggscatter(Heights,y="dheight",x="mheight",add = "reg.line",add.params = list(color = "blue", fill = "li
```



In a simple linear regression, variance should be constant.

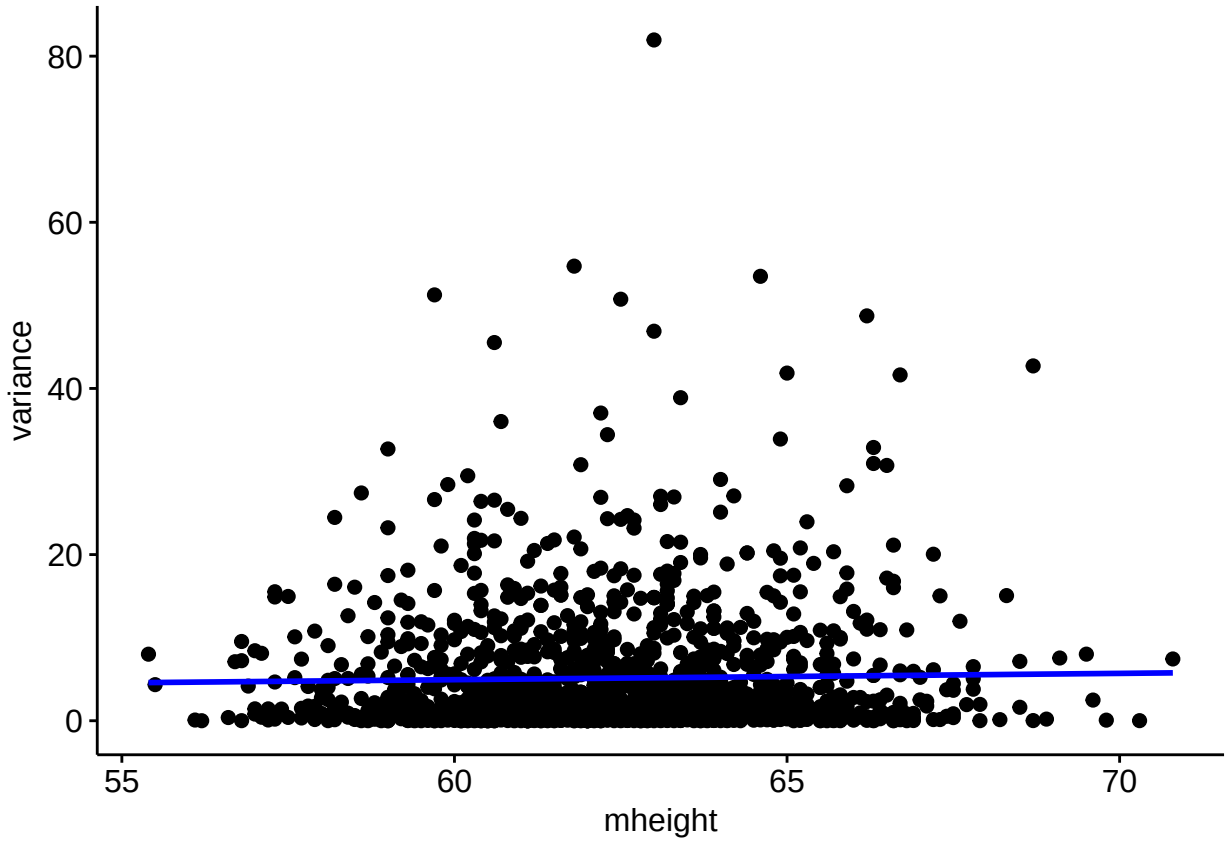
```
summary(fit)$sigma^2.
```

```
## [1] 5.136167
```

To verify the conditional variance is constant I am going to start dealing with residuals of our regression line. So, first, I create a dataset of our residuals. By squaring the residuals I can use the underlying machinery of the regression model to generate a conditional variance, and model how it changes conditioned on mheight.

We are essentially using another regression model to the process variance of our first regression model by asking the expectation of variance given mheight.

```
data <- Heights
data$residuals <- fit$residuals
data$variance <- fit$residuals^2
ggscatter(data,y="variance",x="mheight",add = "reg.line",add.params = list(color = "blue", fill = "ligh
```



It appears we have constant variance within our linear model. This fits our expectation that in a simple linear regression the residuals will have a mean of zero and a constant variance.

(d) A mother's height is 1 standard deviation (for mothers) above the average height for mothers. Her daughter's height is estimated to be ——— (fill in the blank with a number not a symbol) standard deviations (for daughters)

We would say that the daughter's height is equal to 0.4646328 standard deviations, or rho.

```
rho <- rcorr(Heights$dheight,Heights$mheight,type="spearman")
rho$r[1,2]
```

```
## [1] 0.4646328
```