Statistics and Linguistic Applications
Hale
February 15, 2010

Chapter 2 of Vasishth's notes dealt with raindrops hitting Left or Right stones and proportions of balls that are red out of 40 selected from a box. He revealed that we can use the binomial theorem to work out exact probabilities for these events – knowing the true population distribution ensures that we have all the answers. We can even combine probabilities of getting nearby values to establish a region about whose inclusion of the population parameter we can be as sure as we want to be – a confidence interval.

R helps to practice dealing with the situation where you don't have all the answers; where all you have is a random sample and it's unclear exactly which population that sample came from.

Sampling involves Error

In the ideal world, you would measure once and that would be The value. But in the real world, there is variation.

A **statistical model** conceives of the observed values as the addition to the true value of some **error**.

$$\mu + \varepsilon$$

Taking the mean $\bar{X}$ of a sample $X = X_1 ... X_n$ implies averaging over both addends. But if each measurement really does come from the same population the μ are the same. Thus,

$$\bar{X} = \mu + \bar{\varepsilon}$$

There is some historical discussion about this in Vasishth's Appendix 9.

Dice

With dice, we have all the answers. Let D be a random variable corresponding to a six-sided, fair die.

$$\begin{aligned} E[D] &= \sum_{d \in D} dP(d) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} \\ &= 3.5 \end{aligned}$$

The expected number on the face of the die is between 3 and 4. What is the variance and SD?

$$\begin{aligned} Var[D] &= E[D^2] - (E[D])^2 \\ Stddev[D] &= \sqrt{Var[D]} \end{aligned}$$

$$\begin{aligned} E[D] &= \tfrac{7}{2} = 3.5 \\ E[D]^2 &= \tfrac{49}{4} = 12.25 \\ E[D^2] &= \tfrac{1}{6} + 4 \times \tfrac{1}{6} + 9 \times \tfrac{1}{6} + 16 \times \tfrac{1}{6} + 25 \times \tfrac{1}{6} + 36 \times \tfrac{1}{6} = \tfrac{91}{6} = 15.16666666 \\ E[D^2] - (E[D])^2 &= \tfrac{35}{12} = 2.916666 \\ Stddev[D] &= \sqrt{\tfrac{35}{12}} = 1.70783 \end{aligned}$$

In the statistical model of dice, we have that

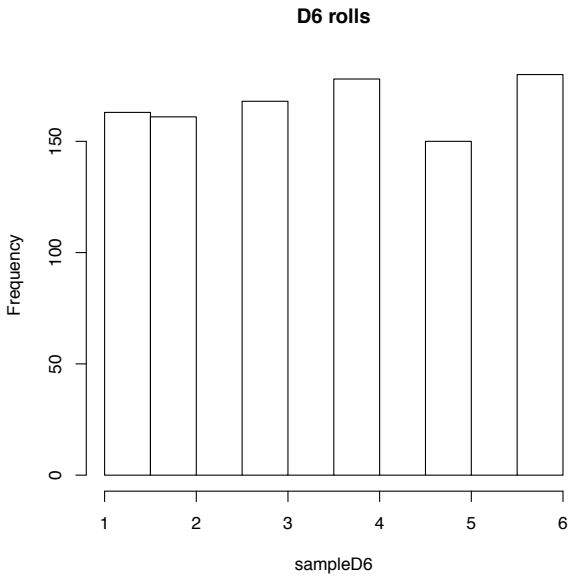
$$D = 3.5 + \varepsilon$$

and because of the way dice are constructed we know that no error is more likely than another. Deviations are are equiprobable:

ε	probability
-2.5	$\frac{1}{6}$
-1.5	$\frac{1}{6}$
-0.5	$\frac{1}{6}$
0.5	$\frac{1}{6}$
1.5	$\frac{1}{6}$
2.5	$\frac{1}{6}$

Histogram of the scores:

```
> sampleD6 <- ceiling(runif(n = 1000, min = 0, max = 6))
> hist(sampleD6, main = "D6 rolls", freq = T)
```



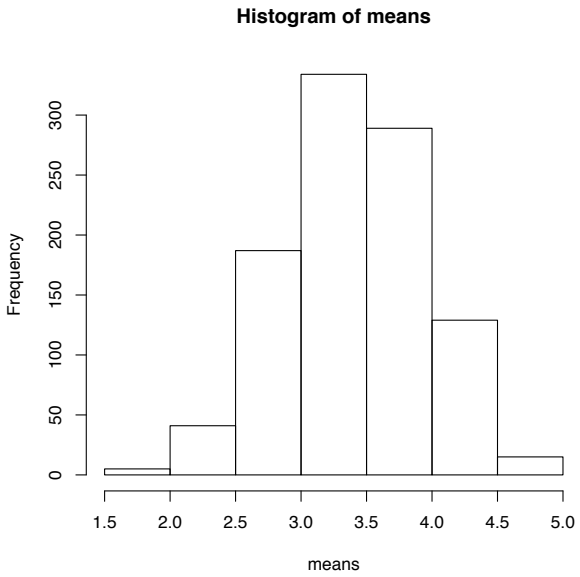
How close was the sample to theory?

```
> var(sampleD6)

[1] 2.931971

> sd(sampleD6)

[1] 1.7123
```

Sampling means

Do a little sample (say, roll 10 D6) then ask 'what is the mean of the sample just obtained?' Such a process might be repeated a thousand times in an effort to determine the true population mean, if we didn't have all the answers already.

```
> means <- c()
> for (i in c(1:1000)) {
+     sampleD6 <- ceiling(runif(n = 10, min = 0, max = 6))
+     currentmean <- mean(sampleD6)
+     means <- append(means, currentmean)
+ }

> hist(means)
```



There is drastically less variablility (*smaller SD*) in the 1000 means than in the scores for single rolls. Of course the mean stays the same.

```
> mean(means)

[1] 3.4621

> sd(means)

[1] 0.5363593
```

Central Limit Theorem

The histogram for individual die rolls is flat (*uniform* distribution) but the histogram for means of samples of length 10 is bell-shaped (*Normal* distribution). This amazing phenomenon is a consequence of the Central Limit Theorem, stated in 4.2 of Vasishth's notes.

Furthermore, there is a precise square-root relationship between the SD of the sampling distribution of the sample mean and the population SD. Vasishth shows you the derivation in Appendix 7, line 55. The proof hinges on the fact that $Var[bX] = b^2 Var[X]$. When b is the "averaging" factor $\frac{1}{n}$ used to find the mean of a sample of size n, this leads to an extra n in denominator, whose square root is taken during the tumultuous passage between variance and standard deviation.