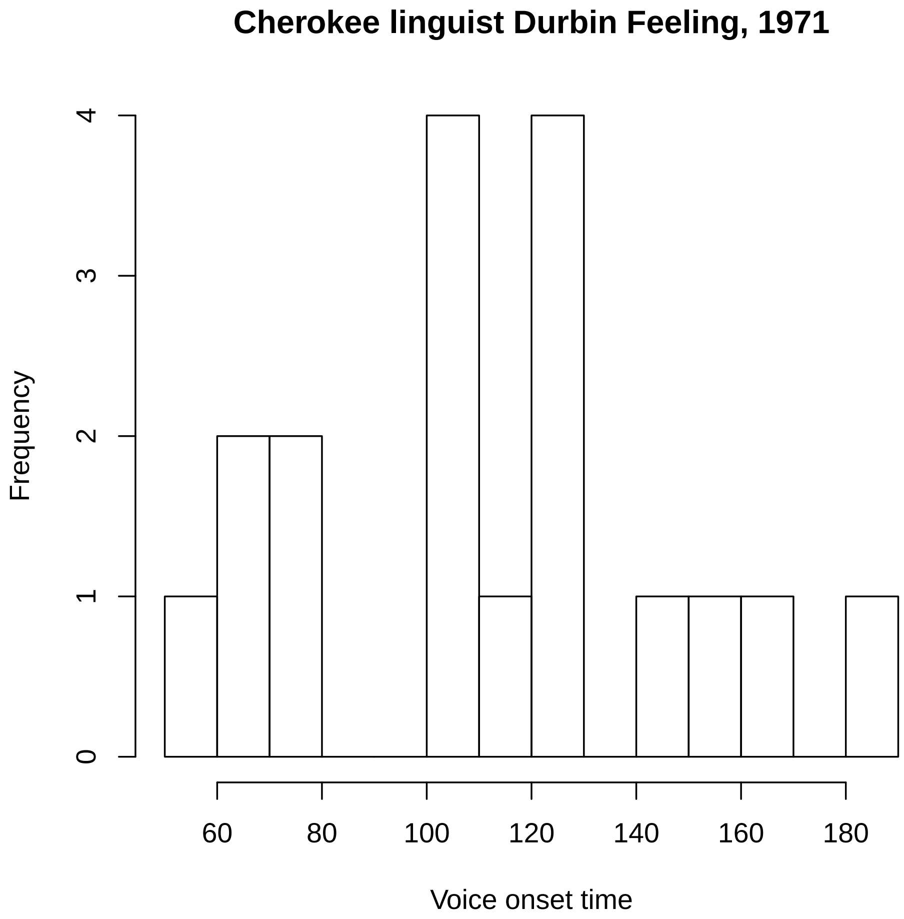
Statistics and Linguistic Applications
Hale
February 3, 2010

1 The Normal approximation

Review of frequency

Remember frequency? That's how often a particular level appears in a data set. Take Keith Johnson's measurements of Cherokee consonant voice onset times (VOT) as an example.

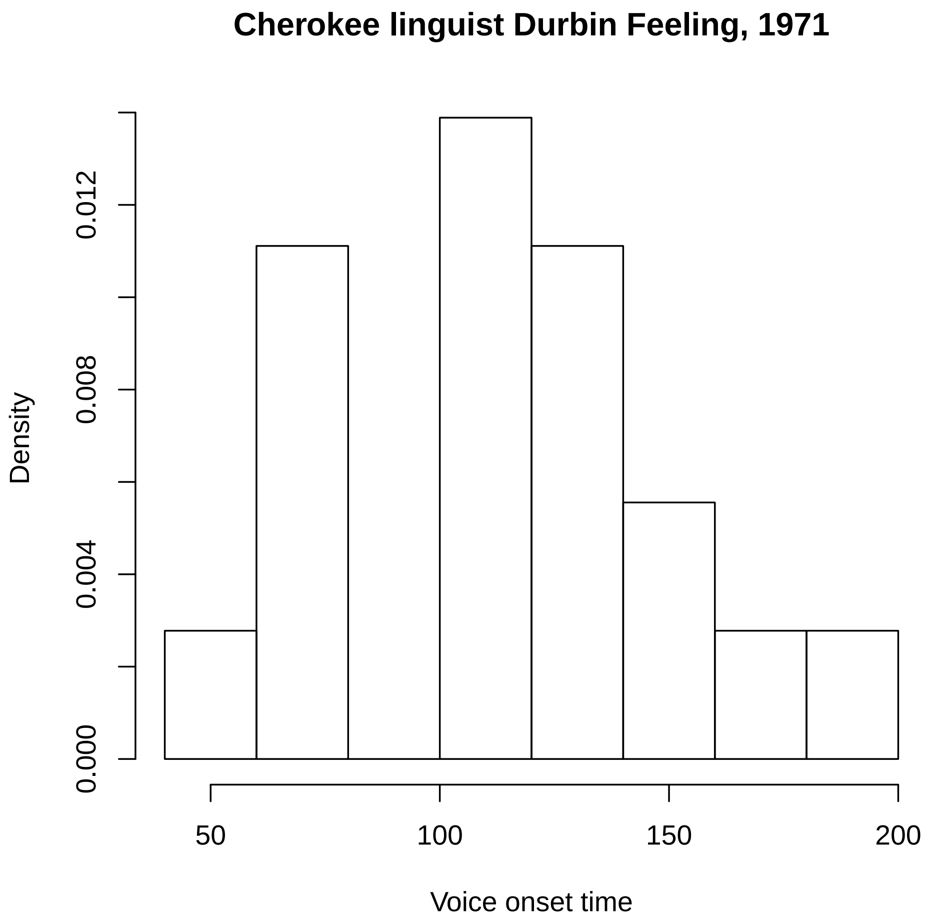
```
> vot71 = c(67, 127, 79, 150, 53, 65, 75, 109, 109, 126, 129, 119,
+     104, 153, 124, 107, 181, 166)
> hist(vot71, xlab = "Voice onset time", main = "Cherokee linguist Durbin Feeling, 1971",
+     breaks = 10)
```



Keith Johnson's data set has four observed VOTs in the interval between 100 and 110 msec (The `breaks=` option to the `hist` function is a request to divide up the data into ten pieces.)

Note that the vertical axis is in terms of counts of how many times Johnson observed a VOT in that interval. Dividing by the number of observations (18) gives the **relative frequency**. Use the `freq=F` option to have R do this for you.

```
> hist(vot71, xlab = "Voice onset time", main = "Cherokee linguist Durbin Feeling, 1971",
+     freq = F)
```

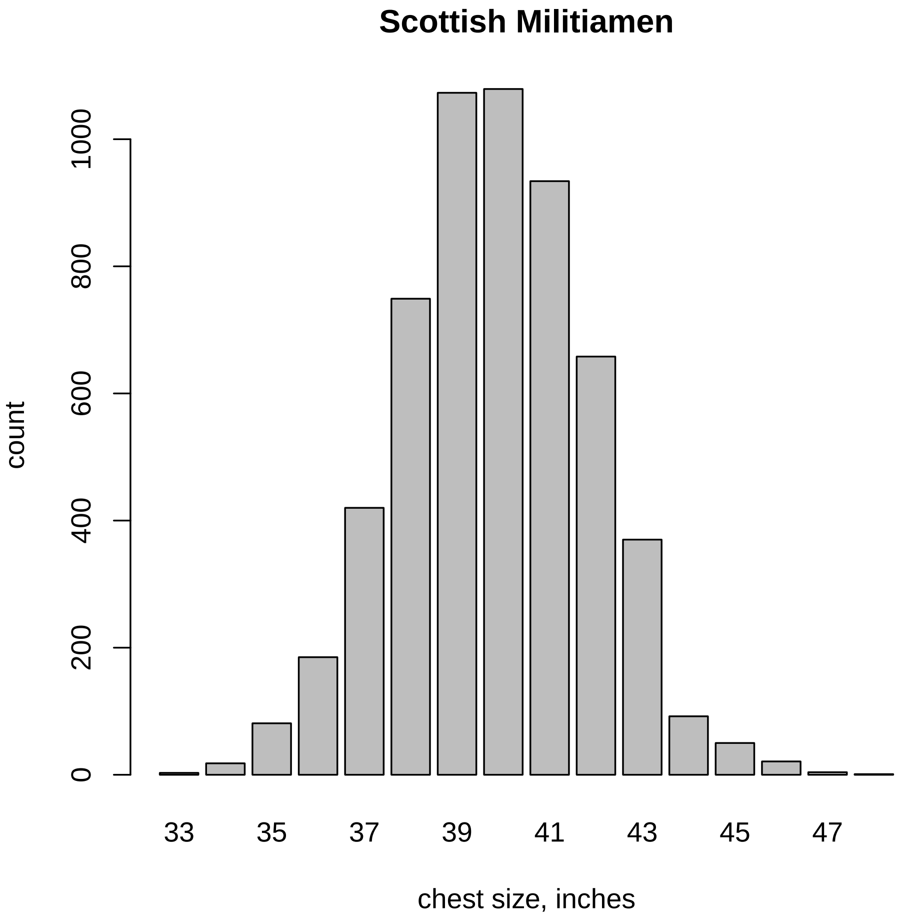


Same picture, different units on the up/down axis. The sizes of the intervals defining the left/right axis matter; it's easier for values to fall into wider intervals. For this reason, it is the **area** and not just the height of the bar that matters. When the up/down axis units are fractions of the data set, rather than actual counts, then the areas of all the bars adds up to 1.0.

The Ideal Histogram

Around 1870, Adolph Quetelet (Belgium) and Francis Galton (England) were finding that the frequencies of natural, individual differences like chest size and height all had the same shape.

```
> quetelet <- read.table("chest-sizes.txt", header = T)
> barplot(quetelet$Count, names.arg = quetelet$Chest, main = "Scottish Militiamen",
+     xlab = "chest size, inches", ylab = "count")
```



This shape turned out to be the **normal curve**. Due to the work of de Moivre, Adrian, Gauss there is an expression for the normal,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(x-\mu)^2/(2\sigma^2)} \tag{1}$$

I include equation 1 only to show you that it really exists; rather than using this expression we can work with diagrams and use the computer to calculate values (they involve

e the base of the natural logarithm, about 2.71828182...

π the ratio of a circle's circumference to its diameter, about 3.141596...

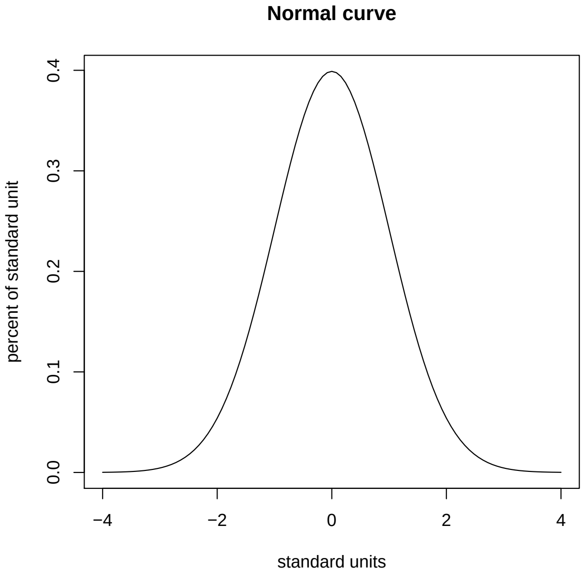
σ how spread out the normal curve shall be (standard deviation)

μ where the normal curve is centered (mean)

)

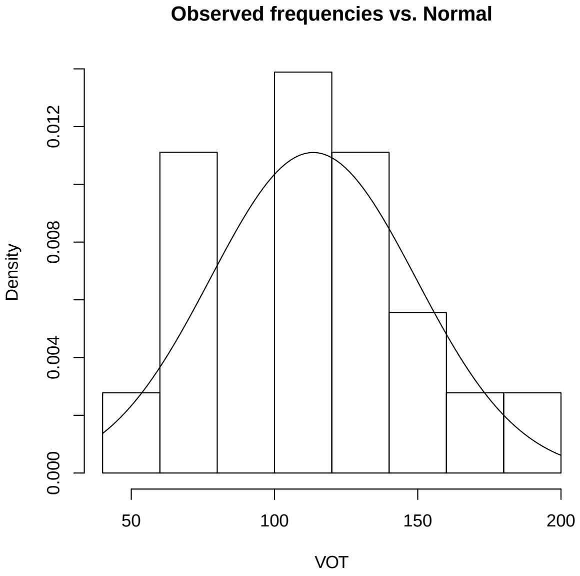
Since e and π are constants, the shape is entirely defined by choices for the **parameters** σ and μ. The R command `dnorm` calculates the density function $f(x)$.

```
> curve(dnorm(x), from = -4, to = 4, xlab = "standard units", ylab = "percent of standard unit",
+     main = "Normal curve")
```



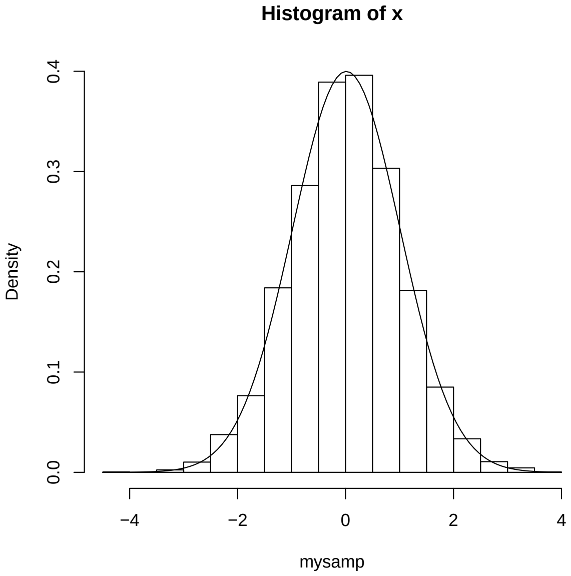
This ideal histrogram can sometimes describe linguistic data like VOT, too. Use Dalgaard's (page 32) little function `hist.with.normal` to compare Durbin Feeling's VOTs with the the ideal.

```
> source("hist.with.normal.R")
> hist.with.normal(vot71, xlab = "VOT", main = "Observed frequencies vs. Normal")
```



Vaguely normal. Of course there are very few observations. If your source is truly normal, larger and larger samples lead to histograms that more closely approximate the ideal curve. Simulate this by asking R to randomly sample from the normal distribution 10,000 times.

```
> mysamp <- rnorm(10000)
> hist.with.normal(mysamp)
```



Properties of the Normal

The Normal curve has some cool properties

symmetric The curve is symmetric about zero, i.e. the part to the right of zero is a mirror image of the part to the left.

unimodal In the Normal, the mean and the mode are the same. As a consequence also of symmetry, the median is also the same as these other two measures of central tendency.

standardizable Using Z-scores, all Normal measurements can be rephrased as having mean= 0 and standard deviation= 1.

By integrating equation 1 or using R's `pnorm` function we can establish facts like:

```
> pnorm(1) - pnorm(-1)

[1] 0.6826895

> pnorm(2) - pnorm(-2)

[1] 0.9544997
```

$$\begin{array}{l}\text{the area under the normal curve between } -1 \text{ and } +1 \text{ is about } 68\% \\ \text{the area under the normal curve between } -2 \text{ and } +2 \text{ is about } 95\%\end{array} \tag{2}$$

This is analogous to a relative-frequency histogram: the area within certain bars corresponds to the proportion of the data set that achieved those scores. Of course, to use facts about the *standard* normal we must first convert measurements to standard units (Z-scores).

Using these properties to make a Normal approximation

We can approximate an actual data set by replacing its actual histogram with the theoretical Normal curve. Then, just by knowing the Normal, we know what percentages of the set should fall within given intervals.

Height example

For instance, say we know that in the US Public Health Service Health and Nutrition Examination Survey, women's heights follow a Normal curve, just like Galton's nobles. The average height is 63.5 inches with standard deviation of 2.5 inches.

By symmetry of the Normal curve, 34.13% should achieve heights 1 standard deviation *above* the mean, and 34.13% should achieve heights 1 standard deviation *below* the mean.

So, in 100 US women, there ought to be about 68 whose height is between $63.5 - 2.5 = 61$ inches and $63.5 + 2.5 = 66$ inches. An even greater majority (95 of them) should be between 58.5 and 68.5 inches. If we sampled 500 US woman, 475 (95%) should be between those heights.

I say "ought" because we are approximating nature with the simplicity of equation 1.

Math anxiety example from Kranzler

Suppose that a test of math anxiety was given to a large group of persons, the scores are assumed to be from a normally distributed population with mean 50 and standard deviation 10. Approximately what percentage of persons earned scores

1. below 50? If the mean is 50, then by symmetry half the scores should be below this value.
2. above 60? Sixty is one standard deviation up from the mean. By symmetry, one SD is about 34%. So, we're talking 50% plus another 34%, that's 84% which should score below 60. But we want *above* 60. Since percentages must add up to 100, subtract 84 from 100 to yield the answer, about 16%.
3. below 30? Thirty is two standard deviations down from 50. So the part that we're NOT in is half, plus two standard deviations. That's $50\% + (95/2)\% = 50 + 47.5 = 97.5\%$. The scores below 30 must be in the remaining 2.5%.
4. above 80? The relevant fact is not on this handout. Ask R `1-pnorm(80,mean=50,sd=10)`
5. between 40 and 60? That's one SD on either side of the mean! So, 68%
6. between 30 and 70?
7. between 60 and 70? That's the difference $+2\text{SD} - +1\text{SD}$. Two standard deviations above the mean accounts for forty-seven and a half percent of the scores. One SD only accounts for 34 percent. So, the scores between 60 and 70 should be $47.5 - 34 = 13.5\%$

VOT

Q. A particular phonetic theory PT holds that the true VOT for Cherokee consonants has mean 150 and standard deviation 36. Given also that VOT is normally distributed, how many examples should fall in the interval between 100 and 110 msec?

A. 110 is $(150 - 110)/36 = 1.11111....$ standard deviations below average. 100 is 1.388 SDs below. The proportion less than 1.1111 SDs `pnorm(-1.1111)` is approximately 0.13. The proportion less than 1.388 `pnorm(-1.388)` is 0.08. In between those two, there was an increase of 0.05. So, about one twentieth of a sample should fall between 100 and 110 msec VOT, according to PT. (But in fact, it was more like one fifth in Keith Johnson's data!)

2 Now you Try

- Suppose that the scores in the `german-proficiency` file (posted on class website) have been obtained after testing a group of 50 children for proficiency in spoken German. Draw a histogram, then make a determination about whether the data are normally distributed or not. Check if the data satisfy the properties in (2).
- Look up and use the builtin R function `qqnorm` to visually confirm or cast doubt on your determination. Does `hist.with.normal` point to the same conclusion?
- How many students would the Normal approximation lead us to expect would score less than fifty-five but still above the mean?