The Normal Distribution

When can we use the simple mean and SD to summarize a list of numbers? One time is when the data are approximately normal. By this, we mean that the distribution looks roughly like the normal curve. First a short review of the normal curve.

Normal curve The standard Normal curve is

$$\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$$

It is typically represented as ϕ. It is symmetric about 0 and has inflection points at +1 and -1. The area under the normal curve from $-\infty$ to z is expressed as $\Phi(z)$:

$$\Phi(z) \;=\; \int_{-\infty}^{z} \phi(x)dx$$

1. Show that the area below $-z$ (take $z > 0$) under the normal curve is the same as the area above z under the curve. That is, show that $\Phi(-z) = 1 - \Phi(z)$.

2. Show that the area between $-z$ and z (for $z > 0$) under the normal curve can be expressed as $\Phi(z) - \Phi(-z)$. This area is approximately .68 when $z = 1$ and .95 when $z = 2$.

3. Show that the area between $-z$ and z (for $z > 0$) under the normal curve can be expressed as $2\Phi(z) - 1$, or equivalently $1 - 2\Phi(-z)$.

The Normal(μ, σ^2) curve has the same shape as the standard normal curve. However, it is symmetric about μ with points of inflection at $\mu + \sigma$ and $\mu - \sigma$. The equation for this curve is:

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

The area between a and b under the $\mathcal{N}(\mu, \sigma^2)$ can be found by using the standard normal curve. In particular, the area between $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$ under the standard normal, i.e.

$$\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

is the same as the area between a and b under the $\mathcal{N}(\mu, \sigma^2)$. This means that we can convert to standard units to find areas under normal curves.

4. Show by a change of variables that the area between a and b is $\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$.

Normal Approximation We often find data that occurs in nature, such as the heights of adult men, follow the normal distribution. Adolphe Quetelet (1796-1874) was one of the first scientist to recognize this phenomenon, and he went about fitting normal curves to lots of different datasets to prove his point.

On the accompanying page, you will find data to which Quetelet fitted the normal distribution. The table gives the chest circumferences of Scottish soldiers (in inches) taken from Stigler's History of Statistics.

The mean chest-size if 40.5 inches and the SD is 2.0 inches.

One way to fit the normal distribution involves the 68-95-99% rule. That is, 68% of the data should be within one SD of the mean, 95% within 2SDs, and 99% within 3SDs, if the data are roughly normal. **Check the 68-95-99% rule for the Scottish soldiers chest sizes**.

Normal Quantiles Another, better, way to check the appropriateness of the normal approximation, is to compare the quantiles of the data to that of the normal distribution.

Consider our data, $x_1, \ldots, x_n$. Let x_q represent the qth sample quantile. That is, at least nq of the x_i are less than or equal to x_q and at least $n(1-q)$ are greater than or equal to x_q.

The **sample cumulative distribution function**, also called the **empirical cumulative distribution function**, gives us the quantiles. It is the function, F_n:

$$F_n(x_q) = q.$$

Well, the inverse of this function gives us the quantiles,

$$F_n^{-1}(q) = x_q.$$

Note that Φ is the cumulative distribution function for the standard normal, and z_q are the standard normal quantiles. Fill in the table with the standard normal quantiles:

z_q	-2	-1	0	1	2
q					

If we plot the pairs of points (z_q, x_q), i.e. plot the standard normal quantiles against the sample quantiles, then if the data are approximately normal, the points should fall on a line.

Why? Because $F_n \approx \Phi$, so

$$\begin{aligned} F_n^{-1}(q) &\approx \Phi^{-1}(q) \\ x_q &\approx z_q \end{aligned}$$

What if F_n is approximately normal, but not standard normal?
What if F_n isn't normal? Take a look at the pictures in your text.