Stat 321 - Day 23
Normal Distributions

Earlier studied a variety of families of discrete probability distributions- binomial, hypergeometric, geometric, negative binomial, and Poisson. Now we will study a variety of families of continuous probability distributions that have applications to data. We will begin with the most important- *normal* (also known as Gaussian) *distributions*.

Example: Miscellaneous Measurements
The Minitab worksheet `MiscMeasurements.mtw` contains sample data on four variables:

- maximum head breadths (in millimeters) on 84 skulls of Etruscan males, measured by anthropologists studying whether Estruscans were native Italians or had immigrated
- chest circumference (in inches, to the nearest inch) of 5732 Scottish soldiers, measured by Belgian mathematician Quetelet who sought to apply mathematical models to physical data
- heights of elderly females (in centimeters), randomly selected from a community in a study of osteoporosis
- scores on a 20-question calculus placement exam administered by a college

(a) Examine a histogram of the distribution of each of these variables (`Graph> Histogram`). Comment on similarities in shape among these distributions.

(b) Draw a smooth curve that approximates (models) the shape of these four distributions of data.

(c) While the shapes are similar, what features differ among these data sets?

(d) Use Minitab to calculate the sample mean and sample standard deviation of these four samples of data. Record them in the table below:

	head breadths	chest circumferences	elderly heights	placement scores
sample mean				
sample std. dev.				

These mound-shaped, symmetric distributions are often modeled by ***normal distributions***. The pdf for the normal distribution with parameters μ (mean) and σ (standard deviation) is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, $-\infty < x < \infty$. Note that this distribution is unimodal and symmetric about its mean μ. The inflection points of the normal curve correspond to x = μ+σ and x = μ-σ.

This pdf does not have a closed-form expression for its integral, so we can not integrate analytically in order to find probabilities for normal distributions. We will use tabulated values and/or Minitab to calculate normal probabilities.

(e) For each of the four data sets, determine the proportion of measurements that fall within one standard deviation of the mean. [*Hints*: For example, with the Etruscan skull measurements, determine that the values of 143.77±5.97 are (137.80, 149.74). Then count how many measurements fall within that interval by: `MTB> let c5=(c1>137.8 & c1<149.74)` and `MTB> tally c5`.] Record these in the table below:

	mean - std. dev.	mean + std. dev.	number in there	proportion
head breadths				
chest circumfer				
heights				
placement scores				

That these proportions are all similar suggests that probability calculations involving normal distributions depend on units of standard deviations away from the mean.

The normal distribution with parameters μ=0 and σ=1 is called a *standard normal distribution*, denoted by Z. Any normal random variable X can be *standardized* by: Z=(X-μ)/σ. For a specific value x, its standardized value is called its *z-score*. Cumulative probabilities for the standard normal distribution have been extensively tabulated.

(f) Use Minitab to produce standardized values of the Etruscan head breadth measurements:

```
MTB> let c11=(c1-143.77)/5.97
```

Count how many and determine what proportion of the standardized scores fall below -1:

```
MTB> let c12=(c11<-1)
MTB> count c12
```

Verify that this is the same count and proportion of the head breadths that fall below 137.8 (the value for which the z-score equals -1):

```
MTB> let c13=(c1<137.8)
MTB> count c13
```

This exercise should convince you that the standard normal distribution is sufficient for finding probabilities involving any normal distribution. In other words, $P(X \le k) = P[Z \le (k-\mu)/\sigma]$.

Example: Birth Weights

Birthweights of babies in the United States can be modeled by a normal distribution with mean μ=3250 grams and standard deviation σ=550 grams. Those weighing less than 2500 grams are considered to be of low birthweight.

(g) Draw a sketch of this normal distribution. Please label the axis, and estimate the scale as well as you can based on the mean and standard deviation. Shade in the region whose area corresponds to the probability that a baby will have a low birthweight.

(h) Based on this shaded region (remembering that the total area under the normal curve is one), make an educated guess as to the proportion of babies born with a low birthweight.

(i) Calculate the *z*-score for a birthweight of 2500 grams.

(j) Look up this *z*-score in a table of standard normal probabilities to determine the proportion of babies are born with a low birthweight (which is the probability that a randomly selected baby would be born with a low birthweight).

(k) What proportion of babies would the normal distribution predict as weighing more than 10 ounces (4536 grams) at birth. [*Hints*: Always start with a sketch of the normal curve and the area you are looking for. Then calculate the *z*-score. Finally, recognize that the tabled value for this *z*-score is not exactly what you're looking for but is closely related to it.]

(l) Describe two different ways that you could have used the table to answer (e). [*Hint*: Make use of symmetry.]

(m) Determine the probability that a randomly selected baby weighs between 3000 and 4000 grams at birth. [*Hints*: Again draw a sketch, and determine what to do with the tabled values for the two relevant *z*-scores.]

Data from the *National Vital Statistics Report* indicates that there were 3,880,894 births in the United States in 1997. A total of 291,154 babies were of low birth weight, while 2,552,852 babies weighed between 3000 and 4000 grams.

(n) Calculate the observed proportions in each of these two groups, and comment on how well the normal calculations in (d) and (g) approximate these values.

(o) How little would a baby have to weigh to be among the lightest 2.5% of all newborns? [*Hints*: Once again start with a sketch. You will need to read the table "in reverse," looking up the area in the middle of the table and reading backwards to find the relevant z-score. Then you will have to un-convert the z-score back to the birthweight scale.]

(p) How much would a baby have to weigh to be among the heaviest 10% of all newborns?

You can use Minitab to perform normal calculations directly by using `Calc> Probability Distributions> Normal`, entering the mean and standard deviation, and then clicking on either "cumulative probability" to find a probability or on "inverse cumulative probability" to find a percentile. You can also use the java applet available from our course webpage.

(q) Use Minitab and/or the applet to verify all of your calculations for the normal birthweight model above.