Activity: Normal Distributions

Concepts: The normal distribution, the standard normal distribution, z-scores, computing probabilities and percentiles from normal distributions, the empirical rule.

Prerequisites: The student should have an understanding of continuous probability distributions.

Just as we earlier studied a variety of families of discrete probability distributions (binomial, hypergeometric, uniform, etc.), we will now study a variety of families of continuous probability distributions that have applications to data. First let's recall what we know about continuous probability distributions:

- Continuous random variable: takes all possible values in an interval
- Probability density function (pdf): $f(x)$ such that $P(a \le X \le b) = \int_a^b f(x)dx$
  - Need $f(x) \ge 0$ for all x and $\int_{-\infty}^{\infty} f(x)dx = 1$ Note: $P(X=x)=0$ and $P(X \ge x) = P(X>x)$
- Cumulative distribution function: $F(x) = P(X \le x)$ for all x
  - $F(x) = \int_{-\infty}^{x} f(y)dy$ Note: $P(a \le X \le b) = F(b)-F(a)$
- Percentiles: the $(100p)^{th}$ percentile is the value of x (call it x_p), that has probability p falling below, i.e., $P(X \le x_p)=p$. (Set the integral of the pdf or the cdf to p and solve for x_p.)
- Expected Value: $E(X)=\int_{-\infty}^{\infty} xf(x)dx$ Variance: $V(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = E(X^2)-[E(X)]^2$

We will begin with the most important family of continuous distribution - the *normal* (also known as Gaussian) *distributions*.

Scenario: Miscellaneous Measurements
The Minitab worksheet `MiscMeasurements_Student.mtw` (located on the shared drive) 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 1000 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 or `hist c1-c4`). Note that you can select Window > Tile to rearrange the graphs. Comment on similarities in shape among these distributions.

(b) What are the key differences among these distributions? Suggest two characteristics that could be used as "parameters" of this distribution (what two things do you think best differentiate the distributions from each other).

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

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. You will be asked (in your homework assignment) to show that the inflection points of the normal curve correspond to $x = \mu+\sigma$ and $x = \mu-\sigma$.

(c) Use Minitab (type `MTB> describe c1-c4`) to calculate the sample mean and sample standard deviation of these four samples of data. Record them in the table below and suggest models for each distribution:

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

The normal 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. Luckily we have alternatives. First, we will "standardize" the data by subtracting the mean and dividing by the standard deviation. Then we will use tabulated values and/or Minitab to calculate normal probabilities. To see why this will work try the following:

(d) 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$\pm$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 $\mu=0$ and $\sigma=1$ is called a *standard normal distribution*, denoted by Z. Any normal random variable X can be *standardized* by: $Z=(X-\mu)/\sigma$. Note that the Z value gives the number of standard deviations away from the mean for the corresponding X value. For a specific value *x*, its standardized value is called its *z-score*. Cumulative probabilities for the standard normal distribution have been extensively tabulated.

(e) 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
```

(f) Produce a histogram and numerical summaries (`describe c11`) for the standardized Etruscan measurements. Describe the shape, center, and spread (location of inflection points) of this distribution.

(g) Now produce standardized values of the Scottish chest measurement data. (Hint: Another way to get standardized values is to use pull down menus in Minitab: Go to Calc|Standardize and then select the column to standardize, fill out how you want to standardize the data, and specify in what column you want to put the standardized scores.) Examine the histogram and numerical summaries of the standardized data. Describe the shape, center, and spread (location of inflection points) of this distribution.

Calculating Probabilities and Percentiles from Normal Distributions:

This exercise should convince you that the standard normal distribution is sufficient for finding probabilities involving any normal distribution. In other words, $P(X \leq k) = P[Z \leq (k-\mu)/\sigma]$. We will use the fact that $P(X \leq x)=P(Z \leq (x-\mu)/\sigma)$ to find $P(X \leq x)$. We will first standardize the observation and then find $P(Z \leq z)$ where these probabilities have been tabulated in Tables Va an Vb (on pages 656-7 of your text). We can also denote $P(Z \leq z)=\Phi(z)$.

Scenario: Birth Weights
Birthweights of babies in the United States can be modeled by a normal distribution with mean $\mu=3250$ grams and standard deviation $\sigma=550$ grams. Those weighing less than 2500 grams are considered to be of low birthweight.

(a) 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.

(b) 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.

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

(d) 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).

(e) 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 values for this *z*-score may not be exactly what you're looking for but are closely related to it.]

(f) 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.]

(g) 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.]

(h) 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.

(i) Use Minitab to verify all of your calculations for the normal birthweight model above.

The Empirical Rule:

(a) Determine (using the table of standard normal probabilities or Minitab) the probability that a normally distributed random variable falls within one standard deviation of its mean.

(b) Is this probability close to the proportions of sample data falling within one standard deviation of its mean (for the Etruscan skull data, Scottish chest circumferences, elderly women's heights, and placement exam scores) that you found previously?

(c) Repeat (a) for falling within *two* standard deviations of its mean.

(d) Repeat (a) for falling within *three* standard deviations of its mean.

Your calculations should confirm what is sometimes called the *empirical rule*: with normally distributed data, 68% fall within one standard deviation of the mean, 95% within two standard deviations, and almost all (99.7%) within three standard deviations.

Scenario: SAT vs. ACT Scores

Scores on the Scholastic Aptitude Test (SAT) follow roughly a normal distribution with mean μ=1000 and standard deviation σ =200. Scores on the American College Test (ACT) follow roughly a normal distribution with mean μ =20.6 and standard deviation σ=6.2. Suppose Bobby scores 1180 on the SAT but Kathy scores 28 on the ACT.

(a) To compare these two scores, we can *standardize* them. Subtract the corresponding mean and divide by the corresponding standard deviation to see "how many standard deviations above the mean" each student falls compared to their peers.

Bobby: Kathy:

(b) Based on these *z*-scores, which student would you say performed better? Explain how this relates to the probabilities lying below these *z* values.