Lecture 1
Review of Fundamental Statistical Concepts

Measures of Central Tendency and Dispersion

A word about notation for this class:

Individuals in a population are designated Y_i, where the index "i" ranges from 1 to N, and N is the total number of individuals in the population. Individuals in a *random sample* taken from a population are also denoted Y_i, but in this case the index "i" ranges from 1 to n, where n is the total number of individuals in the *sample*.

Greek letters will be used for population parameters (e.g. μ = population mean; σ^2 = population variance), while Roman letters will be used for estimates of population parameters, based on random sampling (e.g. $\overline{Y}$ = sample mean $\approx$ population mean = μ; s^2 = sample variance $\approx$ population variance = σ^2).

Basic formulas:

Mean or average (a measure of central tendency)

Population mean: $$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}$$

Sample mean: $$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Variance (a measure of dispersion of individuals about the mean)

Population variance: $$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

Sample variance: $$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}$$

The quantities $(Y_i - \overline{Y})$ are called the *deviations.*

Standard deviation (a measure of dispersion in the original units of observation)

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

Sample standard deviation: $s = \sqrt{s^2}$

Coefficient of variation

In some situations, it is useful to express the standard deviation in units of the population mean. For this purpose, we have a quantity called the coefficient of variation:

Population coefficient of variation: $CV = \frac{\sigma}{\mu}$

Sample coefficient of variation: $CV = \frac{s}{\bar{Y}}$

Measures of dispersion of *sample* means

Another important population parameter we will work with in this class is the *sample variance of the mean* ($\sigma^2_{\bar{Y}}$). If you repeatedly sample a population by taking samples of size n, the variance of those sample means is what we call the sample variance of the mean. It relates very simply to the population variance, in this way:

Variance of the mean: $\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n}$

We can *estimate* $\sigma^2_{\bar{Y}}$ for a population by taking r independent, random samples of size n from that population, calculating the sample means $\bar{Y}_i$, and then calculating the variance of those sample means. In other words, if $\bar{Y}_i$ is the *mean* of the ith sample and $\bar{\bar{Y}}$ is the overall mean for all r samples, then what we find is:

$$s^2_{\bar{Y}} = \frac{\sum_{i=1}^{r} (\bar{Y}_i - \bar{\bar{Y}})^2}{r-1} \cong \sigma^2_{\bar{Y}}$$

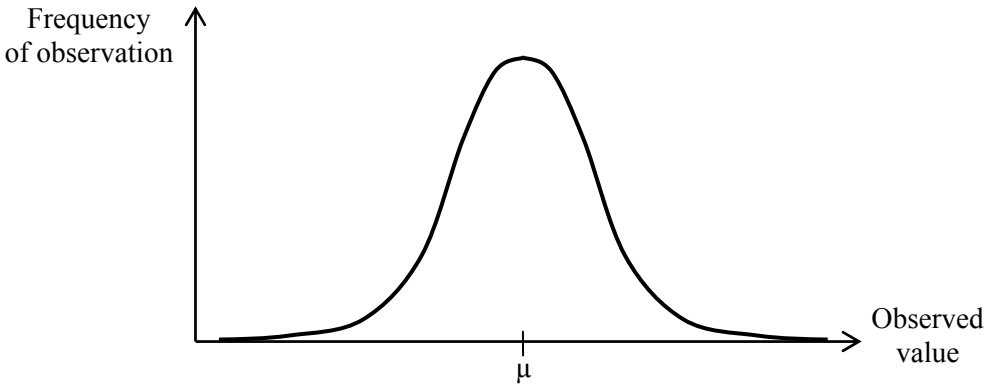
The square root of $\sigma^2_{\bar{Y}}$ is called the standard deviation of a mean, or more often the *standard error*.

Sample Standard Error: $s_{\bar{Y}} = \sqrt{s^2_{\bar{Y}}} = \frac{s}{\sqrt{n}}$

As with the standard deviation, this is a quantity in the original units of observation. As you will see, the standard error is extremely useful due to the role it plays in determining confidence intervals and the powers of tests.

The Normal Distribution

If you measure a quantitative trait on a population of meaningfully related individuals, what you often find is that most of the measurements will cluster near the population mean (μ). And as you consider values further and further from μ, individuals exhibiting those values become rarer. Graphically, such a situation can be visualized in terms of a frequency distribution, as shown below:



Some basic characteristics of this kind of distribution are:

1) The maximum value occurs at μ (i.e. the most probable value of an individual pulled randomly from the population is μ; another way of saying this is that the *expected* value of an individual pulled randomly from this population is μ);
2) The dispersion is symmetric about μ (i.e. the mean, median, and mode of the population are equal); and
3) The "tails" asymptotically approach zero.

A distribution which meets these basic criteria is known as a *normal distribution*.

The following conditions tend to result in a normal distribution of a quantitative trait:

1) There are many factors which contribute to the observed value of the trait;
2) These many factors act independently of one another; and
3) The individual effects of these factors are additive and of comparable magnitude.

As it turns out, a great many variables of interest are approximately normally distributed. Indeed, the normal distribution is observed for characters in complex systems of all kinds: Biological, socioeconomic, industrial, etc.

The bell-shaped normal distribution is also known as a Gaussian curve, named after Friedrich Gauss who figured out the formal mathematics underlying functions of this type. Specifically, a normal probability density function of mean μ and standard variance σ^2 is described by the expression:

$$Z(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{Y-\mu}{\sigma}\right]^2}$$

where Z(Y) is the height of the curve at a given observed value Y. Notice that the location and shape of a normal probability density function are uniquely determined by only two parameters, μ and σ^2. By

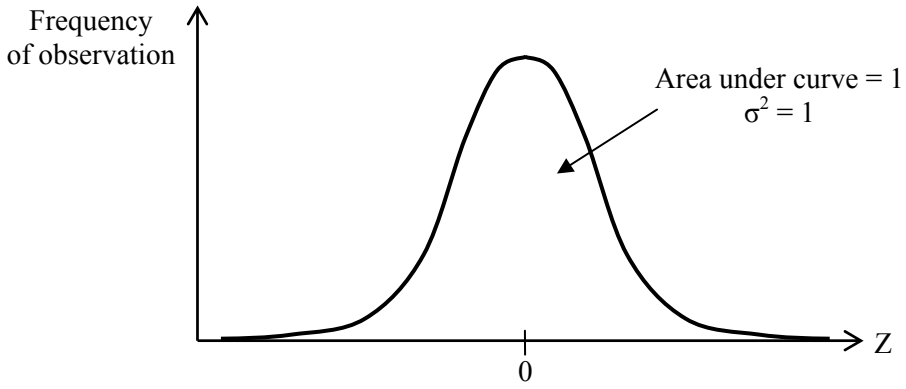
varying the value of μ, one can center Z(Y) anywhere on the x-axis. By varying σ^2, one can freely adjust the width of the central hump. All of the statistical techniques we will discuss in this class are based on the idea that many systems we study in the real world can be modeled by this theoretical function Z(Y). Such techniques fall into the broad category of *parametric statistics*, because the ultimate objectives of these techniques are to estimate and compare the theoretical parameters (in this case, μ and σ^2) which best explain our observations.

If we set μ = 0 and $\sigma^2 = 1$, we obtain an especially useful normal probability density function known as the standard normal curve [N(0,1)]:

$$Z(Y)_{[\mu=0,\sigma^2=1]} = \frac{1}{\sqrt{2\pi}} e^{-\frac{Y^2}{2}} \equiv N(0,1)$$

A word about notation: Rather than Y, it is traditional to use the letter Z to represent a random variable drawn from the standard normal distribution:

$$N(0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$



As with all probability density functions, the total area under the curve equals 1. On page 612 of your book (Appendix A4), you will find a table of the properties of this curve. For any given positive value of Z, the table reports the area under the curve to the right of Z. This is useful because the area to the right of Z **is** the theoretical probability of randomly picking an individual from N(0,1) whose value is greater than Z.

How does this help us in the real world? It helps us because ANY normal distribution can be standardized (i.e. any normal distribution can be converted into N(0,1)). The way this is done is quite simple:

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

Subtracting μ from each observation Y_i shifts the mean of the distribution to 0. Dividing by σ changes the scale of the x-axis from the original units of observation to *units of standard deviation* and thus makes the standard deviation (and the variance) of the distribution equal to 1. What this means is that for any unique individual Y_i from a normal distribution of mean μ and variance σ^2, there is a corresponding

unique value Z_i (i.e. a normal score) in the standard normal curve. And since we know the theoretical probability of picking an individual of a certain value at random from N(0,1), we now have a way of determining the probability of picking an individual of a certain value at random from *any* normally distributed population. Some examples:

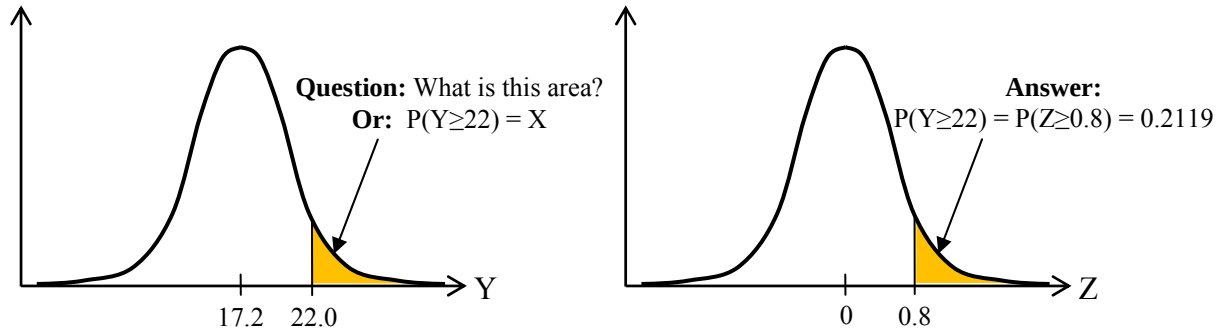
Question 1: From a normally distributed population of finches with mean weight (μ) = 17.2 g and variance (σ^2) = 36 g^2, what is the probability of randomly selecting an individual finch weighing more than 22 g?

Solution: To answer this, first convert the value 22 g to its corresponding normal score:

$$Z_i = \frac{Y_i - \mu}{\sigma} = \frac{22g - 17.2g}{6g} = 0.8$$

From Table A14, we see that 21.19% of the area under N(0,1) lies to the right of Z = 0.8. Therefore, there is a 21.19% chance of randomly selecting an individual finch weighing more than 22 g from this population. In other words, **22 g is not an unusual weight for a finch in this population**.

It helps to visualize these problems graphically, especially as they get more complicated:



Incidentally, we also see, simply by symmetry, that we have a 21.19% chance of randomly selecting an individual finch weighing less than 12.4 g. Do you see why?

Question 2: From a normally distributed population of finches with mean weight (μ) = 17.2 g and variance (σ^2) = 36 g^2, what is the probability of randomly selecting a *sample of 20 finches* with an average weight of more than 22 g?

Solution: The difference between this question and the previous one is that the previous question was asking the probability of selecting an *individual* of a certain value at random while this question is asking for the probability of selecting a *sample* of a certain average value at random. For individuals, the appropriate distribution to consider is the normal distribution of the population of *individuals* (μ = 17.2 g and σ^2 = 36 g^2). But for samples of size n = 20, the appropriate distribution to consider is the normal distribution of *sample means for sample size n = 20* (μ = 17.2 g and $\sigma^2_{\bar{Y}(n=20)} = \frac{\sigma^2}{n} = \frac{36g^2}{20} = 1.8g^2$).

With this in mind, we proceed as before:

$$Z_i = \frac{\bar{Y}_i - \mu}{\sigma_{\bar{Y}(n=20)}} = \frac{22g - 17.2g}{1.34g} = 3.6$$

From Table A14, we see that only 0.02% of the area under N(0,1) lies to the right of Z = 3.6. Therefore, there is a mere 0.02% chance of randomly selecting a sample of 20 finches with an average weight of more than 22 g from this population. In other words, **22 g is an extremely unusual mean weight for a sample of twenty finches in this population**.

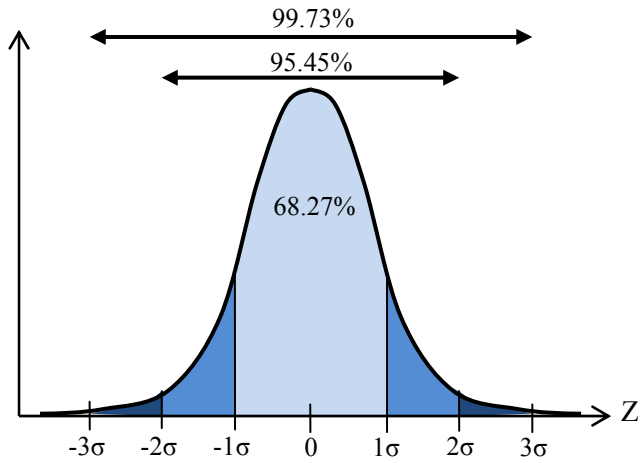
So, with a simple transformation of location and scale, *any* normal distribution, whether of individuals or of sample means, can be transformed into N(0,1), thereby allowing us to determine how unusual a given individual or sample is. Recall that the x-axis of the standard normal curve is in units of standard deviations. Our minds are not used to thinking in terms of units of dispersion, but it is an incredibly powerful way to think. To give you an intuitive feeling for such units, consider the following:

In a normal frequency distribution,

$\mu \pm 1\sigma$ contains 68.27% of the items
$\mu \pm 2\sigma$ contains 95.45% of the items
$\mu \pm 3\sigma$ contains 99.73% of the items

Thought of in another way,

50% of the items fall between $\mu \pm 0.674\sigma$
95% of the items fall between $\mu \pm 1.960\sigma$
99% of the items fall between $\mu \pm 2.576\sigma$



With these basic benchmarks in place, the results from the two examples above make a lot of sense. A 22 g finch is not unusual because it is less than one standard deviation from the mean. But a sample of 20 finches with a mean weight of 22 g is highly unusual because this sample mean is more than three standard errors from the mean.

One final word about the importance and wide applicability of the normal distribution: The **central limit theorem** states that, as sample size increases, the distribution of sample means drawn from a population of **<u>any</u>** distribution will approach a normal distribution with mean μ and variance σ^2/n.

Testing for Normality **[ST&D pp. 566-567]**

One is justified in using Table A14 (and, as you will see, t-tests and ANOVAs) if and only if the population or sample under consideration is "normal." Such statistical tables and techniques are said to "assume" normality of the data. Do not be misled by this use of the word. As a user of these tables and techniques, you do not simply "assume" that your data are normal; you test for it. Normality is spoken of as an "assumption," but in fact it is a criterion which must be met for the analysis to be valid. In this class, we will be using the Shapiro-Wilk test for assessing normality. See pages 566-567 in your text for

a good description of this technique. Below is Figure 24.2 from your book (page 566), with some supplemental annotation and discussion to help in understanding the test:

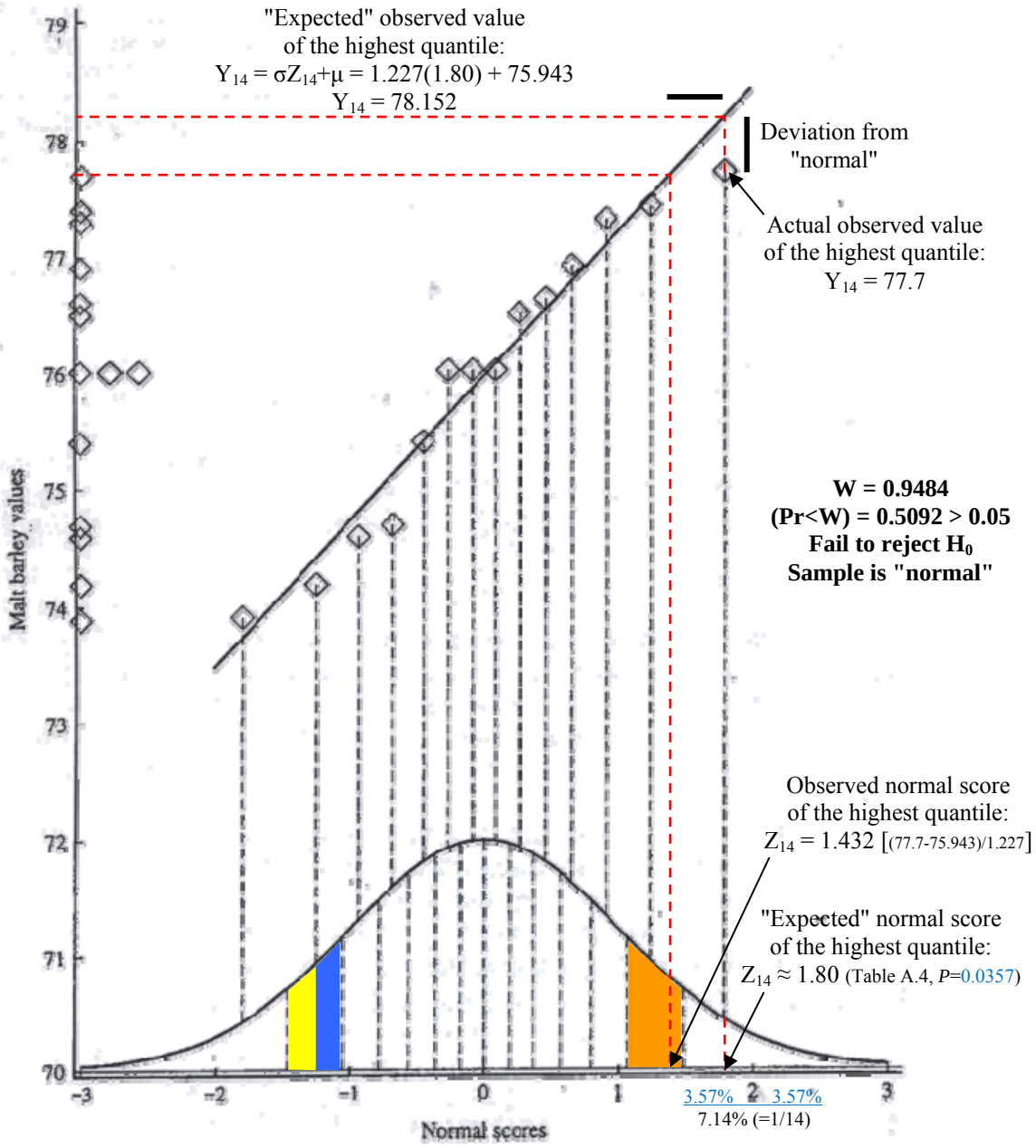


Figure 24.2 A normal probability plot (a.k.a. quantile-quantile or Q-Q plot). This is a graphic tool for visualizing deviation from normality. The Shapiro-Wilk test assigns a probability to such deviation, providing an "objective" determination of normality.

Since the dataset consists of 14 values (n = 14) [77.7, 76.0, 76.9, 74.6, 74.7, 76.5, 74.2, 75.4, 76.0, 76.0, 73.9, 77.4, 76.6, and 77.3], the area under N(0,1) is divided into 14 equal portions. This means that each portion, like the one indicated in orange above, has an area of 1/14 = 0.0714 square units. The normal score (Z) which splits a portion in half (by area) is considered the "expected" value for that portion. In the figure above, this means that the yellow area and the blue area are equal (1/28 = 0.0357 square units

each). The "expected" value of the second Z score is ≈ -1.24 (Table A4, Z value corresponding to a probability of 0.1071=0.0714+0.0357). The 14 "expected" normal scores are then transformed into the original units of observation ($Y_i = \sigma Z_i + \mu$), thereby generating a perfectly straight line with slope σ and intercept μ.

So, each data point in the sample has a corresponding "expected" normal score (e.g. while the actual normal score for 77.7 is $Z_i = \frac{77.7 - 75.943}{1.227} = 1.432$, its expected normal score is $Z_{14} \approx 1.80$), and a normal probability plot is essentially just a scatter-plot of these paired values. You can think of the Shapiro-Wilk test as essentially a test for correlation to the normal line. If the sample is perfectly normal, the scatter-plot of observed values vs. expected normal scores will fall exactly on the normal line and the Shapiro-Wilk test statistic W (similar to a correlation coefficient) will equal 1. Complete lack of correlation (i.e. a completely non-normal distribution) will yield a test statistic W equal to 0.

How much deviation is too much? For this class, to reject the null hypothesis of the test (H_0: The sample is from a normally distributed population), the probability of obtaining, *by chance*, a value of W *from a truly normal distribution* that is less than the observed value of W must be less than 5%.

In this case, W = 0.9484 and Pr<W = 0.5092 > 0.05; so we fail to reject H_0. There is no evidence, at this chosen level of significance, that the sample is from a non-normal population.