Chapter 4: An Introduction to Statistical Modeling

Day 1 :

The PICTURE

In Chapters 1 and 2, we learned some basic methods for analyzing the pattern of variation in a set of data. In Chapter 3, we learned that in order to take statistics to the next level, we have to design studies to answer specific questions (raised, perhaps, by the exploratory and analytical techniques studied in Chapters 1 and 2). We also learned something about designed studies.

We are now ready to move to the next level of statistical analysis: **statistical inference**. Statistical inference uses a sample from a population to draw conclusions about the entire population. The material of Chapter 3 enables us to obtain the sample in a statistically valid way. In order to obtain valid inference and to quantify the precision of the results obtained, we will need to study some of the models and probabilistic concepts found in Chapter 4.

Preview:

- Motivation: statistical inference
- Density histograms

- Population histograms and density curves
- Random variables
- Distribution models
- The binomial and normal distributions
- The power of models
- The Central Limit Theorem
- Identifying probability distributions
- Transformations to normality

Statistical Inference

Recall from Chapter 3 that the **Target Population** was defined as a collection of sampling units about which we want to draw conclusions. From now on, we'll drop the "target", and refer to the target population as the population.

Statistical Inference is the use of a subset of the population (the sample) to draw conclusions about the entire population. More specifically, we'll use measurements taken from the sampling units to draw our conclusions, so that when we speak of the population, we will mean the population of measurements.

A Word on Population Sizes

All known populations of actual sampling units are finite: they have a finite number of sampling units. However, some have a very large number of sampling units. It is mathematically easier and statistically effective to assume such populations are infinite.

In addition, some conceptual populations really are infinite: think of the population consisting of the numbers of dots that turn up in all possible tosses of a particular six-sided die (in this case the sampling unit is the toss of the die).

For these reasons, and because it is a little more difficult to deal with finite populations statistically, in this course we will consider all populations to be infinite.

Some Fine Print on Inference

The validity of inference is related to the way the data are obtained, and to the stationarity of the process producing the data.

For valid inference the units on which observations are made must be obtained using a **probability sample**. The simplest probability sample is a **simple random sample (SRS)**.

Density Histogram:

Histogram for data in which area, rather than height of bar, represents frequency.

- This allows proper representation of histograms with unequal interval widths.
- For a density histogram the bar height is the **density** of the bar: the relative frequency/(unit interval length).

- The total area of the bars equals 1.

Figure **??** shows a density (top) and frequency (bottom) histogram of the heights of 105 high school students. The only difference in the two histograms is the units on the vertical axis.

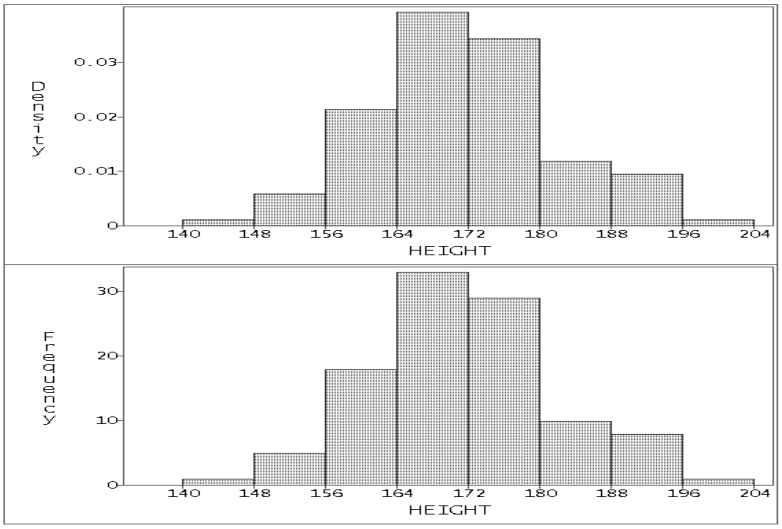

Figure 1: *A density (top) and frequency (bottom) histogram of the heights of 105 high school students.*

Population Histogram:

Frequency and density histograms are ways to represent the pattern of variation in data, which is usually a subset of measurements from a population. A **population histogram** is designed to represent the pattern of variation in a population. If there are a finite number of values in the population, the population histogram is a density histogram with one bar for each value in the population.

Recall, however, that we are assuming infinite populations, and defining population histograms for infinite populations requires a little more care.

Suppose that the set of measurement values is discrete (finite or countable: the set of all the integers is an example of the latter). Then, the population histogram is like a density histogram with one bar for each measurement value in the population. The area of the bar for a measurement value corresponds to the "proportion" of population measurements that take that value. Since the population is infinite, this "proportion" is defined as a limit.

Example 1

Consider the population consisting of all possible rolls of a single six-sided die. Then the population histogram will consist of six bars: one for each of the outcomes of 1, 2, 3, 4, 5, or 6 dots.

We may think of obtaining the population measurements by repeatedly rolling the die. Let $N_n(\{1\})$ denote the number of times a "1" occurs in the first n rolls. Then the proportion of the first n measurements that result in a "1" is $P_n(\{1\}) = N_n(\{1\})/n$. As we make more and more rolls, $P_n(\{1\})$ will approach a limiting value, $P(\{1\})$, which we define to be the "population proportion" of measurements that take the value 1. This is the area of the bar for measurement value 1. The areas of the other bars may be thought of in the same way.

Notes:

(a) The "population proportions" we have just defined are also known as **probabilities**. Thus, in Example 1, $P(\{1\})$ is the probability the die comes up 1. In this context, it is often useful to interpret the "population proportion" having a certain measurement value as the probability

that a randomly-selected measurement has that value.

(b) If the die is fair in Example 1, the area of each bar will equal 1/6. The corresponding population histogram is shown in Figure 4.2, p. 151 of the text.

What's the IDEA?

The **population (or probability) histogram** displays the pattern of variation of a population of discrete measurements. It displays one bar for each measurement value, and the area of the bar equals the population proportion (or probability) of that value.

Suppose we sample randomly and repeatedly from the population characterized by a population histogram, and after every new measurement is sampled, we form a density histogram of the measurements obtained so far, using the same set of bars as in the population histogram. Then the sequence of density histograms will converge to the population histogram. Figure 4.4, p. 154 of the text illustrates the convergence of density histograms for tosses of a fair die to the population histogram.

Density Curve

The pattern of variation of the population measurements is described differently if the set of measurement values is continuous (for our purposes, this means it consists of an interval). In this case, we model the population measurements by means of an idealized curve called a **density**

curve. The area under the density curve between two real numbers a and b is interpreted as the "population proportion" of measurements that take values between a and b. Here "population proportion" is taken in the same limiting sense discussed earlier. Figure 4.6, p. 176 gives a graphical representation of probability as area under a density curve.

Example 2

The **exponential density curve** is often used to model the population of lifetimes of electronic components. Its equation is

$$\begin{aligned} p(y) &= 0,\ y \leq 0, \\ &= e^{-y},\ y > 0 \end{aligned}$$

Suppose we feel the population of lifetimes (in years) of all transistors of a given type is given by the exponential density curve. The proportion of all transistors that last between 1 and 2 years is then the area under the curve between $y = 1$ and $y = 2$, which is given by the integral

$$\int_1^2 e^{-y} dy = -e^{-y} \,|_1^2 = e^{-1} - e^{-2} \approx 0.2325$$

Notes:

(a) As in the discrete case, we equate "population proportion" with probability. So in Example 2, we would say that the probability a transistor randomly chosen from the population lasts

between 1 and 2 years is 0.2325.

(b) The fact that for continuous measurements population proportion or probability is defined in terms of area under the density curve, implies that the population proportion of measurements taking any specific value is 0. For instance, according to the exponential distribution model in Example 2, the population proportion of lifetimes equal to 1 hour is

$$\int_1^1 e^{-y}dy = 0.$$

This is a reflection of the idealized world of the continuous distribution model in which no two measurements (here, transistor lifetimes) are exactly the same.

(c) However, the height of the density curve tells relatively how likely a measurement chosen randomly from the population is to occur very near any point. For example, if a and b are two real numbers, and if the density curve is twice as high at a as at b, then the population proportion of measurements very close to a is twice the population proportion of measurements very close to b.

To make this concrete, consider Example 2, and note that the population proportion having lifetimes within 0.01 hours of 1 hour is

$$\int_{0.99}^{1.01} e^{-y}dy \approx (0.02)(e^{-1}),$$

while the population proportion having lifetimes within 0.01 hours of 2 hours is

$$\int_{1.99}^{2.01} e^{-y}dy \approx (0.02)(e^{-2}).$$

Therefore, the population proportion of lifetimes very close to 1 is approximately

$$(0.02)(e^{-1})/(0.02)(e^{-2}) = e^{-1}/e^{-2} = p(1)/p(2)$$

times the population proportion of lifetimes very close to 2. Since $e^{-1}/e^{-2} = e \approx 2.71818$, we conclude that the population proportion of lifetimes very close to 1 is approximately 2.71818 times the population proportion of lifetimes very close to 2.

What's the IDEA?

The **density curve** displays the pattern of variation of a population of continuous measurements. The population proportion (or probability) of measurements taking values in any interval equals the area under the density curve over the interval. The relative values of the density curve at two different points can also be interpreted as the relative proportion of measurements (or relative probabilities of obtaining a measurement) very near the points.

Suppose we sample randomly and repeatedly from the population characterized by a density curve, and after every new measurement is sampled, we form a density histogram of the measurements obtained so far. If the number of bars in the sequence of density histograms are allowed to increase in a reasonable way, the density histograms will converge to the density curve of the population. Figure 4.7, p. 177 of the text illustrates this.

Day 2 :

Random Variables

We have been discussing measurements selected randomly from a population. The name given to such measurements is **random variable**. To emphasize the fact that they are random quantities, random variables are denoted by upper-case Roman letters, such as Y or Z. In statistical inference, the sample data obtained are assumed to be random variables independently chosen from the population.

Though randomness in random variables can result from such causes as measurement error, the source of their randomness that is of greatest interest (and probably of greatest import) is their random selection from the population.

It is important to distinguish between random variables and the values they take. We will often say something like, "Suppose $Y_1, Y_2, \ldots, Y_n$ is a **random sample**." By this we mean that the data consist of independent measurements (random variables) $Y_1, Y_2, \ldots, Y_n$ randomly selected from the population. The capital Ys tell us that these are random variables whose values will be unknown until the random selection is done. The values they take once the selection has been done will be denoted by small ys: $y_1, y_2, \ldots, y_n$.

Distribution Models

The pattern of variation displayed by the population histogram or density curve is called the **distribution model** (or just **distribution**) of the population measurement (random variable). In the discrete case, the distribution model consists of the probabilities the measurement takes on each possible value. In the continuous case, it consists of the density curve.

Just as the pattern of variation of a set of data is visually summarized by a frequency or density histogram, the pattern of variation of the population is visually summarized by the population histogram or density curve.

And just as the pattern of variation of a set of data is numerically summarized by measures of location such as the mean or median and measures of spread such as the standard deviation or IQR, we can define numerical measures of location and spread to summarize the pattern of variation of the population.

The most common measure of location for a population measurement is the **population mean**, which can be thought of as the place where the population histogram (or density curve) "balances." This interpretation is analogous to the interpretation of the mean of a data set.

The most common measure of spread for a population measurement is the **population standard deviation**, whose formula is similar to that of the standard deviation of a set of data. The **population variance** is equal to the square of the population standard deviation.

Population versus Sample Measures

Be careful not to confuse population measures, such as the mean, variance or standard deviation, with their sample counterparts. As the name implies, population measures summarize the entire population, while sample measures summarize a set of data taken from the population. To avoid confusion, we use Greek letters for population quantities. Thus, (1) The population mean is denoted μ, the Greek letter "mu", which is equivalent to our letter m, and (2) The population standard deviation is denoted σ, the Greek letter "sigma," which is equivalent to our letter s. The population variance is denoted σ^2. (Recall that the sample mean, standard deviation and variance are denoted $\overline{y}$, s and s^2, respectively.)

What's the IDEA?

Populations are large collections of measurements about which we are going to want to draw conclusions. Populations are modeled mathematically using **distribution models**. Graphically, these take the form of **population (or probability) histograms** or **density curves**.

A measurement taken randomly from the population is called a **random variable**. The pattern of variation of the random variable is summarized by the same distribution model that describes the population.

To make **inference** about the population, we will obtain a sample of measurements from the population. These measurements are characterized as n indpendent random variables. Once the measurements are obtained, the randomness of the random variables disappears, and the numbers

obtained become **data**.

The Binomial Distribution Model I: An Example

The binomial distribution model is used in drawing inferences about a population proportion: that is, the proportion of a population that has some characteristic.

Example 3

For instance, suppose the population of interest is a large (in fact, we're going to assume it is infinite) production lot of 12 ounce containers of frozen orange juice concentrate, and the characteristic of interest is the acceptability of the concentrate as defined by a set of quality measures. Since we cannot test all containers in the lot (there are too many, and besides, testing destroys the product), we will test a random sample of n containers and use the results to estimate the population proportion of all containers that are acceptable. Estimation of this sort is the type of statistical inference we're leading up to here. But before we can tell how to do estimation, we need to develop some theory on the binomial distribution model.

To make things simple, suppose we take a random sample of $n = 3$ containers, and suppose the true population proportion of acceptable containers is p. Then, in terms of the acceptability of the containers in the sample, there are 8 possible outcomes (A denotes acceptable, U denotes unacceptable):

	Container		
Outcome	1	2	3
1	A	A	A
2	A	A	U
3	A	U	A
4	U	A	A
5	A	U	U
6	U	A	U
7	U	U	A
8	U	U	U

Because we are assuming the population is infinite (see the notes below), the probability each sampled container is acceptable is p, and the probability it is unacceptable is therefore $1 - p$. In addition, the containers are sampled randomly from the population, and therefore whether any one is acceptable or not is independent of the quality of the other sampled containers. Because of independence, the probability of any of the eight outcomes in the table is computed as the product of the probabilities of the outcomes for each individual container. So, for example, the probability of outcome 3 is

$$P(\{\text{AUA}\}) = P(\{\text{A}\})P(\{\text{U}\})P(\{\text{A}\}) = p \times (1-p) \times p = p^2(1-p).$$

All eight probabilities may be computed this way. the results are listed in the following table:

	Container			
Outcome	1	2	3	Probability
1	A	A	A	p^3
2	A	A	U	$p^2(1-p)$
3	A	U	A	$p^2(1-p)$
4	U	A	A	$p^2(1-p)$
5	A	U	U	$p(1-p)^2$
6	U	A	U	$p(1-p)^2$
7	U	U	A	$p(1-p)^2$
8	U	U	U	$(1-p)^3$

It turns out that in order to make the best use of the sample data to estimate the population proportion p, we need not consider all eight outcomes listed. All we need consider is the number of

acceptable cans in the sample, which we will denote Y. Y is a measurement from the population (random variable) that can take the values 0, 1, 2 or 3. Here are the values of Y for the various juice container samples.

	Container				
Outcome	1	2	3	Probability	Y
1	A	A	A	p^3	3
2	A	A	U	$p^2(1-p)$	2
3	A	U	A	$p^2(1-p)$	2
4	U	A	A	$p^2(1-p)$	2
5	A	U	U	$p(1-p)^2$	1
6	U	A	U	$p(1-p)^2$	1
7	U	U	A	$p(1-p)^2$	1
8	U	U	U	$(1-p)^3$	0

The distribution model for Y is (here $p_Y(y)$ denotes the probability $Y = y$: i.e., there are exactly y acceptable cans in the sample):

$$p_Y(y) = \begin{cases} (1-p)^3, & y = 0 \\ 3p(1-p)^2, & y = 1 \\ 3p^2(1-p), & y = 2 \\ p^3, & y = 3 \end{cases}$$

Notice that $p_Y(1)$ consists of outcomes 5, 6, and 7 and $p_Y(2)$ consists of outcomes 2, 3, and 4, which explains the factor 3 in the distribution model.

Let's recapitulate:

The measurement Y was obtained by

1. Taking three independent measurements from the population,

2. Of which each measurement could take one of two values (A or U),

3. With the same probability of obtaining each value (p for A and $1-p$ for U) at each measurement.

Y was then the total occurrences of one of the two values (A) in the three measurements.

The distribution model derived here is common enough to have a name. We say that Y has a **binomial distribution with parameters 3 and** p, written $Y \sim b(3, p)$. The 3 refers to the number of original measurements in the sample, and the p refers to the probability of obtaining an A for any of the original measurements.

NOTE: In Example 3, we assumed the population (the lot of OJ containers) was infinite. In practice, of course, lots of OJ containers are not infinite. If the population is finite, the number of acceptable OJ containers in the sample no longer has a binomial distribution, and here is why.

Suppose the population has N OJ containers, of which k (equivalently, a proportion $p = k/N$) are acceptable. Then the first container sampled is acceptable with probability p. If the first container is acceptable, there are now $k-1$ acceptable among the remaining $N-1$, so the second container is acceptable with probability $(k-1)/(N-1)$. If the first container is not acceptable, the second is acceptable with probability $k/(N-1)$. Either way, the probability the second is acceptable changes, so the number of acceptable containers in the sample does not have a binomial

distribution.

However, if the population size, N, is large relative to the sample size, then the binomial will be a good approximation to the number of acceptable containers in the sample. It is in this sense that the binomial can be used in finite populations.

The Binomial Distribution Model II: The General Case

We get the general case of the binomial distribution model by considering n measurements (instead of 3) in the sample. Here's how it goes.

1. Take n independent measurements from the population,
2. Of which each measurement can take one of two values (instead of A or U, the common terminology is "success" and "failure"),
3. With the same probability of obtaining each value (p for "success" and $1-p$ for "failure") at each measurement.

If Y denotes the number of "successes" in the n measurements, Y has a binomial distribution with parameters n and p, written $Y \sim b(n,p)$. The distribution model is

$$p_Y(y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}, \ y = 0, 1, \ldots, n,$$

where the values

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n-y)!}$$

are called binomial coefficients.

NOTE: Here ! means factorial, so that for any integer k, $k!$ is the product $k(k-1)(k-2)\cdots(3)(2)(1)$ ($0!$ is defined to be 1). Thus, if, as in Example 3, $n = 3$,

$$p_Y(1) = \binom{3}{1} p(1-p)^2 = \frac{3!}{1!(3-1)!}p(1-p)^2 = \frac{3 \cdot 2 \cdot 1}{(1)(2 \cdot 1)}p(1-p)^2 = 3p(1-p)^2,$$

which is the value obtained in the example.

The mean of the $b(n,p)$ distribution model is np and the variance is $np(1-p)$.

Decision-Making Using the Binomial Distribution Model

One stage of a manufacturing process involves a manually-controlled grinding operation. Management suspects that the grinding machine operators tend to grind parts slightly larger rather than slightly smaller than the target diameter, while still staying within specification limits. To verify their suspicions, they sample 150 within-spec parts and find that 93 have diameters above the target diameter. Is this strong evidence in support of their suspicions?

To find out, we begin by supposing that there is no tendency to grind to larger or smaller diameters than the target diameter. Then the number of the 150 parts, Y, having diameters larger than the target diameter will have a $b(150, 0.5)$ distribution. In this case, the probability of finding 93 or more parts with diameters larger than the target diameter is

$$P(Y \geq 93) =$$

$$\frac{150!}{93!57!}(0.5)^{93}(0.5)^{57} + \frac{150!}{94!56!}(0.5)^{94}(0.5)^{56}$$

$$+ \ldots +$$

$$\frac{150!}{149!1!}(0.5)^{149}(0.5)^{1} + \frac{150!}{150!0!}(0.5)^{150}(0.5)^{0}$$

$$= 0.0021.$$

Thus, **if** there is no tendency to grind to larger or smaller diameters, they would observe as many as 93 of 150 sampled parts having diameters greater than the target in only 21 of 10000 samples.

What's the IDEA?

The binomial distribution model $b(n, p)$ models the number of successes in n independent trials where the probability of success equals p at each trial. The parameter p is often identified with the population proportion of successes.

Day 3 :

The Power of Models

- Quantifiers of data
- Extend range of conclusions

The Normal Distribution Model

The most important and famous of all distribution models is the normal distribution model. It's density curve is the famous "bell curve," given by the equation

$$p_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}, -\infty < y < \infty$$

As you might expect from the notation, the population mean is μ and the population variance is σ^2. A measurement Y from this population is said to have a normal distribution with parameters (or mean and variance) μ and σ^2, written $Y \sim N(\mu, \sigma^2)$.

Computing Normal Probabilities

All probabilities from any normal distribution can be reduced to probabilities from a standard normal (i.e. $N(0,1)$) distribution. Specifically, if $Y \sim N(\mu, \sigma^2)$, then

$$Z = \frac{Y - \mu}{\sigma} \sim N(0,1).$$

So,

$$P(a < Y < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right).$$

Example 4

The ability of a process to produce products that satisfy engineering specifications can be assessed by a **process capability study**. Here is a simple example:

Assembly line specifications at an auto manufacturing plant call for the assembly time of a wire harness module to take between 750 and 1000 seconds. Times recorded on the line for a large sample of these assemblies have a mean of 911 and a standard deviation of 42. Further, the assembly times in the sample have the bell-shaped histogram typical of data from the normal distribution.

If we assume the population of assembly times has a $N(911, 42^2)$ distribution, and if Y represents a random assembly time from the population, we can estimate the within-spec proportion of all assembly times as

$$\begin{aligned} P(750 < Y < 1000) &= \\ P\left(\frac{750-911}{42} < Z < \frac{1000-911}{42}\right) &= \\ P(-3.83 < Z < 2.12) &= 0.9829 \end{aligned}$$

Thus, we estimate the 98.29% of wire harness module assembly times are within spec.

The Central Limit Theorem

The Central Limit Theorem (CLT) is the most important theorem in statistics. In words, it says:

As long as the population standard deviation is finite, the distribution of the mean (or sum) of independently chosen data from that population gets closer and closer to a normal distribution as the sample size increases.

Mathematical Statement of the Central Limit Theorem

Suppose that $Y_1, Y_2, \ldots$ are independent random variables having a distribution with mean μ and variance $\sigma^2 < \infty$. Let

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

be the mean of the first n random variables. It can be shown (and is shown in the text) that the random variable $\overline{Y}$ has mean μ and variance σ^2/n (hence, standard deviation $\sigma/\sqrt{n}$).

Let Z_n be the standardized mean: $Z_n = \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}$. Then

$$\lim_{n\to\infty} P(a < Z_n < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz.$$

That is, as n gets larger, the distribution of Z_n gets closer and closer to a $N(0,1)$.

Example 4.5

It is known from past experience that the variance in the weight of cereal put in "28 ounce" boxes of a certain breakfast cereal product is close to 0.36 ounces, but that the mean weight is less predictable. Thus, the mean is estimated daily by randomly sampling cereal boxes from production and taking the mean of the weights of their contents. Production supervisors like to calculate the mean to within 0.1 ounces of its true value.

If 25 cereal boxes are used to compute the mean (call it $\overline{Y}_{25}$), and if we assume $\sigma = \sqrt{0.36} = 0.6$, then the probability the estimate is within 0.1 ounces of the true mean weight (call it μ) is

$$P(-0.1 < \overline{Y}_{25} - \mu < 0.1) =$$

$$P\left(\frac{-0.1}{\sigma/\sqrt{25}} < \frac{(\overline{Y}_{25} - \mu)}{\sigma/\sqrt{25}} < \frac{0.1}{\sigma/\sqrt{25}}\right) =$$

$$P\left(\frac{-0.1}{0.6/\sqrt{25}} < \frac{(\overline{Y}_{25} - \mu)}{0.6/\sqrt{25}} < \frac{0.1}{\sigma/\sqrt{25}}\right) \approx$$

$$P\left(\frac{-0.1}{0.12} < Z_{25} < \frac{0.1}{0.12}\right) =$$

$$P\left(-0.83 < Z_{25} < 0.83\right) = 0.5953$$

What do you think will happen if we increase the number of cereal boxes used in computing the mean to 100?

The Normal Approximation to the Binomial Distribution

First note that by multiplying both numerator and denominator by n, we can write

$$Z_n = \frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}}.$$

Next, note that if $W \sim b(n, p)$, we can write $W = \sum_{i=1}^{n} Y_i$, where $Y_1, Y_2, \ldots Y_n$ are independent $b(1, p)$ random variables. Also note that the mean and standard deviation of the Y_i are $\mu_Y = p$ and $\sigma_Y = \sqrt{p(1-p)}$, respectively. Then if n is large enough, the CLT says that

$$Z_n = \frac{\sum_{i=1}^{n} Y_i - n\mu_Y}{\sigma_Y \sqrt{n}}$$

$$= \frac{W - np}{\sqrt{np(1-p)}}$$

has approximately a $N(0,1)$ distribution.

So if n is large enough, a standardized binomial random variable (subtract its mean then divide by its standard deviation) has approximately a $N(0,1)$ distribution.

How large does n have to be? Detailed guidelines are given in the text on p. 196, but values of n satisfying $np \geq 10$ and $n(1-p) \geq 10$ will give good results for almost all applications.

A Better Normal Approximation to the Binomial Distribution

The continuity correction can make the CLT approximation to the binomial more accurate. The continuity correction consists of adding or subtracting 0.5 from the endpoints of the interval (k, m) when finding $P(k \leq Y \leq m)$:

$$P(k \leq Y \leq m)$$

$$= P(k - 0.5 \leq Y \leq m + 0.5)$$

$$= P\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{m + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$\approx P\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{m + 0.5 - np}{\sqrt{np(1-p)}}\right),$$

where $Z \sim N(0,1)$.

Example 5

Recall the following problem:

One stage of a manufacturing process involves a manually-controlled grinding operation. Management suspects that the grinding machine operators tend to grind parts slightly larger rather than slightly smaller than the target diameter, while still staying within specification limits. To verify their suspicions, they sample 150 within-spec parts and find that 93 have diameters above the target diameter. Is this strong evidence in support of their suspicions?

And its solution: Suppose that there is no tendency to grind to larger or smaller diameters than the target diameter. Then the number of the 150 parts, Y, having diameters larger than the target diameter will have a $b(150, 0.5)$ distribution. In this case, the probability of finding 93 or more parts with diameters larger than the target diameter is

$$P(Y \geq 93) =$$

$$\frac{150!}{93!57!}(0.5)^{93}(0.5)^{57} + \frac{150!}{94!56!}(0.5)^{94}(0.5)^{56}$$

$$+ \ldots +$$

$$\frac{150!}{149!1!}(0.5)^{149}(0.5)^{1} + \frac{150!}{150!0!}(0.5)^{150}(0.5)^{0}$$

$$= 0.0021.$$

Thus, **if** there is no tendency to grind to larger or smaller diameters, they would observe as many as 93 of 150 sampled parts having diameters greater than the target in only 21 of 10000 samples.

We will use the CLT with the continuity correction to approximate $P(Y \geq 93)$. First, the continuity correction:

$$P(Y \geq 93) = P(Y \geq 93 - 0.5) = P(Y \geq 92.5)$$

Now, by assumption, $p = 0.5$, so

$$P(Y \geq 92.5) = P\left(\frac{Y - (150)(0.5)}{\sqrt{(150)(0.5)(1-0.5)}} \geq \frac{93 - 0.5 - (150)(0.5)}{\sqrt{(150)(0.5)(1-0.5)}}\right) \approx$$

$$P(Z \geq 2.86) = 0.0021,$$

which equals the exact value to four decimal places. Note: if we don't use the continuity correction, the CLT approximation gives an approximate probability of 0.0012, not nearly as close.

What's the IDEA?

The Central Limit Theorem is a surprising result that says: regardless of the distribution of a population, as long as the population standard deviation is finite, the distribution of the mean (or sum) of independently chosen data from that population gets closer and closer to a normal distribution as the sample size increases.

Day 4 :

Assessing Normality

A quick and simple check: **68-95-99.7 rule**.

Identifying Common Distributions

A Q-Q plot is a plot to decide if it is reasonable to assume a set of data are drawn from a known distribution model (called a candidate distribution model). Details of how to construct a Q-Q plot are found in the text. Since the computer will construct the plot for you, these details are not important for you to know. What is important is the idea behind them, which we present now.

We suppose the data consist of n observations, and that these are, in increasing order, $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$. The k^{th} **quantile rank** of the candidate distribution model is the point $q_{(k)}$ below which lies the proportion k/n of the population values. If the data comes from the candidate distribution, then a plot of the pairs $(y_{(k)}, q_{(k)})$ on a scatterplot should roughly follow a straight line. This is the Q-Q plot. In question.

Transformations to Normality

- If the data are positive and skewed to the right, $\ln(Y)$ or $\sqrt{Y}$ should look more normal.
- If the data vary by more than 1 or 2 orders of magnitude, try analyzing $\ln(Y)$, for positive data, or $-1/Y$.

- If the data consist of counts, try analyzing $\sqrt{Y}$.

- If the data are proportions and the ratio of the largest to smallest proportion exceeds 2, try the logit transformation:

  $\ln(Y/(1-Y))$.