MODULE 13

PROBABILITY INTRODUCTION

PROBABILITY is the "language" used to describe the proportion of times that a random event will occur. The language of probability is at the center of statistical inference (see Modules 14 and 15). Only a minimal understanding of probability is required to understand most basic inferential methods, including all of those in this course. Thus, only a short, example-based, introduction to probability is provided here.[1]

13.1 Probability of Individuals

The most basic forms of probability assume that items are selected randomly. In other words, simple probability calculations require that each item, whether that item is an individual or an entire sample, has the same chance of being selected. Thus, in simple intuitive examples it will be stated that the individuals were "thoroughly mixed" and more realistic examples will require randomization.[2]

If every item has the same chance of being selected, then the probability of an event is equal to the proportion of items in the event out of the entire population. In other words, the probability is the number of items in the event divided by the total number of items in the population.

For example, the probability of selecting a red ball from a thoroughly mixed box containing 15 red and 10 blue balls is equal to $\frac{15}{25} = 0.6$ (i.e., 15 individuals ("balls") in the event ("red") divided by the total number of individuals ("all balls in the box"); Figure 13.1-Left). Similarly, the probability of randomly selecting a woman from a room with 20 women and 30 men is 0.4 ($= \frac{20}{50}$; Figure 13.1-Right). In both examples, the calculation can be considered a probability because (i) individuals were randomly selected and (ii) a proportion of a total was computed.

[1] A deeper understanding of probability is required to understand more complex inferential methods beyond those in this course.
[2] See Module 3 for methods to randomly select or allocate individuals.

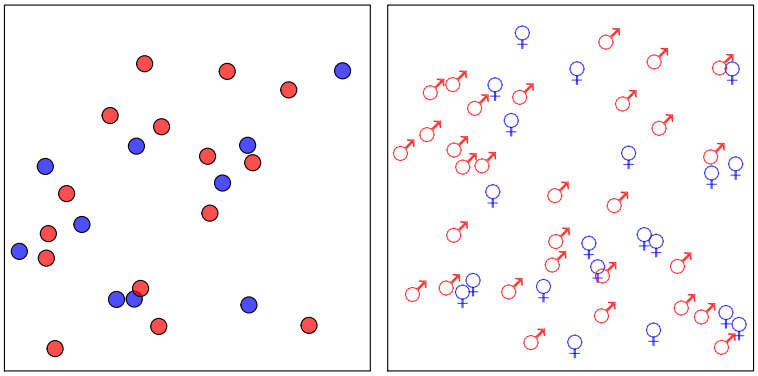

Figure 13.1. Depictions of a 'box' with 15 red balls and 10 blue balls (Left) and a 'room' with 30 men and 20 women (Right).

The two previous examples are simple because the selection is from a small, discrete set of items. Probabilities may be computed for a continuous variable if the distribution of that variable is known for the entire population. For example, the probability that a random individual is greater than 71 inches tall can be calculated if the distribution of heights for all individuals in the population is known (or reasonably approximated). For example, as shown in Module 8, if it can be assumed that heights is $N(66, 3)$, then the proportion of individuals in the population taller than 71 inches tall is 0.0478 (Figure 13.2).[3] This result is a probability because (i) the individual was randomly selected and (ii) the proportion of all individuals of interest in the entire population was found.

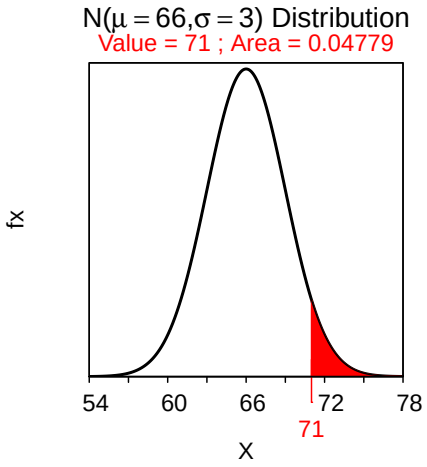

Figure 13.2. Calculation of the probability that a randomly selected individual from a $N(66, 3)$ population will have a height greater than 71 inches.

13.2 Probability of Statistics

The probability of a statistic computed from a random sample can also be found because the Central Limit Theorem (CLT) explains the distribution of statistics from all possible samples from a population Module

[3] As computed with `distrib(71,mean=66,sd=3,lower.tail=FALSE)`.

12. Probability calculations from sampling distributions will be the basis for making statistical inferences in Modules 14 and 15. These calculations are introduced here.

If the sample size is large enough, then the CLT states that the sampling distribution of sample means is approximately normal and the methods from Module 8 may then be used to compute probabilities. Therefore questions such as "what is the probability of observing a sample mean of less than 95 mm from a sample of $n = 50$ from Square Lake?" can be answered. This question is answered by first recalling that, for the length of all fish in Square Lake, μ =98.06 and σ =31.49. Because $n = 50$ is greater than 30, the CLT says that the distribution of the sample means from these samples is $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{50}})$ or $\bar{x} \sim N(98.06, 4.835)$. Thus, the proportion of samples of $n = 50$ from Square Lake with an $\bar{x} < 95$ mm is 0.2634, which comes from computing the area less than 95 on a $N(98.06, 4.835)$ distribution (Figure 13.3-Left).[4]

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(50)) )
[1] 0.2634127
```

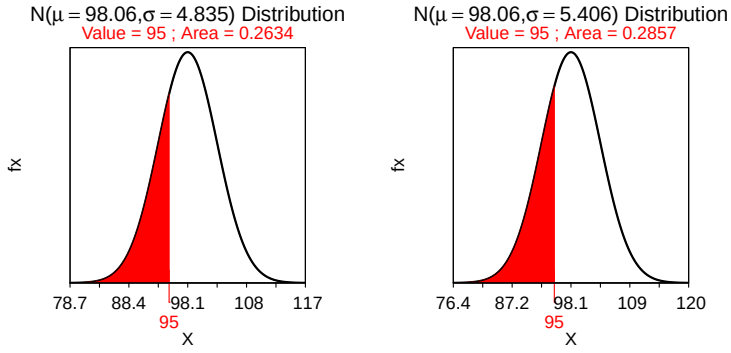

Figure 13.3. Proportion of sample means less than 95 mm on a $N(98.06, 4.84)$ (Left) and $N(98.06, 5.406)$ (Right) distribution.

Consider another question – "what is the probability of observing a sample mean of more than 95 mm in a sample of $n = 40$ from Square Lake?" At first glance it may appear that this question can be answered from the work done for the previous question. However, the sample sizes differ between the two questions and, because the sampling distribution depends on the sample size, a different sampling distribution is used here. Because $n > 30$ the sampling distribution will be $\bar{x} \sim N(98.06, \frac{34.19}{\sqrt{40}})$ or $\bar{x} \sim N(98.06, 5.406)$ (Note the different value of the SE). Thus, the answer to this question is the area to the right of 95 on a $N(98.06, 5.406)$ or 0.7143 (Figure 13.3-Right).

```
> ( distrib(95,mean=98.06,sd=34.19/sqrt(40),lower.tail=FALSE) )
[1] 0.714319
```

◇ **Always check what sample size is being used – if the sample size changes, then the sampling distribution changes.**

[4] Notice that the standard error of $\bar{x}$ is put into the `sd=` argument of `distrib()`. Recall that a standard error really is a standard deviation, it is just named differently (see Section 12.1.1). R has no way of knowing whether the question is about an individual or a statistic; it requires the dispersion in either case and calls both of them `sd=`.

Consider two more Square Lake example questions. First, "what is the probability of observing a sample mean of more than 95 mm in a sample of n =10 from Square Lake?" This question is again about a statistic, but because $n < 15$ and the population is not known to be normal it is not known that the sampling distribution will be normal. Thus, this question cannot be answered. Second, "What is the probability that a fish will have a length less than 85 mm?" This question is about an individual, not a statistic as in the previous questions. Thus, the population distribution, NOT the sampling distribution, is appropriate here. However, this question also cannot be answered because the population distribution is not known to be normally distributed.

Two points are illustrated with the last two questions. First, population distributions are used for questions about individuals and sampling distributions are used for questions about statistics. Second, if the distribution is not known to be normal, no matter which distribution is used, then the probability cannot be computed.[5]

One issue you may have noticed is that these calculations require knowing the mean, standard deviation, and shape (if $n < 30$) of the population. However, the population usually cannot be "seen" (recall Module 1) and, thus, it is uncomfortable to assume so much is known about the population. The only appropriate response to this concern is that we are building towards being able to make inferences with statements based on probabilities that take into account sampling variability. These questions, while not yet realistic, will help you to better understand sampling distibutions for when they are needed to make inferences in later modules.

13.3 A Process for Handling Probability Questions

As seen in the previous two sections, probability questions may use either the population distribution or the sampling distribution. To properly answer these questions it is important to determine (i) which of these two distributions to use, (ii) whether that distribution is normal or not, and (iii) the specific characteristics (i.e., mean and dispersion) of that distribution.

The type of distribution to use is dictated by whether the question is about an individual or about a statistic. Questions about individuals require using the population distribution, whereas questions about statistics require using the sampling distribution. Information about the population distribution, such as whether it is normally distributed or not and what the mean and standard deviation are, will be provided in the background information provided. In contrast, specifics about the sampling distribution must be identified from applying the rules of the Central Limit Theorem to information provided in the background. For both distributions, the probability question cannot be answered if the distribution is not normal. Both distributions are centered on μ, but the population distribution uses the standard DEVIATION as a measure of dispersion, whereas the sampling distribution uses the standard ERROR.

[5] At least with the techniques in this course.