

Stat 321 - Day 31
Sampling Distributions of Statistics

A *random sample* is a sequence of random variables X_1, X_2, \dots, X_n that are independent and identically distributed. A *statistic* is a function of the random variables in a random sample. Each statistic is itself a random variable and therefore has its own probability distribution that describes how it would vary under repeated random sampling. The probability distribution of a statistic is sometimes called a *sampling distribution*.

Example: Grade Point Average

Suppose that a student named Oliver has a .5 probability of getting an A, a .3 probability of getting a B, and a .2 probability of getting a C in a class. Suppose further that this probability distribution holds *independently* for each of two classes that he is taking this term. Let X_1 denote the number of grade points (A = 4 points, B = 3 points, C = 2 points) that he receives course 1 and similarly for X_2 .

(a) Calculate the expected value of Oliver's grade points in a single course.

Consider the random variables

- \bar{X} = mean grade points in the two courses
- R = range (absolute value of difference) of grade points in the two courses
- M = maximum number of grade points in the two courses

These random variables are called *statistics*, because they are functions of the values in a random sample. Each statistic is itself a random variable and therefore has its own probability distribution that describes how it would vary under repeated random sampling. The probability distribution of a statistic is sometimes called a *sampling distribution*.

(b) The following table lists all of Oliver's possible grades in these two courses and the probabilities of those outcomes. In each case, determine the value of these three random variables:

course 1	course 2	probability	sample mean	sample range	sample max
A	A	$(.5)(.5)=.25$			
A	B	$(.5)(.3)=.15$			
A	C	$(.5)(.2)=.10$			
B	A	$(.3)(.5)=.15$			
B	B	$(.3)(.3)=.09$			
B	C	$(.3)(.2)=.06$			
C	A	$(.2)(.5)=.10$			
C	B	$(.2)(.3)=.06$			
C	C	$(.2)(.2)=.04$			

- (c) Report the probability (sampling) distribution of the sample mean grade points by listing its possible values and the probability of each:

\bar{x}				
$p(\bar{x})$				

- (d) Determine the expected value of the sample mean grade points. How does it compare to the expected grade points in a single course?

- (e) Report the probability (sampling) distribution of the sample range R by listing its possible values and the probability of each:

r		
$p(r)$		

- (f) Determine the expected value of the sample range.

- (g) Report the probability (sampling) distribution of the sample maximum M by listing its possible values and the probability of each:

m		
$p(m)$		

- (h) Determine the expected value of the sample maximum.

Four-Course Sample:

- (i) Now suppose that Oliver is taking $n=4$ courses. If you were to enumerate all possible grades in these four courses (as in the table above b), how many outcomes would there be?

The exact analysis above is only feasible with such a small sample size (such as $n=2$). With larger sample sizes we can use simulation to investigate the long-run behavior of statistics such as the sample mean \bar{X} , sample range R, and sample maximum M.

- (j) Use Minitab to simulate 1000 repetitions of his grades in these 4 courses. First put the values 2, 3, and 4 into c1 and then their respective probabilities .2, .3, and .5 into c2. Then use:

```
MTB> random 1000 c3 c4 c5 c6;
SUBC> discrete c1 c2.
MTB> tally c3-c6
```

[Note: c3 contains the grade points in his first course, c4 for his second course, and so on. The 1000 rows constitute the 1000 different samples.]

- (k) Do the tallies reasonably approximate Oliver's grade point distribution? Explain.

- (l) Now use Minitab to calculate the mean of his grade points for these 1000 repetitions:

```
MTB> rmean c3-c6 c7
MTB> name c7 'average'
MTB> tally c7
```

Record the approximate probability distribution for \bar{X} in the table:

\bar{x}	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00
$p(\bar{x})$									

- (m) Produce a histogram and descriptive statistics (`hist c7, desc c7`) for these 1000 simulated GPA's. Comment on:

- the shape of the histogram
- the mean of these sample means (is it still similar to μ ?)
- Oliver's probability of attaining a perfect 4.00 GPA (is it more or less likely than when $n=2$?)
- Oliver's probability of achieving at least a 3.0 GPA (is it more or less likely than when $n=2$?)

Larger Sample Sizes:

Now you will simulate the (approximate) sampling distribution of the sample mean \bar{X} in a sample of $n=10$ courses (that Oliver might take in one year) and in a sample of $n=40$ courses (that Oliver might take in his college career).

(n) Simulate 1000 repetitions of Oliver's grade points in a sample of 10 courses:

```
MTB> random 1000 c11-c20;
```

```
SUBC> discrete c1 c2.
```

Convince yourself that the distribution in any course comes close to 50% A's, 30% B's, and 20% C's:

```
MTB> tally c11-c20
```

Calculate the sample mean grade points for these 1000 samples:

```
MTB> rmean c11-c20 c21
```

```
MTB> name c21 'sample means (n=10)'
```

Examine a histogram and descriptive statistics:

```
MTB> histogram c21
```

```
MTB> describe c21
```

Comment on the shape and record these values:

shape:

mean of sample means:

standard deviation of sample means:

(o) Is the mean of these sample means close to μ (the mean grade points in any one course), which you previously found to equal 3.3?

(p) The standard deviation of these sample means is about how many times smaller than σ (the standard deviation of grade points in any one course)?

(q) Repeat this analysis by simulating 1000 repetitions of a 40-course college career for Oliver. Comment particularly on how the shape, mean, and standard deviation compare between the $n=40$ and $n=10$ cases.

shape:

mean of sample means:

standard deviation of sample means:

comparison with $n=10$ case: