

STAT 325 – Handout 19
Sampling Distributions (2.5, 2.6)

- A **random sample** is a sequence of random variables X_1, X_2, \dots, X_n that are *independent* and *identically* distributed.
 - This property is often abbreviated as i.i.d.
 - The number n is called the **sample size**.
- 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, describing how it would vary under repeated random sampling.
 - The probability distribution of a statistic is called a **sampling distribution**.

Example 19-1: Grade Point Average

Suppose that a student named Marius has a .3 probability of getting an A, a .5 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, variance, and standard deviation of Marius’s grade points in a single course.

Consider the statistics:

- \bar{X} = average (mean) grade points in the two courses
- M = maximum number of grade points in the two courses

The following table lists all of Marius’s possible grades in these two courses.

b) Determine the probabilities of these 9 possible outcomes, and record them in the table. Also report the values of these two statistics for each possible outcome:

course 1	Course 2	probability	sample mean	sample max
A	A			
A	B			
A	C			
B	A			
B	B			
B	C			
C	A			
C	B			
C	C			

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) Determine the variance and SD of the sample mean grade points. How do they compare to their counterparts for grade points in a single course?

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

M				
$p(m)$				

g) Determine the expected value, variance, and SD of the sample maximum.

Now suppose that you want to investigate Marius's academic performance over a year in which he takes 10 courses.

h) If you were to list all possible outcomes (grade permutations) for those 10 courses, how many would there be?

It's no longer feasible to enumerate all possible outcomes, but we can rely on *simulation* to approximate the sampling distributions of these statistics. The following R code, also available from our course website, performs such a simulation:

```
# start with N = number of repetitions, n = number of courses
# also start with pa = Pr(A), pb = Pr(B), pc = Pr(C)
#
grpts = rep(NA, times = n)
GPA = rep(NA, times = N)
GPmax = rep(NA, times = N)
for (i in 1:N) {
  rand = runif(n,0,1)
  for (j in 1:n) {
    if (rand[j] < pa) {grpts[j] = 4}
    if ((rand[j] >= pa) & (rand[j] < pa+pb)) {grpts[j] = 3}
    if (rand[j] >= pa+pb) {grpts[j] = 2}
  }
  GPA[i] = mean(grpts)
  GPmax[i] = max(grpts)
}
hist(GPmax); table(GPmax)
mean(GPmax); sd(GPmax)
hist(GPA); table(GPA)
mean(GPA); sd(GPA)
```

i) Explain the difference between the (`i in 1:N`) and the (`j in 1:n`) loops.

j) Explain what the `GPmax` and `GPA` vectors do.

k) Run this code for 100,000 simulated years of 10 courses per year. Report the approximate sampling distribution, mean, and SD of the sample maximum.

l) What do you notice about the (approximate) sampling distribution of the sample mean GPA? Comment on its shape, mean, and SD. How do these compare to their counter-parts with a sample size of $n = 2$?

m) Use the simulation results to approximate the probability that Marius's GPA will be at least 3.0. Then do the same for a GPA of 3.25.

n) Comment on how these probabilities in the $n = 10$ case compare to the $n = 2$ case.

o) Increase the sample size (number of courses) to 40, representing an entire college career. Before you run the simulation, predict what you will see with regard to the distribution of sample maximum and sample mean.

p) Run a simulation with 100,000 simulated college careers. Comment on what the simulation reveals about the sampling distributions of the sample maximum and sample mean.

q) Again use the simulation results to approximate the probability that Marius's GPA will be at least 3.0. Then do the same for a GPA of 3.25.

r) Comment on how these probabilities in the $n = 40$ case compare to the $n = 10$ case.

Example 19-2: Fast-food service time

Suppose again that the service time for a randomly selected customer at a particular fast-food restaurant follows an exponential distribution with mean 1.25 minutes. Let the random variable T represent this service time, and let $\bar{T} = \sum_{i=1}^n T_i / n$ denote the average service time in a random sample of n customers.

a) Report the mean and standard deviation of T .

b) Use R to simulate the waiting times for $N = 100,000$ *samples*, using each of the following sample sizes for number of *customers*: $n = 1$, $n = 5$, $n = 25$, $n = 100$. For each sample size, comment on the shape of the sampling distribution of \bar{T} and report the mean and SD of the sample means.

c) Comment on how the sampling distribution of \bar{T} changes as the sample size increases.

Theoretical result:

Let X_1, X_2, \dots, X_n be i.i.d. from *any* probability distribution. Denote $E(X_i)$ by μ and $\text{Var}(X_i)$ by σ^2 . Let

$$\bar{X} = \sum_{i=1}^n X_i / n \text{ for some positive integer } n \text{ (sample size).}$$

a) Use properties of expectation to determine $E(\bar{X})$.

b) Use properties of variance to determine $\text{Var}(\bar{X})$ and $\text{SD}(\bar{X})$.

c) Now suppose that the X_i 's have a normal distribution. What do you know about the distribution of \bar{X} in this case? Explain.

Your simulations and theoretical derivations from last time lead to the following result, the most important in all of probability and statistics:

Central Limit Theorem (CLT):

- Let X_1, X_2, \dots, X_n be i.i.d. with $\mu = E(X_i)$ and $\sigma^2 = \text{Var}(X_i)$. Also let $\bar{X} = \sum_{i=1}^n X_i / n$ denote the sample mean. Then the sampling distribution of \bar{X} has:
 - $E(\bar{X}) = \mu$
 - Be careful in reading this statement, which speaks of 3 different means:
 - The sample mean, \bar{X}
 - The population mean, μ
 - The mean of the sample means, $E(\bar{X})$
 - $\text{Var}(\bar{X}) = \sigma^2/n$, so $\text{SD}(\bar{X}) = \sigma/\sqrt{n}$
 - Averages vary less than individual values.
 - SD decreases proportionally to the square root of sample size.
 - An approximately *normal* distribution for large values of n
 - Regardless of the distribution of the X_i 's
 - Exactly normal for any n if the X_i 's are normally distribution
 - Becomes closer and closer to normal as the sample size n increases
 - Also closer to normal for X_i 's that are closer to normal
 - Corollary: The distribution of the *sum* of independent random variables also approaches a normal distribution as the sample size increases, with $E(\text{Sum}) = n\mu$ and $\text{Var}(\text{sum}) = n\sigma^2$.

Example 19-3: Grade Point Averages (cont.)

Reconsider Marius, who has a .3 probability of earning an A, .5 probability of a B, and .2 probability of a C in any one course. Recall that the expected value for the number of grade points in any one course is 3.1, and the SD is 0.7. Assume that he takes 40 courses in his college career and that the courses are independent.

a) Specify the (approximate) distribution of Marius's GPA over these 40 courses. Also draw a well-labeled sketch of this sampling distribution.

b) Suppose that Marius must return his scholarship money if his overall GPA is less than 3.0. Determine the probability that he will have to return his scholarship money. Also indicate the area corresponding to this probability on your sketch in a).

c) Suppose that Marius will win an award if his overall GPA is 3.25 or higher. Determine the probability that he will win this award.

d) Recall that you approximate the probabilities in b) and c) with an R simulation of 100,000 repetitions of Marius's college career. Were the results similar to your answers to b) and c)?

Example 19-4: Manufacturing potato chips

Suppose that the weights of bags of potato chips coming off an assembly line are normally distributed with mean $\mu = 12$ ounces and standard deviation $\sigma = 0.4$ ounces.

- a) Determine the probability that one randomly selected bag weighs less than 11.9 ounces.
- b) If you take a random sample of 10 bags, would you expect the probability of their sample mean weight being less than 11.9 ounces to be greater or less than the probability found in (a)? Explain, without performing the calculation.
- c) Calculate the probability asked about in the previous question. [*Hint*: Draw and label a sketch of the sampling distribution and shade the region whose area corresponds to this probability.] Does this probability indicate that a sample mean as small as 11.9 ounces would be surprising if the population mean were really 12 ounces?
- d) Repeat this analysis, for a sample of 100 randomly selected bags.
- e) What is the smallest sample size for which the probability of the sample mean being less than 11.9 ounces is less than .01? [*Hints*: Find the first percentile of the standard normal distribution as the value z such that $P(Z \leq z) = .01$. Set this percentile equal to the z -score from standardizing 11.9 and solve for n .]

f) If you were told that a consumer group had weighed randomly selected bags and found a sample mean weight of 11.9 ounces, would you doubt the claim that the true mean weight of all of the potato chip bags is 12 ounces? On what unspecified information does your answer depend? Explain.

g) Which of your above answers to would be affected if the distribution of the weights of the bags was not normal but was rather skewed?

h) Find a value k such that the probability of the sample mean weight of 1000 randomly selected bags being between $12-k$ and $12+k$ is roughly 0.95. In other words, between what two \bar{x} values do the middle 95% of the \bar{x} values fall?

i) Determine the smallest sample size for which the probability is .95 that the sample mean falls within $\pm .05$ of 12 ounces (i.e., between 11.95 and 12.05).