Sampling

Marc Meredith*

*Introduction to Statistical Methods

Week 3

- During the first week of class we discussed the stages of measurement:
  1. Specifying the information to be collected
  2. Selecting the units from the population to collect this information from
  3. Recording the information available about the selected units
- Sampling, the term we use to describe step #2 in the process, is our focus this week

Agenda for week:

- What is sampling and how is it done in practice?
- What is the sampling distribution of a statistic?
- How are some known sampling distributions affected by the number of random variables sampled?
- How is knowledge of the sampling distribution of a statistic used to construct a confidence interval over its realization?
- What are asymptotics?
- What do we learn from the Law of Large Numbers (LLN) and Central Limit Theorem (CLT)?

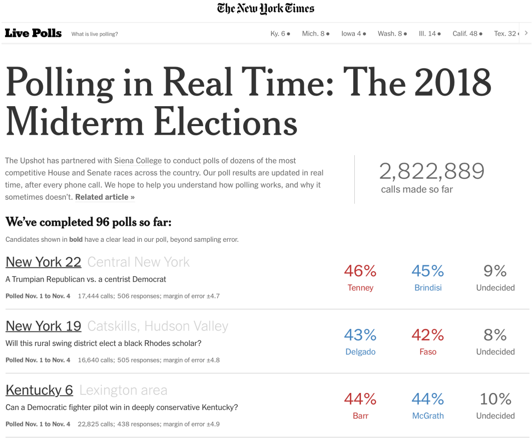
Key takeaways:

- Sampling allows us to make inferences about the characteristics of populations without having to conduct a census
- Our methods of sampling depend on characteristics of the population, our quantities of interest, and how we are going to record information about units selected into the sample
- Confidence intervals on a statistic generally are reduced when the statistic aggregates the information contained in more random variables, although there are diminishing returns to sample size

Key takeaways (continued):

- Knowing the sampling distribution of a statistic generally is essential to produce a meaningful confidence interval on its realization
- When the conditions of the LLN are met, the sample mean should converge the true underlying mean of the random variables used to construct it
- When the conditions of the CLT are met, the sampling distribution of a sample mean will be approximated well by the normal distribution

Our goal is to be able to use R to estimate and properly interpret this by the end of next week:



The New York Times

Live Polls What is live polling? Ky. 6 ● Mich. 8 ● Iowa 4 ● Wash. 8 ● Ill. 14 ● Calif. 48 ● Tex. 32 ●

Polling in Real Time: The 2018 Midterm Elections

The Upshot has partnered with Siena College to conduct polls of dozens of the most competitive House and Senate races across the country. Our poll results are updated in real time, after every phone call. We hope to help you understand how polling works, and why it sometimes doesn't. **Related article »**

2,822,889 calls made so far

We've completed 96 polls so far:

Candidates shown in **bold** have a clear lead in our poll, beyond sampling error.

New York 22 Central New York
A Trumpian Republican vs. a centrist Democrat
Polled Nov. 1 to Nov. 4 17,444 calls; 506 responses; margin of error ±4.7
46% Tenney 45% Brindisi 9% Undecided

New York 19 Catskills, Hudson Valley
Will this rural swing district elect a black Rhodes scholar?
Polled Nov. 1 to Nov. 4 16,640 calls; 505 responses; margin of error ±4.8
43% Delgado 42% Faso 8% Undecided

Kentucky 6 Lexington area
Can a Democratic fighter pilot win in deeply conservative Kentucky?
Polled Nov. 1 to Nov. 4 22,825 calls; 438 responses; margin of error ±4.9
44% Barr 44% McGrath 10% Undecided

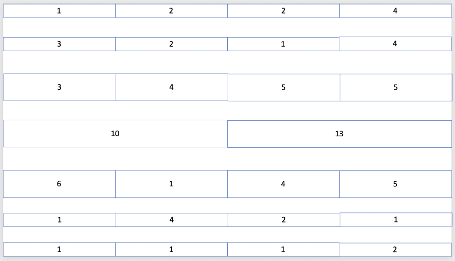
See: `https://nyti.ms/2oOvOL2`

- We often are interested in learning about a quantity of interest about a population
  - E.g., the percentage of Americans who approve of the president's job performance
- Sometimes with a goal of making comparisons about this quantity of interest between two subpopulations
  - E.g., the difference in the percentage of American men and women who approve of the president's job performance
- Sampling is the process through which we select the units from the population over which we (wish to) enumerate data informing us about this quantity

- Two broad approaches to sampling:
  1. Enumerate data from every unit in the population
     - E.g., a census
  2. Enumerate data from a subset of the units in the population
     - E.g., a sample

- Comparative advantages of a census:
  - Calculating many quantities of interest is extremely straightforward
  - Know the exact value of the quantity of interest, even for the smallest subpopulations within the population
- Comparative advantages of a sample:
  - Feasible and cost efficient, even when the population is large (even infinitely large)
  - Depending on the sampling frame (e.g., the method of selecting units into the sample), statistics may be used to draw probabilistic conclusions about the likelihood that the quantity of interest in the population is contained in some range

Sampling example: How many people live in this neighborhood?



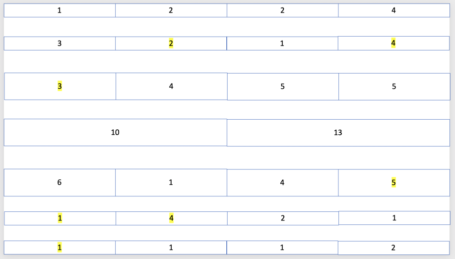
1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

Note: Each number represents the number of people living in that house

- If we conducted a census we would learn that 86 people live in this city
  - 23 people live in the two large houses
  - 33 people live in the eight medium houses
  - 32 people live in the sixteen small houses
- The question is how much of this information could be learned if we only enumerated this information for a subset of the houses

Sampling example: Using the data enumerated in a sample (i.e., houses highlighted in yellow) to learn information about the population requires knowledge of why those units were sampled

- A sampling frame describes the process through which units in the population are selected into the sample



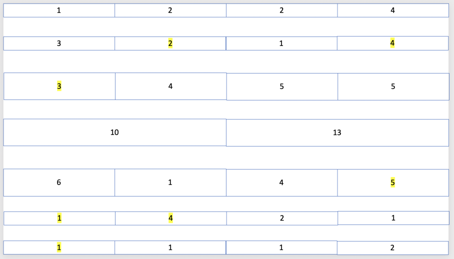
1	2	2	4
3	**2**	1	**4**
3	4	5	5
10		13	
6	1	4	**5**
1	**4**	2	1
1	1	1	2

Note: Each number represents the number of people living in that house

- Sampling frames that all data scientists should know about:
  - Simple random sampling
  - Stratified sampling
  - Quota sampling
  - Cluster sampling
  - Convenience sampling

Sampling example: An example of a simple random sample

- Every unit within the population is equally likely to be sampled
- E.g., select 7 units, with each unit having a probability of $\frac{7}{26}$ of being selected

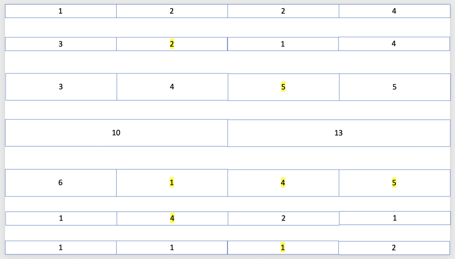


1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

Note: Each number that is shaded yellow represents a house that is selected into the sample

Sampling example: Another example of a simple random sample

- Every unit within the population is equally likely to be sampled
- E.g., select 7 units, with each unit having a probability of $\frac{7}{26}$ of being selected



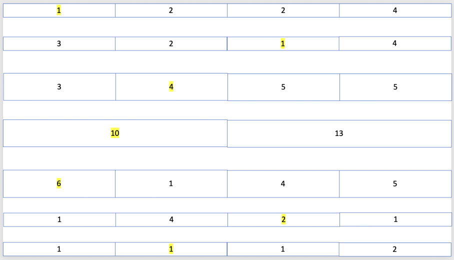
1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

Note: Each number that is shaded yellow represents a house that is selected into the sample

- Total number of people enumerated:
  - Sample 1 $=$ 20 people
  - Sample 2 $=$ 22 people
- One reason for the difference is the types of houses sampled in two samples:
  - Sample 1 $=$ 0 large, 2 medium, and 5 small houses
  - Sample 2 $=$ 0 large, 4 medium, and 3 small houses
- A stratified sample can be used to prevent this from happening
  - Each unit is assigned to a strata
  - Every unit within a strata is equally likely to be sampled, but units from different strata may be sampled at different rates

Sampling example: An example of a stratified random sample

- Every unit within a strata (here defined as a street) is equally likely to be sampled
- E.g., sample 1 unit from each street

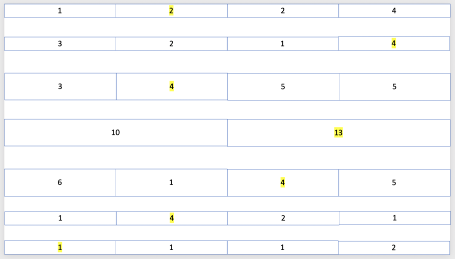


1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

Note: Each number that is shaded yellow represents a house that is selected into the sample

Sampling example: Another example of a stratified random sample

- Every unit within a strata (here defined as a street) is equally likely to be sampled
- E.g., sample 1 house from each street



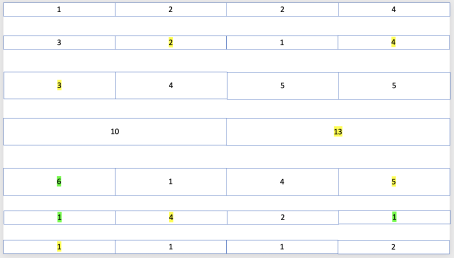
1	**2**	2	4
3	2	1	**4**
3	**4**	5	5
10		**13**	
6	1	**4**	5
1	**4**	2	1
1	1	1	2

Note: Each number that is shaded yellow represents a house that is selected into the sample

- Stratified sampling requires units to be classified into a strata before sampling begins
    - E.g. you know which street a house is located on before you start the process of enumerating its population
- When this isn't feasible, a quota sample may help ensure good representation of observations from different groups
    - Prior to sampling, define a goal of how many observations to collect from each group
        - Where a unit's group would define the strata if the characteristic was known prior to sampling
    - Stop selecting a unit into the sample once the goal for the unit's group has been reached

Sampling example: An example of a quota random sample

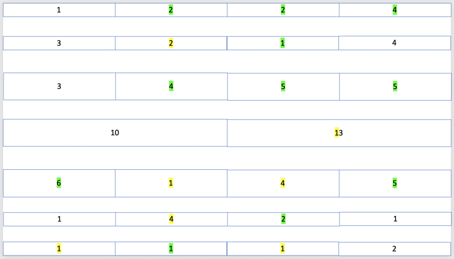
- Sample from a group until the group's quota has been met
- E.g., sample 1 large, 2 medium, and 4 small houses



Note: Each number that is shaded yellow represents a house that is selected before group quota is met, while each number that is shaded green represents a house that is selected after group quota is met

Sampling example: Another example of a quota random sample

- Sample from a group until the group's quota has been met
- E.g., sample 1 large, 2 medium, and 4 small houses

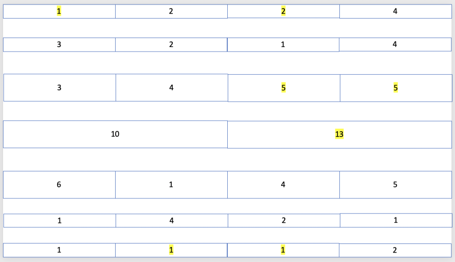


Note: Each number that is shaded yellow represents a house that is selected before group quota is met, while each number that is shaded green represents a house that is selected after group quota is met

- Sometimes it is less costly to select observations into a sample that are "close to each other"
  - Either in terms of distance or some other characteristic
  - E.g., one you sample one house on a street that might reduce the cost of sampling other houses on that street
- A cluster sample may be appropriate when this is the case
- In a cluster sample:
  - Every unit is assigned to a cluster
  - First clusters are selected that contain all of the units that could be sampled
  - Units are then selected from within these selected clusters

Sampling example: An example of a clustered random sample

- Every cluster is equally likely to be sampled
- Every unit from within a cluster is equally likely to be sampled
- E.g., sample $\frac{4}{7}$ of the streets and sample $\frac{1}{2}$ of the houses on those streets

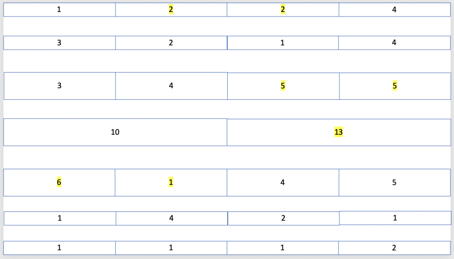


1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

Note: Each number that is shaded yellow represents a house that is selected into the sample

Sampling example: Another example of a clustered random sample

- Every cluster is equally likely to be sampled, a then units are equally likely to be sampled from sampled clusters
  - E.g., sample $\frac{4}{7}$ of the streets and sample $\frac{1}{2}$ of the houses on those streets



1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

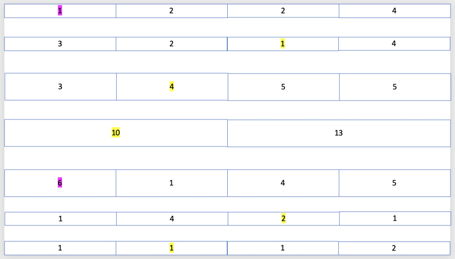
Note: Each number that is shaded yellow represents a house that is selected into the sample

- It is desirable that a sampling frame generates a sample that is likely to be:
    - Representative of the underlying population
    - Has a sufficient number of units from subpopulations of particular interest
    - Has a large number of observations
    - Is collected inexpensively
- Sampling frames tend to have different comparative advantages with respect to these properties

- Applying any of these sampling frames is unrealistic in many real world situations
  - While a sampling frame selects the units that would ideally be sampled, the data scientist often cannot enumerate the necessary information from all of the sampled units
- A convenience sample refers to a sampling frame in which the ease of enumerating a unit is, at least partially, responsible for determining which units get enumerated
  - Phrase also sometimes used to differentiate between a sampling frames in which the potentially sampled units are selecting into that status, as opposed to being selected by the data scientist

Sampling example: An example of a convenience sample

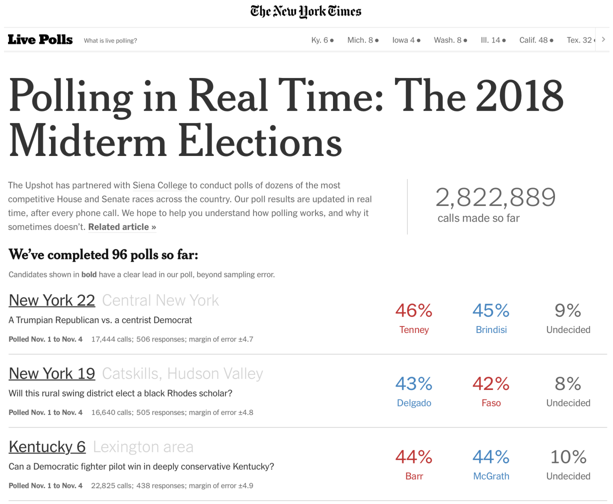
- Attempt to sample one unit from each strata (here defined as a street)
- Successfully contact 1 large, 1 medium, and 3 small houses



1	2	2	4
3	2	1	4
3	4	5	5
10		13	
6	1	4	5
1	4	2	1
1	1	1	2

Note: Each number that is shaded yellow represents a house that is successfully sampled, while each number that is shaded purple represents a house that is unsuccessfully sampled

- General strategy for analyzing data from a convenience sample:
  1. Identify the types of enumerate units that are overrepresented or underrepresented relative to their share of the population
  2. Put less weight on the types of enumerated units that are overrepresented, and more weight on the types of enumerated units that are underrepresented, relative to their share of the population
- In example on the previous slide, put weight of 2 on the large house, 8 on the medium house, and $\frac{16}{3}$ on the small houses
  - Because collected data on 1 of the 2 large houses, 1 of the 8 medium houses, and 3 of the 16 small houses
- Requires a strong assumption that the units of a given type that are enumerated are representative of the units of that same type that are not enumerated



The New York Times

Live Polls What is live polling? Ky. 6 ● Mich. 8 ● Iowa 4 ● Wash. 8 ● Ill. 14 ● Calif. 48 ● Tex. 32

Polling in Real Time: The 2018 Midterm Elections

The Upshot has partnered with Siena College to conduct polls of dozens of the most competitive House and Senate races across the country. Our poll results are updated in real time, after every phone call. We hope to help you understand how polling works, and why it sometimes doesn't. **Related article »**

2,822,889
calls made so far

We've completed 96 polls so far:

Candidates shown in **bold** have a clear lead in our poll, beyond sampling error.

New York 22 Central New York

A Trumpian Republican vs. a centrist Democrat

Polled Nov. 1 to Nov. 4 17,444 calls; 506 responses; margin of error ±4.7

46% Tenney 45% Brindisi 9% Undecided

New York 19 Catskills, Hudson Valley

Will this rural swing district elect a black Rhodes scholar?

Polled Nov. 1 to Nov. 4 16,640 calls; 505 responses; margin of error ±4.8

43% Delgado 42% Faso 8% Undecided

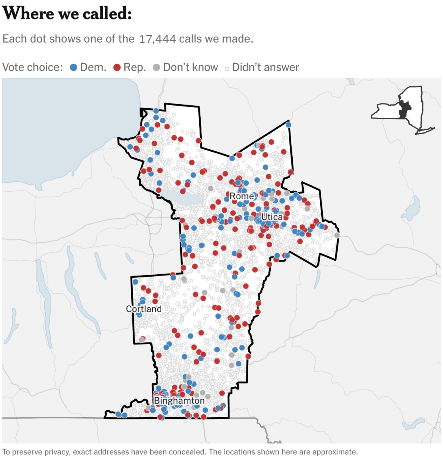
Kentucky 6 Lexington area

Can a Democratic fighter pilot win in deeply conservative Kentucky?

Polled Nov. 1 to Nov. 4 22,825 calls; 438 responses; margin of error ±4.9

44% Barr 44% McGrath 10% Undecided

See: `https://nyti.ms/2oOvOL2`



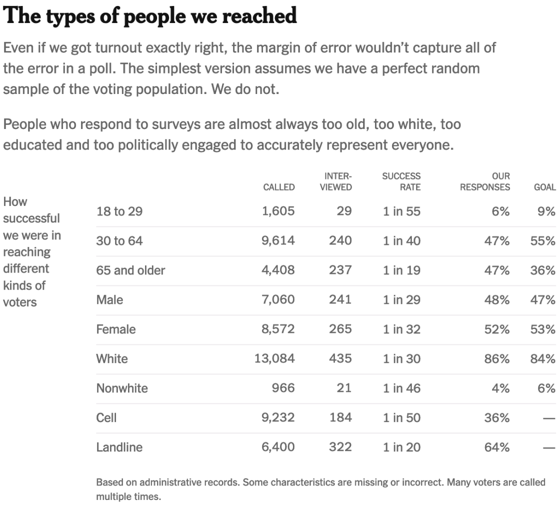
Where we called:

Each dot shows one of the 17,444 calls we made.

Vote choice: ● Dem. ● Rep. ● Don't know ○ Didn't answer

To preserve privacy, exact addresses have been concealed. The locations shown here are approximate.

See: `https://nyti.ms/2oOvOL2`



The types of people we reached

Even if we got turnout exactly right, the margin of error wouldn't capture all of the error in a poll. The simplest version assumes we have a perfect random sample of the voting population. We do not.

People who respond to surveys are almost always too old, too white, too educated and too politically engaged to accurately represent everyone.

		CALLED	INTER-VIEWED	SUCCESS RATE	OUR RESPONSES	GOAL
How successful we were in reaching different kinds of voters	18 to 29	1,605	29	1 in 55	6%	9%
	30 to 64	9,614	240	1 in 40	47%	55%
	65 and older	4,408	237	1 in 19	47%	36%
	Male	7,060	241	1 in 29	48%	47%
	Female	8,572	265	1 in 32	52%	53%
	White	13,084	435	1 in 30	86%	84%
	Nonwhite	966	21	1 in 46	4%	6%
	Cell	9,232	184	1 in 50	36%	—
	Landline	6,400	322	1 in 20	64%	—

Based on administrative records. Some characteristics are missing or incorrect. Many voters are called multiple times.

See: `https://nyti.ms/2oOvOL2`

- The remainder of this lecture will derive properties of what we should expect to occur in a sample generated using a simple random sample
  - Simplifies math, while illustrating general principles that apply to other forms of sampling
- Two general principles when another form of sampling is used:
  1. Put more weight on observations that had a lower probability of being selected into the sample than observations that had a higher probability of being selected into the sample
     - Enumerated units that were less likely to be sampled are more representative of the units that were not enumerated than those enumerated units that were more likely to be sampled
  2. Measures of uncertainty should be larger when data are collected using a clustered sampling frame than a simple random sample
     - Because of the homophily of "close observations"

- Last week, we established the concept of a r.v.
- R.v.s often are used to model the outcomes of interest that is being enumerated using a sampling frame
  - E.g., who the kth sampled person is going to vote for can be thought of as a Bernoulli r.v. Z_k that is equal to 1 if person k will support the Republican candidate and 0 if person k will support the Democratic candidate
- We frequently want to generate a statistic, Y_n, that if a function of two or more r.v.
  - I.e., $Y_n = T(X_1, X_2, ..., X_n)$
  - Examples:
    1. $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (i.e., the sample mean)
    2. $Y_n = min(X_1, X_2, ..., X_n)$ (i.e., the sample minimum)
- We refer to the pdf of a statistic as its sampling distribution
  - Because a statistic is comprised of r.v.s, it also is a r.v.

- The sample mean is one of the most commonly used statistics
  - We'll use the notation $\bar{Y}_n$ to represent $\frac{1}{n}\sum_{i=1}^{n} X_i$
- Two properties of $\bar{Y}_n$ when we assume that $X_1, X_2, ..., X_n$ are independent and identically distributed (iid) with $E[X_i] = \mu_x$ and $var(X_i) = \sigma_x^2$
  1. $E[\bar{Y}_n] = \mu_x$
  2. $var(\bar{Y}_n) = \frac{\sigma_x^2}{n}$

- $E[\bar{Y}_n] =$
  $E[\frac{1}{n}\sum_{i=1}^{n} X_i] =$
  $\frac{1}{n}E[\sum_{i=1}^{n} X_i] =$
  $\frac{1}{n}\sum_{i=1}^{n} E[X_i] =$
  $\frac{1}{n}\sum_{i=1}^{n} \mu_x =$
  $\frac{1}{n}\mu_x \sum_{i=1}^{n} 1 =$
  $\frac{1}{n}\mu_x n = \mu_x$

- $var(\bar{Y}_n) =$

  $var(\frac{1}{n}\sum_{i=1}^{n} X_i) =$

  $\frac{1}{n}^2 var(\sum_{i=1}^{n} X_i) =$

  - Because $var(aX) = a^2 var(X)$

  $\frac{1}{n}^2 \sum_{i=1}^{n} var(X_i) =$

  - Because independent

  $\frac{1}{n}^2 \sum_{i=1}^{n} \sigma_x^2 = =$

  $\frac{1}{n}^2 \sigma_x^2 \sum_{i=1}^{n} 1 = \frac{1}{n}^2 \sigma_x^2 n = \frac{\sigma_x^2}{n}$

- Implications of these formulas:
  1. The sample mean, on average, will equal the expected value of the underlying r.v.s used to construct it
  2. How much that we expect that this mean will deviate from this average is decreasing in the sample size
- The power of these formulas is that it is easy to know the expected value and variance of averages derived from any sample of iid r.v.s, as long as we know the mean and variance of the underlying r.v.s
  - Don't need to know the underlying pdf of the r.v.s
- But cannot directly assess anything about the likelihood that the realization of $\bar{Y}_n$ falls into a specific range without knowledge of its sampling distribution
  - For now, trust me that this one of the most important things we would like to assess about a statistic

Ways of learning about a sampling distribution:

1. Derive the sampling distribution based on the distribution of the underlying r.v.s combined to generate the statistic
2. Apply Chebyshev's Inequality to learn about the "worse case" scenario
3. Apply asymptotics
4. Simulation

- We are going to build up our intuition about sampling distributions by thinking about rolling dice
- Let Z_k be the r.v. that is equal to the number that comes up when we roll die k, so
  $p(Z_k = 1) = \frac{1}{6}$
  $p(Z_k = 2) = \frac{1}{6}$
  $p(Z_k = 3) = \frac{1}{6}$
  $p(Z_k = 4) = \frac{1}{6}$
  $p(Z_k = 5) = \frac{1}{6}$
  $p(Z_k = 6) = \frac{1}{6}$
- We are going to investigate the sampling distribution of $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$
  - I.e., the mean number rolled when we n dice
- Use to show how we can learn about sampling distribution based on the distribution of the underlying r.v.s combined to generate the statistic

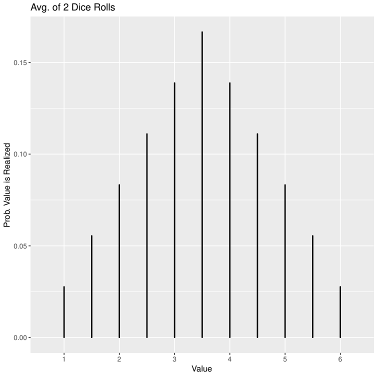
- Lets first think about $\bar{Y}_2 = \frac{Z_1+Z_2}{2}$
- We learned last week that $p(A \cap B) = p(A)p(B)$ when r.v.s A and B are independent
- So $p(Z_1 = c \cap Z_2 = d) = \frac{1}{36}$ for any $c, d \in \{1,2,3,4,5,6\}$
  - Because Z_1 and Z_2 are independent, this means that $p(Z_1 = c \cap Z_2 = d) = p(Z_1 = c)p(Z_2 = d) = \frac{1}{6}\frac{1}{6} = \frac{1}{36}$
- Define $k = c + d$
- We solve for $p(\bar{Y}_2 = \frac{k}{2})$ by multiplying the number of feasible combination of c and d that equal k by $\frac{1}{36}$

- We can solve for $p(\bar{Y}_2)$ by finding the number of squares that correspond to a given values of $\bar{Y}_2$ in this table and multiplying that number by $\frac{1}{36}$. For example, $p(Y_2 = 3) = \frac{5}{36}$:

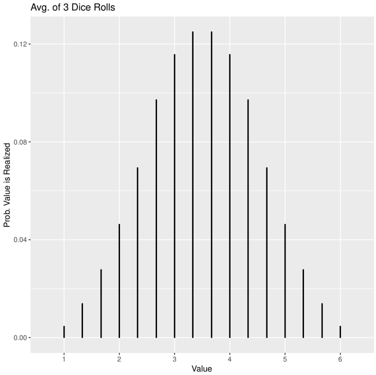
		$Z_2 =$					
		1	2	3	4	5	6
	1	$\bar{Y}_2 = 1$	$\bar{Y}_2 = \frac{3}{2}$	$\bar{Y}_2 = 2$	$\bar{Y}_2 = \frac{5}{2}$	$\underline{Y_2 = 3}$	$\bar{Y}_2 = \frac{7}{2}$
	2	$\bar{Y}_2 = \frac{3}{2}$	$\bar{Y}_2 = 2$	$\bar{Y}_2 = \frac{5}{2}$	$\underline{Y_2 = 3}$	$\bar{Y}_2 = \frac{7}{2}$	$\bar{Y}_2 = 4$
$Z_1 =$	3	$\bar{Y}_2 = 2$	$\bar{Y}_2 = \frac{5}{2}$	$\underline{Y_2 = 3}$	$\bar{Y}_2 = \frac{7}{2}$	$\bar{Y}_2 = 4$	$\bar{Y}_2 = \frac{9}{2}$
	4	$\bar{Y}_2 = \frac{5}{2}$	$\underline{Y_2 = 3}$	$\bar{Y}_2 = \frac{7}{2}$	$\bar{Y}_2 = 4$	$\bar{Y}_2 = \frac{9}{2}$	$\bar{Y}_2 = 5$
	5	$\underline{Y_2 = 3}$	$\bar{Y}_2 = \frac{7}{2}$	$\bar{Y}_2 = 4$	$\bar{Y}_2 = \frac{9}{2}$	$\bar{Y}_2 = 5$	$\bar{Y}_2 = \frac{11}{2}$
	6	$\bar{Y}_2 = \frac{7}{2}$	$\bar{Y}_2 = 4$	$\bar{Y}_2 = \frac{9}{2}$	$\bar{Y}_2 = 5$	$\bar{Y}_2 = \frac{11}{2}$	$\bar{Y}_2 = 6$

- Trying to make grids like this quickly becomes challenging as we roll more dice
- So this is a good example of a problem where R can be helpful to solve for the sampling distribution
- We'll next go through two different .R scripts
  - Rolling2Dice.R: Generates the sampling distribution when averaging 2 dice rolls
  - RollingNDice.R: Extends Rolling2Dice.R to generate the sampling distribution when averaging *n* dice rolls

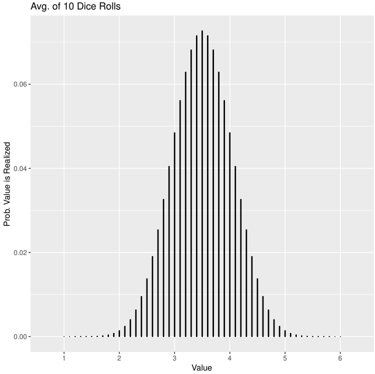
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300 dice rolls illustrates some general features of sampling distributions



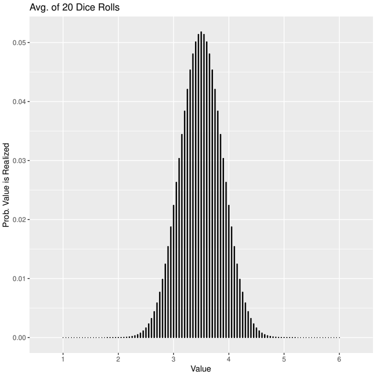
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



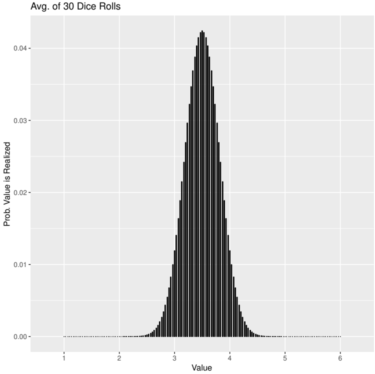
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



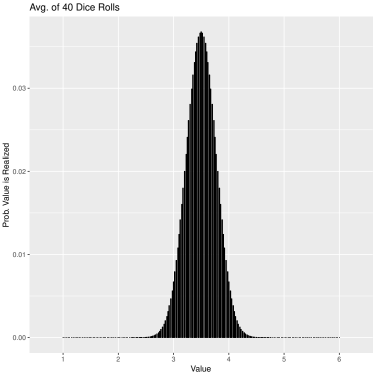
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



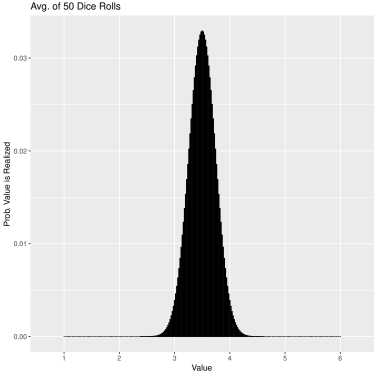
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



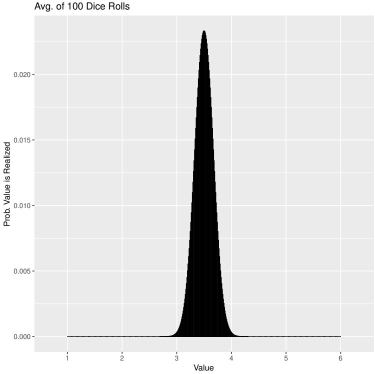
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



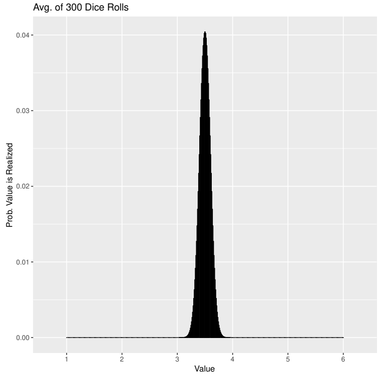
- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



- Looking at the sampling distributions associated with averaging of 2, 3, 10, 20, 30, 40, 50, 100, and 300, dice rolls illustrates some general features of sampling distributions



Takeaways from observing how this sampling distribution changes as n increases:

- The probability of the statistic being close to the mid-point of the sampling distribution increases as more observations are added into the sample
- The range of the sampling distribution decreases as more observations are added into the sample
- The impact of adding an additional observations on the sampling distribution appears to be more consequential when the baseline number of observations is small
- The sampling distribution moves from looking clumpy and non-bell shaped to something that looks smooth and bell shaped as the number of observations increases

- One of the main reason that we want to enumerate the sampling distribution of a statistic is so we can construct a CI on its realization
  - As well engage in hypothesis testing
- We can think about constructing CIs on a statistic in three ways:
  1. Fix the sample size, n, and the bounds, lb and ub, on the CI and figure out the probability, $1 - \alpha$, that the statistic, Y_n, is contained in that CI
  2. Fix the sample size, n, and the probability, $1 - \alpha$, that the statistic, Y_n, is contained in that CI and figure out what the bounds, lb and ub, on the CI
  3. Fix the bounds, lb and ub, and the probability, $1 - \alpha$, that the statistic, Y_n, is contained in that CI and figure out what sample size, n, should be

- To illustrate the first way of thinking about a CI, we can use the code at the bottom of the dice rolling simulation to report the probability that $p(3.25 \leqslant \bar{Y}_n \leqslant 3.75)$ based on the sampling distributions
    - $p(3.25 \leqslant \bar{Y}_{30} \leqslant 3.75) = 0.576$
        - I.e., there is a 57.6 percent chance that the average value on 30 dice will be between 3.25 and 3.75
    - $p(3.25 \leqslant \bar{Y}_{100} \leqslant 3.75) = 0.864$
        - I.e., there is a 86.4 percent chance that the average value on 100 dice will be between 3.25 and 3.75
    - $p(3.25 \leqslant \bar{Y}_{300} \leqslant 3.75) > 0.990$
        - I.e., there is a 99.0 percent chance that the average value on 300 dice will be between 3.25 and 3.75
- Easy to calculate any CI by changing the values on *lb* and *ub* within the R code

- As we discussed last week, the second way of thinking about a CI is more complicated because there often are distinct CIs that a statistic has the same probability of realizing a value within
  - E.g., $p(3.25 \leqslant \overline{Y}_{100} \leqslant 3.75) = p(3.31 \leqslant \overline{Y}_{100} \leqslant 3.9) = 0.864$
- The symmetric two-sided CI in which there is an equal probability of a statistic being above or below the CI is most commonly reported
- We also sometimes identify the one-sided CI in which focus entirely on cases in which the statistic is above or below the CI
  - We can use the code at the bottom of the dice rolling simulation to see that $p(3.22 < \overline{Y}_{100} < 3.78) = (\overline{Y}_{100} < 3.71) = p(\overline{Y}_{100} > 3.29) \approx 0.9$

- The third way of thinking about a CI is potentially useful when designing a sampling frame
- A common way to approach sampling is to think about how many units need to be sampled so that a statistic will be within an certain number of units of the truth with high probability
    - E.g., I want to survey enough people so that there is a 95 percent chance that the share of people who say that they will vote for a political candidate in my sample is within 3 percent of share of people who will vote for a political candidate in the broader population
- For example, we might want to roll enough dice that $p(3.25 < \bar{Y}_n < 3.75) = 0.95$
    - We can adjust the number of dice being averaged in the simulation until we find that $p(3.25 < \bar{Y}_{171} < 3.75) \approx 0.95$

- For reasons that will be apparent shortly, we frequently are interested in finding a CI on a statistic Y_n with a standard normal sampling distribution
  - I.e., $Y_n \sim N(0,1)$
- A reminder that last week we showed that when $Y_n \sim N(0,1)$, then $p(\Phi^{-1}(\frac{\alpha}{2}) < Y_n < \Phi^{-1}(1-\frac{\alpha}{2})) = 1-\alpha$
  - Where $\Phi^{-1}()$ is the inverse normal CDF
    - E.g, $\Phi^{-1}(1.96) = 0.975$ because there is a 97.5 percent chance that the realization of a standard normal random variable is 1.96 or less

- Asymptotics refers to the properties of a statistic as the number of r.v.s sampled gets very large
  - Technically as its get infinitely large
- We are going to talk about two important theorems that rely on asymptotics:
  1. The Law of Large Numbers
     - Defines conditions under which our sample mean asymptotically converges to the population mean
  2. The Central Limit Theorem
     - Defines conditions under which the distribution of the sample mean asymptotically converges to a normal r.v.
- Both are true for a wide range of underlying r.v.'s that were averaged to create a sample mean

The Law of Large Numbers (LLN):

- In mathematical notation:
  - Let $X_1, X_2, ..., X_n$ be iid r.v.'s with $E[X_i] = \mu_x$ and $var(X_i) = {\sigma_x}^2 < \infty$
  - Define: $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - $lim_{n \to \infty} p(\mid \bar{Y}_n - \mu_x \mid < \epsilon) = 1 \ \forall \epsilon$
- In English: The deviation between a population mean and the average of an infinite number of r.v.s with that population mean will almost certainly be less than any number that is even slightly larger than zero, as long as the variance of these r.v.s isn't too large

Why is the LLN important?

- As long as we sample enough observations, the mean in our sample will equal the population mean
    - Implies that we can learn the value of an (unknown) population mean from the sample mean in an extremely large sample
- We don't have to know much about the underlying distribution of the r.v.'s to apply it
- The Slutsky Theorem also tells us that for any continuous function g() if $Y_n \to \mu_x$ then $g(Y_n) \to g(\mu_x)$
    - So, for example, if $Y_n \to \mu_x$ then $2Y_n^2 + 6 \to 2{\mu_x}^2 + 6$

- We will use Chebyshev's Inequality to prove the LLN
- Chebyshev's Inequality states that for any r.v. X:
    - $p(\mid X - E[X] \mid \geqslant k\sigma_x) \leqslant \frac{1}{k^2}$
- In English: What is the highest probability that a the realization of a r.v. X could be more than k standard deviations away units away from its mean
- Intuition:
    - The variance is the expected squared difference between a random variable and its mean
    - Only so many observations can be far away from the mean without driving up the variance

- We can apply Chebyshev's Inequality to put a "worse case" CI on any sampling distribution
- Note that:
    - $p(\mid X - E[X] \mid \geqslant k\sigma_x) \leqslant \frac{1}{k^2} \implies$ $p(\mid X - E[X] \mid < k\sigma_x) > 1 - \frac{1}{k^2}$
- So no sampling distribution exists such that:
    - $p(\mid \bar{Y}_n - E[\bar{Y}_n] \mid < 2\sigma_{\bar{Y}_n}) > 1 - \frac{1}{2^2} = \frac{3}{4}$
    - $p(\mid \bar{Y}_n - E[\bar{Y}_n] \mid < 3\sigma_{\bar{Y}_n}) > 1 - \frac{1}{3^2} = \frac{8}{9}$
    - $p(\mid \bar{Y}_n - E[\bar{Y}_n] \mid < 4\sigma_{\bar{Y}_n}) > 1 - \frac{1}{4^2} = \frac{15}{16}$

- The previous slide shows that no matter what the sampling distribution of the statistic $\bar{Y}_n$ the probability that $\bar{Y}_n$ is within
  - Two standard deviations of its mean with a probability of at least $\frac{3}{4}$
  - Three standard deviations of its mean with a probability of at least $\frac{8}{9}$
  - Four standard deviations of its mean with a probability of at least $\frac{15}{16}$
- The problem with Chebyshev's Inequality is that these CIs often are extremely conservative
  - For example, we showed in a previous class that if r.v. is distributed normal then the probability that its realization is within two standard deviations of its mean is 0.955

Proof of the LLN:

- Chebyshev's Inequality states that $p(\mid \bar{Y}_n - \mu_x \mid \geqslant k\sigma_{\bar{Y}_n}) \leqslant \frac{1}{k^2}$
- Define $k = \frac{\sqrt{n}\epsilon}{\sigma_x} \implies p(\mid \bar{Y}_n - \mu_x \mid \geqslant \frac{\sqrt{n}\epsilon}{\sigma_x}\sigma_{\bar{Y}_n}) \leqslant \frac{{\sigma_x}^2}{n}$
  - Where ϵ is any positive number
- Earlier we showed that $var(\bar{Y}_n) = \frac{{\sigma_x}^2}{n} \implies \sigma_{\bar{Y}_n} = \frac{\sigma_x}{\sqrt{n}}$
- So $p(\mid \bar{Y}_n - \mu_x \mid \geqslant \frac{\sqrt{n}\epsilon}{\sigma_x}\frac{\sigma_x}{\sqrt{n}}) \leqslant \frac{1}{(\frac{\sqrt{n}\epsilon}{\sigma_x})^2} \implies$

  $p(\mid \bar{Y}_n - \mu_x \mid \geqslant \epsilon) \leqslant \frac{{\sigma_x}^2}{\epsilon^2 n}$
- $lim_{n\to\infty}\frac{\frac{{\sigma_x}^2}{\epsilon^2}}{n} = 0$
  - $\sigma_x, \epsilon > 0$ and $\sigma_x < \infty \implies 0 < \frac{{\sigma_x}^2}{\epsilon^2} < \infty$
- So $lim_{n\to\infty}p(\mid \bar{Y}_n - \mu_x \mid \geqslant \epsilon) \leqslant 0$
  - Which can be rewritten as $lim_{n\to\infty}p(\mid \bar{Y}_n - \mu_x \mid < \epsilon) \geqslant 1$

- What is important to take away from this proof?
  - We never truly have an infinite sample, so hard to know exactly when it applies
  - While the proof used the fact that the underlying r.v.'s were independent, it still applies as long as there isn't too much autocorrelation (non-independence)
    - On the homework, we'll use a simulation to show that a "random walk" is just too much autocorrelation
  - It requires that our underlying r.v.'s have a finite variance
  - We'll also need to apply the Central Limit Theorem to understand something about the sampling distribution

The Central Limit Theory (CLT):

- In mathematical notation:
  - Let $X_1, X_2, ..., X_n$ be iid r.v.'s with $E[X_i] = \mu_x$ and $var(X_i) = {\sigma_x}^2 < \infty$
  - Define: $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - As $n \to \infty$, the distribution of $\sqrt{n}(\bar{Y}_n - \mu_x) \to N(0, {\sigma_x}^2)$
- In English: There are many types of r.v.s that, when averaged together, produce a statistic that has a sampling distribution that can be well approximated by the normal distribution as the number of r.v.s that go into the average gets large

- Like the proof of the LLN, the CLT is proved for the case of an infinite sample
  - I am skipping this proof because it relies on mathematics that are beyond the scope of this class
- Because samples are never infinitely big in practice, the normal only approximates the true underlying sampling distribution
- How good this approximation is depends on the combination
  - The sample size
  - The variance of the r.v.s used to construct the mean
  - How much autocorrelation (non-independence) there is between the r.v.s used to construct the mean
- We can start to apply the logic of the CLT when the sample size is around 40 when the variance of the underlying r.v.s is small and the r.v.s are independent, but need bigger samples when the data are less nice

- Last week, we discussed that if $Q \sim N(\mu, \sigma^2)$ then $a + bQ \sim N(a + \mu, b^2\sigma^2)$
- So if the CLT implies that $\sqrt{n}(\bar{Y}_n - \mu_x) \to N(0, {\sigma_x}^2)$, it is also the case that:
  1. Let $a = 0$ and $b = \frac{1}{\sigma_x} \implies \sqrt{n}(\frac{\bar{Y}_n - \mu_x}{\sigma_x}) \sim N(0, 1)$
  2. Let $a = 0$ and $b = \frac{1}{\sqrt{n}} \implies \bar{Y}_n - \mu_x \sim N(0, \frac{{\sigma_x}^2}{n})$
  3. Let $a = \mu_x$ and $b = \frac{1}{\sqrt{n}} \implies \bar{Y}_n \sim N(\mu_x, \frac{{\sigma_x}^2}{n})$
  4. Let $a = n\mu_x$ and $b = \sqrt{n} \implies n\bar{Y}_n = \sum_{i=1}^{n} X_i \sim N(n\mu_x, n{\sigma_x}^2)$

- In lieu of a formal proof of the CLT, I will apply representation #3 of the CLT from the previous slide to approximate two exact sampling distributions that we have already derived
    1. The probability that between 40 to 50 people approve of a politician in a random sample of 100 people when 45 percent of the underlying population approves of the politician
        - Earlier we showed that $p(40 \leqslant \bar{Y}_{100} \leqslant 50) = 0.731$
    2. The probability that the average of 100 dice rolls is between 3.25 and 3.75
        - Earlier we showed that $p(3.25 \leqslant \bar{Y}_{100} \leqslant 3.75) = 0.864$
- The CLT says that we should get almost exactly the same result when we approximate the known exact sampling distributions with a normal distribution

- Let X_i be a r.v. that is equal to one if respondent i approves of the politician and zero if respondent i does not approve of the politician
  - X_i is Bernoulli so $E[X_i] = \pi = \frac{9}{20}$ and $var(X_i) = \pi(1-\pi) = \frac{9}{20}\frac{11}{20}$
- The CLT implies that $\bar{Y}_{100} = \frac{1}{100}\sum_{i=1}^{100} X_i \sim N(\frac{9}{20}, \frac{9*11}{20^2*100})$
- So $p(.395 \leqslant \bar{Y}_{100} \leqslant .505) =$

$$p(\sqrt{100}\frac{.395-.45}{\sqrt{.45*.55}} \leqslant \sqrt{100}\frac{\bar{Y}_{100}-.45}{\sqrt{.45*.55}} \leqslant \sqrt{100}\frac{.505-.45}{\sqrt{.45*.55}}) =$$

$$\Phi(\sqrt{100}\frac{.505-.45}{\sqrt{.45*.55}}) - \Phi(\sqrt{100}\frac{.505-.45}{\sqrt{.45*.55}}) =$$

- The CLT tells us that $Z = \sqrt{100}\frac{\bar{Y}_{100}-.45}{\sqrt{.45*.55}} \sim N(0,1)$

0.8655375 - 0.1344625 $=$ 0.731075

- Solved using "pnorm(sqrt(100)*(.505-.45)/sqrt(.45*.55)) - pnorm(sqrt(100)*(.395-.45)/sqrt(.45*.55)"

- Notice that I set the bounds on $\bar{Y}_{100}$ to be 0.395 and 0.505 to test whether between 40 and 50 people approve
  - Rather than 0.4 and 0.5
- This relates to a general issue with using a continuous distribution, like the normal, to approximate a process that only can take on a discrete number of outcomes
  - While a normal r.v. can take on a value of 40.43, there is no possibility that 40.43 is contained in the exact sampling distribution of $\bar{Y}_{100}$
- Thus, you need to think about what are the feasible outcomes of the exact sampling distribution and how you are going to map infeasible realizations of the normal r.v. into feasible outcomes
  - I.e., rounding infeasible outcomes to the closest feasible outcome so that a realization of 39.51, 39.99, and 40.43 all correspond with the outcome of 40 people approving of the politician

- I am now going to have you apply CLT to solve for the probability that the average of 100 dice rolls is between 3.25 and 3.75
- To do this, you'll first need to solve for $E[D_i]$ and $var(D_i)$
    - Let D_i be a r.v. that is equal to the value of die roll i
- Then you'll use these values in conjunction with the CLT to solve for $p(3.25 < \frac{1}{n}\sum_{i=1}^{100} D_i < 3.75)$

- What does the CLT imply about $\bar{Y}_{100} = \frac{1}{n}\sum_{i=1}^{100} D_i$ given that $E[D_i] = \frac{7}{2}$ and $var(D_i) = \frac{35}{12}$?
  - $\bar{Y}_{100} \sim N(\frac{7}{2}, \sqrt{\frac{35}{12*100}})$
- So $p(3.25 < \bar{Y}_{100} < 3.75) =$

  $$p(\frac{3.25 - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}} < \frac{\bar{Y}_{100} - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}} < \frac{3.75 - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) =$$

  $$\Phi(\frac{3.75 - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) - \Phi(\frac{3.25 - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) =$$

  - Because $\frac{\bar{Y}_{100} - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}} \sim N(0, 1)$

  $0.932298 - 0.06770196 = 0.8645961$

  - Solved using “pnorm((3.755 - 3.5)/sqrt(35/1200)) - pnorm((3.245 - 3.5)/sqrt(35/1200))”
  - Would get 0.8567651 if used “pnorm((3.75 - 3.5)/sqrt(35/1200)) - pnorm((3.25 - 3.5)/sqrt(35/1200))”

- I also am going to show you how to calculate two other CI using the CLT for the dice rolling example
    - What is a symmetric 90 percent CI on $\bar{Y_{100}}$
    - How large should n be so that $p(3.25 < \bar{Y_n} < 3.75) = 0.99$
- These will be useful examples for one of the problems on your homework

- $p(lb < \bar{Y}_{100} < ub) = .9 \implies$
  $p(\frac{lb-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}} < \frac{\bar{Y}_{100}-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}} < \frac{ub-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) = .9 \implies$
  $\Phi(\frac{ub-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) - \Phi(\frac{lb-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) = .9 \implies$
  - Because $\frac{\bar{Y}_{100}-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}} \sim N(0,1)$

  $\Phi(\frac{ub-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) = .95$
  - Because symmetric, $\Phi(\frac{lb-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}}) = 1 - \Phi(\frac{ub-\frac{7}{2}}{\sqrt{\frac{35}{12*100}}})$

$$\Phi\left(\frac{ub - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}}\right) = .95 \implies$$

$$\Phi^{-1}\left(\Phi\left(\frac{ub - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}}\right)\right) = \Phi^{-1}(.95) \implies$$

$$\frac{ub - \frac{7}{2}}{\sqrt{\frac{35}{12*100}}} = \Phi^{-1}(.95) \implies$$

$$ub = \frac{7}{2} + \Phi^{-1}(.95) * \sqrt{\frac{35}{12*100}} = 3.780912$$

- Solved using “7/2 + qnorm(.95)*sqrt(35/(12*100))”

- $p(3.25 < \bar{Y}_n < 3.75) = .99 \implies$

  $p(\frac{3.25-\frac{7}{2}}{\sqrt{\frac{35}{12n}}} < \frac{\bar{Y}_n-\frac{7}{2}}{\sqrt{\frac{35}{12n}}} < \frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}}) = 0.99 \implies$

  $\Phi(\frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}}) - \Phi(\frac{3.25-\frac{7}{2}}{\sqrt{\frac{35}{12n}}}) = 0.99 \implies$

  - Because $\frac{\bar{Y}_n-\frac{7}{2}}{\sqrt{\frac{35}{12n}}} \sim N(0,1)$

  $\Phi(\frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}}) = 0.995 \implies$

  - Because symmetric, $\Phi(\frac{3.25-\frac{7}{2}}{\sqrt{\frac{35}{12n}}}) = 1 - \Phi(\frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}})$

  $\Phi^{-1}(\Phi(\frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}})) = \Phi^{-1}(.995)$

$$\Phi^{-1}(\Phi(\frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}})) = \Phi^{-1}(.995) \implies$$

$$\frac{3.75-\frac{7}{2}}{\sqrt{\frac{35}{12n}}} = \Phi^{-1}(.995)$$

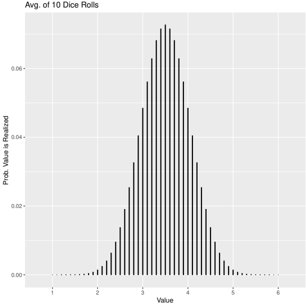
$$\sqrt{n} = \frac{\Phi^{-1}(.995)\sqrt{\frac{35}{12}}}{3.75-\frac{7}{2}} \implies$$

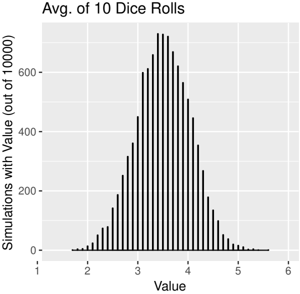
$$n = \frac{35}{12}(\frac{\Phi^{-1}(.995)}{3.75-\frac{7}{2}})^2 = 309.6285$$

- Solved using: “((qnorm(.995)*sqrt(35/12))/(3.75 - 3.5))*((qnorm(.995)*sqrt(35/12))/(3.75 - 3.5))”

- Last week we introduced the concept of a Monte Carlo simulation
- Monte Carlo simulations can be used to approximate the sampling distribution of a statistic
  - By repeatedly constructing the statistic in question in order to approximate the sampling distribution
  - Relates to asymptotics because the simulated sampling distribution will converge to sampling distribution as the statistic is simulated an infinite number of times
- Particularly useful when
  - Hard to use math to solve for the exact sampling distribution
  - Either hard or inappropriate to apply asymptotics to approximate the sampling distribution
  - Relatively easy to program a computer to repeatedly construct the statistic in question

- SimulateRollingNDice.R shows how we can simulate the distribution of averaging the roll of dice using R in order to approximate this distribution
  - Closely approximate the true sampling distribution when simulating 10,000 statistics
  - E.g., with the given seed, we find that $\bar{Y}_{10}$ is between 3.25 and 3.75 in 0.351 of simulations, while $p(3.25 \leqslant \bar{Y}_{10} \leqslant 3.75) = 0.352$





Key takeaways:

- Our methods of sampling depend on characteristics of the population, our quantities of interest, and how we are going to record information about units selected into the sample
- Confidence intervals on a statistic generally are reduced when the statistic aggregates the information contained in many random variables, although there are diminishing returns to sample size
- Knowing the sampling distribution of a statistic generally is essential to produce a meaningful confidence interval on its realization
- When the conditions of the LLN are met, the sample mean should converge the true underlying mean of the random variables used to construct it

Key takeaways (continued):

- When the conditions of the CLT are met, the sampling distribution of a sample mean will be approximated well by the normal distribution
- Haven't yet clearly established why sampling allows us to make inferences about the characteristics of populations without having to conduct a census
  - This week largely started from the premise that we knew characteristics of the population and figured out what a sample was likely to look like
  - But then unclear why we needed to collect a sample in the first place
- Next week will show how the skills that we are developing this week can be applied to learn about the characteristics of a population based on what is observed in a sample