Section 3.4 - Towards Statistical Inference

Statistics 104

Autumn 2004

Sampling Variability

In a survey or an experiment, we are trying to find out about features of a population (usually numerical) and they are unknown.

These unknown features are known as **Parameters**.

Example: Presidential candidate preferences

Possible parameters:

% of Massachusetts registered voters preferring

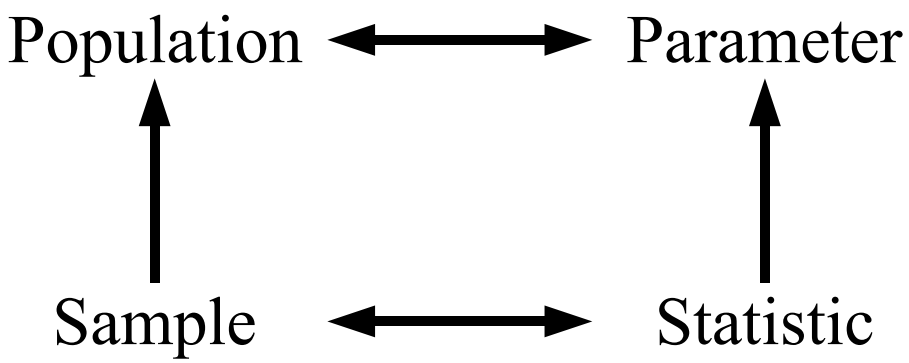
- Michael Badnarik (Libertarian)
- George Bush (Republican)
- David Cobb (Green-Rainbow)
- John Kerry (Democrat)
- Other (write-in candidates, e.g. Ralph Nader)
- Won't vote

Now perform the study. Use the data to estimate parameter values. These estimates are known as **Statistics**.

Example: Do survey in this class. Various statistics that can be calculated from this data are

% of Massachusetts registered voters preferring Bush, Kerry, etc.

Note that **Statistics** depend only on the data, not the parameter values.



Example: A sample of 1785 people were asked if they attended church or synagogue in the previous week. 1035 of the 1785 didn't (750 did).

- Population: All adults
- Sample: the 1785 questioned
- Parameter: Proportion of all all adults who attended church or synagogue in the previous week. (Call it p)
- Statistic: The proportion of the 1785 questioned who attended church of synagogue in the previous week.

$$\hat{p} = \frac{750}{1785} = 0.42$$

Key question:

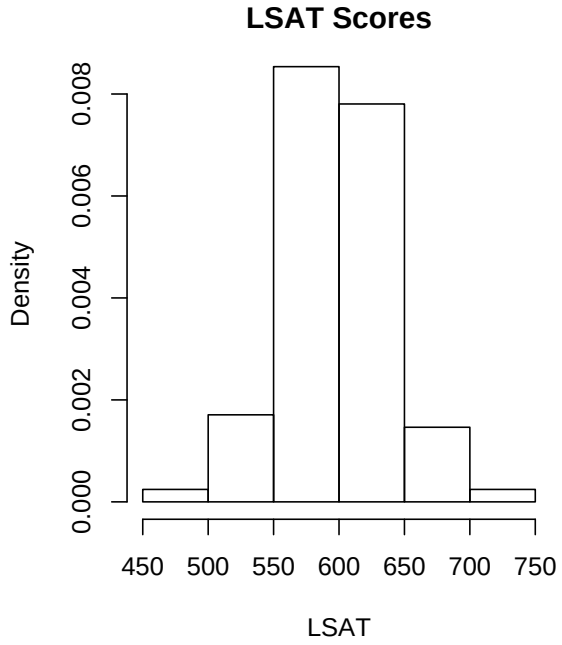
How accurate is the estimate as a guess of the parameter?

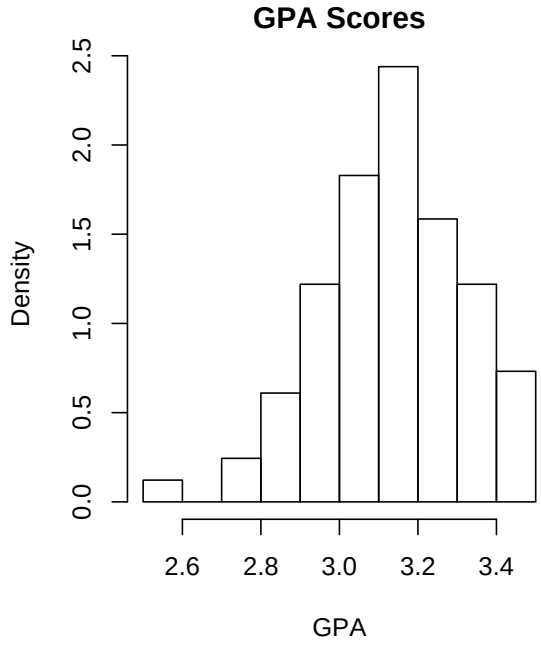
If you take a second sample, you will get a difference response.

Example: 1973 Law School Admissions

82 schools participated in large study of admissions practices

Two measures were reported for each school, average LSAT score (LSAT), and grade point average (GPA) of all members of the incoming class.





	μ	σ^2	σ
$LSAT$	597.55	1463.27	38.25
GPA	3.135	0.03547	0.1883

Lets see if we take SRS with $n = 15$ and $n = 30$.

Sample statistics for 10 different samples with $n = 15$

Sample	$\bar{x}$	s	Sample	$\bar{x}$	s
1	599.0667	36.62643	6	594.6667	41.77092
2	590.2667	48.36538	7	601.4667	30.31470
3	593.6667	51.88816	8	598.6000	49.77636
4	602.0667	40.68602	9	613.5333	33.24985
5	573.0667	50.32816	10	600.6667	39.88137

Sample statistics for 10 different samples with $n = 30$

Sample	$\bar{x}$	s	Sample	$\bar{x}$	s
1	589.7667	34.55199	6	594.5333	37.68539
2	596.7667	43.94643	7	605.8000	32.39987
3	604.0667	36.97617	8	593.0000	36.25865
4	605.6667	29.34966	9	605.5667	39.70207
5	590.2333	43.91346	10	603.0667	46.62612

We would like to get an understanding of what all the different possible SRS can give. We can think about what the statistic calculated from each possible sample is.

Sampling Distribution:

> The distribution of values of a statistic in all possible samples of the same size from the population.

The number of SRS of size n from a population of size N is

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

So for the law school data set, if $n = 15$,

$$\binom{82}{15} \approx 10^{16}$$

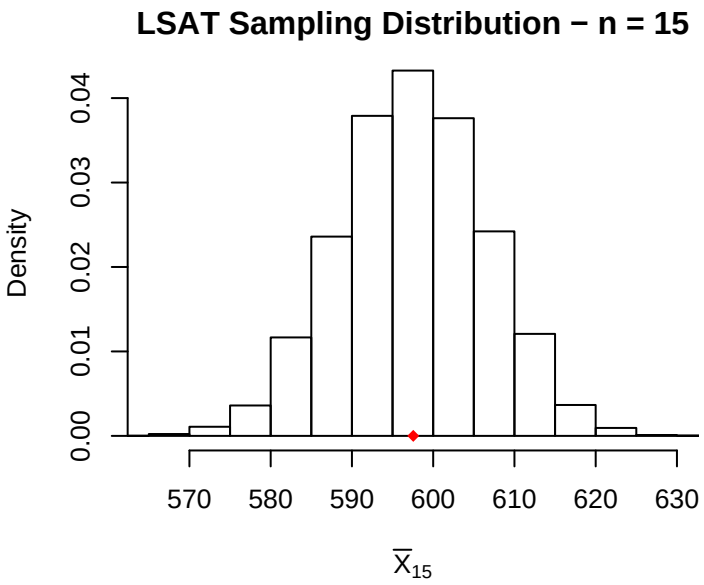
while if $n = 30$,

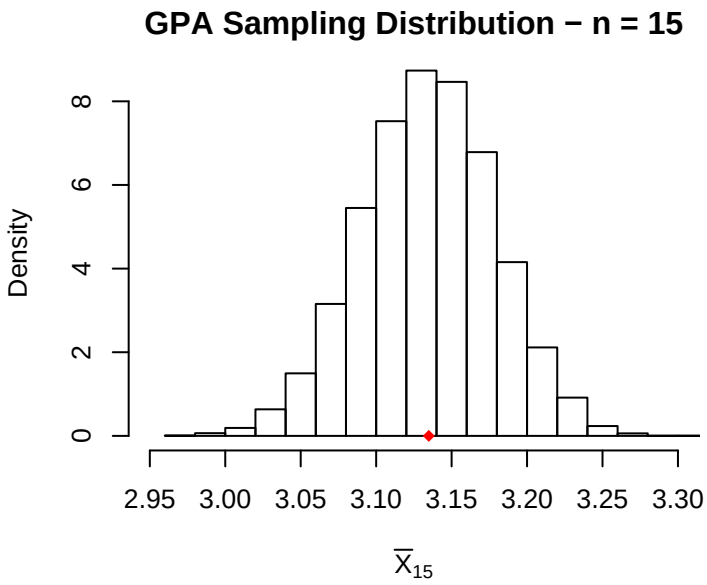
$$\binom{82}{30} \approx 2 \times 10^{22}$$

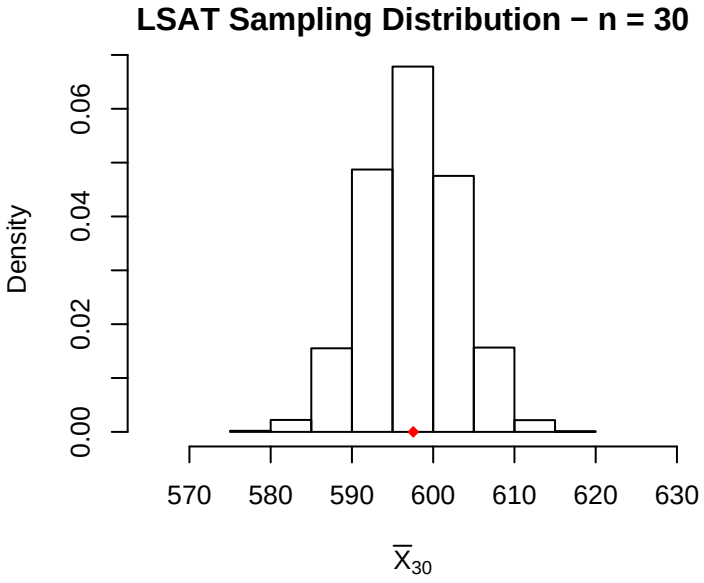
Now we can't actually determine the sampling distribution exactly since there are so many different possible samples. Also the only way to do it would be if we knew the whole population.

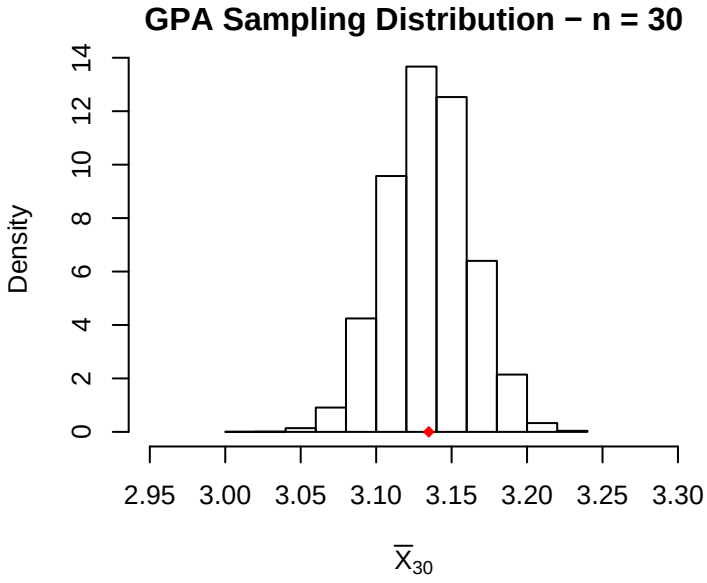
Even though we can't determine the distribution exactly, we will be able to make some statements about the sampling distribution.

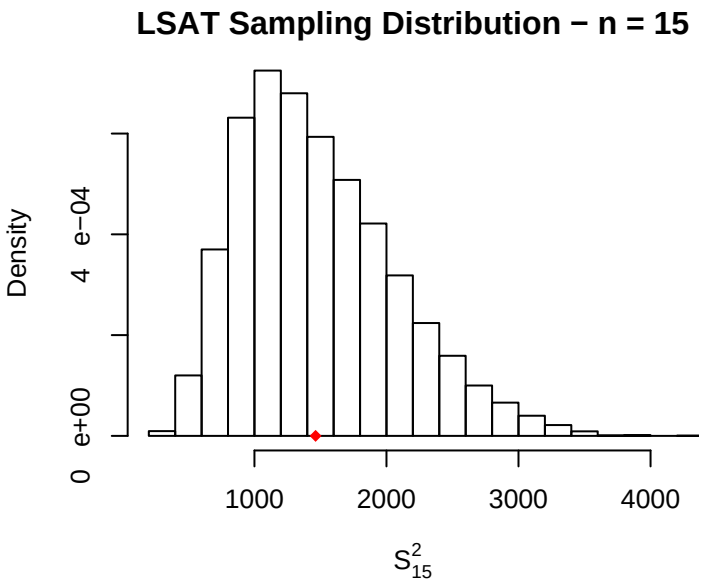
The following plots show Monte Carlo estimates of the sampling distribution of $\bar{X}, S^2$, and S for $n = 15, 30$. Monte Carlo was used here to describe the sampling distributions as there are too many possible samples to enumerate them all. Each histogram is based on 10,000 samples. The red diamond shows the true parameter value.

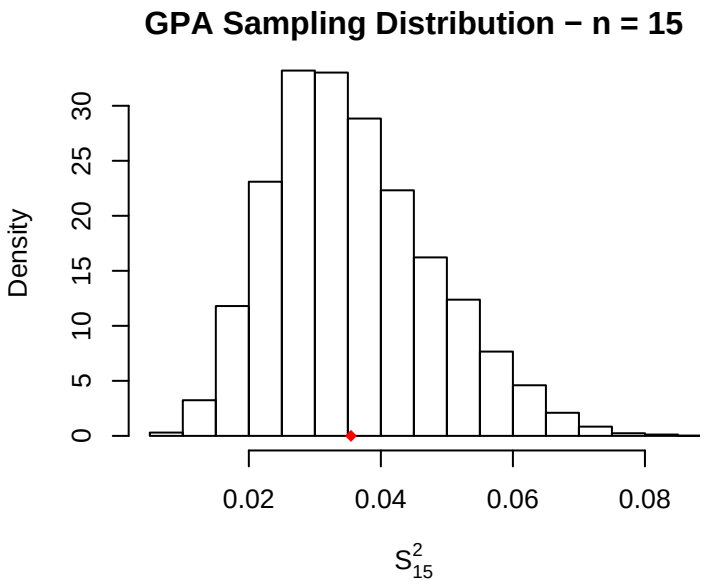

LSAT Sampling Distribution − n = 15
Density
0.00
0.01
0.02
0.03
0.04
570
580
590
600
610
620
630
$\overline{X}_{15}$

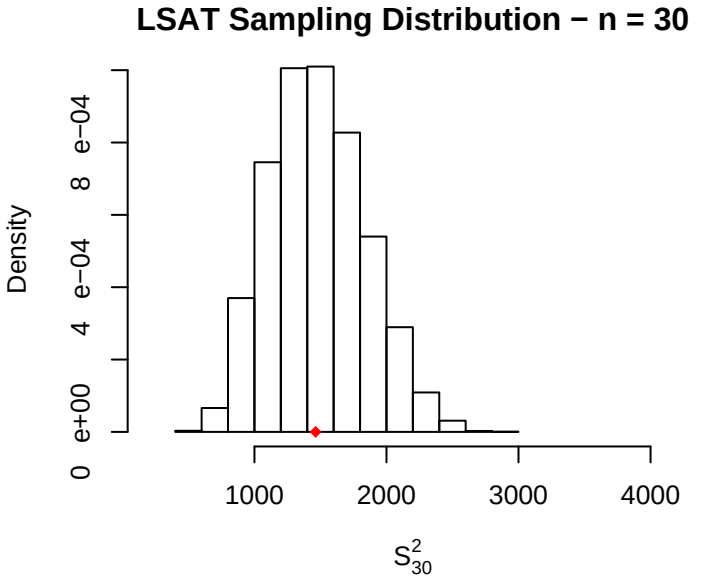

GPA Sampling Distribution − n = 15
Density
0
2
4
6
8
2.95
3.00
3.05
3.10
3.15
3.20
3.25
3.30
$\overline{X}_{15}$

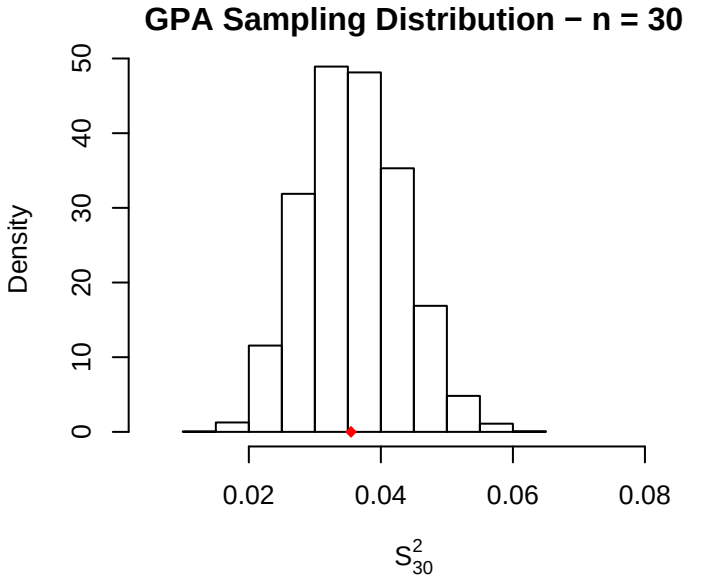

LSAT Sampling Distribution − n = 30
Density
0.00
0.02
0.04
0.06
570
580
590
600
610
620
630
$\overline{X}_{30}$

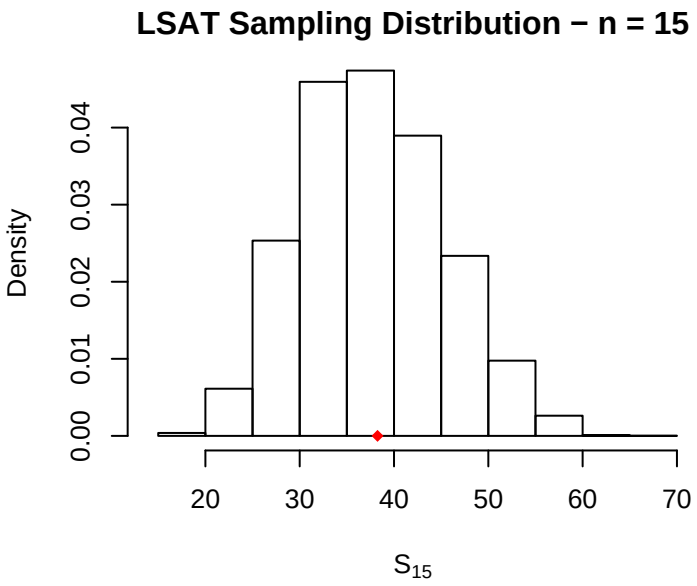

GPA Sampling Distribution − n = 30
Density
0
2
4
6
8
10
12
14
2.95
3.00
3.05
3.10
3.15
3.20
3.25
3.30
$\overline{X}_{30}$

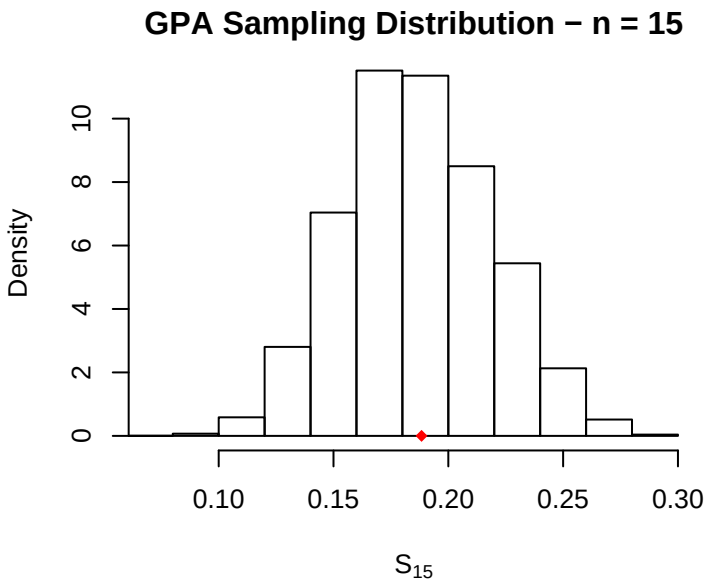

LSAT Sampling Distribution − n = 15
Density
0
e+00
4
e−04
1000
2000
3000
4000
S^2_{15}

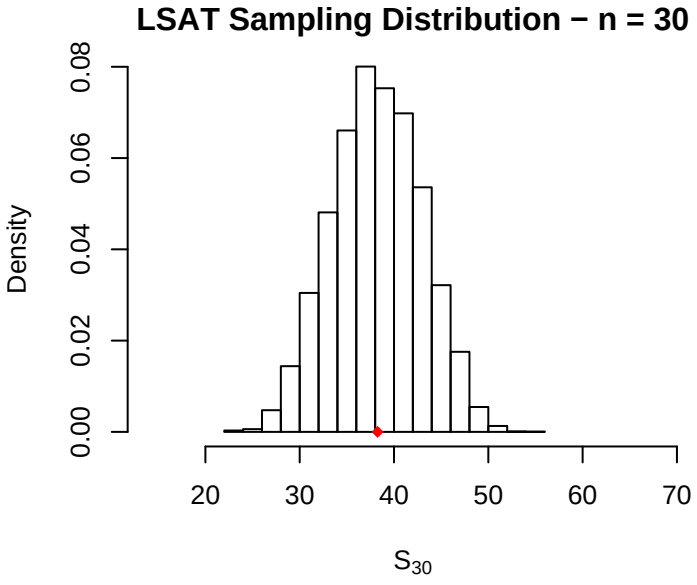

GPA Sampling Distribution − n = 15
Density
0
5
10
15
20
25
30
0.02
0.04
0.06
0.08
S^2_{15}

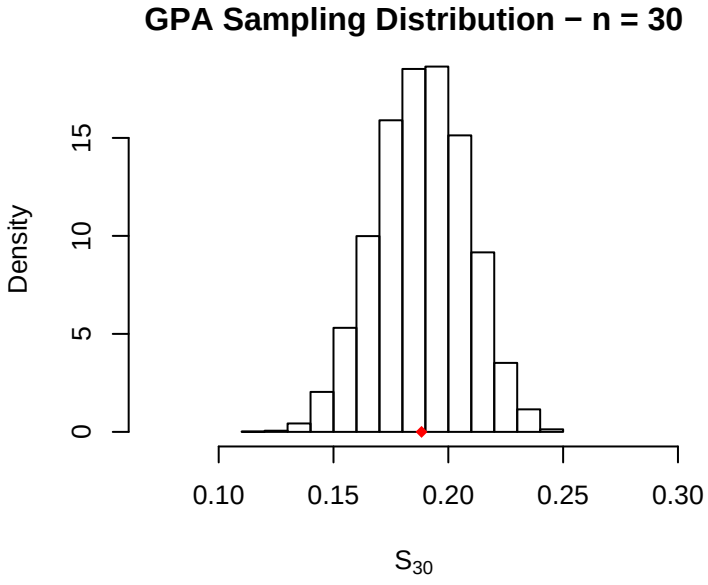

LSAT Sampling Distribution − n = 30
Density
0
e+00
4
e−04
8
e−04
1000
2000
3000
4000
S^2_{30}


GPA Sampling Distribution − n = 30
Density
0
10
20
30
40
50
0.02
0.04
0.06
0.08
S^2_{30}


LSAT Sampling Distribution − n = 15
Density
0.00
0.01
0.02
0.03
0.04
20
30
40
50
60
70
S_{15}


GPA Sampling Distribution − n = 15
Density
0
2
4
6
8
10
0.10
0.15
0.20
0.25
0.30
S_{15}


LSAT Sampling Distribution − n = 30
Density
0.00
0.02
0.04
0.06
0.08
20
30
40
50
60
70
S_{30}


GPA Sampling Distribution − n = 30
Density
0
5
10
15
0.10
0.15
0.20
0.25
0.30
S_{30}

Notice that for all three statistics $(\bar{x}, s^2, s)$ and both variables, the spread of the sampling distribution is smaller for the bigger sample. This should be expected, as with the bigger sample, we have more information and therefore should be able to make a better guess.

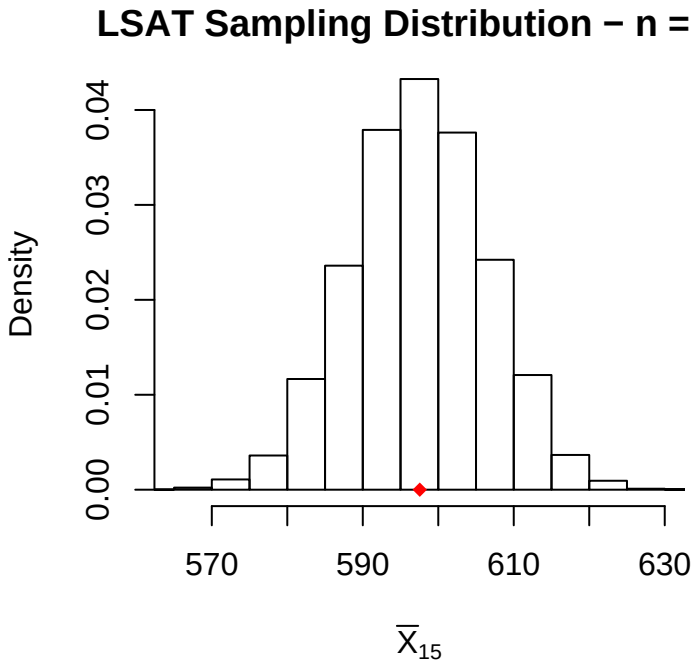
Properties of interest for a sampling distribution:

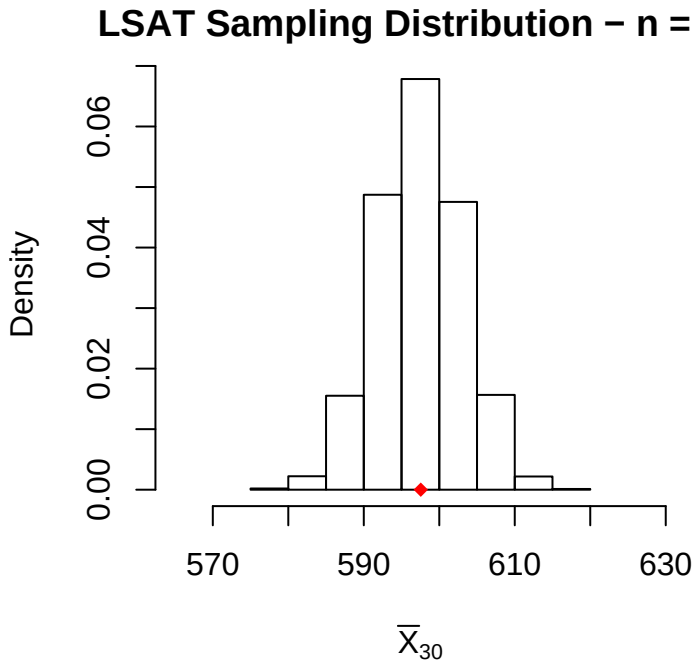
- Shape: Unimodal or bimodal, skewed or symmetric. In many situations, the sampling distribution looks normal.

- Center: The mean is often the measure of center, since sampling distributions are often approximately normal, but the median can also be of interest.

- Spread: The standard deviation is the usual measure of spread for a sampling distribution.

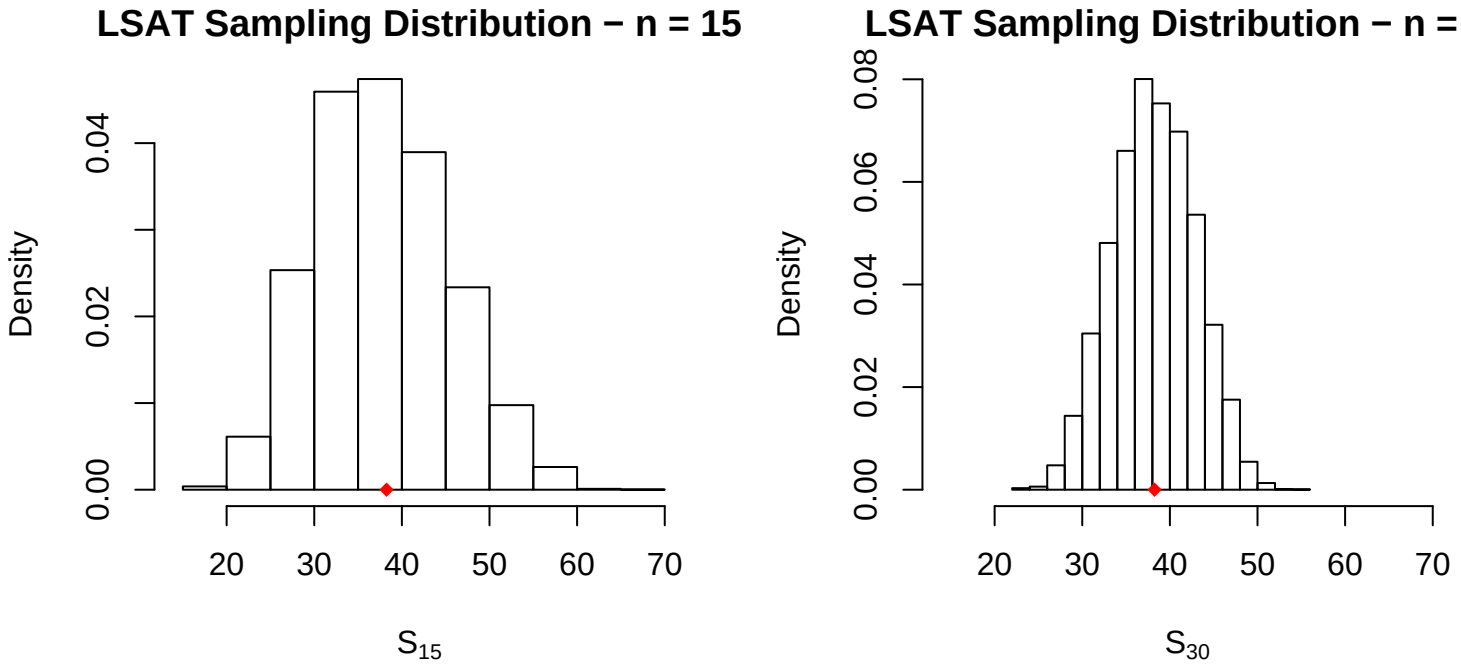
Lets look at the histograms of the sampling distributions for the sample mean of the LSAT scores. (True mean of LSAT $\mu = 597.55$)





Sample Size	$n = 15$	$n = 30$
Average	597.56	597.49
Std Dev	8.97	5.59

Lets look at the histograms of the sampling distributions for the sample standard deviation of the LSAT scores. (True standard deviation of LSAT $\sigma = 38.25$)



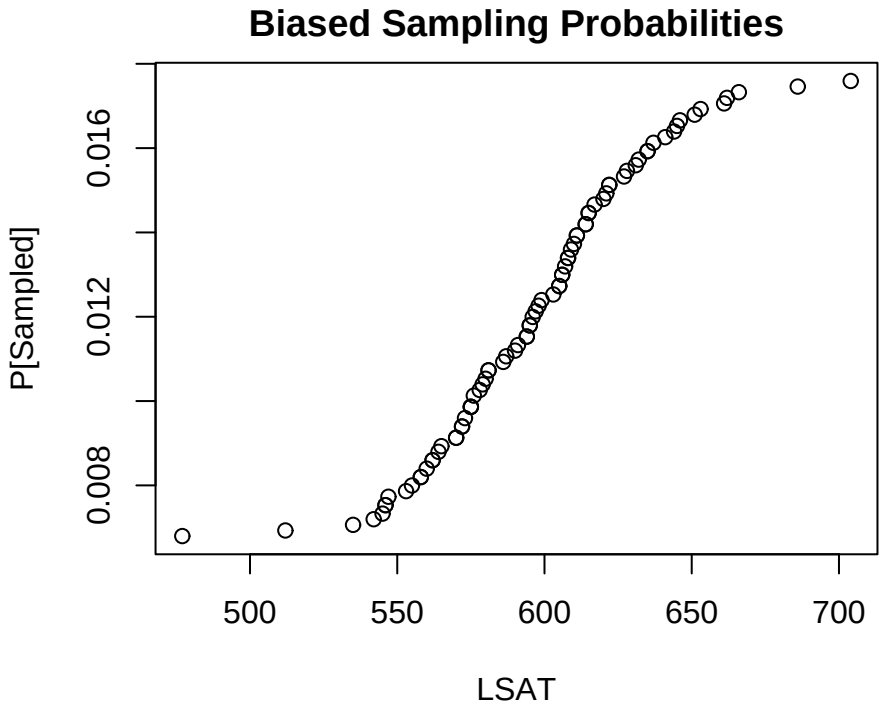
Sample Size	$n = 15$	$n = 30$
Average	37.73	38.18
Std Dev	7.56	4.69

Now lets suppose that a poor sampling scheme was used. As we have seen in the examples, poor sampling schemes can lead to bad answers.

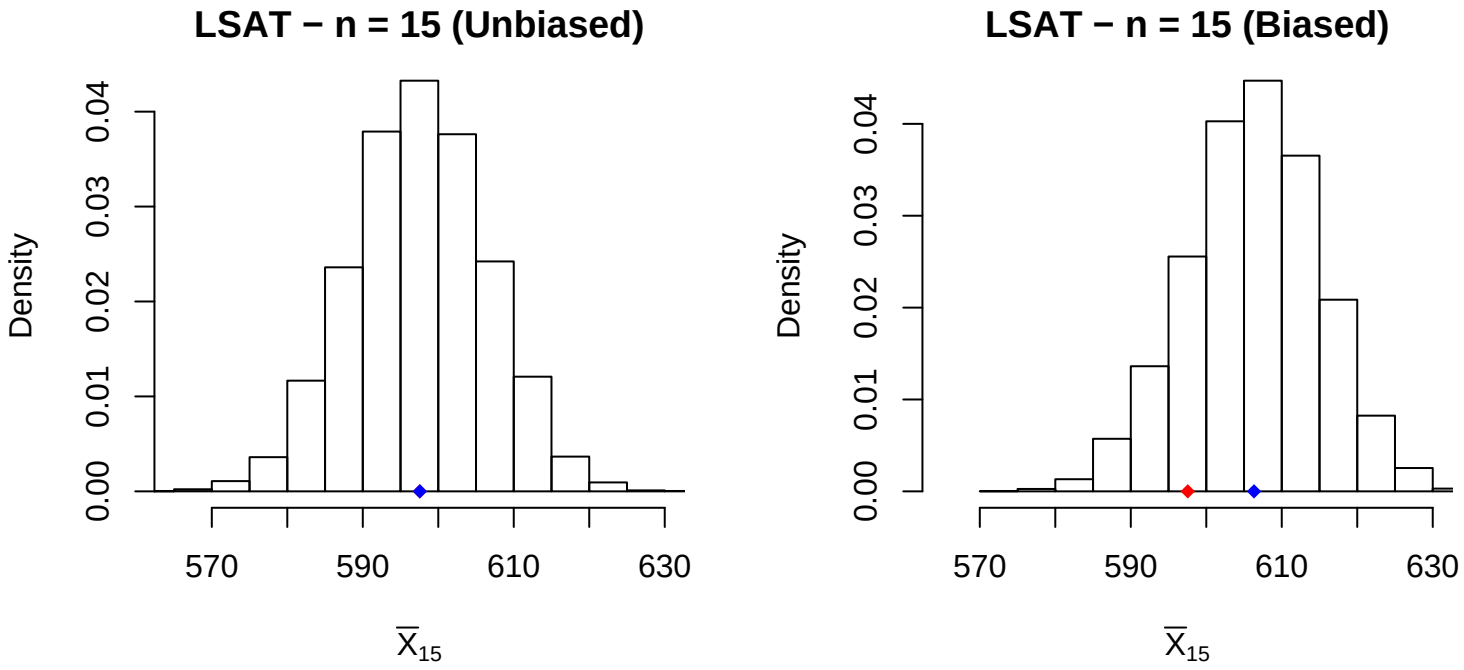
In the following plots, the biased sampling scheme has a higher selection probability for schools with higher LSAT scores. The score with the highest LSAT is 2.5 times more likely to be sampled than the school with the lowest LSAT

$$P[\text{Sampled}] \propto \text{Rank} + 50$$

The mean of the 10,000 simulated sample statistics is given by the blue diamond in the following plots

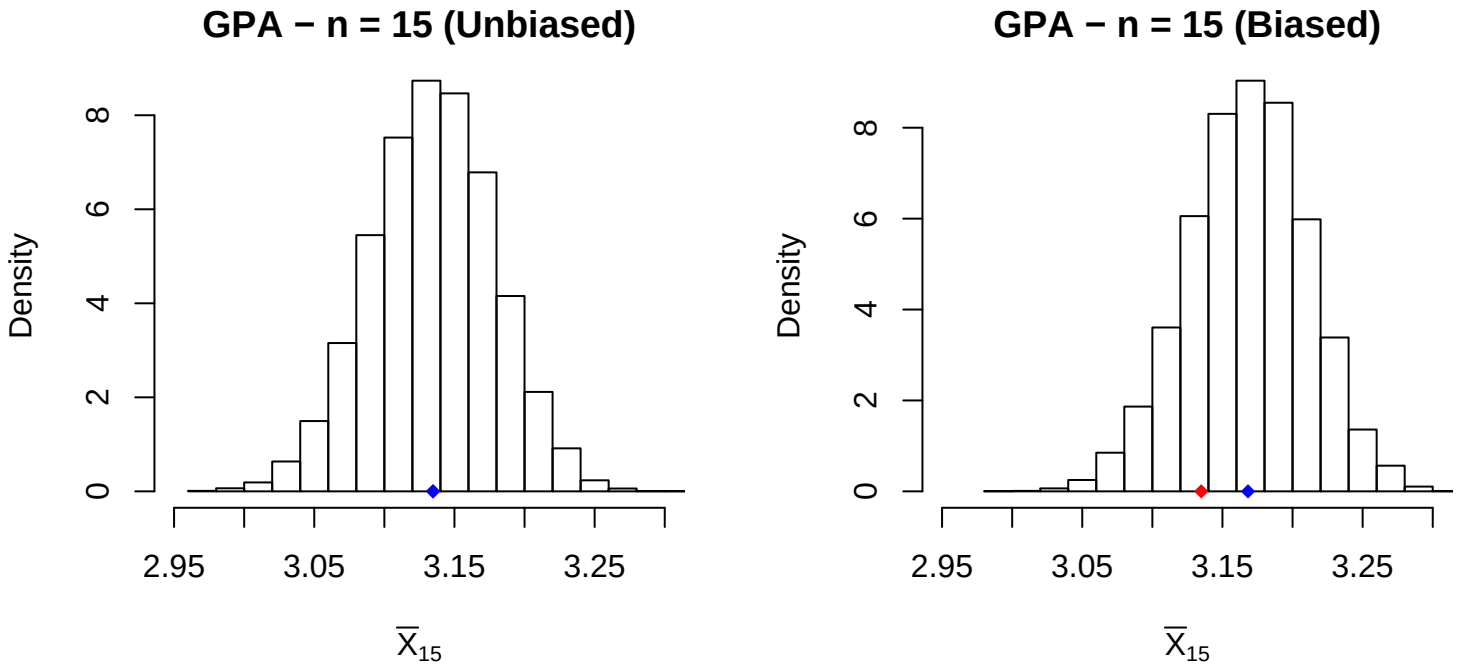


Mean LSAT $n = 15$ ($\mu = 597.55$)



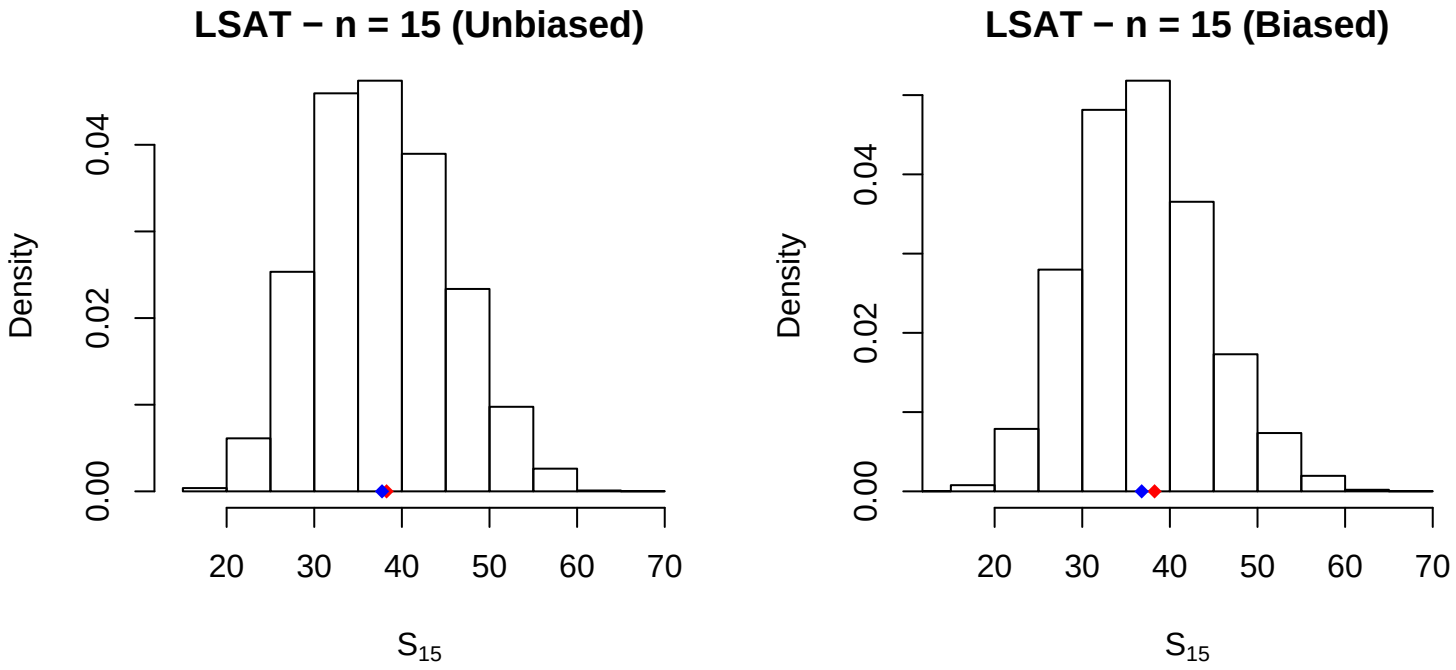
Sampling scheme	Unbiased	Biased
Average	597.56	606.32
Std Dev	8.97	8.80

Mean GPA $n = 15$ ($\mu = 3.135$)



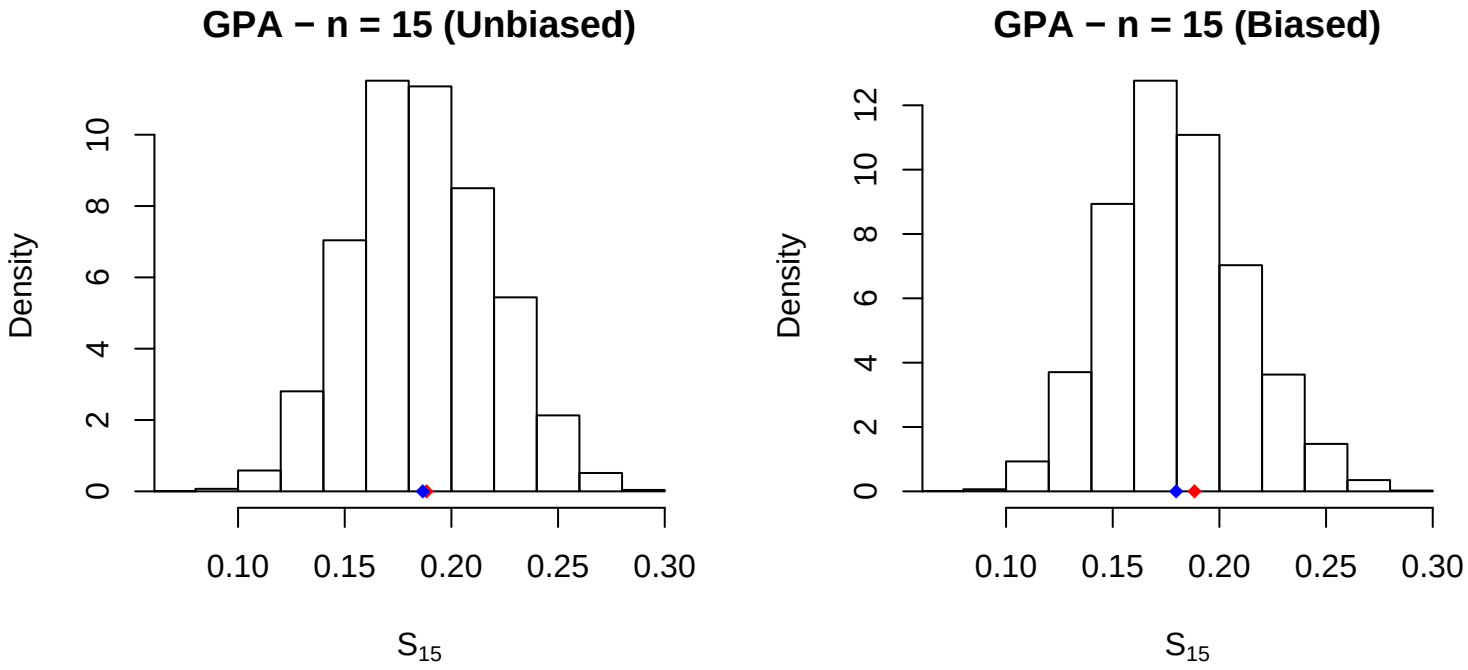
Sampling scheme	Unbiased	Biased
Average	3.135	3.168
Std Dev	0.0442	0.0427

Standard deviation LSAT $n = 15$ ($\sigma = 38.25$)



Sampling scheme	Unbiased	Biased
Average	37.73	36.78
Std Dev	7.56	7.34

Standard deviation GPA $n = 15$ ($\sigma = 0.1883$)



Sampling scheme	Unbiased	Biased
Average	0.1866	0.1797
Std Dev	0.0322	0.0312

Bias:

Concerns the center of the sampling distribution. A statistic is said to be unbiased if the mean of its sampling distribution is equal to the true value of the parameter. Otherwise, it is said to be biased.

$$\text{Bias} = \text{Mean(Sampling Distribution)} - \text{Parameter}$$

In the law school examples, the two SRS schemes ($n = 15$ and 30) lead to unbiased estimates of μ (by $\bar{x}$) and σ^2 (by s^2). The estimate of σ (by s) has a small negative bias.

However the biased sampling scheme leads to biased estimates of all three parameters. For the means of the two variables, estimates of the bias are

	Bias	% Bias	Z-score
LSAT	8.77	1.47	0.98
GPA	0.033	1.05	0.77

In the poll investigating whether the Ohio drinking age should be raised, it appears that the poll had positive bias. More people in the poll wanted to raise the drinking age than in the actually voting population.

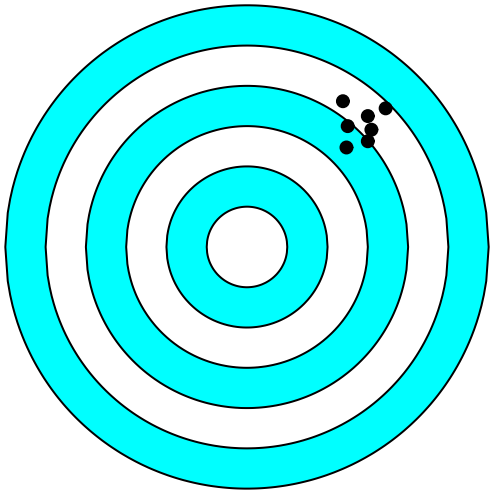
Variability of a statistic:

Measured by the standard deviation of the sampling distribution. The spread depends on the population, the sampling design, and the sample size n. The bigger n, the smaller the spread.

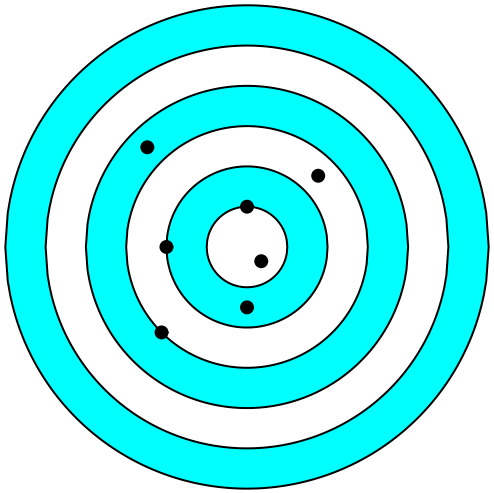
Individual estimates of parameters have

Bias
+
Random Variation

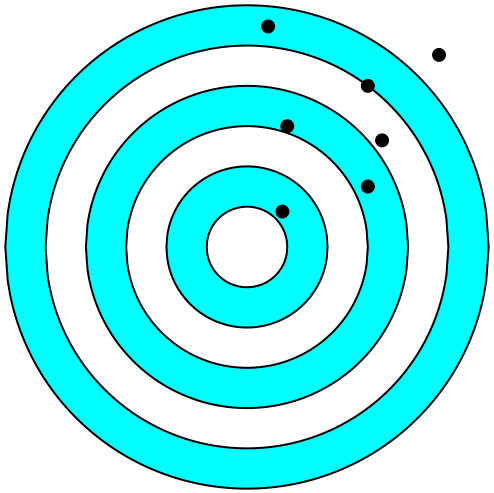
High bias, low variability



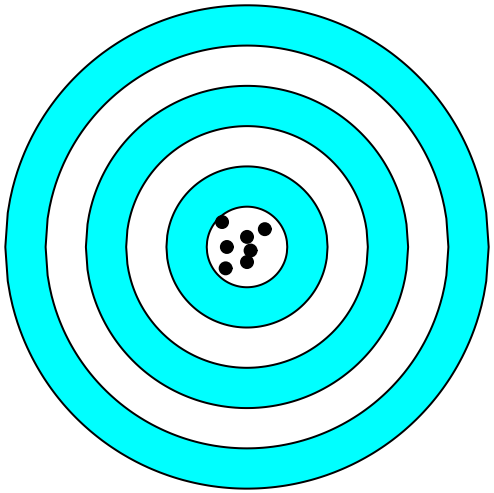
Low bias, high variability



High bias, high variability



Low bias, low variability



Generally you want

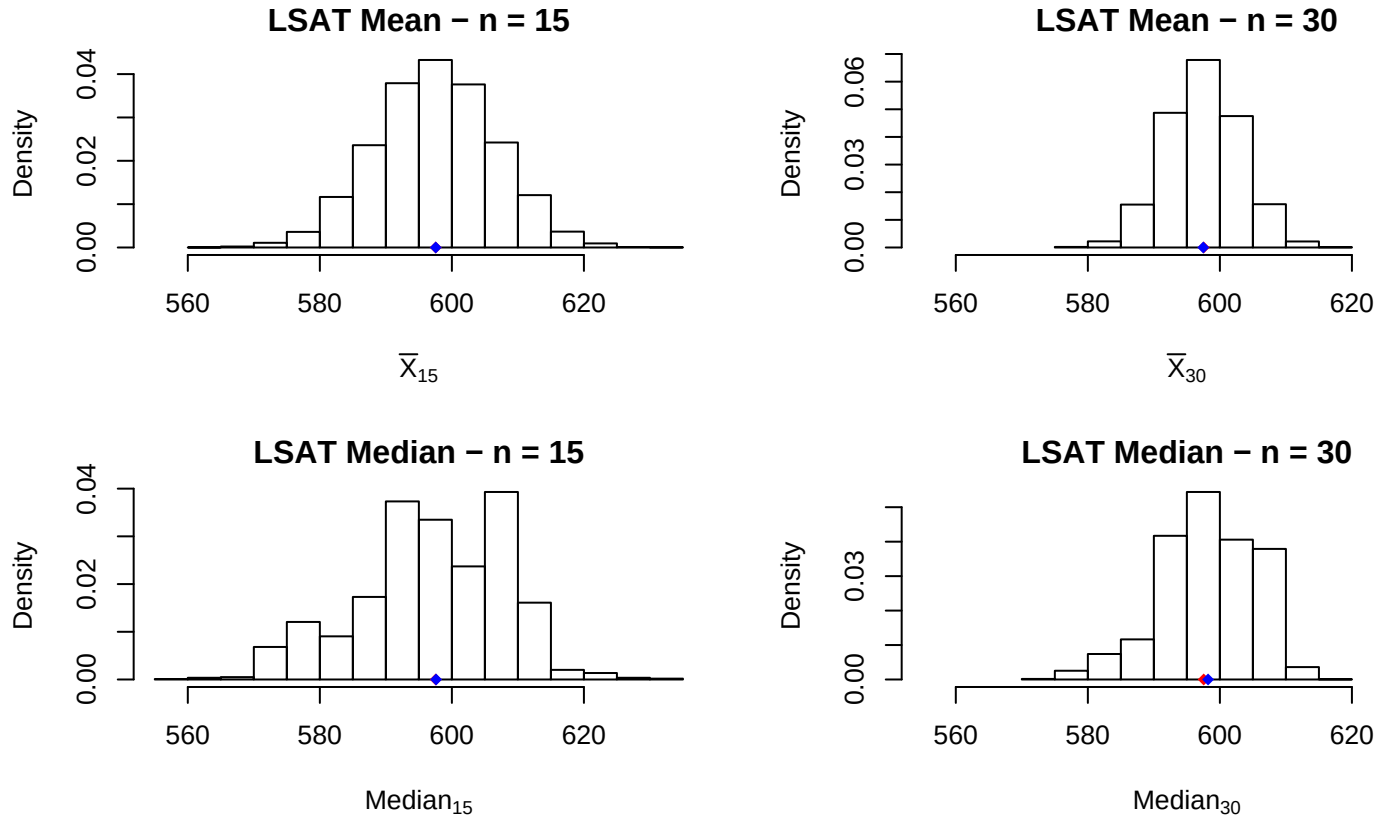
$$\text{Low Bias} + \text{Low Variability}$$

To reduce bias, use random sampling. If we have an accurate list of the population, SRS (plus other random sampling schemes) will give unbiased estimates (or close to unbiased estimates in some cases) of the parameters.

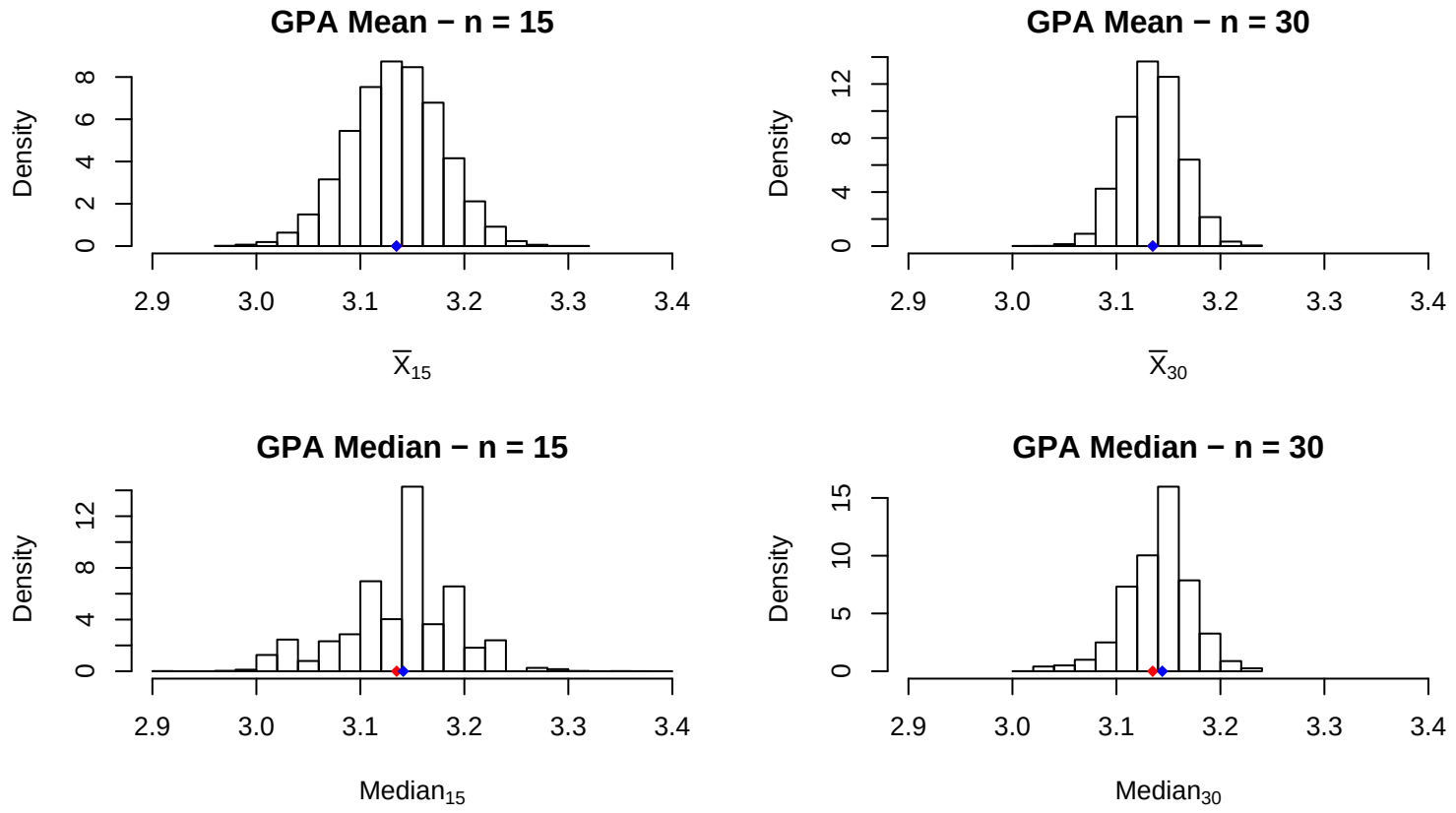
To reduce the variability of a statistic from a SRS, increase the sample size.

Thought also needs to be put into the estimator. It is possible to have two unbiased estimators for a parameter. Then you would want to use the one with lower sampling variability.

For a population with a symmetric distribution (e.g. a normal), the population mean = the population median. The sample mean is a better estimator for the population mean than the sample median, as its sampling distribution has a smaller standard deviation, even though both happen to be unbiased estimators.



Now suppose that the population has a skewed distribution (so the population mean $\neq$ the population median) and you still wish to estimate the population mean. In this case, you would want to use the sample mean has the sample median is an biased estimate of the population mean.



(Note the difference between the mean and median isn't very big here since the distribution of GPA isn't strongly skewed.)

Effect of Population Size

The variability of a statistics from a random sample does not depend on the size of the population, as long as the population is at least 100 times larger than the sample.

A consequence of this is that asking 1000 people in the Boston area (pop: 5.8 million) has about as much precision as asking 1000 people nationally (pop: 290 million).

For the law school example, since the sampling fractions are fairly large, this actually decreases the standard deviation of the sampling distribution.