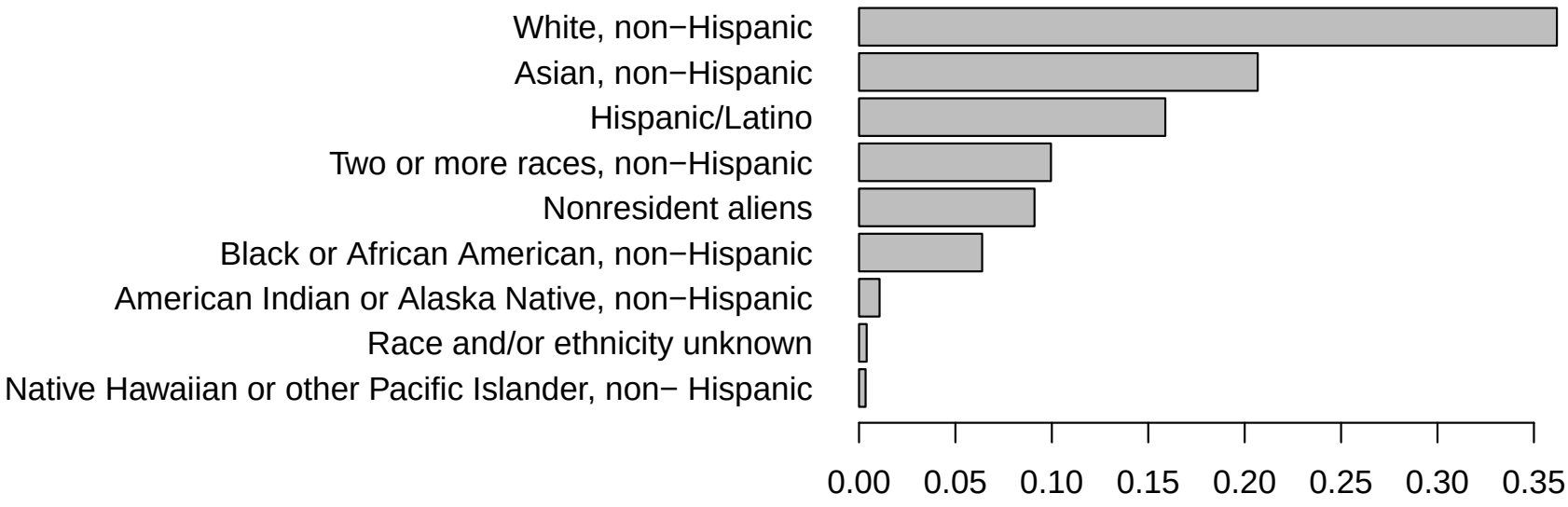
Sampling and chance variability

Data Science 101 Team

The population of Stanford undergraduate students

	Racial/Ethnic category	Degree-Seeking First-Time First Year	Degree-Seeking Undergraduates	Total Undergraduates (both degree- and non-degree-seeking)
1	Nonresident aliens	172	640	642
2	Hispanic/Latino	287	1117	1117
3	Black or African American, non-Hispanic	130	449	449
4	White, non-Hispanic	601	2545	2545
5	American Indian or Alaska Native, non-Hispanic	19	75	75
6	Asian, non-Hispanic	374	1454	1454
7	Native Hawaiian or other Pacific Islander, non- Hispanic	7	24	24
8	Two or more races, non-Hispanic	137	700	700
9	Race and/or ethnicity unknown	11	28	28
10	Total	1738	7032	7034

The population of Stanford undergraduate students



Learning from a sample

- Suppose we did not have this comprehensive dataset with exact information on the **population** of Stanford undergraduate students and we were interested e.g. in the proportion of students who are "Black or African American".
- This is called a parameter. Here the parameter is 6.4%, but it is *unknown* to us. There can be more than one parameter: We may also be interested in the proportion of students who are "Hispanic / Latino".
- We could try to estimate this starting from a **sample**: We can sample 100 students at random and use the proportion of Black or African American students in our sample as an **estimate** for our parameter.
- How big should the sample be? What would be a good strategy to obtain it?

Would sampling work?

We can try a thought experiment using our complete data and see what the result of a sampling experiment would be.

```
set.seed(123)
ssize<-50
observation <- sample(UGRace,ssize,replace=FALSE)
SamplePropB <- sum(observation=="Black or African American, non-Hispanic")/ssize
SamplePropB
```

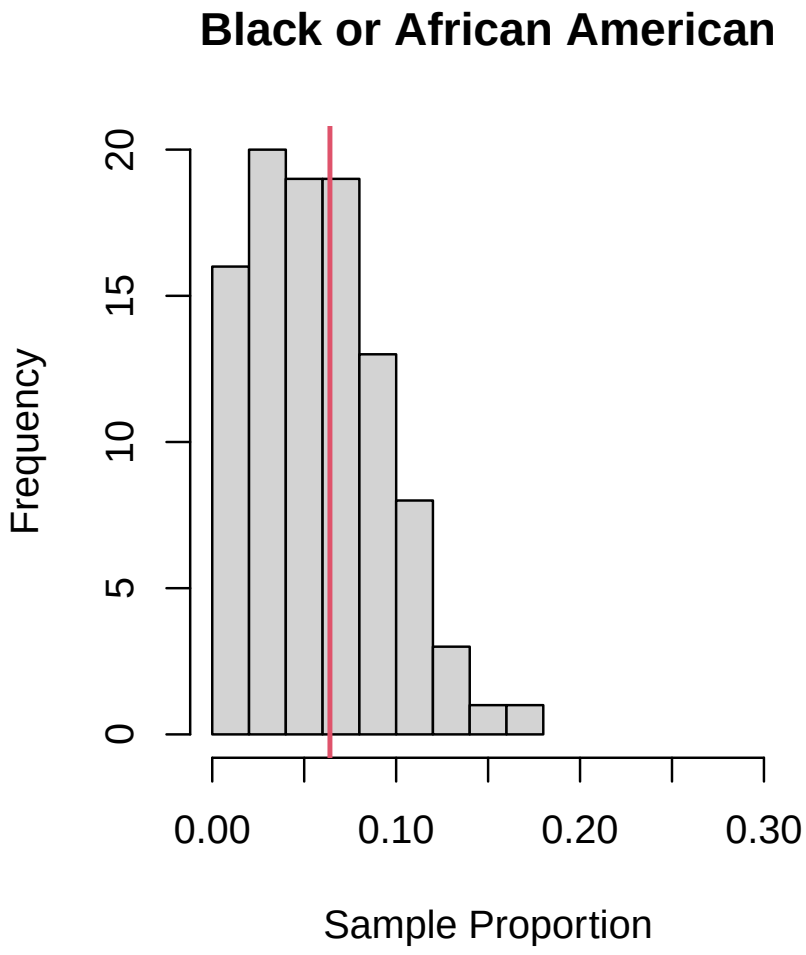
```
[1] 0.08
```

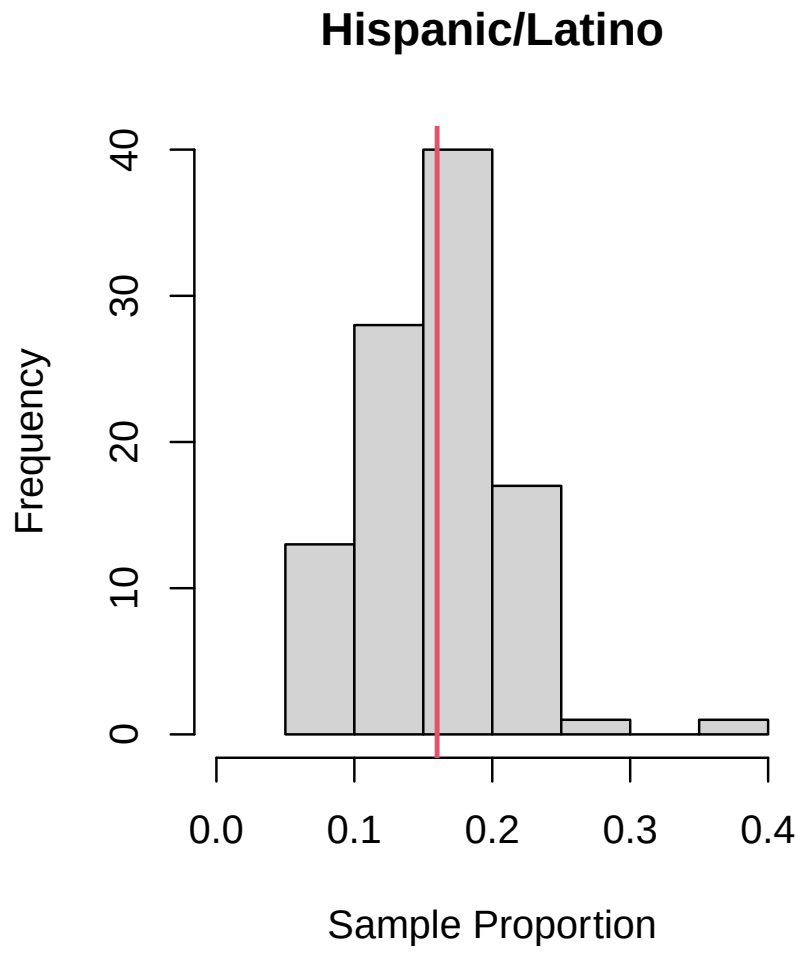
```
SamplePropH <-sum(observation=="Hispanic/Latino")/ssize
SamplePropH
```

```
[1] 0.18
```

Repeating the thought experiment 100 times

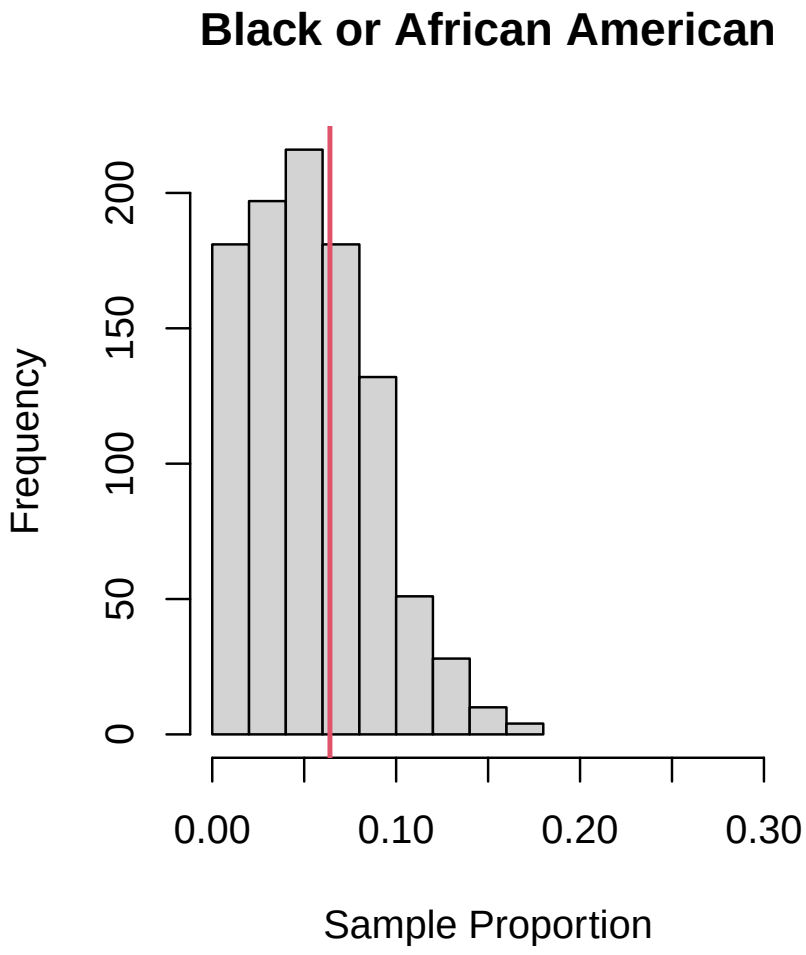
Sampling proportions for 100 samples of size 50

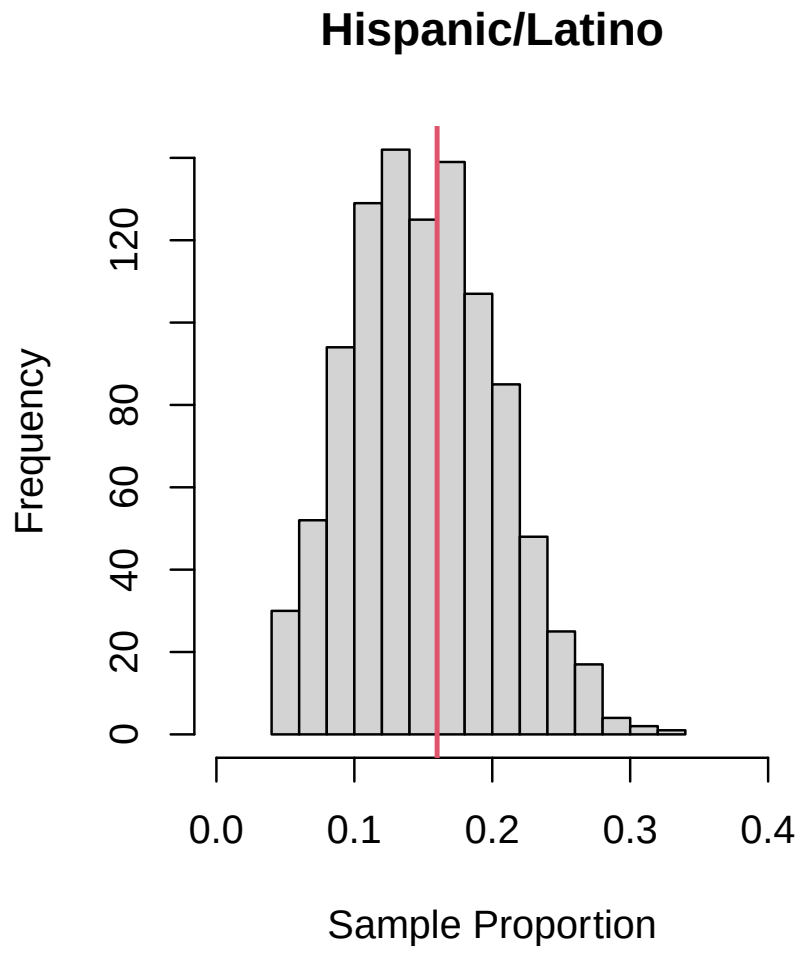




Repeating the thought experiment 1000 times

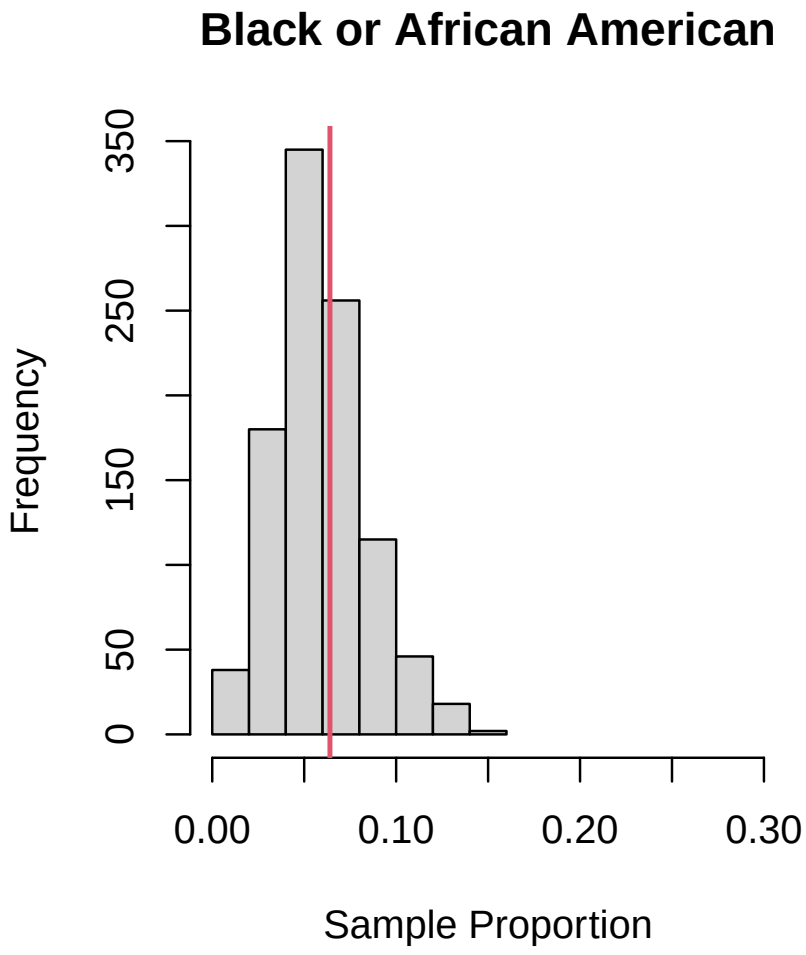
Sampling proportions for 1000 samples of size 50

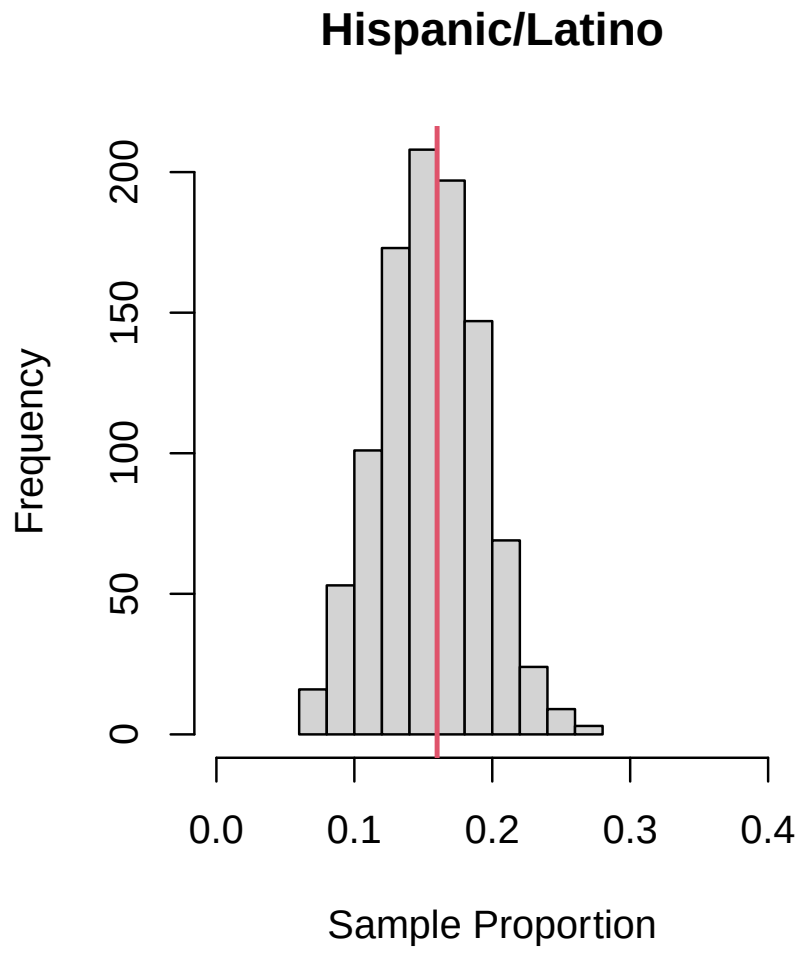




Using a larger sample size

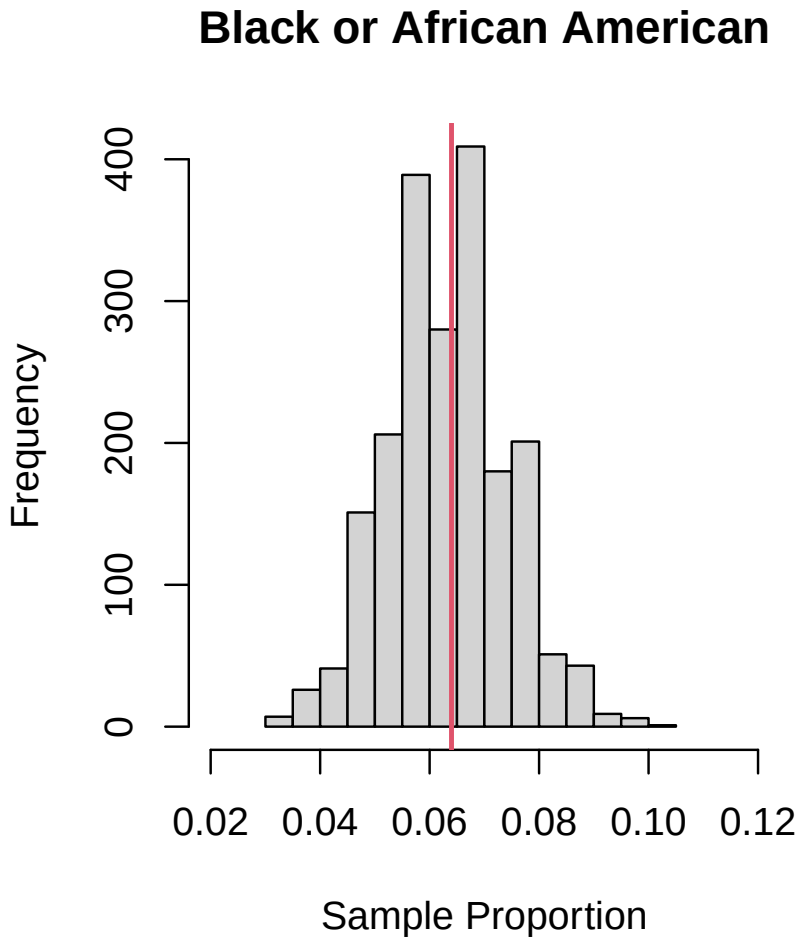
Sampling proportions for 1000 samples of size 100

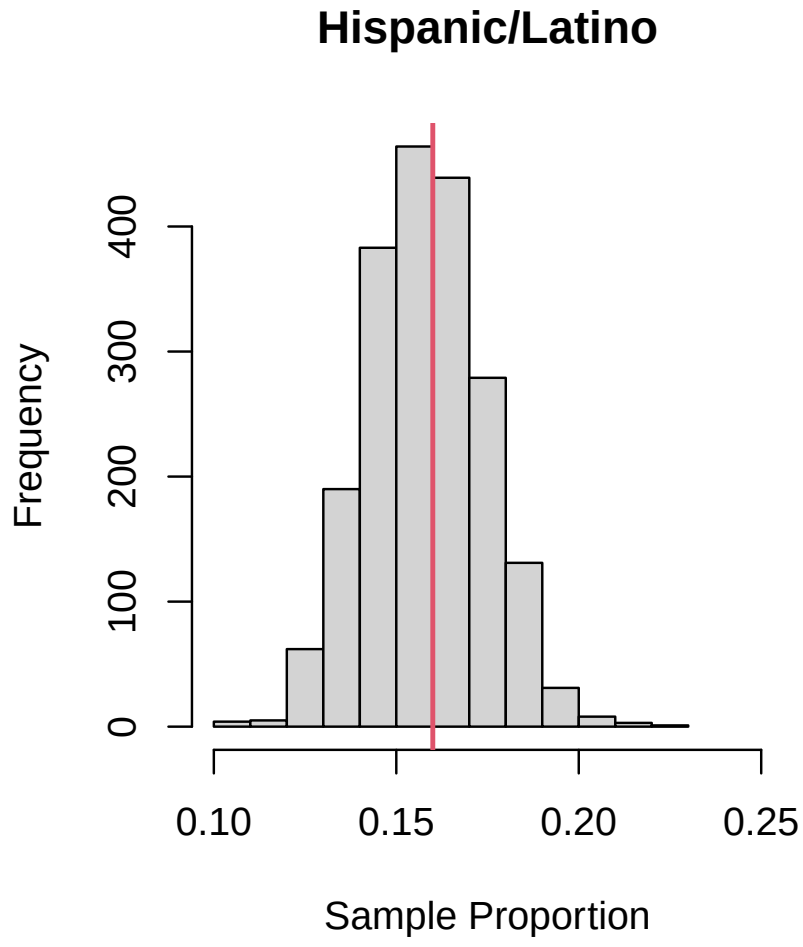




Using an even larger sample size

Sampling proportions for 2000 samples of size 500





Things to note

- Each sample gives us potentially a different result: there is **sampling variability**.
- We need to remember this sampling variability when **interpreting the sampling results**.
- A larger sample size resulted in smaller variability.
- Key point (explained in more detail later): The accuracy of the estimate depends on the size of the sample, not on the size of the population. So even a relatively small sample (100 or 1000) produces an estimate that is close to the parameter of a very large population of (say) 100 million subjects.

Sampling correctly is very important

- When sampling voters to estimate the President's approval rating, it is tempting to sample 1000 voters in your hometown.
- This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.
- This will introduce bias, i.e. the sampling will favor a certain outcome.
- **Selection bias**: A sample of convenience makes it more likely to sample certain subjects than others.
- **Non-response bias**: Parents are less likely to answer a survey request at 6 pm because they are busy with children and dinner.
- **Voluntary response bias**: Websites that post reviews of businesses are more likely to get responses from customers who had very bad or very good experiences.

Sampling Designs

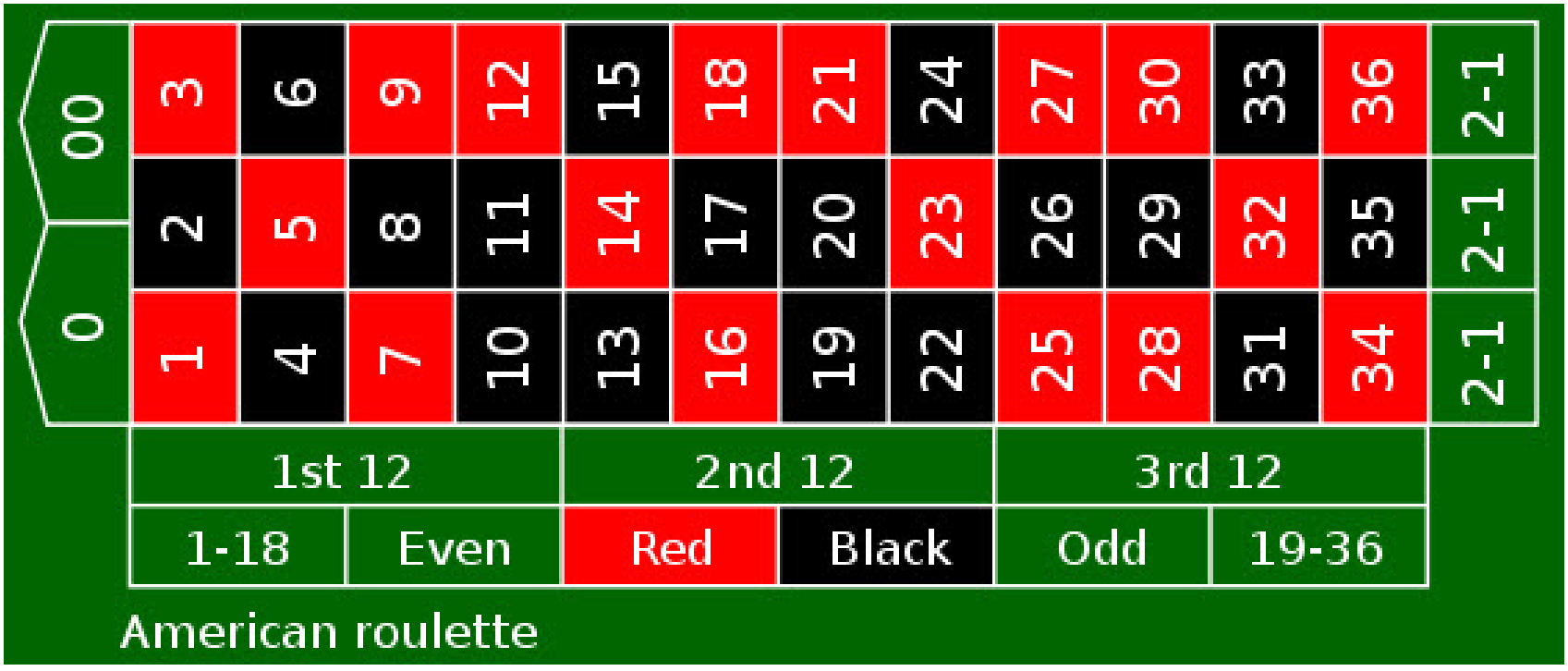
The best methods for sampling use chance in a planned way:

- **Simple Random Sample**: Select subjects at random without replacement.
- **Stratified Random Sample**: Divides the population into groups of similar subjects called strata (e.g. urban, suburban, and rural voters). Then one chooses a simple random sample in each stratum and combines these.

Learning from Data

- Much of the data we work with does not contain information on all the members of a group of interest, but only on a subset.
- We often want to infer from the data something that goes beyond the subset we have measured.
- We typically do not expect to learn everything about a population from a sample.
- Rather we want to learn about the value of a population summary from the corresponding sample summary.
- Example: In the cell-phone study, the **population** was "every potential driver", and our sample were **drivers involved in a collision in Toronto between 7/1/1994 and 8/31/1995**. We were interested in the risk of collision when using phones.

Roulette



- Each spin of the roulette wheel yields either a number from 0 to 36 or “00”.
- Each number also has a color.
- Casino always wins on a “0” or “00”.
- Some variability in what colors are, at least online.

Let's make R play roulette

```
values = c("00", 0:36)
play<-sample(values,1)
play
```

```
[1] "00"
```

Suppose we want to bet on red. Let's make a function that takes a possible value from `values` and tells us whether our bet on red wins or not.

```
red_values = c( 1,  3,  5,  7,  9,  12, 14, 16, 18,
               21, 23,  25, 27, 28, 30, 32, 34,  36)
red_bet = function(spin_value) {
    return(spin_value %in% red_values)
}
red_bet(play)
```

```
[1] FALSE
```

The probability of red

- The probability of a red outcome is 18/38 (this means that when spinning the wheel many times, the fraction of red outcomes will stabilize at 18/38).
- Suppose that we ignored this: Could we estimate it from a **sample**, using the **sample proportion of red outcomes**?

Estimating the probability of red

Recall that the probability of red is $18/38 = 0.474$.

Let's play 5 times:

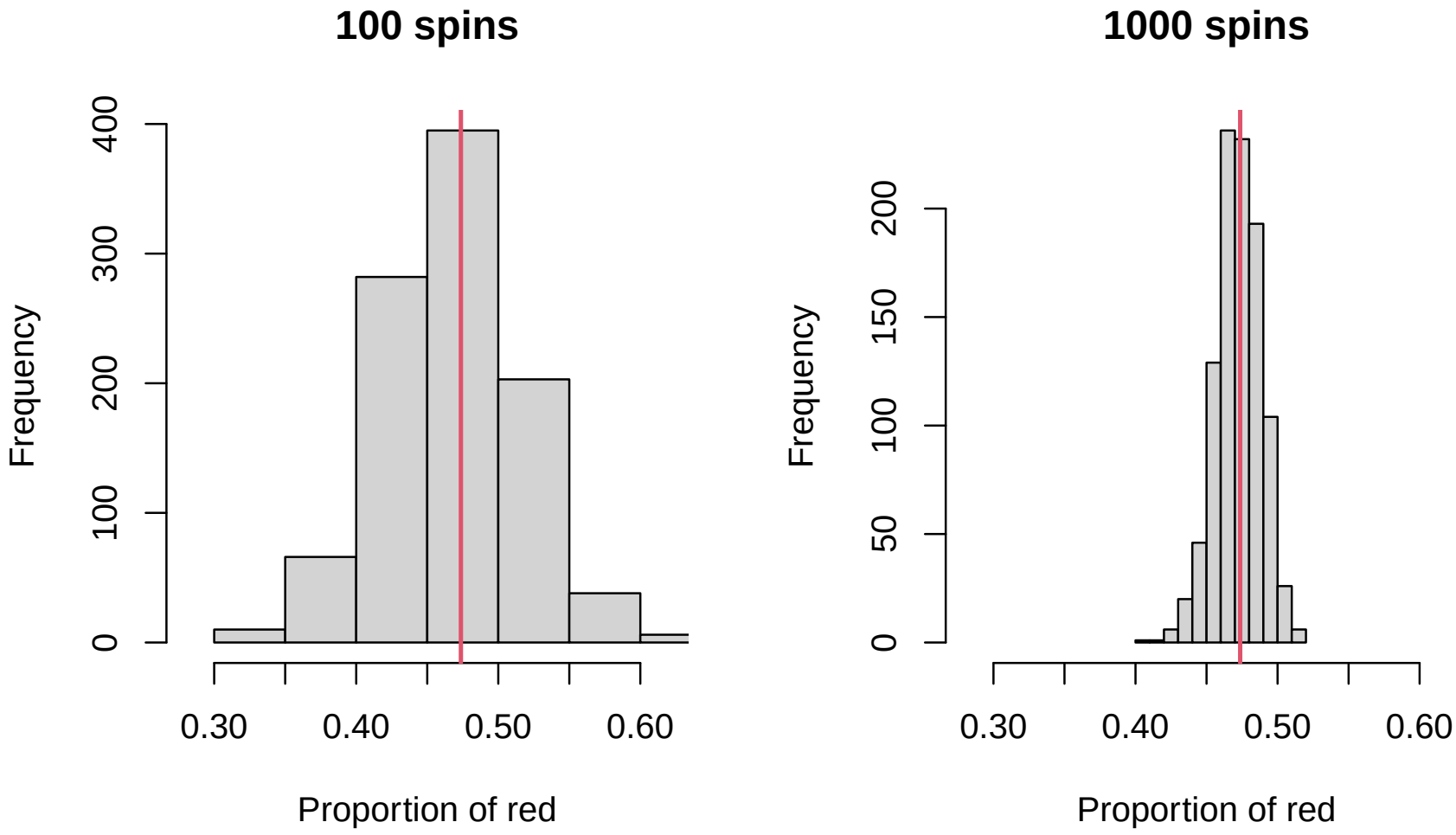
```
[1] FALSE  TRUE  TRUE FALSE FALSE
```

Let's look at the proportion of times in which red comes up in 50 gambles:

```
[1] 0.44
```

Sampling variability of the proportion of red wins

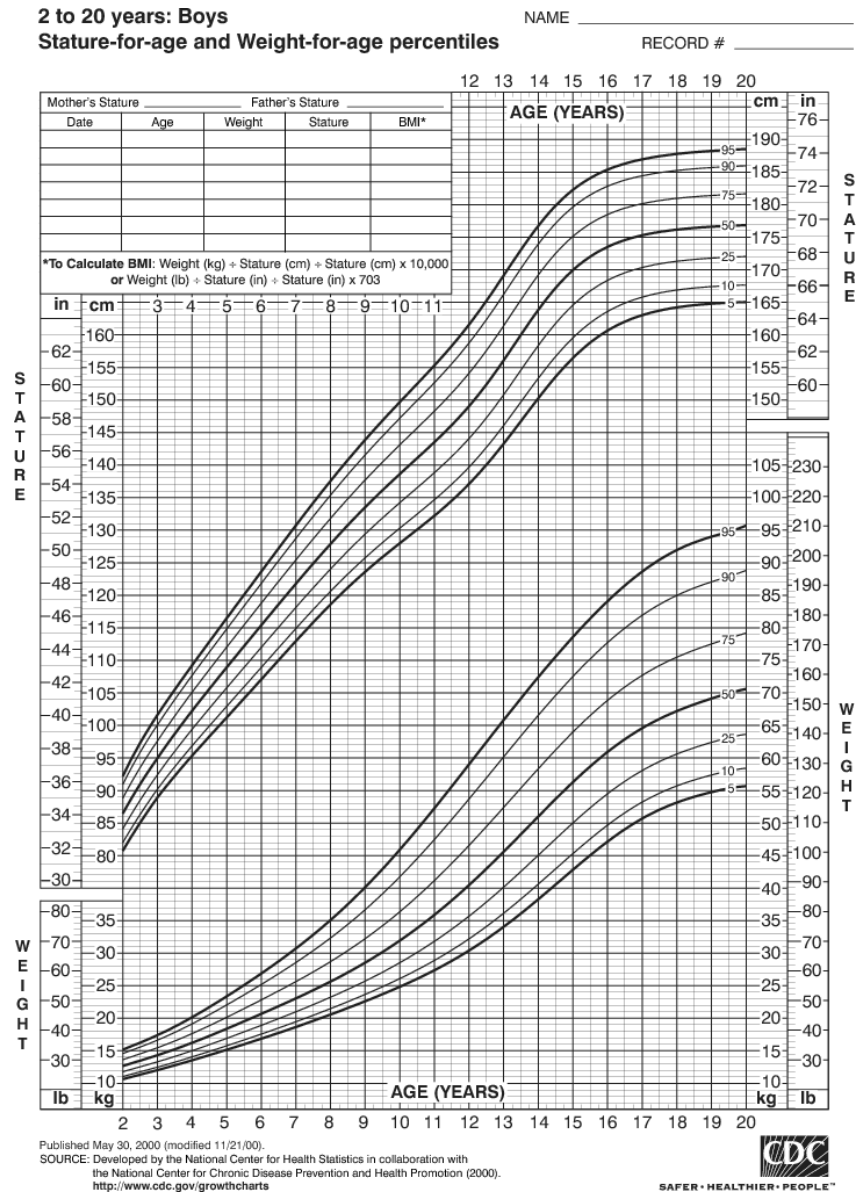
Histograms of 1000 experiments each of 100 or 1000 spins



Take home messages

- Just as we saw from the experiment with Stanford undergraduate population, sample summaries do give us information on the value of the population summaries.
- Each sample potentially has a different value: there is sample variability in our estimate of the population summary.
- The larger the sample, the more accurate is our estimate based on the sample summary.

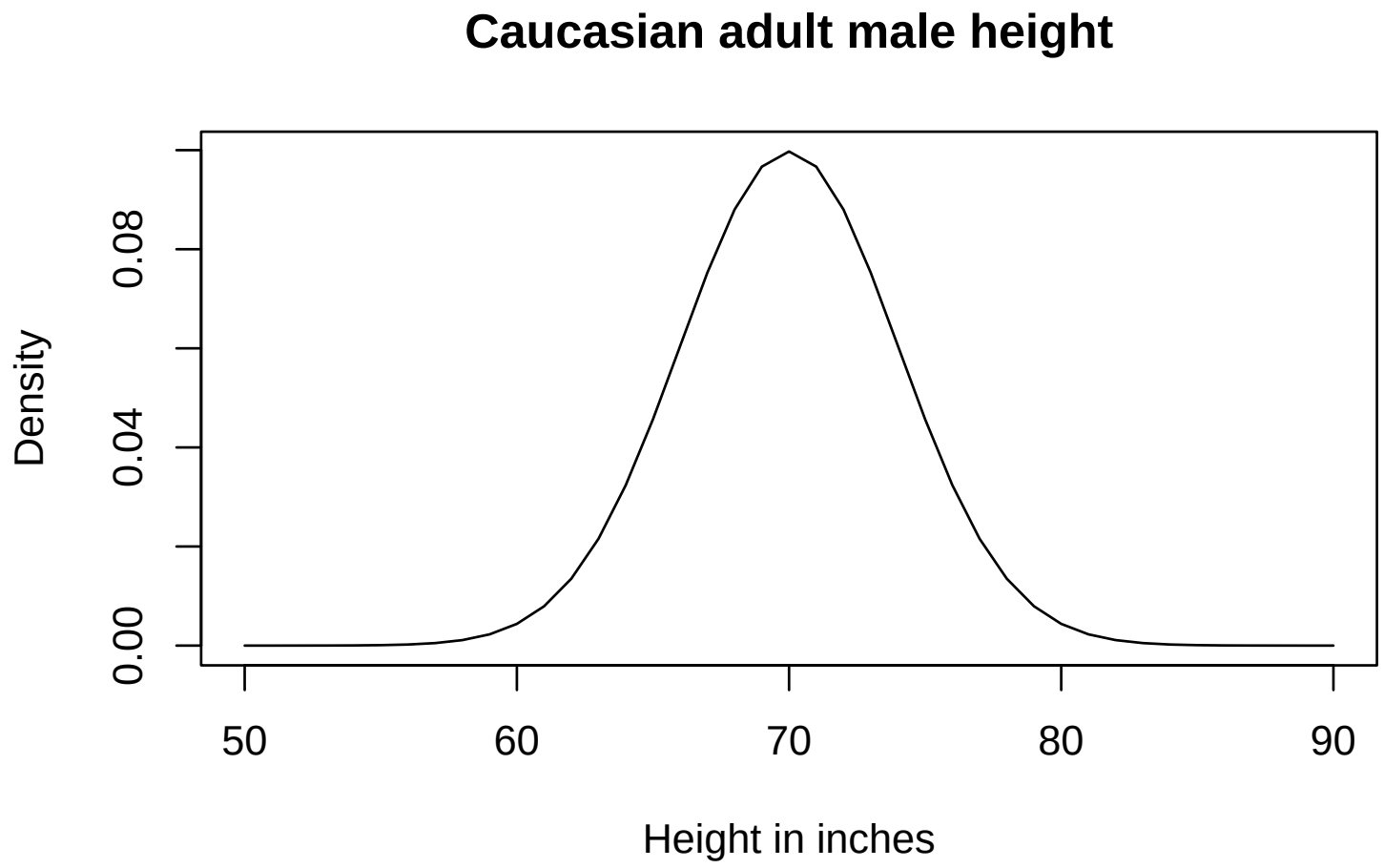
An abstract population



An abstract population

- **Population**: Caucasian males in an age group - one cannot record them all; there is no time limit; we are really referring to an abstract concept.
- **Variables**: Stature and weight.
- **Populations summaries**: 5,10,25,50,75,90,95 percentiles.
- How did the CDC established these values?

A model for the height of US "Caucasian" adult males



Where does this comes from?

Normal distribution

We say that a variable X has Normal distribution with mean μ and standard deviation σ

$$X \sim \mathcal{N}(\mu, \sigma)$$

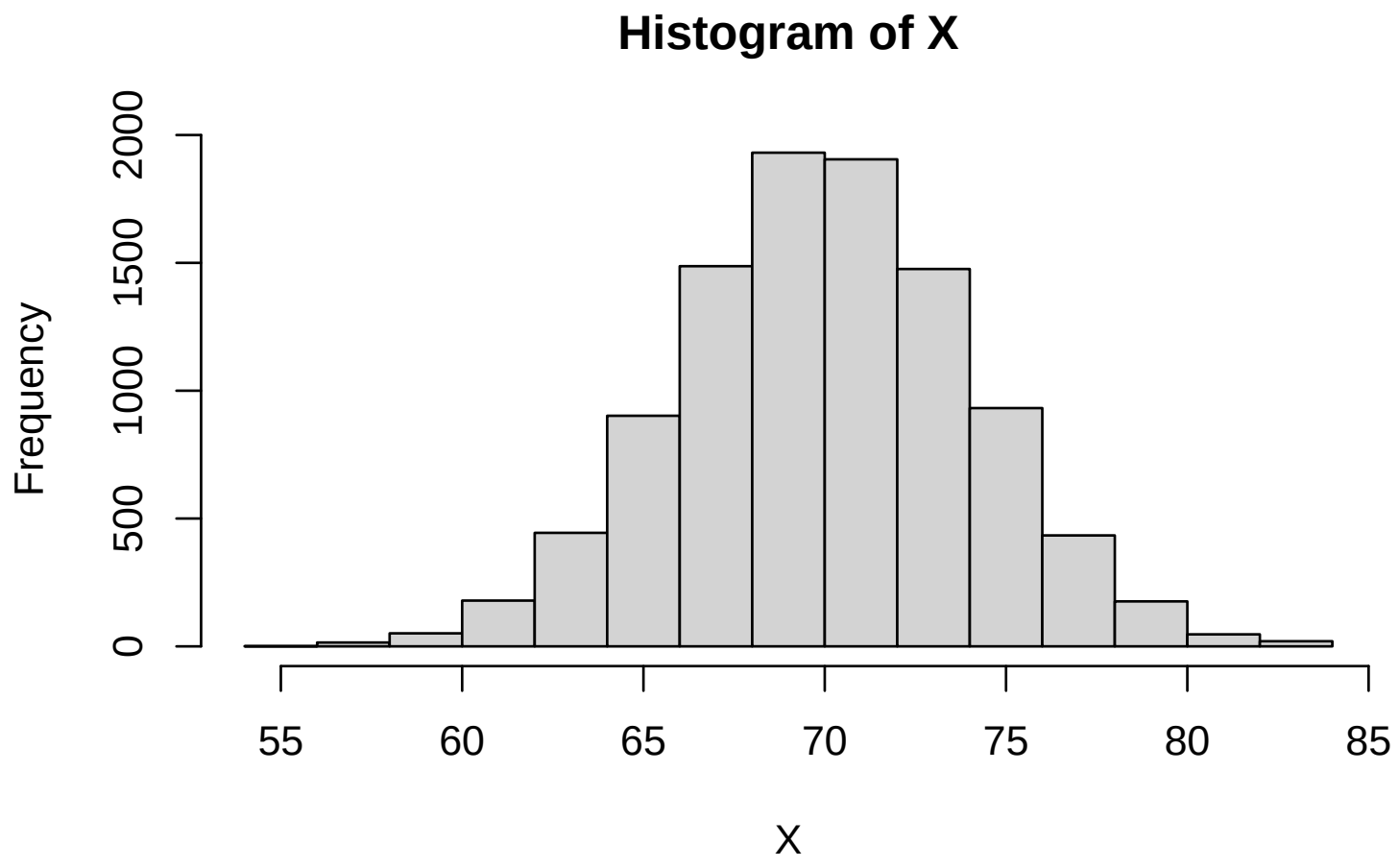
if for a very large sample the "histogram" of all the outcomes of X can be described by the function

$$\text{density}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The mean of X is equal to μ
- The variance of X is equal to σ^2

Generating observations from a Normal in R

```
X<-rnorm(10000,mean=70,sd=4)
hist(X)
```



The histogram of a large sample is close to the density