Estimating with Uncertainty

Question: With limited resources, which of the following Options would give you a more accurate reflection of the Parameters?:

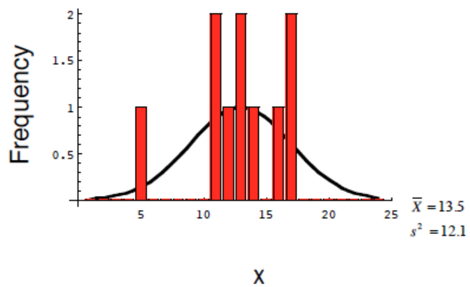
1. 1000 samples, each of 10 individuals

or

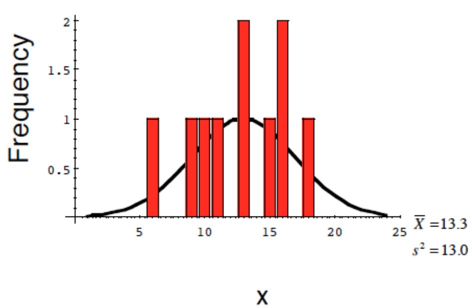
2. 10 samples, each of 1000 individuals?

Frequency distributions NOT sampling distributions

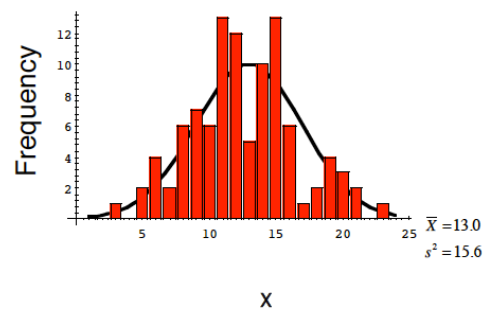
Sample size 10 from Normal distribution with μ=13 and σ^2=16



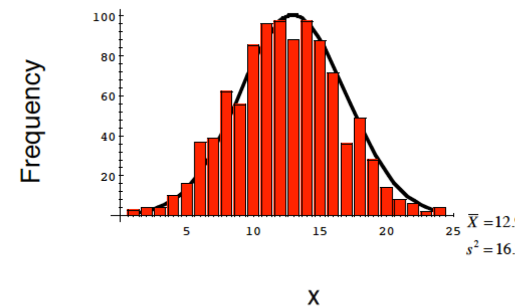
Another sample of 10 from same distribution



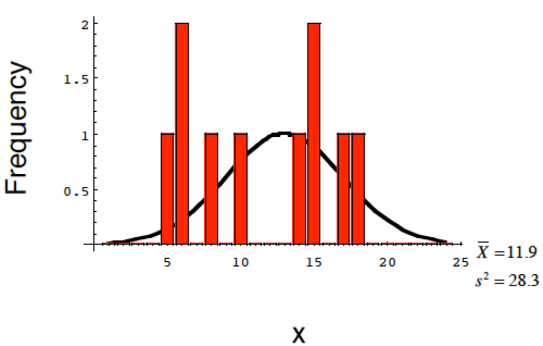
A sample of 100 from the same population distribution



A sample of 1000 from the same population distribution



A third sample of 10 from the same distribution



Estimating with uncertainty

n	$\bar{X}$	s^2
10	13.5	12.1
10	13.3	13.0
10	11.9	28.3
100	13.0	15.6
1000	12.9	16.3

Now we understand the concept of the sampling distribution of an estimate, so let's move on to an example….

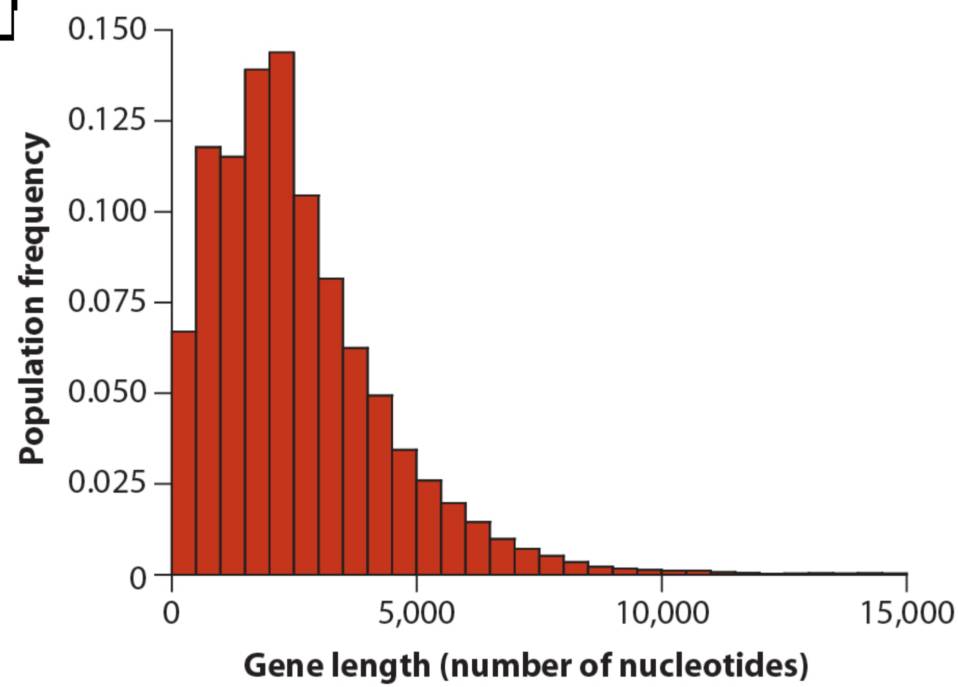
Example: The length of human genes (page 96). From build 35, there are 20,290 genes. Since we ***know*** the size of *all* predicted Genes (**the entire population of gene lengths)**, we are in the unique situation of calculating parameters directly from the data instead of inferring their values from samples. We can compare the values we get from samples to the real, true parameter values.

http://phylo.bio.ku.edu/biostats/geneLenDemo.html

Estimating with uncertainty

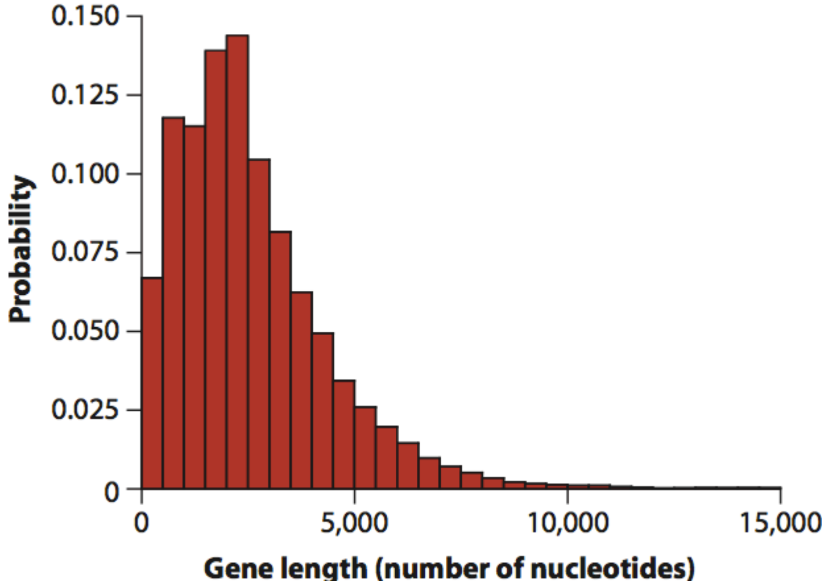
Example: The length of human genes. From build 35, there are 20,290 genes. Since we *know* the size of *all* predicted genes, we are in the unique situation of calculating parameters directly from the data instead of inferring their values from samples.

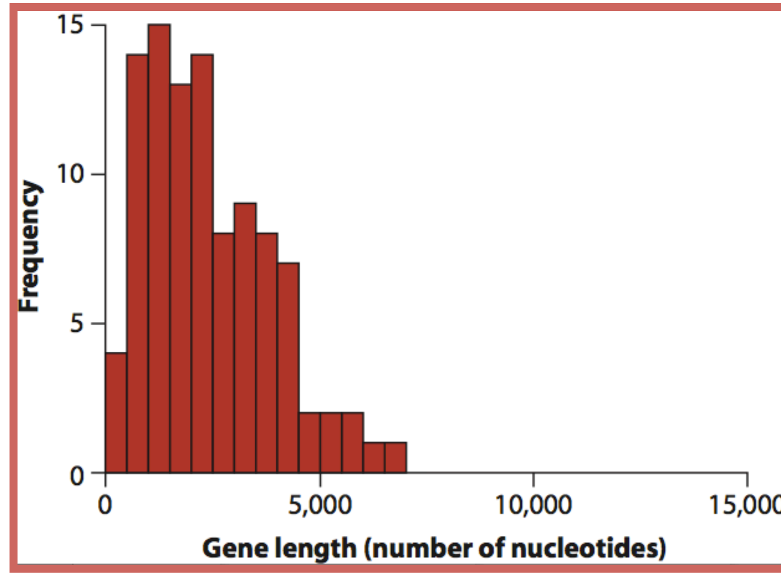
Name	Parameter	Value
Mean	μ	2622.0
S.D.	σ	2036.9



Estimating with uncertainty

Example: The length of human genes. From build 35, there are 20,290 genes. Since we *know* the size of *all* predicted genes, we are in the unique situation of calculating parameters directly from the data instead of inferring their values from samples.



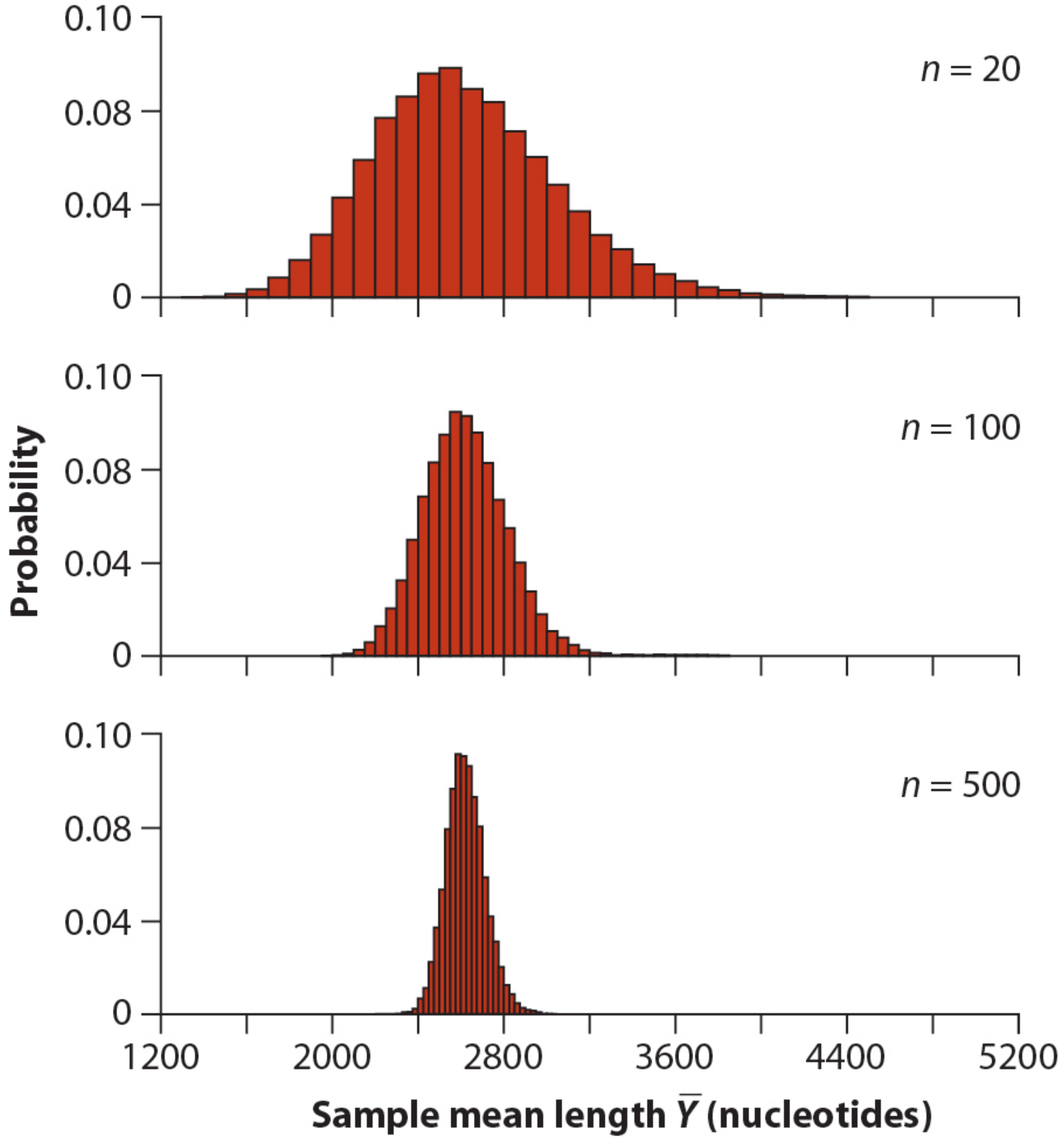


Name	Parameter	Value
Mean	μ	2622.0
S.D.	σ	2036.9

Name	Parameter	Value
Mean	$\bar{X}$	2411.8
S.D.	s	1463.5

n = 100

Estimating with uncertainty



Estimating with uncertainty

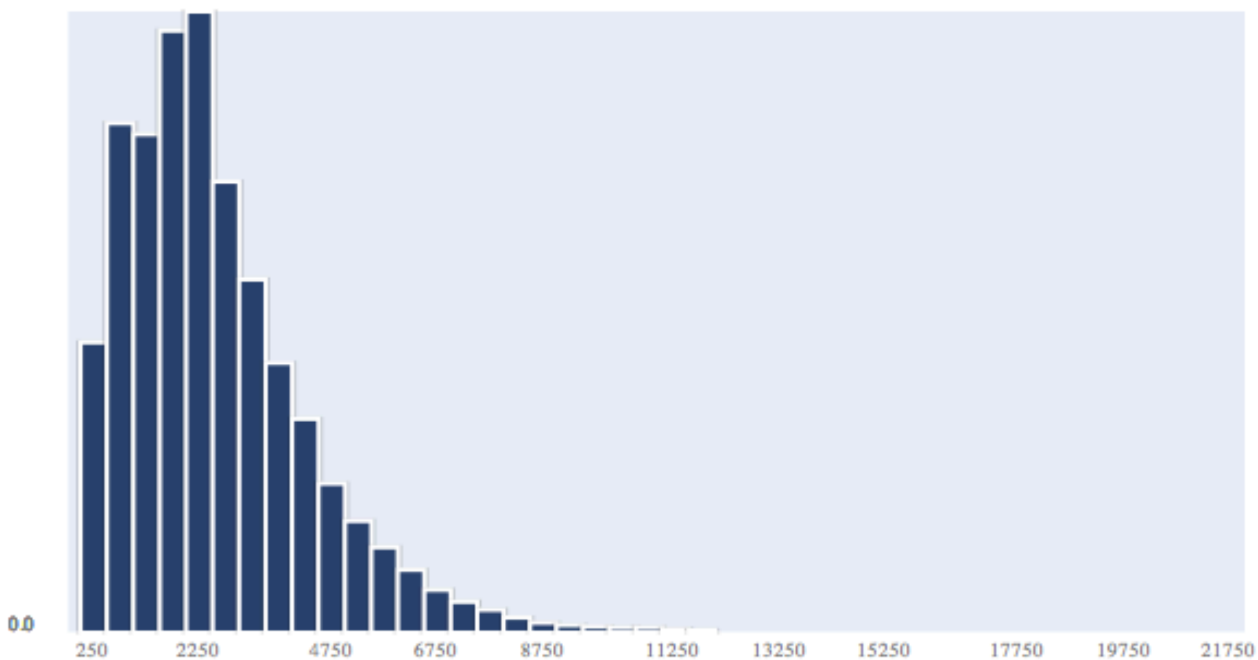
Question: With limited resources, would you rather have:

1. 1000 samples, each of 10 individuals

   or

2. 10 samples, each of 1000 individuals?

Gene lengths in human genes (some long genes were excluded from consideration to make it easier to make these graphs):



Population size: 20287
Population mean length: $\mu = 2614.36$
Population standard deviation of length: $\sigma = 1897.30$

http://phylo.bio.ku.edu/biostats/geneLenDemo.html

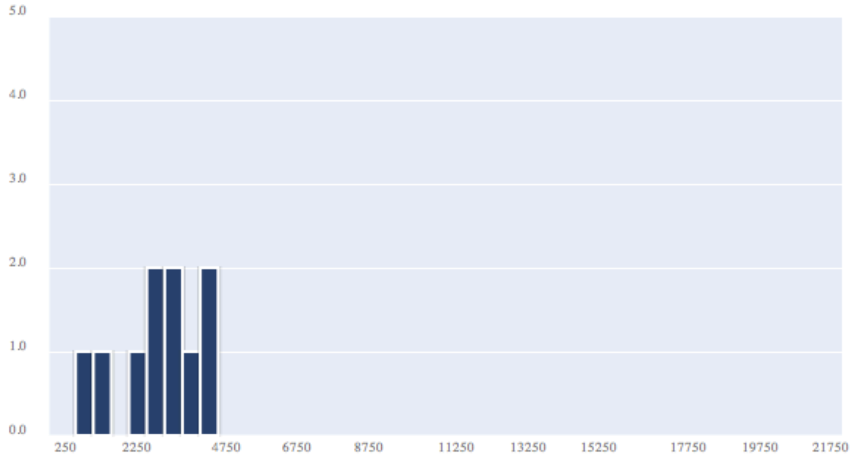
Question: With limited resources, would you rather have:

1. 1000 samples, each of 10 individuals **or** 2. 10 samples, each of 1000 individuals?

1000 samples of 10 individuals

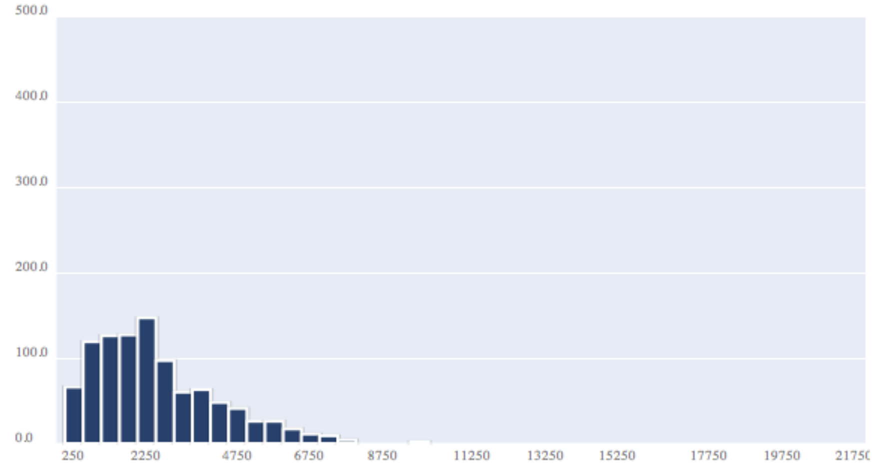
10 samples of 1000 individuals

Gene lengths in the last random sample:

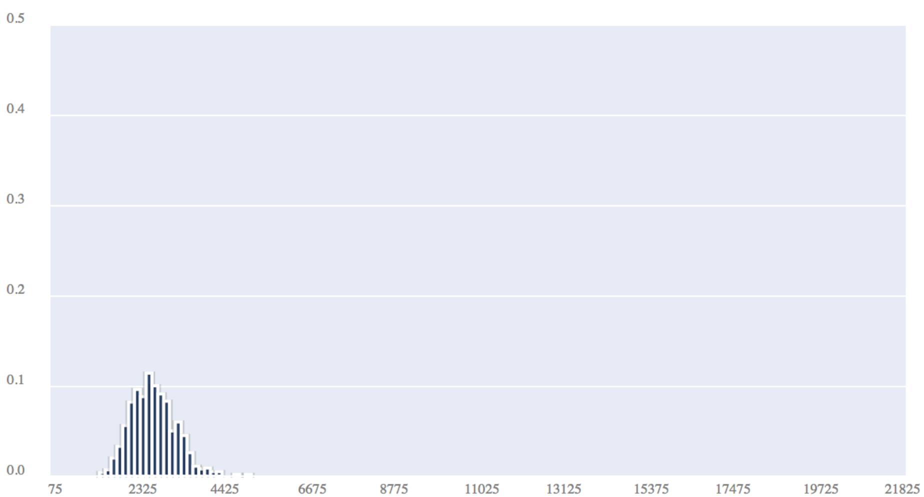


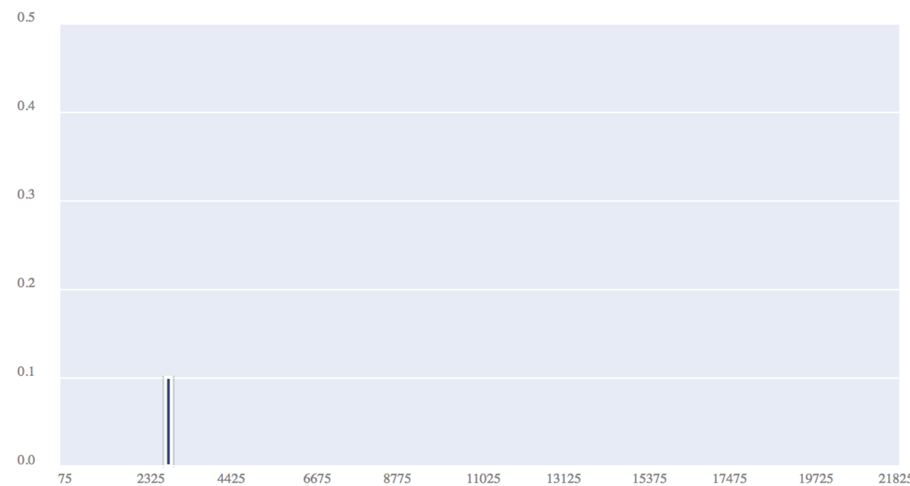
Mean length in sample: $\bar{Y}$ = 2812.90
Sample standard deviation: s = 1193.99

Gene lengths in the last random sample:

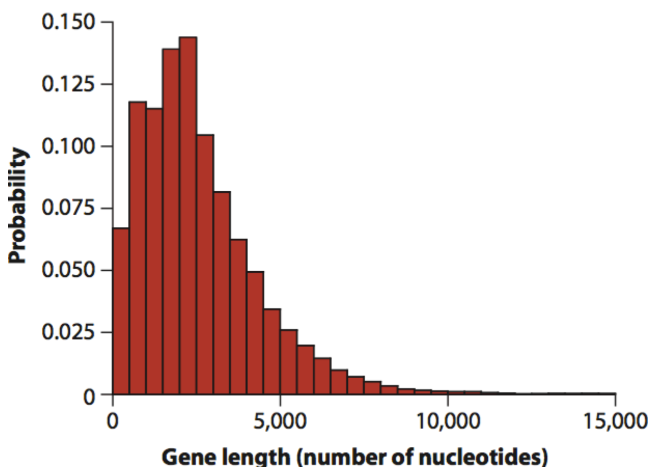


Mean length in sample: $\bar{Y}$ = 2633.77
Sample standard deviation: s = 1988.42

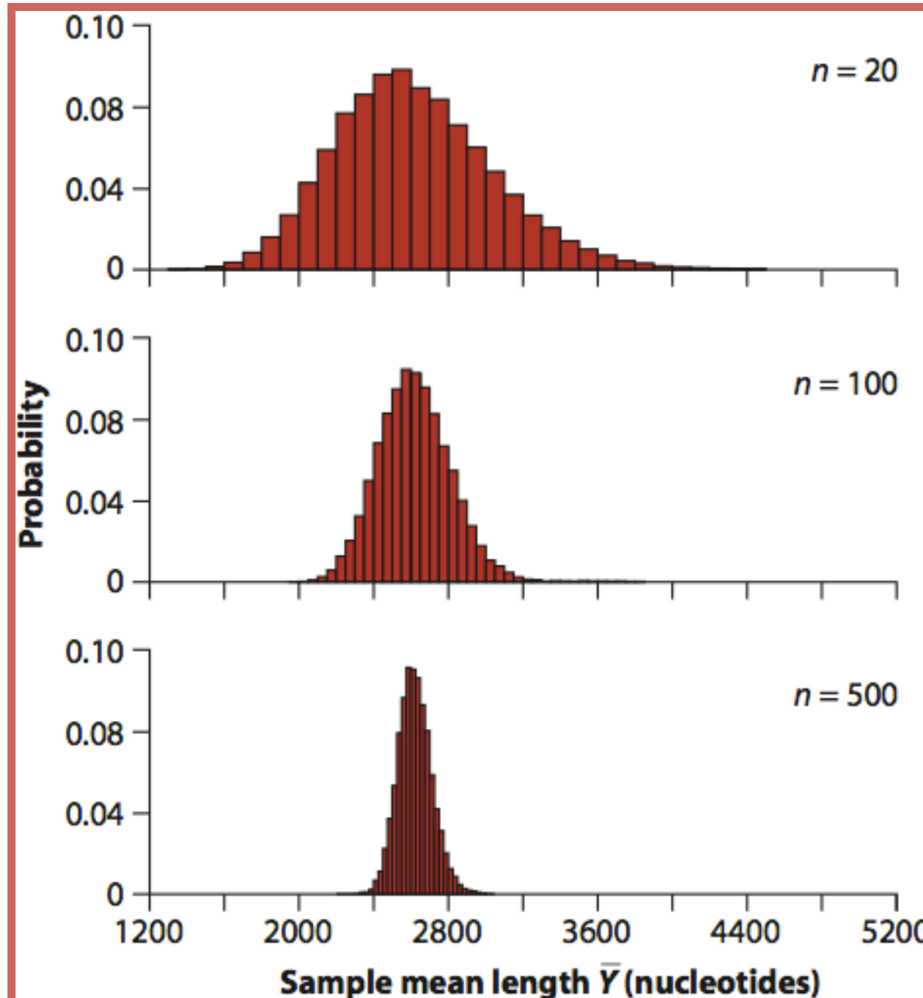




<u>Example:</u> The length of human genes (page 84). From build 35, there are 20,290 genes. Since we *know* the size of *all* predicted genes, we are in the unique situation of calculating parameters directly from the data instead of inferring their values from samples.



Name	Parameter	Value
Mean	μ	**2622.0**
S.D.	σ	**2036.9**

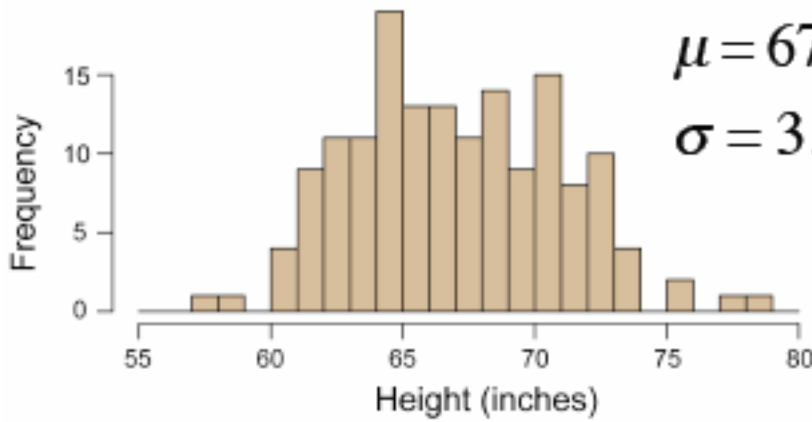


<u>Standard Error</u> <u>(of the mean)</u>**<u>:</u>** the standard deviation of the sampling distribution of some statistic

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

*Some statistic could be: mean, difference between two means, correlation coefficient etc

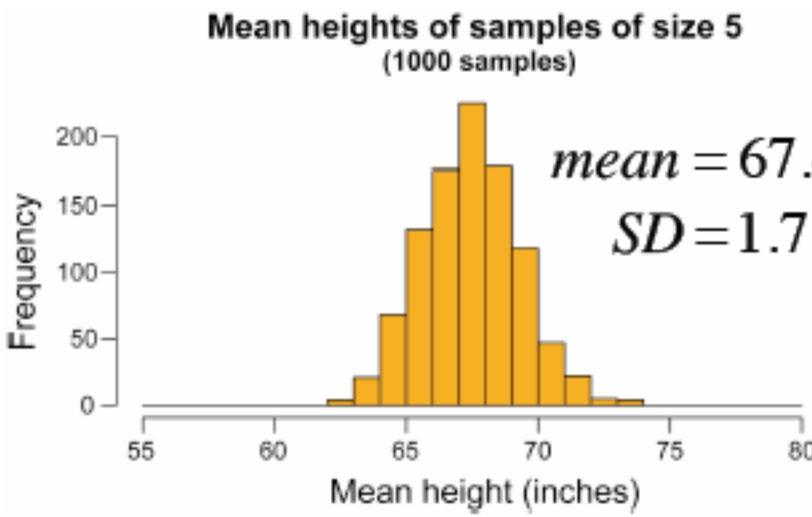
Measuring Uncertainty



$$\mu = 67.4$$
$$\sigma = 3.9$$

$$\mu_{\bar{Y}} = \mu = 67.4$$

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.9}{\sqrt{5}} = 1.7$$



$$mean = 67.4$$
$$SD = 1.7$$

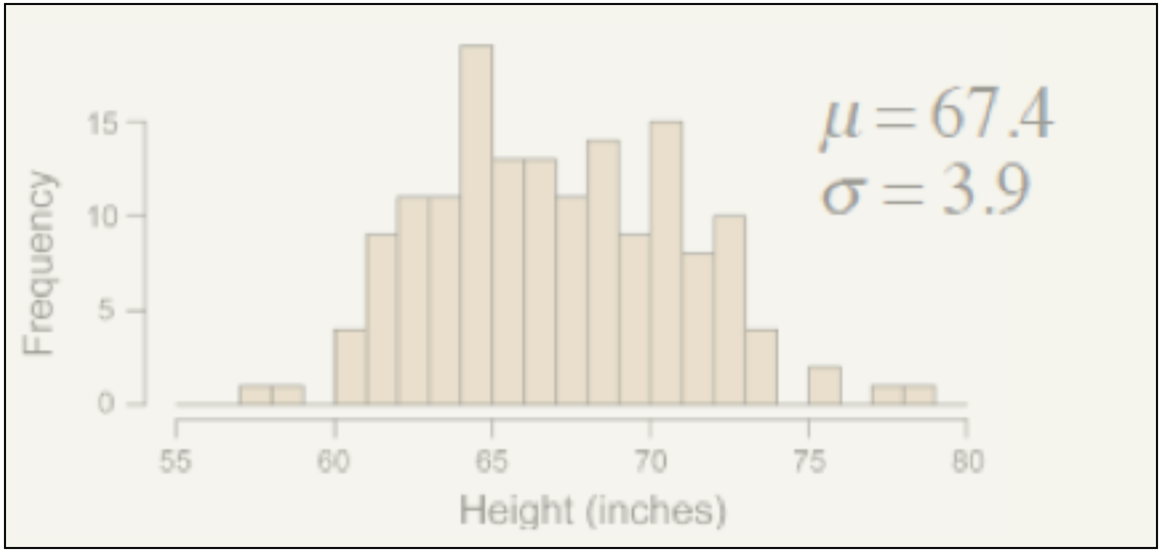
The math works!

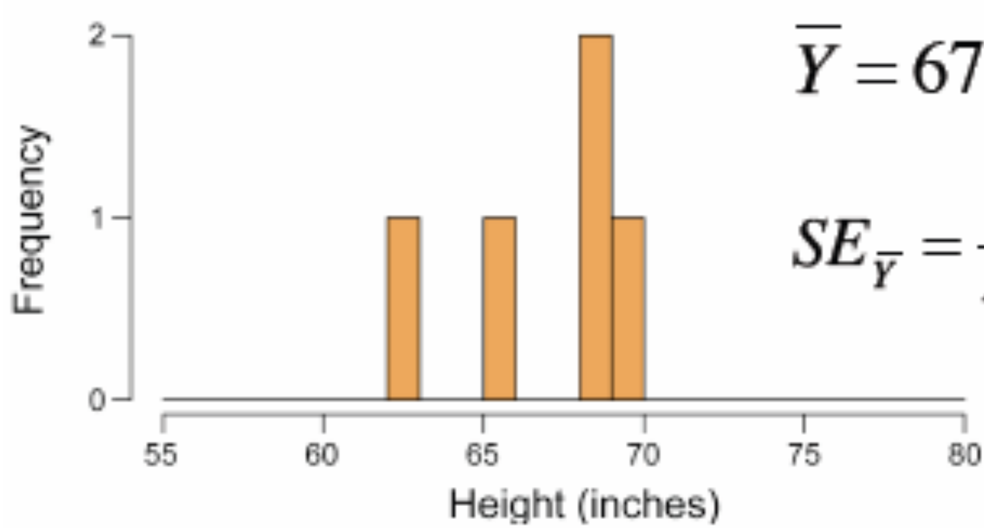
The problem is,
we rarely know σ.

Estimate of the standard error (of the mean):

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

Measuring Uncertainty





$$\overline{Y} = 67.1 \qquad s = 3.1$$

$$SE_{\overline{Y}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{5}} = 1.4$$

We use this as an estimate of $\sigma_{\overline{Y}}$