

ECE 20875

Python for Data Science

Chris Brinton and David Inouye

estimation and sampling

why sample?

- Most analysis problems do not let you work with the whole **population**, e.g.,
 - *How many engines have a defect?* Cannot take apart every engine to find out
 - *What is the average height of people in Indiana?* Would be nearly impossible to measure every person in the state
 - *What is the difference in commute times between people in Indianapolis and people in Chicago?* Again, cannot ask everyone in both cities
- We are often left trying to learn facts about a population by only studying a subset of that population, i.e., a **sample**

how to sample?

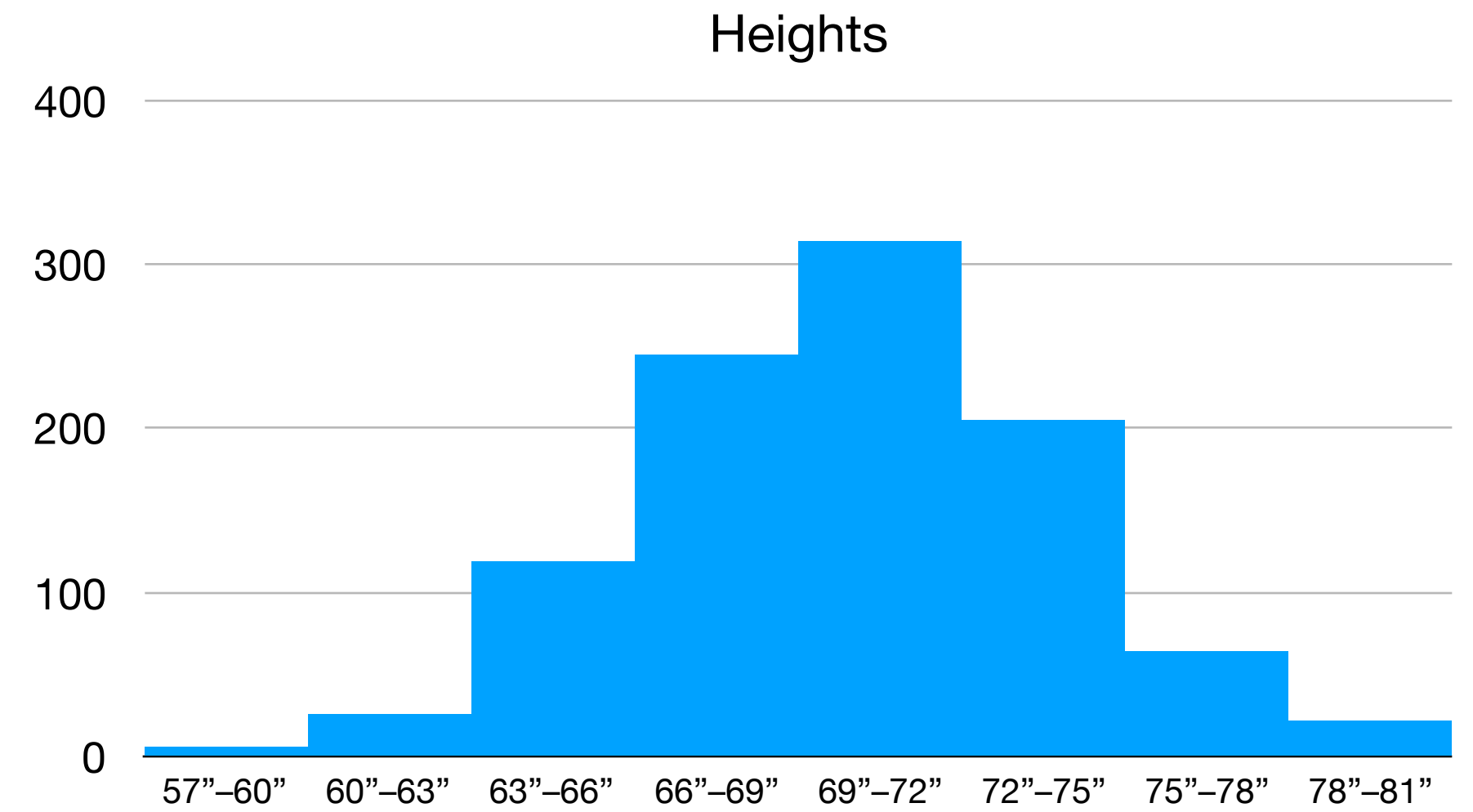
- Many strategies. Some common techniques:
 - **Simple Random Sampling (SRS):** Select S elements from a population P so that each element of P is equally likely to appear in S . **Easiest to analyze**, but **can make it hard to represent rare samples** (rare groups won't show up).
 - **Stratified Sampling:** Subdivide population P into subgroups P_1, P_2 , etc. where each subgroup represents a distinct attribute (e.g., breaking a population up by cities). Do SRS within the subgroups, and combine the result. **Ensures representation of each subgroup**, but **can be hard to set up**.
 - **Cluster Sampling:** Group population into random clusters (not specific subgroups like in stratified sampling). Select clusters at random, add all elements from selected clusters to sample. **Easier to conduct** than SRS, but **adds more variability**.
- We will focus mainly on SRS in this course

statistic vs parameter

- We differentiate between attributes of the population and the sample
- Numbers which summarize a population are called **parameters**
 - Population mean (μ), variance (σ^2), median, etc.
- Numbers which summarize a sample are called **statistics**
 - Sample mean (\bar{x}), variance (s^2), median, etc.
 - The statistics are not guaranteed to be close to the parameters (why?)
- **Estimation** is the problem of making educated guesses for parameters given sample data
 - Key question: How close is our estimate to the true parameter?

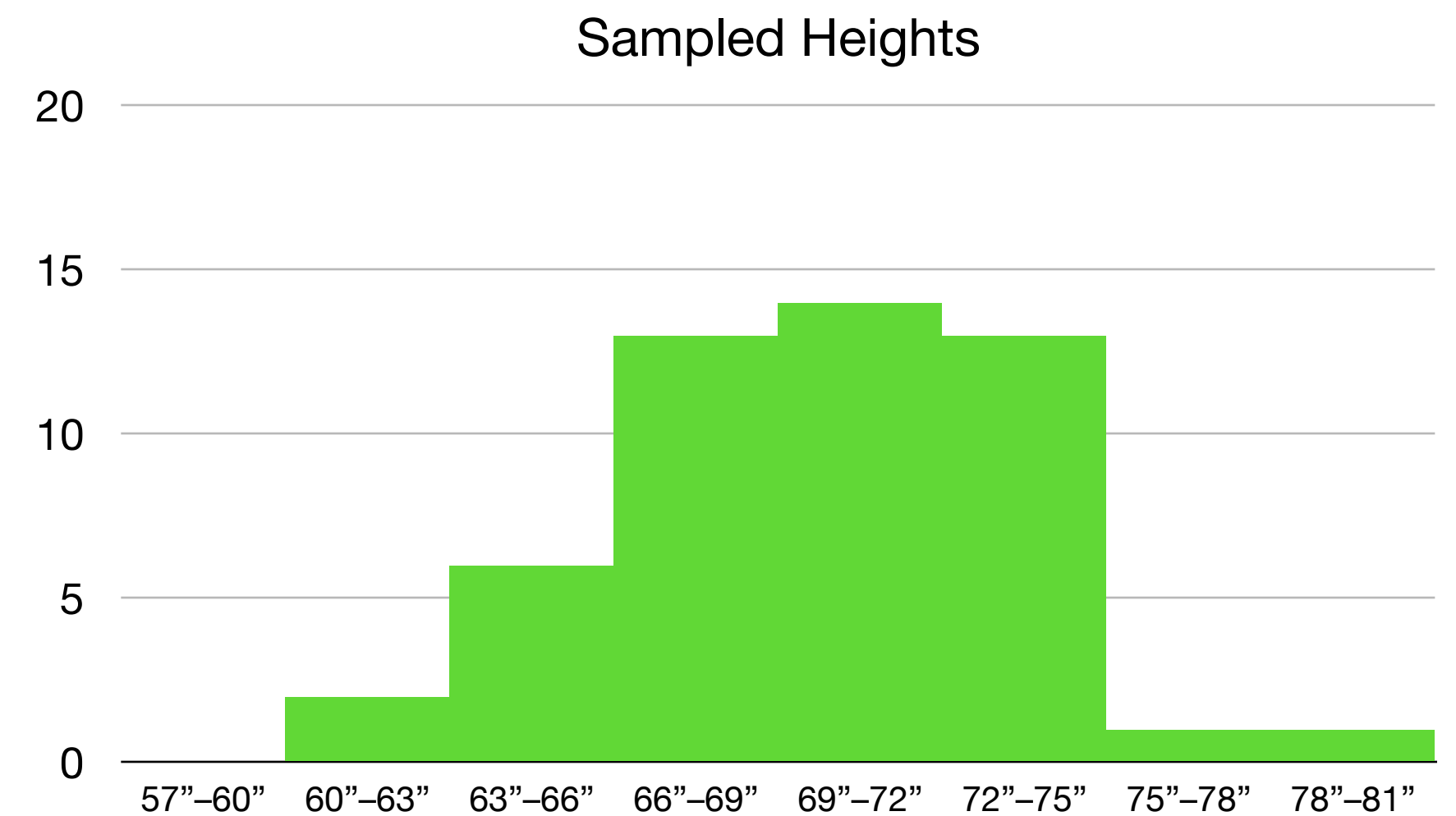
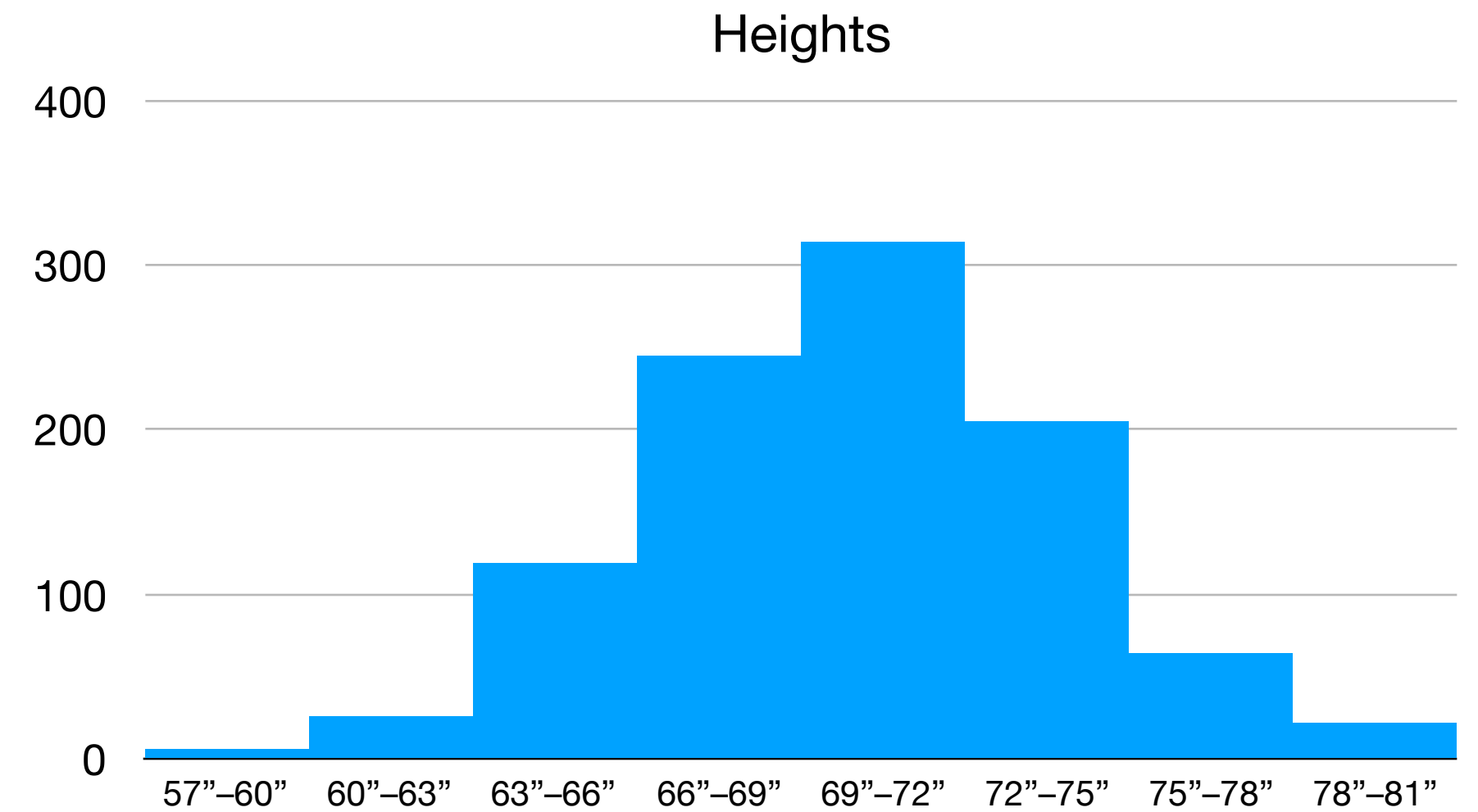
sampling

- Let's consider a population of 1000 people whose heights we have measured



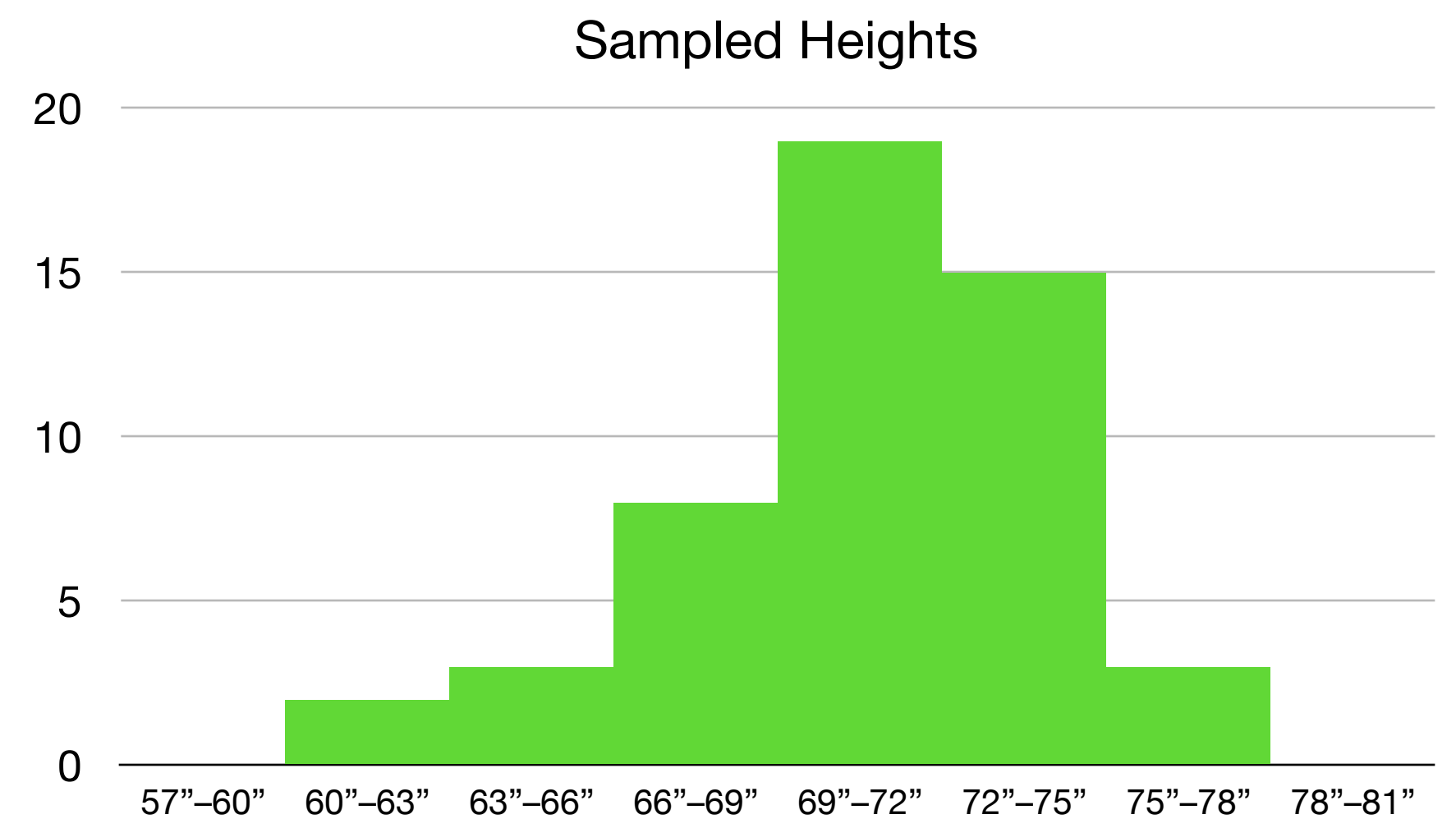
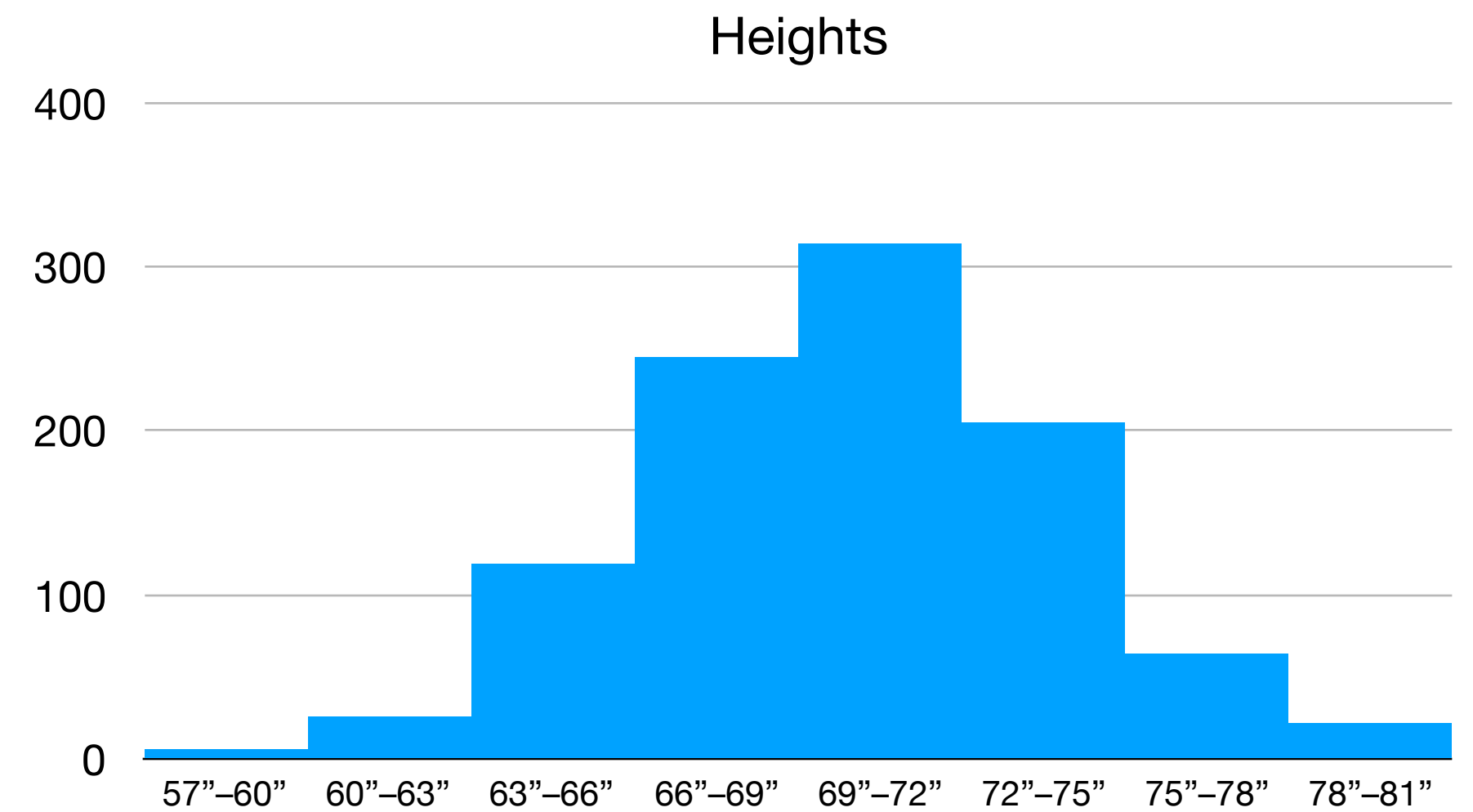
sampling

- Let's consider a population of 1000 people whose heights we have measured
- What if we sample $n = 50$ of them at random?
- Don't get exactly the same distribution



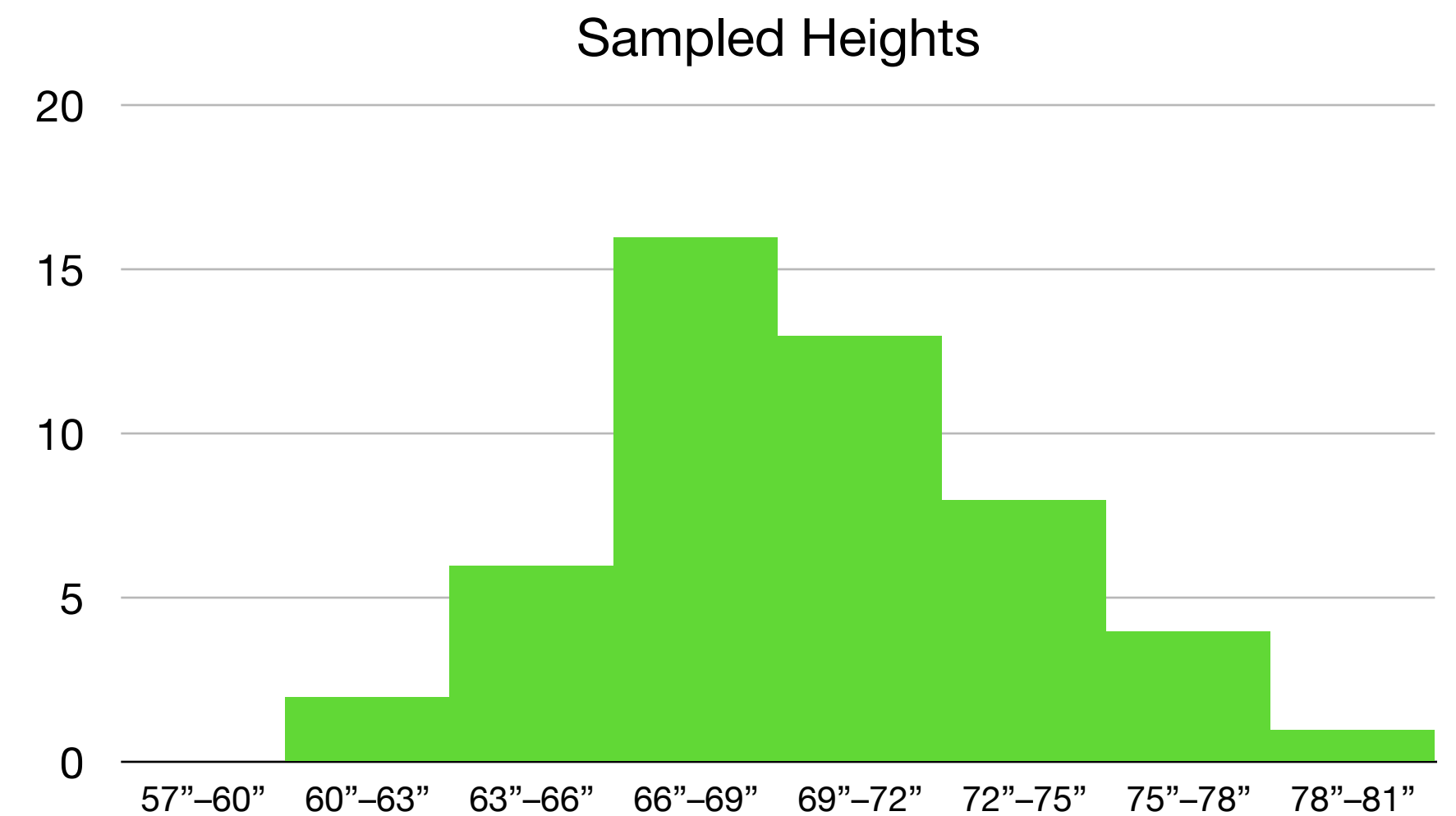
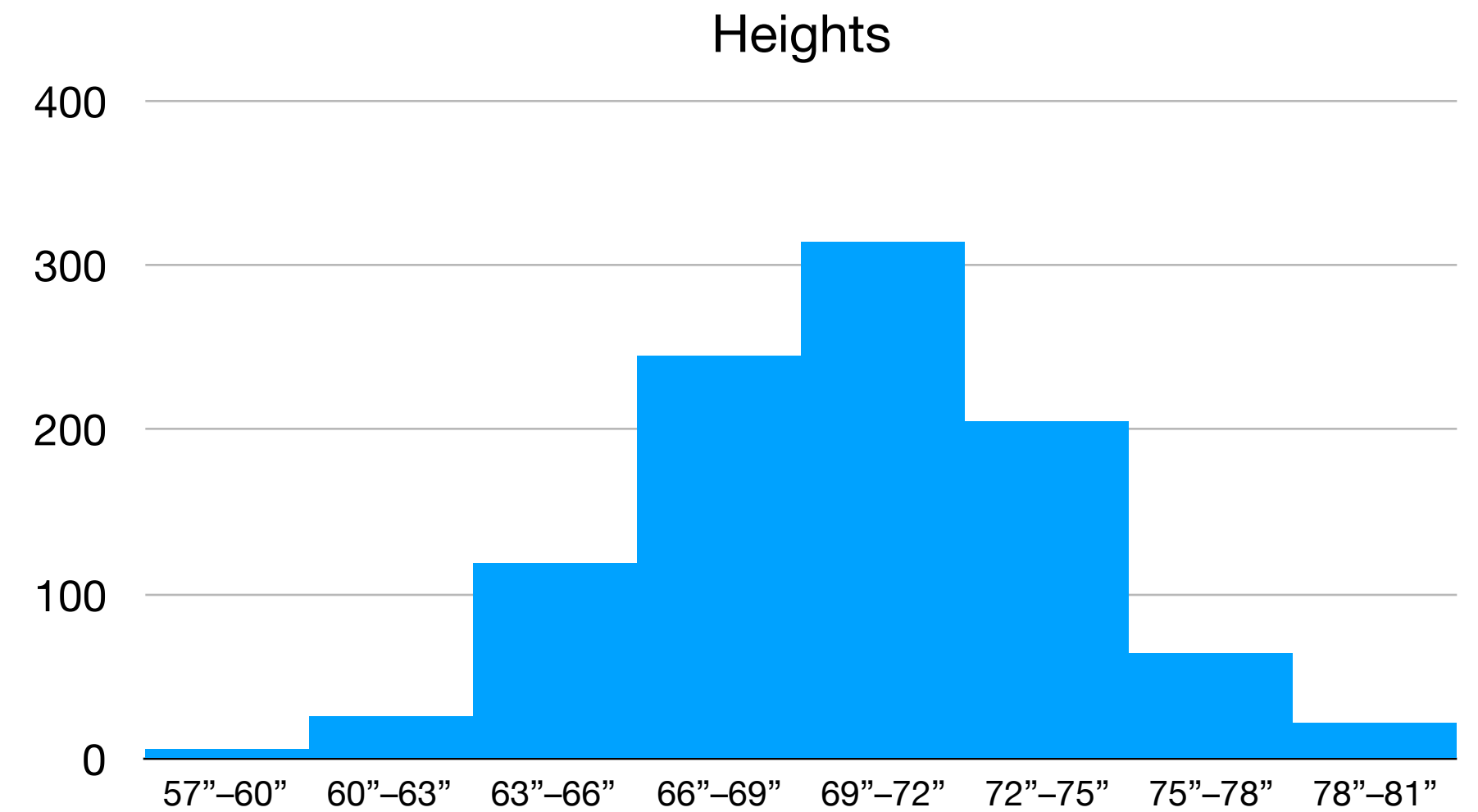
sampling

- Let's consider a population of 1000 people whose heights we have measured
- What if we sample $n = 50$ of them at random?
 - Don't get exactly the same distribution
- What if we sample again?



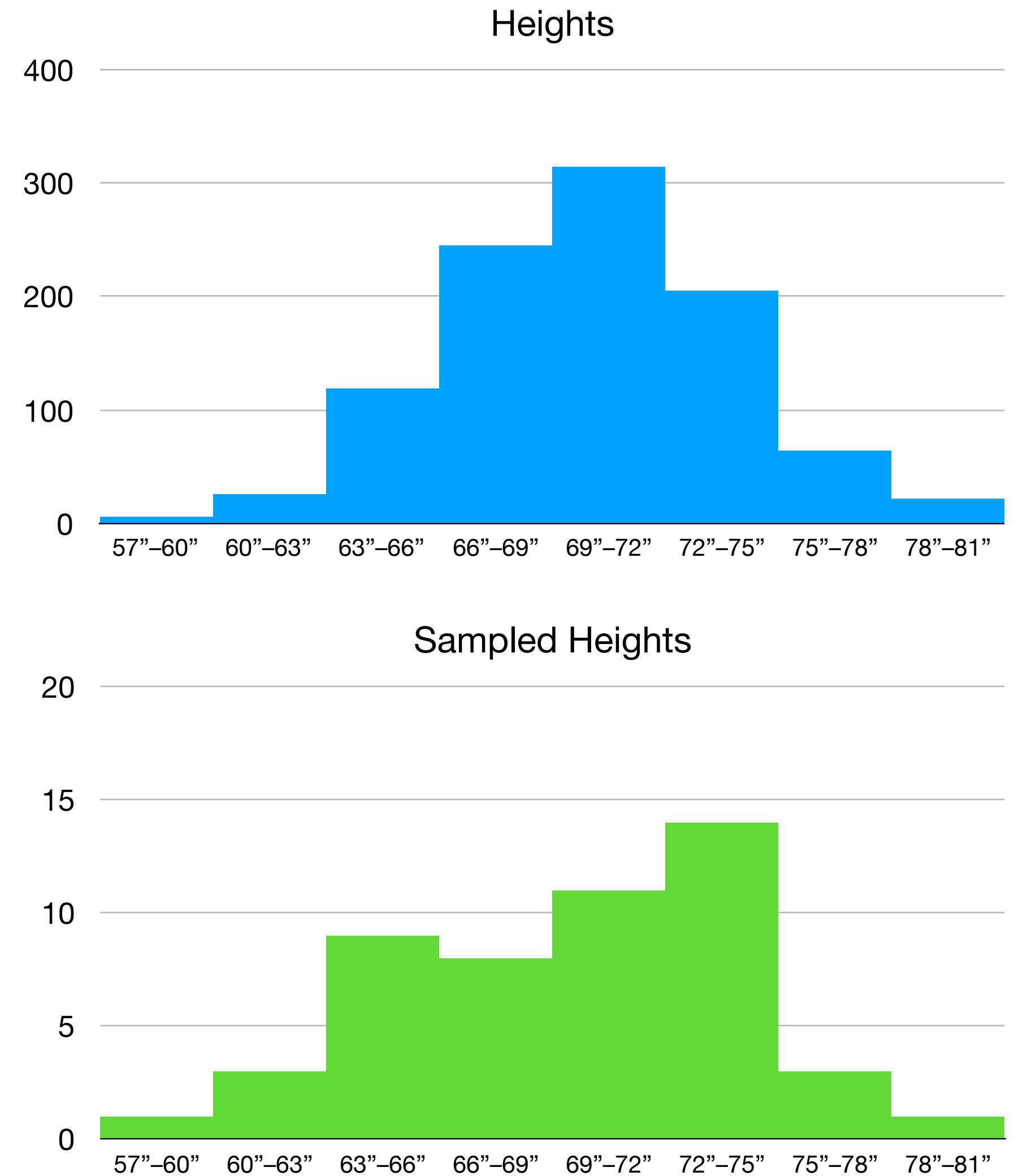
sampling

- Let's consider a population of 1000 people whose heights we have measured
- What if we sample $n = 50$ of them at random?
 - Don't get exactly the same distribution
- What if we sample again?
- And again?



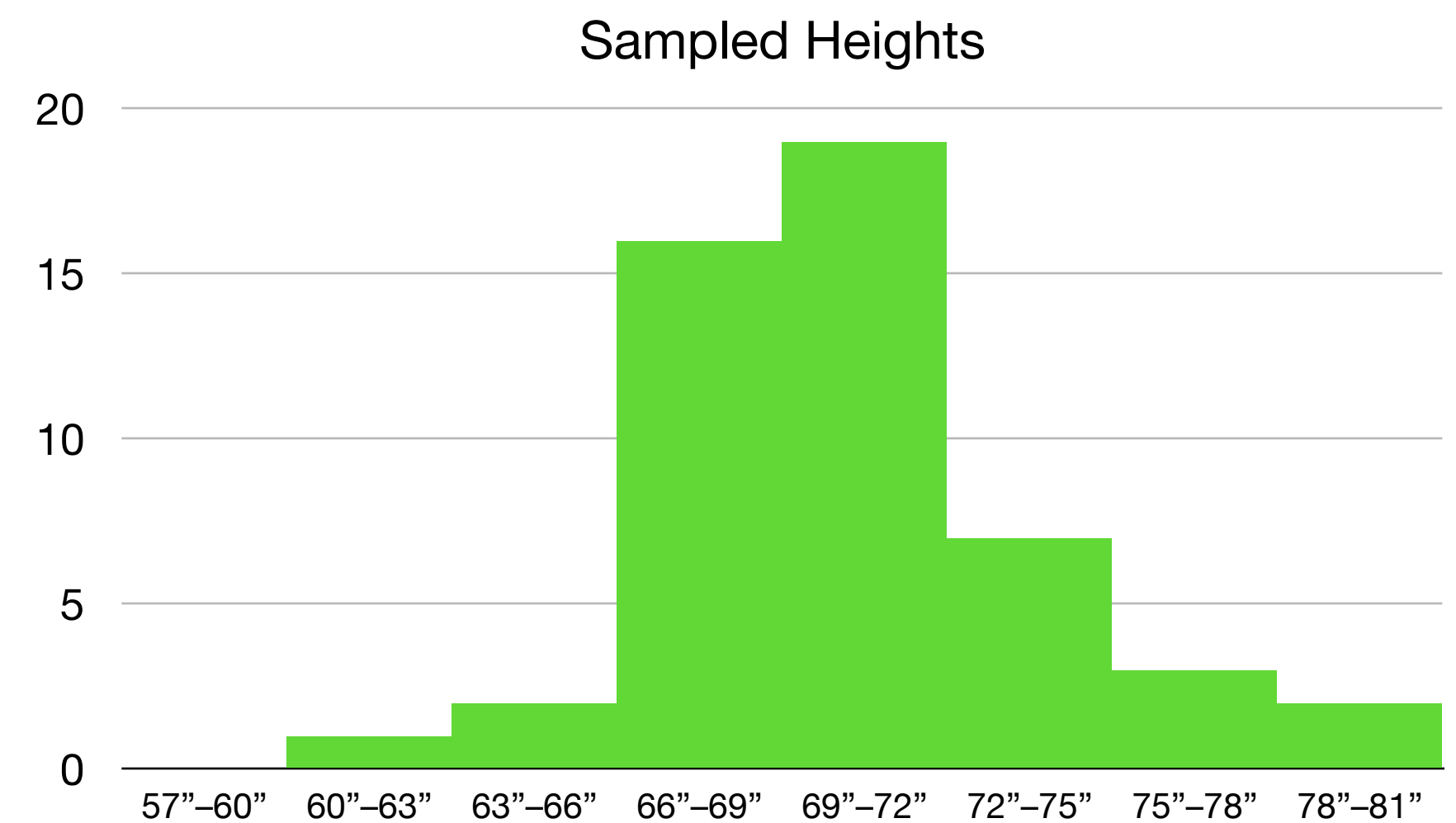
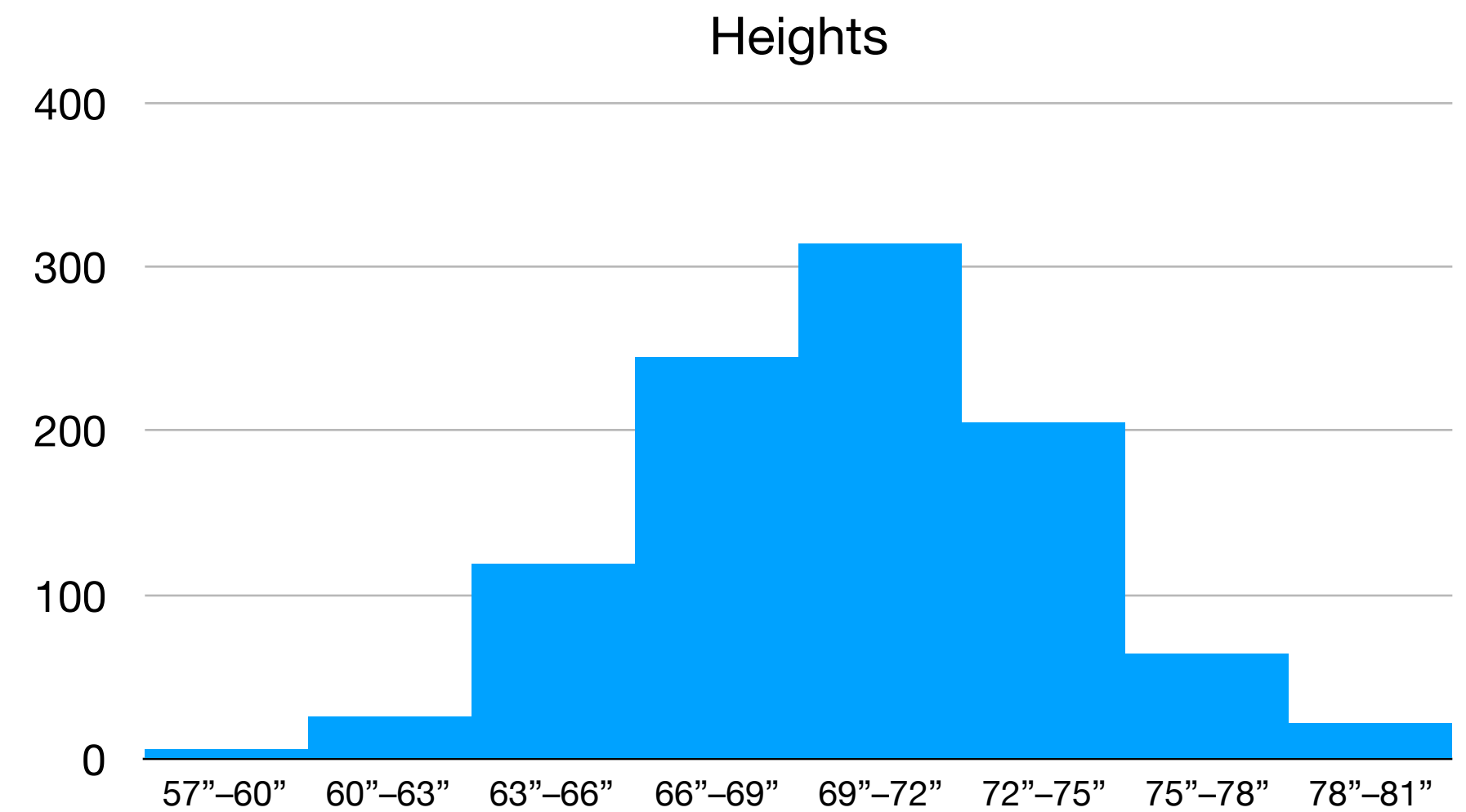
sampling

- Let's consider a population of 1000 people whose heights we have measured
- What if we sample $n = 50$ of them at random?
 - Don't get exactly the same distribution
- What if we sample again?
- And again?



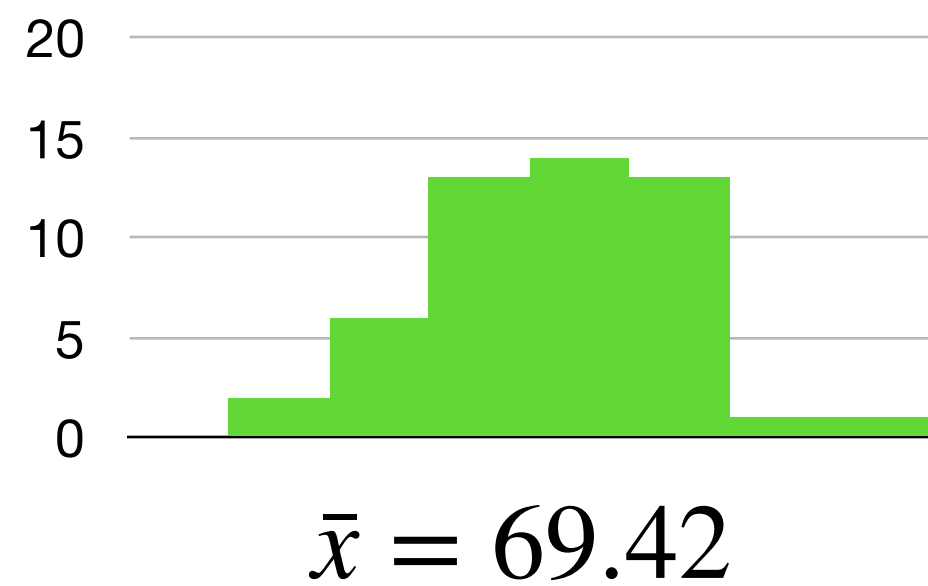
sampling

- Let's consider a population of 1000 people whose heights we have measured
- What if we sample $n = 50$ of them at random?
 - Don't get exactly the same distribution
- What if we sample again?
- And again?



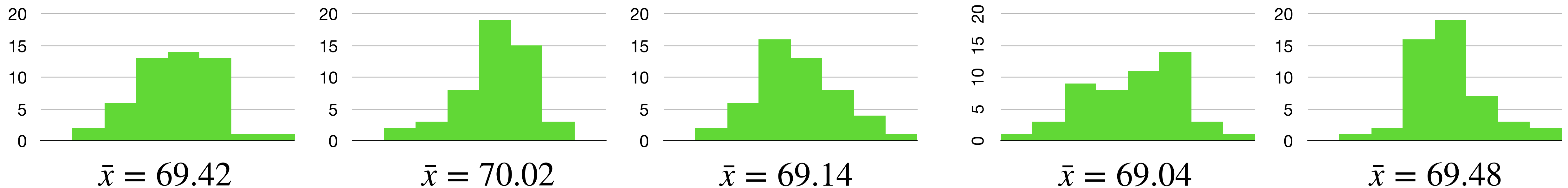
estimate the mean

- What if we want to estimate the mean (μ) of a population?
- Can sample



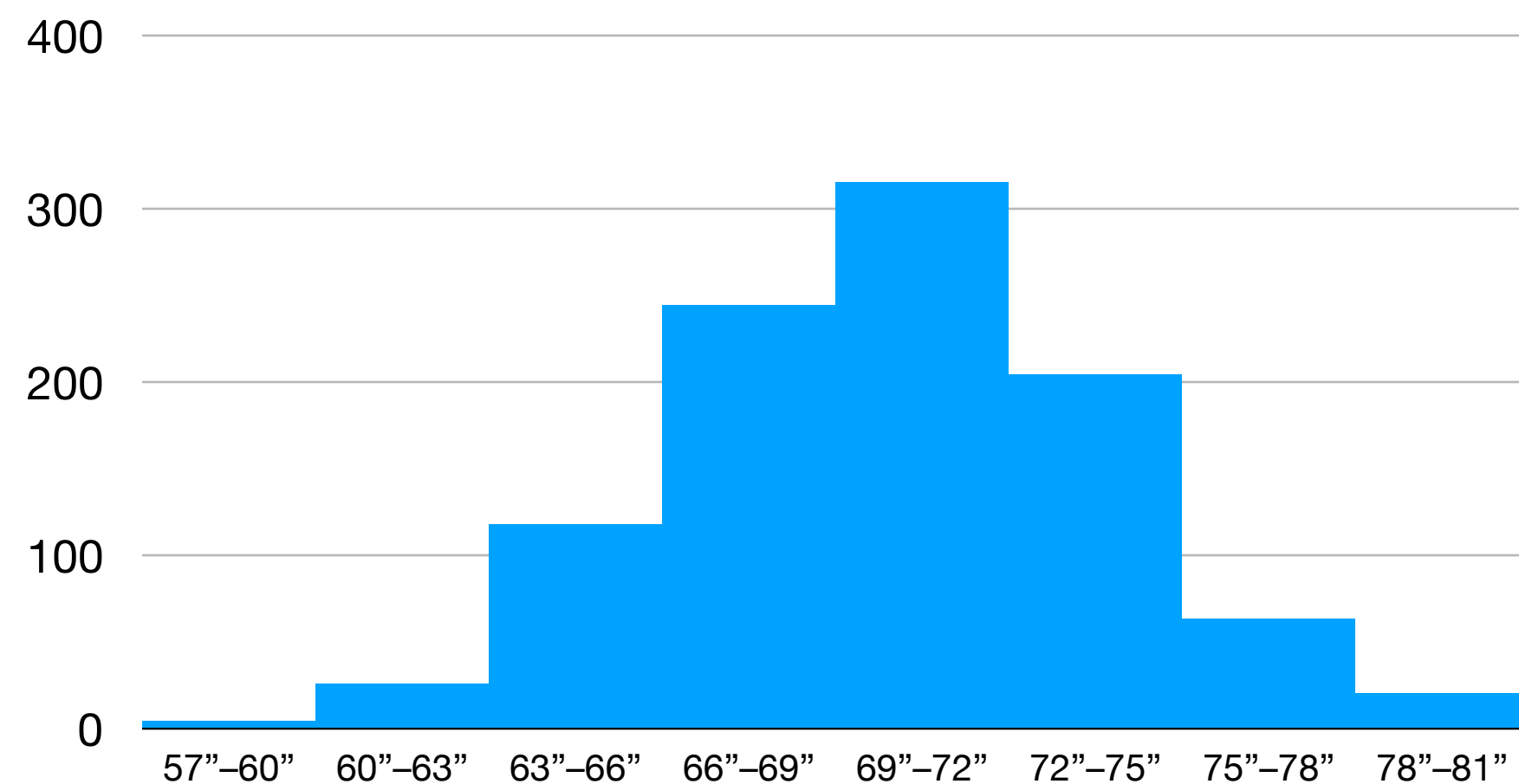
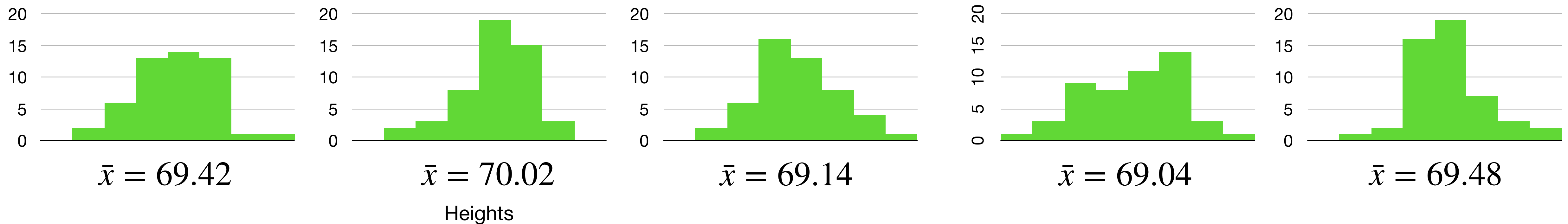
estimate the mean

- What if we want to *estimate* the mean (μ) of a population?
- Can sample, and repeat the experiment



estimate the mean

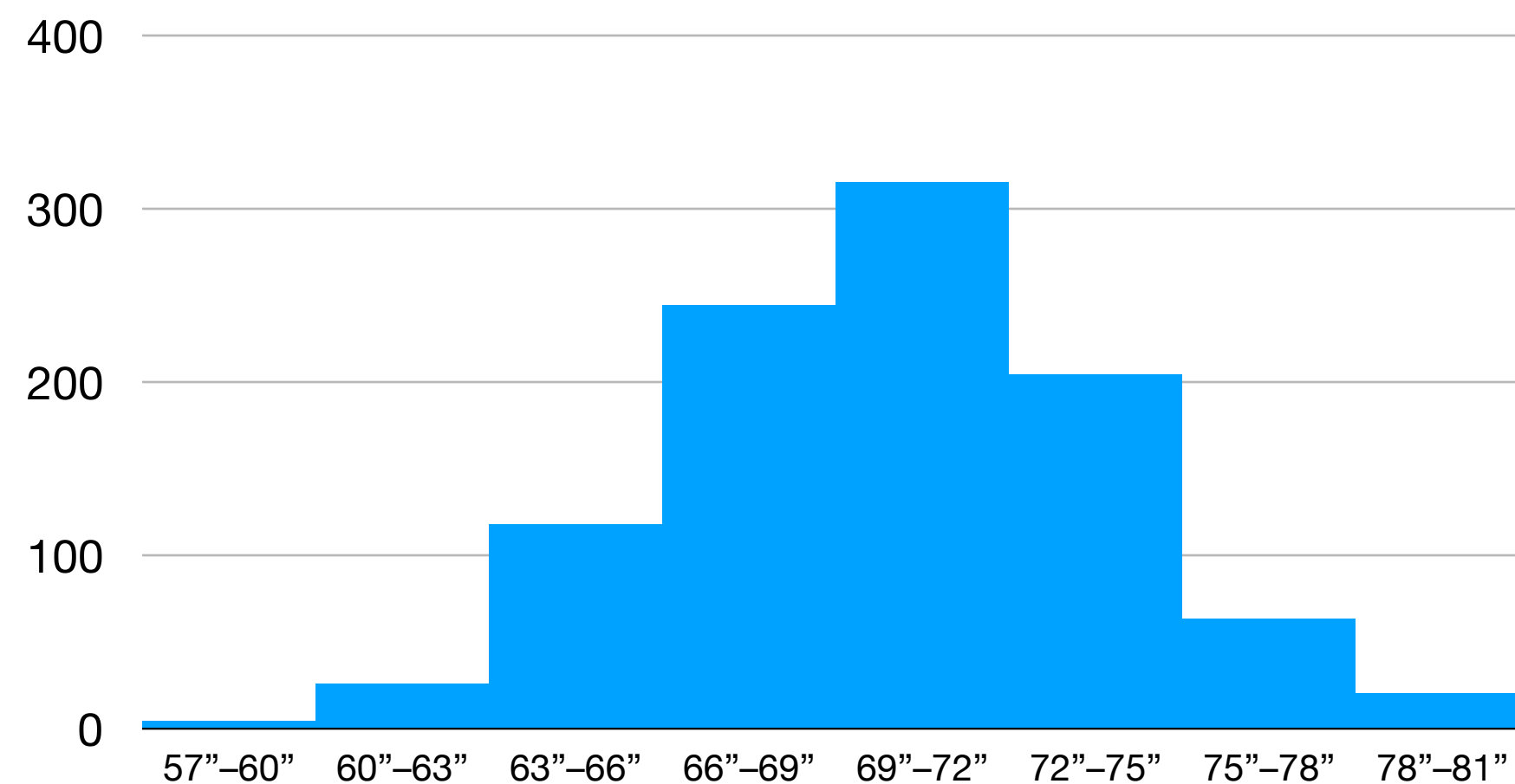
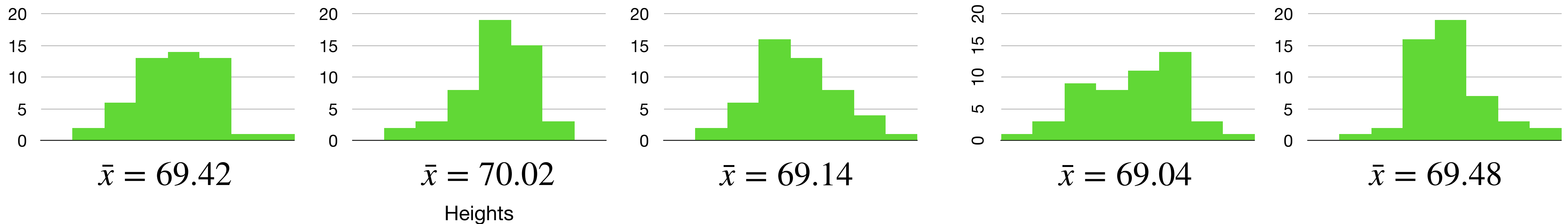
- What if we want to *estimate* the mean (μ) of a population?
- Can sample, and repeat the experiment



$$\mu = 69.436$$

estimate the mean

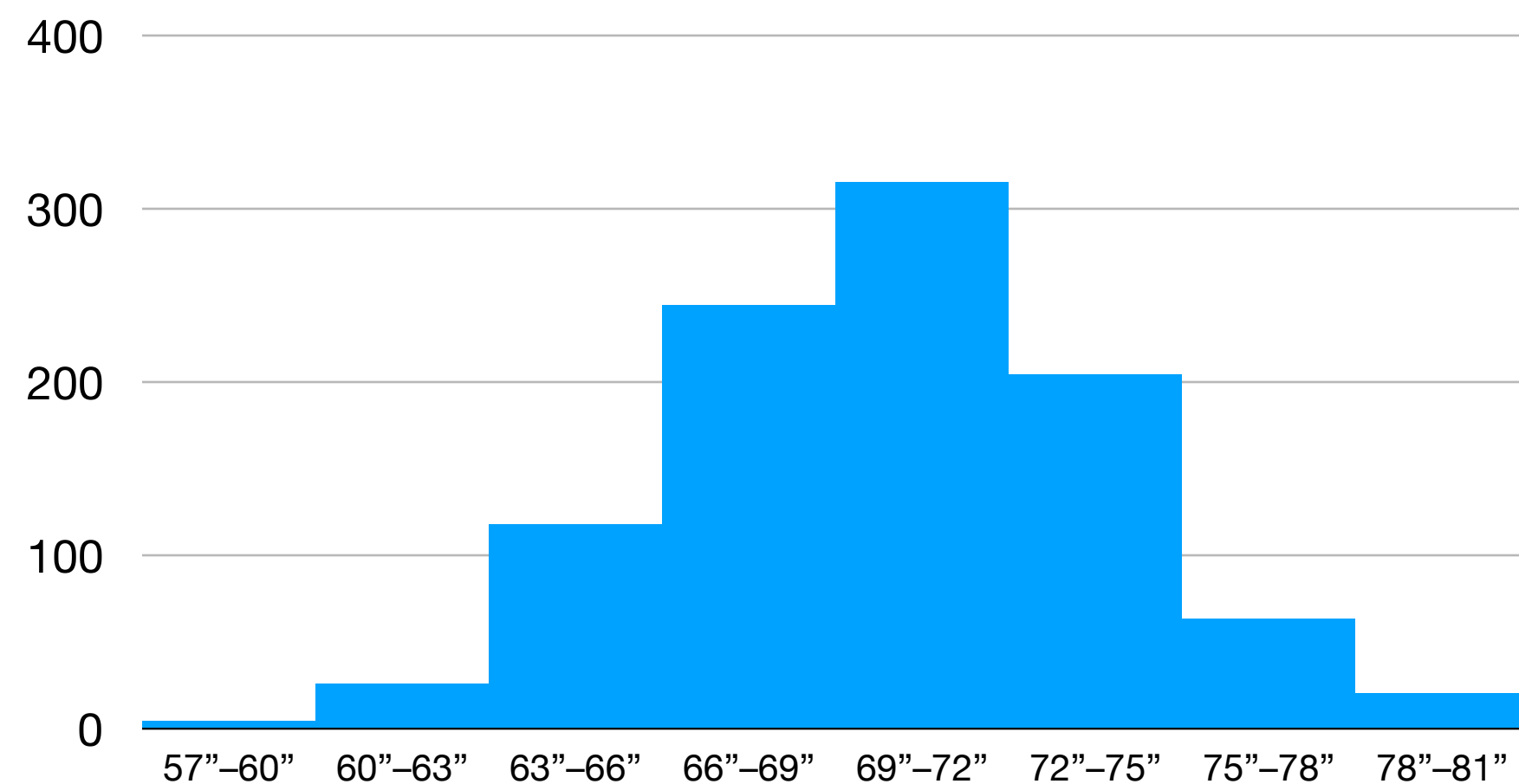
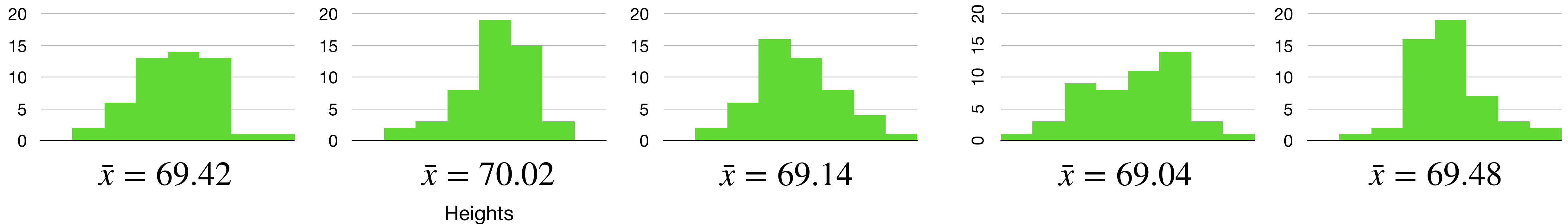
- What if we want to *estimate* the mean (μ) of a population?
- Can sample, and repeat the experiment



- Estimate μ of population using the sample \bar{x}
- How good is this estimate?
- Use the mean squared error (MSE)

how good is our estimate?

- What if we want to *estimate* the mean (μ) of a population?
- Can sample, and repeat the experiment



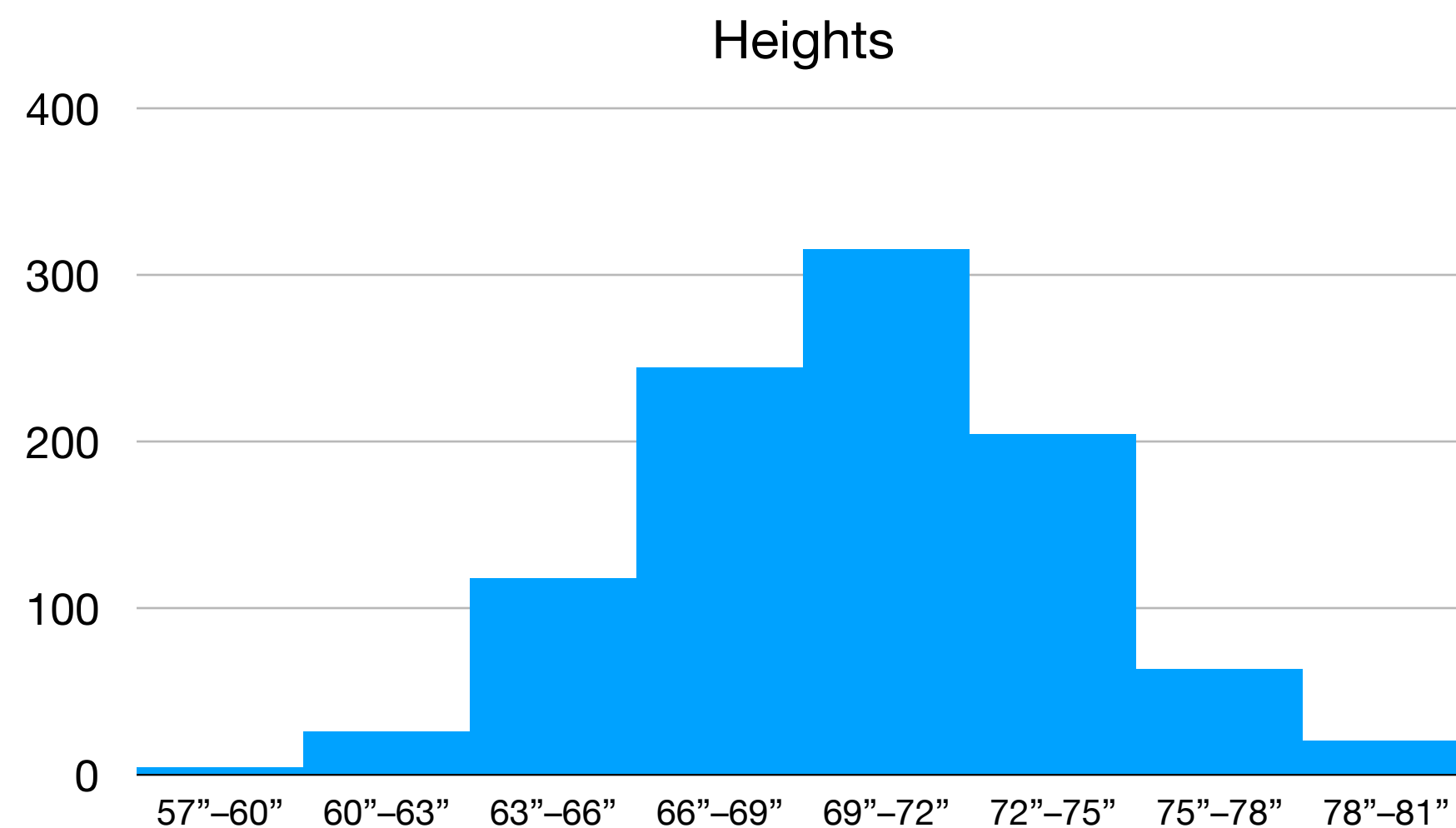
Population $\mu = 69.436$

$$\mathbf{MSE} = \frac{1}{N} \sum_i (\bar{x}_i - \mu)^2$$

MSE of estimates: .118

how good is our estimate?

- What about with smaller samples, e.g., $n = 10$?
- Some \bar{x} 's: [68.6, 67.3, 68.7, 68.9, 69.0, 71.5, 69.8, 67.4, 70.0, 70.8]
- Still pretty good estimates, but not quite as good



Population $\mu = 69.436$

$$\mathbf{MSE} = \frac{1}{N} \sum_i (\bar{x}_i - \mu)^2$$

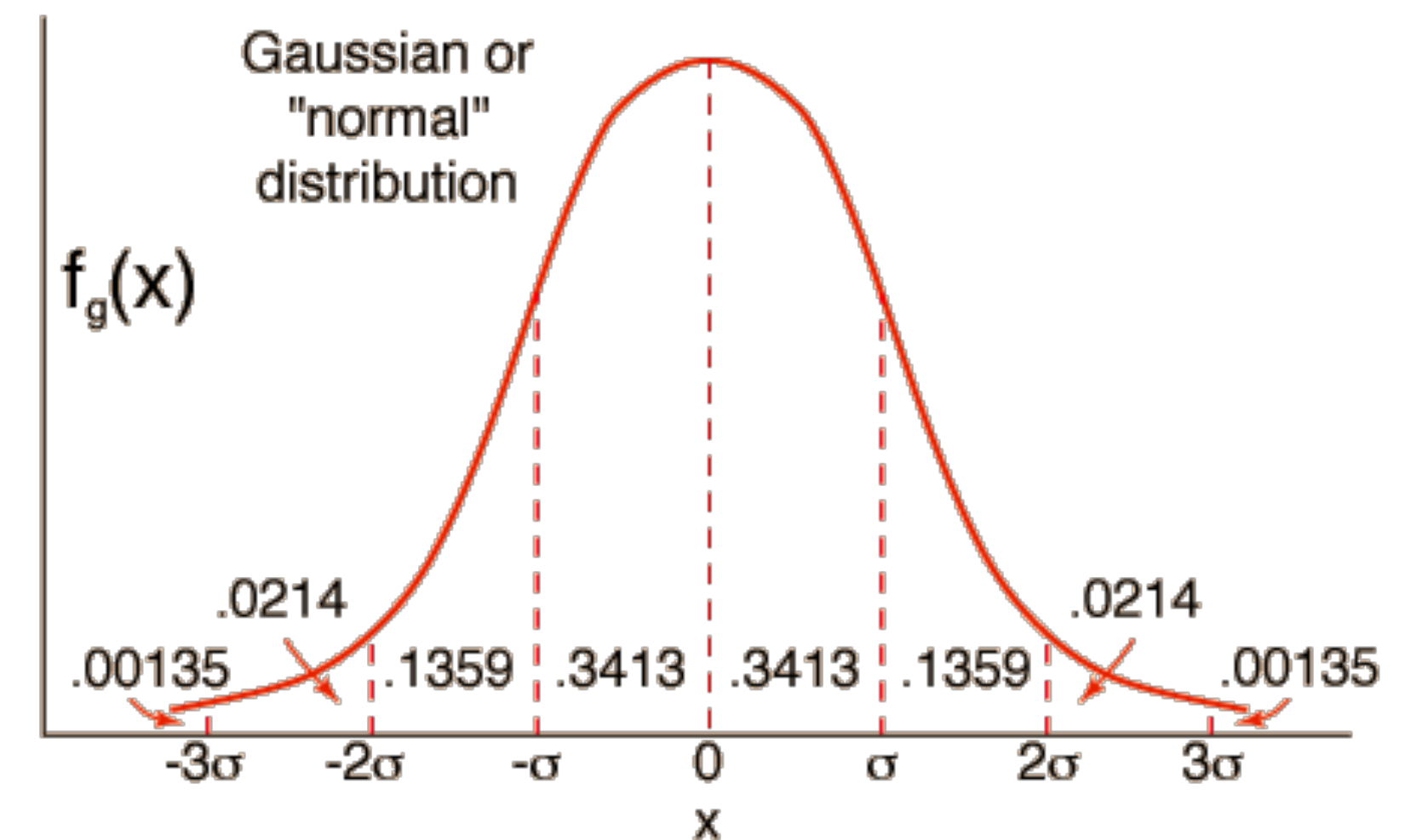
MSE of estimates: 1.70

other useful statistics

- Sample variance (s^2) and standard deviation (s):

$$s^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2, \quad s = \sqrt{s^2}$$

- Quantifies the dispersion of the dataset around the mean
- Why divide by $N - 1$ instead of N ?
 - Consider the case where $N = 1$ (i.e., one sample), what would be the estimate of s^2 ?
 - Only $N - 1$ degrees of freedom when we are using \bar{x} as the estimate of μ
 - For large N this does not matter much though
- Typically, s^2 is a better estimate of σ^2 than s is of σ . There are several tricks to improve the estimates, but we'll usually just use s directly.

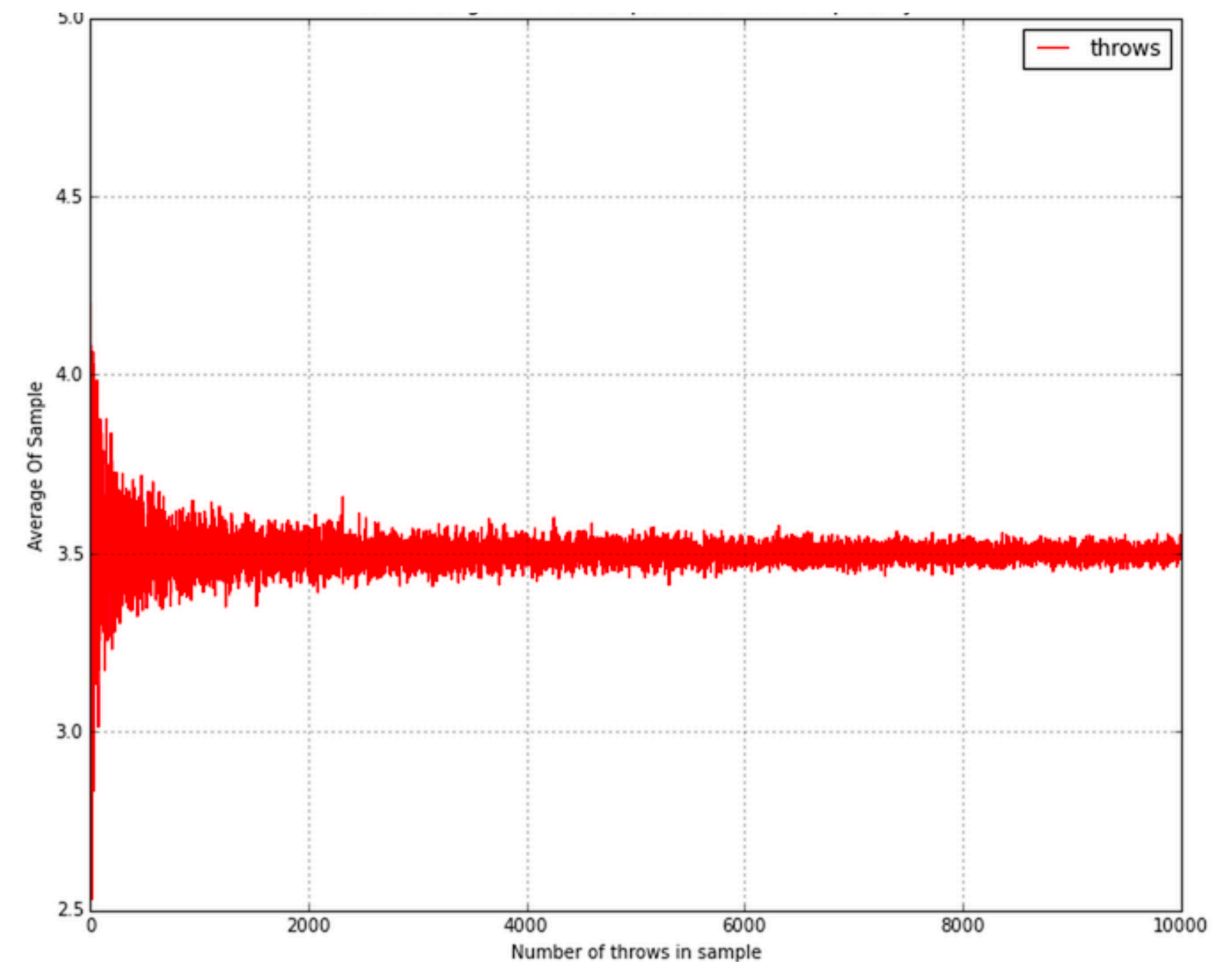


the law of large numbers

- Empirically, we have observed that \bar{x} can be a good estimator for μ
- What we are observing is the **law of large numbers**
 - If X_1, X_2, \dots, X_n are independent and identically distributed (**iid**) random variables, then

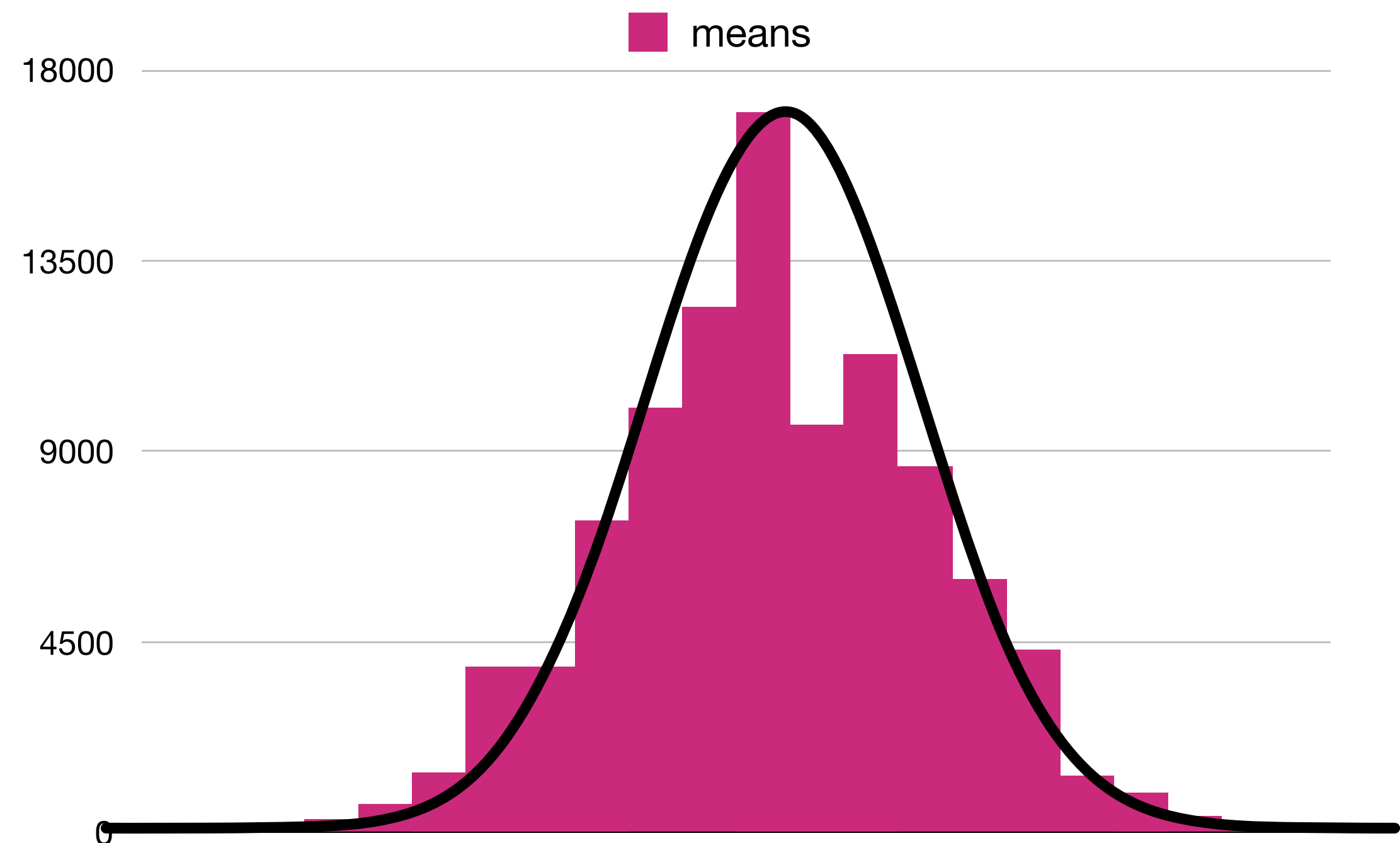
$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu \text{ as } n \rightarrow \infty$$

- In other words, the average of a large number of samples should be close to the population mean
- But any single sample X_i may still be a bad estimate
- What can I say about how good my estimate is?



sampling distribution

- We can also look at the distribution of a sample statistic, e.g., the mean \bar{x}
- This is called a **sampling distribution**
 - View the statistic itself as a random variable, and take samples of this variable by running experiments
- Sampling distribution of the sample mean shown on the right
 - It appears to be normally distributed!

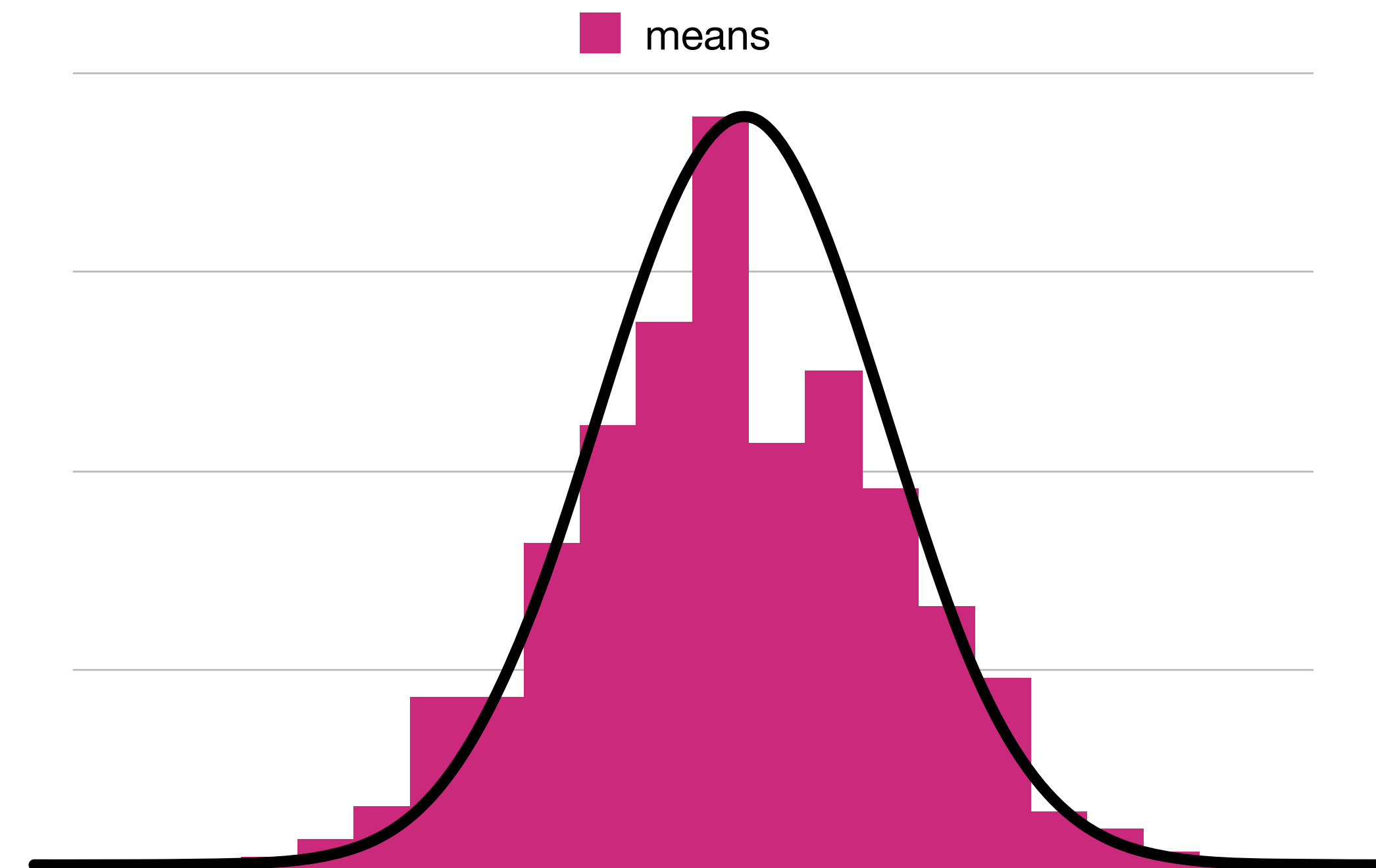


Each data point is the \bar{x} of one experiment

- Average of \bar{x} 's = 69.437
- Standard deviation of \bar{x} 's = 1.17

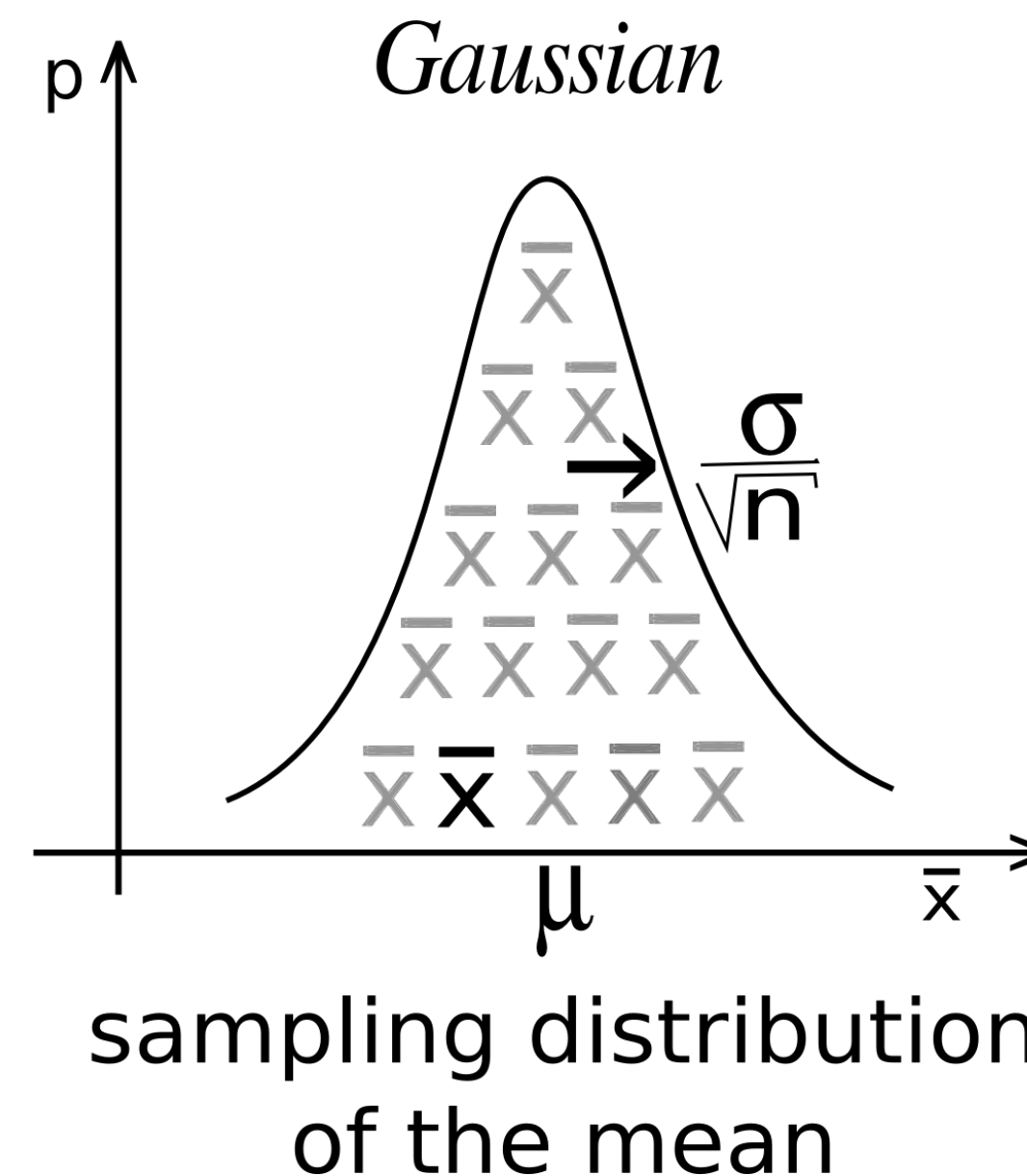
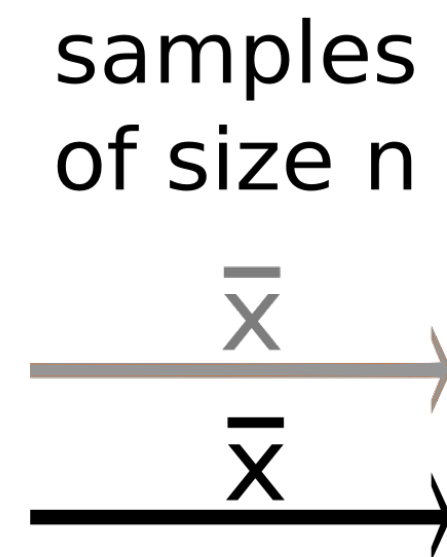
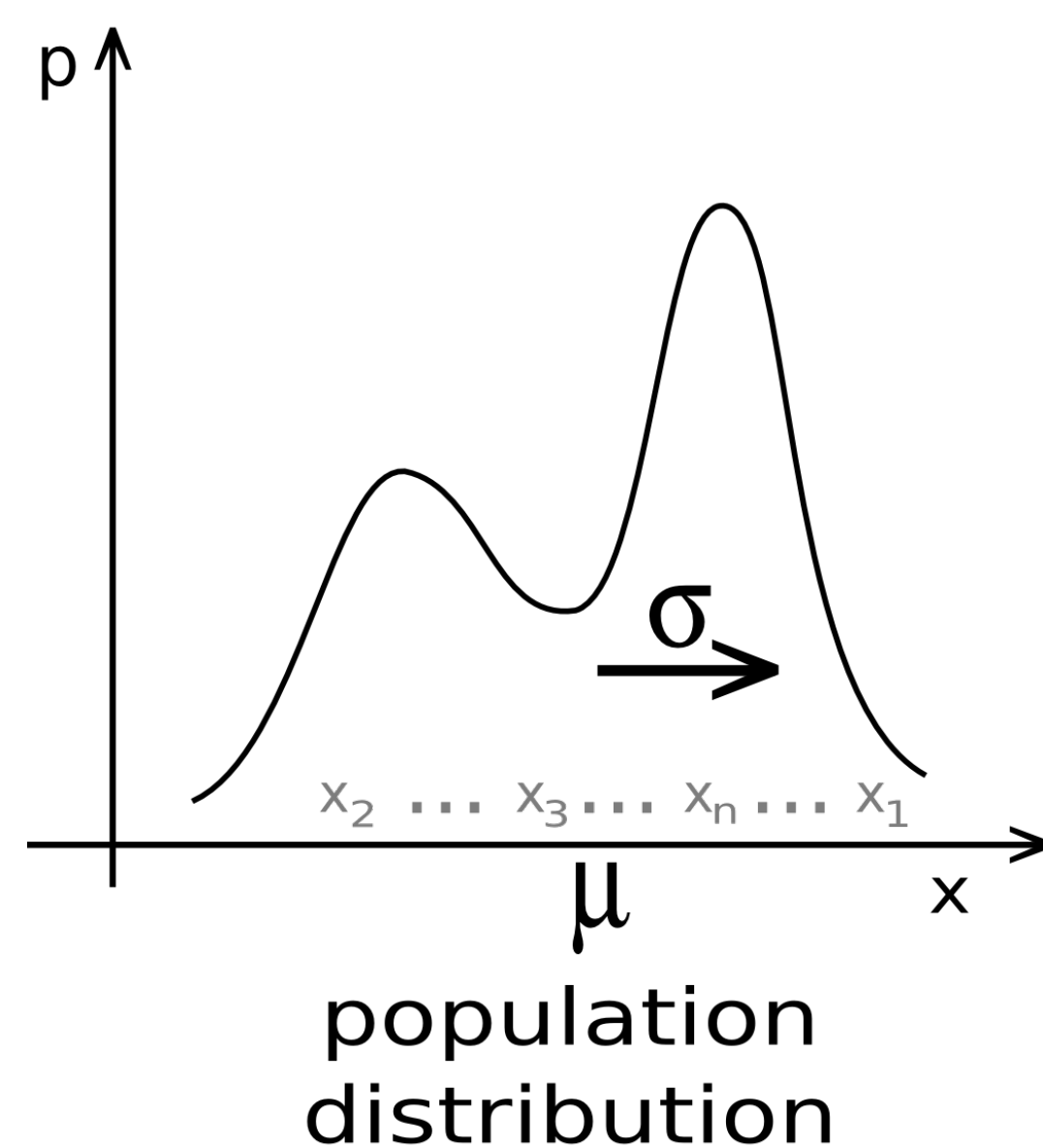
central limit theorem

- The sampling distribution of the sample mean is approximately normal
- This is crystallized as the **central limit theorem**
 - If X_1, X_2, \dots, X_n are iid random variables, then $\bar{X}_n \rightarrow \mathcal{N}(\mu, \sigma^2/n)$
 - If I take multiple samples from the same distribution, the means tend toward a normal distribution centered on the population mean
- There are some other convergence conditions that we won't get into here



in the limit

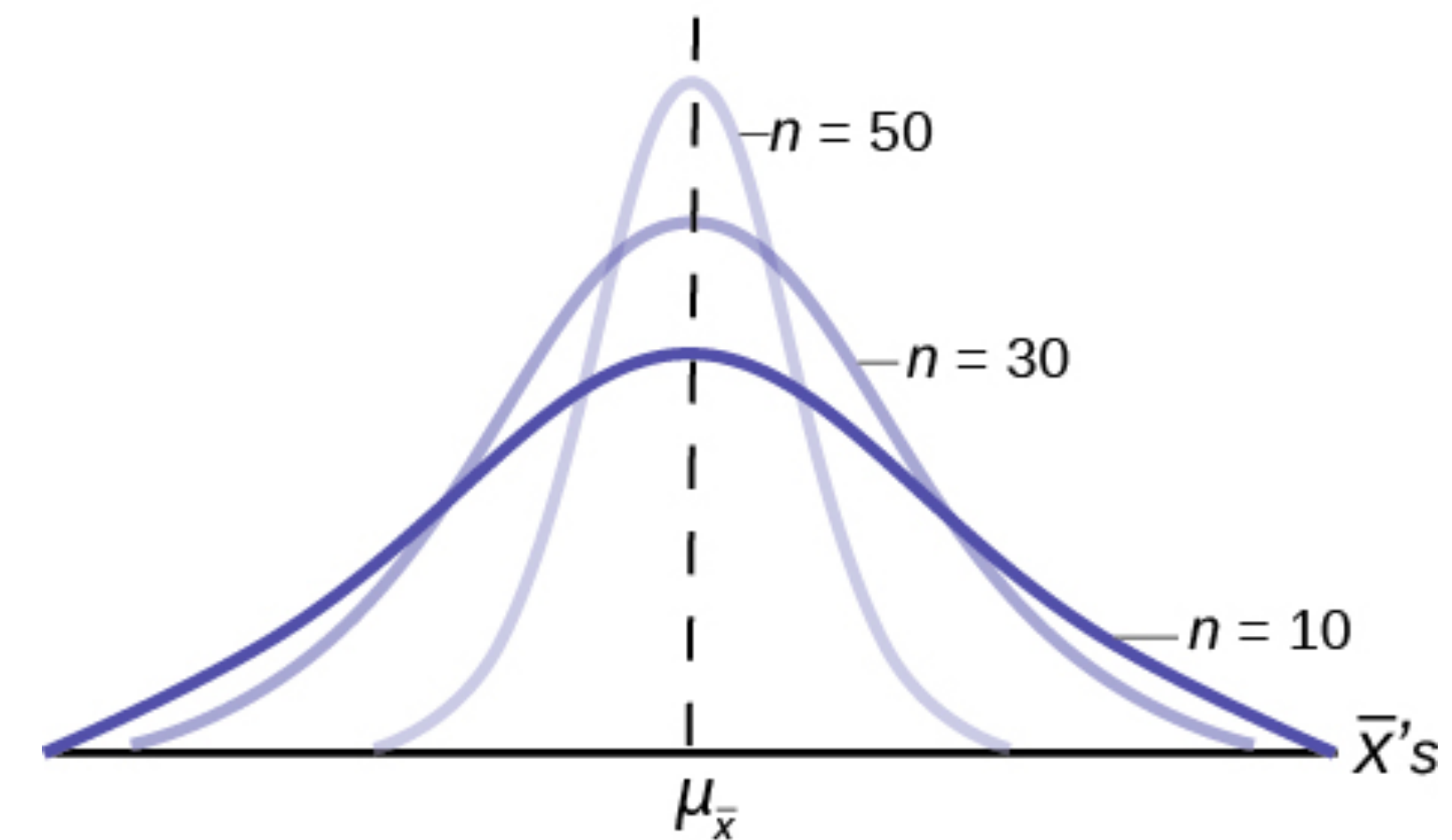
- Let's reason directly about the sampling distribution, as if we could repeat the experiment an infinite number of times
- Mean of sampling distribution: μ (the mean of the population)
- Variance of sampling distribution: σ^2/n (population variance decaying with n)



- We can approximate the population variance σ^2 by the sample variance s^2 when the size of samples n is large

how does this help us?

- Variance of sampling distribution: σ^2/n
 - The bigger the n (the bigger the samples used to generate the means), the smaller the variance of the sampling distribution (the more tightly clustered the means are)
 - In other words, the bigger your sample, the closer your sample mean is likely to be to the true mean
- Implication: if we have a sample mean (or means), we can use properties of the sampling distribution to let us judge ...
 - how good the estimates are (**confidence intervals**)
 - how likely a sample is to be an outlier (**hypothesis testing**)



example

Suppose that the number of YouTube videos Bob watches each day follows a Binomial distribution with 50 trials and a success probability 0.2. What is the distribution of the mean number of videos watched among a random sample of 100 days in Bob's life (assuming the days are independent)?

Note that if $X \sim \text{Bin}(n, p)$, then $\mu_X = np$ and $\sigma_X^2 = np(1 - p)$.

solution

The number of videos Bob watches in a single day follows $X \sim \text{Bin}(50, 0.2)$. Thus, $\mu_X = 50 \cdot 0.2 = 10$ and $\sigma_X^2 = 50 \cdot 0.2 \cdot 0.8 = 8$.

But we are not interested in X , we are interested in the sampling distribution \bar{X}_n over 100 samples. By the CLT, we know

$$\bar{X}_{100} \rightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(10, \frac{8}{100}\right) = \mathcal{N}(10, 0.08)$$

Even though X is Binomial, \bar{X} is Gaussian (for a sufficiently large number of samples).