Lecture 9 - Sampling Distributions and the CLT

Sta102/BME102

February 15, 2016

Colin Rundel

Variability of Estimates

Mean

Sample mean ($\bar{X}$):

$$\bar{X} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Population mean (μ):

$$\mu = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The sample mean ($\bar{X}$) is a *point estimate* of the population mean (μ) - this estimate may not be perfect, but if the sample is good (representative of the population) it should be close - today we will discuss how close.

Variance

Sample Variance (s^2)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Population Variance (σ^2) -

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Similarly, the sample variance (s^2) is a *point estimate* of the population variance (σ^2). For a decent sample, this should also be close to the population variance.

Parameter estimation

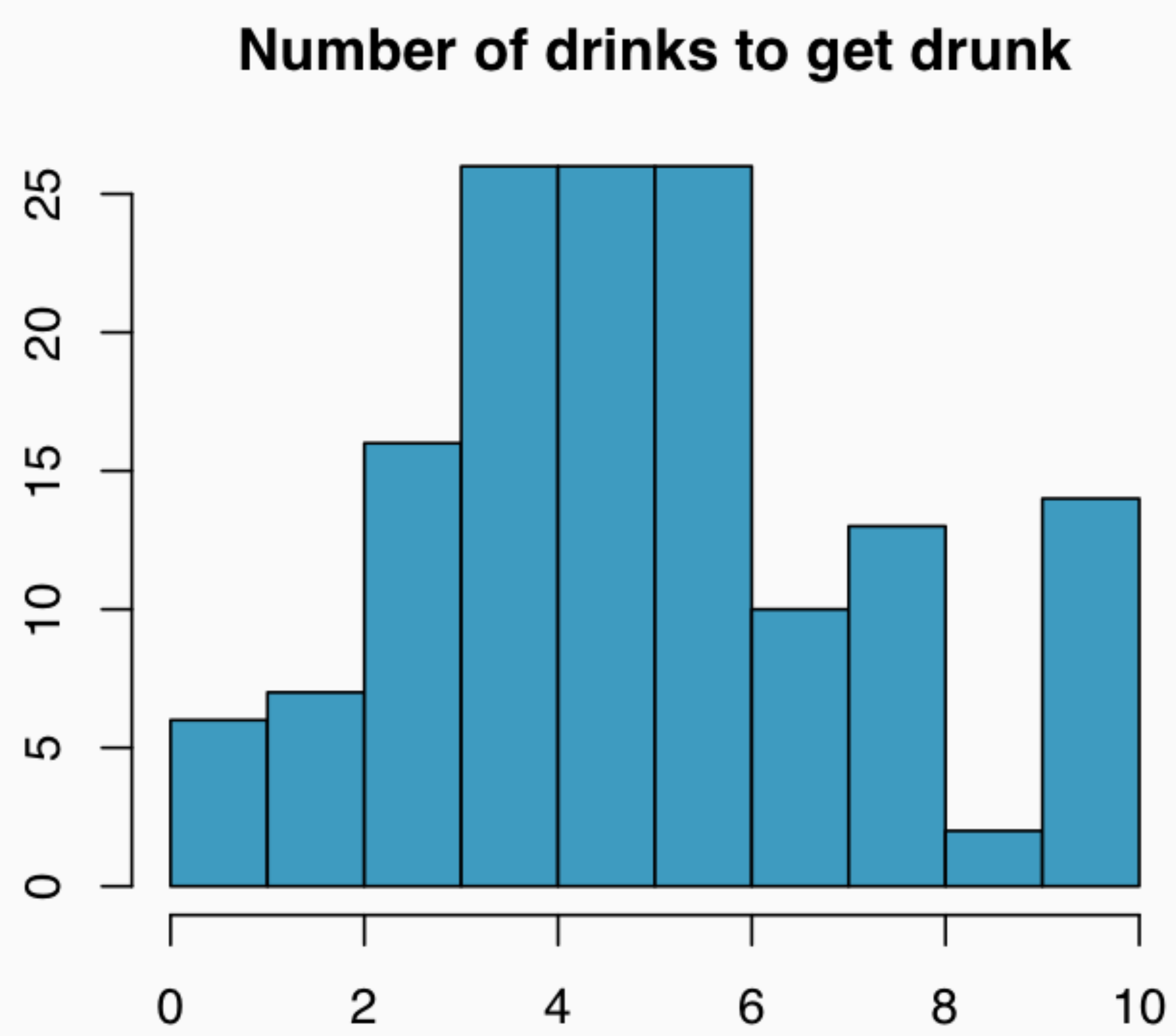
We are usually interested in knowing something about *population parameters*.

Since full populations are difficult (or impossible) to collect data on, we use *sample statistics* as *point estimates* for unknown population parameters of interest.

- Sample statistics vary from sample to sample.
- Quantifying how much sample statistics vary provides a way to estimate the *margin of error* associated with our point estimates.
- First we will look at how much point estimates vary from sample to sample.

Estimate the avg. # of drinks it takes to get drunk

We would like to estimate the average (self reported from students in a Duke Statistics class) number of drinks it takes a person get drunk, we will assume that this is population data:



$$\mu = 5.39 \qquad \sigma = 2.37$$

Activity

- Use RStudio to generate 10 random numbers between 1 and 146 (with replacement)

```
sample(1:146, size = 10, replace = TRUE)
```

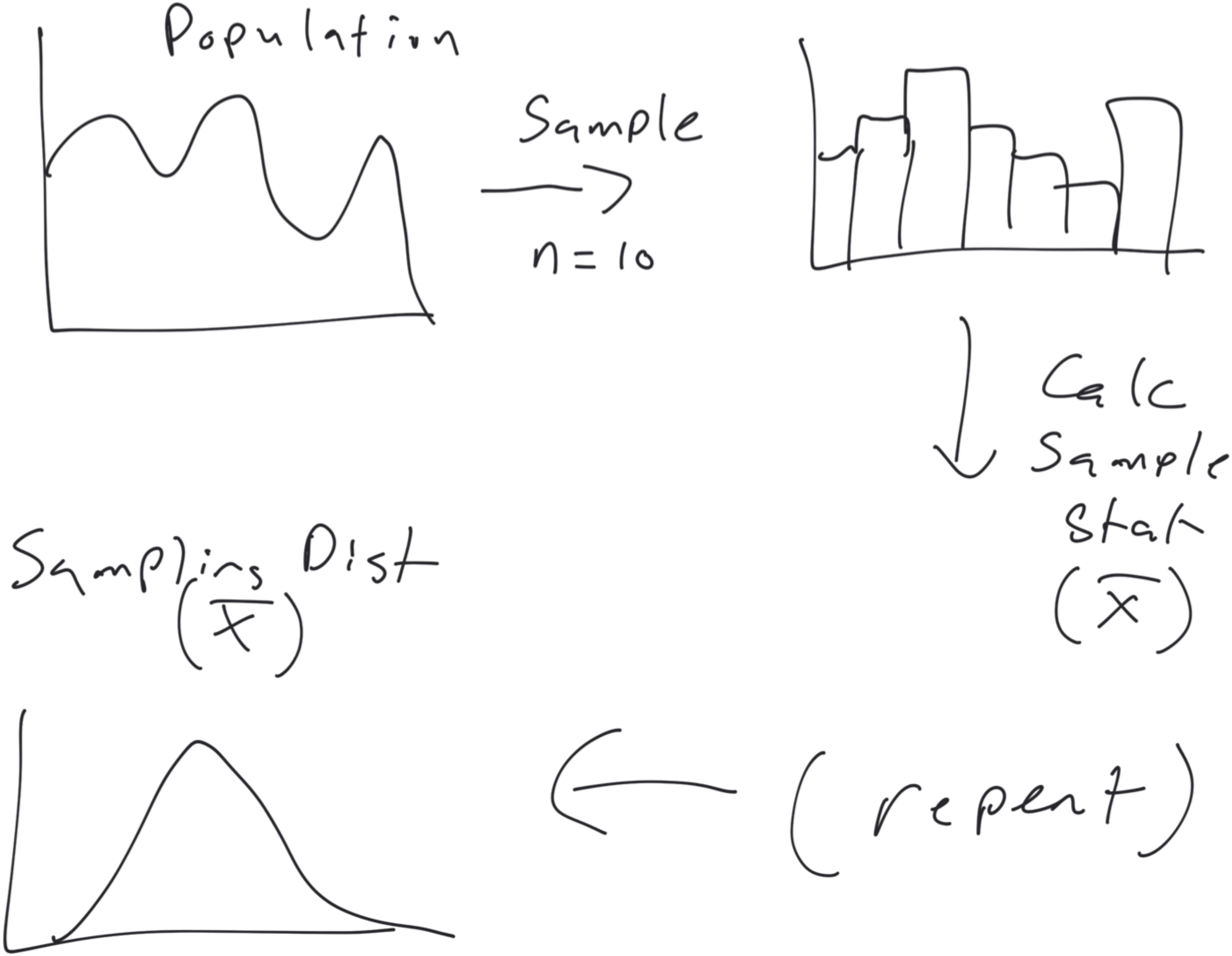
- If you don't have a computer, ask a neighbor to generate a sample for you.

- Using the handout find the 10 data points associated with your sampled values then
    - Calculate the sample mean of these 10 values
    - Round this mean to 1 decimal place
    - Report it using `http://bit.ly/Sta102_CLT`

```r
sample(1:146, size = 10, replace = TRUE)
```

```
## [1]  17  91  89  92 126  94   2  34  98  76
```

1	7	21	6	41	6	61	10	81	6	101	4	121	6	141	4
2	5	22	2	42	10	62	7	82	5	102	7	122	5	142	6
3	4	23	6	43	3	63	4	83	6	103	6	123	3	143	6
4	4	24	7	44	6	64	5	84	8	104	8	124	2	144	4
5	6	25	3	45	10	65	6	85	4	105	3	125	2	145	5
6	2	26	6	46	4	66	6	86	10	106	6	126	5	146	5
7	3	27	5	47	3	67	6	87	5	107	2	127	10		
8	5	28	8	48	3	68	7	88	10	108	5	128	4		
9	5	29	0	49	6	69	7	89	8	109	1	129	1		
10	6	30	8	50	8	70	5	90	5	110	5	130	4		
11	1	31	5	51	8	71	10	91	4	111	5	131	10		
12	10	32	9	52	8	72	3	92	0.5	112	4	132	8		
13	4	33	7	53	2	73	5.5	93	3	113	4	133	10		
14	4	34	5	54	4	74	7	94	3	114	9	134	6		
15	6	35	5	55	8	75	10	95	5	115	4	135	6		
16	3	36	7	56	3	76	6	96	6	116	3	136	6		
17	10	37	4	57	5	77	6	97	4	117	3	137	7		
18	8	38	0	58	5	78	5	98	4	118	4	138	3		
19	5	39	4	59	8	79	4	99	2	119	4	139	10		
20	10	40	3	60	4	80	5	100	5	120	8	140	4		


Population
Sample
n = 10
Calc
Sample
Stat
($\bar{x}$)
(repeat)
Sampling Dist
($\bar{x}$)

Sampling distribution

What we just constructed is called a *sampling distribution* - it is an empirical distribution of sample statistics ($\bar{X}$ in this case).

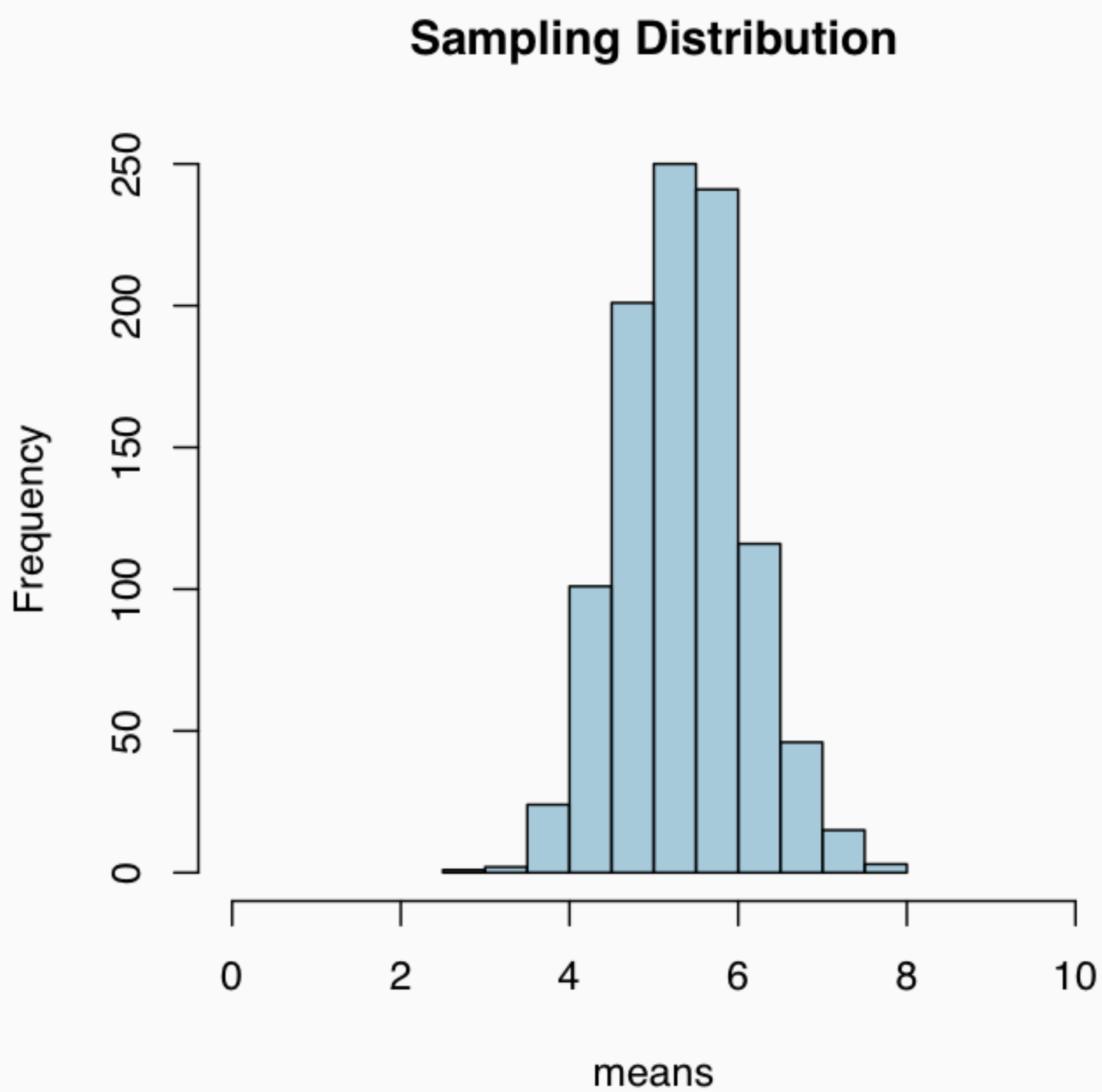
$\bar{X} \sim$ Dist

$E(\bar{X}) = \mu$

$Var(\bar{X}) = \downarrow$ as $n \uparrow$

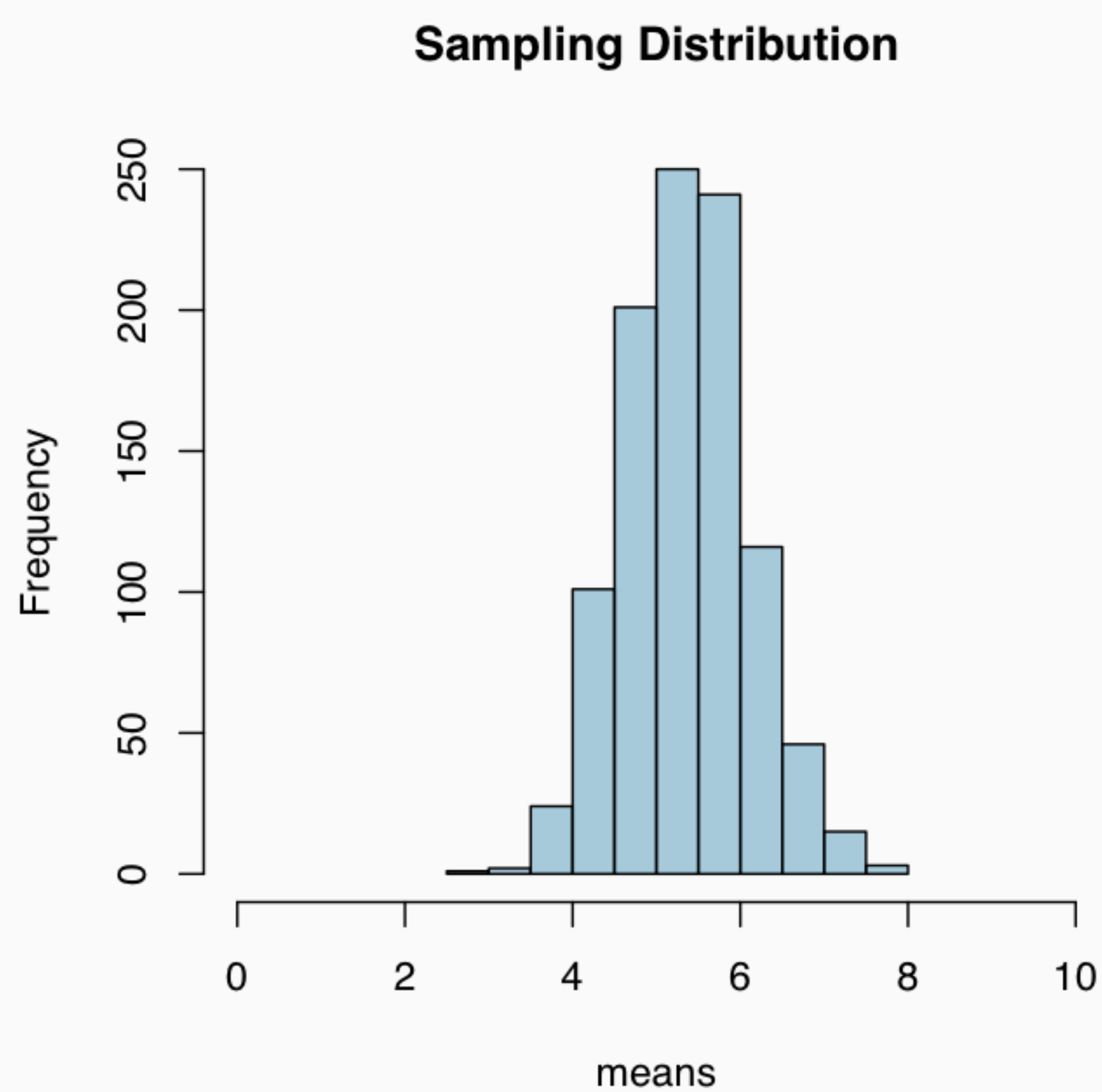
Increasing number of samples

If we increase the number of $\bar{X}$s we calculated to 1000 the sampling distribution looks like:



Increasing number of samples

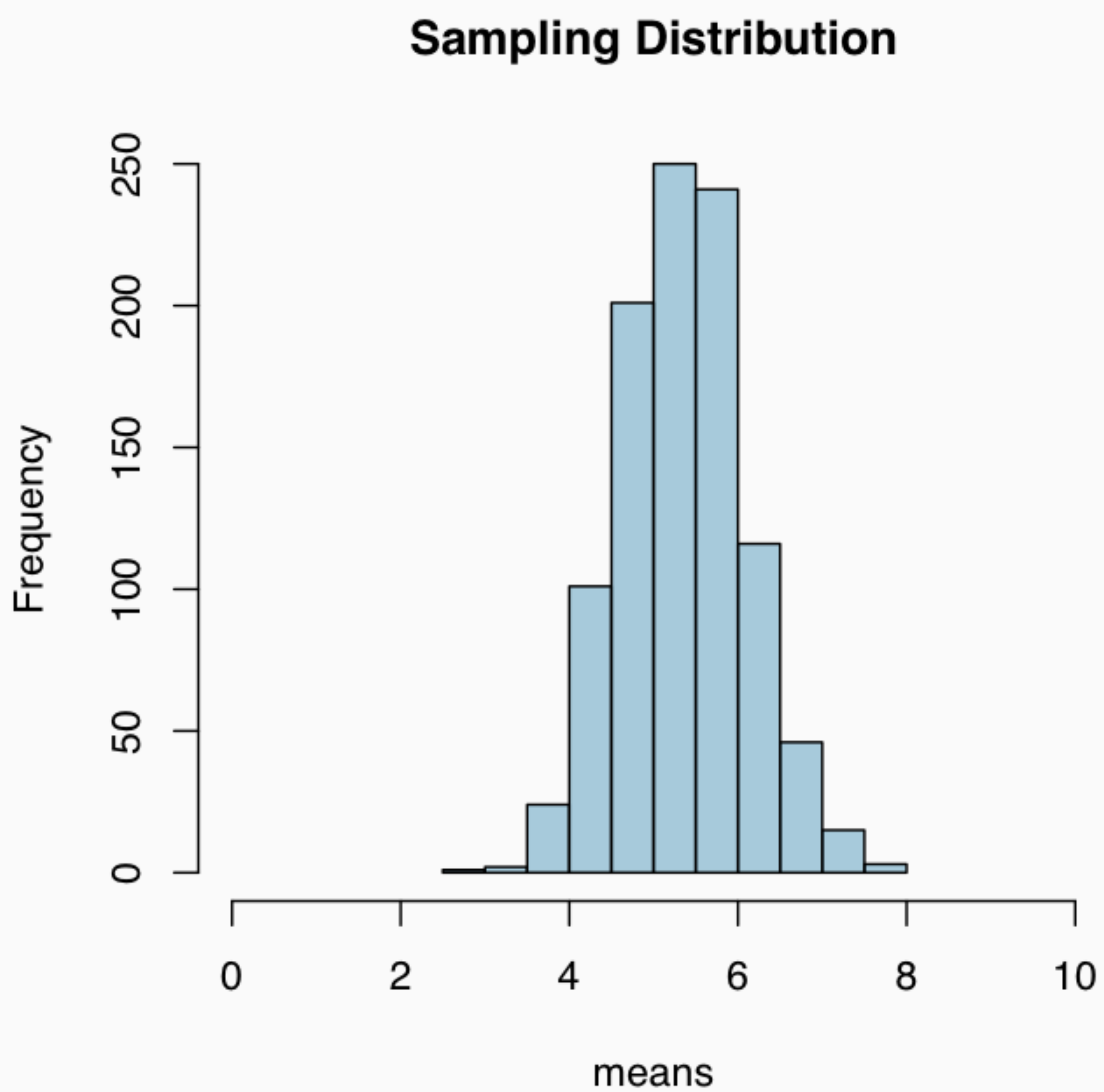
If we increase the number of $\bar{X}$s we calculated to 1000 the sampling distribution looks like:

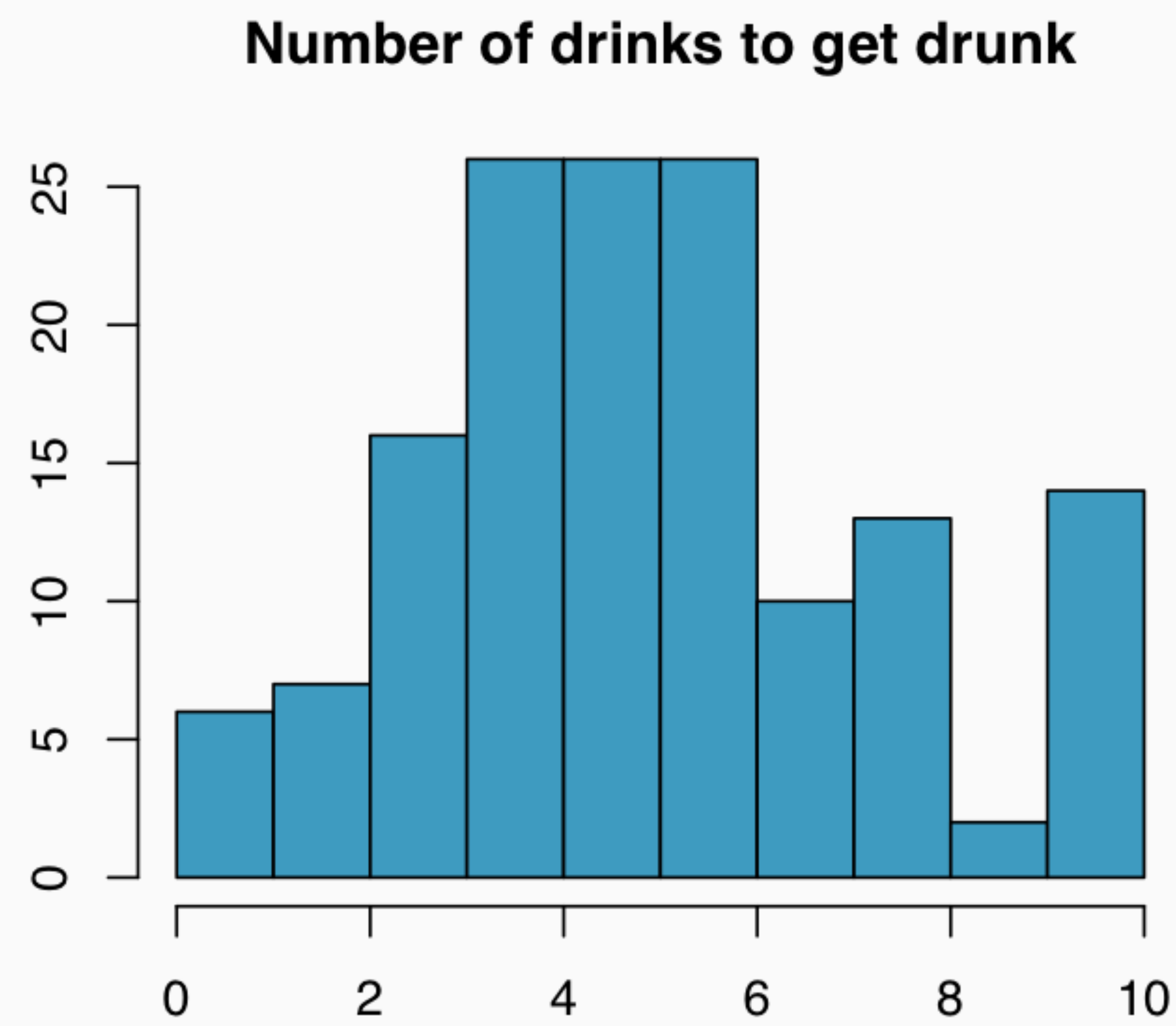


$$\text{avg}(\bar{X}) = 5.4 \qquad SD(\bar{X}) = 0.74$$

Increasing number of samples

If we increase the number of $\bar{X}$s we calculated to 1000 the sampling distribution looks like:



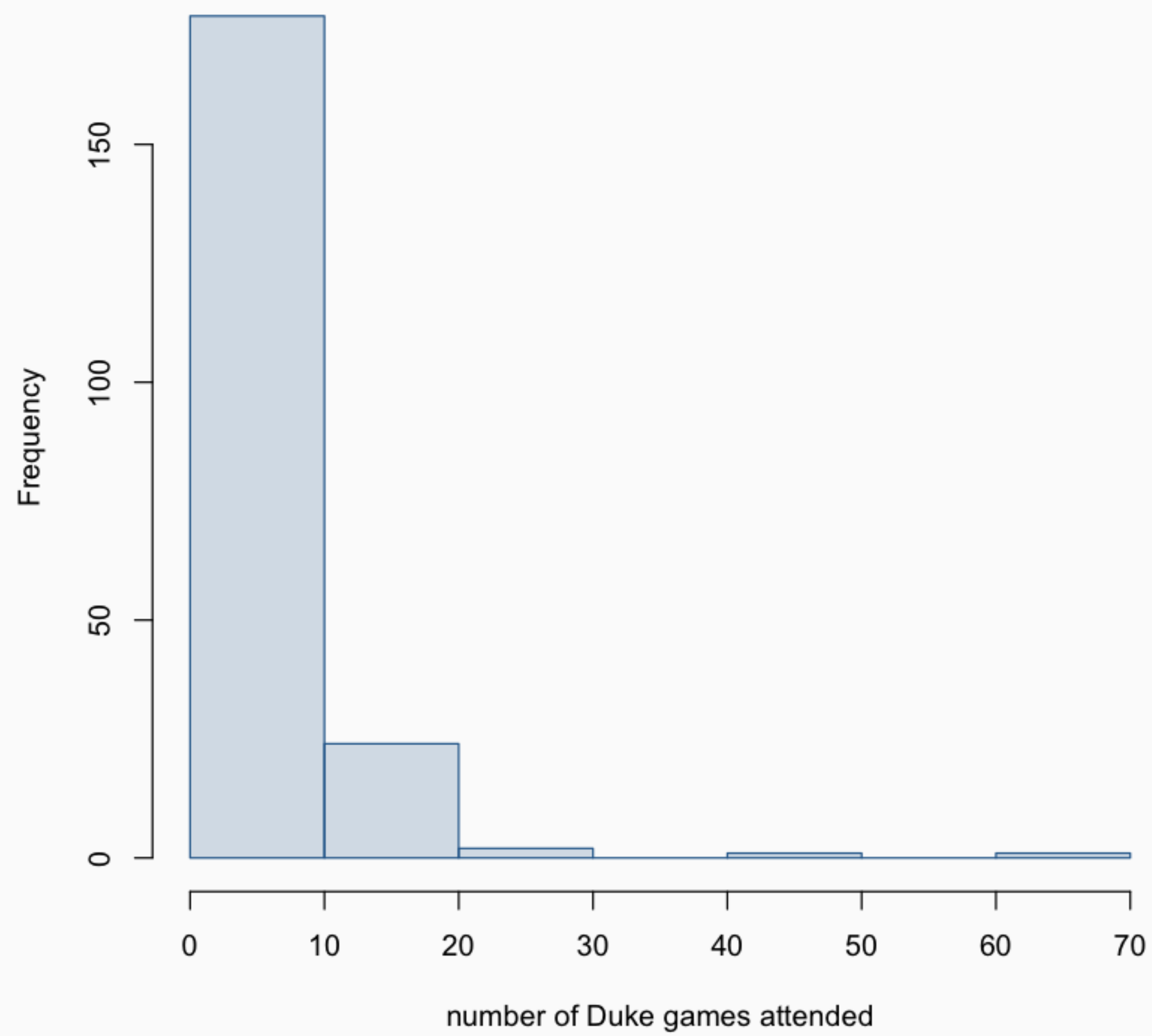


$\text{avg}(\bar{X}) = 5.4 \qquad SD(\bar{X}) = 0.74$

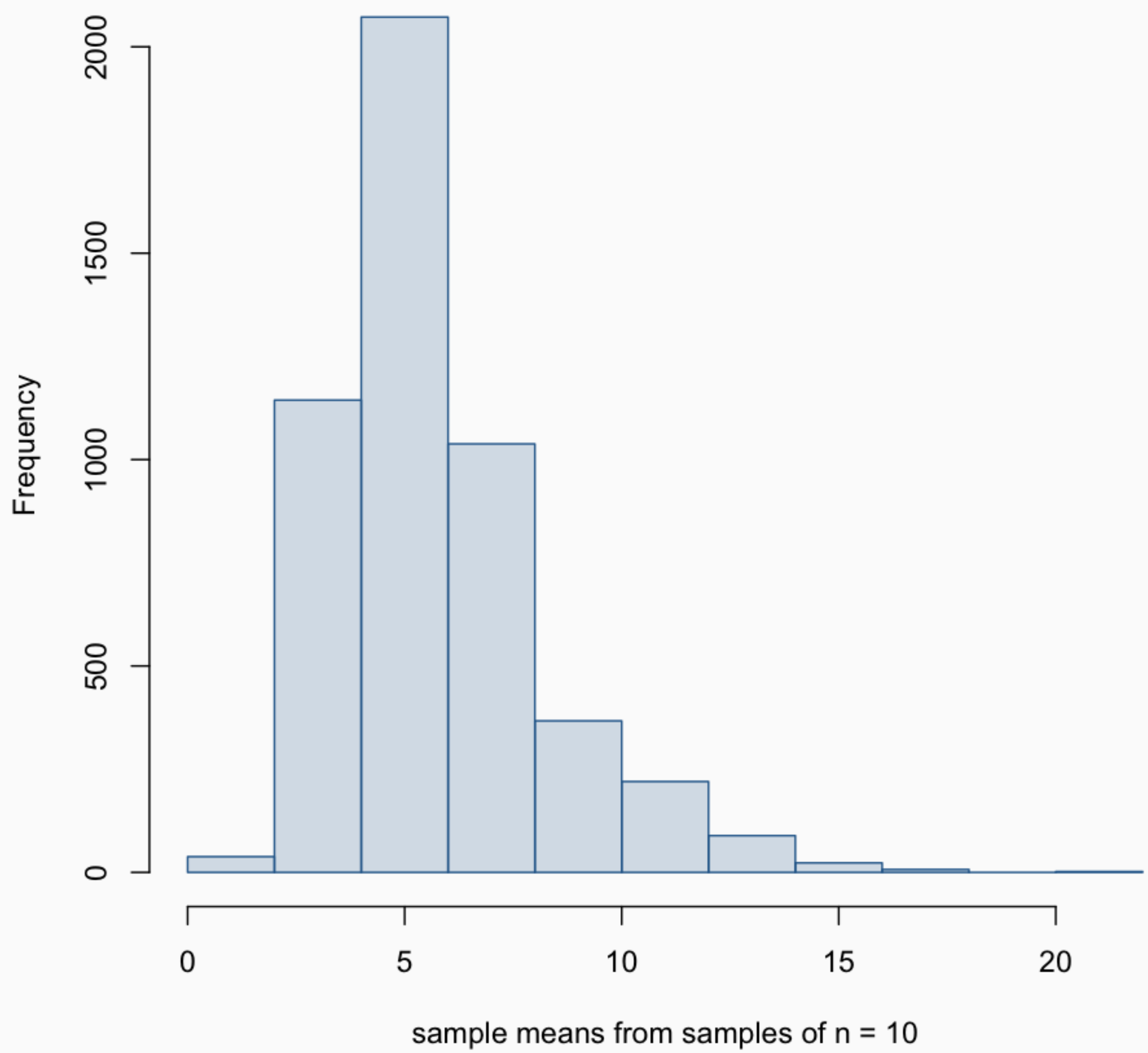
$\mu = 5.39 \qquad \sigma = 2.37$

Average number of Duke games attended

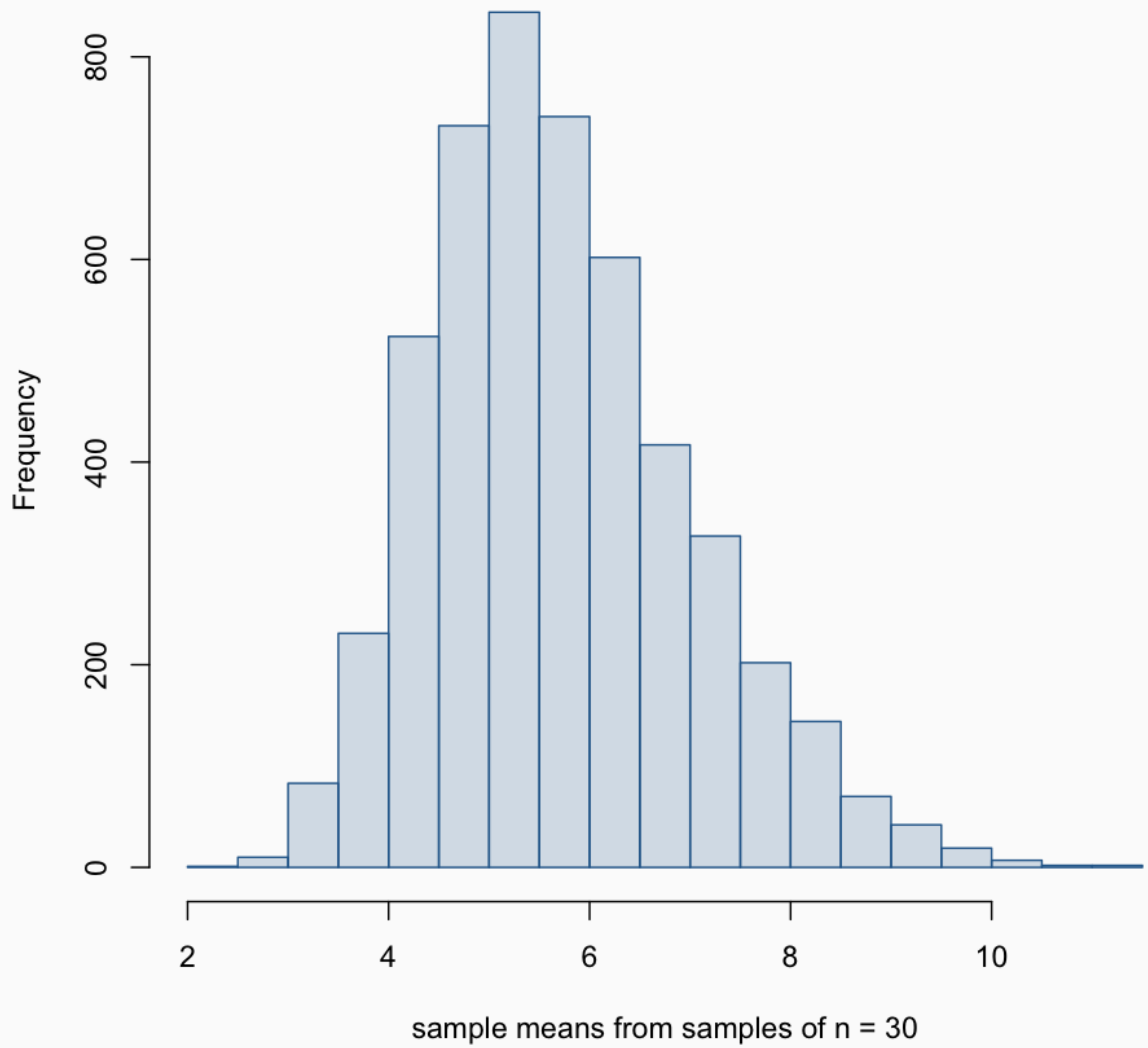
Next let's look at the population data for the number of basketball games attended:



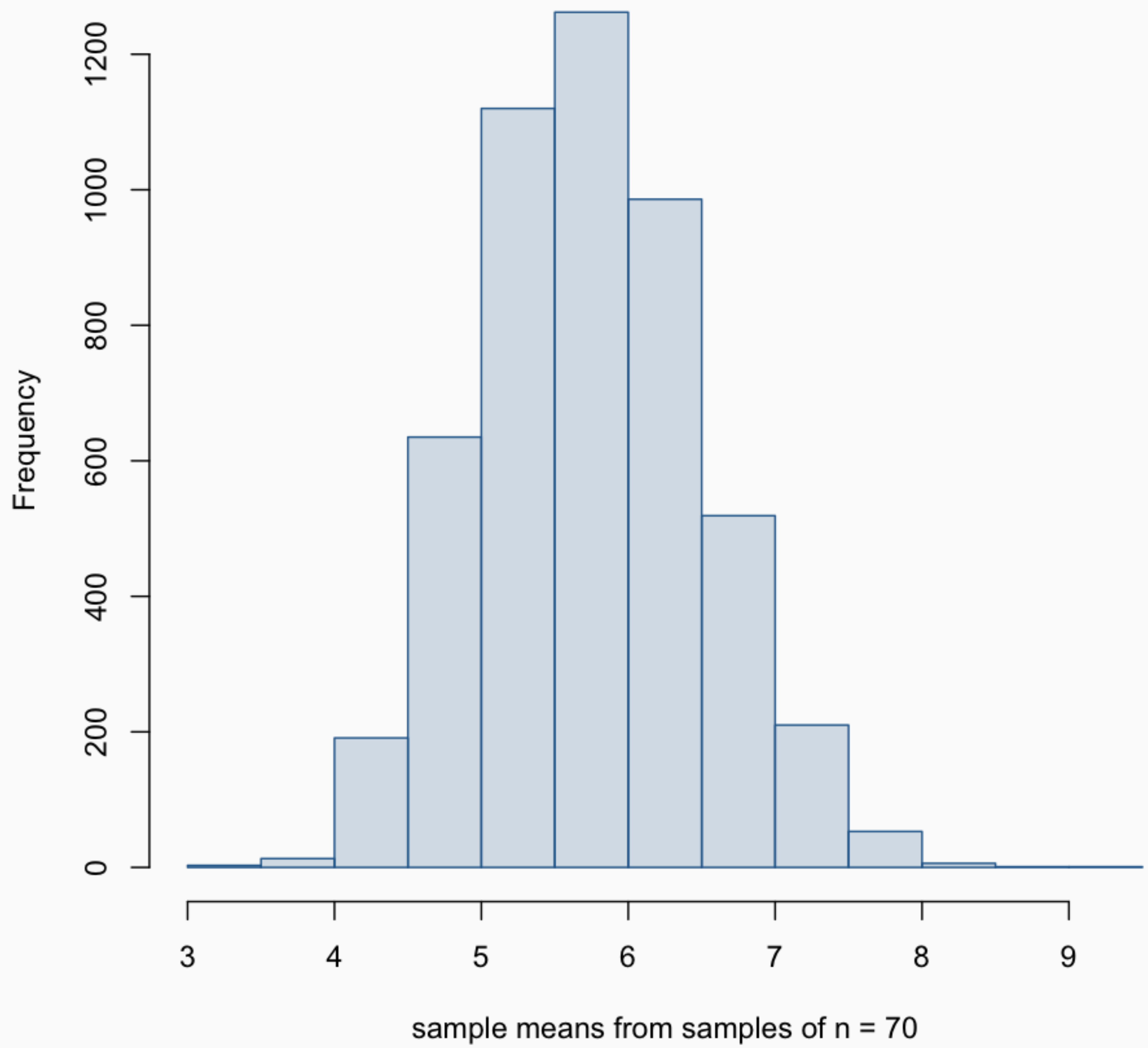
Sampling distribution of $\bar{x}$ when n = 10



Sampling distribution of $\bar{x}$ when n = 30



Sampling distribution of $\bar{x}$ when n = 70



General Patterns

As the sample size, n, increases the sampling distribution of $\bar{x}$:

- Centered at μ
- $Var(\bar{x})$ shrinks as $n \uparrow$
- $\bar{X} \sim N$, for large enough n

Sums of iid Random Variables

Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} D$ where D is some distribution with

$$E(X_i) = \mu \text{ and } Var(X_i) = \sigma^2.$$

Sums of iid Random Variables

Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} D$ where D is some distribution with

$$E(X_i) = \mu \text{ and } Var(X_i) = \sigma^2.$$

If we define $S_n = X_1 + X_2 + \cdots + X_n$ then what is expected value and variance of S_n?

$$E(S_n) = E\left(\sum x_i\right) = \sum_{i=1}^{n} E(x_i) = \sum_{i=1}^{n} \mu = n\mu$$

$$Var(S_n) = Var\left(\sum x_i\right) = n\sigma^2$$

Average of iid Random Variables

Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} D$ where D is some distribution with

$$E(X_i) = \mu \text{ and } Var(X_i) = \sigma^2.$$

If we define $\overline{X}_n = (X_1 + X_2 + \cdots + X_n)/n = S_n/n$ then what is the expected value and variance of $\overline{X}_n$?

$$E(\bar{X}) = E(S_n/n) = \frac{1}{n} E(S_n) = \frac{n\mu}{n} = \mu$$

$$Var(\overline{X}) = Var(S_n/n) = \frac{1}{n^2} Var(S_n)$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Central Limit Theorem

Central limit theorem - sum of iid RVs (S_n)

The distribution of the *sum* of n independent and identically distributed random variables X is approximately normal when n is large.

$$S_n \sim N\left(\mu = n\, E(X),\, \sigma^2 = n\, Var(X)\right)$$

Central Limit Theorem

Central limit theorem - sum of iid RVs (S_n)

The distribution of the *sum* of n independent and identically distributed random variables X is approximately normal when n is large.

$$S_n \sim N\left(\mu = n\,E(X),\, \sigma^2 = n\,Var(X)\right)$$

Central limit theorem - avergae of iid RVs ($\bar{X}$)

The distribution of the *average* of n independent and identically distributed random variables X is approximately normal when n is large.

$$\bar{X} \sim N\left(\mu = E(X),\, \sigma^2 = Var(X)/n\right)$$

Standard Error

We will be seeing the Central Limit Theorem throughout the rest of the course in a variety of different guises (different summary statistics / point estimates - depending on the data and mode of inference).

One common feature we will be looking at is the uncertainty of the *sampling distribution*. This is given a special name when we discuss the standard deviation, which we call the *Standard Error*.

$$SE = \sqrt{\frac{Var(X)}{n}} = \frac{SD(X)}{\sqrt{n}}$$

CLT - Conditions

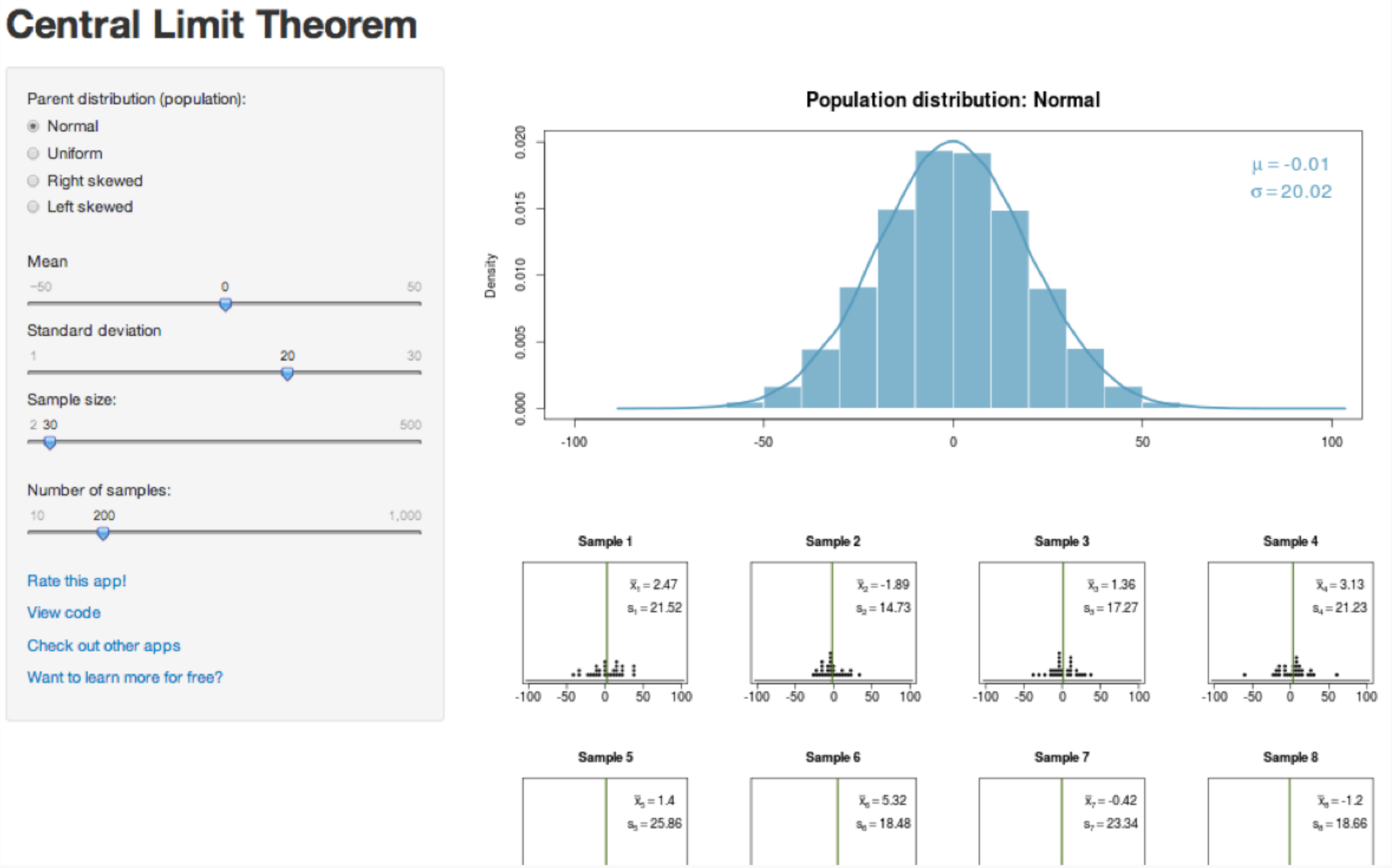
Certain conditions are required for the CLT to apply:

1. *Independence:* Sampled observations must be independent and identically distributed.

   Not true for samples collected without replacement, but approximately correct if
   - random sampling/assignment is used, and
   - $n < 10\%$ of the population.
2. *Sample size/skew:* the population distribution must be nearly normal *or* the sample size must be large (the less normal the population distribution, the larger the sample size needs to be).

   Usually checked using the sample data - assume that the distribution of the sample is similar to the population distribution.

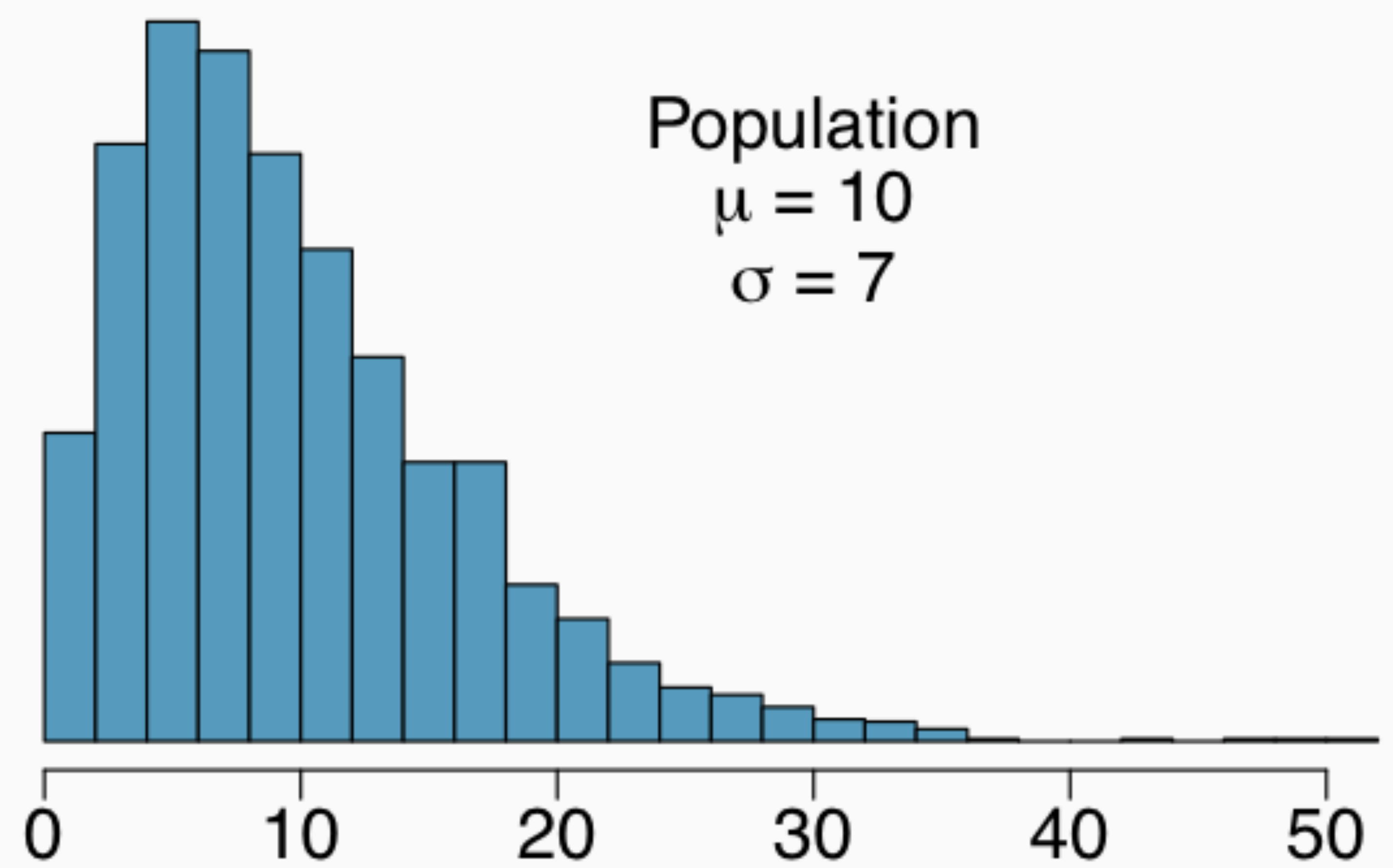
CLT - Simulation



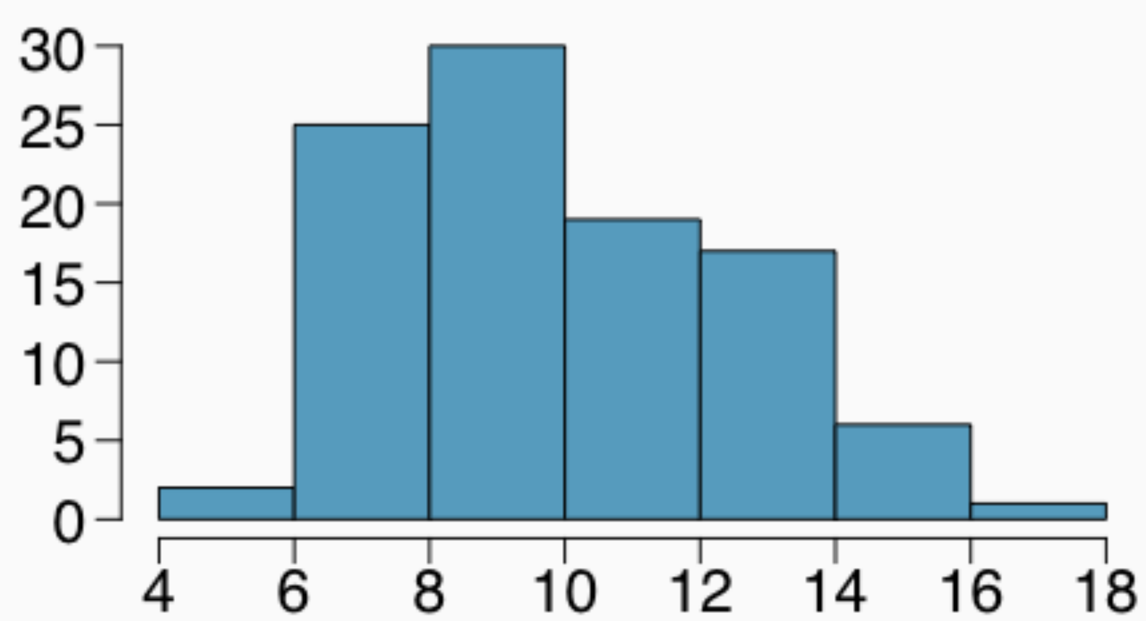
https://gallery.shinyapps.io/CLT_mean/

Review

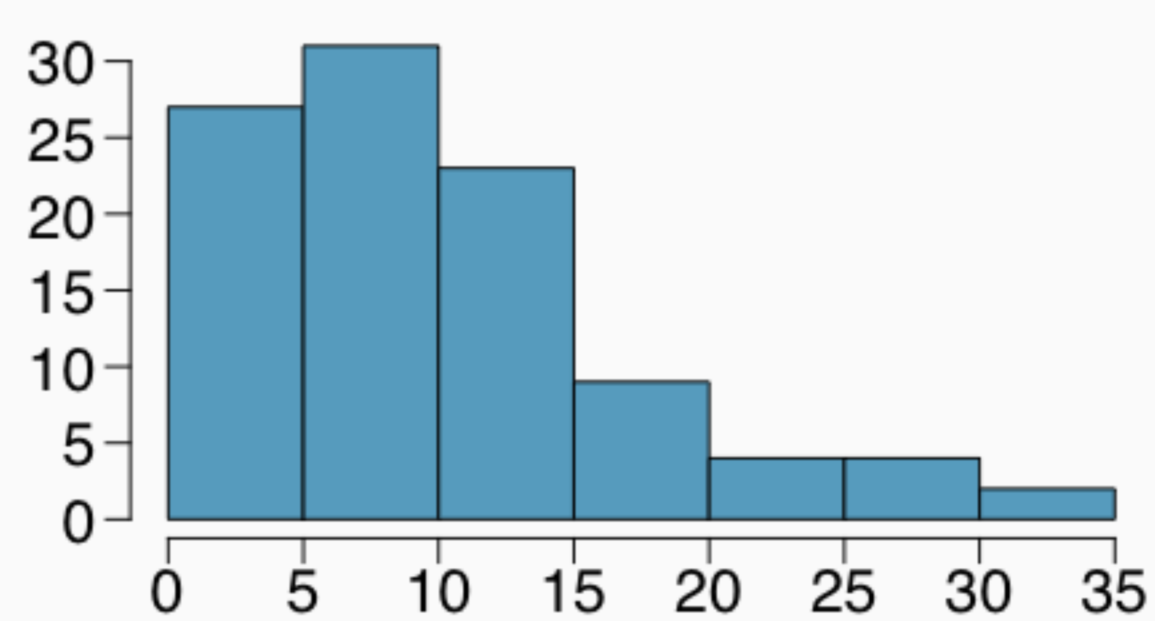
To the right is a plot of a population distribution. Match each of the following descriptions to one of the three plots below.



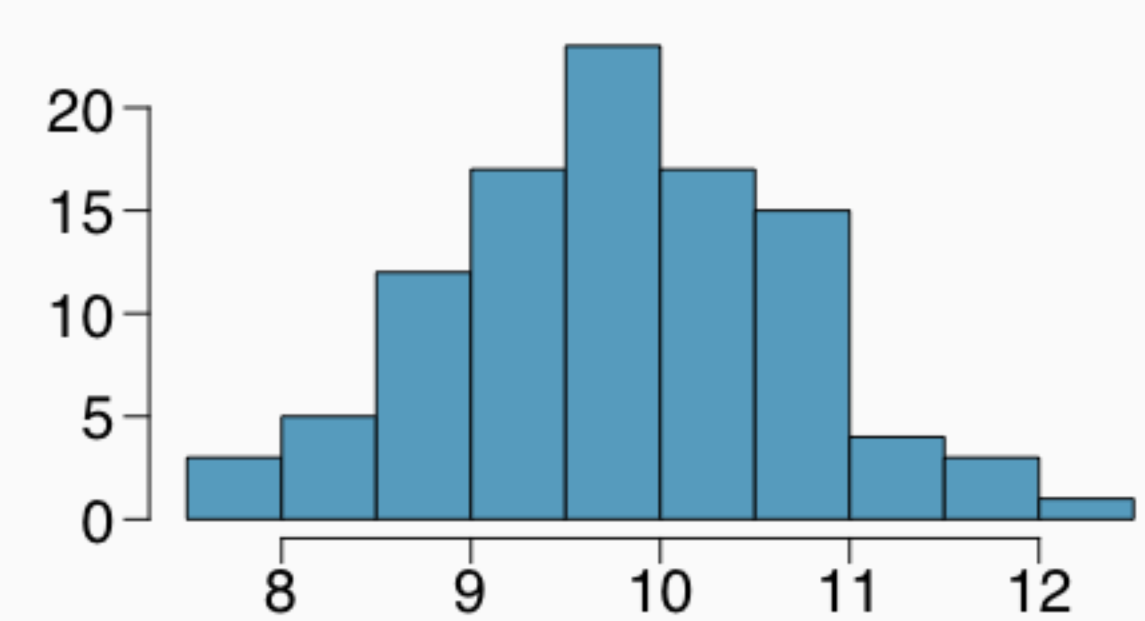
1. a single random sample of 100 observations from this population B
2. a distribution of 100 sample means from random samples with size 7 = n A
3. a distribution of 100 sample means from random samples with size 49 = n C



Plot A



Plot B



Plot C

Confidence intervals

Confidence intervals

Using only a point estimate to estimate a parameter is like fishing in a murky lake with a spear, while a confidence interval is like a fishing net.

If we report a point estimate, we probably will not hit the exact population parameter. If we report a range of plausible values – *a confidence interval* – we have a good shot at capturing the parameter.

Confidence intervals and the CLT

We have a point estimate, $\bar{X}$, for the population mean μ, but we want to design a “net” to have a reasonable chance of capturing μ.

Confidence intervals and the CLT

We have a point estimate, $\bar{X}$, for the population mean μ, but we want to design a “net” to have a reasonable chance of capturing μ.

- The CLT tells us that $\bar{X}$ is a sample from $N(\mu,\ \sigma/\sqrt{n})$.

Confidence intervals and the CLT

We have a point estimate, $\bar{X}$, for the population mean μ, but we want to design a "net" to have a reasonable chance of capturing μ.

- The CLT tells us that $\bar{X}$ is a sample from $N(\mu,\ \sigma/\sqrt{n})$.

- Therefore, 95% of the time a sample's mean ($\bar{X}$) will be within 2 SEs ($2\sigma/\sqrt{n}$) of μ.

Confidence intervals and the CLT

We have a point estimate, $\bar{X}$, for the population mean μ, but we want to design a "net" to have a reasonable chance of capturing μ.

- The CLT tells us that $\bar{X}$ is a sample from $N(\mu,\ \sigma/\sqrt{n})$.
- Therefore, 95% of the time a sample's mean ($\bar{X}$) will be within 2 SEs ($2\sigma/\sqrt{n}$) of μ.
- Then for 95% of samples from the population, μ will be with in 2 SEs of $\bar{X}$.

Example - Cardinals

A transect was sampled 50 times by counting the number of cardinals seen when walking a 1 mile path in the Duke forest. The mean of these samples was 13.2. Estimate the true average number of cardinals along this path, assuming the population distribution is nearly normal with a population standard deviation of 1.74.

Example - Cardinals

A transect was sampled 50 times by counting the number of cardinals seen when walking a 1 mile path in the Duke forest. The mean of these samples was 13.2. Estimate the true average number of cardinals along this path, assuming the population distribution is nearly normal with a population standard deviation of 1.74.

The 95% confidence interval is defined as

$$\textit{point estimate} \pm 2 \times SE$$

.

Example - Cardinals

A transect was ~~sampled~~ walked 50 times by counting the number of cardinals seen when walking a 1 mile path in the Duke forest. The mean of these samples was 13.2. Estimate the true average number of cardinals along this path, assuming the population distribution is nearly normal with a population standard deviation of 1.74.

$\mu = 15$

$\mu = 13$

The 95% confidence interval is defined as

$$\textit{point estimate} \pm 2 \times SE$$

$$\bar{X} = 13.2 \qquad \sigma = 1.74 \qquad SE = \frac{\sigma}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} = 0.25$$

$$\bar{x} \pm 2\frac{\sigma}{\sqrt{n}} = 13.2 \pm 2 \times 0.25$$

$$= (12.7, 13.7)$$

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* $\pm 2 \times SE$.

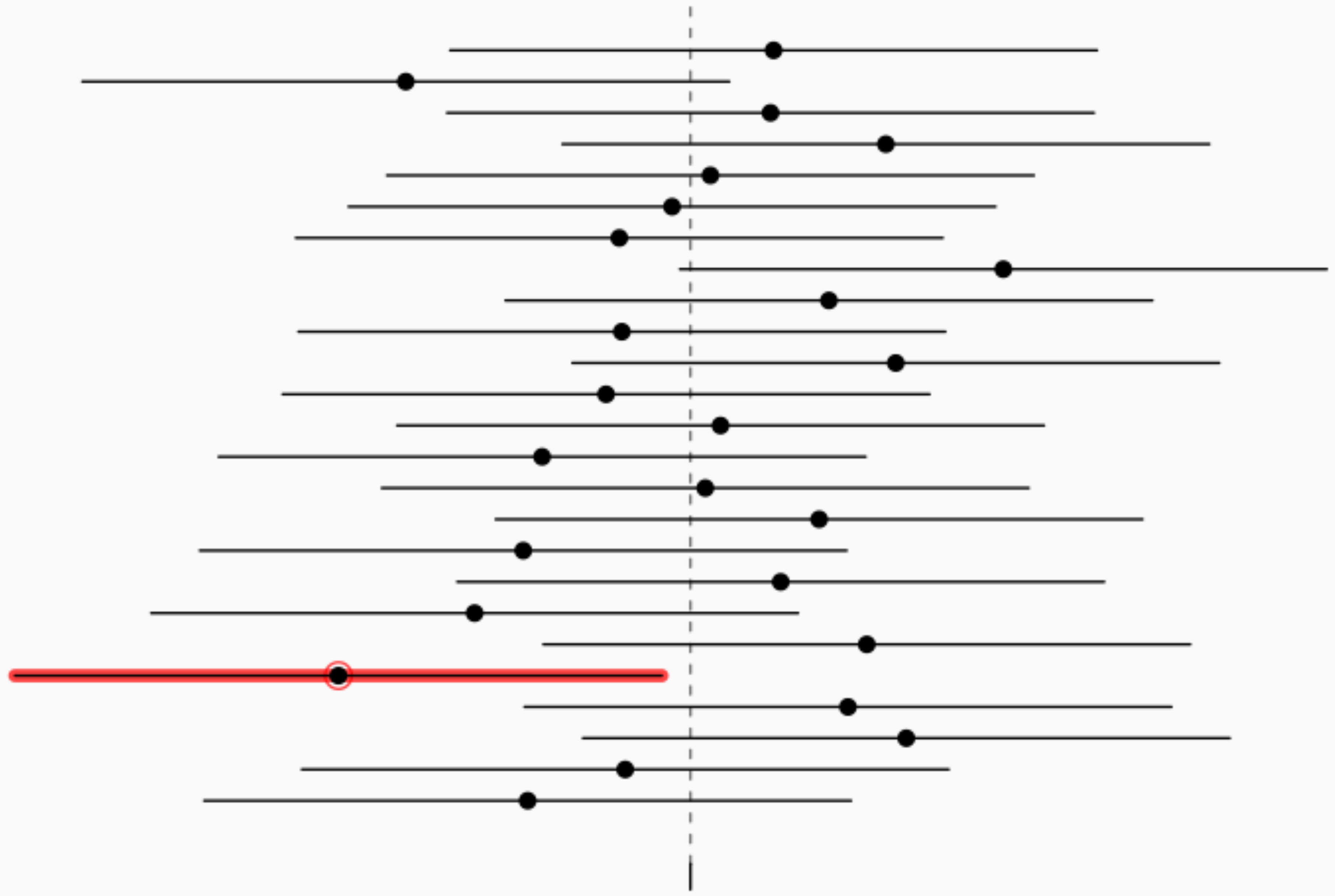
Then about 95% of those intervals would contain the true population mean (μ).

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate* $\pm 2 \times SE$.

Then about 95% of those intervals would contain the true population mean (μ).

The figure on the right shows this process with 25 samples, in this case 24 of the calculated confidence intervals contain the true population average.



A more general confidence interval

A Confidence interval is constructed using the general formula:

$$\textit{point estimate} \pm \textit{CV} \times \textit{SE}$$

A more general confidence interval

A Confidence interval is constructed using the general formula:

$$\textit{point estimate} \pm CV \times SE$$

Conditions when the point estimate is $\bar{X}$:

1. *Independence*: Observations in the sample must be independent
   - random sample/assignment
   - $n < 10\%$ of population
2. *Normality*: nearly normal population distribution or large enough sample
3. *Population Variance*: so far we've assumed this is known, this is almost never true. We'll talk about a more general approach next time.

Changing the confidence level

In general,

$$point\ estimate \pm CV \times SE$$

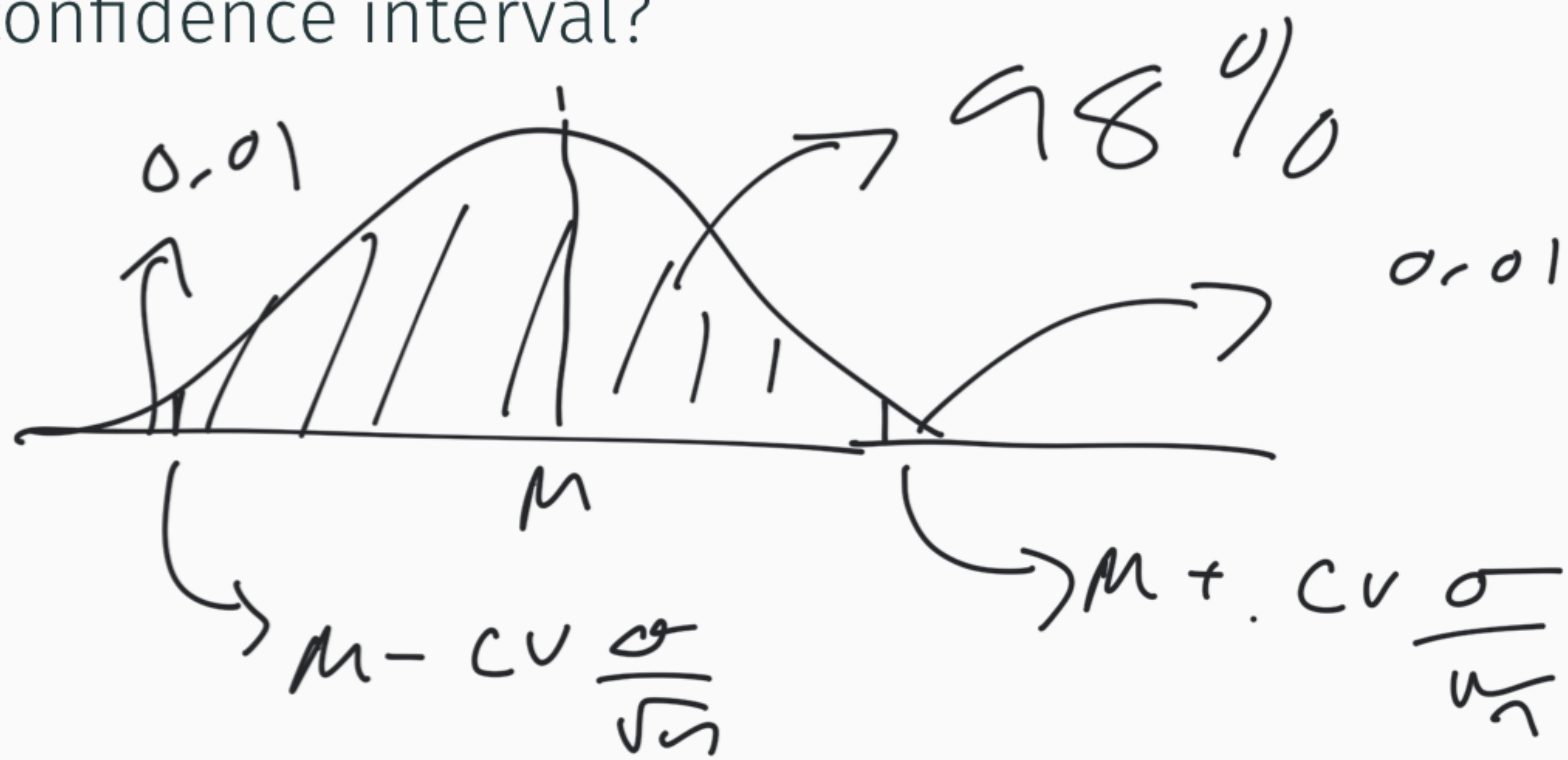
- In order to change the confidence level all we need to do is adjust the critical value in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

If the conditions for the CLT are met then,

- For a 95% confidence interval, $CV = Z^{\star} = 1.96$.
- Using the Z table it is possible to find the appropriate $Z^{\star}$ for any desired confidence level.

Example - Calculating $Z^{\star}$

What is the appropriate value for $Z^{\star}$ when calculating a 98% confidence interval?



$$\Phi(x) = 0.01 \quad \text{or} \quad 1 - \Phi(x) = 0.01$$

Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

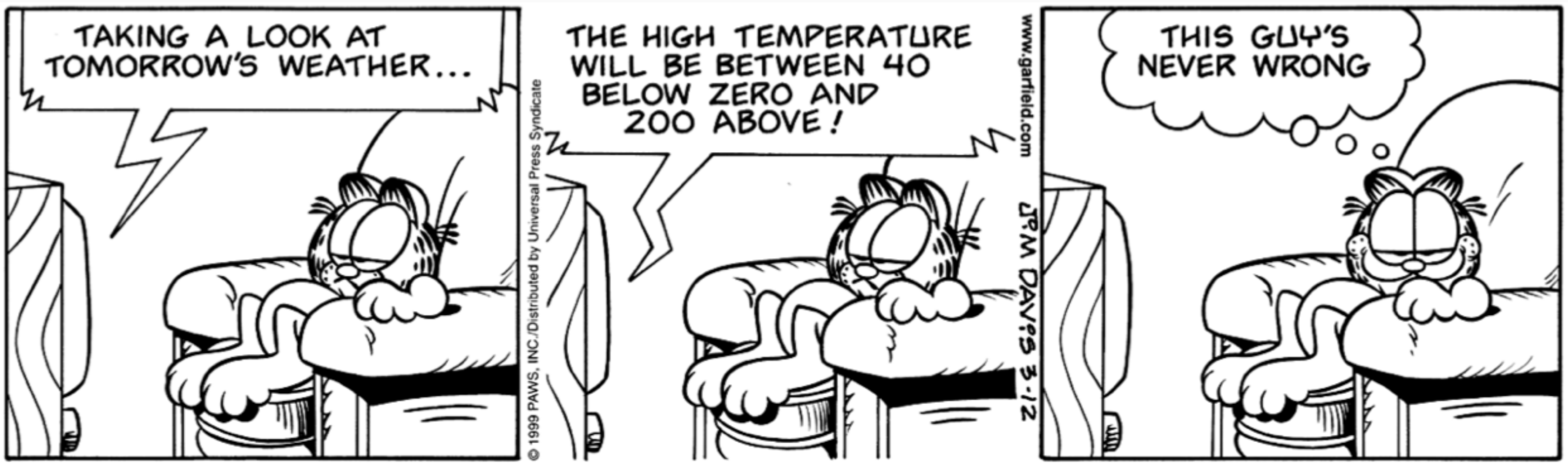
Can you see any drawbacks to using a wider interval?

Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

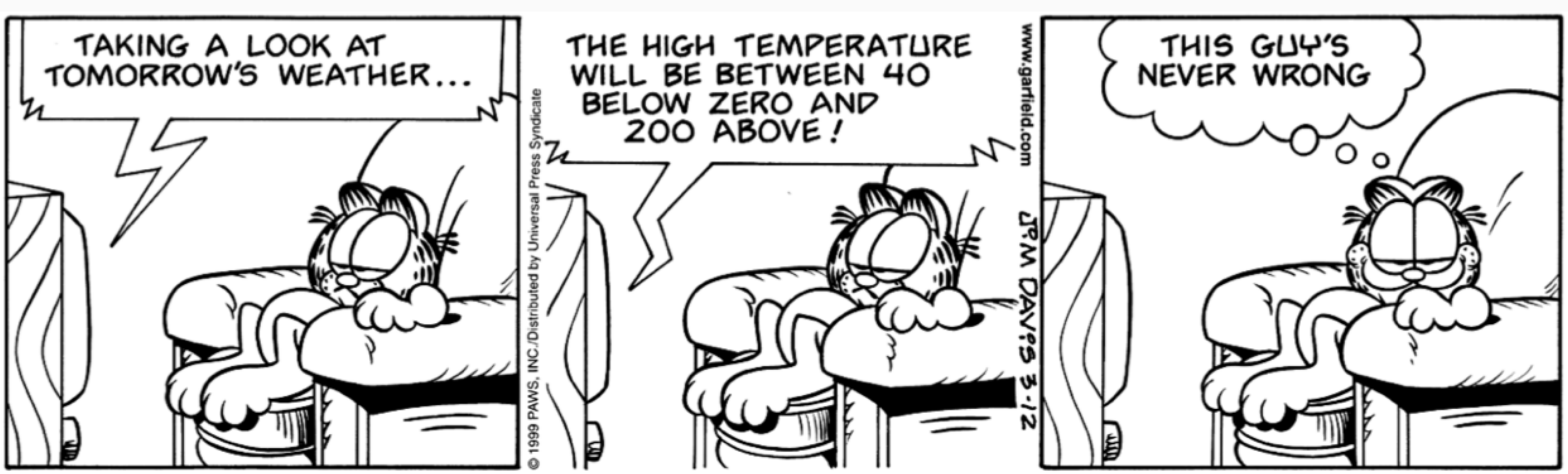


Width of an interval

If we want to be very certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Example - Sample Size

Coca-Cola wants to estimate the per capita number of Coke products consumed each year in the United States, in order to properly forecast market demands they need their margin of error to be 5 items at the 95% confidence level. From previous years they know that $\sigma \approx 30$. How many people should they survey to achieve the desired accuracy? What if the requirement was at the 99% confidence level?

Example - Sample Size

Coca-Cola wants to estimate the per capita number of Coke products consumed each year in the United States, in order to properly forecast market demands they need their margin of error to be 5 items at the 95% confidence level. From previous years they know that $\sigma \approx 30$. How many people should they survey to achieve the desired accuracy? What if the requirement was at the 99% confidence level?

At the 95% and 99% confidence levels $Z^\star$ is 1.96 and 2.58 respectively. Therefore,

$$M.E = CV \cdot SE = 5$$

$$Z^* \frac{30}{\sqrt{n}} = 5$$

Common Misconceptions

1. The confidence level of a confidence interval is the probability that the interval contains the true population parameter.

Common Misconceptions

1. The confidence level of a confidence interval is the probability that the interval contains the true population parameter.

   This is incorrect, CIs are part of the frequentist paradigm and as such the population parameter is fixed but unknown. Consequently, the probability any given CI contains the true value must be 0 or 1 (it does or does not).

Common Misconceptions

2. A narrower confidence interval is always better.

Common Misconceptions

2. A narrower confidence interval is always better.

   This is incorrect since the width is a function of both the confidence level (CV) and the standard error.

Common Misconceptions

3. A wider interval means less confidence.

Common Misconceptions

3. A wider interval means less confidence.

   This is incorrect since it is possible to make very precise statements with very little confidence.