Lecture 10

Part 2 Probability and Distributions

Before We Start

- So far, we have learned a lot about the principles of probability and distribution.
  - My question is: if you flip a **fair** coin 1000 times, according to the Law of Large Numbers we introduced, you should theoretically get around 500 heads.
  - However, let's say in reality, after many flips, for example, after flipping 800 or 900 times, luck hasn't been on your side and most of the results are heads.
  - In this case, what do you think is the probability of getting heads on your next flip?

Before We Start

- According to the Law of Large Numbers, is the next flip more likely to be tails, i.e., with a probability greater than 0.5?
  - Because the Law of Large Numbers tells us that over many trials, the overall relative frequency of heads should be 0.5.
- The above statement is not correct.
  - Based on what we've learned, since each flip is independent of the others, for a fair coin, the probability of heads coming up is always 0.5, regardless of when you flip it.
  - For repeated independent trials, past results do not affect future outcomes.
  - This means that if you lose several gambling games in a row, it does not imply that your chances of winning in the next game are higher.

Before We Start

- Does this contradict the Law of Large Numbers? No, it doesn't.
  - The Law of Large Numbers tells us that the overall probability of heads will approach 0.5, but extreme cases with low probability, such as many consecutive heads for this case, are still possible; it's just that the likelihood of such an event occurring is relatively low!

We have talked about population distribution, let's talk about sampling distribution today.

Statistical Inference

- **Statistical inference:** Determining the behaviour of a population by studying a sample from that population.
  - Population parameters describe populations, e.g., population mean μ and population variance σ^2.
  - Population parameters usually unknown.
  - To find out more about them, we take a random sample of data.
  - Calculate statistics from sample, e.g., sample mean $\bar{X}$ and sample variance s^2.
  - Use $\bar{X}$ and s^2 to estimate μ and σ^2.

Statistical Inference

- So we use sample statistics to make inferences about population parameters.
- Suppose we take a new sample.
- We would probably get a different $\bar{X}$ and s^2.
- Sample statistics are random variables, so there is variability associated with them.
- We would like to know how they differ from sample to sample.

Sampling Distribution

- **Sampling distribution:** The distribution of a sample statistic that would arise if we were to repeatedly take random samples from a population.
- Knowing the sampling distribution of a sample statistic can tell us how accurately we are estimating a population parameter.
- We will focus on the sampling distribution of the sample mean and the sampling distribution of the sample proportion.

Example 1

- Suppose we roll a fair four-sided die and let X be the number that comes up.
- The population can be thought of as an infinite number of rolls of the die.
- Let's repeatedly draw samples of size 2, i.e., roll the die twice.
- Find the sampling distribution of the sample mean $\bar{X}$.

Probability Distribution of X

x	1	2	3	4
$p(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

- Remember that probability distributions represent populations.
- We can calculate the population mean and population variance from the probability distribution in the usual way.

Mean and Variance of X

x	1	2	3	4
$p(x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

$$\mu = E(X) = \sum_{\text{all } x}(x \times p(x)) = \frac{5}{2}$$

$$\sigma^2 = V(X) = E(X^2) - (E(X))^2 = \frac{15}{2} - (\frac{5}{2})^2 = \frac{5}{4}$$

Sampling Distribution of $\bar{X}$

- Let's calculate the sample mean for all possible samples of size 2:

		Second roll			
		1	2	3	4
	1	$\bar{x} = 1$	$\bar{x} = 1.5$	$\bar{x} = 2$	$\bar{x} = 2.5$
First	2	$\bar{x} = 1.5$	$\bar{x} = 2$	$\bar{x} = 2.5$	$\bar{x} = 3$
roll	3	$\bar{x} = 2$	$\bar{x} = 2.5$	$\bar{x} = 3$	$\bar{x} = 3.5$
	4	$\bar{x} = 2.5$	$\bar{x} = 3$	$\bar{x} = 3.5$	$\bar{x} = 4$

- Note that each possible sample of size 2 has equal probability of being drawn.

Sampling Distribution of $\bar{X}$

- We can derive the sampling distribution of $\bar{X}$ from the previous table:

$\bar{x}$	1	1.5	2	2.5	3	3.5	4
$p(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

- Because we could completely list all possible samples of size 2, this sampling distribution could be determined *exactly*.

Mean of $\bar{X}$

$\bar{x}$	1	1.5	2	2.5	3	3.5	4
$p(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

$$E(\bar{X}) = \sum_{\text{all } \bar{x}} (\bar{x} \times p(\bar{x})) = \frac{5}{2}$$

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

Variance of $\bar{X}$

$\bar{x}$	1	1.5	2	2.5	3	3.5	4
$p(\bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

$$V(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 = \frac{55}{8} - (\frac{5}{2})^2$$

$$\Rightarrow V(\bar{X}) = \frac{V(X)}{2} = \frac{\sigma^2}{2}$$

Example 2

- What happens when we can't list all possible samples of a certain size?
- One approach is to repeatedly draw samples,calculate sample statistics and approximate the sampling distribution.
- Suppose that $X \sim U(0, 100)$.
- X is continuous, so we clearly cannot list all possible samples (of any size!).
- Let's run a little experiment...

Experiment

1. Generate a sample of data.
2. Display this data in a histogram.
3. Analyze some sample statistics.
4. Repeat with different sample sizes (n).
5. See what happens.

R Code Example

$$n = 20$$

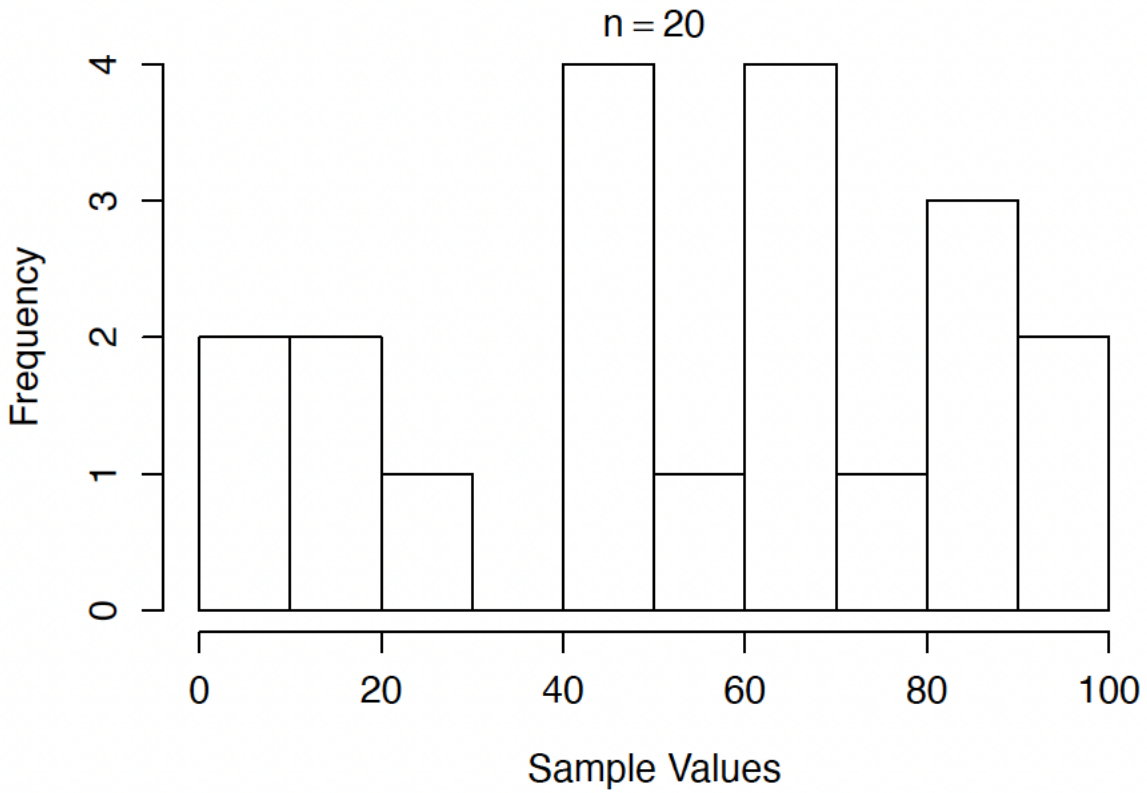
```
> n <- 20  # Sample size.
> unif.data <- runif(n,min=0,max=100)
> hist(unif.data,breaks=10)
> mean(unif.data)
[1] 54.17618
> var(unif.data)
[1] 924.8788
```

R Output Example

$$n = 20$$

```
> unif.data
 [1] 84.157044 47.885988 66.039728
 [4] 53.900445 43.459939 65.143125
 [7] 11.370314  2.755706 47.823901
[10] 65.105942 96.605612 66.360292
[13]  7.824728 88.192319 87.973162
[16] 76.299556 24.684743 96.668611
[19] 10.943103 40.329382
```

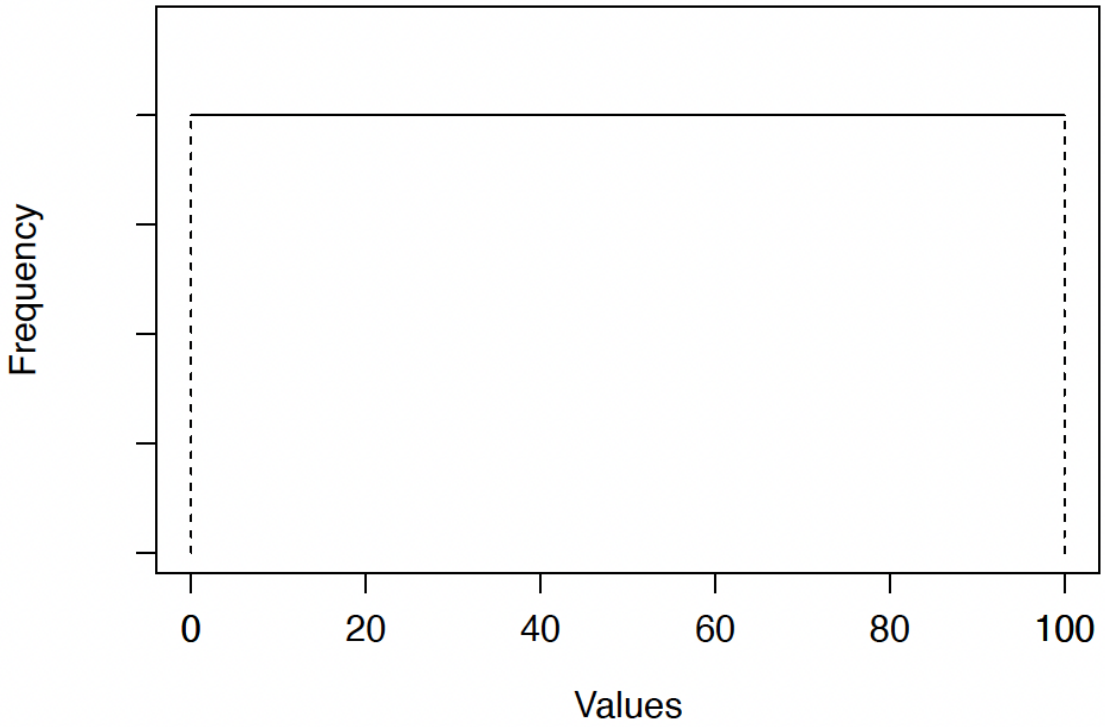
$n = 20$



Sample Mean	54.1762
Sample Variance	924.8788

Histogram

- Why does the histogram not look more like a uniform distribution?



Sample Statistics

- Why is the sample mean not equal (or close) to:

$$E(X) = \frac{a+b}{2} = \frac{0+100}{2} = 50?$$

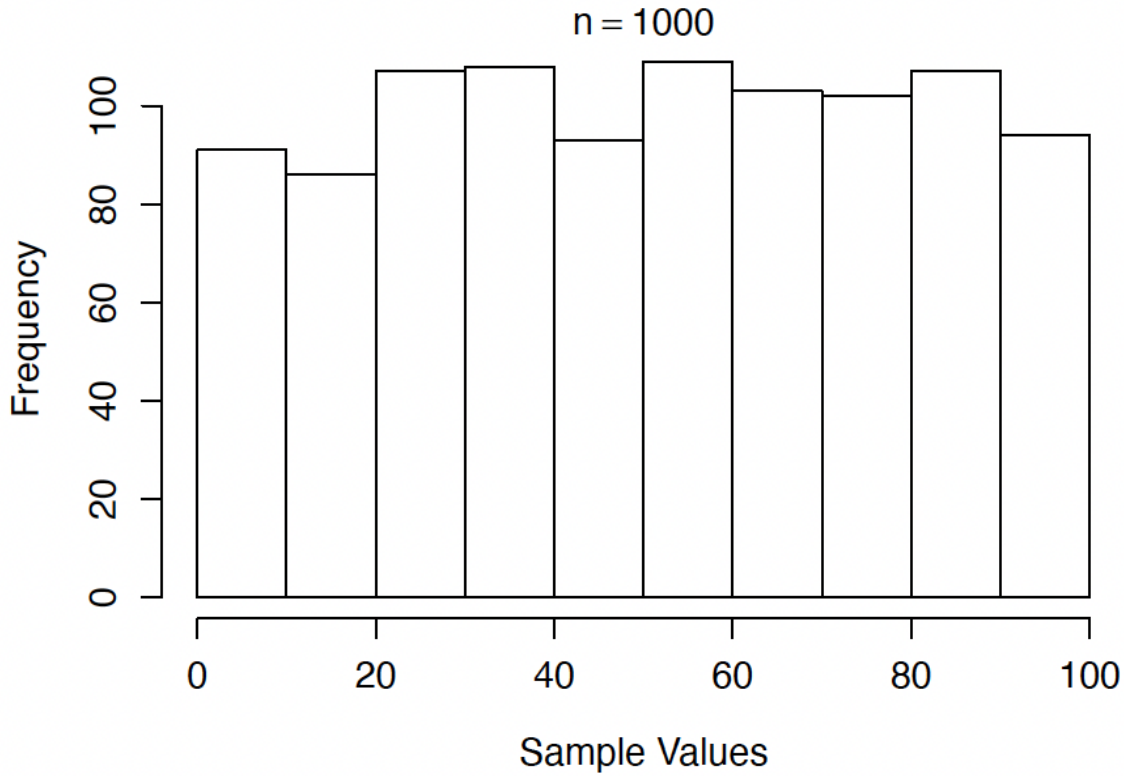
- Why is the sample variance not equal (or close) to:

$$V(X) = \frac{(b-a)^2}{12} = \frac{(100-0)^2}{12} = 833.3333?$$

$n = 1000$

```
> n <- 1000  # Sample size.
> unif.data <- runif(n,min=0,max=100)
> hist(unif.data,breaks=10)
> mean(unif.data)
[1] 50.71992
> var(unif.data)
[1] 802.7182
```

$n = 1000$

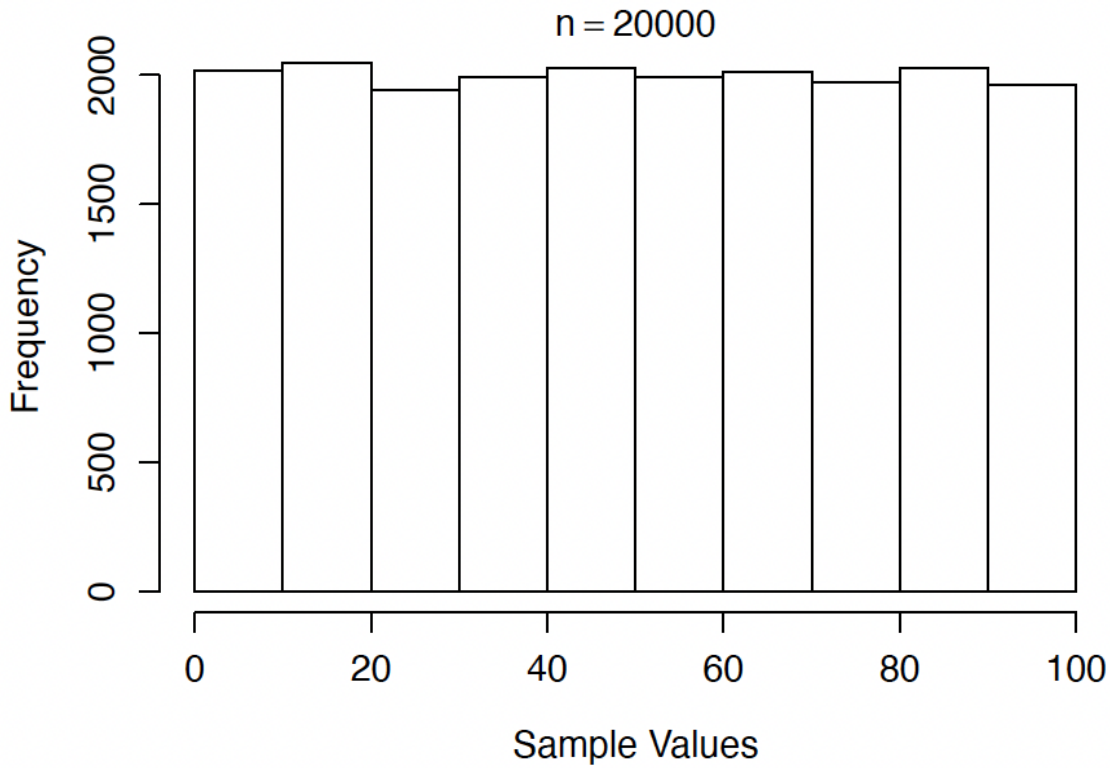


Sample Mean	50.7199
Sample Variance	802.7182

$n = 20000$

```
> n <- 20000  # Sample size.
> unif.data <- runif(n,min=0,max=100)
> hist(unif.data,breaks=10)
> mean(unif.data)
[1] 49.93417
> var(unif.data)
[1] 833.4439
```

$n = 2000$



Sample Mean	49.9342
Sample Variance	833.4439

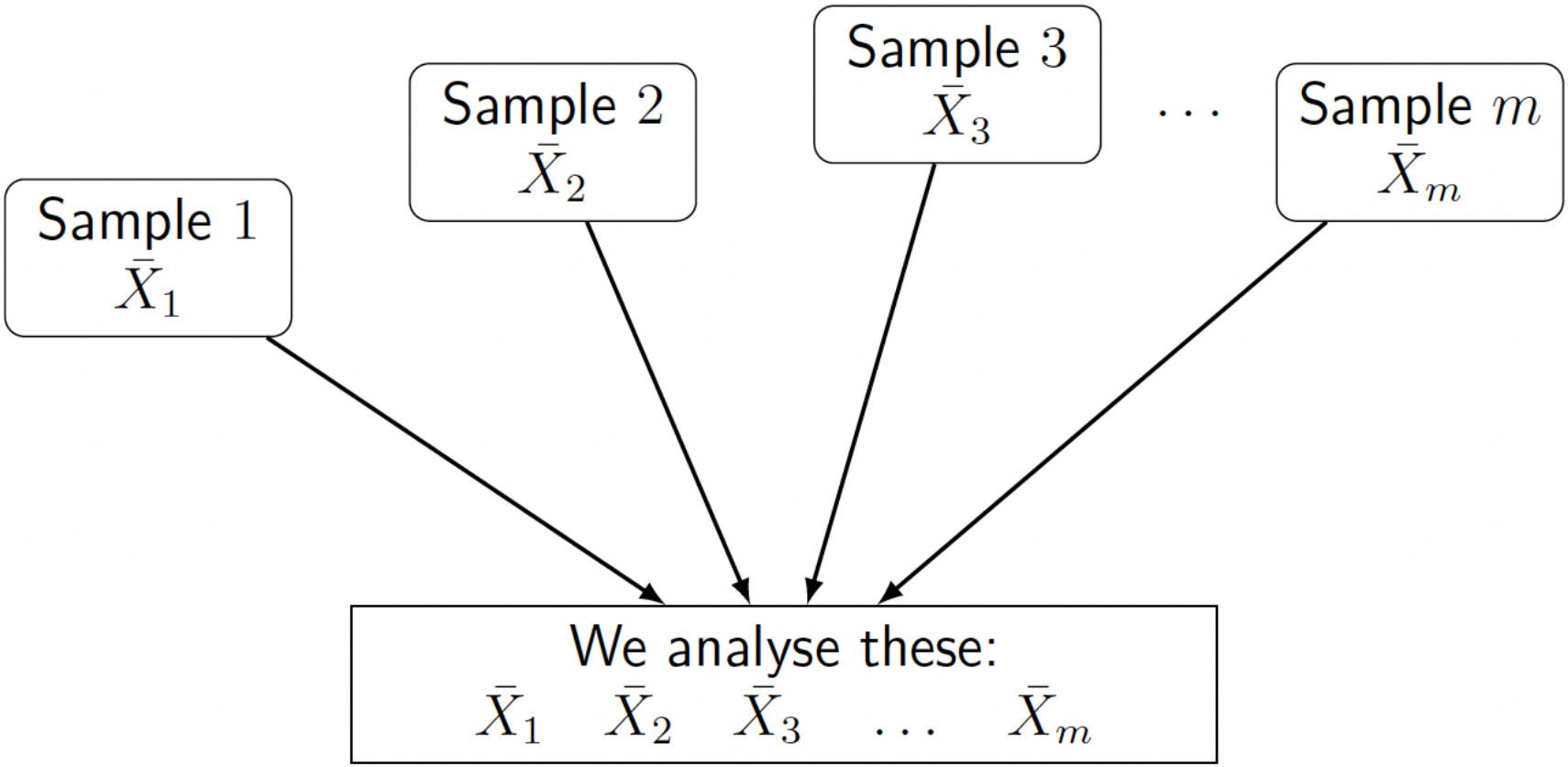
Summary

- As n increases, the histogram of the sample values begins to look more like the distribution of the population (i.e., the uniform distribution).
- As n increases, the sample mean and the sample variance converge towards the population mean and population variance, respectively.

Another Experiment

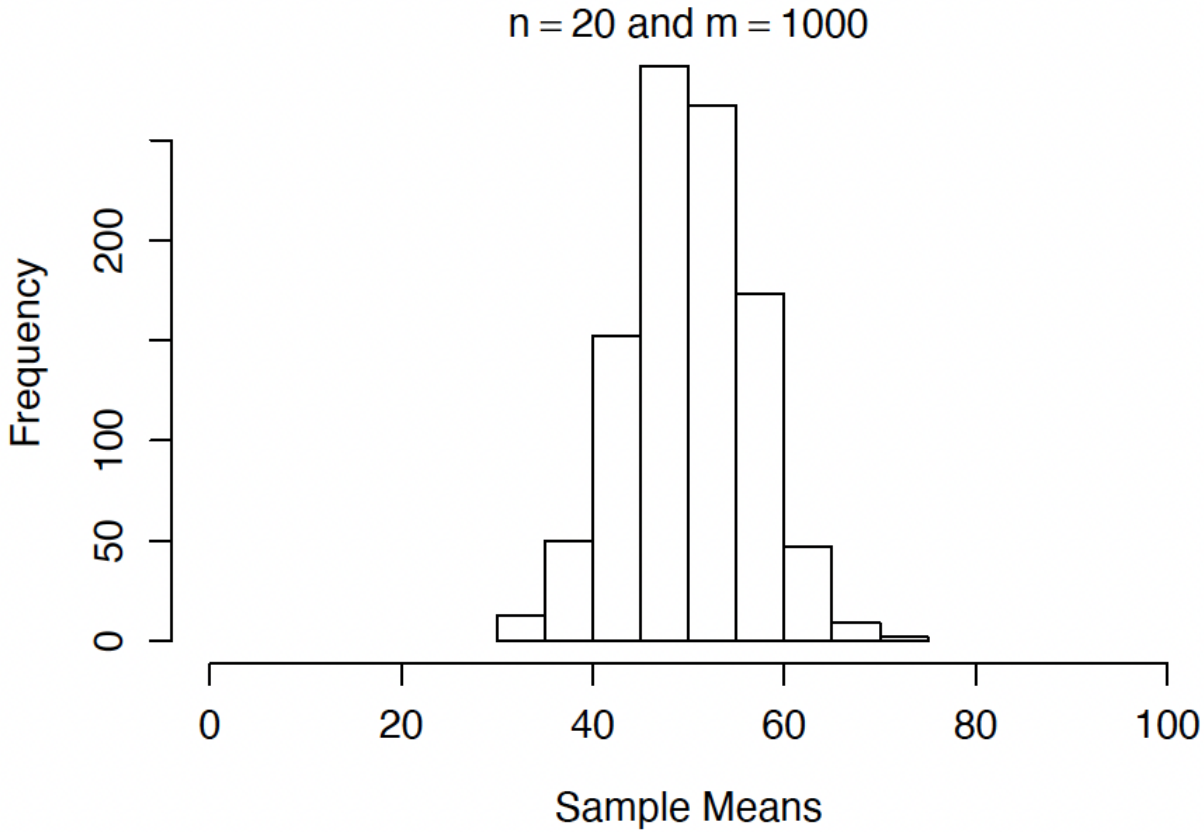
- Let's now redo the previous experiment, but this time generating multiple samples at the same time.
- We'll examine the sample means of the samples and analyze their characteristics (e.g., histograms and sample statistics).
- See what happens when we change vary the sample size n.

Another Experiment



Histogram for Sample Means

$n = 20$ is sample size, $m = 1000$ is number of samples



Mean of Sample Means	49.9882
Variance of Sample Means	42.8857

Summary

- The sampling distribution of the sample mean looks like a normal distribution!
- This phenomenon, called the **Central Limit Theorem**, will occur regardless of the underlying distribution of the population (in this case, a uniform distribution).
- The mean of the sampling distribution of the sample mean is equal to μ, the population mean of X.

Mean of the Sample Mean

- Let $X_1, X_2, \ldots, X_n$ denote our random sample.
- We know that the X_i are independent and that $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ for each i.
- The sample mean is given by:

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

- Therefore, we can use the laws of expected value to calculate $\mu_{\bar{X}} = E(\bar{X})$.

Mean of the Sample Mean

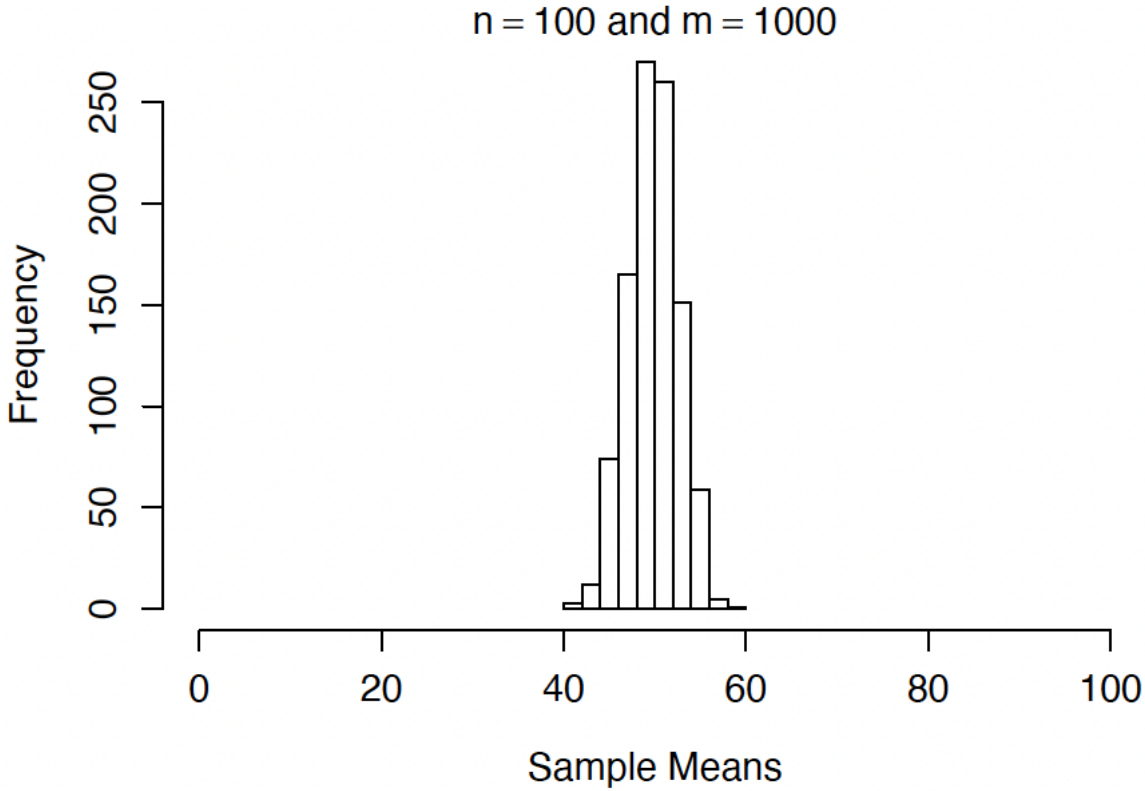
$$
\begin{aligned}
\mu_{\bar{X}} = E(\bar{X}) &= E\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) \\
&= \frac{1}{n}E(X_1 + \ldots + X_n) \\
&= \frac{1}{n}(E(X_1) + \ldots + E(X_n)) \\
&= \frac{1}{n}(\mu + \ldots + \mu) \\
&= \frac{1}{n} \times n\mu \\
&= \mu
\end{aligned}
$$

$n = 100$ and $m = 100$

- Let's see what happens when we increase the sample size (i.e., the number of observations in each sample) from $n = 20$ to $n = 100$.
- We'll leave the number of samples to be $m = 1000$, as before.

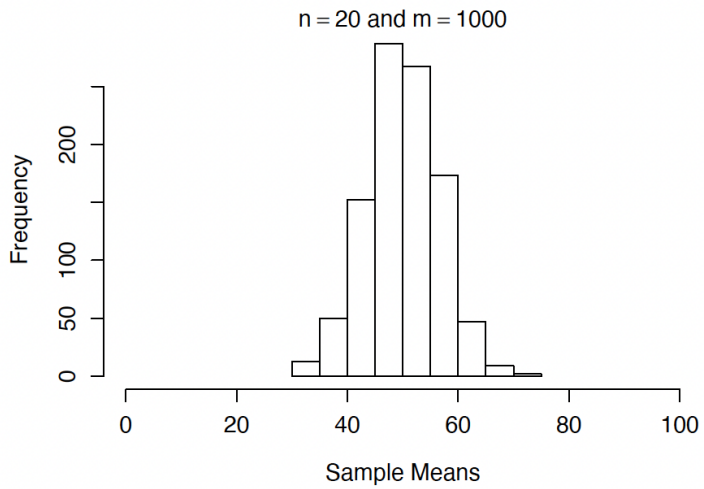
Histogram for Sample Means

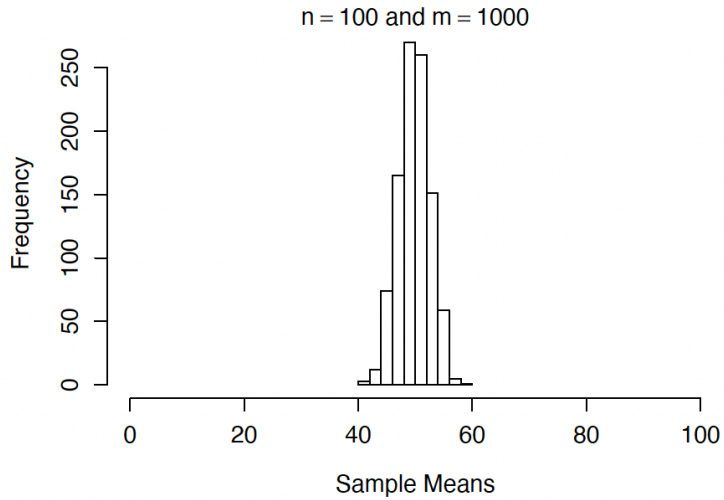
$n = 100$ is sample size, $m = 1000$ is number of samples



Mean of Sample Means	49.8028
Variance of Sample Means	7.5377

Comparison of $n = 20$ to $n = 100$





	$n = 20$	$n = 100$
Mean of Sample Means	49.9882	49.8028
Variance of Sample Means	42.8857	7.5377

Summary

- The sample variance of the sample means gets smaller as the sample size n increases!
- The variance of the sampling distribution of the sample mean is equal to $\frac{\sigma^2}{n}$, the population variance of X divided by n.
- The standard deviation of the sampling distribution of the sample mean is equal to $\frac{\sigma}{\sqrt{n}}$.
- Note: The standard deviation of a sample statistic is called the **standard error (SE)** of the statistic.

Variance of the Sample Mean

$$\sigma^2_{\bar{X}} = V(\bar{X}) = V\left(\frac{1}{n}(X_1 + \ldots + X_n)\right)$$

$$= \frac{1}{n^2}V(X_1 + \ldots + X_n)$$

$$= \frac{1}{n^2}(V(X_1) + \ldots + V(X_n))$$

$$= \frac{1}{n^2}(\sigma^2 + \ldots + \sigma^2)$$

$$= \frac{1}{n^2} \times n\sigma^2$$

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

Central Limit Theorem

- Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2 (i.e., the X_i's are independent and $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ for $i = 1, \ldots, n$). Then:

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu, \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\right) \quad \text{as} \quad n \to \infty$$

- Which means if we standardize:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z \sim N(0,1) \quad \text{as} \quad n \to \infty$$

How Large Should n Be?

- Generally, it depends on the population distribution of X.
  - If X has a normal distribution, then the sample mean has a normal distribution for *all* sample sizes.
  - If X has a distribution that is close to normal, the approximation is good for small sample sizes (e.g., $n \approx 20$).
  - If X has a distribution that is far from normal, the approximation requires larger sample sizes (e.g., $n > 50$).

Example 1

- Consider the marks of all students who took an economics test. If marks are normally distributed, with mean equal to 72 and standard deviation equal to 9, find:

(a) The probability that any one student will have a mark over 78.
(b) The probability that a sample of 10 students will have an average mark over 78.

Solution - Part (a)

- Let X be the mark of a randomly selected student.
- Then $X \sim N(\mu = 72, \sigma^2 = 9^2)$.

$$
\begin{aligned}
P(X > 78) &= P\left(\frac{X - \mu}{\sigma} > \frac{78 - 72}{9}\right) \\
&= P(Z > 0.67) \\
&= 1 - P(Z < 0.67) \\
&= 1 - 0.7486 \\
&= 0.2514
\end{aligned}
$$

Solution - Part (a)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

Solution - Part (b)

- Let $\bar{X}$ be the average mark of the sample of 10 students.
- By CLT, $\bar{X} \sim N\left(\mu_{\bar{X}} = \mu = 72, \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{9^2}{10}\right)$.

$$
\begin{aligned}
P(\bar{X} > 78) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{78 - 72}{\frac{9}{\sqrt{10}}}\right) \\
&= P(Z > 2.11) \\
&= 1 - P(Z < 2.11) \\
&= 1 - 0.9826 \\
&= 0.0174
\end{aligned}
$$

Solution - Part (b)

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952

Comparing Parts (a) and (b)

- Why does the probability drop from 25.14% to 1.74%?
  - The standard deviation of sample means (or the *standard error*) is always smaller than the standard deviation of single observations.
- Interpretation:
  - Sample means have less variability than single observations.
- E.g., much more likely to find an individual who is very smart, than a random sample of 10 students who are all very smart.