Lecture 14
Sampling distribution and Central limit theorem

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Statistical inference

- The goals of **statistical inference**:
  - How to draw conclusions about a population using only a subset of the population
  - How to estimate population parameters and quantify the amount of confidence we can place in them
  - How to use this information to make decisions/conclusions

- **Notation:**
  - **Population** - Entire group of items/individuals we want information about
  - **Sample** - The part of the population we actually examine in order to gather information
  - A **parameter** is a number that describes the population.
  - A **statistic** is a number that describes a sample.

Statistical inference

- Parameter and statistic
  - A **parameter** is a fixed number, but we do not know its actual value.
  - The value of a **statistic** is known after we take a sample, but it can vary from sample to sample.
  - We often use a **statistic** to ***estimate*** an unknown population **parameter**.

Populations and parameters

- Population parameters: parameters to model for a population
- Sample statistics (or statistics): summaries of sample data to estimate the population parameters

Name	Statistic	Parameter
Mean	$\bar{y}$	μ (mu, pronounced "meeoo," not "moo")
Standard deviation	s	σ (sigma)
Correlation	r	ρ (rho)
Regression coefficient	b	β (beta, pronounced "baytah"[7])
Proportion	$\hat{p}$	p (pronounced "pee"[8])

Statistical inference

- **Parameter and statistic**
  - A **parameter** is a fixed number, but we do not know its actual value.
  - The value of a **statistic** is known after we take a sample, but it can vary from sample to sample.
  - We often use a **statistic** to ***estimate*** an unknown population **parameter**.

- Statistical inference of **proportion** and **mean**
  - **Statistical inference** draws conclusions about a population on the basis of data from a sample.
  - It also provides us with a statement of how much **confidence** we can place in our conclusions.
  - We are in particular interested in what **proportion** of the population has a certain opinion or trait, as well as the **mean** value a certain variable takes in the population.

Sampling distribution of a proportion

- What proportion of Korean residents support the current President? (approval rating 지지율)
  - Randomly sample 1000 people and ask them if they support the current President (remember? simple random sampling)
  - Let's say you found 690 (i.e., 69%) of 1000 people showed supports.

  - The proportion of the population is denoted p (Parameter).
  - The proportion of the sample is written $\hat{p}$ (Statistic).
  - The number 690 is the count (Statistic).

Sampling distribution of a proportion

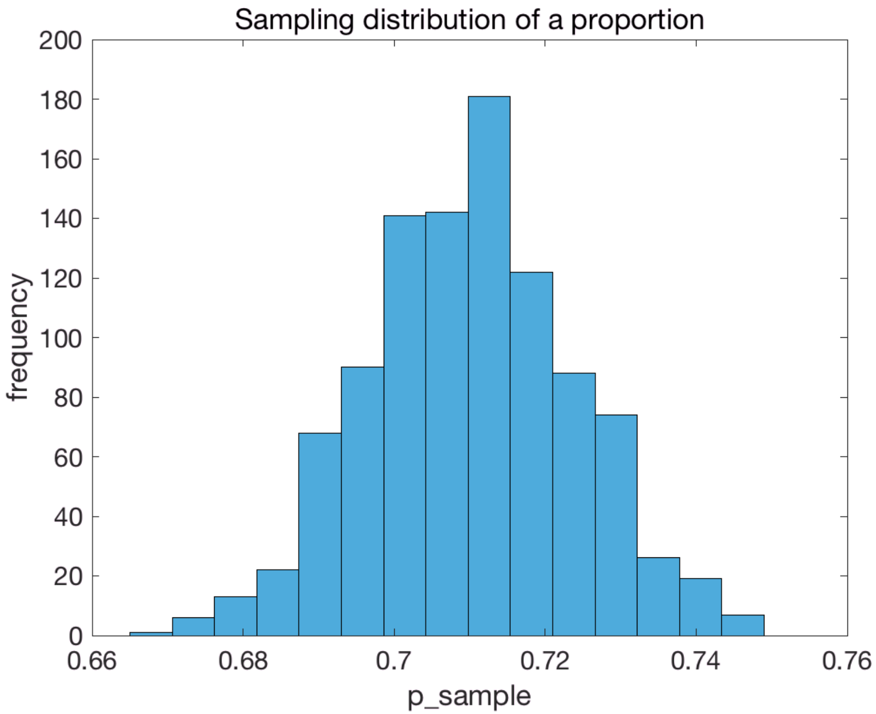
- p is unknown, but $\hat{p} = 0.69$.
- We can use $\hat{p}$ to estimate p.
- But the other survey could show 0.72, and the other survey may show 0.68, etc.
- Let's simulate the whole Korean population (51.25M).

Simulation of simple random sampling and sampling distribution

```
p = 0.71; % This is a parameter for the real approval rating.
          % This value is unknown in reality, but let's assume that I'm God, and so I know it.

population = zeros(51250000,1); % Korean population 51.2M
population(1:round(51250000*p)) = 1; % People who approve the current president

% Simple random sampling

for i = 1:1000 % let's say we did the survey 1000 times

    sample = population(randperm(numel(population), 1000)); % simple random sampling
    p_sample(i,1) = sum(sample==1)/numel(sample); % get the sample statistics for the approval rating

end
```

Sampling distribution of a proportion

```
histogram(p_sample, 15);
set(gca, 'fontsize', 15);
title('Sampling distribution of a proportion');
xlabel('p\_sample');
ylabel('frequency');
```




```
fprintf('\nMean = %2.2f, Standard deviation = %2.2f', mean(p_sample), std(p_sample));
```

```
Mean = 0.71, Standard deviation = 0.01
```

- This is a “sampling distribution”.
- This is different from the sample distribution, which refers to the distribution of the sample, a display of the actual data.

Sampling distribution of a proportion

- In practice, we only sample the population once. However, we want to understand what would happen if we sampled the population *repeatedly*.
- We want to know what **values** the statistic can take and how often it takes them, i.e. we want to know the distribution of the statistic.
- This will give us an indication of how well a statistic estimates a parameter.

- The sampling distribution of a statistic is the distribution of its value in all possible samples of the same size from the same population.
- We use mathematical models (e.g. the normal model) to understand the sampling distribution.
- We say a statistic (used to estimate a parameter) is unbiased when the mean of its sampling distribution is equal to the true value of the parameter.

Sampling distribution of a proportion: Normal model

- To use a Normal model to model the sampling distribution, we need two parameters, *mean* and *standard deviation*.

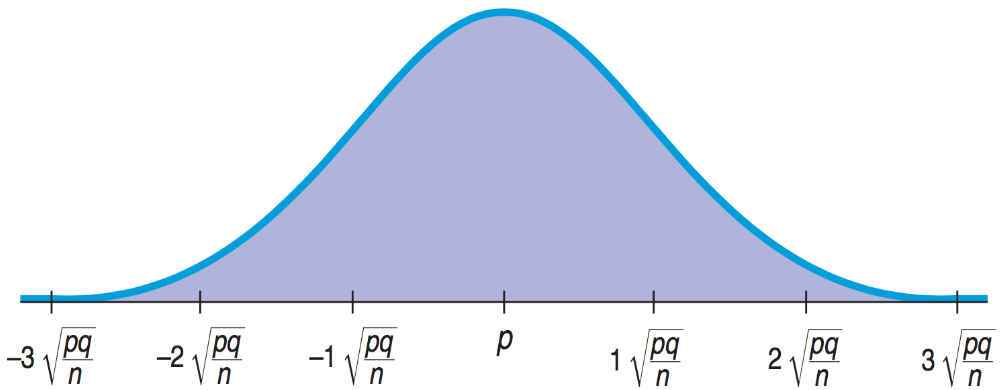
- In the simulation, we know the mean, *p*, but we cannot know the standard deviation until we do simulation. How can we know the standard deviation mathematically?

- From the Binomial model, we know the standard deviation of the *number* of an outcome (e.g., success) is $\sqrt{npq}$, and here, we want to know the standard deviation for the *proportion* of the outcome, which should be the value divided by *n*.
  - $\sigma(\hat{p}) = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$
  - The sampling distribution of a proportion can be modeled with a Normal model, $N(p, \sqrt{\frac{pq}{n}})$

Sampling distribution of a proportion: Normal model

- In our simulation, p = .71, n = 1000, SD = $\sqrt{\frac{0.71 \times (1-0.71)}{1000}}$ = 0.0143

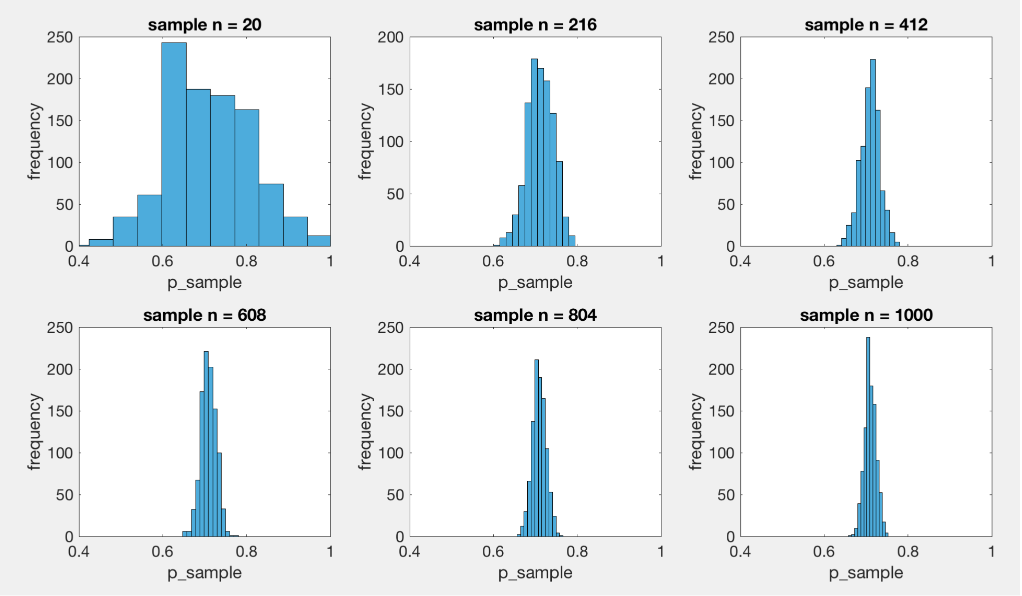
$$N\left(p, \sqrt{\frac{pq}{n}}\right).$$



Sampling distribution of a proportion: Normal model

Simulation of simple random sampling and sampling distribution

```
p = 0.71; % This is a parameter for the real approval rating.
          % This value is unknown in reality, but let's assume that I'm God,

population = zeros(51250000,1); % Korean population 51.2M
population(1:round(51250000*p)) = 1; % People who approve the current preside

% Simple random sampling
iter = 1000;
sample_n = linspace(20, 800, 6);
for i = 1:numel(sample_n)
    for j = 1:iter % let's say we did the survey 1000 times
        sample = population(randperm(numel(population), sample_n(i))); % simp
        p_sample{i}(j) = sum(sample==1)/numel(sample); % get the sample stat
    end
end
```

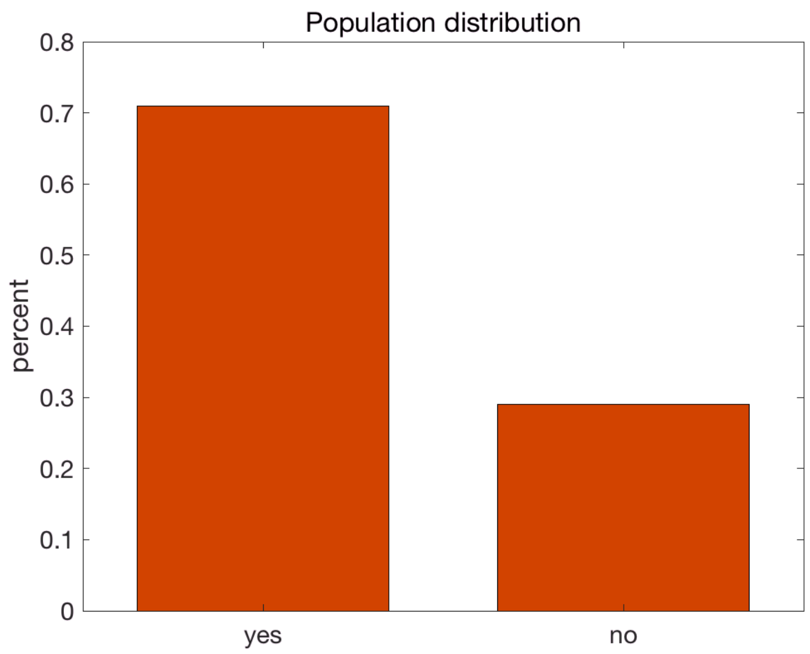



```
For sample n = 20, Mean = 0.71, Standard deviation = 0.1035, SD from a normal model = 0.1015
For sample n = 216, Mean = 0.71, Standard deviation = 0.0308, SD from a normal model = 0.0309
For sample n = 412, Mean = 0.71, Standard deviation = 0.0230, SD from a normal model = 0.0224
For sample n = 608, Mean = 0.71, Standard deviation = 0.0179, SD from a normal model = 0.0184
For sample n = 804, Mean = 0.71, Standard deviation = 0.0164, SD from a normal model = 0.0160
For sample n = 1000, Mean = 0.71, Standard deviation = 0.0138, SD from a normal model = 0.0143
```

Population distribution

- The population distribution is the probability model derived from information on all members of the population.
- Any individual chosen at random from this population will follow the same probability model.
- In the example of approval rating, the **population distribution** is the distribution of Yes's and No's in the population, and is described by the parameter p.

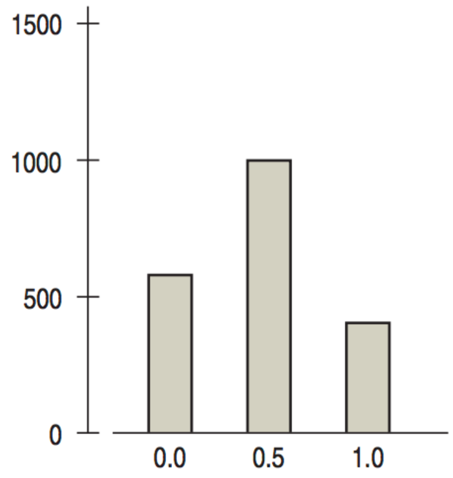
```
% population distribution

bar([.71, 1-.71], .7, 'facecolor', [0.7608 0.3020 0]);
set(gca,'XTickLabel', {'yes', 'no'}, 'fontsize', 15)
title('Population distribution')
ylabel('percent');
```



When does the Normal model work for sampling distribution of a proportion?

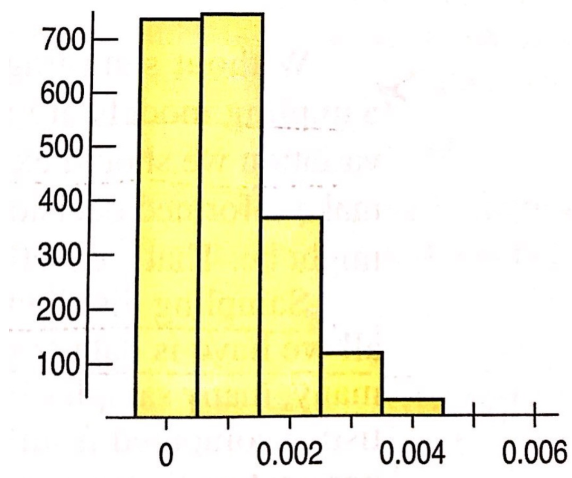
- Sampling distribution

  with sample size n = 2



- Sampling distribution when p = 0.001

  with sample size n = 1000

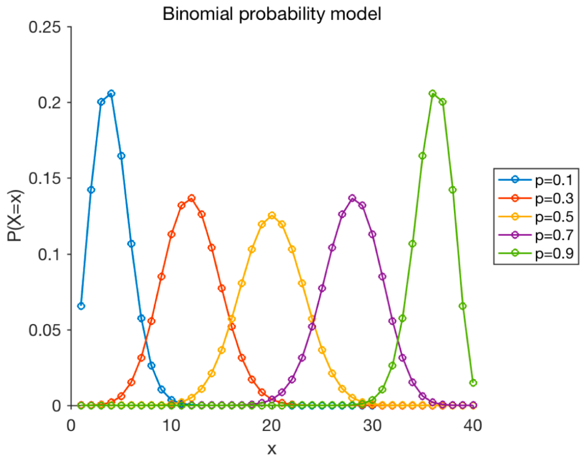


When does the Normal model work for sampling distribution of a proportion?

- Remember…

Approximating the binomial with a normal model

- Plot for Binom(n, p), where n = 40



- Some of them look exactly like the normal distribution.
- Usually, when $np \geq 10$ and $nq \geq 10$, the binomial model is approximately Normal, which can simplify the calculation of the probability.

Assumptions and Conditions

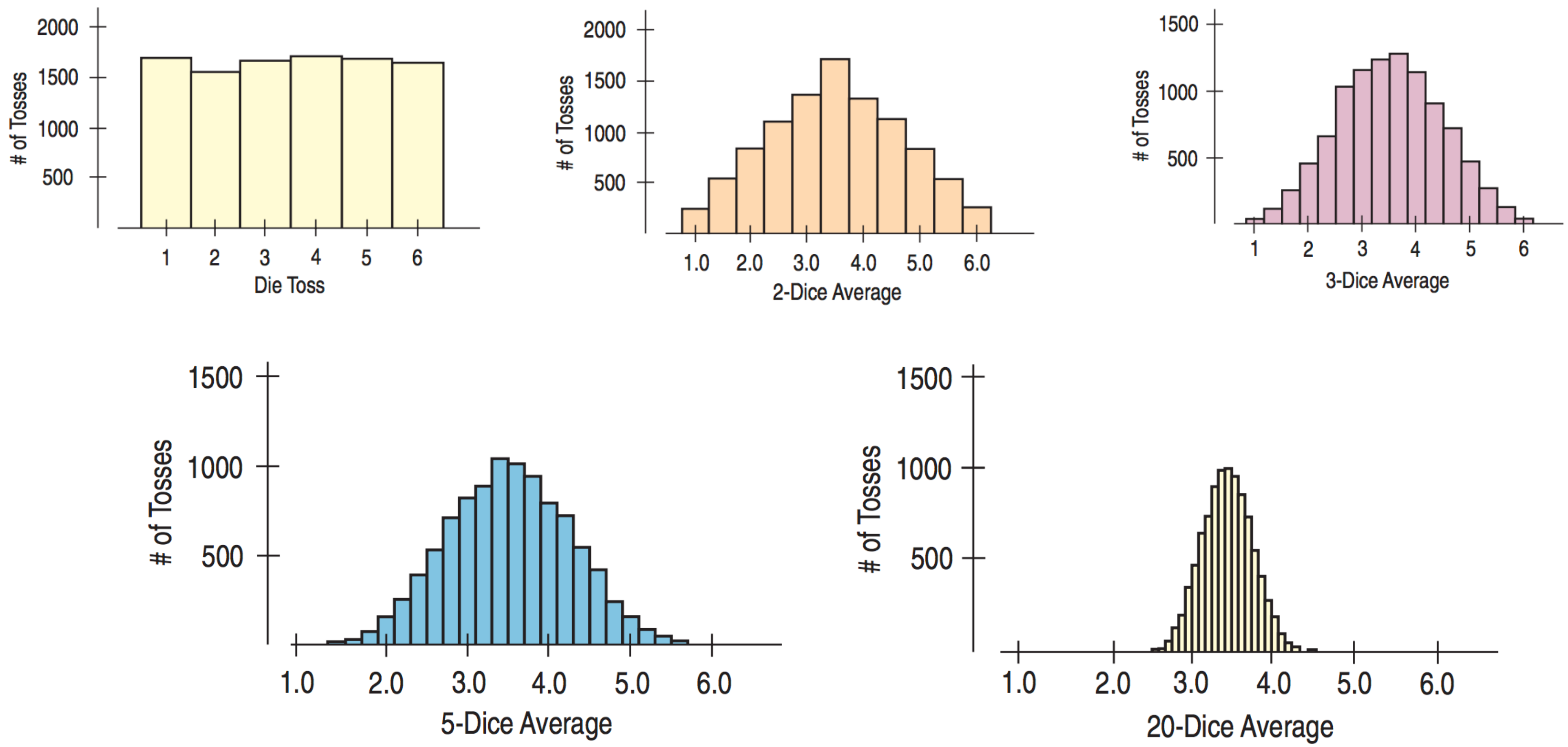
- **Independence Assumption:** The individuals in the samples must be independent of each other.

- We can't know if this assumption is true or not for sure, but we can check the following ***conditions*** that provide information about the assumption.

- Conditions:
  - **Randomization Condition:** random assignment (experiment), random sampling (survey)
  - **10% Condition:** The sample size, *n*, must be no larger than 10% of the population. If you sample more than about 10% of the population, the remaining individuals are no longer truly independent of each other. The sampling distribution will have a smaller standard deviation.
  - **Success/Failure Condition:** The sample size has to be big enough so that we expect at least 10 successes and at least 10 failures (*np* and *nq* should be > 10)

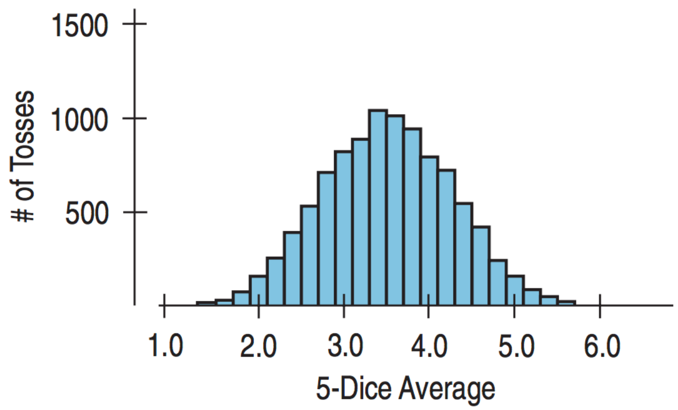
Sampling distribution of a Mean

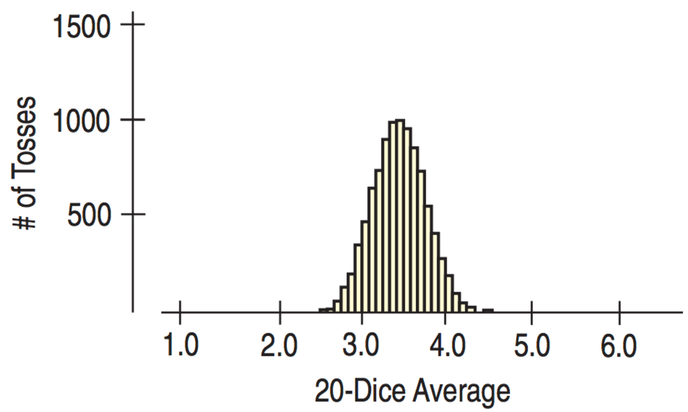
- A simple simulation: If we toss one fair die 10,000 times, what should the histogram of the numbers on the face of the die look like?



Sampling distribution of a Mean

- A simple simulation: If we toss one fair die 10,000 times, what should the histogram of the numbers on the face of the die look like?
- Sample size (=number of dice) gets larger, each sample average is more likely to be close to the population mean.
- And we see the Normal shape clearly, and the spread becomes smaller.





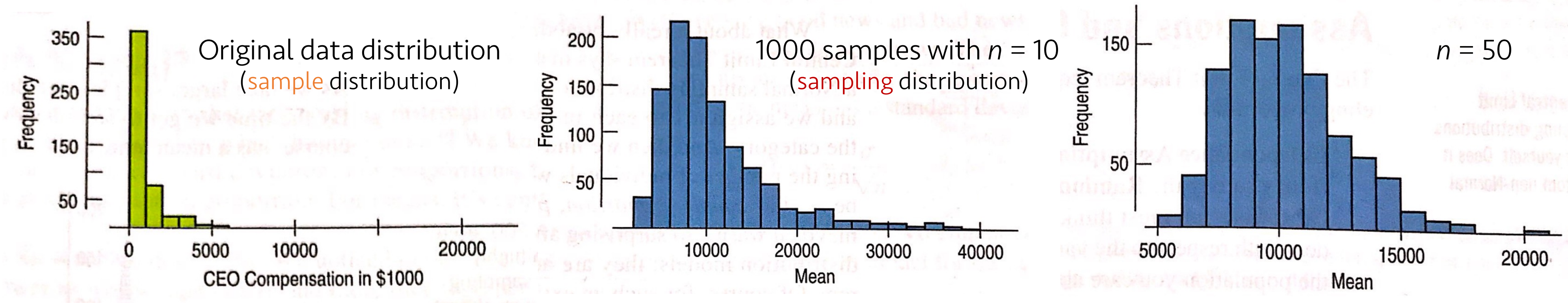
Central Limit Theorem

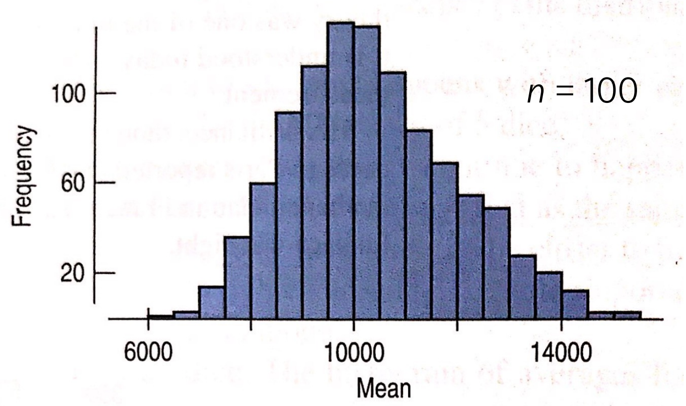
- What we saw with dice simulation is true for means for repeated samples for almost every situation.

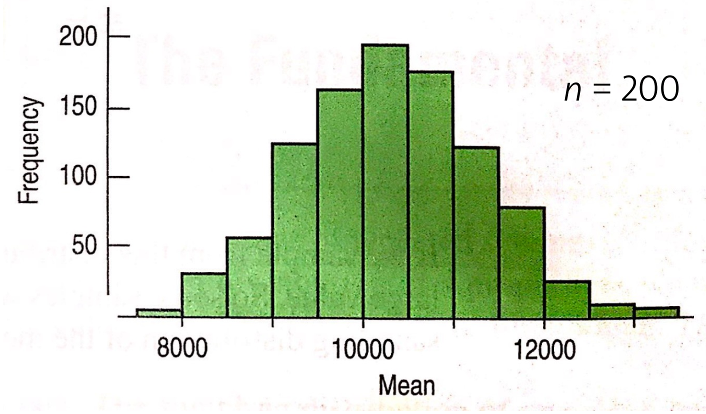
- **Central Limit Theorem:**
  - Definition: The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.
  - Only assumption: the *independence* assumption
  - **This works no matter how the data are distributed.**
  - This is proved by Pierre-Simon Laplace in 1810
  - This is one of the most fundamental theorem of Statistics.

Central Limit Theorem

- Important fact: *it works regardless of the shape of the population distribution!* Even if we sample from a skewed or bimodal population...
- Example: the CEO compensation data:





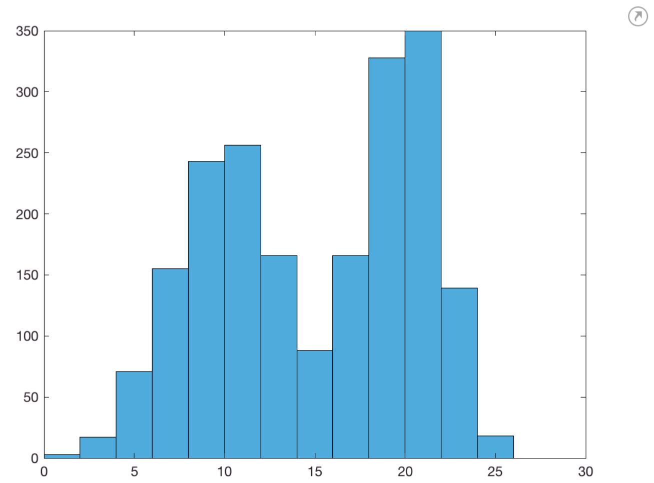


Central Limit Theorem

- Important fact: *it works regardless of the shape of* from a skewed or bimodal population…
- Example: bimodal distribution

Simulate the central limit theorem with bimodal data

```
% data generation
a = [normrnd(10,3,1000,1);normrnd(20,2,1000,1)];
histogram(a);
```

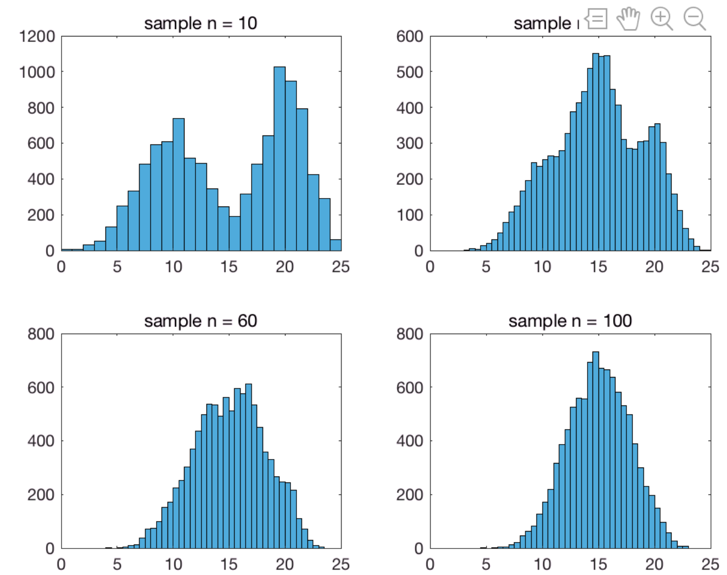


random sampling

```
% random sampling n = 10, 30, 60, 100

sample_n = [10 30 60 100];
for iter = 1:10000
    for i = 1:numel(sample_n)
        m{i}(iter,1) = mean(a(randperm(numel(a),i)));
    end
end
```

figure

```
for i = 1:numel(m)
    subplot(2,2,i);
    histogram(m{i});
    set(gca, 'xlim', [0 25]);
    title(['sample n = ' num2str(sample_n(i))]);
end
```



Which Normal?

- Again, to use a Normal model to model the sampling distribution, we need two parameters, *mean* and *standard deviation*.

- **Mean:** the sampling distribution is centered at the population mean, μ.
- **Standard deviation:** As we saw in the sampling distribution of a proportion, the standard deviation gets smaller as we average more and more samples.
  - How much smaller?
    - Good news: the standard deviation falls as the sample size grows.
    - Bad news: it doesn't drop as fast as we might like. It only goes down by the square root of the sample size.
- $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population.

Review: Which Normal?

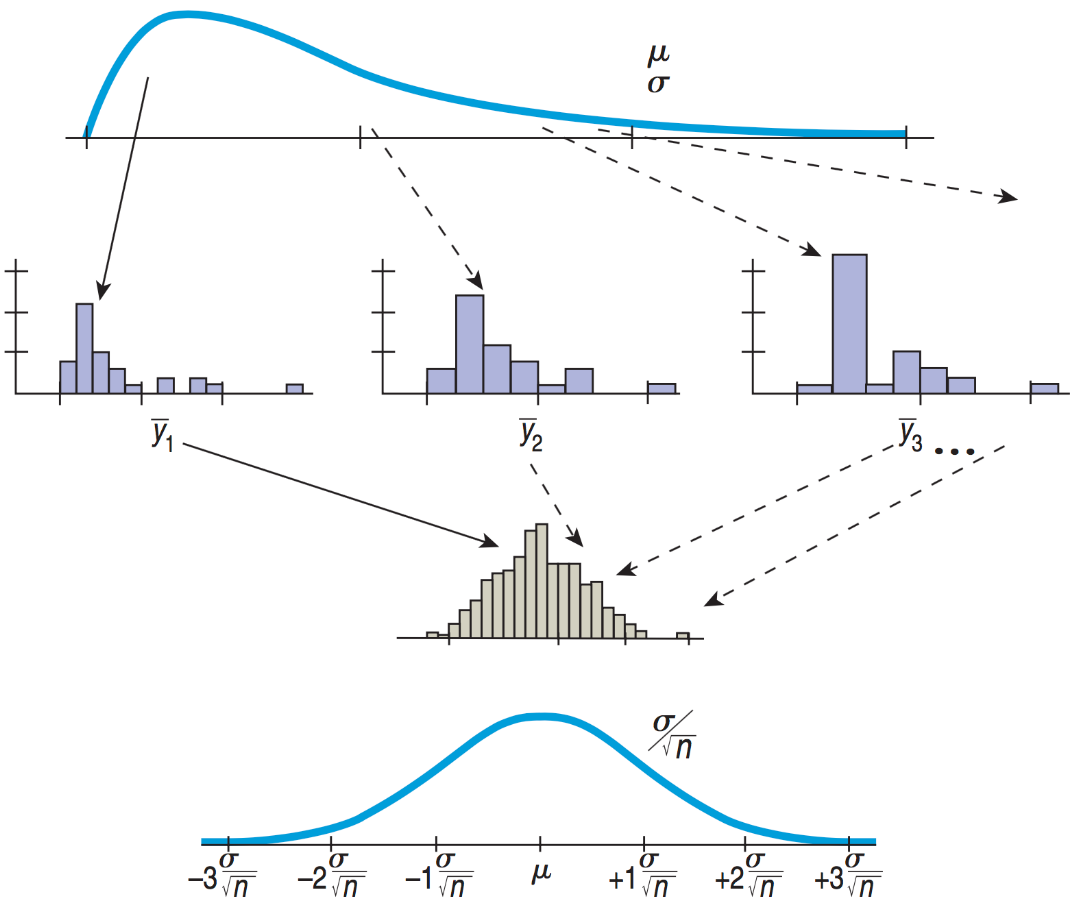
- For categorical data, sample proportion, $\hat{p}$
  - $Mean(\hat{p}) = p, SD(\hat{p}) = \sqrt{\frac{pq}{n}}$

- For quantitative data, sample mean, $\bar{y}$
  - $Mean(\bar{y}) = \mu, SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$

Let’s summarize:

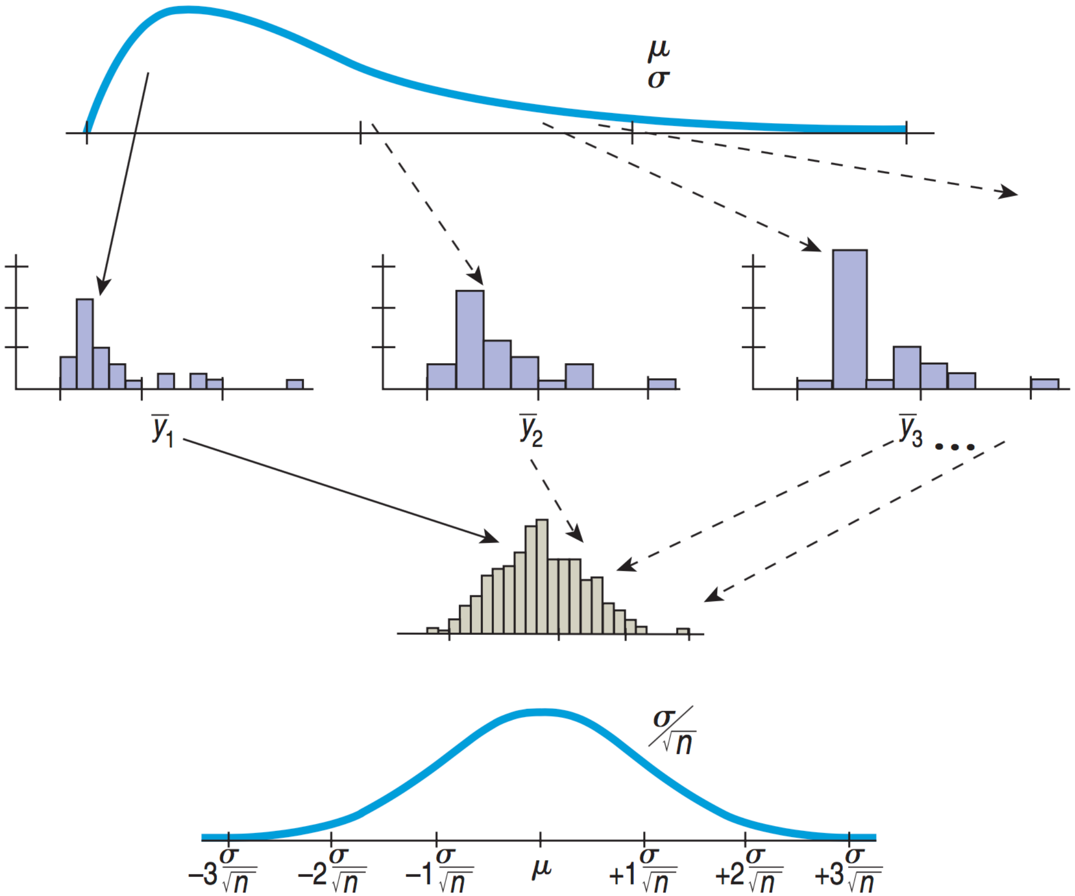
- The statistic itself is a *random variable*.
- A different random sample would have given a different statistic results.
- This sample-to-sample variability is what generates the sampling distribution.
- Fortunately, for the mean and the proportion, the **Central Limit Theorem** tells us that we can model their sampling distribution directly with a Normal model.

Let's summarize:



- Population model, that we cannot observe.
- We draw one real sample (solid line) of size n and show its histogram and summary statistics.
- We imagine (or simulate) drawing many other samples (dashed lines).
- We (imagine) gathering all the means into a histogram.
- The CLT tells us we can model the shape of this histogram with a Normal model, $N(\mu, \frac{\sigma}{\sqrt{n}})$

But… can we actually know the mean and SD of the sampling distribution?



- Population model, that we cannot observe.
- We draw one real sample (solid line) of size n and show its histogram and summary statistics.
- We imagine (or simulate) drawing many other samples (dashed lines).

No.. Then, what should we do?

To be continued…

- We (imagine) gathering all the means into a histogram.
- The CLT tells us we can model the shape of this histogram with a Normal model, $N(\mu, \frac{\sigma}{\sqrt{n}})$

Key Points

Chapter 18: Sampling Distribution Models

- A parameter is a number that describes the population.
- A statistic is a number that describes a sample.
- The sampling distribution of a statistic is the distribution of its value in all possible samples of the same size from the same population.
- **Central Limit Theorem:** The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.
- This works no matter what the original data's distribution is.
- The sampling distribution of a proportion can be modeled with a Normal model, $N(p, \sqrt{\frac{pq}{n}})$
- The sampling distribution of a mean can be modeled with a Normal model, $N(\mu, \frac{\sigma}{\sqrt{n}})$