SJSU SAN JOSÉ STATE UNIVERSITY

DATA 220
Mathematical Methods for Data Analytics

Dr. Mohammad Masum

Probability Distributions

- A random variable is a numerical outcome of a probability experiment (usually denote by capital letters; e.g., X, Y, Z)
  - If the random variables possible values can be counted or listed, then it is Discrete random variable
    - E.g., number of siblings, number that comes up on a roll of a die
  - If the random variables can take any value in an interval, then it is continuous random variable
    - e.g., the height of a randomly
- A probability distribution provides the probabilities of occurrences of different possible outcome for an experiment

Probability model

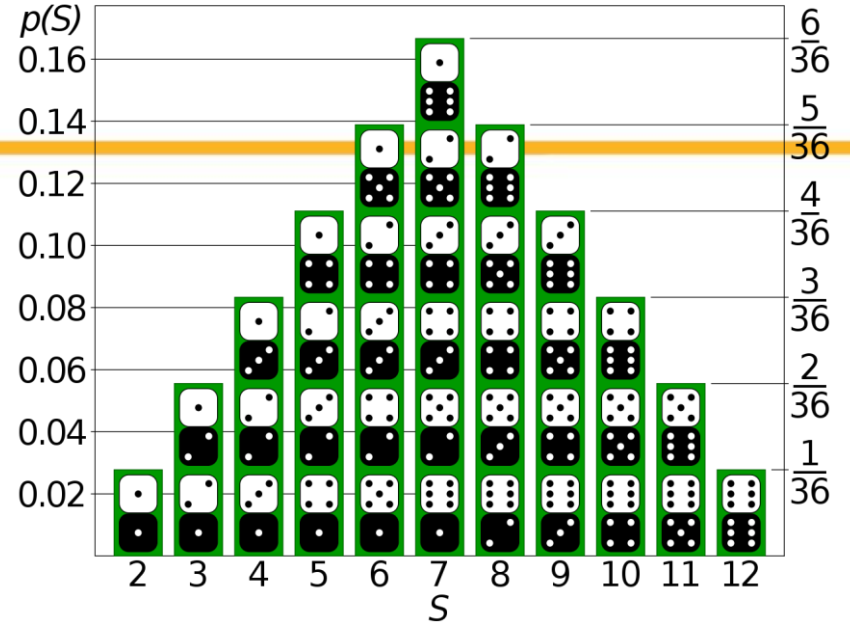
Event	HH	HT	TH	TT
Prob.	.25	.25	.25	.25

Probability distribution for random variable X = #heads

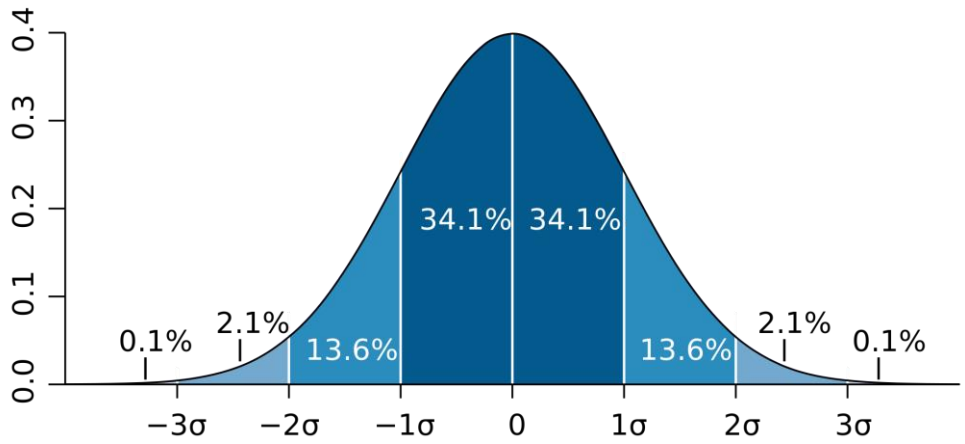
X	0	1	2
P(X)	.25	.5	.25

Probability Distributions

- Probability Mass Function- describe probabilities for a set of exclusive discrete events
  - E.g., coin toss, dice roll, number of people who visits a website in a day, the number of patients who show up at a clinic in an hour, the number of students who attend a class on a given day
- Probability Density (Distribution) Function – describe probabilities for a set of continuous events
  - E.g., IQ score, height, time between earthquakes, rainfall, insurance claim, waiting time



The PMF for probability distribution for the sum *S* of counts from rolling two dice



The PDF for normal distribution

https://en.wikipedia.org/wiki/Probability_distribution

Probability Distributions

- Properties of **PMF**: PMF of a discrete random variable X:
  - PMF is non-negative: $P(X = x) = f(x) > 0, if\ x \in S$
  - PMF sums to one: $\sum_{x \in S} f(x) = 1$
  - PMF is defined for discrete random variable only

Probability Distributions

- Properties of **PDF**: PDF of a continuous random variable X is an integrable function satisfying:
  - PDF is non-negative: $f(x) > 0$
  - PDF integrates to one: The area under the curve is 1
  - PDF is defined for continuous random variable only

Probability Distributions

- Cumulative distribution function (CDF) of a RV –
  - maps the values of the RV to their cumulative probabilities
  - gives the probability that the RV is less than or equal to a specific value

Discrete

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t)$$

Probability mass function

Cumulative distribution function

Probability density function

Continuous

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Probability Distributions

Example: Assume, X is a continuous random variable with pdf $f(x) = 3x^2 \;;\; 0 < x < 1$

Find the $P\left(\frac{1}{2} < x < 1\right) = ?$

Probability Distributions

```
1 from scipy import stats
2 import numpy as np
3 import matplotlib.pyplot as plt
```

```
1 # X is a random variable
2 X = stats.norm(0,1)
```

```
1 X.pdf(3) # P(X = 3)
```

```
0.004431848411938007 5
```

```
1 X.cdf(3) # P(X <= 3)
```

```
0.9986501019683699
```

```
1 X.mean()
```

```
0.0
```

```
1 X.var()
```

```
1.0
```

```
1 X.std()
```

```
1.0
```

```
1 X.rvs() # generate a random sample from X
```

```
-0.1971801818211948 4
```

```
1 X.rvs(10) #generate 10 random sample from X
```

```
array([ 1.06504812, -1.6329481 , -0.47913653, -0.81573339, -0.75866539,
       -0.82628638,  0.3432747 , -0.36121554,  0.78722887, -1.23467131])
```

Expected Value and Variance of a RV (Discrete)

- The expected value of a RV is denoted by $E[X]$
- The expected value is the sum, over all possible values of x, of x times its probability:

$$E[X] = \sum_{all\ possible\ x} x\ P(X = x)$$

- Example: expected value of a roll of a die is:
- The variance of a RV is denoted by $Var[X] or\ \sigma_X^2$ and defined by

$$\sigma_X^2 = \mathrm{E}\left[(\mathrm{X} - \mu_X)^2\right] = \sum_{all\ possible\ x} (x - \mu_X)^2\ P(X = x)$$

- Find the variance and std for rolling one die:

Probability Distributions – Bernoulli

- **A Bernoulli distribution** is the discrete probability distribution of a random variable with one of two possible outcomes: success (*p*) or failure (*1-p*)
  - E.g., will you get a A+ in this course? Tossing a coin; a team will win a championship or not? A rolled die will show 6 or others
- Assumptions
  - A single trail
  - Two possible outcomes:
    - *P(success) = p*
    - *P(failure) = 1-p*

pmf of the Bernoulli distribution

$$P(X = x) = p^x(1-p)^{1-x}$$
$$for\ x = 0\ or\ 1$$

Probability Distributions – Bernoulli

pmf of the Bernoulli distribution

$$P(X = x) = p^x(1-p)^{1-x}$$
$$for\ x = 0\ or\ 1$$

The prevalence of a certain disease in the general population is 10%. If we randomly select a person from this population, we can have only two possible outcomes (diseased or healthy person). We call one of these outcomes (diseased person) success and the other (healthy person), a failure.

The probability of success (p) or diseased person is 10% or 0.1. So, the probability of failure (q) or healthy person = 1-p = 1-0.1 = 0.9.

Probability Distributions – Bernoulli

pmf of the Bernoulli distribution

$$P(X = x) = p^x(1-p)^{1-x}$$
$$for\ x = 0\ or\ 1$$

$$\text{mean}, \mu = p\ \&\ var, \sigma^2 = p(1-p)$$

Use case: Bernoulli distribution is the foundation for many other distributions, and it is particularly useful for modeling binary events

-- Modeling click-through rates: In online advertising, a Bernoulli distribution can be used to model the probability that a user clicks on an ad, which can be used to optimize ad placement and targeting.

Probability Distributions – Binomial

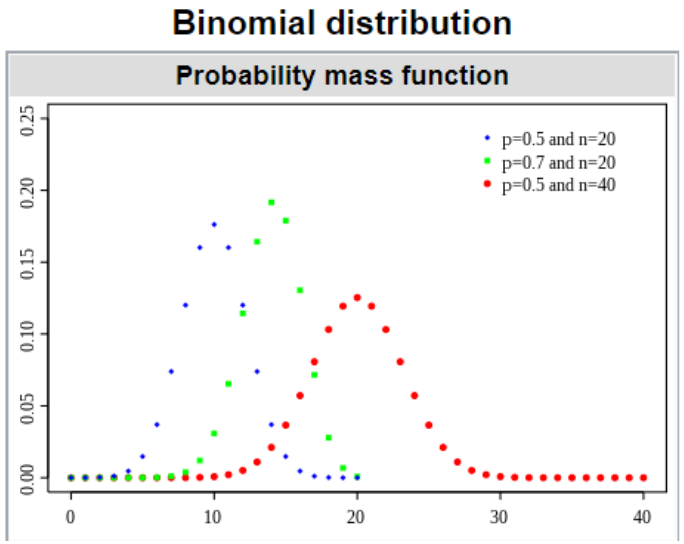
- **A binomial distribution** is the discrete probability distribution of number of successes in a sequence of *n* independent trails; with two potential outcome: success (*p*) or failure (*1-p*)
  - Examples: flipping a coin, whether a customer will purchase a product, whether a student will pass a course
- Assumptions
  - Two possible outcomes per trails: success or failure
  - The probability of success is same for each experiment
  - Experiments are independent
  - Number of experiments, *n*, fixed

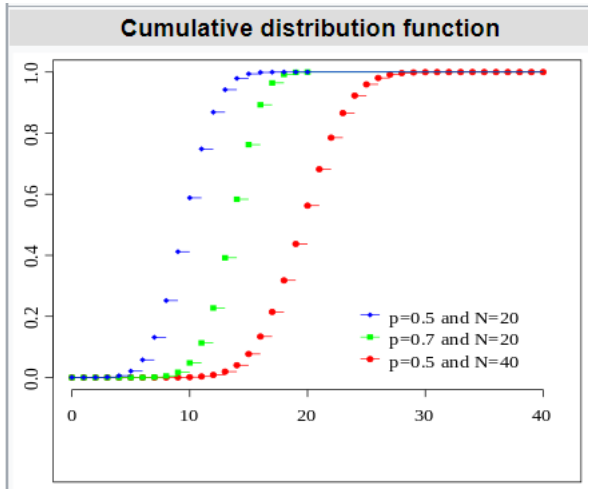
pmf of the binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where X is a random variable of number of successes in n trails

Probability Distributions – Binomial





pmf of the binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\Pr(X \le k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

The binomial distribution is approximately normal for larger n and for p not too close to 0 or 1; thumb rule if np> 15 and n(1-p)>15

Probability Distributions – Binomial

Use case: Binomial distribution is useful in modeling discrete events with a fixed number of trials, such as the number of defective items in a batch or the number of customers who buy a product.

Modeling election results: In a political election, we can model the number of voters who choose a particular candidate as a binomial distribution, where the number of trials is the total number of voters, and the probability of success is the candidate's support rate.
Quality control: In manufacturing, we can use a binomial distribution to model the number of defective products in a sample of a certain size, which can be used to estimate the quality of a production line.

Probability Distributions – Binomial

pmf of the binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

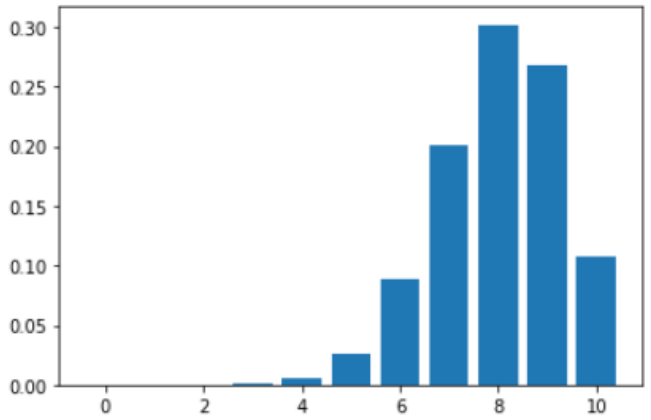
Where X is a random variable of number of successes in n trails

Suppose that 80% of adults with allergies report symptomatic relief with a specific medication. If the medication is given to 10 new patients with allergies, find the probabilities –

a. The medication is effective in all 10 patients
b. The medication is not effective to any patients
c. The medication is effective in exactly three?
d. The medication is effective in at least 3 patients?

Probability Distributions – Binomial

```
1 # of n and p
2 n = 10
3 p = 0.8
4 n_exp = list(np.arange(11))
5 # list of pmf values
6 dist = [binom.pmf(exp , n, p) for exp in n_exp ]
7 plt.bar(n_exp, dist)
8 plt.show()
```



Probability mass function

$$P(x) = \frac{n!}{x! * (n - x)!} * p^x * (1 - p)^{n - x}$$

```
1 pd.DataFrame((n_exp, dist), index = ['x', "p(x)"])
```

	0	1	2	3	4	5	6	7	8	9	10
x	0.000000e+00	1.000000	2.000000	3.000000	4.000000	5.000000	6.00000	7.000000	8.00000	9.000000	10.000000
p(x)	1.024000e-07	0.000004	0.000074	0.000786	0.005505	0.026424	0.08808	0.201327	0.30199	0.268435	0.107374

Mean = n*p
Variance = np(1-p)

```
1 #mean and variance
2 mean, var = binom.stats(n, p)
3 print("mean: {} and variance: {}".format(mean, var))
```

```
mean: 8.0 and variance: 1.5999999999999996
```

```
1 binom.cdf(3, n, p)
```

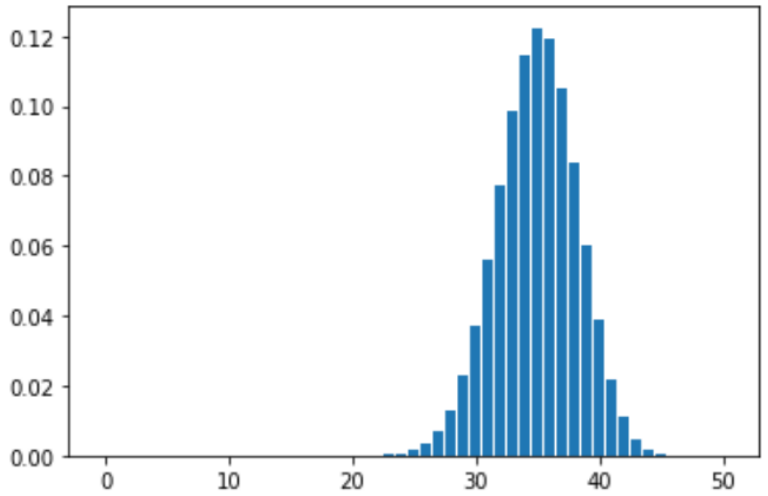
```
0.0008643583999999986
```

Probability Distributions – Binomial

Approximately 70% of DATA 220 students do their homework in time for it to be collected and graded. Each student does homework independently. In a DATA 220 class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.

Probability Distributions – Binomial

Approximately 70% of DATA 220 students do their homework in time for it to be collected and graded. Each student does homework independently. In a DATA 220 class of 50 students, what is the probability that at least 40 will do their homework on time? Students are selected randomly.




```
[  1  print("probability of at least 40 students will do their HW on time: {}".format(1-binom.cdf(39, n, p)))
```

```
probability of at least 40 students will do their HW on time: 0.07885062482305638
```

Probability Distributions – Poisson

- **A Poisson distribution** is the discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval time or space.
  - e.g., number of typos in a printed page, number of car accidents in a day

pmf of the passion distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where X is a random variable, k is the number of occurrences and $\lambda > 0$

$$E(X) = Var(X) = \lambda$$

Suppose a call center receives an average of 4 customer service calls per minute. What is the probability that the call center receives exactly 6 calls in a given minute?

Let X be the random variable representing the number of calls received in a minute. Since the number of calls can be any non-negative integer, X follows a Poisson distribution with parameter λ = 4. The probability mass function for X is:

$$P(X = 6) = \frac{4^6 e^{-4}}{6!} \approx 0.104$$

https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Poisson

- **A Poisson distribution** is the discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval time or space.
  - e.g., number of typos in a printed page, number of car accidents in a day

pmf of the passion distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where X is a random variable, k is the number of occurrences and $\lambda > 0$

$$E(X) = Var(X) = \lambda$$

Example: On a particular river, overflow floods occur once every 100 years on average. Calculate the probability of k = 0, 1, 2, 3, 4, 5, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate.

Given, average event rate one overflow flood per 100 years, $\lambda = 1$

$$P(X = k = 0) = \frac{1^0 e^{-1}}{0!} \approx 0.368$$

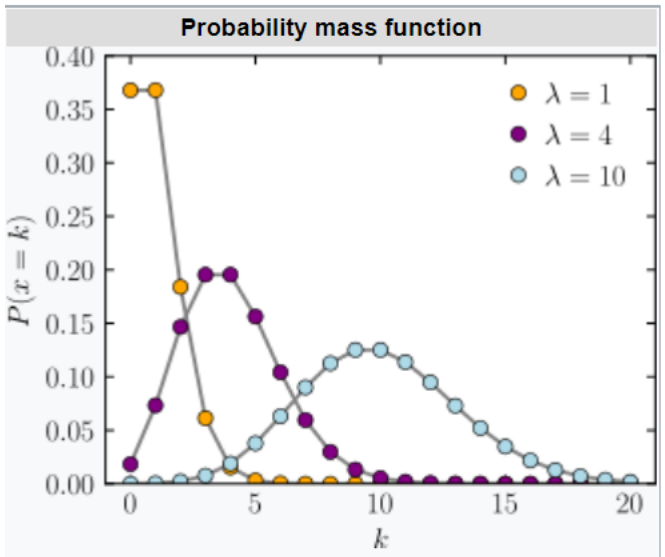
```
1 print("probability of 0 overflow floods in a 100 year interval:", poisson.pmf(0,1))
```

```
probability of 0 overflow floods in a 100 year interval: 0.36787944117144233
```

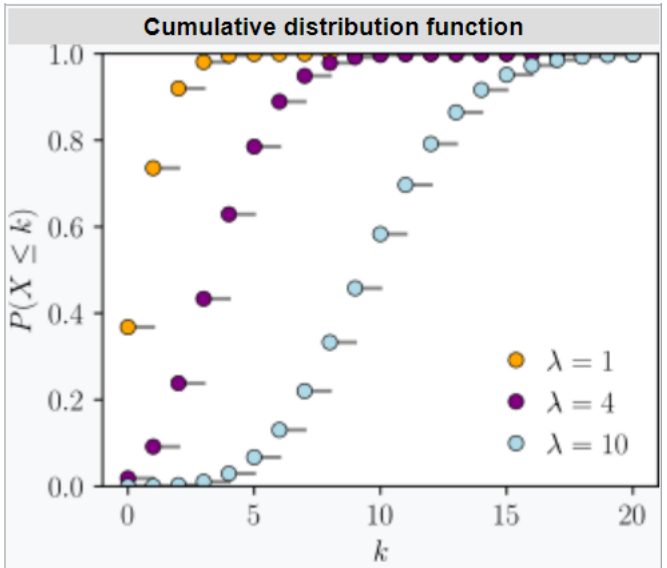
https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Poisson

Poisson Distribution



The horizontal axis is the index k, the number of occurrences. λ is the expected rate of occurrences. The vertical axis is the probability of k occurrences given λ. The function is defined only at integer values of k; the connecting lines are only guides for the eye.



The horizontal axis is the index k, the number of occurrences. The CDF is discontinuous at the integers of k and flat everywhere else because a variable that is Poisson distributed takes on only integer values.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$P(X \leq k) = e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!}$$

The Poisson distribution is approximately normal for large λ

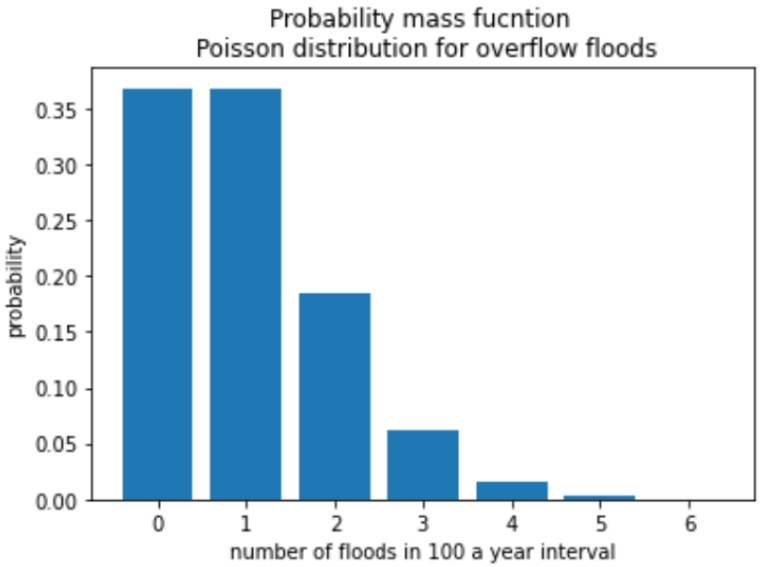
Probability Distributions – Poisson

Example: On a particular river, overflow floods occur once every 100 years on average. Calculate the probability of k = 0, 1, 2, 3, 4, 5, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate.

Probability Distributions – Poisson

Example: On a particular river, overflow floods occur once every 100 years on average. Calculate the probability of k = 0, 1, 2, 3, 4, 5, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate.

	0	1	2	3	4	5	6
k	0.000000	1.000000	2.00000	3.000000	4.000000	5.000000	6.000000
p(k)	0.367879	0.367879	0.18394	0.061313	0.015328	0.003066	0.000511

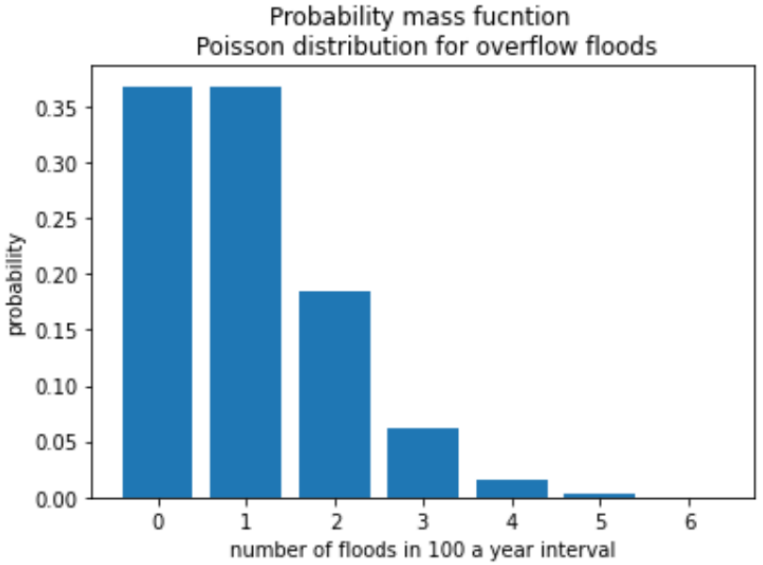


Probability Distributions – Poisson

Example: On a particular river, overflow floods occur once every 100 years on average. Calculate the probability of k = 0, 1, 2, 3, 4, 5, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate.

What is the probability of at least 2 overflow floods in a 100 year interval?

	0	1	2	3	4	5	6
k	0.000000	1.000000	2.00000	3.000000	4.000000	5.000000	6.000000
p(k)	0.367879	0.367879	0.18394	0.061313	0.015328	0.003066	0.000511



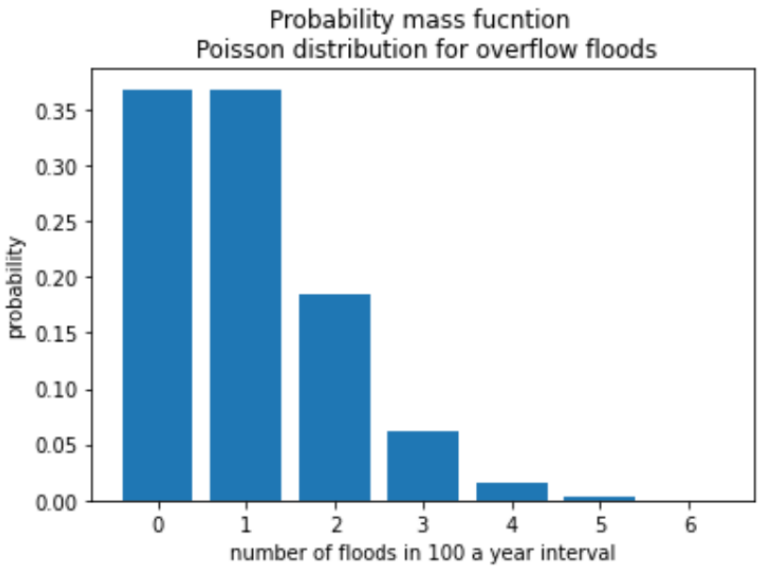
Probability Distributions – Poisson

Example: On a particular river, overflow floods occur once every 100 years on average. Calculate the probability of k = 0, 1, 2, 3, 4, 5, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate.

	0	1	2	3	4	5	6
k	0.000000	1.000000	2.00000	3.000000	4.000000	5.000000	6.000000
p(k)	0.367879	0.367879	0.18394	0.061313	0.015328	0.003066	0.000511

```
1 print("probability of at least 2 overflow floods in a 100 year interval: {}".format(1-poisson.cdf(1,1)))
```

```
probability of at least 2 overflow floods in a 100 year interval: 0.26424111765711533
```



Probability Distributions – Geometric

- The **geometric distribution** is the discrete probability distribution that expresses the number of trails needed to get the first success in repeated Bernoulli trails
- Assumptions:
  - Trials are independent
  - Each trails have two possible outcomes : success (*p*) and failure (*1-p*)
  - The probability for success *p* is fixed for every trails

pmf of the geometric distribution

$$P(X) = (1-p)^{x-1}\, p$$

For instance, if the first success occur at the xth trail then all the previous trails (first (x-1) trails) must fail

$$\boldsymbol{\mu} = \frac{\mathbf{1}}{\boldsymbol{p}} \;\&\; \boldsymbol{\sigma^2} = \frac{\mathbf{1-p}}{\boldsymbol{p}}$$

https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Geometric

- The **geometric distribution** is the discrete probability distribution that expresses the number of trials needed to get the first success in repeated Bernoulli trails
- Assumptions:
  - Trails are independent
  - Each trails have two possible outcomes : success (*p*) and failure (*1-p*)

Modeling first success in a series of trials: In a series of trials where each has a probability of success p, the number of trials it takes to achieve the first success can be modeled using a geometric distribution, which can be used to estimate the expected time until a success occurs.
Modeling survival analysis: In medical research, the time it takes for a patient to experience a certain event (such as disease recurrence or death) can be modeled using a geometric distribution, which can be used to estimate survival probabilities and analyze treatment efficacy.

https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Geometric

- The **geometric distribution** is the discrete probability distribution that expresses the number of trails needed to get the first success in repeated Bernoulli trails
- Assumptions:
  - Trails are independent
  - Each trails have two possible outcomes : success (*p*) and failure (*1-p*)

Example: A doctor is seeking an antidepressant for a newly diagnosed patient. Suppose that, of the available anti-depressant drugs, the probability that any particular drug will be effective for a particular patient is $p = 0.6$. What is the probability that the first drug found to be effective for this patient is the first drug tried, the second drug tried, and so on? What is the expected number of drugs that will be tried to find one that is effective?

https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Geometric

$$P(X = 1) = \ (1 - .6)^{1-1} \ * \ .6 = \ .6$$

$$P(X = 2) = ?$$

```
1 stats.geom.pmf(1, 0.6)
```

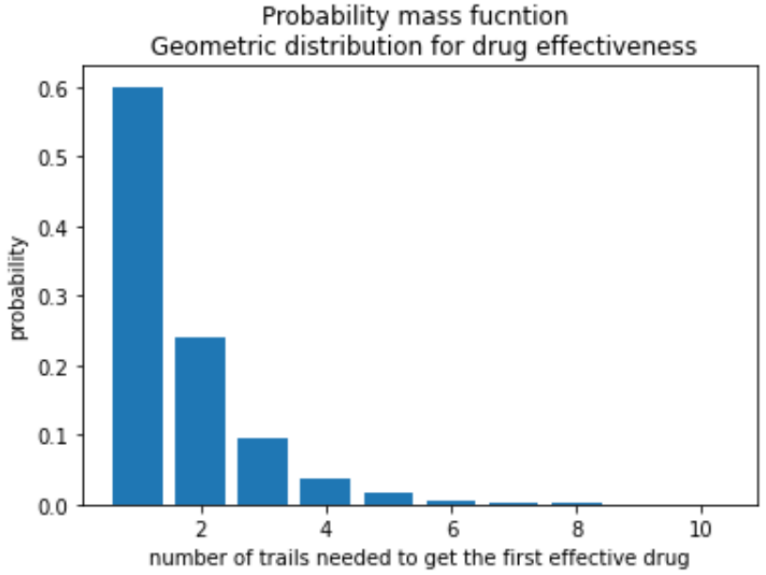
```
0.6
```

https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Geometric

Example: A doctor is seeking an antidepressant for a newly diagnosed patient. Suppose that, of the available anti-depressant drugs, the probability that any particular drug will be effective for a particular patient is $p = 0.6$. What is the probability that the first drug found to be effective for this patient is the first drug tried, the second drug tried, and so on? What is the expected number of drugs that will be tried to find one that is effective?

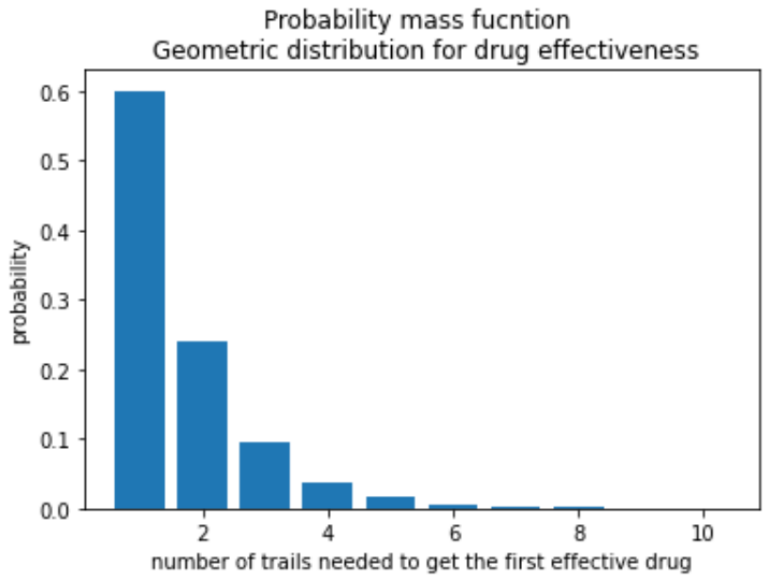
- Geometric distribution follows right skewness shape
- Starts @ 1 and no upper bounds



https://en.wikipedia.org/wiki/Poisson_distribution

Probability Distributions – Geometric

- What is the probability that the first drug found to be effective for the patient on or before 4th tried



Example: A doctor is seeking an antidepressant for a newly diagnosed patient. Suppose that, of the available anti-depressant drugs, the probability that any particular drug will be effective for a particular patient is $p = 0.6$. What is the probability that the first drug found to be effective for this patient is the first drug tried, the second drug tried, and so on? What is the expected number of drugs that will be tried to find one that is effective?

Probability Distributions – Geometric

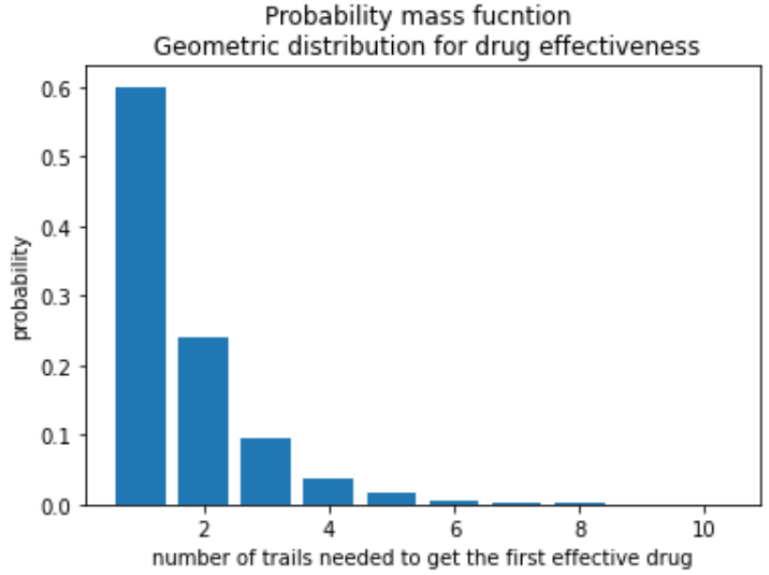
- What is the probability that the first drug found to be effective for the patient on or before 4th tried

$$P(X \leq 4) = P(X = 1) + .. + P(X = 4)$$

```
1 stats.geom.cdf(4,.6)
```

```
0.9744
```

Example: A doctor is seeking an antidepressant for a newly diagnosed patient. Suppose that, of the available anti-depressant drugs, the probability that any particular drug will be effective for a particular patient is $p = 0.6$. What is the probability that the first drug found to be effective for this patient is the first drug tried, the second drug tried, and so on? What is the expected number of drugs that will be tried to find one that is effective?



Probability Distributions – Geometric

Example: A person conducting telephone surveys and there is a 9% chance of reaching an adult who will complete the survey.

What is the probability that the first survey completed occurs on the 3rd call?

$$P(X = 3) = 0.074529$$

```
[0]  1 stats.geom.pmf(3,.09)

0.074529
```

Probability Distributions – Negative Binomial

Example: A person conducting telephone surveys and there is a 9% chance of reaching an adult who will complete the survey.

What is the probability the 3rd completed survey occurs on the 10th call?

Probability Distributions – Negative Binomial

Example: A person conducting telephone surveys and there is a 9% chance of reaching an adult who will complete the survey.

What is the probability the 3rd completed survey occurs on the 10th call?

- The **negative binomial distribution** is the discrete probability distribution that expresses the number of trials needed to get the rth success
- Assumptions:
  - Trails are independent
  - Each trails have two possible outcomes : success (p) and failure ($1-p$)
  - Number of trails is fixed

pmf of the negative binomial distribution

$$P(X = x) = p \times \binom{x-1}{r-1} p^{r-1} (1-p)^{(x-1)-(r-1)}$$

Where X is a random variable and represents the number of trails needed to get the first rth success.

Probability Distributions – Negative Binomial

Recall

pmf of the binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where X is a random variable of number of successes in n trails

First (x-1) trials result in (r-1) successes

pmf of the negative binomial distribution

$$P(X = x) = p \times \binom{x-1}{r-1} p^{r-1} (1-p)^{(x-1)-(r-1)}$$

Where X is a random variable and represents the number of trails needed to get the first rth success.

Example: A person conducting telephone surveys and there is a 9% chance of reaching an adult who will complete the survey.

What is the probability the 3rd completed survey occurs on the 10th call?

Probability Distributions – Negative Binomial

Example: A person conducting telephone surveys and there is a 9% chance of reaching an adult who will complete the survey.

What is the probability the 3rd completed survey occurs on the 10th call?

```
1 stats.nbinom.pmf(10,3,0.09)
```

```
0.0187363671071346
```

pmf of the negative binomial distribution

$$P(X = x) = \binom{x-1}{r-1} p^r \ (1-p)^{x-r}$$

Where X is a random variable and represents the number of trails needed to get the first rth success.

Continuous Probability Distributions – Normal

- The normal distribution (aka Gaussian) is a continuous probability distribution for real valued random variable X with pdf:

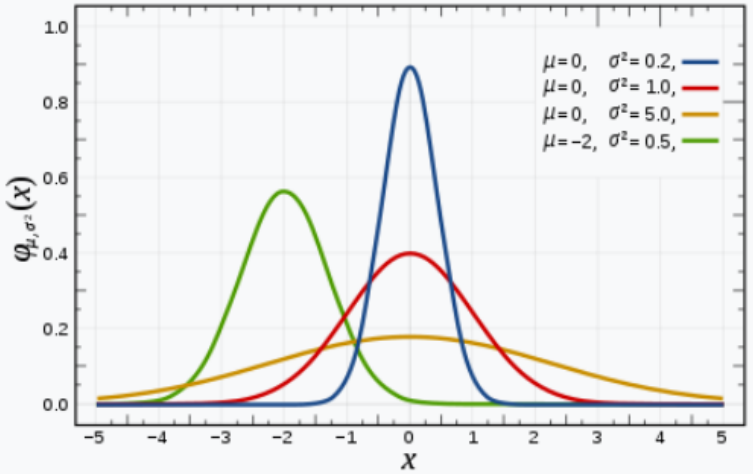
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Continuous Normal Distributions – Normal

- Normal distribution
  - Symmetric, bell shaped, and unimodal
  - Mean, median and mode are equal
  - Follows the empirical rule
  - Mean (or mode) define the location of the peak and population std defines the spread
    - when the curve becomes wide and flat
    - When the curve becomes tall and narrow

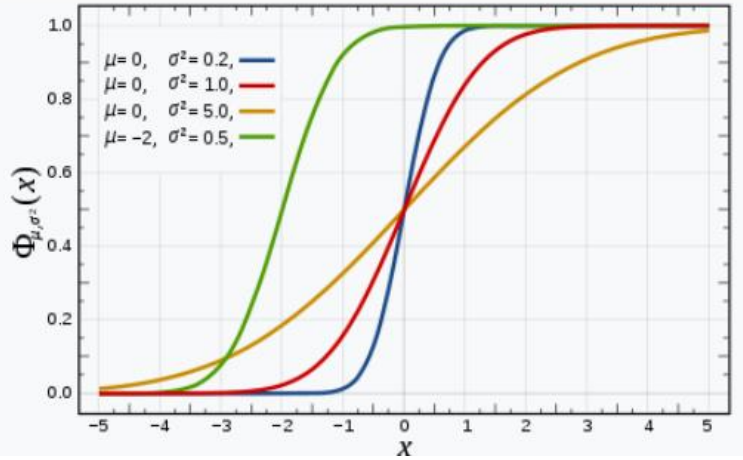
Normal distribution

Probability density function



The red curve is the *standard normal distribution*

Cumulative distribution function



Continuous Probability Distributions – Normal

- Standard Normal distribution
  - $N(0,1)$
  - Mark the cut-off values for the standard normal curve
- Find area under the curve of normal and standard normal distribution
  - Find $P(a < x < b)$

For standard normal distribution, find area under the curve: P(-1.2 < x < .75)

```
1 stats.norm(0,1).cdf(-1.2)
```

```
0.11506967022170822
```

```
1 stats.norm(0,1).cdf(.75)
```

```
0.7733726476231317
```

```
1 stats.norm(0,1).cdf(.75) - stats.norm(0,1).cdf(-1.2)
```

Learn: how to use normal distribution table to find the probabilities

Probability Theory

- Sampling distribution
- Law of Large Number
- The Central Limit Theorem

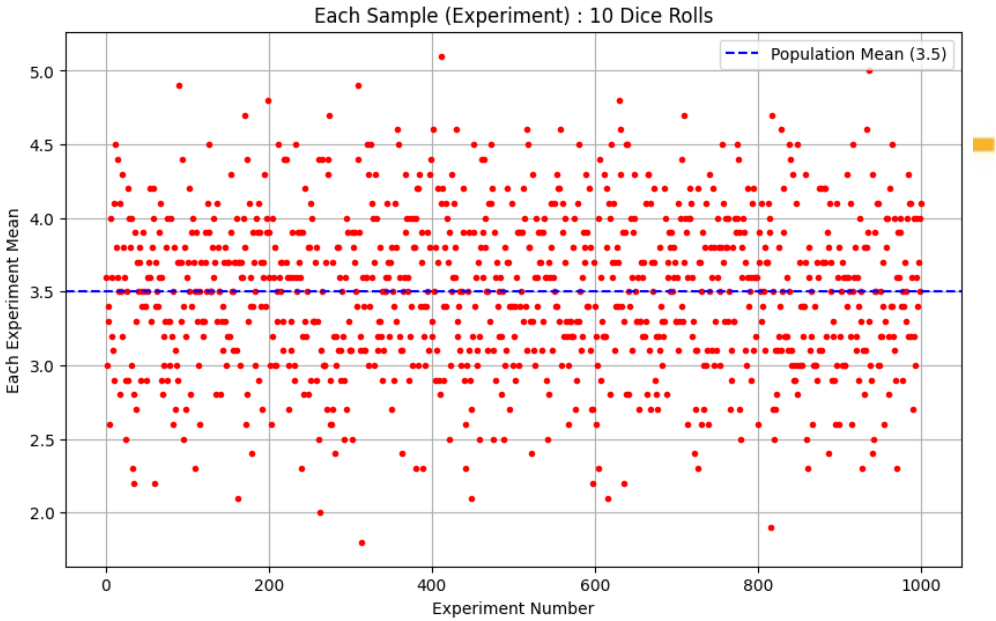
- A **sampling distribution** is the probability distribution of a statistic based on a random sample drawn from a population -- represents the distribution of a statistic (such as the mean or standard deviation) that would be obtained from all possible samples of a certain size drawn from the same population
  - Allows to make inference about the population – estimating population parameters

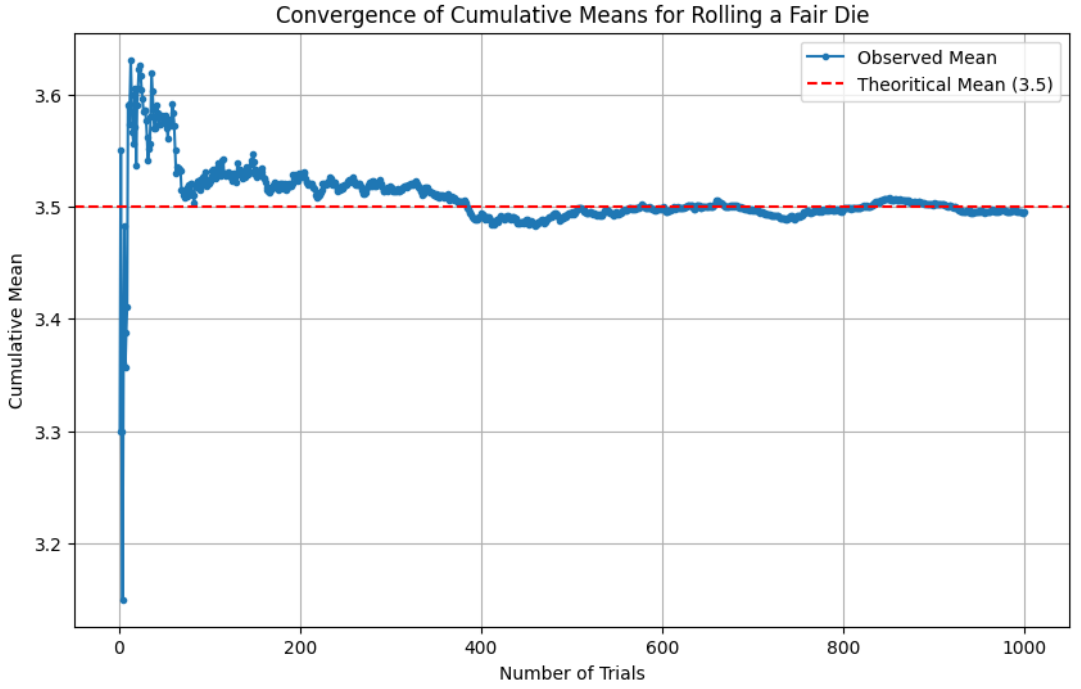
Mean and Standard Deviation of a Sample Mean: Let $\bar{x}$ represent the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then the sampling distribution of $\bar{x}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Example: The population mean for IQ tests for adults is 100 with a standard deviation of 16. Suppose you take samples of size 36 and find the sample mean for each sample. What is the mean and standard deviation of the sampling distribution?

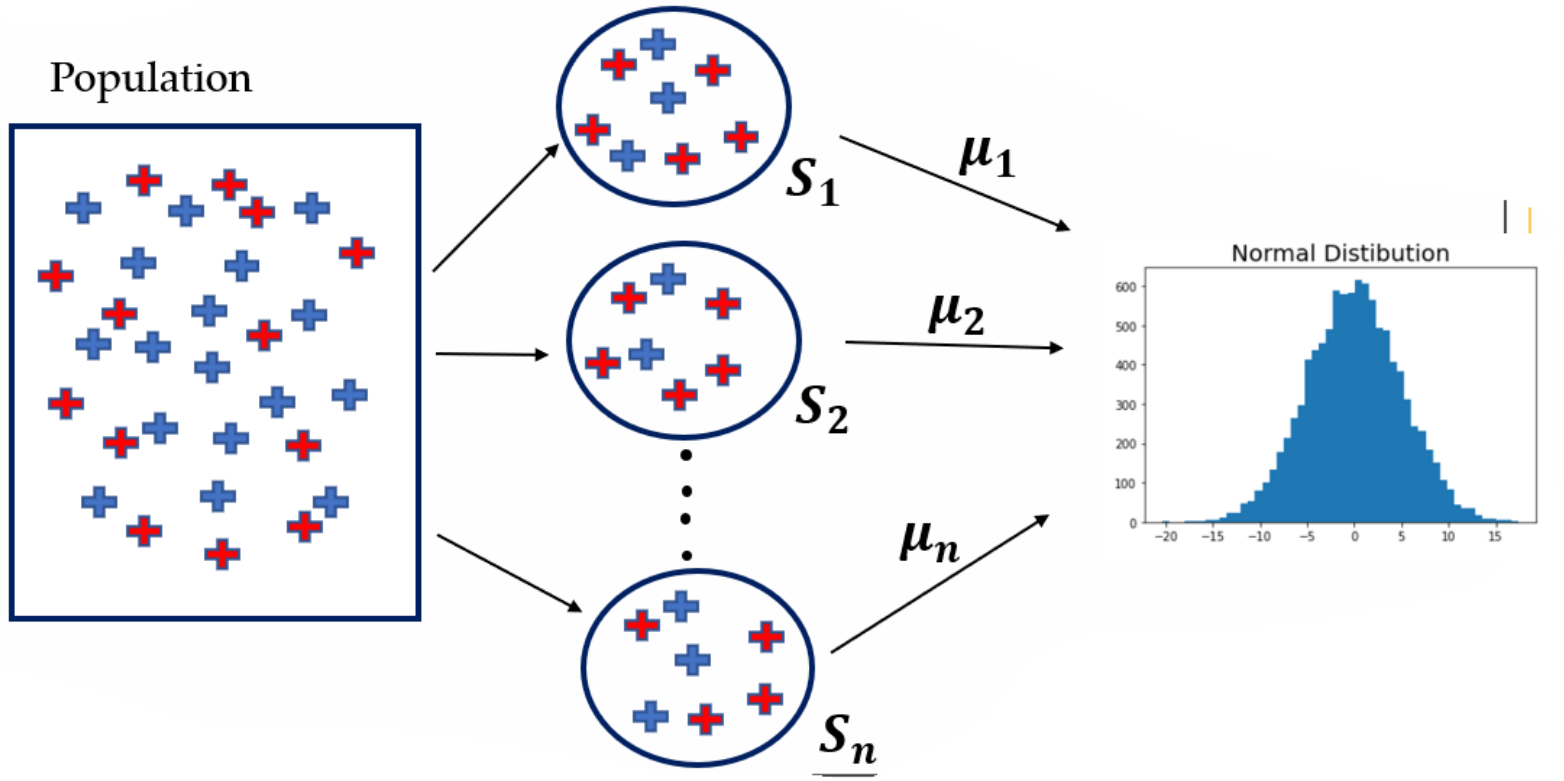
- **Law of Large Number** – as the number of experiment (trails) repetitions increases, the average of the sample mean better approximate the true population mean
  - Why we need it? – one sample or experiment sensitive to sampling variability or noise

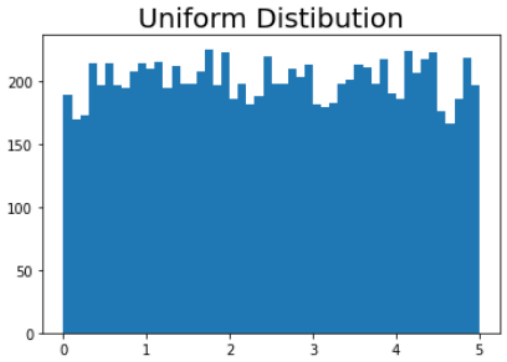
$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{X_i}{n} = \bar{X}$$

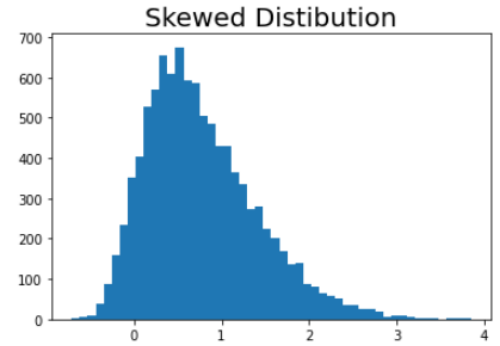


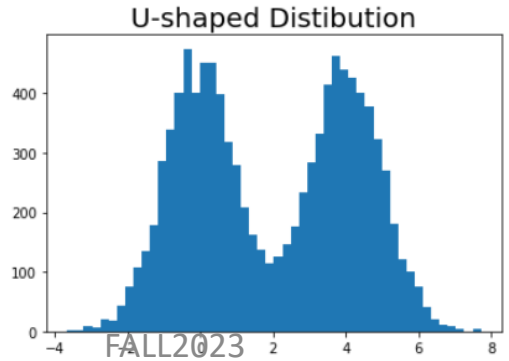


The Central Limit Theorem: Let $\bar{x}$ be the mean of a large (n > 30) simple random sample from a population with mean μ and standard deviation σ. Then $\bar{x}$ has an **approximately normal distribution,** with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

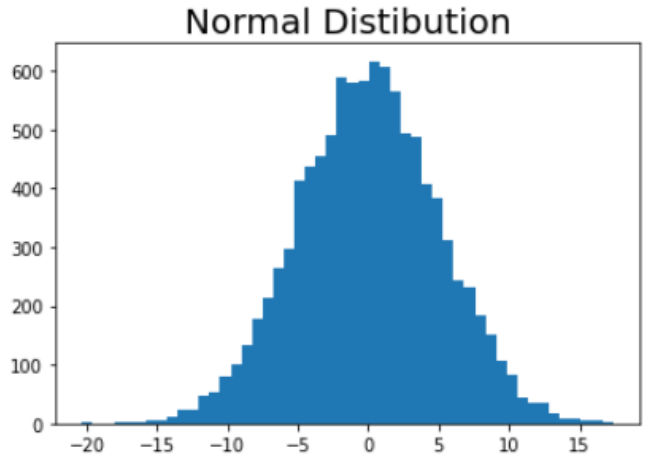




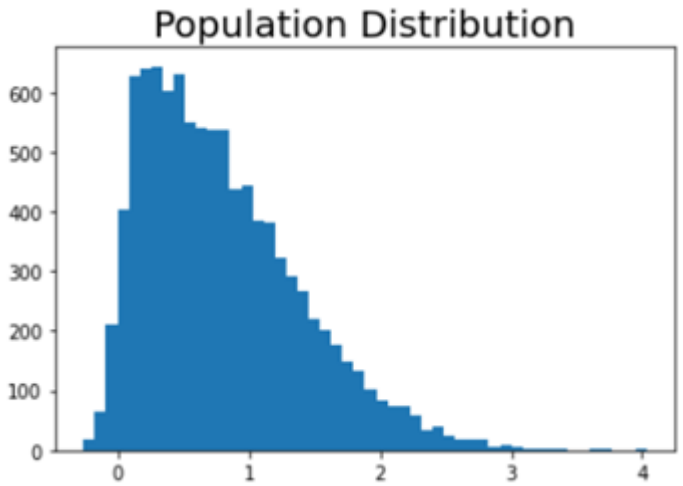


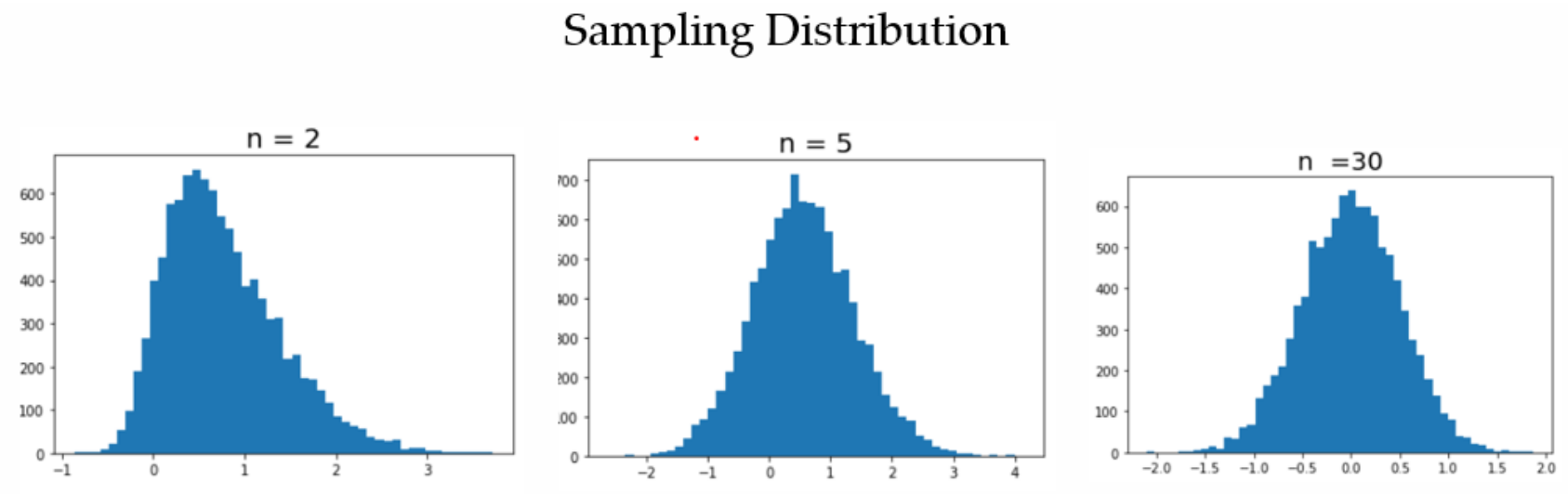


The CLT tells that — whatever the distribution of the population (uniform, skewed, U-shaped) if we take a minimum sample size of 30 then the mean of the sample distribution will be approximately normally distributed.

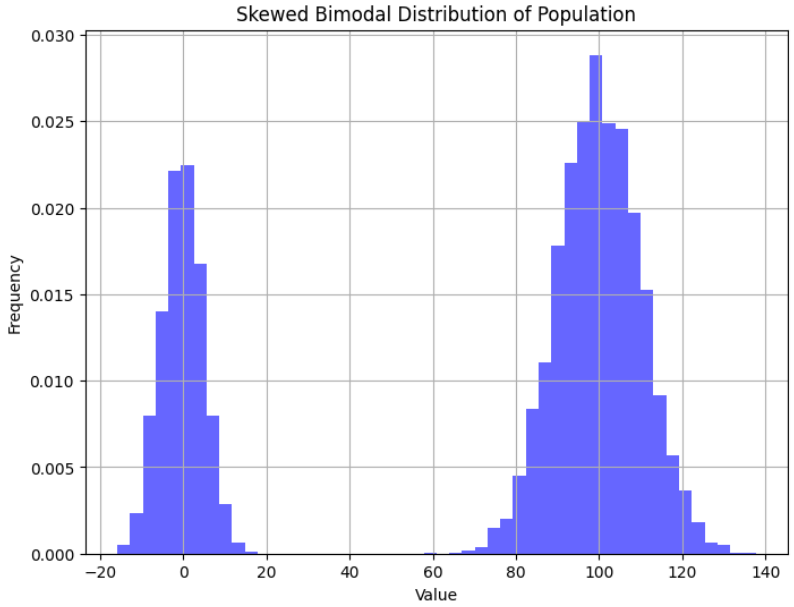


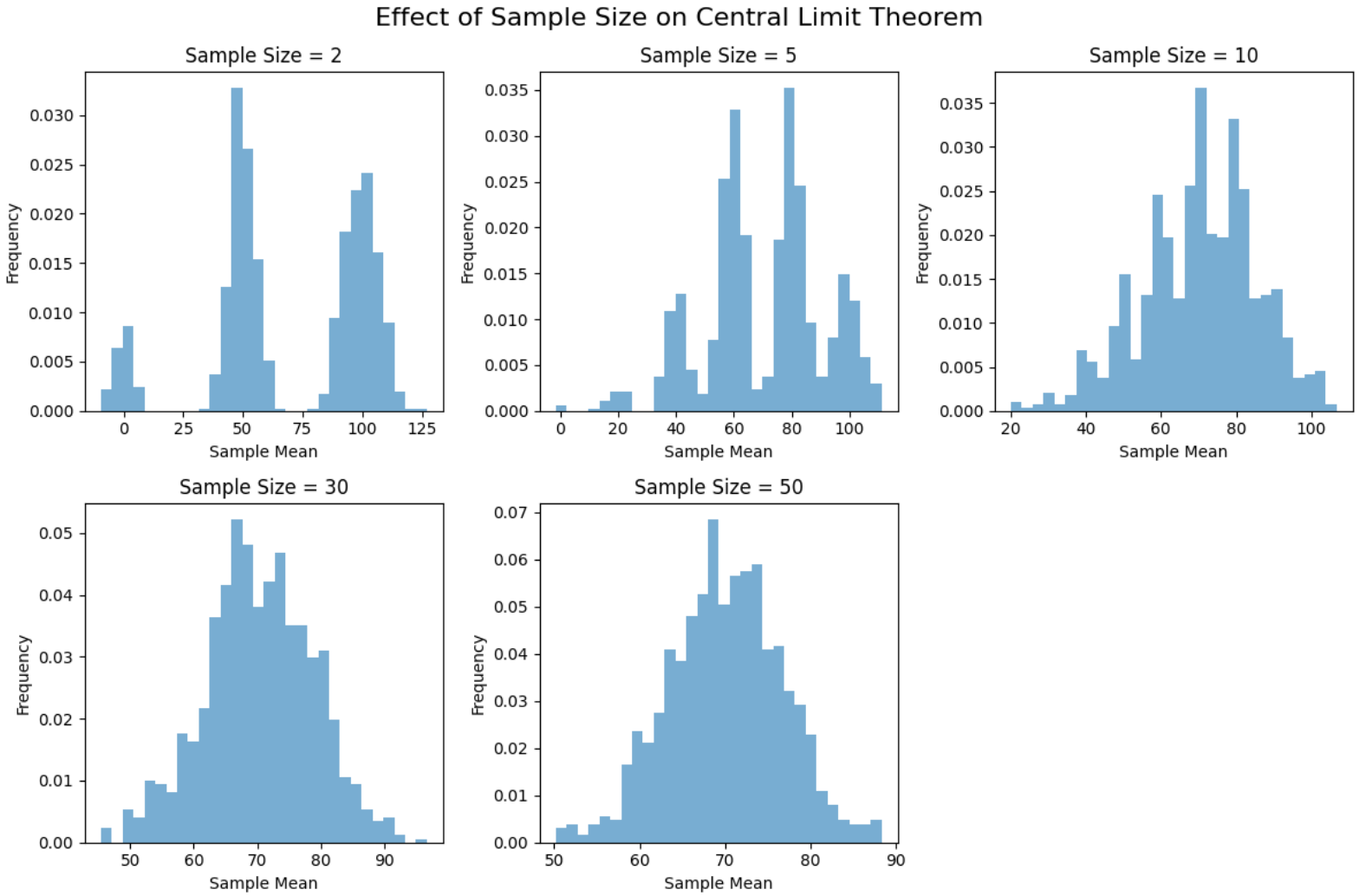
Effect of Sample Size in Central Limit Theorem





Effect of Sample Size in Central Limit Theorem





Sampling Distributions

In California in November 2010, the gubernatorial race pitted the Republican candidate Meg Whitman against the Democratic candidate, Jerry Brown. The exit poll on which TV networks relied for their projections found that, after sampling 3889 voters, 53.1% said they voted for Brown, 42.4% for Whitman, and 4.5% for other/no answer (www.cnn.com/ELECTION/2010). At the time of the exit poll, the percentage of the entire voting population (nearly 9.5 million people) that voted for Brown was unknown. In determining whether they could predict a winner, the TV net-

Sampling Distributions

In California in November 2010, the gubernatorial race pitted the Republican candidate Meg Whitman against the Democratic candidate, Jerry Brown. The exit poll on which TV networks relied for their projections found that, after sampling 3889 voters, 53.1% said they voted for Brown, 42.4% for Whitman, and 4.5% for other/no answer (www.cnn.com/ELECTION/2010). At the time of the exit poll, the percentage of the entire voting population (nearly 9.5 million people) that voted for Brown was unknown. In determining whether they could predict a winner, the TV net-

Q1: How close can we expect a sample percentage to be to the population percentage? For instance, if 53.1% of 3889 sampled voters supported Brown, how close to 53.1% is the percentage of the entire population of 9.5 million voters who voted for him

Q2: How does the sample size influence our analysis? For instance, could we sample 100 instead of 3889 voters? On the other hand, are 3889 voters enough or we need tens of thousands of voters in the sample

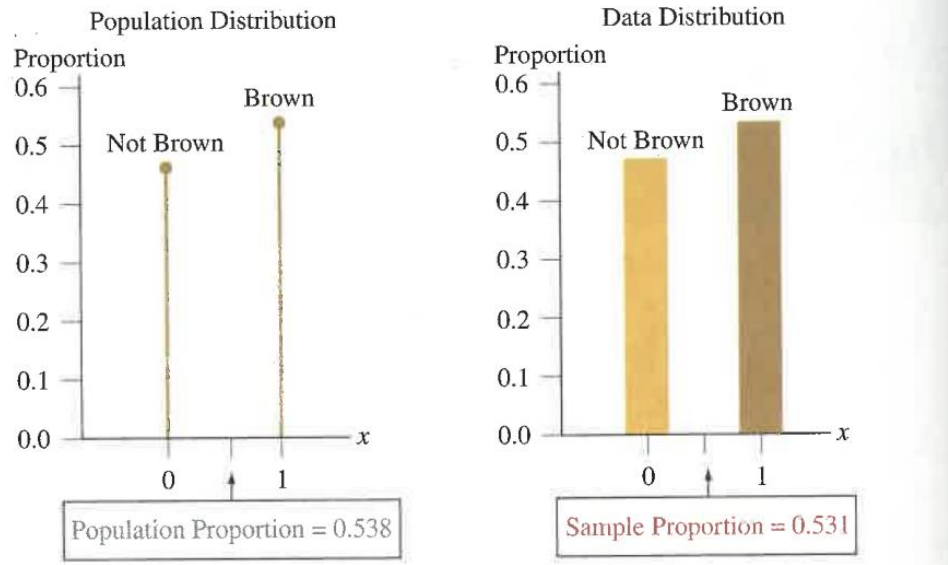
Sampling Distributions

- A sampling distribution shows how sample statistics vary

- A voters' vote is categorical random variable with binary outcome
  - X = vote outcome = 1 or 0
    - 1 indicated voted for Brown
    - 0 otherwise

Sampling Distributions

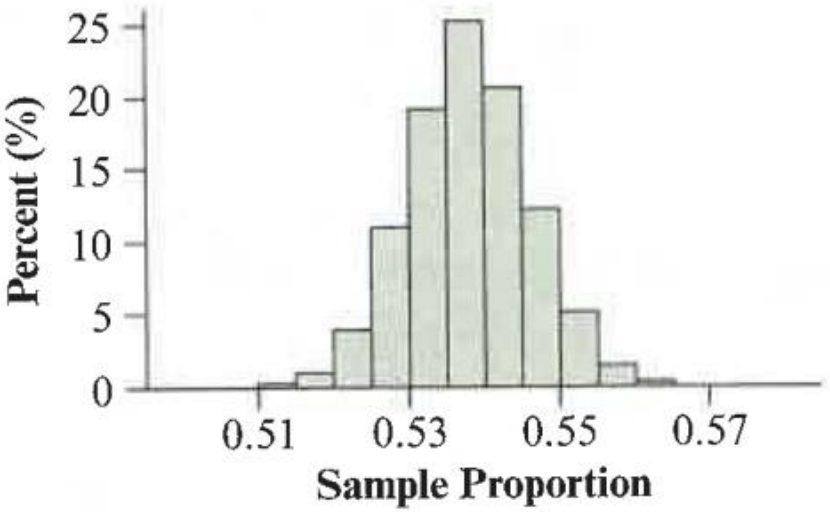
- The population proportion (53.8%) and the sample proportion (53.1%) are very close to each other
- What about other exit polls?
  - Other polls may provide very close proportions like 54.1%, 52.9%.., and also can provide 40%, 60% which are very far away from the population proportion
- To answer, we need to learn **sampling distribution.**



▲ **Figure 7.1** The Population (9.5 Million Voters) and Data (n = 3889) Distributions of Candidate Preference (0 = Not Brown, 1 = Brown). **Question** Why do these look so similar?

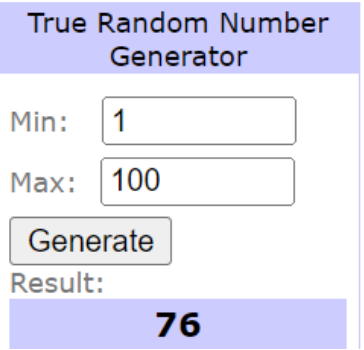
Sampling Distributions

- The **Sampling Distribution** show all possible values of the sample proportion and how often the sample proportion are expected to occur in random sampling



Sampling Distributions

- Simulating the sampling distribution for a sample proportion
  - Simulating a number between 1 and 100 using software or website like random.org
  - We carry out the simulation based on the actual population proportion (53.8%, for simplicity consider 54%)



1 2 3 …..53 54 — Vote for Brown

55 56 …. 100 — Vote for other candidates

Sampling Distributions

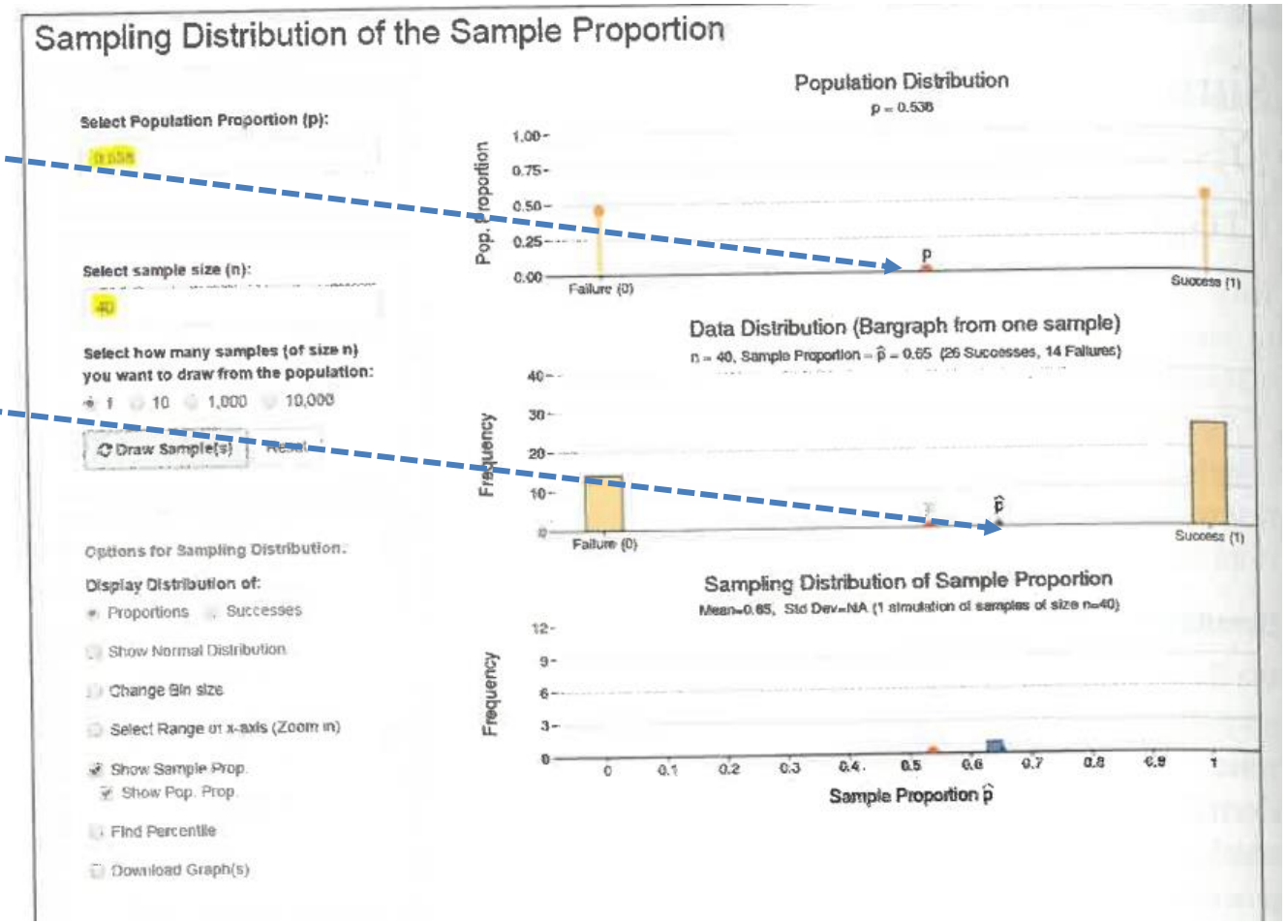
- Consider small sample size:
  - n = 6: 30, 50, 67, 3, 33, 90 ➔ 1, 1, 0, 1, 1, 0 ➔ 4/6 = 66.66%
  - n = 6: 80, 5, 67, 83, 57, 8 ➔ 0, 1, 0, 0, 0, 1 ➔ 2/6 = 33.33%
- With smaller sample size there will be significant variability in the sampling proportion

Sampling Distributions

- We can simulate the polls with larger sample size and check the variability
  - With n = 40
  - With actual sample size in the exit poll, n = 3889

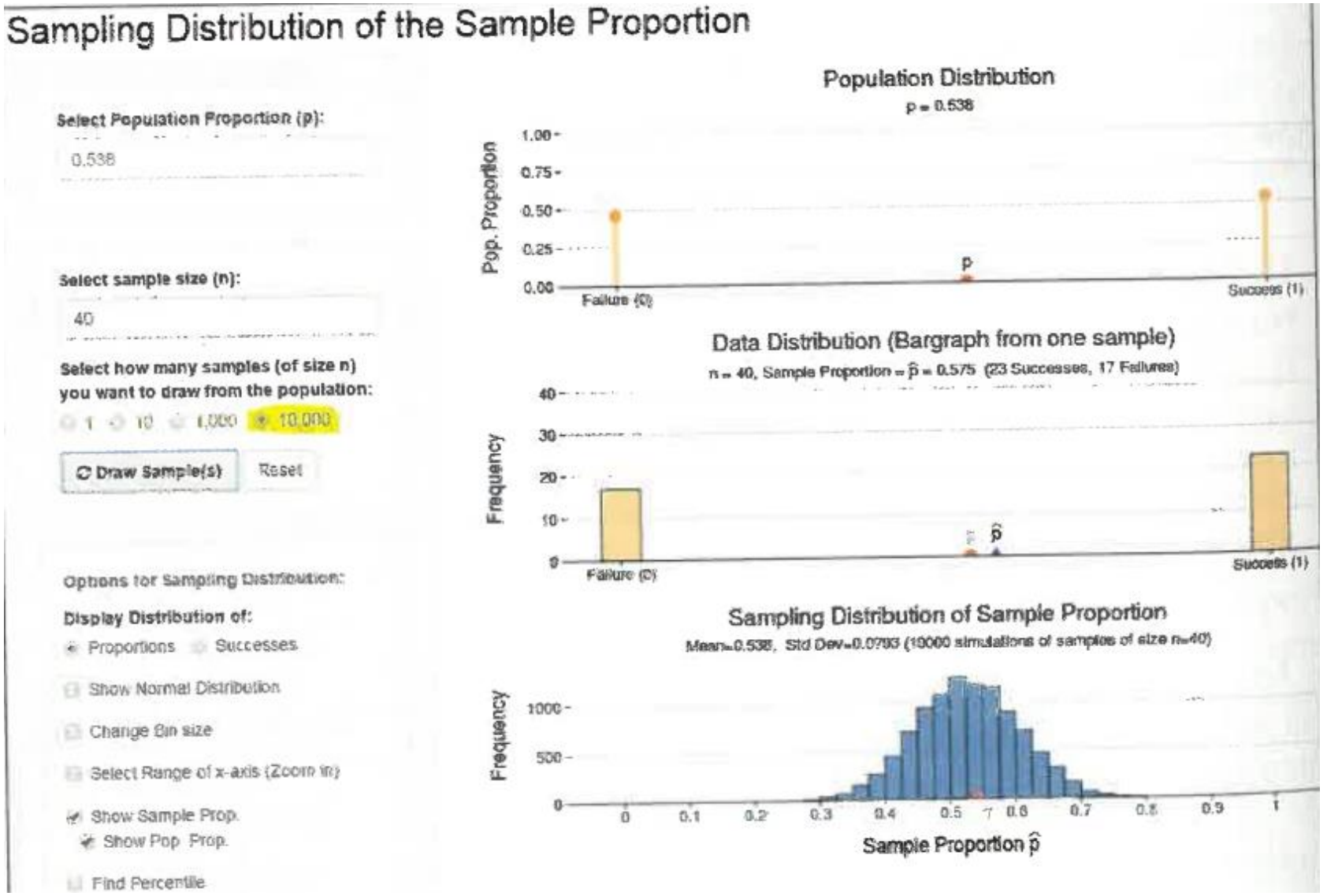
Sampling Distributions

- Population proportion (p = 0.538)
- Data distribution for the sampled drawn of size, n = 40 ➔ sample proportion 0.65
- Location of the sample proportion

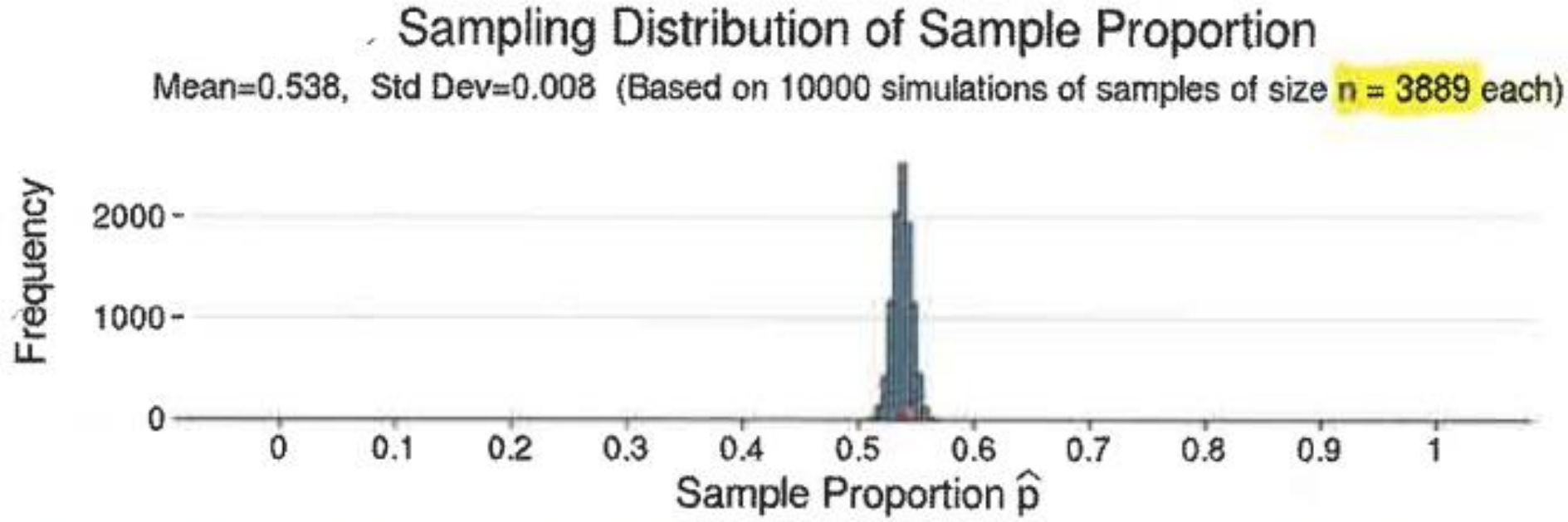


Sampling Distributions

- Result of taking 10,000 samples of size, n = 40
- Population proportion 0.538
- 10000 sample proportions with mean 0.538 and Sd 0.079
  - All proportions fall between 0.3 and 0.75



Sampling Distributions



▲ **Figure 7.4** Simulated Sampling Distribution When Taking 10,000 Samples of Size $n = 3889$. **Question** With a mean of 0.538 and standard deviation of 0.008, find an interval within which almost all sample proportions will fall when sampling $n = 3889$ voters.

Sampling Distributions

Question to Explore

Election results showed that 53.8% of the population of all voters voted for Brown. What was the mean and standard deviation of the sampling distribution of the sample proportion who voted for him? Interpret these two measures.

Sampling Distributions

- Mean and standard deviation of the sampling distributions of the sample proportion

For a random sample of size *n* from a population with proportion *p* in a particular category, then the sampling distribution of the sample proportion has

$$mean = p$$

$$standard\ deviation = \sqrt{\left(\frac{p\ (1-p)}{n}\right)}$$

Sampling Distributions

Question to Explore

Election results showed that 53.8% of the population of all voters voted for Brown. What was the mean and standard deviation of the sampling distribution of the sample proportion who voted for him? Interpret these two measures.

- Population proportion, $p = 0.538$

➔ sampling distribution of the sample proportion = $p = 0.538$

Sampling Distributions

Question to Explore

Election results showed that 53.8% of the population of all voters voted for Brown. What was the mean and standard deviation of the sampling distribution of the sample proportion who voted for him? Interpret these two measures.

- Population proportion, $p = 0.538$

➔ sampling distribution of the sample proportion = $p = 0.538$

- $standard\ deviation = \sqrt{(\frac{p\ (1-p)}{n})}$

$= \sqrt{(\frac{0.538\ \ (0.462)}{3889} = 0.008}$

Sampling Distributions

Picture the Scenario

Let's now conduct an analysis that uses the actual exit poll of 3889 voters for the 2010 California gubernatorial election. In that exit poll, 53.1% of the 3889 voters sampled said they voted for Jerry Brown.

Questions to Explore

a. Is it reasonable to assume a normal shape for the sampling distribution of the sample proportion resulting from exit polls such as this one?

b. Given that the actual population proportion supporting Brown was 0.538, what are the values of the sample proportion we would expect to observe from exit polls such as this one?

c. Based on the results of this exit poll, would you have been willing to predict Brown as the winner on election night while the votes were still being counted?

Picture the Scenario

Let's now conduct an analysis that uses the actual exit poll of 3889 voters for the 2010 California gubernatorial election. In that exit poll, 53.1% of the 3889 voters sampled said they voted for Jerry Brown.

Questions to Explore

a. Is it reasonable to assume a normal shape for the sampling distribution of the sample proportion resulting from exit polls such as this one?

b. Given that the actual population proportion supporting Brown was 0.538, what are the values of the sample proportion we would expect to observe from exit polls such as this one?

c. Based on the results of this exit poll, would you have been willing to predict Brown as the winner on election night while the votes were still being counted?

Picture the Scenario

Let's now conduct an analysis that uses the actual exit poll of 3889 voters for the 2010 California gubernatorial election. In that exit poll, 53.1% of the 3889 voters sampled said they voted for Jerry Brown.

Questions to Explore

a. Is it reasonable to assume a normal shape for the sampling distribution of the sample proportion resulting from exit polls such as this one?

b. Given that the actual population proportion supporting Brown was 0.538, what are the values of the sample proportion we would expect to observe from exit polls such as this one?

c. Based on the results of this exit poll, would you have been willing to predict Brown as the winner on election night while the votes were still being counted?

1. Yes
2. As it is normally distributed and the SD from the previous example is 0.008, we can tell almost 100% of the data will fall between 3- standard deviation

Sampling Distributions

3. We don't know the actual population proportion; however, the sample proportion is given: 53.1%. On the election day, sample proportion is the best estimator of the population proportion.

So, we can estimate the expected standard deviation

- $standard\ deviation = \sqrt{(\frac{p\ (1-p)}{n})}$

$$= \sqrt{(\frac{0.531\ \ (0.469)}{3889} = 0.008}$$

Sampling Distributions

3. We don't know the actual population proportion; however, the sample proportion is given: 53.1%. On the election day, sample proportion is the best estimator of the population proportion.

So, we can estimate the expected standard deviation

- $standard\ deviation = \sqrt{(\frac{p\ (1-p)}{n}}$

$$= \sqrt{(\frac{0.531\ \ (0.469)}{3889} = 0.008$$