BUDT 730
Data, Models and Decisions

Lecture 04

Probability Distributions (II)

Prof. Sujin Kim

Ch5
Standard Probability Distributions

Binomial Distribution

Introduction

- Several distributions for random variables commonly occur in a variety of business applications
    - We’ve already discussed the **General Discrete**
    - **Binomial**
    - **Normal**
    - Text also discusses the **Poisson** and **Exponential** distributions, which are very important for simulation. We will skip them in this course.
- Our goal is to understand the properties of the Normal and Binomial distributions and how to work with them in Excel

Example: Supermarket Spending

- Historical data suggests that customer spending at a supermarket is randomly distributed.
- It is also known that the **probability** that one customer spends **at least $100 is 0.3**.
- If 500 customers shop in a given day, what is the mean number of customers who spend at least $100?
- Intuitively, the answer is approximately

$$500*0.3=150$$

- What is the standard deviation?

Example: Supermarket Spending

- $X_i = 1$, if the ith customer spends at least \$100

  $= 0$, otherwise.
- Note that $P(X_i = 1) = 0.3$ and $P(X_i = 0) = 0.7$
- What is the value of $E(X_i)$?

$$E(X_i) = 1 * 0.3 + 0 * 0.7 = 0.3$$

- # of customers who spend at least \$100 is

$$Y_{500} = X_1 + X_2 + \cdots + X_{500}$$

- Then,

$$E(Y_{500}) = E(X_1) + E(X_2) + \cdots E(X_{500}) = 0.3 * 500 = 150$$

Binomial Experiment

- Consider a trial with two outcomes: success or failure. The probability of a success on each trial is p

Note: This is called the Bernoulli trial.

- Define X_i to be the random outcome of the ith trial.
  - $X_i = 1\ (success)$ with probability p

    $= 0\ (failure)$, otherwise.
- Define Y_n to be the **number of successes** $= \sum_{i=1}^{n} X_i$
- Then, $E(Y_n) = n * p$
- Assume that X_i's are independent. What is the distribution of Y_n?

Binomial Distribution

Binomial Distribution (n, p)

- Y$\sim Binomial(n, p)$ = the **number of successes** in *n* independent (Bernoulli) trials.

$$\boxed{P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0{,}1{,}2\ldots, n}$$

- $E(Y) = np$
- $Var(Y) = np(1-p)$
- $Stdev(Y) = \sqrt{np(1-p)}$

- Real-world examples
  - Number of “no shows” for travel or appointments
  - Number of responses for survey emails or credit card invitation letters

Binomial Distribution in Excel

- Binomial Distribution Calculations in Excel
  - BINOM.DIST(x, *n*, *p*, cumulative)
    - cumulative = TRUE (or 1): $P(X \leq x)$
    - cumulative = FALSE (or 0): $P(X = x)$

Example: Supermarket Spending

Q1: If 500 customers shop in a given day, what is the mean number of customers who spend at least $100? What is the standard deviation?

$$Y_{500} \sim binomial\ (500, 0.3)$$

Example: Supermarket Spending

Q1: If 500 customers shop in a given day, what is the mean number of customers who spend at least $100? What is the standard deviation?

$$Y_{500} \sim binomial\ (500, 0.3)$$

$$E(Y_{500}) = np = 500(0.3) = 150$$

$$Stdev(Y_{500}) = \sqrt{np(1-p)} = \sqrt{500 * 0.3 * 0.7} = 10.25$$

Example: Supermarket Spending

Q2: What is the probability that at least 30% of customers (=150) spend at least $100?

We would like to compute

Note that BINORM.DIST function can handle either $P(Y_{500} = x)$ or $P(Y_{500} \leq x)$.

Example: Supermarket Spending

Q2: What is the probability that at least 30% of customers (=150) spend at least $100?

We would like to compute $P(Y_{500} \geq 150).$

Note that BINORM.DIST function can handle either $P(Y_{500} = x)$ or $P(Y_{500} \leq x)$.

$$P(Y_{500} \geq 150)$$
$$= 1 - P(Y_{500} < 150)$$
$$= 1 - P(Y_{500} \leq 149)$$

= 1 – BINOM.DIST(149, 500, 0.3, TRUE) = 0.5169=51.69%

Quiz 5

The service manager for a new appliance store reviewed sales records of the past 20 sales of new microwaves to determine the number of warranty repairs he will be called on to perform in the next 90 days. Corporate reports indicate that the probability that any one of their new microwaves needs a warranty repair in the first 90 days is 0.05. The manager assumes that calls for warranty repair are independent of one another and is interested in predicting the number of warranty repairs he will be called on to perform in the next 90 days for this batch of 20 new microwaves sold.

a) What type of probability distribution will most likely be used to analyze warranty repair needs on new microwaves in this situation?
b) What is the expected number of the new microwaves sold that will require a warranty repair in the first 90 days?
c) What is the standard deviation of the number of the new microwaves sold that will require a warranty repair in the first 90 days?
d) What is the probability that at most 3 microwaves will require a warranty repair in the first 90 days?

The service manager for a new appliance store reviewed sales records of the past 20 sales of new microwaves to determine the number of warranty repairs he will be called on to perform in the next 90 days. Corporate reports indicate that the probability that any one of their new microwaves needs a warranty repair in the first 90 days is 0.05. The manager assumes that calls for warranty repair are independent of one another and is interested in predicting the number of warranty repairs he will be called on to perform in the next 90 days for this batch of 20 new microwaves sold.

a) What type of probability distribution will most likely be used to analyze warranty repair needs on new microwaves in this situation?

Binomial (n, p)=Binomial (20, 0.05)

b) & c) What are the expected number and the standard deviation of the new microwaves sold that will require a warranty repair in the first 90 days?

$$Mean = \ n * p \ = 20 * 0.05 = 1$$

$VAR = np(1 - p) = 20 * 0.05 * (1 - 0.05)$ = 0.95

$$Stdev = \ \sqrt{20 * 0.05 * (1 - 0.05)} = 0.975$$

d) What is the probability that at most 3 microwaves will require a warranty repair in the first 90 days?

P (X <= 3) =BINOM.DIST (3, 20, 0.05,1) = 0.984

Example: Airline Overbooking

- Airlines often overbook flights because there is a high probability that there will be ‘no shows’
- We would like to assess the benefits and drawbacks of airline overbooking.
- Consider the scenario in which we estimate the no-show rate to be 10% for a flight with 200 seats
- Let X be the number of shows
  - X ~ binomial(n, p = 0.90)

Example: Airline Overbooking

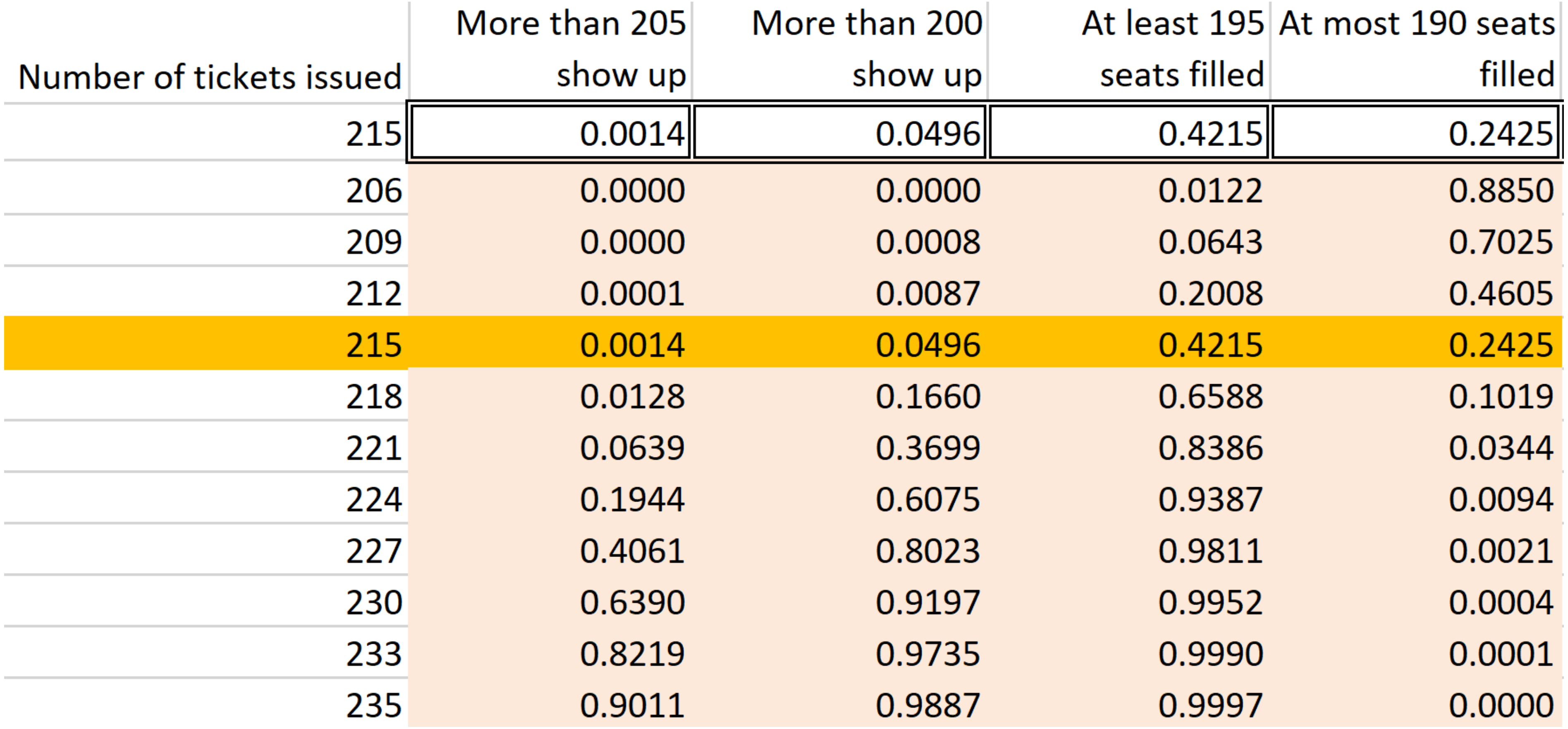
- Suppose that the airline sold 215 tickets
- Calculate the probability that
  - more than 205 passengers show up; that
  - more than 200 passengers show up; that
  - at least 195 seats are filled; and that
  - at most 190 seats are filled.
- Use the BINOM.DIST function to determine the probabilities.

Example: Airline Overbooking

- n=215, p=0.9
- Calculate the probability that
  - more than 205 passengers show up: 1-BINOM.DIST(205,n,p,1) = 0.0014
  - more than 200 passengers show up: 1-BINOM.DIST(200,n,p,1) = 0.0496
  - at least 195 seats are filled: 1-BINOM.DIST(194,n,p,1) = 0.4214
  - at most 190 seats are filled: BINOM.DIST(190,n,p,1) = 0.2425

Example: Airline Overbooking

For a flight with 200 seats how many tickets should the airline company issue?



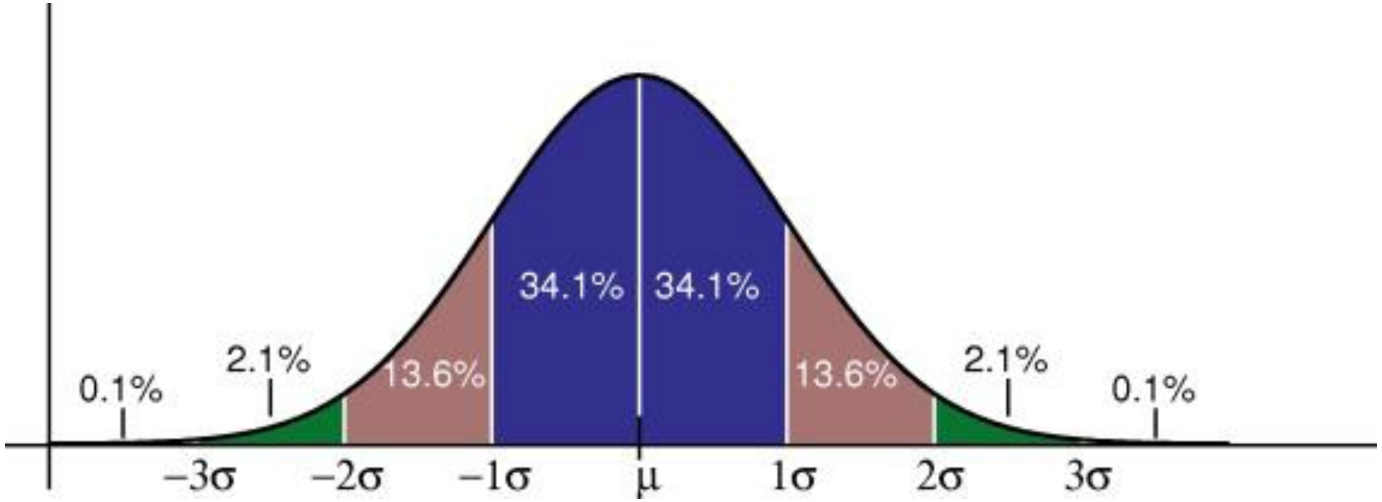
Number of tickets issued	More than 205 show up	More than 200 show up	At least 195 seats filled	At most 190 seats filled
215	0.0014	0.0496	0.4215	0.2425
206	0.0000	0.0000	0.0122	0.8850
209	0.0000	0.0008	0.0643	0.7025
212	0.0001	0.0087	0.2008	0.4605
215	0.0014	0.0496	0.4215	0.2425
218	0.0128	0.1660	0.6588	0.1019
221	0.0639	0.3699	0.8386	0.0344
224	0.1944	0.6075	0.9387	0.0094
227	0.4061	0.8023	0.9811	0.0021
230	0.6390	0.9197	0.9952	0.0004
233	0.8219	0.9735	0.9990	0.0001
235	0.9011	0.9887	0.9997	0.0000

Example: Airline Overbooking

- This model, including the data table, could be used to find the right tradeoff between selling too many tickets (bumping passengers) and not selling enough tickets (empty seats).
- However, it does make the assumption that passengers behave independently of one another and that each has the same probability of not showing up.
- The independence assumption would be violated, for example, if a family of four has tickets. Presumably, they would all show up or none of them would show up.
- If you don't make the binomial assumptions, then the model might be more realistic, but it would be considerably more difficult.

Ch5
Standard Probability Distributions

Normal Distribution



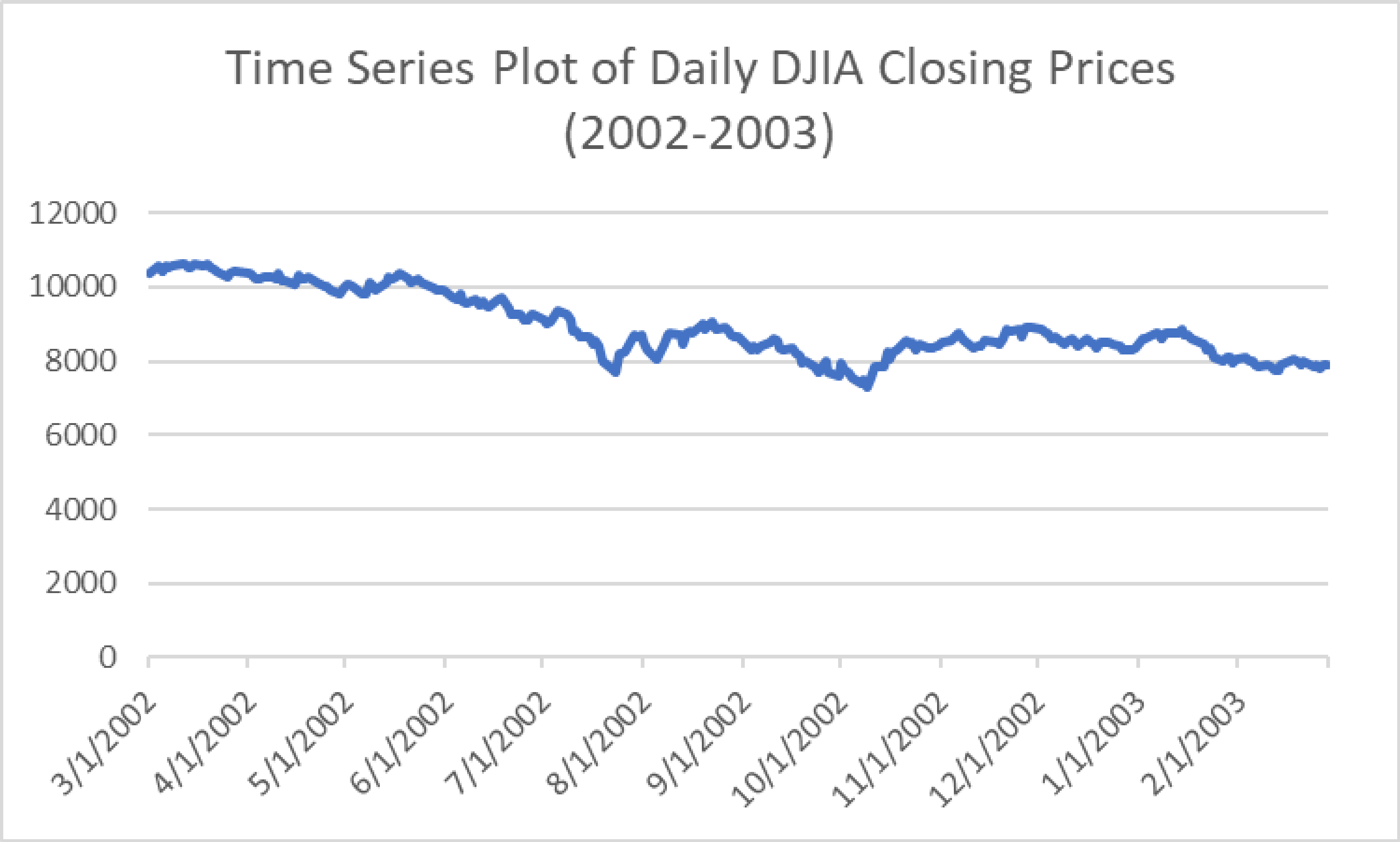
Learning Objectives

- **Normal Distribution** (Ch5)
  - Normal Distribution
  - Learn the properties of the Normal distribution
  - Compute the probability of normal distribution using Excel.
  - Learn how to interpret the result.
- We will use
  - DJIA 2002_2003.xlsx
  - DJIA 2018_2019.xlsx

Normal Distribution

- The single most important distribution in statistics, an assumption of normality is behind most of the procedures we use
  - For example, statistical inference and linear regression
- Real-world examples: investment returns, demand

Example: Dow Jones Industrial Average



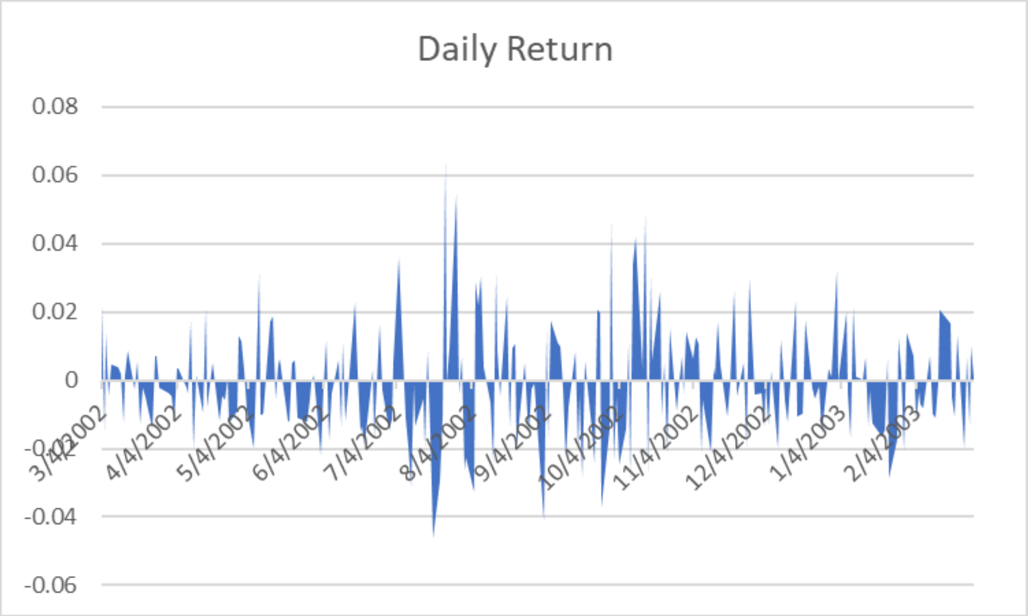
To obtain DJIA data:

- Go to: finance.yahoo.com
- Download into spreadsheet format
- Gives closing values

Example: Dow Jones Industrial Average

- What is the chance of a positive daily return?
- What is the risk of a negative daily return?
- We can calculate the daily returns:

$$\boxed{DR_{Today} = \frac{DJIA_{Today} - DJIA_{Yesterday}}{DJIA_{Yesterday}}}$$

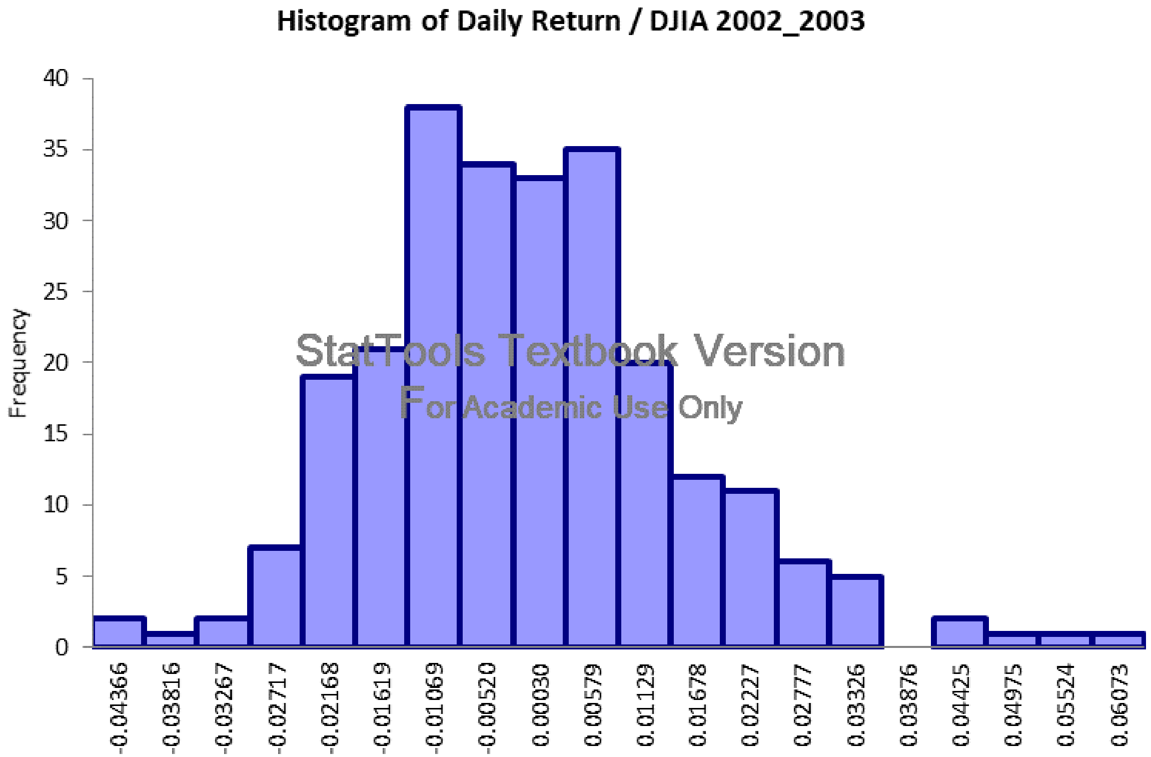


Example: Dow Jones Industrial Average

- What is the chance of a positive daily return?
- What is the risk of a negative daily return?

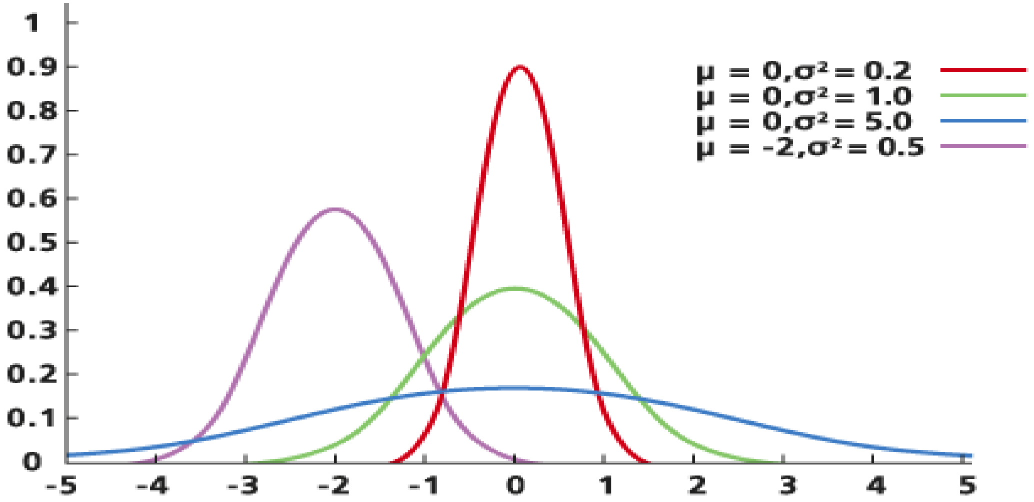
We can answer these questions if we assume daily returns are i.i.d.

In particular, we assume a Normal probability model for the daily returns


Histogram of Daily Return / DJIA 2002_2003

Normal Distribution (9/20(M))

- Continuous, symmetric, bell-shaped distribution
- Defined by two parameters: the **mean μ** and the **standard deviation σ** (denoted by $\boldsymbol{N(\mu,\sigma)}$)
- Changing the standard deviation makes the curve more or less spread out
- Range is $(-\infty, \infty)$ +∞; however, only a relatively small range is likely to occur.



The Normal PDF

- Possible values range from -∞ to +∞; however, only a relatively small range is likely to occur
    - Those within 3 standard deviations of the mean (99.7%)
- The normal PDF is actually quite complex, in spite of its nice bell-shaped appearance
    - Infeasible to integrate exactly, so we use tables or computers to tell us the probabilities

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty.$$

Standardizing: Z-Values

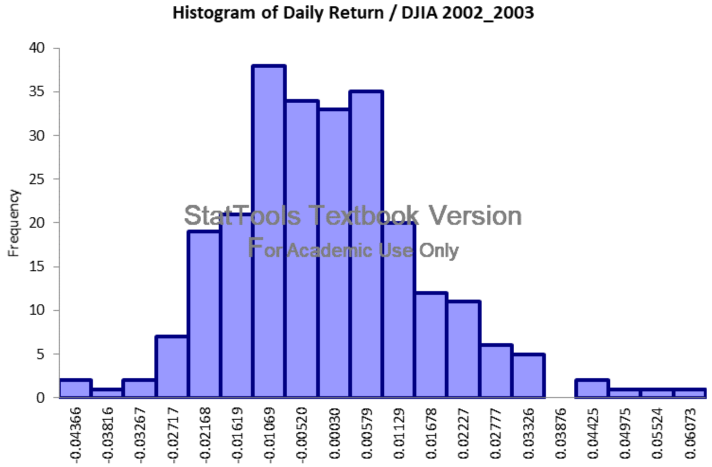
- There are infinitely many normal distributions, one for each pair of μ and σ.
- The **standard normal distribution** is a special case, which has mean 0 and standard deviation 1, so we denote it by $\boldsymbol{N(0,1)}$or the Z distribution
- To standardize, subtract the mean and divide by the standard deviation:

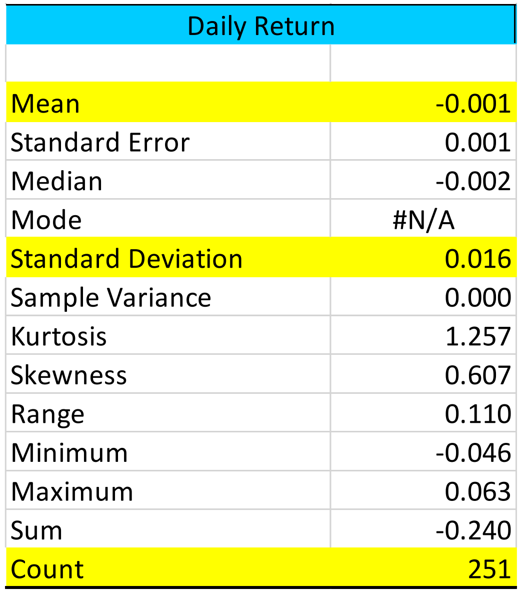
$$\boxed{\boldsymbol{Z = \frac{X-\mu}{\sigma}}}$$

when X is normally distributed, Z is $N(0,1)$.

- Standardizing allows measuring variables with different means and/or standard deviations on a single scale.

Example: Dow Jones Daily Returns





Daily Return	
Mean	-0.001
Standard Error	0.001
Median	-0.002
Mode	#N/A
Standard Deviation	0.016
Sample Variance	0.000
Kurtosis	1.257
Skewness	0.607
Range	0.110
Minimum	-0.046
Maximum	0.063
Sum	-0.240
Count	251

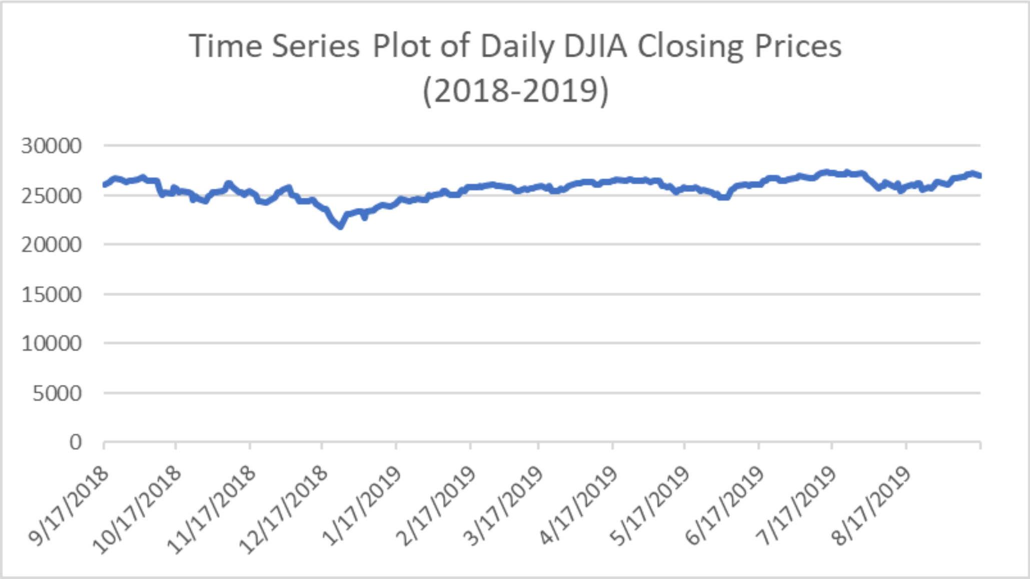
- Are the daily returns of the DJIA normally distributed?
- Since the histogram is well-approximated by a bell-shaped distribution, we assume that the Normal probability model provides a good approximation to the DJIA daily returns
- Moreover, we assume a Normal distribution with mean $\mu = -0.001$ and standard deviation $\sigma = 0.016$

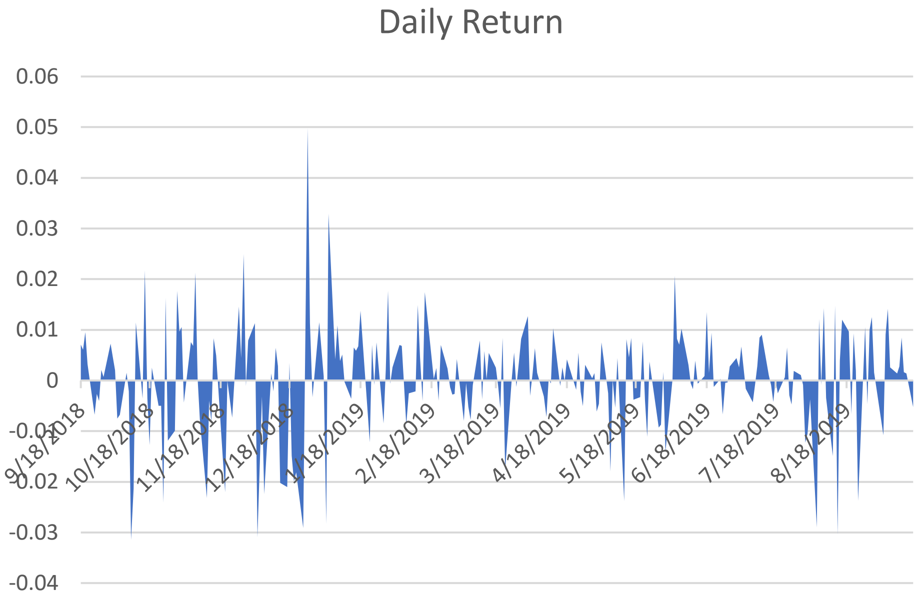
In-class Exercise: DJIA 8/2018-9/2019

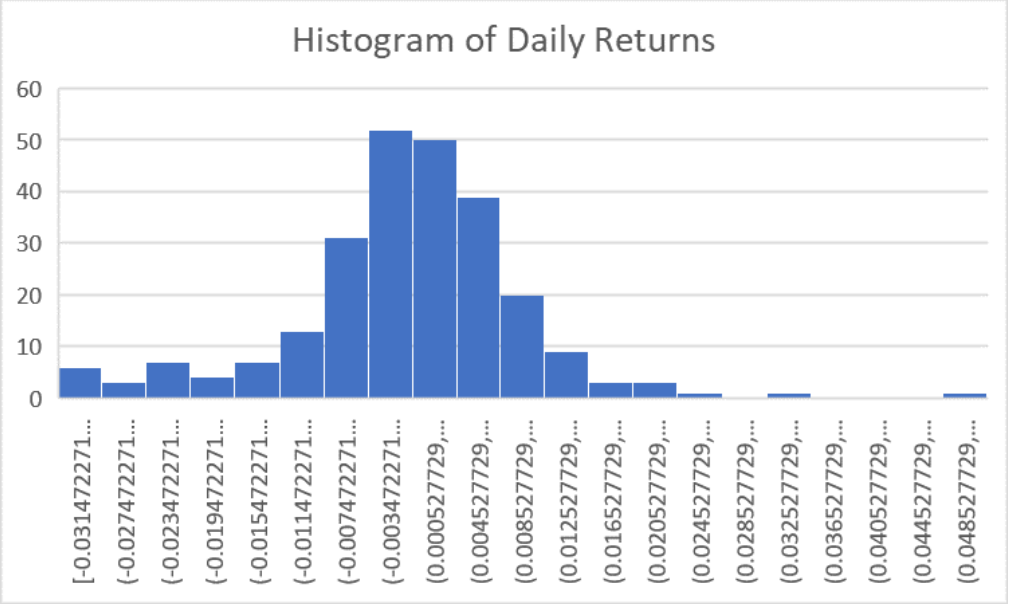
DJIA 2018_2019.xlsx

- Create time series plot of closing prices and daily return
- Create a histogram of daily returns
- Are the daily returns of the DJIA normally distributed?

DJIA 8/2018-9/2019







Normal Distribution Calculations in Excel: Cumulative Probability

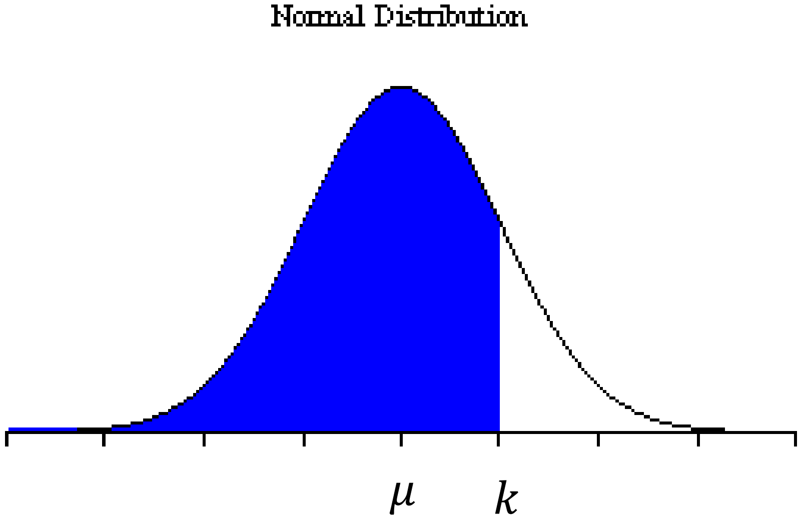
- Suppose we have a normal random variable with mean μ and standard deviation σ

$$X \sim N(\mu, \sigma)$$

- To find the **cumulative probability** in Excel:

$$P(X \ \leq \ k)$$
$$= \ NORM.DIST(k, \mu, \sigma, 1 \ (or\ TRUE))$$

The "1" makes sure it is the cumulative probability that you get

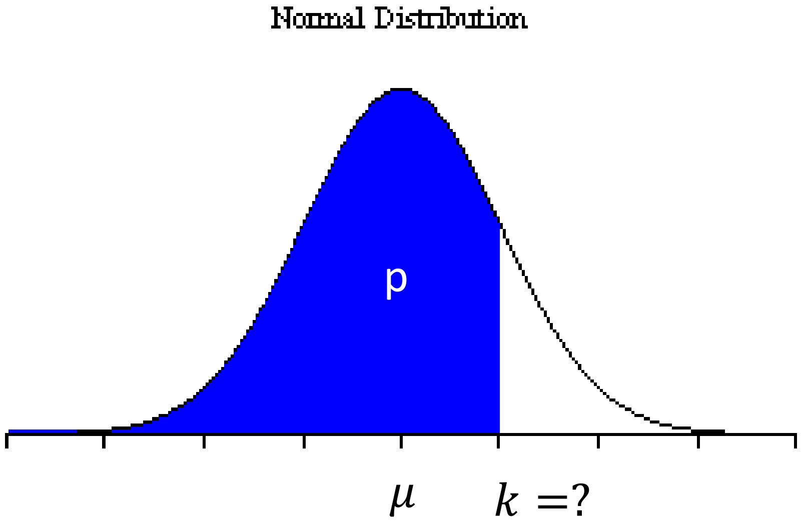


Note: In continuous distribution,

$$P(X \ \leq \ k)= P(X < \ k)$$

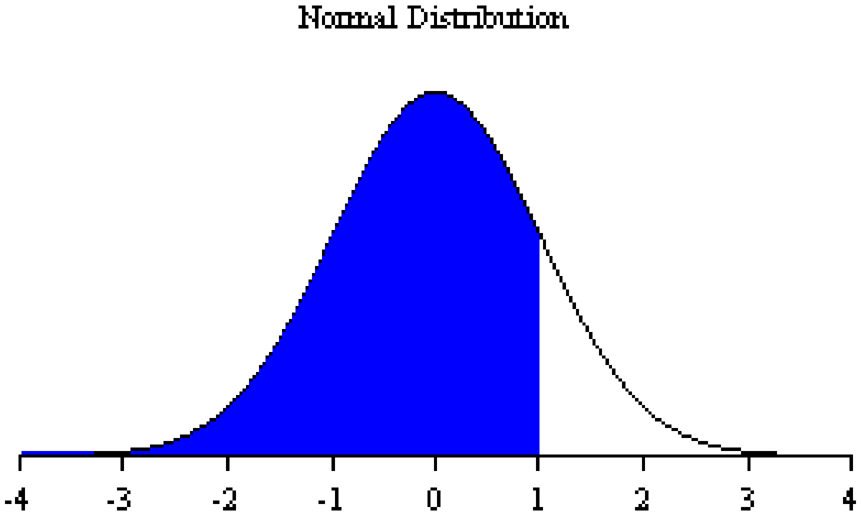
Normal Distribution Calculations in Excel: Percentile

- Suppose we have a normal random variable with mean μ and standard deviation σ
$$X \sim N(\mu, \sigma)$$
- Given a probability, p, what is the corresponding value k such that the probability of being below k *is* p:
$$P(X < k?) = p$$
- k corresponds to the **p*100th percentile** of X.
- To find the percentile in Excel:
$$k = NORM.INV(p, \mu, \sigma)$$



Normal Distribution Calculations in Excel

- For standard normal, use NORM.S.DIST(z, 1) and NORM.S.INV(p) functions
  - The “S” stands for standardized $(\mu = 0, \sigma = 1)$
  - To find the cumulative probability: $P(Z \le k) = NORM.S.DIST(k)$
  - Given a probability, p, the p*100th percentile of $Z = NORM.S.INV(p)$



Example: DIJA Questions

We assume the daily returns of the DJIA follow normal distribution:

$$\mathrm{X} = Normal(\mu = -0.001, \sigma \ = \ 0.016)$$

- What is the probability that DJIA daily return is positive?

- What is the probability that daily return of DJIA is more than 5%?

- What are the lowest 5% of all possible DJIA returns?

In other words, what return would be less than 95% of all returns?

P(X < k?) = 5% => k is the 5th percentile

Example: DIJA Questions

We assume the daily returns of the DJIA normal distributed:
$\mathrm{X} = N(\,\mu = -0.001, \sigma \;=\; 0.016)$

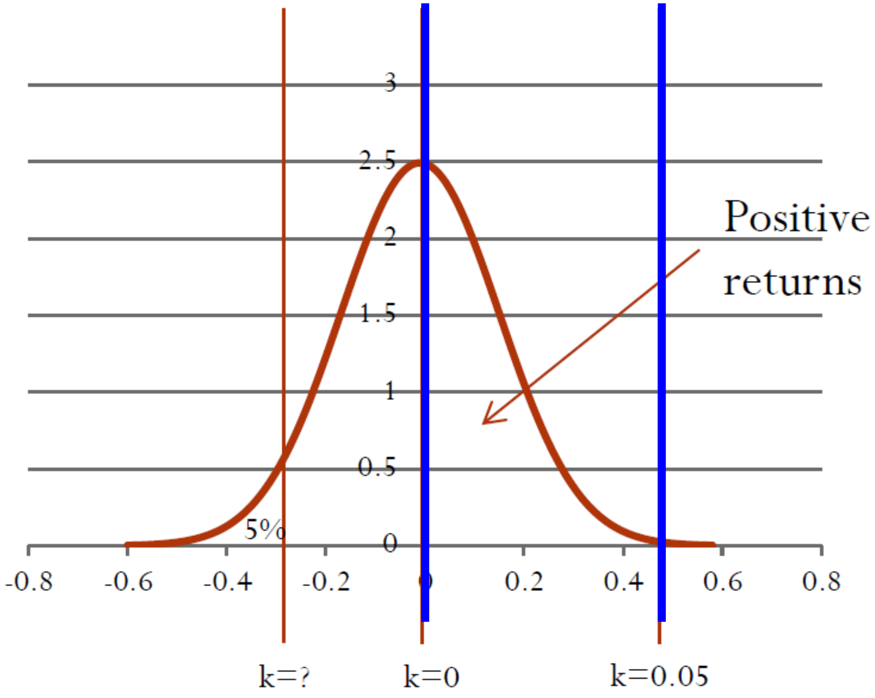
- What is the probability that DJIA daily return is positive?
  - ○ P(X > 0) = 1- NORM.DIST(0,-0.001,.016,1) = 47.50%
  - ○ $P(X \;>\; 0) = \; P\left(\frac{X-(-0.001)}{0.016} > \frac{0-(-0.001)}{0.016}\right)$

$$= P\left(Z > \frac{0.001}{0.016}\right)$$

$$= 1 - \; NORM.S.DIST(0.0625{,}1) \;=\; 47.50\%$$

- What is the probability that daily return of DJIA is more than 5%?
  - ○ P(X > 0.05) = 1- NORM.DIST(0.05,-0.001,.016,1) = 0.07%
  - ○ $P(X \;>\; 0.05) \;= P\left(Z > \frac{0.05+0.001}{0.016}\right)$

$$= 1 - \; NORM.S.DIST(3.1875{,}1) \;=\; 0.07\%$$



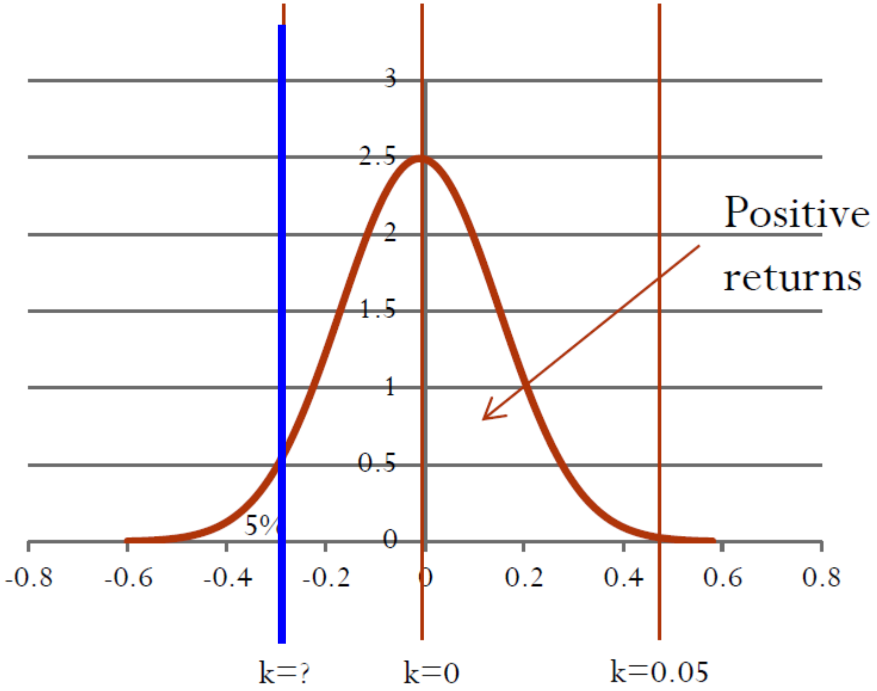
Example: DIJA Questions

We assume the daily returns of the DJIA normal distributed: $X = N(\mu = -0.001, \sigma = 0.016)$

- What are the lowest 5% of all possible DJIA returns?

In other words, what return would be less than 95% of all returns?

- ○ P(X < k?) = 5%
- ○ k = NORM.INV(0.05,-0.001,.016) = -2.73%
- ○ The lowest 5% of all possible DJIA returns are –2.73% or lower.
- ○ $P\left(\frac{X-(-0.001)}{0.016} < \frac{k-(-0.001)}{0.016}\right) = 5\%$
- ○ $\frac{k-(-0.001)}{0.016}$=NORM.S.INV(0.05)=-1.645



Example: DIJA Questions

- What is the probability that daily return of DJIA is between 1% and 3%

- What return would be less than 25% of all returns?

DIJA Questions

Answer the following questions:

- What is the probability that daily return of DJIA is between 1% and 3%

  P(0.01<X<0.03) = P(X<0.03) – P(X<0.01)

  = NORM.DIST(0.03, -0.001,.016, TRUE)

  – NORM.DIST(0.01, -0.001,.016, TRUE) =0.2195

- What return would be less than 25% of all returns?

  Find k such that P(X<k)= 0.75 => 75th percentile

  k = NORM.INV(0.75, -0.001,.016) =0.0098

- Also try to solve these problems using standard normal (Z)

Normal Approximation

Normal Approximation to Binomial

Normal Approximation to Sum of RVs

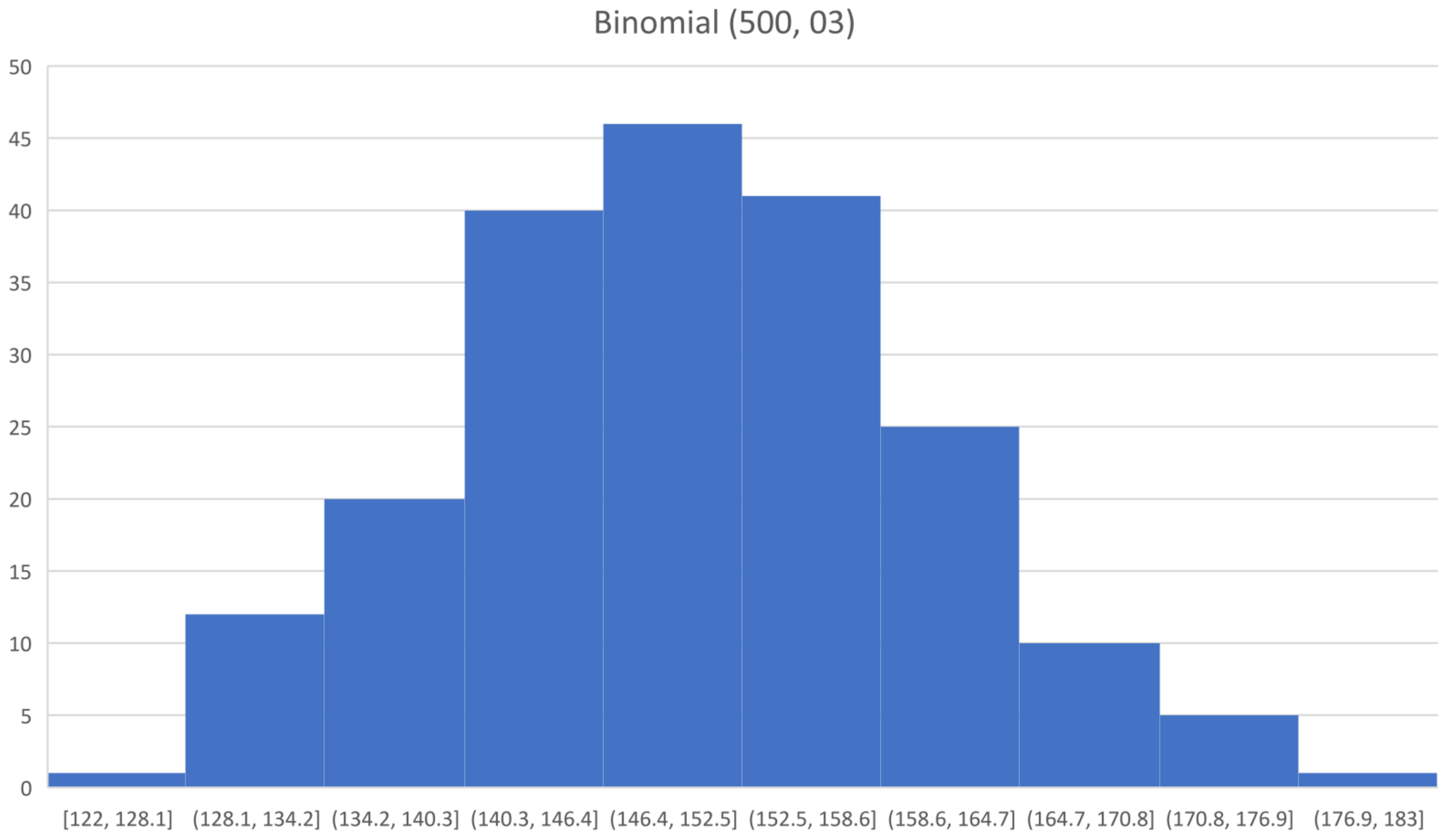
Example: Supermarket Spending (Probability Version)

- Historical data suggests that customer spending at a supermarket is randomly distributed. It is also known that the **probability** that one customer spends **at least $100 is 0.3**.
- Question: What is the probability that at least 30% of customers (=150) spend at least $100?

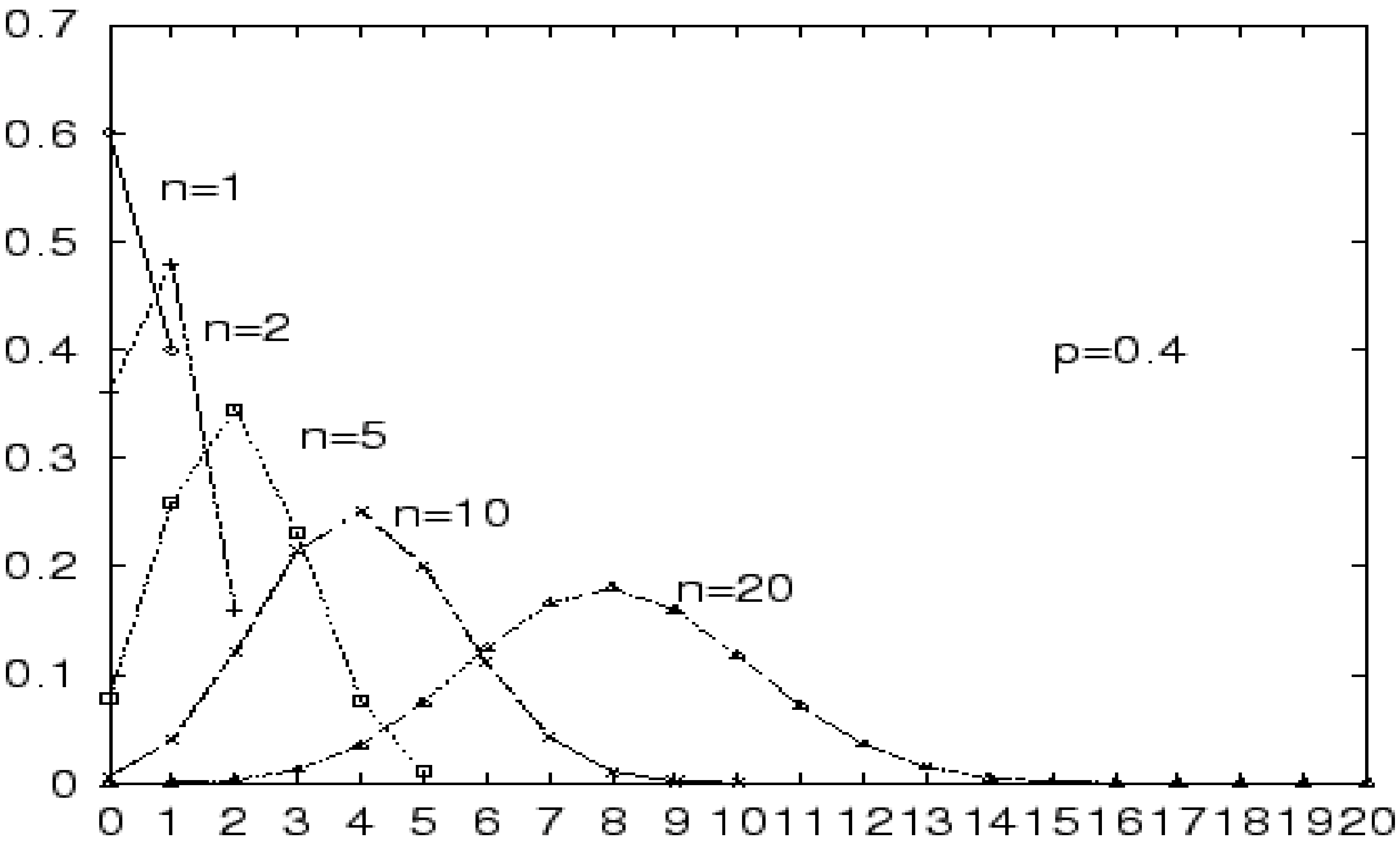
$$P(Y_{500} \geq 150) = 1 - P(Y_{500} \leq 149)$$

= 1 – BINOM.DIST(149, 500, 0.3, TRUE) = 0.5169=51.69%

Binomial (500, 0.3)

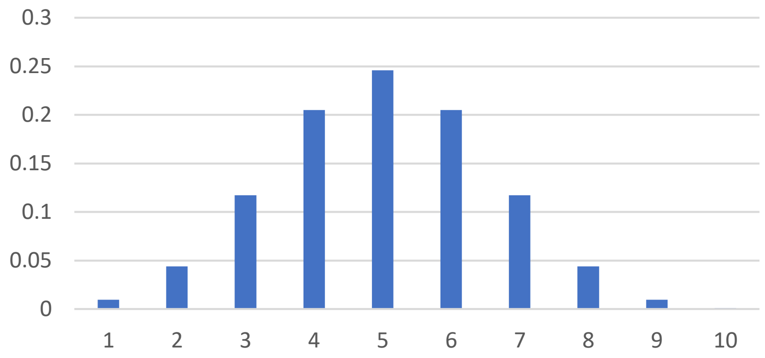


Binomial Probability Distribution

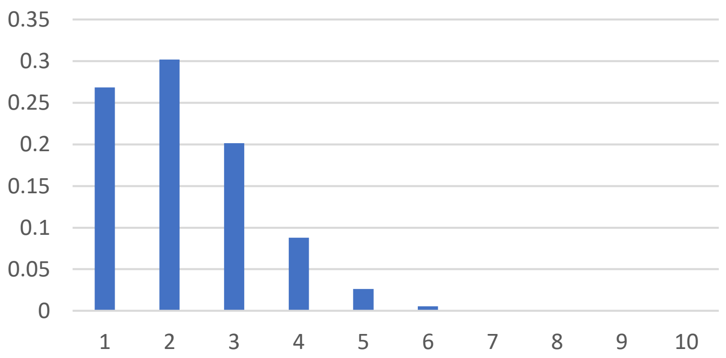


Binomial Probability Distribution

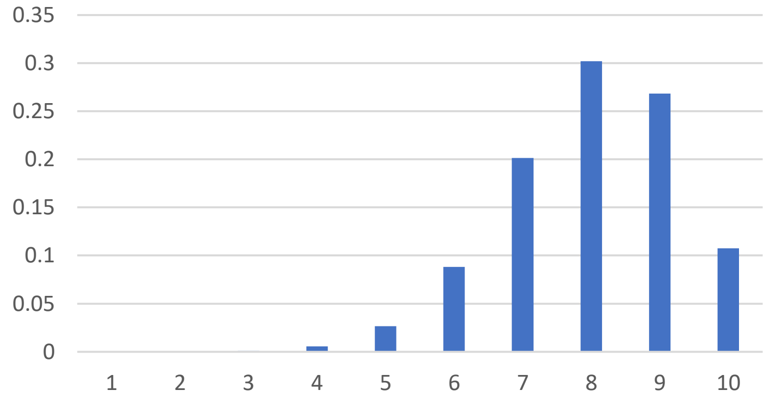
Binomial(10 , 0.5)



Binomial(10, 0.2)

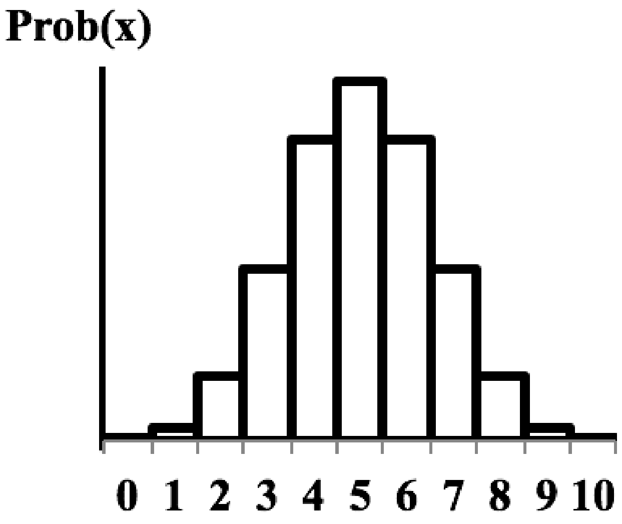


Binomial(10, 0.8)

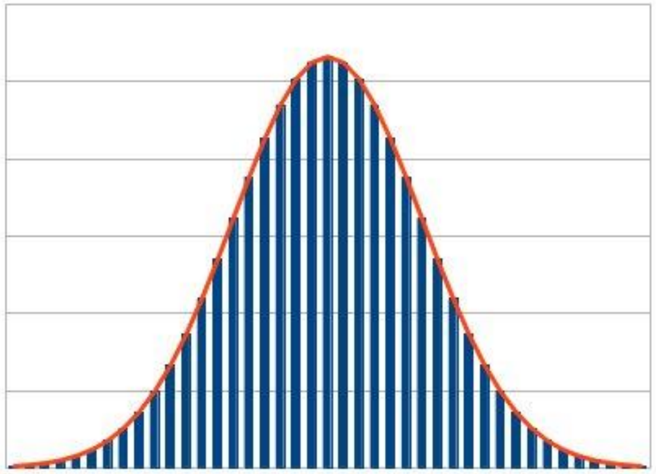


The Normal Approximation to the Binomial

- The graph of binomial probabilities begins to look symmetric and bell-shaped when *n* is fairly large and *p* is not too close to 0 or 1 (np>5 or n(1-p)>5).
- The normal distribution provides a very good approximation to the binomial under these conditions.



$$\boldsymbol{Binomial\ (n, p)}$$



$$\boldsymbol{Normal\ \left(np, \sqrt{np(1-p)}\right)}$$

Example: Supermarket Spending (Probability Version)

- Historical data suggests that customer spending at a supermarket is randomly distributed. It is also known that the **probability** that one customer spends **at least $100 is 0.3**.
- Question: What is the probability that at least 30% of customers (=150) spend at least $100? Apply the normal approximation.

$$P(Y_{500} \geq 150) = 1 - P(Y_{500} \leq 149)$$

= 1 – BINOM.DIST(149, 500, 0.3, TRUE) = 0.5169=51.69%

$\approx 1 - NORM.DIST$(149, 150, 10.25, TRUE)=53.89%

Sum of Independent Normal RVs

- Suppose that X and Y are independent normal random variables such that

$$X \sim Normal\ (\mu_X, \sigma_X)\ and$$
$$Y \sim Normal\ (\mu_Y, \sigma_Y).$$

- Then,

$$X + Y \sim Normal(\mu_X, +\mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2})$$

Example: Supermarket Spending (Revenue Version)

- Suppose that customer spending at the supermarket is independent and identically distributed (i.i.d.) with the mean of $85 and the standard deviation of $30.

- If 500 customers shop in a given day, what is the probability distribution of the revenue of the supermarket on that day?

Sum of i.i.d. Random Variables

- The total amount spent by 500 customers is

$$Y_{500} = X_1 + X_2 + \cdots + X_{500}$$

- Then,

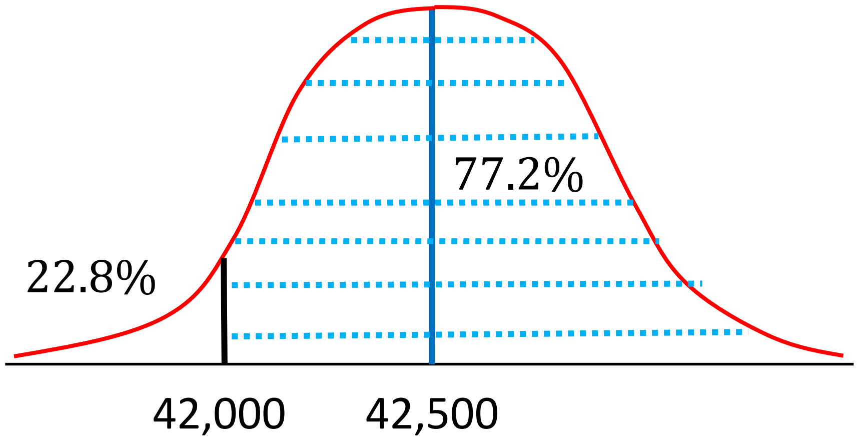
$$Y_{500} \approx Normal(\$42{,}500, \$671)$$

  - Y_{500} is approximately normally distributed.
  - If X_i's are normally distributed, Y_{500} is normally distributed.
  - In general, the distribution of a sum of large number (≥ 30) of independent and identically distributed (i.i.d.) random variables is approximately normal

Sum of i.i.d. Random Variables

- What is the probability that the daily revenues is at least $42,000?

$$
\begin{aligned}
P(Y_{500} \geq \$42{,}000) &= P(Normal(\$42{,}500, \$671) \geq 42{,}000) \\
&= 1 - P(Normal(\$42{,}500, \$671) < 42{,}000) \\
&= 1 - \boldsymbol{NORM.DIST(42,000, 42,500, 671, 1)} \\
&= 1 - 22.8\% = 77.2\%
\end{aligned}
$$


77.2%
22.8%
42,000
42,500