# EE 381 Probability & Statistic with Applications to Computing (Fall 2020)

## Lecture 9
## Continuous Random Variables (p2)

Duc H. Tran, PhD

# Continuous Random Variable’s topics

- Probability distributions, Expectation
- Variance, Covariance & Correlation
- **Uniform Distribution, Normal Distribution**
- **Other Distributions, Central Limit Theorem**

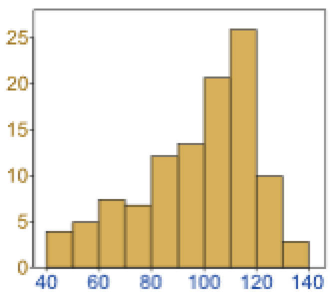
# Midterm 2’s Announcement

- Date: Thursday 11/12/20
- Time: 9:30AM -11:40AM
- Coverage: Lectures 1-9, HWs 1-9.
- Tuesday class on that week (11/10/20) will be used as the review section.

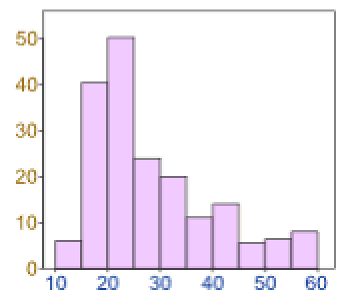
# The Normal Distribution

- Recall: Data can be “distributed” (spread out) in different ways.

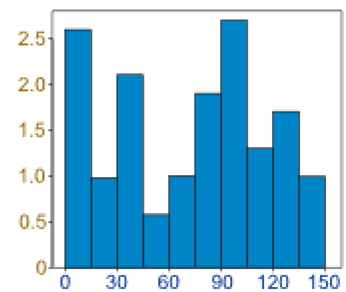
It can be spread out more on the left

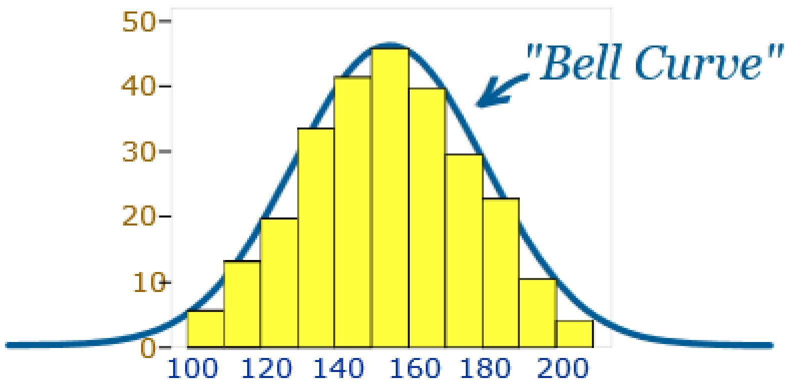



Or more on the right




Or it can be all jumbled up




A Normal Distribution

In cases where data tends to be symmetric around a central value, that data is normally distributed.

We call it normal distribution.



These figures are extracted from: https://www.mathsisfun.com/data/standard-normal-distribution.html

# The Normal Distribution

- The *normal distribution*, also called *Gaussian distribution*, is one of the most important examples of continuous probability distribution.
- It show that data near the mean are more frequent in occurrence than data far from the mean.
- Its density function is given by:

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

Where $\mu$ and $\sigma$ are the mean and standard deviation, respectively.

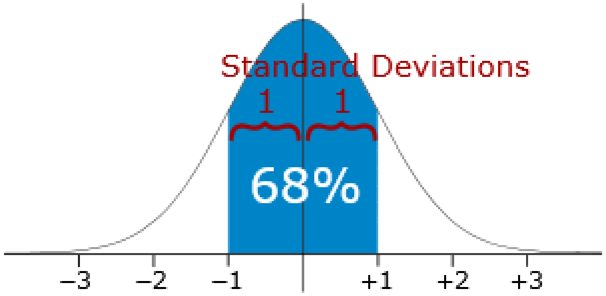
- The corresponding distribution function is given by:

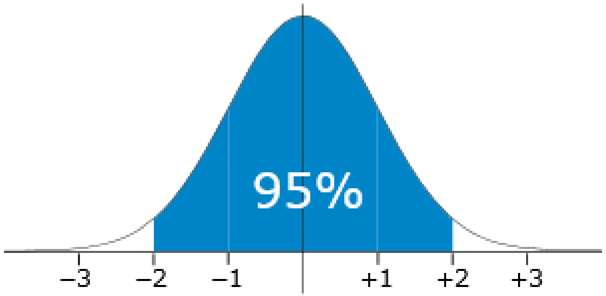
$$F(x) = P(X \leq x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\sigma^2}} \, \mathrm{dv}$$

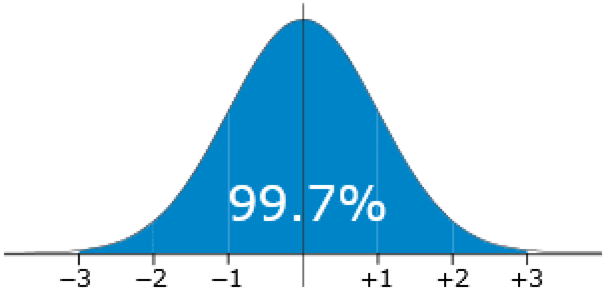
If *X* has the above distribution function *F(x)* , we say that random variable *X* is normally distributed with mean $\mu$ and variance $\sigma^2$.

# The Normal Distribution

- Recall: The standard deviation is a measure of how spread out numbers are.
- In normally distributed data, we found that:




68% of values are within 1 standard deviation of the mean

95% of values are within 2 standard deviation of the mean

99.7% of values are within 3 standard deviation of the mean

These figures are extracted from: https://www.mathsisfun.com/data/standard-normal-distribution.html

# The Normal Distribution

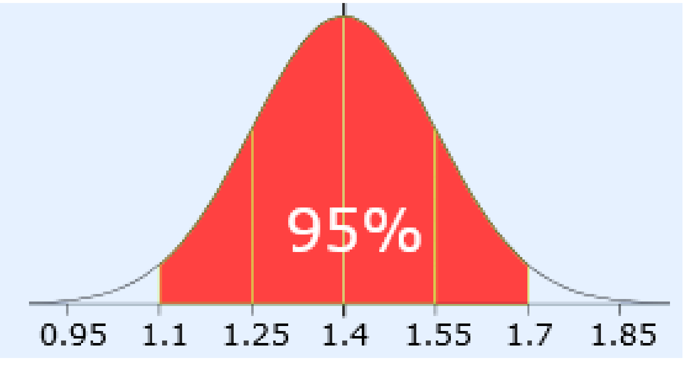
**Example**: 95% of student at a school are between 1.1m and 1.7m tall. Assuming this data is normally distributed, calculate the mean and standard deviation.

**Solution**:

The mean is: *(1.1 + 1.7)/2 = 1.4m*

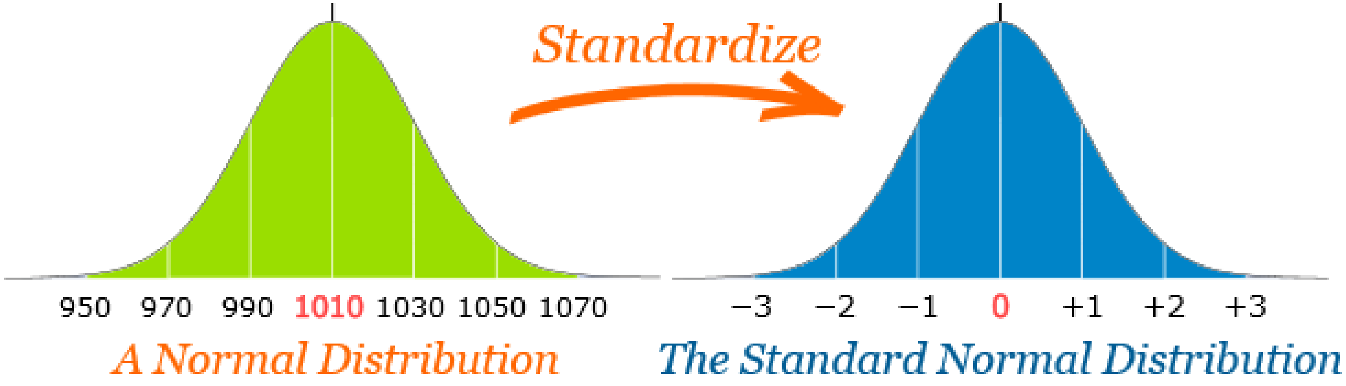
95% is 2 standard deviation either side of the mean (a total of 4 standard deviations) so:

Standard deviation: *(1.7-1.1)/4 = 0.15m*



# The Normal Distribution

- It would be more useful to convert a normal distribution to a standard normal distribution: easier and faster to calculate the required probability (based on Standard Normal Distribution table, Appendix C)
- This process is called standardizing.




These figures are extracted from: https://www.mathsisfun.com/data/standard-normal-distribution.html

# The Normal Distribution

- If we let Z be the standardized variable corresponding to X, i.e., if we let:

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

Then the mean of $Z$ is 0, and the variance is 1.

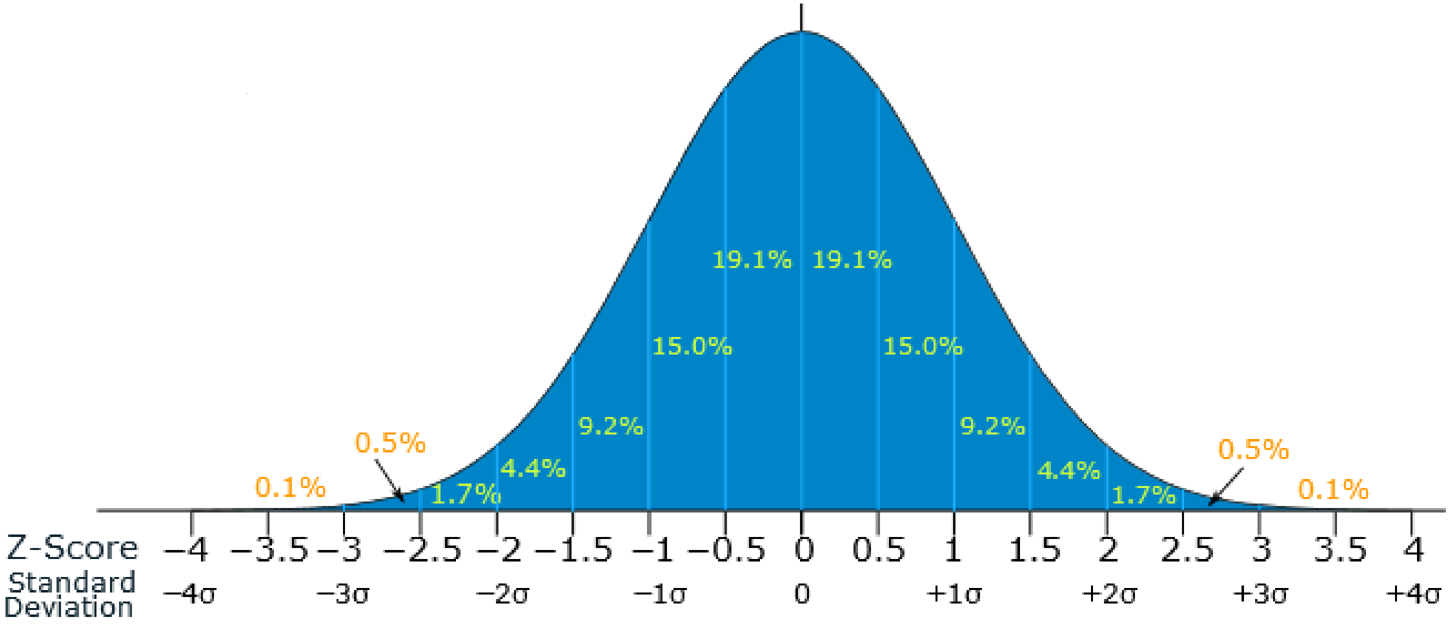
In such cases, the density function for Z will be:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

This is often called *standard normal density function*. Its corresponding distribution function is given by:

$$F(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{z} e^{-\frac{u^2}{2}} du$$

# The Normal Distribution




Standard normal distribution curve

These figures are extracted from: https://www.mathsisfun.com/data/standard-normal-distribution.html

# The Normal Distribution

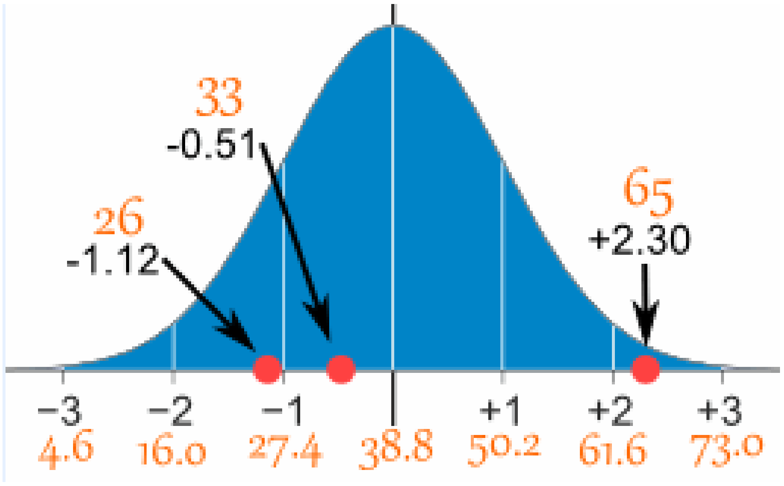
**Example**: A survey of daily travel time had these results (in minutes):

26, 33, 65, 28, 34, 55, 25, 44, 50, 36, 26, 37, 43, 62, 35, 38, 45, 32, 28, 34.

The mean is 38.8 minutes, and the standard deviation is 11.4 minutes. Convert them to standardized values.

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

| x | $\frac{x - \mu}{\sigma}$ | z (z-score) |
|---|---|---|
| 26 | $\frac{26 - 38.8}{11.4}$ | $= -1.12$ |
| 33 | $\frac{33 - 38.8}{11.4}$ | $= -0.51$ |
| 65 | $\frac{65 - 38.8}{11.4}$ | $= +2.30$ |
| ... | ... | ... |



# The Normal Distribution

- The value of $z$ of the standardized variable $Z$ is also called *standard score*.
- The function $F(z)$ is related to the extensive tabulated error function, $erf(z)$. We have:

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du \quad \text{and} \quad F(z) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

# The Normal Distribution

**Example**: Find the area under the standard normal curve:

a) Between *z* = *0* and *z* = *1.2*

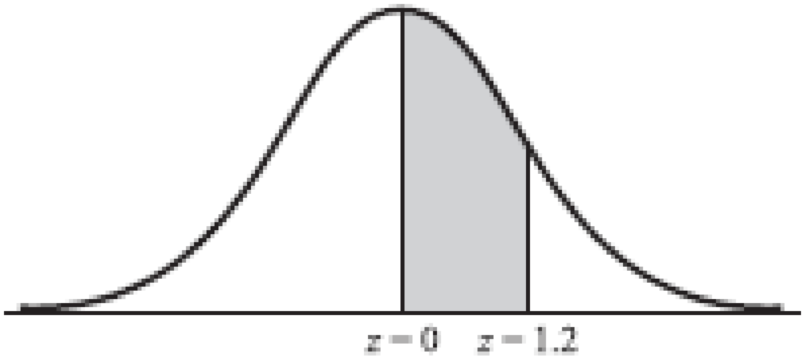
b) Between *z* = -*0.68* and *z* = *0*

c) Between *z* = -*0.46* and *z* = *2.21*

d) Between *z* = *0.81* and *z* = *1.94*

**Solution**

a)



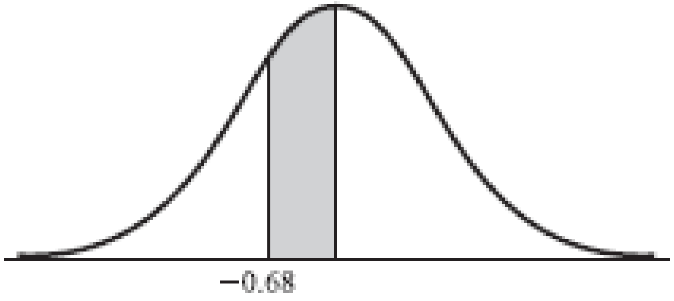

$$P(0 \leq Z \leq 1.2) = \frac{1}{\sqrt{2\pi}}\int_0^{1.2} e^{-u^2/2}\, du = 0.3849$$

Or using the standardized normal distribution table (Appendix C).

# The Normal Distribution

**Example (ctn.)**:

b) Between *z* = *-0.68* and *z* = *0*

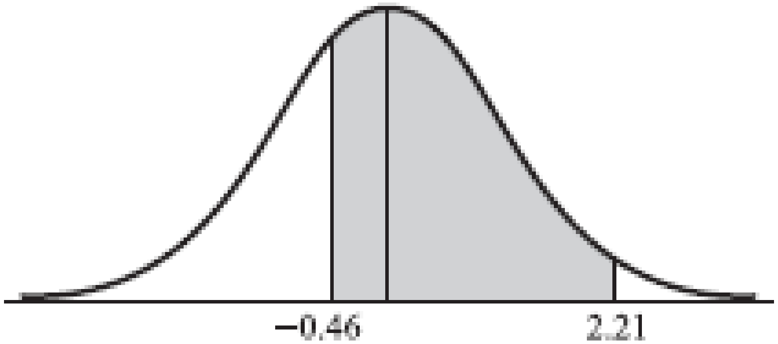



$$P(-0.68 \leq Z \leq 0) = \frac{1}{\sqrt{2\pi}} \int_{-0.68}^{0} e^{-u^2/2} du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{0}^{0.68} e^{-u^2/2} du = 0.2517$$

Or using the standardized normal distribution table (Appendix C).

c) Between *z* = *-0.46* and *z* = *2.21*

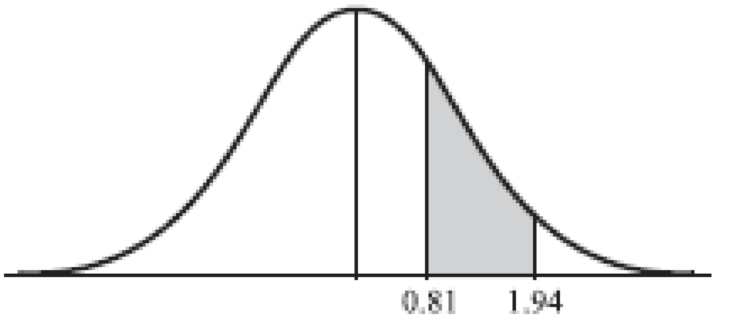



$$\begin{aligned} \text{Required area} &= (\text{area between } z = -0.46 \text{ and } z = 0) \\ &\quad + (\text{area between } z = 0 \text{ and } z = 2.21) \\ &= (\text{area between } z = 0 \text{ and } z = 0.46) \\ &\quad + (\text{area between } z = 0 \text{ and } z = 2.21) \\ &= 0.1772 + 0.4864 = 0.6636 \end{aligned}$$

# The Normal Distribution

**Example (ctn.)**:

b) Between $z = 0.81$ and $z = 1.94$




$$\begin{aligned}\text{Required area } &= (\text{area between } z = 0 \text{ and } z = 1.94) \\ &\quad - (\text{area between } z = 0 \text{ and } z = 0.81) \\ &= 0.4738 - 0.2910 = 0.1828\end{aligned}$$

# In-class exercise

If "area" refers to that under the standard normal curve, find the values of $z$ such that:

a) Area between 0 and $z$ is 0.35

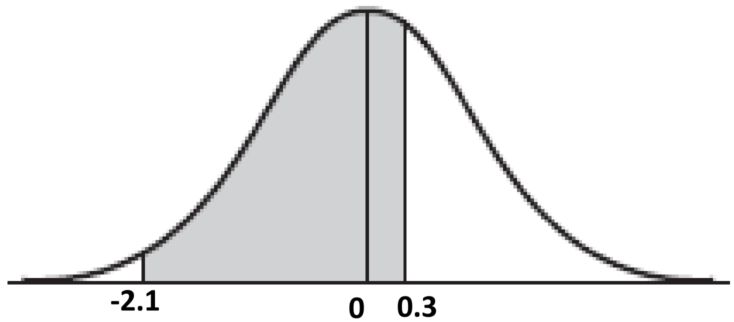
b) Area to the left of $z$ is 0.8665

# The Normal Distribution

**Example:** The mean weight of 500 male students at a certain college is 151 lb and the standard deviation is 15 lb. Assuming that the weights are normally distributed, find how many students weigh:

a) Between 120 and 155 lb.

b) More than 185 lb.




**Solution**:

a) Standardizing:

120 lb in standard units = (119.5-151)/15 = -2.1

155 lb in standard units = (155.5 – 151)/15 = 0.3

Required proportion of students

= (area btw $z=-2.1$ and $z=0$) + (area btw $z=0$ and $z=0.3$)

=0.4821 + 0.1179 = **0.6**

Then the number of students weighing between 120 and 155lb is **500*0.6=300**

# The Normal Distribution

**Example (ctn.):**

b) Number of students weigh more than 185 lb.

Students weighing more than 185 lb must weigh at least 185.1 lb. (Let's choose 185.5 lb).

185.5 lb in standard units = (185.5-151)/15 = 2.30

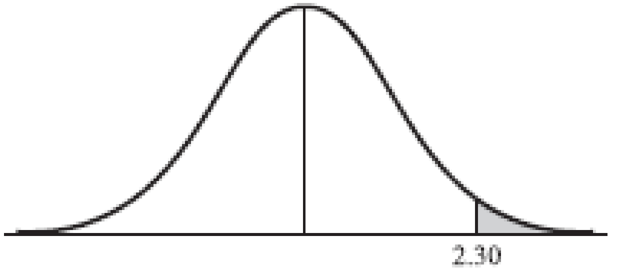
Required proportion of students

= (area to the right of $z=2.30$)

= (area to the right of $z=0$) - (area btw $z=0$ and $z=2.30$)

=0.5 - 0.4893 = **0.0107**

Then the number of students weighing more than 185lb is **500*0.0107=5**



# The Normal Distribution

**Continuity correction factor**: we are using continuous distribution (take all real number) to approximate discrete distribution which only take integers, so we have to have a small correction.

Continuity correction factor table:

- If $P(X = n)$ use $P(n - 0.5 < X < n + 0.5)$
- If $P(X > n)$ use $P(X > n + 0.5)$
- If $P(X \geq n)$ use $P(X > n - 0.5)$
- If $P(X < n)$ use $P(X < n - 0.5)$
- If $P(X \leq n)$ use $P(X < n + 0.5)$

# Some properties of the Normal Distribution

| | |
|---|---|
| Mean | $\mu$ |
| Variance | $\sigma^2$ |
| Standard Deviation | $\sigma$ |
| Moment Generating Function | $M(t) = e^{ut+\frac{\sigma^2 t^2}{2}}$ |

# Relation between Binomial and Normal Distributions

- If $n$ is large and if neither $p$ nor $q$ is too close to zero, the binomial distribution can be closely approximated by a normal distribution with standardized random variable given by:

$$Z = \frac{X - np}{\sqrt{npq}}$$

where $X$ is the random variable giving the number of successes in $n$ Bernoulli trials and $p$ is the probability of success.

- In practice, the approximation is very good if both $np$ and $nq$ are greater than 5.
- The binomial distribution approaches the normal distribution can be described by:

$$\lim_{n\to\infty} P\left(a \leq \frac{X - np}{\sqrt{npq}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{u^2}{2}} du$$

This also means that the standardized random variable $\frac{X-np}{\sqrt{npq}}$ is *asymptotically normal* (which means it converge to normal distribution) as $n \to \infty$.

# Relation between Binomial and Normal Distributions

**Example**: Find the probability of getting between 3 and 6 heads inclusive in 10 tosses of a fair coin by using:

a) The binomial distribution. Recall $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

b) The normal approximation to the binomial distribution

Solution:

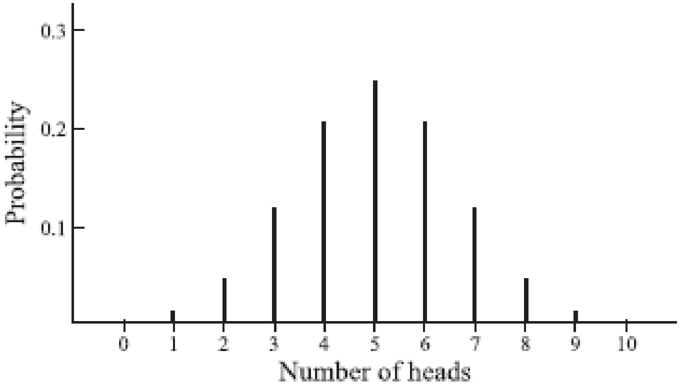
a) Let X be the random variable giving the number of heads that will turn up in 10 tosses.

$$P(X=3) = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \frac{15}{128} \qquad P(X=4) = \binom{10}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^6 = \frac{105}{512}$$

$$P(X=5) = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 = \frac{63}{256} \qquad P(X=6) = \binom{10}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4 = \frac{105}{512}$$

The required probability:

$$P(3 \le X \le 6) = \frac{15}{128} + \frac{105}{512} + \frac{63}{256} + \frac{105}{512} = \frac{99}{128} = 0.7734$$



# Relation between Binomial and Normal Distributions

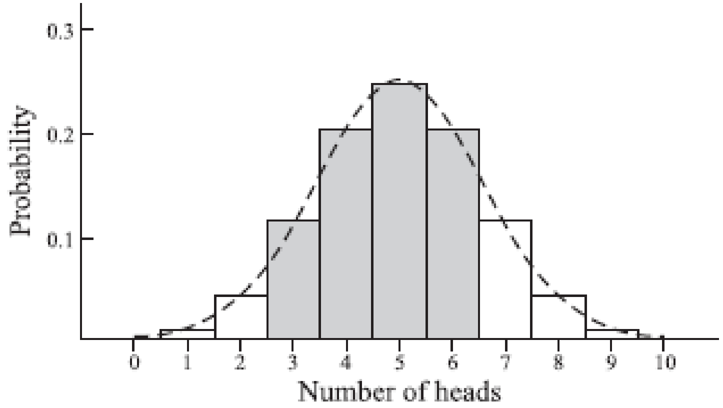
**Example (ctn.)**:

b) Let treat the data as if they were continuous. Then the required probability is the sum of shaded area. $\mu = np = 10 * \frac{1}{2} = 5$ and $\sigma = \sqrt{npq} = 1.58$.

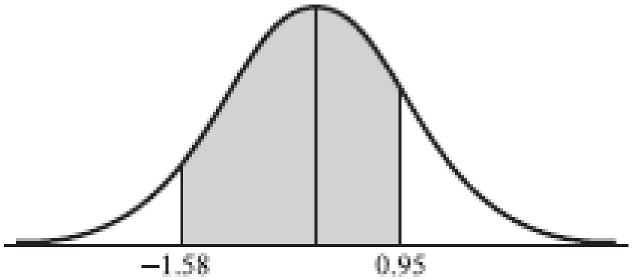
Treating as continuous data, 3 to 6 heads can be considered as 2.5 to 6.5 heads.

*2.5 in standard units = (2.5-5)/1.58 = -1.58*

*6.5 in standard units = (6.5-5)/1.58 = 0.95*




*P = (area btw z=-1.58 and z=0) + (area btw z=0 and z=0.95) = 0.4429+0.3289 =* ***0.7718***



# The Poisson Distribution

- The *Poisson distribution* is a **discrete probability distribution** that expresses the probability of a given number of events occurring in a fixed interval of time or space.
- The *Poisson distribution* is a probability function that results from the *Poisson experiment*. The experiment has the following properties:
  - ✓Events occur independently.
  - ✓The average number of successes that occurs in a specified region (fixed interval of time, a space, a length, an area, a volume…) is known.
  - ✓The probability that a success will occur is proportional to the size of the region.

**Example**: We can use Poisson distribution to model events such as:

- The number of patients arriving in an emergency room btw 10 and 11pm.
- The number of cancers in a certain area.
- The number of phone calls being made at any given time.
- …

# The Poisson Distribution

- Let X be a discrete random variable that can take on the values 0, 1, 2,… such that the probability function of X is given by

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0,\ 1,\ 2,\ \ldots$$

where $\lambda$ is the given mean ($\mu$) of $X$.

$e = 2.71828$, which is the base of the natural logarithm system

- This is called the *Poisson distribution*, and $X$ is said to be *Poisson distributed*.

| | |
|---|---|
| Mean | $\mu = \lambda$ |
| Variance | $\sigma^2 = \lambda$ |
| Standard Deviation | $\sigma = \sqrt{\lambda}$ |
| Moment Generating Function | $M(t) = e^{\lambda(e^t - 1)}$ |

# The Poisson Distribution

**Example**: The average number of homes sold by the ABC Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

**Solution**:

2 homes are sold per day on average → $\mu = \lambda = 2$

We want to find the probability that 3 homes will be sold tomorrow → *x = 3*

*Poisson distribution:*

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

$$P(X = 3) = f(3) = \frac{2^3 2.71828^{-2}}{3!} = 0.18$$

Thus, the probability of selling 3 homes tomorrow is 0.18.

# In-class Exercise

The average number of homes sold by the ABC Realty company is 2 homes per day. What is the probability that this company will *sell fewer than 2 homes in the next day*???

# Relation between the Binomial and Poisson Distributions

- Consider the *rare event* in binomial distribution: $n$ is large ($n \geq 50$), probability $p$ of occurrence of an event is close to 0, and $q = 1\text{-}p$ is close to 1.
- For such cases, the binomial distribution is very closely approximated by the Poisson distribution with $\lambda = np$.

# Relation between the Poisson and Normal Distributions

- If X is the Poisson random variable and $\frac{(X-\lambda)}{\sqrt{\lambda}}$ is the corresponding standardized random variable, then:

$$\lim_{\lambda \to \infty} P\left(a \leq \frac{(X-\lambda)}{\sqrt{\lambda}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{u^2}{2}}\, du$$

In other words, the Poisson distribution approaches the normal distribution as $\lambda \to \infty$.

This also means that the random variable $\frac{(X-\lambda)}{\sqrt{\lambda}}$ is *asymptotically normal*.

# Central Limit Theorem

- This theorem basically shows that there are other distributions besides the Binomial and Poisson that have the normal distribution as the limiting case.

**Theorem 4-2 (Central Limit Theorem)**

Let $X_1, X_2, \dots, X_n$ be independent random variables that are identically distributed (i.e., all have the same probability function in the discrete case or density function in the continuous case) and each of them have finite mean $\mu$ and variance $\sigma^2$.

Then if $S_n = X_1 + X_2 + \cdots + X_n$ where *n = 1, 2, …*

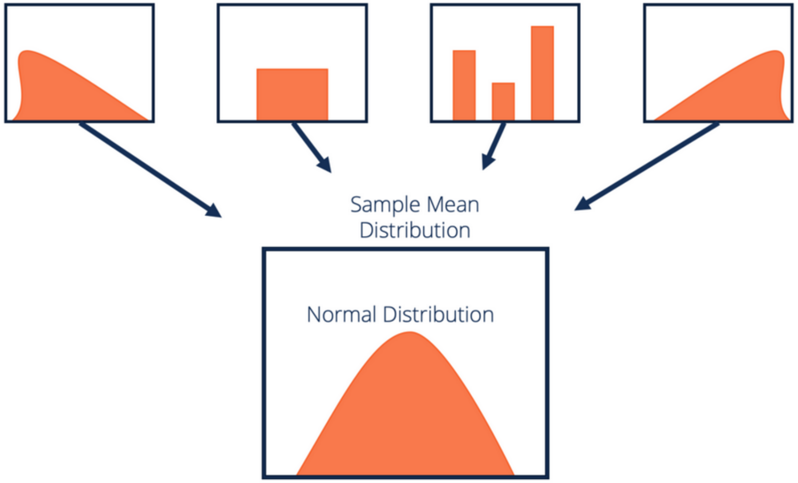
$$\lim_{n\to\infty} P\left(a \leq \frac{(S_n - n\mu)}{\sigma\sqrt{n}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{u^2}{2}}\, du$$

This also means that the random variable $\frac{(S_n - n\mu)}{\sigma\sqrt{n}}$, which is the standardized variable corresponding to $S_n$, is *asymptotically normal*.

# Central Limit Theorem

**Example**:

- An investor is interested in estimating the return of ABC stock market index that is comprised of 100,000 stocks. Large size → can't analyze each stock independently → use random sampling to get an estimate of the overall return of the index.
- The investor picks random samples of the stocks, with each sample comprising at least 30 stocks. The samples must be random, and any previously selected samples must be replaced in subsequent samples to avoid bias.
- The distribution of the sample means will move toward normal as the value of *n* increases. The average return of the stocks in the sample index estimates the return of the whole index of 100,000 stocks, and the average return is normally distributed.






Source: https://corporatefinanceinstitute.com/resources/knowledge/other/central-limit-theorem/

# Central Limit Theorem

The central limit theorem has many applications in different fields:

- Political/election polls: estimate the percentage of people who support a particular candidate.
- Confidence interval : used to calculate the mean family income for a particular region.
- Machine learning.
- Communication.
- …

# Some common and useful distributions

**Uniform Distribution**

A random variable X is said to be uniformly distributed in $a \leq x \leq b$ if its density function is:

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

The distribution is called a uniform distribution.

The distribution function is given by:

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

The mean and variance are, respectively,

$$\mu = \frac{1}{2}(a+b) \qquad \sigma^2 = \frac{1}{12}(b-a)^2$$

# Some common and useful distributions

**Exponential distribution**

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

The mean and variance are, respectively,

$$\mu = \frac{1}{\alpha} \qquad \sigma^2 = \frac{1}{\alpha^2} \qquad M(t) = \frac{\alpha}{\alpha - t}$$

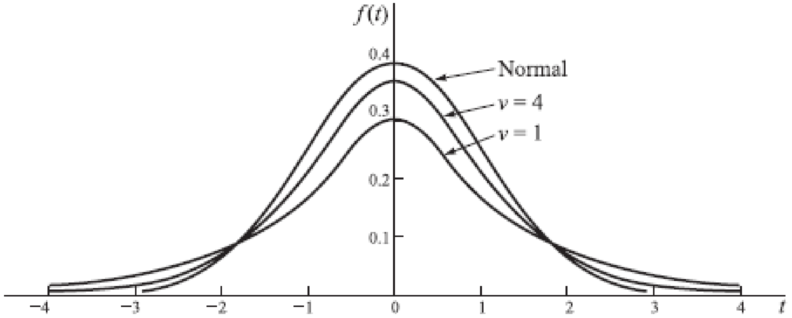
# Some common and useful distributions

**Student's t distribution**

If a random variable has the density function:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2} \qquad -\infty < t < \infty$$

It is said to have Student's t distribution, with *v* degree of freedom. If *v* is large , the graph of *f(t)* closely approximates the standard normal curve.




The mean and variance are, respectively,

$$\mu = 0 \qquad \text{and} \qquad \sigma^2 = \frac{\nu}{\nu-2} \qquad (\nu > 2).$$

# Some common and useful distributions

- Cauchy distribution
- Gamma Distribution
- Beta Distribution
- Chi-square distribution
- …

# Reference

Notes, equations, and figures in the lecture are based on or taken from materials in the course textbook:

"Probability and Statistics", by Spiegel, Schiller and Srinivasan, ISBN 987-007-179557-9 (McGraw-Hill/Schaun's)