Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Pr of. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr
http://avesis.yildiz.edu.tr/zeyneb/

Random Variables and Probability Distributions (cont'd)

Binomial Distribution

- A sequence of binary random variables X_1, $X_2, \ldots, X_n$ is called Bernoulli trials
  - if they all have the same Bernoulli distribution and are independent.
- The random variable representing the number of times the outcome of interest occurs in n Bernoulli trials (i.e., the sum of Bernoulli trials) has a Binomial(n, θ) distribution,
  - where θ is the probability of the outcome of interest (a.k.a. the probability of success).

Binomial Distribution

- A binomial distribution is defined by the number of Bernoulli trials n and the probability of the outcome of interest θ for the underlying Bernoulli trials.
- The pmf of a Binomial(n,θ) specifies the probability of each possible value (integers from 0 through n) of the random variable.
- The theoretical (population) mean of a random variable Y with Binomial(n,θ) distribution is $\mu = n\theta$.
- The theoretical (population) variance of Y is $\sigma^2 = n\theta(1-\theta)$.

Binomial Distribution

- A binomial experiment is one that has the following properties:
  - The experiment consists of n identical trials.
  - Each trial results in one of two outcomes.
    - We will label one outcome a success and the other a failure.
  - The probability of success on a single trial is equal to p, and p remains the same from trial to trial.
  - The trials are independent;
    - that is, the outcome of one trial does not influence the outcome of any other trial.
  - The random variable y is the number of successes observed during the n trials.

Binomial Distribution -Example

- A large power utility company uses gas turbines to generate electricity.

  The engineers employed at the company monitor the reliability of each turbine

  – that is, the probability that the turbine will perform properly under standard operating conditions over a specified period of time.

  The engineers wanted to estimate the probability a turbine will operate successfully for 30 days after being put into service.

  The engineers randomly selected 75 of the 100 turbines currently in use and examined the maintenance records.

  They recorded the number of turbines that did not need repairs during the 30-day time period.
- Is this a binomial experiment?

Binomial Distribution -Example

- For solution, we check this experiment against the five characteristics of a binomial experiment.
  - Are there identical trials?
    - The 75 trials could be assumed identical only if the 100 turbines are the same type of turbine, are the same age, and are operated under the same conditions.
  - Does each trial result in one of two outcomes?
    - Yes. Each turbine either does or does not need repairs in the 30-day time period.
  - Is the probability of success the same from trial to trial?
    - No. If we let success denote a turbine "did not need repairs," then the probability of success can change considerably from trial to trial.
      - For example, suppose that 15 of the 100 turbines needed repairs during the 30-day inspection period.
        - Then p, the probability of success for the first turbine examined, would be 85/100=0.85.
        - If the first trial is a failure (turbine needed repairs), the probability that the second turbine examined did not need repairs is 85/99=0.859.
      - Suppose that after 60 turbines have been examined, 50 did not need repairs and 10 needed repairs.
        - The probability of success of the next (61st) turbine would be 35/40=0.875.

Binomial Distribution -Example

- Were the trials independent?
  - Yes, provided that the failure of one turbine does not affect the performance of any other turbine.
    - However, the trials may be dependent in certain situations. For example,
      - suppose that a major storm occurs that results in several turbines being damaged.
      - Then the common event, a storm, may result in a common result, the simultaneous failure of several turbines.
- Was the random variable of interest to the engineers the number of successes in the 75 trials?
  - Yes. The number of turbines not needing repairs during the 30-day period was the random variable of interest.

- This example shows how the probability of success can change substantially from trial to trial in situations in which the sample size is a relatively large portion of the total population size.
- This experiment does not satisfy the properties of a binomial experiment.

Binomial Distribution

- Although it is possible to approximate $P(y)$, the probability associated with a value of y in a binomial experiment, by using a relative frequency approach, it is easier to use a general formula for binomial probabilities.
- The probability of observing y successes in n trials of a binomial experiment is

$$P(y) = \frac{n!}{y!\,(n-y)!}\theta^{y}(1-\theta)^{n-y}$$

where

n	= number of trials
θ	= probability of success on a single trial
$1 - \theta$	= probability of failure on a single trial
y	= number of successes in n trials
$n!$	= $n(n\text{-}1)(n\text{-}2) \ldots (3)(2)(1)$

Binomial Distribution - Example

- A new variety of turf grass has been developed for use on golf courses, with the goal of obtaining a germination rate of 85%.
- To evaluate the grass, 20 seeds are planted in a greenhouse so that each seed will be exposed to identical conditions.
- If the 85% germination rate is correct,
  - what is the probability that 18 or more of the 20 seeds will germinate?
  - what is the average number of seeds that will germinate in the sample of 20 seeds ?
  - what is the variance of seeds that will germinate in the sample of 20 seeds ?
  - what is the standard deviation of seeds that will germinate in the sample of 20 seeds ?

Binomial Distribution - Example

- **Solution:**
- $P(y) = \frac{n!}{y!(n-y)!}\theta^{y}(1-\theta)^{n-y}$
- $n = 20, \quad \theta = 0.85, \quad y = 18, 19, \text{ and } 20$
- $P(y = 18) = \frac{20!}{18!(20-18)!}(0.85)^{18}(1-0.85)^{20-18} = 0.229$
- $P(y = 19) = \frac{20!}{19!(20-19)!}(0.85)^{19}(1-0.85)^{20-19} = 0.137$
- $P(y = 20) = \frac{20!}{20!(20-20)!}(0.85)^{20}(1-0.85)^{20-20} = 0.038$
- $P(y \geq 18) = P(y = 18) + P(y = 19) + P(y = 20) = 0.405$
- The following commands in R will compute the binomial probabilities:
  - To calculate $P(X = 18)$, use the command dbinom(18, 20, 0.85)
  - To calculate $P(X \leq 17)$, use the command pbinom (17, 20, 0.85)
  - To calculate $P(X \geq 18)$, use the command 1 - pbinom(17, 20, 0.85)

Binomial Distribution - Example

- The average number of seeds that will germinate in the sample of 20 seeds is

$$\mu = n\theta = 20 \times 0.85 = 17$$

- The variance of seeds that will germinate in the sample of 20 seeds is

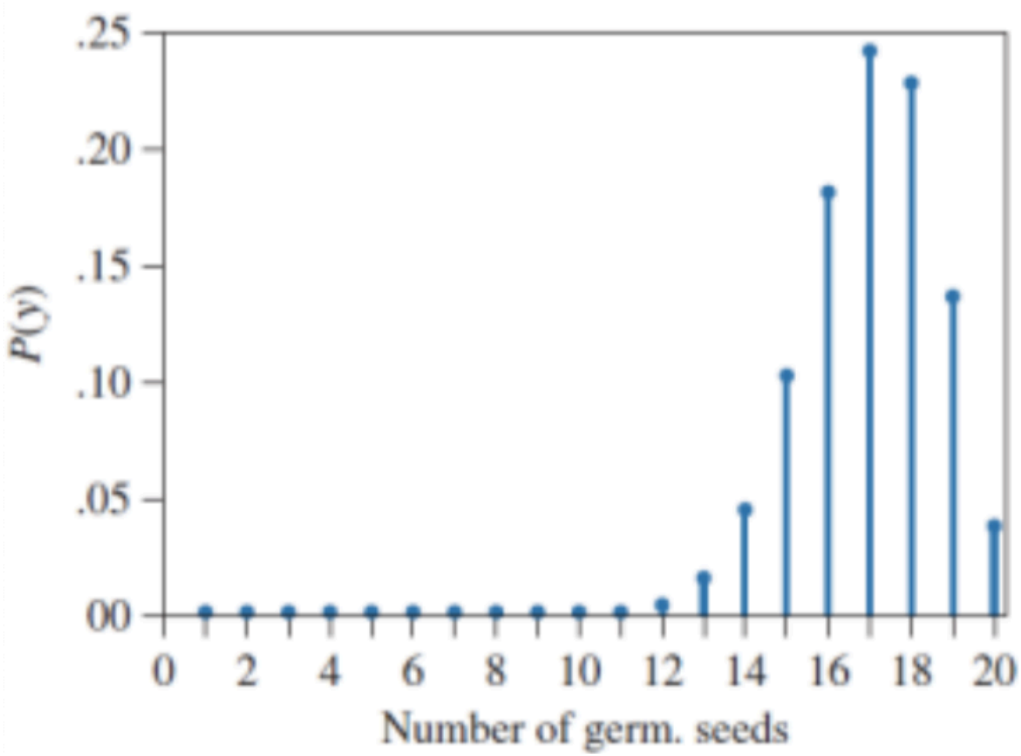
$$\sigma^2 = n\theta(1-\theta) = 20 \times 0.85\ (1 - 0.85) = 2.55$$

- The standard deviation of seeds that will germinate in the sample of 20 seeds is

$$\sigma = \sqrt{n\theta(1-\theta)} = \sqrt{\sigma^2} = \sqrt{\sigma^2} = 1.60$$

Binomial Distribution - Example

- Suppose we examine the germination records of a large number of samples of 20 seeds each.
- If the germination rate has remained constant at 85%, then the average number of seeds that germinate should be close to 17 per sample.



If in a particular sample of 20 seeds we determine that only 12 had germinated, would the germination rate of 85% seem consistent with our results?

- Using a computer software program, we can generate the probability distribution for the number of seeds that germinate in the sample of 20 seeds, as shown in the Figure

Binomial Distribution - Example

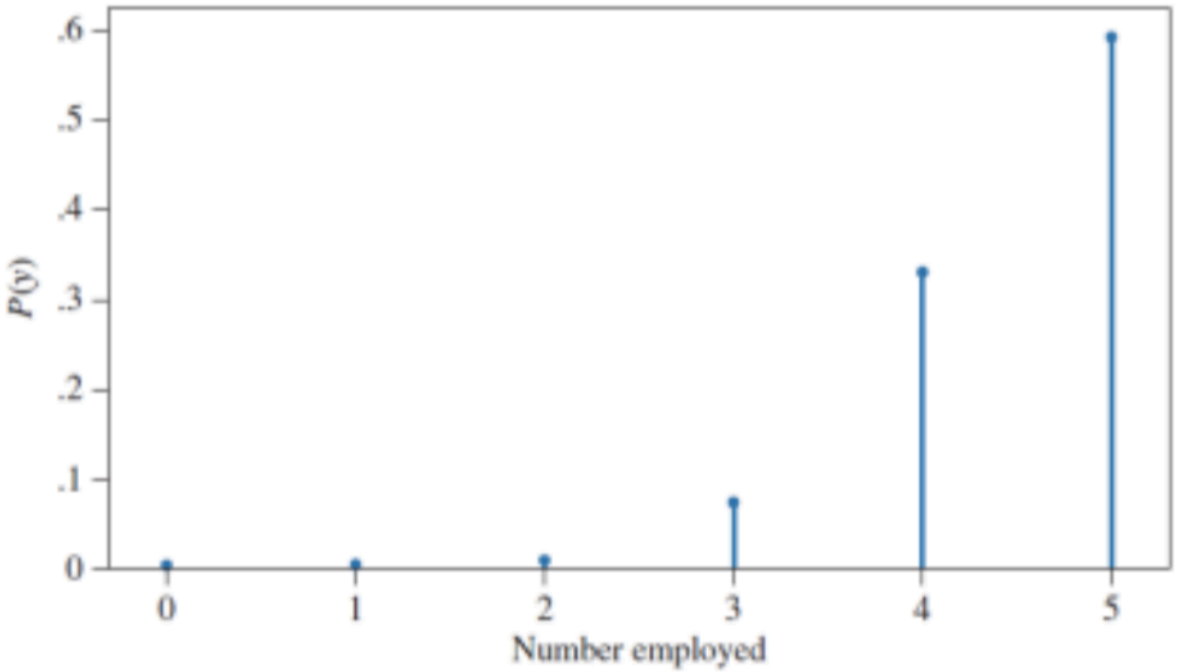
- Suppose that a sample of households is randomly selected from all the households in the city in order to estimate the percentage in which the head of the household is unemployed.
- To illustrate the computation of a binomial probability, suppose that the unknown percentage is actually 10% and that a sample of $n = 5$ (we select a small sample to make the calculation manageable) is selected from the population.
- What is the probability that all five heads of households are employed?

Binomial Distribution - Example

- **Solution:**
- We must carefully define which outcome we wish to call a success.
  - For this example, we define a success as being employed.
- Then the probability of success when one person is selected from the population is $\theta = 0.9$ (because the proportion unemployed is 0.1).
- We wish to find the probability that $y = 5$ (all five are employed) in five trials.
- $P(y) = \frac{n!}{y!(n-y)!}\theta^{y}(1-\theta)^{n-y}$
- $P(y = 5) = \frac{5!}{5!(5-5)!}(0.9)^{5}(1-0.9)^{5-5} = 0.59$

Binomial Distribution - Example

- The binomial probability distribution for $n = 5$, $\theta = 0.9$ is shown in the figure.



– Here, the probability of observing five employed in a sample of five is shown to be 0.59.

Poisson Distribution

- In 1837, S. D. Poisson developed a discrete probability distribution, suitably called the Poisson distribution, which has as one of its important applications the modeling of events of a particular time over a unit of time or space
- For example, the number of automobiles arriving at a toll booth during a given 5-minute period of time.
  - The event of interest would be an arriving automobile, and the unit of time would be 5 minutes.

Poisson Distribution

- A second example would be the situation in which an environmentalist measures the number of PCB particles discovered in a liter of water sampled from a stream contaminated by an electronics production plant.
  - The event would be a PCB particle discovered.
  - The unit of space would be 1 liter of sampled water.

Poisson Distribution

- Let y be the number of events occurring during a fixed time interval of length t or a fixed region R of area or volume $m(R)$.
- Then the probability distribution of y is Poisson, provided certain conditions are satisfied:
  - Events occur one at a time; two or more events do not occur precisely at the same time or in the same space.
  - The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a nonoverlapping time period or region of space;
    - that is, the occurrence (or nonoccurrence) of an event during one period or in one region does not affect the probability of an event occurring at some other time or in some other region.
  - The expected number of events during one period or in one region, μ, is the same as the expected number of events in any other period or region.

Poisson Distribution

- Assuming that the above conditions hold, the Poisson probability of observing y events in a unit of time or space is given by the formula

$$P(y) = \frac{\mu^y e^{-\mu}}{y!}$$

where e is a naturally occurring constant approximately equal to 2.71828 and μ is the average value of y.

Poisson Distribution - Example

- A large industrial plant is being planned in a rural area. As a part of the environmental impact statement, a team of wildlife scientists is surveying the number and types of small mammals in the region.
- Let y denote the number of field mice captured in a trap over a 24-hour period.
- Suppose that y has a Poisson distribution with $\mu = 2.3$; that is, the average number of field mice captured per trap is 2.3.
  - What is the probability of finding exactly four field mice in a randomly selected trap?
  - What is the probability of finding at most four field mice in a randomly selected trap?
  - What is the probability of finding more than four field mice in a randomly selected trap?

Poisson Distribution - Example

- The probability that a trap contains exactly four field mice is computed to be

$$P(y=4)=\frac{\mu^{y}e^{-\mu}}{y!}=\frac{(2.3)^{4}e^{-2.3}}{4!}=\frac{(27.9841)(0.10002588)}{24}=0.1169$$

- The probability of finding at most four field mice in a randomly selected trap is,

$P(y \le 4) = P(y=0) + P(y=1) + P(y=2) + P(y=3) + P(y=4)$

$P(y \le 4) = 0.1003 + 0.2306 + 0.2652 + 0.2033 + 0.1169 = 0.9163$

- The probability of finding more than four field mice in a randomly selected trap, using the idea of complementary events, is

$P(y > 4) = 1 - P(y \le 4) = 1 - 0.9163 = 0.0837$

Thus, it is a very unlikely event to find five or more field mice in a trap.

Poisson Distribution - Example

- The Poisson probabilities can be computed using the following R commands.

$P(y = 4) = \text{dpois}(4, 2.3) = 0.1169022$

$P(y \leq 3) = \text{ppois}(3, 2.3) = 0.7993471$

$P(y > 4) = 1 - P(y \leq 4) = 1 - \text{ppois}(4, 2.3) = 0.08375072$

```
> dpois(4, 2.3)
[1] 0.1169022
> ppois(3, 2.3)
[1] 0.7993471
> 1 - ppois(4, 2.3)
[1] 0.08375072
```

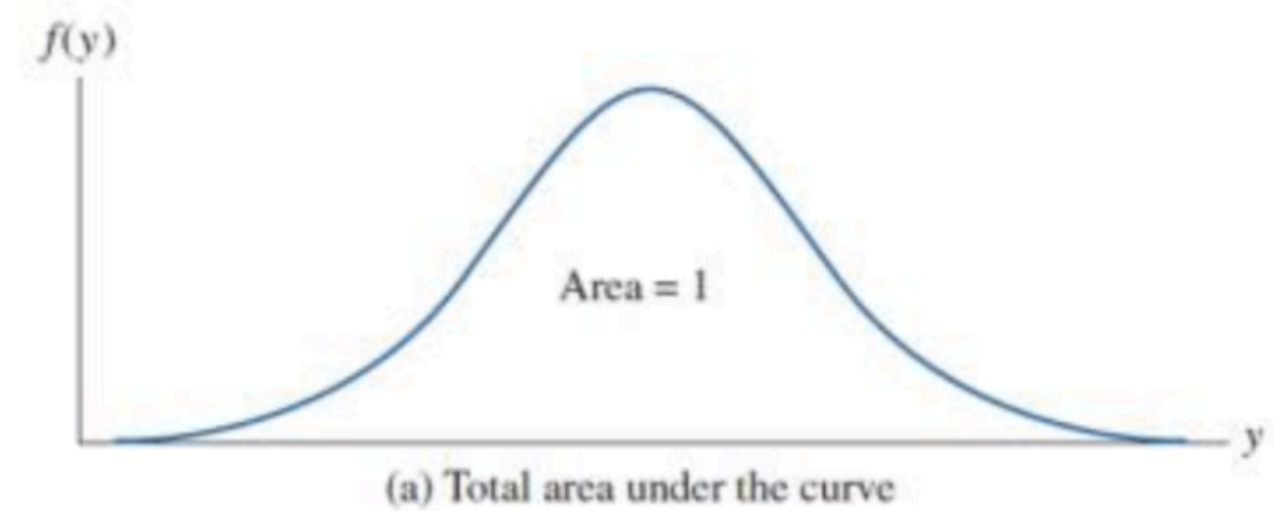
- When n is large and θ is small in a binomial experiment, $n \geq 100$, $\theta \leq 0.01$, and $n\theta \leq 20$,
  - the Poisson distribution provides a reasonable approximation to the binomial distribution.
- In applying the Poisson approximation to the binomial distribution, use $\mu = n\theta$.

Continuous probability distributions

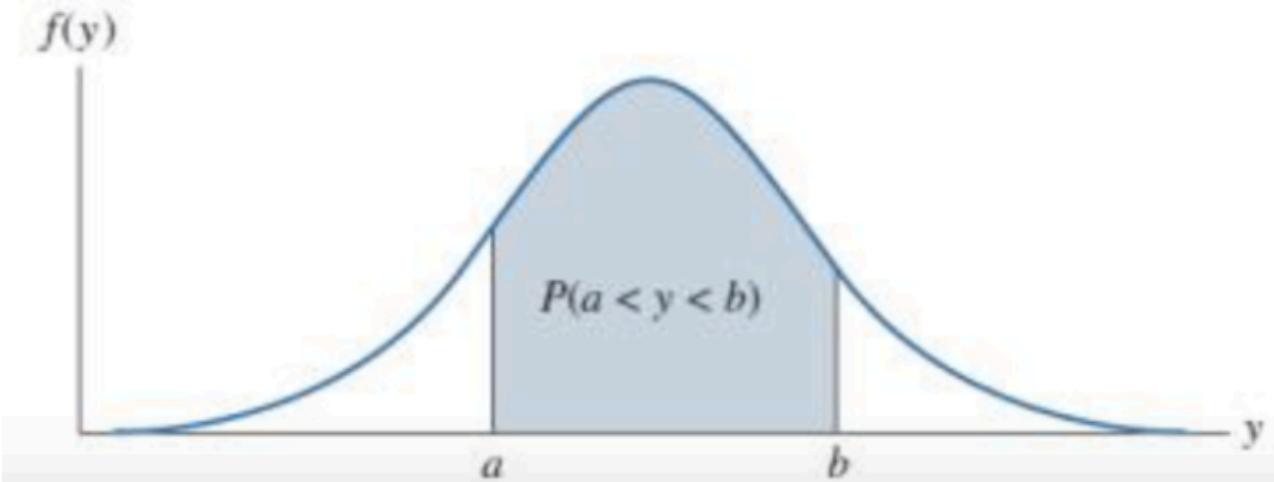
- For discrete random variables, the pmf provides the probability of each possible value.
- For continuous random variables, the number of possible values is uncountable, and the probability of any specific value is zero.
- For these variables, we are interested in the probability that the value of the random variable is within a specific interval from x_1 to x_2;
  - we show this probability as $P(x_1 < X \leq x_2)$.

Continuous probability distributions

- Probability distribution for a continuous random variable

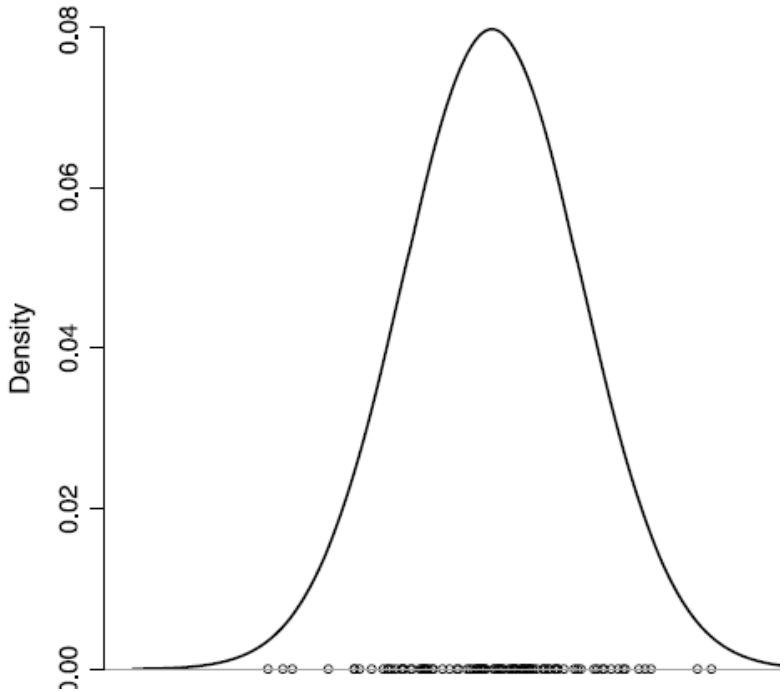


(a) Total area under the curve



Continuous probability distributions

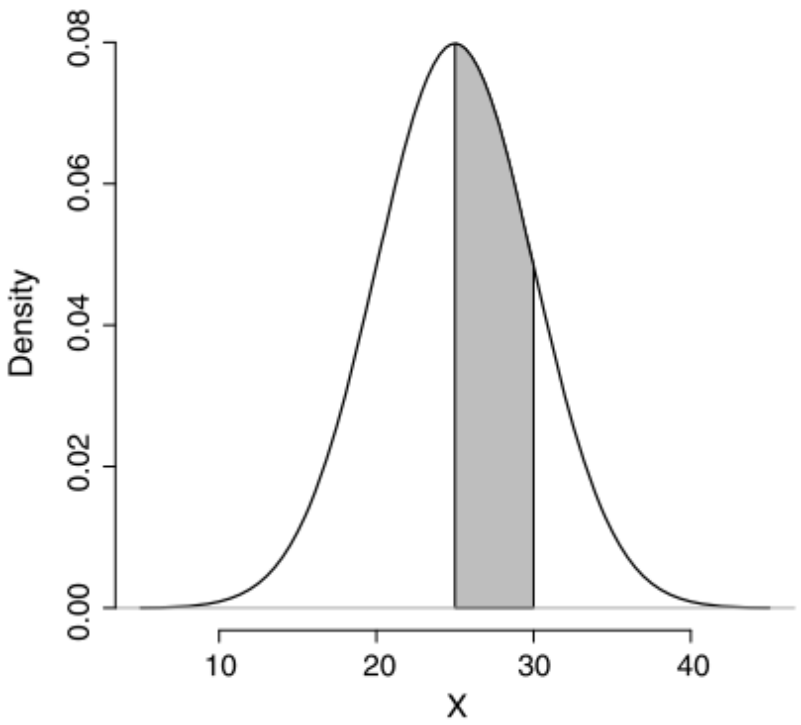
- For continuous random variables, we use probability density functions (pdf) to specify the distribution.
- Using the pdf, we can obtain the probability of any interval.



- The assumed probability distribution for BMI (Body Mass Index), which is denoted as X, along with random sample of 100 values, which are shown as circles along the horizontal axis

Continuous probability distributions

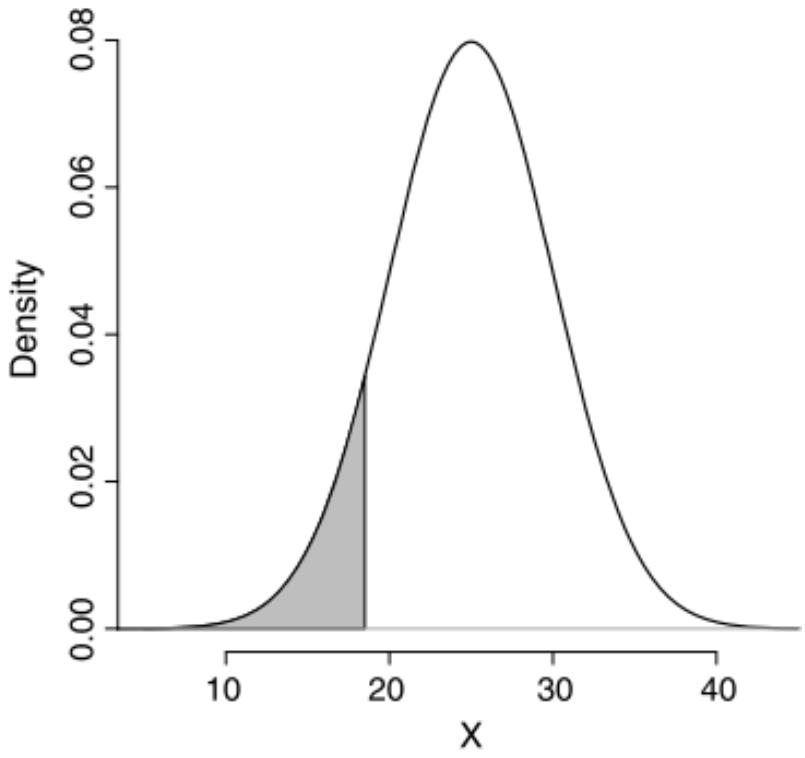
- The total area under the probability density curve is 1.
- The curve (and its corresponding function) gives the probability of the random variable falling within an interval.



- This probability is equal to the area under the probability density curve over the interval.
- The shaded area is the probability that a person's BMI is between 25 and 30.
- People whose BMI is in this range are considered as overweight.
- Therefore, the shaded area gives the probability of being overweight

Lower tail probability

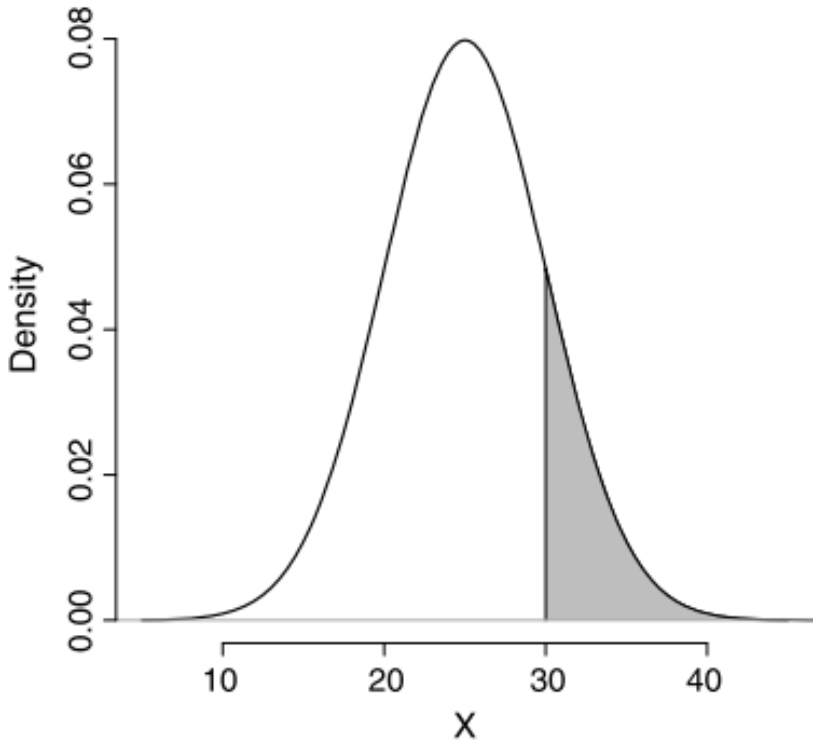
- The probability of observing values less than or equal to a specific value x, is called the lower tail probability and is denoted as $P(X \leq x)$



- This probability is found by measuring the area under the curve to the left of x.
- For example, the shaded area in the left panel of the figure is the lower tail probability of having a BMI less than or equal to 18.5 (i.e., being underweight), $P(X \leq 18.5)$.

Upper tail probability

- The probability of observing values greater than x, is called the upper tail probability and is denoted as $P(X > x)$



- This probability is found by measuring the area under the curve to the right of x.
- For example, the shaded area in the right panel of the figure is the upper tail probability of having a BMI greater than 30 (i.e., being obese), $P(X > 30)$.

Probability of intervals

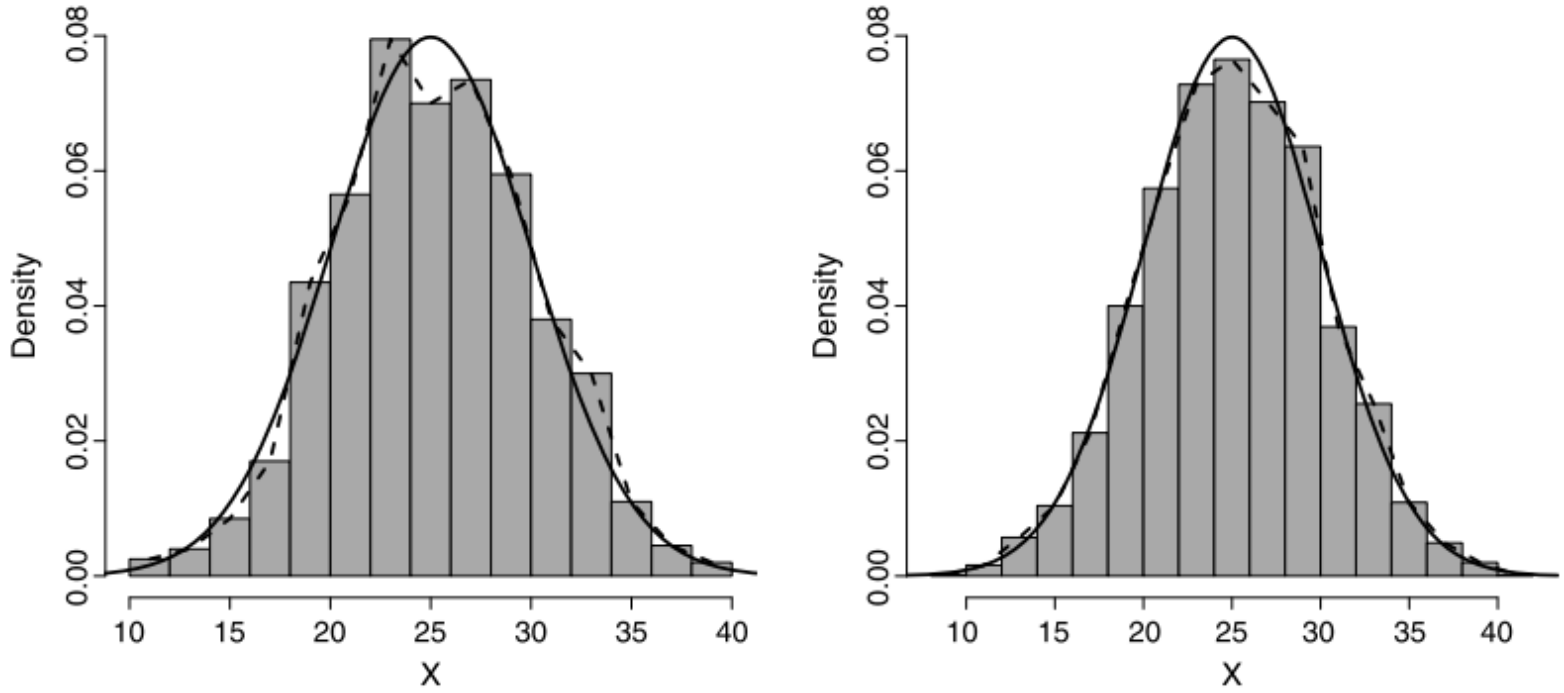
- The probability of any interval from x_1 to x_2, where $x_1 < x_2$, can be obtained using the corresponding lower tail probabilities for these two points as follows:

  $P(x1 < X \leq x2) = P(X \leq x2) - P(X \leq x1).$

- For example, suppose that we wanted to know the probability of a BMI between 25 and 30.
- This probability $P(25 < X \leq 30)$ is obtained by subtracting the lower tail probability of 25 from the lower tail probability of 30:

  $P(25 < X \leq 30) = P(X \leq 30) - P(X \leq 25).$

Probability Density Curves and Density Histograms



- *Left panel*: Histogram of BMI for 1000 observations.
  - The *dashed line* connects the height of each bar at the midpoint of the corresponding interval
  - The *smooth solid curve* is the density curve for the probability distribution of BMI
- *Right panel*: Histogram of BMI for 5000 observations.
  - The histogram and its corresponding *dashed line* provide better approximations to the density curve

    - Recall that the height of each bar is the density for the corresponding interval, and the area of each bar is the relative frequency for that interval.
    - The density histogram and the dashed line, which shows the density for each interval based on the observed data, provide reasonable approximations to the density curve.
    - Also, the area of each bar, which is equal to the relative frequency for the corresponding interval, is approximately equal to the area under the curve over that interval.

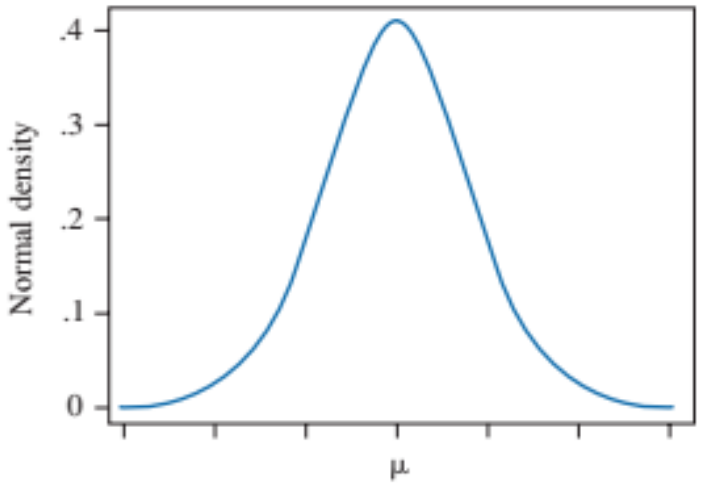
Normal distribution

- A normal distribution and its corresponding pdf are fully specified by the mean μ and variance σ^2.
- A random variable X with normal distribution is denoted $X \sim N(\mu,\sigma^2)$,
  - where μ is a real number, but σ^2 can take positive values only.
- The normal density curve is always symmetric about its mean μ, and its spread is determined by the variance σ^2.
- A normal distribution with a mean of 0 and a standard deviation (or variance) of 1 is called the standard normal distribution and denoted $N(0,1)$.

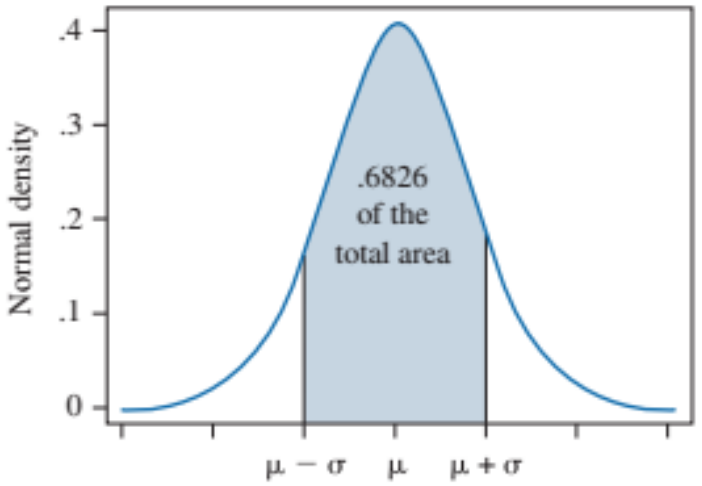
The 68-95-99.7% rule

- The 68–95–99.7% rule for normal distributions specifies that
  - 68% of values fall within 1 standard deviation of the mean:
    $P(\mu - \sigma < X \leq \mu + \sigma) = 0.68$
  - 95% of values fall within 2 standard deviations of the mean:
    $P(\mu - 2\sigma < X \leq \mu + 2\sigma) = 0.95$
  - 99.7% of values fall within 3 standard deviations of the mean:
    $P(\mu - 3\sigma < X \leq \mu + 3\sigma) = 0.997$

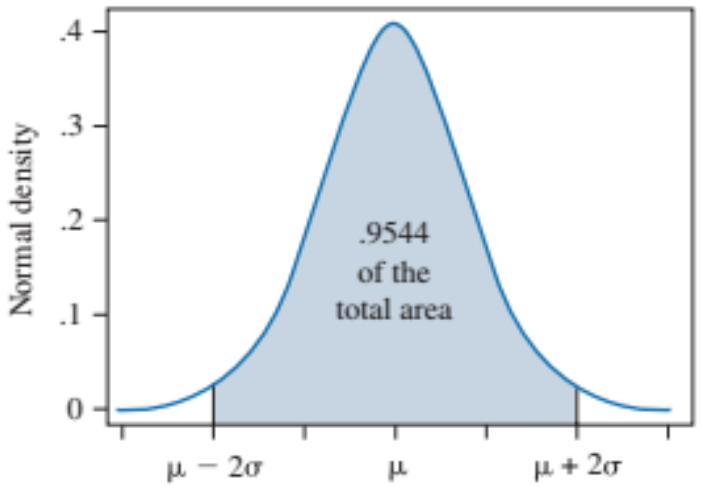
The 68-95-99.7% rule



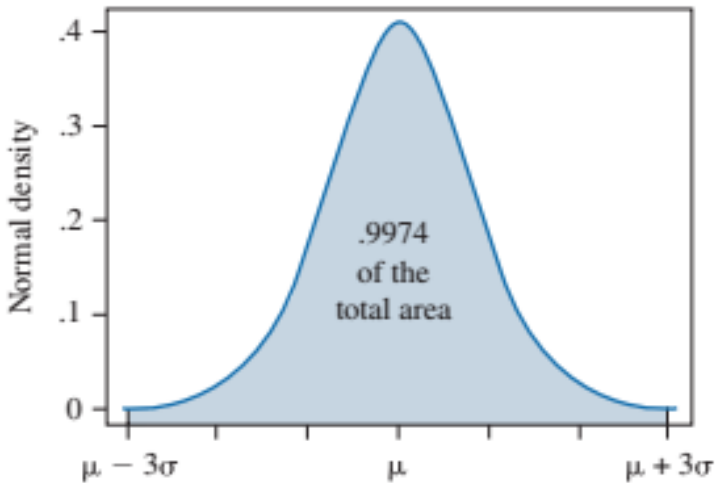
(a) Normal curve



(b) Area under normal curve within 1 standard deviation of mean



(c) Area under normal curve within 2 standard deviations of mean



(d) Area under normal curve within 3 standard deviations of mean

Example

- For example, suppose we know that the population mean and standard deviation for Systolic blood pressure (SBP) are $\mu = 125$ and $\sigma = 15$, respectively.
  - That is, $X \sim N(125, 15^2)$,
    - where X is the random variable representing SBP.
- Therefore, the probability of observing an SBP in the range $\mu \pm \sigma$ is 0.68:
  $P(125 - 15 < X \leq 125 + 15) = P(110 < X \leq 140) = 0.68.$
- This probability corresponds to the central area shown in the Fig. b in the previous slide.

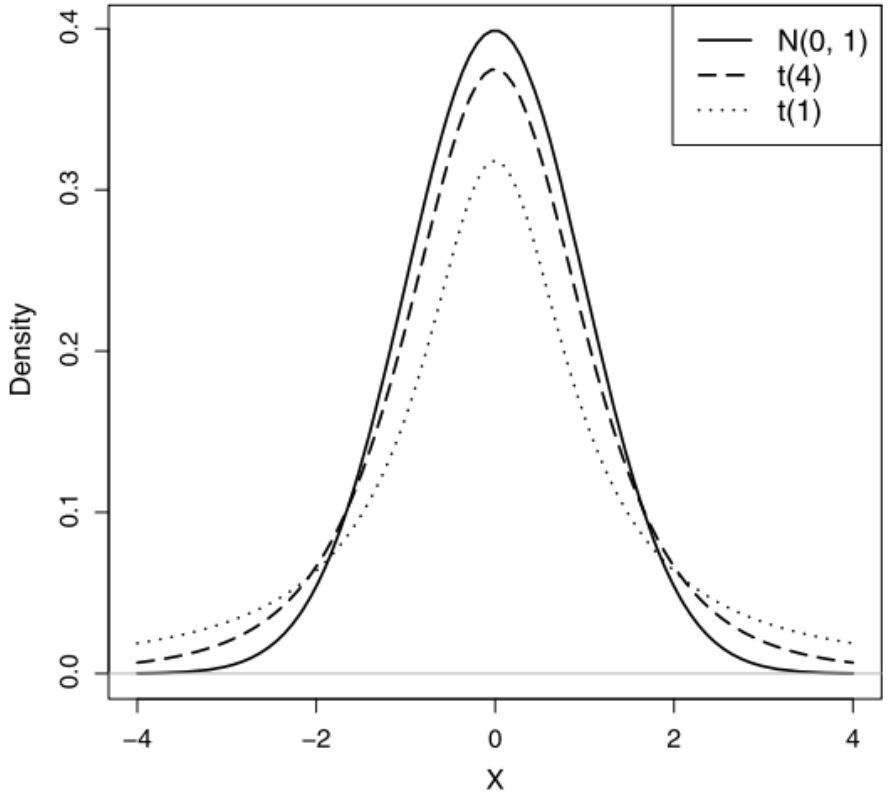
Example

- The probability of observing an SBP in the range $\mu \pm 2\sigma$ is 0.95:
  $P(125 - 2 \times 15 < X \leq 125 + 2 \times 15) = P(95 < X \leq 145) = 0.95.$
- This probability is shown in the Fig. c in the previous slide.
- Lastly, the probability of observing an SBP is in the range $\mu \pm 3\sigma$ is 0.997:
  $P(125 - 3 \times 15 < X \leq 125 + 3 \times 15) = P(80 < X \leq 170) = 0.997.$

- Therefore, we rarely (probability of 0.003) expect to see SBP values less than 80 or greater than 170.

Student's t-distribution

- Another continuous probability distribution that is used very often in statistics is the Student's *t*-distribution or simply the *t*-distribution.



- Comparing the pdf of a standard normal distribution to *t*-distributions with 1 degree of freedom and then with 4 degrees of freedom.
- The *t*-distribution has heavier tails than the standard normal;
  - however, as the degrees of freedom increase, the *t*-distribution approaches the standard normal

Student's t-distribution

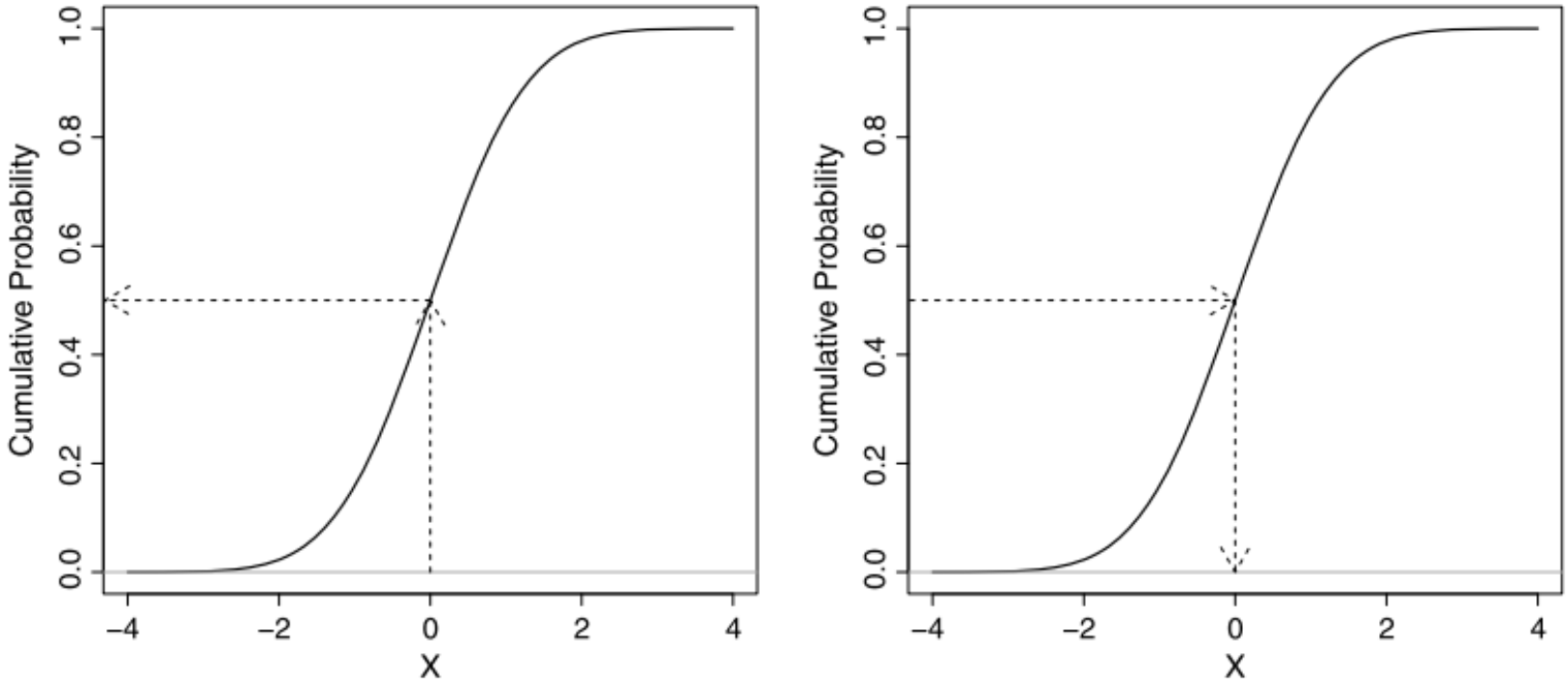
- A *t*-distribution is specified by only one parameter called the degrees of freedom, *df*.
- The *t*-distribution with *df* degrees of freedom is usually denoted as $t(df)$ or tdf, where *df* is a positive real number $(df > 0)$.
- The mean of this distribution is $\mu = 0$,
- The variance is determined by the degrees of freedom parameter, $\sigma^2 = df/(df - 2)$,
  - which is of course defined when $df > 2$.

Cumulative distribution function

- We saw that by using lower tail probabilities, we can find the probability of any given interval.
- Indeed, all we need to find the probabilities of any interval is a function that returns the lower tail probability at any given value of the random variable: $P(X \leq x)$.
- This function is called the cumulative distribution function (cdf) or simply the distribution function.

Quantiles

- We can use the cdf plot in the reverse direction to find the value of the random variable for a given lower tail probability.



Quantiles

- In previous slide:
  - **Left panel**:
  - Plot of the cdf for the standard normal distribution, $N(0,1)$.
    - The cdf plot of the cdf can be used to find the lower tail probability.
      - For instance, following the *arrow* from $x = 0$ (on the horizontal axis) to the cumulative probability (on the vertical axis) gives us the probability $P(X \leq 0) = 0.5$.
  - **Right panel**:
  - Given the lower tail probability of 0.5 on the vertical axis, we obtain the corresponding quantile $x = 0$ on the horizontal axis

Scaling and shifting random variables

- If $Y = aX + b$, then

$$\mu_Y = a\mu_X + b$$
$$\sigma_Y^2 = a^2\sigma_X^2$$
$$\sigma_Y = |a|\sigma_X$$

- The process of shifting and scaling a random variable to create a new random variable with mean zero and variance one is called standardization.
  - For this, we first subtract the mean μ and then divide the result by the standard deviation σ.

$$Z = (X - \mu)/\sigma$$

    - If $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0, 1)$.

Adding/subtracting random variables

- If $W = X + Y$, then

$$\mu_W = \mu_X + \mu_Y$$

- If the random variables X and Y are independent, then we can find the variance of W as follows:

$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2$$

  - If $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$, then assuming that the two random variables are independent, we have

$$W = X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Adding/subtracting random variables

- If we subtract Y from X, then

$$\mu_W = \mu_X - \mu_Y$$

- If the random variables X and Y are independent, then we can find the variance of W as follows:

$$\sigma_W^2 = \sigma_X^2 + \sigma_Y^2$$

  - If $X \sim N(\mu_X, \sigma_X^2)$, and $Y \sim N(\mu_Y, \sigma_Y^2)$, then assuming that the two random variables are independent, we have

$$W = X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Estimation

Parameter Estimation

- The objective of statistics is to make inferences about a population based on information contained in a sample.
- Populations are characterized by numerical descriptive measures called parameters.
- Typical population parameters are the mean μ, the median M, the standard deviation σ, and a proportion π.
- Most inferential problems can be formulated as an inference about one or more parameters of a population.

Parameter Estimation

- Methods for making inferences about parameters fall into one of two categories:
  - estimate the value of the population parameter of interest
  - test a hypothesis about the value of the parameter
- These two methods of statistical inference involve different procedures, and they answer two different questions about the parameter.
  - In estimating a population parameter, we are answering the question
    - "What is the value of the population parameter?"
  - In testing a hypothesis, we are seeking an answer to the question
    - "Does the population parameter satisfy a specified condition? "

Parameter Estimation

- We discussed
  - using random variables to represent characteristics of a population
    - (e.g., BMI, disease status).
  - some commonly used probability distributions for discrete and continuous random variables.
- We are specifically interested in population mean and population variance of a random variable.
  - These quantities are unknown in general.
- We refer to these unknown quantities as parameters.
- Here, we use parameters μ and σ^2 to denote the unknown population mean and variance respectively.
  - Note that for all the distributions we discussed previously, the population mean and variance of a random variable are related to the unknown parameters of probability distribution assumed for that random variable.
    - Indeed, for normal distributions $N(\mu,\sigma^2)$, which are widely used in statistics, the population mean and variance are exactly the same parameters used to specify the distribution.

Parameter Estimation

- In this lecture, we discuss statistical methods for parameter estimation.
  - Estimation refers to the process of guessing the unknown value of a parameter (e.g., population mean) using the observed data.
- For this, we will use an estimator, which is a statistic.
  - A statistic is a function of the observed data only.
- Sometimes we only provide a single value as our estimate.
  - This is called point estimation.
    - Point estimates do not reflect our uncertainty when estimating a parameter.
    - We always remain uncertain regarding the true value of the parameter when we estimate it using a sample from the population.
- To address this issue, we can present our estimates in terms of a range of possible values.
  - This is called interval estimation.

Convention

- We use $X_1, X_2, \ldots, X_n$ to denote n possible values of X obtained from a sample randomly selected from the population.
- We treat $X_1, X_2, \ldots, X_n$ themselves as n random variables because their values can change depending on which n individuals we sample.
- We assume the samples are independent and identically distributed (IID).
- While theoretically we can have many different samples of size n, we usually have only one such sample in practice.
- We use $x_1, x_2, \ldots, x_n$ as the specific set of values we have observed in our sample.
- That is, x_1 is the observed value for X_1, x_2 is the observed value X_2, and so forth.

Point estimation - Population Mean

- Sometimes we only provide a single value as our estimate.
  - This is called point estimation.
    - We use $\hat{\mu}$ and $\hat{\sigma}^2$ to denote the point estimates for and μ and σ^2.
- For a population of size N, μ is calculated as.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Given n observed values, $X_1, X_2, \ldots, X_n$, from the population, we can estimate the population mean μ with the sample mean:

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

  - In this case, we say that $\bar{X}$ is an estimator for μ.
- As our sample (the n representative members from the population) changes, the value of this estimator (sample mean) can also change.

Point estimation - Population Mean

- We usually have only one sample of size n from the population, x_2, … , x_n .
- Therefore, we only have one value for $\bar{X}$, which we denote

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where x_i is the ith observed value of X in our sample, and $\bar{x}$ is the observed value of $\bar{X}$.

- As an example,
  - {consider the study* to estimate the population mean for body temperature among healthy people. From a sample of $n = 148$ people, they estimated the unknown population mean with the sample mean $\hat{\mu} = \bar{x} = 98.25$. This estimate is lower than the commonly believed value of 98.6°F.}
    - [The sample size for this study was relatively small. We would expect that as the sample size increases, our point estimate based on the sample mean would become closer to the true population mean.]

*Mackowiak, P.A., Wasserman, S.S., Levine, M.M.: A critical appraisal of 98.6°F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold AugustWunderlich. JAMA 268, 1578–1580 (1992)

Law of Large Numbers (LLN)

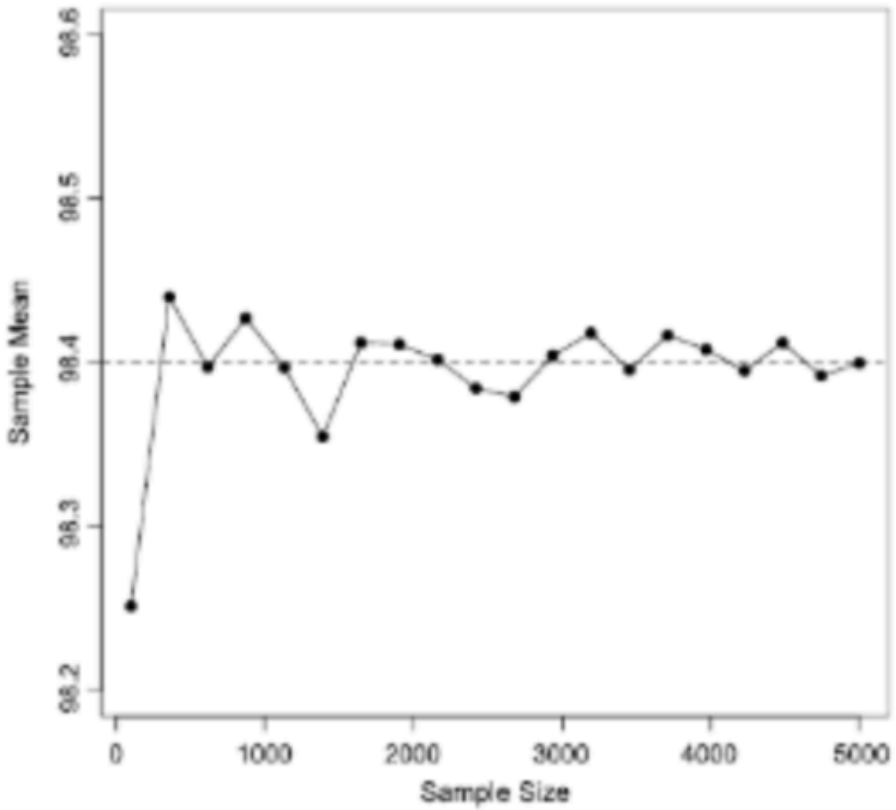
- The Law of Large Numbers (LLN) indicates that (under some general conditions such as independence of observations) the sample mean converges to the population mean ($\bar{X}_n \rightarrow \mu$) as the sample size n increases ($n \rightarrow \infty$).
- Informally, this means that the difference between the sample mean and the population mean tends to become smaller and smaller as we increase the sample size.
- The LLN provides a theoretical justification for the use of sample mean as an estimator for the population mean.

Law of Large Numbers (LLN)

- The Law of Large Numbers is true regardless of the underlying distribution of the random variable.
  - Therefore, it justifies using the sample mean $\bar{X}$ to estimate the population mean for continuous random variables, discrete random variables, whose values are counts (i.e., nonnegative integers), and for discrete binary variables, whose possible values are 0 and 1 only.
- For count variables, the mean is usually referred to as the rate (e.g., rate of traffic accidents).
- For binary random variables, the mean is usually referred to as the proportion of the outcome of interest (denoted as 1).
  - Hence, we sometimes use the notation p instead of $\bar{x}$ for the sample mean of binary random variables.

Law of Large Numbers (LLN)

- Suppose the true population mean for normal body temperature is 98.4°F.
- Here, the estimate of the population mean is plotted for different sample sizes.



- As the sample size is increased, the sample mean $\bar{X}$ converges to the population mean μ.
- For the temperature example, by increasing n, $\bar{X} \rightarrow \mu = 98.4$

Point estimation - Population Variance

- The population variance is the average of squared deviations of each observation x_i from the population mean μ and denoted as σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Given n randomly sampled values $X_1, X_2, \ldots, X_n$ from the population and their corresponding sample mean $\bar{X}$, we can estimate the variance as :

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

- However, this estimator tends to underestimate the population variance.

Point estimation - Population Variance

- To address this issue, a more commonly used estimator for σ^2 is the sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

  - This is the sum of squared deviations from the sample mean divided by $n-1$ instead of n.
    - Dividing by $n-1$ instead of n increases the value of the estimator by a small amount, which is enough to avoid underestimation associated with the more natural estimator.
- Therefore, the sample variance is the usual estimator of the population variance.
  - Likewise, the sample standard deviation S, $(\sqrt{S^2})$, is our estimator of the population standard deviation σ.
- We regard the estimator S^2 as a random variable since it changes as we change the sample.

Point estimation - Population Variance

- However, in practice, we usually have one set of observed values, $x_1, x_2, \ldots, x_n$, and therefore, only one value for S^2, denoted as s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- For binary random variables with 0 and 1 values, we can show that the population variance σ^2 is equal to $\mu(1-\mu)$, where μ is the population mean (proportion).
  - (See the Bernoulli distribution)
- Therefore, after we estimate the population mean μ using the sample mean (proportion) $\bar{x} = p$, we can use it to estimate the population variance instead of estimating σ^2 separately:

$$s^2 = p(1-p)$$