USC Department Of Translational Genomics
Keck Medicine of USC

TRGN 527: Applied Data Science and Bioinformatics

UNIT I. Introduction and Basic Data Science

Week 3 - Lecture 3

Enrique I. Velazquez Villarreal, M.D., Ph.D., M.P.H., M.S. | **Assistant Professor**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

Leader of the USC Bioinformatics Core – *USC CaRE2 Health Equity Center*

David W. Craig, Ph.D. | **Professor and Vice Chair**

Dept. of Translational Genomics

USC | Keck School of Medicine | Norris Comprehensive Cancer Center

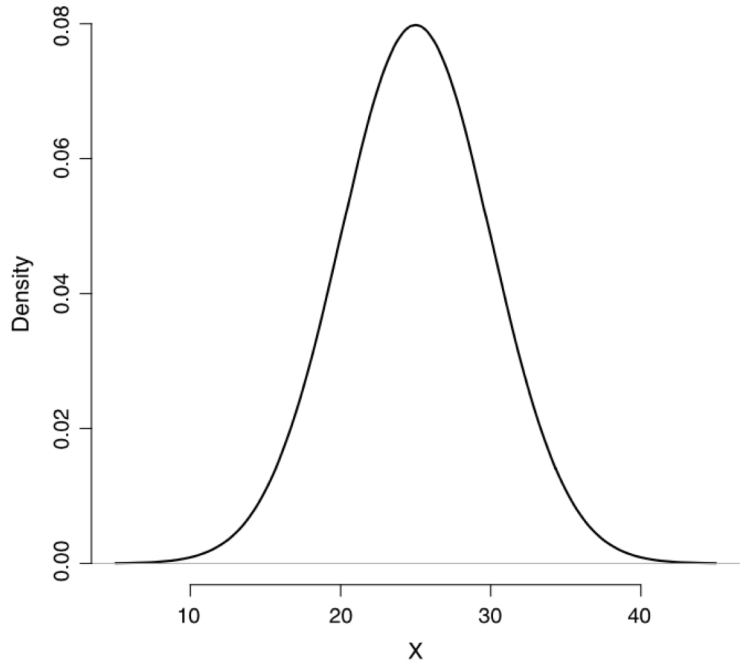
Co-Director, Institute of Translational Genomics

Topics

- Continuous Probability Distributions

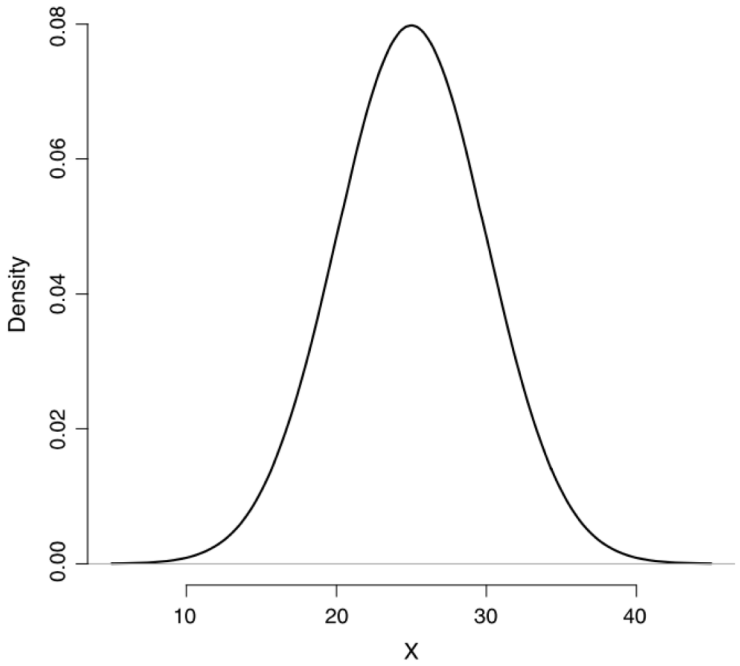
Continuous Probability Distributions

- 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.
- Instead of talking about the probability of any specific value x for continuous random variable X:
  - we talk about the probability that the value of the random variable is within a **specific interval** from x1 to x2; we show this probability as $P(x1 < X \leq x2)$

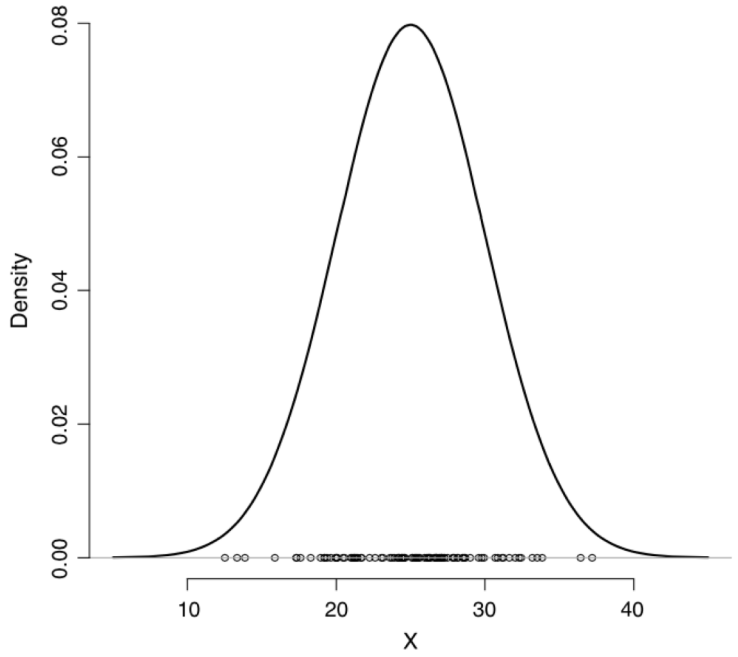


Continuous Probability Distributions

- 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.
  - As an example, consider the continuous random variable X representing the body mass index of the US population



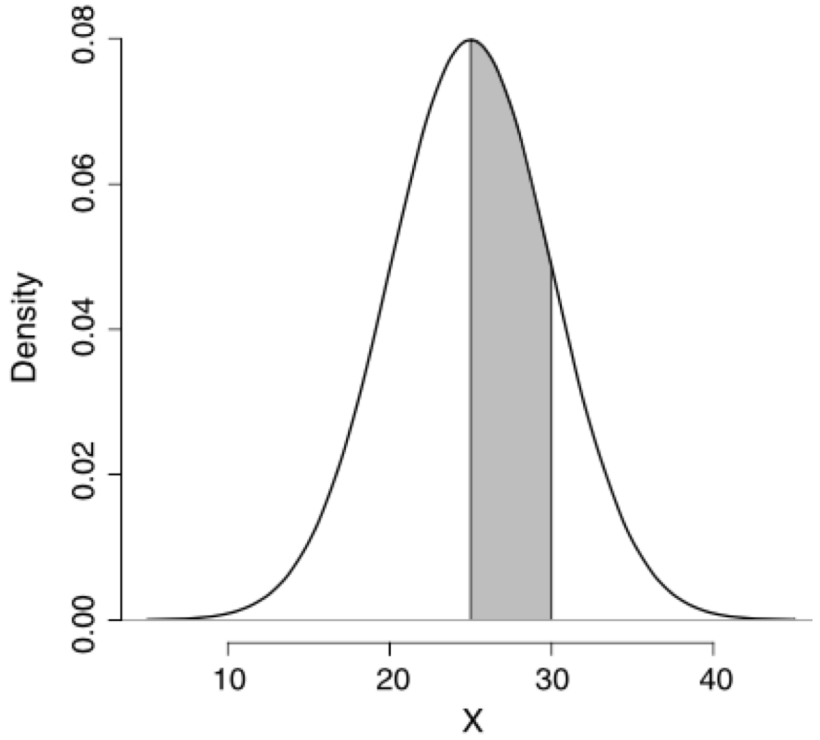
The assumed probability distribution for BMI.
The density curve shown in this figure can be used to find the probability that the value of the random variable falls within an interval.



The assumed probability distribution for BMI, which is denoted as X, along with random sample of 100 values, which are shown as *circles* along the horizontal axis.

Continuous Probability Distributions

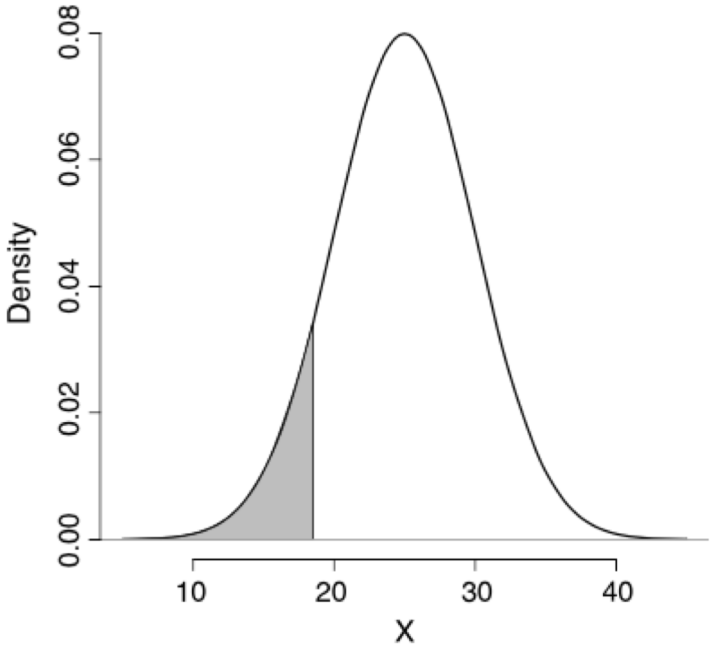
- The *shaded area* is the probability that a person's BMI is between 25 and 30.
- People whose BMI is in this range are considered as overweight.
- The *shaded area* gives the probability of being overweight



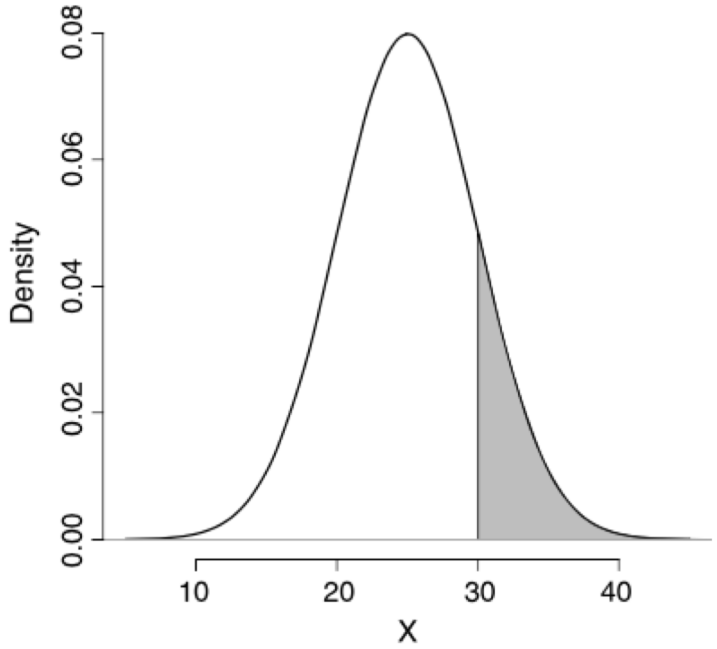
This probability is shown as the shaded area under the probability density curve between 25 and 30.
Now suppose that we shrink the interval from $25 < X \leq 30$ to $28 < X \leq 30$.
The shaded area under the curve would decrease, and the probability of the interval becomes smaller.

Continuous Probability Distributions

- Similar to the discrete distributions, 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 ≤ x).



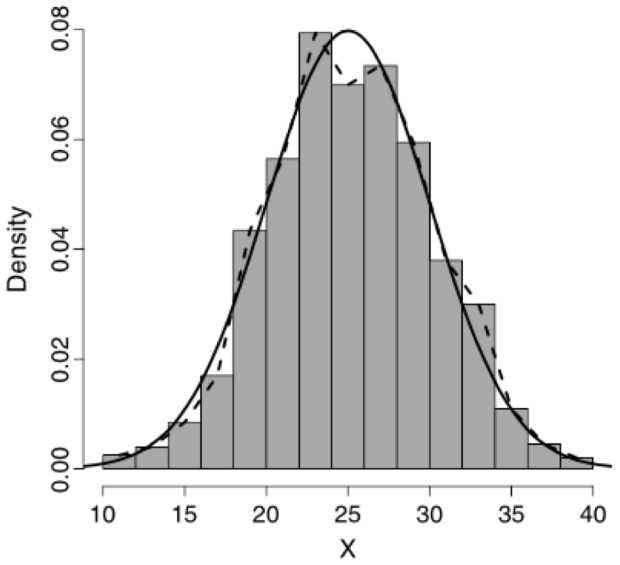
Left panel: The lower tail probability of 18.8, P (X ≤ 18.8)

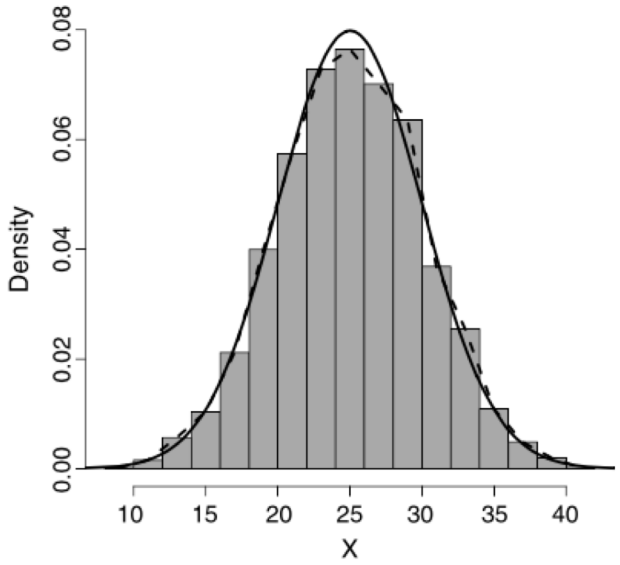


Right panel: The upper tail probability 30, P (X > 30)

Probability Density Curves and Density Histograms

- An analogous comparison can be made between density curves for probability distributions and density histograms for data.





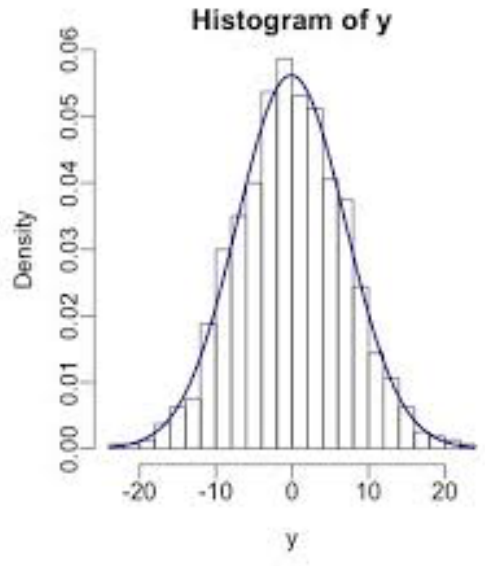
Left panel: Histogram of BMI for 1000 observations.
The *dashed line* connects the height of each bar at the midpoint of the corresponding interval.
The *smooth solid curve* is the density curve for the probability distribution of BMI.

Right panel: Histogram of BMI for 5000 observations.
The histogram and its corresponding *dashed line* provide better approximations to the density curve

Normal Distribution

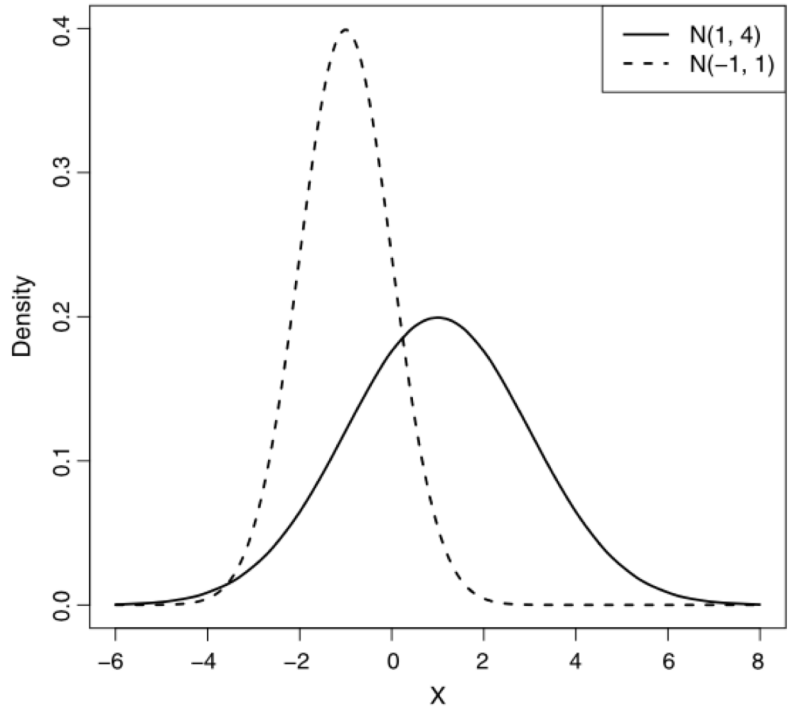
- The bell-shaped normal distribution is iconic in traditional statistics.

- The fact the distributions of sample statistics are often normally shaped has made it a powerful tool in the development of mathematical formulas that approximate those distributions.



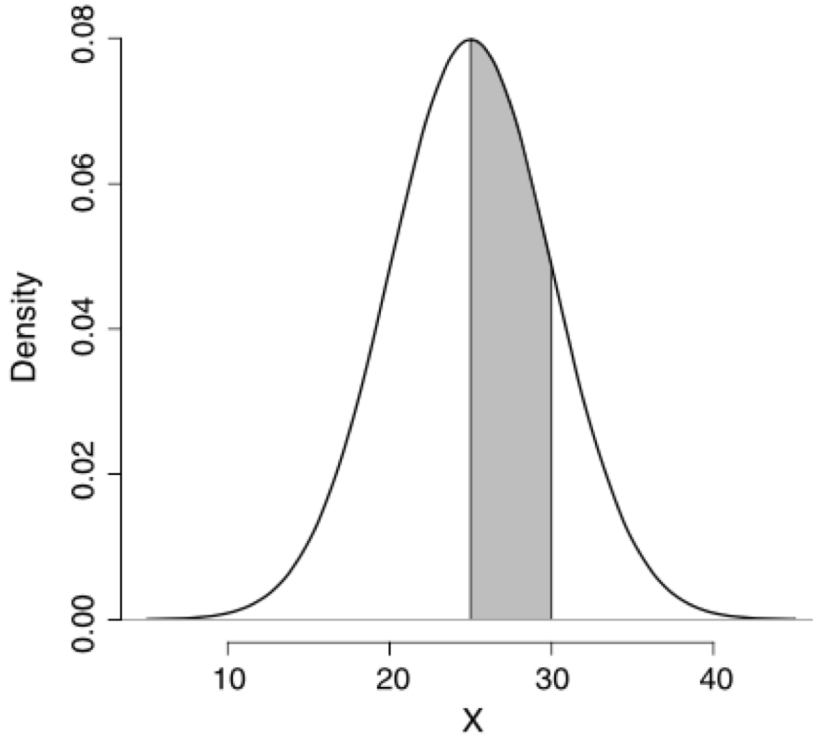
Normal Distribution

- A **normal distribution** and its corresponding pdf are fully specified by the mean μ and variance σ2.
- A random variable X with normal distribution is denoted $X \sim N(\mu,\sigma2)$, where μ is a real number, but σ2 can take positive values only.
- The normal density curve is always symmetric about its mean μ, and its spread is determined by the variance σ2.



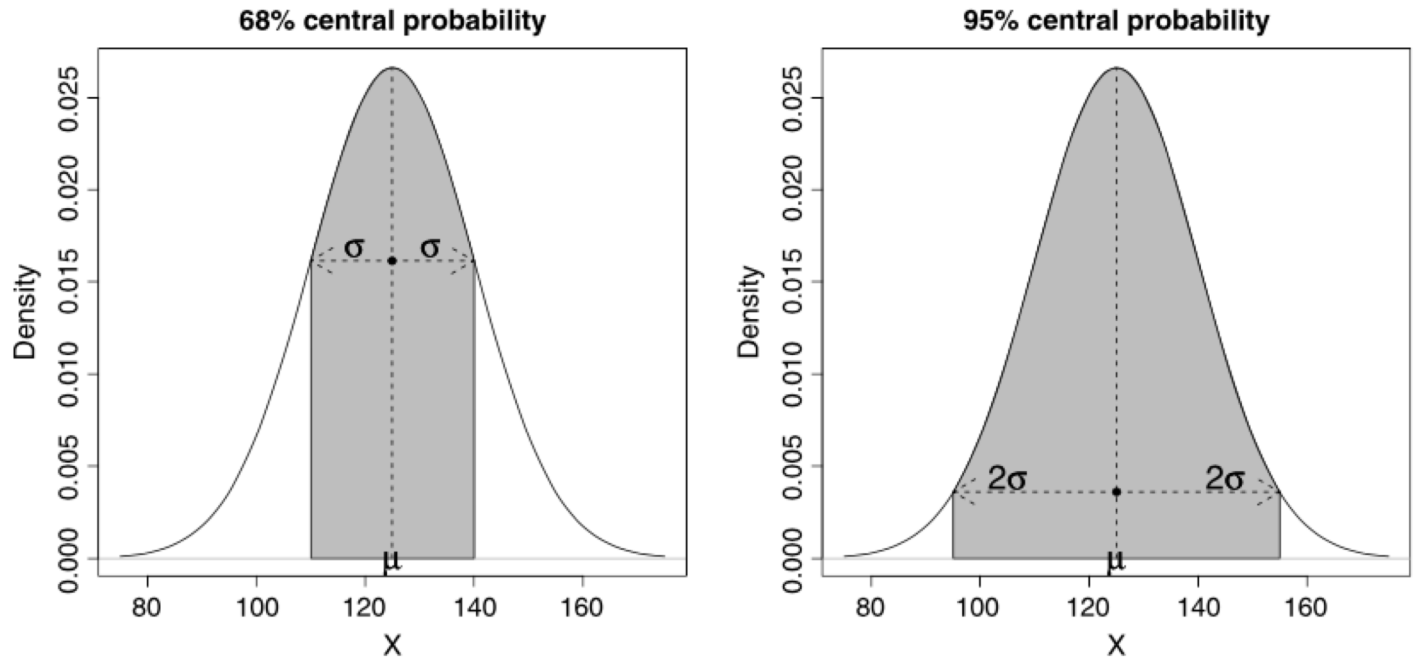
Normal Distribution

- Consider the probability distribution function and its corresponding probability density curve we assumed for BMI below. As we move from left to right, the height of the density curve increases first until it reaches a point of maximum (peak), after which it decreases toward zero.
- We say that the probability distribution is **unimodal**.
  - Because the height of the density curves reduces to zero symmetrically as we move away from the center, we say that the probability distribution is symmetric.



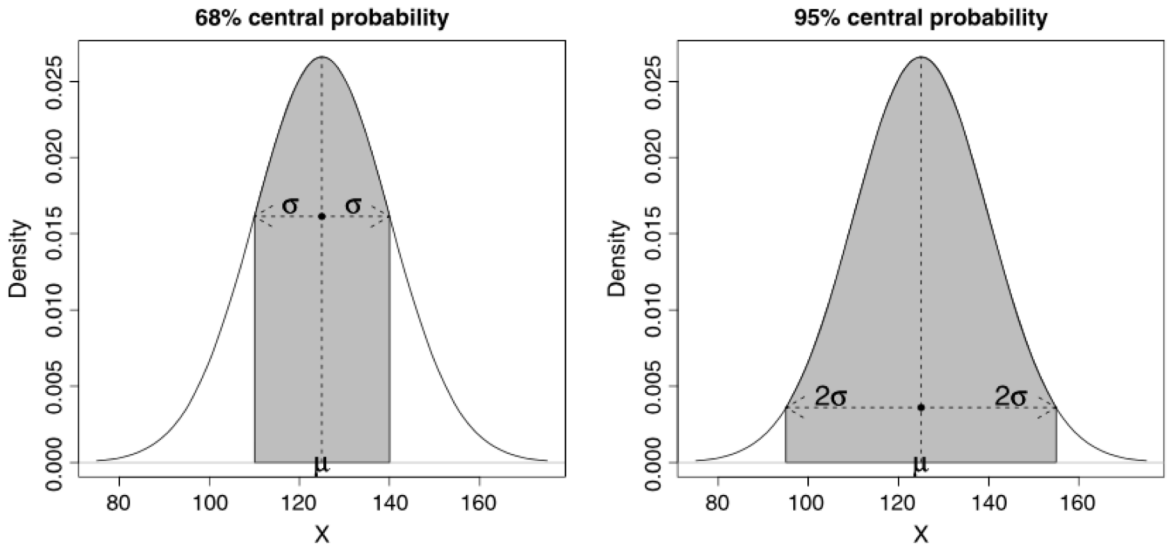
Normal Distribution

- 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: $P(\mu-3\sigma < X \leq \mu+3\sigma)=0.997$.



Normal Distribution

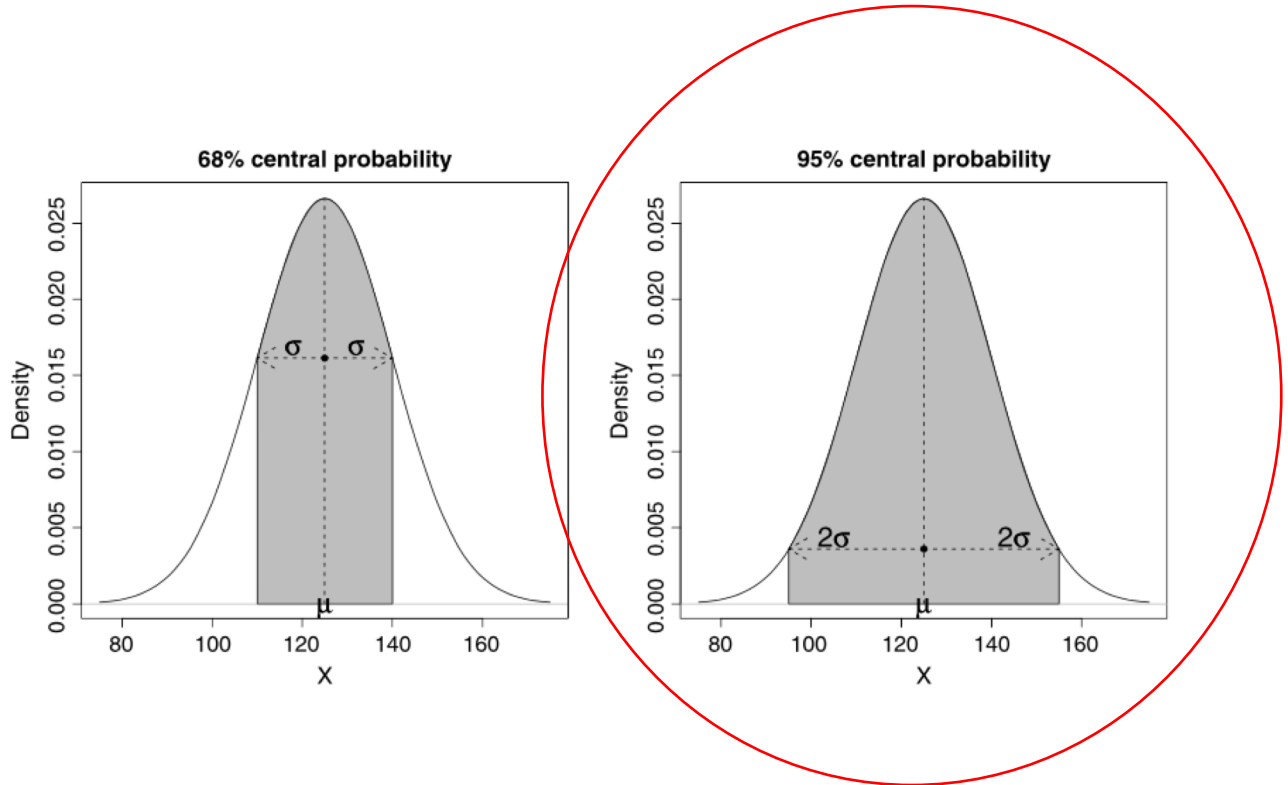
- For example, suppose we know that the population mean and standard deviation for SBP are $\mu = 125$ and $\sigma = 15$, respectively.
- That is, $X \sim N(125,152)$, where X is the random variable representing SBP.
- Therefore, the probability of observing an SBP in the range $\mu \pm \sigma$ is 0.68:
- $P(125 - 15 < X \leq 125 + 15) = P(110 < X \leq 140) = 0.68$.
- This probability corresponds to the central area shown below.



Illustrating the *68–95–99.7% Rule* for systolic blood pressure, which is assumed to have a normal distribution: $X \sim N(125, 152)$. According to the rule, we would expect 68% of the observations to fall within 1 standard deviation of the mean (*left panel*) and 95% of the observations to fall within 2 standard deviations of the mean (*right panel*).
Likewise, nearly 99.7% of observations to fall with in 3 standard deviations of the mean (not shown here)

Normal Distribution

- Likewise, 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 red circle – figure – below.
- 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.



Normal Distribution - Exercises

- Suppose that BMI in a specific population has a normal distribution with mean of 25 and variance of 16: X ~ N (25, 16).
- Then we can simulate 5 values from this distribution using the rnorm() function:

````
## Distributions:
# Normal Distributions
```{R}
rnorm(5, mean = 25, sd = 4)
```
````

```
[1] 29.83693 25.33622 20.22832 35.80395 28.87671
```

Normal Distribution - Exercises

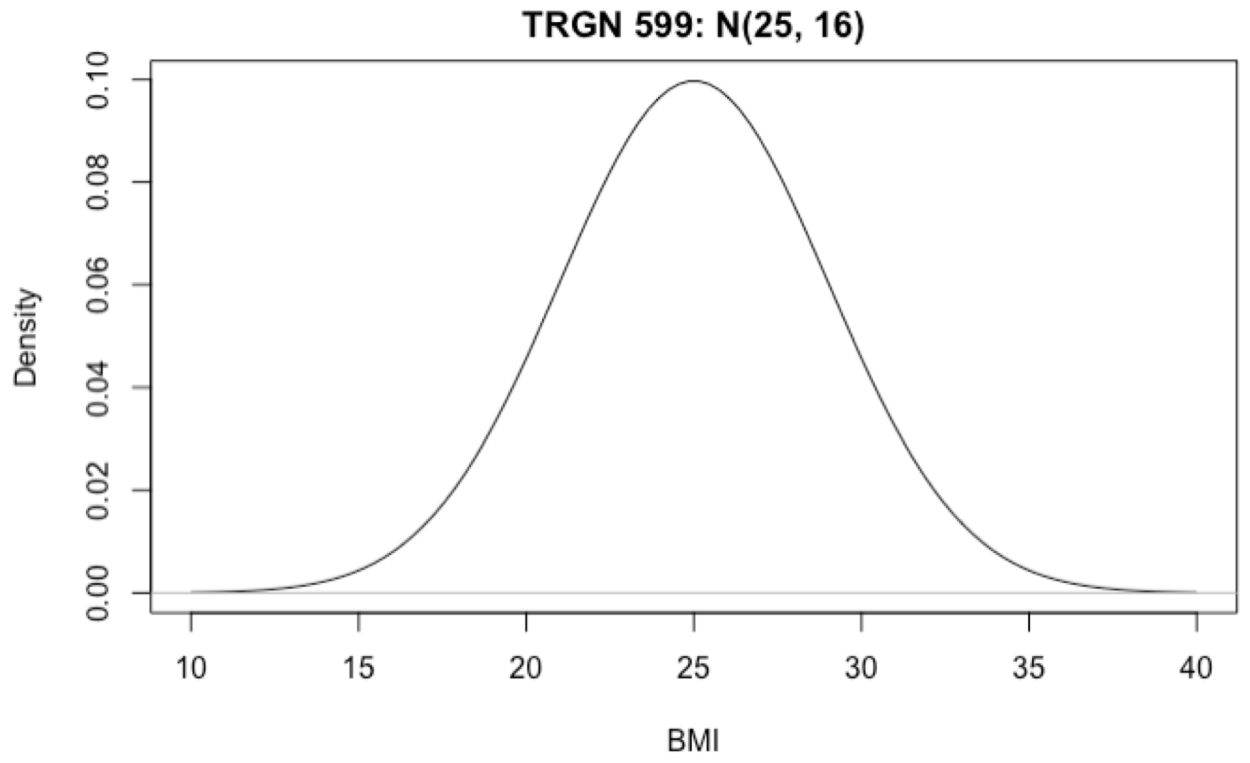
- These numbers can be regarded as BMI values for five randomly selected people from this population.
- In the *rnorm*() function, the first parameter is the number of samples, the second parameter is the mean, and the third parameter is the standard deviation (not the variance).
- Now let us plot the pdf of this distribution.
- A normal random variable can take any value from $-\infty$ to ∞.
- However, according to the *68–95–99.7% rule* approximately 99.7% of the values fall within the interval [13, 37] (i.e., within 3 standard deviations of the mean).
- Therefore, the interval [10, 40] is wide enough to plot the distribution:

````
15
16 ```{R}
17 x <- seq(from = 10, to = 40, length = 100)
18 ```
19
````

Normal Distribution - Exercises

- Here, vector `x` is a sequence of length 100 from 10 to 40.
- We can then find and plot the density for each point in the vector `x`:

```
19
20 ```{R}
21 fx <- dnorm(x, mean = 25, sd = 4)
22 plot(x, fx, type = "l", xlab = "BMI",
23      ylab = "Density", main = "TRGN 599: N(25, 16)")
24 abline(h = 0, col = "gray")
25 ```
```




```
26
```

Normal Distribution - Exercises

- The dnorm() function returns the height of the density curve at a specific point and requires the parameters of the mean and the standard deviation sd. In the plot() function, we are using type="l" to plot the points as a continuous line (curve).
- Recall that for continuous variables, the probability of a specific value is always zero.
- Instead, for continuous variables, we are interested in the probability of observing a value in a given interval.
- For instance, the probability of observing a BMI less than or equal to 18.5 is the area under the density curve to the left of 18.5.
- In R, we find this probability with the cumulative distribution function pnorm():

````
26
27 ```{R}
28 pnorm(18.5, mean = 25, sd = 4,
29        lower.tail = TRUE)
30 ```
````

```
[1] 0.05208128
```

Normal Distribution - Exercises

- Once again, we can find the upper tail probability P (X > 22) by setting the option lower.tail=FALSE.
- The qnorm() returns the quantile for normal distributions.
- For example, the 0.05 quantile for the above distribution is:

````
31
32 ```{R}
33 qnorm(0.05, mean = 25, sd = 4,
34       lower.tail = T)
35 ```
````

```
[1] 18.42059
```

Normal Distribution - Exercises

- We can find the probability of a BMI between 25 and 30 by subtracting their lower tail probabilities, $P(25 < X \leq 30) = P(X \leq 30) - P(X \leq 25)$:

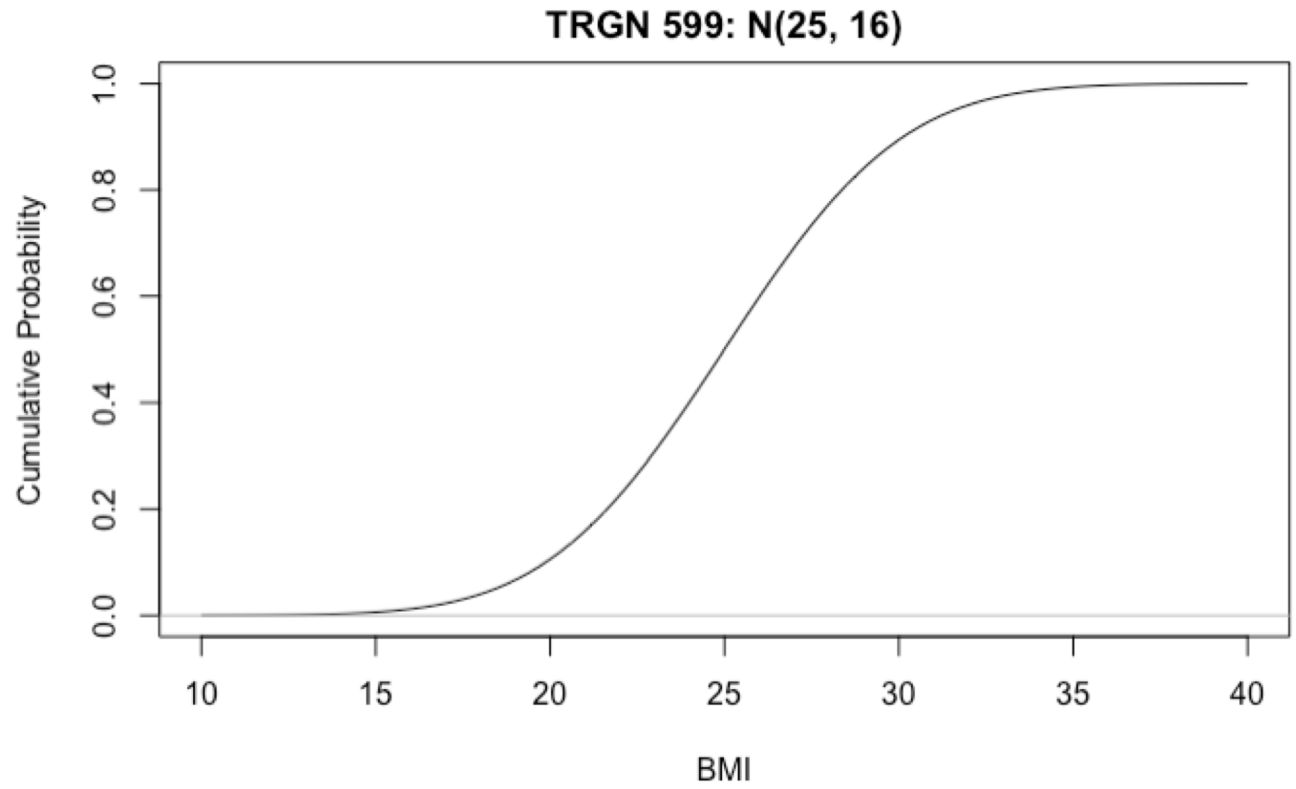
````
36
37 ```{R}
38 pnorm(30, mean = 25, sd = 4) -
39   pnorm(25, mean = 25, sd = 4)
40 ```

[1] 0.3943502
````

Normal Distribution - Exercises

- We can also create a plot of the cdf by using vector x as input to pnorm() function:

```
41
42 ```{R}
43 Fx <- pnorm(x, mean = 25, sd = 4)
44 plot(x, Fx, type = "l", xlab = "BMI",
45      ylab = "Cumulative Probability",
46      main = "TRGN 599: N(25, 16)")
47 abline(h = 0, col = "gray")
48 ```
```



Normal Distribution – Exercises – Rmarkdown

file:///Users/enriquevelazquez/Documents/R_working_directory/Rmarkdown_Week_3_Lecture_3.html

Rmarkdown_Week_3_Lecture_3

Enrique I. Velazquez Villarreal, MD, PhD, MPH, MS

1/23/2019

The following code is explained in the Week 3 Lecture 3

Distributions:

Normal Distributions

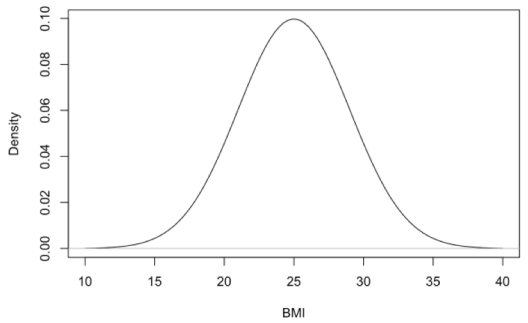
```
rnorm(5, mean = 25, sd = 4)
```

```
## [1] 29.21249 26.89937 23.73754 29.40790 25.78634
```

```
x <- seq(from = 10, to = 40, length = 100)
```

```
fx <- dnorm(x, mean = 25, sd = 4)
plot(x, fx, type = "l", xlab = "BMI",
     ylab = "Density", main = "TRGN 599: N(25, 16)")
abline(h = 0, col = "gray")
```

TRGN 599: N(25, 16)




```
pnorm(18.5, mean = 25, sd = 4,
        lower.tail = TRUE)
```

```
## [1] 0.05208128
```

```
qnorm(0.05, mean = 25, sd = 4,
      lower.tail = T)
```

```
## [1] 18.42059
```

```
pnorm(30, mean = 25, sd = 4) -
  pnorm(25, mean = 25, sd = 4)
```

```
## [1] 0.3943502
```

```
Fx <- pnorm(x, mean = 25, sd = 4)
plot(x, Fx, type = "l", xlab = "BMI",
     ylab = "Cumulative Probability",
     main = "TRGN 599: N(25, 16)")
abline(h = 0, col = "gray")
```

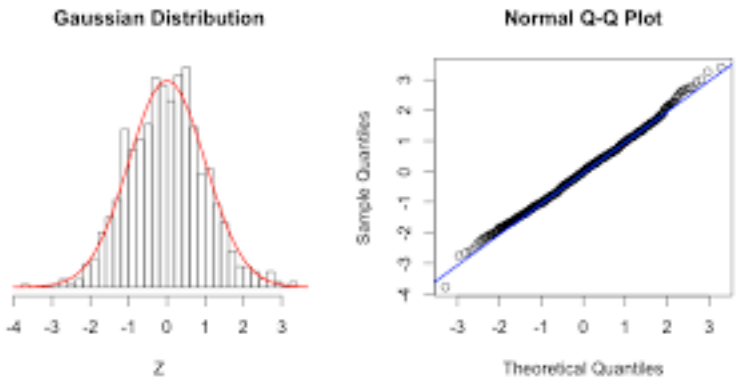
TRGN 599: N(25, 16)

Key ideas in Normal Distribution

- The normal distribution was essential to the historical development of statistics, as it permitted mathematical approximation of uncertainty and variability.

- While raw data is typically not normally distributed, errors often are, as are average and totals in large samples.

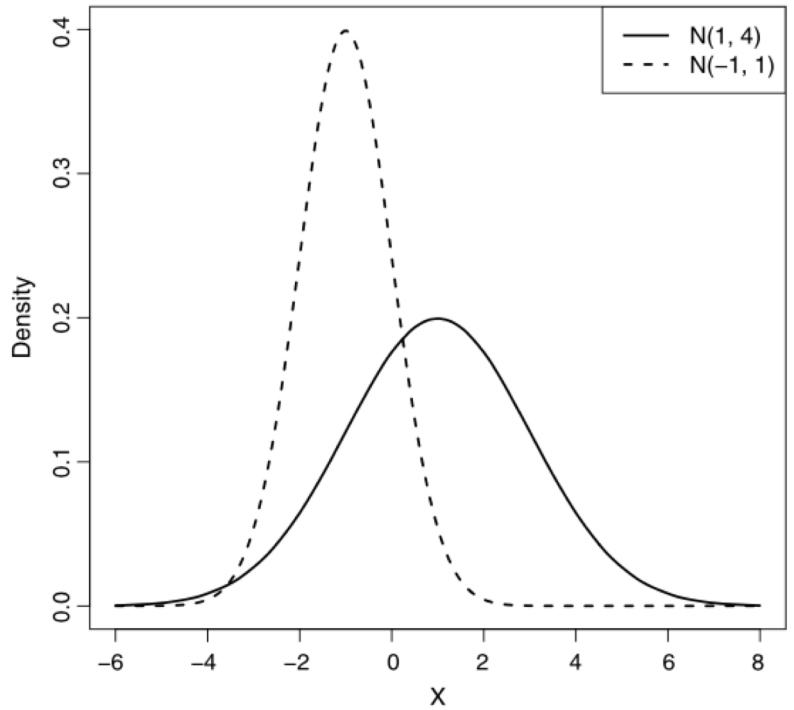
- To convert data to z-scores, you subtract the mean of the data and divide by the standard deviation; you can then compare the data to a normal distribution.



Key Terms of Distribution of a Statistic

- Central limit theorem
  - The tendency of the sampling distribution to take on a normal shape as sample size rises.

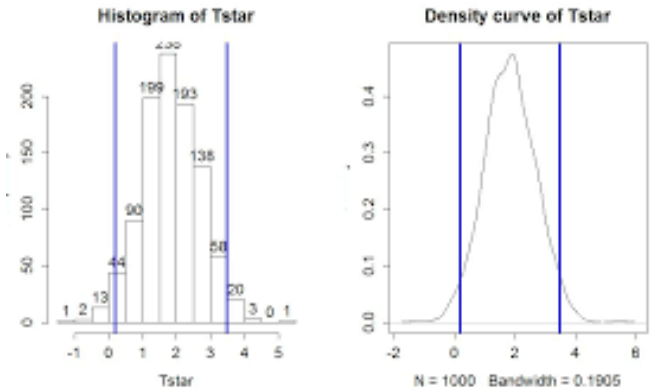
- Standard error
  - The variability (standard deviation) of a sample statistic over many samples (not to be confused with standard deviation, which, by itself, refers to variability of individual data values).



Confidence Intervals

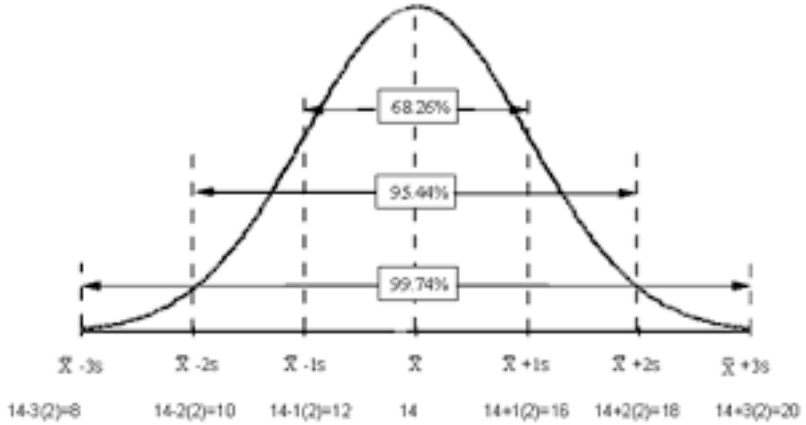
- Confidence level
  - The percentage of confidence intervals, constructed in the same way from the same population, expected to contain the statistic of interest.

- Interval endpoints
  - The top and bottom of the confidence interval.



Key Terms in Normal Distribution

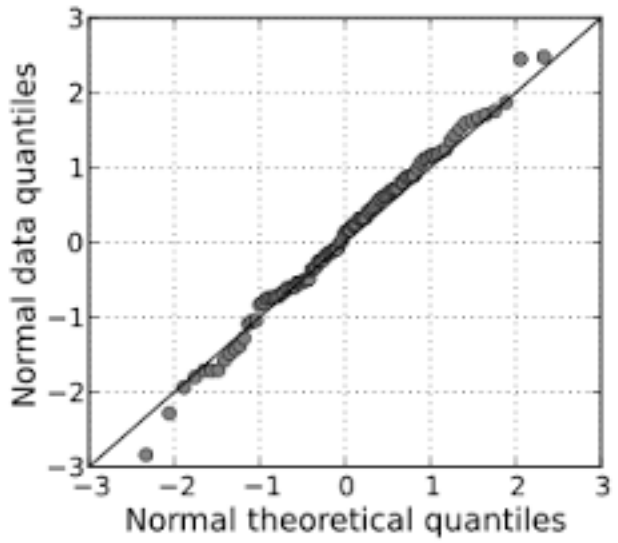
- Error
  - The difference between a data point and a predicted or average value.
- Standardize
  - Subtract the mean and divide by the standard deviation.
- Z-score
  - The result of standardizing an individual data point.



Key Terms in Normal Distribution

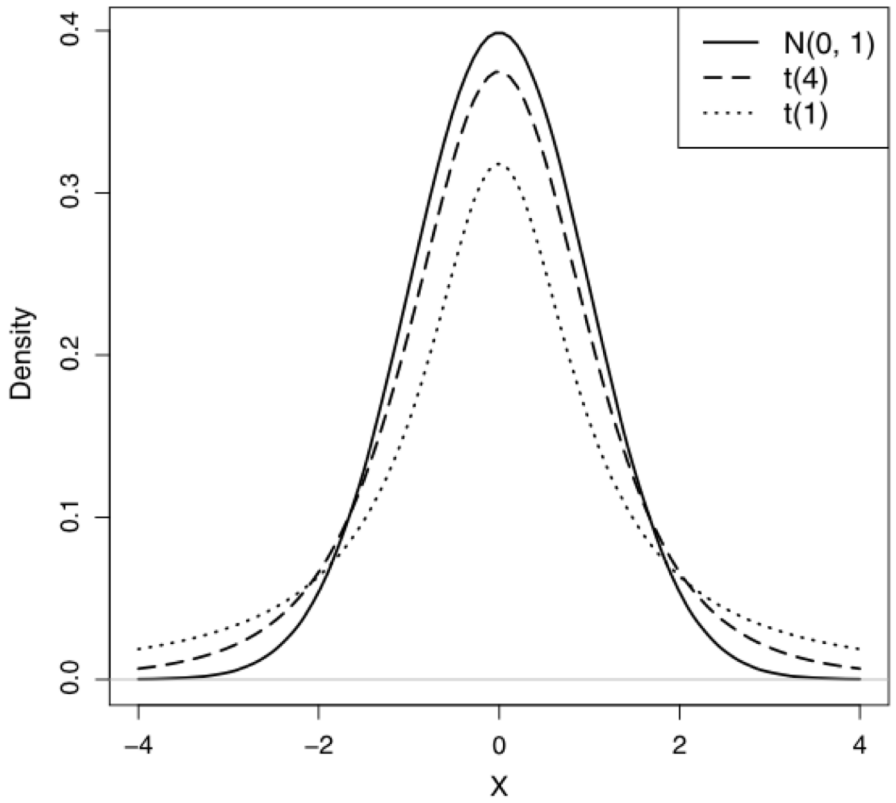
- Standard normal
  - A normal distribution with mean = 0 and standard deviation = 1.

- QQ-Plot
  - A plot to visualize how close a sample distribution is to a normal distribution.



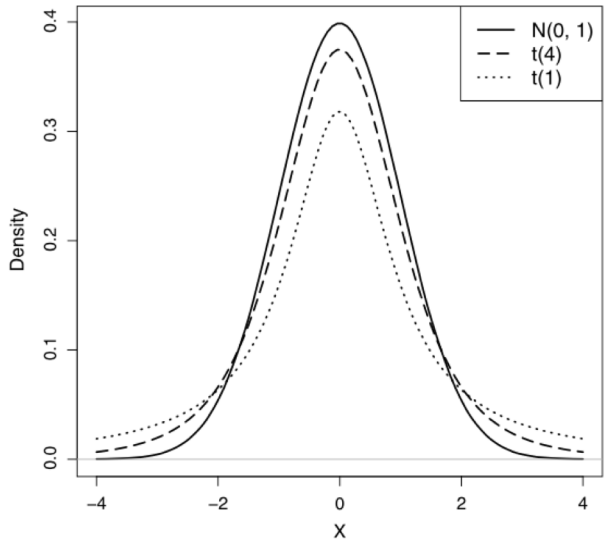
Student's t-distribution

- Another continuous probability distribution that is used very often in statistics is the **Student's** t-**distribution** or simply the t-**distribution**.
- The t-distribution especially plays an important role in testing hypotheses regarding the population mean.
  - For example, testing the hypothesis that whether the average body temperature of healthy people is the widely accepted value of 98.6°F involves the t-distribution.



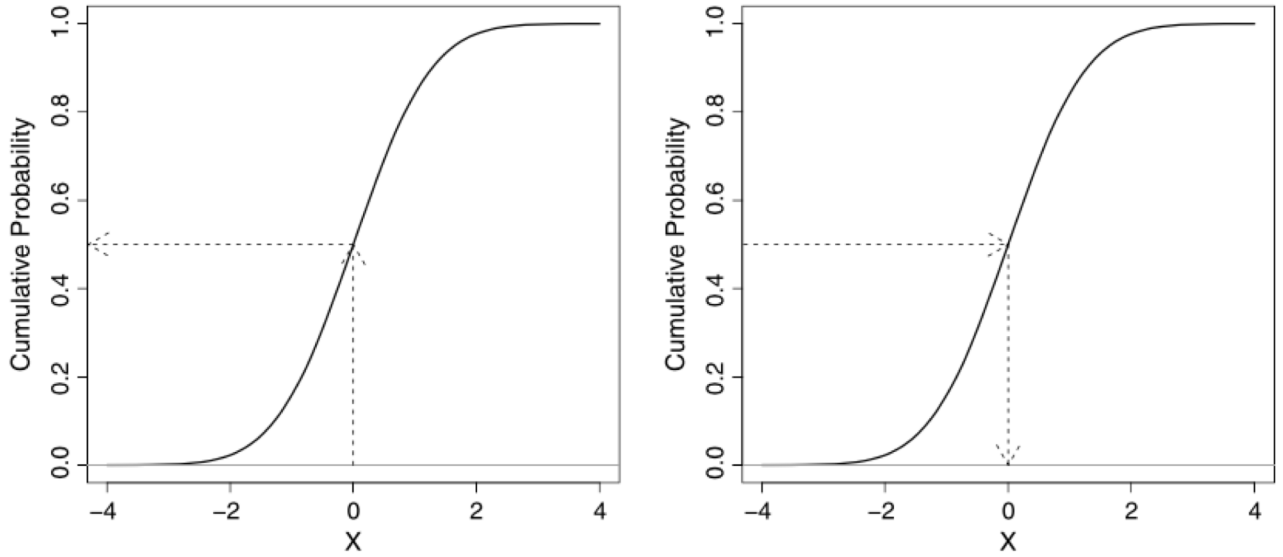
Student's t-distribution

- 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.
  - df is a positive real number (df > 0).
  - The mean of this distribution is μ = 0, and the variance is determined by the degrees of freedom parameter, σ 2 = df /(df − 2), which is of course defined when df > 2.
- 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



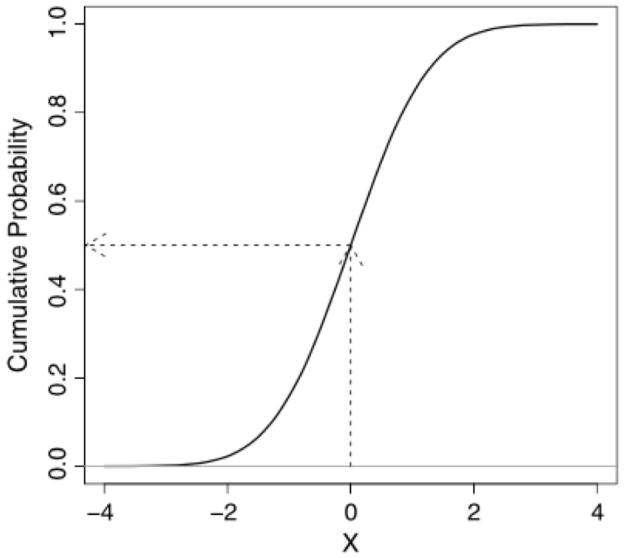
Cumulative Distribution Function and Quantiles

- We saw that by using lower tail probabilities, we can find the probability of any given interval.
- This is true for all probability distributions (discrete or continuous).
- 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.
- This function is called the **cumulative distribution function** (cdf) or simply the **distribution function**.
- For the value x of the random variable X, the cumulative distribution function returns P (X ≤ x).

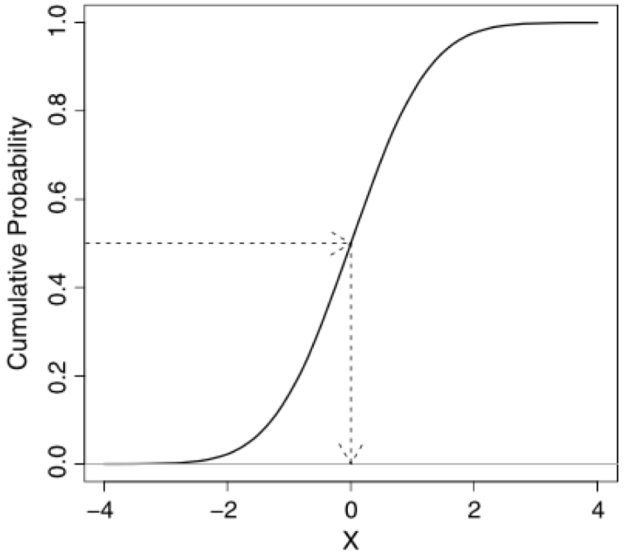


Cumulative Distribution Function

- 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 ≤ 0) = 0.5.

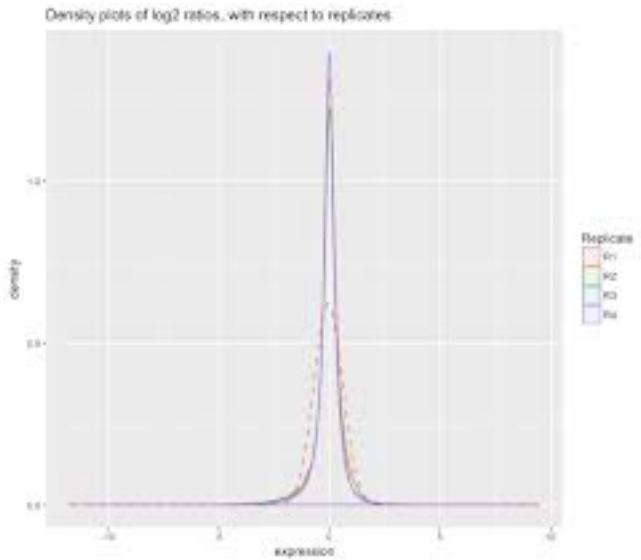


- Given the lower tail probability of 0.5 on the vertical axis, we obtain the corresponding quantile x = 0 on the horizontal axis



Long-Tailed Distributions

- Despite the importance of the normal distribution historically in statistics, and in contrast to what the name would suggest, data is generally not normally distributed.
- Most data is not normally distributed.
- Assuming a normal distribution can lead to underestimation of extreme events (“black swans”).



Key Terms for Long-Tail Distributions

- Tail
  - The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency.

- Skew
  - Where one tail of a distribution is longer than the other.