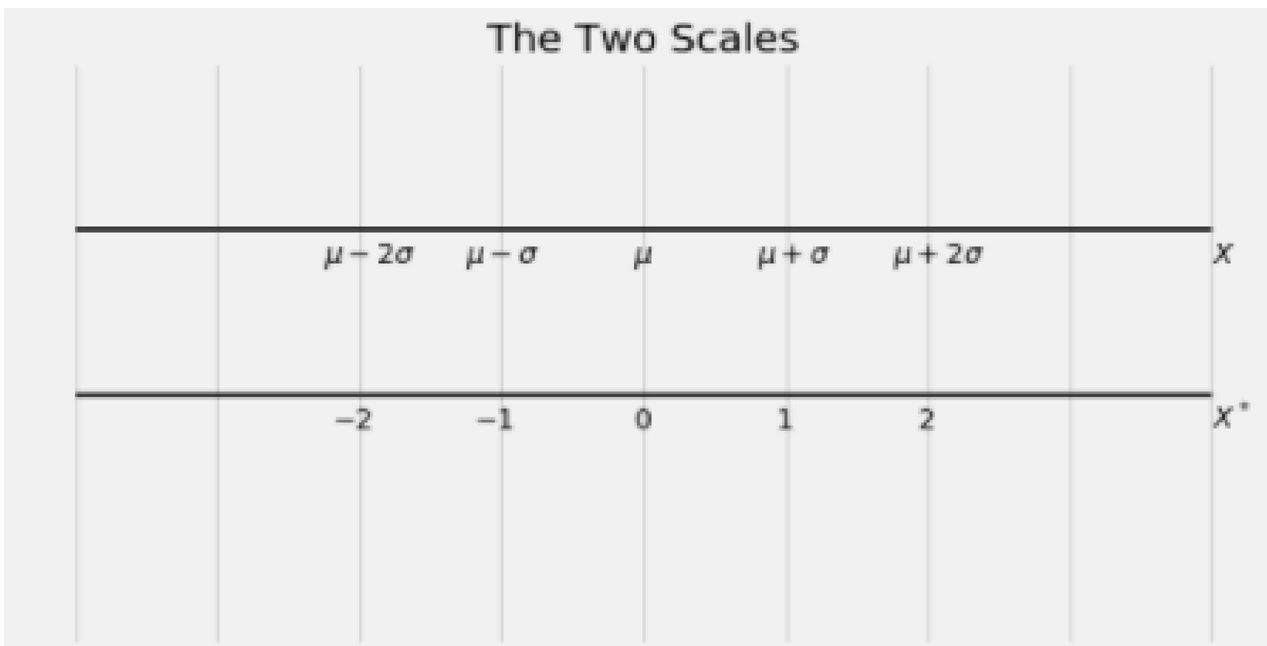
Probability and Mathematical Statistics in Data Science

Lecture 22: Section 8.3: Normal Approximation Section 8.4: How Large is Large

Standard Units

- $E(X) = \mu$ $SD(X) = \boldsymbol{\sigma}$ **Example:** $\mu = 170$ $\boldsymbol{\sigma} = 20$
- $$X^* = \frac{X - \mu}{\sigma} \qquad\qquad X = X^*\sigma + \mu$$

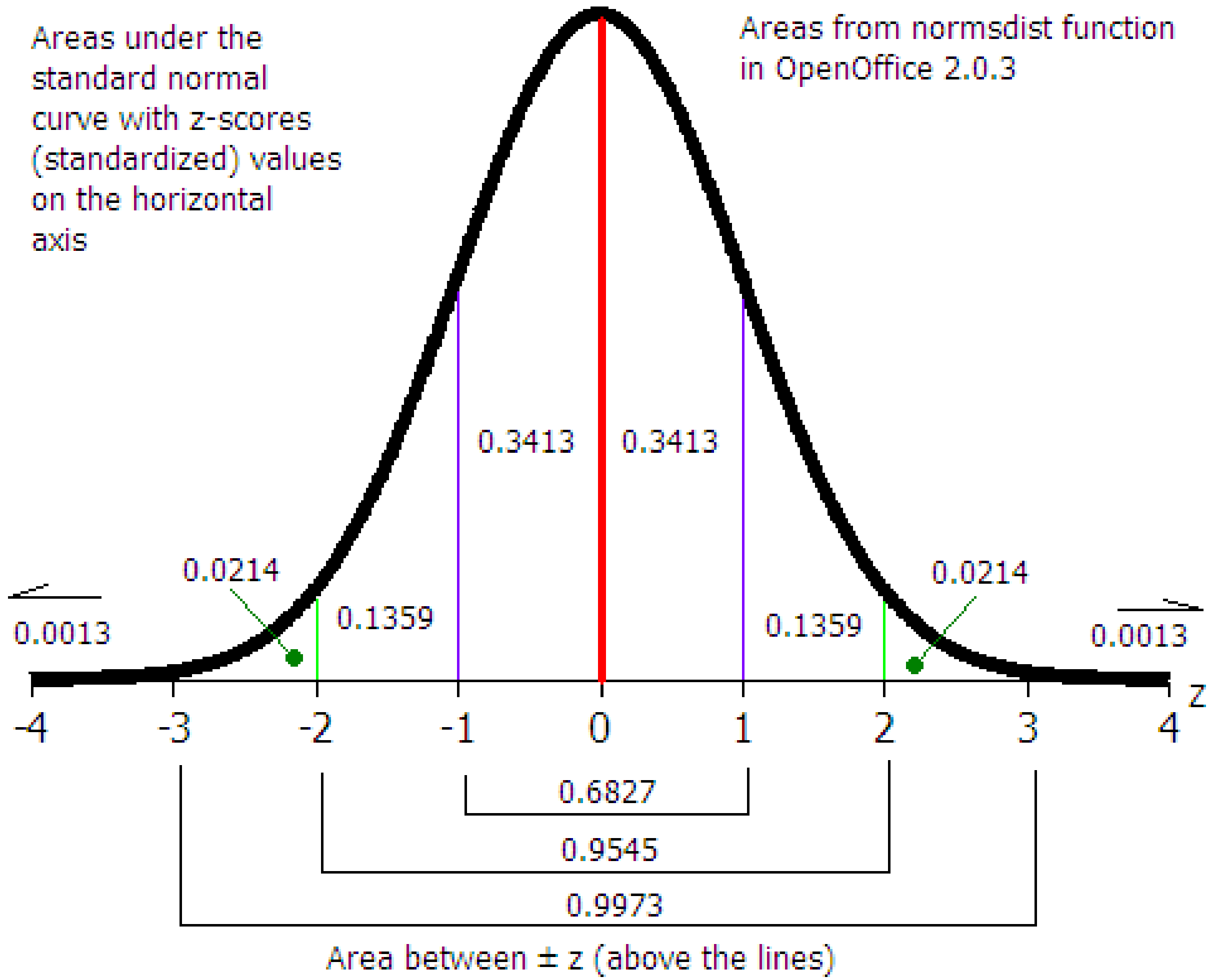


The Standard Normal Curve

1. The standard normal curve has a mean of 0 and standard deviation of 1.

2. We convert our data values to the their corresponding values on the standard normal curve.

3. This enables us to get exact probabilities of certain events.



Standardized Scores: Z-Scores

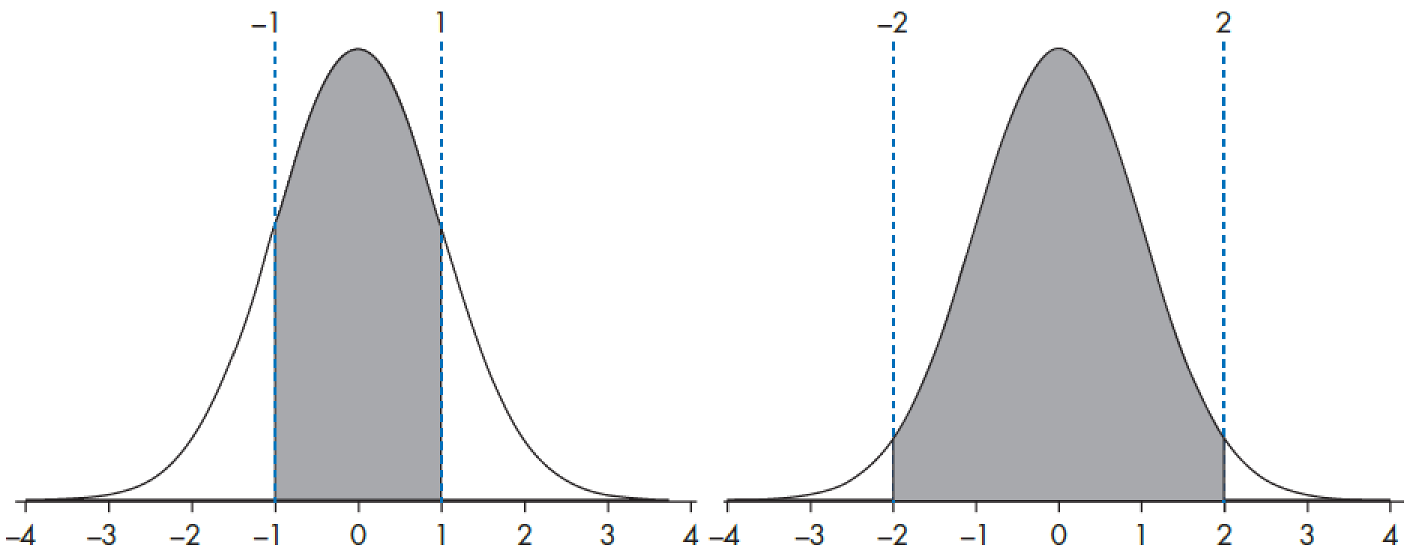
- We might be interested in measuring how many standard deviations a particular measurement value is from the mean.

- The measurement we calculate is known as a z-score, a type of standardized score calculated as follows:

$$\boldsymbol{z} = \frac{X - \mu}{\sigma}$$

The calculated z-score is a standardized score, a measurement value from what is known as the **standard normal curve (distribution)**.

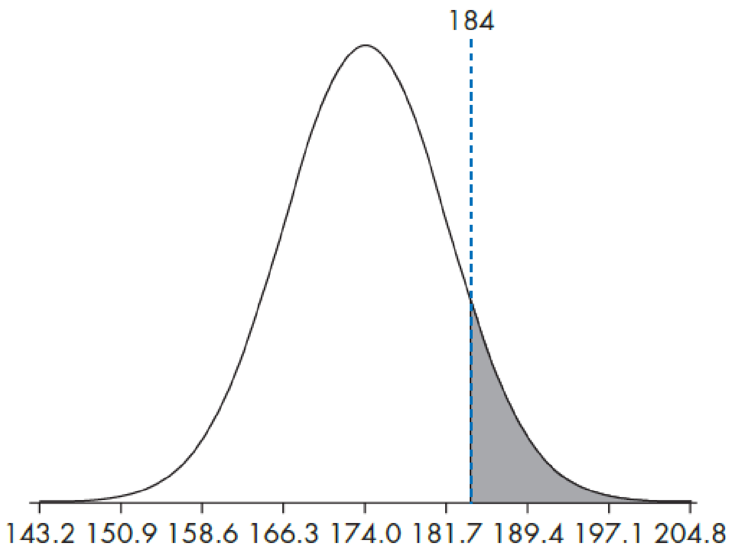
The Empirical Rule and the Standard Normal Distribution

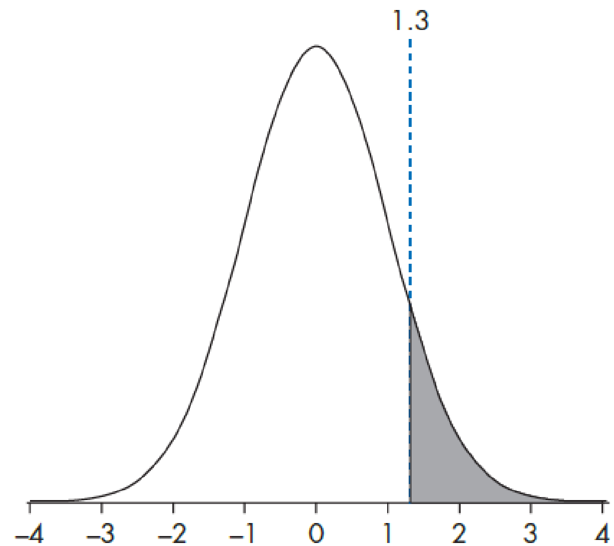


The standard normal distribution has a mean equal to 0 and a standard deviation equal to 1

- 68% of measurement values will have z-scores between -1 and +1
- 95% of measurement values will have z-scores between -2 and +2
- 99.7% of measurement values will have z-scores between -3 and +3

NHANES Men's Height





Mean: 174 cm **Standard Deviation**: 7.7 cm **Value**: 184

z-score = (value – mean)/standard deviation

z-score = (184 – 174)/7.7 = 1.3

The height value of 184 is 1.3 standard deviations above the mean height of 174

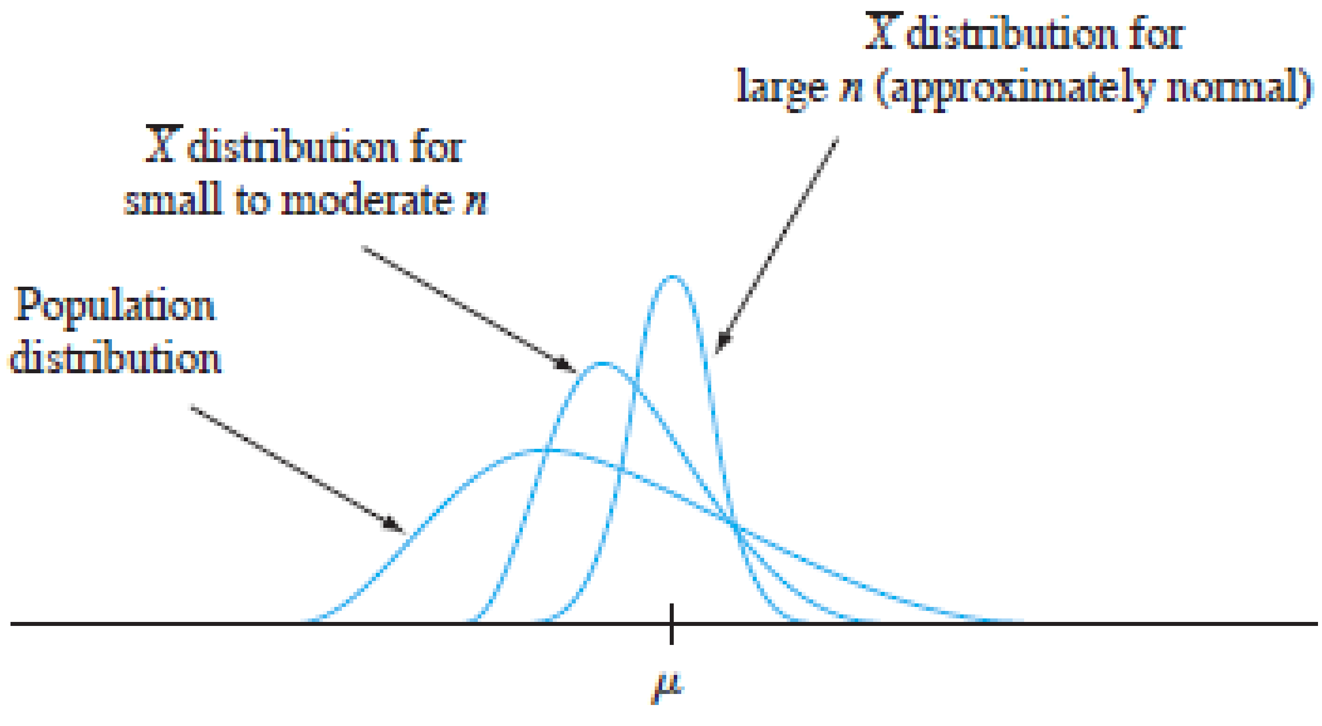
Example: Devore page 244

- The time that it takes a randomly selected rat of a certain subspecies to find its way through a maze is a normally distributed random variable with mean= 1.5 min and standard deviation= .35 min.

- Suppose five rats are selected. Let $X1, \ldots, X5$ denote their times in the maze. Assuming the Xi's to be a random sample from this normal distribution, what is the probability that the total time $S_0 = X1 + \ldots + X5$ for the five is between 6 and 8 min?

- What is the probability the sample mean is at most 2?

The Central Limit Theorem – How Large is Large?



Central Limit Theorem: How Large is Large?

Let $X_1, X_2, X_3, \ldots, X_n$ be an IID sequence from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, then sample mean $\bar{X}$ approximately follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

The larger the sample size n is, the better the approximation.

Rule of Thumb: if $n \geq 30$, Central Limit Theorem can be used.

The Normal Approximation to the Binomial

- When dealing with a large number of trials in a Binomial situation, making direct calculations of the probabilities becomes tedious (or outright impossible).

- Fortunately, the Normal model comes to the rescue…

Sampling Distribution Model for Proportions – Mean and Standard Error

- Let Y - Binom(n,*p*) where n is the number of trials and p is the probability of success.

$$\hat{p} = \frac{Y}{n}.$$

So,

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{E(Y)}{n} = \frac{np}{n} = p,$$

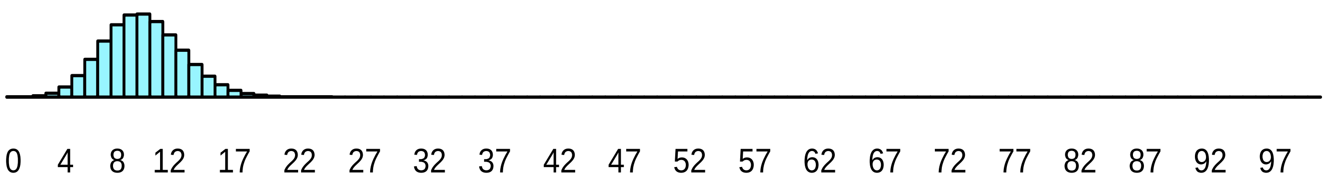
the true proportion of the population. And

$$SD(\hat{p}) = SD\left(\frac{Y}{n}\right) = \frac{SD(Y)}{n} = \frac{\sqrt{npq}}{\sqrt{n^2}} = \sqrt{\frac{npq}{n^2}} = \sqrt{\frac{pq}{n}}.$$

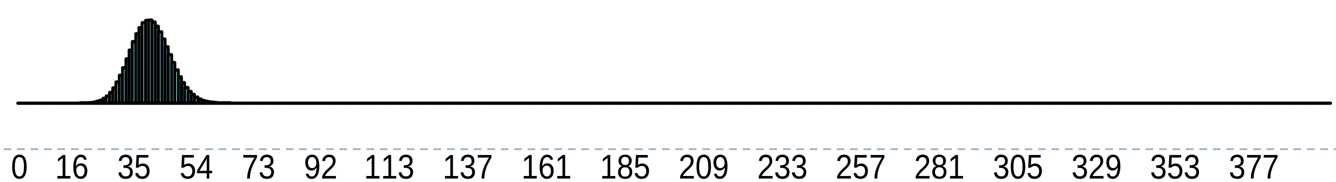
When p is small

n=25

n=100



n=400



The Normal Approximation to the Binomial

- As long as the **Success/Failure Condition** holds, we can use the Normal model to approximate Binomial probabilities.

  - **Success/failure condition**: A Binomial model is approximately Normal if we expect at least 10 successes and 10 failures:

$$np \geq 10 \text{ and } nq \geq 10$$

The Normal Approximation to the Binomial

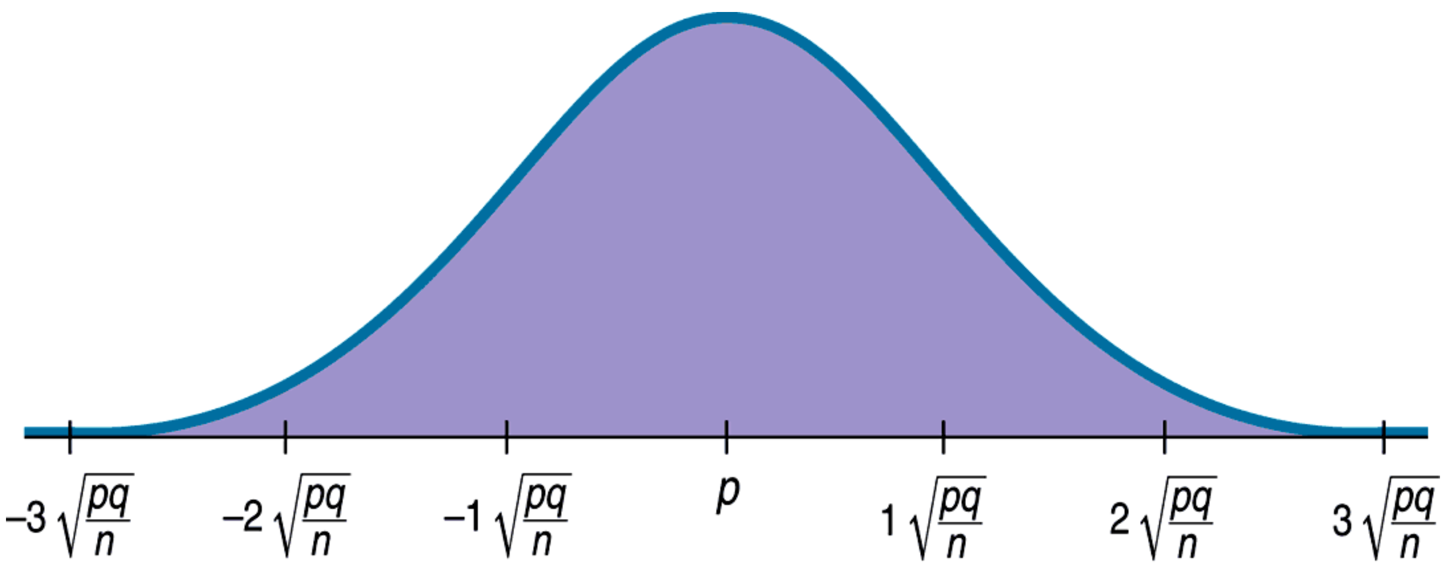
- Condition: np>=10, n(1-p)>=10

- Binomial can be closely approximated by a normal distribution with standardized variable

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{X - np}{\sqrt{npq}}$$

The Central Limit Theorem for Sample Proportions

- A picture of what we just discussed is as follows:



What to Expect of Sample Proportions

- **Example :** Suppose **40% of all voters** in U.S. favor candidate X. Pollsters take a sample of 2400 people.