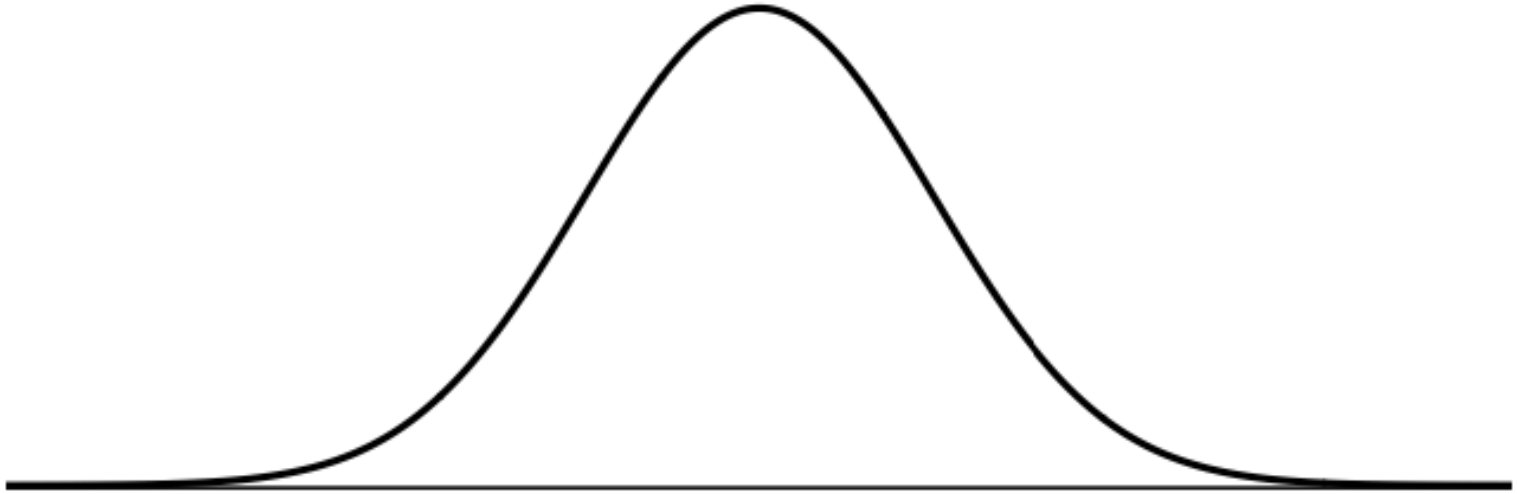
Week 4: Distributions

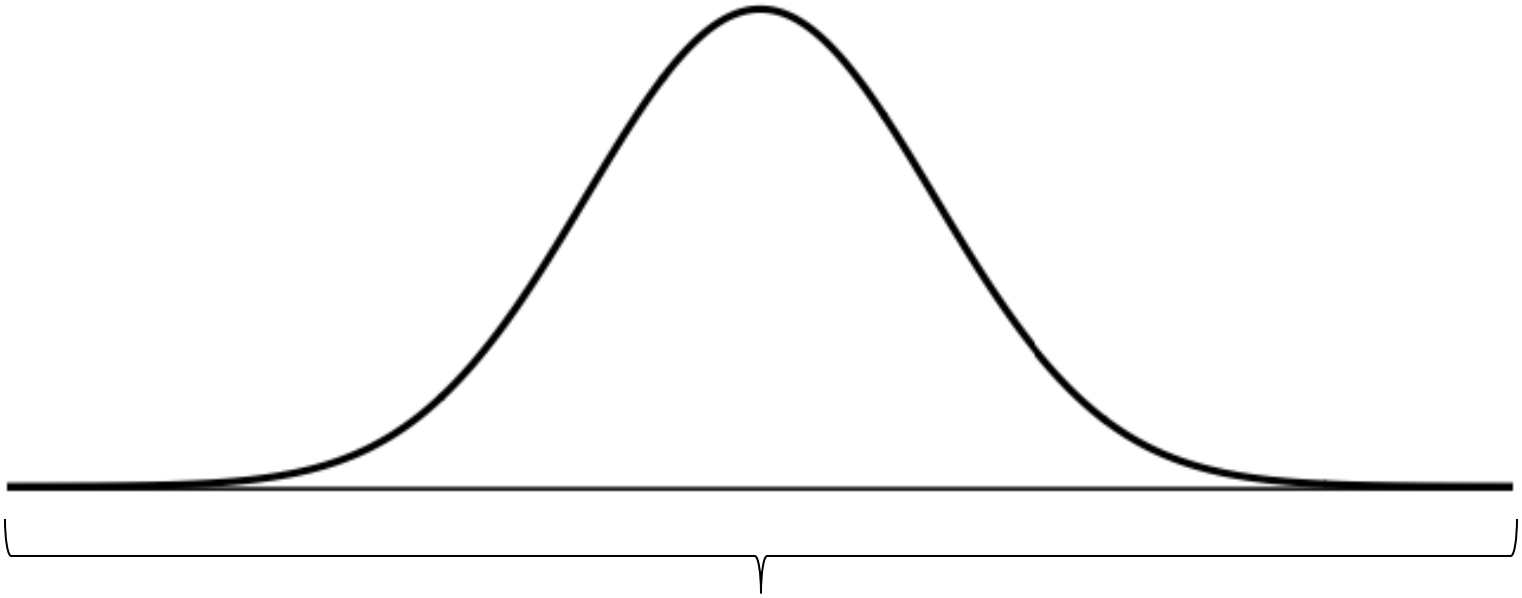
Outline

1. Normal distribution
2. Z scores
3. Generating probabilities, percentiles.
4. Bernoulli distribution
5. Binomial distribution
6. Assignment 2

Normal Distribution

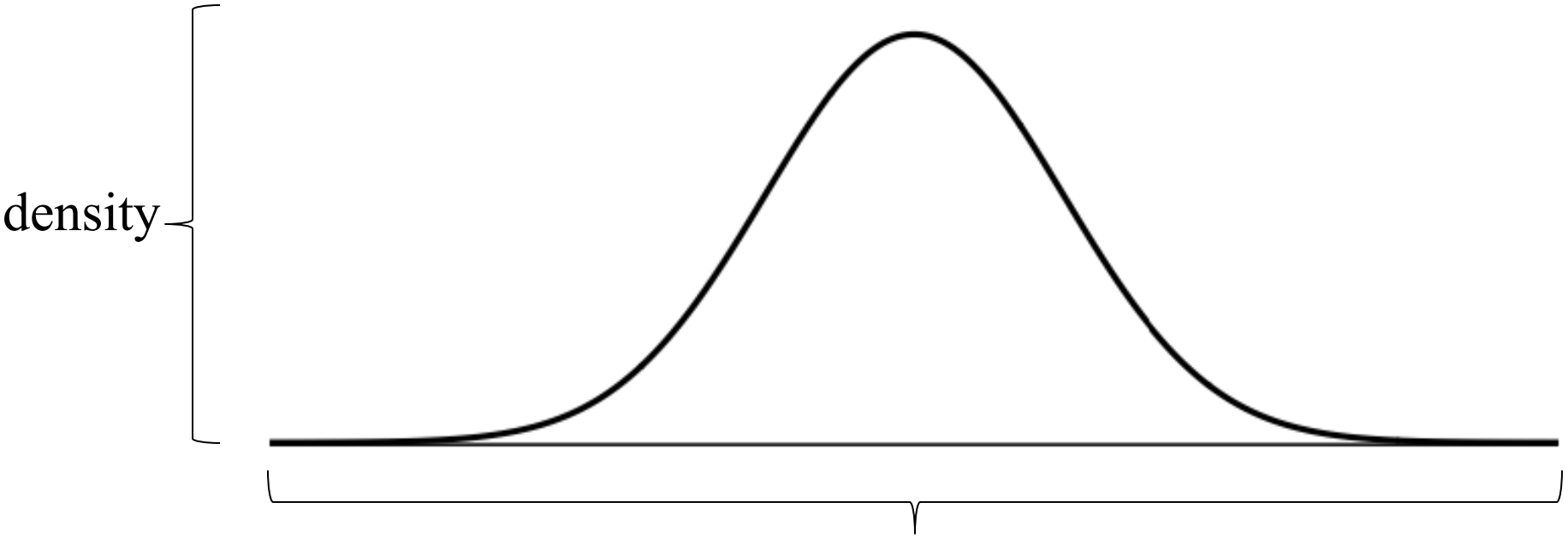


Normal Distribution



Values of some variable (from min to max)

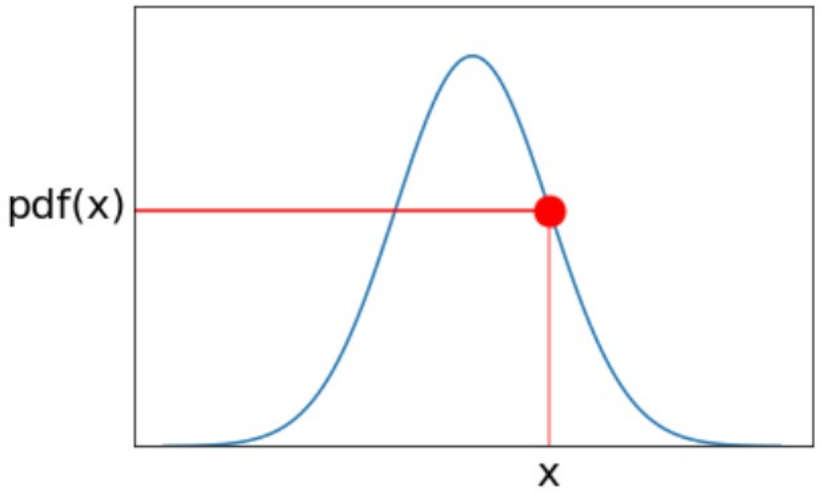
Normal Distribution



Probability Density Function (PDF)

- When you look at a density plot, what does the y-axis represent?
- The "density" is actually derived from the "probability density function" (pdf) for the normal distribution:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$



Exercise in R…..
- Histogram and "dnorm"

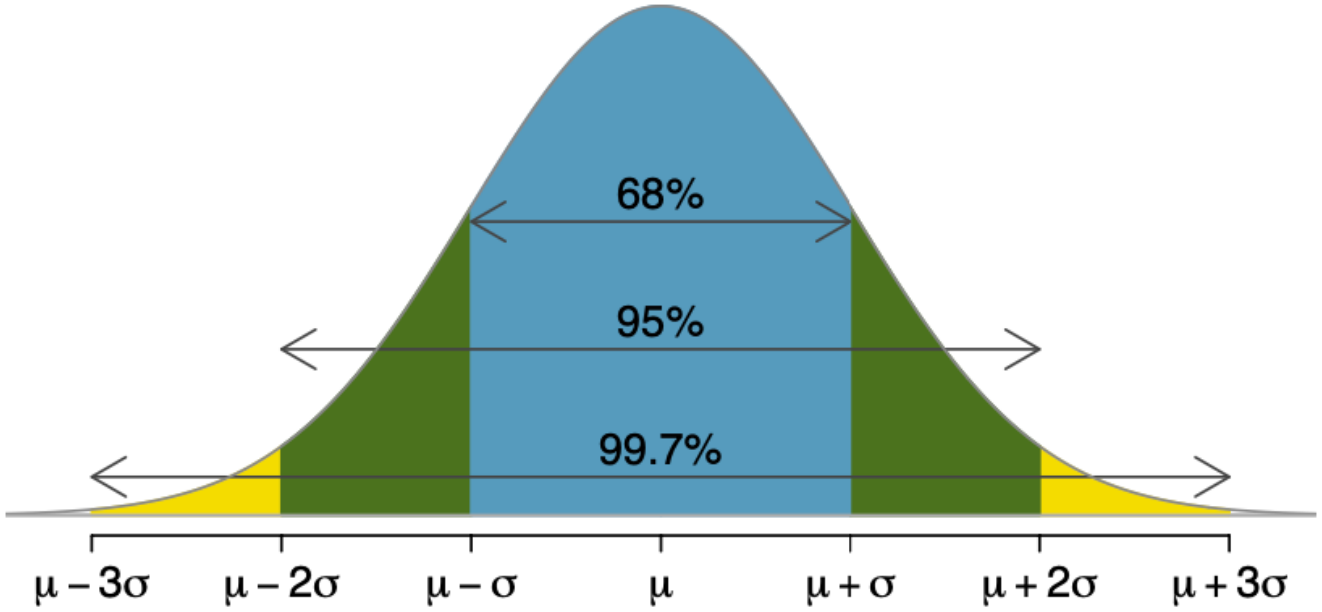
Probability Density Function (PDF)

- Careful: *A "probability density" is not a probability*.

  - Doesn't have an intuitive value or meaning (like prob.)
  - Intuition: an instantaneous likelihood of an event at a particular value of X.
  - Or the "probability per unit of measurement" (values on the X-axis).
  - In a density plot, best used for relative frequencies/comparisons – how frequently x=6 occurs relative to how frequently x=20 occurs.

Probability Density Function (PDF)

- Two core parameters in normal distribution: **mean** and **variance** (or **s.d.**); Distribution written as: $N(\mu,\sigma)$
  - μ = *mean;* σ^2*=variance;* σ = *s.d.*
  - $N(\mu=0,\sigma=1)$: *What special name does this one have*?
    - "Standard normal distribution"
  - $N(\mu=19,\sigma=4)$ or more simply, $N(19, 4)$
- Also, <u>total area under the curve</u> = **1**
- Properties:



Cumulative Distribution Function (CDF)

- The more intuitive use of the normal relates to the "cumulative distribution function" (CDF).

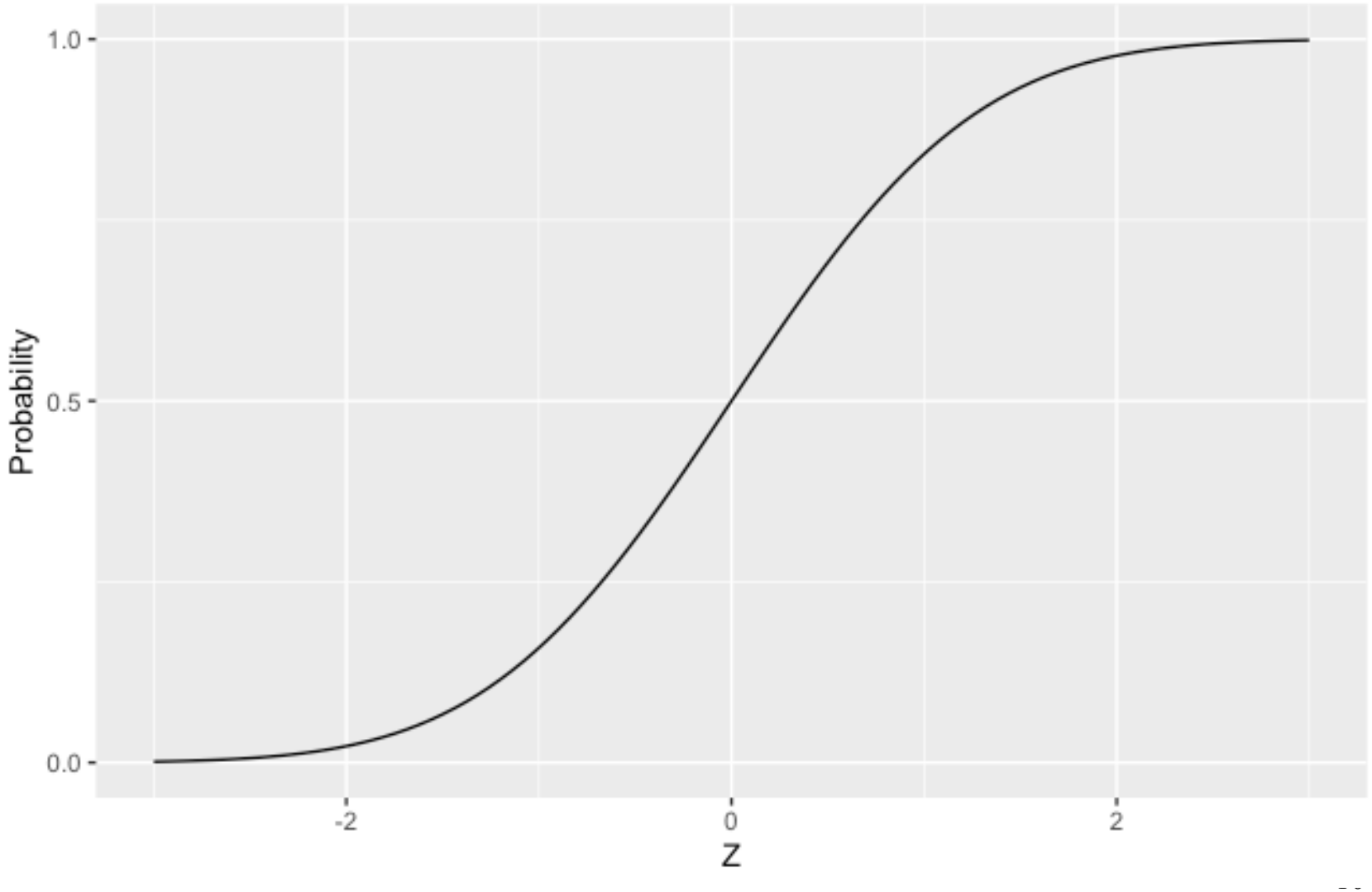
$$F(x) = \int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

- It's the area under the curve to the *left* of some value, x. (Hence the integral)

- That area is a *probability*, i.e., $P(X \leq x_i)$

- What's the probability that Bobby scored lower on the SAT than Suzie?
- What the probability that Suzie scored between 1200 and 1400?

CDF for the Normal



Cumulative Distribution Function (CDF)

- Note/relevance: PDFs and CDFs are the foundation of maximum likelihood theory and estimation, which you'll learn more about in this class with logit and probit, and other classes with other models (Poisson/count models, duration models, ordered probit, multinomial logit, etc.).

- Notice the convention: $F(x) = \mathbf{\Phi}(x) = \text{CDF}$

  $f(x) = \phi(x) = \text{PDF}$

- We can calculate $\text{P}(X \leq x_i)$ using "pnorm" in R.

Examples using the CDF

- Use graphs! OpenIntro

- Example: For SAT scores, mean=1100, sd=200. Bobby gets a 950. What's the probability that someone chosen at random scores below Bobby?
  - Graph it
  - What's another way of interpreting this quantity?

- What fraction scored higher than Bobby?
  - Graph it
  - How do we use the pnorm command now?

Standardizing using Z-scores

- We can standardize variables using Z-scores.

- Why might we want to standardize variables?

- A z-score is a transformation of a variable that changes the units of measurement (e.g., age in years, income in dollars, women's representation in %) into "standard deviation units."

THE Z-SCORE

The Z-score of an observation is the number of standard deviations it falls above or below the mean. We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

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

- In other words, a z-score transformation makes the transformed variable a "standard normal distribution": N(mean=0, sd=1)
- Example in R using women's representation

Redo examples by converting to Z-scores

- Example: For SAT scores, mean=1100, sd=200. Bobby gets a 950. What's the probability that someone chosen at random scores below Bobby?
  - Graph it
  - Convert to Z-score
  - Use R
  - What's another way of interpreting this quantity?

- What fraction scored higher than Bobby?
  - Graph it
  - Convert to Z-score
  - Use R

Age Examples

- In the adult population (18 and older), mean age is 51 and sd=17.

- What is the probability that a random adult is less than 30 years old?
  - Graph it
  - Convert to Z
  - Use pnorm in R

- P(older than 75)

- P(between 40 and 60)

Age Examples: The Reverse…..

- In the adult population (18 and older), mean age is 51 and sd=17.

- Suzie is at the 45th percentile in age. How old is she?
  - Graph it
  - Convert to Z using "qnorm"
  - Do some algebra to solve.

Bernoulli Distribution

- Suitable for binary outcomes (success or fail, war or not war, vote D or R, change in govt or no change)
  - Relevance: Foundation for logit and probit

$$\hat{p} = \frac{\#\text{ of successes}}{\#\text{ of trials}} = \frac{1+1+1+0+1+0+0+1+1+0}{10} = 0.6$$

BERNOULLI RANDOM VARIABLE

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1-p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad \sigma = \sqrt{p(1-p)}$$

Bernoulli and Geometric

- Geometric distribution: How many trials does it take before you observe a success?

- P(success) = 0.6, P(fail) = 0.4

- P(first event or trial "drawn" is a success) =
- P(second event drawn is the first success) =
- P(third event is the first success) =
- P(nth event is the first success) =

- This distribution applies to "independently and identically distributed" (iid) random variables.
  - Relevance: this assumption applies to regression modeling, related what we assume about Y and the errors.

Bernoulli and Geometric

- General rule:

GEOMETRIC DISTRIBUTION

If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

- How many trials on avg. will it take for a first success when P(success) = 0.9?
- When P(success) = .2?

Bernoulli and Geometric

- How about this one:

- P(success) = 0.6, P(fail) = 0.4

- What is the probability that we'll a get our first success *within* one of the first four trials/events? That is, that our first success comes at trial 1 *or* trial 2 *or* trial 3 *or* trial 4?

Binomial Distribution

- Relevance: Foundation for logit and probit (binary outcomes), modeling the prob(Y=1).

- Binomial distribution has three key quantities:

  k = exact number of successes (1=success, 0=failure)

  n = number of independent trials or draws

  p = probability of success

- Binomial distribution describes the prob of having exactly k successes in n independent trials with prob(success) = p.

Binomial Example

- Vote choice in 2024. 1=Vote for Trump, 0=Vote for Biden
- Let's estimate that P(1 [Trump]) = 0.45

- We randomly select 5 voters. What is the probability that exactly one voter will vote for Trump?

- Use this general formula for the "single scenario." What do we have to do to that quantity to get the probability?

$$p^k(1-p)^{n-k}$$

Binomial

- Now we need a general formula for the number of ways we get *k* successes (when *k* is 1, 2, 3, 4, 7, etc).

- Number of ways to get *k* successes in *n* trials:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Use this formula for prior example:
  - How many ways to get 1 Trump voter (Note: Y=1 is "success" in binary outcome models, so success=Trump in our example)
  - How many ways to get 3 Trump voters?
  - How many ways to get 5 Trump voters?

Binomial

- Now combine our two general formulas into one general formula: number of instances x prob(single scenario)

BINOMIAL DISTRIBUTION

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

The mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)}$$

Binomial

- Now combine our two general formulas into one general formula: number of instances x prob(single scenario)

BINOMIAL DISTRIBUTION

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

The mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)}$$

IS IT BINOMIAL? FOUR CONDITIONS TO CHECK.

(1) The trials are independent.
(2) The number of trials, n, is fixed.
(3) Each trial outcome can be classified as a *success* or *failure*.
(4) The probability of a success, p, is the same for each trial.

Another Binomial Example

- Same Trump example. P(Vote Trump) = .45
- Now we have 100 randomly selected voters.
- What's the probability that 51 of those voters vote for Trump?

- Check our math using the dbinom function in R.