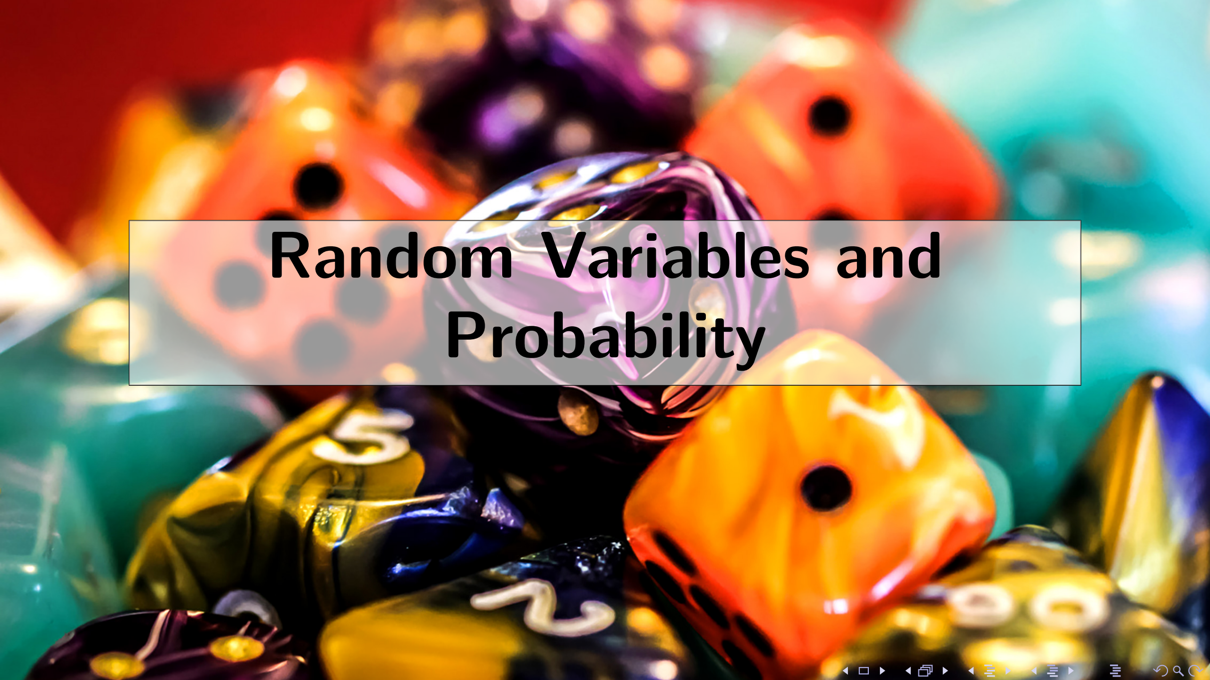
Random Variables and Probability



Today:

- Talk about a little bit of math;
- Introduce the ideas of random variables and probability;
- Discuss Bernoulli, Binomial, and Normal random variables;
- Compute these distributions in R.

Motivation – reasoning about voters...

- Cavalier Johnson: a political candidate running a campaign for Mayor of Milwaukee;
- Johnson's campaign claims that he will win the election – in this case that means getting more than 50% of votes from Milwaukee voters;
- How to assess the Johnson campaign claim? Conduct an experiment:
  - **Sample** n voters from Milwaukee;
  - Record the number of voters who will vote for Johnson.

A little bit of math...

- **Sum** – shorthand to write adding up a bunch of stuff:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5;$$

A little bit of math...

- **Sum** – shorthand to write adding up a bunch of stuff:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5;$$

- **Function** – a relationship between variables with a **dependent variable** that depends on an **independent variable(s)**;
  - We write the dependent variable as f, y, or p (but not always!), the independent variable as x, and the relationship between them as $f(x)$;
  - Example: $f(x) = 2x + 3$;
  - Example: $f(x, z) = 2x + 3z$;

A little bit of math...

- **Sum** – shorthand to write adding up a bunch of stuff:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5;$$

- **Function** – a relationship between variables with a **dependent variable** that depends on an **independent variable(s)**;
  - We write the dependent variable as f, y, or p (but not always!), the independent variable as x, and the relationship between them as $f(x)$;
  - Example: $f(x) = 2x + 3$;
  - Example: $f(x, z) = 2x + 3z$;
  - Example: $y(x_1, x_2, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$;

A little bit of math...

- **Sum** – shorthand to write adding up a bunch of stuff:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5;$$

- **Function** – a relationship between variables with a **dependent variable** that depends on an **independent variable(s)**;
  - We write the dependent variable as f, y, or p (but not always!), the independent variable as x, and the relationship between them as $f(x)$;
  - Example: $f(x) = 2x + 3$;
  - Example: $f(x, z) = 2x + 3z$;
  - Example: $y(x_1, x_2, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$;
  - Example – from our classes on causal inference: $Y_i(T)$;
  - Example – from our classes on causal inference: $Y_i(T, C, C, T, C, T, C)$;

A little bit of math...

- **Sum** – shorthand to write adding up a bunch of stuff:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5;$$

- **Function** – a relationship between variables with a **dependent variable** that depends on an **independent variable(s)**;
  - We write the dependent variable as f, y, or p (but not always!), the independent variable as x, and the relationship between them as $f(x)$;
  - Example: $f(x) = 2x + 3$;
  - Example: $f(x, z) = 2x + 3z$;
  - Example: $y(x_1, x_2, \beta_0, \beta_1, \beta_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$;
  - Example – from our classes on causal inference: $Y_i(T)$;
  - Example – from our classes on causal inference: $Y_i(T, C, C, T, C, T, C)$;

- **Exponentiation** – we write $\exp\{x\}$ and this is shorthand for 2.718^x.

What are random variables?

- **Sample space** for an experiment or random trial = all possible outcomes or results of that experiment;

- **Random variable**:
  - A variable whose values depend on outcomes of a random phenomenon;
  - A variable whose values have a relative likelihood governed by a probability distribution;

- **Probability distribution**: a mathematical function that gives the probabilities that the different possible outcomes occur. Must satisfy three rules:
  1. Probabilities are > 0;
  2. If you add up the probabilities for all possible outcomes they must equal 1;
  3. To find the probability of two mutually exclusive events happening you add up their individual probabilities.

What are random variables: examples.

- The outcome of a FAIR coin toss:
  - Sample space is: $\{\text{Heads}, \text{Tails}\}$;
  - Probability distribution is: $\{0.5, 0.5\}$;

- The outcome of rolling a six sided dice:
  - Sample space is: $\{1, 2, 3, 4, 5, 6\}$;
  - Probability distribution is: $\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$;

- The number of heads you get if you toss a coin 10 times:
  - Sample space is: $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$;
  - Probability distribution is?

- The number of voters who will vote for Johnson:
  - If N people live in Milwaukee the sample space is: $\{0, 1, 2, 3, \ldots, N\}$;
  - Probability distribution is?

Bernoulli Random Variables

- A **Bernoulli random variable** is used to model the outcome of an experiment that can be answered as a **boolean** – typical answers could be:
  - Yes or No;
  - Success or Failure;
  - True or False;
  - 1 or 0;
- Has a probability distribution with a single parameter: π, the probability of Yes/Success/True/1;
- Let's take the outcomes to be 0 and 1. Then we can write this as a function:

$$p(k = 1, \pi) =$$

Bernoulli Random Variables

- A **Bernoulli random variable** is used to model the outcome of an experiment that can be answered as a **boolean** – typical answers could be:
  - Yes or No;
  - Success or Failure;
  - True or False;
  - 1 or 0;

- Has a probability distribution with a single parameter: π, the probability of Yes/Success/True/1;

- Let's take the outcomes to be 0 and 1. Then we can write this as a function:

$$p(k = 1, \pi) = \pi$$
$$p(k = 0, \pi) =$$

Bernoulli Random Variables

- A **Bernoulli random variable** is used to model the outcome of an experiment that can be answered as a **boolean** – typical answers could be:
  - Yes or No;
  - Success or Failure;
  - True or False;
  - 1 or 0;
- Has a probability distribution with a single parameter: π, the probability of Yes/Success/True/1;
- Let's take the outcomes to be 0 and 1. Then we can write this as a function:

$$
\begin{aligned}
p(k=1,\pi) &= \pi \\
p(k=0,\pi) &= 1-\pi \\
p(k,\pi) &=
\end{aligned}
$$

Bernoulli Random Variables

- A **Bernoulli random variable** is used to model the outcome of an experiment that can be answered as a **boolean** – typical answers could be:
  - Yes or No;
  - Success or Failure;
  - True or False;
  - 1 or 0;
- Has a probability distribution with a single parameter: π, the probability of Yes/Success/True/1;
- Let's take the outcomes to be 0 and 1. Then we can write this as a function:

$$p(k = 1, \pi) = \pi$$
$$p(k = 0, \pi) = 1 - \pi$$
$$\text{or: } p(k, \pi) = \pi^k (1 - \pi)^{1-k} \text{ for } k = 0 \text{ or } k = 1.$$

Binomial Random Variables

- A **binomial random variable** is used to model the sum of n Bernoulli experiments – how many successes after n tries?

- Has a probability distribution with three parameters:
  1. n, the number of tries;
  2. k, the number of successes;
  3. π, the probability of each success;

- If we had a single try we could have either $k = 1$ success or $k = 0$ successes:

$$p(k = 1, n = 1, \pi) = \pi$$
$$p(k = 0, n = 1, \pi) = 1 - \pi$$
$$\text{or: } p(k, n = 1, \pi) = \pi^k (1 - \pi)^{1-k}.$$

Binomial Random Variables

- A **binomial random variable** is used to model the sum of n Bernoulli experiments – how many successes after n tries?
- Has a probability distribution with three parameters:
  1. n, the number of tries;
  2. k, the number of successes;
  3. π, the probability of each success;
- If we had two tries we could have $k = 2, 1$ or 0 successes:

$$p(k = 2, n = 2, \pi) = \pi^2$$
$$p(k = 1, n = 2, \pi) = \pi(1 - \pi) + (1 - \pi)\pi$$
$$p(k = 0, n = 2, \pi) = (1 - \pi)^2$$
$$\text{or: } p(k, n = 2, \pi) \propto \pi^k (1 - \pi)^{2-k}.$$

Binomial Random Variables

- A **binomial random variable** is used to model the sum of n Bernoulli experiments – how many successes after n tries?

- Has a probability distribution with three parameters:
    1. n, the number of tries;
    2. k, the number of successes;
    3. π, the probability of each success;

- If we had three tries we could have $k = 3, 2, 1$ or 0 successes:

$$p(k, \pi) \propto \pi^k (1 - \pi)^{3-k}.$$

Binomial Random Variables

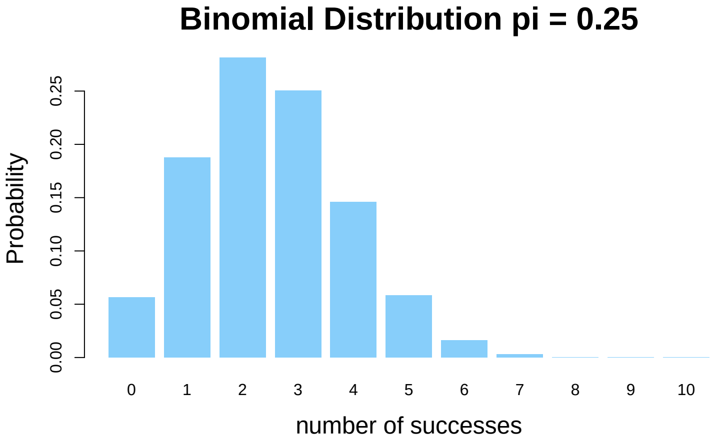
- A **binomial random variable** is used to model the sum of n Bernoulli experiments – how many successes after n tries?

- Has a probability distribution with three parameters:
  1. n, the number of tries;
  2. k, the number of successes;
  3. π, the probability of each success;

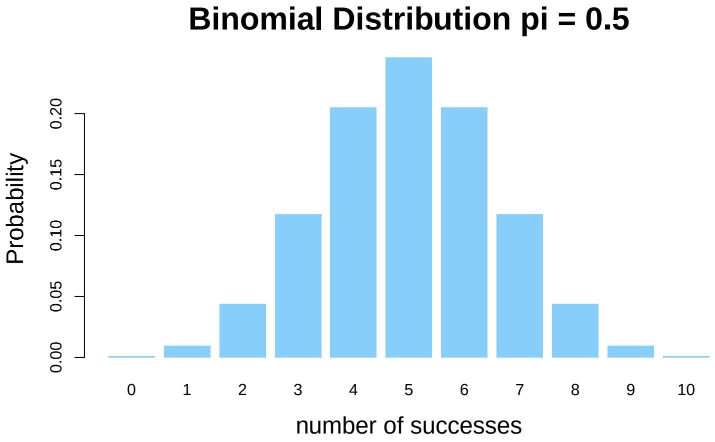
- If we had n tries we could have $k = n, \ldots, 3, 2, 1$ or 0 successes:

$$p(k, \pi) \propto \pi^k (1 - \pi)^{n-k}.$$

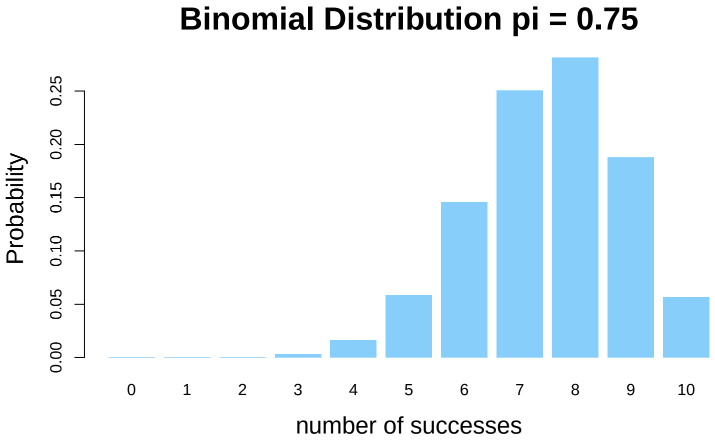
Binomial Random Variables



Binomial Random Variables



Binomial Random Variables



Normal Random Variables

- **Normal random variables** can be used to model just about any interval or ratio data (so continuous data) – why?

Normal Random Variables

- **Normal random variables** can be used to model just about any interval or ratio data (so continuous data) – why?

- **Reminder – the central limit theorem**: As sample size gets bigger the sampling distribution of a sample statistic increasingly follows a normal distribution and the standard deviation of that normal distribution gets smaller;

Normal Random Variables

- **Normal random variables** can be used to model just about any interval or ratio data (so continuous data) – why?

- **Reminder – the central limit theorem**: As sample size gets bigger the sampling distribution of a sample statistic increasingly follows a normal distribution and the standard deviation of that normal distribution gets smaller;

- Has a probability distribution with two parameters: μ, the mean and σ the standard deviation:

$$N(x, \mu, \sigma) = \frac{1}{\sqrt{2 * 3.1416 * \sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\};$$

Normal Random Variables

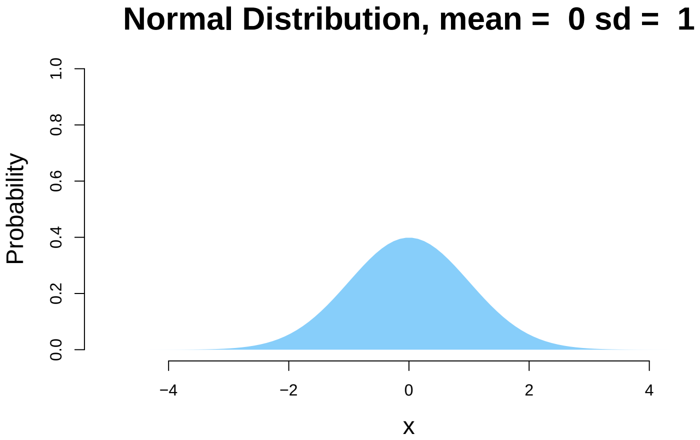
- **Normal random variables** can be used to model just about any interval or ratio data (so continuous data) – why?
- **Reminder – the central limit theorem**: As sample size gets bigger the sampling distribution of a sample statistic increasingly follows a normal distribution and the standard deviation of that normal distribution gets smaller;
- Has a probability distribution with two parameters: μ, the mean and σ the standard deviation:

$$N(x, \mu, \sigma) = \frac{1}{\sqrt{2 * 3.1416 * \sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\};$$

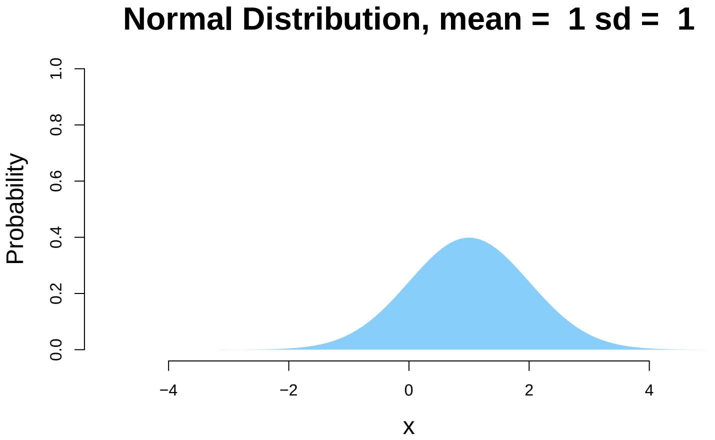
- When $\mu = 0$ and $\sigma = 1$ we call it the **standard normal**:

$$N(x, \mu = 0, \sigma = 1) = \frac{1}{\sqrt{2 * 3.1416}} \exp\left\{-\frac{x^2}{2}\right\}.$$

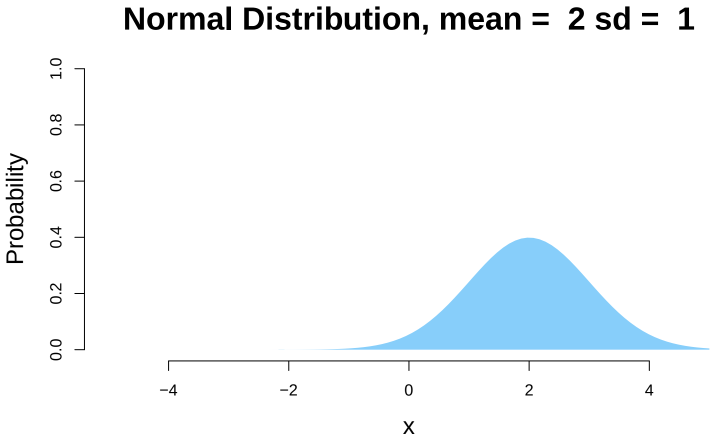
Normal – what happens as mean changes?



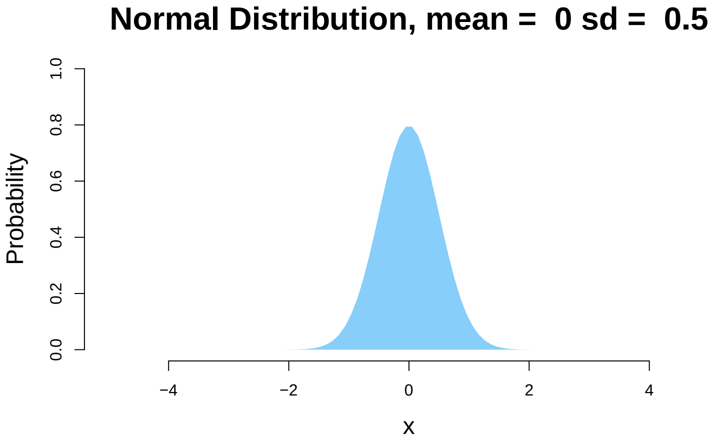
Normal – what happens as mean changes?



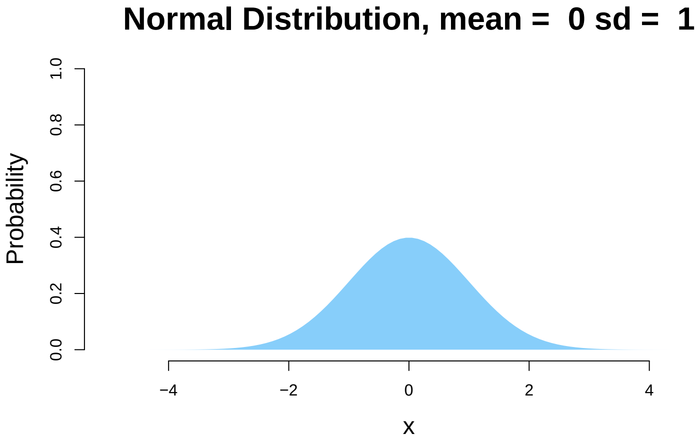
Normal – what happens as mean changes?



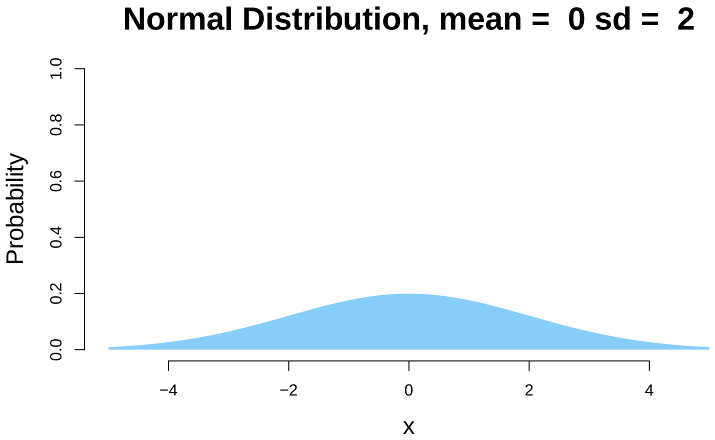
Normal – what happens as standard deviation changes?



Normal – what happens as standard deviation changes?



Normal – what happens as standard deviation changes?



Motivation – reasoning about voters...

- Cavalier Johnson: a political candidate running a campaign for Mayor of Milwaukee;
- Johnson's campaign claims that he will win the election – in this case that means getting more than 50% of votes from Milwaukee voters;
- How to assess the Johnson campaign claim? Conduct an experiment:
  - **Sample** n voters from Milwaukee;
  - Record the number of voters who will vote for Johnson.

If we see that k out of n voters say they will vote for Johnson, what is the probability his campaign's claim is correct?

Why should we care?

Probability is used for all hypothesis testing and statistical modeling.