Lesson 1. Introduction, Random Variables and Distributions

1 Introduction

- **Probability** is the study of random events
- **Statistics** is the study of how to collect, organize, analyze, and interpret numerical information from data
  - **Descriptive statistics** involves organizing, picturing, and summarizing information from samples or populations
  - **Inferential statistics** involves using information from a sample to draw conclusions regarding the population

Example 1. It is a common belief that the normal, healthy human temperature is 98.6°F.

When a temperature measurement is made, the measurement's value may vary from exactly 98.6°F.

A probability question: Suppose the temperature measurements are normally distributed with mean 98.6°F and standard deviation 1 °F.

A statistics question: Suppose we do not have any idea what the normal temperature of a healthy human is, but we observe three measurements of 98.5, 96.0, and 100.1.

- A goal of statistics is to turn data into useful information
- **Statistical conclusions are uncertain, but statisticians insist on quantifying the uncertainty**
- Probability is the mathematical tool used in this quantification
- In this course, we will...
  - learn how to use and assess statistical regression models
  - employ statistical software to implement and analyze these models
  - learn how to present statistical analysis in both a technical and non-technical format
  - learn about the limitations of statistical analysis
- But first, a brief probability review

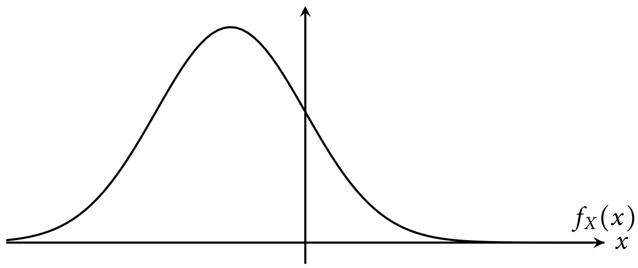
2 Random variables

- A **random variable** is a variable that takes on its values by chance
- Examples of random variables:
  - X = number of heads out of 10 coin flips
  - Y = temperature of a healthy human
- Notation conventions:
  - Uppercase letter (e.g., X, Y, Z) to denote a random variable
  - Lowercase letter (e.g., x, y, z) to denote an observation of a random variable (i.e., a data value)
- A random variable is **continuous** if it can take on a continuum of values

3 Distributions

- The **distribution** of a random variable is a mathematical description of how the observations of a random variable vary
- For a continuous random variable X, we can represent its distribution two ways
- The **cumulative distribution function (cdf)** $F_X(a)$ gives the probability that X is less than or equal to a
  - In other words,
- The **probability density function (pdf)** $f_X(a)$ gives the relative likelihood of X being near a
  - In particular,
- The cdf and pdf are related as follows:
- The cdf $F_X(a)$ answers the question: "How often does X have a value less than a?"
- What about the reverse question: "What value is X less than $(100 \times p)$% of the time?"
  - We can use the inverse of the cdf to answer this question
- The p-**quantile** $F_X^{-1}(p)$ is the value a such that $P(X \leq a) = p$
  - In other words,
- We will generally use cdfs for calculations and pdfs for visualization

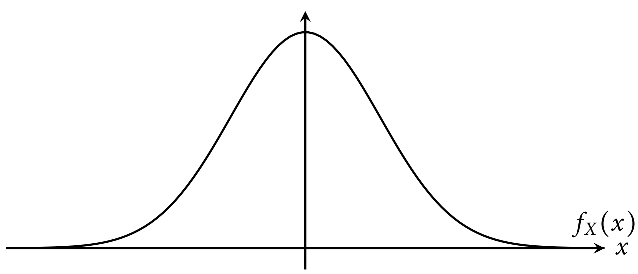
4 Calculating and visualizing probabilities with distributions



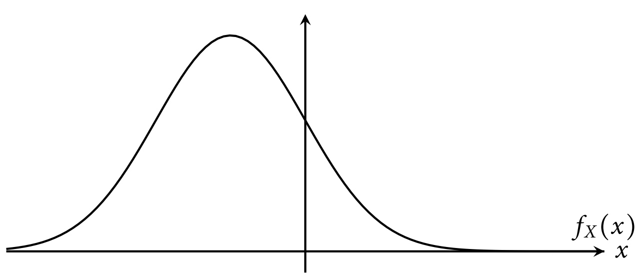
- Area under $f_X(x)$ from $x = -\infty$ to $x = \infty$:

- Therefore,

$$P(X \leq a) + P(X \geq a) =$$

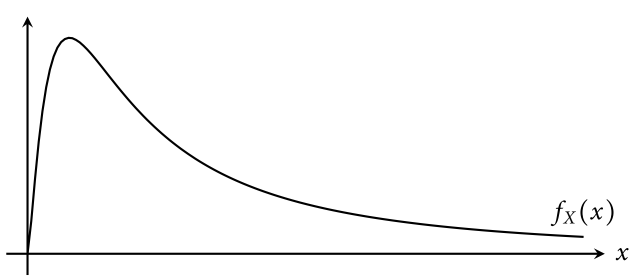


- If $f_X(x)$ is symmetric around 0 and $a > 0$:



- Using the area under $f_X(x)$ from $x = a$ to $x = b$:

$$P(a \leq X \leq b) =$$

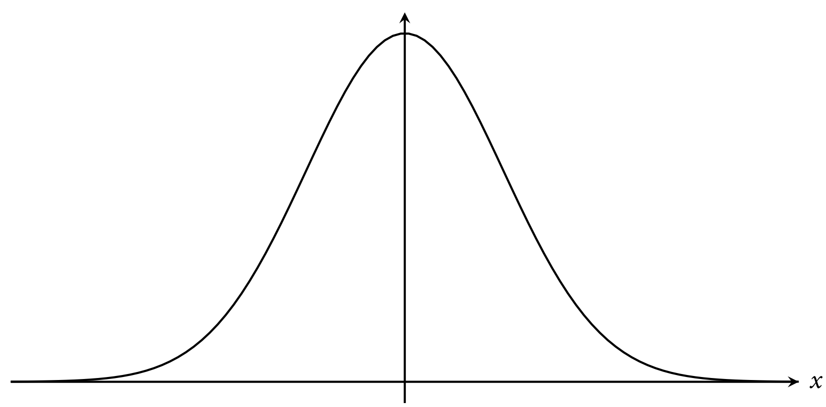


- Suppose the area under $f_X(x)$ from $x = -\infty$ to $x = a$ is p

5 Some families of distributions

- Let's brainstorm – what are some families of distributions you remember from SM239?

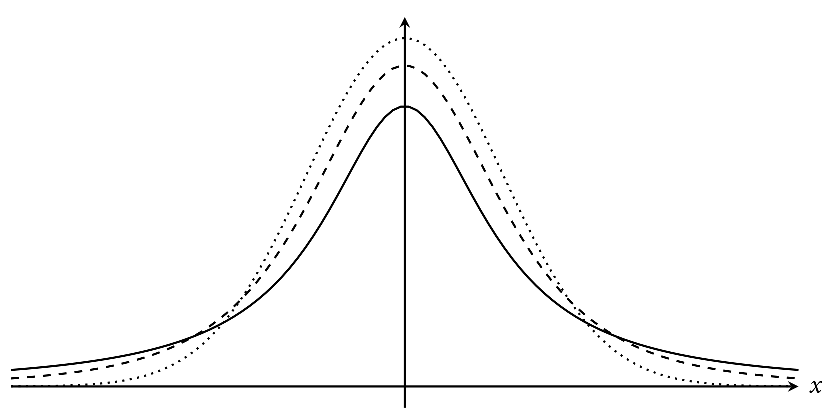
- The **normal distribution** $N(\mu, \sigma^2)$ with mean μ and variance σ^2
  - The pdf of $N(\mu, \sigma^2)$ is symmetric around μ and bell-shaped
  - The **standard normal distribution** $N(0, 1)$ is the special case with $\mu = 0$ and $\sigma^2 = 1$



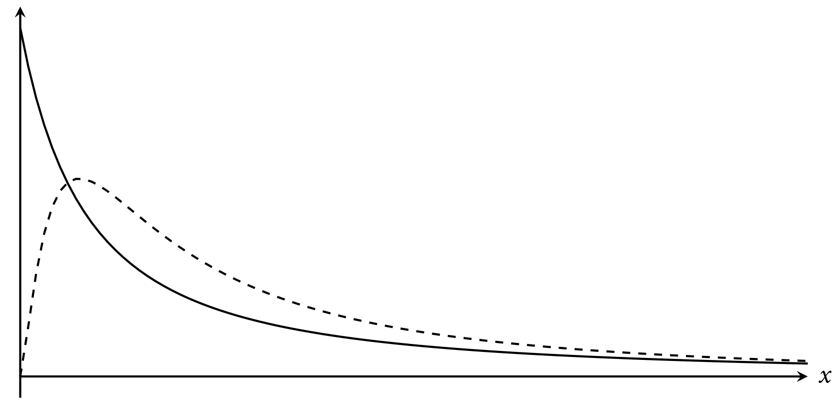
- The t-**distribution** $t(d)$ with d degrees of freedom
  - Like $N(0, 1)$, the pdf of $t(d)$ is symmetric around 0 and bell-shaped
  - Compared to $N(0, 1)$, the pdf of $t(d)$ has "heavier tails"

  $\Longrightarrow$

  - As d increases, $t(d)$ approaches $N(0, 1)$



- The F-**distribution** $F(d_1, d_2)$ with d_1 and d_2 degrees of freedom
    - Unlike $N(\mu, \sigma^2)$ and $t(d)$, $F(d_1, d_2)$ only takes positive values
    - The F-distribution is related to the t-distribution: $t(d)^2 \sim F(1, d)$



6 Exercises

Problem 1. Let X be a random variable that follows the t-distribution with 8 degrees of freedom. It turns out that $F_X(1.4) = 0.90$ and $F_X(-1.4) = 0.10$. Compute the following:

a. $P(X < 1.4)$
b. $P(X > 1.4)$
c. $P(X < -1.4)$
d. $P(-1.4 < X < 1.4)$