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

A probability question: Suppose the distribution of temperature measurements is Normal 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 and distributions

- A **random variable (r.v.)** is a *function* that maps an outcome to a real number
- Examples of r.v.s:
  - X = number of heads out of 10 coin flips
  - Y = temperature of a healthy human
- For this course, we will follow these notation conventions:
  - r.v. $\longleftrightarrow$ capital letter
  - observation (i.e., a data value; an "instance" or "realization" of a r.v.) $\longleftrightarrow$ lowercase letter
- The **distribution** of a r.v. is a mathematical description of how instances of a r.v. vary
  - The distribution tells us the probabilities of events (collections of outcomes) associated with the r.v.
- We can represent a distribution with a PMF/PDF or with a CDF
  - Cumulative distribution function (CDF):
  - Probability density function (PDF):
  - Calculating probabilities:
- We will generally use **CDFs for calculations** and **PDFs for visualization**

3 Some families of distributions

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

- The Normal distribution: Normal(μ, σ^2)

- The t-distribution with df "degrees of freedom": $t(df)$

  - Like Normal$(0, 1)$, the t-distribution is symmetric about 0 and bell-shaped
  - Compared to Normal$(0, 1)$, the t-distribution has "heavier tails"

    $\Rightarrow$

  - As df increases, the t-distribution approaches Normal$(0, 1)$

4 Exercises

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

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