CHAPTER 4: RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

- **random variable**: a variable that takes on values in some random fashion

- **$S$, the set of values the random variable takes on** is called the **support** (or range)

- **a discrete random variable** has a countable support; a countable set is either a finite set or a set that has a one-to-one correspondence with the natural numbers

- **a continuous random variable** has an interval support

- **the probability distribution** of a random variable describes how probabilities are assigned to values in the support

- **probability mass function (pmf) $f$** for a discrete rv: $f(x) = P(X = x)$

  - **cumulative distribution function $F$**: $F(x) = P(X \leq x)$

- **mean** (or expected value) of a discrete random variable: $\mu = E(X) = \sum_{x \in S} xf(x)$

- **the second moment** of a discrete rv: $E(X^2) = \sum_{x \in S} x^2 f(x)$

- **variance** of a discrete random variable: $\sigma^2 = \text{Var}(X) = \sum_{x \in S} (x - \mu)^2 f(x)$

  or equivalently, $\sigma^2 = E(X^2) - \mu^2$

- **standard deviation** of a rv: $\sigma = \sqrt{\sigma^2}$

- **binomial experiment** (or setting):
  1. a fixed number $n$ of trials
  2. two possible outcomes on each trial: “success” or “failure”
  3. the outcomes across the trials are mutually independent
  4. $p$, the probability of success, is the same (constant) for each trial

  The total number of successes in a binomial experiment is a discrete random variable $X$ which has pmf $f$ given by

  $$f(x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n$$

  Write $X \sim \text{binomial}(n, p)$ to denote a binomial random variable with parameters $n$ and $p$. For a binomial rv, $\mu = np$ and $\sigma^2 = np(1 - p)$

- Other prominent families of discrete random variables include the geometric, the negative binomial, the hypergeometric, the uniform, and the Poisson families
• probability distributions for continuous random variables

**probability density function (pdf)** $f$ with $f(x) \geq 0$ and $\int f(x)dx = 1$

areas over intervals represent probabilities: $P(a \leq X \leq b) = \int_a^b f(x)dx$

**mean** (or expected value): $\mu = E(X) = \int x f(x)dx$

**variance**: $\sigma^2 = \int (x - \mu)^2 f(x)dx$

• **normal distribution**

  – a normal random variable has a bell-shaped pdf:

  $$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

  – We write $X \sim N(\mu, \sigma)$

• **standard normal** rv: If $X \sim N(\mu, \sigma)$, then $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$.

• the **empirical rule** is based on probabilities associated with a standard normal rv:

  $$P(-1 < Z < 1) = .6826894809$$
  $$P(-2 < Z < 2) = .954499876$$
  $$P(-3 < Z < 3) = .9973000656$$

• steps for finding normal probabilities:
  1) sketch distribution and shade area
  2) convert boundaries of shaded area for $x$ values to $z$ values if table is to be used
  3) use standard normal table or TI-83, making use of symmetry if needed

• know how to find **percentiles** for a normal rv, that is, find the $x$-value that corresponds to a given probability

• assessing whether data are from (an approximate) normal distribution:
  1. histogram or stem plot should exhibit be roughly symmetric and mound-shaped
  2. use the empirical rule
  3. IQR/s $\approx 1.3$
  4. check linearity of a normal probability plot

• approximating binomial probabilities with normal probabilities
  rule of thumb: if $n \geq 9$ (odds for success) and $n \geq 9$ (odds for failure), then $P(r \leq X \leq s) \approx P(r - .5 \leq Y \leq s + .5)$,
  where $X \sim \text{binomial}(n, p)$ and $Y \sim N(\mu = np, \sigma = \sqrt{np(1-p)})$
Let denote a random sample from some population with mean $\mu$ and standard deviation $\sigma$. Note that the $X_i$'s are independent, identically distributed random (i.i.d.) variables.

the **sample mean**: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

the sample standard deviation: $S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$

the **sample variance**: $S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$

Note that (prior to observing the random variables $X_1, X_2, \ldots, X_n$) the sample mean $\overline{X}$ and the sample variance $S^2$ are themselves random variables. Hence it makes sense to talk about their expected values. The following facts are important.

$\mu_{\overline{X}} = E(\overline{X}) = \mu$
(The expected value of the sample mean equals the population mean.)

$E(S^2) = \sigma^2$  (The expected value of the sample variance equals the population variance.)

$\sigma^2_{\overline{X}} = Var(\overline{X}) = \frac{\sigma^2}{n}$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. (Also see technical note below.)

* Note. In the case where the population is of finite size $N$, the sample size $n$ is greater than $N/20$, and the sampling is done “without replacement”, then $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$.

**The Central Limit Theorem.** Let $X_1, X_2, \ldots, X_n$ denote a random sample from some population with mean $\mu$ and standard deviation $\sigma$. When $n$ is sufficiently large, the sample mean $\overline{X}$ will have an approximate normal distribution. Equivalently, the standardized random variable $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ will have an approximate standard normal distribution.

**Chebyshev's Inequality.** For any random variable $X$ with mean $\mu$ and standard deviation $\sigma$,

$$P(|X - \mu| < k\sigma) \geq 1 - k^{-2}$$

for any $k > 1$.

**(Weak) Law of Large Numbers.** Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with mean $\mu$ and standard deviation $\sigma$. Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\overline{X} - \mu| < \epsilon) = 1.$$