3.4 Functions of Random Variables

We study the mean and variance of a function of random variables, with special interest in the sample mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

One Random Variable, $X$ (continuous case; discrete is similar)

Multiplying by a constant $a$:

The mean of $aX$ is

$$\mu_{aX} = E(aX) = \int_{-\infty}^{\infty} (ax)f(x) \, dx = \int_{-\infty}^{\infty} [a^2(x - \mu_X)^2] f(x) \, dx = \sigma^2_{aX}$$

The variance of $aX$ is

$$\sigma^2_{aX} = E[(aX - a\mu_X)^2] = E[a^2(X - \mu_X)^2] = \int_{-\infty}^{\infty} [a^2(x - \mu_X)^2] f(x) \, dx$$

Adding a constant $a$:

The mean of $X + a$ is

$$\mu_{X+a} = E(X + a) = \int_{-\infty}^{\infty} (x + a)f(x) \, dx = \int_{-\infty}^{\infty} [(X + a) - (\mu_X + a)]^2 f(x) \, dx = \sigma^2_{X+a}$$

The variance of $X + a$ is

$$\sigma^2_{X+a} = E([X + a - (\mu_X + a)]^2) = \int_{-\infty}^{\infty} [a^2(x - \mu_X)^2] f(x) \, dx$$

Two Independent Random Variables, $X$ and $Y$ (discrete; continuous is similar)

$X$ and $Y$ are independent random variables $\iff$ for all sets of numbers $S$ and $T$,

$$P(X \in S \text{ and } Y \in T) = \cdots$$

(If we understand "$X \in S$" and "$Y \in T$" to be events, then this definition is like the one for independent events.) More generally, $X_1, \cdots, X_n$ are independent $\iff$ for all sets $S_1, \cdots, S_n$, $P(X_1 \in S_1 \text{ and } \cdots \text{ and } X_n \in S_n) = P(X_1 \in S_1) \cdots P(X_n \in S_n)$.
Mean of $X + Y$:

$$
\mu_{X+Y} = E(X + Y)
= \sum_x \sum_y (x + y)P(X = x, Y = y)
= \ldots
= \ldots
$$

Variance of $X + Y$:

$$
\sigma^2_{X+Y} = E \left[ (X + Y - \mu_{X+Y})^2 \right]
= \sum_x \sum_y [(x + y) - (\mu_X + \mu_Y)]^2 P(X = x, Y = y)
= \sum_x \sum_y [(x - \mu_X) + (y - \mu_Y)]^2 P(X = x)P(Y = y), \text{ by independence}
= \sum_x \sum_y [(x - \mu_X)^2 + 2(x - \mu_X)(y - \mu_Y) + (y - \mu_Y)^2] p(x)p(y), \text{ using } P(X = x) = p(x)
= \sum_x (x - \mu_X)^2 p(x) \sum_y p(y) + 2 \sum_x \sum_y (xy - \mu_Y x - \mu_X y + \mu_X \mu_Y) p(x)p(y) + \sum_x p(x) \sum_y (y - \mu_Y)^2 p(y)
= \sigma^2_X + 2 \left[ \sum_x xp(x) \sum_y yp(y) - \mu_Y \sum_x xp(x) \sum_y p(y) - \mu_X \sum_y yp(y) + \mu_X \mu_Y \sum_x p(x) \sum_y p(y) \right]
+ \sigma^2_Y
= \sigma^2_X + 2[\mu_X \mu_Y - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y] + \sigma^2_Y
= \sigma^2_X + \sigma^2_Y.
$$

Generalize to Many Independent Random Variables, $X_1, \ldots, X_n$

- The mean of $X_1 + \cdots + X_n$ is $\mu_{X_1+\cdots+X_n} = \ldots$
- The variance of $X_1 + \cdots + X_n$ is $\sigma^2_{X_1+\cdots+X_n} = \ldots$

Now we can handle any linear combination, $c_1X_1 + \cdots + c_nX_n$, of independent random variables.

e.g. (p. 114 #1 (c)) If $X$ and $Y$ are independent with $\mu_X = 9.5, \mu_Y = 6.8, \sigma_X = .4$, and $\sigma_Y = .1$, then find the mean and standard deviation of $X + 4Y$. 
Independence and Simple Random Samples

Before sampling, we can think of each item in a simple random sample as a ________________.

We suppose \( X_1, \cdots, X_n \) are ________________ if they’re from a simple random sample. (So, from §1.1, \( n \) is small compared to the population size \( N \): \( n < \______________ \).) They all have the same distribution as the population, so their distributions are the same: \( X_1, \cdots, X_n \) are independent and identically distributed (i.i.d.).

The Mean and Variance of a Sample Mean

Suppose \( X_1, \cdots, X_n \) are a simple random sample from a population with mean \( \mu \) and variance \( \sigma^2 \). Then the \( \{X_i\} \) are i.i.d., and, before sampling, \( \bar{X} \) is a ________________.

- The mean of \( \bar{X} \) is \( \mu_{\bar{X}} = \)

- The variance of \( \bar{X} \) is \( \sigma^2_{\bar{X}} = \)

- The standard deviation of \( \bar{X} \) is \( \sigma_{\bar{X}} = \)

To summarize, \( \mu_{\bar{X}} = \mu \) and \( \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \) (Don’t ________________.)

e.g. We use the average of four weighings on a lab scale, \( \bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4) \), where the \( X_i \)s are i.i.d. with standard deviation \( \sigma \), because \( \sigma_{\bar{X}} = \)

Standard Deviations of Nonlinear Functions of Random Variables

Propagation of Error

We want to estimate the standard deviation of a nonlinear function \( f \) of a random variable \( X \). Recall the Taylor series expansion of an infinitely differentiable function \( f(x) \) near \( x = a \):

\[
f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f^{(3)}(a)}{3!}(x - a)^3 + \cdots
\]

The first two terms serve as an approximation: \( f(x) \approx \______________ \) (draw)

So to estimate \( \sigma_{f(X)} \), we can use \( \sigma_{f(X)} \approx \______________ \). This is the propagation of error formula.

e.g. (p. 112 Example #3.29) Suppose the radius \( R \) of a circle is measured to be 5.43 cm, with \( \sigma_R = .01 \) cm. Estimate the area of the circle, \( A = \pi R^2 \), and estimate \( \sigma_A \).
The Taylor approximation can be \[ \frac{\partial f}{\partial x} \approx \frac{\partial f}{\partial y} \approx \frac{\partial f}{\partial z} \approx \] so use no more than two significant digits for \( \sigma_f(X) \).

**Review (or Preview) of Partial Derivatives**

The *partial derivative* of a function of several variables is its derivative with respect to \[ \frac{\partial f}{\partial x} \approx \frac{\partial f}{\partial y} \approx \frac{\partial f}{\partial z} \approx \] with the others \[ \frac{\partial f}{\partial y} = \frac{\partial f}{\partial z} = \] .

\[ f(x,y,z) = x^3 + x^2 y + x z^3 \implies \frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = \frac{\partial f}{\partial z} = \]

\[ V = \pi r^2 h \implies \frac{\partial V}{\partial r} = \frac{\partial V}{\partial h} = \]

**Multivariate Propagation of Error**

More generally, consider a multivariable function \( f(X_1, \cdots, X_n) \). If \( X_1, \cdots, X_n \) are independent random variables with small standard deviations \( \sigma_{X_1}, \cdots, \sigma_{X_n} \), then

\[
\sigma_{f(X_1, \cdots, X_n)} \approx \sqrt{\left( \frac{\partial f}{\partial X_1} \right)^2 \sigma_{X_1}^2 + \cdots + \left( \frac{\partial f}{\partial X_n} \right)^2 \sigma_{X_n}^2}
\]

where, in practice, we evaluate the partial derivatives at \( (X_1, \cdots, X_n) \).

This is the *multivariate propagation of error* formula. It can help decide which \[ \frac{\partial f}{\partial X_1} \approx \frac{\partial f}{\partial X_2} \approx \frac{\partial f}{\partial X_3} \approx \] are most responsible for random variation in a quantity calculated from several measurements.

\[ T = 1.203 PV \] (when \( P \) is measured in kilopascals, \( T \) is measured in kelvins, and \( V \) is measured in liters).

(a) Assume \( P \) is measured to be 242.52 kPa, with \( \sigma_P = .03 \) kPa, and \( V \) is measured to be 10.103 L with, \( \sigma_V = .002 \) L. Estimate \( T \) and \( \sigma_T \).

(b) Which would reduce \( \sigma_T \) more: reducing \( \sigma_P \) to .01 kPa or reducing \( \sigma_V \) to .001 L?