1. **Variance:** Suppose $X$ is a random variable on a probability space $(S, P)$ and with expected value $E(X) = m$. The variance of $X$ is the expected squared deviation from $m$, defined by

$$V(X) \equiv E([X - m]^2).$$

(Eq. 1)

Evaluating (1) over $S$, we see

$$V(X) = \sum_{w \in S} [X(w) - m]^2 P(w)$$

Evaluating (1) over Range($X$) we get

$$V(X) = \sum_{a_i \in \text{Range}(X)} (a_i - m)^2 P(X = a_i) = \sum_{a_i \in \text{Range}(X)} (a_i - m)^2 f_X(a_i).$$

- **Fact 1:** Another (possibly easier) way to evaluate variance is $V(X) = E(X^2) - m^2$. We get this from (1) by $E([X-m]^2) = E(X^2 - 2mX + m^2) = E(X^2) - 2mE(X) + m^2$, and the fact that $m = E(X)$.

- **Fact 2:** $V(aX + b) = a^2V(X)$. Think of multiplication by $a$ as a “scale change” and addition by $b$ as “shifting” the measurements implied by $X)$). Then - e.g. - doubling $X$ multiplies variance by 4; shifting does not effect variance (why?? this should be intuitive from (1)).

- **Fact 3:** Given two random variables $X$ and $Y$ defined on the same sample space $S$, the covariance of $X$ and $Y$ is defined by

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y).$$

If the covariance of $X$ and $Y$ is zero we say that $X$ and $Y$ are uncorrelated.

- **Fact 4:** If $X$ and $Z$ are independent they are uncorrelated (so $\text{cov}(X, Y) = 0$), but not conversely; as shown by this simple example: Let $\mathcal{E}$ be the experiment of tossing a fair coin twice (equally likely prob.), and taking $X$ = the number of Heads, $Y$ = the number of Tails, and $Z = (X - Y)^2$. Now check that $X$ and $Z$ are uncorrelated. They are clearly not independent because $Z = (X - Y)^2 = (X - (2 - X))^2 = (2X - 2)^2$ is a function of $X$ - if I tell you $X$, you know $Z$.

- **Fact 5:** The variance of a sum satisfies

$$V(X + Z) = V(X) + V(Z) + 2[E(XZ) - E(X)E(Z)] = V(X) + V(Y) + 2\text{cov}(X, Y).$$

By Fact 3, $V(X + Z) = V(X) + V(Z)$ for independent random variables (but that equation does not imply independence). By induction, if $X_1, \ldots, X_n$ are pairwise independent,

$$V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n).$$

(Eq. 2)
• Fact 5 (variance of the geometric r.v.): Let $W_1$ be the wait for the first success in Bernoulli trials with success probability $P$. Then $V(W_1) = (1 - P)/P^2$. This was proved by first showing

$$
\sum_{n=1}^{\infty} n(n - 1)P(W_1 = n) = \sum_{n=1}^{\infty} n(n - 1)P(1 - P)^{n-1} = \frac{2(1 - P)}{P^2}.
$$

This sum is easily seen to be $E(W_1^2) - E(W_1)$. Since $V(W_1) = E(W_1^2) - [E(W_1)]^2$ we have

$$
V(W_1) = \frac{2(1 - P)}{P^2} + \frac{1}{P} - \frac{1}{P^2},
$$

using $E(W_1) = 1/P$.

• Fact 6 (variance of the negative binomial r.v.): Let $W_k$ be the wait for the $k$th success in Bernoulli trials with success probability $P$. Then $V(W_k) = k(1 - P)/P^2$. This implies the identity

$$
\sum_{n=k}^{\infty} (n-k/P)^2 P(W_k = n) = P^k \sum_{n=k}^{\infty} (n-k/P)^2 \binom{n-1}{k-1} (1-P)^{n-k} = \frac{k(1 - P)}{P^2}.
$$

The proof is probabilistic: We use the fact that $W_k = X_1 + \cdots + X_k$, where $X_1$ is the wait for the first success and $X_{i+1}$ is the wait for the first success after the $i$-th; each $X_i$ is geometric (so $V(X_i) = (1 - P)/P^2$) and they are independent so by (2), the variance of $W_k$ is $k(1 - P)/P^2$.

• Fact 7 (variance of the binomial r.v.): Let $S_n$ be the number of successes in $n$ Bernoulli trials with success probability $P$. Then $V(S_n) = nP(1-P)$. This implies the identity

$$
\sum_{k=0}^{n} (k-nP)^2 P(S_n = k) = \sum_{k=0}^{n} (k-nP)^2 \binom{n}{k} P^k(1-P)^{n-k} = nP(1-P)
$$

and is proved using indicators: $S_n = X_1 + \cdots + X_n$ where $X_i$, the indicator (of success) for the $i$th trial, has $V(X_i) = P(1-P)$ and by (2), $V(S_n)$ is $nP(1-P)$.

2. Variance of an Average: Let $X$ be a random variable on the sample space $(S, P)$ of an experiment $\mathcal{E}$. Write $m = E(X)$ for the mean and $\sigma^2 = V(X)$ for the variance of $X$. $\mathcal{E}$ is performed independently $n$ times and $X_i$ is the value of $X$ on the $i$th trial (note that $E(X_i) = m$ and $V(X_i) = \sigma^2$). Let

$$
A_n = \frac{X_1 + \cdots + X_n}{n}
$$

denote the average of the $n$ observed values of $X$. Clearly

$$
E(A_n) = m \text{ and } V(A_n) = \frac{\sigma^2}{n}.
$$

We observe that the variance $V(A_n) \to 0$ as $n \to \infty$, and this suggests that $A_n$ is a random variable that converges (in some sense) to its mean $m$. This is the content of the important Law of Large Numbers. This observation is formalized by using the next result.
3. **Tchebycheff’s Inequality** Let $X$ be a random variable on $(S, P)$ with mean $E(X) = m$, variance $V(X) = \sigma^2$, and frequency function $f_x$, and let $\epsilon > 0$ be any constant. Then

$$P(|X - m| \geq \epsilon) \leq \frac{V(X)}{\epsilon^2}. \quad (4)$$

This gives a quantitative sense to the observations that

- small variance implies that values of $X$ far from the mean are unlikely and
- if it is likely that $X$ has values that are far from the mean, then the variance must be large.

The proof uses the fact that $\text{Range}(X)$ is the union of $B = \{a_i : |a_i - m| \geq \epsilon\}$ and $B^c = \{a_i : |a_i - m| < \epsilon\}$. By definition $((a_i - m)/\epsilon)^2 \geq 1$ for $a_i \in B$. Therefore since $f_X(a_i) = P(X = a_i)$,

$$P(|X - m| \geq \epsilon) = P(B) = \sum_{a_i \in B} f_X(a_i) \leq \sum_{a_i \in B} \frac{(a_i - m)^2}{\epsilon^2} f_X(a_i)$$

$$\leq \sum_{a_i \in \text{Range}(X)} \frac{(a_i - m)^2}{\epsilon^2} f_X(a_i) = \frac{V(X)}{\epsilon^2}. \quad (5)$$

4. **(*) Law of Large Numbers** Let $\epsilon > 0$ be given. Apply (4) to $X = A_n$ and use (3) to see

$$\text{Prob}(|A_n - m| \geq \epsilon) \leq \frac{V(A_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \quad (5)$$

or, subtracting both sides of (5) from 1,

$$\text{Prob}(|A_n - m| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2} \uparrow 1.$$  

Thus, the random variable $A_n$ (the average of $n$ observations of $X$), converges to $m$ (the expected value of $X$).

An interesting special case is when $X = I_B$ is the indicator of an event $B \subseteq S$ which has probability $P(B)$. Then $X$ has expected value $m = P(B)$ and variance $\sigma^2 = P(B)[1 - P(B)]$. Also writing

$$X_i = \begin{cases} 
1 & \text{if } B \text{ occurs on the } i^{th} \text{ trial} \\
0 & \text{otherwise}
\end{cases}$$

for the value of $X$ on the $i^{th}$ trial,

$$A_n = \frac{X_1 + \cdots + X_n}{n} \rightarrow P(B);$$

in fact by (5),

$$\text{Prob}(|A_n - P(B)| \geq \epsilon) \leq \frac{P(B)[1 - P(B)]}{n\epsilon^2}. \quad (*)$$
Thus, the fraction of the $n$ repetitions in which $B$ occurs (the relative frequency of $B$) converges to the probability of $B$.

The relation expressed in (*) allows us to test the value we assigned to $P(B)$ by comparing it to the observed relative frequency of $B$ in $n$ trials. For example suppose a die is tossed $n = 600$ times and that the event $B = \{\text{the die is a one}\}$ occurred on 150 of the trials. Assuming the die to be fair, $P(B) = 1/6$. We are told that $A_n = 150/600$, so $\varepsilon = 1/4 - 1/6 = 1/12$ in (*), and the right-hand side of (*) evaluates to $1/30$. Equation (*) says that if $P(B)$ really equals $1/6$, such a large number (150) of occurrences of $B$ in $n = 600$ tosses would only happen with probability less than $1/30$. We may in fact have seen this unlikely event, but it is easier to believe that the die is biased in favor of showing a one (i.e., $P(B) > 1/6$).

In fact we will say more: The inequality (*) is equivalent to

$$\text{Prob}(\lvert A_n - P(B) \rvert < \varepsilon) \geq 1 - \frac{P(B)[1 - P(B)]}{n\varepsilon^2}.$$  

The right hand side is interpreted as the confidence that $P(B)$ is closer to the observed value of $A_n$ than $\varepsilon = 1/12$: in our example we are $1 - 1/30 = 29/30 = 96\frac{2}{3}\%$ confident that the die is biased in favor of a 1.

___________________________________________________________

[WE WILL NOT COVER THE REMAINING TOPICS THIS SEMESTER, though you are allowed to read through, if you wish]

5. **Generating Functions** Let $a_0, a_1, \ldots$ (or briefly $\{a_i\}$) denote an infinite sequence of real numbers. Its generating function is defined by

$$A(s) = \sum_{k=0}^{\infty} a_k s^k = a_0 + a_1 s + \cdots + a_k s^k + \cdots$$  

(6) For example

$$A(s) = \frac{1}{1 - s/2} = \sum_{k=0}^{\infty} \frac{s^k}{2^k}$$

is the generating function of $\{1, 1/2, 1/4, \ldots\}$, the sequence of powers of $1/2$. Generating functions take a discrete object (a sequence of numbers) and give back a continuous function on which calculus may be used. Application of continuous tools is very important in discrete mathematics. Generating functions are one such example.

- **Fact 1:** $A(0) = a_0$ and $A(1) = \sum_{k=0}^{\infty} a_k$, the first element of the sequence and the sum of the elements, respectively (just make the substitutions in (8)).

- **Fact 2:** $A'(1) = \sum_{k=1}^{\infty} k a_k s^{k-1}|_{s=1} = \sum_{k=1}^{\infty} k a_k$ (differentiate each term of the sum in (6) and substitute).
6. **Convolutions** Let \( A(s) = \sum_{k=0}^{\infty} a_k s^k \) and \( B(s) = \sum_{k=0}^{\infty} b_k s^k \) be the generating functions of the sequences \( \{a_i\} \) and \( \{b_i\} \), respectively. If you multiply \( A(s) \) and \( B(s) \) and collect terms with the same power of \( s \), you get
\[
A(s)B(s) = a_0 b_0 + (a_0 b_1 + a_1 b_0)s + (a_0 b_2 + a_1 b_1 + a_2 b_0)s^2 + \cdots + (a_0 b_k + \cdots + a_k b_0)s^k + \cdots.
\]
Observe that \( A(s)B(s) \) is a generating function \( C(s) = \sum_{k=0}^{\infty} c_k s^k \) of the sequence \( \{c_i\} \) whose elements are defined by
\[
c_k = a_0 b_k + a_1 b_{k-1} + \cdots + a_{k-1} b_1 + a_k b_0.
\] (7)
This procedure of using (7) to create a new sequence \( \{c_i\} \) from two given sequences \( \{a_i\} \) and \( \{b_i\} \) is called convolution. We say \( \{c_i\} \) is the convolution of \( \{a_i\} \) and \( \{b_i\} \) and we write
\[
\{c_i\} = \{a_i\} \ast \{b_i\}.
\]
The generating function \( C(s) \) of the convolution of two sequences is the product \( A(s)B(s) \) of their generating functions.

6. **Counting Binary Trees**: We will discuss two important applications that illustrate the power of generating functions in discrete problems. The first is to count binary trees. Let \( B_n \) denote the set of rooted binary trees with \( n \) nodes, and let \( b_n \) denote \( |B_n| \), the size of \( B_n \). We have seen that \( b_1 = 1, b_2 = 2, b_3 = 5, \) and \( b_4 = 14, \) etc., and agreed to take \( b_0 = 1 \) (for the empty tree). We also derived the fact that
\[
b_n = b_0 b_{n-1} + b_1 b_{n-2} + \cdots + b_{n-2} b_1 + b_{n-1} b_0,
\] (8)
the term \( b_k b_{n-k-1} \) counting binary trees with \( k \) nodes in the left subtree. We will (I) find the generating function \( B(s) = \sum_{i=0}^{\infty} b_i s^i \) of the sequence \( \{b_i\} \) and (II) compute the coefficient of \( s^n \), namely \( b_n \).

(I) Multiply equation (8) above by \( s^n \) and sum (on both sides of =) from \( n = 1 \) to obtain
\[
\sum_{n=1}^{\infty} b_n s^n = \sum_{n=1}^{\infty} (b_0 b_{n-1} + \cdots + b_{n-2} b_1 + b_{n-1} b_0) s^n = s \sum_{n=1}^{\infty} c_{n-1} s^{n-1},
\] (9)
where in the last sum we write
\[
c_{n-1} = b_0 b_{n-1} + \cdots + b_{n-1} b_0.
\]
Observe (see (7)) that \( c_{n-1} \) above is the \( (n-1)^{st} \) term of the convolution \( \{b_i\} \ast \{b_i\} \), so that \( C(s) = B(s)B(s) \), and we see from (9) that
\[
B(s) - 1 = sC(s) = s (B(s))^2,
\]
the minus 1, because the left hand sum in (9) is \( B(s) \), except the \( n = 0 \) term is missing, and \( b_0 = 1 \). Rearranging terms we get
\[
s(B(s))^2 - B(s) + 1 = 0
\] (10)
a quadratic equation in \( B(s) \). Solving for \( B(s) \) gives
\[
B(s) = \frac{1 \pm \sqrt{1 - 4s}}{2s},
\]
and we reject the positive root because it makes the right side infinite at \( s = 0 \).

(II) Using Newton’s generalized Binomial theorem we see that
\[
(1 - 4s)^{1/2} = \sum_{j=0}^{\infty} (-4s)^j \binom{1/2}{j} = 1 + \sum_{j=1}^{\infty} (-4s)^j \frac{1/2}{j}
\]
and therefore, that
\[
B(s) = \frac{1 - (1 - 4s)^{1/2}}{2s} = -\frac{\sum_{j=1}^{\infty} (-4s)^j \frac{1/2}{j}}{2s} = -\frac{1}{2} \sum_{j=1}^{\infty} (-4)^j s^{j-1} \binom{1/2}{j}
\]
In this expansion \( s^n \) occurs in the \( j = n + 1 \) term, so that \( b_n \) (the coefficient of \( s^n \)) satisfies
\[
b_n = -\frac{(-4)^{n+1}}{2} \binom{1/2}{n+1} = -\frac{(-4)^{n+1}}{2} \left[ \frac{\left( \frac{1}{2} - 1 \right) \left( \frac{1}{2} - 2 \right) \cdots \left( \frac{1}{2} - n \right)}{(n+1)!} \right]
\]
which simplifies to
\[
b_n = \frac{1}{n+1} \left( \frac{2n}{n} \right)
\]
as the number of rooted binary trees with \( n \) nodes.

This was a nontrivial calculation, but not conceptually difficult. You might like to think about determining \( b_n \) without having the useful tool of generating functions.

7. **Generating Functions for Integer Random Variables:** The second important application of generating functions is in Probability. We begin with some basic ideas.

Let \( X \) be a random variable whose range is a subset of \( \{0, 1, \ldots\} \) and write \( p_i = f_X(i) = \text{Prob}(X = i) \) for its probabilities. We use this sequence of probabilities to define \( \phi_X \), the generating function of \( X \):
\[
\phi_X(s) = \sum_{k=0}^{\infty} p_k s^k = \sum_{k=0}^{\infty} \text{Prob}(X = k)s^k
\]
(13)
Note that this sum is an expectation, \( E(s^X) \). By Fact 1, \( \phi(0) = p_0 \) and \( \phi(1) = 1 \).

- **Fact 3: Mean and Variance:** Furthermore by Fact 2, \( \frac{\phi'(s)}{s|_{s=1}} = \sum_{k=1}^{\infty} kp_k = E(X) \). In fact if we differentiate (13) twice and evaluate at \( s = 1 \), we see
\[
\phi''(s)|_{s=1} = \sum_{k=1}^{\infty} k(k-1)p_k = \sum_{k=1}^{\infty} k^2 p_k - \sum_{k=1}^{\infty} kp_k = E(X^2) - E(X).
\]
Adding \( E(X) - [E(X)]^2 \) to both sides of the above equation we have
\[
\phi''(s)|_{s=1} + \phi'(s)|_{s=1} - (\phi'(s)|_{s=1})^2 = V(X).
\]
(14)
Example 1: Let \( X \) be the indicator of success in a B-trial with success probability \( P \). By (13) its generating function is
\[
\phi_X(s) = 1 - P + Ps.
\]
Use Fact 3 to see (again) that \( E(X) = \phi'(1) = P \) and that \( V(X) = P(1 - P) \).

Example 2: Let \( X \) be the score on a toss of a fair die. By (13) its generating function is
\[
\phi_X(s) = \sum_{k=0}^{\infty} \text{Prob}(X = k)s^k = \frac{s + s^2 + s^3 + s^4 + s^5 + s^6}{6}.
\]
Let \( Y \) be the score on a toss of a second fair die and \( Z = X + Y \). Using (13) and the probabilities for \( Z \), \( \phi_Z(s) = \sum_{k=0}^{\infty} \text{Prob}(Z = k)s^k \) satisfies
\[
\phi_Z(s) = \frac{s^2 + 2s^3 + 3s^4 + 4s^5 + 5s^6 + 6s^7 + 5s^8 + 4s^9 + 3s^{10} + 2s^{11} + s^{12}}{36}.
\]

Fact 4: Generating Functions for Independent Sums: Let \( X \) and \( Y \) be random variables with \( \text{Prob}(X = k) = a_k \) and \( \text{Prob}(Y = k) = b_k \) and let \( Z = X + Y \). Then
\[
\{Z = k\} = \bigcup_{i=0}^{k} (\{X = i\} \cap \{Y = k - i\})
\]
and if \( X \) and \( Y \) are independent, \( c_k = \text{Prob}(Z = k) \) satisfies
\[
c_k = \sum_{i=0}^{k} \text{Prob}(\{X = i\} \cap \{Y = k - i\})
\]
\[
= \sum_{i=0}^{k} \text{Prob}(X = i)\text{Prob}(Y = k - i) = \sum_{i=0}^{k} a_i b_{k-i};
\]
From (7), \( \{c_i\} \) is seen to be the convolution \( \{a_i\} * \{b_i\} \), so
\[
\phi_Z(s) = \phi_X(s)\phi_Y(s)
\]
for independent sums. This extends by induction to the sum \( Z = X_1 + \cdots + X_n \) of independent random variables giving
\[
\phi_Z(s) = \phi_{X_1}(s)\phi_{X_2}(s) \cdots \phi_{X_n}(s).
\]
You should check that \( \phi_Z(s) = (\phi_X(s))^2 \) in the previous Example 2 with dice (note \( Z = X + Y \) and \( \phi_X = \phi_Y \)).

These facts combine to give the generating functions for two familiar random variables.
(a) **Negative Binomial Generating Function:** Let $W_k$ be the number of Bernoulli trials needed for $k$ successes, with $\mathcal{P}$ denoting the success probability. First we take $k = 1$, so $W_1$ is the geometric random variable with \( \text{Prob}(W_1 = n) = \mathcal{P}(1 - \mathcal{P})^{n-1}, \ n = 1, 2, \ldots \). Using this in (13) we see

\[
\phi_{W_1}(s) = \frac{\mathcal{P}s}{1 - (1 - \mathcal{P})s}
\]

It is instructive to verify that \( \phi'_{W_1}(s)|_{s=1} = 1/\mathcal{P} \) and that (14) gives \( V(W_1) = (1 - \mathcal{P})/\mathcal{P}^2 \).

As usual, we write $W_k = X_1 + \cdots + X_k$, $X_1$ the number of trials needed for the first success and $X_{i+1}$ the number of trials after the $i^{th}$ success that are needed for the next success. Since the $X_i$ are independent geometrics, we use (15) inductively to obtain

\[
\phi_{W_k}(s) = (\phi_{W_1}(s))^k = \left( \frac{\mathcal{P}s}{1 - (1 - \mathcal{P})s} \right)^k
\]

and again, it is instructive to verify that $E(W_k) = k/\mathcal{P}$ and $V(W_k) = k(1 - \mathcal{P})/\mathcal{P}^2$.

(b) **Binomial Generating Function** Let $S_n$ be the number of successes in $n$ Bernoulli trials with success probability $\mathcal{P}$. Its generating function is

\[
\phi_{S_n}(s) = (1 - \mathcal{P} + \mathcal{P}s)^n.
\]

You can derive this: (A) by applying the binomial theorem to $\phi_{S_n}(s) = \sum_{i=0}^{n} \binom{n}{i} \mathcal{P}^i (1 - \mathcal{P})^{n-i} s^i$, or (B), by noting that $S_n = X_1 + \cdots + X_n$, $X_i$ the indicator of success on the $i^{th}$ trial, and using (14) along with the fact (Example 1) that $\phi_{X_i}(s) = (1 - \mathcal{P} + \mathcal{P}s)$. It is instructive to use (16) to compute the mean and variance of $S_n$. 