A Remark on the Notation for Means and Variances

The **mean**, or **expected value**, or **expectation** of a random variable $X$ can be denoted as

- $\mu_X$
- $\mu(X)$
- $E(X)$ (Here “$E$” means “expectation”)

The **variance** of a random variable $X$ can be denoted as

- $\sigma_X^2$
- $\sigma^2(X)$
- $\text{Var}(X)$

Four Rolls of a Die (1)

The two properties on the previous slide are very useful since you can find the mean and variance for $X_1 + X_2 + \cdots + X_n$ without knowing the distribution of $X_1, X_2, \ldots, X_n$.

**Example:** What is the mean and variance for the sum of the number of spots one gets when rolling a die 4 times?

**Approach 1**

- Let $S_4$ be the total number of spots in 4 rolls.
- **Possible values of $S$:** 4, 5, 6, ..., 23, 24
- **Distribution of $S_4$?**
  - e.g., $P(S_4 = 15) = ?$
  - How many ways are there to have a sum of 15 in 4 rolls?
  - $6^4 = 1296$ possible outcomes, too many to enumerate
- **Is there an easier way?**

### Many Rolls of a Die

The second approach can be easily generalized to more rolls. Consider the total number of spots $S_n$ got in $n$ rolls of a die, and let $X_i$ be the number of spots got in the $i$th roll, for $i = 1, 2, \ldots, n$. Then

$$S_n = X_1 + X_2 + \cdots + X_n$$

and all the $X_i$’s have a common distribution with mean 3.5 and variance 35/6. The mean and variance of $S_n$ are hence

- $E(S_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = 3.5 \times n$
- $\text{Var}(S_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) = \frac{35}{12} \times n$

since $X_i$’s are independent of each other.

The mean and variance $S_n$ can be found without first working out the distribution of $S_n$. 

### Four Rolls of a Die — Approach 2

Let $X_1, X_2, X_3$, and $X_4$ be respectively the number of spots in the 1st, 2nd, 3rd, and 4th roll.

- **Observe** that $S_4 = X_1 + X_2 + X_3 + X_4$
- $X_1, X_2, X_3$, and $X_4$ have a common distribution:
  - value | 1 | 2 | 3 | 4 | 5 | 6
  - probability | 1/6 | 2/6 | 3/6 | 4/6 | 5/6 | 6/6
- **mean:** $E(X_1) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{2}{6} + 3 \cdot \frac{3}{6} + 4 \cdot \frac{4}{6} + 5 \cdot \frac{5}{6} + 6 \cdot \frac{6}{6} = 3.5$
- $\text{Var}(X_1) = \frac{1}{6} \cdot \frac{1}{6} + \frac{2}{6} \cdot \frac{2}{6} + \frac{3}{6} \cdot \frac{3}{6} + \frac{4}{6} \cdot \frac{4}{6} + \frac{5}{6} \cdot \frac{5}{6} + \frac{6}{6} \cdot \frac{6}{6} - E(X_1)^2 = \frac{35}{36}$
- $X_2, X_3,$ and $X_4$ have the same mean and variance as $X_1$ since they have a common distribution
- So $E(S_4) = E(X_1) + E(X_2) + E(X_3) + E(X_4) = 3.5 + 3.5 + 3.5 + 3.5 = 14.$
- **Since** $X_1, X_2, X_3,$ and $X_4$ are independent, we have
  - $\text{Var}(S_4) = \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \text{Var}(X_4)$
  - $= \frac{35}{36} + \frac{35}{36} + \frac{35}{36} + \frac{35}{36} = 35/36.$
Sum and Mean of i.i.d. Random Variables

The rolling die example demonstrates a common scenario for many problems: suppose \(X_1, X_2, \ldots, X_n\) are i.i.d. random variables with mean \(\mu\) and variance \(\sigma^2\).

- Here, “i.i.d.” = “independent, and identically distributed”, which means that \(X_1, X_2, \ldots, X_n\) are independent and have identical probability distributions.

The mean and variance of \(S_n = X_1 + X_2 + \cdots + X_n\) are then

\[
\mathbb{E}(S_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) = \mu \times n = n\mu
\]

\[
\text{Var}(S_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) = \sigma^2 \times n = n\sigma^2
\]

- Observe \(\text{Var}(S_n) = n\sigma^2 \geq \text{Var}(X_i) = \sigma^2\), the sum of \(X_i\)’s has greater variability than a single \(X_i\) does.

Lecture 14&15 - 7

Properties of Correlation \(\rho\)

Let \(\rho\) be the correlation of random variables \(X\) and \(Y\). \(\rho\) has very similar properties with the sample correlation \(r\).

- \(-1 \leq \rho \leq 1\)
- If \(X\) and \(Y\) are independent, then \(\rho = 0\)
  (But when \(\rho = 0\), \(X\) and \(Y\) may not be independent.)
- If \(\rho > 0\) then when \(X\) gets big, \(Y\) also tends to get big, and vice versa. In this case,
  \(\text{Var}(X + Y) > \text{Var}(Y) + \text{Var}(X)\).
- If \(\rho < 0\) then when \(X\) increases, \(Y\) tends to decrease, and vice versa. In this case,
  \(\text{Var}(X + Y) < \text{Var}(Y) + \text{Var}(X)\).
- If \(\rho = 1\) or \(-1\), then there exists constants \(a\) and \(b\) such that \(Y\) always equals \(aX + b\).

Lecture 14&15 - 9

What if Not Independent?

In general, if \(X\) and \(Y\) are NOT independent, then

\[
\text{Var}(X + Y) = \text{Var}(Y) + \text{Var}(X) + 2\rho\sigma(X)\sigma(Y).
\]

Here, \(\rho\) is the correlation between \(X\) and \(Y\), which is defined analogously to the (sample) correlation \(r\).

\[
\text{sample correlation } r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)
\]

\[
\text{correlation } \rho = \mathbb{E} \left[ \left( \frac{X - \mu X}{\sigma X} \right) \left( \frac{Y - \mu Y}{\sigma Y} \right) \right]
\]

- We’ll NEVER compute \(\rho\) in STAT220.
  The formula is FYI only.

Correlation between two independent variables is zero.

Lecture 14&15 - 10

A Statistical Model of Simple Random Sampling

Consider a population comprised of \(N\) individuals, indexed from 1 to \(N\). Each individual has a numerical characteristic such that \(x_i\) is the numerical characteristic of the \(i\)th individual.

Example. The population is the 50,000 people age 25 and over in this town, indexed from 1 to \(N = 50,000\). Let \(x_i\) be the years of schooling of the \(i\)th individual in the population.

When a single individual is selected at random from the population (everyone has \(1/N\) chance to be selected), how many years of schooling \(X\) did he/she get?

- \(X\) is a random variable
- What is the probability distribution of \(X\)?

\[
p_x = P(X = x) = \frac{\text{# of people who have got } x \text{ years of schooling}}{N}
\]

Lecture 14&15 - 11

Review of Simple Random Samples

Suppose \(X_1, X_2, \ldots, X_n\) are \(n\) draws at random without replacement from a population of size \(N\). That is,

1. In the first draw, everyone has \(1/N\) chance to be selected
2. In the second draw, each of the remaining \(N - 1\) has \(1/(N - 1)\) chance to be selected
3. 
4. In the \(n\)th draw, each of the remaining \(N - n + 1\) has \(1/(N - n + 1)\) chance to be selected

Then \(\{X_1, X_2, \ldots, X_n\}\) is called a simple random sample (SRS) of size \(n\).

Lecture 14&15 - 12
Properties of Simple Random Samples

1. Every \( X_i \) has the same probability distribution (the population distribution \( X \)).

2. The \( X_i \)'s are (nearly) independent
   - Since we usually sample without replacement, draws are not independent.
   - As long as the sample size \( n \) is small (< 10% relative to the population size \( N \), the dependencies among sampled values are small and are generally ignored.
   - When sampling from an infinite population (\( N = \infty \)), the \( X_i \)'s are independent.

Due to the reasons above, we often assume observations \( X_1, X_2, \ldots, X_n \) in a simple random sample are i.i.d. from some (population) distribution.

Lecture 14&15 - 13

Properties of the Sample Mean

So far we have shown that: the sample mean \( \bar{X}_n \) of i.i.d random variables with mean \( \mu \) and variance \( \sigma^2 \) has the following properties:

1. \( \mathbb{E}(\bar{X}_n) = \mu \). \( \bar{X}_n \) is an unbiased estimator for \( \mu \).
2. \( \text{Var}(\bar{X}_n) = \sigma^2/n \). The larger \( n \) is, the less variable \( \bar{X}_n \) is.
3. Weak Law of Large Numbers: As \( n \) gets large
   \[ \bar{X}_n \rightarrow \mu. \]
   Intuitively, this is clear from the mean and the variance of \( \bar{X}_n \); the “center” of the distribution \( \bar{X}_n \) is \( \mu \), and the “spread” around it becomes smaller and smaller as \( n \) grows.
4. The distribution of \( \bar{X}_n \), called the sampling distribution of the sample mean, depends on the distribution of \( X_i \).
   - hard to find in general, except for a few cases
   - When \( n \) is large, we have Central Limit Theorem!

Lecture 14&15 - 15

Central Limit Theorem (CLT)

Let \( X_1, X_2, \ldots \) be a sequence of i.i.d. random variables (discrete or continuous) with mean \( \mu \) and variance \( \sigma^2 \). Then, when \( n \) is large,

- the distribution of the sample mean
  \[ \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \]
  is approximately
  \[ N \left( \mu, \frac{\sigma^2}{n} \right). \]
- the distribution of the sum \( X_1 + X_2 + \cdots + X_n \) is approximately
  \[ N(n\mu, n\sigma^2). \]

Lecture 14&15 - 16

If \( X_i \)'s are i.i.d., with the distribution

<table>
<thead>
<tr>
<th>value</th>
<th>1</th>
<th>2</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Probability histogram for the distribution of \( X_1 \):

![Probability histogram for the distribution of \( X_1 \)](value of the sum of the draws)

If \( X_i \)'s are i.i.d., with the distribution

<table>
<thead>
<tr>
<th>value</th>
<th>1</th>
<th>2</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Probability histogram for the distribution of \( S_{25} = X_1 + \cdots + X_{25} \):

![Probability histogram for the distribution of \( S_{25} \)](value of the sum of the draws)

If \( X_i \)'s are i.i.d., with the distribution

<table>
<thead>
<tr>
<th>value</th>
<th>1</th>
<th>2</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Probability histogram for the distribution of \( S_{50} = X_1 + \cdots + X_{50} \):

![Probability histogram for the distribution of \( S_{50} \)](value of the sum of the draws)

If \( X_i \)'s are i.i.d., with the distribution

<table>
<thead>
<tr>
<th>value</th>
<th>1</th>
<th>2</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>1/3</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Probability histogram for the distribution of \( S_{100} = X_1 + \cdots + X_{100} \):

![Probability histogram for the distribution of \( S_{100} \)](value of the sum of the draws)
Example: For the years of schooling example, it is known that the population distribution has mean $\mu = 11.8$ and variance is $\sigma^2 = 12.96$. For a sample of size 400, by CLT, the sample mean $\overline{X}_n$ is approximately

$$N\left(11.8, \sqrt{\frac{12.96}{400}}\right) = N(11.8, 0.18).$$

- Find the probability that the sample mean is between 11.8 ± 0.36.

Lecture 14&15 - 19

Summary: Means and Sums of i.i.d. Random Variables

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean $\mu$ and variance $\sigma^2$.

Let $S_n = X_1 + X_2 + \cdots + X_n$ and $\overline{X}_n = S_n/n$ be respectively the sum and the sample mean of $X_1, X_2, \ldots, X_n$.

So far we have shown that $S_n$ and $\overline{X}_n$ have the following properties

<table>
<thead>
<tr>
<th></th>
<th>$S_n$</th>
<th>$\overline{X}_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>expected value</td>
<td>$\mathbb{E}(S_n) = n\mu$</td>
<td>$\mathbb{E}(\overline{X}_n) = \mu$.</td>
</tr>
<tr>
<td>variance</td>
<td>$\text{Var}(S_n) = n\sigma^2$</td>
<td>$\text{Var}(\overline{X}_n) = \sigma^2/n$</td>
</tr>
<tr>
<td>sampling distribution for small $n$</td>
<td>no general form</td>
<td>no general form</td>
</tr>
<tr>
<td>approximate sampling distribution for large $n$</td>
<td>$N(n\mu, \sqrt{n}\sigma)$</td>
<td>$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$</td>
</tr>
</tbody>
</table>

Lecture 14&15 - 21

Bernoulli Random Variables (1)

A random variable $X$ is said to be a Bernoulli random variable if it takes two values only: 0 and 1.

- $p = P(X = 1)$ is called the probability of success
- Then $P(X = 0)$ must be $1 - p$ since $X$ is either 0 or 1.
- So the distribution of a Bernoulli random variable with probability $p$ of success must be

<table>
<thead>
<tr>
<th>value of $X$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability</td>
<td>$1 - p$</td>
<td>$p$</td>
</tr>
</tbody>
</table>

- Mean and variance:

$$\mathbb{E}(X) = 0 \cdot (1 - p) + 1 \cdot p = p,$$

$$\text{Var}(X) = 0^2 \cdot P(X = 0) + 1^2 \cdot P(X = 1) - \mathbb{E}(X)^2$$

$$= 0 \cdot (1 - p) + 1 \cdot p - p^2 = p(1 - p)$$

Lecture 14&15 - 22

Bernoulli Random Variables (2)

Bernoulli distribution arises when a random phenomenon has only two possible outcomes, e.g.,

- heads or tails in one coin tossing: $X = 1$ if heads, $X = 0$ if tails
- success or failure in a trial: $X = 1$ if success, $X = 0$ if failure
- whether a product is defected: $X = 1$ if defected, $X = 0$ if not
- whether a person uses iPhone: $X = 1$ if yes, $X = 0$ if not

Lecture 14&15 - 23

Binomial Distribution (1)

A random variable $Y$ is said to have a Binomial distribution $B(n, p)$, denoted as $Y \sim B(n, p)$, if it is a sum of $n$ i.i.d. Bernoulli random variables, $X_1, X_2, \ldots, X_n$, with probability $p$ of success.

Binomial distribution arises when we count the number of “successes” in a series of $n$ independent “trials”, e.g.,

- number of heads when tossing a coin $n$ times (“success” = heads)
- # of defected items in a batch of size 1000 (“success” = defected)
- # of iPhone users in a SRS from a huge population (“success” = iPhone user)

Lecture 14&15 - 24
Mean and Variance of Binomial

Recall a Binomial random variable $Y \sim B(n, p)$ are sums of i.i.d. Bernoulli random variables $X_1, X_2, \ldots, X_n$, with probability $p$ of success. The mean and variance of $Y$ are thus

$$E(Y) = E(X_1) + E(X_2) + \cdots + E(X_n) = p + p + \cdots + p = np$$

$$\text{Var}(Y) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) = p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)$$

since $X_i$’s are i.i.d. with mean $p$ and variance $p(1-p)$.

What about the distribution of $Y$? E.g., What is $P(Y = 3)$?

Lecture 14&15 - 25

Binomial Formula

The distribution of a Binomial distribution $B(n, p)$ is given by the binomial formula. If $Y$ has the binomial distribution $B(n, p)$ with $n$ trials and probability $p$ of success per trial, the probability to have $k$ successes in $n$ trials, $P(Y = k)$, is given as

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n.$$ 

Why the binomial formula is true?
See the next slide for an example.

Lecture 14&15 - 27

Factorials and Binomial Coefficients

The notation $n!$, read $n$ factorial, is defined as

$$n! = 1 \times 2 \times 3 \times \cdots \times (n-1) \times n$$

e.g.,

$$1! = 1, \quad 3! = 1 \times 2 \times 3 = 6, \quad 4! = 1 \times 2 \times 3 \times 4 = 24.$$ 

By convention, $0! = 1$.

The binomial coefficient: 

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- which is the number of ways to choose $k$ items, regardless of order, from a total of $n$ distinct items
- $\binom{n}{k}$ is read as “$n$ choose $k$”.

e.g.,

$$\binom{4}{2} = \frac{4!}{2! \times 2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = 6, \quad \binom{4}{4} = \frac{4!}{4! \times 0!} = \frac{4!}{4! \times 1} = 1$$

Lecture 14&15 - 26

Why is the Binomial Formula True? (Optional)

Let $Y$ be the number of success in 4 independent trials, each with probability $p$ of success. So $Y \sim B(4, p)$.

- To get 2 successes $(Y = 2)$, there are 6 possible ways:

  SSFF, SFSS, SFSF, FSSF, FSFS, FFSS

  in which “SSFF” means success in the first two trials, but not the last two, and so on.

  As trials are independent, by the multiplication rule,

  $$P(\text{SSFF}) = P(S)P(S)P(F)P(F) = p^2(1-p)^2$$

  $$P(\text{SFSS}) = P(S)P(F)P(S)P(F) = p(1-p)p(1-p) = p^2(1-p)^2$$

  Observe all 6 ways occur with probability $p^2(1-p)^2$, because all have 2 successes and 2 failures

  So $P(Y = 2) = (\# \text{ of ways}) \times (\text{prob. of each way}) = 6 \cdot p^2(1-p)^2$

Lecture 14&15 - 28

Example

Four fair dice are rolled simultaneously, what is the chance to get (a) exactly 2 aces? (b) exactly 3 aces? (c) 2 or 3 aces?

- A trial is one roll of a die. A success is to get an ace.
- Probability of success $p = 1/6$
- number of trials $n = 4$ is fixed in advance
- Are the trials independent? Yes!
- So $Y$ # of aces got has a $B(4, 1/6)$ distribution

(a) $P(Y = 2) = \frac{4!}{2!2!} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^2 = \frac{25}{216}$

(b) $P(Y = 3) = \frac{4!}{3!1!} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^1 = \frac{5}{324}$

(c) $P(Y = 2 \text{ or } Y = 3) = P(Y = 2) + P(Y = 3) = \frac{25}{216} + \frac{5}{324} = 0.131$

Lecture 14&15 - 30
Requirements to be Binomial (1)
To be a Binomial random variable, check the following
1. the number of trials \( n \) must be fixed in advance,
2. \( p \) must be identical for all trials
3. trials must be independent

Q1: A SRS of 50 from all UC undergrads are asked whether or not he/she is usually irritable in the morning. \( X \) is the number who reply yes. Is \( X \) binomial?
- a trial: a randomly selected student reply yes or not
- prob. of success \( p \) = proportion of UC undergrads saying yes
- number of trials = 50
- Strictly speaking, NOT binomial, because trials are not independent
- Since the sample size 50 is only 1% of the population size (\( \approx 5000 \)), trials are nearly independent
- So \( X \) is approximately binomial, \( B(n = 50, p) \)

Lecture 14&15 - 31

Requirements to be Binomial (2)
Q2 John tosses a fair coin until a head appears. \( X \) is the count of the number of tosses that John makes. Is \( X \) binomial?
- one trial = one toss of the coin
- number of trials is not fixed
- NOT binomial
Q3 Most calls made at random by sample surveys don’t succeed in talking with a live person. Of calls to New York City, only \( 1/12 \) succeed. A survey calls 500 randomly selected numbers in New York City. \( X \) is the number that reach a live person. Is \( X \) binomial?
- one trial = a call that reach a live person
- number of trials \( n = 500 \)
- probability of success \( p = 1/12 \)
- Independent trials? Huge population, so (nearly) independent
- \( X \sim B(500, 1/12) \)

Lecture 14&15 - 32

CLT for Counts and Proportion
Let \( X_1, X_2, \ldots \) be a sequence of i.i.d. Bernoulli random variables with probability \( p \) of success. So \( X_i \) has mean \( \mu = p \) and variance \( \sigma^2 = p(1-p) \). Then
- The sum \( S_n = X_1 + X_2 + \cdots + X_n \) now is the count of \( X_i \)'s that take value “1”, and has a binomial distribution \( B(n, p) \).
  As \( n \) gets large, the distribution of \( S_n \) is approximately
  \[ N(n\mu, \sqrt{n}\sigma) = N(np, \sqrt{np(1-p)}) \].
- The sample mean \( \bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \) is just the proportion of \( X_i \)'s that take value “1.” As \( n \) gets large, the distribution of \( \bar{X}_n \) is approximately
  \[ N \left( \mu, \frac{\sigma}{\sqrt{n}} \right) = N \left( p, \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right) \].

Lecture 14&15 - 33

Example: Twitter Users
Suppose 20% of the internet users use Twitters. If a SRS of 2500 internet users are surveyed, what is the probability that the percentage of Twitter users in the sample is over 21%?