STATS 100A: Two or More Random Variables

Ying Nian Wu

Department of Statistics
University of California, Los Angeles

Some pictures are taken from the internet.
Credits belong to original authors.
Recall Example 2 in Part 1: Sample a random person from a population of 100 people, 50 males and 50 females. 30 males are taller than 6 ft, 10 females are taller than 6 ft.


Example 2: A male, B tall.

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>taller than 6 ft</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>shorter than 6 ft</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ P(A) = \frac{|A|}{|\Omega|} = \frac{50}{100} = 50\%. \]

\[ P(B) = \frac{|B|}{|\Omega|} = \frac{30 + 10}{100} = 40\%. \]

\[ P(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{30}{100} = 30\%. \]

Probability = population proportion.
Population proportion

Experiment $\rightarrow$ outcome $\rightarrow$ number

Example 2: $A$ male, $B$ tall.

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>taller than 6 ft</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>shorter than 6 ft</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{30}{40} = 75\%.$$  

Among tall people, what is the proportion of males?

$$P(B|A) = \frac{|A \cap B|}{|A|} = \frac{30}{50} = 60\%.$$  

Among males, what is the proportion of tall people?

Conditional probability $=$ proportion within sub-population.
Example 2: $X \in \{ \text{male (1), female (0)} \}$, $Y \in \{ \text{tall (1), short (0)} \}$.

$$p(x, y) = P(X = x, Y = y).$$

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>taller than 6 ft</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>shorter than 6 ft</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

$p(1, 1) = .3$, $p(1, 0) = .2$, $p(0, 1) = .1$, $p(0, 0) = .4$. 
Example 2: \( X \in \{ \text{male, female} \}, \ Y \in \{ \text{tall, short} \}. \)

\[
p(x, y) = P(X = x, Y = y).
\]

<table>
<thead>
<tr>
<th></th>
<th>male</th>
<th>female</th>
</tr>
</thead>
<tbody>
<tr>
<td>taller than 6 ft</td>
<td>30</td>
<td>10</td>
</tr>
<tr>
<td>shorter than 6 ft</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
P(X = x) = p_X(x) = \sum_y p(x, y).
\]

\[
P(Y = y) = p_Y(y) = \sum_x p(x, y).
\]
Example 2: $X \in \{ \text{male, female}\}$, $Y \in \{ \text{tall, short}\}$.

$$p(x, y) = P(X = x, Y = y).$$

$$P(X = x|Y = y) = p_{X|Y}(x|y) = p(x, y)/p(y).$$

$$P(Y = y|X = x) = p_{Y|X}(y|x) = p(x, y)/p(x).$$

Chain rule: $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$. 
Example 2: $X \in \{ \text{male, female} \}, \ Y \in \{ \text{tall, short} \}$.

\[
p(x, y) = P(X = x, Y = y).
\]

\[
p(y) = \sum_x p(x, y) = \sum_x p(x)p(y|x).
\]
Example 2: \( X \in \{ \text{male, female} \}, Y \in \{ \text{tall, short} \} \).

\[
p(x, y) = P(X = x, Y = y).
\]

\[
p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')}.
\]
Independence

\[ P(A|B) = P(A). \]
\[ P(A \cap B) = P(A)P(B). \]

\( X \in \{ \text{male, female} \}, \ Y \in \{ \text{college, not} \} \)

\[
\begin{array}{c|c|c|c}
\text{College degree} & \text{M} & \text{F} & \text{Total} \\
\hline
20 & 20 & & 40 \\
50 & 50 & & 100 \\
\hline
\end{array}
\]

\[
\begin{array}{c|c|c|c}
\text{No college degree} & \text{M} & \text{F} & \text{Total} \\
\hline
50 & 50 & & 100 \\
\hline
\end{array}
\]

\[ p(y|x) = p(y). \]
\[ p(x, y) = p(x)p(y|x) = p(x)p(y). \]
Recall Example 6: Rare disease example

1% of population has a rare disease.
A random person goes through a test.
If the person has disease, 90% chance test positive.
If the person does not have disease, 90% chance test negative.
If tested positive, what is the chance he or she has disease?

\[ P(D) = 1\% . \]
\[ P(+|D) = 90\% , \ P(−|N) = 90\% . \]
\[ P(D|+) =? \]
\[ X \in \{D, N\} . \ Y \in \{+, −\} . \]
Example 6: Rare disease example

\[ P(D|+) = \frac{9}{9+99} = \frac{1}{12}. \]

\[ p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x)p(y|x)}{\sum_{x'} p(x')p(y|x')} . \]

\[ p(x): \text{ prior belief. } p(x|y): \text{ posterior belief.} \]
Discrete distribution

$N$: number of people in population.
$N(x, y)$: number of people with eye color $x$ and hair color $y$.
$N(x) = \sum_y N(x, y)$: number of people with eye color $x$.
$N(y) = \sum_x N(x, y)$: number of people with hair color $y$. 
Joint and marginal

\[ p(x, y) = \frac{N(x, y)}{N}. \]

\[ p(x) = \frac{N(x)}{N} = \sum_{y} \frac{N(x, y)}{N} = \sum_{y} p(x, y). \]

\[ p(y) = \frac{N(y)}{N} = \sum_{x} \frac{N(x, y)}{N} = \sum_{x} p(x, y). \]
Conditional

\[
p(x|y) = \frac{N(x, y)}{N(y)} = \frac{N(x, y)/N}{N(y)/N} = \frac{p(x, y)}{p(y)}.
\]

\[
p(y|x) = \frac{N(x, y)}{N(x)} = \frac{N(x, y)/N}{N(x)/N} = \frac{p(x, y)}{p(x)}.
\]
Rules

Marginalization: \( p(y) = \sum_x p(x, y) \).
Conditioning: \( p(x | y) = p(x, y) / p(y) \).
Chain rule: \( p(x, y) = p(x)p(y | x) \).
Expectation

\[ E[h(X, Y)] = \sum_{x,y} h(x, y) p(x, y). \]

Population average or long run average.

\[
\frac{1}{N} \sum_{x,y} h(x, y) N(x, y) = \sum_{x,y} h(x, y) \frac{N(x, y)}{N} = \sum_{x,y} h(x, y) p(x, y) = E[h(X, Y)].
\]
Expectation

\[
\mathbb{E}(X) = \sum_{x,y} xp(x, y) = \sum_{x}x \sum_{y} p(x, y) = \sum_{x} xp(x).
\]

same for \( \mathbb{E}[h(X)] \).

\[
\text{Var}(h(X, Y)) = \mathbb{E}[(h(X, Y) - \mathbb{E}[h(X, Y)])^2].
\]
Two continuous random variables

\[ X = \text{height}, \quad Y = \text{weight}. \]

\[ f(x, y) = \frac{P(X \in (x, x + \Delta x), Y \in (y, y + \Delta y))}{\Delta x \Delta y} = \frac{N(x, y)/N}{\Delta x \Delta y}. \]

density = probability / size
Probability density function
density = prob / size

\[ f(x) = \frac{P(X \in (x, x + \Delta x))}{\Delta x} = \frac{N(x)/N}{\Delta x} = \frac{\sum y N(x, y)/N}{\Delta x} = \frac{\sum y f(x, y)\Delta x\Delta y}{\Delta x} = \int f(x, y)dy. \]

\[ f(y) = \int f(x, y)dx. \]
Joint and marginal densities

Sample points under the surface, collapse on the plane.
Conditional density

density = \frac{\text{prob}}{\text{size}}

\begin{align*}
f(y|x) &= \frac{P(Y \in (y, y + \Delta y) | X \in (x, x + \Delta x))}{\Delta y} \\
&= \frac{N(x, y)/N(x)}{\Delta y} = \frac{N(x, y)/N}{(N(x)/N)\Delta y} \\
&= \frac{f(x, y)\Delta x\Delta y}{f(x)\Delta x\Delta y} = \frac{f(x, y)}{f(x)}.
\end{align*}

\[ f(x|y) = \frac{f(x, y)}{f(y)}. \]
Conditional density
Rules

Marginalization: $f(y) = \int f(x, y) \, dx$.
Normalization (conditioning): $f(x \mid y) = f(x, y) / f(y)$.
Factorization (chain rule): $f(x, y) = f(x) f(y \mid x)$.
$f(y \mid x)$: prediction. $f(x \mid y)$: inference.
If \((X, Y) \sim p(x, y)\), then
\[
\mathbb{E}(h(X, Y)) = \sum_x \sum_y h(x, y)p(x, y).
\]

If \((X, Y) \sim f(x, y)\), then
\[
\mathbb{E}(h(X, Y)) = \int \int h(x, y)f(x, y)dxdy.
\]

\[
\text{Var}(h(X, Y)) = \mathbb{E}[(h(X, Y) - \mathbb{E}[h(X, Y)])^2].
\]
Population average or long run average of $h(X, Y)$.

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i, Y_i) = \frac{1}{n} \sum_{\text{cells}} h(x, y) n f(x, y) \Delta x \Delta y$$

$$\rightarrow \int \int h(x, y) f(x, y) dx dy.$$
Conditional expectation and variance

Recall $\mathbb{E}(Y) = \int y f(y) dy$.

$$h(x) = \mathbb{E}[Y | X = x] = \int y f(y | x) dy.$$  

Regression, prediction.

$$\text{Var}(Y | X = x) = \mathbb{E}[(Y - h(X))^2 | X = x] = \int (y - h(x))^2 f(y | x) dy.$$
Bivariate Normal

\[ X \sim N(0, 1), \]
\[ Y = \rho X + \epsilon; \ \epsilon \sim N(0, 1 - \rho^2), \ (|\rho| \leq 1). \]

\( \epsilon \) is independent of \( X \). Given \( X = x \), \( Y = \rho x + \epsilon \).
The distribution of points within a vertical slice at $x$.

$$\mathbb{E}(Y|X = x) = \mathbb{E}(\rho x + \epsilon) = \rho x.$$ 

Regression towards the mean ($\rho < 1$), e.g., son’s height given father’s height.

$$\text{Var}(Y|X = x) = \text{Var}(\rho x + \epsilon) = \text{Var}(\epsilon) = 1 - \rho^2.$$ 

$$[Y|X = x] \sim \text{N}(\rho x, 1 - \rho^2).$$
Bivariate Normal

\[ f(x, y) = f(x) f(y|x) \]

\[ = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{x^2}{2} \right) \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp \left( -\frac{(y-\rho x)^2}{2(1-\rho^2)} \right) \]

\[ = \frac{1}{2\pi \sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy) \right]. \]

symmetric in \((x, y)\)
Let $\mu_X = \mathbb{E}(X)$, $\mu_Y = \mathbb{E}(Y)$, we define the covariance

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

It is defined for both discrete and continuous random variables.
Covariance

\[(X_i, Y_i) \sim f(x, y), \ i = 1, \ldots, n.\]

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i; \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.
\]

\[
\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).
\]
Covariance

\[
\text{Cov}(X, Y) \equiv \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).
\]

I, III: \((X_i - \bar{X})(Y_i - \bar{Y}) > 0.\)

II, IV: \((X_i - \bar{X})(Y_i - \bar{Y}) < 0.\)
Covariance

\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]
\]
\[
= \mathbb{E}[XY - \mu_X Y - X \mu_Y + \mu_X \mu_Y]
\]
\[
= \mathbb{E}(XY) - \mu_X \mathbb{E}(Y) - \mu_Y \mathbb{E}(X) + \mu_X \mu_Y
\]
\[
= \mathbb{E}(XY) - \mu_X \mu_Y
\]
\[
= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).
\]

Clearly, \( \text{Cov}(X, X) = \text{Var}(X) \) and \( \text{Cov}(Y, Y) = \text{Var}(Y) \).
Linearity

\[
\text{Cov}(aX + b, cY + d) \\
= \mathbb{E}[(aX + b - \mathbb{E}(aX + b))(cY + d - \mathbb{E}(cY + d))] \\
= \mathbb{E}[a(X - \mathbb{E}(X))c(Y - \mathbb{E}(Y))] = ac\text{Cov}(X, Y).
\]

Covariance depends on units (meter/foot, kilogram/pound).

\[
\text{Cov}(X + Y, Z) = \mathbb{E}[(X + Y - \mathbb{E}(X + Y))(Z - \mathbb{E}(Z))] \\
= \mathbb{E}[(X - \mathbb{E}(X) + Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))] \\
= \mathbb{E}[(X - \mathbb{E}(X))(Z - \mathbb{E}(Z))] + \mathbb{E}[(Y - \mathbb{E}(Y))(Z - \mathbb{E}(Z))] \\
= \text{Cov}(X, Z) + \text{Cov}(Y, Z).
\]
Correlation

Standardize: \( X \rightarrow (X - \mu_X)/\sigma_X, Y \rightarrow (Y - \mu_Y)/\sigma_Y. \)

\[
\text{Cov} \left( \frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y} \right) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \text{Corr}(X, Y).
\]
Correlation

\[ \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}). \]

\[ \text{Var}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2; \quad \text{Var}(Y) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \]

\[ \text{Corr}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}. \]
Correlation

Centralize: \( \tilde{X}_i = X_i - \bar{X} \); \( \tilde{Y}_i = Y_i - \bar{Y} \).

\[
\text{Corr}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i}{\sqrt{\sum_{i=1}^{n} \tilde{X}_i^2} \sqrt{\sum_{i=1}^{n} \tilde{Y}_i^2}}.
\]
Centralize: $\tilde{X}_i = X_i - \bar{X}$; $\tilde{Y}_i = Y_i - \bar{Y}$.

$$\text{Corr}(X, Y) = \frac{\sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i}{\sqrt{\sum_{i=1}^{n} \tilde{X}_i^2} \sqrt{\sum_{i=1}^{n} \tilde{Y}_i^2}} = \frac{\langle X, Y \rangle}{|X| |Y|} = \cos \theta.$$
Correlation and regression

Strength of linear relationship:

\[
\frac{\|e\|^2}{\|Y\|^2} = \frac{\sum_i e_i^2}{\sum_i (Y_i - \bar{Y})^2} = \sin^2 \theta = 1 - \cos^2 \theta = 1 - \rho^2.
\]
Correlation and regression

Regression line:

\[ \hat{Y} - \bar{Y} = \beta_1 (X - \bar{X}). \]

\[ \hat{Y} = \beta_1 X + (\bar{Y} - \beta_1 \bar{X}) = \beta_1 X + \beta_0. \]
Correlation and regression

\[ \rho = 1 \]

\[ \rho = -1 \]
Correlation and regression

\[ \rho = 0.8 \]

\[ \rho = -0.8 \]

\[ \theta \]
Correlation and regression
Correlation and regression

![Scatter plots](image)

**Relationship between Height and Weight**

- Males Regression Line: $y = 224.50 + 5.96x$
- Females Regression Line: $y = 246.01 + 5.99x$
Independence

\[ P(A \cap B) = P(A)P(B). \]
\[ p(x, y) = p_X(x)p_Y(y); \quad p(y|x) = p_Y(y). \]
\[ f(x, y) = f_X(x)f_Y(y); \quad f(y|x) = f_Y(y). \]

\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) \\
= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p_X(x)p_Y(y) \\
= \sum_x (x - \mu_X)p_X(x) \sum_y (y - \mu_Y)p_Y(y) \\
= \left( \sum_x xp_X(x) - \mu_X \right) \left( \sum_y yp_Y(y) - \mu_Y \right) = 0.
\]
Conditional independence

Shared cause: [siblings | parent]

\[
p(x, y | z) = p(x | z)p(y | z) \\
f(x, y | z) = f(x | z)f(y | z)
\]

Markov: [future | present, past], [child | parent, grandparent]

\[
p(y | x, z) = p(y | z) \\
f(y | x, z) = f(y | z)
\]
Correlation

Correlation = 0.81

Correlation = -0.81

Correlation = 0

Correlation = 0
Let $X$ be a uniform distribution over $[-1, 1]$. Let $Y = X^2$. Then $X$ and $Y$ are not independent. However, $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0$, and $\mathbb{E}(X) = 0$. Thus $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$. 

![Graph showing $y = x^2$ with $x$ ranging from -1 to 1]
Bivariate normal

\[ X \sim N(0, 1), \]
\[ Y = \rho X + \epsilon; \quad \epsilon \sim N(0, 1 - \rho^2), \]
\[ \mathbb{E}(Y) = \mathbb{E}(\rho X + \epsilon) = 0. \]

\(\epsilon\) and \(X\) are independent.

\[ \text{Var}(Y) = \text{Var}(\rho X + \epsilon) = \rho^2 \text{Var}(X) + \text{Var}(\epsilon) = 1. \]
\[ \text{Cov}(X, Y) = \mathbb{E}(XY) = \mathbb{E}[X(\rho X + \epsilon)] = \rho \mathbb{E}(X^2) + \mathbb{E}(X\epsilon) = \rho. \]
\[ \mathbb{E}(X\epsilon) = \mathbb{E}(X)\mathbb{E}(\epsilon) = 0. \]
Variance of sum

\[ \mathbb{E}(X + Y) = \sum_x \sum_y (x + y)p(x, y) = \sum_x \sum_y xp(x, y) + \sum_x \sum_y yp(x, y) = \mathbb{E}(X) + \mathbb{E}(Y). \]

\[ \text{Var}(X + Y) = \mathbb{E}[((X + Y) - \mu_{X+Y})^2] \]
\[ = \mathbb{E}(((X - \mu_X) + (Y - \mu_Y))^2] \]
\[ = \mathbb{E}((X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] \]
\[ = \mathbb{E}((X - \mu_X)^2] + \mathbb{E}((Y - \mu_Y)^2] + 2\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \]
\[ = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y). \]

If \( X \) and \( Y \) are independent, then \( \text{Cov}(X, Y) = 0 \), and
\[ \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \]
Variance of sum

\[ \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i^2 = Var(X) = \frac{1}{n} |\bar{x}|^2 \]
Variance of sum

\[ x + y = \text{const.} \]

For \( \text{Cov} > 0 \), the relationship between \( x \) and \( y \) is positive correlation.

For \( \text{Cov} < 0 \), the relationship between \( x \) and \( y \) is negative correlation.

\[ \theta \]
Average of iid

\[ X_1, X_2, ..., X_n \sim f(x) \] independently.

**independent and identically distributed, iid**

\[
\bar{x}, n = 1 \quad \bar{x}, n = 2 \quad \bar{x}, \text{large } n
\]

<table>
<thead>
<tr>
<th>( x_1 ) ( x_2 )</th>
<th>small</th>
<th>large</th>
<th>small</th>
<th>small</th>
<th>medium</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>large</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variance becomes smaller, distribution becomes smoother.
Average of iid

\[ x_1 \backslash x_2 \]
\[
\begin{array}{ccc}
\text{small} & \text{small} & \text{medium} \\
\text{large} & \text{medium} & \text{large}
\end{array}
\]
Average of iid

- On average girls grades better,
- Girls grades more consistent,
- Fewer top scoring girls

- On average girls grades much better,
- Girls grades similarly variable,
- More top scoring girls

- On average girls grades slightly better,
- Girls grades much more consistent,
- Many fewer top scoring girls
Sum and average of iid

\( X_i \sim f(x), \ i = 1, ..., n, \) iid: independent and identically distributed.

\[
S = \sum_{i=1}^{n} X_i. \quad \bar{X} = \frac{S}{n}.
\]

\[
\mathbb{E}(X_i) = \mu; \quad \text{Var}(X_i) = \sigma^2, \ i = 1, ..., n.
\]

\[
\mathbb{E}(S) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mu.
\]

\[
\text{Var}(S) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = n\sigma^2.
\]

\[
\mathbb{E}(\bar{X}) = \frac{\mathbb{E}(S)}{n} = \mu.
\]

\[
\text{Var}(\bar{X}) = \frac{\text{Var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.
\]
Law of large number

\[ \mathbb{E}(\bar{X}) = \frac{\mathbb{E}(S)}{n} = \mu. \]

\[ \text{Var}(\bar{X}) = \frac{\text{Var}(S)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0. \]

\[ \bar{X} \to \mu, \text{ in probability}. \]

\[ P(|\bar{X} - \mu| < \epsilon) \to 1, \forall \epsilon > 0. \]

Average \to \text{expectation}. 
Law of large number

Special case:

\[ X = \sum_{i=1}^{n} Z_i, \quad Z_i \sim \text{Bernoulli}(p) \text{ iid.} \]

\[ \mathbb{E}(X) = np; \quad \text{Var}(X) = np(1 - p). \]

\[ \mathbb{E}(X/n) = p; \quad \text{Var}(X/n) = p(1 - p)/n \rightarrow 0. \]

\[ X/n \rightarrow p, \text{ in probability.} \]

Frequency → probability.

\( X/n \) is average of \( Z_i \). Probability is expectation of \( Z_i \).
Law of large number

Special case:

Keep flipping a fair coin, frequency $\to 1/2$.

Intuition: most of $2^n$ sequences have frequencies close to $1/2$. 
Law of large number

Special case: $X_i \sim \text{Uniform}[0, 1]$, iid, $i = 1, \ldots, n$.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \rightarrow \mathbb{E}(X_i) = \frac{1}{2}.$$  

$$P(|\bar{X} - 1/2| < \epsilon) \rightarrow 1, \ \forall \epsilon > 0.$$  

**Intuition:** $(X_1, \ldots, X_i, \ldots, X_n)$ is a random point in $\Omega = [0, 1]^n$, $n$-dimensional unit cube.  

$A = \{(x_1, \ldots, x_i, \ldots, x_n) : |\bar{x} - 1/2| < \epsilon\}$ is the central diagonal piece.  

$P(A)$ is the volume of $A$. $P(A) \rightarrow 1$.  

No matter how small $\epsilon$ is, the volume of the central diagonal piece is almost the same as the volume of the whole $n$-dimensional unit cube $\Omega$.  

**Most of the points in $\Omega$ belong to $A$.** Concentration of measure.
Law of large number

\[ \bar{x}, n = 1 \]
\[ \bar{x} = 1/2 \]

\[ \bar{x}, n = 2 \]
\[ \bar{x} = 1/2 \]

\[ \bar{x}, n = 3 \]
Most of the points in $\Omega$ belong to $A$. Concentration of measure.
Suppose $(x_1, ..., x_i, ..., x_n)$ describes a physical system, e.g., $n = 10^{23}$ molecules.
It evolves \textbf{deterministically} over time, by traversing with $\Omega$.
\textbf{Ergodic}: it traverses every point in $\Omega$ with equal number of visits in the long run.
At any \textbf{random moment}, $(x_i, ..., x_i, ..., x_n) \sim \text{Unif}(\Omega)$.
Then most likely it will be in $A$, with fixed statistical properties (e.g., temperature, pressure, magnetism).
Central limit theorem

\[ X = \sum_{i=1}^{n} \epsilon_i, \quad \epsilon_i \sim \text{Bernoulli}(1/2) \text{ iid.} \]

\[ X \sim \text{Binomial}(n, 1/2). \quad \mu = \mathbb{E}(X) = n/2; \quad \sigma^2 = \text{Var}(X) = n/4. \]

\[ P \left( Z = \frac{X - n/2}{\sqrt{n}/2} = z \right) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z^2}{2} \right) \frac{2}{\sqrt{n}} = f(z) \Delta z. \]
Central limit theorem

Repeat and plot histogram

\[ S = \sum_{i=1}^{n} X_i. \]

\[ \mathbb{E}(X_i) = \mu; \ Var(X_i) = \sigma^2, \ i = 1, \ldots, n. \]

\[ S \sim N(n\mu, n\sigma^2). \]
$6^n$ equally likely sequences $\rightarrow 6^n$ equally likely sums $\rightarrow$ histogram.
Central limit theorem

\[ S = \sum_{i=1}^{n} X_i. \quad \bar{X} = S/n. \]

\[ \mathbb{E}(X_i) = \mu; \quad \text{Var}(X_i) = \sigma^2, \quad i = 1, \ldots, n. \]

\[ S \sim N(n\mu, n\sigma^2). \quad \bar{X} \sim N(\mu, \sigma^2/n). \]
Central limit theorem

Universal, regardless of the distribution of each $X_i$.

\[ S \sim N(n\mu, n\sigma^2). \quad \bar{X} \sim N(\mu, \sigma^2/n). \]

\[ Z = \frac{S - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \]
Probability models

Bayes network, graphical model: Part 1.
Poisson process: Part 2.
Brownian motion: Part 2.

\[ X_{t+\Delta t} = X_t + \sigma \sqrt{\Delta t} \epsilon_t, \]

where \( \mathbb{E}(\epsilon_t) = 0, \text{Var}(\epsilon_t) = 1, \) and \( \epsilon_t \) are iid.

Stochastic differential equation, diffusion

\[ X_{t+\Delta t} = X_t + \mu \Delta t + \sigma \sqrt{\Delta t} \epsilon_t, \]

\[ dX_t = \mu dt + \sigma dB_t. \]

Imagine 1 million particles moving.
Take home message

As long as you can count (and average)
(1) Population of equally likely possibilities
   Probability = population proportion
(2) Large sample of repetitions
   Frequency (sample proportion) \(\approx\) probability
   (a) Probability: population proportion, long run frequency
   (b) Expectation: population average, long run average
   (c) Conditional: sub-population, when something happens
   Continuous: discretize, infinitesimal analysis