# MAS 420: Probability Review

Adam Rohde
Department of Statistics
University of California, Los Angeles

September 26, 2022

# What is probability?

Probability is a mathematical construct to represent processes in the real world that involve randomness or uncertainty.

Consider a random process that selects an outcome from a set of possible outcomes.

The probability of an event $A$ occurring represents how frequently $A$ would occur among many repetitions of the process. (Frequentist view)

An alternative interpretation is that the probability of an event $A$ occurring represents a degree of belief in a proposition. (Bayesian view)

We'll take the frequentist view for this discussion.

# Recommended resources

These slides focus on probability, random variables, and relationships between random variables. We will not be dealing with data, estimation, or statistics.

Useful resources

- Foundations of Agnostic Statistics - Aronow and Miller (2019) [link] - part 1
- Causal Inference: the Mixtape - Cunningham (2021) [link] - chapter 2
- Statistical Rethinking (Videos) - McElreath (2022) [link] - these take a Bayesian perspective

These slides draw heavily from Aronow and Miller (2019).

# Sample space and events

**Definition (Sample Space)**

A sample space, $\Omega = \{\omega_1, \omega_2, ...\}$, is the set of all possible outcomes, $\omega$, of a random generating process.

**Definition (Event)**

An event, $A \subset \Omega$, is a subset of the sample space.

**Example**

Consider rolling a single six sided die.
The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
Events we might be interested in are $A = \{\omega \in \Omega : \omega \text{ is even}\}$, $B = \{1, 6\}$, or $C = \{3\}$.

# Event space and probability functions

## Definition (Event Space)

A set $S$ of subsets of $\Omega$ is an event space if it is

- non-empty: $S \neq \emptyset$
- closed under complements: $A \in S \implies A^c \in S$
- closed under countable unions: if $A_1, A_2, \cdots \in S \implies A_1 \cup A_2 \cup \cdots \in S$

## Definition (Probability Function)

A probability function assigns a number to every event in the event space: $P : S \to \mathbb{R}$.

## Example

Consider rolling a single six sided die and the event $B = \{1, 6\}$:
$P(B) = P(\text{the roll is 1 or 6}) = 1/3$

# Kolmogorov axioms

How do we ensure that the numbers we assign to events align with our basic intuitions about how probabilities should work?

## Definition (Kolmogorov Axioms)

- Non-negativity: $P(A) \geq 0, \forall A \in S$ Events have non-negative probability.
- Normalization to 1: $P(\Omega) = 1$ The probability that one of the outcomes occurs is 1.
- Additivity: if $A, B \in S$ are disjoint (i.e., $A \cap B = \emptyset$), then $P(A \cup B) = P(A) + P(B)$ The probability of non-overlapping events is equal to the sum of their individual probabilities.

## Example

Consider rolling a single six sided die.
$P(\text{the roll is 1}) = 1/6$
$P(\text{the roll is 1,2,3,4,5, or 6}) = 1$
$P(\text{the roll is 1 or 6}) = P(\text{the roll is 1}) + P(\text{the roll is 6}) = 1/6 + 1/6 = 1/3$

# Basic properties

## Basic properties of probability

For $A, B \in S$,

- $A \subset B \implies P(A) \leq P(B)$
- $A \subset B \implies P(B \backslash A) = P(B) - P(A)$
- $P(\emptyset) = 0$
- $0 \leq P(A) \leq 1$
- $P(A^c) = 1 - P(A)$

## Example

Consider rolling a single six sided die.
$P(\text{the roll is 1}) = P(\text{the roll is 1 or 6}) - P(\text{the roll is 6}) = 1/3 - 1/6 = 1/6$
$P(\text{the roll is not 1}) = 1 - P(\text{the roll is 1}) = 1 - 1/6 = 5/6$

# Joint and conditional probabilities

## Definition (Joint probability)

For $A, B \in S$, the joint probability of $A$ and $B$ is $P(A \cap B)$. The joint probability is the probability of the intersection of $A$ and $B$, i.e., the probability that both $A$ and $B$ occur.

## Addition rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

## Definition (Conditional probability and Bayes Rule)

For $A, B \in S$ and $P(B) > 0$, $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
This implies the multiplication rule: $P(A \cap B) = P(A|B)P(B)$.
These imply Bayes rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

# Law of total probability

**Law of total probability**

If $A_1, A_2, \ldots$ form a partition of $\Omega$ and $B \in S$, then

- $A_1 \cup A_2 \cup \cdots = \Omega$, $A_i \cap A_j = \emptyset$, and $\sum_i P(A_i) = 1$
- $P(B) = \sum_i P(B \cap A_i)$
- $P(B) = \sum_i P(B|A_i)P(A_i)$

The probability of $B$ is a weighted average of conditional probabilities.

We can also use this to write Bayes rule as $P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_i P(B|A_i)P(A_i)}$

# Independence of events

**Definition (Independence of events)**

For $A, B \in S$, $A$ and $B$ are independent if $P(A \cap B) = P(A)P(B)$.

We can then see that for independent events we have
$P(A)P(B) = P(A \cap B) = P(A|B)P(B)$ by the multiplication rule.

Which means that $A$ and $B$ are independent iff $P(A) = P(A|B)$.

If knowing whether or not $B$ occurred tells us nothing about whether or not $A$ occurred, the two events are independent.

*(Events are mutually exclusive if the occurrence of one event excludes the occurrence of the other. Mutually exclusive events cannot happen at the same time. So, typically, mutually exclusive events are not independent events. For mutually exclusive events,* $P(A|B) = 0$ *even if* $P(A) > 0$*. For independent events, these are equal.)*

# Random variables

A random variable is function that maps the outcomes in the sample space to numeric outcomes: $X : \Omega \to \mathbb{R}$.

**Example**

Consider flipping a fair coin. The sample space is $\Omega = \{\text{heads, tails}\}$. A random variable representing the coin toss would be $X = \begin{cases} 0 \text{ if tails} \\ 1 \text{ if heads} \end{cases}$

**Example**

Consider rolling a six sided die. The sample space is really $\Omega = \{\text{face with one dot, face with two dots, ...}\}$. A random variable representing the die roll would be

$$X = \begin{cases} 1 \text{ if face with one dot} \\ 2 \text{ if face with two dots} \\ \dots \end{cases}$$

# Discrete random variables

## Definition (Discrete random variable)

A random variable is discrete if its range is a countable set, that is, if it can only take on a finite number of values.

## Definition (Probability mass function (PMF))

For a discrete random variable $X$, the probability mass function is $f(x) = Pr(X = x), \forall x \in \mathbb{R}$. Further, $\sum_x f(x) = 1$.

## Example

Consider a potentially biased coin flip. The sample space is $\Omega = \{\text{heads, tails}\}$ and $X$ is 1 for heads but 0 for tails. Suppose the probability of heads is $0 \leq p \leq 1$. Then the PMF is

$$f(x) = Pr(X = x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

# Cumulative distribution functions

**Definition (Cumulative distribution function (CDF))**

For a random variable $X$, the CDF is $F(x) = Pr(X \leq x), \forall x \in \mathbb{R}$.

**Properties of CDFs**

For a random variable $X$ with CDF $F$,

- $F$ is non-decreasing: $\forall x_1, x_2 \in \mathbb{R}$, if $x_1 < x_2$, then $F(x_1) \leq F(x_2)$.
- $\lim_{x \to -\infty} F(x) = 0$
- $\lim_{x \to \infty} F(x) = 1$
- $\forall x \in \mathbb{R}, 1 - F(x) = Pr(X > x)$

# Continuous random variables

## Definition (Continuous random variable)

A random variable is continuous if there exists a non-negative function $f : \mathbb{R} \to \mathbb{R}$ such that the CDF of $X$ is $F(x) = Pr(X \leq x) = \int_{-\infty}^{\infty} f(u)du, \forall x \in \mathbb{R}$. Continuous random variables can take on uncountably infinite numbers of different values.

## Definition (Probability density function (PDF))

For a continuous random variable $X$ with CDF $F$, the PDF is $f(x) = \frac{d}{du}F(u)\Big|_{u=x}, \forall x \in \mathbb{R}$.

## Example

Consider human heights. Height can take on infinitely many values between say zero and 10 feet. You are not either 5 feet tall or 6 feet tall; you can be lots of heights. But different heights are more and less likely. Height is typically modeled with a normal distribution centered some where between 5 and 6 feet.

# Continuous random variables

## Properties of continuous random variable

For a continuous random variable $X$ with PDF $f$,

- $\forall x \in \mathbb{R}, f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $Pr(X = x) = 0$
- $Pr(X < x) = Pr(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)du$
- $Pr(X > x) = Pr(X \geq x) = 1 - F(x) = \int_{x}^{\infty} f(u)du$
- $Pr(a < X < B) = Pr(a \leq X \leq B) = F(b) - F(a) = \int_{a}^{b} f(x)dx$

# Joint distributions

## Definition (Joint PMF and CDF for Discrete RVs )

For discrete random variables $X, Y$, the joint PMF is $f(x, y) = Pr(X = x, Y = y), \forall x, y \in \mathbb{R}$. The joint CDF is $F(x, y) = Pr(X \leq x, Y \leq y), \forall x, y \in \mathbb{R}$

## Definition (Joint PDF and CDF for Continuous RVs )

For continuous random variables $X, Y$, joint CDF is $F(x, y) = Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv, \forall x, y \in \mathbb{R}$. The joint PDF is $f(x, y) = \frac{\delta^2 F(u,v)}{\delta u \delta v}\Big|_{u=x, v=y}, \forall x, y \in \mathbb{R}$.

## Example

Consider human heights and weight. These variables relate to one another and there are different probabilities for the different combinations these variables can take. Again, these might be modeled with a joint distribution - perhaps a bivariate normal distribution.

# Marginal and conditional distributions

## Definition (Marginal PMF and PDF)

For discrete random variables $X, Y$, the marginal PMF is
$f_Y(y) = Pr(Y = y) = \sum_x f(x, y), \forall y \in \mathbb{R}$.
For continuous random variables $X, Y$, the marginal PDF is
$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx, \forall y \in \mathbb{R}$.

## Definition (Conditional PMF and PDF)

For discrete random variables $X, Y$, the conditional PMF of $Y$ given $X = x$ is
$f_{Y|X}(y|x) = Pr(Y = y|X = x) = \frac{Pr(X=x, Y=y)}{Pr(X=x)} = \frac{f(x,y)}{f_X(x)}, \forall y \in \mathbb{R}$ and possible $x$.
For continuous random variables $X, Y$, the conditional PDF of $Y$ given $X = x$ is
$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \forall y \in \mathbb{R}$ and possible $x$.

## Multiplicative law

$f_{Y|X}(y|x)f_X(x) = f(x, y)$

# Independence of random variables

## Definition (Independence of random variables)

For random variables $X, Y$ with joint PDF or PMF $f$, $X, Y$ are independent if, $\forall x, y \in \mathbb{R}$, $f(x, y) = f_X(x) f_Y(y)$. Write $X \perp\!\!\!\perp Y$ to denote that they are independent.

## Implications of independence (part 1)

All of the following statements are equivalent:

- $X \perp\!\!\!\perp Y$
- $f(x, y) = f_X(x) f_Y(y)$, $\forall x, y \in \mathbb{R}$
- $f_{Y|X}(y|x) = f_Y(y)$
- $\forall D, E \subset \mathbb{R}$, the events $\{X \in D\}$ and $\{Y \in E\}$ are independent.
- For all functions $g$ of $X$ and $h$ of $Y$, $g(X) \perp\!\!\!\perp h(Y)$.

e.g., height and weight would be dependent - as you get taller you tend to weigh more; height and whether you like ice cream are probably independent

# Expected values

Distributions are often summarized with measures of their center and spread.
Expected values are the most common measures of the center of probability distributions.
They provide information on where typical values for the variable are centered.

**Definition (Expected values)**

For a discrete RV $X$ with PMF $f$ the expected value is $\mathbb{E}[X] = \sum_x xf(x)$.
For a continuous RV $X$ with PDF $f$ the expected value is $\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$.

**Example**

Let $X$ be a Bernoulli random variable (0 or 1, possibly representing a coin flip) with probability of 1 (heads) $0 \leq p \leq 1$. Then $\mathbb{E}[X] = \sum_{x=0}^{1} xf(x) = 0 \times (1-p) + 1 \times p = p$.

e.g., an expected value for height of, say, 5 foot 6 inches tells us something about typical heights that is different than if the expected value for height were 7 feet

# Expected values

## Properties of expected values

- Expectation of a function of a RV
  - For a discrete RV $X$, $\mathbb{E}[g(X)] = \sum_x g(x)f(x)$.
  - For a continuous RV $X$, $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.
- $\mathbb{E}[c] = c, \forall c \in \mathbb{R}$
- $\mathbb{E}[cX] = c\mathbb{E}[X], \forall c \in \mathbb{R}$
- Linearity of expectations: $\forall a, b, c \in \mathbb{R}$, $\mathbb{E}[a + bX + cY] = a + b\mathbb{E}[X] + c\mathbb{E}[Y]$.

# Variance

Variance is the most common measure of a distribution's spread or variability. It measures how far apart values for the variable typically are: the expected value of the squared difference between observed values and the mean.

**Definition (Variance)**

Variance of a RV $X$ is $V[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

We can alternatively write

$$
\begin{aligned}
V[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] && \text{expanding out} \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2] && \text{by linearity of expectations} \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 && \text{since } \mathbb{E}[X] \text{ is a constant} \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned}
$$

So $V[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

# Variance and standard deviations

## Properties of variance

- $V[c] = 0, \forall c \in \mathbb{R}$
- $V[c + X] = V[X], \forall c \in \mathbb{R}$
- $V[cX] = c^2 V[X], \forall c \in \mathbb{R}$

## Definition (Standard deviation)

Standard deviation of a RV $X$ is $SD[X] = \sqrt{V[X]}$. $SD[X]$ will have the same units as $X$; $V[X]$ does not since it deals with squared deviations.

## Example

Let $X$ be a Bernoulli random variable (0 or 1, possibly representing a coin flip) with probability of 1 (heads) $0 \leq p \leq 1$.
First, $\mathbb{E}[X^2] = p \times 1^2 + (1 - p) \times 0^2 = p$. We have from above that $\mathbb{E}[X] = p$.
So $V[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p)$; $SD[X] = \sqrt{p(1 - p)}$.

# Mean squared error

We often want to understand how well an RV approximates some value $c$. To do so we want a metric that measures how far $X$ is from $c$ on average.

**Definition (Mean squared error (MSE) )**

Mean squared error for RV $X$ and constant $c$ is $\mathbb{E}[(X-c)^2]$.

We can alternatively write

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[X^2 - 2cX + c^2] \\ &= \mathbb{E}[X^2] + \left(\mathbb{E}[X]^2 - \mathbb{E}[X]^2\right) - 2c\mathbb{E}[X] + c^2 \\ &= \left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right) + \left(\mathbb{E}[X]^2 - 2c\mathbb{E}[X] + c^2\right) \\ &= V[X] + (\mathbb{E}[X] - c)^2 \quad \text{MSE is composed of variance, expected value, and } c.\end{aligned}$$

Note that $\arg\min_{c\in\mathbb{R}} \mathbb{E}[(X-c)^2] = \arg\min_{c\in\mathbb{R}} \left(V[X] + (\mathbb{E}[X]-c)^2\right) =$ $\arg\min_{c\in\mathbb{R}} \left((\mathbb{E}[X]-c)^2\right) = \mathbb{E}[X]$.

So the expected value of $X$ is the value for $c$ that minimizes MSE.

# Covariance

We often want to measure the extent to which two RVs move together or to describe the relationship between two variables (e.g., consider the relationship between height and weight). We can generalize variance to this end.

**Definition (Covariance)**

The covariance between two RVs $X, Y$ is $\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.

Again, we can rewrite covariance as $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
Large positive covariance means that high values of $X$ appear when there are high values for $Y$. (Height and weight have a positive covariance. Taller people tend to weigh more.) Large negative covariance means that high values of $X$ appear when there are low values for $Y$. Note that covariance depends on the distributions of both $X$ and $Y$ and so covariance will differ in magnitude for different pairs of RVs.

# Properties of covariance

## Properties of covariance

- $V[X+Y] = V[X] + 2\text{Cov}[X,Y] + V[Y]$
- $V[a+bX+cY] = b^2V[X] + 2bc\text{Cov}[X,Y] + c^2V[Y]$
- $\text{Cov}[c,X] = \text{Cov}[X,c] = \text{Cov}[c,d] = 0, \forall c,d \in \mathbb{R}$
- $\text{Cov}[X,Y] = \text{Cov}[Y,X]$
- $\text{Cov}[X,X] = V[X]$
- $\text{Cov}[a+bX, c+dY] = bd\text{Cov}[X,Y], \forall a,b,c,d \in \mathbb{R}$
- $\text{Cov}[X+W, Y+Z] = \text{Cov}[X,Y] + \text{Cov}[X,Z] + \text{Cov}[W,Y] + \text{Cov}[W,Z]$

# Correlation

Like standard deviation rescales variance, correlation rescales covariance. Correlation is positive when covariance is positive and negative when covariance is negative. Where as covariance can take any value, correlation is bounded by -1 and 1.

## Definition (Correlation)

The correlation between two RVs $X, Y$ is $\rho[X, Y] = \frac{\text{Cov}[X,Y]}{\sqrt{V[X]}\sqrt{V[Y]}}$.

## Correlation and *linear* dependence

- $\rho[X, Y] \in [-1, 1]$
- $\rho[X, Y] = 1 \iff Y = a + bX$ for some $a, b \in \mathbb{R}, b > 0$
- $\rho[X, Y] = -1 \iff Y = a - bX$ for some $a, b \in \mathbb{R}, b > 0$

Note that $\rho[X, Y] = 0$ does not mean $X \perp\!\!\!\perp Y$. E.g., $X \sim U(-1, 1)$, $Y = X^2$. This is a non-linear relationship. [R demo]

# Properties of Correlation and Independence of RVs

## Properties of Correlation

- $\rho[X, Y] = \rho[Y, X]$
- $\rho[X, X] = 1$
- $\rho[a + bX, c + dY] = \rho[X, Y]$ if $b, d > 0$ or $b, d < 0$
- $\rho[a + bX, c + dY] = -\rho[X, Y]$ if $b < 0 < d$ or $d < 0 < b$

## Implications of independence (part 2)

If $X, Y$ are independent RVs ($X \perp\!\!\!\perp Y$), then

- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- $\text{Cov}[X, Y] = 0$
- $\rho[X, Y] = 0$
- $V[X + Y] = V[X] + V[Y]$

# Conditional expectations

Conditional expectations are a key way to understand how one variable relates to another. They let us describe how the center of one RV's distribution changes when we condition on some value of the other variable.

**Definition (Conditional expectations)**

For discrete RVs, $\mathbb{E}[Y|X=x] = \sum_y y f_{Y|X}(y|x)$ for all possible values of $x$.
For continuous RVs, $\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$ for all possible values of $x$.

**Linearity of conditional expectations**

For RVs $X, Y$ and functions $g, h$, $\mathbb{E}[g(X) + h(X)Y|X=x] = g(x) + h(x)\mathbb{E}[Y|X=x]$.

# Conditional expectation functions

For every value of $x$, $\mathbb{E}[Y|X=x]$ maps $Y$ to the conditional mean of $Y$ (a single value). So we can consider this a type of function that takes in values of $x$ and outputs the conditional expectation of $Y$ when $X=x$.

**Definition (Conditional expectation functions (CEF))**

For RVs $X, Y$ with joint PMF/PDF $f$, the conditional expectation function of $Y$ given $X=x$ is $G_Y(x) = \mathbb{E}[Y|X=x]$ for all possible values of $x$. We usually write $\mathbb{E}[Y|X]$ to denote the CEF.

**Law of iterated expectations**

$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$

# Properties of conditional expectation functions

## Properties of conditional expectation functions

For RVs, $X, Y$ and $\epsilon = Y - \mathbb{E}[Y|X]$,

- $\mathbb{E}[\epsilon|X] = 0$ the deviations are centered on zero conditional on $X$
- $\mathbb{E}[\epsilon] = 0$ the deviations are centered on zero
- for function $g$, $\text{Cov}[g(X), \epsilon] = 0$ there is no linear relationship between X and the deviations

$\mathbb{E}[\epsilon|X] = \mathbb{E}[Y - \mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0$
$\mathbb{E}[\epsilon] = \mathbb{E}[\mathbb{E}[\epsilon|X]] = \mathbb{E}[0] = 0$
I'll leave the last property to you.

# Best predictors

Suppose $X, Y$ are RVs with some joint distribution and that we observe that $X = x$.
What is the best guess for the value of $Y$ (where best means lowest MSE)?
That is, what function $g$ of $X$ minimizes $\mathbb{E}[(Y - g(X))^2]$?

**CEF is the best predictor**

For RVs $X, Y$, the CEF, $\mathbb{E}[Y|X]$, is the best (minimum MSE) predictor of $Y$ given $X$. (Proof on page 75 of Aronow and Miller (2019))

So the CEF is the best way to approximate $Y$ when we know $X$. Thus, the CEF is a good target when we are trying to study how $X$ relates to $Y$. But the CEF can be extremely complicated, since there are no restrictions on the distributions of $Y$ or $X$.

# Best linear predictors

We can say more if we limit our prediction functions to linear functions of $X$. That is, restricting to $g(X) = a + bX$ for some values of $a, b \in \mathbb{R}$.

What is the best guess for the value of $Y$ (i.e., $g$ that minimizes $\mathbb{E}[(Y - g(X))^2]$) using linear $g$?

## Best linear predictor (BLP)

For RVs $X, Y$, the best (minimum MSE) linear predictor of $Y$ given $X$ is $g(X) = \alpha + \beta X$, where

$$\alpha = \mathbb{E}[Y] - \frac{\text{Cov}[X, Y]}{V[X]}\mathbb{E}[X] \text{ and } \beta = \frac{\text{Cov}[X, Y]}{V[X]}$$

(Proof on page 77 of Aronow and Miller (2019))

The BLP has an important relationship with OLS regression.
We will call $\beta$ the regression coefficient.

# Properties of BLP

## Properties of BLP

For RVs $X, Y$ and $\epsilon = Y - g(X)$, where $g(X) = \alpha + \beta X$ is the BLP,

- $\mathbb{E}[\epsilon] = 0$
- $\mathbb{E}[X\epsilon] = 0$
- $\text{Cov}[X, \epsilon] = 0$

While $\mathbb{E}[\epsilon|X] = 0$ for the CEF this is not true for the BLP. We only have the weaker statement $\mathbb{E}[\epsilon] = 0$.
Also, there is no covariance between $X$ and the deviations.

## BLP for independent RVs

For independent RVs $X, Y$, the BLP of $Y$ given $X$ is $\mathbb{E}[Y]$.

# Covariance, correlation, and the regression coefficient

Covariance, correlation, and the regression coefficient from the BLP have the following relationship. Note these are all capturing linear associations between $X$ and $Y$.

$$\beta = \frac{\text{Cov}[X, Y]}{V[X]} = \frac{\sqrt{V[Y]}}{\sqrt{V[X]}}\rho[X, Y]$$

See Aronow and Miller (2019) for multivariate generalizations, proofs, and more detail.

# Possible topics for next Section

Review of matrix algebra
Review of statistics
Review of OLS