Joint Probability

Lecture 3

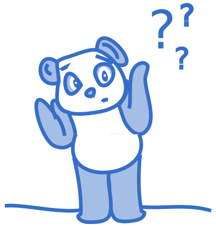
Please, sign in on iClicker

Outline

1. Joint Distributions
2. Independence and Dependence for Random Variables

1. Joint Distributions

- So far, we have only considered one random variable at a time which has an univariate distribution.
- However, we often have more than one random variable.
- A joint distribution is a probability distribution involving two or more random variables.



Coins come again!

- Consider two independent fair coins (i.e., two independent Bernoulli random variables!).

Random Variable Setup

- Define the following random variables:

$$X = \text{First coin's outcome.}$$
$$Y = \text{Second coin's outcome.}$$

- The joint distribution of this process is the following:

X/Y	`H`	`T`
`H`	0.25	0.25
`T`	0.25	0.25

Computing Probabilities

- Each cell of the previous joint distribution is computed as:

$$
\begin{aligned}
P(X = \texttt{H} \cap Y = \texttt{H}) &= P(X = \texttt{H}) \cdot P(Y = \texttt{H}) \quad \because \text{independen} \\
&= 0.5 \cdot 0.5 \quad \because \text{fair coins} \\
&= 0.25
\end{aligned}
$$

Can we have an univariate setup?

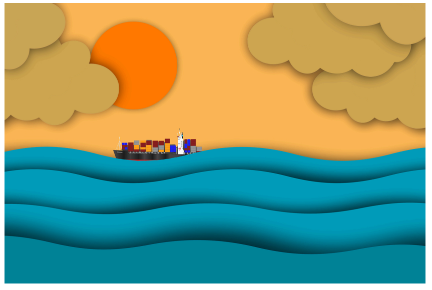
- Alternatively, we can define the following random variable:

$$Z = \text{Outcomes obtained when tossing two independent coi}$$

Outcome	Probability
HH	0.25
HT	0.25
TH	0.25
TT	0.25

1.1. Example: Length of Stay Versus Gang Demand

- Consider an example that a Vancouver port faces with gang demand.
- When a ship arrives, they request a certain number of gangs to unload the ship.
- We will work with the following joint distribution of length of stay (LOS) of a ship and its gang demand (Gangs).



PMFs for Gangs and LOS

Number of Gangs	Probability
1	0.2
2	0.4
3	0.3
4	0.1

LOS	Probability
1	0.25
2	0.35
3	0.20
4	0.10
5	0.10

Now, we might wonder…

> What is the probability that a ship requires 4 gangs **AND** will stay in port for 5 days?

- The information provided by both separate PMFs (Gangs and LOS) is not sufficient to answer this question.
  - These PMFs are called marginal distributions.
- We would need to know the joint distribution between LOS and Gangs.

The joint distribution…

- We need a probability for every possible combination of the number of Gangs and LOS.
- In this case, $5 \times 4 = 20$ probabilities.

	Gangs = 1	**Gangs = 2**	**Gangs = 3**	**Gangs = 4**
LOS = 1	0.00170	0.04253	0.12471	0.08106
LOS = 2	0.02664	0.16981	0.13598	0.01757
LOS = 3	0.05109	0.11563	0.03203	0.00125
LOS = 4	0.04653	0.04744	0.00593	0.00010
LOS = 5	0.07404	0.02459	0.00135	0.00002

> What is the probability that a ship requires 4 gangs **AND** will stay in port for 5 days?

Now, we might wonder…

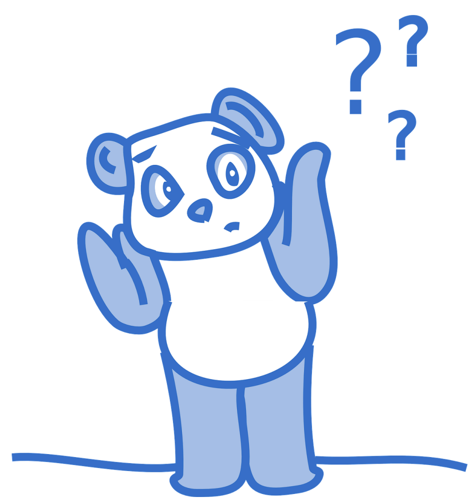
> Could the 20 numbers in the joint distribution be **ANY** probabilities between 0 and 1?

- No, we have the following restrictions:
  - They are restricted by the fact that they will need to add up to 1 (recall the Axiom of Probability for sample space!).
  - Joint distribution need to be consistent with marginal distributions.

1.2. Calculating Marginal Distributions from the Joint Distribution

Let us define the following:

- In a random system/process with more than one random variable, the distribution of a standalone variable is called a marginal distribution.



Calculating Marginal Distributions from the Joint Distribution

- In the case of discrete random variables, we add up the probabilities of the corresponding standalone outcomes.

	Gangs = 1	Gangs = 2	Gangs = 3	Gangs = 4
LOS = 1	0.00170	0.04253	0.12471	0.08106
LOS = 2	0.02664	0.16981	0.13598	0.01757
LOS = 3	0.05109	0.11563	0.03203	0.00125
LOS = 4	0.04653	0.04744	0.00593	0.00010
LOS = 5	0.07404	0.02459	0.00135	0.00002

Let us start with the marginal distribution of LOS...

- We can compute $P(\mathrm{LOS} = 1)$.
- There are four ways this could happen:
  - $\mathrm{LOS} = 1$ and $\mathrm{Gangs} = 1$.
  - $\mathrm{LOS} = 1$ and $\mathrm{Gangs} = 2$.
  - $\mathrm{LOS} = 1$ and $\mathrm{Gangs} = 3$.
  - $\mathrm{LOS} = 1$ and $\mathrm{Gangs} = 4$.
- These events are disjoint.

So we know

$$
\begin{aligned}
P(\text{LOS} = 1) = P(&\text{LOS} = 1 \cap \text{Gangs} = 1) + \\
&P(\text{LOS} = 1 \cap \text{Gangs} = 2) + \\
&P(\text{LOS} = 1 \cap \text{Gangs} = 3) + \\
&P(\text{LOS} = 1 \cap \text{Gangs} = 4) \\
= 0.25.
\end{aligned}
$$

	Gangs = 1	Gangs = 2	Gangs = 3	Gangs = 4
LOS = 1	0.00170	0.04253	0.12471	0.08106
LOS = 2	0.02664	0.16981	0.13598	0.01757
LOS = 3	0.05109	0.11563	0.03203	0.00125
LOS = 4	0.04653	0.04744	0.00593	0.00010
LOS = 5	0.07404	0.02459	0.00135	0.00002

We have $P(\text{LOS} = 1)$...

- But we would also need $P(\text{LOS} = 2)$, $P(\text{LOS} = 3)$, etc.
- Thus, we add up each row from our joint distribution.

R Code | **Output**

```r
rowSums(joint_distribution) %>%
  kable(col.names = "Probability", align = "c") %>%
  kable_styling(font_size = 30) %>%
  column_spec(1, bold = TRUE)
```

Now for Gangs!

R Code | Output

```r
colSums(joint_distribution) %>%
  kable(col.names = "Probability", align = "c") %>%
  kable_styling(font_size = 30) %>%
  column_spec(1, bold = TRUE)
```

- Note both marginals computed from the joint are consistent with our initial marginals.

iClicker Question

Answer **TRUE** or **FALSE**:

We obtain a marginal distribution by summing the rows of a joint distribution; therefore, each row of a joint distribution must sum to 1.

A. TRUE

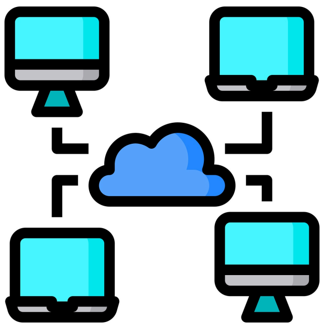
B. FALSE

Answer

- It is **FALSE**.
- The entire table sums to one, not individual rows.

2. Independence and Dependence Concepts

- A big part of Data Science is about modeling the relationship between the variables in our datasets.
  - What factors influence number of tickets sold for a game?
  - Some factors (team performance) may be dependent, while others (weather) are independent.



2.1. Independence

- Let X and Y be two random variables.
- X and Y are independent if knowing something about one of them tells us nothing about the other:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y), \text{for all } x \text{ and}$$

- We would only need the marginals to obtain their joint distribution.

Product of Expectation for Independent Random Variables

- If X and Y are independent, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

- Contrast this to expectation of sum of random variables

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y),$$

which does not require the random variables to be independent.

Going back to the two coins!

- Recall we had this joint distribution:

$$X = \text{First coin's outcome}$$
$$Y = \text{Second coin's outcome.}$$

X/Y	H	T
H	0.25	0.25
T	0.25	0.25

Step 1: Obtaining the Marginals from the Joint

X/Y	`H`	`T`	**Marginals for X**
`H`	0.25	0.25	0.5
`T`	0.25	0.25	0.5
Marginals for Y	0.5	0.5	

Step 2: Applying the Independence Property via the Marginals

X/Y	`H`	`T`	Marginals for X
`H`	0.25	0.25	0.5
`T`	0.25	0.25	0.5
Marginals for Y	0.5	0.5	

Since

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y), \text{for all } x, y \in \{$$

We can see that the two coin flips are independent:

Let us check another two-coin case…

$$X = \text{First coin's outcome}$$
$$Y = \text{Second coin's outcome.}$$

X/Y	H	T
H	0.2	0.6
T	0.05	0.15

Computing the Marginals

X/Y	`H`	`T`	**Marginals for X**
`H`	0.2	0.6	0.8
`T`	0.05	0.15	0.2
Marginals for Y	0.25	0.75	

Applying the Independence Property via the Marginals

X/Y	`H`	`T`	**Marginals for X**
`H`	0.2	0.6	0.8
`T`	0.05	0.15	0.2
Marginals for Y	0.25	0.75	

We can see that the these two coins are also independent!
Because

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y), \text{for all } x, y \in \{$$

But there is no independence in this other two-coin case!

$$X = \text{First coin's outcome}$$
$$Y = \text{Second coin's outcome.}$$

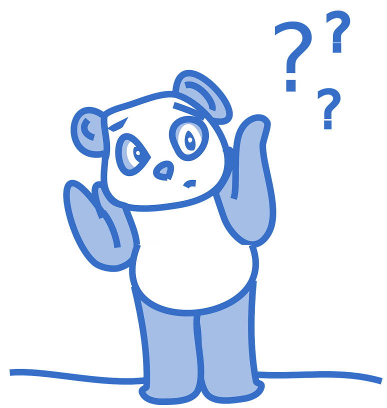
X/Y	`H`	`T`
`H`	0.5	0
`T`	0	0.5

- These two coins are completely dependent!

2.2. Measures of Dependence

- Let us ask ourselves the following:

> What if two random variables are not independent?



> Is there some measure of dependence?

2.2.1. Covariance and Pearson's Correlation

- Covariance is one common way of measuring dependence between two numeric random variables.
- It measures the amount of dependence and direction:

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Going back to our cargo ship example!

R Code | Output

```r
joint_distribution %>%
  kable(align = "cccc") %>%
  kable_styling(font_size = 30) %>%
  column_spec(1, bold = TRUE)
```

- For a larger LOS, there are larger probabilities associated with a smaller gang demand.

Exercise: Compute the Covariance of LOS and Gangs

- Recall that
  $\mathrm{Cov}(\mathrm{LOS}, \mathrm{Gangs}) = \mathbb{E}(\mathrm{LOS} \cdot \mathrm{Gangs}) - \mathbb{E}(\mathrm{LOS})\mathbb{E}(\mathrm{Gangs}$

Step 1: Calculate the Marginal PMFs

R Code | **Output**

```r
Marginal_PMF_LOS <- tribble(
  ~n_days, ~p,
  1, 0.25,
  2, 0.35,
  3, 0.2,
  4, 0.1,
  5, 0.1
)
Marginal_PMF_LOS

Marginal_PMF_Gangs <- tribble(
  ~n_gangs, ~p,
  1, 0.2,
  2, 0.4,
  3, 0.3,
  4, 0.1,
)
Marginal_PMF_Gangs
```

Step 2: Compute $\mathbb{E}(\text{LOS})$ and $\mathbb{E}(\text{Gangs})$

R Code | Output

```r
E_LOS <- sum(Marginal_PMF_LOS$n_days * Marginal_PMF_LOS$p)
E_LOS

E_Gangs <- sum(Marginal_PMF_Gangs$n_gangs * Marginal_PMF_Gangs$p)
E_Gangs
```

Hence:

$$\mathbb{E}(\text{LOS}) = 2.45$$

$$\mathbb{E}(\text{Gangs}) = 2.3.$$

Step 3: Compute $\mathbb{E}(\mathrm{LOS} \cdot \mathrm{Gangs})$

- Melting `joint_distribution` (manually!)

R Code | **Output**

```r
joint_distribution <- data.frame(
  LOS = c(rep(1, 4), rep(2, 4), rep(3, 4), rep(4, 4), rep(5, 4)),
  Gangs = rep(1:4, 5),
  p = c(
    0.00170, 0.04253, 0.12471, 0.08106,
    0.02664, 0.16981, 0.13598, 0.01757,
    0.05109, 0.11563, 0.03203, 0.00125,
    0.04653, 0.04744, 0.00593, 0.00010,
    0.07404, 0.02459, 0.00135, 0.00002
  )
)
joint_distribution
```

Computing the Crossed Expected Value

R Code Output

```r
E_LOS_Gangs <- sum(joint_distribution$LOS *
  joint_distribution$Gangs *
  joint_distribution$p)
E_LOS_Gangs
```

Thus:

$$\mathbb{E}(\text{LOS} \cdot \text{Gangs}) = 4.89956.$$

Computing the Covariance

$$
\begin{aligned}
\mathrm{Cov}(\mathrm{LOS}, \mathrm{Gangs}) &= \mathbb{E}(\mathrm{LOS} \cdot \mathrm{Gangs}) - \mathbb{E}(\mathrm{LOS})\mathbb{E}(\mathrm{Gangs}) \\
&= 4.89956 - [(2.45)(2.3)] \\
&= -0.73544.
\end{aligned}
$$

- Indeed, we can see that the covariance between LOS and Gangs is negative.
- A negative sign indicates that an increase in LOS is associated with a decrease in Gangs.

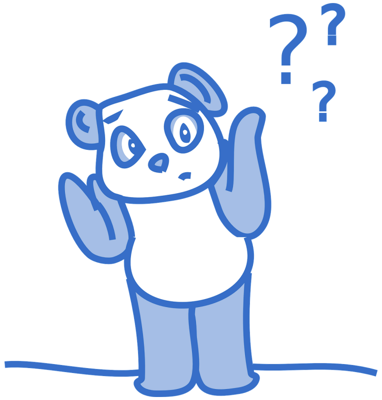
iClicker Question

Answer **TRUE** or **FALSE**:

Covariance can be negative, but not the variance.

A. TRUE

B. FALSE



Answer

- It is **TRUE**.
- Covariance can also have a negative sign.
- Nonetheless, it will not be restricted between -1 and 1.
- On the other hand, variance's mathematical definition will always make it non-negative:

$$\text{Var}(X) = \mathbb{E}\{[X - \mathbb{E}(X)]^2\}.$$

iClicker Question

Answer **TRUE** or **FALSE**:

Without any further assumptions between random variables X and Y, covariance is calculated as

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

Computing $\mathbb{E}(XY)$ requires the joint distribution, but computing $\mathbb{E}(X)\mathbb{E}(Y)$ only requires the marginals.

A. TRUE

B. FALSE

Answer

- It is **TRUE** for any class of random variables.
- For discrete random variables, with $P(X = x, Y = y)$ being the joint distribution along with the marginals $P(X = x)$ and $P(Y = y)$, we define the following:

$$\mathbb{E}(XY) = \sum_{x,y} xy \cdot P(X = x, Y = y)$$

$$\mathbb{E}(X) = \sum_{x} x \cdot P(X = x)$$

$$\mathbb{E}(Y) = \sum_{y} x \cdot P(Y = y).$$

Covariance Drawback

Pearson's Correlation Coefficient

- Pearson's correlation standardizes the distances according to the standard deviations σ_X and σ_Y of X and Y, respectively.

$$\begin{aligned}\operatorname{Corr}(X, Y) &= \mathbb{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right] \\ &= \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}.\end{aligned}$$

- Note that $-1 \leq \operatorname{Corr}(X, Y) \leq 1$.

Pearson's Correlation Scale

- -1 means a perfect negative linear relationship between X and Y.
- 0 means no linear relationship (however, this does not mean independence!).
- 1 means a perfect positive linear relationship.

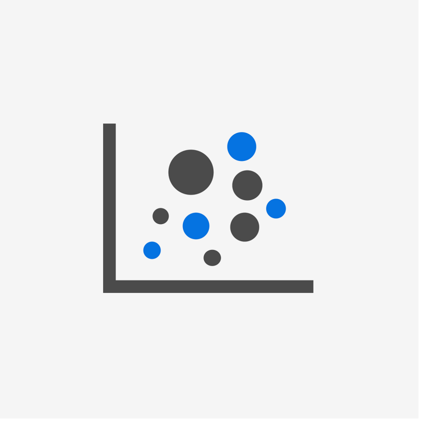
Correlation coefficient is invariant to scaling

- If we multiply X by 10, then the correlation of X and Y remains the same:

$$
\begin{aligned}
\operatorname{Corr}(10X, Y) &= \frac{10 \operatorname{Cov}(X, Y)}{\sqrt{10^2 \operatorname{Var}(X) \operatorname{Var}(Y)}} \\
&= \operatorname{Corr}(X, Y)
\end{aligned}
$$

2.2.2. Kendall's τ_K

- Pearson's correlation measures linear dependence.
- But many relationships between real-world variables are not linear.
  - Ordinal variables: variables have natural, ordered categories



Example of ordinal data

Movie	Critic X Rank	Critic Y Rank
A	1	1
B	2	9
C	3	3
D	4	5
E	5	10

- The rankings of two variables are consistent, but the actual values differ non-linearly

Definition of Kendall's τ_K rank coefficient

- Kendall's τ_K can be used to measure the ordinal association between two variables.
- It measures agreement between each pair of observations (x_i, y_i) and (x_j, y_j) with $i \neq j$:

Agreement means

$$
\begin{gathered}
x_i < x_j \quad \text{and} \quad y_i < y_j, \\
\text{or} \\
x_i > x_j \quad \text{and} \quad y_i > y_j;
\end{gathered}
$$

which gets a positive sign.

Definition of Kendall's τ_K

Disagreement means

$$
x_i < x_j \quad \text{and} \quad y_i > y_j,
$$
$$
\text{or}
$$
$$
x_i > x_j \quad \text{and} \quad y_i < y_j;
$$

which gets a negative sign.

Formal Definition

- Kendall's τ_K averages the amount of agreement and disagreement by taking the difference between the number of agreement and number of disagreement pairs.
- The formal definition with n data pairs is

$$\tau_K = \frac{\text{Number of agreement pairs} - \text{Number of disagreement pairs}}{\binom{n}{2}}.$$

- Kendall's τ_K is between -1 and 1, and measures dependence's strength (and direction).

First Example

- We will create a dataset called `non_linear_function` with $n = 21$ where:

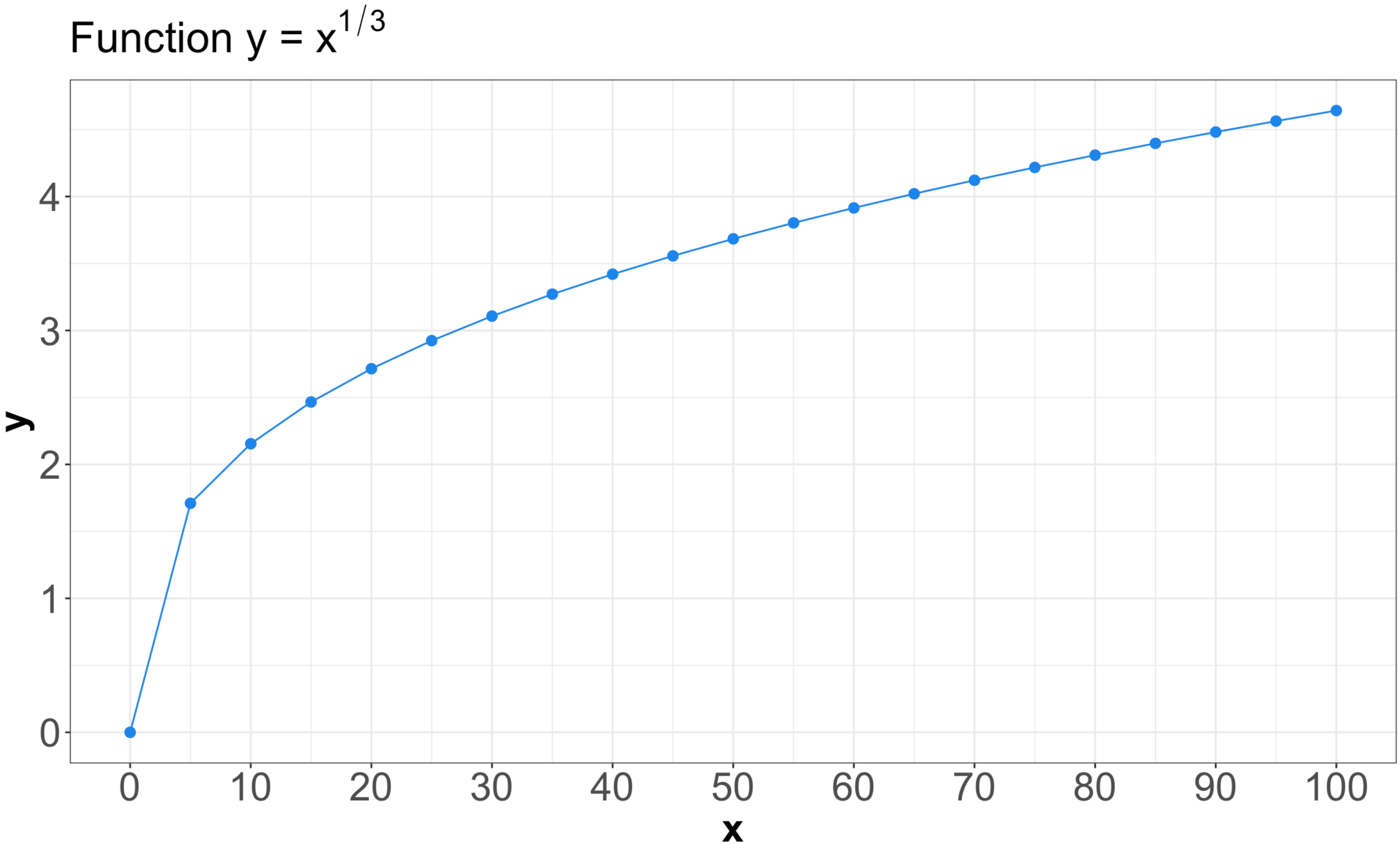
$$y = x^{1/3}.$$

Coding Up `non_linear_function`

R Code | Output

```r
non_linear_pairs <- tibble(
  x = seq(from = 0, to = 100, by = 5),
  y = x^(1 / 3)
)
non_linear_pairs
```

Plotting `non_linear_function`



Computing Correlation Metrics

R Code Output

```r
tribble(
  ~Pearson, ~Kendall,
  round(cor(non_linear_pairs, method = "pearson")[1, 2], 4),
  round(cor(non_linear_pairs, method = "kendall")[1, 2], 4)
) %>%
  knitr::kable(align = "cc")
```

Second Example

- We will create a dataset called `parabola_pairs` with $n = 21$ where:

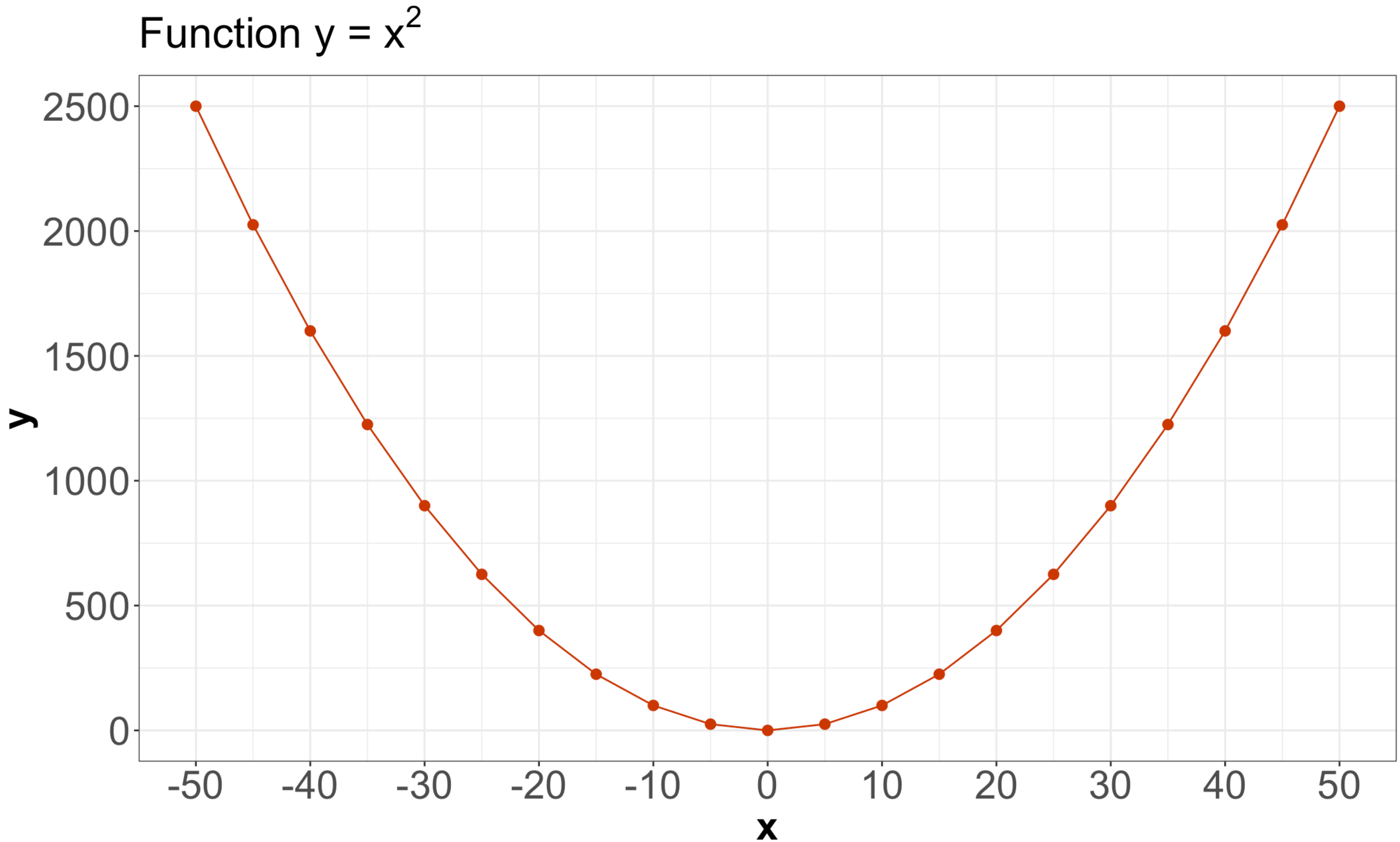
$$y = x^2.$$

Coding Up parabola_pairs

R Code | Output

```r
parabola_pairs <- tibble(
  x = seq(from = -50, to = 50, by = 5),
  y = x^2
)
parabola_pairs
```

Plotting `parabola_pairs`



Computing Correlation Metrics

R Code | Output

```r
tribble(
  ~Pearson, ~Kendall,
  round(cor(parabola_pairs, method = "pearson")[1, 2], 4),
  round(cor(parabola_pairs, method = "kendall")[1, 2], 4)
) %>%
  knitr::kable(align = "cc")
```

- Patterns like a parabola are not monotonically increasing or decreasing.
- Thus, neither Pearson nor Kendall's τ_K will capture the parabola pattern.

2.3. Variance of a Sum Involving Two Non-Independent Random Variables

- Suppose X and Y are not independent random variables.
- We have

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y).$$

If X and Y are independent,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Then,

$$\begin{aligned}\operatorname{Cov}(X, Y) &= \mathbb{E}(XY) - [\mathbb{E}(X)\mathbb{E}(Y)] \\ &= [\mathbb{E}(X)\mathbb{E}(Y)] - [\mathbb{E}(X)\mathbb{E}(Y)] = 0.\end{aligned}$$

Therefore,

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

Today's Learning Goals

You should be able to…

- Calculate **marginal distributions** from a **joint distribution**.
- Describe **independent** RVs and understand the important properties for independent RVs.
- Calculate and describe **covariance**.
- Calculate and describe two mainstream **correlation** metrics: Pearson's correlation and Kendall's τ_K.