Lecture 5

Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/09/14

Slides Updated: 2023-09-26

Agenda

1. A Preview of Multivariate Analysis
2. Marginal and Conditional Probability Distributions
3. Independent Random Variables
4. The EV of a function of RVs
5. Covariance of two RVs

Multivariate Analysis

- First part of course focused on univariate analysis
  - I.e., **one variable**
- We developed tools to summarize one variable
  - Tables, figures, and functions
  - Central tendency and dispersion

Multivariate Analysis

- However, multivariate is helpful to develop a **theory**
  - How can we draw **inferences** about a *population* from a *sample*?

```
require(tidyverse)

df <- read_rds('https://github.com/jbisbee1/PSCI_8356/raw/main/Lectures/Data/sc_debt.Rds')
```

Looking at the data

- Always *always* **always** look at your data!

```
df
```

```
## # A tibble: 2,546 × 16
##    unitid instnm       stabbr grad_…¹ control region preddeg
##     <int> <chr>        <chr>    <int> <chr>   <chr>  <chr>
##  1 100654 Alabama A &… AL       33375 Public  South… Bachel…
##  2 100663 University … AL       22500 Public  South… Bachel…
##  3 100690 Amridge Uni… AL       27334 Private South… Associ…
##  4 100706 University … AL       21607 Public  South… Bachel…
##  5 100724 Alabama Sta… AL       32000 Public  South… Bachel…
##  6 100751 The Univers… AL       23250 Public  South… Bachel…
##  7 100760 Central Ala… AL       12500 Public  South… Associ…
##  8 100812 Athens Stat… AL       19500 Public  South… Bachel…
##  9 100830 Auburn Univ… AL       24826 Public  South… Bachel…
## 10 100858 Auburn Univ… AL       21281 Public  South… Bachel…
## # … with 2,536 more rows, 9 more variables: openadmp <int>,
## #   adm_rate <dbl>, ccbasic <int>, sat_avg <int>,
## #   md_earn_wne_p6 <int>, ugds <int>, costt4_a <int>,
## #   selective <dbl>, research_u <dbl>, and abbreviated
## #   variable name ¹grad_debt_mdn
```

Looking at the data

- What are the **units of observation**?
- What are the **variables**?
  - What is the definition of a variable?

Looking at the data

Name	Definition
unitid	Unit ID
instnm	Institution Name
stabbr	State Abbreviation
grad_debt_mdn	Median Debt of Graduates
control	Control Public or Private
region	Census Region
preddeg	Predominant Degree Offered: Assocates or Bachelors
openadmp	Open Admissions Policy: 1=Yes, 2=No, 3=No 1st time students
adm_rate	Admissions Rate: proportion of applications accepted
ccbasic	Type of institution*
sat_avg	Average SAT scores
md_earn_wne_p6	Average Earnings of Recent Graduates
ugds	Number of undergraduates
costt4_a	Average cost of attendance (tuition-grants)
selective	Institution admits fewer than 10% of applications, 1=Yes, 0=No
research_u	Institution is a research university, 1=Yes, 0=No

*

Looking at data

- As scholars, you probably have several versions of the three fundamental questions buzzing!
    1. What can we say about the data **we have**?
    2. What can we say about the data **we don't have**?
    3. What can we say about the data **we'd expect to see**?

Question 1

- How many schools are selective (have admissions rates less than 10%)?

```
df %>%
  mutate(sel = ifelse(adm_rate < .1,1,0)) %>%
  count(sel)
```

```
## # A tibble: 3 × 2
##     sel     n
##   <dbl> <int>
## 1     0  1563
## 2     1    25
## 3    NA   958
```

Question 2

- What is the average admissions rate for schools in the United States?

```
df %>%
  summarise(avg_adm_rate = mean(adm_rate,na.rm=T))
```

```
## # A tibble: 1 × 1
##   avg_adm_rate
##          <dbl>
## 1        0.679
```

- What do we need to assume in order to believe this result?

Question 3

- If we draw a school at random, what is the probability that it is selective?

```
set.seed(123)
df %>%
  sample_n(size = 1) %>%
  select(adm_rate)
```

```
## # A tibble: 1 × 1
##   adm_rate
##      <dbl>
## 1    0.970
```

Other questions?

- Relationships!
  - Are public universities "better"?
  - Do more selective schools produce grads who make more money?
  - Are more expensive schools more selective?
  - ...?
- These are all questions involving **two variables**
  - Could be more! (Are selective schools in New England more expensive?)
- Welcome to **multivariate analysis**

Theories

- Before turning to the data, it is useful to think about your assumptions
- Are public universities "better"?
  - What do we mean by "better"?
  - What do we think the answer is?
  - **Why** do we think this?

Let's investigate

- Are public universities "better"?
- Look at two variables:
  - `control`
  - `sat_avg`

```
df %>%
  select(control,sat_avg)
```

```
## # A tibble: 2,546 × 2
##    control sat_avg
##    <chr>     <int>
##  1 Public      939
##  2 Public     1234
##  3 Private      NA
##  4 Public     1319
##  5 Public      946
##  6 Public     1261
##  7 Public       NA
##  8 Public       NA
```

Multivariate

- Let's use some tools to look at them more closely

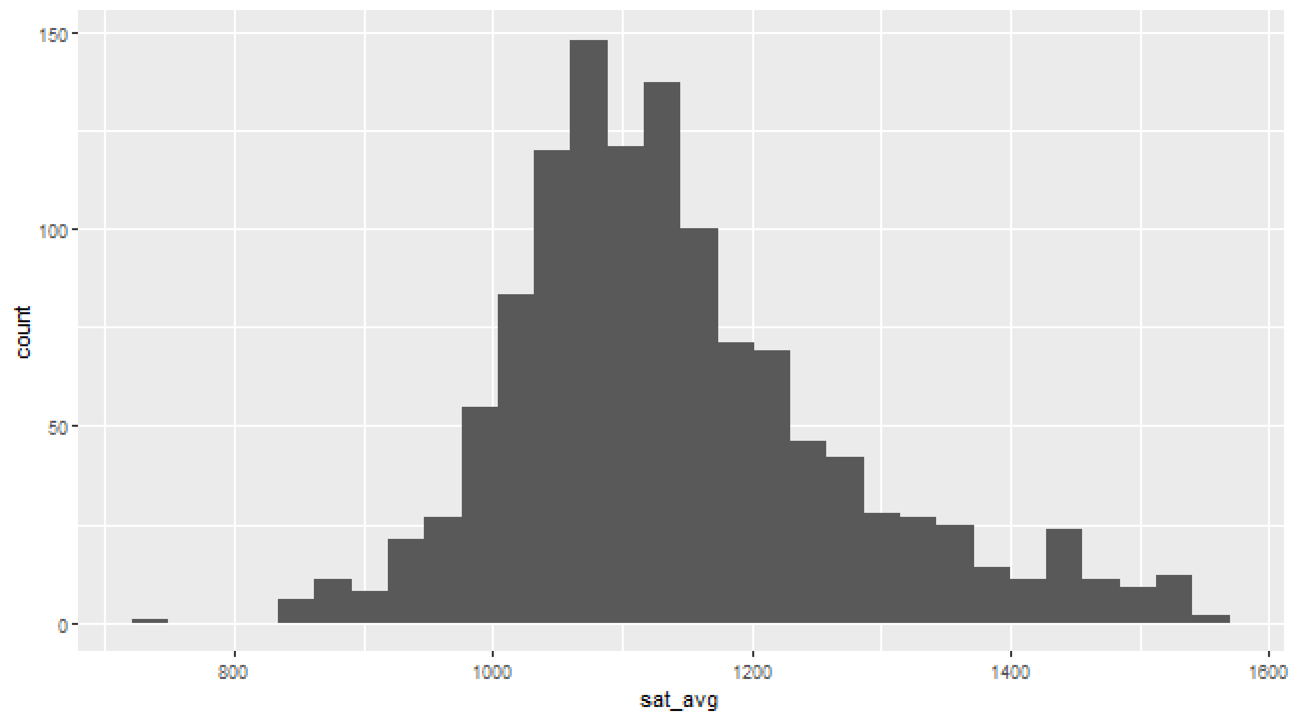
```
df %>%
  count(control)
```

```
## # A tibble: 2 × 2
##   control     n
##   <chr>   <int>
## 1 Private  1320
## 2 Public   1226
```

Multivariate

- Let's use some tools to look at them more closely

```
df %>%
  ggplot(aes(x = sat_avg)) +
  geom_histogram()
```



Multivariate

- How might we begin to answer our research question?

```
df %>%
  group_by(control) %>%
  summarise(avg_sat = mean(sat_avg,na.rm=T))
```

```
## # A tibble: 2 × 2
##   control avg_sat
##   <chr>     <dbl>
## 1 Private   1151.
## 2 Public    1123.
```

- Does this answer our question?

Back to abstract notation!

- To answer, we want a more principled way of talking about relationships
- Running example: hypothetical congressional election
  - GOP has a 73% chance of winning control of the House
  - GOP has an 18% chance of winning control of the Senate
- Two random variables, Y_1 and Y_2, one for each chamber

Multivariate Example

- Y_1 and Y_2 take on the value 1 if the GOP wins control of the associated chamber, and 0 otherwise
  - Refresher: what type of RVs are these?
  - Bernoulli experiments where "success" is GOP winning control
- Denote any particular realization of these RVs as the "ordered pair" (y_1, y_2)
  - $(y_1, y_2) = (y_2, y_1)$ iff $y_1 = y_2$
- What is $P(Y_1 = 1)$? What about $P(Y_2 = 1)$?
  - $P(Y_1 = 1) = 0.73$; $P(Y_2 = 1) = 0.18$

Multivariate Example

- What is the probability that Republicans win both chambers?
- Use set notation! A is the **intersection** of the events $Y_1 = 1$ and $Y_2 = 1$
    - $A = (Y_1 = 1 \cap Y_2 = 1)$
- So what is $P(A)$?
    - 0.73*0.18 = 0.1314?
- Not necessarily! Use the multiplicative law
    - $P(A) = P(Y_1 = 1)P(Y_2 = 1|Y_1 = 1)$
    - If $P(A) = 0.73 * 0.18$, it must be that control of the two chambers are **independent events**
    - Refresh: definition of an independent event?
    - $P(Y_1 = 1 \cap Y_2 = 1) = P(Y_1 = 1)P(Y_2 = 1)$

Joint Probability Distribution

- So...*are* these independent events?
- An example of two independent events

	$Y_1 = 0$	$Y_1 = 1$	**Totals**
$Y_2 = 0$	0.22	0.60	0.82
$Y_2 = 1$	0.05	0.13	0.18
Totals	0.27	0.73	1

- An example of two **dependent events**

	$Y_1 = 0$	$Y_1 = 1$	**Totals**
$Y_2 = 0$	0.25	0.57	0.82
$Y_2 = 1$	0.02	0.16	0.18
Totals	0.27	0.73	1

Joint Probability Distribution

- Why was one **independent** and the other **dependent**?
- In the first table, each cell divided by either the row or column total is the same as the marginal probability (subject to rounding)
  - $0.22/0.27 \approx 0.82$
  - $0.05/0.27 \approx 0.18$
  - $0.22/0.82 \approx 0.27$
  - $0.13/0.18 \approx 0.82$
- In the second table, this relationship **broke**
  - Relate back to the definitions!

Joint Probability Distribution

- Just as we did with univariate probability distributions, **joint probability distributions** are the probabilities associated with all possible values of Y_1 and Y_2
  - Denote as $P(Y_1 = y_1, Y_2 = y_2)$ or just $P(y_1, y_2)$
  - We can imagine these as functions, although in the preceding example, it is easier to just show as a table
- Note that the axioms from the univariate world apply here
  - Axiom 1: $p(y_1, y_2) \geq 0 \; \forall \; y_1, y_2$
  - Axiom 2: $\sum_{y1,y2} p(y_1, y_2) = 1$
- Joint probability distributions can have **distribution functions**
  - $F(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2), \;\; -\infty < y_1 < \infty, -\infty < y_2 < \infty$
  - Often referred to as the **joint cumulative distribution function** or **joint CDF**

Joint CDFs

- For two discrete RVs like in our example, this is $F(y_1, y_2) = \sum_{t_1 \leq y_1} \sum_{t_2 \leq y_2} p(t_1, t_2)$
- For two continuous RVs, we say they are **jointly continuous** if their *joint distribution function is continuous in both arguments*
  - That is, if there exists a nonnegative function $f(y_1, y_2)$ such that:
  - $F(y_1, y_2) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} f(t_1, t_2) dt_2 dt_1$ for $-\infty < y_1 < \infty, \ -\infty < y_2 < \infty$
  - then Y_1 and Y_2 are jointly continuous and the function $f(y_1, y_2)$ is the **joint probability density function** or **joint PDF**

Example

- Let's say we want to calculate the probability that two jointly continuous random variables fall into particular intervals
- $P(a < Y_1 \leq b,\ c < Y_2 \leq d) = \int_c^d \int_a^b f(y_1, y_2) dy_1 dy_2$
- Show that this is equivalent to $F(b, d) - F(b, c) - F(a, d) + F(a, c)$

Marginal Probability Distributions

- NB: all **bivariate** events ($Y_1 = y_1, Y_2 = y_2$) are **mutually exclusive**
- Thus, the **univariate** event $Y_1 = y_1$ can be thought of as the **union** of bivariate events
  - The union is taken *over all possible values for* y_2
- Example: let's roll two 6-sided dice
  - $P(Y_1 = 1) = p(1,1) + p(1,2) + \cdots + p(1,6)$
  - $P(Y_1 = 1) = 6 * \frac{1}{36} = \frac{1}{6}$
- Genericallly: $P(Y_1 = y_1) = \sum_{\forall y_2} p(y_1, y_2)$
- Test: What is the marginal probability for $Y_2 = y_2$?
  - $P(Y_2 = y_2) = \sum_{\forall y_1} p(y_1, y_2)$
- Denote $p_1(y_1)$ as the **marginal probability function** of the *discrete* random variable Y_1

Continuous Case

- **Marginal density function** for continuous RV Y_1 is:
    - $f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_2$
- Test: what is the marginal density function for Y_2?
    - $f_2(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) dy_1$

Conditional Probability Distributions: Discrete

- Recall: $P(A \cap B) = P(A)P(B|A)$ due to the **multiplicative law**
- The bivariate event (y_1, y_2) can be re-written as the **intersection** of two events: $Y_1 = y_1$ and $Y_2 = y_2$
  - Thus: $p(y_1, y_2) = p_1(y_1)p(y_2|y_1)$
  - or $p(y_1, y_2) = p_2(y_2)p(y_1|y_2)$
- NB: $p(y_1|y_2) = P(Y_1 = y_1|Y_2 = y_2)$
  - or $p(y_1|y_2) = \frac{P(Y_1=y_1, Y_2=y_2)}{P(Y_2=y_2)}$
  - or $p(y_1|y_2) = \frac{p(y_1,y_2)}{p_2(y_2)}$ for $p_2(y_2) > 0$ (why?)
- The **conditional distribution function** of Y_1 given $Y_2 = y_2$ is $P(Y_1 \leq y_1|Y_2 = y_2) = F(y_1|y_2)$
- The associated CDF is $f(y_1|y_2) = \frac{f(Y_1,y_2)}{f_2(y_2)}$

Independent Random Variables

- Previous content was hurried in order to bring us here...**how to make inferences from samples**
- Recall that independent events A and B imply $P(A \cap B) = P(A)P(B)$
- Also remember our example of an event involving two random variables: $(a < Y_1 \leq b) \cap (c < Y_2 \leq d)$
    - This event can be **decomposed** to two events: $a < Y_1 \leq b$ and $c < Y_2 \leq d$
- If Y_1 and Y_2 are **independent**, then:
    - $P(a < Y_1 \leq b,\ c < Y_2 \leq d) = P(a < Y_1 \leq b)P(c < Y_2 \leq d)$
- The joint probability of two independent RVs can be written as the **product of their marginal probabilities**

Independent Random Variables

- Generalizing to $F(y_1, y_2) = F_1(y_1)F_2(y_2) \; \forall \; (y_1, y_2)$
  - where $F(y_1, y_2)$ is the joint CDF for Y_1 and Y_2
  - and $F_1(y_1)$ is the CDF for Y_1, and $F_2(y_2)$ is the CDF for Y_2
- Thus, if Y_1 and Y_2 are independent:
  - **Discrete RVs**: $p(y_1, y_2) = p_1(y_1)p_2(y_2)$
  - **Continuous RVs**: $f(y_1, y_2) = f_1(y_1)f_2(y_2)$
- Thus, further, $f(y_1, y_2) = g(y_1)h(y_2)$
  - where $g(\cdot)$ and $h(\cdot)$ are non-negative functions
  - In English, if we want to prove two RVs are independent, we can do so by finding two functions that satisfy these properties

Expectations of functions of RVs

- Recall from the univariate world that we can show the expected value of a function of a random variable $g(Y)$ was
  - **Discrete RVs**: $E[g(Y)] = \sum_y g(y)p(y)$
  - **Continuous RVs**: $E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$
- We can do the same in the multivariate world with a function of several random variables
  - **Discrete**: $E[g(Y_1, Y_2, \ldots, Y_k)] = \sum_{y_k} \ldots \sum_{y_2} \sum_{y_1} g(y_1, y_2, \ldots, y_k)p(y_1, y_2, \ldots, y_k)$
  - **Continuous**: $E[g(Y_1, Y_2, \ldots, Y_k)] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(y_1, y_2, \ldots, y_k)f(y_1, y_2, \ldots, y_k)dy_1 dy_2 \ldots dy_k$

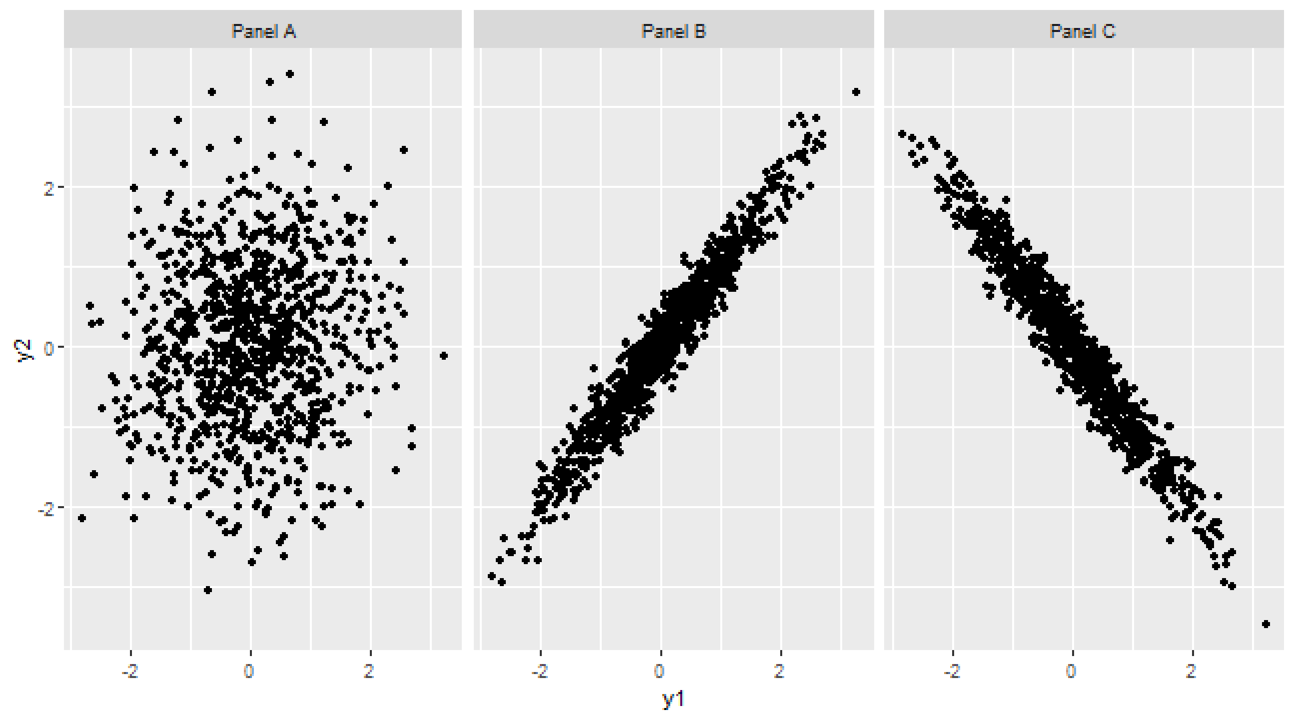
Expectations of functions of RVs

- Rules of expectations also work here
  - Pull out constants: $E[cg(Y_1, Y_2)] = cE[g(Y_1, Y_2)]$
  - Distribute expectations: $E[g_1(Y_1, Y_2) + \cdots + g_k(Y_1, Y_2)] = E[g_1(Y_1, Y_2)] + \cdots + E[g_k(Y_1, Y_2)]$
- These allow a powerful result in which
  - If Y_1 and Y_2 are independent
  - And if $g(Y_1)$ and $h(Y_2)$ are functions of only Y_1 and Y_2
  - Then $E[g(Y_1)h(Y_2)] = E[g(Y_1)]E[h(Y_2)]$

Covariance of Two RVs

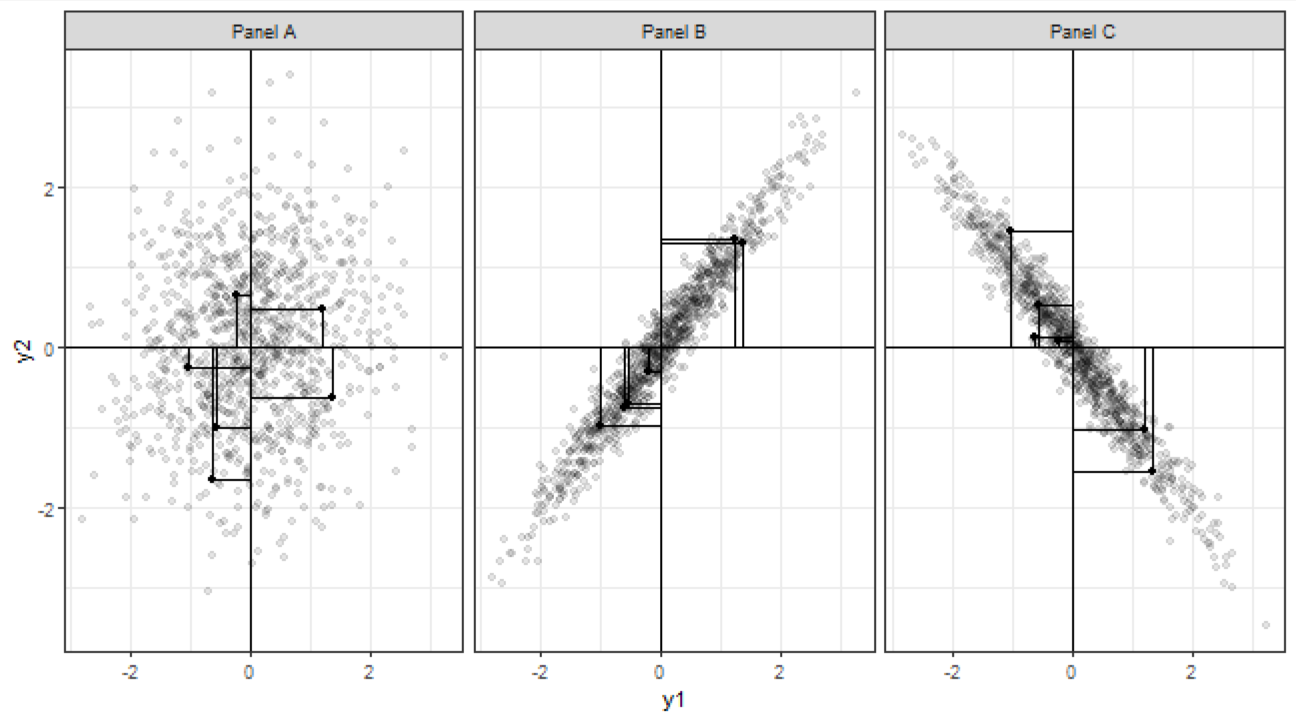
- If we say that Y_1 and Y_2 are **independent**, we are saying
    - **Discrete**: joint probability is equal to *the product of their individual probability functions*
    - **Continuous**: joint PDF is equal to *the product of their individual PDFs*
- But what if Y_1 and Y_2 **are** related?
    - That is, given what we know about the value of Y_1, we can make better than a random guess about Y_2
- We can **describe** how much the two processes are related with the property of **covariance**
    - $COV(Y_1, Y_2) \equiv E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$

Examples



Covariance

- Let's think about two quantities: $(y_1 - \mu_1)$ and $(y_2 - \mu_2)$



Covariance

- Think through what these lines represent
  - How much a randomly chosen point **deviates** from its mean
- Note two patterns from the points chosen in each panel
  - In panel A: bigger deviations in y_1 are sometimes associated with bigger deviations in y_2, but not always
  - In panel A: in some cases the y_1 deviation is positive and the y_2 deviation is negative, but not always
  - In panels B and C: bigger deviations in y_1 are consistently associated with bigger deviations in y_2
  - In panel B: positive deviations in y_1 are associated with positive deviations in y_2, and negative deviations in y_1 are associated with negative deviations in y_2
  - In panel C: positive deviations in y_1 are associated with negative deviations in y_2, and vice versa

Covariance

- How can we summarize these conclusions more efficiently? Take the product of the y_1 and y_2 deviations
  - $(y_1 - \mu_1)(y_2 - \mu_2)$
  - In panel A, this product is sometimes positive and sometimes negative
  - In panel B, this product is always positive
  - In panel C, this product is always negative
- And how can we **further** summarize these conclusions?
  - Take the **expectation**!
  - $COV(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$

Covariance

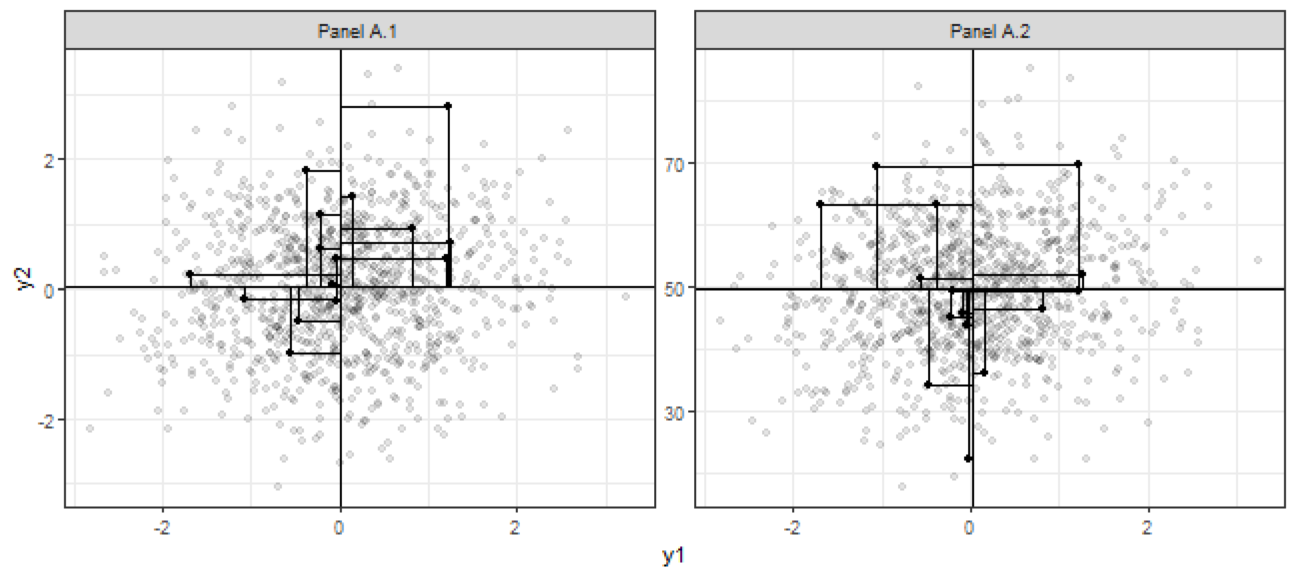
- Let's calculate!

```
toplot %>%
  group_by(facet) %>%
  summarize(cov = mean((y1-mean(y1))*(y2-mean(y2))))
```

```
## # A tibble: 3 × 2
##   facet       cov
##   <chr>     <dbl>
## 1 Panel A  0.0865
## 2 Panel B  0.978
## 3 Panel C -0.983
```

Covariance

- But what if we change the scale?




```
res <- toplot2 %>%
  group_by(facet) %>%
  summarize(cov = mean((y1-mean(y1))*(y2-
mean(y2))))
```

```
## # A tibble: 2 × 2
##   facet          cov
##   <chr>        <dbl>
## 1 Panel A.1 0.0865
## 2 Panel A.2 1.30
```

Correlation

- We need to make this *scale invariant*
- **Standardize** by the product of the two RVs' standard deviations
  - $\rho(Y_1, Y_2) = \frac{COV(Y_1, Y_2)}{\sigma_1 \sigma_2}$
- Can you prove that $-1 \leq \rho \leq 1$?
- Summing up:
  - Independence of Y_1 and Y_2 implies that $COV(Y_1, Y_2) \approx 0$
  - Or more accurately, $\rho(Y_1, Y_2) \approx 0$
- NB: these are useful tools for measuring the strength of a *linear* relationship
  - Not so good for other types of relationships, like **curvelinear**