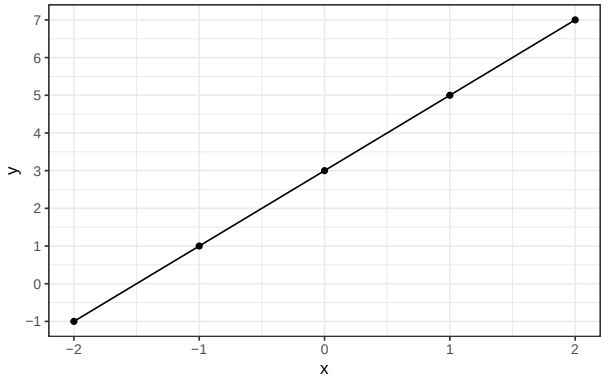
Lecture 2

Review: Linear Equations

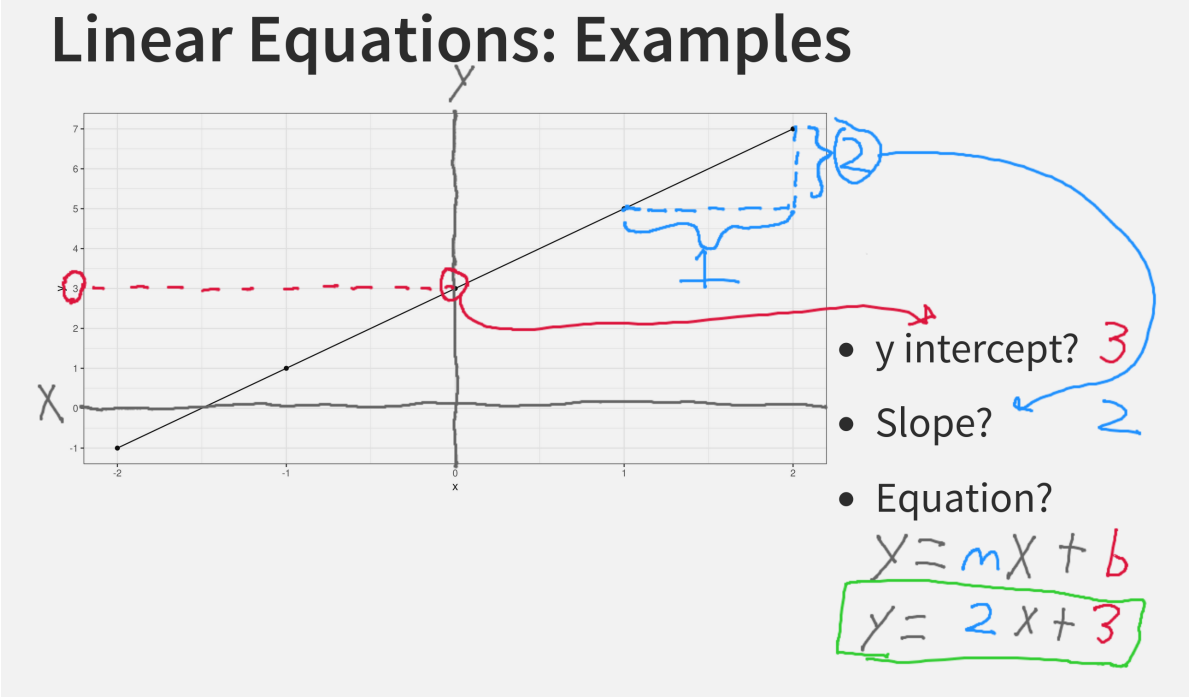
- In 1 dimension given by $y = mx + b$
- x is the independent variable
- y is the dependent variable
- m is the slope
- b is the y intercept
- (-b/m) is the x intercept, but usually not important in econometrics

Linear Equations: Examples



- y intercept?
- Slope?
- Equation?

Linear Equation Solution



Slope and intercept interpretation

You rent a car. You estimate the cost per day to rent and operate the car is given by the equation

- $Price = 30 + 0.2 * miles$

Interpret the slope and y intercept in words.

NOTE: "Interpret" in this class always refers to a basic explanation of a formula or coefficient which directly uses the numeric values and that a (reasonably) normal person can understand

Slope and intercept interpretation - Answer

$Price = 30 + 0.2 * miles$

The generic interpretation of the y intercept is the value of y when x=0.

Substituting in our variable names, this is the price paid when no miles are driven

Settings miles=0 in the formula we get a price of 30.

Putting into normal language, our car costs $30/day to rent when we don't drive at all.

Slope and intercept interpretation - Answer

$Price = 30 + 0.2 * miles$

Generically, the slope is how much y increases for each 1 unit increase in x.

Here x is in miles and y is in dollars. So every additional mile we drive the car increases the cost by 0.2 dollars.

Stated like a functioning human, each additional mile we drive the car costs twenty cents.

Slope and intercept interpretation - Answer

If we put these together, there's a fixed cost of $30/day to rent the car, plus we pay 20 cents per mile to operate it. This is sufficient for interpretation. If we want to get fancy, $30 is our fixed cost and 20 cents is our variable cost.

Linear Equations: Higher Dimensions

- In multiple dimensions will be given by $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$
  - β_0 is the intercept, y is the dependent variable
  - Now we have k dependent variables, x_k, each with their own slope β_k

(Optional) Linear transformations as a matrix

- In matrix notation: $Y = X\beta$
  - Y is an n by 1 column vector (where n individuals are observed)
  - X is an n by k matrix
  - β is a k by 1 column vector (where k parameters are used)
- The definition of matrix multiplication makes this work out perfectly

(Optional) Matrix multiplication Example

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

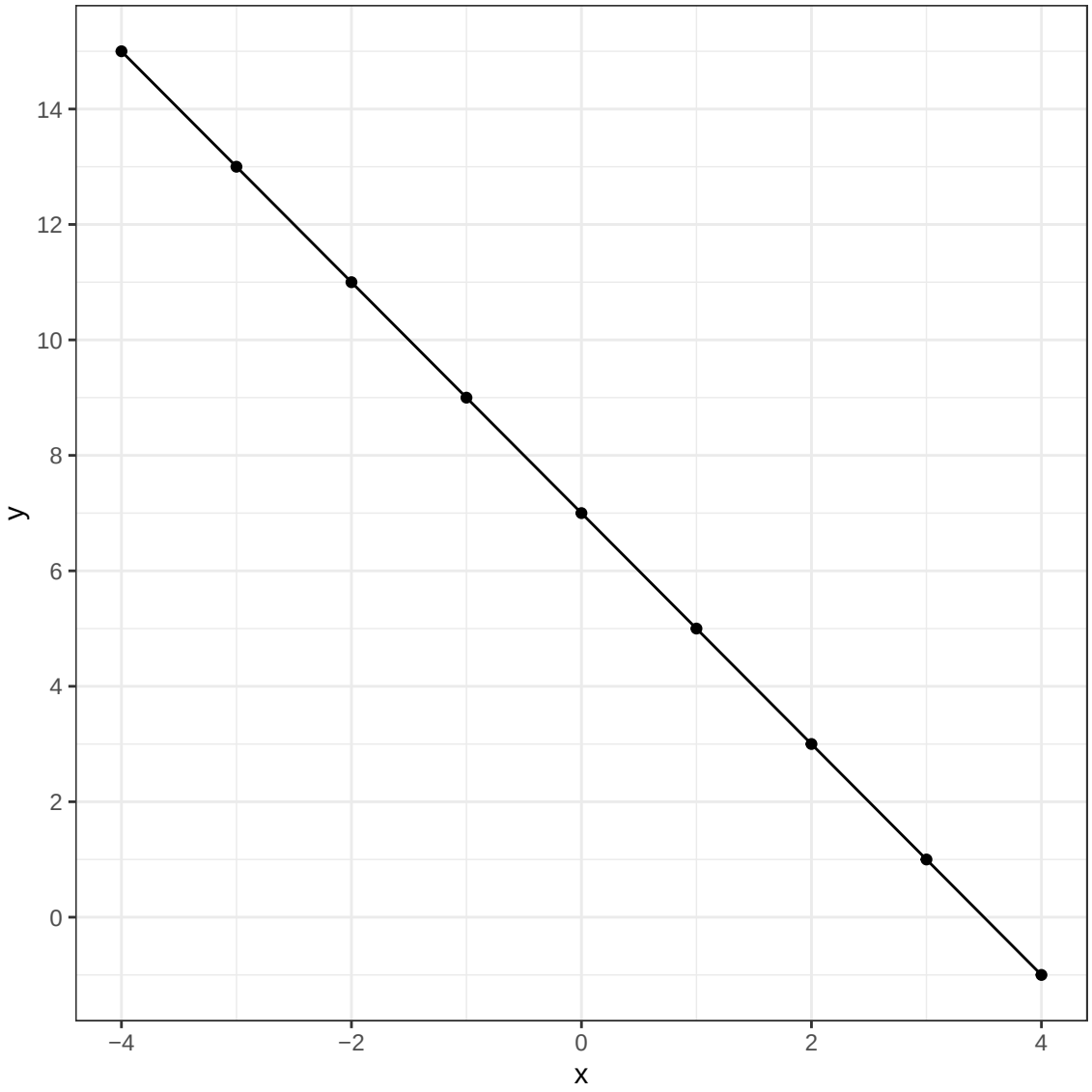
$$\begin{bmatrix} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} \\ y_3 = \beta_0 + \beta_1 x_{13} + \beta_2 x_{23} \end{bmatrix}$$

where, e.g. x_{21} is x_2 for individual 1

Dropping subscripts: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Higher Dimensions: Example

Linear Equations: Question



- What is the formula for this line?

Answer:

- $y = mx + b$
- $(0, 7)$ gives y intercept: $b = 7$

- $(0, 7) \rightarrow (1, 5)$ gives slope: $m = -2$ (we can use any 2 points by taking $\frac{\Delta y}{\Delta x}$)
- $y = -2x + 7$

Stats Review

Data types

The names/categories aren't important to know, but the interpretation is. This will show up frequently in class :

- **Interval**: a 1 unit change is the same everywhere (dollars)
- **Ordinal**: higher numbers are better, but the difference isn't easily interpretable (e.g. you code 0 for a HS dropout, 1 for HS degree, 2 for some college, 3 for bachelor's degree, etc)
- **Categorical**: the number is just an identifier with no numeric meaning (zip code)

- A 10% increase in revenue has a very different interpretation from a 10% increase in education or zip code.

Data types - a binary variable

You have data on employees at your company. You find the following average relationship between gender (0 for male and 1 for female) and height (in cm):

$height = 176 - 14 * gender$

Interpret the slope of -14 in this context

Binary variable - answer

a "1 unit increase" in gender means changing gender from 0 to 1 since there are only two possible values. For this encoding, this means going from male to female. If we want to interpret this properly, it means that the average female employee at the company is 14 centimeters shorter than the average male.

The intercept of 176 is the average height for males in the company in cm.

Data strutures

- **Cross-sectional**: A snapshot of data at a given time.Students' grades in this classroom at end of semester.
- **Time Series**: Study 1 unit over time. Your grade every week in this class.
- **Panel Data** : Combination of the above two. Each student's grade every week.

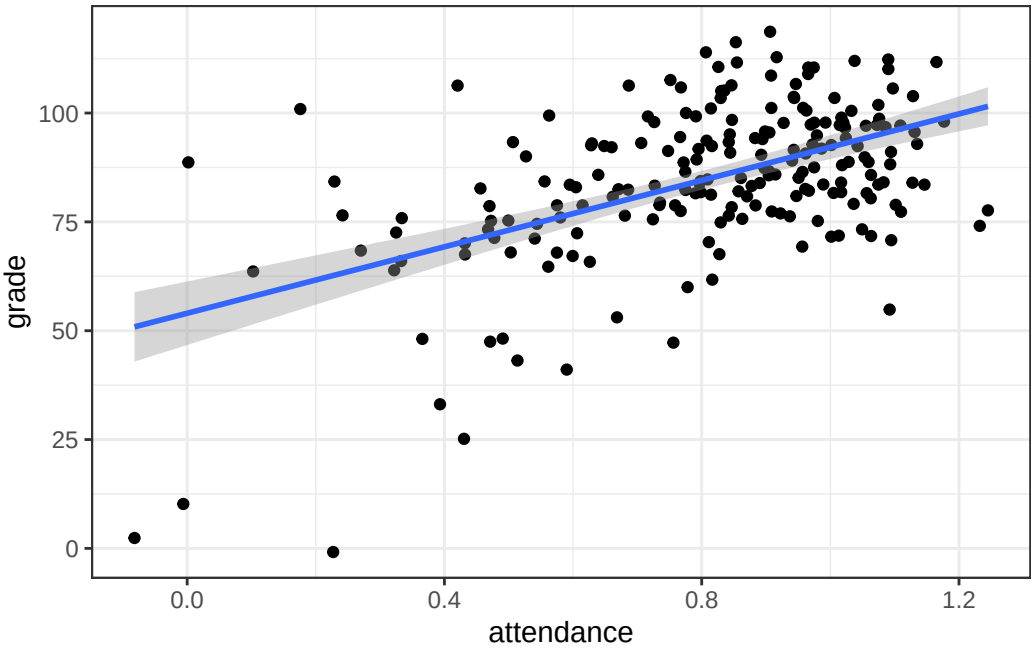
- There are other distinctions based on whether the same individuals are followed over time
- These will determine the subscripts of our regression equations later in the course

Data structures: Cross Section

Attendance vs grades, simulated from anonymized data

```
     student  baseline attendance     grade
  1:       1 0.9507893  0.9657565 110.46481
  2:       2 0.7784678  1.0628186  80.41383
  3:       3 1.0457980  1.0900670 112.30689
  4:       4 1.1704702  0.8975915  87.35327
  5:       5 0.9979116  0.3249460  72.55599
 ---
205:     205 1.0530886  1.0166823  84.04216
206:     206 1.0585210  0.4560729  82.68328
207:     207 1.0844113  0.8926517  90.41293
208:     208 1.0179260  1.2323755  74.08424
209:     209 0.6371130  0.8295373  74.87772
```

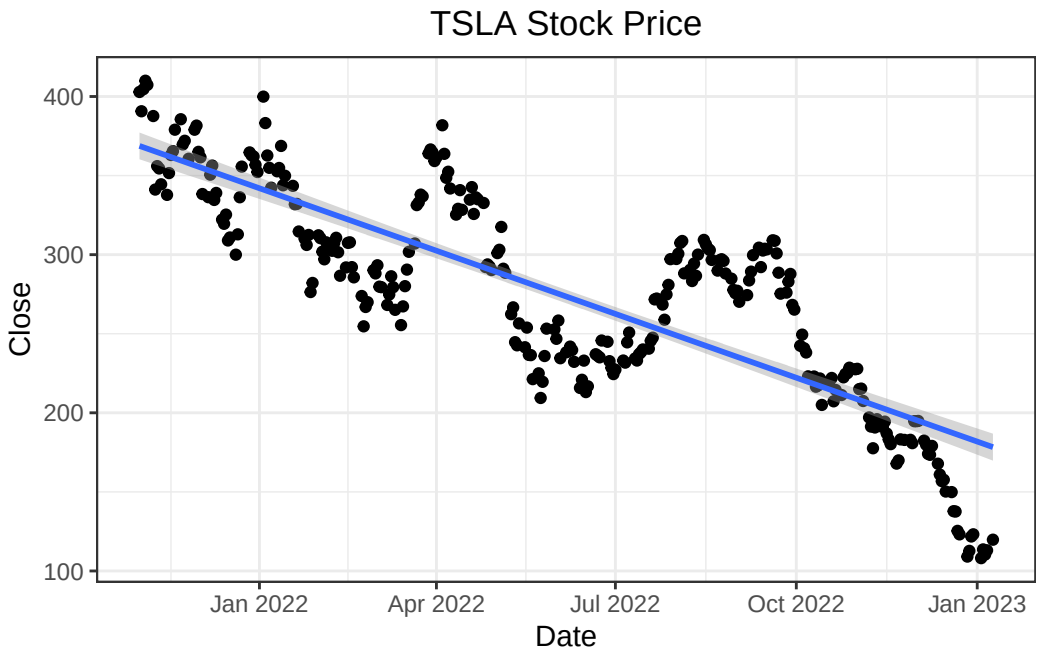
Data Structures: Cross Section Plot



Data Structures: Time Series

```
           Date    Close
  1: 2021-11-01 402.8633
  2: 2021-11-02 390.6667
  3: 2021-11-03 404.6200
  4: 2021-11-04 409.9700
  5: 2021-11-05 407.3633
 ---
295: 2023-01-03 108.1000
296: 2023-01-04 113.6400
297: 2023-01-05 110.3400
298: 2023-01-06 113.0600
299: 2023-01-09 119.7700
```

Data Structures: Time Series Screenshot



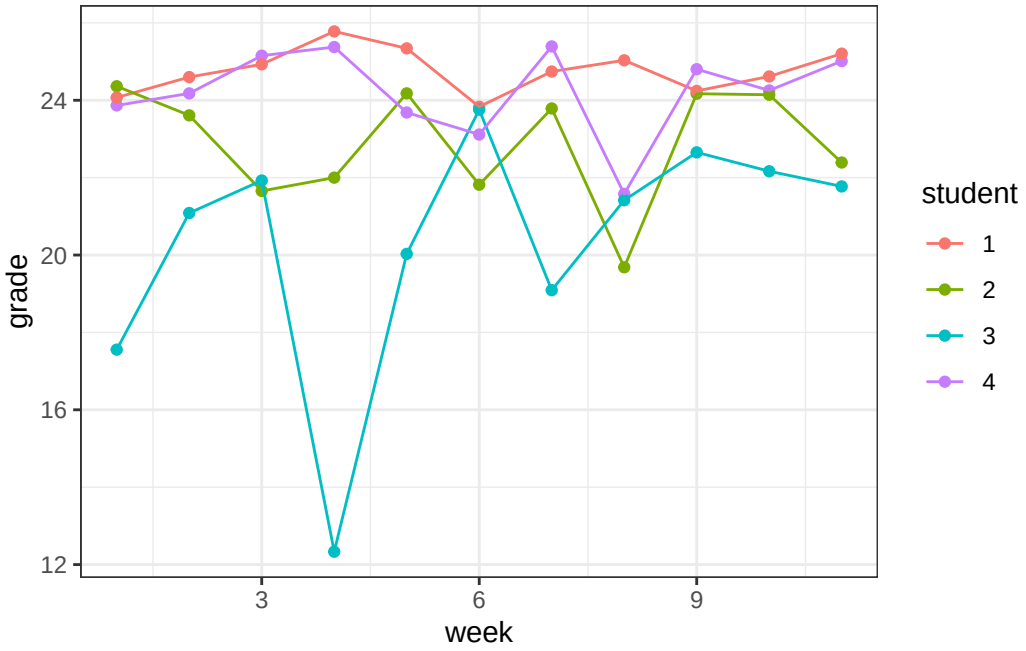
Data Structures: Panel Screenshot

Data is simulated from actual anonymized data

	student	week	attendance	grade	Roster	TA	Reason
1:	159	11	0	25.73623	3	ThatOtherGuy	
2:	80	6	1	19.49742	6	ThatGuy	
3:	138	5	1	24.64185	3	ThatOtherGuy	
4:	64	5	1	25.76792	3	ThatOtherGuy	
5:	208	2	0	21.76407	3	ThatOtherGuy	Holiday

2306:	124	7	0	24.60940	6	ThatGuy	
2307:	31	3	1	22.27998	1	ThatGuy	
2308:	128	3	1	23.49010	1	ThatGuy	
2309:	148	9	1	21.88609	2	ThatOtherGuy	
2310:	188	4	0	23.07566	1	ThatGuy	Canceled

Panel Data



Populations vs Samples

- When conducting research we need to define the population we're interested in
  - For average salary by educational attainment, do we include 16 year olds? 70 year olds? Unemployed individuals? Part-time?
- For a **population**, a measure of interest is called a **parameter** (average income in the US)
- When data is limited we need to estimate this **statistic** using a **sample**

Populations vs Samples

- mean individual income in the US is \$57,143 in 2021 (from FRED St Louis).
- If we randomly sample 1000 individuals we may end up getting an average of \$51,000 in this specific sample.
- Here $\mu = 57143, \hat{\mu} = 51000$
  - A hat is used to indicate an estimate
- parameters are true values, statistics are estimates

Common Sample Statistics: mean

- An individual observation of outcome x is labeled as $x_i,\ i = 1, 2, ..., n$
- You could also label them, e.g. $income_{sam}, income_{fred}$, etc.
- Mean: $E[X] = \mu$ for population, $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$ for sample
  - E[] refers to the expectation, or weighted average, of a random variable
- The mean is the average, or center, of the distribution

Sample Statistics Question:

```
    x
1:  1
2:  5
3: -7
4: 12
5: 15
```

- $\bar{x} = ?$

Sample Statistics Answer

- $x_1 = 1, x_2 = 5, x_3 = -7, x_4 = 12, x_5 = 15$
- $\sum x_i = x_1 + x_2 + ... + x_5 = 26$
- $n = 5 \implies \bar{x} = 26/5 = 5.2$
  - If this is a population we call this $\mu = E[X]$ instead of $\bar{x}$

Common Sample Statistics: Variance

- Variance: $\sigma^2 = E[(x - \mu)^2]$ for population
  - $= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n}$
  - Where $\bar{x}$ was calculated as before
- This is the average squared distance from the mean
  - Why squared?
- This measures the dispersion, or spread, of a distribution, rather than the center

Common Sample Statistics: Variance

- For Samples: $s^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2}{n-1}$ for sample
- We divide by n-1 instead of n to make the result unbiased. This only matters for small samples.
- Unbiased means that on average $s^2 = \sigma^2$ (ie if we repeat an experiment many times)
- Standard deviation: $\sigma = \sqrt{\sigma}^2, s = \sqrt{s}^2$

Calculating Standard Deviation

```
   x
1:  1
2:  5
3: -7
4: 12
5: 15
```

- Same dataset, but now we want to calculate standard deviation. Steps?

Calculating Standard Deviation

- First calculate variance, the average (squared) distance from the mean:
  - subtract $\bar{x}$ from each observation
  - Square the result
  - Take the mean of this result

Calculating Standard Deviation: Tabular Calculation

```
   x xbar difference squaredDifference
1:  1  5.2       -4.2             17.64
2:  5  5.2       -0.2              0.04
3: -7  5.2      -12.2            148.84
4: 12  5.2        6.8             46.24
5: 15  5.2        9.8             96.04
```

```
[1] 308.8
```

- This is the sum of squared differences. To get the variance divide by $n = 5$ if a population (σ^2), or $n = 4$ if a sample (s^2)
  - $\sigma^2 = 61.76, s^2 = 77.2$

Calculating Standard Deviation: Tabular Calculation

- Finally, squared units are weird, so take the square root to get the standard deviation (σ or s)
  - $\sigma = 7.86, s = 8.79$

A note on higher moments

- The information in the variance and standard deviation is captured in $E[X^2]$, the second moment.
- $E[X^3]$ gives the skewness of a distribution
- $E[X^4]$ gives the tail weight (kurtosis)
- Given $E[X], E[X^2], ...E[X^n]$ for n from 1 to ∞ fully specifies all "well-behaved" distributions
  - Analogous to Taylor's theorem

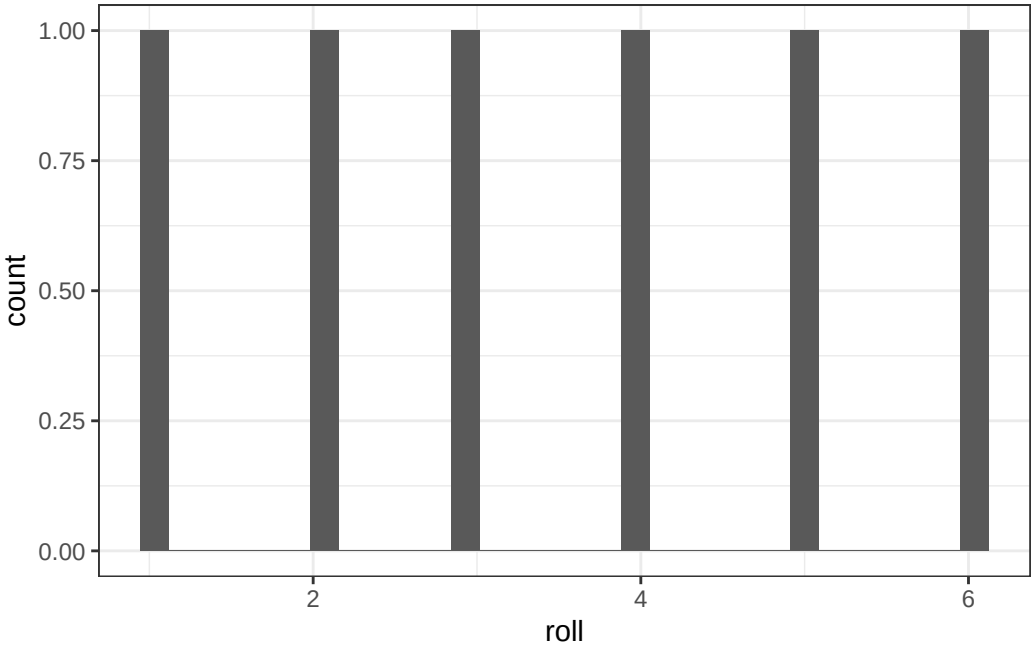
Random Variables

- A random variable is used to represent a random event
- The variable itself is typically denoted as a capital letter, and the outcome a lowercase one
  - X is the roll of a die. $X = 5$ means a specific roll was a 5. $X = x$ means a specific roll was x
- Random Variable Operations:
  - Measure the probability of a specific event: $P(X = 5)$
  - Measure the average value $E[X]$
  - Transform them: X^2 is the value of a squared dice roll. $X + Y$ is the sum of 2 dice

Random Variable Representation: Dice

- Suppose we toss a six-sided die. How can we represent this event mathematically?
- We can specify all possible outcome and their associated probability
- outcomes: {1,2,3,4,5,6}
- probabilities: $\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\}$
- How to visualize?

Random Variable Visualization: Histograms



Discrete Random Variables

- If a random variable has a countable number of outcomes (like a dice roll or a coin flip) it is called discrete
  - The function that generates the histogram is called the probability mass function, $p(x)$. $\sum p(x) = 1$
- Two events are independent if their joint probability is the product of their individual probabilities
  - $P(X \cap Y) = P(X) * P(Y)$
  - Means that knowing the value of one random variable gives us no information about the other
  - **Extremely** strong assumption!

Adding random variables: Dice

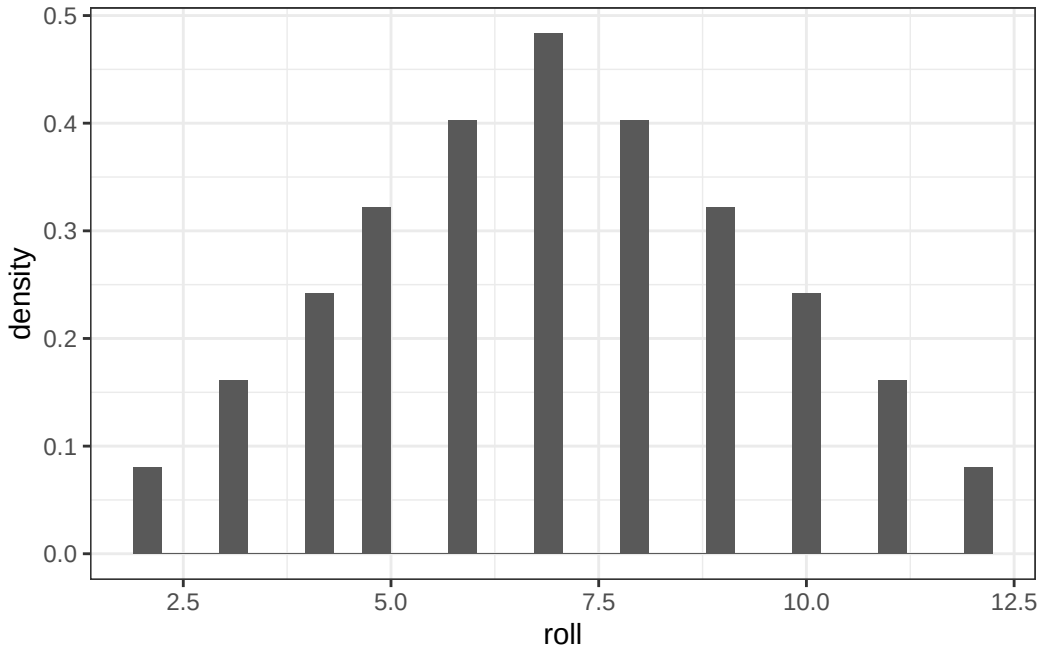
- A roll of a 6 sided die is a random variable. $p(x) = \frac{1}{6}, x \leq 6 \;\; x \in \mathbb{N}$
- If X is the event of a roll of 1 die, and Y of a second die, then $X + Y$ gives the probability associated with the outcome of rolling two dice
- What is the probability that $X + Y = 7$?
- Calculation is called a convolution. Notice the symmetry!

Calculating Standard Deviation: Tabular Calculation

- We can list out all possible permutations of the two dice by using a table
- Because of independence, the probability of each permutation is identical and equal to $\frac{1}{6} * \frac{1}{6} = \frac{1}{36}$
- The table below gives the value of $X + Y$ for the corresponding values of X and Y:

- to find $P(X = 7)$ just count the number of 7's in the table at divide by 36. Note that they're all on the diagonal (as is every other possible value)
- Formula is convoluted (hence the name convolution): $p(x = 7) = \sum_{x=1}^{x=6} p(x) * p(7 - x)$

Adding random variables: histogram



Adding random variables: Expectation

- Adding random variables is hard, but their mean is easy to calculate
  - Weighted average: $E[X] = \sum xp(x)$
- Single die has mean $\frac{1+2+3+4+5+6}{6} = 3.5$
- Expectation is linear: $E[\sum \alpha X_i] = \alpha \sum E[X_i]$
- Two dice $(X_1 + X_2)$ has mean $3.5 + 3.5 = 7$

Adding random variables: Variance

- For independent events, $var(X+Y) = \sigma^2_{x+y} = \sigma^2_x + \sigma^2_y$
- but $\sigma_{x+y} \neq \sigma_x + \sigma_y$
  - and $\sigma^2_{\alpha x} = \alpha^2 \sigma^2_x$
- In general, $\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y + 2\rho\sigma_x\sigma_y$
  - ρ is the correlation coefficient

Continuous Random Variables

- Variables that take uncountable values are called continuous random variables (e.g. height, distance)
  - The function that generates the histogram is called the probability density function, $f(x)$. $\int_{-\infty}^{\infty} f(x)dx = 1$
  - the value at a single point does not give a probability. We need to take the area under the curve
- The function $F(x) = P(X \leq x)$ is called the cumulative distribution function (CDF). It is $\int_{-\infty}^{x} f(t)dt$

Adding random variables: Normal Distribution

- General formula for adding two random variables is a nightmare (don't memorize): $f_{x+y}(z) = \int_{-\infty}^{\infty}(\int_{-\infty}^{t} f(z-x)dx)f(t)dt$
- One random variable let's us easily calculate a mean (and sum) of random variables:
- $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{(\frac{x-\mu}{\sigma})^2}{2}}$
  - This is a **normal** random variable with mean μ and standard deviation σ

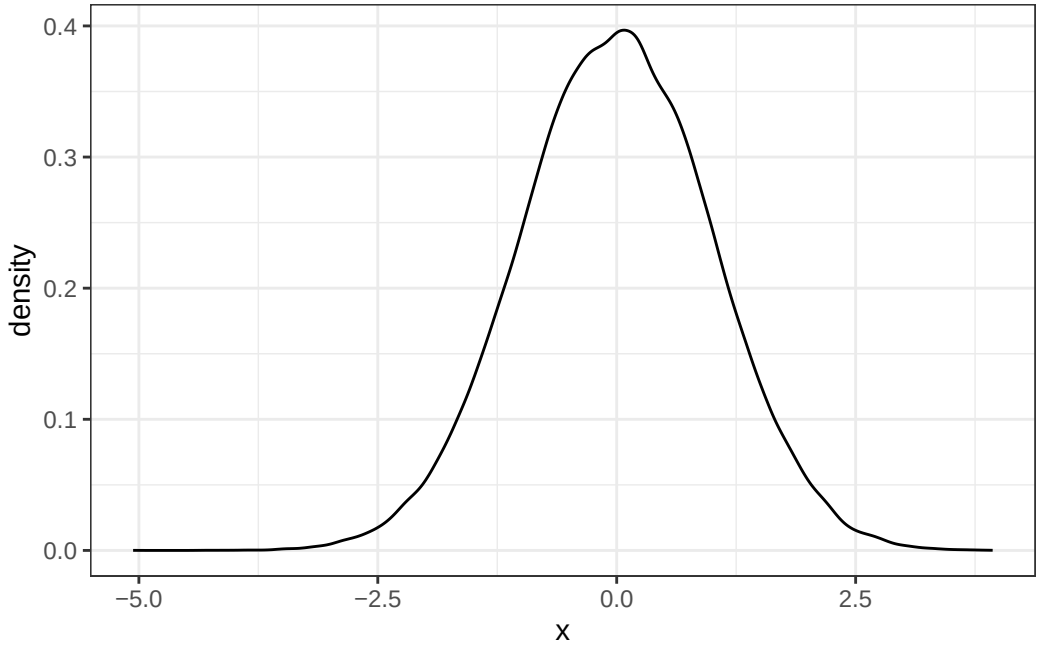
Adding random variables: Normal Distribution

- If X is normal we write this as $X \sim N(\mu, \sigma)$
- $\frac{X_1+X_2+\ldots+X_n}{n} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
  - Almost no other distributions have this property ("closed under convolutions")

Normal Distribution Usage

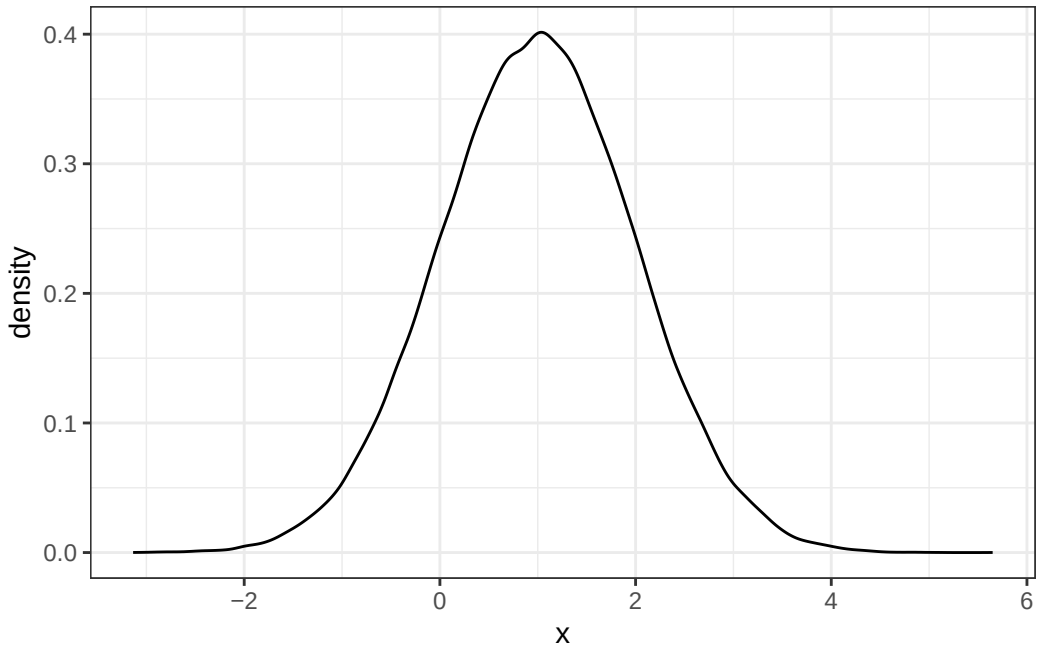
- Given a normal distribution with mean 0 and standard deviation 1, what is the probability that $-1 < x < 1$?
- Calculate using the area under the curve from -1 to 1
- Can't do by hand. Use either a table or the function pnorm in R ($\Phi(z)$)
- 68% of data lies between -1 and +1 standard deviation (=pnorm(1)-pnorm(-1))
- 95% of data is between 2 standard deviations of mean, and 99.7% within 3 standard deviations

Normal distribution plot

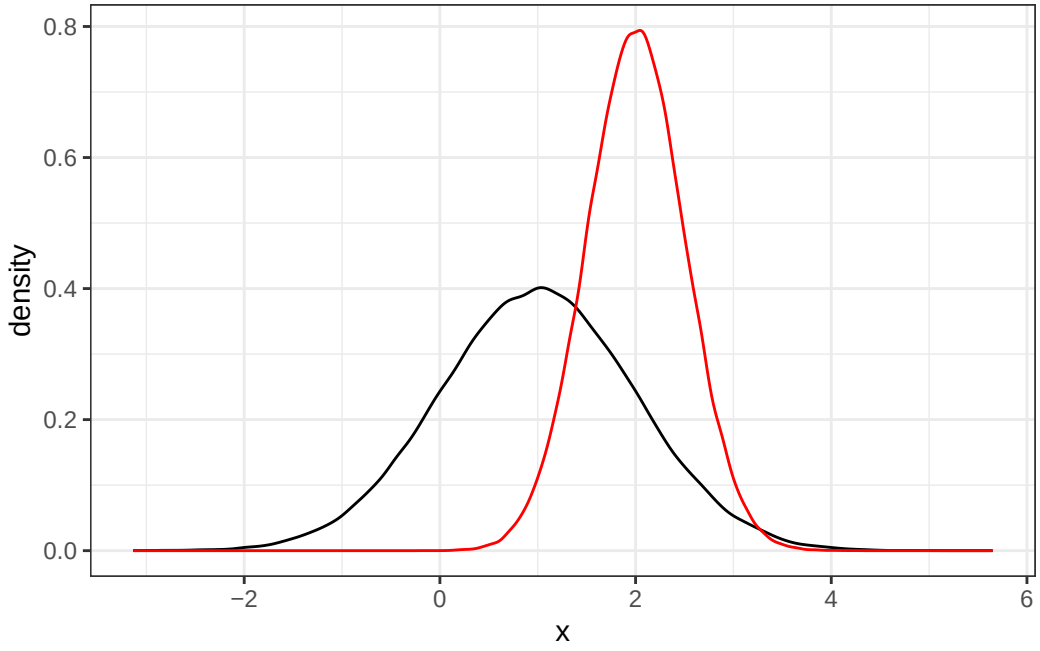


Interpreting Mean and SD

The following is a normal distribution with mean 1 and standard deviation 1. What will the graph look like if instead we have mean 2 and standard deviation 1/2?



Interpreting Mean and SD - Result



Random variables: Population vs Sampling: graph

https://mcfortran.shinyapps.io/sampling/

Sample Statistics as Random variables

- Our sample statistic, e.g. $\bar{x}$ will change every time we use a different sample
- Each individual observation, x_i is a random variable. Our sample statistic is then a transformation of this vector of x_is: $\bar{x} = (x_1 + x_2 + ...x_n)/n$
- The sampling distribution described above will be different from the population distribution (each x_i is just a random draw from the population distribution)

Sampling Moments

- Sampling distributions have their own mean and variance (and other moments)
- $E[\bar{x}] = \mu$
- $var(\bar{X}) = \frac{var(X)}{n} = \frac{\sigma^2}{n}$
  - $sd(\bar{x}) = \frac{\sigma}{\sqrt{n}}$
- How to calculate the probability distribution?

Sampling Moments

- as $n \to \infty$ $\bar{x}$ approaches a distribution with mean μ and variance 0
- We can state this a bit more precisely using the central limit theorem
- **Caveat**: we have assumed our samples are drawn **independently**. If observations are related to each other this is violated

Administrative Miscellanea

- Homework 2 due next Wednesday before class
- Problem set 1 due Friday at midnight
- Quiz 1 next Wednesday in class
- Finish review, start with bivariate OLS today
- I'll skip on programming lab 2 that was scheduled for Monday
  - Just make sure you're set up for the first problem set
  - Please come to office hours (or schedule time) if having issues

Central Limit Theorem

- If we have independent, identically distributed (iid) random samples from a population with finite variance the Central Limit Theorem applies:
- $lim_{n\to\infty}\bar{x} = N(\mu, \frac{\sigma}{\sqrt{n}})$
- Once samples get large our **sampling** distribution becomes normally distributed with declining variance, regardless of the shape of our population distribution
- Many consider this to be the most beautiful theorem in mathematics

Central Limit Theorem - Illustration

https://mcfortran.shinyapps.io/sampling/

Central Limit Theorem - Intuition 1

- When we added dice, we had natural symmetry. If our distributions are **identical** this will always arise
  - And if they are **independent** then we just need to multiply the probabilities piecewise
  - So i.i.d. distributions should result in this symmetry
  - Outliers become exponentially more unlikely and smoothed around the center

Central Limit Theorem - Intuition 2

- Moments ($E[X^n]$) had easy math. So calculate every moment and translate this back to probabilities
  - This is done with a moment generating function: the Laplace transform of the probability function
  - $M_X(t) = E[e^{tx}] \implies E[X^n] = \frac{\partial^n M_X(t)}{\partial t^n}|_0$
    - Far beyond this course, but a common technique in graduate level math
- This is how you formally prove the central limit theorem - The third moments and higher all decline very quickly, leaving only the mean and variance

Central Limit Theorem - Intuition 3

- The normal distribution stays normal when taking averages
  - The normal distribution is the only **finite variance** distribution with this property
  - So any non-normal distribution will end up gravitating towards a normal distribution
  - Formula: e^{-x^2} is its own Fourier transform (related to the Laplace transform)
    - The fourier transform makes convolutions easier to calculate. See convolution theorem.

Cross Moments: covariance and correlation

- For two random variables (or columns of data) the average product of the two, $E[XY]$, captures important information
- Like variance, we subtract out the mean to make it more interpretable

Cross Moments: covariance and correlation

- We define $cov(X, Y) = E[(x - \mu_x)(y - \mu_y)]$
- If X and Y are both above their mean, the product is positive. If they're both below the mean it's also positive
- If one value is above the mean and the other is below their product is negative
- This then gives a measure of how closely associates X and Y are. If they move in the same direction it is positive, opposite directions is negative. "Independent" is zero

Cross Moments: covariance and correlation

- As long as $cov(X, Y) \neq 0$, if we learn the value of X, we also learn something about Y
  - If X and Y are independent, the covariance is 0. The reverse is not true though!

Cross moments: correlation

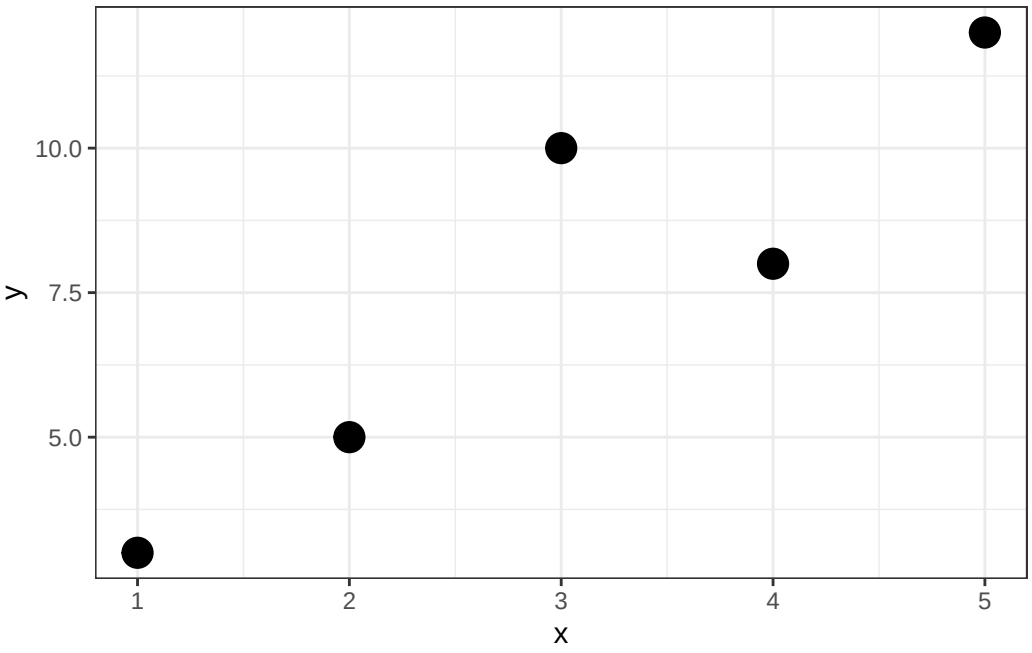
- Standardize further: standardize covariance to be in the range $[-1, 1]$ by dividing by $\sigma_x \sigma_y$. Called correlation
- $\rho \equiv cov(X, Y)/(\sigma_x \sigma_y)$
- The sample statistic for ρ is usually called r instead of $\hat{\rho}$

Correlation: visualization

https://mcfortran.shinyapps.io/correlation/

Correlation: Calculation

```
   x  y
1: 1  3
2: 2  5
3: 3 10
4: 4  8
5: 5 12
```



- Is $r > 0, r < 0$, or, $r = 0$?

Correlation: Calculation

- Steps to calculate?
- $cov(x, y) = E[(x - \bar{x})(y - \bar{y})]$:
  - Calculate the mean of x and of y
  - subtract the mean from each observation
  - multiply the two results
  - take the average
- Once we have covariance, standardize to get r (or ρ)
  - $cov(x, y)/(\sigma_x \sigma_y)$
  - σ calculated as before: subtract out the mean from each observation, square it, and take the average

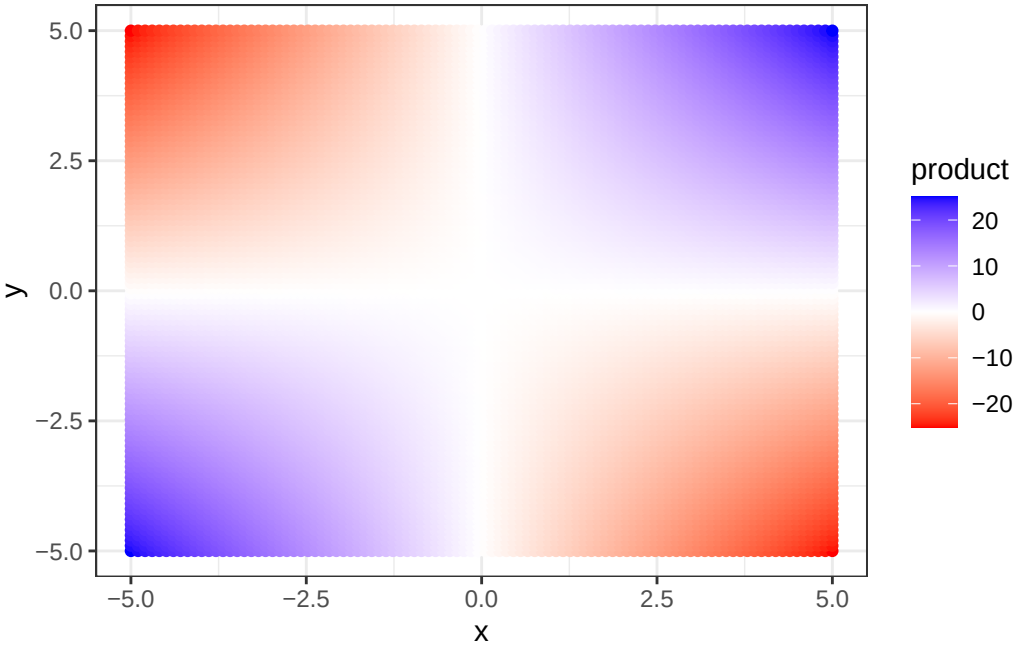
Correlation: Calculation

```
   x  y xbar ybar x-xbar y-ybar product
1: 1  3    3  7.6     -2   -4.6     9.2
2: 2  5    3  7.6     -1   -2.6     2.6
3: 3 10    3  7.6      0    2.4     0.0
4: 4  8    3  7.6      1    0.4     0.4
5: 5 12    3  7.6      2    4.4     8.8

cov(x,y):  4.2

   x  y xbar ybar x-xbar y-ybar product devx  devy
1: 1  3    3  7.6     -2   -4.6     9.2    4 21.16
2: 2  5    3  7.6     -1   -2.6     2.6    1  6.76
3: 3 10    3  7.6      0    2.4     0.0    0  5.76
4: 4  8    3  7.6      1    0.4     0.4    1  0.16
5: 5 12    3  7.6      2    4.4     8.8    4 19.36

SD x:  1.414214

SD y:  3.261901

r:  0.9104655
```

Correlation: Observation

- Once we subtract out the mean of x and y, we are always on the standard x-y plane.
- We can now calculate the contribution to covariance by looking at x*y for each point.
- An observation at point (x,y) after subtracting $\bar{x}, \bar{y}$ will always contribute the same amount to the covariance
- Points near y=x or y=-x have the largest impact, and points that are further from the center

Correlation: Observation



- More extreme observations (outliers) are highly influential
- Note exactly on y=x may contribute little if close to (0,0), but they also contribute little to both σ_x and σ_y, so ρ may not be affected much