Class 2d: Review of concepts in Probability and Statistics

Business Forecasting

Methods of Qualitative Forecasting

Delphi Method

- A structured communication process to reach a consensus for complex, uncertain and long terms forecasting tasks
  1. Select a group of experts
  2. Invite them to the study. They are anonymous and don't talk to each other!
  3. Ask them to answer a questionnaire
  4. Get initial responses
  5. Compile them into summary
  6. Send them summary and get their feedback with refined answers
  7. Reiterate until consensus is reached or no further improvement

Example: Determining AI threats

- What are the risks of AI developments?
- Panel of experts from academia and industry
  - Computer scientists, engineers, CEOs of AI companies, ethic experts
- Send them questionnaires asking about potential threats
- Compile responses into summary and send them back
- Get more rounds of responses until consensus
- Identify the most probable risks

Brainstorming

- Creative technique for generating ideas.
- Encourages free thinking and building on suggestions.
- Appropriate for exploring possibilities.
    - Form a group (no need for experts)
    - State the problem
    - Encourage ideas, no matter how crazy
    - Build and combine each others' ideas
    - Document the ideas and synthesize them

Example: Enhancing Employee Engagement

- Tech company's HR department.
- Representatives from HR, IT, and different departments.
- Generate ideas for a mobile app to enhance employee engagement.
- Write them down and implement the relevant ones

Panel of Experts

- Assemble knowledgeable individuals
    - At the same time and spot
- They meet, offer insights and expertise, and discuss
- Aid in well-informed decisions.
- Sometimes ends up with a report with conculsions

Example: Environmental Policy Formulation

- Government agency want to find identify and address most pressing environmental issues
- Environmental scientists, economists, conservationists, and policymakers.
- Discuss policy options.
- Create comprehensive environmental policies.

Focus Groups

- Gather diverse participant - not necessarily experts
- Share perceptions, attitudes, and opinions.
- Provide qualitative data and consumer insights.

Example: Market Research for a New TV SHOW

- Proposing a new TV Show and trying to see how well it will do
- Participants from various demographics.
- Understand consumers' preferences and perceptions about the TV show
- Fine-tune the product and marketing strategy.

Types of Data

Longitudinal Data

- Observations are collected for the same subject (entity) over a period of time
- Same as time series data
- Example: Tracking a company's annual revenue and number of employees over several years

Longitudinal Data Example

Show 5 entries

Year	Revenue	Employees
2018	50000	50
2019	52000	55
2020	55000	60
2021	58000	65
2022	60000	70

Showing 1 to 5 of 5 entries

Previous 1 Next

- Another Example: Share of people with Diabetes in Mexico in years 2010, 2015, 2020

Cross-Sectional Data

- Observations are collected at a single point in time
- Example: A survey of customers' satisfaction with a product and likelihood of repurchase at a certain point in time

Cross-Sectional Data Example

Show 5 entries

Customer_ID	Satisfaction_Score	Repurchase_Likelihood
1	7	Likely
2	8	Unlikely
3	5	Likely
4	9	Likely

Showing 1 to 4 of 4 entries

Previous 1 Next

- Another Example: Share of people with Diabetes in 2010 in Mexico, USA, Canada, Brazil

Panel Data

- Combines both longitudinal and cross-sectional data
- Observations are collected for multiple subjects over multiple points in time
- Example: Tracking the annual revenue and number of employees of several companies over a few years

Panel Data Example

Show 5 entries

Year	Company	Revenue	Employees
2018	A	50000	50
2018	B	52000	55
2018	C	55000	60
2019	A	58000	65
2019	B	60000	70

Showing 1 to 5 of 15 entries

Previous 1 2 3 Next

- Another Example: Share of people with Diabetes in Mexico, USA, Canada, Brazil, each country in years 2010, 2015, 2020

Q1

Show 5 entries

Month	Cryptocurrency	Market_Cap
Jan	Bitcoin	60000
Jan	Ethereum	40000
Jan	Dogecoin	10000
Feb	Bitcoin	62000
Feb	Ethereum	41000

Showing 1 to 5 of 36 entries

Previous 1 2 3 4 5 … 8 Next

Panel data

- Multiple time observation per subject (currency) and multiple subjects

Q2

Show 5 entries

Country	Population_Millions	GDP_Billions	Internet_Users_Millions
United States	331	21433	246
China	1439	15308	904
India	1380	3160	560
Brazil	213	1848	126
Russia	145	1690	116

Showing 1 to 5 of 5 entries

Previous 1 Next

Cross-sectional data

- Single (time) observation per subject (user), many subjects

Q3

Show 5 entries

Year	Electric_Car_Sales
2020	20000
2021	30000
2022	40000
2023	50000
2024	60000

Showing 1 to 5 of 5 entries

Previous 1 Next

Longitudinal data

- Multiple (time) observations of a single subject

Variable Types

Variable Types

We have two general types: Categorical and Numerical variables

Categorical Variables

- Variables that can be divided into one or more groups or categories.
    - **Ordinal:** These variables can be logically ordered or ranked.
        - *Variable:* Customer Satisfaction Survey Results
        - *Example:* Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied
    - **Nominal:** These variables cannot be ordered or ranked.
        - *Variable:* Social Media Platforms Used
        - *Example:* Facebook, Instagram, Twitter, LinkedIn, TikTok, Snapchat

Numerical Variables

- Variables that hold numeric value and ordering is possible
  - **Discrete:** These variables can only take certain values
    - *Example*: Number of App Downloads from App Store
    - *Example*: Number of children you have
    - *Example*: Size of coke products: 0.33L, 0.5L, 1L, 2.25L

Numerical Variables

- Variables that hold numeric value and ordering is possible
  - **Discrete:** These variables can only take certain values
    - *Example*: Number of App Downloads from App Store
    - *Example*: Number of children you have
    - *Example*: Size of coke products

- **Continuous:** These variables can take any value within a range
  - *Example*: Time spent on a Webpage
  - *Example*: Exchange rate between MXN and USD

- What's the main difference between ordinal and discrete?
  - We could say 1=Very unsatisfied, 2=Unsatisfied
  - But we cannot say that very unsatisfied has half of satisfaction of person who is just unsatisfied!
  - We can order, but these numbers don't have meaning in terms of distance between them

Mexican Health Survey

- Representative sample of the Mexican population n=37858.
- We will use it to investigate market for Ozempic

Show 5 entries

age	gender	weight	location_type	diabetes	Mother_diabetes	Difficulty_walking
51	Male	77.4657	Urban	0	1	A lot of difficulty
41	Female	80.0499	Urban	0	0	A lot of difficulty
44	Male	87.1874	Urban	0	1	No difficulty
68	Female	54.9827	Urban	0	0	No difficulty
52	Female	34.3283	Urban	0	0	A lot of difficulty

Showing 1 to 5 of 37,858 entries

Previous 1 2 3 4 5 … 7,572 Next

- *Age*: Numerical, Discrete
- *Gender*: Categorical, Nominal
- *Weight*: Numerical, Continuous
- *Location_type*: Categorical, Nominal
- *Diabetes*: Categorical, Nominal
- *Mother_diabetes*: Categorical, Nominal
- *Difficulty_walking*: Categorical, Ordinal

Summarizing Data

Graphical summaries

Categorical variables

Frequency Tables

Frequency table: present the absolute frequencies (counts) and relative frequencies (shares) of each category.

- Categories are mutually exclusive and collectively exhaustive
- Relative frequency of category i: $p_i = \frac{n_i}{N}$
  - n_i is count of category i
  - N is total count in the sample

Show 8 entries

Location

Category	n_i	p_i
Rural	9899	0.261
Urban	27959	0.739
Total	37858	1

Showing 1 to 3 of 3 entries

Previous 1 Next

Show 8 entries

Difficulty Waking

Category	n_i	p_i
A lot of difficulty	1639	0.043
Impossible	183	0.005
No difficulty	31269	0.826
Some difficulty	4767	0.126
Total	37858	1

Bar Charts

Bar charts visually represents the frequency count of each category

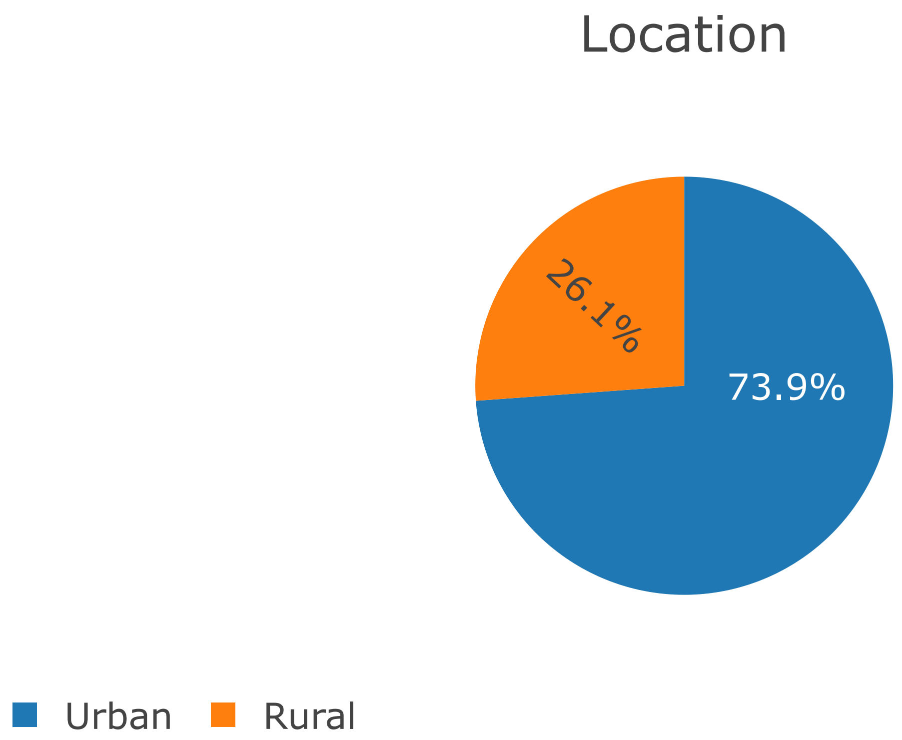
Bar Charts

Bar charts visually represents the frequency count of each category

More Creative Bar Chart

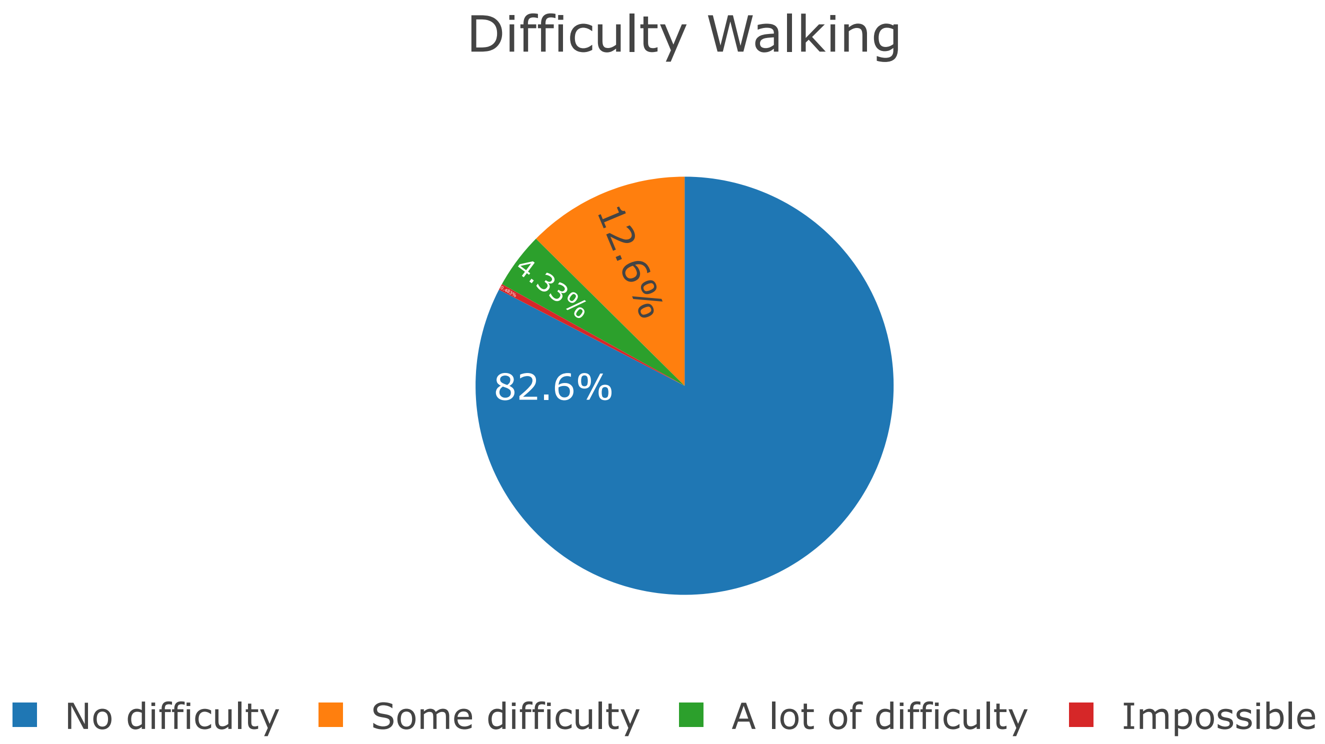
Pie Charts

Pie chart: Each slice is proportional to the category's frequency



Pie Charts

Pie chart: (Angle of) Each slice is proportional to the category's frequency



My favorite pie chart

Frequency Distribution

Suppose we survey people age 30-50 how many partners they had in their life.

- What's the distribution of partners?
- Calculate relative frequencies
- Show them on a bar graph

Data

Show 6 entries

Number_of_partners	n_i	p_i
0	5	0.033
1	9	0.06
2	13	0.087
3	22	0.147
4	27	0.18
5	19	0.127

Showing 1 to 6 of 22 entries

Previous 1 2 3 4 Next

Distribution

Frequency Distribution

We can also show frequency of age of people who have diabetes from our data

Frequency Distribution

Compare it to the age distribution in the adult population (20+)

Numerical Variables: Continuous

- What about continuous values? Why can't we do the same?

Show 6 entries

weight	n_i	p_i
30.3745	1	0.0000264
30.4593	1	0.0000264
30.5235	1	0.0000264
30.6135	1	0.0000264
30.7581	1	0.0000264
30.9106	1	0.0000264

Showing 1 to 6 of 36,297 entries

Previous 1 2 3 4 5 ... 6,050 Next

- Most values never repeat, so they have very low relative frequency

Histograms

Solution: Group similar values together

- Construct intervals and show how many observations are in a given interval

Process

1. Decide how many intervals
2. And how wide they are
3. Then calculate the absolute and relative frequencies of each interval
4. Plot it with bars

My approach

- I want k (example k=5) equal intervals

- Divide the range of the data into k equal intervals

    - *Range* is max-min of the data

```
# Calculate max and min
max_value <- max(Health_data$weight)
min_value <- min(Health_data$weight)

# Calculate the difference
range <- max_value - min_value
```

```
## [1] "Range= 190.8078 - 30.3745 = 160.4333"
```

- With 5 intervals, each will be 32kg wide
- The first one starts at the minimum value (30.3745)
- The last one ends at the maximum value (190.8078)
- Calculate how many observations I have in each interval and what's the relative frequency

Histograms

- Midpoint represents middle of the interval - center of the bar
- P_i is cumulative frequency: share of observations in this or smaller interval
  - *Example*: $P_{(62.46-94.55)} = 0.911$
  - *Interpretation*: 91.1% of people have weight lower than 94.55kg

Show 6 entries

Interval	Midpoint	n_i	p_i	P_i
30.37 - 62.46	46.42	10068	0.2659411	0.2659411
62.46 - 94.55	78.5	24430	0.6453061	0.9112472
94.55 - 126.63	110.59	3206	0.0846849	0.9959321
126.63 - 158.72	142.68	143	0.0037773	0.9997094
158.72 - 190.81	174.76	11	0.0002906	1

Showing 1 to 5 of 5 entries

Previous 1 Next

Histogram with 10 Classes

Now, let's increase the number of classes to 10.

Show 6 entries

Interval	Midpoint	n_i	p_i	P_i
30.37 - 46.42	38.4	796	0.0210259	0.0210259
46.42 - 62.46	54.44	9272	0.2449152	0.2659411
62.46 - 78.5	70.48	15742	0.415817	0.6817581
78.5 - 94.55	86.53	8688	0.2294891	0.9112472
94.55 - 110.59	102.57	2661	0.070289	0.9815362
110.59 - 126.63	118.61	545	0.0143959	0.9959321

Showing 1 to 6 of 10 entries Previous 1 2 Next

Histogram with 100 Classes

Show 6 entries

Interval	Midpoint	n_i	p_i	P_i
30.37 - 31.98	31.18	8	0.0002113	0.0002113
31.98 - 33.58	32.78	11	0.0002906	0.0005019
33.58 - 35.19	34.38	7	0.0001849	0.0006868
35.19 - 36.79	35.99	16	0.0004226	0.0011094
36.79 - 38.4	37.59	24	0.0006339	0.0017433
38.4 - 40	39.2	24	0.0006339	0.0023772

Showing 1 to 6 of 100 entries

Previous 1 2 3 4 5 … 17 Next

- Helps to see the distribution and outliers
- Is more always better?
- With smaller intervals, histogram tends to the **probability density function**

Probability Density Function (PDF)

Definition

- **Probability Density Function (pdf)** describes the probability distribution of a continuous random variable.
- It **does not** give probability at a given value (this is always 0 for continous variable)
- It shows which in which intervals that variable the most often appears
- It is used to calculate the probability of the random variable being in a given interval
- Area under it always adds up to 1

Example

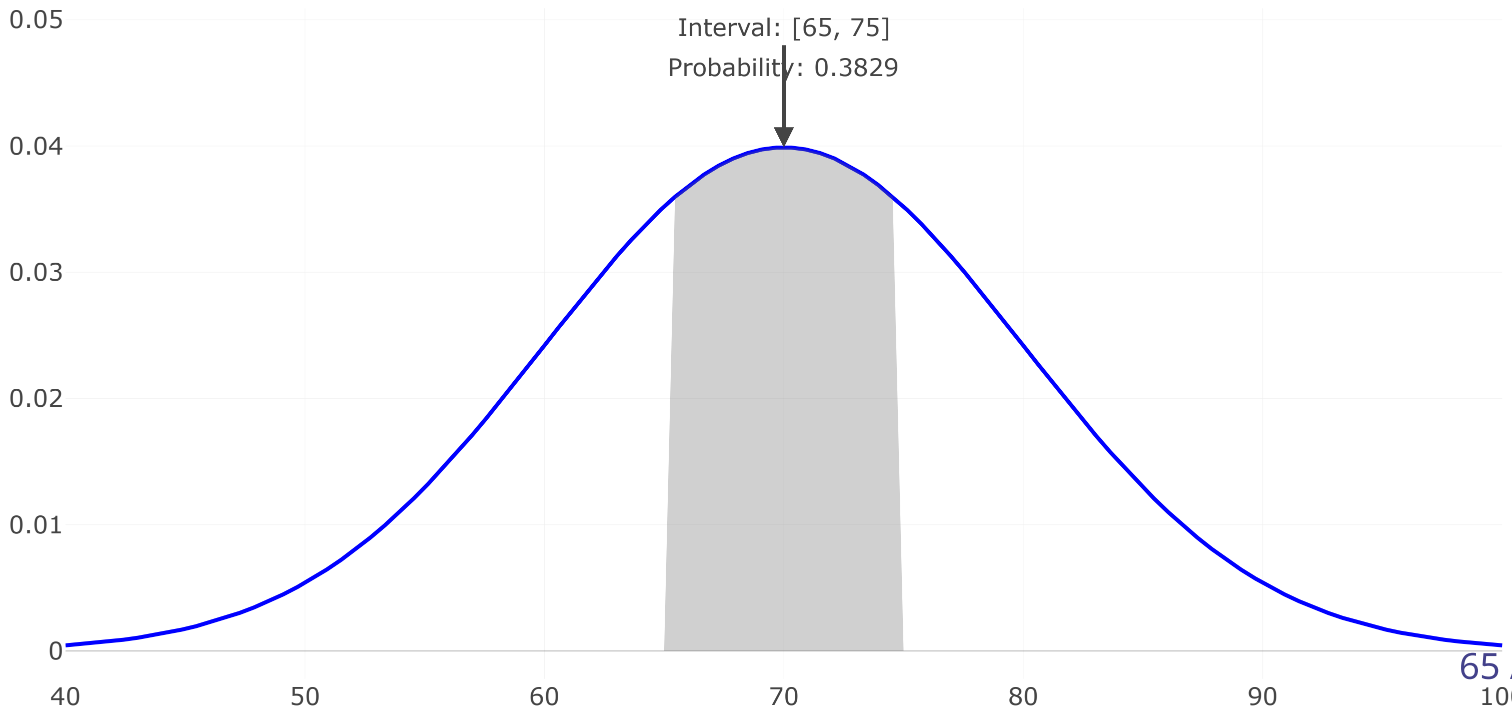
We have a random variable X representing the weight of adults in Mexican population. The PDF of X helps to describe the likelihood of finding a person of a specific weight within a range (e.g., between 58kg and 60kg).

How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

What is the share of population with weight between 65kg and 75kg?

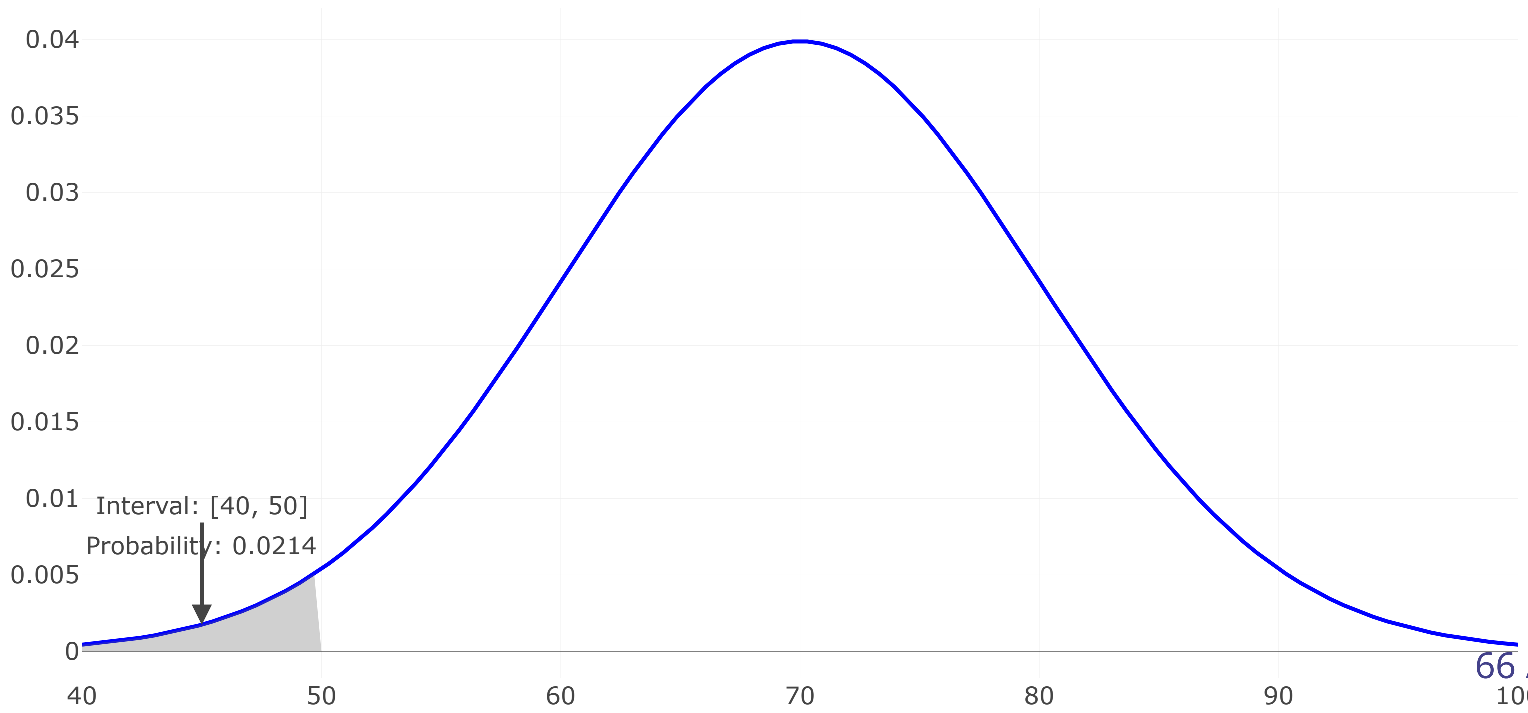


How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

What is the share of population with weight between 40 and 50kg?

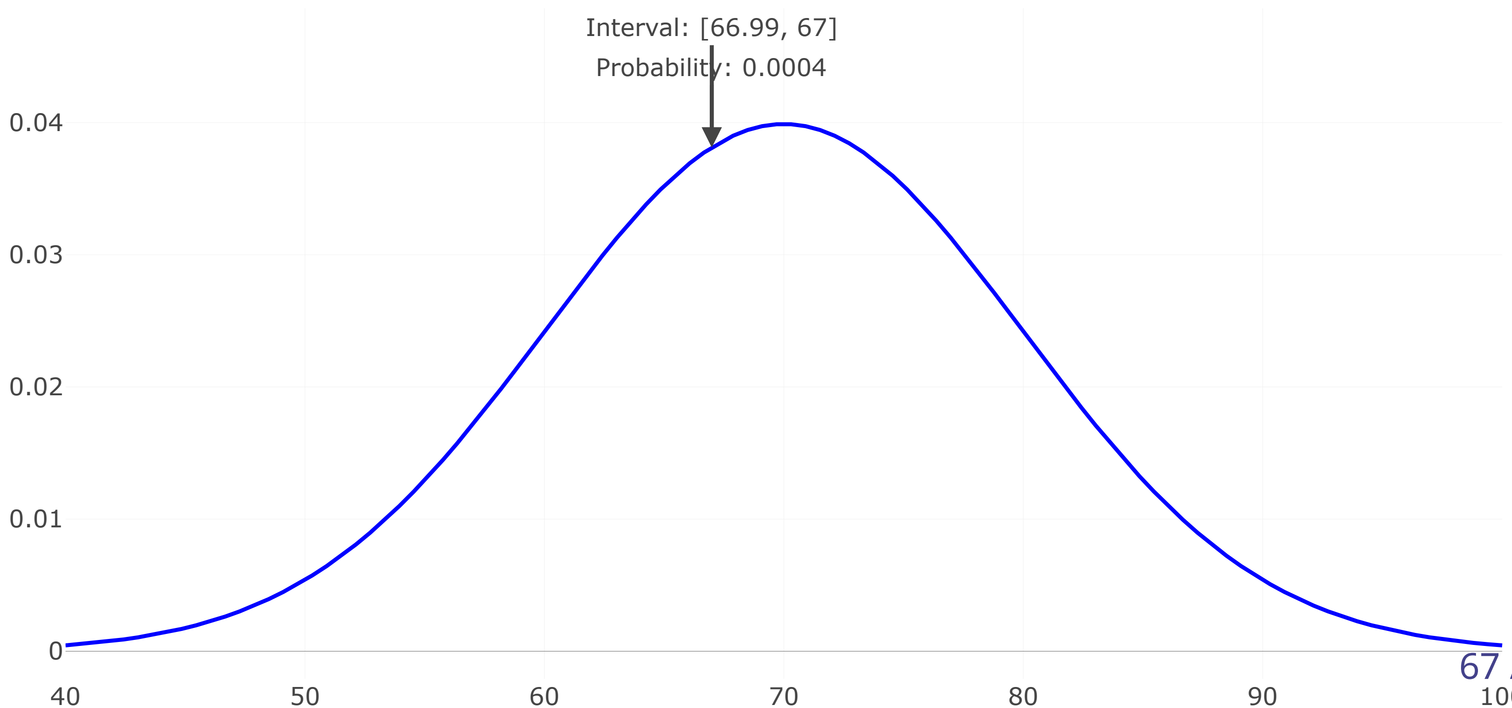


How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

What is the share of population with weight between 66.99 and 67 kg?

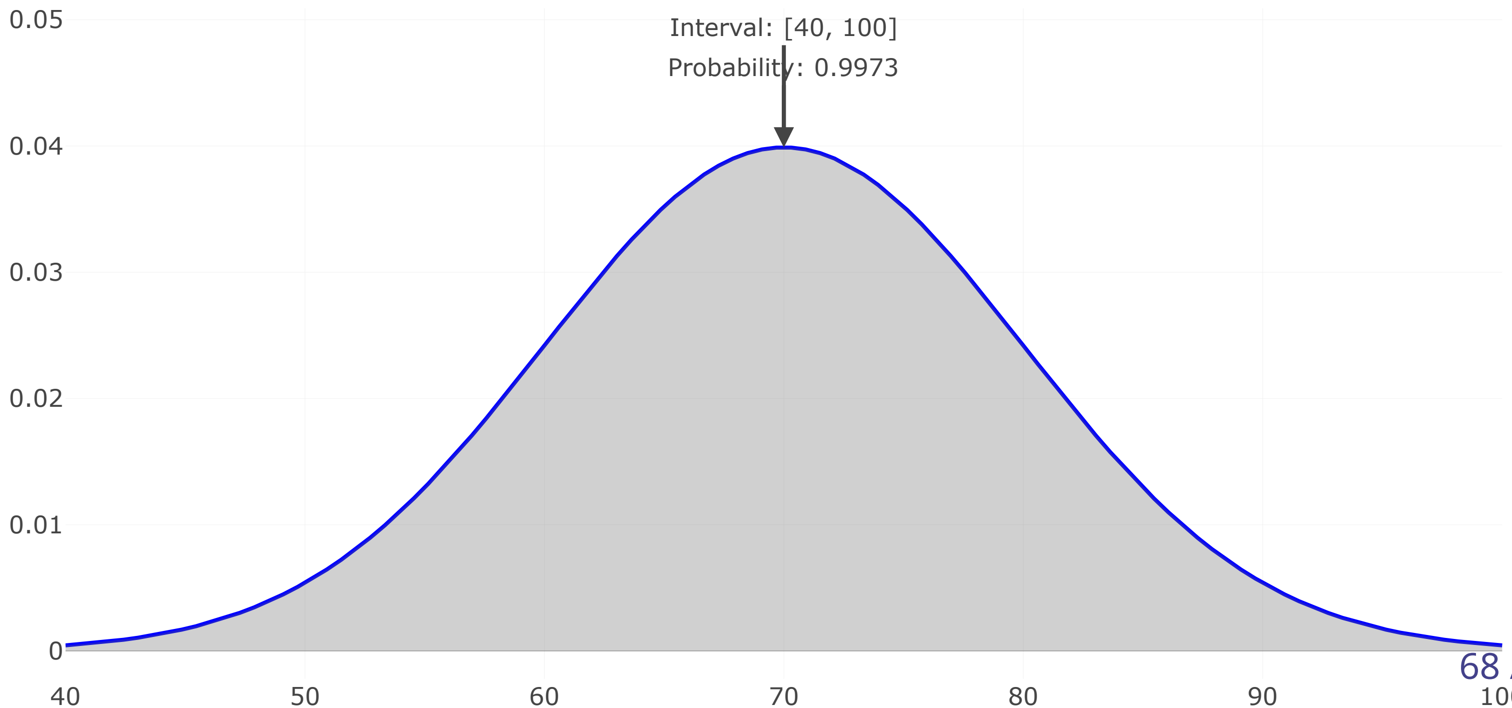


How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

What is the share of population with weight between 40 and 100 kg?



Distribution Shapes: Modality

Which is uniformaly distributed

1. Weights of Adult Females
2. Salaries in Mexico
3. Airbnb prices in CDMX
4. Birthdays of Classmates (day of the month)

Distribution Shapes: Skewness

Age at death

What if we want to calculate proportion of people who weight less or equal to 50kg?

Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) gives the probability that a random variable X will take on a value less than or equal to a specific value.

For a continuous random variable X with PDF f(x), the CDF F(x) is defined as:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = P(X \leq x)$$

Characteristics:

- The CDF starts (for minus infinity) at 0 (minimum)
- It approaches 1 as x approaches infinity (maximum)
- It is non decreasing
- It is right continuous

Example 1: Normal Variable (weight in the population)

$F(50) = \int_{-\infty}^{50} f(t)\,dt = P(X \leq 50) = 0.02$

Example 2: Normal Variable (weight in the population)

$F(72) = \int_{-\infty}^{72} f(t)\,dt = P(X \leq 72) = 0.58$

Example 3: Normal Variable (weight in the population)

$F(102) = \int_{-\infty}^{102} f(t)\, dt = P(X \leq 102) = 0.99$

Never integrate a CDF!

Empirical CDF

What if we only have a sample and we don't know the true pdf?

Intuition on how it comes up:

Empirical CDF

What if we only have a sample and we don't know the true pdf?

Intuition on how it comes up:

Empirical CDF

$$ECDF(x) = \frac{\sum I(w_i \leq x)}{N} = \frac{\text{Number of people with weight lower than x}}{N}$$

- $I(w_i < x) = 1$ if weight of person i is lower than x (*Indicator Function*)
- N is total number of people (*Sample Size*)
- Share of people with weight lower than x

- So how do we calculate share of people with weight=<50kg?
  $P(weight \leq 50) = ECDF(50)$
- What about more than 100?

- Is $F_x(X)$ a valid distribution function?
- What's the probability that the rent is larger than 10 000?

Summarizing Data

Comparisions and Associations

Comparisions

- Descriptive and visual comparisons
- NOT declaring statistically significant differences, just eyeballing
- That's coming next

Comparing categorical variables

Do people living in rural areas are more likely to have diabetes?

- We have two categorical variables
- We can use frequency table to see how diabetes is distributed among the two types of areas:

	No Diabetes	Has Diabetes
Rural	8906	993
Urban	24780	3179

Comparing categorical variables

Do people living in rural areas are more likely to have diabetes?

- Are relative frequencies more helpful?
- Share of each subgroup within the sample

	No Diabetes	Has Diabetes	Total
Rural	0.24	0.03	0.27
Urban	0.65	0.08	0.73
Total	0.89	0.11	1.00

- Can we compare numbers in the *Has Diabetes* column?
- **Marginal frequencies** are total probabilities by group

Table of frequency

- We want to compare whether someone living in rural area is more likely to have diabetes than someone living in urban area
- So we want to see whether:

$$P(Diabetes_i = 1|Area_i = Rural) > P(Diabetes_i = 1|Area_i = Urban)$$

- We want to look at the **relative conditional frequencies**
- They are usually in **contingency tables**
    - Share with diabetes within urban sample
    - Share with diabetes within rural sample

	No Diabetes	Has Diabetes
Rural	0.90	0.10
Urban	0.89	0.11

$$P(Diabetes_i = 1|Area_i = Rural) = \frac{P(Diabetes_i = 1 \cap Area_i = Rural)}{P(Area_i = Rural)} \approx \frac{0.03}{0.03 + 0.24} \approx 0.1$$

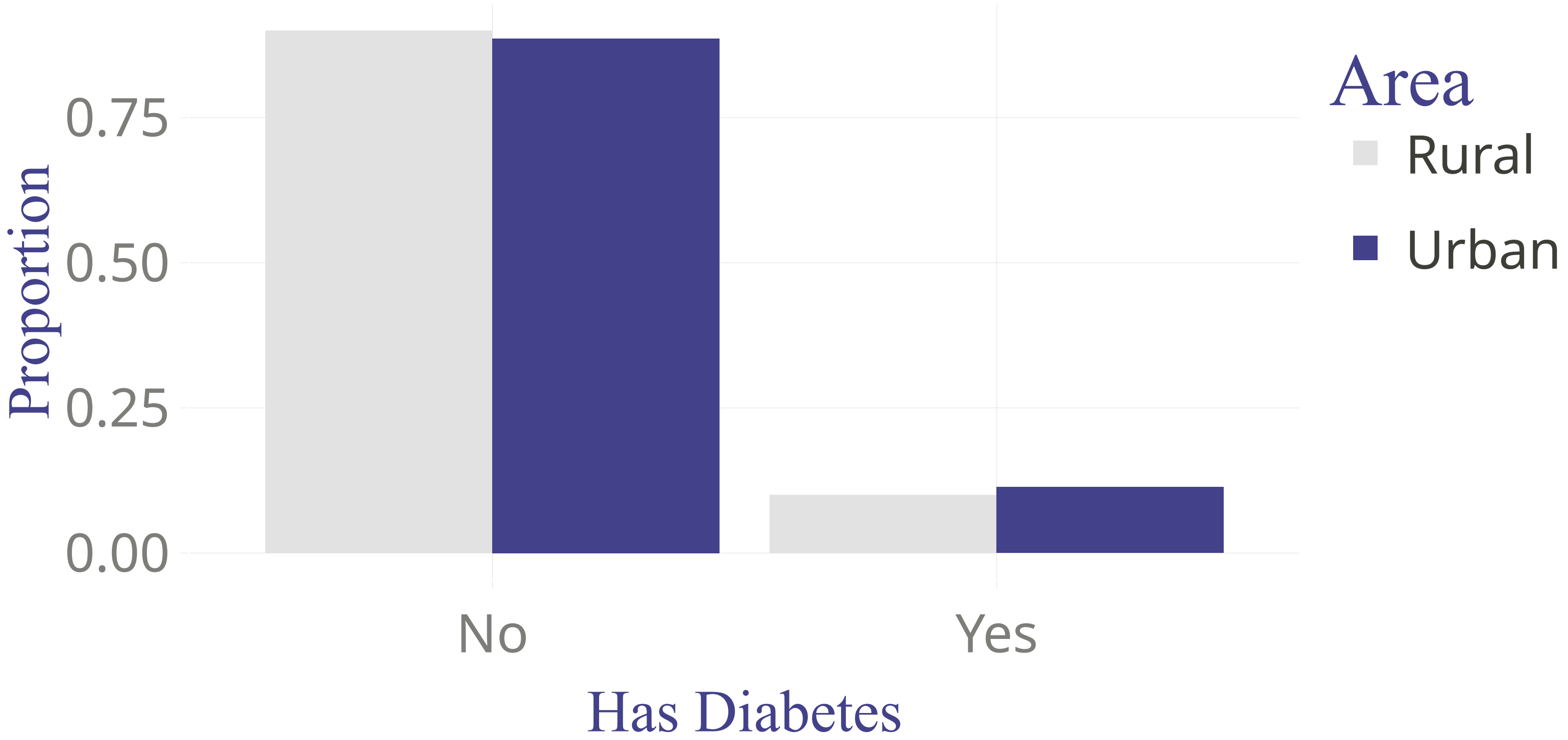
Or:

$$P(Diabetes_i = 1|Area_i = Rural) = \frac{\text{Number live in Rural \& Have diabetes}}{\text{Number live in Rural}} = \frac{993}{993 + 8906} \approx 0.1$$

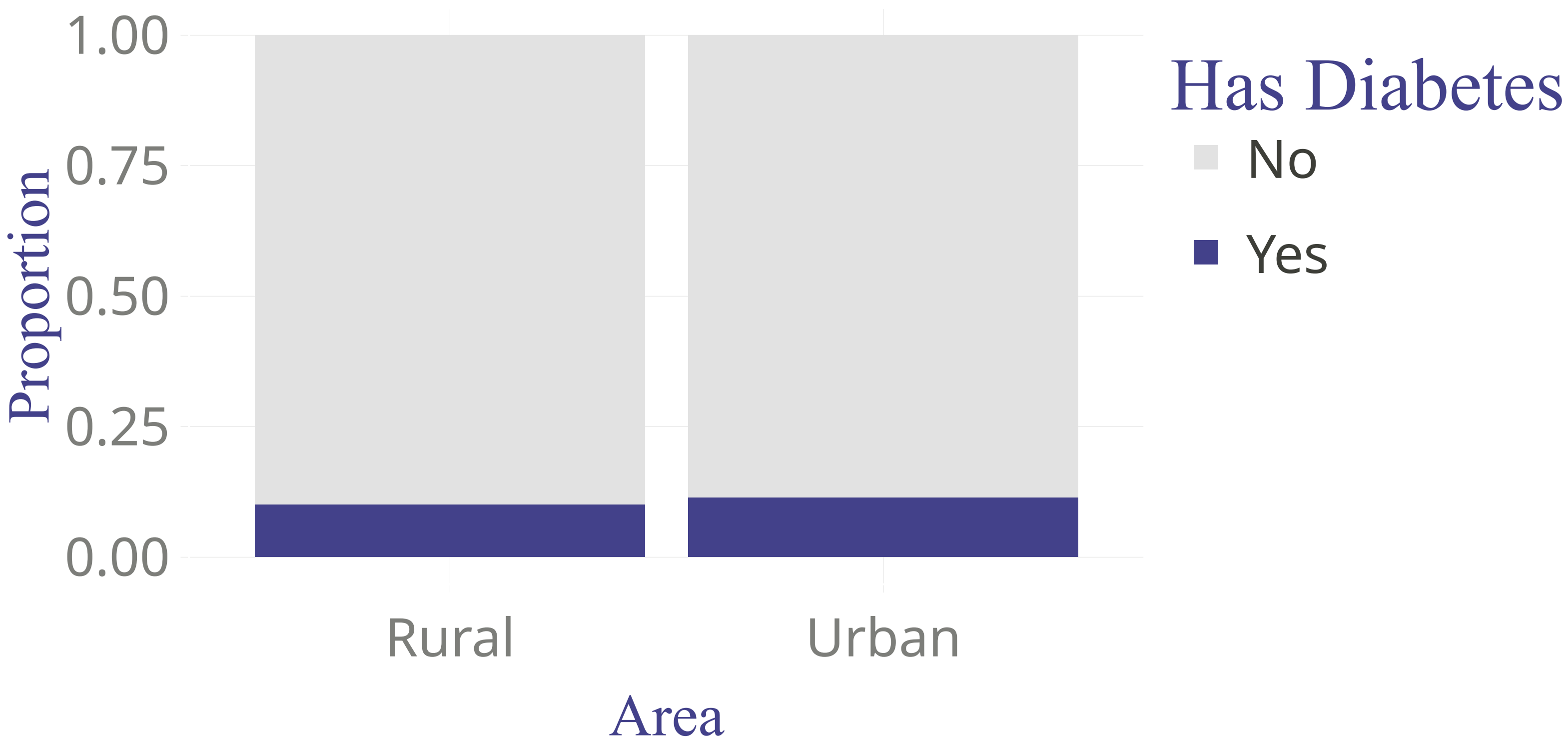
	No Diabetes	Has Diabetes
Rural	0.90	0.10
Urban	0.89	0.11

- What about marginal frequencies here?
  - Row sums should add up to 1
    - $P(Diabetes_i = 1|\text{Area=Rural}_i) + P(Diabetes_i = 0|\text{Area=Urban}_i)$
  - Column sums are meaningless
    - $P(Diabetes_i = 1|\text{Area=Rural}_i) + P(Diabetes_i = 1|\text{Area=Urban}_i)$

- We can visualize it on a barplot



- Or better on a **stacked barplot**



- *Stacked barplot* clearly shows the distribution of diabetes within each group

Practice

- Are you more likely to have diabetes if your mother had diabetes?
- By how much?

	No Diabetes	Has Diabetes
Mother No Diabetes	25270	2427
Mother Has Diabetes	8283	1721

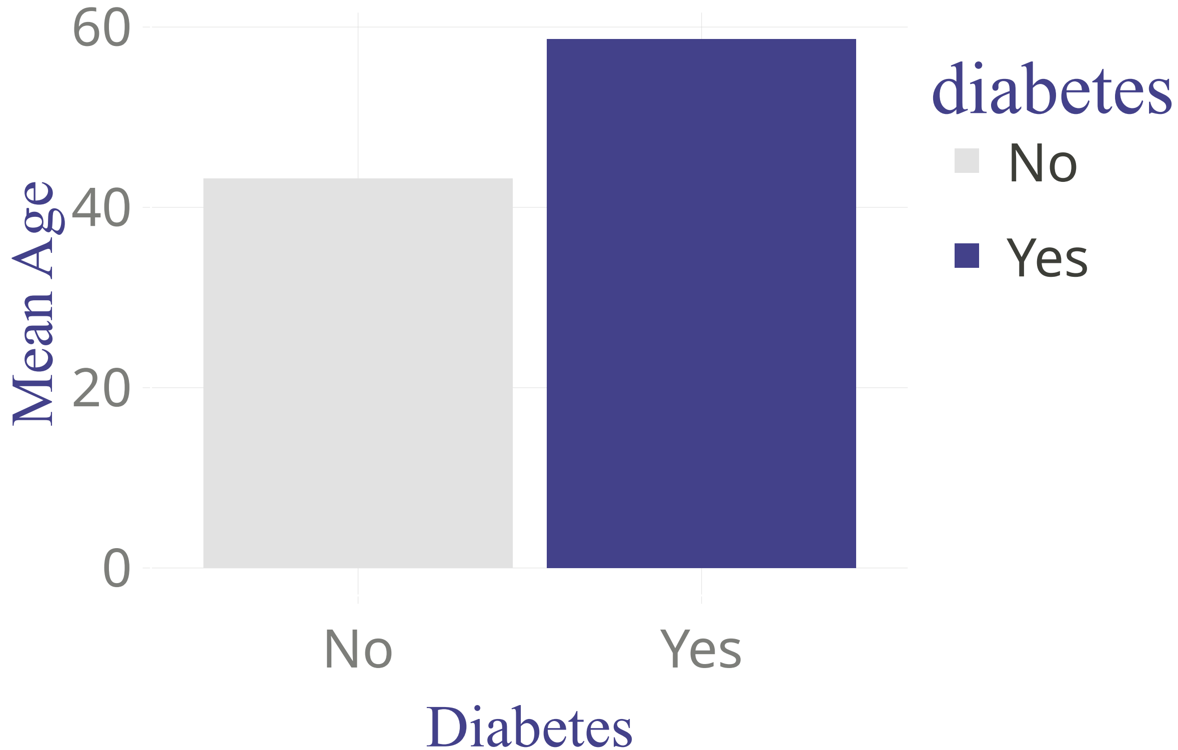
Practice

	No Diabetes	Has Diabetes
Mother No Diabetes	0.91	0.09
Mother Has Diabetes	0.83	0.17

- Does it mean that having diabetic mother **causes** higher change of having diabetes?

One quantitative and one categorical

- For quantitative variables we can compare some summary statistics
  - Are people with diabetes older than people without it?
  - *Example* means in two subpopulations



One quantitative and one categorical

- Or we can do Box and Whiskers plots as before
- Or we can compare the whole distributions of frequencies

One quantitative and one categorical

- For continuous variables we can use the same methods (except frequency distribution)
- Instead, we can compare densities or histograms
- Are men heavier than women?

Associations: Two Quantitative Variables

- Likely people would subscribe to the website to lose weight
- But do these people have resources?
- What is the relationship between Body Mass Index (BMI) and Income?
- More generally, how to measure association between two quantitative variables
- Association between qualitative variables is measured with contingency tables

Associations

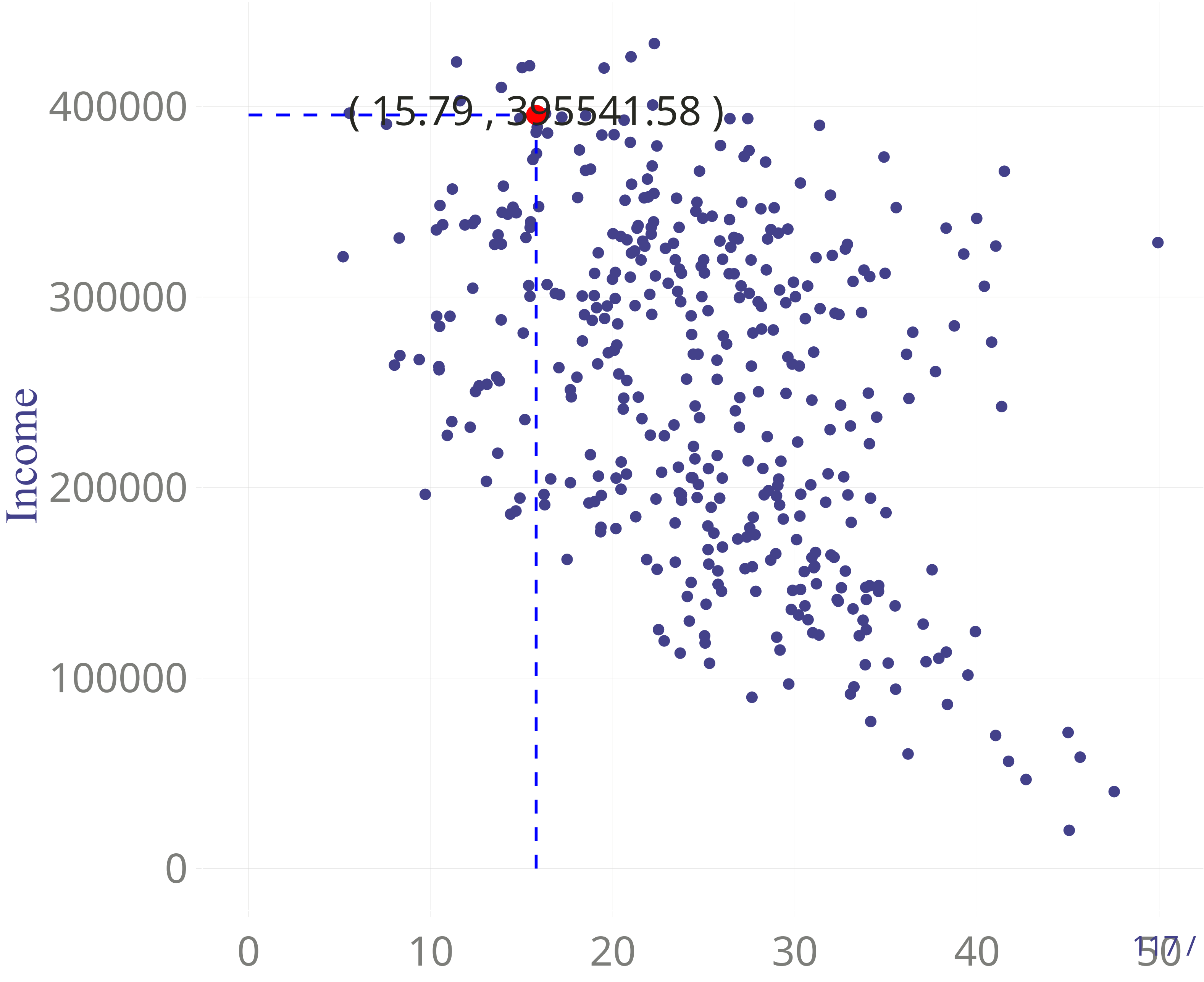
- Suppose we surveyed people from Guadalajara and CDMX about their BMI, education and income.
- Scatter plots show associations between two quantitative variables
  - We put variables of interest (*example*: Y and X) on the axis
  - We place observation on the cartesian plane using their values of variable X and Y: $\{(x_1, y_1), (x_2, y_2)..\}$
- In our case:
  - X axis is BMI
  - Y axis is Income
  - An individual i is placed on these axis based on $(BMI_i, Income_i)$

Show 4 entries

City	BMI	Education	Income
Mexico City	19.52	17.5	420224.44
Mexico City	22.16	15.3	368793.49
Mexico City	36.47	11.3	281512.52
Mexico City	24.56	13.4	344991.58

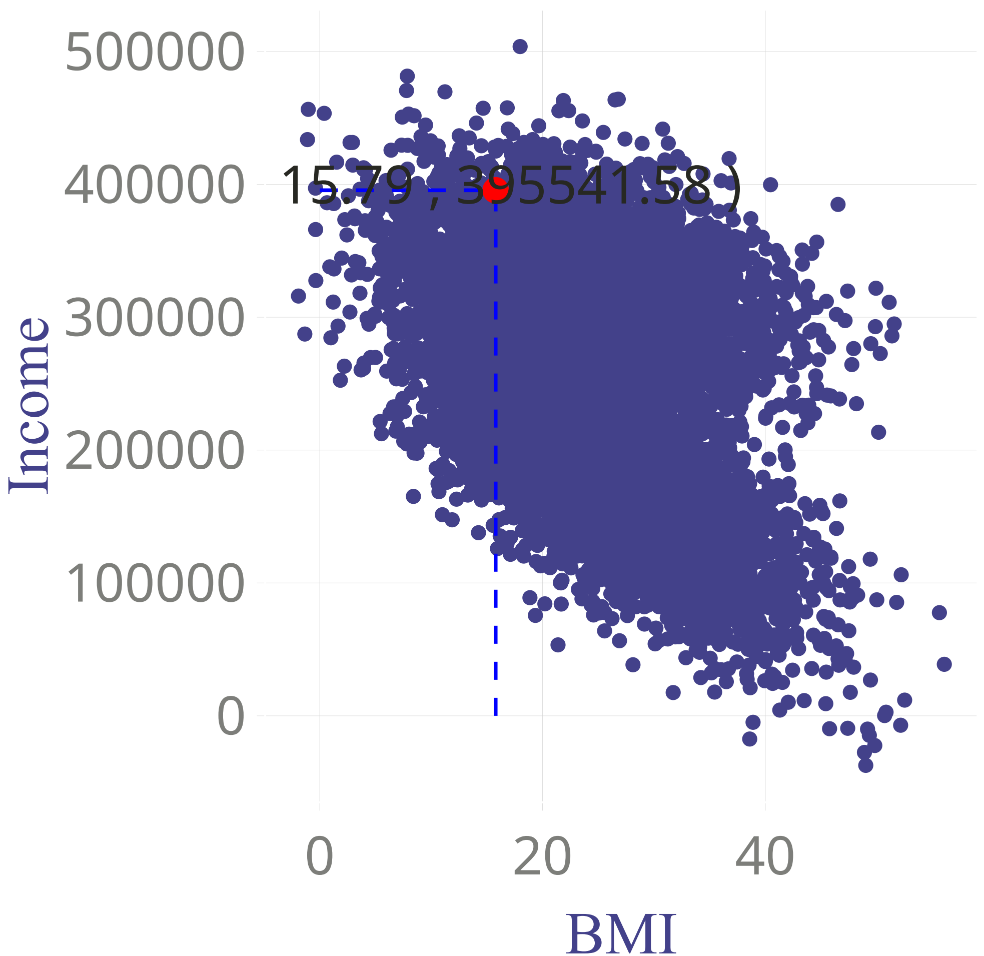
Showing 1 to 4 of 400 entries

Previous 1 2 3 4 5 … 100 Next


(15.79 , 395541.58)
Income
400000
300000
200000
100000
0
0
10
20
30
40
50

Associations

- Scatterplots become very messy if you have a lot of observations



Associations

- If n is larger, better to use binscatter:
  - Group x variable into quantiles (ex: 10 deciles)
  - Calculate average of y in each decile
  - Plot

```
## Call: binsreg
##
## Binscatter Plot
## Bin/Degree selection method (binsmethod)  =  User-specified
## Placement (binspos)                       =  Quantile-spaced
## Derivative (deriv)                        =  0
##
## Group (by)                         =  Full Sample
## Sample size (n)                    =  10000
## # of distinct values (Ndist)       =  3214
## # of clusters (Nclust)             =  NA
## dots, degree (p)                   =  0
## dots, smoothness (s)               =  0
## # of bins (nbins)                  =  10
```

Assocations

- Would you say that the relationship is stronger in Guadalajara or in Mexico City?

- How to measure the strength of the relationship?

Associations

Covariance

- **Covariance** measures the strength of the relationship between two variables.

$$\text{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)$$

And it's sample equivalent is:

$$\hat{\text{Cov}}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

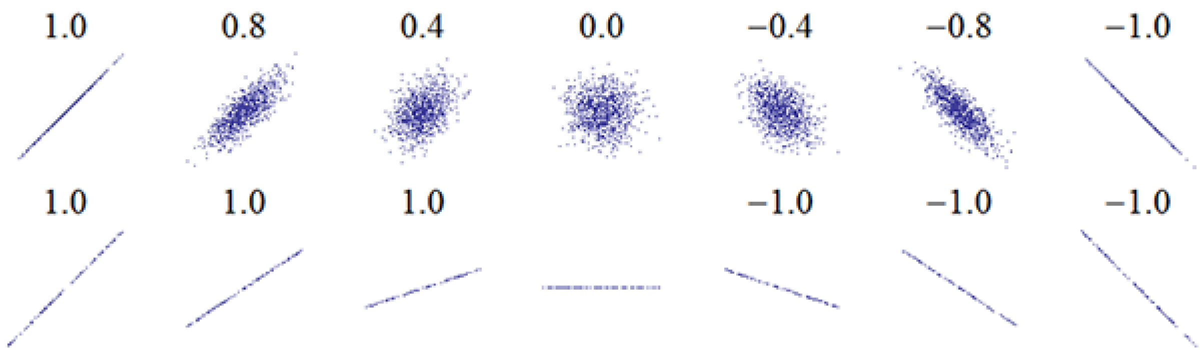
- Covariance whether the two variables move together

- Covariance increases when:
  - The relationship is stronger
  - The deviations of variables are larger

We use the Correlation coefficient to quantify the strength and direction of a relationship between two variables. $e.g.$, think about height and weight, or hours of sleep and irritability.

● The Pearson product-moment correlation coefficient is scale free and it ranges between -1 and 1.
● It is typically denoted by r, for sample data or by ρ (the greek symbol Rho), to indicate the population value.
● You have probably examined XY scatterplots to visualize this type of bivariate relationship, and have begun to evaluate the 2 dimensional attributes of the scattercloud to gain a sense of direction and strength of the relationship.

● Often, introductory textbooks show a figure like the following which depicts a series of XY scatterplots reflecting correlation patterns of differing size and sign. This one is the Wikipedia illustration.



● A correlation of -1 means that the X and Y variables have a perfect negative relationship and the data points fit a straight line with a negative slope.
● Similarly, a correlation of +1 means that X and Y have a perfect positive relationship and fall on a line with positive slope.

Source: https://shiny.rit.albany.edu/stat/rectangles/

Covariance

Covariance

- What has stronger relationship with Income: BMI or Years of Education?

- BMI has larger covariance
- But we can't compare covariances of different variables
- Covariance depends on the scales (or units) of the variable
- All else equal, larger standard deviation implies larger covariance
  - The squares are just bigger

Reminder

We often use it to calculate variance of a sum or difference of two random variables

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$$

Reminder: if a is a constant

$$E(aX) = aE(X) \quad and \quad E(a+X) = E(X) + a$$

And

$$E(X+Y) = E(X) + E(Y)$$

More on that in the homework!

Correlation

- **Correlation measures** the strength of a linear relationship between two variables.
- It ranges between -1 and 1

Population Correlation coefficient:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

Sample Correlation coefficient:

$$\hat{\rho}(X,Y) = \frac{\hat{\mathrm{Cov}}(X,Y)}{s_X \cdot s_Y}$$

Where $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Correlation

- Correlation is preferred over covariance because it's **scale-independent** and easier to interpret.
- Suppose that instead of measuring income (Y variable) in MXN , we measure it in Dollars.
  - Z income in dollars $Z = \frac{Y}{16}$
  - Is $Cov(X, Z) = Cov(X, Y)$?

$$
\begin{aligned}
cov(X, Z) &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(z_i - \mu_Z) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(\frac{y_i}{16} - \frac{\mu_Y}{16}) \\
&= \frac{1}{16}\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y) \\
&\neq cov(X, Y)
\end{aligned}
$$

Correlation

- Correlation is preferred over covariance because it's **scale-independent** and easier to interpret.
- Suppose that instead of measuring income (Y variable) in MXN , we measure it in Dollars.
  - Z income in dollars $Z = \frac{Y}{16}$
  - Is $\rho(X, Z) = \rho(X, Y)$?

$$\rho(X, Z) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(z_i - \mu_Z))}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2} \cdot \sqrt{\sum_{i=1}^{N}(z_i - \mu_Z)^2}}$$

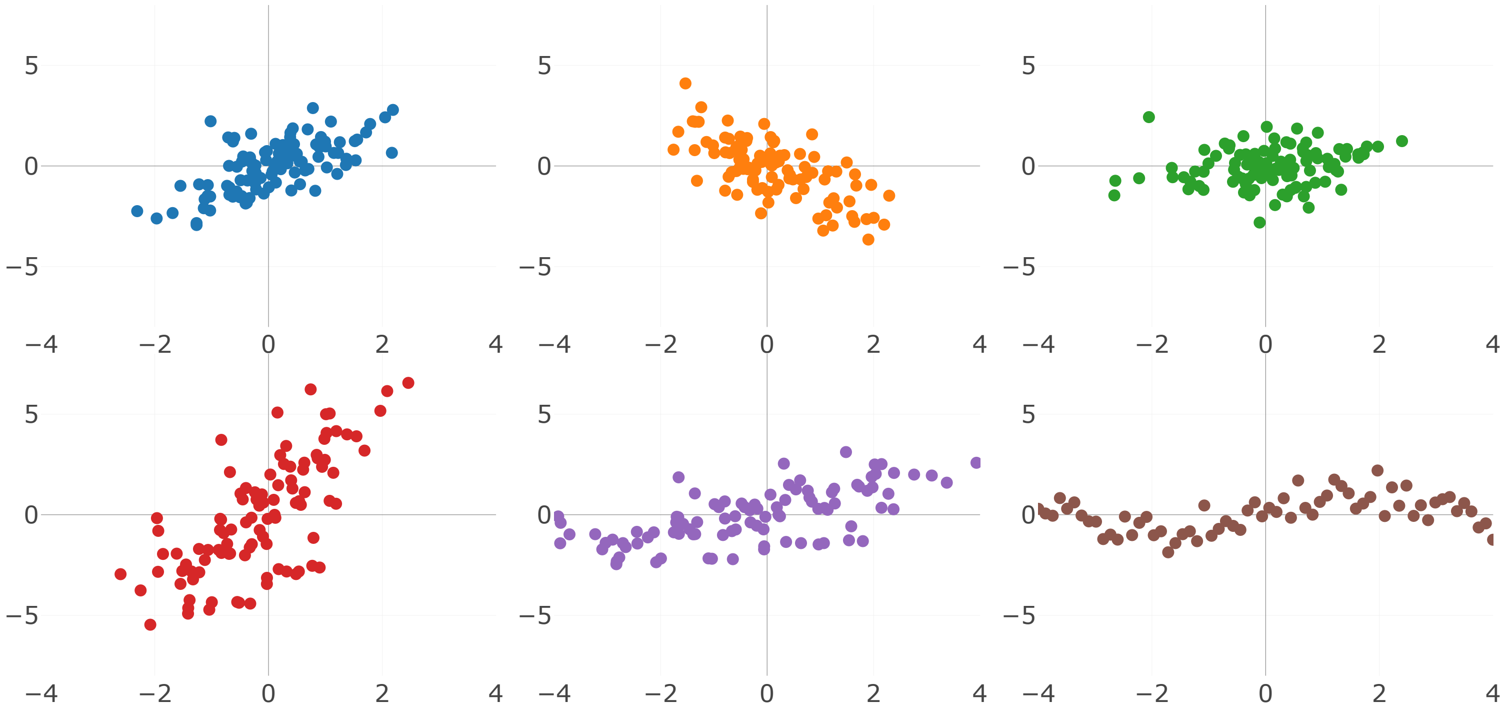
$$= \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}(x_i - \mu_X)(\frac{y_i}{16} - \frac{\mu_Y}{16})}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2} \cdot \sqrt{\sum_{i=1}^{N}(\frac{y_i}{16} - \frac{\mu_Y}{16})^2}}$$

$$= \frac{\frac{1}{16}\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\frac{1}{16}\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \mu_Y)^2}}$$

$$= \rho(X, Y)$$

Correlation

- Correlation with education is actually stronger



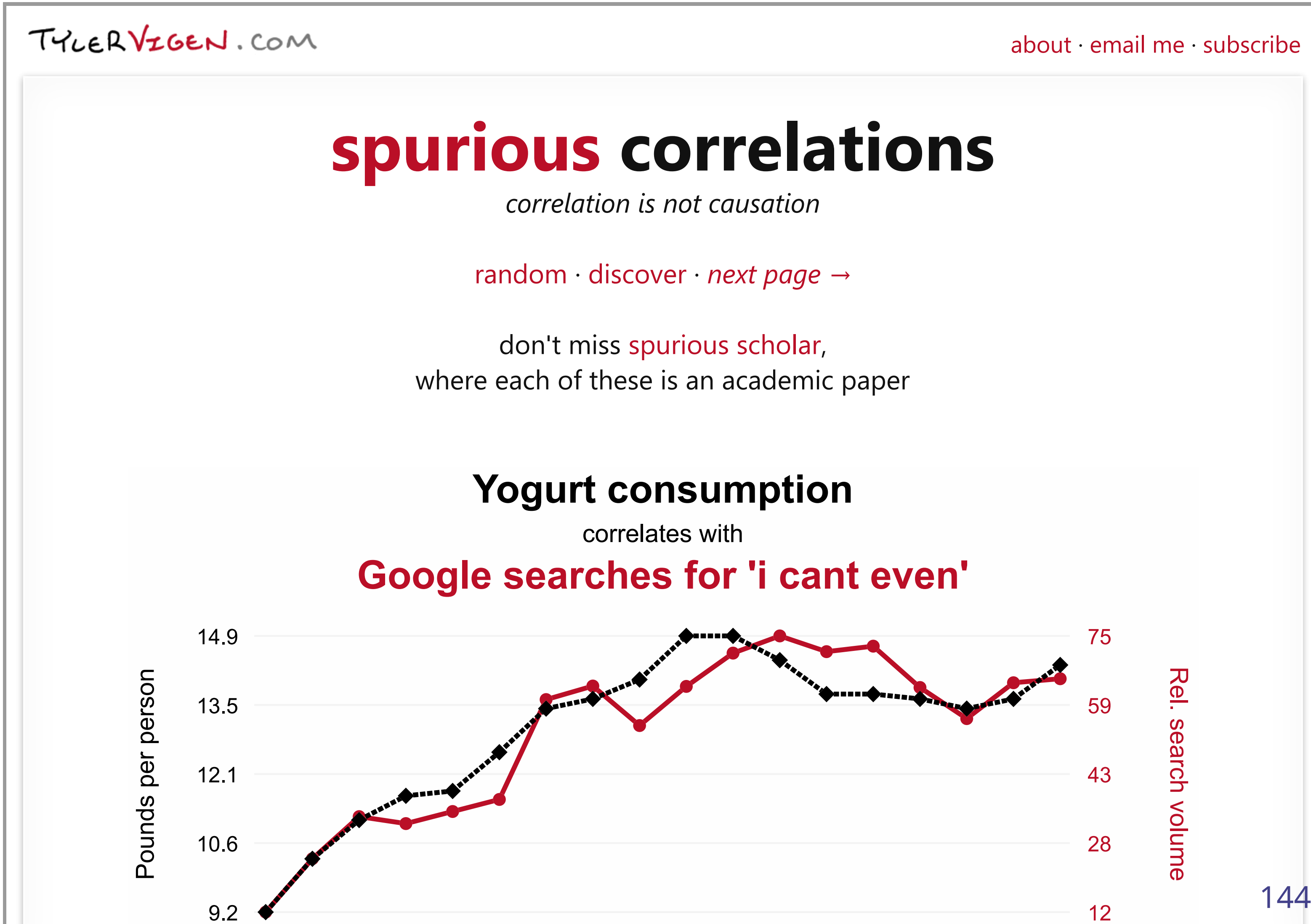
5
0
−5
−4
−2
0
2
4

Correlation

1. Correlation is a value between -1 and 1: $-1 \leq \rho(X, Y) \leq 1$.
2. Perfect positive correlation: $\rho = 1$. Perfect negative correlation: $\rho = -1$.
3. No linear correlation: $\rho = 0$, but this doesn't imply independence.
4. Correlation measures **linear** relationships; nonlinear relationships might not be accurately captured.
5. Correlation doesn't imply causation; a relationship could be coincidental.

Causality vs Correlation

Causality vs Correlation



Causality vs Correlation

- Less obvious examples
- You look at historical data from some media campaign
- You notice that people who were more exposed to ads were less likely to buy that product
- What can you conclude?
- Are people who were exposed to ads similar to people who were not?
- Maybe they were targeted in the first place because they are less likely to buy and you want to change it?

Causality vs Correlation

- Less obvious examples
- Education usually correlates with Income (correlation)
- Does it mean that if decide to get a degree, you will earn more? (causality)