# CDFs and PDFs

30 January 2019

PLSC 309

# Expected Value of a PMF

### Expected value of a Discrete Random Variable

If $X$ takes outcomes $x_1$, ..., $x_k$ with probabilities $P(X = x_1)$, ..., $P(X = x_k)$, the expected value of $X$ is the sum of each outcome multiplied by its corresponding probability:

$$
\begin{aligned}
E(X) &= x_1 \times P(X = x_1) + \cdots + x_k \times P(X = x_k) \\
&= \sum_{i=1}^{k} x_i P(X = x_i) \qquad (2.71)
\end{aligned}
$$

The Greek letter $\mu$ may be used in place of the notation $E(X)$.

# Variance of a PMF

### General variance formula

If $X$ takes outcomes $x_1$, ..., $x_k$ with probabilities $P(X = x_1)$, ..., $P(X = x_k)$ and expected value $\mu = E(X)$, then the variance of $X$, denoted by $Var(X)$ or the symbol $\sigma^2$, is

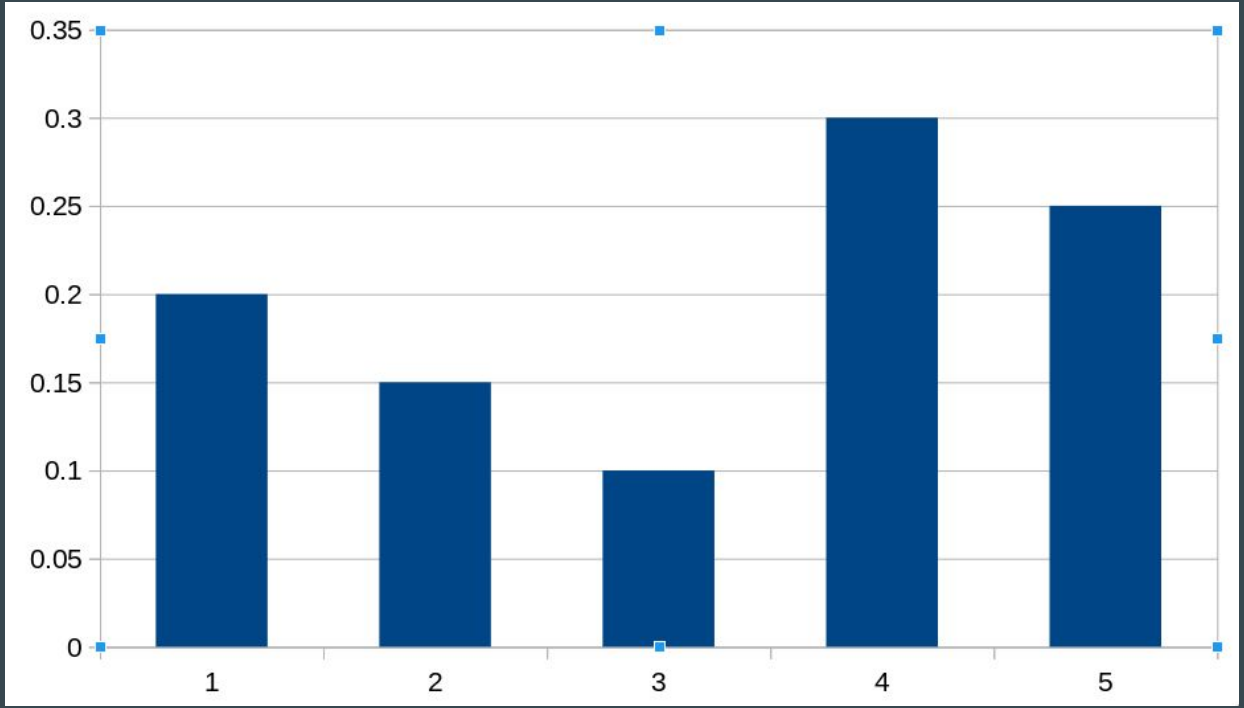
$$\begin{aligned}\sigma^2 &= (x_1 - \mu)^2 \times P(X = x_1) + \cdots \\ &\qquad \cdots + (x_k - \mu)^2 \times P(X = x_k) \\ &= \sum_{j=1}^{k} (x_j - \mu)^2 P(X = x_j) \qquad (2.72)\end{aligned}$$

The standard deviation of $X$, labeled $\sigma$, is the square root of the variance.

# Example

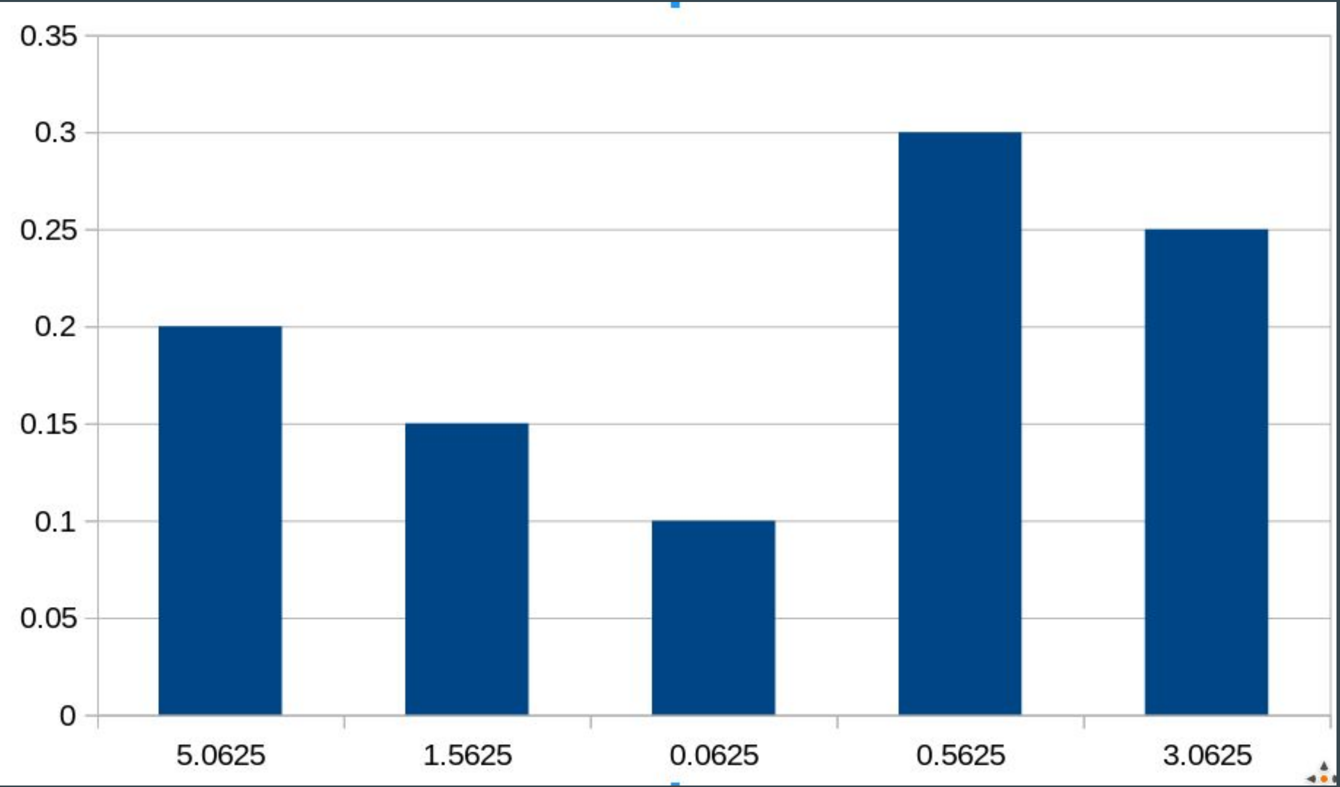
| X | P(X=x) |
|---|---|
| 1 | 0.2 |
| 2 | 0.15 |
| 3 | 0.1 |
| 4 | 0.3 |
| 5 | 0.25 |

# Example EV




$\mu = 3.25$

# Example variance

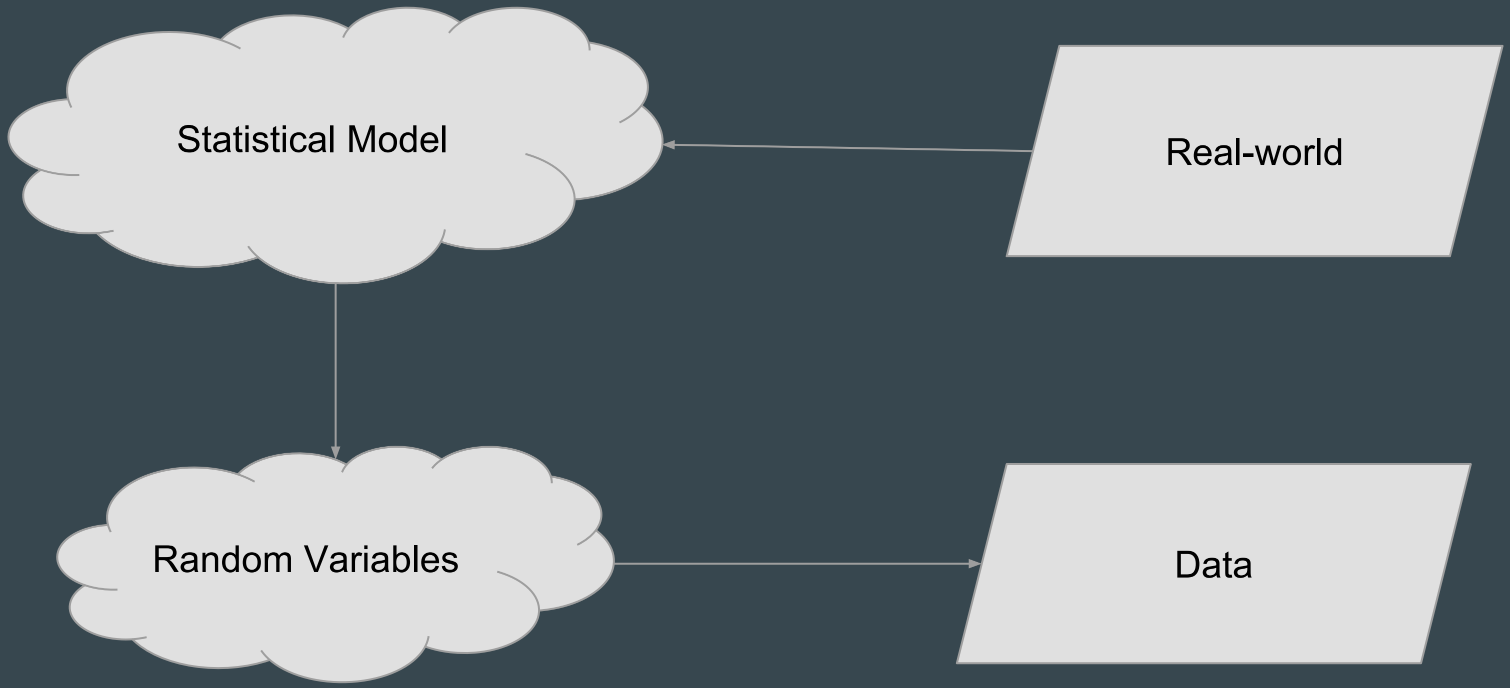



Var(X) = 2.1875

# Review

- Probability is the chance of future events happening
    - Or certain processes unfolding in a certain way
- When we think probabilistically, we are thinking *infinitely*
- We want to do this, because data analysis is about making an argument that your *small slice of data* says something about the *vast quantity of potential data in the real-world*

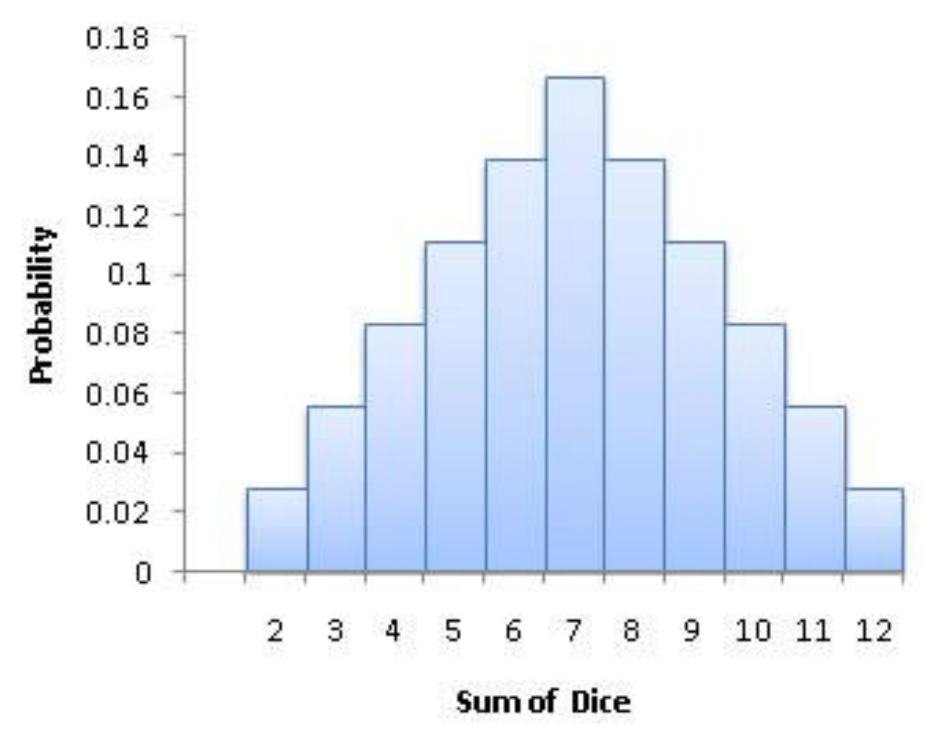
# Review



# Review

- Random variables are *predictable* in aggregate
- They are *uncertain*
- In other words, they *vary* (hence: variable)
- The mathematical function that describes this variability is a *probability distribution*
    - The probability distribution for discrete or ordinal data is called a *Probability Mass Function (PMF)*

# Back to Sigma-notation

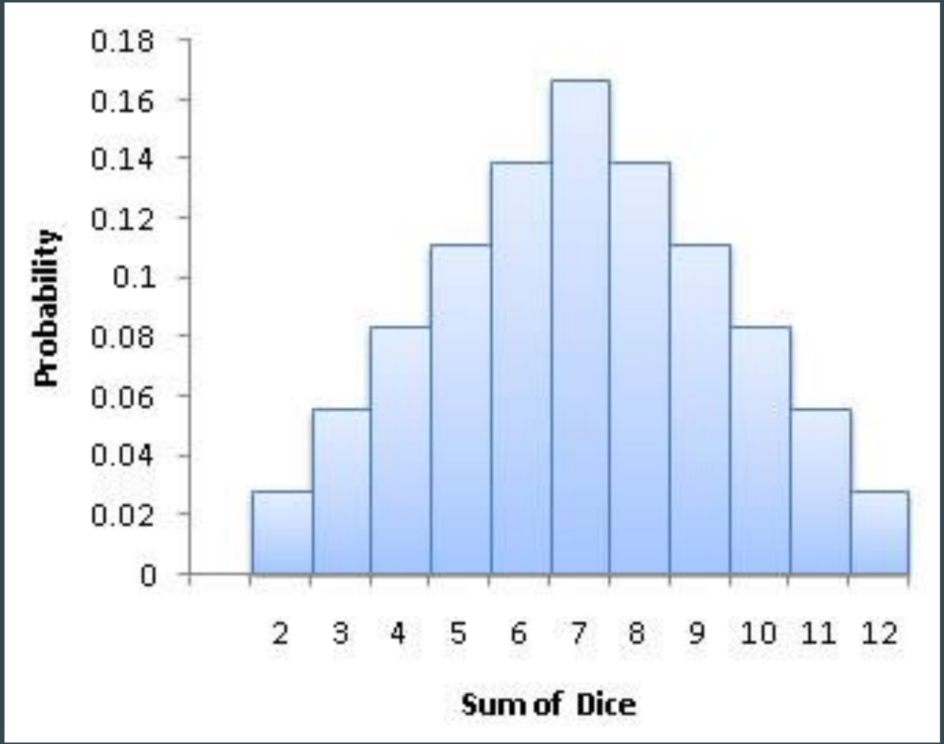



$\Omega$: 2-12 (x-axis)

$F$: $\sum(x_i)/N$

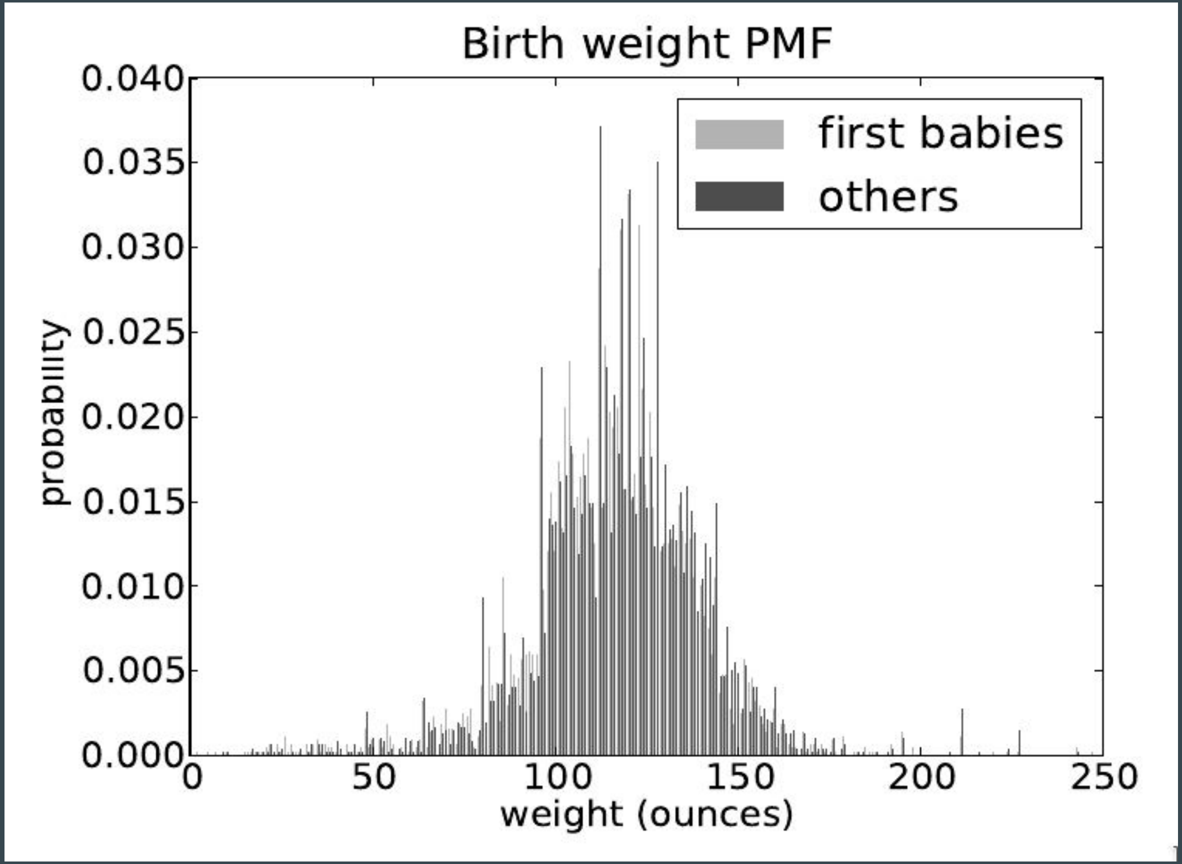
P: PMF

# PMF

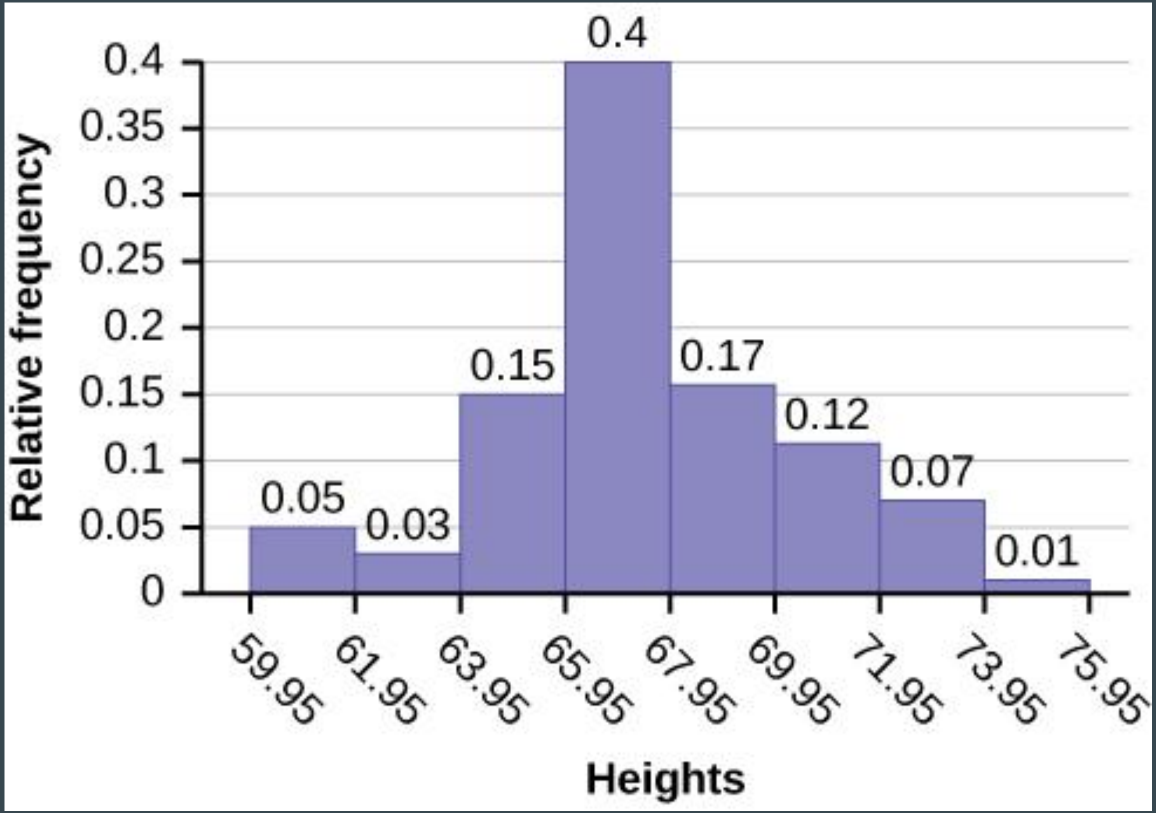



- Each value of x, has a corresponding value f(x)
- f(x) is the chance that x happens
  - f(2) = 3%
  - f(7) = 17%
- Straight forward for small range of values
- What's f(x<7)?

# PMF with a large range of values



# Percentiles



# Percentiles

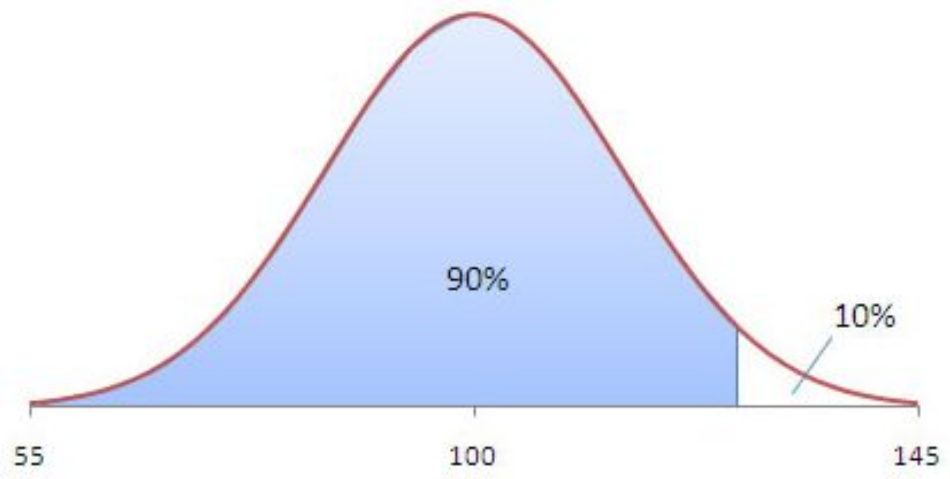
Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:


That means you are at the **80th percentile**.

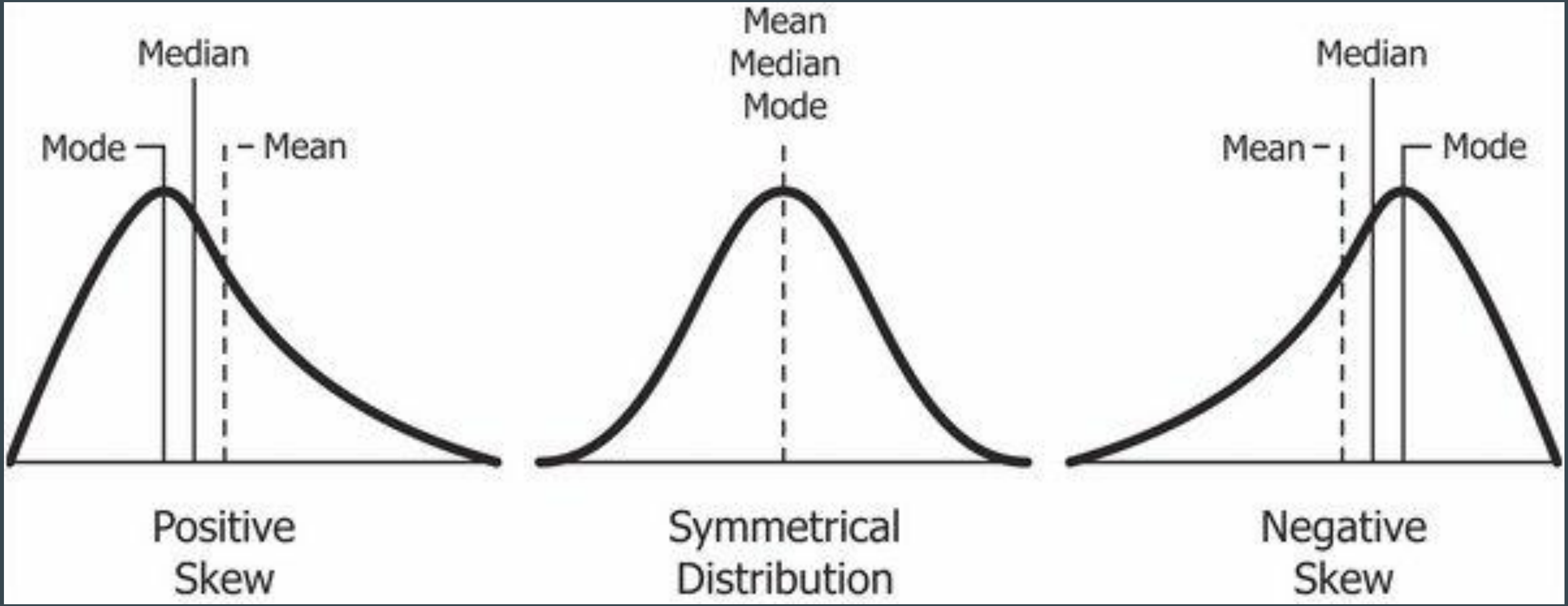
If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

# Percentiles




- The 90th percentile is
  - P(x=.9) or
  - P(x=.89) or...
  - P(x=.01)
- $P(X < x) = .9$

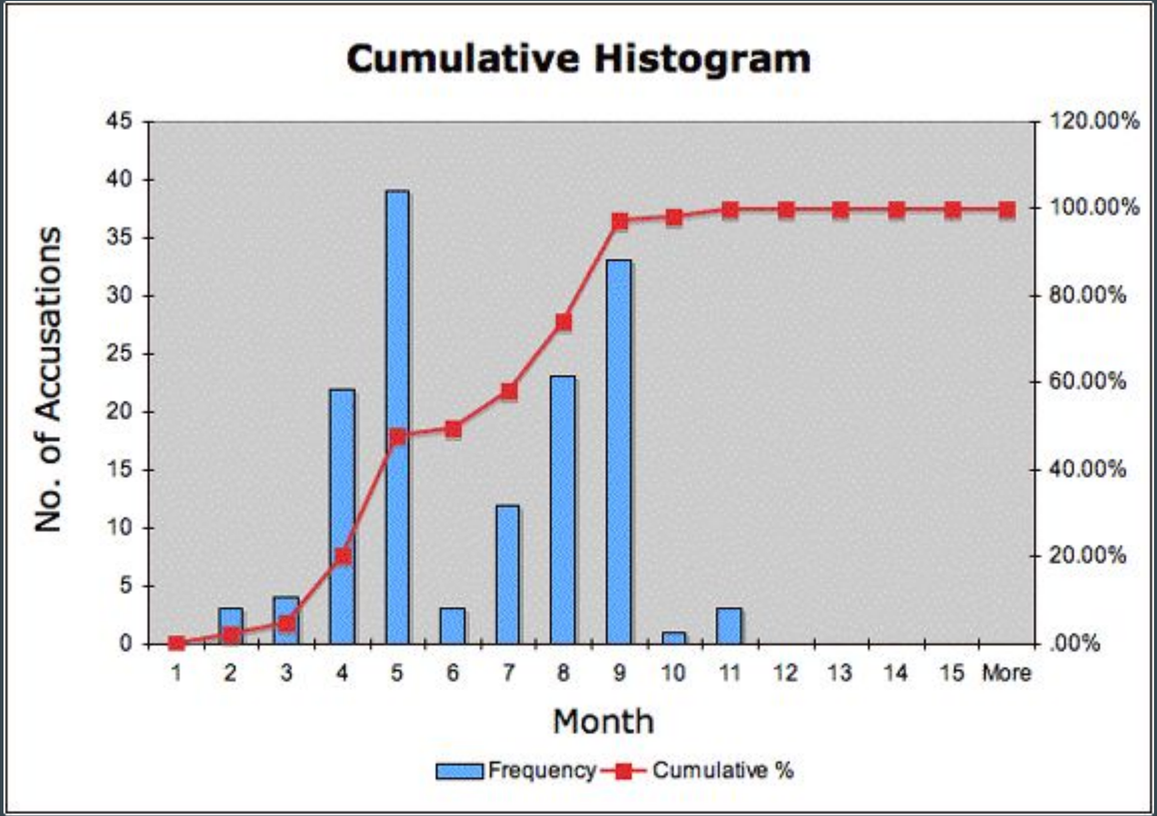
# What do percentiles tell us?



# Cumulative Distribution Function

- Instead of $P(X=x)$ (PMF)
- CDF: $P(X < x)$
- In other words this is a graph where
  - X-axis is the range of possible values
  - Y-axis is the *percentile*
- Note that for a discrete variable, $P(X < 2) = P(X=0) + P(X=1) + P(X=2)$

# CDFs


Cumulative Histogram
No. of Accusations
45
40
35
30
25
20
15
10
5
0
120.00%
100.00%
80.00%
60.00%
40.00%
20.00%
.00%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
More
Month
Frequency
Cumulative %

# CDF properties

- CDF never decreases
  - $P(X <2)$ cannot be less than $P(X<1)$
- As X approaches its minimum value, CDF approaches 0
- As X approaches its maximum value, CDF approaches 1

# CDF Example

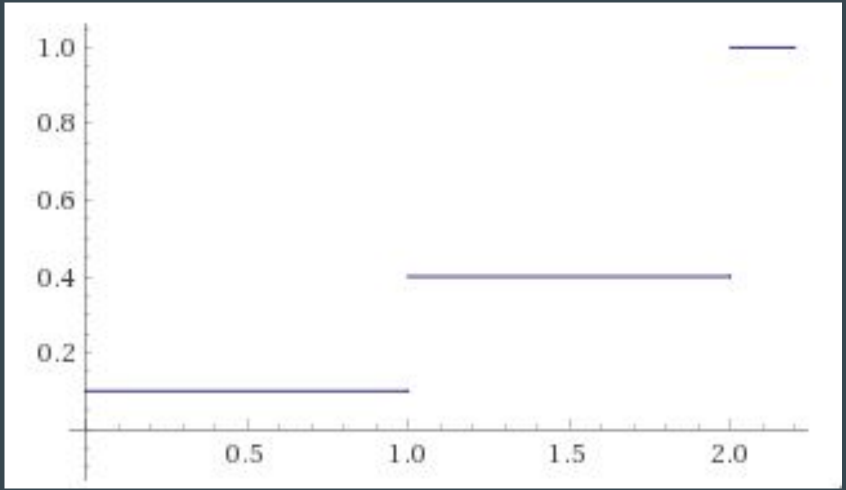
| Number of seats R's gain in Senate | Probability |
|---|---|
| 0 | 0.10 |
| 1 | 0.30 |
| 2 | 0.60 |

# CDF Example

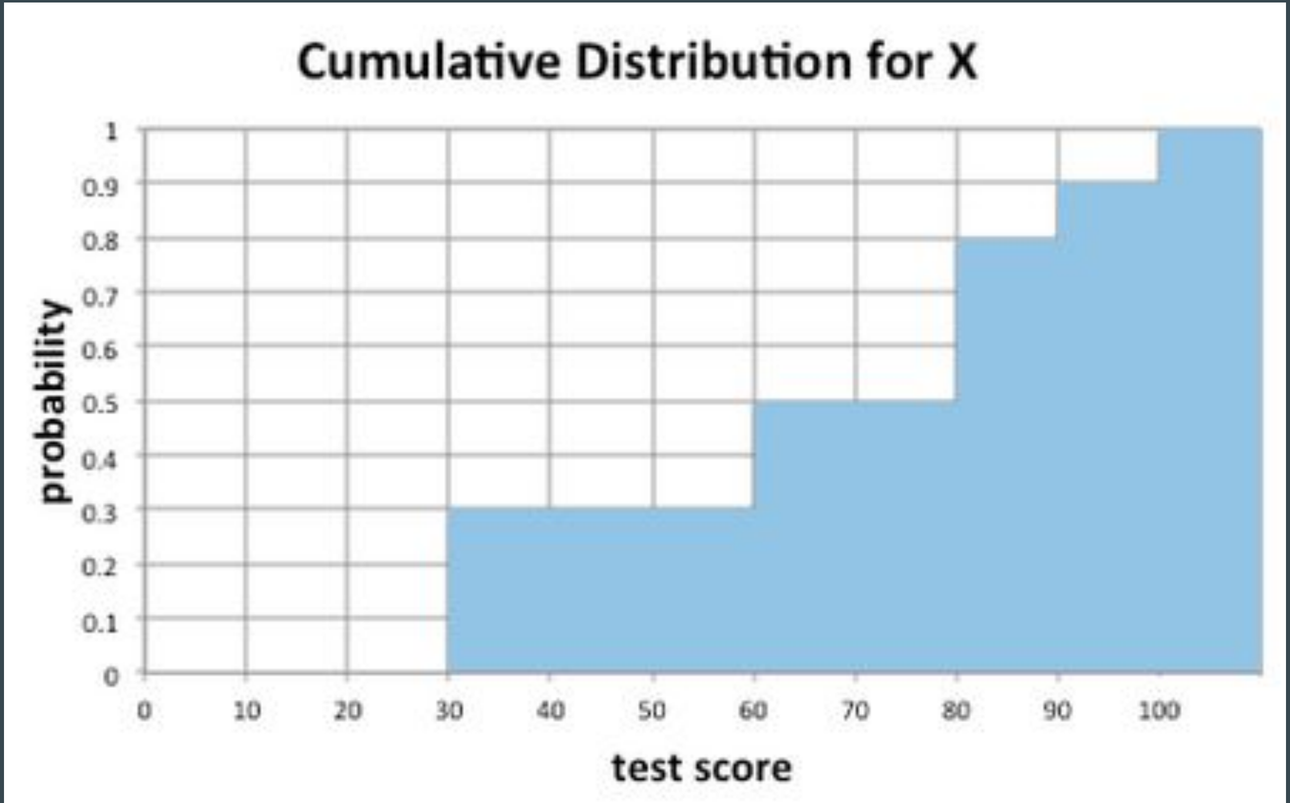
| Number of seats R's gain in Senate | Probability |
|---|---|
| X < 0 | 0.10 |
| X < 1 | 0.40 |
| X < 2 | 1 |

# CDF Example

$$f(x) = \begin{cases} 0.1 & 0 < x < 1 \\ 0.4 & 1 < x < 2 \\ 1 & 2 < x \end{cases}$$

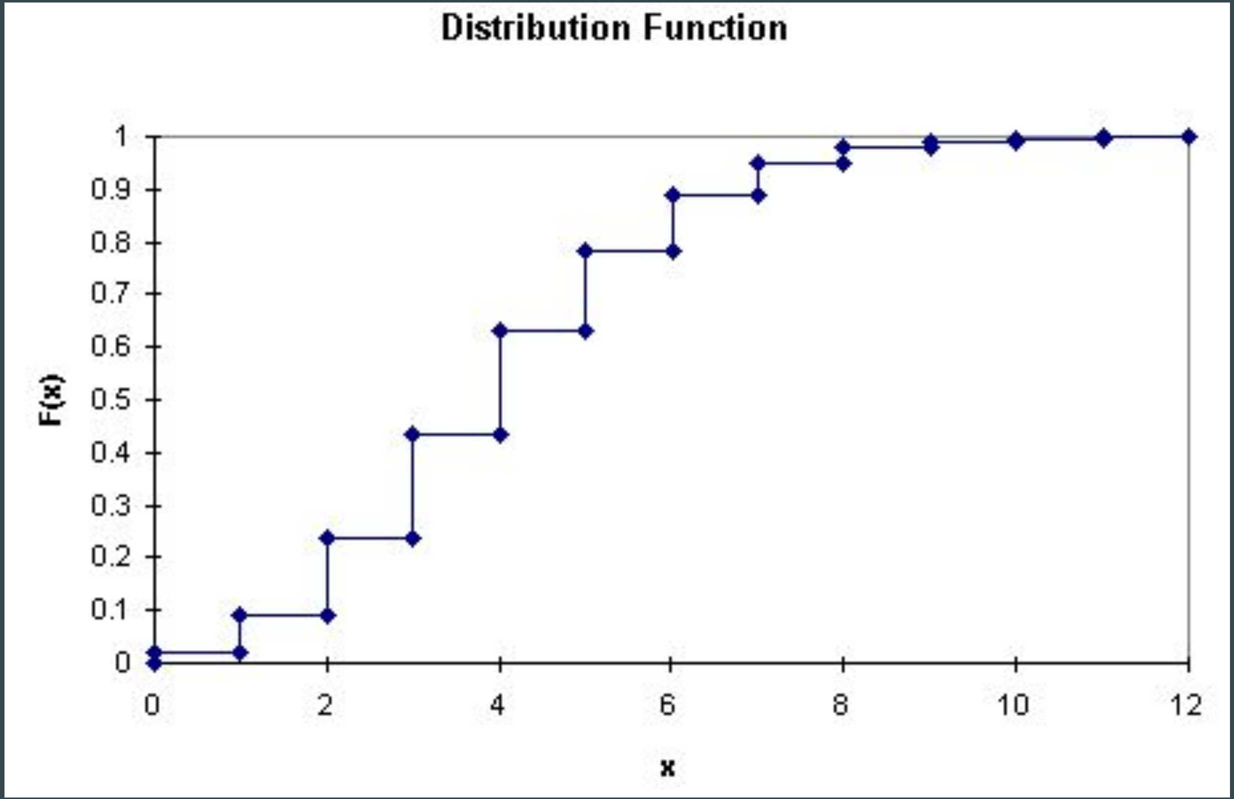


# $P(X < x)$

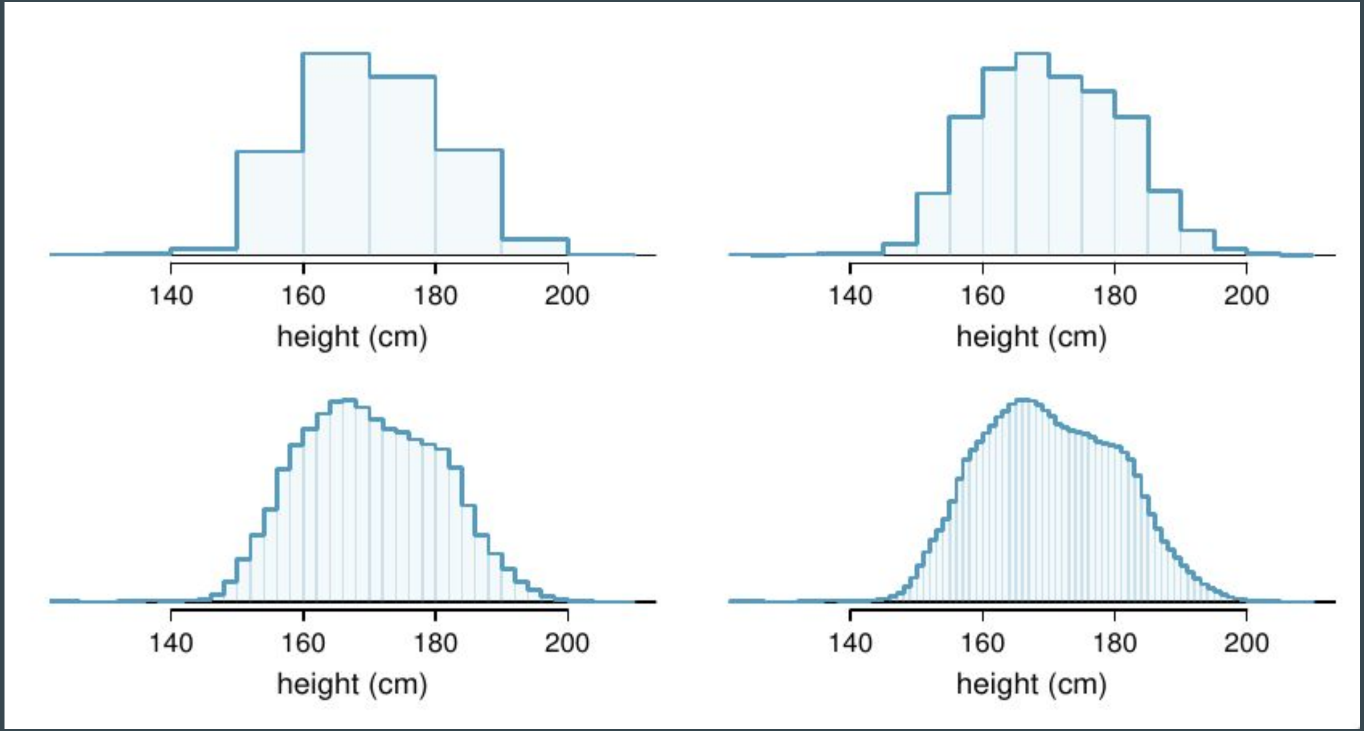


Cumulative Distribution for X
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
probability
0
10
20
30
40
50
60
70
80
90
100
test score

# P(X < x)

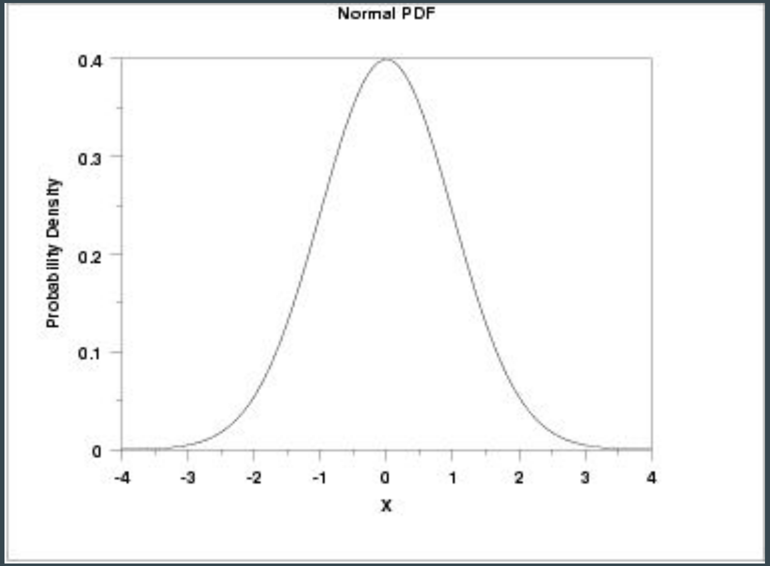

Distribution Function
F(x)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
x

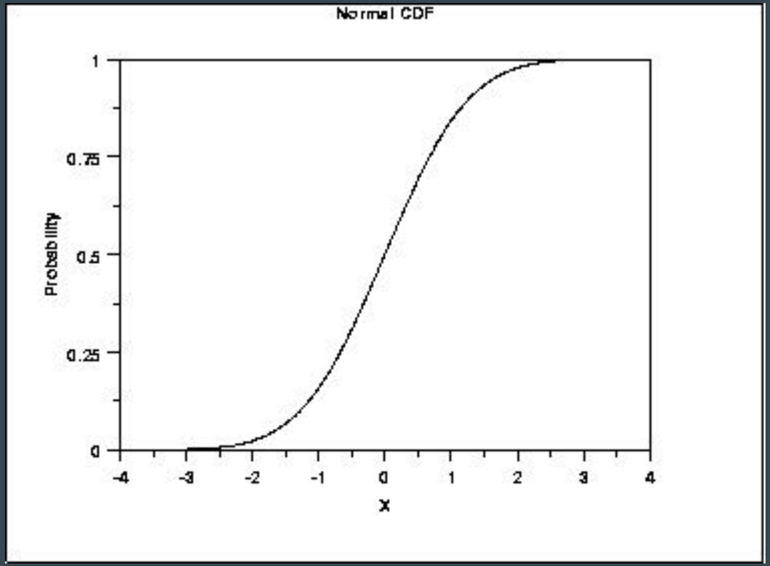
# What about continuous variables?



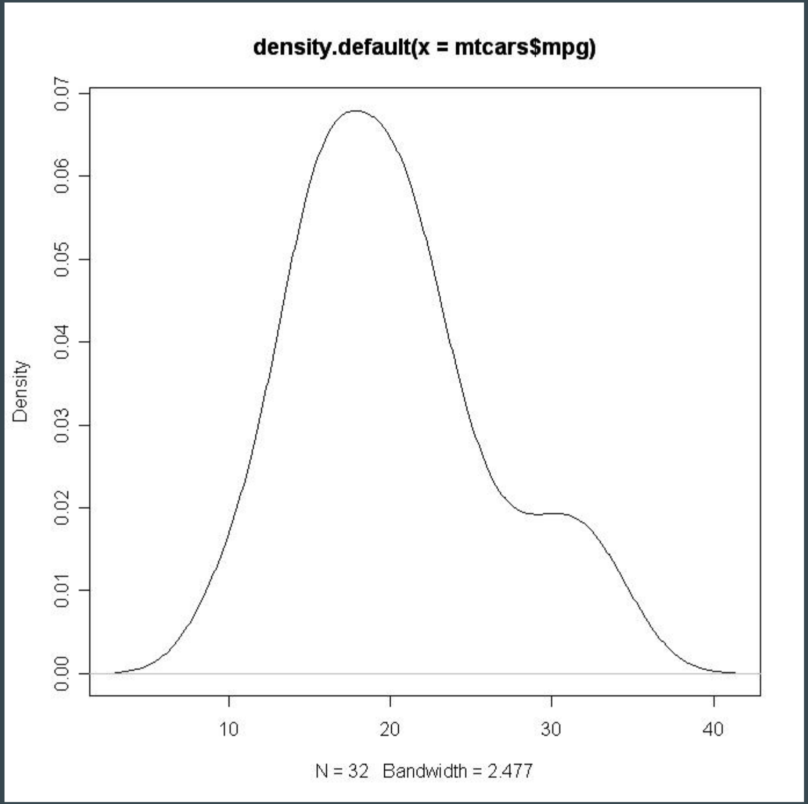
# Probability Density Function (PDF)

- PDF is the same as PMF, but for continuous variables



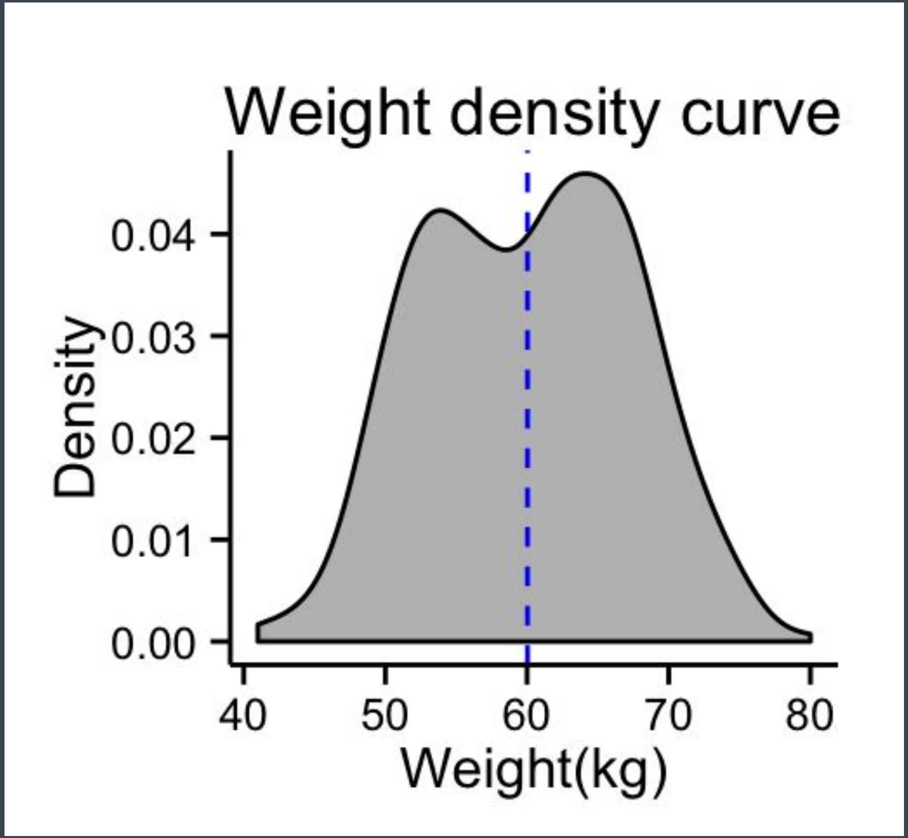



# PDF example

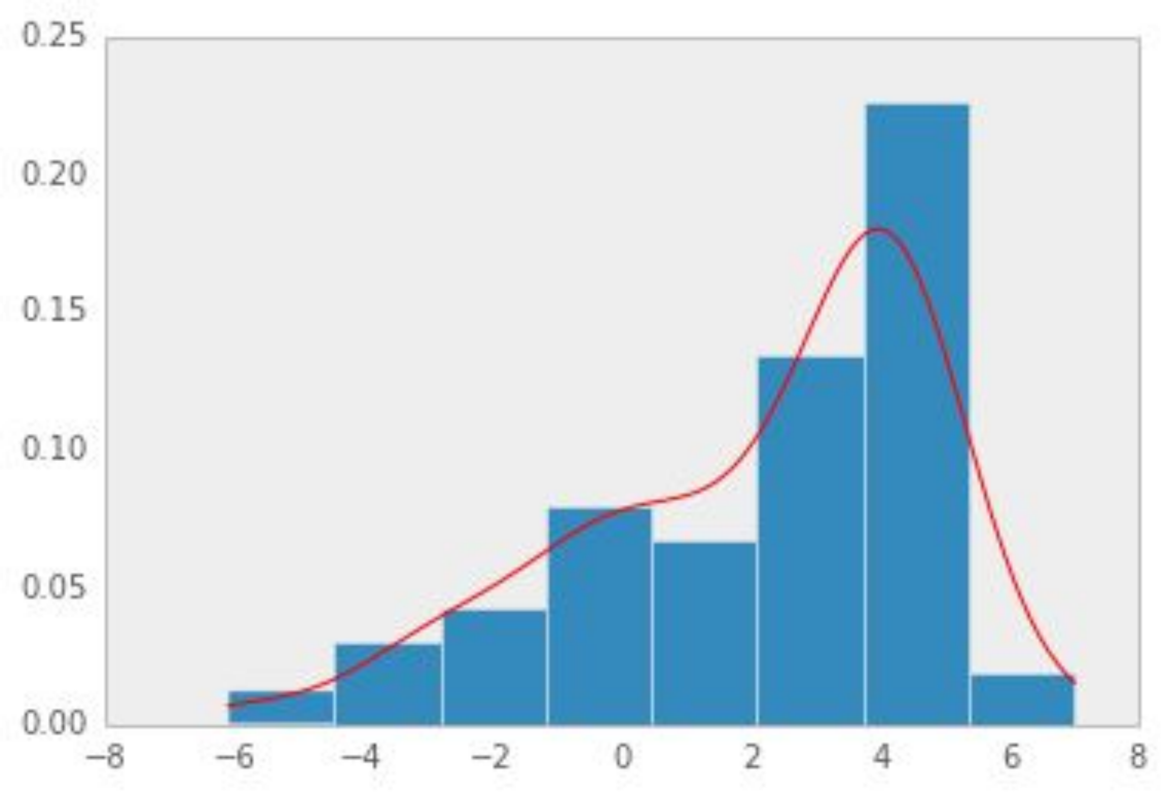


density.default(x = mtcars$mpg)
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
10
20
30
40
N = 32   Bandwidth = 2.477

# PDF example


Weight density curve
Density
0.04
0.03
0.02
0.01
0.00
40
50
60
70
80
Weight(kg)

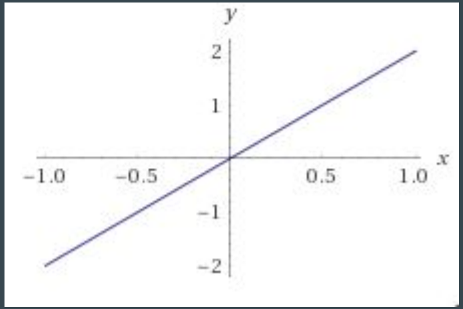
# But how can we calculate the PDF?



# Back to functions

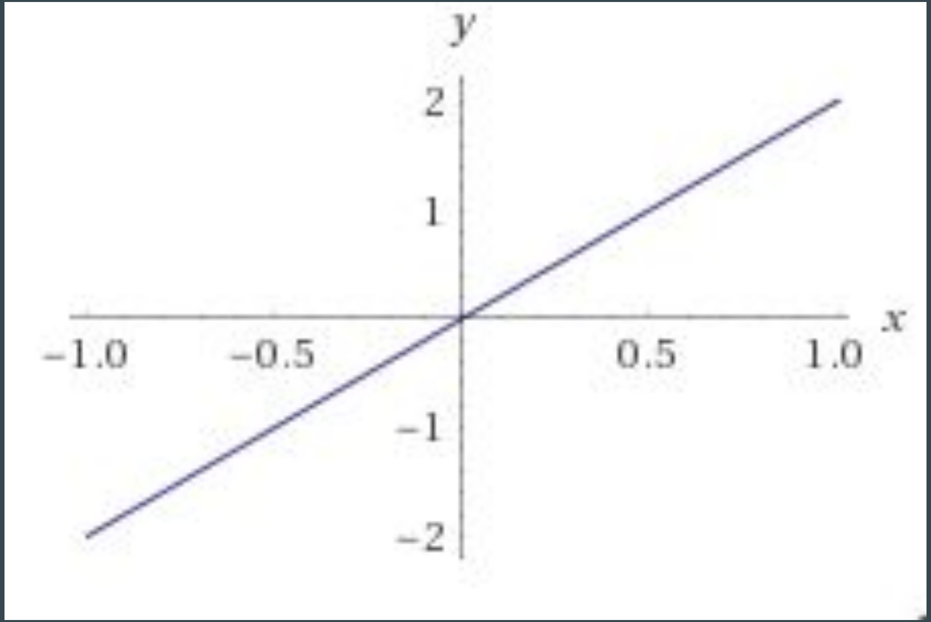
- A function has three parts:
  - Input
  - Transformation
  - Output
- Example: $f(x) = 2x$

| Input | Transformation | Output |
|---|---|---|
| 2 | 2(2) | 4 |
| 3 | 2(3) | 6 |
| 4 | 2(4) | 8 |



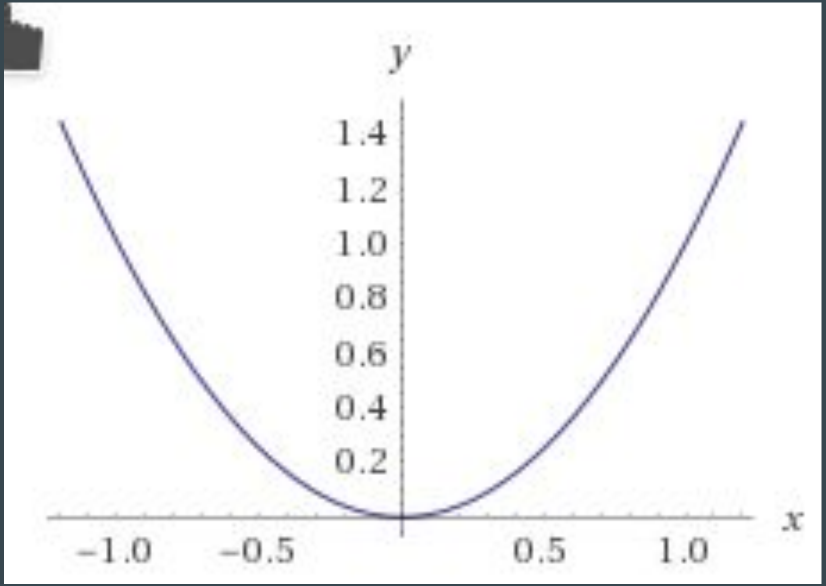
# Single variable functions are lines

The function: $y = 2x$ or $f(x) = 2x$ is a *single-variable* function over a two dimensional space

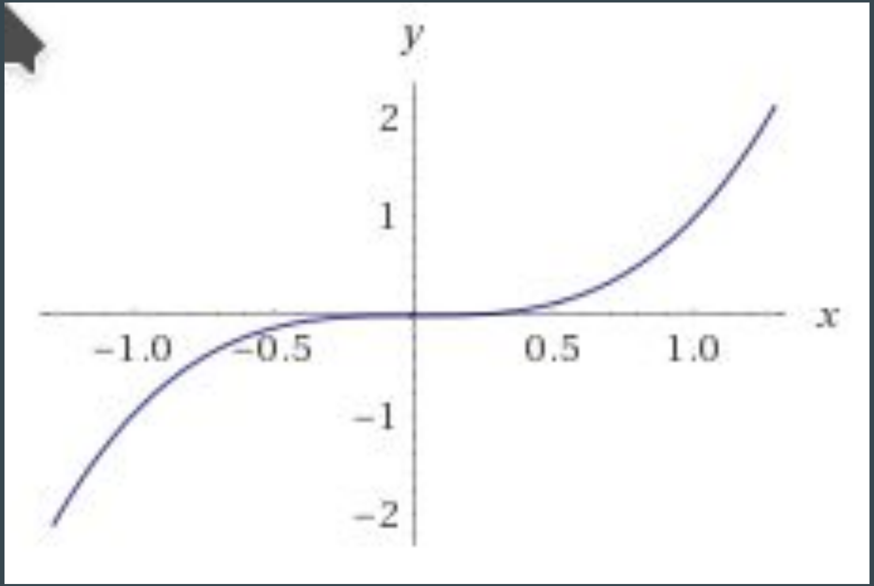


# Exponential functions curve lines

$f(x) = x^2$

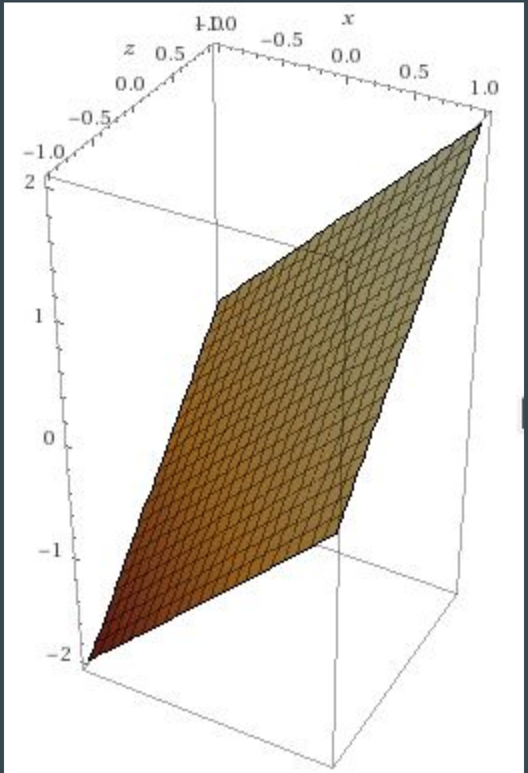



$f(x) = x^3$



# Multivariable functions are planes

$f(X_1, X_2) = X_1 + X_2$



# Review

- PMFs express the *probability distribution* for discrete and ordinal data
- PDFs express the *probability distribution* for continuous data
  - X-axis: possible values of x
  - Y-axis: probability that x happens
- CDFs express the *cumulative probability distribution* for all types of data
  - X-axis: possible values of x
  - Y-axis: percentile