Linear Models Lecture 1

Robert Gulotty

University of Chicago

May 30, 2023

Goals

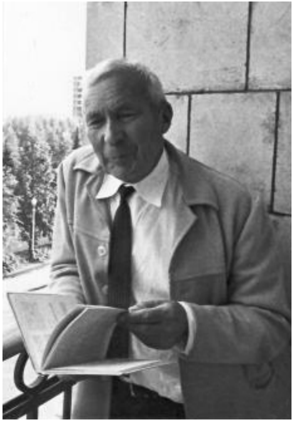
- We will be modeling data.
  - Summarizing complex data with less complicated formula (data reduction).
  - Drawing inferences about causal processes.
  - Making predictions.
- To do so, we will either
  - Make direct assumptions about the way the data was generated.
  - Use statistical theory to approximate the way the data was generated.
- In both cases, we use random variables called 'statistics', their distributions and their relations.
- Today we review the mathematics of random variables, their distributions, and their summaries.

Basic Ideas of Probability

Probability models random phenomenon with three ingredients:

1) **Sample space**: {all possible outcomes of a random process}
2) **Events**: {subsets of all possible outcomes of a random process}
3) **Probability Law**: a function $P(\cdot)$ which gives the relative chance of an event, a non-negative number.

Founder of Axiomatic Probability



Kolmogorov (1933).

Sample Spaces: All Things that Can Happen

Definition

The ***sample space*** *is the set of all things that can occur. This set is often referred to by the symbol* Ω *or* S.

Examples:

1) Discrete: The outcome of two dice
    - $\Omega = \{$ ⚀⚀, ⚀⚁, $\ldots,$ ⚅⚅ $\}$
2) Continuous: Value of a house
    - $\Omega = \{t : t \in \mathbb{R}\}$
3) Continuous: Survival of Monarch
    - $\Omega = \{t : 0 \leq t \leq 120\}$

Probability Law

- Probability: chance of event.
  - A function $P : \Omega \rightarrow [0, 1]$
  - Describes relative likelihood of events.
- Given an event A, $P(A)$ quantifies the chance it occurs.
- Let $E = \{x : x \in (a,\ b)\}$ be an event.

$$P(E) = \int_{x \in E} f_X(x) dx$$

where $f_X(x)$ is the *probability density function*, e.g.

  - Normal distribution: dnorm(x, μ, sd)
  - t distribution: dt(x, df)
  - F distribution: df(x, df1, df2)
- Note: the probability of continuous events are only positive for *intervals*.

Example: Candidate heights

- What is the probability a male candidate would be at least 72 inches tall by chance?

$$\Omega = \{h : 20 \leq h \leq 108\}, \quad E = \{h : 72 \leq h \leq 108\}$$

- Male heights are close to a normal distribution with mean (μ) 70 inches with a standard deviation (σ) of 3 inches.

$$f_H(h) = dnorm(h, \mu = 70, sd = 3)$$

$$\begin{aligned} P(E) &= \int_{72}^{108} f_H(h) dh \\ &= integrate(dnorm, lower = 72, upper = 108, mean = 70, sd = 3) \quad = .252 \end{aligned}$$

Aside on Calculating integrals

- In R there is a function pnorm(x) that calculates the integral of the normal distribution from $-\infty$ to x.
- We can then rewrite our problem in terms of these functions.

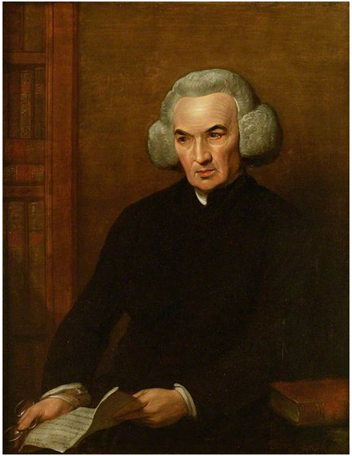
$$
\begin{aligned}
P(E) &= \int_{72}^{108} f_H(h)dh \\
&= \int_{-\infty}^{108} f_H(h)dh - \int_{-\infty}^{72} f_H(h)dh \\
&= F(H \leq 108) - F(H \leq 72) \qquad \textit{where F() is the anti-derivative of f()} \\
&= pnorm(108, 70, 3) - pnorm(72, 70, 3) \\
&= 1 - 0.747
\end{aligned}
$$

Example: Short and Tall

- What is the probability a male candidate would either at least 72 inches tall or below 65 inches?

$$
\begin{aligned}
E_1 &= \{h : 72 \leq h \leq 108\} \\
E_2 &= \{h : 20 \leq h \leq 65\} \\
E &= E_1 \cup E_2 \\
P(E) &= P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \\
P(E_1 \cup E_2) &= \int_{20}^{65} f_H(h)dh + \int_{72}^{108} f_H(h)dh - 0 \\
&= [F(H \leq 65) - F(H \leq 20)] + 0.25 \\
&= [pnorm(65, 70, 3) - 0] + 0.25 \\
&= .048 + 0.25
\end{aligned}
$$

Formalizer of Bayes' Rule



Price (1784).

Definition of Conditional Probability

The *conditional probability* of an event A given an event B is given by:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{for all } P(B) \neq 0$$

$P(A \cap B)$ is called the joint probability.
We can rearrange this to form the "Multiplication Rule":

$$P(A \cap B) = P(A|B)P(B)$$

Examples of Conditional probability

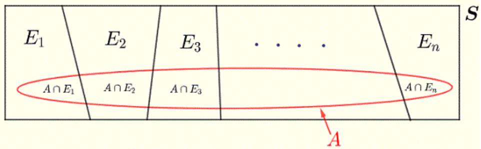
If A is Trump runs again, and B is the event that Trump wins:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$P(B|A^c) = \frac{P(B \cap A^c)}{P(A^c)} = \frac{0}{P(A^c)} = 0$$

Law of Total Probability

- E_1, E_2 are mutually exclusive if $E_1 \cap E_2 = \emptyset$
- If a given set of mutually exclusive events, $E_1, E_2, E_3 \ldots E_n$, their union forms the sample space Ω we say $E_1, E_2, E_3 \ldots E_n$ *partitions* the sample space.



- Divide and conquer: turn big problem into small problems:

Theorem

Law of Total Probability: Given such a partition, the probability of any event A is:

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \ldots + P(A|E_n)P(E_n)$$

Proof of Law of Total Probability

Proof.

$$
\begin{aligned}
P(A) &= P(A \cap \Omega) && \text{because } A \subseteq \Omega \\
P(A) &= P(A \cap (E_1 \cup E_2 \cup \ldots \cup E_n)) && \text{E partitions } \Omega \\
P(A) &= P((A \cap E_1) \cup (A \cap E_2) \cup \ldots \cup (A \cap E_n)) && \text{Distributive Law} \\
P(A) &= P(A \cap E_1) + P(A \cap E_2) + \ldots P(A \cap E_n) && E_i \cap E_j = \emptyset \\
P(A) &= P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \ldots + P(A|E_n)P(E_n) && \text{multiplication rule}
\end{aligned}
$$

$\square$

Independence

Definition

Suppose we have two events E_1, E_2. We say these events are independent if:

$$P(E_1 \cap E_2) = P(E_1)P(E_2)$$

- To discover the systematic component of the experiment, we need to make assumptions about what affects what.
- The most important assumption is the notion of independence.

Independence and Information

Does one event provide information about another event?

Definition

If B does not change the probability that A occurs, A is independent of B.

$$P(A|B) = P(A)$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$$

Independence is symmetric: if F is independent of E, then E is independent of F

Conditional Independence

Definition

Let E_1 and E_2 be two events. We will say that the events are conditionally independent given E_3 if

$$P(E_1 \cap E_2 | E_3) = P(E_1 | E_3) P(E_2 | E_3)$$

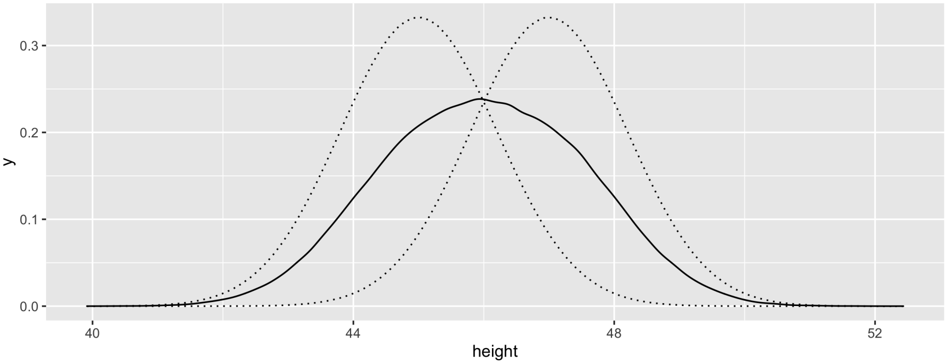
Amusement Park Example

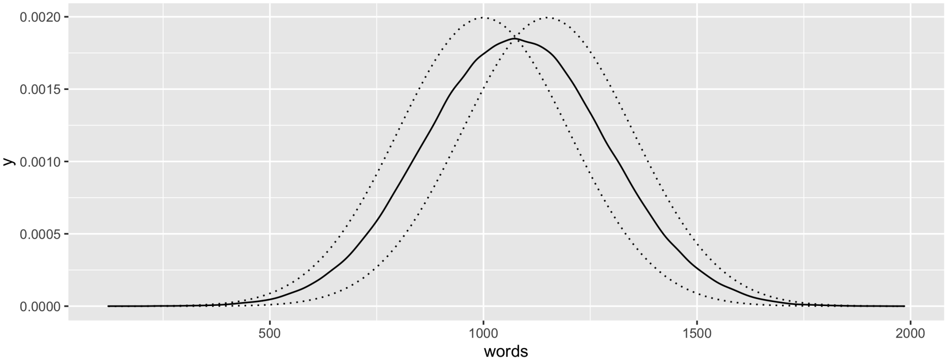
- There is a strong correlation between child height and literacy (a spurious relationship).
- Suppose there is an Amusement Park popular with 1st graders (6-7 year olds).
- Can we rely on the children understanding this sign?

Statistical Model of Height and Literacy

- x is the height of a child where $f(x) = N(\mu = 33 + z * 2, s = 1.2)$.
- y is $\#$ words the child knows $f(y) = N(\mu = 100 + z * 150, s = 200)$.
- z is the child's age, where $f(z) = \begin{cases} 6 \text{ with probability } 1/2 \\ 7 \text{ with probability } 1/2 \end{cases}$
- The probability that a first grader can read the "you must be this tall sign" and is at least 48" tall:

$$P(A \cap B) \text{ where } A = \{x : x > 48\},\ B = \{y : y > 1400\}$$


0.3
0.2
0.1
0.0
y
40
44
48
52
height


0.0020
0.0015
0.0010
0.0005
0.0000
y
500
1000
1500
2000
words

Using Law of Total Probability

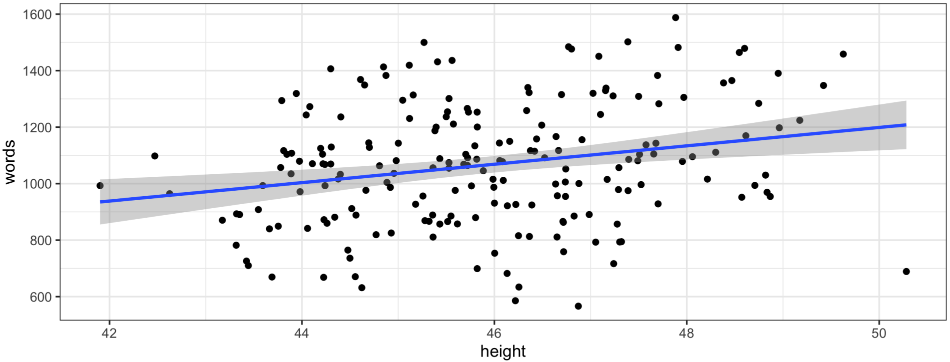
$$
\begin{aligned}
A &= \{x : x > 48\} \\
P(A) &= P(A|Z=6) * P(Z=6) + P(A|Z=7) * P(Z=7) \qquad \text{(Law of Total Probability)} \\
&= [\int_{48}^{\infty} \phi(x|45, 1.2)dx] * P(Z=6) + [\int_{48}^{\infty} \phi(x|47, 1.2)dx] * P(Z=7) \\
&= (1 - pnorm(48, 45, 1.2)) * 1/2 + (1 - pnorm(48, 47, 1.2)) * 1/2 \\
&= 0.104
\end{aligned}
$$

By similar reasoning, $B = \{y : y > 1400\}$, $P(B) = .06$.

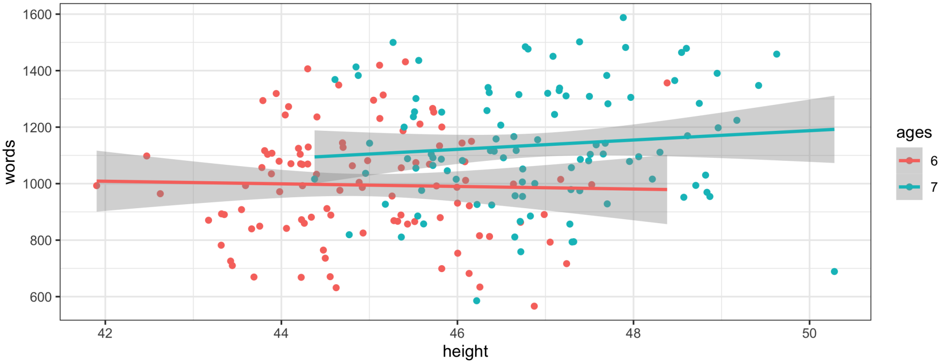
Using conditional independence

We know *height* and *literacy* are conditionally independent.

$$\begin{aligned}
P(A \cap B) &= P(A \cap B|Z=6)P(Z=6) + P(A \cap B|Z=7)P(Z=7) \\
&= P(A|Z=6)P(B|Z=6)P(Z=6) + P(A|Z=7)P(B|Z=7)P(Z=7) \\
&= (1 - pnorm(48, 45, 1.2)) * (1 - pnorm(1400, 1000, 200)) * 1/2 + \\
&\quad (1 - pnorm(48, 47, 1.2)) * (1 - pnorm(1400, 1150, 200)) * 1/2 \\
&= .01
\end{aligned}$$



1600
1400
1200
1000
800
600
words
42
44
46
48
50
height



words
height
1600
1400
1200
1000
800
600
42
44
46
48
50
ages
6
7

	DV: height	
	(1)	(2)
words	0.002***	0.0002
	(0.001)	(0.0004)
ages		1.976***
		(0.183)
Constant	44***	32.8***
	(0.552)	(1.130)
Observations	200	200
R^2	0.057	0.407

Random Variables

Given a sample space Ω, and a probability law P:

Definition

A random variable is a function that assigns real numbers (usually) to events in a sample space Ω.

$$X(\omega) : \Omega \to \mathbb{R}$$

Connecting random variables to probability

- X assigns some numbers to events.

$$X(\omega) : \Omega \to \mathbb{R}$$

- Remember, probability assigns a chance to an event.

$$P(\omega) : \Omega \to \mathbb{R}$$

- What are the probabilities associated with $X(\omega)$?

Cumulative Distribution Function: F(x)

Random variables are characterized by cumulative distribution functions.

Definition

Cumulative Distribution function. For a continuous random variable X define its cumulative distribution function $F(\cdot)$ as,

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

PDFs, CDFs, OLS

Gauss (1809)

Probability Density Function: f(x)

Definition

If X is a continuous random variable, the probability density function of X is the function $f_X(x)$ that satisfies.

$$
\begin{aligned}
F_X(x) &= P(X \leq x) \\
&= P(X \in (-\infty,\ x)) \\
&= \int_{-\infty}^{x} f_X(t)dt \qquad x \in \mathbf{R}
\end{aligned}
$$

$$
\begin{tikzcd}
\text{PDF} \arrow[rr, bend left, "\text{Integrate}"] & & \text{CDF} \arrow[ll, bend left, "\text{Differentiate}"]
\end{tikzcd}
$$

Example: continuous uniform

Definition

Y has a uniform distribution on the interval (a, b) if

$$f_Y(y) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq y \leq b \\ 0 & \text{otherwise} \end{cases}$$

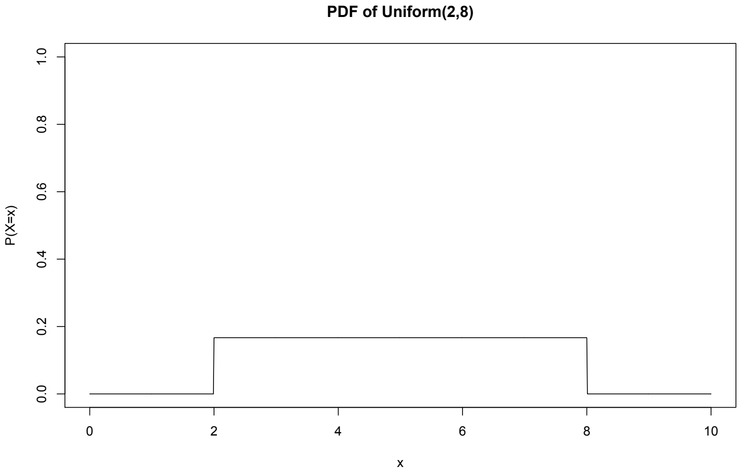
$$F_Y(y) = \begin{cases} 1 & \text{if } b < y \\ \frac{y-a}{b-a} & \text{if } a < y < b \\ 0 & \text{if } y < a \end{cases}$$

Example: Uniform

Example: Suppose that we are waiting for Comcast to show up and install our cable package. They say that they may arrive between 2:00 and 8:00. Without any further information, you may have no reason to suspect any particular time over any other.

$$X \sim U(2, 8)$$

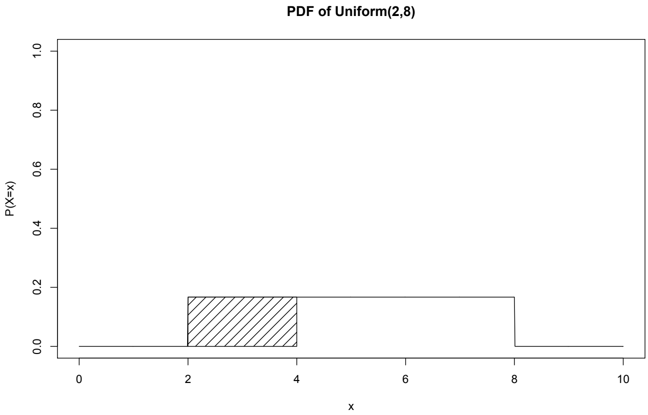
Example: PDF of continuous uniform



Example: continuous uniform

What is the probability that the cable installation truck arrives before 4:00?

Example: CDF of continuous uniform ($P(X \leq 4)$)



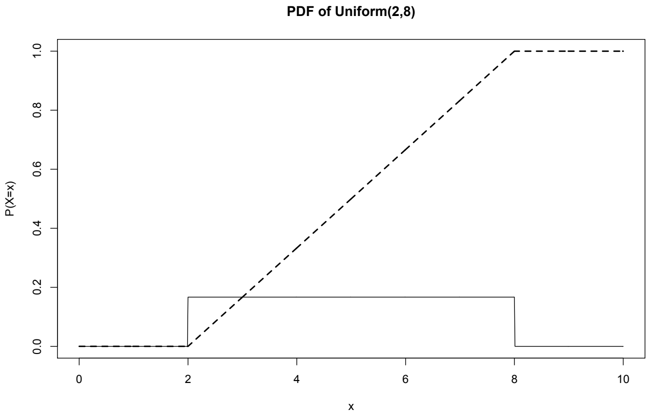
Example: continuous uniform

$$P(Y \leq y) = \int_{x=-\infty}^{y} f(x)dx = \int_{x=-\infty}^{y} \frac{1}{b-a}dx = \int_{x=a}^{y} \frac{1}{b-a}dx$$

$$\frac{y}{b-a} - \frac{a}{b-a} = \frac{y-a}{b-a}$$

$$= \frac{4-2}{8-2}$$

Example: continuous uniform



Definition of Expectation

What can we expect from a trial?

The expectation is the **value** of random variable weighted by the **probability** of observing that outcome.

Definition

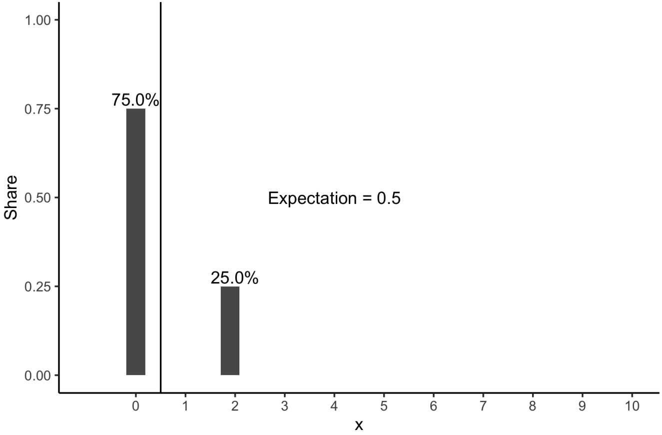
Expected Value: define the expected value of a function X as,

$$E[X] = \sum_{x:p(x)>0} xp(x) \qquad \textit{when } x \textit{ is discrete}$$

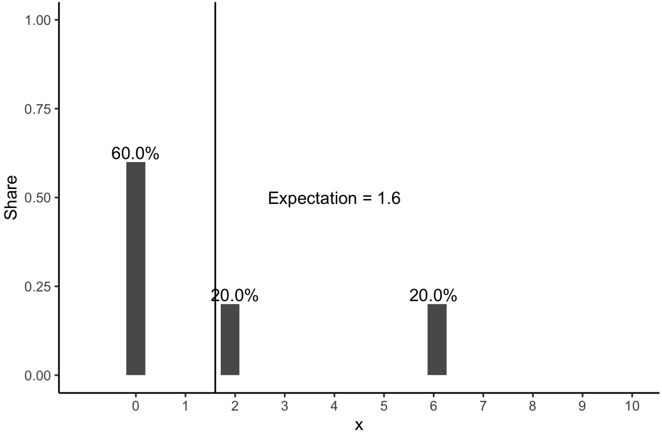
$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \qquad \textit{when } x \textit{ is continuous}$$

In words: for all values of x with $p(x)$ greater than zero, take the sum/integral of values times weights.

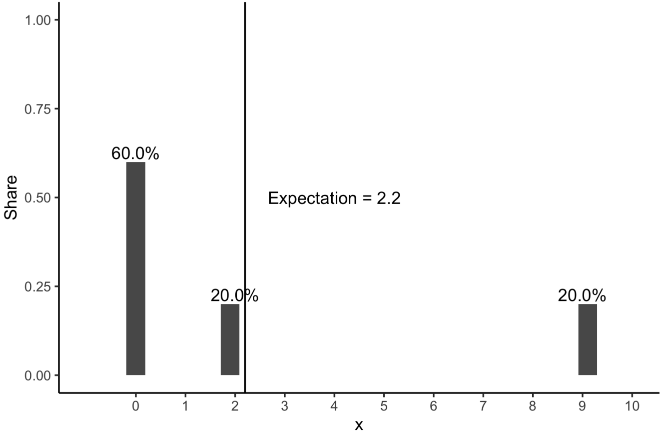
$$E(X) = \sum_{x:p(x)>0} x * p(x) = 0 * .75 + 2 * .25 = \frac{1}{2}$$



$$E(X) = \sum_{x:p(x)>0} x * p(x) = 0 * .6 + 2 * .2 + 6 * .2 = 1.6$$



$$E(X) = \sum_{x:p(x)>0} x * p(x) = 0 * .6 + 2 * .2 + 9 * .2 = 2.2$$



Expectation Properties

$$
\begin{aligned}
E[X+Y] &= E[X] + E[Y] \\
E[a] &= a \\
E[aX] &= aE[X] \\
E[E[X]] &= E[X] \\
E[XY] &\neq E[X] \times E[Y]
\end{aligned}
$$

Properties of the Expectation

$$E[a] = a$$

Proof: Suppose Y is a random variable such that $Y = a$ with probability 1 and $Y = 0$ otherwise:

$$\begin{aligned} E[Y] &= \sum_{y:p(y)>0} yp(y) \\ &= ap(Y = a) + 0 * p(Y = 0) \\ &= a * 1 + 0 \\ &= a \end{aligned}$$

Justification of Expectation

If we want to predict y with no other information, and our prediction is called π, one standard for prediction is to minimize the mean-square error:

$$\begin{aligned} M &= \int (y-\pi)^2 f(y)dy \\ &= E[(y-\pi)^2] \\ &= E[y^2 - 2\pi y + \pi^2] \\ &= E[y^2] - E[2\pi y] + E[\pi^2] \\ &= E[y^2] - 2\pi E[y] + \pi^2 \end{aligned}$$

Using calculus to minimize:

$$\begin{aligned} \frac{\partial M}{\partial \pi} &= -2E[y] + 2\pi \\ \pi &= E[y] \end{aligned}$$

Suppose $X \sim Uniform(3,5)$. What is $E[X]$?

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \int_{-\infty}^{3} x0dx + \int_{3}^{5} x\frac{1}{5-3}dx + \int_{5}^{\infty} x0dx \\
&= 0 + \frac{x^2}{4}|_3^5 + 0 \\
&= 0 + 5^2/4 - 3^2/4 + 0 \\
&= 4
\end{aligned}$$

Corollary

Suppose X is a continuous random variable. Then,

$$E[aX+b] = aE[X] + b$$

Proof.

$$\begin{aligned} E[aX+b] &= \int_{-\infty}^{\infty} (ax+b)f(x)dx \\ &= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx \\ &= aE[X] + b \times 1 \end{aligned}$$

□

Second Moment: Variance

Expected value is a measure of central tendency.
What about spread? Variance

- For each value, we might measure distance from center
  - Distance, squared $d(x, E[x])^2 = (x - E[x])^2$
- Then we might take weighted average of these distances by taking an expectation.

Two formulas for Variance

$$
\begin{aligned}
E[(X - E[X])^2] &= \sum_x (x - E[X])^2 p(x) \\
&= \sum_x \left(x^2 - E[X]x - xE[X] + E[X]^2\right) p(x) \\
&= \sum_x \left(x^2 - 2E[X]x + E[X]^2\right) p(x) \\
&= \sum_x x^2 p(x) - \sum_x 2xE[X]p(x) + \sum_x E[X]^2 p(x) \\
&= \sum_x x^2 p(x) - 2E[X]\sum_x xp(x) + \sum_x E[X]^2 p(x) \\
&= E[X^2] - 2E[X]^2 + E[X]^2 \\
&= E[X^2] - E[X]^2 = \text{Var}(X)
\end{aligned}
$$

Definition of Variance

Definition

The variance of a random variable X, $var(X)$, is

$$\begin{aligned} var(X) &= E[(X - E[X])^2] \\ &= \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx \\ &= E[X^2] - E[X]^2 \end{aligned}$$

- We will define the standard deviation of X, $\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$
- $\mathrm{var}(X) \geq 0$.
- We use σ^2 to indicate variance.

Variance Corollary

Corollary

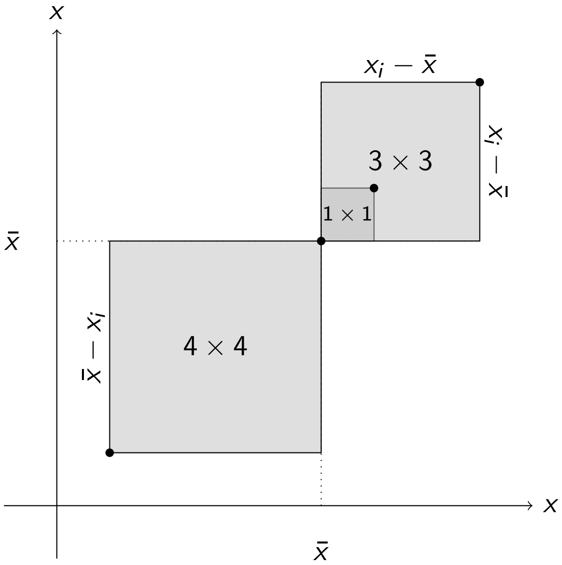
$Var(aX + b) = a^2 Var(X)$

Proof: Define $Y = aX + b$. We know that $Var(Y) = E[(Y - E[Y])^2]$.

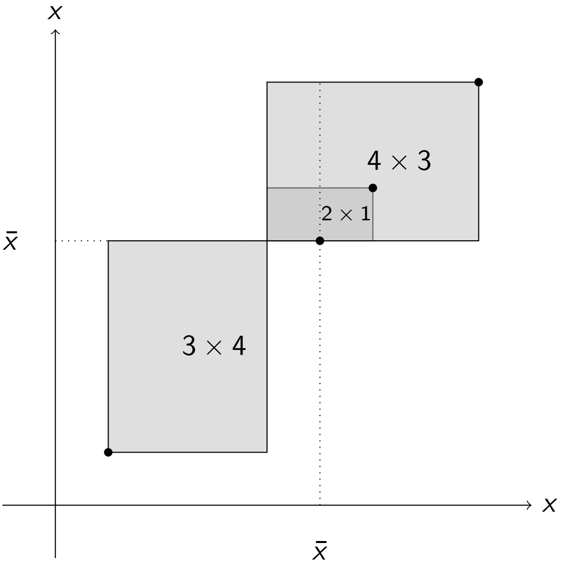
$$
\begin{aligned}
&= E[((aX + b) - E[aX + b])^2] \\
&= E[((aX + b) - (aE[X] + b))^2] \\
&= E[(aX - aE[X])^2] \\
&= E[(a^2X^2 - 2a^2XE[X] + a^2E[X]^2)] \\
&= a^2E[X^2] - 2a^2E[X]^2 + a^2E[X]^2 \\
&= a^2(E[X^2] - E[X]^2) \\
&= a^2Var(X)
\end{aligned}
$$

Often the sample (co)variance appears in an alternative form:

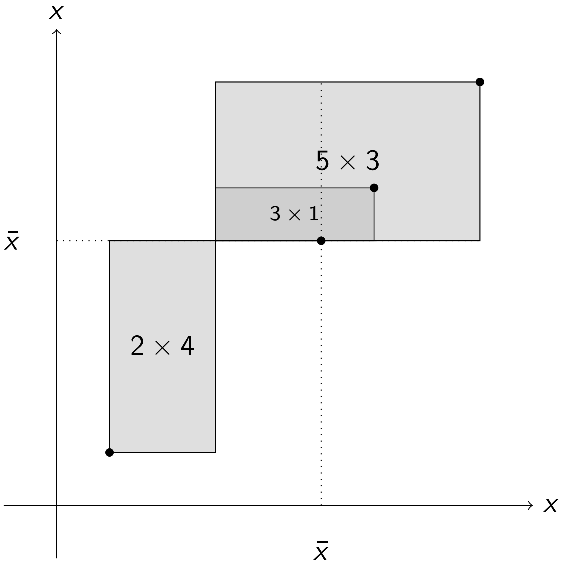
$$\begin{aligned}
\widehat{Var}(X) &= \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - \bar{x}) \\
&= \frac{1}{N}\sum_i x_i^2 - \bar{x}x_i - x_i\bar{x} + \bar{x}\bar{x} \\
&= \frac{1}{N}\sum_i x_i^2 - \bar{x}\frac{1}{N}\sum_i x_i - \bar{x}\frac{1}{N}\sum_i x_i + \frac{1}{N}\sum_i \bar{x}\bar{x} \\
&= \frac{1}{N}\sum_i x_i^2 - \bar{x}\frac{1}{N}N\bar{x} - \bar{x}\frac{1}{N}\sum_i x_i + \bar{x}\bar{x} && (\textstyle\sum_i x_i = N\bar{x}) \\
&= \frac{1}{N}\sum_i x_i^2 - \bar{x}\frac{1}{N}\sum_i x_i && (\bar{x}\bar{x} \text{ cancel}) \\
&= \frac{1}{N}\sum_i (x_i - \bar{x})x_i
\end{aligned}$$



$$\widehat{Var}(X) = \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - \bar{x})$$
$$= \frac{(-4 * -4) + (1 * 1) + (3 * 3)}{4}$$
$$= \frac{26}{4}$$



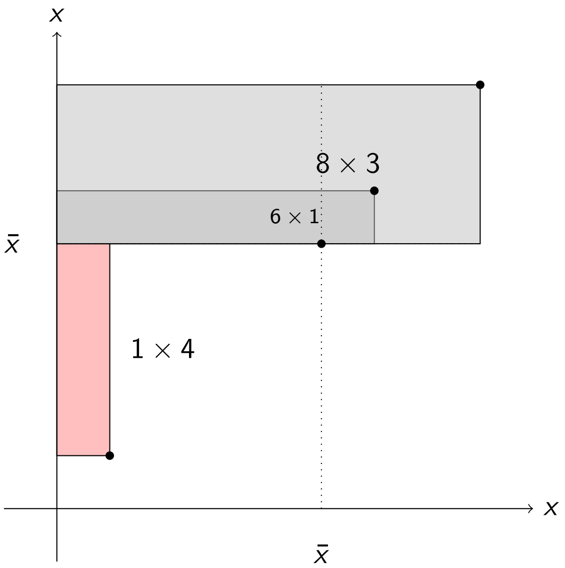
$$
\begin{aligned}
\widehat{Var}(X) &= \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - (\bar{x} - 1)) \\
&= \frac{(-4 * -3) + (1 * 2) + (3 * 4)}{4} \\
&= \frac{26}{4}
\end{aligned}
$$



$$\widehat{Var}(X) = \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - (\bar{x} - 2))$$

$$= \frac{(-4 * -2) + (1 * 3) + (3 * 5)}{4}$$

$$= \frac{26}{4}$$



$$\widehat{Var}(X) = \frac{1}{N}\sum_i (x_i - \bar{x})(x_i - (\bar{x} - \bar{x}))$$

$$\widehat{Var}(X) = \frac{1}{N}\sum_i (x_i - \bar{x})(x_i)$$

$$= \frac{(-4*1) + (1*6) + (3*8)}{4}$$

$$= \frac{26}{4}$$

Example of Variance: Uniform

$X \sim \text{Uniform}(0, 1)$. What is $Var(X)$?

$$E[X^2] = \int_0^1 X^2 \frac{1}{1-0} dx = \frac{X^3}{3}\Big|_0^1$$
$$= \frac{1}{3}$$
$$E[X]^2 = \left(\frac{1}{2}\right)^2$$

$$Var(X) = E[X^2] - E[X]^2$$
$$= \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Named Distributions: Normal

Definition

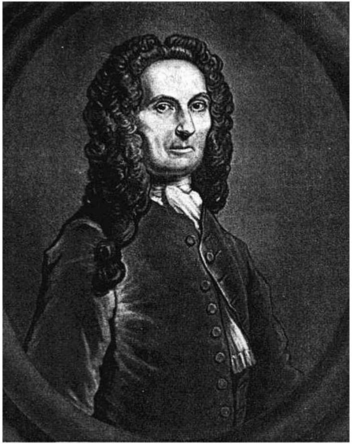
Suppose X is a random variable with $X \in \mathbf{R}$ and probability density function

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

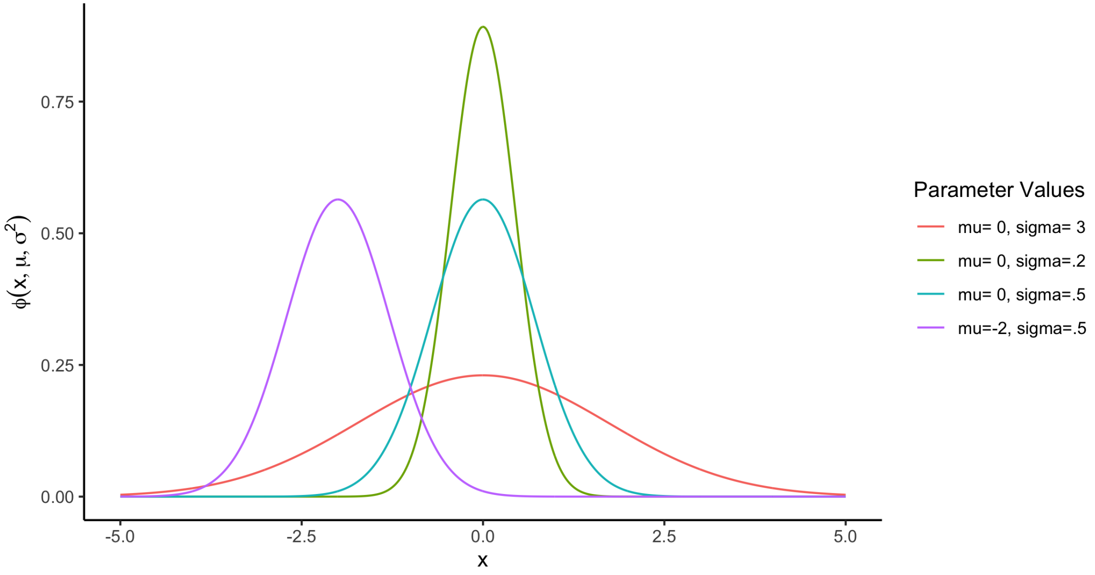
Then X is a normally distributed random variable with parameters μ and σ^2.
Equivalently, we'll write

$$X \sim Normal(\mu, \sigma^2)$$

Discoverer of Normal Distribution



De Moivre (1711)



0.75
0.50
0.25
0.00
$\phi(x, \mu, \sigma^2)$
-5.0
-2.5
0.0
2.5
5.0
x
Parameter Values
mu= 0, sigma= 3
mu= 0, sigma=.2
mu= 0, sigma=.5
mu=-2, sigma=.5

Named Distributions: χ^2

Definition

Suppose X is a continuous random variable with $X \geq 0$, with PDF

$$f(x) = g(n/2)x^{n/2-1}e^{-x/2}$$

Then we will say X is a χ^2 distribution with n degrees of freedom. Equivalently,

$$X \sim \chi^2(n)$$

- $X = \sum_{i=1}^{N} Z^2$, where $Z \sim N(0, 1)$

Chi-Squares, p-values, histograms

Pearson (1901)

Student's t-Distribution

Definition

Suppose $Z \sim$ *Normal*$(0, 1)$ *and* $U \sim \chi^2(n)$. *Define the random variable* Y *as,*

$$Y = \frac{Z}{\sqrt{\frac{U}{n}}}$$

If Z *and* U *are independent then* $Y \sim t(n)$, *with PDF*

$$f(x) = h(n)\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

We will use the t-distribution extensively for test-statistics

Using the t-Distribution

Suppose we take N iid draws,

$$X \sim \text{Normal}(\mu, \sigma^2)$$

Define our data set $\boldsymbol{x} = (x_1, \ldots, x_N)$
Calculate:

$$\bar{x} = \sum_{i=1}^{N} \frac{x_i}{N}$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{N}}$$

$$t \sim Student's\ t(N-1)$$

Example

10 measurements from a normally distributed x to test whether $\mu = 80$.

Step 1: Calculate the sample mean,

$$\bar{x} = \frac{83 + 93 + 147 + 102 + 104 + 151 + 114 + 62 + 79 + 87}{10} = 102.2$$

Step 2: Calculate the sample standard deviation,

$$s^2 = \frac{(-19.2^2 + -9.2^2 + 44.8^2 - 0.2^2 + 1.8^2 + 48.8^2 + 11.8^2 - 40.2^2 - 23.2^2 - 15.2^2)}{9} = 818.8$$

Step 3: Calculate test statistic,

$$t = \frac{102.2 - H_0}{\frac{\sqrt{818.8}}{\sqrt{10}}}, \quad \frac{102.2 - 80}{\frac{\sqrt{818.8}}{\sqrt{10}}} = 2.4533$$

Step 4: Calculate p-value,

$$2 * (1 - pt(2.4533, 9)) = .0365$$

Joint PDFs

If we want to know the probability of a set of joint events $(x, y) \in A$

$$P((X, Y) \in A) = \int_{y \in A} \int_{x \in A} f_{X,Y}(x, y) dx dy$$

We can also calculate the PDFs of X and Y individually (these are the marginal distributions):

$$f_X(x) = \int_y f_{X,Y}(x, y) dy, \qquad f_Y(y) = \int_x f_{X,Y}(x, y) dx$$

Example of Joint Probability Density: Roof

$$f_{X,Y}(x,y) = x + y, \quad \textit{for } x,\ y\ \in [0,1]$$

the marginal densities $f_X(x)$ and $f_Y(y)$ are

$$f_X(x) = \int_0^1 (x+y)dy = xy + \frac{y^2}{2}\Big|_0^1 = x + \frac{1}{2}$$

$$f_Y(y) = \int_0^1 (x+y)dx = \frac{x^2}{2} + yx\Big|_0^1 = \frac{1}{2} + y$$

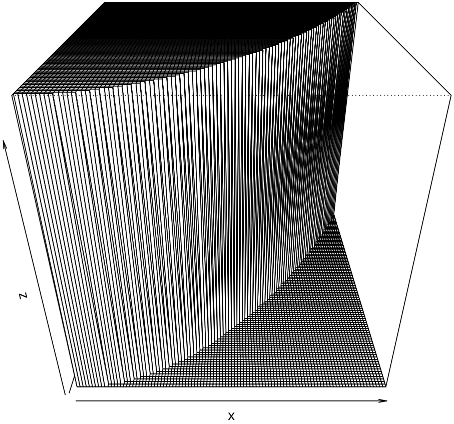
Example of Joint Probability Density: Quarter Circle

$$f_{X,Y}(x,y) = 3/2 \quad \text{for } x^2 \leq y \leq 1 \text{ and } 0 \leq x \leq 1$$

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy = \int_{x^2}^{1} \frac{3}{2}dy \\
&= \frac{3}{2}y|_{x^2}^{1} = \frac{3}{2}(1-x^2) \\
f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx = \int_{0}^{\sqrt{y}} \frac{3}{2}dx \\
&= \frac{3}{2}x|_{0}^{\sqrt{y}} = \frac{3}{2}\sqrt{y}
\end{aligned}$$

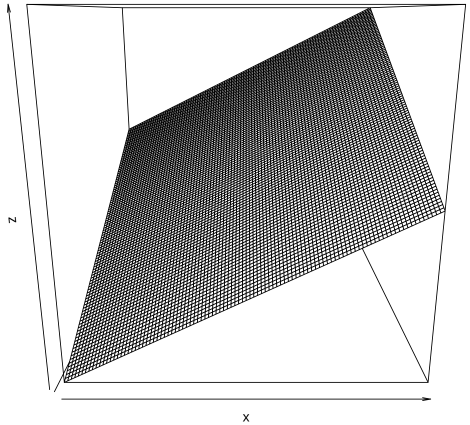
Plotting the Quarter Circle Distribution

```
x <- seq(0, 1, 0.01)
y <- seq(0, 1, 0.01)
z <- outer(x, y, function(x, y){ ifelse(y> x^2, 2/3,0)})
persp(x, y, z)
```


z
x

Plotting the Roof Distribution $f(x, y) = x + y$

```
x <- seq(0, 1, 0.01)
y <- seq(0, 1, 0.01)
z <- outer(x, y, function(x, y){ x + y })
persp(x, y, z)
```


z
x

Covariance

Definition

For jointly continuous random variables X and Y define, the covariance of X and Y as,

$$
\begin{aligned}
Cov(X, Y) &= E[(X - E[X])(Y - E[Y])] \\
&= E\left[XY - E[X]Y - E[Y]X + E[X]E[Y]\right] \\
&= E[XY] - 2E[X]E[Y] + E[E[X]E[Y]] \\
&= E[XY] - E[X]E[Y]
\end{aligned}
$$

Note, $E[XY] = \int_x \int_y xyf(x,\ y)dydx$

Correlation

Definition

Define the correlation of X and Y as,

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Covariance facts

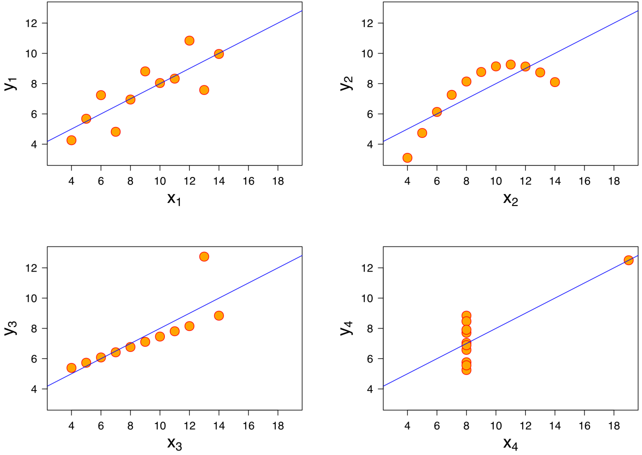
Variance is the covariance of a random variable with itself.

$$
\begin{aligned}
Cov(X,X) &= E[XX] - E[X]E[X] \\
&= E[X^2] - E[X]^2
\end{aligned}
$$

Correlation measures the linear relationship between two random variables

$$
\begin{aligned}
E(XY) &= \sigma_{XY} + \mu_X \mu_Y \\
E(X^2) &= \sigma_X^2 + \mu_X^2
\end{aligned}
$$

Correlation= .816



Correlation and Covariance

Suppose $X = Y$

$$
\begin{aligned}
Cor(X, Y) &= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \\
&= \frac{Var(X)}{Var(X)} \\
&= 1
\end{aligned}
$$

Suppose $X = -Y$

$$
\begin{aligned}
Cor(X, Y) &= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \\
&= \frac{-Var(X)}{Var(X)} \\
&= -1
\end{aligned}
$$

Example covariance X, Y: $f(x, y) = X + Y$

Suppose X and Y have joint probability distribution $x + y$ for $x,\ y \in [0, 1]$.

$$
\begin{aligned}
E[XY] &= \int_0^1 \int_0^1 xy(x+y)dxdy \\
&= \int_0^1 \int_0^1 (x^2y + y^2x)dxdy \\
&= \int_0^1 (\frac{x^3y}{3} + \frac{y^2x^2}{2}|_0^1)dy \\
&= \int_0^1 (\frac{y}{3} + \frac{y^2}{2})dy \\
&= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}
\end{aligned}
$$

$$
\begin{aligned}
E[X] &= \int_0^1 \int_0^1 x(x+y)dxdy \\
&= \int_0^1 \frac{x^3}{3} + y\frac{x^2}{2})|_0^1 dy \\
&= \int_0^1 (\frac{1}{3} + \frac{y}{2})dy \\
&= \frac{y}{3} + \frac{y^2}{4}|_0^1 \\
&= \frac{1}{3} + \frac{1}{4} = \frac{7}{12}
\end{aligned}
$$

Example: $X + Y$

$$
\begin{aligned}
\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\
&= \frac{1}{3} - \frac{7}{12} * \frac{7}{12} \\
&= \frac{1}{3} - \frac{49}{144} = -\frac{1}{144}
\end{aligned}
$$

$$
\begin{aligned}
\text{Cor}(X, Y) &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\
&= \frac{-\frac{1}{144}}{\frac{11}{144}} \\
&= \frac{-1}{11}
\end{aligned}
$$

Conditional Probability Distribution Function

Definition

Suppose X and Y are random variables with joint PDF $f(x, y)$. Then define the conditional probability distribution $f(x|y)$ as

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}$$

$$f(x|y)f_Y(y) = f(x, y)$$

Examples of Conditional Distributions (roof)

- Roof Distribution:

$$f_{X,Y}(x,y) = x + y$$
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{x+y}{\frac{1}{2}+y}$$

- Quarter circle distribution:

$$f_{X,Y}(x,y) = 3/2 \quad \text{for } y^2 \leq x \leq 1 \text{ and } 0 < y < 1$$
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{3/2}{\frac{3}{2}\sqrt{y}}$$

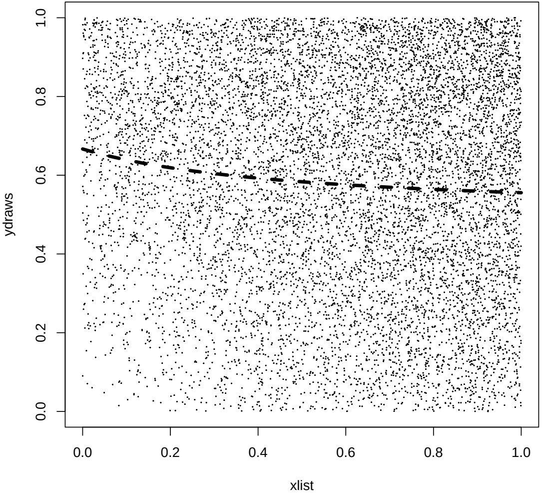
Conditional Expectation

$$\mathbb{E}(Y|X) = \int y \cdot f_{Y|X}(y|x)dy$$

$\mathbb{E}(Y|X)$ is the *best* way to predict Y given X.

Example of Conditional Expectation

$$
\begin{aligned}
E(Y|X) &= \int_0^1 y f_{Y|X}(y|x) dy \\
&= \int_0^1 y \left(\frac{x+y}{\frac{1}{2}+x}\right) dy \\
&= \frac{1}{x+\frac{1}{2}} \int_0^1 y(x+y) dy \\
&= \frac{1}{x+\frac{1}{2}} \left(\frac{y^2 x}{2} + \frac{y^3}{3}\right)\Big|_0^1 \\
&= \frac{2+3x}{3+6x}
\end{aligned}
$$


ydraws
0.0
0.2
0.4
0.6
0.8
1.0
xlist

Height (X) and Age (Z): Covariance

We said the height of a child is distributed $f(x) = N(\mu = 33 + z * 2, s = 1.2)$.
Recall $f(x, z) = f(x|z)f(z)$ and that $f(x) = f(x|z = 6).5 + f(x|z = 7).5$

$$
\begin{aligned}
Cov(x, z) &= E[xz] - E[x]E[z] \quad \textit{from definition} \\
&= \sum \int xzf(x, z)dx - \int xf(x) \sum zf(z) \\
&= (\int x * 6f(x|z = 6)f(z = 6)dx + \int x * 7f(x|z = 7)f(z = 7)dx) - 46 * 6.5 \\
&= (\int x * 6f(x|z = 6).5dx + \int x * 7f(x|z = 7).5dx) - 46 * 6.5 \\
&= 3 * \int xf(x|z = 6)dx + 3.5 * \int xf(x|z = 7)dx) - 46 * 6.5 \\
&= (3 * 45 + 3.5 * 47) - 46 * 6.5 = 299.5 - 299 \\
&= 0.5
\end{aligned}
$$

Height (X) and Age (Z): Covariance

Given

$$f(x) = N(\mu = 33 + z * 2, s = 1.2)$$

$$Var(Z) = p * (1 - p) = .25$$

Then

$$\frac{Cov(x, z)}{var(z)} = .5/.25 = 2$$

$$E(X|Z) = 33 + z * 2$$

Height (X) and Age (Z): Variances

$$\sigma^2 + E(X)^2 = E(X^2)$$

Correlation requires Var(X):

$$f(x) = .5 * N(45,\ 1.2) + .5 * N(47,\ 1.2) \qquad \textit{by assumption.}$$
$$E(x) = \sum p_i E[x_i] \qquad \textit{when we have a mixture of distributions with weights } p$$
$$Var(x) = \sum p_i E[x_i^2] - [\sum p_i E[x_i]]^2$$
$$Var(x) = .5 * E(X_1^2) + .5 * E(X_2^2) - (.5 * E(X_1) + .5 * E(X_1))^2$$
$$Var(x) = .5 * (\sigma_1^2 + E(X_1)^2) + .5 * (\sigma_2^2 + E(X_2)^2) - (.5 * E(X_1) + .5 * E(X_2))^2$$
$$Var(x) = .5 * (1.2^2 + 45^2) + .5 * (1.2^2 + 47^2) - (.5 * 45 + .5 * 47)^2$$
$$Var(x) = 2.44$$

Height (X) and Age (Z): Correlation

The correlation:

$$Cor(x,z) = \frac{Cov(x,z)}{\sqrt{Var(x)Var(z)}}$$

$$Cor(x,z) = \frac{.5}{\sqrt{2.44 * .25}}$$

$$Cor(x,z) = 0.64$$

Note, we saw the regression coefficient: Cov(x,z)/Var(z)=2.

We will show that:

$$r_{x,z} * \frac{s_x}{s_z} = b_1$$

$$0.64 * \frac{\sqrt{2.44}}{\sqrt{0.25}} = 2$$

Independence and Covariance

Proposition

Suppose X and Y are independent. Then

$$Cov(X, Y) = 0$$

Independence and Covariance

Proof.

Suppose X and Y are independent.

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

Calculating $E[XY]$

$$\begin{aligned}
E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\
&= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy \\
&= E[X]E[Y]
\end{aligned}$$

Iterated Expectations (LIE)

Proposition

Suppose X and Y are random variables. Then

$$E[X] = E[E[X|Y]]$$

- Inner Expectation is $E[X|Y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$.
- Outer expectation is over y.
- This is analogous to the law of total probability.

Iterated Expectations

Proof.

$$\begin{aligned}
E[E[X|Y]] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dy dx \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x,y) dy dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= E[X]
\end{aligned}$$

□

LIE Example: Proxy Fighting

- Suppose the US is seeking a local ally, but these come in three equally probable kinds, "tough", "average" and "weak".
- The US would be willing to give \$11,000 per fighter if the group is strong, \$10,000 if the group is average, and \$0 if they are weak.
  - $E[price_{USA}] = E[E[price_{USA}|type]] = \sum_{type} E[price_{USA}|type] * p(type)$
  - $E[price_{USA}] = \frac{1}{3} * (11k) + \frac{1}{3} * (10k) + \frac{1}{3} * (0) = 7,000$
- The strong group will fight if they are paid \$10,000, the average group will fight if they are paid \$6,000, and the weak group would fight for \$50.
  - $E[price_{proxy}] = E[E[price_{proxy}|type]] = \sum_{type} E[price_{proxy}|type] * p(type)$
  - $E[price_{proxy}] = \frac{1}{3} * (10k) + \frac{1}{3} * (6k) + \frac{1}{3} * (50) = 5,350$
- If neither group knows their type, then they will be able to sell their services.

LIE Example: Adverse Selection

- Suppose that the proxy knows their type but the US does not.
- If the US offers its average valuation of 7,000, it is only taken by the average and weak types.
- But, if tough proxies will not offer their services, the US would pay at most:

$$E[price_{USA}] = \frac{1}{2}10,000 + \frac{1}{2}0 = 5,000$$

- So even the average group will not fight.