

Linear Models Lecture 1

Robert Gulotty

University of Chicago

May 29, 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.

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.

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

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

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.

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

Example: Short and Tall

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

$$E_1 = \{h : 72 \leq h \leq 108\}$$

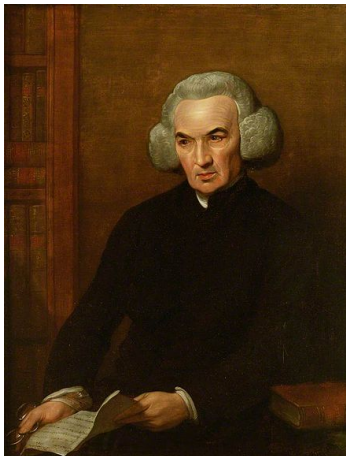
$$E_2 = \{h : 20 \leq h \leq 65\}$$

$$E = E_1 \cup E_2$$

$$P(E) = P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

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

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

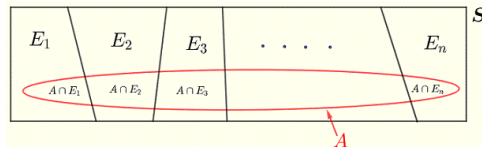
If A is Trump runs again, and B is the event that Trump wins:

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

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

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, $E_1, E_2, E_3 \dots E_n$, their union forms the sample space Ω we say $E_1, E_2, E_3 \dots E_n$ *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:

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

Proof of Law of Total Probability

Proof.

$$P(A) = P(A \cap \Omega)$$

because $A \subseteq \Omega$

$$P(A) = P(A \cap (E_1 \cup E_2 \cup \dots \cup E_n))$$

E partitions Ω

$$P(A) = P((A \cap E_1) \cup (A \cap E_2) \cup \dots \cup (A \cap E_n))$$

Distributive Law

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + \dots + P(A \cap E_n)$$

$E_i \cap E_j = \emptyset$

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

multiplication rule



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

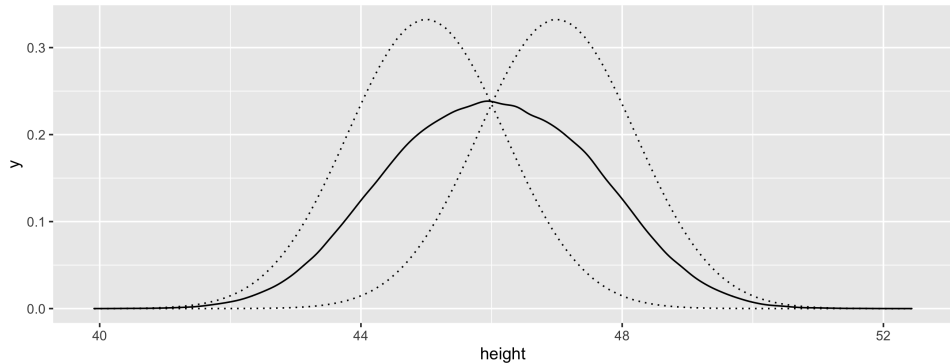
- There is a strong correlation between child height and literacy (a spurious relationship).
- Suppose there is an Amusement Park popular with 1st graders (6-7 year olds).
- Can we rely on the children understanding this sign?

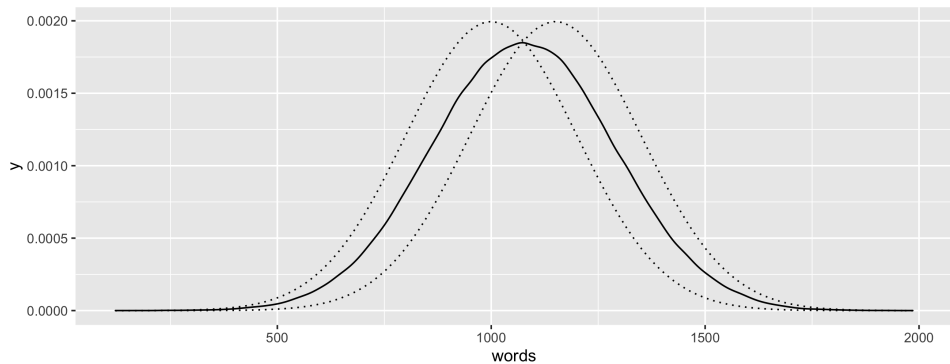


Statistical Model of Height and Literacy

- 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:

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





Using Law of Total Probability

$$A = \{x : x > 48\}$$

$$P(A) = P(A|Z = 6) * P(Z = 6) + P(A|Z = 7) * P(Z = 7) \quad (\text{Law of Total Probability})$$

$$= \left[\int_{48}^{\infty} \phi(x|45, 1.2) dx \right] * P(Z = 6) + \left[\int_{48}^{\infty} \phi(x|47, 1.2) dx \right] * P(Z = 7)$$

$$= (1 - \text{pnorm}(48, 45, 1.2)) * 1/2 + (1 - \text{pnorm}(48, 47, 1.2)) * 1/2$$

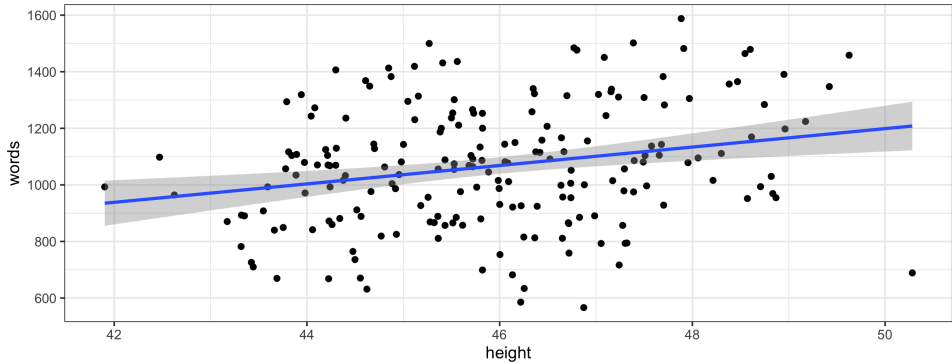
$$= 0.104$$

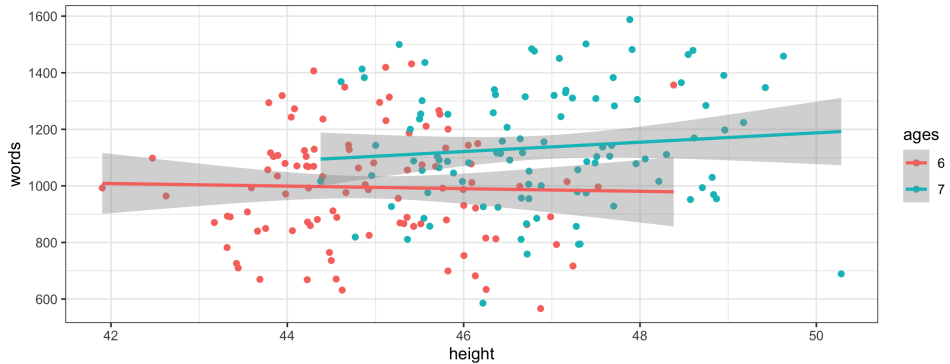
By similar reasoning, $B = \{y : y > 1400\}$, $P(B) = .06$.

Using conditional independence

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

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





	<i>DV: height</i>	
	(1)	(2)
words	0.002*** (0.001)	0.0002 (0.0004)
ages		1.976*** (0.183)
Constant	44*** (0.552)	32.8*** (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 .

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

Connecting random variables to probability

- X assigns some numbers to events.

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

- Remember, probability assigns a chance to an event.

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

- 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



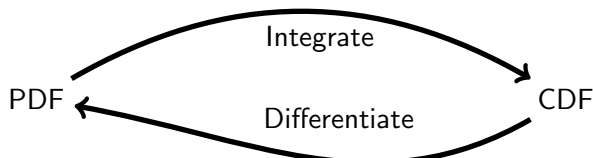
Gauss (1809)

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.

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

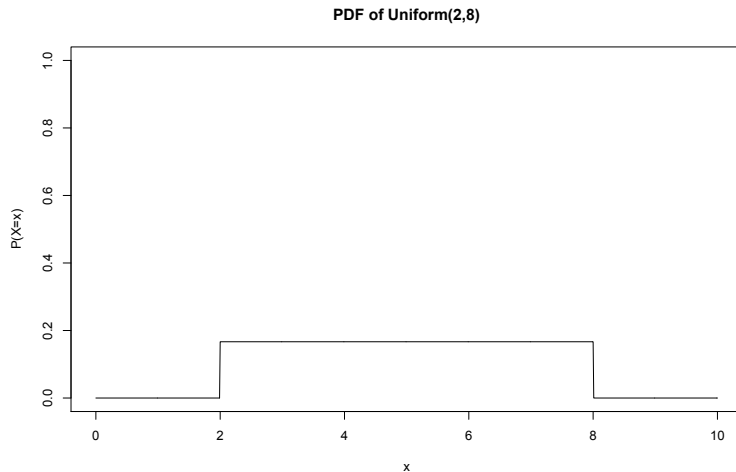


Example: Uniform

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.

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

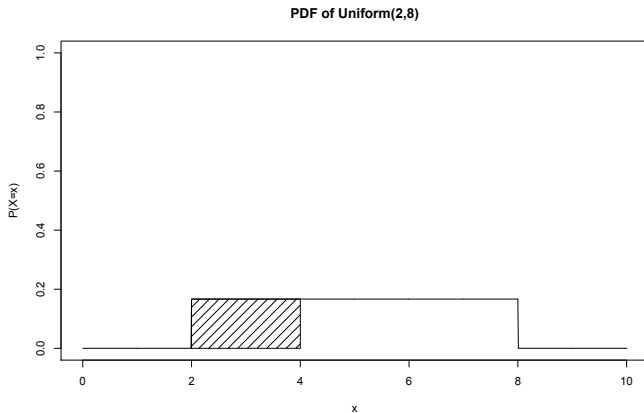
Example: PDF of continuous uniform



Example: continuous uniform

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

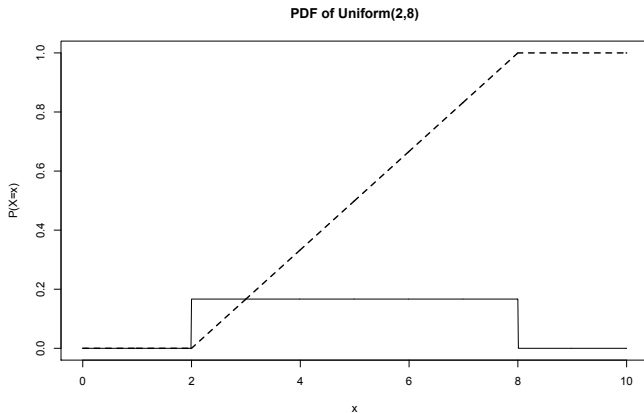
Example: CDF of continuous uniform ($P(X \leq 4)$)



Example: continuous uniform

$$\begin{aligned}
 P(Y \leq y) &= \int_{x=-\infty}^y f(x) dx = \int_{x=-\infty}^y \frac{1}{b-a} dx = \int_{x=a}^y \frac{1}{b-a} dx \\
 &= \frac{y}{b-a} - \frac{a}{b-a} = \frac{y-a}{b-a} \\
 &= \frac{4-2}{8-2}
 \end{aligned}$$

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

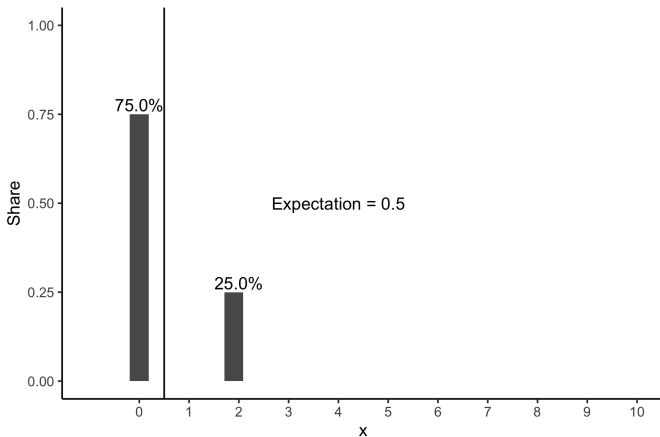
Expected Value: define the expected value of a function X as,

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

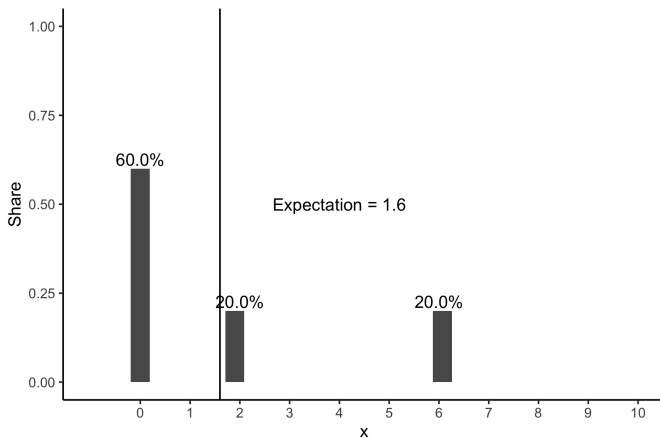
$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \quad \text{when } x \text{ is continuous}$$

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

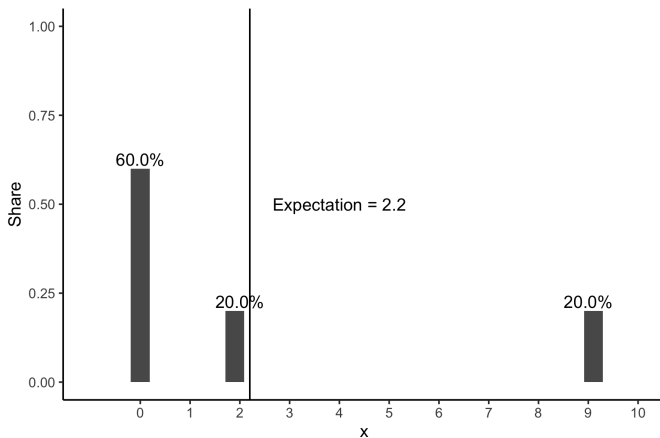
$$E(X) = \sum_{x:p(x)>0} x * p(x) = 0 * .75 + 2 * .25 = \frac{1}{2}$$



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



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



Expectation Properties

$$E[X + Y] = E[X] + E[Y]$$

$$E[a] = a$$

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

$$E[E[X]] = E[X]$$

$$E[XY] \neq E[X] \times E[Y]$$

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

$$\begin{aligned}
 E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
 &= \int_{-\infty}^3 x0dx + \int_3^5 x\frac{1}{5-3}dx + \int_5^{\infty} x0dx \\
 &= 0 + \frac{x^2}{4}\Big|_3^5 + 0 \\
 &= 0 + 5^2/4 - 3^2/4 + 0 \\
 &= 4
 \end{aligned}$$

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

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

Definition of Variance

Definition

The variance of a random variable X , $\text{var}(X)$, is

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

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

Variance Corollary

Corollary

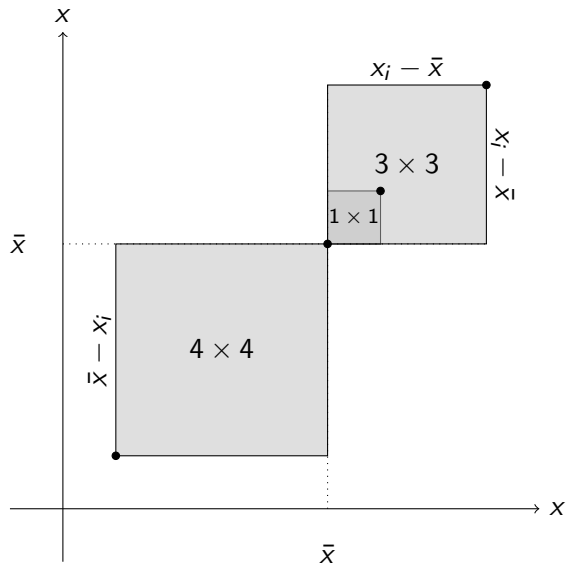
$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

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

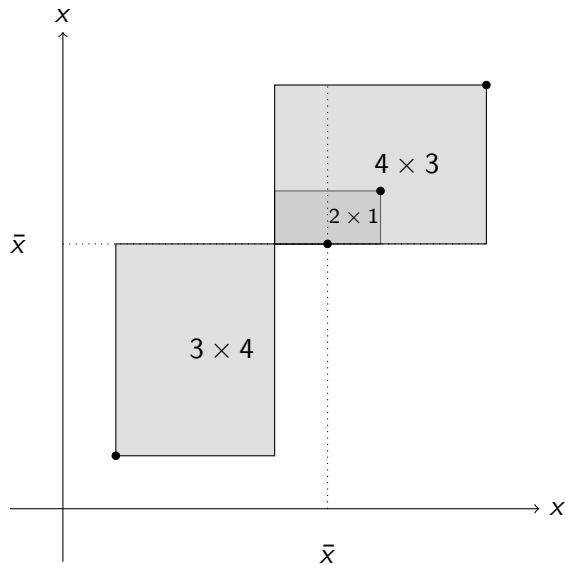
$$\begin{aligned} &= E[((aX + b) - E[aX + b])^2] \\ &= E[((aX + b) - (aE[X] + b))^2] \\ &= E[(aX - aE[X])^2] \\ &= E[(a^2X^2 - 2a^2XE[X] + a^2E[X]^2)] \\ &= a^2E[X^2] - 2a^2E[X]^2 + a^2E[X]^2 \\ &= a^2(E[X^2] - E[X]^2) \\ &= a^2\text{Var}(X) \end{aligned}$$

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

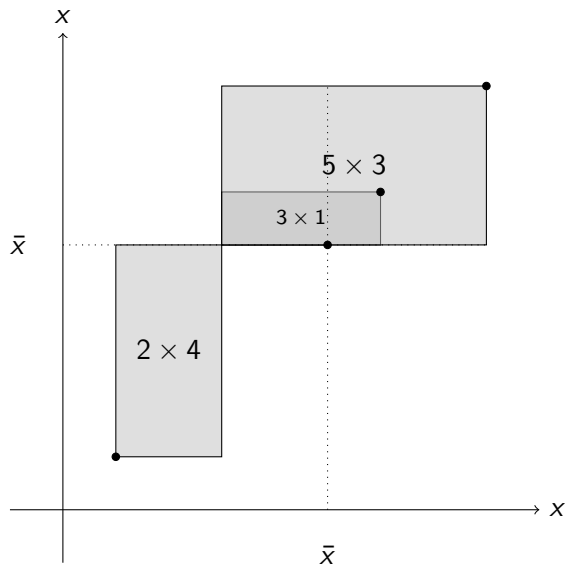
$$\begin{aligned}
 \widehat{Var}(X) &= \frac{1}{N} \sum_i (x_i - \bar{x})(x_i - \bar{x}) \\
 &= \frac{1}{N} \sum_i x_i^2 - \bar{x}x_i - x_i\bar{x} + \bar{x}\bar{x} \\
 &= \frac{1}{N} \sum_i x_i^2 - \bar{x} \frac{1}{N} \sum_i x_i - \bar{x} \frac{1}{N} \sum_i x_i + \frac{1}{N} \sum_i \bar{x}\bar{x} \\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



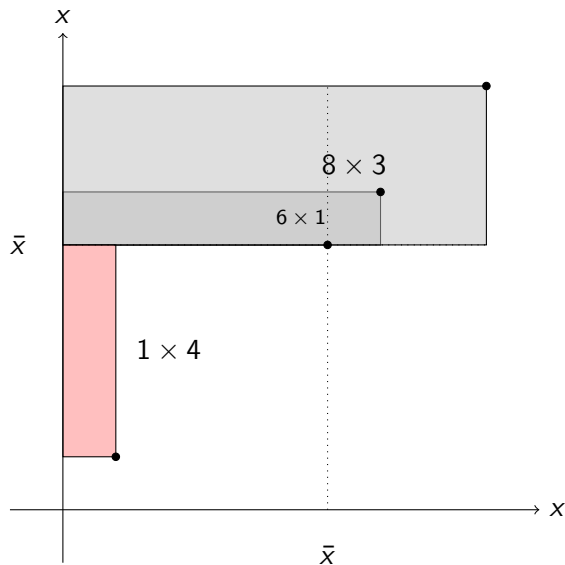
$$\begin{aligned}\widehat{Var}(X) &= \frac{1}{N} \sum_i (x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{(-4 * -4) + (1 * 1) + (3 * 3)}{4} \\ &= \frac{26}{4}\end{aligned}$$



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



$$\begin{aligned}
 \widehat{Var}(X) &= \frac{1}{N} \sum_i (x_i - \bar{x})(x_i - (\bar{x} - 2)) \\
 &= \frac{(-4 * -2) + (1 * 3) + (3 * 5)}{4} \\
 &= \frac{26}{4}
 \end{aligned}$$



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

$$\begin{aligned} \widehat{Var}(X) &= \frac{1}{N} \sum_i (x_i - \bar{x})(x_i) \\ &= \frac{(-4 * 1) + (1 * 6) + (3 * 8)}{4} \\ &= \frac{26}{4} \end{aligned}$$

Example of Variance: Uniform

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

$$\begin{aligned} E[X^2] &= \int_0^1 X^2 \frac{1}{1-0} dx = \frac{X^3}{3} \Big|_0^1 \\ &= \frac{1}{3} \end{aligned}$$

$$E[X]^2 = \left(\frac{1}{2}\right)^2$$

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

Named Distributions: Normal

Definition

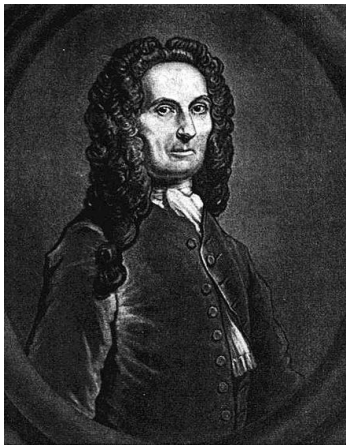
Suppose X is a random variable with $X \in \mathbf{R}$ and probability density function

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

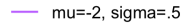
Then X is a **normally** distributed random variable with parameters μ and σ^2 .
Equivalently, we'll write

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

Discoverer of Normal Distribution



De Moivre (1711)



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 $Z \sim \text{Normal}(0, 1)$ 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 $\mathbf{x} = (x_1, \dots, x_N)$

Calculate:

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

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

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

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

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):

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

Example of Joint Probability Density: Roof

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

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

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

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

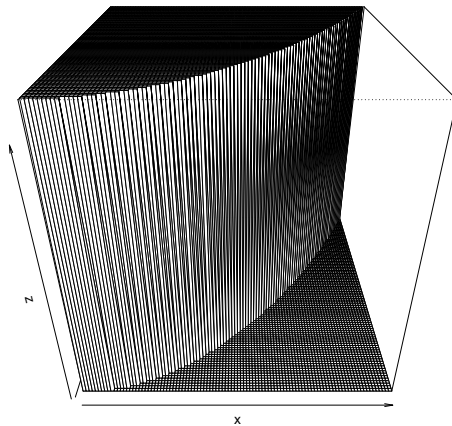
Example of Joint Probability Density: Quarter Circle

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

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y) dy = \int_{x^2}^1 \frac{3}{2} dy \\ &= \frac{3}{2} y \Big|_{x^2}^1 = \frac{3}{2} (1 - x^2) \\ f_Y(y) &= \int_{-\infty}^{\infty} f(x,y) dx = \int_0^{\sqrt{y}} \frac{3}{2} dx \\ &= \frac{3}{2} x \Big|_0^{\sqrt{y}} = \frac{3}{2} \sqrt{y} \end{aligned}$$

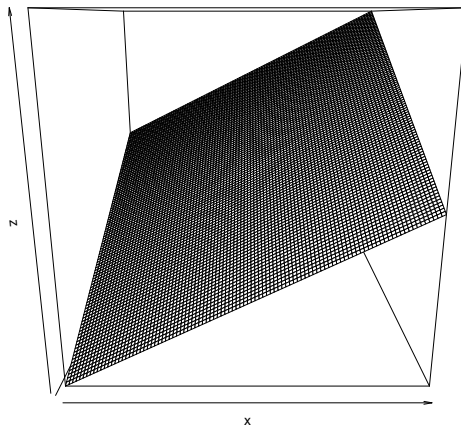
Plotting the Quarter Circle Distribution

```
x <- seq(0, 1, 0.01)
y <- seq(0, 1, 0.01)
z <- outer(x, y, function(x, y){ ifelse(y > x2, 2/3, 0)})
persp(x, y, z)
```



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)
```



Covariance

Definition

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

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

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

Correlation

Definition

Define the correlation of X and Y as,

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

Covariance facts

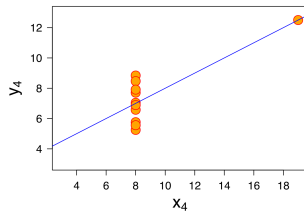
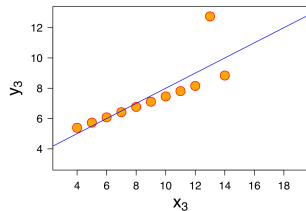
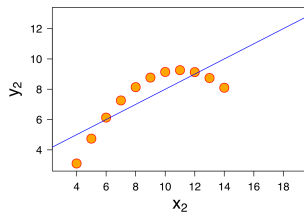
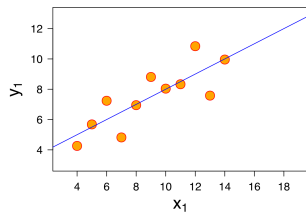
Variance is the covariance of a random variable with itself.

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

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 $X = Y$

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

Suppose $X = -Y$

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

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

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

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

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

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:

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

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{3/2}{\frac{3}{2}\sqrt{y}}$$

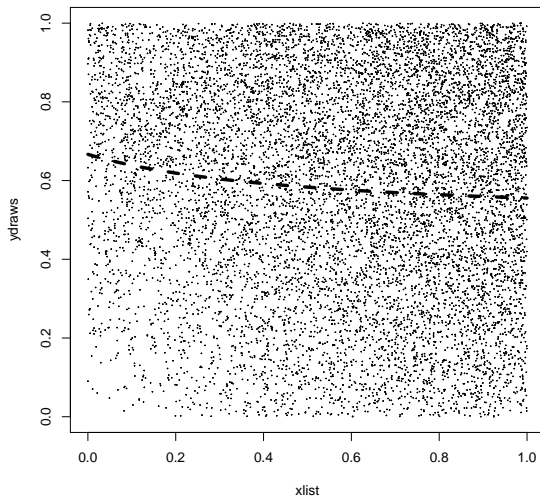
Conditional Expectation

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

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

Example of Conditional Expectation

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



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 $f(x) = f(x|z = 6).5 + f(x|z = 7).5$

$$\text{Cov}(x, z) = E[xz] - E[x]E[z] \quad \text{from definition}$$

$$= \sum \int xzf(x, z)dx - \int xf(x) \sum zf(z)$$

$$= \left(\int x * 6f(x|z = 6)f(z = 6)dx + \int x * 7f(x|z = 7)f(z = 7)dx \right) - 46 * 6.5$$

$$= \left(\int x * 6f(x|z = 6).5dx + \int x * 7f(x|z = 7).5dx \right) - 46 * 6.5$$

$$= 3 * \int xf(x|z = 6)dx + 3.5 * \int xf(x|z = 7)dx - 46 * 6.5$$

$$= (3 * 45 + 3.5 * 47) - 46 * 6.5 = 299.5 - 299$$

$$= 0.5$$

Height (X) and Age (Z): Covariance

Given

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

$$\text{Var}(Z) = p * (1 - p) = .25$$

Then

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

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

Height (X) and Age (Z): Variances

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

Correlation requires $\text{Var}(X)$:

$$f(x) = .5 * N(45, 1.2) + .5 * N(47, 1.2) \quad \text{by assumption.}$$

$$E(x) = \sum p_i E[x_i] \quad \text{when we have a mixture of distributions with weights } p$$

$$\text{Var}(x) = \sum p_i E[x_i^2] - [\sum p_i E[x_i]]^2$$

$$\text{Var}(x) = .5 * E(X_1^2) + .5 * E(X_2^2) - (.5 * E(X_1) + .5 * E(X_1))^2$$

$$\text{Var}(x) = .5 * (\sigma_1^2 + E(X_1)^2) + .5 * (\sigma_2^2 + E(X_2)^2) - (.5 * E(X_1) + .5 * E(X_2))^2$$

$$\text{Var}(x) = .5 * (1.2^2 + 45^2) + .5 * (1.2^2 + 47^2) - (.5 * 45 + .5 * 47)^2$$

$$\text{Var}(x) = 2.44$$

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

$$\text{Cov}(X, Y) = 0$$

Independence and Covariance

Proof.

Suppose X and Y are independent.

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

Calculating $E[XY]$

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

Iterated Expectations (LIE)

Proposition

Suppose X and Y are random variables. Then

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

- Inner Expectation is $E[X|Y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx$.
- 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.

Next Steps

- Read and work through "Estimators1", "Estimators2", "Estimators3.pdf"
- Finish working through "ProofsAdvice.pdf"

Law of Total Variance

Suppose X and Y are random variables. Then

$$\text{var}[X] = E[\text{var}[X|Y]] + \text{var}(E[X|Y])$$

The first term is the average of the variation for each value of Y . The second term is the variation across the averages within each value of Y .

Estimators

- We observe realizations from a vector of random variables $\mathbf{x} = (x_1, x_2 \dots x_n)$.
- When we construct a function $z(\mathbf{x})$ of observed data, they are called 'statistics'.
- If that function is used to infer the value of a parameter of a distribution, it is called an 'estimator'.
- All estimators are random variables with distributions.
 - $\bar{x}(\mathbf{x}) = \frac{\sum x_i}{n}$ is the sample mean and estimates the true mean (μ).
 - $s^2(\mathbf{x}) = \frac{\sum (x_i - \bar{x})^2}{n-1}$ is the sample variance and estimates the true variance (σ^2).
- An estimator $\hat{\theta}$ is evaluated by its expectation and variance.
 - Is $E(\hat{\theta}) - \theta = 0$? If so, it is called unbiased.
 - Is $var(\hat{\theta}) \leq var(\tilde{\theta})$, where $\tilde{\theta}$ is any other unbiased estimator? If so, it is called efficient.
 - How does $\hat{\theta}_n$ behave as n gets arbitrarily large? Next time.

Preview: Sample Mean is unbiased

- If $\mathbf{x} = (x_1, x_2 \dots x_n)$ are a iid random sample drawn from some distribution with mean μ , and we calculate $\bar{x}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N x_i$, we can show that $E[\bar{x}] - \mu = 0$:

$$\begin{aligned}
 E[\bar{x}(\mathbf{x})] &= E\left[\frac{1}{N} \sum_{i=1}^N x_i\right] \\
 &= \frac{1}{N} \left[\sum_{i=1}^N E[x_i] \right] \\
 &= \frac{1}{N} \left[\sum_{i=1}^N \mu \right] \\
 &= \frac{1}{N} [N\mu] \\
 &= \mu
 \end{aligned}$$