Lecture 15: Chapter 7, Section 1
Random Variables

- Definitions, Notation
- Probability Distributions
- Application of Probability Rules
- Mean and s.d. of Random Variables; Rules

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-4)
  - Displaying and Summarizing (Lectures 5-12)
  - Probability
    - Finding Probabilities (discussed in Lectures 13-14)
    - Random Variables
    - Sampling Distributions
  - Statistical Inference

Definition

Random Variable: a quantitative variable whose values are results of a random process

Looking Ahead: *In Inference, we'll want to draw conclusions about population proportion or mean, based on sample proportion or mean. To accomplish this, we will explore how sample proportion or mean behave in repeated samples. If the samples are random, sample proportion or sample mean are* ***random variables.***

Looking Ahead: *Sample proportion and sample mean are very complicated random variables. We start out by looking at much simpler random variables.*

Definitions

- **Discrete Random Variable:** one whose possible values are finite or countably infinite (like the numbers 1, 2, 3, …)
- **Continuous Random Variable:** one whose values constitute an entire (infinite) range of possibilities over an interval

Notation

Random Variables are generally denoted with capital letters such as X, Y, or Z.

The letter Z is often reserved for random variables that follow a standardized normal distribution.

Example: *A Simple Random Variable*

- **Background**: Toss a coin twice, and let the random variable X be the number of tails appearing.
- **Questions:**
  - What are the possible values of X?
  - What kind of random variable is X?
- **Responses:**
  - Possible values:
  - X is a

Practice: 7.1a p.285

Definitions

- **Probability distribution** of a random variable tells all of its possible values along with their associated probabilities.
- **Probability histogram** displays possible values of a random variable along horizontal axis, probabilities along vertical axis.

Definition

- **Probability distribution** of a random variable tells $\boxed{\text{all}}$ of its possible values along with their associated probabilities.

Looking Back: *Last chapter we considered **individual** probabilities like the chance of getting two tails in two coin tosses. Now we take a more global perspective, considering the probabilities of **all** the possible numbers of tails occurring in two coin tosses.*

Median and Mean of Probability Distribution

- **Median** is the middle value, with half of values above and half below (equal area value on histogram).
- **Mean** is average value ("balance point" of histogram)
- **Mean equals Median** for symmetric distributions

Example: *Probability Distribution of a Random Variable*

- **Background**: The random variable X is the number of tails in two tosses of a coin.
- **Questions:**
  - What are the probabilities of the possible outcomes?
  - What is the probability distribution of X?
- **Responses:** Possible outcomes:

Each has probability so the probability distribution is:

X = Number of tails	0	1	2
Probability			

Non-overlapping "Or" Rule→P(X=1) = ________

Practice: 7.1b p.285

Example: *Probability Distribution of a Random Variable*

- **Background**: We have the probability distribution of the random variable *X* for number of tails in two tosses of a coin.

X = Number of tails	0	1	2
Probability	1/4	1/2	1/4

- **Question:** How do we display and summarize *X*?
- **Response:** Use _

Summarize: (center) mean=median=_

(spread) Typical distance from 1 is a bit less than

(shape)

Notation; Permissible Probabilities and Sum-to-One Rule for Probability Distributions

P(*X*=*x*) denotes the probability that the random variable X takes the value x.

Any probability distribution of a discrete random variable X must satisfy:

- $0 \leq P(X = x) \leq 1$ where x is any value of X
- $P(X = x_1) + P(X = x_2) + \cdots + P(X = x_k) = 1$

  where $x_1, x_2, \cdots, x_k$ are all possible values of X

According to this Rule, if a probability histogram has bars of width 1, their total area must be 1.

Interim Table

To construct probability distribution for more complicated random processes, begin with interim table showing all possible outcomes and their probabilities.

Example: *Interim Table and Probability Distribution*

- **Background:** A coin is tossed 3 times and the random variable X is number of tails tossed.
- **Questions:** What are the possible outcomes, values of X, and probabilities? How do we find probability that $X=1$? $X=2$?
- **Response:**
  - Interim Table:
  - Use
    Rule to combine probabilities

Outcome	X=no.of tails	Probability
HHH	0	1/8
HHT	1	1/8
HTH	1	1/8
THH	1	1/8
HTT	2	1/8
THT	2	1/8
TTH	2	1/8
TTT	3	1/8

Practice: 7.5a p.286

Example: *Probability Distribution and Histogram*

- **Background:** X is number of tails in 3 coin tosses.
- **Question:** What are the probability distribution of X and probability histogram?
- **Response:** Use the interim table to determine probabilities.

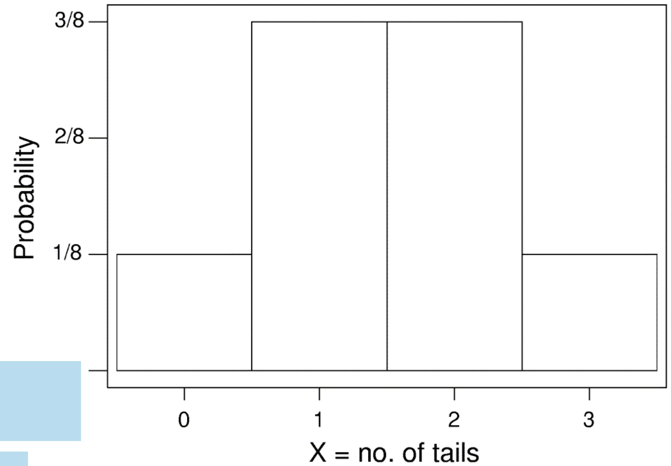
X = Number of tails	0	1	2	3
$P(X = x)$				

Use the probability distribution to sketch the histogram.

Practice: 7.5a-b p.286

Example: *Summaries from Probability Histogram*

- **Background:** Histogram for number of tails in 3 coin tosses.



- **Question:** What does it show?
- **Response:**

  Histogram has
  - Shape:
  - Center: median=mean=
  - Spread: Typical distance from mean a bit less than since 1 and 2 (which are more common) are only 0.5 away from 1.5; 0 and 3 (less common) are 1.5 away from 1.5.

Looking Ahead: *Standard deviation of R.V. to be introduced later on.*

Practice: 7.5c p.286

Definition *(Review)*

- **Probability:** chance of an event occurring, determined as the
  - Proportion of equally likely outcomes comprising the event; or
  - Proportion of outcomes observed in the long run that comprised the event; or
  - Likelihood of occurring, assessed subjectively.

Looking Back: *Principle of* ***equally likely outcomes*** *was used to establish coin-flip probabilities. For other R.V.s, like household size, the distribution has been constructed for us based on* ***long-run observations****.*

Example: *Different Ways to Assess Probabilities*

- **Background**: Census Bureau reported distribution of U.S. household size in 2000.

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** What is the difference between how these probabilities have been assessed, and the way we assessed probabilities for coin-flip examples?
- **Response:** Coin-flip probabilities are based on (two equally likely faces).

  Household probabilities are based on (all households in U.S. in 2000).

Probability Rules *(Review)*

Probabilities must obey

- Permissible Probabilities Rule
- Sum-to-One Rule
- "Not" Rule
- Non-Overlapping "Or" Rule
- Independent "And" Rule
- General "Or" Rule
- General "And" Rule
- Rule of Conditional Probability

Example: *Permissible Probabilities Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** How do these probabilities conform to the Permissible Probabilities Rule?
- **Response:**

Example: *Sum-to-One Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** According to the "Sum-to-One" Rule, what must be true about the probabilities in the distribution?
- **Response:** According to the Rule, we have

Example: *"Not" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** According to the "Not" Rule, what is the probability of a household *not* consisting of just one person?
- **Response:**

Example: *Non-Overlapping "Or" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** According to the Non-overlapping "Or" Rule, what is the probability of having fewer than 3 people?
- **Response:** The probability of having fewer than 3 people is P(X<3)

=

Example: *Independent "And" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** Suppose a polling organization has sampled two households at random. According to the Independent "And" Rule, what is the probability that the first has 3 people and the second has 4 people?
- **Response:** The probability that the first has 3 people and the second has 4 people is

P(X1=3 and X2=4)

=

where we use X1 to denote number in 1st household,
X2 to denote number in 2nd household.

Practice: 7.8e p.287

Example: *General "Or" Rule*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** Suppose a polling organization has sampled two households at random. According to the General "Or" Rule, what is the probability that one or the other has 3 people?
- **Response:** The events overlap: it is possible that both households have 3 people. P(X1=3 or X2=3) =

where we apply the Independent "And" Rule for P(X1=3 and X2=3).

Example: *Rule of Conditional Probability*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** Suppose a polling organization samples only from households with fewer than 3 people. What is the probability that a household with fewer than 3 people has only 1 person?
- **Response:**

  P(X=1 given X<3) =

Mean and Standard Deviation of Random Variable

- **Mean of discrete random variable *X***

$$\mu = x_1 P(X = x_1) + \cdots + x_k P(X = x_k)$$

Mean is **weighted average** of values, where each value is weighted with its probability.

- **Standard deviation of discrete random variable *X***

$$\sigma = \sqrt{(x_1 - \mu)^2 P(X = x_1) + \cdots + (x_k - \mu)^2 P(X = x_k)}$$

Standard deviation is "typical" distance of values from mean. Squared standard deviation is the **variance**.

Looking Back: *Greek letters are used because these are the mean and standard deviation of* ***all*** *the random variables' values.*

Example: *Mean of Random Variable*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** What is the mean household size?
- **Response:** $1(0.26)+2(0.34)+\ldots 7(0.01) =$ is the mean household size.

Looking Back: *Median is 2 (has 0.5 at or below it). Mean is greater than median because distribution is skewed right. Also, mean is less than the "middle" number, 4, because smaller household sizes are weighted with higher probabilities.*

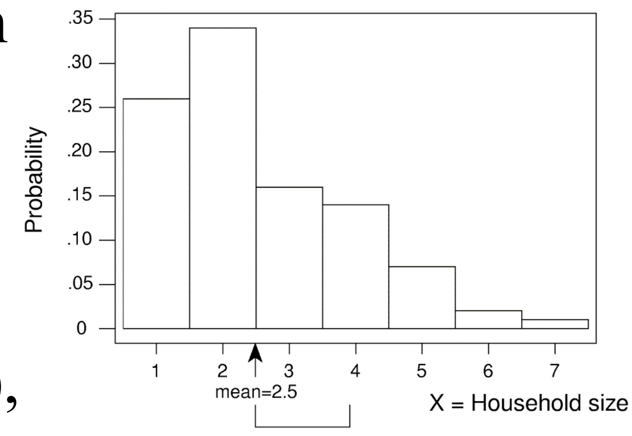
Practice: 7.8g-h p.288

Example: *Standard Deviation of R.V.*

- **Background**: Household size in U.S. has

X	1	2	3	4	5	6	7
$P(X = x)$	0.26	0.34	0.16	0.14	0.07	0.02	0.01

- **Question:** What is the standard deviation of household sizes (typical distance from the mean, 2.5)?
(a) 0.014 (b) 0.14 (c) 1.4 (d) 14.0
- **Response:** The typical distance of household sizes from their mean, 2.5, is the closest are 0.5 away (2 and 3), the farthest is 4.5 away (7). *(Or calculate by hand or with software).*



Standard deviation =

A Closer Look: *Skewed right→ most of the spread arises from values above the mean, not below.*

Practice: 7.15c p.288

Rules for Mean and Variance

- Multiply R.V. by constant→its mean and standard deviation are multiplied by same constant [or its abs. value, since s.d.>0]
- Take sum of two independent R.V.s→
  - mean of sum = sum of means
  - variance of sum = sum of variances

(variance is *squared* standard deviation)

Looking Ahead*: These rules will help us identify mean and standard deviation of sample proportion and sample mean.*

Example: *Mean, Variance, and SD of R.V.*

- **Background**: Number *X* rolled on a die has

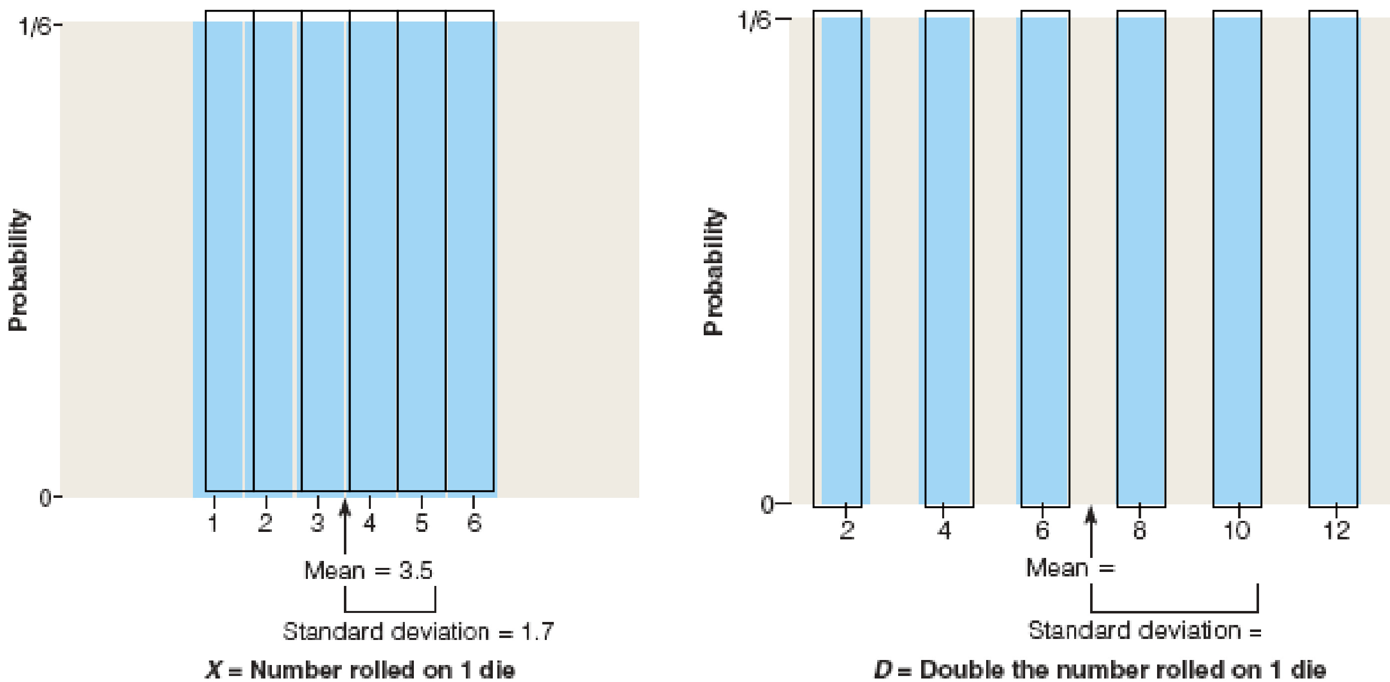
X=no. rolled	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

- **Question:** What are the mean, variance, and standard deviation of *X*?
- **Response:**
  - **Mean:** same as median (because symmetric)
  - **Variance:** (found by hand or with software)
  - **Standard deviation:** (square root of variance)

Practice: 7.12 p.288

Example: *Mean and SD for Multiple of R.V.*

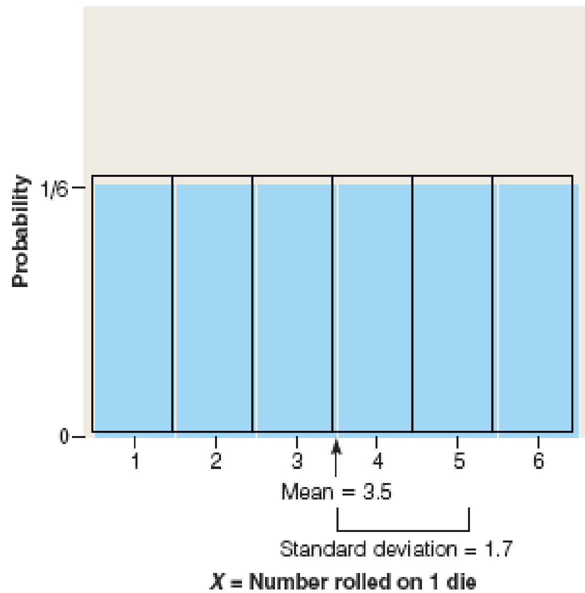
- **Background**: Number *X* rolled on a die has mean 3.5, s.d. 1.7.

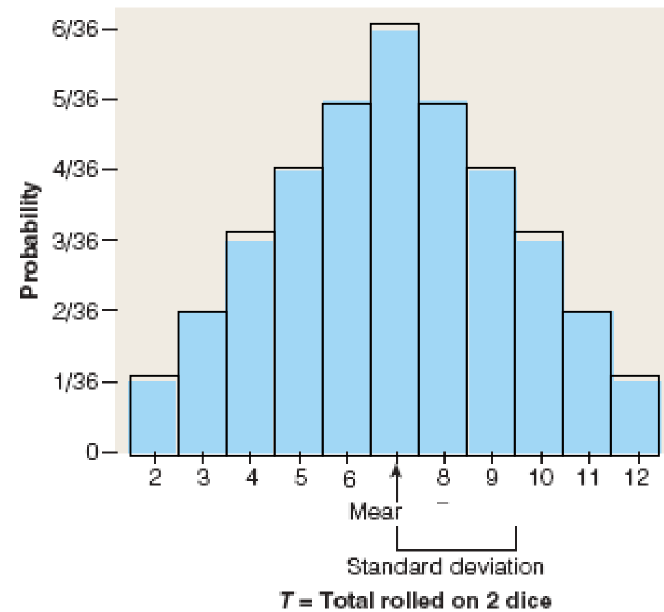


- **Question:** What are mean and s.d. of double the roll?
- **Response:** For double the roll, mean is ____________
  s.d. is

Example: *Mean and SD for Sum of R.V.s*

- **Background**: Numbers $X1$, $X2$ on 2 dice each have mean 3.5, variance 2.92.

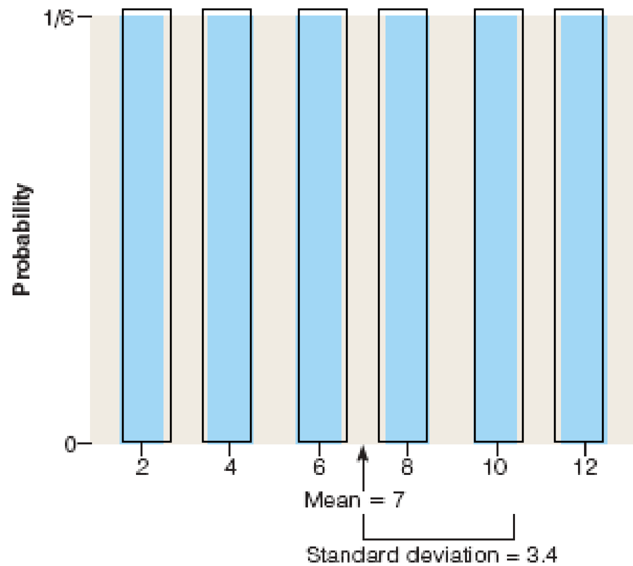




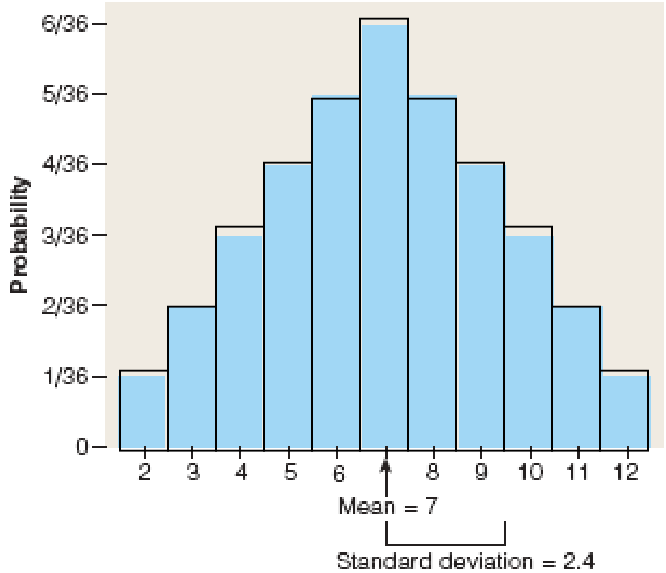
- **Question:** What are mean, variance, and s.d. of total on 2 dice?
- **Response**: Mean ______, variance ______ s.d. ______

Example: *Doubling R.V. or Adding Two R.V.s*

- **Background**: Double roll of a die: mean=7, s.d.= 3.4.
  Total of 2 dice: mean=7, s.d.= 2.4.



D = Double the number rolled on 1 die

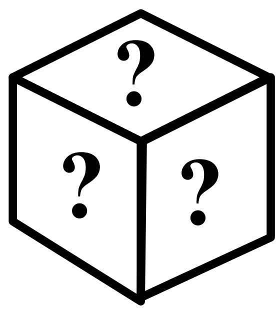


T = Total rolled on 2 dice

- **Question**: Why is double roll more spread than total of 2 dice?
- **Response**: Doubling roll of 1 die makes [2(1)=2 or 2(6)=12] more likely; totaling 2 dice tends to have low and high rolls "cancel each other out".

Example: *Doubling R.V. or Adding Two R.V.s*

- This is the key to the benefits of sampling many individuals: The average of their responses gets us closer to what's true for the larger group.
- If the numbers on a die were unknown, and you had to guess their mean value, would you make a better guess with a single roll or the average of two rolls?



Lecture Summary
(Random Variables)

- Random variables
  - Discrete vs. continuous
  - Notation
- Probability distributions: displaying, summarizing
- Probability rules applied to random variables
- Constructing distribution table
- Mean and standard deviation of random variable
- Rules for mean and variance