Random Variables and Discrete Distributions

Lecture 8

Reading: Sections 9.1 – 9.4, 9.6

(Optional: 9.5)

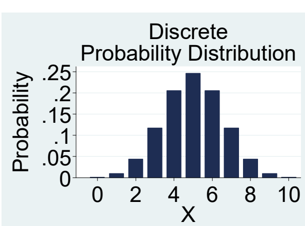
1

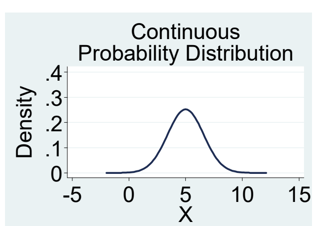
Random Variables

- Random variable: Assigns a number to each outcome of an experiment
  - Often denoted with a capital letter: W, X, Y, Z
  - Discrete random variable: Takes a countable (finite) number of values
    - # of cats in a randomly selected household
  - Continuous random variable: Takes an uncountable (infinite) number of values
    - mean # of cat for random sample of households

2

Distributions of Random Variables





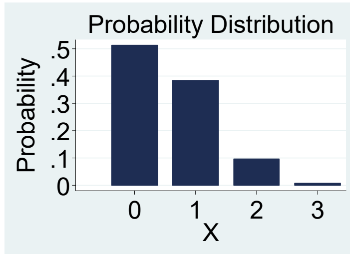
- Discrete distribution gives probability: height of bar
- Continuous distribution gives density: area under density function f(x) is probability
- For sample statistics, which are random variables, distributions can tell us how much sampling noise

3

Summer Job: Sales Call Example

- X: # of sales per day

x	p(x)
0	0.512
1	0.384
2	0.096
3	0.008



Compensation plans:
(A) \$12/day,
(B) \$20/sale or
(C) \$6/day + \$10/sale

Which two things inform the choice amongst these three plans?

4

Expected Value (μ) and Variance (σ^2)

- $\mu = E[X] = \sum xp(x)$
- $\sigma^2 = V[X] = E[(X-\mu)^2] = \sum(x-\mu)^2 * p(x)$

x	p(x)
0	0.512
1	0.384
2	0.096
3	0.008

- $E[X] = 0 * 0.512 + 1 * 0.384 + 2 * 0.096 + 3 * 0.008 = 0.6$
  - Interpretation of 0.6?
- $V[X] = (0-0.6)^2 * 0.512 + (1-0.6)^2 * 0.384 + (2-0.6)^2 * 0.096 + (3-0.6)^2 * 0.008 = 0.48$
- $SD[X] = \sqrt{0.48} = 0.69$

Why is $\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{4} x_i}{4} = \frac{0+1+2+3}{4} = 1.5$ wrong?

5

Laws of Expected Value & Variance

- E[c] = c
- E[X + c] = E[X] + c
- E[c*X] = c*E[X]
- Ex: Y compensation/day
  - **(A)**: E[Y] = E[12] = \$12
  - **(B)**: E[Y] = E[20*X] = 20*0.6 = \$12
  - **(C)**: E[Y] = E[10*X + 6] = \$12

- V[c] = 0
- V[X + c] = V[X]
- V[c*X] = c^2*V[X]
  - **(A)**: V[Y] = V[12] = 0
  - **(B)**: V[Y] = V[20*X] = 400*0.48 = 192; SD[Y] = \$13.86
  - **(C)**: V[Y] = V[10*X + 6] = 100*0.48 = 48; SD[Y] = \$6.93

These reflect the basic laws of summation (discrete) and integration (continuous) and apply to statistics and parameters 6

Linear Combinations of Random Variables

- $E[a + bX_1 + cX_2] = a + bE[X_1] + cE[X_2]$
  - Generalizes for any number of variables
- $V[a + bX_1 + cX_2] = b^2V[X_1] + c^2V[X_2] + 2bcCOV[X_1, X_2]$
  - Recall: $COV[X_1, X_2] = SD[X_1]SD[X_2]\rho$
- Non-linear combinations: expectation does NOT distribute

7

1 Day versus 30 Days: Assuming Performance Independent

	1 Day Total			**30 Day Total**		
	E[Y]	V[Y]	SD[Y]	E[W]	V[W]	SD[W]
(A)	12	0	0	30*12 = 360	0	0
(B)	12	192	13.86	30*12 = 360	30*192 = 5760	75.89
(C)	12	48	6.93	30*12 = 360	30*48 = 1440	37.95

$W = Y_1 + Y_2 + Y_3 + \cdots + Y_{30}$.

$(Y_1 + Y_2 + Y_3 + \cdots + Y_{30})$ is *very different* from $30 * Y$.

Why/how?

8

```
correlate fem_emp_2006 fem_emp_2012 male_emp_2006 male_emp_2012, format;
(obs=34)
             | fem~2006 fem~2012 mal~2006 mal~2012
-------------+------------------------------------
fem_emp_2006 |   1.0000
fem_emp_2012 |   0.9482   1.0000
male_em~2006 |   0.5904   0.4948   1.0000
male_em~2012 |   0.4758   0.6106   0.6498   1.0000

summarize fem_emp_2006 fem_emp_2012 male_emp_2006 male_emp_2012, format;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
fem_emp_2006 |        34       58.92       11.45      22.75      80.88
fem_emp_2012 |        34       60.05       10.43      28.73      77.88
male_em~2006 |        34       74.43        6.31      60.88      88.05
male_em~2012 |        34       72.23        6.34      60.20      85.18
```

OECD data

What is mean and s.d. of the *change* in female ER from 2006 to 2012?

$Mean[F_{12} - F_{06}] = Mn[F_{12}] - Mn[F_{06}] = 60.05 - 58.92 = 1.13$

$V[F_{12} - F_{06}] = V[F_{12}] + V[F_{06}] - 2rSD[F_{12}]SD[F_{06}]$
$= 10.43^2 + 11.45^2 - 2 * 0.9482 * 10.43 * 11.45 = 13.41$

$SD[F_{12} - F_{06}] = \sqrt{13.41} = 3.66$

9

Discrete Probability Distributions

- Some discrete probability distributions:
  - We'll study: Bernoulli, Binomial
  - We'll not study: Poisson, Multinomial, Geometric, Hypergeometric
    - But, using probabilities rules / a tree you should be able to find the probabilities of the geometric distribution but don't worry about its μ and σ^2
- Which appropriate depends on experiment
  - Underlying assumptions must be met

10

Bernoulli

- Bernoulli trial has 2 possible outcomes: "success" "failure"
  - Simplest exp. possible
  - P(success): p
  - P(failure): 1 – p
- Bernoulli random variable: equals 1 if outcome is "success", 0 if "failure"

x	p(x)
0	1 – p
1	p

$E[X] = 0*(1 - p) + 1*(p) = p$

$E[X] = \mu = p$

$V[X] = (1 - p)*(0 - p)^2 + p*(1 - p)^2 = p*(1 - p)$

$V[X] = \sigma^2 = p*(1 - p)$

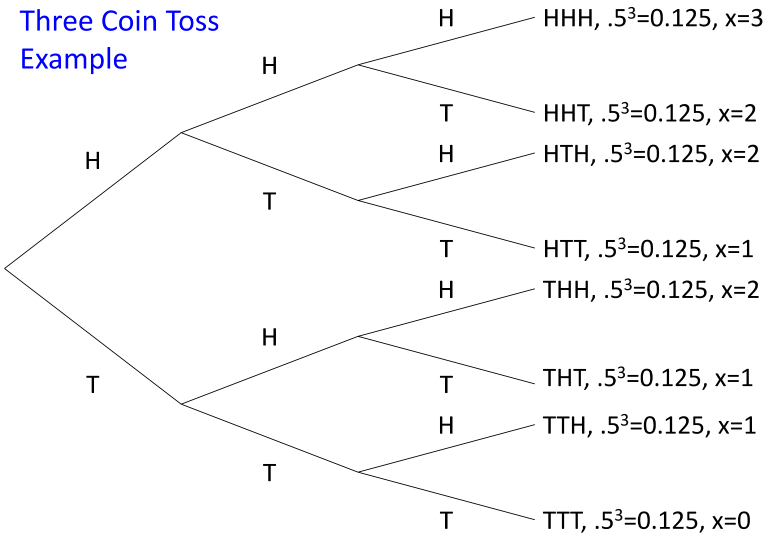
11

Binomial

- Binomial experiment:
  - Fixed # of trials: n
  - Each trial, only two possible outcomes
  - P(success): p
  - P(failure): 1 – p
  - Trials are: independent; Outcome of a trial does not affect outcome in another trial: p constant
- Binomial R.V.: Number of successes in a Binomial experiment
  - Possible values of X?
  - Binomial dist. gives prob. of each value of X
  - # of randomly sampled people who support Mayor John Tory?
  - # students raising hand?

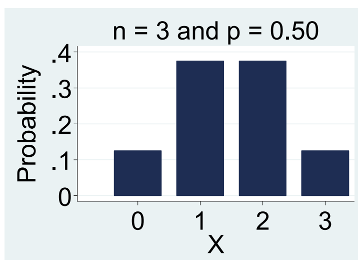
12

Three Coin Toss Example

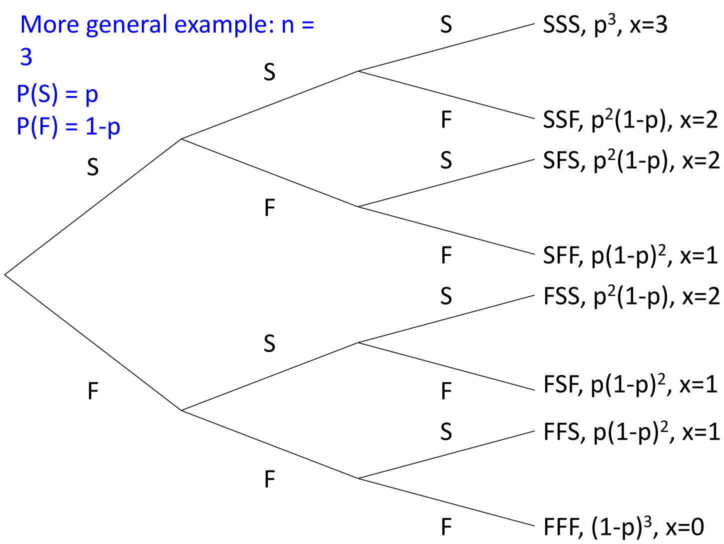


Probability Distribution

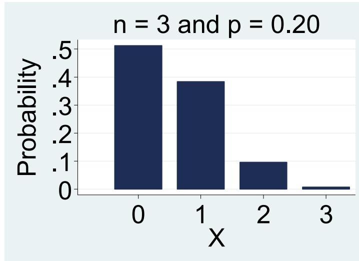
x	p(x)
0	0.125
1	0.375
2	0.375
3	0.125
Total	1

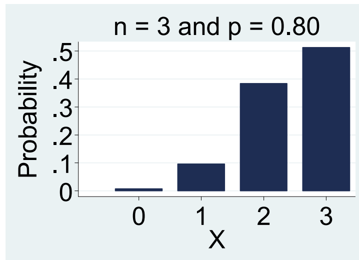


More general example: n = 3
P(S) = p
P(F) = 1-p



x	p(x)
0	$(1 - p)^3$
1	$3*p(1 - p)^2$
2	$3*p^2(1 - p)$
3	p^3
Total	1





Generalizing (when tree unwieldy)

- $P(specific\ sequence\ of\ x\ successes) = p^x(1-p)^{n-x}$ (i.e. tip of one branch)
- How many sequences with x? $C_x^n = \frac{n!}{x!(n-x)!}$
  - "n factorial" = n! = n*(n-1)*(n-2)*…*3*2*1
    - Recall: 1! = 1 and 0! = 1
- Binomial Prob.: $P(X = x) = C_x^n p^x (1-p)^{n-x}$ where $x = 0, 1, \ldots, n$
  - Two parameters: n and p

Table 1. Decile Income Transition Matrix for the 2007-to-2012 panel of Taxfilers

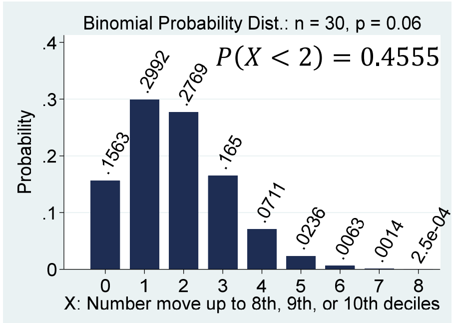
2007 decile	2012 decile: 1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	Total mobility statistics: Immo-bility	Upward	Down-ward
	Percentage										Percentage		
1st	39.7	22.9	11.2	7.6	5.4	4.0	3.1	2.5	2.0	1.5	39.7	60.3	0.0
2nd	13.5	39.4	18.5	10.0	6.2	4.4	3.1	2.2	1.6	1.0	39.4	47.0	13.5
3rd	6.4	14.9	36.3	16.9	9.7	6.1	4.1	2.8	1.8	1.0	36.3	42.4	21.4
4th	4.5	7.2	17.5	27.6	17.5	10.7	6.8	4.3	2.6	1.3	27.6	43.2	29.2
5th	3.1	4.4	8.2	17.0	25.6	17.6	11.3	6.9	4.0	1.8	25.6	41.6	32.7
6th	2.3	3.0	5.1	9.0	16.9	24.3	18.3	11.7	6.5	2.7	24.3	39.3	36.4
7th	1.8	2.1	3.4	5.9	9.5	16.9	24.3	19.6	11.8	4.6	24.3	36.0	39.7
8th	1.4	1.6	2.3	4.0	6.4	9.9	17.2	26.3	22.0	8.8	26.3	30.9	42.8
9th	1.2	1.2	1.6	2.7	4.1	6.4	10.1	18.1	32.5	22.1	32.5	22.1	45.4
10th	1.2	0.9	1.1	1.6	2.4	3.4	5.3	8.6	18.2	57.4	57.4	0.0	42.6

Source: Statistics Canada, Longitudinal Administrative Databank 2007 and 2012, authors' calculations.

Source: Statistics Canada (2016) "The evolution of income mobility in Canada"
https://www150.statcan.gc.ca/n1/pub/75f0002m/75f0002m2016001-eng.htm

Example: 1st decile up to 8th, 9th or 10th

- There's a 6% chance a taxfiler in the 1st decile in 2007 jumps up to the 8th, 9th, or 10th decile in 2012
- In a random sample of 30 taxfilers in 1st decile in 2007, what is the chance less than 2 jump to 8th, 9th or 10th?



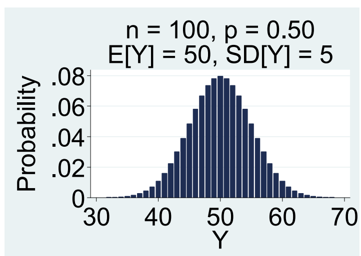
$$P(X = x) = C_x^n p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$P(X = 0) = \frac{30!}{0!(30-0)!} 0.06^0 (1-0.06)^{30-0} = 0.1563$$

$$P(X = 1) = \frac{30!}{1!(30-1)!} 0.06^1 (1-0.06)^{30-1} = 0.2992$$

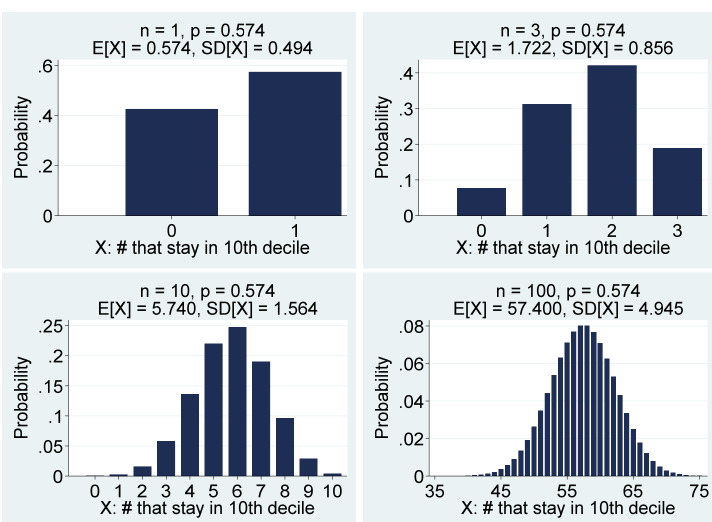
Binomial: Mean and Variance

- Binomial is sum of Bernoulli's: $Y = \sum_{i=1}^{n} X_i$
- $E[Y] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np$
- $V[Y] = V[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} V[X_i] = \sum_{i=1}^{n} p(1-p) = np(1-p)$
- $SD[Y] = \sqrt{np(1-p)}$



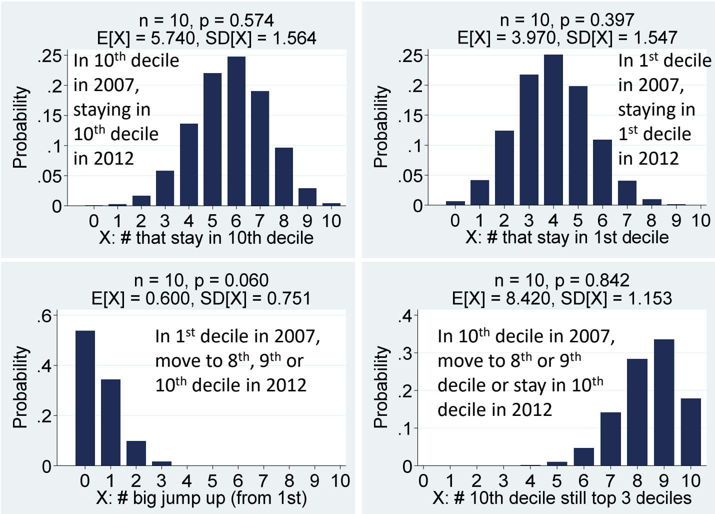
Is getting 75 heads *statistically plausible*?

Does the Empirical Rule apply here, approximately?



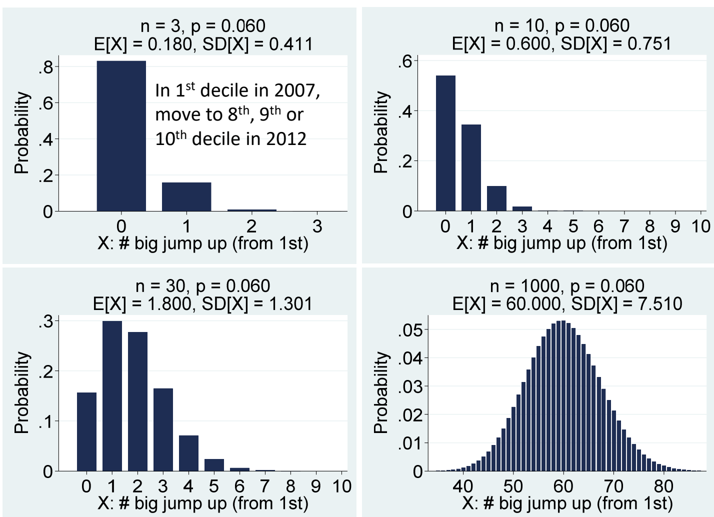
Example: 10th decile in 2007, still in 10th decile in 2012

Meaning of p = 0.574? How does n parameter affect shape?



How does p parameter affect shape?

22



How does n parameter affect shape?

23

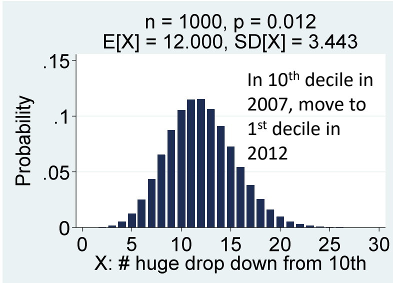
Binomial Distribution Summary

- $X \sim B(n, p)$: two-parameter distribution: n & p
- Shape of distribution depends on parameters
  - positively skewed, negatively skewed, symmetric, or approximately Normal (Bell shaped)
- Probabilities for $x = 0,1,2,\ldots,n$ given by $P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$
- $E[X] = np$
- $V[X] = np(1-p)$ and $SD[X] = \sqrt{np(1-p)}$

24

Finding Binomial Probabilities

Given distribution to right, in a random sample of 1,000 taxfilers in the 10th decile of the income distribution in 2007, what is the chance that 30 or more drop down to the 1st decile of the income distribution in 2012?
$P(X \geq 30) = ?$



Find $P(X = 5)$ w/ basic calculator?

$$P(X = 5) = \frac{1000!}{5!\,(1000-5)!} 0.012^5 (1-0.012)^{1000-5} = 0.01246$$

25

Cumulative Probabilities

- Cumulative probability: $P(X \leq x)$
  - Ex: 10 coin tosses and X counts heads:
    $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.055$
    $P(X > 2) = 1 - P(X \leq 2) = 1 - 0.055 = 0.945$
  - Cumulative probabilities usually most relevant for statistical inference (also in many statistical tables)
    - E.g. 1.2% of all taxfilers in 10th decile in 2007 are in the 1st decile in 2012. In a sample of 1,000, if 20 drop that far (2%), ask "What's the chance *so many* dropped?"
      $P(X \geq 20) = 1 - P(X \leq 19) = 0.0206$

26