Ch5: Discrete Distributions

- HW2 due 10pm
- Download and open: 05-Discrete.xls

22 Sep 2011
BUSI275
Dr. Sean Ho

Trinity Western University

Outline for today

- Example: conditional probabilities
- Discrete probability distributions
  - Finding μ and σ
- Binomial experiments
  - Calculating the binomial probability
  - Excel: BINOMDIST()
  - Finding μ and σ
- Poisson distribution: POISSON()
- Hypergeometric distribution: HYPGEOMDIST()

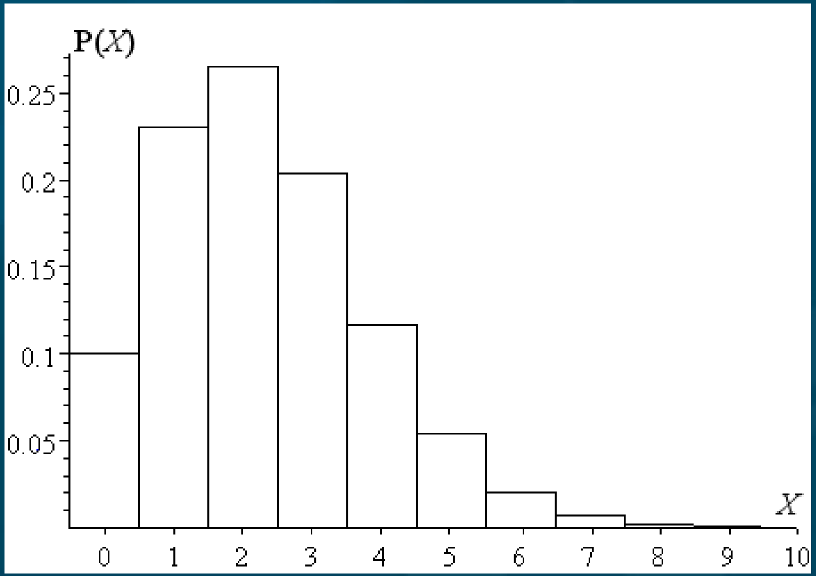
Example: conditional prob.

	Age <18	18-25	>25	Total
Cell phone	40	80	60	180
No cell	40	10	20	70
Total	80	90	80	250

- What fraction are aged 18-25?
- What is the overall rate of cellphone ownership?
- What fraction are minors with cellphone?
- Amongst adults over 25, what is the rate of cellphone ownership?
- Are cellphone ownership and age independent in this study?

Discrete probability distribs

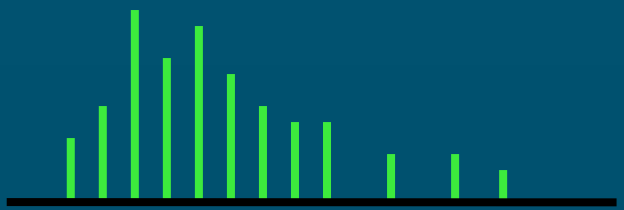
- A random variable takes on numeric values
  - Discrete if the possible values can be counted, e.g., {0, 1, 2, …} or {0.5, 1, 1.5}
  - Continuous if precision is limited only by our instruments
- Discrete probability distribution: for each possible value X, list its probability P(X)
  - Frequency table, or
  - Histogram
- Probabilities must add to 1
  - Also, all $P(X) \geq 0$



Probability distributions

Ch. 5

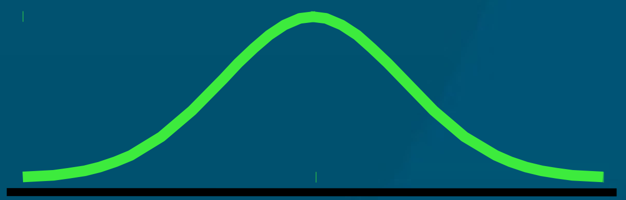
Discrete Random Variable



- Binomial
- Poisson
- Hypergeometric

Continuous Random Variable

Ch. 6



- Normal
- Uniform
- Exponential

Mean and SD of discrete distr.

- Given a discrete probability distribution P(X),
- Calculate mean as weighted average of values:

$$\mu = \sum_X X\, P(X)$$

- E.g., # of email addresses: 0% have 0 addrs; 30% have 1; 40% have 2; 3:20%; P(4)=10%
  - μ = 1*.30 + 2*.40 + 3*.20 + 4*.10 = 2.1
- Standard deviation:

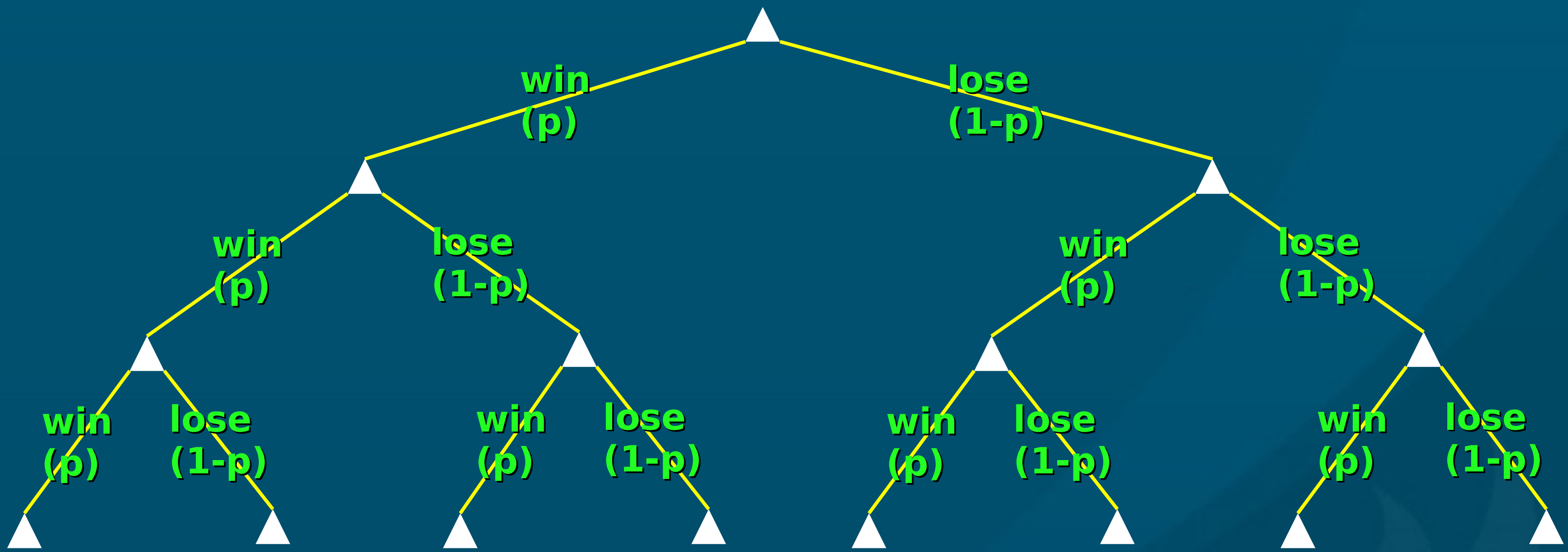
$$\sigma = \sqrt{\sum_X (X - \mu)^2\, P(X)}$$

Binomial variable

- A binomial experiment is one where:
  - Each trial can only have two outcomes: {"success", "failure"}
  - Success occurs with probability p
    - Probability of failure is q = 1-p
  - The experiment consists of many (n) of these trials, all identical
  - The variable x counts how many successes
- Parameters that define the binomial are (n,p)
- e.g., 60% of customers would buy again: out of 10 randomly chosen customers, what is the chance that 8 would buy again?
  - n=10, p=.60, question is asking for P(8)

Binomial event tree

- To find binomial prob. P(x), look at event tree:



- x successes means n-x failures
- Find all the outcomes with x wins, n-x losses:
  - Each has same probability: $p^x(1-p)^{(n-x)}$
  - How many combinations?

Binomial probability

- Thus the probability of seeing exactly x successes in a binomial experiment with n trials and a probability of success of p is:

$$P(x)=\binom{n}{x}(p)^x(1-p)^{(n-x)}, \quad where \quad \binom{n}{x}=C_x^n=\frac{n!}{x!(n-x)!}$$

- "n choose x" is the number of combinations
  - n! ("n factorial") is (n)(n-1)(n-2)...(3)(2)(1), the number of permutations of n objects
  - x! is because the ordering within the wins doesn't matter
  - (n-x)! is same for ordering of losses

Pascal's Triangle

- Handy way to calculate # combinations, for small n:

1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1

x (starts from 0) →

n ↓

- In Excel: COMBIN(n, x)
  - COMBIN(6, 3) → 20

Excel: BINOMDIST()

- Excel can calculate P(x) for a binomial:
  - BINOMDIST(x, n, p, cum)
- e.g., 60% of customers would buy again:
  out of 10 randomly chosen customers,
  what is the chance that 8 would buy again?
  - BINOMDIST(8, 10, .60, 0) → 12.09%
- Set cum=1 for cumulative probability:
  - Chance that at most 8 (≤8) would buy again?
    - BINOMDIST(8, 10, .60, 1) → 95.36%
  - Chance that at least 8 (≥8) would buy again?
    - 1 – BINOMDIST(7, 10, .60, 1) → 16.73%

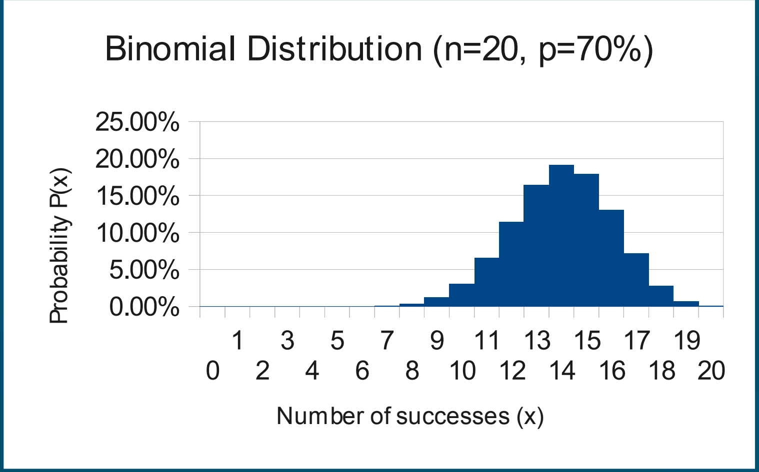
μ and σ of a binomial

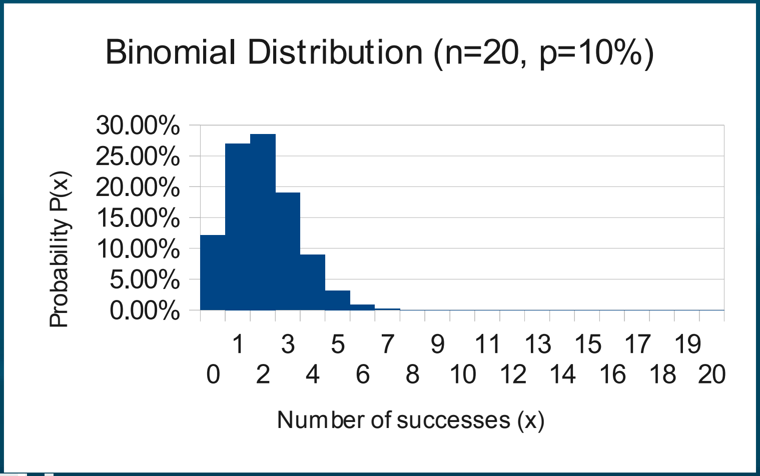
- n: number of trials

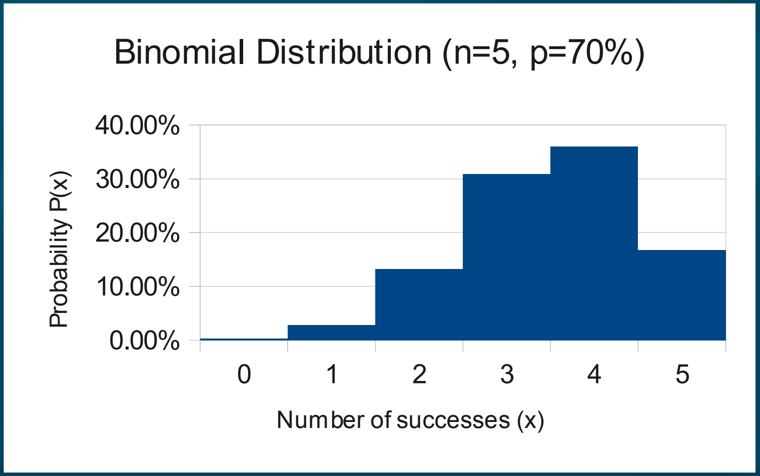
  p: probability of success
- Mean: expected # of successes: $\mu = np$
- Standard deviation: $\sigma = \sqrt{(npq)}$
- e.g., with a repeat business rate of $p=60\%$, then out of $n=10$ customers, on average we would expect $\mu=6$ customers to return, with a standard deviation of $\sigma=\sqrt{(10(.60)(.40))} \approx 1.55$.

Binomial and normal

- When n is not too small and p is in the middle, the binomial approximates the normal:







Poisson distribution

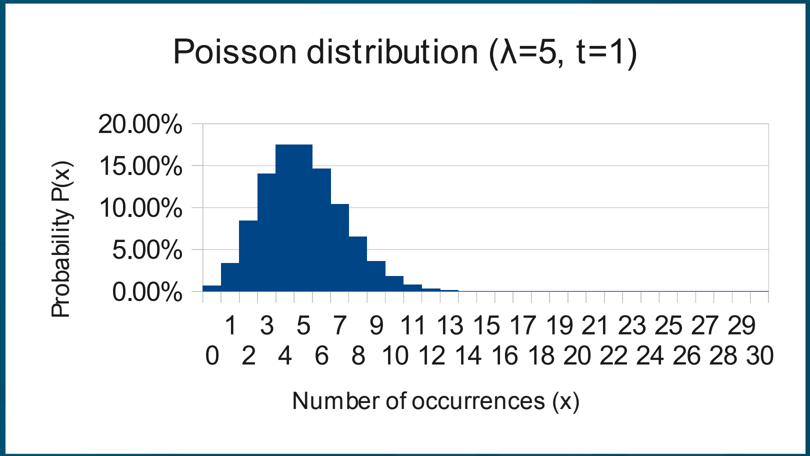
- Counting how many occurrences of an event happen within a fixed time period:
  - e.g., customers arriving at store within 1hr
  - e.g., earthquakes per year
- Parameters: λ = expected # occur. per period
  t = # of periods in our experiment
  - P(x) = probability of seeing exactly x occurrences of the event in our experiment

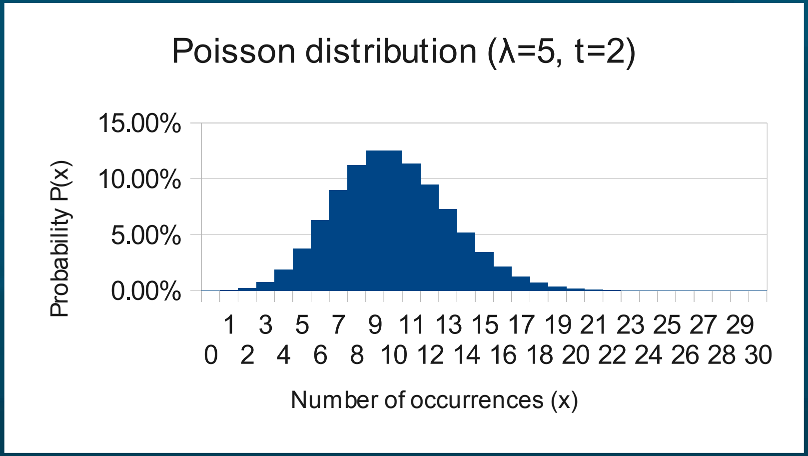
$$P(x)=\frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

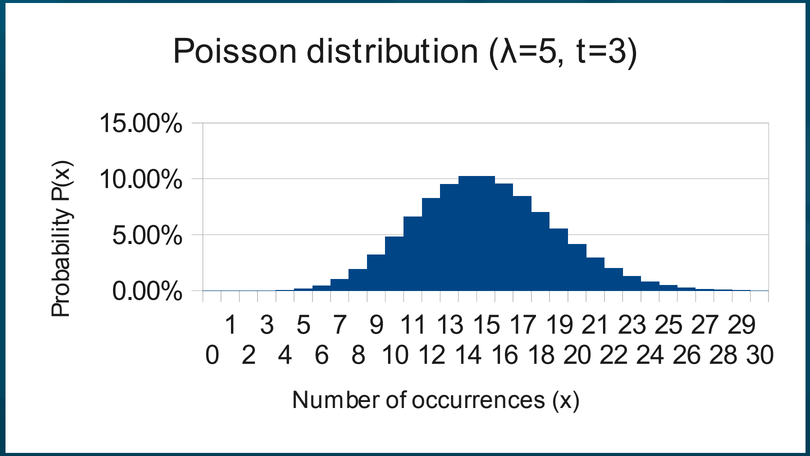
- Mean = λt, and SD = √(λt)

Excel: POISSON()

- POISSON(x, λ*t, cum)
  - Need to multiply λ and t for second param
  - cum=0 or 1 as with BINOMDIST()
- Think of Poisson as the "limiting case" of the binomial as n→∞ and p→0







Hypergeometric distribution

- n trials taken from a finite population of size N
- Trials are drawn without replacement:
  the trials are not independent of each other
  - Probabilities change with each trial
- Given that there are X successes in the larger population of size N, what is the chance of finding exactly x successes in these n trials?

$$P(x) = \frac{\binom{X}{x}\binom{N-X}{n-x}}{\binom{N}{n}} \qquad \left(\text{recall } \binom{n}{x} = \frac{n!}{x!(n-x)!}\right)$$

Hypergeometric: example

- In a batch of 10 lightbulbs, 4 are defective.
- If we select 3 bulbs from that batch, what is the probability that 2 out of the 3 are defective?
  - Population: N=10, X=4
  - Sample (trials): n=3, x=2

$$P(2) = \frac{\binom{4}{2}\binom{10-4}{3-2}}{\binom{10}{3}} = \frac{\left(\frac{4!}{2*2}\right)\left(\frac{6!}{1*5!}\right)}{\left(\frac{10!}{3!*7!}\right)} = \frac{(3!)(6)}{\left(\frac{10*9*8}{3!}\right)} = \frac{3}{10}$$

- In Excel: HYPGEOMDIST(x, n, X, N)
  - HYPGEOMDIST(2, 3, 4, 10) → 30%

TODO

- HW2 (ch2-3): due tonight at 10pm
  - Remember to format as a document!
  - HWs are to be individual work
- Form teams and find data
  - Email me when you know your team
- Dataset description due Tue 4 Oct