Ch5-6: Common Probability Distributions

- *HW3* ***due Thu 10pm***
- *Dataset* ***description due next Tue 7Feb***

31 Jan 2012

Dr. Sean Ho

busi275.seanho.com

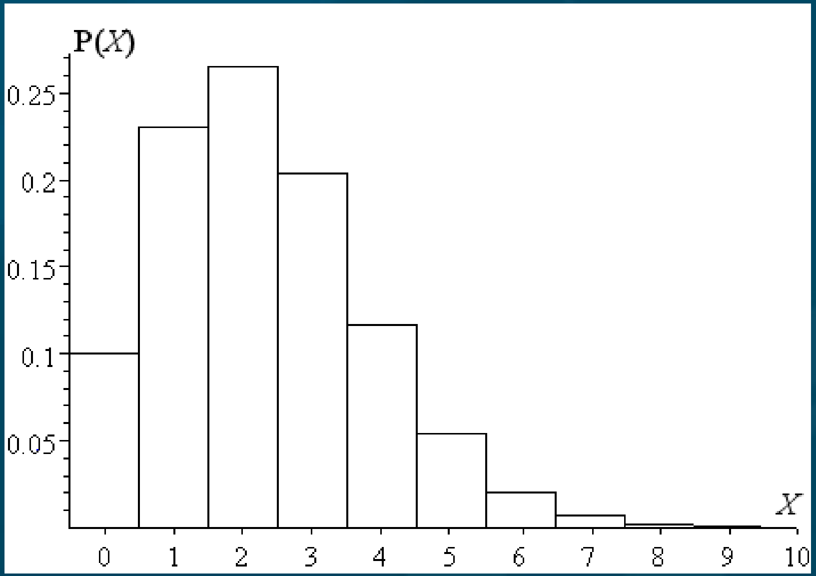
TRINITY WESTERN UNIVERSITY

Outline for today

- Discrete probability distributions
  - Finding μ and σ
  - Binomial experiments: BINOMDIST()
  - Poisson distribution: POISSON()
  - Hypergeometric: HYPGEOMDIST()
- Continuous probability distributions
  - Normal distribution: NORMDIST()
    - Cumulative normal
    - Continuity correction
    - Standard normal
  - Uniform distribution
  - Exponential distribution: EXPONDIST()

Discrete probability distribs

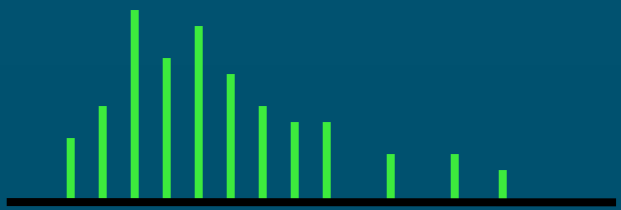
- A random variable takes on numeric values
  - Discrete if the possible values can be counted, e.g., {0, 1, 2, …} or {0.5, 1, 1.5}
  - Continuous if precision is limited only by our instruments
- Discrete probability distribution: for each possible value X, list its probability P(X)
  - Frequency table,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



Probability distributions

Ch. 5

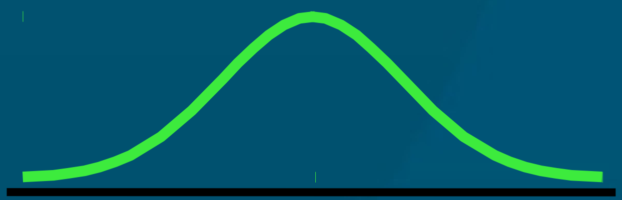
Discrete Random Variable



- **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:

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

Outline for today

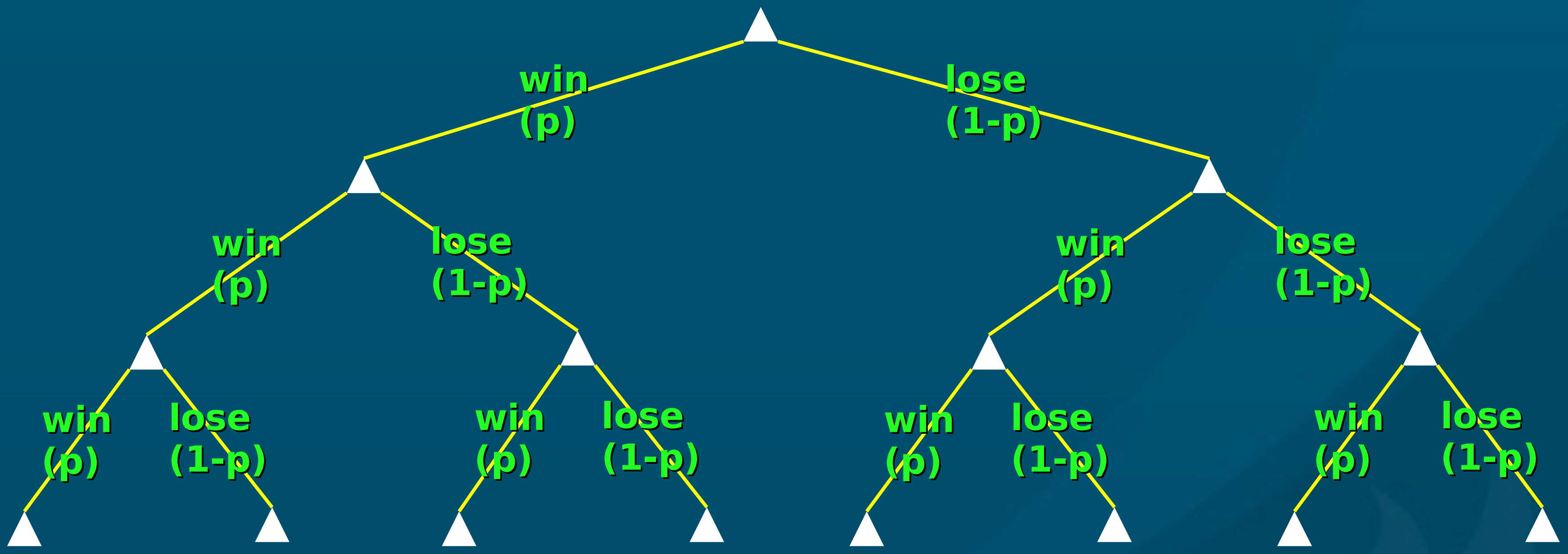
- Discrete probability distributions
  - Finding μ and σ
  - Binomial experiments: BINOMDIST()
  - Poisson distribution: POISSON()
  - Hypergeometric: HYPGEOMDIST()
- Continuous probability distributions
  - Normal distribution: NORMDIST()
    - Cumulative normal
    - Continuity correction
    - Standard normal
  - Uniform distribution
  - Exponential distribution: EXPONDIST()

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:

$$P(x) = \binom{n}{x} (p)^x (1-p)^{(n-x)}$$

- Three parts:
  - Number of combinations: “n choose x”
  - Probability of x successes: p^x
  - Probability of n-x failures: $(1 – p)^{n – x}$

Number of combinations

- The first part, pronounced "n choose x", is the number of combinations with exactly x wins and n-x losses
- Three ways to compute it:
  - Definition: $$\binom{n}{x} = C_x^n = \frac{n!}{x!\,(n-x)!}$$
    - n! ("n factorial") is (n)(n-1)(n-2)...(3)(2)(1), the number of permutations of n objects
  - Pascal's Triangle (see next slide)
  - Excel: COMBIN()

Pascal's Triangle

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

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)

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

Excel: BINOMDIST()

- Excel can directly calculate P(x) for a binomial:
  - BINOMDIST(*x, n, p, cum*)
- e.g., if 60% of customers would buy again, then 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%

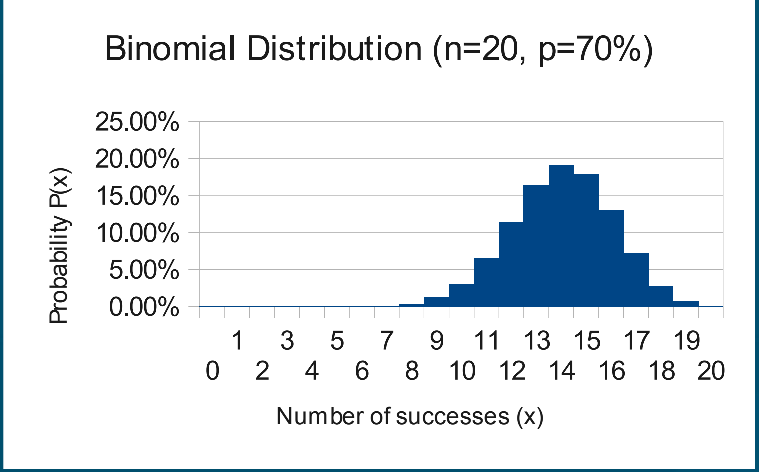
μ and σ of a binomial

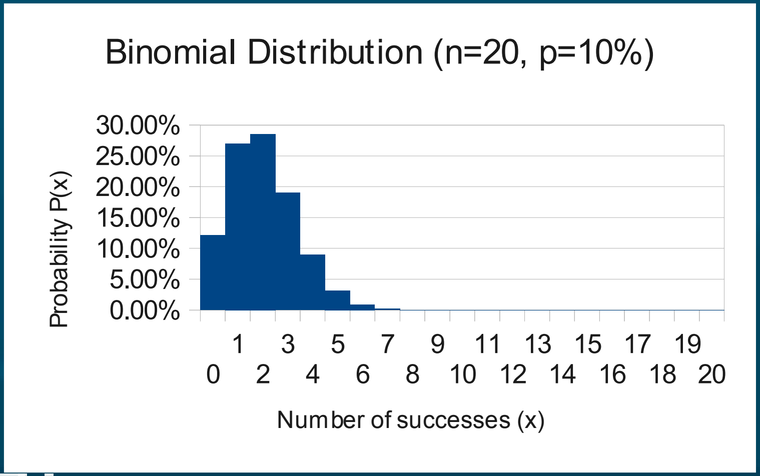
- n: number of trials

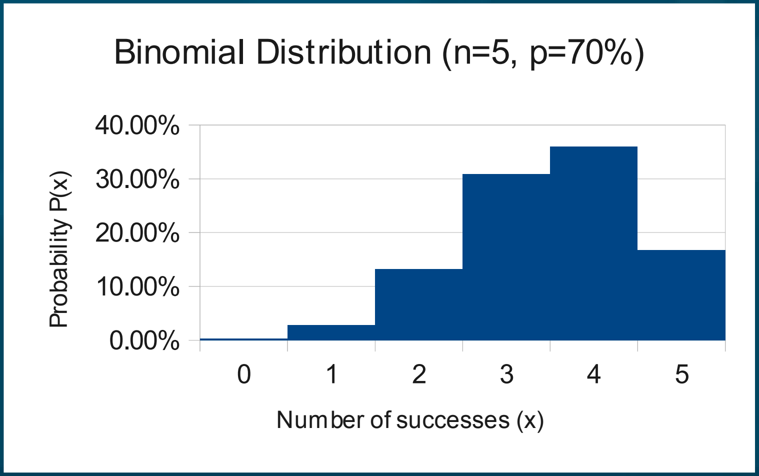
  p: probability of success
- Mean: expected # of successes: $\mu = np$
- Standard deviation: $\sigma = \sqrt{npq}$
- e.g., with a repeat business rate of $p=60\%$, then out of $n=10$ customers, on average we would expect $\mu=6$ customers to return, with a standard deviation of $\sigma=\sqrt{10(.60)(.40)}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

Binomial and normal

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







Outline for today

- Discrete probability distributions
  - Finding μ and σ
  - Binomial experiments: BINOMDIST()
  - Poisson distribution: POISSON()
  - Hypergeometric: HYPGEOMDIST()
- Continuous probability distributions
  - Normal distribution: NORMDIST()
    - Cumulative normal
    - Continuity correction
    - Standard normal
  - Uniform distribution
  - Exponential distribution: EXPONDIST()

Poisson distribution

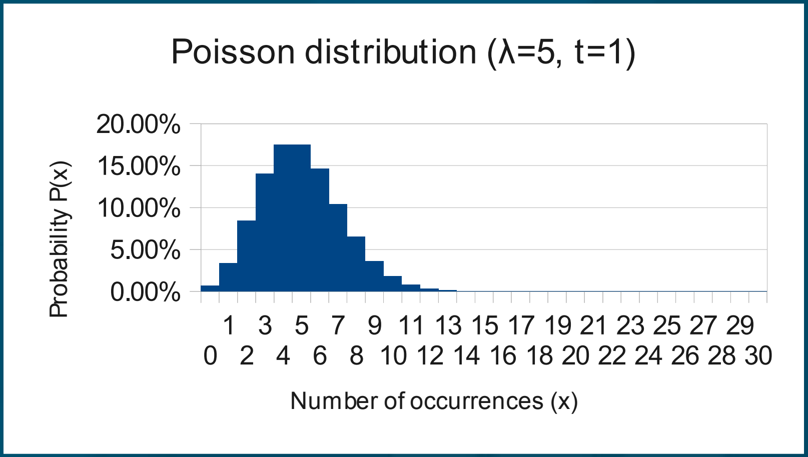
- Counting how many occurrences of an event happen within a fixed time period:
  - e.g., customers arriving at store within 1hr
  - e.g., earthquakes per year
- Parameters: λ = expected # occur. per period
  t = # of periods in our experiment
  - $P(x)$ = probability of seeing exactly x occurrences of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

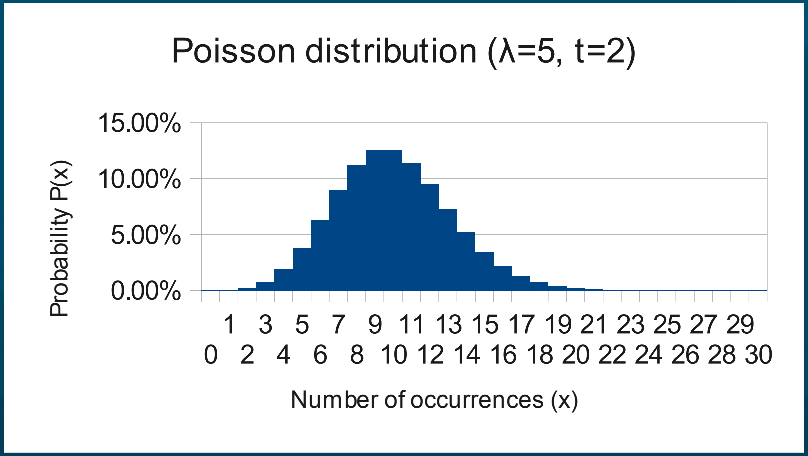
$$P(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

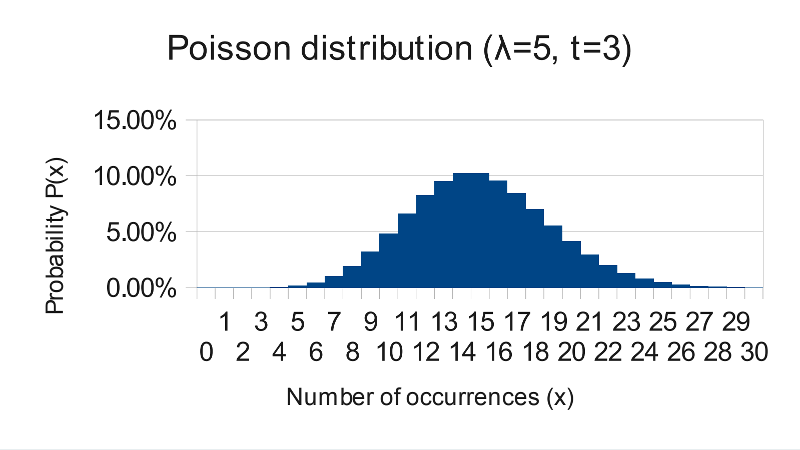
- Mean = λt, and SD = $\sqrt{(\lambda t)}$

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?

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

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%

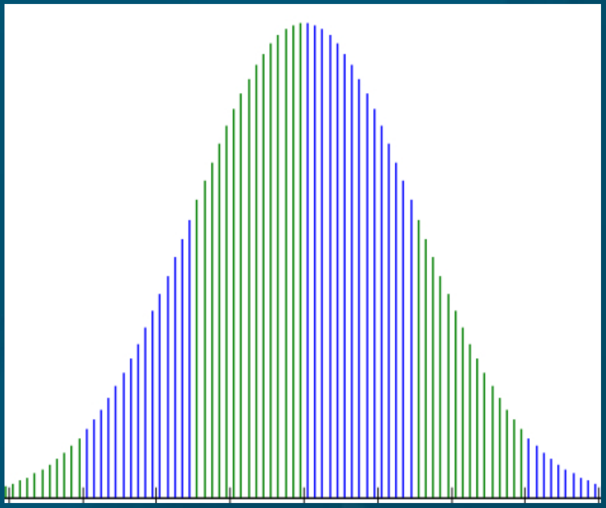
Outline for today

- Discrete probability distributions
  - Finding μ and σ
  - Binomial experiments: BINOMDIST()
  - Poisson distribution: POISSON()
  - Hypergeometric: HYPGEOMDIST()
- Continuous probability distributions
  - Normal distribution: NORMDIST()
    - Cumulative normal
    - Continuity correction
    - Standard normal
  - Uniform distribution
  - Exponential distribution: EXPONDIST()

Normal distribution

- The normal “bell” curve has a formal definition:

$$N(\mu,\sigma)(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



- Mean is μ, standard deviation is σ
- Drops exponentially with z-score
- Normalized so total area under curve is 1
- Excel: NORMDIST(x, μ, σ, cum)
  - e.g., exam has μ=70, σ=10.
    What is probability of getting a 65?
  - =NORMDIST(65, 70, 10, 0) → 3.52%

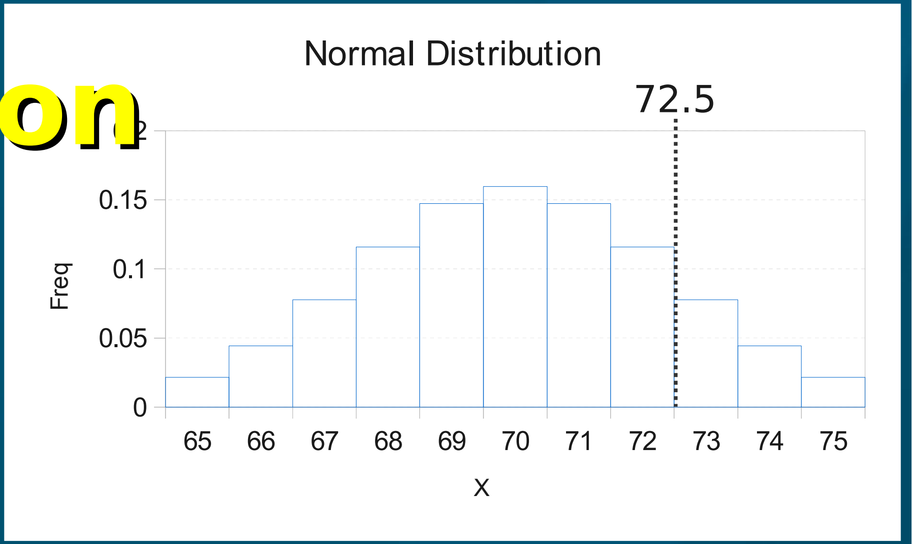
Cumulative normal

- Usually, we are interested in the probability over a range of values:
  - Area of a region under the normal curve
- The cumulative normal gives area under the normal curve, to the left of a threshold:
  - e.g., exam with $\mu=70$, $\sigma=10$. What is probability of getting below 65?
  - =NORMDIST(65, 70, 10, 1) → 30.85%
  - e.g., getting between 75 and 90?
  - =NORMDIST(90, 70, 10, 1) – NORMDIST(75, 70, 10, 1) → 28.58%

Inverse function

- Excel can also find the threshold (x) that matches a given cumulative normal probability:
  - NORMINV(area, μ, σ)
- E.g., assume air fares for a certain itinerary are normally distrib with σ=\$50 but unknown μ. The 90^{th} percentile fare is at \$630. What is the mean air fare?
  - We have: NORMINV(.90, μ, 50) = 630, so
  - =630 – NORMINV(.90, 0, 50) → μ=\$565.92

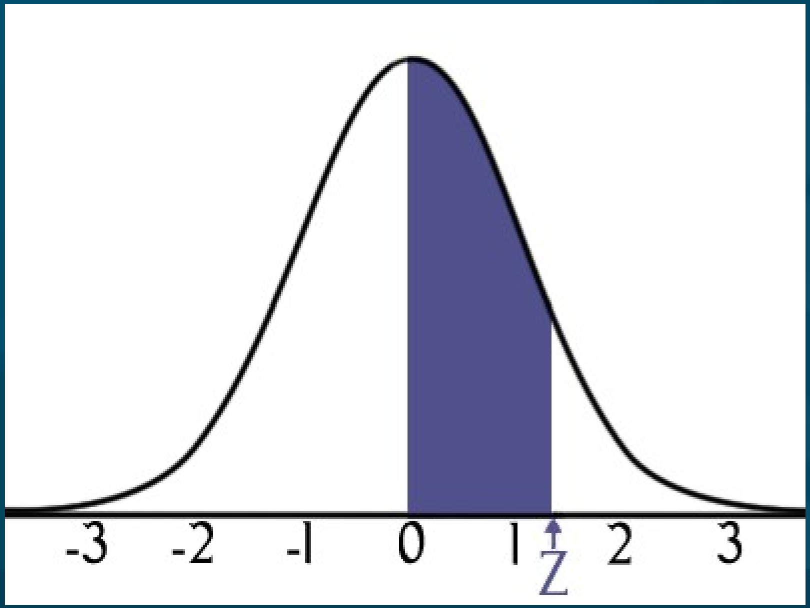
Continuity correction



- For discrete variables (e.g., integer-valued):
  - e.g., # of students per class, assumed to be normally distributed with μ=25, σ=10
- The range can be inclusive or exclusive:
  - Probability of a class having fewer than 10?
    - <10: excludes 10
  - At least 30 students? ≥30: includes 30
- Edge of the bar is at ±0.5 from the centre
  - <10: =NORMDIST(9.5, 25, 10, 1) → 6.06%
  - ≥30: =1-NORMDIST(29.5, 25, 10, 1) → 32.6%

Standard normal

- There is a whole family of normal distributions, with varying means and standard deviations
- The standard normal is the one that has $\mu=0$, $\sigma=1$
- This means z-scores and x-values are the same!
- In Excel: NORMSDIST(x) (cumulative only) and NORMSINV(area)



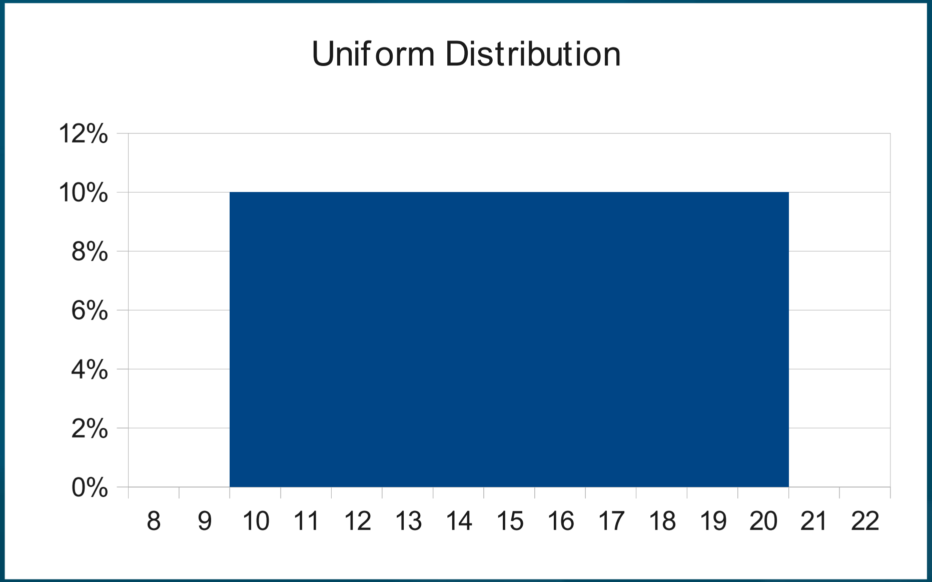
Outline for today

- Discrete probability distributions
  - Finding μ and σ
  - Binomial experiments: BINOMDIST()
  - Poisson distribution: POISSON()
  - Hypergeometric: HYPGEOMDIST()
- Continuous probability distributions
  - Normal distribution: NORMDIST()
    - Cumulative normal
    - Continuity correction
    - Standard normal
  - Uniform distribution
  - Exponential distribution: EXPONDIST()

Uniform distribution

- With a uniform distribution,
  all values within a range are equally likely
  - e.g., roll of a fair die:
    {1,2,3,4,5,6} all have probability of 1/6
  - Range is from a to b:

$$U(x) = \begin{Bmatrix} \dfrac{1}{b-a} & \text{if } a \leqslant x \leqslant b \\ 0 & \text{otherwise} \end{Bmatrix}$$



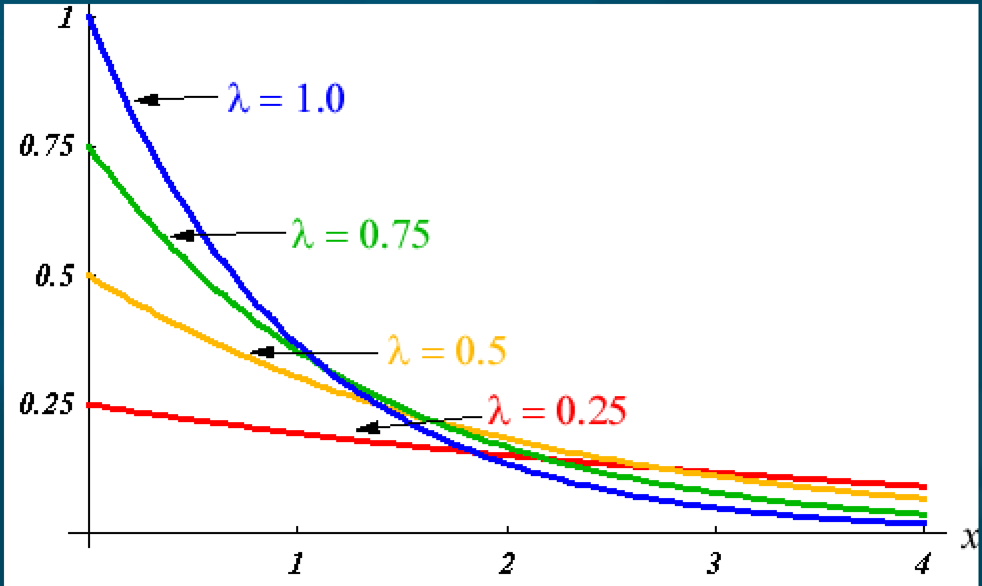
- $\mu = (a+b)/2$, $\sigma = \sqrt{(b-a)^2/12}$

Exponential distribution

- Time between occurrences of an event
  - e.g., time between two security breaches
- Exponential density: probability that the time between occurrences is exactly x is:

$$E(x) = \lambda e^{-\lambda x}$$

  - $1/\lambda$ = mean time between occurrences
  - Need both $x, \lambda > 0$
- EXPONDIST(x, λ, cum)
  - Density: cum=0



Exponential probability

- Exponential probability (cumulative distribution) is the probability that the time between occurrences is less than x:

$$P\left(0 \le x \le a\right) = 1 - e^{-\lambda a}$$

  - Excel: EXPONDIST(x, λ, 1)
- e.g., average time between purchases is 10min. What is the probability that two purchases are made less than 5min apart?
  - EXPONDIST(5, 1/10, 1) → 39.35%
  - Don't forget to convert from 1/λ to λ

TODO

- HW3 (ch4): due this Thu 26Jan
- Proposal meetings this week
  - Submit proposal ≥24hrs before meeting
- Dataset description next week: 7Feb
  - If using existing data, need to have it!
  - If gather new data, have everything for your REB application: sampling strategy, recruiting script, full questionnaire, etc.
- REB application in two weeks: 14Feb (or earlier)
  - If not REB exempt, need printed signed copy