How was spring break?

1 - I slept the whole time

5 - I worked the whole time

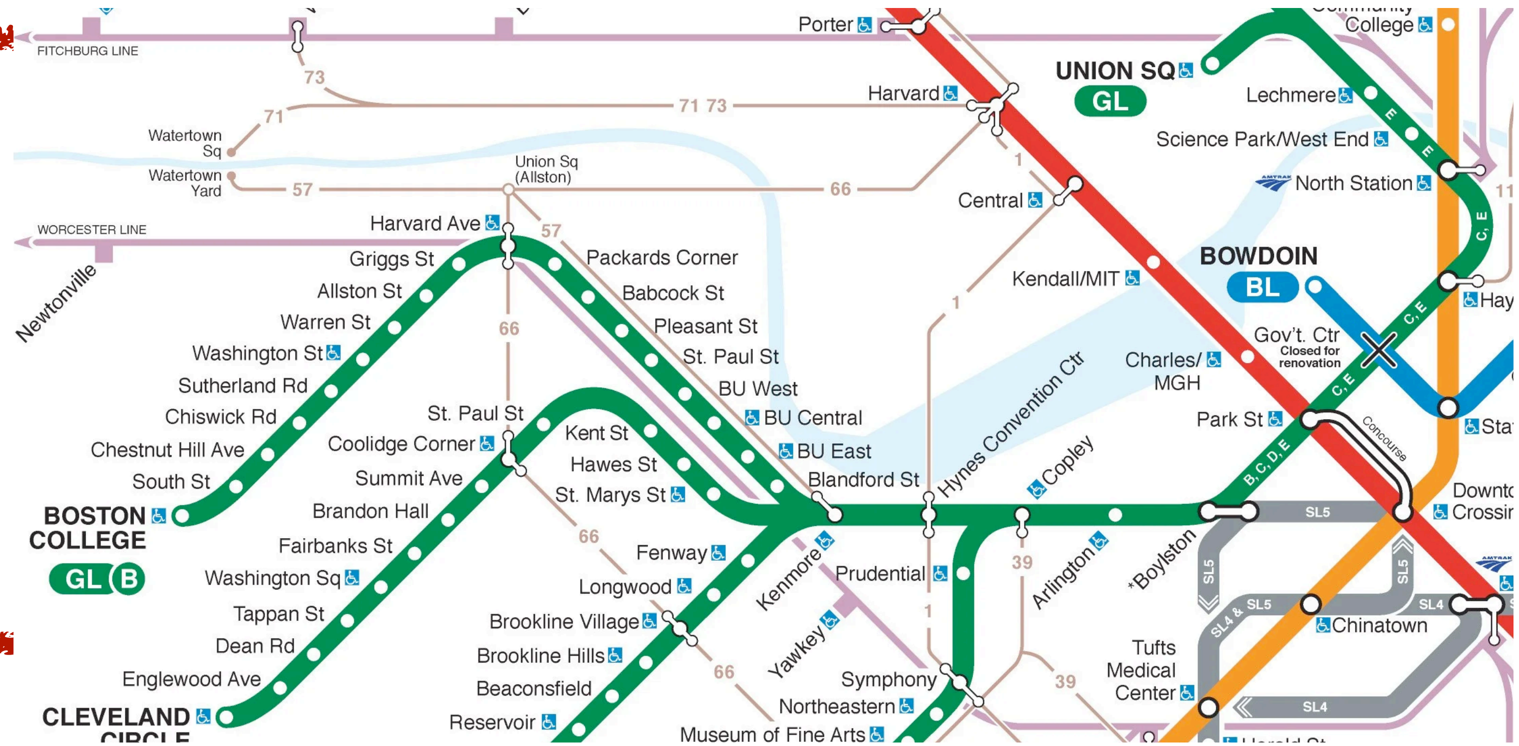
10 - I had parties the whole time

Normal distributions, central limit theorem, cumulative distribution function

GL Green Line Service Changes

Beginning Monday, March 21, Green Line service between North Station and Lechmere will resume, and regular service on the Union Branch will begin.

- B and C Branch trains will terminate at Government Center
- D Branch trains will terminate at North Station
- E Branch trains will terminate at Union Square



Distributions and probability mass functions

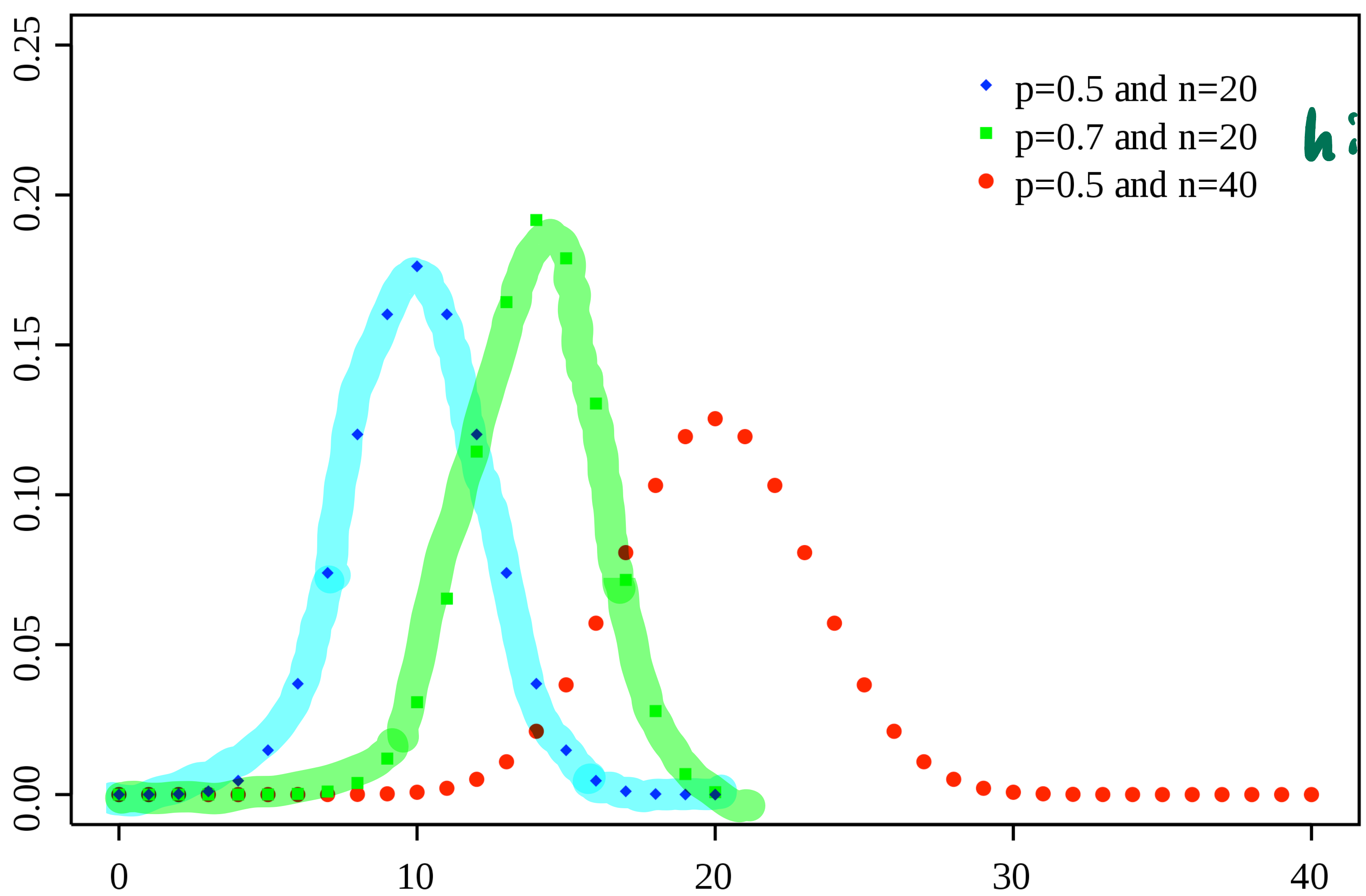
- "All distributions have a **probability mass function.** This tells us how the mass is distributed across outcomes."

- For a binomial distribution, the shape of this function depends on p → likelihood of success and n → # of trials

- For a poisson distribution, the shape of this function depends on λ → rate of occurrence

Distributions and probability mass functions

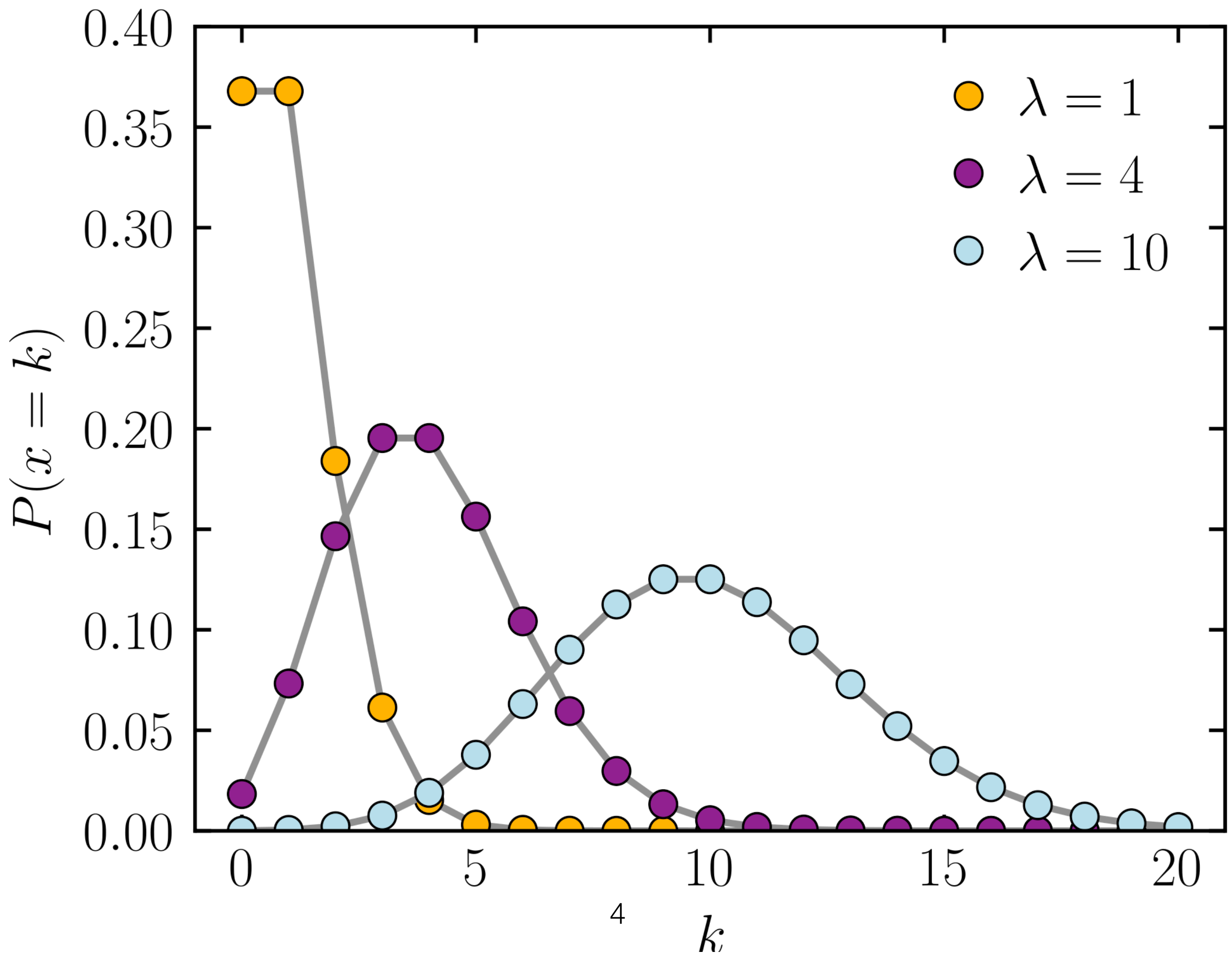
- For a binomial distribution, the shape of this function depends on
  ____p____ and ____n____



higher p
value

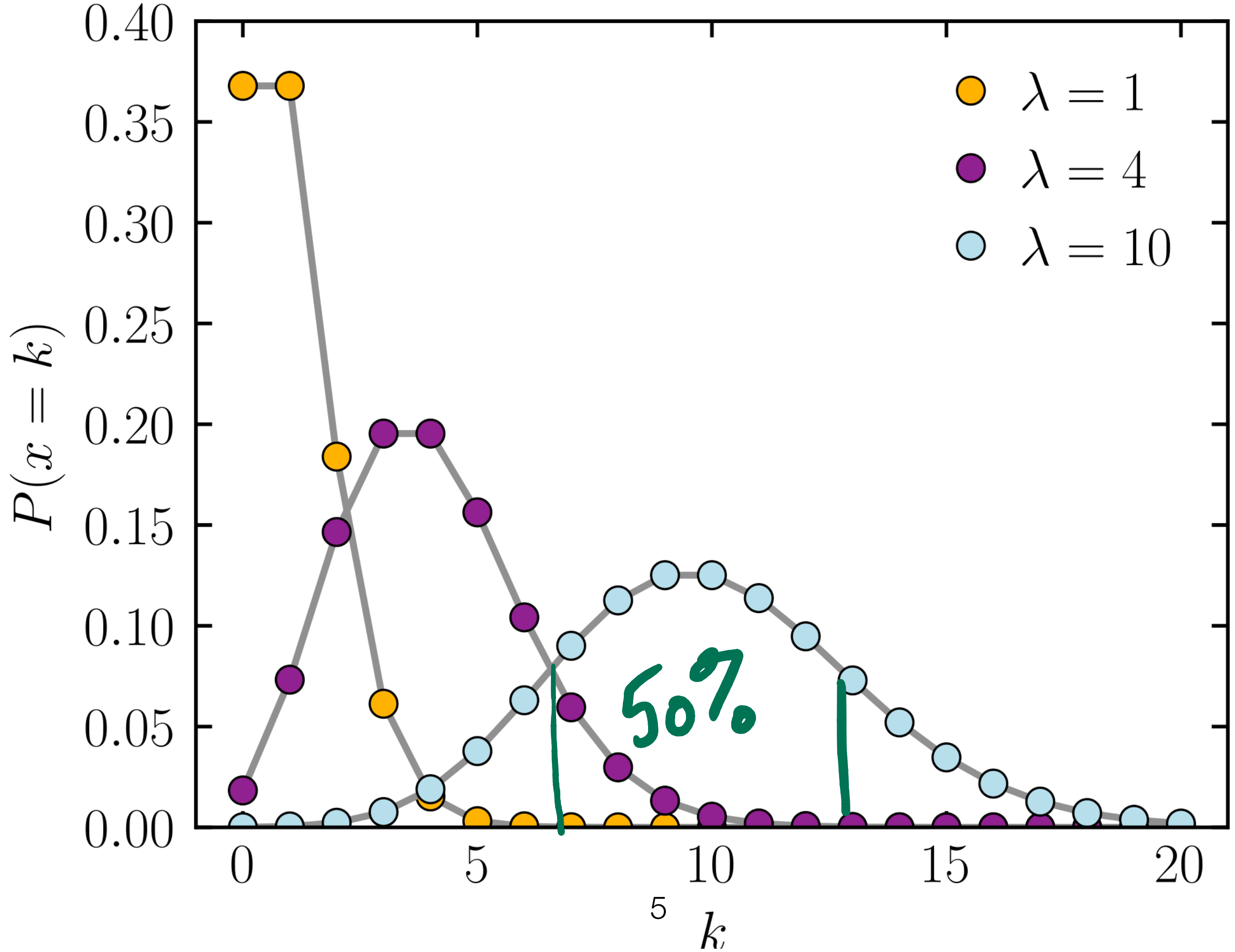
Distributions and probability mass functions

- For a poisson distribution, the shape of this function depends on $\underline{\lambda}$



Distributions and probability mass functions

- The area under the curve for a probability mass function sums to: one
- This means that one of the outcomes will happen



Distributions and probability mass functions

- We want to pick the best probability distribution for the events that we're observing to create the best model/the model that is closest to the ground truth.

- Binomial distributions: discrete r.v.s - only success/failure, # of successes in a given # of trials

- Poisson distributions: discrete r.v.s. - only happened/didn't happen, # of occurrences in a time span

- both of these are for discrete variables
  ↳ B(p, 1 ... 3) to sum to one

ICA Question 1: distributions [binomial, poisson, neither]

What is the best distribution to model each of the following events that you observe in the world?

A. Number of new trains observed at the Park Street station - binomial*
B. Number of trains to arrive at Union ~~Station~~ Square in the span of 10 minutes - poisson
C. Times that it takes for the Green Line to travel between Lechemere and Union Square - neither
D. Number of donuts that Felix buys from Union Square donuts per hour - poisson
E. Diameters of donuts bought from Union Square donuts today - neither
F. Number of Union Square donuts that are filled in a box of 12 donuts
↳ binomial ↳ n

Finally, write down something that you observe in the world around you and what kind of distribution best fits it.

A. binomial

↳ new / not new

↳ newness of trains is independent

this lecture

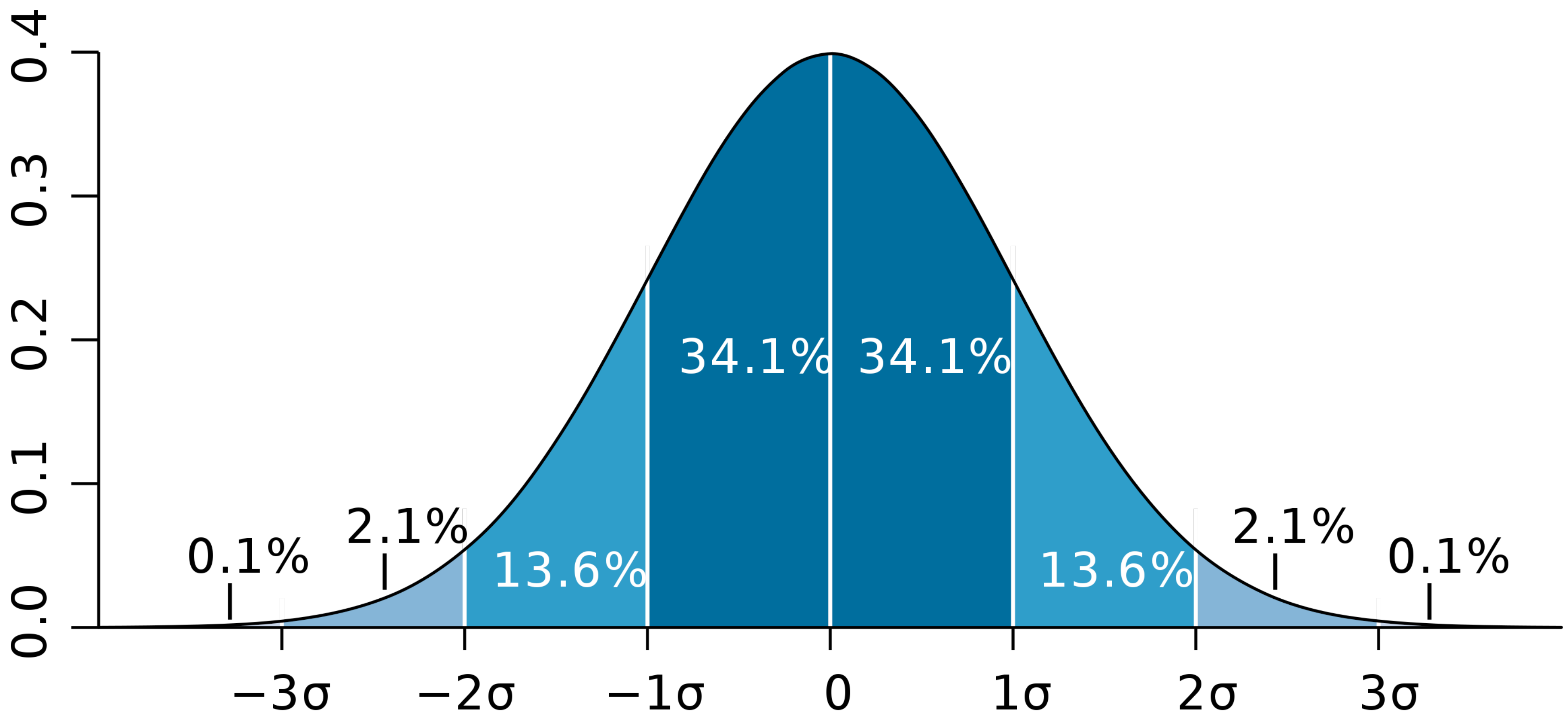
for all train examples, assume that train newness / travel time / etc are independent

Normal Distributions

- A **normal (or Gaussian)** distribution is a bell-shaped curve.

std dev variance

- Previously, we used it to ground what σ means (and σ^2)



Normal Distributions

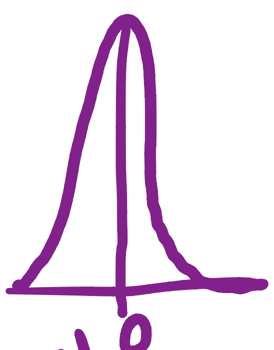
- A **normal** distribution is a bell-shaped curve that models a random variable that has a default mean, an expected variation about that mean, and whose values are real-valued

SD σ

- in contrast to discrete

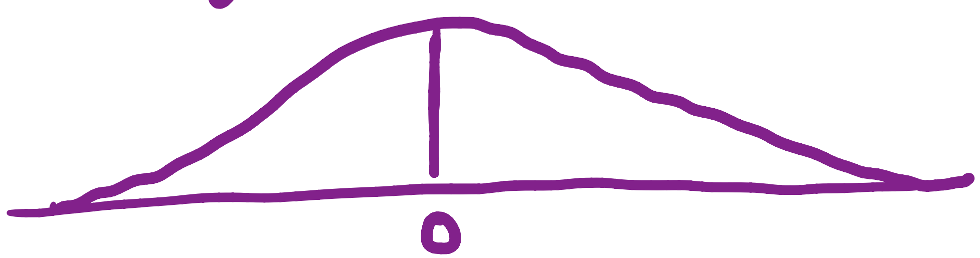
- heights
- times

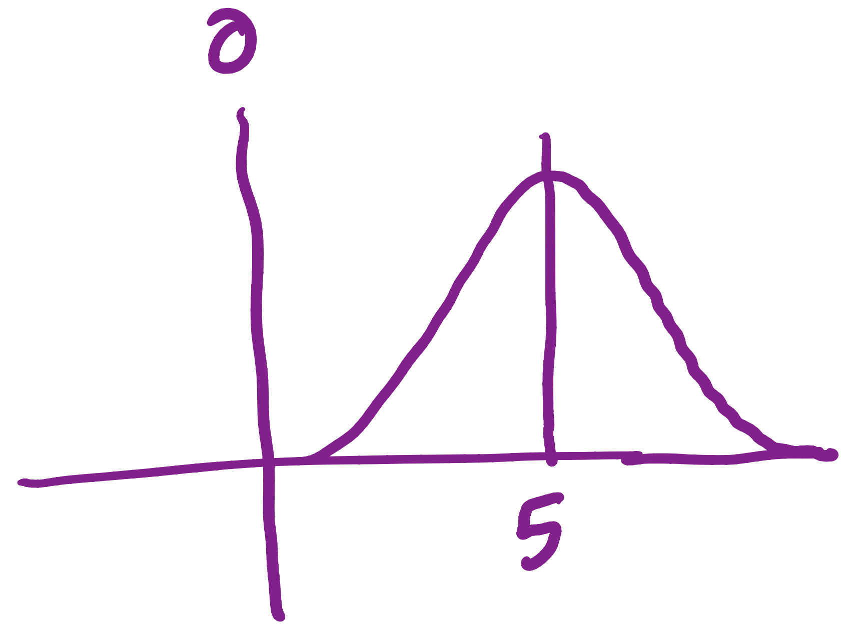
small σ



small σ

large σ





Normal Distributions

- A **normal** distribution is a bell-shaped curve that models a **real-valued random variable**.

D vs. mass

- It is defined by the **probability *density* function** $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

f(5)

- Where μ is the mean (expected value) and σ is the standard deviation of the random variable

- If we say that $X \sim N(5,4)$, this means:

X is a r.v. w/ mean 5 and variance 4

ICA Question 2: Normal distributions and pdf

The times between North Station and Union Square have a mean of 5 minutes and a variance of 4 minutes.

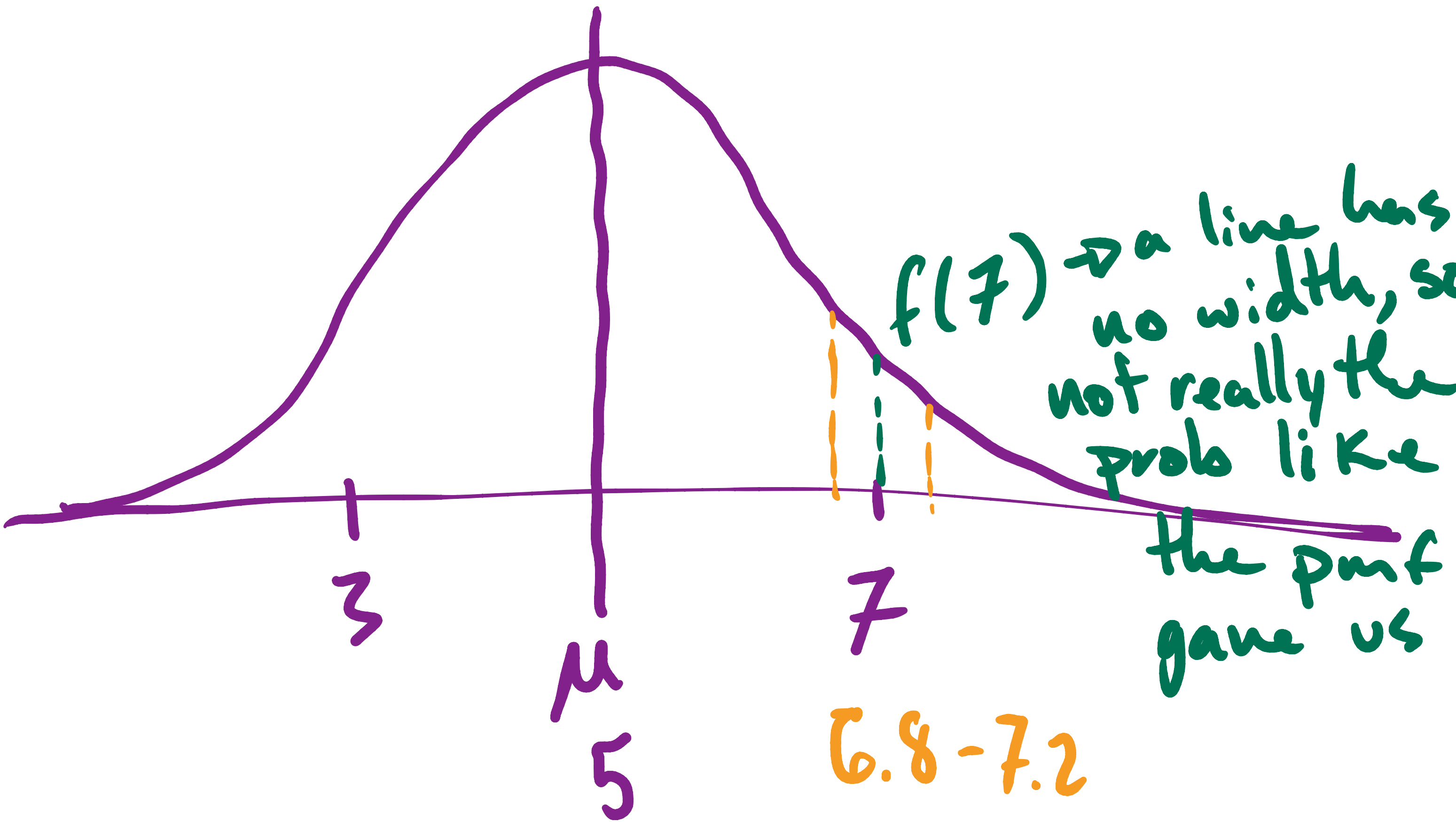
What is f(5) if $X \sim N(5,4)$ and $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$?

σ μ

yes, this is all in the exponent!

0.1997

Pdfs (real-valued r.v.)



Central Limit Theorem

- The **central limit theorem** says that if a population has a mean μ and a standard deviation σ, if we take enough samples, with replacement, the samples' means will be normally distributed.

- Requirements:

- a coin flip - heights
- die rolls

  - observations must be independent from/dependent on one another
  - the mean and the variance must be defined and finite
  - observed random variables must/don't have to be normally distributed

- sample size? sufficiently big enough

ICA Question 3: central limit theorem

```
# for coin flipping
import random
# for graphing
import matplotlib.pyplot as plt

# central limit theorem
def flip_coin(times):
    return [random.randint(0, 1) for t in range(times)]

samples = YOUR NUMBER HERE
times = YOUR OTHER NUMBER HERE
averages = []
for sample_num in range(samples):
    coin_flips = flip_coin(times)
    averages.append(sum(coin_flips) / len(coin_flips))

plt.hist(averages)
plt.show()
```

Cumulative Distribution Functions

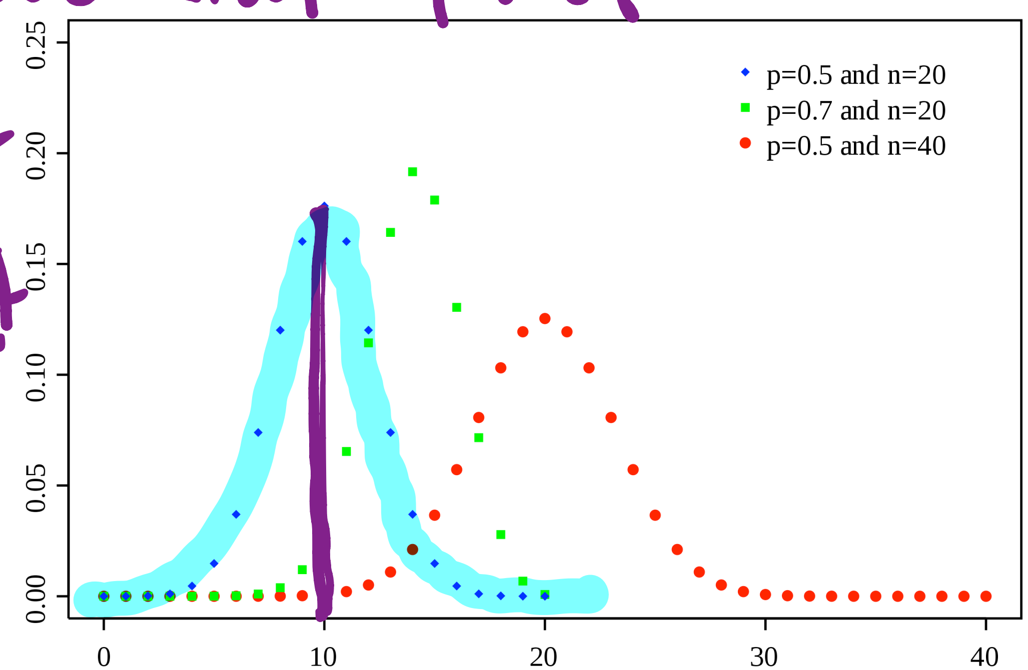
- A probability mass function or a probability density function tells us "what is the probability that a random variable will take this value according to the underlying distribution"

$f(10, \ldots)$

- A **cumulative distribution function** tells us the probability that a random variable will take a value less than or equal to a target value → x-axis

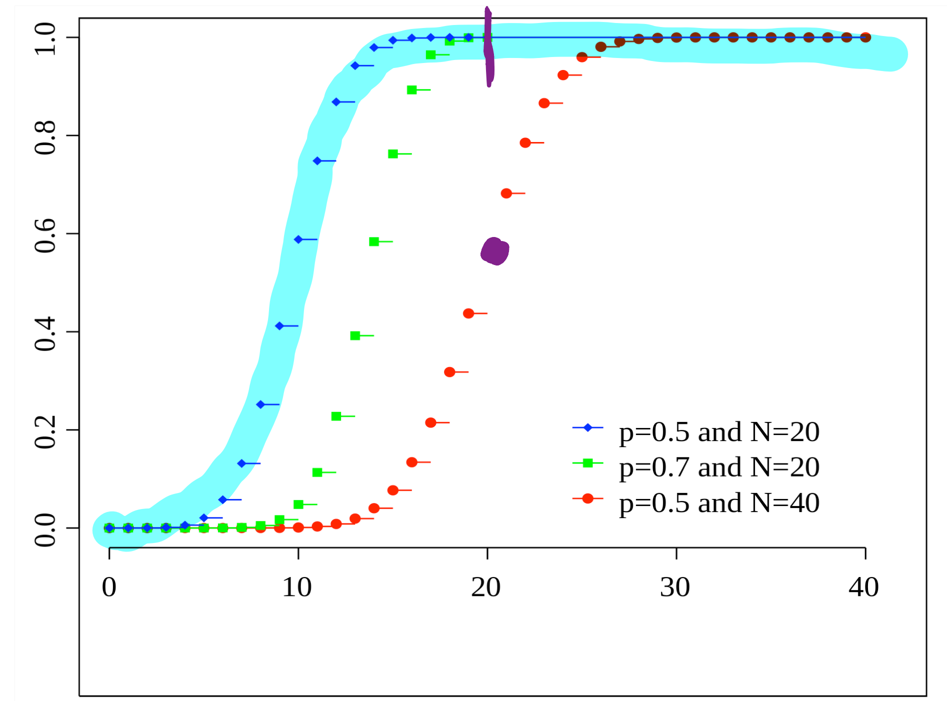
binomial -pmf

width:1 times height



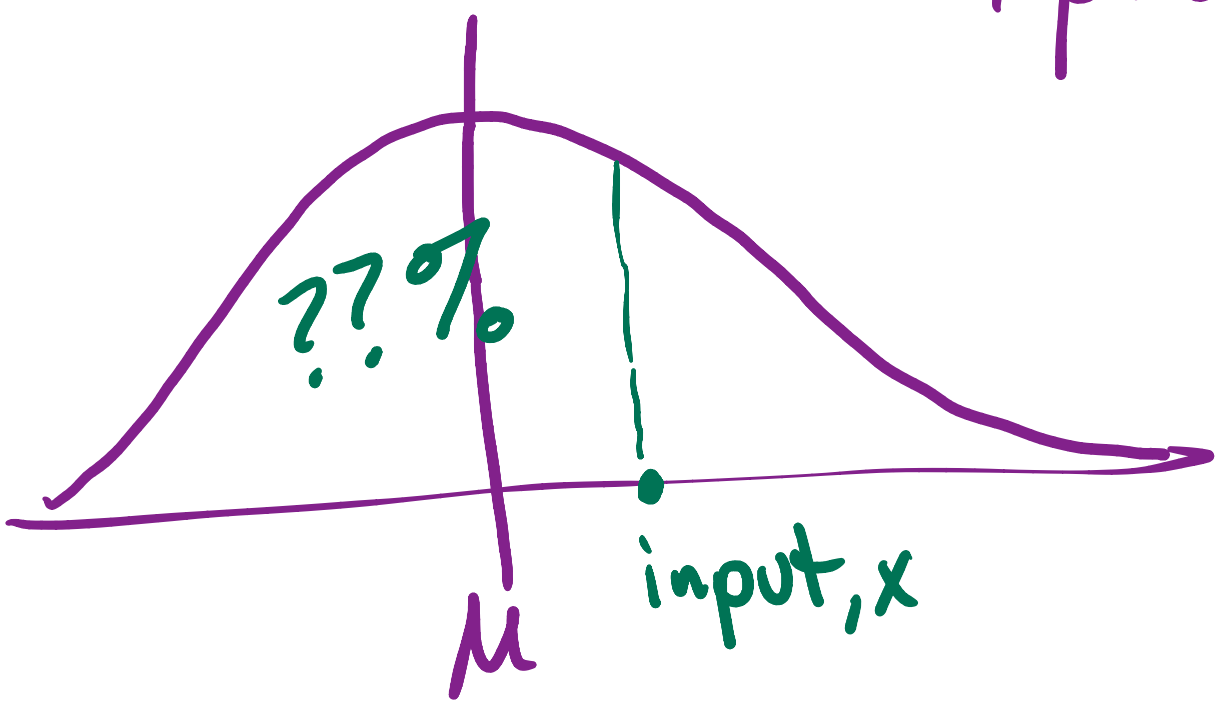
1

0



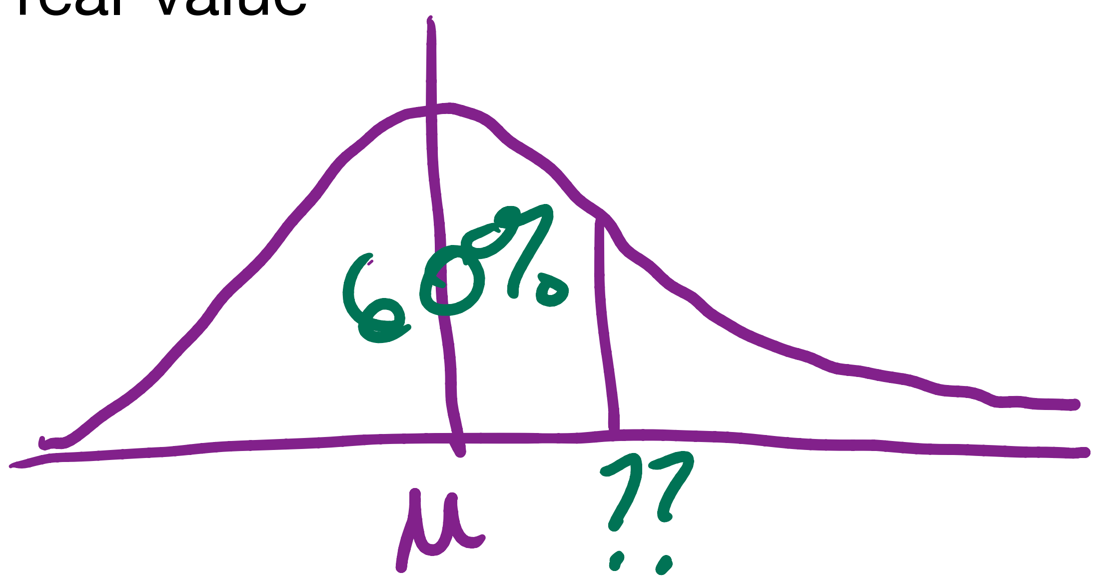
Cumulative Distribution Functions

- A **cumulative distribution function** tells us the probability that a random variable will take a value less than or equal to a target value.
- Input: a real-value 5.5
- Output: a percentage chance /probability



Percent Point Functions

- A **percent point function** tells us the value of x for which some percentage of the normal distribution is at or under that value.
- Input: a percentage chance
- Output: a real-value



ICA Question 4: cdf/ppf

The times between North Station and Union Square have a mean of 5 minutes and a variance of 4 minutes. If $X \sim N(5,4)$, then, using scipy, answer the following questions:

1) If I know that I'm very lucky and I expect my travel time to be in the **bottom** 10% of times, how long should I budget for my trip?

   ↳ ppf for .1 → 2.44

2) What percentage of trains can I expect to take 5 - 8 minutes (inclusive) for this trip?

   ↳ cdf (8) - cdf (5) → .43 → 43%

Mini-project clarifications

- (There's a pinned piazza post with these as well) → will be this afternoon
- submission format? length? no constraints on format/length.
- examples? unfortunately none :(
- thoroughness of the application of math topics? scale this appropriately. If you are working individually, expect to spend about 25 minutes actually writing down the explanation/grounding for each math topic.
- how much time? approx 200 minutes outside of class time — we are assuming that you may need to review the topics from class but that you do not need to learn them from scratch in this estimate

Mini-project clarifications

- what kind of scenarios to consider? up to you! pick ones that demonstrate the math topics that you've chosen well
- is it necessary to code the demonstrations or is it okay to just explain the concept? it is not necessary to code the demonstrations
- what can be included in the research? math topics? code implementation? calculations? yes. Anything you need to do that is not creating your actual deliverable
- is it necessary to collect actual data? no, but fabricated data should be reasonable

Mini-project clarifications

- More questions? post on the pinned piazza post and/or come to Felix's office hours!

Schedule

ICA passcode: "green"

Turn in ICA 16 on Canvas (make sure that this is submitted by 2pm!)

HW 6's final due date is on Tuesday. No late day deductions for this HW!

HW 7 will be released on Thursday, it is due on April 3rd. You'll need content from Thursday's/ next Monday's lecture for this HW.

Test 3 (your last in-class test (**Test 4** is during your final exam slot)) is the Thurs. after this one

Mon	Tue	Wed	Thu	Fri	Sat	Sun
March 21st Lecture 16 - normal distributions	**Felix OH Calendly** **HW 6 due @ 11:59pm**	**Felix OH Calendly**	**Felix OH Calendly** Lecture 17 - hypothesis testing			
March 28th Lecture 18 - t-tests, experimental bias	**Felix OH Calendly**	**Felix OH Calendly**	**Felix OH Calendly** **Test 3 (HW 5/6)**			**HW 7 due @ 11:59pm**

More recommended resources on these topics

- Probability density functions: YouTube, 3Blue1Brown -- Why "probability of 0" does not mean "impossible" | Probabilities of probabilities, part 2

- why approximating a normal CDF is hard: Wikipedia, https://en.wikipedia.org/wiki/Normal_distribution#Numerical_approximations_for_the_normal_CDF_and_normal_quantile_function

- Central Limit Theorem: YouTube, Central limit theorem | Inferential statistics | Probability and Statistics | Khan Academy