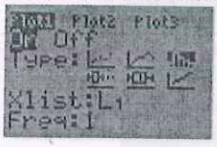
9.4 – Statistics: Introduction, Measures of Central Tendency (Center)

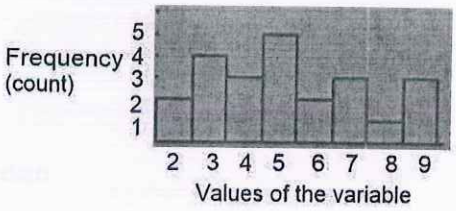
Given a set of data, it is always helpful to view the distribution (histogram):

2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 8, 9, 9, 9

1) 2nd, '+' (Mem), ClearAllLists, <enter>
2) STAT, Edit
3) Enter the data set into list 1 (L1)
4) 2nd, Y=, set as follows:

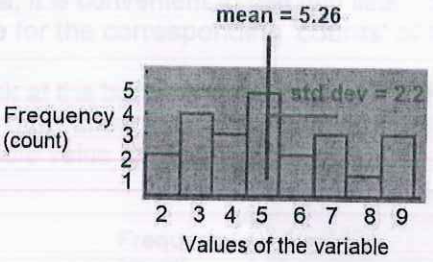


Histogram



4) 2nd, Mode to exit
5) y=, clear out any equations
6) Zoom, 9:StatPlot
7) Window, Xscl=1
8) Graph

But it is also useful to have a single number which summarizes something about a data distribution. Such a number is called a ***statistic***.



Some numbers, such as 'mean' are measures of central tendency (or center)

Some numbers, such as 'standard deviation' are measures of dispersion (or spread)

Measures of central tendency (center):

Mean

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Median

- Arrange data in order.
- Median is middle value.
- If n is even, median is average of two middle values.

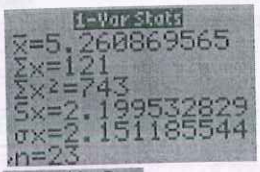
Mode

- Most frequently occurring value or values.
- Can be more than one mode (bimodal, etc.)

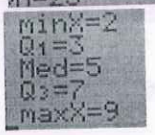
If you have a complete data set and can enter it into a calculator, the calculators '1-Var Stats' function can provide mean and median:

1) Enter data into L1
2) Stats, right arrow to 'CALC'
3) 1-Var Stats

Result...



← mean

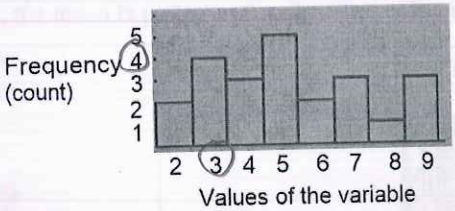


← median

Grouped data: Sometimes, we don't have a complete data set, but we have information on groups of data. If we were just given the histogram, we know that there are 4 data values somewhere between 2.5 and 3.5, but we don't know exactly what those values are. They might be 3, 3, 3, 3 or they might be 2.7, 2.8, 3.1, 3.2.

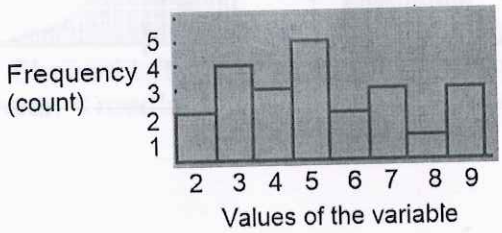
Even with grouped data, we can obtain estimated statistics. To enter the data, it is convenient to use two lists: one for the data values (L1) and one for the corresponding 'counts' of the data values.

Look at the bar for the data value 3. There is a count of 4 data values. We don't know their exact values, but we will assume they are all at the middle value for this bar which is '3'.

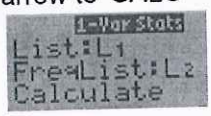


So we can enter this histogram's grouped data into lists L1, and L2:

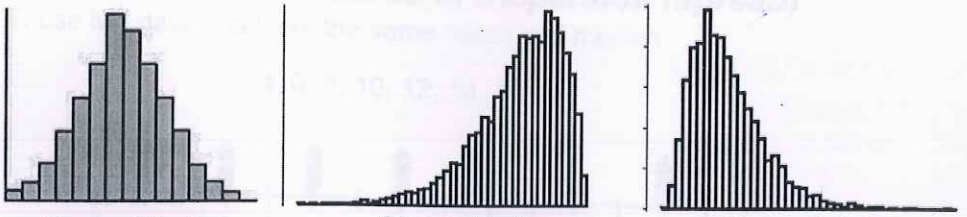
data values L1	counts L2
2	2
3	4
4	3
5	5
6	2
7	3
8	1
9	3



1) Enter data into L1 and counts into L2
2) Stats, right arrow to 'CALC'
3) 1-Var Stats



Shape: Symmetric or skewed



Symmetrical

Skewed left
(tail toward lower values)

Skewed right
(tail toward higher values)

How mean, median, and skew are related

What is the mean and median of this data set?

6 7 8 9 10

$$\overline{X} = \frac{6+7+8+9+10}{5} = 8$$

$$Median = 8$$

If we move the '10' farther away from the mean, how does mean and median change?

6 7 8 9 20

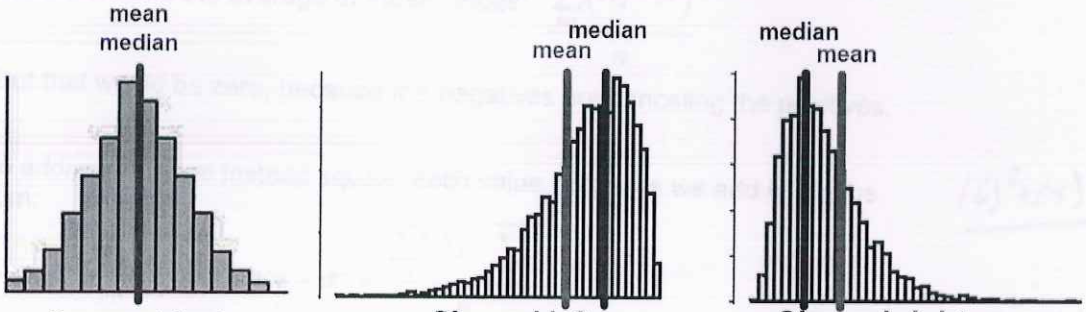
$$\overline{X} = \frac{6+7+8+9+20}{5} = 10$$

$$Median = 8$$

The median is unaffected, but the outlier 'pulls the mean towards itself'

The median is unaffected, but the outlier 'pulls the mean towards itself'

In a perfectly symmetric distribution, the mean and median are aligned, but in skewed distributions, the mean is pulled in the direction of the tail:



Symmetrical

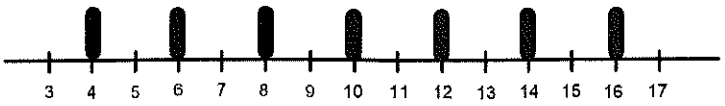
Skewed left
(mean < median)

Skewed right
(mean > median)

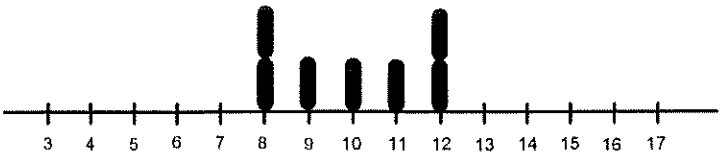
9.5 – Statistics: Measures of Dispersion (spread)

These two data sets have the same mean and median...

4, 6, 8, 10, 12, 14, 16

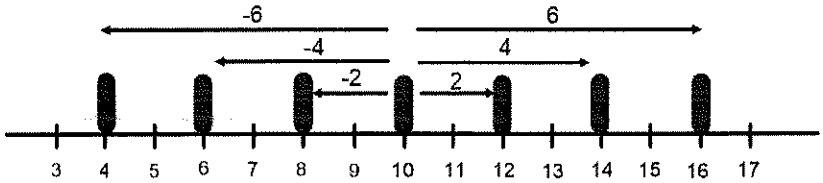


8, 8, 9, 10, 11, 12, 12



...but the data in the top set is spread out further from the mean.

We might ask, 'what is the typical, or average, distance data points are away from the mean?'



$\overline{X} = 10$

$\frac{6+4+2-6-4-2}{6} = 0$

For each point, we could compute the difference between the data value and the mean... $X - \overline{X}$

...and then take the average of these values: $\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}{n}$

But that would be zero, because the negatives are canceling the positives.

To address this, we instead square each value before as we add it into the sum:

$$variance = \sigma^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}$$

$\frac{(6)^2+(4)^2+(2)^2+(-6)^2+(-4)^2+(-2)^2}{6}$

$\sigma^2 = 18,66\overline{6}$

The result is the ***variance*** which is a statistic that gives the average distance squared that the data points are spread around the mean.

There are two problems with variance as a measure of the spread:
1) It isn't 'intuitive' to think in terms of 'distances squared'
2) The variance is in terms of units squared. For example, if this data represented lengths of fish in cm, the variance would be in cm^2

To get a better measure of dispersion (or spread) that is more intuitive and matches the units of the data set, we just take the square-root of the variance. This is called ***standard deviation:***

$$standard\ deviation = \sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}}$$

Standard deviation represents the average distance that the data points are from the mean.

Measures of Dispersion

There are many different ways to represent dispersion. A few of the more commonly used are:

1. **Range**-- the difference between the largest value and the smallest value.
2. **Variance**--deviation from the mean (averaged square deviation)

$$\sigma^2 = \frac{\left(x_1 - \overline{X}\right)^2 + \left(x_2 - \overline{X}\right)^2 + \ldots + \left(x_n - \overline{X}\right)^2}{n} = \sum \frac{\left(x_i - \overline{X}\right)^2}{n}$$

3. **Standard Deviation**--square root of the variance, 'the average distance that data points are from the mean':

$$\sigma = \sqrt{\frac{\left(x_1 - \overline{X}\right)^2 + \left(x_2 - \overline{X}\right)^2 + \ldots + \left(x_n - \overline{X}\right)^2}{n}} = \sqrt{\sum \frac{\left(x_i - \overline{X}\right)^2}{n}}$$

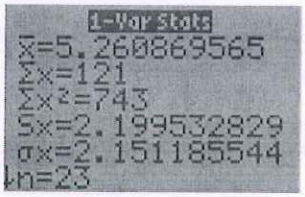
In general, a relatively small standard deviation indicates that the measures tend to cluster close to the mean.

Relatively high standard deviation shows that the measures are widely scattered from the mean.

Using a calculator to compute standard deviation:

1) Enter the data either into L1, or for grouped data, into L1 and counts in L2.
2) STAT, right arrow, CALC, 1-Var Stats

Results...



The calculator gives two different versions of standard deviation:

$\longleftarrow s$

$\longleftarrow \sigma$

The difference between s and σ is subtle. Usually, you should use the s value.

ɔpulations vs. samples from populations

Usually, when we are looking at data that has been collected, it has been collected as being a representative **sample** from a large population about which we are trying to draw conclusions.

For example, let's say we are looking at some data measured using the students in this room. We can calculate statistics about this data, but we would usually be measuring this room's students because we are trying to learn something about ***all*** finite math students. That would make 'all finite math students' the population, and this particular class' students a **sample** from this population.

If your data is a sample - use s

If your data is the entire population - use σ

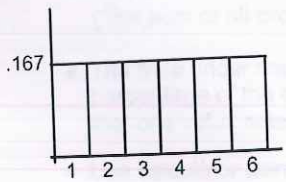
The difference between s and σ is that in the s version we actually divide by n-1, not n:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}}$$

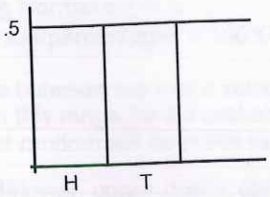
If you want to know more about this difference, please read the posted explanation 'Why is it n-1 instead of n?' on www.mrfelling.com in the 'extras' section of our class page.

9.6: The Normal Distribution and Z-scores

Uniform probability distributions (probability density functions)

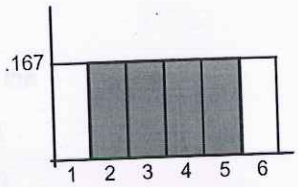


Probability of rolling a number



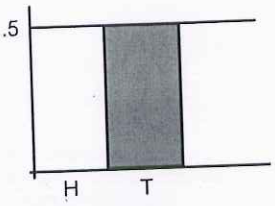
Probability of single coin flip

Area under probability density curve represents probability of an event:



Probability of rolling a number

$P(2,3,4,5) = \frac{4}{6} = \frac{2}{3}$

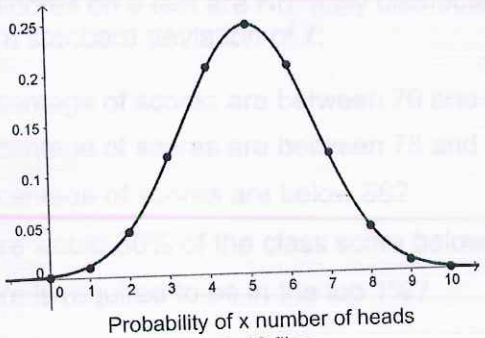


Probability of single coin flip

$P(T) = \frac{1}{2}$

Toss a fair coin 10 times. What is the probability of getting no heads? 1 head? 2 heads?

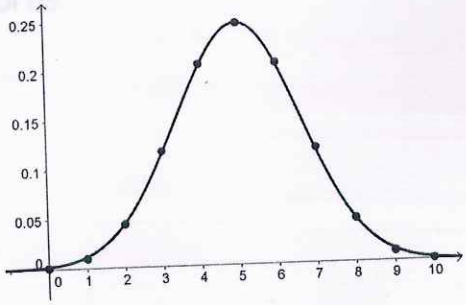
$b(10,k;0.5)$
$b(10,0;0.5) = .0010$
$b(10,1;0.5) = .0098$
$b(10,2;0.5) = .0439$
$b(10,3;0.5) = .1172$
$b(10,4;0.5) = .2051$
$b(10,5;0.5) = .2461$
$b(10,6;0.5) = .2051$
$b(10,7;0.5) = .1172$
$b(10,8;0.5) = .0439$
$b(10,9;0.5) = .0098$
$b(10,10;0.5) = .0010$



Probability of x number of heads in 10 flips

Data from many experiments follow this common pattern:

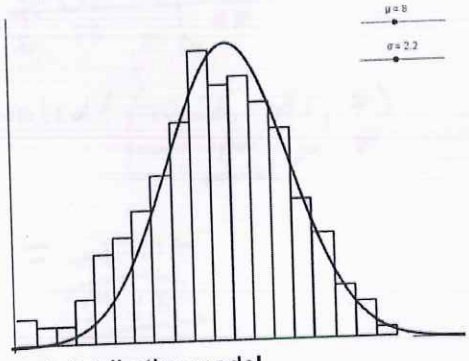
- Heights of adults
- Weights of adults
- Test scores
- Coin tossing



Normal distribution (or Gaussian distribution)

Normal distribution model

Datasets like these can be modeled using a ***Normal Distribution Model:***



Normal distribution model

Don't need to know, but in case you're interested here is the function model for a Normal distribution:

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

(geogebra demo 'normalparameters.ggb')

English alphabet letters for data...

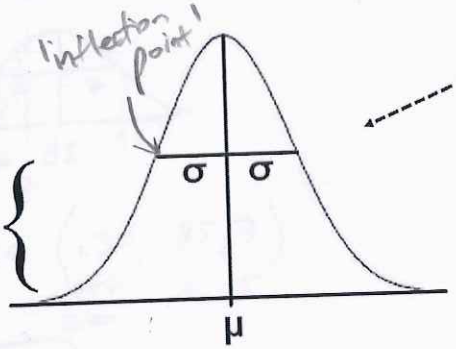
$\bar{x}$ = mean
s = standard deviation

Greek letters for models (populations)...

μ = mean (mu 'mew')
σ = standard deviation (sigma)

These are called the ***parameters*** of the model.

Every time you work a problem involving a Normal distribution, you should sketch the distribution and mark the mean +/- 1 standard deviation, like this:



(draw the 'ninja')

60.7% of the way up (a little more than halfway up is good for sketching).

Normal distribution model

The Normal distribution model is a ***probability density function***, which means:

- The area under the <u>whole</u> Normal curve is 1.
  (The sum of all probabilities/percentages = 100%)
- The area under the curve between two x or z values is the percentage of the data in this range (or the probability that one value selected at random will be in this range).
- Use calculator **normalcdf(**lower, upper, mean, std dev)
  (x or z data range ----> area/percentage/probability)
- Use calculator **invNorm(**area to the ***left***, mean, std dev)
  (area/percentage/probability ----> x or z data range)

(easiest to see how this works by looking at examples)...

Example: The scores on a test are Normally distributed, with a mean of 85 and a standard deviation of 7.

a) What percentage of scores are between 70 and 86?

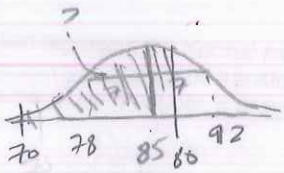
b) What percentage of scores are between 78 and 92?

c) What percentage of scores are below 88?

d) What score would 30% of the class score below?

e) What score is required to be in the top 1%?

a)

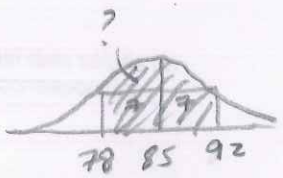


normalcdf(70, 86, 85, 7)
lower bound, upper bound, μ, σ

= .5407

54.13

b)

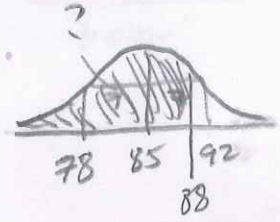


normalcdf(78, 92, 85, 7)
LB, UB, μ, σ

= .6827

68.38

c)

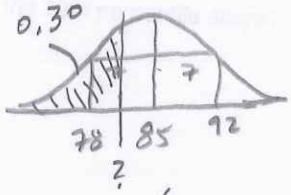


normalcdf(−999, 88, 85, 7)

= .6659

66.62

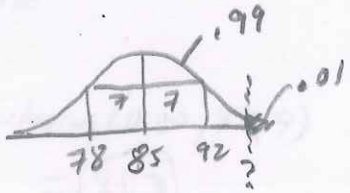
d)



invNorm(.30, 85, 7)
left area, μ, σ

= 81.3 (81)

e)

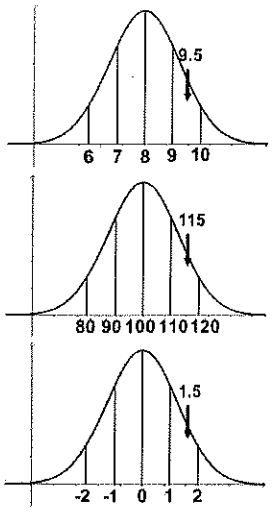


invNorm(.99, 85, 7)
(left area)

= 101.3 (101)

Different data set normal curves have different means and standard deviations, but the same shape.

It's useful to standardize by converting values to numbers of standard deviations from the mean.



$$Z = \frac{9.5-8}{1} = 1.5$$

$$Z = \frac{115-100}{10} = 1.5$$

Z-score

$$Z = \frac{x-\mu}{\sigma}$$

x = *original data*

μ = *mean of the data distribution*

σ = *standard deviation*

A Z-score is the number of standard deviations away from the mean this data is.

Positive z-score = above the mean.
Negative z-score = below the mean.

Comparing results from different datasets

Scores for college-bound students on the SAT and ACT tests are unimodal and symmetrical with the following means and standard deviations:

SAT: $\bar{x}$ = 1500, s = 250 **ACT**: $\bar{x}$ = 20.8, s = 4.8

Which of the following students has a better score?

Student 1: SAT score of 2030 **Student 2**: ACT score of 32

Both seem substantially above the mean. We can calculate z-scores to see which is more standard deviations above the mean.

student

$z = \frac{2030-1500}{250}$

$z = 2.12$ std dev above the mean

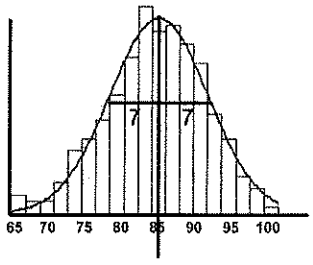
student 2

$z = \frac{32-20.8}{4.8}$

$z = 2.33$ std dev above the mean

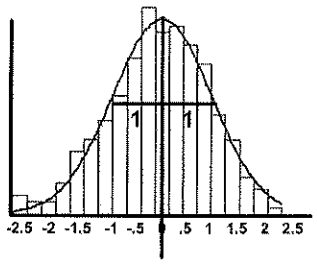
"better"

We can represent the Normal model data values using the original data values (x) or we can standardize by transforming all the x data values into corresponding z-scores. This produces a **Standard Normal Distribution.**



85

Normal Distribution
(x values)
N(85, 7)

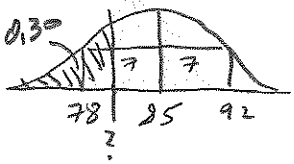


0

Standard Normal Distribution
(z values)
N(0, 1)

Example: The scores on a test are Normally distributed, with a mean of 85 and a standard deviation of 7.

a) What is the 30th percentile score?



invNorm(0.30, 85, 7)
= 81.3

b) Find the z-scores for tests of 78, 85, and 92.

$\left(z = \frac{x-\mu}{\sigma}\right)$

$z_{78} = \frac{78-85}{7} = \frac{-7}{7} = -1\sigma$

$z_{85} = \frac{85-85}{7} = \frac{0}{7} = 0$

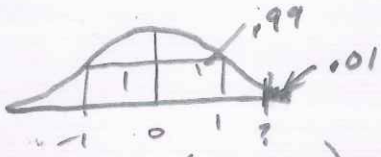
$z_{92} = \frac{92-85}{7} = \frac{7}{7} = +1\sigma$

Z score so use $\mu=0$ $\sigma=1$

c) If a score is 1.5 standard deviations below the mean, what percentage of scores are below this score? $z=-1.5$

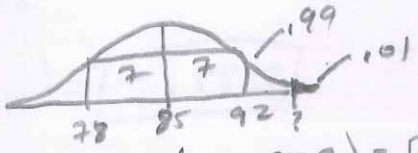
normalcdf(−99, −1.5, 0, 1)
= .0668 [6.68]

d) How many standard deviations do you need to be above the mean to be in the top 10%?



invNorm (.99, 0, 1) = [+2.32 σ]

e) What score on the test would put you in the top 10%?



invNorm (.99, 85, 7) = [101.3]

Example: A data set is Normally distributed.

a) What percentage of the data is within 1 standard deviation of the mean?

b) What percentage of the data is within 2 standard deviations of the mean?

c) What percentage of the data is within 3 standard deviations of the mean?

a)

normalcdf(−1, 1, 0, 1) = .6827 [68%]

b)

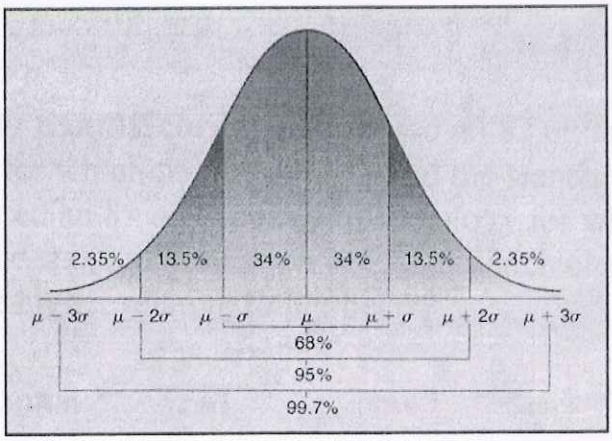
normalcdf(−2, 2, 0, 1)
= .9545 [95%]

c)

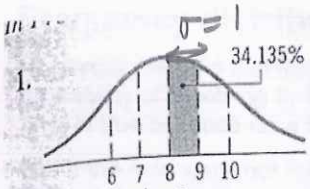
normalcdf(−3, 3, 0, 1)
= .9973 [99.73]

Normal distribution model

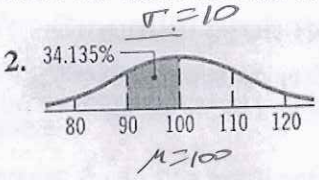
The 68-95-99.7 Rule... *(memorize the 68-95-99.7 % values)*



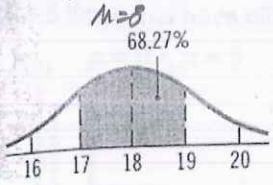
#1-4 Determine μ and σ by inspection for each normal curve.

1. 

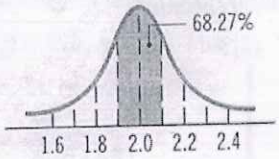
μ = 8

2. 

μ = 100

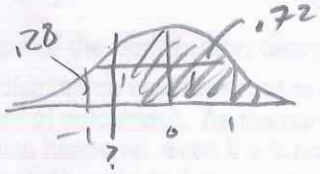
3. 

μ = 18
σ = 1

4. 

μ = 2.0
σ = 0.1

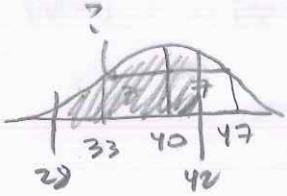
Find a z-score so that 72% of the population is to the right of z.



invNorm(.28, 0, 1) = −.5828

#16 **Life expectancy of clothing** If the average life of a certain make of clothing is 40 months with a standard deviation of 7 months, what percentage of these clothes can be expected to last from 28 months to 42 months? Assume that clothing lifetime follows a normal distribution.

μ = 40 mo
σ = 7 mo



normalcdf(LB, UB, μ, σ)
normalcdf(28, 42, 40, 7)
= .5692
56.9%

Comparing Exam Scores Beth scored an 82 on the final exam in biology for which the mean is 73 and the standard deviation is 9. She scored an 89 on the final in sociology for which the mean is 81 and the standard deviation is 15. In which class was Beth's exam score higher relative to her peers?

$z_{bio} = \frac{82-73}{9}$

$= 1\sigma$

$z_{socio} = \frac{89-81}{15}$

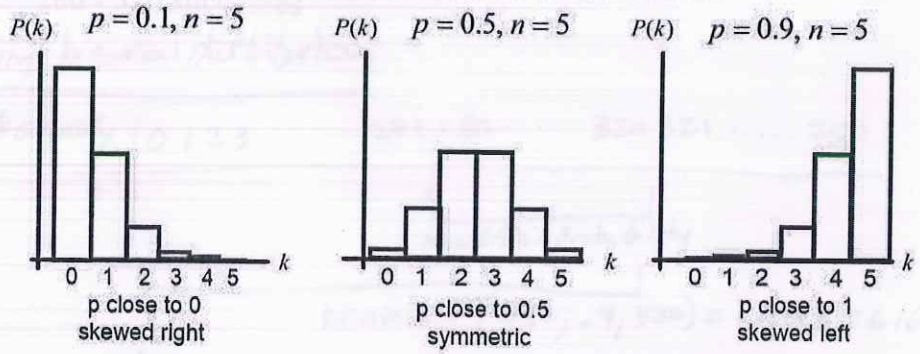
$= 0.67\sigma$

better

Frequency distribution for Binomial probabilities / Bernoulli Trials

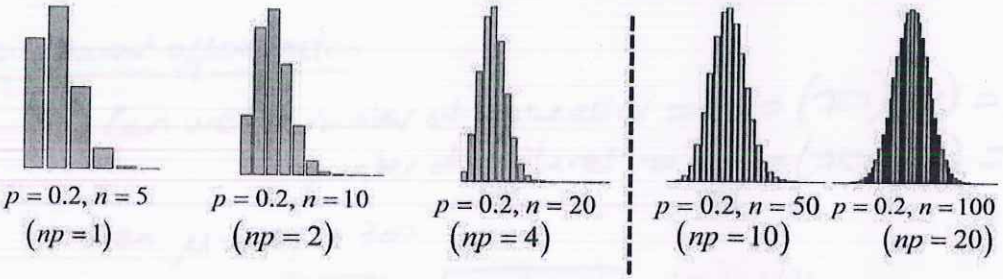
Yesterday, we said that flipping a coin 10 times and determining the probability of receiving 0, 1, 2, etc. heads forms a Normal distribution. That is true because for a fair coin p=0.5.

But if the coin were not fair, the Binomial distribution would not be normal: imagine flipping coins 5 times that have different probabilities of landing on heads:



So the shape of the distribution depends upon the probability of success, *p*.

The Binomial distribution also turns out to depend upon another factor: the sample size (n, the total number of outcomes). As the number of outcomes goes up, a strange and unexpected thing happens: even if p is not close to .5, and the distribution is skewed, the distribution becomes more and more symmetrical as n increases, and begins to resemble a Normal shape:



When $np \geq 10$, *and* $nq \geq 10$ the distribution can be approximated by a Normal distribution.

A Binomial distribution can be approximated by a Normal distribution if:

- p (probability of success) is close to 0.5

  -or-

- n is large (at least 10 'successes' and at least 10 'failures', np>10 and nq>10)

For Binomial probabilities that can be approximated by Normal distributions:

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

where $n = number\ of\ trials$

$p = probability\ of\ success$

$q = probability\ of\ failure\ (q = 1 - p)$

24. Suppose a binomial experiment consists of 750 trials and the probability of success for each trial is 0.4. Approximate the probability of obtaining the number of successes indicated by using a normal curve.

280 - 320 successes

Using binomial distribution

# successes x	0 1 2 ... 279	280 ... 320	321 ... 750
P		reach this probability	

binomcdf(750, .4, 320) = .9363616796

binomcdf(750, .4, 279) = .0628137219

$= .9363616796 - .0628137219$

$= \boxed{.8735479577}$

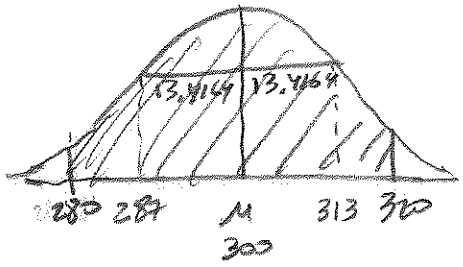
Using Normal approximation

Can we? number of successes $= np = (750)(.4) = 300 \geq 10$ ✓

number of failures $= nq = (750)(.6) = 450 \geq 10$ ✓ Yes

then $\mu = np = 300$

$\sigma = \sqrt{npq} = \sqrt{750(.4)(.6)} = 13.4164$

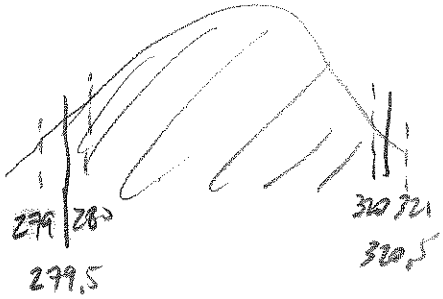


P = normalcdf(280, 320, 300, 13.4164)

$= \boxed{.86396}$

(a little low, but close)

Some argue that we should include the part of the distribution between 279–280 and 320–321:



P = normalcdf(279.5, 320.5, 300, 13.4164)

$= \boxed{.8734826265}$

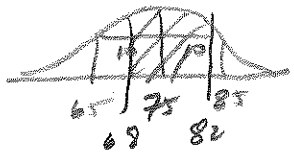
$.873547$

binomial answer

(off accurate to 3 decimal places)

#20 In Math 135 the average final grade is 75.0 and the standard deviation is 10.0. The professor's grade distribution shows that 15 students with grades from 68.0 to 82.0 received Cs. Assuming the grades follow a normal distribution, how many students are in Math 135?

$\mu = 75.0$
$\sigma = 10.0$



normalcdf(68, 82, 75, 10)
= .516

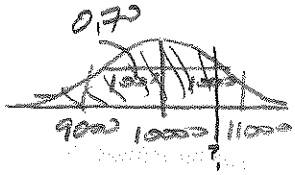
So 51.6% of students should have this grade range. Then:

$\frac{15}{n} = .51607\ldots$

$n = \frac{15}{.51607\ldots} = 29.066$ about 29 students

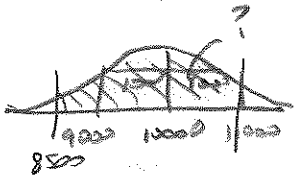
#18 **Movie Theater Attendance** The attendance over a weekly period of time at a movie theater is normally distributed with a mean of 10,000 and a standard deviation of 1000 persons. Find:

a. the number in the lowest 70% of the attendance figures



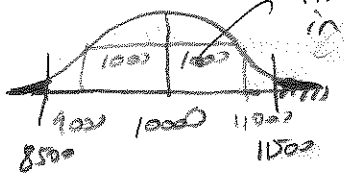
invNorm(0.70, 10000, 1000)
= 10524 persons

b. the percent of attendance figures that falls between 8500 and 11,000 persons.



normalcdf(8500, 11000, 10000, 1000)
= .77454
77.5%

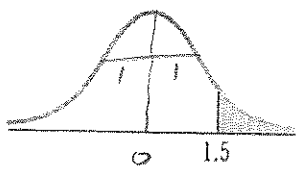
c. the percent of attendance figures that differs from the mean by 1500 persons or more.



find area in middle = normalcdf(8500, 11500, 10000, 1000)
= .866385...
shaded = 1 − .866385... = .13361
13.4%

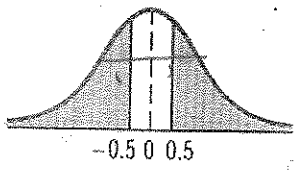
Find the area of the shaded region under the Normal curve:
(no μ, σ given, so assume these are z-scores)

11.



= normalcdf(1.5, 999, 0, 1)
= .0668

12.



middle = normalcdf(−0.5, 0.5, 0, 1)
= .38292...
shaded = 1 − .38292...
= .617075