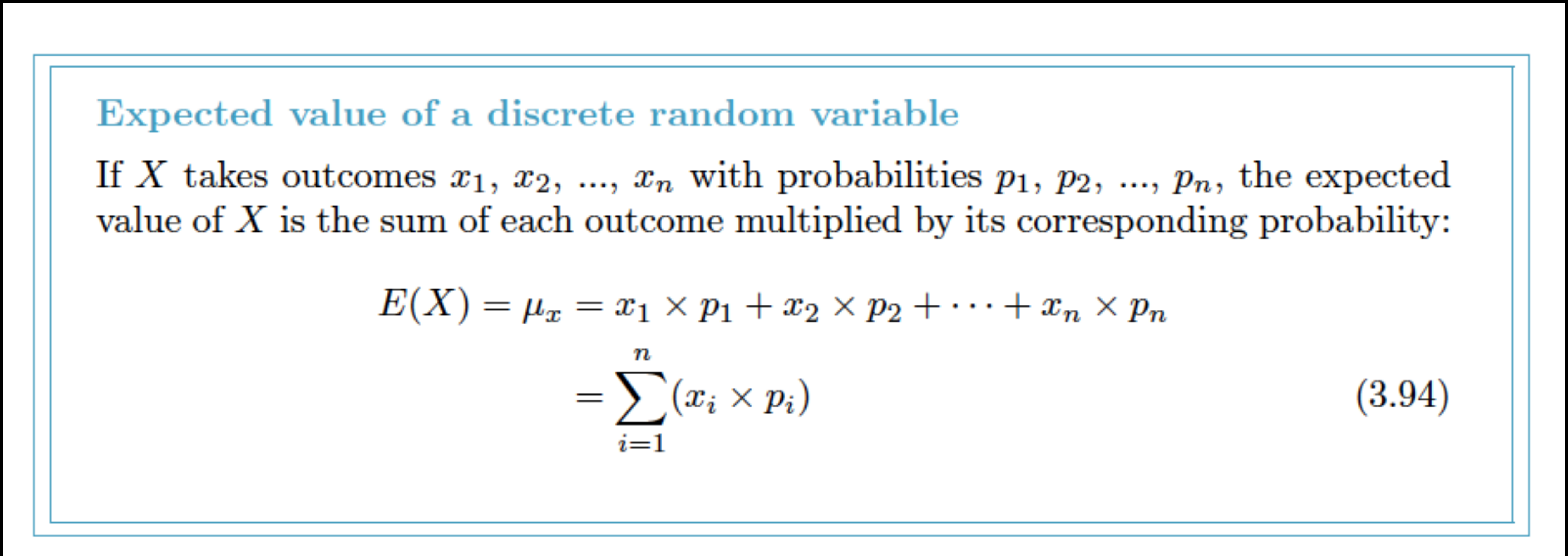
Welcome to Week #6!

Quick review from last time…

Expectation

"Expected" value of an average outcome is more heavily weighted to events with a higher probability of occurring.



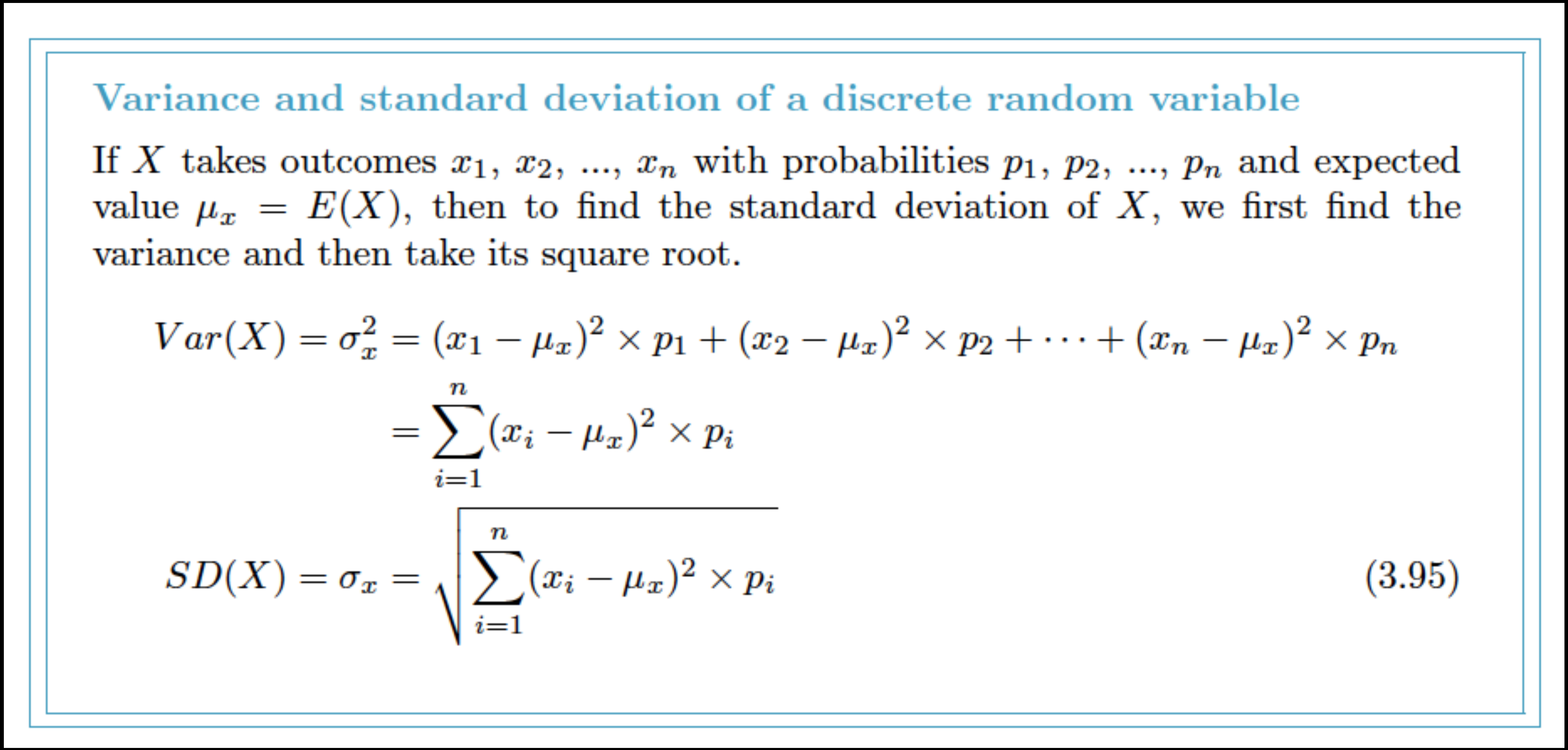
Expected value of a discrete random variable

If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$
\begin{aligned}
E(X) = \mu_x &= x_1 \times p_1 + x_2 \times p_2 + \cdots + x_n \times p_n \\
&= \sum_{i=1}^{n} (x_i \times p_i) \qquad (3.94)
\end{aligned}
$$

Variability

Contribution to the variance is greater from outcomes that have a larger probability of occurring.



Variance and standard deviation of a discrete random variable

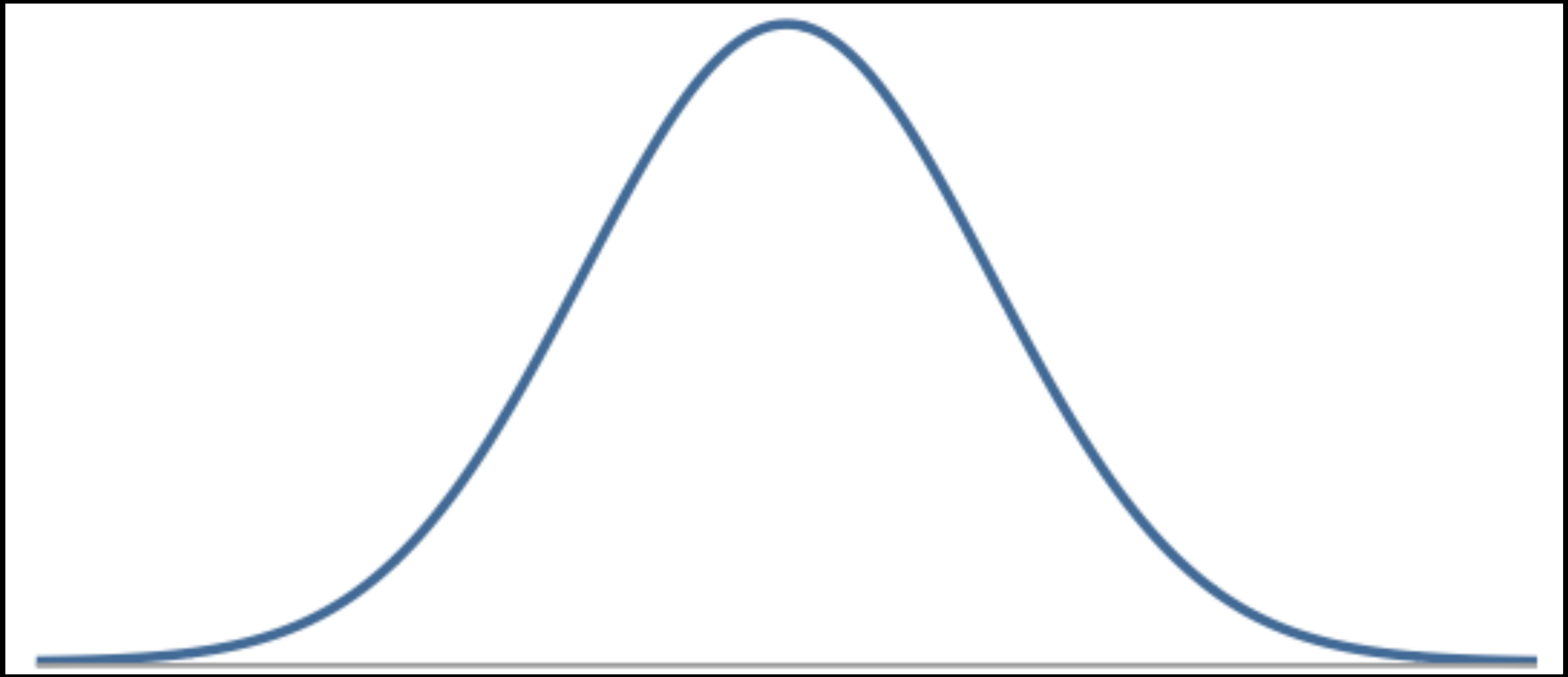
If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n and expected value $\mu_x = E(X)$, then to find the standard deviation of X, we first find the variance and then take its square root.

$$\begin{aligned} Var(X) = \sigma_x^2 &= (x_1 - \mu_x)^2 \times p_1 + (x_2 - \mu_x)^2 \times p_2 + \cdots + (x_n - \mu_x)^2 \times p_n \\ &= \sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i \end{aligned}$$

$$SD(X) = \sigma_x = \sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i} \qquad (3.95)$$

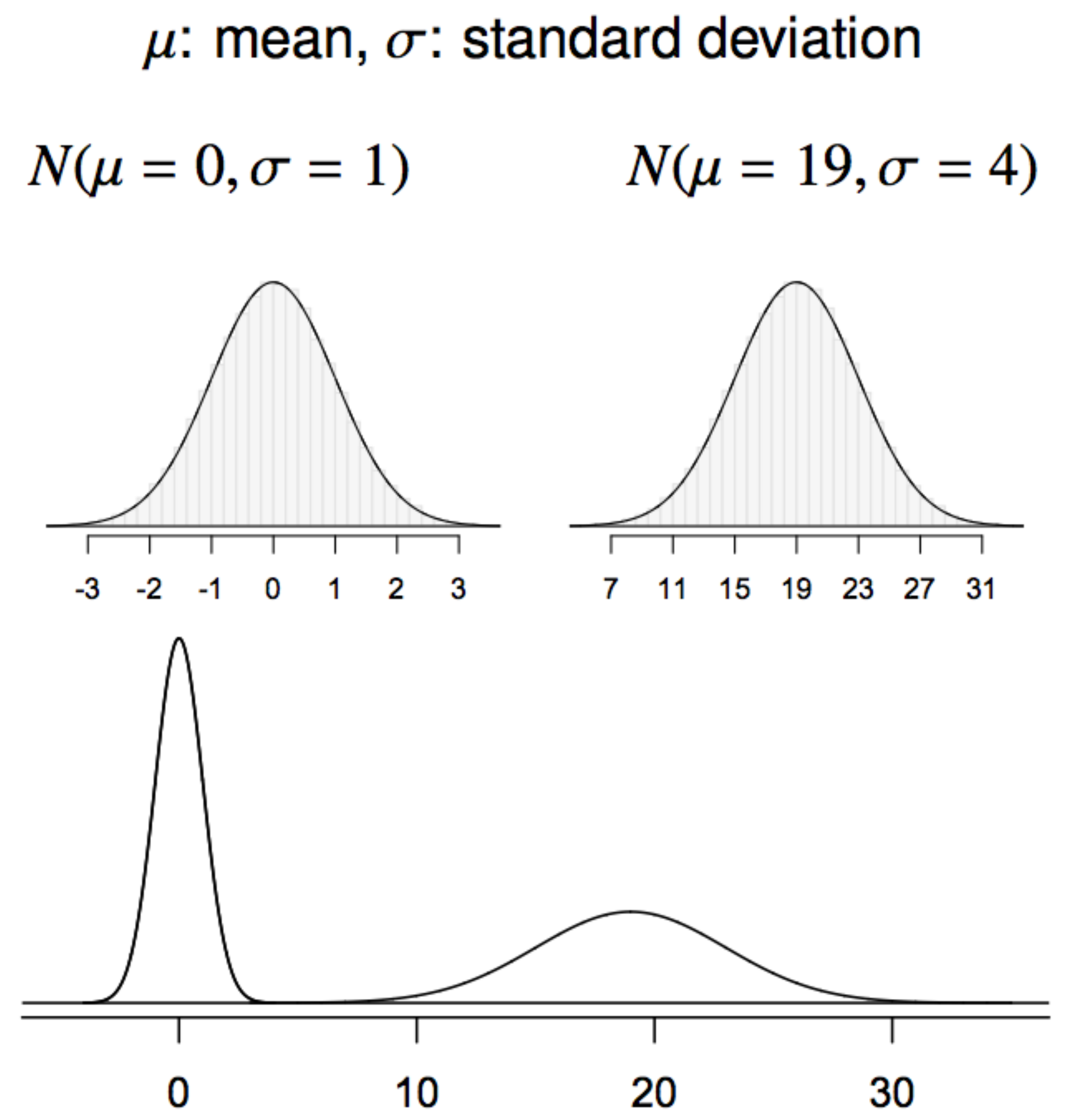
The Normal distribution

In Chapter 3, we look at the Normal distribution. The Normal distribution is the most famous continuous distribution.



To find areas under curves, we generally use a table or technology (i.e. calculator, stat program, etc.).

Normal distributions with different parameters

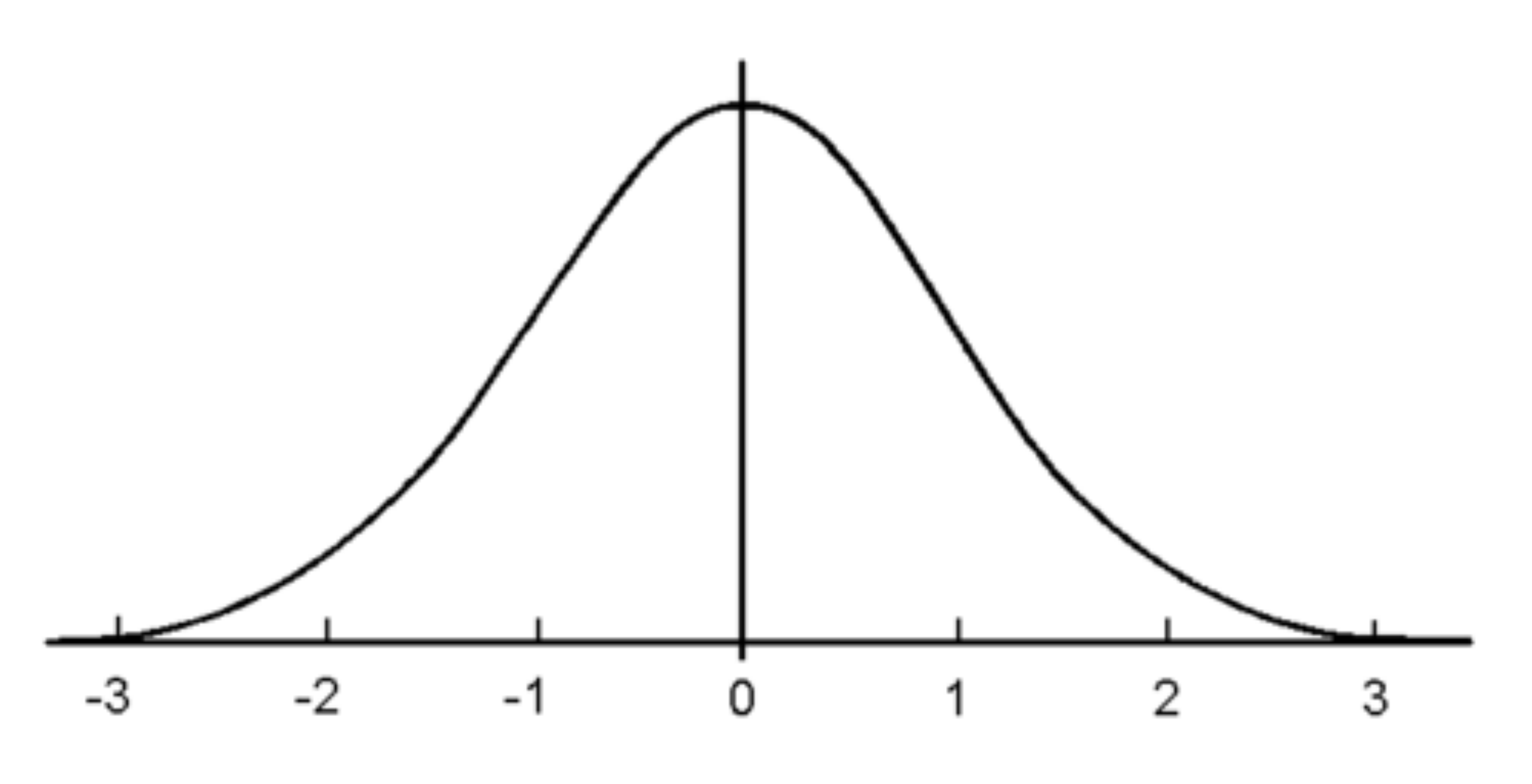


The Standard Normal Curve

$$Z = \frac{(\text{observation - mean})}{\text{SD}}$$

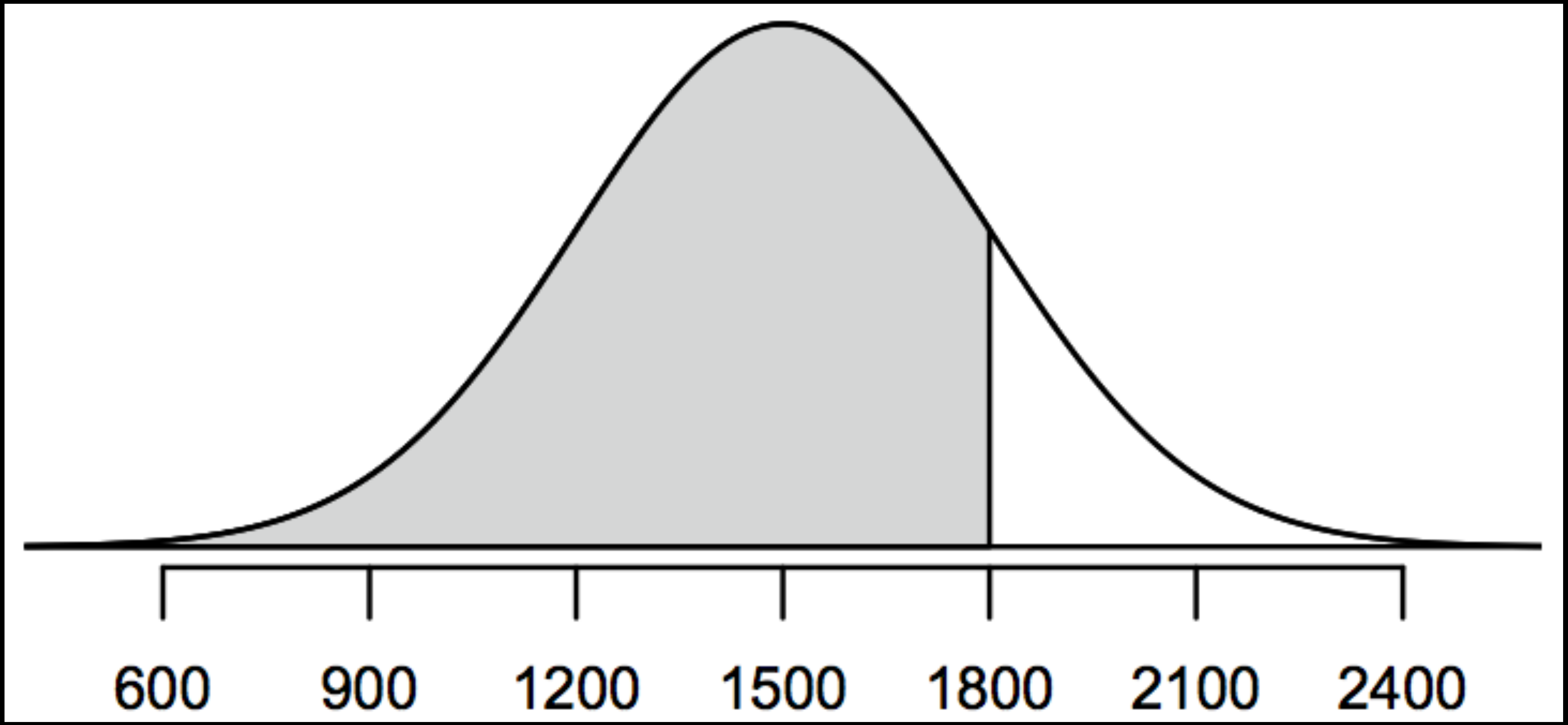
What units are on the horizontal axis?

Z-scores! ← A way to compare normal distributions



Percentiles

Percentile is the percentage of observations that fall below a given data point.
Graphically, percentile is the area below the probability distribution curve to the left of that observation.

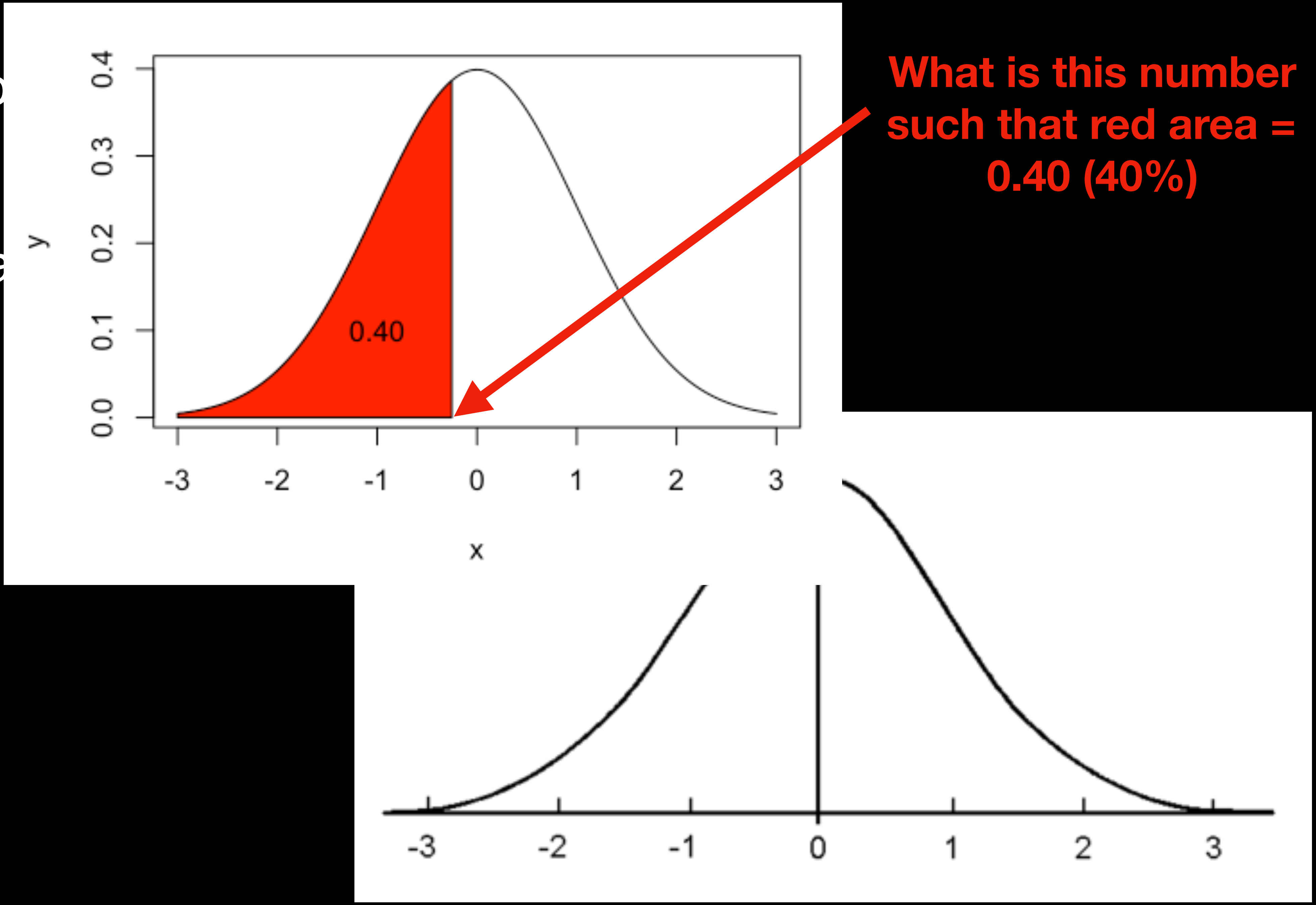


Finding percentiles from the standard normal curve

What Z-score corresponds to the 50th percentile?
i.e. P(Z < ?) = 0.5 Z =

What Z-score co
i.e. P(Z < ?) = 0.

What Z-score ha
i.e. P(Z < ?) = 0.



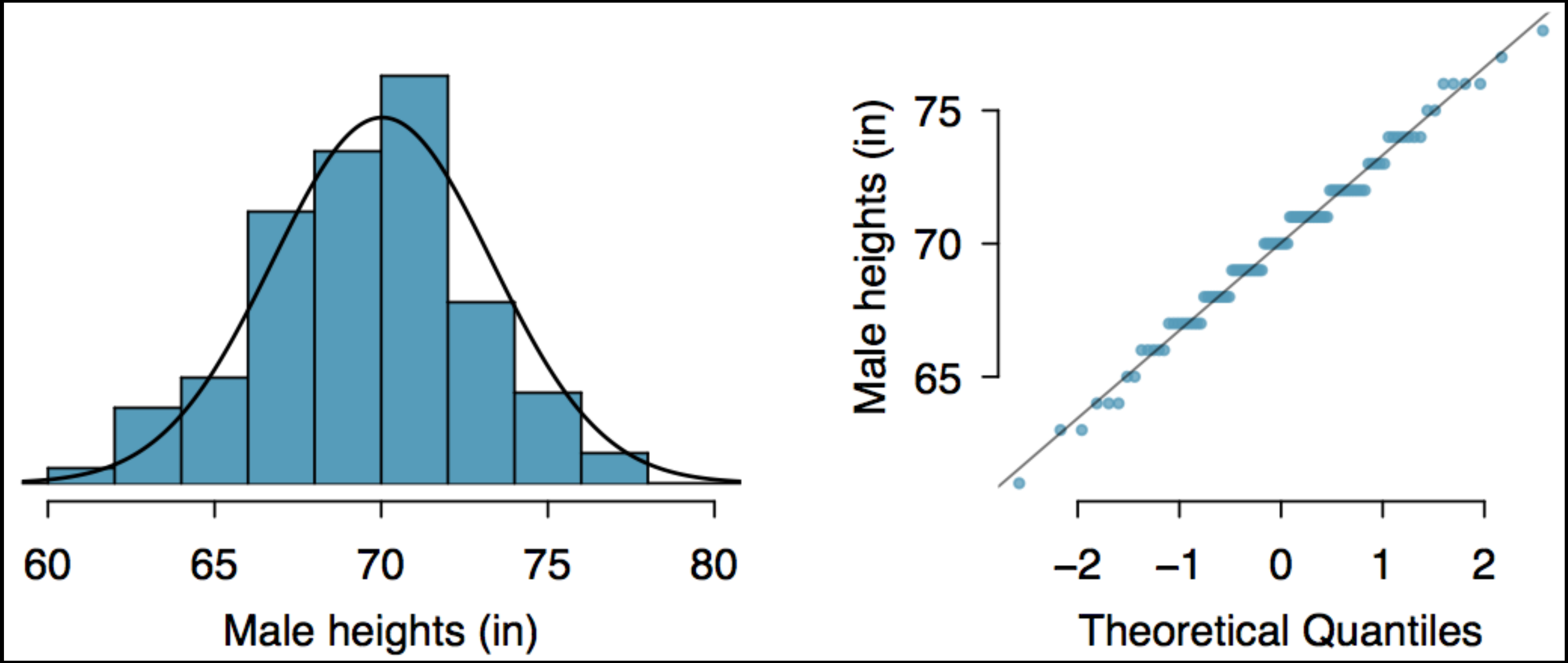
What is this number such that red area = 0.40 (40%)

dnorm
pnorm
qnorm

in R

Is it Normal? The Normal probability plot

A histogram and normal probability plot of a sample of 100 male heights.



A few more normal distribution examples

For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.

Step 1: Plot the distribution

Step 2: Plot the measurement in question

Step 3: Estimate

Step 4: Calculate

Step 5: Check answer with intuition

A few more normal distribution examples

Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test. Tom takes the test and scores 585. Will he be admitted to this university?

Step 1: Plot the distribution

Step 2: Plot the measurement in question

Step 3: Estimate

Step 4: Calculate

Step 5: Check answer with intuition

The Binomial formula

The Binomial formula

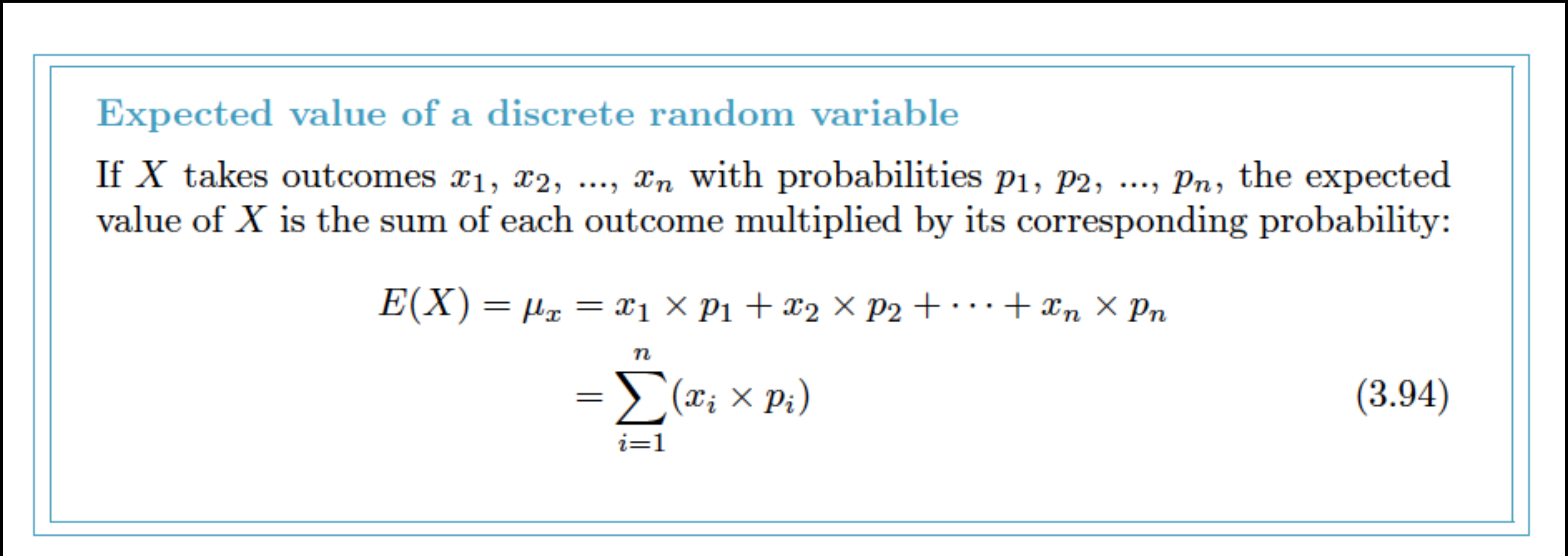
… before MATH, lets look at an example in R…

The Binomial formula: now some math

Plagiarized from OIS:

If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$\mu = E[X] = \mathrm{SUM}_i(\mathrm{Prob}_i \text{ x } p_i)$



Expected value of a discrete random variable

If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned} E(X) = \mu_x &= x_1 \times p_1 + x_2 \times p_2 + \cdots + x_n \times p_n \\ &= \sum_{i=1}^{n} (x_i \times p_i) \qquad (3.94) \end{aligned}$$

The Binomial formula: now some math

Plagiarized from OIS:

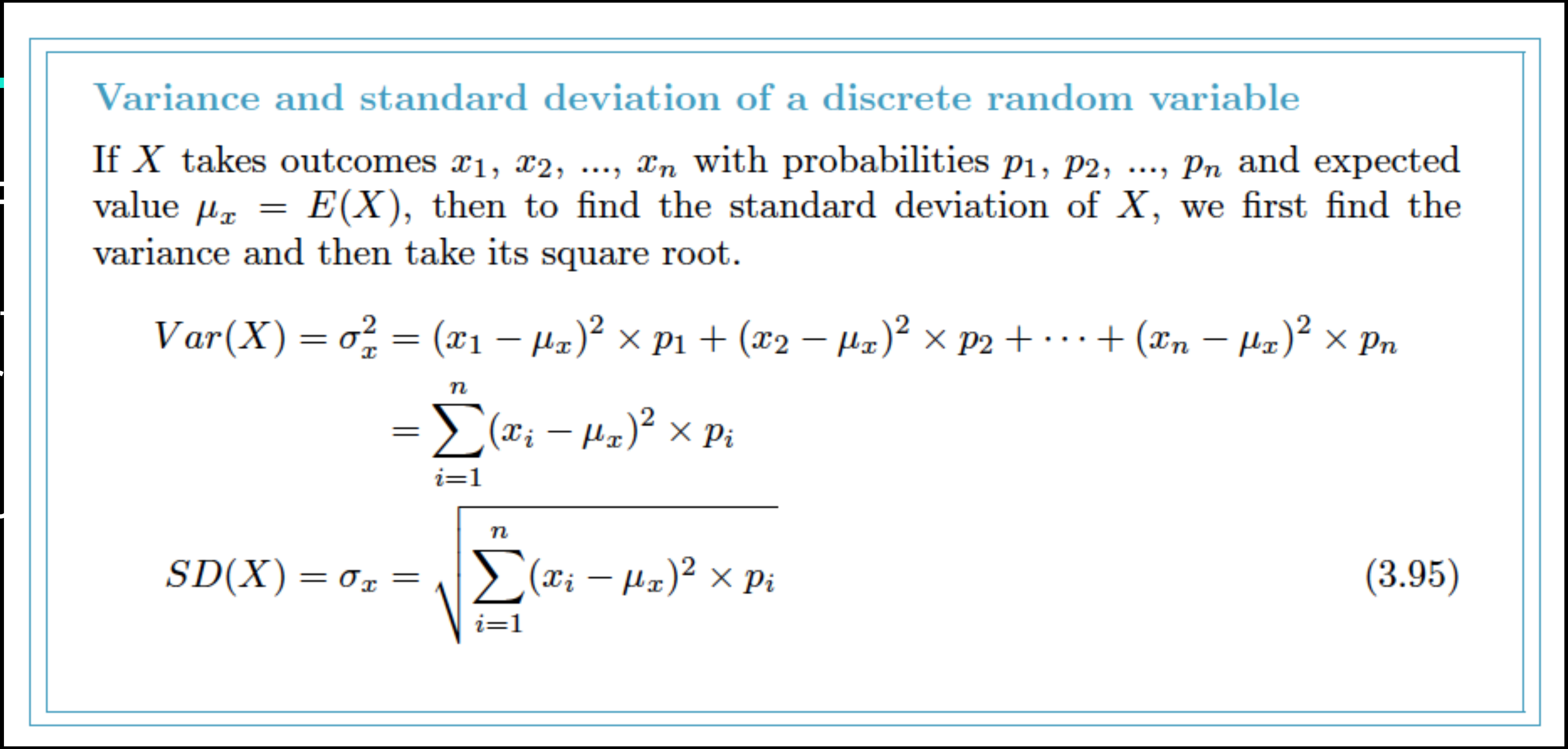
If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$
\begin{aligned}
\mu = E[X] &= SUM_i(Prob_i \times p_i) \\
&= P(X = 0) \times 0 + P(X = 1) \times 1 \\
&= (1 - p) \times 0 + p \times 1 \\
&= 0 + p = p
\end{aligned}
$$

p = prob_sample ~0.5

Similarly, the variance of X can be computed:

$$\sigma^2 = SUM_i(Prob_i \times Var_i)$$



Variance and standard deviation of a discrete random variable

If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n and expected value $\mu_x = E(X)$, then to find the standard deviation of X, we first find the variance and then take its square root.

$$\begin{aligned} Var(X) = \sigma_x^2 &= (x_1 - \mu_x)^2 \times p_1 + (x_2 - \mu_x)^2 \times p_2 + \cdots + (x_n - \mu_x)^2 \times p_n \\ &= \sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i \end{aligned}$$

$$SD(X) = \sigma_x = \sqrt{\sum_{i=1}^{n} (x_i - \mu_x)^2 \times p_i} \qquad (3.95)$$

Similarly, the variance of X can be computed:

σ^2 = $\text{SUM}_i(\text{Prob}_i \text{ x Var}_i)$

The Binomial formula: now some math

Plagiarized from OIS:

If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned}\mu = E[X] &= \mathrm{SUM}_i(\mathrm{Prob}_i \times p_i) \\ &= P(X = 0) \times 0 + P(X = 1) \times 1 \\ &= (1 - p) \times 0 + p \times 1 \\ &= 0 + p = p\end{aligned}$$

p = prob_sample ~0.5

Similarly, the variance of X can be computed:

$$\begin{aligned}\sigma^2 &= \mathrm{SUM}_i(\mathrm{Prob}_i \times \mathrm{Var}_i) \\ &= P(X = 0)(0 - p)^2 + P(X = 1)(1 - p)^2 \\ &= (1-p)p^2+p(1-p)^2 \\ &= p(1-p)\end{aligned}$$

The standard deviation is = $\sqrt{p(1 - p)}$

The Binomial formula: now some math

Lets say we want to know what is the probability of getting our first success on the nth trial?

We've had n-1 failures:

P(failure & failure &...) for n-1 times

Laws of probability dictate we multiply:

P(failure & failure &...) = $P(\text{failure})^{n-1}$

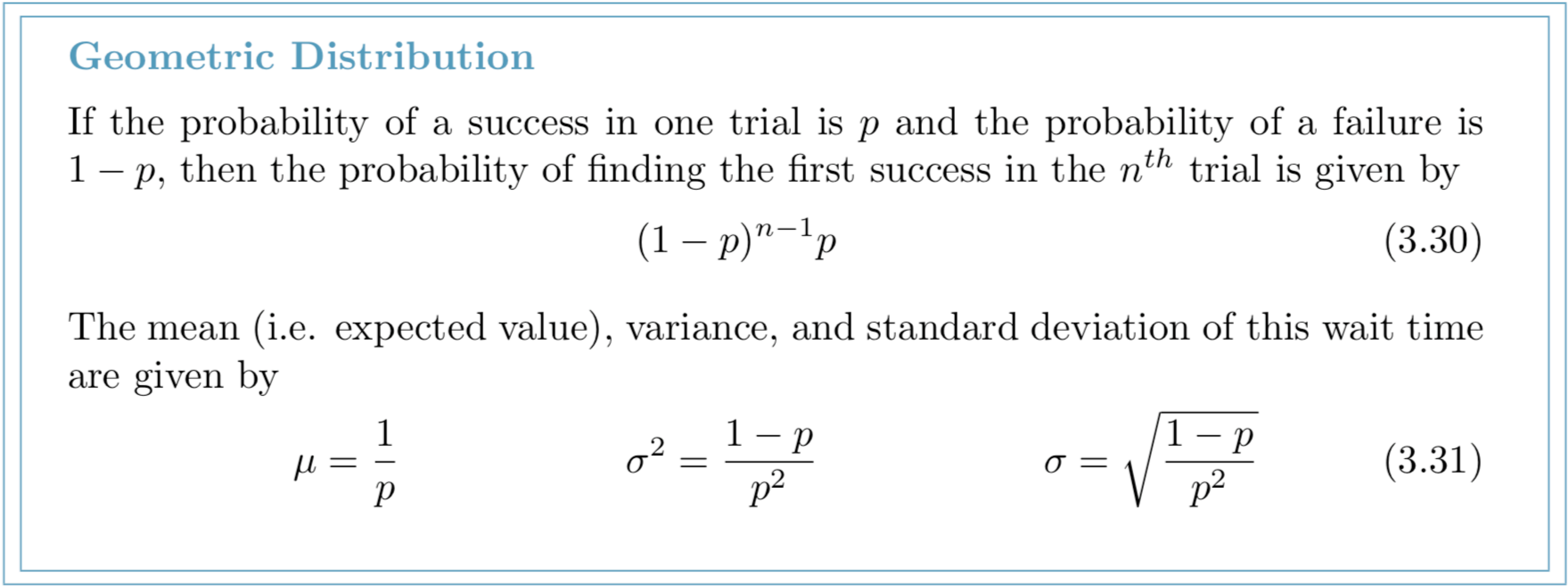
Then we have a success, so that is P(success):

P(success on nth try) = $P(\text{failure})^{n-1}$ X P(success)

To be consistent with OIS, lets define the probability of success as p:

P(success on nth try) = $(1-p)^{n-1}$ X p

The Binomial formula: From the Geometric Distribution



Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n^{th} trial is given by

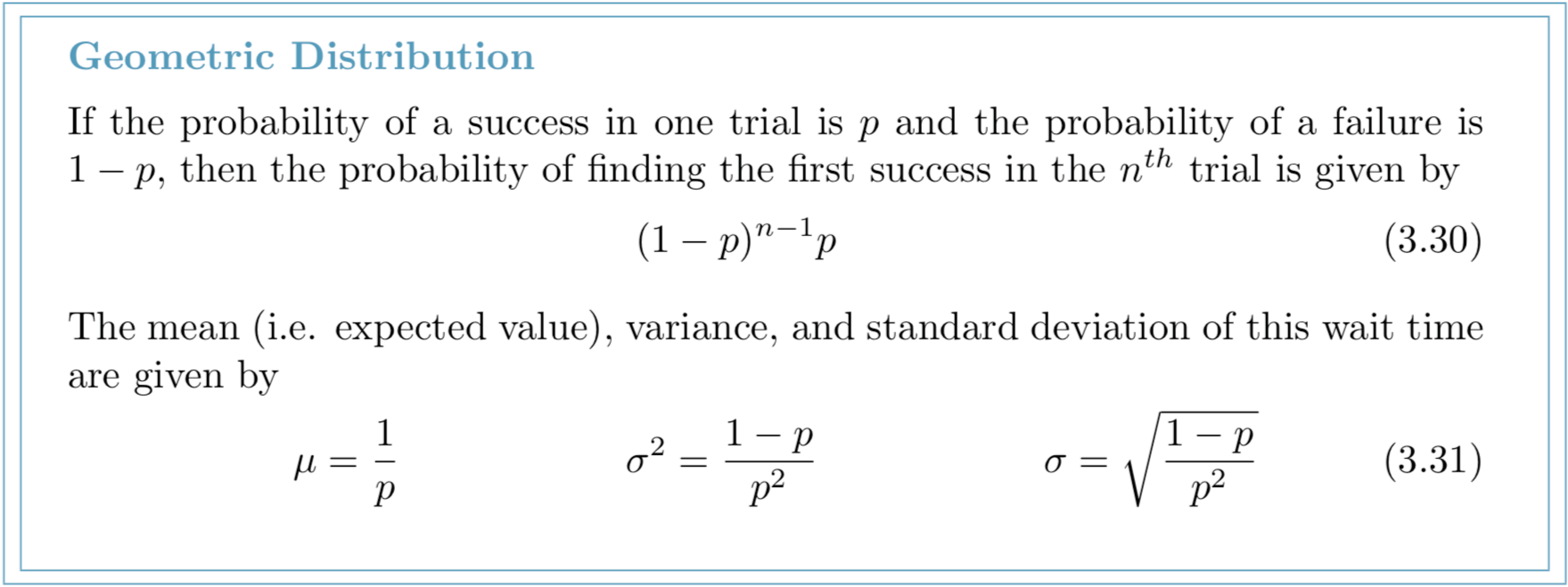
$$(1-p)^{n-1}p \tag{3.30}$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.31}$$

Example in R!

The Binomial formula: From the Geometric Distribution



Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n^{th} trial is given by

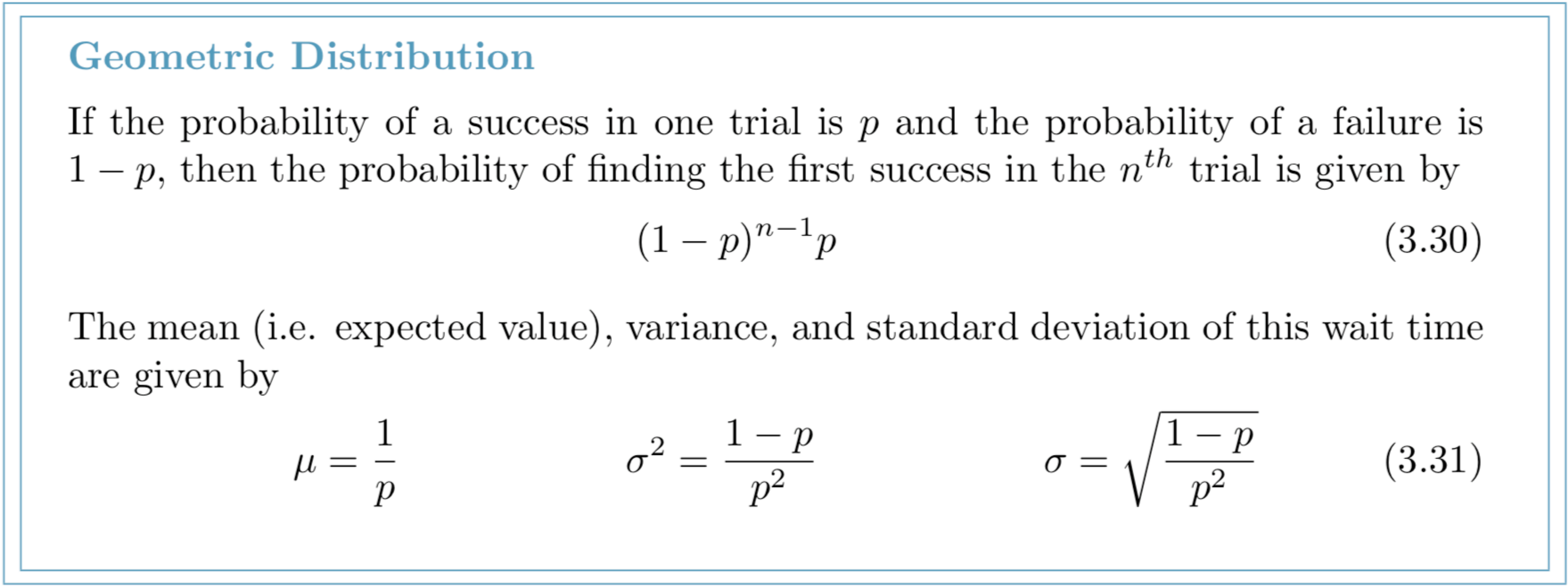
$$(1-p)^{n-1}p \tag{3.30}$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.31}$$

But what about the more general case of getting a certain number, “k”, of successes in “n” trials?

The Binomial formula: From the Geometric Distribution



Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p \tag{3.30}$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu=\frac{1}{p} \qquad \sigma^2=\frac{1-p}{p^2} \qquad \sigma=\sqrt{\frac{1-p}{p^2}} \tag{3.31}$$

But what about the more general case of getting a certain number, “k”, of successes in “n” trials?

Binomial distribution!

The Binomial formula: Factorials

n! = factorial(n)

(note: 0! = 1)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$ = choose(n,k)

$$\binom{3}{2} = \frac{3*2*1}{(2*1)*(3-2)}$$

“3 choose 2”

$$\binom{7}{3} = \frac{7*6*5*4*3*2*1}{(3*2*1)*[(7-3)*(7-3-1)*(7-3-2)*(7-3-3)]}$$

“7 choose 3”

The Binomial distribution: An example

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k\ (1-p)^{(n-k)}$$

e.g. If the probability of a severe lung condition for a smoker = 0.3, what is the distribution of number of cases of severe lung condition among 4 randomly chosen friends who smoke?

Find the probabilities where k = 0, 1, 2, 3, 4 using the binomial formula for each value of k. Note that n and p are fixed.

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k\ (1 - p)^{(n-k)}$$

e.g. If the probability of a severe lung condition for a smoker = 0.3, what is the distribution of number of cases of severe lung condition among 4 randomly chosen friends who smoke?

$$P(k\ \textit{successes in } n\ \textit{trials}) = \binom{n}{k} p^k\,(1-p)^{(n-k)}$$

e.g. If the probability of a severe lung condition for a smoker = 0.3, what is the distribution of number of cases of severe lung condition among 4 randomly chosen friends who smoke?

$$P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^{k} \, (1-p)^{(n-k)}$$

e.g. If the probability of a severe lung condition for a smoker = 0.3, what is the distribution of number of cases of severe lung condition among 4 randomly chosen friends who smoke?

$$P(k \textit{ successes in } n \textit{ trials}) = \binom{n}{k} p^k \, (1-p)^{(n-\boxed{k})}$$

Varies between 0 & n=4

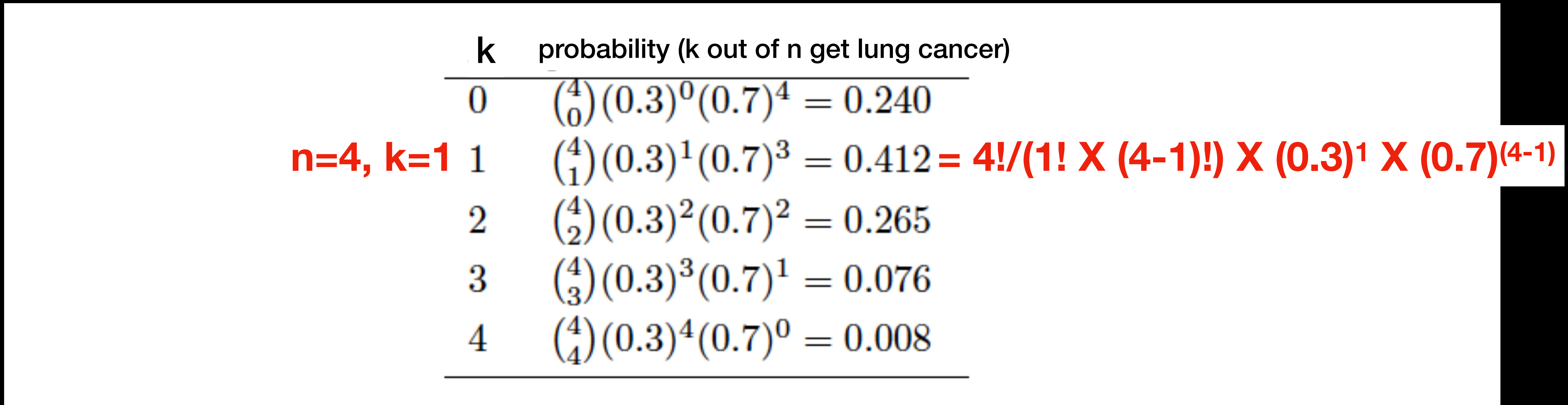
e.g. If the probability of a severe lung condition for a smoker = 0.3, what is the distribution of number of cases of severe lung condition among 4 randomly chosen friends who smoke?

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k \, (1-p)^{(n-k)}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

e.g. If the probability of a severe lung condition for a smoker = 0.3, what is the distribution of number of cases of severe lung condition among 4 randomly chosen friends who smoke?

Find the probabilities where k = 0, 1, 2, 3, 4 using the binomial formula for each value of k.
The entire *distribution* is defined below. Note that, correcting for rounding error, the probabilities must add
to 1.

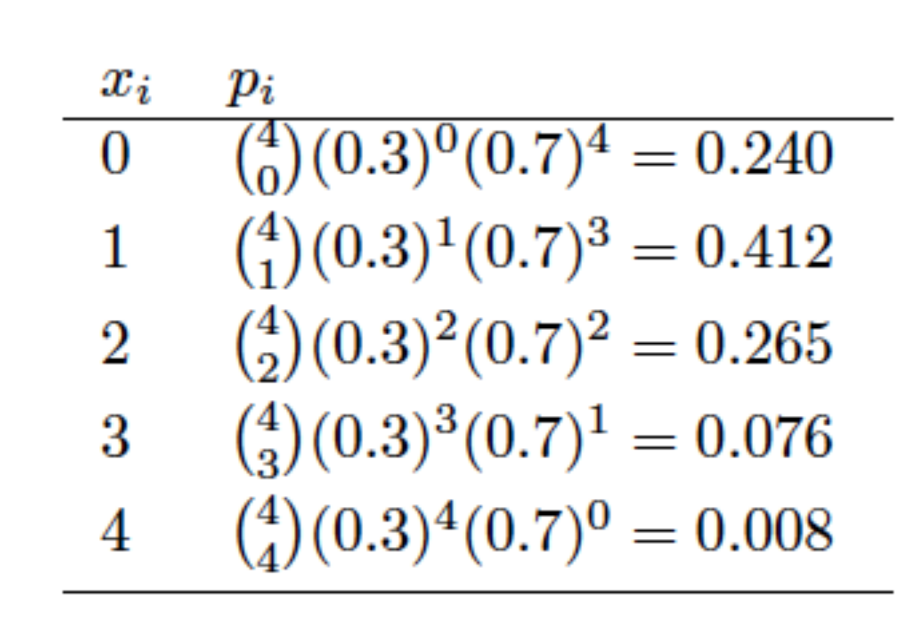


	k	probability (k out of n get lung cancer)	
	0	$\binom{4}{0}(0.3)^0(0.7)^4 = 0.240$	
n=4, k=1	1	$\binom{4}{1}(0.3)^1(0.7)^3 = 0.412$	= 4!/(1! X (4-1)!) X $(0.3)^1$ X $(0.7)^{(4-1)}$
	2	$\binom{4}{2}(0.3)^2(0.7)^2 = 0.265$	
	3	$\binom{4}{3}(0.3)^3(0.7)^1 = 0.076$	
	4	$\binom{4}{4}(0.3)^4(0.7)^0 = 0.008$	

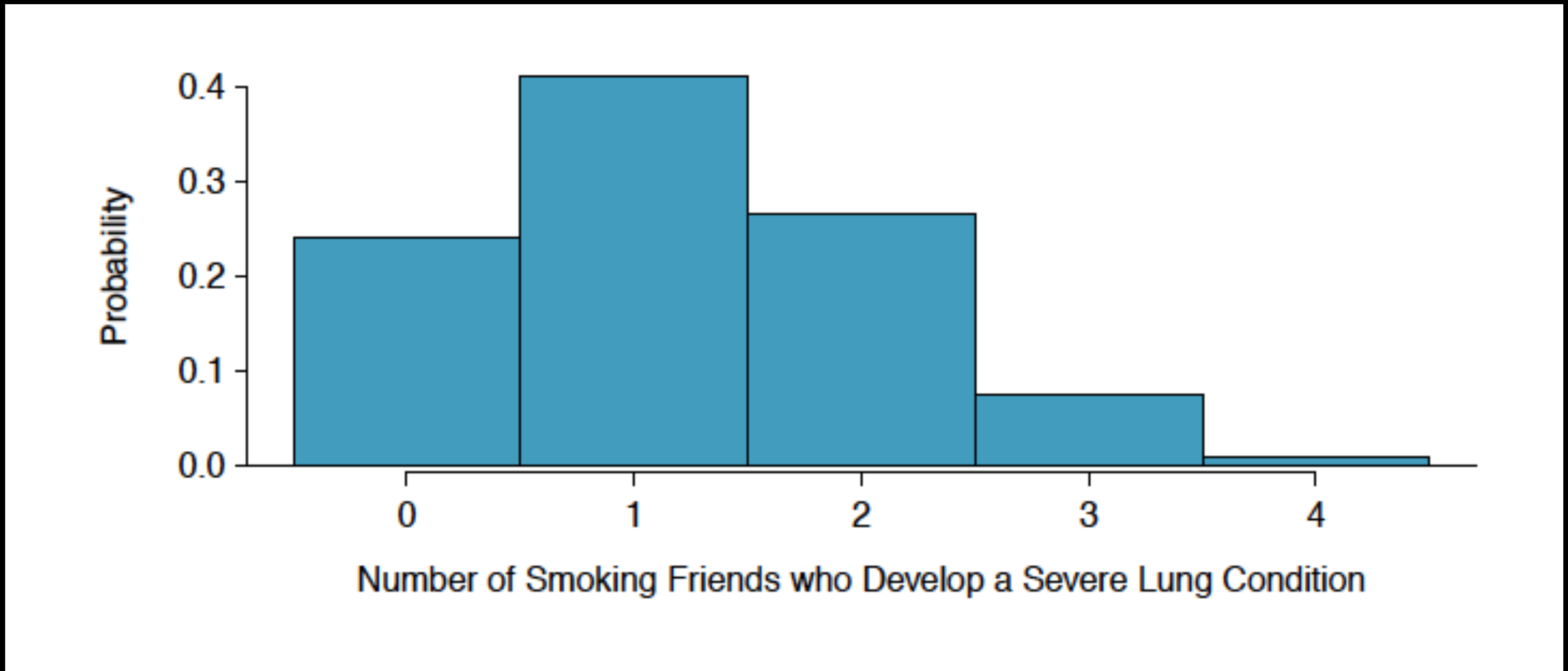
The Binomial distribution (cont.)

Once the probabilities of each value are calculated using the binomial formula, a probability histogram can be drawn in order to visualize the distribution. Like any distribution, the binomial distribution has a mean and a standard deviation.

Example of rolling a dice in R!



x_i	p_i
0	$\binom{4}{0}(0.3)^0(0.7)^4 = 0.240$
1	$\binom{4}{1}(0.3)^1(0.7)^3 = 0.412$
2	$\binom{4}{2}(0.3)^2(0.7)^2 = 0.265$
3	$\binom{4}{3}(0.3)^3(0.7)^1 = 0.076$
4	$\binom{4}{4}(0.3)^4(0.7)^0 = 0.008$



“k” successes

The Binomial distribution (cont.)

Recall the formulas from the previous chapter for calculating mean and standard deviation of a probability distribution.

$$E(X) = \mu_X = \sum x_i p_i$$

$$Var(X) = \sigma_X^2 = \sum (x_i - \mu_x)^2 p_i$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

Fortunately, for the binomial distribution with parameters n and p, there exist short-cut formulas for finding the mean and standard deviation.

Mean or Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 100 Americans, how many would you expect to be obese?

Easy enough, 100 x 0.262 = 26.2.

Or more formally, $\mu = np$ = 100 x 0.262 = 26.2.

But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Mean and Standard deviation of a binomial distribution

Mean $\mu = np$ $\sigma = \sqrt{np(1-p)}$ Standard Deviation

Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

Note: Mean and standard deviation of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Unusual observations

Using the notion that observations that are more than 2 standard deviations away from the mean are considered unusual and the mean and the standard deviation we just computed, we can calculate a range for the plausible number of obese Americans in random samples of 100.

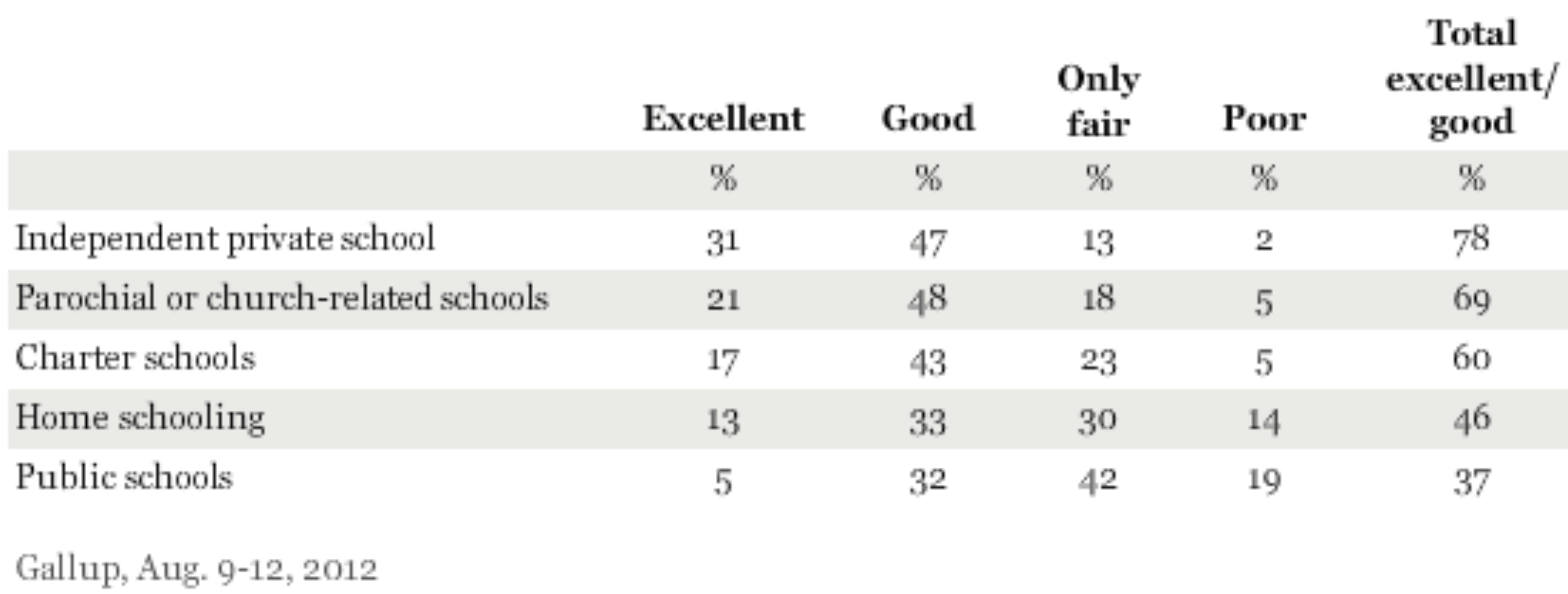
$$26.2 \pm (2 \times 4.4) \rightarrow (17.4, 35.0)$$

Practice

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) Yes (b) No

Hint: what is the mean? What is the SD?



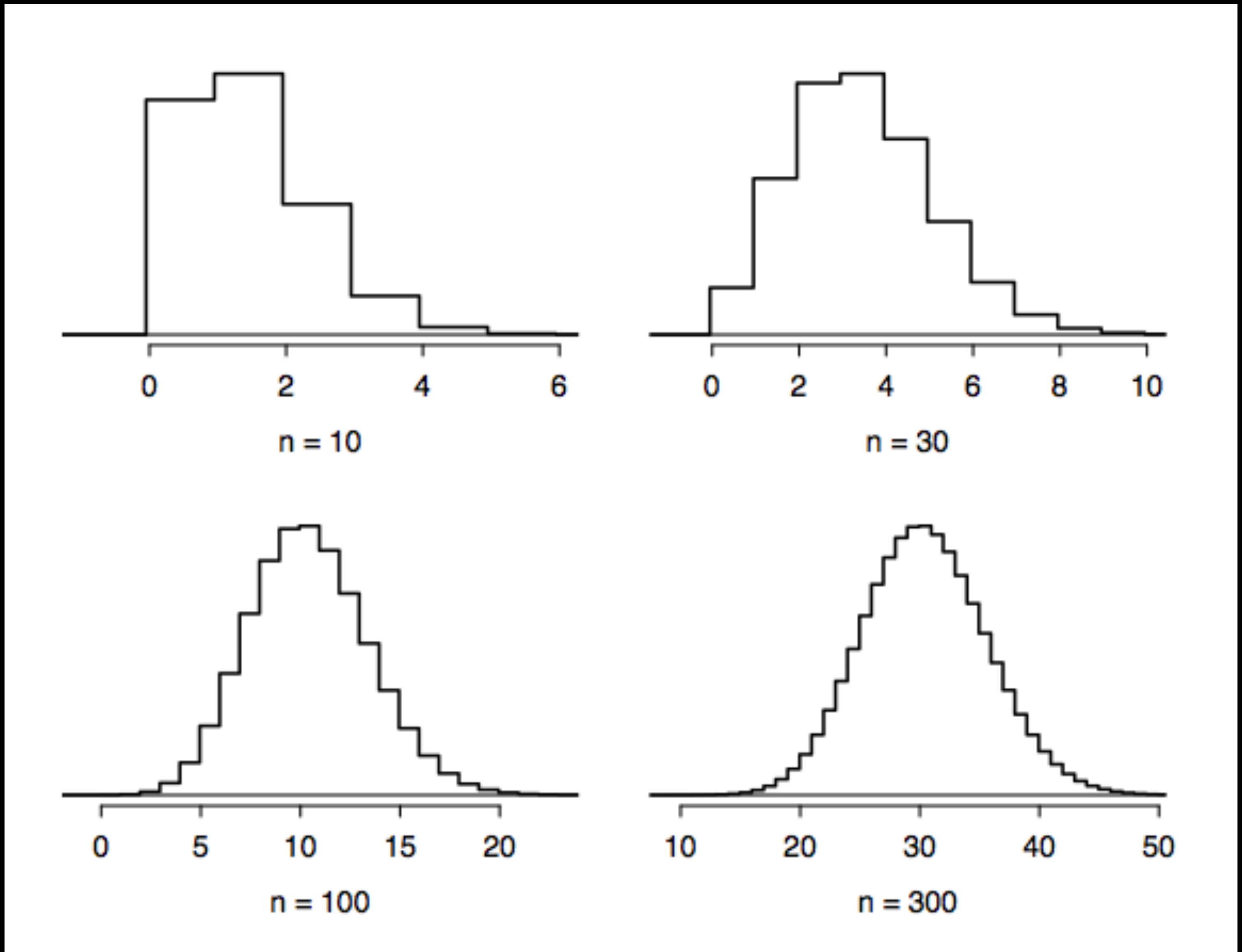
	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

Gallup, Aug. 9-12, 2012

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

Distributions of number of successes

Hollow histograms of samples from the binomial model
where **p = 0.10** and n = 10, 30, 100, and 300.
What happens as n increases?



See this applet with sliders for n and p to see how shape binomial distribution changes as n and p change:

http://www.stat.berkeley.edu/~stark/Java/Html/BinHist.htm

Note: the scales on the histograms are different!

Binomial to normal

A study found that approximately 25% of Facebook users are considered power users (i.e. they submit much more content than the average user).

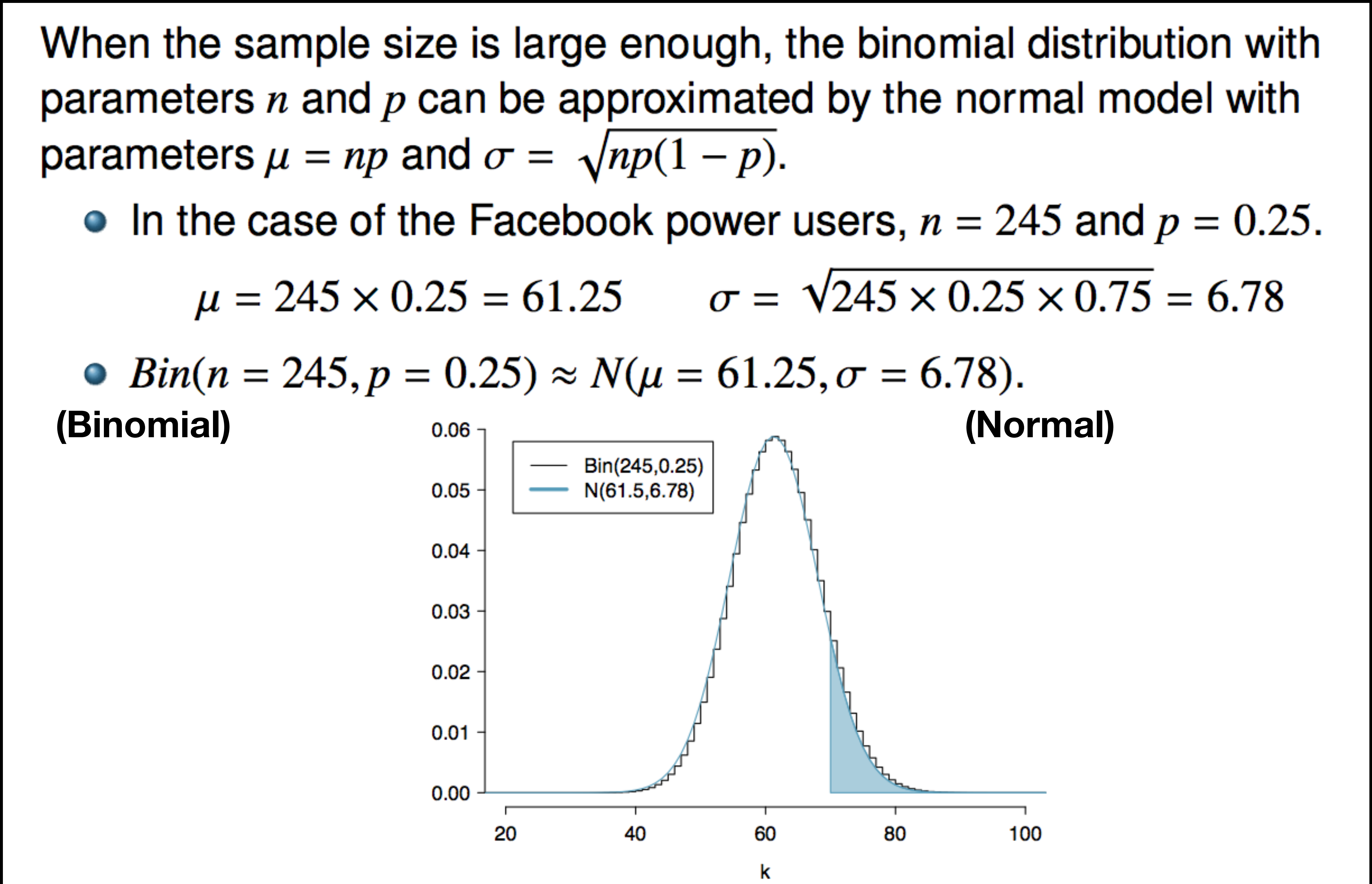
The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$\begin{aligned} P(X \geq 70) &= P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \ldots \text{ or } K = 245) \\ &= P(K = 70) + P(K = 71) + P(K = 72) + \ldots + P(K = 245) \end{aligned}$$

This seems like an awful lot of work...

Normal approximation to the binomial



When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.

(Binomial)

(Normal)

Bin(245,0.25)
N(61.5,6.78)
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
k

Normal approximation to the binomial

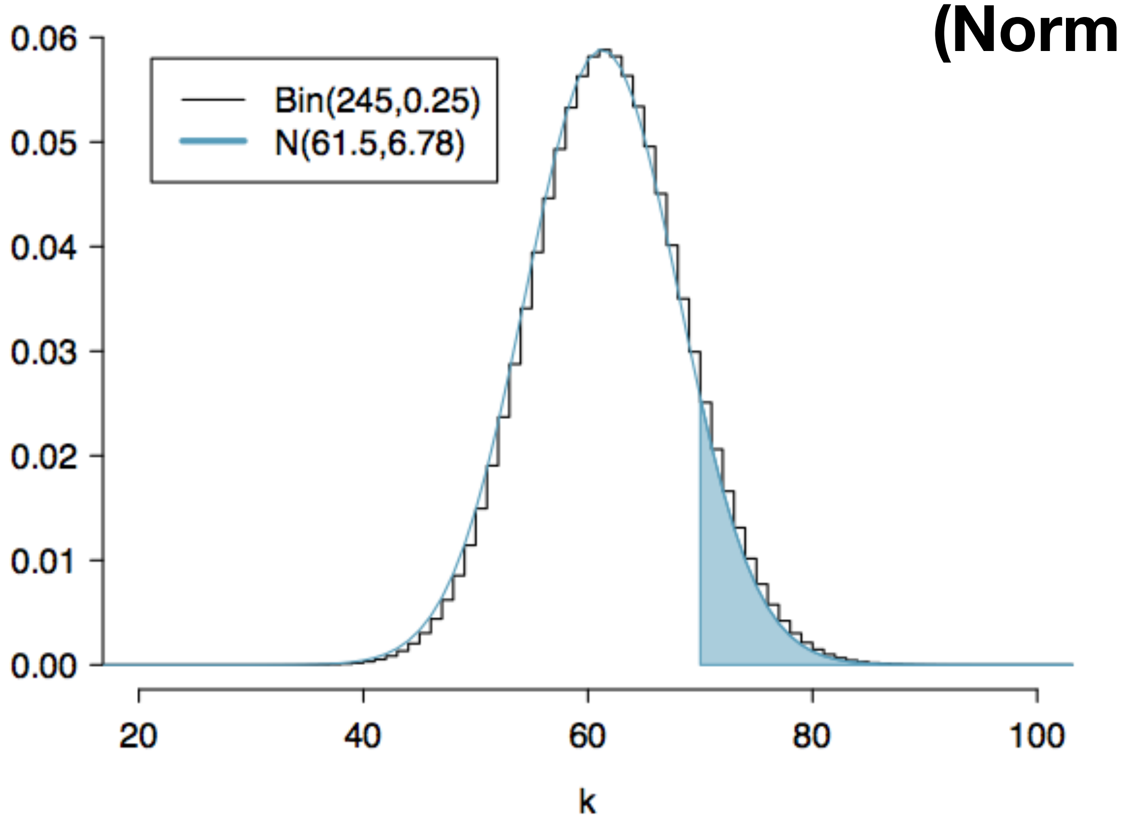
When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

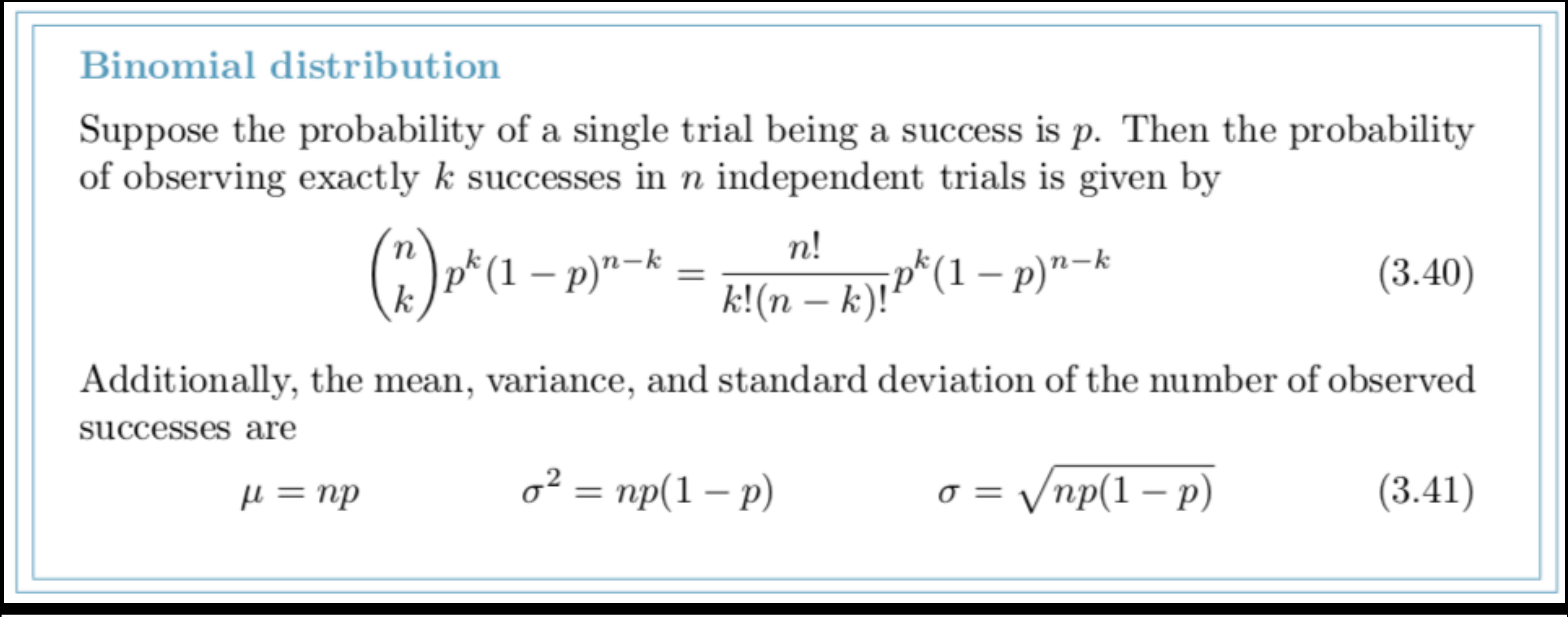
$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.

(Binomial) **(Normal)**



Slight Tangent: The Negative Binomial distribution



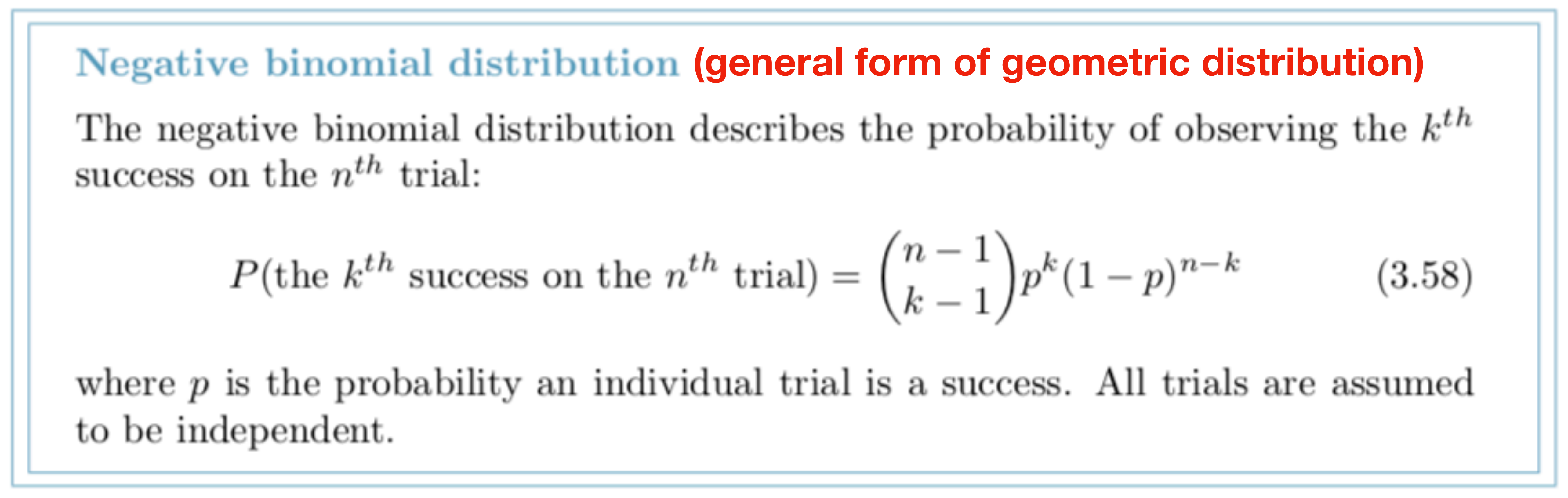
Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \qquad (3.40)$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)} \qquad (3.41)$$



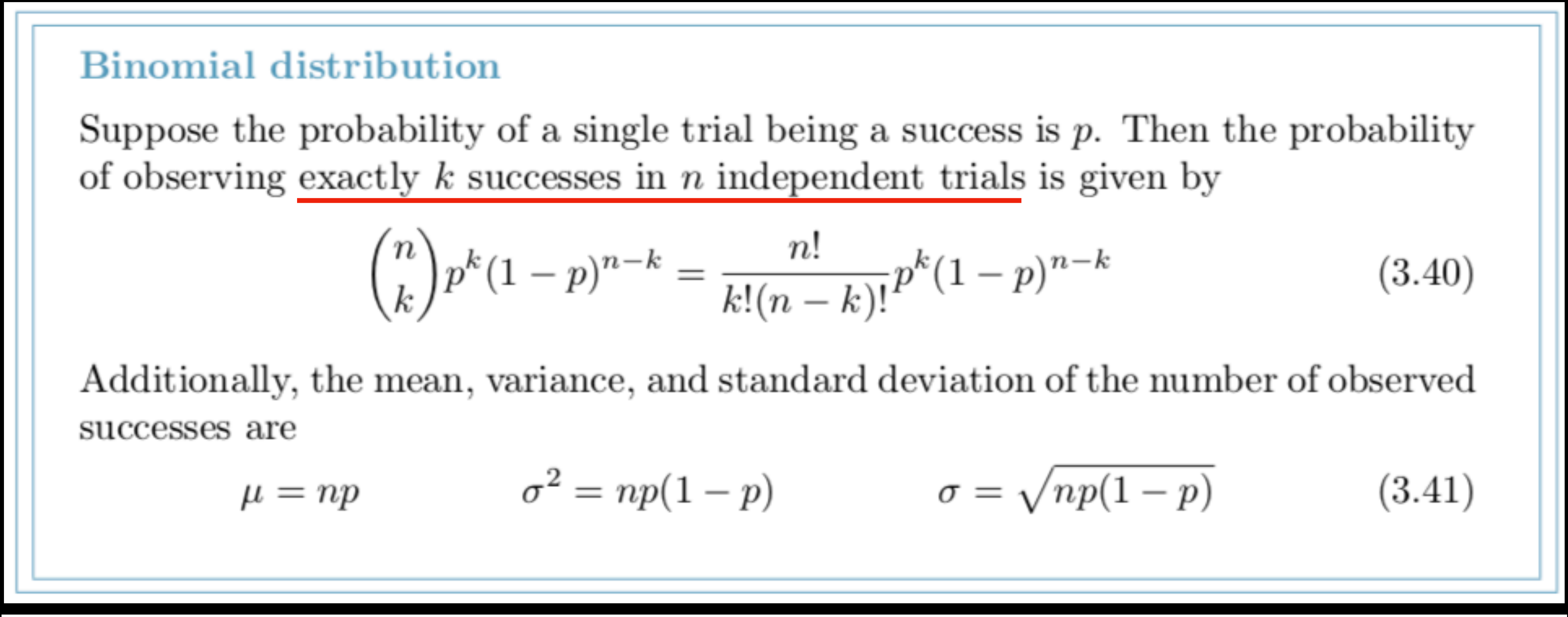
Negative binomial distribution (general form of geometric distribution)

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \qquad (3.58)$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

Slight Tangent: The Negative Binomial distribution



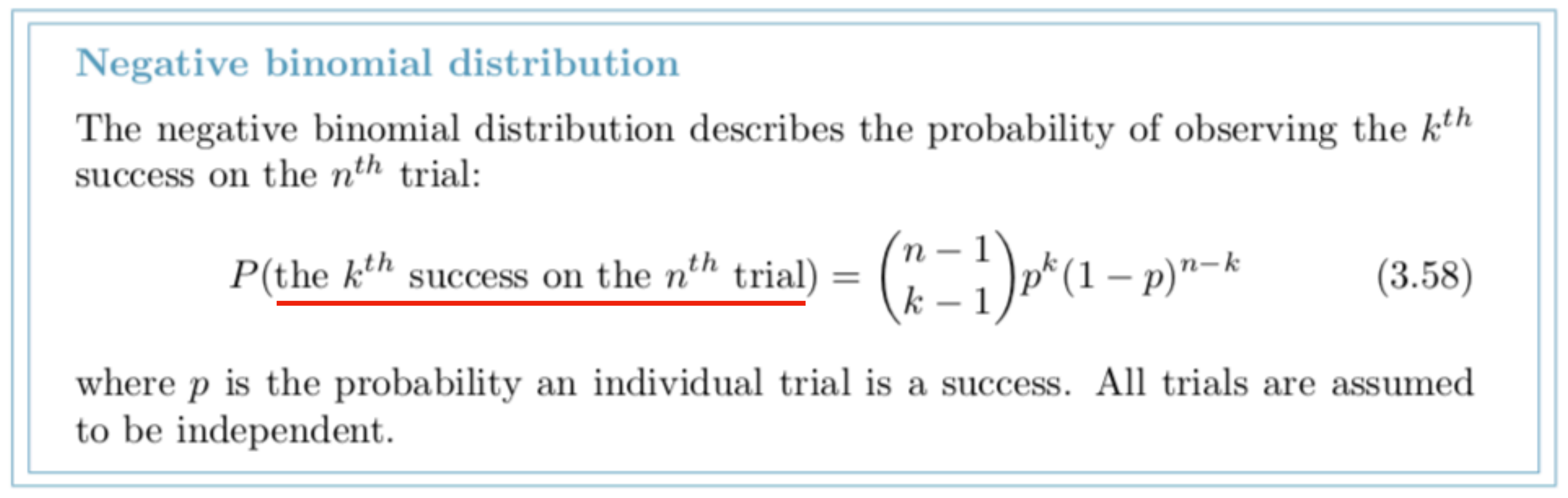
Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{3.40}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)} \tag{3.41}$$



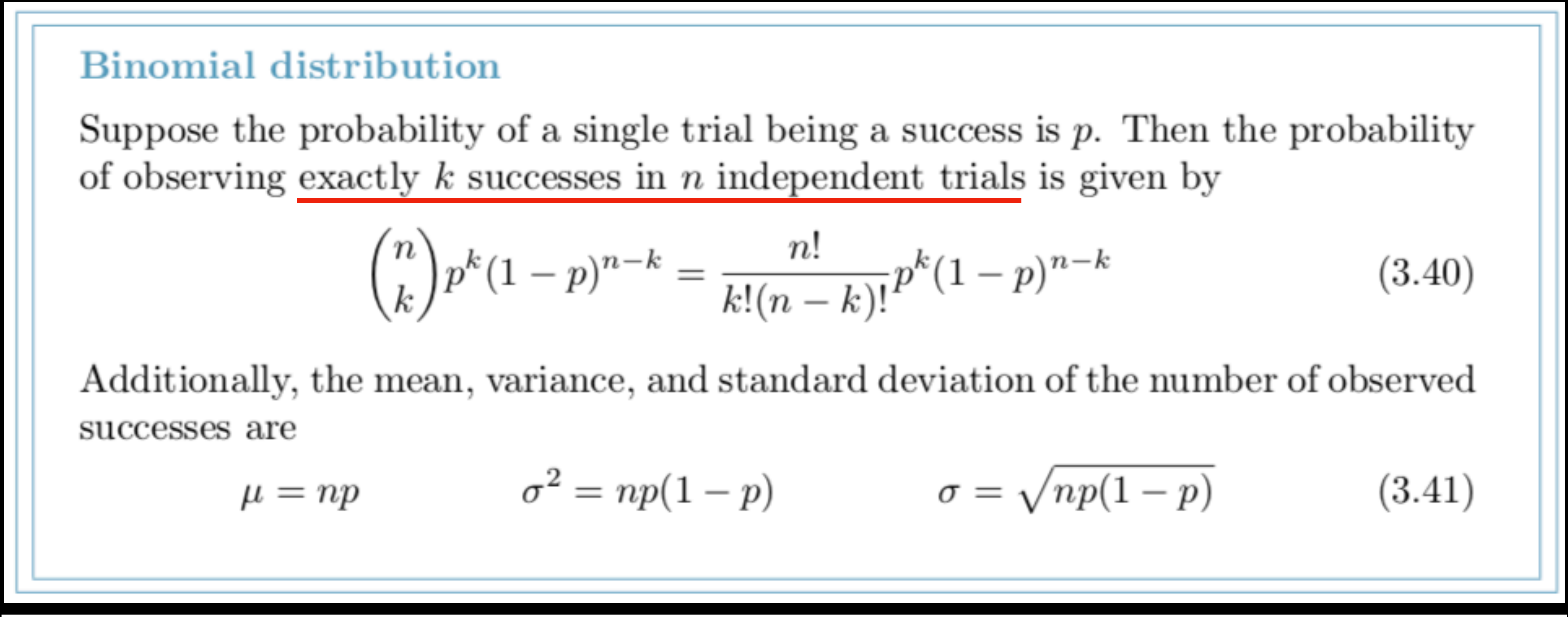
Negative binomial distribution

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$P(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \tag{3.58}$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

Slight Tangent: The Negative Binomial distribution



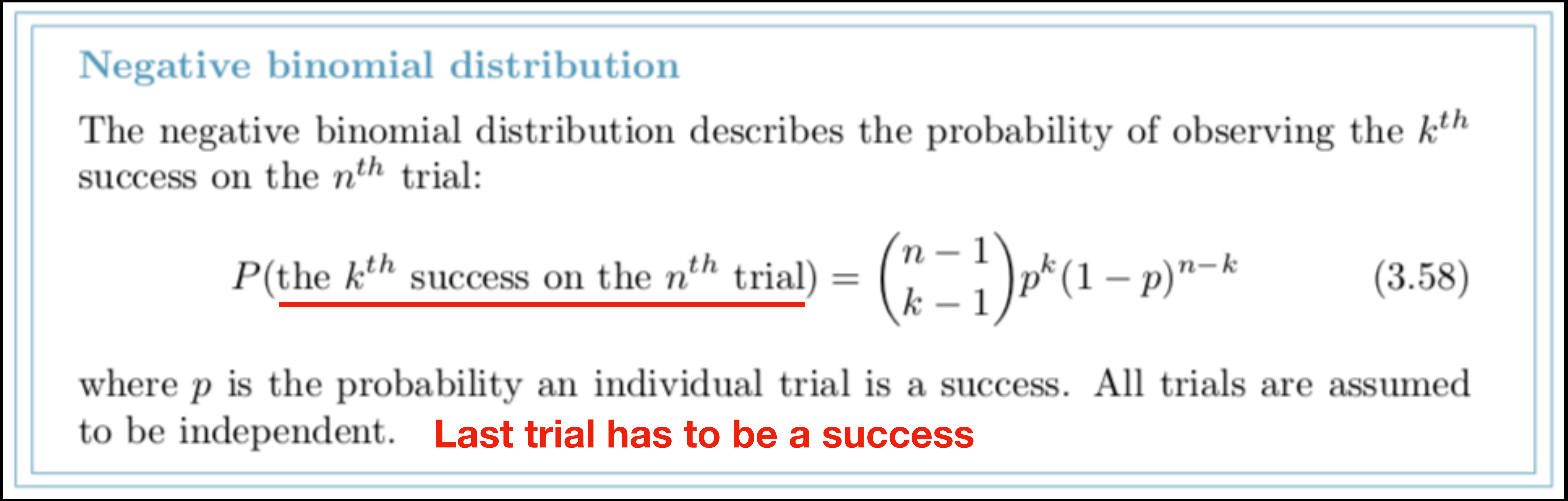
Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{3.40}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)} \tag{3.41}$$



Negative binomial distribution

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$P(\underline{\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \tag{3.58}$$

where p is the probability an individual trial is a success. All trials are assumed to be independent. Last trial has to be a success

Practice

3.29 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

Practice

3.29 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

What is p?

What is a "trial" here? What is a "success"?

What is n?

What is the probability we want to know?

Practice

3.29 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

What is p? p = 0.7

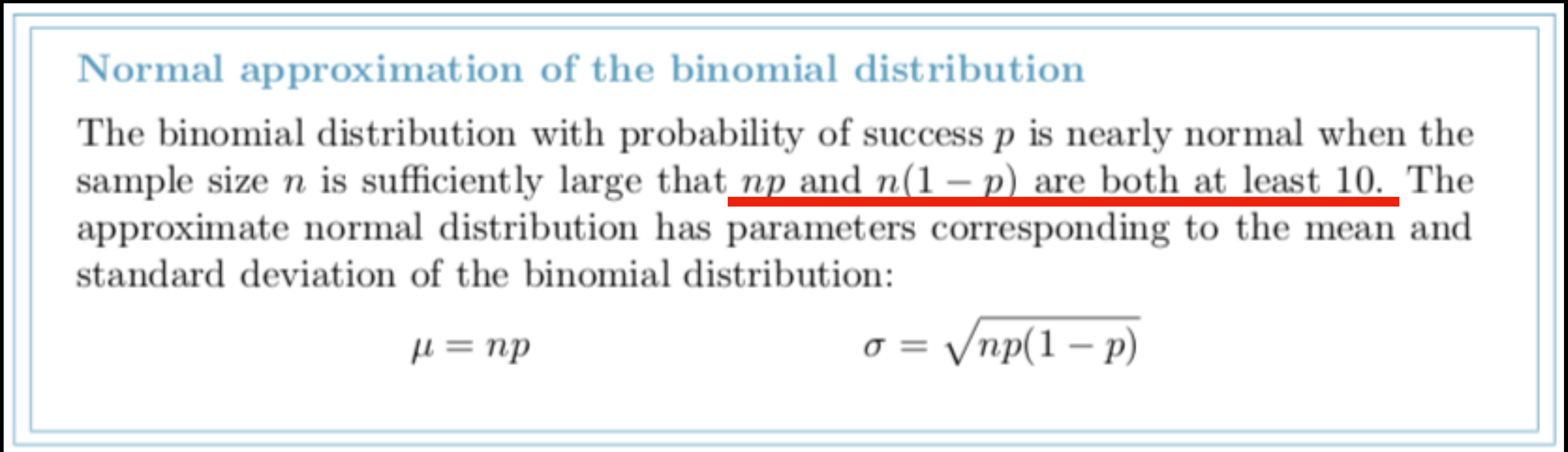
What is a "trial" here? What is a "success"?

What is n?

What is the probability we want to know?

Practice

3.29 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?



Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

p = 0.7
n = 2500

n*p = 1750
n*(1-p) = 750

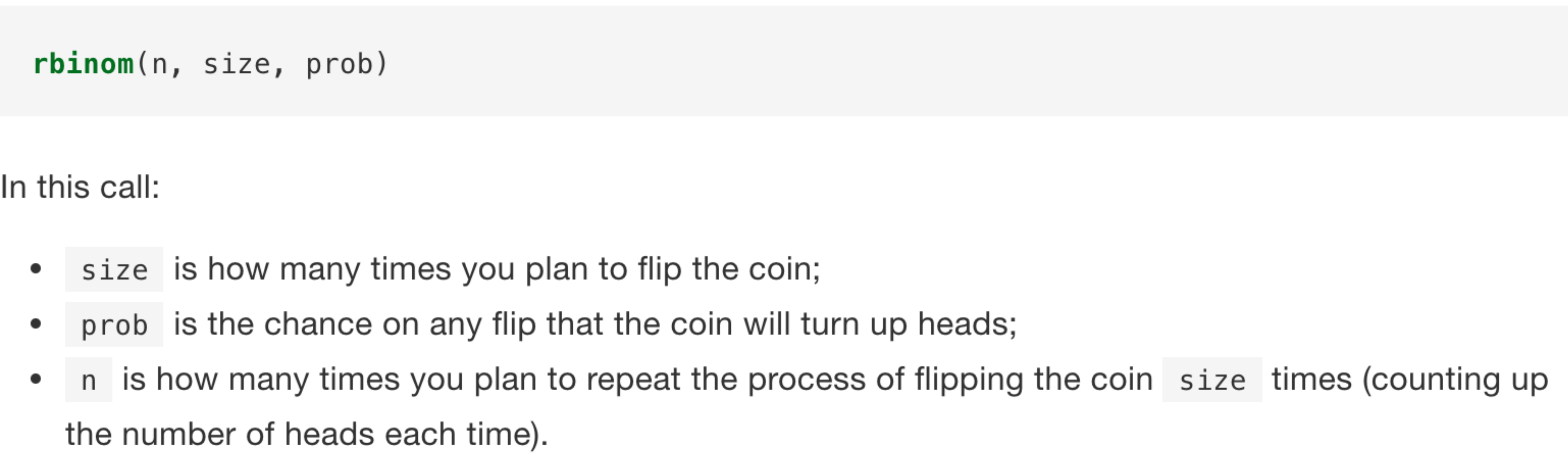
So, we can use normal distribution with the above definitions.

Practice

3.29 University admissions. Suppose a university announced that it admitted 2,500 students for the following year's freshman class. However, the university has dorm room spots for only 1,786 freshman students. If there is a 70% chance that an admitted student will decide to accept the offer and attend this university, what is the approximate probability that the university will not have enough dormitory room spots for the freshman class?

In R!

FYI: Simulations in R



```
rbinom(n, size, prob)
```

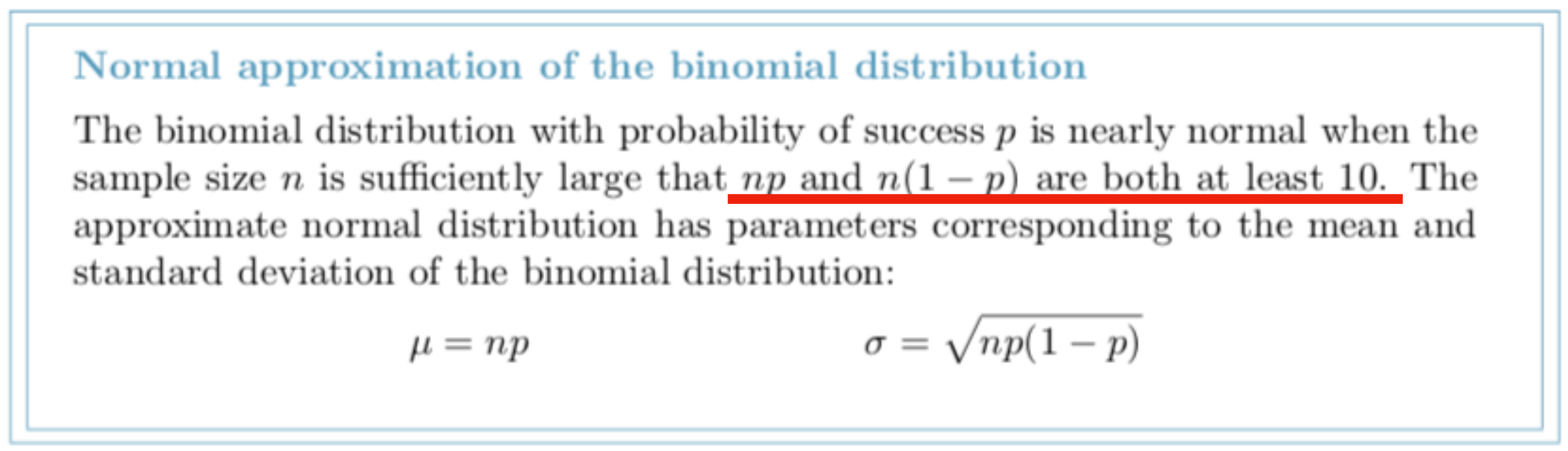
In this call:

- `size` is how many times you plan to flip the coin;
- `prob` is the chance on any flip that the coin will turn up heads;
- `n` is how many times you plan to repeat the process of flipping the coin `size` times (counting up the number of heads each time).

flip 1 fair coin 100 times: rbinom(n=1, size = 100, prob = 0.5)
flip 1 unfair coin 10 times: rbinom(n=1, size=10, prob=0.1)

flip 20 fair coins 10 times: rbinom(n=20, size=10, prob = 0.5)

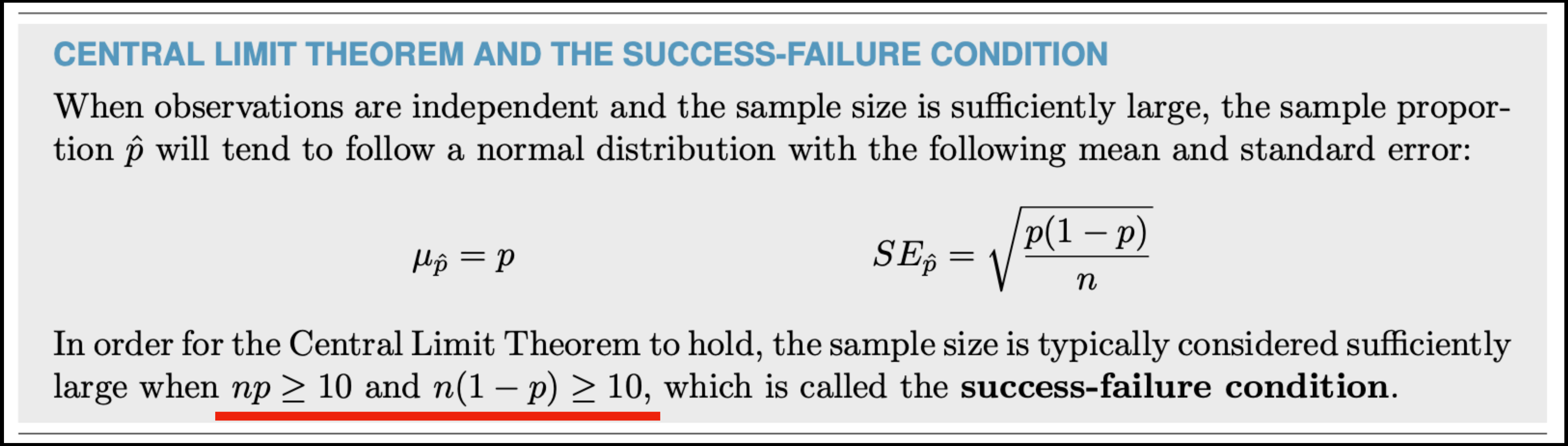
Foundations for Inference - How well can we really know anything?



Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$



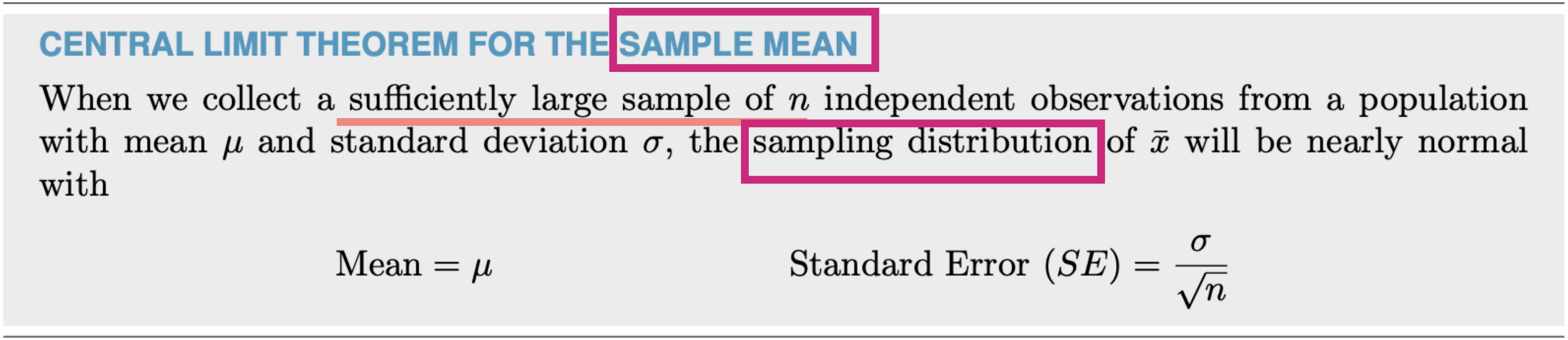
CENTRAL LIMIT THEOREM AND THE SUCCESS-FAILURE CONDITION

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with the following mean and standard error:

$$\mu_{\hat{p}} = p \qquad\qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when $np \geq 10$ and $n(1-p) \geq 10$, which is called the **success-failure condition**.

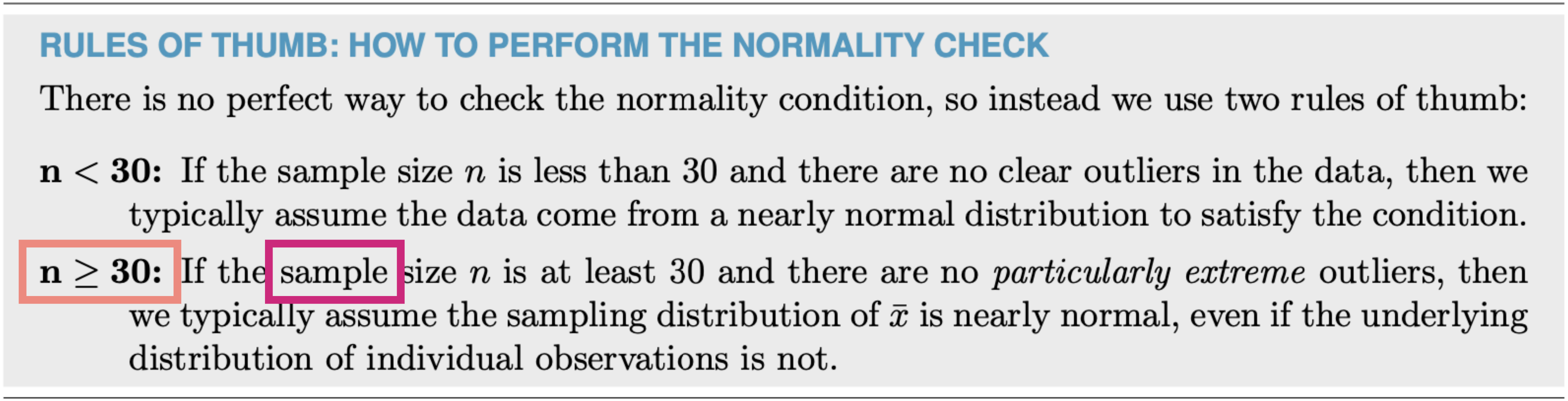
https://openstax.org/books/introductory-business-statistics/pages/7-3-the-central-limit-theorem-for-proportions



CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will be nearly normal with

$$\text{Mean} = \mu \qquad\qquad \text{Standard Error } (SE) = \frac{\sigma}{\sqrt{n}}$$



RULES OF THUMB: HOW TO PERFORM THE NORMALITY CHECK

There is no perfect way to check the normality condition, so instead we use two rules of thumb:

$\mathbf{n < 30}$: If the sample size n is less than 30 and there are no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

$\mathbf{n \geq 30}$: If the sample size n is at least 30 and there are no *particularly extreme* outliers, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

TO R!

Key Insights about how well we know the “average” number representing a sample:

Let’s say we want to know the average observation (sample mean) from a random sample taken from a population (usually the case, very rarely can we sample the entirety of the population):

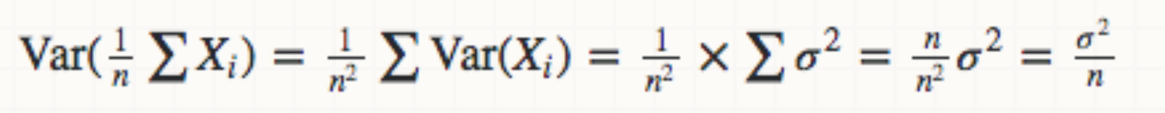
IF the samples are independent (e.g. randomly sampled)
IF the sample size is “large enough” (typically > 30 observations)
IF the underlying population distribution is not strongly skewed

This is admittedly a bit “hand-wavy” and not rigorous (stay tuned for your future stats classes!)

THEN

1. The “average” value of this population mean is the sample mean
2. The error on the measurement of the mean is given by the “standard error”: SE = s/n$^{1/2}$

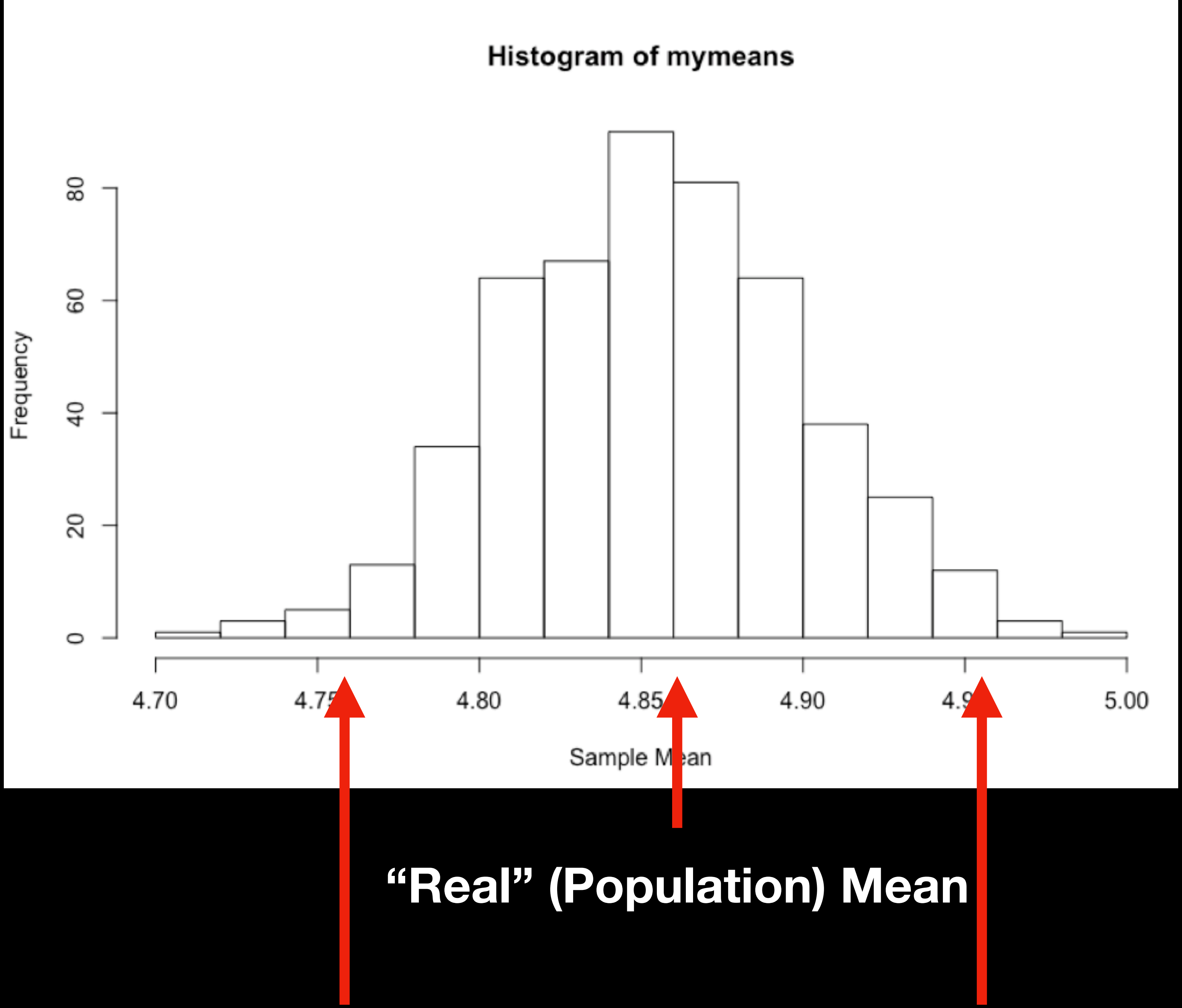
If you are curious, this comes from:



$$\mathrm{Var}\left(\tfrac{1}{n}\sum X_i\right) = \tfrac{1}{n^2}\sum \mathrm{Var}(X_i) = \tfrac{1}{n^2} \times \sum \sigma^2 = \tfrac{n}{n^2}\sigma^2 = \tfrac{\sigma^2}{n}$$

Where “s” is the standard deviation of the sample & n is the number of samples

Confidence Intervals

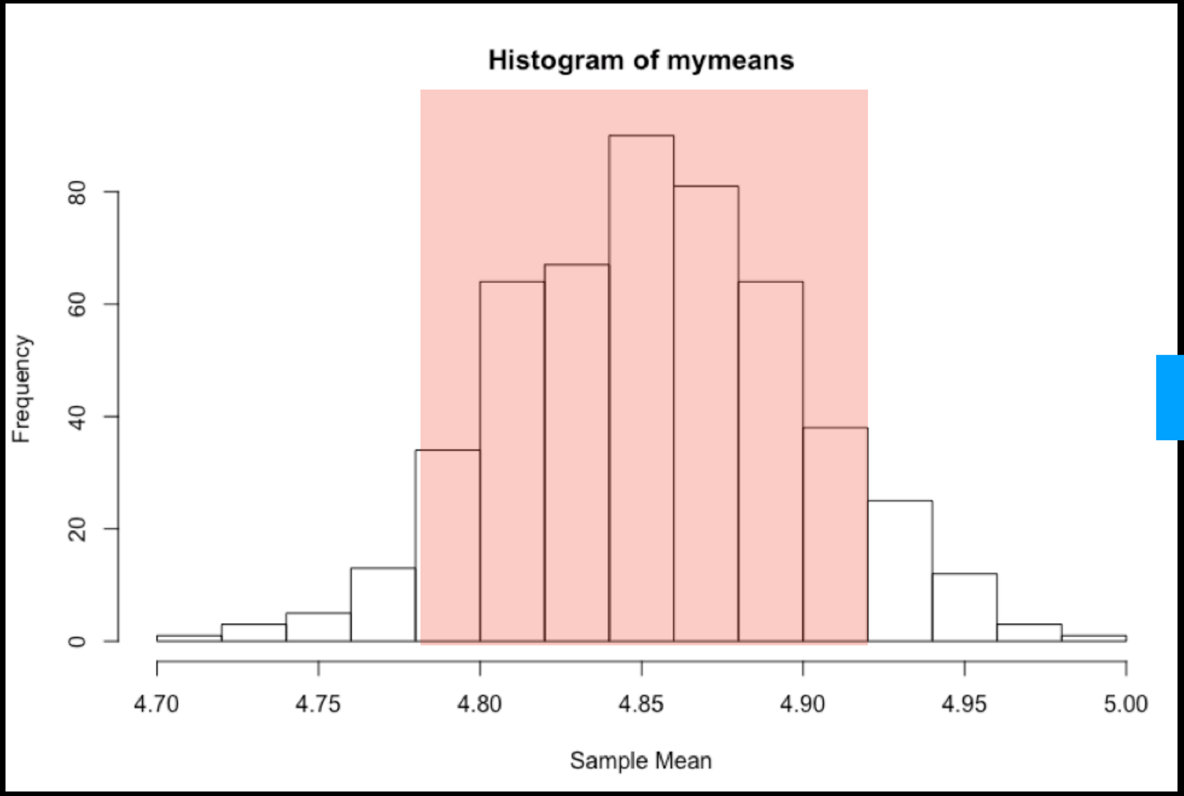


“Real” (Population) Mean

But I’m pretty sure the mean is somewhere in this interval

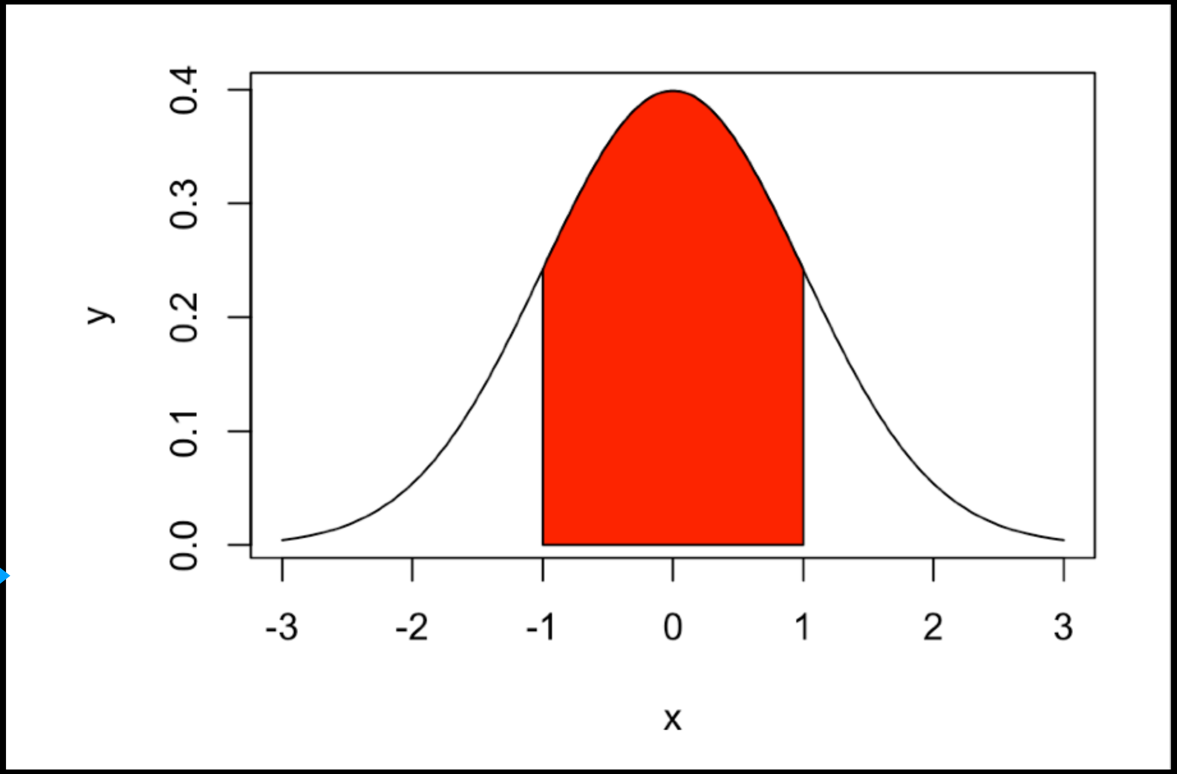
Confidence Intervals

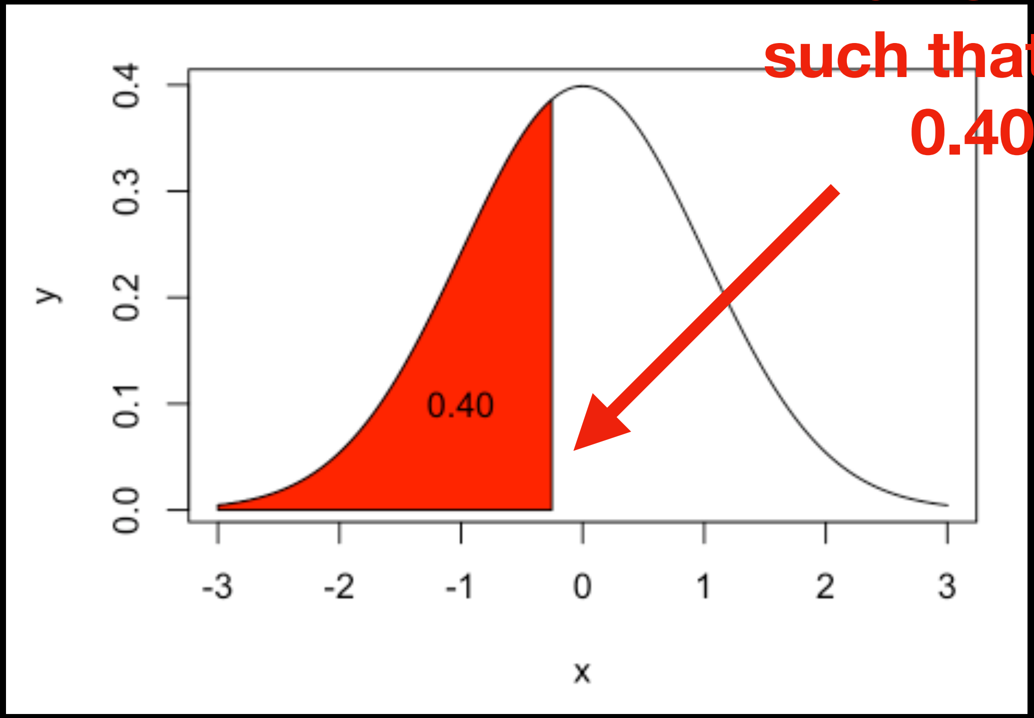
~Normal



Where are we 95% confident the population mean is in between?

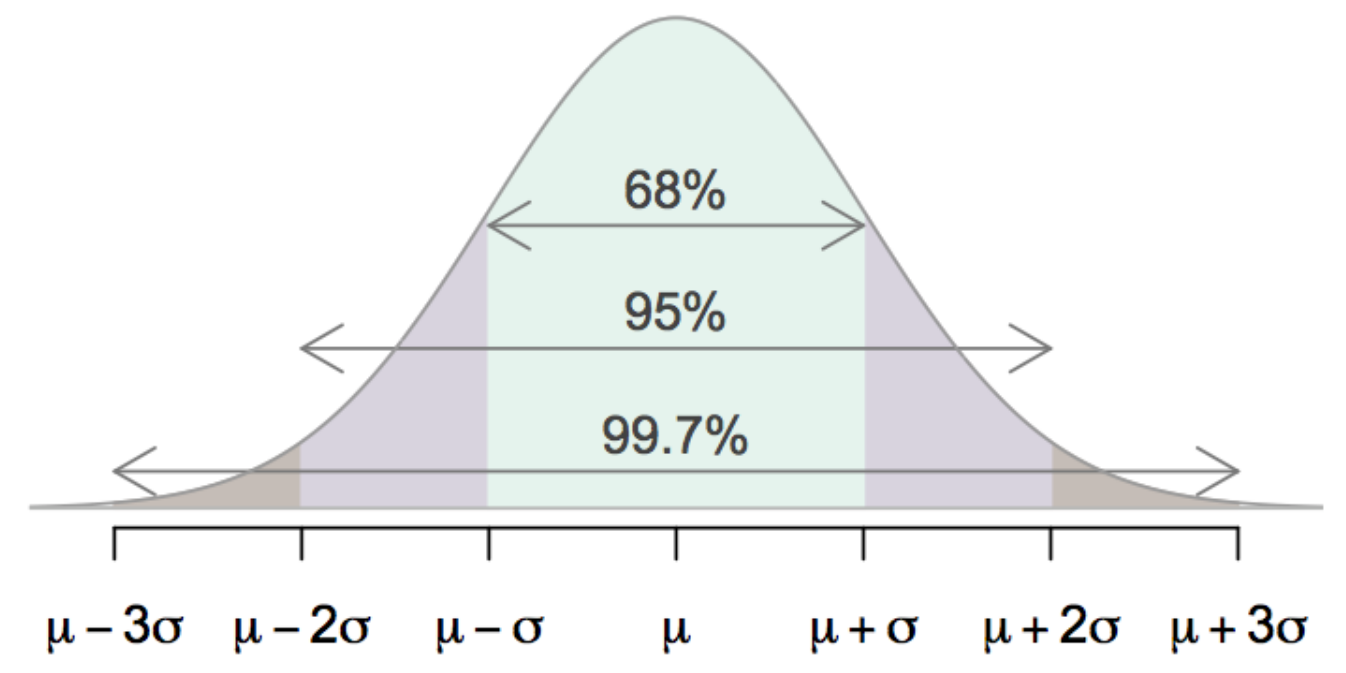
Recall:



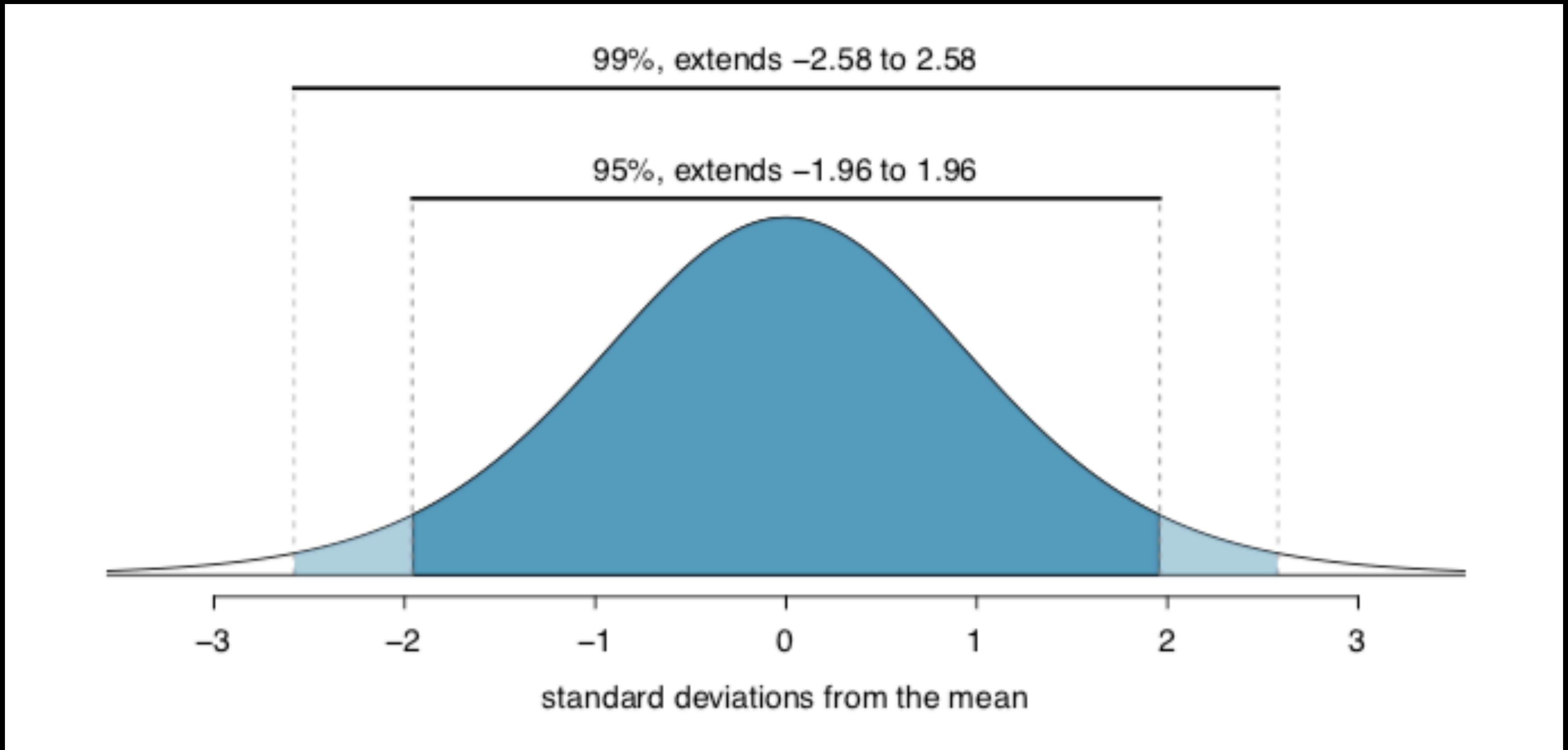


What is this number such that red area = 0.40 (40%)

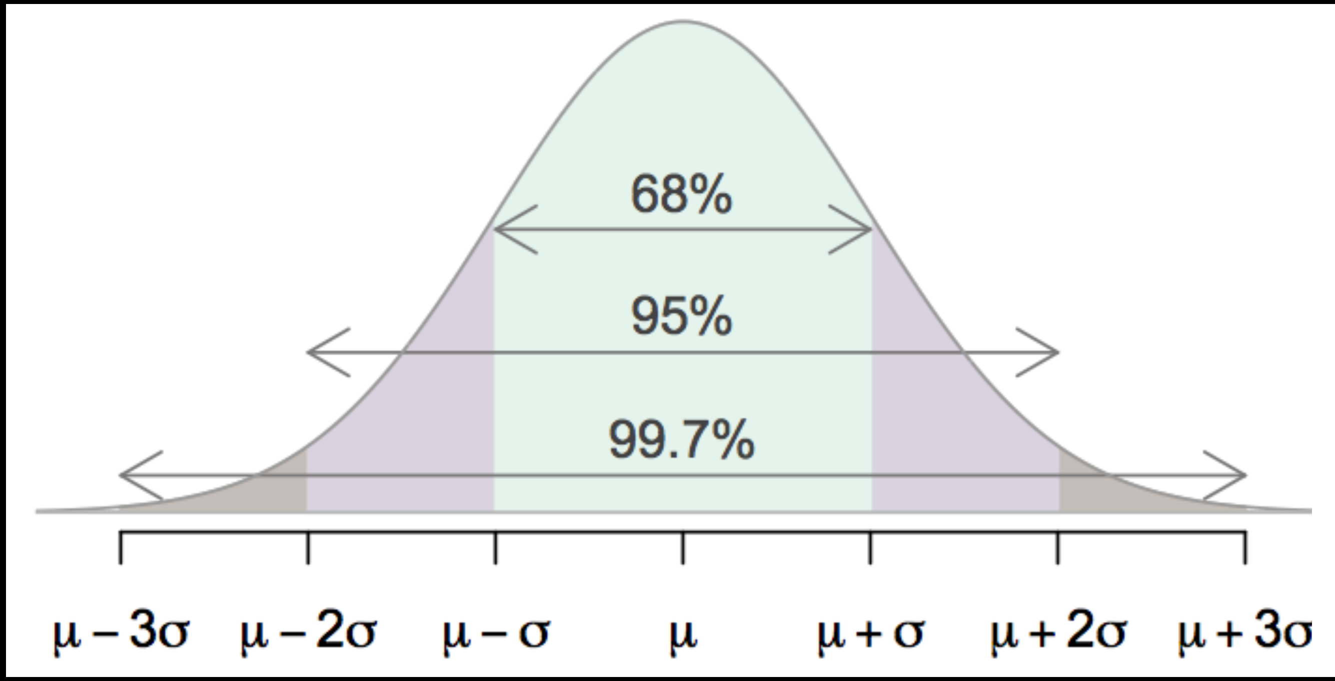
Confidence Intervals

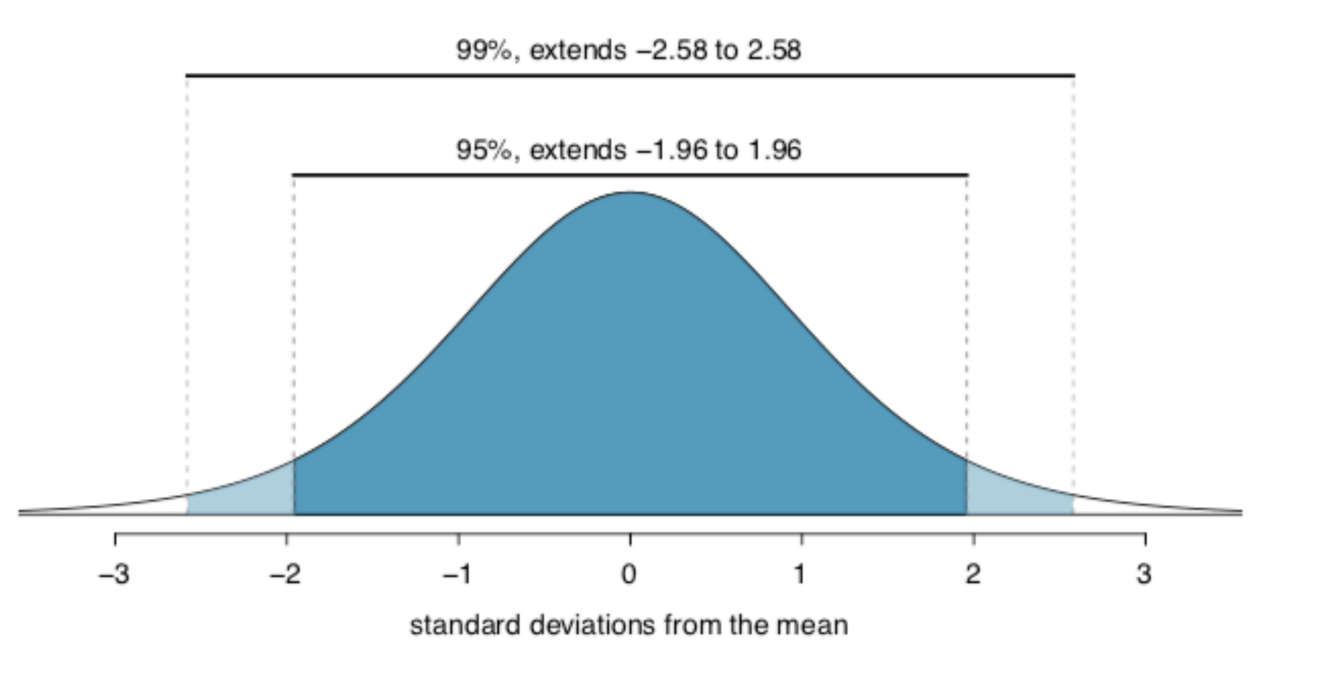


95% of the distribution is within approximately 2 SE from the mean



Confidence Intervals





Practically: We say the 95% confidence interval for a population's mean is:

sample mean +/- 1.96 X SE

Indeed, we can do this for any confidence interval requested in R.

Confidence Intervals: In general

point estimate ± z* x SE

“Point estimate” can be a mean, proportion, difference of means…

- In a confidence interval, *z* x SE* is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z* in the above formula.

Confidence Intervals: In general

point estimate ± z* x SE

- In a confidence interval, *z* x SE* is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.

- In order to change the confidence level we need to adjust z* in the above formula.

- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

- For a 95% confidence interval, z* = 1.96.

- However, using the standard normal (z) distribution, it is possible to find the appropriate z* for any confidence level.

Overview of next 2 Classes

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses. ?

   For a single mean this will look like:

   H_0: $\mu =$ null value ?

   ? H_A: $\mu <$ or $>$ or $\neq$ null value

2. Check assumptions and conditions
3. Calculate a test statistic ? and a p-value ?
4. Make a decision, and interpret it in context

- If p-value $< \alpha$, reject H_0,

  ? there is sufficient evidence for [H_A]

- If p-value $> \alpha$, do not reject H_0,

  there is not sufficient for evidence for [H_A]

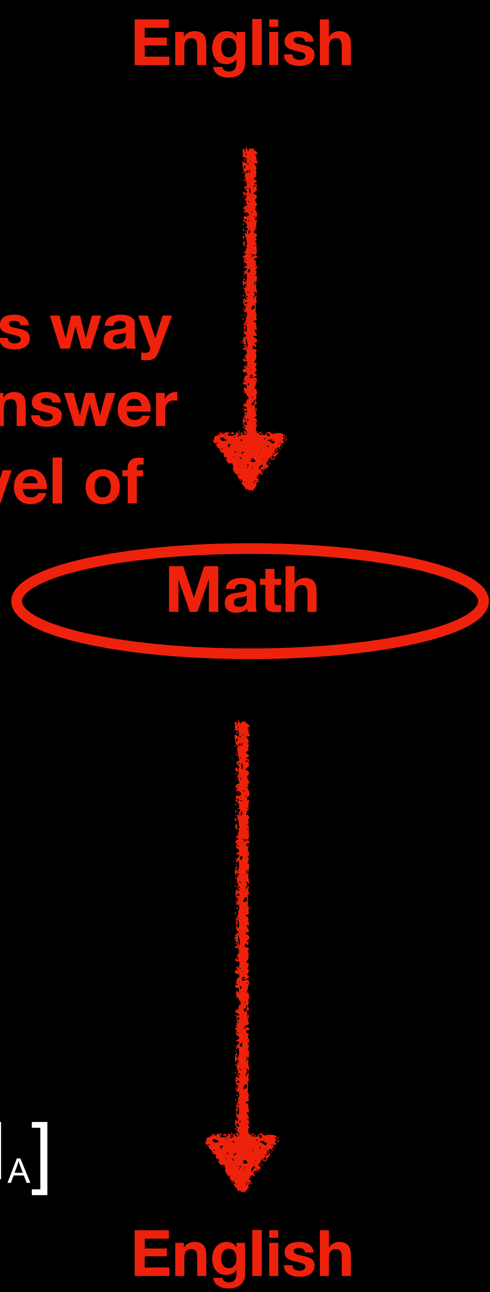
Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.
   For a single mean this will look like:
   H_0: μ = null value
   H_A: μ < or > or $\neq$ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]


English

Math

English

Provides a rigorous way to determine the answer with a specific level of confidence.

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.
   For a single mean this will look like:
   H_0: μ = null value
   H_A: μ < or > or ≠ null value

How do we frame our question into the "null" and "alternative" hypothesis framework? What are these different hypotheses?

2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

What distributions can we use to explore our sample?

Is our sample large or small?

e.g. if we are asking a question about sample means, do we expect our sample means to be normally distributed?

Use a normal distribution (Ch 5)?
t-distribution (Ch 7)?
Chi-square (Ch 6)?

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.
   For a single mean this will look like:
   H_0: μ = null value
   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

Calculate a number using our chosen distribution (e.g. the normal distribution) to see how “weird” a parameter of our sample is.

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Draw a "hard line" to determine if we can reject or we fail to reject the "null hypothesis"

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

We have actually been doing this mathematically already, you just didn't know!

- If p-value $< \alpha$, reject H_0,

  there is sufficient evidence for [H_A]

- If p-value $> \alpha$, do not reject H_0,

  there is not sufficient for evidence for [H_A]

Hypothesis Testing Framework (Ch. 5-7)

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value

2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

Let's look at some examples!

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Hypotheses: Definition

In statistics a hypothesis means a very specific thing (slightly different then, for example, a science definition): it is a claim to be tested

H_0, Null Hypothesis: the "default", "standard" or currently accepted claim, the currently accepted value for a parameter. We start this process by assuming this is true.

H_A, Alternative Hypothesis: the "research" hypothesis, or claim we need to test

Possible Outcomes:

(1) We say we "reject the null hypothesis" - i.e. H_A is *more* true than H_0

(2) We say we "fail to reject the null hypothesis"

Note: we *cannot* say that H_A or H_0 is true, only that one is *more likely* to be true than the other.

Examples of stating Hypothesis: Practice #1

It is believed a candy machine makes peanut butter cups that are on average 5g. After maintenance, a worker claims the machine no longer makes the cups at a weight of 5g. What are H_0 and H_A? How do we write them in a statistical format?

The "default" or "previously assumed" claim is the <u>null hypothesis</u>

The <u>alternative hypothesis</u> is the claim to be tested

with math

$$H_0: \mu = 5g$$

$$H_A: \mu \neq 5g$$

population mean

Examples of stating Hypothesis: Practice #2

A company has stated their ping-pong machine makes ping-pongs that are 6mm in diameter. A worker believes the machine no longer makes ping-pongs of this size and samples 100 ping-pongs to perform a hypothesis test with 99% confidence. What are H_0 and H_A?

Think on it for a moment!

Examples of stating Hypothesis: Practice #3

Doctors believe that the average teen sleeps on average no longer than 10 hours per day. A researcher believes that teens on average sleep longer. What are H_0 and H_A?

Examples of stating Hypothesis: Practice #4

The school board claims that at least 55% of students bring an iPhone to school. A teacher believes this number is too high and randomly samples 25 students to test. What are H_0 and H_A?

Examples of stating Hypothesis: Practice #5

A super fan of shopping says that on average buying socks on ebay is cheaper than in person at their local shop. A price comparison study has shown that prices for new socks are on average the same or more expensive on ebay as in their local store. Our shopper wants to setup a statistical test to see if their intuition is right. What are H_0 and H_A?

Summary: Set the hypothesis

The general outline of the process:

1. Set the hypotheses.
   For a single mean this will look like:
   H_0: μ = null value
   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

How do we frame our question into the “null” and “alternative” hypothesis framework? What are these different hypotheses?

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Summary: Set the hypothesis

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ $<$ or $>$ or $\neq$ null value

tell us something about what tests we will perform (more in examples)

2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value $< \alpha$, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value $> \alpha$, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Summary: Set the hypothesis

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value
2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,

  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,

  there is not sufficient for evidence for [H_A]

Are we interested in a hypothesis about the population mean (μ)?

Proportion (p)? (Ch. 5)

Difference of 2 means and/or paired data (μ_1 - μ_2)? (Ch. 7)

Difference between observations and theorized results? (Ch. 6)

(more in examples)

Hypothesis Testing: Where we are going

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value

2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,
  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,
  there is not sufficient for evidence for [H_A]

Picking appropriate distributions and applying - Rest of Ch 5, and 6 & 7

Hypothesis Testing: Where we are going

The general outline of the process:

1. Set the hypotheses.

   For a single mean this will look like:

   H_0: μ = null value

   H_A: μ < or > or ≠ null value

2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,

  there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,

  there is not sufficient for evidence for [H_A]

(a) normal, large sample

(b) normal?, small sample

(c) observations & theory

<u>Test Statistics</u>

(a) Z-score -> P(Z)

(b) T-Score -> P(T)

(c) χ^2 -> P(χ^2)

Hypothesis Testing: Where we are going

The general outline of the process:

1. Set the hypotheses.

For a single mean this will look like:

H_0: μ = null value

H_A: μ < or > or ≠ null value

2. Check assumptions and conditions
3. Calculate a test statistic and a p-value
4. Make a decision, and interpret it in context

- If p-value < α, reject H_0,

there is sufficient evidence for [H_A]

- If p-value > α, do not reject H_0,

there is not sufficient for evidence for [H_A]

Test Statistics

(a) Z-score -> P(Z)

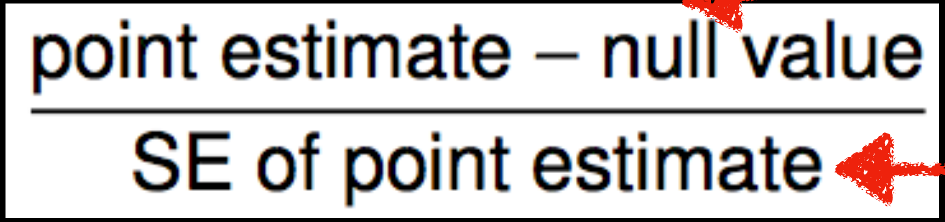
(b) T-Score -> P(T)

(c) χ^2 -> P(χ^2)

Compare Z-score, T-score or χ^2 to our level of significance - α - to see if we can reject the null hypothesis (if the p-value of our test statistic < α)

Anatomy of a test statistic

The general form of a test statistic is



$$\frac{\text{point estimate} - \text{null value}}{\text{SE of point estimate}}$$

Only tricks are:

(1) picking what the point and null values are based on our hypotheses

(2) what the form of the standard errors is based on what our underlying distribution looks like (normal, t-distribution, χ^2)

This construction is based on

- identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
- standardizing that difference using the standard error of the point estimate.

These two ideas will help in the construction of an appropriate test statistic for count data.