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:

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

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

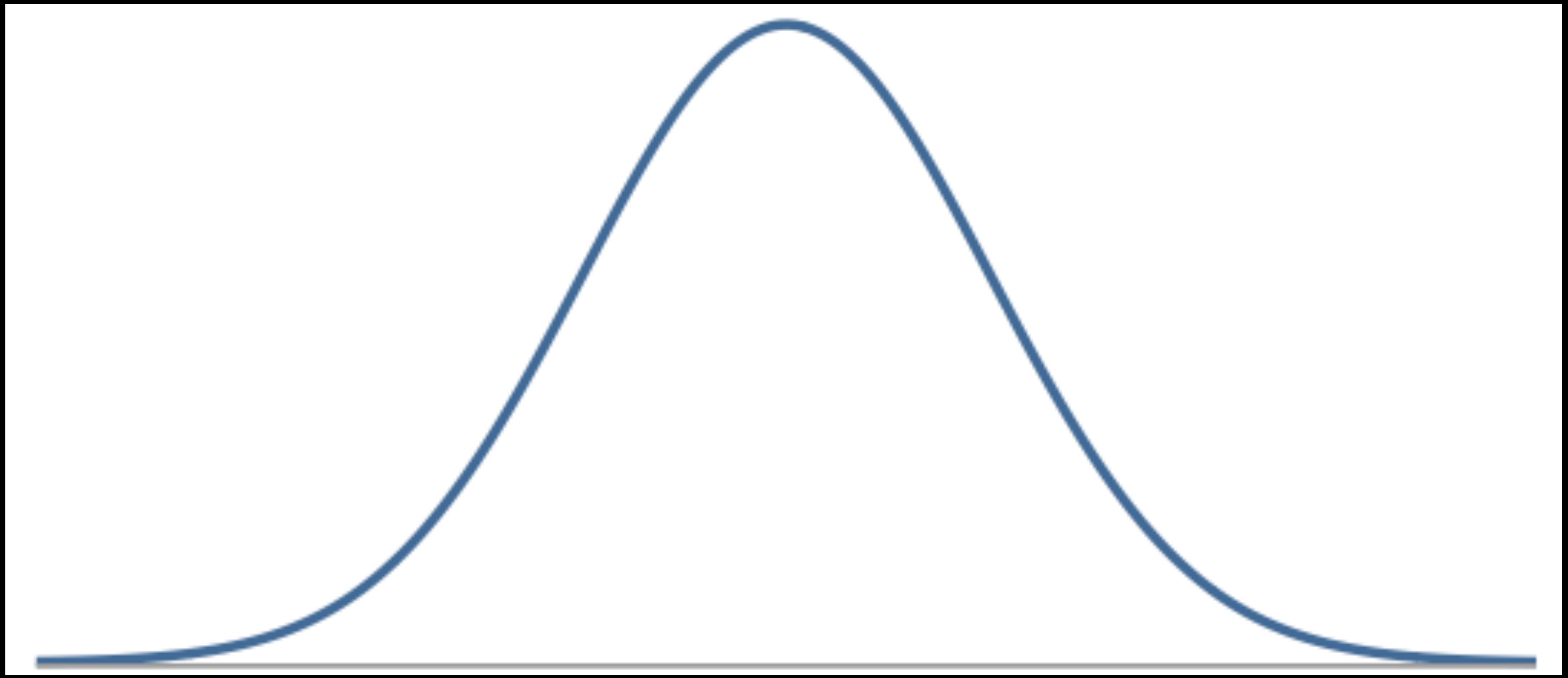
If X takes outcomes x_1, x_2, ..., x_n with probabilities p_1, p_2, ..., p_n and expected value $\mu_x = E(X)$, then to find the standard deviation of X, we first find the variance and then take its square root.

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

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

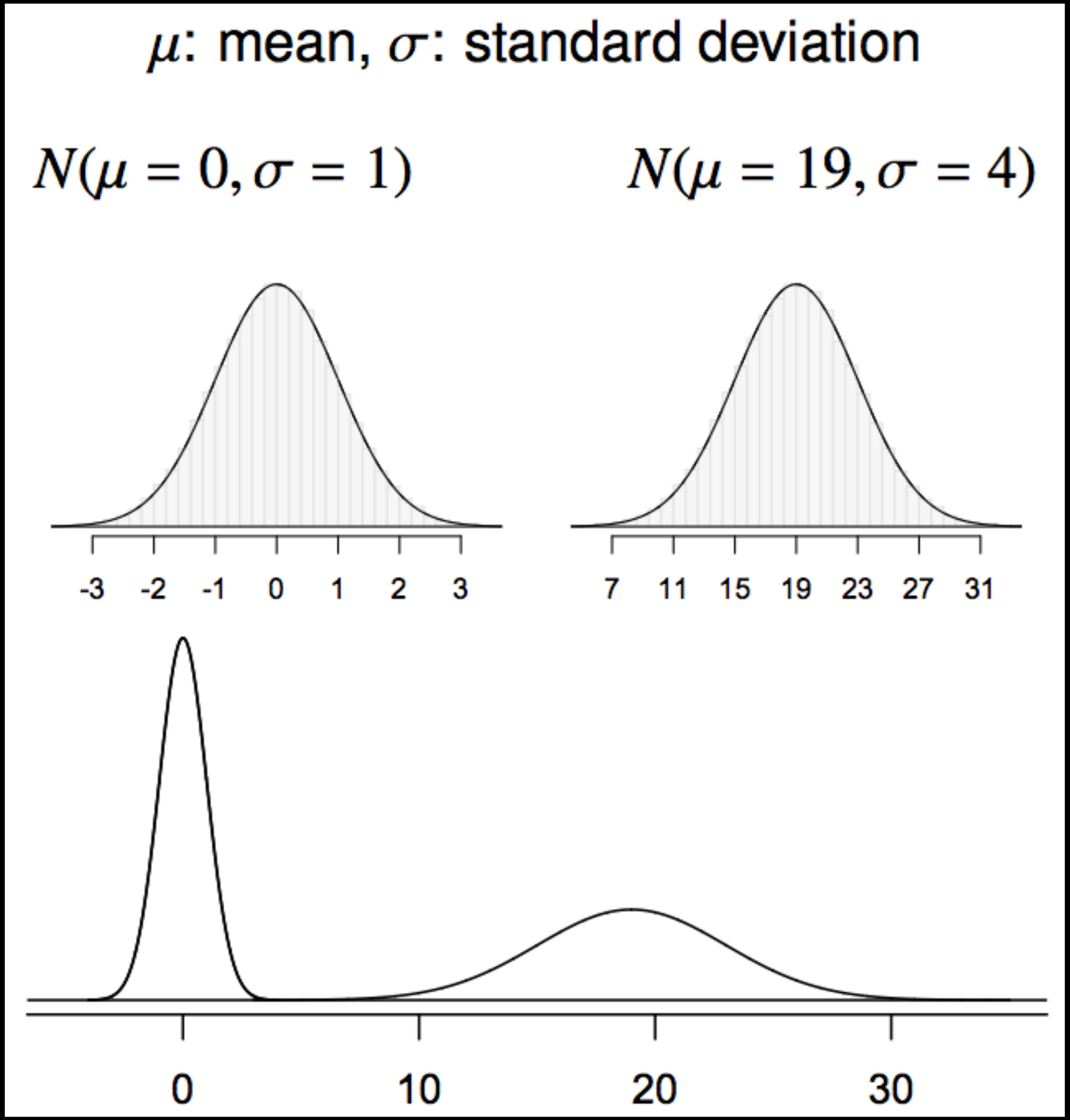
The Normal distribution

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

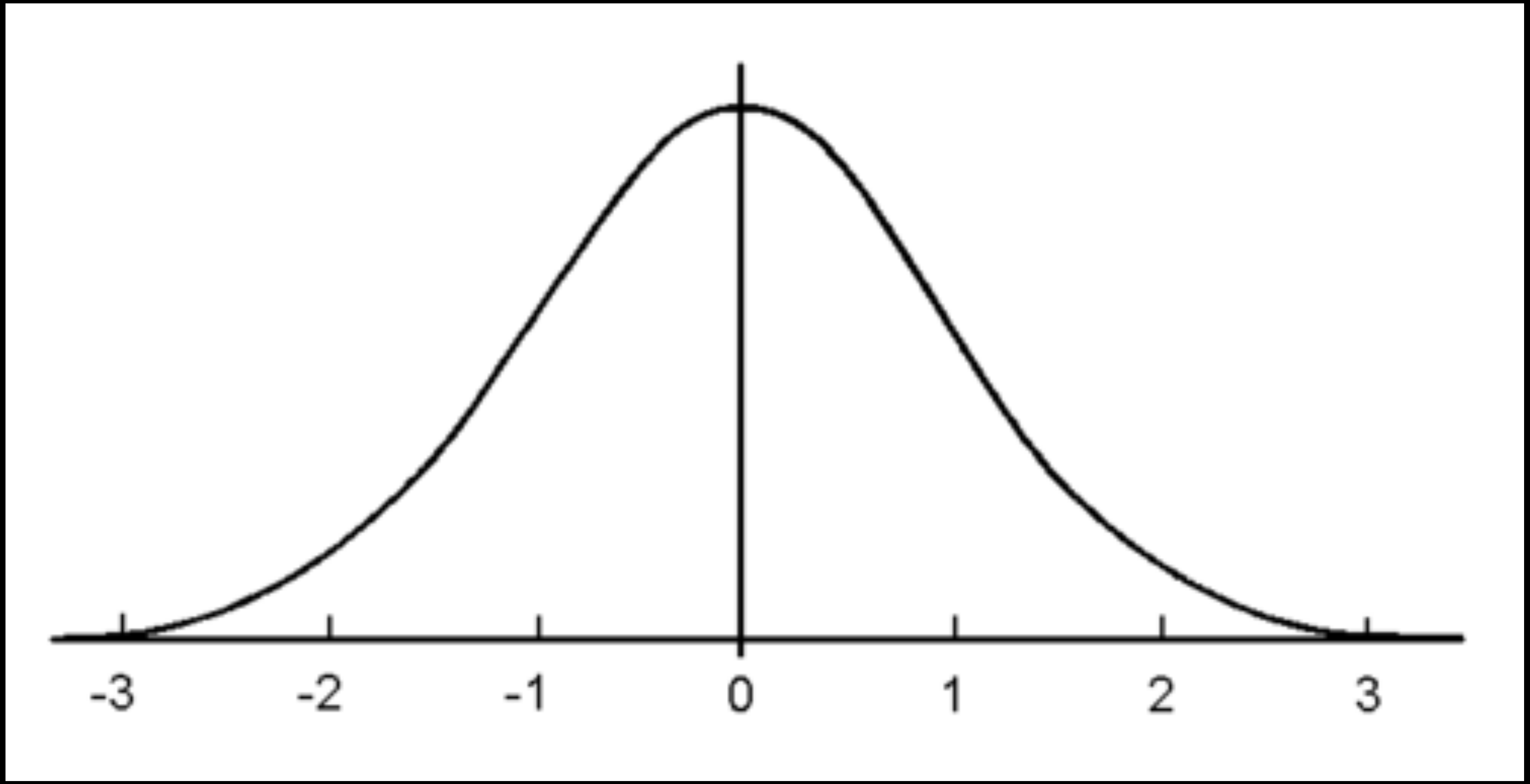


The Standard Normal Curve

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

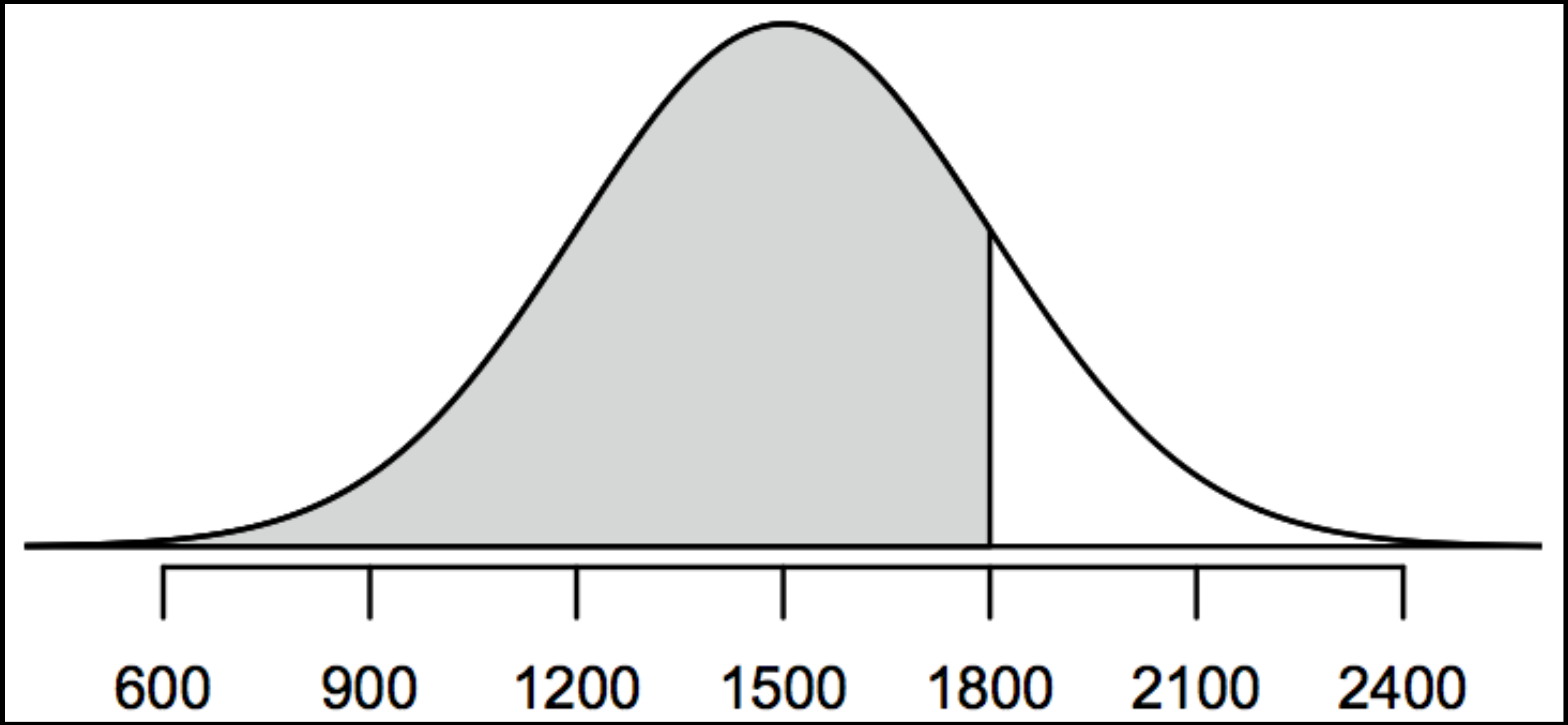
What units are on the horizontal axis?
Z-scores!

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.

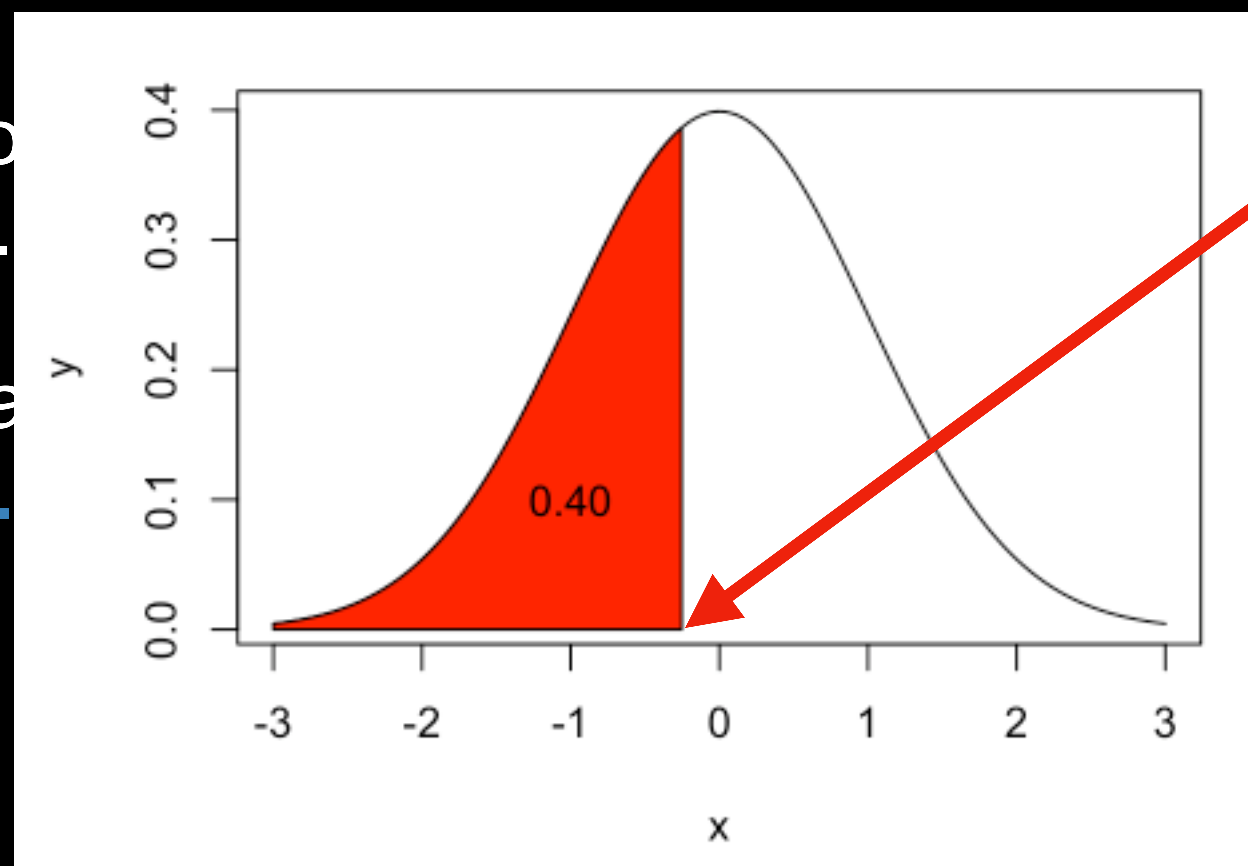


Finding percentiles from the standard normal curve

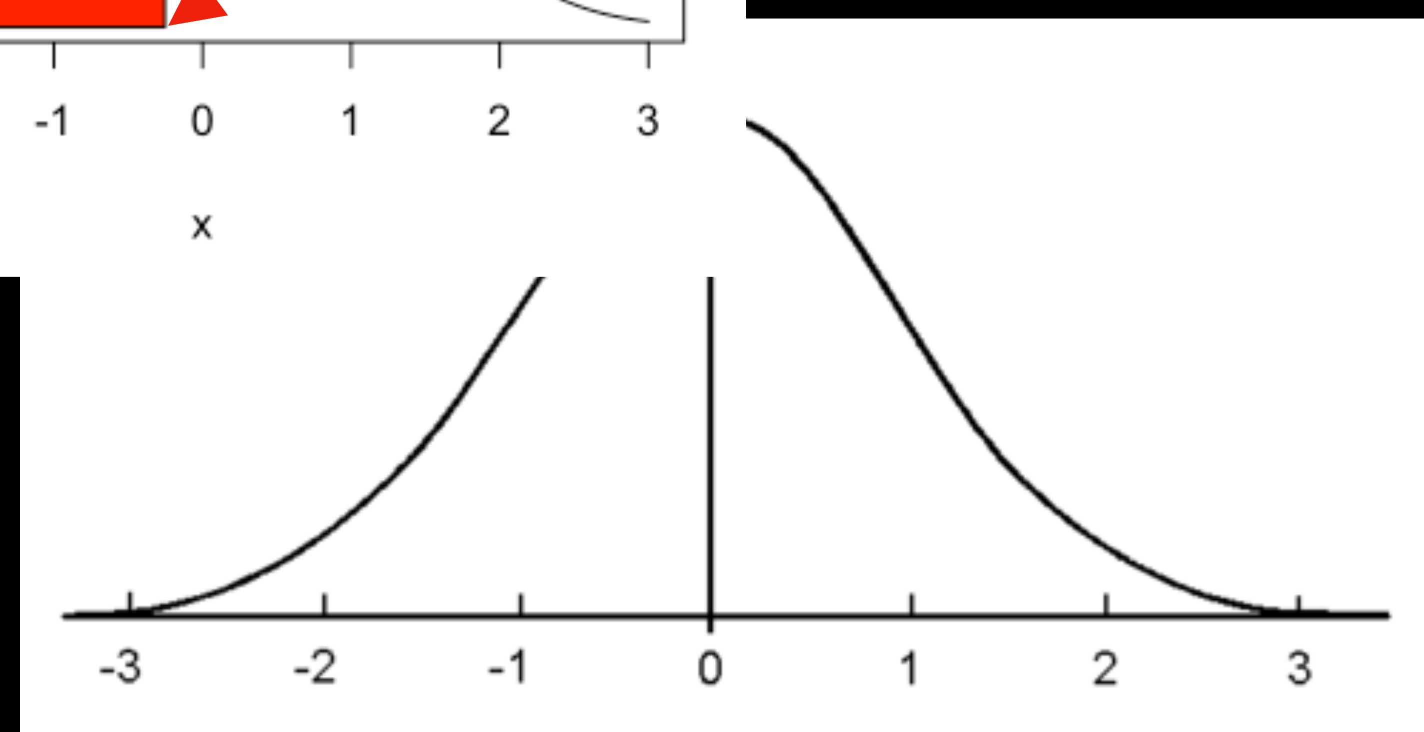
What Z-score corresponds to the 50th percentile?
i.e. P(Z < ?) = 0.5 Z =

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%)

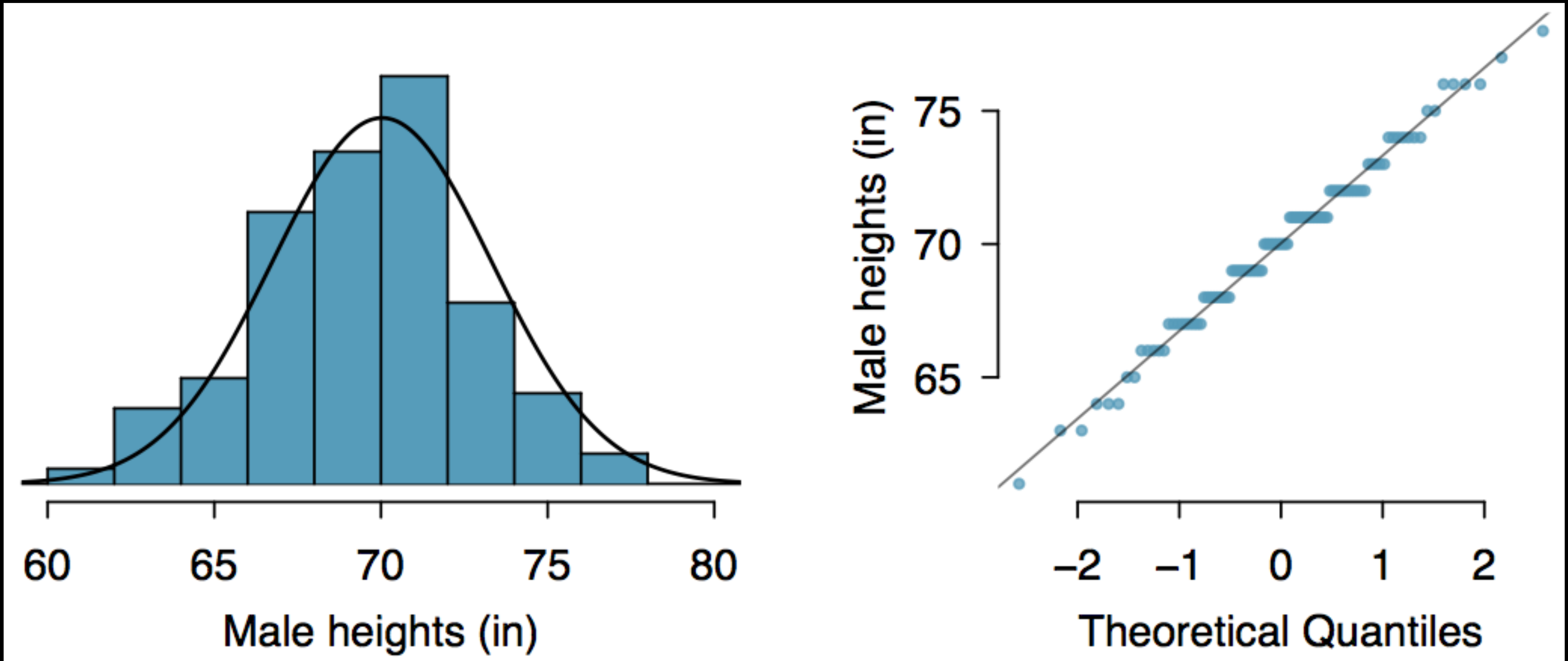


dnorm
pnorm
qnorm

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

A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is traveling at more than 100 km/hr?

Step 1: Plot the distribution

Step 2: Plot the measurement in question

Step 3: Estimate

Step 4: Calculate

Step 5: Check answer with intuition

With R!

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

$\mu = E[X] = SUM_i(Prob_i \times p_i)$

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:

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

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:

$\sigma^2 = SUM_i(Prob_i \text{ x } Var_i)$

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

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

p = prob_sample ~0.5

Similarly, the variance of X can be computed:

$$
\begin{aligned}
\sigma^2 &= SUM_i(Prob_i \times Var_i) \\
&= P(X = 0)(0 - p)^2 + P(X = 1)(1 - p)^2 \\
&= (1-p)p^2+p(1-p)^2 \\
&= p(1-p)
\end{aligned}
$$

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(failure)$^{n-1}$

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

P(success on nth try) = P(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

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

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.

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

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?

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

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?

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

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.

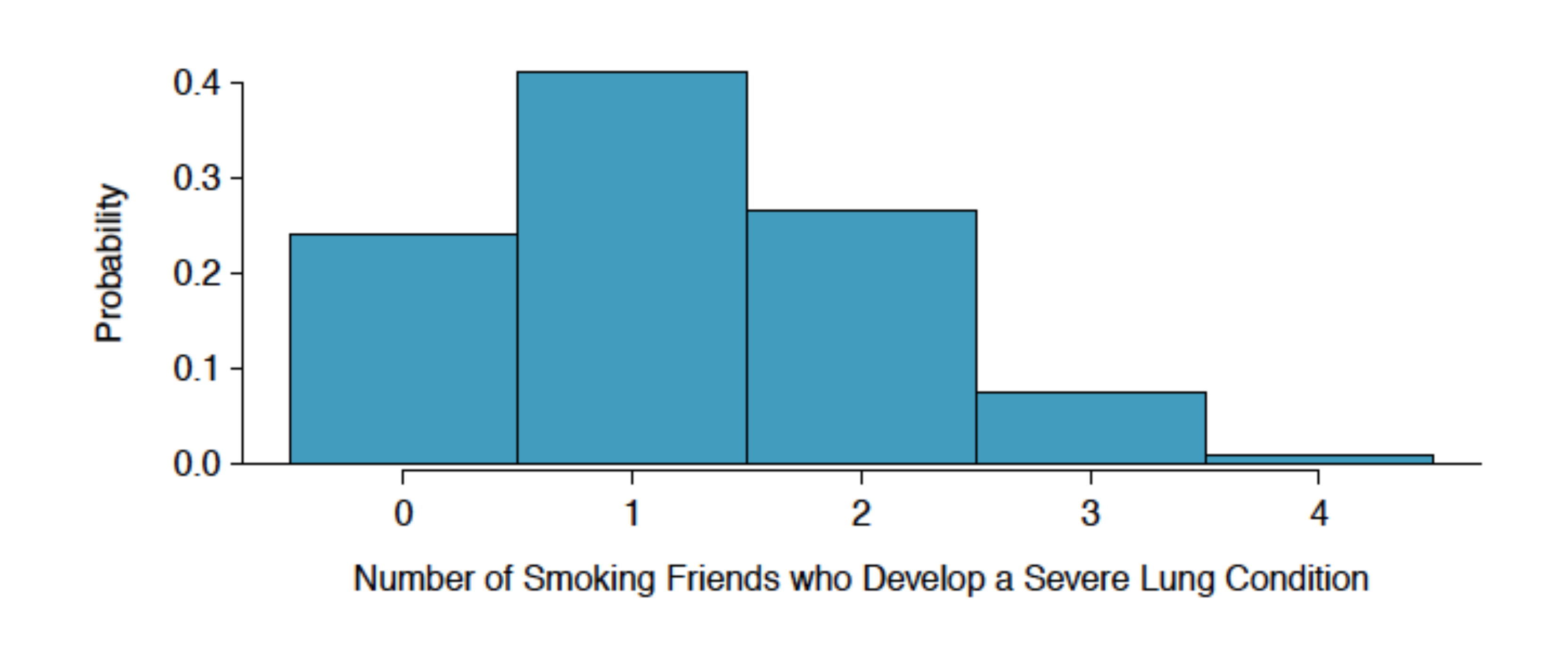
	k	probability (k out of n get lung cancer)	
	0	$\binom{4}{0}(0.3)^0(0.7)^4 = 0.240$	
n=4, k=1	1	$\binom{4}{1}(0.3)^1(0.7)^3 = 0.412$	$= 4!/(1! \times (4-1)!) \times (0.3)^1 \times (0.7)^{(4-1)}$
	2	$\binom{4}{2}(0.3)^2(0.7)^2 = 0.265$	
	3	$\binom{4}{3}(0.3)^3(0.7)^1 = 0.076$	
	4	$\binom{4}{4}(0.3)^4(0.7)^0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

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, $100 \times 0.262 = 26.2$.

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

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?

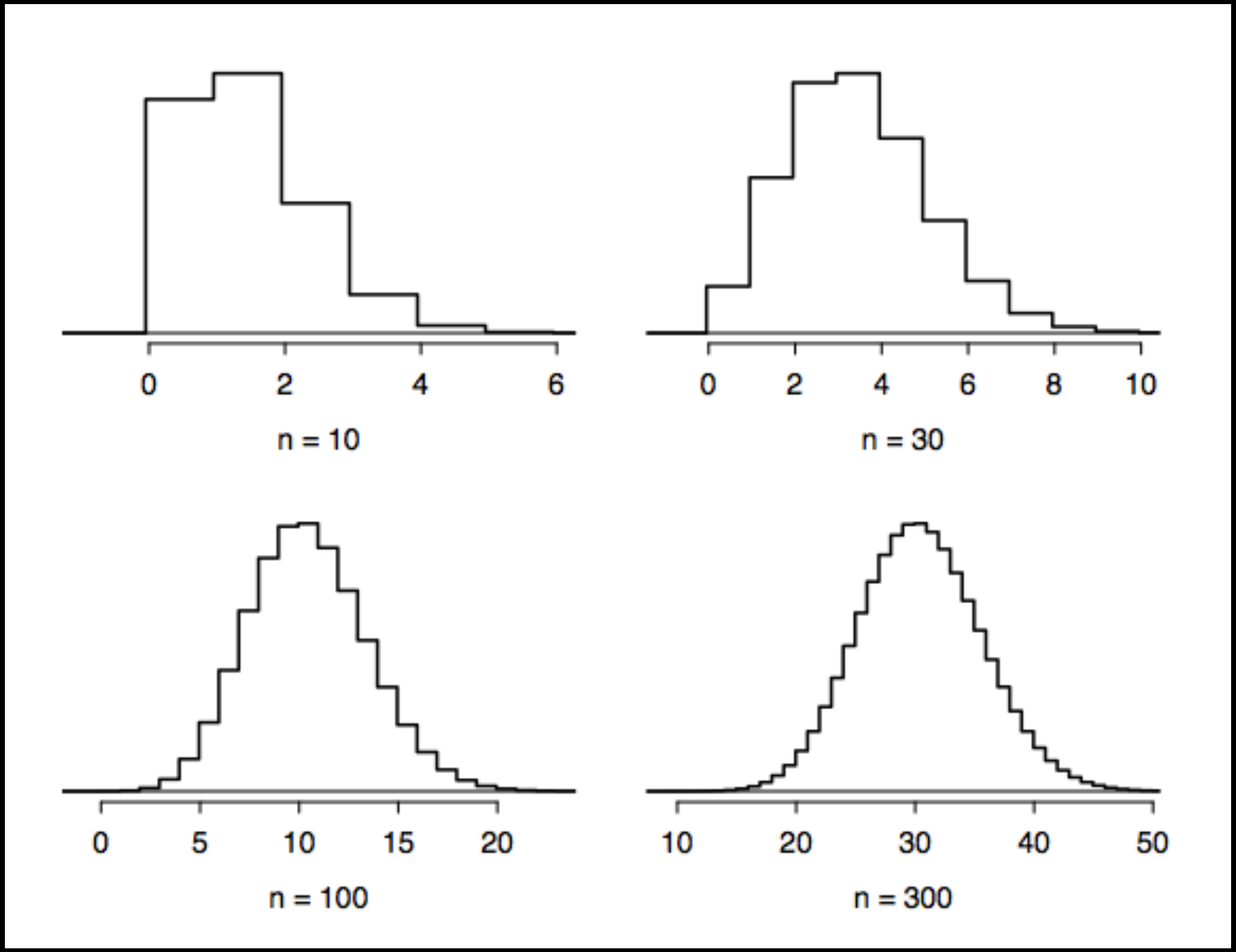
	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

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?



Note: the scales on the histograms are different!

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

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

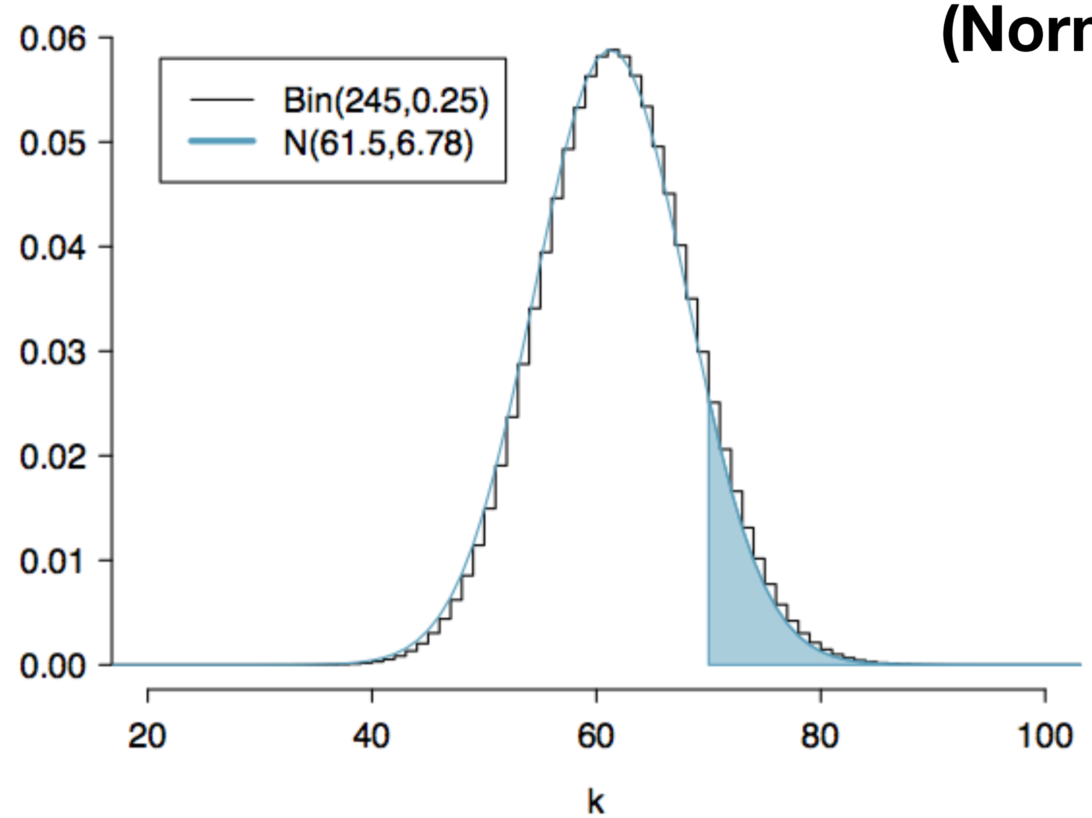
When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- 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)**



Normal approximation to the binomial

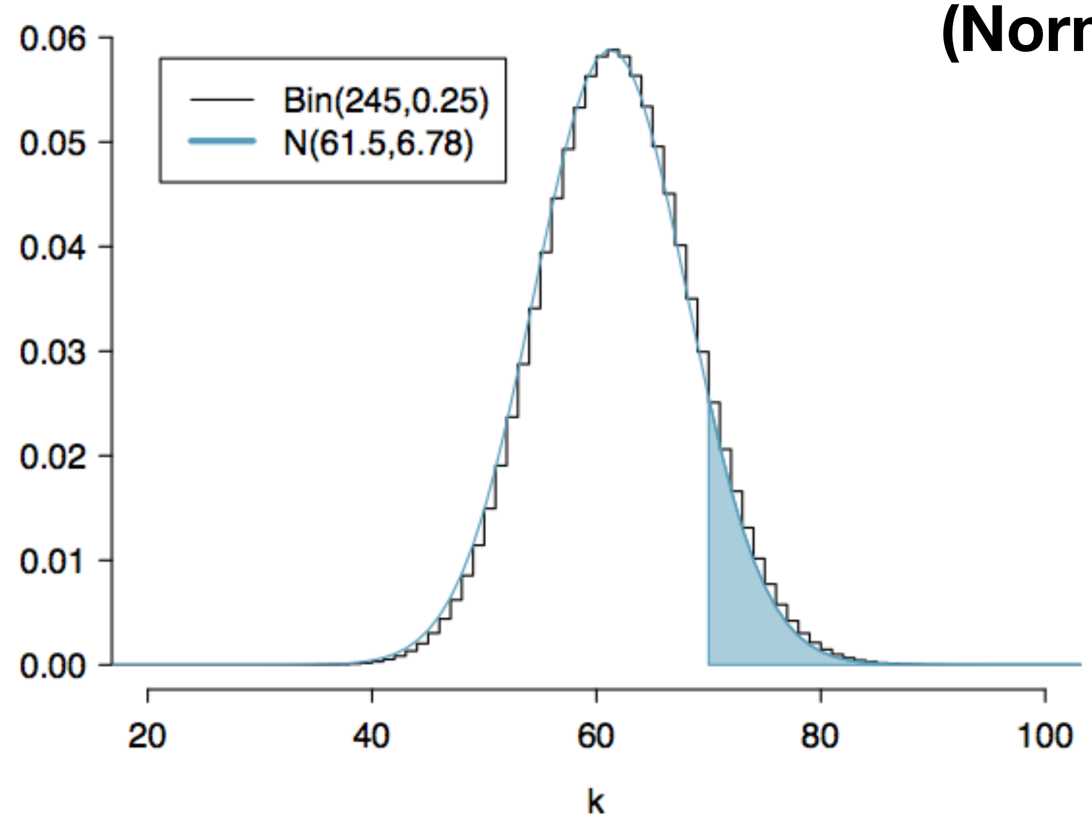
When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

- 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

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

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 (general form of geometric distribution)

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

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

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:

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

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:

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

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)

Practice

When telephones were just invented, there was a probability of 80% of success in any attempt to make a telephone call. What is the probability of 7 successful calls in 10 attempts?

Practice

A manufacturer of silver pistons finds that on the average, 12% of her pistons are rejected because they are either oversize or undersize.

What is the probability that a batch of 10 pistons will contain:

(i) no more than 2 rejects?

(ii) at least 2 rejects?