Announcements:

- Week 5 quiz begins at 4pm today and ends at 3pm on *Wed*
- If you take more than 20 minutes to complete your quiz, you will only receive partial credit. (It doesn't cut you off.)
- Today: Sections 8.5 to 8.7; skip 8.8

Homework: Due Wed, Feb 20th

Chapter 8, # 60a + 62a (count together as 1), 74, 82

Sections 8.5 to 8.7: *CONTINUOUS RANDOM VARIABLES*

- Find probabilities for *intervals*, not single values.
- X = a continuous random variable, can take any value in one or more intervals.
- $P(a < X < b)$ = proportion of the population with values in the interval (a to b).

We will cover 3 situations:

1. Uniform random variable

   Example: Buses run every 10 minutes, X = time you wait

2. Normal random variable

   Example: X = height of randomly selected woman

3. Normal approximation for a binomial random variable

   Example: X = number who favor candidate in large poll

   Note: X is actually discrete, but for large n is approximated by continuous distribution in this situation.

For each of these, you should be able to find probabilities like the following, where *a* and *b* are *fixed numbers*, X is a *random variable* of specified type: Let X = height of woman

- $P(a < X < b)$; Example: $P(65 < X < 68)$ = Proportion of women *between* 65 and 68 inches
- $P(X < a)$; Example: $P(X < 70)$ = Proportion of women *shorter* than 70 inches
- $P(X > b)$; Example: $P(X > 66)$ = Proportion of women *taller* than 66 inches

Note: For *continuous random variables*, $>$ ("greater than") and $\geq$ ("greater than or equal to") are the same because the probability of X equaling an exact value is essentially 0.

For *discrete random variables* (such as binomial) approximated by normal, that's not the case. It will matter whether it is $>$ or $\geq$.

UNIFORM RANDOM VARIABLES:

Equally likely to fall anywhere in an interval.

Example: What time of day were you born?

- X = exact time a randomly selected child is born (Natural, not Cesarean!).
- Assume equally likely to be anytime in 24 hours.
- $X = 0$ is midnight, $X = 6$ is 6:00am, $X = 7.5$ is 7:30am, etc.

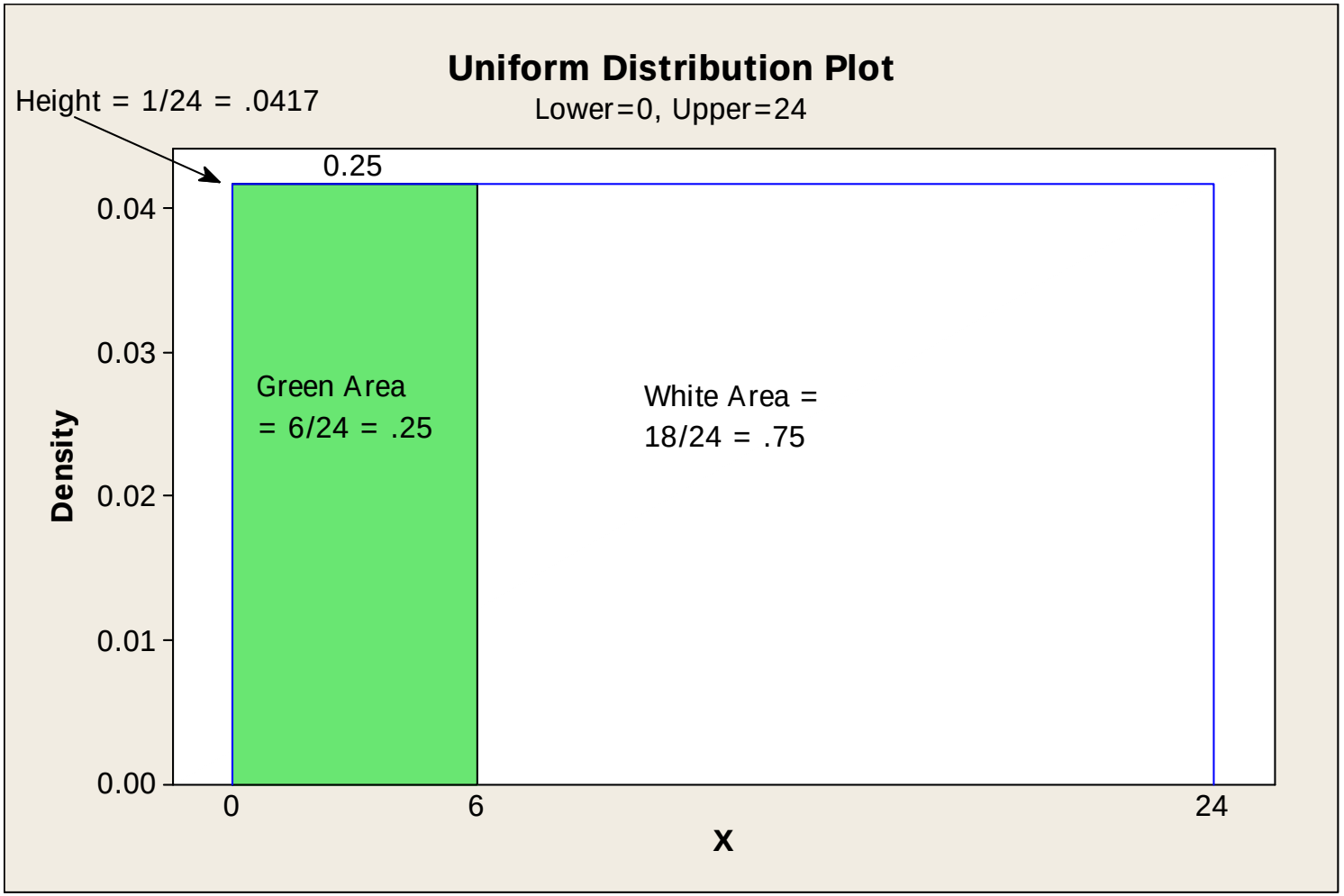
For instance:

$P(0 < X < 6)$

= Probability of being born between midnight and 6am

$$= \frac{6 \text{ hours}}{24 \text{ hours}} = \frac{1}{4} \text{ or } .25.$$

Picture of pdf (to be defined) showing $P(0 < X < 6)$ as green shaded region. Note that green region takes up ¼ of total blue rectangle. Area in blue rectangle = 1 = $P(0 < X < 24)$



GENERAL DEFINITION:
CONTINUOUS RANDOM VARIABLE:

The probability density function (A different pdf abbreviation!) for a *continuous* random variable X is denoted as f(x), and is the formula for a "curve" such that:

1. Total area under the curve = 1
2. $P(a < X < b)$ = area under the curve between a and b.

SPECIAL CASE: Uniform random variable (flat "curve")
Pdf for uniform random variable from L (lower) to U (upper) is:

$f(x) = \frac{1}{U - L}$ for all x between L and U,

f(x) = 0 otherwise (for values outside of the range L to U)

Example:

Assume birth times are uniform, 0 to 24, so f(x) = 1/24 for all x between 0 and 24, and f(x) = 0 otherwise.

Probability for uniform random variables:

$P(\text{a} < X < \text{b}) = \frac{b-a}{U-L}$ = area in rectangle from a to b

Example:

L = 0, U = 24, $P(\text{a} < X < \text{b}) = \frac{b-a}{24-0} = \frac{b-a}{24}$ (See picture)

$P(6 < X < 10) = 4/24 = 1/6$

= probability of being born between 6am and 10am.

NOTATION :

- f(x) is the pdf for the continuous random variable X.
- It is a function such that:

$$P(a < X < b) = \int_a^b f(x)dx$$

- The mean μ, variance σ^2 and standard deviation σ for X are:

$$\mu = \int_{-\infty}^{\infty} xf(x)dx \qquad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \text{ and } \sigma = \sqrt{\sigma^2}$$

Won't need calculus, will use tables, R Commander, Excel.

Parameters are fixed numbers associated with a pdf.
Example: Binomial parameter is p = probability of "Success."

UNIFORM DISTRIBUTION between L and U:

- f(x) = $\frac{1}{U-L}$ for any x between L and U, and 0 otherwise.
- Area between any two numbers a and b is $\frac{b-a}{U-L}$
- L and U are the *parameters* for a uniform distribution.

Mean and standard deviation for a uniform random variable:

- Mean is half way between L and U = $\frac{L+U}{2}$
- Standard deviation is $\sqrt{\frac{(U-L)^2}{12}}$ (not obvious how to find it)

For births:

- Mean is 24/2 = 12 (noon), may not be of much interest here!
- Standard deviation = 6.93 hours, like an "average distance" from noon, averaged over all births.

NORMAL RANDOM VARIABLES

- The mean μ and standard deviation σ are the only two *parameters* for a normal random variable.

- pdf (and thus all probabilities) *completely defined* once you know mean μ and standard deviation σ:

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

Examples: Think of the values of the following for yourself:

1. How many hours you slept last night.
2. Your height.
3. Your verbal SAT score. (Compare to other UCI students)

These are all approximately normal random variables, so you can determine where you fall relative to everyone else if you know μ and σ.

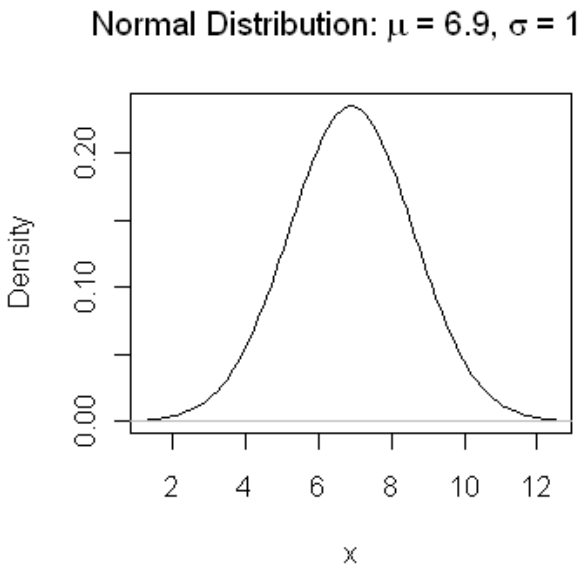
Random variable:	μ	σ
Sleep hours for students:	6.9 hours,	1.7 hours
Women's heights:	65 inches	2.7 inches
Men's heights	70 inches	3.0 inches
Verbal SAT scores, UCI students	563*	75
Verbal SAT scores, all test-takers	500	112

*Note that SAT means differ by school at UCI. You can see them here for 2002 to 2011:
http://www.oir.uci.edu/adm/IA24-fall-fr-mean-sat-by-school.pdf?R=246423

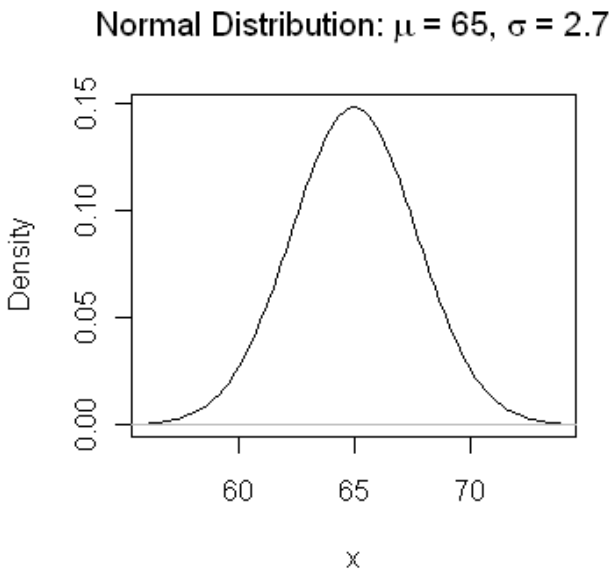
Source for all test-takers is for 2010:
http://professionals.collegeboard.com/profdownload/sat-percentile-ranks-2010.pdf

Pictures of these:

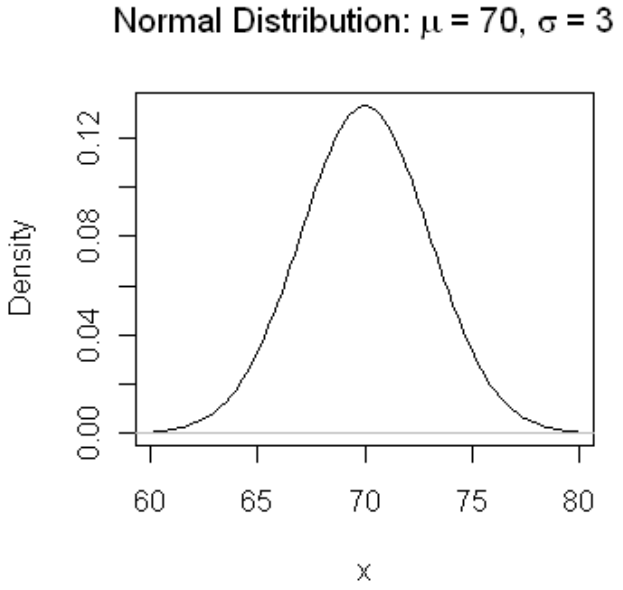
Hours of sleep



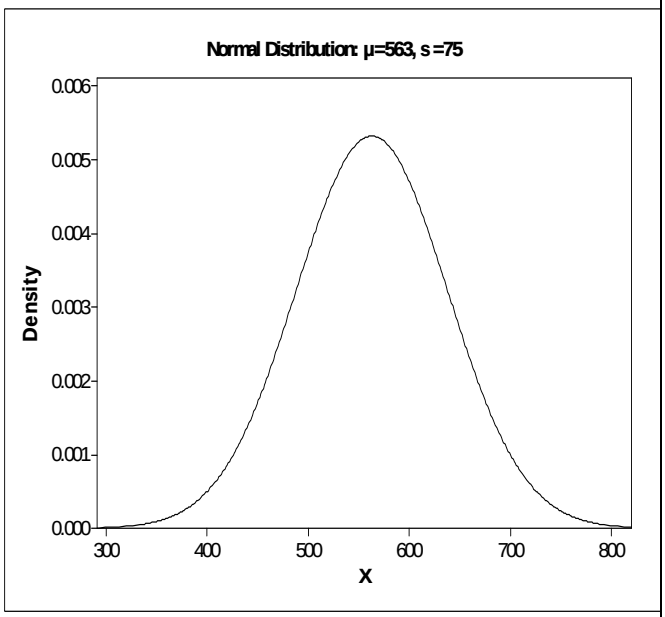
Female heights



Male heights



UCI Verbal SAT scores



What is the same and what is different about these pictures?

HOW TO FIND PROBABILITIES FOR NORMAL RANDOM VARIABLES

Two methods; in both cases you need to know *mean* **μ**, *standard deviation* **σ**, and value(s) of interest **k**:

Method 1: Convert value(s) of interest to z-scores, then use computer *or* Table A.1, which is inside the back cover of the book and on pages 668-669. (*Will need this for exams unless you have a calculator that finds normal curve probabilities.*)

Method 2: Use computer directly. (Excel or R Commander).

Often you will need Rules 1 and/or 2 from Chapter 7 as well.

Always draw a picture so you know if your answer makes sense!

Method 1 (Example: What proportion sleeps > 8 hours?)

- k is a value of interest (Ex: $k = 8$)
- μ and σ are the mean and standard deviation (6.9, 1.7)

Step 1: Convert k to a z-score, which is ***standard normal*** with $\mu = 0$ and $\sigma = 1$:

$$z = \frac{k - \mu}{\sigma} \qquad \text{Ex: } z = \frac{8 - 6.9}{1.7} = .647$$

Step 2:

Look up z in Table A.1, or use R Commander or Excel to find area above or below z. $P(Z > .647) = .259$

Table A.1 gives areas *below* z. Here is a small part of the left hand side of the table:

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

Examples (pictures of some of these shown in class):

$P(z < -2.24) = .0125$

$P(z > +2.24) = .0125$

$P(-2.24 < z < 2.24) = 1 - (.0125 + .0125) = 1 - .025 = .975$

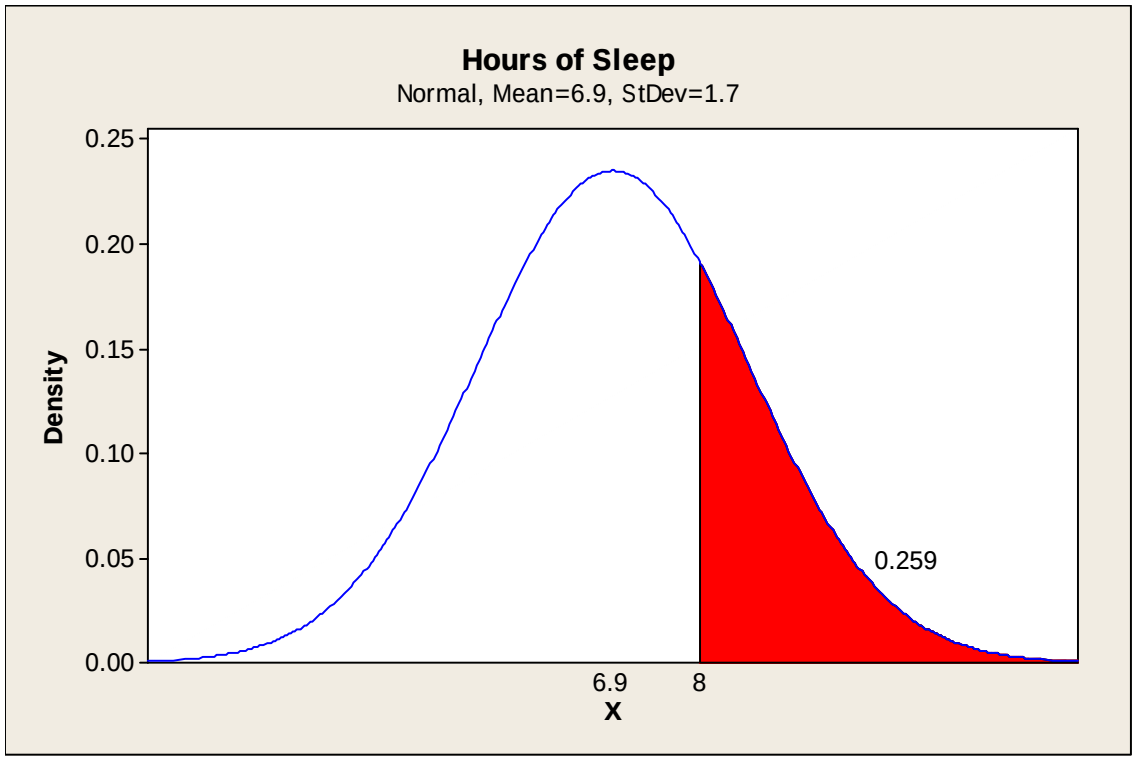
$P(-1.96 < z < 1.96) = 1 - (.025 + .025) = 1 - .05 = .95$

This last one is where the **mean ± 2 s.d.** part of the Empirical Rule comes from!

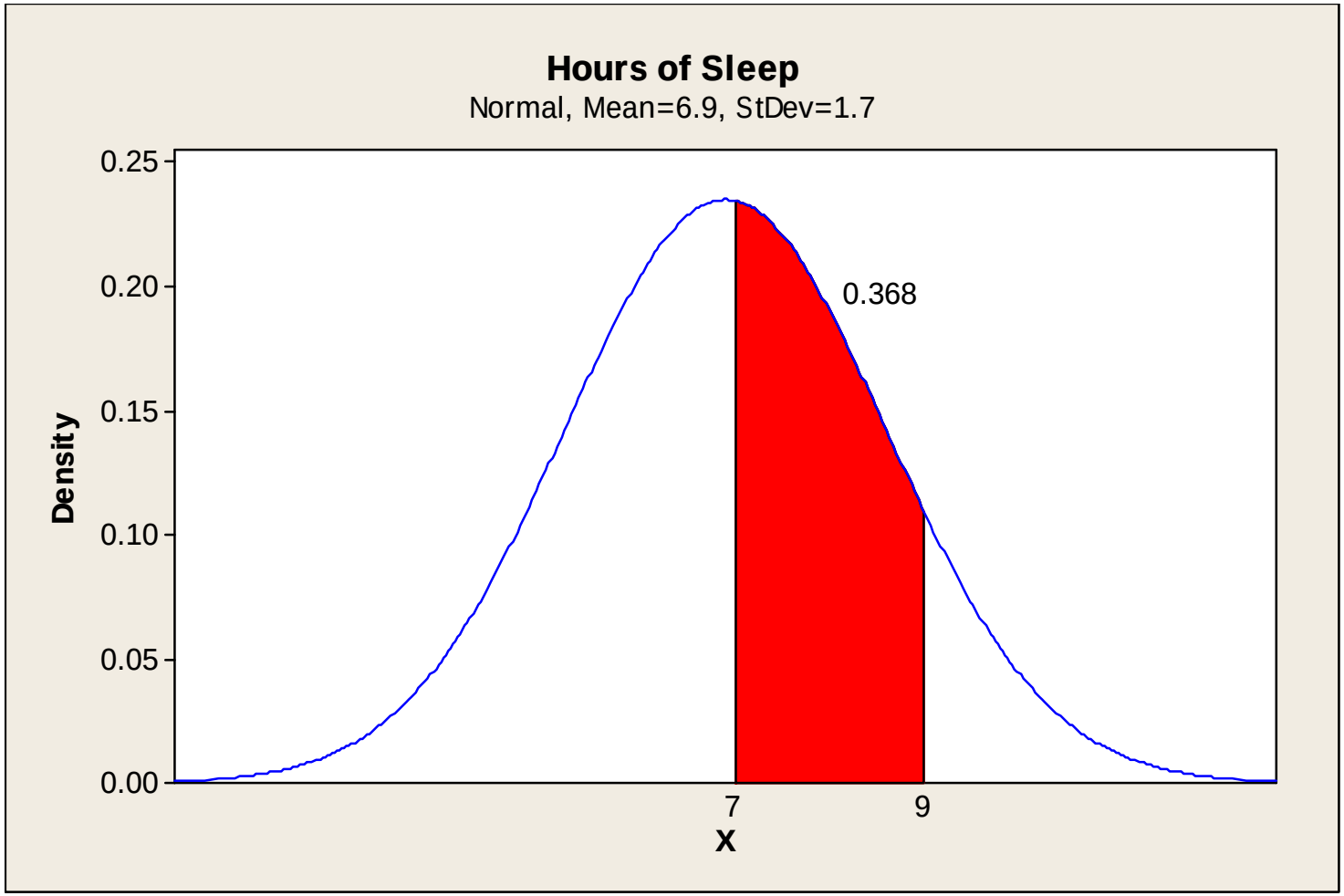
Technically, it is **mean ± 1.96 s.d.** that covers 95% of the values; we round to 2.

Some pictures for hours of sleep

- Mean = 6.9 hours, standard deviation = 1.7 hours
- P($X > 8$) = proportion who sleep more than 8 hours = .259
- Same as P($Z > .647$); from Table A.1, P($Z > .65$) = .2578



P(7 < X < 9) = proportion who sleep between 7 and 9 hours = .368



Here are some useful relationships for normal curve probabilities (a, b, d are numbers); remember that the total area under the curve from $-\infty$ to ∞ is 1.

See Figures 8.8 to 8.11 on pgs 284-285:

1. $P(X > a) = 1 - P(X \leq a)$
2. $P(a < X < b) = P(X \leq b) - P(X \leq a)$
3. $P(X > \mu + d) = P(X < \mu - d)$
4. $P(X < \mu) = .5$

Method 2: Use computer

Using R Commander (see "how to use R for *Chapter **2***" on website): *Distributions* → *Continuous distributions* → *Normal distribution* → *Normal probabilities*

- Enter variable value, mu, sigma, then choose *lower tail* or *upper tail*.
- Result shown in output window.

Using Excel: These are found under the Statistical functions. Can find z-score first, then use *=NORM**S**DIST(z)*, gives area *below* the number z, for standard normal.

- Example: *=NORMSDIST(1.96)* gives .975

Or, don't find z-score first. Use *=NORMDIST(k,mean,sd,true)*

- Note there is no "S" between NORM and DIST
- Gives area *below* *k* (*true* says you want cdf) for normal distribution with specified mean and standard deviation.

Example: Sleep hours, with mean $\mu = 6.9$ and $\sigma = 1.7$.

What proportion of students sleep *more than 8 hours*?
Use value = 8, $\mu = 6.9$, $\sigma = 1.7$, upper tail.

R Commander result: 0.2587969 (about 26%)

Excel gives proportion *less than* 8 hours:
NORMDIST(8,6.9,1.7,true) = .741203

- Use complement rule from Chapter 7:

$$P(X > 8) = 1 - P(X \leq 8)$$

  - Proportion *more than* 8 hours = 1 – .741203 = .258797 (same as result from R Commander).

What proportion of students get the recommended 7 to 9 hours of sleep? Picture showed that it was about .368, or 36.8%.

Get what we need from R Commander:

- Proportion *less than 9 hours* is .8916
- Proportion *less than 7 hours* is .5234
- Proportion between 7 and 9 hours is .8916 – .5234 = .3682 or about 36.8%

See Section 8.6 for practice in finding proportions for normal random variables.

Main rule to remember: Area (proportion) under entire normal curve is 1 (or 100%). Draw a picture!!

Working backwards: Find the cutoff for a certain proportion

Example: What *z*-value has 95% (.9500) of the standard normal curve below it?

Method 1: Table A.1. Find .9500 in body of table, then read *z*.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
...										
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633

Result: It's between $z = 1.64$ and $z = 1.65$, so use $z = 1.645$

What is the amount of sleep that only 5% of students exceed?

In general, $X = z\sigma + \mu$, so $X = 1.645(1.7) + 6.9 = 9.7$ hours

Method 2: Using R Commander:
Distributions → *Continuous distributions* → *Normal distribution* → *Normal quantiles*

Enter proportion of interest, mean, standard deviation, and upper or lower tail.

Ex: Height with 30% of women *above* it.
Enter .3, 65, 2.7, upper.
(Proportion of interest = .3, mean = 65, st. dev. = 2.7, want upper tail.)
Result is 66.41588.

Conclusion is that about 30% of women are taller than 66.42 inches

Section 8.7: USING NORMAL DISTRIBUTION TO APPROXIMATE BINOMIAL PROBABILITIES

Example from last lecture:
Political poll with $n = 1000$.
Suppose *true* $p = .48$ in favor of a candidate.
X = number in poll who say they support the candidate.
X is a binomial random variable, $n = 1000$ and $p = .48$.

- n trials = 1000 people
- *"success"* = support, *"failure"* = doesn't support
- Trials are *independent*, knowing how one person answered doesn't change others probabilities
- p remains fixed at .48 for each random draw of a person

Mean = np = (1000)(.48) = **480**.

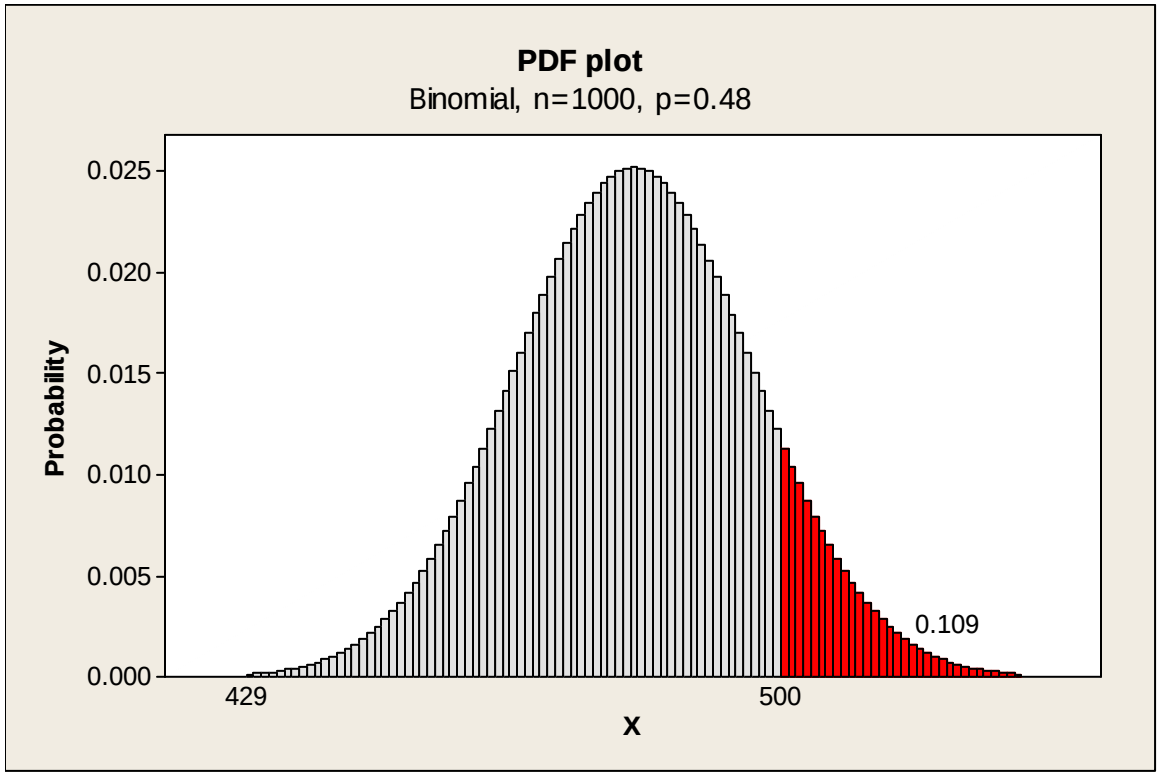
Standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{1000(.48)(.52)} = \mathbf{15.8}$

What is the probability that *at least half* of the *sample* support the candidate?
(Remember only 48% of population supports him or her.)

$P(X \geq 500) = P(X = 500) + P(X = 501) + \ldots + P(X = 1000)$.

Using Excel: $1 - P(X \leq 499) = 1 - .891 = .109$.

Picture of the binomial pdf for this situation; each tiny rectangle covers one value, such as 500, 501, etc. Shaded area of .109 is area of all rectangles from 500 and higher.



See next slide for interpretation.

- In polls of 1000 people in which 48% favor something, the poll will say *at least half favor it* with probability of .109, i.e. just over .10 or in just over 10% of polls.

- To find the probability, the computer had to sum the areas of all of the red rectangles. There is a better way, especially if doing this by hand!

NORMAL APPROXIMATION FOR BINOMIAL RANDOM VARIABLE

If X is a binomial random variable with n trials and success probability p, and if n is large enough so that np and $n(1-p)$ are both at least 5 (better if at least 10), then X is *approximately a normal random variable* with:

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

Therefore

$$P(X \le k) \approx P(z \le \frac{k - np}{\sqrt{np(1-p)}})$$

In other words, these are *almost* equivalent:

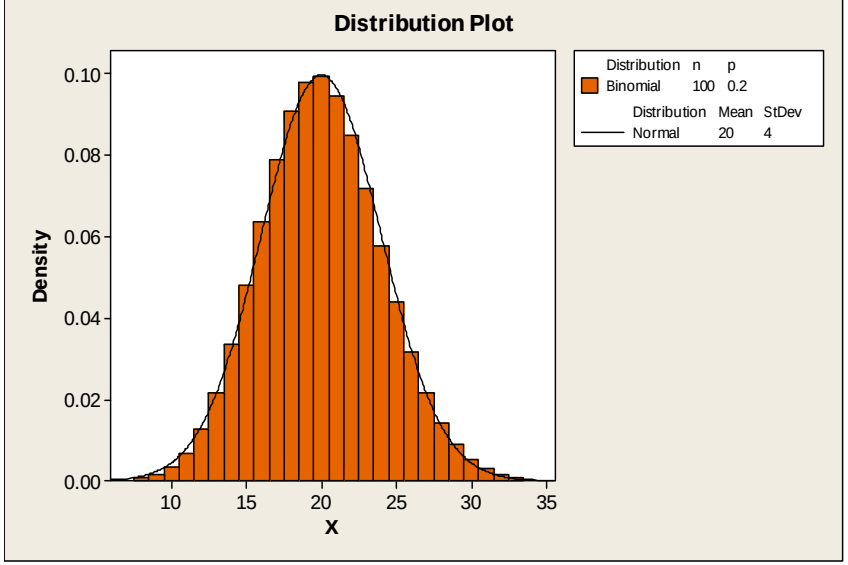
- Adding probabilities for all values from 0 to k for binomial random variable with n, p

- Finding area under curve to the left of k for normal random variable with

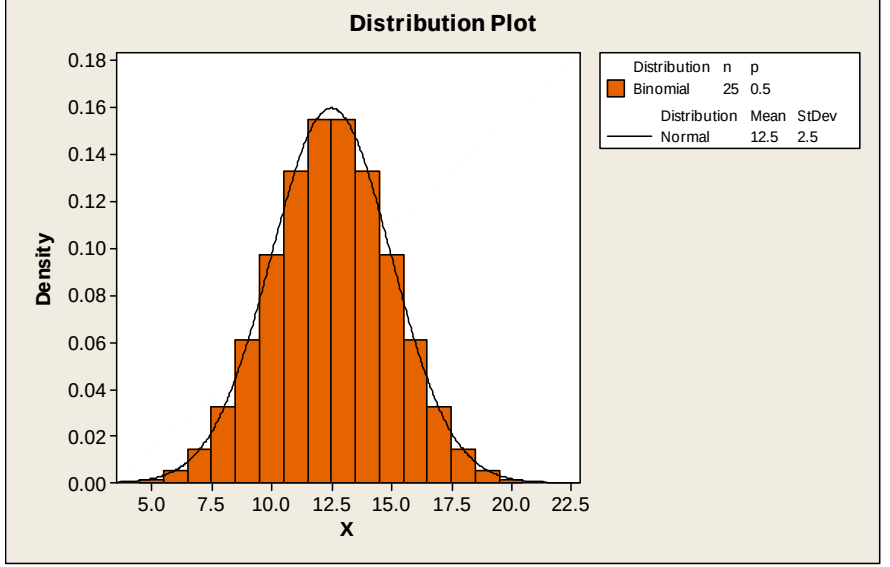
$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Comparing binomial & normal for some values of n and p:

n = 100, p = .2; μ = np = 20, σ = 4



n = 25, p = .5; μ = np = 12.5, σ = 2.5



Shaded rectangles show the binomial probabilities for each value on the x axis; smooth bell-shaped curves show the normal distribution with the same mean and standard deviation as the binomial.

Poll example, we found exact binomial probability:

A poll samples 1000 people from a population with 48% who have a certain opinion. X = number in the *sample* who have that opinion. What is the probability that a *majority* (at least 500) of the *sample* have that opinion? Exact: .109

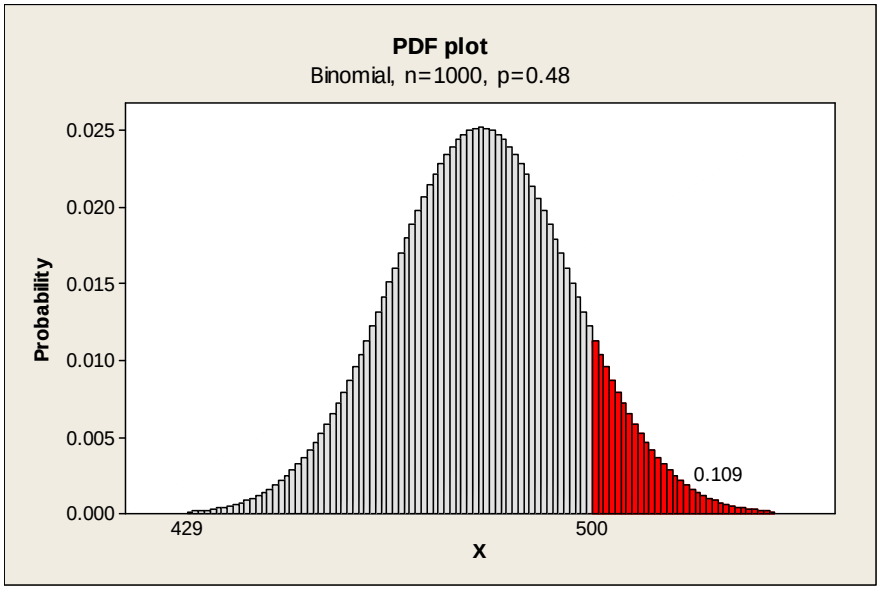
Binomial with $n = 1000$ and $p = .48$, $\mu = 480$ and $\sigma = \sqrt{1000(.48)(.52)} = 15.8$

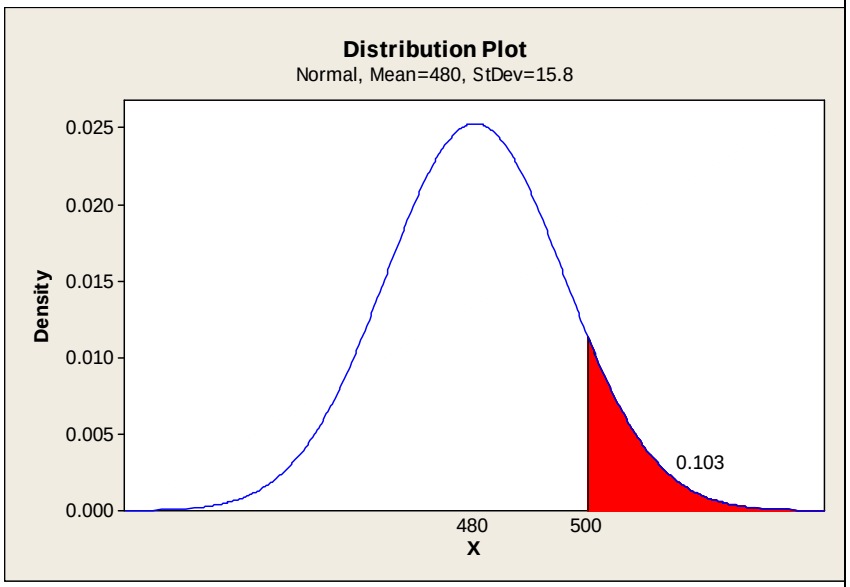
Normal approximation:

$$P(X \geq 500) \approx P(z \geq \frac{500 - 480}{15.8}) = P(z \geq 1.2658) = .103$$

Picture on next page….

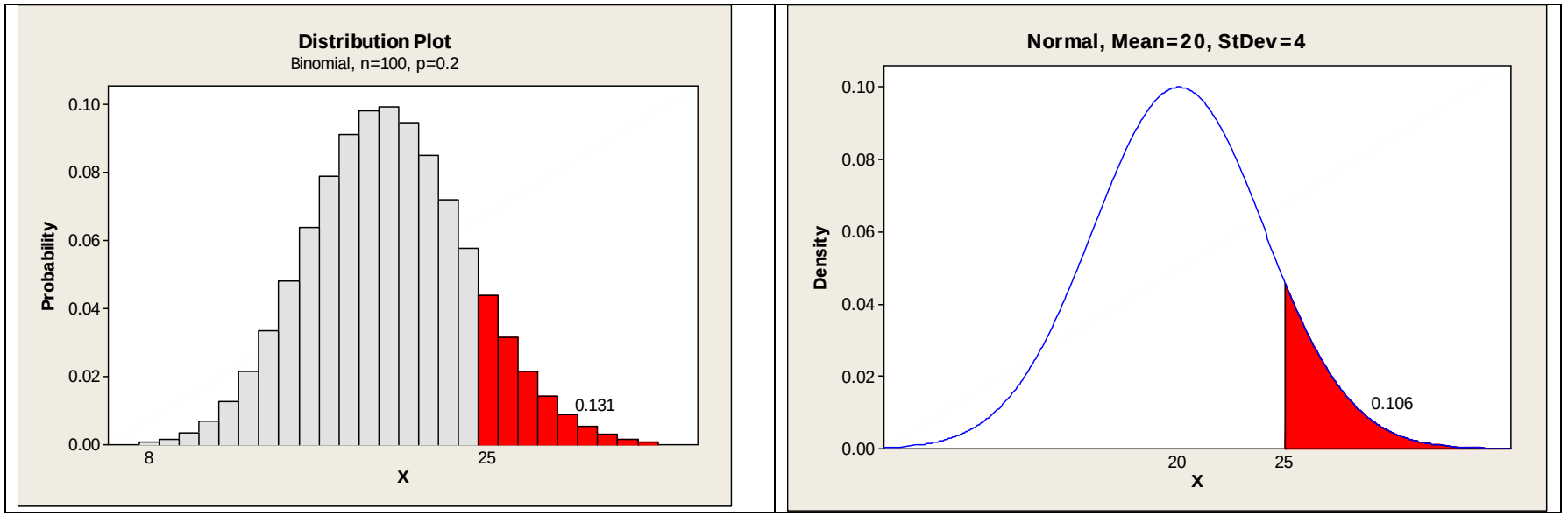
Comparing exact binomial and normal approximation: $n = 1000$ and $p = .48$





CONTINUITY CORRECTION

Example with smaller n (fewer rectangles):
$n = 100, p = .2; \mu = 20, \sigma = 4$



Not very accurate! A more accurate place to start is either 0.5 above or below k, depending on the desired probability. Note that binomial rectangle starts at 24.5, not at 25.

Ex: $n = 100$ and $p = .2$, probability of at least 25 successes:

Exact binomial probability of at least 25 successes is **0.1313**.

Find P(X > 24.5) for normal X with $\mu = 20$ and $\sigma = 4$.

Why? Normal $P(X \geq 25) = 0.1056$; but $P(X \geq \mathbf{24.5}) = \mathbf{0.1303}$.

In general for smallish n, normal approximation of binomial:

$P(X \leq k) \approx P(z \leq \frac{k + .5 - np}{\sqrt{np(1-p)}})$	(Start at upper end of k rectangle)
$P(X \geq k) \approx P(z \geq \frac{k - .5 - np}{\sqrt{np(1-p)}})$	(Start at lower end of k rectangle)