CSSS/SOC/STAT 321
Case-Based Statistics I

Random Variables & Probability Distributions II: Continuous Distributions

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle

Continuous Random Variables

Sor far, our discussion of probability has been restricted to *discrete* variables

But what about continuous random variables like:

1. Height of a child. We can measure height really precisely with the right equipment.
2. Time until the next bus. We can split a second finer and finer.
3. Exchange rate. How much the dollar is worth in euros.
4. Gross Domestic Product. Total value of all goods and services.

Can't even list all the outcomes of these variable, so we can't list the probability of each outcome.

In general, we won't see repitition of the same exact value ever.
All frequencies are 0 or 1.

Example: Waiting for the train

Suppose your city has a subway that runs very regularly.
Every ten minutes there is a train.

Like most subway riders, you show up at the subway unaware of the scheduled time for the next train.

How long will your wait for the next train be in minutes?

Call you wait X:

X is continuous; you can chop it into tiny fractions of a second.

X is also rigidly bounded. It can't be less than 0, or more than 10.

Example: Waiting for the train

X lies somewhere between 0 and 10 minutes.

What is the probability that X is some particular value?

For exmaple, what is the probability that the train will arrive in exactly 3 minutes 25.00000000. . . seconds?

Example: Waiting for the train

X lies somewhere between 0 and 10 minutes.

What is the probability that X is some particular value?

For exmaple, what is the probability that the train will arrive in exactly 3 minutes 25.00000000. . . seconds?

Zero. That is,

$$\mathrm{P}(X = 3 : 25.0000000\ldots) \approx 0$$

Example: Waiting for the train

Why? There is an uncountable infinity of possible arrival times between 0 and 10 minutes.

If we split the total probability of train arrival ($= 1$) into an infinite number of pieces, each piece will be about 0.

In general, the probability that a continuous variable will take on an exact value is always 0.

(Note that we now refer to $\mathrm{P}(X)$ instead of $\Pr(X)$.
We use a different notation for the probability of continuous variables.)

Example: Waiting for the train

We cannot talk about the probability of specific values of continuous distribution

Instead, focus on the probability that X lies in a specific interval.

For example, what is the probability that the train will arrive at or after 1 minute has passed, but before 5 minutes?

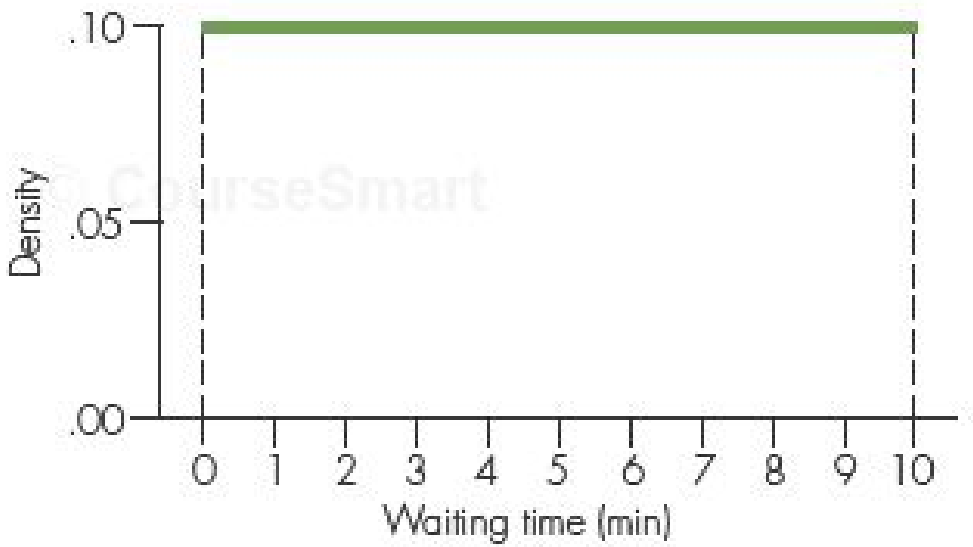
$$\mathrm{P}(1 \leq X < 5) > 0$$

Probabilities over intervals of continuous variables are positive, so we can calculate this. But we need to think a bit about the shape of the distribution

The Uniform Distribution

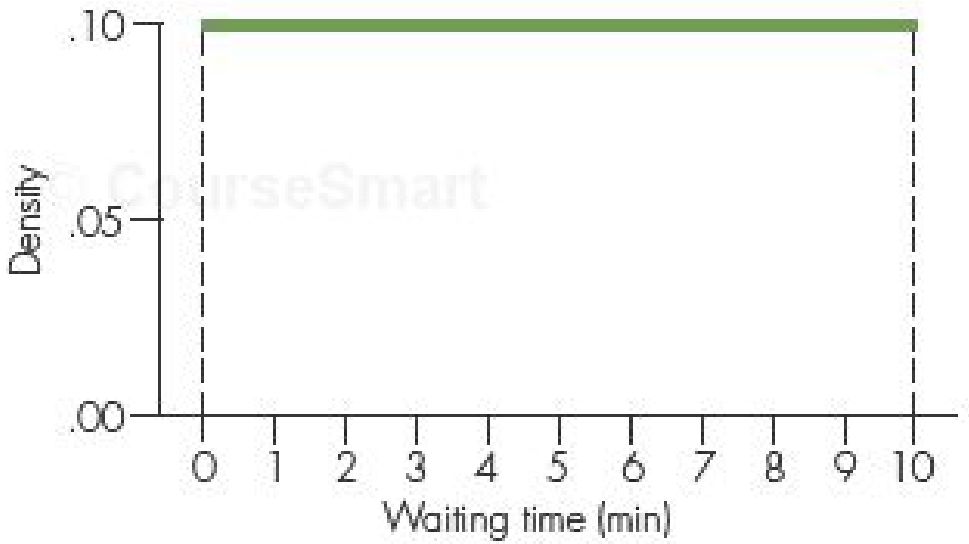
In the train example, there is no reason to consider the train more likely to arrive at any particular moment.

This is a rare case where all of the possible outcomes of a continuous variable are equally, or Uniformly, likely:



This is the uniform distribution's pdf.

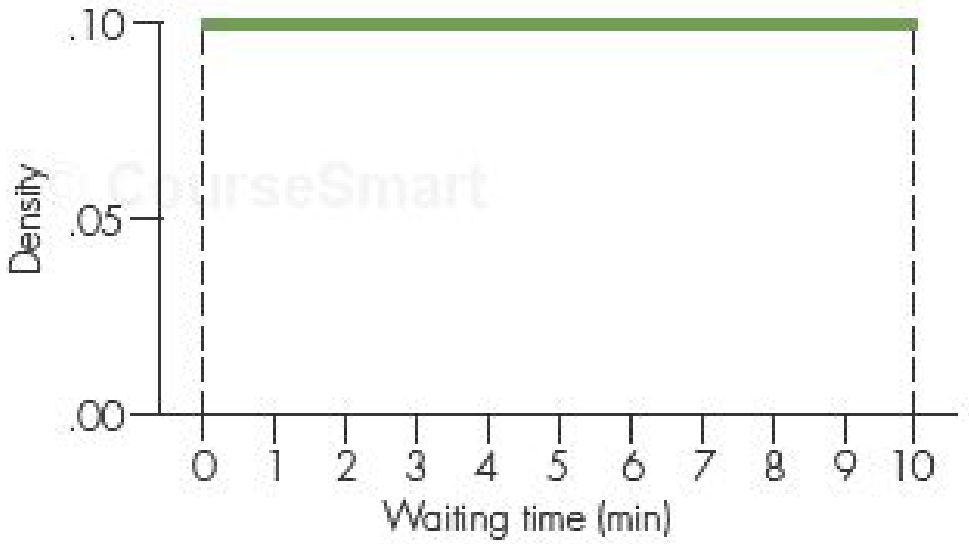
The Uniform Distribution



An unusual, and very simple, continuous distribution

It has two parameters: A, the minimum possible value of x, and B, the maximum possible value

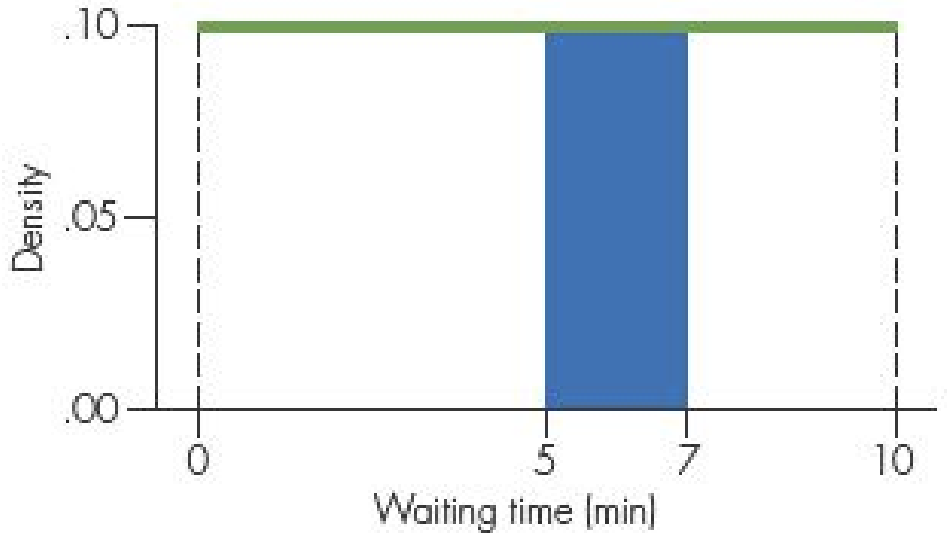
The Uniform Distribution



The pdf of the Uniform:

$$\mathrm{P}(X) = \frac{1}{B - A}$$

The Uniform Distribution



The cdf of the Uniform, over the sub-interval from a to b:

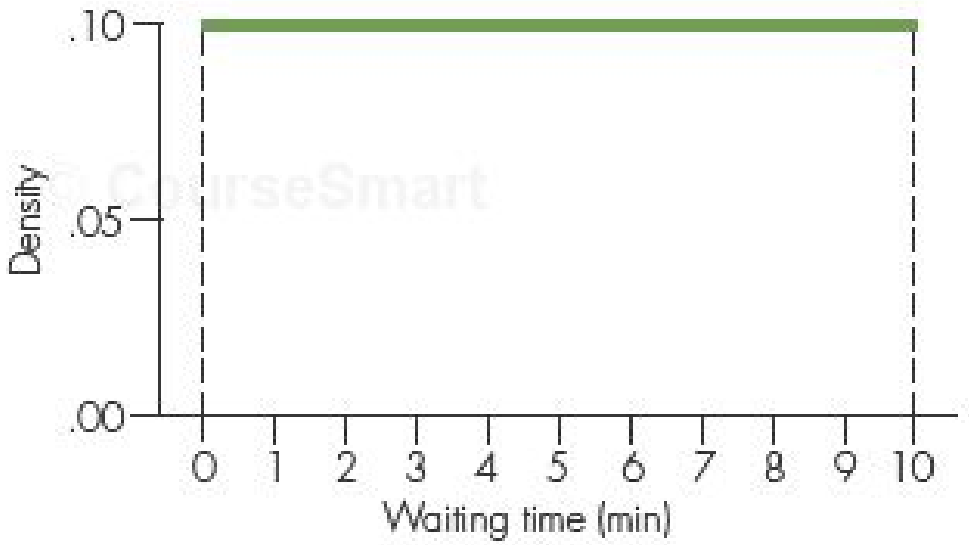
$$\mathrm{P}(a < x \leq b) = \frac{b-a}{B-A}$$

The Uniform Distribution

In our example, the probability the train will arrive between minute 1 and minute 5 is

$$\mathrm{P}(1 \leq X < 5) = \frac{5-1}{10-0} = 0.4$$

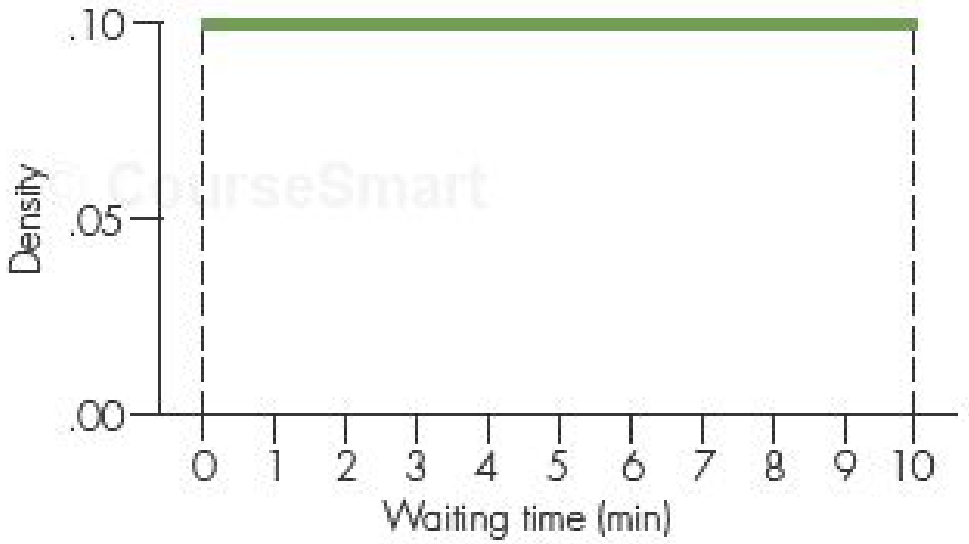
The Uniform Distribution



We can summarize any distribution by its expected value and variance. For the uniform, these are:

$$\mathrm{E}(X) = \frac{1}{2}(a+b)$$

The Uniform Distribution



We can summarize any distribution by its expected value and variance. For the uniform, these are:

$$\begin{aligned} \mathrm{E}(X) &= \frac{1}{2}(a+b) \\ \mathrm{Var}(X) &= \frac{1}{12}(b-a)^2 \end{aligned}$$

CDFs for continuous distributions

Because the uniform distribution is a rectangle, it's easy to calculate the areas that correspond to probabilities for an interval of X

But most continuous distributions follow complex curves

In general, we need calculus to find the CDF of a continuous distribution

$$\mathrm{P}(a \leq x < b) = \int_a^b \mathrm{P}(x)\mathrm{d}x$$

This says the cdf of a continuous distribution is the integral of the pdf

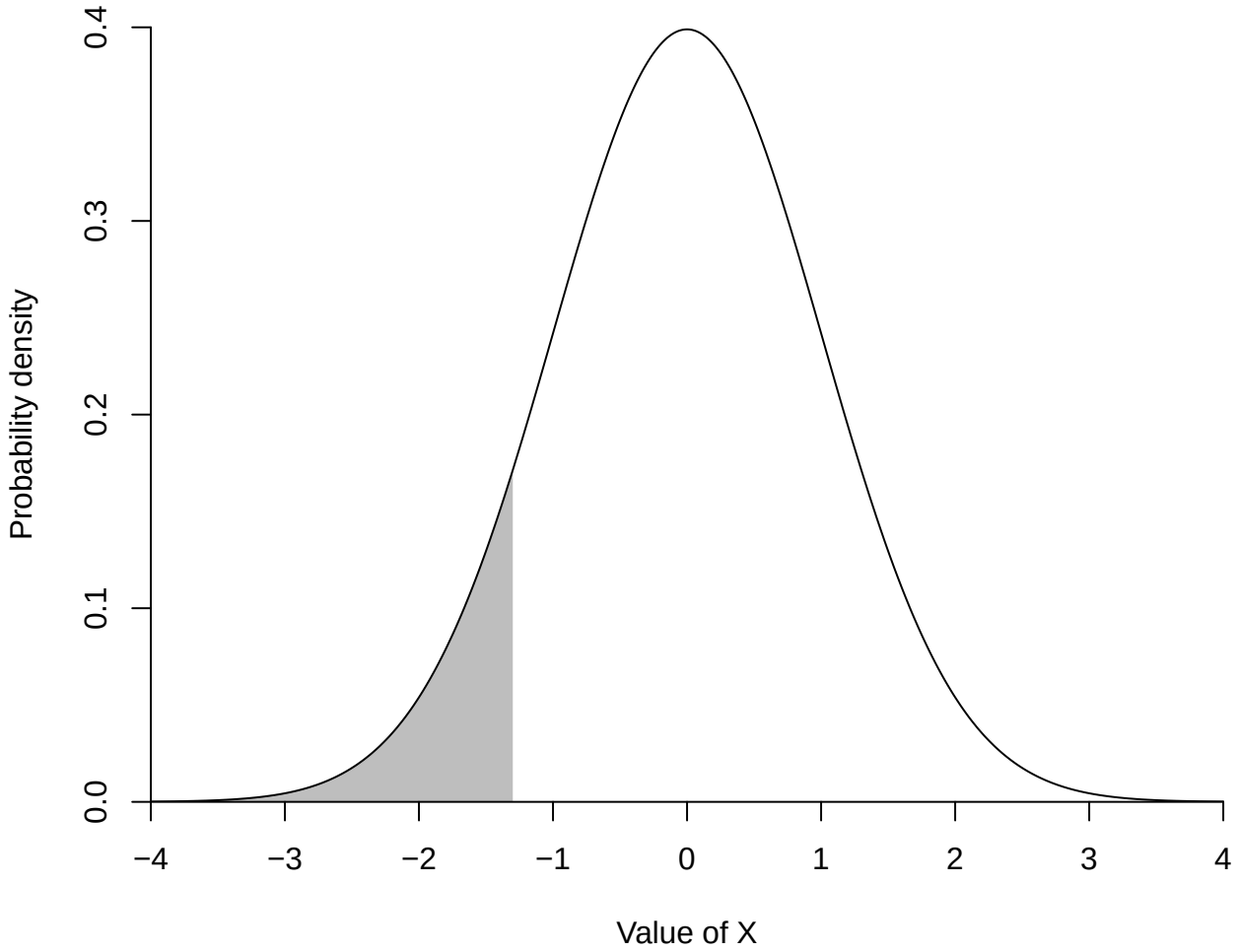
The integral gives us the *area* under the probability curve between a and b

CDFs for continuous distributions

You won't be required to do any calculus yourself;
`R` or `Stata` will do this easily

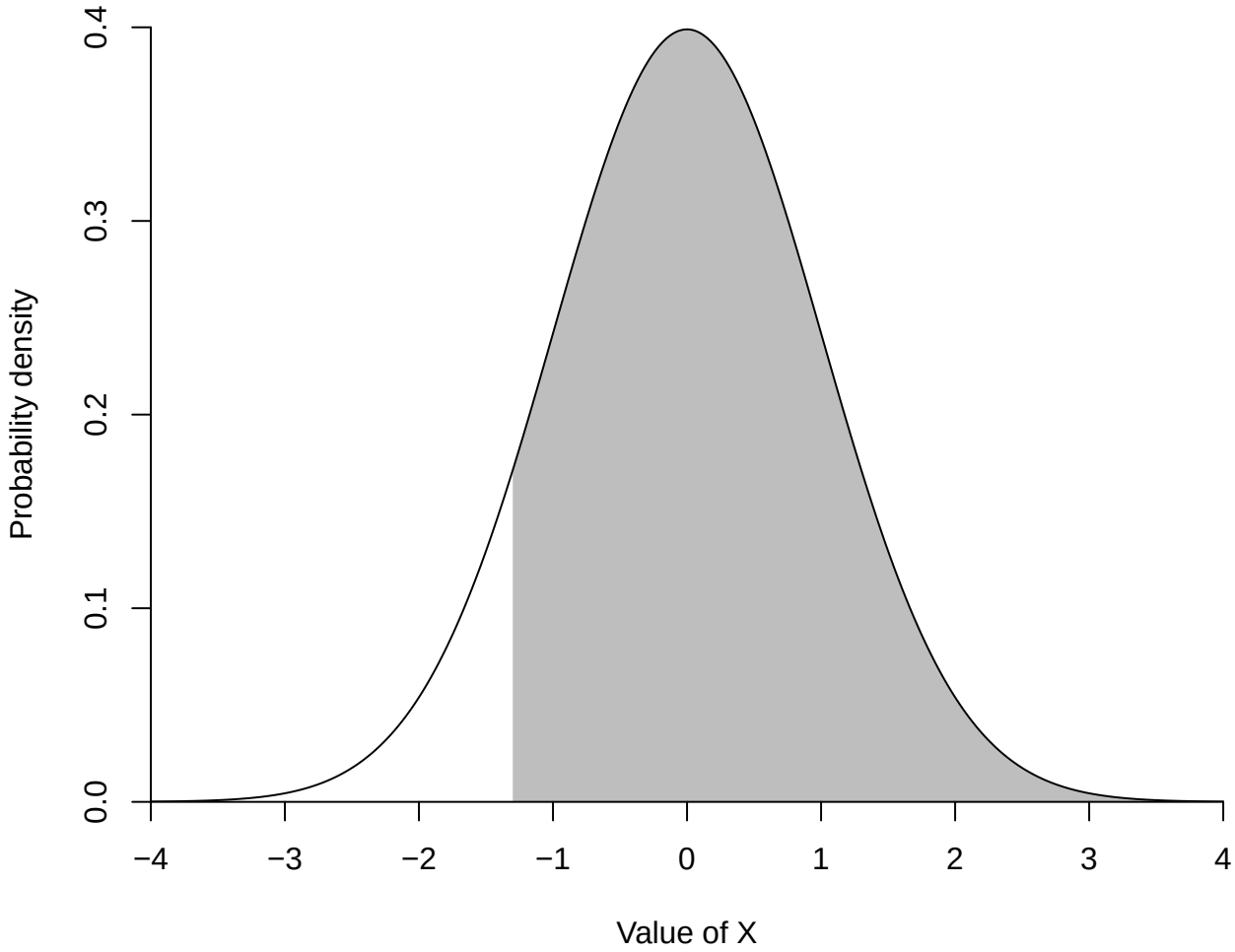
But even with a tool to solve the calculus problem,
we still need to think about what quantity we want to calculate

In this context, there are three helpful rules of continuous CDFs that apply to most continuous distributions



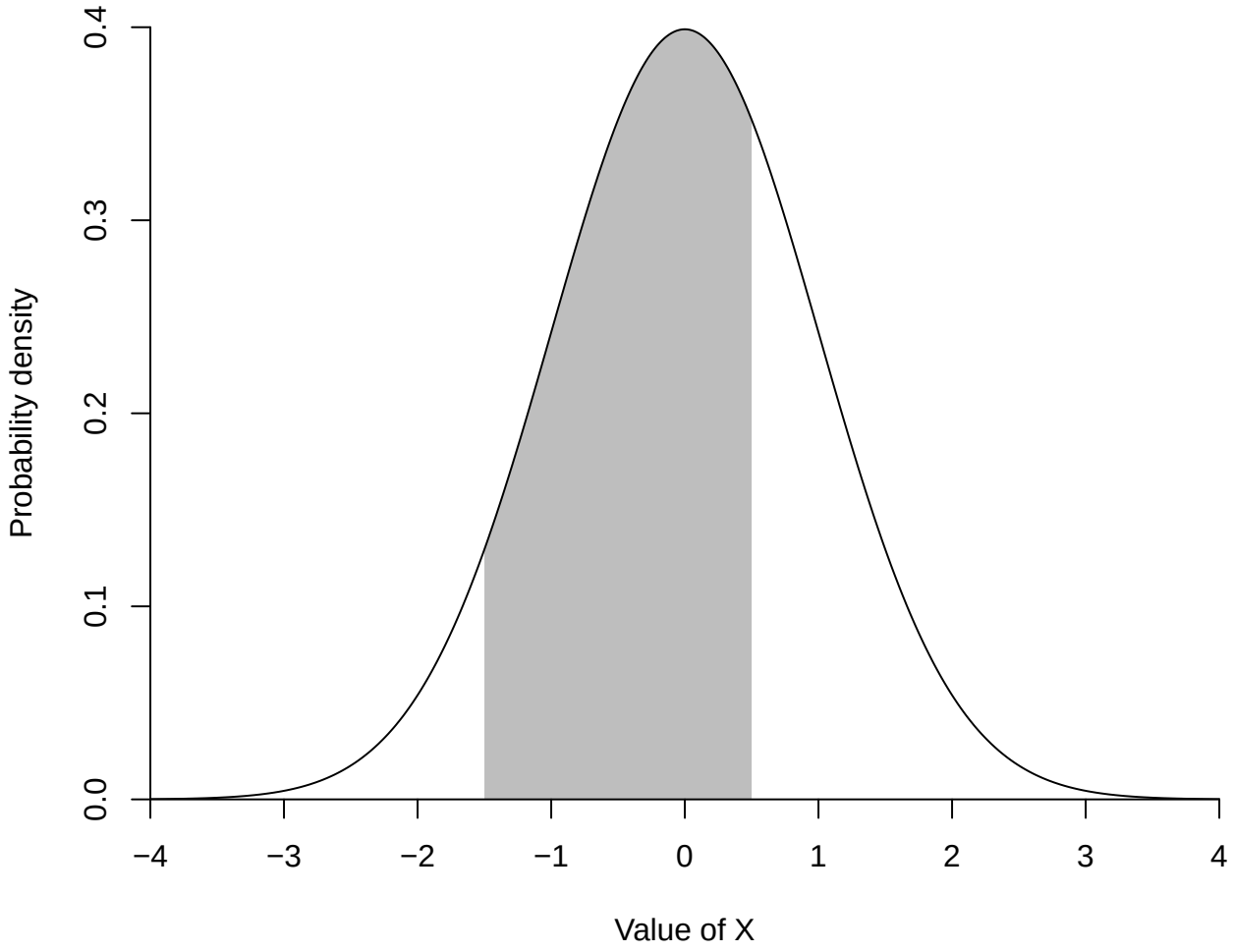
We can often use one tail to calculate the other tail

Suppose we wanted to know $\mathrm{P}(X > -1.3)$, but all we knew was $\mathrm{P}(X < -1.3) = 0.097$, or the area in gray



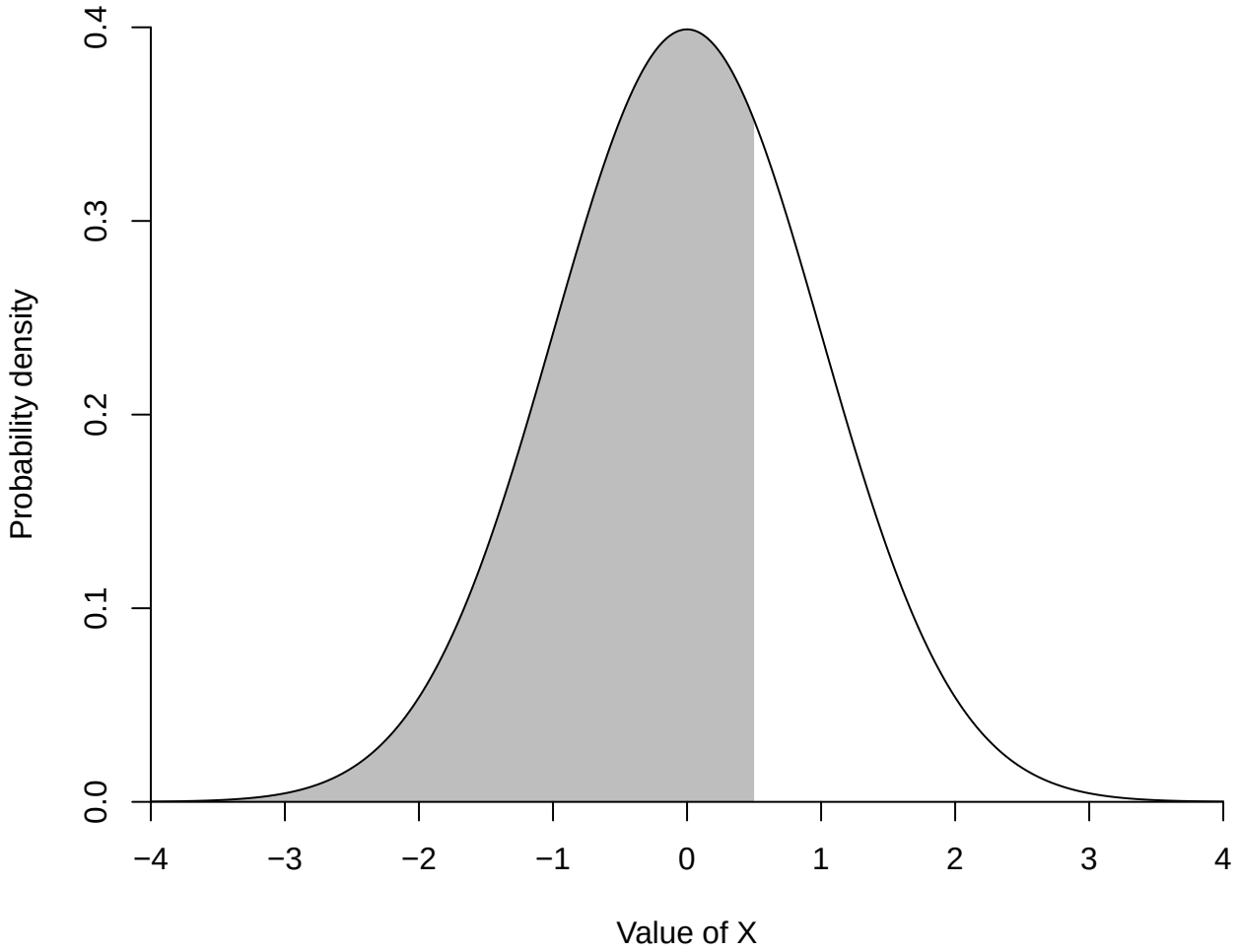
Rule 1:
$\mathrm{P}(X > a) = 1 - \mathrm{P}(X \leq a)$

So $\mathrm{P}(X > -1.3) = 1 - 0.097 = 0.903$, or the new area in gray



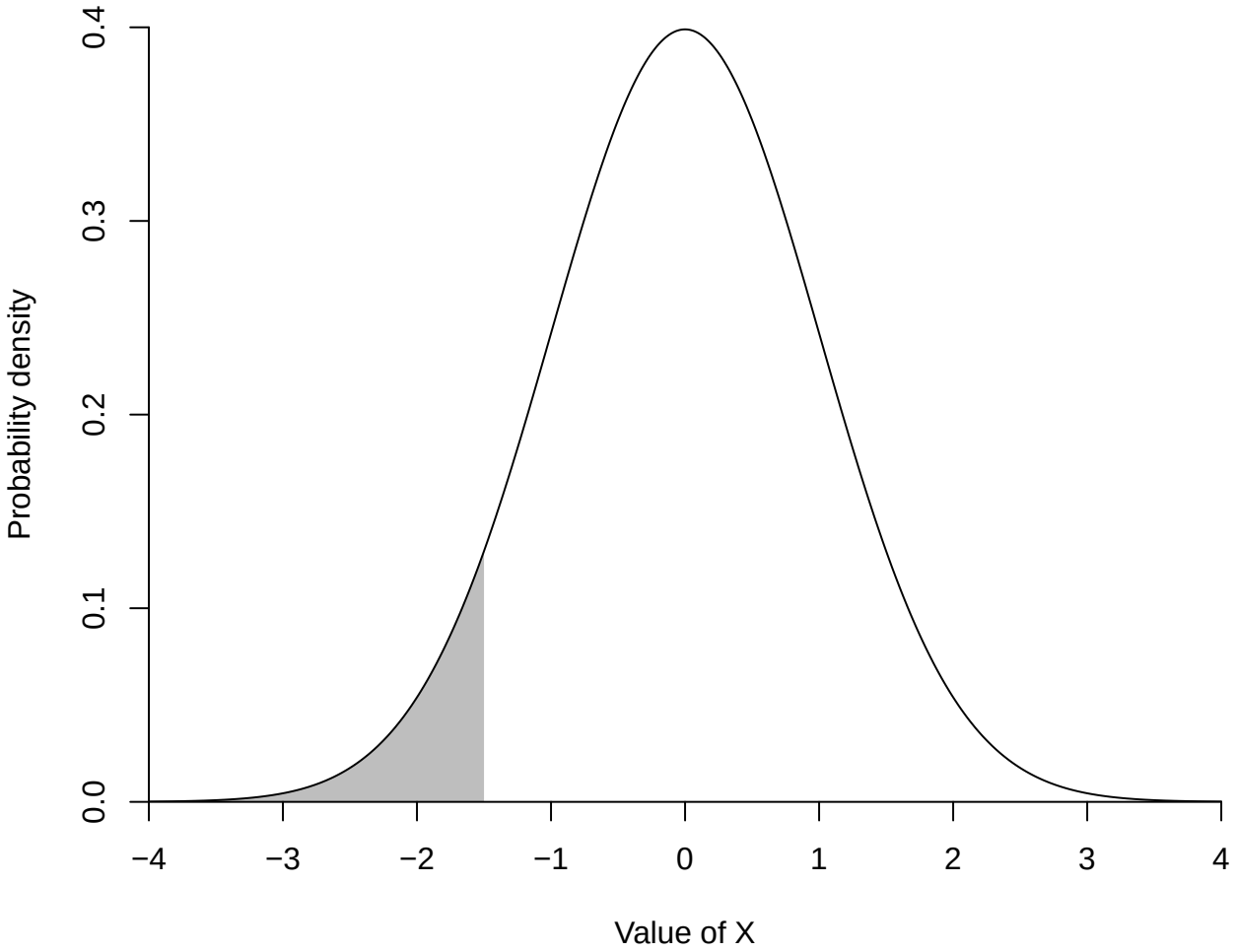
What if we wanted to know an area in the middle of the curve?

That is, what if we wanted to know $\mathrm{P}(a \leq X < b)$, or in this case, $\mathrm{P}(-1.5 < X \leq 0.5)$?

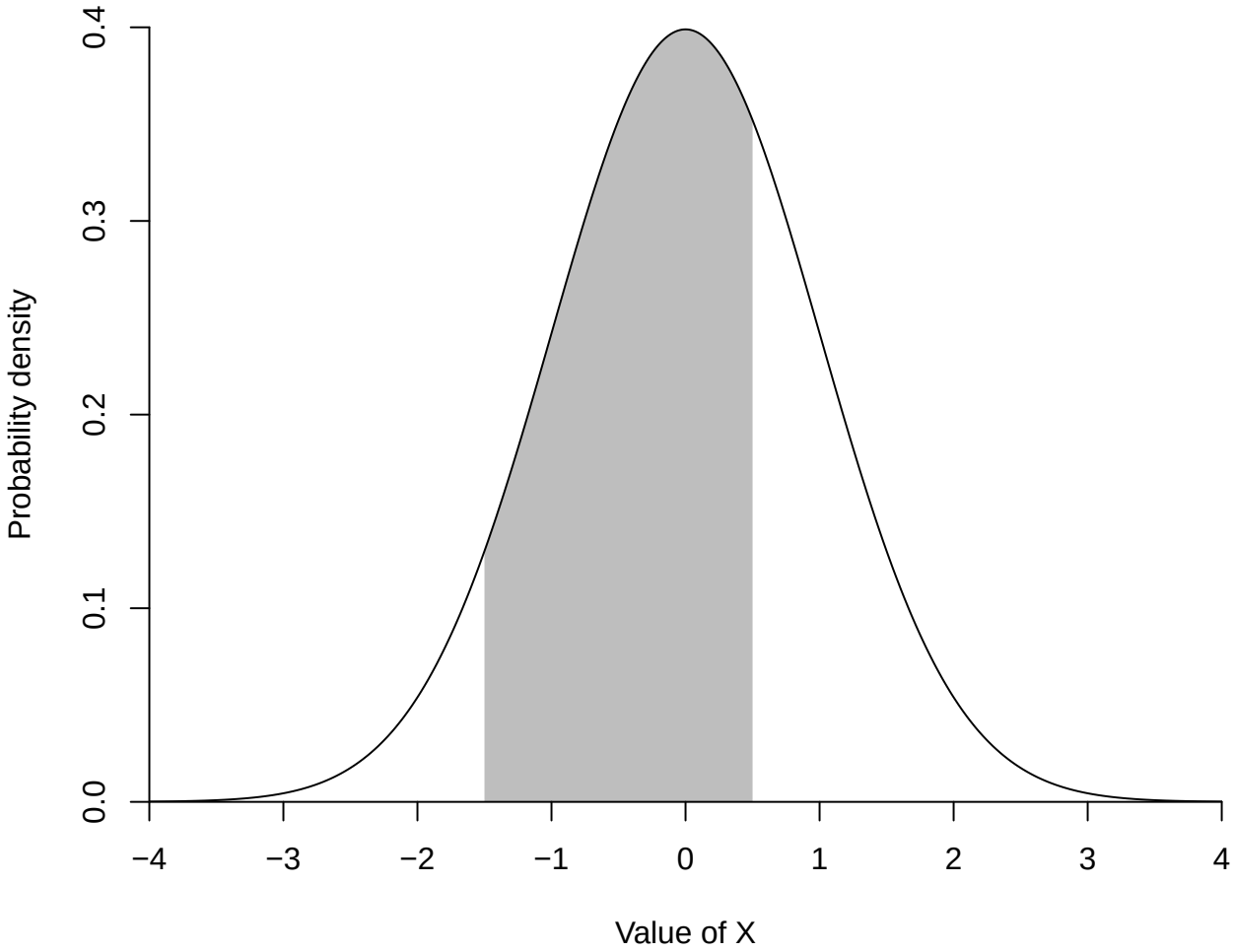


Rule 2: $\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a)$

That is, start with the whole area from $-\infty$ to $b\ldots$

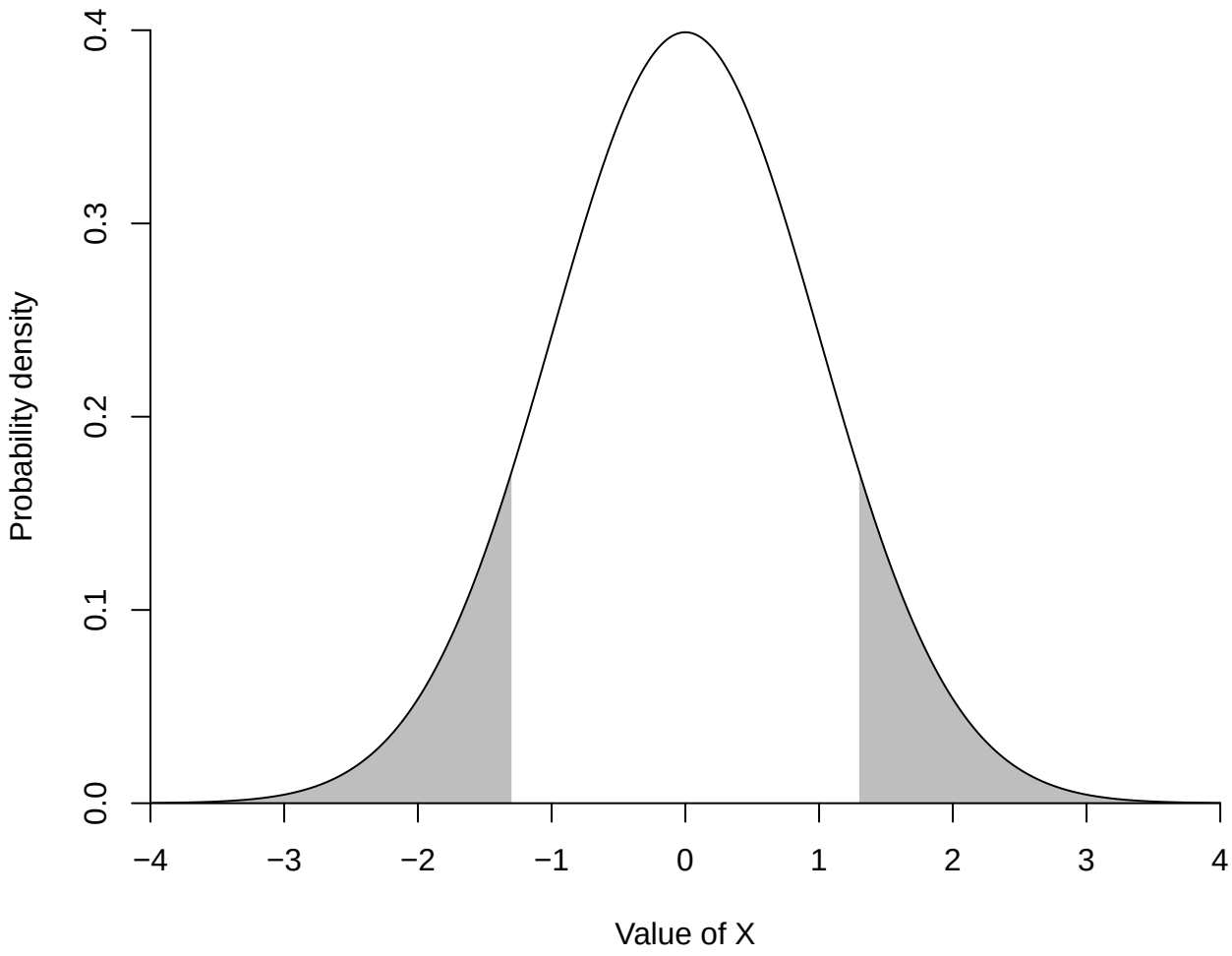


and subtract off the area from $-\infty$ to a



leaving just the part we want!

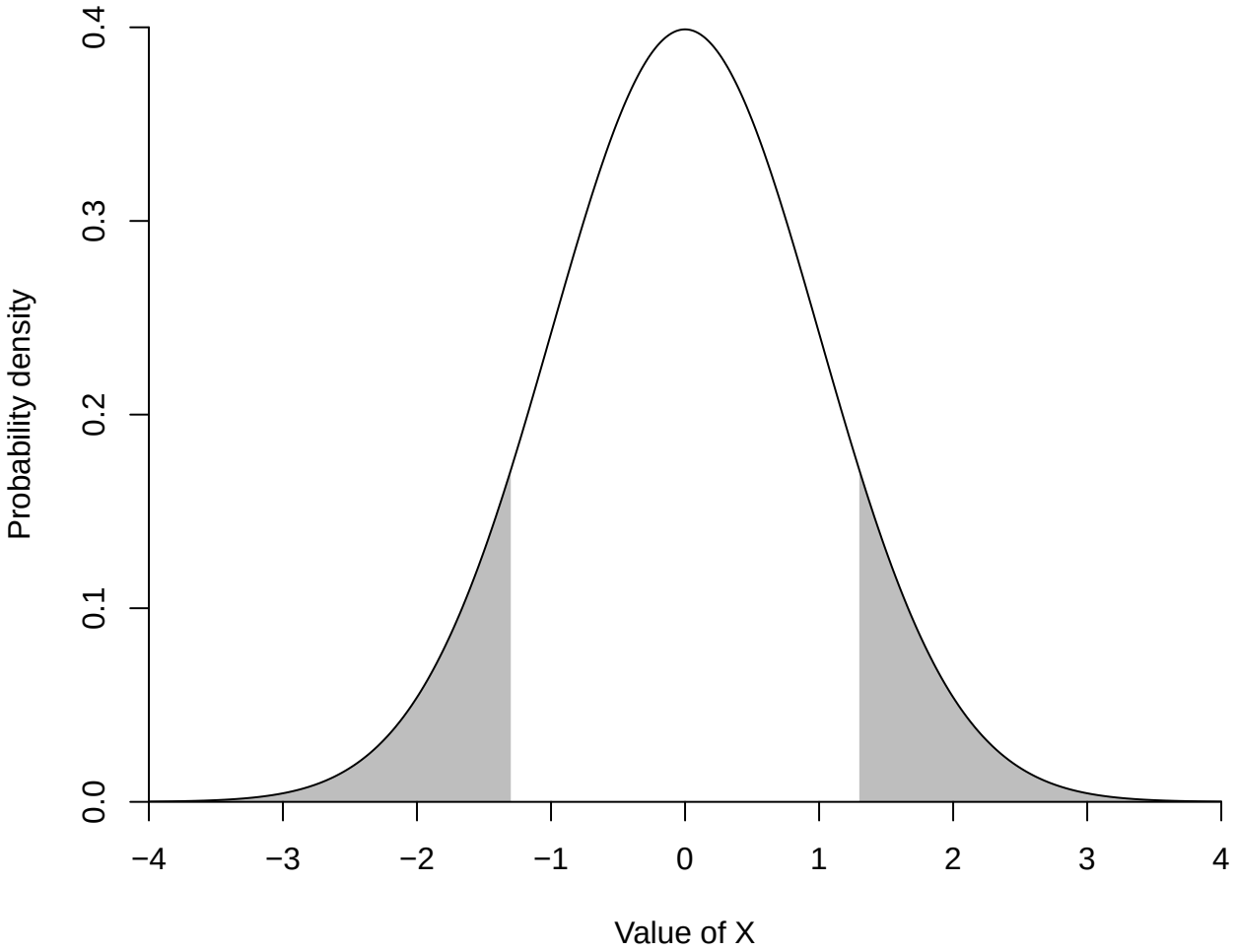
$$\mathrm{P}(-1.5 < X \leq 0.5)$$
$$\mathrm{P}(X \leq 0.5) - \mathrm{P}(X \leq -1.5)$$
$$0.691 - 0.067$$
$$0.625$$



The above rules always work. For symmetric distributions, we have a third rule as well

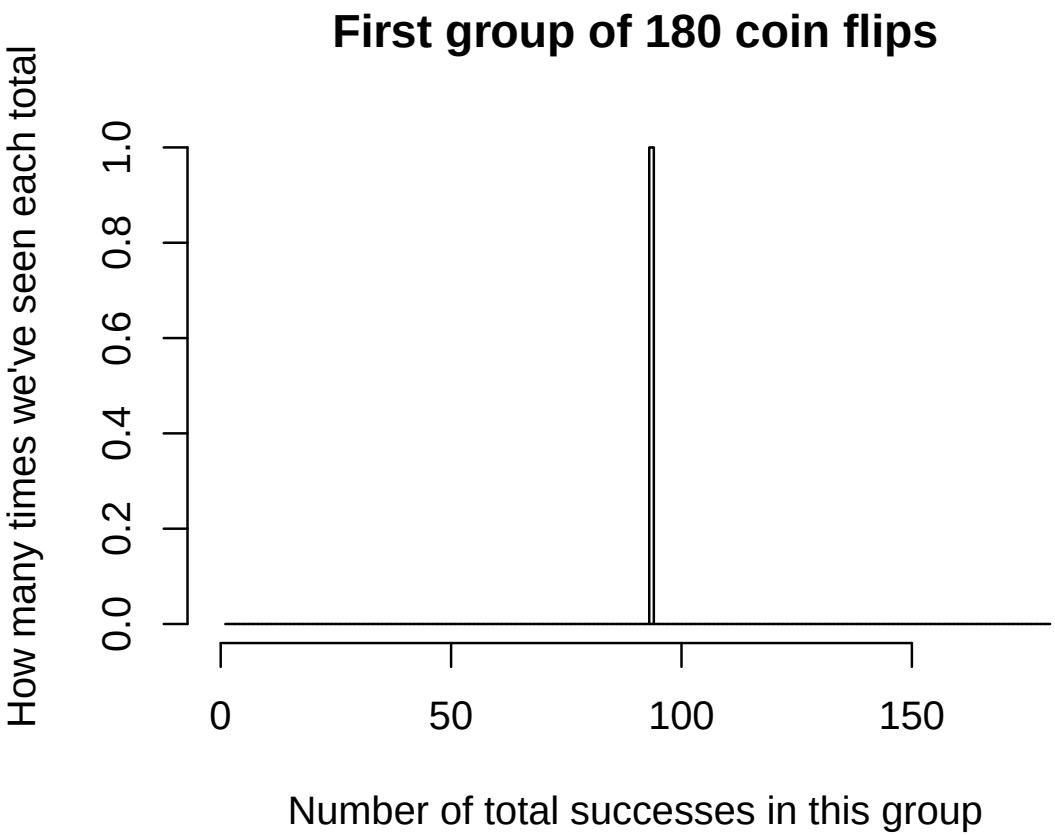
Here we have $\mathrm{P}(X \leq -1.3)$ and $\mathrm{P}(X \geq 1.3)$

The two areas at left are the same, and so must these probabilities



Rule 3: For symmetric distributions,
$\mathrm{P}(X \leq \mu - d) =$
$\mathrm{P}(X \geq \mu + d)$

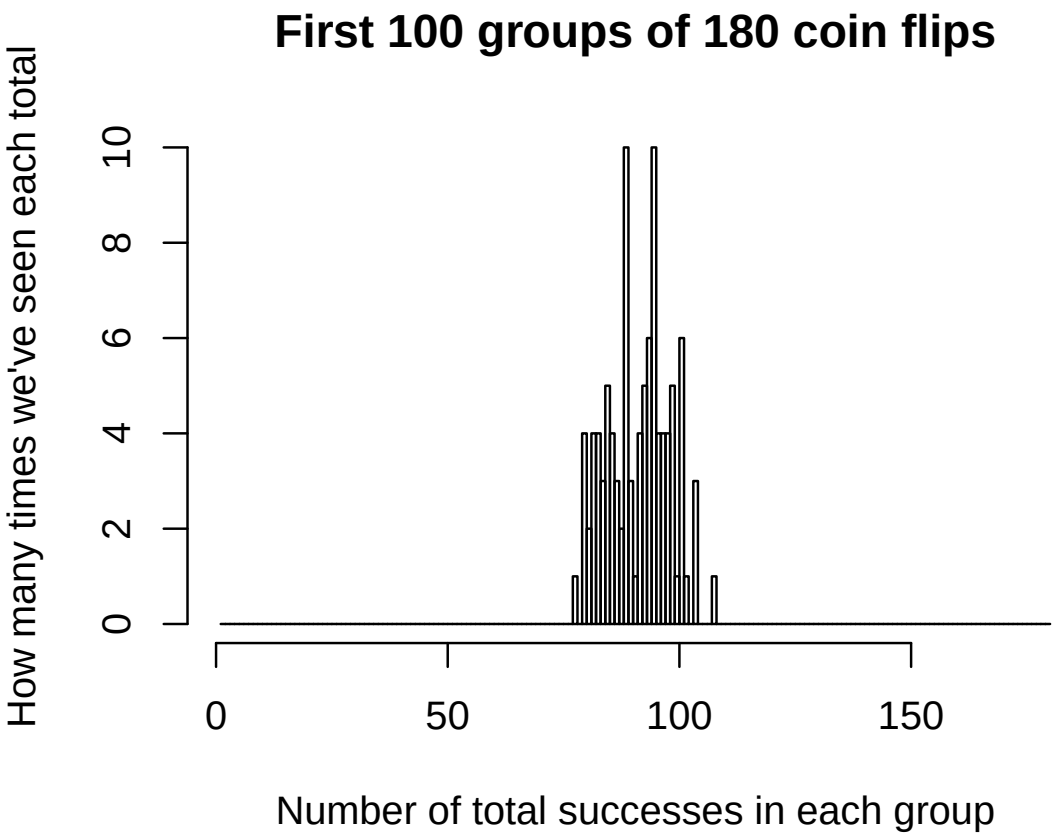
For any symmetric distribution, tails equally far from the mean have the same area, and hence values as extreme as $\mu \pm d$ are equally likely



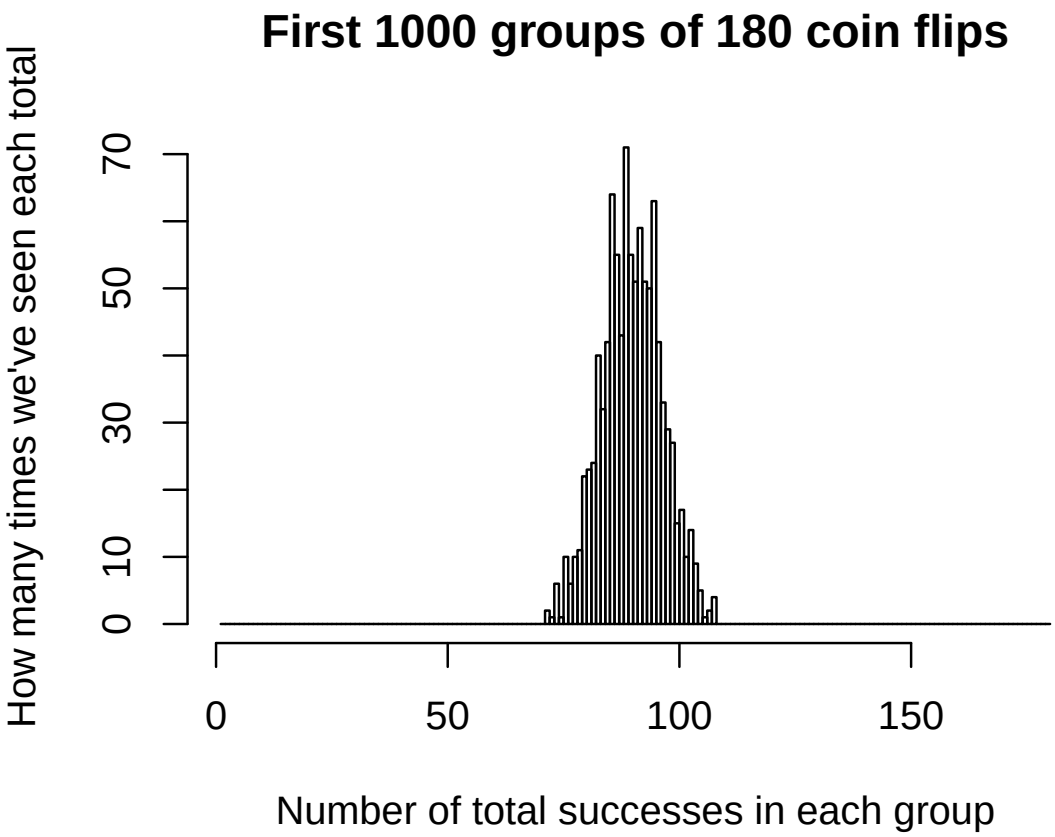
A quick detour back to discrete distributions. . .

Let's go back to our "sum of coin flips" binomial distribution, but with 180 coins to flip in each trial

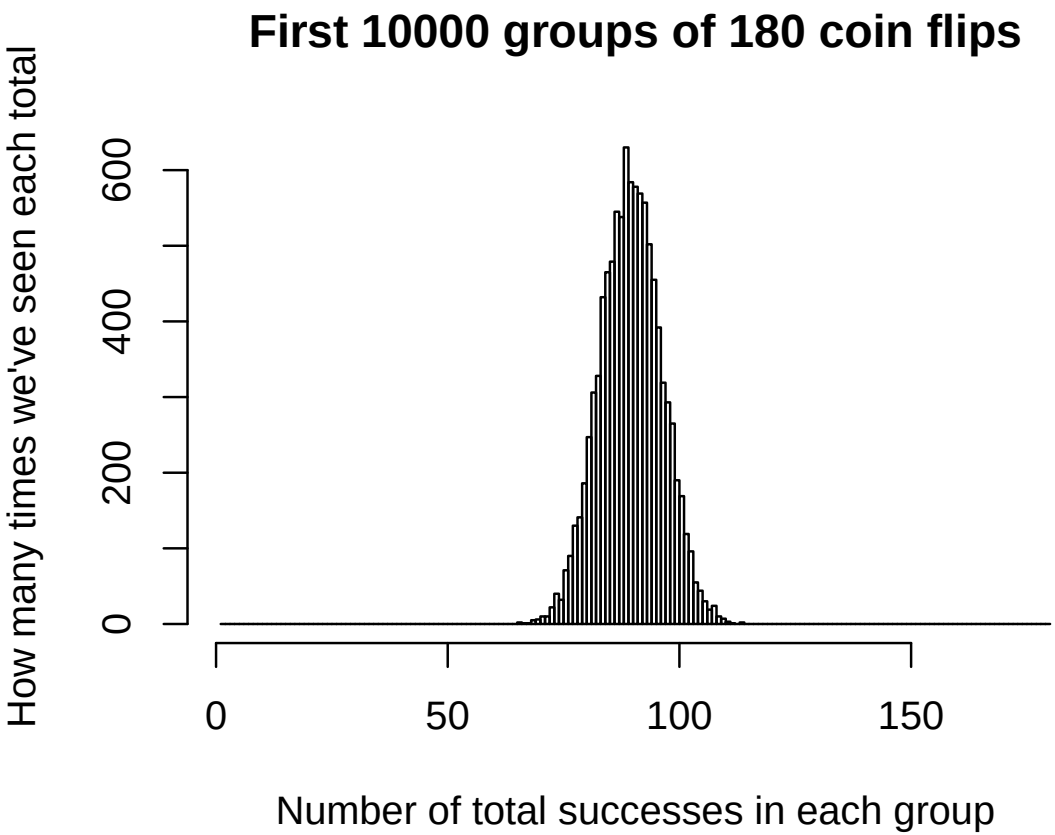
On our first flip we might get 94 heads out of 180 trials



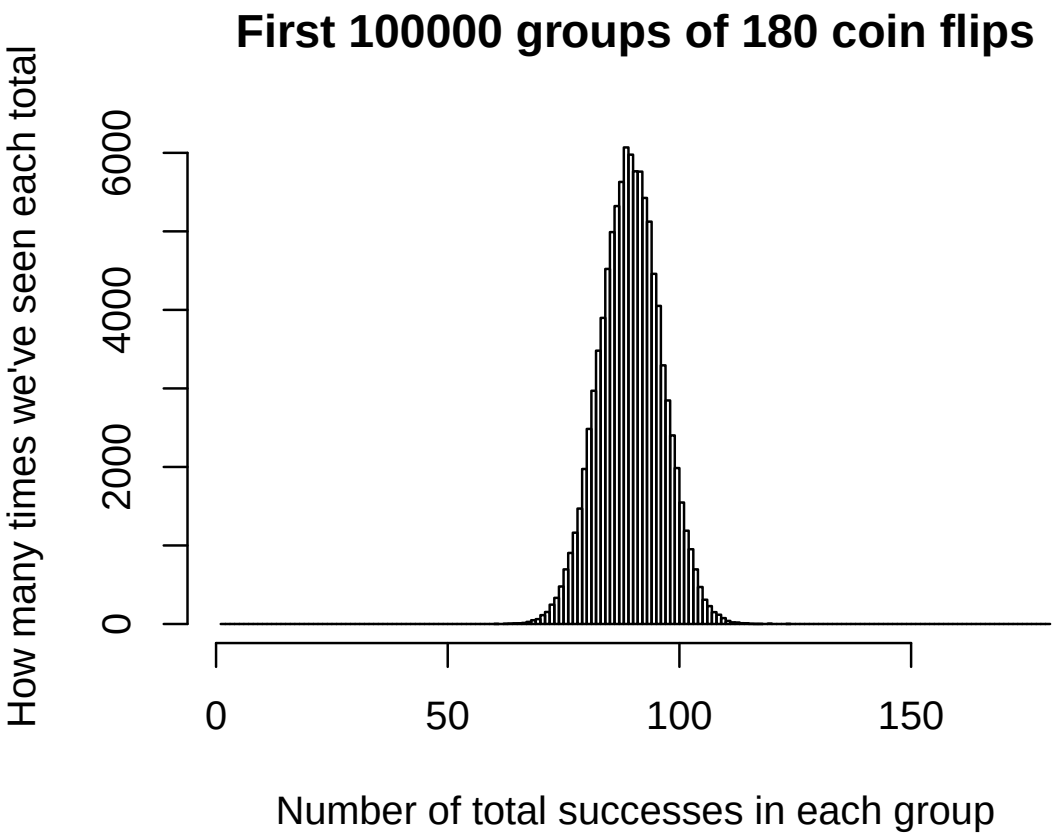
Skipping ahead, here is our distribution after 100 sets of flips. . .



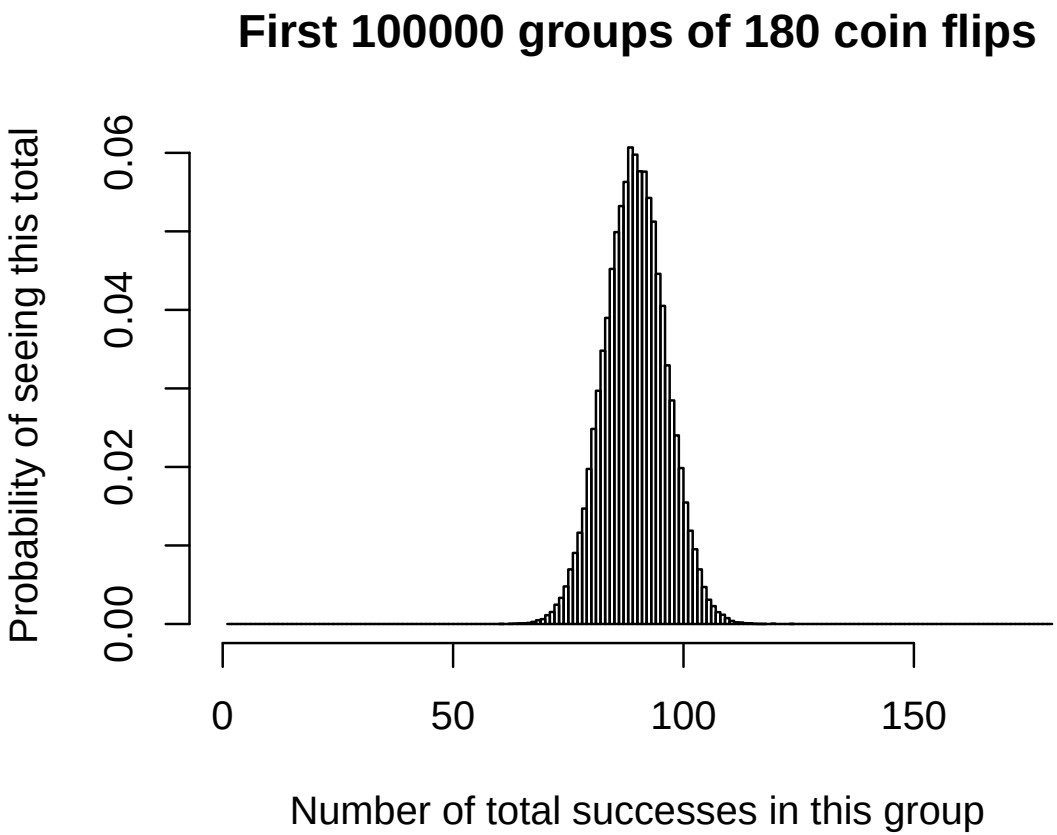
Here is our distribution after 1000 sets of flips...



Here is our distribution after 10,000 sets of flips...



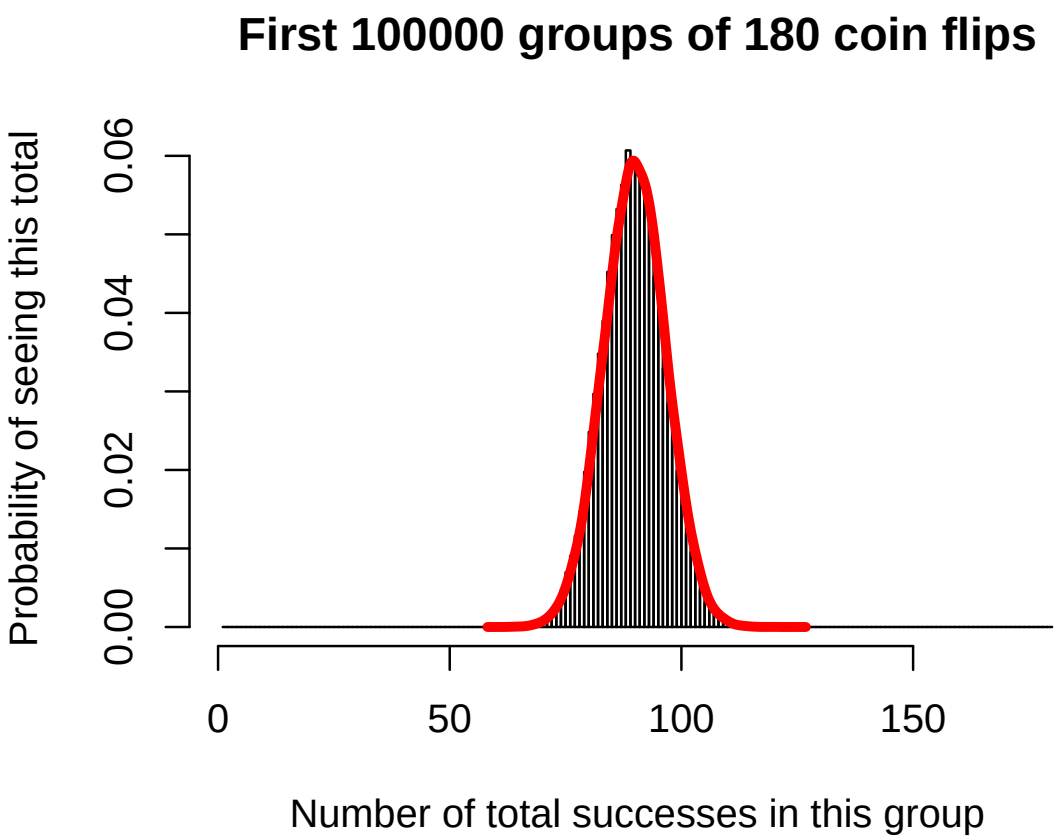
Here is our distribution after 100,000 sets of flips...



Once again, we can divide by the total number of simulation runs to get probabilities

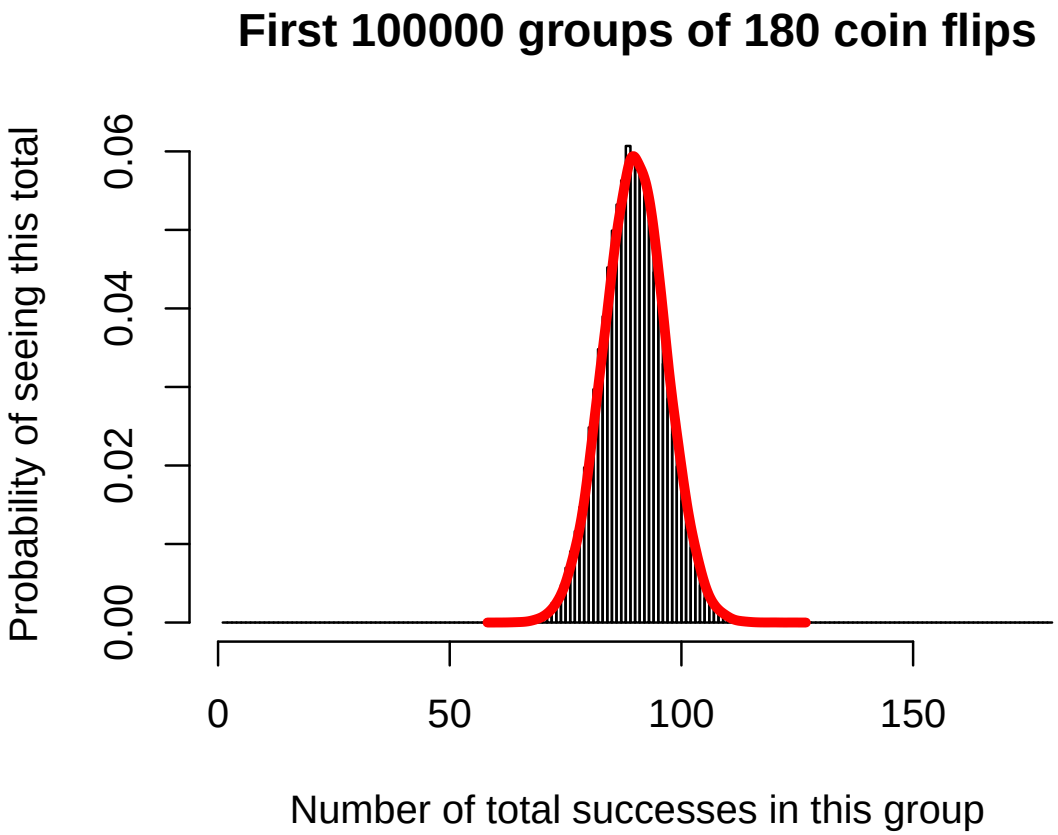
Notice that extreme values are vanishingly rare

And that a bell is smoothly traced out by the histogram's bars



If we trace out this smooth curve, we get the probability density in red

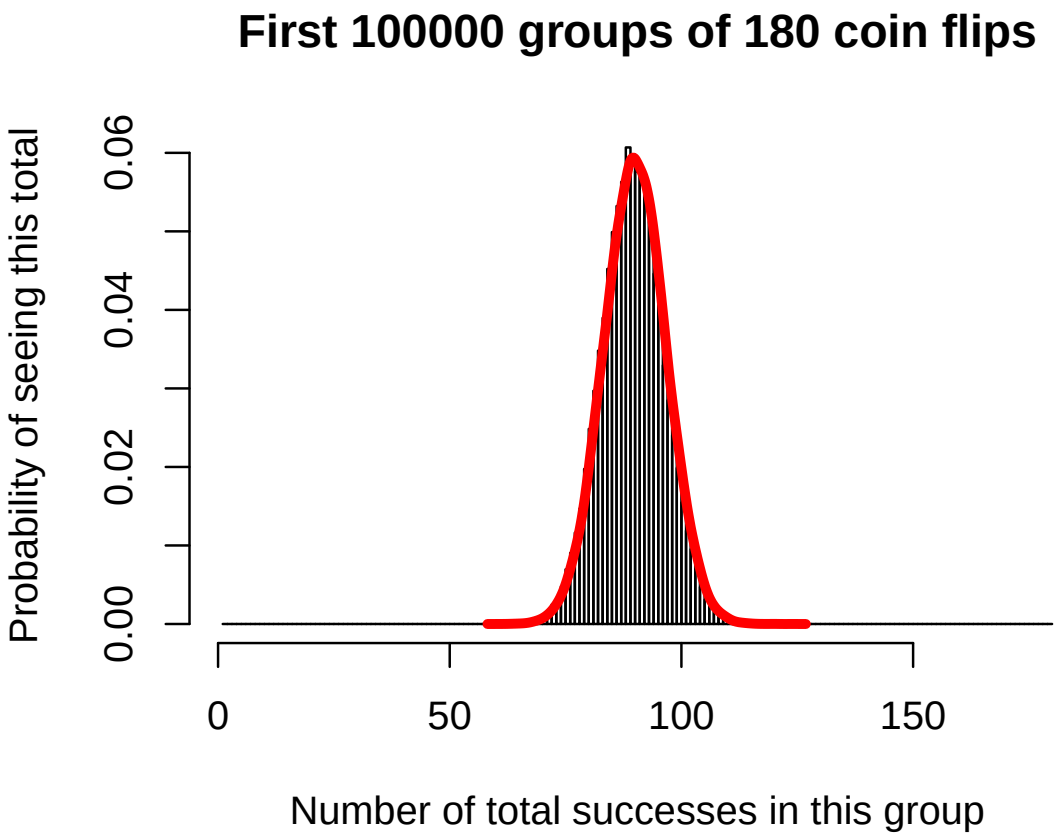
We've discovered something fundamental: if you add together many independent random variables, even RVs as simple as coin flips, their sum approximates a bell curve



Central Limit Theorem

The more independent random variables we sum together, the more closely their sum approximates a Normal distribution.

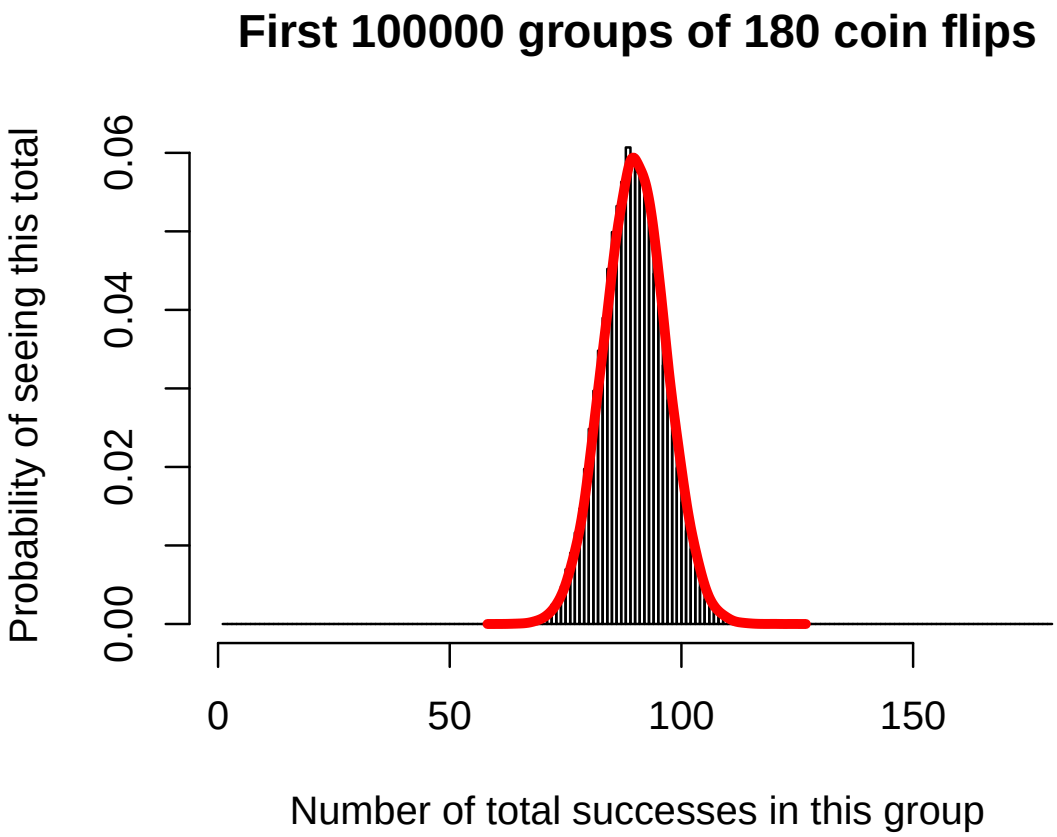
Example: if we ask everyone in America if they have a job, and add their responses together, we get the unemployment rate. The unemployment rate may be approximately Normally distributed



Central Limit Theorem

The more independent random variables we sum together, the more closely their sum approximates a Normal distribution.

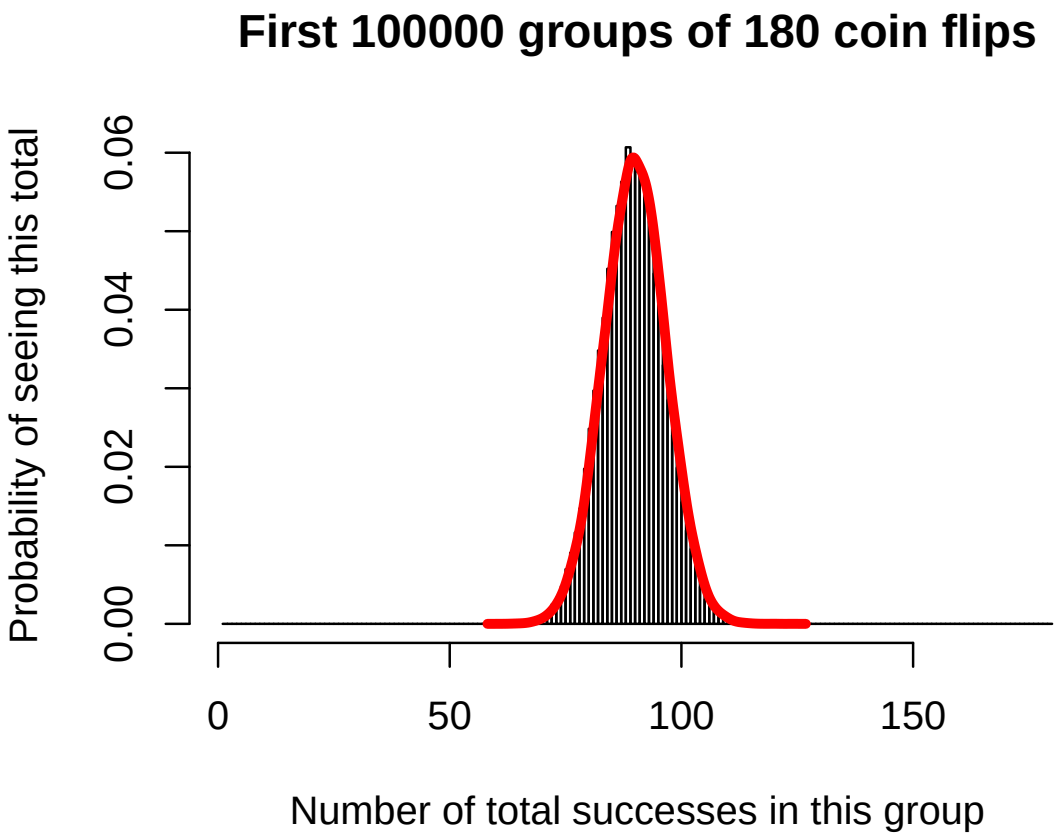
Notice that the Normal distribution only holds in this special case



Central Limit Theorem

The more independent random variables we sum together, the more closely their sum approximates a Normal distribution.

It is a distribution named *Normal, not the "normal" distribution you see in the world*



Central Limit Theorem

The more independent random variables we sum together, the more closely their sum approximates a Normal distribution.

It's also called the Gaussian distribution, and is just one possible distributions out of thousands known to statisticians

But it's very useful in intro statistics!

The Normal Distribution

Suppose we have a large number of variables with unknown distributions

Suppose further they are *independent*; ie, uncorrelated with each other

Let us call these variables $x_{1i}, x_{2i}, x_{3i}, \ldots, x_{ki}$

They might be how much each American i spends on each product & service k for sale in the economy

Now suppose we add together the spending of each American to create X_i, how much each American spends on everything combined.

What is the distribution of X?

The Normal Distribution

According to the Central Limit Theorem, as $k \to \infty$,
X will follow the Normal distribution:

$$\mathrm{P}(X|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(X-\mu)^2}{2\sigma^2}\right]$$

The Normal Distribution

According to the Central Limit Theorem, as $k \to \infty$,
X will follow the Normal distribution:

$$\mathrm{P}(X|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(X-\mu)^2}{2\sigma^2}\right]$$

$$\mathrm{E}(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$$

The Normal Distribution

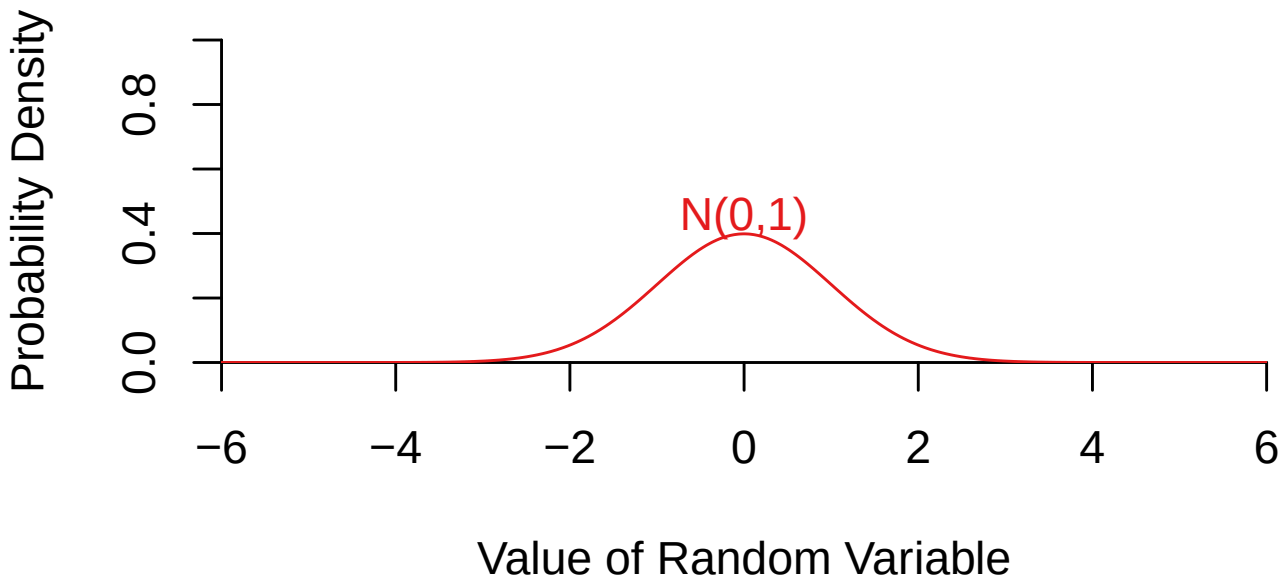
According to the Central Limit Theorem, as $k \to \infty$, X will follow the Normal distribution:

$$\mathrm{P}(X|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(X-\mu)^2}{2\sigma^2}\right]$$

$$\mathrm{E}(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$$

The Normal distribution is continuous and symmetric, with positive probability on X from $-\infty$ to ∞

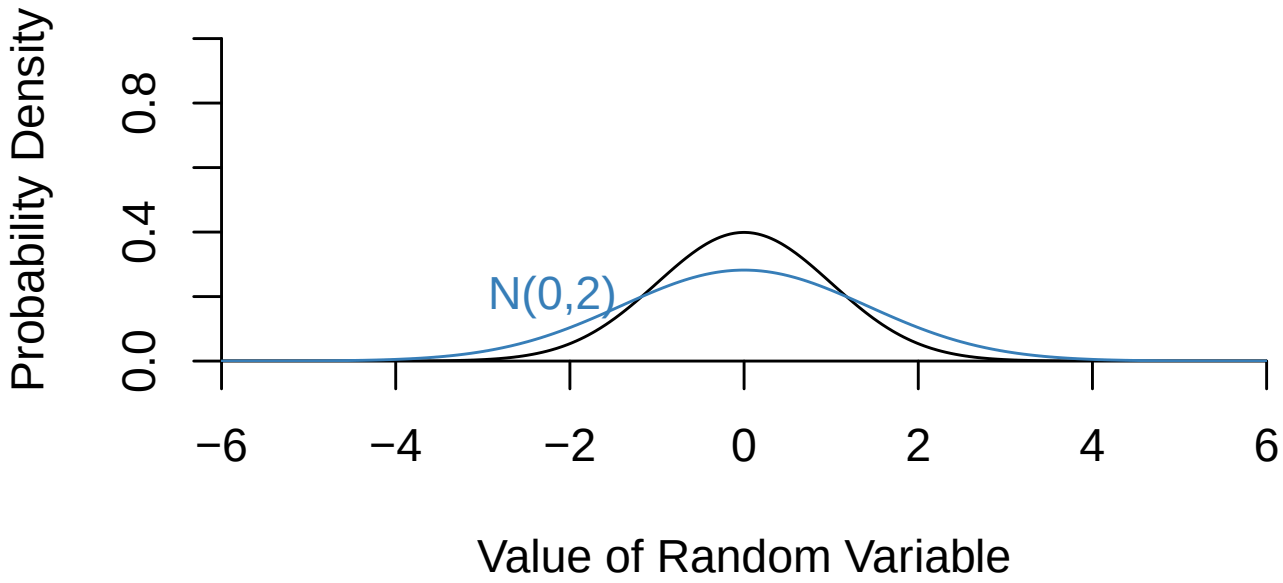
Examples of the Normal Distribution



This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

Known as the Standard Normal. Also the Bell Curve.

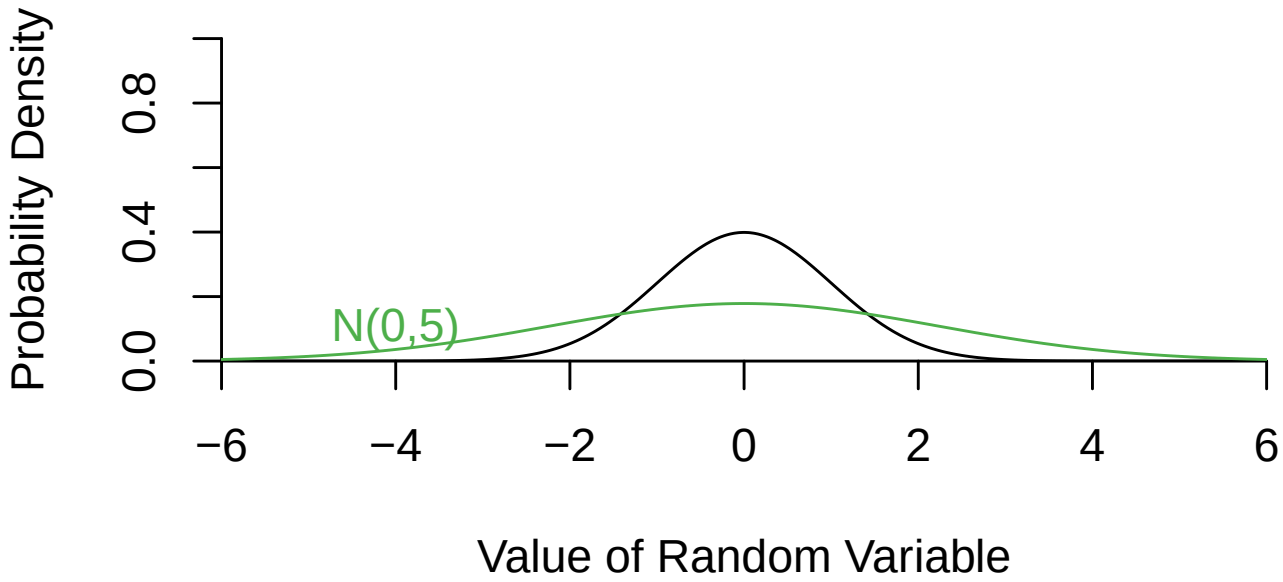
Examples of the Normal Distribution



This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 2$.

The larger variance has spread out the distribution.

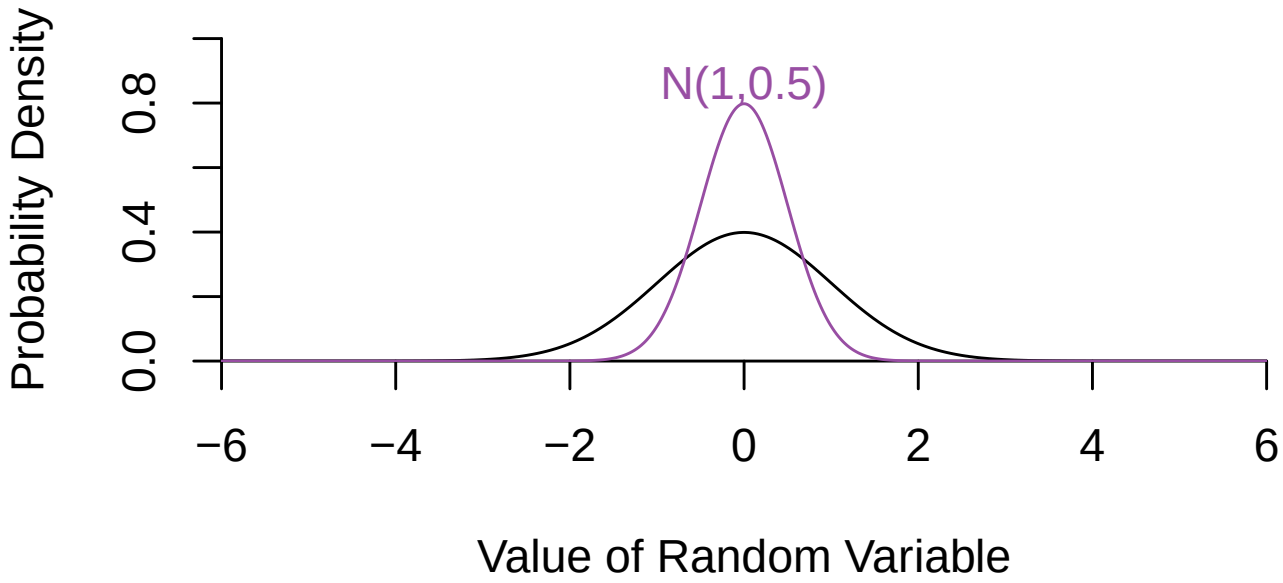
Examples of the Normal Distribution



This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 5$.

The larger variance has spread out the distribution even more

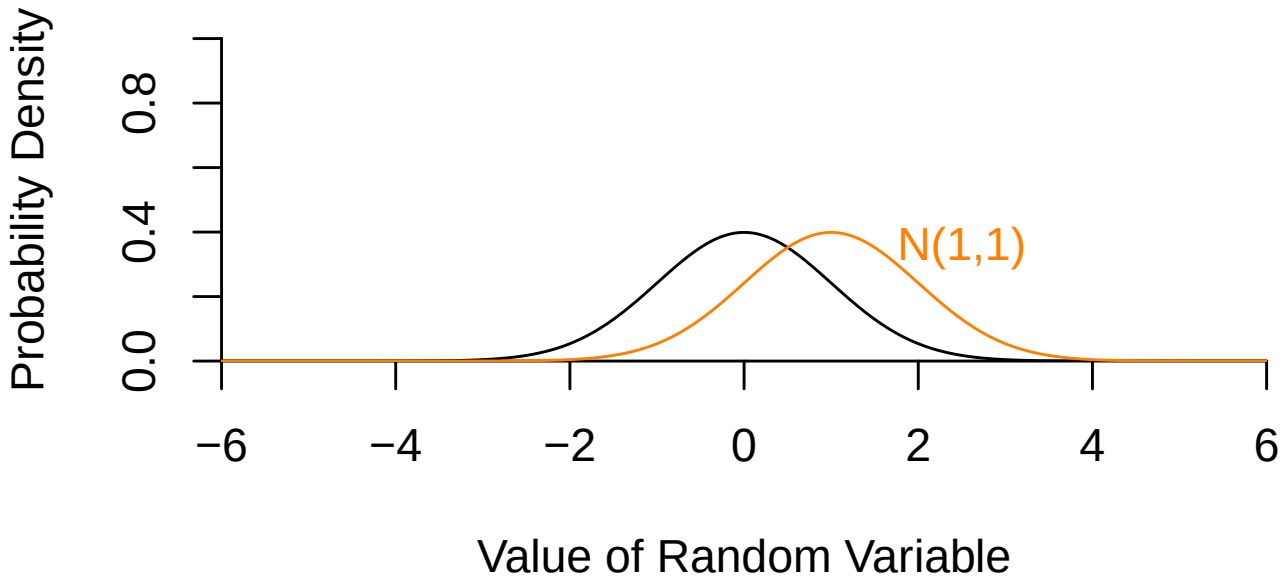
Examples of the Normal Distribution



This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 0.5$.

Smaller variance tightens distribution to a spike over the mean

Examples of the Normal Distribution



This is the Normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$.

Increasing the mean just shifts the distribution rightward

Additional properties of the Normal distribution

The Normal distribution has some nice properties:

1. Easy to add two Normal variables.

   If $X \sim \text{Normal}(\mu_X, \sigma_X^2)$ and $Y \sim \text{Normal}(\mu_Y, \sigma_Y^2)$, then

   $X + Y = \text{Normal}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

   Note this means the standard deviation of $X + Y$ is $\sqrt{\sigma_X^2 + \sigma_Y^2}$.

Additional properties of the Normal distribution

The Normal distribution has some nice properties:

1. Easy to add two Normal variables.

   If $X \sim \text{Normal}(\mu_X, \sigma_X^2)$ and $Y \sim \text{Normal}(\mu_Y, \sigma_Y^2)$, then

   $X + Y = \text{Normal}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

   Note this means the standard deviation of $X + Y$ is $\sqrt{\sigma_X^2 + \sigma_Y^2}$.

2. The Normal with mean 0 and variance 1 serves as a useful reference distribution. The Normal(0,1) is known as the Standard Normal distribution

Comparing distributions with different moments

Normally distributed variables can have widely varying means μ and variances σ^2

This raises a question: if we compare two values from two different Normal distributions, how do we decide which is "more extreme"?

For example, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%

Comparing distributions with different moments

Normally distributed variables can have widely varying means μ and variances σ^2

This raises a question: if we compare two values from two different Normal distributions, how do we decide which is "more extreme"?

For example, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%
2. A 65% on a test with a mean of 30% and a standard deviation of 25%

Comparing distributions with different moments

Normally distributed variables can have widely varying means μ and variances σ^2

This raises a question: if we compare two values from two different Normal distributions, how do we decide which is "more extreme"?

For example, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%
2. A 65% on a test with a mean of 30% and a standard deviation of 25%
3. A 38% on a test with a mean of 25% and a standard deviation of 5%

z-scores

To solve this sort of problem,
it helps to *standardize* a normal variable to have the same mean and variance

That is, we convert each score to a common metric, called a z-score, in which the mean is 0, and each unit is a standard deviation move away from zero

For random variable x with mean μ and variance σ^2, the z-score is:

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

Notice that while the original variable $X \sim \text{Normal}(\mu, \sigma^2)$,

the z-score is $Z \sim \text{Normal}(0, 1)$

z-scores: Example

So, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%

$$z = \frac{x - \mu}{\sigma} = \frac{0.9 - 0.8}{0.06} = 1.67$$

z-scores: Example

So, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%

$$z = \frac{x - \mu}{\sigma} = \frac{0.9 - 0.8}{0.06} = 1.67$$

2. A 65% on a test with a mean of 30% and a standard deviation of 25%

$$z = \frac{x - \mu}{\sigma} = \frac{0.65 - 0.3}{0.25} = 1.4$$

z-scores: Example

So, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%

$$z = \frac{x - \mu}{\sigma} = \frac{0.9 - 0.8}{0.06} = 1.67$$

2. A 65% on a test with a mean of 30% and a standard deviation of 25%

$$z = \frac{x - \mu}{\sigma} = \frac{0.65 - 0.3}{0.25} = 1.4$$

3. A 38% on a test with a mean of 25% and a standard deviation of 5%

$$z = \frac{x - \mu}{\sigma} = \frac{0.38 - 0.25}{0.05} = 2.6$$

z-scores: Example

So, which is more impressive?

1. A 90% on a test with a mean of 80% and a standard deviation of 6%

$$z = \frac{x - \mu}{\sigma} = \frac{0.9 - 0.8}{0.06} = 1.67$$

2. A 65% on a test with a mean of 30% and a standard deviation of 25%

$$z = \frac{x - \mu}{\sigma} = \frac{0.65 - 0.3}{0.25} = 1.4$$

3. A 38% on a test with a mean of 25% and a standard deviation of 5%

$$z = \frac{x - \mu}{\sigma} = \frac{0.38 - 0.25}{0.05} = 2.6$$

All three scores are impressive, but the student with 38% should be proudest.

z-scores and percentiles

What are the (theoretical) percentiles of the three test scores?

That is, what percentage of test-takers did student 1 beat?
student 2? student 3?

We can easily look up the percentile of a z-score (using your textbook or `R`)

For our example,

μ	σ	x	z	percentile
0.80	0.06	0.90	1.67	95th
0.30	0.25	0.65	1.40	92nd
0.25	0.05	0.38	2.60	99th

(Note to the curious:
To calculate these percentiles, I just used `pnorm(z)` in `R`.)

z-scores and critical values

If we can go from z-scores to percentiles,
we can also go from percentiles to z-scores

Suppose you took the third exam,
with the mean of 25% and the standard deviation of 5%.

How well would you have to score to be in the top 20% of the class?

To answer this, we first need to find the z^*, or critical value,
that marks the 80th percentile.

z-scores and critical values

we first need to find the z^*, or critical value, which marks the 80th percentile.

If we look this up in `R`, using `qnorm(0.8)`,
we find the desired $z^* = 0.84$

What actual test score does this correspond to?

Note that if $z = (x - \mu)/\sigma$, then

$$x^* \quad = \quad z^*\sigma + \mu$$

z-scores and critical values

we first need to find the z^*, or critical value, which marks the 80th percentile.

If we look this up in `R`, using `qnorm(0.8)`,
we find the desired $z^* = 0.84$

What actual test score does this correspond to?

Note that if $z = (x - \mu)/\sigma$, then

$$\begin{aligned} x^* &= z^*\sigma + \mu \\ &= 0.84 \times 0.05 + 0.25 \end{aligned}$$

z-scores and critical values

we first need to find the z^*, or critical value, which marks the 80th percentile.

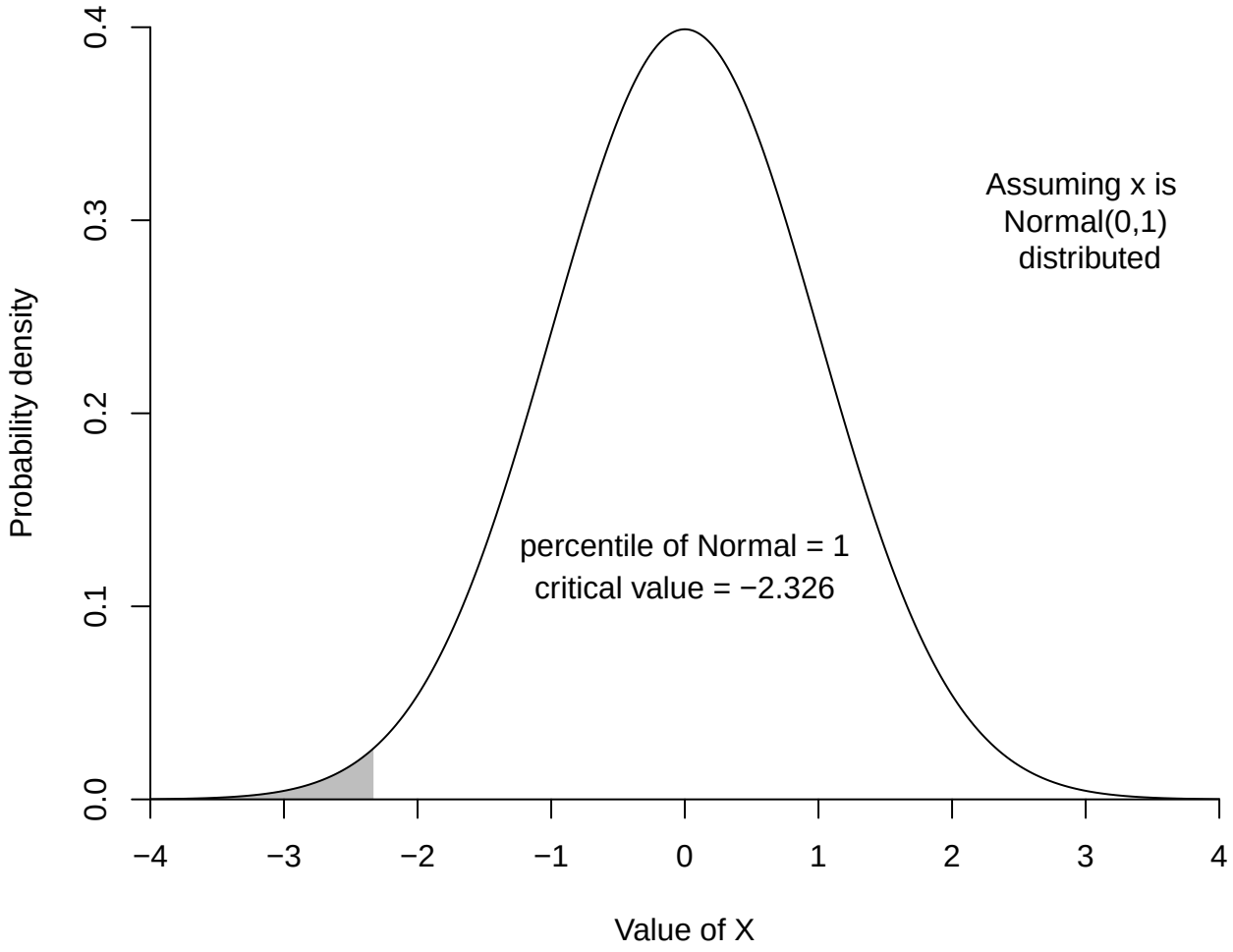
If we look this up in `R`, using `qnorm(0.8)`,
we find the desired $z^* = 0.84$

What actual test score does this correspond to?

Note that if $z = (x - \mu)/\sigma$, then

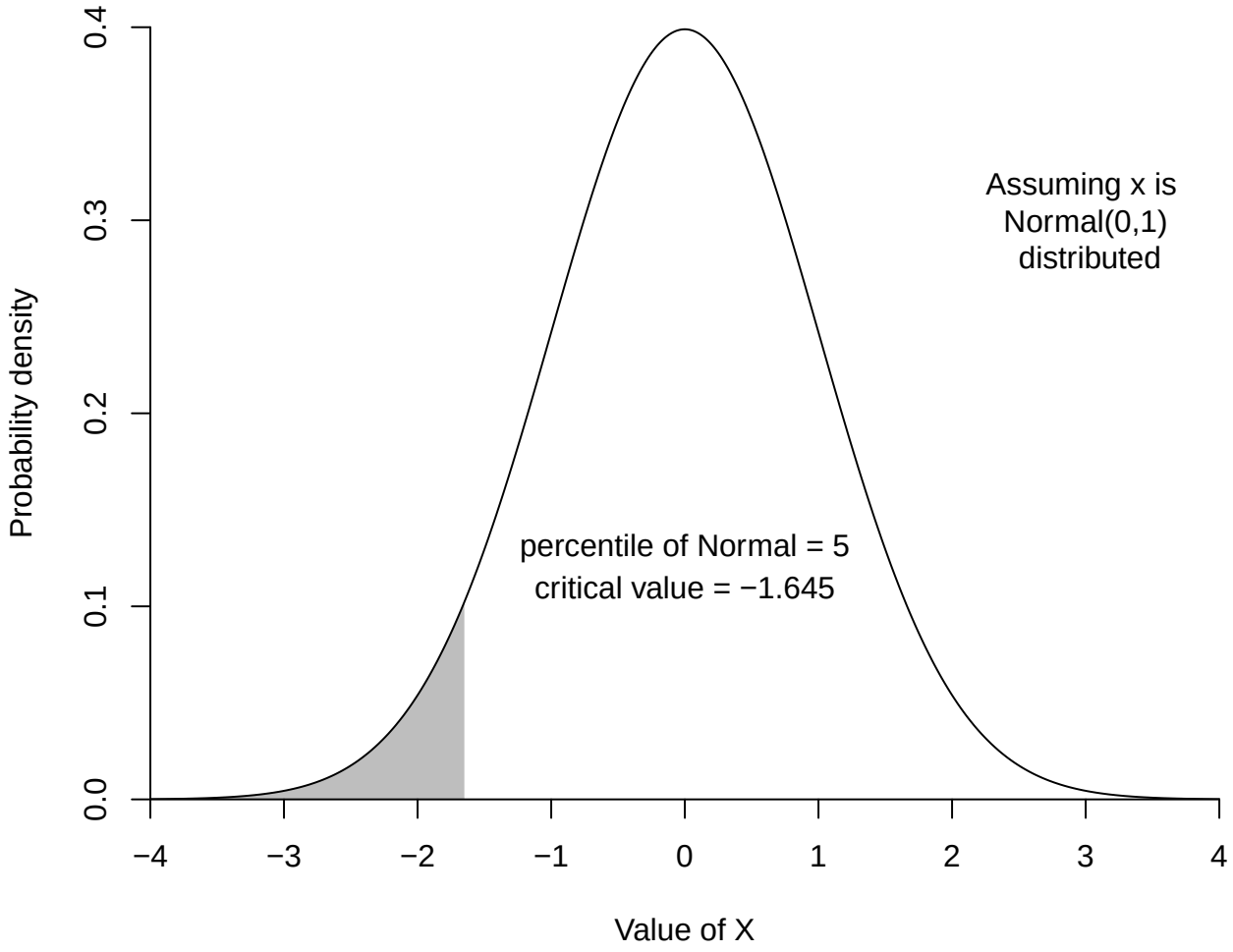
$$\begin{aligned} x^* &= z^*\sigma + \mu \\ &= 0.84 \times 0.05 + 0.25 \\ &= 29.2\% \end{aligned}$$

Upshot: if we know the theoretical distribution of a Normal variable,
we can freely convert between the variable, z-scores, and percentiles



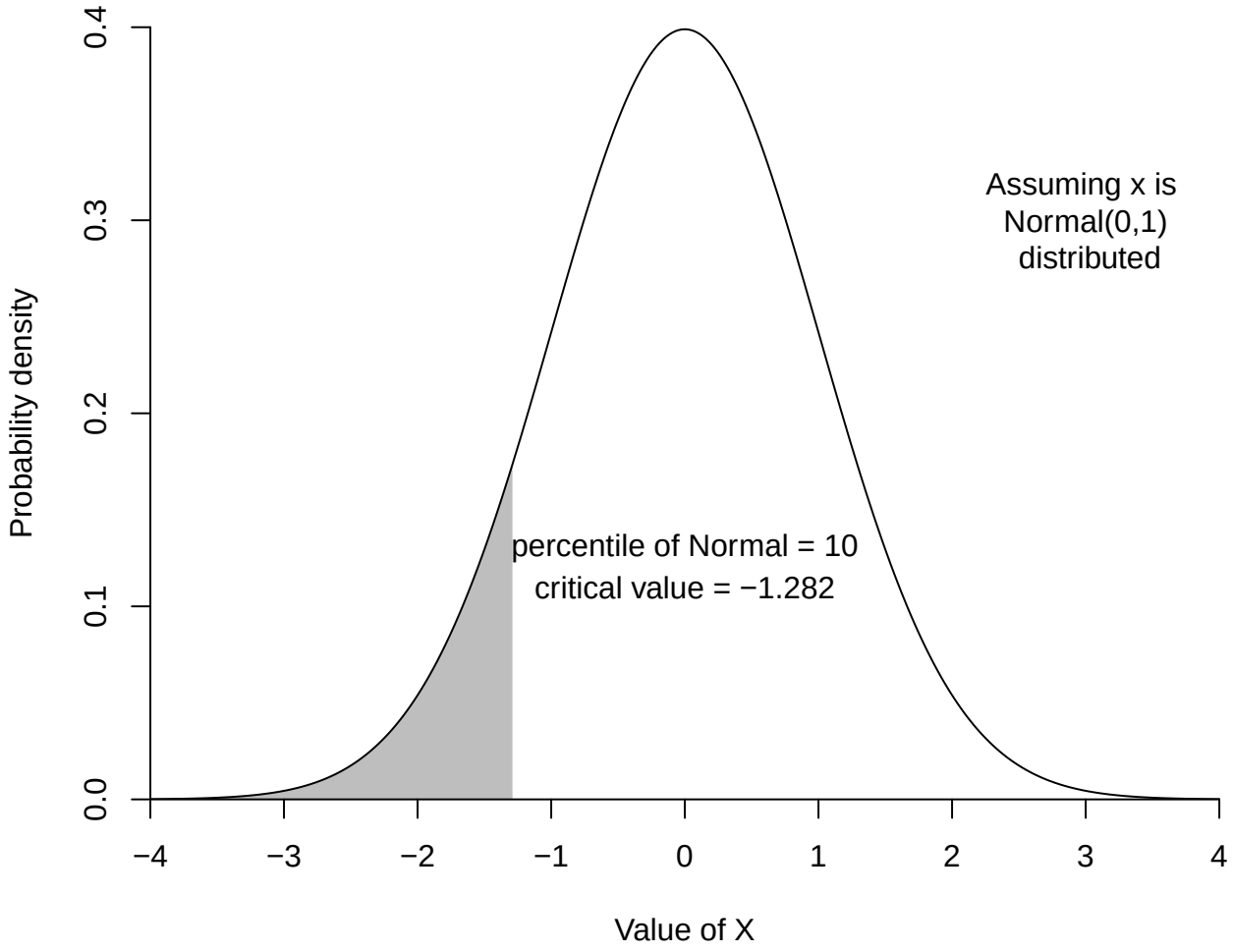
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 1% of the distribution



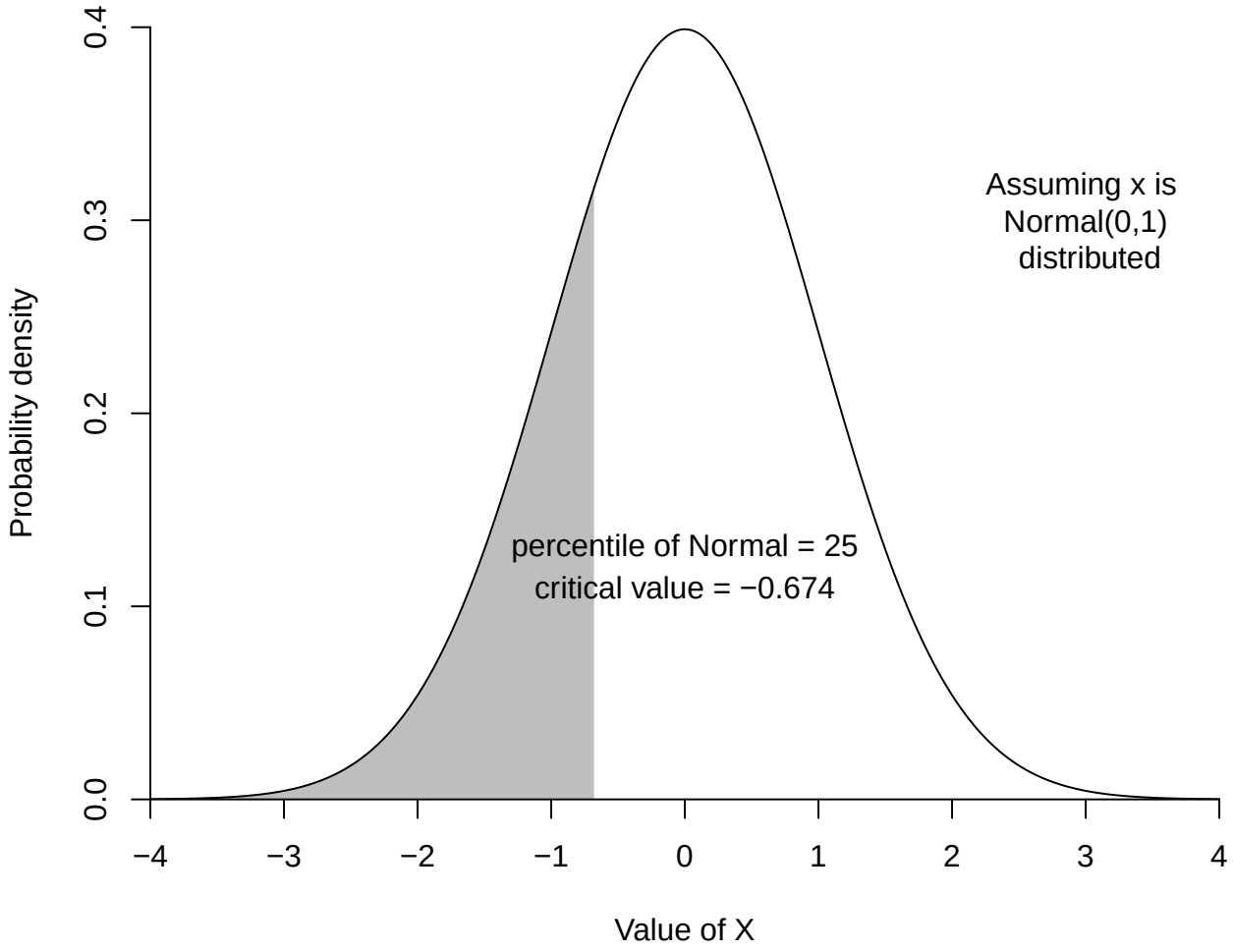
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 5% of the distribution



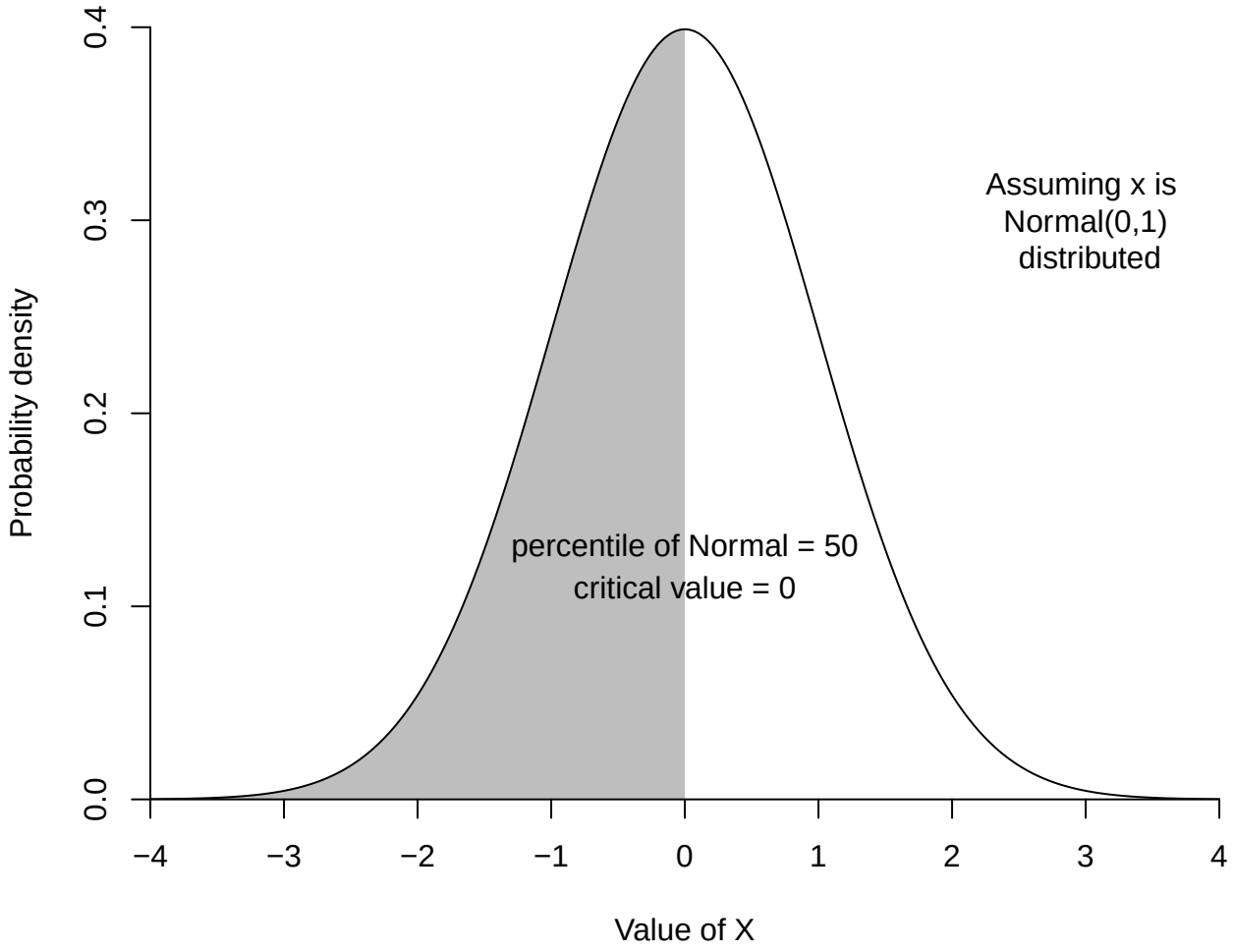
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 10% of the distribution



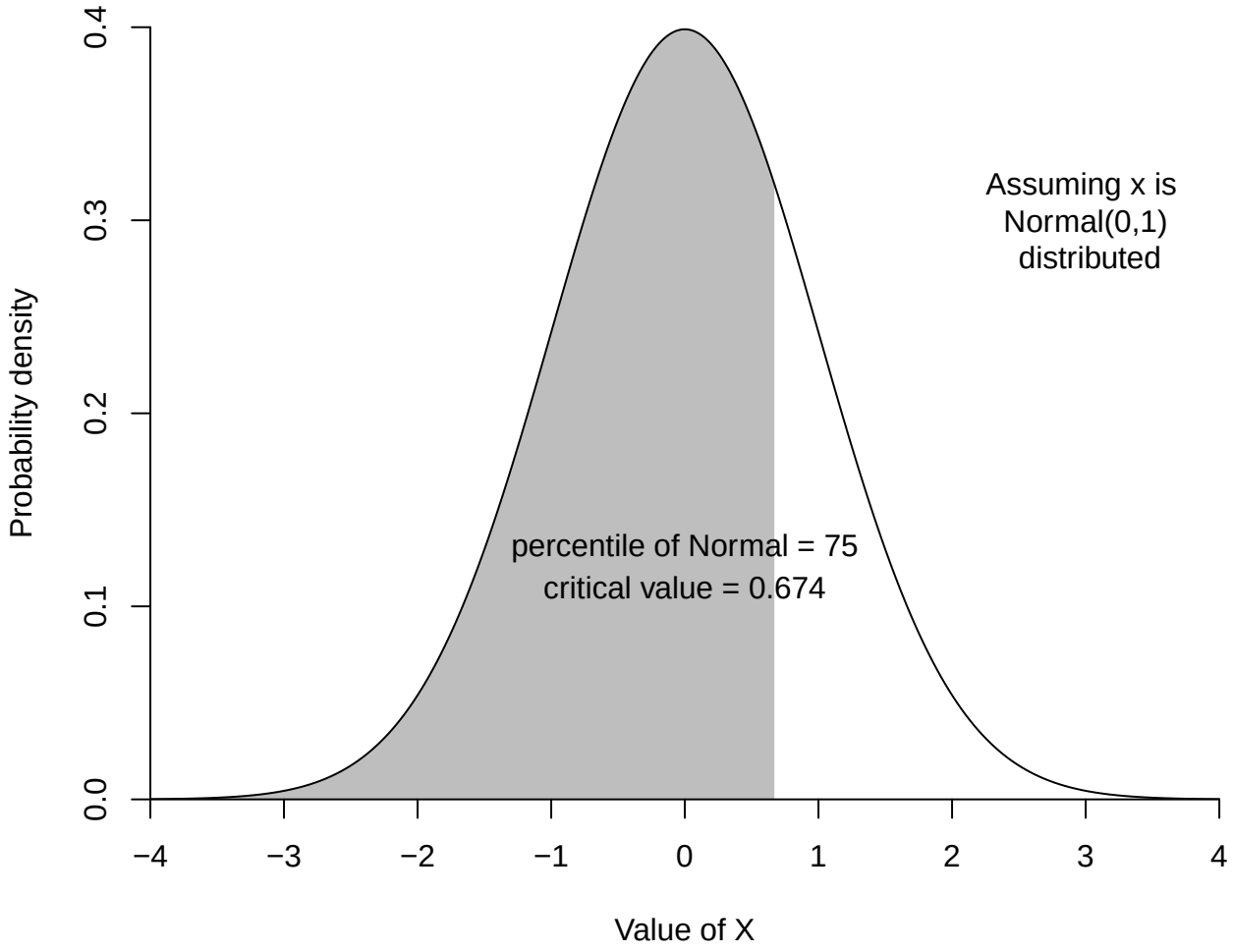
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 25% of the distribution



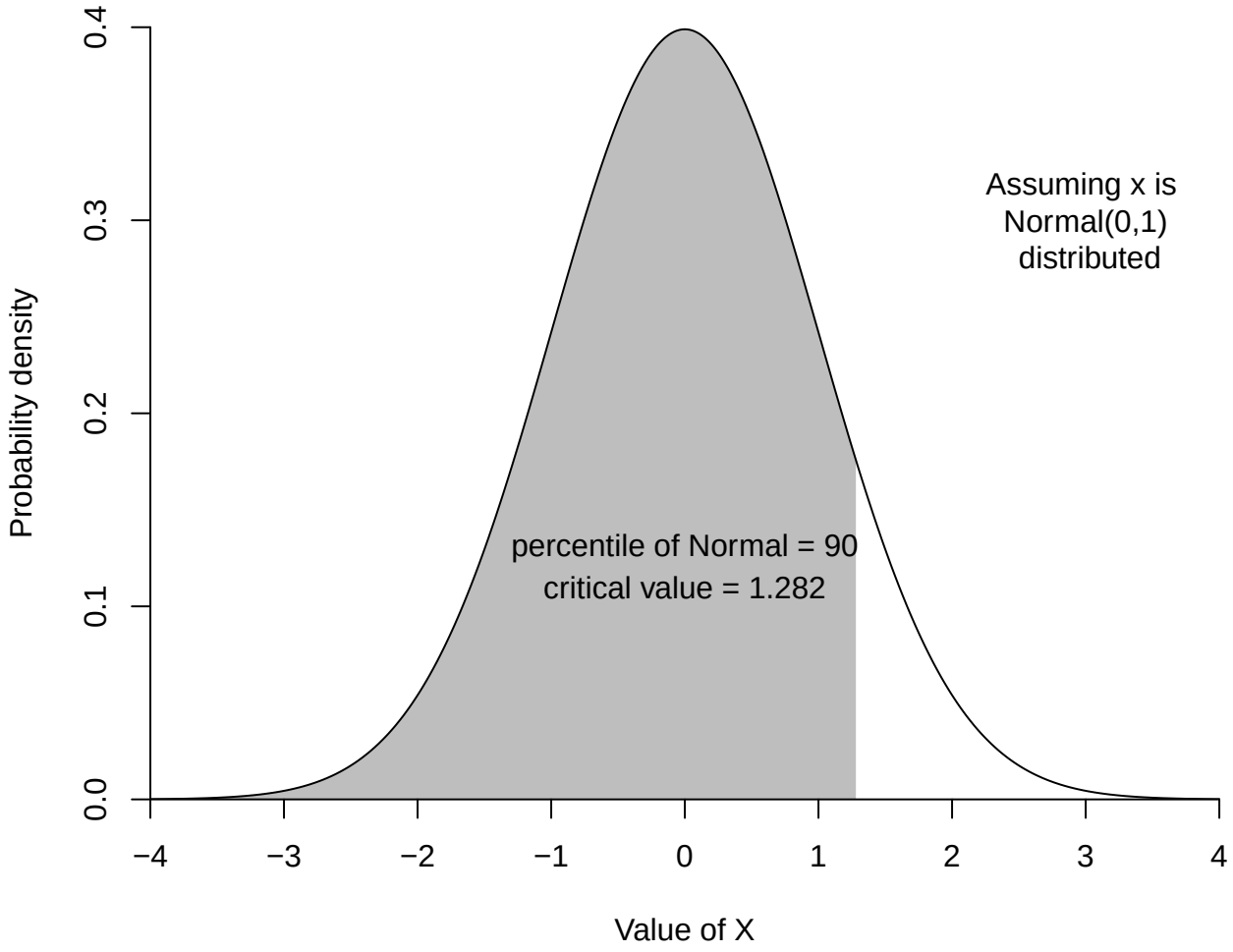
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 50% of the distribution



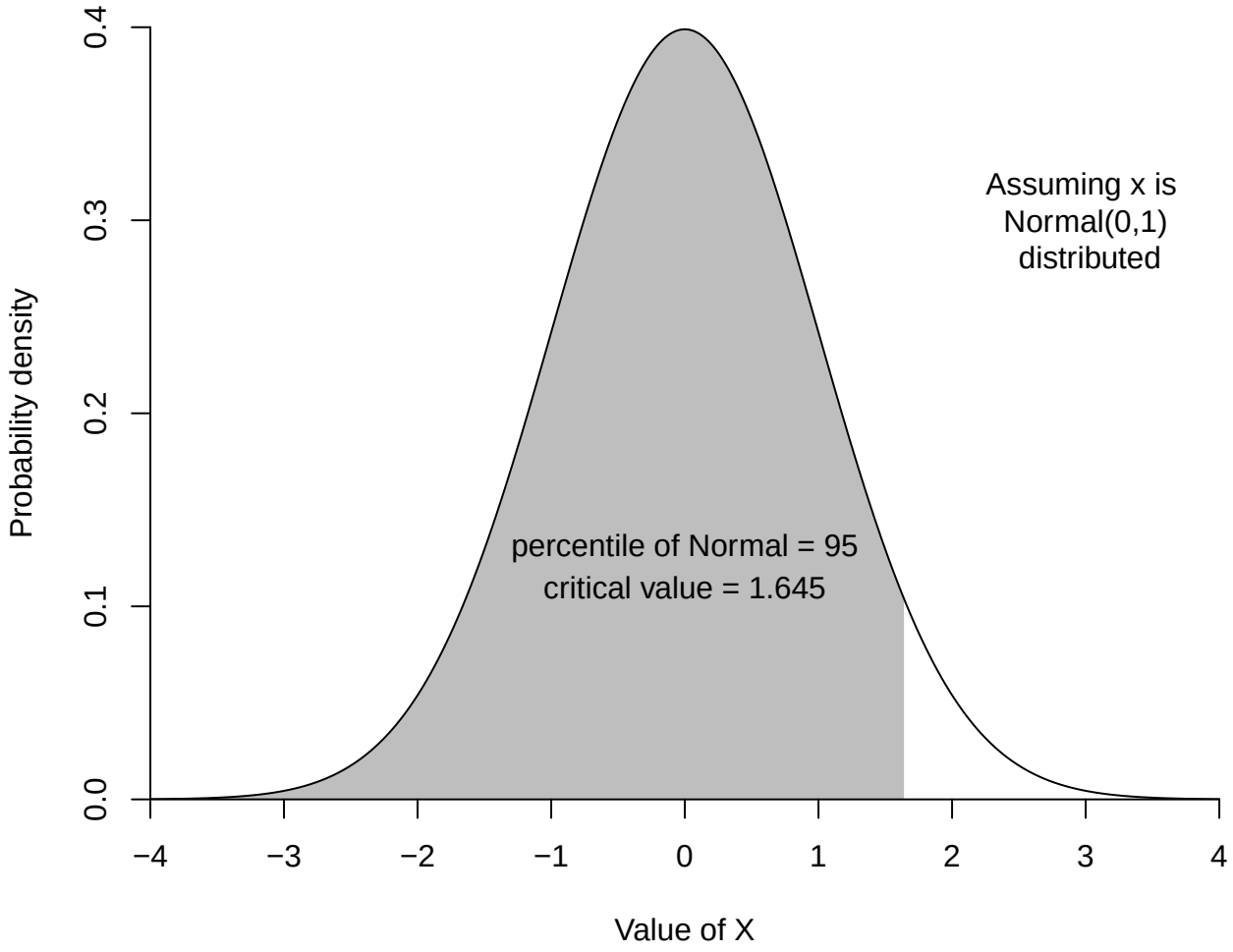
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 75% of the distribution



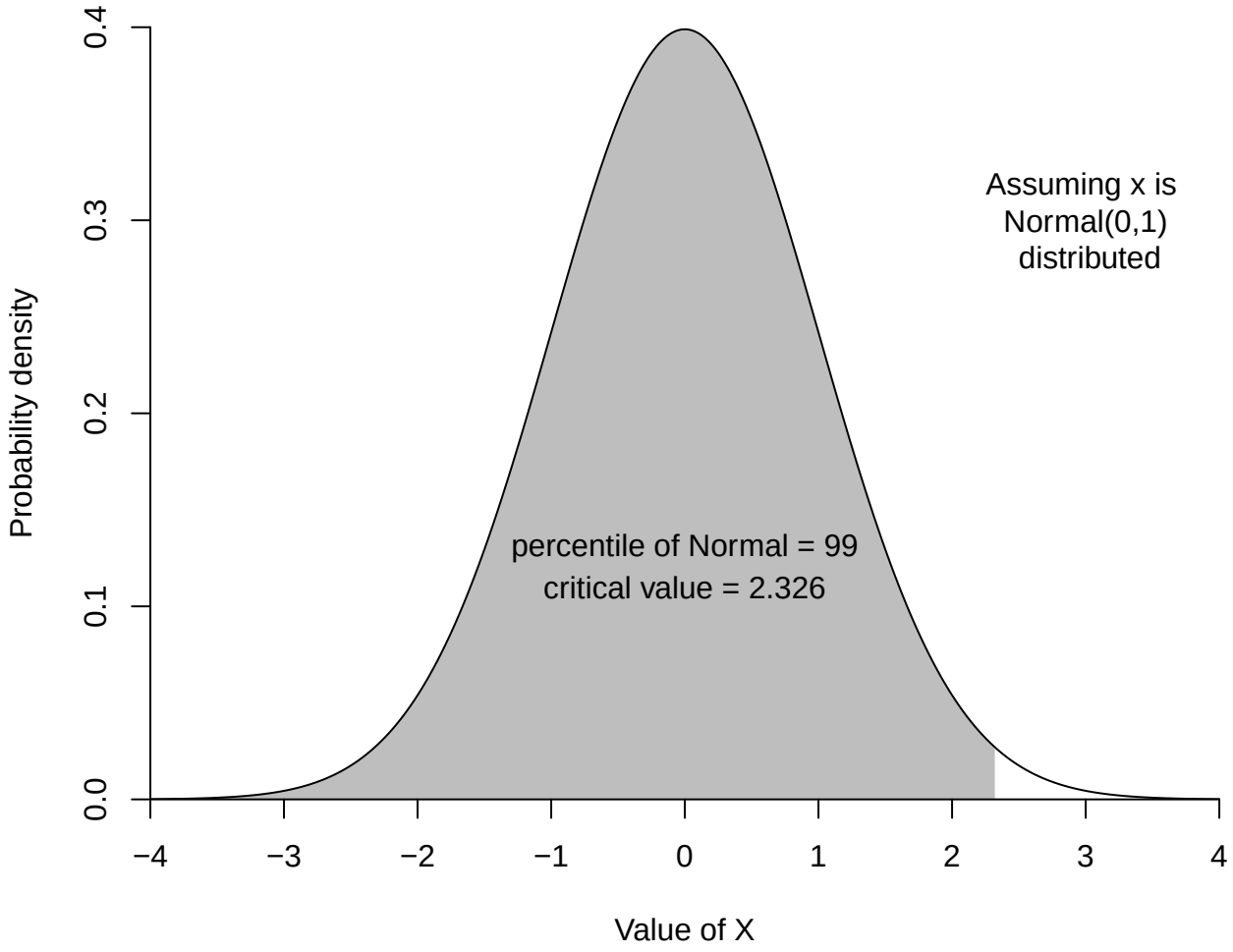
Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 90% of the distribution



Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 95% of the distribution



Let's take a quick survey of the percentiles and critical values of the standard Normal

Here is the first 99% of the distribution

Unemployment example

Let's apply this framework to a real world variable.

Unemployment example

Let's apply this framework to a real world variable.

Unemployment in the US reached 9.6% in 2010, but varied across states

Unemployment example

Let's apply this framework to a real world variable.

Unemployment in the US reached 9.6% in 2010, but varied across states

Suppose that the average state had 9.6% unemployment, but that the standard deviation across states was 2.2.

Unemployment example

Let's apply this framework to a real world variable.

Unemployment in the US reached 9.6% in 2010, but varied across states

Suppose that the average state had 9.6% unemployment, but that the standard deviation across states was 2.2.

If we suppose unemployment is Normally distributed,
this leads to a Normal(μ=9.6, σ^2=4.84) distribution

Unemployment example

Let's apply this framework to a real world variable.

Unemployment in the US reached 9.6% in 2010, but varied across states

Suppose that the average state had 9.6% unemployment, but that the standard deviation across states was 2.2.

If we suppose unemployment is Normally distributed,
this leads to a Normal(μ=9.6, σ^2=4.84) distribution

(Quick check: why did I write 4.84 instead of 2.2?

Unemployment example

Let's apply this framework to a real world variable.

Unemployment in the US reached 9.6% in 2010, but varied across states

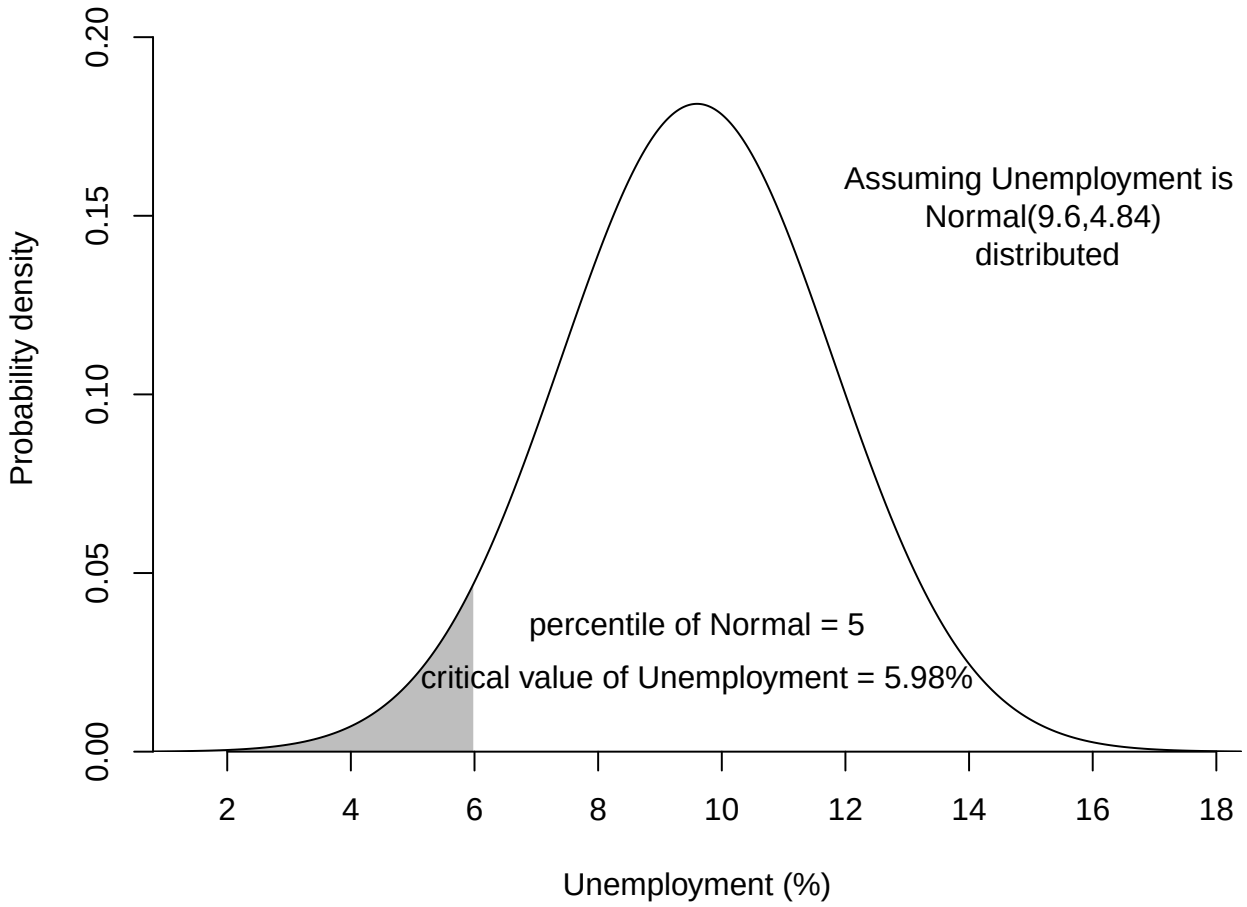
Suppose that the average state had 9.6% unemployment, but that the standard deviation across states was 2.2.

If we suppose unemployment is Normally distributed,
this leads to a Normal(μ=9.6, σ^2=4.84) distribution

(Quick check: why did I write 4.84 instead of 2.2?

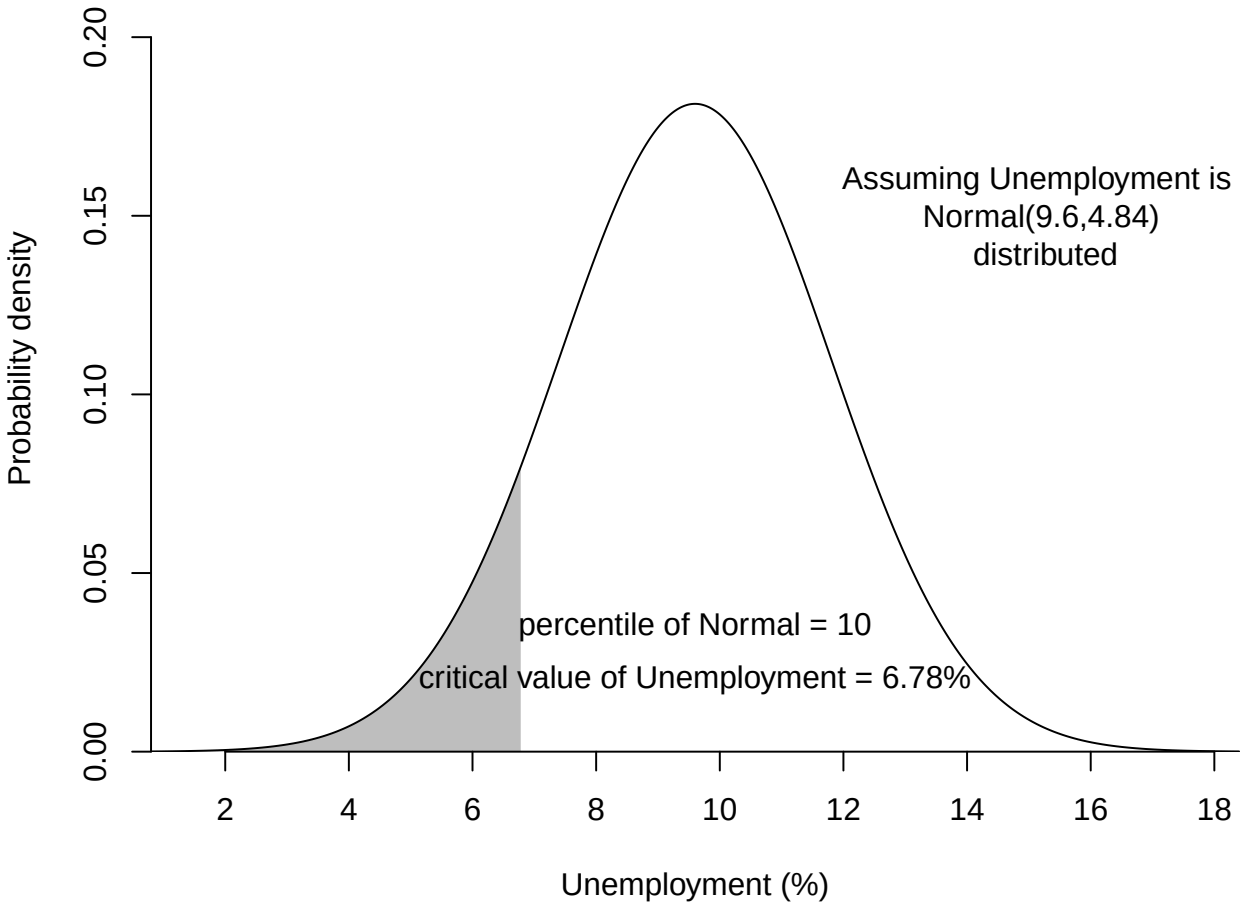
Answer: statisticians usually use the variance when we write the distribution in Normal(μ,σ^2)

Your text often uses the standard deviation σ instead – be careful to note which is being used!)



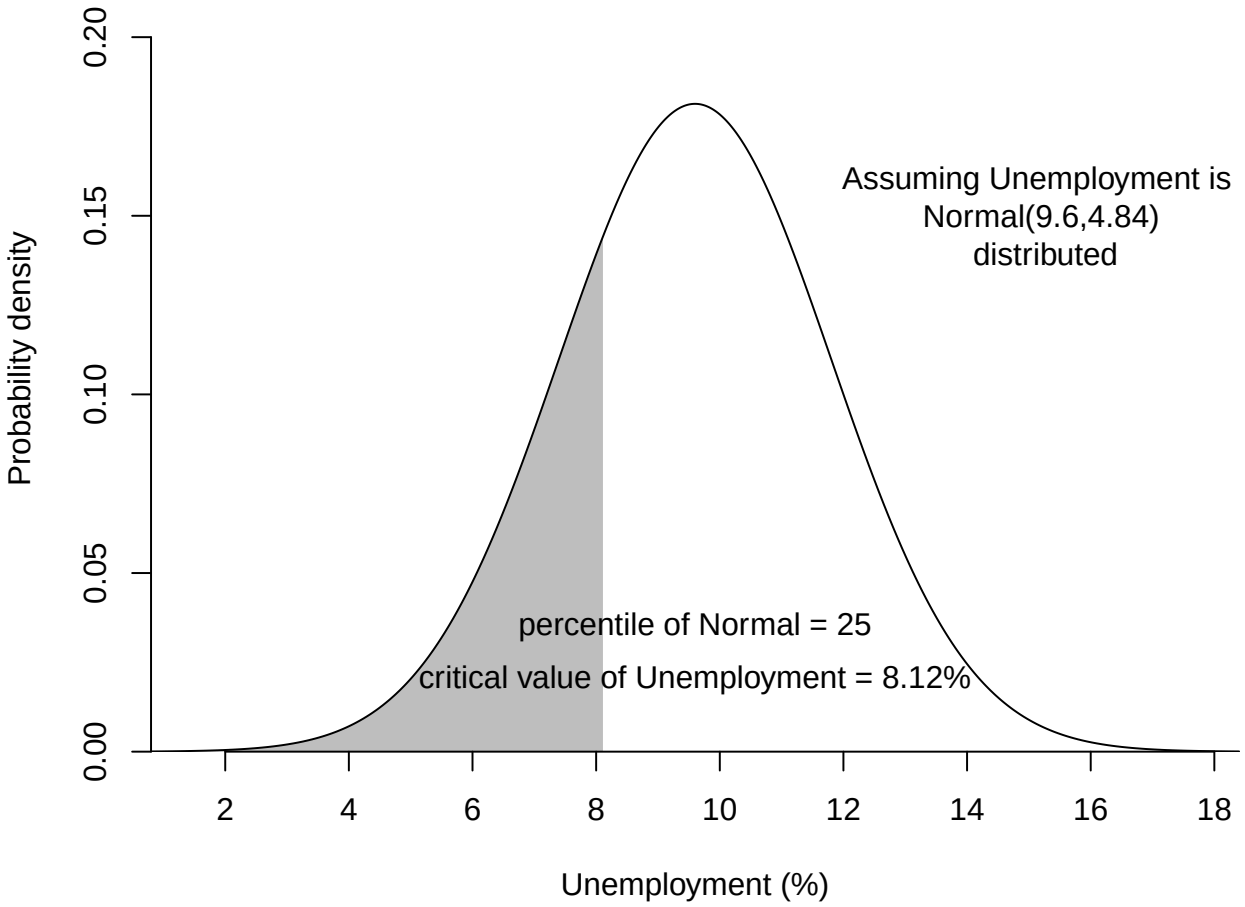
Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 5% of states should be below 5.98% unemployment



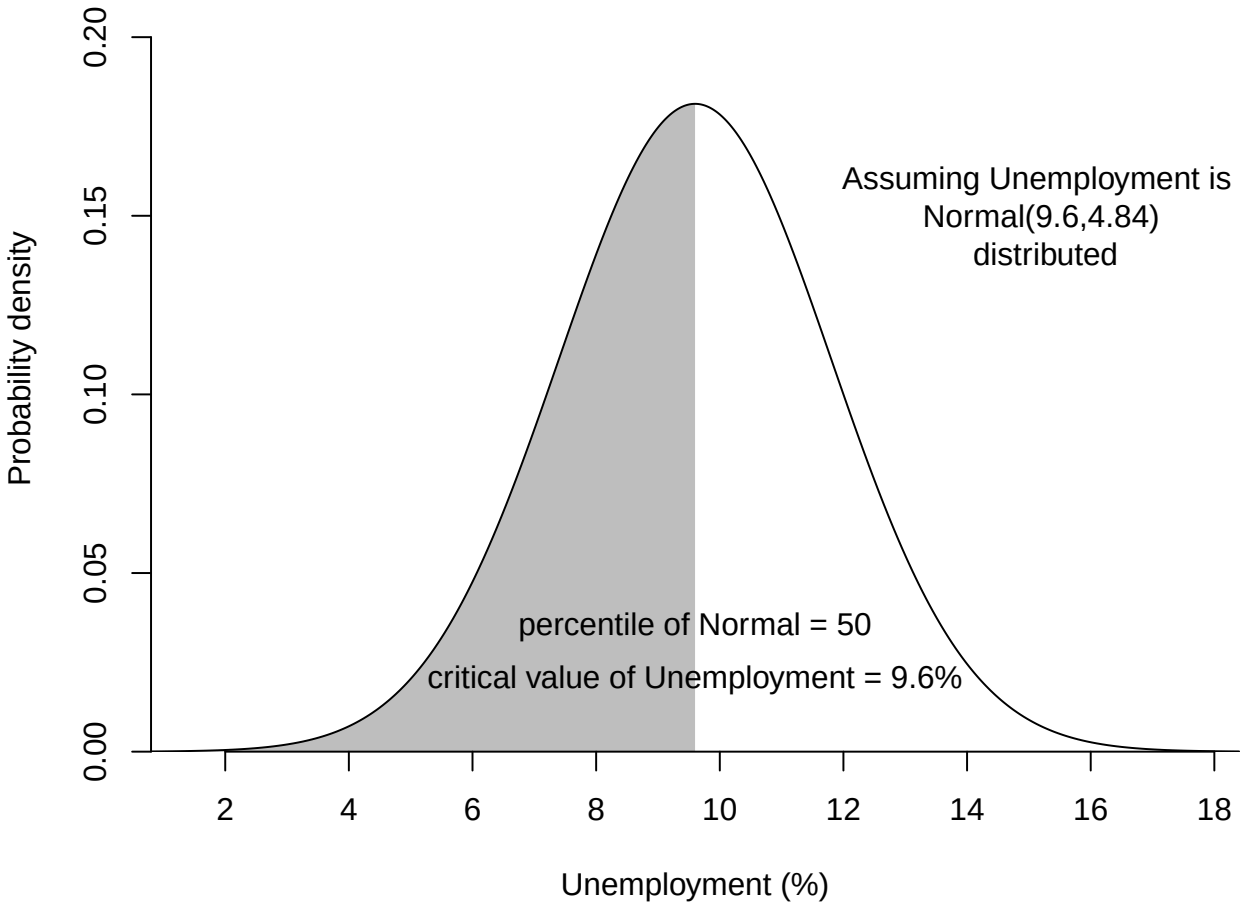
Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 10% of states should be below 6.78% unemployment



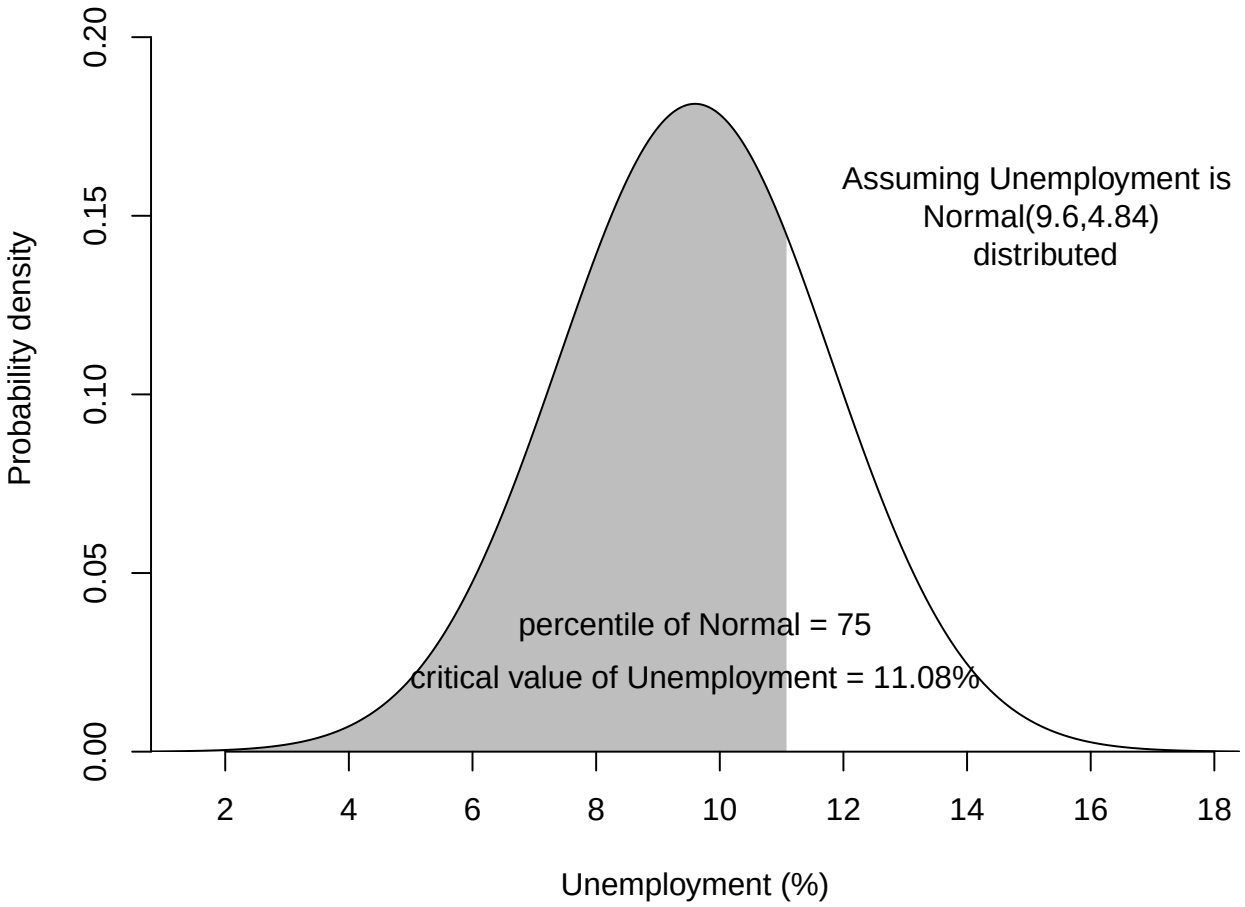
Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 25% of states should be below 8.12% unemployment



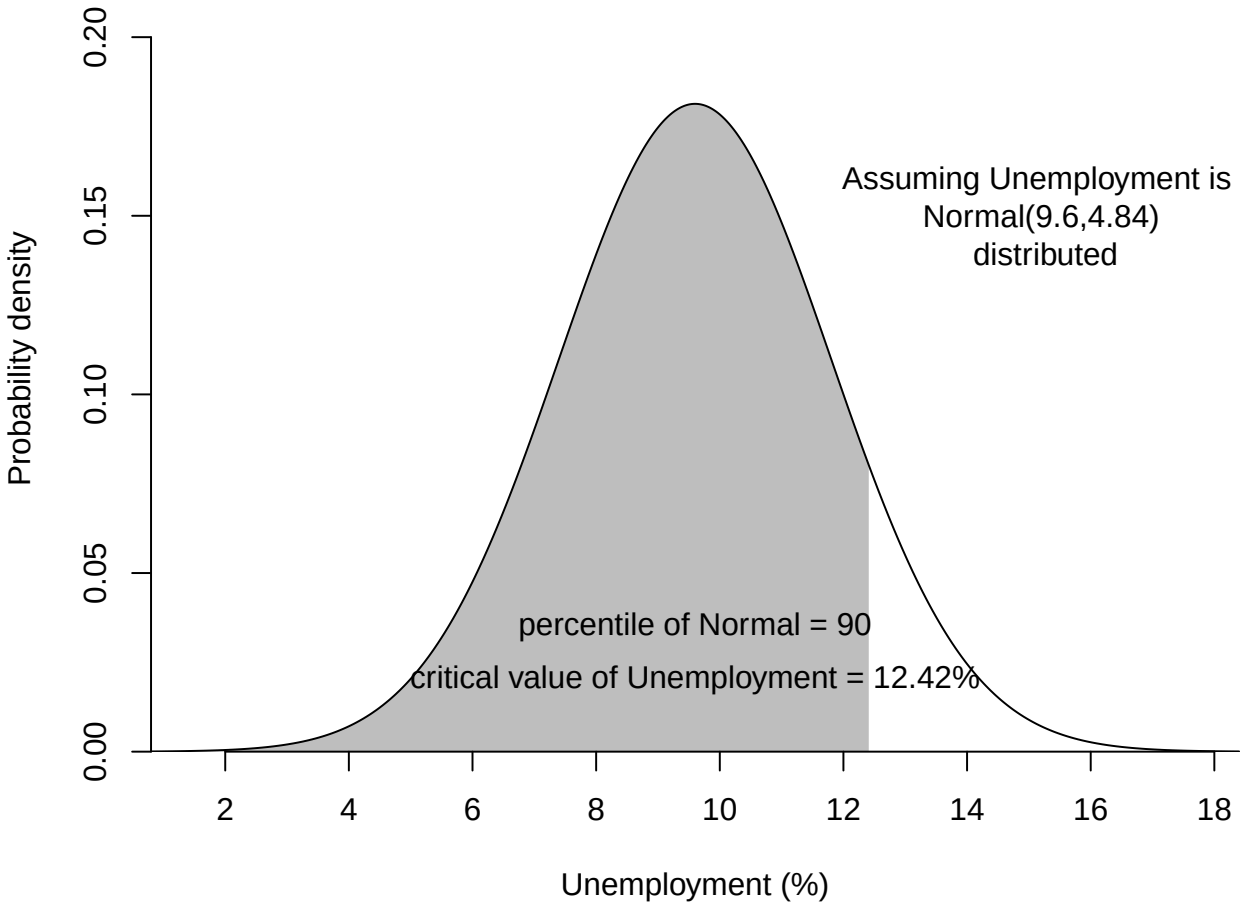
Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 50% of states should be below 9.6% unemployment



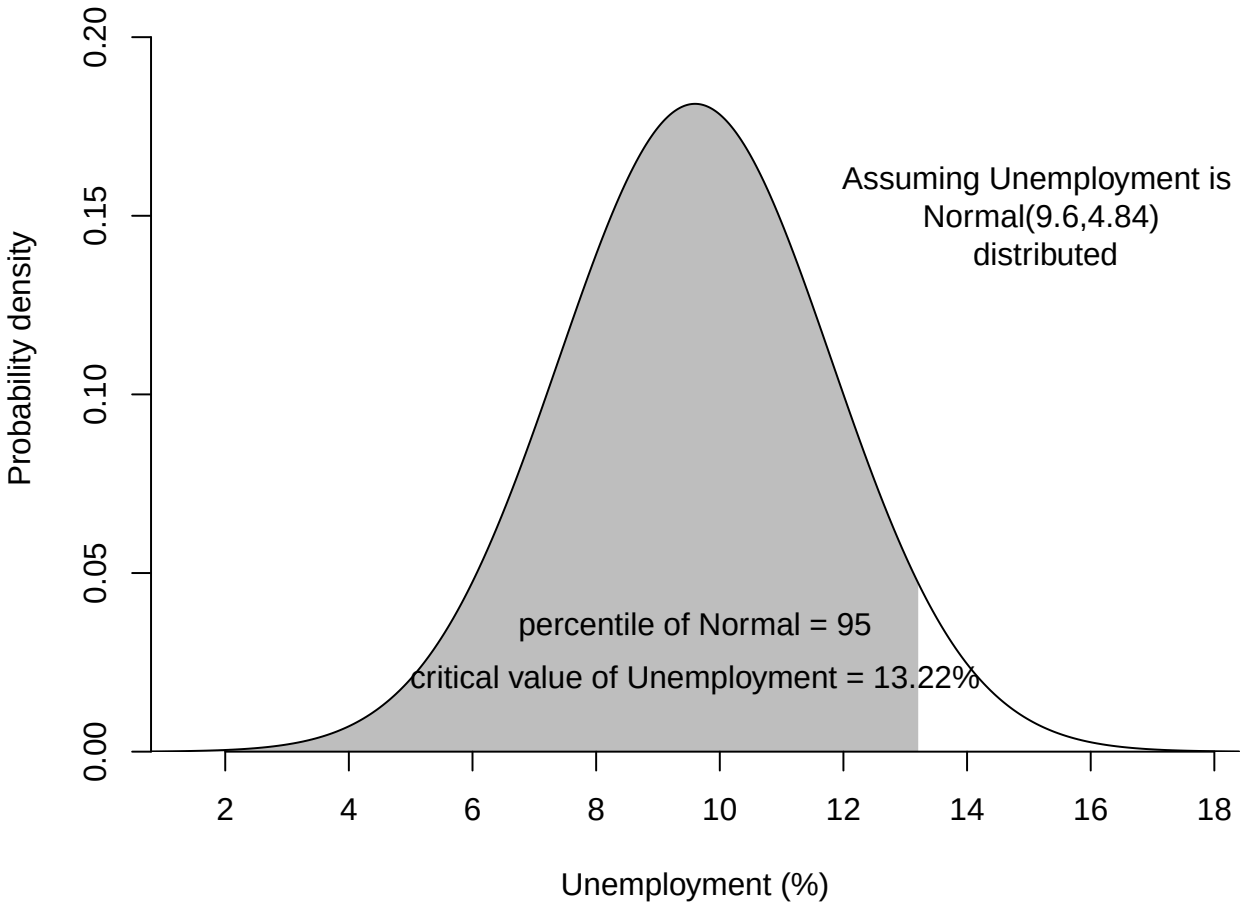
Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 75% of states should be below 11.08% unemployment



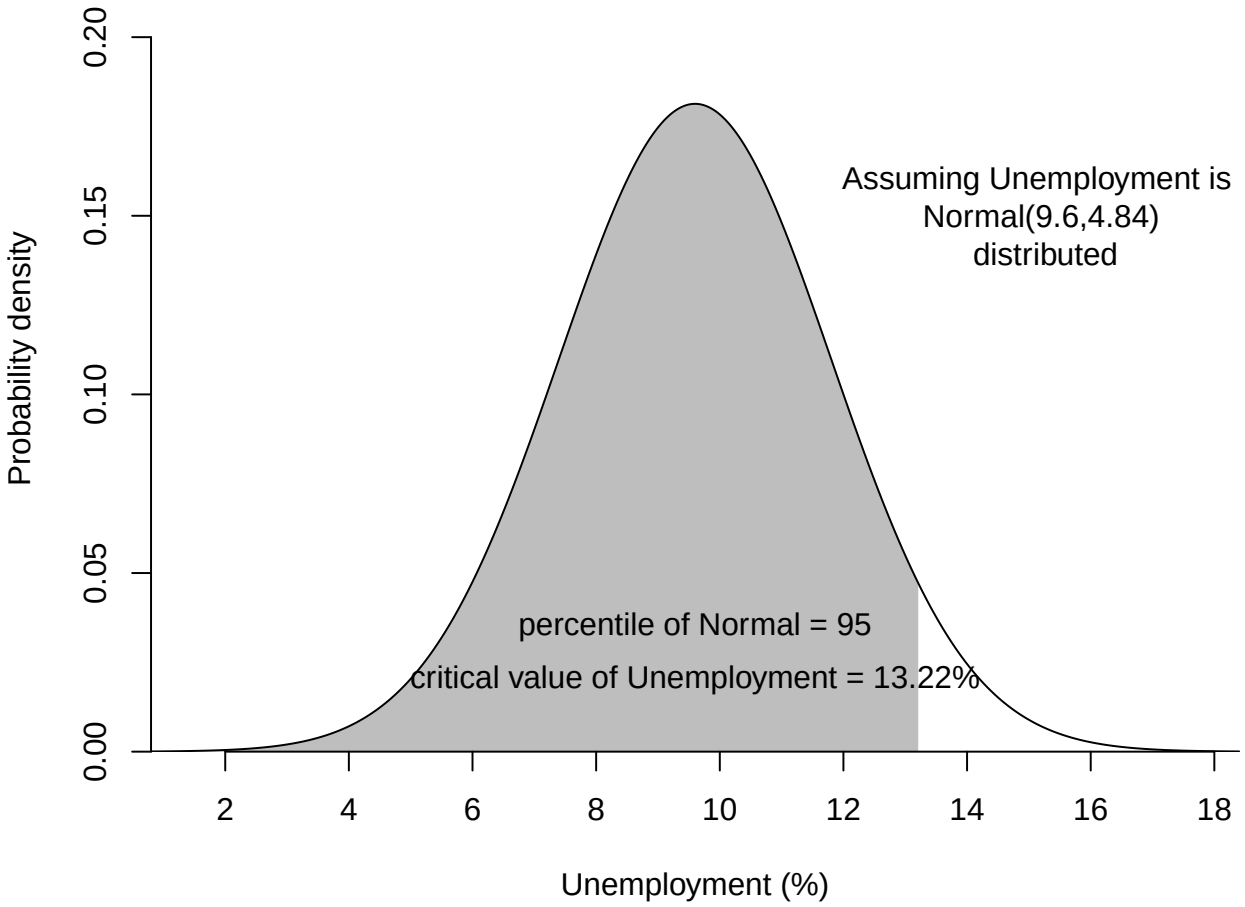
Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 90% of states should be below 12.42% unemployment

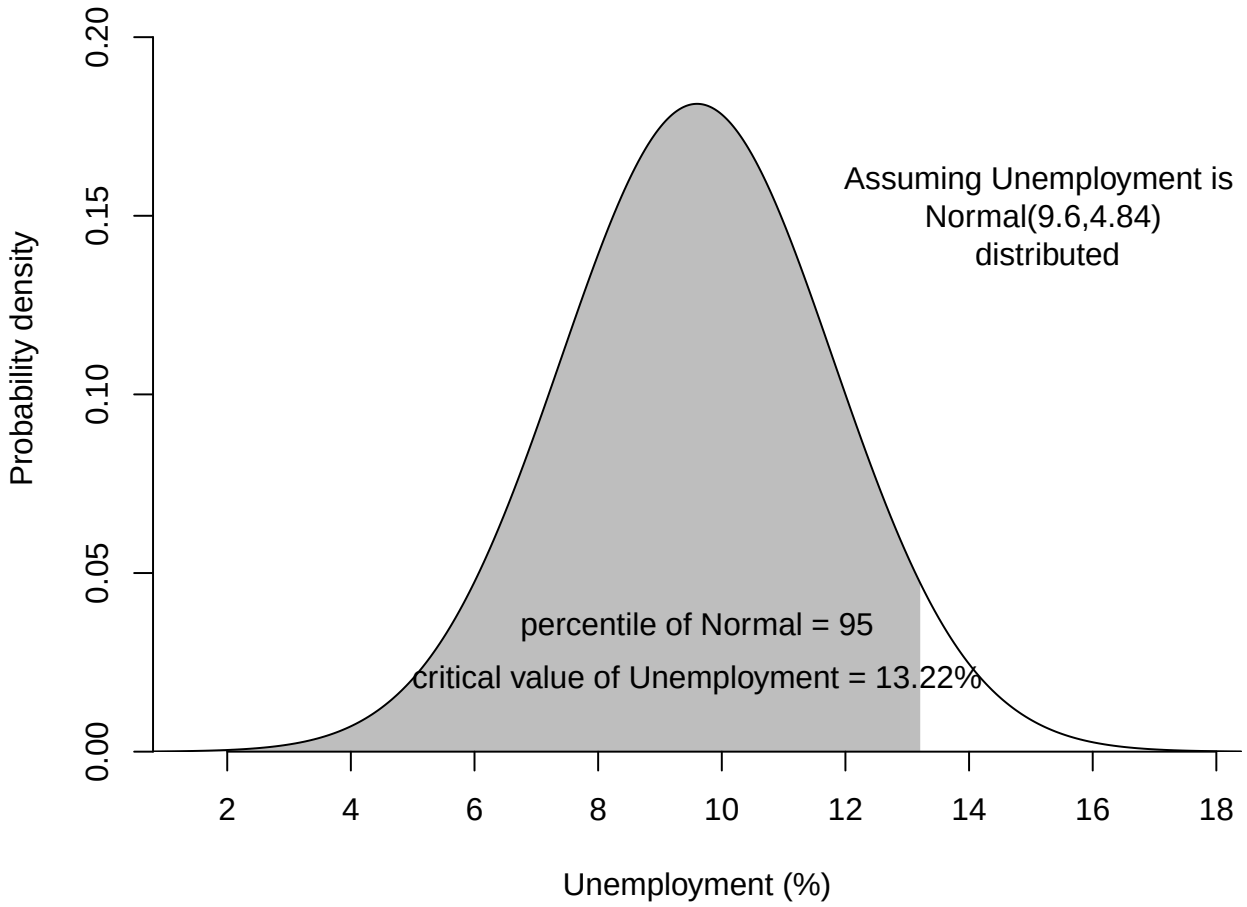


Using the given distribution, I have calculated, in `R`, the critical values for various percentiles

I find that 95% of states should be below 13.22% unemployment

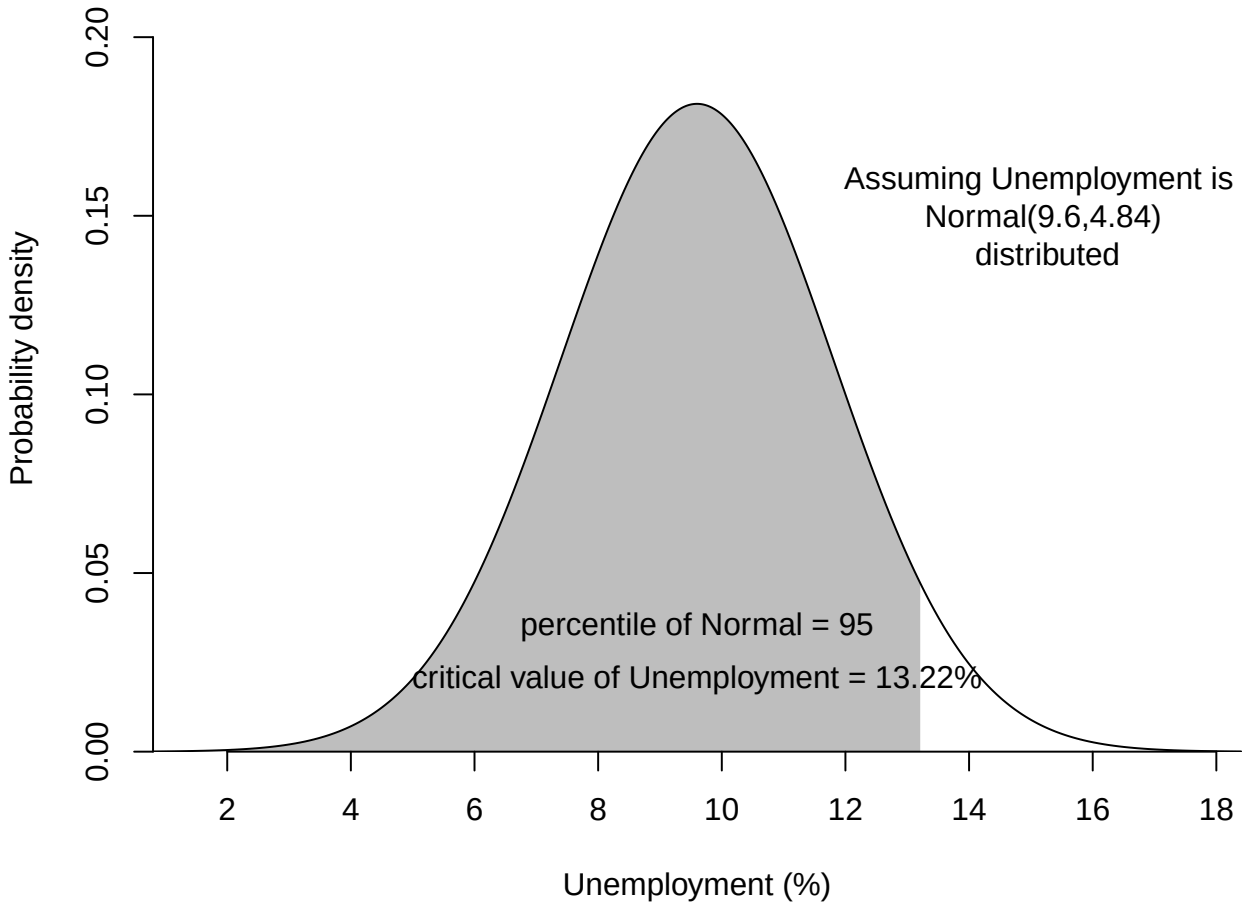


Using the given distribution, I have calculated, in R, the critical values for various percentiles.
How?



Using the given distribution, I have calculated, in R, the critical values for various percentiles. **How?**

In math: $z^* \sigma + \mu$



Using the given distribution, I have calculated, in R, the critical values for various percentiles. **How?**

In math: $z^* \sigma + \mu$

In R:

```
qnorm(0.95)*2.2 +
9.6
```

z-scores and probability intervals

Suppose you wanted to summarize the range of most probable outcomes for a theoretical Normal distribution.

For example, if the mean male height in the US is 5 ft 10 in, and the standard deviation is 3 in, *and* height is Normally distributed,

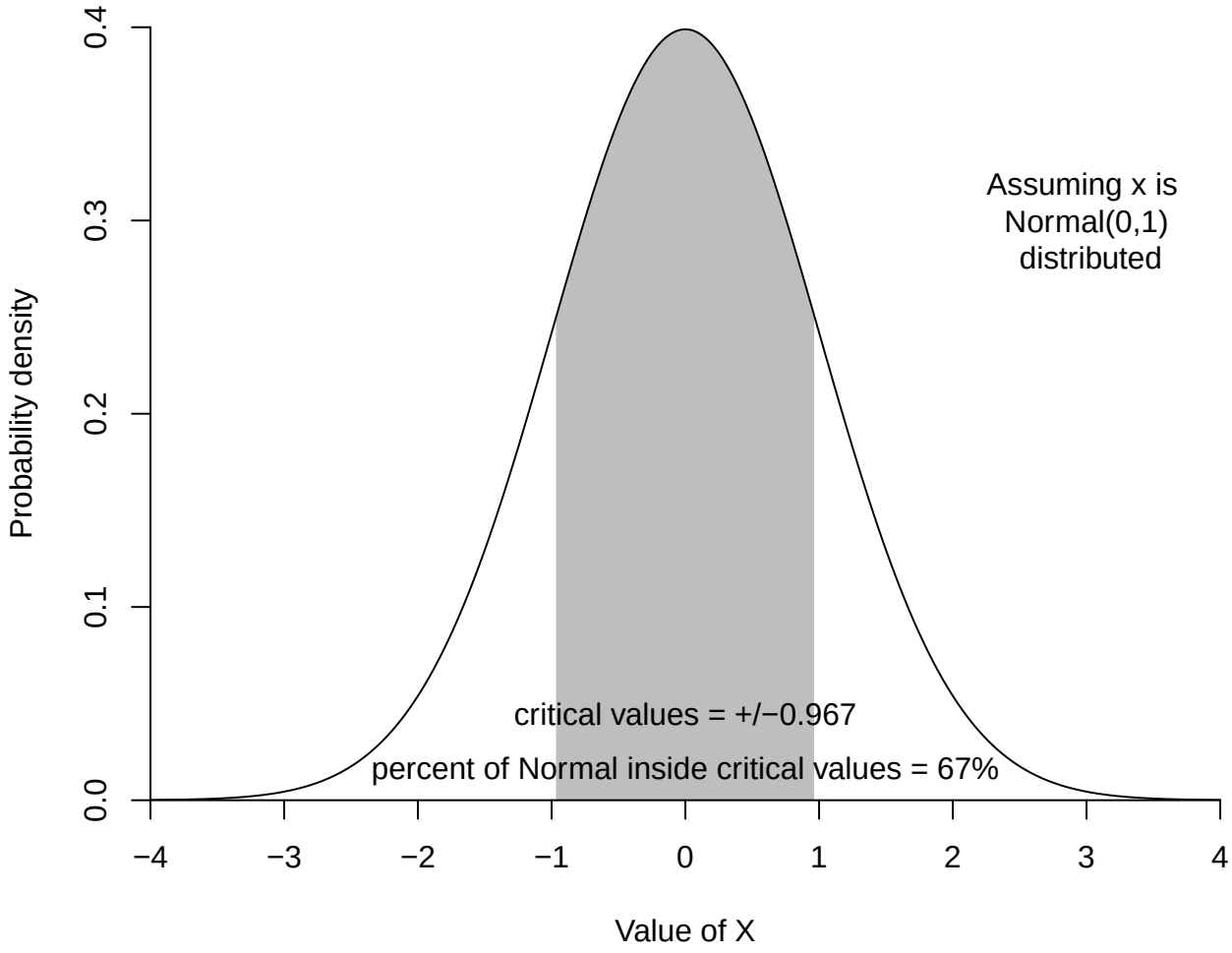
1. What critical values of height bound two-third of all men?
2. What critical values of height bound 95% of all men?
3. What critical values of height bound 99% of all men?

z-scores and probability intervals

What critical values of height bound 95% of all men?

Slightly tricky:

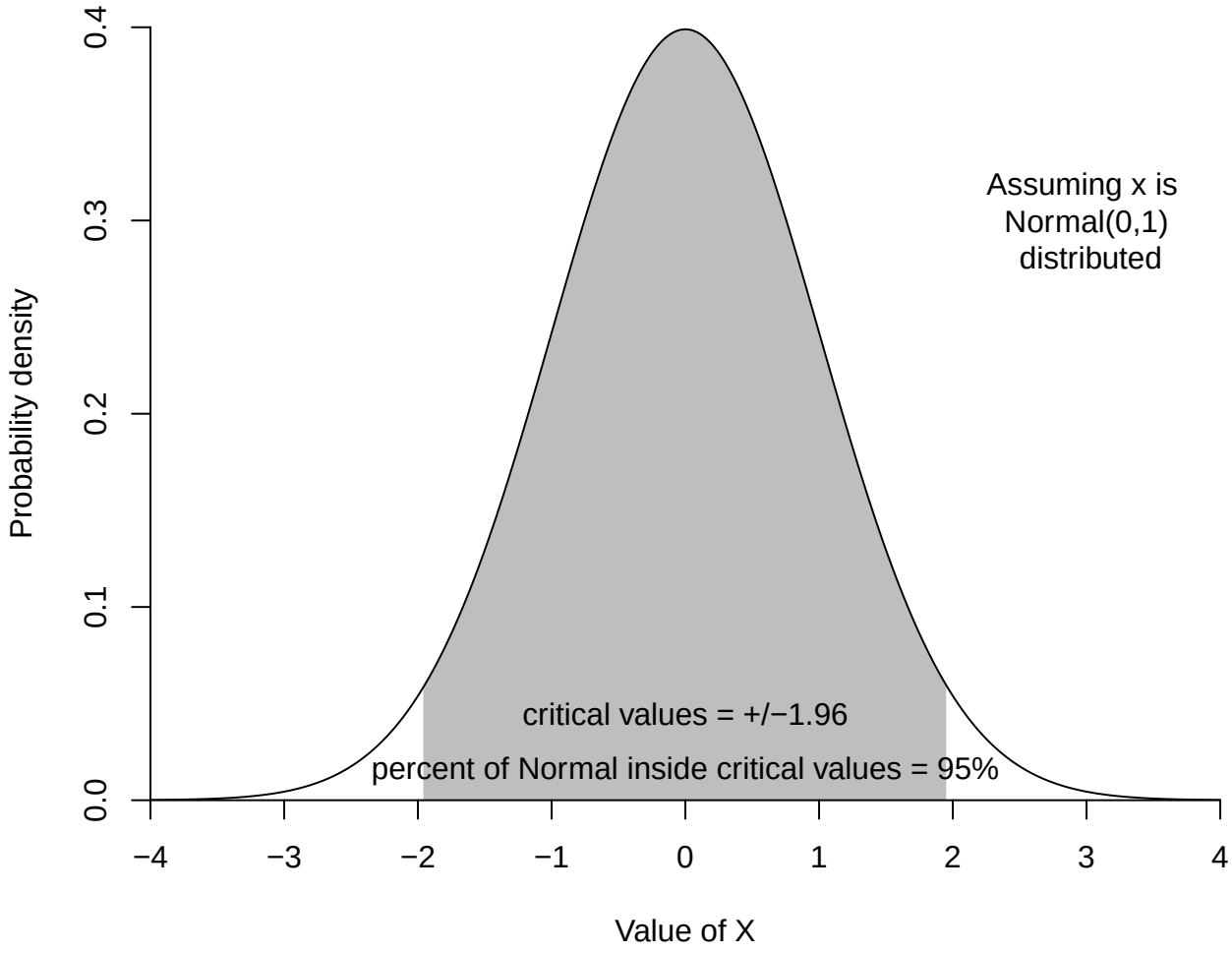
We need the critical values for the 2.5th and 97.5th percentiles (Why?)



The critical values for the 67%, 95%, and 99% intervals are memorable

Two-thirds of a Normal distribution lies within ≈ 1 standard deviation of the mean

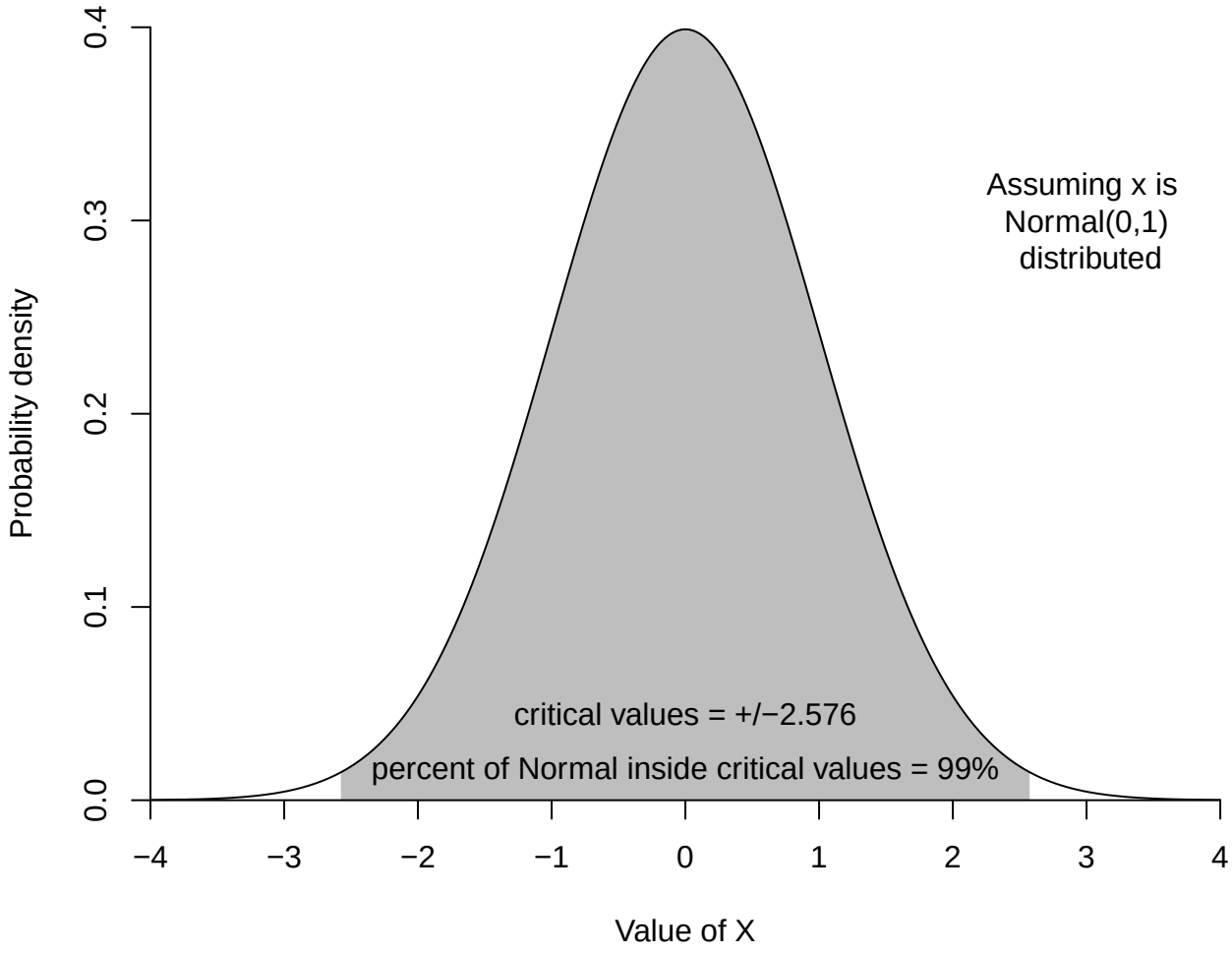
(Remember: z-scores are in standard deviation units!)



The critical values for the 67%, 95%, and 99% intervals are memorable

Two-thirds of a Normal distribution lies within ≈ 2 standard deviation of the mean

(Remember: z-scores are in standard deviation units!)



The critical values for the 67%, 95%, and 99% intervals are memorable

Two-thirds of a Normal distribution lies within ≈ 2.5 standard deviation of the mean

(Remember: z-scores are in standard deviation units!)

If the mean male height in the US is 5 ft 10 in, and the standard deviation is 3 in, *and* height is Normally distributed,

1. What critical values of height bound two-third of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 0.967 \approx$ 5 ft 7 to 6 ft 1.

If the mean male height in the US is 5 ft 10 in, and the standard deviation is 3 in, *and* height is Normally distributed,

1. What critical values of height bound two-third of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 0.967 \approx$ 5 ft 7 to 6 ft 1.

2. What critical values of height bound 95% of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 1.96 \approx$ 5 ft 4 to 6 ft 4

If the mean male height in the US is 5 ft 10 in, and the standard deviation is 3 in, *and* height is Normally distributed,

1. What critical values of height bound two-third of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 0.967 \approx$ 5 ft 7 to 6 ft 1.

2. What critical values of height bound 95% of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 1.96 \approx$ 5 ft 4 to 6 ft 4

3. What critical values of height bound 99% of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 2.576 \approx$ 5 ft 2 in to 6 ft 6 in

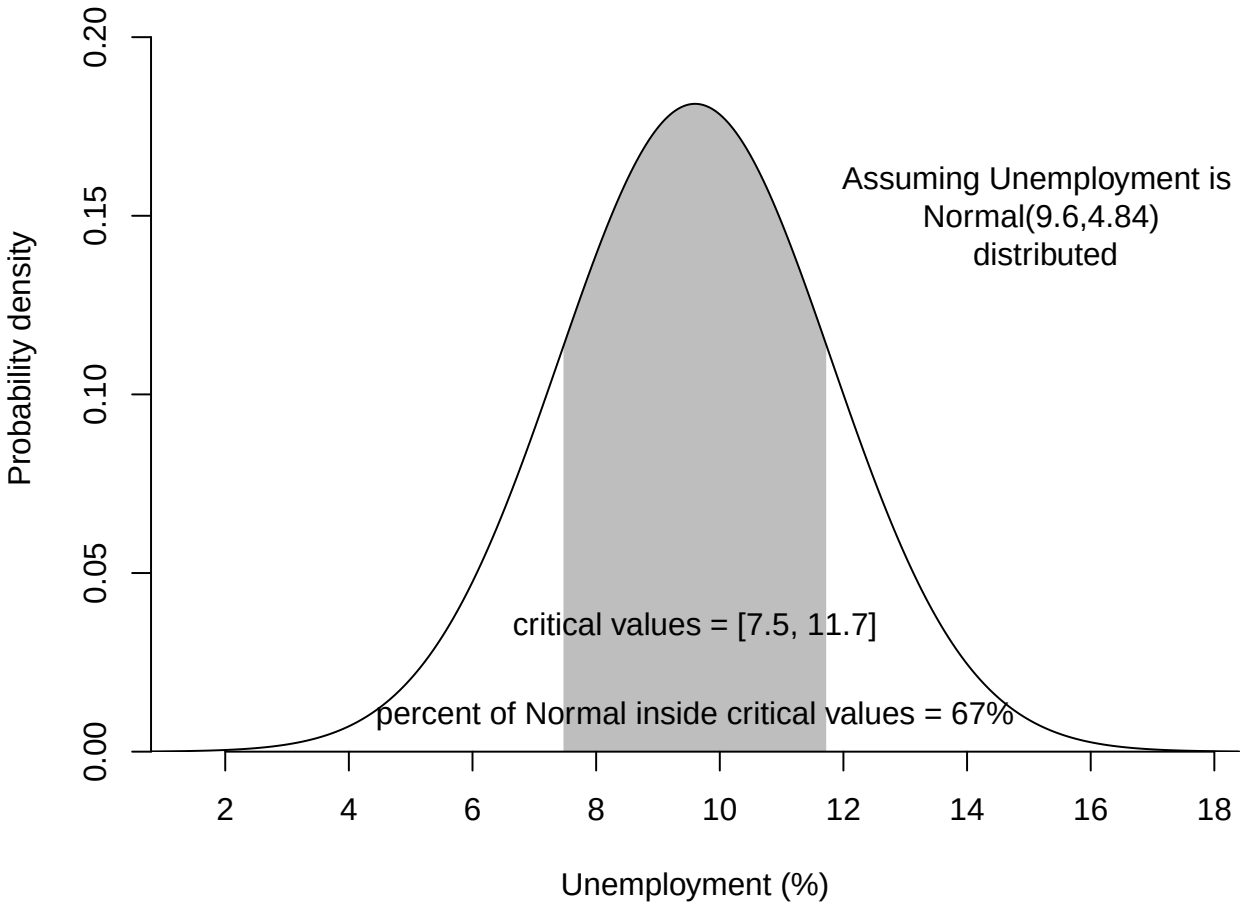
If the mean male height in the US is 5 ft 10 in, and the standard deviation is 3 in, *and* height is Normally distributed,

1. What critical values of height bound two-third of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 0.967 \approx$ 5 ft 7 to 6 ft 1.

2. What critical values of height bound 95% of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 1.96 \approx$ 5 ft 4 to 6 ft 4

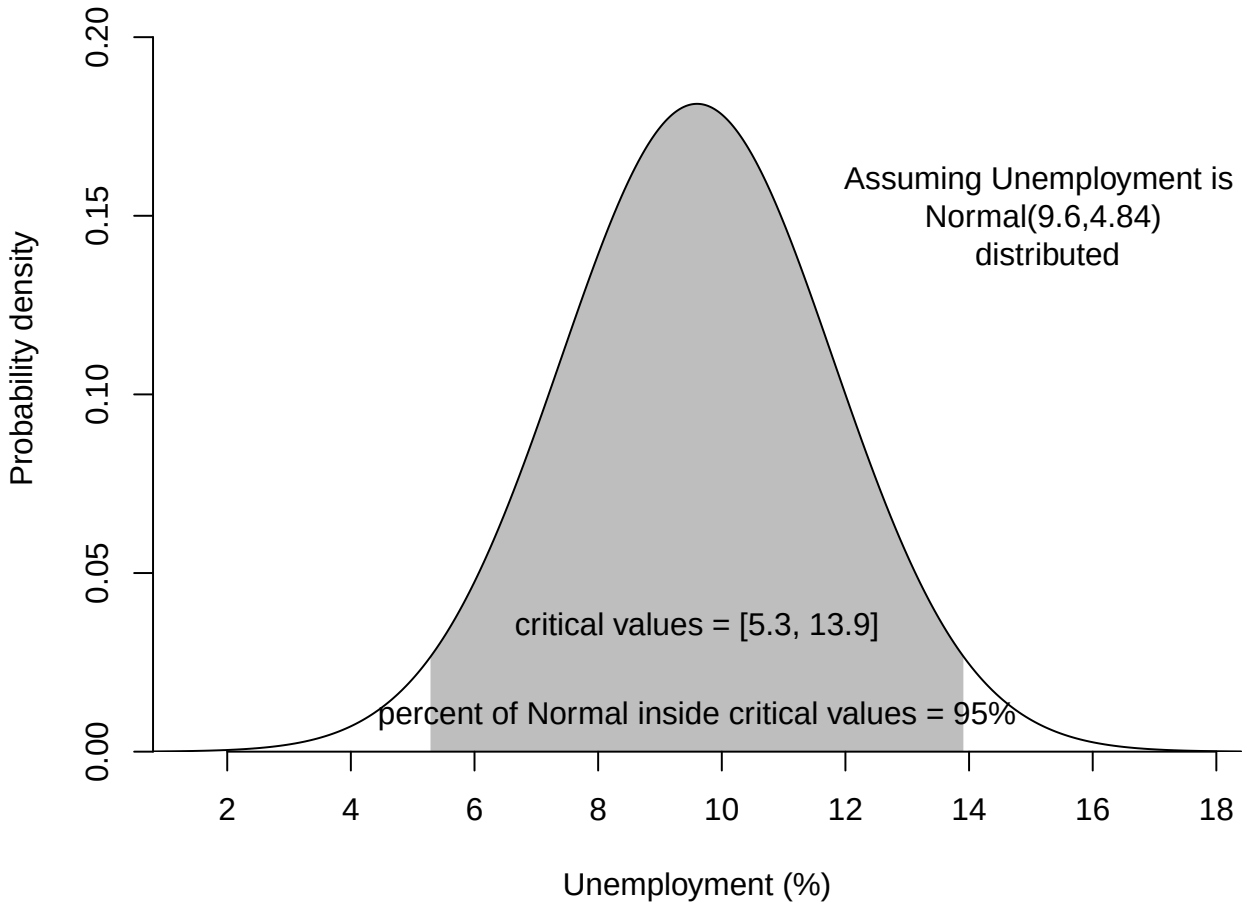
3. What critical values of height bound 99% of all men?
   $70 \text{ inches } \pm 3 \text{ inches } \times 2.576 \approx$ 5 ft 2 in to 6 ft 6 in

Warning! These statements hold only for variables that really are Normal. *Not* for all data.



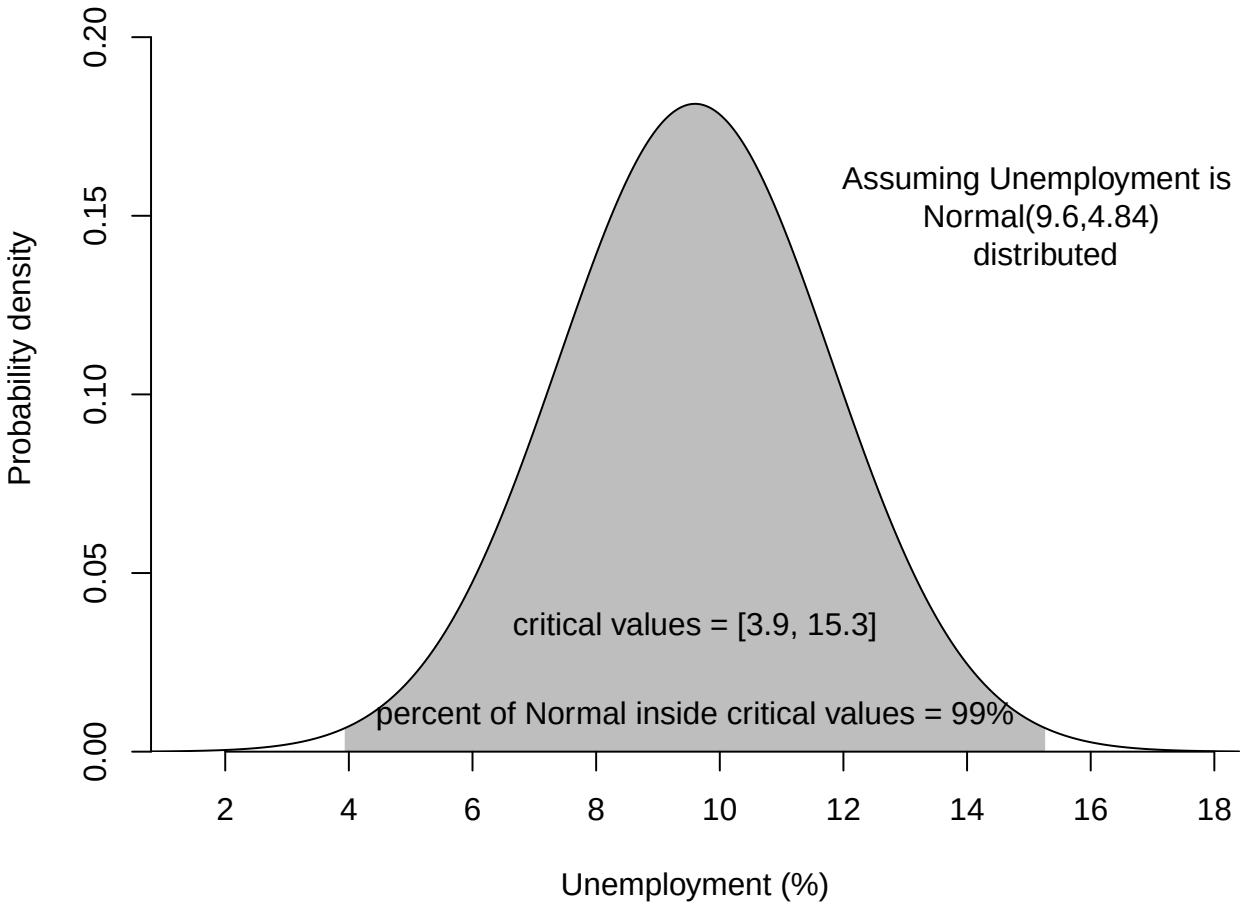
We can apply the same logic to the unemployment example

At left is the 67% interval



We can apply the same logic to the unemployment example

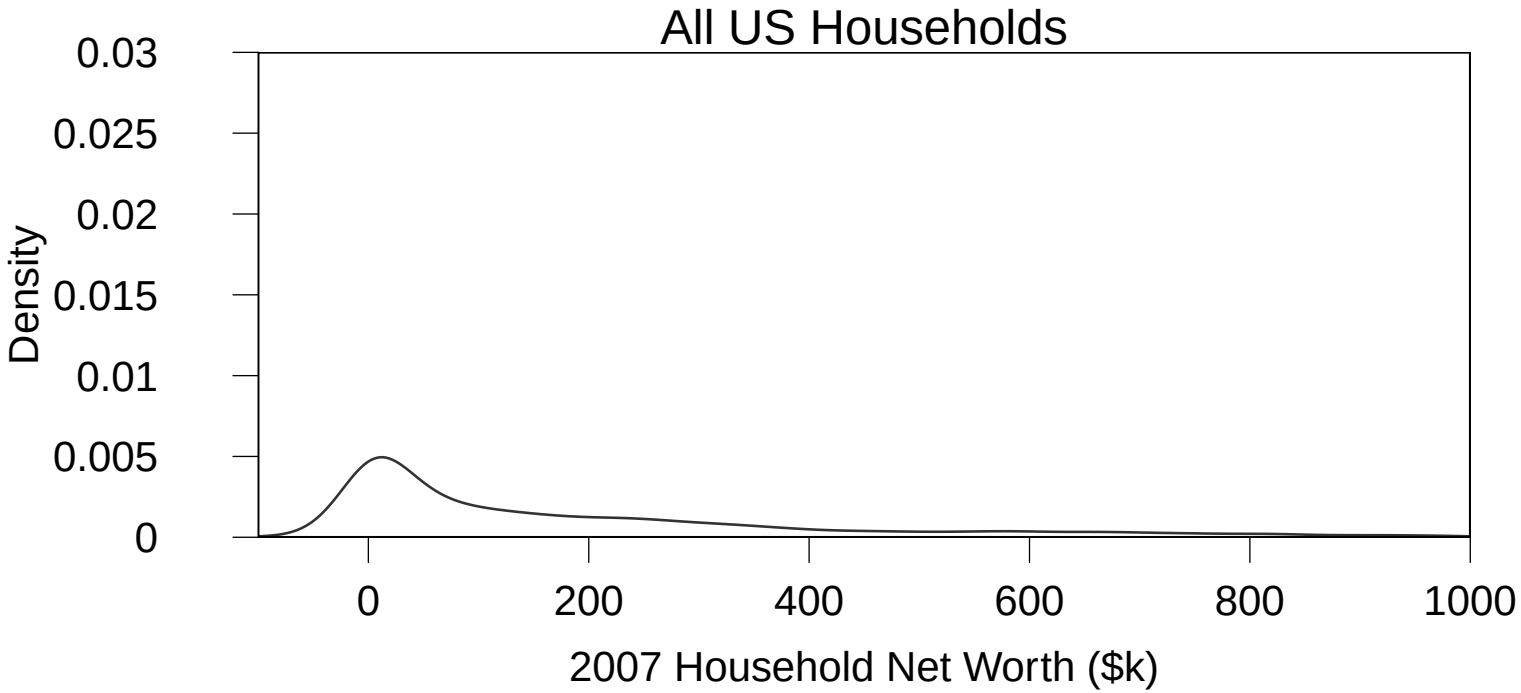
At left is the 95% interval



We can apply the same logic to the unemployment example

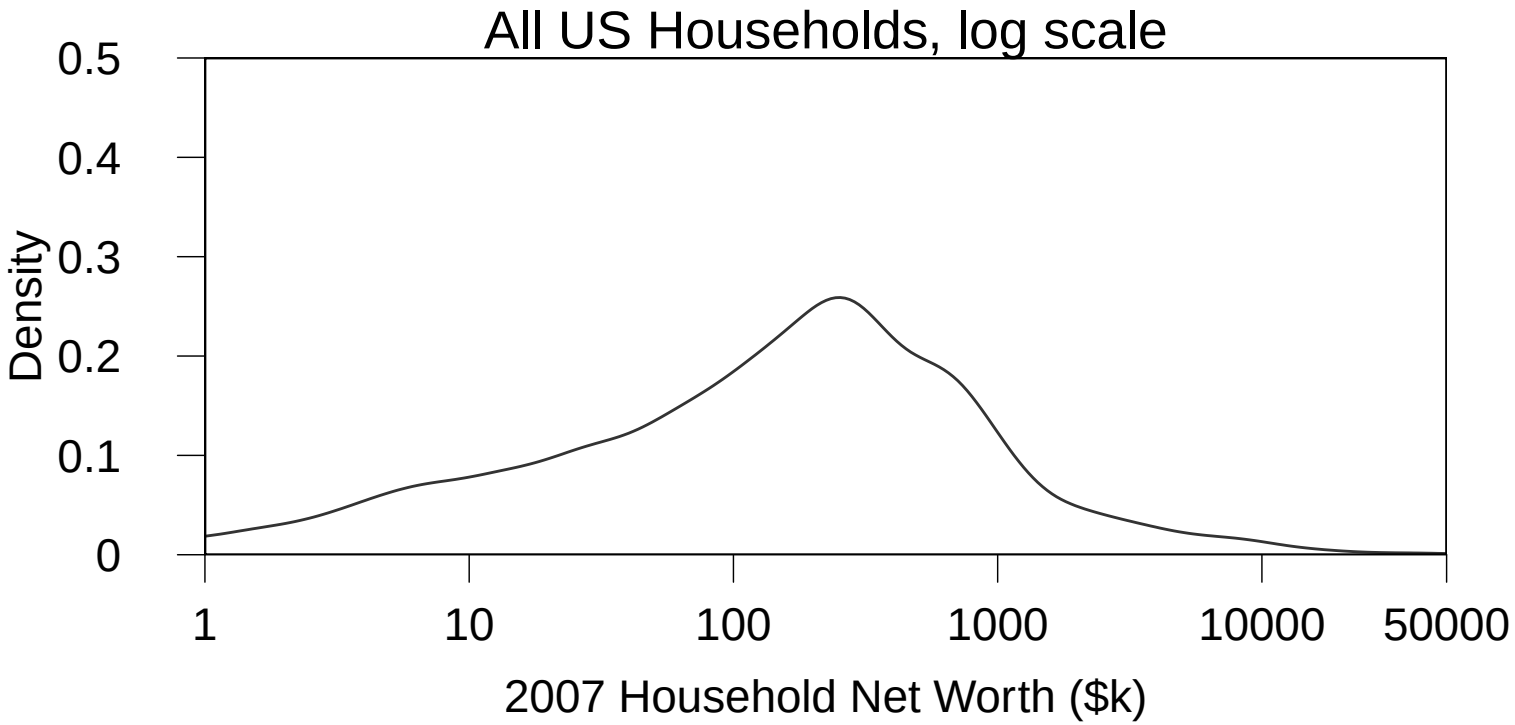
At left is the 99% interval

Comparing two distributions



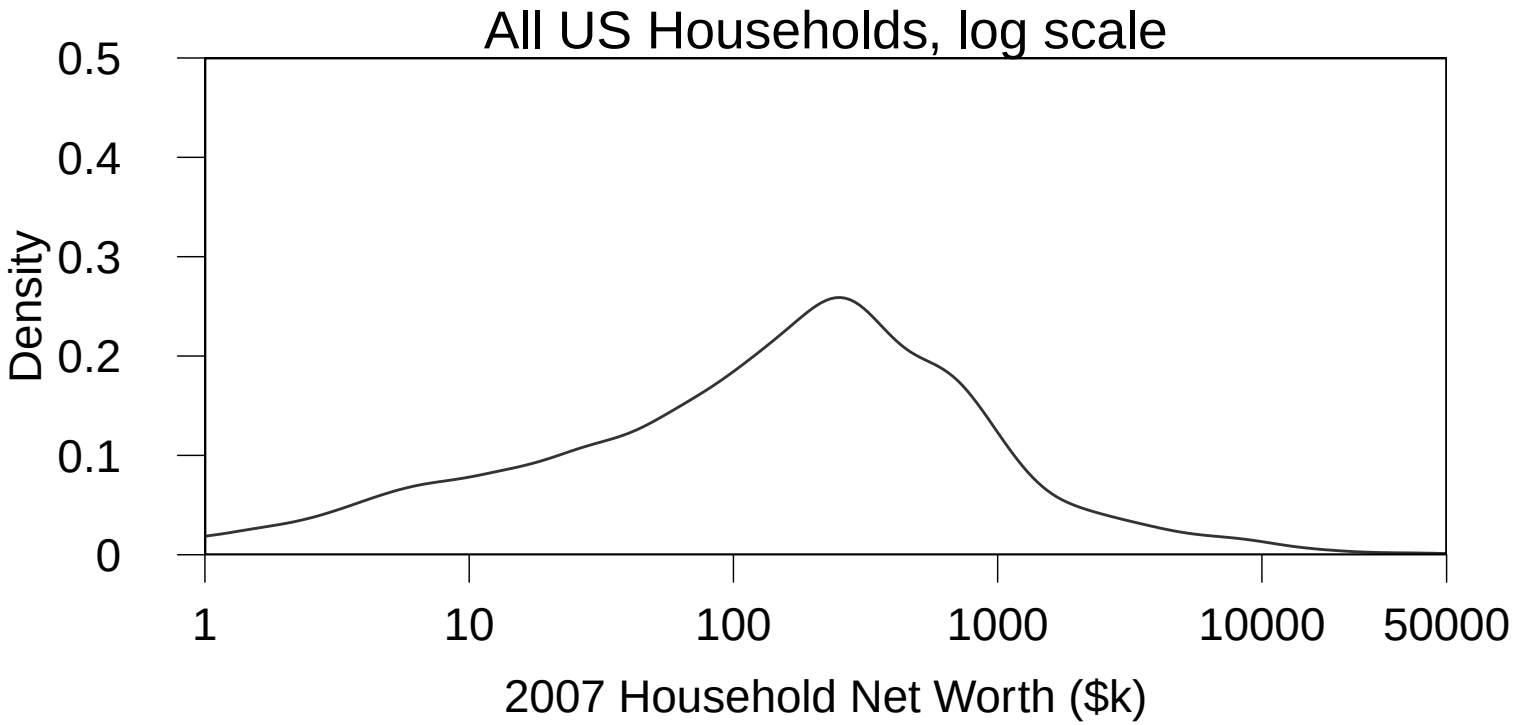
Recall the wealth example, which considered the net worth of 10000 American households in 2007

Comparing two distributions



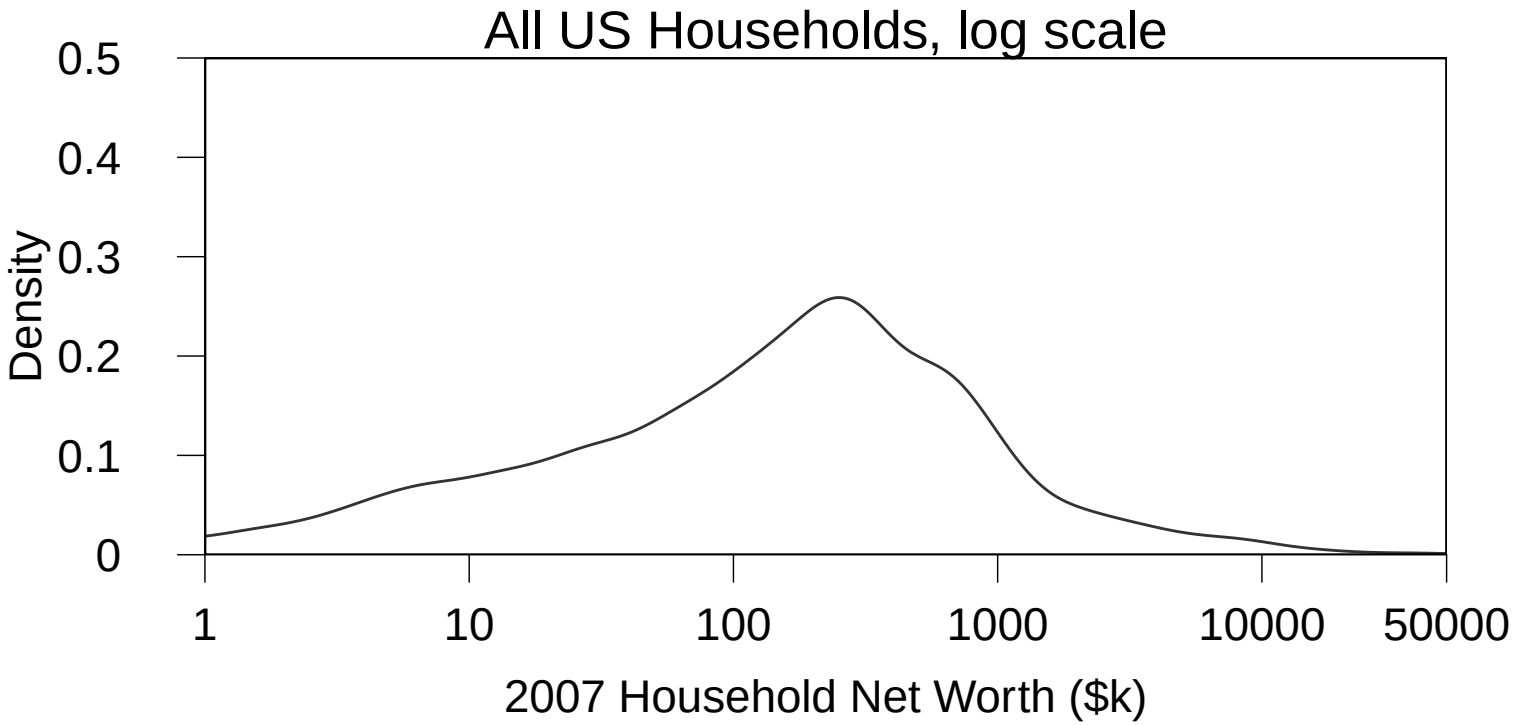
The data were skewed right on a linear scale

Comparing two distributions



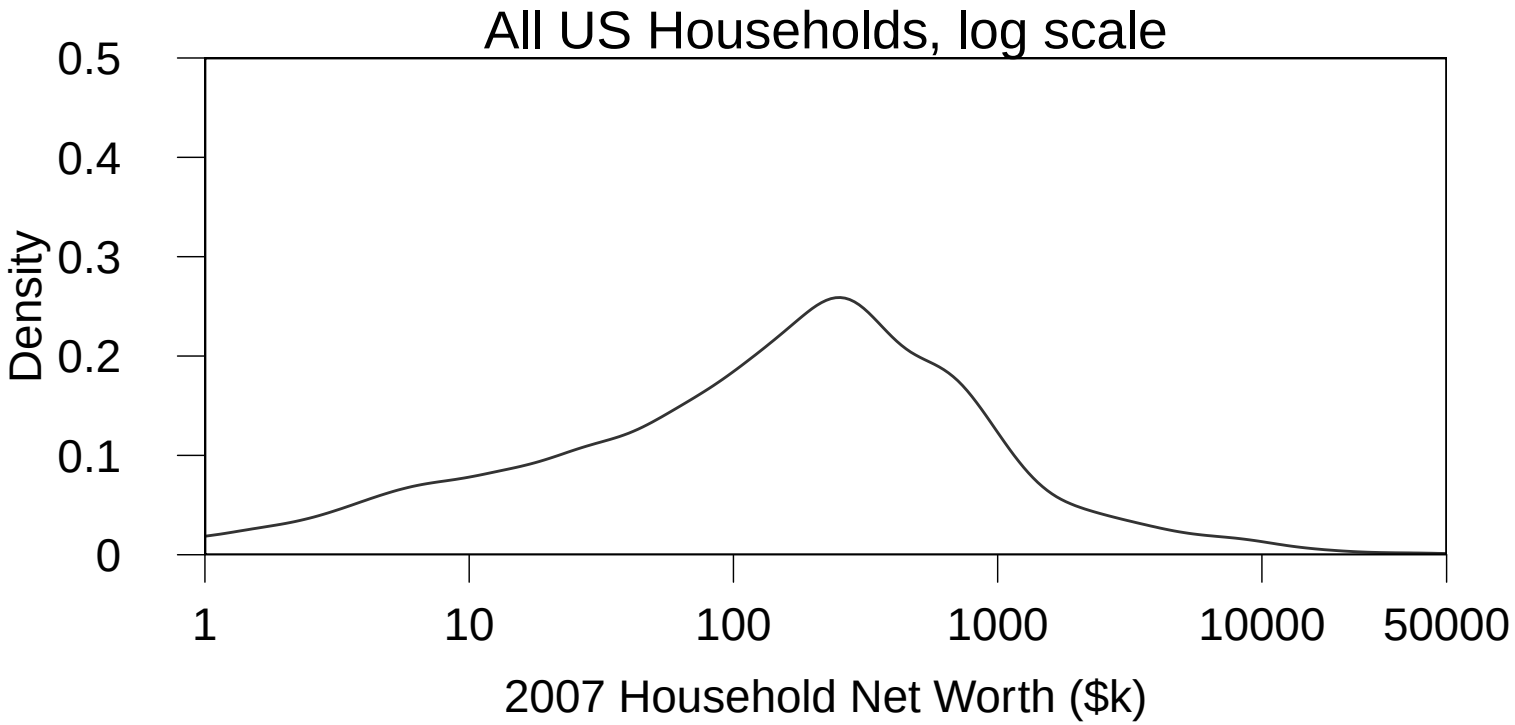
But appeared roughly symmetrically distributed on a log scale

Comparing two distributions



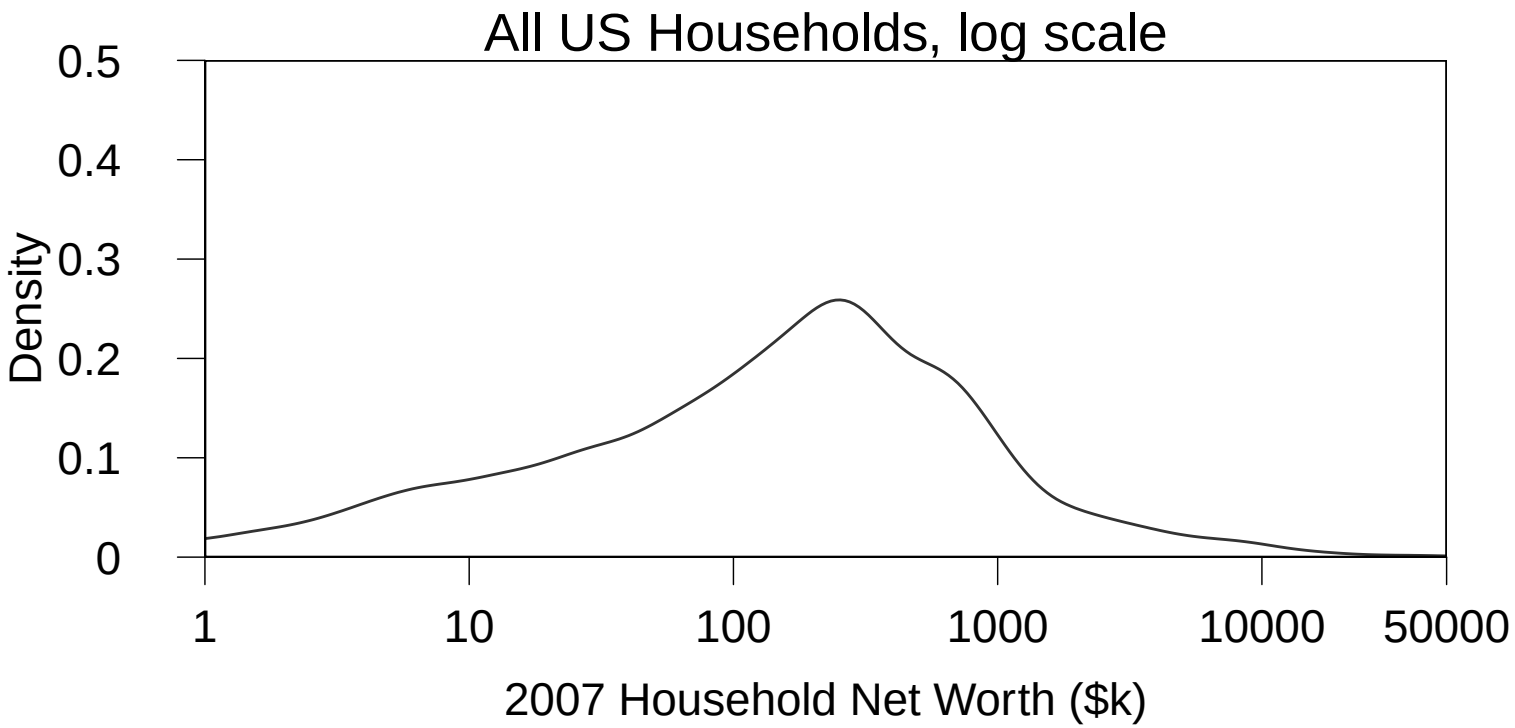
Are the logged wealth data approximately Normally distributed?

Comparing two distributions



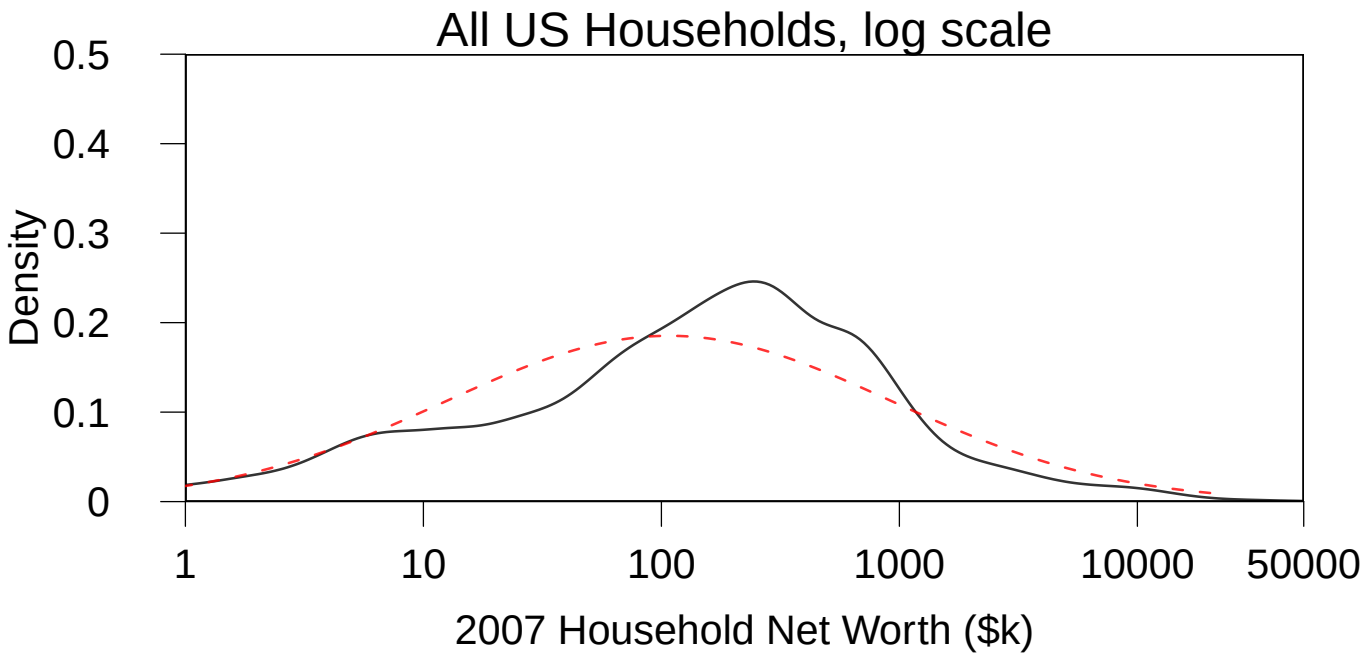
The mean log wealth is 4.62, and the variance of log wealth is 4.55.

Comparing two distributions



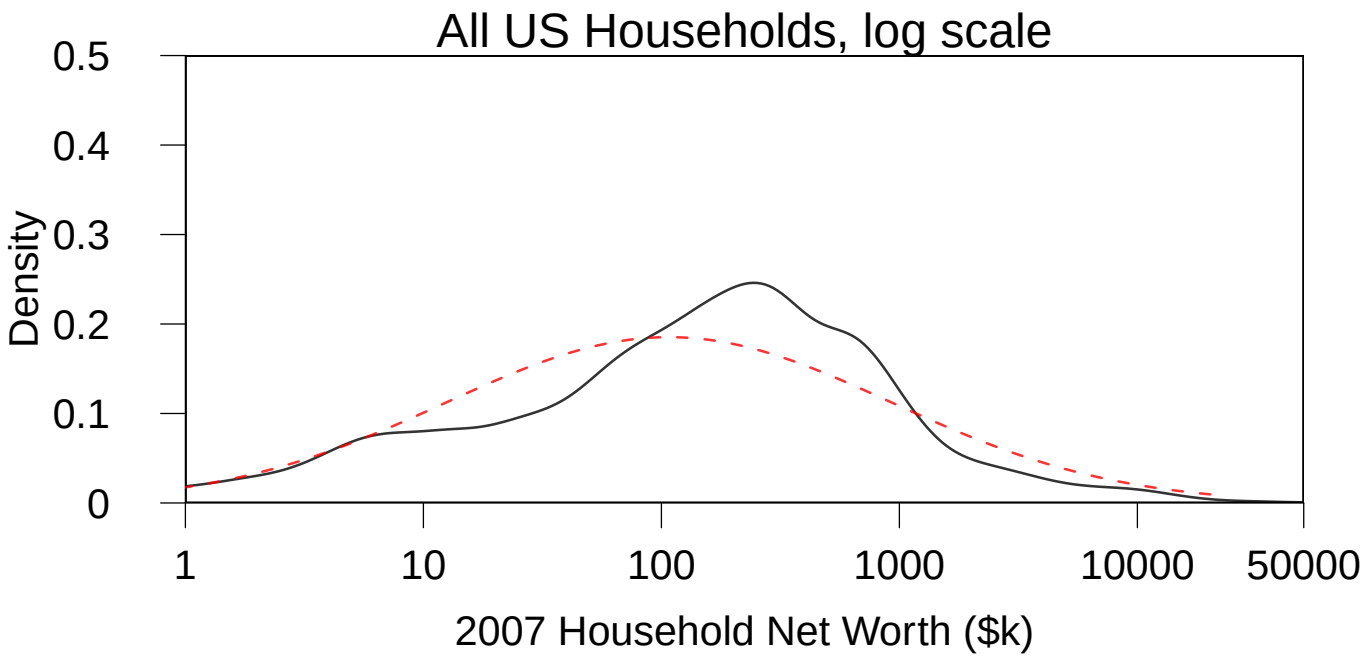
So could $\log(\text{wealth}) \sim \text{Normal}(4.66, 4.55)$?

Comparing two distributions



Compare a $\log(\text{wealth}) \sim \text{Normal}(4.66, 4.55)$ pdf to the density of the sample data.

Comparing two distributions



Looks pretty close, but with some differences.
Is there a better way to highlight them?

QQplots

One method of comparing distributions is a quantile-quantile plot, or QQplot

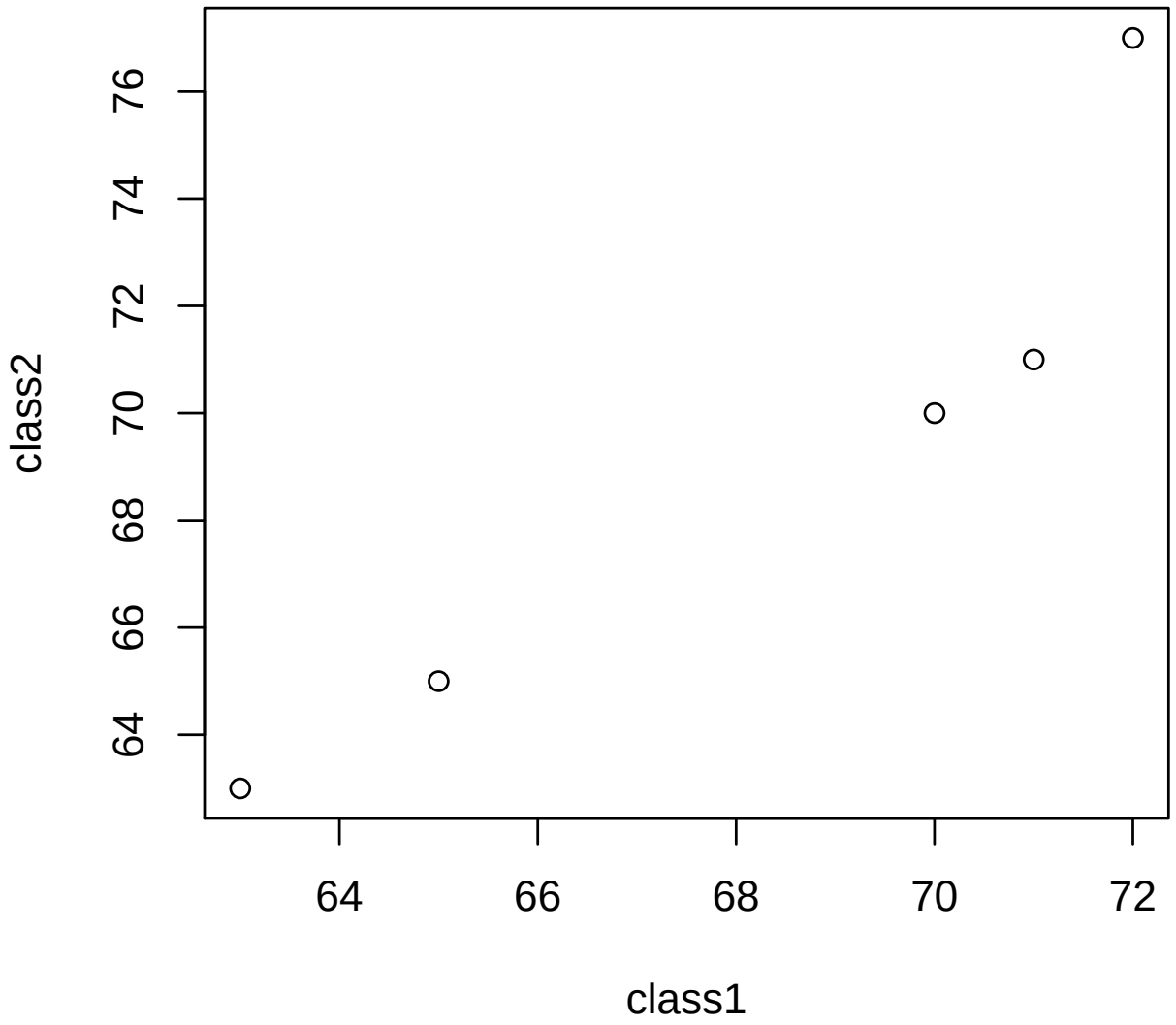
Imagine we compare the heights of students in two five person classes

Suppose four shortest people in each class have the same height, but the tallest person in Class1 is really tall:

Height1	Height2
63	63
65	65
70	70
71	71
72	77

These five datapoints are also the $0.2, 0.4, 0.6, 0.8, 1.0$ quantiles for each class

QQplots



Let's plot the values (quantile for class 1, quantile for class 2) for each quantile in a scatterplot

This QQplot helps detect differences in the distributions

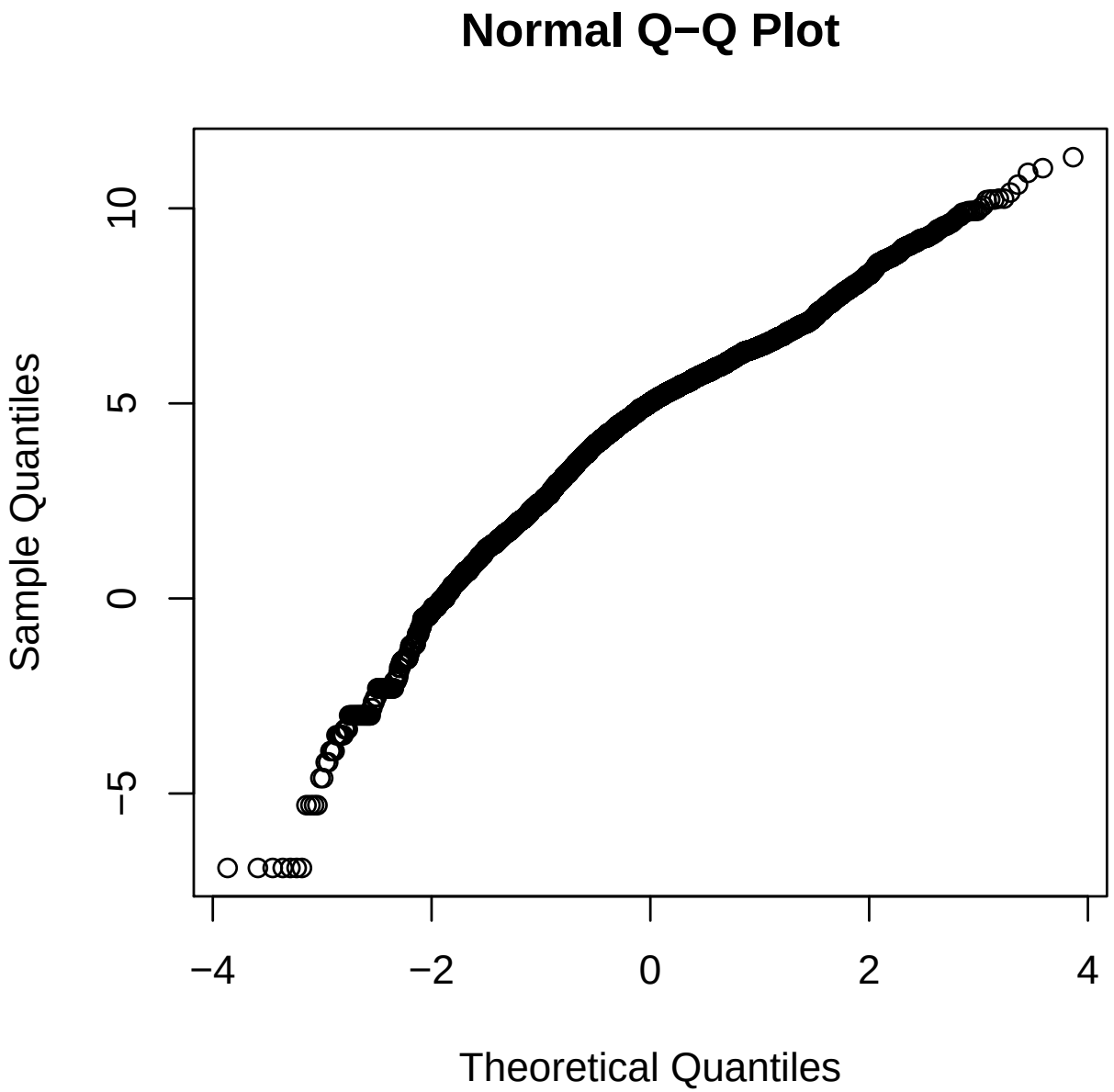
QQplots against a theoretical distribution

QQplots are most useful for plotting empirical quantiles against a theoretical distribution

That is, how does the 10th percentile of log(wealth) line up to the 10th percentile of a $\text{Normal}(4.66, 4.55)$? The 20th percentile? Etc.

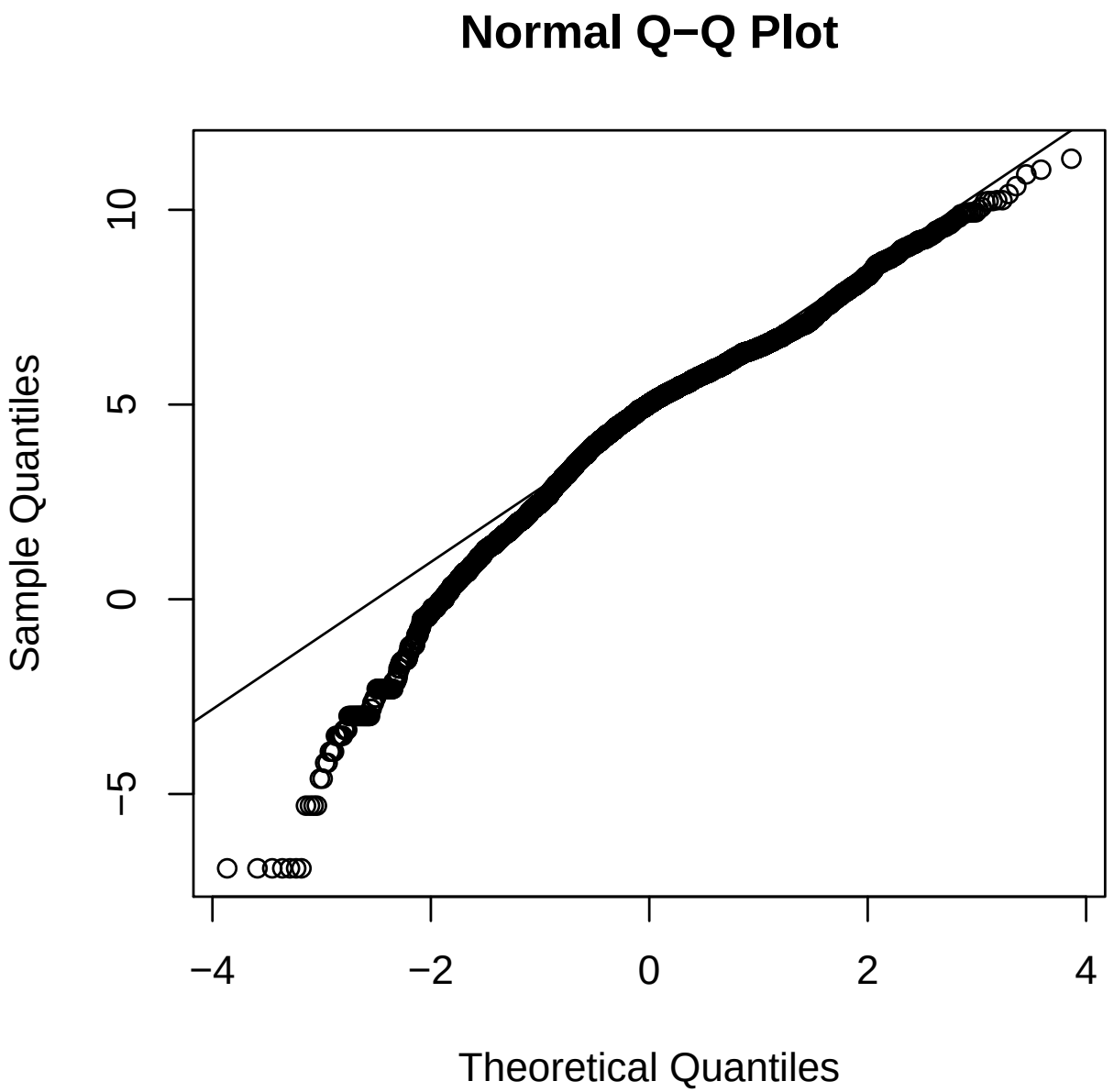
We just plot as many quantiles as we can, and look for deviations from a straight line

QQplots



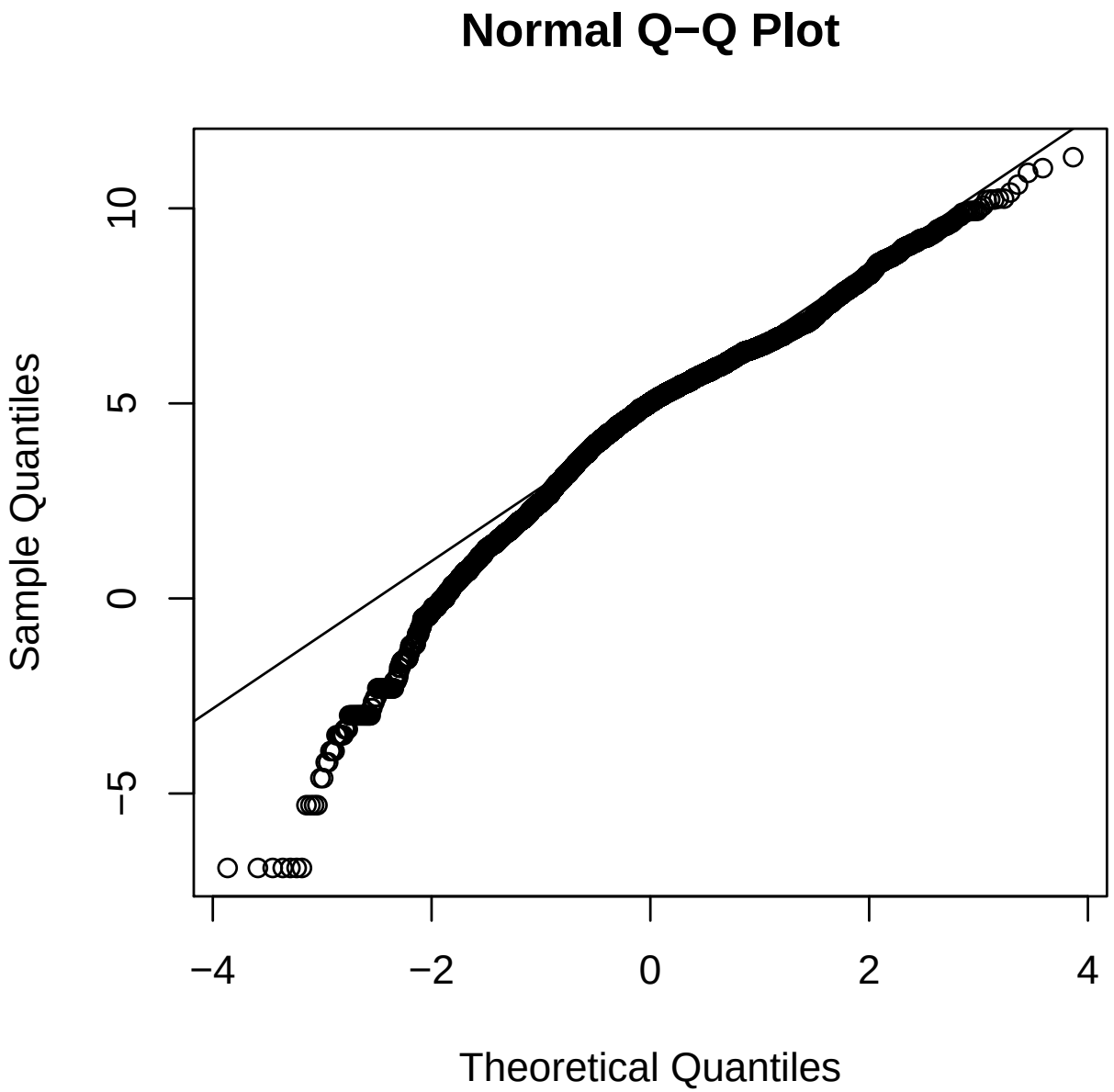
What does this shape suggest?

QQplots



Is the wealth data log-normal?

QQplots



Is the wealth data log-normal? In what way might it deviate?

What features would a more appropriate distribution have?