PM 592
Regression Analysis for Public Health Data Science

Week 2

Probability

Probability

The Normal Distribution

Other Distributions

Lecture Objectives

- Describe the normal distribution and ways it is used.
- Use the normal distribution to calculate probabilities.
- Explain the Central Limit Theorem and its applications.
- Describe other distribution types and their use.

1. Review

- ✓ Study types
- ✓ Variable types
- ✓ Methods for exploratory data analysis

Let X be a random variable, measured on a sample of N subjects from a target population.

There is a lot we don't know about X for the population:

- What is the mean of X?
- What is the variation in X?

X could be something like:

- Observed airspeeds of European unladen swallows
- Number of words tweeted from a user's account in a given day
- Amount of calories individuals consume at a Vegas buffet

But we can use **statistics** from our sample to make an inference about what population **parameters** might be.

Data characteristic	Measure	Population parameter	Sample statistic	Unit
Central tendency	Mean	μ	$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$	Original scale
Dispersion	Variance	σ^2	$S^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N-1}$	Square of original scale
Dispersion	Standard deviation	σ	$S = \sqrt{S^2}$	Original scale

Descriptive statistics are focused on just describing our sample.
Inferential statistics involves the estimation of population parameters.

We will discuss the following distributions:

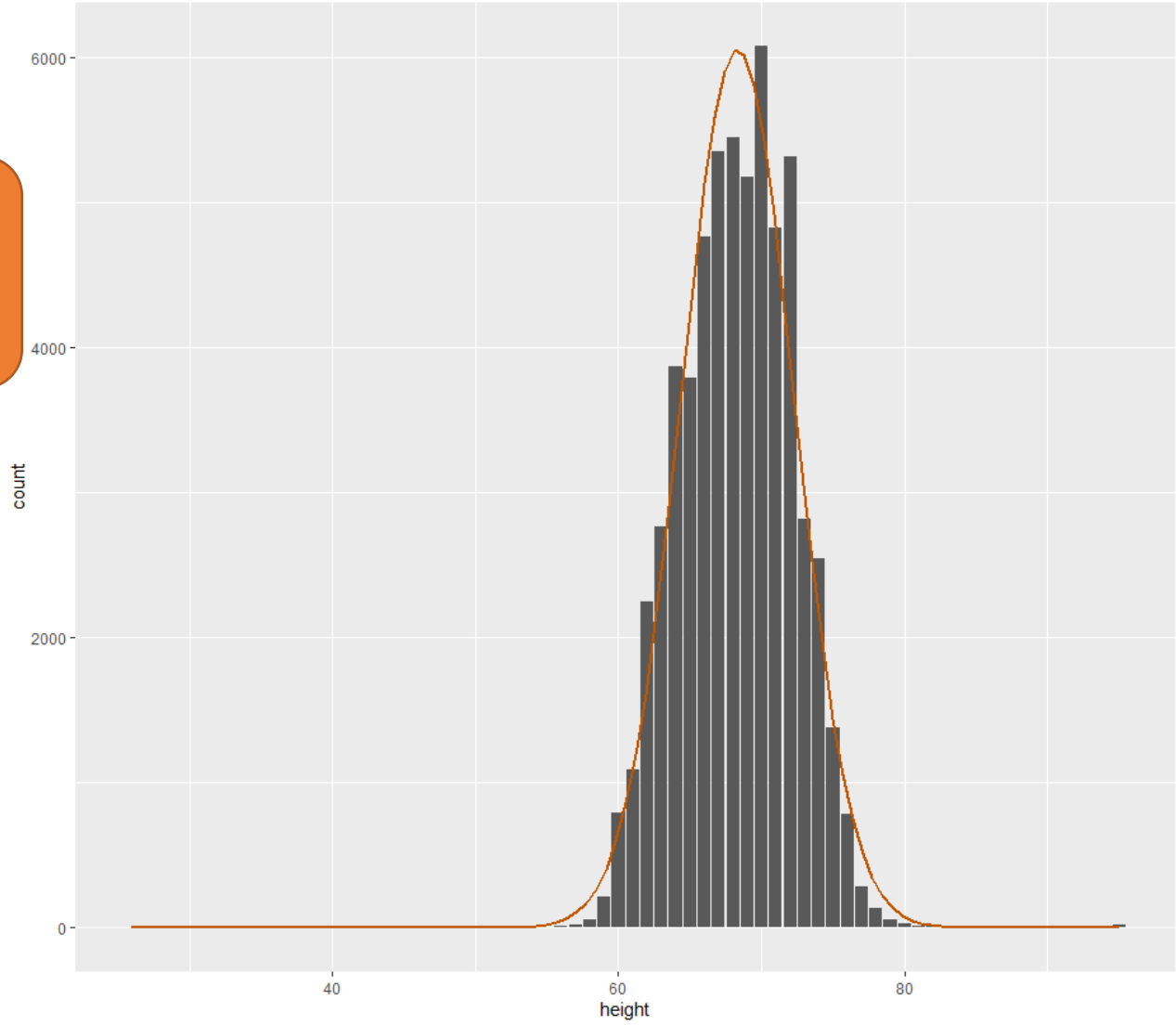
- Normal
- T
- Chi-square
- F
- Geometric (Bernoulli)
- Binomial

The normal distribution

- Perhaps the most important distribution
- X~ N(μ, σ)

In theory, one needs to know the population mean and standard deviation to define the normal distribution.

Many phenomena follow the normal distribution. On the right is the height distribution of approximately 60,000 OKCupid users, with a normal curve overlaid.



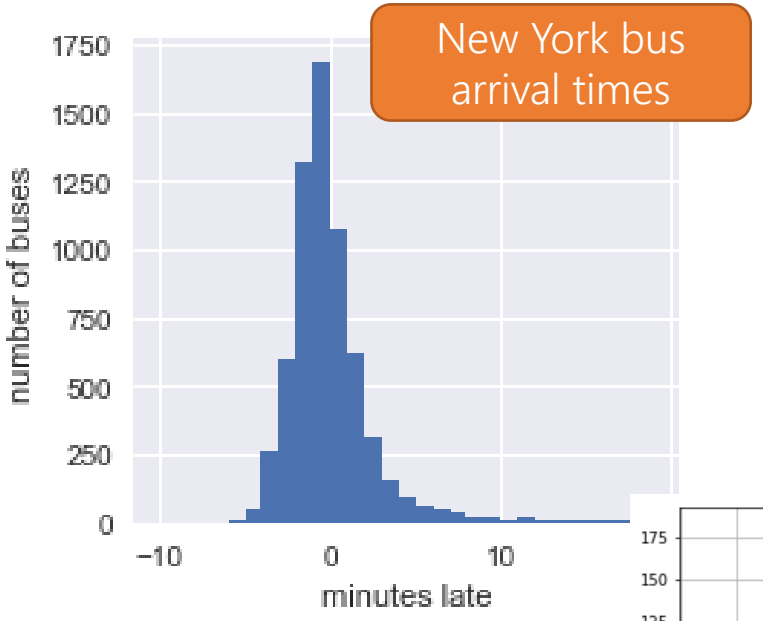
The normal distribution is frequently used:

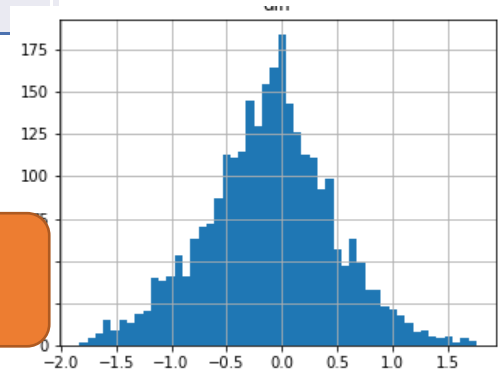
- To **approximate a distribution** of a continuous variable X.
- To compute **z-scores**.
- To find **probabilities** that a score X falls in a certain range.
- To approximate the **sampling distribution of the mean**.

The normal distribution is important because of how common it is!

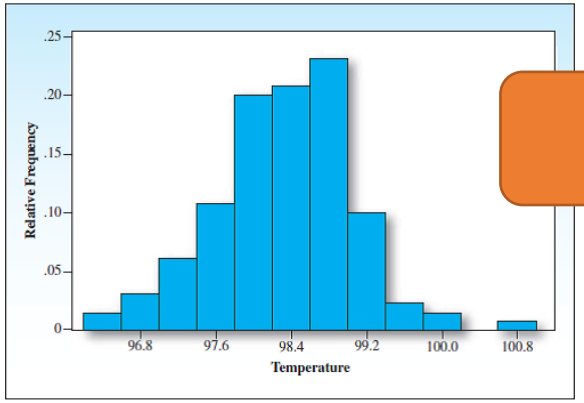
And even if something doesn't directly follow a normal distribution, we can frequently approximate its distribution as normal.

This makes it easy to compute probabilities.





Sentiment in Londoners' tweets about the NHS.



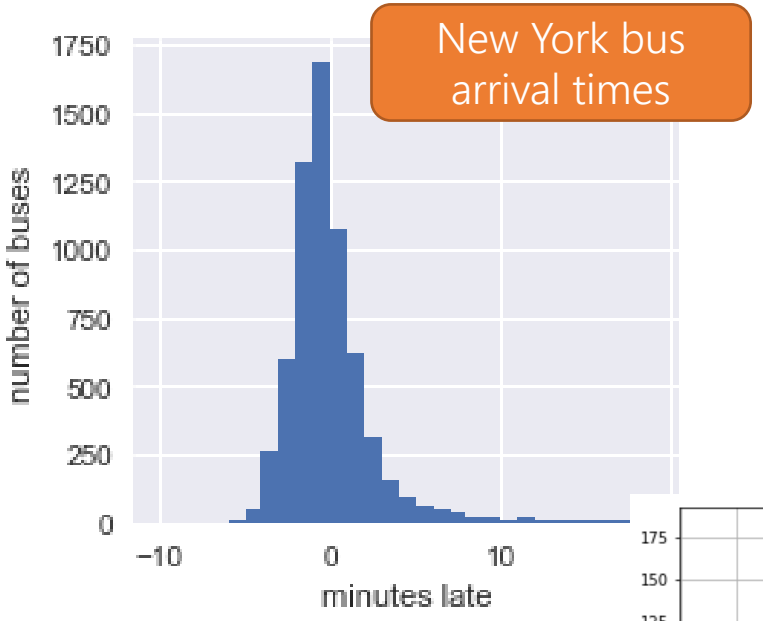
Body temperature of healthy individuals.

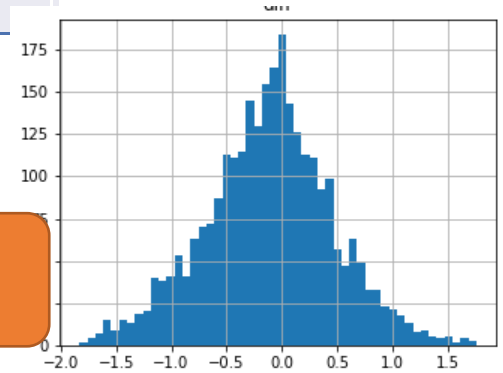
One problem:

All these distributions have different means and standard deviations.

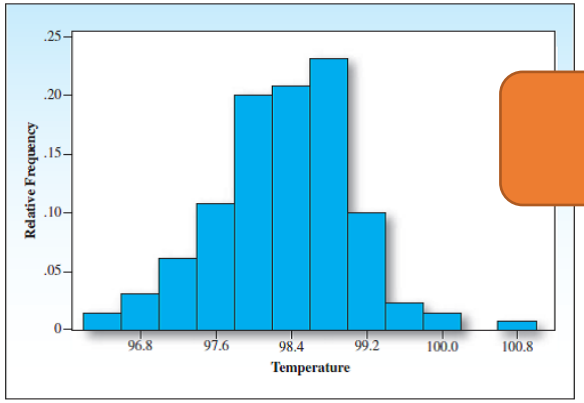
Therefore we frequently want to convert our observed distribution to the <u>standard normal distribution</u>.

The standard normal distribution has a mean of 0 and standard deviation of 1.





Sentiment in Londoners' tweets about the NHS.

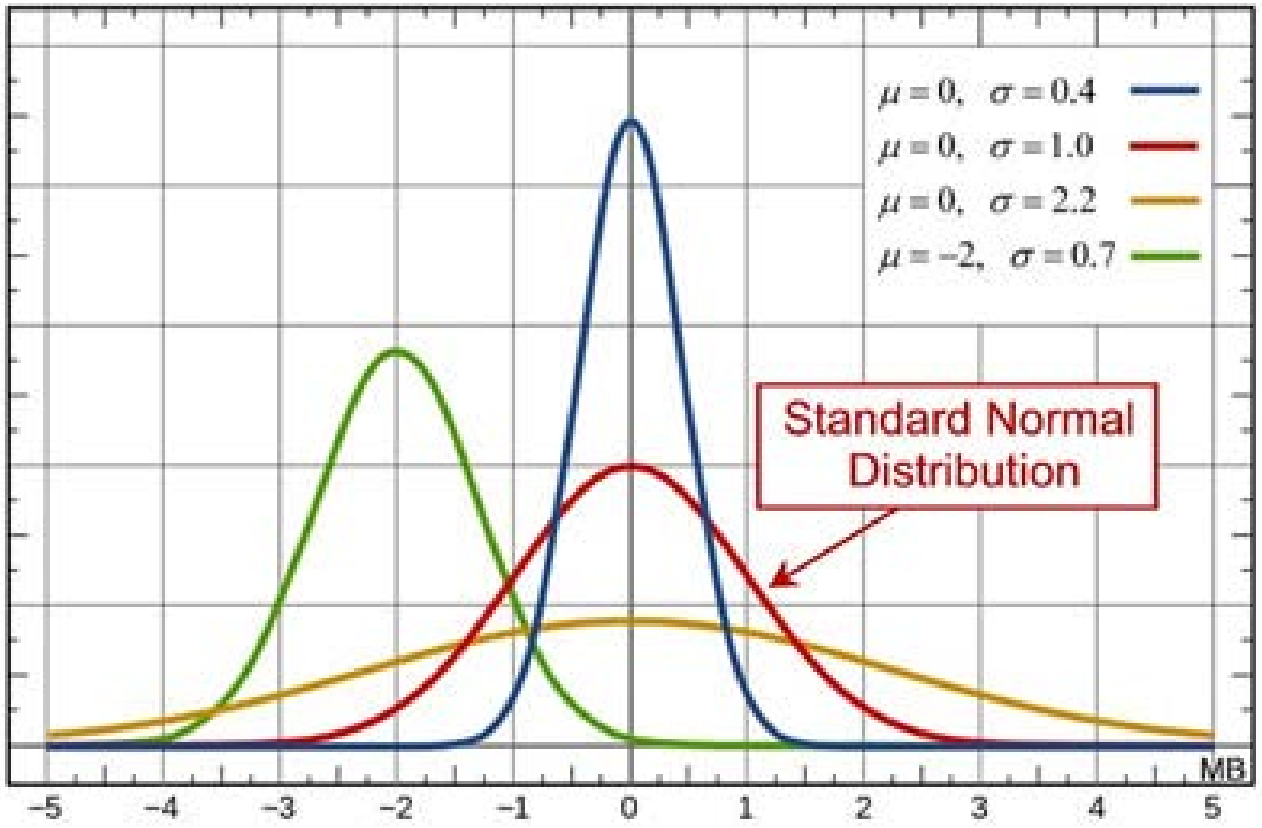


Body temperature of healthy individuals.

We want to convert these to a standard normal distribution because we know a lot about that particular distribution.

For any distribution of variable *X*, we can convert the raw scores to **standardized scores** (aka "Z" scores) by the following:

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



The normal distribution is frequently used:

- To **approximate a distribution** of a continuous variable X.
- To compute **z-scores**.
- To find **probabilities** that a score X falls in a certain range.
- To approximate the **sampling distribution of the mean**.

Suppose we observe heights (in inches) of these OkCupid users.

```
> okc %>% select(height)
# A tibble: 59,946 x 1
   height
    <dbl>
 1     75
 2     70
 3     68
 4     71
 5     66
 6     67
 7     65
 8     65
 9     67
10     65
# ... with 59,936 more rows
```

We can create z-scores manually, or use a function.

```
> mean(okc$height, na.rm=T)
[1] 68.3016

> sd(okc$height, na.rm=T)
[1] 3.944652

> # Create height z-scores
> okc <-
+   okc %>%
+   mutate(height_z = (height - mean(height, na.rm=T))/sd(height, na.rm=T))

> okc %>% select(starts_with("height"))
# A tibble: 59,946 x 2
   height height_z
    <dbl>    <dbl>
 1     75   1.70
 2     70   0.431
 3     68  -0.0765
 4     71   0.684
 5     66  -0.583
 6     67  -0.330
 7     65  -0.837
 8     65  -0.837
 9     67  -0.330
10     65  -0.837
# ... with 59,936 more rows
```

Someone 75" tall is 1.70 standard deviations taller than average.

Someone 65" tall is 0.837 standard deviations shorter than average.

This tells us information about each person *relative to others in the sample* and can be useful when comparing several variables on different scales.

```
> okc %>% select(height, height_z)
# A tibble: 59,946 x 2
   height height_z
    <dbl>    <dbl>
 1     75   1.70
 2     70   0.431
 3     68  -0.0765
 4     71   0.684
 5     66  -0.583
 6     67  -0.330
 7     65  -0.837
 8     65  -0.837
 9     67  -0.330
10     65  -0.837
# ... with 59,936 more rows
```

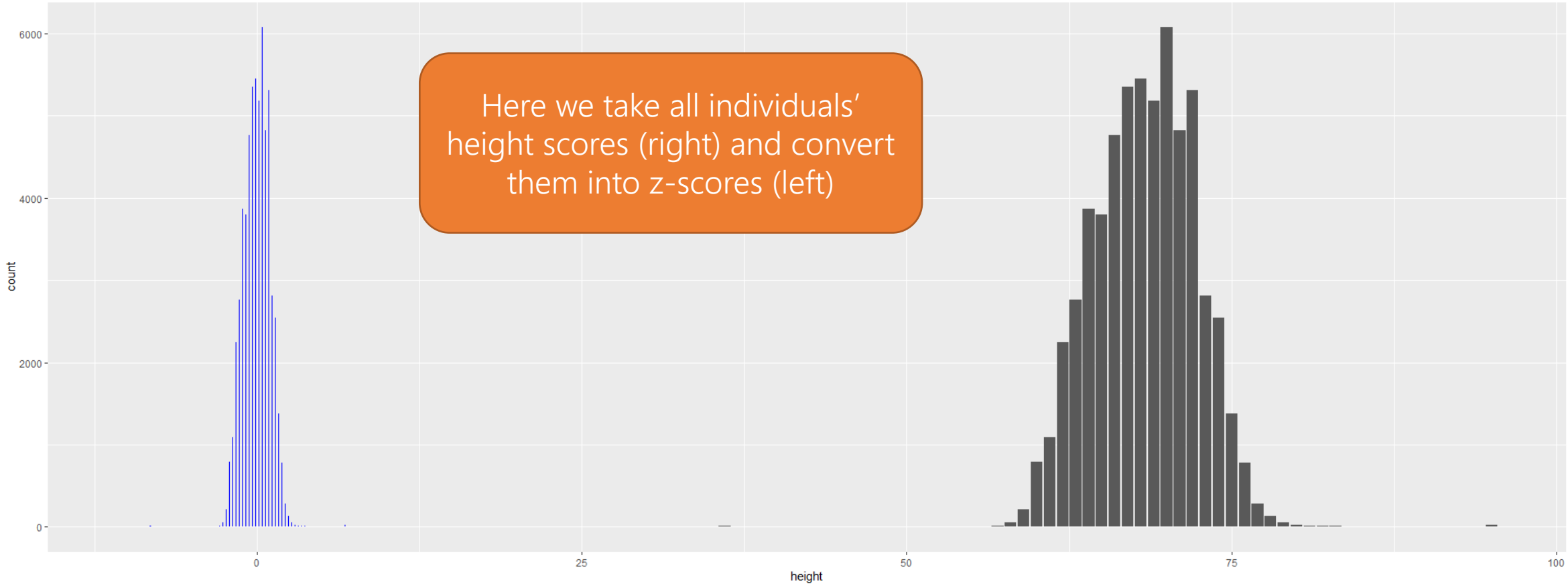
We can create z-scores manually, or use a function.

```
> scale(data$ftime)
…
attr(,"scaled:center")
[1] 68.3016
attr(,"scaled:scale")
[1] 3.944652
```

This effectively does two things:

1) **Shifts** the distribution.

2) **Scales** the distribution.



Practice Question 1

What can we glean from this analysis?

```
> mean(cereals$calories)
[1] 106.8831

> cereals <-
+   cereals %>%
+   mutate(calories_cent = calories - mean(calories))

> cereals %>%
+   select(name, starts_with("calories")) %>%
+   filter(name %in% c("Cheerios", "Frosted Mini-Wheats", "Raisin Bran"))
# A tibble: 3 x 3
  name                calories calories_cent
  <chr>                  <dbl>         <dbl>
1 Cheerios                 110          3.12
2 Frosted Mini-Wheats      100         -6.88
3 Raisin Bran              120         13.1
```

a) The mean calories per serving of cereals in this data set is 100.

b) Cheerios has 3.12 standard deviations more calories per serving than the average cereal.

c) Cheerios has 3.12 more calories per serving than the average cereal.

d) Frosted Mini-Wheats is the best cereal.

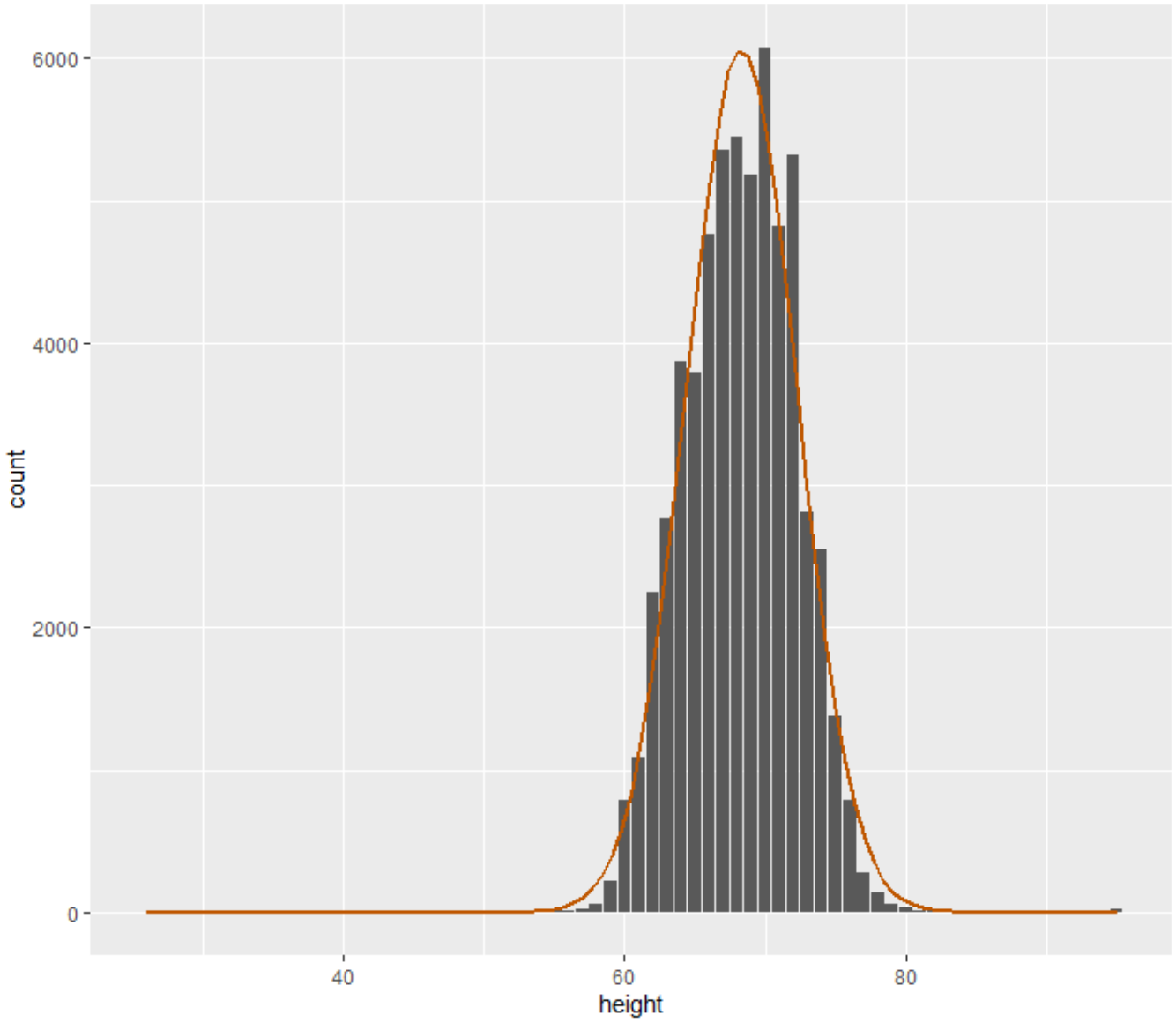
The normal distribution is frequently used:

- To **approximate a distribution** of a continuous variable X.
- To compute **z-scores**.
- To find **probabilities** that a score X falls in a certain range.
- To approximate the **sampling distribution of the mean**.

As we saw before, the mean height was 68.3”with a standard deviation of 3.9”.

Is this distribution perfectly normal? **No.**

Can we approximate the distribution as being normal? **Yes.**



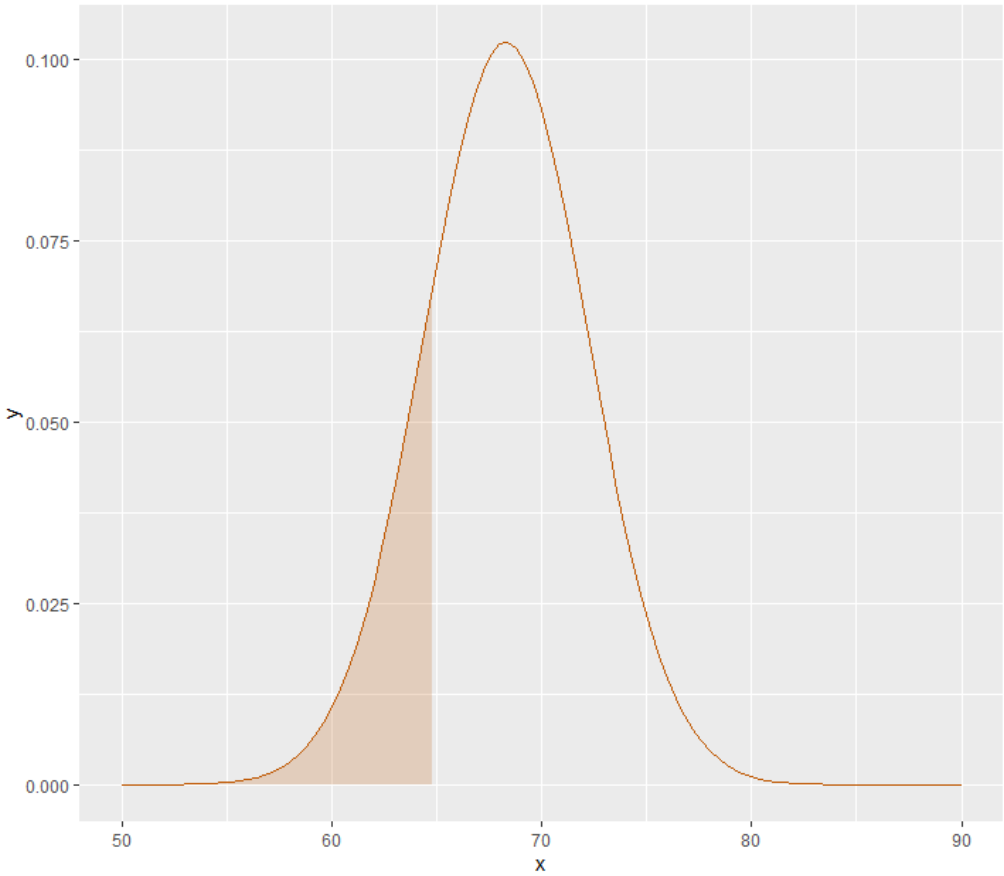
By approximating distributions as normal, we can find probability values corresponding to a given z-value.

Assume height is normally distributed – what is the probability an OkCupid user has a height < 65”?

X = 65

Z = (65-68.3)/3.9 = -0.846

```
> pnorm(-0.846)
[1] 0.1987764
```



There’s a 19.88% chance of someone being shorter than 65”.

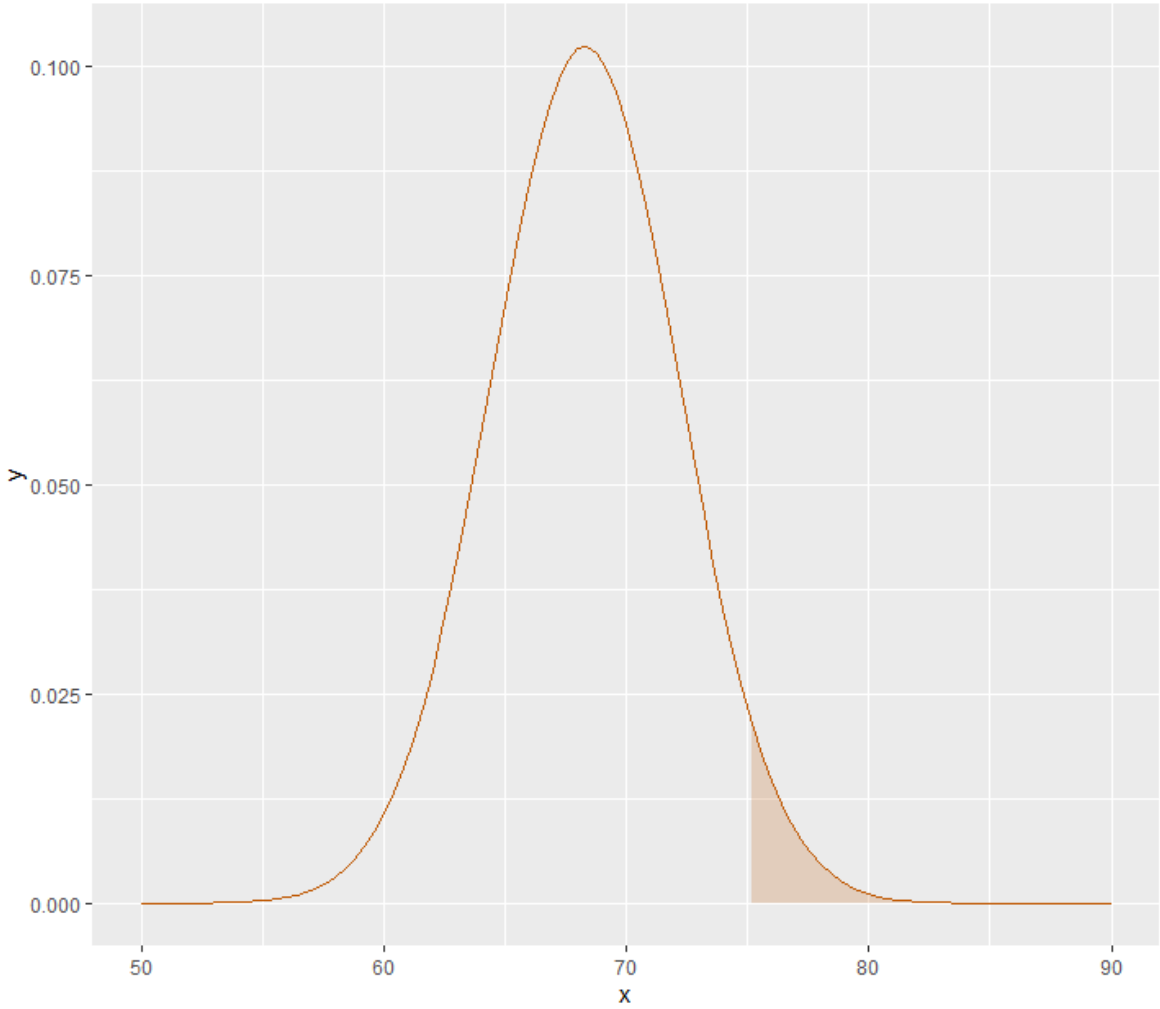
Assume height is normally distributed – what is the probability an OkCupid user has height > 75”?

X = 75

Z = (75-68.3)/3.9 = 1.72

```
> pnorm(1.72)
[1] 0.9572838

> 1-pnorm(1.72)
[1] 0.04271622
```

There’s a 95.7% chance someone is shorter than 75”, so there is a 4.3% chance someone is taller.

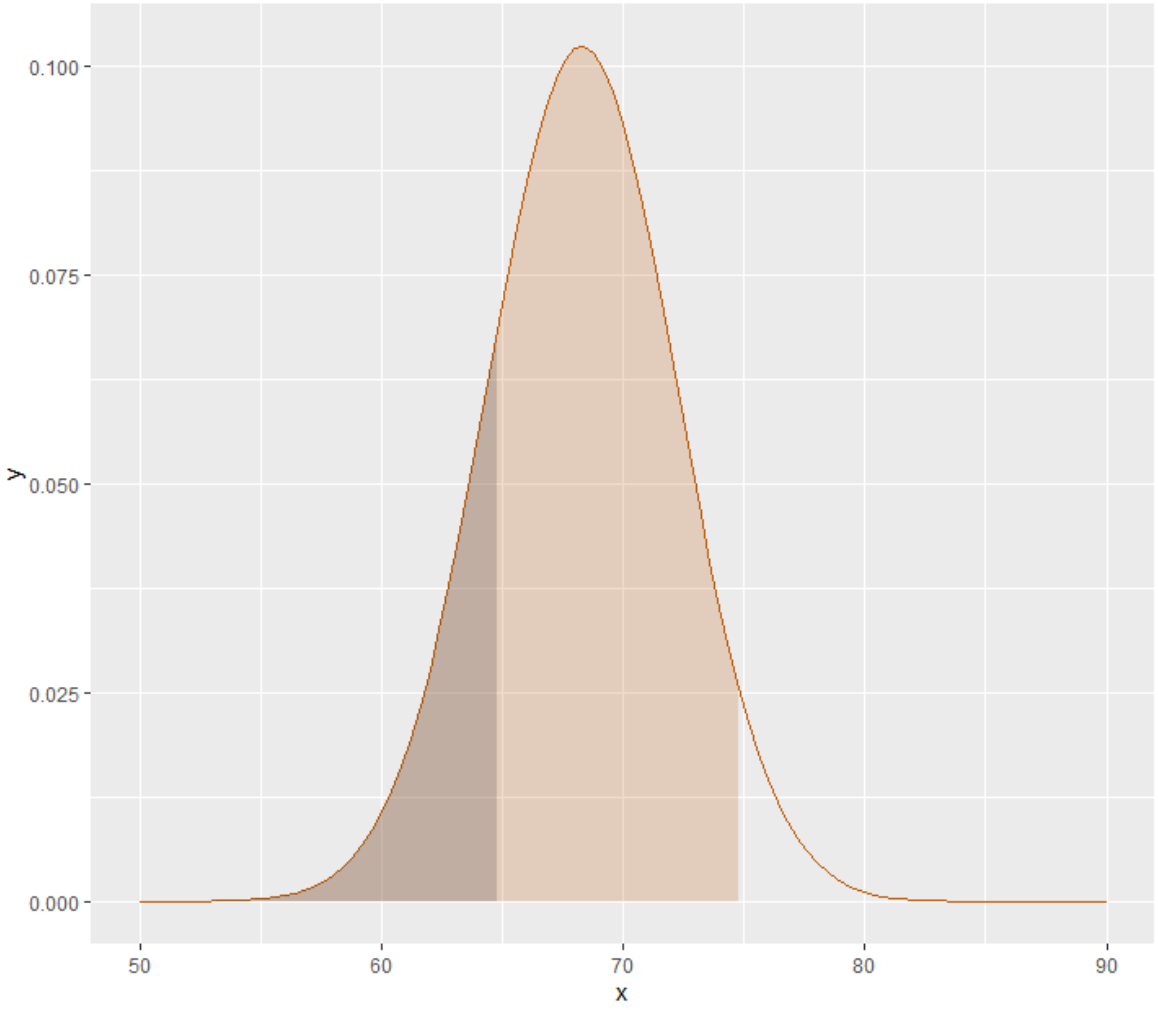


Assume height is normally distributed – what is the probability an OkCupid user would have height between 65" and 75"?

$P(X<75) = 0.957$

$P(X<65) = 0.199$

$P(65<X<75) = 0.957-0.199 = 0.758$

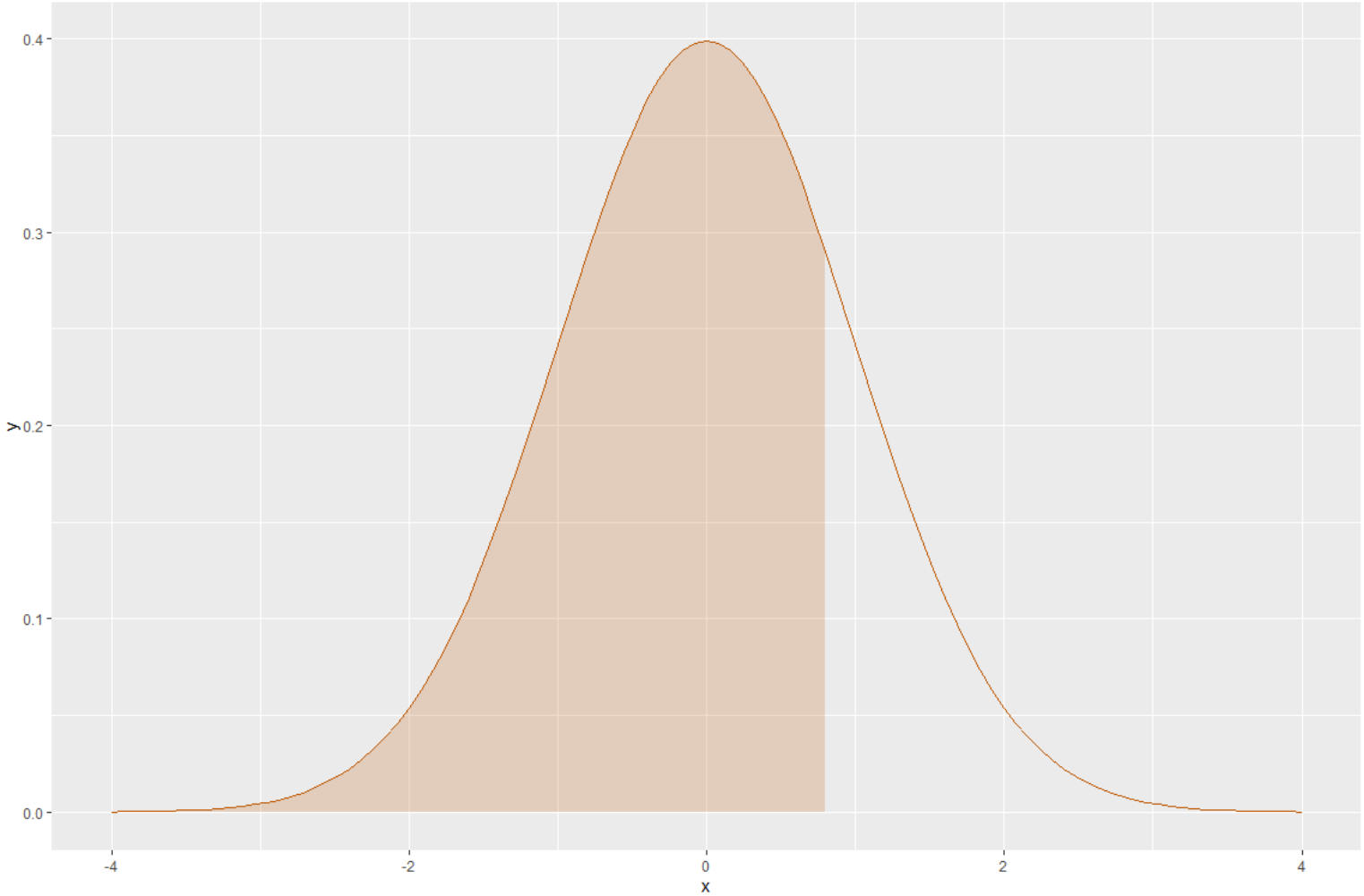


There's a 75.8% chance an individual would be between 65" and 75".

What is the Z-value corresponding to the 80th percentile?

$P(X<Z) = 0.80$

```
> qnorm(0.80)
[1] 0.8416212
```



Practice Question 2

If somebody scored in the 99th percentile on a cognitive test, what would their z-score be?

The normal distribution is frequently used:

- To **approximate a distribution** of a continuous variable X.
- To compute **z-scores**.
- To find **probabilities** that a score X falls in a certain range.
- To approximate the **sampling distribution of the mean**.

Americans' average height is 66". Is there any evidence that the mean height of OkCupid users is different from 66"?

The Z-test is the most basic type of statistical hypothesis test. Is the mean of our population equal to a hypothesized mean?

$$H_0: \mu = 66$$
$$H_A: \mu \neq 66$$

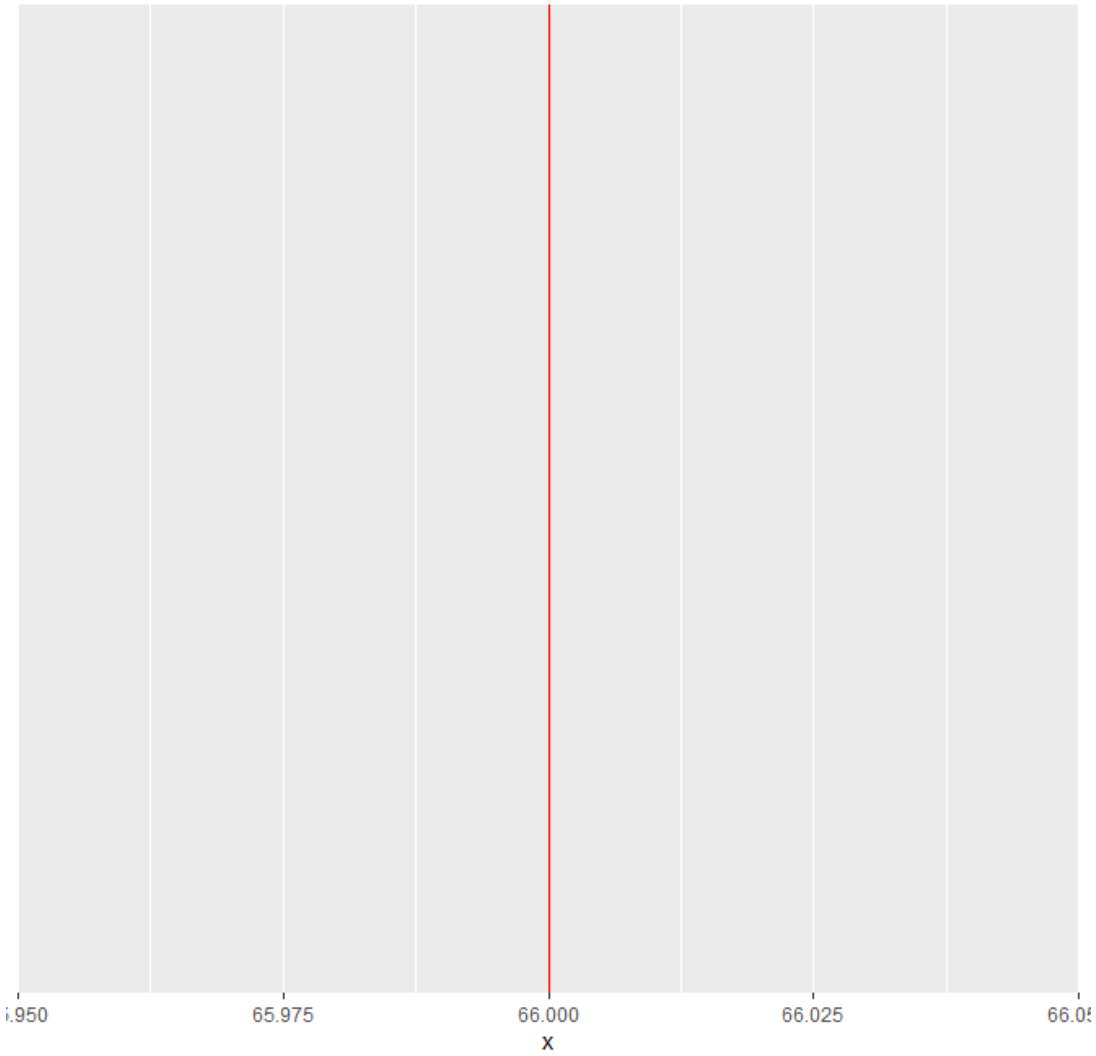
We can tackle this in three steps:

1) Assume that we're looking at a sample that DOES come from a population with $\mu = 66$.

2) Examine the distribution of means we would observe <u>if</u> we took a sample mean from this population.

3) Figure out how unlikely it would be to observe the mean we observed under the null hypothesis.

1) Assume that we're looking at a sample that DOES come from a population with $\mu = 66$.



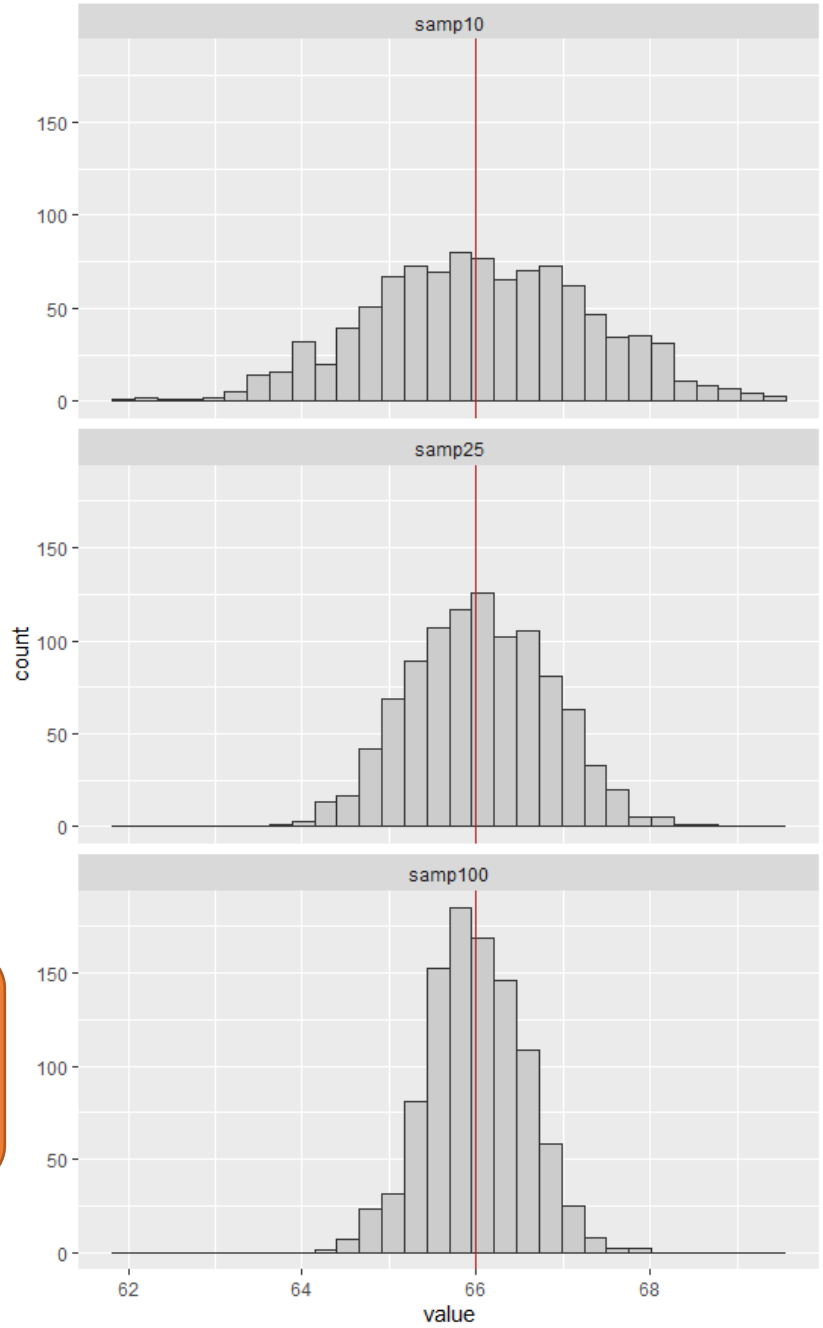
2) Examine the distribution of means we would observe <u>if</u> we took a sample mean from this population.

This is called the **sampling distribution of the mean**. It depends on the mean, the standard deviation, and the sample sizes we observe.

If we took a sample of individuals' heights from a distribution with

$$\mu = 66, \sigma = 3.9$$

We could expect to obtain the following distributions (assuming we sampled 10, 25, and 100 individuals to arrive at those means):



Our samples more accurately reflect the true mean when we have larger sample size!

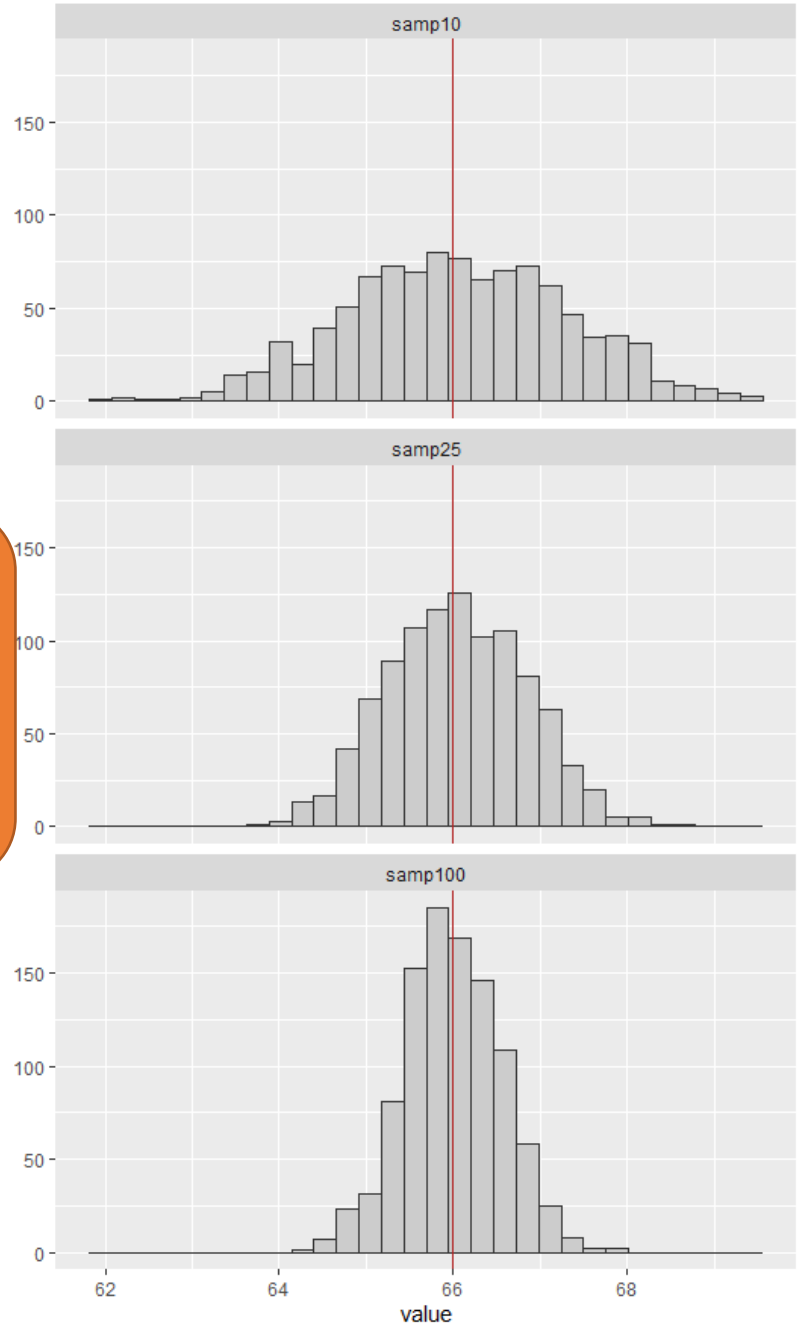
These sampling distributions have their own mean and standard deviation

We expect our sample mean to accurately reflect our population mean.

$$\mu_{\bar{x}} = \mu,$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

And the "standard error" is how accurately we estimate our sample mean. As *n* goes up, the standard error goes down and our estimates are more precise.

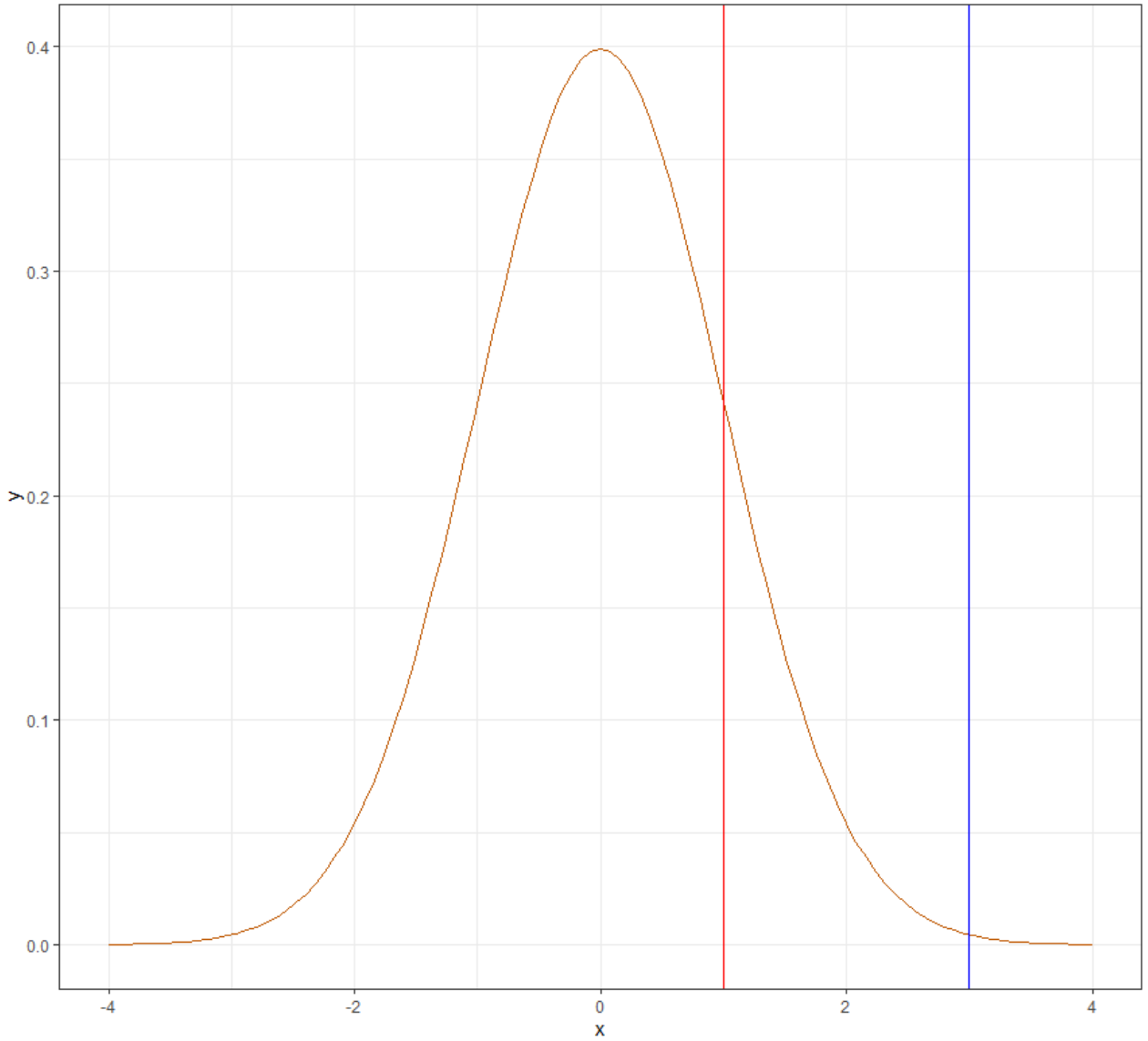


3) Under these conditions, figure out how unlikely it would be to observe the mean we observed under the null hypothesis.

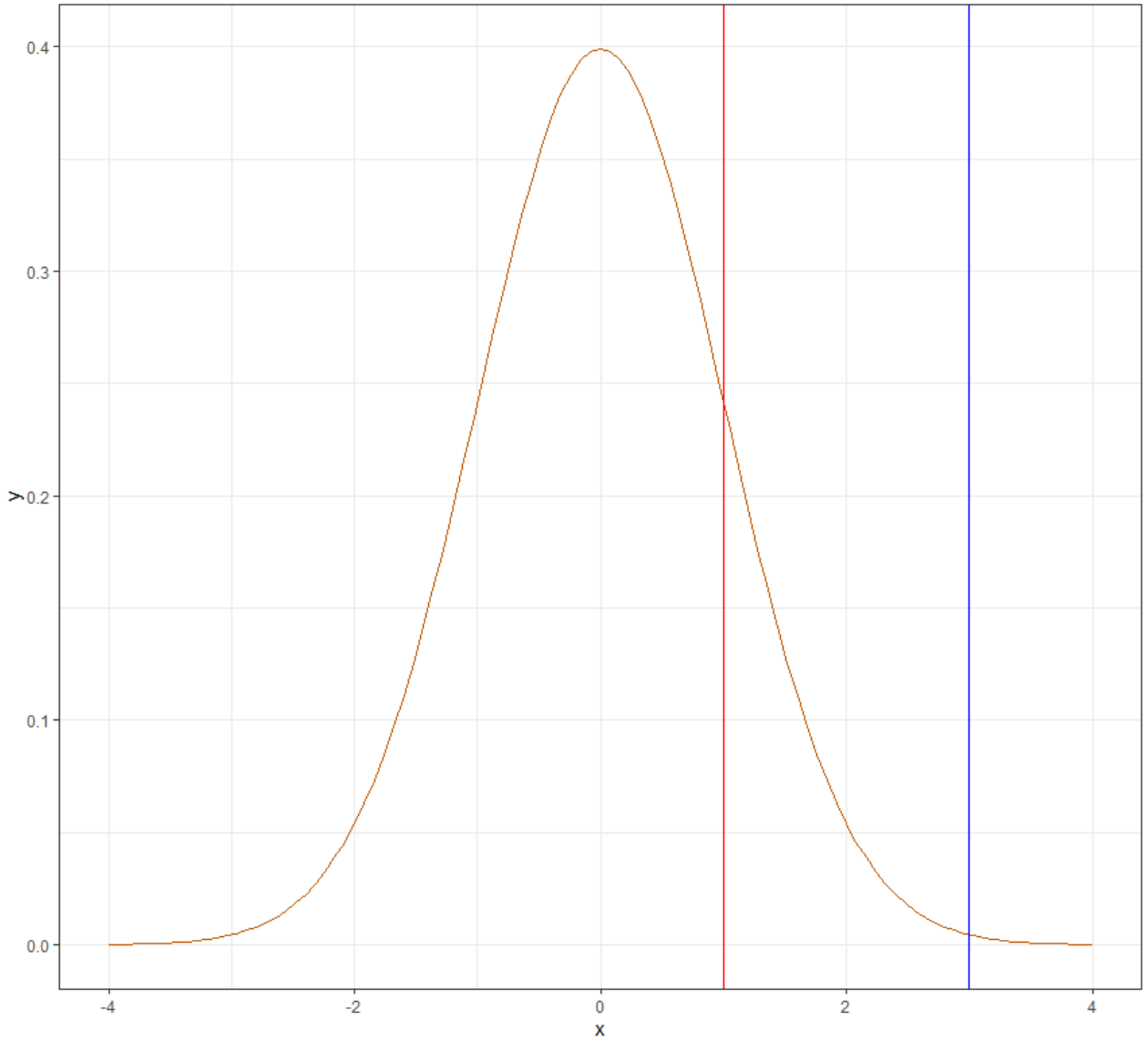
We know how we would expect our sample means to be distributed under the null hypothesis.

How weird would it be to observe our sample mean (or a sample mean more extreme) under these conditions?

If this was our sampling distribution under H_0, would it be stranger to observe an observed mean at the red line, or the blue line?



The **blue line** represents a mean that would be very unlikely to be observed under the null hypothesis.



The red and blue lines represent possible observed Z values.

$$Z_{obs} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

Assume we know the population standard error is 0.016.

In our example,

$$Z_{obs} = \frac{68.3 - 66}{0.016} = 144$$

This is a very high Z value!

It would be incredibly unlikely to observe this sample mean (68.3) if the true mean was 66.

We conclude our sample comes from a population with a mean different from 66.

What is the probability that H_0 would be true and we would observe a sample mean more extreme than what we observed?

This is the 2-sided p-value corresponding to a Z-score of 144.

```
> 2*(1-pnorm(abs(144)))

[1] 0
```

p<.001. We reject the null hypothesis. The average height of OkCupid users in our data set is different from 66".

Example

A policeman observed the speed of 16 cars on a stretch of road. Is there evidence that these cars are, on average, travelling at 55 MPH? Assume we know the population standard deviation is 10 MPH.

```
> speeds
 [1] 57 69 66 53 59 69 50 55 65 48 55 54 68 58 58 68
> mean(speeds)
[1] 59.5
```

Example

A policeman observed the speed of 16 cars on a stretch of road. Is there evidence that these cars are, on average, travelling at 55 MPH? Assume we know the population standard deviation is 10 MPH.

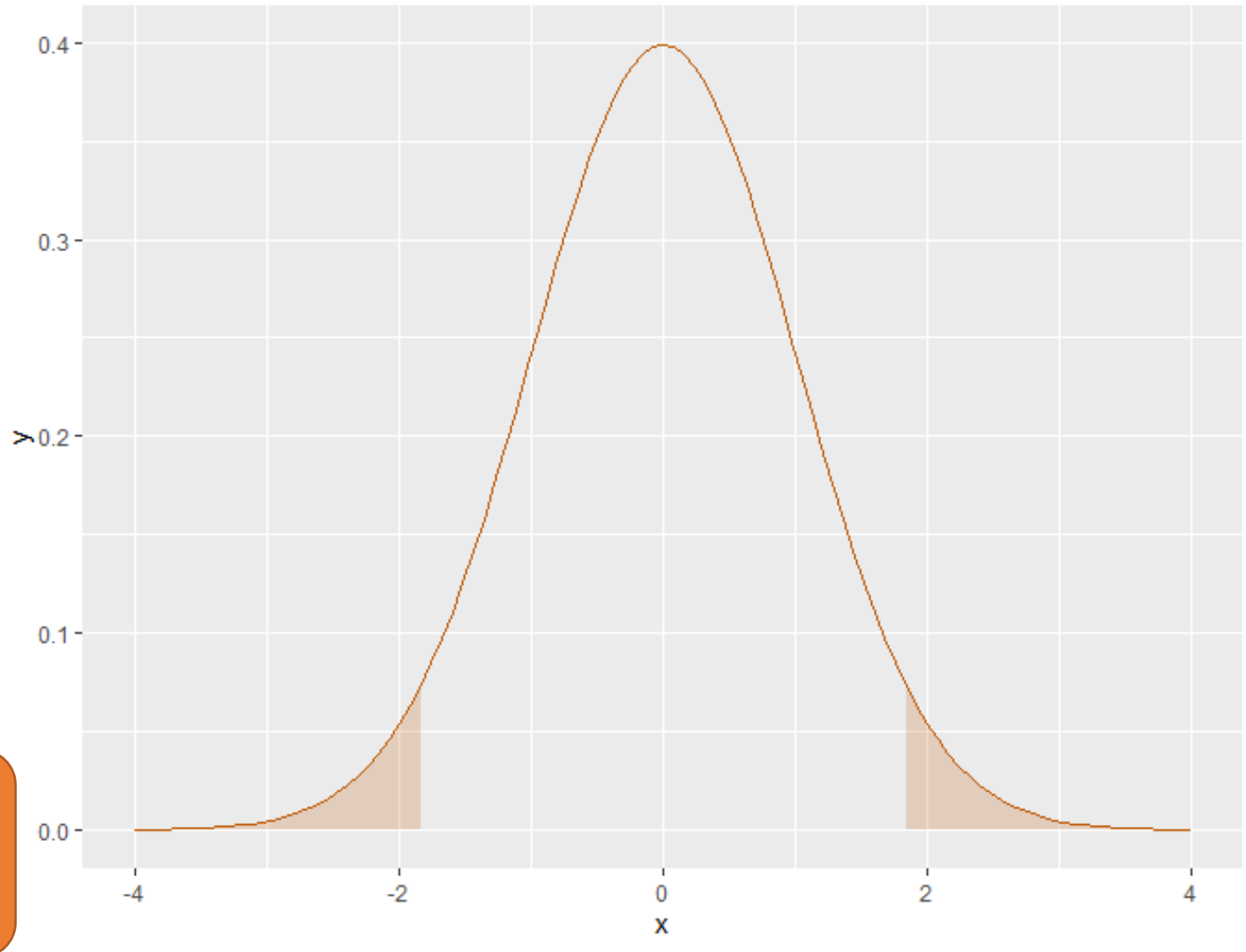
```
> speeds
 [1] 57 69 66 53 59 69 50 55 65 48 55 54 68 58 58 68
> mean(speeds)
[1] 59.5
```

$$Z_{obs} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{59.5 - 55}{\frac{10}{\sqrt{16}}} = \frac{4.5}{2.5} = 1.8$$

2. The Normal Distribution

$$Z_{obs} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{59.5 - 55}{\frac{10}{\sqrt{16}}} = \frac{4.5}{2.5} = 1.8$$

```
> 2*(1-pnorm(abs(1.8)))

[1] 0.07186064
```



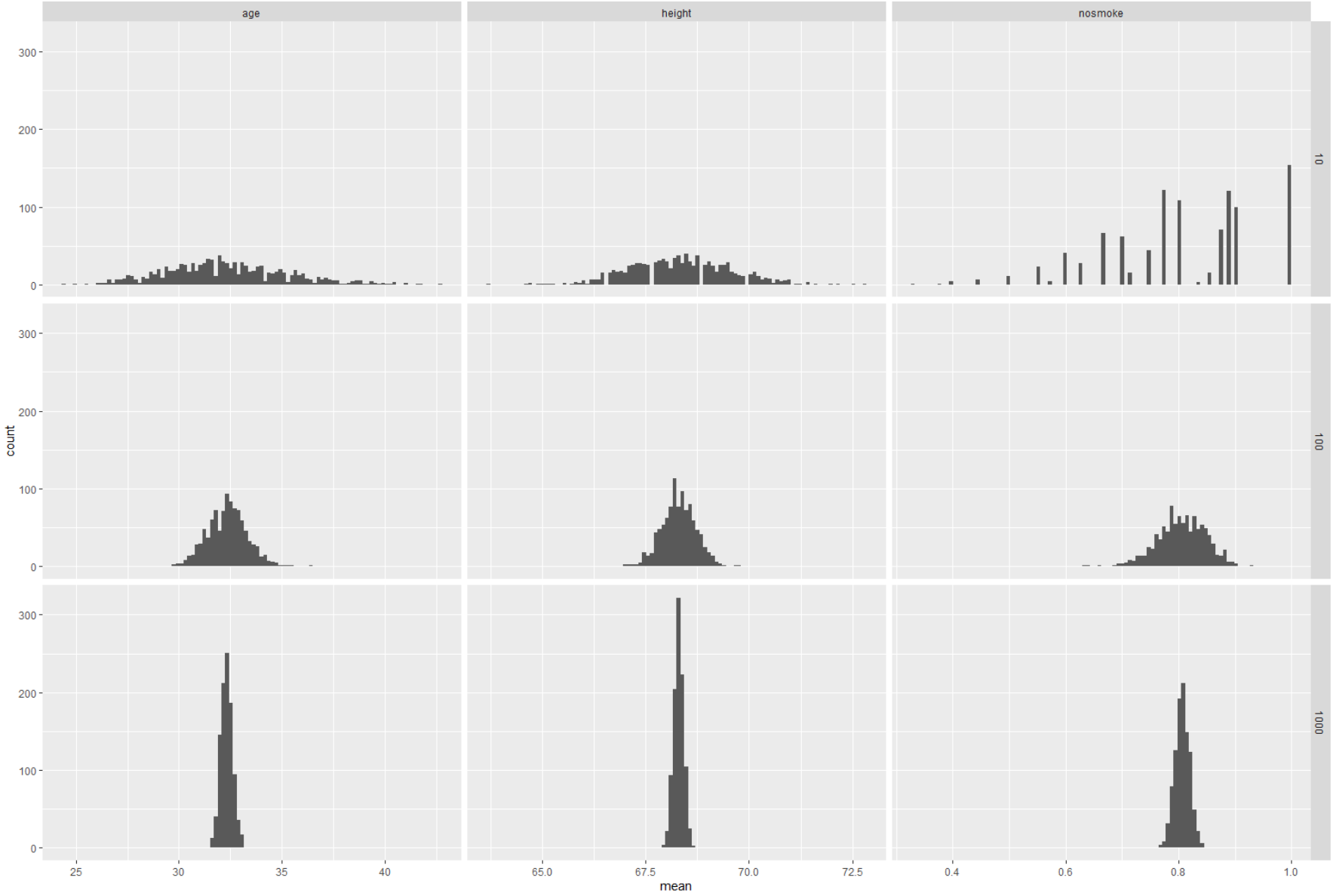
At alpha=.05 we fail to reject H_0; these cars do appear to come from a population with mean 55MPH.

The Central Limit Theorem

The sampling distribution of the mean of X ($\bar{X}$) will approach a normal distribution regardless of the actual distribution of X if the sample mean is based on a large sample size.

Regardless of the actual distribution of X, $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{N}}}$ has a standard normal distribution.

2. The Normal Distribution

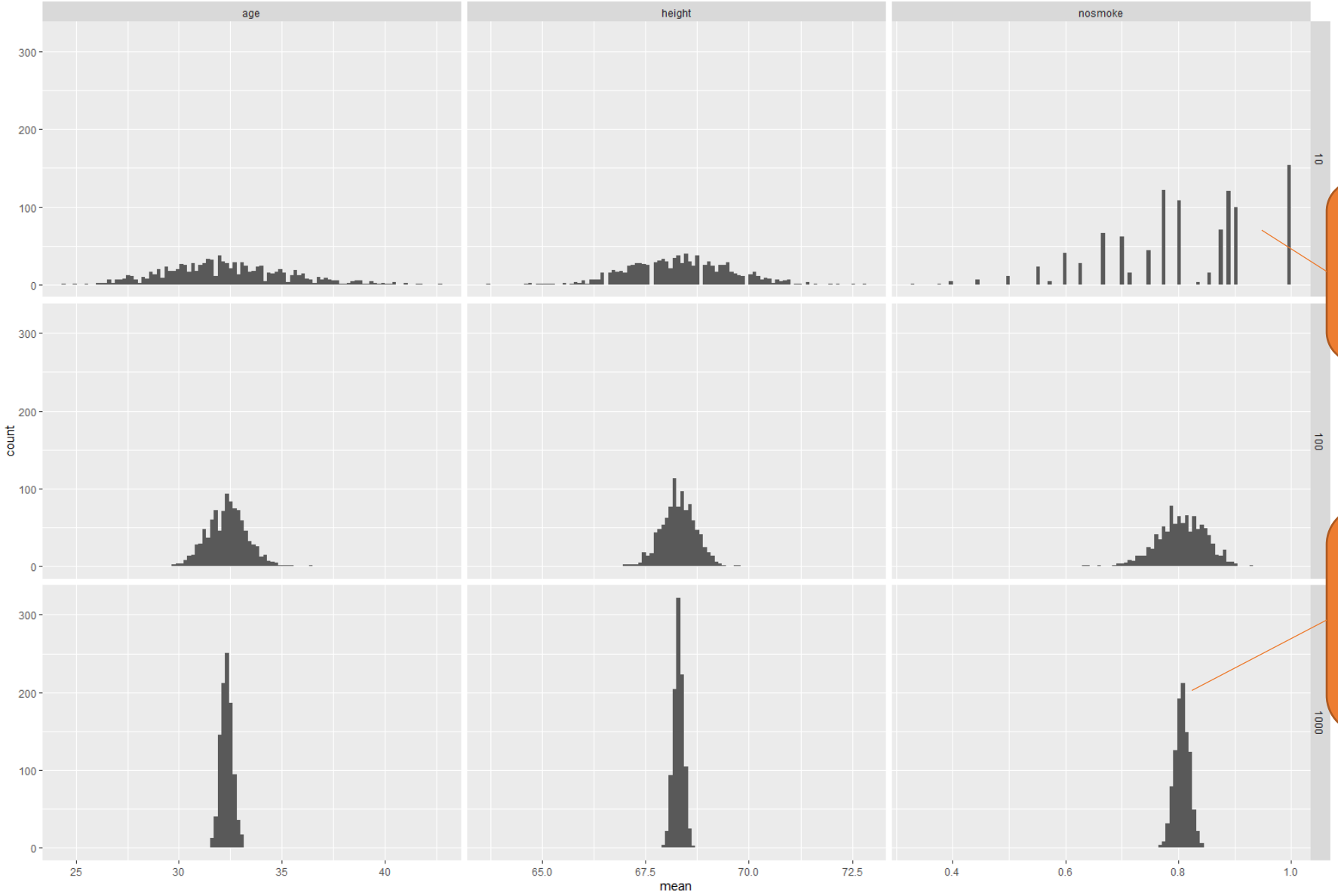


We can see this in action.

Age has a skewed distribution, height is relatively normal, and “nosmoke” is a binary variable.

However, if we take large enough samples of these variables from the population, the sampling distribution of the mean approaches a normal distribution.

2. The Normal Distribution



The distribution of all possible means we could obtain by sampling N=10 people from the population is still non-normal.

But the distribution of all possible sample means we could obtain by sampling N=1000 people from the population looks much more normal.

See OpenIntro 5.1 for more on the Central Limit Theorem

What does this mean?

Because of the central limit theorem, when our sample size is “large” we can still use our typical statistical methods like the Z and t-test even when the distribution of *X* is not normal.

How large is “large”?

- For small deviations from normality, N > 25 or so.
- For large deviations from normality, N in the hundreds.

Therefore in exploratory data analysis, we often want to see how well a data sample fits the normal distribution. We can accomplish this by:

- Boxplots (from last class)
- Normal probability plots
- Statistics such as skeweness and kurtosis

Skewness and kurtosis

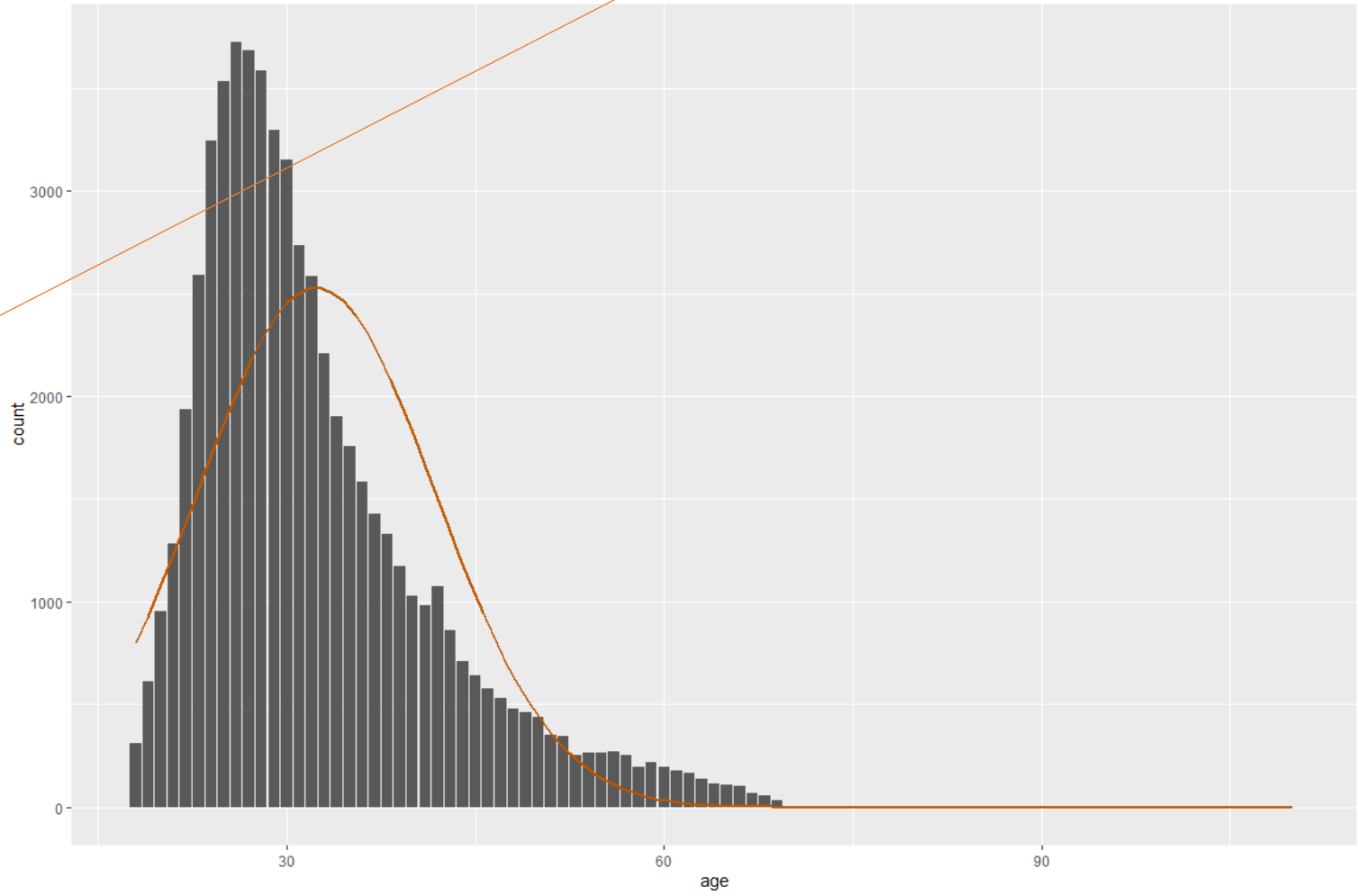
- Skewness is the extent to which a distribution is symmetrical (negative = left skew, positive = right skew).
- Kurtosis is the extent to which a distribution is peaked vs. flat (negative = flatter, positive = peaked)
- A guideline is that values greater than 1 (skewness) or 3 (kurtosis) in magnitude reflect considerable departure from normality.

```
> okc %>% select(height, age) %>% describe()
       vars     n  mean   sd median trimmed  mad min max range  skew kurtosis   se
height    1 59937 68.30 3.94     68   68.33 4.45  26  95    69 -0.07     1.68 0.02
age       2 59946 32.34 9.45     30   31.09 7.41  18 110    92  1.27     1.57 0.04
```

2. The Normal Distribution

```
> okc %>% select(height, age) %>% describe()
       vars     n  mean   sd median trimmed  mad min max range  skew kurtosis   se
height    1 59937 68.30 3.94     68   68.33 4.45  26  95    69 -0.07     1.68 0.02
age       2 59946 32.34 9.45     30   31.09 7.41  18 110    92  1.27     1.57 0.04
```

The 1.27 skew for age indicates that age has some considerable positive-skew.



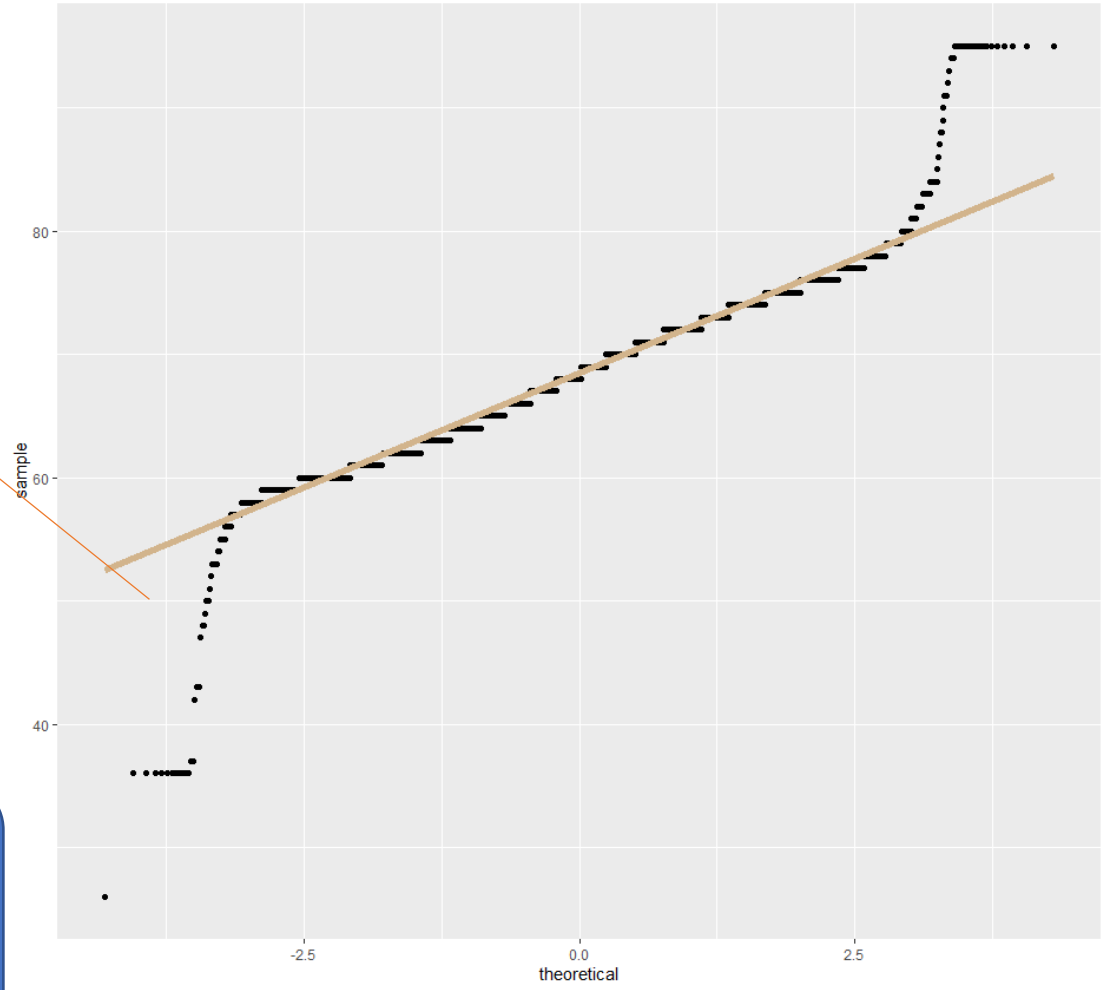
QQ Plot

- The X values are ordered from smallest to largest
- The ith ordered X observation is plotted against the value of a standard normal random variable corresponding to the $(i/n)^{th}$ percentile
- If the variable is normally distributed, the QQ plot will follow a straight line

The QQ plot for height

The low values of height are lower than we'd expect under a truly normal distribution. Likewise, the high values are higher than we'd expect.

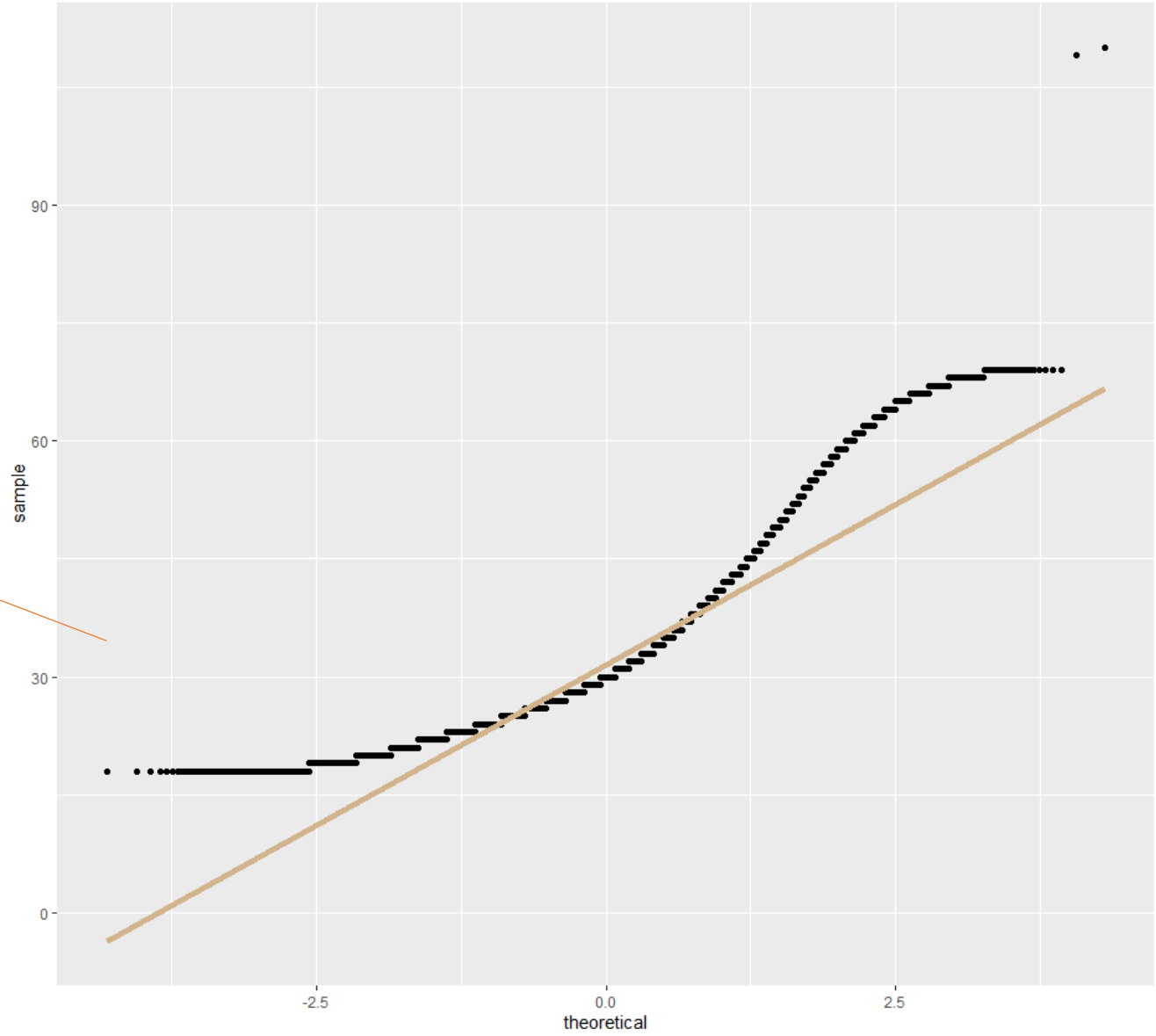
This is not surprising considering there was some evidence of kurtosis.



As far as QQ plots go, this does show some departure from normality but in practice the tails usually do show some slight departure. It would only mildly violate the assumption of normality.

The QQ plot for age

Age shows a more concerning departure from normality. This is probably a moderate departure.



Recap

- The normal distribution can be used:
  - To approximate a distribution of a continuous variable X.
  - To compute z-scores.
  - To find probabilities that a score X falls in a certain range.
  - To approximate the sampling distribution of the mean.
- Many parametric statistical tests rely on the sampling distribution of the mean being normal
- When sample size is large enough (central limit theorem) the sampling distribution of the mean will be normal

Recap

- Determine whether a distribution is normal or violates normality
- Transform scores from a distribution into a z-score
- Given a normal distribution, find the probability that a score X falls in a certain range
- Given a normal distribution, find the value that corresponds with a certain area under the curve
- Describe why the Central Limit Theorem is important for statistical testing

Test Yourself

True or false? On average, users that are “seeing someone” are younger than users in all other relationship categories.

```
_______________________
Group variables            status

-- Variable type: numeric ------------------------------------------------------------------------------------
# A tibble: 6 x 12
  skim_variable status         n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>         <chr>                    <chr>


1 age           available              0             1  33.9  9.18    18    27  32    39     109 ▇▂▂▁▁
2 age           married                0             1  36.5  9.59    20    29  35    42      66 ▅▇▅▃▂
3 age           seeing someone         0             1  29.5  6.90    18    25  28    32      68 ▇▇▅▁▁
4 age           single                 0             1  32.4  9.52    18    26  30    37     110 ▇▂▁▁▁
5 age           unknown                0             1  35.5  9.99    22    28  36.5  38      57 ▅▂▇▅▂
6 age           <NA>                   0             1  42.3 20.3     26    31  36    50.5    65 ▇▇▁▁▇
```

Test Yourself

True or false? On average, users that are “seeing someone” are younger than users in all other relationship categories.

Users who are seeing someone have a mean age of 29.5 years, which is lower than the mean for all other groups.

```
──────────────────────
Group variables            status

-- Variable type: numeric ------------------------------------------------------------------------------
# A tibble: 6 x 12
  skim_variable status         n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>         <chr>                    <chr>


1 age           available              0             1  33.9  9.18    18    27  32    39     109 ▅▂▂▁
2 age           married                0             1  36.5  9.59    20    29  35    42      66 ▃▅▅▃▂▁
3 age           seeing someone         0             1  29.5  6.90    18    25  28    32      68 ▅▅▂▁
4 age           single                 0             1  32.4  9.52    18    26  30    37     110 ▅▂▁▁
5 age           unknown                0             1  35.5  9.99    22    28  36.5  38      57 ▅▂▅▂▂
6 age           <NA>                   0             1  42.3 20.3     26    31  36    50.5    65 ▅▅▂▅
```

Test Yourself

True or false? The least amount of variation occurred for individuals who were “seeing someone.”

```
_______________________
Group variables            status

-- Variable type: numeric ------------------------------------------------------------------------------------
# A tibble: 6 x 12
  skim_variable status         n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>         <chr>                    <chr>


1 age           available              0             1  33.9  9.18    18    27  32    39     109 ▇▂▂▁▁
2 age           married                0             1  36.5  9.59    20    29  35    42      66 ▅▇▅▃▂
3 age           seeing someone         0             1  29.5  6.90    18    25  28    32      68 ▇▇▅▁▁
4 age           single                 0             1  32.4  9.52    18    26  30    37     110 ▇▂▁▁▁
5 age           unknown                0             1  35.5  9.99    22    28  36.5  38      57 ▅▂▇▅▂
6 age           <NA>                   0             1  42.3 20.3     26    31  36    50.5    65 ▇▇▁▁▇
```

Test Yourself

True or false? The least amount of variation occurred for individuals who were "seeing someone."

The distribution of age for users who were "seeing someone" is 6.90 years, which is the smallest of all the status categories.

```
_____________________
Group variables            status

-- Variable type: numeric ------------------------------------------------------------------------------
# A tibble: 6 x 12
  skim_variable status         n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>         <chr>                     <chr>


1 age           available              0             1  33.9  9.18    18    27  32    39     109 ▇▅▂▁▁
2 age           married                0             1  36.5  9.59    20    29  35    42      66 ▅▇▅▃▂
3 age           seeing someone         0             1  29.5  6.90    18    25  28    32      68 ▇▇▂▁▁
4 age           single                 0             1  32.4  9.52    18    26  30    37     110 ▇▅▂▁▁
5 age           unknown                0             1  35.5  9.99    22    28  36.5  38      57 ▅▃▇▅▂
6 age           <NA>                   0             1  42.3 20.3     26    31  36    50.5    65 ▇▇▂▁▇
```

Test Yourself

IQ scores are distributed with mean 100 and SD 15. Suppose someone has an IQ of 89. How would you use pnorm/qnorm to compute that person's IQ percentile?

Test Yourself

IQ scores are distributed with mean 100 and SD 15. Suppose someone has an IQ of 89. How would you use pnorm/qnorm to compute that person's IQ percentile?

Option 1: pnorm(89, mean=1000, sd=1)

Option 2: pnorm(-0.733)

Note: (89-100)/15 = -0.733

Test Yourself

True or false? We can infer in the population that those who used drugs are, on average, 0.28 years younger than those who never used drugs.

```
> profiles <- profiles %>%
+   mutate(any_drugs = if_else(drugs == "never", 0, 1) %>%
+            factor(labels = c("Never Used", "Any Use")))

> t.test(age ~ any_drugs, data = profiles)

 Welch Two Sample t-test
data:  age by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 3.318760 3.731443
sample estimates:
mean in group Never Used    mean in group Any Use
   33.19521                 29.67011

> t.test(scale(age) ~ any_drugs, data = profiles)
 Welch Two Sample t-test
data:  scale(age) by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 0.3510883 0.3947456
sample estimates:
mean in group Never Used    mean in group Any Use
            0.09044088              -0.28247608
```

Test Yourself

True or false? We can infer in the population that those who used drugs are, on average, 0.28 years younger than those who never used drugs.

P<0.05 so we can infer that there is a statistically significant difference in age between drug use groups. The mean age for those who never used is 33.20, vs. 29.67 for those who had any use. This difference is 3.53 years.

```
> profiles <- profiles %>%
+   mutate(any_drugs = if_else(drugs == "never", 0, 1) %>%
+            factor(labels = c("Never Used", "Any Use")))

> t.test(age ~ any_drugs, data = profiles)

 Welch Two Sample t-test
data:  age by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 3.318760 3.731443
sample estimates:
mean in group Never Used    mean in group Any Use
   33.19521                 29.67011

> t.test(scale(age) ~ any_drugs, data = profiles)
 Welch Two Sample t-test
data:  scale(age) by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 0.3510883 0.3947456
sample estimates:
mean in group Never Used    mean in group Any Use
            0.09044088              -0.28247608
```

Test Yourself

True or false? We are 95% confident that, on average, those who never used drugs are between 3.32 and 3.73 years older than those who have used drugs.

```
> profiles <- profiles %>%
+   mutate(any_drugs = if_else(drugs == "never", 0, 1) %>%
+            factor(labels = c("Never Used", "Any Use")))

> t.test(age ~ any_drugs, data = profiles)

 Welch Two Sample t-test
data:  age by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 3.318760 3.731443
sample estimates:
mean in group Never Used    mean in group Any Use
   33.19521                 29.67011

> t.test(scale(age) ~ any_drugs, data = profiles)
 Welch Two Sample t-test
data:  scale(age) by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 0.3510883 0.3947456
sample estimates:
mean in group Never Used    mean in group Any Use
            0.09044088              -0.28247608
```

Test Yourself

True or false? We are 95% confident that, on average, those who never used drugs are between 3.32 and 3.73 years older than those who have used drugs.

The first t-test shows that the 95% CI on the difference between means is (3.32, 3.73).

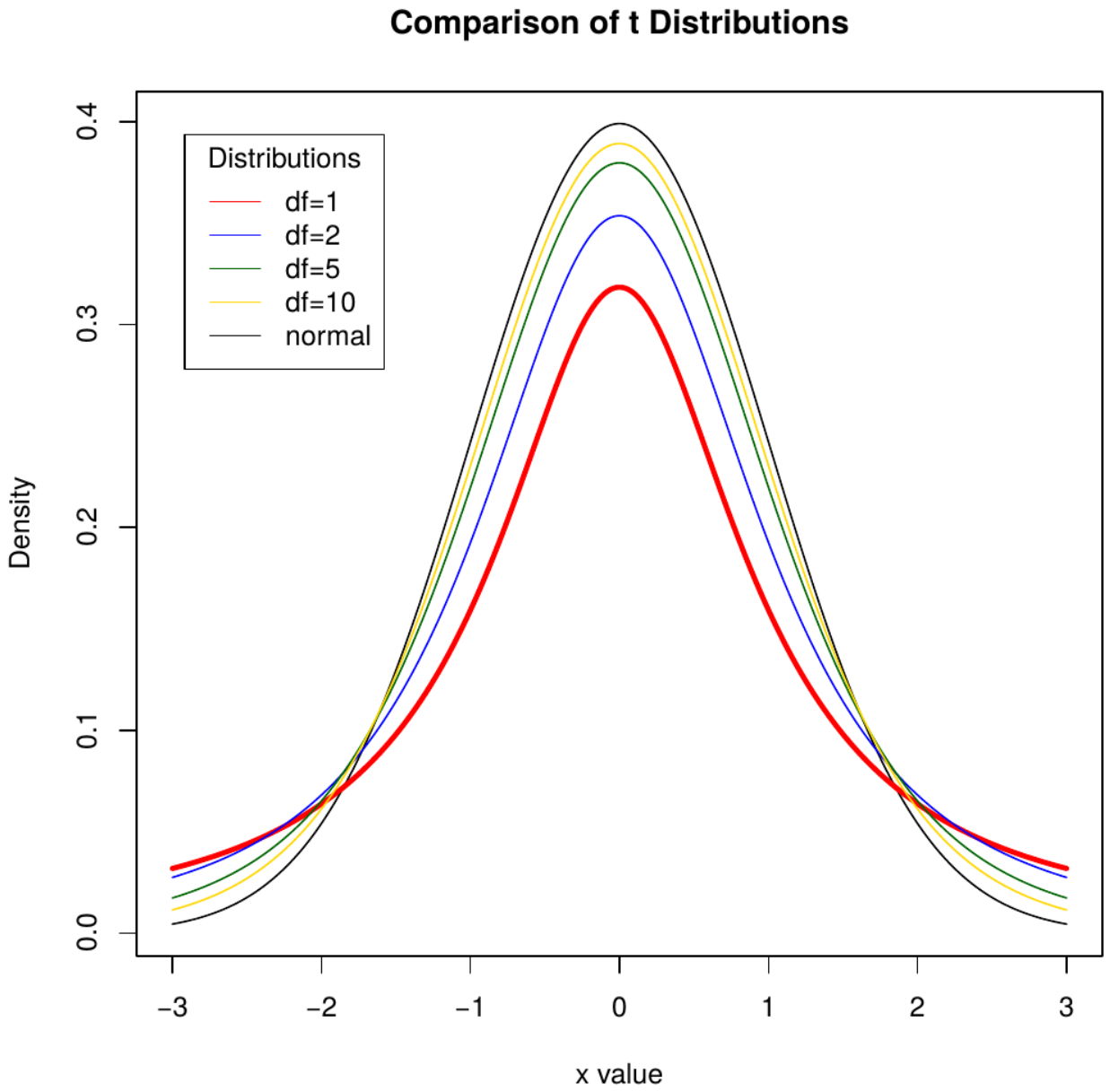
```
> profiles <- profiles %>%
+   mutate(any_drugs = if_else(drugs == "never", 0, 1) %>%
+            factor(labels = c("Never Used", "Any Use")))

> t.test(age ~ any_drugs, data = profiles)

 Welch Two Sample t-test
data:  age by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 3.318760 3.731443
sample estimates:
mean in group Never Used    mean in group Any Use
   33.19521                 29.67011

> t.test(scale(age) ~ any_drugs, data = profiles)
 Welch Two Sample t-test
data:  scale(age) by any_drugs
t = 33.487, df = 13780, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 0.3510883 0.3947456
sample estimates:
mean in group Never Used    mean in group Any Use
            0.09044088              -0.28247608
```

There are other distributions that are also important as they are the distributions for test statistics when we conduct inferential analysis.

Distribution	Distribution of…
T-distribution	• Sampling distribution of the mean when the population standard deviation is unknown. • Regression parameters.
Chi-square	• Sum of squared normal distributions. • Test statistic for contingency tables. • Test statistic for maximum likelihood estimates.
F-distribution	• The ratio of chi-square variables.
Bernoulli	• The proportion of successful trials.

The t-distribution

- Symmetric, unimodal, and similar to the standard normal distribution
- The shape depends on the degrees of freedom
- When we estimate the population variance (i.e., σ is unknown), we use the t-statistic for testing hypotheses regarding a mean difference



The t-distribution is typically used:

- In a test of whether a sample mean is equal to a hypothesized value μ_0.

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{N}}$$

- In a test of whether the mean of two groups is equal ($\overline{X_D}$ is the mean difference in groups, s_D is the standard deviation of the mean difference).

$$T = \frac{\overline{X_D} - \mu_0}{s_D/\sqrt{N}}$$

- These tests have N-1 df.

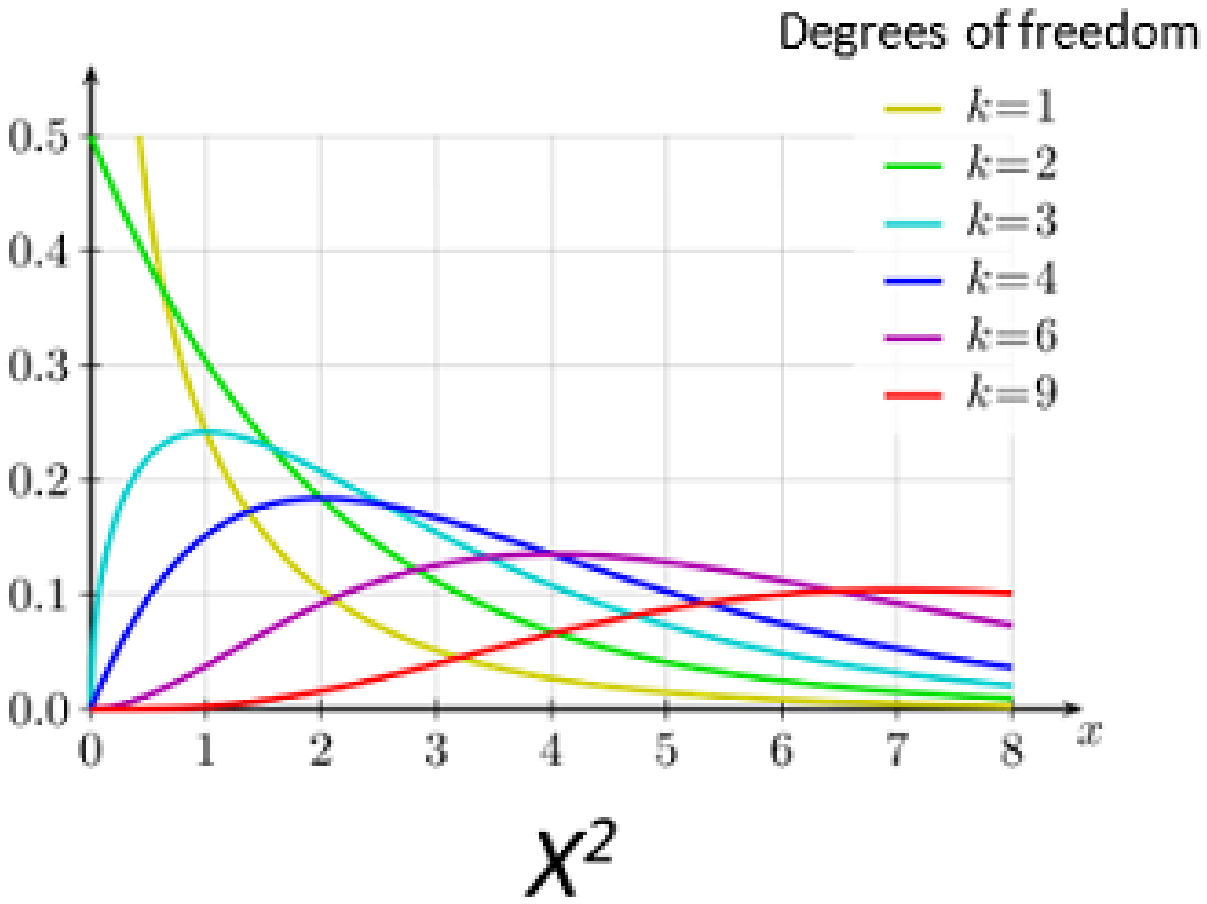
Practice Question 3

a) Use the `pt` function in R to find the 2-sided p-value corresponding to t_{10}=2.25 (df = 10).

b) Use the `qt` function in R to find the t value corresponding to a 2-sided p-value of 0.05 (on 8 degrees of freedom).

The chi-square (χ^2) distribution

- Non-symmetric, ranges from 0 to ∞, shape depends on degrees of freedom
- A χ^2 distribution with 1 df is equivalent to Z^2
- A χ^2 distribution with *k* df is equivalent to the sum of squares of *k* independent standard normal $\sum_{i=1}^{k} Z_i^2$
- Typically used to describe the distribution of a variance



Practice Question 4

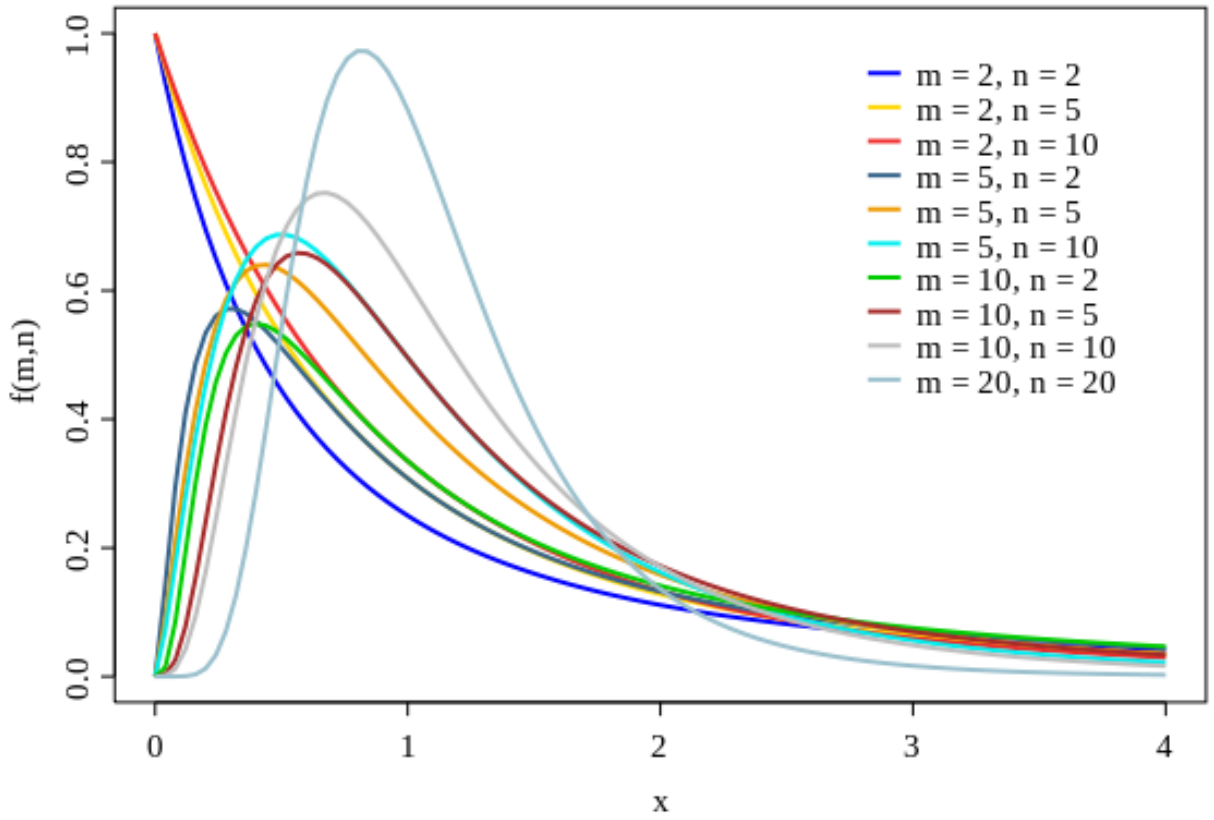
a) Use the `pchisq` function in R to find the p-value corresponding to $\chi^2_{10}=5.25$.

b) Use the `qchisq` function in R to find the χ^2 value corresponding to a p-value of 0.05 (on 8 degrees of freedom).

Chi-square tests are non-directional; we always use the area on the right side of the tail to compute p-values.

The F distribution

- Non-symmetric, ranges from 0 to ∞, shape depends on numerator and denominator degrees of freedom
- Typically used for ANOVA tests.



Practice Question 5

On your own, figure out the functions to use to answer the following questions:

a) Find the p-value corresponding to $F_{10,1}$=2.25.

b) Find the F value corresponding to a p-value of 0.05 (on 10 numerator and 1 denominator df).

F tests are non-directional; we always use the area on the right side of the tail to compute p-values.

Bernoulli Distribution

- Used to describe the probability of a dichotomous outcome.

$$P(X = 1) = p = 1 - q$$

$$Var(X) = pq$$

❑ For what value of p is the variance the highest?

Binomial Distribution

- Used to describe the probability of a certain number of successes (k) in a certain number of trials (n).

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

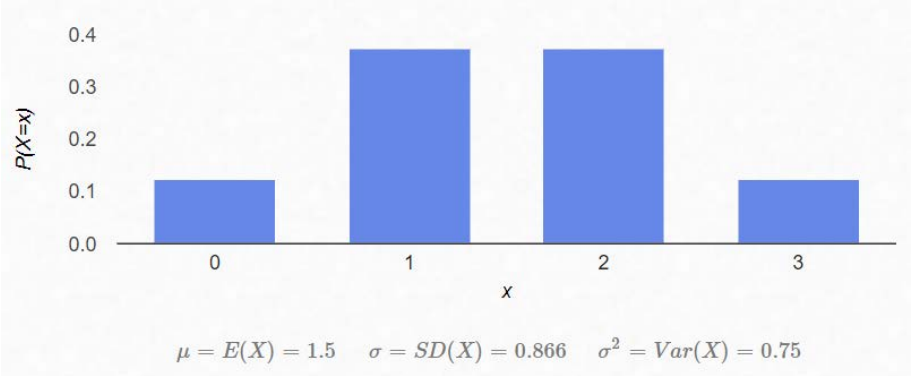
Example: The probability of flipping a coin three times and seeing a total of two heads is:

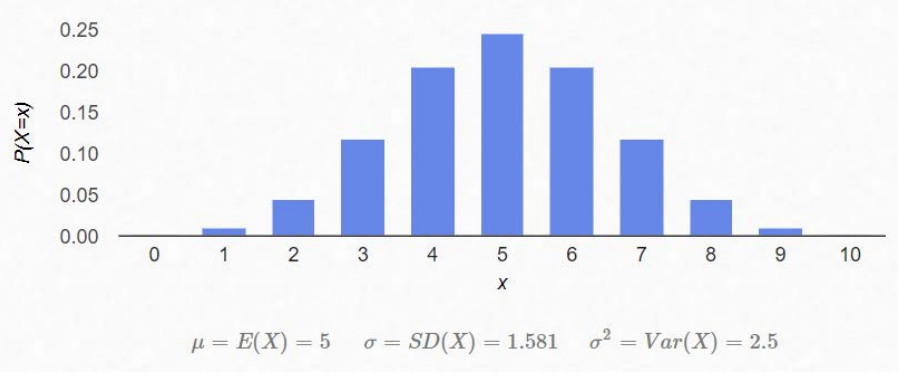
$$P(Y = 2) = \binom{3}{2} .5^2\,(1-.5)^{3-2} = \frac{3*2*1}{(2*1)(1)}(.25)(.5) = .375$$

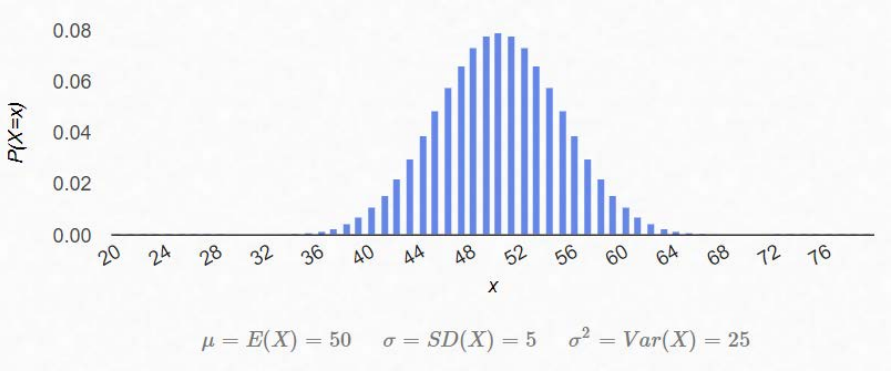
Some notes on the **binomial distribution**:

- The Bernoulli distribution is a special case of the binomial distribution when n=1.
- As the number of trials increases, the binomial distribution begins to look increasingly more like a normal distribution.

The figures on the right show a binomial distribution for the number of heads when flipping a coin 3, 10, and 100 times.







Recap

- Given any distribution from lecture, be able to calculate area under the curve given a raw score
- Explain the relationship between the Bernoulli and binomial distributions
- Calculate probabilities and variances with a Bernoulli distribution
- Calculate probabilities with a binomial distribution

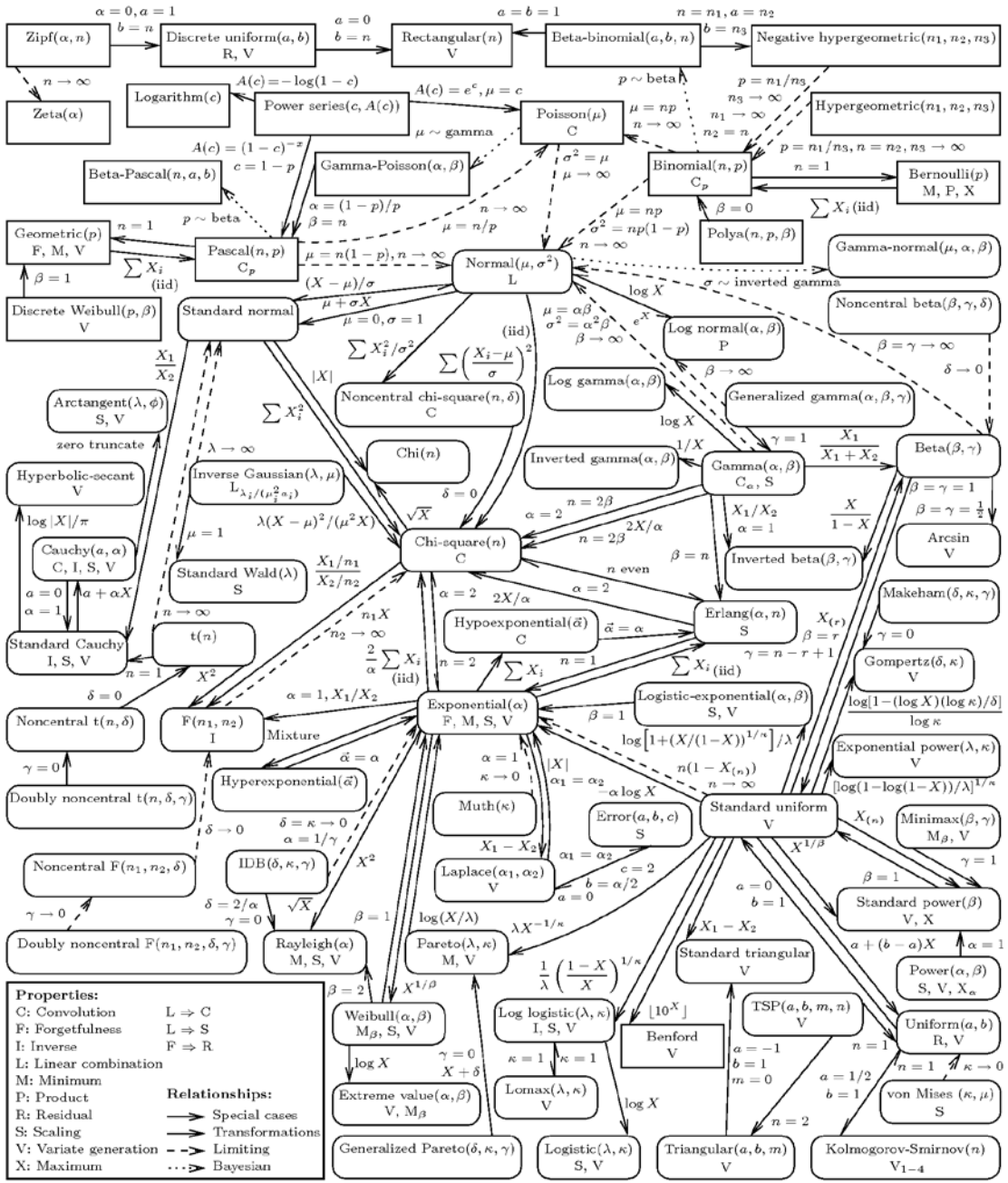
The number of distributions available to us may be daunting, but we typically only focus on a couple.

Distributional assumptions are at the heart of most statistical tests.

Perhaps the definitive reference for the relationships among distributions.

http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

Leemis, L. M., & McQueston, J. T. (2008). Univariate distribution relationships. The American Statistician, 62(1), 45-53.



Additional Reading

- P-values
  https://www.youtube.com/watch?v=JQc3yx0-Q9E
- The Central Limit Theorem
  https://www.youtube.com/watch?v=YAlJCEDH2uY
- Using pnorm and qnorm in R
  https://diggingdeeperwithstats.wordpress.com/2021/05/21/visual-guide-to-pnorm-dnorm-qnorm-and-rnorm-functions-in-r/

Packages and Functions

- `mean`
- `qnorm`
- `pnorm`
- `qt`
- `pt`
- `qchisq`
- `pchisq`
- `qf`
- `pf`