Module 4: Inference on a normal population

Jake Ferguson

Module goals

1. Introduce the central limit theorem
2. Understand some basic operations on probability used in inference
3. Calculate Wald confidence intervals of estimates
4. Run one- and two-sample t-test

Random variables

A variable whose outcome depends on a random phenomenon is called a **random variable**.

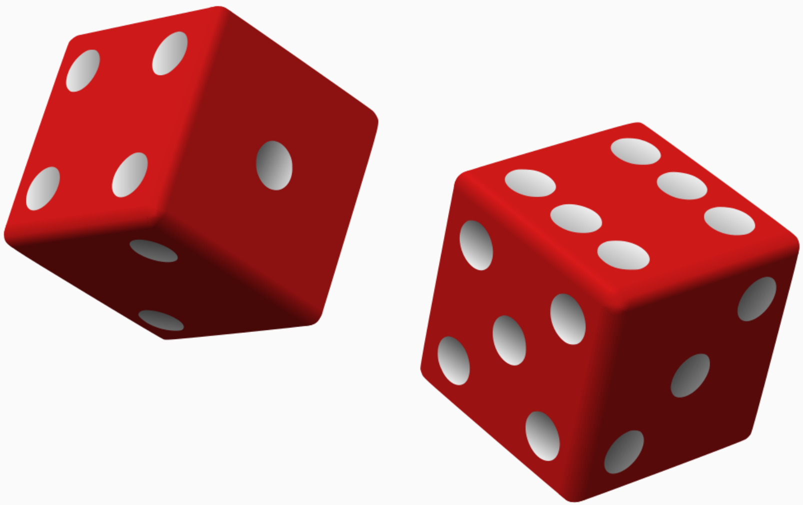
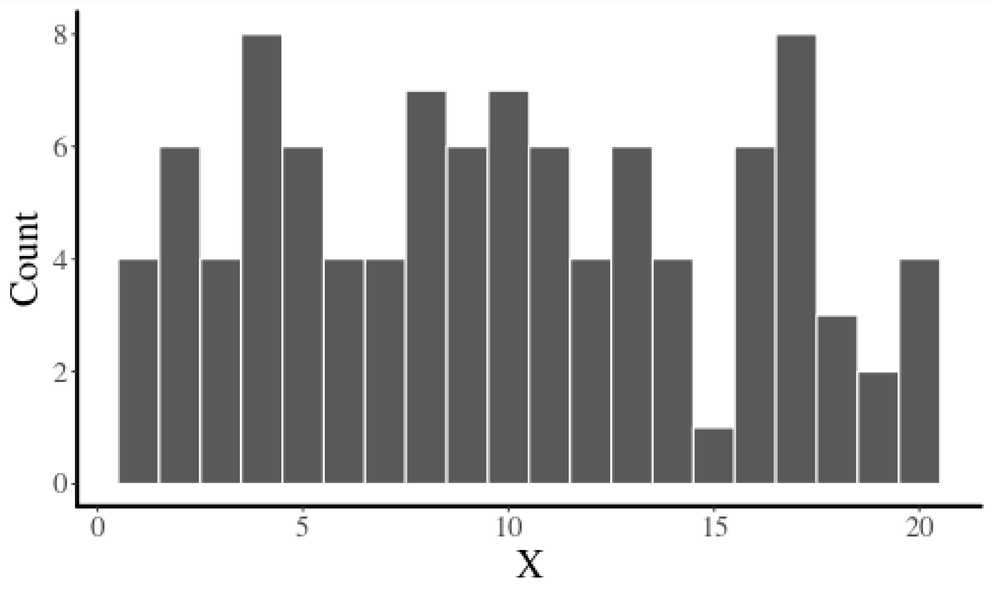


image source: wikimedia.org

A random sample

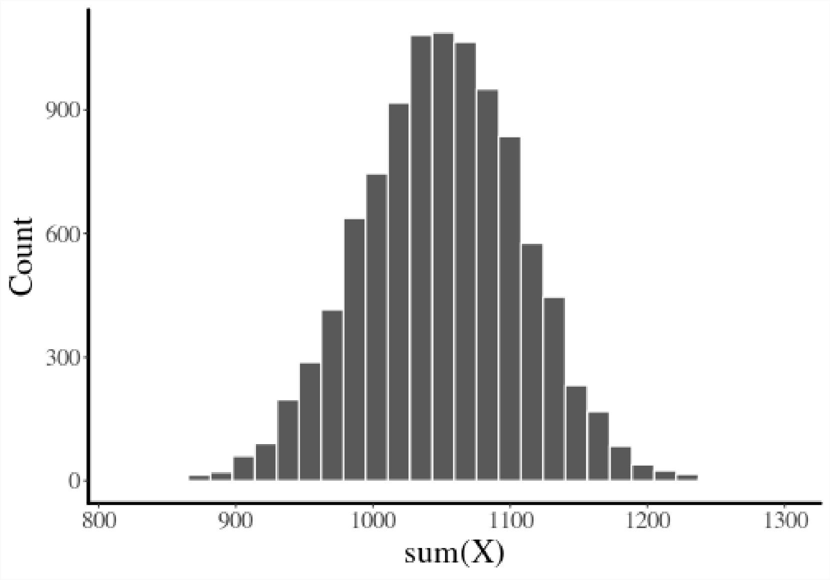
```
x ← sample(1:20, 100, replace=T)
```



Sums of random variables

Many measurements we make and statistics are a sum of random variables.

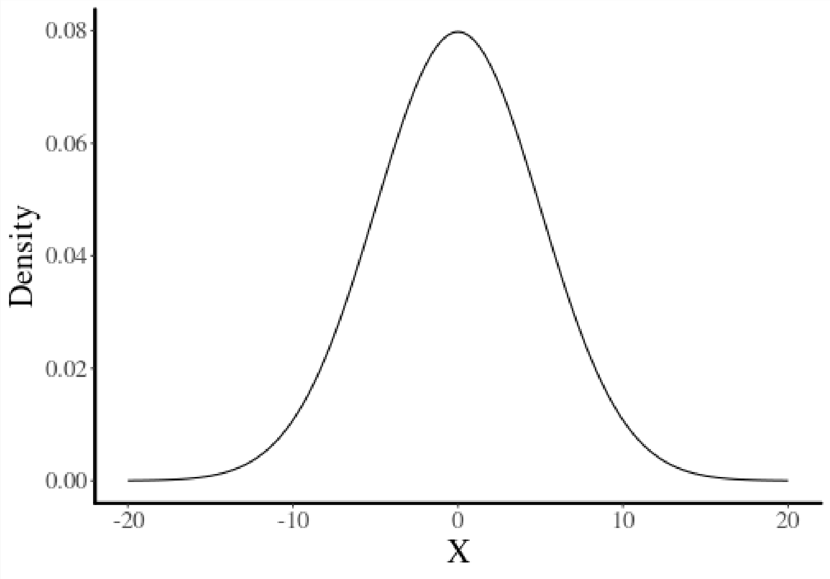
```
x ← numeric(10000)
for(i in 1:10000) {
  x[i] ← sum(sample(1:20, 100, replace=T))
}
```



Collections of such variables display remarkable regularity properties. This is the **Central Limit Theorem** (CLT).

The normal distribution

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



μ is the mean

σ is the standard deviation

The mean and standard deviation

The sample mean $(\bar{x})$ is the estimate of the population mean (μ).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sample variance (s^2) is the estimate of the population variance (σ^2).

$$\bar{x} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

In R use the `mean` and `var` functions:

```
mean(x)
```

```
## [1] 1050.277
```

```
var(x)
```

```
## [1] 3394.525
```

Excercise 4A

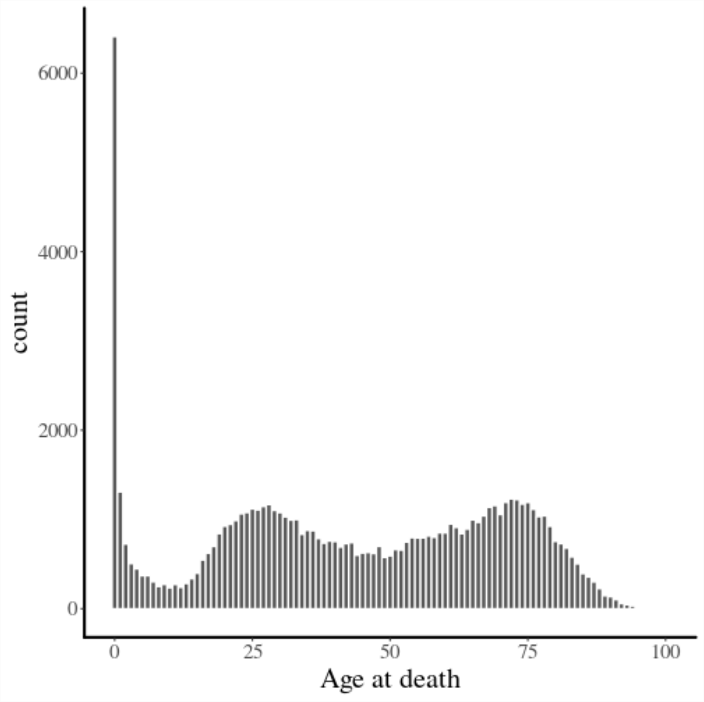
Estimating means and standard deviations

The sampling distribution

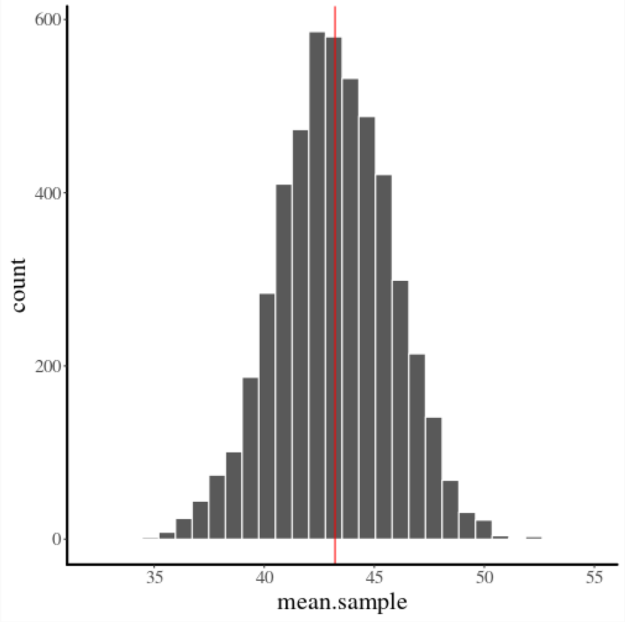
The sampling distribution describes the plausible values of the outcome if we conducted the sampling procedure again. It is a hypothetical construct.

Example: the 1918 flu

These are the deaths due to spanish flu in Switzerland



Repeatedly take samples of size 100 from the population to look at the distribution of the estimates

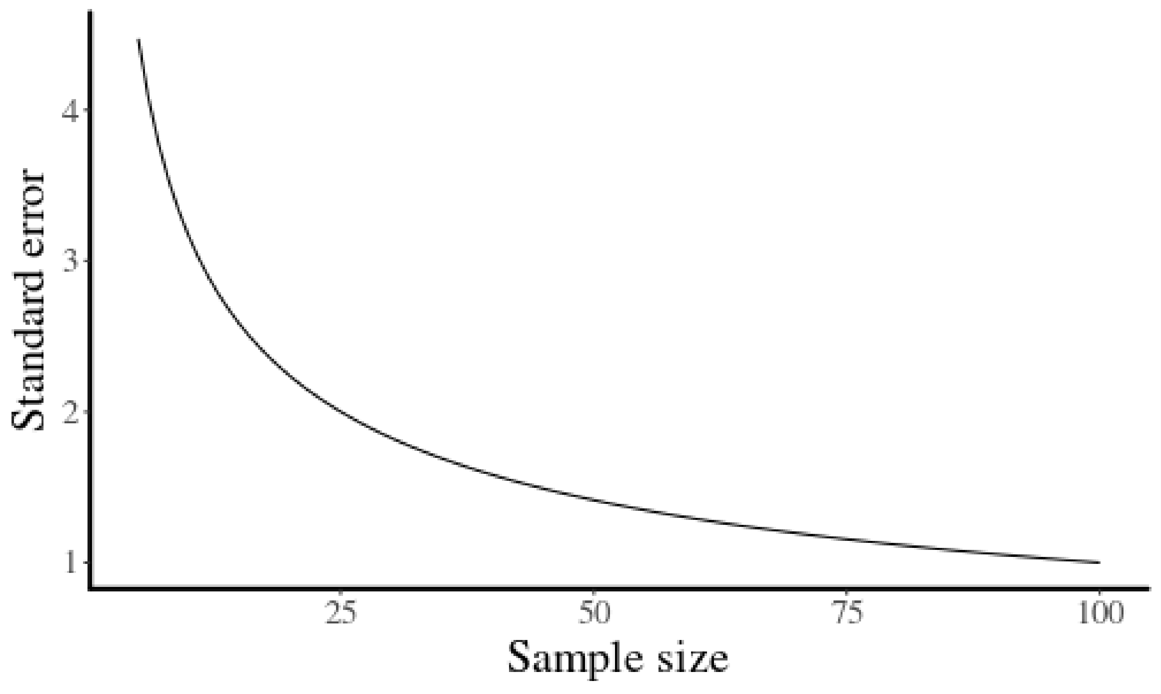


The standard error

The standard deviation of the sampling distribution is so important that it has a special name, the **standard error** $(\sigma_{\bar{x}})$.

$$\sigma_{\bar{x}}^2 = \frac{s^2}{n}$$

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



Excercise 4B

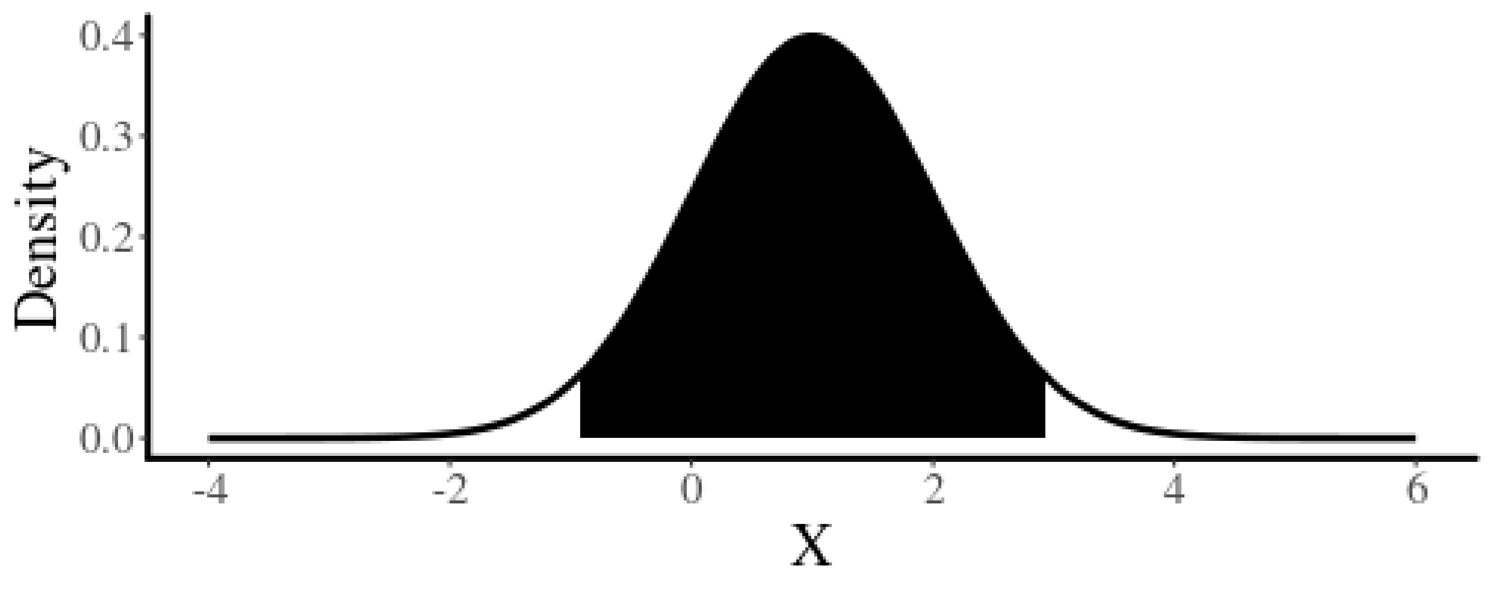
Simulating sampling distributions

Inference with known σ

We and estimate the mean horn length of a unicorn, $\bar{x}$. The standard error determines how much uncertainty is in $\bar{x}$. How confident can we be that the population horn average (i.e., the true value) is close to our estimate?

A confidence interval gives a range of values that will contain the true value some specified proportion of the time.

A 95% confdence interval when the population variance is known, is $\bar{x} \pm 1.96 \cdot \sigma_{\bar{X}}$



Example in R

The Hawai'ian monk seal data. Calculated the confidence interval of the behavior differences:

```
xbar <- mean(diff.dat)
n <- length(diff.dat)
se <- sd(diff.dat)/sqrt(n)

xbar - qnorm(0.975)*se
```

```
## [1] -0.03272467
```

```
xbar + qnorm(0.975)*se
```

```
## [1] 0.007792686
```

image: Hawaii Marine Animal Response

The z-statistic

Given a sample with mean $\bar{x}$ and standard error $\sigma_{\bar{x}}$, the z-statistic is normally distributed

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Example

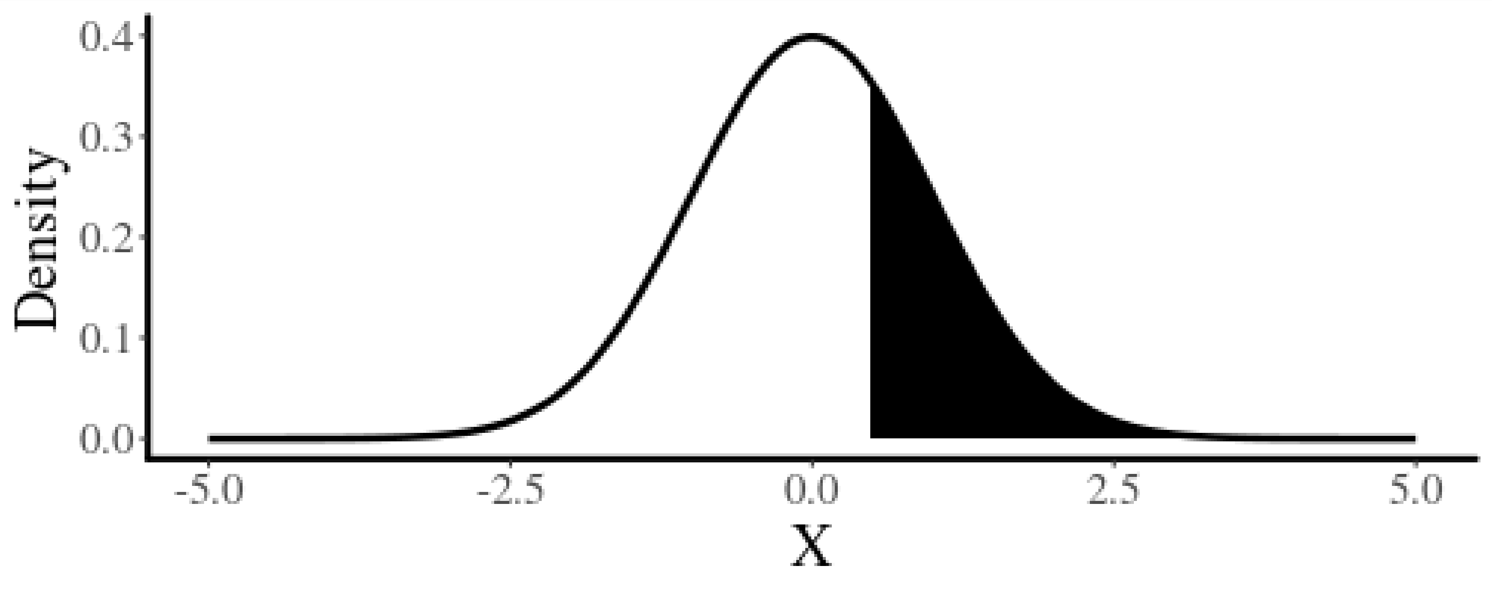
Take a random sample of $n = 80$ babies in the US and get a mean birth weight of 3370 g. This population is well studied and known to have a mean of $\mu = 3339$ and standard devation of $\sigma = 573$

$$\begin{aligned} Z &= \frac{3370 - 3339}{573/\sqrt{80}} \\ &= 0.48 \end{aligned}$$

Getting p-values

Now what is the probability that we could have drawn a sample with this average weight or larger from our population?

$$P[Z > 0.48] = ?$$




```
pnorm(q=0.48, lower.tail=F)
```

```
## [1] 0.3156137
```

Excercise 4C

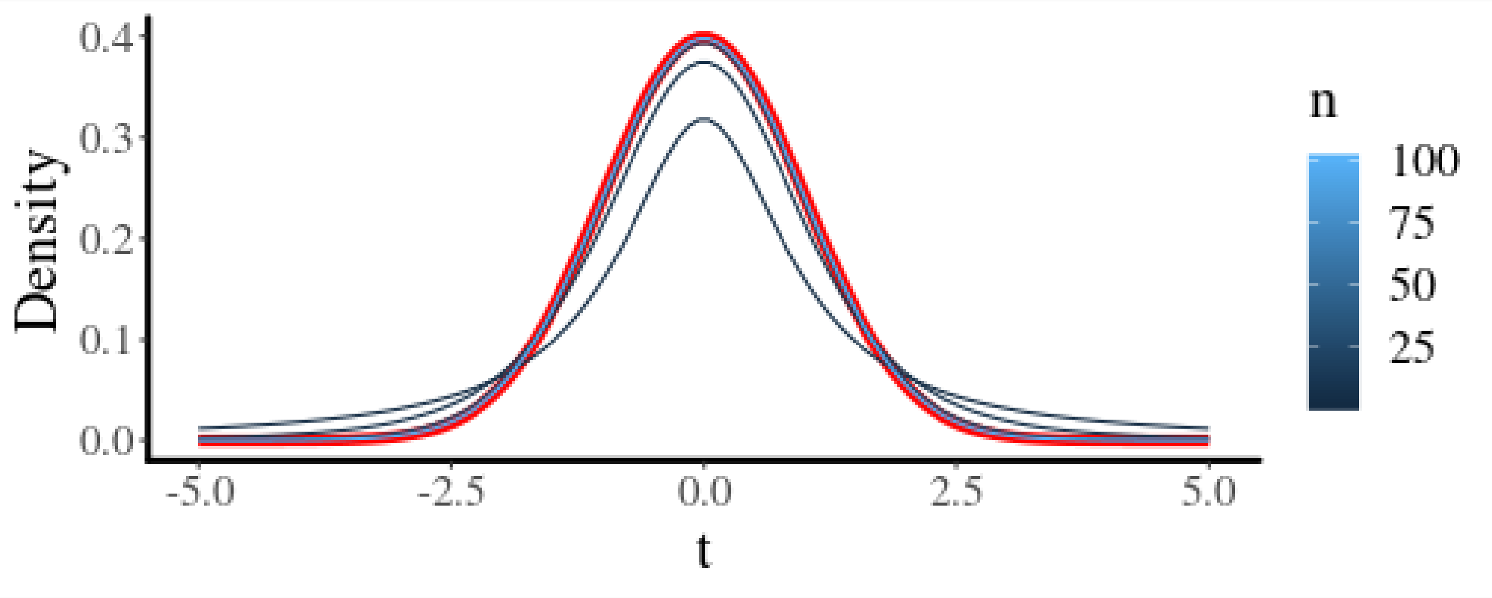
Calculating confidence intervals, z-scores, and p-values

Inference when σ is unknown

When the variance is unknown, we need to account for the additional uncertainty due to estimating σ.

The t-distribution

$$t = \frac{\bar{x} - \mu}{s}$$



As the sample size increases, the t-distribution converges to the z-distribution

Wald confidence intervals

The additional uncertainty in the t-distribution influences the confidence interval. We need to specify the sample size used to estimate s using an argument called the degrees of freedom.

```
xbar <- mean(diff.dat)
n <- length(diff.dat)
se <- sd(diff.dat)/sqrt(n)

xbar - qt(0.975, df=n-1)*se
```

```
## [1] -0.03367423
```

```
xbar + qt(0.975, df=n-1)*se
```

```
## [1] 0.008742242
```

Compare this to the interval that assumes s is known: (-0.0327, 0.0078)

image: USFWS

Excercise 4D

Calculating Wald confidence intervals

The one-sample or paired t-test

The Hawai'ian monk seal data. Test if the difference between treatments is 0.

Do this by hand

```
xbar <- mean(diff.dat)
xbar
```

```
## [1] -0.01246599
```

```
n <- length(diff.dat)
se <- sd(diff.dat)/sqrt(n)
t <- (mean(diff.dat) - 0)/se
t
```

```
## [1] -1.206046
```

```
2*pt(t, df=n-1)
```

```
## [1] 0.2382623
```

Or use the R function `t.test`:

```
t.test(diff.dat)
```

```
##
##      One Sample t-test
##
## data:  diff.dat
## t = -1.206, df = 27, p-value = 0.2383
## alternative hypothesis: true mean is not e
## 95 percent confidence interval:
##  -0.033674225  0.008742242
## sample estimates:
##   mean of x
## -0.01246599
```

The two-sample t-test

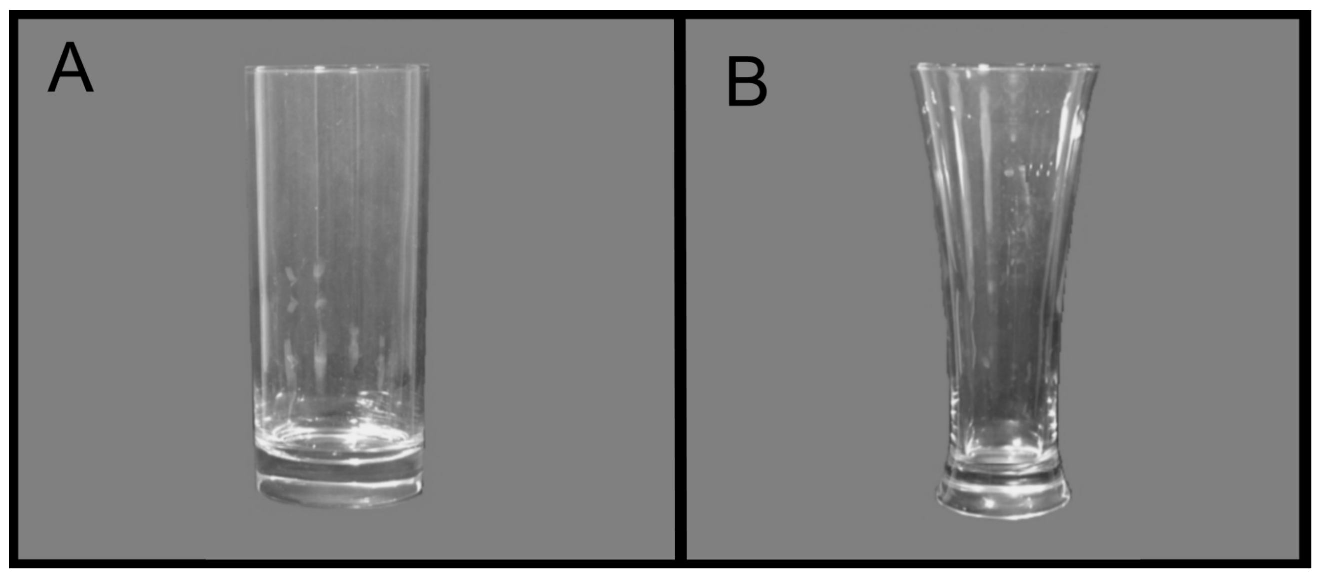
$$t = \frac{\bar{x}_1 - \bar{x}_2 - H_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

image: http://thirstyfortea.com

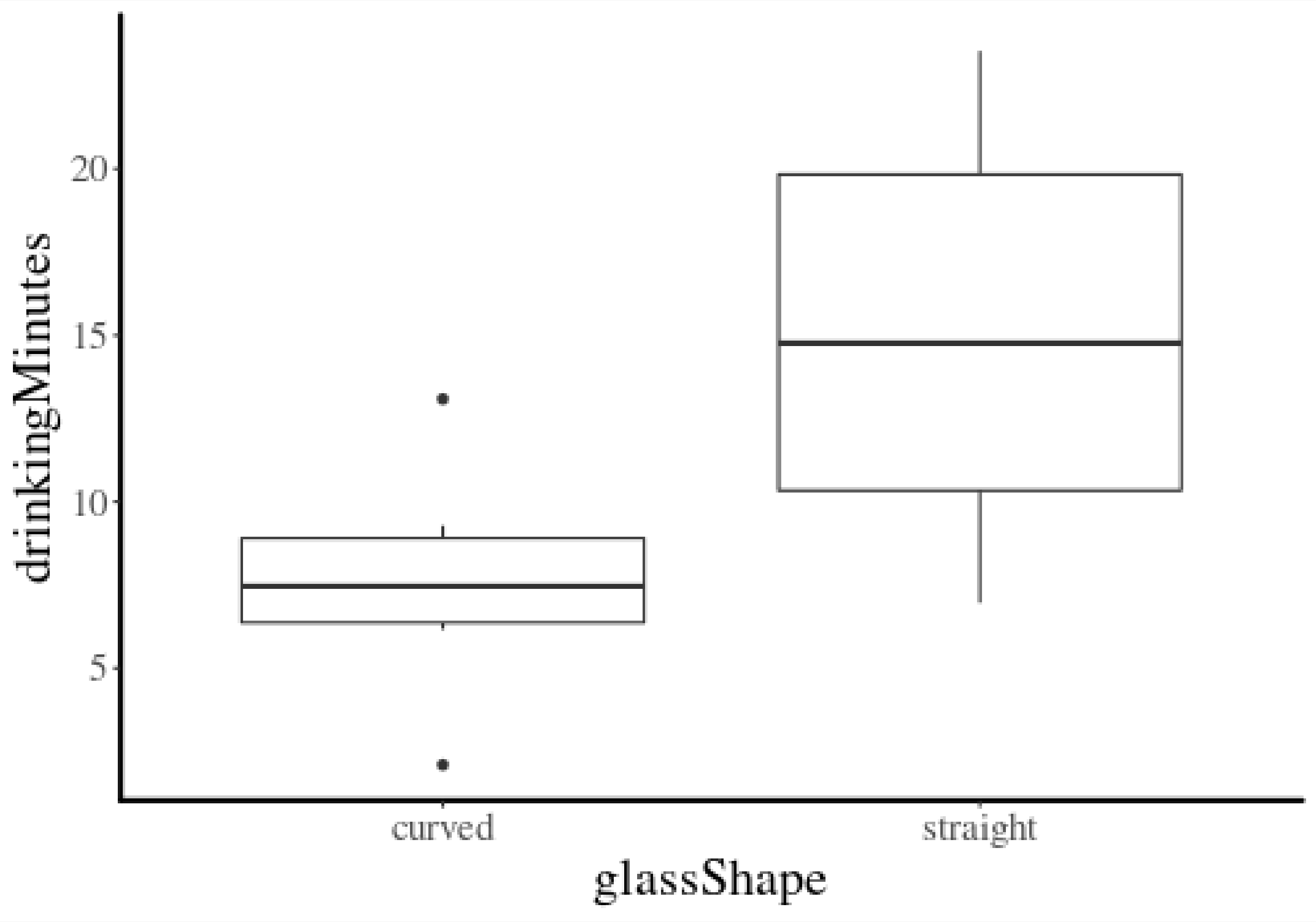
Example

Does the shape of a glass affect the speed a beer is consumed?



Attwood AS, Scott-Samuel NE, Stothart G, Munafò MR (2012) Glass Shape Influences Consumption Rate for Alcoholic Beverages. PLoS ONE 7(8): e43007. https://doi.org/10.1371/journal.pone.0043007

The data



A two-sample `t.test`

In R we define a response and predictor variables using formula:

`response ~ predictor`

```
t.test(drinkingMinutes ~ glassShape, data=glass.dat)
```

```
##
##     Welch Two Sample t-test
##
## data:  drinkingMinutes by glassShape
## t = -3.5111, df = 13.637, p-value = 0.003585
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.802186  -2.837148
## sample estimates:
##   mean in group curved mean in group straight
##               7.593333              14.913000
```

Excercise 4E

One- and two-sample t-tests

Summary

- We used the CLT to determine the sampling distribution
- We linked the sample to the population with confidence intervals
- We tested whether the mean of a sample was equal to a specific value (the null hypothesis)
- We looked at one- and two-sample tests