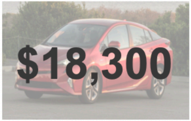
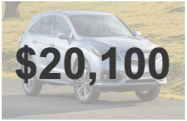
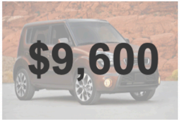
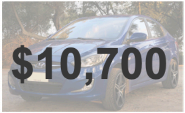
IMS 19: Inference for a single mean

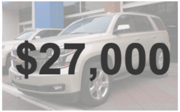
Example











You'd like to know for how much an average car from Awesome Autos sells for. Let's say that you are only able to randomly sample five cars from Awesome Auto.

You observe a mean value of $\overline{x} = 17140$, and a standard deviation of $s_x = 7170.29$

Mathematical model for a mean

Fundamental Theorem of Statistics for Means

When we collect a sufficiently large sample ($n \geq 30$) of independent observations from a population with mean μ and standard deviation σ_x, the sampling distribution of $\overline{x}$ will be nearly normal with

$$\overline{x} \sim N\left(\mu, \sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}\right)$$

where $\sigma_{\overline{x}}$ is called the **standard error**.

Mathematical model for a mean with small samples

If our sample size is small (i.e. $n \leq 30$), then we must check

- **Independence**: Context specific. We are looking for a random sample.
- **Normality:** When a sample is small, we also require that the sample observations come from a normally distributed population. We can relax this condition more and more for larger and larger sample sizes. This is vague, but essentially we just check that there are no clear outliers in the data.

Mathematical model for a mean with small samples

If our data meets the independence and normality assumption and ***we know*** σ_x

$$\frac{\overline{x} - \mu}{\sigma_x / \sqrt{n}} \sim N(0, 1)$$

However, we typically ***do not know*** σ_x, and thus we estimate it using s_x. Where s_x is the estimator we learned in the beginning of the course.

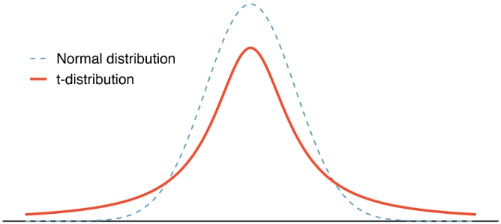
$$T = \frac{\overline{x} - \mu}{s_x / \sqrt{n}} \sim t(df = n - 1)$$

t-distribution

$$T = \frac{\overline{x} - \mu}{s_x/\sqrt{n}} \sim t(df = n - 1)$$

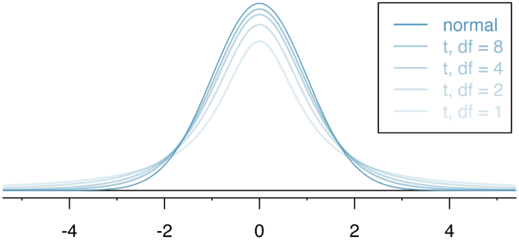
- T is known as the **T-score** or the **test statistic**
- The Z score and T score are computed in the exact same way and are conceptually identical: each represents how many standard errors the observed value is from the null value.

t-distribution



- The t-distribution is very similar to the standard normal distribution (i.e. $N(0,1)$) but it has longer tails.
- The longer tails reflect that s_x is a random quantity estimating σ_x.

t-distribution



- The **degrees of freedom (df)** describes the shape of the t-distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal distribution.
- The degrees of freedom is a function of sample size $df = n - 1$
- Visualize the degrees of freedom by clicking here

Hypothesis Testing with a single mean

It is recommended to default to using a t-distribution when using a hypothesis test for a single mean. The steps are generally the same as before:

1. Frame the research question in terms of hypotheses.
    - μ
2. Collect data (check conditions)
    - Independence and approx normal.
3. Model the randomness that would occur if the null hypothesis was true.
    - Calculate the test statistic T.
4. Analyze the data.
    - Use $T \sim t(df = n - 1)$ to get the p-value.
    - Calculate tail area with `pt(test_stat, df = n-1)`
5. Form a conclusion.

Example

Recall our Awesome Auto data set. $n = 5$, $\overline{x} = 17140$, and $s_x = 7170.29$. Suppose you hear that the Awesome Auto dealership typically sells cars for 20000. You decide to test this claim.

a) Write the hypotheses in symbols.

b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

c) Find and interpret the p-value in this context.

d) What is the conclusion of the hypothesis test when using $\alpha = 0.05$?

Example

a) Write the hypotheses in symbols.

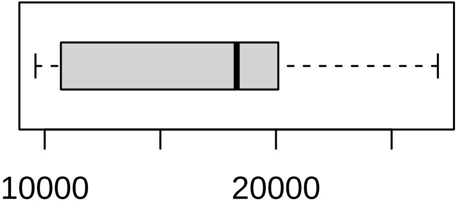
$$H_0 : \mu = 20000$$

$$H_A : \mu \neq 20000$$

Example

b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

- Conditions: Independence, Normality (no extreme outliers)



- Test Statistic under H_0

$$T = \frac{\overline{x} - \mu_0}{\sigma_x / \sqrt{n}} = \frac{17140 - 20000}{7170.29/\sqrt{5}} \approx -0.892$$

- Degrees of Freedom

$$df = 5 - 1 = 4$$

Example

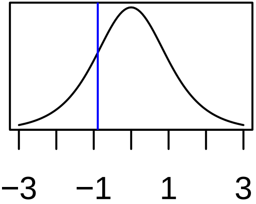
c) Find and interpret the p-value in this context.

THIS IS THE ONLY CODE BOX YOU NEED TO ADJUST:

```
my_df <- 4
null <- 20000
my_se <-  7170.29/sqrt(5)
my_mean <- 17140
my_test_stat <- (my_mean - null)/my_se
```

To draw a picture with a vertical line at your test statistic (T), use this code:

```
curve(dt(x, df = my_df), to = -3, from = 3,
      yaxt="n", ylab = "", xlab = "")
abline(v = my_test_stat, col = "blue")
```



Example

Calculate the left tail area of your test statistic:

```r
pt(my_test_stat, df= my_df)
```

```
[1] 0.2114265
```

P-value for this setting:

```r
2*pt(my_test_stat, df= my_df)
```

```
[1] 0.422853
```

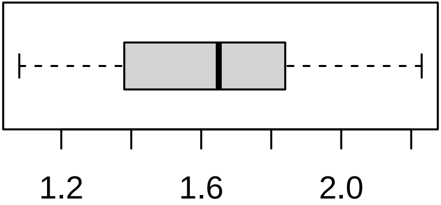
Example

d) What is the conclusion of the hypothesis test?

Our p-value is approximately 0.423, which is larger than $\alpha = 0.05$. We conclude in favor of H_0. It seems possible that the true mean sales price of cars at Awesome Auto could be 20000.

Example

Data on top speeds measured on a laboratory race track for 26 Sagebrush lizards. Scientists wanted to determine if the mean speed of Sagebrush lizards was the same as Western fence lizard (2.32 meters per second). Below is a boxplot and some summary statistics.



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. std.dev
1.080   1.388   1.650   1.613   1.830   2.230   0.324
```

a) Write the hypotheses in symbols.

b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

c) Find and interpret the p-value in this context.

d) What is the conclusion of the hypothesis test?

Confidence Intervals

To create confidence intervals using a single mean with small (and large) samples. Check the same **independence** and **normality** conditions as before.

Confidence Intervals for a mean with t-distribution

$$lower = \overline{x} - t^*_{df} s_{\overline{x}}$$

$$upper = \overline{x} + t^*_{df} s_{\overline{x}}$$

where $s_{\overline{x}} = \frac{s_x}{\sqrt{n}}$

Confidence Intervals

- t^*_{df} represents the number of standard errors away from the mean.
- t^*_{df} changes for each degrees of freedom (df)
- Use code to find the values for t^*_{df} to construct a 95% confidence interval with df $= 18$.

```r
# use qt() to find the t-cutoff (with 95% in the middle)
qt(0.025, df = 18)
```

```
[1] -2.100922
```

```r
qt(0.975, df = 18)
```

```
[1] 2.100922
```

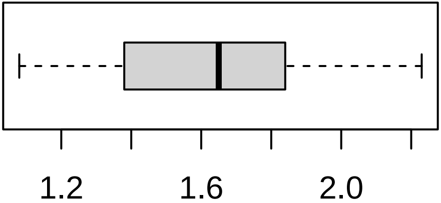
Confidence Intervals

- CIs can only be used with *two-sided* test hypotheses.
- If the CI contains μ_0 (i.e. the value under H_0) then H_0 is supported.
- CIs and p-values will agree for a two-sided test if they use the same α

P-Value Check	CI Check	Conclusion
p-value $\leq \alpha$	$\mu_0 \notin$ [lower, upper]	Against H_0
p-value $> \alpha$	$\mu_0 \in$ [lower, upper]	Favor H_0

Practice Problem

Data on top speeds measured on a laboratory race track for 26 Sagebrush lizards. Scientists wanted to determine if the mean speed of Sagebrush lizards was the same as Western fence lizard (2.32 meters per second). Below is a boxplot and some summary statistics.



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.080   1.387   1.650   1.613   1.830   2.230
```

a) Construct a 95% confidence interval.

b) What would the conclusion of the test be based on this interval?

Practice Problem

Elevated mercury concentrations are an important problem for both dolphins and other animals. We want to investigate the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan. Summary statistics: $\overline{x} = 4.4$, $s_x = 2.3$, $min = 1.7$, and $max = 9.2$.

a) Write the hypotheses in symbols.

b) Check conditions, then calculate the test statistic, T, and the associated degrees of freedom.

c) Find the p-value.

d) Calculate a 90% CI.

e) What is the conclusion of the hypothesis test?

Practice Problem

[IMS 19.16] Working backwards, I. A 95% confidence interval for a population mean, μ, is given as (18.985, 21.015). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 36 observations. Calculate the sample mean, the margin of error, and the sample standard deviation. Assume that all conditions necessary for inference are satisfied. Use the t-distribution in any calculations.

t.test() in R

If we have access to the raw data, we can preform the t-test for a single mean using the `t.test()` function

Important function options/arguments:

- `mu`: the value under H_0
- `alternative`: describes the direction of the alternative hypothesis, the options are `"two.sided"`, `"less"`, `"greater"` than `mu`.
- `conf.level`: if creating a confidence interval, this dictates what confidence level to use.

t.test() in R

```r
awesome_auto <- c(18300, 20100, 9600, 10700, 27000)

t.test(awesome_auto, mu = 20000,
       alternative = "two.sided",
       conf.level = .95)
```

```

    One Sample t-test

data:  awesome_auto
t = -0.8919, df = 4, p-value = 0.4229
alternative hypothesis: true mean is not equal to 20000
95 percent confidence interval:
  8236.914 26043.086
sample estimates:
mean of x
    17140
```