THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Probability Review 2

Lecture 3

STA 371G

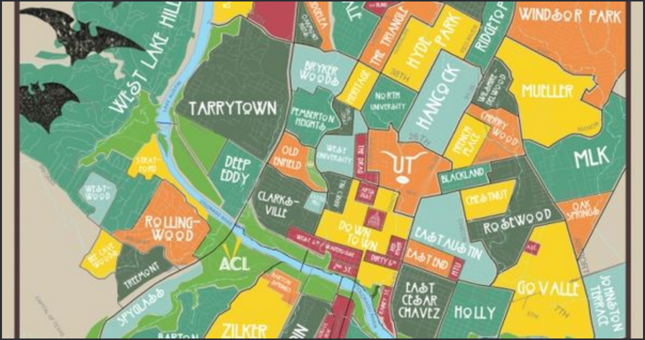
Sample vs Population

Find out the average house price in Austin.

Sample vs Population

Find out the average house price in Austin.

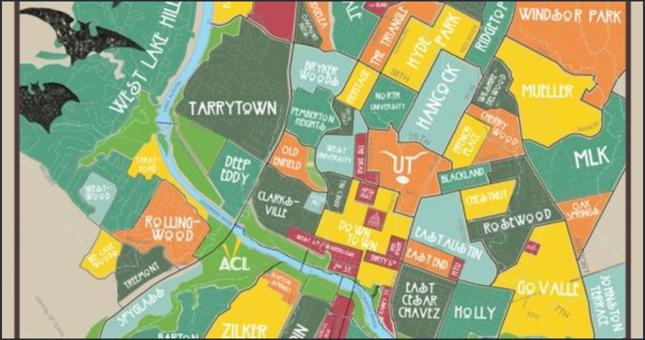
How would you do that?



Sample vs Population

Find out the average house price in Austin.

How would you do that?

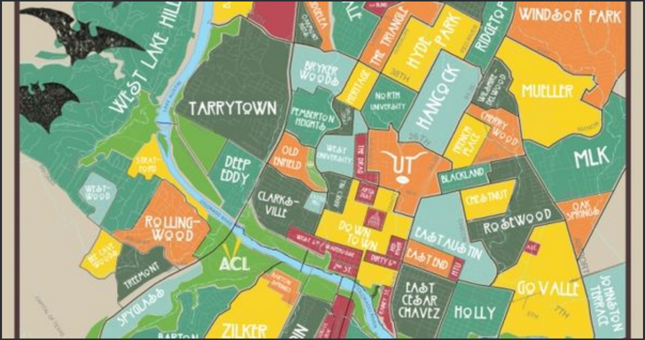


Look at each house price?

Sample vs Population

Find out the average house price in Austin.

How would you do that?



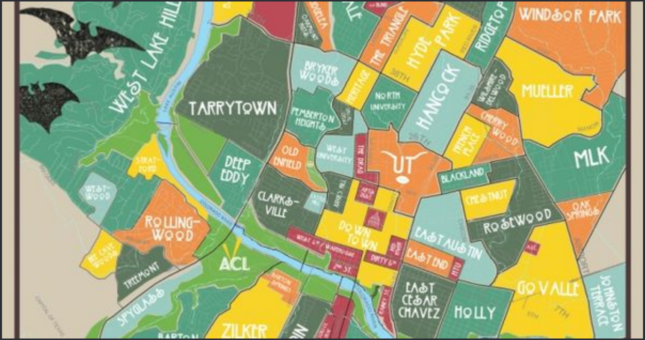
Look at each house price?

360,000 houses in Austin!

Sample vs Population

Find out the average house price in Austin.

How would you do that?



Look at each house price?

360,000 houses in Austin!

Can we do something smarter?

Sample vs Population

A smarter approach:

Sample vs Population

A smarter approach:

- Pick n houses randomly (e.g. $n = 100$)

Sample vs Population

A smarter approach:

- Pick n houses randomly (e.g. $n = 100$)
- Take the average of the prices of these n houses

Sample vs Population

A smarter approach:

- Pick n houses randomly (e.g. $n = 100$)
- Take the average of the prices of these n houses
- Hope that your estimate is close to the true price average.

Sample vs Population

A smarter approach:

- Pick n houses randomly (e.g. $n = 100$)
- Take the average of the prices of these n houses
- Hope that your estimate is close to the true price average.

Just like making polls to predict election results!

Sample vs Population

	Population	**Sample**
Members	all house prices	prices you picked
Average	population mean	sample mean
Variance	population variance	sample variance

Sample vs Population

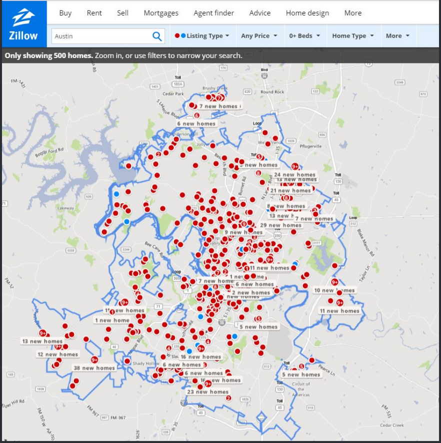
	Population	**Sample**
Members	all house prices	prices you picked
Average	population mean	sample mean
Variance	population variance	sample variance

Estimating a population parameter (population mean) based on a sample statistic (sample mean).

Collecting a sample

Zillow.com, “Austin, TX.”

- Click “More Map”
- Select 15 houses, note their prices in an R script.
- Do not discard any price, use the first 15
- Try to represent different regions



Only showing 500 homes. Zoom in, or use filters to narrow your search.

Collecting a sample

Your R script should look like this

```r
# Create a vector of house prices (You should have 15 price data)
sample_house_prices <- c(327000,276000,513000)
# Calculate sample statistics
sample_mean <- mean(sample_house_prices)
sample_variance <- var(sample_house_prices)
sample_standard_deviation <- sd(sample_house_prices)
# Sample mean of first 5 houses
sample_mean_5 <- mean(sample_house_prices[1:5])
# Print them to console
cat("Sample Mean", sample_mean)
cat("Sample Variance", sample_variance)
cat("Sample Standard Deviation", sample_standard_deviation)
cat("Sample Mean of first 5 houses",sample_mean_5)
```

Sampling Distribution

On Learning Catalytics, enter your results.

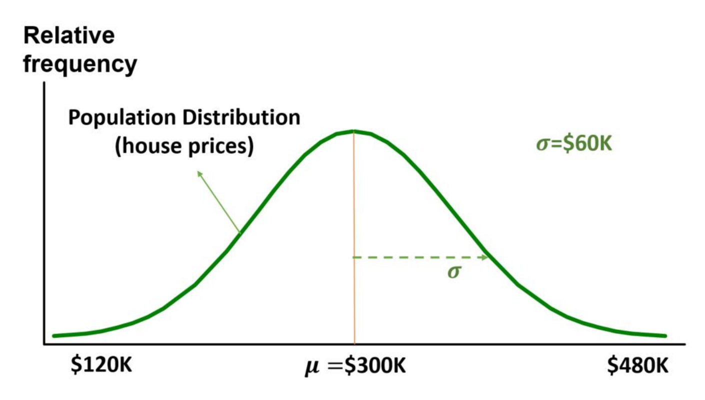
And here is what they look like…

Sampling Distribution

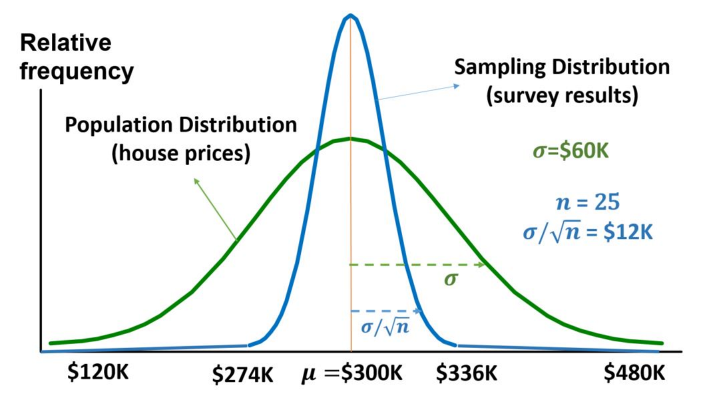
Distribution of your answers ⟶ Sampling distribution

Statistic	**Population**	**Sample Mean**
Mean	μ	μ
Standard Deviation	σ	$\sigma/\sqrt{n}$

Sampling Distribution



Sampling Distribution



Sampling Distribution

Assume $\mu =$\$300K, $\sigma =$\$60K.

	n	$\sigma/\sqrt{n}$	3 std. dev. range (99.7%)
Survey 1	25	\$12K	\$264K \$336K
Survey 2	100	\$6K	\$282K \$318K
Survey 3	3600	\$1K	\$297K ... \$303K

Sampling Distribution

Let’s compare sample mean of 5 houses vs 15 houses.

What do you expect to see?

t Distribution

We often do not know population variance and use sample variance instead.

In that case, the sample mean will have a *t* distribution.

Hypothesis Testing

Hypothesis: Average house price in Austin is $1M.

Hypothesis Testing

Hypothesis: Average house price in Austin is $1M.

Your survey on 25 houses: Average house price is $\$305K$.

Hypothesis Testing

Hypothesis: Average house price in Austin is \$1M.
Your survey on 25 houses: Average house price is $\$305K$.

Questions, questions...

Hypothesis Testing

Hypothesis: Average house price in Austin is $1M.

Your survey on 25 houses: Average house price is $\$305K$.

Questions, questions...

- Would you reject the hypothesis? Why?

Hypothesis Testing

Hypothesis: Average house price in Austin is $1M.

Your survey on 25 houses: Average house price is $\$305K$.

Questions, questions...

- Would you reject the hypothesis? Why?
- Is it possible that, out of bad luck, you picked the cheapest houses?

Hypothesis Testing

Hypothesis: Average house price in Austin is $1M.
Your survey on 25 houses: Average house price is $\$305K$.

Questions, questions...

- Would you reject the hypothesis? Why?
- Is it possible that, out of bad luck, you picked the cheapest houses?
- Would you be more comfortable with your conclusion if you had 1000 houses in your survey?

Hypothesis Testing

Hypothesis: Average house price in Austin is $1M.
Your survey on 25 houses: Average house price is $*305K*.

Questions, questions...

- Would you reject the hypothesis? Why?
- Is it possible that, out of bad luck, you picked the cheapest houses?
- Would you be more comfortable with your conclusion if you had 1000 houses in your survey?
- When should you reject the hypothesis? When not?

P-Value

Your sample mean: $305K.

P-Value

Your sample mean: \$305K.

$H_0 : \mu = \$1M$ (Null hypothesis)
$H_1 : \mu < \$1M$ (Alternative hypothesis)

P-Value

Your sample mean: \$305K.

$H_0 : \mu = \$1M$ (Null hypothesis)
$H_1 : \mu < \$1M$ (Alternative hypothesis)

The *P*-value is “the probability of observing such an extreme (\$305K or less) sample statistic given the null hypothesis is true.”

P-Value

Your sample mean: $305K.

$H_0 : \mu = \$1M$ (Null hypothesis)
$H_1 : \mu < \$1M$ (Alternative hypothesis)

The *P*-value is “the probability of observing such an extreme ($305K or less) sample statistic given the null hypothesis is true.”

- P-value $\leq \alpha$, reject the null hypothesis

P-Value

Your sample mean: \$305K.

$H_0 : \mu = \$1M$ (Null hypothesis)
$H_1 : \mu < \$1M$ (Alternative hypothesis)

The P-value is "the probability of observing such an extreme (\$305K or less) sample statistic given the null hypothesis is true."

- P-value $\leq \alpha$, reject the null hypothesis
- P-value $> \alpha$, do not reject the null hypothesis

P-Value

Your sample mean: \$305K.

$H_0 : \mu = \$1M$ (Null hypothesis)
$H_1 : \mu < \$1M$ (Alternative hypothesis)

The *P*-value is “the probability of observing such an extreme (\$305K or less) sample statistic given the null hypothesis is true.”

- P-value $\leq \alpha$, reject the null hypothesis
- P-value $> \alpha$, do not reject the null hypothesis

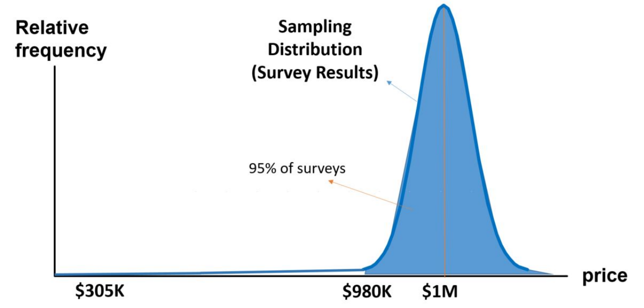
α is usually chosen as 0.05 prior to sampling.

P-Value

If the null hypothesis were true…

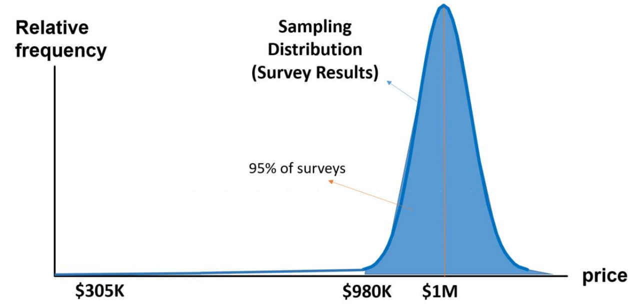
P-Value

If the null hypothesis were true…



P-Value

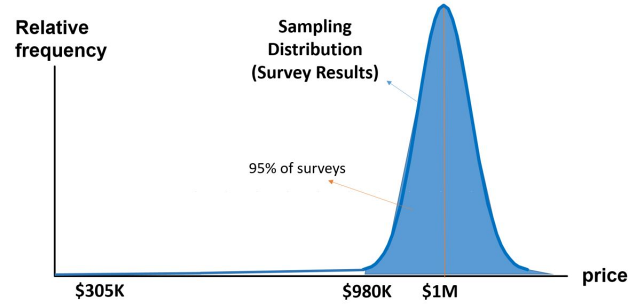
If the null hypothesis were true...



P-value is smaller than 10^{-100}, while $\alpha = 0.05$.

P-Value

If the null hypothesis were true...



P-value is smaller than 10^{-100}, while $\alpha = 0.05$.
Rather than thinking you are cursed, you simply reject the hypothesis!

P-Value

Learning Catalytics...

P-Value

Learning Catalytics…

Your sample mean = $305K.

P-Value

Learning Catalytics...

Your sample mean = \$305K.

H_0: Average house price is \$390K.
Would you reject the hypothesis?
P-value = 0.01, α=0.05

P-Value

Learning Catalytics...

Your sample mean = \$305K.

H_0: Average house price is \$390K.
Would you reject the hypothesis?
P-value = 0.01, α=0.05

H_0: Average house price is \$320K.
Would you reject the hypothesis?
P-value = 0.34, α=0.01

P-Value

Learning Catalytics...

Your sample mean = $305K.

H_0: Average house price is $390K.
Would you reject the hypothesis?
P-value = 0.01, α=0.05

H_0: Average house price is $320K.
Would you reject the hypothesis?
P-value = 0.34, α=0.01

Confidence Interval

Sample mean is not equal to the population mean, but “close.”

Confidence Interval

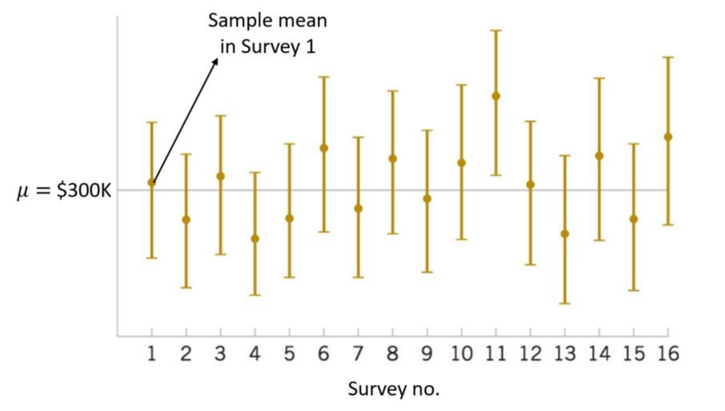
Sample mean is not equal to the population mean, but “close.”

Confidence interval is a range that includes the population mean with a certain level of “confidence.”

Confidence Interval

Sample mean is not equal to the population mean, but “close.”

Confidence interval is a range that includes the population mean with a certain level of “confidence.”



Confidence Interval

Add the following to your R script

```r
# Calculate 95% confidence interval (default)
avg_price_ci_95 <- t.test(sample_house_prices)
# Calculate 99% confidence interval
avg_price_ci_99 <- t.test(sample_house_prices, conf.level = 0.99)
# Display results
cat("95% confidence interval is:", avg_price_ci_95$conf.int)
cat("99% confidence interval is:", avg_price_ci_99$conf.int)
```

Enter your results on Learning Catalytics.