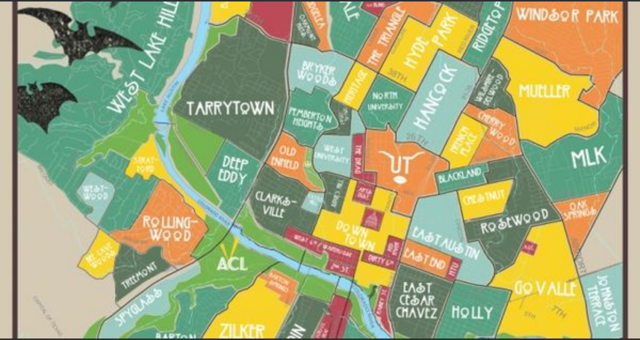
THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Probability Review 2

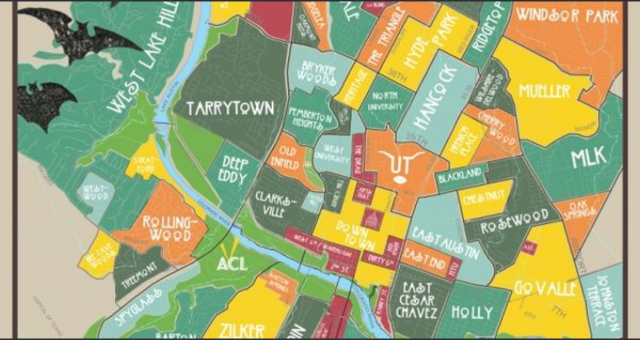
Lecture 3

STA 371G

How would you figure out the average house price in Austin?

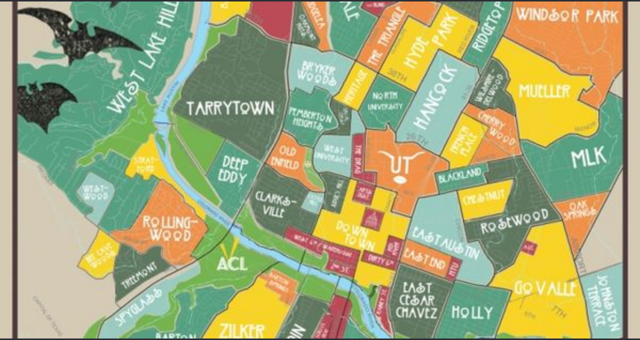


How would you figure out the average house price in Austin?



Look up each house price?

How would you figure out the average house price in Austin?



Look up each house price?

There are 360,000 houses in Austin — is there a better way?

Sample vs Population

A faster approach:

Sample vs Population

A faster approach:

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

Sample vs Population

A faster approach:

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

Sample vs Population

A faster approach:

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

Sample vs Population

A faster approach:

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

Just like making polls to predict election results!

Sample vs Population

	Population	**Sample**
Members	all houses	houses you selected
Mean	population mean (μ)	sample mean ($\hat{\mu}$)
SD	population SD (σ)	sample SD ($\hat{\sigma}$)

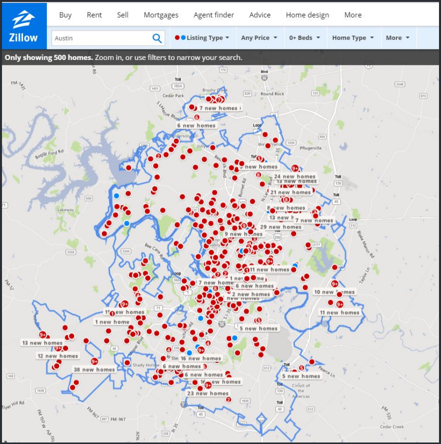
Sample vs Population

	Population	Sample
Members	all houses	houses you selected
Mean	population mean (μ)	sample mean ($\hat{\mu}$)
SD	population SD (σ)	sample SD ($\hat{\sigma}$)

We will estimate a population parameter (population mean) based on a sample statistic (sample mean).

Collecting a sample

- Go to `zillow.com` and search for Austin, TX.
- Click “More Map.”
- Select 15 houses (try to get houses from all over town in a representative way), noting their prices in an R script.
- Do not discard any price, use the first 15 you find.



Collecting a sample

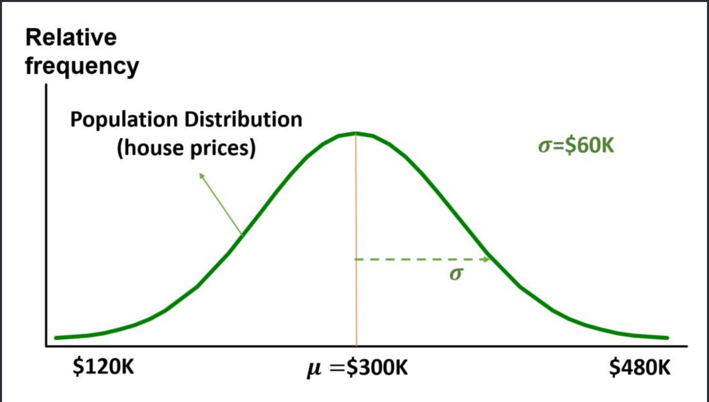
```r
# Create a vector of house prices (you should have 15 prices)
sample_house_prices <- c(327000, 276000, 513000)
# Calculate sample statistics
sample_mean <- mean(sample_house_prices)
sample_variance <- var(sample_house_prices)
sample_standard_deviation <- sd(sample_house_prices)
# Sample mean of first 5 houses
sample_mean_5 <- mean(sample_house_prices[1:5])
```

Sampling Distribution

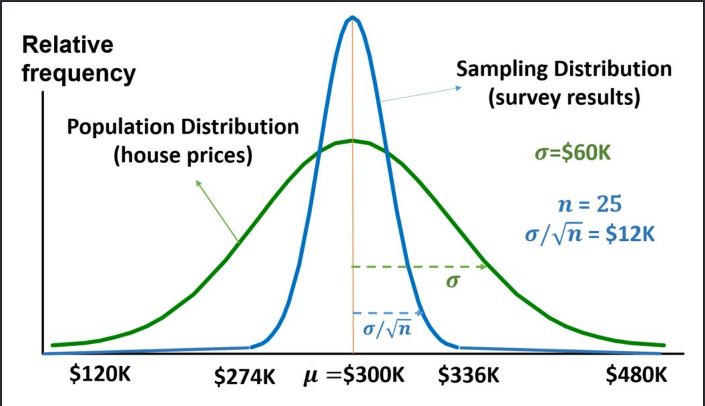
Distribution of your answers $\longrightarrow$ Sampling distribution

Statistic	**Population**	**Sampling Distribution**
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 SD 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.

- 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$.

- 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.

- 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 $\$305K$.

- 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-values

Your sample mean: $\hat{\mu} = \$305K$

p-values

Your sample mean: $\hat{\mu} = \$305K$

- $H_0 : \mu = \$1M$ (null hypothesis)

The *p*-value is “the probability of observing such an extreme ($\leq \$305K$) sample statistic if in fact the null hypothesis is true.”

p-values

Your sample mean: $\hat{\mu} = \$305K$

- $H_0 : \mu = \$1M$ (null hypothesis)
- $H_1 : \mu < \$1M$ (alternative hypothesis)

The p-value is "the probability of observing such an extreme ($\leq \$305K$) sample statistic if in fact the null hypothesis is true."

p-values

Your sample mean: $\hat{\mu} = \$305K$

- $H_0 : \mu = \$1M$ (null hypothesis)
- $H_1 : \mu < \$1M$ (alternative hypothesis)

The p-value is "the probability of observing such an extreme ($\leq \$305K$) sample statistic if in fact the null hypothesis is true."

p-values

Your sample mean: $\hat{\mu} = \$305K$

- $H_0 : \mu = \$1M$ (null hypothesis)
- $H_1 : \mu < \$1M$ (alternative hypothesis)

The p-value is "the probability of observing such an extreme ($\leq \$305K$) sample statistic if in fact the null hypothesis is true."

- If $p < \alpha$, reject the null hypothesis

p-values

Your sample mean: $\hat{\mu} = \$305K$

- $H_0 : \mu = \$1M$ (null hypothesis)
- $H_1 : \mu < \$1M$ (alternative hypothesis)

The p-value is “the probability of observing such an extreme ($\leq \$305K$) sample statistic if in fact the null hypothesis is true.”

- If $p < \alpha$, reject the null hypothesis
- If $p \geq \alpha$, do not reject the null hypothesis

p-values

Your sample mean: $\hat{\mu} = \$305K$

- $H_0 : \mu = \$1M$ (null hypothesis)
- $H_1 : \mu < \$1M$ (alternative hypothesis)

The p-value is "the probability of observing such an extreme ($\leq \$305K$) sample statistic if in fact the null hypothesis is true."

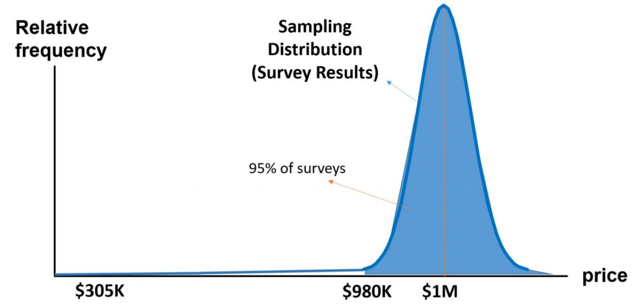
- If $p < \alpha$, reject the null hypothesis
- If $p \geq \alpha$, do not reject the null hypothesis

α is usually chosen as 0.05 prior to sampling.

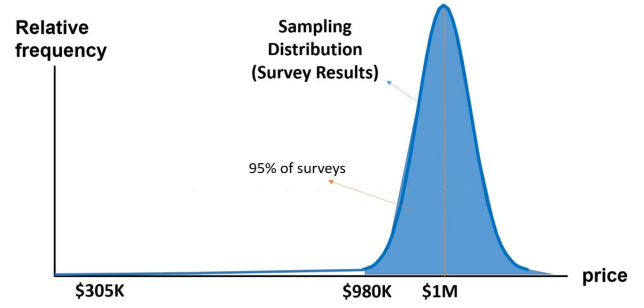

"A ROLLICKING COMEDY."
-JOHN ANDERSON, VARIETY
WRITTEN, DIRECTED BY AND STARRING LAKE BELL
IN A WORLD...
SPEAK UP AND LET YOUR VOICE BE HEARD
WINNER
SUNDANCE
www.inaworldmovie.com
SONY
COMING SOON

Let’s assume we are in a world where H_0 is true…

Let's assume we are in a world where H_0 is true...

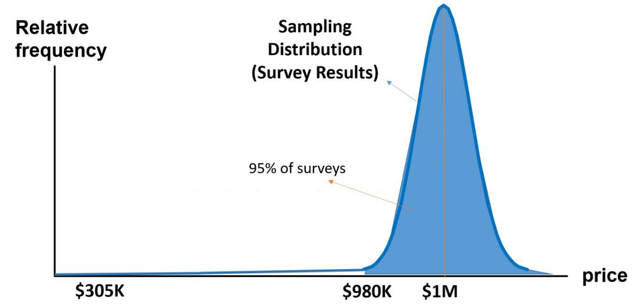


Let's assume we are in a world where H_0 is true...



What is the probability of a sample where $\hat{\mu} \leq \$305K$?

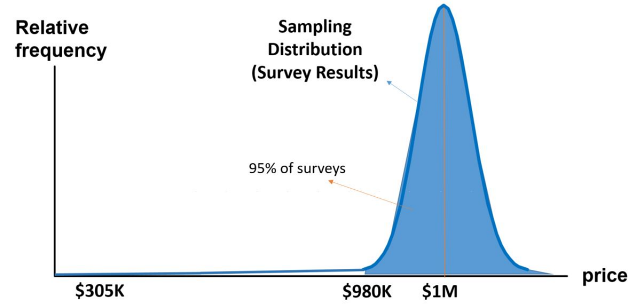
Let's assume we are in a world where H_0 is true...



What is the probability of a sample where $\hat{\mu} \leq \$305K$?

This is p, the area to the left of $\$305K$. $p < 10^{-100}$!

Let's assume we are in a world where H_0 is true...



What is the probability of a sample where $\hat{\mu} \leq \$305K$?

This is p, the area to the left of $\$305K$. $p < 10^{-100}$!

Since $p < \alpha = 0.05$, reject the null hypothesis!

Confidence Intervals

The sample mean is probably not exactly equal to the population mean, but it's almost certainly "close."

Confidence Intervals

The sample mean is probably not exactly equal to the population mean, but it’s almost certainly “close.”

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

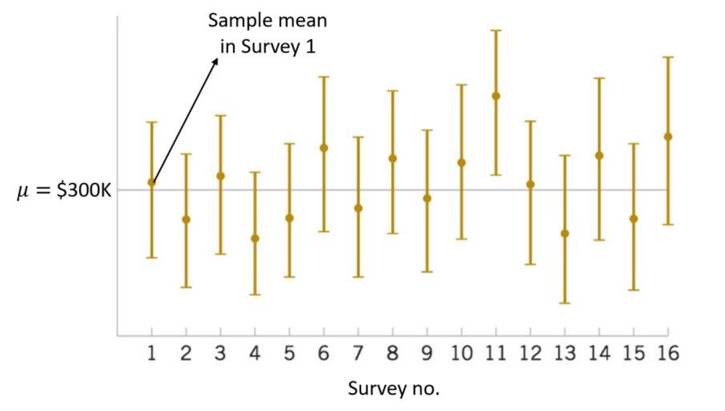
Confidence Intervals

The sample mean is probably not exactly equal to the population mean, but it's almost certainly "close."

A confidence interval is a range that includes the population mean with a certain level of "confidence."

Confidence Intervals

Each sample will have a different 95% confidence interval, and 95% of such intervals will contain the population mean:



Confidence Interval

```
# Calculate 95% confidence interval (default)
avg_price_ci_95 <- t.test(sample_house_prices, conf.level=0.95)
# Calculate 99% confidence interval
avg_price_ci_99 <- t.test(sample_house_prices, conf.level=0.99)
```