THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

# Probability Review 2

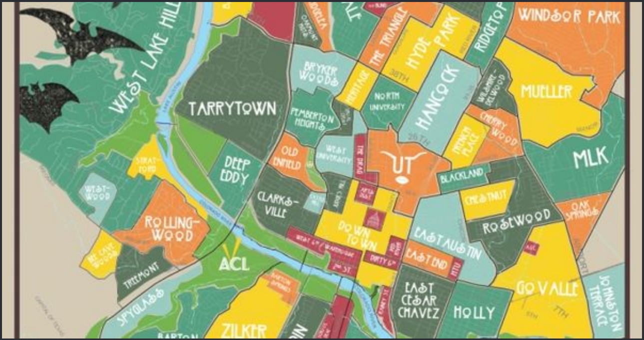
## Lecture 2

### STA 371G

# 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:

- 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

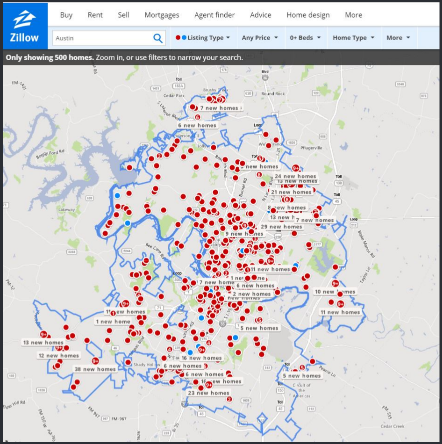
| | Population | Sample |
|---|---|---|
| Members | all house prices | prices you picked |
| Average | population mean | sample mean |
| Variance | population variance | sample variance |

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



# Collecting a sample

Your R script should look like this

```
# 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 | $\mu$ | $\mu$ |
| Standard Deviation | $\sigma$ | $\sigma/\sqrt{n}$ |

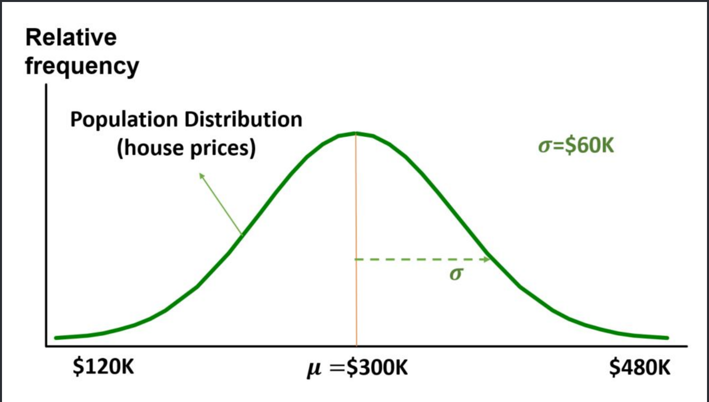
What would you expect if you had 10 000 houses in your survey?

# Sampling Distribution

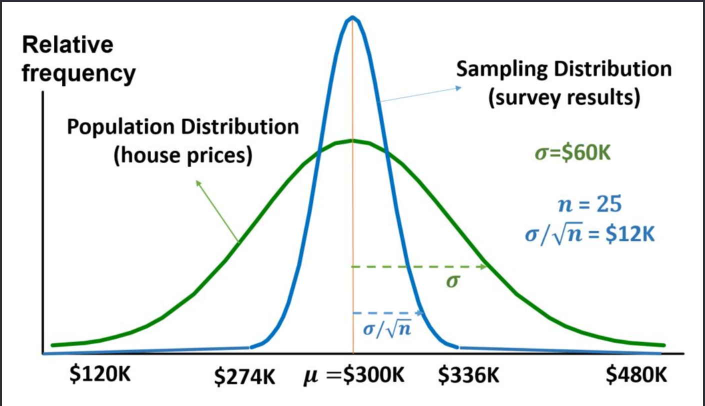
Let's compare sample mean of 5 houses vs 15 houses.

What do you expect to see?

# 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 | \$274K ............ \$336K |
| Survey 2 | 100 | \$6K | \$282K ....... \$318K |
| Survey 3 | 3600 | \$1K | \$297K ... \$303K |

# *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: The average house price in Austin is \$1M.
Your survey on 25 houses: Average price is \$305$K$.

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.

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

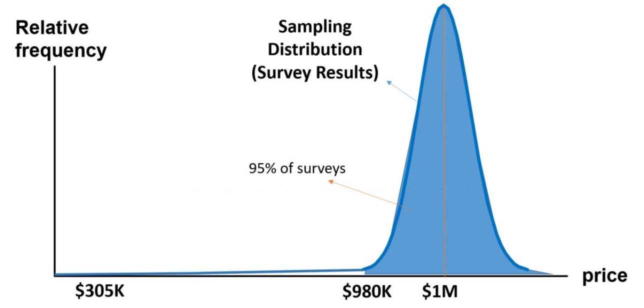
The *P*-value is "the probability of observing such an extreme (\$305K or less) sample statistic given the null hypothesis is true."

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

$\alpha$ is usually chosen as $0.05$ prior to sampling.

# 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!

# Confidence Interval

Content Here

# Confidence Interval 2

Content Here