THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

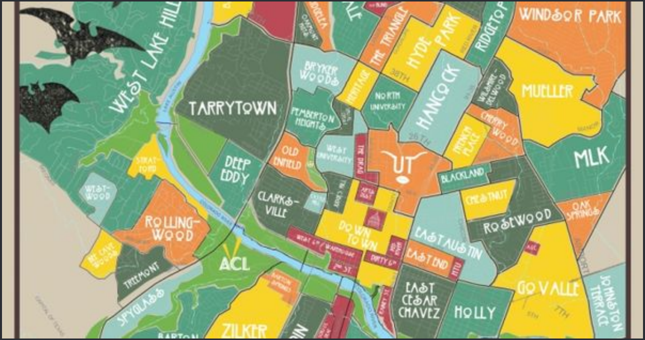
Probability Review 2

Lecture 2

STA 371G

Sample vs Population

Suppose you need to find out what the average house price in Austin is. How would you do that?



Planning to look into each house price? You better start now!

Because there are 360,000 houses in Austin!

Can we do something smarter?

Sample vs Population

A smarter approach would be to

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

Just like doing polls to predict election results!

Sample vs Population

Estimating a population parameter (in this case mean) based on a sample statistic (in this case, sample mean).

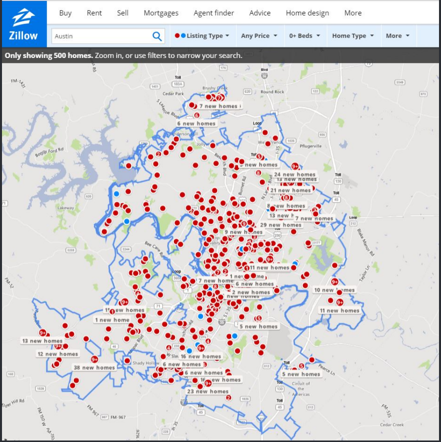
- Population ← all houses in Austin
- Sample ← *n* houses you picked
- Population mean ← price average of all houses in Austin
- Sample mean ← average price of *n* houses in your sample

We could also estimate other population parameters, such as variance using the sample variance.

Collecting a sample

On Zillow.com, type “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



Austin

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.

We will plot their histogram later.

Sampling Distribution

Everyone found a different sample mean, which one is correct?
None.

Sample mean (your answers) itself has a distribution, separate from the house price distribution in Austin. This is called sampling distribution.

Expected value of the sample mean = Population mean
Standard deviation of the sample mean (s) = Standard deviation of the population (σ) / $\sqrt{n}$

The higher the sample size (n), the lower the standard deviation of the sample mean (s).

Sampling Distribution

Assume the average house price in Austin (μ) is \$300K and the standard deviation $\sigma =$\$60K. (You don't know these!)

(Assuming normal distribution), 99.7% of the houses are between $\$[120K, 480K]$.

Sample mean's mean \$300K, standard deviation $\$60K/\sqrt{n}$.

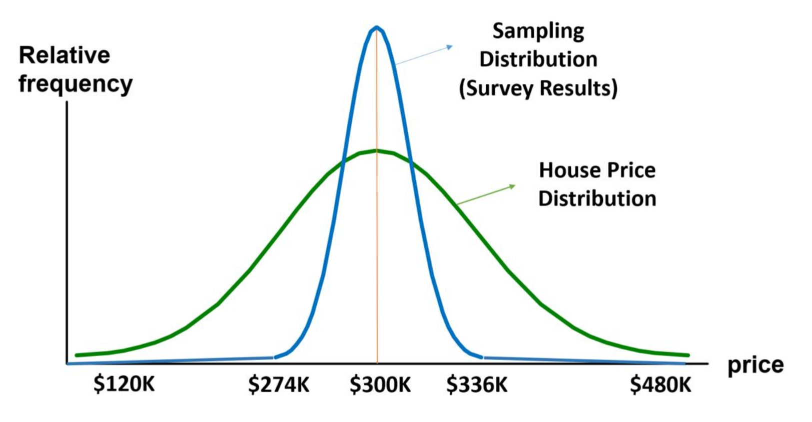
$n = 25$ houses, $s = 60K/5 = 12K$, 99.7% of surveys in $\$[274K, 336K]$.
$n = 100$ houses, $s = 60K/10 = 6K$, 99.7% of surveys $\$[282K, 318K]$.

Sampling Distribution

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

What do you expect to see?

Sampling Distribution



t Distribution

When population is normally distributed with μ and σ^2, sample mean has a normal distribution mean μ and variance σ^2/n.

If we don't know σ (we often don't!), how can we use sample mean's distribution?

We use sample variance (s) instead. In that case, the sample mean will (after normalization) have a t distribution.

Hypothesis Testing

Hypothesis: The average house price in Austin is $1M.

Your survey on 30 houses: Average 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

$H_0 : \mu = \$1M$ (Null hypothesis)
$H_1 : \mu < \$1M$ (Alternative hypothesis)
Assume that your sample mean is \$305K.

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

α is usually chosen as 0.05 prior to sampling.

P-Value

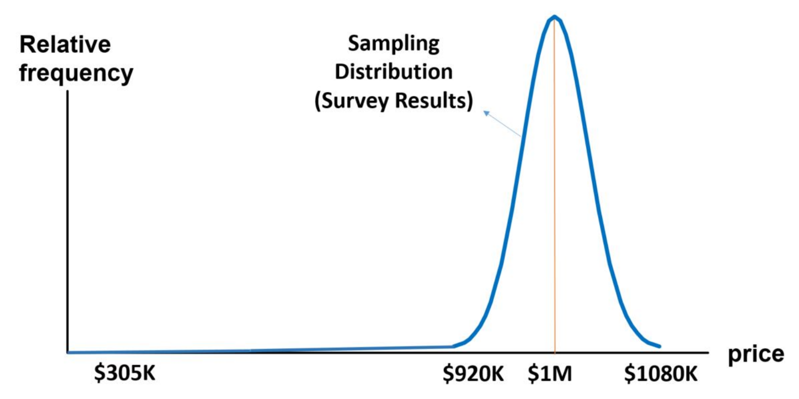
Consider the null hypothesis true with $\mu = \$1M$ and $\sigma = \$200K$. For $n = 25$, the sampling distribution has a mean $\$1M$ and the standard deviation $\$40000$.

Approximately 95% of the surveys will give a result in $\$[\,920K, 1080K\,]$.

In fact, P-value of a sample mean of \$305K is 5×10^{-143}.

Rather than thinking you are cursed, you simply reject the hypothesis!

P-Value


Relative frequency
Sampling Distribution (Survey Results)
price
$305K
$920K
$1M
$1080K

Confidence Interval

Content Here

Confidence Interval 2

Content Here