Module 5: Statistics

PDAT 610G – Truman State University

Statistics collects and uses data to draw conclusions in the face of uncertainty.

1. Identify the research objective
2. Collect information
3. Organize and summarize information graphically and numerically
   (*Descriptive Statistics*)
4. Draw conclusions from the information
   (*Inferential Statistics*)
   (a) Make a decision
   (b) Solve a problem
   (c) Design or evaluate a product or process

Statistical methods describe and account for the effects of variability.

- *Variability* encompasses anything that causes repeated measurements to produce different results.
  - Random error from a measurement device or process.
  - Sources of bias that systematically skew results.
  - Effects of randomization in methods themselves.

The way data is gathered affects the conclusions you can draw.

Retrospective study:

Uses existing sources of data (gathered for a different purpose).

Observational study:

Collects new data, but does not attempt to influence values of the response or explanatory variables.

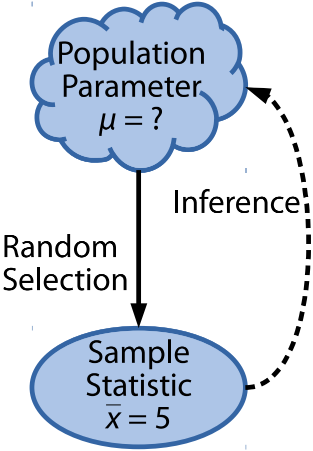
Designed experiment:

Applies *treatments* to experimental units, then measures the effect on the response variable.

Examples/Question on Inferential/Descript or types of studies???

Inferential statistics uses sample data to learn about the underlying population.

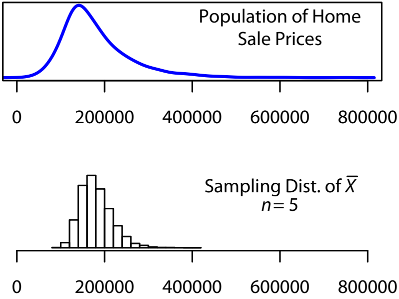
- We don’t know about the *whole* population.
- We can calculate the *sample* mean.
- The sample mean estimates the true population mean.
- Understanding the variability in the population, we can calculate the likely accuracy of our estimate.



The Central Limit Theorem

The Central Limit Theorem describes the distribution of all possible sample means obtainable from a population.

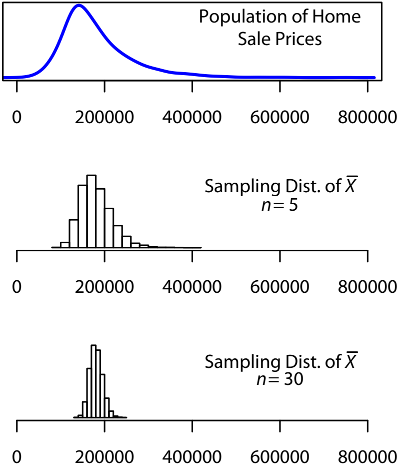
- The sampling distribution of $\bar{X}$ is centered on the population mean: $\mathrm{E}(\bar{X}) = \mu$.



The Central Limit Theorem

The Central Limit Theorem describes the distribution of all possible sample means obtainable from a population.

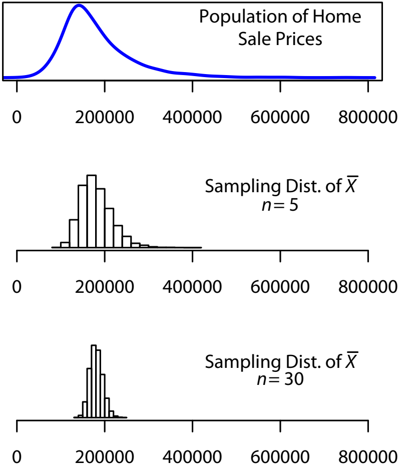
- The sampling distribution of $\bar{X}$ is centered on the population mean: $\mathrm{E}(\bar{X}) = \mu$.
- As n increases, sample means cluster more closely around the population mean: $\mathrm{VAR}(\bar{X}) = \sigma^2/n$.



The Central Limit Theorem

The Central Limit Theorem describes the distribution of all possible sample means obtainable from a population.

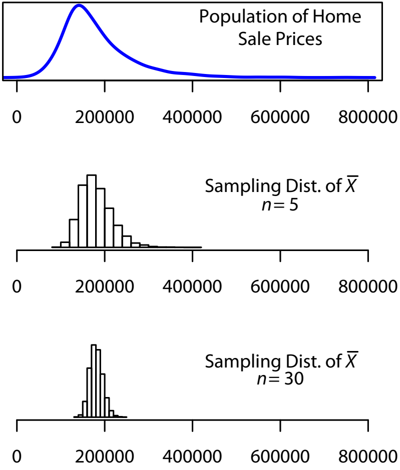
- The sampling distribution of $\bar{X}$ is centered on the population mean: $\mathrm{E}(\bar{X}) = \mu$.
- As n increases, sample means cluster more closely around the population mean: $\mathrm{VAR}(\bar{X}) = \sigma^2/n$.
- As n increases, the sampling distribution of $\bar{X}$ approaches a normal distribution.



The Central Limit Theorem

The Central Limit Theorem describes the distribution of all possible sample means obtainable from a population.

- The sampling distribution of $\bar{X}$ is centered on the population mean: $\mathrm{E}(\bar{X}) = \mu$.
- As n increases, sample means cluster more closely around the population mean: $\mathrm{VAR}(\bar{X}) = \sigma^2/n$.
- The distribution of $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ approaches the standard normal as $n \to \infty$.



Logic of Hypothesis Testing: If observed data is not consistent with your assumptions, question your assumptions.

- *Null Hypothesis* (H_0): Represents *status quo*, or your default assumption.

Logic of Hypothesis Testing: If observed data is not consistent with your assumptions, question your assumptions.

- *Null Hypothesis* (H_0): Represents *status quo*, or your default assumption.
- *Alternate Hypothesis* (H_1, sometimes H_A): An alternative to the null hypothesis.

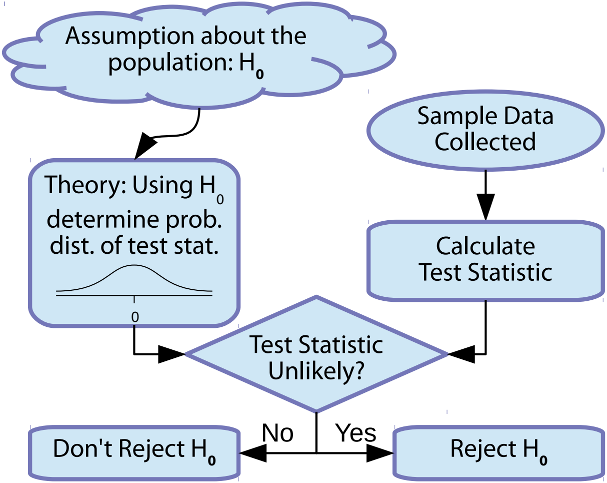
Logic of Hypothesis Testing: If observed data is not consistent with your assumptions, question your assumptions.

- *Null Hypothesis* (H_0): Represents *status quo*, or your default assumption.
- *Alternate Hypothesis* (H_1, sometimes H_A): An alternative to the null hypothesis.
- *p*-value: If you believe H_0 is true, how likely would your observed sample data be?

Logic of Hypothesis Testing: If observed data is not consistent with your assumptions, question your assumptions.

- *Null Hypothesis* (H_0): Represents *status quo*, or your default assumption.
- *Alternate Hypothesis* (H_1, sometimes H_A): An alternative to the null hypothesis.
- *p*-value: If you believe H_0 is true, how likely would your observed sample data be?
- If your observed sample data appears unlikely enough, that should lead you to question H_0.

A *hypothesis test* uses a *test statistic* to evaluate hypotheses.



Hypotheses examples???

The One-Sample *t*-Test (*P*-Value)

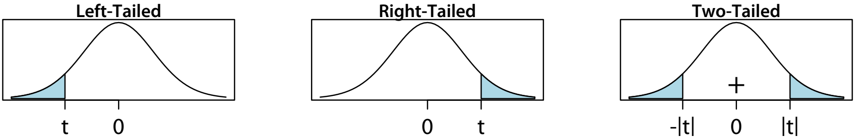
- Normal population or large sample ($n \geq 30$).
- σ *unknown*.

Step 1: Determine H_0: $\mu = \mu_0$ and H_1.

Step 2: Select α. (prior to analysis!)

Step 3: Compute $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

Step 4: Calculate the *P*-value (using *t*-dist. with df = $n - 1$).

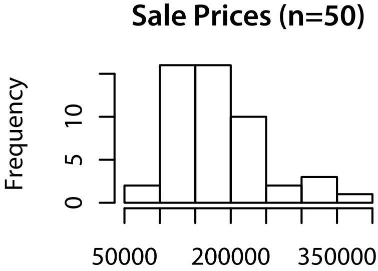


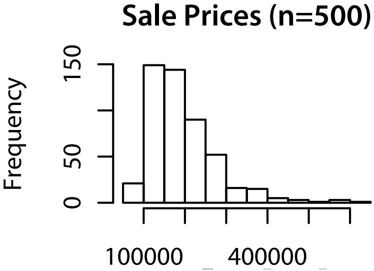
Step 5: If $P < \alpha$, reject H_0. Otherwise, fail to reject H_0.

Step 6: Interpret the results in words related to your data.

Some samples for examples…

```r
library(AmesHousing)
library(tidyverse)
# Use only houses built at or after 1950 (why later)
ames <- make_ordinal_ames()
ames.1950 <- ames %>% filter(Year_Built >= 1950)
set.seed(20404)   # For repeatability.
ames.50 <- sample_n(ames.1950, 50)
ames.500 <- sample_n(ames.1950, 500)
```


Sale Prices (n=50)

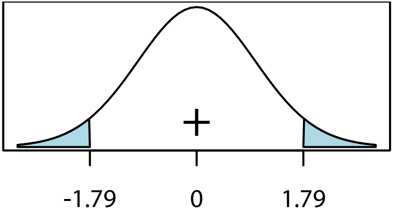

Sale Prices (n=500)

Is the mean sale price different than $200,000?

$H_0: \mu = 200000, H_1: \mu \neq 200000.$

We'll use $\alpha = 0.05$.

$$t = \frac{183084.1 - 200000}{66892.08/\sqrt{50}} = -1.788156.$$



```
mean(ames.50$Sale_Price)

## [1] 183084.1

sd(ames.50$Sale_Price)

## [1] 66892.08

# Calculate the p-value
pt(q = -1.788156, df=49) +
  (1-pt(q = 1.788156, df=49))

## [1] 0.0799349
```

Is the mean sale price different than $200,000?

```
t.test(x=ames.50$Sale_Price, mu=200000, alternative="two.sided")

##
## ^^IOne Sample t-test
##
## data:  ames.50$Sale_Price
## t = -1.7882, df = 49, p-value = 0.07993
## alternative hypothesis: true mean is not equal to 200000
## 95 percent confidence interval:
##  164073.6 202094.6
## sample estimates:
## mean of x
##  183084.1
```

Is the mean sale price different than $200,000?

- If we believe the population mean sale price was $200,000, there would be a 7.99% chance to see data as extreme as ours.

- Since $p = 0.0799 > 0.05$, we don't reject H_0.

- **A yes/no answer based on the chosen α:**
  "There is not enough evidence at the 5% significance level to conclude that the mean sale price of all houses in Ames is different than $200,000."

- **Interpret the p-value more directly:**
  "There may be weak evidence ($p = 0.0799$) to conclude that the mean sale price of all houses in Ames is different than $200,000."

Confidence intervals give an estimate for unknown population parameters.

- **Interpretation:** We can say with 95% confidence that the mean sale price of all homes in Ames is between $164073.58 and $202094.62.
- **Meaning:** For 95% of all possible random samples, this method will give an interval that contains the true population mean.
- Since the confidence interval overlaps $200,000, we can't conclude that the mean sale price is different than $200,000.

Question: Replace 50 with 500

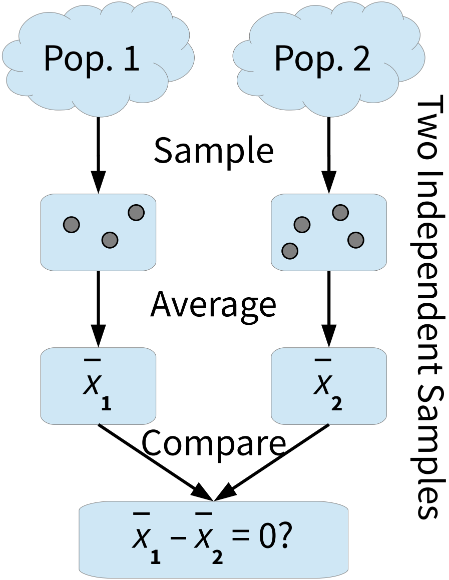
```
t.test(x=ames.500$Sale_Price, mu=200000, alternative="two.sided")

##
## ^^IOne Sample t-test
##
## data:  ames.500$Sale_Price
## t = -0.89389, df = 499, p-value = 0.3718
## alternative hypothesis: true mean is not equal to 200000
## 95 percent confidence interval:
##  189447.3 203953.0
## sample estimates:
## mean of x
##  196700.2
```

Question: p-values

A test of population *means* is used to compare values of a numeric variable on different populations.

- **Question:**
  Is there sufficient evidence to conclude that the mean for Population 1 (μ_1) is different from the mean for Population 2 (μ_2)?
- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 - \mu_2 < 0$ (left-tailed)
  $H_1 : \mu_1 - \mu_2 > 0$ (right-tailed)
  $H_1 : \mu_1 - \mu_2 \neq 0$ (two-tailed)

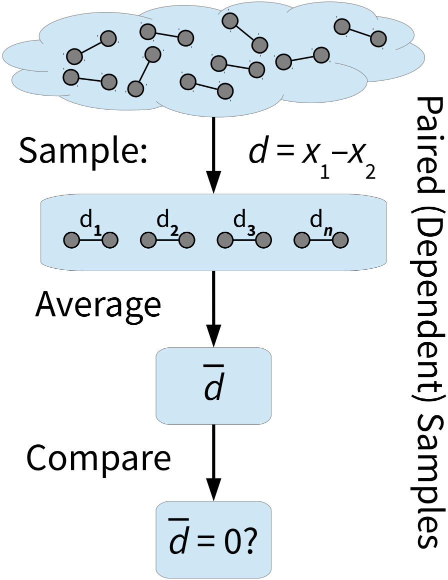
Indep. Samples: Compare Sample Means



- Blood pressures are compared between two groups. The control received a placebo, and the treatment groupo received a new drug.
- Web site visitors are randomly shown one of two versions of a news story. Does time spent on the page differ?

Paired Samples: Compare Individuals

- Pediatric patient temperatures are measured under the tongue and under the arm. Are the under-arm temperatures lower?
- Two types of reflective paint are tested on highway lane markings in various parts of the state. Do they wear differently?



Sample: $d = x_1 - x_2$

d_1 d_2 d_3 d_n

Average

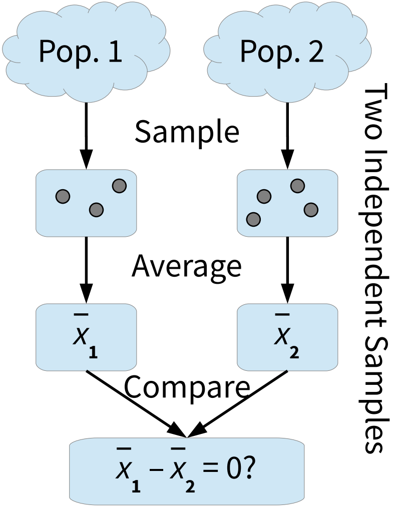
$\bar{d}$

Compare

$\bar{d} = 0?$

Paired (Dependent) Samples

Where possible, pairing gives a more powerful test.



Pop. 1

Pop. 2

Sample

Average

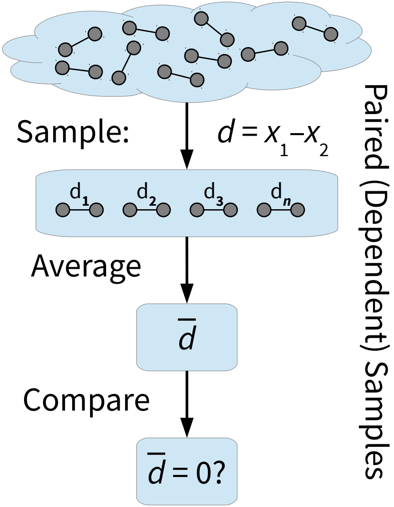
$\bar{x}_1$

$\bar{x}_2$

Compare

$\bar{x}_1 - \bar{x}_2 = 0?$

Two Independent Samples



Sample:

$d = x_1 - x_2$

d_1 d_2 d_3 d_n

Average

$\bar{d}$

Compare

$\bar{d} = 0?$

Paired (Dependent) Samples

Two Types of Tests

- Independent Samples
  - Two samples from different populations.
  - Choice of one sample not related to the choice of the other.

Two Types of Tests

- Independent Samples
  - Two samples from different populations.
  - Choice of one sample not related to the choice of the other.
- Paired (Dependent) Samples
  - There is a natural, one-to-one pairing of individuals in each population.
  - Pairs are selected randomly.
  - When pairing makes sense, it increases the power of the test. Differences aren't washed out by averaging.

Question? Which samples are independent and which are paired?

- Do the students in a social studies class perform better, on average, on the unit pre-test or the unit post-test?
- Students are assigned randomly to one of two classes. Each class is given a different curriculum. Did one class perform better on average?

Question? Which samples are independent and which are paired?

- Do men or women more efficiently metabolize alcohol? Subjects are asked to consume an alcoholic drink, and the alcohol concentration in their blood is measured two hours later.
- In order to determine which of two brands of special reflective paint wears better on roads, state highway officials decide to paint stripes with each brand of paint at several different locations and measure the length of time before the paint loses its reflectivity.

The Independent-Sample (Welch's) *t*-Test

Step 1: H_0: $\mu_1 = \mu_2$, and H_1: $\mu_1 \neq \mu_2$, $\mu_1 > \mu_2$, or $\mu_1 < \mu_2$.

Step 2: Select α. (prior to analysis)

Step 3: Criteria: independent random samples, each

- from a normal population, **OR**
- with sample size ≥ 30.

Step 4: Test statistic: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$.

Step 5: Compute *p*-value using a *t*-distribution with

- df $= \min\left(n_1 - 1, n_2 - 1\right)$ (by hand), or
- df $= \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ (by computer)

Step 6: If $P < \alpha$, reject H_0. Otherwise, fail to reject H_0.

Step 7: Interpret the results in words.

Another Solution: The Pooled Sample Standard Deviation.

- If we can assume $\sigma_1 = \sigma_2 = \sigma$, we can get a more powerful test.
- We *pool* s_1 and s_2 to approximate σ:
$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$
- $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}}.$
- $\text{df} = n_1 + n_2 - 2.$

Does remodelling make a difference in sale price?

```
ames.50$Remodeled <-
  ames.50$Year_Built != ames.50$Year_Remod_Add
t.test(Sale_Price ~ Remodeled, data=ames.50)

##
## ^^IWelch Two Sample t-test
##
## data:  Sale_Price by Remodeled
## t = -1.6996, df = 24.773, p-value = 0.1017
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -73570.039   7061.369
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            174438.0            207692.3
```

Does remodelling make a difference in sale price?

```
ames.500$Remodeled <-
  ames.500$Year_Built != ames.500$Year_Remod_Add
t.test(Sale_Price ~ Remodeled, data=ames.500)

##
## ^^IWelch Two Sample t-test
##
## data:  Sale_Price by Remodeled
## t = -3.9202, df = 214.87, p-value = 0.000119
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -53129.57 -17578.00
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            186235.4            221589.2
```

Does remodelling make a difference in sale price?

```
ames.1950$Remodeled <-
  ames.1950$Year_Built != ames.1950$Year_Remod_Add
t.test(Sale_Price ~ Remodeled, data=ames.1950)
```

```
##
## ^^IWelch Two Sample t-test
##
## data:  Sale_Price by Remodeled
## t = -9.1247, df = 1238.2, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -41610.48 -26883.75
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            184312.8            218559.9
```

Conclusions in words.

From our sample of 50 homes:

> *There is weak to no evidence* $(p = .1017)$ *that mean sale price of remodeled homes is different from that of unremodeled homes.*

From our sample of 500 homes:

> *There is very strong evidence* $(p = .000119)$ *the mean sale price of remodeled homes is different from that of unremodeled homes.*

For *all* home sales, what does $p = .0000000000000022$ mean?

Hypothesis Tests on Paired Samples

- Compute $d = x_1 - x_2$ for each data point.
  (Note the order of the subtraction and what the sign of d means for your problem.)
- Hypotheses
  - $H_0 : \mu_d = 0.$
  - $H_1 : \mu_d \neq 0, \mu_d < 0,$ or $\mu_d > 0.$
- Criteria
  - Sample data are matched pairs.
  - The population of differences (d's) is normally distributed (without outliers), **or** the # of *pairs* is large ($n \geq 30$).
- Do one-variable t-test on the new variable d.

Hypothesis Tests on Paired Samples

Step 0: Compute $d = x_1 - x_2$ for each data point.

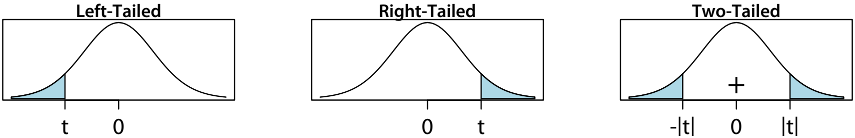
Step 1: H_0: $\mu_d = 0$, and H_1: $\mu_d \neq 0$, $\mu_d < 0$ or $\mu_d > 0$.

Step 2: Select α. (prior to analysis)

Step 3: Num. *pairs* large ($n \geq 30$), or d's approx. normal?

Step 4: Compute $t = \dfrac{\bar{d} - 0}{s/\sqrt{n}}$.

Step 5: Compute p-value (t-dist. with df $= n - 1$).



Step 6: If $P < \alpha$, reject H_0. Otherwise, fail to reject H_0.

Step 7: Interpret the results in words.

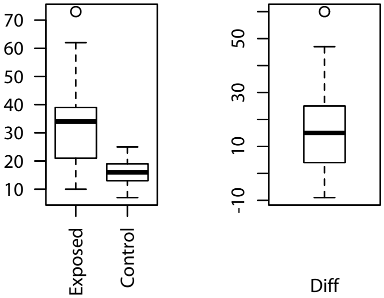
Example: Paired Blood Lead Level

Children whose parents worked in a lead-related factory were paired with ”control” children from the same neighborhoods, and blood lead level was measured.

```r
library(PairedData)
data(BloodLead)
BloodLead$Diff <-
  BloodLead$Exposed - BloodLead$Control
head(BloodLead)
```

```
##   Pair Exposed Control Diff
## 1  P01      38      16   22
## 2  P02      23      18    5
## 3  P03      41      18   23
## 4  P04      18      24   -6
## 5  P05      37      19   18
## 6  P06      36      11   25
```

```r
boxplot(BloodLead[,2:3], las=2)
boxplot(BloodLead$Diff, xlab="Diff")
```



Is mean blood level higher for children with parents in lead-related factories?

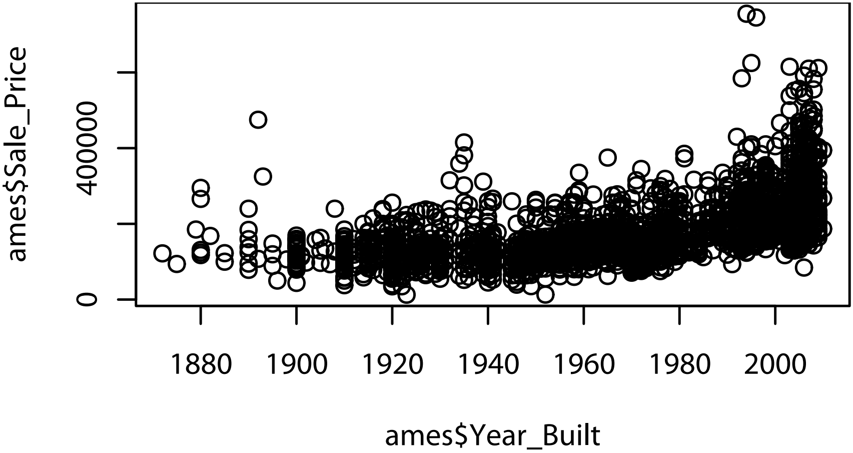
```
t.test(x=BloodLead$Exposed, y=BloodLead$Control,
  alternative="greater", paired=TRUE)
```

```
##
## ^^IPaired t-test
##
## data:  BloodLead$Exposed and BloodLead$Control
## t = 5.783, df = 32, p-value = 0.000001018
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  11.29201      Inf
## sample estimates:
## mean of the differences
##                 15.9697
```

“There is strong evidence (p = 0.000001) that children whose parents word in lead-related factory have high blood lead levels, on average.”

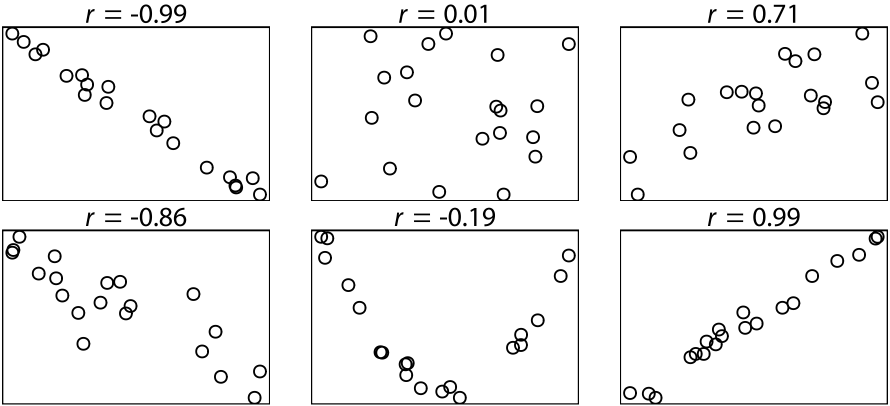
Is sale price related to the year built?

```
plot(x=ames$Year_Built, y=ames$Sale_Price)
```



The *correlation coefficient* measures the *linear* relationship between two variables.

The correlation coefficient *r* is standardized between -1 and 1:



The **cor** function calculates correlation.

```
# Correlation between two vectors.
cor(ames$Year_Built, ames$Sale_Price)

## [1] 0.5584261

# Correlation between all columns of data frame or matrix.
# Make a demo data frame with three columns:
ames.temp <- ames[, c("Year_Built", "Year_Remod_Add", "Sale_Price")]
# Calculate pairwise correlation between all three variables:
cor(ames.temp)

##                Year_Built Year_Remod_Add Sale_Price
## Year_Built      1.0000000      0.6120953  0.5584261
## Year_Remod_Add  0.6120953      1.0000000  0.5329738
## Sale_Price      0.5584261      0.5329738  1.0000000
```

Simple linear regression finds the equation of the line that best fits the data.

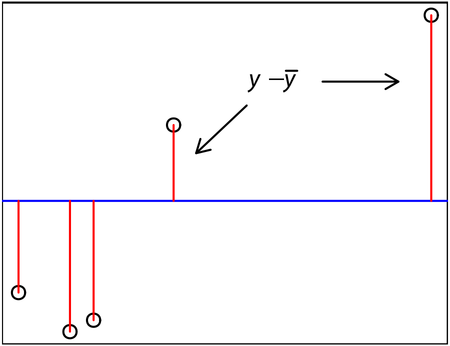
- We assume a theoretical model:

Y	$=$	β_1	x	$+$	β_0	$+$	ε
Response Variable		Slope	Explanatory Variable		Intercept		Random “Error”

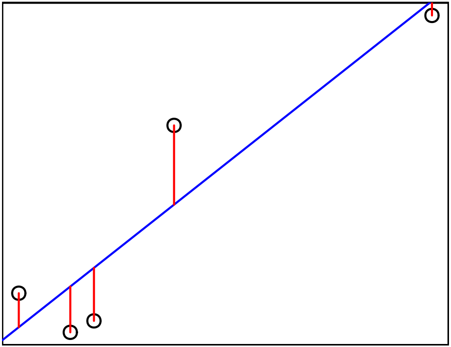
- Obtain estimates from the data: $\hat{y} = b_1 x + b_0$.
- Coefficients b_1 and b_0 chosen to minimize the sum of squared residuals: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

The Idea: Residuals are smaller when the line fits the data well.

Bid fit: large residuals.



Better fit: smaller residuals.



Minimizing $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ ensures all residuals relatively small.

The `lm` command does linear regression in **R**.

```
fit.sale.year <- lm(Sale_Price ~ Year_Built, data=ames)
summary(fit.sale.year)
```

```
##
## Call:
## lm(formula = Sale_Price ~ Year_Built, data = ames)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -147394  -41227  -14502   23093  540805
##
## Coefficients:
##                 Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -2726883.30    79834.60  -34.16   <2e-16 ***
## Year_Built      1474.96       40.49   36.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66280 on 2928 degrees of freedom
## Multiple R-squared:  0.3118,^^IAdjusted R-squared:  0.3116
## F-statistic:  1327 on 1 and 2928 DF,  p-value: < 2.2e-16
```

Coefficient estimates give the regression equation.

```
##                 Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) -2726883.301 79834.60077 -34.15666 1.626675e-215
## Year_Built      1474.964    40.49253  36.42558 6.293708e-240
```

The `Estimate` column gives the slope and intercept for the equation.

$$\widehat{\texttt{Sale\_Price}} = 1474.964\,\texttt{Year\_Built} - 2726883.301$$

The Pr column tests whether coefficients are statistically significantly different from zero.

```
##                  Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) -2726883.301 79834.60077 -34.15666 1.626675e-215
## Year_Built      1474.964    40.49253  36.42558 6.293708e-240
```

- H_0: $\beta_1 = 0$. *The slope of the regression line is equal to zero.*
- H_1: $\beta_1 \neq 0$. *The slope of the regression line is not equal to zero.*
- p-value: $p = 6.2937078 \times 10^{-240} < 0.05$. *Reject H_0!*
- Conclusion in words: (**Note:** Check assumptions first–see below.) There is sufficient evidence that the slope of the regression line between `Year_Built` and `Sale_Price` is different from zero. In other words, there is sufficient evidence that there is at least *some* linear relationship between year built and sale price.

The R^2 value indicates how well the model predicts Y.

```
## Multiple R-squared: 0.3118  Adjusted R-squared: 0.3116
```

- The R^2 value indicates the percentage of variation in Y that's explained by its linear relationship to x.
- "Approximately 31.18% of the variation in `Sale_Price` is explained by its relationship to `Year_Built`."
- You'll see more about adjusted R^2 later.

Steps in Simple Linear Regression

1. Make a graph. Does a linear fit seem appropriate?
2. Calculate the regression model.
3. Evaluate regression diagnostics—is the model valid? (This is a more formal version of step 1.)
4. Use the model!
    - Interpret coefficients.
    - Make predictions.
    - Examine points with interesting residuals.

Using our home price equation…

$$\hat{y} = 1474.964x - 2726883.30.$$

- Interpret the slope: the change in y when x changes by one unit.

  "For every year newer a house is, we expect its sale price to go up by 1474.96, on average."

- Plug in x to make predictions.

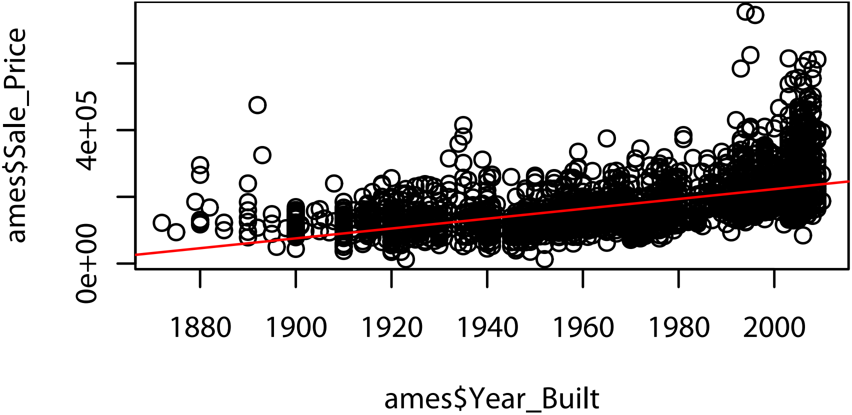
  $$\hat{y} = 1474.964(1971) - 2726883.30 \approx 180270.74.$$

  "A home built in 1971 has an expected same price of $180,270.74."

- "The highest-selling home built before 1900 had been renovated in 1992 with overall quality listed as 'very excellent…'"

A graph of the regression line and data.

```
plot(x=ames$Year_Built, y=ames$Sale_Price)
abline(fit.sale.year, col="red")
```



Warning: Not a great fit!

A valid regression model satisfies four assumptions (stated informally here):

Linearity The relationship between x and the expected value of Y is linear.

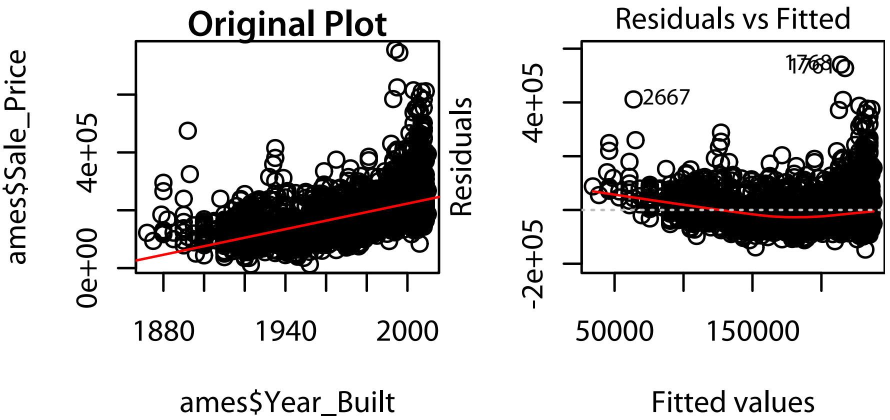
Constant Variance The variability of residuals is the same near one x value as it is near any other x value. (Homoscedasticity)

Normality Near any fixed x, residuals are normally distributed. [Note: Becomes less important for large samples. One source suggests $n > 30 + 20(\#\text{ exp. vars})$.]

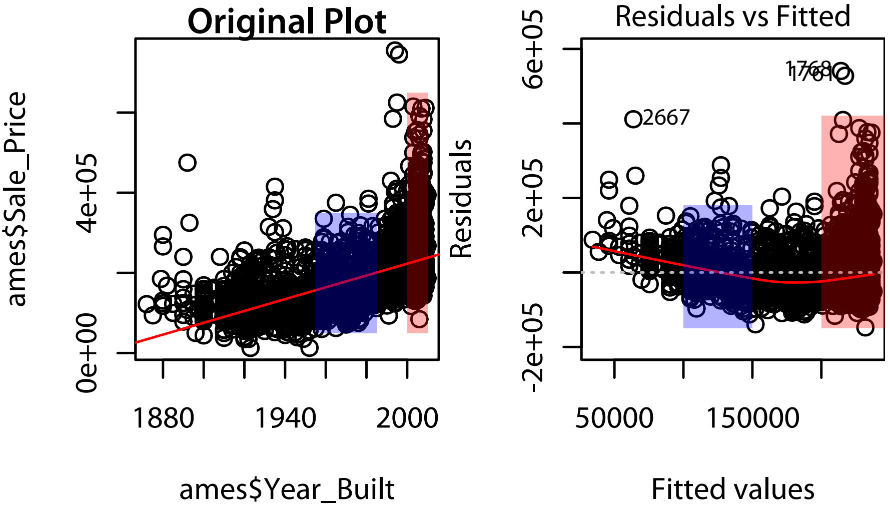
Independence Residuals are independent of one another.

Sale prices show non-linearity.

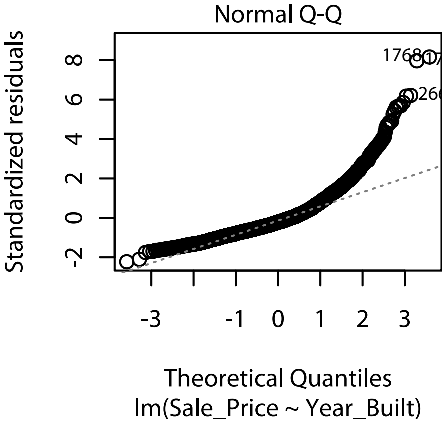
- The regression line is not centered within the cloud of data points.
- The residual plot is not centered on the dotted zero line.



Sale prices show non-constant variance.



The Normal Q-Q plot compares residuals to an ideal normal distribution.

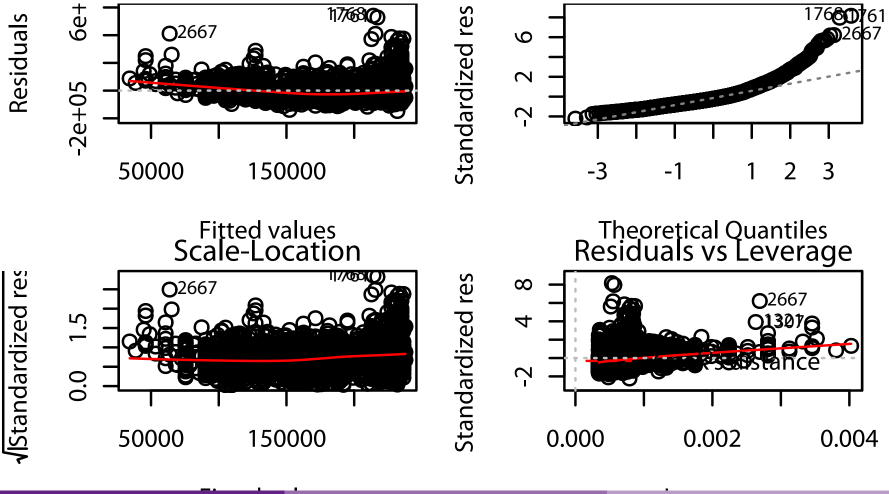


If residuals are normally distributed, the plot of residuals vs. theoretical quantiles will appear linear.

This plot shows evidence of non-normal residuals!

The `plot` command generates diagnostics.

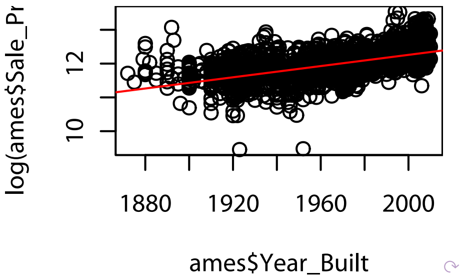
```
par(mfrow=c(2,2), mar=c(5,4,0,1))
plot(fit.sale.year)
```



Sometimes a log transformation of *Y* creates a better model.

- Model: $\log(Y) = \beta_1 x + \beta_0 + \varepsilon$
- The logarithm function “pulls in” extreme data points.
- For home prices, we could argue directly that we want a model that isn’t unduely influenced by extremely expensive homes.

```r
fit.log <-
  lm(log(Sale_Price) ~ Year_Built,
  data=ames)
plot(x = ames$Year_Built,
  y=log(ames$Sale_Price))
abline(fit.log, col="red")
```



Multiple linear regression uses more than one explanatory variable to predict Y.

- The model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$.
- Let's predict price using both square footage and year built.

```
fit.sale.sqft.year <- lm(Sale_Price ~ First_Flr_SF + Year_Built, data=ames)
summary(fit.sale.sqft.year)$coefficients
```

```
##                 Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)  -2042141.3692 68194.691266 -29.94575 4.165985e-172
## First_Flr_SF      101.1349     2.705082  37.38700 1.944663e-250
## Year_Built       1068.1304    35.049905  30.47456 2.114008e-177
```

```
summary(fit.sale.sqft.year)$adj.r.squared
```

```
## [1] 0.5339375
```

```
summary(fit.sale.sqft.year)
```

```
##
## Call:
## lm(formula = Sale_Price ~ First_Flr_SF + Year_Built, data = ames)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -434097  -32813   -9218   23910  420116
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.042e+06  6.819e+04  -29.95   <2e-16 ***
## First_Flr_SF  1.011e+02  2.705e+00   37.39   <2e-16 ***
## Year_Built    1.068e+03  3.505e+01   30.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54540 on 2927 degrees of freedom
## Multiple R-squared:  0.5343,^^IAdjusted R-squared:  0.5339
## F-statistic:  1679 on 2 and 2927 DF,  p-value: < 2.2e-16
```

Logistic regression uses one or more explanatory variables to predict the likelihood of a Yes/No categorical variable.

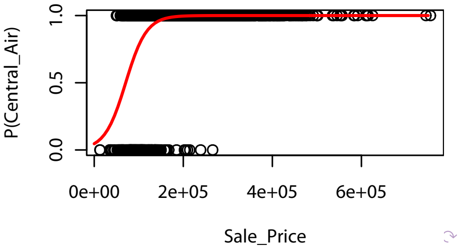
- Suppose p represents the probability that a house has central air conditioning.
- Perhaps houses with a high sale price have a higher likelihood of central air.
- Logistic regression models the *log odds* as a linear function of one or more variables: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + + \cdots + \varepsilon, \; or \; p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\varepsilon)}}$

Probability of central air increases with price.

- The `glm` command stands for “generalized linear model,” which is a type of model that includes logistic regression.
- The linear function $\beta_0 + \beta_1 x_1 + \dots$ is connected to the response via a “link” function.
- Specifying `family=binomial` has default `link=logit`.

```
fit.ac.logistic <-
  glm(Central_Air ~ Sale_Price,
      data=ames,
      family=binomial)
coef(fit.ac.logistic)

##   (Intercept)    Sale_Price
## -3.025133e+00  4.280822e-05
```

$$p = \frac{1}{1 + e^{-(-3.025 + 0.00004281 \text{Sale\_Price})}}$$



More predictive modeling in PDAT 613…

- Multiple Linear Regression:
  - More on variabile significance.
  - Comparing between models.
  - Categorical (indicator) variables as predictors.
- Logistic Regression:
  - More on logistic regression and categorization.
  - Interpretation of elements of the model itself.
- Other predictive modeling techniques.
- Techniques like *cross-validation* to evaluate how well a model would predict *Y*, not only for existing data, but also for new data that might be collected in the future.