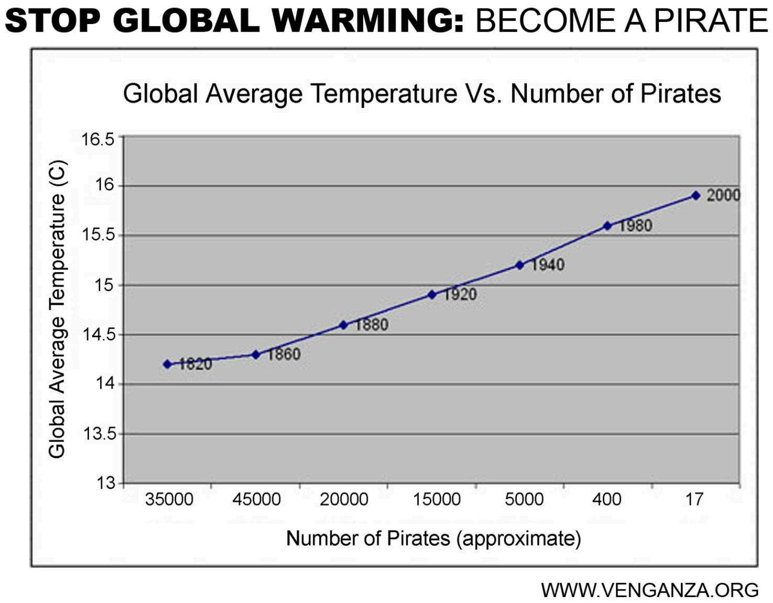
Experimental Design and One-Way ANOVA

36-600

Experiments and Causation

- Throughout this course we have tried to uncover the association between a set of predictor variables $\mathbf{x}$ and a response variable Y, if it exists
- However, even if a statistically significant association does exist…
  - **…association does not imply causation!**



(image credit)

Case Study 1: Industrial Experiments

(The following case studies and the material on experimental design were provided by Professor Zach Branson)

- Apple wants to test the water durability of their laptops
    - they randomly sample 100 identical laptops for study, and pour water on half of them
    - 20 of 50 "treatment" (water-doused) keyboards continued to work, as opposed to 50 of 50 "control" keyboards
- Did the water *cause* the keyboards to break?

- **Yes**: the laptops were otherwise identical...the only difference was the treatment

Case Study 2: Clinical Trials

- The Food and Drug Administration wanted to determine whether a new drug alleviated hypertension
  - they randomly picked 100 people with hypertension...
  - ...and placed 50 people each into the treatment and control groups
  - 30 of 50 treated people had alleviated hypertension, as opposed to 10 of 50 in the control group
  - a two-sample population proportions test yielded a p-value of 0.0001
- Did the new drug *cause* alleviated hypertension?

- **We cannot be sure.**
- Let's say it turns out that, totally by accident, the 50 people in the treatment group had health insurance, while the 50 people in the control group did not
  - so perhaps it was the drug, or the insurance (or the fact that those without insurance were poorer, or...)
  - randomization leads to identical treatment and control groups, but only *on average*..."unlucky" randomization can happen

Case Study 3: Epidemiology

- We wish to study the effect of smoking on lung health
- Ideally, we would run an experiment in which we randomly place people into a smoking treatment group, and into a non-smoking control group
    - however, it is not ethical to force people to smoke
- So we randomly select 5000 smokers and 5000 non-smokers to study, and we find that the smokers have worse lung health
- Does smoking *cause* the observed deterioration of lung function?

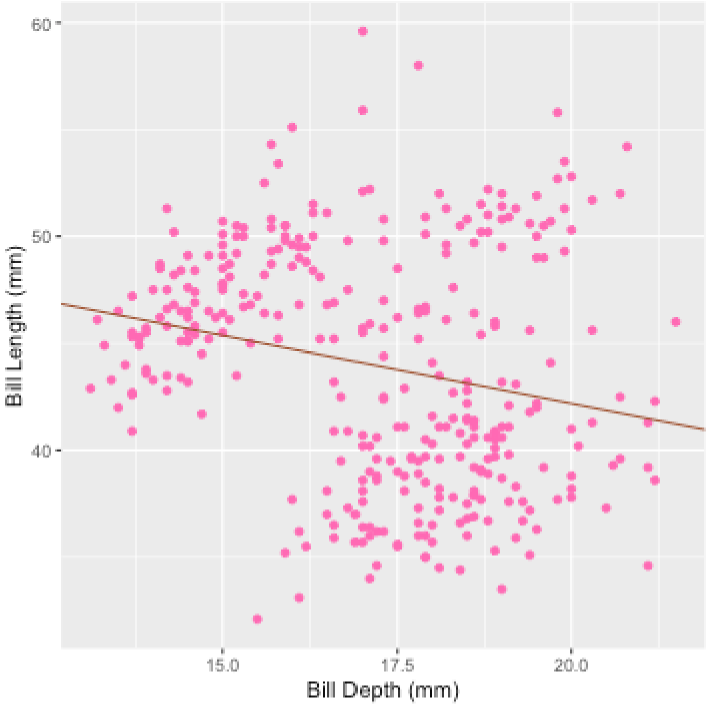
- **Again, we cannot be sure.**
- For instance, smokers tend to be older, poorer, and to not have insurance
- Can we mitigate this issue?

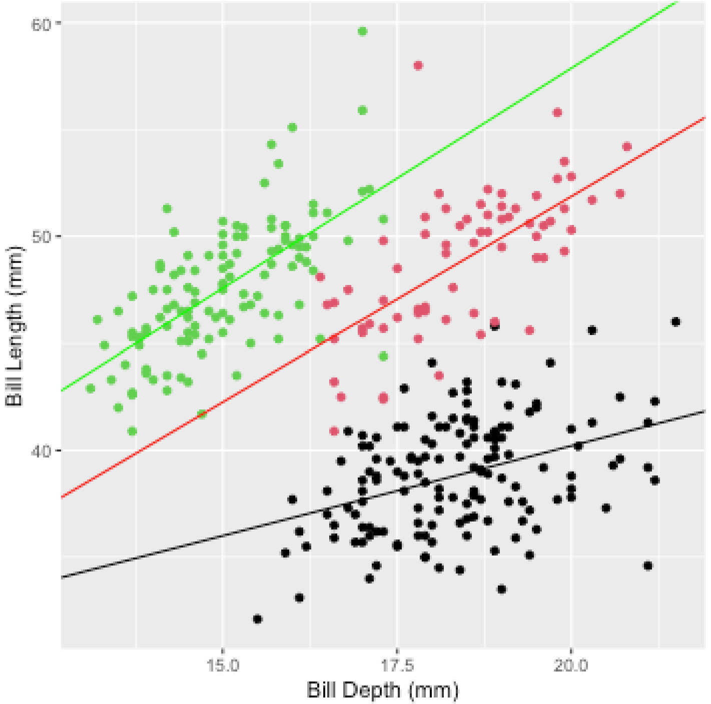
$\rightarrow$ if we can identify subsets of treatment and control groups that have similar age, income, education, insurance, etc., and find similar results as before, we are in a position to argue more strenuously for causality

Experiment Design

- Let's assume that we want to divide N people into two groups, a treatment group and a control group...(assume N is an even number)
- How might we do this?
- Via *Bernoulli trials*: assign people to groups effectively via coin flips
  - issue: the group sizes can end up being very different
- Via *complete randomization*: pick exactly $N/2$ people at random to be in the treatment group
  - issue: while this resolves the group-size issue observed with Bernoulli trials, there is still the issue of covariates
  - for instance, perhaps we want to run a clinical trial that includes both smokers and non-smokers: it could turn out that one group ends up with many more smokers than the other
  - the "covariate issue" can be dealt with when analyzing the data...or during the experiment design stage
- Via *block randomization*: identify a potentially problematic covariate, divide people into groups on the basis of that covariate (e.g., smokers vs. non-smokers), and then perform complete randomization within each covariate group...this scheme can be extended to multiple covariates: e.g., male smokers, female non-smokers, etc.
  - covariate examples: gender, socioeconomic status, geographic location, medical risk factors, education, etc.

Digression: Why Is Identifying Covariates Important?





How to Analyze Designed Experiment Data

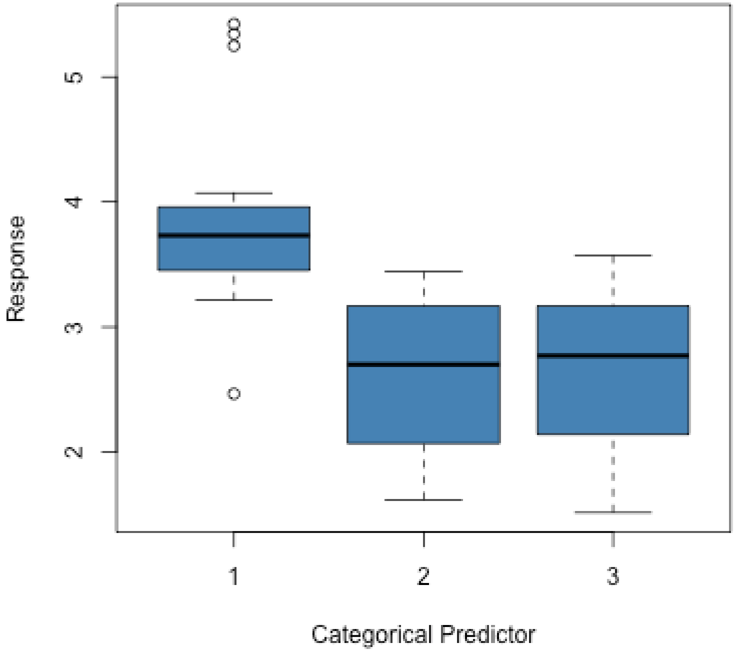
- Let's assume that our data consists of a predictor variable with k categories (or treatment groups) and a response variable.
- If $k = 2$ and the response variable is normally distributed within each group, we can do a two-sample t test
- If $k = 2$ and the response variable is not normally distributed, but the distributions are known or can be assumed, we can utilize other hypothesis tests like the population proportions test
    - this is realm of so-called A/B testing…the test that is used depends on the distribution of the response values for groups A and B
- If $k > 2$ and the response variable is normally distributed within each group (with equal variances), we can do one-way analysis of variance (or ANOVA)
    - if there are two categorical predictors, we can do a two-way ANOVA
- If $k > 2$ and the response variable is normally distributed within each group (with equal variances), *and* there is another continuous predictor that can be treated as a covariate (e.g., age), we can do (one-way) analysis of covariance (or ANCOVA)
- Etc.
- **NOTE:** if your work potentially involves the design of experiments, consider taking 36-749, *Experimental Design for Behavioral and Social Sciences* (offered every fall)

The One-Way ANOVA Setting

- Recall that a statistical model is a description of the data-generating process
- In the simple linear regression setting, our data consists of a continuously valued predictor variable $\mathbf{x}$ that we attempt to relate to the values of response variable $\mathbf{Y}$
  - however, what if instead the values of the predictor variable are discretely valued…and specifically, what if they represent groups (or categories)?
- If there are $k > 2$ groups we would utilize one-way *analysis of variance* (or one-way *ANOVA*)
  - "one-way" simply indicates that (in our chosen setting) there is only one (categorical) predictor variable

Why Not Use Simple Linear Regression?

- Because categories/groups may have no natural numerical order
- If we were to apply linear regression alone, then switch the placement of groups 1 and 2 and apply linear regression again, the slope would change!



Why Not Use Simple Linear Regression?

- Let's define the predictor variable `x` as a *factor variable*
- This causes `R` to change the definition of the model...
    - a factor variable with k levels is split into $k-1$ so-called *dummy variables*; the other level becomes the so-called *reference level*
- The linear regression model becomes

$$Y_i = \beta_0 + \mathcal{I}_{x_i=2}\beta_2 + \mathcal{I}_{x_i=3}\beta_3 + \epsilon_i$$

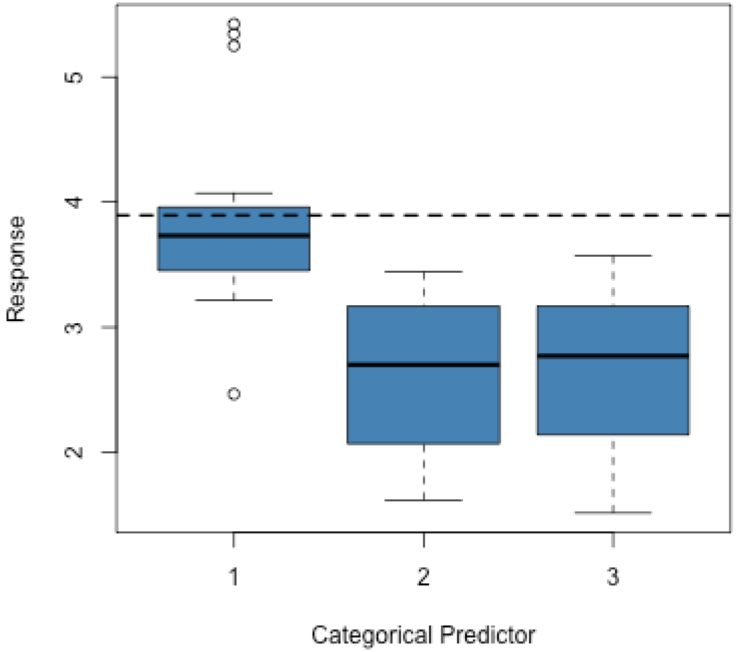
- $\mathcal{I}$ is the *indicator function*, and it takes on value 1 if the condition is true and 0 otherwise
    - for instance, $\mathcal{I}_{x_i=4}$ is 0 if $x_i = 3$ or 5 and 1 if $x_i = 4$
- We can rewrite the model in a form that might be more intuitive:

$$Y_i = \begin{cases} \beta_0 & x_i = 1 \\ \beta_0 + \beta_2 & x_i = 2 \\ \beta_0 + \beta_3 & x_i = 3 \end{cases}$$

- If we change the ordering of the groups, the β_i's might change, but only because we perhaps define a new reference level

Why Not Use Simple Linear Regression?

- The dashed line is β_0 and represents the predicted response for group 1
  - the negative coefficients for `x2` and `x3` in the `lm()` output indicate that the model is predicting that the means in groups 2 and 3 are smaller than the mean in group 1
- So…why not use simple linear regression? It turns out, we *do* use it…but the model definition changes since `x` is categorical and not quantitative
- So what then is ANOVA?
  - it is simply a mechanism for running a hypothesis test on linear regression output



One-Way Analysis of Variance

- For one-way ANOVA, the statistical model is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

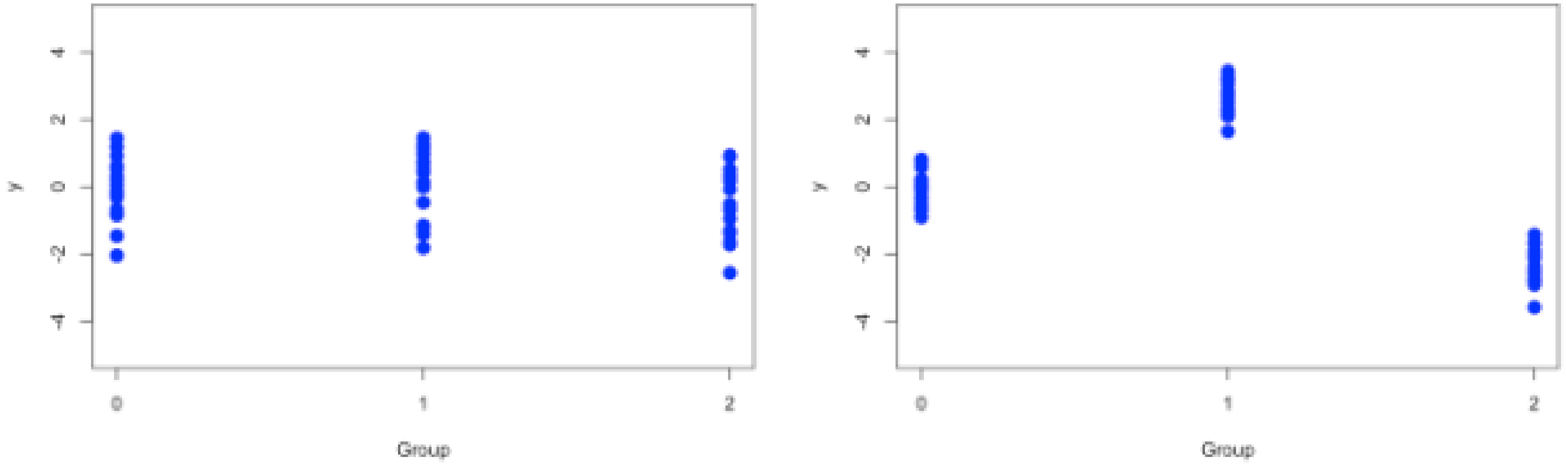
  - i denotes the treatment group (there are k groups overall)
  - j denotes an observed datum within group i
  - μ is the overall mean response
  - τ_i is the *deterministic* (i.e., not random) effect of treatment in group i
  - ϵ_{ij} are the error terms, assumed to be independent, normally distributed, and of constant variance σ^2...thus $Y_{ij} \sim \mathcal{N}(\mu + \tau_i, \sigma^2)$

ANOVA: Goal

- The goal of ANOVA is to perform the hypothesis test

$$H_o : \tau_1 = \cdots = \tau_k = 0 \quad \text{vs.} \quad \text{at least one value differs from zero}$$

  - to reject the null, the value of one or more of the τ_i's has to be large with respect to σ



- The left figure represents a situation in which we would fail to reject the null hypothesis, while the right figure represents a situation in which we would reject the null

ANOVA Example

```
anova(lm(Y~x))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          2 12.879  6.4395  11.116 0.0002609 ***
## Residuals 29 16.800  0.5793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- the first column shows $k-1$ and $n-k$ (so we can determine that $k=3$ and $n=32$)
- the second column shows the SST (top) and SSE (bottom) (see appendix)
- the third columns shows the MST (= SST /($k-1$); top) and MSE (= SSE /($n-k$); bottom) (again, see appendix)
- the fourth column shows $F = MST/MSE$
- the fifth and last column shows the p-value
- We observe that $p \ll \alpha = 0.05$, so we reject the null hypothesis and conclude that at least one of the means is different from the others

But Which Mean is Different?

- To try to determine *which* of the means is different from the others, we can use a "post-hoc" test such as the Tukey HSD (honest significant difference) test

```
TukeyHSD(aov(Y~x)) # one quirk: it won't work with anova() output, just aov() output
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Y ~ x)
##
## $x
##            diff        lwr        upr     p adj
## 2-1 -1.27593333 -2.0039372 -0.5479295 0.0004657
## 3-1 -1.26000000 -2.2306718 -0.2893282 0.0088999
## 3-2  0.01593333 -0.9846122  1.0164789 0.9991476
```

- the first column of output shows the groups being compared (2 vs. 1, etc.)
- the second column gives the observed mean difference
- the third and fourth columns provide confidence intervals on the *true* mean difference…if the interval does not contain zero, we conclude the means are different
- the last column reinforces the confidence interval by providing a p-value (where the null is that the true difference is zero)

But Which Mean is Different?

```
TukeyHSD(aov(Y~x)) # one quirk: it won't work with anova() output, just aov() output
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Y ~ x)
##
## $x
##            diff        lwr        upr     p adj
## 2-1 -1.27593333 -2.0039372 -0.5479295 0.0004657
## 3-1 -1.26000000 -2.2306718 -0.2893282 0.0088999
## 3-2  0.01593333 -0.9846122  1.0164789 0.9991476
```

- What do we conclude here?
    - group 1 has a different mean from group 2, and from group 3…while the means in groups 2 and 3 are not significantly different

But Which Mean is Different?

- A final point to make is that the Tukey HSD test attempts to correct for *multiple comparisons*, i.e., for running many separate hypothesis tests
- If you run many tests, then you are more likely, by chance, to see p-values that are less than α even if the null is always true
- The Tukey HSD test attempts to control the "family-wise error rate" such that if all the nulls are correct, the probability of seeing $p < \alpha$ occur *once* is α
    - *however*, the algorithm for controlling the rate of false positives tends to be overly conservative
    - we will simply point out here that alternative test schemes are available, e.g., Dunnett's test, that one might want to explore using, particularly when the number of groups is large

Appendix: Sum of Squares of Errors and of Treatment Groups

- We break the total sum of squared differences between each datum Y_{ij} and the overall mean $\bar{Y}$ into two pieces

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2+\sum_{i=1}^{k}n_i(\bar{Y}_{i\bullet}-\bar{Y})^2=SSE+SST$$

  - $\bar{Y}_{i\bullet}$ is the sample mean of the data of group i
  - n_i is the sample size in group i
  - SSE is the sum of squares of the errors (where the "error" is how far each datum is from its group mean)
  - SST is the sum of squares for each treatment group (how far each group mean is from the overall mean)

Appendix: Hypothesis Testing

- We can form two statistics:

$$\frac{SSE}{\sigma^2} \quad \text{and} \quad \frac{SST}{\sigma^2}$$

- Under the null hypothesis that $\tau_1 = \cdots = \tau_k = 0$, we can write that

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-k} \quad \text{and} \quad \frac{SST}{\sigma^2} \sim \chi^2_{k-1}$$

  - χ^2_ν is a chi-square distribution for ν degrees of freedom

- The following ratio defines a random variable that is sampled from an F distribution:

$$\frac{SST/(k-1)}{SSE/(n-k)} = \frac{MST}{MSE} = F \sim F_{k-1,n-k}$$

  - $k-1$ and $n-k$ are the number of *numerator* and *denominator* degrees of freedom, respectively

- We reject the null hypothesis if the value of F is (far) larger than its mean value, $(n-k)/(n-k-2)$ (which is ≈ 1 if $n \gg k$)