Study design
What to sample
where, when, and how many

Well collected data can make inference simple and poorly collected data can make inference impossible.
Question:

- What is the overarching idea?
- What is the specific question you want to answer?
Metric:

• What exactly will you measure and how?

• E.g. Happiness, droughts, intelligence, speed, water depth, ...?
What depth?
Where, when, how?
Where, when, how?
**Target Population**

- What places or populations do you want to make a conclusion about.
- Sample frame = All of the things you can actually sample (e.g. the target population of UW students except for those on leave this term).
Sample Design

• How to select a sample from your population (sample frame)?
You want your sample to be representative of the larger target population. One way to do this is to use random sampling where everyone has the same probability of being included in the sample.
The urn contains all of people in your sample frame, and now you choose a random sample without replacement (just like the bags of marbles in class).
Confounding

- Can you answer your question with the data?
In November 2007, the container ship *Cosco Busan* released 54,000 gallons of bunker fuel oil into San Francisco Bay. The accident oiled shoreline near spawning habitats for the largest population of Pacific herring on the west coast of the continental United States. We assessed the health and viability of herring embryos from oiled and unoiled locations that were either deposited by natural spawning or
Conclusions?
Other Examples

• People who eat salmon tend to live longer.
• Salmon populations in the Columbian declined during the period of major dam construction.
• The field where fertilizer was applied had plants than the field where fertilizer was not applied.
What can we do?

• **Careful interpretation**
  – Correlation ≠ Causation!
  – Oil spill example. These sites are different but can’t say anything about other sites.

• **Experiment** vs Observation study
  – You choose who or what get’s the treatment.
  – Vitamin E example and heart disease.
  – Hormone replacement and heart disease.
  – But...
Experiments are not always possible

- Effects of CO2 on global warming.
- Smoking and lung cancer.
- US tax policy and economic growth.
- Effects of large oil spills on Fish.
- Affects of urbanization in Puget Sound on salmon populations.
Sample Size:

- Power analysis
- Homework example.
Evaluate the effectiveness of a new teaching approach
Evaluate the effectiveness of a new teaching approach

```r
p_values <- numeric(10000)
for(i in 1:10000){
  groupA <- rnorm(20,4.16,1.31)
  groupB <- rnorm(20,5.16,1.31)
  p_values[i] <- t.test(groupA,groupB)$p.value
}

> mean(pilotDat)
[1] 4.164

Effect size = 1
> sd(pilotDat)
[1] 1.312675
```
Evaluate the effectiveness of a new teaching approach

```r
sampleSizes <- c(5,10,15,20,25,30,35,40,45,50)
power_vals <- numeric(10)
for(j in 1:10){
  sampSize <- sampleSizes[j]
p_values <- numeric(10000)
  for(i in 1:10000){
    groupA <- rnorm(sampSize,4.16,1.31)
    groupB <- rnorm(sampSize,5.16,1.31)
    p_values[i] <- t.test(groupA,groupB)$p.value
  }
  power_vals[j] <- length(p_values[abs(p_values)<=0.05])/10000
}
```
Evaluate the effectiveness of a new teaching approach

```r
> groupA <- rnorm(30,4.16,1.31)
> groupB <- rnorm(30,5.16,1.31)
```
Evaluate the effectiveness of a new teaching approach

> groupA <- rnorm(30,4.16,1.31)
> groupB <- rnorm(30,5.16,1.31)

> mean(groupB) - mean(groupA)
[1] 1.519446

> mean(groupB) - mean(groupA) + 2*sqrt((var(groupA)+var(groupB))/30)
[1] 2.309244

> mean(groupB) - mean(groupA) - 2*sqrt((var(groupA)+var(groupB))/30)
[1] 0.7296479
Examples from homework turned in.

• (0.52, 1.48)
• (0.50, 1.23)
• (-0.26, 1.72)
• (0.20, 1.46)

• Notice: we expect about 80% of the confidence intervals to not include 0.

• Or, equivalently, we expect to reject the null hypothesis (that there is not difference) about 80% of the time (assuming the true difference is 1, our estimate of stdev is about right, and we reject when a t-test produces a p-value < 0.05).
Central Limit Theorem

Here is a very skewed distribution. So a sample from this distribution will tend to have a lot of small values with a few very large values.
The central limit theorem says that if you take many different samples and calculate the mean for each, then the distribution of the means will be Normal as the sample size gets large. For very non-normal data, the sample size has to get pretty big (as is the case here). Notice the with a sample size of 5, the distribution of means still don’t look particularly normal (bell shaped).
Central Limit Theorem

As we increase the sample size the distribution starts to look more and more normal.
Central Limit Theorem

sample size = 25
Central Limit Theorem

sample size = 50
Central Limit Theorm

Here for a sample size of 100 the distribution looks very normal. Also notice that the variability of the means is much smaller (much less spread in the distribution).
The standard deviation of these means is called the standard error (se). An estimate of this is seen above in the equation. If you are looking at the difference between two means, the variance (i.e. standard deviation squared) is called the pooled variance. And the variance above is replaced with groupA_variance + groupB_variance. This, along with the fact the distribution is approximately normal (bell shaped) allows us to calculate the approximate confidence interval (sampleAvg ± 2*se).
End