Week 3: Displaying data and Exploratory Data Analysis
**Cauchy Distribution** – disease epidemics in which there is suddenly a big pulse of disease and then things calm down or foraging behavior with sudden shifts in location to gain large amounts of food.

\[ \text{population.cauchy} = \text{rcauchy}(1000, 0, 1) \]
I have no idea

population.odd = c(runif(1000, 0, 5), 1/runif(1000, 0, 1), rnorm(100, 0, 1))
Food Poisoning at a Festival

population.odd <- rgamma(1000, 1, .1)
• A summary statistic calculated from a sample of a population is, in and of itself, **random**.

• The **mean** of a sample is often a pretty good summary statistic – it’s a reasonable thing to do with your sample data. (Note that we could have done other things – max-min, top 5 minus mean of bottom five).

• The mean of a sample is more likely to be close to the mean of the population with bigger sample sizes. (**Accuracy**)  

• Particularly odd (extreme) estimates are more likely with smaller sample sizes (longer tails on the distribution of the mean). (**Precision**)  

• The distribution of the mean is generally **Gaussian** – even if the distribution of the population is not.  

• **We know something about a sample mean even when we don’t know anything about the population.** This is more than handy – it’s essential.

• For populations that are far from normal, the LCT still works in most cases. **Trouble arises with long, fat tails** in the population (intuitive once you think about it). How can you tell if your population has long fat tails? (1) previous knowledge or data (2) look at the sample data (not just the mean) – one of the many things we’ll cover in graphing / exploratory data analysis.

• Regression to the mean; Lack of independence; Sums of independent variables; other things beyond the mean
Regression to the Mean

• “On any task that contains both luck and skill, people who score above the mean are likely to have been luckier than people who score below the mean. Since luck does not hold from trial to trial, people who score above the mean can be expected to do worse on a subsequent trial.”

• More to do with base rates than CLT.

• Person who scored 750 on SAT could be lucky, unlucky, or perfect true score of 750.

• Only people 751-800 were unlucky. People scoring 650-749 lucky. There are a lot more of the people who could have been lucky – most likely possibility is that they were lucky.
What types of simulations have we done so far?
One of these things is not like the others …
Simulations –

1) Hands on simulations from a marbles and “fish.” Knowing the truth about the population!

2) Simulate a population on computer and do the exact same thing as above. Generate population that you know something about and then see how some sort of statistical approach performs.

3) When I did the distributions of prob(extreme event) – this was a simulation of simulations!

4) Have some data and we don’t know the truth about the population, simulate other samples that are similar to get more information about other values of summary statistic that our sample might have taken. In this case we don’t know the truth. Bootstrap.

E4E – Mix some of the above ideas so we could simulation populations (we know the truth) and then test how well the bootstrap (a simulation procedure when we don’t know the truth) works.
Graphing Principles

Exploratory Data Analysis
Goal: Understand your data!
- What is the general shape of each variable?
- Are there any unusual observation?
- What are relationships between your variables?
- How well does a model fit your data?

Final Graphical Display
1) Display as much of the actual data as possible
2) Display uncertainty in estimates
3) Answer a question with the graph
4) Question -> Graph -> Statistical Analysis
The Histogram

1. What kind of population distribution do the data come from?
2. Where are the data located?
3. How spread out are the data?
4. Are the data symmetric or skewed?
5. Are there outliers in the data?
1. Is there a difference between groups?
2. Does the location differ between subgroups?
3. Does the variation differ between subgroups?
4. Are there any outliers?
5. Is the distribution symmetric?

- Particularly good for large data sets
- Beware of multiple ways that a boxplot can be designed
x1 <- rnorm(15,4,1)
x2 <- rnorm(20,2,1)
data <- list(pools=x1,riffles=x2)
par(bty="l")
stripchart(data,vertical=T,method="jitter",ylab="Fish density",pch=16)
boxplot(data, add=T, outline=F, boxwex=0.4)
```r
x1BS <- numeric(10000)
x2BS <- numeric(10000)
xDiff <- numeric(10000)
for(i in 1:10000){
  mean1 <- mean(sample(x1,replace=T))
  mean2 <- mean(sample(x2,replace=T))
  x1BS[i] <- mean1
  x2BS[i] <- mean2
  xDiff[i] <- mean1-mean2
}
q1 <- quantile(x1BS,prob=c(0.025,0.975))
q2 <- quantile(x2BS,prob=c(0.025,0.975))
points(x=c(1,2),y=c(mean(x1),mean(x2)),pch=16,cex=2,col="blue")
segments(x0=c(1,2),y0=c(q1[1],q2[1]),x1=c(1,2),y1=c(q1[2],q2[2]),lwd=4,col="blue")
```
The Effect Size?

Graphs from homework – Difference in means and a CI around that difference.

```
xDiff[i] <- mean1-mean2

# CI for the difference (non overlapping CI's not the same as a CI for the difference)
> quantile(xDiff,prob=c(0.025,0.975))
  2.5%    97.5%
  1.837049 3.035392
```
Flipflop versus DailySeasonal mean(flipflop, na.rm=T)

How to treat “NA” in the bootstrap???
The Scatterplot

1. Are variables $X$ and $Y$ related?
2. Are variables $X$ and $Y$ linearly related?
3. Are variables $X$ and $Y$ non-linearly related?
4. Does the variation in $Y$ change depending on $X$?
5. Are there outliers?
1. Are there any shifts in location? Is parameter stationary over space or time?
2. What is the magnitude of the shifts in location?
3. Is there a distinct pattern in the shifts in location?
Graphing Principles

Exploratory Data Analysis
Goal: Understand your data!
- What is the general shape of each variable?
- Are there any unusual observations?
- What are relationships between your variables?
- How well does a model fit your data?

Final Graphical Display
1) Display as much of the actual data as possible
2) Display uncertainty in estimates
3) Answer a question with the graph
4) Question -> Graph -> Statistical Analysis

Every graph is an answer to a question!!
SWAM Model Overview

Overlay geospatial datalayers with sample extents

- Significant Variables
- Predictive model of population distribution

Salmon Population Data

Geospatial (Habitat) Datalayers

Road Density

Overlay geospatial datalayers with sample extents

Statistical analyses

- Significant Variables
- Predictive model of population distribution
Oregon Plan Integrated Monitoring

- Spawning Surveys
- Habitat Surveys
- Juvenile Surveys
- Water Quality
Sample Sites

Yellow triangles = Index sites
Pink squares = random surveys
What’s the question?
What’s the answer?
What's the question? What's the answer?
Species distributions across SE Alaska
What’s the question?
What’s the answer?
What’s the question?
What’s the answer?
FIGURE 3. Comparisons of mean length and date at capture in the estuary for juvenile steelhead by stock, production type, and year. Panels (A) and (B) show plots of mean length and peak migration date, with model-based 95% credible intervals. The dots indicate individual fish. Gray bars and gray dots represent hatchery fish, open bars and black dots represent unclipped fish. Panels (C) and (D) pertain to the main effects for length and timing. The points are the estimated
Visualizing Uncertainty
Predicting sea level rise
And then there’s ...
Fig. 3. Coho salmon summer populations, spring migrants, and overwinter survival for the treatment and reference streams in the Alsea and Nestucca basins, pretreatment and posttreatment. Solid bars, treatment pre; open bars, reference pre; diagonally hatched bars, treatment post; horizontally hatched bars, reference post.
Fig. 4. Maximum weekly maximum temperature (mean ± SD) for each of the classes with temperature data in the National Ecological Framework for Canada (Marshall et al., 1996) (a) ecoregions (n = 7) and (b) ecodistricts (n = 25); Ecological Land Classification of Ontario (Crim, 2002) (c) ecoregions (n = 6) and (d) ecodistricts (n = 27); and Aquatic Ecosystem Classification (Wichers et al., 2004) (e) aquatic ecosystem units (n = 29) where stream temperature data were available. The ecological classes are arranged along a north–south gradient (left to right), see references for labelled maps of the ecological classes.
Fig. 3  Seasonal temperature trend estimates across 18 stream sites in the northwest U.S. Error bars denote ± 1 SE; asterisks denote trends that are significantly different from zero at $\alpha \leq 0.05$. Arrows indicate SE values that extend beyond the Y-axis scale.