STAT 100 Chance and Data Analysis Nov. 22, 2002
Revised.

Today: Review topics
Logic of Confidence Intervals – Resting Pulse = 72/min?
Testing Hypotheses with Confidence Intervals

Review topics: This is the same list I sent you in e-mail – am posting for the record. Send me an e-mail about the one or two things you most want me to review. (weldon@sfu.ca)

Major Headings:
Unexplained Variation (UV)
Numerical Summary of Data
Sampling
Role of Simulation in Data Analysis
Role of Graphics in Data Analysis
Experimental Design
Probability
Prediction & Regression
Quality Control
Testing Hypotheses

More detailed headings

Unexplained Variation (UV)
Numerical Summary of Data
- Mean
- Standard Deviation
- Median and Quartiles
- Interquartile Range (IQR)
- Percentiles
- Range (problem of dependence on sample size)
- Sorting Tables
- Forming Indices to collapse data
  - (e.g. birth-death data, economic indicators, colour matching)
- Calibration (e.g. essay marking, earthquake age)
- Evaluating Course Marks ("Curving"?)
- Correlation
  - Positive and Negative and 0

- Sampling
    - Random Sampling
        - sampling with and without replacement
    - Variability of Averages
        - square root law
        - sampling with and without replacement
    - Variability of Proportions
        - Proportions are averages
    - Estimation of parameters
        - Earthquake article
        - Confidence Intervals
    - Survival Analysis
        - Censored data (observation window)
        - Hazard
    - Randomized Response Technique
    - Lotteries
        - Average return
        - Binomial Model
        - Carry-over and no-carry-over types
    - Wild animal populations
    - Estimation of population minimum (difficult!)
        - Travelling Salesman Problem
        - Nearest Neighbour Approach
        - Optimization
- Role of Simulation in Data Analysis
    - Sports Leagues
    - Investments
        - Portfolio Diversification
        - Illusion of Persistent Trends
        - Risk vs Variability
    - Insurance
        - Spreading the risk
        - Needs for profitability, premiums
    - Variability in 0-1 data
    - Random walk with variable step sizes
- Role of Graphics in Data Analysis
    - Time Series
        - Seasonality and seasonal adjustment
        - Intervention

- Dotplot
- Scatterplot (for two variables)
- Smoothing of time series
  - Lowess
  - Moving Average
- Straight line summary of correlated data
  - Intercept and Slope
- Zipf's Law
- Visualization of correlation
- 3-D scatter plots
- Probability plots

Experimental Design

- Randomization
- Blocking
- Control Group
- Double Blind
- Assignment of treatment by Investigator
- Causality – how to prove it
- Lurking Variables in Observational Studies (Berkeley Graduate Admissions)
- Cost considerations

Probability

- Long Run Relative Frequency
- Law of Averages (counts vs proportions in long run)
- Rare Events – Lotteries
- Binomial Distribution Model
  - Sampling of categories
  - Approximate Normality
- Normal Distribution Model
  - 1,2,3 SD proportions
  - Use for describing distribution of averages and sums
- Geometric Distribution Model
  - Use for survival & duration data
  - Probability Model vs Empirical Model
- Weibull Distribution Model
  - Use for survival & duration data

Prediction & Regression

- Linear Regression Lines
  - Prediction Using Averages of Vertical Strips
  - Least Squares (Minimum sum of squared deviations)

Prediction errors

Using a Numerical Index to predict a category (colour-matching)

Quality Control

management by exception

profiting from reduced variability

control charts – timing of exceptions and elimination of causes

Testing Hypotheses

Logic of learning from surprise

Comparison of two populations vs two sample distributions

Between-Sample vs Within-Sample variability

Hypothesis Testing – a demonstration with some details

Resting Heart Rate (Pulse)

Participation voluntary (as is attendance!)

Measure pulse over a 30 second period. Any difference in average pulse by sex?

Report frequency by hands (actual data from class):

	Females	Males
20-22 (21)	0	0
23-25 (24)	0	0
26-28 (27)	0	3
29-31 (30)	0	4
32-34 (33)	6	8
35-37 (36)	7	12
38-40 (39)	8	5
41-43 (42)	3	1
44-46 (45)	1	4
47-49 (48)	1	1
50-52 (51)	0	0

We can summarize this data by means and SDs (using midpoint data) and graphically by a dotplot.:

```
mean-F   37.7
 SD-F     3.8
mean-M   35.8
 SD-M     5.1
```

```
Character Dotplot


                         .        :
                  :      :        :
                  :      :        :      .
                  :      :        :      :       .      .
-----+---------+---------+---------+---------+---------+-female
                         :
                         :
                  :      :
                  :      :        .
   .      :       :      :        :              :
   :      :       :      :        :      .       :      .
-----+---------+---------+---------+---------+---------+-male
  28.0      32.0      36.0      40.0      44.0      48.0
```

There appears to be a shift from male to female in these distributions. But since they would hardly ever be exactly the same, the question is not whether they are different but whether they are different enough to infer that a population difference exists.
Suppose we assume that the male students reporting are a sample of a larger population of male students and similarly for females. Would average pulse be different in these larger groups, given this data?

A first attempt to answer this question might be to look at the mean and SD from each group. This allows us to focus on the numerical summary of each distribution.

We had

```
mean-F   37.7
 SD-F     3.8
mean-M   35.8
 SD-M     5.1
```

We could say female pulses were 37.7 ± 3.8 and males 34.1 ± 5.1. As we already know from the dotplot, the two sample distributions overlap. But the question is still of interest – what we want to know is if there is a tendency for the female pulses to be higher and the males pulses to be lower. One way to make this precise is to ask if the population means differ. Our evidence for this is the sample means. But in considering the difference of sample means, we need to compare this difference with the variability of the sample means. Using the square root law, we have

mean-F 37.7
SD of the mean-F = 3.8/ $\sqrt{26}$ = .75
mean-M 34.1
SD of the mean-M = 5.1/ $\sqrt{38}$ = .83

So we really have 37.7 ± .75 as our estimate of the population mean for females, and 35.8 ± .83 as our estimate of the population mean for males. This suggests a difference in the population means, but it is still not clear whether it is large enough to discount sampling variability (instead of a real difference in population means.)

> Note the important difference between the SD of the pulses in each group, and the SD of the mean pulse in each group. Also, an important point in the discussion of estimation of populatiojn parameters (such as the population mean) is the following:
>
> Random sampling is an "unbiased" procedure – that is, one tends to get the right thing on average. More precisely, the average value of the sample mean if we were to repeat the sampling process many times, would be exactly the population mean. Now when we are considering using the sample mean to estimate the poulation mean, since we know it is "right" on average, the precision of the estimate only depends upon its variability. So the square root law is the key to how good the sample mean is as an estimate of the population mean – it tells you the variability of the sample mean.

Now to return to our consideration of the difference between the female and male pulse distributions ....

We have 37.7 ± .75 for females
35.8 ± .83 for males

Before we make a final judgement on whether this apparent difference is evidencve ffor a difference in the two populations (or females and males), there are two details that could be leading us astray:

1. mean ±SD only includes 68% of the distribution. So there still might be some overlap.

2. The difference in means will vary more than the difference in either mean. Since we are really intersted in whether the population difference is 0 or not, this could be important.

The solution to 1. is to look at mean ± 2 SDs which will include the true population mean 95% of the time (this needs more explanation).

Addressing 1. we have

37.7 ± 2(.75) = 37.7 ± 1.5 pulses for females
35.8 ± 2(.83) = 35.8 ± 1.7 pulses for males

and so now our difference in population means lookslike it might possibly be 0. (difference between 37.7 and 35.8 attributable to sampling variation).

These [mean ± 2 SDs] are called **95% Confidence Intervals** for the population means.

The solution to 2. is to compute the SD of the difference of means (this needs more explanation).

The SD of the difference is calculated as $\sqrt{.75^2+.86^2}=1.1$ and the sample difference is 37.7-35.8 = 1.9 so we could say the difference is 1.9 ± 2(1.1) or 1.9 ± 2.2.

This last interval does contain 0, so a 0 population difference is credible, and the data does not prove a difference in population means exists. We can conclude that the means of the two population distributions have not been shown to be different. (It is not true that we have shown the difference IS 0.)

Homework
To prepare for the continued discussion of this topic, please read Tanur pp 68-76, "The Importance of Being Human". In addition to the idea that overlapping distributions can still be distinguishable, there is the important use of a scatterplot to distinguish subgroups in two-variable data.