ESD.86
Descriptive Statistics and Statistical Graphics

Dan Frey
Associate Professor of Mechanical Engineering and Engineering Systems
Plan for Session

• Descriptive statistics
• Visual perception / cognition
• Tufte's paper
  – Statistical graphics
  – Statistical thinking
Measures of Central Tendency

• Arithmetic mean
  – an unbiased estimate of

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

• Median

\[
\mu = E(x) = \int_{S} xf_x(x)dx
\]

• Mode – The most frequently observed value in the sample

\[
\begin{align*}
X_{\frac{n+1}{2}} & \text{ if } n \text{ is odd} \\
\frac{1}{2} \left( X_{\frac{n}{2}} + X_{\frac{n}{2}+1} \right) & \text{ if } n \text{ is even}
\end{align*}
\]
Confidence Intervals

• Assuming a given distribution and a sample size $n$ and a given value of the parameter $\theta$ the 95% confidence interval from $U$ to $V$ is s.t. the estimate of the parameter $\hat{\theta}$

$$\Pr(U < \hat{\theta} < V | \theta) = 95\%$$

• The confidence interval depends on the confidence level, the sample variance, and the sample size
Concept Question

• Here is a sample of data
• This Matlab code computes a 95% confidence interval for the mean
• Then it runs a t-test on the data and the hypothesis that it was from a normal distribution with a mean greater than or equal to the lower confidence bound
• What p-value will the test return?

```matlab
sample=[1.2 3.1 4.5 2.3 0.8 7.2];
[h,p,ci] = ttest(sample);
[h2,p2] = ttest(sample,ci(1),0.05,'left')
```

1) exactly 0.95
2) exactly 0.975
3) more than 0.975
4) less than 0.95
5) none of the above
Concept Question

• This Matlab code repeatedly generates simulated "data" (5 samples) from a normally distributed population with a known mean and computes a 95% confidence interval
• How often will the 95% confidence interval include the mean of the distribution from which the samples came?

```matlab
pm=1;
for i=1:1000
    sample=random('Normal',pm,1,1,5);
    [h,p,ci] = ttest(sample);
    in_int(i)=(ci(1)<=pm)*(ci(2)>=pm);
end
mean(in_int)
```

1) 95% of the time
2) a little less than 95%
3) a little more than 95%
4) Not enough information
Bayesian Credible Interval

• In Bayesian statistics, a credible interval is a posterior probability interval...

• For example, a statement such as "following the experiment, a 90% credible interval for the parameter $t$ is 35-45" means that the posterior probability that $t$ lies in the interval from 35 to 45 is 0.9.
Measures of Dispersion

- Population Variance
  \[ \text{VAR}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \]

- Sample variance
  \[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \]
  – an unbiased estimate of \[ \sigma^2 = E((x - E(x))^2) \]

- \( n^{th} \) central moment
  \[ E((x - E(x))^n) \]

- \( n^{th} \) moment about \( m \)
  \[ E((x - m)^n) \]
Skewness and Kurtosis

- **Skewness**  \( E((x - E(x))^3) \)
  - Positively skewed distribution

- **Kurtosis**  \( E((x - E(x))^4) \)
  - Positive kurtosis
Correlation Coefficient

• Sample

\[ r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)S_X S_Y} \]

\[ S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \]

• Which is an estimate of

\[ \frac{E((x - E(x))(y - E(y))))}{\sigma_x \sigma_y} \]
But What Does it Mean?

Pick a Box! Are they all the same?

Correlation 0.7

Correlation 0.7

Correlation 0.7

Correlation 0.7

The Human Brain

• The human brain is tremendously complex
• It possesses many highly specialized component parts each associated with specific tasks
• The functioning of the human brain gives rise to
  – Consciousness
  – Intelligence (ability to learn, plan...)
  – Emotions
  – Decision making
• “The mind … accomplishes remarkable feats
  no engineer can duplicate…”

Evolution and the Brain

• “The mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced…”

Three Brains in One

- **Reptilian Complex**
  - digestion, reproduction, circulation, breathing, "fight or flight" response

- **Limbic System**
  - houses primary centers of emotion
  - hippocampus -- important aspects of long term memory

- **Neocortex**
  - processing senses
  - logic
  - language
  - motor control

Figure by MIT OCW.
Neocortex

• 2mm thick
• 6 layers
• About the size of a newspaper (unfolded)  
  – So it has to be wrinkled up to fit in the skull
• About $3 \times 10^{10}$ neurons
• About 1000 synapses per neuron
• Each neuron capable of 200 cycles / sec  
  – 5 million times slower than a computer
Visual Cortex

About $\frac{1}{3}$ of the cerebral cortex.

Retina – a $2\frac{1}{2}$-D data stream
$\sim$ a million nerve fibers

The fiber pathways are two-way… They carry as much information down from higher conceptual areas as from lower sensory areas...

Humans are much better than computers at classifying objects based on complex, noisy, ambiguous images.

Certain classes of things have specialized areas.

The same areas involved in “seeing” a face or place are active in thinking about it.

Face Recognition

• “saccade” = a fast motion of the eye (~3 per second)
• Your eye can be tracked as it looks at faces
• Your eyes are apparently very active in seeing a face
• The activity depends on familiarity

Images of eye tracking during face recognition removed due to copyright restrictions.
Change Blindness

• “…focused attention is needed to detect change”*
• You cannot assume that information flowing into your brain will be attended to; it may be ignored completely
• Is it also true of graduate students?
• Is it true of a professors?
• Is it possible to do good research when key information fails to be acknowledged?

Asch Conformity Experiments

• Subjects asked to participate in a "vision test"
• In reality, all but one of the participants were confederates of the experimenter
• A high proportion (~33%) conformed to the erroneous majority view
• Many subjects showed extreme discomfort, but conform anyway
A Hypothesis

• Facts about the brain provide some insight into problems in:
  – statistical inference
  – research methods
  – decision making

• Cognitive sciences and psychology also provide insights into the remedies
Edward R. Tufte's Paper

• "Visual and Statistical Thinking: Displays of Evidence for Making Decisions"
• What are its key points?
  – History
  – Research practices
  – Statistics per se

Graphics Press LLC, Cheshire CT.
The General Argument

1. (Tufte) "An essential analytic task in making decisions based on evidence is to understand how things work – mechanism, trade-offs, process and dynamics, cause and effect. That is, intervention-thinking and policy-thinking demand causality-thinking."

2. (Tufte) "Making decisions based on evidence requires the appropriate display of that evidence. Good displays of that data help to reveal knowledge relevant to understanding mechanism, process and dynamics, cause and effect. That is, displays of statistical data should directly serve the analytic task at hand."

3. (Frey) Replace the instances of the term "making decisions" with "doing high quality ESD research." The sentences above are still correct yet have different implications.
The Cholera Epidemic in London, 1854

- Dot plot of cholera instances suggests a spatial pattern
- Centered on a water pump

Map removed due to copyright restrictions.

What about the Exceptions?

- Very few instances in the work house and brewery despite proximity to the suspected pump
- Also some instances far from the pump

Map removed due copyright restrictions.
Aggregation of Data

Figures removed due copyright restrictions.

Aggregation in time

Spatial Aggregation
The Challenger Decision

Upon ignition, smoke leaked from this joint. A flame burned through 59 seconds later.

Rubber O-rings, nearly 38 feet (11.6 meters) in circumference; 1/4 inch (6.4 mm) thick.

The field joint that leaked.

Courtesy of NASA.

Figure by MIT OCW.
Right Conclusion, Poor Communication

**PRIMARY CONCERNS**

**FIELD JOINT - HIGHEST CONCERN**

- EROSION PENETRATION OF PRIMARY SEAL REQUIRES RELIABLE SECONDARY SEAL FOR PRESSURE INTEGRITY
  - IGNITION TRANSIENT - (0-600 MS)
    - (0-170 MS) HIGH PROBABILITY OF RELIABLE SECONDARY SEAL
    - (170-330 MS) REDUCED PROBABILITY OF RELIABLE SECONDARY SEAL
    - (330-600 MS) HIGH PROBABILITY OF NO SECONDARY SEAL CAPABILITY
  - STEADY STATE - (600 MS - 2 MINUTES)
    - IF EROSION PERETRATES PRIMARY O-RING SEAL - HIGH PROBABILITY OF NO SECONDARY SEAL CAPABILITY
    - BENCH TESTING SHOWED O-RING NOT CAPABLE OF MAINTAINING CONTACT WITH METAL PARTS GAP OPENING RATE TO MEOP
    - BENCH TESTING SHOWED CAPABILITY TO MAINTAIN O-RING CONTACT DURING INITIAL PHASE (0-170 MS) OF TRANSIENT

Source: NASA
Confusion about Evidence, Burden of Proof, and Significance

Absence of evidence ≠ evidence of absence
Would this Have Been Convincing?

Image removed due to copyright restrictions.
Plot of O-ring damage index vs. Temperature at time of launch.

- Continuous measure of degradation
- Plotted versus the suspected variable
- Shows the range over which the prediction is needed
- The lack of data in that range is the most striking feature
What about Feynman's Demonstration?

• It did show the phenomenon at issue
• But it also confounds many factors
• It doesn't show consistency with most of the data (non-failures)

Photo and diagrams removed due to copyright restrictions.
load carsmall

gscatter(Weight,MPG,Model_Year,"','xos')

figure(2)
xvars = [Weight Displacement Horsepower];
yvars = [MPG Acceleration];
gplotmatrix(xvars,yvars,Model_Year,"','xos')
Histograms

• A graph of continuous data
• Approximates a pdf in the limit of large $n$
Box and Whisker Plot

load fisheriris
s1 = meas(51:100,3);
s2 = meas(101:150,3);
boxplot([s1
    s2],'notch','on','labels',
    {'versicolor','virginica'})
Florence Nightingale's Polar-Area Diagrams

Source: Nightingale, F. "Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army." 1858.
Next Steps

• Wednesday, 11 April
  – Session on Regression (no pre-read)

• Friday, 13 April
  – Session to support the term project
  – Be prepared to stand up and talk for 5 minutes about your ideas and your progress

• 16-17 April, No classes (Patriot's Day)

• Wednesday 18 April, ANOVA