TA: Jason Fischer

W411 Turner Hall
602-499-8701
fischer9@illinois.edu
IB 372
Lab 1: Introduction to Statistics

Fall 2010
Thanks to Steve Paton - Smithsonian Tropical Research Institute for providing original spanish version of this file...
What are statistics?
Why do we use them?
What are statistics?
Why do we use them?

How do you define real?
What are statistics?

- Statistics are numbers used to:

  **Describe** and **draw conclusions** about DATA

- These are called *descriptive* (or “univariate”) and *inferential* (or “analytical”) statistics, respectively.
Statistic vs. Parameter

Formally (and somewhat confusingly):

• A **statistic** is a measure of some attribute of a **sample**.

• Whereas, a **parameter** is the real and unique measure of the attribute for the whole **population**.

• Usually, the population is too big to measure, so in practice, **statistics represent parameters**. (thus, even “descriptive” stats are usually inferential too)
Variables

• A variable is anything we can measure/observe
• Three types:
  – Continuous:
  – Discrete:
  – Categorical:

• Dependence in variables:

  “Dependent variables depend on independent ones”
Variables

• A variable is anything we can measure/observe
• Three types:
  – Continuous: values span an uninterrupted range (e.g. height)
  – Discrete: only certain fixed values are possible (e.g. counts)
  – Categorical: values are qualitatively assigned (e.g. low/med/hi)

• Dependence in variables:

  “Dependent variables depend on independent ones”
Descriptive Statistics
Descriptive statistics

Techniques to summarize continuous and discrete data

Numerical
- Mean
- Variance
  - Standard deviation
  - Standard error
- Median
- Mode
- Skew
- etc.

Graphical
- Histogram
- Boxplot
- etc.
The Bell Curve

- Central tendency – clustering of data around a particular value (mean, median, mode)

- Variability of data (skew, variance)
The *Mean*: Most important measure of “central tendency”

**Population Mean**

\[
\mu = \frac{\sum_{i=1}^{N} X_i}{N}
\]
The **Mean**: Most important measure of “central tendency”

**Sample Mean**

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]
Additional central tendency measures

**Median**: the 50\(^{th}\) percentile

\[
M = \begin{cases} 
X_{(n+1)/2} & \text{(n is odd)} \\
\frac{X_{n/2} + X_{(n/2)+1}}{2} & \text{(n is even)}
\end{cases}
\]

**Mode**: the most common value

1, 1, 2, 4, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 10, 12, 15
The Bell Curve

Where are the mean, median, and mode?
The Bell Curve

Where are the mean, median, and mode?
Variance:
Most important measure of “dispersion”

\[ \sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \]

Population Variance
Variance:
Most important measure of “dispersion”

Sample Variance

\[ s^2 = \frac{\sum (X_i - \overline{X})^2}{n - 1} \]

From now on, we’ll ignore sample vs. population. But remember:

We are almost always interested in the population, but can measure only a sample.
Additional *dispersion* measures

*Standard deviation:* average distance from the mean

\[ s = \sqrt{s^2} \]
>99%

~95%

~68%

Amazing!

Handy!

Important!
You're three standard deviations above the norm.

Love letter from a statistician.

Um... Thanks?
Additional *dispersion* measures

**Standard deviation:**
average distance from the mean

\[ s = \sqrt{s^2} \]  
(duh!)

**Standard error:**
the accuracy of our estimate of the population mean

\[ SE = \frac{s}{\sqrt{n}} \]

Bigger sample size (n) \( \rightarrow \) smaller error

**Range:** total span of the data \((X_{\text{max}} - X_{\text{min}})\)
Additional dispersion measures

**Skew:** unequal distribution of data around the area of central tendency
- separates the measures of central tendency

**Range:** total span of the data \((X_{\text{max}} - X_{\text{min}})\)
“Graphical Statistics”
The Friendly Histogram

- Histograms represent the *distribution* of data
- They allow you to visualize the *mean*, *median*, *mode*, *variance*, and *skew* at once!
Constructing a Histogram is Easy

<table>
<thead>
<tr>
<th>X (data)</th>
<th>Value</th>
<th>Frequency (count)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Histogram of X
Interpreting Histograms

- Mean?
- Median?
- Mode?
- Skew?
- Shape?
Interpreting Histograms

- Mean?  
  $= 9.2$
- Median?  
  $= 6.5$
- Mode?  
  $= 3$
- Standard deviation?  
  $= 8.3$
- Variance?  
- Skew?  
  (which way does the “tail” point?)
- Shape?
An alternative: Boxplots
Boxplots also summarize a lot of information...

Within each sample:

- 75% percentile
- Median
- 25% percentile
- “Outliers”

Compared across samples:

Weight (kg)

- SC
- Ba
- Se
- Es
- Da
- Is
- Fe

Island
Break!
Inferential Statistics: Introduction
**Inference**: the process by which we draw conclusions about an unknown based on evidence or prior experience.

In statistics: make conclusions about a population based on samples taken from that population.

*Important: Your sample must reflect the population you’re interested in, otherwise your conclusions will be misleading!*
Statistical Hypotheses

• Scientific vs. Statistical hypotheses

• Very often presented in pairs:
  – Null Hypothesis ($H_0$):
    Usually the “boring” hypothesis of “no difference”
  – Alternative Hypothesis ($H_A$)
    Usually the interesting hypothesis of “there is an effect”

• Statistical tests attempt to (mathematically) reject the null hypothesis
Significance

• Your sample will never match $H_0$ perfectly, even when $H_0$ is in fact true.

• The question is whether your sample is different enough from the expectation under $H_0$ to be considered significant.

• If your test finds a significant difference, then you reject $H_0$. 
**p-Values Measure Significance**

The *p*-value of a test is the probability of observing data at least as extreme as your sample, assuming $H_0$ is true.

- If $p$ is very small, it is unlikely that $H_0$ is true.
  
  (in other words, if $H_0$ were true, your observed sample would be unlikely)

- How small does $p$ have to be?
  - It’s up to you (depends on question)
  - 0.05 is a common cutoff
    - *If* $p<0.05$, *then there is less than 5% chance that you would observe your sample if the null hypothesis was true.*
Errors in Hypothesis Testing

• **Type I Error**: Reject $H_0$ when $H_0$ is actually true
  – i.e. You find a difference when really there is none
  – The probability of **Type I error** is called the “significance level” of a test, and denoted $\alpha$

• **Type II Error**: Accept $H_0$ when $H_0$ is actually false
  – i.e. There really is a difference, but you conclude there is none
  – The probability of **Type II error** is denoted $\beta$

  (and $[1 – \beta$ is called the “power” of the test)
‘Proof’ in statistics

• Failing to reject (i.e. “accepting”) $H_0$ does *not* prove that $H_0$ is true!

• And accepting $H_A$ doesn’t prove that $H_A$ is true either!

  Why?

• Statistical inference tries to draw conclusions about the population from a small sample
  – By chance or poor collection methods, the samples may be misleading
  – Example: if you always accept $H_0$ at $p=0.05$, then *1 in 20 times you will be wrong!*
Assumptions of inferential statistics

- All inferential tests are based on assumptions
  - If your data cannot meet the assumptions, the test results may be invalid!
- In particular:
  - Inferential tests assume *random sampling*
  - Many tests assume the data fit a *theoretical distribution* (often normal)
    - These are “parametric tests”
    - Luckily, there *are* non-parametric alternatives
The Normal Distribution
aka “Gaussian” distribution

• Occurs frequently in nature

• *Especially* for measures that are based on *sums*, such as:
  – sample means
  – body weight
  – “error”
  (*aka “the Central Limit Theorem”*)

• *Many* statistics are based on the assumption of normality
  – You *must* make sure your data are normal, or try something else!

Sample normal data:
Histogram + theoretical distribution
(i.e. sample vs. population)
Properties of the Normal Distribution

• Symmetric
  Mean = Median = Mode

• Theoretical percentiles can be computed exactly
  ~68% of data are within 1 standard deviation of the mean
  >99% within 3 s.d.
  “skinny tails”
What if my data aren’t Normal?

• It’s OK!
• Although lots of data are Gaussian (because of the CLT), many simply aren’t.
  – Example: Fire return intervals
• Solutions:
  – Transform data to make it normal (e.g. take logs)
  – Use a test that doesn’t assume normal data
    • Don’t worry, there are plenty
    • Especially these days...
• Many stats work OK as long as data are “reasonably” normal
Inferential Statistics: Methods
Student’s $t$-Test
Student’s \( t \)-test

• Several versions, all using inference on a sample to test whether the true population mean (\( \mu \)) is different from \_
  
  – The one-sample version tests whether the population mean equals a specified value, e.g.
    
    \[ H_0: \mu = 0 \]
    
  – The two-sample version tests whether the means of two populations are equal
    
    \[ H_0: \mu_1 = \mu_2 \]
$t$-Test Methods

• Compute $t$
  
  – For one-sample test:  
    
    $$t = \frac{\bar{x} - \mu}{SE}$$ 
    
    Remember:
    
    $$SE = \frac{s}{\sqrt{n}}$$

  – For two-sample test:
    
    $$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}$$ 
    
    $$S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
t-Test Methods

- Once you have $t$, use a look-up table or computer to check for significance.
- Significance depends on “degrees of freedom” (basically, sample size).
- Bigger difference in means and bigger sample size both improve ability to reject $H_0$. 

![Table of t-distribution critical values](image)
How does the t-Test work?

• Statistical magic!
• Student figured out the “t-distribution” shown to the left.
  – Given a sample mean and df, we can see where the mean is likely to be.
• If the null-hypothesis mean is very unlikely to fall under the curve, we reject $H_0$.

Reject $H_0$ if your $t$ is in one of the red regions!
Try it!

Work through the Excel exercises for one- and two-sample $t$-tests now
ANOVA
ANOVA: ANalysis Of VARIance

• Tests for significant effect of 1 or more categorical factors on a dependent variable
• Each factor may have 2 or more levels
• Can also test for interactions between factors

• For just 1 factor with 2 levels, ANOVA = t-test
  – So why can’t we just do lots of t-tests for more complicated experiments?
ANOVA: Concept

• Despite the name, ANOVA really looks for difference in means between groups (factors & levels)

• To do so, we partition the variability in our data into:
  – (1) The variability that can be explained by factors
  – (2) The leftover unexplained variability (error or residual variability)

Total variability = Variability due to factors + error
(we only have to calculate two of these values)
Example: We study tree growth rates on clay vs. sand vs. loam soil (10 trees each)

How many factors?

How many levels for each factor?

What is $H_0$?
ANOVA example continued

Example: We study tree growth rates on clay vs. sand vs. loam soil (10 trees each)

Square = clay soil
Diamond = sand soil
Triangle = loam soil
‘replicate’ = plot
‘y’ = growth
ANOVA example continued

- First find total variability using *Sum of Squares*
  - Find overall mean (horizontal line)
  - Each “square” is the distance from one data point to the mean, squared
  - *Total Sum of Squares* (*SST*) is the sum of all the squared deviations
ANOVA example continued

• Now measure variability unexplained by factor of interest (soil)
  – Find means for each level
  – *Error Variability (SSE)* is the sum of all squared deviations from these level means

  – *Which is greater—SSE or SST?*

The remaining variability is due to soil factor (say, *SSF*). It’s easy to compute, since

\[
\text{SST} = \text{SSE} + \text{SSF}
\]
ANOVA example continued

• Next, we calculate *degrees of freedom* ($df$)
  – $df$ is based mainly on sample size
  – Every time you estimate a parameter from your data, you lose one $df$
    • Example: Since we computed the mean of our 30 observations, we only need to know 29 of the values now to determine the last one!

• For our example we have:
  
  $df_{SST} = 30 - 1 = 29$
  $df_{SSE} = (10 - 1) + (10 - 1) + (10 - 1) = 27$
  $df_{SSF} = df_{SST} - df_{SSE} = 2$
ANOVA example continued

• From SS and df, we compute *Mean Square (MS)* variability

• Finally (!) we test whether the variability explained by our factor is significant, relative to the remaining variability
  – The ratio $\text{MS}_{\text{soil}}/\text{MS}_{\text{error}}$ is $F$
  – By statistical magic, we can look up the probability of observing such a large $F$ just by chance.

*In other words, we find the p-value associated with $H_0$: Soil has no effect on growth*

• We can then go back and see which groups differ (e.g. by *t*-test)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil</td>
<td>99.2</td>
<td>2</td>
<td>49.6</td>
<td>4.24</td>
<td>0.025</td>
</tr>
<tr>
<td>Error</td>
<td>315.5</td>
<td>27</td>
<td>11.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>414.7</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What do we conclude?
Try it!

Work through the Excel exercise for ANOVA now
Chi-square ($\chi^2$) Test

• In biology, it’s common to measure frequency (or count) data

• $\chi^2$ is used to measure the deviation of observed frequencies from an expected or theoretical distribution:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where: $O$ is the observed frequency (# of events, etc.)

$E$ is the expected frequency under $H_0$
$\chi^2$ Example

• Again we’ll also need to know degrees of freedom
• For the $\chi^2$ test,
  \[
  \text{df} = \text{number of groups} - 1
  \]
• Then (again by statistical magic), we can look up how big $\chi^2$ needs to be (the critical value) to reject $H_0$ at a given significance level
Imagine we conduct an experiment to determine the food preference of rodents, with the following results:

<table>
<thead>
<tr>
<th>Food</th>
<th># Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuna</td>
<td>31</td>
</tr>
<tr>
<td>Peanut butter</td>
<td>69</td>
</tr>
<tr>
<td>Fresh fish</td>
<td>35</td>
</tr>
<tr>
<td>Cheese</td>
<td>65</td>
</tr>
</tbody>
</table>

\[ n = 200 \]

A reasonable *expectation* (our null hypothesis) is:

\[ H_0 = \]
Imagine we conduct an experiment to determine the food preference of rodents, with the following results:

<table>
<thead>
<tr>
<th>Food</th>
<th># Eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuna</td>
<td>31</td>
</tr>
<tr>
<td>Peanut butter</td>
<td>69</td>
</tr>
<tr>
<td>Fresh fish</td>
<td>35</td>
</tr>
<tr>
<td>Cheese</td>
<td>65</td>
</tr>
</tbody>
</table>

\( \chi^2 \) Example

A reasonable *expectation* (our null hypothesis) is:

\[ H_0 = \text{Rodents like all foods equally well} \]

Thus, under \( H_0 \) our expected frequency for each food is:

\[ \frac{200}{4} = 50 \]
χ² Example

First, we draw up a contingency table:

<table>
<thead>
<tr>
<th></th>
<th>tuna</th>
<th>PB</th>
<th>fish</th>
<th>cheese</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>31</td>
<td>69</td>
<td>35</td>
<td>64</td>
<td>200</td>
</tr>
<tr>
<td>Expected</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>200</td>
</tr>
</tbody>
</table>

Then we compute χ²:

\[
\chi^2 = \frac{(31 - 50)^2}{50} + \frac{(69 - 50)^2}{50} + \frac{(35 - 50)^2}{50} + \frac{(65 - 50)^2}{50} = 22.0
\]
The critical value for df = 3, \( \alpha = 0.05 \) is \( \chi^2_{0.05,3} = 7.815 \).

Since our \( \chi^2 \) is greater than \( \chi^2_{\text{critical}} \), we reject \( H_0 \). Our results differ significantly from the expectation under \( H_0 \), suggesting that there actually is a food preference in the studied rodents.
Try it!

Work through the Excel exercise for the Chi-square test now
Correlation

• *Correlation* measures the strength of the relationship between two continuous variables
  – *When X gets larger, does Y consistently get larger (or smaller)?*

• Often measured with Pearson’s correlation coefficient
  – Usually just called “correlation coefficient”
  – Almost always represented with the letter $r$
Correlation

Computing Pearson’s correlation coefficient:

\[ r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} \]

-1 ≥ r ≥ 1

Amount that X and Y vary together

Total amount of variability in X and Y
Correlation Examples
**Correlation Cautions**

- *Correlation does not imply causality!*  
  (note: doesn’t matter which data are X vs. Y)
- *$r$ can be misleading* 
  – it implies nothing about *slope*  
  – it is blind to outliers and obvious nonlinear relationships

![Same $r$ in each panel!](image-url)
Try it!

Work through the Excel exercise for correlations now
Regression

• Unlike correlation, *regression does* imply a functional relationship between variables
  – The *dependent* variable *is a function* of the *independent variable(s)*
  – What kind of variables are these?

• In regression, you propose an algebraic model, then find the “best fit” to your data
  – Most commonly, the model is a simple line (“*Y is a linear function of X*”)
There are many possible relationships between two variables...
Regression

- We’ll focus on simple linear regression of one dependent (Y) and one independent (X) variable
- The model is:

\[ Y = a + bX + \varepsilon \]

- \( Y \) = values of the dependent variable
- \( X \) = values of the independent variable
- \( a, b \) = “regression coefficients” (what we want to find)
- \( \varepsilon \) = residual or error
In regression we always plot the independent variable on the x-axis.

Potential Regression Outcomes:

- **Positive Relationship**: $b > 0$
  - Y vs. X diagram

- **Negative Relationship**: $b < 0$
  - Y vs. X diagram

- **No Relationship**: $b = 0$
  - X-axis only
How do we “fit” a Regression?

- Most common method is “least-squares”
  - Find $a$ and $b$ to minimize the (squared) distances of data points from the regression line

\[
\sum (Y - a - bX)^2 = \text{min}
\]

\[
\hat{Y}_5 - \bar{Y}_5 = \epsilon_5
\]
How do we “fit” a Regression?

Find individual residuals (ε):

\[ Y_i - \hat{Y}_i = \varepsilon_i \]

Then the sum of all (squared) residuals is:

\[ \sum (Y_i - \hat{Y}_i)^2 \]

A computer or clever mathematician can find the \( a \) and \( b \) that minimize this expression (producing the “best fit” line).
Regression: Is the fit significant?

• We usually ask whether the model is significant by testing $b$.

\[ Y = a + bX + \varepsilon \]

• The null hypothesis:
  \[ H_0: b = 0 \] (the line may as well be flat)

against the alternative:
  \[ H_A: b \neq 0 \] (the best fit isn’t a flat line)

• Luckily (statistical magic), this can be tested with the $t$-test!
  – Depends on degrees of freedom
    (increase sample size to improve significance)
Regression: How good is the fit?

Perfect fit - all the points on the line

Good fit

OK fit (I’ll take it!)
Regression: How good is the fit?

• Formally, we often measure the fit with the coefficient of determination, $R^2$.
• $R^2$ is the proportion of variation in Y “explained” by the regression
  – Values range from 0 to 1
  – 0 indicates no relationship, 1 indicates perfect relationship

*Note:* In simple linear regression, yes, $R^2$ is actually the same as Pearson’s correlation coefficient ($r$) squared. But this is just a special case—regressions get much more complicated. Don’t get in the habit of confusing the two statistics!
Regression Reservations

• Again, regression *does* imply causality (unlike correlation), but importantly, *it still does not test a causal biological relationship*
  – Some other variable might affect both X and Y
  – You could even have the relationship backwards!

• Be careful extrapolating regression lines:
  – beyond the original data range, or
  – to other populations
Mycoplasmal conjunctivitis and bird feeding
Change in house finch abundance (finches/party hour)

East

$y = 0.1345x - 1.0045$

Change in feeder density (people/km²)
## Regression Example

<table>
<thead>
<tr>
<th>Altitude (m)</th>
<th>Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25.0</td>
</tr>
<tr>
<td>50</td>
<td>24.1</td>
</tr>
<tr>
<td>190</td>
<td>23.5</td>
</tr>
<tr>
<td>305</td>
<td>21.2</td>
</tr>
<tr>
<td>456</td>
<td>20.6</td>
</tr>
<tr>
<td>501</td>
<td>20.0</td>
</tr>
<tr>
<td>615</td>
<td>18.5</td>
</tr>
<tr>
<td>700</td>
<td>17.2</td>
</tr>
<tr>
<td>825</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Which is the independent variable?  
What is $H_0$?
Try it!

Work through the Excel exercise for simple linear regression now
Choosing the appropriate test

• Ask yourself what kind of variables you are testing

• Ask whether you are looking for differences among means or whether the variables are related