# Overview

## Statistics of Well-Behaved Data (44 slides)

Probability Concepts
    7 slides (count doesn't include title, goals, or R lesson), 1 R lesson
Measures of Location and Spread
    7 slides, 1 R lesson
Functions of Random Variables
    8 slides, 1 R lesson
Confidence Intervals
    10 slides, 1 R lesson
Hypothesis Testing
    12 slides, 1 R lesson

## Statistics of Real Data (24 slides)

Identifying Outliers and Non-Gaussian Data
    7 slides, 1 R lessons
Effect of Outlier on Location
    7 slides, 1 R lesson
Effect of Outliers on Spread
    6 slides, 1 R lessons
Confidence Interval for the Trimmed Mean
    4 slides, 1 R lesson

## Regression with Well-Behaved Data (44 slides)

Introduction to Curve Fitting
    7 slides, 1 R lesson
Least-Squares Fit to a Straight Line
    10 slides, 1 R lesson
Generalized Least-Squares
    10 slides, 1 R lesson
Non-Linear Least-Squares
    11 slides, 1 R lesson
Confidence and Correlation
    6 slides, 1 R lesson

## Regression with Real Data (18 slides)

Problems and Warnings
    5 slides, no R lesson
Diagnostics and Remedial Measures
    10 slides, 1 R lesson
Robust Regression
    3 slides, 1 R lesson

# Slide-by-Slide Headings

## Probability Concepts

- randomness/random variable/types of random variables
- probability/discrete random variables/continuous random variables
- binned random variables
- statistical independence/probability of measuring multiple statistically independent events
- probability density function/discrete random variable
- continuous random variable
- probability over ranges

## Measures of Location and Spread

- mode/median/mean
- normal pdf parameters
- skewed Gaussian
- estimation of the mode/estimation of the median
- estimation of the mean
- variance/standard deviation
- location and spread for several distributions

## Functions of Random Variables

- linear functions
- sums of a single random variable
- pdf of the average
- how are counts averaged?
- the Central Limit Theorem
- non-linear functions
- propagation of variance

## Confidence Intervals

- cumulative distribution function
- probability intervals
- confidence intervals
- confidence interval of the average
- visualization of confidence limits
- why doesn't the interval with s work?
- the t-distribution
- confidence intervals using t
- revisiting the required number of replicates

## Hypothesis Testing

- z-test for one average
- graphical interpretation
- single-sided z-test for one average
- t-test for one average
- t-test example
- single-sided t-test example
- F-statistic
- how well does $s^2$ approximate $\sigma^2$?
- example F-test
- t-test for two averages
- example t-test for two averages
- p-values

## Identifying Outliers and Non-Gaussian Data

- why do we want Gaussian-distributed data?/what hope do we have that the distribution is Gaussian?/what goes against this hope?
- how are outliers and non-Gaussian distributions recognized?
- fallacy of the three sigma edit
- box plots
- probability plot
- histogram

## Effect of Outliers on Location

- order statistics and L-estimators
- effect of a single outlier on the mean of six numbers
- the median and its robustness
- comments on the median
- the trimmed mean
- robustness of the trimmed mean
- the Winsorized mean

## Effect of Outliers on Spread

- effect of a single outlier on the standard deviation of six numbers
- median absolute deviation (MAD)
- robustness of the MAD
- inter-quartile range (IQR)
- robustness of the IQR
- effect of a mixed distribution

## Confidence Interval of the Trimmed Mean

- confidence interval of the trimmed mean with known σ
- experimental variance of Winsorized data
- experimental variance of trimmed data
- t-limits for the trimmed mean

## Introduction to Curve Fitting

- an equation parameter is the end goal of the experiment
- an optimal strategy for testing multiple hypotheses
- curve fitting with a normal pdf
- curve fitting with counting
- curve fitting with measured data
- probability for a set of measurements
- method of maximum likelihood

## Least-Squares Fit to a Straight Line

- probability for $y = a_0 + a_1x$
- total probability and chi-square
- minimizing chi-square
- solving for the intercept
- solving for the slope
- un-weighted least-squares
- example un-weighted least-squares
- chi-square minimum for the example
- weighting with counting error
- iteration gives excellent results with counting error

## Generalized Least-Squares

- what is a linear least-squares
- linearizable equations
- the general linear equation
- notational simplification
- matrix least-squares
- errors in the coefficients
- example matrix least-squares
- alternative matrix formalism
- associated matrix algebra

## Non-Linear Least-Squares

- the basic idea
- a two-parameter equation
- two-parameter example
- replicate data sets
- effect of noise on chi-square
- another view of chi-square space
- initial parameter estimates
- grid search
- gradient search
- Taylor's Series Expansion
- Marquardt Algorithm

## Confidence and Correlation

- confidence interval of the slope
- example of a slope t-test
- confidence interval of the intercept
- coefficient of determination (multiple $R^2$)
- assumptions about and limitation of $R^2$/other thoughts

## Problems and Warnings

- unequal contributions to chi-square space
- false minima in chi-square space
- false minima with identical functions
- simple conceptual problems
- more complex conceptual problems

## Diagnostics and Remedial Measures

- residuals
- model departures examined via residuals/diagnostics for residuals
- incorrect regression function
- detecting randomness
- variable error
- non-independence of residuals
- detecting non-normality
- detecting outliers/ futility of the three sigma edit
- identifying outliers using the hat matrix and residuals
- identifying outliers using studentized deleted residuals

## Robust Regression

- Tukey's tri-median
- least absolute deviation regression
- robust regression using MM-estimators

# Description of Programming Projects

1. **Ion-Count Simulation:** run and understand a pre-written program
2. **Standard Deviation Investigations:** (a) Process a set of data for location and spread using a variety of R functions. (b) Demonstrate that degrees of freedom need to be N-1.
3. **Root Sum of Squares:** a Monte Carlo simulation of the propagation of variance
4. **Confidence Interval for the t-Statistic:** a Monte-Carlo investigation into t-intervals
5. **Comparing Two Distributions:** uses synthetic data sets created by the student to perform F- and t-tests
6. **Distribution of Log(Counts):** uses a Monte Carlo simulation to compare the distribution of log(counts) to a Gaussian distribution
7. **Deprecated**
8. **Comparing the Mean, Median, Trimmed Mean and Winsorized Mean:** write a function computing the Winsorized mean. Use a data set with and without an outlier to investigate the robustness of the location estimators.
9. **Comparing the Standard Deviation, MAD and IQR:** use a set of data with and without an outlier to investigate the robustness of the spread estimators
10. **Deprecated**
11. **Confidence Limit of a Trimmed Mean:** compare a trimmed mean to a specified value and show its ruggedness toward an outlier
12. **Probability of an Average:** show that the average maximizes the total probability for estimating the location of the data distribution. Compare this value to the probability of obtaining the "true" value.
13. **Least-Squares Fit to Counting Data:** given a set of data use an iterative least-squares to fit the data
14. **Least-Squares Fit to Absorption Spectra:** given a set of data solve the hard way using matrix algebra. Solve the problem the easy way using the R function lm().
15. **Chi-Square Space and a Non-Linear Fit:** given a set of data from a Lorenzian function make initial parameter guesses and map out chi-square space in the vicinity of the guess
16. **Comparing Two Calibration Curves:** given two sets of data obtain the coefficients and errors. Compare the two intercepts and slopes using the t-test. Report the $R^2$ value.
17. **Dropping Residuals:** write a program that uses the hat matrix and studentized deleted residuals to identify outliers
18. **Ignoring Outliers:** given a set of data compare three robust methods to ordinary least-squares fits with and without the outlier