Food for Thought

Intelligent Data Binning

Wed.Feb.06.2013

Use diverse methods to solve commonly-occurring problems.
Recognize opportunities for mathematical & statistical analysis.
Demonstrate modern tools and utilities.
Present benefits of data visualization.

Today's Recipe

- ***Hummus***
  - Serves: 4 Equipment: food processor
- **Ingredients**
  - 1 (15 oz) can (425 grams) chickpeas
  - 1/4 cup lemon juice (1 large lemon)
  - 1/4 cup tahini (Krinos)
  - 1/2 garlic clove, minced
  - 2 tablespoons (30 ml) olive oil, plus more for serving
  - 1/2 to 1 teaspoon salt (to taste)
  - 1/2 teaspoon (2 grams) cumin
  - 2-3 tablespoons (28 – 42 ml) water
  - paprika for dusting on top

Common Data Analysis Problems

- What's the best distribution for this set of data points?
- What if there is no good fit to the data set?
- How should *outliers* be defined and counted?
  - In statistics, an outlier is an observation that is numerically distant from the rest of the data.
- What if the data sets are conditionally dependent?
- There simply is no universally good way to figure out high-dimensional distributions from scratch.
- Mathematica notebooks: plotCurves.nb, plotHistograms.nb

Estimating Probability Densities

- Ways of estimating densities:
  - Histograms
  - Kernel density estimation
    - KDE is a non-parametric way to estimate the probability density function of a random variable -- making data smooth and continuous.
  - Local polynomials
  - Series expansions
  - Splines
  - Wavelets and Fourier series
- →Work with probability *mass* function not probability density function because the data is discrete.

The Data Binning Problem

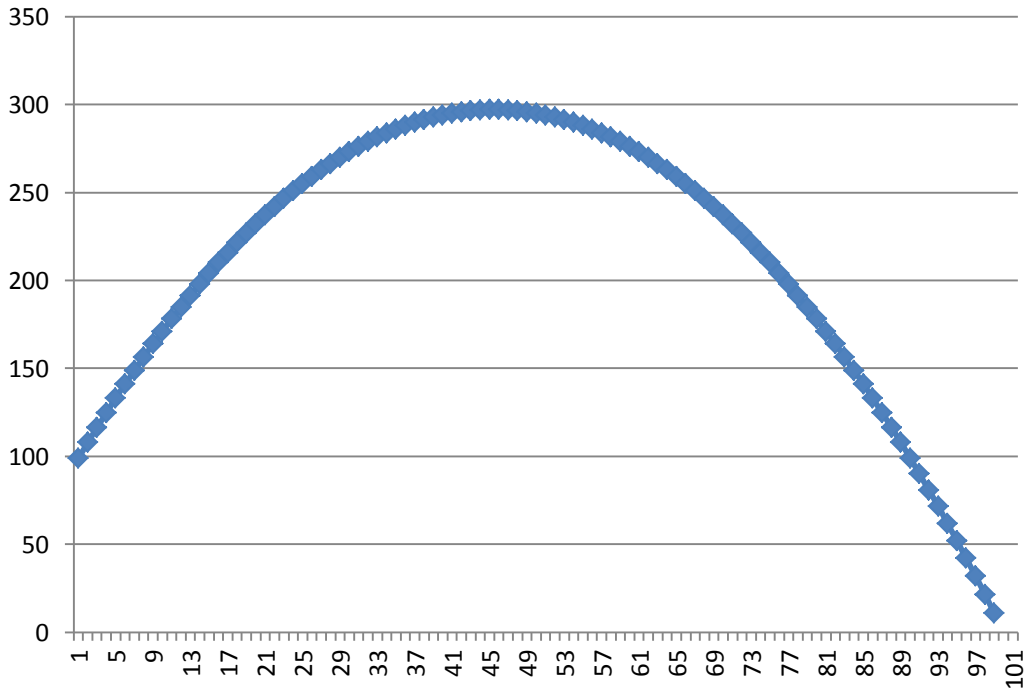
- There is no "best" number of bins. Different bin sizes reveal different data features that depend on actual data distribution and analysis goals.
- Bias-variance trade-off:
  - when using a large number of very small bins, the minimum bias in an estimate of any density decreases, but the *variance* in the estimates increases.
- ***Claim***: Best bin width for both continuous & discrete data demands tradeoff:
  - maximize number of bins (minimize bin width): $\mathbb{R} \rightarrow r$
  - minimize number of empty bins: $\mathbb{N} \rightarrow n$
  - $\{0 < r \leq \max(p)\} * \{0 \leq n \lesssim \text{number of points}\}$

Uniform Binning Methods

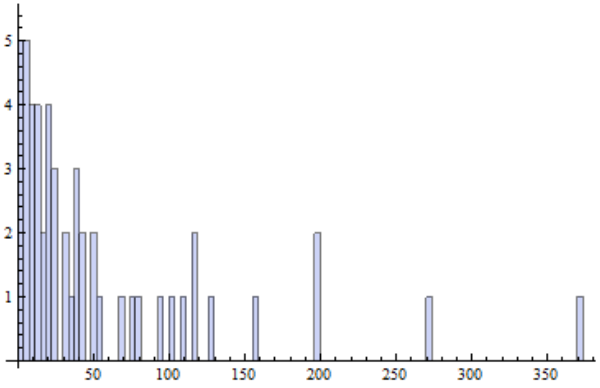
- *m* (say, 100) equally-spaced bins
- (Max-Min) / *n*
- Square root of standard deviation
- Square root of mean
- **Max Bins * Min Empty Bins**
- Freedman-Diaconis: Twice the interquartile range divided by the cube root of sample size.
  - $h_{opt} \propto n^{1/3}$

Data Binning Approach

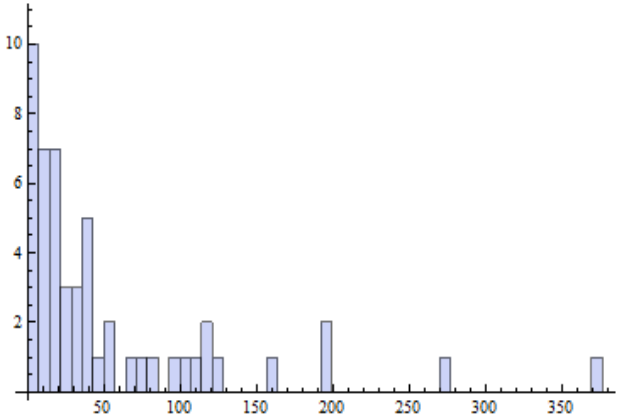
- Max function of 2 variables: {$\mathbb{R}\rightarrow r$, $\mathbb{N}\rightarrow n$: r*n}



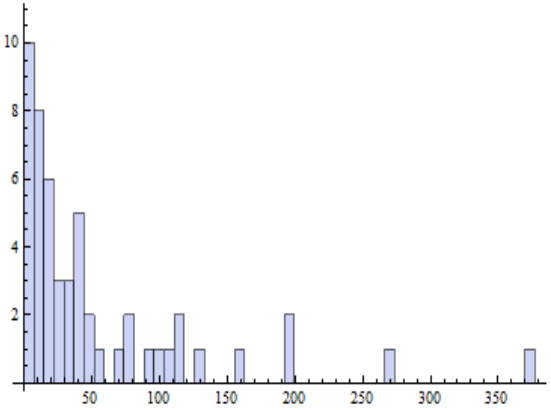
Adenocarcinoma.EGF Bin size=3.6973 Bin method: (Max-Min) / 100



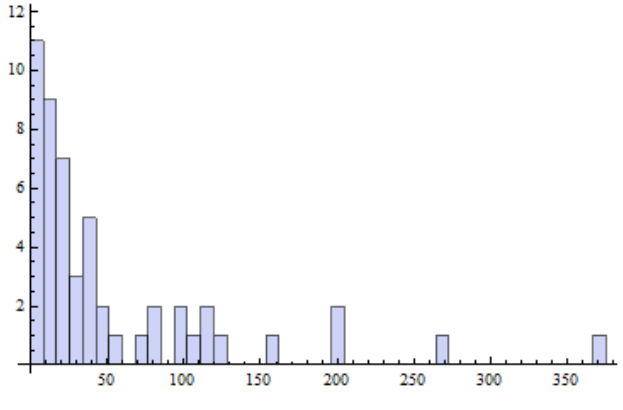
Adenocarcinoma.EGF Bin size=7.1102 Bin method: (Max-Min) / nPoints



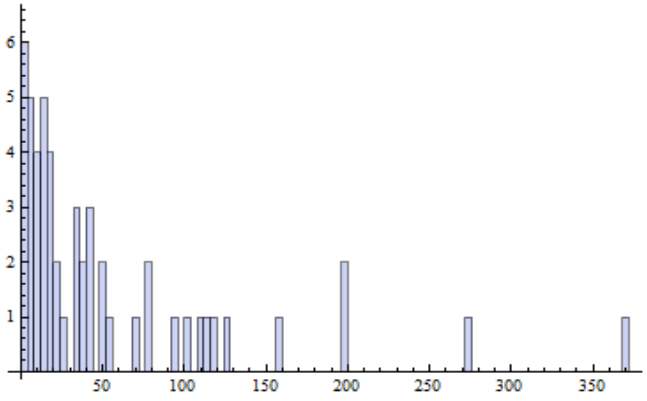
Adenocarcinoma.EGF Bin size=7.3953 Bin method: Square root of Mean



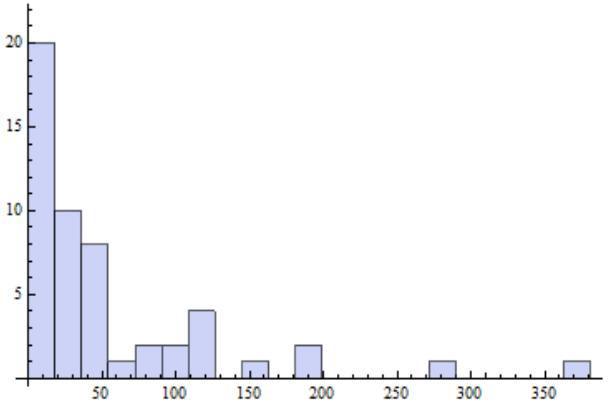
Adenocarcinoma.EGF Bin size=8.5264 Bin method: Square root of StdDev



Adenocarcinoma.EGF Bin size=3.9973 Bin method: MaxBins / MinEmptyBins



Adenocarcinoma.EGF Bin size=18.1259 Bin method: Freedman-Diaconis



What is *Mathematica*?

- ***Mathematica*** is a computational engine for use in scientific, engineering, mathematics, and technical computing.
  - computer algebra, symbolic/numeric computation, visualization, statistics
- Since 2008, it supports parallel computing grid technologies.
  - Windows HPC Server, Microsoft Compute Cluster Server, Sun Grid
- Since 2010, it supports CUDA and OpenCL GPU hardware and can automatically generate C code.
  - Intel C++ Compiler or Visual Studio 2010
- Integrated with ***Wolfram Alpha*** (2009), an online service that answers factual queries directly by computing the answer from a knowledge base of curated, structured data instead of providing a list of documents or web pages.
- Installed in our PC and Mac Labs.

Mathematica v9 Features

- Histogram[], DistributionFitTest[], PDF[], ListPlot[], ChiSquareDistribution[], SmoothHistogram[].
- Example: ***plotHistograms.nb***
  - Histogram[data, {BinSizes[[n]]}, "Count"]
  - disFitObj = DistributionFitTest[data, ChiSquareDistribution[v], "PearsonChiSquare"];
  - Show[SmoothHistogram[data], Plot[PDF[disFitObj["FittedDistribution"], x], {x, 0, 400}, PlotStyle -> Red], PlotRange -> All]
- KDE is implemented by SmoothKernelDistribution[].
- Symbolic estimation by KernelMixtureDistribution[].

What is C♯?

- **C♯** is a strongly typed programming language with imperative, declarative, functional, generic, object-oriented, event-driven, reflective, concurrent, and component-oriented paradigms.
  - Microsoft Visual Studio 2012 Ultimate installed in our PC labs.
- The .NET Framework provides the ***Common Language Runtime*** (CLR), a managed execution run-time environment, which runs the code and provides services that make the development process easier.
- Anders Hejlsberg designed the CLR (2000), which drove the design of the C♯ language itself.

C♯ v4.0 Features

- **Generics**:

```
List<MyClass> = new List<MyClass>();
```

- **Variance**:
  - *Covariance*: an IEnumerable<A> is considered an IEnumerable<B> if A has a reference conversion to B (out)
    - Covariant type parameters enable you to make assignments that look like ordinary polymorphism:

```
IEnumerable<Derived> d = new List<Derived>();
IEnumerable<Base> b = d;  // Every Derived is a Base.
```

  - *Contravariance*: (convert from general type to specialized type)
    - contravariant type parameters can be used as parameter types:

```
Action<Base> b = (target) => { Console.WriteLine(target.GetType().Name); };
Action<Derived> d = b;
d(new Derived());
```

C♯ v4.0 Features

- **Functional programming**:
  - A ***lambda expression*** is an anonymous method used to create *delegates* or *expression tree* types and is the preferred way to write inline code. They allow writing local functions that can be passed as arguments or returned as the value of function calls.
    - An anonymous method provides an unnamed function block.
    - A delegate is a type-safe function pointer.
    - An expression tree allows translation of executable code into data.
- **L**anguage **In**tegrated **Q**uery (LINQ): ***CytokineSample.sln***
  - extension methods, implicitly typed variables, object initializers, anonymous types
  - LinQ to SQL, LinQ to XML, LinQ to Objects: treat data as sets!
  - LINQ's query operators work with any collection of data that implements the IEnumerable<T> interface.

C♯ Code - LINQ

```
var query = (from item in results
                where (item.ClinicalType == cType) && (item.Sample == sample)
                group item by new { item.ClinicalType, item.Sample } into g
                select new {
                   g.Key.ClinicalType,
                   g.Key.Sample,  // lambda expressions
                   Avg_IL8 = g.Average(i => i.IL8),
                   Avg_IL10 = g.Average(i => i.IL10),
                   Avg_VEGF = g.Average(i => i.VEGF),
                   Avg_IFNG = g.Average(i => i.IFNG),
                   Avg_TNFA = g.Average(i => i.TNFA),
                   Avg_MCP1 = g.Average(i => i.MCP1),
                   Avg_EGF = g.Average(i => i.EGF)
                }).First();
```

Summary

- Treat the biomarker data set as discrete data "islands" with *no outliers*.
- Optimize bin width as 2-variable function:
  - product of bin width and number of empty bins
- Each "bin" is treated as a random variable.
- *Mathematica* and *C♯* LINQ are modern tools for analyzing small to moderate size data sets.