Epilogue

The complete data analysis story

Given a dataset D with k features and n instances.

- For each continuous column
  - Summary statistics: identify the mean,variance, mode. Compute robust version of these statistics if data has outliers.
  - Visualize the data as histogram or line plot or any other fancy plot
  - Choose among known parametric distribution of density
    - Uniform, Gaussian, exponential, chi-squared or heavy-tailed versions Gamma, t-distribution, beta if quantity between 0 and 1
    - Or non-parametric options like
      - Empirical CDF or Kernel density estimators

Analysing single continuous column (continued)

- Fit parameters using Maximum likelihood estimator or use Bayesian if you have prior knowledge of the column
  - Compute the square error of the fit using a set-aside validation dataset.
  - Think about whether the fit is biased and consistent
- If you have any hypothesis about the domain, use hypothesis tests about its mean or median

If column is discrete

- Summarize the data in terms of number of distinct values, mode, frequency of each column
- Visualize the frequencies as a pie chart or other charts
- If columns is on counts, choose among known parametric distributions for counts
  - Binomial, Poisson, Geometric, Negative binomial, Zipfian.
- Fit parameters using MLE or Bayes

Relationship among continuous column pairs

- Visualize: Inspect scatter plot
- Summarize: Measure the sample covariance matrix and covariance melahun
- Reason about their independence.
    - There are hypothesis tests for checking for independence of two columns which we missed.

Whole data analysis

- If one of the columns is of special interest as a response variable, fit a regression model with that column as target
  - Choose among linear regression and kernel regression models
  - Opt for robust regression is data is suspected to have outliers
  - For linear regression: estimate the square error of the fitted parameters.
    - Hypothesis tests can be used to test the assumptions of normality of the residuals.

If the number of columns is very large

- Reduce the number of dimensions using Principal component analysis

- Visualize the data using t-sne or PCA or other projection methods.

Multivariate normal?

- Fit a multivariate normal if possible
  - Goodness of fit tests exist for this
- Inspect the contours of density

- Compute linear projections of the data (also normal) for inspection into lower dimensional distributions.

If one of the columns is time

- Visualize the data as a time-series
- Summarize the data using auto-correlation functions, remove trend and seasonality, establish stationarity of the series.
- Fit auto-regressive or MA model or a combined ARIMA model
- Inspect the fitted model
- Forecast future values

If more than one dataset sample is available

Perform hypothesis tests on whether they are from the same distribution:

- If both are assumed to be normal: check for equality of means assuming their variance is the same
- If the samples are paired, perform paired-t-test to check if difference of one is greater than another assuming normal
- If nothing is known about the distribution perform two-sample test. Can also be extended to paired samples.