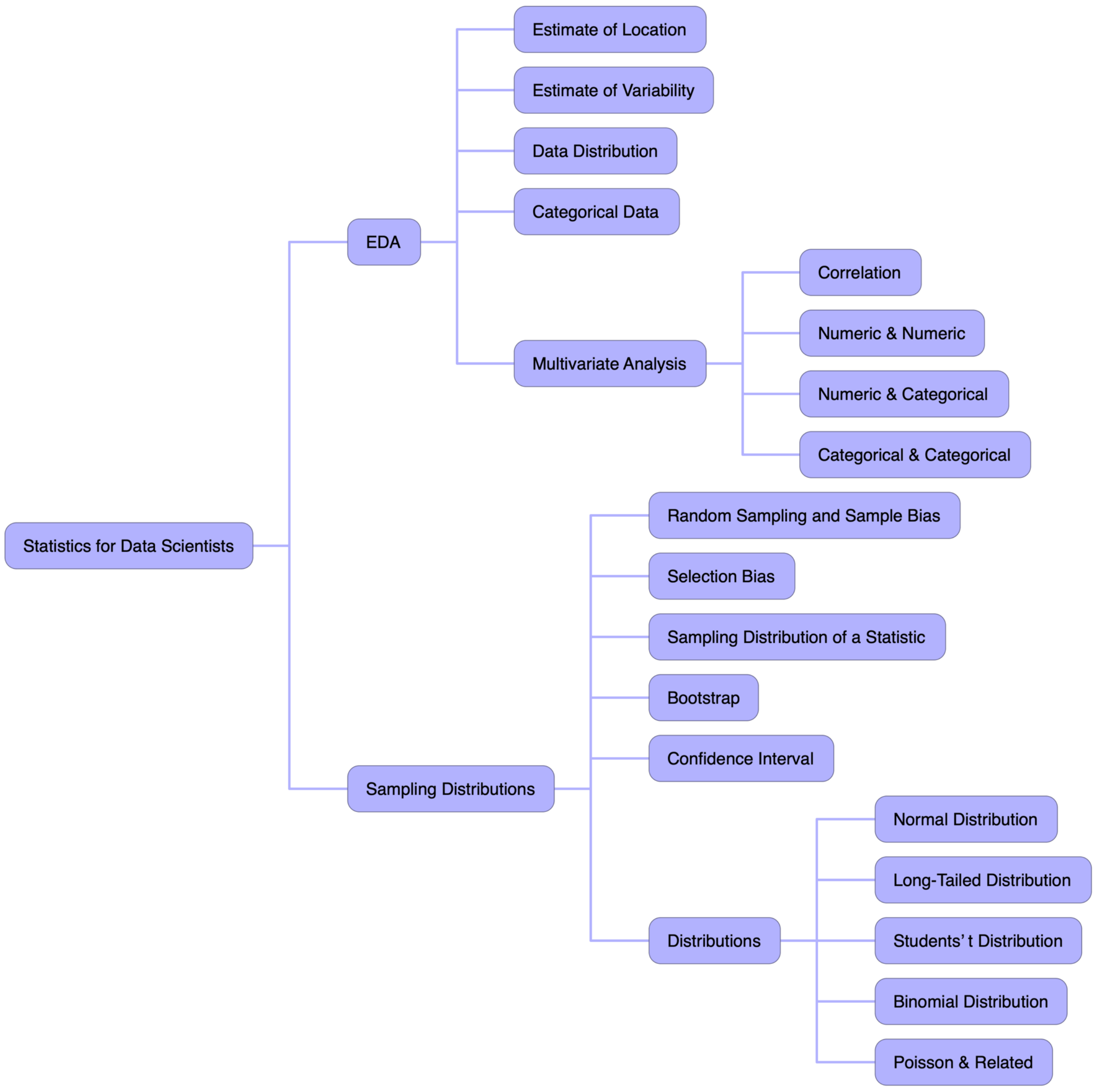
This article is mainly my notes from the book Practical Statistics for Data Scientists by Peter Bruce and Andrew Bruce. Some concepts are from my own interview experiences. It covers the highly frequent concepts in the data scientist interviews.
I will gradually add additional concepts from A/B testing and machine learning.



Part 1 Exploratory Data Analysis

Estimate of Location

Mean: Mean value, trimmed Mean(largest and smallest value omitted), Weighted mean.
Median: Median, Weighted Median.

Estimate of Variability

Deviation (errors, residuals)
Variance(mean-squared-error)

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{N - 1}$$

Standard Deviation (the square root of the variance)

$$\text{Standard Deviation} = s = \sqrt{\text{Variance}}$$

Mean Absolute Deviation(MAD), the mean of the absolute value of the deviations from the mean.
It is also called l1 norm or manhattan norm.

$$\text{Mean Absolution Deviation} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

Median Absolute Deviation from the Median

$$\text{Median Absolution Deviation} = \text{Median}(|x_1 - m|, |x_2 - m|, ..., |x_N - m|)$$

Interquartile Range: Range between 75% percentile and 25% percentile, also called IQR.

Data Distribution

Single Variable

Numeric Values:

Boxplot, Frequency Table, Histogram, Density Plot

Categorical Values:

Mode, Expected Value, Bar Charts, Pie Charts (statisticians try to avoid pie charts for visual purpose).

Coefficient

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

Scatterplots

Multivariate

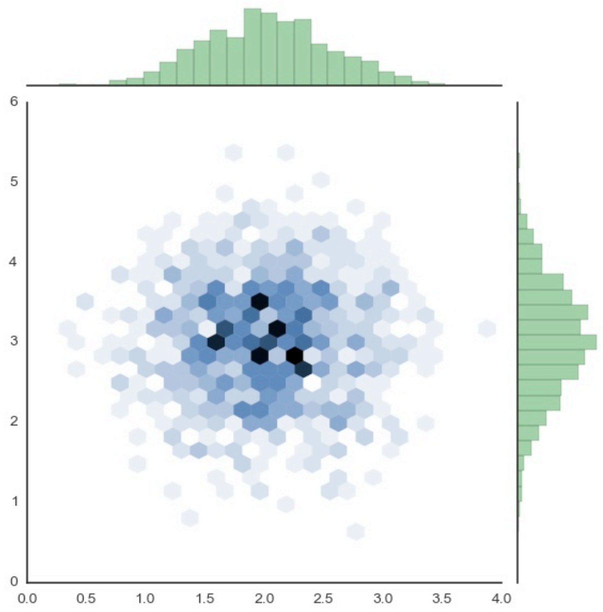
Contingency Table

A tally of counts between two or more categorical variables.

	Male	Female
Blue	2	4
Red	3	5

Hexagonal Binning

A plot of two numeric variables with the records binned into hexagons.



Contour Plots

A plot showing the density of two numeric variables like a topographical map.

Violin Plots

Similar to a boxplot but showing the density estimate.

Sampling Distribution

Basic Concepts:

Sample, population, N, Random Sampling, Stratified Sampling, Sample Bias, Data Quality

Data Quality: Completeness, Consistency of Format, Cleanliness, Accuracy of individual data points, and Representativeness.

Statistical Bias: measurement or sampling errors that are systematic and produced by the measurement or sampling process.
Let's take gun shot range as an example.
Error due to random chance: The gun shots are not in the right center.
Error due to bias: There are more shots in the top right corner.

Data Snooping: Extensive hunting through data in search of something interesting.

Massive Search Effect: Bias or non-reproducibility resulting from repeated data modeling, or modeling data with large numbers of predictor variables.
Massive Search Effect is one type of form of selection bias, others include cherry-picking data, selection of time intervals that accentuate a particular statistical effect and stopping an experiment when the results look "interesting".

Regression to the Mean
Extreme observations tend to be followed by more central ones.

Sampling Distribution of a Statistic

Data Distribution: The frequency distribution of individual values in a dataset.

Sampling Distribution: The frequency distribution of a sample statistic over many samples or resamples.

Central Limit Theorem: The tendency of the sampling distribution to take on a normal shape as sample size rises(even if the population is not normal shaped).

Standard Error: The variability (standard deviation) of a sample statistic over many samples.

$$\text{Standard Error} = SE = \frac{s}{\sqrt{n}}$$

So, to decrease the standard error for 2 times, we need a 4 times larger sample size.

Bootstrap Sample A sample taken with replacement(put the ball back) from an observed dataset.

Resampling The process of taking repeated samples from observed data; includes both bootstrap and permutation(shuffling) procedures.

Confidence Interval 90% confidence interval: It is the interval that encloses the central 90% of the bootstrap sampling distribution of a sample statistic. So the [(1-0.90)/2]% of the B resample results are trimmed from this interval.

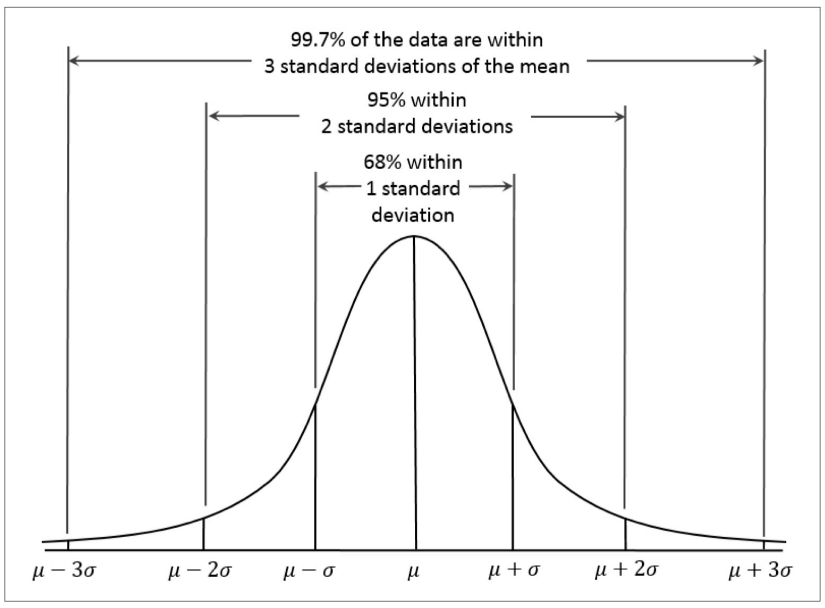
The higher the level of confidence, the wider the interval.
The smaller the sample, the wider the interval.

P-value
p-value is the probability we get this sample or a more extreme sample under H0.
P value can be calculated by computing the test statistics(Z-score) and get the probability of two tails in the t-distribution.

Increasing the sample size will tend to result in a smaller P-value only if the null hypothesis is false.

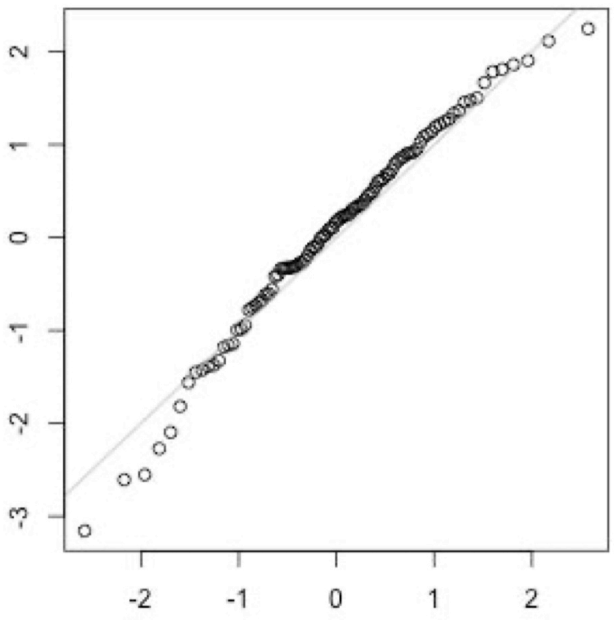
Distributions

Normal Distribution



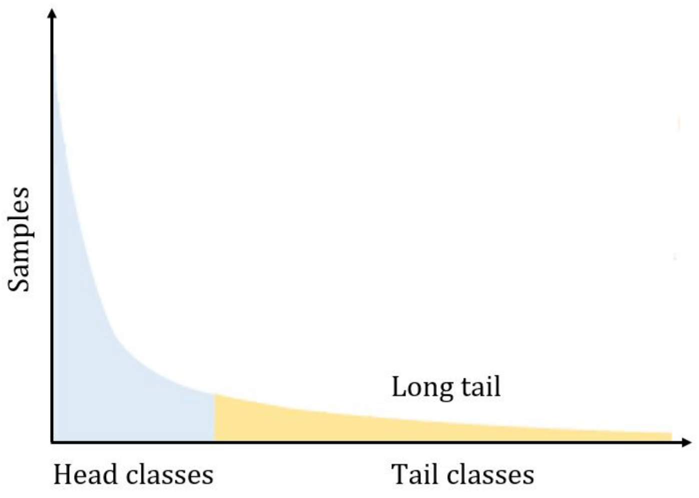
> *Practical Statistics for Data Scientists, Peter Bruce & Andrew Bruce.*

QQ plot



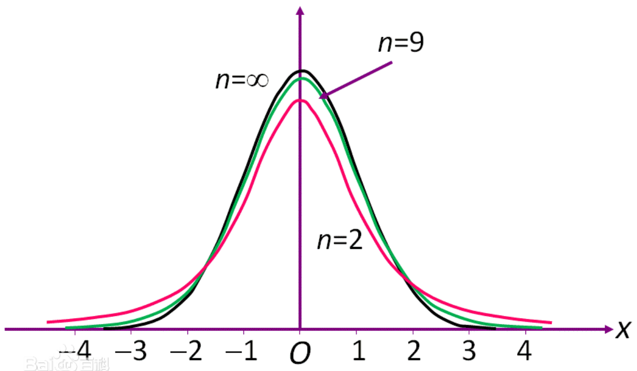
After Z-Score normalization(standardization), put the Z-score on y-axis and the corresponding quantile of a normal distribution for that value's rank on the x-axis, we have a QQ plot.

Long-Tailed Distribution



Student's T Distribution

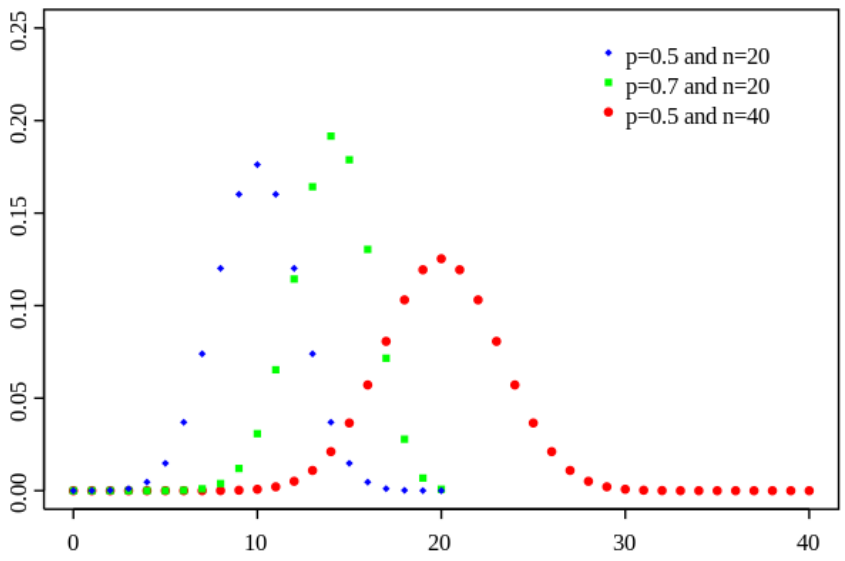
$$f(x)=\frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$$



T distribution has thicker tails. T-test can reduce outliers' impact on standard deviations, thus reducing the sample size demand.

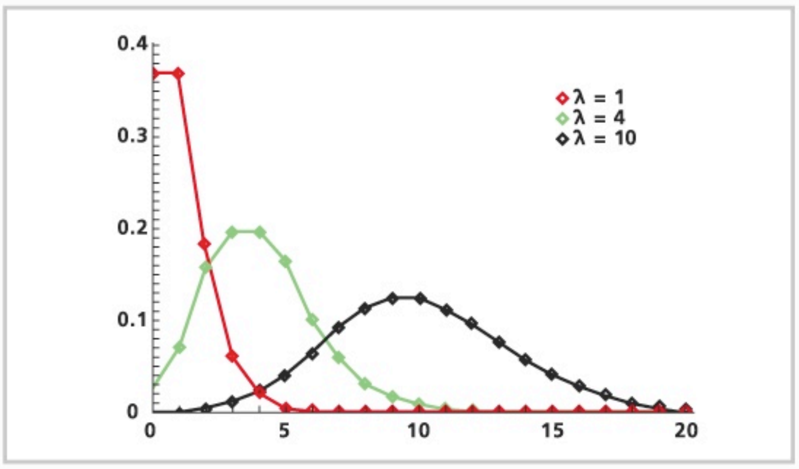
Binomial Distribution

$$\Pr(x)=\frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x} \quad \mu=n\pi \quad \sigma=\sqrt{n\pi(1-\pi)}$$



Poisson Distribution

$$P(x=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$



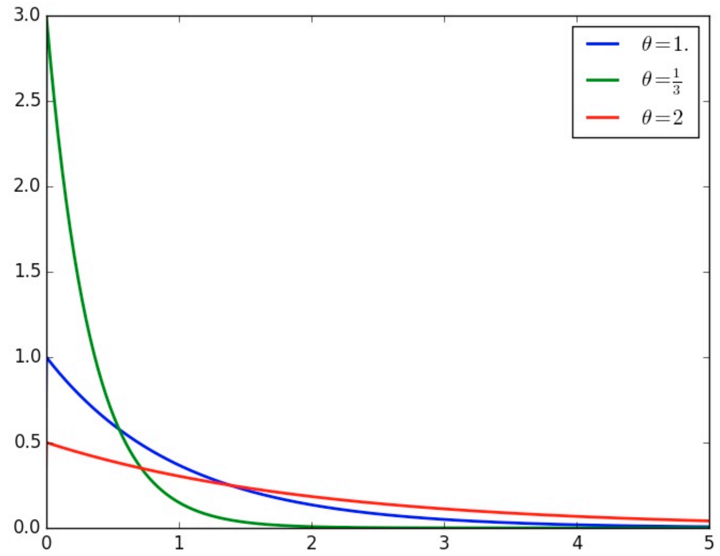
泊松分布

Poisson Distribution is used to describe in a unit period of time, what is the possibility that a random independent event happens.

Exponential Distribution

$$P(X > t) = P(N(t) = 0) = \frac{(\lambda t)^0 e^{-\lambda t}}{0!}$$

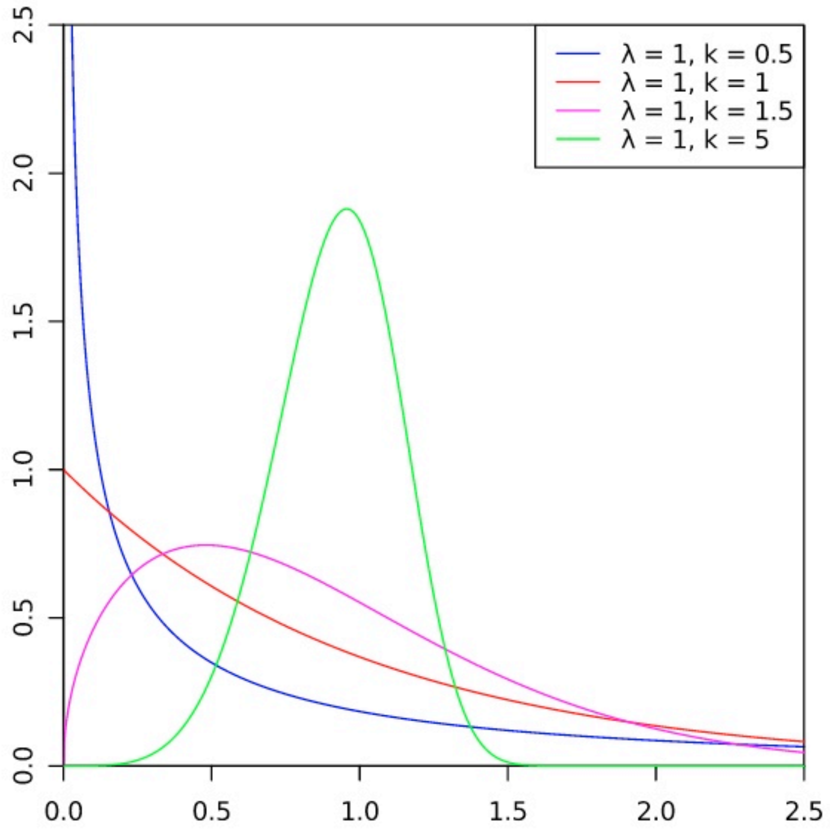
$$= e^{-\lambda t}$$



Exponential distribution is used to describe the length time periods/breaks among independent random events. E.g. the time intervals passengers enter the airport. It can also be used in the durance distribution of systems.

Weibull Distribution

$$f(t) = \frac{\beta t^{\beta-1}}{\eta^{\beta}} e^{-\left(\frac{t}{\eta}\right)^{\beta}}$$



Weibull distribution has more parameters. We can understand it as describing the distribution of a product probability of breaking.