**Table of Contents**

[Exploratory Data Analyst 3](#_Toc54535821)

[Describing Data 3](#_Toc54535822)

[Probability 4](#_Toc54535823)

[Data and Sampling Distributions 5](#_Toc54535824)

[Random variable 5](#_Toc54535825)

[Binomial random variables 6](#_Toc54535826)

[Multiple Random Variables 6](#_Toc54535827)

[Random Sampling and Sample Bias 7](#_Toc54535828)

[Sampling Distribution of a Statistic 7](#_Toc54535829)

[Confidence Intervals 7](#_Toc54535830)

[Normal (Gaussian) distribution 8](#_Toc54535831)

[Common Distributions 9](#_Toc54535832)

[Statistical Experiments and Significance Testing 10](#_Toc54535833)

[A/B Testing 10](#_Toc54535834)

[Hypothesis Test 10](#_Toc54535835)

[Statistical Significance and P-Values 11](#_Toc54535836)

[Hypothesis tests Example 11](#_Toc54535837)

[ANOVA (Analysis of Variance) 12](#_Toc54535838)

[chi-square (χ2) Test 13](#_Toc54535839)

[Finding Sample Size 14](#_Toc54535840)

[Inference for Averages 14](#_Toc54535841)

[Inference for two matched populations 14](#_Toc54535842)

[Inference for two independent populations 14](#_Toc54535843)

Resource 1: Practical Statistics for Data Scientist

*(“Practical Statistic for Data Scientists by Peter Bruce and Andrew Bruce (O’Reilly). Copyright 2017, 978-1-491-95296-2”)*

Resource 2: All of Statistics

*(“All of Statistics by Wasserman, Springer Science & Business Media, 2004, 978-1-4419-2322-6”)*

Resource 3: USC Courses

*(GSBA 545 course at USC, Professor: Tony Lin)*

# Exploratory Data Analyst

## Describing Data

1. Data Types
   1. Numeric
      1. Continuous
      2. Discrete
   2. Categorical
      1. Binary
      2. Ordinal
         1. Has an explicit ordering
2. Rectangular Data
   1. Data frame
   2. Feature (column)
      1. Attribute, input, predictor, variable
   3. Records (row)
      1. Case, example/instance, observation, pattern, sample
   4. Outcome (Y)
      1. Dependent variable, response, target, output
3. Estimates of Location
   1. Mean
      1. Mean
      2. Weighted mean
      3. Trimmed mean - Average of value after dropping a fixed a number of extreme values
   2. Median
      1. Feature: Robust – Not sensitive to extreme values (outliers)
   3. Mode
      1. Most commonly occurring category/value
   4. Outlier
4. Estimates of Variability (Dispersion)
   1. Measure whether the data values are tightly clustered or spread out
   2. Deviations
      1. Difference between the observed values and the estimate of location
      2. Synonyms: errors, residuals
   3. Variance
      * 1. Degree of freedom - The number of independent values that a statistical analysis can estimate
        2. Df = n-1 when using samples
           1. If use n 🡪 underestimate the true value (biased estimate)
        3. Standard deviation 🡪 the square root of the variance
      1. Mean absolute deviation
      2. Median absolute deviation
   4. Range
      1. Interquartile Range 🡪 Difference between 75th and 25th
   5. Percentile
      1. The value such that P percent of the values <= this value or less and (100-P) percent >= this value
5. Measures of symmetry
   1. Skewness
      1. Skewness to right 🡪 K > 0, mean > median
   2. Tail
      1. The long narrow portion of a frequency distribution, where relatively extreme values occur at low frequency

## Probability

1. Terminology
   1. Sample space – The set of possible outcomes of an experiment.
   2. Event – A possible outcome from a random process
   3. Probability – Percentage of time the event is expected to happen
      1. P(A) = Event / Sample Space
2. Rule One
   1. P(S) = 1
   2. For any event A, P(A) ∈ [0, 1]
   3. The complement of A
      1. P(AC) = 1 – P(A)
   4. The union of events A and B
      1. A or B
      2. A ∪ B
   5. The intersection of A and B
      1. A and B
      2. A ∩ B or AB
3. Rule Two
   1. For any events A and B
      1. P(A or B) = P(A) + P(B) – P(A and B)
   2. If event A and B are mutually exclusive (disjoint)
      1. P(A and B) = ∅
   3. If events A and B are independent
      1. P(A and B) = P(A) \* P(B)
   4. Conditional probability – The chance that the event A occurs among the event B has already happened
      1. P(A|B)
         1. If P(B) > 0
      2. For any pair of events A and B
         1. P(A and B) = P(A|B)P(B) = P(B|A)P(A)
      3. For independent events
         1. P(A|B) = P(A) and P(B|A) = P(B)
   5. Choose k elements from n – how many distinct ways?

|  |  |  |
| --- | --- | --- |
|  | Order Matter | Order Doesn’t Matter |
| Without Replacement |  |  |
| With Replacement |  |  |

1. Bayes’ Theorem
   1. Let A1, …, Ak be a partition of sample space such that P(Ai) > 0 for each i. If P(B)>0 then for each i = 1,…,k
      1. The following example is when k = 2
      2. A1 🡪 A & A2 🡪 Ac

# Data and Sampling Distributions

## Random variable

1. A process that generates numerical outcomes with known probabilities
   1. Example: Flip a fair coin twice and let X be the number of heads
      1. X is the random variable and its distribution P(X=x)
      2. X 🡪 P(X=x)
      3. 0 🡪 0.25, 1 🡪 0.5, 2 🡪 0.25
   2. Cumulative distribution function (CDF)
      1. The frequency with which the random variable assumes values less than the input
      2. P (X <= x)
      3. Rules
         1. P(x < X <= y) = F(y) – F(x)
         2. P(X > x) = 1 – F(x)
   3. Probability mass function (PMF)
      1. P (X = x)
   4. Probability density function (PDF)
      1. Specify the probability of the random variable (continuous) falling within a particular range of values, as opposed to taking on any one value.
2. Expected Value – Average value of all possible outcomes
3. Standard deviation (SD) – the amount of variability for a subject set of data from the mean
   1. Standard error of the mean (SEM) – measures how far the sample mean of the data is likely to be from the true population mean – smaller than SD

## Binomial random variables

1. Discrete Random Variable
2. Bernoulli trial – A random variable that results in either a “success” or “failure”
   1. Binomial random variable – A random variable that counts the number of “successes” that occur
3. Sums and percentages for independent Bernoulli trials
   1. Assumptions
      1. n independent trials
      2. Chance of “success” each time: p
   2. Formulas
      1. X = total number of successes
      2. E(X) = np
   3. Proportion of successes
      1. = proportion of “successes” in n trials

## Multiple Random Variables

1. Independent random variables 🡪 Xi’s
   1. Sum of n random variables 🡪 Sn = X1 + X2 + … + Xn
      1. E(Sn) = nE(X)
      2. SE(Sn) = sqrt(n) \* SE(X)
   2. Average of n random variables
2. Dependent random variables
   1. Correlation is not causation
   2. Correlation (ρ) 🡪 measures how the outcomes of X are related to the outcomes of Y
      1. Corr(X, Y) 🡪 Correlation coefficient
      2. Range 🡪 [-1, 1]
      3. Positively / negatively correlated
   3. Covariance (σXY) 🡪 measure of the strength of the correlation between two or more sets of random variables
      1. Cov(X,Y) = corr(X, Y) \* SE(X) \* SE(Y)
      2. Always between -1 and +1
   4. If X and Y are independent random variables
      1. Cov(X, Y) = corr(X, Y) = 0
3. Random variables with different distributions
   1. Example where aX+bY can be used:
      1. Sum of X and Y: a=1, b=1
      2. Difference between X and Y: a=1, b=-1
      3. Average of X and Y: a=0.5, b=0.5

## Random Sampling and Sample Bias

1. Population – The larger data set or idea of a data set
   1. Parameter = probability of all population (unknown)
2. Sample – A subset of the population
   1. Random sample – drawing elements into a sample at random (every item had a known probability of being selected)
      1. Avoid the problem of sample bias
      2. Can either do or not do replacement
   2. Simple random sample – a random sample where every item had an equal probability of being selected (without stratifying)
      1. Stratified sampling – dividing the population into strata and randomly sampling from each strata
   3. Sample statistic – numerical fact about the sample
   4. Sample bias – A sample that misrepresents the population

## Sampling Distribution of a Statistic

1. Central Limit Theorem
   1. Explanation
      1. If n is “large”, the distribution of possible values for Sn will have a normal distribution, even if the outcomes for the original X’s don’t
      2. Make sure X’s are independent
      3. The tendency of the sampling distribution to take on a normal shape as sample size rises
   2. Application: binomial distribution (1 or 0)
      1. CTR (click-through rate) = p
      2. Sample: 1000 users
      3. => MEAN = np
      4. => VAR = np(1-p)
2. Standard Error
   1. The variability (sd) of a sample statistic over many samples
      1. SD refers to variability of individual data values
   2. As the sample size increases, the standard error decreases
   3. We usually use bootstrap to estimate standard error

## Confidence Intervals

1. Two types of statistical inference
   1. Tests of statistical significance (aka hypothesis tests)
      1. Seeing if the unknown population parameter might be equal to a certain value
   2. Confidence intervals
      1. Making “reasonable guesses” for the unknown population parameter
2. Explanation for CI
   1. The probability that the true value lies within a certain interval
   2. Ex: 95% confidence interval 🡪 95% of all population who XXX is a% to b%

1. Confidence level 🡪 95 % above 🡪 used to calculate Z
   1. The percentage of confidence intervals
   2. The chance that the sample proportion (which is random) was “close to” the population parameter
   3. The higher the level of confidence, the wider the interval
2. Bootstrap is an effective way to construct confidence intervals

## Normal (Gaussian) distribution

1. Estimate percentiles from the average and SD -> Know the shape of the distribution
   1. Means controls where the center is
   2. Standard deviation controls the spread
2. Key Terms
   1. Error – the difference between a data point and a predicted or average value
   2. Standardize – subtract the mean and divide by the standard deviation
   3. Z-score – the result of standardizing an individual data point
   4. Standard normal distribution
   5. QQ-plot – A plot to visualize how close a sample distribution is to a normal distribution
      1. A scatterplot with the observed quantities on the x-axis and the distribution’s quantiles on the y-axis
      2. Prefer fall on the diagonal straight line 🡪 closed to normal
3. Factors
4. Symmetric
   * 1. For and x, P(Z > x) = P(Z < -x)
     2. Median = mean
5. Range
   * 1. 68% of data are within +/- SD of mean
     2. 95% of data are within +/- 2 SD of mean
     3. 99.7% of data are within +/- 3 SD of mean
6. Finding the percentage less than value x
7. Finding the value x where there is p% of data smaller than x
8. Student’s t-Distribution
   1. Normal distribution 🡪 T distribution when N<30
      1. Also when you don’t know the σ
9. We can always use t-test, we just can’t use z-test when n<30
10. T distribution has only one parameter: degree of freedom (df = N-1)
    * + 1. Degrees of freedom 🡪 allow the t-distribution to adjust to different sample sizes, statistics, and number of groups
11. Central limit theorem 🡪 The tendency of the sampling distribution to take on a normal shape as sample size rises
12. Sample statistics are often normally distributed
    1. Confidence interval
       1. Normal distribution
       2. T distribution
    2. We use t-distribution if we don’t know the population variance
       1. T-distribution is more conservative
       2. We use more Z for Bernoulli because we know the Var[X] = np(1-p)

## Common Distributions

1. Binomial Distribution - Distribution of number of successes in x trials
   1. Key ideas
      1. With two possible outcomes; one with probability p and the other with probability 1-p
      2. With large n, and provided p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution
   2. Bernoulli
      1. Each customer has same 10% to buy
         1. P = 10%
         2. E(X) = p = 10%
         3. Var(X) = p(1-p) = 0.09
   3. Binomial
      1. 20 customers come in, how many people going to buy?
         1. E(X) = np = 20 \* 0.1 = 20%
         2. Var(X) = np(1-p) = 20 \* 0.09 = 1.8
   4. Negative Binomial
      1. How many people until 20th sales?
         1. E(X) = k/p = 20/0.1 = 200
         2. Var(X) = k(1-p)/(p^2)
   5. Geometric
      1. How many customers until the first sale? 🡪 k = 1
         1. E(X) = 1/p = 1/0.1 = 10
         2. Var(X) = (1-p)/(p^2)
2. Poisson and related distributions
   1. Key terms
      1. Lambda – the mean number of events that occurs in a specified interval of time or space
         1. The variance for a poisson distribution is also lambda
      2. Exponential distribution – The frequency distribution of the time or distance from one event to the next event
      3. Poisson distribution – the frequency distribution of the number of events in sampled units of time or space
   2. Exponential
      1. Mean 🡪 1/lamda; Variance 🡪 1/(lamda^2)
      2. Average 5 minutes per customers
      3. E(X) = 5 🡪 lamda = 1/E(X) = 1/5
      4. When we create new random variable, we can use rand() to create F(X) 🡪 use F(X) and lamda to calculate x
   3. Poisson-Exponential
      1. How many customers come 60 minutes?
      2. E(X) = lamda \* t = 1/5 \* 60 = 12
   4. Gamma
      1. How long until 100 customers coming? (sum of exponential)
      2. E(X) = n/lamda = 100/(1/5) = 500

# Statistical Experiments and Significance Testing

## A/B Testing

1. Key Terms
   1. A/B testing is the process of comparing two variations of a page element, usually by testing users' response to variant A vs variant B, and concluding which of the two variants is more effective.
   2. Randomly assigning subjects to treatments
   3. Subject are assigned randomly to the groups
      1. Subject 🡪 The items (web visitors, patients, etc) that are exposed to treatments
      2. Control group 🡪 No treatment / standard treatment
         1. We need the control group because we need to keep “other things are equal”
      3. Treatment group 🡪 Specific treatment
   4. Use test statistic as the metric to measure the effect of the treatment
      1. Binary variable 🡪 buy or don’t buy / click or no-click / fraud or no fraud

## Hypothesis Test

1. Key Terms
   1. Hypothesis test (significance test)
      1. Learn whether random chance might be responsible for an observed effect
         1. Must be random chance in assignment of subjects
         2. Must a true difference between groups A and B
      2. Create
   2. Null hypothesis 🡪 A statement that the data are due only to luck
      1. We want to prove the null hypothesis wrong
   3. Alternative hypothesis 🡪 A statement that describes that there is a relationship between two selected variables (not luck)
2. Resampling
   1. Resampling
      1. The process of taking repeated samples from observed data; includes both bootstrap and permutation (shuffling) procedures
   2. Bootstrap sample
      1. A sample taken with replacement from an observed data set
   3. Permutation test
      1. The procedure of coming two or more samples together, and randomly (or exhaustively) reallocating the observations to resamples
         1. An exhaustive permutation test
            1. All the possible ways it could be divided 🡪 only for small sample sizes
         2. An bootstrap permutation test
            1. With replacement

## Statistical Significance and P-Values

1. Key terms
   1. P-value
      1. The probability of getting new results MORE EXTREME than the actual data, if the null hypothesis were true.
      2. Probability of obtaining a result at least as unusual / extreme, given that the null hypothesis was assumed
      3. If p is “small”, we reject the null hypothesis and say it’s not luck
   2. Alpha
      1. Significance level 🡪 The probability of study rejecting the null hypothesis, given that the null hypothesis was assumed
      2. Or “what is the probability of a result this extreme?”
   3. Statistically Significant
      1. When p-value < 0.05 = alpha
   4. Type Error
      1. Type I error 🡪 Incorrectly rejects a true null hypothesis
      2. Type II error 🡪 Fails to reject a null hypothesis which is really false
   5. Test statistic - a measure of how incompatible the data are with the null hypothesis
      1. Hypothesis Test --- Test Statistic
         1. Z-Test 🡪 Z-score
         2. T-Test 🡪 T-score
         3. ANOVA 🡪 F-statistic
         4. Chi-Square Test 🡪 Chi-square statistic

## Hypothesis tests Example

1. For one population proportion
   1. Recall: sample proportion successes observed among n independent with population proportion p
   2. Process
      1. H0 🡪 population proportion = p0
      2. Calculate
         1. Compute
      3. If p-value < 0.05 => reject H0, can’t be explained by luck
2. For two population proportions
   1. Example:
      1. You have two large simple random samples from populations of “yes/no” data
      2. You wonder, could the two populations have the same percentage of “yes”?
   2. Process
      1. H0 🡪 p1=p2
      2. p-value
      3. Make a conclusion based on p-value
   3. Confidence Intervals

## ANOVA (Analysis of Variance)

1. Definition:
   1. Instead of an A/B test, you have multiple simple random samples from various populations (A-B-C-D)
   2. You believe all populations are normally distributed with the same population SDs
   3. You want to know if the population means are all equal, or if at least one is different
2. F test for equality of multiple means (AVOVA)
   1. H0: all k populations have same mean and sd
   2. Collect data

* 1. Calculate “Sum Squares” (SS):
  2. Calculate “Mean Squares” (MS):
     + 1. Variance across group means
       2. Variance due to residual error
  3. The higher F, the lower p-value

1. F test for equality of two standard deviations
   1. Definition
      1. You have two random samples from normally distributed populations
      2. You have sample SD’s s1 and s2
      3. You want to know if the population SD’s s1 and s2 are equal
   2. Step:
      1. State H0: pop. SD’s are equal, s1 = s2
      2. Get sample SD’s s1 and s2
      3. Compute
      4. Get p-value by comparing calculated Fdf1,df2 distribution, where
         1. df1 = (sample size used to calculate s1) – 1
         2. df2 = (sample size used to calculate s2) – 1
      5. Make a conclusion based on p-value

## chi-square (χ2) Test

1. Key terms
   1. Chi-square statistic
      1. A measure of the extent to which some observed data departs from expectation
   2. Expectation or expected
      1. How we would expect the data to turn out under some assumption, typically the null hypothesis
   3. The chi-square (χ2) random variable
      1. Define Z1, Z2, … , Zk ~ k independent standard normal random variables. Then Ck = Z12 + Z22 + … + Zk2 has a chi-square (χ2) distribution with k degree of freedom??? (Not understand)
2. The χ2 test of independence
   1. When to use:
      1. You have two or more samples from various populations; each sample has two or more categories of outcomes
      2. You want to know if all populations have the same distribution (i.e., chances are independent of population)
   2. Calculation Process

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. State H0: variables are independent | | | | |
| Observed Counts | **Headline A** | **Headline B** | **Headline C** |
| Click | 14 | 8 | 23 |
| No-Click | 986 | 992 | 988 |
| N | 1000 | 1000 | 1000 |
| Sum(Click) / sum(No-click) = 34/3000 = 11.33% | | | |
| 1. **Calculate Expected Value** | | | | |
| Expected Counts | **Headline A** | **Headline B** | **Headline C** |
| Click | 11.33 | 11.33 | 11.33 |
| No-Click | 988.67 | 988.67 | 988.67 |
| N | 1000 | 1000 | 1000 |
| 1. **Check that all E’s > 5; if not, group categories** | | | |
| 1. **Calculate the Pearson residual 🡪** | | | |
| Parson Residual | **Headline A** | **Headline B** | **Headline C** |
| Click | 0.792 | -0.990 | 0.198 |
| No-Click | -0.085 | 0.106 | -0.021 |
| 1. **Computer chi-square statistic ( C )** | | | |
|  | | | |
| 1. **df = (r-1) \* (c-1) = (3-1) \* (2-1) = 2 df** | | | |
| 1. **p-value = CHIDIST(1.666, 2) = 0.4348** | | | |
| 1. **Reject H0: Not independent** | | | |

1. Inference for standard deviation of a normal population
   1. (1-α)% confidence interval for σ?
2. Inference with F distribution??? (Don’t understand)
   1. Let Cn and Cd be independent chi-squared random variables with degrees of freedom equal to n and d, respectively
   2. Then F = (Cn / n) / (Cd / d) is said to have an F distribution with n, d degrees of freedom
   3. Properties of the Fn,d distribution:
      1. F > 0
      2. If d is large, E(F) ~ 1
      3. As n, d both 🡪 infinity, SE(F) 🡪 0

------------------------------------------------------------------------------------------------

## Finding Sample Size

1. K
   1. want the SE to be no larger than k
   2. #% confidential interval to have width of +/- K

## Inference for Averages

## Inference for two matched populations

1. Matched-pair data, there are two samples, but every item in the first sample can be logically matched up with exactly one item in a second sample
2. Take differences between the paired values as a single population

## Inference for two independent populations

1. H0: no effect, difference = 0
2. Where σ’s might be different
   1. Calculate
   2. Df = smaller of (n1-1) and (n2-1)
3. Where σ’s are known to be equal
   1. When we might believe that the population SD’s must be equal
      1. Draw two samples from the same population
      2. We think that two population are identical
   2. Calculate
   3. Df = n1 + n2 – 2