MET CS688 C1

Web Analytics and Mining

Zlatko Vasilkoski

Statistical Methods With R

Statistical Methods with R

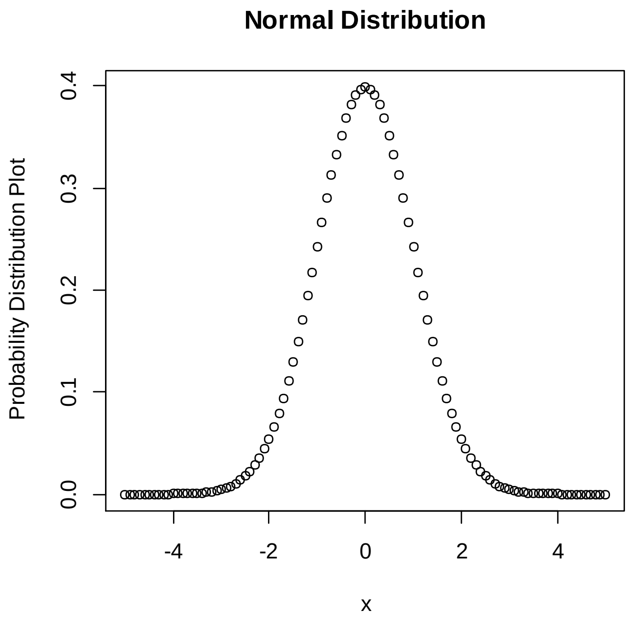
Probability in R

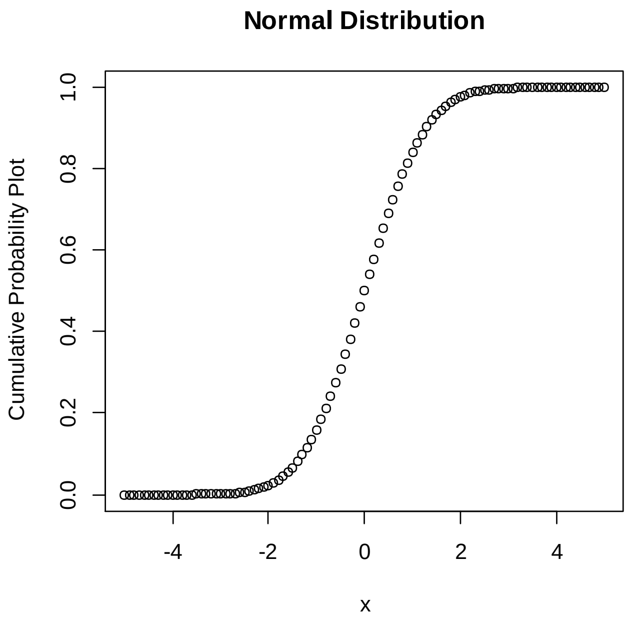
- Several probability distributions
  - Normal Distribution
    - dnorm(x) – Generates Probability Distribution for sequence x.
    - pnorm(x) – Generates Cumulative Probability for sequence x.
    - qnorm(x) – Find Probability from given Cumulative Distribution x (gives 0 for x=0.5).
    - rnorm(n) – Generates n random numbers according to the Normal Distribution.
  - Binomial Distribution
    - dbinom – Generates Probability Distribution
    - pbinom – Generates Cumulative Probability
    - qbinom – Find Probability from given Cumulative Distribution
  - Chi-Squared Distribution
    - dchisq – Generates Probability Distribution
    - pchisq – Generates Cumulative Probability
  - T Distribution
    - dt – Generates Probability Distribution
    - pt – Generates Cumulative Probability
    - qt – Find Probability from given Cumulative Distribution

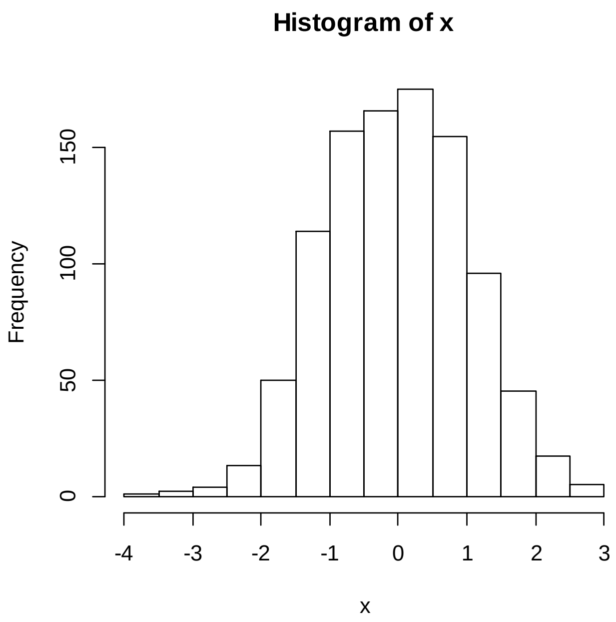
```r
# Probability
rm(list=ls()); cat("\014") # clear all
# Normal Distribution ====
x <- seq(from=-5, to=5, by=0.1); y <- dnorm(x); plot(x,y) # Probability Distribution Plot
x <- seq(from=-5, to=5, by=0.1); y <- pnorm(x); plot(x,y) # Cumulative Probability Plot
Prob <- 0; x <- pnorm(Prob);  Prob==qnorm(x) # Find Probability from Cumulative Distribution
x <- rnorm(1000); hist(x) # Generated random numbers according to the Normal Distribution

# T Distribution ====
x <- seq(from=-5, to=5, by=0.1); y <- dt(x,df=10); plot(x,y) # Probability Distribution Plot
x <- seq(from=-5, to=5, by=0.1); y <- pt(x,df=10); plot(x,y) # Cumulative Probability Plot
Prob <- 0; x <- pt(Prob,df=10);  Prob==qt(x,df=10) # Find Probability from Cumulative Distribution
x <- rt(1000,df=10); hist(x) # Generated random numbers according to the T Distribution

# Binomial Distribution ====
Ntrials <- 100 # number of trials
Prob <- 0.7 # probability of success for a single trial.
x <- seq(from=0, to=Ntrials, by=1); y <- dbinom(x, size=Ntrials, prob=Prob); plot(x,y) # Probability Distribution Plot
x <- seq(from=0, to=Ntrials, by=1); y <- pbinom(x, size=Ntrials, prob=Prob); plot(x,y) # Cumulative Probability Plot
Prob1 <- 0.7; x <- pbinom(Prob1, size=Ntrials, prob=Prob1);  Prob1==qbinom(x, size=Ntrials, prob=Prob) # Find Probability from Cumulative Distribution
x <- rbinom(1000, size=Ntrials, prob=Prob); hist(x) # Generated random numbers according to the Binomial Distribution

# Chi-Squared Distribution ====
x <- seq(from=-5, to=5, by=0.1); y <- dchisq(x,df=10); plot(x,y) # Probability Distribution Plot
x <- seq(from=-5, to=5, by=0.1); y <- pchisq(x,df=10); plot(x,y) # Cumulative Probability Plot
```

Probability and R - Normal Distribution

- Normal Distribution
  - dnorm(x) – Generates Probability Distribution for sequence x.
  - pnorm(x) – Generates Cumulative Probability for sequence x.
  - qnorm(x) – Find Probability from given Cumulative Distribution x (gives 0 for x=0.5).
  - rnorm(n) – Generates n random numbers according to the Normal Distribution.



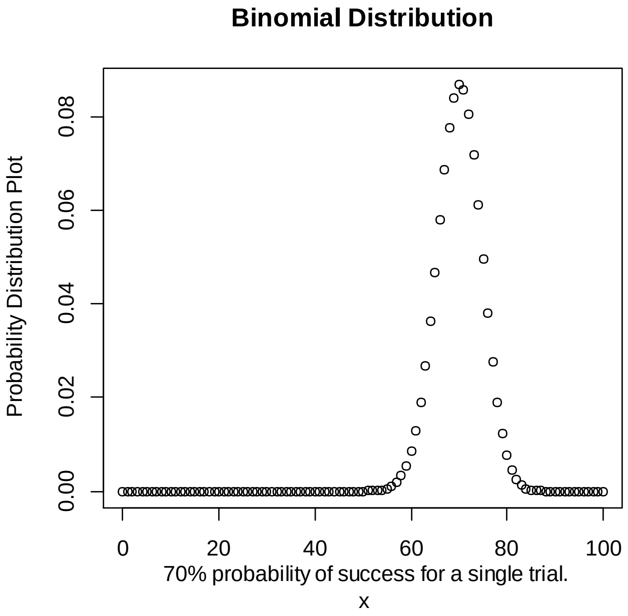


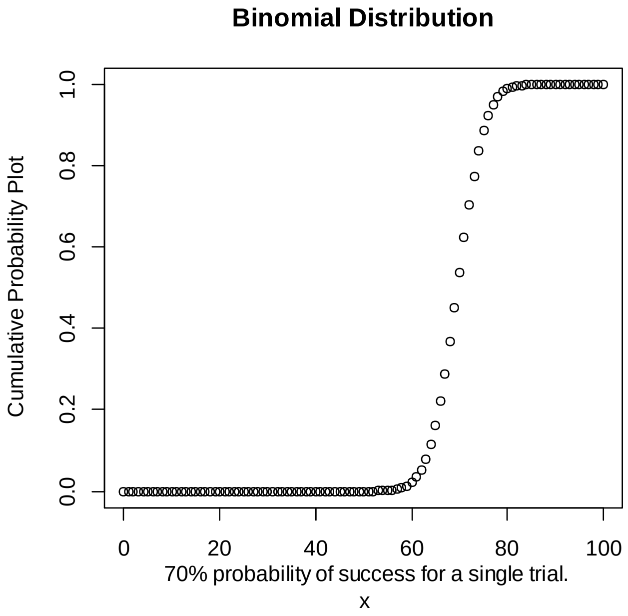


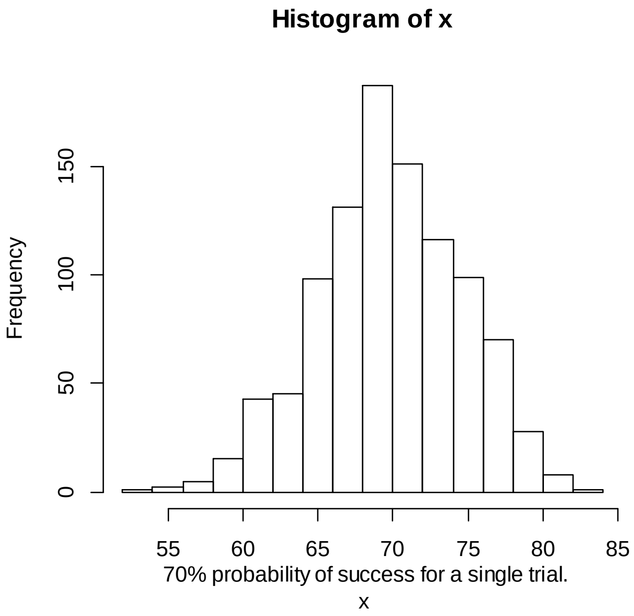
Probability and R - Binomial Distribution

- Binomial Distribution $P(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
  - dbinom – Generates Probability Distribution.
  - pbinom – Generates Cumulative Probability.
  - qbinom – Find Probability from given Cumulative Distribution.
  - rbinom (n) – Generates n random numbers according to the Binomial Distribution.

p – probability of a **successful** event
k – number of **successful** events
n – total number of **independent** trials



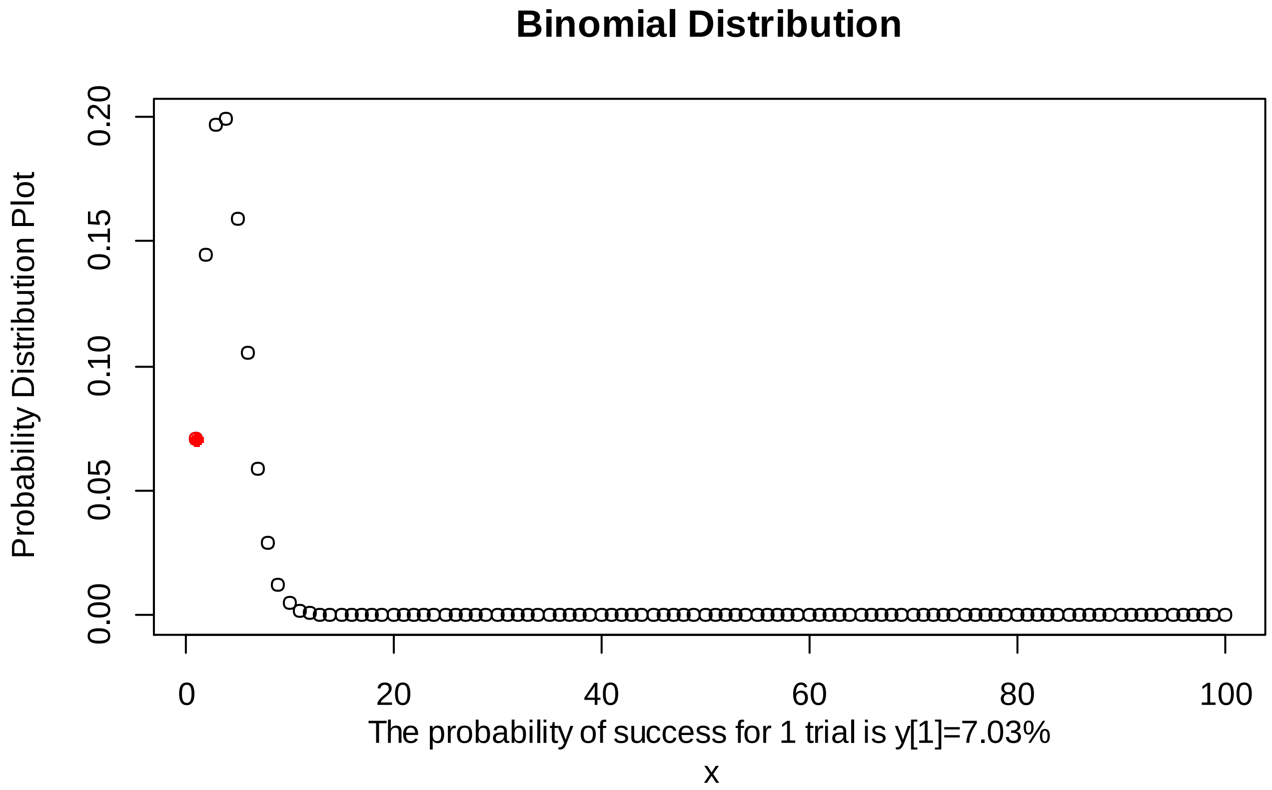




Binomial Distribution Example with R

Example: Newsfeed & Binomial Distribution

- Newsfeed and ads stories are being shown to a customer. Every story within a Newsfeed has a 4% chance of being an ad. Assume the probability distribution of ads being shown to a customer is binomial. What is the chance a user will be shown only a single ad in 100 stories?
- To find the answer use Binomial Distribution (dbinom) with
  - Ntrials (n) = 100 # number of trials (stories)
  - Prob (p) = 4/100 # probability of success for a single trial
  - Nevent (k) = 1 # shown only a single ad (**successful** event)
- dbinom(Nevent, size=Ntrials, prob=Prob) = 7.03%

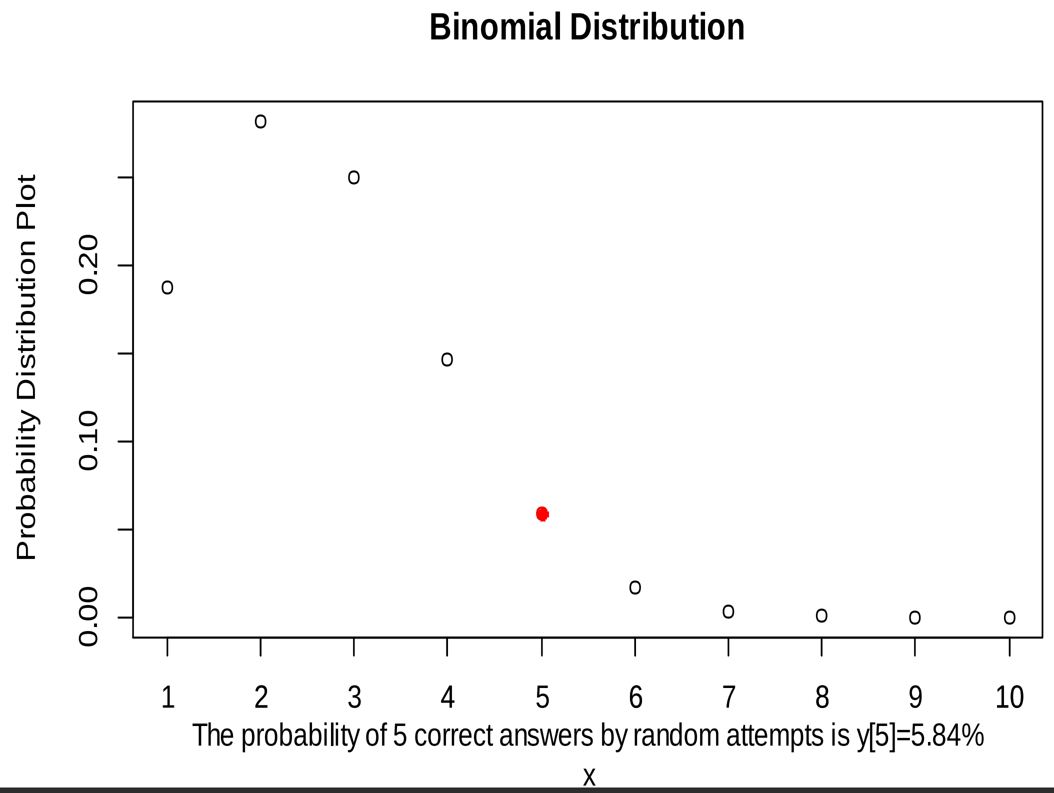



```r
Ntrials <- 100 # number of trials (stories)
Prob <- 4/100 # probability of success for a single trial
Nevent <- 1 # shown only a single ad
yy <- dbinom(Nevent, size=Ntrials, prob=Prob) # 7% chance a user will be shown only a single ad in 100 stories
x <- seq(from=1, to=Ntrials, by=1); y <- dbinom(x, size=Ntrials, prob=Prob);
plot(x,y, main = "Binomial Distribution", ylab = "Probability Distribution Plot") # Probability Distribution Plot
points(Nevent, yy, col='red', pch = 19)
strg <- paste0("The probability of success for ", Nevent, " trial is y[", Nevent,"]=", round(y[Nevent]*100, 2),"%")
mtext(strg, side=1, line=2)
print(strg)
```

Binomial Distribution Example with R

Quiz Example:

- There were 10 multiple choice questions on a Quiz. Each question has 4 possible answers, and only one of them is correct. Find the probability of having 5 correct answers if you attempt to answer every question at random.
- To find the answer use Binomial Distribution (dbinom) with
  - Ntrials = 10 # Number of multiple-choice questions
  - Prob = 1/4=0.25 # probability of one correct answers out of 4
  - Nevent = 5 # Number of correct answers at random
- dbinom(Nevent, size=Ntrials, prob=Prob) = 5.84%
- The probability of 5 correct answers by random attempts is y[5]=5.84%.
- This means 6 students out of 100 will have 1/2 of the questions answered correctly.



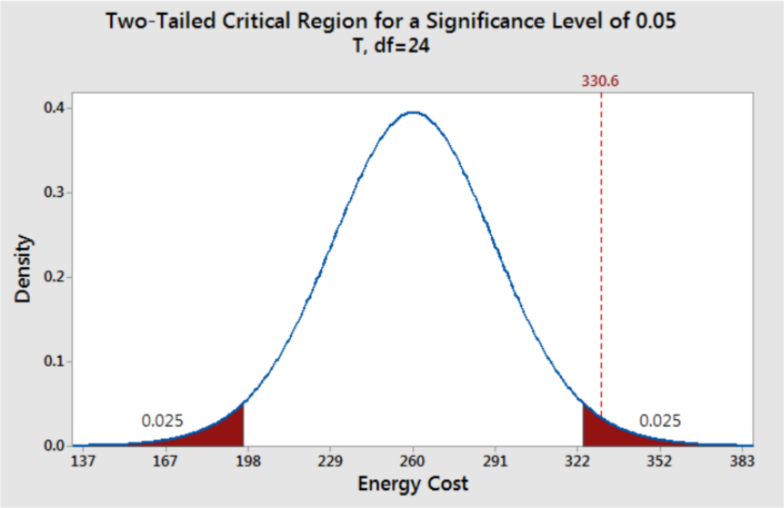

```r
Ntrials <- 10 # Number of multiple-choice questions
Prob <- 1/4 # probability of one correct answers
Nevent <- 5 # Number of correct answers at random
yy <- dbinom(Nevent, size=Ntrials, prob=Prob) # 0.06=6% or 6 students out of 100 will have 1/2 of the questions answered correctly
x <- seq(from=1, to=Ntrials, by=1); y <- dbinom(x, size=Ntrials, prob=Prob);
plot(x,y, main = "Binomial Distribution", ylab = "Probability Distribution Plot") # Probability Distribution Plot
axis(side = 1, at = 1:Ntrials)
points(Nevent, yy, col='red', pch = 19)
strg <- paste0("The probability of ", Nevent, " correct answers by random attempts is y[", Nevent,"]=", round(y[Nevent]*100, 2),"%")
mtext(strg, side=1, line=2)
print(strg)
```

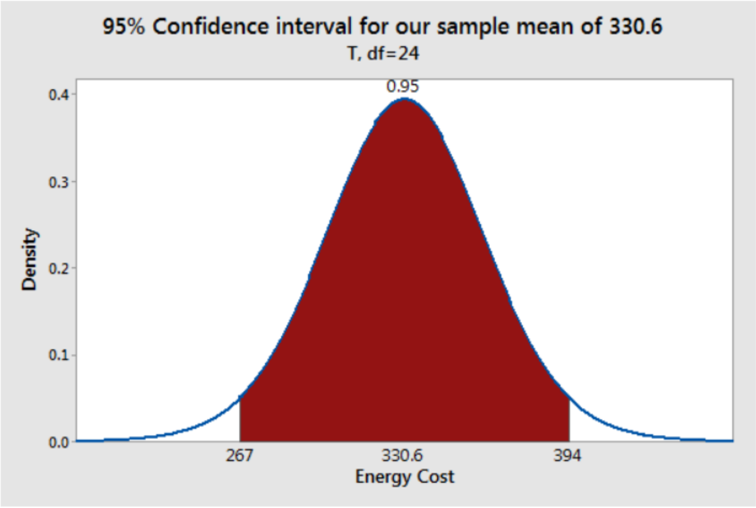
Statistical Methods - Hypothesis testing

Hypothesis testing. Retain or reject hypothesis based on measurements of observed samples. The decision is often based on a statistical mechanism called hypothesis testing.

- For example, determine whether the means of two groups are equal to each other.
    - The **null hypothesis** is that the two means are equal.
- Retaining or rejecting the hypothesis is based on the **P-value**.
    - The **p.value** is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.
    - If **p.value < 0.05** then **Reject the Null Hypothesis**. The analyzed sample is statistically significant (sample statistic is unusual enough relative to the null hypothesis).
    - Conclusion: **Means are not the same.**

- **Type 1 error** (false positives) - falsely rejecting a null hypothesis when the null hypothesis is true.
- α - **significance level** of hypothesis testing - the probability of committing a type 1 error.
- **Type 2 error** (false negatives)





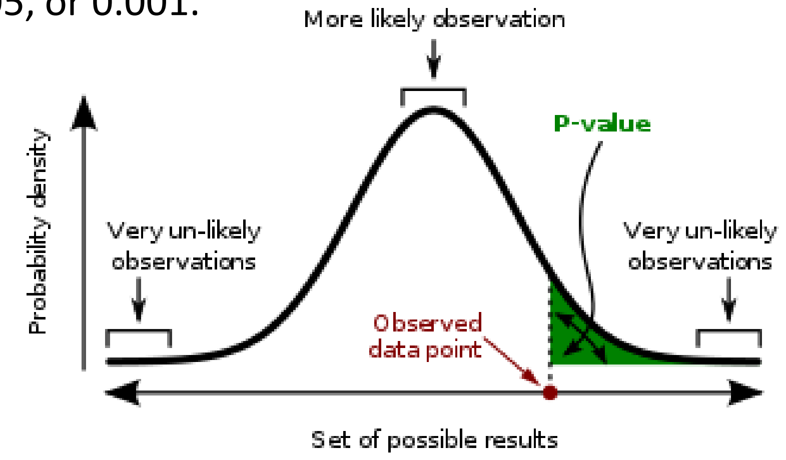
Hypothesis testing – p Value

p Value - The probability of observing a more extreme result assuming the null hypothesis is true.

- Used in the context of null hypothesis testing in order to quantify the idea of statistical significance of evidence (reductio ad absurdum argument).
- **The smaller the p-value**, **the higher the significance** because
  - Observed outcome would be very unlikely under the null hypothesis.
  - It tells the investigator that the null hypothesis may not adequately explain the observation.
- The null hypothesis is **rejected** if the p-Value (probability) is less than or equal to a small, but arbitrarily pre-defined threshold value (**level of significance**).
  - **If $p \leq \alpha$ reject the Null Hypothesis.**
  - The level of significance α is arbitrary but commonly set to 0.05, 0.01, 0.005, or 0.001.

Example:
A **null hypothesis** states that a certain statistics **T** is a normal distribution **N(0,1)** (with mean zero and variance 1).

- We **retain** the null hypothesis if there is <u>NO significant difference</u> between **T** and **N** distributions!
- The **rejection** of this null hypothesis (sufficient evidence against it) could mean that
  - The mean of T is not zero or
  - the variance of T is not 1 or
  - T is not normally distributed



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Hypothesis testing Example

Example: Use t-Test to determine whether the means of two sequences x and y are equal to each other.

- Generate 2 random numbers sequences x and y, according to the Normal Distribution.
  - rnorm(n)
- Take the null hypothesis to be:
  - **The two means of x and y are equal.**
- Obtaining a p-value that is > 0.05 we fail to reject the null hypothesis.

```
# 1) Hypothesis testing ====
# t-Test - Determine whether the means of two groups are equal to each other.====
# The null hypothesis is that the two means are equal.
x = rnorm(10); y = rnorm(10)
t.test(x,y)
```

```
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = 1.2569, df = 16.856, p-value = 0.2259
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.325012  1.281347
sample estimates:
 mean of x  mean of y
 0.1155109 -0.3626568
```

Statistical Methods - Statistical inference

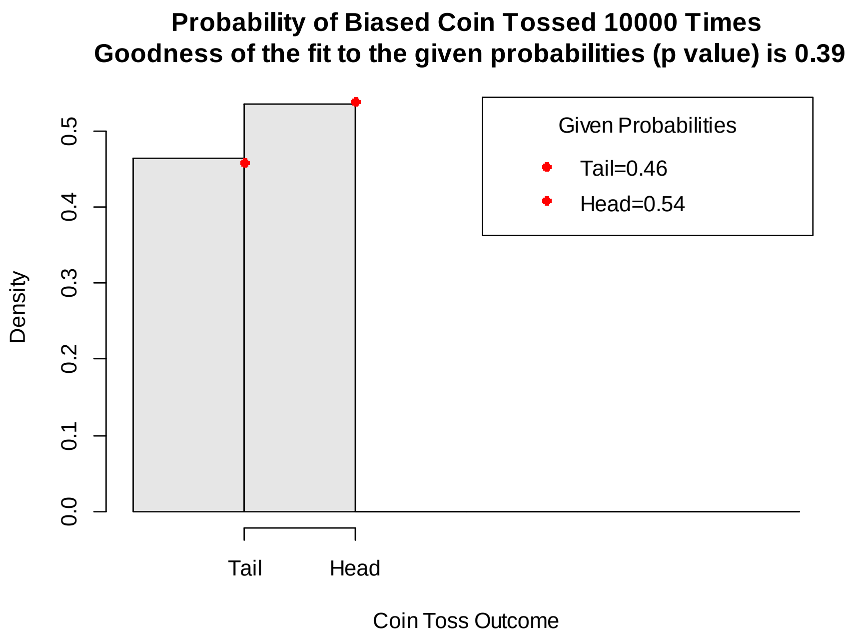
- ANOVA - a statistical tool used in several ways to develop and confirm an explanation for the observed data. Exploratory data analysis, employs an additive data decomposition.
- ANOVA Null Hypothesis (assumption): Y has normal distribution for each categorical group in X.
- ANOVA is useful for comparing (testing) three or more means (groups or variables) for statistical significance.

- ANOVA Example
  - Study effects of tea (X) on weight loss (Y).
    - individuals randomly split into smaller groups and drinking different kinds of tea (X).
    - individuals (X) split into groups based on an attribute they possess.
  - Blood pressure Y among 3 age groups X.
    - ANOVA Null Hypothesis is: Y has normal distribution for each category (age group).
    - Check for normality
    - Check for equal variance
    - ANOVA uses F (Fisher) – statistic, which is simply a ratio of two variances (how far is the data are scattered from the mean).
    - For one-way ANOVA, the ratio of the between-group variability to the within-group variability follows an F-distribution when the null hypothesis is true.

Hypothesis Testing Examples: Biased Coin

Example:

- A coin is sold for cheaters which is weighted unevenly so that the probability of a head is 0.54 versus the probability of a tail which is 0.46.
- Create a code that can prove/disprove the hypothesis if a coin is biased.
- Generate 10,000-coin tosses and create a plot of the probability.




```r
# Answer:
# Create a vector of biased probabilities and their names.
Pnames <- c("Tail","Head")
P <- c(0.46, 0.54) # Given Biased Probabilities

# Use R's function "sample()" to generate the 10,000 coin tosses.
# Using sampling with replacement and specify the vector of
# biased probabilities "P" as arguments of the function "sample()".

Num.Samples <- 1e4
throws <- sample(1:2, Num.Samples, replace=TRUE, prob=P )
TT <- table(throws)
# throws
# tails   heads
# 4643    5357

TTP <- table(throws)/Num.Samples # Coin Toss Probabilities
# throws
# tails   heads
# 0.4643  0.5357

# Tests the goodness of the sample fit to the given probabilities using chi-squared test.
ChiSq <- chisq.test(TT, p = P)
# Chi-squared test for given probabilities
ChiSq
# data:  TT
# X-squared = 0.74436, df = 1, p-value = 0.3883

# The P value is 0.38 (not bellow 0.05) - so the fit to the given probabilities is good.
P.Value <- round(ChiSq$p.value,2)

# Histogram of the distribution
hist(throws, probability=TRUE, col=gray(.9), breaks=c(0:6),
     main=paste0("Probability of Biased Coin Tossed ",Num.Samples," Times",
                 "\n Goodness of the fit to the given probabilities (p value) is ",P.Value),
     xlab = "Coin Toss Outcome")
points(P, col = "red", pch=16)
legend("topright", paste0(Pnames,"=",P),
       pch = 16, col = "red", title = "Given Probabilities", inset = .02)
```

Probability & Bayesian Theorem

Few different formulations

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Probability someone who definitely has an allergy would make the claim that they do.

Probability that someone actually has a food allergy give they say they do.

General probability that someone has a food allergy

$$P(A|C) = \frac{P(C|A)\,P(A)}{P(C)}$$

probability someone would claim to have a food allergy

Prior Probability

$$P(A|B) = \frac{\boxed{P(B|A)\,P(A)}}{\boxed{P(B|A)P(A)} + \boxed{P(B|\neg A)P(\neg A)}}$$

Same formula as the prior probability, but this time for not-A.

Bayesian Theorem

Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available (it is a dynamic analysis of a sequence of data). Bayesian inference, where H is hypothesis and E is an observed event is:

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$

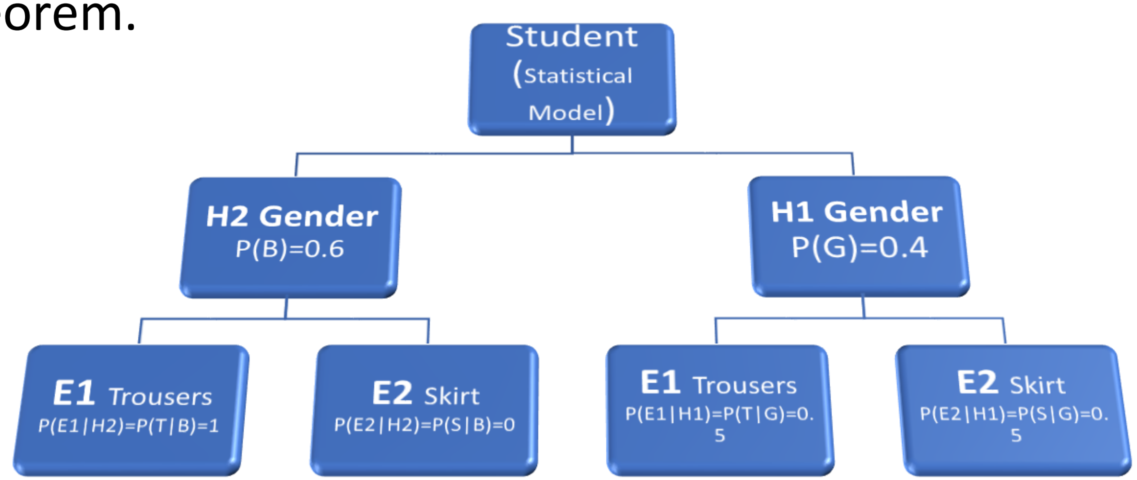
- Posterior probability $P(H|E)$ - Estimate of the probability for the hypothesis H given the observed evidence E.
- A prior probability $P(H)$ - estimate of the probability of the hypothesis H before the current evidence E observed.
- The "likelihood function" $P(E|H)$ is derived from a statistical model for the observed data.
- The evidence probability $P(E)$(marginal likelihood or "model evidence") - Prob of observing event E.

For 2 events the Bayesian theorem has the following form:

$$P(H_1|E) = \frac{P(E|H_1)}{\boldsymbol{P_{tot}(E)}} P(H_1) = \frac{P(E|H_1)}{\boldsymbol{P(E|H_1)P(H_1)+P(E|H_2)P(H_2)}} P(H_1)$$

Bayesian Theorem – Example 1

Suppose there is a mixed school having 60% boys and 40% girls as students. The girls wear trousers or skirts in equal numbers; the boys all wear trousers. An observer sees a (random) student from a distance; all the observer can see is that this student is wearing trousers. What is the probability this student is a girl? The correct answer can be computed using Bayes' theorem.



$H_1 = G\,;\, H_2 = B\,;$

$P(H_1) = P(G) = 0.4$

$P(E_1|H_1) = P(T|G) = 0.5$

$P(E_2|H_1) = P(S|G) = 0.5$

$E_1 = T\,;\, E_2 = S$

$P(H_2) = P(B) = 0.6$

$P(E_1|H_2) = P(T|B) = 1$

$P(E_2|H_2) = P(S|B) = 0$

- $P(H_1|E_1) = \frac{P(E_1|H_1)}{P_{tot}(E)} P(H_1) = \frac{P(E_1|H_1)}{P(E_1|H_1)P(H_1)+P(E_1|H_2)P(H_2)} P(H_1)$
- $P(H_1|E) = \frac{P(E|H_1)}{P(E|H_1)P(H_1)+P(E|H_2)P(H_2)} P(H_1) = \frac{0.5*0.4}{0.4*0.5\,+\,1*0.6} = 0.25$
- $P(G|T) = 0.25$

Bayesian Theorem – Example 2

You're about to get on a plane to London. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in London?

H_1-rain,

H_2 - no rain,

E = yyy (yes, yes, yes).

Assume $P(H_1) = 1/4, P(H_2) = 3/4$

- $P(E|H_1) = P(yyy|R) = \left(\frac{2}{3}\right)^3$ and $P(E|H_2) = P(yyy|!\,R) = \left(\frac{1}{3}\right)^3$
- $P(H_1|E) = \frac{P(E|H_1)}{P(E|H_1)P(H_1)+P(E|H_2)P(H_2)} P(H_1) = \frac{\left(\frac{2}{3}\right)^3 \frac{1}{4}}{\left(\frac{2}{3}\right)^3 \frac{1}{4}+\left(\frac{1}{3}\right)^3 \frac{3}{4}} = \frac{8}{11}$

Bayesian Classification

How do we implement it in our SMS Classification?

- Create Train ham/spam frequency table from Zipf's dtm containing:
  - | Term | Spam Freq | Ham Freq | Occurrence |
- Problem: What to do about the terms that are not in the training set? Simple rule:
  - assign a very small probability to terms that are not in the training set.
- Estimate the "spamminess" of a term given its context.

Another Example: Automatic Document Classification

- Look at the words used in the documents and treat the presence or absence of each word as a **feature**. This would give you as many features as there are words in your vocabulary.
- Naïve Bayes—an extension of the Bayesian classifier—is a popular algorithm for the document-classification problems.
- $P(H_i|E) = P(H_i)$ where H_i is the class i (Spam or Ham).
- Classification Score $\frac{P(E|H_i)}{P(E)}$ is based on $P(H_i|E) < P(H_j|E)$ or $P(H_i|E) > P(H_j|E)$
- Use the right side of the formula to get the value on the left. Do this for each class and compare the two probabilities.
  1. Calculate $P(H_i)$ - frequency of class i by adding up how many times that class (spam/ham) appears, normalized (dividing) by the total number of SMS.
  2. Calculate $P(E|H_i)$-rewrite it as $P(E_0|H_i) \cdot P(E_1|H_i) \cdot P(E_2|H_i) \cdots$, we can calculate probability like this if we assume that all the words are independently likely (conditional independence).
  3. Note: No need to calculate $P(E)$ since they are all the same, and we only need to compare if

     $P(H_i|E) < P(H_j|E)$ or $P(H_i|E) > P(H_j|E)$

Automatic Document Classification.

- Classification Score (no division by $P(E)$) based on $S = P(H_i|E) = P(E|H_i) \cdot P(H_i)$
  - $P(H_i)$ – Prior probability, equals the ratio of the number of Spam/Ham SMS.
  - $P(E|H_i) = P_{new} \cdot P_{occurrence}$ – Calculate as product of 2 terms
    1. $P_{new} = P_0^n$ – If new words encountered, assign to them a very small occurrence (prior) probability $P_0 = 10^{-6}$.
       - Count the number n of new Test SMS words (not present in Train data) and take $P_0 = 10^{-6}$.
    2. $P_{occurrence}$– Product of word occurrences (feature) in SMS.
       - Multiply the occurrence of the existing Test SMS words present in Train data.

```
Browse[2]> msg.match
[1] "have"      "havent"    "love"      "send"      "tomorrow" "too"       "you"
Browse[2]>
```

```
Browse[2]> msg.match
 [1] "bucks"    "couple"   "have"      "havent"   "love"      "might"    "send"      "tomorrow" "too"       "you"
Browse[2]>
```

```
> SMS.Test.Corpus[[1]]$content
[1] "i havent forgotten you i might have a couple bucks to send you tomorrow k i love ya too"
>
```