Statistics and Probability

Dr. Subrata Das

Adjunct Faculty, Villanova School of Business and Northeastern
Research and Development Consultant, MIT Lincoln Lab
Founder & President, Machine Analytics, Cambridge, MA

sdas@machineanalytics.com

https://www.linkedin.com/in/subrata-das-1293354

Outline

- Probability and Bayesian Probability
- Statistics and Hypotheses Testing
- Selected Probability Distributions
- Regression Analysis
- Cluster Analysis
- Stochastic Process Modeling

Probability Theory

- Probability provides a quantitative description of the likely occurrence of a particular event.
- Probability of an event A, denoted as $p(A)$, is conventionally expressed on a scale from 0 to 1.
- Three approaches that provide guidelines on how and what values to assign to probabilities: classical, relative frequency, and axiomatic.
- While the classical and axiomatic approaches provide guidance on how to assign values to probabilities, the relative frequency approach specifies what values to assign as follows: probability of an event *A* is defined as ratio of the outcome of *A* to "total number" of trials in a random experiment.
- Subjective probability describes an individual's personal judgment about how likely a particular event is to occur.
  - Not therefore based on any precise computation, but assessment by a subject matter expert based on his/her prior experience or perception of likelihood.

Sample Data

Outlook	Temperature	Humidity	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

Random Variables

- Random variable is a function defined over an event space and its value is determined by the outcome of an event.
- Range of a 'discrete' random variable (also called its states) is finite or denumerable:

$$Weather \in \{Sunny, Rainy, Snowy\}$$

- Probability distribution of a random variable is a function whose domain contains the values that the random variable can assume, and whose range is a set of values associated with the probabilities of the elements of the domain:

$$p(Sunny) = 0.55,\ p(Rainy) = 0.15,\ p(Snowy) = 0.3$$

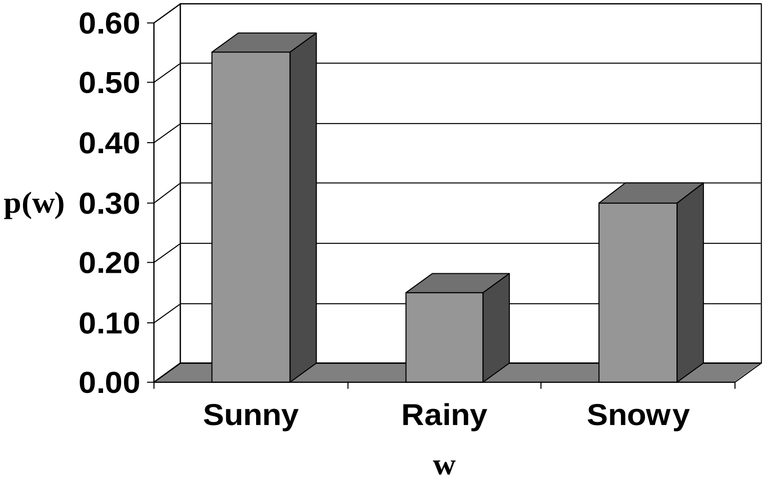
Forms of Variables

- Quantitative or numerical variables
  - Observations are measured on a continuous scale or numbers (e.g. Temperature, Height, Weight)
  - Discrete and continuous numerical variables are often differentiated by determining whether the variables are related to a count or a measurement
- Qualitative or categorical variables
  - Observations are measured on a discrete set of values (e.g. Occupation, Day of the Week, Gender, Size)
  - Nominal categorical variables has no inherent order (e.g. M, F) whereas ordinal categorical variables has an inherent rank or order (e.g. short, medium, tall).

Probability Distribution

- Discrete random variable
  - Variable X with outcomes

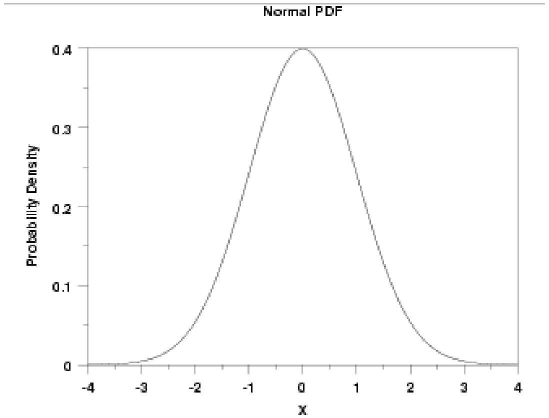
    $$X \in \{A_1, \dots, A_n\}$$
  - $$p(A_i) \geq 0, \sum_{i=1}^{n} p(A_i) = 1$$



- Continuous random variable
  - Probability density function (pdf) over continuous variable, e.g. $x \in [-4, 4]$
  - $$\int_{-4}^{+4} p(x)dx = 1$$
  - $$p(a \leq x \leq b) = \int_a^b p(x)dx$$

Gaussian density:

$$p(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$



Conditional and Joint Probabilities

- Conditional Probability
  - Probability of the event A given the event B
  - Denoted as $p(A|B)$
  - Probability of rain given cloud: $p(Rain|Cloud)$
- Joint Probability
  - Probability of both the events A and B
  - Denoted as $p(A,B)$
  - Probability of rain and game: $p(Rain,Game)$

Bayesian Probability

- Expresses degree of belief of an individual (and hence subjective probability) in the occurrence of an uncertain event.
- Contrasts to frequentism, which assigns probabilities according to relative frequency of occurrence.
- Revises prior estimates of probabilities, based on additional experience and information via Bayes' formula: $\boxed{p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}}$

Example Bayesian Computation

- Random Variables:
  - D : patient has certain disease X (e.g. cancer)
  - S : patient exhibits certain symptom Y (e.g. rash)
- Probabilities:
  - p(D) : prior probability that the patient has X
  - p(S) : prior probability that the patient exhibits Y
- Conditional Probabilities:
  - p(D | S) : probability that the patient has X if exhibits Y
  - p(S | D) : probability that patient exhibits Y having X

2x2 Matrix

	D	not D
S	True Positive $p(S \mid D) = 0.90$	False Positive (Type II Error) $p(S \mid \text{not } D) = 0.01$
not S	Missed Detection (Type I Error) $p(\text{not } S \mid D) = 0.10$	Correct Rejection $p(\text{not } S \mid \text{not } D) = 0.99$

$p(D \mid S) = ?$ given $p(D) = 0.01$

Probability of having disease X given the observed symptom Y and the prior probability of having disease X

Bayes Computation

$$p(D \mid S) = p(S \mid D) \times p(D) / p(S)$$

$$\begin{aligned} p(S) &= p(S \,\&\, D) + p(S \,\&\, \text{not } D) \\ &= p(D) \times p(S \mid D) + p(\text{not } D) \times p(S \mid \text{not } D) \\ &= 0.01 \times 0.90 + 0.99 \times 0.01 \\ &= 0.009 + 0.0099 \\ &= 0.0189 \end{aligned}$$

$$\begin{aligned} p(D \mid S) &= p(S \mid D) \times p(D) / p(S) \\ &= 0.90 \times 0.01 / 0.0189 \\ &= 0.4762 \end{aligned}$$

Exercise

- 1% of women have breast cancer
- 80% of mammograms detect breast cancer when it is there (true positive)
- 9.6% of mammograms detect breast cancer when it's not there (false positive)
- Now suppose you get a positive test result. What are the chances you have cancer?

Outline

- Probability and Bayesian Probability
- Statistics and Hypotheses Testing
- Selected Probability Distributions
- Regression Analysis
- Cluster Analysis
- Stochastic Process Modeling

Descriptive vs. Inferential

- Descriptive statistics
  - Summarize and describe the information content of data that have been collected
  - Distribution, measure of central tendency, measure of dispersion
- Inferential statistics
  - Draws valid inferences about a population based on a sample
  - Make a generalization, test hypothesis, estimate, prediction or decision

Dependence vs. Interdependence

- Dependence methods
  - Use independent variable(s) to predict dependent variable(s)
  - Linear, logistics and kernel regressions, auto-regression, factor analysis, survival analysis
- Interdependence methods
  - Variables involved are independent variables
  - Hierarchical and k-means clustering, linear discriminant analysis, multidimensional scaling

Statistical Hypothesis

- Assumption about a population parameter.
- Verify the hypothesis on a random sample of the population.
- Null hypothesis H_0 – common view try to reject
- Alternative hypothesis H_1 – logical negation of H_0 that the researcher really thinks the cause or phenomenon.
- Significance tests generate 95% or 99% likelihood that the results do not fit the null hypothesis, then it is rejected, favoring the alternative.
- Ex: Null hypothesis that two population means are equal.

Hypothesis Testing Steps

- State null and alternative hypotheses
- Select appropriate test statistic and its probability distribution
- Select level of significance
- Delineate regions of rejection
- Calculate test statistic
- Make a decision regarding null hypothesis

An Example

The mean emission of all engines of new design needs to be below 20 ppm if the design needs to meet the new standard. Ten engines are manufactured for testing purposes, and the emission level of each is determined. The emission data is

15.6 16.2 22.5 20.5 16.4 19.4 16.6 17.9 12.7 13.9

Does the data supply sufficient evidence to conclude that this type of engine meet the new standard? Assume error tolerance is 0.05.

Source: Khan Academy

Select Hypotheses

- A two-tail is used when the QA examiner has no idea which direction the study will go and thus interested in both direction. On the other hand, one-tail test is used when the direction of the assumption is of interest.
- In the context of the example it makes to find the evidence of less than 20 ppm, left tail.

$$H_0: \mu \geq 20 \text{ ppm}$$

$$H_1: \mu < 20 \text{ ppm}$$

Select Test Statistic

- If population standard deviation σ is known: $Z = \dfrac{\overline{X} - \mu}{\sigma}$

  where $\bar{X}$ is sample mean and μ is the population mean

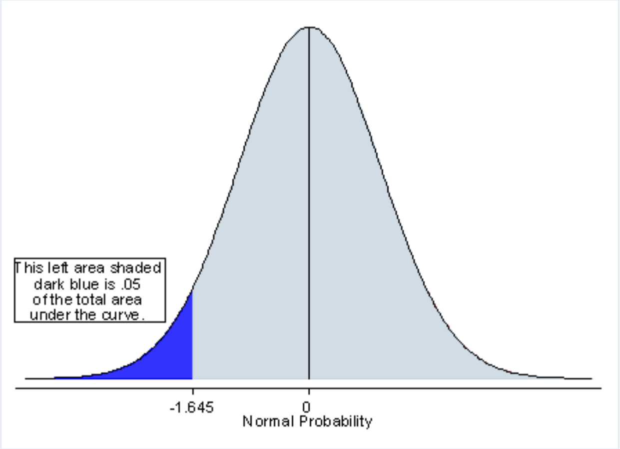
- If population standard deviation σ is not known: $t = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$

- When the sample size is small then the Student's *t* distribution will be used:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n-1}}$$

Test Statistic

- Level of significance: 0.01
- Region of rejection:
- Calculate test statistic: t = ?
- Make a decision: ?



Chi-Square (χ^2)

- Non-parametric test to investigate whether distributions of categorical variables differ from one another.
- One way or single sample chi-square goodness of fit is to determine whether a significant difference exists between an observed and some theoretical expected distribution (frequencies of occurrence).
- Two way chi-square test of independence is to determine whether a significant difference exists between the distributions in two or more categories with two or more groups.

Chi-Square Goodness-of-Fit

- Null hypothesis: Observed and expected distributions of the variable *Outlook* are the same.
- Chi-square statistic = $\sum_i \frac{(O-E)^2}{E}$
- Predetermined level of significance = 95%
- Degrees of freedom = (3-1)x(3-1) = 4

Outlook	Expected	Observed
sunny	5	8
overcast	4	2
rain	5	4

$$\chi^2 = \sum_n \frac{(Observed - Expected)^2}{Expected}$$
$$= \frac{9}{5} + \frac{4}{4} + \frac{1}{5}$$
$$= 3.0$$

Chi-Square Goodness-of-Fit

- Value 3.0 lies between 2.20 and 3.36 and the corresponding probability is p is between 0.7 and 0.5, which is less than 0.95
- Hence the null hypothesis that the two distributions are the same is rejected.

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

Chi-Square Test of Independence

- Is *Outlook* a good predictor of *Class* or is there a significant difference exists between the distributions in *Outlook* and *Class*.
- Null hypothesis is that two distributions are independent

Class / Outlook	play	don't play	Row Subtotal
sunny	2	3	5
overcast	4	0	4
rain	3	2	5
Column Subtotal	9	5	Total = 14

Outlook	Class
sunny	play
sunny	don't play
sunny	don't play
sunny	don't play
sunny	play
overcast	play
overcast	play
overcast	play
overcast	play
rain	don't play
rain	don't play
rain	play
rain	play
rain	play

Chi-Square Test of Independence

- Row subtotals and column subtotals must have equal sums, and total expected frequencies must equal total observed frequencies.
- Note that we are computing expectation with a view that if those total numbers were exactly the total number of observations in the past then what would we expect

$$
\begin{aligned}
& p(Outlook = sunny \,\&\, Class = play) \\
&= p(Outlook = sunny) \times p(Class = play) \\
&= \left(\sum_{Class} \big(p(Outlook = sunny) \times p(Class) \big) \right) \times \\
&\quad \left(\sum_{Outlook} \big(p(Class = play) \times p(Outlook) \big) \right) \\
&= \big(p(sunny) \times p(play) + p(sunny) \times p(don't\ play) \big) \times \\
&\quad \big(p(play) \times p(sunny) + p(play) \times p(overcast) + p(play) \times p(rain) \big) \\
&= (\text{Row subtotal for } sunny/14) \times (\text{Column subtotal for } play/14)
\end{aligned}
$$

$$
\begin{aligned}
& Exp(Outlook = sunny \,\&\, Class = play) \\
&= 14 \times p(Outlook = sunny) \times p(Class = play) \\
&= (\text{Row subtotal for } sunny/14) \times (\text{Column subtotal for } play/14) \\
&= \frac{9 \times 5}{14}
\end{aligned}
$$

Chi-Square Test of Independence

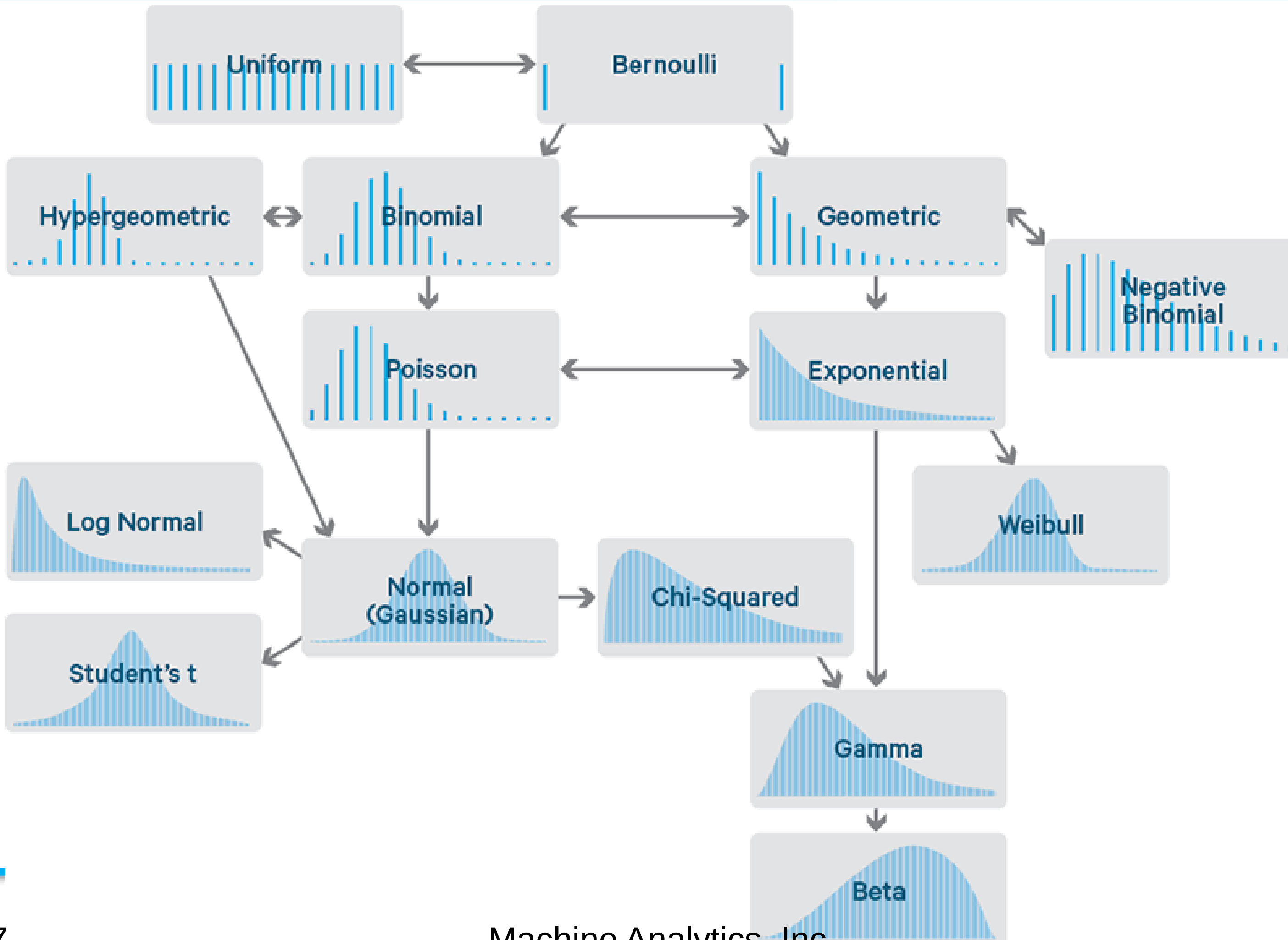
- Chi-square statistic = $\sum_i \frac{(O-E)^2}{E}$
- Level of significance = 95%
- Degrees of freedom = (3-1)x(2-1) = 2
- Chi-sq value ? Is less / greater than <Table Value>, so we would accept / reject the null hypothesis that there is there a significant difference exists between the distributions in *Outlook* and *Class*.

Observed	Expected	$(O-E)^2/E$
2	3.21	?
3	1.79	?
4	2.57	?
0	1.43	?
3	3.21	?
2	1.79	?

Outline

- Probability and Bayesian Probability
- Statistics and Hypotheses Testing
- Selected Probability Distributions
- Regression Analysis
- Cluster Analysis
- Stochastic Process Modeling

Common Probability Distributions



Ref: Sean Owen, 2015

Probability Distributions in Detail

- Continuous
  - Normal
  - Lognormal
  - Exponential
  - Weibull
  - Beta
  - Dirichlet
  - Gamma
- Discrete
  - Binomial
  - Multinomial
  - Poisson

Normal Distribution

- A normal (or Gaussian) random variable X with mean μ and variance σ^2 is described through pdf:

$$f\left(X;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

- For general n-dimensional variable (Σ is covariance):

$$f\left(X;\mu,\Sigma\right)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)}$$

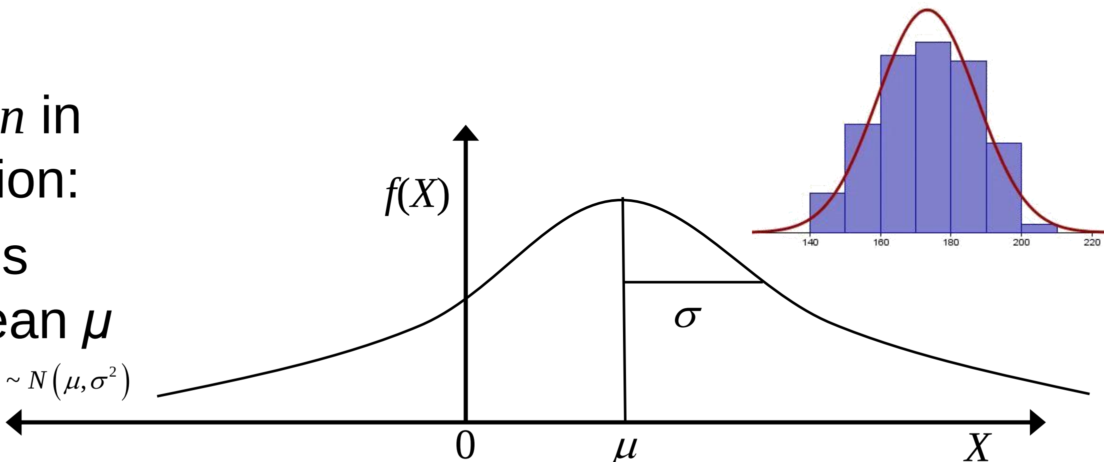
- Special case (standard normal distribution):

$$\mu=0,\sigma^2=1$$

- Normal is a limiting case of n in a discrete binomial distribution:

$$\mu=Np,\sigma^2=Npq,q=1-p$$

- When a random variable X is distributed normally with mean μ and variance σ^2, we write: $X\sim N\left(\mu,\sigma^2\right)$

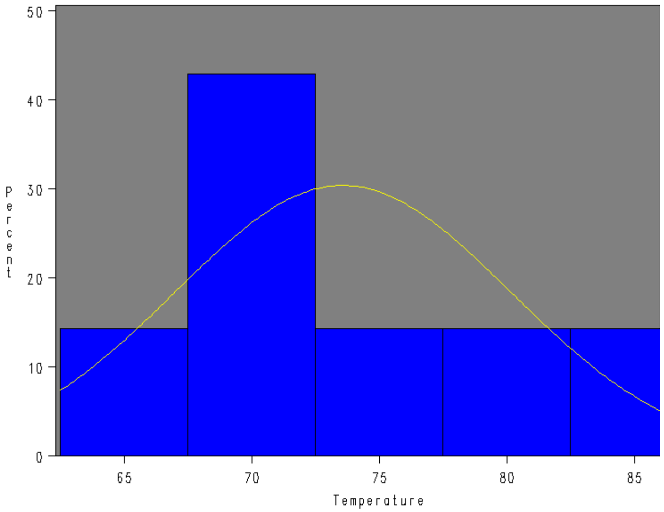


Normal Distribution

Temperature	Humidity
75	70
80	90
85	85
72	95
69	70
72	90
83	78
64	65
81	75
71	80
65	70
75	80
68	80
70	96

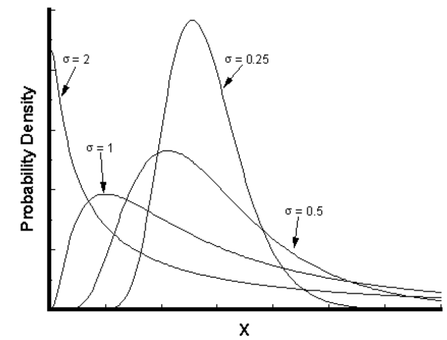
Variable: Temperature (Temperature)

Basic Statistical Measures			
Location		Variability	
Mean	73.57143	Std Deviation	6.57167
Median	72.00000	Variance	43.18681
Mode	72.00000	Range	21.00000
		Interquartile Range	11.00000



Lognormal Distribution

- Good for modeling the lives of units whose failure modes are of a fatigue-stress nature.
- A random variable X is distributed lognormally if the logarithm of X, $\ln(X)$, is normally distributed.
- Say $X' = \ln(X)$ and μ and σ are respectively mean and variance of X':
- Equal probabilities under the normal and lognormal *pdf*s, and incremental areas should also be equal:
- Relation between $f(X)$ and $f(X')$:



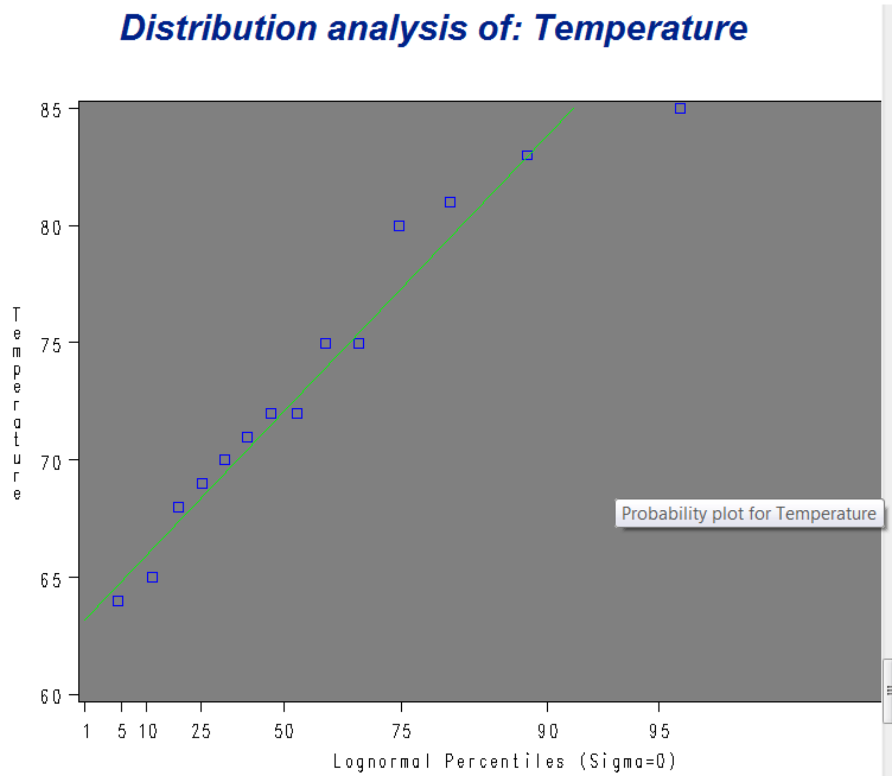
$$f(X') = \frac{1}{\sqrt{2\pi}\sigma'} e^{-\frac{1}{2}\left(\frac{X'-\mu'}{\sigma'}\right)^2}$$

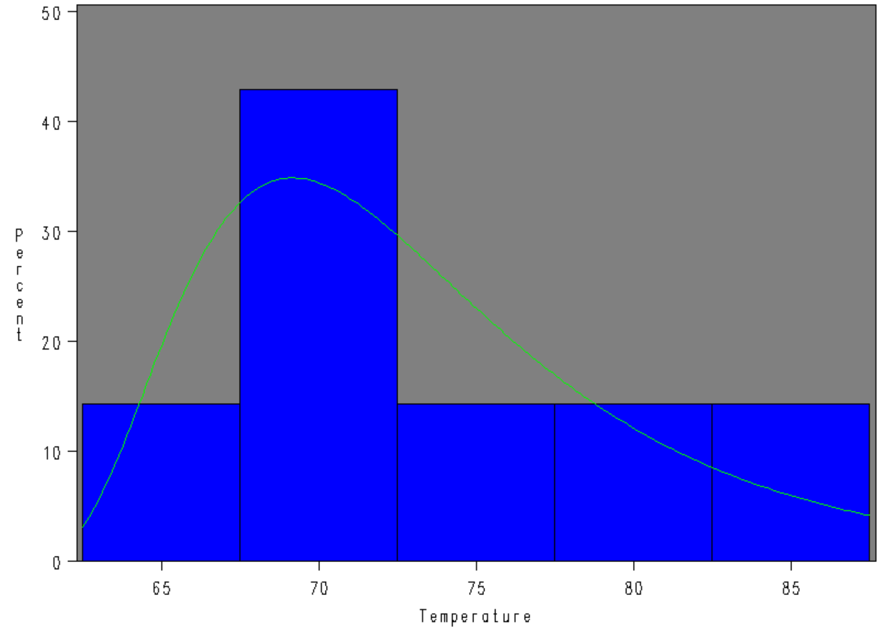
$$f(X)dX = f(X')dX'$$

$$X' = \ln(X) \Rightarrow dX' = \frac{1}{X}dX \Rightarrow f(X) = \frac{f(X')}{X}$$

$$\boxed{f(X) = \frac{1}{X\sqrt{2\pi}\sigma'} e^{\frac{1}{2}\left(\frac{\ln(X)-\mu'}{\sigma'}\right)^2} \quad X \geq 0,\ f(X) \geq 0}$$

Lognormal





Goodness-of-Fit Tests for Lognormal Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.11385708	**Pr > D**	>0.500
Cramer-von Mises	**W-Sq**	0.03099351	**Pr > W-Sq**	>0.500
Anderson-Darling	**A-Sq**	0.24558213	**Pr > A-Sq**	>0.500

Exponential Distribution

- Suppose that the amount of time a customer support agent spends on a phone call is exponentially distributed with mean 5 minutes, so λ = 1/5. What is the probability that an agent will spend more than 10 minutes in a call?
- Density:

$$f(X;\lambda) = \lambda e^{-\lambda X} \quad X \geq 0$$
$$= 0 \quad X < 0$$
$$E[X] = \frac{1}{\lambda}, Var[X] = \frac{1}{\lambda^2}$$

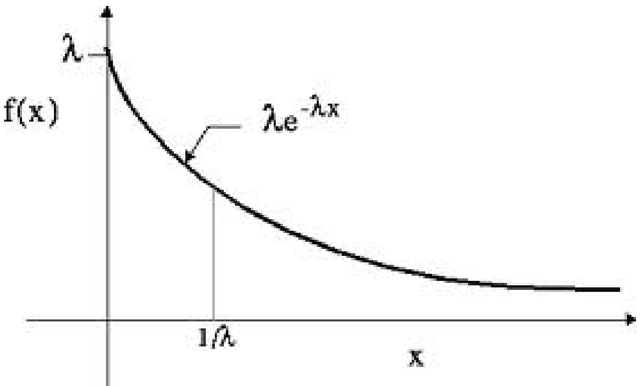
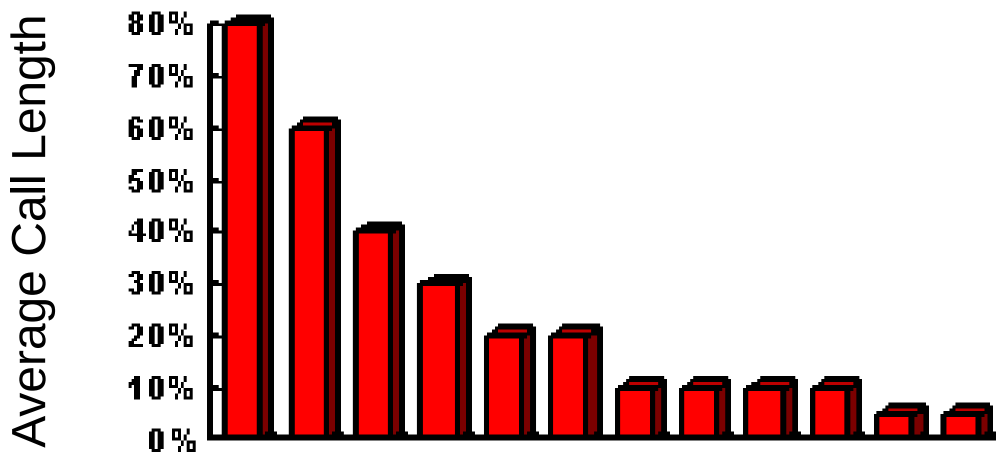


Figure 6. Exponential *pdf*

- Memoryless: $p(X > s+t \mid X > s) = p(X > t)$
  - Probability that an agent will spend more than 10 minutes on a call given that the agent is still on the call after 5 minutes.

Fitted Exponential Distribution

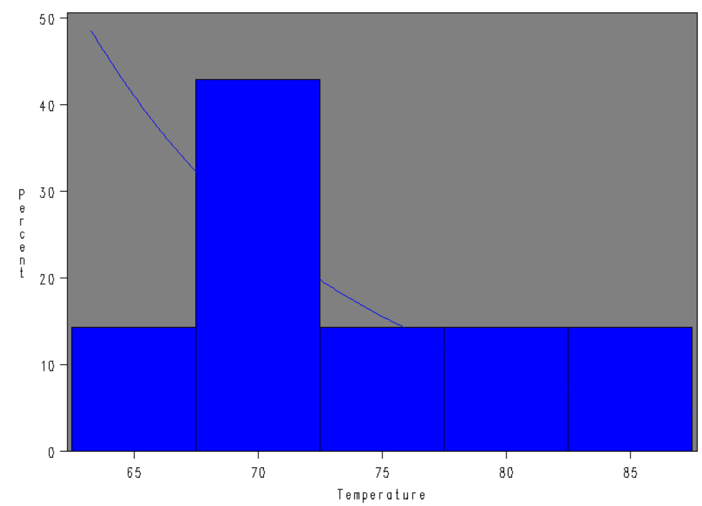


Probability of Call Length

Exponential distribution of call center call length
Source: Internet

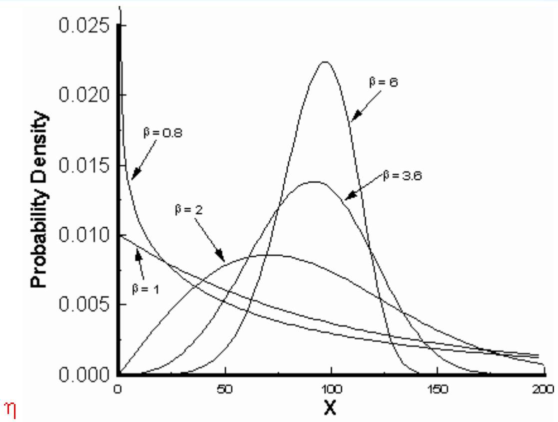
Fitted Distribution for Temperature

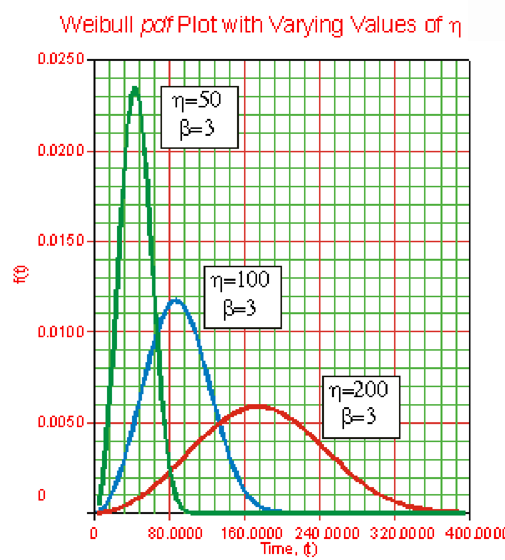
Parameters for Exponential Distribution		
Parameter	**Symbol**	**Estimate**
Threshold	Theta	63.26374
Scale	Sigma	10.30769
Mean		73.57143
Std Dev		10.30769

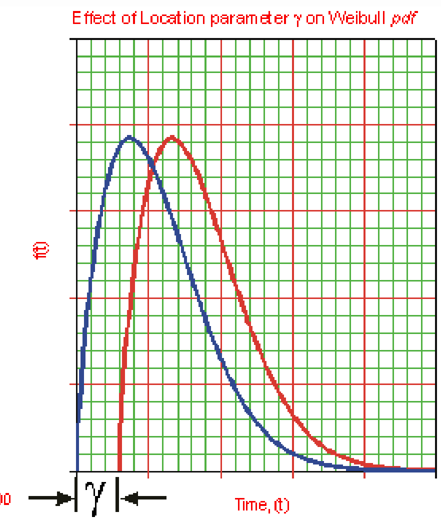


Weibull

- PDF: $f\left(X;\beta,\theta,\delta\right)=\frac{\beta}{\theta}\left(\frac{X-\delta}{\theta}\right)^{\beta-1}e^{-\left(\frac{X-\delta}{\theta}\right)^{\beta}}\quad X\geq\delta$
  - β is the shape parameter
  - θ is the scale parameter
  - δ is the location parameter
- Special cases:
  - $\beta = 1$: Exponential distribution
  - $\beta = 2$: Rayleigh distribution
  - $3 \leq \beta \leq 4$: Approximates normal distribution

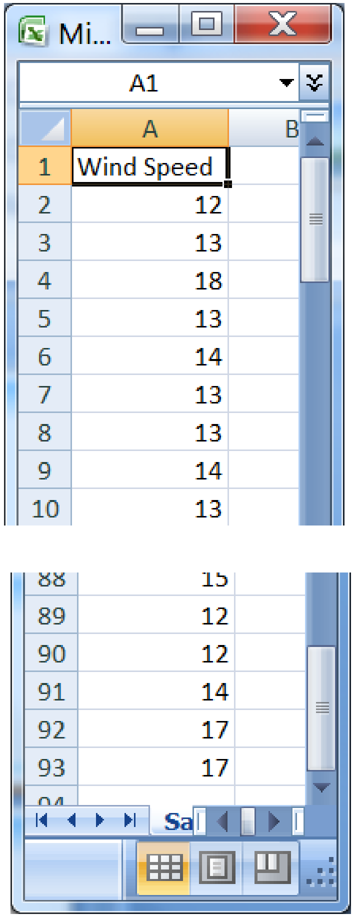


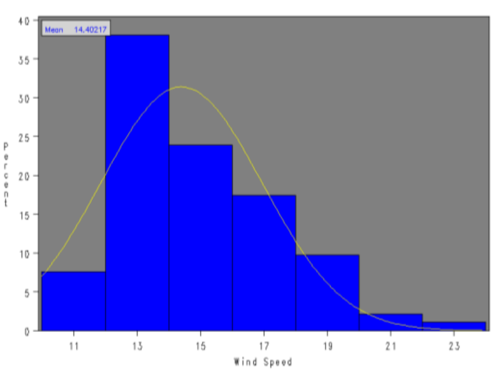




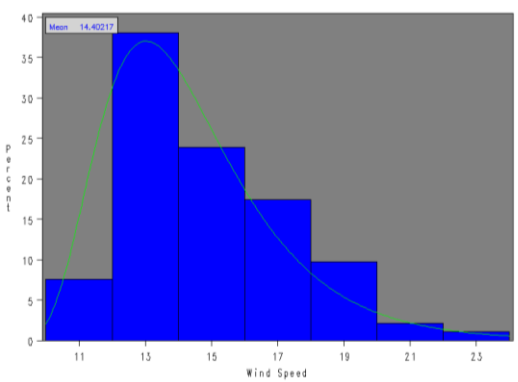
Fitted Weibull DIstribution

- San Francisco Wind Speed Data Jun-Aug 1965

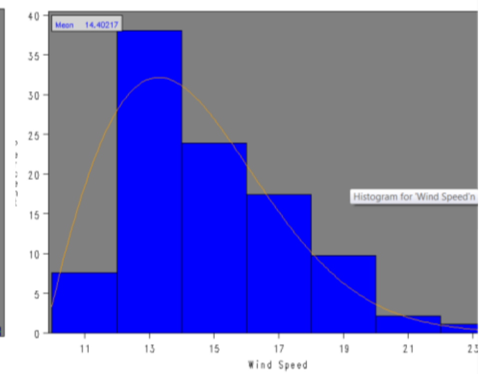




Normal



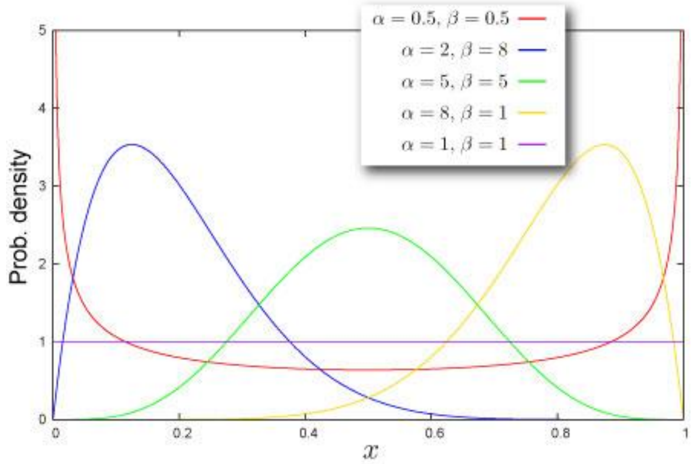
Lognormal



Weibull

Beta Distribution

- Used to model continuous data with values between 0 and 1.
- The distribution function $F(X)$ for the beta distribution has no closed form solution.
- Standard univariate beta distribution:



$$f\left(X;\alpha,\beta\right)=\frac{X^{\alpha-1}\left(1-X\right)^{\beta-1}}{\mathrm{B}\left(\alpha,\beta\right)} \quad 0 \le X \le 1$$

$$\mathrm{B}\left(\alpha,\beta\right)=\frac{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}{\Gamma\left(\alpha+\beta\right)}$$

$$\mu=\mathrm{E}\left[X\right]=\frac{\alpha}{\alpha+\beta},\ \sigma^2=Var\left[X\right]=\frac{\alpha\beta}{\left(\alpha+\beta\right)^2\left(\alpha+\beta+1\right)}$$

Gamma Distribution

- Standard univariate Gamma distribution:

$$f(X;\gamma) = \frac{X^{\gamma-1}e^{-X}}{\Gamma(\gamma)} \qquad X \geq 0;\ \gamma > 0$$

- Gamma function is defined by:

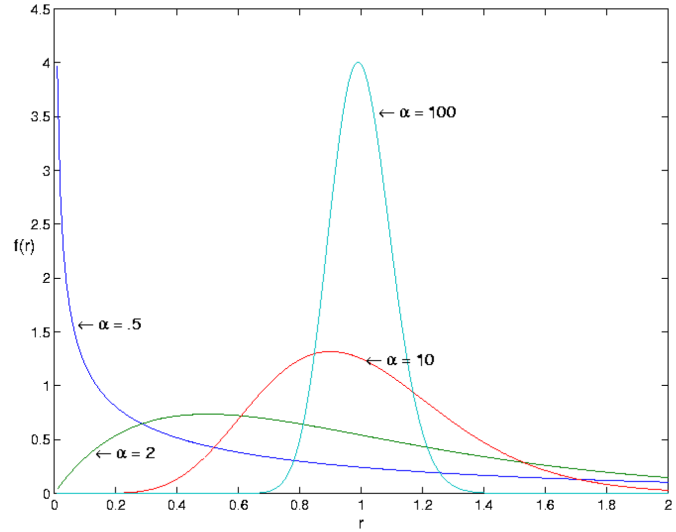
$$\Gamma(\gamma) = \int_0^{\infty} x^{\gamma-1}e^{-x}dx \qquad \gamma \in (0,\infty)$$

- Good for modeling highly skewed variables.

$$\Gamma(\gamma+1) = \gamma\Gamma(\gamma) \qquad \gamma > 0$$

$$\Gamma(k) = (k-1)! \qquad k \text{ is a positive integer}$$

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$



Dirichlet Distribution

- Generalization of Beta distribution:

$$f\left(X_1,\ldots,X_n;\alpha_1,\ldots,\alpha_n\right)=\frac{1}{Beta\left(\alpha_1,\ldots,\alpha_n\right)}\prod_{i=1}^{n}X_i^{\alpha_i-1}$$

$$X_i\geq 0,\alpha_i\geq 0\quad\text{and}\sum_i X_i=1$$

- The parameter α_i can be interpreted as the prior observation counts for events governed by the probability representing the variable X_i.

$$\mu_i=\mathrm{E}\left[X_i\right]=\frac{\alpha_i}{\alpha_1+\ldots+\alpha_n}=\frac{\alpha_i}{\sum_j\alpha_j}$$

$$\sigma_i^2=Var\left[X_i\right]=\frac{\alpha_i\left(\sum_j\alpha_j-\alpha_i\right)}{\left(\sum_j\alpha_j\right)^2\left(\sum_j\alpha_j+1\right)}$$

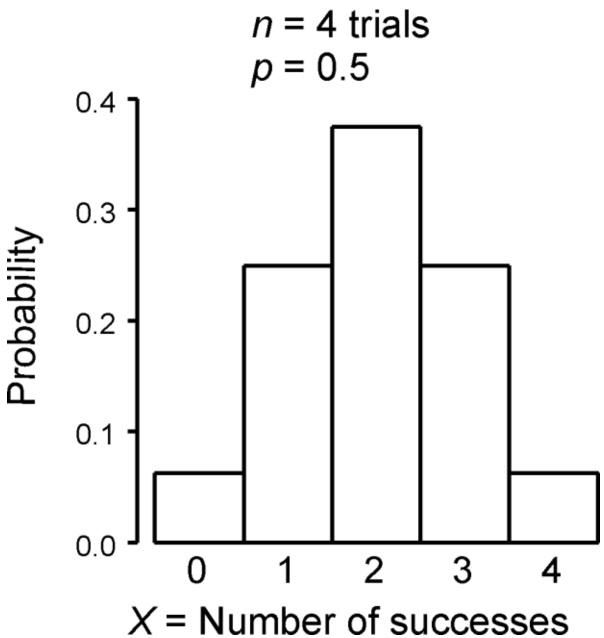
- Normalization constant:

$$Beta\left(\alpha_1,\ldots,\alpha_n\right)=\frac{\prod_{i=1}^{n}\Gamma\left(\alpha_i\right)}{\left(\sum_{i=1}^{n}\alpha_i\right)}$$

Binomial Distribution

- Suppose the probability of *success* and *failure* in any one trial is given by the fixed probabilities p and $q = 1-p$.
- Binomial distribution provides the probability of the number of successes and failures in n such independent 2-outcome trials.
- Mean is np and variance is $np(1-p)$

$$f\left(X;n,p\right)=\binom{n}{X}p^{X}\left(1-p\right)^{n-X}$$



Multinomial Distribution

- Suppose the probability of each outcome in any one k-outcome trial is given by the fixed probabilities $p_1,\ldots,p_k$.
- The multinomial distribution is a generalization of the binomial distribution giving the probability of each combination of outcomes in n independent trials of such k-outcome process.
- Distribution of each n k-outcome trials:

$$f\left(X_1,\ldots,X_k;n,p_1,\ldots,p_k\right)=\frac{n!}{X_1!\ldots X_k!}p_1^{X_1}\ldots p_k^{X_k} \qquad X_i \geq 0, p_i \geq 0 \text{ and } \sum_i p_i = 1$$

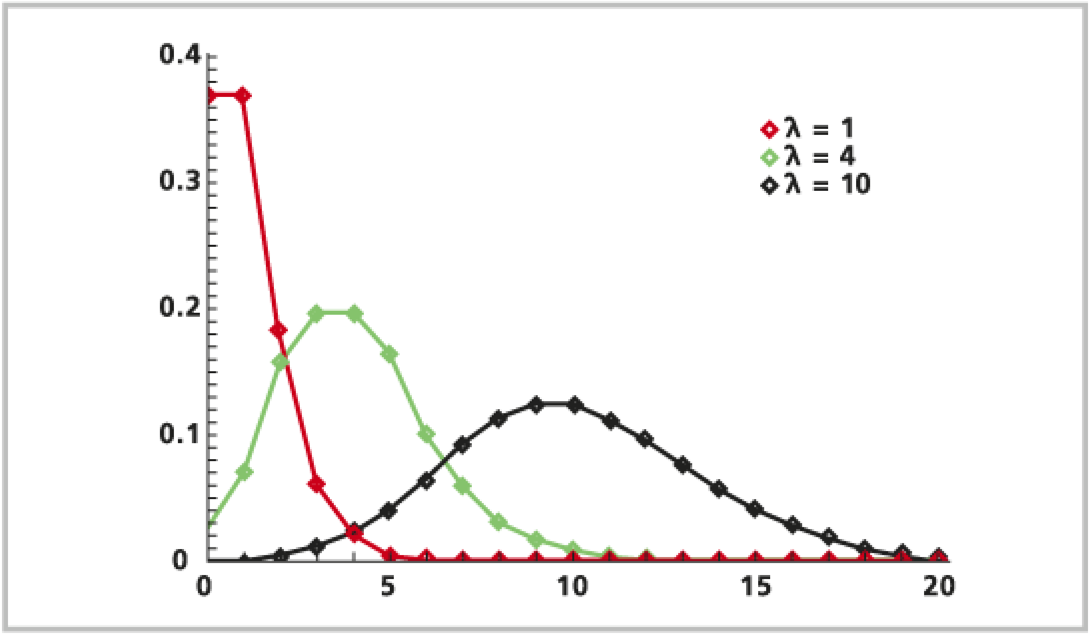
Poisson Distribution

- Expected “count” of a number of events across time or over an area
  - The number of calls received during the first hour in a call center.
  - The number of customers arrived at a supermarket over the weekend.
- Underlying assumptions:
  - Probability of observing a single event over a small interval is approximately proportional to the size of that interval.
  - Probability of two events occurring in the same narrow interval is negligible.
  - Probability of an event within a certain interval does not change over different intervals.
  - Probability of an event in one interval is independent of the probability of an event in any other non-overlapping interval.

Poisson Distribution

- Probability density function: $$f(X;\lambda)=\frac{\lambda^X e^{-\lambda}}{X!}$$

- Mean is λ

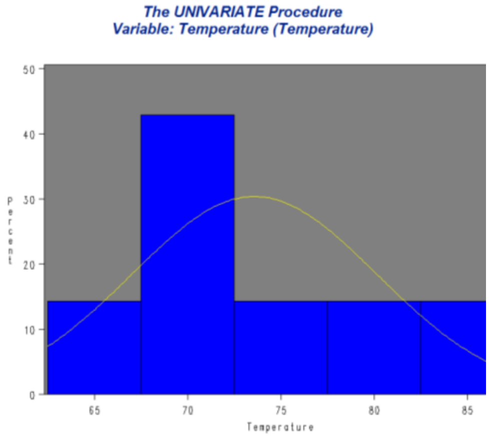


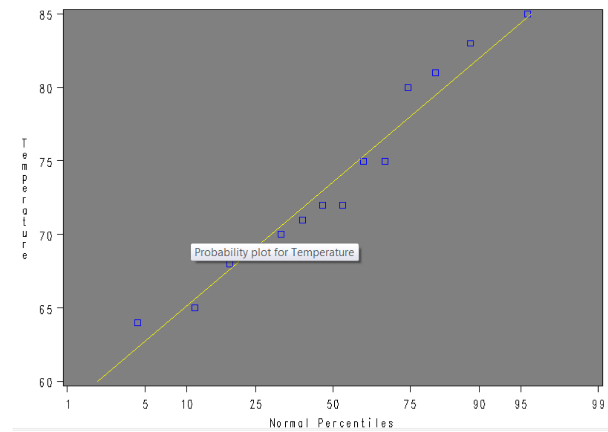
Goodness-of-Fit Test

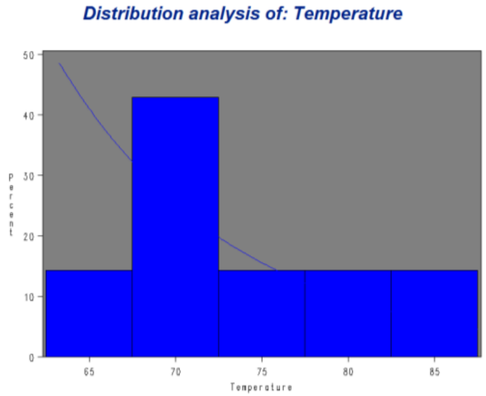
- Probability Plot
- Chi-Square Test
- Kolmogorov-Smirnov Test

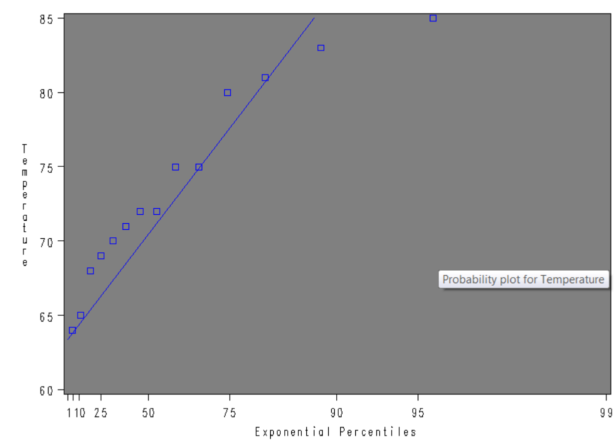
Probability Plot

- Percentile is the value of a variable below which a certain percent of observations fall
- Just look at the plotted points, and see how well they fit the normal line. Process data is normally distributed if they fit well.









Kolmogorov-Smirnov Test

- Continuous observations with cumulative density function (*cdf*) F: $x_1, \ldots, x_n$
- Null hypothesis: $F(0)$ is a known cdf $H_0 : F(x) = F_0(x), \quad \textit{for all } x$
- Empirical cumulative distribution: $\hat{F}(x) = \dfrac{\#(i : x_i < x)}{n}$
- Kolmogorov-Smirnov test statistics: $D_n = \sup_x \left| \hat{F}(x) - F_0(x) \right|$
- Null distribution of the statistic D_n can be obtained by simulation or, for large samples, using the Kolmogorov-Smirnov's distribution function.

Kolmogorov-Smirnov Test

- Test statistic D can be compared to the critical value from a statistical table. If D is larger than the critical value, then we reject the hypothesis that the data set was drawn from the theoretical distribution $F(0)$; otherwise we do not reject the hypothesis
- This shows that $D = 0.16$ (0.22) is well in the middle of the distribution and so the data do not contradict the null hypothesis that the discrepancies are normally distributed with zero mean and variance equal to one.

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.16592298	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.05096146	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.30214105	Pr > A-Sq	>0.250

Goodness-of-Fit Tests for Exponential Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.22553608	Pr > D	0.157
Cramer-von Mises	W-Sq	0.14975221	Pr > W-Sq	0.103
Anderson-Darling	A-Sq	0.78160576	Pr > A-Sq	0.116

Conjugate Prior

- Problem of choosing a prior probability distribution
  - Realistic vs. a mathematical function that simplifies the analytic computation of posterior
  - Posterior belongs to the same functional family as the prior

Conjugate Prior p(X)	Likelihood p(Z|X)	Posterior p(X|Z)
$Normal(\mu_0, \sigma_0^2)$	$Normal(\mu, \sigma^2)$, known σ^2	$Normal(\mu_1, \sigma_1)$
$Beta(p; r, s)$	$Binomial(n; N, p)$	$Beta(p; r + n, s + N - n)$
$Gamma(\lambda; r, s)$	$Poisson(\lambda; n)$	$Gamma(\lambda; r + n, s + 1)$
$Dirichlet(p_1, \ldots, p_k; \alpha_1, \ldots, \alpha_k)$	$Multinomial(n_1, \ldots, n_k; p_1, \ldots, p_k)$	$Dirichlet(p_1, \ldots, p_k; n_1+\alpha_1, \ldots, n_k+\alpha_k)$
$Gamma(\lambda; r, s)$	$Exponential(\lambda; n)$	$Gamma(\lambda; r+n, s+\Sigma x_i)$

Bayesian Inference with Conjugates

- Consider prior distribution of variable p representing the probability of a head is Beta:

$$f\left(p;\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}p^{\alpha-1}\left(1-p\right)^{\beta-1}\qquad 0\le p\le 1$$

- Likelihood for obtaining n heads with probability of a head being p is binomial:

$$f\left(n;N,p\right)=\binom{N}{n}p^{n}\left(1-p\right)^{N-n}$$

- The posterior is Beta:

$$f\left(p;\alpha+n,\beta+N-n\right)=\frac{\Gamma\left(\alpha+n+\beta+N-n\right)}{\Gamma\left(\alpha+n\right)\Gamma\left(\beta+N-n\right)}p^{\alpha+n-1}\left(1-p\right)^{\beta+N-n-1}\qquad 0\le p\le 1$$

Outline

- Probability and Bayesian Probability
- Statistics and Hypotheses Testing
- Selected Probability Distributions
- Regression Analysis
- Cluster Analysis
- Stochastic Process Modeling

Regression Analyses

- Multiple Linear
- Nonlinear
- Logistic
- Generalized Linear Model
- Kernel

Data for Logistic Regression

Outlook	Temperature	Humidity	Windy	Class	Class_0_1	Temp_0_1
sunny	75	70	true	play	1	1
sunny	80	90	true	don't play	0	1
sunny	85	85	false	don't play	0	1
sunny	72	95	false	don't play	0	0
sunny	69	70	false	play	1	0
overcast	72	90	true	play	1	1
overcast	83	78	false	play	1	1
overcast	64	65	true	play	1	0
overcast	81	75	false	play	1	1
rain	71	80	true	don't play	0	0
rain	65	70	true	don't play	0	0
rain	75	80	false	play	1	0
rain	68	80	false	play	1	0
rain	70	96	false	play	1	0

Simple Linear Regression

- Models the relationship between two variables by fitting a linear equation to observed data.

$$\boxed{Y = a + bX}$$

- One variable (X) is an explanatory variable, and the other one (Y) is a dependent variable. Slope is b and a is intercept.
- Least-squares is the most common method for fitting equation where the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line.

Simple Linear Regression

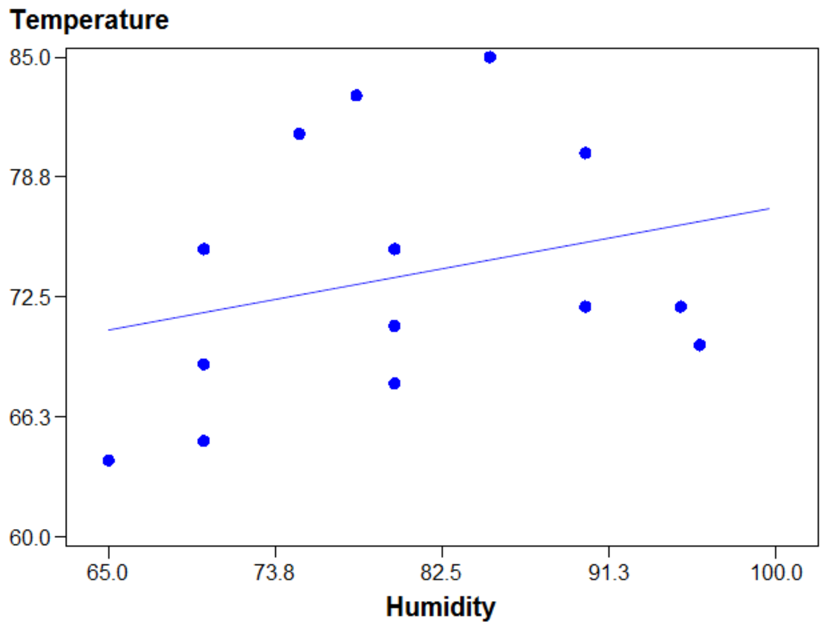
- N data point: $(y_1, x_1), \ldots, (y_n, x_n)$
- Minimize: $\sum_{i=1}^{n} (y_i - a - bx_i)^2$

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{Cov(X,Y)}{Var(X)}$$

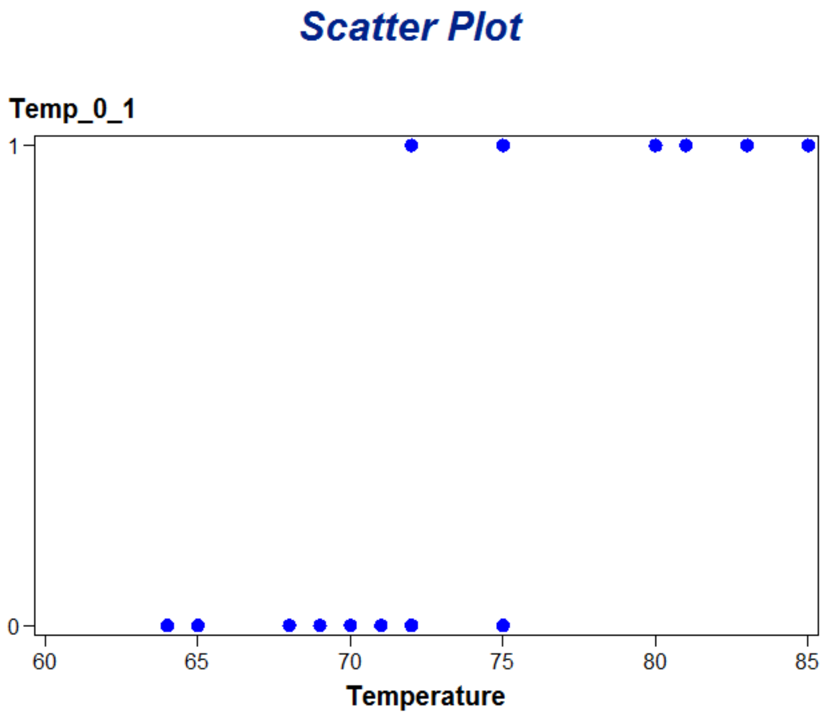
$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

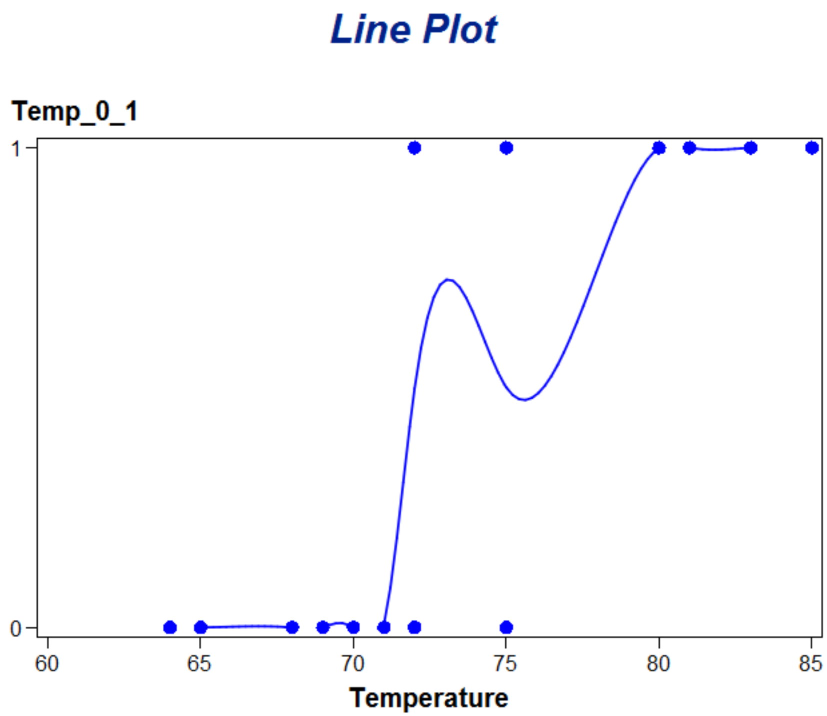


Temperature	Humidity
75	70
80	90
85	85
72	95
69	70
72	90
83	78
64	65
81	75
71	80
65	70
75	80
68	80
70	96

Logistics Regression

- Regression in which the dependent variable is binary.
- Check these Temperature vs binary Outlook plots.





Logit Equation

- Case 1:
  - Probabilities would be above 1 or below 0 for some *X*

$$p\left(Y=1\,|\,X\right)=a+bX$$

- Case 2:
  - Ratio is positive but would be ∞ for some *X*

$$\frac{p}{1-p}=a+bX$$

$$\boxed{p=\frac{e^{a+bX}}{1+e^{a+bX}}}$$

- Case 3:
  - Log values are between 0 and 1

$$\log\left(\frac{p}{1-p}\right)=a+bX$$

Parameter Estimation

- Maximum likelihood methods to estimate the parameters a and b of a logistic model with binary response.
  - $Y = 1$ with probability p
  - $Y = 0$ with probability 1-p
- For each $Y_i = 1$ the probability pi appears in the product. Similarly, for each $Y_i = 0$ the probability $1 - p_i$ appears in the product.
- Maximize the log likelihood and solve for a and b.

$$L(a,b;Data) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}$$

$$= \prod_{i=1}^{n}\left(\frac{e^{a+bX_i}}{1+e^{a+bX_i}}\right)^{Y_i}\left(\frac{1}{1+e^{a+bX_i}}\right)^{1-Y_i}$$

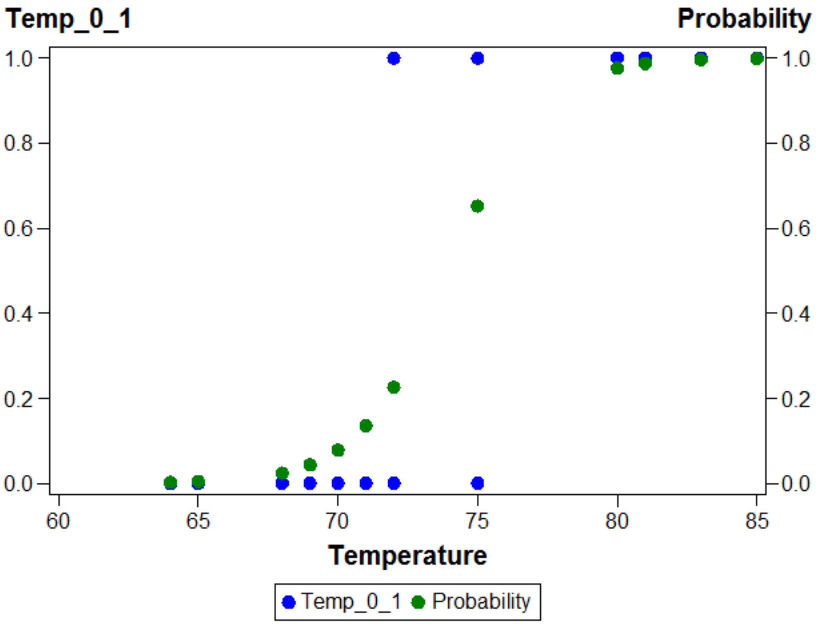
$$= \prod_{i=1}^{n}\frac{\left(e^{a+bX_i}\right)^{Y_i}}{1+e^{a+bX_i}}$$

$$\log\left(L(a,b;Data)\right)$$

Logistic Regression Fitted

Temp.	Temp_0_1	log(*p*/(1-*p*))	*p*
75	1	?	0.34782
80	1	?	0.02336
85	1	?	0.00107
72	0	?	0.77455
69	0	?	0.95677
72	1	?	0.77455
83	1	?	0.00370
64	0	?	0.99798
81	1	?	0.01269
71	0	?	0.86472
65	0	?	0.99624
75	0	?	0.34782
68	0	?	0.97629
70	0	?	0.92244

$$\log\left(\frac{p}{1-p}\right) = 45.94 - 0.62X$$



Factor Analysis

- Obtains a small set of uncorrelated variables from a large set of variables to gain insight to categories.
  - Clusters variables into homogeneous sets.
  - Creates new variables (i.e. factors or latent variables).
- Exploratory
  - No pre-defined idea of the structure or variable dimensions in a set of variables.
- Confirmatory
  - To test specific hypothesis about the structure or the number of dimensions underlying a set of variables.

Factor Model

- *m* independent, and hence orthogonal, factors:

$$
\boxed{\begin{aligned}
&X_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \ldots + \lambda_{1m}F_m + e_1 \\
&X_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \ldots. + \lambda_{2m}F_m + e_2 \\
&\ldots \\
&X_n = \lambda_{n1}F_1 + \lambda_{n2}F_2 + \ldots + \lambda_{nm}F_m + e_n
\end{aligned}}
$$

- Parameters L_{ij}'s are referred to as loadings; L_{ij} is called the *loading* of variable X_i on factor F_j.
- Loadings range from -1 to 1, representing degree to which each of the variables correlates with each of the factors.
- Use principal components to decide the number of factors.

Deriving Principal Components

- The first component vector a_1 is the linear combination a_1X with maximum variance

$$Var(F_1) = Var(a_1^T X) = a_1^T Cov(X) a_1$$

such that $\sum a_1^2 = 1$

- The second component vector a_2 is the linear combination a_2X with maximum variance

$$Var(F_2) = Var(a_2^T X) = a_2^T Cov(X) a_2$$

such that $\sum a_2^2 = 1$ and so on …

- Eigenvalue is the amount of variance in the data described by the factor; it helps to choose the number of factors.

Factor Analysis Example

- Thurstone 20 boxes data:
  http://life.bio.sunysb.edu/morph/

i	x	y	z	i	x2	y2	z2	xy	xz	yz	sqrt(x2+y2	sqrt(x2+z2	sqrt(y2+z2	2(x+y)	2(x+z)	2(y+z)	logx	logy	logz	xyz	sqrt(x2+y2+z2)	ex	ey	ez
1	3	2	1	1	9	4	1	6	3	2	3.605551	3.162278	2.236068	10	8	6	0.477121	0.30103	0	6	3.741657387	20.08554	7.389056	2.718282
2	3	2	2	2	9	4	4	6	6	4	3.605551	3.605551	2.828427	10	10	8	0.477121	0.30103	0.30103	12	4.123105626	20.08554	7.389056	7.389056
3	3	3	1	3	9	9	1	9	3	3	4.242641	3.162278	3.162278	12	8	8	0.477121	0.477121	0	9	4.358898944	20.08554	20.08554	2.718282
4	3	3	2	4	9	9	4	9	6	6	4.242641	3.605551	3.605551	12	10	10	0.477121	0.477121	0.30103	18	4.69041576	20.08554	20.08554	7.389056
5	3	3	3	5	9	9	9	9	9	9	4.242641	4.242641	4.242641	12	12	12	0.477121	0.477121	0.47712	27	5.196152423	20.08554	20.08554	20.08554
6	4	2	1	6	16	4	1	8	4	2	4.472136	4.123106	2.236068	12	10	6	0.60206	0.30103	0	8	4.582575695	54.59815	7.389056	2.718282
7	4	2	2	7	16	4	4	8	8	4	4.472136	4.472136	2.828427	12	12	8	0.60206	0.30103	0.30103	16	4.898979486	54.59815	7.389056	7.389056
8	4	3	1	8	16	9	1	12	4	3	5	4.123106	3.162278	14	10	8	0.60206	0.477121	0	12	5.099019514	54.59815	20.08554	2.718282
9	4	3	2	9	16	9	4	12	8	6	5	4.472136	3.605551	14	12	10	0.60206	0.477121	0.30103	24	5.385164807	54.59815	20.08554	7.389056
10	4	3	3	10	16	9	9	12	12	9	5	5	4.242641	14	14	12	0.60206	0.477121	0.47712	36	5.830951895	54.59815	20.08554	20.08554
11	4	4	1	11	16	16	1	16	4	4	5.656854	4.123106	4.123106	16	10	10	0.60206	0.60206	0	16	5.744562647	54.59815	54.59815	2.718282
12	4	4	2	12	16	16	4	16	8	8	5.656854	4.472136	4.472136	16	12	12	0.60206	0.60206	0.30103	32	6	54.59815	54.59815	7.389056
13	4	4	3	13	16	16	9	16	12	12	5.656854	5	5	16	14	14	0.60206	0.60206	0.47712	48	6.403124237	54.59815	54.59815	20.08554
14	5	2	1	14	25	4	1	10	5	2	5.385165	5.09902	2.236068	14	12	6	0.69897	0.30103	0	10	5.477225575	148.4132	7.389056	2.718282
15	5	2	2	15	25	4	4	10	10	4	5.385165	5.385165	2.828427	14	14	8	0.69897	0.30103	0.30103	20	5.744562647	148.4132	7.389056	7.389056
16	5	3	2	16	25	9	4	15	10	6	5.830952	5.385165	3.605551	16	14	10	0.69897	0.477121	0.30103	30	6.164414003	148.4132	20.08554	7.389056
17	5	3	3	17	25	9	9	15	15	9	5.830952	5.830952	4.242641	16	16	12	0.69897	0.477121	0.47712	45	6.557438524	148.4132	20.08554	20.08554
18	5	4	1	18	25	16	1	20	5	4	6.403124	5.09902	4.123106	18	12	10	0.69897	0.60206	0	20	6.480740698	148.4132	54.59815	2.718282
19	5	4	2	19	25	16	4	20	10	8	6.403124	5.385165	4.472136	18	14	12	0.69897	0.60206	0.30103	40	6.708203932	148.4132	54.59815	7.389056
20	5	4	3	20	25	16	9	20	15	12	6.403124	5.830952	5	18	16	14	0.69897	0.60206	0.47712	60	7.071067812	148.4132	54.59815	20.08554

Factor Analysis Example

Eigenvalues of the Covariance Matrix: Total = 3735.11207 Average = 186.755604

	Eigenvalue	Difference	Proportion	Cumulati
1	3080.89185	2595.39059	0.8248	0.82
2	485.50126	325.98415	0.1300	0.95
3	159.51711	154.60797	0.0427	0.99
4	4.90913	2.87626	0.0013	0.99
5	2.03288	0.70813	0.0005	0.99
6	1.32475	0.69356	0.0004	0.99
7	0.63118	0.41889	0.0002	0.99
8	0.21229	0.12361	0.0001	1.00
9	0.08868	0.08583	0.0000	1.00
10	0.00285	0.00281	0.0000	1.00
11	0.00005	0.00001	0.0000	1.00
12	0.00003	0.00002	0.0000	1.00
13	0.00001	0.00001	0.0000	1.00

14 factors will be retained by the MINEIGEN

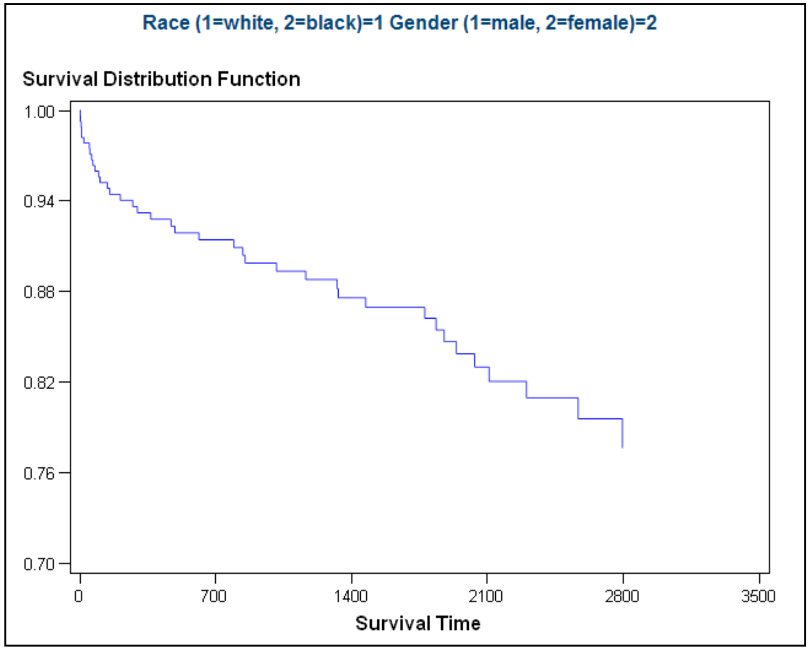
	Eigenvectors						
	1	2	3	4	5	6	
x2	0.11281	-0.00634	-0.01278	-0.12379	0.73942	-0.39254	-0.18
y2	0.02455	0.19797	-0.09994	0.03093	0.13498	0.57957	0.06
z2	0.00791	0.06718	0.21073	0.25736	0.02036	-0.03242	0.28
xy	0.05563	0.14135	-0.06807	-0.09091	0.36437	0.48628	-0.18
xz	0.03466	0.06716	0.22095	-0.04774	0.14007	-0.18310	0.53
yz	0.01337	0.10988	0.14954	0.04121	-0.05525	0.07239	0.24
sqrt (x2+y2)	0.01345	0.01841	-0.01028	-0.00862	0.12432	0.01804	-0.00
sqrt (x2+z2)	0.01317	0.00701	0.02079	0.02098	0.11933	-0.06186	0.03
sqrt (y2+z2)	0.00430	0.03662	0.01461	0.04195	0.04491	0.13231	0.12
2(x+y)	0.03494	0.06526	-0.03206	-0.03446	0.33002	0.17983	-0.03
2(x+z)	0.03149	0.03210	0.10205	0.03011	0.25646	-0.16812	0.34

Survival Analysis

- Time to event analysis – the time from the beginning of an observation period (e.g. surgery) to an event (e.g. death, end of the study) or loss of contact/withdrawal from the study.
- A censored subject may or may not have an event after the end of observation time; right censoring: at the time of observation, the relevant event had not yet occurred.
- Kaplan-Meier (product-limit)/Life Table estimators of the survivor and hazard functions for the sample as a whole, or for separate subgroups; they are not multivariate regression models.
- Cox's semi-parametric proportional hazard model for continuous time data.

Survival Analysis Example

- Kidney transplant patients.



Kidney transplant data: Variables represented in the dataset are as follows (Source: Medical College of Wisconsin):

Observation number
Time to death or on-study time
Death indicator (0=alive, 1=dead)
Gender (1=male, 2=female)
Race (1=white, 2=black)
Age in years

SAMPLE

1	1	0	1	1	46
2	5	0	1	1	51
3	7	1	1	1	55
….					
524	3430	0	1	2	28
525	1	0	2	1	41
526	2	1	2	1	60
…..					

Outline

- Probability and Bayesian Probability
- Statistics and Hypotheses Testing
- Selected Probability Distributions
- Regression Analysis
- Cluster Analysis
- Stochastic Process Modeling

Statistical Clustering

- Hierarchical
- Partitional (*k*-means)
- Multidimensional scaling

Agglomerative vs. Divisive

- Agglomerative (aka bottom-up)
  - Start with all data points in their own group
  - Repeatedly merge two groups that have smallest dissimilarity until there is only one cluster
- Divisive (aka top-down)
  - Start with all points in one cluster
  - Repeatedly split a group into two resulting biggest dissimilarity until all points are in their own cluster
- Agglomerative is simpler

Hierarchical Clustering

- Two approaches:
  - Bottom up or agglomerative : grouping small clusters into larger ones
  - Top down or divisive: splitting big clusters into smaller ones.
- Bottom up algorithm:
  - **Input:** N items to cluster
  - **Output:** Hierarchical partitions of items.
  - **Step 1:** Assign each item to a cluster, so initially there will be N clusters, each containing just one item. Compute “distances” (similarities) between clusters.
  - **Step 2:** Find the closest pair of clusters and merge them into a single cluster, so there will be one cluster less.
  - **Step 3:** Compute distances between the new cluster and each of the old clusters.
  - **Step 4:** Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

Distance between clusters

- Single-linkage (connectedness or minimum)
  - Distance between two clusters to be equal to the shortest distance from any member of one cluster to any member of the other cluster.
- Complete-linkage clustering (diameter or maximum)
  - Distance between two clusters to be equal to the greatest distance from any member of one cluster to any member of the other cluster.
- Average-linkage clustering
  - Distance between two clusters to be equal to the average distance from any member of one cluster to any member of the other cluster.

Ward's distance

- Distance between two clusters A and B depends on by how much the sum of squares will increase the two clusters are merged.
- Minimize the Sum of Squares (SS) of any two clusters that can be formed at each step.

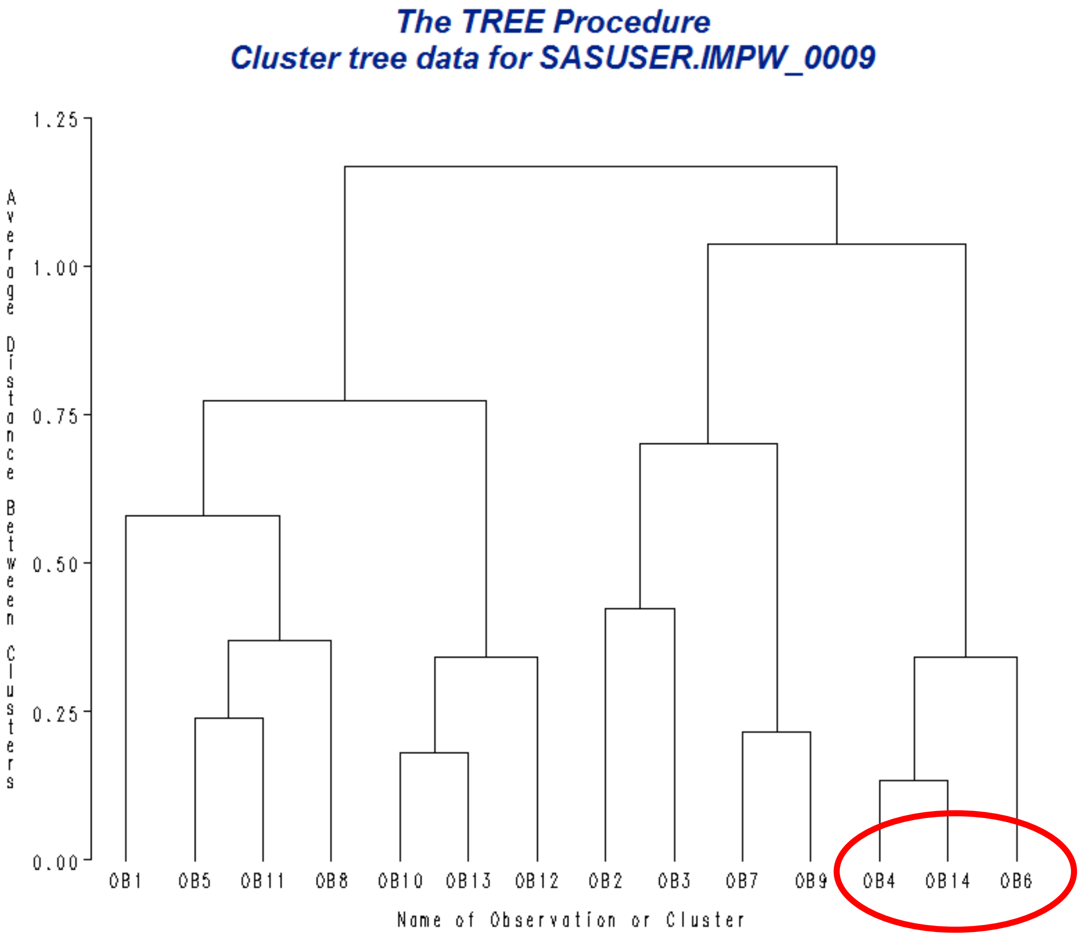
$$\Delta(A,B) = \sum_{z_i \in A \cup B} \left\| z_i - \overline{z} \right\|^2 - \sum_{x_j \in A} \left\| x_j - \overline{x} \right\|^2 - \sum_{y_j \in B} \left\| y_j - \overline{y} \right\|^2$$

$$= \frac{n(A)+n(B)}{n(A)n(B)} \left\| \overline{x} - \overline{y} \right\|^2$$

where $\overline{x}$ represents cluster-centers and n(*) represents the number of elements in the cluster.

Hierarchical Cluster

No.	Temp.	Humidity
1	75	70
2	80	90
3	85	85
4	72	95
5	69	70
6	72	90
7	83	78
8	64	65
9	81	75
10	71	80
11	65	70
12	75	80
13	68	80
14	70	96



Pros & Cons

- Hierarchical structure is informative
- Complexity is higher as compared to k-means
- The crucial question is how many clusters?

k-means Clustering

- Unsupervised, non-deterministic, flat (non-hoerarchical) clustering technique.
- Algorithm:
  - **Input:** Set of N items and number K of centroids
  - **Output:** K clusters
  - **Step 1:** Place K points into the space represented by the items that are being clustered. These points represent initial cluster centroids. Good practice is to place them as far from each other as possible.
  - **Step 2:** Assign each object to the cluster that has the closest centroid.
  - **Step 3:** When all objects have been assigned, recalculate the positions of the K centroids.
  - **Step 4:** Repeat Steps 2 and 3 until the centroids no longer change.

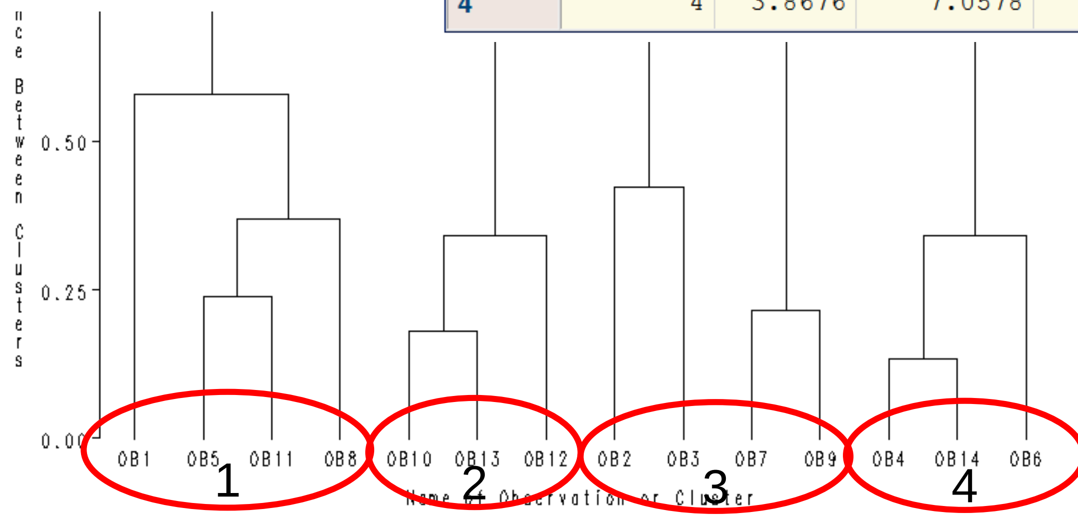
k-means Clustering

No.	Temp.	Humidity
1	75	70
2	80	90
3	85	85
4	72	95
5	69	70
6	72	90
7	83	78
8	64	65
9	81	75
10	71	80
11	65	70
12	75	80
13	68	80
14	70	96

Cluster tree c

Cluster Means		
Cluster	Temperature	Humidity
1	68.25000000	68.75000000
2	71.33333333	80.00000000
3	83.00000000	79.33333333
4	73.50000000	92.75000000

Cluster Summary				
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Rad Exce
1	4	3.9476	6.8648	
2	3	2.4833	3.6667	
3	3	3.8944	6.0093	
4	4	3.8676	7.0578	



Pros & Cons

- Faster to compute that hierarchical clusters
- Difficult to determine the value of K
- Doesn't work well with non-convex clusters
- Different initial positions of centroids yield different final clusters

Outline

- Probability and Bayesian Probability
- Statistics and Hypotheses Testing
- Selected Probability Distributions
- Regression Analysis
- Cluster Analysis
- Stochastic Process Modeling

Time and Regression

- ARMA/ARIMA
- ARCH/GARCH

Autoregressive Moving Average (ARMA)

- Process if stochastic in that it evolves in time according to probabilistic laws.
- Gaussian stationary processes ARMA(p, q):

$$\boxed{x_t = \underbrace{a_1 x_{t-1} + \ldots + a_p x_{t-p}}_{\text{Autoregressive (AR)}} + \underbrace{\varepsilon_t - b_1 \varepsilon_{t-1} - \ldots - b_q \varepsilon_{t-q}}_{\text{MovingAvergae (MA)}}}$$

- where ε_t is a sequence of uncorrelated random variables with zero mean and variance σ^2.
- The basic principle consists of computing the values taken by the innovation $\hat{\varepsilon}_t$ of the stochastic process:

$$\left(x(t); t = 1, 2, \ldots\right)$$

Parameter Estimation

- Likelihood of innovations:

$$(2\pi)^{-n/2}\left(\prod_{t=1}^{n}\sigma_t\right)^{-1}\exp\left\{-\frac{1}{2}\sum_{t=1}^{n}(\hat{\varepsilon}_t/\sigma_t)^2\right\}$$

$\sigma_t = h_t\sigma$ is the standard deviation of $\hat{\varepsilon}_t$

- Maximizing above with respect to the parameters $a_1,\ldots,a_p,b_1,\ldots,b_q$ is equivalent to minimizing

$$\left(\prod_{t=1}^{n}h_t^2\right)^{1/n}\sum_{t=1}^{n}(\hat{a}_t/h_t)^2$$

Transformation to Kalman Filter

- State Vector: $X_t = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{bmatrix}$ $m = \max(p, q+1)$

$$T = \begin{pmatrix} a_1 & 1 & 0 & \dots & 0 \\ a_2 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & 0 & 0 & \dots & 1 \\ a_m & 0 & 0 & \dots & 0 \end{pmatrix}_{m\times m}$$

- State-space representation:

$$\boxed{\left.\begin{aligned} X_t &= TX_{t-1} + Hb_t \\ x_t &= ZX_t \end{aligned}\right\} t = 1, \dots, n, \dots}$$

$$H = \begin{pmatrix} 1 & -b_1 & \dots & -b_{m-1} \end{pmatrix}_{1\times m}$$

$$Z = \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}_{m\times 1}$$

$$\alpha_t = \frac{ZL_t}{h_t^2}$$

- Evaluation of $\hat{\varepsilon}_t$ via Kalman filter recursion:

$$\hat{\varepsilon}_t = x_t - Z\hat{X}_t$$

$$\hat{X}_{t+1} = T\hat{X}_t + K_t\left(\hat{\varepsilon}_t / h_t^2\right)$$

$$K_{t+1} = K_t - \alpha_t TL_t$$

$$L_{t+1} = TL_t - \alpha_t K_t$$

$$h_{t+1}^2 = h_t^2\left(1 - \alpha_t^2\right)$$