COMP7/8150: Fundamentals of Data Science

Week 4: Statistical / Probabilistic Methods

Max H. Garzon

Data Life Cycle

- **Problem Definition/goal**
  - Identify/specify goals of the data analysis
  - commit to specific deliverables
- **Data pre-processing**
  - Identify appropriate data
  - Acquire data (gather, lookup, understand)
- **Data processing**
  - Identify methods (gather, cleanse, store)
  - Carry out the analysis (patterns, trends, predictions?)
- **Data post-processing**
  - Visualize and present
  - Deploy and evaluate. Iterate, if necessary

Learning Objectives

- **To identify and characterize major concepts and models for data understanding and processing**
  - **Descriptive statistics, plots**
    data presentation/visualization
  - **Probabilistic models (samples, probability measures)**
    understanding, predicting

- **To identify simulation methods :**
  - **visualizing data (cognitive load, HCI concepts) Clustering**
  - **Understanding data (chunking, clustering, modeling)**

Where does data come from?

- Real-world
  - Observations
  - Experiments to produce it
- Usually we do not know the full corpus of data (population), only have samples
- Can use models (approximations) to understand the corpus of data
- Various kinds of models
  - Statistical (mean, std, moments)
  - Machine Learning (clusters, patterns, ..)
  - Computational intelligence

Statistical models

http://www.youtube.com/watch?v=ooOdP1BJxLg

$$y = f(x) + \varepsilon$$

- Two major components:
  - Deterministic component f(x), describing the relationship between a dependent variable (y) and independent variable(s) (x) or predictor(s) it depends on.
  - A random component ε, representing the deviation (due to measurement error) between the actual observation *y* and its predicted value f(x).

.. Statistical models: examples

- First order linear model:
  - $y = \alpha + \beta x + \varepsilon$
  - $f(x) = \alpha + \beta x$
- Usually we do not know the coefficients (parameters α and β)
- Key question: how to find the "best estimate" based on observations (x_1, y_1), (x_2, y_2), ... (x_n, y_n).
- Random error component ε is assumed to follow a certain probability distribution.

.. Examples of a statistical model

- Special/simplest case: f(x) = constant = μ.

  One population model:
  - $y = \mu + \varepsilon$
- This model is useful to describe a random sample of size n $\{y_1, y_2, \ldots, y_n\}$ taken from a population with mean μ.
- Usually we do not know the parameter μ.
- The random error component ε is assumed to follow certain probability distribution with additional parameters of interest.

Statistical/random Experiments

- Sample space Ω (elementary events)
  set of all possible outcomes (population)
- A probability space is defined by a σ-algebra of events (subsets E of Ω for which probabilities below make sense) E_i that are *measured* by a probability of occurrence $p(A_i)$ so that even for composite events E_i :
  - $p(\cup^{\cdot} E_i) = \Sigma_i\, p(E_i)$ (disjoint countable unions)
  - $p(\Omega)=1$, so that $p(E^c)=p(\Omega - E)= 1 - p(E)$;

  In particular, $p(\Phi)=0$ and
  $p(E \cup F)= p(E) + p(F) – p(E \cap F)$.

.. Experiments: Examples

- Toss a fair coin $\Omega=?; p(E)=?$
- Toss a fair coin 10 times ? ?
- Roll an unloaded die ? ?
- Select an integer number randomly in an interval [a,b]
- Select an interval at random in the real line
- Take a measurement of a person’s height
- Go to a bus stop and record a wait time for bus
- Select a random page in a book and count the number of typos $\Omega=?; p(e)=?$
- Take a set of persons’ heights and a histogram every 5”

Common probability models

- Some popular probability distributions are useful to model the outcomes of certain experiments.
- We can classify them into two classes:
  - Discrete distributions: Discrete Uniform, Binomial, Poisson, Geometric, Hypergeometric, Negative Binomial …
  - Continuous distributions: Uniform, Exponential, Normal, Chi-square, Gamma, Student-t, F, …

Random variables (RVs)

- A **random variable X** is a numerical measurement X(e) of the elementary events e in a random experiment in which events

  {X=x} = (X=cx) = { e in Ω: X(e) = x }

  are measurable events for every value x.
- Usually, the outcome is not fixed and so its value is determined by chance.
  We can describe its chances by appropriate probability distributions
- Random variables are denoted using capital letters such as X, Y, Z, ...
- Example: Toss a coin ten times. Let

  X=number of heads observed.

  Find p(X=4). How about its `expected value' and its `variability' arising from uncertainty?

Probability distribution of dRV

http://www.youtube.com/watch?v=Fvi9A_tEmXQ&feature=related

- A **discrete random variable** (dRV) has a countable number of possible values (finite or infinite but enumerable by 1,2,3, ..)
  - Examples (quantitative)
    X=number of phone calls you received on a given day
    Y=Roll two dice and observe the sum of the faces up
  - Examples (qualitative/categorical): color of hair, gender
- Every dRV X gives rise to a prob distribution of X on the same sample space Ω with
  $p_X(x)=Pr(X=x)$ [will usually drop subindex X]
- Cumulative pdf $F(x) = Pr(X \leq x)$
- Can understand X via (a plot of) the pdf, particularly if we can find some approximation using a known pdf

Mean and variance of a RV

http://www.youtube.com/watch?v=j__Kredt7vY&feature=relmfu

- We can model random variable X by its probability distribution $X \sim p(x) = Pr(X=x)$.
- Measures for its "behavior" can be
  - Central tendency
    Expected value/mean/mode
    $\mu = E(X) = \Sigma_x \; p(x)\, x$
    Higher Moments $E(X^r) = \Sigma\, x^r\, p(x^r)$
    (skewedness r=3; kurtosis r=4)
  - Dispersion
    Variance $\sigma^2 = Var(X) = \Sigma_x (x-\mu)^2\, p(x)$
    Standard deviation std $\sigma = \sqrt{\sigma^2}$

    Percentiles q: smallest x so that $P(X \le x) \ge q$
- The mean/variance/std of the most popular distributions is known (check with R, Python, …)

Other visualizations of a RV

http://www.youtube.com/watch?v=j__Kredt7vY&feature=relmfu

- Suppose that we can model random variable X with probability distribution $X \sim p(x) = \Pr(X=x)$.
- Visualizations of its "behavior" can be
  - Its Mode

    Expected value/mean

    $\mu = E(X) = \Sigma_x \; p(x)\, x$

    Higher Moments $E(X^r) = \Sigma\, x^r\, p(x^r)$
  - Its Dispersion

    Variance $\sigma^2 = Var(X) = E[(x-\mu)^2] = \Sigma_x (x-\mu)^2\, p(x)$

    Standard deviation std $\sigma = \sqrt{\sigma^2}$

    Percentiles q: smallest x so that $P(X \leq x) \geq q$
- The mean/variance/std of the most popular distributions is known (check with R, Python, …)

Discrete Probability Distributions

- Common discrete distributions are:
  - Discrete Uniform
  - Bernoulli and Binomial
  - Poisson
  - Geometric
  - Hypergeometric
  - Negative Binomial
- You need to know the key parameters of each distribution so that you can choose appropriately for a given data set.

.. Discrete Uniform Distribution

- You choose a digit (0..9) randomly and X=random digit selected.
  - p(x)=Pr(X=x)=0.1, for x=0, 1, 2, …, 9.
- You toss a fair die and

  X=number that comes up.
  - p(x)=Pr(X=x)=1/6, for x=1, 2, 3, 4, 5, 6.
- In general, X=an integer number randomly selected between two integers A and B.

  What are the probability distribution, mean, and variance for X ?

General Properties of the Discrete Uniform Distribution (dUD)

- Let X=an integer number randomly selected between two integers A and B.
- formula for its probability distribution, mean and variance:
  - $p(x)=Pr(X=x) = 1/C$, where $C=(B-A+1)$.
  - $E(X) = \mu = (A+B)/2$.
  - $Var(X) = \sigma^2 = (C^2-1)/12$.
- For random digit example, verify that

  $E(X)= 4.5$, $Var(X)=99/12=8.25$.

Binomial Distribution

http://www.youtube.com/watch?v=Edm--LTH4SM

- The experiment consists of *n* identical repetition of trials with only two outcomes (Bernoulli) each: Success (X=1) or (X=0).
- The probability of success on a single trial is *p* and **remains constant** from trial to trial. The probability of failure is $q = 1 - p$.
- The trials are **independent**.
- We are interested in a RV
  X= number of successes in n trials.
- What are the probability distribution, mean, and variance of X ?

Binomial Distribution-Examples

- You toss a fair coin 10 times and
  X = number of heads up
- You roll a fair die 4 times and
  X = number of times 1 appeared
- A restaurant accept 25 reservations and
  X = number of confirmed reservations
- To indicate that X follows a binomial distribution, write $X \sim B(n,p)$

Binomial Probability Distribution

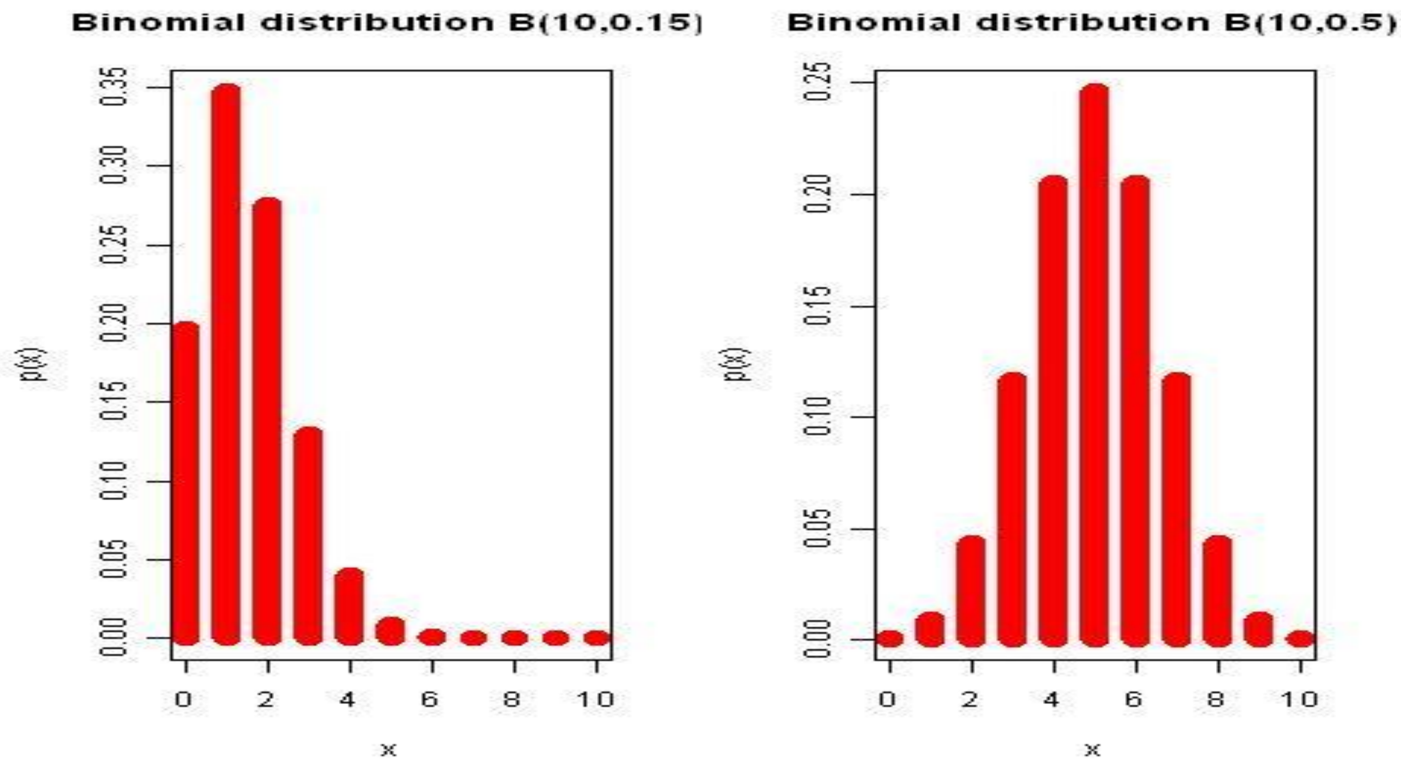
$X \sim B(n, \pi)$

$$p(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}, \quad 0, 1, \cdots, n.$$

1. In R: size=n, prob=π, distribution function=*binom*.
2. $E(X) = n\pi$.
3. $Var(X) = n\pi(1-\pi)$.

R uses π instead of p.
Both notations will be used.

Plot of Binomial distribution



Computing probabilities in R

- R can be useful to compute various common probability distributions for discrete and continuous random variables.
- In general, for distribution named xxx, you can use
- dxxx = probability distribution function, P(X=x).
- **p**xxx = cumulative probability distribution, P(X≤x).
- **q**xxx = percentile of probability distribution.
- **r**xxx = generate a random element from xxx distribution.

Binomial probabilities in R

- Syntax:
  - `dbinom(x, size, prob, log = FALSE)`
  - `pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)`
  - `qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)`
  - `rbinom(n, size, prob)`
- Note: optional arguments (with default values) can be (and usually are) omitted (or need to know the defaults.)

.. Binomial Distribution: Examples

- You toss a fair coin 10 times and
  X=number of heads up
- X follows a binomial distribution,
  $X \sim B(n,p)$, n=10, p=0.5
- Finding Pr(X=4) using R:
  - > dbinom(4, 10, 0.5)
  - [1] 0.2050781
- Finding Pr(X≤4) using R:
  - > pbinom(4, 10, 0.5)
  - [1] 0.3769531

.. Binomial Distribution: Examples

- You roll a fair die 4 times and
  X=number of times 1 appeared.
- $X \sim B(n,p)$, $n=4$, $p=1/6$.
- Finding $Pr(X \le 2)$ using R:

```
> pbinom(2, 4, 1/6)
[1] 0.9837963
```

- Finding $Pr(X=x)$ for $x=0,1,\ldots,4$ using R:
  - `> dbinom(0:4, 4, 1/6)`

```
[1] 0.482253086 0.385802469
0.115740741 0.015432099
0.000771605
```

[5 possible values]

Poisson distribution: Examples

- X=number of phone calls received per day.
- X=number of traffic accidents in one week at a certain location.
- X=number of typos found in one page of a report.
- In general X is the number of events that occur in a period of time or space during which an average of λ such events are expected to occur.
- Notation: $X \sim \text{Poisson}(\lambda)$

Poisson Probability Distribution

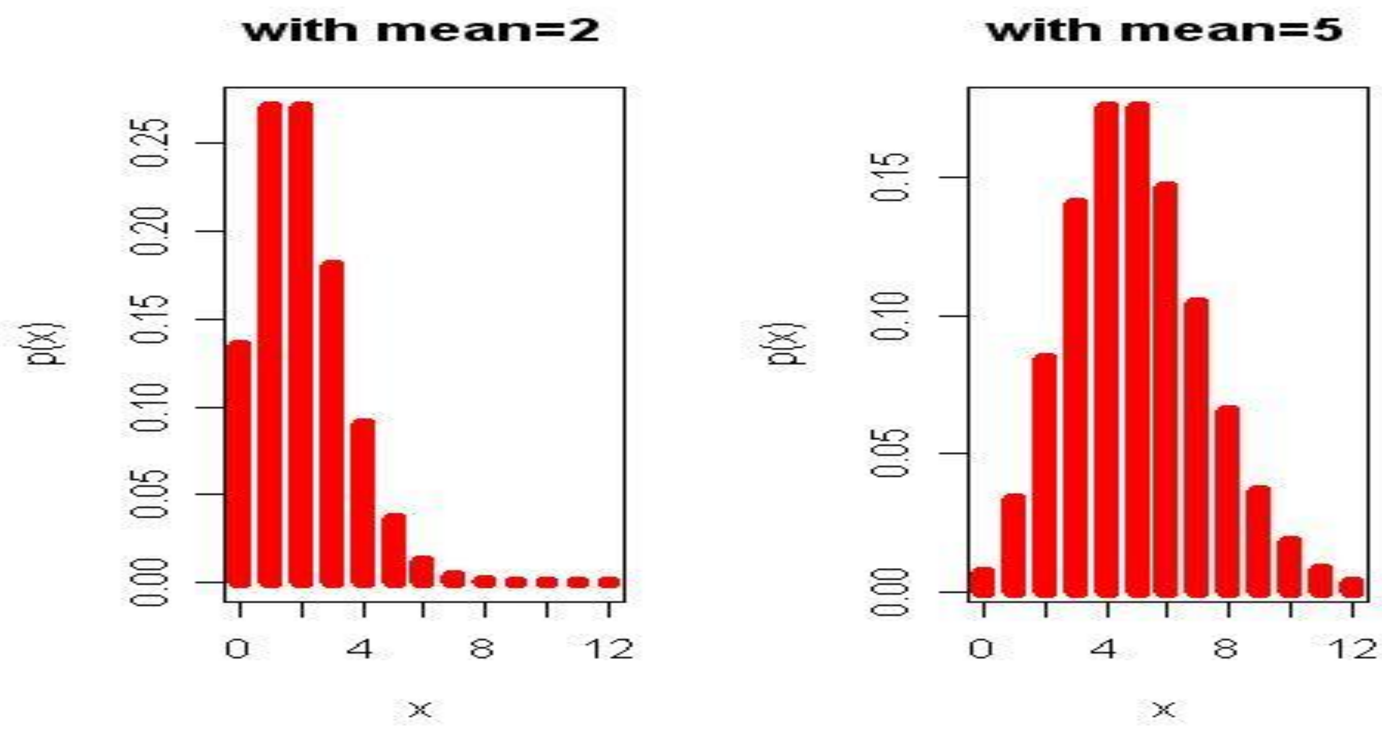
Poisson distribution

$X \sim Poisson(\lambda)$, if the p.d.f. of X is

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \cdots$$

1. In R: lambda=λ, distribution function=$pois$.
2. $E(X) = \lambda$.
3. $Var(X) = \lambda$.

.. Poisson distributions



Computing Poisson in R

- Syntax:
  - `dpois(x, lambda, log = FALSE)`
  - `ppois(q, lambda, lower.tail = TRUE, log.p = FALSE)`
  - `qpois(p, lambda, lower.tail = TRUE, log.p = FALSE)`
  - `rpois(n, lambda)`
  - Note: optional arguments (with default values) can be (and usually are) omitted.

.. Poisson Distribution: App

- The average number of traffic accidents on a certain section of highway is two per week.

  What is the probability of exactly one accident during a one-week period?
  - X ~ Poisson(2).
    $Pr(X = 1) = 2^1 e^{-2}/1! = 0.2707$; or
  - Using R to compute it:
    ```
    > dpois(1,2)
    [1] 0.2706706
    > 2^1*exp(-2)
    [1] 0.2706706
    ```

.. Poisson Distribution: App

- The average number of phone calls received is five per day
- What is the probability of at most two calls received in day?
  - X ~ Poisson(5)
  - Finding p(X ≤ 2) using R:

```
> dpois(0:2,5)
   [1] 0.006737947 0.033689735 0.084224337
> ppois(2,5)
   [1] 0.1246520
> sum(dpois(0:2,5))
   [1] 0.1246520
```

Simulation of random numbers from discrete distribution

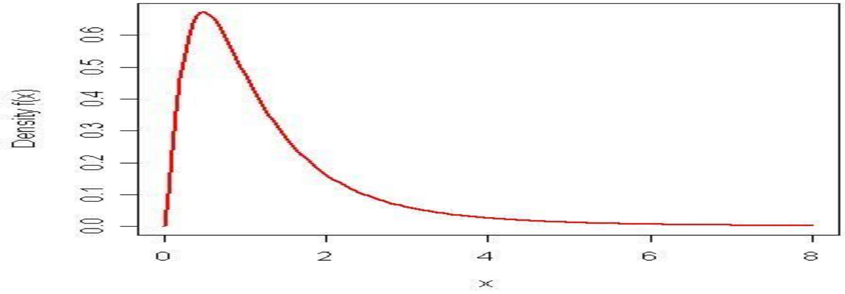
- Sampling from a binomial distribution:
  - ```
    > rbinom(20, size=4, p=0.5)
    [1] 2 3 1 1 1 1 1 0 0 2 2 3 2 3 1 3 0 1 3 2
    ```
  - ```
    > rbinom(20, size=4, p=0.5)
    [1] 3 1 2 3 4 3 3 3 1 1 2 4 0 2 3 2 2 3 2 1
    ```
- Sampling from a Poisson distribution:
  - ```
    > rpois(20, lambda=2)
    [1] 1 2 4 3 1 1 0 2 0 1 4 0 4 1 3 2 3 1 5 1
    ```
    ```

Continuous RVs (cRVs)

- Continuous random variables can assume the uncountably many values corresponding to points on a Euclidean line interval.
- Need define carefully the events bc the laws of probability need to hold, e.g. selecting a random # in unit interval [0,1] cannot have p=0 for single points bc probability of full event p(Ω)=1.
- Examples:
  - Heights, weights
  - Length of life of a particular product
  - Quantifiable experimental lab error

.. cRVs

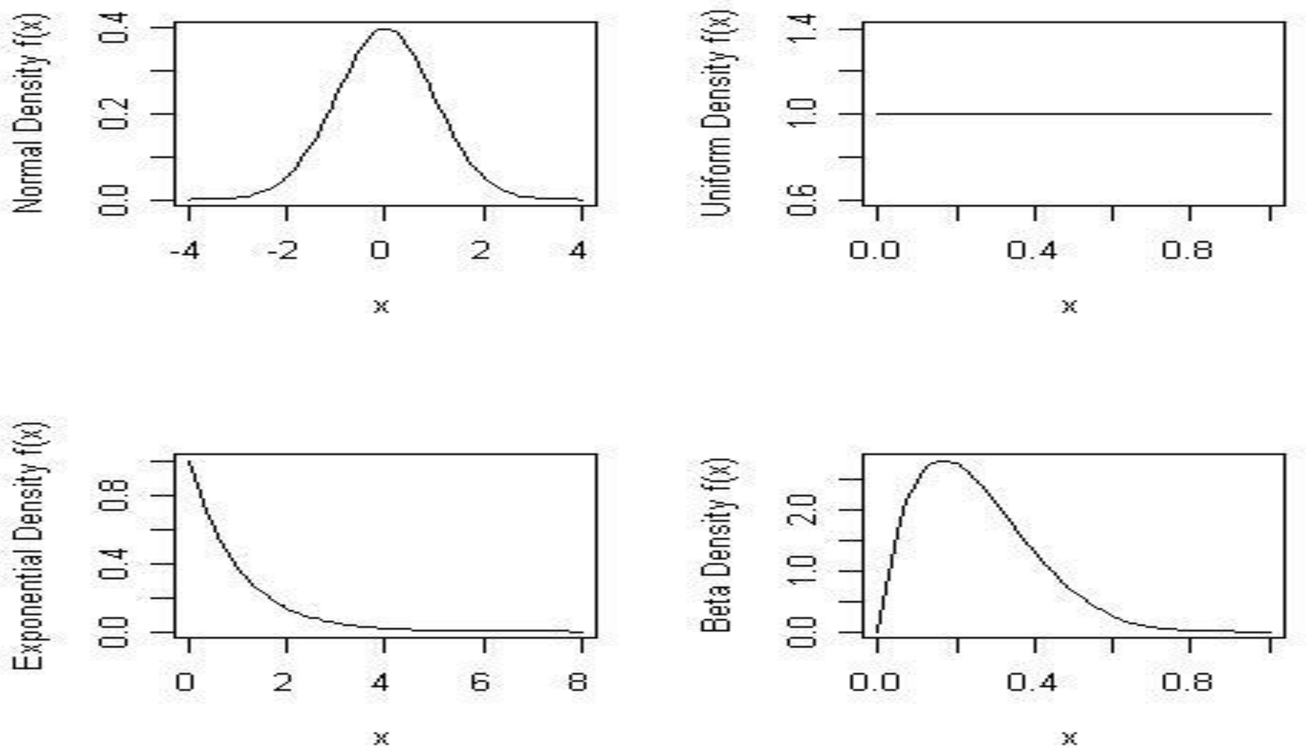
- The density function f(X), defined over interval(s) of real numbers now describes the probability distribution (pdf) of a continuous random variable *X*.
- *f* (*X*) is called the probability distribution, or probability density function for *X*.



Common Continuous Distributions

- Uniform distribution
- Normal distribution
- Exponential distribution
- Chi-squared distribution
- Gamma distribution
- Student-t distribution
- F-distribution

Plot of various pdf's



Continuous distributions in R

d*dist*(x,*parameter*)	density at x
p*dist*(x,*parameter*)	cumulative distribution function to x
q*dist*(p,*parameter*)	inverse cdf
r*dist*(n,*parameter*)	generates n random numbers from distribution

dist	Distribution	Parameters	Defaults
beta	beta	shape1, shape2	-, -
cauchy	Cauchy	loc, scale	0, 1
chisq	chi-square	df	-
exp	exponential	rate	1
f	F	df1, df2	-, -
gamma	Gamma	shape	-
lnorm	log-normal	mean, sd (of log)	0, 1
logis	logistic	loc, scale	0, 1
norm	normal	mean, sd	0, 1
t	Students t	df	-
unif	uniform	min, max	0, 1

Plotting pdf's in R

```
par(mfrow = c(2,2))
x <- c(-40:40)*0.1
y <- dnorm(x)
plot(x, y, type="l", ylab="Normal Density f(x)")
x <- c(0:40)/40
y <- dunif(x)
plot(x, y, type="l", ylab="Uniform Density f(x)")
x <- c(0:80)*0.1
y <- dexp(x,1)
plot(x, y, type="l", ylab="Exponential Density f(x)")
x <- c(0:40)/40
y <- dbeta(x,2,6)
plot(x, y, type="l", ylab="Beta Density f(x)")
```

Uniform Distribution

$X \sim U(A, B)$ with p.d.f.

$$p(x) = \frac{1}{B - A}, \quad A \leq x \leq B.$$

Simple property:

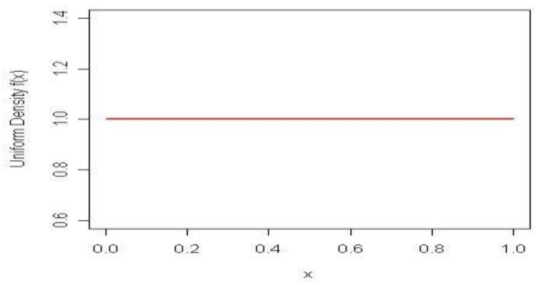
1. In R: min=A, max=B, distribution function=$unif$
2. $E(X) = \frac{A+B}{2}$
3. $Var(X) = \frac{(B-A)^2}{12}$

Special case: $A = 0$, $B = 1$.

Special U(0,1) Distribution

$X \sim U(0,1)$ with p.d.f.

$$p(x) = 1, \quad 0 \leq x$$



Simple property:

1. In R: default values min=0, max=1, distribution function=uni
2. $E(X) = \frac{1}{2}$.
3. $Var(X) = \frac{1}{12}$.

Normal distribution

http://www.youtube.com/watch?v=e-K0LQeEexk&feature=relmfu

$X \sim N(\mu, \sigma^2)$ with p.d.f.

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

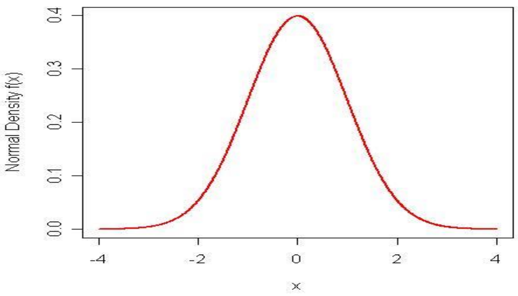
Simple property:

1. In R: mean=μ, sd=σ, distribution function=$norm$.
2. $E(X) = \mu$
3. $Var(X) = \sigma^2$

Standard N(0,1) distribution

Standard normal distribution: $Z \sim N(0,1)$ with p.d.f.

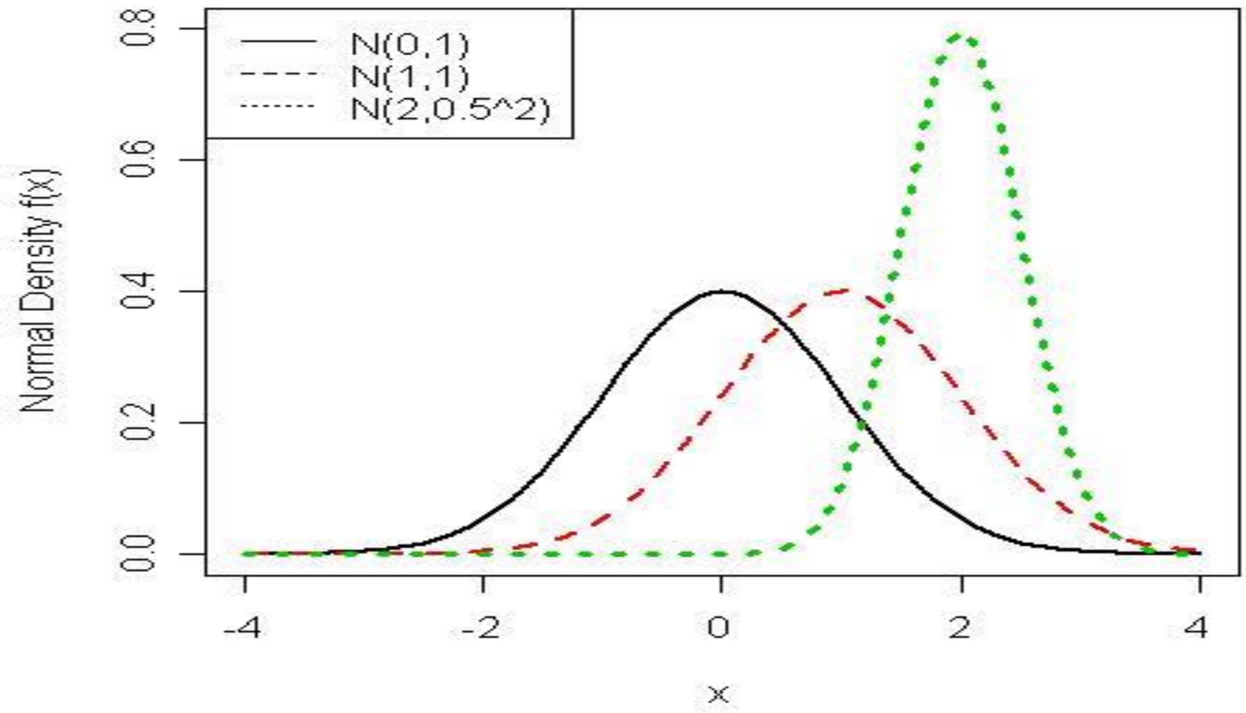
$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$



Simple property:

1. In R: default mean=0, sd=1, distribution function=$norm$.
2. $E(Z) = 0$
3. $Var(Z) = 1$

Plot of normal distributions



Normal Distributions in R

```
x <- c(-40:40)*0.1
y1 <- dnorm(x,0,1)
y2 <- dnorm(x,1,1)
y3 <- dnorm(x,2,0.5)
y <- cbind(y1,y2,y3)
matplot(x, y, type="l",
ylab="Normal Density f(x)",
lwd=c(2.5,2.5,3.5))
legend("topleft",
c("N(0,1)","N(1,1)","N(2,0.5^2)"),
lty=c(1,2,3))
```

.. Normal probabilities in R

- `dnorm(x, mean = 0, sd = 1, log = FALSE)`
- `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`
- `qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)`
- `rnorm(n, mean = 0, sd = 1)`

Sampling pd's from X ~N(0,1)

- The Student's t-distribution with n degrees of freedom: emerges from taking a sample fo size n iid from X~N(0, 1) to estimate X's mean, when σ is unknown. Get a `fatter" distribution that tends to X as n grows larger and larger.
- The Chi-distribution χ^2 likewise emerges from taking the distance (sum of squares) from the origin of the sample vector of X of dimension n. Get a `fatter" distribution that tends to X as n grows larger and larger

Computing general normal probability distributions

- For $X \sim N(\mu, \sigma^2)$,
  $P(a < X < b)$ is the area under the appropriate normal curve.
- We standardize a distribution by converting values into z-scores, i.e., rescaling it to differences from the mean μ (in standard deviation σ units)

$$z = \frac{x - \mu}{\sigma}$$

Computing normal probability Example 1: Z ~N(0,1)

```
>#P(Z <= 1)
> pnorm(1)
    [1] 0.8413447
> #P(Z <= -1.96)
> pnorm(-1.96)
    [1] 0.02499790
> #P(Z >1) = 1 - P(Z <= 1)
> 1-pnorm(1)
    [1] 0.1586553
> pnorm(1, lower.tail=FALSE)
    [1] 0.1586553
> #P(-2 < Z  <= 2) = P(Z <= 2) - P(Z <= -2)
> pnorm(2)-pnorm(-2)
    [1] 0.9544997
```

Computing normal probability Example 2: $X \sim N(\mu, \sigma^2)$.

Let X be the SAT Math score, which is normally distributed with $X \sim N(\mu, \sigma^2)$.

- If we assume that the mean score is 500 and standard deviation is 100. Find P(X< 700), P(200 < X <=800), and the score for the 99-th percentile.

```
> #X ~ N(500, 100^2)
> #P(X < 700)
> pnorm(700, mean=500, sd=100)
      [1] 0.9772499
> #P(200 < X <= 800)= P(X <= 800) - P(X<=200)
> pnorm(800, mean=500, sd=100)-pnorm(200, mean=500, sd=100)
      [1] 0.9973002
> #Finding 99 percentile
> qnorm(0.99, mean=500, sd=100)
      [1] 732.6348
```

Problems to consider

Let X be the SAT Math score, assuming it is normally distributed with $X \sim N(\mu, \sigma^2)$

1. Assume that the mean score is 500 but the std is unknown. If the 99-th percentile is 725, find its standard deviation.
2. If both the mean score and its std are unknown but we know the 99-th percentile is 725 and 60-th percentile is 550, find its mean and standard deviation.

Significance of mean and std

- Chebyshev's Inequality for RV X with $\mu=E(X)$:

  $Pr(|X-\mu| \geq k\sigma) \leq 1/k^2$, for all $k > 0$

- In particular,

  $Pr(|X-\mu| \geq \sqrt{2}\ \sigma) \leq ½$

  $Pr((|X-\mu| \geq 2\sigma) \leq ¼$

  (75% of the data lies within 2 significance units of the mean)

  $Pr((|X-\mu| \geq 5\sigma) \leq 0.04$

  (96% of the data lies within 5 significance units of the mean)

- Quite general but poor bound, can be tightened with knowledge of specific X

- Markov's inequality:

  $Pr(|X| \geq a) \leq \mu/a$, for all $a > 0$

  e.g. no more than 1/5 of any population can have more than 5 times the average income.

Sampling

- In practice, full distributions are rarely fully known; we can only sample them and try to infer/approximate the true pdf's
- Statistics (such as sample mean, sample proportion, or sample variance) vary from sample to sample and hence they are random variables (on a differnt sample space
- The probability distributions for statistics are called sampling distributions
- Important to calculate/know them for

  assessing *how accurate* approximations are

.. Sampling the normal distribution

```
> rnorm(5, 0, 1)
    [1] -0.88614298  0.08983277  1.82616196 -
1.83386114 -0.03037911
> rnorm(5, 1, 0.5)
[1] 0.6464083 1.0125589 0.3020117 1.4480203 0.8132631
> options(digits=2)  # set digits for display
> rnorm(5, 0, 1)
    [1] -0.21 -0.37 -1.05  0.36 -0.48
> rnorm(5, 1, 0.5)
    [1] 1.30 1.91 0.82 1.99 0.97
> rnorm(5, mean=1, sd=0.5)
    [1] 0.77 1.46 2.48 1.28 0.56
```

.. Sampling: uniform/exp distribution

- Sampling from a uniform distribution
  - `> options(digits=2)   # set digits for display`
  - `> runif(n=10, min=0, max=10)`

    [1] 7.48 8.67 8.03 7.31 1.71 8.27 0.84 0.80 5.50 8.83
  - `> runif(8)`

    [1] 0.83 0.05 0.97 0.66 0.27 0.86 0.17 0.65
- Sampling from an exponential distribution
  - `> rexp(n=10, rate=3)`

    [1] 0.181 0.054 0.085 0.233 0.149 0.768 0.019 0.274 0.907 1.720
  - `> rexp(n=10, rate=1/3)`

    [1] 5.33 12.28 6.49 10.61 3.05 7.69 0.33 1.91 0.44 0.29

.. Sampling: distributions

http://www.youtube.com/watch?v=NBRp6HuN_wk&feature=relmfu

- Approximated with simulation techniques:
  - Study the distribution of its sample mean
  - Use R to simulate several (say, 1000) random samples of size n from various distributions.
- How can we be sure these samples are realistic?
- Follows from the Central Limit Theorem

.. Sampling: distributions

- Have unknown distribution $X(\mu,\sigma)$
- A sample is
  - A sequence $\{X_1,X_2, \ldots X_i, \ldots\}$ of independent, identically distributed (i.i.d.) $X_i \sim X(\mu,\sigma)$; OR
  - The actual outcomes obtained in a finite initial segment of such
- Many new random variables can be defined from a sample $S \sim \{X_1,X_2, \ldots X_n\}$, such as

$S_n = X_1+X_2+ \ldots +X_n$ or $A_n= (X_1+X_2+ \ldots +X_n)/n$

  - What are their distributions like?
  - Is mean of the sample $\bar{x}$ = E(A) = μ of original? How about variances σ^2, std's σ? S_n ?

.. Sampling: distributions

http://www.youtube.com/watch?v=NBRp6HuN_wk&feature=relmfu

- Central Limit Theorem (CLT)
  - The sample means $\bar{x}_n = E(A_n)$ approach the standard normal distribution with mean μ, regardless of their original distribution X.
  - The z-scores converge to the standard normal distribution, regardless of the original X.
- Law of Large Numbers (LLN)

  With virtual certainty, $E(A_n) \rightarrow \mu$ as $n \rightarrow \infty$
  - Weak (in probability):

    $\Pr(|A_n - \mu| \geq \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$ **(for all** $\varepsilon > 0$**)**
  - Strong (almost surely/always):

    $\Pr(\lim E(A_n) = \mu) = 1$, i.e. $E(A_n) \rightarrow \mu$ a.surely.

Central Limit Theorem (CLT): the **Sample** Mean

Let $\bar{x}$ be the sample mean of size n from a population with mean=μ, and s.d. =σ. When the sample size n is large,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1).$$

If σ is unknown, we can estimate it by its sample variance, s^2, and

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \approx N(0, 1).$$

.. CLT: B(n,p)

Let $X \sim B(n, \pi)$ and $\hat{\pi} = X/n$ be the sample proportion. When the sample size n is large,

$$z = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \approx N(0, 1).$$

.. Sampling: distributions

- Approximated with simulation techniques:
  - Study the distribution of its sample mean
  - Use R to simulate several (say, 1000) random samples of size n from various distributions.
- How can we be sure these samples are realistic/suitable for inference?
  size n >=32 (large) or size n>=13 (small)
- Can be proved formally by Central Limit Theorem.

Verify the distribution of sample means via simulation

1. We use R to simulate a random sample of n (say, n=25 or n=100) variates from any distribution (e.g. normal, binomial, Poisson,…)
2. For the sample obtained, compute its sample mean,
3. Repeat Step 1 and Step 2 for 1000 times to produce 1000 $\bar{x}$ different and z-scores.
4. Compute the average $\bar{x}$ and standard deviation of 1000 sample means. Compare them with CLT.
5. Plot the histogram of $\bar{x}$ and z's.

About the CLT and LLNs

https://en.wikipedia.org/wiki/Central_limit_theorem

- CLT was proposed by de Moivre (1733) and later proved by Laplace (1812), Chebyshev (1890s) and Lyapunov (1901).

- "I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along."

Sir Francis Galton (1889)

Practice problem

- Write R code to implement the previous simulation procedure.
- You can also standardize the values of sample means by

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

and plot the histogram of z's.

Questions?

Model assumptions

- Usually, a statistical method assumes that the error component is a random sample from a (normal) population with zero mean and constant variance.
  - elements of the error term are independent, identically distributed (i.i.d.) ~ $N(0,\sigma^2)$.
- Two questions:
  - Are the methods sensitive to a model failure ?
  - How to check the validity of the assumption ?

Assuring independence

- Failure of independence assumption is a common reason for invalid statistical inference for most statistical methods.
- It is possible (but usually hard) to check the independence of the observed data. One should concentrate on the issue of randomization in experimental design and the sampling plan used.
- We can use some statistical methods (so called robust) to rely less on the data independence assumption.

Checking for normality

- Normality assumption is very common for most statistical methods.
- Because of the CLT, most statistical methods rely less on the data normality assumption.
- When the sample size is small (say n<10), it is impossible to check normality of the distribution.
- When the sample size is moderately large, it is possible to detect gross departure from normality.

How ?

Do a normal probability plot.

Normal probability plot in R

- **Description**
- qqnorm produces a normal plot of the values in y.
- qqline adds a line to a normal plot which can detect departure to normality.
- qqplot produces a QQ plot of two datasets.

- If the plot shows gross departure from the (theoretical) diagonal line, that is strong evidence against the normality assumption.

Normal probability plots in R

- Checking (approximate) normality of student t-distribution with various pdf's
  - We will discuss t-distribution in the next chapter
  - The larger pdf, the closer to the distribution of N(0,1) the distribution of t is known to be.
- We generate a random sample of size 200 from different t-distributions and check its normality:

```
par(mfrow = c(1,2))
y <- rt(200, df = 2); qqnorm(y); qqline(y, col = 2)
y <- rt(200, df = 30); qqnorm(y); qqline(y, col = 2)
```

Normality plots of t(2), t(30)