Random Variables

In set 3 of the notes, we study the basic probability theory, where the outcome of an experiment can be anything. In order to make analytical analysis, one need to transform the outcome of an experiment to numerical values.

Note: Categorical variables can also represented by binary values.

Random variable:

- For a given sample space S of some experiment, a random variable (rv) is any rule that associates a number with each outcome in S. In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.
- We usually let **X** stand for the random variable.

Examples:

Consider the experiment of tossing a fair coin three times independently. Define the random variable X to be the number of heads obtained in the three tosses. A complete enumeration of the value of X for each point in the sample space is:

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X(s)	3	2	2	2	1	1	1	0

Consider the experiment that we toss coin until first head shows up. Define the random variable to be the number of Tails before the first Head shows up.

s	H	TH	TTH	TTTH	TTTTH	….
X(s)	0	1	2	3	4	….

Consider the experiment of tossing a fair dice two times independently. Define the random variable X to be the summation of the two value.

Consider the experiment where we measure the chemical reaction time. We can define a random variable by an identity function.

Two basic types of Random variable:

Discrete: is an rv whose possible values is countable.

Continuous: is an rv whose possible values consists either of all numbers in a single interval on the number line or all numbers in a disjoint union of such intervals, and no possible value of the variable has positive probability, that is, P(X = c) =0 for all possible c.

Probability distribution: The probability distribution of a random variable is tells us the randomness of this random variable. This randomness is complete determined by the probability function on S, and the random variable function X(s).

Probability distribution of a discrete random variable

Terminology and Notation:

- Capital letters, such as X, are used to represent the random variable.
- Lower case letters, like x, refer to particular values taken by the random variable.

Definiton:

For any real value x:

$$P(X=x)=P(\{s:X(s)=x\})=P(\text{all } s\in S:X(s)=x)\ .$$

For any set B of real values:

$$P(X\in B)=P(\{s:X(s)\in B\})=P(\{\text{all } s\in S:X(s)=x\})=\sum_{x\in B}P(X=x)\ .$$

The **probability mass function** of a discrete variable is defined as $p(x)=P(X=x)$.

Note: The probability distribution of discrete rv also satisfies all the axioms.

Examples:

Consider also the experiment of tossing a fair coin three times independently. Define the random variable X to be the number of heads obtained in the three tosses. The rv is:

s	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X(s)	3	2	2	2	1	1	1	0

The pmf is:

x	0	1	2	3
p(x)	1/8	3/8	3/8	1/8

Consider the experiment that we toss coin until first head shows up. Define the random variable to be the number of Tails before the first Head shows up.

s	H	TH	TTH	TTTH	TTTTH	….
X(s)	0	1	2	3	4	….

The pmf is:

x	0	1	2	…..
p(x)	1/2	1/4	1/8	…..

Consider a group of five potential blood donors—a, b, c, d, and e—of whom only a and b have type O+ blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified. Let the rv be the number of typings necessary to identify an O+ individual. Then the pmf of Y is

x	1	2	3	4
p(x)				

Note: the pmf completely determines the randomness of the discrete random variable. Or we can claim that we know everything about this rv once we have the knowledge of its pmf.

Cumulative distribution function (cdf): F(x) of a discrete rv variable X with pmf p(x) is defined for every number x by

$$F(x)=P(X\leqslant x)=\sum_{y\leqslant x} p(y)$$

Note: the cdf is not only defined on the possible values of X, but any value on R.

Example:
Consider also the experiment of tossing a fair coin three times independently. Define the random variable X to be the number of heads obtained in the three tosses. The cdf of this rv is

x	$(-\infty,0)$	[0,1)	[1,2)	[2,3)	[3, ∞)
F(x)	0	1/8	4/8	7/8	8/8

The plot of this function is a **step** function:

Note: cdf is a exactly equivalent representation of pmf. Given the cdf, we can also retrieve the pmf using $p(x)=F(x)-F(x\text{-})$ = the jump at x in the cdf plot.

Expected values (population mean)

Consider a census data, recall the formula for population mean:

$$mean = \text{average of all data points} = \frac{\sum_{\text{all possible } x \text{ values}} x\,N(x)}{N},$$

which is the weighted average based on frequency.

Definition of the expected value of a discrete rv:

$$E(X) = \mu_X = \sum_{\text{all possible } x} xp(x).$$

Example

Let X be number of children born up to and including the first boy. Assume p is the probability of having a toy in each birth, then $p(x) = p(1-p)^x$ for all positive integer x. Then the mean is

$$E(X) = \sum_{x=0}^{\infty} xp(1-p)^x = \frac{1}{p}.$$

Note: one need to be careful about infinite summation.

Sometimes interest will focus on the expected value of some function h(X) rather than on just E(X). Define a new rv Y=h(X), then

$$E(Y) = \sum_{\text{all possible } y} y\,P(Y=y) = \sum_{\text{all possible } y} y\,P(X \in \{x : h(x) = y\}) = \sum_{\text{all possible } x} h(x)\,p(x).$$

Proposition: the expected value of a function of discrete rv X is

$$E(h(X)) = \sum_{\text{all possible } x} h(x)\,p(x).$$

Example:
Consider also the experiment of tossing a fair coin three times independently. Define the random variable X to be the number of heads obtained in the three tosses. What is E(3-X)?

Proposition: $E(a*X+b) = a*E(X)+b$.

Population variance

Definition of the expected value of a discrete rv:

$$Var(X) = \sigma_x^2 = E(X-\mu_X)^2 = \sum_{\text{all possible x}} (x-\mu_X)^2\,p(x).$$

And the standard deviation is defined as $\sigma_X = \sqrt{\sigma_X^2}$.

Alternative formula $Var(X) = E(X^2) - (E(X))^2 = \sum_{\text{all possible } x} x^2 p(x) - \mu_X^2$.

Example

Let X be number of children born up to and including the first boy. Assume p is the probability of having a boy in each birth, then $p(x) = p(1-p)^{x-1}$ for all positive integer x. Then the variance is

$$Var(X) = \sum_{x=1}^{\infty} \left(x - \frac{1}{p}\right)^2 p(1-p)^x = \frac{1-p}{p^2}.$$

Proposition: $Var(a * X + b) = a^2 * Var(X)$.

Discrete Distribution I: Binomial and Bernoulli distributions

Suppose a pmf function is completely determined by some quantity. Such a quantity is called a parameter of the distribution. The collection of all probability distributions for different values of the parameter is called a family of probability distributions. In the section, we study some popular discrete distribution families.

Bernoulli trial: A Bernoulli trail is an experiment with two, and only two, possible outcomes. And A random variable X has a Bernoulli(p) distribution if its pmf follows:

$$P(X=1) = p, P(X=0) = 1-p,$$

where 0 and 1 stand for two different outcomes (usually called failure and success).

The mean and variance of a Bernoulli(p) random variable are easily seen to be EX = (1)(p) + (0)(1 − p) = p and VarX = $(1-p)^2 p + (0-p)^2 (1-p) = p(1-p)$.

Binomial experiment: consists of n indepedent Bernoulli trials with same parameter p. Let X be the number of successes among all these n trials, then X has a Binomial distribution, denoted as Bin(n, p).

pmf of Binomial distribution Bin(n,p):

$$p(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ if } x=1,\ldots,n$$

cdf of Binomial distribution Bin(n,p):

$$F(x;n,p) = \sum_{y \le x} \binom{n}{y} p^y (1-p)^{n-y}.$$

There is no simple formula for the cdf.

Mean and variance of Binomial distribution Bin(n,p)

$$E(X)=np\,;Var(X)=np(1-p).$$

Discrete Distribution II: hypergeometric and negative binomial

Suppose there are N balls (M red balls, and N-M black balls) in a urn. Every time, you draw a ball from the urn. Let X be the totall number of red balls among n draws.

If every time after you draw the ball, you actually put the ball back into the urn, i.e. draw with replacement, then X follows a Bin(n, M/N) distribution.

If every time after you draw the ball, you don't put the ball back into the urn, i.e. draw without replacement, then X follows the so-called **hypergeometric distribution** hm(n,M,N).

Alternative interpretation: Binomial experiment is drawing from an infinite population, and hypergeometric experiment is drawing from a finite population.

pmf of hm(n,M,N):

$$p(x;n,M,N)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},\ \text{if}\ \max(0,n-N+M)\leqslant x\leqslant\min(n,M)$$

Mean and variance of hm(n,M,N)

$$E(X)=\frac{nM}{N}\,;Var(X)=n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right).$$

What is the difference comparing with mean and variance of binomial rv?

Example: Five individuals from an animal population thought to be near extinction in a certain region have been caught, tagged, and released to mix into the population. After they have had an opportunity to mix, a random sample of 10 of these animals is selected.
(1) Let X be the number of tagged animals in the second sample. If there are actually 25 animals of this type in the region, what is the probability that (a) X = 2 ? (b) X is less than 2 ?
(2)S uppose the population size N is not actually known, so the value x is observed and propose a way to estimate N based on x.

Negative Binomial experiment: consists of indefinite number of indepedent Bernoulli trials with same parameter p. The experiment stops when r successes have been observed.
Let X be the number of failures that precede the r^{th} success, then X has a Negative Binomial distribution, denoted as NBin(r, p).

pmf of Nbin(r,p):

$$p(x;r,p)=\binom{x+r-1}{r-1}p^r(1-p)^x$$, if x is a nonnegative integer.

Mean and variance of hm(n,M,N)

$$E(X)=\frac{r(1-p)}{p};Var(X)=\frac{r(1-p)}{p^2}.$$

Discrete Distribution III: Poisson distribution

Poisson distribution: is a discrete distribution with pmf

$$p(x;\mu)=\frac{e^{-\mu}\mu^x}{x!}$$, for all nonnegative integer x.

Note: This is a legitmate pmf, which follows by the Taylor expansion of exponential function.

Mean and variance of poisson distribution

$$EX=VarX=\mu.$$

Why poisson distribution is important?

It is a limit of binomial distribution: $\lim_n \text{bin}(x;n,p_n)=p(x;\mu),\text{if } np_n=\mu.$

It is additive: the sum of two independent poisson rv's is still a poisson rv.

Poisson process.

R provides four utility functions for each of the many commonly used distributions:
r- for data simulation,
d- for probability density function (pdf),
p- for cumulative distribution function (cdf), and
q- for quantiles (inverse of cdf).

Sample code:

```
rbinom(7,5,0.6); rpois(10,5.5); rhyper(7,9,6,5) dbinom(0:5,5,0.6); dpois(0:10,5.5); dhyper(0:5,9,6,5)
pbinom(0:5,5,0.6); cumsum(dbinom(0:5,5,0.6)) qpois(c(0,.25,.5,.75,1),5.5); ppois(0:10,5.5)
```

Probability distribution of a continuous random variable

Continuous random variable can takes uncountable many possible values, and its probability for any single value is 0, that is, P(X = x) =0 for all possible x. Therefore, it is impossible to use probability mass function to characterize the distribution of a continuous variable.
We use a “population histogram” to discribe the randomness of a continuous rv.

Definiton: Given a continuous rv X, its **probability density function** (pdf) is the nonnegative function f that satisfies

$$P(X\in(a,b))=\int_a^b f(x)\,dx, \text{for any real values } b>a.$$

Geometrically, the probability value is the area under the density curve.

Properties:

- $\int_{-\infty}^{\infty} f(x)dx = P(X\in \mathrm{R}) = 1$
- $P(X\in[a,b])=\int_a^b f(x)\,dx.$
- If pdf is known, any probability calculation can be done by integral, e.g.,
  $P(X \text{ is smaller than a, or between b and c})=\int_{-\infty}^{a} f(x)\,dx+\int_b^c f(x)\,dx.$

Example: Let X be the distance between two randomly chosen consecutive cars on a freeway**.** Assume that X has a pdf as:

$$f(x)=\lambda e^{-\lambda(x-a)} \text{ if } x>a; \text{ and } f(x)=0, otherwise.$$

Verify the f(x) is a valid pdf, and calculate P(X< b).

Cumulative distribution function (cdf): F(x) of a continuous rv variable X with pdf p(x) is defined for every number x by

$$F(x)=P(X\leqslant x)=\int_{-\infty}^{x} f(y)\,dy$$

Propositions:

- F(x) is a continous, non-decreasing function;
- F(∞) = 1 and F($-\infty$) =0;

Example: Let X be the distance between two randomly chosen consecutive cars on a freeway**.** Assume that X has a pdf as:

$$f(x)=\lambda e^{-\lambda\cdot(x-a)} \text{ if } x>a; \text{ and } f(x)=0, otherwise.$$

Compute and plot the cumulative density function.

cdf and pdf are essentially the same presentation of distribution.

- $P(a<X\leqslant b)=F(b)-F(a)$
- $P(X>a)=1-F(a)$
- $f(x)=F'(x)$
- Compute the quantile by inversing cdf: F(m)=1/2; F(Q1)=0.25; F(Q3)=0.75.

Definition: The expected value of a continuous rv X, or h(X) is defined as, respectively,

$$E(X)=\int xf(x)dx; Eh(X)=\int h(x)f(x)dx.$$

Its variance is defined as $Var(X)=\int [x-E(X)]^2 f(x)dx$. Its standard deviation is defined as $\sqrt{Var(X)}$.

The following properties still hold for continuous random variable:

- The variance can be computed as $Var(X)=E(X^2)-(EX)^2$.
- $E(a*X+b)=a*E(X); Var(a*X+b)=a^2*Var(X)$.

Continuous Distribution I: Uniform distribution

Uniform distribution U(a,b): is a continous distribution with pdf

$$f(x;a,b)=\frac{1}{(b-a)} \text{ if } a<x<b, \text{and } f(x;a,b)=0 \text{ otherwise}.$$

The uniform distribution has a flat density curve and

$$E(X)=\frac{b-a}{2} \text{ and } Var(X)=\frac{1}{12}(b-a)^{2.}$$

What is the cdf of U(a,b)?

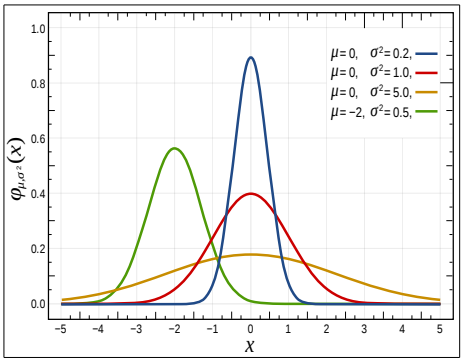
What is E(min(X-a, b-X))?

Continuous Distribution II: Normal distribution

Normal distribution $N(\mu,\sigma^2)$: is a continous distribution with pdf

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)} \text{ for all } x,$$

where μ is any real value, and σ is positive.



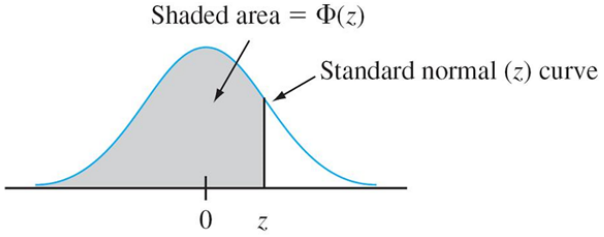
The normal distribution has a symmetric, bell-shape density curve and

$$E(X)=\mu \text{ and } Var(X)=\sigma^2.$$

Normal distribution is one the most important distribution. However the probability calculation, or the cdf for normal rv doesn't have a closed formula.

Calculation for standard Normal distribution $N(0,1^2)$

Notation: We use Z to denote the standard normal, that is normal distribution with mean 0 and unit variance; use $\phi(z)$ to denote the pdf of Z; and use $\Phi(z)=\int_{-\infty}^{z}\phi(u)du$ to denote the cdf of Z.

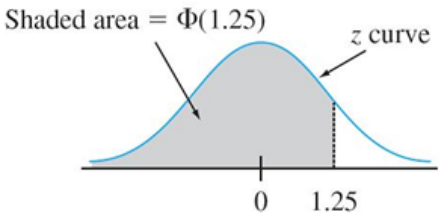


Normal Z-Table: $\Phi(z)=\int_{-\infty}^{z}\phi(u)du$ can not be computed by hand. Z-table provides values of $\Phi(z)$ for different z up to 2 decimals.

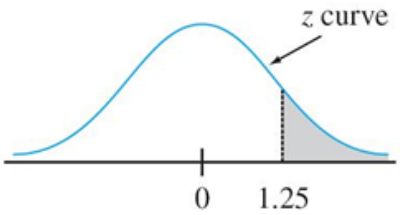
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.40	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879

Probability calculation:

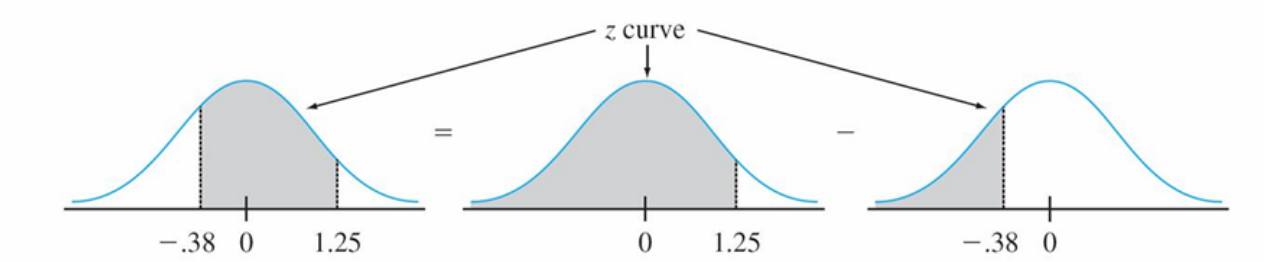
(a) $P(Z \leq 1.25) =$



(b) $P(Z > 1.25) =$

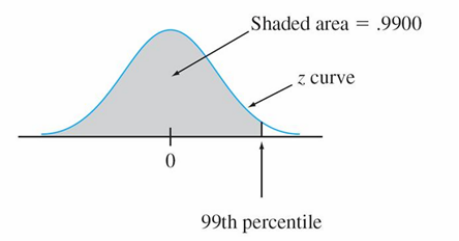


(c) $P(-.38 \leq Z \leq 1.25)$

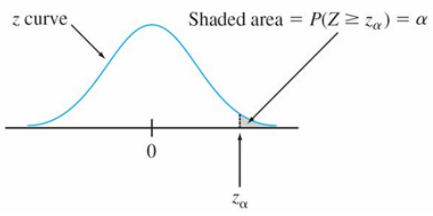


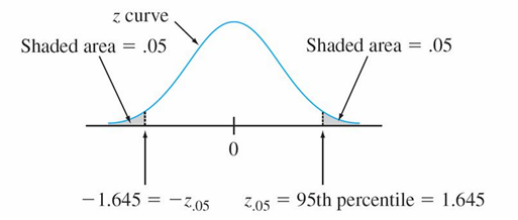
- **68.27%** of the probability is between -1 and 1.
- **95.45%** of the probability is between -2 and 2.
- **99.73%** of the probability is between -3 and 3.

Percentile calculation:



Notation: z critical values Z_α





Calculation for General Normal distribution $N(\mu, \sigma^2)$

Important connection between general normal and standard normal rvs:
If X follows $N(\mu, \sigma^2)$, then $(X-\mu)/\sigma$ follows standard normal distribution. Therefore, any calculation for general normal can be reduced to calculation of standard normal.

Probability calculation:

- $P(X<x)=P(Z<z)$ where $z=(x-\mu)/\sigma$.
- $P(x_1<X<x_2)=P(z_1<Z<z_2)$ where $z_i=(x_i-\mu)/\sigma$.
- **68.27%** of the observations fall within 1 standard deviation of the mean.
- **95.45%** of the observations fall within 2 standard deviations of the mean.
- **99.73%** of the observations fall within 3 standard deviations of the mean.

Example Assume that the wingspan of adult dragonflies is normally distributed with μ = 4 inches and σ = 0.25 inches. Let X represent the wingspan of a randomly chosen adult dragonfly.

a) Find the probability the wingspan of a randomly selected adult dragonfly is less than 4.3 inches.

First calculate the z score: $z = \dfrac{4.3-4.0}{0.25} = 1.20$ Look up cumulative probability for 1.20.

Solution: $P(X < 4.30) = P(Z < 1.20) = 0.8849$

b) Find the probability that a randomly selected adult dragonfly has a wingspan of more than 4.3 inches.

Solution: $P(X > 4.30) = P(Z > 1.20) = 1 - P(Z < 1.20) = 1 - 0.8849 = 0.1151$

c) Find the probability that a randomly selected adult dragonfly has a wingspan between 3.61 and 4.73 inches.

Find the z scores: $z_a = \dfrac{3.61-4.00}{0.25} = -1.56$ **and** $z_b = \dfrac{4.73-4.00}{0.25} = 2.92$

Solution:

$P(3.61 < X < 4.73) = P(-1.56 < Z < 2.92) = P(Z < 2.92) - P(Z < -1.56)$ = .9982 – .0594 = .9388

d) Find the probability that a randomly selected adult dragonfly has a wingspan less than 3.8 or greater than 4.2 inches.

Find the z scores: $z_a = \dfrac{3.8-4.00}{0.25} = -0.80$ **and** $z_b = \dfrac{4.2-4.0}{0.25} = 0.80$

Solution:

$$P(X < 3.8 \quad or \quad X > 4.2) = P(Z < -0.80) + \left(1 - P(Z < 0.80)\right) = .2119 + \left(1 - .7881\right) = .4238$$

Percentile (critical value) calculation: $Z_\alpha * \sigma + \mu$

Example

What wingspan is longer than 90% of all adult dragonfly wingspans?

The wingspan that is longer than 90% of all adult dragonfly wingspans is ____ inches.

Between what two lengths do the middle 95% of all adult dragonfly wingspans fall?

Approximating the Binomial Distribution (Central limit Theorem):

Let X be a binomial random variable Bin(n,p). Then, X has approximately a normal distribution with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, if sample size is large enough, namely $np \geq 10$ and $n(1-p) \geq 10$.

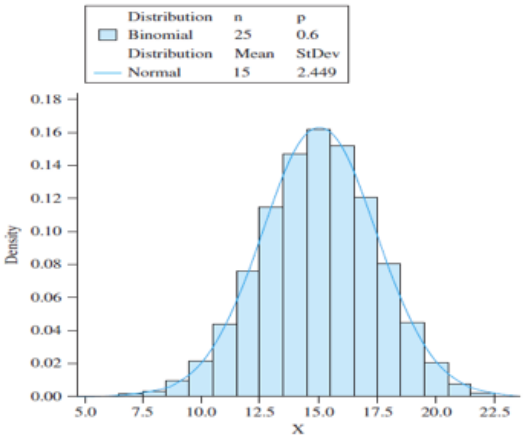


Figure 4.25 Binomial probability histogram for $n = 25$, $p = .6$ with normal approximation curve superimposed

Example: A coin is tossed 100 times. Estimate the probability that the number of heads lies between 40 and 60 (the word "between" in mathematics means inclusive of the endpoints).

Solution: The expected number of heads is 100·1/2 = 50, and the variance for the number of heads is 100 · 1/2 · 1/2 = 5. Thus, since n = 100 is reasonably large, we have

$$P(39.5 \leqslant X \leqslant 60.5) \approx P(40 \leqslant N(50,5) \leqslant 60) = P(-2.1 \leqslant Z \leqslant 2.1) \approx 0.9642$$

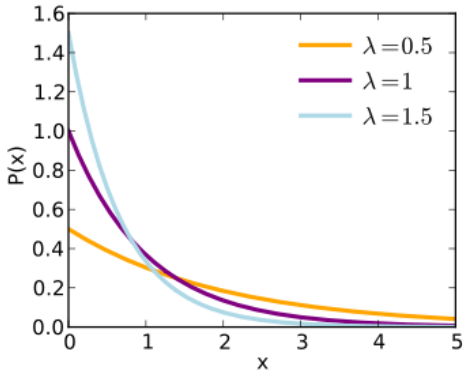
The actual value is .96480, to five decimal places.

Continuous Distribution III: Exponential and Gamma distribution

Exponential distribution: is a continous distribution with pdf

$f(x;\lambda) = \lambda e^{-\lambda x}$ if $x > 0$; and $f(x;\lambda) = 0$ otherwise,

given positive parameter λ.



Mean, variance, cdf, quantile and etc

$EX =$

$Var(X) =$

$F(x) =$

$M =$

Memoryless propery:

$P(X \geqslant t + t_0 | X \geqslant t_0)$

=

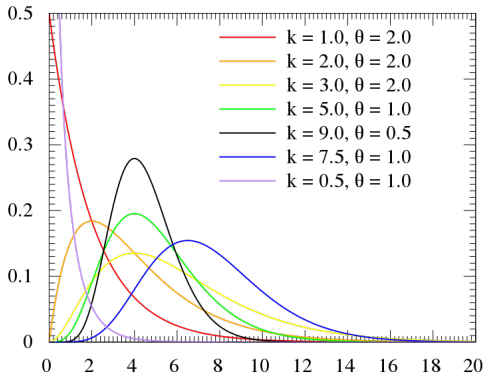
= $P(X \geqslant t)$

Gamma distribution $\text{Gamma}(\alpha,\beta)$: is a continous distribution with pdf

$$f(x;\alpha,\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} \text{ if } x>0; \text{ and } f(x;\lambda)=0 \text{ otherwise,}$$

given positive parameters $\alpha \text{and } \beta,$ where

$$\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}\,dx$$



Mean, variance

$$EX=\alpha\beta$$

$$Var(X)=\alpha\beta^2$$

Chi-squared distribution is a special case of Gamma distribution

$$\chi^2_{\nu}=\text{Gamma}(\alpha=\nu/2,\beta=2).$$

Probability Plots

Question: How can you decide if a set has normal distribution or exponential distribution?

- It is risky to assume data follows certain distribution without actually inspecting the data.
- Histograms and box plots can help, since it reveals the shape of the data distribution
- However, sometimes we need a more sensitive way to judge the adequacy of a Normal model (say if we assume the data follow normal distribution).

- The most useful tool for assessing it is another graph, the quantile plot (or probability plot).

Idea: Comparing Distribution quantile and Sample quantile.

Sample Percentile
Order the n sample observations from smallest to largest. Then the ith smallest observation in the list is taken to be the $[100(i-0.5)/n]^{th}$ percentile.

Probability plot
$[100(i-.5)/n]^{th}$ percentile of the distribution versus *i^{th} smallest sample observation*

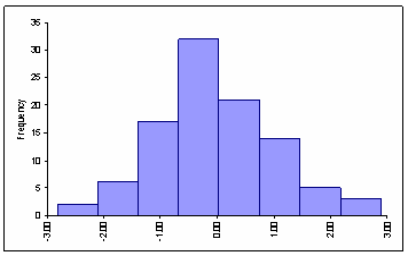
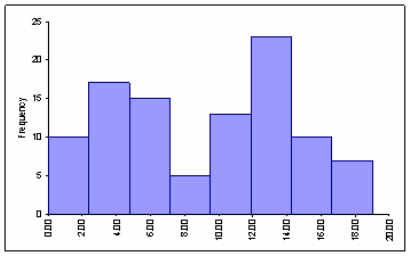
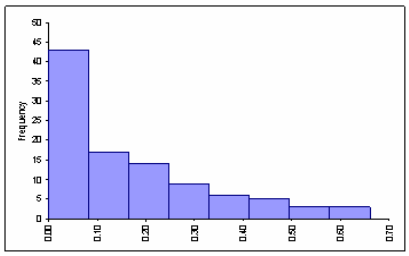
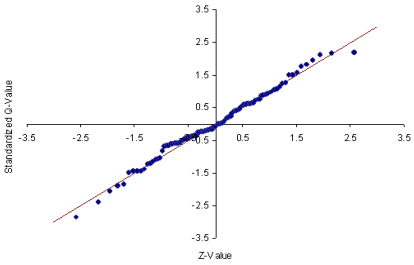
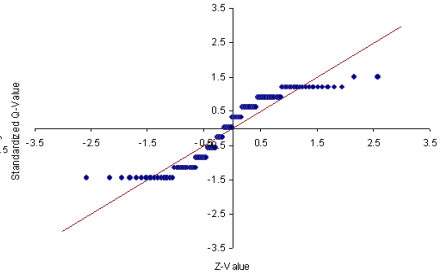
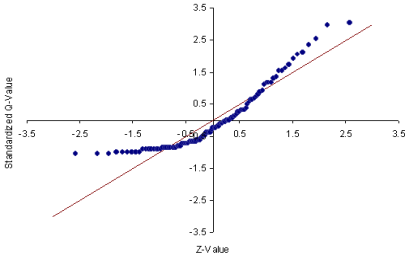
Diagnose
If the sample percentiles are close to the corresponding population distribution percentiles, the plotted points will then fall close to a 45-degree diagonal line. Substantial deviations of the plotted points from a diagonal line cast doubt on the assumption that the distribution under consideration is the correct one.

Normal probabilty Plots

[100(i-.5)/n]th percentile of the ***standard normal distribution*** versus *ith smallest sample observation*

Diagnose

If the true distribution is indeed a normal distribution, **not necessary standard normal**, the plotted points will then fall close to a straight line. Substantial deviations of the plotted points from a straight line cast doubt on the assumption that the distribution under consideration is the correct one.

Normal	Bimodal	Right skewed
		
		
Points on the line	"S" shape	"C" type shape

R provides four utility functions for each of the many commonly used distributions:
Sample code:

runif(7,0,1); rnorm(10,5.5,36); rexp(7,5); rgamma(10, shape=2, scale=3)

x<-rnorm(100)
qqnorm(x);qqline(x)