Probability Distributions, Confidence Intervals

CS 700
Jana Kosecka

Review

- Statistical Summarization of data
- Mean, median, mode, variance, skewness
- Quantiles, Percentiles,
- Issues of robustness
- Suitability of different metrics (harmonic vs, arithmethic mean, mean vs. mode)
- Histograms

Continuation

- Previous summarization obtained only based on some sample of the data from the population
- How confident are we in the measurements
- Need to understand sources of errors
- Typically making some assumption about their characteristic probability distributions
- Next review of some distribution
- Follow up estimation of confidences

Review of Probability Concepts

- Classical (theoretical) approach:

$$\frac{\text{No. Ways Event } A \text{ Can Occur}}{\text{Total Number of Events}}$$

process has to be known!

- Empirical approach (relative frequency):

$$\frac{\text{No. Times Result } A \text{ Occurred in the Experiment}}{\text{Total Number of Observations}}$$

- The relative frequency converges to the probability for a large number of experiments.

4

Review of Probability Rules

1. A probability is a number between 0 and 1 assigned to an event that is the outcome of an experiment: $P[A] \in [0,1]$
2. Complement of event A.
$$P[A] = 1 - P[\overline{A}]$$
3. If events A and B are mutually exclusive then
$$P[A \text{ or } B] = P[A] + P[B]$$
$$P[A \text{ and } B] = 0$$

Review of Probability Rules (cont'd)

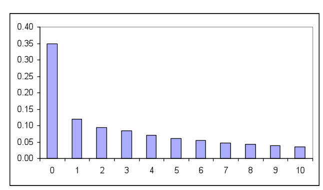
4. If events $A_1, \ldots, A_N$ are mutually exclusive and collectively exhaustive then:
$$\sum_{i=1}^{N} P[A_i] = 1$$
5. If events A and B are not mutually exclusive then:
6. Conditional Probability:
$$P[A \text{ or } B] = P[A] + P[B] - P[A \text{ and } B]$$
$$P[A \mid B] = \frac{P[A \text{ and } B]}{P[B]} = \frac{P[B \mid A]P[A]}{P[B]}$$

Review of Probability Rules (cont'd)

7. If events A and B are independent (i.e., P[A] = P[A|B] and P[B]=P[B|A]) then:
$$P[A \text{ and } B] = P[A,B] = P[A]P[B]$$
8. If events A and B are not independent then
$$P[A \text{ and } B] = P[A \mid B]P[B] = P[B \mid A]P[A]$$
9. Theorem of Total Probability: if events $A_1, \ldots, A_N$ are mutually exclusive and collectively exhaustive then
$$P[B] = \sum_{i=1}^{N} P[B \mid A_i]P[A_i]$$

Discrete Probability Distribution

- Distribution: set of all possible values and their probabilities.



Number of I/Os per Transaction	Probability
0	0.350
1	0.120
2	0.095
3	0.085
4	0.070
5	0.060
6	0.054
7	0.048
8	0.043
9	0.040
10	0.035
	1.000

- Cumulative distribution
$$F(x) = \Pr[X \le x] = \sum_{xi \le x} P(X = x_i) = \sum_{xi \le x} p(x_i)$$

Moments of a Discrete Random Variable

- Expected Value:

$$\mu = E[X] = \sum_{\forall i} X_i P[X_i]$$

- k-th moment:

$$\mu = E[X^k] = \sum_{\forall i} X_i^k P[X_i]$$

Number of I/Os per Transaction	Probability	For First Moment (average)	For Second Moment
0	0.350	0.000	0.000
1	0.120	0.120	0.120
2	0.095	0.190	0.380
3	0.085	0.255	0.765
4	0.070	0.280	1.120
5	0.060	0.300	1.500
6	0.054	0.324	1.944
7	0.048	0.336	2.352
8	0.043	0.344	2.752
9	0.040	0.360	3.240
10	0.035	0.350	3.500
	1.000	**2.859**	**17.673**

mean (2.859)
second moment (17.673)

9

Central Moments of a Discrete Random Variable

- k-th central moment:

$$E[(X-\overline{X})^k] = \sum_{\forall i} (X_i - \overline{X})^k P[X_i]$$

- The variance is the second central moment:

$$\sigma^2 = E[(X-\overline{X})^2] = E[X^2 + (\overline{X})^2 - 2X\overline{X}]$$
$$= E[X^2] + (\overline{X})^2 - 2(\overline{X})^2 =$$
$$= E[X^2] - (\overline{X})^2$$

10

Central Moments of a Discrete Random Variable

Number of I/Os per Transaction	Probability	For First Moment (average)	For Second Moment	For Second Central Moment
0	0.350	0.000	0.000	2.8609
1	0.120	0.120	0.120	0.4147
2	0.095	0.190	0.380	0.0701
3	0.085	0.255	0.765	0.0017
4	0.070	0.280	1.120	0.0911
5	0.060	0.300	1.500	0.2750
6	0.054	0.324	1.944	0.5328
7	0.048	0.336	2.352	0.8231
8	0.043	0.344	2.752	1.1365
9	0.040	0.360	3.240	1.5085
10	0.035	0.350	3.500	1.7848
	1.000	**2.859**	**17.673**	**9.4991**

average (2.859)
variance (9.4991)

11

Properties of the Mean

- The mean of the sum is the sum of the means.
- If X and Y are independent random variables, then the mean of the product is the product of the means.

$$E[X+Y] = E[X] + E[Y]$$

$$E[XY] = E[X]E[Y]$$

12

Discrete Random Variables

- Binomial
- Hypergeometric
- Negative Binomial
- Geometric
- Poisson

13

The Binomial Distribution

- Distribution: based on carrying out independent experiments with two possible outcomes:
  - Success with probability p and
  - Failure with probability $(1-p)$.
- A binomial r.v. counts the number of successes in n trials.
- Probability that we get k success in n trials is

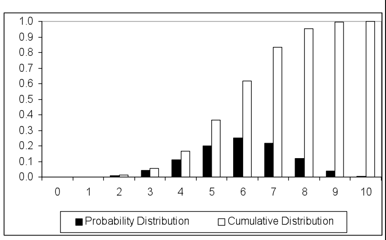
$$P[X=k]=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

14

The Binomial Distribution

Success Probability **0.6 (p)**
Number of Attempts **10 (n)**

Number of Attempts (k)	Probability k successful attempts in n	Cumulative
0	0.000105	0.000105
1	0.001573	0.001678
2	0.010617	0.012295
3	0.042467	0.054762
4	0.111477	0.166239
5	0.200658	0.366897
6	0.250823	0.617719
7	0.214991	0.832710
8	0.120932	0.953643
9	0.040311	0.993953
10	0.006047	1.000000

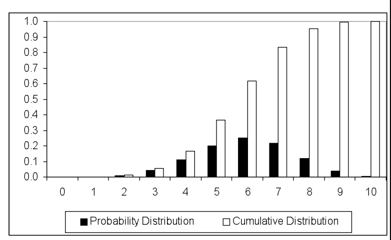


15

The Binomial Distribution

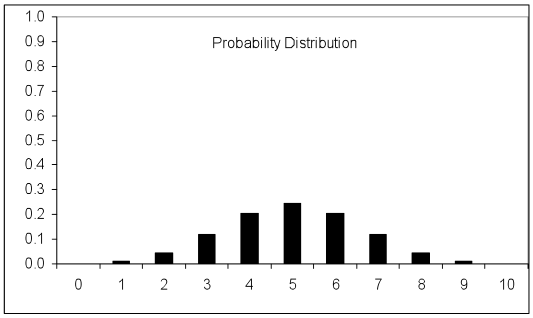
Success Probability **0.6 (p)**
Number of Attempts **10 (n)**

Number of Attempts (k)	Probability k successful attempts in n	Cumulative
0	0.000105	0.000105
1	0.001573	0.001678
2	0.010617	0.012295
3	0.042467	0.054762
4	0.111477	0.166239
5	0.200658	0.366897
6	0.250823	0.617719
7	0.214991	0.832710
8	0.120932	0.953643
9	0.040311	0.993953
10	0.006047	1.000000



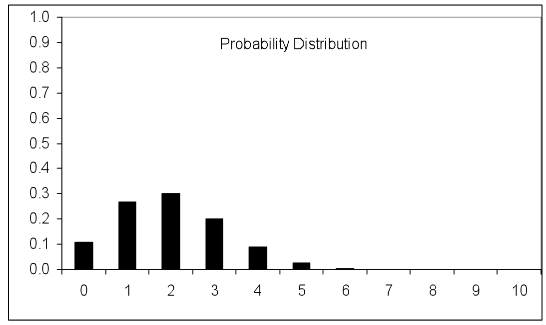
16

Shape of the Binomial Distribution



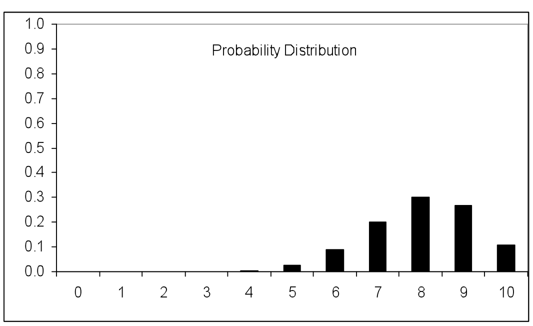
p = 0.5 symmetric for any *n*

Shape of the Binomial Distribution



p = 0.2 right skewed

Shape of the Binomial Distribution



p = 0.8 left skewed

Moments of the Binomial Distribution

- Average: *n p*
- Variance: $np(1-p)$
- Standard Deviation: $\sqrt{np(1-p)}$
- Coefficient of Variation:

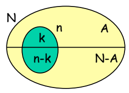
$$\frac{\sqrt{np(1-p)}}{np} = \sqrt{\frac{1-p}{np}}$$

Hypergeometric Distribution

- Binomial was based on experiments with equal success probability (n-draws with replacements)
- Hypergeometric: not all experiments have the same success probability (n-draws without replacements)
- Given a sample size of *n* out of a population of size *N* with *A* known successes in the population, the probability of k successes is

choose k successes out of A successes in the population

choose (n-k) failures from N-A failures in the population



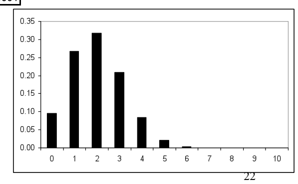
$$P[X = k] = \frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}}$$

total # of possible samples

Hypergeometric Distribution

No. successes in sample k	sample size n	no. successes in population A	population size N	
0	20	10	100	0.09511627
1	20	10	100	0.26793316
2	20	10	100	0.31817063
3	20	10	100	0.20920809
4	20	10	100	0.08410730
5	20	10	100	0.02153147
6	20	10	100	0.00354136
7	20	10	100	0.00036793
8	20	10	100	0.00002300
9	20	10	100	0.00000078
10	20	10	100	0.00000001

In Excel:
Pr[X=k]=HYPGEOMDIST (k,n,A,N)



Moments of the Hypergeometric

- Average: $\frac{nA}{N}$
- Standard Deviation: $\sqrt{\frac{nA(N-A)}{N^2}}\sqrt{\frac{N-n}{N-1}}$
- If the sample size is less than 5% of the population, the binomial is a good approximation for the hypergeometric.

Negative Binomial Distribution

- Probability of success is equal to *p* and is the same on all trials.
- Random variable X counts the number of trials until the *k*-th success and r failures is observed.
- Keep on observing until predefined number *r* of failures occurred *X ~ NB(r,p)*
- As opposed to binomial *X~B(n,p)*

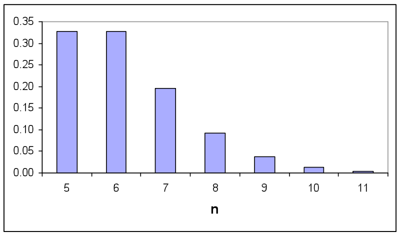
$$P[X = k] = \binom{k+r-1}{k-1}(1-p)^r p^k$$

- If r is integer waiting time in Bernoulli process

Negative Binomial Distribution

Success probability	0.8

k	n	Prob[X=n]
1	1	0.800000
1	2	0.160000
1	3	0.032000
1	4	0.006400
5	5	0.327680
5	6	0.327680
5	7	0.196608
5	8	0.091750
5	9	0.036700
5	10	0.013212
5	11	0.004404



In Excel:
Pr [X=n] = NEGBINOMDIST (n-k,k,p)
Pr [X=r+k] = NEGBINOMDIST (r,k,p)

Moments of the Negative Binomial Distribution

- Average: $\frac{k}{p}$
- Standard Deviation: $\sqrt{\frac{k(1-p)}{p^2}}$
- Coefficient of Variation: $\sqrt{\frac{1-p}{k}}$

Geometric Distribution

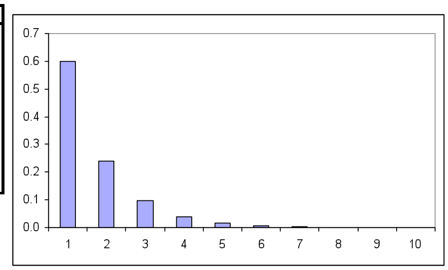
- Special case of the negative binomial with *k*=1.
- Probability of failures until the first success
- Probability that the first success occurs after *n* trials is

$$p[X = n] = p(1-p)^{n-1} \qquad n = 1,2,...$$

Geometric Distribution

Success probability	0.6

n	P[X=n]
1	0.6000
2	0.2400
3	0.0960
4	0.0384
5	0.0154
6	0.0061
7	0.0025
8	0.0010
9	0.0004
10	0.0002



Moments of the Geometric Distribution

- Average: $\frac{1}{p}$
- Standard Deviation: $\sqrt{\frac{1-p}{p^2}}$
- Coefficient of Variation: $\sqrt{1-p} \le 1$

Poisson Distribution

- Used to model the number of arrivals over a given interval, e.g.,
  - Number of requests to a server
  - Number of failures of a component
  - Number of queries to the database.
- A Poisson distribution usually arises when arrivals come from a large number of independent sources.

Poisson Distribution

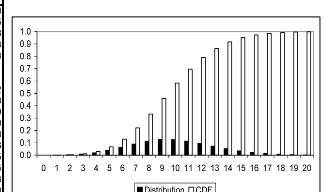
- Distribution: $P[X=k] = \frac{\lambda^k e^{-\lambda}}{k!} \quad k = 0,1,\ldots,\infty$
- Parameter λ number of expected events during time interval
- Counting arrivals in an interval of duration *t*:

$$P[k \text{ arrivals in } [0,t)] = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \quad k = 0,1,\ldots,\infty$$

- Average = λ

Poisson Distribution

Lambda	10	
K	Poisson Distribution	CDF
0	0.00005	0.0000
1	0.00045	0.0005
2	0.00227	0.0028
3	0.00757	0.0103
4	0.01892	0.0293
5	0.03783	0.0671
6	0.06306	0.1301
7	0.09008	0.2202
8	0.11260	0.3328
9	0.12511	0.4579
10	0.12511	0.5830
11	0.11374	0.6968
12	0.09478	0.7916
13	0.07291	0.8645
14	0.05208	0.9165
15	0.03472	0.9513
16	0.02170	0.9730
17	0.01276	0.9857
18	0.00709	0.9928
19	0.00373	0.9965
20	0.00187	0.9984



In Excel:
P[X=k] = POISSON (k,λ,FALSE)
P[X≤k] = POISSON (k,λ,TRUE)

Continuous Random Variables

Relevant Functions

- Probability density function (pdf) of r.v. X: $f_X(x)$

$$P[a \le X \le b] = \int_a^b f_X(x)dx$$

- Cumulative distribution function (CDF):

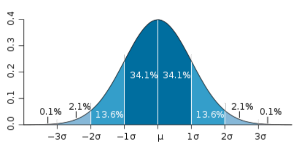
$$F_X(x) = P[X \le x]$$

- Tail of the distribution (reliability function):

$$R_X(x) = P[X > x] = 1 - F_X(x)$$

Continuous Probability Distribution

- Distribution provides probability for all possible values
- Normal distribution, Gaussian distribution, Bell curve



- Cummulative probability distribution

$$F(x) = \Pr[X \le x] = \int_{-\infty}^{x} f(t)dt$$

Moments

- k-th moment: $E[X^k] = \int_{-\infty}^{+\infty} x^k f_X(x)dx$
- Expected value (mean): first moment

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f_X(x)dx$$

- k-th central moment:

$$E[(X-\mu)^k] = \int_{-\infty}^{+\infty} (x-\mu)^k f_X(x)dx$$

- Variance: second central moment

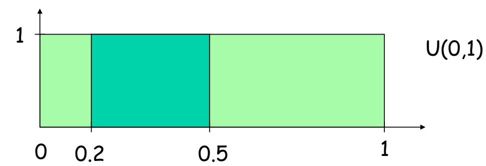
$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx$$

The Uniform Distribution

- pdf: $f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$
- Mean: $\mu = \frac{a+b}{2}$
- Variance: $\sigma^2 = \frac{(b-a)^2}{12}$

37

The Uniform Distribution



P[0.2<X<0.5]=(0.5-0.2)x1.0=0.3

38

The Normal Distribution

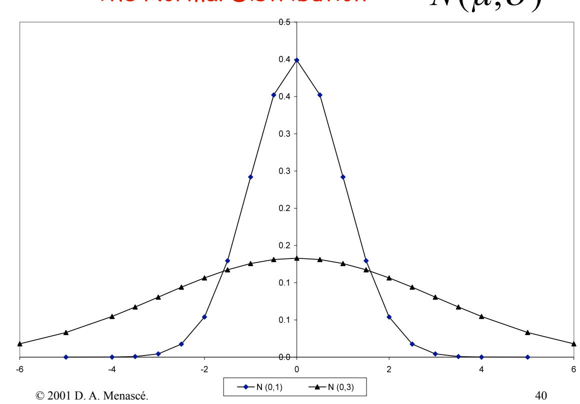
- Many natural phenomena follow a normal distribution.
- The normal distribution can be used to approximate the binomial and the Poisson distributions.
- Two parameters: mean and standard deviation.

$$f(x) = N(\mu, \sigma)$$

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

39

The Normal Distribution $N(\mu, \sigma)$



40

The Standard Normal Distribution

- Standard – zero mean and unit variance
- To use tables for computing values related to the normal distribution, we need to standardize a normal r.v. as

$$Z = \frac{X - \mu}{\sigma}$$

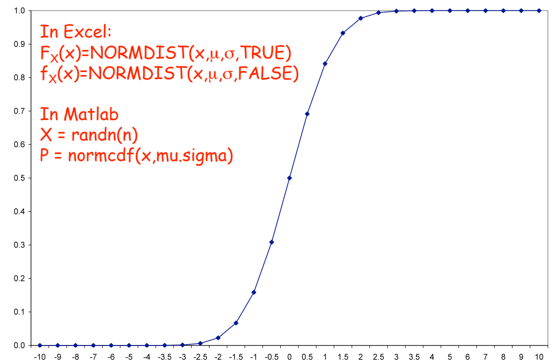
standard normal score

- Given X, compute a Z value z.
- Find the area value in a Table (Prob [0<Z<z]).

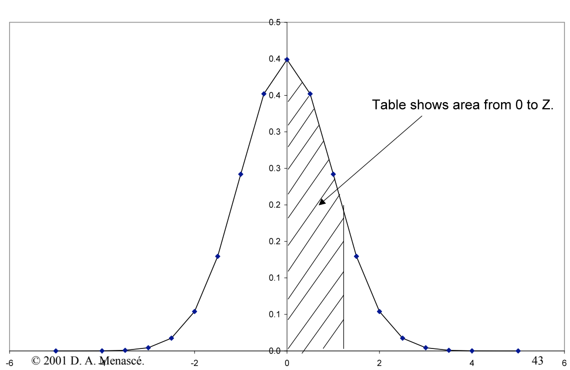
Normal CDF

In Excel:
$F_X(x)$=NORMDIST(x,μ,σ,TRUE)
$f_X(x)$=NORMDIST(x,μ,σ,FALSE)

In Matlab
X = randn(n)
P = normcdf(x,mu.sigma)



Using Normal Tables



The Normal as an Approximation to the Binomial Distribution

- The normal can approximate the binomial if the variance of the binomial (works for large n)

$$np(1-p) \geq 10$$

- Binomial:

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

- Transformation:

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

- To avoid exact calculations for large n

The Normal as an Approximation to the Binomial Distribution

- Consider a binomial r.v. X with average 50 and variance 25. What is

$$P[50 \le X \le 60]?$$

- Transformation: $Z = \frac{X-50}{\sqrt{25}} = \frac{60-50}{5} = 2.0$

- Using the table, the area between 50 and 60 for Z=2.0 is 0.4772. So,

$$P[50 \le X \le 60] = 0.4772$$

The Normal as an Approximation to the Poisson Distribution

- The normal can approximate the Poisson distribution if $\lambda > 5$.

- Poisson: $\mu = \lambda$

$$\sigma = \sqrt{\lambda}$$

- Transformation:

$$Z = \frac{X-\lambda}{\sqrt{\lambda}}$$

The Lognormal Distribution

- It is a random variable such that its natural logarithm has a normal distribution.

$$f_X(x) = \frac{1}{x\sqrt{2\pi}\sigma_{\ln X}} e^{-(1/2)\left[(\ln x - \mu_{\ln X})/\sigma_{\ln X}\right]^2} \qquad x > 0$$

Y = ln X (X and Y are r.v.'s) and Y = N(μ,σ)

- Suitable for effect which have multiplicative factors (e.g. long term discount factor as product of short term discounts, attenuation of a wireless channel)

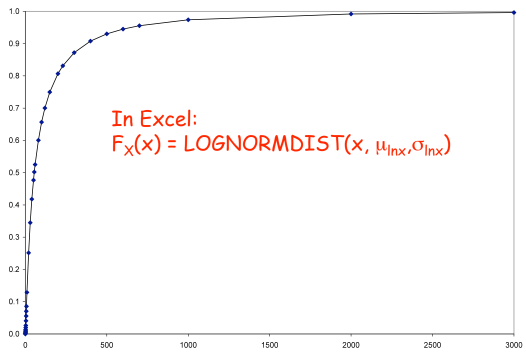
The Lognormal distribution

- Mean: $E[X] = e^{\mu_{\ln X} + \sigma^2_{\ln X}/2}$

- Standard Deviation:

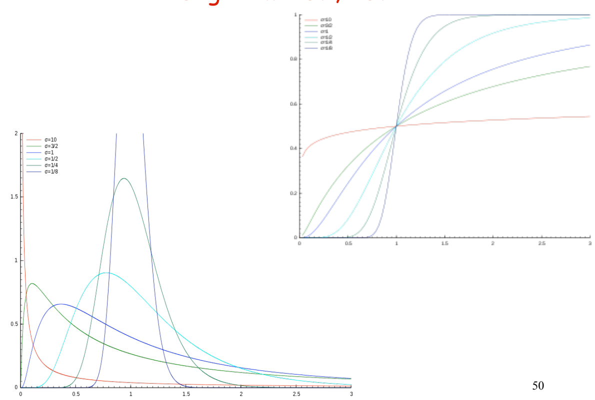
$$\sigma = \sqrt{e^{2\mu_{\ln X} + \sigma^2_{\ln X}} .(e^{\sigma^2_{\ln x}} - 1)}$$

Lognormal CDF



In Excel:
$F_X(x)$ = LOGNORMDIST(x, μ_{lnx},σ_{lnx})

Lognormal DF, CDF



The Exponential Distribution

- Widely used in queuing systems to model the inter-arrival time between requests to a system.
- If the inter-arrival times are exponentially distributed then the number of arrivals in an interval *t* has a Poisson distribution and vice-versa.

$$f_X(x) = \lambda e^{-\lambda.x}$$

- CDF

$$F_X(x) = 1 - e^{-\lambda.x} \qquad x \geq 0$$

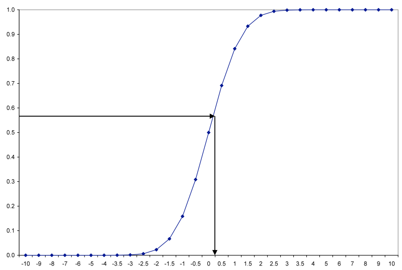
The Exponential Distribution

- Mean and Standard Deviation:

$$\mu = \sigma = 1/\lambda$$

- The COV is 1. The exponential is the only continuous r.v. with COV=1.
- The exponential distribution is "memoryless." The distribution of the residual time until the next arrival is also exponential with the same mean as the original distribution.

Generation of Random Variables



- randomly generate a number u = U(01,)
- x = F^{-1} (u) where F is the CDF

Examples of CDFs and Their Inverse Functions

Exponential	$F(x) = 1 - e^{-x/a}$	$-a\mathrm{Ln}(1-u)$
Pareto	$F(x) = 1 - x^{-a}$	$\frac{1}{(1-u)^{1/a}}$
Geometric	$F(x) = 1 - (1-p)^x$	$\left\lceil \frac{\mathrm{Ln}(u)}{\mathrm{Ln}(1-p)} \right\rceil$

Confidence Interval for the Mean

- The sample mean is an estimate of the population mean.
- Problem: given *k* samples of the population (with *k* sample means), get a single estimate of the population mean.
- Only probabilistic statements can be made:
- E.g. we want mean of the population but can get only mean of the sample
- k samples, k estimates of the mean
- Finite size samples, we cannot get the true mean
- We can get probabilistic bounds

Confidence Interval for the Mean

$$\Pr[c_1 \le \mu \le c_2] = 1 - \alpha$$

where,
(c_1, c_2): confidence interval
α : significance level
$100(1-\alpha)$: confidence level (usually 90 or 95%)
$1-\alpha$: confidence coefficient.

How to determine confidence interval ?
e.g. use 5% and 95% percentiles on sample means as bounds
Significance level e.g. 0.1

Confidence for the mean

- Issue how to estimate confidence interval ?
- E.g. take k samples, estimate k-means, sort them in increasing order take
- To estimate 90% confidence interval, use 5-percentile and 95-percentile of the sample means as confidence bounds
- Possible to estimate it from single sample
- Thanks to central limit theorem – statement about distribution of sample mean

Central Limit Theorem

- If the observations in a sample are independent and come from the same population that has mean μ and standard deviation σ then the sample mean for large sample has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

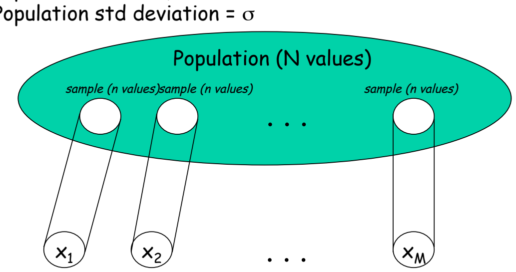
$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

- The standard deviation of the sample mean is called the *standard error.*
- Different from standard deviation
- As sample size increases the standard error goes down

58

Central Limit Theorem

Population mean = μ
Population std deviation = σ



Average of $x_1, \ldots, x_M = \mu$
Standard deviation of $x_1, \ldots, x_M = \sigma$ /sqrt(n)

59

Confidence Interval

- 100 (1-α)% confidence interval for the population mean:

$$(\bar{x} - z_{1-\alpha/2} s/\sqrt{n}, \bar{x} + z_{1-\alpha/2} s/\sqrt{n})$$

$\bar{x}$: sample mean
s: sample standard deviation
n: sample size
$z_{1-\alpha/2}$: (1-α/2)-quantile of a unit normal variate (N(0,1)).

60

Example of Confidence Interval Computation

CPU Time (msec)
5.76
2.67
3.77
2.27
2.83
1.05
2.61
1.06
5.78
3.51
2.77
1.83
1.77
1.19
2.21
24.80
1.80
1.34
1.28
1.21
2.15
1.09
1.34
32.07

n	24	
sample mean	4.51	
sample std	7.56	
alpha	0.1	
conf level	90	
1-(alpha/2)	0.95	
z0.95	1.645	from a Normal Table
c1	1.97	
c2	7.04	

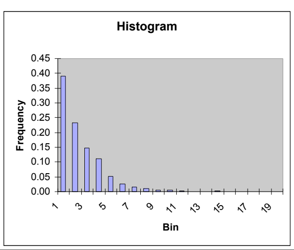
With 90% confidence the population mean is in the interval 1.97 7.04

Quantile-Quantile (Q-Q plots)

- Used to compare distributions
- E.g. compare empirical with theoretical distribution
- Plot the quantiles against each other
- "Equal shape" is equivalent to "linearly related quantile functions."
- A Q-Q plot is a plot of the type $(Q_1(p), Q_2(p))$ where $Q_1(p)$ is the quantile function of data set 1 and $Q_2(p)$ is the quantile function of data set 2.
- The values of p are $(i\text{-}0.5)/n$ where n is the size of the smaller data set.

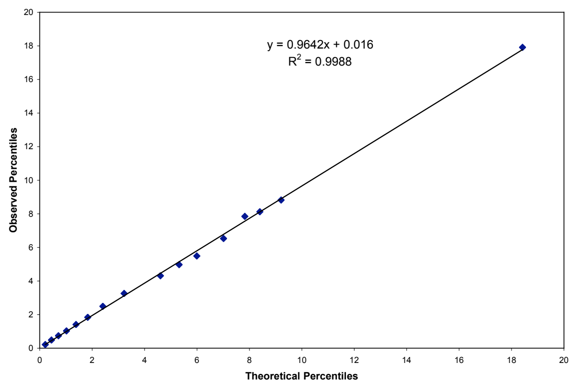
Example of a Quantile-Quantile Plot

- One thousand values are suspected of coming from an exponential distribution (see histogram)
- The quantile-quantile plot is pretty much linear, which confirms the conjecture.



Data for Quantile-Quantile Plot

qi	yi	xi
0.100	0.22	0.21
0.200	0.49	0.45
0.300	0.74	0.71
0.400	1.03	1.02
0.500	1.41	1.39
0.600	1.84	1.83
0.700	2.49	2.41
0.800	3.26	3.22
0.900	4.31	4.61
0.930	4.98	5.32
0.950	5.49	5.99
0.970	6.53	7.01
0.980	7.84	7.82
0.985	8.12	8.40
0.990	8.82	9.21
1.000	17.91	18.42



Theoretical Q-Q Plot

- Compare one empirical data set with a theoretical distribution.
- Plot (x_i, $Q_2([i\text{-}0.5]/n)$) where x_i is the $[i\text{-}0.5]/n$ quantile of a theoretical distribution ($F^{-1}([i\text{-}0.5]/n)$) and $Q_2([i\text{-}0.5]/n)$ is the i-th ordered data point.
- If the Q-Q plot is reasonably linear the data set is distributed as the theoretical distribution.

What if the Inverse of the CDF Cannot be Found?

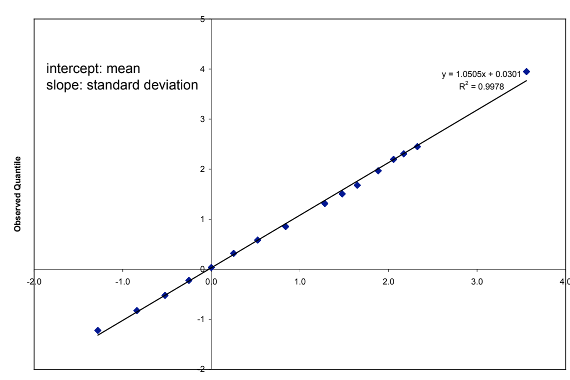
- Use approximations or use statistical tables
  - Quantile tables have been computed and published for many important distributions
- For example, approximation for N(0,1):

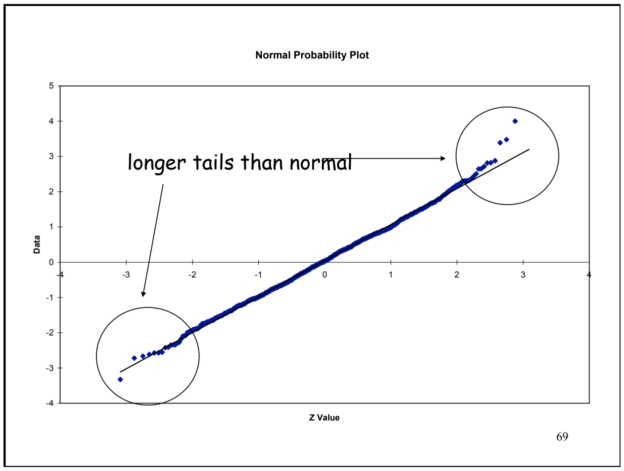
$$x_i = 4.91[q_i^{0.14} - (1-q_i)^{0.14}]$$

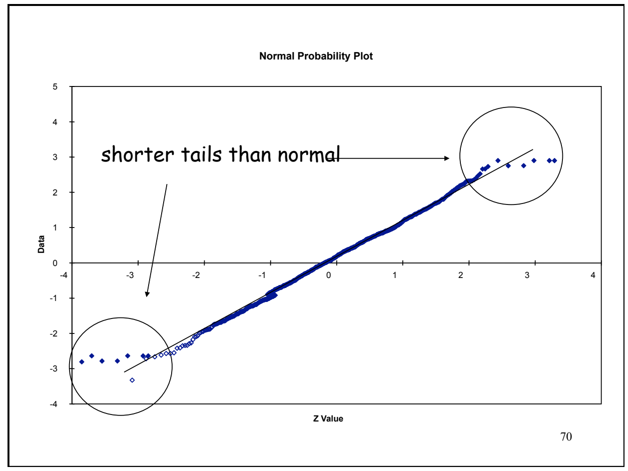
- E.g. to compute x for 95% quantile,

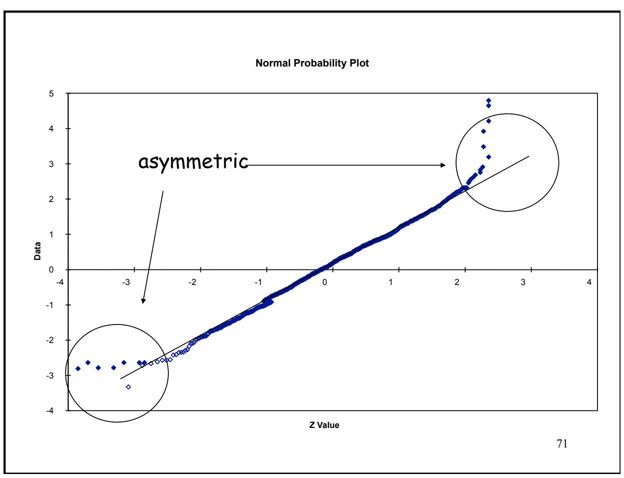
$$q_i = 0.95, x_i = 1.64$$

- For N(μ,σ) the x_i values are scaled as $\mu + \sigma x_i$ before plotting




Normal Probability Plot
longer tails than normal
Data
Z Value
69


Normal Probability Plot
shorter tails than normal
Data
Z Value
70


Normal Probability Plot
asymmetric
Data
Z Value
71