Internet Measurement and Data Analysis (8)

Kenjiro Cho

2011-11-16

review of previous class

Class 7 Measuring the diversity and complexity of the Internet

- sampling
- statistical analysis
- histogram
- exercise: histogram, CDF

today's topics

Class 8 Distributions

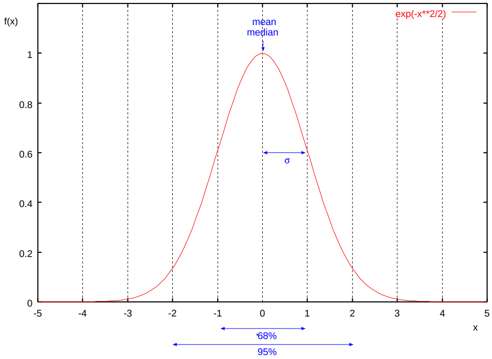
- normal distribution and other distributions
- confidence intervals
- statistical tests
- exercise: generating distributions, confidence intervals
- **assignment 2**

various distributions

- normal distribution
- exponential distribution
- power-law distribution

normal distribution (1/2)

- also known as gaussian distribution
- defined by 2 parameters: μ:mean, σ^2:variance
- sum of random variables follows normal distribution
- standard normal distribution: $\mu = 0, \sigma = 1$
- in normal distribution
  - 68% within $(mean - stddev, mean + stddev)$
  - 95% within $(mean - 2 * stddev, mean + 2 * stddev)$



normal distribution (2/2)

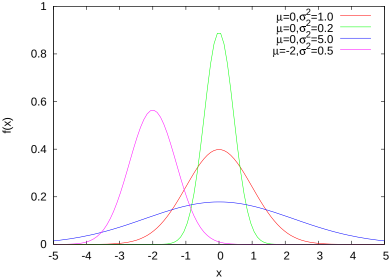
probability density function (PDF)

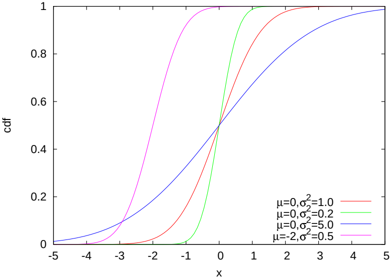
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

cumulative distribution function (CDF)

$$F(x) = \frac{1}{2}(1 + erf\,\frac{x-\mu}{\sigma\sqrt{2}})$$

$$\mu : mean, \sigma^2 : variance$$





exponential distribution

the time interval of independent events occurring at a constant average rate follows exponential distribution

- e.g., call intervals of telephone, intervals of TCP sessions

probability density function (PDF)

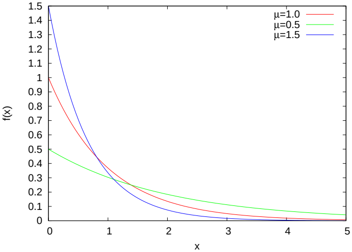
$$f(x) = \lambda e^{-\lambda x}, (x \geq 0)$$

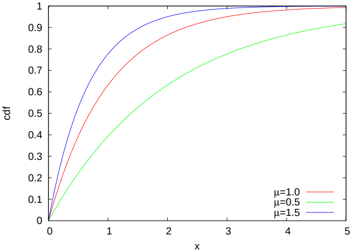
cumulative distribution function (CDF)

$$F(x) = 1 - e^{-\lambda x}$$

$$\lambda > 0 : \textit{rate parameter}$$

$$\textit{mean} : E[X] = 1/\lambda, \textit{variance} : Var[X] = \lambda^{-2}$$





power-law distribution

Zipf's law

- an empirical law found in frequency distributions of "rank data" in 1930's
- the share of n-th ranked item is roughly $1/n$ of the top share
- many observations in social science, natural science, and data communications
  - e.g., word frequency in English text, city population, wealth distribution
  - file size distribution, network traffic
- long-tail in linear scale, heavy-tail in log-log scale

pareto distribution: often used in networking research

pareto distribution

probability density function (PDF)

$$f(x) = \frac{\alpha}{\kappa}(\frac{\kappa}{x})^{\alpha+1}, (x > \kappa, \alpha > 0)$$

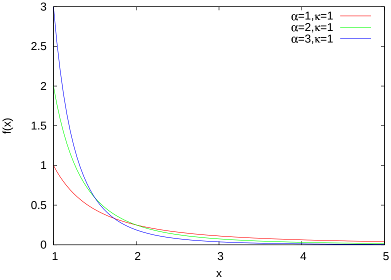
cumulative distribution function (CDF)

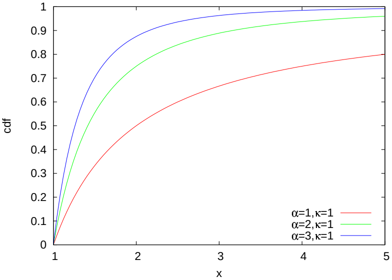
$$F(x) = 1 - (\frac{\kappa}{x})^{\alpha}$$

$$\kappa : \textit{minimum value of } x, \alpha : \textit{pareto index}$$

$$\textit{mean} : E[X] = \frac{\alpha}{\alpha - 1}\kappa, (\alpha > 1)$$

if $\alpha \leq 2$, variance $\rightarrow \infty$. if $\alpha \leq 1$, mean and variance $\rightarrow \infty$.





CCDF

Complementary Cumulative Distribution Function (CCDF)
in power-law distribution, the tail of distribution is often of interest

ccdf: probability of observing x or more

$$F(x) = 1 - P[X <= x]$$

- plot ccdf in log-log scale
  - to see the tail of the distribution or scaling property

plotting CCDF

to plot CDF

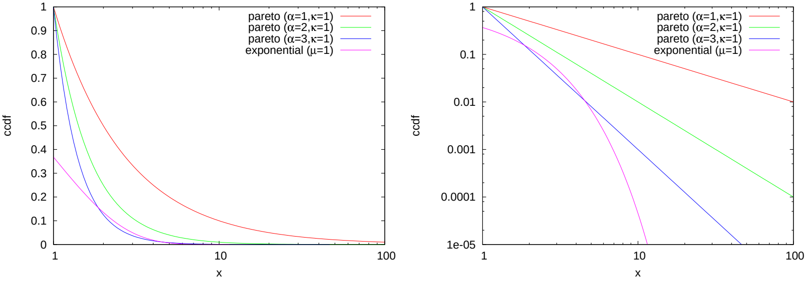
- sort $x_i, i \in \{1, \ldots, n\}$ by value
- plot $(x_i, \frac{1}{n}\sum_{k=1}^{i} k)$
- Y-axis is usually in linear scale

to plot CCDF

- sort $x_i, i \in \{1, \ldots, n\}$ by value
- plot $(x_i, 1 - \frac{1}{n}\sum_{k=1}^{i} k)$
- both X-axis and Y-axis are in log scale

CCDF of pareto distribution

- log-linear (left)
  - exponential distribution: straight line
- log-log (right)
  - pareto distribution: straight line



confidence interval

- confidence interval
  - provides probabilistic bounds
  - tells how much uncertainty in the estimate
- confidence level, significance level

$$Prob\{c_1 \le \mu \le c_2\} = 1 - \alpha$$

$$\begin{array}{ll} (c1, c2): & \textit{confidence interval} \\ 100(1-\alpha): & \textit{confidence level} \\ \alpha: & \textit{significance level} \end{array}$$

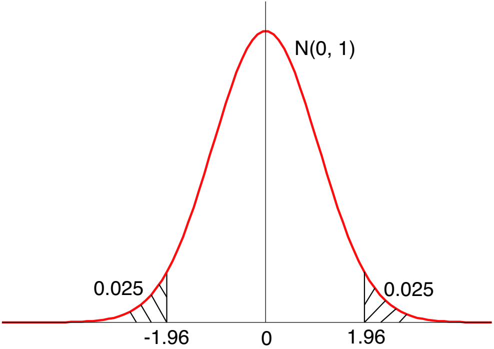
- example: with 95% confidence, the population mean is between $c1$ and $c2$
- traditionally, 95% and 99% are often used for confidence level

95% confidence interval

sample mean from normal distribution $N(\mu, \sigma)$ follows normal distribution $N(\mu, \sigma/\sqrt{n})$

95% confidence interval corresponds to the following area in the standard normal distribution

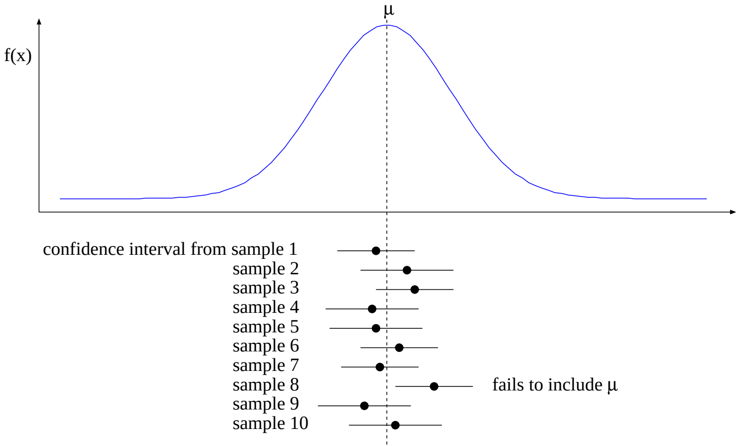
$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma\sqrt{n}} \leq 1.96$$



standard normal distribution N(0, 1)

illustration of confidence interval

- confidence level 90% means 90% samples will contain population mean in their confidence intervals



confidence interval for mean

when sample size is large, confidence interval for population mean is

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

here, $\bar{x}$:sample mean, s:sample standard deviation, n:sample size,α:significance level

$z_{1-\alpha/2}$:$(1-\alpha/2)$-quantile of unit normal variate

- for 95% confidence level: $z_{1-0.05/2} = 1.960$
- for 90% confidence level: $z_{1-0.10/2} = 1.645$
- example: 5 measurements of TCP throughput
  - 3.2, 3.4, 3.6, 3.6, 4.0Mbps
  - sample mean $\bar{x} = 3.56$Mbps, sample standard deviation $s = 0.30$Mbps
  - 95% confidence interval:

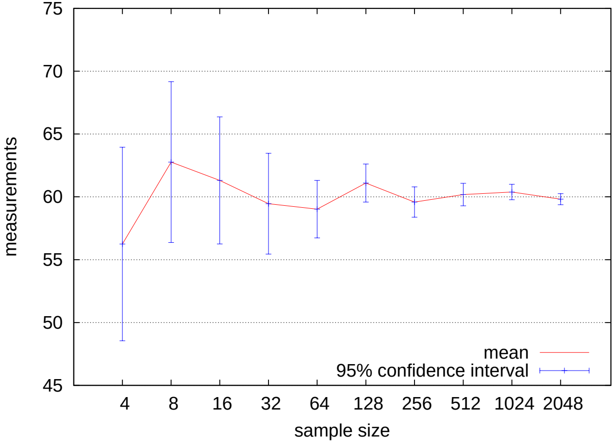
    $$\bar{x} \mp 1.96(s/\sqrt{n}) = 3.56 \mp 1.960 \times 0.30/\sqrt{5} = 3.56 \mp 0.26$$

  - 90% confidence interval:

    $$\bar{x} \mp 1.645(s/\sqrt{n}) = 3.56 \mp 1.645 \times 0.30/\sqrt{5} = 3.56 \mp 0.22$$

confidence interval for mean and sample size

confidence interval becomes smaller as sample size increases



confidence interval with varying sample size

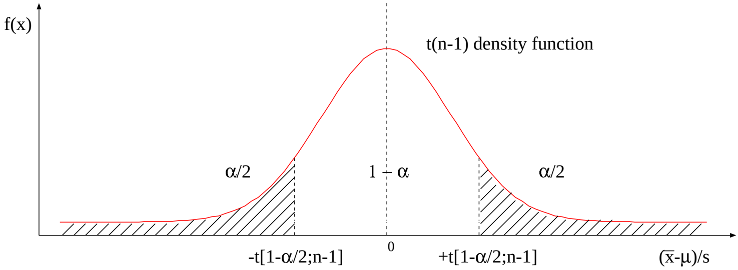
confidence interval for mean when sample size is small

when sample size is small (< 30), confidence interval can be constructed only if population has normal distribution

- $(\bar{x} - \mu)/(s/\sqrt{n})$ for samples from normal population follows $t(n-1)$ distribution

$$\bar{x} \mp t_{[1-\alpha/2;n-1]}\, s/\sqrt{n}$$

here, $t_{[1-\alpha/2;n-1]}$:$(1-\alpha/2)$-quantile of a t-variate with $(n-1)$ degree of freedom



example: confidence interval for mean when sample size is small

- example: in the previous TCP throughput measurement, confidence interval should be calculated using $t(n-1)$ distribution
  - 95% confidence interval, $n = 5$: $t_{[1-0.05/2,4]} = 2.776$

    $$\bar{x} \mp 2.776(s/\sqrt{n}) = 3.56 \mp 2.776 \times 0.30/\sqrt{5} = 3.56 \mp 0.37$$

  - 90% confidence interval, $n = 5$: $t_{[1-0.10/2,4]} = 2.132$

    $$\bar{x} \mp 2.132(s/\sqrt{n}) = 3.56 \mp 2.132 \times 0.30/\sqrt{5} = 3.56 \mp 0.29$$

other confidence intervals

- for population variance
  - chi-square distribution with degree of freedom $(n-1)$
- for ratio of sample variances
  - F distribution with degree of freedom (n_1-1, n_2-1)

how to use confidence interval

applications

- provide confidence interval to show possible range of mean
- from sample mean and stddev, compute how many trials are needed to satisfy a given confidence interval
- repeat measurement until a given confidence interval is reached

sample size for determining mean

- how many observations n is required to estimate population mean with accuracy $\pm$r% and confidence level $100(1-\alpha)\%$?
- perform preliminary test to obtain sample mean $\bar{x}$ and standard deviation s
- for sample size n, confidence interval is $\bar{x} \mp z\frac{s}{\sqrt{n}}$
- desired accuracy of $r\%$

$$\bar{x} \mp z\frac{s}{\sqrt{n}} = \bar{x}(1 \mp \frac{r}{100})$$

$$n = (\frac{100zs}{r\bar{x}})^2$$

- example: by preliminary test for TCP throughput, the sample mean is 3.56Mbps, sample standard deviation is 0.30Mbps. how many observations will be required to obtain accuracy (< 0.1Mbps) with 95% confidence?

$$n = (\frac{100zs}{r\bar{x}})^2 = (\frac{100 \times 1.960 \times 0.30}{0.1/3.56 \times 100 \times 3.56})^2 = 34.6$$

inference and hypothesis testing

the purpose of hypothesis testing

- a method to statistically test a hypothesis on population using samples

inference and hypothesis testing: both sides of the coin

- inference: predict a value to be within a range
- hypothesis testing: whether a hypothesis is accepted or rejected
  - make a hypothesis about population, compute if the hypothesis falls within the 95% confidence interval
  - accept the hypothesis if it is within the range
  - reject the hypothesis if it is outside of the range

example: hypothesis testing

when flipping N coins, we have 10 heads. In this case, can we accept a hypothesis of $N = 36$? (here, assume the distribution follows normal distribution with $\mu = N/2, \sigma = \sqrt{n}/2$)

- hypothesis: 10 heads for $N = 36$
- hypothesis testing for 95% confidence level

$$-1.96 \leq (\bar{x} - 18)/3 \leq 1.96 \quad 12.12 \leq \bar{x} \leq 23.88$$

10 is outside of the 95% confidence interval so that the hypothesis of $N = 36$ is rejected

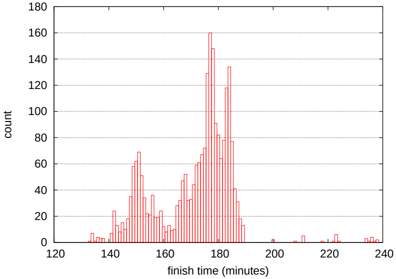
discarding outliers

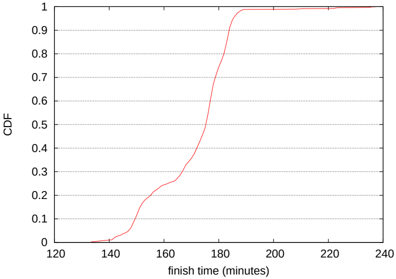
outliers should not be discarded blindly. investigation needed, which sometimes leads to new findings

- Chauvenet's criterion: heuristic method to reject outliers
  - calculate sample mean and standard deviation from sample size n
  - assuming normal distribution, determine the probability p of suspected data point
  - if $n \times p < 0.5$, the suspicious data point may be discarded
  - note: when $n < 50$, s is not reliable. the method should not apply repeatedly
- example: 10 delay measurements: 4.6, 4.8, 4.4, 3.8, 4.5, 4.7, 5.8, 4.4, 4.5, 4.3 (sec). is it ok to discard 5.8sec?
  - $\bar{x} = 4.58, s = 0.51$
  - $t_{sus} = \frac{x_{sus} - \bar{x}}{s} = \frac{5.8 - 4.58}{0.51} = 2.4$, 2.4 times larger than s
  - $P(|x - \bar{x}| > 2.4s) = 1 - P(|x - \bar{x}| < 2.4s) = 1 - 0.984 = 0.016$
  - $n \times p = 10 \times 0.016 = 0.16$
  - $0.16 < 0.5$: we may discard 5.8sec

previous exercise: histogram and CDF

- distribution of finish time of a city marathon (from Class 2)
- plot a CDF this time





previous exercise: histogram and CDF (cont'd)

- distribution of finish time of a city marathon (from Class 2)
- plot a CDF this time

original:

```
# Minutes Count
133 1
134 7
135 1
136 4
137 3
138 3
141 7
142 24
...
```

add cumulative count:

```
# Minutes Count CumulativeCount
133 1 1
134 7 8
135 1 9
136 4 13
137 3 16
138 3 19
141 7 26
142 24 50
```

exercise: generating normally distributed random numbers

- using a uniform random number generator function (e.g., rand in ruby), create a program to produce normally distributed random numbers with mean u and standard deviation s.

box-muller transform

basic form: creates 2 normally distributed random variables, z_0 and z_1, from 2 uniformly distributed random variables, u_0 and u_1, in $(0, 1]$

$$z_0 = R\cos(\theta) = \sqrt{-2\ln u_0}\cos(2\pi u_1)$$

$$z_1 = R\sin(\theta) = \sqrt{-2\ln u_0}\sin(2\pi u_1)$$

polar form: approximation without trigonometric functions
u_0 and u_1: uniformly distributed random variables in $[-1, 1]$,
$s = u_0^2 + u_1^2$ (if $s = 0$ or $s \geq 1$, re-select u_0, u_1)

$$z_0 = u_0\sqrt{\frac{-2\ln s}{s}}$$

$$z_1 = u_1\sqrt{\frac{-2\ln s}{s}}$$

random number generator code by box-muller transform

```
# usage: box-muller.rb [n [m [s]]]
n = 1 # number of samples to output
mean = 0.0
stddev = 1.0

n = ARGV[0].to_i if ARGV.length >= 1
mean = ARGV[1].to_i if ARGV.length >= 2
stddev = ARGV[2].to_i if ARGV.length >= 3

# function box_muller implements the polar form of the box muller method,
# and returns 2 pseudo random numbers from standard normal distribution
def box_muller
  begin
    u1 = 2.0 * rand - 1.0  # uniformly distributed random numbers
    u2 = 2.0 * rand - 1.0  # ditto
    s = u1*u1 + u2*u2    # variance
  end while s == 0.0 || s >= 1.0
  w = Math.sqrt(-2.0 * Math.log(s) / s) # weight
  g1 = u1 * w  # normally distributed random number
  g2 = u2 * w  # ditto
  return g1, g2
end
# box_muller returns 2 random numbers.  so, use them for odd/even rounds
x = x2 = nil
n.times do
  if x2 == nil
    x, x2 = box_muller
  else
    x = x2
    x2 = nil
  end
  x = mean + x * stddev  # scale with mean and stddev
  printf "%.6f\n", x
end
```

assignment 2: normal distribution, histogram and confidence interval

- the purpose is to understand normal distribution and confidence interval
- assignment
  1. generate 10 sets of normally distributed numbers with varying sample size.
  2. create 2 histogram plots for sample size 128 and 1024
  3. compute confidence interval of mean for the 10 sets, and make a plot
- items to submit
  1. 2 histogram plots
  2. a plot of interval estimation for the 10 sample sets
- submission format: a single PDF file including 3 plots (2 histogram plots and 1 interval estimation plot)
- submission method: upload the PDF file through SFC-SFS
- submission due: 2011-12-03

assignment details

1. generate 10 sets of normally distributed numbers with varying sample size.
   - use the box-muller code in today's exercise
   - use your height in cm for mean, and half of your foot size in cm for standard deviation
   - with varying sample size $n = \{4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$.
2. create 2 histogram plots for sample size 128 and 1024
   - confirm that the generated random numbers follow normal distribution
   - select appropriate bin size for histograms using commonly used boundaries for heights (e.g., 1cm, 2cm, 5cm, etc)
3. compute confidence interval of mean for the 10 sets, and make a plot
   - confirm that confidence interval changes according to sample size.
   - for each of the 10 sample sets, compute the confidence interval of mean. Use confidence level 95%, confidence interval $\mp 1.960 \frac{s}{\sqrt{n}}$.
   - plot the results of the 10 sets in a single graph; X-axis for sample size n in log-scale, Y-axis for mean and confidence interval in linear scale. (the plot should look similar to slide 17).

summary

Class 8 Distributions

- normal distribution and other distributions
- confidence intervals
- statistical tests
- exercise: generating distributions, confidence intervals
- **assignment 2**

next class

Class 9 Measuring traffic of the Internet (11/18 9:25-10:55 e11)

- traffic measurement
- exercise: traffic measurement

Class 10 Hot topics (11/18 11:10-12:40 e11)