Module 2: Introduction to Statistics

Niko Kaciroti, Ph.D.
BIOINF 525 Module 2: W16
University of Michigan

Course Info

- Instructor
  - Niko Kaciroti: nicola@umich.edu
    - Office: 300 N. Ingalls Bldg, 10-floor, #1027NW
    - Office hours: Th. 11:00pm to 12:00pm or by appointment
- Teaching Assistant
  - Hongyang Li: hyangl@umich.edu
    - Office hours: Wed. 1:00pm-2:30pm
- Data from TROPHY Study will be used for the class
  - Feasibility of treating prehypertension with an angiotensin-receptor blocker. *New England Journal of Medicine.* 2006; 354:1685-97.
- Grading
  - Homework
    - Students may form study groups to discuss class notes and homework assignments but must hand in their own work

Topic

- Probability distributions
- Quantifying central values and variability in the data
  - Mean, SD, Quartiles, Inter Quartile Range (IQR)
- Graphical display of data
  - Histograms, Boxplot
- Normal distribution
  - Q-Q Plots for Normality
  - Shapiro Wilks Test for Normality
  - Features of normal distribution/Center limit theorem (CLT)
- Other commonly used distribution
  - T-distribution, χ^2 distribution, $\mathcal{F}$ distribution

Statistical Programs

- Some commonly used statistical programs:
  - SAS
  - SPSS
  - Stata
  - R/Rstudio
- R is an open source program, and will be used in the lab.
  It is available for downloaded at: http://cran.r-project.org/
  Rstudio at: http://www.rstudio.com/ide/download/
- Reference Book : ***The R Book*** 2nd edition by Michael J. Crawley

Variable Types

- Discrete variables (or categorical data)
  - May take values as an integer or belong to a set number of categories
    - ✓ # Number of sunny days in January (0,1,2,…,31)
    - ✓ Gender (0="Female", 1="Male")
    - ✓ Race (1="White", 2="Black", 3="Asian", 4="Others")
- Continuous variables
  - May take any real value within a defined range
    - ✓ Height/Weight (6.11ft,/191.12lbs)
    - ✓ Blood Pressure (100.12mmHg)

Randomness and Probability

- The concepts of randomness and probability are central to the field of statistics
- Most experiments are not perfectly reproducible: some experiments are more accurate, some are less
- We will outline the basic ideas of *probability distributions*, which are used to measure the degree of uncertainty and reproducibility

Probability Distributions

- Probability distribution describes how data points are distributed. That is, what is the likelihood (or relative frequency) that a certain value occurs

- It is a mathematical function that *assigns* some probability to each of the possible outcome values of a random variable (X)†
  - Probability mass function (**pmf**): Is used for discrete variables
  - Probability density function (**pdf**): Is used for continuous variables

†X is a random variable if it can take different values, each with some probability.
  - E.g., X indicates the outcome of tossing a coin (Tail/Head).

PMF for Discrete Variables

- **Pmf** is a function that gives the probability that X equals to some value k, *Pr(X=k)*.
  - E.g. Tossing a coin, the pmf is:

    Pr(X=Tail)=p=0.5

    Pr(X=Head)=1-p=0.5

  - Rolling a dice, the pmf is:

    $Pr(X=k)=p_k=1/6$ for k=1,2,…,6.

  - Sum of all probabilities must be 1:

$$\sum_{k=1}^{n} Pr(X = k) = 1$$

PDF for Continuous Variables

- The probability density function *f(x)* is a bit more complicated when x is continuous. It is defined as the limit when $\delta x \rightarrow 0$ of :

$$Pr(x \leq X < x+\delta x) / \delta x = f(x)$$

Here $Pr(x \leq X < x+\delta x)$ is the probability that X lies between *x* and $x+\delta x$ where δx is small

- Another useful function describing the distribution is the Cumulative Density Function (CDF), which gives the probability that X is less than or equal to *x*:

$$F(x)=P(X \leq x)=\int_{-\infty}^{x} f(x)dx$$

Why It's Important to Learn About Different Distributions?

- It is necessary to decide which are the appropriate:
  - Descriptive measures (to best analyze/present the data)
  - Statistical test(s) when testing hypothesis
- Helps in deciding whether to transform your data
  - Does the data have characteristics that might make it inconsistent with the assumptions about the distributions used in a statistical test?

Commonly Used Distributions

(Continuous Variables)

- ***Normal*** distribution with mean μ and variance σ^2
- ***t*** distribution with *k* degrees of freedom (aka Student's t distribution)
- **Chi-Square** (χ^2) distribution with *k* degrees of freedom
- $\mathcal{F}$ distribution with *(n, m)* degrees of freedom
  - **degrees of freedom (*df*)** is the number of values in a statistics that are free to vary (not constrained).
  - *df*=n-#of constrains

Normal, *t*, χ^2, and $\mathcal{F}$ distributions are widely used for statistical testing (more on these later). These tests are derived based on normal distribution assumption, which makes the normal distribution vital to statistics.

Commonly Used Distributions

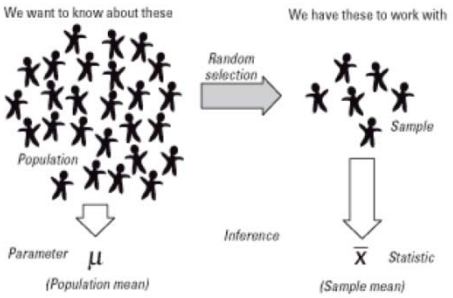
(Discrete Variables)

- Bernoulli distribution (Yes/No)
  - Tossing a coin, does it land "Tail": 1="Yes"/0="No")
- Binomial distribution (Number of successes in n trials)
  - Number of "tails" when tossing a coin n times (e.g. 5 times out of 10)
- Multinomial distribution (belonging to some categories)
  - Race 1=W, 2=B, 3=A, 4=Other
- Poisson distribution (count data: 0,1,2,….)
  - # of goals is a soccer game (0,1,2,….)
  - # of points in a football game???

Topic

- Probability distributions
- **Quantifying central values and variability in the data**
  - **Mean, SD, Quartiles, Inter Quartile Range (IQR)**
- Graphical display of data
  - Histograms, Boxplot
- Normal distribution
  - Q-Q Plots for Normality
  - Shapiro Wilks Test for Normality
  - Features of normal distribution/Center limit theorem (CLT)
- Other commonly used distribution
  - T-distribution, χ^2 distribution, $\mathcal{F}$ distribution

Sampling: What is the Difference Between a Sample and a Population?



- **Population** (or target population): It consists of all people or things that you want to describe (*N*)
  - *E.g.:* All males in graduate school at U of M in the academic year 2015-16
- **Sample**: Representative subset of the population. Taking a sample (at random) of *n* out of *N* individuals to estimate some characteristics (height) for the population
  - Sample: 50 males in graduate school at U of M in 2015 - 16

Sampling Example: Two samples of n=10 students

	Sample 1 (height)	Sample 2 (height)
	5.2	5.5
	5.4	5.5
	5.5	5.6
	5.6	5.6
	5.7	**5.7**
	5.8	5.8
	5.9	5.8
	6.0	5.9
	6.1	5.9
mean height	5.7	5.7

What is different between sample 1 and sample 2?

- Sample 2 is more homogenous. All heights are within 2 inches from the mean.

Quantifying Central Values and Variability in the Data

Mean, Variance, Standard Deviation

Let $x_1, x_2, \ldots, x_n$, be a random sample from a population.

- **Mean,** is defined as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Variance,** is a measure of variability around the mean defined as the "average" of the squared deviations from the mean:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \quad or \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- **Standard deviation (sd),** is square root of variance. It is measured in the same unit as the mean:

$$sd = \sqrt{s^2} = s$$

- **Standard error (se),** often is referred to as a measure of precision and is defined as:

$$se(\bar{x}) = \frac{s}{\sqrt{n}}$$

Quintiles, Percentiles, Median, Quartiles, IQR

- ***q*-Quintiles:** Are points that divide data into *q* equal subsets
- **Percentiles (*100-quintile*):** It is the value below which a certain percent of observations fall. *E.g.*, the 20th percentile is the value below which 20% of the observations are found
- **Median**: It is a value that splits the data into two equal parts (Median= 50 percentile, 2-quintile)
- **Quartiles**: Are values that divide the data into four equal parts Q1=25th percentile, Q2=50th percentile (median),Q3=75th percentile
- **Inter Quintile Range (IQR)**: IQR=Q3-Q1

Calculating Percentiles

Data (n=11)	Sorted	
11	*11*	
35	*12*	
31	*22*	← Q1
24	*24*	
12	*30*	
57	*31*	← Median/Q2
32	*32*	
30	*35*	
22	*44*	← Q3
77	*57*	
44	*77*	

Mean=32.3

SD=18

The p^{th} percentile of n values is the $p*n^{th}$ value in the sorted data (round to the nearest integer in the list 1,2,…,n of order)

Median: 0.5*11=5.5 (the 6th)→ 31

Q1: 0.25*11=2.75 (the 3rd) → 22

Q3: 0.75*11=8.25 (the 8th) → 44

IQR = Q3-Q1: 44-22=22

Calculating Percentiles

(Not sensitive to outliers)

Data (n=11)	Sorted	
11	*11*	
35	*12*	
31	*22*	← Q1
24	*24*	
12	*30*	
57	*31*	← Median/Q2
32	*32*	
30	*35*	
22	*44*	← Q3
77	*57*	
44	*777*	
Mean=32.3	(96.9)	
SD=18	(215.4)	

Would the mean or median change if there was a large outlier?

Median: 0.5*11=5.5 (the 6th)→ 31

Q1: 0.25*11=2.75 (the 3rd) → 22

Q3: 0.75*11=8.25 (the 8th) → 44

IQR = Q3-Q1: 44-22=22

Topic

- Probability distributions
- Quantifying central values and variability in the data
  - Mean, SD, Quartiles, Inter Quartile Range (IQR)
- **Graphical display of data**
  - **Histograms, Boxplot**
- Normal distribution
  - Q-Q Plots for Normality
  - Shapiro Wilks Test for Normality
  - Features of normal distribution/Center limit theorem (CLT)
- Other commonly used distribution
  - T-distribution, χ^2 distribution, $\mathcal{F}$ distribution

Graphical Display of Data

- A visual or graphical display of data is a useful tool for understanding and summarizing the data. It should always be the first step on statistical analysis.
  - It is very useful for data quality control, checking for errors, unusual values or outliers
  - It is also useful for understanding the distributions of the data, which will help in choosing the appropriate statistical models
- Two important graphical display of data are:
  - Histogram
  - Boxplot

Histograms

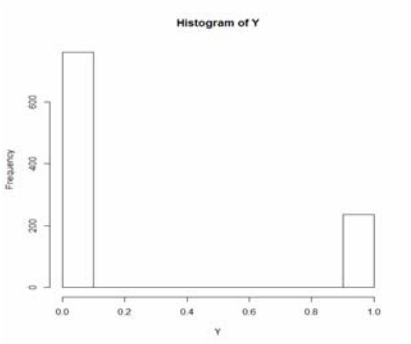
- A **histogram** is a graphical representation of the probability distribution for a given variable. It displays the frequencies of observations occurring in certain ranges of values
  - For discrete measures it shows the frequency of values in each category
  - For continuous measure it shows the frequency of values occurring in small intervals covering the whole range

Example: Histogram for Discrete Measures

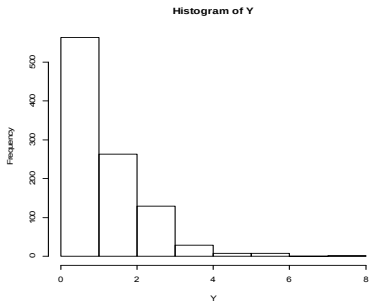
Simulated Data

- In R

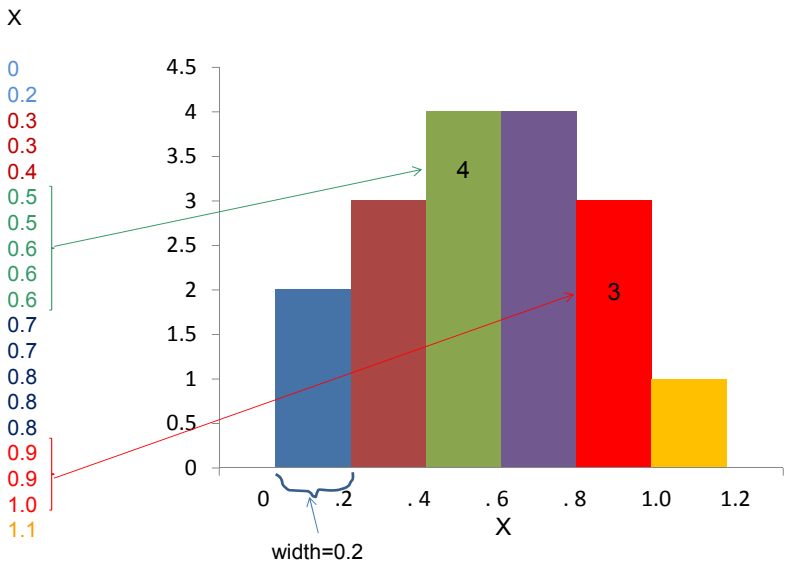
Y<-rbinom(1000,1,.25)
hist(Y)



Y<-rpois(1000,1.5)
hist(Y)

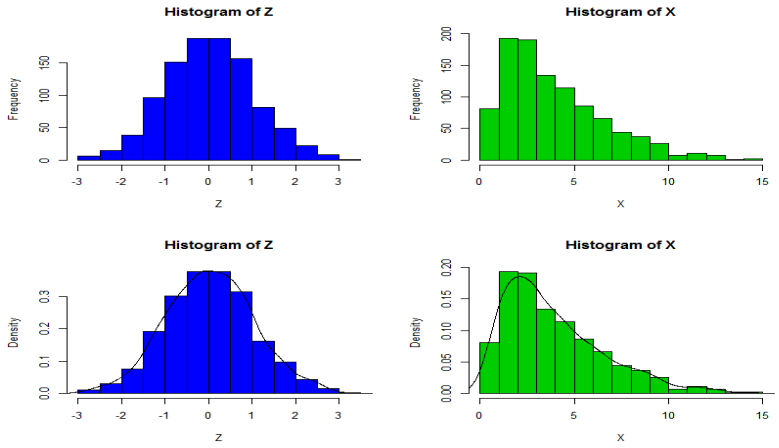


Histogram for continuous Measures



Example: Histogram of continuous Measure

Simulated Data



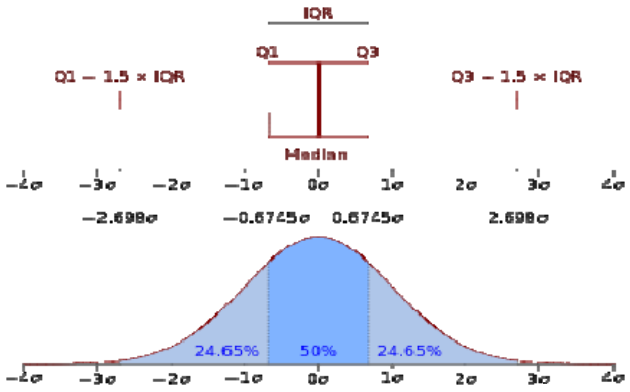
What are some differences between blue and green histograms?

Boxplots

- What is a boxplot
  - A **box plot** or **boxplot** (**box-and-whisker plot**) is a graphical tool for conveying information on location, variation, and symmetry of the data.
  - It is also used for detecting and illustrating location and variation differences between two or more groups of data **(side-by-side boxplot)**
- Why it is useful?

  It is an efficient visual tool to summarize the main characteristics of the data: Mean/median, quintiles, spread, symmetry and outliers.

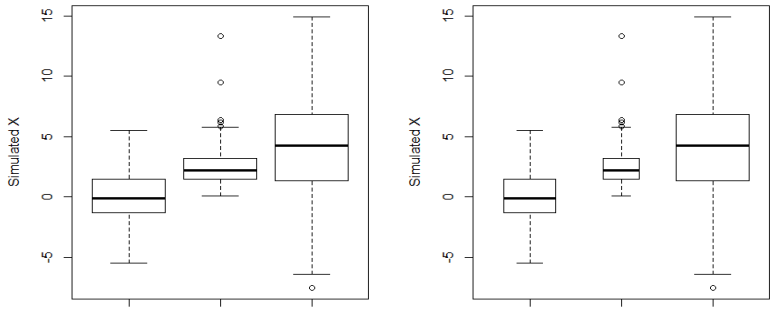
Boxplot



*If the data are not normal the length of the whisker is up to the min/max value

Side-by-Side Boxplot Examples Using R

(Simulated Data)



- In the second panel the larger the n the wider the box
- Which one has largest median? Smallest IQR?

Topic

- Probability distributions
- Quantifying central values and variability in the data
  - Mean, SD, Quartiles, Inter Quartile Range (IQR)
- Graphical display of data
  - Histograms, Boxplot
- **Normal distribution**
  - **Q-Q Plots for Normality**
  - **Shapiro Wilks Test for Normality**
  - **Features of normal distribution/Center limit theorem (CLT)**
- Other commonly used distribution
  - T-distribution, χ^2 distribution, $\mathcal{F}$ distribution

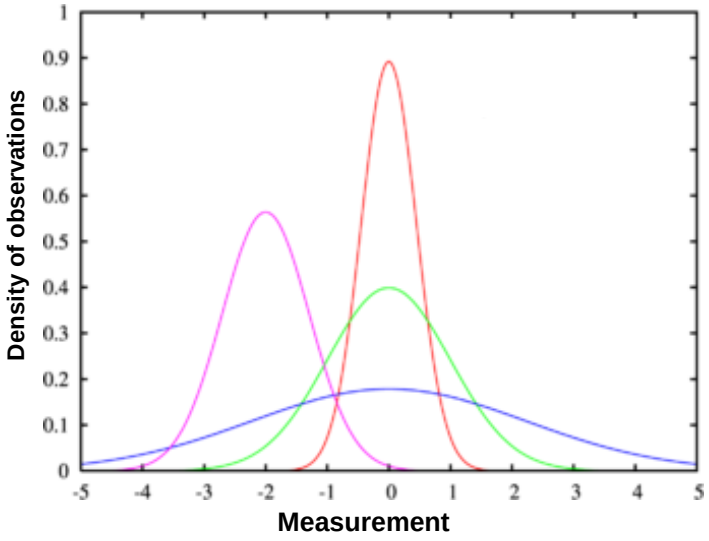
Normal Distribution

- Normal distribution is the most used distribution (and a building block) in Statistics. It is used to describe and summarize real life data, and also to perform statistical testing
- If *X* is normally distributed we write:

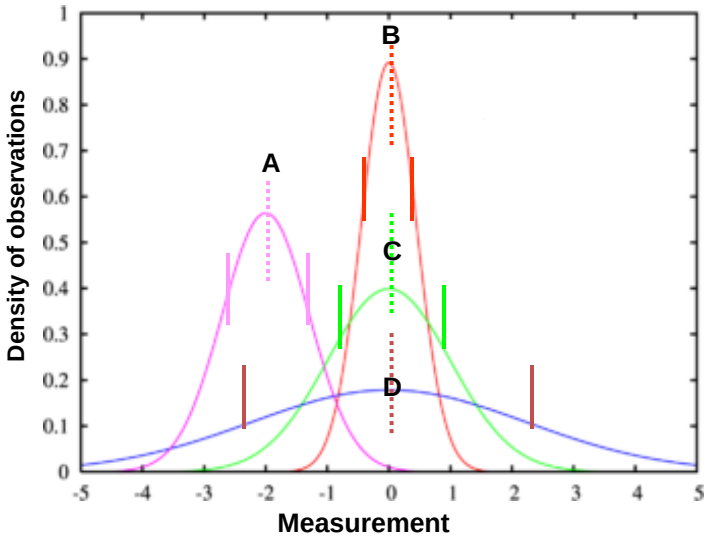
$$X \sim N(\mu, \sigma^2)$$

  - "~" stands for *'is distributed as'*
  - μ is the mean parameter: $\mu = \int x f(x) dx$
  - σ^2 is the variance parameter: $\sigma^2 = \int (x - \mu)^2 f(x) dx$
  - σ is standard deviation (SD)
- Pdf of *X* is: $f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

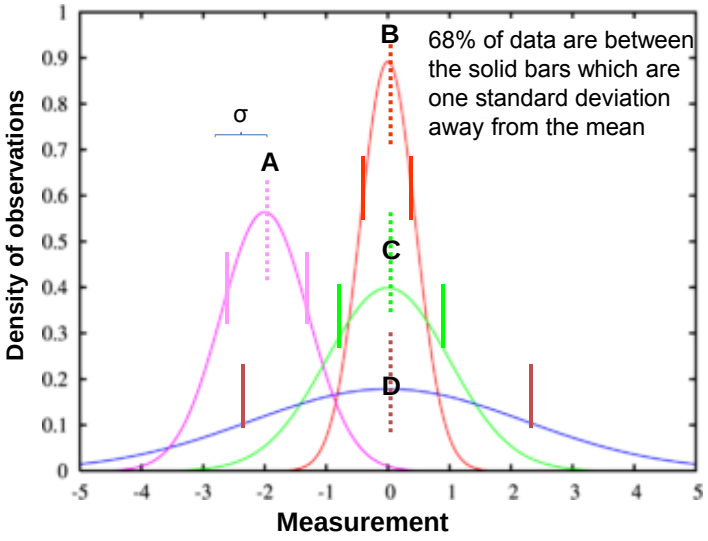
PDF of Normal Distributions



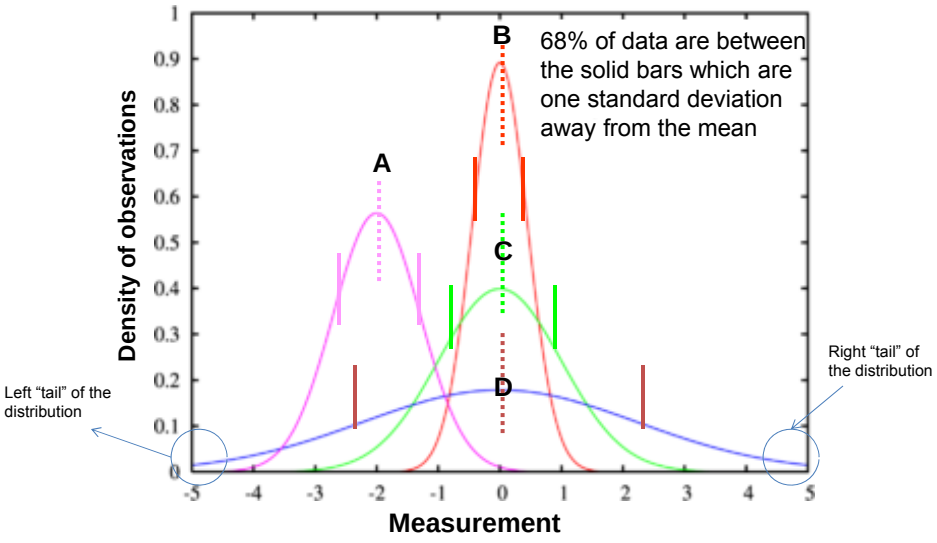
Does the Mean (μ) Alone Define the Distribution?



Standard Deviation (σ)



Tail of the Distribution



Characteristics of Normal Distribution

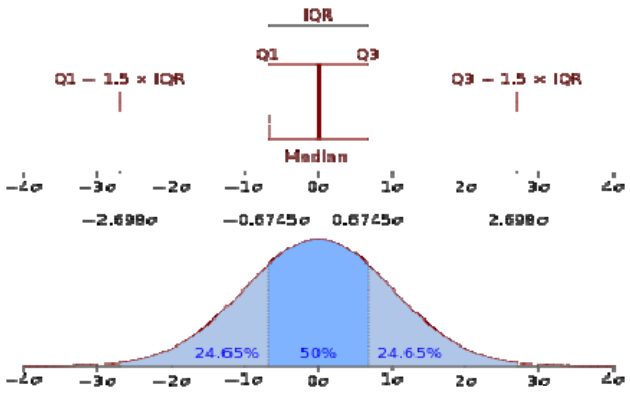
- It is a symmetrical bell-shaped curve
  - Most frequently occurring points around μ (mean/median)
  - Points of inflection at μ - σ (standard deviation) and μ + σ
  - 68/95/99 Rule
    - ✓ 68% of the data are within μ - σ and μ + σ

      $P(\mu - \sigma < X < \mu + \sigma) \approx .68$
    - ✓ 95% of the data are within μ - 2σ and μ + 2σ

      $P(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = .95$
    - ✓ 99% of the data are within μ - 3σ and μ + 3σ

      $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx .99$
  - Symmetric to the mean
    - ✓ Skewness=0 (Left and right "tail" of the distribution are the same)

Outliers for Normal Distribution

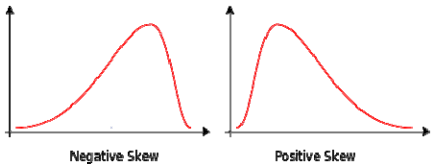


Mild outliers: < Q1 - 1.5*IQR but > Q1 - 3*IQR
> Q3 + 1.5*IQR but < Q3 + 3*IQR

Extreme outliers <Q1 - 3*IQR or >Q3 + 3*IQR

Skewness Coefficient

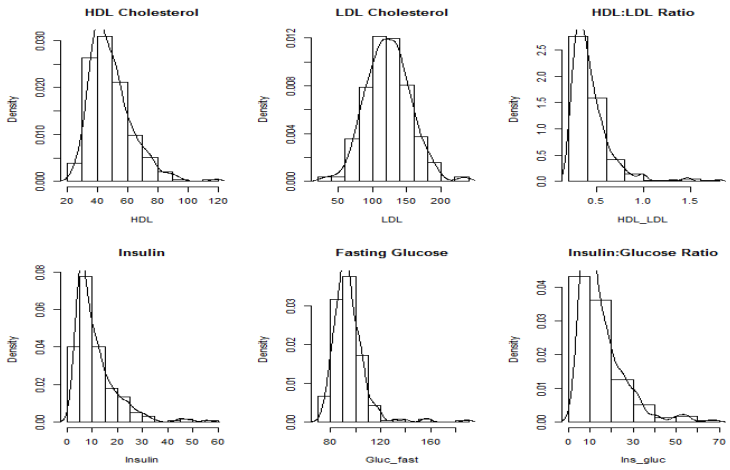
- Skewness is a measure of symmetry. A simple measure of Skewness is: Skewness=(mean-median)/sd



Why it is useful: If skewness is not 0, it indicates that the data are not normally distributed and a transformation may be needed.

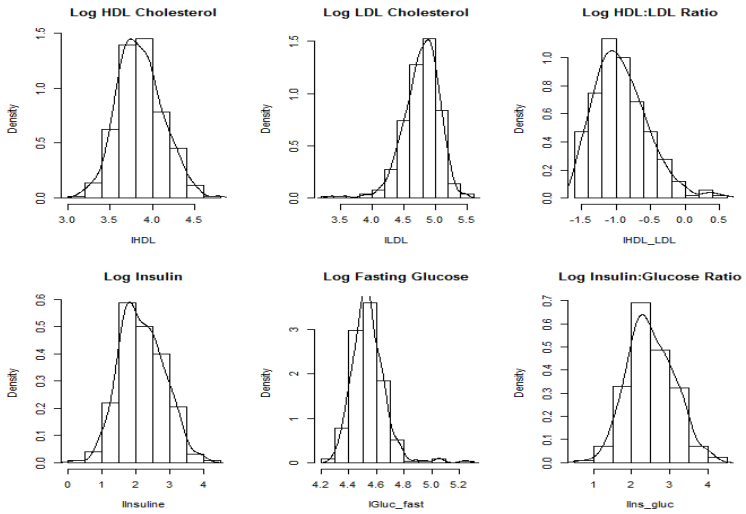
Physiological Distributions

(Data from TROPHY Study*)



*Feasibility of treating prehypertension with an angiotensin-receptor blocker. *New England Journal of Medicine*. 2006; 354:1685-97.

Log-transform Physiological Distributions



Which of These Samples Might Have a Normal Distribution?

A. Height of 50 males in graduate school at U of M in February 2013
   ✓ Yes.

B. Household income of 10,000 randomly chosen people in the US
   x No. Skeweness > 0 due to few high earners can stretch the tail to the right.

C. Shoes sizes for shoes from a factory that produces equal numbers of shoes in sizes 6 to 10
   x No.

Testing for Normality

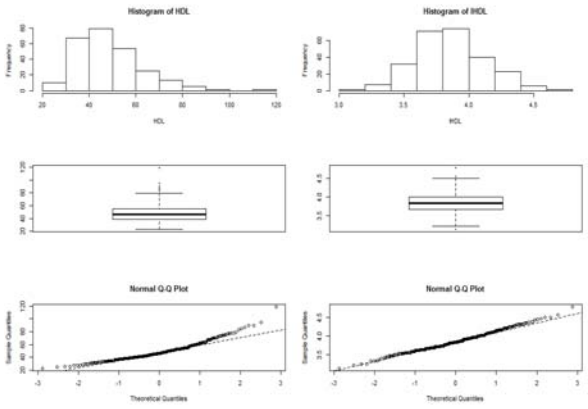
- Plot to "test" for normality
  - Histograms: Bell shape, symmetry
  - Boxplot: Symmetry, outliers
  - Q-Q plot (Quantile-Quantile Plot)

    The simplest visual test of normality is the 'q-q plot'. This plots the quantiles of the observed data against the quantiles from a normal distribution.

    If the data are normally distributed then the q-q plot will follow a straight line. Departures from normality show up as various sorts of non-linearity (e.g. S-shapes or banana shapes).
- Shapiro-Wilks test
  - Formal statistical test for normality

Plots to Test for Normality

(TROPHY data)



Shapiro Wilks for Normality

- Shapiro Wilks test is used to formally test for normality

  In R: shapiro.test(HDL)

  shapiro.test(lHDL)

Shapiro-Wilk normality test

data: HDL

W = 0.9305, **p-value = 1.394e-09**

data: lHDL

W = 0.9926, **p-value = 0.2322**

- Conclusion:

  If p-value < .05, then reject the hypothesis that data are normal

Shapiro Wilks for Normality

- Shapiro Wilks test is used to formally test for normality

  In R: shapiro.test(HDL)

  shapiro.test(lHDL)

Shapiro-Wilk normality test

data: HDL

W = 0.9305, **p-value = 1.394e-09** → HDL is not normal

data: lHDL

W = 0.9926, **p-value = 0.2322** → No evidence that Log-HDL is not normal

- Conclusion:

  If p-value < .05, then reject the hypothesis that data are normal

Linear Transformation of Normally Distributed Random Variables are Normally Distributed

1. **If X, *is a* normal random variables, with means μ, and standard deviations σ, then the linear transformed variable will also be normally distributed**

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

2. **If X_1, X_2 are two independent normal random variables, with means μ_1, μ_2 and standard deviations σ_1, σ_2, then their linear combination will also be normally distributed**

$$a_1X_1 + a_2X_2 \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2)$$

Central Limit Theorem (CLT)

- If, $x_1, x_2, \ldots, x_n$ is a sequence of independent identically distributed random variables, each having mean μ and variance σ^2, then the CLT states that as the size (*n)* of the sample gets large, the distribution of $\bar{x}$, (or $\sum_i x_i$) becomes normally distributed with E(Y)= μ, Var(Y)=$\frac{\sigma^2}{n}$.

$$\bar{x} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n}) \qquad or \qquad \sum_i x_i \xrightarrow{d} N(n\mu, n\sigma^2)$$

- Why is CLT important: If *n* is fairly large (*n > 30*), then following CLT one can implement statistical tests related to the $\bar{x}$ (or $\sum_i x_i$) that are based on normal distributed theory even if x_i are not normally distributed.

Illustration of Center Limit Theorem Using R

(Simulated data in R)

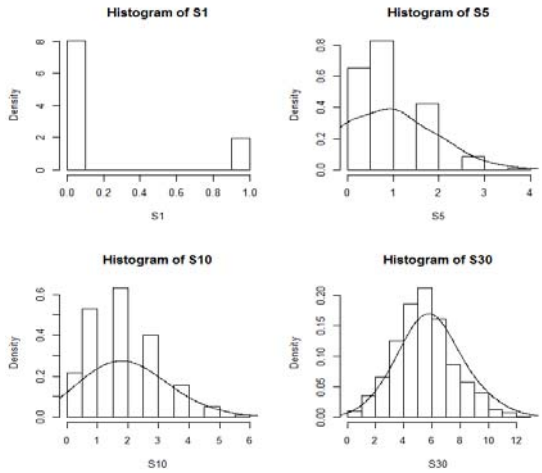
X_i ~ *Bernoulli(.2)*

$S1=X_1$

$S5=X_1+X_2+\ldots+X_5$

$S10=X_1+X_2+\ldots+X_{10}$

$S30=X_1+X_2+\ldots+X_{30}$



Application of CLT: 95% Confidence Intervals

- From CLT: $\bar{x} \xrightarrow{d} N(\mu, \frac{\sigma^2}{n})$ then (from 68/95/99 rule)

  $Pr[\mu - 1.96*\sigma/\sqrt{n} < \bar{x} < \mu + 1.96*\sigma/\sqrt{n}] = 0.95$

  or

  $Pr[\bar{x} - 1.96*se(\bar{x}) < \mu < \bar{x} + 1.96*se(\bar{x})] = 0.95$

  where $se(\bar{x})$ is an estimate for $\sigma/\sqrt{n}$.

- $[\bar{x} - 1.96*se(\bar{x}), \bar{x} + 1.96*se(\bar{x})]$ is referred as the 95%Confidebce Interval (95%CI) for μ

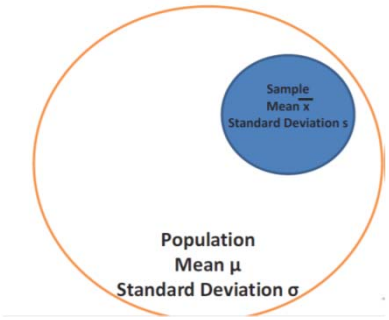
- More general: If $\hat{\beta}$ is an estimate for β then:

  $[\hat{\beta} - 1.96*se(\hat{\beta}), \hat{\beta} + 1.96*se(\hat{\beta})]$ is the 95%CI for β

- Why are Confidence Intervals (CI) important?

CI are Important When Making Inferences About a Population Based on a Sample

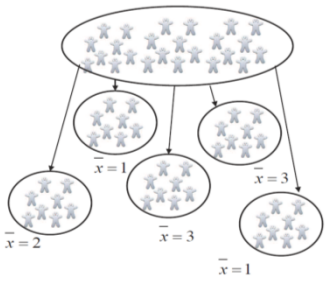
- **Statistical Inference:** Drawing conclusion based on data. Estimating the characteristics or properties of a population derived from the analysis of a sample drawn from it.
  - What do $\bar{x}$ and s tell us about μ and σ?



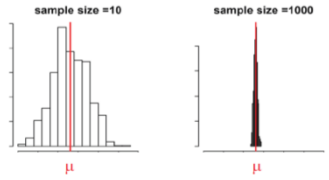
How Close are Sample Values to Population Values?

Depends on Sampling Variability

The observed sample mean will vary from sample to sample



The distribution of the sample mean is centered on the population mean



The spread of the distribution of the sample mean depends on the SD of the population and on the sample size:

$$SE = \frac{SD}{\sqrt{n}}$$

16

- Confidence intervals are used to indicate the reliability of an estimate

Topic

- Probability distributions
- Quantifying central values and variability in the data
  - Mean, SD, Quartiles, Inter Quartile Range (IQR)
- Graphical display of data
  - Histograms, Boxplot
- Normal distribution
  - Q-Q Plots for Normality
  - Shapiro Wilks Test for Normality
  - Features of normal distribution/Center limit theorem (CLT)
- **Other commonly used distribution**
  - **T-distribution, χ^2 distribution, $\mathcal{F}$ distribution**

T-distribution

- Let, $x_1, x_2, \ldots, x_n$, be independently normally distributed with mean 0 and sampling variance s. Then the t-distribution is derive as:

$$t_{n-1} = \frac{\bar{x}}{s/\sqrt{n}} = \frac{\bar{x}}{se(\bar{x})}$$

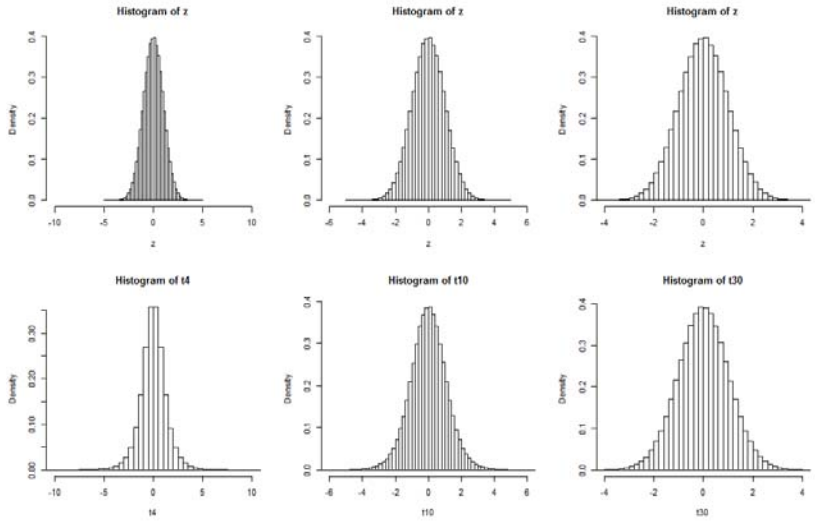
- The t distribution is very similar to the normal distribution except that, for small samples it has longer tails. As the sample size gets large, the t distribution approaches the normal distribution

- In many applications we will modify/transform the data to implement the t-distribution. In other words, we will convert something unknown into something whose properties are well established (t-distribution). E.g. if $x_1, x_2, \ldots, x_n$ has mean μ then

$$t_{n-1} = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Illustration of T-distribution

(Simulated Data in R)



Statistical Tests using T-distribution

- T-distribution and t-test (or Student's t-test)
  - One sample t-test,
  - Two sample t-test
  - Paired t-test

χ^2 distribution

- If, $Z_1, Z_2, \ldots, Z_k$ have a standard normal distribution (μ =0, σ =1) then

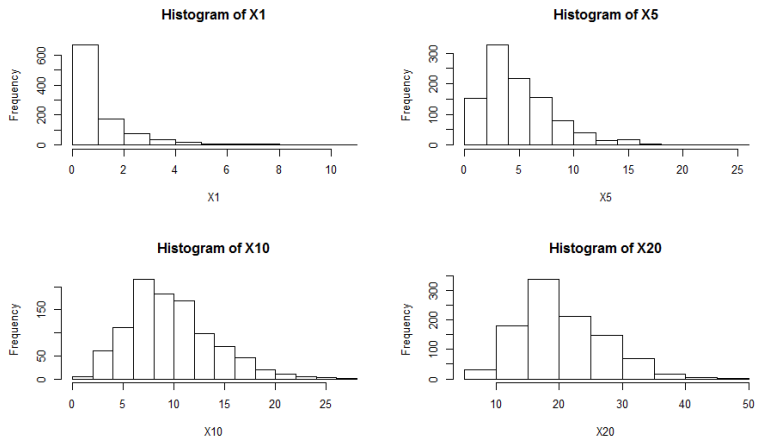
$$X = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$$

has a Chi-Square distribution with *k* degrees of freedom.

- The mean is: E(X)=k
- The variance is: Var(X)=2k

- χ^2 distribution is used in χ^2-test, most notably for categorical data analysis.

χ^2 distribution



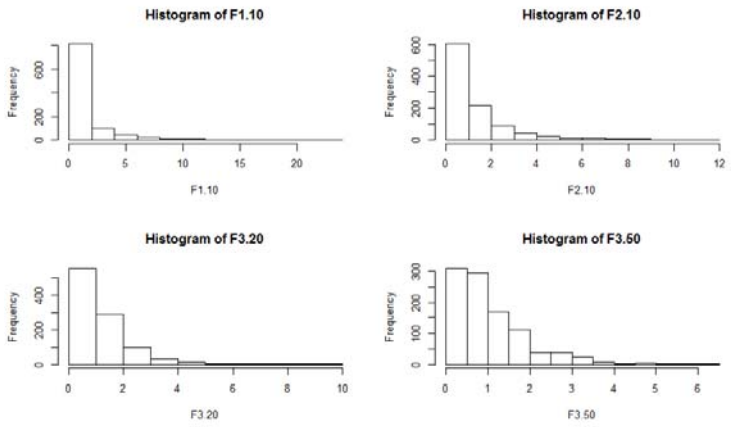
$\mathcal{F}$ distribution

- If χ_n^2 and χ_m^2 are independent following Chi-Square distribution with *n* and *m* degrees of freedom, then their scaled ratio has an $\mathcal{F}$ distribution with (*n, m*) degrees of freedom:

$$\mathcal{F}_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

- $\mathcal{F}$ distribution is used in the F-test, most notably regression analysis or the analysis of variance

$\mathcal{F}$ distribution



Few Summary Points

Probability distribution: It describes how the data are distributed, giving the likelihood that a certain value occurs

- Commonly used distributions:

Categorical data
- Binomial distribution
- Multinomial
- Poisson

Continues data
- Normal
- t-distribution
- Chi-Square (χ^2) distribution
- $\mathcal{F}$ distribution

Few Summary Points

Graphics:

- Histogram → Examine distribution of a random variable
- Boxplot → Examine the symmetry, spread, outliers
- Side-by-side Boxplots → Compare distributions (median/IQR)

Quantifying central points and variability:

- Mean → Measure of center, sensitive to outliers
- Median → Measure of center, not sensitive to outliers
- SD → Measure of spread, sensitive to outliers
- IQR → Measure of spread, not sensitive to outliers

Few Summary Points

Test for Normality:

- Histogram → Symmetric bell-shape
- Boxplot → Symmetric, no outliers
- Q-Q Plot → Visual comparison with a normal distribution
- Shapiro-Wilks test → Statistical test for normality