Statistics Vocabulary

Statistics **is** **a collection of mathematical methods to (SODI):**
- **State or formulate precise questions to investigate**
- **Obtain (create or collect) data relevant to the question**
- **Describe (summarize, display or analyze) data (Descriptive)**
- **Infer general conclusions from specific data (Inferential)**

Population (N) **the whole group of individuals to be studied**
- **Census** — **data collected about the population**
- **Frame** — **complete listing of the population**
- **Parameters** — **facts about the (whole) population**

Sample (n) **a subgroup of the population**
- **Survey** — **data collected about the sample**
- **Statistics** — **facts about the (whole) sample**

Data are values of Variables (X) measuring individuals characteristics:
- **Nominal** — **completely qualitative; no ordering**
- **Ordinal** — **ordered qualitative; no arithmetic**
- **Interval** — **quantitative; differences meaningful; fake zero**
- **Ratio** — **quantitative; ratios meaningful; real zero**
- **Discrete** — **small set of numeric values (no fractions)**
- **Continuous** — **large range of numeric values (fractions ok)**

Study **is a purely observational investigation**
Experiment **is an investigation of the effect of some modification**
- **Control groups expose confounding due to lurking variables**
- **Double-blind studies and placebos eliminate expectation errors**
- **Correlation versus causation requires additional controls**

Data is collected in three ways:
- **"Now" or Cross-sectional** — **data gathered at a point in time**
- **"Past" or Retrospective** — **data gathered over a past period**
- **"Future" or Prospective** — **data to be gathered in the future.**

Representative samples usually chosen randomly:
- **Convenience** — **individuals selected "by what's there"**
- **Systematic** — **every kth individual is selected**
- **Random** — **each individual has equal chance of selection**
- **Simple random** — **each possible sample has equal chance**
- **Stratified** — **population partitioned into (relevant) strata, individuals proportionately selected by strata**
- **Cluster** — **population partitioned into (irrelevant) clusters, clusters randomly chosen, whole cluster used.**

Possible errors:
- **chance or sampling error**
- **systematic bias: under-coverage, non-response, interviewer, etc.**

Descriptive Statistics

Frequency Distribution (Table) of a variable shows number with each value.

Graphical Presentations of Frequency Distributions:

- **Histogram:** best to determine "shape" of frequency distribution
  (Bar Chart: bars separated and graphs non-frequency data)
- **Dot:** draw bars with dots or other shapes
- **Stem-Leaf:** frequency of units digit
- **Pareto:** frequency of qualitative data (ordered by frequency)
- **Circle Plot (Pie Chart)** is best for parts of a whole
- **Line Plot (ogive or polygon)** connects dots, best for time-line data
- **Scatter Plot** is used to compare or correlate two variables

- **Box Plot (Fenced)** shows the 5 number summary (outlier limits)
  (provides alternative "shape" of the distribution)

Shapes of Data Distribution

- Normal is Symmetric and Bell-shaped
- Skewed may be (Tailed) Left or Skewed (Tailed) Right
- Bi-Modal and Multi-modal ("bumpy")

Measures of Center (Typical value):

Mode (MO)	most numerous value
Median (M or Q_2)	half-way point value (ordered list)
Mid-Range (MR)	average of Hi & Lo values
Mean (μ:mu or $\bar{x}$:x-bar)	average of all values

Measures of Spread or Dispersion or Position (Variety of values):

Quartiles (Q_1, Q_3)	1st & 3rd quarter-way point values
Inter-Quartile Range (IQR)	difference, $Q_3 - Q_1$
Range (R)	difference Hi − Lo
5 Number Summary	{ Lo, Q_1, M=Q_2, Q_3, Hi }
Standard Deviation (σ or s)	"rms average difference from mean"
Variance (σ^2 or s^2)	standard deviation squared (σ:sigma)

If the population distribution is: $Lo=X_1 \le X_2 \le X_3 \le \ldots \le X_{N-1} \le X_N=Hi$
(or n for sample), the basic parameters or statistics are the following:

Median:	$Q_2 = X_{(N+1)/2}$
First Quartile:	$Q_1 = X_{(N+1)/4}$
Third Quartile:	$Q_3 = X_{3(N+1)/4}$
Mid-Range:	$(X_N + X_1)/2$
Range:	$R = X_N - X_1$
Mean:	μ (or $\bar{x}$) $= (X_1 + X_2 + X_3 + \ldots + X_{N-1} + X_N)/N$
Variance:	$\sigma^2 = [(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots + (X_N - \mu)^2]/N$
	$s^2 = [(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + \ldots + (X_n - \bar{x})^2]/(n-1)$

Rules of thumb for any data (Chebyshev's Theorem):

75% of the data within: $\mu \pm 2\sigma$
89% of the data within: $\mu \pm 3\sigma$
96% of the data within: $\mu \pm 5\sigma$

Rules of thumb for Normal sample data:

*** $\mu \approx \bar{x} \approx M=Q2 \approx MR \approx MO$ "the norm"
*** $\sigma \approx s \approx ({}^3/_4)IQR \approx ({}^1/_4)R$ "the spread"
${}^1/_2$ of the data within: $\mu \pm ({}^2/_3)\sigma \approx M \pm ({}^1/_2)IQR$
*** ${}^2/_3$ of the data within: $\mu \pm \sigma \approx M \pm ({}^3/_4)IQR$
82% of the data within: $\mu \pm ({}^4/_3)\sigma \approx M \pm IQR$
*** 95% of the data within: $\mu \pm 2\sigma \approx M \pm ({}^3/_2)IQR$
99.7% of the data within: $\mu \pm 3\sigma$

Outliers are unusual occurances. For Normal data:

5% of the data outside: $\mu \pm 2\sigma \approx M \pm (1.5)IQR$
1% of the data outside: $\mu \pm ({}^8/_3)\sigma \approx M \pm (2)IQR$
1 in a million outside: $\mu \pm 5\sigma \approx M \pm (3.75)IQR$
1 in a billion outside: $\mu \pm 6\sigma \approx M \pm (4.5)IQR$

Measures of Data Distribution Shape

Coefficient of Variation: $CV = \sigma/|\mu|$; $\bar{x}/s$
Coefficient of Skewness: $SI = (\mu - M)/\sigma$; $(\bar{x} - M)/s$
Symmetric: $SI \approx 0$ Skewed Left: $SI<0$ Skewed Right: $SI>0$

The Z-score of X is number of standard deviations from the mean:

$z(X) = (X - \mu) / \sigma$ (population) $z(X) = (X - \bar{x}) / s$ (sample).

The Percentile of X_i : **The X_i of the Pth percentile:**

$$P(X_i) = \frac{100(\ \#\text{-values} < X_i)}{\{N,n\}} \ (\blacktriangledown) \qquad i = \frac{(\ \{N,n\} + 1) \times P}{100} \ (\blacktriangle)$$

Frequency Tables for Population or Sample with k distinct values:

k distinct values: $X_1\ X_2\ ,\ \ldots\ , X_{k-1}\ , X_k$
each occurring $f_1\ ,\ f_2\ ,\ \ldots\ ,\ f_{k-1}\ ,\ f_k$ times,
with total frequency $f_1 + f_2 + \ldots + f_{k-1} + f_k = N$ or n.
Arrange in table; graph with Histogram (Relative if using %'s).
Mean: $\{\mu,\bar{x}) = (X_1f_1 + X_2f_2 + \ldots + X_kf_k) / \{N,n\}$
Variance: $\{\sigma^2, s^2\} = [(X_1 - \mu)^2f_1 + \ldots + (X_k - \mu)^2f_k] / \{N,n-1\}$

If too many (or continuous) values, group (summarize) values into classes and construct frequency table for each class.

Classes defined by lower and upper limits (non-overlapping)
Class midpoint is average of consecutive lower limits
Class width is difference of consecutive lower limits

If original data is unavailable, use midpoint as X_i for entire class.

Correlation and Regression

For bi-variate raw data (draw scatter plot of ordered pairs):
$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, (N for population)

The coefficient of correlation is: (ρ, rho, for populations)

$$r = \frac{(X_1 - \overline{x})(Y_1 - \overline{y}) + (X_2 - \overline{x})(Y_2 - \overline{y}) + \ldots + (X_n - \overline{x})(Y_n - \overline{y})}{(s_x)(s_y)(n - 1)}$$

If r is close to +1:
the data pairs are linearly correlated with positive slope.
If r is close to -1:
the data pairs are linearly correlated with negative slope.
If r is close to 0:
the data pairs are not linearly correlated.

The "best fit" least-squares linear function is: $Y = aX + b$
with $a = rs_y / s_x$ & $b = \overline{y} - a\,\overline{x}$
where $\overline{x}$, s_x & $\overline{y}$, s_y are for x-data & y-data.

TI-83/84 Techniques

Random Numbers:	Seed-#,STO>,"MATH-PRB-1=rand",Enter "MATH-PRB-5=randInt(1,n)",Enter
Enter Data List:	STAT, EDIT, 1:Edit, Enter data: one column (L#) for simple raw data two columns (L#s) for frequency table two columns (L#s) for regression.
Sort Data List:	STAT, EDIT, 2:SortA, Enter L#, Enter.
Clear Data List:	Position cursor on L#, Clear, Enter.
Calculate Mean, Median, etc:	
(1 L#)	STAT, CALC, 1:1-Var Stats, "Enter L#", Enter
(2 L#s for Freq)	STAT, CALC, 1:1-Var Stats, "Enter L#s", Enter
Draw Graphs:	STATPLOT (2nd Y=) HiLite "1:", Enter HiLite "On", HiLite-Graph, Enter L#(s), Enter Window&Graph or Zoom-9(ZoomStat)
HiLite-Graph Choices:	1=Scatter, 2=XY-Line, 3=Histogram 4=Fenced-BoxPlot, 5=BoxPlot, 6=NormalityPlot

Coefficient of Correlation and Least-Squares Regression Line (Y = aX+b):
CATALOG (2ND 0); Scroll DiagnosticOn, Enter×2.
STAT, CALC, 4:LinReg, Enter L#s, Enter

Introduction to Naive Set Theory

A set is a collection of elements determined by a (perhaps implied) rule for membership. Write "s is an element of S" as: $s \in S$.

Finite sets are exhibited as lists enclosed within {...}. Order is irrelevant.
the set of the first 5 whole numbers is {1,2,3,4,5}.
the set of US Presidents = {Washington,Adams,…,Obama}.
Some sets are finite but too large to list. Some sets are infinite.

To count, conceive of the counting process as multi-case or multi-step:
multi-case: total count is the sum of the disjoint, case counts;
multi-step: total count is the product of the sequential step counts.

A set B is a subset of a set A, written $B \subset A$, if every element in B is in A.
The set of bachelor US Presidents is a subset.
The set with no elements is called the null set , Ø or {}..
The set of women US Presidents is the null set.

Given sets A and B, define the union (or) and intersection (&) as:
$A \cup B$ = AorB = {all elements in A or B (or both) }
$A \cap B$ = A&B = {only elements in both A and B }
A and B are disjoint if $A \cap B = A\&B = \emptyset$.
Given sets A and B, define the product as the set of ordered pairs:
$A \times B$ = { all (a,b) where a in A and b in B }

A relation is a rule which assigns to some elements of a domain (input) set, D , one or more elements of a range (output) set, R. Write this as:
$$R \subseteq D \times R.$$
A function is a rule which assigns to every element of a domain (input) set, D , precisely one element of a range (output) set, R. Write this as:
$$f: D \to R.$$
If D is the set of all people and R is the set of all dates, then birth-date is a function from D to R; while marriage-date is a relation, not a function.

A permutation is an ordered selection from a finite set. Permutations are shown as lists enclosed within <...>.
A combination is an unordered selection from a finite set. Combinations are shown as lists enclosed within {...} - combinations are subsets.
If set A has N elements:
the number of permutations with N elements selected is:
N factorial = N! = N*(N-1)*...*2*1 = N-pick-N
the number of permutations with R elements selected is:
$${}_NP_R = {}^{N!}/_{(N-R)!} = \text{N-pick-R}$$
the number of combinations with R elements selected is:
$${}_NC_R = {}^{N!}/_{(R!\,(N-R)!)} = \text{N-choose-R}$$

Probability Vocabulary

Probability requires a sample space, S, of outcomes of a random process.
Each execution of the random process is called a trial.
An event, A, is a result of the random process. A is a subset of S, $A \subset S$.
An outcome is an atomic event. An outcome is an element of S.

Probability measures the "frequency" or "likelihood" of an event.
If A is an event, the probability of A is P(A) with $0 \leq P(A) \leq 1$:
P(certain event) = P(something happens) = P(S) = 1
P(impossible event) = P(nothing happens) = P(Ø) = 0
P(rare event) < 0.05 (sometimes < 0.025 or < 0.01)

If all outcomes of the process are equally-likely, then:
P(A) = (# outcomes in A) / (total # of outcomes).
For any process, for a large number of trials, then approximately:
P(A) = (# trials producing A) / (total # of trials).
The Law of Large Numbers says the two definitions are consistent.

Given any two events, A, B:
AorB = $A \cup B$ = all outcomes in A or B or both
A&B = $A \cap B$ = all outcomes in both A and B
A^C = "Not A" = Complement of A = all outcomes not in A = $A' = \overline{A}$.

The laws of probability are rules for calculating unknown probabilities from known probabilities. The fundamental law is:
$$P(AorB) = P(A) + P(B) - P(A\&B)$$
It follows that: $P(A) + P(A^C) = 1$

Two events, A, B, are disjoint if: $A\&B = \emptyset$ or $P(A\&B) = 0$
(Think of disjoint events as mutually exclusive.)

Two events, A, B, are independent if: $P(A\&B) = P(A)*P(B)$
(Think of independent events as events which do not affect each other, like successive flips of a coin or rolls of a die.)

The conditional probability of "B given A", written P(B|A), is the probability that B will occur assuming that A has occurred. It is defined as:
$$P(B|A) = P(A\&B) / P(A)$$

If P(A) and P(B|A) are known, calculate: $P(A\&B) = P(B|A) * P(A)$

For any two events, A,B: $P(B) = P(B|A)*P(A) + P(B|A^C)*P(A^C)$ (Bayes).

If A,B are independent, then: P(B|A) = P(B), thus, P(B) does not change whether or not A is known to have occurred.

Discrete Random Variables

Discrete Random Variable, X, is the result of a random process with finite numeric outcomes, X, where each outcome has probability, P(x) = P{X=x} :

$$P(x) \geq 0 \qquad \Sigma\, P(x) = 1$$

P is the probability distribution function. The probability of an event is the sum of the probabilities of the outcomes which comprise the event.

The mean or average or expected value of a discrete random variable is:

$$\mu = \mu_x = E(X) = \Sigma\, x\, P(x)$$

The expected value is the expected average of a large number of trials. It may be used to determine if a bet is fair: E(X) = 0 ; or if a possible gain is worth the risk of loss: E(X) > 0 or < 0.

The variance and standard deviation, σ, of a discrete random variable is:

$$\sigma^2 = \sigma_x^2 = \Sigma\, (x - \mu_x)^2\, P(x)$$

The standard deviation measures likely (or average) closeness to the mean.

To calculate μ's and σ's on the TI-83, enter the probablity distribution like a frequency distribution and use: STAT - CALC - 1: 1-Var Stats (L_1,L_2)

Binomial probability distribution applies to a random variable, X, based on:
1. n independent trials with binary outcomes, {Success, Failure};
2. for each trial, P(Success) = p and P(Failure) = q = 1 – p;
3. X counts the number of successes:

$$P(X=x) = P(x) = {}_nC_x\, p^x\, q^{(n-x)} = \left({}^{n!}/_{x!\,(n-x)!} \right) p^x\, q^{(n-x)}$$

For a binomial random variable: $\mu = np$ $\sigma^2 = np(1-p)$
On the TI-83, calculate P(X = x) as: 2nd-Vars, binompdf(n, p, x)
On the TI-83, calculate P(X ≤ x) as: 2nd-Vars, binomcdf(n, p, x)

Poisson probability distribution applies to a random variable, X, counting the number of times an event occurs over an interval (usually of time) when the "average" or mean rate of occurrence over the interval, μ , is known:

$$P(X=x) = \mu^x\, e^{-\mu} / x! \qquad \sigma^2 = \mu$$

Continuous Random Variables

Continous Random Variable, X, is the result of a random process with continuous, infinite numeric outcomes, spanning a range, R. It has a probability density function, p(x), defined on R and satisfying:

$p(x) \geq 0$ **Area(under p(x) over all of R) = 1**

The probability that $a \leq X \leq b$ = $P(a \leq X \leq b) = P\{ a \leq X \leq b \}$ = Area under p(x) from a to b. For continuous random variables, $P(X=a) = 0$.

Uniform probability density defined on $\{a \leq x \leq b\} = [a,b]$ as:

$p(x) = 1 / (b - a)$ $\mu = (a+b)/2$

Exponential probability density (waiting times) is defined on $\{ x \geq 0 \}$, as:

$p(x) = ae^{-ax}$, where $a = {}^{1}/_{\mu}$, $Prob\{X \leq x\} = 1 - e^{-ax}$

Normal probability density (bell curve - most data) is defined on {all x} as:

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

The standard normal distribution (Z-score table) has $\mu = 0$ and $\sigma = 1$.

For an arbitrary normal random variable, translate x into a Z-score:

$Z(x) = (x - \mu) / \sigma$ **and lookup area in the Z-score table.**

$Prob\{ X < a \}$ = area for Z(a). $Prob\{ X > a \} = 1 -$ area for Z(a).
$Prob\{ a < X < b \}$ = area for Z(b) – area for Z(a).

Remember: $Prob \{\mu - \sigma < X < \mu + \sigma\} \approx 0.68$; $Prob \{\mu - 2\sigma < X < \mu + 2\sigma\} \approx 0.95$.

On the TI-83, calculate the $Prob\{ a < X < b \}$:

- **2nd-Vars, normalcdf(a, b, μ, σ);**
- **for negative infinity, $\{X < b\}$, use a = at least 5σ below μ;**
- **for positive infinity, $\{X > a\}$, use b = at least 5σ above μ.**

To find the value of x for which the $Prob\{ X < x \} = q$:

- **Use Z-score process in reverse, or,**
- **2nd-Vars, invNorm(q, μ, σ), yields the x.**

To find the value of x for which $Prob\{ X > x \} = q$:

- **Use Z-score process in reverse, or,**
- **2nd-Vars, invNorm(1 – q, μ, σ), yields the x.**

Inferential Statistics: Central Limit Theorem

Scenario: Sample of size n selected from Population of size N:

population mean = μ population proportion = p
population standard deviation = σ .

The sample mean = $\bar{x}$, sample proportion = $\bar{p}$, and sample standard deviation = s are all random variables. As such, there are:

$\mu(\bar{x})$ = mean of the sample mean
$\mu(\bar{p})$ = mean of the sample proportion
$\mu(s)$ = mean of the sample standard deviation
$\sigma(\bar{x})$ = standard deviation of the sample mean
$\sigma(\bar{p})$ = standard deviation of the sample proportion
$\sigma(s)$ = standard deviation of the sample standard deviation.

Warning! There's: s, σ, $\sigma(\bar{x})$, $\sigma(\bar{p})$, $\sigma(s)$. Do not confuse them.
When σ of the population is known, it should be used.
When σ is unknown:
For large samples, use s and assume normal distribution.
For small samples, use s and the t-distribution.

For any large ($N \geq 10{,}000$) population and for any sufficiently large ($n \geq 100$) simple random sample, the distribution of the sample mean, $\bar{x}$, is approximately normal with:

mean of sample mean = $\mu(\bar{x}) = \mu_x = \mu$
standard deviation of sample mean = $\sigma(\bar{x}) = \sigma_x = \sigma/\sqrt{n}$.

If the population is normal, then the sample mean is always normal.

For a Binomial Distribution with n independent binary trials (prob = p), with n is sufficiently large ($np(1-p) > 10$ or $np > 5$ and $n(1 - p) > 5$), then the distribution is approximately normal with:

mean = np standard deviation = $\sqrt{np(1 - p)} \leq \sqrt{n/4}$.

For any large population with population proportion, p, and for any sufficiently large, but not too large ($n < .05N$) simple random sample, the distribution of the sample proportion, $\bar{p}$, is approximately normal with:

mean of sample proportion = $\mu(\bar{p}) = p$

sigma of sample proportion = $\sigma(\bar{p}) = \sqrt{p(1 - p)/n} \leq 1/\sqrt{4n}$.

Estimating an unknown population parameter

	Proportion	Mean
Unknown parameter	p	μ
sample size	n	n
statistic	$\bar{p}$	$\bar{x}$
confidence interval	$\bar{p} \pm E$	$\bar{x} \pm E$
standard deviation of statistic (σ_s)	$1/\sqrt{4n}$	$\sigma/\sqrt{n}$ $s/\sqrt{n}$
margin of error (E)	(# of σ_s)$\times\sigma_s$	
confidence level	$1 - \alpha$	
# of σ_s	$Z(\alpha/2)$	$Z(\alpha/2)$; $t(\alpha/2,n-1)$

(TI-83: $Z(\alpha/2)$ = invNorm(α/2,0,1))

Hypothesis Testing of Population Parameters
(using symbols from above)

Test Hypothesis (H) will be one of { $<$?, $\neq$?, $>$? }

Null Hypothesis (H_0) is always { = ? }

Test Statistics (ts)	$(\bar{p} - ?)/\sigma_s$	$(\bar{x} - ?)/\sigma_s$
Test Hypothesis	$p > ?$	$\mu > ?$
P-Value (cv)	$Z(1-\alpha)$	$Z(1-\alpha)$ $t(1-\alpha,n-1)$
Reject H_0	ts > cv	ts > cv
Test Hypothesis	$p < ?$	$\mu < ?$
P-Value (cv)	$Z(\alpha)$	$Z(\alpha)$ $t(\alpha,n-1)$
Reject H_0	ts < cv	ts < cv
Test Hypothesis	$p \neq ?$	$\mu \neq ?$
P-Value (cv)	$Z(\alpha/2)$	$Z(\alpha/2)$ $t(\alpha/2,n-1)$
Reject H_0 if	ts > +cv / < -cv	ts > +cv / < -cv