Research Methods and Statistics for Economics

Biwei Chen

Overview

"Statistics is the grammar of science." This course teaches statistical language of data science, practices data analytical skills in statistical software, and applies statistical methods to conducting research in economics. The main objective is to learn just enough statistics to make scientific decisions in general and address data-driven economics, finance, and policy research questions in particular.

1. What is Statistics? Why do we need Statistics?
2. What is Data Analysis? How to best summarize and visualize data?
3. What is Probability? What are the essential rules, laws, and models?
4. What is Statistical Inference? How to perform Statistical Inference?
5. What are the relationships between Data, Statistics, and Probability?
6. How to combine Probability and Statistics in Data and Regression Analyses?
7. What constitute Empirical Economics Research? How to apply Statistical Methods?

Themes and Topics

Part I: Statistics and Data Analysis ▪ The Big Question/Picture ▪ Data Structure and Quality ▪ Data Collection and Sampling ▪ Exploratory Data Analysis EDA ▪ Data Cleaning and Management	Part II: Probability Theory and Models ▪ Counting Rules and Event Relations ▪ Joint and Conditional Probabilities ▪ Addition Law and Multiplication Law ▪ Law of Total Probability and Bayesian Law ▪ Random Variables - Probability Distributions
Part III: Essential Statistical Inference ▪ Random Sample (IID) ▪ Sampling Distribution ▪ Parameter Estimation ▪ Confidence Intervals ▪ Hypotheses Testing	Part IV: Research Methods and Regression ▪ Empirical Economic Research ▪ Simple Linear Regression ▪ Multiple Linear Regression ▪ Probabilistic Regressions ▪ Estimation and Inference

Foundational Laws & Theorem

1. Addition (Union) Law: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

2. Multiplication (Intersection) Law: $P(AB)=P(A)P(B|A)=P(B)P(A|B)$

3. Law of Total Probability: $P(A)=P(AB)+P(AB^C)=P(B)P(A|B)+P(B^C)P(A|B^C)$

4. Bayes' Theorem/Law (Inverse Conditional Probability): $P(A|B)=P(A)P(B|A)/P(B)$

5. Law of Large Numbers (IID): Convergence in Probability of the Sample Statistic to Population

6. Central Limit Theorem (IID): Convergence in Distribution of the Sample Mean to Normal

$\bar{x} \sim N(\mu, \sigma^2/n)$ $(\bar{x}-\mu) \sim N(0, \sigma^2/n)$ $n^{0.5}(\bar{x}-\mu) \sim N(0, \sigma^2)$ $n^{0.5}(\bar{x}-\mu)/\sigma \sim N(0, 1)$

Confidence Intervals: 90%, 1.645SE around the point estimate; 95%, 1.96SE; 99%, 2.576SE

Lecture 1 The Essence of Statistics

Biwei Chen

I. What is Statistics?

1. Karl Pearson: Statistics is the grammar/language of science.
2. Statistics is the art and science of collecting, analyzing, presenting, and interpreting data.
3. Hal Varian: The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades.

II. Why Statistics (more than ever before)?

1. Demand for effective/scientific decision making (personal, professional, business, public policy)
2. Supply of data—big data (large size, high dimension, complex structure, continuous)
3. The high objective in this course: conduct empirical research in economics

III. How to Apply Statistics? An Anatomy

1. Blackbox f(.): Input (data) → output (analysis)
2. Technology: Statistical inference (sampling, estimation, hypothesis testing)
3. Foundation/principles: Probability theory, including rules, laws, and models

IV. What is Data? "Data are the world's most valuable resources."

1. Definition: Data are information for learning and decision making.
2. Data sources (collection): web, administration, experiment, survey
3. Data storage: data table consists of observations (ID) in rows and variables (features) in columns
4. Data structure: cross-section (xsec), time series (tseries), panel or longitudinal (xt data), spatial
5. Data type (variables): qualitative (categorical/factor labels) and quantitative (numeric values)
6. Data measurement scales: nominal, ordinal, interval, ratio.
7. Data quality: validity (content), reliability (stability), comparability, coverage (unbiased selection)
8. Ethical and legal issues (two principles): 1) confidentiality/privacy; 2) ownership and usage rights

V. How to Collect (useful) Data? Data Sampling

1. Population: The set of all observations relevant for the analysis.
2. Sample: The subset for which data is collected. Purpose: samples can be extrapolated to the population. Biased samples won't be representative enough of the population.
3. Why practice sampling? 1) Feasibility; 2) Cost (time and money); 3) Update
4. Sampling approaches: 1) Random; 2) Non-random (criterion or rule-based)
5. Potential problems in sampling: Sampling biases (non-sampling error, population overrepresented or underrepresented in the data): 1) Exclusion; 2) Self-selection; 3) Non-response; 4) Survivorship
6. Random sampling is the process that most likely leads to representative samples.
7. Sampling error/noise (randomness): decrease as the size of a sample increases.

VI. How to Analyze Data? Data Analysis or Analytics

1. Descriptive or exploratory statistics
2. Modeling, estimating, testing, forecasting
3. Descriptive, predictive, prescriptive

Lecture 2: Exploratory Data Analysis

I. EDA or Descriptive Data Analysis

1. What is EDA? describes the features of the variables of interest in the data.
2. Why EDA? 1) check data cleaning; 2) guide subsequent analysis; 3) provide context of the results of subsequent analysis; 4) answer simple questions and ask additional questions
3. How to perform EDA? Routines: 1) focus on the most vital variable(s); 2) list frequencies (qualitative) and histograms (quantitative); 3) check for extreme values ("outliers"); 4) provide summary statistics; 5) explore further if necessary (cross section comparison, time series data)

II. Basic Concepts: Data Frequencies, Probability, and Distribution

1. Frequencies describes the occurrence of specific values (or groups of values) of a variable.
   1) Absolute frequency is the number of observations (counts)
   2) Relative frequency is the percentage, or proportion (%)
2. Probability is a measure of the likelihood of an event; its value is always between 0 and 1. With data, probabilities are relative frequencies; the "events" are assigned specific values of the variable.
3. Distribution (of a variable) shows the frequency of each potential value of the variable in the data.

III. Summary (Descriptive) Statistics: One Variable

1. Centrality and location: mean (arithmetic, geometric, weighted), median, and mode
   1) Arithmetic (simple average=equal weighted) vs weighted mean (e.g., S&P500)
   2) Geometric: What is the average compounded rate of return over multiple periods?
   3) Median: the midpoint (odd #) or the average of two midpoints (even # of obs)
   4) Mode: the most frequently observed value in the sample (least used measure)
2. Dispersion: range (max-min), variance, standard deviation, coefficient of variation (SD/E(X))
   1) Range (max–min) and standardized range: (max–min)/[(max+min)/2]
   2) Percentile: how the data are spread over the intervals from the smallest to the largest value
   3) Quartile: first quartile Q_1 is the 25th percentile. Interquartile range ($IQR=Q_3–Q_1$)
   4) Variance var(X): population σ^2 versus sample s^2 (average squared distance from the mean)
   5) Standard deviation σ_X: SD not comparable across two samples with different units or levels
   6) Coefficient of variation $\sigma_X/E(X)$: comparable across samples with different units
3. Asymmetry: skewness (deviations from the mean or the mean-median difference) **E(X)–Median/SD**
   1) Symmetric: no skewness (deviations fall roughly equally on both sides of the mean)
   2) Symmetric distribution → Mean=Median=Mode (for distribution with a single mode)
   3) Positively skewed: deviations are more pronounced/extreme for observations mean > median
   4) Negatively skewed: deviations are more pronounced/extreme for observations mean < median
4. Extremum: kurtosis measures peakedness of the distribution in relation to the shape of the tails (thickness of the tails). How frequently do data take extreme values relative to the central?
   1) Mesokurtic (k=3): normal distribution. Excess kurtosis=kurtosis minus three.
   2) Leptokurtic (k>3): distributions have more frequent occurrences of extreme values (fat tails)
   3) Platykurtic (k<3): distributions have less frequent occurrences of extreme values

Note: 1) Symmetric distribution → zero skewness, but not vice versa. 2) Variance and kurtosis are both affected by the presence of extreme values, but they measure different phenomenon. High (low) variance and low (high) kurtosis can be present at the same time for the same distribution. Moments: mean, variance, skewness, kurtosis, and higher moments.
Note: What are the causes for "outliers"? 1) Mistakes in digits or units of measurement; 2) Inherent (e.g., superstars)

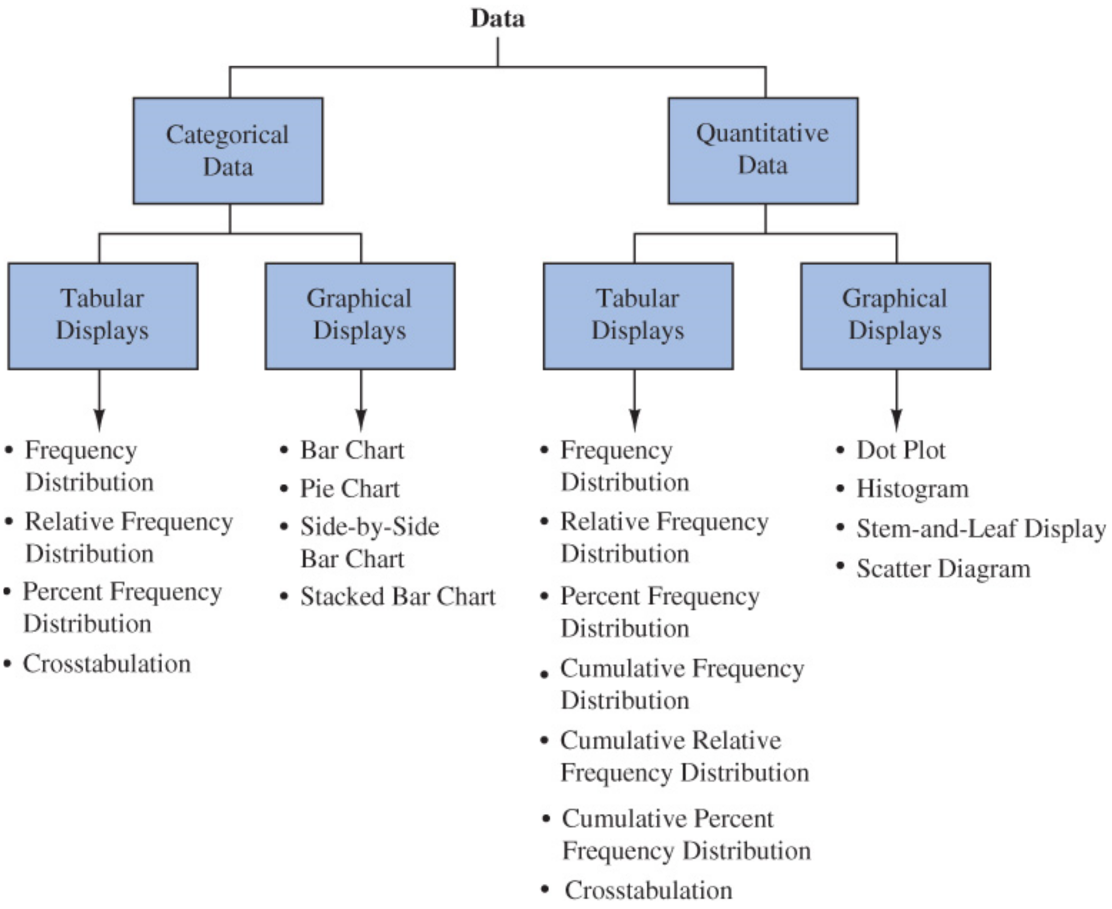
Note: z-scores and the empirical rule for normal distribution can help detect extreme values (e.g., for z>2 or 3). Chebyshev's theorem provides a lower bound for the mass of observations in the distribution.

IV. Summary (Descriptive) Statistics: Two Variables

1. Covariance: an absolute-value measure of co-movement between two variables.
2. Correlation: a standardized measure of the linear relationship between two variables.
3. Correlation coefficient [-1, 1]: 1) positive (>0.3); 2) negative (<-0.3); 3) weak or zero (-0.3, 0.3)
4. Both measures can be applied to quantify linear relationship in time series or cross-section data.

V. Data Visualization

1. Formats: histogram (discrete), density plot (continuous), box plot, and violin plot
   1) Histogram: a bar graph showing the frequency of each value of a variable
   2) Density plot: continuous version of histogram (kernel density estimates)
   3) Box plot: a visual representation of many quantiles and extreme values
   4) Violin plot mixes elements of a bot plot and a density plot (rotated)
   5) Dot plot and stem-and-leaf display: both are simple histogram
2. Relationships: 1) scatter plot; 2) linear or nonlinear fit; 3) high-dimension plots
3. Guidelines for data visualization: 1) purpose; 2) focus; 3) audience



Graph Source: Camm et al. (2024) CH2, Essential of Statistics for Business and Economics

Lecture 3: Probability Concepts and Rules

I. Counting Rules: Combination and Permutation

1. Rule of product: 3 shirts and 4 pant, how many outfits? This belongs to a multi-step decision.
2. Combinations (no order): how many ways to choose x (subsets) out of n? C=n!/x!(n–x)!
3. Permutations (with order): how many ways to order x out of n? P=n!/(n–x)! and C=P/x!
4. Example: What are the C and P for choosing three out of {a, b, c, d}? C=4, P=24.
   How many ways to order 52 cards? 52 factorials.

II. Basic Concepts

1. A single experiment is a repeatable procedure that leads to the occurrence of a single outcome from a set of possible outcomes. A single outcome s is referred to as a sample point or element.
2. Outcomes set or universal set: all possible distinct outcomes of the experiment. Sample space (S): the universal set containing all possible outcomes/objects. $S=\{s_1, s_2, \ldots, s_n\}$
3. Event (A): any possible subset of the outcomes set ($A\subseteq S$). Depict in words, notation, diagram.
   1) An elementary event is an event that contains a single sample point s.
   2) A sure event always occurs (universal set). An impossible/null (empty set) event never occurs.
4. A joint event refers to an outcome of an experiment where the sample point is two dimensional.
5. A set is a collection of objects, where an "object" is a generic term that refers to the elements (or members) of the set. The notation $a\in A$ denotes that the object (element) a belongs to the set A.
6. Subset ($B\subseteq A$): B is a subset of A if every member of set B is also a member of set A. Furthermore, B is a proper (strict) subset of A if A contain at least one member not in B. $B\subset A$

II. Event (Set) Relationships via Venn (Euler) Diagrams

1. Union ($A\cup B$): the event containing all sample points in A or B or both. $A\cup B=B\cup A$.
2. Intersection ($A\cap B$): the event containing the sample points in both A and B. $A\cap B=B\cap A$
3. Empty set Ø (impossible event) has no element. Universal set S consists of all possible outcomes.
4. Complement A^C: the event consisting of all elements not in A but in S (event A does not occur), $A^C=\{b \mid b\notin A, b\in S\}$. $A\cup A^C=S$, $A\cap A^C=\emptyset$, $(A\cup B)^C=A^C\cap B^C$, $(A\cap B)^C=A^C\cup B^C$
5. Difference: an event A minus B, denoted by A\B, consists of all elements in A but not in B.
   $A\backslash B=\{a \mid a\in A, a\notin B\}$. $A\backslash B=A\cap B^C$. $A\backslash B \neq B\backslash A$ (non-commutative). $A\backslash B = B\backslash A$ if and only if A=B
6. Mutually exclusive events: if one event occurs, then the other cannot (only one). Two events are disjoint if no element in common. $A\cap B=\emptyset$. Exhaustive: one of the two events must occur $A\cup B=S$.

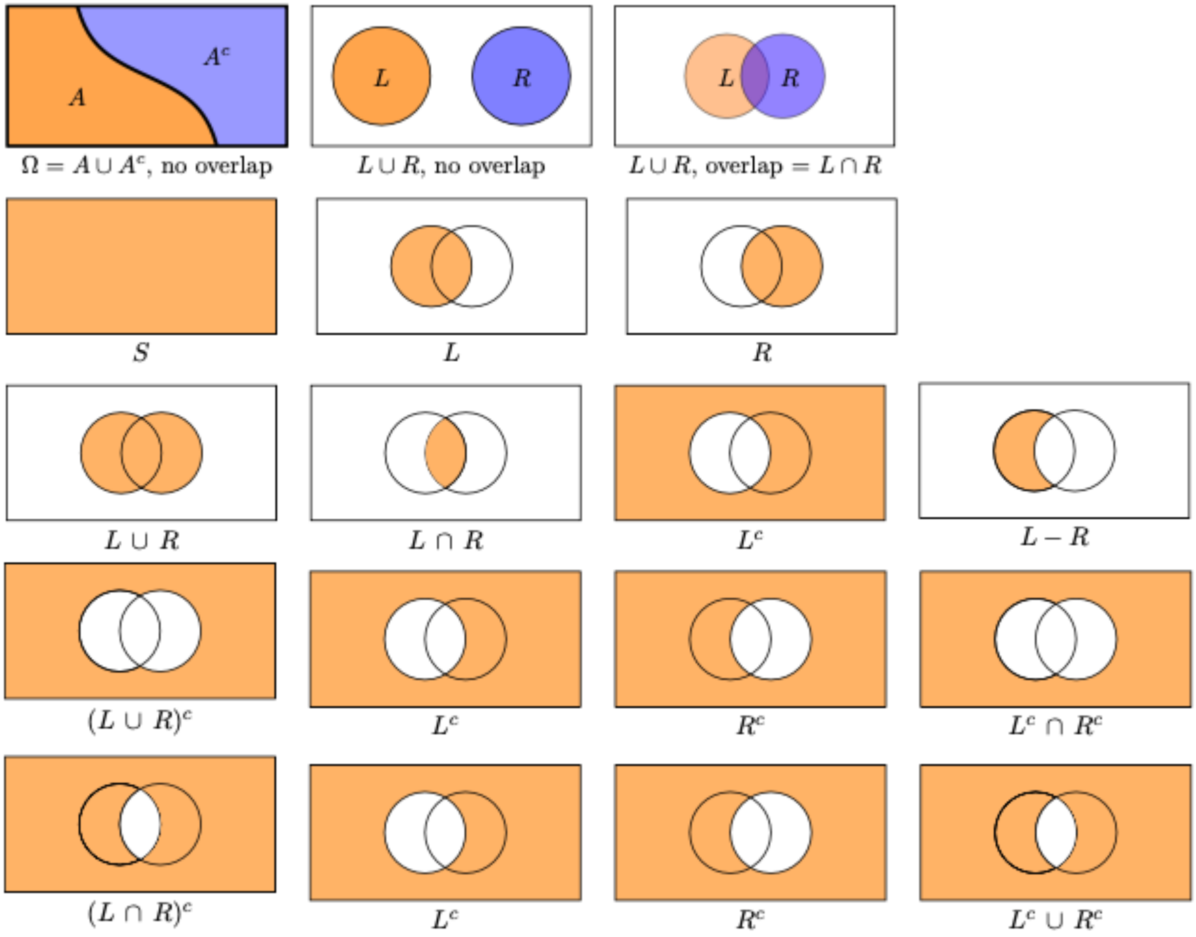
IV. Events and Probabilities

1. A probability is a real number assigned to events in a sample space to represent their likelihood of occurrence. Sample space: $S=\{s_1, s_2, \ldots, s_n\}$. Outcome: $s\in S$, $\sum_i P(s_i)=1$, $\sum_{(s\in A)} P(s)=1$
2. P(A) denotes the probability of the event $A\subseteq S$. $P(A)= \sum_{(s\in A)} P(s)$
3. Axioms (by construction): 1) Non-negativity $0\leq P(A)\leq 1$; 2) Additivity $P(A\cup B)=P(A)+P(B)$ if $A\cap B=\emptyset$ (for disjoint events); 3) Inclusion-exclusion principle $P(A\cup B)=P(A)+P(B)–P(A\cap B)$.
4. Probability of a subset: $0=P(\emptyset)\leq P(A)\leq P(B)\leq P(S)$ if $A\subseteq B\subseteq S$
5. Complement: $P(A^C)=1–P(A)$ or $P(A)+P(A^C)=1$ or $P(A\cup A^C)=P(S)=1$
6. Mutually exclusive events: $P(A\cap B)=0$ since $A\cap B=\emptyset$. → Two independent events with positive probability cannot be mutually exclusivity (proof by definition).
7. Joint probability P(AB)=P(BA), marginal prob P(A) and P(B), conditional prob $P(A|B) \neq P(B|A)$

V. Probability Rules and Laws: how to calculate the probability of relational events

1. Addition law (union): $P(A\cup B)=P(A)+P(B)-P(A\cap B) \rightarrow P(A)+P(B)\geq P(A\cup B)\geq P(A\cap B)$
2. Multiplication law (intersection): $P(AB)=P(A)P(B|A)=P(B)P(A|B)=P(BA)$
3. Independent events: 1) $P(A\cap B)=P(AB)=P(A)P(B)$; 2) $P(A|B)=P(A)$ and $P(B|A)=P(B)$.
4. Conditional events: $P(A|B)=P(AB)/P(B)$ and $P(B|A)=P(BA)/P(A)$. Given the occurrence of B, A happens within the context of B. Conditional probability is calculated as a proportion of the joint probability to the marginal probability: Narrowing down the focus to the event being conditioned.
   1) If A and B are independent, then $P(A|B)=P(A)$ and $P(B|A)=P(B)$
   2) If B provides useful information about A, then $P(A|B)$ will be larger or smaller than $P(A)$
5. Law of total probability $P(A)=P(AB_1)+P(AB_2)=P(B_1)P(A|B_1)+P(B_2)P(A|B_2)$. [$B_i$ partitions S]
6. Bayes' theorem/rule/formula: $P(A|B)=P(A)P(B|A)/P(B)$ to flip/invert conditional probabilities
   1) Enhanced version: $P(A_1|B)=P(A_1)P(B|A_1)/P(B)$ where $P(B)=P(A_1)P(B|A_1)+P(A_2)P(B|A_2)$
   2) Philosophy: Prior probabilities are $P(A_i)$, which is often subjective; and posterior probability is $P(A_i|B)$ in which B represents new evidence for updating the prior probabilities.

Intersection is what they have in common; union is what they have altogether.



$\Omega = A \cup A^c$, no overlap

$L \cup R$, no overlap

$L \cup R$, overlap $= L \cap R$

S

L

R

$L \cup R$

$L \cap R$

L^c

$L - R$

$(L \cup R)^c$

L^c

R^c

$L^c \cap R^c$

$(L \cap R)^c$

L^c

R^c

$L^c \cup R^c$

Source: Jeremy Orloff and Jonathan Bloom (2014) Introduction to Probability and Statistics

Lecture 4: Probability Distributions

I. Basic Concepts

1. Random variable: a function that maps the outcome (random experiment) to a real number.
   1) Discrete: the range is countable subset of the real line
   2) Continuous: the range is any uncountable subset of the real line
2. Cumulative distribution function (CDF): a function F(x) from the real number to the interval [0,1].
   1) $F(a)=P(X\le a)$ right-continuous; 2) $0\le F(x)\le 1$; 3) $F(a)\le F(b)\ \forall\ a\le b$ non-decreasing
3. Probability density function (PDF) f(x): assigns a probability to any range of realization of a RV.
   1) Discrete (PMF): $0\le f(x)=P(X=x)\le 1$; $\sum_i P(x_i)=1$; $P(a<X\le b)=F(b)-F(a)=\sum_{(a,b]}P(X=x_i)$, $(a<b)$.
   2) Continuous (PDF): $f(x)\ge 0$; $P(a<X\le b)=F(b)-F(a)=\int_{(a,b]}f(x)\le 1$; $dF(x)/dx=f(x)$; $\int f(x)dx=1$.
   3) What is the probability of getting a number out of an infinite-dimension dice? "Ball dice"

II. Probability Model and Moments

1. Probability model specifies a family of densities $f(x;\theta)$, defined over the range of the random variable X, one density function for each value of the parameter θ for $\theta\in\Theta$ (parameter space).
   1) Full feature: the parameters define a family of distributions (CDF & PDF).
   2) Partial features: moments and central moments; but moments do NOT determine F(x) uniquely
2. Mean (μ), variance (σ^2), and moments
   1) Mean: the weighted average. $E(x)=\sum xP(x)$; $E(x)=\int xf(x)dx$; $\bar{x}=\sum x/n$
   2) Variance: the weighted squared distance from the mean. $V(X)=E[(X-\mu)^2]=E(X^2)-E^2(X)$ $V(x)=\sigma^2=\sum(x-\mu)^2P(x)$; $V(x)=\int(x-\mu)^2f(x)dx$; $S^2=\sum(x-\bar{x})^2/(n-1)$
   3) Covariance: $Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$; $s_{xy}=\sum(x-\bar{x})(y-\bar{y})/(n-1)$
   4) Raw moments (μ_k'): Discrete $E(x^k)=\sum x^k P(x)$; Continuous $E(x^k)=\int x^k f(x)dx$
   5) Central moments (μ_k): Discrete $\mu_k=\sum(x-\mu)^kP(x)$; Continuous $V(x)=\int(x-\mu)^kf(x)dx$

III. Properties of E(.), V(.), Cov(.), Cor(.)

1. $E(X+Y)=E(X)+E(Y)$; $E(aX+bY+c)=aE(X)+bE(Y)+c$; $E[g(x)]=\sum g(x)P(x)=\int g(x)f(x)dx$
2. $V(aX+c)=a^2V(X)$; $V(X+Y)=V(X)+V(Y)+2COV(X,Y)$; $V(X-Y)=V(X)+V(Y)-2COV(X,Y)$
3. $V(aX+bY+cZ)=a^2V(X)+b^2V(Y)+c^2V(Z)+2abCOV(X,Y)+2acCov(X,Z)+2bcCov(Y,Z)$
4. $\sigma_{XY}=Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]=E(XY)-E(X)E(Y)$; $Cov(aX+b, cY+d)=acCov(X, Y)$
5. $\rho_{XY}=Cor(X,Y)=\sigma_{XY}/\sigma_X\sigma_Y$. Independence → No correlation ($\rho_{XY}=0$), but NOT vice versa
6. Variance-Covariance-Correlation Matrix: diagonal + off-diagonal = N(N+1)/2 elements

V. Conditional Mean, Variance, and Moments

1. Conditional mean $E(Y|X=x)=\sum y\bullet f(y|x)$ and $E(Y|X=x)=\int y\bullet f(y|x)dy$
2. Conditional variance $V(Y|X=x)=\sum[y-E(Y|X=x)]^2\bullet f(y|x)$ and $V(Y|X=x)=\int[y-E(Y|X=x)]^2\bullet f(y|x)dy$
3. Raw moments $E(Y^r|X=x)=\int y^r\bullet f(y|x)dy$, r=1, 2, …,
4. Central moments $E\{[Y-E(Y|X=x)]^r|\ X=x\}=\int[y-E(y|x)]^r\bullet f(y|x)dy$

https://en.wikipedia.org/wiki/Mean
https://en.wikipedia.org/wiki/Variance
https://en.wikipedia.org/wiki/Covariance
https://en.wikipedia.org/wiki/Central_moment
https://en.wikipedia.org/wiki/Moment_(mathematics)
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

V. Discrete Random Variables (Theoretical Distributions)

Distribution	PMF	Scenarios (random experiment)	E(X)	V(X)
Uniform	$P(x)=1/n$	equal chance of each among n outcomes	$(n+1)/2$	$(n^2-1)/12$
Bernoulli	$P(X=1)=p$	A single trial with success rate p	p	$p(1-p)$
Binomial	$P(X=k)=$ $C(n, k)p^k(1-p)^{n-k}$	k successes among n independent trials (the sum of independent Bernoulli trials)	np	$np(1-p)$
Geometric	$P(k)=(1-p)^{k-1}p$	the first success comes up on the nth trial	$(1-p)/p$	$(1-p)/p^2$
Negative Binomial	$P(k)=C(n-1, k-1)\bullet$ $p^{k-1}(1-p)^{n-k}p$	the kth success is on the nth trials (exactly k-1 successes among first n-1 trials)	$k(1-p)$ $/p$	$k(1-p)/p^2$
Hypergeometric	$P(k)=C(m, k)\bullet C$ $(N-m, n-k)/C(N, n)$	N items: m type success, N-m type failure k successes in n trials without replacement	nm/N	$n(m/N)(N-m/N)$ $(N-n/N-1)$
Multinomial	$P(k)=(n!/x_1!x_2!...x_k!)$ $p_1^{x1}p_2^{x2}... p_k^{xk}$	Models the probability of counts for each side of a k-sided dice rolled n times.	$E(x_i)$ np_i	$V(x_i)$ $Np_i(1-p_i)$
Poisson	$P(k)= \lambda^k e^{-\lambda}/k!$	k occurrences in a given interval, λ mean	λ	λ

Note: "e"–Euler's number. Taylor expansion $e^{\lambda}=\sum_{[0, \infty]} \lambda^k/k!$

VI. Continuous Random Variables (Theoretical Distributions)

Distribution	PDF	Description	E(X)	V(X)
Uniform U(a, b)	$f(x)=1/(b–a)$	Equal chance taking a value between a and b	$(a+b)/2$	$(b–a)^2/12$
Normal $N(\mu, \sigma^2)$ Standard Normal	$f(x)=(1/2\pi\sigma^2)^{-0.5}$ $\exp\{-0.5(x-\mu)^2/\sigma^2\}$	Model "error/noise/shocks" in econometrics Standardized normal: $Z=(X–\mu)/\sigma \sim N(0,1)$	μ 0	σ^2 1
Lognormal lnN(.)	$X=\exp(\mu+\sigma Z)$	If the log of X is normal, then X is lognormal	$e^{\mu+\sigma2/2}$	
Exponential (λ)	$f(x)=\lambda e^{-\lambda x}$, $x>=0$	Model durations (intervals) $F(x)=1–e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
Pareto (m, α)	$f(x)=\alpha m^{\alpha}/x^{\alpha+1}$	Power-law. $F(x)=1– m^{\alpha}/x^{\alpha}$, $x>=m$		
Chi-square $\chi^2(k)$	$\chi^2(k)=\sum_{[1, k]}Z_i$	Sum of squared independent standard normal	k	$2k$
$F=(U/m)/(V/n)$	$F_{m, n}$	Ratio of two independent chi-square U and V	$n/n-2$	
Student t	$T=Z/sqrt(V/m)$	Small sample standard normal with dof=n-1		

V. Probability Approximations

1. Binomial distribution can be approximated by a normal distribution when n is large enough ($np>=5$ and $nq>=5$). Binomial RV converges to normal distribution when n increase: $\mu=np$, $\sigma^2=(npq)^2$
2. Poisson distribution is an accurate approximation ($\lambda=np$) to the binomial $X\sim B(n, p)$ when n is large and p is small. Poisson RV converges to exponential distribution when
3. The empirical rule: ~68.26% within one SD of the mean; ~95.44% within 2SDs; ~99.7% within 3SDs. CIs: 90% C.I. is 1.6SE interval around the estimate; 95% C.I. is 1.96SE; 99% C.I. is 2.6SE

VI. Distributions for Hypothesis Testing

1. Z distribution (standard normal, applied in the central limit theorem)
2. t distribution (testing single coefficient significance; testing equality of two population means)
3. F distribution (joint test of regression significance; ratio of two population variance)
4. χ^2 distribution (test the population variance; normality; maximum likelihood tests)

What is the difference of $E(X)=\sum x/N$ and $E(X)=\sum x\bullet p(x)$? Answer: no difference. When each observation is weighted equally, $p(x)=1/N$. When some observations are more frequent, e.g., occur n times, $p(x)=n/N$.

Lecture 5: Statistical Inference I

I. What is Statistical Inference (SI)? Why? How?

The question for statistical inference: What is the true value of a statistic of the population?

1. Statistical inference is the act of generalizing from the data to some more general patterns.
    1) Goal: learn about the population, or general pattern, that the data represents
    2) Function: uncover the unknown true value(s)/parameter(s) in the population
2. Process: 1) consider a statistic; 2) compute its estimated value; 3) infer its value in the population
3. Assess external validity: 1) define the population of interest; 2) define the population the data represents; and 3) compare the two to assess external validity of our inference
4. No statistical methods can save a bad sample. How to obtain a representative sample?

II. Random Sample and Sampling

1. Random sample assumption (IID): Independence and Identical Distribution
2. Sampling refers to a procedure to select a number of objects from a larger set, the target population
3. The sampling procedure gives rises to a random sample (IID) when
    1) The probability of selecting any one of the population objects is the same
    2) The selection of one does not affect it is not affected by the selection of others

III. Repeated Samples and Sampling Distribution

1. Repeated samples: each observed dataset can be viewed as a sample drawn from the population.
2. Given an IID sample of the data $x=(x_1, x_2, \ldots, x_n)$, a statistic is a function of the sample (data). Not all statistics are created equal, some are useful for estimating parameters and/or testing hypotheses.
    1) A statistic is a random variable because it is a function of the data (sample). Dependent on it.
    2) With repeated samples, the statistic has a distribution and its value differs across samples
    3) **The sampling distribution of a statistic is the distribution of the statistic across repeated samples of the same size. Standard error is the standard deviation of the sampling statistic.**
3. Sampling distribution for common sampling statistics, given $x_i \sim IID(\mu, \sigma^2)$ and $p_i \sim IID(p, pq/n)$
    1) $\bar{x}=\sum x/n \sim N(\mu, \sigma(\bar{x}))$ where $\sigma(\bar{x})=\sigma/n^{0.5}$ for $n<5\%N$, otherwise $\sigma(\bar{x})=(\sigma/n^{0.5})[(N-n)/(N-1)]^{0.5}$
    2) $\bar{p} = I/n \sim N(p, \sigma(\bar{p}))$ where $\sigma(\bar{p})=[p(1-p)/n]^{0.5}$ for $n<5\%N$, otherwise apply FPCF.
    3) $s^2=\sum(x-\bar{x})^2/n-1$: $(n-1)\bullet s^2/\sigma^2 \sim \chi^2(n-1)$, given $x_i \sim N(\mu, \sigma^2)$

IV. Estimation: Concepts, Methods, and Quality

1. Estimator vs estimate: 1) An estimator ($\hat{}\theta\hat{}$) is a function (formula) for computing the statistic; 2) An estimate is the numerical value of the statistics given a particular data sample.
2. Different samples produce different estimates, but they can share the same estimator.
3. Point and interval estimation: **interval estimate = point estimate ± margin of error**
4. Sampling statistic is a point estimator. SE is calculated to construct confidence intervals.
5. Estimation methods: 1) Method of moments; 2) Least square; 3) Maximum likelihood; 4) Bayesian
6. Estimation quality ($\hat{}\theta\hat{}$): 1) Unbiasedness (accuracy); 2) Efficiency (precision); 3) Consistency

V. Properties of the sampling distribution (with large sample size n, asymptotic)

1. Unbiasedness: sample average converges to population mean as n increases to infinity
2. Asymptotic normality: the sampling distribution of the sample mean is approximately normal
3. Root-n convergence: the S.E. is inversely proportional to the square root of the sample size n

Lecture 6: Statistical Inference II

I. Theoretical Foundations for Statistical Inference: The distribution of sample means approaches the normal distribution as the sample size gets larger, regardless of the shape of the population distribution (CLT). The spread of the sampling distribution decreases as the sample size increases (LLN).

1. Asymptotics describes the behavior of statistics as the sample size limits to infinity.
2. Law of Large Numbers LLN (X~IID): a sample average can be brought as close as the average in the population (the true mean) from which it is drawn, by enlarging the sample size.
   1) WLLN: Convergence in probability $\bar{x}=\sum x/n \rightarrow \mu$, as $n\uparrow$, for IID x, with probability one as n goes to infinity. The LLN assures us that the estimated value is likely close to the true value, but it does not tell us how close it is. This is where the CLT can help.
   2) Examples: toss a coin for P(H); throw a dice to estimate the average value. Applications: how can we estimate the true probability of a biased coin? A biased dice? Refer to online simulation.
3. Central Limit Theorem CLT (X~IID): In large samples, the estimated average (sample average) is distributed approximately normal around the true expected value (population mean), and the variance of this distribution equals the variance of the variable divided by the sample size.
   $\bar{x} \sim N(\mu, \sigma^2/n)$ $(\bar{x}-\mu) \sim N(0, \sigma^2/n)$ $n^{0.5}(\bar{x}-\mu) \sim N(0, \sigma^2)$ $n^{0.5}(\bar{x}-\mu)/\sigma \sim N(0, 1)$

II. Interval Estimation for Population Mean and Proportion

1. How confident are we in the estimated sampling statistics and its distribution that represent the true value in the population? The CI gives the range of values where we think that the true value falls with a (1–α) % likelihood of the (particular) sample estimate.
2. Confidence interval with confidence level (1–α)% containing a true parameter: **$\text{Prob}(\theta \in CI_{1-\alpha})=1-\alpha$**
3. Steps to construct confidence intervals based on a given sample
   1) Estimate the parameter(s) with the point estimate(s), which is the sampling statistic
   2) Calculate the standard error via the formula (for unknown σ^2, substitute with s^2)
   3) 90% CI is 1.645SE around the point estimate; 95% CI, 1.96SE; 99% CI, 2.576SE
4. Formula: a 95% CI is the range [point estimate–1.96•SE, point estimate+1.96•SE]
5. Sample size requirement (given a desired margin of error and a planning value *p**)

   1) $\text{ME} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n = \frac{(z_{\alpha/2})^2\sigma^2}{\text{ME}^2}$ 2) $\text{ME} = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \rightarrow n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{\text{ME}^2}$

Distribution	**Sample Size n**	**Population**	**Statistics**	**1-α Confidence Interval**
x_i ~iid $N(\mu, \sigma^2)$ or x_i ~iid (μ, σ^2)	Any	σ^2 Known	$\bar{x}-\mu/(\sigma/n^{0.5}) \sim z$	$[\bar{x}-z_{\alpha/2}(\sigma/n^{0.5}), \bar{x}+z_{\alpha/2}(\sigma/n^{0.5})]$
	Small	σ^2 Unknown	$\bar{x}-\mu/(s/n^{0.5}) \sim t(n-1)$	**$[\bar{x}-t_{\alpha/2}(s/n^{0.5}), \bar{x}+t_{\alpha/2}(s/n^{0.5})]$**
$p_i \sim$ iid $(p, pq/n)$	Large	p Known	$\bar{p}-p/(pq/n)^{0.5} \sim z$	$[\bar{p}-z_{\alpha/2}\sigma(\bar{p}), \bar{p}+z_{\alpha/2}\sigma(\bar{p})]$
	Small	p Unknown	Independent + np>10 + n(1-p)>10	
x_i ~iid $N(\mu, \sigma^2)$	Any	μ Unknown	$(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$	$[(n-1)s^2/\chi^2_2 , (n-1)s^2/\chi^2_1]$

Practical advice: If the population is normally distributed N (μ, σ^2), the confidence interval built, given an unknown σ^2, is exact and can be used for any sample size. If the population is not normal, the CI will be approximate. In such case, the quality of approximation depends on both the distribution of the population and the sample size. In most applications, n>=30 is adequate when using the CI to develop an interval estimate of a population mean. If the population distribution is highly skewed or includes outliers, larger sample sizes are necessary.

III. Simulation or Bootstrap Methods for Constructing Sampling Distribution and Confidence Interval
Simulation (subsample random sampling) vs Bootstrap (full sample random sampling with replacement)

Lecture 7: Hypotheses Testing I

I. Basic Question and Decision Rule

1. Does the sample estimate provide strong enough evidence against its population parameter?
2. Examples: 1) Is the population mean equal to a specific value? 2) Is the mean-difference large?
3. Types: 1) One-sided (strictly positive/negative outcomes); 2) Two-sided (non-zero outcomes)
4. The decision rule (reject or not reject) in statistical testing is comparing the test statistic to a critical value of a chosen significance level in the distribution under the null hypothesis.

II. Elements of Hypothesis Testing

1. **Null hypothesis** H_0 (population true value or region) versus **Alternative H_A** (the complement)
2. A **test statistic** is a measure of the distance of the estimated value of the statistics from what its true value would be if the null hypothesis were true. Example: standardized sample mean/prop.
3. **Significance level α** (preset) is the max prob of a false positive being tolerated given H_0 is true
4. **Critical value** corresponds to a pre-chosen significance level **α** under the null H_0 distribution

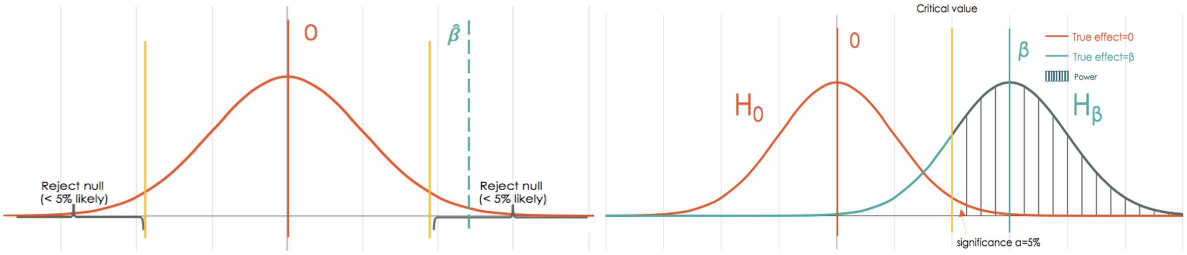
III. HT (Significance Test) Procedures

1. Establish the null and alternative hypotheses (mutually exclusive)
2. Specify the test statistic (z, t, F, χ^2) and its distribution under H_0
3. Choose a significance level α (rejection region) and critical value
4. Calculate the test statistic (or corresponding p-value) from the data
5. Conclude the test (reject or not reject) by the test statistic and the critical value

The **p-value** informs us of the probability of a false positive (or false rejection) in sample estimation under H_0 (is it equivalent to the size of the test?). The p-value for a test is the smallest significance level at which we can reject the null hypothesis given the value of the test statistics in the sample. Smaller p-value leads to stronger evidence against the null (negative or no effect). **Decision: Reject H_0 if the p-value=<α.**

IV. HT Errors: 1) Type I error is FP (FNR=**α** under H_0); 2) Type II error is FN (FNR=**β** under H_A)

HT Decision	Null hypothesis H_0	Alternative hypothesis H_A
Fail to reject H_0	True Negative TN (Correct)	False Negative FN **(FNR=β)**
Reject H_0	False Positive FP **(FPR=α)**	True Positive TP (Correct)

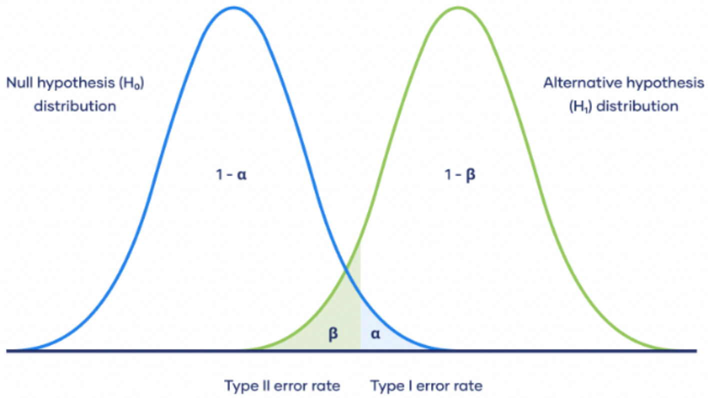


Source: J-PAL. Question: In the early stage of COVID-19 testing, which type of error must policy makers prioritize?

1. **The size of the test,** under H_0 (θ=0, negative), is the probability of a false positive decision.
2. **The power of the test**, given H_A (θ≠0, positive), is the probability of rejecting a false negative.
3. **The level of significance** (pre-chosen **α**) is the max prob of a false positive we can tolerate. A smaller level of significance (larger critical value) will make it harder to reject the null.
4. **What is the relationship between Type-I and Type-II errors? What affect the power of a test?** https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Lecture 8: Hypotheses Testing II

I. Types of Errors in Statistical Testing and Decision Making



https://www.scribbr.com/statistics/type-i-and-type-ii-errors/

Type I error rate is α
P(Type I error)=P(Reject H_0 | H_0 is true)

Type II error rate is β.
P(Type II error)=P(Fail to reject H_0 | H_A)

There is a trade-off between α and β: increase one will decrease the other.

For statisticians, a Type I error is usually worse. In practical terms, however, either type of error could be worse depending on your research context. $n>(z_\alpha+z_\beta)^2\sigma^2/(\mu_0-\mu_A)^2$

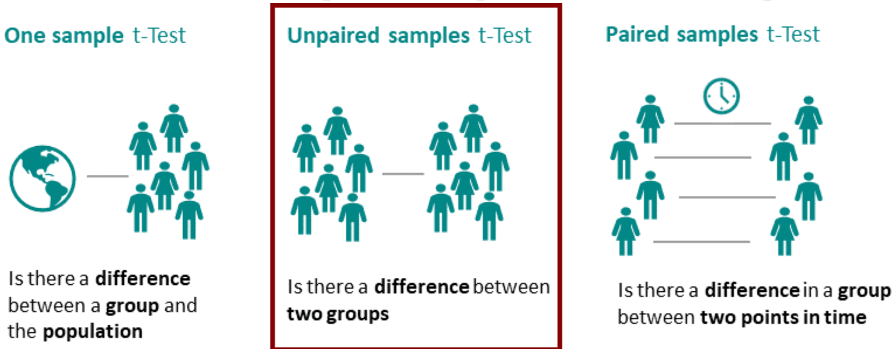
Principle: Establish the null hypothesis in a way that it can be straightforward to reject.

II. Hypotheses Testing for the Difference in Means H_0: $\mu_1-\mu_2=d_0$

Distribution $\bar{x}_1$ $\bar{x}_2$	n_1 n_2	σ_1 σ_2	Test Statistic under H_0
Normal	any	known	$z=(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)/\sigma(\bar{x}_1-\bar{x}_2)$ $\sigma(\bar{x}_1-\bar{x}_2)=(\sigma_1^2/n_1+\sigma_2^2/n_2)^{0.5}$
Normal	large	unknown	$z=(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)/\sigma(\bar{x}_1-\bar{x}_2)$ $\sigma(\bar{x}_1-\bar{x}_2)=(s_1^2/n_1+ s_2^2/n_2)^{0.5}$
Normal	small	unknown $\sigma_1=\sigma_2$	$t=(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)/s(\bar{x}_1-\bar{x}_2)$ $s(\bar{x}_1-\bar{x}_2)=(s_1^2/n_1+ s_2^2/n_2)^{0.5}$
Normal, paired $d=x_1-x_2$	any $n_1=n_2$	unknown	$t(n-1)=(dbar-d_0)/(s/n^{0.5})$ $s_d=(1/n-1)(\Sigma d_i-E(d))^{0.5}$

$(s_1^2/n_1+s_2^2/n_2)^2 \div (s_1^2/n_1)^2/(n_1-1) +(s_2^2/n_2)^2/(n_2-1)$, dof~$n_1+n_2-2$

III. Variants of t-test: Unpaired (independent) vs Paired (dependent)

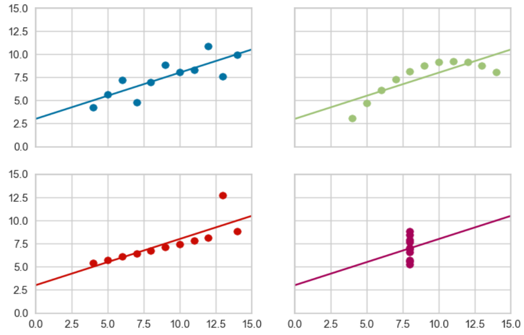

One sample t-Test
Is there a difference between a group and the population
Unpaired samples t-Test
Is there a difference between two groups
Paired samples t-Test
Is there a difference in a group between two points in time

https://datatab.net/tutorial/paired-t-test

Lecture 9 Linear Regression Models

I. Simple Regression Model

1. Definition: Regression analysis is a statistical method that uncovers the average value of Y conditional on X. (Regression model is conditional mean E[Y|X], a linear function of X.)
2. Function: 1) estimate and quantify the relationship between Y (explained or dependent) and X (explanatory or independent); 2) to forecast (or predict) new Ys out of sample in decision making
3. Model specification for cross-sectional data: Y_i= a+bX_i+e_i=E[Y|X]+e_i where $e_i \sim$ NIID(0, σ^2)
4. Estimation method and formula **E[Y]≠E[Y|X] & E[e|X]=0**
    1) Model parameters: a is the intercept coefficient, b is the slope coefficient
    2) Regression line is E[Y|X]=a+bX, since E[e|X]=E[e]=0 (e, X uncorrelated)
    3) Random error/disturbance/shock e_i is a white noise E(e_i)=0 and V(e_i)=σ^2
    4) Ordinary least square (OLS) estimation: mini SSE=$\sum e_i^2$=$\sum[Y_i$–E($Y_i|X_i$)$]^2$
    5) OLS formula for the estimators: b=C(X, Y)/V(X) and a=E(Y)–b•E(X)
    6) Unbiased estimator of σ^2: **s^2=$\sum(\boldsymbol{y_i - \hat{y}_i})^2$ /$(\boldsymbol{n-2})$=SSE/(n-2)=MSE**
5. Classical OLS assumptions and BLUE properties
    1) Linearity, E(e|X)=0, NIID (homoskedasticity, no autocorrelation), no multicollinearity
    2) Provided OLS assumptions, OLS is the Best Linear Unbiased Estimator B.L.U.E.



https://www.scikit-yb.org/en/latest/api/anscombe.html

- Among all possibilities, is there a unique line fits exactly all data points?
- How to find a line or curve that can best fit all the data points in a scatter plot?
- What does it mean by best? Is there a standard/measure?
- By minimizing the total sum of squared errors, the OLS method provides a simple yet "best" solution.
- Linear regression approximates the average slope of the nonlinear pattern

II. Simple Regression Diagnosis

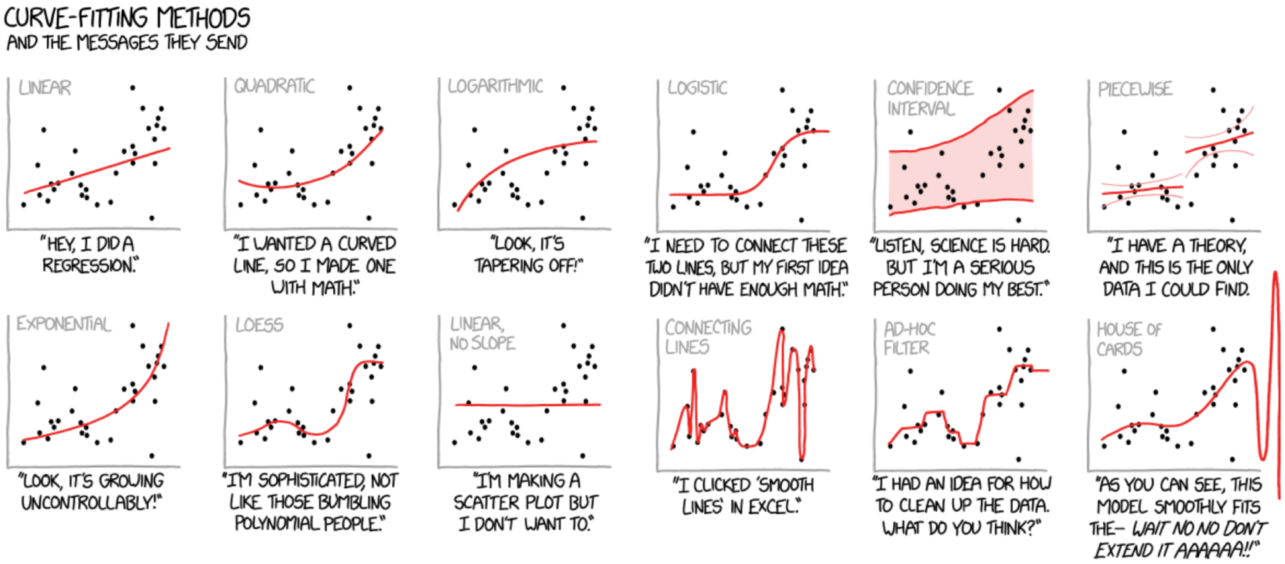
1. Measure of the goodness-of-fit: **R-square**
    1) How much variation in Y can be explained by X?
    2) Variance decomposition: Var(Y)=Var(a+bX)+Var(e), since Cov=0←E(e|X)=0
    3) Coef of determination **R^2=V(a+bX)/V(Y)=1–V(e)/V(Y)=SSR/SST=1–SSE/SST**. R^2=**ρ^2**
    4) **SST=SSR+SSE. SST $= \sum(\mathbf{y_i} - \bar{\mathbf{y}})^2$, SSR $= \sum(\hat{\mathbf{y}}_\mathbf{i} - \bar{\mathbf{y}})^2$, SSE $= \sum(\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i})^2$**
2. Residual test: can we find systematic patterns in the residual? Violation of NIID.
    1) Residual e_i=Y_i–E($Y_i|X_i$): the difference between Y_i and Y_i hat. Assumed NIID.
    2) In cross-section, plot e against X: is the variance constant over X? (homoskedasticity)
    3) In time series, plot e_t against $e_{t-1:}$ is the residuals autocorrelated over time? (autocorrelation)
3. Model predictions (interpolation vs extrapolation) and forecast errors (for model comparison)
    1) In-sample estimate E($Y_i|X_i$): given a specific X_i in the sample, what is the predicted Y_i?
    2) Out-of-sample prediction E($Y_i|X_j$)$_:$ for a new X_j not in the sample, what is the predicted Y_j?
    3) Prediction errors: MSE=RSS/(n–2)=$\sum[Y_i$-E($Y_i|X_i$)$]^2$/(n–2), MAE=$\sum|Y_i$-E($Y_i|X_i$)|/(n–2)

III. Simple Linear Regression: Statistical Inference

1. Standard errors of regression coefficient estimates: $SE(b)=(1/n^{0.5})\bullet SD(e)/SD(X)$
2. Robust standard error (heteroskedasticity-robust): The White-Huber “sandwich” formula
3. Confidence interval of regression coefficients: 95% $CI(b)=[b–1.96\bullet SE(b), b+1.96\bullet SE(b)]$ such that we can expect its true value to lie with 95% confidence in the interval estimated by the data
4. CI for $\hat{}y_i$: 95% $CI(\hat{}y_i)=[\hat{}y_i–1.96\bullet SE(\hat{}y_i), \hat{}y_i+1.96\bullet SE(\hat{}y_i)]$, $SE(\hat{}y)=SD(e)[1/n+(x_i– \bar{x})^2/n\sigma^2]^{0.5}$
5. Prediction interval for y_p: 95% $PI(y_p)=[y_p–1.96\bullet SPE(y_p), y_p+1.96\bullet SPE(y_p)]$
6. Significance test of regression coefficient H_0: b=0. Apply the t-test: $t(n-2)=b/SE(b)$
7. External validity of a regression analysis: Can we apply the sample result to a new context?
8. Observational data (invalid for causal inference) vs experimental data (controlled effect)

*IV. Regression Model Specification

1. Functional forms: linear vs nonlinear. Can we transform variables to linear regression?
2. Transformation of variables: in levels, in differences, in logs or semi-logs (interpretation vs fit)
3. Influential observations (“outliers”): In EDA, they should be excluded if errors
4. Measurement errors (wrong values, recording noise) → biased estimation
5. Why modeling nonlinearity? Points of the story (Helwig, 2021).
   1) Just because a model fits “good” doesn’t mean that it is a good model.
   2) Linear regression models may not reveal “true” relationships in the data
   3) More complex models can provide better in-sample fit than the linear model
   4) A “good” model survives out-of-sample test in forecasting (gold standard)
6. Nonlinear regression: polynomial; piecewise linear spline; generalized additive models (GAM)
   1) Polynomial regression captures a certain amount of curvature in a nonlinear relationship.
   2) Splines provide a way to smoothly interpolate between fixed points called knots.
   3) Piecewise linear spline produces connected line segments.
   4) GAM: a technique to automatically fit a spline regression.
7. Nonparametric regression (smoothing): local averaging; local regression; kernel regression.
8. Locally weighted scatterplot smoothing (lowess or loess): a smooth curve fit around a bin scatter.
9. Observation weighted regression (Weighted Least Square – WLS estimation)


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

https://therbootcamp.github.io/appliedML_2019Jan/_sessions/Fitting/Fitting_practical.html

V. Multiple Regression and Statistical Inference

1. Why include multiple variables in a linear regression? 1) omitted variables; 2) higher-order terms; 3) interaction between two variables (potentially for higher-order terms)
2. Multiple linear regression model: Y_i=a+b_1X_{1i}+ b_2X_{2i}+…+b_kX_{ki}+e_i where e_i~NIID(0, σ^2)
3. Significance tests (t) for individual coefficient estimate: H_0: a=0 or b_k=0. The t-statistic.
4. Joint significant test (F-test) for the regression: H_0: b_1=b_2…=b_k=0. The F-statistic.
5. Confidence interval for the coefficient estimates

*VI. Estimation Criteria (the true population parameter is α and its corresponding estimator is a)

1. Unbiasedness: E(a)=α, the expectation of an estimator is equal to its population parameter
2. Consistency: V(a) → 0, as sample size increase to infinity, variance disappears
3. Efficiency: V(a)=<V(a'), smaller variance among competing estimators

Note that being unbiased is a precondition for an estimator to be consistent.

Lecture 10: Probabilistic Regressions

Central Question: How can one estimate and forecast the probability of an event, based on other information related to the event?

I. Linear Probability Model

1. Suppose Y is a binary label to document the occurrence of an event. Y_0=0 or Y_1=1
2. Unconditional probability E(Y)=Y_0P(Y=Y_0)+Y_1P(Y=Y_1)=P(Y=1)
3. Conditional probability E(Y|X)= Y_0P(Y=Y_0|X)+Y_1P(Y=Y_1|X)=P(Y=1|X)
4. LPM: YP_i=P(Y=1|X)=α+β_1X_{1i} + β_2X_{2i}+…+e_i where e_i~NIID(0, σ^2)
5. Disadvantage: The fitted values/line of YP_i can go out of probability bound [0, 1]
6. Solution: change the linearity to a probability function, as in the logit and probit

II. Logit and Probit Models

1. The Logit: YP_i=P(Y=1|X)=G(X)=exp(α+β_1X_{1i} + β_2X_{2i}+…)/1+ exp(α+β_1X_{1i} + β_2X_{2i}+…)
2. The Probit: YP_i=P(Y=1|X)=F(X)=Prob(Z=<α+β_1X_{1i} + β_2X_{2i}+…) where Z ~ iid N(0, 1)
3. Both G(.) and F(.) are cumulative probability distribution functions (nondecreasing up to one)
4. Logistic regression does not assume the residuals are normally distributed nor constant variance
5. Estimation method: maximum likelihood numerical optimization
6. The predicted YP= Prob(Z=a+b_1X_{1i} + b_2X_{2i}+…+b_kX_{ki}<z)

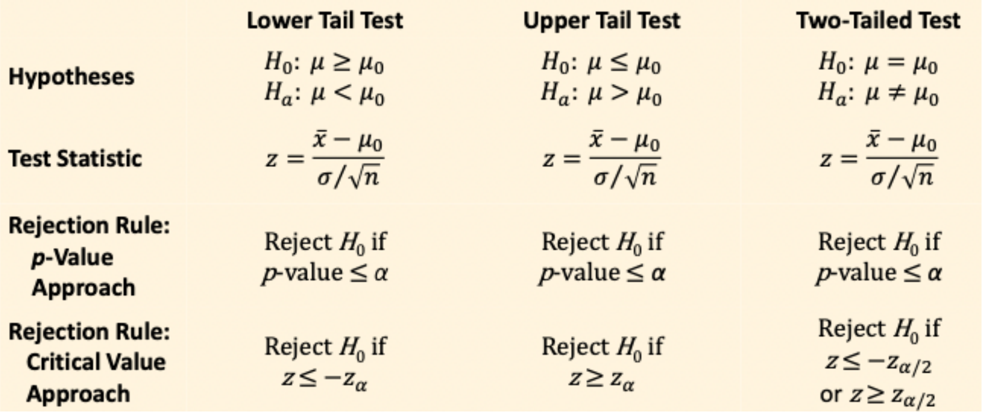
III. Statistical Inference and Model Evaluation

1. Peuso-R^2 (McFadden) goodness of fit: 1-lnLM_1/lnLM_0 where LM_0 is the intercept-only model
2. Residual assessment: the deviance residual is useful for checking if individuals are not well fit
3. Likelihood ratio (LR) test: whether the observed difference in model fit is statistically significant
4. Prediction errors, classification errors, confusion matrix, receiver operating characteristics (ROC)

Reference: Logistic Regression https://uc-r.github.io/logistic_regression

Appendix

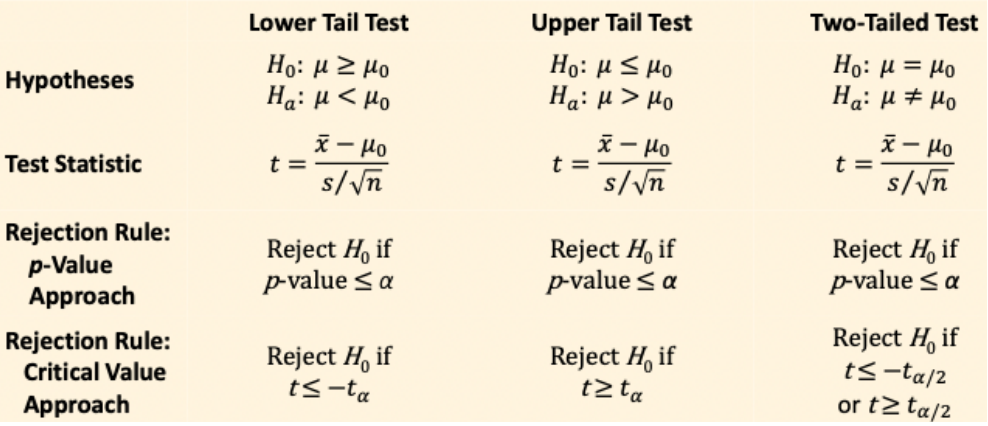
I. Hypothesis Testing about a Population Mean via z distribution



	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0\colon \mu \geq \mu_0$ $H_a\colon \mu < \mu_0$	$H_0\colon \mu \leq \mu_0$ $H_a\colon \mu > \mu_0$	$H_0\colon \mu = \mu_0$ $H_a\colon \mu \neq \mu_0$
Test Statistic	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \leq -z_\alpha$	Reject H_0 if $z \geq z_\alpha$	Reject H_0 if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$

Source: Camm et al. (2024) CH9

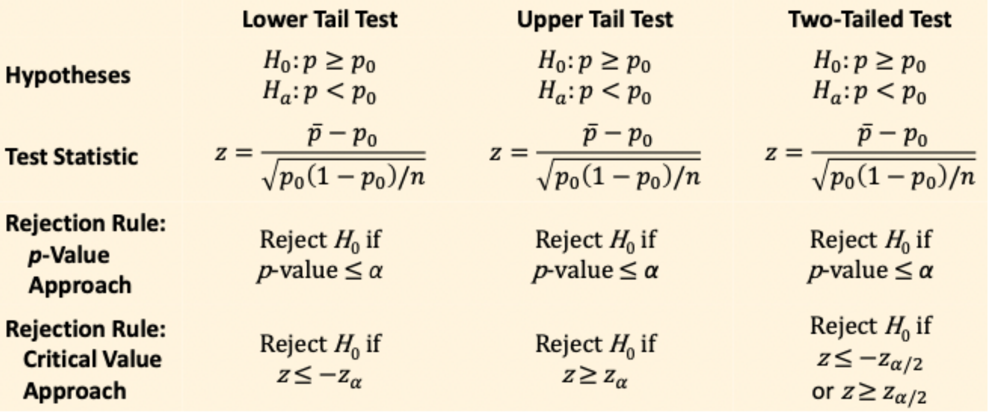
III. Hypothesis Testing about a Population Mean via t distribution



	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0\colon \mu \geq \mu_0$ $H_a\colon \mu < \mu_0$	$H_0\colon \mu \leq \mu_0$ $H_a\colon \mu > \mu_0$	$H_0\colon \mu = \mu_0$ $H_a\colon \mu \neq \mu_0$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $t \leq -t_\alpha$	Reject H_0 if $t \geq t_\alpha$	Reject H_0 if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

Source: Camm et al. (2024) CH9

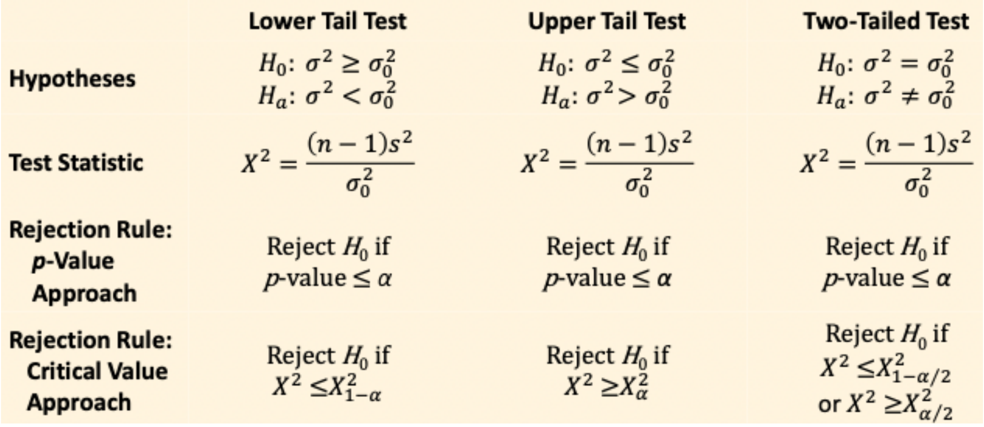
III. Hypothesis Testing about a Population Proportion via z distribution



	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0\colon p \geq p_0$ $H_a\colon p < p_0$	$H_0\colon p \geq p_0$ $H_a\colon p < p_0$	$H_0\colon p \geq p_0$ $H_a\colon p < p_0$
Test Statistic	$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$	$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$	$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $z \leq -z_\alpha$	Reject H_0 if $z \geq z_\alpha$	Reject H_0 if $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$

Source: Camm et al. (2024) CH9

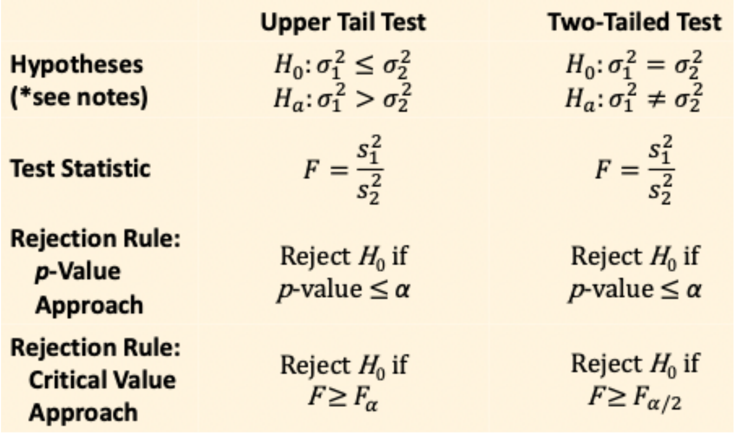
IV. Hypothesis Testing about a Population Variance via χ^2 distribution



	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0\colon \sigma^2 \geq \sigma_0^2$ $H_a\colon \sigma^2 < \sigma_0^2$	$H_0\colon \sigma^2 \leq \sigma_0^2$ $H_a\colon \sigma^2 > \sigma_0^2$	$H_0\colon \sigma^2 = \sigma_0^2$ $H_a\colon \sigma^2 \neq \sigma_0^2$
Test Statistic	$X^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$X^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$X^2 = \frac{(n-1)s^2}{\sigma_0^2}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $X^2 \leq X_{1-\alpha}^2$	Reject H_0 if $X^2 \geq X_{\alpha}^2$	Reject H_0 if $X^2 \leq X_{1-\alpha/2}^2$ or $X^2 \geq X_{\alpha/2}^2$

Source: Camm et al. (2024) CH11

V. HT about two Population Variances via F distribution



	Upper Tail Test	Two-Tailed Test
Hypotheses (*see notes)	$H_0\colon \sigma_1^2 \leq \sigma_2^2$ $H_a\colon \sigma_1^2 > \sigma_2^2$	$H_0\colon \sigma_1^2 = \sigma_2^2$ $H_a\colon \sigma_1^2 \neq \sigma_2^2$
Test Statistic	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{s_1^2}{s_2^2}$
Rejection Rule: *p*-Value Approach	Reject H_0 if p-value $\leq \alpha$	Reject H_0 if p-value $\leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $F \geq F_\alpha$	Reject H_0 if $F \geq F_{\alpha/2}$

Source: Camm et al. (2024) CH11

Lecture

$E(\bar{x})=\mu$ and $E(\bar{p})=p$ Statistical inference is just not very important with Big Data.

Bootstrap is a resampling method for generating the entire sampling distribution and its statistics.

1) In practice, we only have one dataset (all the data we have) and we would like to uncover the sampling distribution of samples of the same size to that data.
2) To estimate standard error of the statistic, we compute the standard deviation of the estimates of the statistics across available bootstrap samples of the same size.
3) "Pull yourself out of the swamp by your bootstraps." How is that possible?

Power-law (Pareto): distributions with very large extreme values (the filthy rich, mega city size)

Sampling Methods (Reference: IMS, CH2)

1. Simple random sampling: Almost all statistical methods are based on the notion of implied randomness. If data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable.
2. Stratified sampling: The population is divided into groups called strata. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum.
3. Cluster sampling: break up the population into many groups, called clusters. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample.
4. Multistage sampling: rather than keeping all observations in each cluster (a cluster sample), collect a random sample within each selected cluster.

Given a sample of I.I.D. observations from some distribution, inferential statistical analyses are concerned with inferring properties about the population from which the sample was collected.
Independent Identically Distributed RVs (I.I.D.)/Samples

1. Independence: $f(x, y; \theta)=f_x(x; \theta_x)f_y(y; \theta_y)\ \forall(x, y)$ where the marginal distributions can be different
2. Identical distribution: $f(x; \theta_x)\equiv f(y; \theta_y)\ \forall(x, y)$ where the equality sign "≡" indicates all the marginal distributions have the same functional form $f_x(.)=f_y(.)$ and the same unknown parameters $\theta_x=\theta_y=\theta$

Comparison and Correlation

1. Relations between two variables: 1) independent; 2) associated/correlated/dependent; 3) causal
2. Correlation and dependence
   1) Covariance and correlation: measure linear relevance
   2) Independence: joint density = marginal density multiplication
   3) Relation: independence → zero correlation, but not vice versa.

Statistical Applets http://digitalfirst.bfwpub.com/stats_applet/asset/applet_index.html

https://www.rapidtables.com/math/symbols/Basic_Math_Symbols.html

Uncertainty is the nature of the world that we cannot change in making decisions, but being able to measure and deal with it is the way out of ignorance.