BST 210
Applied Regression Analysis

- Welcome -

Lecture 1
Plan for Today

1. Motivations, Examples, History, and Important People
2. Mathematical Development and Generalizations
3. Key Review Concepts
4. Course Preview and Details
5. Break – 4 minutes
6. Introduction to Linear Regression and related methods (with more review of topics as we go)

Why are we interested?

"Humans are constantly sorting the world into categories, predicting how things work, and testing those predictions – the essence of science."

- New Scientist magazine

Regression = Crystal ball ?

How good are we at predicting? At finding associations? At testing?
How frequently are we wrong? How precise are our predictions?
How confident are we in them? How do our predictions vary?

Regression: the 'automobile' of statistical analysis

"… despite its limitations, occasional accidents, and incidental pollution, it and its numerous variations, extensions, and related conveyances carry the bulk of statistical analyses, and are known and valued by nearly all."

- Stephen M. Stigler, historian of statistics

Regression is ubiquitous!

AstraZeneca

29th August 2019

"AstraZeneca has announced that its potential new systemic lupus erythematosus (SLE) medicine, anifrolumab, has met its primary endpoint, achieving a statistically-significant and clinically-meaningful reduction in disease activity."

- Pharma Times

HEALTHCARE FINANCE

JAN 30, 2019

"Addressing SDOH (social determinants of health)? Consider artificial intelligence and machine learning.

The technology is still maturing, but already can be used to make sense of social determinants data -- a must in modern care delivery."

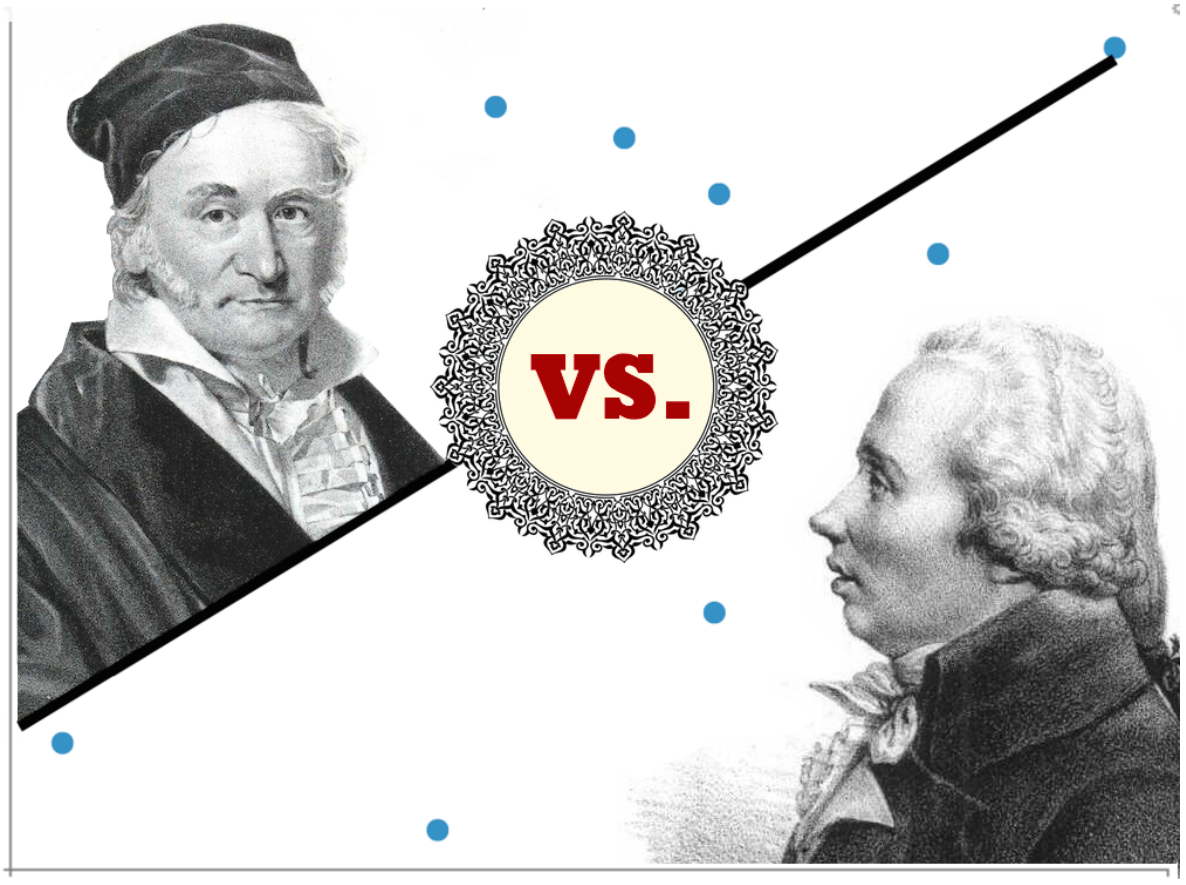
RAND CORPORATION

August 7, 2017

"Largest-Ever Study of Pets and Kids' Health Finds No Link; Findings Dispute Widely Held Beliefs About Positive Effects of Pet Ownership

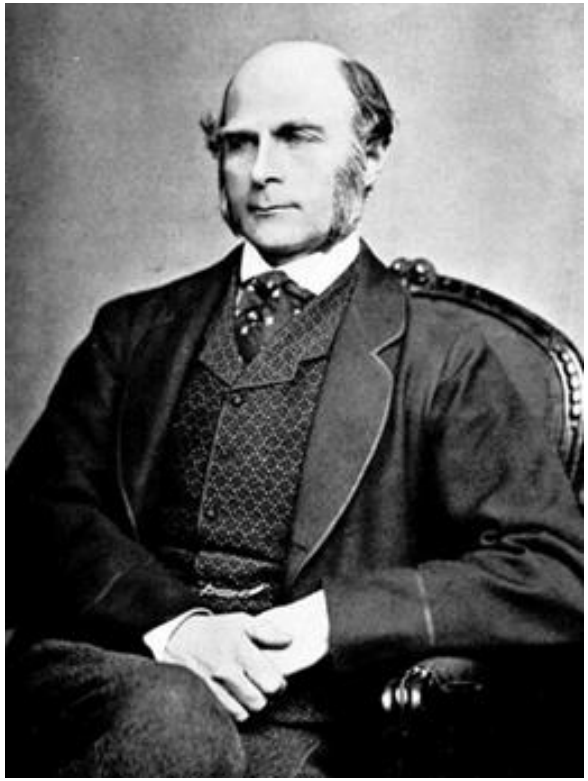
The new research brings advanced statistical tools such as double robust regression analyses to the study of this topic, which the scientists used to account for other factors that may influence a child's health rather than pet ownership, such as family income."

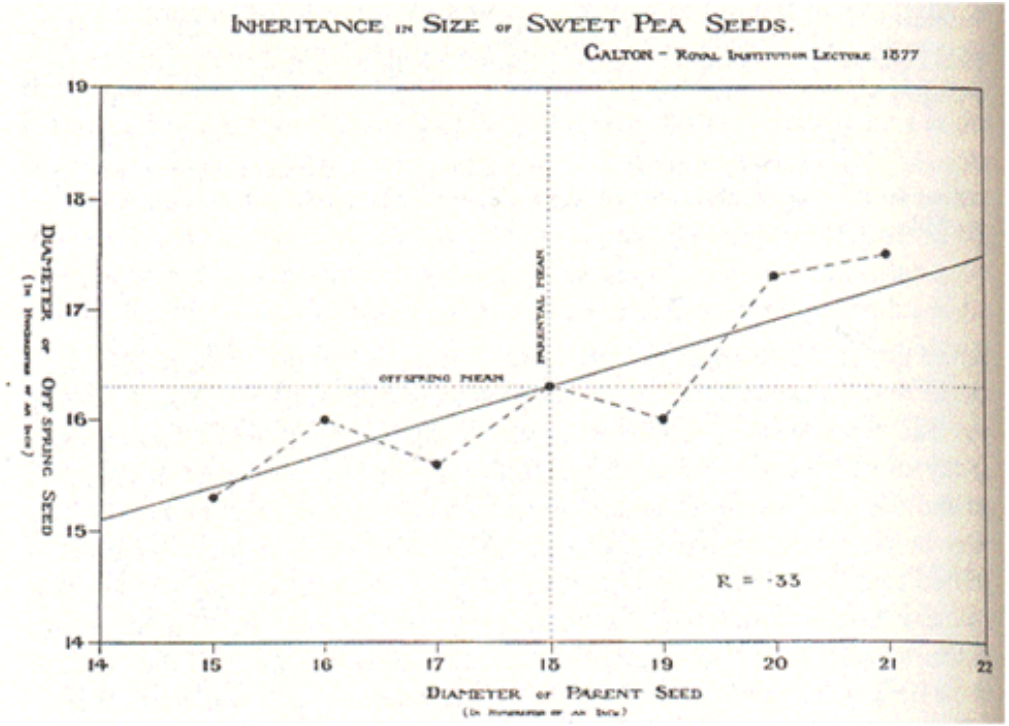
Gauss (1809) or Legendre (1805)?



Least Squares and planetary/comet movement

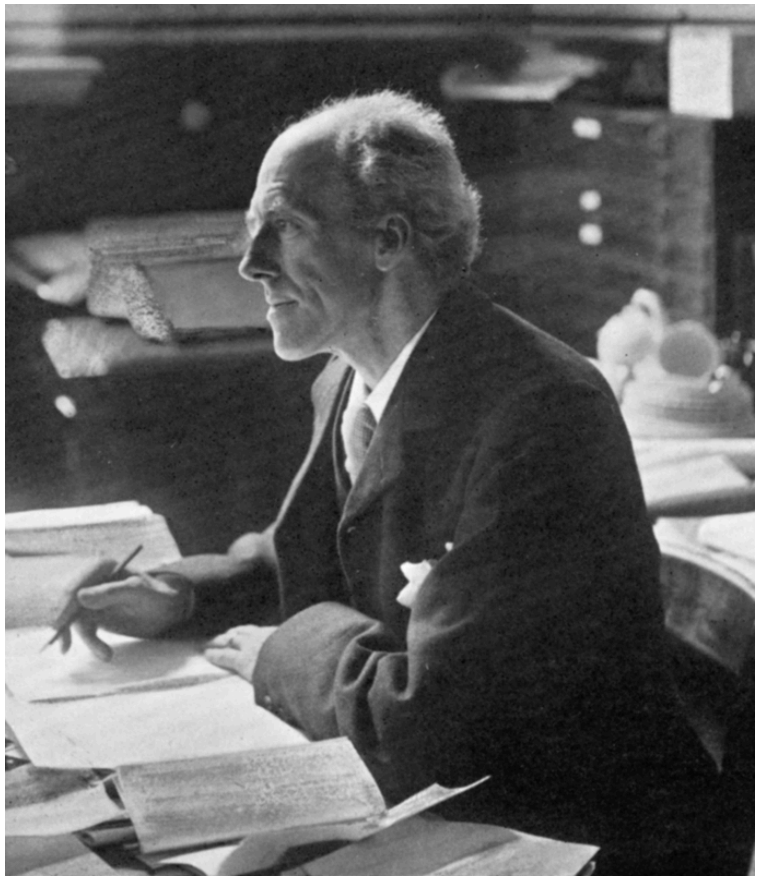
Galton, 1886





***Regression* Towards Mediocrity in Hereditary Stature**

Pearson (1901), then Fisher (1911)

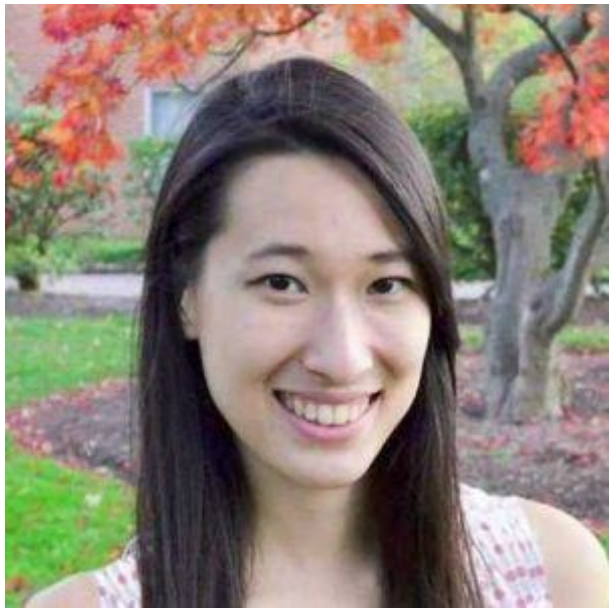


Regression Line with Least Squares

Culminating theory

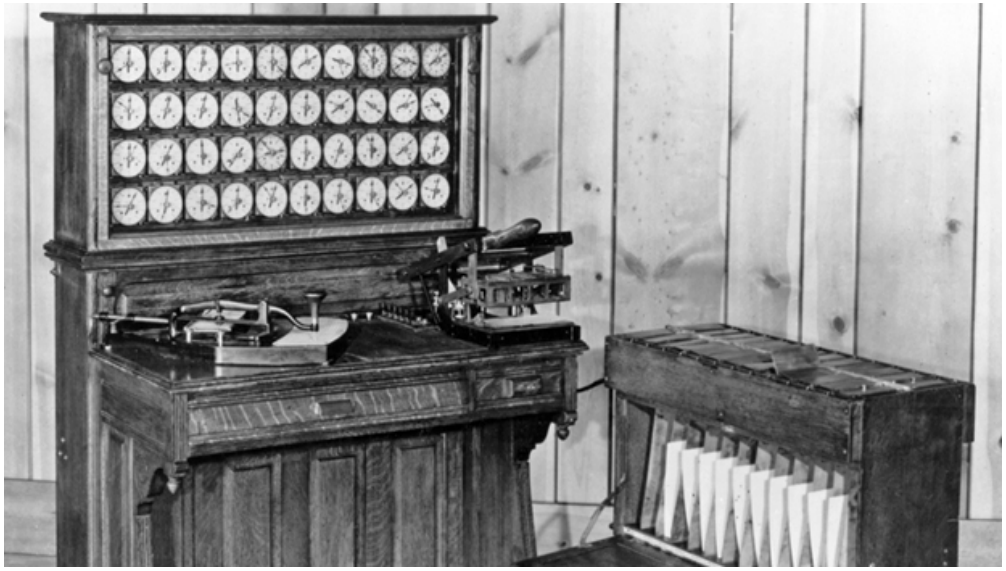
Important to regression in modern times…

our BST 210 TAs!



Beau, Izzy and Christina

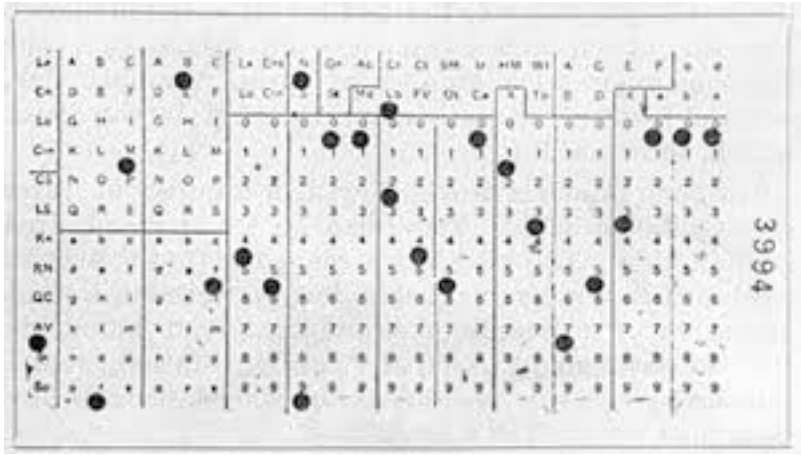
Computing



1920s IBM Punched card tabulator

Brought regression into mainstream

1970s - PCs



"Math is logic of certainty; probability is the logic of uncertainty ." - Joe Blitzstein

$$Y = mx + b \quad \rightarrow \quad Y_i = \alpha + \beta x_i + \varepsilon_i$$

where $E(Y_i|x_i) = \alpha + \beta x_i$

For this idea of 'regressing' one feature on another - adding real world *human variability introduces:*

-Variation
-Spread
-Multiple m's
-Multiple b's

Simple line is not enough – we need representation of this variability!

* Shown here only in general terms – stay tuned for details

In fact we can generalize this further -

$$Y_i = \alpha + \beta\, f(x_i) + \varepsilon_i$$

Where $f(x_i)$ can take many forms: x, x^2 , x^3, $\sin(x)$, so long as the equation is linear in α, β.

* Question: What must be true for ε_i and Y_i ?

Hint: remember Gauss!

Wala! Linear Regression

- *Objective:* model the expected value of a continuous variable, Y, as a linear function of the continuous predictor, X, $E(Y_i) = \beta_0 + \beta_1 x_i$
- *Model structure:* $Y_i = \beta_0 + \beta_1 x_i + e_i$
- *Model assumptions:* Y is is normally distributed, errors are normally distributed, $e_i \sim N(0, \sigma^2)$, and independent, and X is fixed, and constant variance σ^2.
- *Parameter estimates and interpretation:* $\hat{\beta}_0$ is estimate of β_0 or the intercept, and $\hat{\beta}_1$ is estimate of the slope, etc... Do you recall, what is the interpretation of the intercept and the slope?
- *Model fit:* R^2, residual analysis, F-statistic
- *Model selection:* From a plethora of possible predictors, which variables to include?

And we can generalize even further -

$$g(Y_i) = \alpha + \beta\, f(x_i) + \varepsilon_i$$

Where link function g(*) can take many forms:

g(*)	Distribution
- linear	Normal
- logit	Binomial
- log	Poisson
- generalized logit	multinomial

Wala! Generalized Linear Models (GLM)

(…and there are further extensions and links to survival modeling!)

- The data $Y_1, Y_2, ..., Y_n$ are independently distributed, i.e., cases are independent.
- The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,...)
- GLM does NOT assume a linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the transformed response in terms of the link function and the explanatory variables; e.g., for binary logistic regression $logit(\pi) = \beta_0 + \beta X$.
- Independent (explanatory) variables can be even the power terms or some other nonlinear transformations of the original independent variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure, and *overdispersion* (when the observed variance is larger than what the model assumes) maybe present.
- Errors need to be independent but NOT normally distributed.

Which ‘regression’ approach we use depends on the underlying *data generating process*

- Normal (Gaussian)
- Binomial (from Bernoulli)
- Poisson
- Exponential
- Weibull
- Gamma

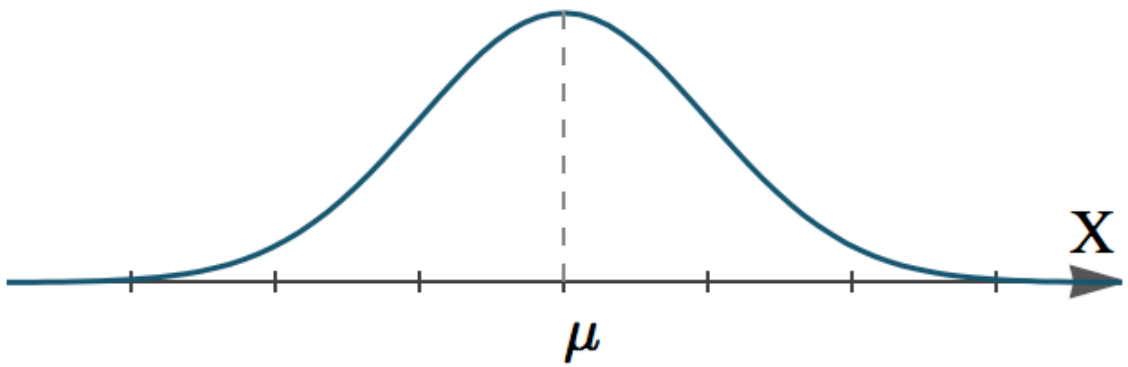
We've now motivated some main themes for this course.

Let's now review several relevant distributions, and elements of hypothesis testing.

Normal (Gaussian) Distribution (pdf)

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

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



Mean: $E[X] = \int_{-\infty}^{\infty} x f(x) dx = \mu$

Variance: $V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \sigma^2$

Beautiful!

It’s the most important continuous distribution -

- Good theoretical properties (sums, differences, and averages of normal X are normally distributed), easy to work with mathematically.
- Closely related to many other distributions (binomial and Poisson distributions are approximately normal in certain circumstances).
- Many naturally occurring variables roughly follow a normal distribution (body weights and heights, blood pressure--even measurement errors).
- Many variables that are right (positively) skewed can be transformed, usually by logs, to normal X (plasma concentrations, cholesterol, blood lead levels, etc.).
- Connection between the size of a sample N and the extent to which a sampling distribution approaches normality. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal (Central Limit Theorem).

Binomial Distribution (pmf - discrete)

$X \sim Bin\ (n, p)$:

- Sum of *n* independent Bernoulli X (0 or 1) trials, all with same probability of success, P(X=1) = p.
- Helps us make inferences about *proportions*. We can think of proportions as averages for Bernoulli.
- X: Counts number of units in a sample of size *n* with a characteristic of interest.
- *p*: probability of observing the characteristic in an individual unit
  q: 1-p, the probability of not observing the characteristic
- Assumptions: A sample of *n* units is drawn from an infinite population *without replacement*; each unit in the population has the same probability *p* of having the characteristic of interest.
- Sample space: {0, 1, 2, ..., n }
  Parameters: *n* = number of trials, *p* = P(success) = P (X = 1)

Binomial Distribution (pmf - discrete)

X ~ Bin (n, p)

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,\ldots,n \quad 0 < p < 1$$

$\binom{n}{x}$ = number of ways of choosing x objects from a total of n without regard to order

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

$x! = 1(2)(3)\ldots(x) \qquad 0! \equiv 1$

Poisson Distribution (pmf-rare events)

$$X \sim \text{Poisson}(\mu)$$

- The Poisson distribution is used for modeling counts of relatively *rare* events – number of cases of cancer in Colorado in 2005, radioactive counts per unit time, number of plankton per aliquot of seawater, ... events occur independently within an interval and between intervals.

- Unlike the Binomial distribution, no upper bound is assumed for the sample space, so the counts are assumed to occur in infinite (or large) populations.

- This distribution helps us to make inferences about *rates*. Rates are usually average occurrences of events over time or of objects over space, such as the number of events per person- time, the density of objects in a specified area, number accidents in Boston in one hour, number of bacterial colonies on an agar culture dish.

- In contrast to the binomial distribution, the Poisson distribution is defined by one parameter, μ, where $\mu > 0$.

Poisson Distribution (pmf-rare events)

$$X \sim \text{Poisson}(\mu)$$

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!} \qquad x = 01, 2, ... \qquad \mu = \lambda t \text{ or } \lambda A$$

- The probability of *x* events occurring (in a time period of length *t* or in an area *A*) with parameter μ is the probability mass function above.
- λ = the rate of occurrence of events per unit time, or the density of objects per unit area or unit time.
- For a period of time *t* or an area *A*, we can interpret λt or λA as the *average* or *expected number* of events or objects for the Poisson distribution, μ.

Hypothesis testing

- **A *hypothesis* is a claim or statement about a population parameter (or parameters).**
- **A *hypothesis test* is a statistical method of quantifying evidence (using sample information) to reach a decision about a hypothesis.**

- 1) State a **"null hypothesis": Ho** This is a claim that is initially assumed to be true. The wording usually states: "There is *no change* between ...," "... no difference...", "... no effect of ...", "...no association ...". H0 is where we place the burden of proof for data—what we would actually like to *disprove*. Form of H0: population characteristic = hypothesized value; e.g. H0: $\mu = 90$ mm Hg

- 2) State an **opposing, or "alternative hypothesis": H1**. This statement contradicts H0; the null and alternative hypotheses cannot both be true. H1 is what we would like to prove to be true. Form of H1: can have the same form as H0 with > or < or $\neq$ in place of =; e.g. H0: $\mu \neq 90$ mm Hg

- 3) Collect data assuming H0 is true. Test that assumption--find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) Apply decision rule about the truth of H0. Recognize the role of chance in our decision-making.

Hypothesis testing

When we make a decision about H0 and H1 from the data, there are 4 possibilities:

1.H0 is true and we say it is true (1-α)

2.H0 is true and we say it's false (Type 1 error: convicting the innocent, α)

3.H0 is false and we don't say it's false (Type II error: letting guilty go free, β)

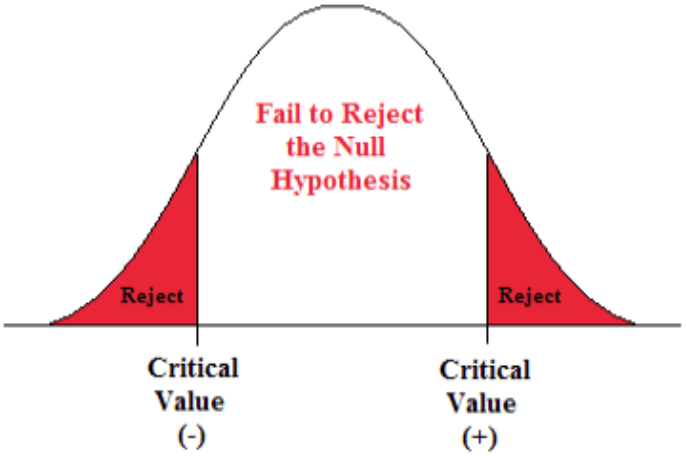
4.H0 is false and we say it's false (Power, 1-β)

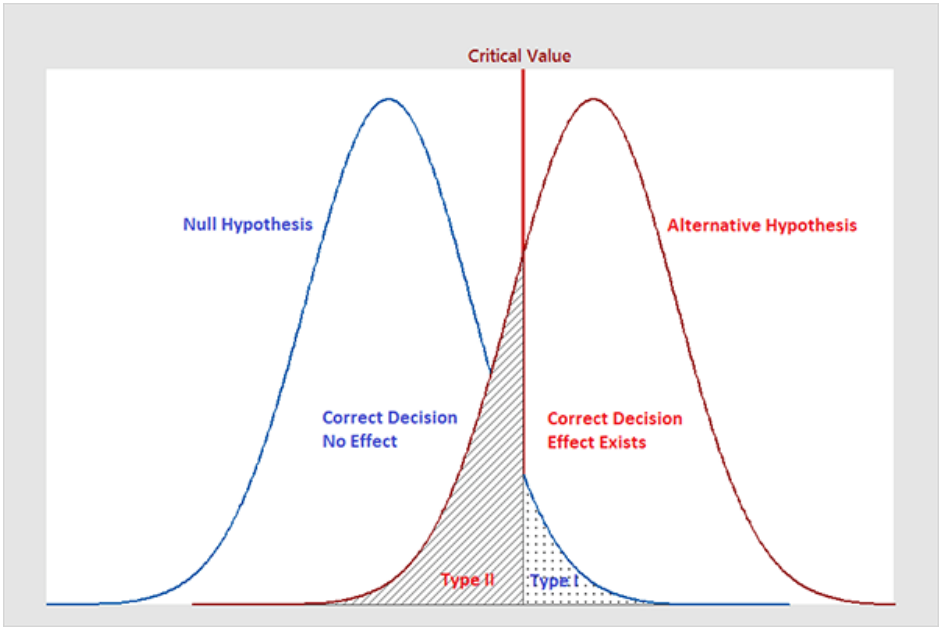
Type I error: Probability of rejecting the null hypothesis (H0) when it is true. "Reject H0 | H0 is true" - usually considered the more serious error

Type II error: Probability of failing to reject the null hypothesis when H0 is false – "Fail to Reject H0 | H0 is false"

	H_0 True	**H_0 False/H_1 True**
Accept/Fail to reject H_0	1. Correct Probability of correct decision = 1-α = Level of confidence	3. Type II Error P(Type II Error) =β
Reject H_0	2. Type I error P(Type I error) = α Level of significance	4. Correct Probability of correct decision = 1-β = Power

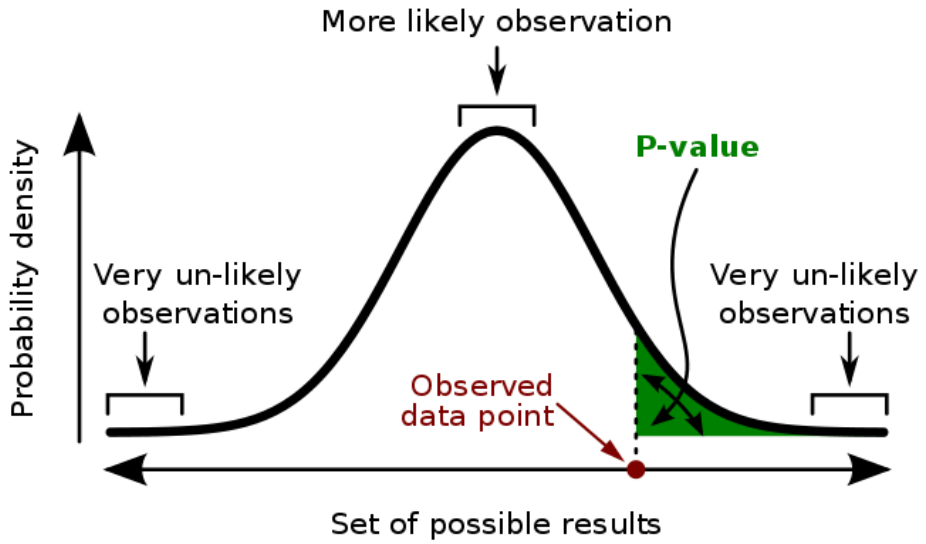
Hypothesis testing





Hypothesis testing

P-value

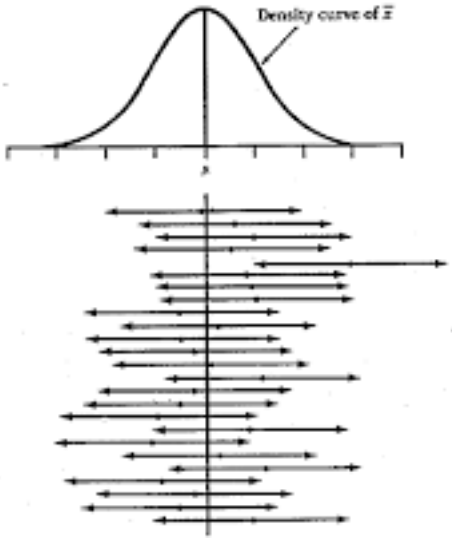


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Controversial!
Prioritize clinical and real world meaningfulness

Hypothesis testing

Confidence Interval



If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean.

Next, a Course Preview

This course will cover the major areas of regression analysis, including methods of analysis for continuous, binary, count, and survival outcomes.

- Descriptive and graphical statistics
- Linear regression, model building and model assessment
- Logistic regression, model building and model assessment, and extensions
- Poisson regression and extensions
- Survival curve estimation and Cox proportional hazards regression

Course Preview continued

We will include modern data analysis methods

- Generalized additive models, to assess linearity of effect of continuous covariates
- Modern methods for model selection
- Propensity scores
- Big picture ideas of data analysis and data presentation
- Ways to think about missing data

Course Details

- Course Design
  - Lectures develop/motivate ideas -> Labs provide practice
  - 7 Homeworks (no late hw, submitted via Canvas by midnight on Mondays), 2 Projects, 2 Exams
  - Extra credit options
  - Computing in R, SAS, Stata interchangeably

- Can ask questions via Canvas
- Refer to Syllabus

Introduction to Linear Regression (with continued review along the way)

Analyses involving linear regression may include consideration of:

- Types of data and descriptive statistics
- Inferential statistics (point estimates, confidence intervals, hypothesis testing)
- Correlation analysis, *t*-tests
- Simple linear regression

Nominal and Ordinal Data

Nominal data are put in *unordered* categories, and are called **binary** or **dichotomous** when there are two categories

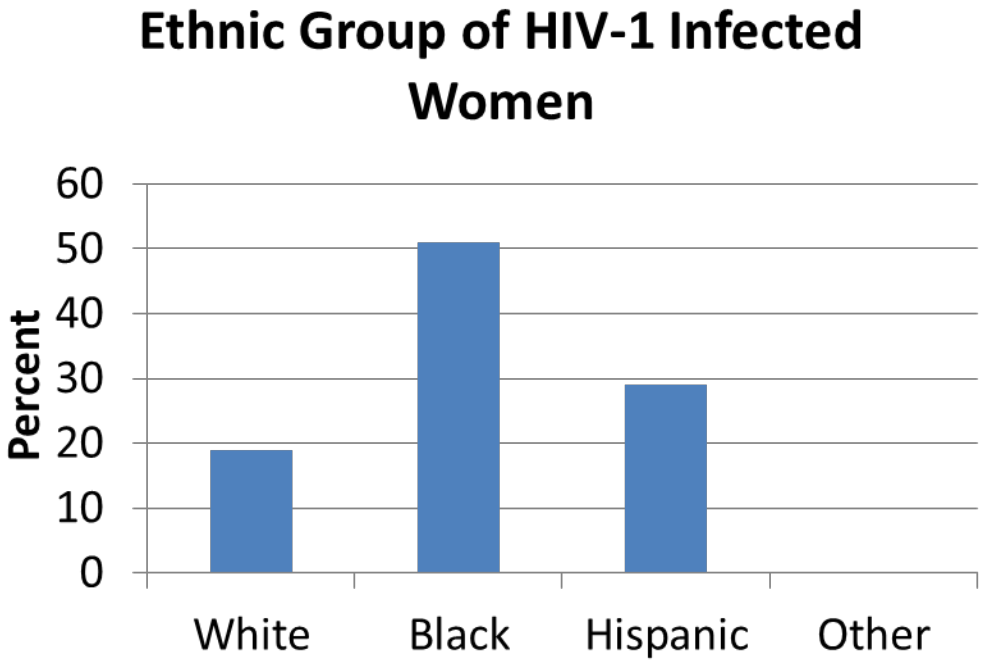
- Examples: Age group (Youth/Adult), Country (US, Canada, Mexico, Spain, France), Seizures/no seizures

Ordinal data are put in *ordered* categories

- Examples: Excellent/Good/Fair/Poor, A/B/C/F

Descriptive Statistics for Nominal or Ordinal Data

- Frequencies, %, proportions
- Often displayed as histograms or bar charts



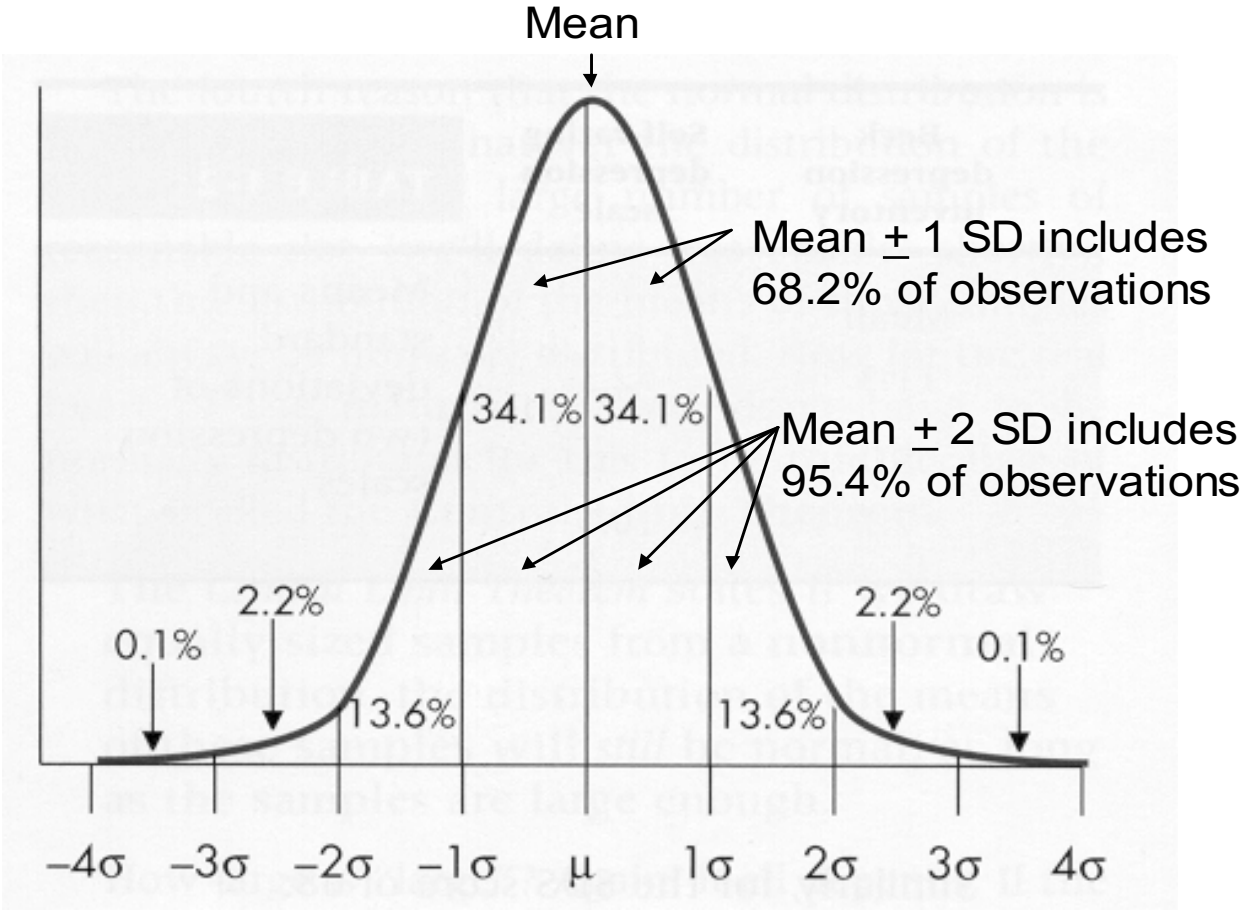
Continuous Data

- *Quantitative* observations
- Examples
  - Temperature (°C)
  - HIV viral load (copies of HIV per mL of blood)
  - IQ (intelligence quotient)
  - Age (really integer)

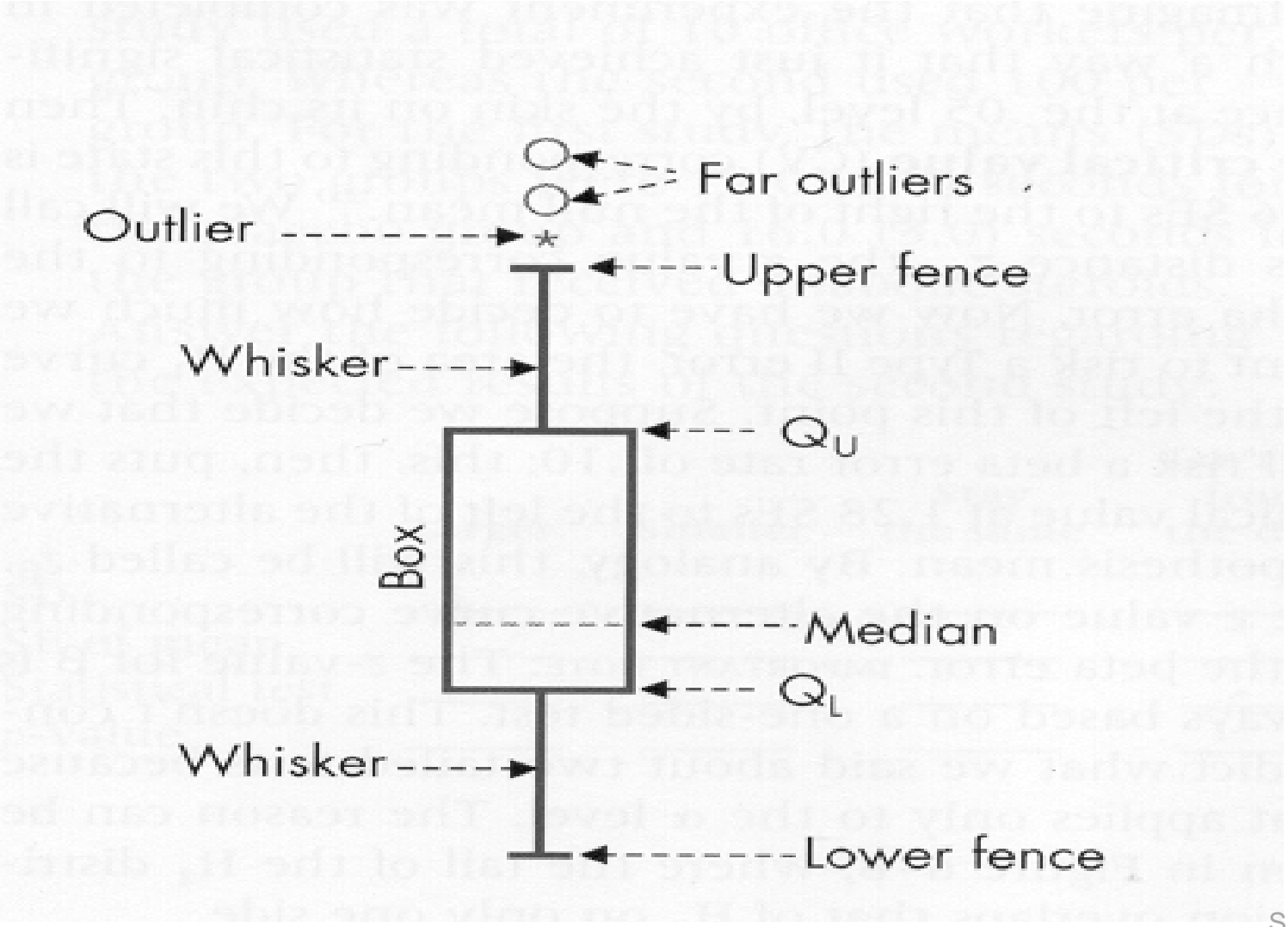
Descriptive Statistics for Continuous Data

- Measures of **central tendency** – mean (sample average), median (50^{th} percentile)
- Measures of **dispersion** – variance, standard deviation, range (minimum to maximum), interquartile range (25^{th} to 75^{th} percentiles)
- Often displayed as histograms or boxplots

The Normal Distribution



Boxplots


Far outliers
Outlier
Upper fence
Whisker
Q_U
Box
Median
Q_L
Whisker
Lower fence

Inferential Stats: Point Estimates

We determine a **point estimate** of the parameter of interest

- For example, the sample mean to be used to estimate the population mean, or fitted regression slope to be used to estimate a population regression slope
- The "best" estimates have **no or minimal bias** and **high precision** (small standard errors)
- Any point estimate is unlikely to equal the true population value

Inferential Stats: Confidence Intervals

- A **confidence interval** specifies the *range* of likely population values based on the sample
- Size of the confidence interval depends on the sample size, data variability, and level of confidence (often 95%)
- **Interpretation of 95% confidence**: In repeated sampling, 95% of confidence intervals include the true population value.

Hypothesis Testing

- Makes a “yes” or “no” decision about whether a **null hypothesis** is true, for example:
  - Is the mean IQ of a group of cardiac surgery infants = 100 (the average value for control infants)? $H_0 : \mu = 100$
  - Is the mean IQ of a group of cardiac infants assigned to one of two treatment groups the same? $H_0 : \mu_1 = \mu_2$

Hypothesis Testing: Errors

- **Type I error** (false positive): incorrectly concluding that there is a difference, when there really is no difference (alpha error)
- **Type II error** (false negative): incorrectly concluding there is no difference when there really is a difference (beta error)
- Often choose $\alpha = 0.05$, $\beta = 0.10$ or 0.20, need power calculations to determine sample size

P Values and What They Mean

- ***P* value** = chance of obtaining observed results as or more extreme results, *assuming that the null hypothesis is true*
  - Measures strength of the evidence taking chance into account
  - Is a poor indicator of the magnitude of the treatment effect
  - Is more of a 'threshold' piece of information
  - Is somewhat arbitrary, based on chosen alpha level (pvalue has become controversial recently)

When You See P < 0.05, What Questions Should You Ask?

- Although the result is statistically significant, is it clinically important?
- How many hypotheses were tested?
  - Testing multiple hypotheses or many subgroups invalidates P values unless they have been adjusted correctly (Bonferroni)
- Are the comparisons fair and unbiased?
- Have we controlled for potential confounders?

When You See P > 0.05, What Questions Should You Ask?

- Are there too few events occurring or too few patients studied to have a good chance to detect a real difference?
- Are there biases that could have obscured real treatment differences?
- Have we collected information on the right variables?
- Could there be measurement error?
- With lower N, should we try nonparametric test?

Confidence Intervals vs. *P* Values

- CI more informative than *P* values
  - CI provides likely range of treatment differences, not just a yes/no decision
- Hypothesis tests – can get hung up on *P*=0.05…what about *P*=0.049, *P*=0.051, *P*=0.06
  - CI avoids issue entirely

What is Regression?

- An approach to modeling the relationship between a **dependent variable** *Y* and one or more **independent variables** (or predictor variables or covariates) *X*
- **Linear regression** is used when the dependent variable *Y* is *continuous*
- **Logistic regression** is used when the dependent variable *Y* is *dichotomous*
- The covariates *X* can be categorical (nominal or ordinal) or continuous

Multiple techniques for continuous outcomes fall into Linear Regression category

- Many health outcomes take on continuous values, such as cholesterol level, malaria parasitemia, and IQ
- The analysis of continuous outcomes has a long history and uses such methods as correlations, *t*-tests, linear regression, and the analysis of variance

Next time:

- We'll delve into the these methods, and discuss when it's most appropriate to employ each one.
- We'll also discuss when and why we want to move toward a linear regression model in place of the other approaches,

…and more!