Probability Theory, Estimation, and Statistical Inference

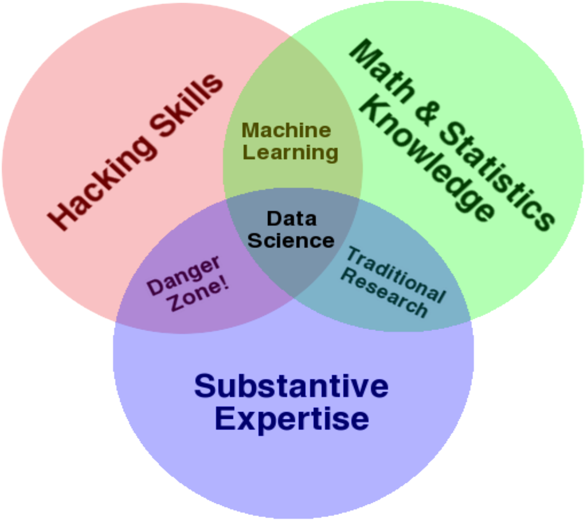
POLS GU4716

Columbia University

What is the reason for this overview?

Photo: Paramount Pictures/Sunset Boulevard/Corbis via Getty Images/Michael Ochs Archives/Getty Images; https://www.vulture.com/

What is data science?



Data Science Venn diagram (Source: Drew Conway)

Basic building blocks for political analysis

- Do not take a ride into the danger zone!
- Think of data in terms random variables (RVs)—stochastic: outcome of a chance experiment.
- Probability: what is likelihood of each realization of RV?
- Interested in correlation & conditional expectation: $E[Y|X]$.
- Estimate parameters employing statistical inference.
- Hypothesis testing.

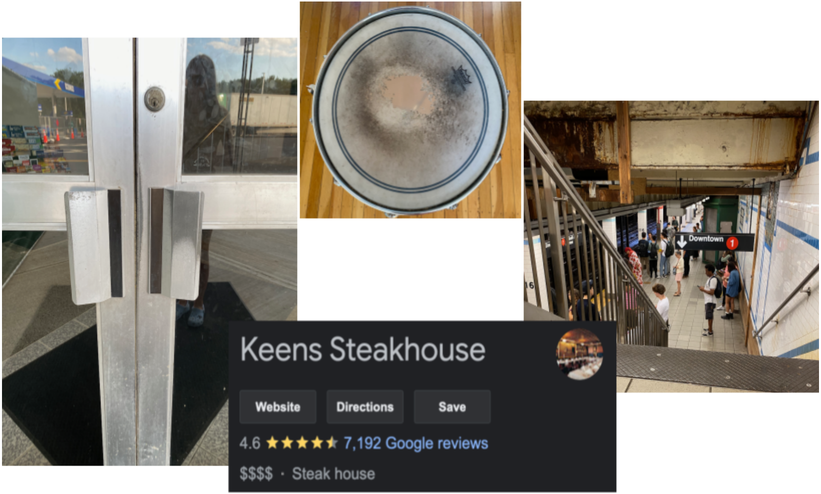
Probability theory basics

- $0 \leq P(A) \leq 1$ for every A.
- If A, B, C constitute an exhaustive set of events, then $P(A + B + C) = 1$.
- If A, B, C are mutually exclusive events, then $P(A + B + C) = P(A) + P(B) + P(C)$.

Random variables

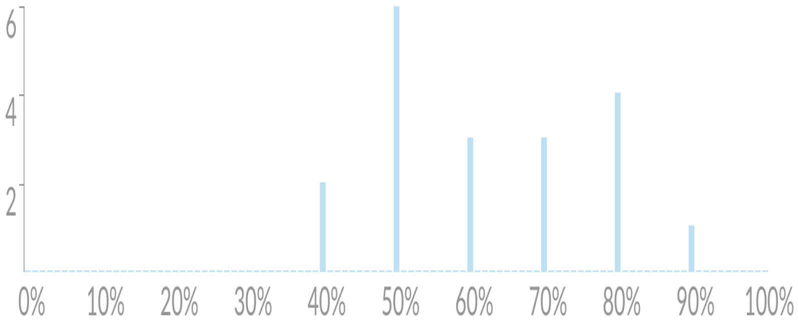
- Two types: **discrete** vs. **continuous**.
- Discrete: takes on a finite number of values
  - Example: self-placement on a 7 point ideological scale
- Continuous: takes on an uncountably infinite number of values—use it to approximate when many values are possible
  - Example: total amount of campaign expenditures
- Care about distributions, esp. how distributions of variables relate to each other and statistics that describe the behavior of distributions.

Distributions are part of everyday life



Distribution of correct answers on pre-test

Average Score	High Score	Low Score	Standard Deviation	Average Time
62%	90%	40%	1.47	04:10



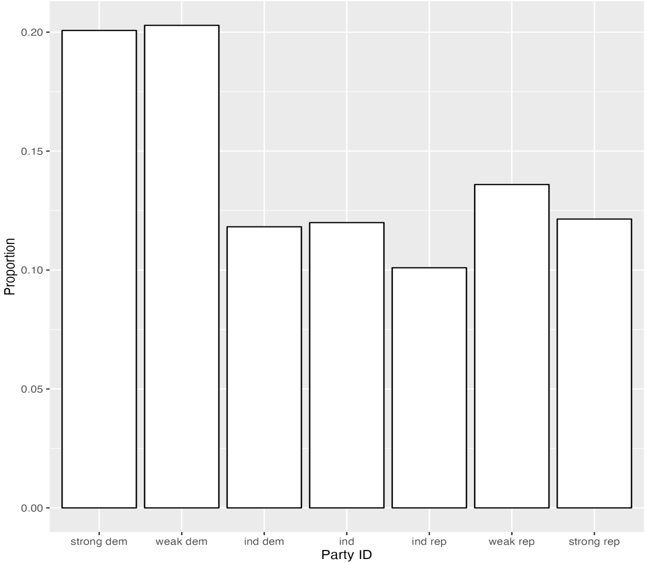
Types of probability functions

- **Probability density function (PDF)**: assigns a probability to each value of RV X.
- Formal definition: a **discrete PDF** for a variable X taking on the values $x_1, x_2, x_3, \ldots, x_n$ is a function such that:

$$f_X(x) = P[X = x_i] \qquad \text{for } i = 1, 2, 3, \ldots, n.$$

  and is zero otherwise.
- Example: coin flip, where heads = 1 and tails = 0.
- NB: $\sum_x f_X(x_i) = 1$

Discrete PDF Example: 2016 ANES Ideology 7 point scale



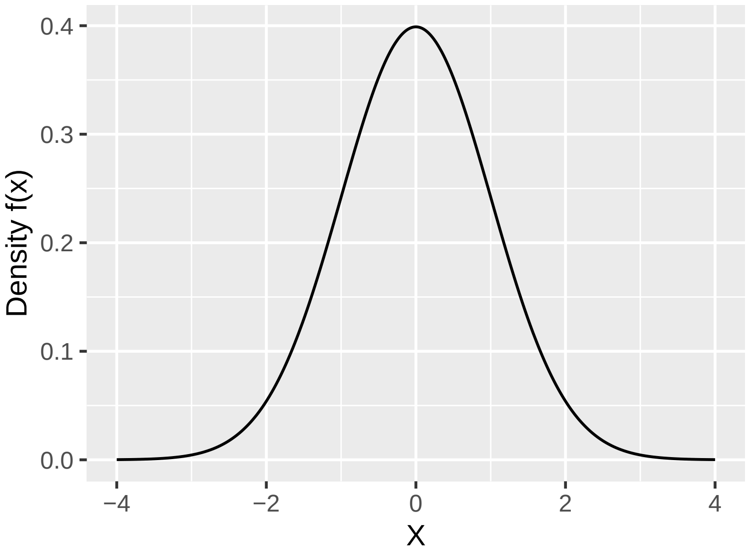
Continuous probability density function

- Harder to formally define since the prob. that a continuous RV takes on a particular value is generally zero.
- The probability that the realization of RV Z lies b/t two values, a and b is given by

$$P[a < Z < b] = \int_a^b f_Z(z)dz$$

- NB: $\int_{-\infty}^{\infty} f_Z(z)dz = 1$

Continuous PDF Example: Standard Normal—N(0,1)



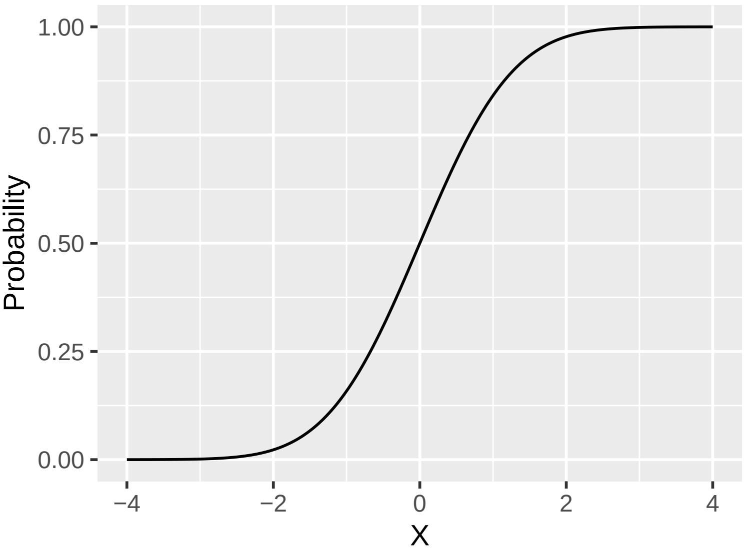
- **Cumulative probability density function (CDF)** gives the probability of a RV being less than or equal to some value: $F(x) = P[X \leq x]$.
- For a discrete RV

$$F(x) = \sum_{x_j \leq x} f(x_j).$$

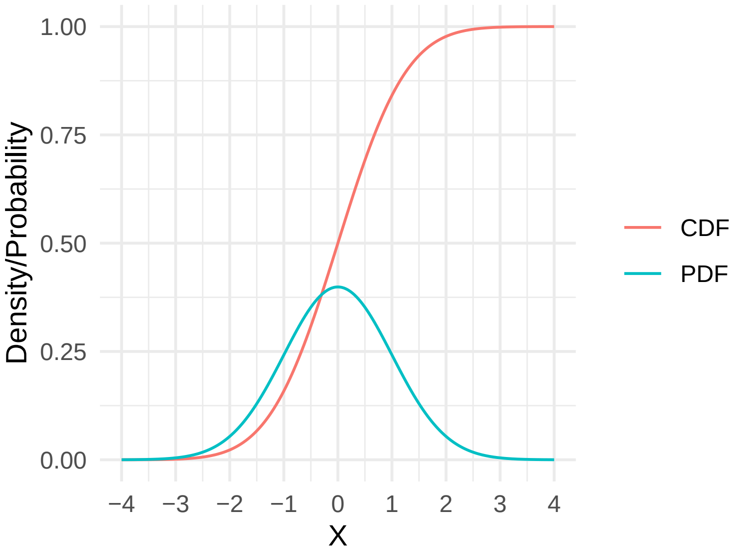
- For a continuous RV

$$\begin{aligned} F_X(x) &= P[X \leq x] \\ &= P[-\infty \leq X \leq x] \\ &= \int_{-\infty}^{x} f_X(u) du. \end{aligned}$$

CDF example



Relationship between a PDF and a CDF

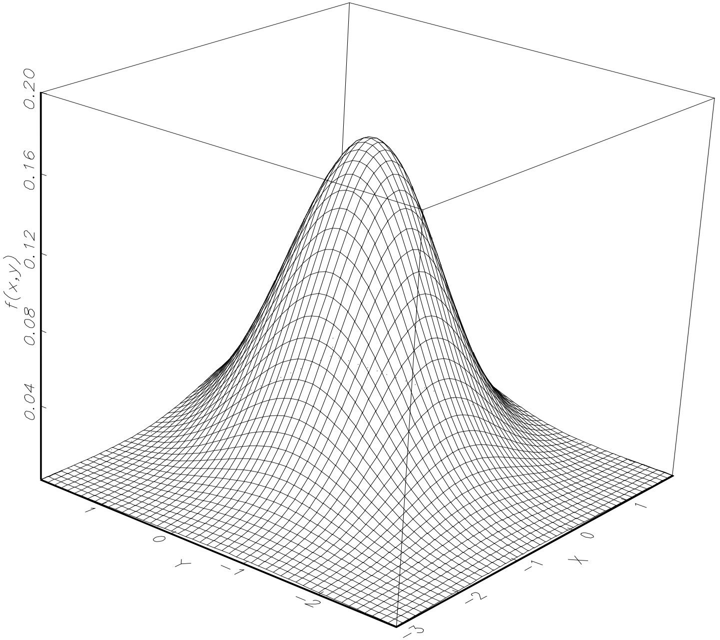


Joint probability density functions

- How does one variable relate to another?
- **Joint probability density functions**:

$$f(x, y) = P[X = x \text{ and } Y = y]$$

Joint PDF



Conditional probability density functions

- **Conditional PDF**: the probability that X has the realization x given that Y has the realization y. The written as

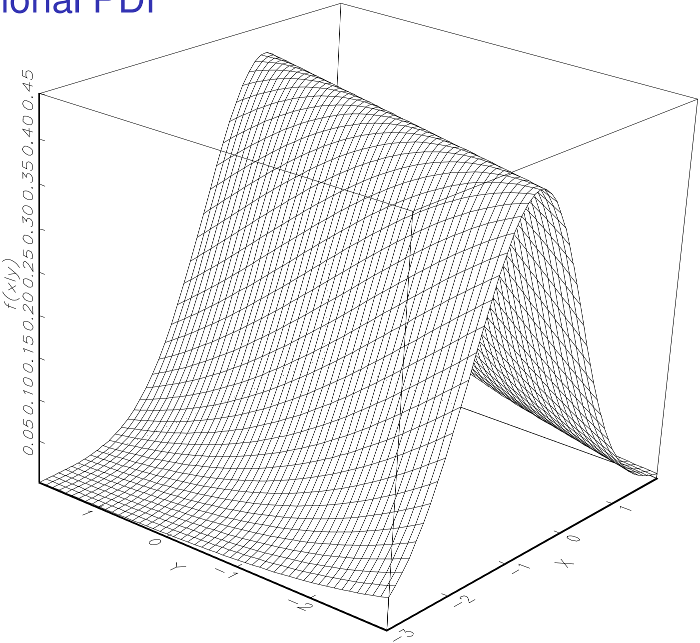
$$f_{X,Y}(x|y) = P(X = x|Y = y) \quad (1)$$

$$= \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad (2)$$

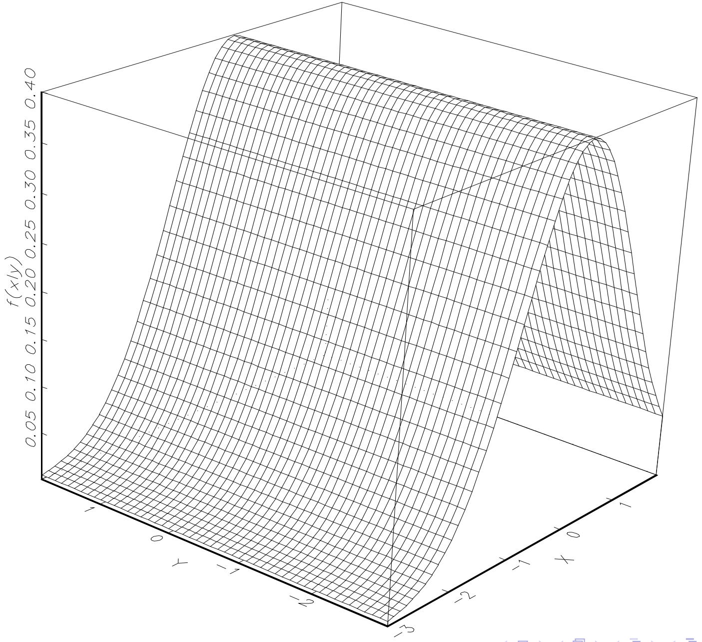
- **Statistical Independence**: X and Y are statistically independent if and only if

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$$

Conditional PDF



Conditional PDF, X and Y independent



Expectations, variance, and covariance

- Central tendencies of RVs.
- Expected value (μ) or population mean:

Expectation as a popuilation parameter

$$E[X] = \sum_{j=1}^{n} x_j f(x_j)$$

if X is discrete and has the possible values $x_1, x_2, \ldots, x_n$, and

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

if X is continuous.

Properties of expected values

- If a and b are constants,

$$E[a] = a$$
$$E[bX] = bE[X]$$
$$E[a + bX] = E[a] + E[bX] = a + bE[X]$$

- If X and Y are independent, then $E(XY) = E(X)E(Y)$

Covariance a correlation.

Population mean & sample mean

How can we obtain good estimation of the population from the sample

- If $x_1, x_2, \ldots, x_N$ is our population, then the population mean is

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

- Use sample mean to estimate the population mean. If $x_1, x_2, \ldots, x_n$ is our sample, then the sample mean is

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

- Expectation/mean: can do a lot theoretically and empirically w/ just these

Variance

- The distribution of values of X around its expected value can be measured by the variance:

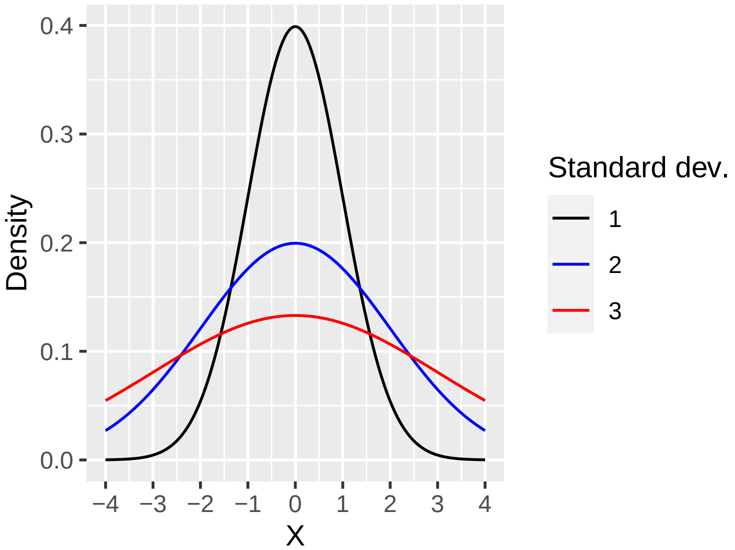
$$\begin{aligned} \mathrm{var}[X] &= E[(X - \mu_X)^2] \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

- **Standard deviation** of X (σ_X,): $\sqrt{\mathrm{var}[X]}$.
- To estimate the sample variance, do

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$$

- Horribly neglected statistic in popular discussions/uses of data analysis

Normal distributions with different variances



Covariance

- We often want to know if to variables are related or covary together.
- The **covariance** of two rvs, X and Y, with means μ_X and μ_Y is defined as

$$\begin{aligned} \mathrm{cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] && (3) \\ &= E[XY] - \mu_X \mu_Y. && (4) \end{aligned}$$

- NB: If X and Y are independent, then $E[XY] = E[X]E[Y]$ implying $\mathrm{cov}(X, Y) = 0$.
- To estimate the sample covariance, do

$$\widehat{\mathrm{cov}}[X, Y] = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

Correlation

- The size of the covariance depends on the units in which X & Y are measured.
- Can use correlation coefficient, which measures statistical association & $\in [-1, 1]$.
- The population correlation coefficient, ρ, is defined as:

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

- To estimate the sample correlation coefficient, do

$$\hat{\rho} = \frac{1}{n-1} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\hat{\sigma}_x \hat{\sigma}_y}$$

Conditional expectation

- The conditional expectation of X, given $Y = y$, is defined as

$$E(X|Y = y) = \begin{cases} \sum_x xf(x|Y = y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} xf(x|Y = y)dx & \text{if } X \text{ is continuous} \end{cases}$$

- Will get into this more with linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- The best linear predictor (BLP) of y_i given values of x_i:

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i.$$

Statistical inference

- Primary goal: estimate distribution of our population of interest using a sample & laws of statistics.
- Random sample: independent & identically distributed (iid) observations
- Are estimates obtained from a sample “good”?
- Statistical inference
- Sample avg. is an unbiased estimator of the true population mean (proof)
- Sample avg. is a RV—take a different sample, likely to get a different average—fiction of repeated samples (frequentist)
- The hypothetical distribution of the sample avg. is the **sampling distribution**.

- Two laws of statistics that can tell us about the precision and distribution of our estimator (in this case the avg.):
  - **Law of Large Numbers** states that, under general conditions, $\bar{Y}$ will be near μ_y with very high probability when N is large.
  - **Central Limit Theorem**: (under general conditions) the distribution of $\bar{Y}$ is well approximated by a normal distribution when N is large.
- Can use these to quantify the degree of uncertainty about estimates to separate signal from noise.

Desirable properties of estimators

- Unbiasedness: expectation of estimator = pop. parameter

$$E[\bar{x}_n] = E[X]$$

- Consistency: estimator converges to parameter as $n \uparrow$
- Efficiency: minumum variance for unbiasedness/consistency

Difference of means

- Recall the SATE: $\frac{1}{n}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$
- Estimate of the SATE:

$$\widehat{\text{SATE}} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n - n_1}\sum_{i=1}^{n}(1 - T_i) Y_i,$$

  where n_1 is size of treatment group, n is full sample size.
- Difference of means in randomized control trial (RCT) is unbiased estimator of SATE (consistent for pop. ATE)

Uncertainty

- For *any* estimate, it is essential to measure the precision w/ which it has been estimated—what is variability of the estimator?
- What is the likelihood that we observe a parameter value by chance?
- Standard deviation of the sampling distribution—can't be directly obtained since comes from hypothetical repeated random sample and/or random treatment assignment
- But can estimate it.

Estimating uncertainty

- Standard error of the sample mean:

$$\hat{\sigma}_{\bar{x}_n} = \sqrt{\frac{1}{n}\hat{\sigma}_X^2}$$

- Standard error of the difference of means:

$$\sqrt{\frac{1}{n}\hat{\sigma}_X^2 + \frac{1}{m}\hat{\sigma}_Y^2},$$

where m indicates the sample size for Y

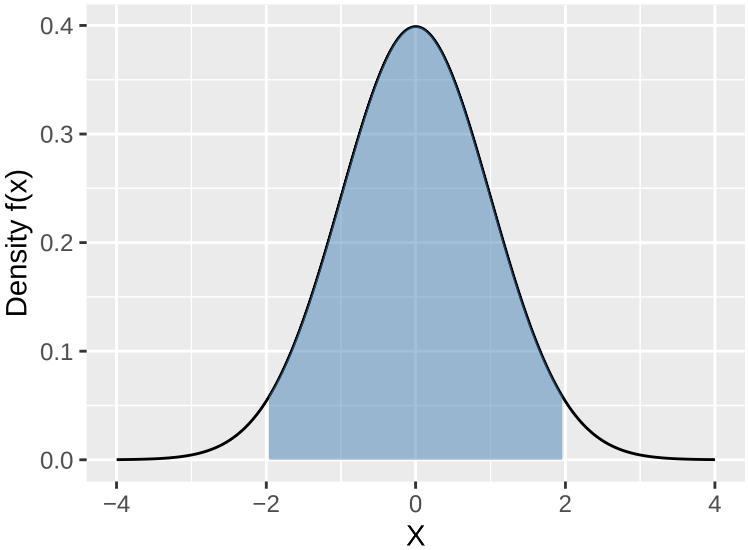
Confidence intervals

- Give a range of values that are likely to include the true value of a parameter
- Construct confidence intervals (CIs) using parameter estimates and standard errors
- E.g., 95% CI for the sample mean:

$$[\bar{x}_n - 1.96 \times \hat{\sigma}_{\bar{x}_n}, \bar{x}_n + 1.96 \times \hat{\sigma}_{\bar{x}_n}]$$

- 1.96 is a critical value; comes from the quantiles of the standard normal distribution; gives us area under curve where 95% of values lie.

Standard Normal—95% area under the curve



Hypothesis testing

- Most (?) common question of interest: does CI for a parameter estimate include 0?
- NB: for many/most estimators, will invoke CLT to say that $\sim N(0,1)$ and use relevant critical values (1.96 for 95%; 1.64 for 90%)
- Very common (but somewhat dubious) related usage: if parameter estimate $> 2 \times$ std. error $\Rightarrow$ "statistically significant" association (t test)
- Other locution: "bounded away from zero"

Regression Analysis

Introduction to regression analysis

- Workhorse tool in political analytics—need it for machine learning and other more sophisticated data science tasks
- Specify a model the indicates how covariates relate to an outcome
- Obtain estimates of coefficients on covariates to conduct inference

Linear regression model

- Recall condit'l expectation and assume linear structure to data generating process (DGS):

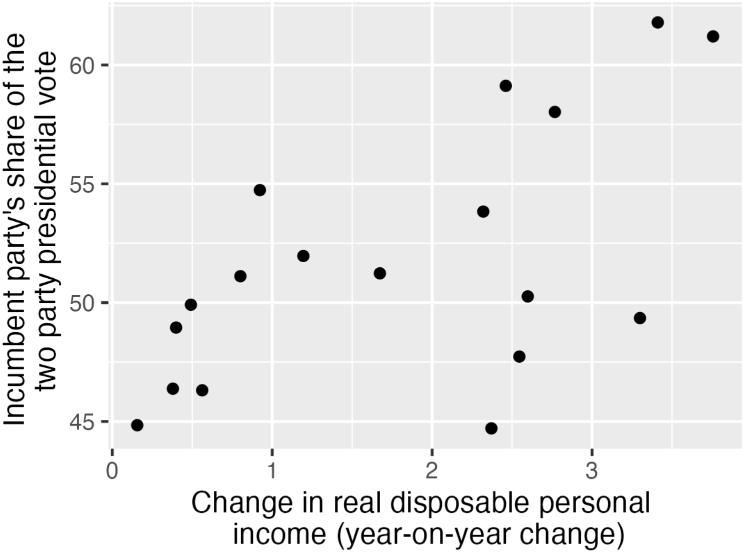
$$E(y_i|x_i) = \beta_0 + \beta_1 x_i.$$

- Modify to include outcomes we can observe and incorporate stuff we cannot:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Assuming $E[\varepsilon_i] = 0$ & $\varepsilon_i \perp\!\!\!\perp x_i$
- Will estimate β_0 and β_1 from data

Economic growth and presidential election outcomes



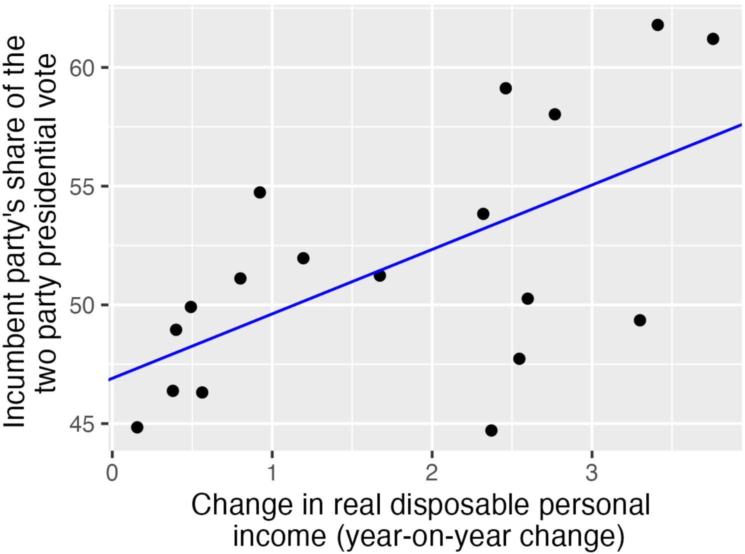
Ordinary Least Squares (OLS)

- OLS is a great option for fitting a line to data
- Choose values for β_0 and β_1 that minimize the sum of squared residuals (vertical deviations from the estimated line):

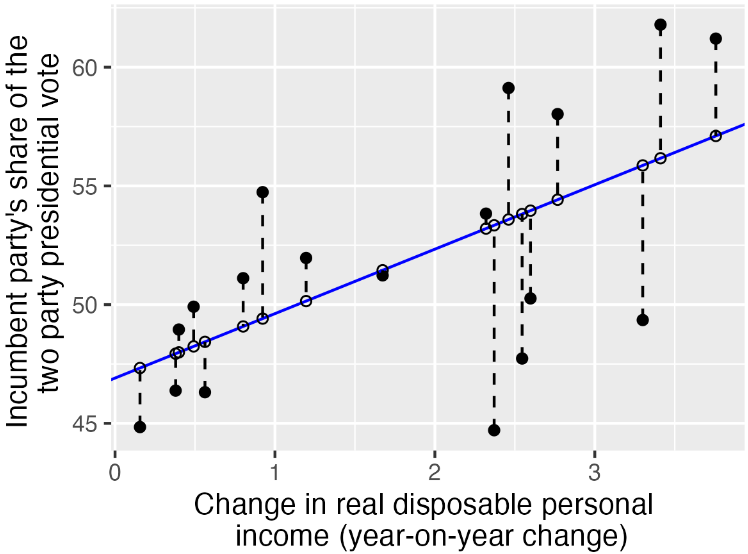
$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

- $$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Estimated regression line



Residuals



Incumbent party's share of the two party presidential vote
60
55
50
45
0
1
2
3
Change in real disposable personal income (year-on-year change)

OLS properties

- With a set of (reasonable) assumptions, OLS estimates have "optimal" properties—unbiased/consistent, efficient
- Interpretation of slope: unit change in x_i gives you β_1 change in y_i
- Also provides estimates of uncertainty of coefficients
- Measures of goodness of fit: $R^2 \in [0, 1]$ and F-test

Economic Voting Model—OLS Regression Results

	(1)				
	Est.	S.E.	2.5%	97.5%	t
(Intercept)	46.906	1.920	42.835	50.977	24.425
Change in DPI	2.714	0.904	0.798	4.631	3.002
N	18				
R-squared	0.36				
Adj. R-squared	0.32				

Violations of OLS assumptions

- If Gauss-Markov assumptions are violated, OLS may lose desirable properties—but remarkably robust
- But can sometimes fix with alternative versions of least squares (e.g., Generalized Least Squares)—but cure can sometimes be worse than the disease!
- Tests of assumptions are available

Multiple regression

- Can extend this model to include lots of covariates (K):

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i$$

- Similar assumptions as with bivariate regression ($\varepsilon_i \perp\!\!\!\perp x_{ki} \forall x_{ki}$), plus concerns about correlation among xs.
- Interpretation of Coefficients: one unit increase in x_k is associated w/ a β_k change in y, holding all other xs constant.
- Can write more compactly using matrix notation:

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i$$

Maximum Likelihood Estimation

- Estimation technique with strong justification from statistical theory (unlike OLS)
- Assume distribution for data—i.e., all observations are drawn from the same distribution (PDF)—assuming DGP
- Given distribution, can write down a function that describes how the data for an individual were generated
- Key assumption: observations in sample are independent and identically distributed (iid)—we get independence by assuming we have a random sample.
- Gives us the likelihood function; ML means we choose parameter estimates that maximize the likelihood of observing the sample that we actually observe.
- Find the maximum of the likelihood function wrt to the parameters of interest.

Flipping coins (Bernoulli trials)

- $$Y = \begin{cases} 1 & \text{for a head} \\ 0 & \text{for a tail} \end{cases}$$
- Let $p = \Pr(Y = 1)$
- Flip coin n times to obtain the sample: $Y_1, \ldots, Y_n$.
- Key assumption: coin flips are independent and prob of observing a head or a tail is the same at each trial.
- From the given sample, we want to estimate p.
- Note that $p \equiv$ population mean:

$$\begin{aligned} E(Y_i) &= \Pr(Y_i = 1) \cdot 1 + \Pr(Y_i = 0) \cdot 0 \\ &= \Pr(Y_i = 1) \\ &= p \end{aligned}$$

- Let L_i = likelihood for observation i—What is likelihood for ith toss?
- For each observation could get a head (w/ prob p) or could get a tail (w/ prob $1 - p$):

$$L_i = p^{Y_i}(1-p)^{(1-Y_i)}$$

- If we have iid sampling then

$$L = \left[p^{Y_1}(1-p)^{(1-Y_1)}\right]\left[p^{Y_2}(1-p)^{(1-Y_2)}\right]\cdots\left[p^{Y_n}(1-p)^{(1-Y_n)}\right]$$
$$= \Pi_{i=1}^{n} p^{Y_i}(1-p)^{(1-Y_i)}.$$

- Take natural log to get log-likelihood—easier to deal w/:

$$\ln L = \sum_{i=1}^{n} \ln L_i = \sum_{i=1}^{n} \ln\left[p^{Y_i}(1-p)^{(1-Y_i)}\right]$$
$$= \sum_{i=1}^{n} \left[Y_i \ln p + (1 - Y_i)\ln(1-p)\right]$$

- Want to choose value for p that maximizes $\ln L$.
- Use calculus: take derivative of likelihood wrt to parameter of interest, set equal to 0 and solve.
- Returning to our likelihood function:

$$\frac{d \ln L}{dp} = \sum_{i=1}^{n} \left[Y_i \frac{1}{p} - (1 - Y_i) \frac{1}{1-p} \right]$$

- Setting the derivative equal to zero and solving gives:

$$\sum_{i=1}^{n}\left[Y_i\frac{1}{p}-(1-Y_i)\frac{1}{1-p}\right]=0$$

$$\frac{\sum_{i=1}^{n}[Y_i(1-p)-(1-Y_i)p]}{p(1-p)}=0$$

$$\sum_{i=1}^{n}[Y_i-pY_i-p+pY_i]=0$$

$$np=\sum_{i=1}^{n}Y_i$$

$$p=\frac{\sum_{i=1}^{n}Y_i}{n}$$

ML for normal linear regression model

- Take our simple regression model from before ($y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$) and add assumption $\varepsilon \sim N(0, \sigma^2)$.
- Implies log likelihood:

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}\right]$$

- Maximize the log likelihood function with respect to β_0 and β_1 and then with respect to σ^2.
- No closed form solution—use hill-climbing algorithm (or gradient decent) for numerical solution
- Results are essentially equivalent to OLS!
- Consistency and efficiency

Logistic regression

- If y_i is binary, can use OLS (well, GLS actually) to estimate parameters via linear probability model (LPM)—but can have problems
- Possible better option: assume different distribution for ε and do ML.
- The logistic function gives rise to the **logit** model:

$$\Pr(y_i = 1) = F(\beta'\mathbf{x}_i) = \frac{e^{\beta'\mathbf{x}_i}}{1 + e^{\beta'\mathbf{x}_i}}.$$

- Using logistic CDF Guarantees predicted values $\in [0, 1]$, but nonlinear model.

Logistic likelihood

- This will give rise to the log-likehood function

$$\ln L = \sum_{i=1}^{n} \left\{ y_i \ln \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}} + (1 - y_i) \ln \left[1 - \frac{e^{\beta' \mathbf{x}_i}}{1 + e^{\beta' \mathbf{x}_i}} \right] \right\}$$

- Will use this for classification in machine learning
- Can extend to outcome with multiple categories.
- If assume normal distribution for errors, get probit model.

Generalized linear model (GLM) approach

- Can estimate models w/ discrete outcomes (and non-continuous outcomes)
- Essentially end up in the same place as with ML, but different way to get there.
- General framework for deriving a wide range of models.
- For logistic regression, assume logit "link"