Midterm Studyguide CPSC 540 :,) 

MATH REVIEW & PROBABILITY THEORY

Probability Spaces (Ω, B, P)

Sample Space (Ω): set of all possible outcomes.
Ex: single dice roll → $\Omega = \{1, 2, 3, 4, 5, 6\}$; coin toss → $\Omega = \{H, T\}$
Event Space (B): Sets of outcomes (events) of interest.
Ex: even numbers on a dice roll $\{2, 4, 6\}$
Conditions:

1. $\emptyset \in \mathcal{B}$
2. $A \in \mathcal{B} \rightarrow A^c \in \mathcal{B}$
3. $A_1, A_2, \ldots A_n \in \mathcal{B} \rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{B}$

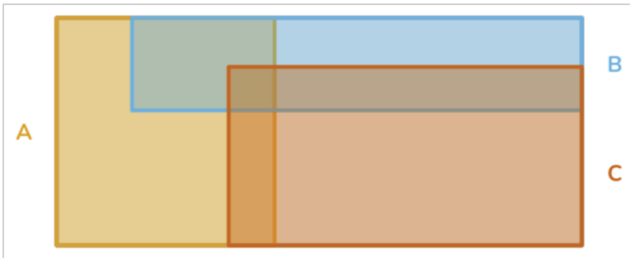
1. *Trivial*: $\{\varnothing, \Omega\}$
2. *Powerset*: all subsets of Ω

Probability Function (P): Assigns probability [0,1] to each event in B
Conditions:

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots \in \mathcal{B}, A_i \cap A_j = \emptyset \rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Set Theory

- $A \cup B = B \cup A$
- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$
- $(A \cup B) \cup C = A \cup (B \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$



Formulas derived from set theory:

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(B \cap A^c) = P(B) - P(A \cap B)$

Conditional Probability "restricts" sample space to event B.

For $A, B \in \mathcal{B}$ and $P(B) \neq 0$:

$$P(A \mid B) = \frac{P(A \cup B)}{P(B)}$$

Independence: P(A|B)=P(A) → knowing B doesn't change probability of A.

Bayes' Theorem: update probability of hypothesis based on new data.

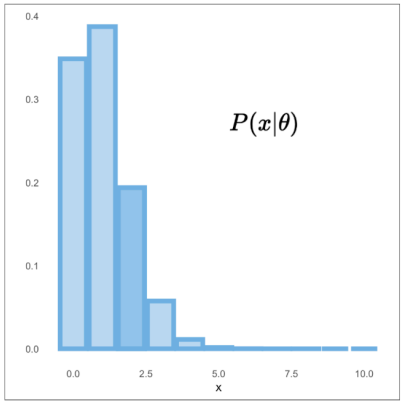
$$P(A \mid B) = \frac{P(A \cup B)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

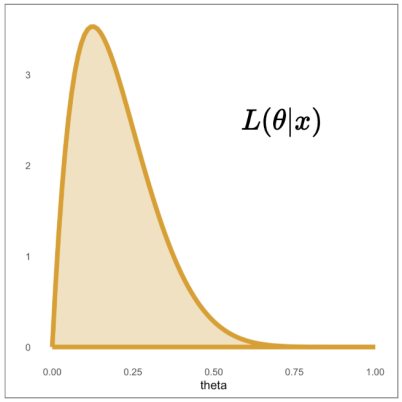
$$P(\text{theory} \mid \text{data}) = \frac{P(\text{data} \mid \text{theory}) \cdot P(\text{theory})}{\text{P(data)}}$$

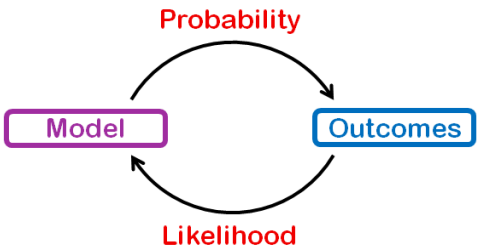
"How did my theory *change* after seeing the data?"

Probability vs Likelihood

- Probability: event is random
- Likelihood: parameters are random

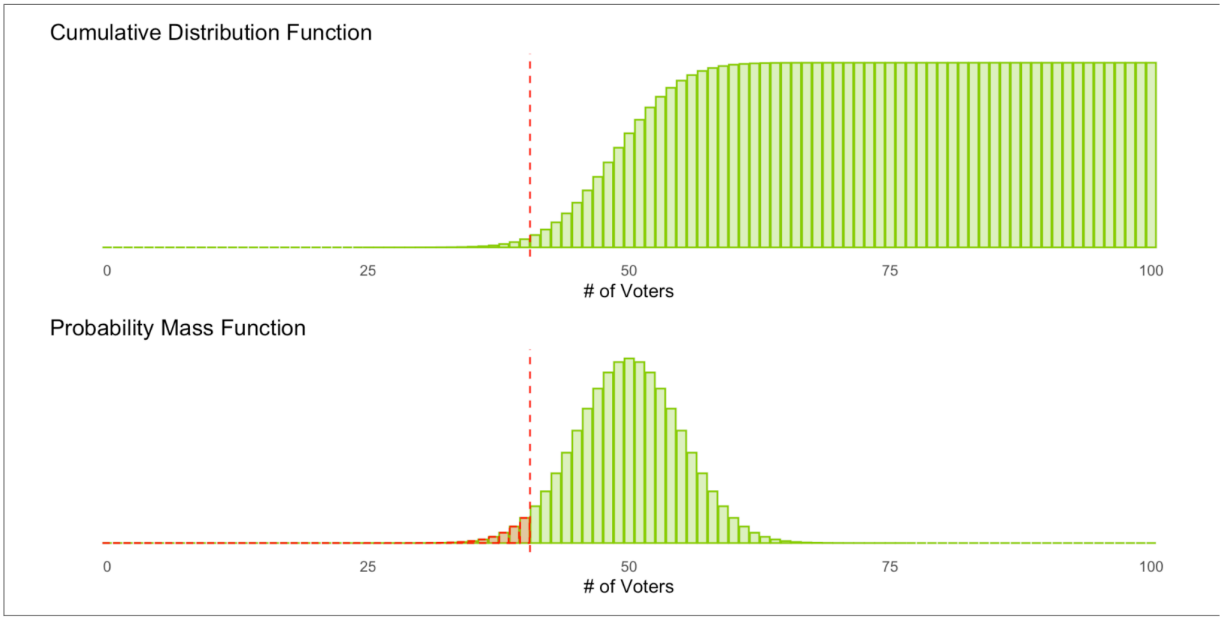


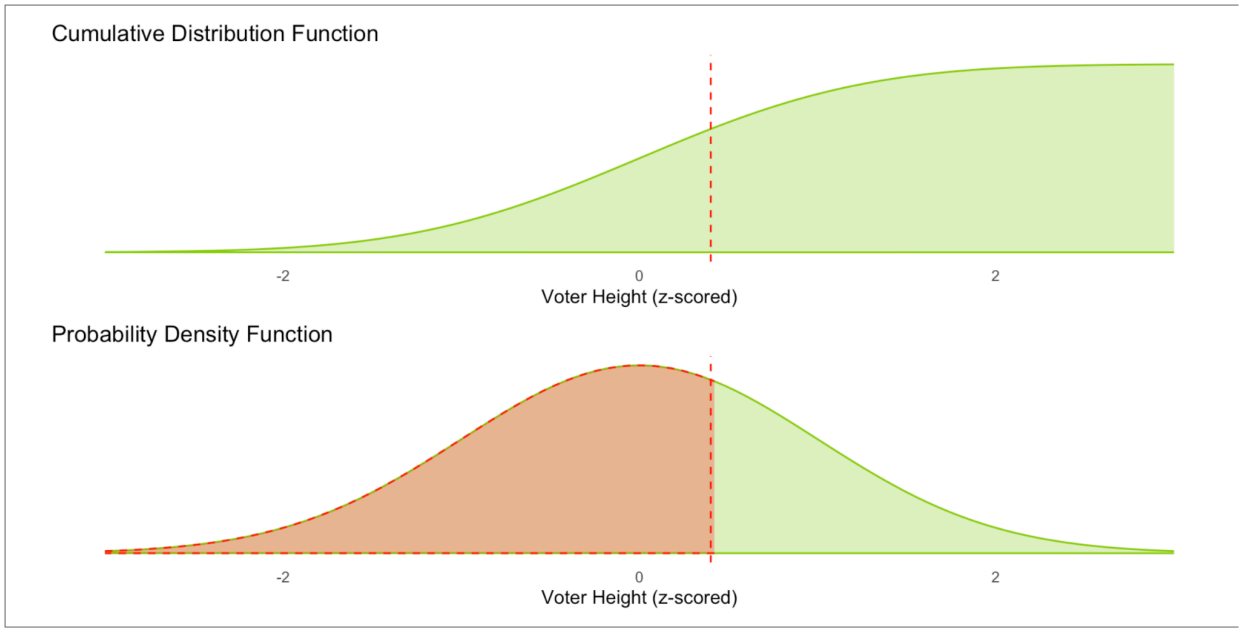




Probability Density Function (PDF): Describes likelihood of different outcomes. Probabilities obtained by integrating over intervals.

Cumulative Distribution Function (CDF): integral of PDF from negative infinity to x.





Expected Value: The average or "typical" value for a random variable. Sum of the possible values weighted by the probability of that value

$$\mathbb{E}(x) = \sum_{i}^{n} \underbrace{x}_{\text{value of x}} \overbrace{f_X(x)}^{\text{probability of x}}$$

Discrete Random Variables: countable values

Continuous Random Variables: uncountable values (e.g., height, time).

- **discrete:** $\mathbb{E}(x) = \sum_i^n x f_X(x) = \sum_i^n x P(X = x)$
- **continuous:** $\mathbb{E}(x) = \int_{-\infty}^{\infty} x f_X(x) dx$

Variance: spread of distribution around its mean.

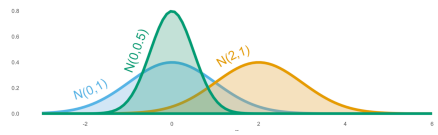
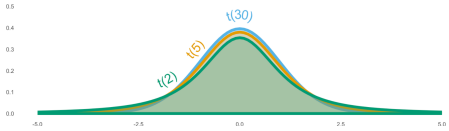
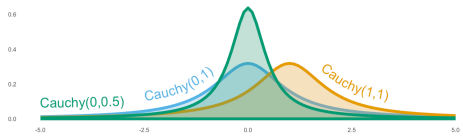
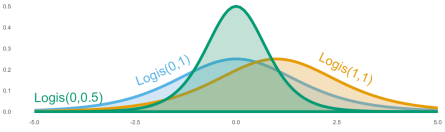
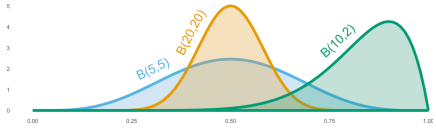
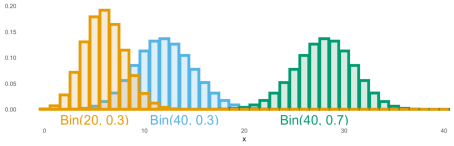
Moments of a distribution are expectations: $\mu_n' = \mathbb{E}X^n$

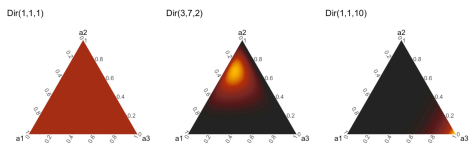
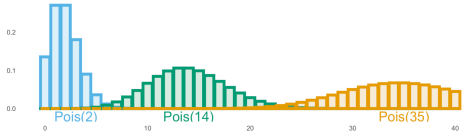
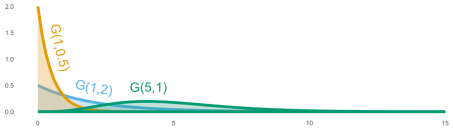
Central Moments replace X with mean-centered value X - μ: $\mu_n = \mathbb{E}(X - \mu)^n$

1. 1st Moment: mean.
2. 2nd Central Moment: variance.
3. 3rd Moment: Skewness (the asymmetry of the distribution).
4. 4th Moment: Kurtosis ("tailedness" of the distribution, indicating how likely extreme values are).

Mean and Variance of Transformations

- $\mathbb{E}(aX + b) = a * \mathbb{E}(X) + b$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $Var(aX + b) = a^2 Var(X)$
- $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$

Distribution	Graph	Properties	Notes
Normal		μ: mean/median/ center; σ: standard deviation	
Student t		ν: degrees of freedom (smaller = higher kurtosis)	when ν>30, distribution close to N(0, 1)
Cauchy		location: center scale: width	undefined mean and variance (same as student-t distribution where ν=1)
Logistic		μ: mean s: scale	
Beta		α: # successes β: #f failures	
Binomial		n: # trials p: P(success)	Bernoulli = binomial where n = 1

Dirichlet		α: simplex of probabilities for each outcome	generalization of Beta distribution when >2 possible outcomes
Poisson		λ: rate/expected number of counts	mean = var = λ. overdispersion: counts have more variance than this (use neg. binomial distribution)
Gamma		k: shape θ: scale	

Graph Theory

Nodes (Vertices): Represent entities.
Edges: Connections between nodes.
Path connects nodes w/o repetition
Cycle returns to original node w/o passing through any node more than once
Important Graph Types
Directed Acyclic Graphs (DAGs): directed edges, no cycles
Cliques: Subsets of nodes where each pair of nodes connected by an edge.
Density: Measures how connected a graph is by comparing # edges to total possible # edges.

Information Theory

KL Divergence: Measure of distance btwn distributions p and q (how much info lost if you approximate p with q?)

$$D_{KL}(p||q) = \underbrace{H(q,p)}_{\text{cross entropy}} - \underbrace{H(p)}_{\text{entropy}}$$

Entropy: measure of chaos/disorder
Cross-entropy: If we think q(x) is probability distribution, but p(x) = *real* distribution, how surprised will we be?

MARKOV CHAIN MONTE CARLO (MCMC)

Goal: Generate samples from target distribution w/o knowing PDF/CDF
Markov Process: graph of states + transitions

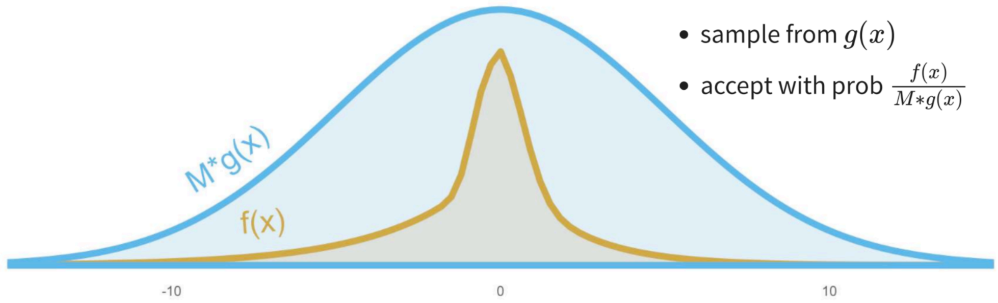
MCMC used when sampling from a probability distribution when direct sampling is difficult.
Chain: sequence of draws from distribution
Draw: one sample/state from chain
Target Distribution: distribution from which we want to sample.
Proposal Distribution: distribution used to generate candidate samples

- Ex: normal and uniform distributions.

Markov Chains: sample depends only on previous one.
Monte Carlo: using random sampling to compute numerical results

- **Pros:** easy, sometimes fast
- **Cons**: not always fast, not always exact

Accept-Reject Sampling: Useful when known PDF/CDF, problematic when exact distribution unknown/complex



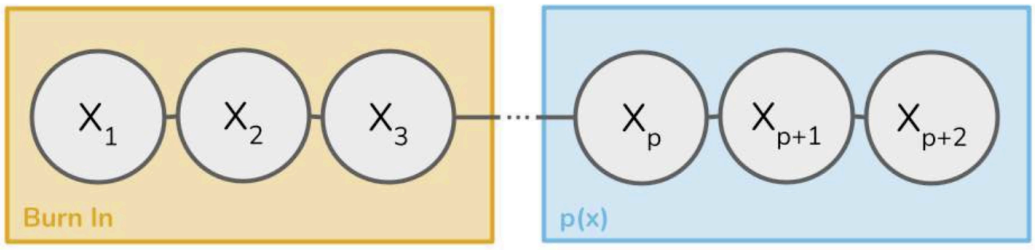
- Samples uncorrelated, throwing away useful info
- Hard to choose g(x)
- More common value in target distribution = more likely to accept

Markov chains enable sampling by using current state to inform next (more efficient in certain situations)

Key Concepts in MCMC

Markov Chain Property: state depends only on last one, allowing convergence towards target distribution.
Steady State: When Markov chain reaches steady state, samples can be taken as if they were from target distribution.
Goal of MCMC: find Markov chain whose steady state (after burn-in) = p(x)

- burn-in can be used for more efficient sampling



Metropolis Algorithm

Proposal: A new state is proposed.

Acceptance Probability: P(accepting new state) based on ratio of probabilities (how likely new state is compared to current one).
P(new state) > P(current state) → accepted
Else → accepted with probability based on ratio.
Ex: King Markov visiting islands, where transition probability between islands proportional to population sizes.

Metropolis-Hastings Algorithm

Extension of Metropolis Algorithm: Adds correction term to account for biases introduced by asymmetrical proposal distribution.

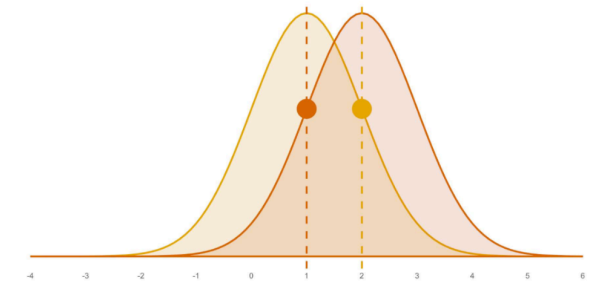
1. generate proposed value
2. calculate P(accepting proposal)
3. accept or reject

Hyperparameters:

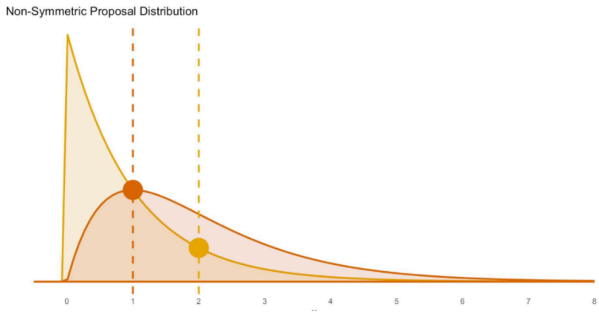
- **Proposal Width**: how far proposed samples are from the current state (standard deviation in normal distribution).
    - How far away from current value we're willing to go
    - Small widths → high autocorrelation
    - Large widths→ low acceptance rates.
- **Burn-in**: how long chain gets to reach stationary distribution (intial samples discarded)
- **Thinning**: how many samples we *thin* to get rid of autocorrelation (take every nth sample)

Symmetric vs Asymmetric Proposals: correction cancels out when using symmetric distributions (e.g., normal distributions). However, when using asymmetric distributions (e.g., gamma distribution), Hastings correction adjusts for bias.

$q(x_n|x_*) = q(x_*|x_n)$



$q(x_n|x_*) \neq q(x_*|x_n)$



Metropolis algorithms may reject many samples, which can lead to inefficiencies.

Gibbs Sampling

Use: For multidimensional distributions where joint sampling is difficult but conditional distributions are easier to sample from.
Key Feature: Always accepts samples since conditional densities are known.
Procedure: Iteratively update each variable conditioned on current values of other variables.
Pros: Efficient for multidimensional sampling with known conditional distributions.
Cons: Requires knowledge of conditional distributions.

Hamiltonian Monte Carlo (HMC)

Goal: Generate efficient proposals using gradient information from the target distribution.
Core Idea: Treat negative log-density as potential energy surface and use physics simulations (momentum and position updates) to generate samples.

- Likely reject if energy not conserved (result of leapfrog step approx)

Steps:

1. Start with a current state.
2. Assign random momentum ("flick") to simulate a particle's movement through the target distribution.
3. Use Leapfrog Integration to discretize the trajectory.
4. Accept/reject based on the ratio of the energy in the current and proposed states.

Hyperparameters:

- **Leapfrog Step Size:** Smaller step sizes result in better approximations but increase computation time.
- **Number of Steps**: More steps explore further but may cause divergence if not tuned correctly.

Challenges:

- **Divergent Transitions:** Occur when the approximated energy does not match the true energy, leading to sample rejections.
- **Steep Curvature:** Difficult for HMC to navigate, causing potential biases in sampling.

Diagnostics and Convergence
Trace Plots: visualization of samples across iterations.

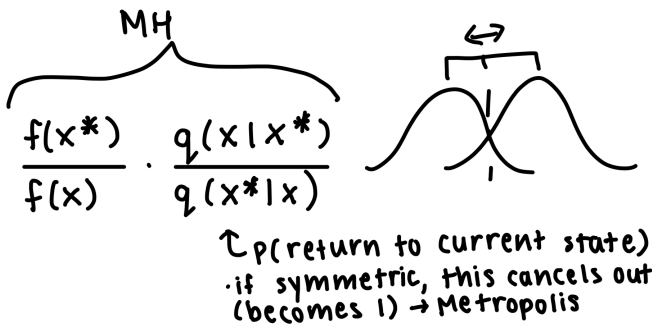
1. start chain from dif values
2. compare samples from chains
3. Look for "fuzzy caterpillar" patterns indicating convergence.

R-hat Statistic: Measures variance within chains relative to the variance between chains.

- $R-hat < 1.1$ indicates good convergence.

Larger proposal widths: fewer accepted samples but enable more exploration.
Smaller widths: more correlated samples.

Algorithm	Accept/Reject	Metropolis	Metropolis-Hastings	Gibbs Sampling	Hamiltonian MC (HMC)
Key Definition	Simple method to generate samples based on whether they meet a certain criterion	Special case of MH using symmetric proposal distributions	Generalization of Metropolis, allows non-symmetric proposals	Samples from multi-dimensional distributions by sampling conditionals	Uses physics (momentum & gradients) to generate proposals efficiently
Steps	1. Generate candidate sample 2. Accept if it meets a certain probability	1. Propose new state from symmetric distribution 2. Compute acceptance probability 3. Accept/reject based on prob. 4. Repeat	1. Propose new state 2. Compute acceptance probability 3. Accept/reject 4. Repeat	1. Update each parameter one at a time, conditional on the others 2. Iterate for all parameters	1. Assign random momentum to a particle in the target distribution 2. Simulate particle movement using Hamiltonian dynamics 3. Accept/reject based on energy conservation
Use Cases	Easy to implement in simple cases where rejection is not costly	Used when proposal distribution is symmetric and well-tuned	More flexible than Metropolis, works with asymm. proposal distributions	Useful for multi-dimensional distributions with known conditionals	Efficient for high-dimensional problems where gradient info is available
Acceptance Criterion	Sample accepted if it meets a threshold	Sample accepted if new state is more probable or randomly selected otherwise	Sample accepted if probability ratio exceeds random draw	Always accepted because conditional distributions are known	Sample accepted if energy between current and proposed states is conserved
Proposal Method	Generate random samples from target distribution	Symmetric proposal distributions (e.g., normal around current state)	Asymmetric or symmetric proposal distributions	Based on conditional distributions	Uses gradients to propose new samples
Pros	Simple and effective for some applications	Simple to implement, doesn't require asymmetric corrections	Works with more complex proposal distributions	Efficient for multidimensional problems, always accepts proposals	Efficient, uses gradient information to generate high-quality proposals
Cons	High rejection rates possible	Can be inefficient if proposal is poor	Requires computing proposal probability	Requires knowledge of conditional distributions	Complex, requires tuning of step size and steps
Computational Cost	Low for each sample, but many samples may be rejected	Moderate, depends on proposal width and acceptance rate	Moderate to high due to more complex calculations	Low computational cost per step, but may require many iterations	High, due to physics simulations
Tuning Parameters	None	Proposal distribution (width)	Proposal distribution (width), acceptance probability	None, but requires knowledge of conditional distributions	Step size, number of leapfrog steps
When to Use	Simple problems with known distributions, low cost of rejection	When proposal distributions are easy to define and symmetric	When asymmetric proposals are more suitable	When conditional distributions are easy to compute	Complex high-dimensional distributions where gradient info is available



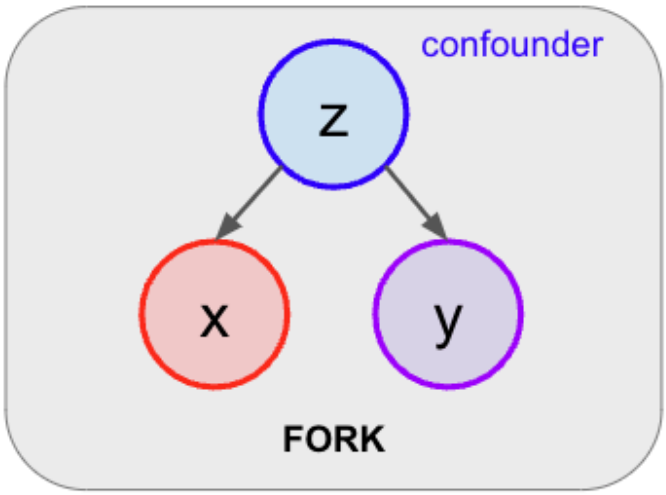
CAUSAL INFERENCE

Prediction vs. Inference vs. Description

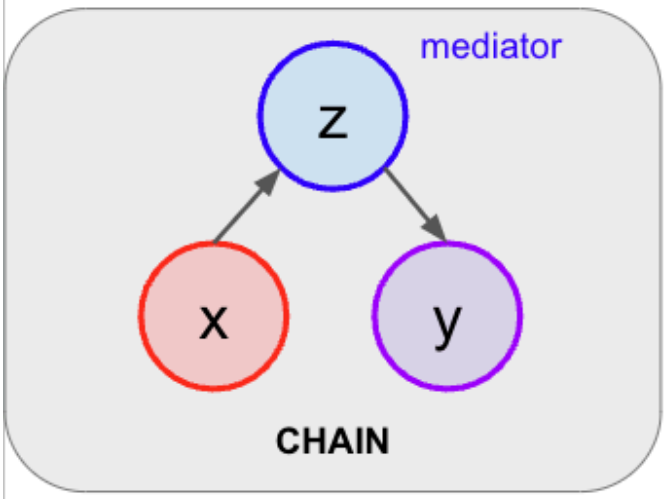
- **Prediction:** get output as close to actual output as possible
- **Description:** describe the data
- **Inference:** about relationships between variables

DAG: Directed Acyclic Graph
Directed: All edges point from one node to another (directionality of cause and effect).
Acyclic: No cycles allowed (variable cannot cause itself directly or indirectly).
Graph: Consists of nodes (variables) and edges (causal effects between variables).
DAGs do not assume the strength or nature (e.g., linearity) of relationships between variables, only that causal relationship exists.

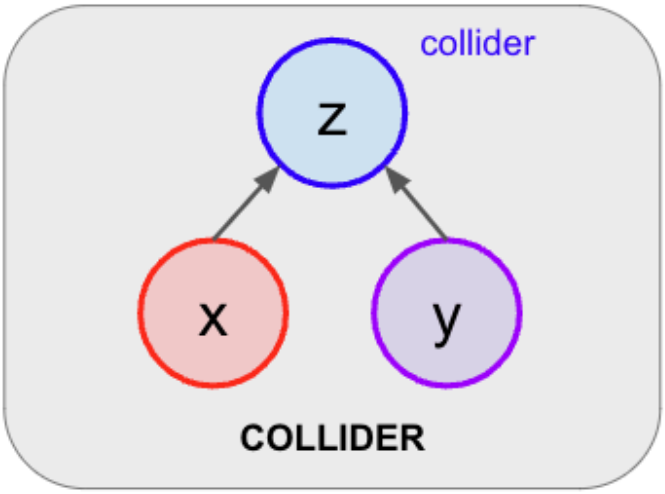
Common DAG Paths



- Z causes both X and Y, making X and Y appear associated even though they don't cause each other
- Ex: Education (Z) influences both income (X) and cognitive test scores (Y), but income doesn't directly affect test scores.
- Control for confounding variable Z to eliminate spurious associations between X and Y.



- X causes Z, and Z causes Y, but X affects Y only through Z.
Ex: Studying (X) improves comprehension (Z), which in turn improves exam scores (Y).
- Do not control for Z if goal is to estimate total effect of X on Y (including indirect effects).
- Do control for Z if you want direct effect of X on Y, bypassing mediator.



- X and Y both cause Z, but controlling for Z can induce artificial association btwn X and Y even though they're independent.
- Ex: College acceptance (Z) depends on both extracurriculars (X) and exam scores (Y), but conditioning on Z creates a false relationship between X and Y.
- Collider bias occurs when conditioning on collider (Z), which should be avoided to prevent misleading associations.

Open Path: Allows association to transmit between variables (e.g., in forks and chains).
Closed Path: No association transmits (e.g., colliders, unless conditioned on).

Backdoor Path: non-causal path that introduces spurious associations between variables.

- identify and close backdoor paths to achieve accurate causal estimates.

Methods for Controlling Variables

Covariates

- adjusts for confounders by including them as covariates in the model.
- Ex: Examining effect of extracurriculars on exam scores while adjusting for college acceptance in a linear regression. Acceptance can act as a collider if improperly controlled for, introducing bias (collider bias).

Subsetting

- subset the data, focusing on specific groups.
- Ex: Instead of adding college acceptance as a covariate, subset data to only include individuals who were accepted or not accepted.

Matching

- Create balanced groups (treatment vs. control) based on confounders, ensuring groups are comparable.
- Ex: In a job training scenario, men were overrepresented in training group compared to control group. To balance this, weights can be assigned based on gender proportions to make groups comparable.

Propensity Scores

- Estimate probability of treatment (e.g., receiving job training) given a set of covariates (e.g., gender, education) by accounting for confounders.
- Conditions:
    - No unmeasured confounders are present.
    - Each subject has a non-zero probability of receiving treatment.

Inverse Probability Weighting (IPW)

- Use propensity scores for unbiased treatment effect estimates.
- weights individuals in treatment and control groups based on the inverse of their propensity scores.
- weighted groups are comparable, even if they differ in key characteristics like gender.
- more unlikely it is that someone should be in their treatment group, the higher the weight that subject has

Dif-in-Dif

- *Assumption*: Parallel Trends

Regression Discontinuity

- **Running Variable**: determines whether you're above or below cutoff
- **Cutoff**: threshold for getting treatment
- **Bandwidth**: how wide around cutoff to look

GENERALIZED LINEAR MODEL

- **Independently and Identically Distributed (IID)**: Data points assumed to be independent of each other and come from same distribution.
- **Linearity**: relationship btwn predictors and outcome is linear in the parameters.
- **Normally Distributed Errors:** Errors assumed to be normally distributed around regression line.
- **Homoscedasticity**: Variance of errors is constant across all levels of the predictors.

Model Fitting Methods

Least Squares Method: Minimizes squared differences btwn observed and predicted values.
Maximum Likelihood Estimation (MLE): choose parameters (e.g., coefficients) that maximize likelihood of observed data.

Generalized Linear Models (GLMs)

Extend linear regression by allowing different types of outcome variables and likelihood functions.
Use when: relationship between predictors and response variable is not linear or when response variable does not follow normal distribution.
Two components:

- **Link Function:** Transforms linear predictor Xβ into a scale appropriate for outcome variable.
    - Depends on range of expected values
- **Likelihood Function**: Specifies how data points are distributed around expected value.
    - Must reflect nature of the data

Common Examples of GLMs

Model Type	Description	Outcome Variable Type	Link Function	Likelihood Function	Common Applications
Linear Regression	Continuous outcome using linear predictors.	Continuous	Identity Link	Normal Distribution	Predicting prices, scores, etc.
Logistic Regression	Binary outcome (e.g., success/failure).	Binary (0 or 1)	Logit Link	Bernoulli Distribution	Classification problems, e.g., spam detection, disease presence
Poisson Regression	Count data where outcome represents event counts.	Count	Log Link	Poisson Distribution	Modeling number of occurrences, e.g., number of emails received, website visits
Gamma Regression	Continuous, positive outcome variables with skewness.	Continuous (Positive)	Inverse Link	Gamma Distribution	Modeling insurance claims, survival times
Binomial Regression	Predicts number of successes in a fixed number of trials.	Binary (Successes/Trials)	Logit Link	Binomial Distribution	Predicting proportions, e.g., voter turnout, pass rates
Negative Binomial Regression	Overdispersed count data where variance > mean.	Count	Log Link	Negative Binomial Distribution	Modeling counts with overdispersion, e.g., RNA sequence data
Robust Student t Regression	Continuous data w outliers by using heavy-tailed dist.	Continuous	Identity Link	Student t Distribution	Outlier-prone data, e.g., financial data with extreme values

Mixed Effects Models (MEMs)

Example: Modeling School Fundraising Impact
Problem: Predict effect of fundraising on standardized test scores for students across different schools.
Why MEMs?

- If we fit separate intercepts and slopes for each school, we're throwing away info and can't generalize findings
- Schools may have different starting points (baseline test scores) and react differently to fundraising (slope variability).
- Consist of fixed and random effects

Fixed effects: Parameters associated w all observed levels of variable (e.g., treatment/control groups).

- Baseline standardized test scores and slope for fundraising are constant across all schools.

Random effects: Parameters drawn from population distribution, which help generalize findings across a broader population (e.g., subjects, schools).

- Each school has its own intercept and slope, which vary around population-level averages.

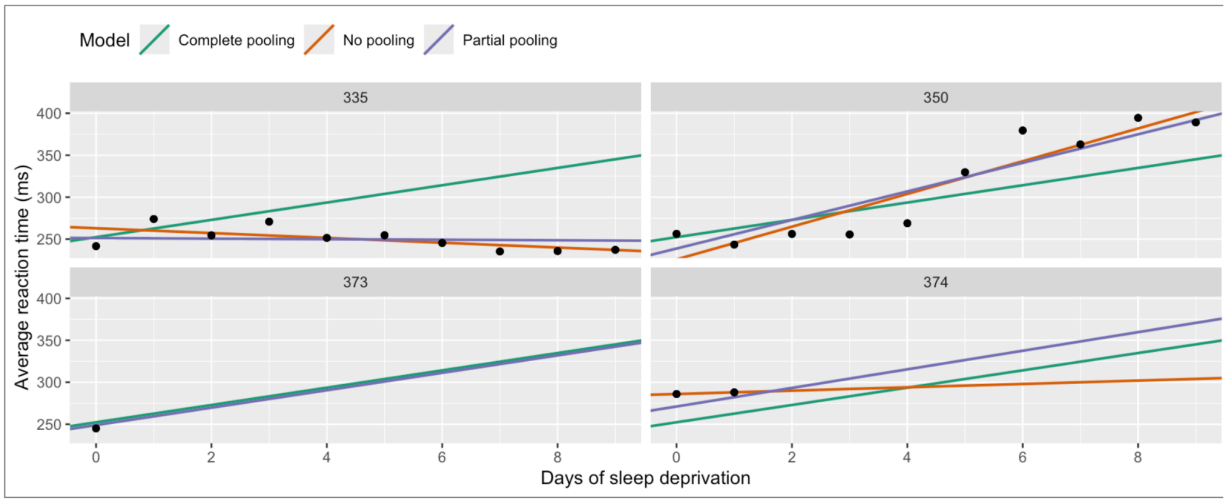
By assuming intercepts and slopes are drawn from a normal distribution, we can generalize findings to other schools.

Example: Sleep Study Data

Three Pooling Strategies

1. **Complete Pooling:** All participants assumed to have same response to sleep deprivation, leading to a single intercept and slope for all.
2. **No Pooling:** Each participant's data is analyzed independently, yielding individual-specific estimates and fitting separate models for each. Maximizes individuality but discards shared information across the population.
3. **Partial Pooling (Random Effects):** Compromise between complete and no pooling, allowing for individual variation while learning from overall population. Applies regularization towards the population mean.

Look for trends across types of pooling in graph below:



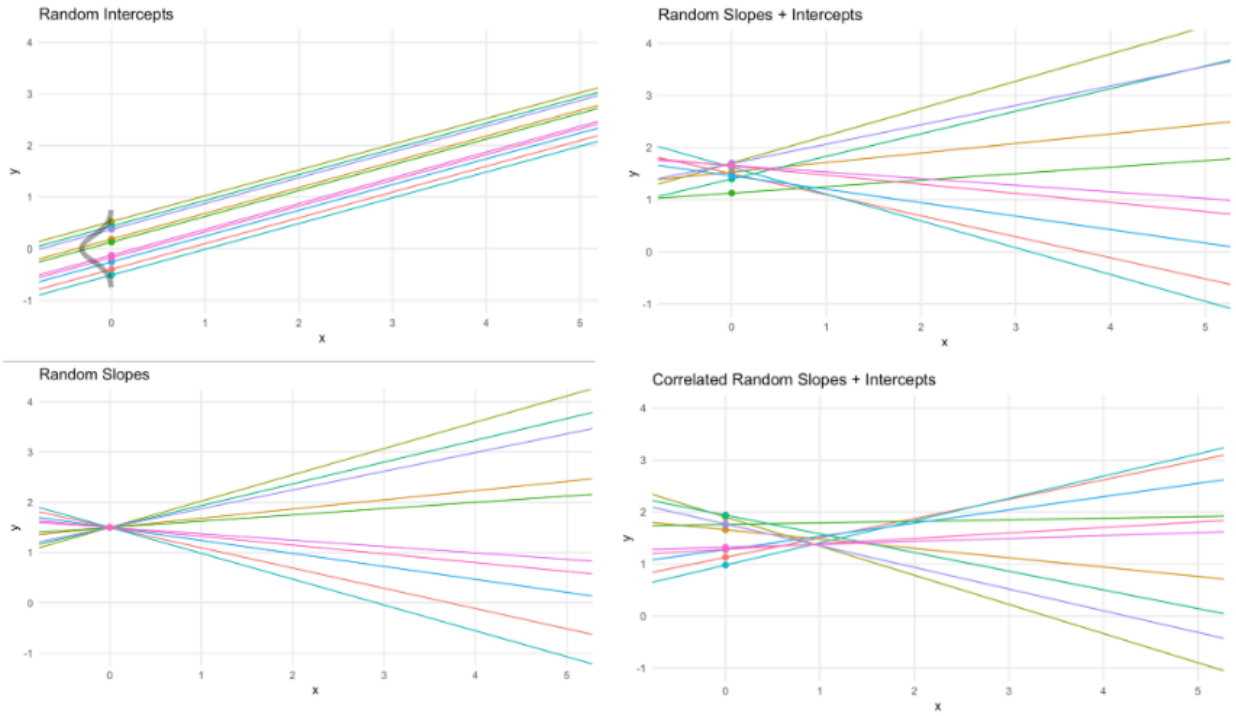
Regularization toward group mean

- Less certain we are (due to little amount of data) or more variance -> pulled more toward mean

Random Intercepts Model: Allows for different starting points (intercepts) but assumes effect of predictor (slope) is same across all subjects.

Random Slopes Model: Allows for different slopes but assumes common intercept across all subjects.

Random Intercepts and Slopes: Combines both to allow subjects to have different starting points and different responses to predictors.



Survival Analysis

Deals with modeling time until an event occurs (e.g., death, customer churn).
Survival Time: time until event of interest occurs.
Censoring: Occurs when event has not been observed for some subjects (e.g., if the study ends before the subject experiences the event).
Kaplan-Meier Curves: Decreasing curve showing how likely it is to survive until time t
Cox Proportional Hazards Model: Estimates death rate instant after time t (based on predictor variables).

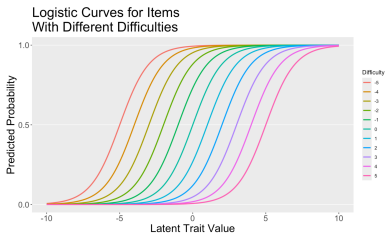
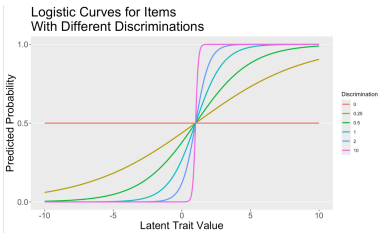
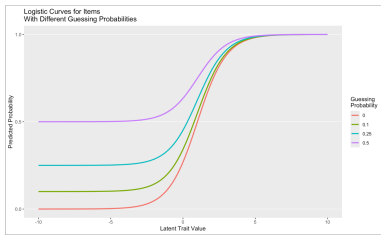
Latent Variable Modeling

Example: spelling test

- difficulty of item
- spelling ability of person

Neither of these things are observed, but we can estimate them using observed correctness of spelling test items.

Rasch (1PL) Model	Rasch (2PL) Model	Rasch (3PL) Model
θ: latent trait value (ability) for person	θ: latent trait value (ability) for person	θ: latent trait value (ability) for person (standard normal dist)
βj: difficulty of item	βj: difficulty of item	βj: difficulty of item
	αj: discrimination of item	αj: discrimination of item

		cj: “guessability” of item 

Higher discrimination=more information an item gives you about someone’s latent trait value

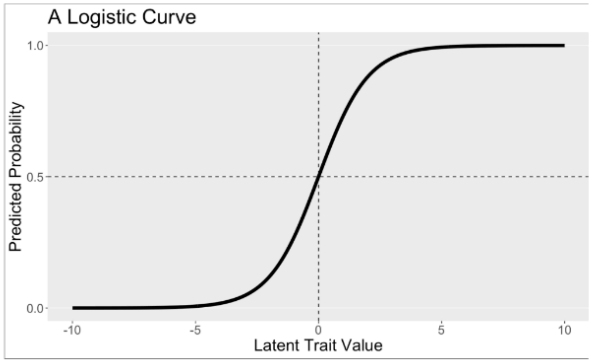
Sum Scores vs. Weighted Scores

When grading exams you can either:

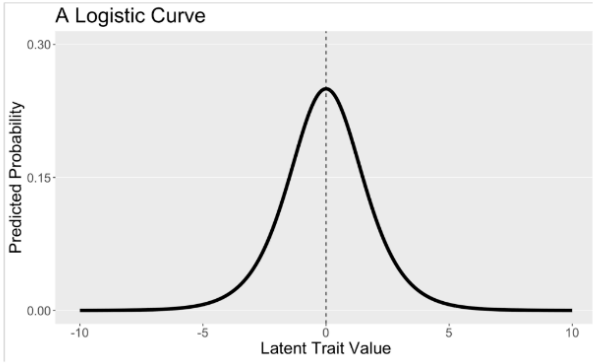
sum score: weight each question equally

- common on tests and surveys
- assume each item gives you same amount of information

weighted score: weight each question differently



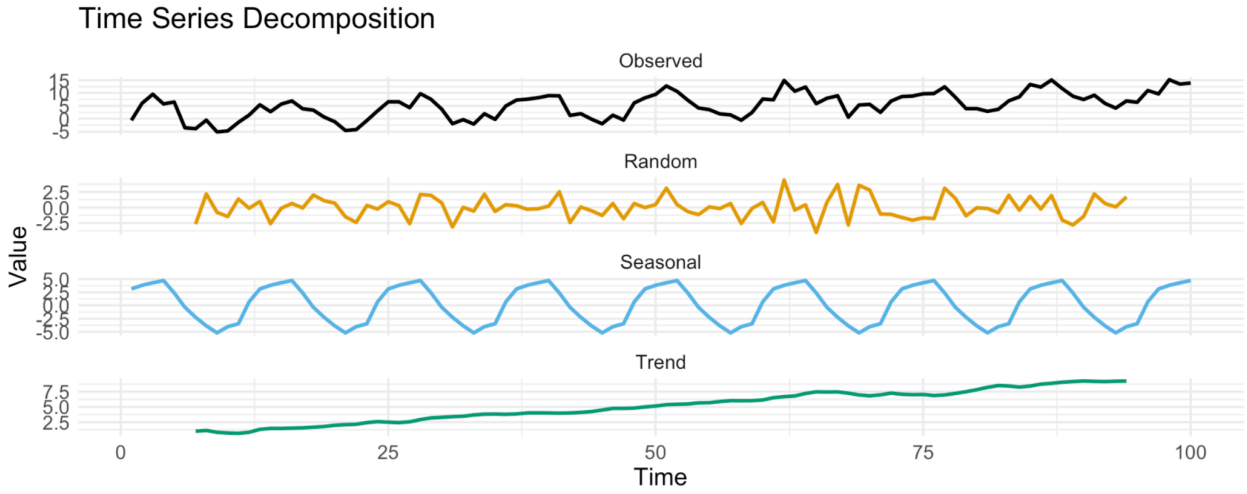
Item Characteristic Curve: steeper slope = more info item gives us about someone’s ability (higher weight it gets in our score)



Item Information Curve: derivative of ICC. How much information an item gives about different latent trait values.

TIME SERIES ANALYSIS

Time series data consists of repeated measurements over time from the same subject or system.
Seasonality: Regular fluctuations due to cyclic patterns like time of year or day of the week.
Trend: Long-term increase or decrease in the data.
Noise: Random variation not explained by trend or seasonality.



Small Time Series: Handled using Mixed Effects Models (MEMs).
Large Time Series: More complex modeling techniques required such as decomposition and ARIMA models.

Autocorrelation and Partial Autocorrelation

Correlation: btwn x1 and y
Autocorrelation: correlation btwn time series and lagged versions of itself.
Autocorrelation Plot: Shows how much past values of time series are related to current value, up to a certain lag (e.g., 30 time steps back).
High Autocorrelation: Indicates strong dependence on previous values. A "fuzzy caterpillar" shape in the plot signifies little to no autocorrelation.
Semi-Partial Correlations: correlation btwn x1 and y, after subtracting influence of x2 on x1
Partial Autocorrelation: correlation btwn y_t and y_{t-2}, after subtracting influence of y_{t-1} on y_t and after subtracting influence of y_{t-1} on y_{t-2}

- Measures direct correlation between the current value and a lagged value, removing the influence of the intermediate lags.

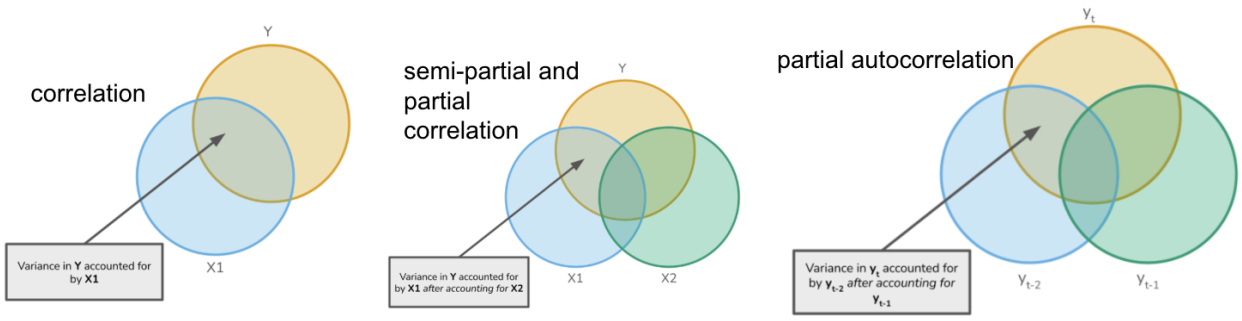
Partial Autocorrelation Function (PACF): Measures correlation between time series and lagged versions of itself, after accounting for correlations at shorter lags.

- Helps determine the number of lags to include in an AR model.

Partial Correlation: correlation btwn x1 and y, after subtracting influence of x2 on x1 and after subtracting influence of x2 on y
Regression coefficients reflect relationship between predictor and outcome, adjusted for other covariates in model.

- For simple linear regression, coefficient is correlation scaled by ratio of standard deviations.

- **Multiple Predictors**: regression coefficients represent semi-partial correlations, which measure relationship btwn predictor and outcome after accounting for other variables.



Stationarity and White Noise

Stationarity: key assumption in time series modeling that allows for more accurate modeling and forecasting.

Stationarity Conditions

1. constant mean
2. constant SD
3. no seasonality (constant autocorrelation)

Examples of Non-Stationary Time Series:

- Changing Variance: Time series with increasing or decreasing variability over time.
- Changing Mean: A time series with a trend where the mean increases or decreases.
- Seasonality: Time series with periodic fluctuations, such as a sine wave pattern.

White Noise = type of stationary time series that meets the following conditions:

1. mean = 0
2. constant SD
3. no autocorrelation

Non-Stationary Time Series

AR Model: use lagged observations as predictors for current observation

- constant autocovariance

Differencing: transform non-stationary time series into stationary one by calculating differences between consecutive time steps

Expected Value and Variance: After differencing, time series may have a constant expected value and variance, making it suitable for time series models.

Autocorrelation and Partial Autocorrelation: Help in identifying appropriate lag structure for AR models).

Moving Average (MA) Models

MA(q): current value predicted based on the error (white noise components) of previous time steps.

- Use past error to help improve/correct current prediction

Additive Time Series Model: Assumes observed time series is sum of trend, seasonal, and noise components

Characteristic Polynomial for AR(p) Models: Generalize stationarity condition by finding roots of the polynomial.

ARIMA

ARIMA(p, d, q): Autoregressive component of p, integrated component (stationarity) of d, and Moving Average component of q
Combines AR and MA components.

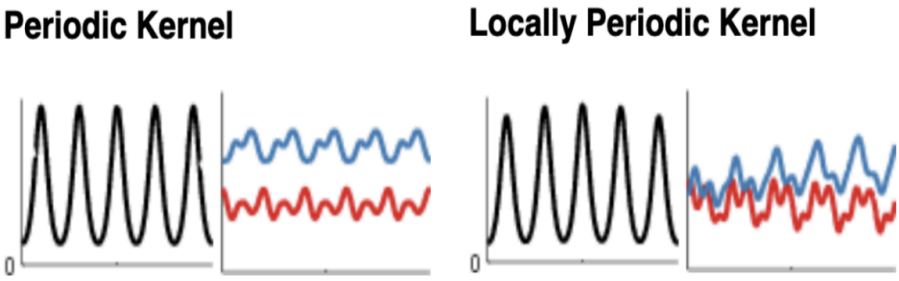
- Includes an integrated (I) component to deal with non-stationary time series.
- Instead of predicting non-stationary time series, we predict the stationary time series composed of differences in lags of time series

Gaussian Processes for Time Series

- Extend normal distributions to functions, where each point in function follows normal distribution, and relationships btwn points are defined by covariance function (or kernel).
- Used for more complex time series modeling where the relationship between data points is non-linear or where there is uncertainty in predictions

Kernel Functions: how close points in time affect one another.

- kernel choice important for capturing the underlying structure of time series
- acts as a prior that encodes assumptions about shape of GP → incorporate observed data by conditioning on it.
- Gives us info on how data covaries



GENERALIZED ADDITIVE MODELS (GAMS)

Extension of Generalized Linear Models (GLMs) that allows for more flexible, nonlinear relationships by fitting smooth functions to predictors.
Purpose: Use when relationship btwn predictors and outcome is not strictly linear, but we want to retain some structure in how variables relate to the outcome.
Non-Linear Extensions in GLMs and Feature Engineering

Creating Nonlinearity in GLMs

1. **Link functions:** Allow the linear model to predict nonlinear outcomes.
2. **Feature engineering:** Manipulating predictor variables (e.g., creating polynomial terms, one-hot encoding categorical variables) to model nonlinear relationships. Anytime you take existing features and transform them to create new ones

Design Matrix: matrix containing all inputs to your model after being processed (e.g. one hot encoding, polynomial basis expansion)

- make more columns from single column; use those columns as predictors

Polynomial Regression

Polynomials: You can raise predictors to powers (e.g., x^2, x^3, etc.) to capture curved relationships in the data.

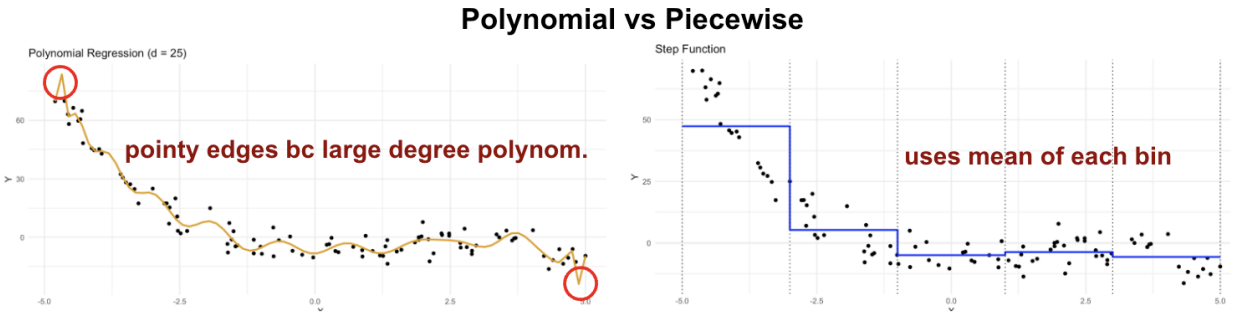
Polynomial regression: Adds polynomial terms to model nonlinear relationships; still a linear model

Ex: simple polynomial transformation can model curvilinear trends, like the effect of vitamin D on health, where effect increases to a point but then decreases.

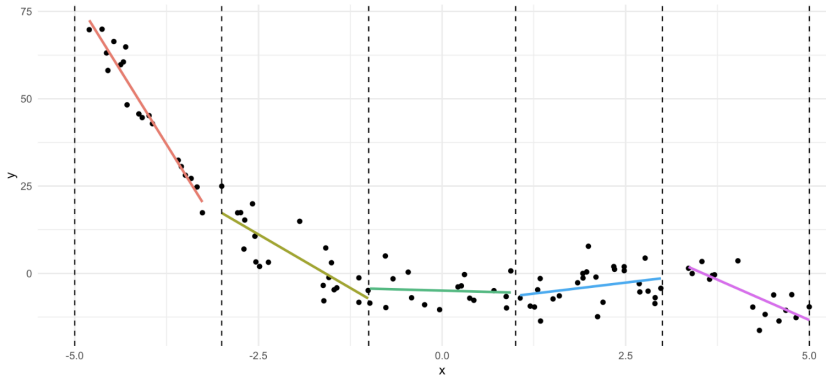
Risks: High-degree polynomials (e.g., x^{25}) can lead to "wiggly" models, meaning overfitting—behave poorly outside data range (leading to unrealistic predictions at the boundaries)

Piecewise Regression

Piecewise regression: Divides data into bins (intervals) and fits separate linear models within each bin.



Dummy Coding: For categorical variables, one-hot encoding or dummy coding used to represent multiple categories as binary columns → transforms categorical variable into numerical form for linear models.



Additive Models

Combines linear modeling with feature engineering by applying smooth functions (**splines**) to predictors, → more flexible model than simple linear regression.

Allow us to retain linearity in the parameters while transforming the predictors nonlinearly

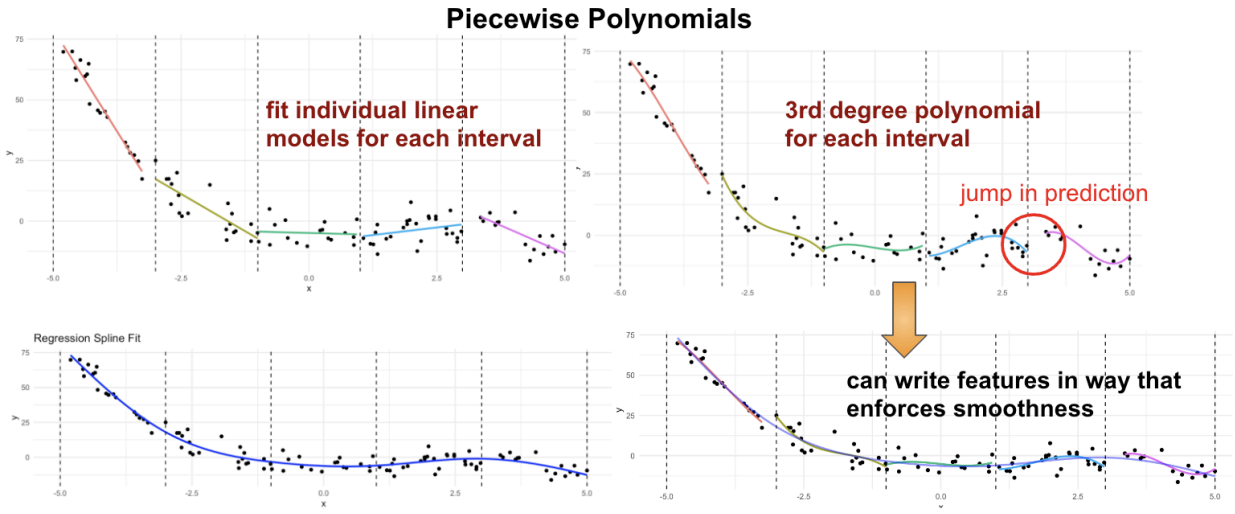
$$y_i = \beta_0 + f(x_i) + \epsilon$$

intercept; error; smooth func. (transforms predictor x)

Knots: Points where model switches between these linear segments.
Continuity at Knots: In some models, you require the lines from adjacent segments to meet at the knots (continuity) so there are no abrupt jumps in model predictions.
Example of Discontinuity: model for income vs. education might show a discontinuity at 13 years of education if modeled w/o continuity constraints.

Regression Splines: Smooth, continuous piecewise polynomial functions of predictor x used to predict y that fit the data in intervals defined by knots.

- **Advantages**: Flexibility w/o requiring high-degree polynomials, reducing risk of overfitting.
- **Ex**: Spline models avoid unrealistic "spikes" often observed (esp. at extremes/ends) in high-degree polynomial models.
- Must choose the right basis functions



Basis Functions and Smoothing

Basis Functions: applied to each predictor to create smooth effects in GAMs.

- Each predictor transformed using set of basis functions, and final smooth curve (actual fitted line)=combo of these transformations (sum of contributions from all basis functions at particular point).
    - Each function multiplied by a coefficient to adjust curve's shape

In piecewise linear models: basis functions represent slope of each segment in model.
For Cubic Splines: create smooth, piecewise polynomial functions across data range.

Controlling Wiggliness in Additive Models

Wiggliness: how much function changes across its range (measured by second derivative).

Penalization: Prevent overly wiggly functions that could result in overfitting (prevent large changes in f'' (rate of change of rate of change)

Regularization Parameter (λ): used to control smoothness of model. Higher values of λ reduce wiggliness by penalizing large second derivatives.

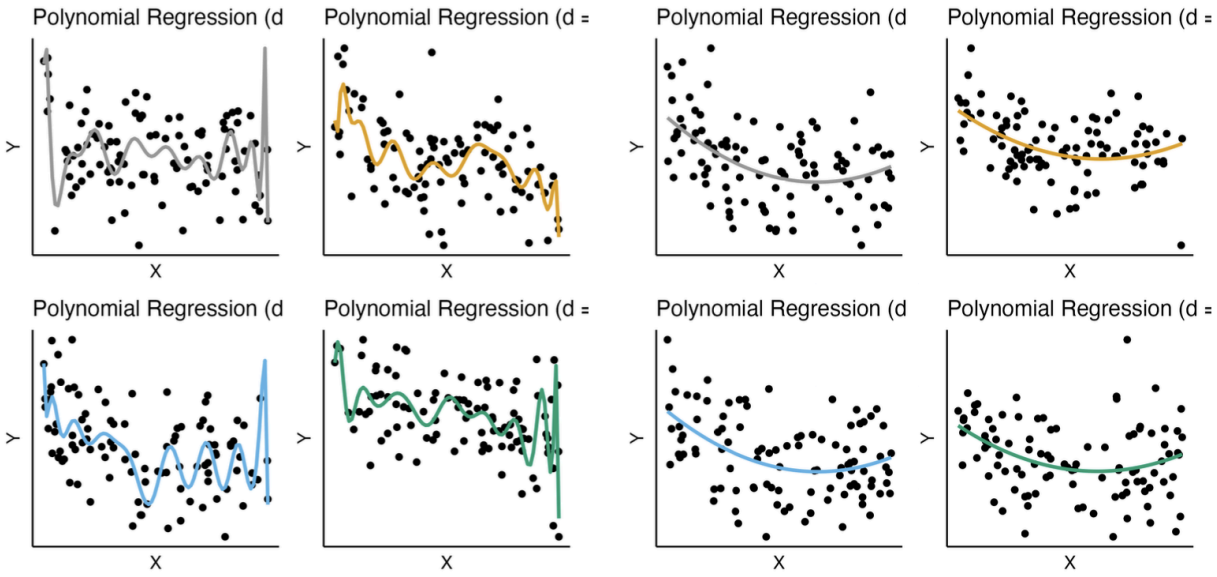
Penalized Likelihood: adjusted by subtracting penalty for wiggliness

- Penalized Likelihood=Likelihood−λ*Wiggliness

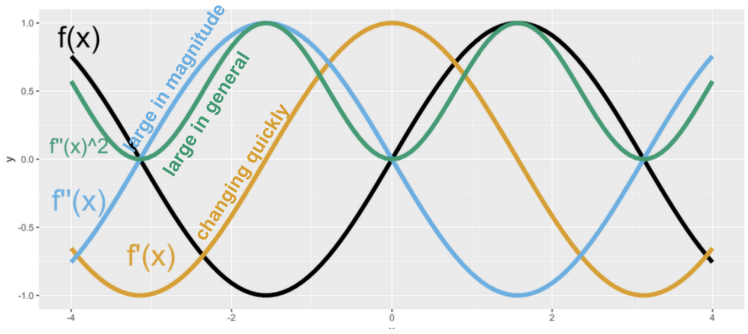
Choosing Regularization Parameter (λ): Larger λ enforce smoother curves by heavily penalizing wiggly functions.

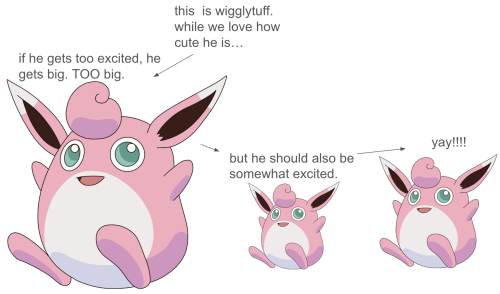
Basis Function Count (k): Controls # basis functions used. More basis functions increase flexibility but also risk overfitting

- # and type of basis functions affect model smoothness.
- Appropriate function = low bias
- Don't overfit (low variance)



(a) High Variance *(b) Low Variance*





JOURNAL CLUB DISCUSSION FROM CLASS (10/9)

---

1. Paper on Synthetic Data in Study of Device Briefing Tool (DBT) in Surgical Settings

Key Focus: Examining impact of a Device Briefing Tool (DBT) on surgical safety and performance using synthetic data.

Study Overview:

- **Purpose**: Assess effect of introducing DBT, which was designed to improve communication and device handling during surgeries, on surgical outcomes
- **Data Used**: Synthetic data based on real hospital records
- **Outcome Measure**: Surgical team performance, specifically non-technical skills (NOTECHS) scores

Key Statistical Methods:

- **Difference-in-Differences (DiD)**:
    - **Used to compare** pre- and post-intervention trends in surgical departments (treatment vs. control)
    - **Assumptions**:

        - **Parallel Trends**: Assumes the trend in outcomes for the control group can serve as the counterfactual for the intervention group
        - **Positivity**: All departments must have had a chance to be assigned to either the treatment or control group
    - **Limitations**:
        - Could not empirically test parallel trends assumption
        - Relied on synthetic data, which raises questions about real-world applicability

Findings:

- **Minimal Negative Effect**: Found a small negative correlation between DBT use and surgical outcomes, but the effect was insignificant
- **Conclusion**: DBT did not have meaningful impact on improving or worsening surgical performance

Challenges and Limitations:

- **Synthetic Data**:
    - Used to protect patient privacy but does not perfectly mimic real-world complexities.
    - **Ethical Concerns**: Can synthetic data be used to guide real clinical decisions?
- **Assumptions**:
    - **Parallel Trends** and **Positivity** are critical, but difficult to verify, especially with synthetic datasets

---

2. Paper on Poisson Regression in Linguistics (Analyzing Gestures)

Key Focus: Application of Poisson regression to analyze count data in linguistics, specifically gestures used in different social contexts.

Study Overview:

- **Purpose**: Evaluate frequency of co-speech gestures by Catalan and Korean speakers based on social context (talking to friend vs. professor)
- **Data**: Count data representing the number of gestures made during speech

Key Statistical Methods:

- **Poisson Regression**:
    - **Used for** count data where outcome = frequency of events (# gestures)
    - **Lambda (λ)**: Model estimates avg # events (gestures) per observation
    - **Link Function**: The logarithm of the expected count modeled as linear combination of predictors (e.g., social context)

Limitations:

- **Overdispersion**:
    - When variance of data exceeds mean, making Poisson model unreliable
    - In linguistics, gesture data often shows greater variability than expected under Poisson assumptions
- **Zero Inflation**:
    - Too many zeros (e.g., no gestures made in some contexts) distort Poisson model.
    - Common issue in datasets where some categories may produce no events for certain individuals or in certain contexts

Findings:

- **Poisson Regression** effectively models count data but requires adjustments for overdispersion and zero inflation.
- **Potential Solutions**:
    - Use of **Negative Binomial Regression** to handle overdispersion
    - **Zero-inflated Poisson models** to address excess zeros

Challenges:

- **Interpretation**:
    - **Log scale** output of Poisson regression can complicate direct interpretation of results
    - Requires careful handling of categorical variables (e.g., social context) to ensure proper model fit

---

3. Paper on Generalized Additive Models (GAMs) for Cardiovascular and Respiratory Data

Key Focus: Using GAMs to model complex, non-linear physiological data, such as cardiovascular and respiratory signals.

Study Overview:

- **Purpose**: Decompose physiological waveforms (e.g., heart and lung interactions) into their respective components using GAMs
- **Data**: Time series of physiological measurements from patients undergoing mechanical ventilation (e.g., pulse pressure, central venous pressure)

Key Statistical Methods:

- **Generalized Additive Models (GAMs)**:
    - **Flexible models** that allow non-linear, smooth functions to represent relationships between variables
    - Useful for modeling **cyclical data**, such as recurring physiological signals (e.g., heartbeats, respiration)

    - **Splines**: GAMs use splines to smooth data, fitting curve rather than straight line to capture non-linearity

Challenges:

- **Cyclic and Non-uniform Data**:
    - GAMs handle cyclic data (e.g., heart-lung interactions) but require special smoothing techniques
    - **Cubic Splines** used to fit smooth functions to physiological signals, but penalizing splines is necessary to prevent overfitting (i.e., too much flexibility)
- **Interactions Between Variables**:
    - GAMs can model interactions btwn continuous variables (e.g., time within the respiratory cycle) and categorical variables (e.g., ventilation type)

Findings:

- **Accurate Decomposition**:
    - GAMs allow for separating effects of cardiac and respiratory cycles in data
    - Useful in medical settings for better understanding physiological signals monitored during surgeries.
- **Potential Application**:
    - Can improve analysis of medical waveforms, such as predicting fluid responsiveness in patients

Limitations:

- **Complexity**:
    - GAMs require careful tuning (e.g., selecting right number of knots for splines) to avoid overfitting
    - Models can become computationally intensive due to flexibility required for fitting smooth functions

---

STATISTICAL INFERENCE

Problem of Inference

- **Inference**: using info/data from sample to draw conclusions about population to learn about real-world relationships
    - **Description**: Summarizing data without making predictions or conclusions
        - Ex: Describing # of blondes in a classroom
    - **Prediction**: Involves predicting outcomes, often used in machine learning
    - **Inference**: Focus on understanding underlying relationships (causal or otherwise) within data

Statistical Inference: Using information (data) from a sample to make conclusions about a population.

- Population often modeled as groups of existing "experimental units" or **Data Generating Process (DGP)**
    - **DGP**: theoretical concept that generates population (where population not just a fixed group but ongoing process, e.g., people named Michael in the USA)
    - Population parameters inferred using sample data (can't directly observe entire population)
- **Ex**: Group of Existing People: All 3.28M Michaels
    - DGP for Michaels: the theoretical process that creates Michaels
    - Inferring the mean height of all people named Michael in the US (3.28 million Michaels vs. an ongoing process that produces Michaels).
    - **Application**: Generalizing findings about a subset (e.g., sampled Michael heights) to a broader population or DGP.

Claim: "I am faster at doing crosswords than you" based on single instance where my time was slightly better than yours

- Problems with such a claim:
    - One data point not sufficient to make strong inferences.
    - Variability in performance could lead to different outcomes
    - observed data could represent any number of underlying distributions or data-generating processes.

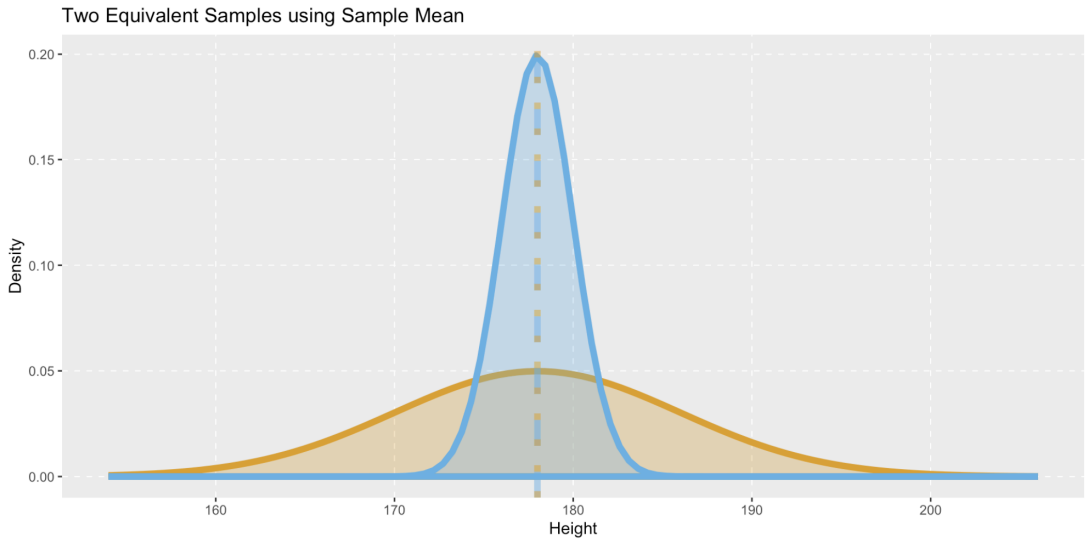
Updated claim: my mean crossword time is faster than yours ($\mu_{me} > \mu_{you}$)

- If we were doing description, sample mean alone suffices (e.g., what is the mean crossword time? my time: 25m 05s; your time: 25m 23s)
- But inference, we want to generalize and know about μ (population), not sample x(bar)

Inference: First Problem

Statistics: functions of data that summarize the data, T(X)

- **Population Statistic**: T(X), where X = random variable (e.g., mean height of people named Michael)
- **Sample Statistic**: T(X), where X = realized sample of X (e.g., mean of 100 randomly sampled heights of people named Michael)
- **Pros**: summarize info about data in easily digestible way, enabling meaningful inferences (ex: New York subway map simplifies complex network)
- **Cons**: often throw out info, potentially losing nuances (e.g., sample mean ignores variability in data).
- When choosing a statistic (e.g., mean), you're implicitly assuming that samples w same mean are equivalent, even though they may differ in variability



Estimation in Statistical Inference

Sample vs. Population

- **Problem**: We can't observe every possible data point in a population → rely on **samples**.
    - Ex: Using sample mean heights of college students to infer true mean height of all college students.
- **Inference Goal**: Move from sample statistics (e.g., sample mean) to population parameters (e.g., population mean, denoted as μ).

Estimation

- **Estimand**: target quantity (e.g., population mean, μ)
- **Estimator**: function/rule (recipe) W(x) used to estimate estimand from sample data (e.g., sample mean)

$$\bar{x} = W(\textbf{heights}) = \frac{1}{N} \sum_{i=1}^{N} \text{height}_i$$

- **Estimate**: actual value produced when applying estimator to sample data (e.g., mean height of 5 sampled Michaels)

$$\bar{x} = W(176, 177, 175, 179, 173) = 176$$

Ex:

- Target: Mean height of all Michaels (estimand).
- Function: Using sample mean formula to estimate.
- Result: Mean height of a sample (e.g., 176 cm).

Choosing Estimators

Two common methods for choosing estimators:

1. **Method of Moments (MOM):** technique where sample moments set = to population moments to create estimators.
    - **Moment**: The expected value of a random variable raised to a power (1st moment = mean, 2nd moment = variance)
    - *review methods section of this document*
        - p^{th} sample moment: $\frac{1}{n}\sum_{i=1}^{n} X_i^p$
        - p^{th} population moment: $\mathbb{E}(X^p)$
2. **Maximum Likelihood Estimation (MLE):** Choose parameter values that maximize likelihood of observed data — **Expectation Maximization (EM)**

    $\arg\max_{\theta} L(\theta|x)$ ← estimate of θ *maximizes likelihood* of data
    - Likelihood: probability of observed data given certain parameter values.
    - **Log-Likelihood**: Used bc logarithms simplify product of probabilities into sum (makes calculus easier)
    - **Ex**:
        - Given observed data points, calculate likelihood of dif parameter values (e.g., mean and variance) to maximize likelihood of the data
        - Normal distribution: MLE for mean = sample mean, and MLE for variance = sample variance

Solution for Normal Distribution:

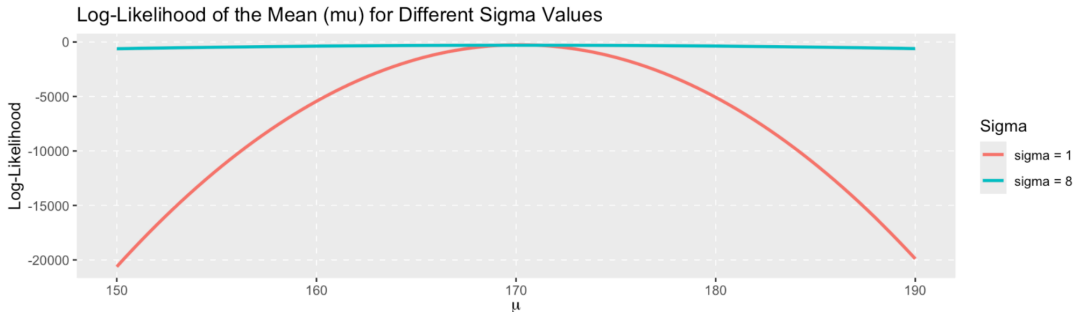
- $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- $\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$

Criteria for Evaluating Estimators

1. **Unbiasedness**:
   - **Intuitive Explanation**: An unbiased estimator doesn't systematically overestimate or underestimate the true parameter.
   - **Mathematical Definition**: An estimator $\hat{\theta}$ is unbiased if $\mathbb{E}(\hat{\theta}) = \theta$
   - **Example**:
     - sample mean $\hat{\mu}$ = unbiased estimator of population mean μ; over repeated samples, its expected value = population mean.
     - sample variance calculated by dividing by n & therefore biased bc underestimates population variance.
       - to correct this, use $n-1$ in the denominator
2. **Consistency**:
   - **Intuitive Explanation**: consistent estimator approximates true parameter better as sample size n increases.
   - **Mathematical Definition**: $\lim_{n\to\infty} \hat{\theta} = \theta$
   - **Example**:
     - Law of Large Numbers ensures that as n grows, sample mean converges to population mean: $\lim_{n\to\infty} \hat{\mu} = \mu$
   - **Weak Law of Large Numbers**:
     - ↑ independent, random samples of X → sample mean of x approaches and converges to expected value:
       for all $\epsilon > 0$, if $\sigma^2 < \infty$
     - $\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$
3. **Efficiency**:
   - **Intuitive Explanation**: efficient estimator has smallest possible variance (provides most precise estimates w minimal data)
   - **Mathematical Definition**:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

   - where $I(\theta)$ is the Fisher Information for θ
   - **Fisher Information**:
     - amount of info a sample contains about a parameter
     - how sensitive log-likelihood function $\ell(\theta|X)$ is to changes in θ



- **Ex**:
  - If everyone in Room A owns the same number of cats, one person's data reveals the entire population's average cat ownership.
  - With a Room B where cat ownership follows a distribution (e.g., Poisson), one individual's cat ownership only partially informs the population mean due to higher variability.

Frequentist vs. Bayesian Approaches

1. **Frequentist Perspective**:
   - Parameters fixed, data is random sample of process P_θ
   - inference relies on repeated samples of X from P_θ
   - Probability: long-run frequency of events.

$$p = \lim_{n \to \infty} \frac{k}{n}$$

2. **Bayesian Perspective**:
   - Parameters are random, data is fixed.
   - Inference based on updating prior beliefs using observed data
   - Probability quantifies uncertainty about parameters based on the data.

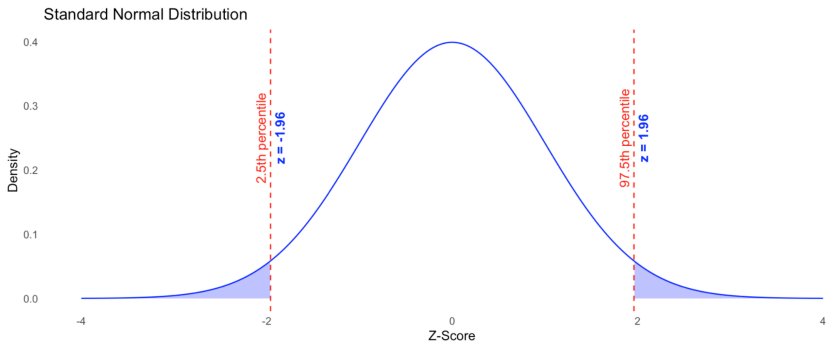
Frequentist Confidence Intervals (CI)

1. **Definition**:
   - interval estimate that, if repeatedly calculated over multiple samples, would capture true parameter $1-\alpha$ % of the time.
   - **General Form**: Point Estimate±Margin of Error
   - **Interpretation**: We are confident in the procedure used to calculate CIs, knowing that $1-\alpha$% of them would contain θ over many samples.
2. **Constructing Confidence Intervals**:
   - Depends on:
     - **Sampling Distribution** of the estimator
     - **Variability in the Data** (σ)
     - **Sample Size** (n)
     - **Desired Confidence Level** $(1-\alpha)$
   - **Known Variance**:
     - When population variance σ^2 is known and n≥30, CI for the mean μ is

$$\underbrace{\bar{x}}_{\text{Point Est}} \pm z_{\alpha/2}\underbrace{\left(\frac{\sigma}{\sqrt{n}}\right)}_{\text{Margin of Error}}$$

- $\bar{x}$: Sample mean
- $z_{\alpha/2}$: Critical value from the standard normal distribution
- n: Sample size

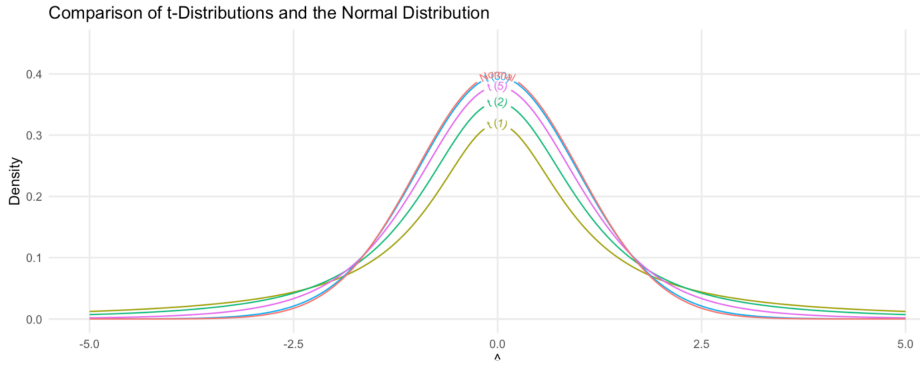
$$\alpha = 0.95; z_{\alpha/2} = 1.96$$



- **Unknown Variance**:
  - For smaller samples and unknown variance, use t-distribution (heavy tails)

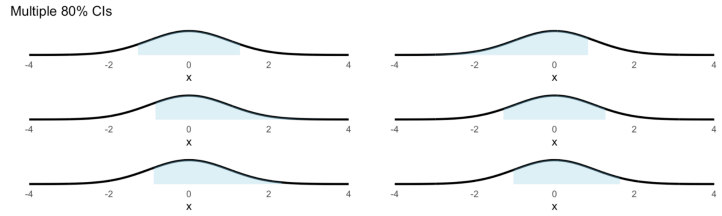
$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$$

  - s: Sample standard deviation
  - $t_{\alpha/2,n-1}$: Critical value from the t-distribution with $n-1$ degrees of freedom
  - implies two sources of uncertainty: uncertain abt sample mean and sample SD



3. **Evaluating Confidence Intervals**:
   - **Coverage**: well-constructed CI has good nominal coverage (matching confidence level), with adjustments for sample size and distribution shape as needed
     - **Nominal** Coverage: $1-\alpha$
     - **Actual** Coverage: $P(lb \leq \theta \leq ub)$
   - **Precision**: narrower CI (w consistent confidence level) preferred; implies greater certainty in parameter estimate

Many intervals where $P(lb \leq \theta \leq ub) = 1 - \alpha$



- Choose the *narrowest* one.

4. **Interpretation Nuances**:
   - Correct CI interpretation focuses on long-run error rates, not probability of any specific interval containing θ.
   - **"Act as if" Principle**: Frequentist CI methods control long-run error rates; therefore, acting as if interval contains θ is reasonable since error rate is known and controlled.

Interval vs. Point Estimates

- **Point Estimate**: Single value (e.g., sample mean) as a best guess for population parameter.
- **Interval Estimate**: Range that provides likely span within which parameter lies (clearer picture bc including uncertainty)

Bootstrapped Confidence Intervals

- **Purpose**: Provides a flexible approach to constructing confidence intervals when traditional assumptions about the population distribution may not hold.
- **Process**:
  1. **Resampling**: Randomly sample w replacement from original dataset, treating sample as approx representation of population.
  2. **Bootstrap Samples**: Create multiple "new" samples (bootstrap samples), each same size as original dataset.
  3. **Calculate Statistic** (e.g., mean, median) for each bootstrap sample.

- **Result**: Produces empirical sampling distribution of the statistic, allowing for CI construction less dependent on strict distributional assumptions

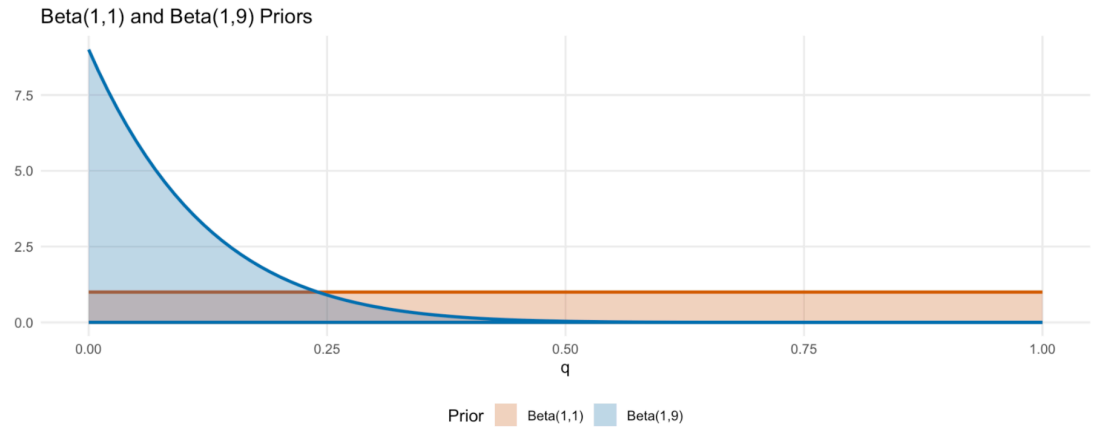
Bayes Rule By Hand

https://cmparlett.shinyapps.io/betaprior/

Beta-Binomial

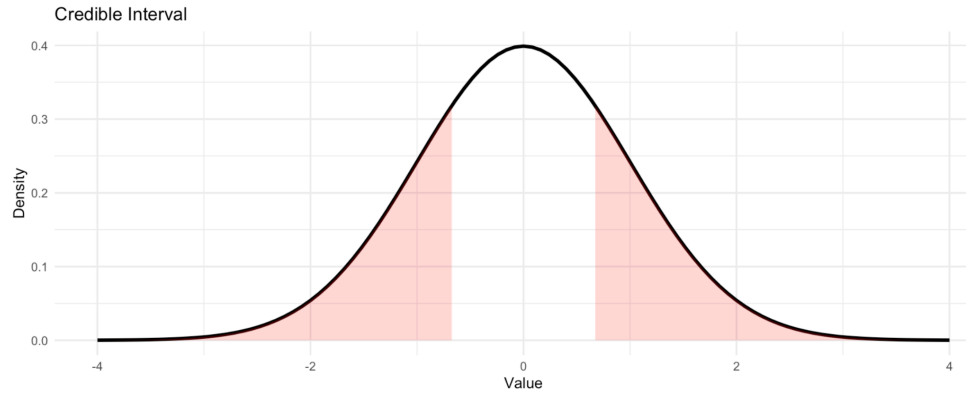
We're interested in estimating q the proportion of days it rains in California. It rained 12 of the last 365 days.

- **Binomial Likelihood:** *[Math Processing Error]*
- **Beta Prior:** $p \sim \text{Beta}(\alpha, \beta) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)}$

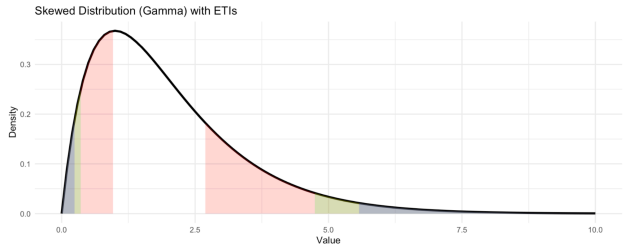


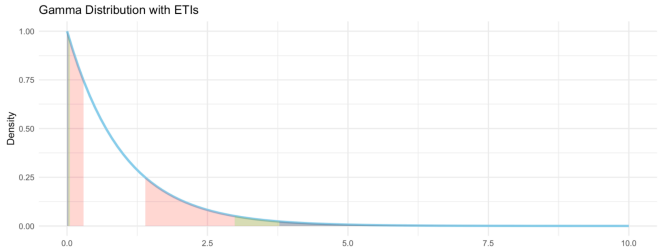
Bayesian Interval Estimation and Summarizing Posteriors

- **Point Estimates and Interval Summaries**:
  - Bayesian **point estimates** derived directly from posterior, using **mean** $\mathbb{E}(p(\theta|x))$ or **median** of posterior samples
  - **Credible Intervals**: "There is a c% chance that θ is between these values"
  -



  - **Equal-Tailed Interval (ETI)**: middle c% of posterior probability, splitting remaining probability symmetrically across both tails

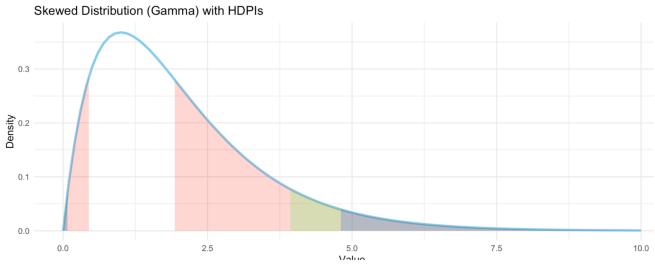


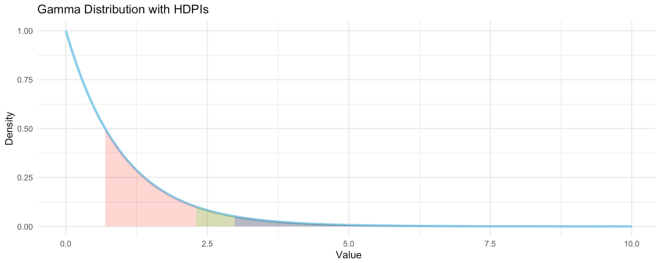


- **Highest Density Interval (HDI)**: shortest interval containing c% of probability mass -> highest density of values within interval
  - useful for skewed distributions (equal-tailed may not be optimal)

  Step 1: Identify the density threshold k such that the interval $\theta : p(\theta \mid x) \geq k$ contains $1 - \alpha$ probability mass.

  Step 2: define the interval $[lb, ub]$ that contains all θ with $p(\theta \mid x) \geq k$





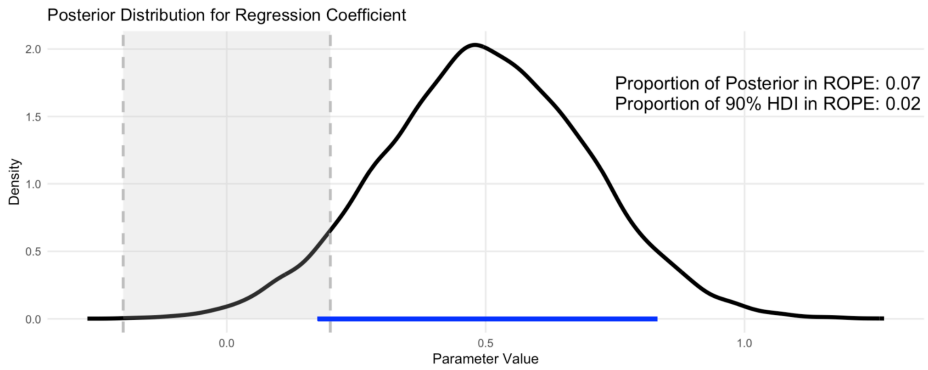
- can use posterior samples from Bayesian model to calculate summary statistics of transformed variables (e.g., squaring parameter values) by transforming posterior samples before summarizing them

Feature	Equal Tailed Interval (ETI)	Highest Density Interval (HDI)
Calculation	Splits excluded values equally between tails.	Contains highest density regions within specified probability mass.
Symmetry	Ideal for symmetric distributions; can be skewed in asymmetric cases.	Adapts to asymmetry by focusing on densest posterior regions.
Mode Inclusion	Does not necessarily include the mode.	Always contains the mode or modes of the posterior distribution.
Interval Width	Width can be greater if posterior is skewed.	Generally narrower for a given confidence level, especially in skewed distributions.
Interpretation	Similar to frequentist Confidence Intervals (CI).	Represents the most credible values within highest density.
Ease of Calculation	Easier to calculate, straightforward approach.	More complex; requires density threshold identification.

Region of Practical Equivalence: range of values that are practically equivalent to no effect

1. Define a ROPE: use domain expertise or a "standard" small value like (1/10)*SD
2. Calculate what % of your Posterior CI overlaps with ROPE

- lots of overlap=evidence for practical equivalence
- little overlap=evidence for non-equivalence
- **Smallest Effect Size of Interest:** smallest effect size that would be meaningful, clinically relevant, or impactful



Null Hypothesis Significance Testing

Reductio Ad Absurdum (Reduction to Absurdity): form of proof by contradiction

- We want to prove X, assume not X, show that it leads to a false, ridiculous, or highly unlikely outcome. Therefore X.
- Steps: 1) claim 2) assume contradiction 3) RAA 4) conclusion

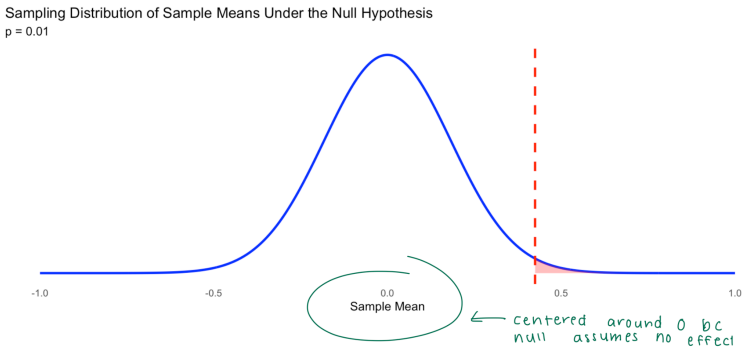
Null Hypothesis: Any hypothesis of "no effect"

Significance Testing: Evaluating likelihood of a hypothesis

Test-Statistic: summary of the data calculated using a sample (stat we do hypothesis testing on).

P-Values $p = p(\text{test statistic} \mid H_0)$

- assuming the null is true and there's no effect, what is P(observing a test-statistic as or more extreme than one we calculated from our data)
- higher p-value=data more compatible w null



Common Misconceptions about P-values:

- p-values are not P(null is true)
- p-values are not P(effect will replicate)
- non-significant p-values do not mean null is true

Directional vs. Non-directional Nulls:

- **Directional Nulls**: Hypothesize effects only in specific direction (e.g., "exercise increases lung capacity") $\mu \geq 0$
- **Non-directional Nulls**: Indifferent to direction of effect, just concerned w/ if effect exists. $\mu = 0$

Frequentist Hypothesis Testing Approaches

Parameter Estimation vs. Hypothesis Testing

- **Parameter Estimation**: **estimating values** of parameters using data, providing point estimates or confidence intervals.
- **Hypothesis Testing**: used to **support theory or hypothesis** about data
    - test if params meet specific criteria (e.g., mean = zero)

Fisherian and Neyman-Pearson Hypothesis Testing

- **Fisherian Hypothesis Testing**:
    - Uses **p-values** as continuous measure of evidence against an H0
    - answers: “is the observed data consistent with null?”

- **Neyman-Pearson Framework**:
    - Divides hypothesis testing process into **null** H0 and **alternative** H1 hypotheses
    - **binary decision** to reject or not reject null
    - Control **type I (false positive)** and **type II (false negative)** errors
    - **critical values** or **significance levels** (commonly $\alpha = 0.05$) to control error rates in long run
    - **FTR H0:** we have not provided evidence that is false, we will not act as if it’s false
    - **Reject H0:** we have provided evidence that is false, we will act as if it’s false

Aspect	Fisherian Hypothesis Testing	Neyman-Pearson Significance Testing
Purpose	Evaluate evidence against H_0 using p-value	Make a decision between H_0 and H_A using test statistic
Steps	1. Choose test 2. Define H_0 3. Calculate p-value 4. Assess significance based on p-value	1. Choose test 2. Define H_0 and H_A 3. Calculate test statistic and critical value 4. Assess significance based on critical value
Role of p-value / Critical Value	p-value is a continuous measure indicating the strength of evidence against H_0	Critical value is a threshold for decision-making, comparing results to a set cut-off
Decision Rule	Reject H_0 if p-value is sufficiently low	Reject H_0 if test statistic is more extreme than the critical value
Error Rates	Does not focus on binary decision making or control error rates	Controls Type I and Type II errors through significance level (α) and power (β)
Philosophy	Inductive inference; aims to measure evidence without making definitive decisions	Frequentist perspective; focuses on long-run error rates and decision rules
Outcome Type	Provides a measure of evidence against H_0, not a strict decision	Provides a binary decision to accept or reject H_0

Modern Null Hypothesis Significance Testing (NHST)

- **NHST as a Hybrid**:
    - Combines aspects of both Fisherian and Neyman-Pearson frameworks, using **p-values** in a decision-making capacity.
    - common misconception: interpreting p-value as probability that null is true, **actually** indicates probability of obtaining results as extreme as observed if H0 holds.

Frequentist Approaches: Type I & II Errors, and Familywise Error Rate

- **Error Types** (remember we get to choose alpha directly)

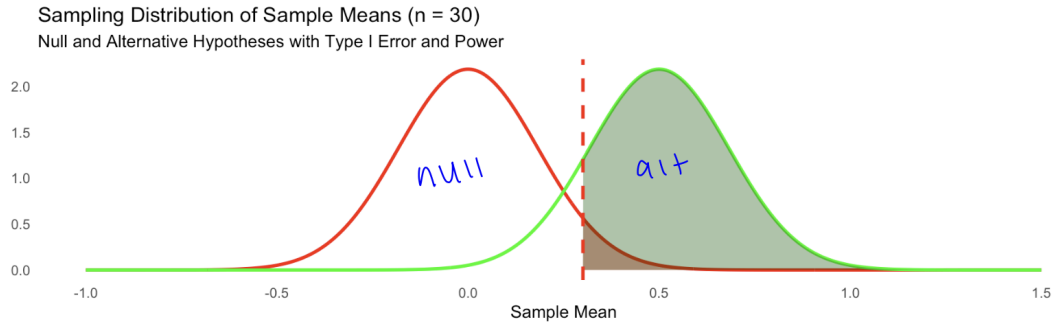
	Fail to Reject H0	Reject H0
H0 True	Correct	Type I Error; FP
H1 True	Type II Error; FN	Correct

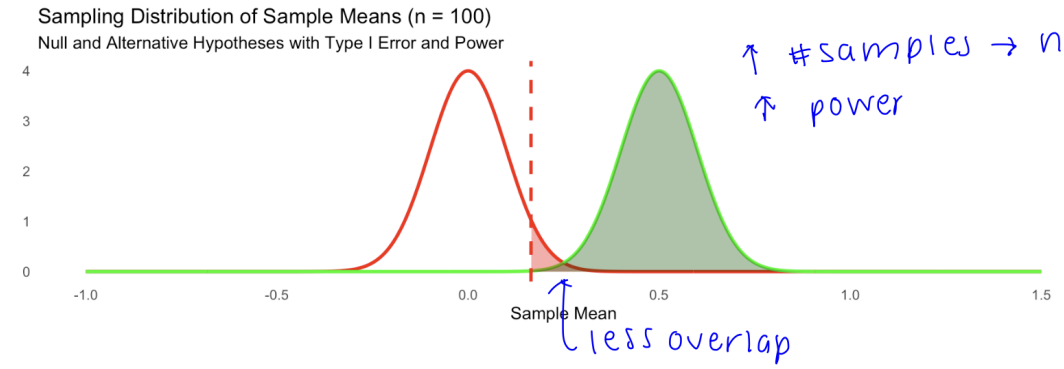
	Fail to Reject H0	Reject H0
H0 True	$1 - \alpha$ Correct	α Type I Error
H1 True	β Type II Error	$1 - \beta$ Power

Power analysis: if there is an effect, how likely are you to detect it (beta); affected by 4 things:

1. sample size
2. population standard deviation
3. effect size
4. α

Power Analysis: n

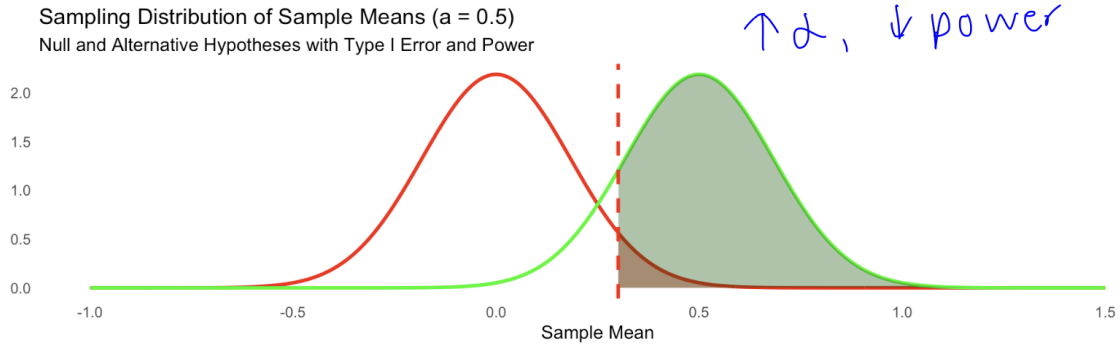




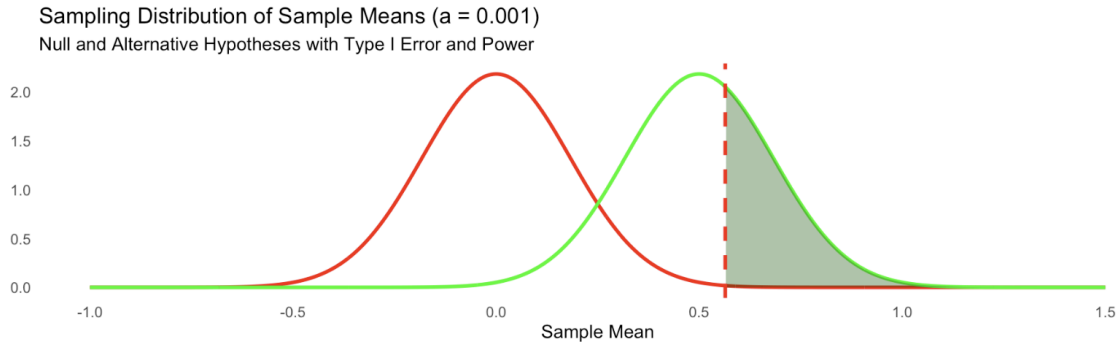
↑ #samples → narrow,
↑ power

↑ less overlap

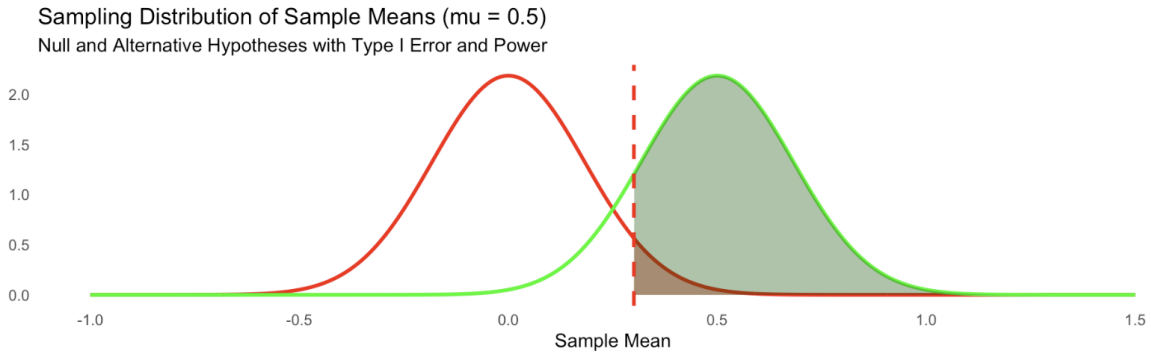
Power Analysis: α

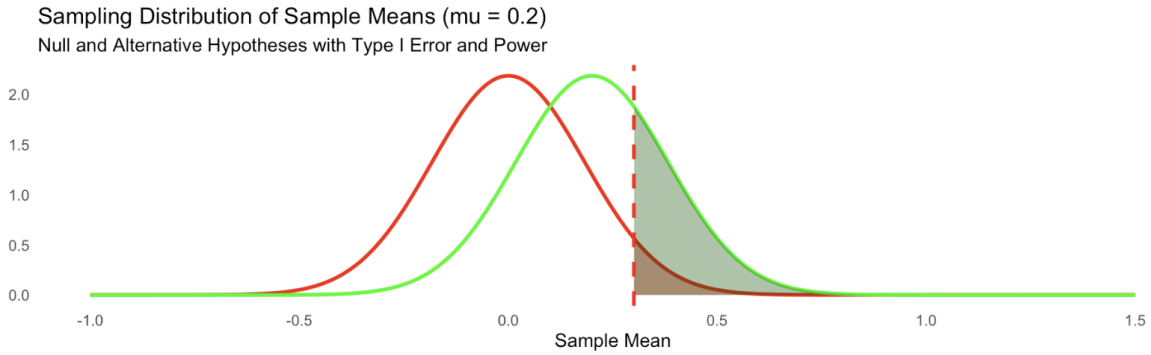


↑ α, ↓ power



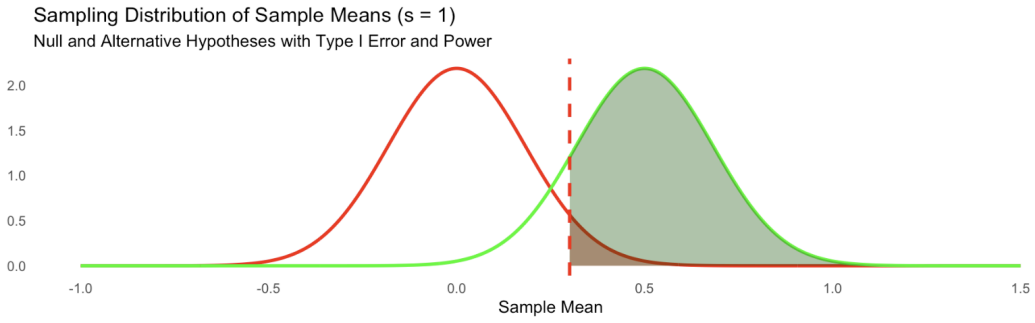
Power Analysis: Effect Size

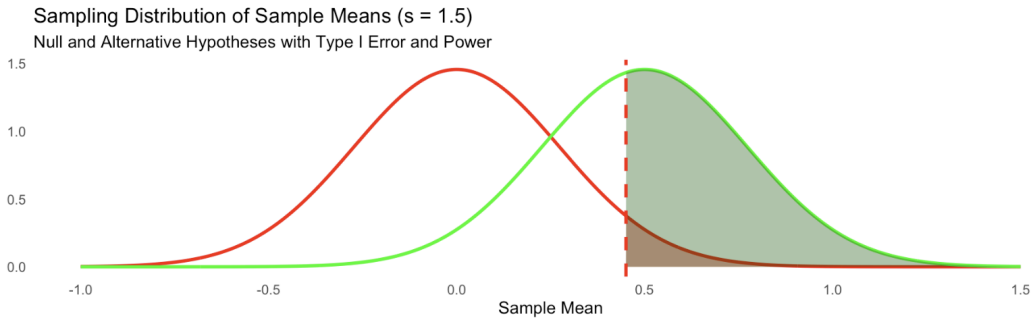




Power Analysis: σ

↑σ, wider sampling dist, ↓ power



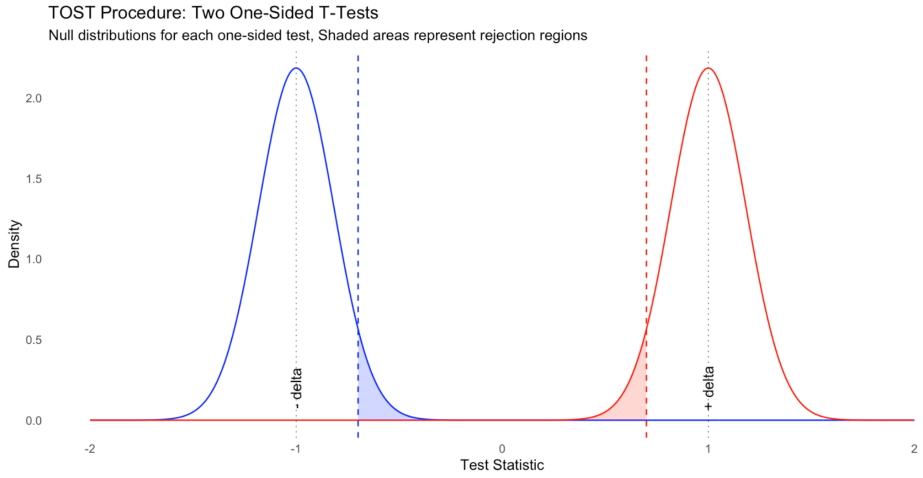


- **Familywise Error Rate**:

- W/ multiple comparisons, familywise error rate increases, potentially inflating chance of at least one Type I error.
- Adjustments:
  - **Bonferroni:** $p_{thresh} = \frac{\alpha}{m}$; where m is the number of tests
  - **Sidak:** $p_{thresh} = 1 - (1 - \alpha)^{\frac{1}{m}}$; where m is the number of tests

Equivalence Testing in Frequentist Analysis

- **Equivalence Testing**:
  - Allows researchers to provide evidence for **null hypothesis** by defining range of values considered practically equivalent to no effect (e.g., mean differences between -1 and +1).
  - **Two One-Sided Tests (TOST)**: Tests whether observed effect falls within this equivalence range, enabling evidence in favor of null hypothesis under pre-specified bounds.



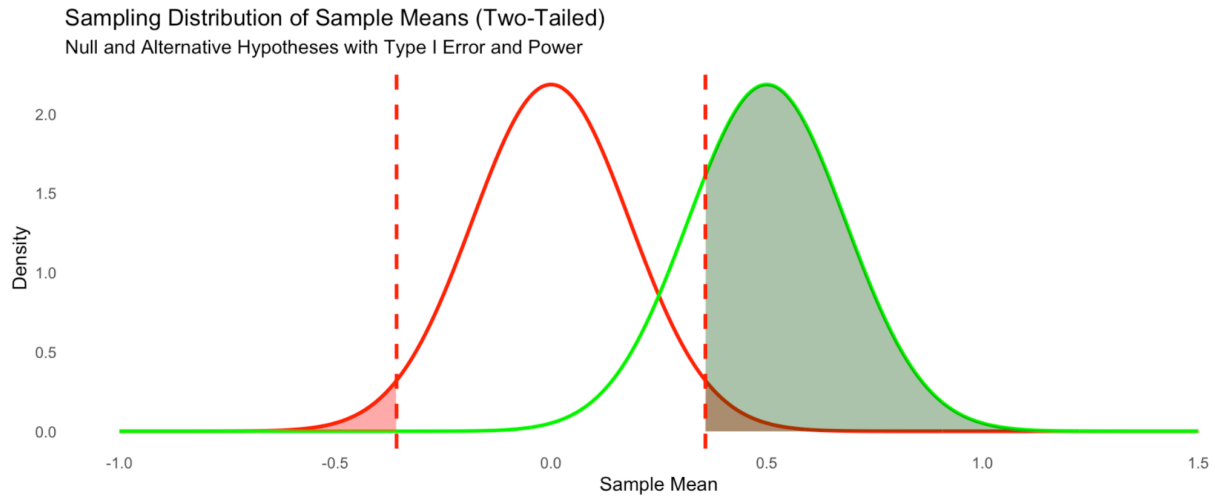
Summary: Bayesian vs. Frequentist Perspectives on Uncertainty

- **Bayesian Perspective**:
  - Uses probability distributions (posterior) to quantify uncertainty about parameter values, with priors incorporating subjective beliefs or domain expertise.
  - Focuses on updating beliefs and allows for direct probabilistic statements about parameters.
- **Frequentist Perspective**:
  - Often centers on hypothesis testing via NHST, with p-values and confidence intervals reflecting long-term probabilities of errors.
  - Lacks direct probabilistic interpretation of parameters but offers robust tools for error control in decision-making.

Key Concepts in Hypothesis Testing

- **One-sided vs. Two-sided Tests**:
  - **One-sided Test**: Concentrates Type I error rate (alpha) on single tail.
  - **Two-sided Test**: Splits alpha equally btwn both tails of sampling distribution.

- Ex: Testing if listening to negative comments affects depression scores. If both increases and decreases in scores are of interest, use two-sided test



Errors in Hypothesis Testing

- **Type I Error (False Positive)**: Incorrectly rejecting a true null hypothesis.
- **Type II Error (False Negative)**: Failing to reject a false null hypothesis.
- **Familywise Error Rates**:
  - When multiple hypotheses are tested simultaneously, the cumulative probability of at least one Type I error increases.
  - **Bonferroni Correction**: Divides alpha by the number of tests to control the familywise error rate.
  - **Sidak Correction**: Uses a more precise calculation, 1−(1−α)1/m1 - (1 - \alpha)^{1/m}1−(1−α)1/m, where m is the number of tests.

Hypothesis Testing with Confidence Intervals

- CI = intuitive way to understand hypothesis tests.
  - If CI includes null value, FTR null.
  - Else, reject null.

Practical Issues in Hypothesis Testing

- **P-hacking**:
  - Manipulating statistical analysis to obtain p-value below given threshold (e.g., 0.05).
  - Commonly done by altering model specifications (e.g., adding or removing covariates).
  - Can cause incorrect conclusions bc invalid statistical inferences.

Equivalence Testing

- **Objective**: support the null hypothesis within defined range of practical equivalence.
- **Two One-Sided Tests (TOST)**:
  - two one-sided t-tests to show that observed effect falls within a specified range around zero.
  - **Interval Null Hypothesis**: Defines an acceptable range (e.g., effects between -1 and 1 are considered equivalent to zero).

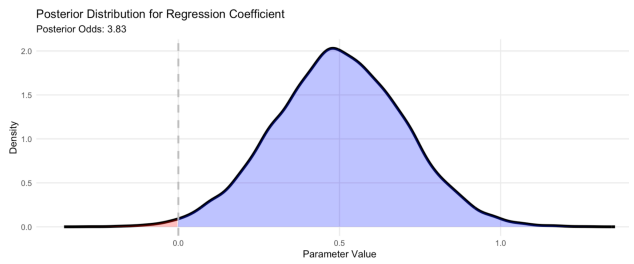
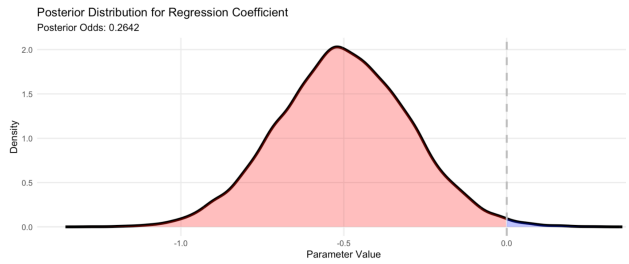
- **Interpretation**:
  - If both tests reject the null (i.e., effect is greater than the lower bound and less than the upper bound), then evidence supports practical equivalence to no effect.

Bayesian Hypothesis Testing

Three main ways to test hypotheses in Bayesian framework:

1. Check if a Credible Interval overlaps with value(s) of interest
2. Posterior Odds
3. Bayes Factors (Likelihood odds)

Posterior odds

Example 1	Example 2
$H_0 : \theta \leq 0; H_A : \theta > 0$ **Posterior Odds:** $\frac{P(H_A \mid \text{data})}{P(H_0 \mid \text{data})} = 3.83$	$H_0 : \theta \leq 0; H_A : \theta > 0$ **Posterior Odds:** $\frac{P(H_A \mid \text{data})}{P(H_0 \mid \text{data})} = 0.79$
interpretation: HA ≈ 4 times more likely than H0	interpretation: HA ≈ 0.2642x more likely than H0; H0 is ≈ 3.785 (1/0.2642)x more likely than HA ❗ We just provided evidence for the Null
	

Bayes Factors (Likelihood odds)

- **Posterior odds:** compare posterior probabilities of two hypotheses (after seeing data)
- **Prior odds**: compare prior probabilities of two hypotheses (before seeing data)
- **Bayes Factors**: how much data changes prior odds when transforming to posterior odds

$$\underbrace{\frac{p(H_A \mid \text{data})}{p(H_0 \mid \text{data})}}_{\text{Posterior Odds}} = \underbrace{\frac{p(\text{data} \mid H_A)}{p(\text{data} \mid H_0)}}_{\text{Bayes Factor}} \cdot \underbrace{\frac{p(H_A)}{p(H_0)}}_{\text{Prior Odds}}$$

- $BF < 1$: seeing the data made H_A *less* plausible
- $BF = 1$: seeing the data did not change the relative plausibility of H_A
- $BF > 1$: seeing the data made H_A *more* plausible

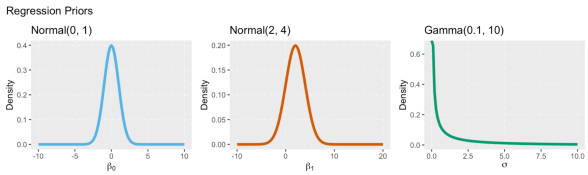
Bayesian Regression Models

Setting Priors in Bayesian Regression

Choosing BLR Priors

$\beta_0 \sim \mathcal{N}(0,1); \beta_1 \sim \mathcal{N}(2,4); \sigma \sim \Gamma(0.1,10)$

We put priors on (*at least*) all parameters in our model.

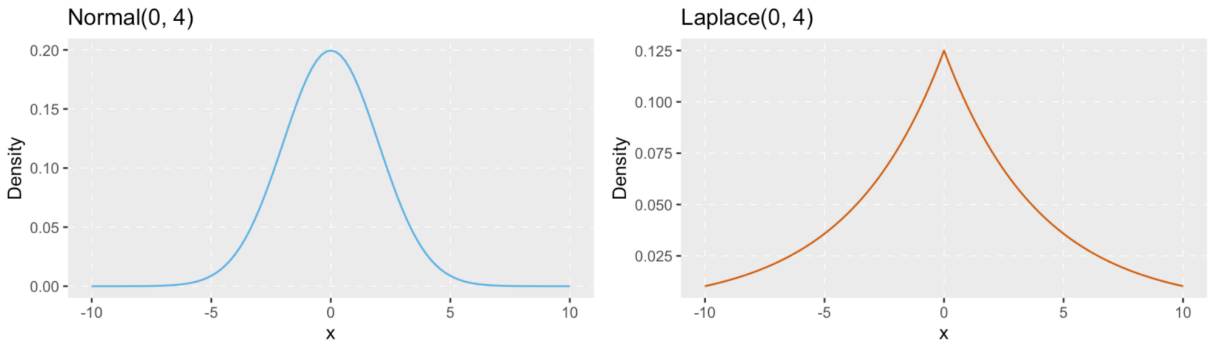


- **Purpose of Priors**: Reflect beliefs about parameter values before seeing the data.
  - Ex: **Age coefficient for income prediction** should reasonably avoid extreme values like -100 or billions.
  - **Common Priors**: Normal for coefficients, Gamma for variances.
- **Distribution Choices**:
  - **Normal**: Common due to symmetry around zero.
  - **Student-t or Cauchy**: Chosen for "fat tails" when extreme parameter values are possible.

Uniform Priors

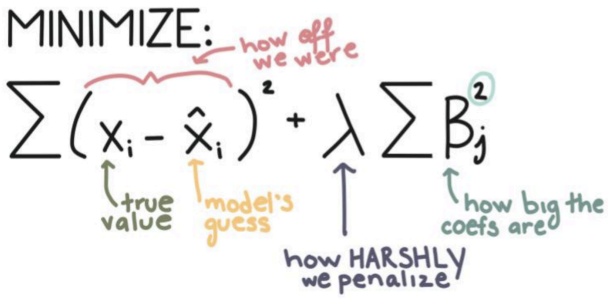
Regularization (in general): discourages overfitting, makes models simpler

Regularizing Priors: keeps parameter values in "reasonable range"



Ridge = Normal Prior, Lasso = Laplacian Prior

Ridge



Other Types of Priors

Aspect	Uninformative/Non-informative Priors	Weakly Informative (Regularizing) Priors	Strongly Informative Priors
Definition	Priors with minimal to no specific information about parameter values	Priors that limit extreme values without strong directional bias	Priors with substantial, specific information about parameters
Purpose	Represent minimal knowledge; aim for objectivity	Regularize to avoid extreme values; provide mild guidance	Encode strong prior beliefs based on existing knowledge
Common Forms	Uniform, Jeffrey's (based on Fisher information)	Normal (centered around zero), Laplacian/Lasso (sparsity-inducing)	Normal, Student-t, or Cauchy with specific means and variances
Examples of Use	Used when there is little to no prior knowledge	Useful when extreme values are unlikely but specifics are unknown	Used in fields with established research (e.g., smoking and lung health)
Impact on Posterior	Minimal influence; lets data "speak for itself"	Moderates posterior, discouraging extreme parameter values	Strong influence, potentially overpowering data
Pros	Simplicity; often seen as "objective"	Prevents overfitting; balances flexibility with realism	Captures rich prior information, aiding convergence
Cons	Can ignore valuable prior knowledge; can be improper (e.g., unbounded)	Limited information; less objective than uninformative priors	Risk of overpowering data; not suitable for new/unexplored areas
Application in Modeling	Often default in Bayesian software	Common in machine learning (Ridge/Lasso regularization)	Used with well-documented parameters or expert input
Interpretation Complexity	Easy to explain (equally likely values)	Slightly more complex, involving trade-offs	High complexity; requires expertise to justify and interpret

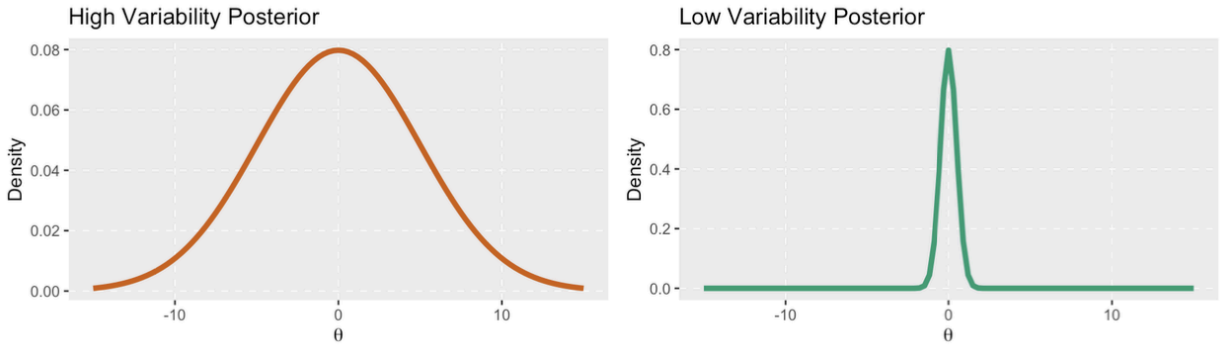
Choosing Priors

- Choose priors compatible with the parameter's range (e.g., only positive for variance).
- Reflect uncertainty with wider priors for less known parameters.
- Consider regularizing priors that avoid extreme values but don't strongly constrain.

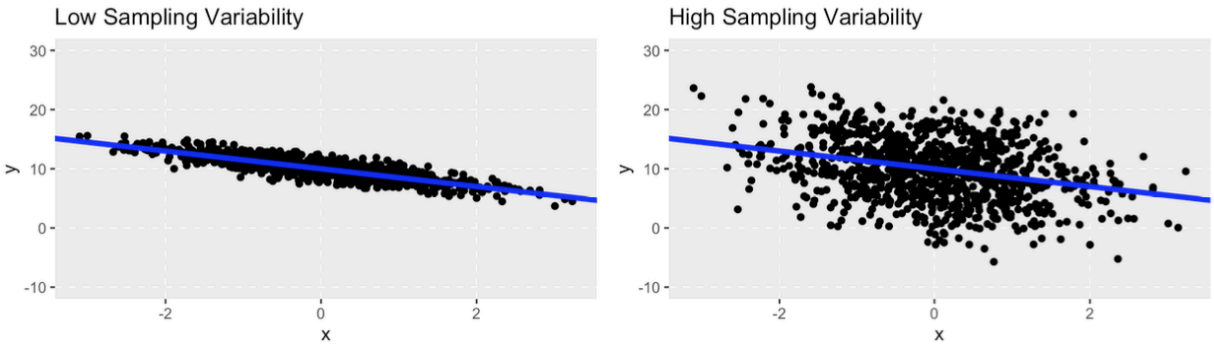
Predictive Distributions

Two sources of variability:

1. Distributional (Prior or Posterior) Variability: How certain are we about the value of the parameters?



2. Sampling (Data) Variability: How much will observed data typically vary from its expected value? (inherent variability in data around true value)



Prior predictive check: simulate samples from prior and make sure it lines up with expectations about reality

Posterior predictive check: simulates new data based on the posterior

- look for systematic discrepancies between real and simulated data (tells about model fit)

What to do with outputs of Bayesian Models:

1. **Manipulating Posterior Draws:** <u>often values more interpretable and reportable to general audience</u>
    - usually use some kind of sampling algorithm like MCMC to get posterior draws from posterior distribution.
        - draw: one individual sample from posterior distribution, contains one value per parameter in θ
        - chain: Markov chain to generate samples from θ, each iteration in chain produces one draw
    - reaching convergence = draws will be draws from $p(\theta|x)$ (our posterior); use them to approximate posterior distribution itself
    - approximate probability of θ being in custom range
    - diagnose chain-specific issues
    - use loss function f(x) to approximate risk/reward of your estimates being wrong

Ex: Reporting District Mean Posteriors vs. Intercept and Offsets

In hierarchical models, results can be reported either as:

1. **Intercept and District Offsets**: global mean (Intercept) and deviations for each district (Offsets).
2. **District Mean Posteriors**: direct posterior distributions of mean for each district.

Why Report District Mean Posteriors?

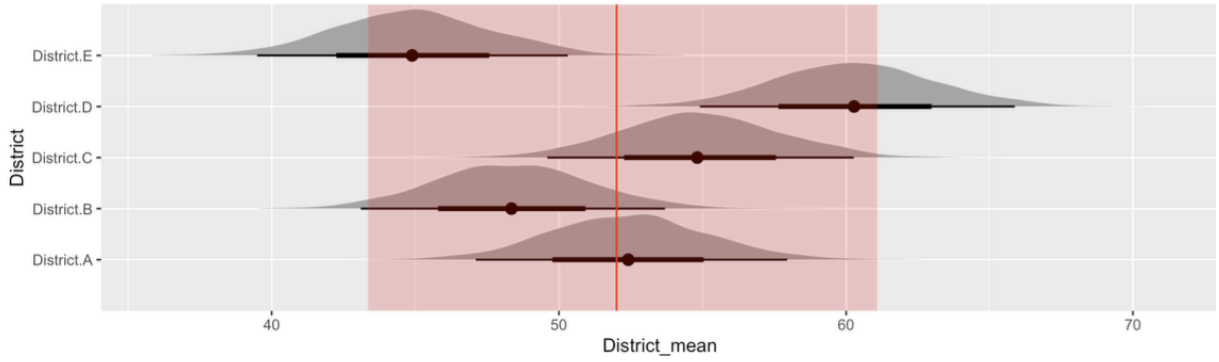
1. **Simplified Interpretation**:
    - Directly provides actionable estimates (e.g., "District A's mean is 75 with 95% credible interval of [70, 80]").
    - Eliminates need to reconstruct means by adding offsets to global intercept.
2. **Reflects Hierarchical Regularization**:
    - Incorporates partial pooling: districts with less data shrink toward overall mean.
    - Captures combined uncertainty in both global and district-specific effects.
3. **Practical Communication**:
    - Easier for non-technical audiences to understand and use in decision-making.

Example:

- **Traditional Reporting**:

- Intercept: 80 (global mean).
- District A Offset: -5.
- District A mean: Requires calculation (80 - 5 = 75).

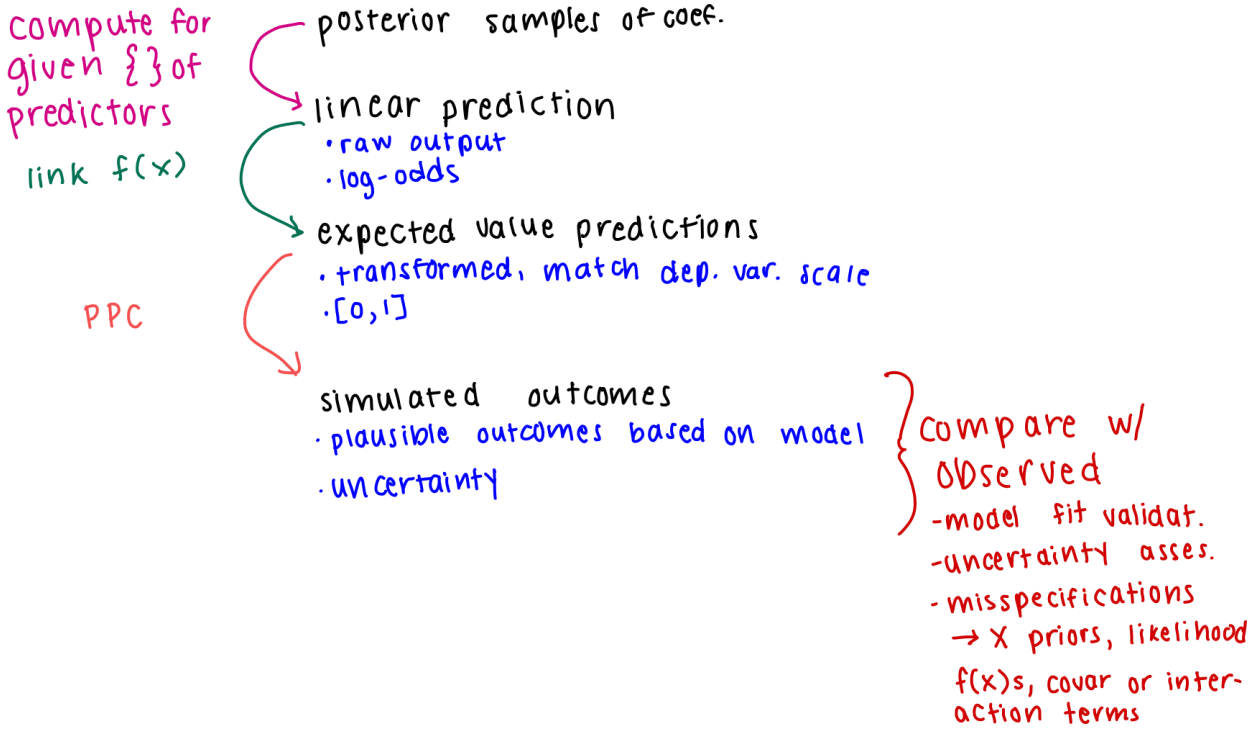
- **District Mean Posterior**:
  - Directly reported as 75 with 95% credible interval of [70, 80].



$$\mathbb{E}(y|\mathbf{X}) = \mu = g^{-1}(\mathbf{X}\beta)$$
$$p(y|x) \sim \pi(\mu, \ldots)$$

- **Linear** Predictions: $\mathbf{X}\beta$
- **Expected Value** Predictions: $\mathbb{E}(y|\mathbf{X}) = g^{-1}(\mathbf{X}\beta)$
- **Posterior** Predictions: $p(\tilde{y}|y)$

Expected Value Predictions: draws of expected values that are linear predictions transformed using link function



2. **Marginal Effects:** help explain/plot effects that are otherwise difficult to communicate.
    - Marginal effects (derivatives): " slope of outcome wrt one of model's predictors"
    - Marginal effects (integrals): "'average' of unit-level effect estimates"