CS 784: Computational Linguistics
Lecture 3: Statistical Methods

Freda Shi

School of Computer Science, University of Waterloo
fhs@uwaterloo.ca

January 16, 2025

[Most slides adapted from Roger Levy.]

Statistics

In Lecture 2.2 we reviewed basics of probability: the logical calculus of uncertainty—a branch of mathematics.

The primary focus of this lecture is **statistics**: the mathematics, science, craft, and art of drawing inferences from data.

Statistics

In Lecture 2.2 we reviewed basics of probability: the logical calculus of uncertainty—a branch of mathematics.

The primary focus of this lecture is **statistics**: the mathematics, science, craft, and art of drawing inferences from data.

The two fields are fundamentally different, but probability is used extensively in statistics.

Parameter Estimation

Consider a binary random variable Y with two possible outcomes (e.g., coin flips), head (1) and tail (0).

Y obeys a **Bernoulli distribution** with parameter θ:

$$P(Y = 1) = \theta, \quad P(Y = 0) = 1 - \theta$$

Parameter Estimation

Consider a binary random variable Y with two possible outcomes (e.g., coin flips), head (1) and tail (0).

Y obeys a **Bernoulli distribution** with parameter θ:

$$P(Y = 1) = \theta, \quad P(Y = 0) = 1 - \theta$$

Parameter estimation figures out what the parameter θ is.

Parameter Estimation

Consider a binary random variable Y with two possible outcomes (e.g., coin flips), head (1) and tail (0).

Y obeys a **Bernoulli distribution** with parameter θ:

$$P(Y = 1) = \theta, \quad P(Y = 0) = 1 - \theta$$

Parameter estimation figures out what the parameter θ is.

In general, we will use $\boldsymbol{y}$ to refer to observed-outcome data and θ to refer to the model parameter(s) to be estimated.

Parameter Estimation

Consider a binary random variable Y with two possible outcomes (e.g., coin flips), head (1) and tail (0).

Y obeys a **Bernoulli distribution** with parameter θ:

$$P(Y = 1) = \theta, \quad P(Y = 0) = 1 - \theta$$

Parameter estimation figures out what the parameter θ is.

In general, we will use $\boldsymbol{y}$ to refer to observed-outcome data and θ to refer to the model parameter(s) to be estimated.

From the parameter-estimation perspective, deep learning is statistics.

Statistical Estimators

Estimator: a procedure for guessing a quantity of interest within a population based on a sample.

Example: The **relative frequency estimator** for θ is the proportion of observed outcomes that are 1:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Statistical Estimators

Estimator: a procedure for guessing a quantity of interest within a population based on a sample.

Example: The **relative frequency estimator** for θ is the proportion of observed outcomes that are 1:

Indicator of estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Statistical Estimators

Estimator: a procedure for guessing a quantity of interest within a population based on a sample.

Example: The **relative frequency estimator** for θ is the proportion of observed outcomes that are 1:

Indicator of estimator

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Data are stochastic, so estimators give random variables.

Statistical Estimators

Estimator: a procedure for guessing a quantity of interest within a population based on a sample.

Example: The **relative frequency estimator** for θ is the proportion of observed outcomes that are 1:

Indicator of estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Data are stochastic, so estimators give random variables.

Bias of an estimator: $\mathbb{E}[\widehat{\theta}] - \theta$.

$$\mathbb{E}[\widehat{\theta}] \overset{expand}{=} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \overset{linearity}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{n}\cdot n\cdot\theta = \theta$$

Statistical Estimators

Estimator: a procedure for guessing a quantity of interest within a population based on a sample.

Example: The **relative frequency estimator** for θ is the proportion of observed outcomes that are 1:

Indicator of estimator

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Data are stochastic, so estimators give random variables.

Bias of an estimator: $\mathbb{E}[\widehat{\theta}] - \theta$.

$$\mathbb{E}[\widehat{\theta}] \overset{expand}{=} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \overset{linearity}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{1}{n}\cdot n\cdot\theta = \theta$$

Variance of an estimator is the ordinary variance:

$$Var[\widehat{\theta}] = \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2] = \frac{\theta(1-\theta)}{n}.$$

Statistical Estimators

Estimator: a procedure for guessing a quantity of interest within a population based on a sample.

Example: The **relative frequency estimator** for θ is the proportion of observed outcomes that are 1:

$$\text{Indicator of estimator} \longrightarrow \widehat{\theta} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i$$

Data are stochastic, so estimators give random variables.

Bias of an estimator: $\mathbb{E}[\widehat{\theta}] - \theta$.

$$\mathbb{E}[\widehat{\theta}] \overset{expand}{=} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \overset{linearity}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{n}\cdot n\cdot\theta = \theta$$

Variance of an estimator is the ordinary variance:

$$Var[\widehat{\theta}] = \mathbb{E}[(\widehat{\theta} - \mathbb{E}[\widehat{\theta}])^2] = \frac{\theta(1-\theta)}{n}.$$

Good estimators have favorable bias-variance trade-offs.

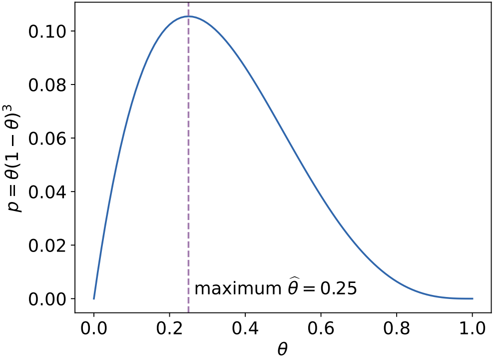
Maximum Likelihood Estimation

$$L(\theta; \boldsymbol{y}) \equiv P(\boldsymbol{y} \mid \theta) \qquad \widehat{\theta}_{MLE} = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

Maximum Likelihood Estimation

$$L(\theta; \boldsymbol{y}) \equiv P(\boldsymbol{y} \mid \theta) \qquad \widehat{\theta}_{MLE} = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

#Toss	**Outcome**
1	T
2	H
3	T
4	T

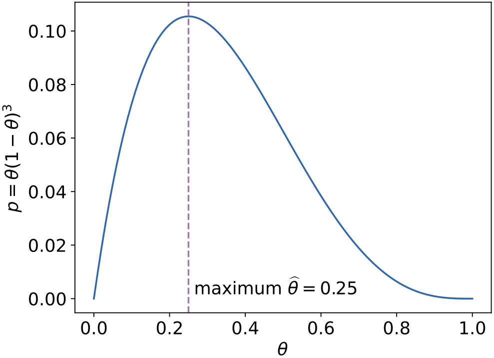


Maximum Likelihood Estimation

$$L(\theta; \boldsymbol{y}) \equiv P(\boldsymbol{y} \mid \theta) \qquad \widehat{\theta}_{MLE} = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

#Toss	**Outcome**
1	T
2	H
3	T
4	T

The **maximum likelihood estimate (MLE)** also turns out to be the relative frequency estimate (RFE).



Maximum Log-Likelihood Estimation

$\log(x)$ is monotonically increasing w.r.t. x, so

$$\arg\max_{\theta} \log L(\theta; \boldsymbol{y}) = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

Maximum Log-Likelihood Estimation

$\log(x)$ is monotonically increasing w.r.t. x, so

$$\arg\max_{\theta} \log L(\theta; \boldsymbol{y}) = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

We usually assume independence of observations, so

$$L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} P(y_i \mid \theta)$$

$$\log L(\theta; \boldsymbol{y}) = \sum_{i=1}^{n} \log P(y_i \mid \theta) = \sum_{i=1}^{n} y_i \log \theta + (1 - y_i) \log(1 - \theta)$$

Maximum Log-Likelihood Estimation

$\log(x)$ is monotonically increasing w.r.t. x, so

$$\arg\max_{\theta} \log L(\theta; \boldsymbol{y}) = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

We usually assume independence of observations, so

$$L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} P(y_i \mid \theta)$$

$$\log L(\theta; \boldsymbol{y}) = \sum_{i=1}^{n} \log P(y_i \mid \theta) = \sum_{i=1}^{n} y_i \log \theta + (1 - y_i) \log(1 - \theta)$$

$\log L(\theta; \boldsymbol{y})$ is more derivative-friendly.

Maximum Log-Likelihood Estimation

$\log(x)$ is monotonically increasing w.r.t. x, so

$$\arg\max_{\theta} \log L(\theta; \boldsymbol{y}) = \arg\max_{\theta} L(\theta; \boldsymbol{y})$$

We usually assume independence of observations, so

$$L(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} P(y_i \mid \theta)$$

$$\log L(\theta; \boldsymbol{y}) = \sum_{i=1}^{n} \log P(y_i \mid \theta) = \sum_{i=1}^{n} y_i \log \theta + (1 - y_i) \log(1 - \theta)$$

$\log L(\theta; \boldsymbol{y})$ is more derivative-friendly.

To minimize it, we can set its derivative w.r.t. θ to 0:

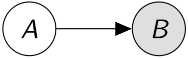
$$\frac{d}{d\theta} \log L(\theta; \boldsymbol{y}) = 0$$

$$\sum_{i=1}^{n} \underbrace{\frac{y_i}{\theta}}_{\substack{\text{added if} \\ y_i = 1}} - \underbrace{\frac{1 - y_i}{1 - \theta}}_{\text{added if } y_i = 0} = 0 \Rightarrow \widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Bayesian Parameter Estimation

The Bayes' rule:

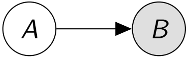
$$P(A \mid B) \propto P(B \mid A) P(A)$$



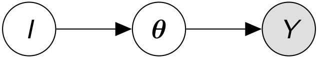
Bayesian Parameter Estimation

The Bayes' rule:

$$P(A \mid B) \propto P(B \mid A) P(A)$$



Assume that the model parameters $\boldsymbol{\theta}$, background knowledge (prior) I, and observed data Y are all random variables.

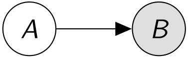


We are interested in the posterior distribution $P(\boldsymbol{\theta} \mid Y, I)$.

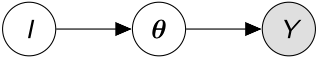
Bayesian Parameter Estimation

The Bayes' rule:

$$P(A \mid B) \propto P(B \mid A) P(A)$$



Assume that the model parameters $\boldsymbol{\theta}$, background knowledge (prior) I, and observed data Y are all random variables.



We are interested in the posterior distribution $P(\boldsymbol{\theta} \mid Y, I)$.

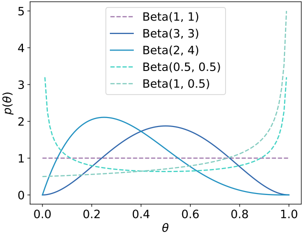
$$\begin{aligned} \overbrace{P(\boldsymbol{\theta} \mid Y, I)}^{\text{posterior}} &\propto P(Y \mid \boldsymbol{\theta}, I) P(\boldsymbol{\theta} \mid I) && \text{(Bayes' rule)} \\ &= P(Y \mid \boldsymbol{\theta}) \underbrace{P(\boldsymbol{\theta} \mid I)}_{\text{prior}} && \text{(conditional independence)} \end{aligned}$$

Example: Beta Distribution for Coin Flips

The Beta distribution express background knowledge I as two "pseudo-count" parameters α_1 and α_2:

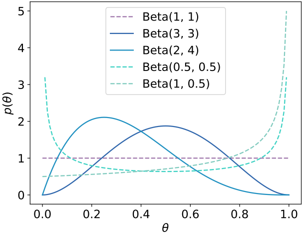
$$P(\theta \mid \alpha_1, \alpha_2) = \frac{\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)}$$

Normalizer $B(\alpha_1, \alpha_2) = \int_0^1 \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} d\theta$



Example: Beta Distribution for Coin Flips

The Beta distribution express background knowledge I as two “pseudo-count” parameters α_1 and α_2:



$$P(\theta \mid \alpha_1, \alpha_2) = \frac{\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}}{B(\alpha_1, \alpha_2)}$$

Normalizer $B(\alpha_1, \alpha_2) = \int_0^1 \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1} d\theta$

Beta distribution is a **conjugate prior** for the Bernoulli likelihood:

$$\overbrace{P(\theta \mid Y, \alpha_1, \alpha_2)}^{\text{posterior}} \propto \overbrace{P(Y \mid \theta)}^{\text{likelihood}} \overbrace{P(\theta \mid \alpha_1, \alpha_2)}^{\text{prior}}$$
$$= \theta^m (1-\theta)^{n-m} \cdot \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1},$$

where m is the number of heads and n is the number of coin flips.

Posterior Prediction

Beta distribution Beta(α_1, α_2)
(suppose $\alpha_1, \alpha_2 > 1$)

Mean
$$\frac{\alpha_1}{\alpha_1 + \alpha_2}$$

Mode
$$\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$$

Posterior Prediction

	Beta distribution $\mathrm{Beta}(\alpha_1, \alpha_2)$ (suppose $\alpha_1, \alpha_2 > 1$)	Posterior (n coin flips, m heads)
Mean	$\frac{\alpha_1}{\alpha_1 + \alpha_2}$	$\frac{m + \alpha_1}{n + \alpha_1 + \alpha_2}$
Mode	$\frac{\alpha_1 - 1}{\alpha_1 + \alpha_2 - 2}$	$\frac{m + \alpha_1 - 1}{n + \alpha_1 + \alpha_2 - 2}$

Point Estimates

A **point estimate** of a model parameter is a **statistic**.

> *A **statistic**... or **sample statistic** is any **quantity** computed from values in a sample which is considered for a statistical purpose.* *[Source: Wikipedia]*

Point Estimates

A **point estimate** of a model parameter is a **statistic**.

> *A **statistic**... or **sample statistic** is any **quantity** computed from values in a sample which is considered for a statistical purpose.* *[Source: Wikipedia]*

We have seen a few examples of point estimates: MLE, RFE, posterior mean, and posterior mode.

Point Estimates

A **point estimate** of a model parameter is a **statistic**.

> *A **statistic**... or **sample statistic** is any **quantity** computed from values in a sample which is considered for a statistical purpose.* *[Source: Wikipedia]*

We have seen a few examples of point estimates: MLE, RFE, posterior mean, and posterior mode.

All of them discarded the information from the curve of $P(\boldsymbol{\theta} \mid Y, I)$.

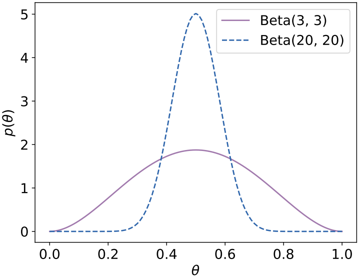
Point Estimates

A **point estimate** of a model parameter is a **statistic**.

> *A **statistic**... or **sample statistic** is any **quantity** computed from values in a sample which is considered for a statistical purpose.* [Source: Wikipedia]

We have seen a few examples of point estimates: MLE, RFE, posterior mean, and posterior mode.

All of them discarded the information from the curve of $P(\boldsymbol{\theta} \mid Y, I)$.



Both distributions have mean 0.5, but do they contain the same information?

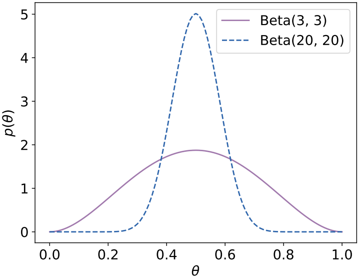
Point Estimates

A **point estimate** of a model parameter is a **statistic**.

> *A **statistic**... or **sample statistic** is any **quantity** computed from values in a sample which is considered for a statistical purpose.* [Source: Wikipedia]

We have seen a few examples of point estimates: MLE, RFE, posterior mean, and posterior mode.

All of them discarded the information from the curve of $P(\boldsymbol{\theta} \mid Y, I)$.



Both distributions have mean 0.5, but do they contain the same information?

Curve shape captures **uncertainty** about parameters.

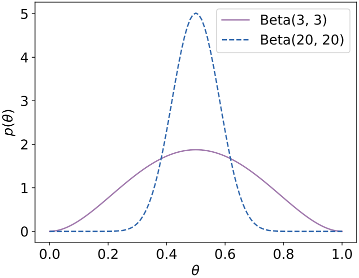
Point Estimates

A **point estimate** of a model parameter is a **statistic**.

> *A **statistic**... or **sample statistic** is any **quantity** computed from values in a sample which is considered for a statistical purpose.* *[Source: Wikipedia]*

We have seen a few examples of point estimates: MLE, RFE, posterior mean, and posterior mode.

All of them discarded the information from the curve of $P(\boldsymbol{\theta} \mid Y, I)$.



Both distributions have mean 0.5, but do they contain the same information?

Curve shape captures **uncertainty** about parameters.

Credible intervals (Bayesian) and **confidence intervals** (frequentist) quantify this uncertainty.

Bayesian Credible Intervals

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

Bayesian Credible Intervals

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

A $(1-\alpha)$ Bayesian **credible interval** (CI) on parameter θ is an interval containing $(1-\alpha)$ of the posterior probability mass.

α: significance level. $* : \alpha = 0.05$, $** : \alpha = 0.01$, $*** : \alpha = 0.001$.

Bayesian Credible Intervals

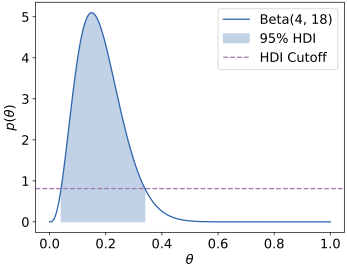
$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

A $(1-\alpha)$ Bayesian **credible interval** (CI) on parameter θ is an interval containing $(1-\alpha)$ of the posterior probability mass.

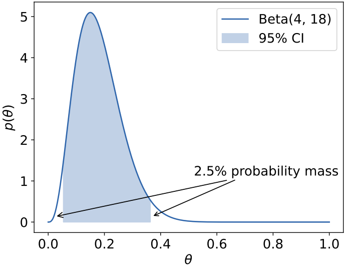
α: significance level. $* : \alpha = 0.05$, $** : \alpha = 0.01$, $*** : \alpha = 0.001$.

Two common standards for Bayesian CI construction:

Highest Posterior Density



Symmetric



Bayesian Credible Intervals

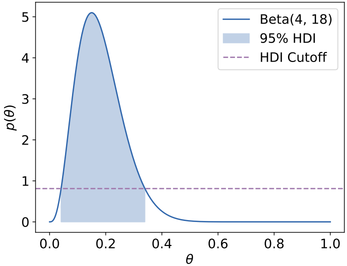
$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

A $(1-\alpha)$ Bayesian **credible interval** (CI) on parameter θ is an interval containing $(1-\alpha)$ of the posterior probability mass.

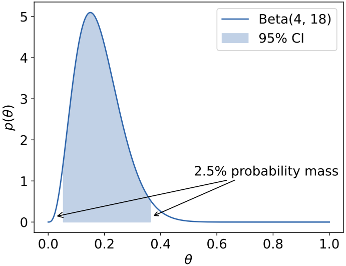
α: significance level. $* : \alpha = 0.05$, $** : \alpha = 0.01$, $*** : \alpha = 0.001$.

Two common standards for Bayesian CI construction:

Highest Posterior Density



Symmetric



Multivariate generalization: interval → region.

Monte Carlo Methods

The posterior distribution $P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$ could be too complex to compute analytically.

To determine the credible interval, we can simulate the posterior distribution using **Monte Carlo** methods.

Monte Carlo Methods

The posterior distribution $P(\theta \mid Y, I) \propto P(Y \mid \theta)P(\theta \mid I)$ could be too complex to compute analytically.

To determine the credible interval, we can simulate the posterior distribution using **Monte Carlo** methods.

Generally speaking, we

- Define a domain of possible inputs
- Generate n i.i.d. random inputs from a probability distribution over the domain
- Perform a computation on each randomly generated input
- Aggregate the results

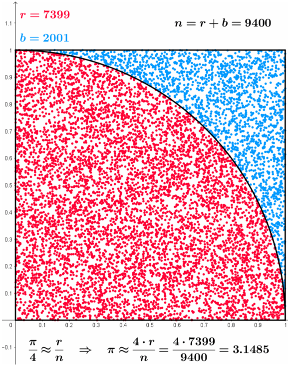
Monte Carlo Methods

The posterior distribution $P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$ could be too complex to compute analytically.

To determine the credible interval, we can simulate the posterior distribution using **Monte Carlo** methods.

Generally speaking, we

- Define a domain of possible inputs
- Generate n i.i.d. random inputs from a probability distribution over the domain
- Perform a computation on each randomly generated input
- Aggregate the results



[Source: Wikipedia]

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.
- For any desirable θ, we can compute $P(Y \mid \theta) P(\theta \mid I)$.

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.
- For any desirable θ, we can compute $P(Y \mid \theta) P(\theta \mid I)$.
- We can sample from $\mathcal{U}(0, 1)$, or any uniform distribution.

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.
- For any desirable θ, we can compute $P(Y \mid \theta) P(\theta \mid I)$.
- We can sample from $\mathcal{U}(0, 1)$, or any uniform distribution.

Suppose we know $u = \max_\theta P(Y \mid \theta) P(\theta \mid I)$, or simply use a large enough u. Repeating the following process approximates sampling from $P(\theta \mid Y, I)$:

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.
- For any desirable θ, we can compute $P(Y \mid \theta) P(\theta \mid I)$.
- We can sample from $\mathcal{U}(0, 1)$, or any uniform distribution.

Suppose we know $u = \max_\theta P(Y \mid \theta) P(\theta \mid I)$, or simply use a large enough u. Repeating the following process approximates sampling from $P(\theta \mid Y, I)$:

- Draw $\theta' \sim \mathcal{U}(l, r)$, where l and r are the bounds of the domain of θ.

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.
- For any desirable θ, we can compute $P(Y \mid \theta) P(\theta \mid I)$.
- We can sample from $\mathcal{U}(0, 1)$, or any uniform distribution.

Suppose we know $u = \max_\theta P(Y \mid \theta) P(\theta \mid I)$, or simply use a large enough u. Repeating the following process approximates sampling from $P(\theta \mid Y, I)$:

- Draw $\theta' \sim \mathcal{U}(l, r)$, where l and r are the bounds of the domain of θ.
- Draw $u' \sim \mathcal{U}(0, u)$.

Monte Carlo Methods: Our Case

$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I)$$

- We can't sample from $P(\theta \mid Y, I)$ directly.
- We can't calculate $pdf(\theta \mid Y, I)$, nor compute its integral (i.e., the cumulative distribution function, CDF); therefore, we can't directly calculate the credible interval.
- For any desirable θ, we can compute $P(Y \mid \theta) P(\theta \mid I)$.
- We can sample from $\mathcal{U}(0, 1)$, or any uniform distribution.

Suppose we know $u = \max_{\theta} P(Y \mid \theta) P(\theta \mid I)$, or simply use a large enough u. Repeating the following process approximates sampling from $P(\theta \mid Y, I)$:

- Draw $\theta' \sim \mathcal{U}(l, r)$, where l and r are the bounds of the domain of θ.
- Draw $u' \sim \mathcal{U}(0, u)$.
- If $u' \leq P(Y \mid \theta') P(\theta' \mid I)$, collect θ'; otherwise discard it (rejection sampling).

Example: Monte Carlo with Rejection Sampling

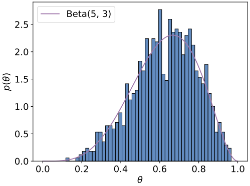
$$P(\theta \mid Y, I) \propto P(Y \mid \theta) P(\theta \mid I), P(Y \mid \theta) = \theta^4 (1-\theta)^2, P(\theta \mid I) = 1$$

For reference, we know that the posterior distribution is $\text{Beta}(5,3)$.

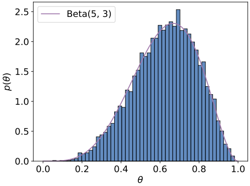
```
import seaborn as sns
import matplotlib.pyplot as plt

def monte_carlo(n_samples=1000, n_heads=4, n_tails=2):
    likelihood = lambda theta: theta ** n_heads * (1 - theta) ** n_tails
    sampled_thetas = []
    while len(sampled_thetas) < n_samples:
        theta = np.random.uniform(0, 1)
        u = np.random.uniform(0, 1)
        if u < likelihood(theta):
            sampled_thetas.append(theta)
    sns.histplot(sampled_thetas, bins=50, stat='density')
    plt.savefig(f'monte-carlo-{n_samples}.pdf', bbox_inches='tight')

np.random.seed(42)  # for reproducibility
monte_carlo(1000), monte_carlo(10000), monte_carlo(100000)
```

Note: a few imports are omitted for brevity.

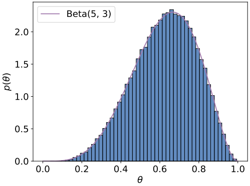
Example: Monte Carlo with Rejection Sampling



1,000 samples

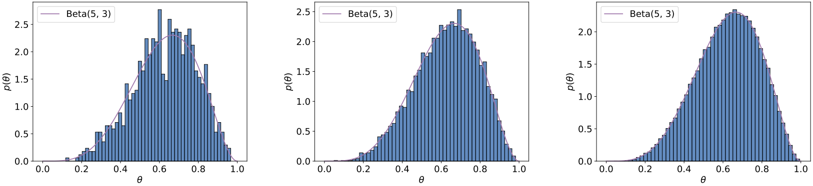


10,000 samples



100,000 samples

Example: Monte Carlo with Rejection Sampling



1,000 samples 10,000 samples 100,000 samples

Warning: rejection sampling could be very inefficient.

In practice, we use **Markov Chain Monte Carlo** (MCMC) with advanced algorithms such as the Metropolis–Hastings algorithm.

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

If $(1-\alpha)$ of these intervals contain the true parameter θ, then Proc is a method for constructing a $(1-\alpha)$ confidence interval.

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

If $(1 - \alpha)$ of these intervals contain the true parameter θ, then Proc is a method for constructing a $(1 - \alpha)$ confidence interval.

For a **normal distribution**,

$$\frac{\widehat{\mu} - \mu}{\sqrt{S^2 / n}} \sim t_{n-1}$$

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

If $(1-\alpha)$ of these intervals contain the true parameter θ, then Proc is a method for constructing a $(1-\alpha)$ confidence interval.

For a **normal distribution**,

$$\text{Sample Mean}\ \frac{\boxed{\widehat{\mu}} - \mu}{\sqrt{S^2 / n}} \sim t_{n-1}$$

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

If $(1-\alpha)$ of these intervals contain the true parameter θ, then Proc is a method for constructing a $(1-\alpha)$ confidence interval.

For a **normal distribution**,

$$\frac{\text{Sample Mean } \widehat{\mu} - \mu}{\underset{\text{Standard Error}}{\sqrt{S^2/n}}} \sim t_{n-1}$$

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

If $(1-\alpha)$ of these intervals contain the true parameter θ, then Proc is a method for constructing a $(1-\alpha)$ confidence interval.

For a **normal distribution**,

Sample Mean $\dfrac{\widehat{\mu} - \mu}{\sqrt{S^2 / n}} \sim t_{n-1}$ Degree of Freedom

Standard Error

Frequentist Confidence Intervals

Key idea: for parameter θ, define a procedure for constructing an interval I_θ from data $\boldsymbol{y}$.

$$I_\theta = \text{Proc}(\boldsymbol{y})$$

Suppose we repeat the procedure many times, each time collecting data $\boldsymbol{y}$ and constructing $I_\theta = \text{Proc}(\boldsymbol{y})$.

If $(1-\alpha)$ of these intervals contain the true parameter θ, then Proc is a method for constructing a $(1-\alpha)$ confidence interval.

For a **normal distribution**,

$$\text{Sample Mean}\ \frac{\boxed{\widehat{\mu}} - \mu}{\underbrace{\boxed{\sqrt{S^2/n}}}_{\text{Standard Error}}} \sim t_{\boxed{n-1}}\ \text{Degree of Freedom}$$

Frequentist confidence interval $= \widehat{\mu} \pm t_{n-1} \cdot \frac{S}{\sqrt{n}}$

Bayesian Hypothesis Testing

Hypothesis: a candidate theory/model for the generative process by which data $\boldsymbol{y}$ come into the world.

Bayesian Hypothesis Testing

Hypothesis: a candidate theory/model for the generative process by which data $\boldsymbol{y}$ come into the world.

Bayesian inference provides a simple toolkit to compare two hypotheses $\{H_i\}$.

$$P(H_i \mid \boldsymbol{y}) \propto P(\boldsymbol{y} \mid H_i) P(H_i)$$

Bayesian Hypothesis Testing

Hypothesis: a candidate theory/model for the generative process by which data $\boldsymbol{y}$ come into the world.

Bayesian inference provides a simple toolkit to compare two hypotheses $\{H_i\}$.

$$P(H_i \mid \boldsymbol{y}) \propto P(\boldsymbol{y} \mid H_i) P(H_i)$$

$$\overbrace{\frac{P(H \mid \boldsymbol{y})}{P(H' \mid \boldsymbol{y})}}^{\text{Posterior odds}} = \overbrace{\frac{P(\boldsymbol{y} \mid H)}{P(\boldsymbol{y} \mid H')}}^{\text{Likehood ratio}} \overbrace{\frac{P(H)}{P(H')}}^{\text{Prior odds}}$$

Bayesian Hypothesis Testing

Hypothesis: a candidate theory/model for the generative process by which data $\boldsymbol{y}$ come into the world.

Bayesian inference provides a simple toolkit to compare two hypotheses $\{H_i\}$.

$$P(H_i \mid \boldsymbol{y}) \propto P(\boldsymbol{y} \mid H_i) P(H_i)$$

$$\overbrace{\frac{P(H \mid \boldsymbol{y})}{P(H' \mid \boldsymbol{y})}}^{\text{Posterior odds}} = \overbrace{\frac{P(\boldsymbol{y} \mid H)}{P(\boldsymbol{y} \mid H')}}^{\text{Likehood ratio}} \overbrace{\frac{P(H)}{P(H')}}^{\text{Prior odds}}$$

We use the **likelihood ratio** as the **Bayes factor**.

Bayesian Hypothesis Testing

Hypothesis: a candidate theory/model for the generative process by which data $\boldsymbol{y}$ come into the world.

Bayesian inference provides a simple toolkit to compare two hypotheses $\{H_i\}$.

$$P(H_i \mid \boldsymbol{y}) \propto P(\boldsymbol{y} \mid H_i) P(H_i)$$

$$\overbrace{\frac{P(H \mid \boldsymbol{y})}{P(H' \mid \boldsymbol{y})}}^{\text{Posterior odds}} = \overbrace{\frac{P(\boldsymbol{y} \mid H)}{P(\boldsymbol{y} \mid H')}}^{\text{Likehood ratio}} \overbrace{\frac{P(H)}{P(H')}}^{\text{Prior odds}}$$

We use the **likelihood ratio** as the **Bayes factor**.

Compared to the posterior odds, the Bayes factor is more robust to the choice of prior.

Empirical Interpretation of Bayes Factors

$$K = \frac{P(\mathbf{y} \mid H_1)}{P(\mathbf{y} \mid H_2)}$$

$\log_{10} K$	K	Strength of evidence
0 to $\frac{1}{2}$	1 to 3.2	Not worth more than a bare mention
$\frac{1}{2}$ to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

[Kass and Raftery, 1995; Table source: Wikipedia]

Example: Bayesian Hypothesis Testing

$$H_1 : P(\theta \mid H_1) = \begin{cases} 1 & \theta = 0.5 \\ 0 & \text{otherwise} \end{cases}$$ Coin is fair

$$H_2 : p(\theta \mid H_2) = 1 \quad 0 \leq \theta \leq 1$$ Coin is not fair*

Example: Bayesian Hypothesis Testing

$$H_1 : P(\theta \mid H_1) = \begin{cases} 1 & \theta = 0.5 \\ 0 & \text{otherwise} \end{cases}$$ Coin is fair

$$H_2 : p(\theta \mid H_2) = 1 \quad 0 \leq \theta \leq 1$$ Coin is not fair*

Data: $\boldsymbol{y} = [0, 1, 1, 0, 1, 1]$

Example: Bayesian Hypothesis Testing

$$H_1 : P(\theta \mid H_1) = \begin{cases} 1 & \theta = 0.5 \\ 0 & \text{otherwise} \end{cases}$$ Coin is fair

$$H_2 : p(\theta \mid H_2) = 1 \quad 0 \leq \theta \leq 1$$ Coin is not fair*

Data: $\boldsymbol{y} = [0, 1, 1, 0, 1, 1]$

$$P(\boldsymbol{y} \mid H_1) = \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 = 0.0156$$

$$P(\boldsymbol{y} \mid H_2) = \int_0^1 \underbrace{\theta^4 (1-\theta)^2}_{P(\boldsymbol{y} \mid \theta)} \cdot \underbrace{1}_{P(\theta|H_2)} \, d\theta = 0.0095$$

Example: Bayesian Hypothesis Testing

$$H_1 : P(\theta \mid H_1) = \begin{cases} 1 & \theta = 0.5 \\ 0 & \text{otherwise} \end{cases}$$ Coin is fair

$$H_2 : p(\theta \mid H_2) = 1 \quad 0 \leq \theta \leq 1$$ Coin is not fair*

Data: $\boldsymbol{y} = [0, 1, 1, 0, 1, 1]$

$$P(\boldsymbol{y} \mid H_1) = \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^2 = 0.0156$$

$$P(\boldsymbol{y} \mid H_2) = \int_0^1 \underbrace{\theta^4 (1-\theta)^2}_{P(\boldsymbol{y} \mid \theta)} \cdot \underbrace{1}_{P(\theta|H_2)} \, d\theta = 0.0095$$

$$\frac{P(\boldsymbol{y} \mid H_1)}{P(\boldsymbol{y} \mid H_2)} = 1.64$$

Power Analysis

For certain data, we have

H_0 is	Accept H_0	Reject H_0
True	Correct (prob. $1-\alpha$)	Type I error (prob. α)
False	Type II error (prob. β)	Correct (prob. $1-\beta$)

α: significance level. $1-\beta$: power.

Power Analysis

For certain data, we have

H_0 is	Accept H_0	Reject H_0
True	Correct (prob. $1-\alpha$)	Type I error (prob. α)
False	Type II error (prob. β)	Correct (prob. $1-\beta$)

α: significance level. $1-\beta$: power.

In a statistical test, we typically control α, which sets up a threshold for decision making, and compute $1-\beta$.

A $1-\beta$ of 0.8 is generally considered good.

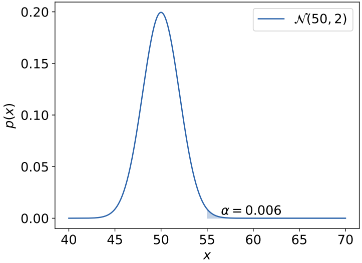
Example: Power Analysis

We are interested in whether the average grade of the class is indifferent from 50. Based on the grades of 25 students with sample standard deviation of 10, we decided to reject H_0 if the average grade is greater than 55.

Example: Power Analysis

We are interested in whether the average grade of the class is indifferent from 50. Based on the grades of 25 students with sample standard deviation of 10, we decided to reject H_0 if the average grade is greater than 55.

What is the significance level α?

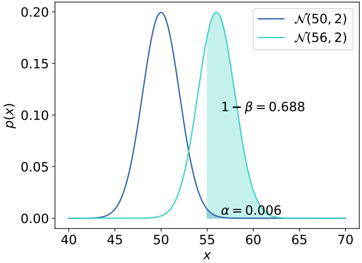


Example: Power Analysis

We are interested in whether the average grade of the class is indifferent from 50. Based on the grades of 25 students with sample standard deviation of 10, we decided to reject H_0 if the average grade is greater than 55.

What is the significance level α?

Suppose the true population mean is 56. What is the power $1 - \beta$?

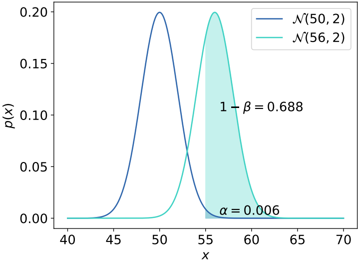


Example: Power Analysis

We are interested in whether the average grade of the class is indifferent from 50. Based on the grades of 25 students with sample standard deviation of 10, we decided to reject H_0 if the average grade is greater than 55.

What is the significance level α?

Suppose the true population mean is 56. What is the power $1-\beta$?



Since the sample size reduces the uncertainty, the power is higher when the sample size is larger **if the effect is indeed present**.

Frequentist Hypothesis Testing

The **Neyman–Pearson paradigm**: formulate two hypotheses about the generative process underlying the data.

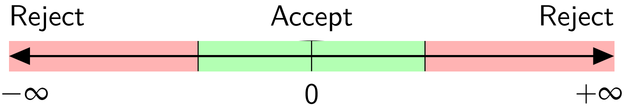
- Null hypothesis H_0.
- Alternative hypothesis H_A within which H_0 is nested.

Frequentist Hypothesis Testing

The **Neyman–Pearson paradigm**: formulate two hypotheses about the generative process underlying the data.

- Null hypothesis H_0.
- Alternative hypothesis H_A within which H_0 is nested.

Define a **rejection region** R such that if $T(\boldsymbol{y}) \in R$, we reject H_0.

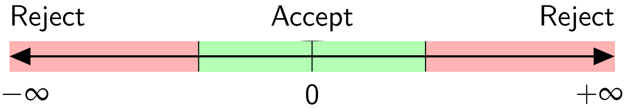


Frequentist Hypothesis Testing

The **Neyman–Pearson paradigm**: formulate two hypotheses about the generative process underlying the data.

- Null hypothesis H_0.
- Alternative hypothesis H_A within which H_0 is nested.

Define a **rejection region** R such that if $T(\boldsymbol{y}) \in R$, we reject H_0.



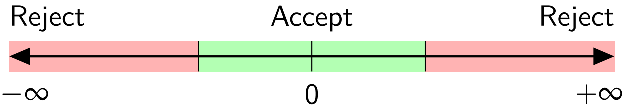
Choose a **test statistic** $T(\boldsymbol{y})$ that is a function of the data.

Frequentist Hypothesis Testing

The **Neyman–Pearson paradigm**: formulate two hypotheses about the generative process underlying the data.

- Null hypothesis H_0.
- Alternative hypothesis H_A within which H_0 is nested.

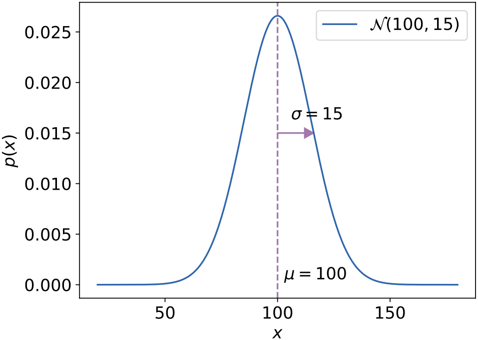
Define a **rejection region** R such that if $T(\boldsymbol{y}) \in R$, we reject H_0.



Choose a **test statistic** $T(\boldsymbol{y})$ that is a function of the data.

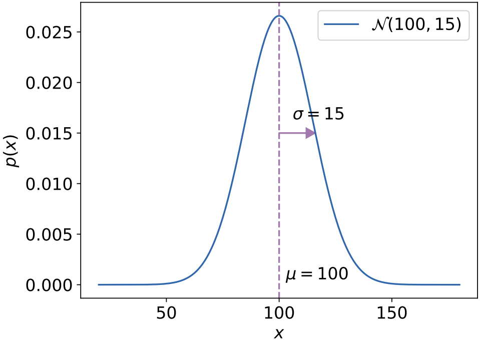
Collect data, compute $T(\boldsymbol{y})$, and compare it to the rejection region.

The Gaussian (or Normal) Distribution



$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

The Gaussian (or Normal) Distribution



$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

Unbiased estimates from a size n sample:

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

The *t*-Test: Three Variants

- **One-sample (Student's) *t*-test:** Does the underline population mean of a sample differ from zero?

The t-Test: Three Variants

- **One-sample (Student's) t-test:** Does the underline population mean of a sample differ from zero?
- **Two-sample t-test (unpaired):** Do the underlying population means of two independent samples differ?

William Sealy Gosset, a.k.a. *Student*

The *t*-Test: Three Variants

- **One-sample (Student's) *t*-test:** Does the underline population mean of a sample differ from zero?
- **Two-sample *t*-test (unpaired):** Do the underlying population means of two independent samples differ?
- **Two-sample *t*-test (paired):** You have a sample of individuals from the population and take measurements from each member of the sample in two different conditions. Do the underlying population means in the two conditions differ?

William Sealy Gosset, a.k.a. *Student*

One-Sample t-Test

Null hypothesis H_0: the mean of the **normally-distributed** population underlying that a sample comes from is $\mu = 0$.

One-Sample t-Test

Null hypothesis H_0: the mean of the **normally-distributed** population underlying that a sample comes from is $\mu = 0$.

Alternative hypothesis H_1: $\mu \neq 0$ (two tails; generally preferred) or $\mu > 0$ (one-tailed; less common).

One-Sample t-Test

Null hypothesis H_0: the mean of the **normally-distributed** population underlying that a sample comes from is $\mu = 0$.

Alternative hypothesis H_1: $\mu \neq 0$ (two tails; generally preferred) or $\mu > 0$ (one-tailed; less common).

Test statistic:

$$t = \frac{\bar{x}}{s/\sqrt{n}}$$

One-Sample t-Test

Null hypothesis H_0: the mean of the **normally-distributed** population underlying that a sample comes from is $\mu = 0$.

Alternative hypothesis H_1: $\mu \neq 0$ (two tails; generally preferred) or $\mu > 0$ (one-tailed; less common).

Test statistic:

$$t = \frac{\bar{x}}{s/\sqrt{n}}$$

Compare t to the t-distribution with $n-1$ degrees of freedom:

Reject H_0 if $|t| > t_{n-1,1-\alpha/2}$.

Two-sample t-Test (Unpaired)

Assumptions: two samples are **i.i.d. normal**.

Two-sample t-Test (Unpaired)

Assumptions: two samples are **i.i.d. normal**.

Null hypothesis H_0: $\mu_1 = \mu_2$.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ (two-tailed); $\mu_1 > \mu_2$ (one-tailed).

Two-sample *t*-Test (Unpaired)

Assumptions: two samples are **i.i.d. normal**.

Null hypothesis H_0: $\mu_1 = \mu_2$.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ (two-tailed); $\mu_1 > \mu_2$ (one-tailed).

If we assume that the two underlying populations have **equal variance** (Student's *t*-test), the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}} \text{ where } s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

Two-sample t-Test (Unpaired)

Assumptions: two samples are **i.i.d. normal**.

Null hypothesis H_0: $\mu_1 = \mu_2$.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ (two-tailed); $\mu_1 > \mu_2$ (one-tailed).

If we assume that the two underlying populations have **equal variance** (Student's t-test), the test statistic is

$$t = \frac{\bar{x_1} - \bar{x_2}}{s_p\sqrt{1/n_1 + 1/n_2}} \text{ where } s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

If we do not assume that the two underlying populations have equal variance, we use the Welch's t-test.

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Paired Two-Sample *t*-Test

Assumptions:

- In a sample of units from a population, for each unit, we have two measurements $\langle x_1, x_2 \rangle$ on the same scale.
- The difference between measurements is **i.i.d. normal**.
- (Sufficient condition: paired measurements are bivariate normal).

Paired Two-Sample *t*-Test

Assumptions:

- In a sample of units from a population, for each unit, we have two measurements $\langle x_1, x_2 \rangle$ on the same scale.
- The difference between measurements is **i.i.d. normal**.
- (Sufficient condition: paired measurements are bivariate normal).

Null hypothesis H_0: $\mu_1 = \mu_2$.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ (two-tailed); $\mu_1 > \mu_2$ (one-tailed).

Paired Two-Sample t-Test

Assumptions:

- In a sample of units from a population, for each unit, we have two measurements $\langle x_1, x_2 \rangle$ on the same scale.
- The difference between measurements is **i.i.d. normal**.
- (Sufficient condition: paired measurements are bivariate normal).

Null hypothesis H_0: $\mu_1 = \mu_2$.

Alternative hypothesis H_1: $\mu_1 \neq \mu_2$ (two-tailed); $\mu_1 > \mu_2$ (one-tailed).

Strategy: compute the difference $d_i = x_{1i} - x_{2i}$ for each unit, and apply the one-sample t-test to the differences.

Linear Regression

We often want a parameterized form to draw inferences about *conditional distributions* $P(Y \mid X_1, \ldots, X_n)$.

Questions we might ask:

- Is there evidence that each X_i predicts Y above and beyond the predictive value of the other X_j?
- Do X_i and X_j have "separate" influences on Y, or do they "interact" in their influences on Y?
- What is the shape of the predictive relationship between Y and X_i?

Linear Regression

We often want a parameterized form to draw inferences about *conditional distributions* $P(Y \mid X_1, \ldots, X_n)$.

Questions we might ask:

- Is there evidence that each X_i predicts Y above and beyond the predictive value of the other X_j?
- Do X_i and X_j have "separate" influences on Y, or do they "interact" in their influences on Y?
- What is the shape of the predictive relationship between Y and X_i?

$$Y = \overbrace{\underbrace{\beta_0}_{\text{“intercept”}} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}^{\text{Predicted Mean}} + \overbrace{\epsilon}^{\text{Noise} \sim \mathcal{N}(0,\sigma)}$$

Parameter Estimation in Linear Regression

There are two major approaches (which are deeply related yet different) in widespread use:

Parameter Estimation in Linear Regression

There are two major approaches (which are deeply related yet different) in widespread use:

- The principle of **maximum likelihood**: pick parameter values that maximizes the probability of data Y
  Choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y \mid \{\beta_i\}, \sigma)$ *as large as possible.*

Parameter Estimation in Linear Regression

There are two major approaches (which are deeply related yet different) in widespread use:

- The principle of **maximum likelihood**: pick parameter values that maximizes the probability of data Y
  Choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y \mid \{\beta_i\}, \sigma)$ *as large as possible.*
  If we augment the X matrix with a column of 1s, the model turns into

  $$Y = X\boldsymbol{\beta} + \epsilon.$$

  The MLE estimate turns out to be

  $$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^\intercal Y$$

Parameter Estimation in Linear Regression

There are two major approaches (which are deeply related yet different) in widespread use:

- The principle of **maximum likelihood**: pick parameter values that maximizes the probability of data Y
  Choose $\{\beta_i\}$ *and* σ *that make the likelihood* $P(Y \mid \{\beta_i\}, \sigma)$ *as large as possible.*
  If we augment the X matrix with a column of 1s, the model turns into
  $$Y = X\boldsymbol{\beta} + \epsilon.$$
  The MLE estimate turns out to be
  $$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^{\mathsf{T}} Y$$
- **Bayesian inference**: put a probability distribution on the model parameters and update it on the basis of the Bayesian rule.
  $$P(\{\beta_i\}, \sigma \mid Y, X) \propto \underbrace{P(Y \mid \{\beta_i\}, \sigma, X)}_{\text{Likelihood}} \underbrace{P(\{\beta_i\}, \sigma)}_{\text{Prior}}$$

Frequentist Hypothesis Testing in Linear Regression

$$(\widehat{\beta}_i - \beta_i)\frac{\sqrt{(X^TX)_{i,i}}}{s} \sim t_{n-m-1},$$

where n is the sample size, m is the number of predictors, and s is the residual standard error.

$$s = \sqrt{\frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$$

Suppose the null hypothesis is $H_0 : \beta_i = 0$, then we will use the t-statistic to test it.

The Decomposition of Variance

Beautiful property of linear models: we can decompose the variance of the dependent variable into two parts.

$$\begin{aligned} Var(Y) &= \sum_j (y_j - \bar{y}^2) \\ &= \underbrace{\sum_j (\widehat{y}_j - \bar{y})^2}_{Var_M(Y)} + \underbrace{\sum_j (y_j - \widehat{y}_j)^2}_{\text{unexplained}} \end{aligned}$$

The Decomposition of Variance

Beautiful property of linear models: we can decompose the variance of the dependent variable into two parts.

$$\begin{aligned} Var(Y) &= \sum_j (y_j - \bar{y}^2) \\ &= \underbrace{\sum_j (\widehat{y}_j - \bar{y})^2}_{Var_M(Y)} + \underbrace{\sum_j (y_j - \widehat{y}_j)^2}_{\text{unexplained}} \end{aligned}$$

Key idea for proof:

$$\sum_j (\widehat{y}_j - \bar{y}) \underbrace{(y_j - \widehat{y}_j)}_{\text{residual}} = 0 \Leftarrow \begin{cases} \sum_j (y_j - \widehat{y}_j) = 0 \\ \sum_j \mathbf{x}_j (y_j - \widehat{y}_j) = 0 \end{cases}$$

Exercise: complete the proof.

Coefficient of Determination (R^2)

For linear models,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

- R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variables.
- R^2 is a measure of the fit of the model.
- R^2 is always between 0 and 1.

Interaction Terms

Usually in real practices, multiple predictors “interact” with each other.

$$\begin{aligned}\text{Sales} =& \beta_0 + \beta_1 \text{Advertising Cost} + \beta_2 \text{Store Size} \\ \text{Sales} =& \beta_0 + \beta_1 \text{Advertising Cost} + \beta_2 \text{Store Size} + \\ & + \beta_3 \text{Advertising Cost} \times \text{Store Size}\end{aligned}$$

Interaction Terms

Usually in real practices, multiple predictors “interact” with each other.

$$\begin{aligned}\text{Sales} =& \beta_0 + \beta_1 \text{Advertising Cost} + \beta_2 \text{Store Size} \\ \text{Sales} =& \beta_0 + \beta_1 \text{Advertising Cost} + \beta_2 \text{Store Size} + \\ & + \beta_3 \text{Advertising Cost} \times \text{Store Size}\end{aligned}$$

The interaction term β_3Advertising Cost × Store Size captures the effect of the interaction between the two predictors.

Interaction Terms

Usually in real practices, multiple predictors “interact” with each other.

$$\begin{aligned}\text{Sales} =&\beta_0 + \beta_1\text{Advertising Cost} + \beta_2\text{Store Size}\\ \text{Sales} =&\beta_0 + \beta_1\text{Advertising Cost} + \beta_2\text{Store Size}+\\ &+\beta_3\text{Advertising Cost} \times \text{Store Size}\end{aligned}$$

The interaction term $\beta_3\text{Advertising Cost} \times \text{Store Size}$ captures the effect of the interaction between the two predictors.

A significantly positive β_3 indicates that more advertising cost is more effective in larger stores.

Interaction Terms

Usually in real practices, multiple predictors “interact” with each other.

$$\begin{aligned}\text{Sales} =&\beta_0 + \beta_1\text{Advertising Cost} + \beta_2\text{Store Size}\\ \text{Sales} =&\beta_0 + \beta_1\text{Advertising Cost} + \beta_2\text{Store Size}+\\ &+ \beta_3\text{Advertising Cost} \times \text{Store Size}\end{aligned}$$

The interaction term $\beta_3\text{Advertising Cost} \times \text{Store Size}$ captures the effect of the interaction between the two predictors.

A significantly positive β_3 indicates that more advertising cost is more effective in larger stores.

Explain the interaction terms first when interpreting the results.

Correlation vs. Causation

$$\text{Reading Time} \sim \text{Height} + \text{Vocabulary Size}$$

Randomly sample people of ages 3–70.

Result: $\beta_{\text{Height}} < 0(^{***}), \beta_{\text{Vocabulary Size}} < 0(^{***})$.

Correlation vs. Causation

$$\text{Reading Time} \sim \text{Height} + \text{Vocabulary Size}$$

Randomly sample people of ages 3–70.

Result: $\beta_{\text{Height}} < 0(^{***}), \beta_{\text{Vocabulary Size}} < 0(^{***})$.

Q: Does this mean that taller people read in a faster speed?

Correlation vs. Causation

$$\text{Reading Time} \sim \text{Height} + \text{Vocabulary Size}$$

Randomly sample people of ages 3–70.

Result: $\beta_{\text{Height}} < 0(^{***}), \beta_{\text{Vocabulary Size}} < 0(^{***})$.

Q: Does this mean that taller people read in a faster speed?

A: Yes and no.

Correlation vs. Causation

$$\text{Reading Time} \sim \text{Height} + \text{Vocabulary Size}$$

Randomly sample people of ages 3–70.

Result: $\beta_{\text{Height}} < 0(^{***}), \beta_{\text{Vocabulary Size}} < 0(^{***})$.

Q: Does this mean that taller people read in a faster speed?

A: Yes and no.

A more plausible explanation:

$$\text{Reading Time} \sim \text{Vocabulary Size} + \text{Age}$$
$$\text{Height} \sim \text{Age}$$

Correlation vs. Causation

$$\text{Reading Time} \sim \text{Height} + \text{Vocabulary Size}$$

Randomly sample people of ages 3–70.

Result: $\beta_{\text{Height}} < 0(^{***}), \beta_{\text{Vocabulary Size}} < 0(^{***})$.

Q: Does this mean that taller people read in a faster speed?

A: Yes and no.

A more plausible explanation:

$$\text{Reading Time} \sim \text{Vocabulary Size} + \text{Age}$$
$$\text{Height} \sim \text{Age}$$

To infer causation, we need to conduct a **controlled experiment**.

`https://www.r-causal.org/`

What's Not Covered?

The χ^2-test, measuring the difference between the observed and expected frequencies of the outcomes of a set of variables.

$$\chi^2_{DoF} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

What's Not Covered?

The χ^2-test, measuring the difference between the observed and expected frequencies of the outcomes of a set of variables.

$$\chi^2_{DoF} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

The F-test, comparing the variances of two samples.

$$F_{k-1,n-k} = \frac{s_1^2}{s_2^2},$$

where n is the sample size, k is the number of predictors, and s_1^2 and s_2^2 are the variances of the two samples.

When should we use frequentist vs. Bayesian methods?

Some philosophy of science:

- Frequentist hypothesis testing generally gives much higher efficiency.

When should we use frequentist vs. Bayesian methods?

Some philosophy of science:

- Frequentist hypothesis testing generally gives much higher efficiency.
- Frequentist hypothesis testing has an asymmetric treatment of the null (H_0) and alternative (H_1) hypotheses.
  The p-value from a dataset D shows how unlikely the dataset was to be produced under H_0.
  In some senses, the alternative hypothesis is never actively used!

When should we use frequentist vs. Bayesian methods?

Some philosophy of science:

- Frequentist hypothesis testing generally gives much higher efficiency.
- Frequentist hypothesis testing has an asymmetric treatment of the null (H_0) and alternative (H_1) hypotheses.
  The p-value from a dataset D shows how unlikely the dataset was to be produced under H_0.
  In some senses, the alternative hypothesis is never actively used!
- Bayesian hypothesis testing overcomes this asymmetric treatment by directly comparing the two hypotheses.

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid H_1)} \frac{P(H_0)}{P(H_1)}$$

When should we use frequentist vs. Bayesian methods?

Some philosophy of science:

- Frequentist hypothesis testing generally gives much higher efficiency.
- Frequentist hypothesis testing has an asymmetric treatment of the null (H_0) and alternative (H_1) hypotheses.
  The p-value from a dataset D shows how unlikely the dataset was to be produced under H_0.
  In some senses, the alternative hypothesis is never actively used!
- Bayesian hypothesis testing overcomes this asymmetric treatment by directly comparing the two hypotheses.

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid H_1)} \frac{P(H_0)}{P(H_1)}$$

- However, we have to specify the prior anyway, which may significantly affect the test results.
  Additionally, it could be slow to compute the posterior distribution.

Next

Morphology, tokenization