Statistics Review

Joshua Agterberg

Johns Hopkins University

Zoom

Outline

Figure: Source: https://sarahmarley.com/2015/07/30/why-statistics-is-not-just-maths/

Notes

Notes available at my website

Outline

Samples and Population

- We have a population distribution f_0 and a *model* $\mathcal{F} = \{f : f \in \mathcal{F}\}$
- Goal: extract some information about f_0 from $\mathcal{F}$.
- Examples:
  - Population follows a $N(\mu_0, \sigma_0^2)$ distribution, and from the set $\mathcal{F} := \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$
  - Population exhibits some probability p_0 of having an attribute (e.g. having COVID-19), and consider $\mathcal{F} = Binomial(n, p), p > 0$.
  - Population follows some continuous distribution f_0 and we set $\mathcal{F} = \{$ all continuous distributions$\}$.

Samples and Population

- We have a population distribution f_0 and a *model* $\mathcal{F} = \{f : f \in \mathcal{F}\}$
- Goal: extract some information about f_0 from $\mathcal{F}$.
- Examples:
  - Population follows a $N(\mu_0, \sigma_0^2)$ distribution, and from the set $\mathcal{F} := \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$
  - Population exhibits some probability p_0 of having an attribute (e.g. having COVID-19), and consider $\mathcal{F} = Binomial(n, p), p > 0$.
  - Population follows some continuous distribution f_0 and we set $\mathcal{F} = \{$ all continuous distributions$\}$.

Samples and Population

- We have a population distribution f_0 and a *model* $\mathcal{F} = \{f : f \in \mathcal{F}\}$
- Goal: extract some information about f_0 from $\mathcal{F}$.
- Examples:
  - Population follows a $N(\mu_0, \sigma_0^2)$ distribution, and from the set $\mathcal{F} := \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$
  - Population exhibits some probability p_0 of having an attribute (e.g. having COVID-19), and consider $\mathcal{F} = \mathit{Binomial}(n, p), p > 0$.
  - Population follows some continuous distribution f_0 and we set $\mathcal{F} = \{$ all continuous distributions$\}$.

Parametric Families

- If the family satisfies $\mathcal{F} := \{f_\theta : \theta \in \mathbb{R}^d\}$, then we say it is *parametric*
- Examples of parametric families:
  - Bernoulli: $X \sim Ber(p), P(X = 1) = p$
  - Binomial: $X \sim Bin(n, p), P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, $Bin(n, p) = \sum_{i=1}^{n} Ber(p)$
  - Normal: $X \sim N(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
    $Bin(n, p) \approx N(np, npq), \frac{N(\mu,\sigma^2)-\mu}{\sigma} = N(0, 1)$
  - Chi-square: $X \sim \chi^2_\nu, \chi^2_\nu = \sum_{i=1}^{\nu} N(0, 1)^2$
  - t-distribution: $X \sim t_\nu, t_\nu = \frac{N(0,1)}{\sqrt{\chi^2_\nu/\nu}}, t_\infty = N(0, 1), t_0 =$ Cauchy (undefined mean and variance)
  - F-distribution $X \sim F_{n,m}, F_{n,m} = \frac{\chi^2_n/n}{\chi_m/m}, t^2_\nu = F_{1,\nu}$
  - Others: Exponential, Poisson, Gamma, Beta, Negative Binomial, ...

Parametric Families

- If the family satisfies $\mathcal{F} := \{f_\theta : \theta \in \mathbb{R}^d\}$, then we say it is *parametric*
- Examples of parametric families:
  - Bernoulli: $X \sim Ber(p), P(X = 1) = p$
  - Binomial: $X \sim Bin(n, p), P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $Bin(n, p) = \sum_{i=1}^{n} Ber(p)$
  - Normal: $X \sim N(\mu, \sigma^2), f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $Bin(n, p) \approx N(np, npq)$, $\frac{N(\mu,\sigma^2)-\mu}{\sigma} = N(0, 1)$
  - Chi-square: $X \sim \chi^2_\nu$, $\chi^2_\nu = \sum_{i=1}^{\nu} N(0, 1)^2$
  - t-distribution: $X \sim t_\nu$, $t_\nu = \frac{N(0,1)}{\sqrt{\chi^2_\nu/\nu}}$, $t_\infty = N(0, 1)$, $t_0 =$ Cauchy (undefined mean and variance)
  - F-distribution $X \sim F_{n,m}$, $F_{n,m} = \frac{\chi^2_n/n}{\chi_m/m}$, $t^2_\nu = F_{1,\nu}$
  - Others: Exponential, Poisson, Gamma, Beta, Negative Binomial, ...

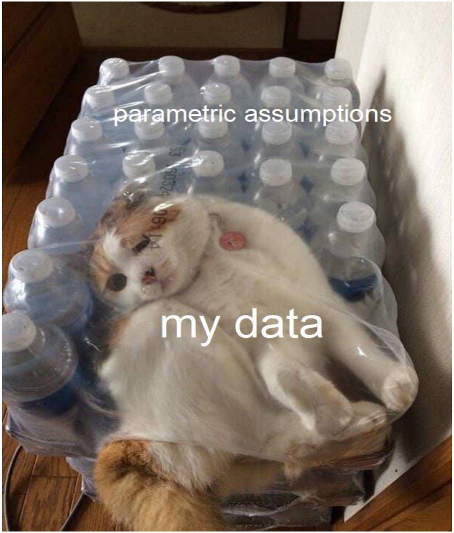

Figure: Source: https://www.facebook.com/statsmemes/photos/a.306077739764526/975685

Nonparametric Families

- If the family $\mathcal{F}$ is infinite-dimensional, we (typically) say it is *nonparametric* (Tsybakov, 2008)
- Semiparametric out-of-scope (Bickel et al., 1998)
- Examples of nonparametric families
  - $\mathcal{F} := \{f : \mathbb{E}_f |X|^2 < \infty\}$; i.e. the set of distributions with finite second moment
  - $\mathcal{F} := \{f : f \text{ is a continuous density}\}$
  - $\mathcal{F} := \{f : f \text{ is infinitely differentiable}\}$
  - $\mathcal{F} := \{f : f$ has the property that $\log f(tx + (1-t)y) \geq t \log f(x) + (1-t) \log f(y)$ (log-concave distributions see Samworth)
  - $\mathcal{F} := \{f : f \text{ is has continuous derivatives up to order } r\}$

Nonparametric Families

- If the family $\mathcal{F}$ is infinite-dimensional, we (typically) say it is *nonparametric* (Tsybakov, 2008)
- Semiparametric out-of-scope (Bickel et al., 1998)
- Examples of nonparametric families
  - $\mathcal{F} := \{f : \mathbb{E}_f|X|^2 < \infty\}$; i.e. the set of distributions with finite second moment
  - $\mathcal{F} := \{f : f \text{ is a continuous density}\}$
  - $\mathcal{F} := \{f : f \text{ is infinitely differentiable}\}$
  - $\mathcal{F} := \{f : f$ has the property that $\log f(tx + (1-t)y) \geq t \log f(x) + (1-t)\log f(y)$ (log-concave distributions see Samworth)
  - $\mathcal{F} := \{f : f \text{ is has continuous derivatives up to order } r\}$

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

Main Ideas

- We observe a sample $X_1, ..., X_n$ iid from f_0.
- Assuming that there is a true parameter θ (e.g. the mean, the variance, etc.), can we use our data to study the true distribution? (Frequentist method). We can:
  - estimate θ,
  - perform a hypothesis test,
  - or find a confidence interval about the true parameter.
- Always pay attention to assumptions! In many cases, assumptions do not hold, but they make our lives easier.
- If we know exact distributions, we can perform inference exactly
- Otherwise, we study asymptotics (van der Vaart, 2000)
- Often much easier to study asymptotic results than finite-sample results (Bickel and Doksum, 2007)

A note on prerequisites

- Much of this material is covered in an introductory statistics course
- However some of it (e.g. R code, material at the end) may be new
- My hope is to leave you with a basic idea of both the mathematics and the philosophy of statistical inference, so that even new material is not difficult
- 553.630 covers statistical theory at the upper undergraduate/graduate level in primarily parametric settings with an emphasis on explicit calculations
- 553.730 covers statistical theory at the graduate level with an emphasis on proving results for parametric families

A note on prerequisites

- Much of this material is covered in an introductory statistics course
- However some of it (e.g. R code, material at the end) may be new
- My hope is to leave you with a basic idea of both the mathematics and the philosophy of statistical inference, so that even new material is not difficult
- 553.630 covers statistical theory at the upper undergraduate/graduate level in primarily parametric settings with an emphasis on explicit calculations
- 553.730 covers statistical theory at the graduate level with an emphasis on proving results for parametric families

A note on prerequisites

- Much of this material is covered in an introductory statistics course
- However some of it (e.g. R code, material at the end) may be new
- My hope is to leave you with a basic idea of both the mathematics and the philosophy of statistical inference, so that even new material is not difficult
- 553.630 covers statistical theory at the upper undergraduate/graduate level in primarily parametric settings with an emphasis on explicit calculations
- 553.730 covers statistical theory at the graduate level with an emphasis on proving results for parametric families

A note on prerequisites

- Much of this material is covered in an introductory statistics course
- However some of it (e.g. R code, material at the end) may be new
- My hope is to leave you with a basic idea of both the mathematics and the philosophy of statistical inference, so that even new material is not difficult
- 553.630 covers statistical theory at the upper undergraduate/graduate level in primarily parametric settings with an emphasis on explicit calculations
- 553.730 covers statistical theory at the graduate level with an emphasis on proving results for parametric families

A note on prerequisites

- Much of this material is covered in an introductory statistics course
- However some of it (e.g. R code, material at the end) may be new
- My hope is to leave you with a basic idea of both the mathematics and the philosophy of statistical inference, so that even new material is not difficult
- 553.630 covers statistical theory at the upper undergraduate/graduate level in primarily parametric settings with an emphasis on explicit calculations
- 553.730 covers statistical theory at the graduate level with an emphasis on proving results for parametric families

Outline

2. Exact Parametric Methods
   - Estimation
   - One-Sample Testing
   - Two-Sample Testing

Exact Parametric Methods

In some cases, if we assume the population has a distribution, we can explicitly characterize the finite-sample distribution

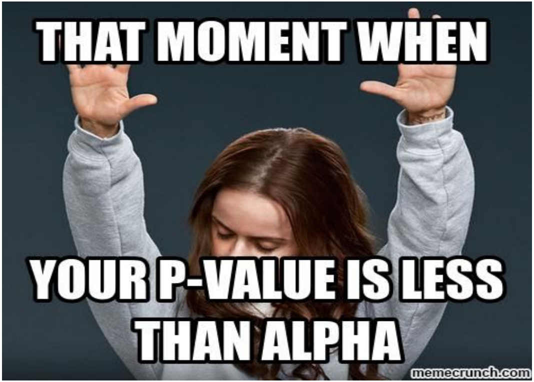

Figure: Source: https://www.pinterest.com/pin/246853623302262497/

Estimation

- Data: $X_i \sim N(\mu, \sigma^2)$
- Estimator: $\hat{\mu} = \bar{X}$
- Distribution: $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ ("proof": $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}}$)
- C.I.: $\bar{X} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$

We know the *exact* distribution when $X_i \sim N(\mu, \sigma^2)$.

Estimation

- Data: $X_i \sim N(\mu, \sigma^2)$
- Estimator: $\hat{\mu} = \bar{X}$
- Distribution: $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ ("proof": $\frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}}$)
- C.I.: $\bar{X} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$

We know the *exact* distribution when $X_i \sim N(\mu, \sigma^2)$.

One-Sample Testing

- If $X_i \sim N(\mu, \sigma^2)$, we want to test whether the mean is equal to μ_0
- Form the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

- Under the null $\mu = \mu_0$, the data $X_i \sim N(\mu_0, \sigma^2)$
- Form the *test statistic* $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$, where s is the *sample standard deviation*
- Reject at level α if $|T| > t_{n-1}(\alpha/2)$

One-Sample Testing

- If $X_i \sim N(\mu, \sigma^2)$, we want to test whether the mean is equal to μ_0
- Form the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

- Under the null $\mu = \mu_0$, the data $X_i \sim N(\mu_0, \sigma^2)$
- Form the *test statistic* $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$, where s is the *sample standard deviation*
- Reject at level α if $|T| > t_{n-1}(\alpha/2)$

One-Sample Testing

- If $X_i \sim N(\mu, \sigma^2)$, we want to test whether the mean is equal to μ_0
- Form the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

- Under the null $\mu = \mu_0$, the data $X_i \sim N(\mu_0, \sigma^2)$
- Form the *test statistic* $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$, where s is the *sample standard deviation*
- Reject at level α if $|T| > t_{n-1}(\alpha/2)$

One-Sample Testing

- If $X_i \sim N(\mu, \sigma^2)$, we want to test whether the mean is equal to μ_0
- Form the hypotheses

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

- Under the null $\mu = \mu_0$, the data $X_i \sim N(\mu_0, \sigma^2)$
- Form the *test statistic* $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$, where s is the *sample standard deviation*
- Reject at level α if $|T| > t_{n-1}(\alpha/2)$

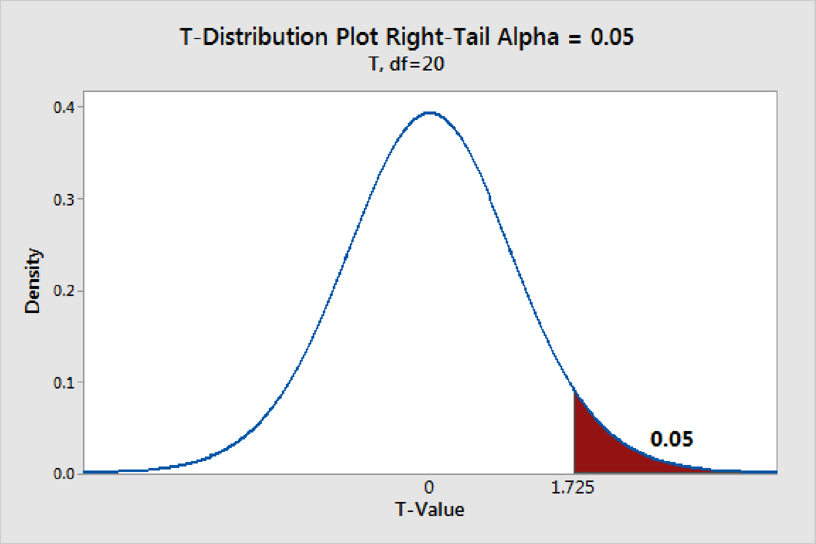

Figure: Source: https://statisticsbyjim.com/hypothesis-testing/one-tailed-two-tailed-hypothesis-tests/

Two-Sample Testing

- $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$
- Want to test $\mu_X = \mu_Y$
- Form the hypothesis:

$$H_0 : \mu_X = \mu_Y$$
$$H_A : \mu_x \neq \mu_Y$$

- Test statistic: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} \sim t_{n-1}$, where

$$s_p^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

- Rejection region: $|T| > t_{n-1}(\alpha/2)$

Two-Sample Testing

- $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$
- Want to test $\mu_X = \mu_Y$
- Form the hypothesis:

$$H_0 : \mu_X = \mu_Y$$
$$H_A : \mu_x \neq \mu_Y$$

- Test statistic: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} \sim t_{n-1}$, where

$$s_p^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

- Rejection region: $|T| > t_{n-1}(\alpha/2)$

Two-Sample Testing

- $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$
- Want to test $\mu_X = \mu_Y$
- Form the hypothesis:

$$H_0 : \mu_X = \mu_Y$$
$$H_A : \mu_x \neq \mu_Y$$

- Test statistic: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} \sim t_{n-1}$, where

$$s_p^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

- Rejection region: $|T| > t_{n-1}(\alpha/2)$

Two-Sample Testing

- $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$
- Want to test $\mu_X = \mu_Y$
- Form the hypothesis:

$$H_0 : \mu_X = \mu_Y$$
$$H_A : \mu_x \neq \mu_Y$$

- Test statistic: $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}} \sim t_{n-1}$, where

$$s_p^2 := \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

- Rejection region: $|T| > t_{n-1}(\alpha/2)$

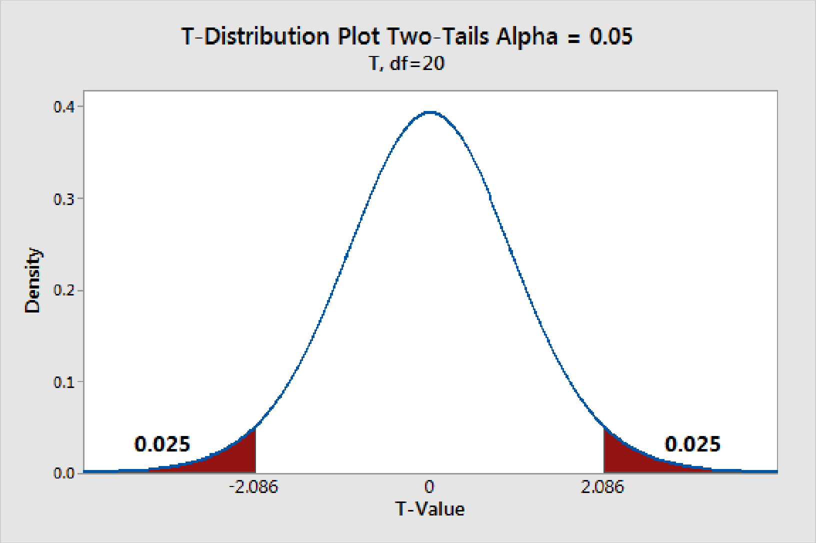

Figure: Source: https://statisticsbyjim.com/hypothesis-testing/one-tailed-two-tailed-hypothesis-tests/

Main Ideas

- Know exact distribution of data
- Calculate its exact distribution under the null H_0 (both cases, we had normal data, and had to estimate σ)
- Test whether we would observe the value of the test statistic under the null hypothesis
- Could do for other parameters of interest (σ^2, multivariate means, covariances)
- Duality between confidence interval and Hypothesis testing

Main Ideas

- Know exact distribution of data
- Calculate its exact distribution under the null H_0 (both cases, we had normal data, and had to estimate σ)
- Test whether we would observe the value of the test statistic under the null hypothesis
- Could do for other parameters of interest (σ^2, multivariate means, covariances)
- Duality between confidence interval and Hypothesis testing

Main Ideas

- Know exact distribution of data
- Calculate its exact distribution under the null H_0 (both cases, we had normal data, and had to estimate σ)
- Test whether we would observe the value of the test statistic under the null hypothesis
- Could do for other parameters of interest (σ^2, multivariate means, covariances)
- Duality between confidence interval and Hypothesis testing

Main Ideas

- Know exact distribution of data
- Calculate its exact distribution under the null H_0 (both cases, we had normal data, and had to estimate σ)
- Test whether we would observe the value of the test statistic under the null hypothesis
- Could do for other parameters of interest (σ^2, multivariate means, covariances)
- Duality between confidence interval and Hypothesis testing

Main Ideas

- Know exact distribution of data
- Calculate its exact distribution under the null H_0 (both cases, we had normal data, and had to estimate σ)
- Test whether we would observe the value of the test statistic under the null hypothesis
- Could do for other parameters of interest (σ^2, multivariate means, covariances)
- Duality between confidence interval and Hypothesis testing

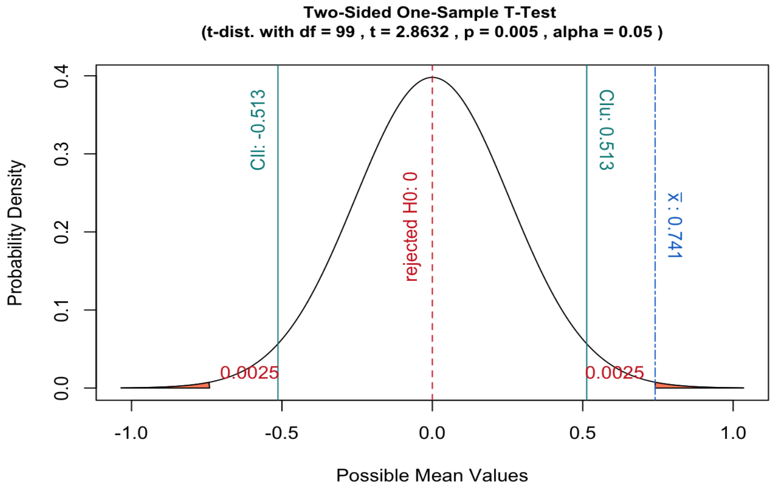

Figure: Source: https://stats.stackexchange.com/questions/220434/hypothesis-testing-why-center-the-sampling-distribution-on-h0

Outline

3. Large-Sample Parametric Methods
   - Central Limit Theorem
   - Estimation and Testing for Proportions
   - Estimation and Testing for More General Parametric Families

Large-Sample Concepts

- In many cases, we do not know exact distribution of the data
- Nevertheless, with enough samples, we can use the asymptotic results from probability theory, namely the Central Limit Theorem

Large-Sample Concepts

- In many cases, we do not know exact distribution of the data
- Nevertheless, with enough samples, we can use the asymptotic results from probability theory, namely the Central Limit Theorem

Central Limit Theorem

- $X_1, ..., X_n \sim F$ iid
- Define

$$S_n := \sum_{i=1}^{n} X_i$$

- Then as $n \to \infty$, we have that

$$\frac{S_n - n\mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

- Idea for inference: if we can write a test statistic in terms of iid summands, then we can use the CLT to perform hypothesis tests

Central Limit Theorem

- $X_1, ..., X_n \sim F$ iid
- Define

$$S_n := \sum_{i=1}^{n} X_i$$

- Then as $n \to \infty$, we have that

$$\frac{S_n - n\mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

- Idea for inference: if we can write a test statistic in terms of iid summands, then we can use the CLT to perform hypothesis tests

Central Limit Theorem

- $X_1, ..., X_n \sim F$ iid
- Define

$$S_n := \sum_{i=1}^{n} X_i$$

- Then as $n \to \infty$, we have that

$$\frac{S_n - n\mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

- Idea for inference: if we can write a test statistic in terms of iid summands, then we can use the CLT to perform hypothesis tests

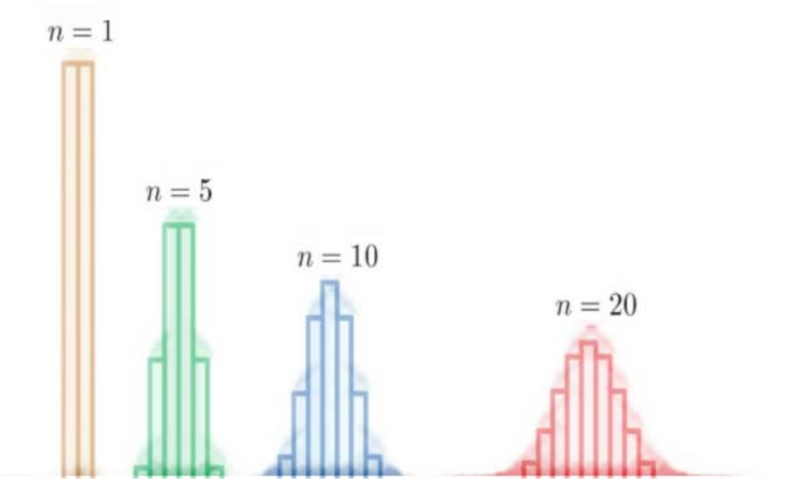

Figure: Source: http://www.marketexpress.in/2016/11/central-limit-theorem-normal-distribution.html

Estimation of Proportions Using the CLT

- For testing proportions (presence or absence of a characteristic), for a fixed sample of size n the distribution is $Binom(n, p)$, where $p = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{person } i \text{ has the characteristic})$
- Want to either estimate p or perform Hypothesis test
- Example
  - $H_0 : \mathbb{P}(\text{drug X works}) = .8$
- Observation: for large n, by CLT

$$\frac{S_n - np}{\sigma/\sqrt{n}} \approx N(0, 1),$$

where $S_n = \sum_{i=1}^{n} X_i$.

Estimation of Proportions Using the CLT

- For testing proportions (presence or absence of a characteristic), for a fixed sample of size n the distribution is $Binom(n, p)$, where $p = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{person } i \text{ has the characteristic})$
- Want to either estimate p or perform Hypothesis test
- Example
  - $H_0 : \mathbb{P}(\text{drug X works}) = .8$
- Observation: for large n, by CLT

$$\frac{S_n - np}{\sigma/\sqrt{n}} \approx N(0, 1),$$

where $S_n = \sum_{i=1}^{n} X_i$.

Estimation of Proportions Using the CLT

- For testing proportions (presence or absence of a characteristic), for a fixed sample of size n the distribution is $Binom(n, p)$, where $p = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{person } i \text{ has the characteristic})$
- Want to either estimate p or perform Hypothesis test
- Example
  - $H_0 : \mathbb{P}(\text{drug X works}) = .8$
- Observation: for large n, by CLT

$$\frac{S_n - np}{\sigma/\sqrt{n}} \approx N(0, 1),$$

where $S_n = \sum_{i=1}^{n} X_i$.

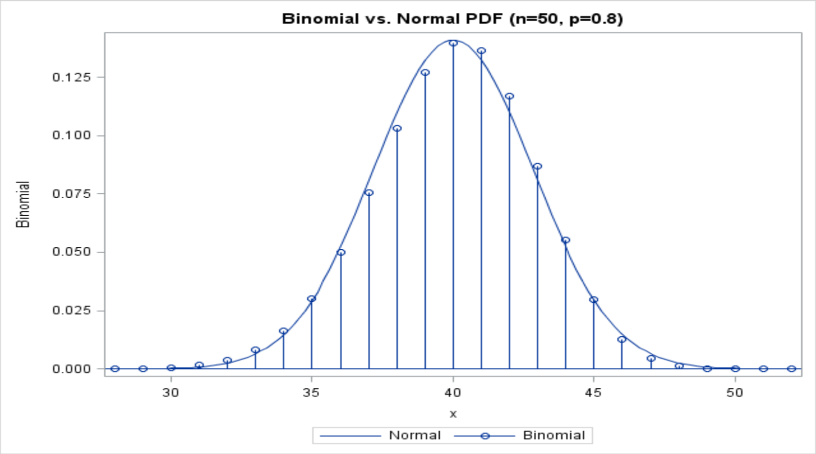

Figure: Source: https://blogs.sas.com/content/iml/2012/03/14/the-normal-approximation-to-the-binomial-distribution-how-the-quantiles-compare.html

One-Sample Testing for Proportions using the CLT

- Hypothesis: $H_0 : p = p_0$ vs. $H_A : p \neq p_0$
- Test statistics: $Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0, 1)$
- Rejection region: $|Z| > z(\alpha/2)$

Two-Sample Testing for Proportions using the CLT

- Hypothesis: $H_0 : p_X - p_Y = D_0$ vs. $H_A : p_X - p_Y \neq D_0$
- Test statistics: $Z = \frac{\hat{p}_X - \hat{p}_Y - D_0}{\sqrt{\frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}}} \sim N(0,1)$
- Rejection region: $|Z| > z(\alpha/2)$

Other Tests

- Can perform tests for variance, goodness-of-fit, etc., using CLT
- Idea is if somehow can write test statistic $T \approx \frac{S_n - n\mu}{s/\sqrt{n}}$, then it is approximately $N(0,1)$.
- Other distributions that arise from asymptotics:
  - ξ^2 distribution (e.g. $T^2 \approx N(0,1)^2 \approx \xi^2(1)$
  - F is a ratio of ξ^2, so comes when analyzing variance
- See notes for more details on other tests
- Type I error: α, and Type II error = $\mathbb{P}$(error if H_0 is false).

Other Tests

- Can perform tests for variance, goodness-of-fit, etc., using CLT
- Idea is if somehow can write test statistic $T \approx \frac{S_n - n\mu}{s/\sqrt{n}}$, then it is approximately $N(0, 1)$.
- Other distributions that arise from asymptotics:
  - ξ^2 distribution (e.g. $T^2 \approx N(0, 1)^2 \approx \xi^2(1)$
  - F is a ratio of ξ^2, so comes when analyzing variance
- See notes for more details on other tests
- Type I error: α, and Type II error = $\mathbb{P}$(error if H_0 is false).

Other Tests

- Can perform tests for variance, goodness-of-fit, etc., using CLT
- Idea is if somehow can write test statistic $T \approx \frac{S_n - n\mu}{s/\sqrt{n}}$, then it is approximately $N(0, 1)$.
- Other distributions that arise from asymptotics:
  - ξ^2 distribution (e.g. $T^2 \approx N(0, 1)^2 \approx \xi^2(1)$
  - F is a ratio of ξ^2, so comes when analyzing variance
- See notes for more details on other tests
- Type I error: α, and Type II error = $\mathbb{P}$(error if H_0 is false).

Other Tests

- Can perform tests for variance, goodness-of-fit, etc., using CLT
- Idea is if somehow can write test statistic $T \approx \frac{S_n - n\mu}{s/\sqrt{n}}$, then it is approximately $N(0, 1)$.
- Other distributions that arise from asymptotics:
  - ξ^2 distribution (e.g. $T^2 \approx N(0, 1)^2 \approx \xi^2(1)$
  - F is a ratio of ξ^2, so comes when analyzing variance
- See notes for more details on other tests
- Type I error: α, and Type II error = $\mathbb{P}$(error if H_0 is false).

Other Tests

- Can perform tests for variance, goodness-of-fit, etc., using CLT
- Idea is if somehow can write test statistic $T \approx \frac{S_n - n\mu}{s/\sqrt{n}}$, then it is approximately $N(0,1)$.
- Other distributions that arise from asymptotics:
  - ξ^2 distribution (e.g. $T^2 \approx N(0,1)^2 \approx \xi^2(1)$
  - F is a ratio of ξ^2, so comes when analyzing variance
- See notes for more details on other tests
- Type I error: α, and Type II error = $\mathbb{P}$(error if H_0 is false).

Other Tests

- Can perform tests for variance, goodness-of-fit, etc., using CLT
- Idea is if somehow can write test statistic $T \approx \frac{S_n - n\mu}{s/\sqrt{n}}$, then it is approximately $N(0, 1)$.
- Other distributions that arise from asymptotics:
  - ξ^2 distribution (e.g. $T^2 \approx N(0, 1)^2 \approx \xi^2(1)$
  - F is a ratio of ξ^2, so comes when analyzing variance
- See notes for more details on other tests
- Type I error: α, and Type II error = $\mathbb{P}$(error if H_0 is false).

Consistency

- Suppose $X_1, ..., X_n$ are iid f_θ, for $\theta \in \Theta$
- Inference on θ is a bit more complicated than just applying the CLT
- Want an estimator $\hat{\theta}$ that uses the data such that $\hat{\theta}_n \to \theta$ in probability, where this means

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) = 0$$

for all $\varepsilon > 0$.

Consistency

- Suppose $X_1, ..., X_n$ are iid f_θ, for $\theta \in \Theta$
- Inference on θ is a bit more complicated than just applying the CLT
- Want an estimator $\hat{\theta}$ that uses the data such that $\hat{\theta}_n \to \theta$ in probability, where this means

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) = 0$$

for all $\varepsilon > 0$.

Consistency

- Suppose $X_1, ..., X_n$ are iid f_θ, for $\theta \in \Theta$
- Inference on θ is a bit more complicated than just applying the CLT
- Want an estimator $\hat{\theta}$ that uses the data such that $\hat{\theta}_n \to \theta$ in probability, where this means

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) = 0$$

for all $\varepsilon > 0$.

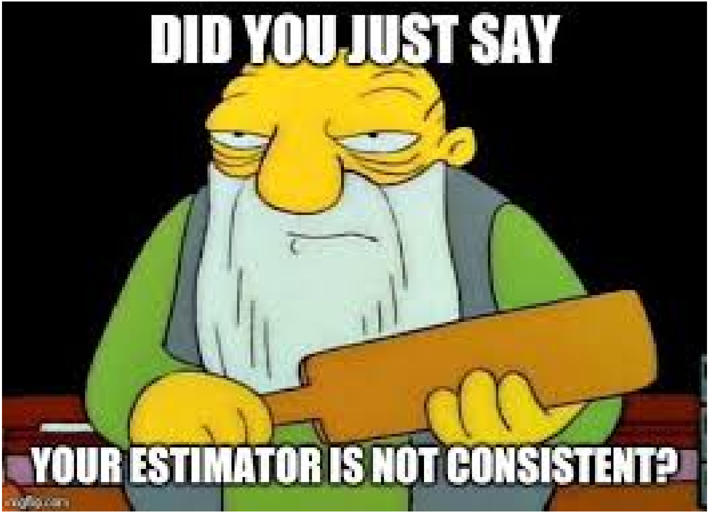

Figure: Source: https://www.facebook.com/StatisticalMemes/

Consistency Example

Example: $X_1, ..., X_n \sim U(0, \theta)$, $\theta > 0$.

Set $\hat{\theta} := \max_{1 \le i \le n} X_i$. Then

$$\begin{aligned}
\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) &= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + \mathbb{P}(\hat{\theta} - \theta > \varepsilon) && (1)\\
&= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + 0 && (2)\\
&= \mathbb{P}(\theta - \varepsilon > \hat{\theta}) = \mathbb{P}(\max_{1 \le i \le n} X_i < \theta - \varepsilon) && (3)\\
&= \mathbb{P}(X_1 < \theta - \varepsilon, ..., X_n < \theta - \varepsilon) && (4)\\
&= \Big(\mathbb{P}(X_1 < \theta - \varepsilon)\Big)^n = \left(\frac{\theta - \varepsilon}{\theta}\right)^n && (5)
\end{aligned}$$

Where (3) is since $\hat{\theta} < \theta$ always, and by definition, (4) is because $\max X_i < c$ if and only if all $X_i < c$, (5) is because the X_i's are iid and the CDF of the uniform distribution.

Consistency Example

Example: $X_1, ..., X_n \sim U(0, \theta)$, $\theta > 0$.
Set $\hat{\theta} := \max_{1 \le i \le n} X_i$. Then

$$\begin{aligned}
\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) &= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + \mathbb{P}(\hat{\theta} - \theta > \varepsilon) && (1) \\
&= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + 0 && (2) \\
&= \mathbb{P}(\theta - \varepsilon > \hat{\theta}) = \mathbb{P}(\max_{1 \le i \le n} X_i < \theta - \varepsilon) && (3) \\
&= \mathbb{P}(X_1 < \theta - \varepsilon, ..., X_n < \theta - \varepsilon) && (4) \\
&= \Big(\mathbb{P}(X_1 < \theta - \varepsilon)\Big)^n = \left(\frac{\theta - \varepsilon}{\theta}\right)^n && (5)
\end{aligned}$$

Where (3) is since $\hat{\theta} < \theta$ always, and by definition, (4) is because $\max X_i < c$ if and only if all $X_i < c$, (5) is because the X_i's are iid and the CDF of the uniform distribution.

Consistency Example

Example: $X_1, ..., X_n \sim U(0, \theta)$, $\theta > 0$.
Set $\hat{\theta} := \max_{1 \le i \le n} X_i$. Then

$$\begin{align}
\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) &= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + \mathbb{P}(\hat{\theta} - \theta > \varepsilon) \tag{1} \\
&= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + 0 \tag{2} \\
&= \mathbb{P}(\theta - \varepsilon > \hat{\theta}) = \mathbb{P}(\max_{1 \le i \le n} X_i < \theta - \varepsilon) \tag{3} \\
&= \mathbb{P}(X_1 < \theta - \varepsilon, ..., X_n < \theta - \varepsilon) \tag{4} \\
&= \Big(\mathbb{P}(X_1 < \theta - \varepsilon)\Big)^n = \left(\frac{\theta - \varepsilon}{\theta}\right)^n \tag{5}
\end{align}$$

Where (3) is since $\hat{\theta} < \theta$ always, and by definition, (4) is because $\max X_i < c$ if and only if all $X_i < c$, (5) is because the X_i's are iid and the CDF of the uniform distribution.

Consistency Example

Example: $X_1, ..., X_n \sim U(0, \theta)$, $\theta > 0$.
Set $\hat{\theta} := \max_{1 \le i \le n} X_i$. Then

$$
\begin{aligned}
\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) &= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + \mathbb{P}(\hat{\theta} - \theta > \varepsilon) && (1)\\
&= \mathbb{P}(\theta - \hat{\theta} > \varepsilon) + 0 && (2)\\
&= \mathbb{P}(\theta - \varepsilon > \hat{\theta}) = \mathbb{P}(\max_{1 \le i \le n} X_i < \theta - \varepsilon) && (3)\\
&= \mathbb{P}(X_1 < \theta - \varepsilon, ..., X_n < \theta - \varepsilon) && (4)\\
&= \Big(\mathbb{P}(X_1 < \theta - \varepsilon)\Big)^n = \left(\frac{\theta - \varepsilon}{\theta}\right)^n && (5)
\end{aligned}
$$

Where (3) is since $\hat{\theta} < \theta$ always, and by definition, (4) is because $\max X_i < c$ if and only if all $X_i < c$, (5) is because the X_i's are iid and the CDF of the uniform distribution.

Consistency Example

- Hence, we see

$$\mathbb{P}|\hat{\theta} - \theta| > \varepsilon) = \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

  which tends to zero for all $\varepsilon > 0$ since the term in the parentheses is less than 1.

- So for $X_1, ..., X_n \sim U(0, \theta)$, $\hat{\theta} = \max_i X_i$ is *consistent* for θ.
- But

$$\mathbb{E}(\hat{\theta}) = \frac{n}{n+1}\theta \qquad \text{(check!)}$$

  which does not equal θ! (it is biased)

Consistency Example

- Hence, we see

$$\mathbb{P}|\hat{\theta} - \theta| > \varepsilon) = \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

which tends to zero for all $\varepsilon > 0$ since the term in the parentheses is less than 1.

- So for $X_1, ..., X_n \sim U(0, \theta)$, $\hat{\theta} = \max_i X_i$ is *consistent* for θ.
- But

$$\mathbb{E}(\hat{\theta}) = \frac{n}{n+1}\theta \qquad \text{(check!)}$$

which does not equal θ! (it is biased)

Consistency Example

- Hence, we see

$$\mathbb{P}|\hat{\theta} - \theta| > \varepsilon) = \left(\frac{\theta - \varepsilon}{\theta}\right)^{n}$$

which tends to zero for all $\varepsilon > 0$ since the term in the parentheses is less than 1.

- So for $X_1, ..., X_n \sim U(0, \theta)$, $\hat{\theta} = \max_i X_i$ is *consistent* for θ.
- But

$$\mathbb{E}(\hat{\theta}) = \frac{n}{n+1}\theta \qquad \text{(check!)}$$

which does not equal θ! (it is biased)

Bias-Variance Tradeoff

- Define the bias: $\mathbb{E}(\hat{\theta}) - \theta$. Say $\hat{\theta}$ is *unbiased* if bias $= 0$
- Define the Mean-Squared Error (MSE):

$$
\begin{aligned}
\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta} - \theta)^2 + 2((\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta))\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right] + (\mathbb{E}\hat{\theta} - \theta)^2 + 2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right] + (\mathbb{E}\hat{\theta} - \theta)^2 \\
&= \text{Variance}(\hat{\theta}) + \text{Bias}^2
\end{aligned}
$$

Bias-Variance Tradeoff

- Define the bias: $\mathbb{E}(\hat{\theta}) - \theta$. Say $\hat{\theta}$ is *unbiased* if bias $= 0$
- Define the Mean-Squared Error (MSE):

$$
\begin{aligned}
\mathbb{E}\left[(\hat{\theta}-\theta)^2\right] &= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}\hat{\theta})^2+(\mathbb{E}\hat{\theta}-\theta)^2+2((\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta))\right] \\
&= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}\hat{\theta})^2\right]+(\mathbb{E}\hat{\theta}-\theta)^2+2\mathbb{E}\left[(\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta)\right] \\
&= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}\hat{\theta})^2\right]+(\mathbb{E}\hat{\theta}-\theta)^2 \\
&= \text{Variance}(\hat{\theta}) + \text{Bias}^2
\end{aligned}
$$

Bias-Variance Tradeoff

- Define the bias: $\mathbb{E}(\hat{\theta}) - \theta$. Say $\hat{\theta}$ is *unbiased* if bias $= 0$
- Define the Mean-Squared Error (MSE):

$$
\begin{aligned}
\mathbb{E}\left[(\hat{\theta} - \theta)^2\right] &= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta)^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + (\mathbb{E}\hat{\theta} - \theta)^2 + 2\big((\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)\big)\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right] + (\mathbb{E}\hat{\theta} - \theta)^2 + 2\mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\hat{\theta})^2\right] + (\mathbb{E}\hat{\theta} - \theta)^2 \\
&= \text{Variance}(\hat{\theta}) + \text{Bias}^2
\end{aligned}
$$

Example Continued

- For $X_1, ..., X_n \sim U(0, \theta)$, we saw $\mathbb{E}\hat{\theta} = \frac{n}{n+1}]\theta$, which is biased
- The variance is $\frac{n}{(n+1)^2(n+2)}\theta^2$ (do this!)
- So the MSE is:

$$\text{Variance} + \text{Bias}^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{1}{n+1}\right)^2 \theta^2$$

- Note that as $n \to \infty$, $MSE \to 0$.

Example Continued

- For $X_1, ..., X_n \sim U(0, \theta)$, we saw $\mathbb{E}\hat{\theta} = \frac{n}{n+1}]\theta$, which is biased
- The variance is $\frac{n}{(n+1)^2(n+2)}\theta^2$ (do this!)
- So the MSE is:

$$\text{Variance} + \text{Bias}^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{1}{n+1}\right)^2\theta^2$$

- Note that as $n \to \infty$, $MSE \to 0$.

Example Continued

- For $X_1, ..., X_n \sim U(0, \theta)$, we saw $\mathbb{E}\hat{\theta} = \frac{n}{n+1}]\theta$, which is biased
- The variance is $\frac{n}{(n+1)^2(n+2)}\theta^2$ (do this!)
- So the MSE is:

$$\text{Variance} + \text{Bias}^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{1}{n+1}\right)^2\theta^2$$

- Note that as $n \to \infty$, $MSE \to 0$.

Example Continued

- For $X_1, ..., X_n \sim U(0, \theta)$, we saw $\mathbb{E}\hat{\theta} = \frac{n}{n+1}]\theta$, which is biased
- The variance is $\frac{n}{(n+1)^2(n+2)}\theta^2$ (do this!)
- So the MSE is:

$$\text{Variance} + \text{Bias}^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(\frac{1}{n+1}\right)^2\theta^2$$

- Note that as $n \to \infty$, $MSE \to 0$.

Method of Moments

- The Method of Moments estimates the sample moments via

$$\hat{\mu}_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

- Examples:
  - $X \sim Poi(\lambda) : \mu_1 = \lambda \Rightarrow \hat{\mu}_1 = \bar{X}$ and $\hat{\lambda} = \bar{X}$
  - $X \sim N(\mu, \sigma^2) : \mu_1 = \mu, \mu_2 = \mu^2 + \sigma^2 \Rightarrow \hat{\mu} = \bar{X}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (biased)
  - $X \sim \Gamma(\alpha, \beta) : \mu_1 = \alpha/\beta, \mu_2 = \frac{\alpha(\alpha+1)}{\beta^2} \Rightarrow \hat{\beta} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}, \alpha = \hat{\beta}\hat{\mu}_1$
  - $X \sim U(0, \theta) : \mu_1 = \theta \Rightarrow \hat{\theta} = 2\bar{X}$ (could make no sense)
- Pros: easy, consistent, asymptotically unbiased
- Cons: could make no sense, not efficient

Method of Moments

- The Method of Moments estimates the sample moments via

$$\hat{\mu}_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

- Examples:
  - $X \sim Poi(\lambda) : \mu_1 = \lambda \Rightarrow \hat{\mu}_1 = \bar{X}$ and $\hat{\lambda} = \bar{X}$
  - $X \sim N(\mu, \sigma^2) : \mu_1 = \mu, \mu_2 = \mu^2 + \sigma^2 \Rightarrow \hat{\mu} = \bar{X}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ (biased)
  - $X \sim \Gamma(\alpha, \beta) : \mu_1 = \alpha/\beta, \mu_2 = \frac{\alpha(\alpha+1)}{\beta^2} \Rightarrow \hat{\beta} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}, \alpha = \hat{\beta}\hat{\mu}_1$
  - $X \sim U(0, \theta) : \mu_1 = \theta \Rightarrow \hat{\theta} = 2\bar{X}$ (could make no sense)
- Pros: easy, consistent, asymptotically unbiased
- Cons: could make no sense, not efficient

Maximum Likelihood

- $\hat{\theta} = \arg\max lik(\theta) = \arg\max \prod_{i=1}^{n} f(X_i|\theta)$
- $\hat{\theta} = \arg\max l(\theta) = \arg\max \sum_{i=1}^{n} \log f(X_i|\theta)$
- Example: $X \sim Poi(\lambda)$: $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$l(\lambda) = \sum_{i=1}^{n} (X_i \log \lambda - \lambda - \log X!)$$

$$= log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X!$$

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n = 0 \Rightarrow \hat{\lambda} = \bar{X}$$

Maximum Likelihood

- $\hat{\theta} = \arg\max lik(\theta) = \arg\max \prod_{i=1}^{n} f(X_i|\theta)$
- $\hat{\theta} = \arg\max l(\theta) = \arg\max \sum_{i=1}^{n} \log f(X_i|\theta)$
- Example: $X \sim Poi(\lambda)$: $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

$$l(\lambda) = \sum_{i=1}^{n} (X_i \log \lambda - \lambda - \log X!)$$

$$= log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X!$$

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n = 0 \Rightarrow \hat{\lambda} = \bar{X}$$

Properties of the MLE

- Asymptotically unbiased
- Consistent (consistency is the least we can ask for!)
- *Efficient*, which means that it achieves the Cramer-Rao Lower Bound, or that

$$\sqrt{nI(\theta)}(\hat{\theta}_{MLE} - \theta) \to N(0, 1),$$

and $I(\theta) := E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]^2$ (Fisher information)

Properties of the MLE

- Asymptotically unbiased
- Consistent (consistency is the least we can ask for!)
- *Efficient*, which means that it achieves the Cramer-Rao Lower Bound, or that

$$\sqrt{nI(\theta)}(\hat{\theta}_{MLE} - \theta) \to N(0, 1),$$

and $I(\theta) := E\left[\frac{\partial}{\partial\theta} \log f(X|\theta)\right]^2$ (Fisher information)

Likelihood Ratio Test

- Hypothesis: $H_0 : \mu = \mu_0; H_A : \mu = \mu_A$
- Test statistic: $\Lambda = \frac{f(X|H_0)}{f(X|H_A)}$ (ratio of likelihoods)
- Rejection region: small value of $\Lambda(X)$
- Most powerful for simple null vs. simple alternative
- Example: $N(\mu, \sigma)$ with σ known
  - $H_0 : \mu = \mu_0; H_A : \mu = \mu_A$
  - $\Lambda = \frac{\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_0)^2]}{\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_A)^2]}$
  - Reject for small $\sum_{i=1}^n (X_i - \mu_A)^2 - \sum_{i=1}^n (X_i - \mu_0)^2 = 2n\bar{X}(\mu_0 - \mu_A) + n\mu_A^2 - n\mu_0^2$.
  - If $\mu_0 > \mu_A$, reject for small value of $\bar{X}$. If $\mu_0 < \mu_A$, reject for large value of $\bar{X}$

Generalized Ratio Test

- Hypothesis: composite null vs. composite alternative
- Test statistic:
  $\Lambda = \frac{\max_{\theta \in H_0} f(X|\theta)}{\max_{\theta \in H_0 \cup H_A} f(X|\theta)} \Rightarrow -2 \log \Lambda \sim \chi^2_{\dim \Omega - \dim \omega_0}$ as $n \to \infty$
- Rejection region: small value of $\Lambda(X)$ or large value of $-2 \log \Lambda$
- Example:
  - $H_0 : \mu = \mu_0$; $H_A : \mu \neq \mu_0$. σ^2 is known
  - $\Lambda(X) = \frac{\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2]}{\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2]}$
  - Reject when
    $-2 \log \Lambda > \chi_1^2(\alpha) \Rightarrow \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi_1^2(\alpha) \Rightarrow \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z(\alpha/2)$

Generalized Ratio Test

- Hypothesis: composite null vs. composite alternative
- Test statistic:
  $\Lambda = \frac{\max_{\theta \in H_0} f(X|\theta)}{\max_{\theta \in H_0 \cup H_A} f(X|\theta)} \Rightarrow -2 \log \Lambda \sim \chi^2_{\dim \Omega - \dim \omega_0}$ as $n \to \infty$
- Rejection region: small value of $\Lambda(X)$ or large value of $-2 \log \Lambda$
- Example:
  - $H_0 : \mu = \mu_0$; $H_A : \mu \neq \mu_0$. σ^2 is known
  - $\Lambda(X) = \frac{\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2]}{\exp[-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2]}$
  - Reject when
    $-2 \log \Lambda > \chi^2_1(\alpha) \Rightarrow \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi^2_1(\alpha) \Rightarrow \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z(\alpha/2)$

Generalized Ratio Test

- Hypothesis: composite null vs. composite alternative
- Test statistic:
  $\Lambda = \frac{\max_{\theta \in H_0} f(X|\theta)}{\max_{\theta \in H_0 \cup H_A} f(X|\theta)} \Rightarrow -2\log\Lambda \sim \chi^2_{\dim\Omega - \dim\omega_0}$ as $n \to \infty$
- Rejection region: small value of $\Lambda(X)$ or large value of $-2\log\Lambda$
- Example:
  - $H_0 : \mu = \mu_0$; $H_A : \mu \neq \mu_0$. σ^2 is known
  - $\Lambda(X) = \frac{\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_0)^2]}{\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2]}$
  - Reject when
    $-2\log\Lambda > \chi^2_1(\alpha) \Rightarrow \left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi^2_1(\alpha) \Rightarrow \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > z(\alpha/2)$

Generalized Ratio Test

- Hypothesis: composite null vs. composite alternative
- Test statistic: $\Lambda = \frac{\max_{\theta \in H_0} f(X|\theta)}{\max_{\theta \in H_0 \cup H_A} f(X|\theta)} \Rightarrow -2\log\Lambda \sim \chi^2_{\dim\Omega - \dim\omega_0}$ as $n \to \infty$
- Rejection region: small value of $\Lambda(X)$ or large value of $-2\log\Lambda$
- Example:
  - $H_0 : \mu = \mu_0$; $H_A : \mu \neq \mu_0$. σ^2 is known
  - $\Lambda(X) = \frac{\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu_0)^2]}{\exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2]}$
  - Reject when $-2\log\Lambda > \chi^2_1(\alpha) \Rightarrow \left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi^2_1(\alpha) \Rightarrow \left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > z(\alpha/2)$

Summary

- When we know exact distributions, we can perform inference exactly, but these are often only in particular situations
- When we have a proportion, we can use the approximation to the normal distribution to perform large-sample tests
- When we assume a parametric family, we can use the MLE within that family and be assured that it is asymptotically normally distributed as well
- The MLE is almost always what we want to use
- For testing, if we can characterize the distribution under the null hypothesis, we can calculate p-values
- More details in my notes and in 553.630 and 553.730

Summary

- When we know exact distributions, we can perform inference exactly, but these are often only in particular situations
- When we have a proportion, we can use the approximation to the normal distribution to perform large-sample tests
- When we assume a parametric family, we can use the MLE within that family and be assured that it is asymptotically normally distributed as well
- The MLE is almost always what we want to use
- For testing, if we can characterize the distribution under the null hypothesis, we can calculate p-values
- More details in my notes and in 553.630 and 553.730

Summary

- When we know exact distributions, we can perform inference exactly, but these are often only in particular situations
- When we have a proportion, we can use the approximation to the normal distribution to perform large-sample tests
- When we assume a parametric family, we can use the MLE within that family and be assured that it is asymptotically normally distributed as well
- The MLE is almost always what we want to use
- For testing, if we can characterize the distribution under the null hypothesis, we can calculate p-values
- More details in my notes and in 553.630 and 553.730

Summary

- When we know exact distributions, we can perform inference exactly, but these are often only in particular situations
- When we have a proportion, we can use the approximation to the normal distribution to perform large-sample tests
- When we assume a parametric family, we can use the MLE within that family and be assured that it is asymptotically normally distributed as well
- The MLE is almost always what we want to use
- For testing, if we can characterize the distribution under the null hypothesis, we can calculate p-values
- More details in my notes and in 553.630 and 553.730

Summary

- When we know exact distributions, we can perform inference exactly, but these are often only in particular situations
- When we have a proportion, we can use the approximation to the normal distribution to perform large-sample tests
- When we assume a parametric family, we can use the MLE within that family and be assured that it is asymptotically normally distributed as well
- The MLE is almost always what we want to use
- For testing, if we can characterize the distribution under the null hypothesis, we can calculate p-values
- More details in my notes and in 553.630 and 553.730

Summary

- When we know exact distributions, we can perform inference exactly, but these are often only in particular situations
- When we have a proportion, we can use the approximation to the normal distribution to perform large-sample tests
- When we assume a parametric family, we can use the MLE within that family and be assured that it is asymptotically normally distributed as well
- The MLE is almost always what we want to use
- For testing, if we can characterize the distribution under the null hypothesis, we can calculate p-values
- More details in my notes and in 553.630 and 553.730

Outline

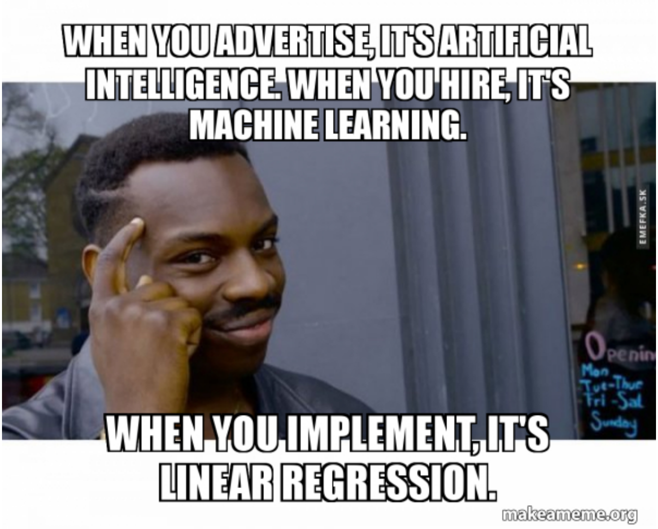

Figure: Source: https://makeameme.org/meme/when-you-advertise-f81897f53a

What is linear regression?

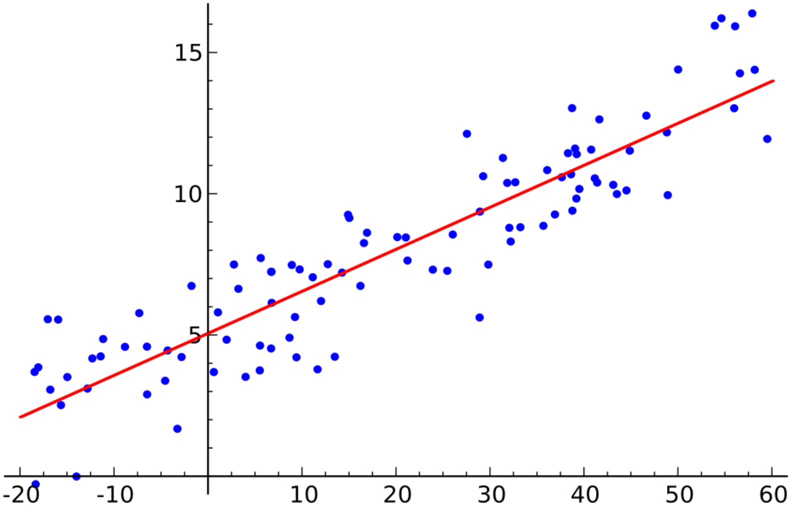

Figure: Source: https://en.wikipedia.org/wiki/Linear regression/media/File:Linear regression.svg

Model

- Write $y = \beta_1 x + \beta_0 + \varepsilon$
- Assume $\varepsilon \sim N(0, \sigma^2)$
- Want to estimate $\hat{\beta}_1$ and $\hat{\beta}_0$
- Closed form solution under this model:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - b\bar{x},$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), \qquad S_{xx} = \sum (x_i - \bar{x})^2$$

Model

- Write $y = \beta_1 x + \beta_0 + \varepsilon$
- Assume $\varepsilon \sim N(0, \sigma^2)$
- Want to estimate $\hat{\beta}_1$ and $\hat{\beta}_0$
- Closed form solution under this model:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - b\bar{x},$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), \qquad S_{xx} = \sum (x_i - \bar{x})^2$$

Model

- Write $y = \beta_1 x + \beta_0 + \varepsilon$
- Assume $\varepsilon \sim N(0, \sigma^2)$
- Want to estimate $\hat{\beta}_1$ and $\hat{\beta}_0$
- Closed form solution under this model:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - b\bar{x},$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), \qquad S_{xx} = \sum (x_i - \bar{x})^2$$

Hypothesis Testing

Can test whether $\beta_1 = 0$ since we have the exact distributions under this model:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / S_{xx}}} \sim N(0,1), \qquad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE / S_{xx}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 \ \overline{x^2} / S_{xx}}} \sim N(0,1), \qquad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE \ \overline{x^2} / S_{xx}}} \sim t_{n-2}$$

$$\frac{\hat{y} - (\beta_0 + \beta_1 x^*)}{\sqrt{MSE(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}})}} \sim t_{n-2}$$

Multivariate Setting

- In practice, we observe many more variables than the univariate setting
- We might observe: Height, Weight, frequency of physical activity, etc.
- How do we do estimation in this setting?

Multivariate Setting

- In practice, we observe many more variables than the univariate setting
- We might observe: Height, Weight, frequency of physical activity, etc.
- How do we do estimation in this setting?

Multivariate Setting

- In practice, we observe many more variables than the univariate setting
- We might observe: Height, Weight, frequency of physical activity, etc.
- How do we do estimation in this setting?

Model

- $Y = \mathbf{X}\beta + \varepsilon$ where

$$Y \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times d}, \beta \in \mathbb{R}^d$$

$$\mathbb{E}(\varepsilon) = 0, \mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 I_d \qquad \text{Gauss-Markov Assumptions}$$

- By convention, we attach a column of all ones to the matrix $\mathbf{X}$ to account for intercept term
- Want to estimate $\beta \in \mathbb{R}^d$ given the observations $\mathbf{X}$ and the response variables Y

Model

- $Y = \mathbf{X}\beta + \varepsilon$ where

$$Y \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times d}, \beta \in \mathbb{R}^d$$

$$\mathbb{E}(\varepsilon) = 0, \mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 I_d \qquad \text{Gauss-Markov Assumptions}$$

- By convention, we attach a column of all ones to the matrix $\mathbf{X}$ to account for intercept term
- Want to estimate $\beta \in \mathbb{R}^d$ given the observations $\mathbf{X}$ and the response variables Y

Model

- $Y = \mathbf{X}\beta + \varepsilon$ where

$$Y \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times d}, \beta \in \mathbb{R}^d$$

$$\mathbb{E}(\varepsilon) = 0, \mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 I_d \qquad \text{Gauss-Markov Assumptions}$$

- By convention, we attach a column of all ones to the matrix $\mathbf{X}$ to account for intercept term
- Want to estimate $\beta \in \mathbb{R}^d$ given the observations $\mathbf{X}$ and the response variables Y

Estimator

$$
\begin{aligned}
\arg\min_{\beta \in \mathbb{R}^d} \|\mathbf{X}\beta - Y\|_2 &= \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{X}\beta - Y\|_2^2 \\
&= \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\langle \mathbf{X}\beta - Y, \mathbf{X}\beta - Y\rangle \\
&= \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\langle \mathbf{X}\beta, \mathbf{X}\beta\rangle - \langle Y, \mathbf{X}\beta\rangle + \langle Y, Y\rangle \\
&= \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2}\langle \mathbf{X}\beta, \mathbf{X}\beta\rangle - \langle Y, \mathbf{X}\beta\rangle
\end{aligned}
$$

Estimator

Define $f : \mathbb{R}^d \to \mathbb{R}$ via $f(\beta) = \frac{1}{2}\langle \mathbf{X}\beta, \mathbf{X}\beta \rangle - \langle Y, \mathbf{X}\beta \rangle$.
We will take the derivative and set it equal to zero.

$$\nabla f = \frac{d}{d\beta}\frac{1}{2}\beta^\top \mathbf{X}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{Y}$$
$$\implies \nabla f = \mathbf{X}^\top \mathbf{X}\beta - \mathbf{X}^\top Y$$
$$\implies \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y$$

provided $\mathbf{X}^\top \mathbf{X}$ is invertible, which happens as long as there is no *collinearity* (i.e. no column of $\mathbf{X}$ is a linear combination of other columns)

Estimator

Define $f : \mathbb{R}^d \to \mathbb{R}$ via $f(\beta) = \frac{1}{2}\langle \mathbf{X}\beta, \mathbf{X}\beta\rangle - \langle Y, \mathbf{X}\beta\rangle$.
We will take the derivative and set it equal to zero.

$$
\begin{aligned}
\nabla f &= \frac{d}{d\beta}\frac{1}{2}\beta^\top \mathbf{X}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{Y} \\
\implies \nabla f &= \mathbf{X}^\top \mathbf{X}\beta - \mathbf{X}^\top Y \\
\implies \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top Y
\end{aligned}
$$

provided $\mathbf{X}^\top \mathbf{X}$ is invertible, which happens as long as there is no *collinearity* (i.e. no column of $\mathbf{X}$ is a linear combination of other columns)

OLS Estimator is BLUE

We have that

$$\begin{aligned}
\mathbb{E}(\hat{\beta}) &= \mathbb{E}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y \right) \\
&= \mathbb{E}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \varepsilon) \right) \\
&= \mathbb{E}\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \varepsilon \right) \\
&= \beta + \mathbb{E}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} \varepsilon) \\
&= \beta
\end{aligned}$$

so $\hat{\beta}$ is *unbiased*.

OLS Estimator is BLUE

Hence, the covariance $\mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^\top\right]$ satisfies

$$
\begin{aligned}
&\mathbb{E}\left[\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y-\beta\right)\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y-\beta\right)^\top\right]\\
&=\mathbb{E}\left[\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta+\varepsilon)-\beta\right)\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta+\varepsilon)-\beta\right)^\top\right]\\
&=\mathbb{E}\left[\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon+\beta-\beta\right)\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon+\beta-\beta\right)^\top\right]\\
&=\mathbb{E}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\varepsilon\varepsilon^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\right)\\
&=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}\left(\varepsilon\varepsilon^\top\right)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}=\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}
$$

OLS Estimator is BLUE

Let $\tilde{\beta}$ be any other linear unbiased estimator, where linear means $\tilde{\beta} = \mathbf{H}Y$ for some $\mathbf{H}$. Since $\tilde{\beta}$ is unbiased,

$$\beta = \mathbb{E}(\tilde{\beta}) = \mathbb{E}(\mathbf{H}Y) = \mathbf{H}\mathbb{E}(\mathbf{X}\beta + \varepsilon) = \mathbf{H}\mathbf{X}\beta \implies \mathbf{H}\mathbf{X} = I_d.$$

We know that $\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathbf{X} = I_d$, so write

$$\mathbf{H} := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{C},$$

for $\mathbf{C}$ satisfying $\mathbf{C}\mathbf{X} = \mathbf{0}$.

OLS Estimator is BLUE

Let $\tilde{\beta}$ be any other linear unbiased estimator, where linear means $\tilde{\beta} = \mathbf{H}Y$ for some $\mathbf{H}$. Since $\tilde{\beta}$ is unbiased,

$$\beta = \mathbb{E}(\tilde{\beta}) = \mathbb{E}(\mathbf{H}Y) = \mathbf{H}\mathbb{E}(\mathbf{X}\beta + \varepsilon) = \mathbf{H}\mathbf{X}\beta \implies \mathbf{H}\mathbf{X} = I_d.$$

We know that $\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]\mathbf{X} = I_d$, so write

$$\mathbf{H} := (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top + \mathbf{C},$$

for $\mathbf{C}$ satisfying $\mathbf{C}\mathbf{X} = \mathbf{0}$.

OLS Estimator is BLUE

Then since

$$\mathrm{Cov}(Y) = \mathrm{Cov}(\mathbf{X}\beta + \varepsilon) = \mathrm{Cov}(\varepsilon) = \sigma^2 I_d,$$

we see

$$\begin{aligned}
\mathrm{Cov}(\tilde{\beta}) &= \mathrm{Cov}(\mathbf{H}Y) = \mathbf{H}\mathrm{Cov}(Y)\mathbf{H}^\top = \sigma^2\mathbf{H}\mathbf{H}^\top \\
&= \sigma^2\left((\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top + \mathbf{C}\right)\left((\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top + \mathbf{C}\right)^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}} \\
&\quad + \mathbf{C}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}} + (\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{C}^\top + \mathbf{C}\mathbf{C}^\top \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}} + \sigma^2\mathbf{C}\mathbf{C}^\top
\end{aligned}$$

since $\mathbf{CX} = \mathbf{0}$.

OLS Estimator is BLUE

- So we have shown for any other estimator $\tilde{\beta}$ that is a linear function of Y and is unbiased that its variance is the variance of $\hat{\beta}$ plus the matrix $\sigma^2 \mathbf{C}\mathbf{C}^\top$
- In particular, $\sigma^2 \mathbf{C}\mathbf{C}^\top$ is a positive semidefinite matrix, meaning the variance of $\tilde{\beta}$ exceeds that of $\hat{\beta}$ by a positive semidefinite matrix
- This is the *Gauss-Markov Theorem*.

OLS Estimator is BLUE

- So we have shown for any other estimator $\tilde{\beta}$ that is a linear function of Y and is unbiased that its variance is the variance of $\hat{\beta}$ plus the matrix $\sigma^2 \mathbf{C}\mathbf{C}^\top$
- In particular, $\sigma^2 \mathbf{C}\mathbf{C}^\top$ is a positive semidefinite matrix, meaning the variance of $\tilde{\beta}$ exceeds that of $\hat{\beta}$ by a positive semidefinite matrix
- This is the *Gauss-Markov Theorem*.

OLS Estimator is BLUE

- So we have shown for any other estimator $\tilde{\beta}$ that is a linear function of Y and is unbiased that its variance is the variance of $\hat{\beta}$ plus the matrix $\sigma^2 \mathbf{C}\mathbf{C}^\top$
- In particular, $\sigma^2 \mathbf{C}\mathbf{C}^\top$ is a positive semidefinite matrix, meaning the variance of $\tilde{\beta}$ exceeds that of $\hat{\beta}$ by a positive semidefinite matrix
- This is the *Gauss-Markov Theorem*.

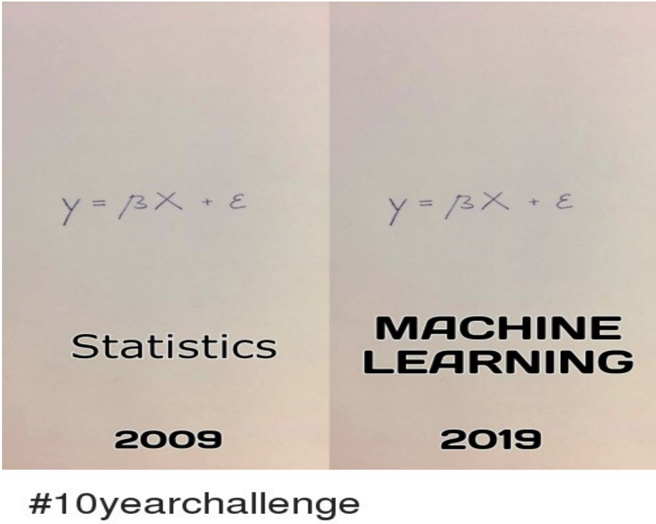

Figure: Source: https://medium.com/nybles/understanding-machine-learning-through-memes-4580b67527bf

Variable Selection Techniques

- Some regression problems have a very large number of predictors $d \geq n$, in which case classical results may not hold
- One way to eliminate this issue is to perform *variable selection*
- Classical techniques include
  - AIC
  - BIC
  - MSE
- AIC and BIC penalize for having too many variables – a variable has to help “enough"
- MSE is agnostic to model choice, but doesn’t penalize for too many variables

Variable Selection Techniques

- Some regression problems have a very large number of predictors $d \geq n$, in which case classical results may not hold
- One way to eliminate this issue is to perform *variable selection*
- Classical techniques include
  - AIC
  - BIC
  - MSE
- AIC and BIC penalize for having too many variables – a variable has to help “enough"
- MSE is agnostic to model choice, but doesn’t penalize for too many variables

Variable Selection Techniques

- Some regression problems have a very large number of predictors $d \geq n$, in which case classical results may not hold
- One way to eliminate this issue is to perform *variable selection*
- Classical techniques include
  - AIC
  - BIC
  - MSE
- AIC and BIC penalize for having too many variables – a variable has to help “enough"
- MSE is agnostic to model choice, but doesn’t penalize for too many variables

Variable Selection Techniques

- Some regression problems have a very large number of predictors $d \geq n$, in which case classical results may not hold
- One way to eliminate this issue is to perform *variable selection*
- Classical techniques include
  - AIC
  - BIC
  - MSE
- AIC and BIC penalize for having too many variables – a variable has to help “enough"
- MSE is agnostic to model choice, but doesn’t penalize for too many variables

Stepwise Regression

1. Initiate $V := \emptyset$
2. for each variable $v \notin V$:
   1. Run a model with all the variables in V and the variable v_i
   2. keep track of the AIC/BIC
3. Find the variable v^* that maximizes AIC, and set $V := V \cup v^*$.
4. go back to step 2

Penalized Regression

- Instead of minimizing the objective $\|\mathbf{X}\beta - Y\|_2^2$, one can add a *regularization term*
- Examples:
  - $\lambda\|\beta\|_1$ (Lasso)
  - $\lambda\|\beta\|_2$ (Ridge)
- Intuitively, it penalizes for higher values of β
- Common in other Machine Learning problems

Penalized Regression

- Instead of minimizing the objective $||\mathbf{X}\beta - Y||_2^2$, one can add a *regularization term*
- Examples:
  - $\lambda||\beta||_1$ (Lasso)
  - $\lambda||\beta||_2$ (Ridge)
- Intuitively, it penalizes for higher values of β
- Common in other Machine Learning problems

Penalized Regression

- Instead of minimizing the objective $\|\mathbf{X}\beta - Y\|_2^2$, one can add a *regularization term*
- Examples:
  - $\lambda\|\beta\|_1$ (Lasso)
  - $\lambda\|\beta\|_2$ (Ridge)
- Intuitively, it penalizes for higher values of β
- Common in other Machine Learning problems

Penalized Regression

- Instead of minimizing the objective $\|\mathbf{X}\beta - Y\|_2^2$, one can add a *regularization term*
- Examples:
  - $\lambda\|\beta\|_1$ (Lasso)
  - $\lambda\|\beta\|_2$ (Ridge)
- Intuitively, it penalizes for higher values of β
- Common in other Machine Learning problems

Outline

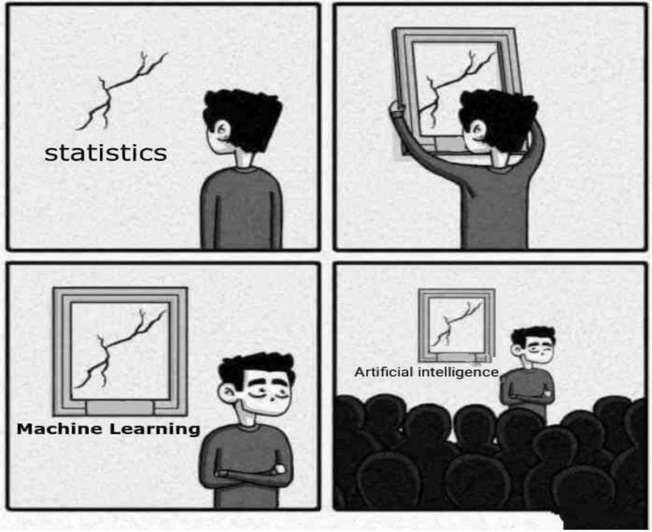

Figure: Source: https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Broadly speaking, there are two areas of machine learning
- Supervised learning:
  - Regression (continuous response)
  - Classification (categorical response variable)
- Unsupervised learning:
  - No specific response variable
  - Dimensionality Reduction
  - Clustering
  - Manifold Learning
- In either case, the resulting inference task may still be hypothesis testing, estimation, or prediction
- Textbooks often focus on estimation and prediction

Crash Course on Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!
- I am happy to discuss this more with anyone

Crash Course on Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!
- I am happy to discuss this more with anyone

Crash Course on Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!
- I am happy to discuss this more with anyone

Crash Course on Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!
- I am happy to discuss this more with anyone

Crash Course on Machine Learning

- Machine Learning can be closer to engineering or closer to statistics
- I believe machine learning should be *principled*, but many just believe it should do well on real problems
- Linear regression is principled, and neural networks work on real problems
- Even still, we do not understand everything about linear regression!
- I am happy to discuss this more with anyone

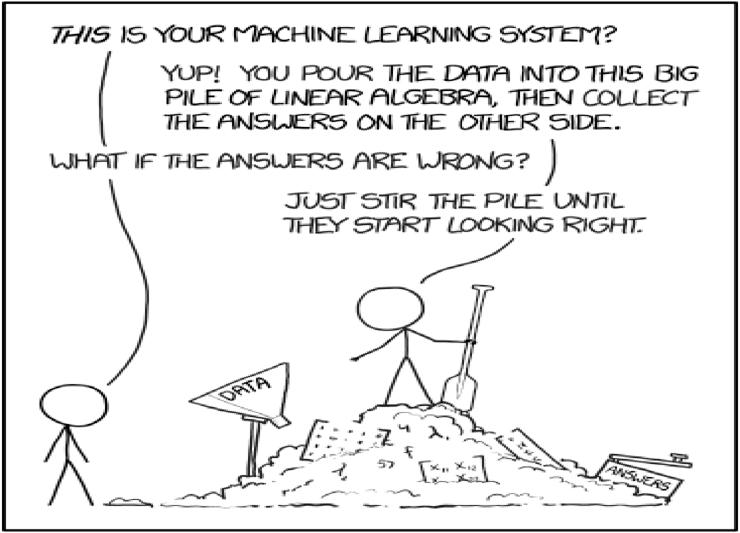

Figure: Source: https://xkcd.com/1838/

Coursework

- 553.630 (Introduction to Statistics) and 553.730 (Statistical Theory) cover parametric statistical theory
- 553.731 (Asymptotic Statistics) and 553.735 (Statistical Pattern Recognition) cover modern statistical theory for statistics, but mostly emphasize general statistical inference as opposed to studying algorithms
- 553.740 (Machine Learning I) and 553.741 (Machine Learning II) cover modern techniques and theory for machine learning including optimization

Coursework

- 553.630 (Introduction to Statistics) and 553.730 (Statistical Theory) cover parametric statistical theory
- 553.731 (Asymptotic Statistics) and 553.735 (Statistical Pattern Recognition) cover modern statistical theory for statistics, but mostly emphasize general statistical inference as opposed to studying algorithms
- 553.740 (Machine Learning I) and 553.741 (Machine Learning II) cover modern techniques and theory for machine learning including optimization

Coursework

- 553.630 (Introduction to Statistics) and 553.730 (Statistical Theory) cover parametric statistical theory
- 553.731 (Asymptotic Statistics) and 553.735 (Statistical Pattern Recognition) cover modern statistical theory for statistics, but mostly emphasize general statistical inference as opposed to studying algorithms
- 553.740 (Machine Learning I) and 553.741 (Machine Learning II) cover modern techniques and theory for machine learning including optimization

Coursework

- 553.761 (Nonlinear Optimization I) and 553.762 (Nonlinear Optimization II) cover optimization and constrained optimization (including gradient descent)
- 553.636 (Introduction to Data Science) covers practical implementation of Machine Learning Algorithms
- Other related courses: 553.792 (Matrix Analysis), 553.632 (Bayesian Statistics), 553.738 (High-dimensional Statistics) etc.

Coursework

- 553.761 (Nonlinear Optimization I) and 553.762 (Nonlinear Optimization II) cover optimization and constrained optimization (including gradient descent)
- 553.636 (Introduction to Data Science) covers practical implementation of Machine Learning Algorithms
- Other related courses: 553.792 (Matrix Analysis), 553.632 (Bayesian Statistics), 553.738 (High-dimensional Statistics) etc.

Coursework

- 553.761 (Nonlinear Optimization I) and 553.762 (Nonlinear Optimization II) cover optimization and constrained optimization (including gradient descent)
- 553.636 (Introduction to Data Science) covers practical implementation of Machine Learning Algorithms
- Other related courses: 553.792 (Matrix Analysis), 553.632 (Bayesian Statistics), 553.738 (High-dimensional Statistics) etc.

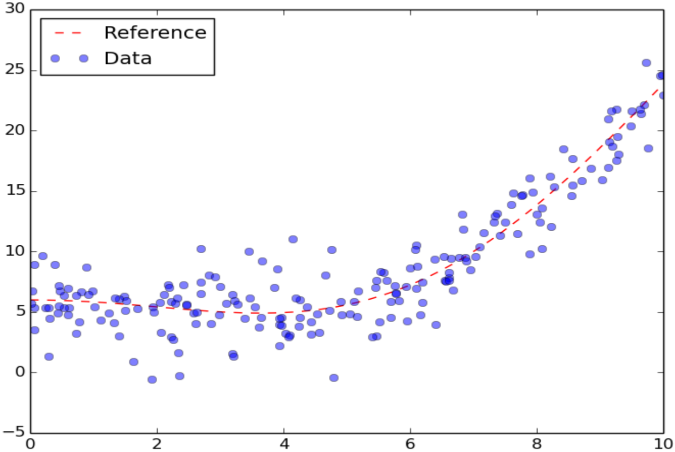

Figure: Source: https://pythonhosted.org/PyQt-Fit/NonParam_tut.html

Regression

- Idea is we have covariates **X** (as in the linear regression case), and seek to discover $Y_i = f(X_i)$ for some function f
- Sometimes f is linear ($f(X_i) = X_i^\top \beta$)
- Sometimes f is more involved (smooth, highly nonlinear, piecewise linear)
- Statistics worries about the statistical properties of an estimator of f; machine learning worries about how to actually do the estimation
- Example:
  - f is a smooth function, and we minimize some objective function to find our estimator $\hat{f}$ (nonparametric)
  - Statistics studies the *statistical* properties of $\hat{f}$
  - Machine Learning studies how to optimize the objective function for $\hat{f}$

Regression

- Idea is we have covariates $\mathbf{X}$ (as in the linear regression case), and seek to discover $Y_i = f(X_i)$ for some function f
- Sometimes f is linear ($f(X_i) = X_i^\top \beta$)
- Sometimes f is more involved (smooth, highly nonlinear, piecewise linear)
- Statistics worries about the statistical properties of an estimator of f; machine learning worries about how to actually do the estimation
- Example:
  - f is a smooth function, and we minimize some objective function to find our estimator $\hat{f}$ (nonparametric)
  - Statistics studies the *statistical* properties of $\hat{f}$
  - Machine Learning studies how to optimize the objective function for $\hat{f}$

Regression

- Idea is we have covariates **X** (as in the linear regression case), and seek to discover $Y_i = f(X_i)$ for some function f
- Sometimes f is linear ($f(X_i) = X_i^\top \beta$)
- Sometimes f is more involved (smooth, highly nonlinear, piecewise linear)
- Statistics worries about the statistical properties of an estimator of f; machine learning worries about how to actually do the estimation
- Example:
  - f is a smooth function, and we minimize some objective function to find our estimator $\hat{f}$ (nonparametric)
  - Statistics studies the *statistical* properties of $\hat{f}$
  - Machine Learning studies how to optimize the objective function for $\hat{f}$

Regression

- Idea is we have covariates **X** (as in the linear regression case), and seek to discover $Y_i = f(X_i)$ for some function f
- Sometimes f is linear ($f(X_i) = X_i^\top \beta$)
- Sometimes f is more involved (smooth, highly nonlinear, piecewise linear)
- Statistics worries about the statistical properties of an estimator of f; machine learning worries about how to actually do the estimation
- Example:
  - f is a smooth function, and we minimize some objective function to find our estimator $\hat{f}$ (nonparametric)
  - Statistics studies the *statistical* properties of $\hat{f}$
  - Machine Learning studies how to optimize the objective function for $\hat{f}$

Regression

- Idea is we have covariates **X** (as in the linear regression case), and seek to discover $Y_i = f(X_i)$ for some function f
- Sometimes f is linear ($f(X_i) = X_i^\top \beta$)
- Sometimes f is more involved (smooth, highly nonlinear, piecewise linear)
- Statistics worries about the statistical properties of an estimator of f; machine learning worries about how to actually do the estimation
- Example:
  - f is a smooth function, and we minimize some objective function to find our estimator $\hat{f}$ (nonparametric)
  - Statistics studies the *statistical* properties of $\hat{f}$
  - Machine Learning studies how to optimize the objective function for $\hat{f}$

Regression

- Idea is we have covariates **X** (as in the linear regression case), and seek to discover $Y_i = f(X_i)$ for some function f
- Sometimes f is linear ($f(X_i) = X_i^\top \beta$)
- Sometimes f is more involved (smooth, highly nonlinear, piecewise linear)
- Statistics worries about the statistical properties of an estimator of f; machine learning worries about how to actually do the estimation
- Example:
  - f is a smooth function, and we minimize some objective function to find our estimator $\hat{f}$ (nonparametric)
  - Statistics studies the *statistical* properties of $\hat{f}$
  - Machine Learning studies how to optimize the objective function for $\hat{f}$

Regression

- Typically study

$$\inf_{f \in \mathcal{F}_0} \sum_{i=1}^{n} \|f(X_i) - Y_i\|_\eta$$

  for η some norm and $\mathcal{F}$ some (computable) function class
- Often want to minimize MSE ($\eta = 2$)
- Examples of ML algorithms to find f above:
  - Random Forests
  - Neural Networks
  - Linear Regression
  - Nonparametric Regression (splines and things)
  - More involved classes of functions

Regression

- Typically study

$$\inf_{f \in \mathcal{F}_0} \sum_{i=1}^{n} \|f(X_i) - Y_i\|_\eta$$

  for η some norm and $\mathcal{F}$ some (computable) function class
- Often want to minimize MSE ($\eta = 2$)
- Examples of ML algorithms to find f above:
  - Random Forests
  - Neural Networks
  - Linear Regression
  - Nonparametric Regression (splines and things)
  - More involved classes of functions

Regression

- Typically study

$$\inf_{f \in \mathcal{F}_0} \sum_{i=1}^{n} \| f(X_i) - Y_i \|_\eta$$

  for η some norm and $\mathcal{F}$ some (computable) function class
- Often want to minimize MSE ($\eta = 2$)
- Examples of ML algorithms to find f above:
  - Random Forests
  - Neural Networks
  - Linear Regression
  - Nonparametric Regression (splines and things)
  - More involved classes of functions

Classification

- Response variable $Y_i \in \{1, ..., K\}$ for some K
- Similar to above, only the norm to minimize is different
- Examples include
  - K-NN
  - Random Forests
  - Logistic Regression
  - Neural Networks
- Similar ideas to Regression

Classification

- Response variable $Y_i \in \{1, ..., K\}$ for some K
- Similar to above, only the norm to minimize is different
- Examples include
  - K-NN
  - Random Forests
  - Logistic Regression
  - Neural Networks
- Similar ideas to Regression

Classification

- Response variable $Y_i \in \{1, ..., K\}$ for some K
- Similar to above, only the norm to minimize is different
- Examples include
  - K-NN
  - Random Forests
  - Logistic Regression
  - Neural Networks
- Similar ideas to Regression

Classification

- Response variable $Y_i \in \{1, ..., K\}$ for some K
- Similar to above, only the norm to minimize is different
- Examples include
    - K-NN
    - Random Forests
    - Logistic Regression
    - Neural Networks
- Similar ideas to Regression

Supervised Learning

- Much of Machine Learning is concerned with the details of the implementation
- For example, one may use *gradient descent* to actually solve the optimization problem
- Other optimization methods exist (second-order methods, etc.)
- One can define a loss function and do some mathematics to figure out how to solve for the optimal $\hat{f}$
- Also works in more “exotic" situations
  - Matrix Completion
  - Subspace Clustering
  - Graph Clustering
  - Tensor Methods

Supervised Learning

- Much of Machine Learning is concerned with the details of the implementation
- For example, one may use *gradient descent* to actually solve the optimization problem
- Other optimization methods exist (second-order methods, etc.)
- One can define a loss function and do some mathematics to figure out how to solve for the optimal $\hat{f}$
- Also works in more “exotic" situations
  - Matrix Completion
  - Subspace Clustering
  - Graph Clustering
  - Tensor Methods

Supervised Learning

- Much of Machine Learning is concerned with the details of the implementation
- For example, one may use *gradient descent* to actually solve the optimization problem
- Other optimization methods exist (second-order methods, etc.)
- One can define a loss function and do some mathematics to figure out how to solve for the optimal $\hat{f}$
- Also works in more “exotic" situations
  - Matrix Completion
  - Subspace Clustering
  - Graph Clustering
  - Tensor Methods

Supervised Learning

- Much of Machine Learning is concerned with the details of the implementation
- For example, one may use *gradient descent* to actually solve the optimization problem
- Other optimization methods exist (second-order methods, etc.)
- One can define a loss function and do some mathematics to figure out how to solve for the optimal $\hat{f}$
- Also works in more “exotic" situations
  - Matrix Completion
  - Subspace Clustering
  - Graph Clustering
  - Tensor Methods

Supervised Learning

- Much of Machine Learning is concerned with the details of the implementation
- For example, one may use *gradient descent* to actually solve the optimization problem
- Other optimization methods exist (second-order methods, etc.)
- One can define a loss function and do some mathematics to figure out how to solve for the optimal $\hat{f}$
- Also works in more “exotic" situations
  - Matrix Completion
  - Subspace Clustering
  - Graph Clustering
  - Tensor Methods

Supervised Learning

- Much of Machine Learning is concerned with the details of the implementation
- For example, one may use *gradient descent* to actually solve the optimization problem
- Other optimization methods exist (second-order methods, etc.)
- One can define a loss function and do some mathematics to figure out how to solve for the optimal $\hat{f}$
- Also works in more “exotic" situations
    - Matrix Completion
    - Subspace Clustering
    - Graph Clustering
    - Tensor Methods

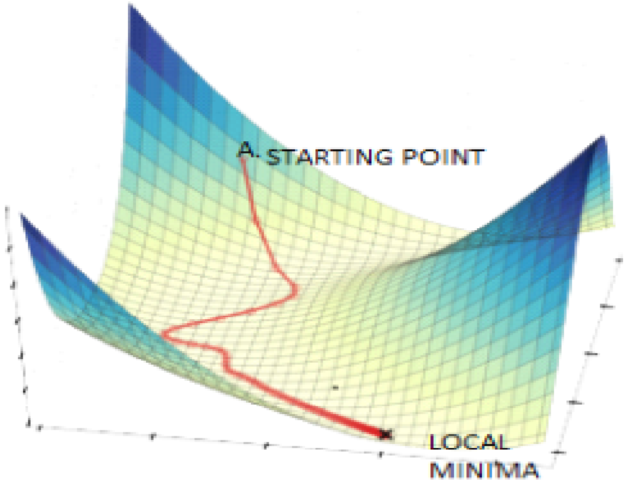

Figure: Source: https://www.datasciencecentral.com/profiles/blogs/alternatives-to-the-gradient-descent-algorithm

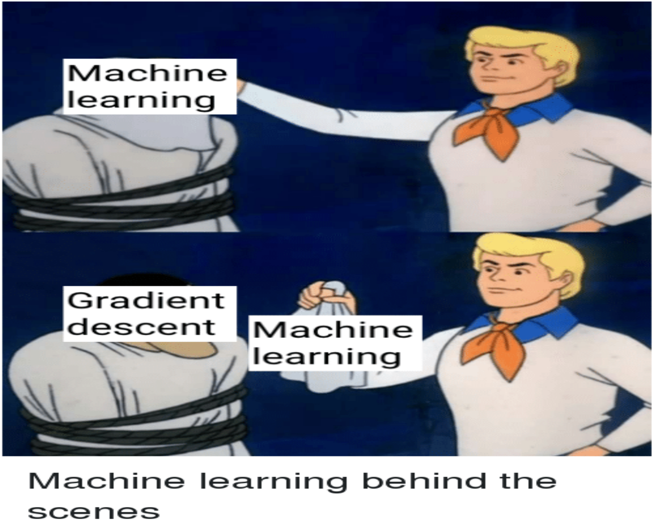

Figure: Source:
https://me.me/i/machine-learning-gradient-descent-machine-learning-machine-learning-behind-the-ea8fe9fc64054eda89232d7ffc9ba60e

Unsupervised Learning

- Want to uncover some hidden structure in the data
- Hidden structure could be:
  - Sparsity
  - Linearity
  - “Smooth" Nonlinearity
  - Clusters
- Algorithms proposed for different assumed structure in the data

Unsupervised Learning

- Want to uncover some hidden structure in the data
- Hidden structure could be:
    - Sparsity
    - Linearity
    - “Smooth" Nonlinearity
    - Clusters
- Algorithms proposed for different assumed structure in the data

Unsupervised Learning

- Want to uncover some hidden structure in the data
- Hidden structure could be:
  - Sparsity
  - Linearity
  - “Smooth" Nonlinearity
  - Clusters
- Algorithms proposed for different assumed structure in the data

Unsupervised Learning

- Want to uncover some hidden structure in the data
- Hidden structure could be:
  - Sparsity
  - Linearity
  - “Smooth" Nonlinearity
  - Clusters
- Algorithms proposed for different assumed structure in the data

Unsupervised Learning

- Want to uncover some hidden structure in the data
- Hidden structure could be:
  - Sparsity
  - Linearity
  - “Smooth" Nonlinearity
  - Clusters
- Algorithms proposed for different assumed structure in the data

Unsupervised Learning

- Want to uncover some hidden structure in the data
- Hidden structure could be:
  - Sparsity
  - Linearity
  - “Smooth" Nonlinearity
  - Clusters
- Algorithms proposed for different assumed structure in the data

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more
- Highly intuitive explanation in terms of covariances and singular value decompositions
- First component of PCA maximizes the variance along that direction

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more
- Highly intuitive explanation in terms of covariances and singular value decompositions
- First component of PCA maximizes the variance along that direction

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more
- Highly intuitive explanation in terms of covariances and singular value decompositions
- First component of PCA maximizes the variance along that direction

PCA

- Principal Components Analysis assumes data are linear combination of underlying variables
- Lots of theory exists in fixed-dimension, high-dimension, and more
- Highly intuitive explanation in terms of covariances and singular value decompositions
- First component of PCA maximizes the variance along that direction

PCA

Suppose $\mathbb{E}(X) = 0$ and $\mathbb{E}(XX^\top) = \Sigma_0 + \sigma^2 I_d$, where Σ_0 is rank $r < d$. Then

$$\mathbb{E}(XX^\top) = \underbrace{\mathbf{UDU}^\top}_{\text{top } r \text{ eigenvectors}} + \underbrace{\mathbf{U}_\perp \mathbf{D}_\perp \mathbf{U}_\perp^\top}_{\text{bottom } d-r \text{ eigenvectors}}$$

$$\frac{1}{n}\sum_{i=1}^n X_i X_i^\top = \underbrace{\hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{U}}^\top}_{\text{top } r \text{ eigenvectors}} + \underbrace{\hat{\mathbf{U}}_\perp \hat{\mathbf{D}}_\perp \hat{\mathbf{U}}_\perp^\top}_{\text{bottom } d-r \text{ eigenvectors}}$$

Idea is that when $d_r - d_{r+1}$ is sufficiently large then

$$\hat{\mathbf{U}} \approx \mathbf{U} \qquad \hat{\mathbf{D}} \approx \mathbf{D}.$$

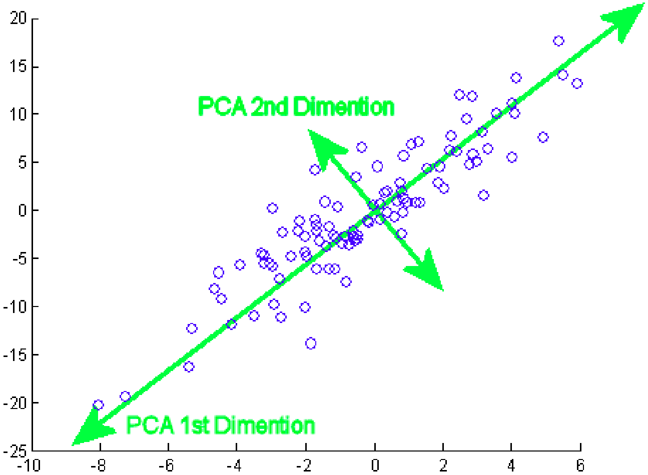

Figure: Source: https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure
- Lots of algorithms exist (see Wiki on nonlinear dimensionality reduction)

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure
- Lots of algorithms exist (see Wiki on nonlinear dimensionality reduction)

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure
- Lots of algorithms exist (see Wiki on nonlinear dimensionality reduction)

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure
- Lots of algorithms exist (see Wiki on nonlinear dimensionality reduction)

Manifold Learning

- Assume we observe $X_i \in \mathbb{R}^D$, where D is very large
- Idea is X_i are noisy observations of a manifold $\mathcal{M} \subset \mathbb{R}^D$, where $\mathcal{M}$ is of dimension $d < D$
- Example: X_i are from the unit sphere in $\mathbb{R}^D$, then $\mathcal{M}$ is of dimension $d = D - 1$
- Manifold Learning seeks to uncover this manifold structure
- Lots of algorithms exist (see Wiki on nonlinear dimensionality reduction)

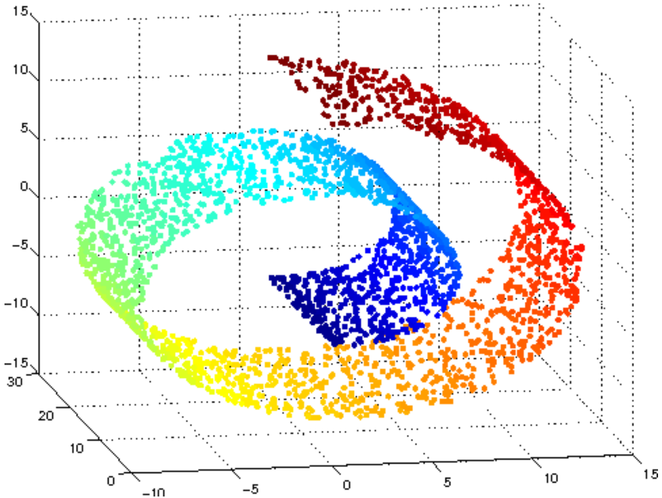

Figure: Source: https://www.semanticscholar.org/paper/Algorithms-for-manifold-learning-Cayton/100dcf6aa83ac559c83518c8a41676b1a3a55fc0/figure/0

Clustering

- Clustering assumes data come from a mixture and seeks to estimate the clusters
- Examples:
  - K-Means (uses only means)
  - Expectation Maximization Algorithm (Mixtures of Gaussians)
  - Spectral Clustering –clusters using eigenvectors of a matrix

Clustering

- Clustering assumes data come from a mixture and seeks to estimate the clusters
- Examples:
  - K-Means (uses only means)
  - Expectation Maximization Algorithm (Mixtures of Gaussians)
  - Spectral Clustering –clusters using eigenvectors of a matrix

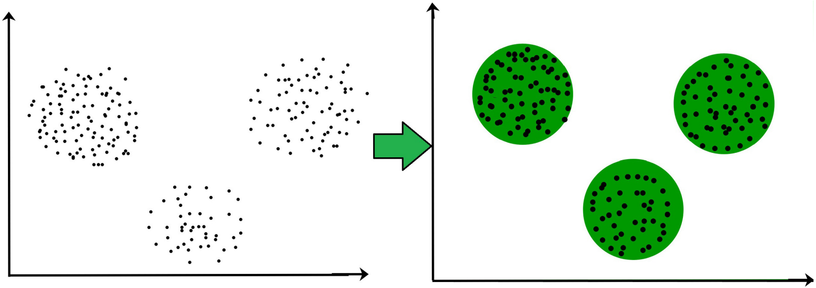

Figure: Source:
https://www.geeksforgeeks.org/clustering-in-machine-learning/

Outline

Nonparametric Statistics

- Recall we had a family of distributions $\mathcal{F}$
- Parametric required that $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$
- Nonparametric Statistics makes no such assumption
- Estimation requires estimating the function f entirely (parametric it is easier, since we just need to estimate a parameter)
- Also nonparametric regression, classification, and hypothesis testing

Nonparametric Statistics

- Recall we had a family of distributions $\mathcal{F}$
- Parametric required that $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$
- Nonparametric Statistics makes no such assumption
- Estimation requires estimating the function f entirely (parametric it is easier, since we just need to estimate a parameter)
- Also nonparametric regression, classification, and hypothesis testing

Nonparametric Statistics

- Recall we had a family of distributions $\mathcal{F}$
- Parametric required that $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$
- Nonparametric Statistics makes no such assumption
- Estimation requires estimating the function f entirely (parametric it is easier, since we just need to estimate a parameter)
- Also nonparametric regression, classification, and hypothesis testing

Nonparametric Statistics

- Recall we had a family of distributions $\mathcal{F}$
- Parametric required that $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$
- Nonparametric Statistics makes no such assumption
- Estimation requires estimating the function f entirely (parametric it is easier, since we just need to estimate a parameter)
- Also nonparametric regression, classification, and hypothesis testing

Nonparametric Statistics

- Recall we had a family of distributions $\mathcal{F}$
- Parametric required that $\mathcal{F} = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$
- Nonparametric Statistics makes no such assumption
- Estimation requires estimating the function f entirely (parametric it is easier, since we just need to estimate a parameter)
- Also nonparametric regression, classification, and hypothesis testing

High-Dimensional Issues

Let $X_1, ..., X_n$ be iid such that $\mathbb{E}X = \mu \in \mathbb{R}^d$ with covariance $\sigma^2 I_d$.

$$\mathbb{P}\left(\|\bar{X} - \mu\| > \varepsilon \right) = \mathbb{P}\left(\|\bar{X} - \mu\|^2 > \varepsilon^2 \right) \leq \frac{\mathbb{E}\left(\sum_{j=1}^d \left[\bar{X}(j) - \mu(j)\right]^2 \right)}{\varepsilon^2}$$

$$= \frac{d\mathbb{E}(\bar{X}(j) - \mu_j)^2}{\varepsilon^2} = \frac{d\sigma^2}{\varepsilon^2}$$

This shows that

$$\mathbb{P}\left(\|\bar{X} - \mu\| > \sigma\sqrt{nd} \right) \leq \frac{1}{n}.$$

When d is very small with respect to n, then this is quite useful. But if $\sigma = 1$ and $d \approx n$, then this bound is uninformative!

Let $X_1, ..., X_n$ be iid such that $\mathbb{E}X = \mu \in \mathbb{R}^d$ with covariance $\sigma^2 I_d$.

$$\mathbb{P}\left(\|\bar{X} - \mu\| > \varepsilon \right) = \mathbb{P}\left(\|\bar{X} - \mu\|^2 > \varepsilon^2 \right) \leq \frac{\mathbb{E}\left(\sum_{j=1}^d \left[\bar{X}(j) - \mu(j)\right]^2 \right)}{\varepsilon^2}$$
$$= \frac{d\mathbb{E}(\bar{X}(j) - \mu_j)^2}{\varepsilon^2} = \frac{d\sigma^2}{\varepsilon^2}$$

This shows that

$$\mathbb{P}\left(\|\bar{X} - \mu\| > \sigma\sqrt{nd} \right) \leq \frac{1}{n}.$$

When d is very small with respect to n, then this is quite useful. But if $\sigma = 1$ and $d \approx n$, then this bound is uninformative!

High-Dimensional Issues

Let $X_1, ..., X_n$ be iid such that $\mathbb{E}X = \mu \in \mathbb{R}^d$ with covariance $\sigma^2 I_d$.

$$\mathbb{P}\left(\|\bar{X} - \mu\| > \varepsilon \right) = \mathbb{P}\left(\|\bar{X} - \mu\|^2 > \varepsilon^2 \right) \leq \frac{\mathbb{E}\left(\sum_{j=1}^{d} \left[\bar{X}(j) - \mu(j) \right]^2 \right)}{\varepsilon^2}$$
$$= \frac{d\mathbb{E}(\bar{X}(j) - \mu_j)^2}{\varepsilon^2} = \frac{d\sigma^2}{\varepsilon^2}$$

This shows that

$$\mathbb{P}\left(\|\bar{X} - \mu\| > \sigma\sqrt{nd} \right) \leq \frac{1}{n}.$$

When d is very small with respect to n, then this is quite useful. But if $\sigma = 1$ and $d \approx n$, then this bound is uninformative!

High-Dimensional Statistics

- High-dimensional statistics can (loosely) be broken down into two areas:
  - Fixed dimension, fixed n results (Wainwright, 2019; Vershynin, 2018)
  - Asymptotics as $n, d \to \infty$ (Random Matrix Theory)
- The idea is we often observe data with many covariates, so either we study what happens with all the covariates or we study how to impose further structure
- Further structure includes sparsity, low-rank assumptions, manifold structure, etc.

High-Dimensional Statistics

- High-dimensional statistics can (loosely) be broken down into two areas:
  - Fixed dimension, fixed n results (Wainwright, 2019; Vershynin, 2018)
  - Asymptotics as $n, d \to \infty$ (Random Matrix Theory)
- The idea is we often observe data with many covariates, so either we study what happens with all the covariates or we study how to impose further structure
- Further structure includes sparsity, low-rank assumptions, manifold structure, etc.

High-Dimensional Statistics

- High-dimensional statistics can (loosely) be broken down into two areas:
    - Fixed dimension, fixed n results (Wainwright, 2019; Vershynin, 2018)
    - Asymptotics as $n, d \to \infty$ (Random Matrix Theory)
- The idea is we often observe data with many covariates, so either we study what happens with all the covariates or we study how to impose further structure
- Further structure includes sparsity, low-rank assumptions, manifold structure, etc.

Thank you!

I am available for questions and Zoom if you have further questions and would like to discuss statistics or anything!

References I

P.J. Bickel and K.A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Number v. 1 in Mathematical Statistics: Basic Ideas and Selected Topics. Pearson Prentice Hall, 2007. ISBN 9780132306379.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Springer New York, 1998. ISBN 9780387984735.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008. ISBN 9780387790527.

A.W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000. ISBN 9780521784504.

References II

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.