STAT 88: Lecture 33

Contents

Warm up:

population dist'n. $E(X_i)=\mu$, $Var(X_i)=\sigma^2$

(a) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, what distribution is $\bar{X}$?

(b) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$ and two samples are independent, what distribution is $\bar{X} - \bar{Y}$?

pop. dist'n. $E(X_i)=p$, $Var(X_i)=p(1-p)$

(c) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(p), (approximately) what distribution is $\bar{X}$?

(d) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(p_X) and $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim}$ Bernoulli(p_Y) and two samples are independent, (approximately) what distribution is $\bar{X} - \bar{Y}$?

(a) $E(\bar{X})=\mu$, $Var(\bar{X})=\frac{\sigma^2}{n}$.

$\bar{X}=\frac{1}{n}(X_1+\cdots+X_n) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ (Exact distribution)

(b) $\bar{X} \sim \mathcal{N}(\mu_X, \frac{\sigma_X^2}{n})$. $\bar{Y} \sim \mathcal{N}(\mu_Y, \frac{\sigma_Y^2}{m})$

indep.

$\Rightarrow \bar{X}-\bar{Y} \sim \mathcal{N}(\mu_X-\mu_Y, \frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m})$

(c) $E(\bar{X})=p$, $Var(\bar{X})=\frac{p(1-p)}{n}$.

$\bar{X} \sim \mathcal{N}(p, \frac{p(1-p)}{n})$ (by CLT)

(d) $\bar{X} \sim \mathcal{N}(p_X, \frac{p_X(1-p_X)}{n})$, $\bar{Y} \sim \mathcal{N}(p_Y, \frac{p_Y(1-p_Y)}{m})$

$\Rightarrow \bar{X}-\bar{Y} \sim \mathcal{N}(p_X-p_Y, \frac{p_X(1-p_X)}{n}+\frac{p_Y(1-p_Y)}{m})$

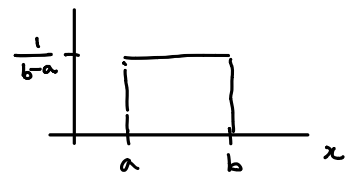
Last time

Continuous probability distributions:

Uniform

Let $X \sim \text{Unif}(a, b)$. Then the density is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$



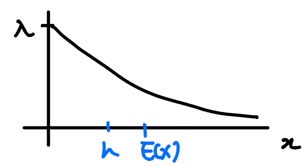
We have $E(X) = (a+b)/2$ and $\text{SD}(X) = (b-a)/\sqrt{12}$.

Exponential

Let $X \sim \text{Exp}(\lambda)$. Then the density is

E.g. X= time until failure of a mechanical device

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$



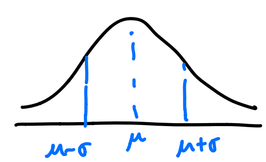
We have $E(X) = \frac{1}{\lambda}$ and $\text{SD}(X) = \frac{1}{\lambda}$. The Half Life is $h = \log 2/\lambda$.

median

Normal

Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then the density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty.$$



We have $E(X) = \mu$ and $\text{SD}(X) = \sigma$. If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and X and Y are independent, then

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).$$

This result extends to linear combinations of independent normal random variables.

10.4. The Normal Distribution

Confidence Interval for the Difference Between Means Suppose you have two independent samples as follows:

- $X_1, X_2, \ldots, X_n$ are i.i.d. with mean μ_X and SD σ_X.
- $Y_1, Y_2, \ldots, Y_m$ are i.i.d. with mean μ_Y and SD σ_Y.

You want to estimate the difference $\mu_X - \mu_Y$. Then $\bar{X} - \bar{Y}$ is an unbiased estimator for $\mu_X - \mu_Y$.

By CLT, we know

- $\bar{X}$ is approximately $\mathcal{N}(\mu_X, \frac{\sigma_X^2}{n})$.
- $\bar{Y}$ is approximately $\mathcal{N}(\mu_Y, \frac{\sigma_Y^2}{m})$.

$W \sim N(\mu_W, \sigma_W^2)$

$P(W - 2\cdot\sigma_W \le \mu_W \le W + 2\cdot\sigma_W) = 95\%$

Then $\bar{X} - \bar{Y}$ is approximately $\mathcal{N}(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$ and an approximate 95% CI for $\mu_X - \mu_Y$ is given by

$$\bar{X} - \bar{Y} \pm 2\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}.$$

$W = \bar{X} - \bar{Y}$.

$\mu_W = \mu_X - \mu_Y$

$\sigma_W = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$

Example: Suppose you have drawn samples of people independently from two cities, and suppose you have collected the following data:

- μ_X The incomes of the 400 sampled people in City X have an average of 70,000 dollars and an SD of 40,000 dollars. $n=400$, $\bar{x} = 70000$, $\sigma_X \approx 40{,}000$
- μ_Y The incomes of the 600 sampled people in City Y have an average of 80,000 dollars and an SD of 50,000 dollars. $m=600$, $\bar{Y} = 80{,}000$. $\sigma_Y = 50{,}000$

Find a 95% CI for the difference between the mean incomes in the two citie

$$(70{,}000 - 80{,}000) \pm 2\sqrt{\frac{40{,}000^2}{400} + \frac{50{,}000^2}{600}} = (-15{,}716, -4{,}824)$$

Test for the Equality of Two Means (A/B Test) We wish to determine if two independent populations have the same mean, i.e. $\mu_X - \mu_Y = 0$.

Example: Suppose you have drawn samples of people independently from two cities, and suppose you have collected the following data:

- The incomes of the 400 sampled people in City X have an average of 70,000 dollars and an SD of 40,000 dollars.
- The incomes of the 600 sampled people in City Y have an average of 80,000 dollars and an SD of 50,000 dollars.

$H_0 : \mu_X = \mu_Y$, the mean income in City X is the same as the mean income in City Y.

$H_A : \mu_Y > \mu_X$. $(\mu_Y - \mu_X > 0)$

Test statistic: $\bar{Y} - \bar{X}$. Our observed value is $\underbrace{10{,}000}_{= 80{,}000 - 70{,}000}$. We reject H_0 if $\bar{Y} - \bar{X}$ is large.

Under H_0, we have

$$\bar{Y} - \bar{X} \sim \mathcal{N}\left(\underset{=}{0}, \frac{\overset{= 40{,}000^2}{\sigma_X^2}}{400} + \frac{\overset{= 50{,}000^2}{\sigma_Y^2}}{600}\right). \qquad \sim N\left(\underbrace{\mu_Y - \mu_X}_{\overset{=}{0} \text{ under } H_0}, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

$$\sim N(0, 2858^2)$$

p-value:

$P(\bar{Y} - \bar{X} \geq 10{,}000)$

$= P\left(\frac{\bar{Y} - \bar{X} - 0}{2858} \geq \frac{10{,}000 - 0}{2858}\right)$

$= P(Z \geq 3.5)$

$= 1 - \Phi(3.5)$

Confidence Interval for the Difference Between Proportions This is a special case of the above where now populations are 0's and 1's.

- $X_1, X_2, \ldots, X_n$ are i.i.d. from Bernoulli(p_X);
- $Y_1, Y_2, \ldots, Y_m$ are i.i.d. from Bernoulli(p_Y).

95% CI: $\bar{X}-\bar{Y} \pm 2\sqrt{\frac{\bar{p}_X(1-\bar{p}_X)}{n} + \frac{\bar{p}_Y(1-\bar{p}_Y)}{m}}$

$\bar{X} - \bar{Y}$ is an unbiased estimator for $p_X - p_Y$:

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(p_X - p_Y, \frac{p_X(1-p_X)}{n} + \frac{p_Y(1-p_Y)}{m}\right).$$

by CLT

Example: Suppose we have independent samples from two cities, where sample sizes are $n = 400$ and $m = 600$ for City X and City Y, and:

$\bar{X} = 0.37$

- 37% of the City X sample are undecided about who they want as President;
- 28% of the City Y sample are undecided about who they want as President.

$\bar{Y} = 0.28$

Find a 95% CI for $p_X - p_Y$.

$\bar{X}-\bar{Y} \pm 2\sqrt{\frac{\bar{p}_X(1-\bar{p}_X)}{n} + \frac{\bar{p}_Y(1-\bar{p}_Y)}{m}}$

$\approx (0.37 - 0.28) \pm 2\sqrt{\frac{0.37(1-0.37)}{400} + \frac{0.28(1-0.28)}{600}}$

$= (0.029, 0.15)$

Test for the Equality of Two Proportions Our hypotheses are:

- $H_0 : p_X = p_Y = p$; here p is just a name we are giving to the common value of p_X and p_Y.
- $H_A : p_X > p_Y$. $P_X - P_Y > 0$

Test statistic: $\bar{X} - \bar{Y}$. Under H_0,

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \frac{p(1-p)}{n} + \frac{p(1-p)}{m}\right).$$

Example: Suppose we have independent samples from two cities, where sample sizes are $n = 400$ and $m = 600$ for City X and City Y, and:

- 37% of the City X sample are undecided about who they want as President;
- 28% of the City Y sample are undecided about who they want as President.

Test $H_0 : p_X = p_Y$ vs $H_A : p_X > p_Y$ at level 5%. $\frac{\overset{\bar{X}}{P_X}(1-\overset{\bar{X}}{P_X})}{n} + \frac{\overset{\bar{Y}}{P_Y}(1-\overset{\bar{Y}}{P_Y})}{m}$

Test statistic: $\bar{X} - \bar{Y}$.

Under H_0, $\bar{X} - \bar{Y} \sim \mathcal{N}(0, \frac{p(1-p)}{n} + \frac{p(1-p)}{m})$

What is p? City X and City Y have common p.

$\Rightarrow$ estimate p by $\hat{p} = \frac{400}{1000} \cdot \bar{X} + \frac{600}{1000} \cdot \bar{Y} = 0.316$

$(\hat{p} = \frac{n}{n+m} \cdot \bar{X} + \frac{m}{n+m} \cdot \bar{Y})$

$\Rightarrow \bar{X} - \bar{Y} \sim \mathcal{N}(0, \frac{0.316(1-0.316)}{400} + \frac{0.316(1-0.316)}{600})$

$\approx \mathcal{N}(0, 0.03^2)$

$X_1, \dots, X_n \sim Ber(P_X) \overset{\text{Under } H_0}{=} Ber(p)$

$Y_1, \dots, Y_m \sim Ber(P_Y) \underset{\text{Under } H_0}{=} Ber(p)$

$\hat{p} = \frac{X_1 + \dots + X_n + Y_1 + \dots + Y_m}{n+m}$

$= \frac{X_1 + \dots + X_n}{n+m} + \frac{Y_1 + \dots + Y_m}{n+m}$

$= \frac{n}{n+m} \cdot \bar{X} + \frac{m}{n+m} \cdot \bar{Y}$

p-val $= P(\bar{X} - \bar{Y} \geq 0.37 - 0.28)$

$= P(\bar{X} - \bar{Y} \geq 0.09)$

$= P(\frac{\bar{X} - \bar{Y} - 0}{0.03} \geq \frac{0.09 - 0}{0.03})$

$= P(Z \geq 3)$

$= 1 - \Phi(3)$

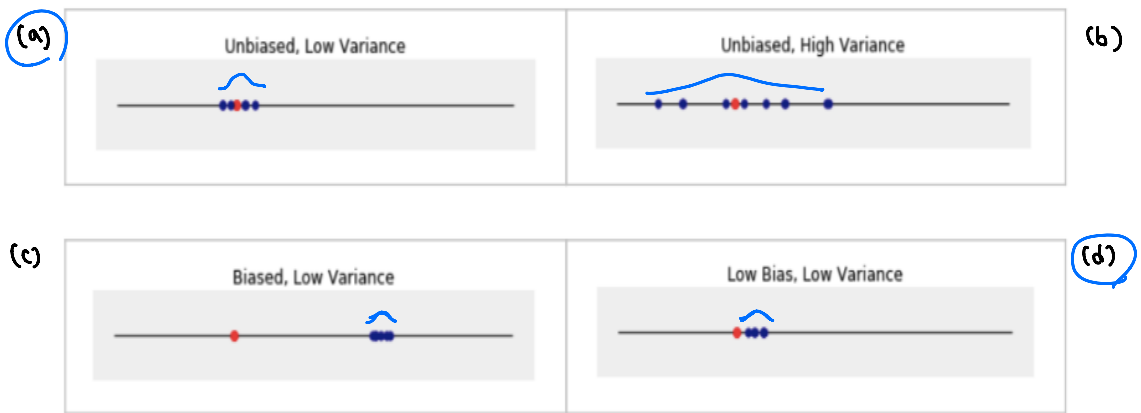
11.1. Bias and Variance

E.g. μ

Suppose we are trying to estimate a constant numerical parameter, θ, and our estimator is the statistic T. Below θ is red and T is blue for different samples.

E.g. $\bar{X}$ fixed random

What are the two best estimators?



Lets make a quantitative analysis.

E.g. $T = \bar{X}$, $\theta = \mu$

Mean Squared Error The error in our estimate is $T - \theta$. Then

$$\mathrm{MSE}_\theta(T) = E_{\underline{\theta}}\left((T - \underline{\theta})^2\right).$$

We are using θ as a subscript to remind us that the expectation is calculated under the assumption that θ is the true value of the parameter.

Think of this as the average distance squared of T from θ. We want $\mathrm{MSE}_\theta(T)$ to be as small as possible.

Decomposition of Error

Deviation:

$$D_\theta(T) = T - E_\theta(T).$$

(deviation of T from the mean)

Is it random or constant?

Bias:

$$B_\theta(T) = \boxed{E_\theta(T)} - \theta.$$

(Unbiased means $E_\theta(T) = \theta$)

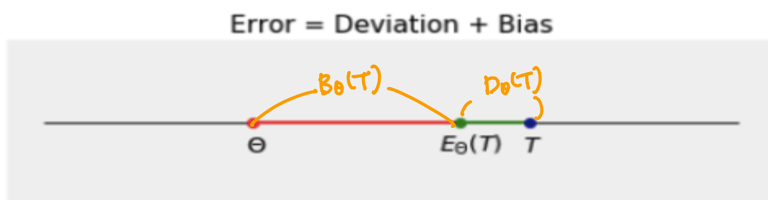
Is it random or constant?

We have a decomposition of the error as the sum of the deviation and the bias:

$$T - \theta = \underbrace{(T - E_\theta(T))}_{=D_\theta(T)} + \underbrace{(E_\theta(T) - \theta)}_{=B_\theta(T)}.$$

$E(T) = c$

$E(c) = c$



What is $E_\theta(D_\theta(T))$? What is $E_\theta(D^2_\theta(T))$?

$$= E(T - E(T)) = E(T) - \underbrace{E(E(T))}_{= E(T)} = 0$$

$$= E((T - E(T))^2) = Var(T)$$

Bias-Variance Decomposition

$$\begin{aligned}
\mathrm{MSE}_\theta(T) &= E_\theta\left((T-\theta)^2\right) \\
&= E_\theta\left((D_\theta(T) + B_\theta(T))^2\right) \\
&= E_\theta\left(D^2_\theta(T) + 2B_\theta(T)D_\theta(T) + B^2_\theta(T)\right) \\
&= E(D^2(T)) + 2E(B(T)\cdot D(T)) + E(B^2(T)) \\
&= Var(T) + 2B(T)\cdot \underbrace{E(D(T))}_{=0} + B^2(T) \\
&= Var(T) + B^2(T)
\end{aligned}$$