Stat 88 lec 33

Warm up 2:00 - 2:10

a) if $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$

what distribution is $\bar{X}$ (include parameters)

$$\bar{X} \sim N\left(p, \frac{pq}{n}\right), \quad q = 1-p$$

b) If $X_1, \ldots, X_{200} \sim \text{Bernoulli}(p_x)$ and $Y_1, \ldots, Y_{300} \overset{iid}{\sim} \text{Bernoulli}(p_y)$ } independent samples

what distribution is $\bar{X} - \bar{Y}$ (include parameters)?

$$\bar{X} - \bar{Y} \sim N\left(p_x - p_y, \frac{p_x q_x}{200} + \frac{p_y q_y}{300}\right)$$

last time

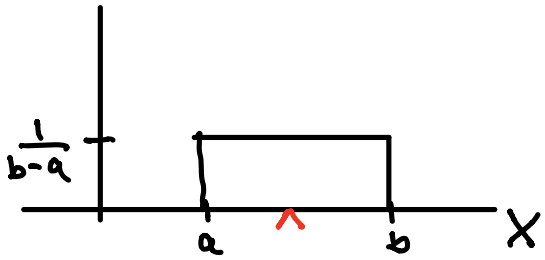
Chap 10 continuous probability distributions

uniform

$X \sim Unif(a,b)$ $\quad f(x) = \begin{cases} \frac{1}{b-a} & a<x<b \\ 0 & else \end{cases}$

$E(X) = \frac{b+a}{2}$

$SD(X) = \frac{(b-a)}{\sqrt{12}}$



Gauss model for measurement error.

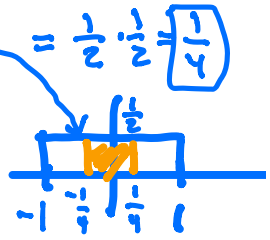
let

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \overset{iid}{\sim} Unif(-1,1)$ are n independent measurement errors.

For example $-.3, .7, .2$, etc

All numbers between -1 and 1 are "equally likely".

What is the chance the absolute value of the error is less than $1/4$? $\rightarrow P(|\varepsilon| < 1/4) =$ area $= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$



exponential

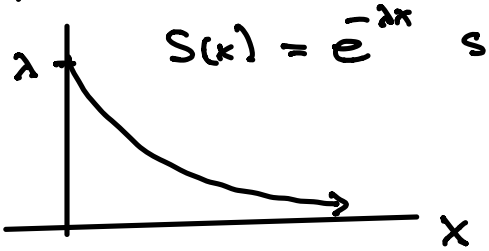
$X \sim Exp(\lambda)$ $\quad f(x) = \lambda e^{-\lambda x}$, $x>0$, $\lambda>0$

$S(x) = e^{-\lambda x}$ survival function

$E(X) = \frac{1}{\lambda}$

$SD(X) = \frac{1}{\lambda}$

$h = \frac{\log(2)}{\lambda}$ ← half life



X = time until failure of a mechanical device.

Normal

$X \sim N(\mu, \sigma^2)$

$E(X) = \mu$

$SD(X) = \sigma$

if $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$
$Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$ } indep

$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_x^2}{m}\right)$

Test $H_0: \mu_x - \mu_y = 0$

$H_1: \mu_x - \mu_y \neq 0$

See whether $0 \in \bar{x} - \bar{y} \pm 2\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$

or p-val $|\bar{x} - \bar{y}|$ is $< .05$

Today

1. Sec 10.4 CI for the difference between proportions and Test for the equality of two proportions
2. extra practice problems
3. Sec 11.1 Bias and Variance.

① Sec 10.4 CI for the difference between proportions

This is a special case of before where now populations are 0 and 1s.

$$\left.\begin{aligned} X_1,\dots,X_n &\overset{iid}{\sim} \text{Bernoulli}(p_x) \\ Y_1,\dots,Y_m &\overset{iid}{\sim} \text{Bernoulli}(p_y) \end{aligned}\right\} \text{ indep samples}$$

$\bar{X}-\bar{Y}$ is an unbiased estimator for $p_x - p_y$

$$\bar{X}-\bar{Y} \sim N\left(p_x-p_y,\ \frac{p_x q_x}{n} + \frac{p_y q_y}{m}\right)$$

Now suppose we have independent samples from two cities, sizes $n=400$, and $m=600$ for city X and City Y.

- 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

we approximate $p_x \approx .37$

$p_y \approx .28$

then a 95% CI for $p_x - p_y$ is

$$(.37-.28) \pm 2\sqrt{\frac{.37(.63)}{400} + \frac{.28(.72)}{600}} = (.029, .15)$$

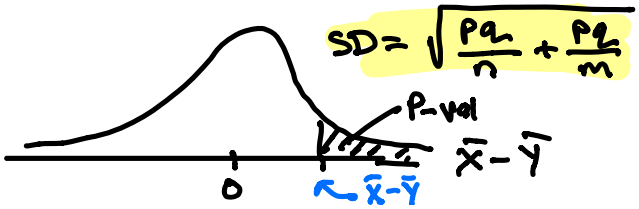
Test for equality of two proportions:

$H_0: p_x = p_y = p$

$H_1: p_x > p_y$ (or $p_x \neq p_y$ 2 sided alternative)

T.S. $\bar{X}-\bar{Y}$

$SD = \sqrt{\frac{pq}{n} + \frac{pq}{m}}$



Assuming the null, we reject null if

$$\text{P-val} = 1 - \Phi\left(\frac{(\bar{x}-\bar{y}) - 0}{\sqrt{\frac{pq}{n} + \frac{pq}{m}}}\right) < .05$$

example 10.5.7

7. A simple random sample of 200 students is taken at University A. Independently, a simple random sample of 300 students is taken at University B.

- 20% ($= \bar{A}$) of the students in Sample A are football fans, whereas 50% ($= \bar{B}$) of the students in Sample B are football fans.

Test the hypothesis

$H_0: P_A = P_B = 30\%$

$H_1: P_B > P_A$

it is just an accident that this is $\Rightarrow 50\% - 20\% = \bar{B} - \bar{A}$

T.S. $\bar{B} - \bar{A}$

$$SD(\bar{B}-\bar{A}) = \sqrt{\frac{(.3)(.7)}{200} + \frac{(.3)(.7)}{300}} = .04$$

$$\text{P-val} = 1 - \Phi\left(\frac{.3 - 0}{.04}\right) = 1 - \Phi(7.5) \approx 0 < .05$$

($\bar{B}-\bar{A}$; $SD(\bar{B}-\bar{A})$)

$\Rightarrow$ reject null.

(2) Extra problems

Ex

Suppose the time until the next earthquake in a particular place is exponentially distributed with rate 1 per year. Find the probability that the next earthquake happens within x days.

$T \sim \text{Exp}(1)$ — rate in years

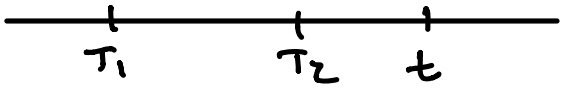
$P\left(T < \frac{x}{365}\right)$ — convert to years

$$= \int_0^{\frac{x}{365}} 1 \cdot e^{-1 \cdot t}\, dt$$

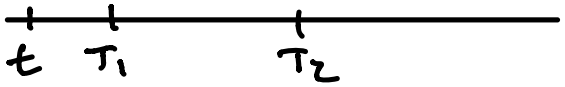
$$= -e^{-\frac{x}{365}} + e^{0}$$

$$= \boxed{1 - e^{-x/365}} \leftarrow F\left(\frac{x}{365}\right)$$

Trick to find cdf of min and max:



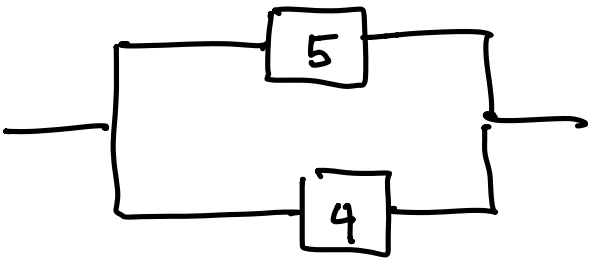
$$P(\max(T_1, T_2) < t) = P(T_1 < t, T_2 < t)$$



$$P(\min(T_1, T_2) > t) = P(T_1 > t, T_2 > t)$$

Ex

An electrical circuit consists of 2 components in the following diagram.



The lifetimes of the components, measured in days, have independent exponential distributions with means given in the diagram

Let T be the lifetime of the circuit.
Find the cdf of T.

$$\left.\begin{aligned} T_1 &\sim \text{Exp}\left(\tfrac{1}{5}\right) \\ T_2 &\sim \text{Exp}\left(\tfrac{1}{4}\right) \end{aligned}\right\} \text{ indep}$$

$T = \max(T_1, T_2)$ since circuit works until both components die.

$$\begin{aligned} F(t) = P(T < t) &= P(T_1 < t, T_2 < t) \\ &= P(T_1 < t)\,P(T_2 < t) \\ &= \boxed{\left(1 - e^{-\frac{1}{5}t}\right)\left(1 - e^{-\frac{1}{4}t}\right)} \end{aligned}$$

⑧ Sec 11.1 Bias and Variance.

Suppose we are trying to estimate a constant numerical parameter, θ, and our estimator is the statistic T.

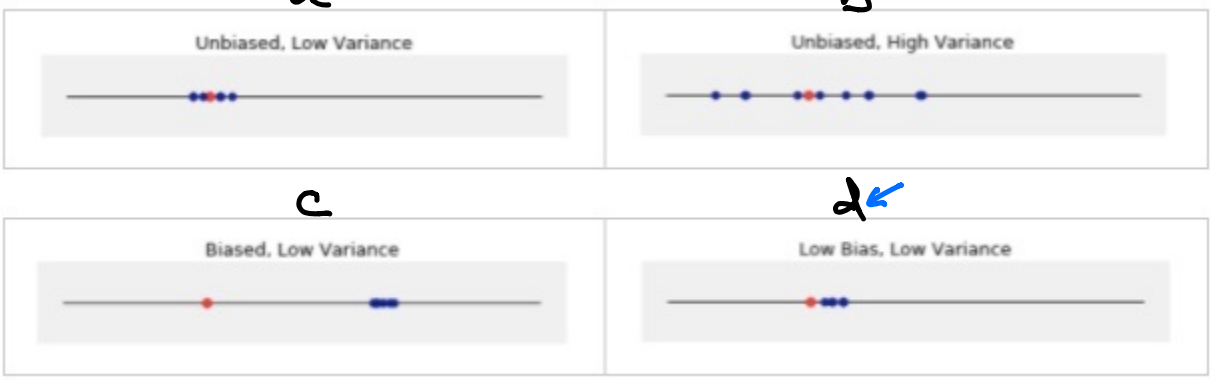
Below θ is red

T is blue for different samples

What are the two best estimators?

(a, d)

a ← ↙



Lets make a quantitative analysis.

mean square error

$$\text{mse}_\theta(T) = E_\theta\left((T-\theta)^2\right)$$

← θ is true Parameter

Think of this as the average distance squared of T from θ, we want $MSE_\theta(T)$ to be as small as possible.

Decomposition of Error

Bias:

let $B_\theta(T) = E_\theta(T) - \theta$ (unbiased means $E_\theta(T) = \theta$)

↖ bias

This is a constant !!

deviation:

let $D_\theta(T) = T - E_\theta(T)$ (deviation of T from the mean)

↖ deviation

This is a RV !!

What is $E_\theta(D_\theta(T))$?

$E(D(T)) = E(T - E(T))$

$= E(T) - E(T) = 0$

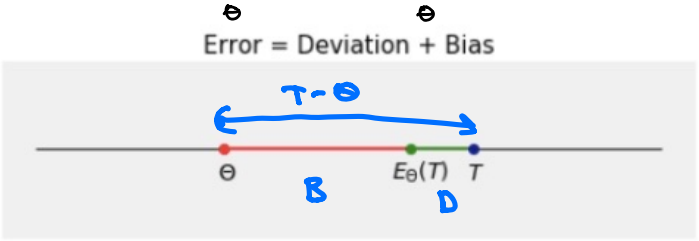
What is $E_\theta(D_\theta^2(T))$?

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

$= Var(T)$

Write $T - \theta = (T - E_\theta(T)) + (E_\theta(T) - \theta)$

$D_\theta(t)$ $\qquad$ $B_\theta(t)$



Bias-Variance Decomposition

$$MSE_\theta(T) = E_\theta\left((T-\theta)^2\right) \qquad (T-\theta)^2 = (D(T)+B(T))^2$$

$$= E_\theta\left((D_\theta(T)+B_\theta(T))^2\right)$$

$$= E_\theta\left(D_\theta^2(T) + 2D_\theta(T)B_\theta(T) + B_\theta^2(T)\right)$$

Simplify

$$= E(D^2(T)) + E(2D(T)B(T)) + E(B^2(T))$$

Constants $\qquad$ Constant

$$= E(D^2(T)) + 2B(T)E(D(T)) + B^2(T)$$

$\parallel$ $\qquad$ $\parallel$

$Var(T)$ $\qquad$ 0

$$= \boxed{Var(T) + B^2(T)}$$