Examples:

① $X_1, X_2, \ldots X_n \sim N(\mu, \sigma^2)$ iid

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ (even if n is small)

② $X_1, \ldots X_n \sim N(\mu_X, \sigma_X^2)$ iid

$Y_1, \ldots Y_m \sim N(\mu_Y, \sigma_Y^2)$ iid

$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$ ← properties of the normal & linear combinations of normal r.v.

③ $X_1, \ldots X_n \sim$ Bernoulli (p_X) iid

$Y_1, \ldots Y_m \sim$ Bernoulli (p_Y)

n, m large.

$\bar{X} \overset{\text{approx by CLT}}{\approx} N(p_X, \frac{p_X q_X}{n})$, $\bar{Y} \approx N(p_Y, \frac{p_Y q_Y}{m})$

$E(X_1) = p_X$

$Var(X_1) = p_X q_X$

Uniform

Exponential

Normal

$f(x) \geq 0$

$\int_{-\infty}^{\infty} f(x)\,dx = 1$

$F(x)$, $(0 \leq F(x) \leq 1)$

$F(x) = \int_{-\infty}^{x} f(t)\,dt$

- continuous & increasing on $(-\infty, \infty)$

$P(a < X < b) = \int_a^b f(t)\,dt$

$= F(b) - F(x)$

$E(X), Var(X), SD(X)$

Example 1

2 independent samples:
Sample 1 from Fresno, size 400
Sample 2 from Irvine size 500

We have average income from sample 1 is $\$70{,}000$ $\bar{X}$
Sample SD = \$40,000

Sample 2: mean $\bar{Y}$ = \$80,000
Sample SD = \$50,000

$Var(\bar{X})$ can be estimated as $\frac{(40{,}000)^2}{400}$

$Var(\bar{Y})$ " " " $\frac{(50{,}000)^2}{500}$

$\bar{X} - \bar{Y} = \pm \$10{,}000$

$SD(\bar{X}-\bar{Y}) \approx \sqrt{\frac{(40000)^2}{400} + \frac{(50000)^2}{500}} = \$3{,}000$

95% CI for the true difference in mean incomes b/w Irvine (Y) & Fresno (X)

$= (\bar{X}-\bar{Y}) \pm 2 \times SD(\bar{X}-\bar{Y})$

$= (-10{,}000 \pm 2 \times 3000) = -10{,}000 \pm 6000$

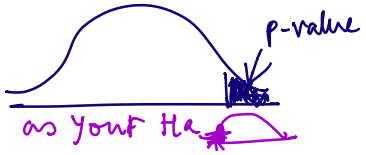
$= (\$-16{,}000, \$-4{,}000)$

Example 2. Perform a hypothesis test on the true mean difference in incomes being 0.

$H_0: \mu_X = \mu_Y$

H_A or H_1: $\mu_X < \mu_Y$ or $\mu_Y - \mu_X > 0$

If you use $\bar{X}-\bar{Y}$ & $\mu_X - \mu_Y < 0$ as your Ha



p-value

$\bar{Y} - \bar{X} \overset{\text{by CLT}}{\approx} N(0, \text{Var}(\bar{Y} - \bar{X}))$

$SD(\bar{Y} - \bar{X}) \approx 3{,}000$ (estimated above)

$$z = \frac{(\bar{Y} - \bar{X}) - E(\bar{Y} - \bar{X} \mid H_0)}{SD(\bar{Y} - \bar{X})} = \frac{10{,}000 - 0}{3000} = \frac{10}{3}$$

$$\text{P-value} = P(z > 10/3) = 1 - \Phi(\tfrac{10}{3}) \approx 0.00043$$
$$\approx 0.04\%$$

Reject the null hypothesis.

C.I. for proportions

2 indep samples from Fresno (size 400) & Irvine (size 500)

proportion in Fresno sample that want Gov. Newsom recalled = 49%

proportion in Irvine sample that want CN recalled is 48%

$\bar{X} = 0.49$ $\quad \text{Var}(\bar{X}) \approx \frac{(0.49)(0.51)}{400}$

$\bar{Y} = 0.48$ $\quad \text{Var}(\bar{Y}) \approx \frac{(0.48)(0.52)}{500}$

$$SD(\bar{X} - \bar{Y}) = \sqrt{\text{Var}(\bar{X}) + \text{Var}(\bar{Y})} \approx 0.0335$$

$\bar{X} - \bar{Y} = 0.49 - 0.48 = 0.01$

95% C.I. for true difference in means (proportions)

(by CLT) $0.01 \pm 2 \times 0.0335$

$= (-0.057, 0.077) = (-5.7\%, 7.7\%)$

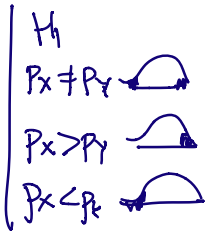
Duality b/w C.I & 2-sided hypothesis tests tells us that we would not reject a null hyp of true mean difference being 0 at 5% significance level since 0 is in this 95% C.I.

Test for equality of proportions

H_0: prop. of Fresno voters favoring recall = prop. of Irvine voters favoring recall.

$p_x = p_y = p$

H_1: $p_x > p_y$ ($p_x - p_y > 0$ )



$T = \bar{X} - \bar{Y}$

I need a value for the common value of p_x, p_y

Estimate this using total sample ($\hat{p}$ ← estimate) $\hat{p} = \frac{\text{total count}}{\text{total sample}}$

(weighted average of $\hat{p}_x$, $\hat{p}_y$)

$$\hat{p} = \frac{(0.46)(400) + (0.48)(500)}{400+500} \approx 0.4644$$

Under H_0, $\bar{X} - \bar{Y} \sim N(0, \sigma^2)$

$$\sigma = \sqrt{\frac{pq}{400} + \frac{pq}{500}} \approx 0.0335$$

$$Z = \frac{(\bar{X} - \bar{Y}) - (0)}{SD(\bar{X} - \bar{Y})} = \frac{0.01 - 0}{0.0335}$$

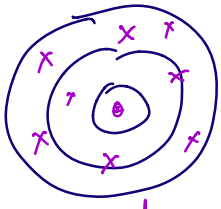
$$P\text{-value} = P\left(Z > \frac{0.01 - 0}{0.0335}\right) \approx 0.382$$

Fail to reject the Null.

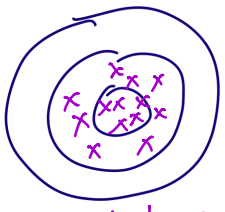
Section 11.1 Bias & Variance

Data to estimate a population parameter

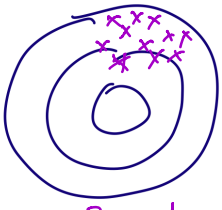
θ : true value , T : estimate



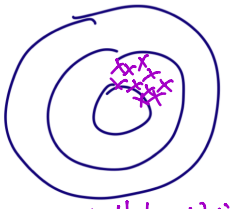
unbiased
high variance



unbiased
low variance



Biased
Low variance
precise (low variance)



low bias
low variance

unbiased $E(T) = \theta$

biased $E(T) \neq \theta$