Lecture 2: we call the underlying dist. of a $\underbrace{\text{statistic}}_{\text{order, mean, etc}}$ its Sampling dist

Goal $\begin{cases} \bar{Y} \text{ Sample mean} \to \text{estimate the unknown population mean } \mu \\ S^2 \text{ Sample variance} \to \text{estimates the unknown population variance } \sigma^2 \end{cases}$

Standardized Random Variable

Let Y have mean μ and standard deviation $\sigma = \sqrt{var(Y)}$

	heights	weights
1	180cm	80kg
2		
3		
4		
⋮		

↓ Standardize both columns

$$Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$$

$$E[Z] = E\left[\frac{Y-\mu}{\sigma}\right] = \frac{1}{\sigma} E[Y-\mu]$$

$$= \frac{1}{\sigma}\left(E[Y]-\mu\right) = 0$$

$$E[a] = a \qquad \Rightarrow E[Z] = 0$$

$$Var(Z) = var\left(\frac{Y-\mu}{\sigma}\right) = E[Z^2] - \underbrace{\left(E[Z]\right)^2}_{=0} = E\left[\left(\frac{Y-\mu}{\sigma}\right)^2\right]$$

$$= \frac{1}{\sigma^2} E\left[\left((y-\mu)^2\right)\right] = 1$$

☆ Properties of variance

$var(a)=0$, $var(by)=b^2 var(y)$, $var(a+by)=b^2 var(y)$

Normal Dist.

$Y \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{Y-\mu}{\sigma} \Rightarrow \begin{cases} E[Z]=0 \\ var(Z)=1 \end{cases}$

μ: location param.
σ^2: shape param.

$Z \sim N(0,1) \Longrightarrow U = \mu + \sigma Z = E[\mu+\sigma Z] = \mu + \sigma E[Z] = \mu$

$var(U) = var(\mu+\sigma Z) = \sigma^2 var(Z) = \sigma^2$

$var(\mu+\sigma Z) = \sigma^2 var(U) = \sigma^2$

Thrm. Any linear combination of normally-distributed random variables is itself a r.v.

$\left.\begin{array}{l} Y_1 \sim N(\mu_1, \sigma_1^2) \\ Y_2 \sim N(\mu_2, \sigma_2^2) \end{array}\right\} \xRightarrow{\text{indep.}} \underbrace{Y_1 + Y_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)}_{\text{No need to use mgf}}$

$Var\left(\frac{Y-\mu}{\sigma}\right) = \underbrace{\frac{1}{\sigma^2}}_{=\sigma^2} var(y) = 1$

$var(y) = E((y-\mu)^2)$

Sampling Dist. of Sample Mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$y_1, y_2, \ldots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$E[\bar{y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[y_i] = \frac{1}{n}(\underbrace{\mu+\mu+\mu}_{n}) = \mu$$

$$Var(\bar{y}) = var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} var(y_i)\right) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \longrightarrow$ is a good estimator? $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$

An estimator $\hat{\theta}$ is unbiased for estimating θ if $E[\hat{\theta}] = \theta$

$E[\bar{y}] = \mu \Rightarrow \bar{y}$ is an unbiased estimator for μ

$var(\bar{y}) = \frac{\sigma^2}{n}$ } less variability as n increases

$n \uparrow \longrightarrow var(\bar{y}) \downarrow \Rightarrow$ estimator becomes more precise

stdizing mean: $\frac{\bar{y} - \overbrace{E[\bar{y}]}^{=\mu}}{\sqrt{var(\bar{y})}} = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

- For normally-distributed observations, then the underlying dist of Z, is always normal exactly, both for small and large sample sizes
  w the statistic

- For all other distributions, Z follows normal dist. approximately whenever sample size is large
  (Central limit theorem)

Ex:
$n = 9$
$\sigma = 4$

Find prob. that $\underbrace{\text{sample mean}}_{\bar{Y}}$ is within 2 units from population mean.

$$P(|\bar{Y} - \mu| \leq 2) = P(-2 \leq \bar{Y} - \mu \leq 2)$$

$$= P(-1.5 \leq Z \leq 1.5)$$

$$= 0.8664$$

$$= 1 - 2P(Z \geq 1.5$$

$$= 1 - 2(0.9668)$$

Note: Sample mean ≠ population mean

$X_1, X_2, \ldots, X_m \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

indep. (what does it mean)

$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

Subtraction is an entirely different quantity

what's the Sampling dist of $\bar{X} - \bar{Y}$?

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{m}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n}\right)$$

$$E[\bar{X} - \bar{Y}] = E[\bar{X}] - E[\bar{Y}] = \mu_1 - \mu_2$$

$$Var[\bar{X} - \bar{Y}] = var(\bar{x}) + var(-\bar{y}) = var(\bar{x}) + 1\, var(\bar{y}) = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}$$

$\because var(\bar{x} + (-\bar{y}))$

$$\therefore \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

$\sigma_1^2 = 2 \quad \sigma_2^2 = 2.5, \; m = n$

Find the sample size so that $\bar{X} - \bar{Y}$ is within
1 unit from $\mu_1 - \mu_2$ with prob 0.95

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

$$P\left(\left|(\bar{x}-\bar{Y}) - (\mu_1 - \mu_2)\right| \le 1\right) = 0.95 \Rightarrow$$

$$P\left[-1 \le (\bar{X}-\bar{Y}) - (\mu_1 - \mu_2) \le 1\right] = 0.95$$

$$P\left[\frac{-1}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}} \le \frac{(\bar{X}-\bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}}\right] = 0.95 \Rightarrow$$

$$P\left[|Z| \le \sqrt{\frac{n}{4.5}}\right] = 0.95 \Rightarrow \sqrt{\frac{n}{4.5}} = 1.96 \Rightarrow n \approx 17.3 \Rightarrow n = 18$$

$$\Rightarrow 1 - P\left[\sqrt{\frac{n}{4.5}} \le |Z|\right] = 0.95$$

$$\Rightarrow P\left[\sqrt{\frac{n}{4.5}} \le |Z|\right] = 0.05$$

$1.$

After break:

Chi-Square Dist

Greek letter

$$Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2(1)$$

$$W \sim \chi^2_{(\nu)}$$ degrees of freedom

$$f(w) = \frac{w^{\frac{\nu}{2}-1} e^{-\frac{w}{2}}}{2^{\nu/2}\,\Gamma(\frac{\nu}{2})}$$

$$\chi^2_{(n)} \equiv \text{Gamma}\left(\frac{n}{2}, 2\right), \quad \alpha = \frac{n}{2}, \beta = 2$$

$$\begin{cases} E[W] = \nu \\ var(w) = 2\nu \end{cases}$$

$$\left.\begin{array}{l} W_1 \sim \chi^2(n) \\ W_2 \sim \chi^2(m) \end{array}\right\} \xRightarrow{\text{indep.}} W_1 + W_2 \sim \chi^2(n+m)$$

$$\left.\begin{array}{l} X_1 \sim \text{Gamma}(\alpha_1, \beta) \\ X_2 \sim \text{Gamma}(\alpha_2, \beta) \end{array}\right\} \xRightarrow{\text{indep}} X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$$

$$Z_1, Z_2, \dots, Z_6 \overset{iid}{\sim} N(0,1)$$

Find $\underline{\underline{b}}$ s.t. $P\left[\sum_{i=1}^{6} Z_i^2 \le b\right] = 0.95$

$$Z_i^2 \sim \chi^2(1) \Rightarrow \sum_{i=1}^{6} Z_i^2 \sim \chi^2(6)$$

$$U = \sum_{i=1}^{6} Z_i^2 \sim \chi^2(6)$$

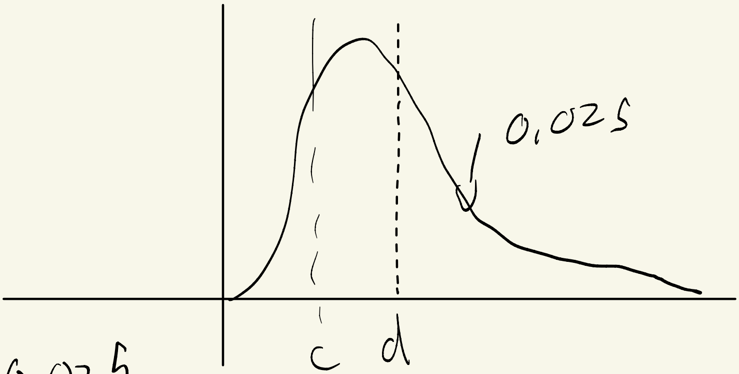
$P(U \le b) = 0.95 \Rightarrow P[U > b] = 0.05$

$\Rightarrow b = 12.5916$

Exp.

$W \sim \chi^2(5)$, Assumption: left and right tails have the same probability.

Find Constants $\underline{c}$ and $\underline{d}$ s.t. $P[c \le W \le d] = 0.95$



$P[W \le c] = 0.025$

$\Rightarrow P[W \geq c] = 0.975 \Rightarrow c = 0.8312$

$P[W \le d] = 0.975 \Rightarrow P[W \geq d] = 0.025 \Rightarrow d = 12.8325$

Finding distribution of S^2
(Finding ↓ sample)

$$y_1, y_2, \dots, y_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \begin{cases} \bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \\ \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)} \end{cases}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Consider $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \mu + \mu - \bar{y})^2$$

$$= \sum\left[(y_i - \mu) - (\bar{y} - \mu)\right]^2$$

$$= \sum\left[(y_i - \mu)^2 - 2(\bar{y} - \mu)(y_i - \mu) + (\bar{y} - \mu)^2\right]$$

$$= \sum(y_i - \mu)^2 - 2\sum(\bar{y} - \mu)(y_i - \mu) + \sum(\bar{y} - \mu)^2$$

$$= \sum(y_i - \mu)^2 - 2(\bar{y} - \mu)\sum(y_i - \mu) + n(\bar{y} - \mu)^2$$

$$= \sum (y_i-\mu)^2 - 2(\bar{y}-\mu)\sum(y_i-\mu) + n(\bar{y}-\mu)$$

$$\sum_{i=1}^{n} (y_i-\mu) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\mu = \sum_{i=1}^{n} y_i - n\mu$$

$$= n\bar{y} - n\mu$$

$$= n(\bar{y}-\mu)$$

$$= \sum (y_i-\mu)^2 - 2n(\bar{y}-\mu)^2 + n(\bar{y}-\mu)^2$$

$$= \sum (y_i-\mu)^2 - n(\bar{y}-\mu)^2$$

$$\therefore \Rightarrow \sum_{i=1}^{n} (y_i-\bar{y})^2 = \sum_{i=1}^{n} (y_i-\mu)^2 - n(\bar{y}-\mu)^2 \text{ identity}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i-\bar{y})^2 \Rightarrow \frac{(n-1)s^2}{\sigma^2} = \sum\left(\frac{y_i-\mu}{\sigma}\right)^2 - \frac{n(\bar{y}-\mu)^2}{\sigma^2}$$

$$= \sum_i (Y_i-\mu)^2 - 2\sum_i (\bar{Y}-\mu)(Y_i-\mu) + \sum_{i=1}^{n} (\bar{Y}-\mu)^2$$

$$= \sum (Y_i-\mu)^2 - 2(\bar{Y}-\mu)\sum (Y_i-\mu) + n(\bar{Y}-\mu)^2$$

$$= \sum (Y_i-\mu)^2 - 2n(\bar{Y}-\mu)^2 + n(\bar{Y}-\mu)^2$$

$$= \sum (Y_i-\mu)^2 - n(\bar{Y}-\mu)^2$$

$-2nA + nA =$
$A(-2n+n) =$
$-nA$

$$\sum_{i=1}^{n} (Y_i-\bar{Y})^2 = \sum_{i=1}^{n} (Y_i-\mu)^2 - n(\bar{Y}-\mu)^2$$ identity

$$S^2 = \frac{1}{n-1}\sum (Y_i-\bar{Y})^2 \Longrightarrow \frac{(n-1)S^2}{\sigma^2} = \sum \left(\frac{Y_i-\mu}{\sigma}\right)^2 - \frac{n(\bar{Y}-\mu)^2}{\sigma^2}$$

$$\frac{(n-1)S^2}{\sigma^2} = \left(\frac{n-1}{\sigma^2}\right) \text{[illegible]} \frac{1}{\cancel{n-1}} \sum (Y_i-\bar{Y})^2$$

Step 2. Consider $\sum \left(\frac{y_i - \mu}{\sigma}\right)^2$

$Y_i \sim N(\mu, \sigma^2) \Rightarrow \frac{Y_i - \mu}{\sigma} \sim N(0,1) \Rightarrow \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(1)$

so $\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 = \underbrace{\chi^2_{(1)} + \chi^2_{(1)} + \dots + \chi^2_{(1)}}_{n \text{ times}} = \chi^2_{(n)}$

$\Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$

Step 3. Consider $\frac{n(\bar{y}-\mu)^2}{\sigma^2}$: $\bar{y} \sim N\left(\mu, \frac{\sigma^3}{n}\right) \Longrightarrow \frac{\bar{y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \sim \chi^2(1)$

$$\therefore \quad \frac{n(\bar{y}-\mu)^2}{\sigma^2} \sim \chi^2(1)$$

Step 4: Merge the results

$$\frac{(n-1)S^2}{\sigma^2} = \sum\left(\frac{y_i-\mu}{\sigma}\right)^2 - \frac{n(\bar{y}-\mu)^2}{\sigma^2}$$

$$\left.\begin{array}{l} W_1 \sim \chi^2(n) \\ W_2 \sim \chi^2(m) \end{array}\right\}^{\text{indep}} \Longrightarrow W_1 + W_2 \sim \chi^2(n+m)$$

$$\underbrace{\frac{(n-1)s^2}{\sigma^2}}_{\chi^2(n-1)} + \underbrace{\frac{n(\bar{y}-\mu)^2}{\sigma^2}}_{\chi^2(1)} = \underbrace{\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2}_{\chi^2(n)}$$

One more solid point

Result is correct only if $\frac{(n-1)s^2}{\sigma^2}$ and $\frac{n(\bar{y}-\mu)^2}{\sigma^2}$ are indep.

Only indep. if S^2 and $\bar{Y}$ are indep.

Thrm: S^2 and $\bar{Y}$ are independent

$$\frac{(n-1)s^2}{\sigma} \sim \chi^2(n-1) \qquad \sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2 \sim \chi^2(n)$$

$$\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2(n-1)$$

losing 1 degree of freedom

when you restrict the parameter by a certain value, then you lose 1 degree of freedom

$$\underbrace{x+y}_{2}+z=2$$

degrees of freedom

$W \sim \chi^2(v) \Rightarrow E[W] = v$

$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ $\therefore$ $E[W] = n-1$

$\Rightarrow E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1$

S^2 is unbiased estimator of σ^2

$\Rightarrow \frac{\cancel{(n-1)}}{\sigma^2} E[S^2] = \cancel{n-1}$

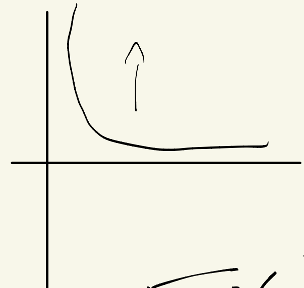
$\boxed{\Rightarrow E[S^2] = \sigma^2}$

Another estimator: $E\left[\frac{1}{n}\sum(y_i-\bar{y})^2\right] \neq \sigma^2$

$\hookrightarrow$ not unbiased estimator

Q. Find an unbiased estimator for σ

Try s to estimate σ



Jensen's inequality: $g(x)$ is convex

$\Rightarrow g(E[X]) \le E[g(x)]$

$g(x) = -\sqrt{x}$ convex $\Rightarrow -\sqrt{s^2}$ convex $\Rightarrow -\sqrt{E[s^2]} \le E[-\sqrt{s^2}]$

Strictly convex $\Rightarrow -\sqrt{\sigma^2} \le -E[s]$

$\Rightarrow E[s] \le \sigma$

$\Rightarrow E[s] < \sigma$

s underestimates σ

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$E[s] = E\left[\sqrt{s^2}\right]$$

$$= \sqrt{\frac{\sigma^2}{n-1}} \cdot E\left[\sqrt{\frac{n-1\, s^2}{\sigma^2}}\right]$$

$$= \sqrt{\frac{\sigma^2}{n-1}} \cdot E\left[\left(\frac{(n-1)s^2}{\sigma^2}\right)^{1/2}\right]$$

$$= \sqrt{\frac{\sigma^2}{n-1}}\, E\left[W^{1/2}\right]$$

☆ $W \sim \chi^2(\nu) \Rightarrow E[W^\alpha]$
$$= \frac{\Gamma\left(\frac{\nu}{2}+\alpha\right) \cdot 2^\alpha}{\Gamma\left(\frac{\nu}{2}\right)}$$

know this by heart

$$\alpha = \frac{1}{2} \Rightarrow E[s] = \sqrt{\frac{\sigma^2}{n-1}} \frac{\Gamma\left(\frac{n-1}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot 2^{1/2}$$

$$= \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot \sigma$$

$$E[s] = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \Rightarrow E[s] = C_n \cdot \sigma \Rightarrow$$

$$\frac{1}{C_n} E[s] = \sigma$$

$$\Rightarrow E\left[\frac{s}{C_n}\right] = \sigma$$

$$\Rightarrow \sigma = \sqrt{\frac{n-1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot S$$

unbiased for estimating σ