MSE 213 - L6 p1

1

RECAP: THE MEAN OF MEANS, THE VARIANCE OF THE MEANS.

$Z_1 = \bar{X}_{s'_1} = 1290\,kg$ / $Z_2 = \bar{X}_{s'_2} = 1310\,kg$

$N_{s'_1} = 50$, $N_{s'_2} = 50$ etc...

$N_{s_1} = 100$ | $N_{s_2} = 100$ | $N_{s_3} = 100$ | etc...

$Y_1 = \bar{X}_{s_1} = 1300\,kg$ | $Y_2 = \bar{X}_{s_2} = 1310\,kg$ | $Y_3 = \bar{X}_{s_3} = 1360\,kg$

$\tilde{s}_1^{\,2} = (200\,kg)^2$ | $\tilde{s}_1^{\,2} = (220\,kg)^2$ | ....

$s_1^{\,2} \approx (200\,kg)^2$ | $s_1^{\,2} \approx (220\,kg)^2$

POPULATION: "ALL CARS IN ITALY"

$N_{pop} = 40$ MILLION

TRUE MEAN OF THE WEIGHT:

$\mu = 1330\,kg$

VARIANCE: $\sigma^2 = (200\,kg)^2$

Every sample containing 100 points/cars has a mean $\bar{X}_{si}$ which we call Y_i. The collection of all these means (there would be 400'000 means) has $E(Y) = \mu$ (naturally the mean of all means gives the population mean)

and $Var(Y) = \dfrac{\sigma^2}{N_s} = \dfrac{\sigma^2}{100}$ i.e. the variance of the means is 100x less than the population variance. (see central limit theorem, CLT.)

The CLT also tells us that Y has an approximately Gaussian distribution.

What if we had chosen $N_{s'} = 50$. Let's call these means Z_i.

Now $E(Z_i) = \mu$ $\quad Var(Z_i) = \dfrac{\sigma^2}{N_{s'}} = \dfrac{\sigma^2}{50}$

What if we combine two samples, say Z_i and Z_j (which would correspond to Y_1), and average those?

$$E\left(\frac{Z_i + Z_j}{2}\right) = \tfrac{1}{2}\left(E(Z_i) + E(Z_j)\right) = \tfrac{1}{2}(\mu + \mu) = \mu = E(Y_i) \rightarrow \text{Same.}$$

$$Var\left(\frac{Z_i + Z_j}{2}\right) = \frac{1}{4} Var(Z_i + Z_j) = \frac{1}{4}\left(Var(Z_i) + Var(Z_j)\right) = \frac{1}{4}\left(\frac{\sigma^2}{50} + \frac{\sigma^2}{50}\right) = \frac{\sigma^2}{100}$$

$= Var(Y_i)$ $\rightarrow$ Same

Var is quadratic!

Z_1 and Z_2 are <u>independent</u>

MSE 213 – L 6 p2

$\boxed{2}$ We measured a sample, and got $\bar{x}_s$ and s^2. What now?

- Our best guess of μ is $\bar{x}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} x_i$
- Our best guess of σ^2 is s^2, where s^2 is the unbiased sample variance and $s^2 = \frac{N_s}{N_s - 1}\tilde{s}^2$ and $\tilde{s}^2 = \frac{1}{N_s}\sum_{i=1}^{N_s}(x_i - \bar{x}_s)^2$ is the simple sample variance because

$E(\tilde{s}^2) = \sigma^2 - \frac{\sigma^2}{N}$ i.e. it has a bias of $-\frac{\sigma^2}{N}$ so it will probably underestimate the true σ.

- Given $\bar{x}_s$ and assuming we know the true σ^2, how confident can I be in my knowledge of the true μ?

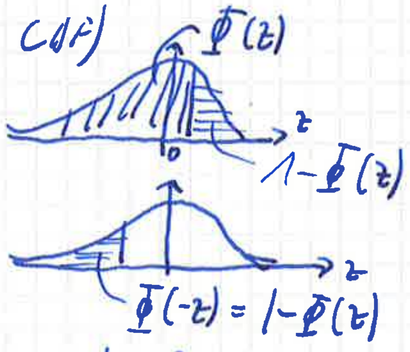
I.e. what is the probability $P(\bar{x}_s - a \leq \mu \leq \bar{x}_s + a)$ given $\bar{x}_s$ and N_s?

$\boxed{3}$

We know, from the CLT: $p(x_s) = \exp\left(-\frac{(\bar{x}_s - \mu)^2}{2(\sigma^2/N_s)}\right) \cdot \frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{N}}$

Hence: $P\left(\frac{\bar{x}_s - \mu}{(\sigma/\sqrt{N_s})} \leq z\right) = \Phi(z)$ (standard normal CDF)

Since :



by symmetry

$$P\left(-z \leq \frac{\bar{x}_s - \mu}{\sigma/\sqrt{N_s}} \leq z\right) = 2\Phi(z) - 1$$

$$= P\left(-\frac{z\sigma}{\sqrt{N_s}} \leq \bar{x}_s - \mu \leq \frac{z\sigma}{\sqrt{N_s}}\right)$$

$$= P\left(-\frac{z\sigma}{\sqrt{N_s}} - \bar{x}_s \leq -\mu \leq \frac{z\sigma}{\sqrt{N_s}} - \bar{x}_s\right)$$

divide by -1, this swaps $\leq$ to $\geq$. Then swap the left/right limit:

$$= P\left(\bar{x}_s - \frac{z\sigma}{\sqrt{N_s}} \leq \mu \leq \bar{x}_s - \frac{z\sigma}{\sqrt{N_s}}\right) = 2\Phi(z) - 1$$

So with $\bar{x}_s = 1300\,\text{kg}$, $N_s = 100$, $\sigma = 200\,\text{kg}$: "μ lies between 1270 and 1330 with a probability... $\left[1330 = \bar{x}_s + \frac{z\sigma}{\sqrt{N}} \Rightarrow z = \frac{30}{200/\sqrt{100}} = 1.5\right]$ $2\Phi(1.5) - 1 = \underline{87\%}$

MSE 213 L 6 p3

$\boxed{3\text{ bis}}$ Also we can say:

With 95% probability $\left[2\Phi(z)-1=0.95 \Rightarrow \Phi(z)=0.975 \Rightarrow z\approx 2\right.$
$\left.\Rightarrow x_s-\frac{z\sigma}{\sqrt{N_s}}=1260\right]$ μ will be between 1260 and 1340

$\Rightarrow$ "The 95% confidence interval is 1260 to 1340."

Hypothesis Testing: Your friend throws 10 times "6" with a die...

$\boxed{4}$

H_A: "The die is fair" $p_1=p_2\dots=p_6=\frac{1}{6}$

$P(10\text{ times }6\,|\,H_A)=\left(\frac{1}{6}\right)^{10}=0.00000017\%$ → seems unreasonable...

H_B: The die has $p_6=0.9$, $p_2=0.1$ others $=0$

$P(10\text{ times }6\,|\,H_B)=(0.9)^{10}=35\%$ → plausible

H_C: $P(10\times 6\,|\,H_C)=0.99^{10}=90\%$ → also plausible... ($p_6=0.99$, $p_5=0.01$)

H_D: $p_6=1$ others $=0$

$P(10\times 6\,|\,H_D)=1^{10}=1\dots$ → also plausible...?

↳ $P(\text{OUTCOME}\,|\,\text{HYPOTHESIS})\neq P(\text{HYPOTHESIS TRUE})$

Hypothesis examples:
- The mean weight of cars in the USA > in Italy → YES/NO
- Smoking makes cancer more likely than not smoking → Y/N
- There is a particle at 125 GeV

5 Hypothesis testing procedure

1) Get a YES/No Hypothesis, as a statement about a probability distribution. → "H_1"

"My friend cheats" → H_1
↳ "The die does not follow a Laplace distribution." → cannot calculate P

2) Get the logical opposite of H_1, i.e. something that covers all other scenarios except H_1.
→ "H_0", the "null hypothesis"

→ "The die does follow a Laplace distribution" → H_0
→ can calculate P

3) Choose a level of "statistical significance" α. This is the chance you will reject H_0 (and believe H_1) ~~even~~ even if it is true.

$\alpha = 5\%$ is common... should it be?

4) Compute $P(\text{outcome} \mid H_0)$ for some measurement/observation.

Is $P < \alpha$ → ~~..~~ reject H_0, believe H_1

Die example: $P(10 \times 6 \mid H_0) = 0.0000017\% < 5\%$ → reject H_0...
→ "Friend cheats"

2 throws only: $P(2 \times 6 \mid H_0) = \frac{1}{36} = 2.8\% < 5\%$ ⇝ would you reject H_0?