

- Hw 2 due Fri
- Midterm June 3
 - 1 sheet notes, handwritten
- OH:
 - My OH W after class
 - Aft's OH Tu Th 12-1p, AUPX 130

Section 4.3

Applications of the normal approximation
(and confidence intervals)

Confidence intervals

Ex) You find a new species of bats in an unexplored cave.

- **Goal:** Determine their mean weight.
- **Model:** Weight of i^{th} bat is $w_i \sim N(\mu, 1)$, where $w_1, w_2, \dots, w_n$ are independent.
 $\uparrow$
unknown mean weight
- Mean weight μ is unknown. Variance is assumed to be 1 for simplicity.

Q1) You observe w_1 what is your estimate of μ ? What can you say about the accuracy of your estimate?

Q2) You observe $w_1, w_2, \dots, w_n$. Then same as Q1.

Q1) You observe $w_1 \sim N(\mu, 1)$

Your estimate of μ is $\hat{\mu} := \underline{w_1}$

estimate of μ

Error: $\hat{\mu} - \mu = \underline{w_1 - \mu} \sim \underline{N(\mu, 1) - \mu} = \underline{N(0, 1)}$

Thus, $P(|\hat{\mu} - \mu| \leq 2) = P(\underline{N(0, 1)} \leq 2) \geq \underline{95\%}$

$$P(|\hat{\mu} - \mu| \leq 2) \geq 95\%$$

I.e., $\mu \in [\hat{\mu} - 2, \hat{\mu} + 2]$ with probability at least 95%.

Confidence interval: We call the interval $[\hat{\mu} - 2, \hat{\mu} + 2]$ a 95% *confidence interval* because it contains the true parameter with probability $\geq 95\%$.

Ex) Your weight w_1 to be 2.5 units.

Then $[2.5 - 2, 2.5 + 2] = [0.5, 4.5]$ is 95%
conf. interval.

Now you observe $w_1, w_2, \dots, w_n \sim N(\mu, 1)$.

Your estimate of μ is $\hat{\mu} := \frac{w_1 + w_2 + \dots + w_n}{n}$

Q) What is the distribution of $\hat{\mu}$?

Fact: Let $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ be independent.
Then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Thus, $w_1 + w_2 + \dots + w_n \sim N(\underbrace{\mu + \mu + \dots + \mu}_{n\mu}, \underbrace{1 + 1 + \dots + 1}_n)$,

and $\hat{\mu} \sim \frac{1}{n} N(n\mu, n) = N(\mu, \frac{1}{n}) = N(\mu, \frac{1}{n})$

$$\hat{\mu} \sim N\left(\mu, \frac{1}{n}\right)$$

So, $\hat{\mu} - \mu \sim \underline{N(\mu, 1/n) - \mu = N(0, 1/n)}$

and $\sqrt{n}(\hat{\mu} - \mu) \sim \underline{\sqrt{n} N(0, 1/n) = N(0, \sqrt{n}^2/n) = N(0, 1)}$

So $P(\overbrace{\sqrt{n}|\hat{\mu} - \mu|}^{N(0,1)} \leq 2) \leq 95\%$

Thus, $\mu \in [\hat{\mu} - \frac{2}{\sqrt{n}}, \hat{\mu} + \frac{2}{\sqrt{n}}]$ with probability at least 95%, i.e.,

$[\hat{\mu} - \frac{2}{\sqrt{n}}, \hat{\mu} + \frac{2}{\sqrt{n}}]$ is a 95% confidence interval.

Ex) After weighing 16 bats, you compute $\hat{\mu} = 7$. Give a 95% confidence interval.

95% conf. interval:

$$\left[\hat{\mu} - \frac{2}{\sqrt{16}}, \hat{\mu} + \frac{2}{\sqrt{16}} \right] = \left[7 - \frac{1}{2}, 7 + \frac{1}{2} \right] = [6.5, 7.5]$$

Challenge philosophy question: Does $\mu \in [6.5, 7.5]$

w.p. $\geq 95\%$?

Confidence intervals with normal approximation

Recall: Normal approximation

$$P(\text{Bin}(n, p) \in [a, b]) \approx P(N(np, np(1 - p)) \in [a, b])$$

In other words, if $S_n \sim \text{Bin}(n, p)$ then $S_n \approx np + \sqrt{np(1 - p)}Z$, where $Z \sim N(0, 1)$.

Rule of thumb: Approximation is accurate when $np(1 - p) \geq 10$.

Q) You have a biased coin that shows heads with unknown probability p . How to estimate p ? Assume $p \in (0,1)$.

A) Flip the coin n times, let S_n be number of heads, so $S_n \sim \text{Bin}(n, p)$.

Take the estimate $\hat{p} := \frac{S_n}{n}$ } proportion of heads

For large n we may use the normal approximation:

$$S_n \approx \frac{np + \sqrt{np(1-p)} Z}{1},$$

$$Z \sim \mathcal{N}(0,1)$$

$$\text{so } \hat{p} - p = \frac{S_n}{n} - p \approx \frac{\frac{np + \sqrt{np(1-p)} Z}{n} - p}{1} = \sqrt{\frac{p(1-p)}{n}} Z$$

Now, find a confidence interval

For any $\epsilon > 0$,

$$P(|\hat{p} - p| \leq \epsilon) \approx P\left(\frac{|\sqrt{p(1-p)} Z|}{\sqrt{n}} \leq \epsilon\right) = P(|Z| \leq \epsilon \sqrt{\frac{n}{p(1-p)}})$$

p is unknown, so we can't calculate RHS, but we can bound it.

Note: $p(1 - p) \leq \frac{1}{4}$ for any $p \in [0,1]$ so $\sqrt{p(1 - p)} \leq \frac{1}{2}$

Plug in the bound $\sqrt{p(1-p)} \leq \frac{1}{2}$ to give

$$P\left(|Z| \leq \frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right) \geq P(|Z| \leq 2\sqrt{n}\epsilon)$$

That chain of inequalities (with normal approx.) gives:

$$P(|\hat{p} - p| \leq \epsilon) \geq P(|Z| \leq 2\sqrt{n}\epsilon)$$

Now choose ϵ . For instance, taking $\epsilon = \frac{1}{\sqrt{n}}$ gives

$$(*) = P(|Z| \leq 2\sqrt{n}\epsilon) = P(|Z| \leq 2) \geq 95\%$$

I.e. $P(|\hat{p} - p| \leq \frac{1}{\sqrt{n}}) \geq 95\%$,

I.e., $[\hat{p} - \frac{1}{\sqrt{n}}, \hat{p} + \frac{1}{\sqrt{n}}]$ is a 95% conf. interval

Q) You flip a biased coin 10,000 times and get 3,000 heads. Give a 99.7% confidence interval for the probability p that the coin lands on heads.

Choose ϵ so that $(*) \geq 99.7\%$. Choose $\epsilon = \frac{3}{2\sqrt{n}}$

$$(*) = P(|Z| \leq 2\sqrt{n} \cdot \frac{3}{2\sqrt{n}}) = P(|Z| \leq 3) = 99.7\%$$

$$\begin{aligned} \text{Then } [\hat{p} - \epsilon, \hat{p} + \epsilon] &= \left[\frac{3000}{10,000} - \frac{3}{2\sqrt{10,000}}, \frac{3000}{10,000} + \frac{3}{2\sqrt{10,000}} \right] \\ &= \left[0.3 - \frac{3}{200}, 0.3 + \frac{3}{200} \right] \end{aligned}$$

is a 99.7% conf. interv.

Ex) Polling.

“Latest poll shows 65% approve of free university tuition with a margin of error of 3 percentage points.”

What does this mean? The poll only sub-sampled the relevant population. Let p be the proportion of the population that approve. Then $[0.65 - 0.03, 0.65 + 0.03]$ is a confidence interval for p (typically 95%).

Q) How many people had to be polled?