* Announcement

① HW9 due today (11:59 PM PT)

② HW10 ~ 11/9
covers Ch9

③ Quiz 8 : Ch8

④ No discussion sections on Tuesday (11/3) (Go Vote!)

STAT 88: Lecture 29

Contents

Last time

A/B testing:

A/B testing is the shorthand for comparing the distributions of two random samples.

$$\text{A} = \text{Control group}; \quad \text{B} = \text{Treatment group}.$$

It follows the same 5 steps for hypothesis testing:

(a) H_0: treatment has no effect on back pain.

(b) H_A: treatment has an effect on back pain.

(c) Test statistic X: # patient in the treatment group who had pain relief.

Under H_0, any difference between treatment and control groups is due to the random assignment of elements to treatment and control, so X follows $\text{HG}(N, G, n)$ where N=total number of patients; G=total number of patients who had pain relief; n=number of patients in the treatment group.

(d) Find p-value.

(e) Reject H_0 iff p-value $\leq 5\%$.

HW10 Q5.

Type-I error: (From warm up in Lecture 28) The *type I error* is the probability of rejecting the null hypothesis H_0 given that it is true.

$\sigma=20$

A population distribution has an SD of 20. You want to test if the population mean is equal to 50:

$$H_0 : \mu = 50 \text{ vs } H_A : \mu < 50.$$

The average of a sample of 64 observations is $\bar{x}$.

$n=64$ " obs-value.

(a) Write down the expression for p-value.

(b) Is p-value random or fixed? Why?

(c) Suppose you reject H_0 if p-value is less than or equal to 5%. Find the region of $\bar{x}$ where you reject H_0.

(d) Find the type-I error at 5% level, i.e. the probability of rejecting the null hypothesis H_0 given that it is true.

$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{64}} = 2.5$

(a) Test statistic $\bar{X}$. Under H_0, $\bar{X} \sim N(50, 2.5^2)$ (by CLT)

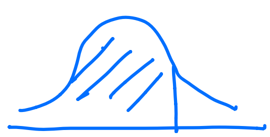
Obs-value $= \bar{x}$

p-value $= P(\bar{X} \le \bar{x})$

$= P\left(\underbrace{\frac{\bar{X}-50}{2.5}}_{= Z} \le \frac{\bar{x}-50}{2.5}\right)$

$= P\left(Z \le \frac{\bar{x}-50}{2.5}\right)$

$= \Phi\left(\frac{\bar{x}-50}{2.5}\right)$



(b) p-value is random b/c $\bar{x}$ changes across different samples

(c) Reject H_0 $\Leftrightarrow$ p-value ≤ 0.05

$\Leftrightarrow \Phi\left(\frac{\bar{x}-50}{2.5}\right) \le 0.05$

$\Leftrightarrow \bar{x} \le 2.5 \cdot \Phi^{-1}(0.05) + 50$

(d) When H_0 is true, your future observation $\bar{x}$ has $N(50, 2.5^2)$

Type-I error $= P(\text{Reject } H_0 \text{ given } H_0 \text{ true})$

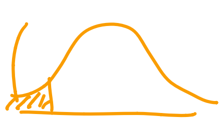
$= P(\bar{X} \le 2.5 \cdot \Phi^{-1}(0.05) + 50)$

$= P\left(\frac{\bar{X}-50}{2.5} \le \Phi^{-1}(0.05)\right)$

$= \Phi(\Phi^{-1}(0.05)) = 0.05.$

standard normal curve

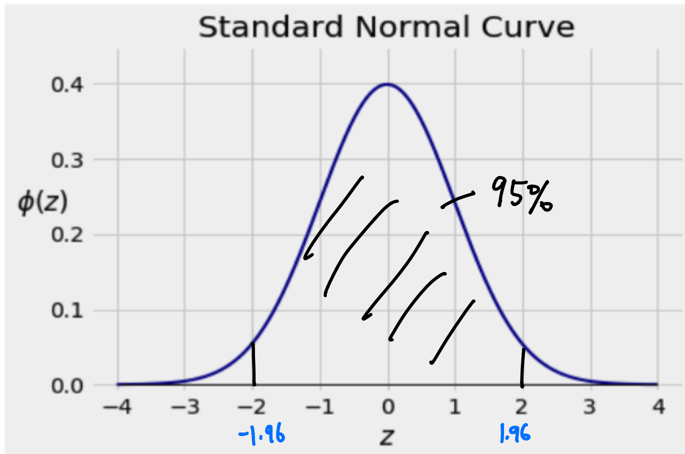
$\Phi(z) = \Phi(\Phi^{-1}(0.05))$



$\rightsquigarrow$ When H_0 is true, you reject H_0 w/ 5% chance due to randomness of your data.

9.3. Confidence Intervals: Method

Preliminary The standard normal curve:



Z has $N(0,1)$, $P(-2<Z<2) = 95\%$

Confidence interval A confidence interval is an interval of estimates of a *fixed* but unknown parameter, based on data in a random sample.

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and SD σ. We know $\bar{X}$ is an unbised estimator of μ (i.e $E(\bar{X}) = \mu$), and SD$(\bar{X}) = \sigma/\sqrt{n}$ is a measure of the average spread of $\bar{X}$.

If n is large, the Central Limit Theorem tells us that the distribution of $\bar{X}$ is roughly normal, so $\sim N(\mu, \frac{\sigma^2}{n})$

$$P\left(-2 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2\right) \approx 0.95.$$

$\rightarrow Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Multiply -1 $\rightarrow$ $-2 < \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} < 2$

$\rightarrow \bar{X} - 2\cdot\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2\cdot\frac{\sigma}{\sqrt{n}}$

We rewrite this equation as follows:

$$P\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

$$\Longleftrightarrow P\left(\mu \in \left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)\right) \approx 0.95.$$

What is random and what is fixed?

$(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}})$ random

μ fixed

The *random* interval

$$\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)$$

(CLT)

is called an approximate 95% confidence interval for μ. It is a random interval because its endpoints depend on the sample mean $\bar{X}$ which is a random variable whose value varies across samples.

Interpretation: the chance that this *random interval* contains the *fixed parameter* is about 95%.

n=64

Example: (From warm up in Lecture 28) A population distribution is known to have $\sigma=20$ an SD of 20. The average of a sample of 64 observations is 55. What is your 95% confidence interval for the population mean?

obs. value = 55

95% CI: $\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)$

obs. value of $\bar{X} = 55$

$\sigma = 20$

$n = 64$

$\rightsquigarrow \left(55 \pm 2 \cdot \frac{20}{\sqrt{64}}\right) = (50, 60)$

$= \text{with replacement}$

<u>Example:</u> (Proportion of undecided voters) In a simple random sample of 400 voters in a state, 23% are undecided about which way they will vote. Find a 95% CI for the proportion of undecided voters in the state.

$X_1, \dots, X_{400} \sim \text{Bernoulli}(p)$ $\begin{cases} 1 & \text{if undecided} \\ 0 & \text{o.w.} \end{cases}$

$\sigma = \sqrt{p(1-p)}$ $\Rightarrow$ $\frac{\sigma}{\sqrt{n}} = SD(\bar{X})$

$X_1, \dots, X_n \sim \text{Pop. Dist'n.}$

$E(X_i) = \mu$, $SD(X_i) = \sigma$

$\Downarrow$

By CLT, $\bar{X} \sim N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$

$P(\mu \in (\bar{X} - 2\cdot\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\cdot\frac{\sigma}{\sqrt{n}})) = 95\%$

95% CI : $(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}})$

Obs-value of $\bar{X} = 0.23$

$n = 400$

$\sigma = \sqrt{p(1-p)} \underset{\text{Approximation}}{\approx} \sqrt{0.23(1-0.23)} = 0.44$

$\rightsquigarrow (0.23 \pm 2 \cdot \frac{0.44}{\sqrt{400}}) = (0.186, 0.274)$

Confidence Level

In above problem, find 99.7% confidence interval.

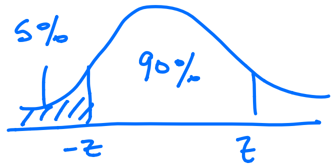
99.7% CI : $(\bar{X} \pm 3 \cdot \frac{\sigma}{\sqrt{n}})$

$= (0.23 \pm 3 \cdot \frac{0.44}{\sqrt{400}})$

$= (0.164, 0.296)$

$\Rightarrow$ Notice greater certainty requires wider interval.

To find 90% confidence interval,



$z = \Phi^{-1}(0.95)$

$= \text{stats.normal.ppf}(0.95)$

$= 1.64$

So 90% CI is

$$\left(\bar{X} - 1.64\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.64\frac{\sigma}{\sqrt{n}}\right).$$

9.4. Confidence Intervals: Interpretation

95% CI for μ:

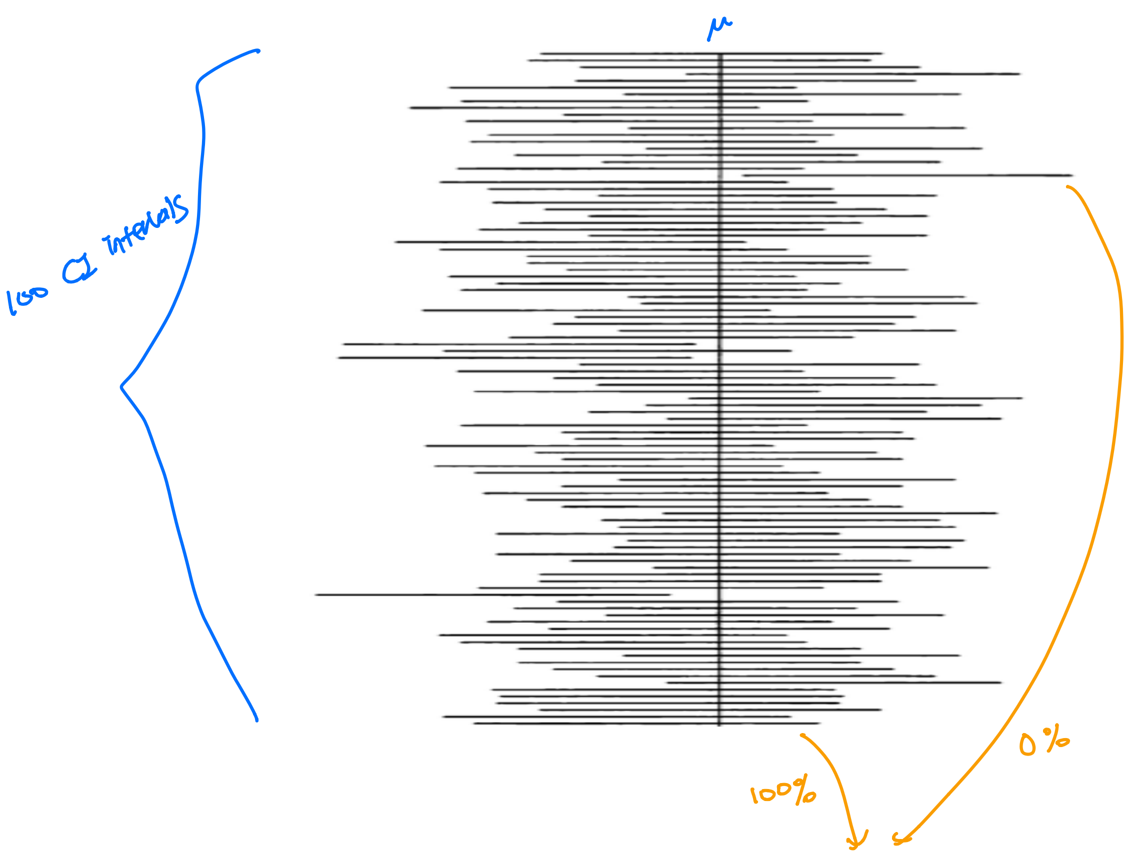
$$\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right).$$

It satisfies the property

$$P\left(\mu \in \left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)\right) \approx 0.95.$$

The probability statement above is interpreted in terms of long run frequencies:

If you repeat the sampling process 100 times, and construct a 95% confidence interval each time, then about 95 of the 100 intervals will contain the parameter μ.

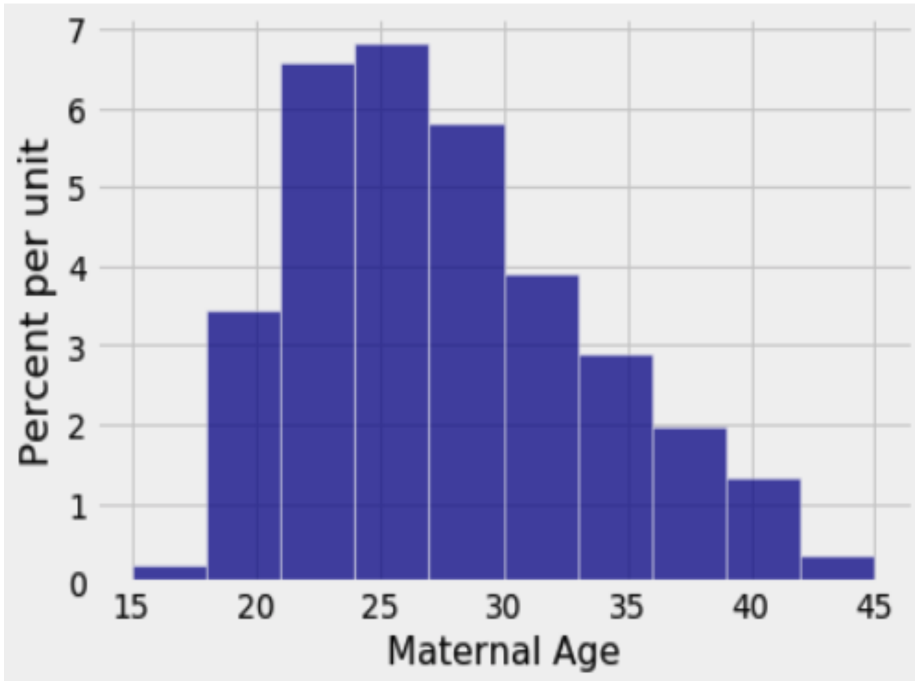


Ex: Suppose your observed instance of 95% CI is $(79, 82)$. What is the chance that $\mu \in (79, 82)$?

Either 0% or 100%. μ is a fixed number, and $(79,82)$ fixed interval, so it either contains μ or not

Comparison with the Bootstrap The interpretation of CI is the same as in Data 8.

Example: Here is a distribution of 1174 maternal ages (years) from a random sample.



The sample mean is about 27.23 years and the sample SD is about 5.8 years. Find the approximate 95% CI of μ and interpret.

This works because $\bar{X}$ is normally distributed by CLT. But if n is too small $\bar{X}$ may not be normal and we have to bootstrap your 95% CI. How do you do this?

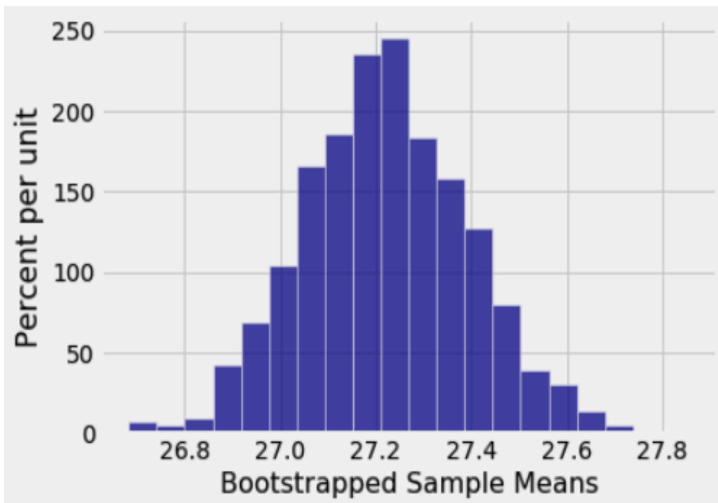
```python
def one_resampled_mean():
    return np.average(births.sample().column('Maternal Age'))
```

We then called this function repeatedly to create an array of 2,000 bootstrap means:

```python
means = make_array()

for i in np.arange(2000):
    means = np.append(means, one_resampled_mean())

Table().with_column('Bootstrapped Sample Means', means).hist(0, bins=2
```



Finally, we found the "middle 95%" of the bootstrapped means. That was our empirical bootstrap 95% confidence interval for the population mean.

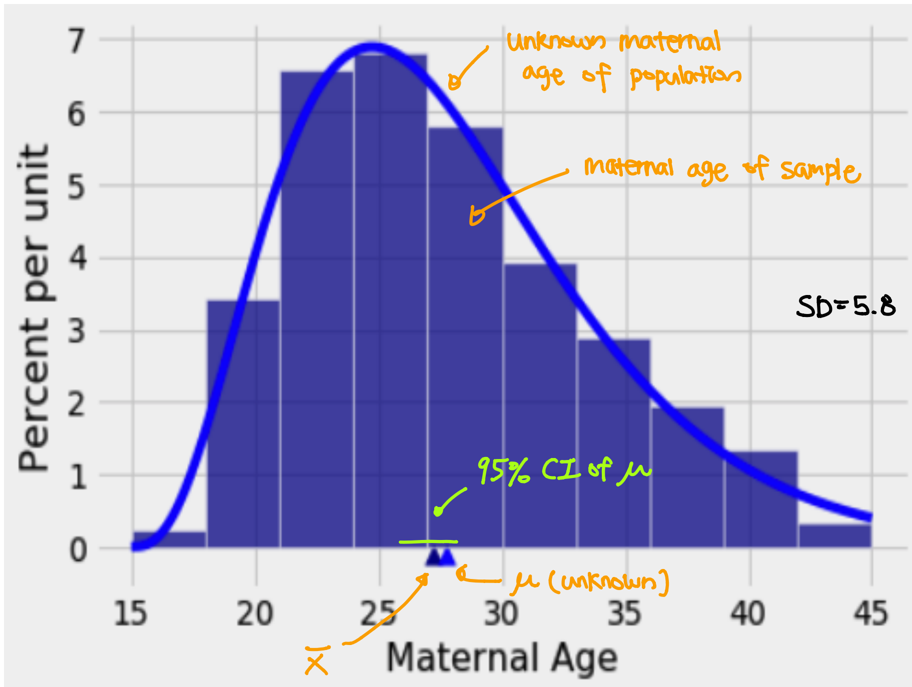
```python
left = percentile(2.5, means)
right = percentile(97.5, means)
left, right
```

```
(26.89182282793867, 27.572402044293014)
```

close to (26.89, 27.57)

What the Confidence Interval Measures

CI is an interval of estimates of μ:



$\bar{X}$ is close to μ. On average it is $\mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ away from μ. Is there a 95% chance that maternal ages are between $(26.89, 27.57)$?