STA220H1: The Practice of Statistics I

Elena Tuzhilina

February 28, 2023

Please turn on your videos :)

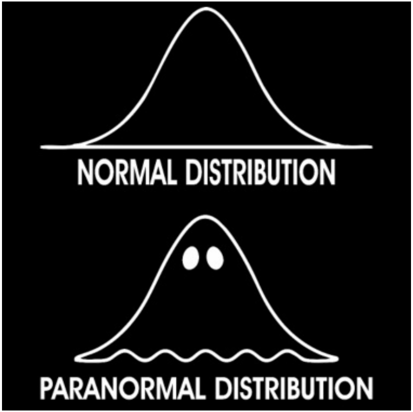


Figure 1: [picture source]

Announcements

1. Submit your regrade requests on Crowdmark by Thursday.
2. Midterm 2 is in two weeks! Same logistics (the review session will be held online this time).

Agenda for today

- Recap: normal distribution, sample mean distribution and CLT
- More about CLT
- Confidence intervals

Recap: expectation and variance

Expectation

- If X is a random variable and a is a number then

const RV

$$E(a \cdot X) = a \cdot E(X)$$

- If Y is also a random variable then

$$E(X + Y) = E(X) + E(Y)$$

Variance

- If X is a random variable and a is a number then

$$Var(a \cdot X) = a^2 \cdot Var(X)$$

- If Y is also a random variable and it is independent of X then

$$Var(X + Y) = Var(X) + Var(Y)$$

Recap: expectation and variance

RV

If $X_1, \ldots, X_n$ are independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ and $\bar{X} = \frac{X_1+\ldots+X_n}{n}$ is the average of these random variables then

RV

$$E(\bar{X}) = \mu \text{ and } Var(\bar{X}) = \frac{\sigma^2}{n}$$

?

$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

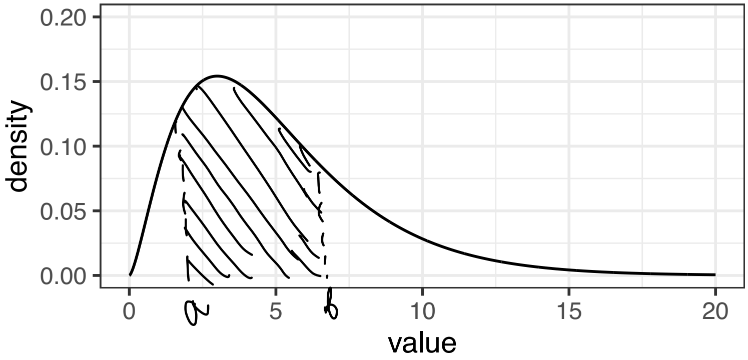
$E(X_i) = p \qquad Var(X_i) = p \cdot (1-p)$

$E(\bar{X}) = p \qquad Var(\bar{X}) = \frac{p(1-p)}{n}$

Recap: density curves

We use **density curves** to describe the distribution of continuous random variables:

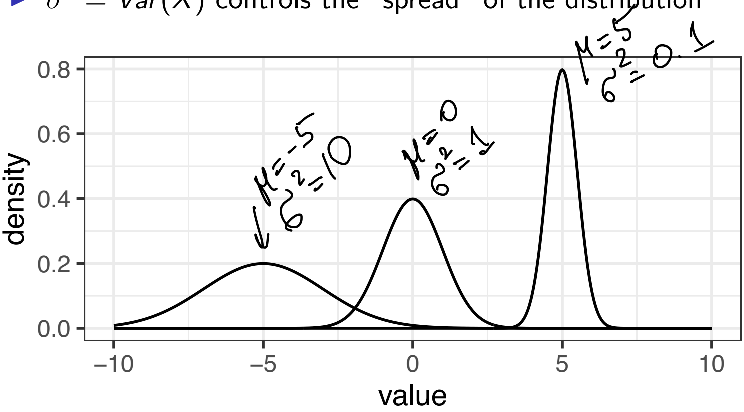
- The total area under the density curve is always 1
- The area under the curve bounded by a and b vertical lines is equal to $P(a \leq X \leq b)$



Recap: normal distribution

Normal random variable $X \sim \mathit{Normal}(\mu, \sigma^2)$ has symmetric, bell-shaped and unimodal distribution.

- $\mu = E(X)$ controls the "center" of the distribution
- $\sigma^2 = Var(X)$ controls the "spread" of the distribution

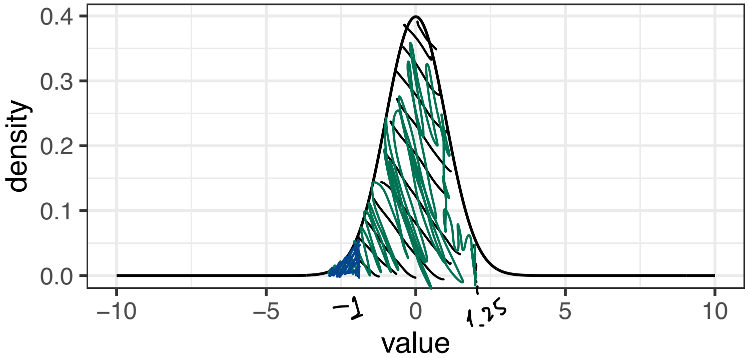


Recap: normal distribution

Standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$.

- To find the probabilities $P(a \leq X \leq b)$ for standard normal we use the distribution table ← $P(X \leq \ldots)$

$P(-1 \leq X \leq 1.25) = P(X \leq 1.25) - P(X \leq -1)$



Recap: normal distribution

- If $X \sim \mathit{Normal}(\mu, \sigma^2)$ we use **standardization**. The transformed variable $Y = \frac{X-\mu}{\sigma}$ has standard normal distribution.

$\mu \quad \sigma^2 \Rightarrow \sigma = 10$

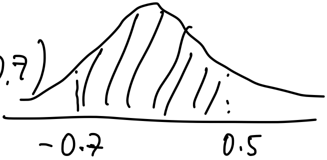
$$\frac{X-1}{10}$$

For example, if $X \sim \mathit{Normal}(1, 100)$

$$P(-6 \le X \le 6) = P\left(\frac{-6-1}{10} \le \frac{X-1}{10} \le \frac{6-1}{10}\right) =$$

$$= P(-0.7 \le Y \le 0.5) =$$

$$= P(Y \le 0.5) - P(Y \le -0.7)$$



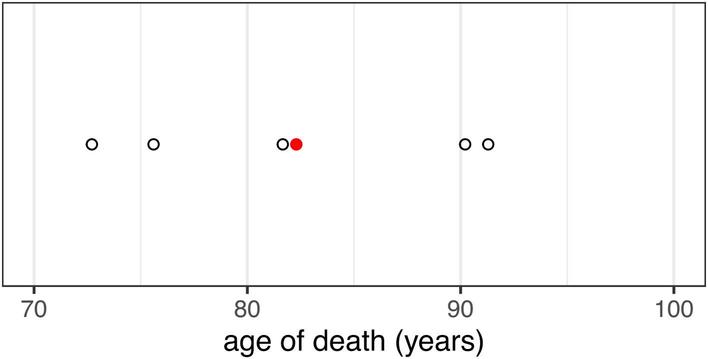
Recap: sample mean ditributoin

We want to study the **population parameter** μ, e.g. the average life expectancy in Canada.

We take a **sample** of n people and compute the average age of death for them.

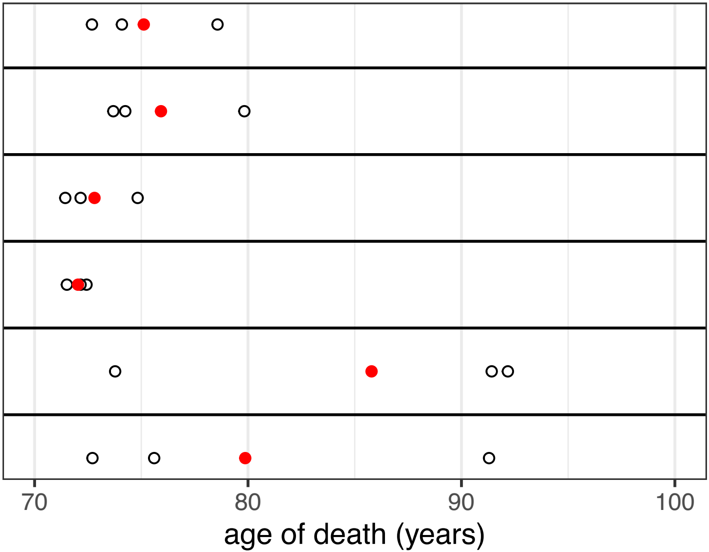
Black dots: sample $x_1, \ldots, x_n$

Red dot: sample mean $\bar{x} = \frac{x_1+\ldots+x_n}{n}$ $\approx \mu$



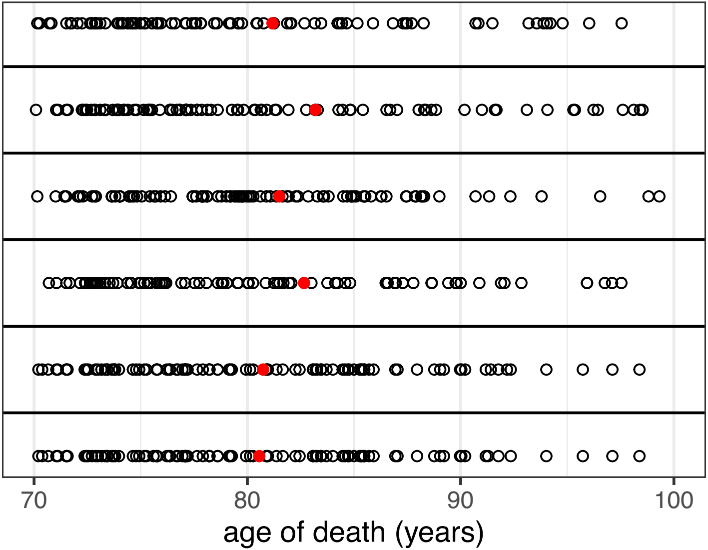
Recap: sample mean ditributoin

If your sample size is small (e.g. $n = 3$), then $\bar{x}$ can significantly vary from sample to sample.



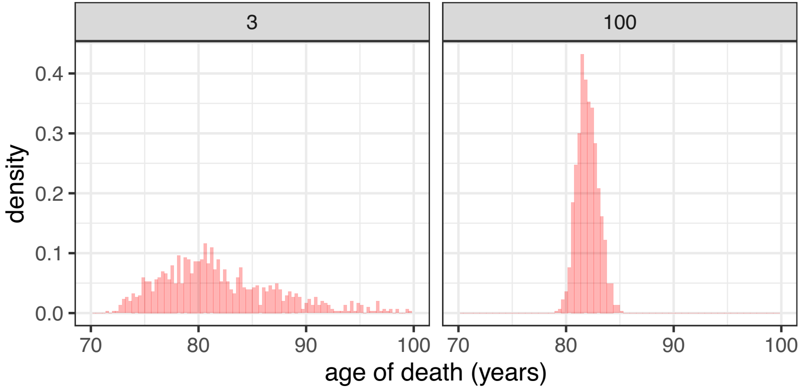
Recap: sample mean ditributoin

If your sample size is large (e.g. $n = 100$), then the variation in $\bar{x}$ is less considerable.



Recap: sample mean ditributoin

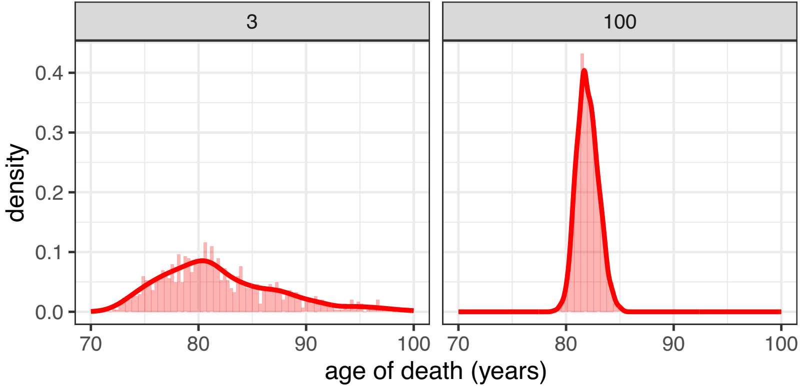
Can we characterize the behavior of $\bar{x}$?



Recap: alternative view

- We have n random variables $X_1, \ldots, X_n \longrightarrow x_1 \ldots x_n$
- We assume that they have the same distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$
- We consider $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$, it is a **random variable**
- Each sample mean $\bar{x}$ is a realization of $\bar{X}$

What is the probability density of $\bar{X}$?



Recap: central limit theorem

Central limit theorem: for n large enough

$$\bar{X} \text{ approximately} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

$\rightarrow 0$

$E(\bar{X}) = \mu$

$var(\bar{X}) = \frac{\sigma^2}{n}$

$Sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

- For different samples $\bar{x}$ will "jump around" μ
- The larger the sample size the closer $\bar{x}$ to μ

Central limit theorem: more examples

CLT applies to almost all types of probability distributions.

Example: the probability to win a lottery ticket is p. Suppose we buy n tickets and compute the **proportion of winning tickets**.

#winning / n

- $X_1, \ldots, X_n \sim Bernoulli(p)$ is the outcome for each ticket
- $Y = X_1 + \ldots + X_n \sim Binomial(n, p)$ is the total number of winning tickets
- $\bar{X} = \frac{Y}{n}$ is the proportion of winning tickets

What is the distribution for the proportion of winning tickets?

Exercise

$X_1 \ldots X_n \sim \text{Bernoulli}(p)$

$\bar{X}$

$E(y) = np$

$Var(y) = n \cdot p(1-p)$

The probability to win in a lottery p. Suppose we buy n tickets.

$E(\bar{X}) = p$

$Var(\bar{X}) = \frac{p(1-p)}{n}$

$X_1 \ldots X_n \sim \text{Bern}(p)$

$E(X_i) = p$

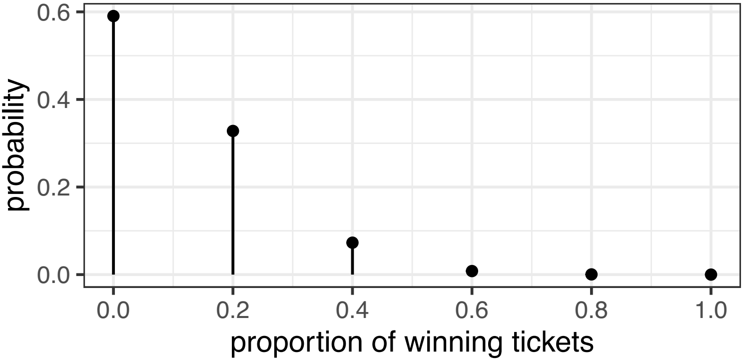
$Var(X_i) = p(1-p)$

$E(\bar{X}) = p \quad Var(\bar{X}) = \frac{p(1-p)}{n}$

Central limit theorem: more examples

$= 0, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, 1$

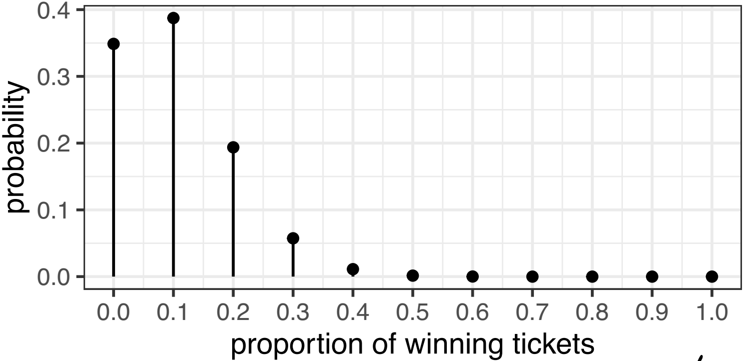
Example: $n = 5$ and $p = 0.1$. Note that $\bar{X}$ has discrete distribution.



0 1 2 3 4 5

Central limit theorem: more examples

Example: $n = 10$ and $p = 0.1$. Note that $\bar{X}$ has discrete distribution.



0 1 2 10

Central limit theorem: more examples

$N\left(0.1, \frac{0.1 \cdot 0.9}{100}\right)$

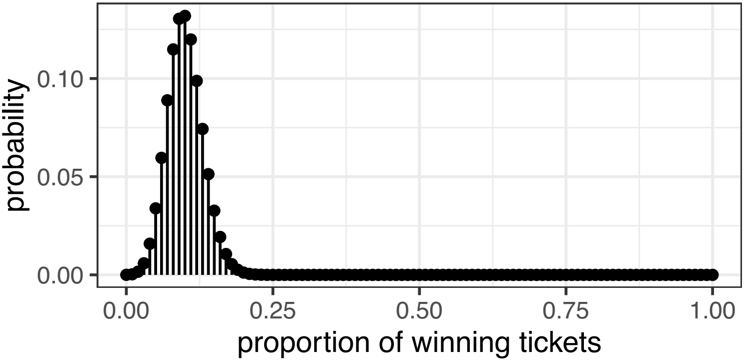
Example: $n = 100$ and $p = 0.1$. Note that if n is large enough

- For different samples the proportion of winning tickets will "jump around" p

$P(\bar{x} < 0.5)$

- The larger the sample size the closer the proportion of winning tickets to p

$0 \quad \frac{1}{100} \quad \frac{2}{100} \quad \cdots \quad \frac{99}{100} \quad 1$



Confidence intervals

We want to study the average life expectancy in Canada. (μ)

We take a sample of $\underset{n}{25}$ people and record their ages of death

```
ages
```

```
## [1] 74.7 82.8 72.6 97.0 84.3 72.8
```

. . . . - - - - - -

We compute the sample mean for these 25 people

```
mean(ages)
```

```
## [1] 82.7
```

$= \bar{x} \simeq \mu$

82.7 ± 5

We claim that it is an **estimate** of the average life expectancy in Canada. *How confident are we in our estimate?*

Confidence intervals

CLT:

$$\bar{X} \text{ approximately} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardization:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \text{ approximately} \sim Normal\left(0, 1\right)$$

$\sim N(0,1)$

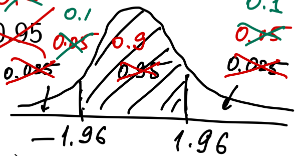
$P(-a \le Z \le a) = 0.95$

Distribution table:

-1.28 $\qquad$ 1.28

$-\cancel{1.64}$ $\qquad$ $\cancel{1.64}$ $\qquad$ $\cancel{0.9}$ $\quad$ 0.8

$$P\left(-\cancel{1.96} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \cancel{1.96}\right) = \cancel{0.95}$$



Interval for μ:

$$P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \ge -1.96 \Rightarrow \bar{X} - \mu \ge -1.96 \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \ge \mu$$

Confidence intervals (known σ)

95% confidence interval for μ:

$$\left[\underbrace{\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}}, \bar{x} + \underbrace{1.96 \cdot \frac{\sigma}{\sqrt{n}}}\right]$$

margin of error

$\sigma = 1 \qquad \bar{x} = 82.7$

$\mu \in \left[82.7 - 1.96 \cdot \frac{1}{5}, \; 82.7 + 1.96 \cdot \frac{1}{5}\right]$

$72.7 - 92.7 \quad 95\%$

$\boxed{81.7 - 83.7} \quad 50\%$

Exercise

How to find 90% confidence interval? How to find 80% confidence interval?

$$\left[\underbrace{\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}}_{\text{margin of error}}, \underbrace{\bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}}\right]$$

$\boxed{60\%}$ ↓

-1.64 $\quad$ $+1.64$ $\quad$ 90%

-1.28 $\quad$ $+1.28$ $\quad$ 80%

Confidence intervals (unknown σ)

95% confidence interval for μ:

$$\left[\bar{x} - 1.96 \cdot \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{s}{\sqrt{n}}\right]?$$

$$x_1 \ldots x_n \longrightarrow \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \rightarrow s = \sqrt{\ldots}$$

$$\bar{x} \simeq \mu \qquad s \simeq \sigma$$

$$\mu \in \left[82.7 - 1.96 \cdot \frac{9.5}{5} , 82.7 + 1.96 \frac{9.5}{5} \right]$$

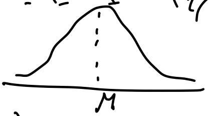
```
sd(ages)
```

```
## [1] 9.5
```
$\longleftarrow$ sample sd

Alternative view

$x_1 \dots x_n \longrightarrow \bar{x}, S, S^2$

1	2	3	4	5	6
1/6	1/6	.	.	.	1/6



- We have n random variables $X_1, \dots, X_n$
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a **random variable** $E(\bar{X}) = \mu$
- Each sample mean $\bar{x}$ is a realization of $\bar{X}$
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a **random variable** $E(S^2) = \sigma^2$
- Each sample variance s^2 is a realization of S^2

$\boxed{E(S^2) \neq \sigma^2}$

$\frac{1}{n}$

$X_1 + \dots + X_n \longrightarrow \bar{X}$

$x_1 + \dots + x_n \longrightarrow \bar{x}$

Confidence intervals (unknown σ)

RV

number

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

number

Standardization:

RV number

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \text{ approximately} \sim t_{n-1}$$

RV

t_{24}

t_3
t_4
t_{100}

- "t distribution with $n - 1$ degrees of freedom"
- It is similar to normal, but not quite...

Confidence intervals (unknown σ)

$t \to N(0,1)$

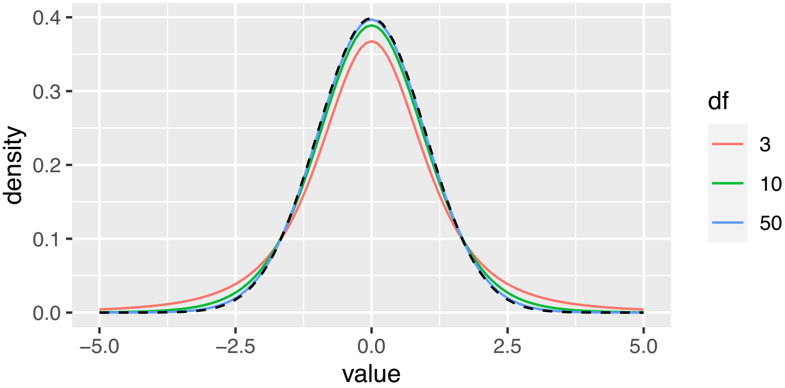
$n > 50$

Normal: $a = 1.96$

t with df = 3: $a = 3.18$

t with df = 10: $a = 2.23$

t with df = 50: $a = 2.01$



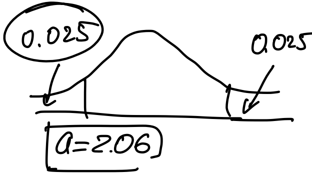
Confidence intervals (unknown σ)

95% : $\boxed{6}$ $a = 1.96$

95% confidence interval for μ:

sample sample

$$\left[\bar{x} - a \cdot \frac{s}{\sqrt{n}}, \bar{x} + a \cdot \frac{s}{\sqrt{n}}\right]$$



Where a is found from the distribution table.

$t_{n-1} = \boxed{t_{24}}$

-1.96 1.96

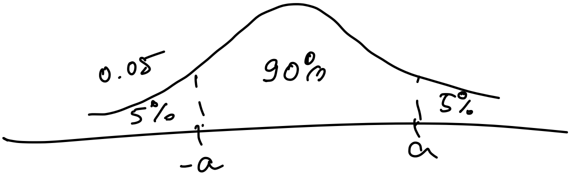
$$\left[82.7 - 2.06 \cdot \frac{9.5}{5}, 82.7 + 2.06 \cdot \frac{9.5}{5}\right]$$

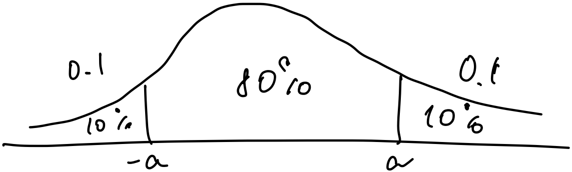
Exercise

$a = 1.71$

$\boxed{a = 1.32}$

How to find 90% confidence interval? How to find 80% confidence interval?





Confidence intervals: more examples

We want to estimate the probability to win in the lottery. (p)

We take a sample of 50 tickets and record the outcomes (1 - win, 0 - lose)

```
tickets
```

```
##  [1] 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0......
```

We compute the sample mean for these 50 outcomes
(i.e. proportion of winning tickets)

```
mean(tickets)
```

```
## [1] 0.08
```
$= \bar{x} \simeq p$

We claim that it is an **estimate** of the probability to win the lottery. *How confident are we in our estimate?*

Confidence intervals: more examples

$\frac{\sigma^2}{n}$

CLT:

$$\bar{X} \text{ approximately} \sim Normal\left(\overset{\mu}{p}, \frac{p(1-p)}{n}\right)$$

$\bar{X} \simeq p$

95% confidence interval for p (known σ):

$$\left[\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

$p(1-p) \simeq \bar{X}(1-\bar{X})$

95% confidence interval for p (unknown σ):

$$\left[\bar{x} - 1.96 \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + 1.96 \cdot \sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]$$

$$\left[0.08 - 1.96 \cdot \sqrt{\frac{0.08 \cdot 0.92}{50}}, 0.08 + 1.96 \sqrt{\ldots}\right]$$

1.64

1.28

Exercise

How to find 90% confidence interval? How to find 80% confidence interval?

TO DO

1. Module 6. Confidence Intervals Part 1 and Module 7. Confidence Intervals Part 2
2. Quiz 7 due Monday (March 6) @ 11:59 PM (EST)
3. Practice Problem Set 7