STA220H1: The Practice of Statistics I

Elena Tuzhilina

February 28, 2023

Please turn on your videos :)

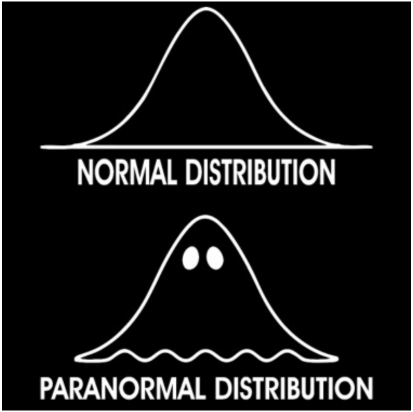


Figure 1: [picture source]

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

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

- 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

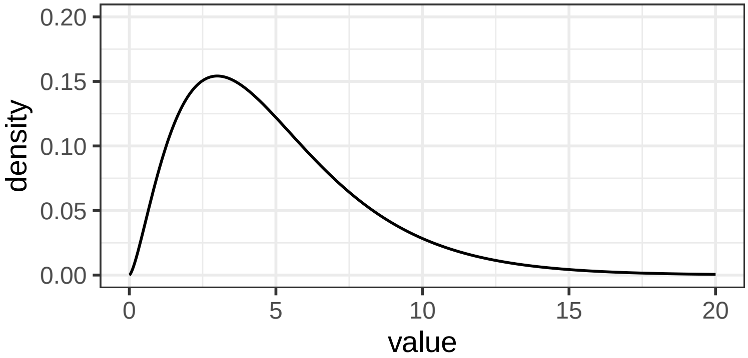
If $X_1, \ldots, X_n$ are independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ and $\bar{X} = \frac{X_1 + \ldots + X_n}{n}$ is the average of these random variables then

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

Recap: density curves

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

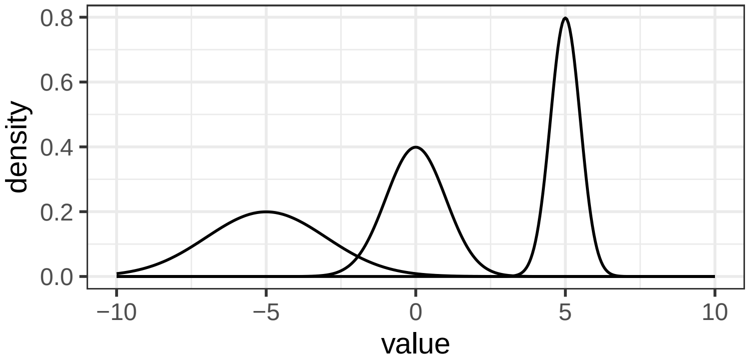
- The total area under the density curve is always 1
- The area under the curve bounded by a and b vertical lines is equal to $P(a \leq X \leq b)$



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
- $\sigma^2 = \mathit{Var}(X)$ controls the "spread" of the distribution

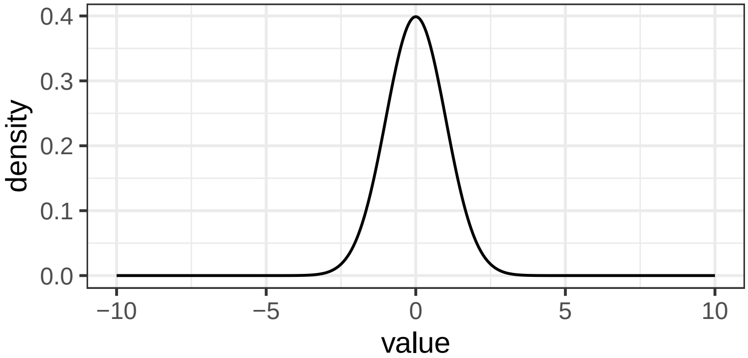


Recap: normal distribution

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

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



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.

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

$P(-6 \leq X \leq 6) =$

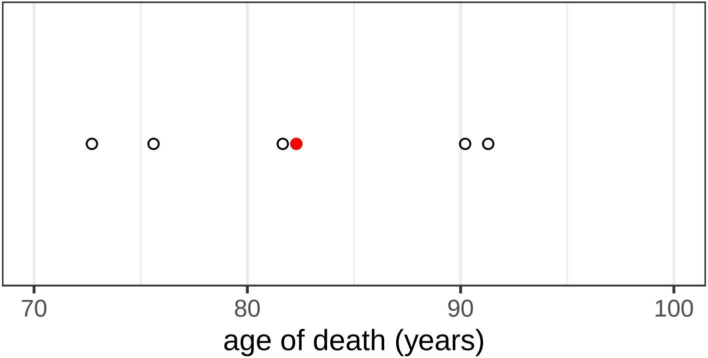
Recap: sample mean ditributoin

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.

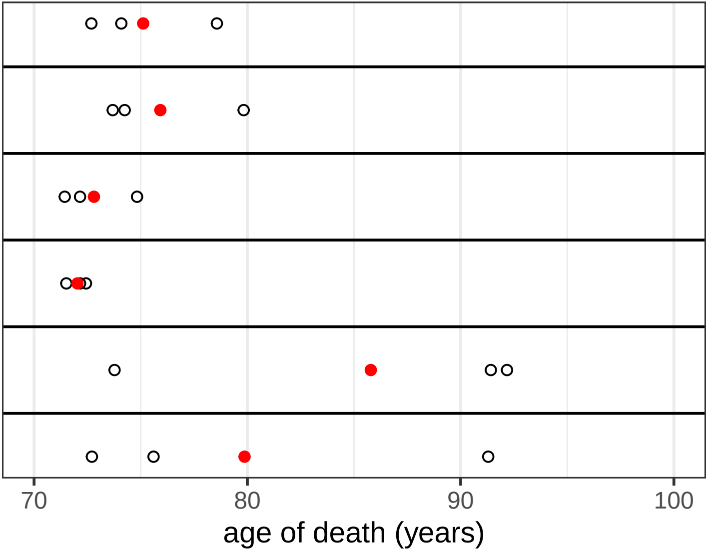
Black dots: sample $x_1, \ldots, x_n$

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



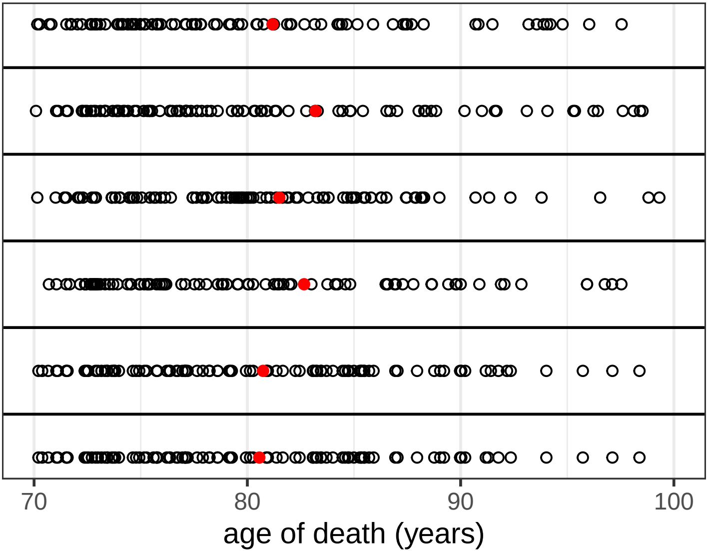
Recap: sample mean ditributoin

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



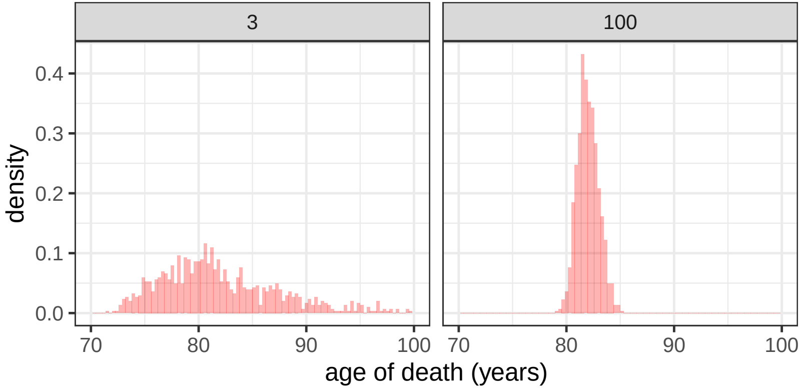
Recap: sample mean ditributoin

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



Recap: sample mean ditributoin

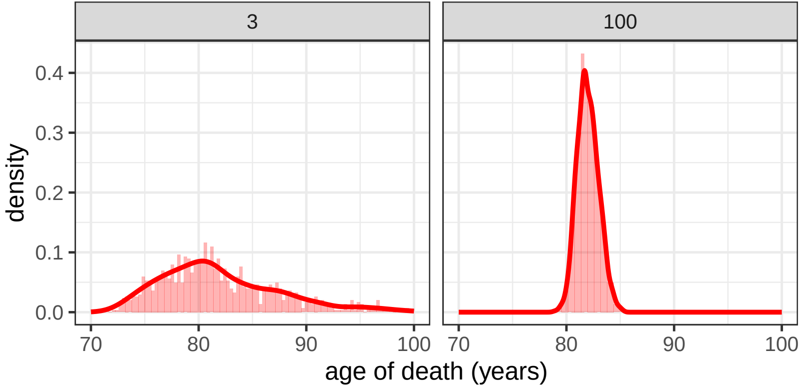
Can we characterize the behavior of $\bar{x}$?



Recap: alternative view

- We have n random variables $X_1, \ldots, X_n$
- 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

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

- 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**.

- $X_1, \ldots, X_n \sim Bernoulli(p)$ 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

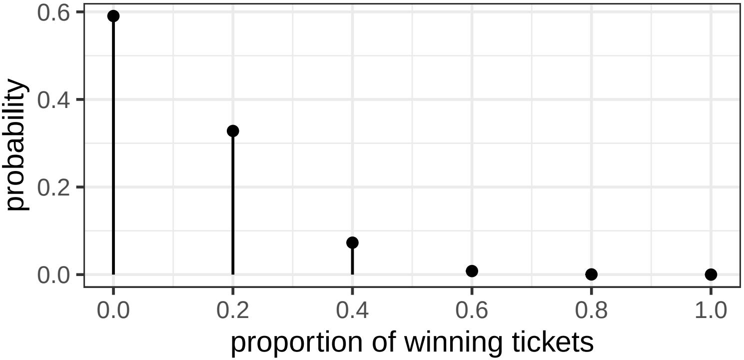
The probability to win in a lottery p. Suppose we buy n tickets.

$E(\bar{X}) =$

$Var(\bar{X}) =$

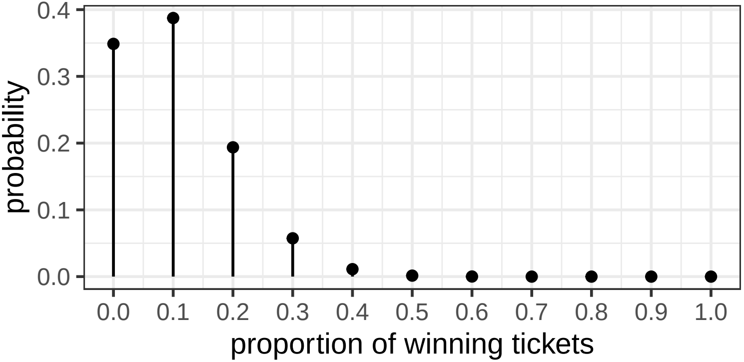
Central limit theorem: more examples

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



Central limit theorem: more examples

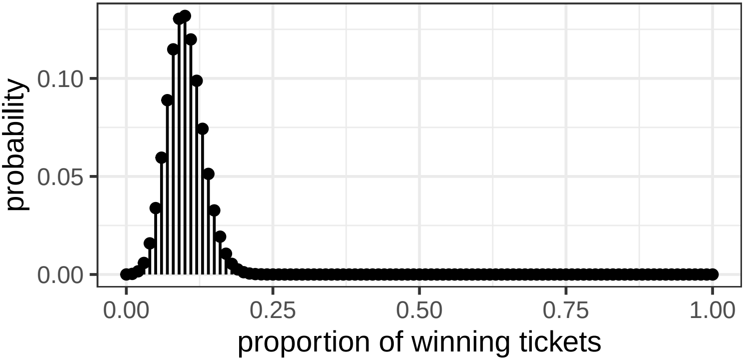
Example: $n = 10$ and $p = 0.1$. Note that $\bar{X}$ has discrete distribution.



Central limit theorem: more examples

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



Confidence intervals

We want to study the average life expectancy in Canada.

We take a sample of 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
```

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

Confidence intervals

CLT:

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

Standardization:

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

Distribution table:

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

Interval for μ:

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

Confidence intervals (known σ)

95% confidence interval for μ:

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

Exercise

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

Confidence intervals (unknown σ)

95% confidence interval for μ:

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

```r
sd(ages)
```

```
## [1] 9.5
```

Alternative view

- We have n random variables $X_1, \ldots, X_n$
- $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a **random variable**
- 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**
- Each sample variance s^2 is a realization of S^2

Confidence intervals (unknown σ)

Standardization:

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

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

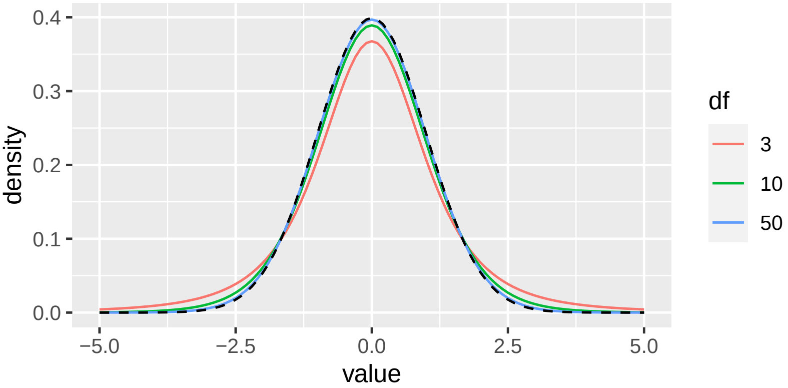
Confidence intervals (unknown σ)

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% confidence interval for μ:

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

Where a is found from the distribution table.

Exercise

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.

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
```

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

CLT:

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

95% confidence interval for p (known σ):

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

95% confidence interval for p (unknown σ):

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

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