

JBMO50

Statistical Computing

Q1 2024-2025

Lecture 2: Sampling distribution

Part I: Basics in probability & statistics

- 1.0. Introduction: Learning from your data

- 1.1. The data generating model in random experiments

} Last time

- 1.2. Population model and parameter(s), sample statistics

- 1.3. The sampling distribution

} Today

- 1.4. Mean, bias, variance, mean squared error

First: binomial trial experiment + brief recap (illustrative exam questions) + pseudo code

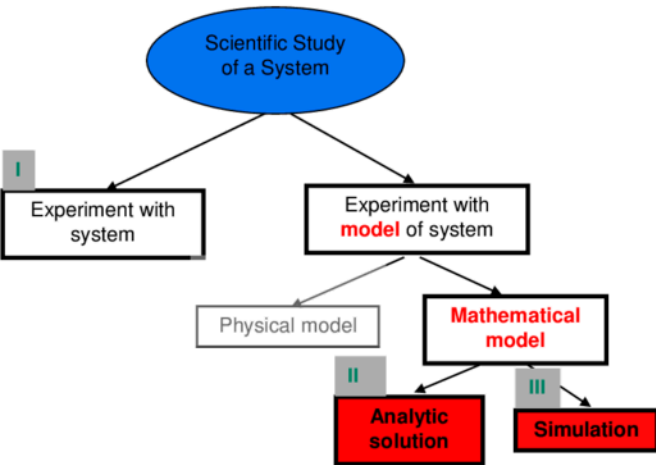
Illustrative exam questions

- The Bernoulli distribution is a special case of the Binomial distribution. TRUE or FALSE.

=> TRUE: $X \sim \text{Bin}(n = 1, \pi)$ is $X \sim \text{Bern}(\pi)$

Illustrative exam question: Answers

- ❑ When tossing a fair coin 6 times, what is the probability of obtaining exactly four times tails in a row?
- ❑ 3 approaches possible to answer this question



Approach 1: not very practical

Approach 2+3: see next slides

Note: Approach 2 gives exact solution; 1 & 3 approximate

Approach 2: Analytical

- Classical definition of probability (*as seen in Data Statistics*):

$$P(A) = \frac{\text{number of outcomes we are interested in}}{\text{total number of outcomes}} = \frac{N(A)}{N}$$


- When tossing a fair coin 6 times, what is the probability of obtaining exactly four times tails **in a row**?
1. How do we calculate $N(A)$ in this question? Explain.
 2. How do we calculate N in this question? Explain.
 3. Calculate the probability of obtaining exactly four times tails (T) **in a row** when tossing a fair coin 6 times.

□ Approach 2: Exact analytical solution

$$P(4T \text{ in a row} | 6 \text{ tosses}) = \frac{\text{Nr of ways } 4T \text{ in a row can happen}}{\text{Total nr of outcomes}}$$

$$= \frac{5}{2^6} = 0.0781$$

Approach 3: Simulation

- ❑ Find an approximate answer using computer simulation
- ❑ First, structure your code on paper, using pseudo code
- ❑ Pseudo code is also a great way (and standard) of communicating your algorithms
- ❑ Then implement your routine (in )

■ Step 1: What is **input**?

- Number of coin tosses : $n = 6$
- Number of tails in a row: $k = 4$
- Number of experiments: S

■ Step 2: What is **output**?

- 'Success' if we observe exactly k times tails in a row, 'fail' else
- Calculate proportion of successes:
$$\frac{\text{number of 'success'}}{S}$$

■ Step 3: describe steps to derive output from input

Input:

- *S*: integer defining the number of experiments
- *nr_tosses*: integer defining the number of coin tosses
- *nr_tails*: integer defining the number of tails in a row

Output: proportion of trials with exactly *nr_tails* times tails in a row

1. Set *Out_vec* a vector of length *S*
2. Repeat *S* times
 - A. Store a random ("heads","tails") sample of size *nr_tosses* in *Out*
 - B. Count the *nr_seq1* with *nr_tails* "tails" in *Out*
 - C. If *nr_seq1* > 0 then
 - Count the *nr_seq2* with (*nr_tails*+1) "tails" in *Out*
 - If *nr_seq2* = *nr_seq1*
 - Out_vec*[*s*] = 0
 - Else
 - Out_vec*[*s*] = 1
 - Else
 - Out_vec*[*s*] = 0
3. Return *sum*(*Out_vec*) / *S*

Note: we assume *nr_tosses* smaller than $2 \cdot \text{nr_tails}$)

```

kTails=function(n,k,S){
  # n is number of coin flips (with  $n < 2*k$  assumed)
  # k is number of tails in a row
  # S is number of times we repeat coin tossing experiment
  outvec <- vector(mode = 'numeric', length = S)
  nrtails <- toString(rep('tail', k))
  nrtailsplusone <- toString(rep('tail', k+1))
  for (i in 1:S){
    Out <- toString(sample(c('head','tail'), n, replace=TRUE))
    nrmatch <- sum(gregexpr(nrtails, Out)[[1]]>0)
    if(nrmatch!=0){
      if(sum(gregexpr(nrtailsplusone, Out)[[1]]>0)>=nrmatch){
        outvec[i] <- 0
      }else{
        outvec[i] <- 1
      }
    }else{
      outvec[i] <- 0
    }
  }
  return(sum(outvec)/S)
}

```

kTails(6,4,10⁵)

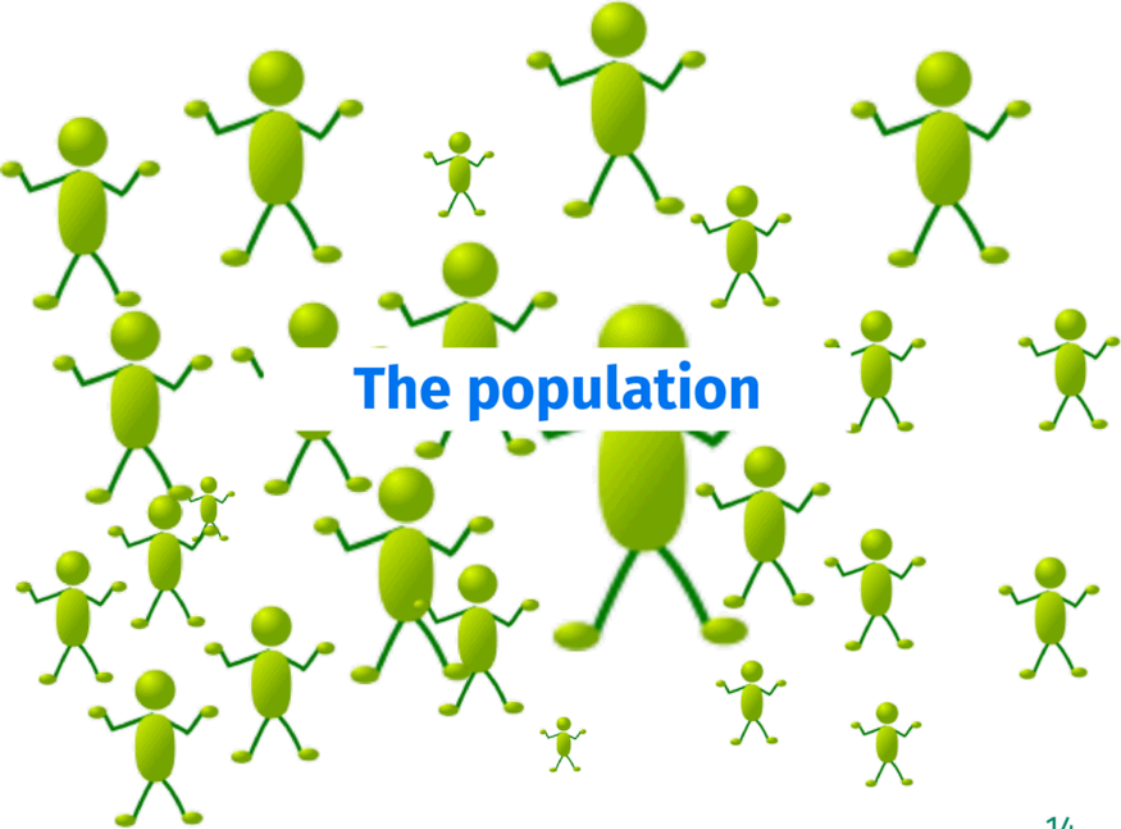
use function for probability of $k = 4$ tails when flipping coin $n = 6$ times and repeating the experiment $S = 100.000$ times

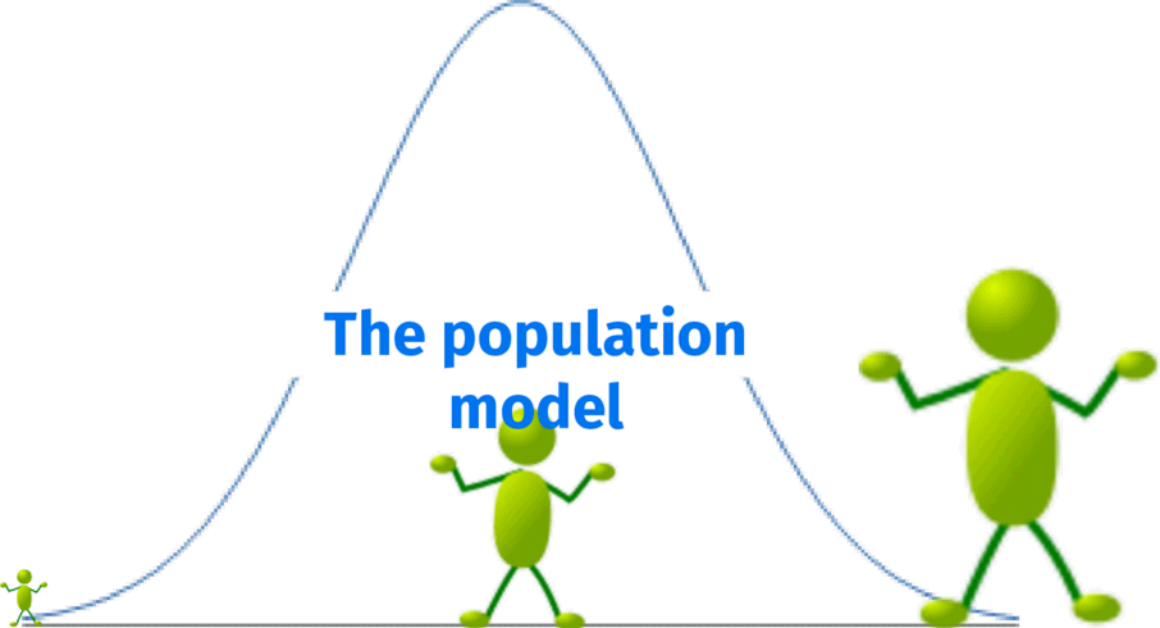
Part I: Basics in probability & statistics

- I.0. Introduction: Learning from your data
- I.1. The data generating model in random experiments
- 1.2. Population model and parameter(s), sample statistics
- I.3. The sampling distribution
 - 1.3.0. Definitions
 - 1.3.1. Example 1: Sampling distribution of the sample mean
 - 1.3.2. Example 2: Sampling distribution of the sample proportion
- I.4. Mean, bias, variance, mean squared error

1.2. Population model and parameter(s), sample statistics

The world through the eyes of a statistician





How statisticians view the world

The population model

- When studying a *property of interest*, e.g. body length, for a *population of interest*, e.g. adult women, a *model for this population* is assumed
 - E.g., body length of adult women is *normally distributed* with population mean equal to 161 cm and variance equal to 36 cm
 - Or, body length adult women $\sim N(161, 36)$
 - Or, body length adult women $\sim N(\mu, \sigma^2)$ with $\mu = 161$ and $\sigma^2 = 36$ where the mean μ and the variance σ^2 are the *parameters* of the population model and N is the assumed probability distribution
- Usually, the population model is unknown so it needs to be *estimated*, meaning that the model parameters need to be estimated (the normal shape is not questioned)

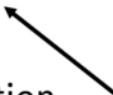
The normal distribution

- Is the statistical model that is assumed to underlie many continuous random variables

$$X \sim N(\mu, \sigma^2)$$

- Also known as the Gaussian distribution

The statistical model
for real-valued
random variables

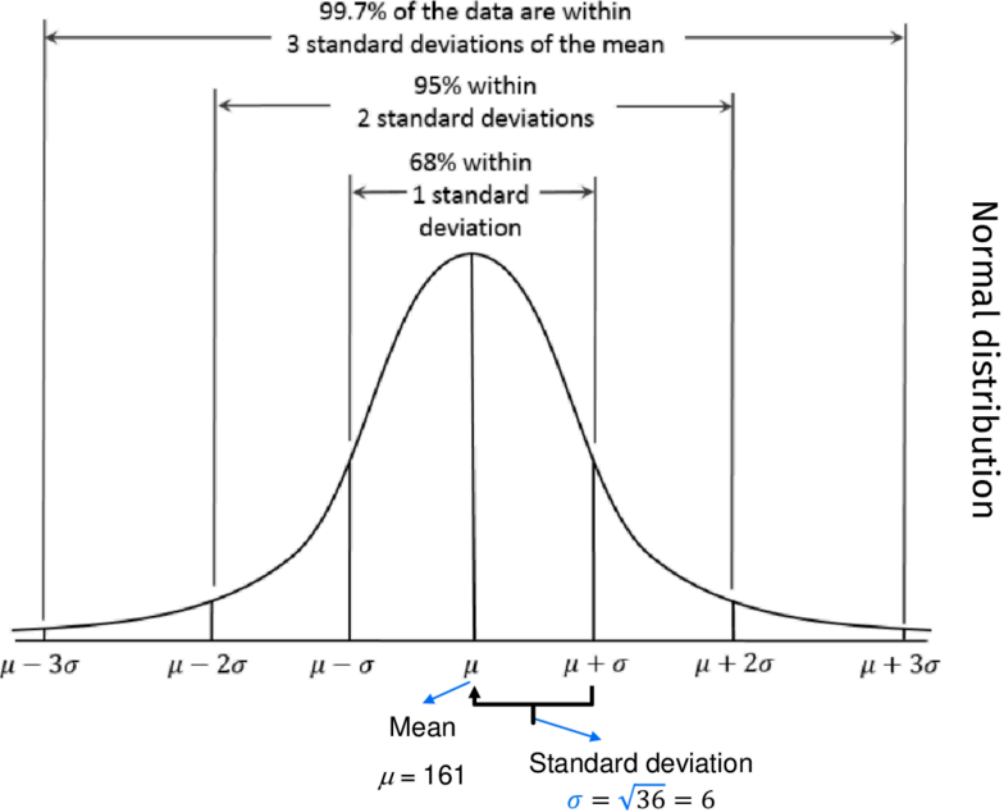


- Has many nice properties, e.g., central limit theorem

Key ingredients of statistical models:

- 1) Shape / Distribution and
- 2) Model parameters

E.g., Normal Shape with parameters μ, σ^2 or Bernoulli distributed with parameter π



IQ scores

- ❑ The population of adult IQ scores based on the Wechler Adult Intelligence scale (WAIS) is normally distributed with mean equal to 100 and standard deviation equal to 15. The category for people with IQ **equal to or less than** 79 is considered borderline/defective.
- ❑ Calculate the probability of someone in the adult population being classified as borderline/defective.
- ❑ *Note: 80 was commonly used as a cutoff for labeling someone mentally disabled*

Answer


□ $IQ \sim N(100, 15^2)$

□ Formula for standardization: $Z_{IQ} = \frac{IQ - \mu}{\sigma}$, with $\mu = 100$ and $\sigma = 15$ and $IQ = 79$

$$\Rightarrow Z_{IQ=79} = \frac{79 - \mu}{\sigma} = \frac{79 - 100}{15} = \frac{-21}{15} = -1.40$$

□ Z table:

-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

□ Or, using  :

```
> pnorm(-1.4)
[1] 0.08075666
```

□ Thus: $P(IQ \leq 79) = 0.081$

Estimating the population model

- Usually a particular distribution is assumed for a random variable of interest (e.g., body length, IQ score, assumed to follow a normal distribution)
- To characterize this variable, only the population parameters need to be inferred. This is, **estimates of the population parameters** are needed.
 - E.g., body length for Dutch adult women (assuming a normal distribution):

$$X \sim N(?, ?)$$

- Any ideas on how to obtain estimates of the population parameters?

Sample statistics as estimators of population parameters

□ Principle:

- To characterize the **population**, a (random) **sample** of the population is taken and the **population parameter** of interest is **estimated** by the **sample statistic**
- E.g., the body length of 500 Dutch adult women is measured and their average length is used to estimate the mean and the sample variance is used to estimate the population variance

Basic concepts: Parameter, Estimator, Estimate

□ What are 'estimators'?

- In statistics, a (point) **estimator** is an approximation of a population parameter that *uses observed data*. It is also called a **statistic**.
- Rule = estimator; its value = **estimate**.

Example 1: Length Dutch adult women

■ Population model?

$$N(\mu, \sigma^2)$$

■ Population
parameters of
interest?

$$\mu, \sigma^2$$

■ Estimator/statistic?

$$\hat{\mu} = \bar{x} \quad (\text{sample mean})$$
$$\hat{\sigma}^2 = s^2 \quad (\text{sample variance})$$

■ Estimate?

$$\hat{\mu} = 161$$
$$\hat{\sigma}^2 = 36$$

Example 2: Coin tossing experiment

Toss a coin. Does head turn up?

55

Yes

53%

No

47%

■ Population model?

$X \sim \text{Bern}(\pi)$

■ Population parameter of interest?

π , probability of heads

■ Estimator/statistic?

$$\hat{\pi} = p = \frac{k}{n} = \frac{\text{nr. heads up}}{\text{total nr of tosses}}$$

■ Estimate?

$\hat{\pi} = 0.53$

■ Crucial questions about estimators/statistics, e.g. sample mean:

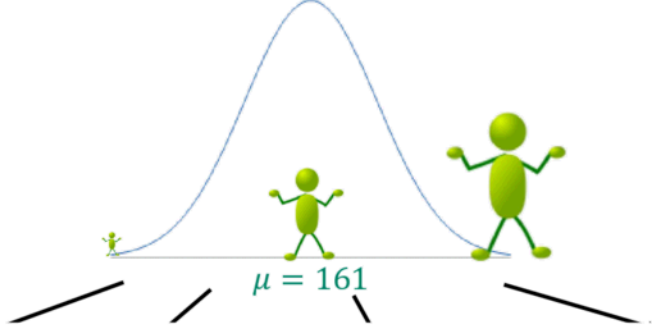
- *How well will the sample mean approximate the population mean?* What, for example, can we expect about the sample mean as an estimate of the population mean?
- This is, **for a particular sample**, how far is the sample statistic from the value of the population parameter?
- More precisely, for a particular sample of given **sample size n** , how far *can* the sample statistic be from the value of the population parameter?
- The answer to these questions is the **sampling distribution**.

1.3. The sampling distribution

1.3.0. Sampling distribution: Definition

□ What?

- The **sampling distribution** is the *probability distribution of the sample statistic* for a *given sample size n*
- It describes how the sample statistic varies over different samples of size n
- This is, *the statistic is considered to be a random variable!*



Observe that the sample average varies over the different samples;

The sample average is itself a random variable!

$\Rightarrow \bar{x} = 163$

$\Rightarrow \bar{x} = 165$

$\Rightarrow \bar{x} = 156$

$\Rightarrow \bar{x} = 163$

(You can generate samples with R as follows: `rnorm(n = 5, mean = 161, sd = 6)`)

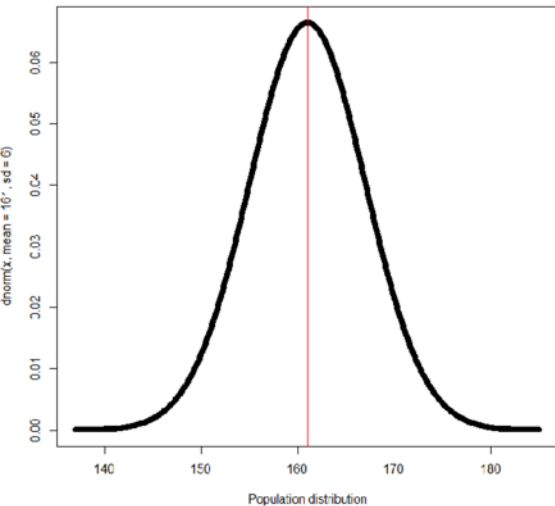
1.3.1. Example 1: Sampling distribution of the sample mean for data from a normal population

- From **statistical theory** we know:

$$\text{If } X \sim N(\mu, \sigma^2) \quad \text{then} \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

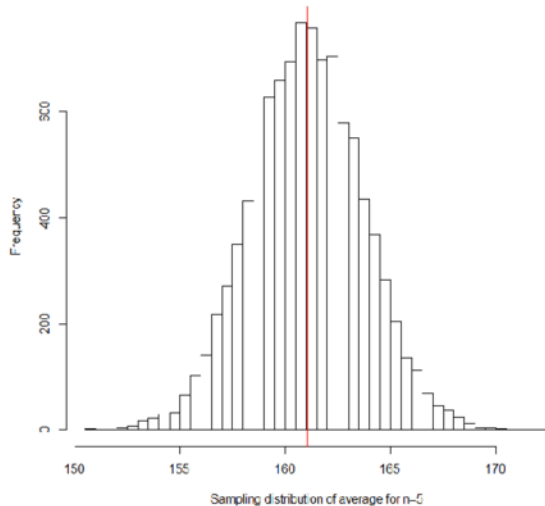
- In words : the **sampling distribution of the average** of a sample of size n taken from a **population that is normally distributed** with mean μ and variance σ^2 is itself a normal distribution with the same mean but smaller variance
- Note: the standard deviation (=square root of variance) of a sample statistic is also called the **standard error** of that statistic
 - So standard error of the sample mean is ...

Population



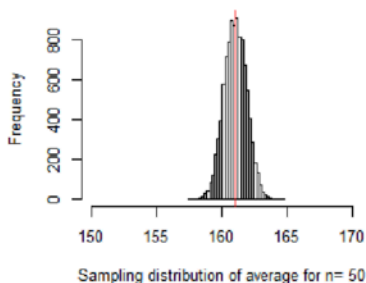
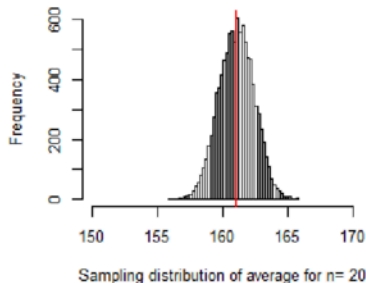
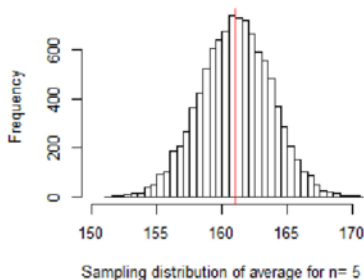
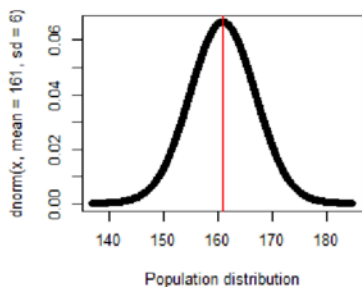
$$X \sim N(161, 36)$$

Sampling distribution

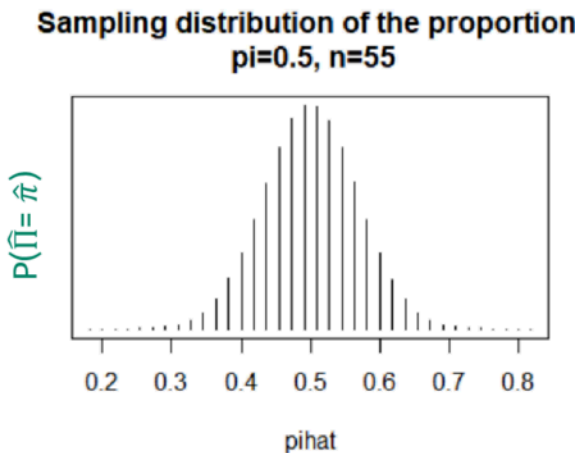
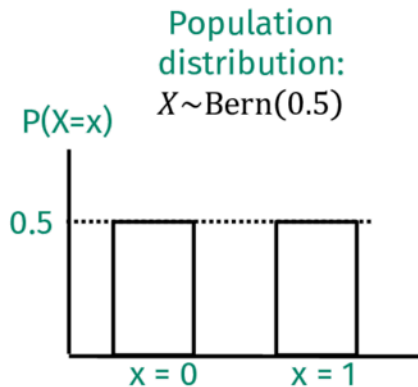


$$\bar{X} \sim N\left(161, \frac{36}{5}\right)$$

Influence of sample size on the sampling distribution



1.3.2. Example 2: Bernoulli population and sampling distribution of the proportion



Why is the sampling distribution so important?

- ❑ Based on a study conducted on a sample with size n for a population of adults being 18-21 years old and having cholesterol level between 200-225, your doctor tells you that your expected survival time is 15 years.
- ❑ Should you worry?
- ❑ This is based on a study conducted on a sample of size $n = 100$
- ❑ And let's assume it has the following sampling distribution

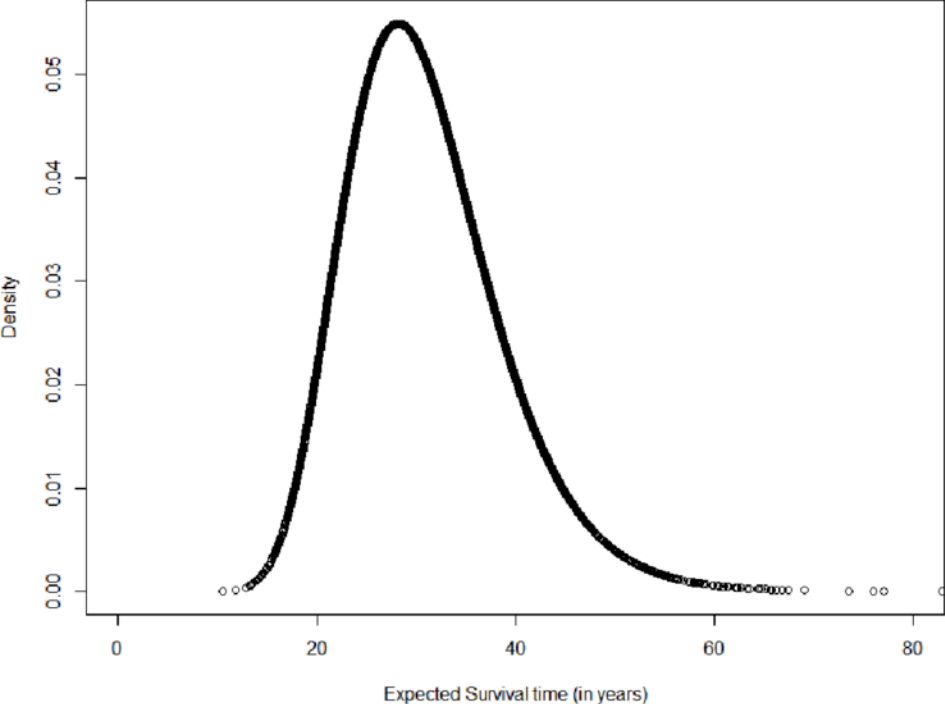


Figure 1.3.2. (Hypothetical) sampling distribution of the expected survival time after diagnosis for $n = 100$

- ❑ The sampling distribution expresses our **uncertainty** about the statistic as an estimate of the population parameter
- ❑ It shows what other values of the statistic could have been obtained when taking **another sample** from the same population and with the same sample size
- ❑ Crucial question: how does the sampling distribution compare to the population parameter of interest?
- ❑ That's up next (in the next lecture).

Part I: Basics in probability & statistics

- 1.0. Introduction: Learning from your data
- 1.1. The data generating model in random experiments

□ 1.2. Population model and parameter(s), sample statistics

□ 1.3. The sampling distribution

□ 1.4. Mean, bias, variance, mean squared error

+ Monte Carlo Simulation (Part II)

} Today

} Next time