Introduction to Statistical Inference

Felipe José Bravo Márquez

August 25, 2021

- The main goal of statistical inference is investigate properties about a target **population**.
- A **population** is the entire group of individuals that we are interested in studying.
- This could be anything from all humans to a specific type of cell.
- The individual elements of the population sometimes are called **units**.
- Example: What is the average height of all people in Chile? Here the population is all the inhabitants of Chile.
- In order to draw conclusions about a **population**, it is generally not feasible to gather all the data about it.
- The special case where you collect data on the entire population is a **census**.

Populations and Samples

- In statistical inference we try to make reasonable conclusions about a population based on the evidence provided by **sample data**.
- A **sample statistic** or simply **statistic** is a quantitative measure calculated from a sample. Examples: the mean, the standard deviation, the minimum, the maximum.

Example

- Taking a sample **survey** can help you determine the percentage of people in a population who have a particular characteristic.
- Nielsen Media Research takes a **survey** so they can get an estimate of the proportion of all U.S. households that are tuned to a particular television program [Watkins et al., 2010].
- The true proportion that Nielsen would get from a survey of **every** household is called a **population parameter**.
- Nielsen uses the proportion in the **sample** as an **estimate** of this parameter.
- Such an estimate from a sample is called a **statistic**.

Sampling Methods

- Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population.
- We do this primarily to save time and effort.
- A good sample is **representative**: it looks like a small version of the population.
- In practice, we can't tell if if a sample is representative, since we can't get all the population data.
- But, we can tell whether a **sampling method** is good or not.
- A sampling method is **biased** if it produces samples such that the estimate from the sample is larger or smaller, on average, than the population parameter being estimated.
- The problems or **biases** of a sampling method can come from two sources:

    1. **Sampling selection bias**: the way in which the sample is built.
    2. **Response bias**: the method for getting a response.

Sample selection bias

Sample selection bias happens when samples tend to result in estimates of population parameters that systematically are too high or too low [Watkins et al., 2010].
It can occur in the following ways:

- **Size bias**: using a method that gives larger units a bigger change of being in the sample.
  - Example: patients who spent more days in a hospital are more likely to be selected for the sample.

- **Voluntary response bias**: letting people volunteer to be in the sample.
  - Example: When a radio program asks people to call in and take sides on some issue, those who care about the issue will be over-represented, and those who don't care as much might not be represented at all.

- **Convenience sample bias**: units are chosen because of convenience.
  - Example: What percentage of the students in your graduating class plan to go to work immediately after graduation?
  - We can use our friends as a quicker and more convenient sample, but almost certainly biased because friends are unlikely to be representative of the entire target population.

Sample selection bias

- **Judgment sampling bias**: selecting the sampling units based on "expert" judgment. Problem: experts might overlook important features of a population.
    - Example: In the early days of US election polling, local "experts" were hired to sample voters in their locale by filling certain quotas (so many men, so many women, so many voters over the age of 40, so many employed workers, and so on).
    - The poll takers used their own judgment as to whom they selected for the poll.
    - It took a very close election (the 1948 presidential election, in which polls were wrong in their prediction) for the polling organizations to realize that quota sampling was a biased method.

Sample selection bias

- **Sampling frame bias**: a sampling frame is the "list" of all population units from which you select the sample. Constructing an inadequate sampling frame is a cause of bias.
    - The problem is that in many problems is extremely hard to make that "list".
    - How would you list all the people using the Internet worldwide or all the ants in a park?
    - For all practical purposes, you can't.
    - There will often be a difference between the population and the sampling frame.
    - A sample might represent the units in the frame quite well, but how well your sample represents the population depends on how well you've chosen your frame.
    - If you start from a bad frame, even the best sampling methods can't save you: bad frame, bad sample [Watkins et al., 2010].

Response bias

These types of bias derive from the method of obtaining the response.

- **Nonresponse bias**: people often refuse to respond to a survey. These people may be different from those who agree to participate.

- **Incorrect response or measurement bias**: the bias might be the result of intentional lying, or come from inaccurate measuring devices, including inaccurate memories of people being interviewed in self-reported data.
  - Example 1: many people don't want to admit that they watch a certain television program.
  - Example 2: patients in medical studies are prone to overstate how well they have followed the physician's orders.
  - Example 3: many people tend to underestimate the amount of time they actually spend with their cell phones.

Response bias

- **Questionnaire bias**: people's opinions may vary depending on the interviewer's tone of voice, the order in which the questions are asked and the wording of the questions, etc..
    - Example: Reader's Digest asked the same 1031 people to respond to these two statements:
        1. I would be disappointed if Congress cut its funding for public television.
        2. Cuts in funding for public television are justified as part of an overall effort to reduce federal spending.
    - Note that agreeing with the first statement is pretty much the same as disagreeing with the second. However:
    - First statement: 54% agreed, 40% disagreed, and 6% didn't know.
    - Second statement: 52% agreed, 37% disagreed, and 10% didn't know. [Barnes, 1995]

Random Samples

- The key idea for building a good sample is to **randomize**, that is, let chance choose the sampling units.
- Selecting your sample by chance is the only method guaranteed to be unbiased.

Simple Random Sampling

- All possible samples of a given fixed size are equally likely.
- All individuals in the population are indexed and randomly drawn with equal probability until the sample size is reached.

Stratified Random Sampling

- We divide the population into subgroups based on shared characteristics (e.g., country, city) that do not overlap and that cover the entire sampling frame.
- These subgroups are called strata.
- Take a simple random sample for each strata proportional to its size.
- It ensures that every strata is properly represented in the sample.

Simple random sample
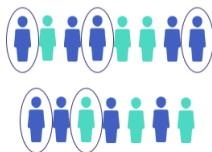
Stratified sample

[1]Figure source:
https://www.scribbr.com/methodology/sampling-methods/
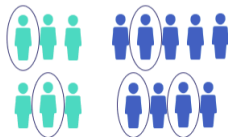
A Formal Definition of Statistical Inference

- The process of drawing conclusions about a population from sample data is known as **statistical inference**.
- From a general point of view, the goal of inference is to **infer** the distribution that generates the observed data.
- Example: Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F?
- However, in most cases we are only interested in inferring some property of F (e.g., its **mean** value).
- Statistical models that assume that the distribution can be modeled with a finite set of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ are called **parametric models**.
- Example: if we assume that the data comes from a normal distribution $N(\mu, \sigma^2)$, μ and σ would be the parameters of the model.

Frequentist Approach

The statistical methods to be presented is this class are known as **frequentist (or classical)** methods. They are based on the following postulates [Wasserman, 2013]:

- Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.
- Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

There is another approach to inference called **Bayesian inference**, which is based on different postulates, to be discussed later in the course.

Point Estimation

- Point estimation is the process of finding the **best guess** for some quantity of interest from a **statistical sample**.
- In a general sense, this quantity of interest could be a parameter in a parametric model, a CDF F, a probability density function f, a regression function r, or a prediction for a future value Y of some random variable.
- In this class we will consider this quantity of interest as a **population parameter** θ.
- By convention, we denote a point estimate of θ by $\hat{\theta}$ or $\hat{\theta}_n$.
- It is important to remark that while θ is an unknown fixed value, $\hat{\theta}$ depends on the sample data and is therefore a random variable.
- We need to bear in mind that the process of sampling is by definition a **random experiment**.

Point Estimation

Formal Definition

- Let $X_1, \ldots, X_n$ be n IID data points from some distribution F.
- A point estimator $\hat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

- The **bias** of an estimator is defined as:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- An estimator is unbiased if $\mathbb{E}(\hat{\theta}_n) = \theta$ or $\text{bias}(\hat{\theta}_n) = 0$

Sampling Distribution

- If we take multiple samples, the value of our statistical estimate $\hat{\theta}_n$ will also vary from sample to sample.
- We refer to this distribution of our estimator across samples as the **sampling distribution** [Poldrack, 2019].
- The sampling distribution may be considered as the distribution of $\hat{\theta}_n$ for all possible samples from the same population of size *n*.
- The sampling distribution describes the variability of the point estimate around the true population parameter from sample to sample.
- We need to bear in mind this is an imaginary concept, since in real situations we can't obtain all possible samples.
- Actually, in most cases we will only work with a single sample.

Standard Error

- The standard deviation of $\hat{\theta}_n$ is called the **standard error** *se*:

$$se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- The standard error tells us about the variability of the estimator between all possible samples of the same size.
- It can be considered as the standard deviation of the sampling distribution.
- It is a measure of the uncertainty of the point estimate.

The Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be a random sample of a population of mean μ and variance σ^2.
- Let's suppose that we are interested in estimating the **population mean** μ (e.g., the mean height of Chilean people).
- A sample statistic we can derive from the data is the **sample mean** $\overline{X_n}$:

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- The sample mean is a **point estimator** of the mean $\overline{X_n} = \hat{\mu}$.
- We can show that the sample mean is an unbiased estimator of μ:

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n} \times \mathbb{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}(n \times \mu) = \mu$$

- The standard error of the sample mean $se(\overline{X_n}) = \sqrt{\mathbb{V}(\overline{X_n})}$ can be calculated as:

$$\mathbb{V}(\overline{X_n}) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} X_i) = \frac{n}{n^2}\mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Then,

$$se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$$

- The formula for the standard error of the mean implies that the quality of our measurement involves two quantities: the population variability σ, and the size of our sample n.

The Standard Error of the Sample Mean

- We have no control over the population variability, but we do have control over the sample size.
- Thus, if we wish to improve our sample statistics (by reducing their sampling variability) then we should use larger samples.
- However, the formula also tells us something very fundamental about statistical sampling.
- That the utility of larger samples diminishes with the square root of the sample size.
- This means that doubling the sample size will not double the quality of the statistics; rather, it will improve it by a factor of $\sqrt{2}$. [Poldrack, 2019]

Sample Variance

- A common problem when calculating $se(\overline{X_n})$ is that, in general, we do not know σ of the population.
- In those cases we can estimate σ using the **sample variance** s:

$$s^2 = \frac{1}{n-1} \sum_{i}^{n} (X_i - \overline{X_n})^2$$

- This is an unbiased estimator of the variance.
- The standard error of the sample mean when the population variance is unknown can be estimated as follows:

$$\hat{se}(\overline{X_n}) = \frac{s}{\sqrt{n}}$$

Population Variance

- There is also the population variance, defined as follows:

$$\sigma^2 = \frac{1}{N}\sum_{i}^{N}(X_i - \overline{X_N})^2$$

- The population variance should only be calculated from population data (all the individuals).
- Note that we are using *N* instead of *n* to denote the entire population rather than a sample.
- If is calculated from a sample, it would be a **biased** estimator of the population variance.

The Sampling Distribution of the Sample Mean

- We discussed earlier that the sampling distribution is an imaginary concept.
- Let's imagine the sampling distribution of the sample mean.
- Imagine drawing (with replacement) all possible samples of size *n* from a population.
- Then for each sample, calculate the sample statistic, which is this case is the sample mean.
- The frequency distribution of those sample means would be the sampling distribution of the mean (for samples of size *n* drawn from that particular population).
- In the next example we will calculate the sampling distribution for a toy example in which the population is known.

- Suppose our entire population is a family of 5 siblings and our property of interest is age measured in years.
- Our population consists of the following 5 values: 2, 3, 4, 5, and 6.
- Let's calculate the population mean μ and the population standard deviation σ.

```
> pop <-c(2,3,4,5,6)
> mean(pop)
[1] 4
> sd.p=function(x){sd(x)*sqrt((length(x)-1)/length(x))}
> sd.p(pop)
[1] 1.414214
```

μ=4 and $\sigma = 1.414214$

- Now, we will use the R library "gtools" to draw all 25 possible samples (with replacement) of size 2.

```
> library(gtools)
> library(tidyverse)
> samp_size <- 2
> samples<-as_tibble(permutations(length(pop), samp_size,
+                                 pop, repeats.allowed=TRUE))
> samples
# A tibble: 25 x 2
      V1    V2
   <dbl> <dbl>
 1     2     2
 2     2     3
 3     2     4
 4     2     5
 5     2     6
 6     3     2
 7     3     3
 8     3     4
 9     3     5
10     3     6
# ... with 15 more rows
```

- We can calculate the sample mean of each sample using the command "mutate":

```
> samples <- samples %>% rowwise() %>%
+   mutate(sample_mean=mean(c(V1,V2)))
> samples
# A tibble: 25 x 3
# Rowwise:
      V1    V2 sample_mean
   <dbl> <dbl>       <dbl>
 1     2     2           2
 2     2     3         2.5
 3     2     4           3
 4     2     5         3.5
 5     2     6           4
 6     3     2         2.5
 7     3     3           3
 8     3     4         3.5
 9     3     5           4
10     3     6         4.5
# ... with 15 more rows
```

The Sampling Distribution of the Sample Mean

- The distribution of these sample means is the **sampling distribution**.
- We can visualize its shape by plotting an histogram:

```
ggplot(samples, aes(x=sample_mean)) +
  geom_histogram(bins = 10, color="black", fill="white")
```
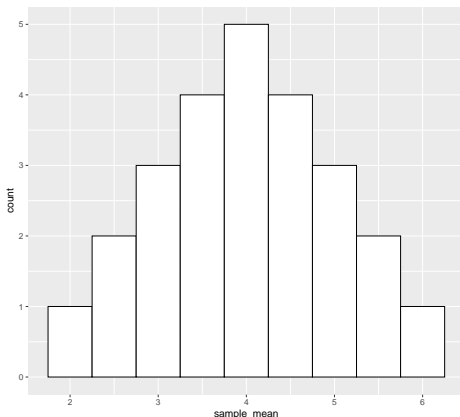
The Sampling Distribution of the Sample Mean

- You may noticed that the histogram is peaked in the middle, and symmetrical.
- This is a consequence of the Central Limit Theorem!!!
- We can see that the population distribution is very different from the sampling distribution:

```
ggplot(data.frame(pop), aes(x=pop)) +
  geom_histogram(bins = 5, color="black", fill="white")
```
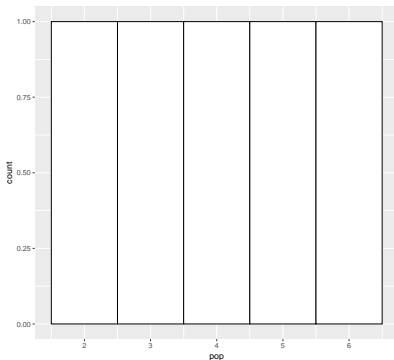
The Sampling Distribution of the Sample Mean

- Let's calculate the mean and the standard deviation of the sample means:

```
> mean(samples$sample_mean)
[1] 4
> sd.p(samples$sample_mean)
[1] 1
```

- We can see that mean of the sampling distribution of the mean $\mu_{\overline{X}}$ equals the population mean μ.
- We can also calculate the theoretical standard error $se = \sigma/\sqrt{n}$

```
> sd.p(pop)/sqrt(samp_size)
[1] 1
```

which is the same as the standard distribution of the sampling distribution of the sample mean.

- We have validated empirically that the sample mean is a good estimator of the population mean and that its standard error can be calculated from the population standard deviation and the sample size.

The Sampling Distribution of the Sample Mean

- The central limit theorem tell us the conditions under which the sampling distribution of the mean is normally distributed or at least approximately normal.
- If the population from which you sample is itself normally distributed, then the sampling distribution of the mean will be normal, regardless of sample size.
- If the population from which you sample is non-normal, the sampling distribution of the mean will still be approximately normal given a large enough sample size.
- What size is sufficient? Some authors say 30 or 40. But if the population distribution is extremely non-normal (i.e. very skewed) you will need more.

Point Estimation of a Proportion

- Suppose we want to estimate the fraction of people who will vote for a certain candidate.
- Our population parameter p corresponds to the true fraction of voters for this candidate.
- We can model a sample of independent voters $X_1, \ldots, X_n$, as Bernoulli distributed random variables with parameter p.
- We interpret $X_i = 0$ as a negative vote and $X_i = 1$ as a positive vote.
- The sample proportion $\hat{p}_n = \frac{1}{n} \sum_i X_i$ is our estimator of p.

Point Estimation of a Proportion

- Then $\mathbb{E}(\hat{p}_n) = \frac{1}{n} \sum_i \mathbb{E}(X_i) = p$, and $\hat{p}_n$ is unbiased.
- The standard error *se* would be

$$se = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$$

- The estimated standard error $\hat{se}$:

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

- By the Central Limit Theorem the sampling distribution of the sample proportion converges to a Normal distribution: $\hat{p}_n \approx N(p, \hat{se}^2)$.
- This is because the sample proportion is actually the sample mean of a binary population.

Consistency

- A good estimator is expected to be unbiased and of minimum standard error.
- Unbiasedness used to receive much attention but these days is considered less important
- Many of the estimators we will use are biased.
- A reasonable requirement for an estimator is that it should converge to the true parameter value as we collect more and more data.
- A point estimator $\hat{\theta}_n$ of a parameter θ is **consistent** if it converges to the true value when the number of data in the sample tends to infinity.

- Theorem: If for an estimator $\hat{\theta}_n$, its *bias* $\to 0$ and its *se* $\to 0$ when $n \to \infty$, $\hat{\theta}_n$, it is a consistent estimator of θ.
- For example, for the sample mean $\mathbb{E}(\overline{X_n}) = \mu$, which implies that the *bias* = 0.
- Then $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$ converges to zero when $n \to \infty$.
- $\overline{X_n}$ is a consistent estimator of the mean.
- For the case of the Bernoulli experiment one has that $\mathbb{E}(\hat{p}) = p \Rightarrow bias = 0$ and $se = \sqrt{p(1-p)/n} \to 0$ when $n \to \infty$.
- Then $\hat{p}$ is a consistent estimator of *p*.

Maximum Likelihood Estimation

- The estimators we have presented so far (e.g., the sample mean, the sample proportion) are intuitive, easy to compute, and consistent.
- Maximum Likelihood Estimation (MLE) is a more general framework for estimating the **parameters** of any **parametric model**.
- In MLE, we assume that the sample data is generated by a given probability distribution (continuous or discrete) parameterized by θ and try to find the value of θ that maximizes the joint probability of the data under that distribution.
- Idea: find the parameter values of the assumed statistical model that make the observed data most probable.
- For example, we can assume that each data point is generated by $N(\mu, \sigma^2)$, then we compute the joint PDF (or PMF) of our data and find the parameter values for μ and σ that maximize that joint density (or mass).

Maximum Likelihood Estimation

- Students learning statistics often ask: how would we ever know that the distribution that generated the data is in some parametric model? [Wasserman, 2013]
- There are cases where background knowledge suggests that a parametric model provides a reasonable approximation.
- Example 1: independent binary experiments (e.g., flipping a coin or voting for a candidate) can be adequately represented with Bernoulli or Binomial distributions.
- Example 2: counts of traffic accidents are known from prior experience to follow approximately a Poisson model.
- In many other cases non parametric methods are preferable, but they are beyond the scope of this course.

Maximum Likelihood Estimation

- Let $X_1, \ldots, X_n$ be IID with PDF (or PMF) $f(x; \theta)$.
- Since we are assuming that our data samples are independent random variables, the joint density (or mass) would be the product of each PDF (or PMF):

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

- We refer to this joint density (or mass) as the **likelihood function**.
- The likelihood function is just the joint density (or mass) of the data, except that we treat it is a function of the parameter θ.
- In MLE, we turn the estimation task into an optimization problem:

$$\max_{\theta} \quad \mathcal{L}_n(\theta) \tag{1}$$

- The maximum likelihood estimator MLE, denoted by $\hat{\theta}_n$, is the value of θ that maximizes $\mathcal{L}_n(\theta)$.

Maximum Likelihood Estimation

- In many cases the log-likelihood is easier to optimize $l_n(\theta)$:

$$l_n(\theta) = \log(\mathcal{L}_n(\theta)) = \log(\prod_{i=1}^{n} f(X_i; \theta)) = \sum_{i=1}^{n} \log(f(X_i; \theta))$$

- Since the logarithm is a monotonic function, the maximum of occurs at the same value of θ.

- If the l_n is differentiable we can find $\hat{\theta}_n$ by setting the derivatives to zero:

$$\frac{\partial l_n}{\partial \theta} = 0$$

- The MLE has many good mathematical properties that go beyond the scope of this course to discuss.

- Some properties worth knowing are that the MLE is **consistent** and is **asymptotically Normal distribuited** (i.e., its sampling distribution converges to a Gaussian).

Maximum Likelihood Estimation

- Example 1: Suppose that $X_1, \ldots, X_n \sim$ Bernoulli(p).
- The probability mass function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p.
- Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$
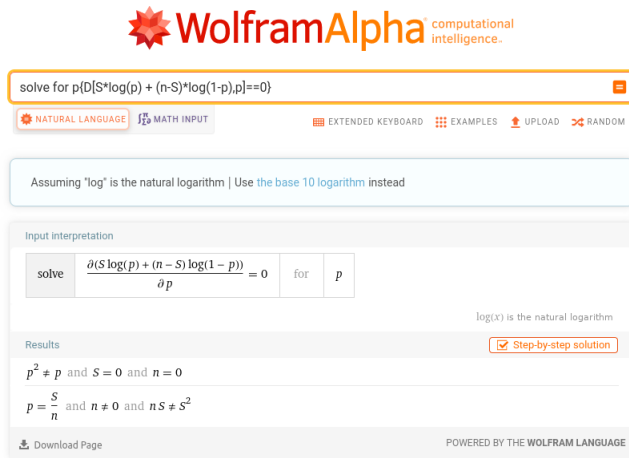
where $S = \sum_i X_i$. Hence,

$$l_n(\theta) = S \log p + (n - S) \log(1 - p).$$

- Take the derivative of $l_n(\theta)$, set it equal to 0 to find that the MLE is $\hat{p}_n = S/n$, which is the sample proportion.

Maximum Likelihood Estimation

- This can be verified by pasting the following formula into WolframAlpha[2]:

```
solve for p{D[S*log(p) + (n-S)*log(1-p),p]==0}
```


---

[2] https://www.wolframalpha.com

- Example 2: Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.
- The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:

$$\mathcal{L}_n(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right) \tag{2}$$

$$= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(X_i - \mu)^2\right) \tag{3}$$

$$= \frac{1}{\sigma^n} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \exp\left(-\frac{n(\overline{X} - \mu)^2}{2\sigma^2}\right) \tag{4}$$

- Where $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the sample mean and $S^2 = \frac{1}{n} \sum_{i=1}^{n}(X_i - \overline{X}^2)$.

Maximum Likelihood Estimation

● The last equality above follows from the fact that

$$\sum_{i=1}^{n}(X_i - \mu)^2 = nS^2 + n(\overline{X} - \mu)^2$$

which can be verified by writing

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X} + \overline{X} - \mu)^2$$

and then expanding the square.

● The log-likelihood is:

$$l_n(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X} - \mu)^2}{2\sigma^2}$$

● Solving the equations $\frac{\partial l_n(\mu,\sigma)}{\partial \mu} = 0$ and $\frac{\partial l_n(\mu,\sigma)}{\partial \sigma} = 0$

● We conclude that $\hat{\mu} = \overline{X}$ and $\hat{\sigma} = S$.

● It can be verified that these are indeed global maxima of the likelihood.

Maximum Likelihood Estimation

- Let's verify this using WolframAlpha:

```
solve for mu{D[-n*log(sigma) -(n*S^2)/(2*sigma^2)-
n*(x_b-mu)^2/(2*sigma^2),mu]==0}
```

Maximum Likelihood Estimation

- Now, if we replace μ by $\overline{X}$, the last expression of l_n cancels out.
- the value of σ can be obtained as follows:

```
solve for sigma{D[-n*log(sigma) -n*S^2/(2*sigma^2),sigma]==0}
```

Confidence Interval

- We know that the value of a point estimator **varies** from sample to sample.
- It is more reasonable to find an interval that is likely to trap the real value of the parameter in the long run.
- The general form of a confidence interval in the following:

  Confidence Interval $=$ Sample Statistic $\pm$ Margin Error

- The wider the interval the more uncertainty there is about the value of the parameter.

Confidence Interval

Definition

- A **confidence interval** for an unknown population parameter θ with a **confidence level** $1 - \alpha$, is an interval $C_n = (a, b)$ where:

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha$$

- In addition $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data.
- The α value is known as the **significance** level, generally taken as 0.05, which is equivalent to working with a confidence level of 95%.
- Significance can be interpreted as the probability of being wrong.

- There is a lot of **confusion** about how to interpret a confidence interval.
- A confidence interval is not a probability statement about θ since θ is a fixed quantity in frequentist inference, not a random variable
- One way to interpret them is to say that if we repeat the **same experiment** many times, the interval will contain the value of the parameter $(1 - \alpha)\%$ of the times.
- This interpretation is correct, but we rarely repeat the same experiment several times.
- A better interpretation: one day I collect data I create a 95% confidence interval for a parameter θ_1. Then on day 2, I do the same for a parameter θ_2 and so repeatedly n times. The 95% of my intervals will contain the actual values of the parameters.

Confidence Interval

- Later in the course, we will discuss Bayesian methods in which we treat θ as if it were a random variable and we do make probability statements about θ.
- In particular, we will make statements like "the probability that θ is in C_n, given the data, is 95 percent."
- However, these Bayesian intervals refer to degree-of-belief probabilities.
- These Bayesian intervals will not, in general, trap the parameter 95 percent of the time.

- We have n independent observations $X_1, \ldots, X_n$ (IID) of distribution $N(\mu, \sigma^2)$.
- Suppose μ is **unknown** but σ^2 is **known**.
- We know that $\overline{X_n}$ is an unbiased estimator of μ.
- By the law of large numbers we know that the distribution of $\overline{X_n}$ is concentrated around μ when n is large.
- By the CLT we know that

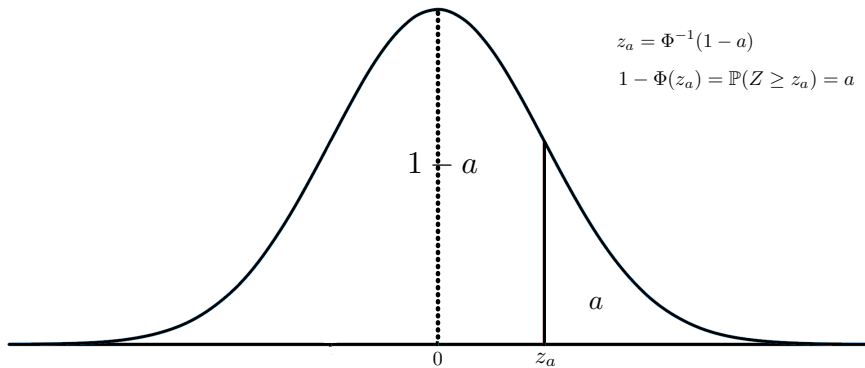$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

when n is large.

- We want to find an interval $C_n = (\mu_1, \mu_2)$ with confidence level $1 - \alpha$:
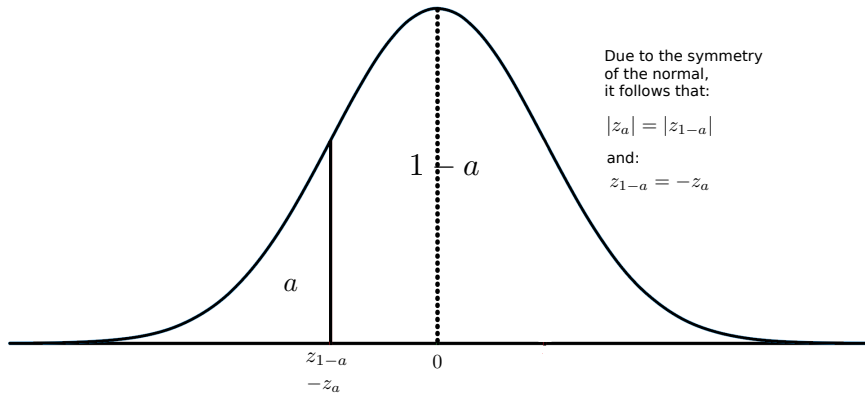
$$\mathbb{P}(\mu_1 \leq \mu \leq \mu_2) = 1 - \alpha$$

- Let $z_a = \Phi^{-1}(1 - a)$, with $a \in [0, 1]$ where Φ^{-1} is the quantile function of a standardized normal.
- This is equivalent to saying that z_a is the value such that $1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$.
- By symmetry of the normal distribution: $z_{\alpha/2} = -z_{(1-\alpha/2)}$.

$$z_a = \Phi^{-1}(1-a)$$
$$1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$$

$1 - a$

a

0 $\quad z_a$

Due to the symmetry
of the normal,
it follows that:

$$|z_a| = |z_{1-a}|$$

and:

$$z_{1-a} = -z_a$$

$1 - a$

a

z_{1-a}
$-z_a$

0

From the diagram, we can say that:

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

$1 - \alpha$

$\alpha/2$

$\alpha/2$

$-z_{\alpha/2}$ 0 $z_{\alpha/2}$
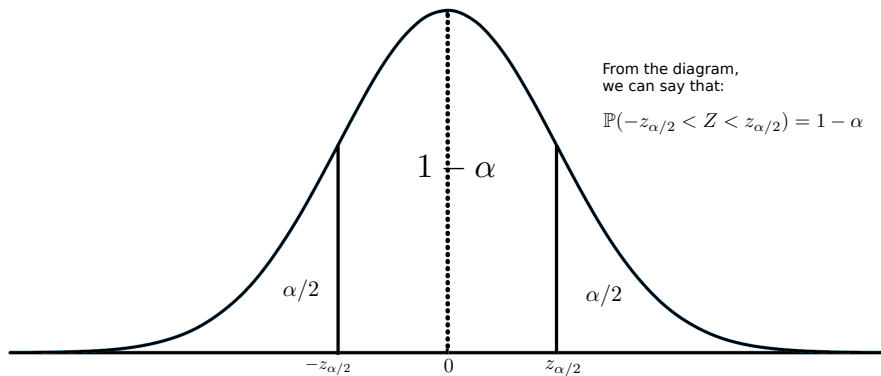
Confidence Interval

- The confidence interval for μ is:

$$C_n = (\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Then $z_{\alpha/2}$ tells us how many times we have to multiply the **standard error** to build the interval.
- The smaller the value of α the larger the value of $z_{\alpha/2}$ and hence the wider the interval.
- Proof:

$$
\begin{aligned}
\mathbb{P}(\mu \in C_n) &= \mathbb{P}(\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\
&= \mathbb{P}(-z_{\alpha/2} < \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) \\
&= \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= 1 - \alpha
\end{aligned}
$$

- Since $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ we can use the quantile function of the normal to calculate confidence intervals in R.

```
> alpha <- 0.05
> xbar <- 5
> sigma <- 2
> n <- 20
> se <-sigma/sqrt(n)
> error <- qnorm(1-alpha/2)*se
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.123477
> right
[1] 5.876523
>
```

T Distribution

- In practice, if we do not know μ we are unlikely to know σ.
- If we estimate σ using *s*, confidence intervals are better build using the **T-student** distribution, especially when the sample size is small.
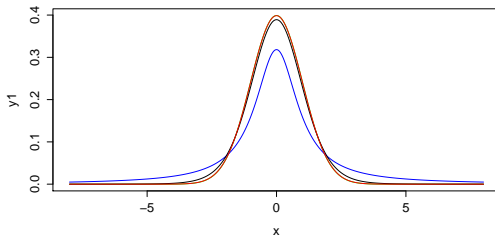
T Distribution

- An R.V. has distribution *t* with *k* degrees of freedom when it has the following PDF:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1 + \frac{t^2}{k})^{(k+1)/2}}$$

- When $k = 1$ it is called **Cauchy** distribution.
- When $k \to \infty$ it converges to a standardized normal distribution.
- The t-distribution has wider tails than the normal distribution when it has few degrees of freedom.

T Distribution

```
x<-seq(-8,8,length=400)
y1<-dnorm(x)
y2<-dt(x=x,df=1)
y3<-dt(x=x,df=10)
y4<-dt(x=x,df=350)
plot(y1~x,type="l",col="green")
lines(y2~x,type="l",col="blue")
lines(y3~x,type="l",col="black")
lines(y4~x,type="l",col="red")
```

- Let $s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$ we have:

$$T = \frac{\overline{X_n} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Let $t_{n-1,a} = \mathbb{P}(T > a)$, equivalent to the quantile function qt evaluated at $(1-a)$.
- The resulting confidence interval is:

$$C_n = (\overline{X_n} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \overline{X_n} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$$

- Since the tails of the t distribution are wider when n is small, the resulting confidence intervals are wider.

T-Distribution Confidence Interval

- Let's calculate a confidence interval for the mean of `Petal.Length` of the **Iris** data with 95% confidence.

```
>data(iris)
>alpha<-0.05
>n<-length(iris$Petal.Length)
>xbar<-mean(iris$Petal.Length)
>xbar
[1] 3.758
>s<-sd(iris$Petal.Length)
>se<-s/sqrt(n)
>error<-qt(p=1-alpha/2,df=n-1)*se
>left<-xbar-error
>left
[1] 3.473185
>right<-xbar+error
>right
[1] 4.042815
```

- Another way:

```
>test<-t.test(iris$Petal.Length,conf.level=0.95)
>test$conf.int
[1] 3.473185 4.042815
```

- Suppose we want to compute the proportion of subjects who will vote for a candidate and we also want a confidence interval for the estimated proportion.
- As showed earlier, the sampling distribution of a the sample proportion follows a Normal distribution.
- The confidence interval C_n for a proportion is:

$$C_n = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Confidence Interval for a Population Proportion

- Example: $1,219$ respondents indicated that they would vote for candidate A in a survey of $3,532$ people.
- Compute a 95% confidence interval for the proportion of voters:

$$\hat{p} = \frac{1219}{3235} = 0.345$$

- and $z_{\alpha/2} = 1.96$

```
> qnorm(1-0.025)
[1] 1.959964
```

$$C_n = 0.345 \pm 1.96\sqrt{\frac{0.345(1-0.345)}{3532}} = (0.329, 0.361)$$

- In R:

```
> prop<-prop.test(1219,3532,correct=FALSE)
> prop$conf.int
[1] 0.3296275 0.3609695
```
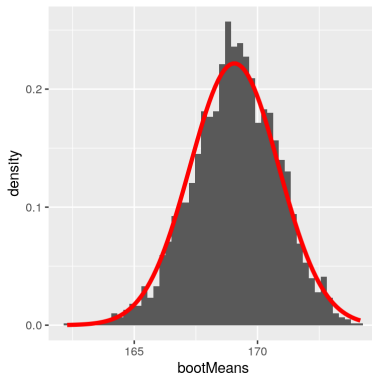
The Boostrap

- We have used our knowledge of the sampling distribution of the mean to compute the standard error of the mean.
- But what if we can't assume that the estimates are normally distributed, or we don't know their distribution?
- The idea of the bootstrap is to use the data themselves to estimate an answer.
- The bootstrap method was conceived by Bradley Efron of the Stanford Department of Statistics, who is one of the world's most influential statisticians.
- The idea behind the bootstrap is that we repeatedly sample from the actual dataset.
- We sample with replacement, such that the same data point will often end up being represented multiple times within one of the samples.

- We then compute our statistic of interest on each of the bootstrap samples, and use the distribution of those estimates as our sampling distribution.
- In a sense, we treat our particular sample as the entire population, and then repeatedly sample with replacement to generate our samples for analysis.
- This makes the assumption that our particular sample is an accurate reflection of the population, which is probably reasonable for larger samples but can break down when samples are smaller.
- This technique can be used to estimate the standard error of any statistic and to obtain a confidence interval (CI) for it.

- Let's use the bootstrap to estimate the sampling distribution of the mean:

- The histogram shows the distribution of means across bootstrap samples, while the red line shows the normal distribution based on the sample mean and standard error.

The Boostrap

- From the figure, we can see that the distribution of means across bootstrap samples is fairly close to the theoretical estimate based on the assumption of normality.
- It doesn't make much sense to use bootstrap for the sample mean because we know the theoretical shape of the sampling distribution.
- The bootstrap would more often be used to generate standard errors and confidence intervals for estimates of statistics (e.g., the median, the standard deviation) where we know or suspect that the normal distribution is not appropriate.
- Bootstrap is especially useful when CI doesn't have a closed form, or it has a very complicated one.

The Boostrap

- The following function implements Bootstrap confidence intervals for a generic R function:

```r
myboot<-function(x,fun,nRuns,sampleSize,alpha){
values<-vector()
for(i in 1:nRuns){
  samp.i<-sample(x, size = sampleSize, replace = T)
  values[i]<-fun(samp.i)
}
point.est <-fun(x)
se <- sd(values)
l.CI <- quantile(values, alpha/2)
u.CI <- quantile(values, 1-alpha/2)


return(c("Point Estimate"=point.est,
         "Standard error"=se,
         "Lower CI limit" = l.CI,
         "Upper CI limit" = u.CI))
}
```

The Boostrap

- The lower and upper CI limits are obtained from the $\alpha/2$ and $(1 - \alpha/2)$ sample quantiles.
- Let's compute a Bootstrap CI for the median of Petal.Length using 6000 runs, a sample size of 64, and a level of significance of 0.05:

```
> myboot(iris$Petal.Length,median,6000,64,0.05)
      Point Estimate       Standard error   Lower CI limit.2.5%
          4.3500000            0.2779297             3.7500000
Upper CI limit.97.5%
          4.7000000
```

Conclusions

- We have introduced many important concepts in statistical inference, in particular, the frequentist approach.
- We have studied estimators and their sampling distributions.
- We introduced a general technique for calculating estimators called maximum likelihood.
- We studied how to measure the uncertainty of an estimator using a confidence interval.
- We have introduced a method of approximating confidence intervals by means of simulations called bootstrap.

Barnes, F. (1995).
Can you trust those polls.

Poldrack, R. A. (2019).
Statistical thinking for the 21st century.
https://statsthinking21.org/.

Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.

Watkins, A. E., Scheaffer, R. L., and Cobb, G. W. (2010).
Statistics: from data to decision.
John Wiley & Sons.