BS1820: Maths and Statistics Foundations for Analytics

Statistics 1

Zhe Liu

Imperial College Business School

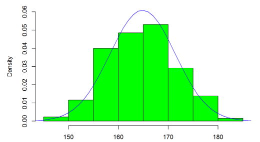
Email: zhe.liu@imperial.ac.uk

Outline

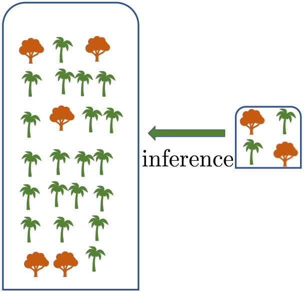
1.1 Introduction

Descriptive Statistics

Heights of the trees in the forest



Inferential Statistics



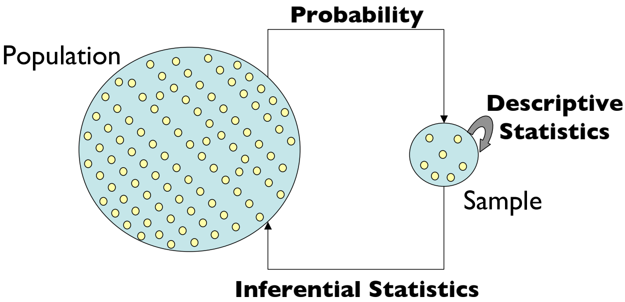
Descriptive Statistics:

- Summarize the **sample** using statistics (e.g., mean, standard deviation, etc.)

Inferential Statistics:

- Infer properties of **population** from random samples

1.1 Introduction



Descriptive Statistics:

- Summarize the **sample** using statistics (e.g., mean, standard deviation, etc.)

Inferential Statistics:

- Infer properties of **population** from random samples

1.1 Introduction

We will learn the following:

- **Parameter estimation**:
  Use sample data to calculate a "best estimate" of an unknown (underlying) population parameter. **E.g.** mean, variance

- **Confidence interval**:
  How closely the sample estimate matches the true parameter of population

- **Hypothesis testing**:
  Determine whether sample outcomes could lead to a rejection of a hypothesis under a pre-specified significance level

- **Regression analysis**:
  - Estimate the relationship between dependent and independent variables
  - Used for prediction / forecasting

1.2 Parameter Estimation

One major problem in statistics is the estimation of unknown parameters.

E.g. Suppose we have observations $X_1, \ldots, X_n$ from n i.i.d. Bernoulli trials. How do we estimate $p := \mathsf{P}(X = 1)$ from these observations?

E.g. Suppose we have observations $Y_1, \ldots, Y_n$ from a $\mathcal{N}(\mu, \sigma^2)$ distribution. How do we estimate μ and σ from these observations?

Definition. An **estimator** $\widehat{\theta}$ is a function of a random sample that we use to estimate the unknown (true) parameter θ. An **estimate** is a particular realization of an estimator based on the observed sample.

Remark: An **estimator** is a function (statistic) and an **estimate** is a number.

1.2 Parameter Estimation

Let us highlight once more that an estimator refers to a statistic (function) that is used as a tool to estimate a parameter based on the data collected.

Example: Suppose we use an estimator $\widehat{\mu}$ to estimate the mean height of trees in a forest.

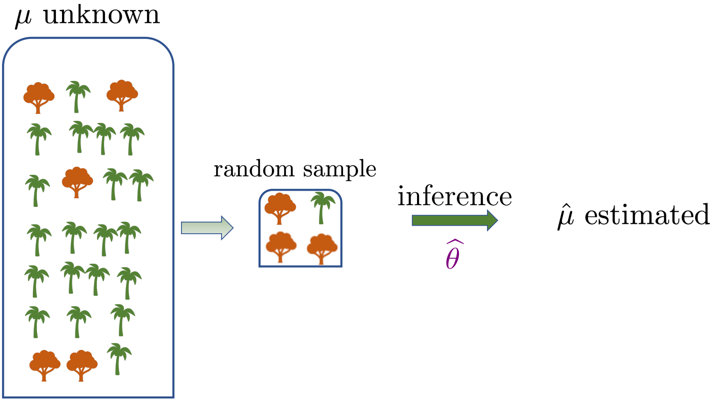
We feed our estimator (a function/statistic) with a random sample and generate an estimate ($\hat{\mu}$) of the mean height while the true mean is μ.

Let the height of trees in the forest be denoted by RV X and $X_1, \ldots, X_n$ is a random sample of size n. Then we may define our estimator as

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

1.2 Parameter Estimation

X is a RV, hence $\widehat{\mu}(\mathbf{X})$ as a function of $X_1, \ldots, X_n$ is also a RV.
But the point estimate $\widehat{\mu}(\mathbf{x})$ is a number.



Remark: One can use different estimators to estimate the same parameter.

The question is what is a **good** estimator?

1.3 Good Estimators

We use **mean squared error** (MSE) to evaluate the **quality** of an estimator $\widehat{\theta}$.

$$\mathsf{MSE}(\widehat{\theta}) = \underbrace{\mathsf{Var}(\widehat{\theta})}_{\textbf{Variance}} + (\underbrace{\mathsf{E}[\widehat{\theta}] - \theta}_{\textbf{Bias}})^2 = \mathsf{Var}(\widehat{\theta}) + \mathsf{Bias}(\widehat{\theta})^{\mathbf{2}}$$

Good estimators are:

- **Unbiased**: expectation equals true parameter

$$\mathsf{Bias}(\widehat{\theta}) := \mathsf{E}[\widehat{\theta}] - \theta = 0$$

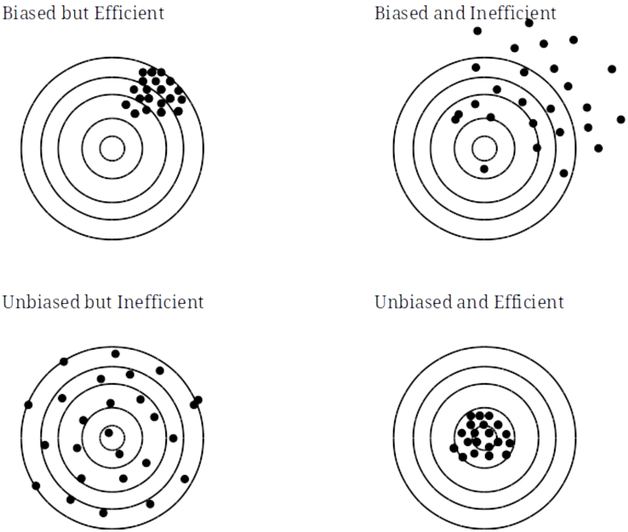
- **Efficient**: has lower variance

$$\mathsf{Var}(\widehat{\theta}) \downarrow$$

The “optimal” one is the “**minimum-variance unbiased** (MVUE) estimator”.

1.3 Good Estimators

There is often a trade-off between **unbiasedness** and **efficiency**!



1.4 Common Estimation Methods

We want to estimate **parameters** θ with observed **data** D. By Bayes' rule:

$$\underbrace{\mathsf{P}(\theta \mid D)}_{\textbf{posterior}} = \frac{\overbrace{\mathsf{P}(D \mid \theta)}^{\textbf{likelihood}} \times \overbrace{\mathsf{P}(\theta)}^{\textbf{prior}}}{\underbrace{\mathsf{P}(D)}_{\textbf{evidence}}}$$

There are two major estimation methods:

Maximum likelihood

- Parameters are fixed but unknown
- Focus on the likelihood $\mathsf{P}(D \mid \theta)$ – the probability of observing the data
- Choose θ that maximizes the likelihood

Bayesian

- Parameters are RVs with some dist.
- Convert prior dist. to posterior dist. using observed data
- All about updating belief of the dist.

1.5 Maximum Likelihood Estimation

Let the observed data be $\mathbf{x} = \{x_1, \ldots, x_n\}$. Define the **likelihood function** as the "likelihood" of observing data $\mathbf{x}$ given underlying parameter θ:

$$L(\theta) := \begin{cases} \mathrm{P}(\mathbf{x} \mid \theta) & \text{for discrete RV} \\ f(\mathbf{x} \mid \theta) & \text{for continuous RV} \end{cases}$$

and the **log-likelihood function**:

$$l(\theta) := \log L(\theta).$$

Remarks:

- The lhs is a function of θ, while the rhs is a conditional PMF/PDF.
- Since $\log(\cdot)$ is an increasing function, maximizing the log-likelihood function is equivalent to maximizing the likelihood function.
- Our goal is to find $\widehat{\theta}$ that **maximizes** the likelihood of observing data $\mathbf{x}$:

$$\max_{\theta} \; L(\theta) \quad \text{or} \quad \max_{\theta} \; l(\theta)$$

1.5 Maximum Likelihood Estimation

Example: Let $X \sim \text{Ber}(p)$ and $\mathbf{x} = \{x_1, \ldots, x_n\}$ observed data of n i.i.d. draw. What is the maximum likelihood estimator $\widehat{p}$?

1.5 Maximum Likelihood Estimation

Example: Let $X \sim \mathsf{Ber}(p)$ and $\mathbf{x} = \{x_1, \ldots, x_n\}$ observed data of n i.i.d. draw. What is the maximum likelihood estimator $\widehat{p}$?

Answer: First write down the (log)-likelihood function:

$$\begin{aligned}
L(p) &= \mathsf{P}(\mathbf{x} \mid p) \\
&= \mathsf{P}(X_1 = x_1, \ldots, X_n = x_n \mid X \sim \mathsf{Ber}(p)) \\
\text{(independence)} \quad &= \prod_{i=1}^{n} \mathsf{P}(X_i = x_i \mid X_i \sim \mathsf{Ber}(p)) \\
&= \prod_{i=1}^{n} p^{x_i} (1-p)^{(1-x_i)}
\end{aligned}$$

$$l(p) = \log L(p) = \left(\sum_{i=1}^{n} x_i\right) \log p + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p)$$

Maximizing the log-likelihood function, we get

$$\widehat{p} = \underset{p}{\operatorname{argmax}}\ l(p) = \frac{\sum_{i=1}^{n} x_i}{n}.$$

What does this mean?

1.6 Bayesian Estimation

Example: Suppose I have 3 coins, with probabilities of observing heads by coin $i = 1, 2, 3$ being 0.25, 0.5 and 0.75, respectively. I gave you a coin and you flipped it once, observing a head. What is the probability that I have given you coin 3?

- Let $\theta \in \{1, 2, 3\}$ be the coin I gave you and $X = 1$ for observing a head and $X = 0$ for observing a tail.
- Then question is: having observed $x = 1$, what is $\mathsf{P}(\theta = 3 \mid x = 1)$?
- What if we use maximum likelihood estimation?

$$\mathsf{P}(x = 1 \mid \theta = 1) = 0.25,\ \mathsf{P}(x = 1 \mid \theta = 2) = 0.5,\ \mathsf{P}(x = 1 \mid \theta = 3) = 0.75$$

- Let's think in the Bayesian way.
  (1) **Prior belief**: $\mathsf{P}(\theta = i) = 1/3,\ i = 1, 2, 3.$
  (2) **Likelihood**: $\mathsf{P}(x = 1 \mid \theta)$ same as above.
  (3) **Posterior belief**: how does our belief about receiving coin 3 change?

1.6 Bayesian Estimation

Example: Suppose I have 3 coins, with probabilities of observing heads by coin $i = 1, 2, 3$ being 0.25, 0.5 and 0.75, respectively. I gave you a coin and you flipped it once, observing a head. What is the probability that I have given you coin 3?

– Recall Bayes' rule:

$$\underbrace{\mathsf{P}(\theta \mid D)}_{\textbf{posterior}} = \frac{\overbrace{\mathsf{P}(D \mid \theta)}^{\textbf{likelihood}} \times \overbrace{\mathsf{P}(\theta)}^{\textbf{prior}}}{\underbrace{\mathsf{P}(D)}_{\textbf{evidence}}} \quad \Rightarrow \quad \textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

1.6 Bayesian Estimation

Example: Suppose I have 3 coins, with probabilities of observing heads by coin $i = 1, 2, 3$ being 0.25, 0.5 and 0.75, respectively. I gave you a coin and you flipped it once, observing a head. What is the probability that I have given you coin 3?

– Recall Bayes' rule:

$$\underbrace{\mathsf{P}(\theta \mid D)}_{\textbf{posterior}} = \frac{\overbrace{\mathsf{P}(D \mid \theta)}^{\textbf{likelihood}} \times \overbrace{\mathsf{P}(\theta)}^{\textbf{prior}}}{\underbrace{\mathsf{P}(D)}_{\textbf{evidence}}} \quad \Rightarrow \quad \textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

coin θ	**prior** $\mathsf{P}(\theta)$	**likelihood** $\mathsf{P}(x = 1 \mid \theta)$	likelihood×prior $\mathsf{P}(x = 1 \mid \theta)\mathsf{P}(\theta)$	**posterior** $\frac{\mathsf{P}(x=1\mid\theta)\mathsf{P}(\theta)}{\mathsf{P}(x)}$
1	$1/3$	0.25		
2	$1/3$	0.50		
3	$1/3$	0.75		
sum	1			

1.6 Bayesian Estimation

Example: Suppose I have 3 coins, with probabilities of observing heads by coin $i = 1, 2, 3$ being 0.25, 0.5 and 0.75, respectively. I gave you a coin and you flipped it once, observing a head. What is the probability that I have given you coin 3?

– Recall Bayes' rule:

$$\underbrace{\mathsf{P}(\theta \mid D)}_{\textbf{posterior}} = \frac{\overbrace{\mathsf{P}(D \mid \theta)}^{\textbf{likelihood}} \times \overbrace{\mathsf{P}(\theta)}^{\textbf{prior}}}{\underbrace{\mathsf{P}(D)}_{\textbf{evidence}}} \quad \Rightarrow \quad \textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

coin θ	**prior** $\mathsf{P}(\theta)$	**likelihood** $\mathsf{P}(x=1 \mid \theta)$	likelihood$\times$prior $\mathsf{P}(x=1 \mid \theta)\mathsf{P}(\theta)$	**posterior** $\frac{\mathsf{P}(x=1 \mid \theta)\mathsf{P}(\theta)}{\mathsf{P}(x)}$
1	$1/3$	0.25	0.0825	
2	$1/3$	0.50	0.1650	
3	$1/3$	0.75	0.2475	
sum	1		0.495	

1.6 Bayesian Estimation

Example: Suppose I have 3 coins, with probabilities of observing heads by coin $i = 1, 2, 3$ being 0.25, 0.5 and 0.75, respectively. I gave you a coin and you flipped it once, observing a head. What is the probability that I have given you coin 3?

- Recall Bayes' rule:

$$\underbrace{\mathsf{P}(\theta \mid D)}_{\textbf{posterior}} = \frac{\overbrace{\mathsf{P}(D \mid \theta)}^{\textbf{likelihood}} \times \overbrace{\mathsf{P}(\theta)}^{\textbf{prior}}}{\underbrace{\mathsf{P}(D)}_{\textbf{evidence}}} \quad \Rightarrow \quad \textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

coin θ	**prior** $\mathsf{P}(\theta)$	**likelihood** $\mathsf{P}(x=1 \mid \theta)$	likelihood×prior $\mathsf{P}(x=1 \mid \theta)\mathsf{P}(\theta)$	**posterior** $\frac{\mathsf{P}(x=1\mid\theta)\mathsf{P}(\theta)}{\mathsf{P}(x)}$
1	1/3	0.25	0.0825	0.167
2	1/3	0.50	0.1650	0.333
3	1/3	0.75	0.2475	0.500
sum	1		0.495	1

- We can repeat and update your belief

1.7 Confidence Intervals

We have so far discussed the **point estimates** for population parameters.

It is often better to provide a range of plausible values for the parameter to allow for a margin of error, this results in **interval estimates**.

A commonly used interval estimate is **confidence interval** (CI).

$$\text{CI} = \text{point estimate} \pm \text{margin of error}$$

We want to find intervals that are very likely to **cover** the true parameter.

1.7 Confidence Intervals

Definition. Given random sample $\mathbf{X}$, a $100(1-\alpha)\%$ **confidence interval** (CI) for the unknown parameter θ is a **random interval** $[L(\mathbf{X}), U(\mathbf{X})]$ such that

$$\mathsf{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]) = 1-\alpha.$$

Remark (Important!):

1. The interval is random because it is based on a random sample $\mathbf{X} = \{X_1, \ldots, X_n\}$.
2. So, if we construct many such intervals based on different random samples, then approximately $100(1-\alpha)\%$ of them will cover the true value of θ.
3. α is called the **significance level**. **E.g.** $\alpha = 0.01, 0.05, 0.1$

1.7 Confidence Intervals

Interpretation of CI:

Suppose a 95% confidence interval for parameter θ is constructed as $[L, U]$. [True/False]

1. If we draw a large sample $\{x_1, \ldots, x_n\}$, then approximately 95% of the sample data will fall into $[L, U]$.
2. There is a 95% probability that θ lies in $[L, U]$.
3. 95% of the time, the interval generated according to this "recipe" will cover the true parameter θ.

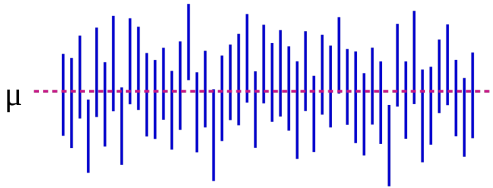
1.7 Confidence Intervals

Interpretation of CI:

Suppose a 95% confidence interval for parameter θ is constructed as $[L, U]$.
[True/False]

1. If we draw a large sample $\{x_1, \ldots, x_n\}$, then approximately 95% of the sample data will fall into $[L, U]$.
2. There is a 95% probability that θ lies in $[L, U]$.
3. 95% of the time, the interval generated according to this "recipe" will cover the true parameter θ.

Remark: We can talk about the probability that a random CI will contain the true parameter, not the probability that a specific CI contains the parameter – once constructed, it either does or does not!



1.7 Confidence Intervals

We now construct CIs for

1. (Normally distributed) **population mean** μ
   - when population standard deviation σ is known
   - when population standard deviation σ is unknown

2. (Arbitrarily distributed but large) **population mean** μ

3. **Population proportion** p

1.8 CI for (Normal) Population Mean

Suppose $X_1, \ldots, X_n$ are i.i.d. random sample of $\mathcal{N}(\mu, \sigma^2)$. Sample mean is a random variable given by

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

– If population standard deviation σ is **known**:

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Hence we can construct the $100(1-\alpha)\%$ CI from:

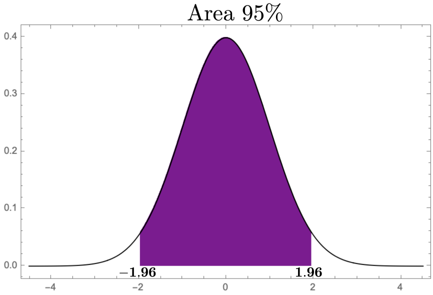
$$1 - \alpha = \mathsf{P}\left(\bar{X} - z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{\sigma}{\sqrt{n}}\right) = \mathsf{P}(-z \leq Z \leq z)$$

where

$$z = z_{\alpha/2} = \mathsf{P}(Z \geq z_{\alpha/2}) = -\Phi^{-1}\left(\frac{\alpha}{2}\right) = \texttt{-qnorm}(\alpha\texttt{/2})$$

1.8 CI for (Normal) Population Mean

Example: $\alpha = 0.05$ corresponds to a 95% CI.



$$0.95 = \mathsf{P}\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \quad \Rightarrow \quad \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Often used z-values:

α	confidence level	$z_{\alpha/2}$
0.1	90%	1.645
0.05	95%	1.96
0.01	99%	2.58

1.8 CI for (Normal) Population Mean

– If population standard deviation σ is **unknown**:

Using sample standard deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

we have

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \text{ (t-distribution with $n-1$ degree of freedom)}$$

Hence we can construct the $100(1-\alpha)\%$ CI from:

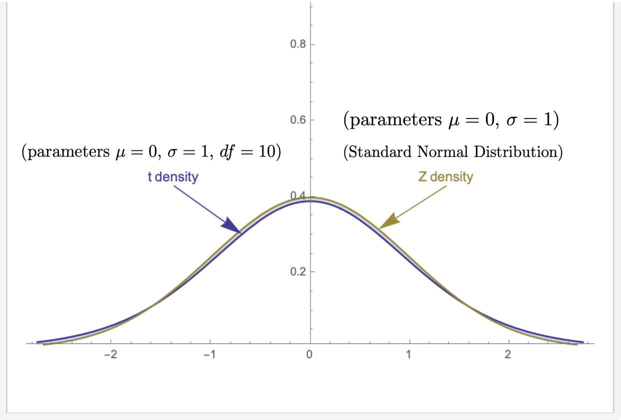
$$1 - \alpha = \mathsf{P}\left(\bar{X} - t\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t\frac{S}{\sqrt{n}}\right) = \mathsf{P}(-t \leq T \leq t)$$

where

$$t = t_{n-1}^{\alpha/2} = \mathsf{P}(T \geq t_{n-1}^{\alpha/2}) = \texttt{qt(1-}\alpha\texttt{/2,n-1)} \quad (\alpha/2 \text{ upper quantile})$$

1.8 CI for (Normal) Population Mean

- t-value is derived from the t-distribution, with parameters: μ, σ and df (dof).
- Shape depends on df $= n - 1$
- When n is large, t-value is close to z-value.



1.9 CI for (Large) Population Mean

Recall by CLT, for large n (e.g. $n > 30$),

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

Hence using S for σ, we can construct an approximated $100(1 - \alpha)\%$ CI from:

$$1 - \alpha = \mathsf{P}\left(\bar{X} - z\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + z\frac{S}{\sqrt{n}}\right) \approx \mathsf{P}(-z \leq Z \leq z)$$

where

$$z = z_{\alpha/2} = \mathsf{P}(Z \geq z_{\alpha/2}) = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$$

1.10 CI for Population Proportion

Suppose we wish to construct a $100(1-\alpha)\%$ CI for a proportion p.

E.g. $p = \mathsf{P}(\text{Head})$ – where coin is possibly biased.

Data: n i.i.d. coin tosses $X_1, \ldots, X_n$ with

$$X_i := \begin{cases} 1, & \text{if } i^{th} \text{ coin toss is head} \\ 0, & \text{otherwise.} \end{cases}$$

We have found the maximum likelihood estimator

$$\widehat{p} := \frac{\sum_{i=1}^{n} X_i}{n}$$

and can show (how?) that

$$\mathsf{E}(\widehat{p}) = p, \quad \mathsf{Var}(\widehat{p}) = \frac{p(1-p)}{n}$$

1.10 CI for Population Proportion

By CLT, for large n we have

$$\widehat{p} \approx \mathcal{N}\left(p,\, \frac{p(1-p)}{n}\right) \quad \Rightarrow \quad Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \approx \mathcal{N}(0,1)$$

Hence we can construct the $100(1-\alpha)\%$ CI from:

$$1 - \alpha = \mathsf{P}\left(\widehat{p} - z\sqrt{\frac{p(1-p)}{n}} \le p \le \widehat{p} + z\sqrt{\frac{p(1-p)}{n}}\right) \approx \mathsf{P}(-z \le Z \le z)$$

Approximating p by $\widehat{p}$, we get an approximate $100(1-\alpha)\%$ CI for p:

$$\left[\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right].$$

Construction of such CIs arises often in survey sampling.

E.g. What percentage, p, of the population will vote for Trump?

1.11 Sample Size Requirement

Given a confidence level $100(1-\alpha)\%$ and **margin of error** w (half width of the CI), what should be the minimum sample size n?

– Normal population mean (σ known): $w = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\dfrac{z_{\alpha/2}\sigma}{w}\right)^2$

– Normal population mean (σ unknown): $w = t_{n-1}^{\alpha/2}\dfrac{S}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\dfrac{t_{n-1}^{\alpha/2}S}{w}\right)^2$

– Large population mean: $w \approx z_{\alpha/2}\dfrac{S}{\sqrt{n}} \quad \Rightarrow \quad n \approx \left(\dfrac{z_{\alpha/2}S}{w}\right)^2$

– Population proportion: $w \approx z_{\alpha/2}\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}} \quad \Rightarrow \quad n \approx \left(\dfrac{z_{\alpha/2}}{w}\right)^2 \widehat{p}(1-\widehat{p})$

Remark: larger **sample size** $\Leftrightarrow$ smaller **margin of error**.